nature genetics

ARTICLES

https://doi.org/10.1038/s41588-019-0484-x

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

Leo Speidel[1], Marie Forest[2], Sinan Shi[1] and Simon R. Myers[1,3]*

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 have 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 1,000 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 has previously been possible. We identify multiple regions under strong positive selection, and multi-allelic traits including hair color, body mass index 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 the underlying genealogical history, which records the historical coalescence, recombination and mutation events producing the observed variation patterns. In practice, several key existing approaches (for example, refs. [1,2]) leverage an underlying coalescent model, which provides a flexible modeling framework and is the limiting behavior of a variety of finite-population models[3,4]. 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[5]. Other approaches[6–11] use more heuristic coalescent approximations, sometimes reducing accuracy: regardless,the existing methods that have been published 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[1,2], while for larger datasets diverse approaches based on data summaries[12–14] or downsampling[15,16] have predominated. In humans, such tools have detected genetic structure and admixture in good agreement with independent evidence[13,17], changes in population size[15,18–20] and introgression with archaic groups, including Neanderthals[21].

We developed a scalable method, Relate, to estimate genome-wide genealogies (Fig. 1; see Methods, and also https://myers-group.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, while being thousands of times faster.

We demonstrate the utility of a genealogy-based analysis by applying Relate to 4,956 haplotypes of the 1,000 Genomes Project (1,000GP) dataset[22,23]. 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[24]. 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)[25] and identifies previously unreported genomic regions under strong positive selection. We find a remarkable enrichment of single nucleotide polymorphisms (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 nonsymmetric 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[26], but leveraging knowledge of ancestral and derived status at each 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

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Université du Québec à Montréal, Montréal, Canada. [3]Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. *e-mail: myers@stats.ox.ac.uk

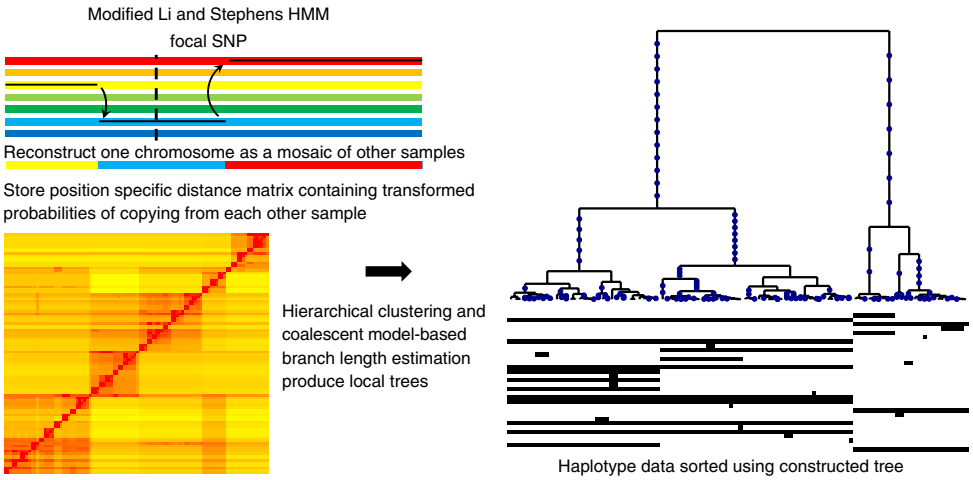

Fig. 1 | Relate method overview. Our method applies a version of the Li and Stephens HMM[15], modified to take ancestral and derived states into account, to calculate at a focal SNP (dotted vertical line) a position-specific distance matrix d (bottom left). Each entry d_{ij} of this matrix stores the 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 (see Methods and 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.

time by only rebuilding trees at a subset of sites along the genome (see Methods and Supplementary Figs. 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 subpopulations 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[27]. We compared the performance to ARGweaver[2], 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 more than four orders of magnitude faster than ARGweaver, for the cases where we were able to apply the latter, and also much faster than RENT+[11] (Fig. 2a,b). Our approach scales linearly in sequence length and quadratically in sample size N, enabling genealogical inference for 10,000 human samples, for example, genome-wide using a compute cluster.

To evaluate accuracy, we compared, 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 (Fig. 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 Fig. 3). Other accuracy measures yielded similar improvements (Supplementary Note). Relate identified repeat mutations and variable mutation rates and was robust to computational rephasing of haplotypes (Supplementary Fig. 4). We next compared Relate's inferred population sizes to those from applying two leading specialist approaches, MSMC[19] and SMC++[20]. For multiple previously tested[19,20] scenarios including oscillating population sizes and bottlenecks similar to those observed in out-of-Africa events of modern humans, Relate obtained more accurate estimates, particularly in the recent past (Fig. 2e and Supplementary Fig. 3). While Relate assumes a single population when estimating branch lengths, when applied to a combined sample from two diverged populations it still performed well in recovering their distinct population histories and estimating their split time(s) (Fig. 2f,g).

Genome-wide human genealogies. We applied Relate to 2,478 1,000GP 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 d (Supplementary Table 1). We found that 86% of all SNPs (>96% of SNPs at >0.2% derived-allele frequency (DAF)) map uniquely to trees, falling to 76% for CpG dinucleotides, which are known to possess strongly elevated mutation rates (Supplementary Fig. 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 Fig. 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 observed a clear out-of-Africa bottleneck for Eurasian populations peaking at 20,000–40,000 YBP (CHB, Chinese in Beijing; 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 60,000 YBP (Fig. 3a,b). This is consistent with recent studies[15,28] 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 (Fig. 3c). We also detected, and dated, very recent separations <10,000 YBP, such as between CHB

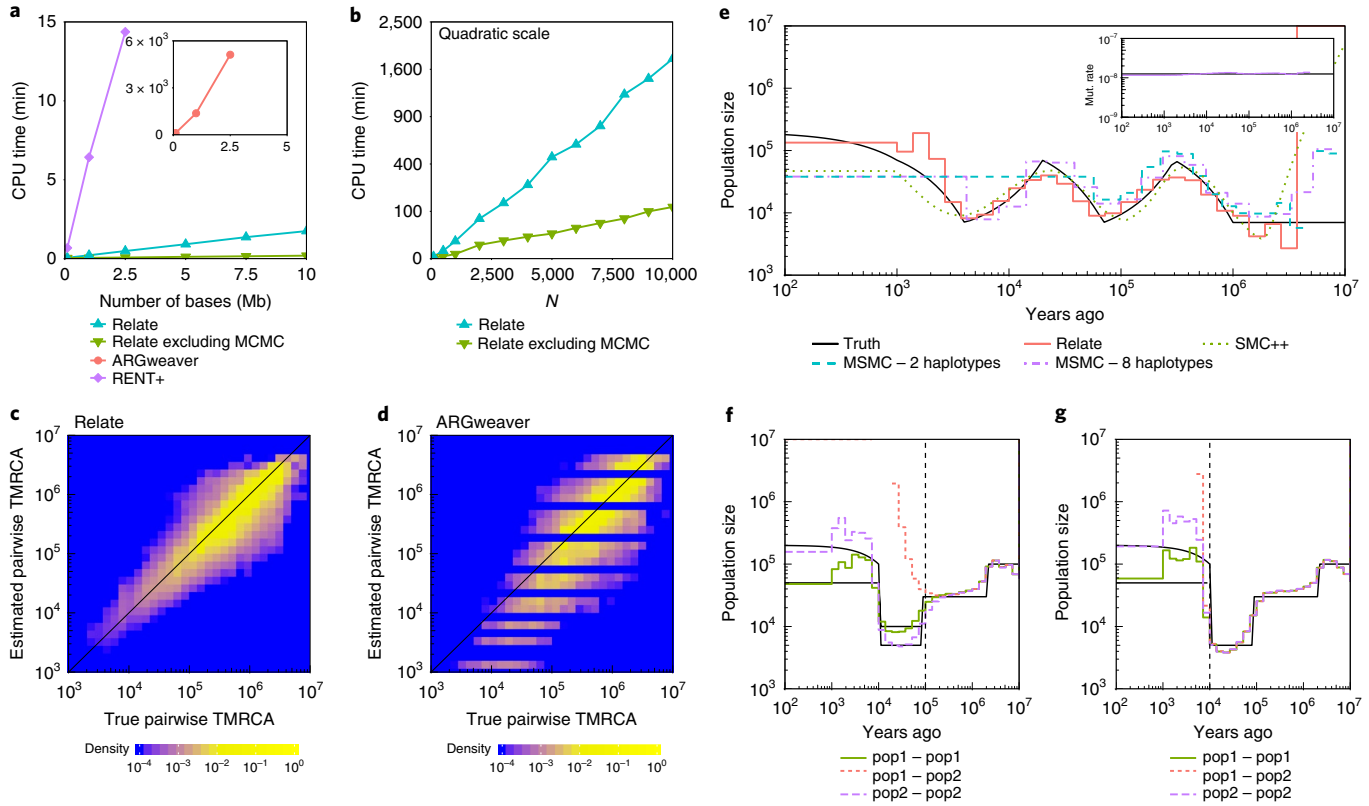

Fig. 2 | Simulated data. a, Runtimes of Relate, RENT+ and ARGweaver in CPU minutes as a function of the number of bases simulated with $N=200$, per base per generation mutation rate $\mu=1.25\times10^{-8}$, diploid effective population size $2N_e=30{,}000$ and recombination rates taken from human chromosome 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 2.5 Mb for each data point. Other parameters are the same as in **a**; the y axis is on a quadratic scale. **c**, True TMRCAs 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, 200 Mb and an oscillating population size[19]. Relate and SMC++ estimates are based on 200 haplotypes, MSMC estimates are obtained from 2 or 8 haplotypes. The 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 10,000 YBP (**g**).

and JPT (JPT, Japanese in Tokyo) or FIN and GBR (FIN, Finnish in Finland) (Fig. 3d,e). Finnish samples exhibit a second bottleneck, around 3,000–9,000 YBP following separation from GBR[29,30], with other population-specific events in, for example, Peruvians and Gujarati individuals (Supplementary Fig. 6). The Finnish bottleneck is thought to have caused enrichment of certain disease-causing gene variants, commonly classified as Finnish heritage diseases[29,30]. A strong bottleneck post-dating separation from Eurasian groups is absent in African populations (LWK, Luhya in Webuye, Kenya and YRI shown, Fig. 3f). All populations show a remarkable increase often to >1,000,000; in the recent past (Supplementary Fig. 6), however, we note possible inaccuracies due to incomplete power to detect rare variants[23] leading to underestimation and computational phasing leading to overestimation (Supplementary Fig. 4).

Exploring the relative mutation rate of particular mutation classes through time confirms, as reported previously[24], a strong elevation in the rate of trinucleotide changes including TCC to 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 (Fig. 4a). Other mutation types show more subtle temporal biases and signatures consistent with GC-biased gene conversion[31] (Supplementary Fig. 6). Overall, these results support the 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 that 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 (Myr) in age and with varying lower end), across human groups (Fig. 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[21,32]. We therefore labeled deep branches possessing at least two derived mutations by whether at least one mutation is shared with the sequenced Denisova[32] or Neanderthal[21,33] genomes (Fig. 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 10,000 YBP, 85–90% are shared with Neanderthal or Denisovan for most Eurasian groups (Fig. 4c and Methods). Any lineages from recent introgression events will show a lower-end age younger than the time of introgression and an upper end older than the split time of the introgressing group, so we expected 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

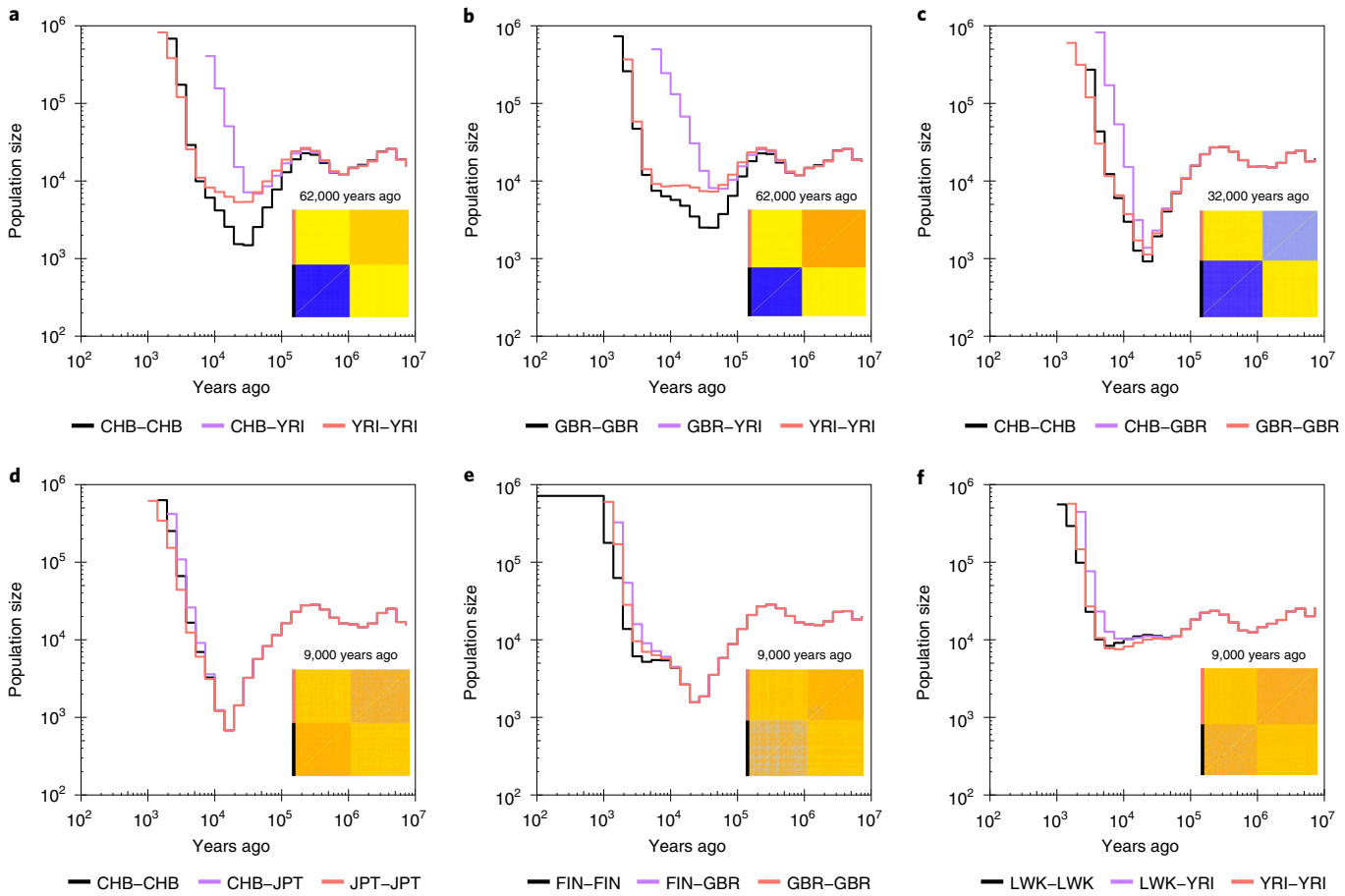

Fig. 3 | Population sizes and split times in 1,000GP. a–f, Relate-based population-specific estimates of population size and cross-population coalescence rates using genome-wide genealogies for CHB and YRI (**a**), GBR and YRI (**b**), CHB and GBR (**c**), CHB and JPT (**d**), FIN and GBR (**e**) and 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 color on the left of each matrix.

left a major, recent impact on non-African populations studied here. An exception is Iberian populations in Spain (IBS), which has more long branches shared with African populations (Supplementary Fig. 6). In the East and South Asian groups, the data suggest a very recent arrival of Denisovan DNA (mainly <15,000 YBP). In non-Africans, Neanderthal sharing remains high for branches with lower-end age younger than ~30,000 YBP. These dates are only lower bounds on the introgression time, and an accurate arrival date of Neanderthal DNA would require estimating a joint genealogy that requires further work. Nevertheless, they are consistent with previous estimates based on linkage disequilibrium (LD)[34], and of direct evidence of hybrids[34,35] 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 underestimates in some cases.

In contrast to non-African groups, sharing with Neanderthal/Denisovans is lower (<20%, Fig. 4c) in African populations and declines towards the present, suggesting minimal recent interactions[21,32]. 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 versus 7,012 mutations occur in African versus 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 Fig. 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 (Fig. 4d, Methods). In panmictic simulations with matched population size histories, we observed no such excess (Supplementary Fig. 6). This gives evidence for ancient but uncharacterized population structure within Africa, as suggested elsewhere[36,37]. 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 modeling how mutations arise and spread, genealogical trees offer the potential to powerfully investigate different modes of natural selection. For example, a recent method, Singleton Density Score (SDS), indirectly tests for differences in tree tip branch lengths between carriers and noncarriers using the density of singletons around a focal SNP[38] and a tree-based analog (trSDS) tests this directly[39]. 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

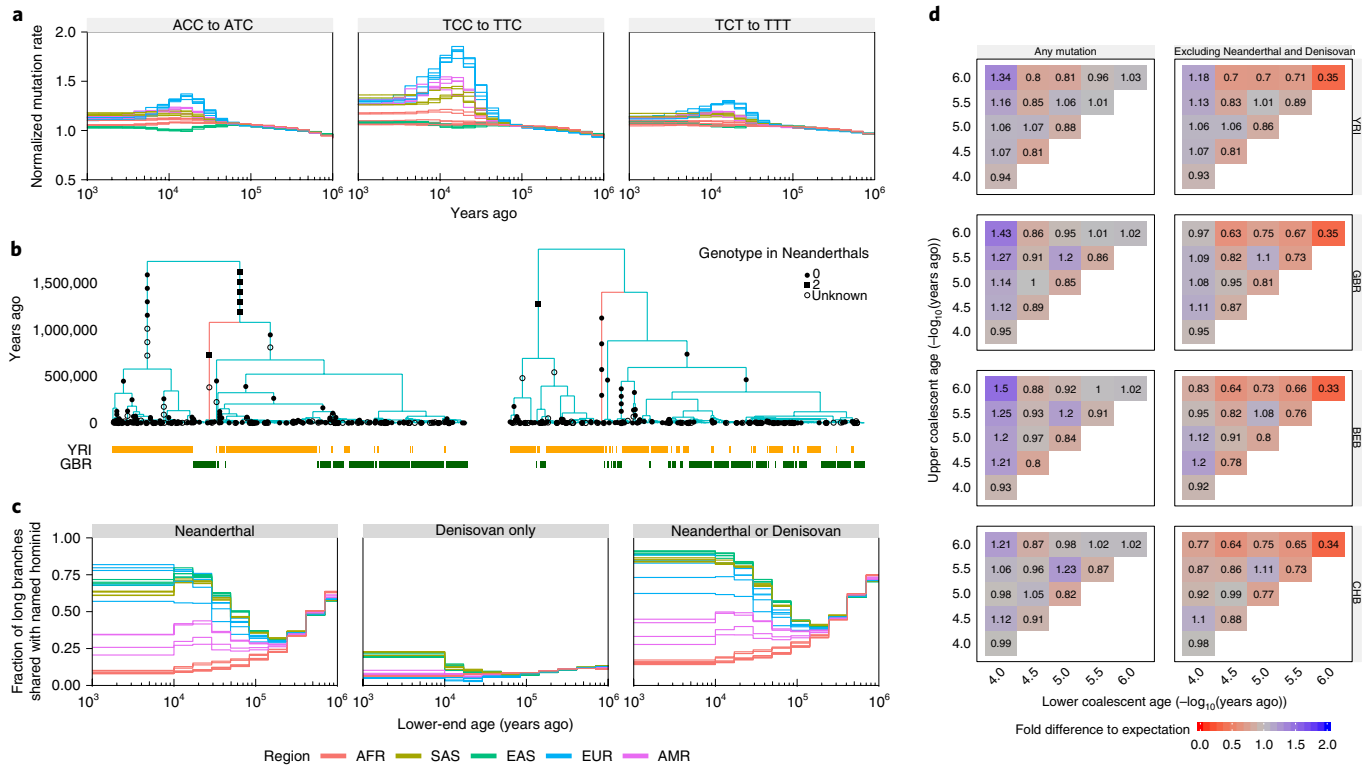

Fig. 4 | Evolution of human mutation rates and evidence for introgression. a, Evolution of mutation rates for three triplet mutations ACC to ATC, TCC to TTC and TCT to TTT (see Supplementary Fig. 7 for all 96 triplet mutations and Methods for normalization). **b**, Marginal trees for subregions on chromosome 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 > 1 million YBP that are shared with Neanderthals (left), Denisovans and not Neanderthals (center) or Neanderthals or Denisovans (right) (Methods). In **a** and **c**, colors 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 events, relative to the expected number of mutations when randomizing topology but fixing ages of coalescence events (Methods). The right column shows mutations not present in Neanderthal or Denisovan samples.

be calculated analytically and is robust in principle to population size changes (Methods).

The simulated data (Fig. 5a) in null no-selection scenarios showed a close match of our P values P_R to the expected uniform distribution. Across a range of selective advantages and SNP frequencies (Fig. 5b and Supplementary Fig. 6), our approach increased 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 yielded a test that is uniformly more powerful than other approaches (Fig. 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[40].

Calculating P_R for SNPs across twenty 1,000GP 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 included the *LCT* (lactase) region associated with lactose tolerance in Europeans and a mutation in the *EDAR* gene in East Asian populations[41,42], with both likely causal variants strongly associated with our most significant mutation ($r^2 \geq 0.8$). We also observed a previously detected strong signal of positive selection in the major histocompatibility complex in GBR[43] (Fig. 5c). Among new regions, we identified selection evidence at the *EDARADD* gene, which interacts with the *EDAR* gene[44] in the formation of hair follicles, sweat glands and teeth[42], in all South Asian populations and Finns, with $P_R < 10^{-6}$ in all other European populations. In 16 out of 35 regions, we identified GWAS catalog hits (odds ratio OR = 6.44; $P = 0.01$), nonsynonymous 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 were found only for African populations.

SNPs in functional parts of the genome are significantly enriched among targets of positive selection (Fig. 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. Many previous studies[45–48] have attempted to test polygenic traits for evidence of directional selection, but confounding due to population stratification[49,50] 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 thinned 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[49,50], but is not expected, in general, to produce genome-wide significant false-positives. At each SNP, we used only the

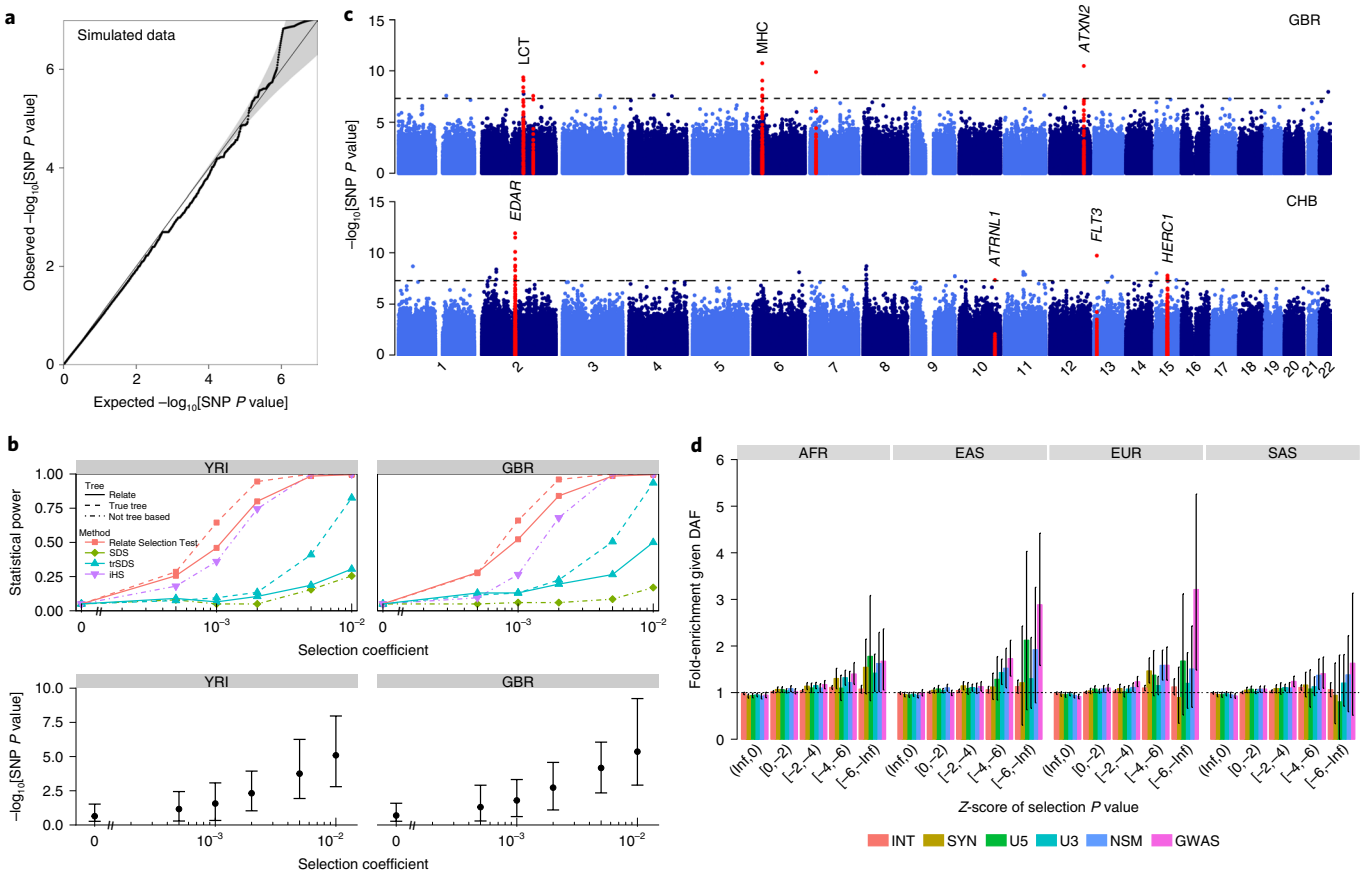

Fig. 5 | Natural selection. a, Quantile–quantile plot of *P* values P_R for selection evidence of SNPs. We simulated 250 Mb for $N = 1{,}000$ haplotypes using the recombination map of chromosome 1 and a bottleneck population size resembling that of non-African populations. **b**, Power simulations using $N = 1{,}000$ haplotypes. We used historical population sizes estimated by Relate for YRI (left) and GBR (right). The top row shows statistical power and the bottom row shows P_R with the mean indicated by circles and the 5th and 95th 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 major histocompatibility complex in GBR. **d**, Mean enrichment of functional annotation among targets of selection, conditional on allele frequency. Error bars show 95% confidence intervals estimated from 1,000 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. Inf, infinity. 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), nonsynonymous mutations (NSM) and GWAS hits (GWAS).

association direction, rather than its strength, to offer additional robustness to potential confounding.

If positive selection influences a trait in a certain direction, for example 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 (Fig. 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[51,52], because to avoid ascertainment effects we condition on SNP allele frequencies at trait-influencing sites. Therefore, we additionally tested 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 color within the UK Biobank[53] (Fig. 6a). As in previous studies[48,54,55], we found a signal for SNPs associated with blonder hair color among European populations. We further observed strong selection towards light brown hair color and against black hair color, which was weaker in South Asians, but not in other groups. Testing based on iHS scores decreased significance by up to four orders of magnitude (Fig. 6a) and some signals became nonsignificant. We applied the same test to test 84 traits: 6 from the UK Biobank and 78 with at least 10 genome-wide significant GWAS catalog association signals in each effect direction, in all populations except recently admixed groups. Sixty-one of these (73%) showed nominal selection evidence ($P < 0.05$) in at least one population (Fig. 6b), with strong geographic clustering. The most significant signal ($P = 6 \times 10^{-14}$) was for SNPs associated with decreased body mass index (BMI) in Utah residents with Northern and Western European ancestry (CEU). The largest number of selection signals were 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 Fig. 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[56]. While several studies have reported genetic differentiation between populations[57–59], evidence for selection remains controversial[39,46–50,57,58,60] and some

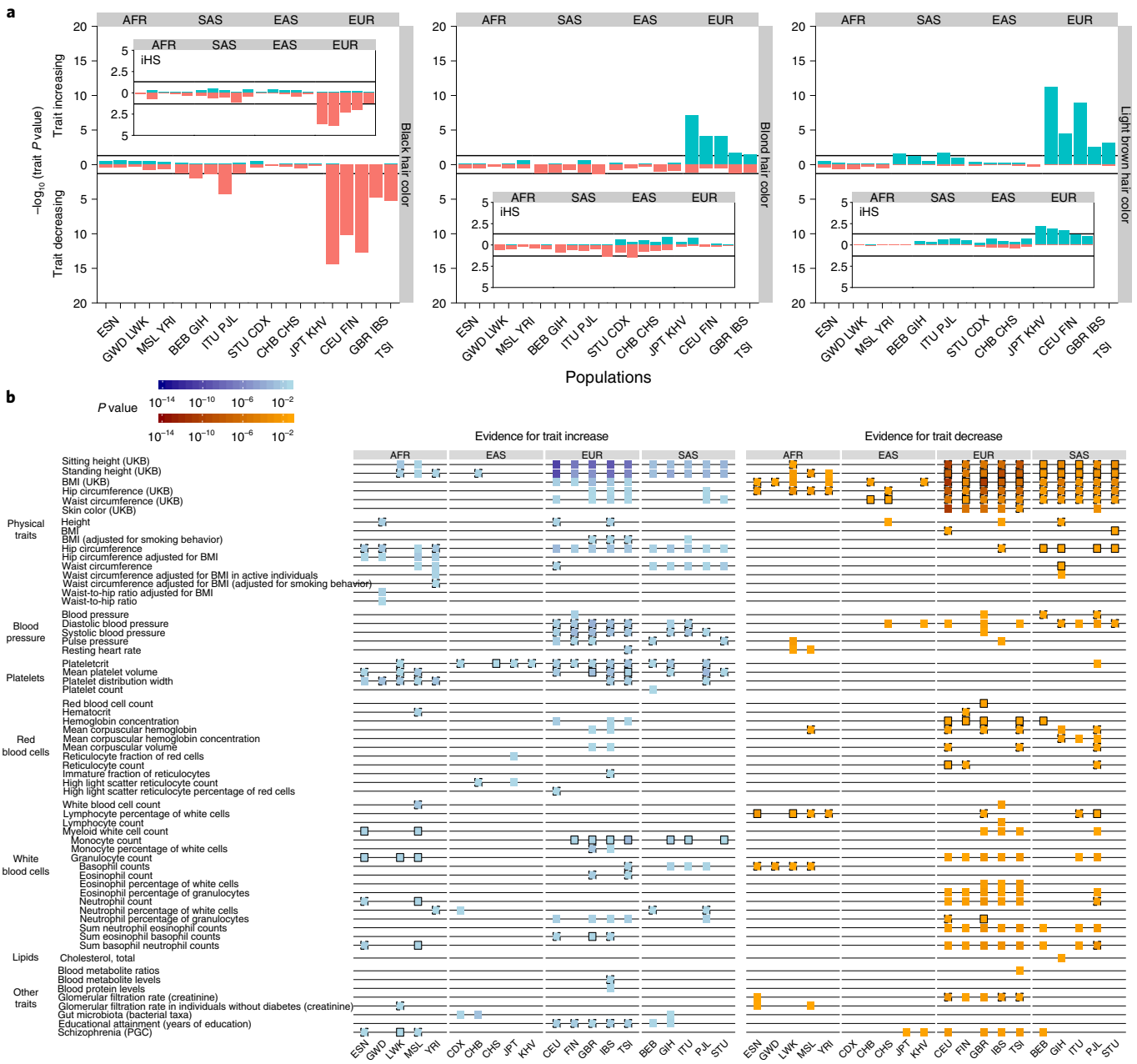

Fig. 6 | Evidence of selection on traits. a, *P* values for evidence of directional selection of black, blond and light brown hair color (see Methods for calculation of *P* values). Insets show *P* values for the same test but using iHS scores instead, where iHS scores are calculated for each population separately for any variant with a minor allele frequency > 5% in that population. Population abbreviations are defined in Supplementary Table 2. **b**, Evidence for directional or bidirectional selection on multi-allelic traits. Each trait is associated with at least ten 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).

studies reporting recent selection on increased height in Europeans appear confounded by subtle population stratification[39,49,50]. Our test found 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 identified strong evidence of selection favoring BMI-decreasing SNPs across almost all populations, with agreement of DAF shifts, indicative of directional selection. For both traits, we detected little evidence of selection in the smaller GWAS catalog 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 saw selection evidence for a variety of blood-related phenotypes, with congruent DAF signals. In Europeans and some South Asians, we detected a strong signal favoring SNPs associated with blood pressure increases, contrary to previous studies suggesting the opposite direction[54,61]. We, moreover, found evidence in many groups for selection favoring 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 1,000GP data analysis, we provided 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[62,63]) 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 found a relatively small (<40) number of clear signals of strong, ongoing selection across multiple human populations. In contrast, we found 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 saw 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[40]. 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 some of the reasons for the complicated assignment of selection signals to specific phenotypes (Supplementary Note).

Relate provides age estimates for mutations and other events, and this has enabled 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, for example, 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[36,37]. 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 the estimated mutational ages obtained here directly (https://zenodo.org/record/3234689).

There are several natural extensions to Relate itself, such as, for example, allowing for increasing sample sizes. A recently developed method, tsinfer[64], 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[2]. This approach might also be useful for efficient statistical phasing and/or imputation of individuals only typed at a subset of markers.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41588-019-0484-x.

Received: 20 February 2019; Accepted: 15 July 2019;
Published online: 2 September 2019

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**, e1004342 (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. Lond. B* **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. & Lunter, G. Demographic inference using particle filters for continuous Markov jump processes. Preprint at *bioRxiv* https://doi.org/10.1101/382218 (2018).
17. 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).
18. Reich, D. D. E. et al. Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
19. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
20. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
21. Green, R. E. et al. A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
22. Sudmant, P. H. et al. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
23. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
24. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl Acad. Sci. USA* **112**, 3439–3444 (2015).
25. 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).
26. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).

27. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
28. Bae, C. J., Douka, K. & Petraglia, M. D. On the origin of modern humans: Asian perspectives. *Science* **358**, eaai9067 (2017).
29. Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559 (2015).
30. 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).
31. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genom. Hum. Genet.* **10**, 285–311 (2009).
32. Meyer, M. et al. A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).
33. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
34. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012).
35. Fu, Q. et al. An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
36. 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. USA* **108**, 15123–15128 (2011).
37. Ragsdale, A. P. & Gravel, S. Models of archaic admixture and recent history from two-locus statistics. *PLoS Genet.* https://doi.org/10.1371/journal.pgen.1008204 (2019).
38. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
39. Edge, M. & Coop, G. Reconstructing the history of polygenic scores using coalescent trees. *Genetics* **211**, 235–262 (2019).
40. 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).
41. Enattah, N. S. et al. Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
42. Hardouin, E. et al. Positive Selection in East Asians for an EDAR Allele that Enhances NF-κB Activation. *PLoS ONE* **3**, e2209 (2008).
43. 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).
44. Sadier, A., Viriot, L., Pantalacci, S. & Laudet, V. The ectodysplasin pathway: from diseases to adaptations. *Trends Genet.* **30**, 24–31 (2014).
45. 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).
46. 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).
47. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
48. Field, Y. et al. Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
49. Sohail, M. et al. Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *eLife* **8**, e39702 (2019).
50. Berg, J. J. et al. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife* **8**, e39725 (2019).
51. Maruyama, T. The age of an allele in a finite population. *Genet. Res.* **23**, 137 (1974).
52. 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).
53. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
54. 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).
55. 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. USA* **111**, 4832–4837 (2014).
56. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
57. Turchin, M. C. et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44**, 1015–1019 (2012).
58. Robinson, M. R. et al. Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47**, 1357–1362 (2015).
59. 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).
60. Crespi, B., Summers, K. & Dorus, S. Adaptive evolution of genes underlying schizophrenia. *Proc. R. Soc. B* **274**, 2801–2810 (2007).
61. Young, J. H. et al. Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1**, e82 (2005).
62. Hinch, A. G. et al. The landscape of recombination in African Americans. *Nature* **476**, 170–175 (2011).
63. Fledel-Alon, A. et al. Variation in human recombination rates and its genetic determinants. *PLoS ONE* **6**, e20321 (2011).
64. Kelleher, J., Wong, Y., Albers, P., Wohns, A. W. & McVean, G. Inferring the ancestry of everyone. Preprint at *bioRxiv* https://doi.org/10.1101/458067 (2018).

Acknowledgements

We thank N. Barton, D. Falush, M. Przeworski, G. Sella, J. Terhorst, P. Palamara, G. Lunter, J. Marchini, S. Hu, C. B. Cole, T. Aid and C. E. West for helpful comments, ideas and suggestions. L.S. acknowledges the support provided through the Engineering and Physical Sciences Research Council (grant number EP/G03706X/1). M.F. acknowledges the support provided through the Natural Sciences and Engineering Research Council of Canada (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). For computation we used the Oxford Biomedical Research Computing 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.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0484-x.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.R.M.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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)[4]. 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 Fig. 1, Supplementary Fig. 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[26] (see Supplementary Fig. 2 for parameter choices). Intuitively, this HMM reconstructs a haplotype as a mosaic of other sample haplotypes along the genome (Supplementary Fig. 1), allowing for mismatching in the copying process and viewing changes in haplotype as recombination events. After applying the HMM, at a focal SNP ℓ, 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, and therefore receiving a smaller distance in this matrix.

We modified the Li and Stephens HMM to account for ancestral and derived states, which is 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 coalesces 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 practice, 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, for example 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 MCMC algorithm to estimate branch lengths. The MCMC algorithm has a coalescent prior assuming a single panmictic population[3].

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 five 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 subpopulations and cross-coalescence rates between populations.

Estimating relative mutation rates through time. We estimate the mutation rate through time for all 96 triplet mutations (Fig. 4a and Supplementary Fig. 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 normalize the mutation rates such that the average rate over time equals 1. In simulations (Supplementary Fig. 4), we show that variable mutation rates among categories can be detected by this approach and approximately dated.

Pre-processing of the 1,000GP dataset. The 1,000GP dataset comprises 2,504 individuals, from 26 populations. We obtain a phased version of the dataset (see Data availability). We next exclude multi-allelic SNPs, and we exclude one individual (two haplotypes) from each population for future applications, and the remaining 2,478 individuals (Supplementary Table 2). We use a genomic mask provided with the 1,000GP 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[21] and Altai[33] Neanderthals (NEA) and a Denisovan (DEN)[32] (see Data availability) 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 1 million YBP. Of any such branches, we calculate the fraction of branches with at least one NEA or DEN mutation. In Fig. 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 Fig. 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 randomize 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 $P(r,s|k)=\sum_{\ell\geq k,h<k} I_{t_\ell\in s} I_{t_h\in r}\frac{2h}{\ell(\ell-1)}\frac{1}{k}$, 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 Fig. 4d, $\log_{10}$ 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 favorable mutations spreading rapidly in a population. One approach to capture this is via the number of lineages ultimately descending from the potentially favorable 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, for example, Ref. [65]). 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 increases 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 1,000GP dataset is based on a subset of all SNPs mapping to trees, and present in three or more copies in the dataset. Specifically, we remove SNPs failing any of the following filters: (1) the number of mutations mapping to that SNP's tree is in the bottom 5th percentile or (2) the fraction of tree branches having at least one SNP is in the bottom 5th percentile. This excludes approximately 7% of SNPs.

Simulation of positive selection. To simulate positive natural selection, we adopt the pipeline outlined in ref. [48]. We first simulate the trajectory of the DAF using simuPOP[66]. 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 be 0.7 (see Supplementary Fig. 7 for other present-day DAFs). We then use mbs2 (ref. [67]; mutation rate $\mu=1.25\times10^{-8}$, constant recombination rate $\rho=5\times10^{-9}$) to simulate a region of 20 Mb, given the DAF trajectory for the central selected SNP. For each nonzero 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 trSDS proposed in ref. [39]. For iHS, SDS and trSDS, we standardize scores using the mean and standard deviation in the neutral case, which is an idealized setting that should favor 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 Fig. 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[68] (ACB, ASW, CLM, MXL, PEL, PUR (see Supplementary Table 2 for abbreviations)), 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 use 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[69] (see Data availability) and the GWAS catalog 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 catalog[70] (see Data availability) to study polygenic adaptation. We use only association signals whose GWAS *P* value is smaller than 5×10^{-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 ten 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 download results for a large-scale GWAS conducted by the Psychiatric Genomics Consortium[71] (see Data availability). We consider SNPs reaching a GWAS *P* value of 5×10^{-8} of which there are 9,138. We intersect this set of SNPs with SNPs segregating in each of the considered populations. As for the GWAS catalog, we identify approximately independent blocks. We then choose 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[53] (see Data availability), 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 conduct 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 compare to our results that are 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Relate-estimated coalescence rates, allele ages and selection *P* values for the 1,000GP can be downloaded from https://zenodo.org/record/3234689. Datasets used in the current study were obtained from the following URLs: 1,000GP phased dataset, https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html (13 January 2017); Genomic mask, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/ (20 July 2017); Human ancestral genome, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/ (20 July 2017); Altai Neanderthal, http://cdna.eva.mpg.de/neandertal/Vindija/VCF/Altai/ (17 February 2018); Vindija Neanderthal, http://cdna.eva.mpg.de/neandertal/Vindija/VCF/Vindija33.19/ (1 May 2018); Denisovan, http://cdna.eva.mpg.de/neandertal/altai/Denisovan/ (2 March 2018); GWAS catalog, https://www.ebi.ac.uk/gwas/api/search/downloads/full (9 November 2017); PGC GWAS study, https://www.med.unc.edu/pgc/results-and-downloads (23 November 2018); HaploReg, http://archive.broadinstitute.org/mammals/haploreg/data/haploreg_v4.0_20151021.vcf.gz (21 October 2017); GTEx eQTL https://storage.googleapis.com/gtex_analysis_v7/single_tissue_eqtl_data/GTEx_Analysis_v7_eQTL.tar.gz (13 January 2019); UK Biobank GWAS summary statistics, http://www.nealelab.is/uk-biobank (4 October 2018); PopHumanScan, https://pophumanscan.uab.cat (13 January 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 January 2017); RENT+, https://github.com/SajadMirzaei/RentPlus (2 October 2017); msprime, https://github.com/tskit-dev/msprime (22 July 2017); msmc, https://github.com/stschiff/msmc2 (14 October 2017); SMC++, https://github.com/popgenmethods/smcpp (14 October 2017); simuPOP, http://simupop.sourceforge.net/ (27 June 2018); mbs, http://www.sendou.soken.ac.jp/esb/innan/InnanLab/ (27 June 2018); SDS, https://github.com/yairf/SDS (27 June 2018), selscan, https://github.com/szpiech/selscan (31 July 2018); hapbin, https://github.com/evotools/hapbin (11 December 2018).

References

65. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent tree. *Stoch. Model.* **14**, 273–295 (1998).
66. Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).
67. 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).
68. 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).
69. 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).
70. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2016).
71. Ruderfer, D. M. et al. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715.e16 (2018).

nature research

Corresponding author(s): Simon R. Myers

Last updated by author(s): 27 June 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No data was collected in this study.
Data analysis	We used our implementation of the proposed method, Relate, which 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)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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 (13 Jan 2017); Genomic mask, 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/ (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 (21 Oct 2017); GTEx eQTL 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)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was predetermined by the data set used (1000 Genomes Project Phase 3). This data set comprises 2504 samples. We confirmed using simulations that this sample size is sufficient for the applications considered in this study.
Data exclusions	We excluded one individual (two haplotypes) from each of the 26 populations for future applications, where this individual was chosen uniformly at random.
Replication	For each application, we confirmed that our findings are consistent known results in the literature. We did not replicate our analysis on an independent data set, because such a data set was not available.
Randomization	Randomisation was performed when necessary using standard techniques that account for ascertainment effects or structure in the data such as linkage-disequilibrium (LD).
Blinding	We were not blinded to group allocation, but randomly permutated samples where necessary.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging