High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability

Pier Francesco Palamara, Jonathan Terhorst, Yun S. Song and Alkes L. Price

Interest in reconstructing demographic histories has motivated the development of methods to estimate locus-specific pairwise coalescence times from whole-genome sequencing data. Here we introduce a powerful new method, ASMC, that can estimate coalescence times using only SNP array data, and is orders of magnitude faster than previous approaches. We applied ASMC to detect recent positive selection in 113,851 phased British samples from the UK Biobank, and detected 12 genome-wide significant signals, including 6 novel loci. We also applied ASMC to sequencing data from 498 Dutch individuals to detect background selection at deeper time scales. We detected strong heritability enrichment in regions of high background selection in an analysis of 20 independent diseases and complex traits using stratified linkage disequilibrium score regression, conditioned on a broad set of functional annotations (including other background selection annotations). These results underscore the widespread effects of background selection on the genetic architecture of complex traits.

Overview of the ASMC method. We developed a new method, ASMC, that estimates the coalescence time (which we also refer to as time to most recent common ancestor, TMRCA) for a pair of chromosomes at each site along the genome. ASMC utilizes a HMM, which is built using the coalescent with recombination (HMRs) to estimate the coalescence time of two homologous chromosomes at each position in the genome, leveraging previous advances in coalescent theory. These methods have been broadly applied to reconstructing demographic histories of human populations. More generally, methods for inferring ancestral relationships among individuals have potential applications for the detection of signatures of natural selection, genome-wide association studies, and genotype calling and imputation. However, all currently available methods for inferring pairwise coalescence times require whole-genome sequencing (WGS) data, and can only be applied to small datasets due to their computational requirements.

Here, we introduce a new method, the ascertained sequentially Markovian coalescent (ASMC), that can efficiently estimate locus-specific coalescence times for pairs of chromosomes using only ascertained SNP array data, which are widely available for hundreds of thousands of samples. We verified the accuracy of ASMC using coalescent simulations, and determined that it is orders of magnitude faster than other methods when WGS data are available. Leveraging the speed of ASMC, we analyzed SNP array and WGS datasets with the goal of detecting signatures of recent positive selection and background selection using pairwise coalescence times along the human genome. We first analyzed 113,851 British individuals from the UK Biobank dataset, detecting 12 loci with unusually high density of very recent coalescence times as a result of recent positive selection at these sites. These include six known loci linked to nutrition, immune response and pigmentation, as well as six novel loci involved in immunity, taste reception and other aspects of human physiology. We then analyzed 498 unrelated WGS samples from the Genome of the Netherlands dataset to search for signals of background selection at deeper time scales and finer genomic resolution. We determined that SNPs in regions with low values of average coalescence time are strongly enriched for heritability across 20 independent diseases and complex traits (average $n = 86,000$), even when conditioning on a broad set of functional annotations (including other background selection annotations).

Results

Overview of the ASMC method. We developed a new method, ASMC, that estimates the coalescence time (which we also refer to as time to most recent common ancestor, TMRCA) for a pair of chromosomes at each site along the genome. ASMC utilizes a HMM, which is built using the coalescent with recombination process; the hidden states of the HMM correspond to a discretized set of TMRCA intervals, the emissions of the HMM are the observed genotypes, and transitions between states correspond to changes in TMRCA along the genome due to historical recombination events. ASMC shares several key modeling components with previous coalescent-based HMM methods, such as the PSMC, the MSMC, and, in particular, the recently developed SMC++. In contrast to these methods, however, the main objective of ASMC is not to reconstruct the demographic history of a set of analyzed samples. Instead, ASMC is optimized to efficiently compute coalescence times along the genome of pairs of individuals in modern datasets. To this end, the ASMC improves current coalescent HMM approaches in two key ways. First, by modeling non-random ascertainment of genotyped variants, ASMC enables accurate processing of SNP array data, in addition to WGS data. Second, by introducing a new dynamic programming algorithm, it is orders of magnitude faster than other coalescent HMM approaches, which enables it to process large volumes of data. Details of the method are described in the Methods; we have released open-source software implementing this method (see URLs).

©Department of Statistics, University of Oxford, Oxford, UK. 2Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. 3Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. 4Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. 5Department of Statistics, University of Michigan, Ann Arbor, MI, USA. 6Department of Statistics, University of California, Berkeley, Berkeley, CA, USA. 7Computer Science Division, University of California, Berkeley, Berkeley, CA, USA. 8Chan Zuckerberg Biohub, San Francisco, CA, USA. *e-mail: palamara@stats.ox.ac.uk; aprice@hsph.harvard.edu
Simulations. We assessed the accuracy of ASMC in inferring locus-specific pairwise TMRCA from SNP array and WGS data via coalescent simulations using the ARGON software\(^1\). In brief, we measured the correlation between the true and inferred average TMRCA for all pairs of 300 individuals simulated using a European demographic model\(^1\), for a 30-Mb region with SNP density and allele frequencies matching those of the UK Biobank dataset (Fig. 1 and Methods). As expected, ASMC achieved high accuracy when applied to WGS data \((r^2 = 0.95)\). When sparser SNP array data were analyzed, the correlation remained high (for example, \(r^2 = 0.87\) for SNP array density in the UK Biobank dataset), and increased with genotyping density. Similar relative results were obtained when comparing the root mean squared error (RMSE) between true and inferred TMRCA at each site, and the posterior mean estimate of TMRCA attained higher accuracy than the maximum a posteriori (MAP) estimate (Supplementary Fig. 1). Inference locus-specific TMRCA is closely related to the task of detecting genomic regions that are identical-by-descent (IBD), which we define as regions for which the true TMRCA is lower than a specified cut-off (other, related definitions have been proposed\(^{32}\)). ASMC attained higher IBD detection accuracy \((area\ under\ the\ precision-recall\ curve)\) than the widely used Beagle IBD method, although it runs considerably faster, making approximations that have only a small impact on accuracy (Fig. 1). Accordingly, we observed that the running time of ASMC was two to four orders of magnitude faster than SMC++ method, which was shown to be more computationally efficient than other coalescent-based methods\(^1\), has asymptotic running time \(O(s^m)\). In comparison, the SMC++ method, which was shown to be more computationally efficient than other coalescent-based methods\(^1\), has a speedup in the analysis of WGS data leverages approximations that do not result in a notable loss of accuracy (Supplementary Fig. 4). The memory cost of ASMC was also efficient compared to SMC++, scaling linearly with \(s\) (Supplementary Fig. 5).

Signals of recent positive selection in the UK Biobank. The computational efficiency of ASMC enables its application to analyses of TMRCA for large datasets. We thus used ASMC to infer locus-specific TMRCA in 113,732 unrelated individuals of British ancestry...
from the UK Biobank, typed at 678,956 SNPs after quality control and phased using Eagle34 (Methods). We note that phasing accuracy in this dataset is very high, with an average switch error rate in the order of 0.3% (one switch error approximately every 10 cM34). We partitioned the data into batches of approximately 10,000 samples each and inferred locus-specific TMRCA for all haploid pairs within each batch, analyzing a total of 2.2 billion pairs of haploid genomes.

We identified genomic regions with an unusually high density of very recent inferred TMRCA events (that is, within the past several thousand years). Such signals are expected at sites undergoing recent positive selection, because a rapid rise in frequency of a beneficial allele causes all individuals with the beneficial allele to coalesce to a more recent common ancestor than under neutral expectation41; approaches to detect selection based on distortions in inferred coalescence times have recently been applied at different time scales41. We thus computed a statistic, DRC150, reflecting the density of recent coalescence (within the past T generations), averaged within 0.05-cM windows. To compute approximate P values, we noted that the DRC150 statistic under the null is approximately Gamma-distributed. We thus obtained approximate P values for the DRC150 statistic by fitting a Gamma distribution to the null 18% of the genome obtained by conservatively excluding 500-kb windows around regions that have previously been implicated in scans for positive selection (Methods). Using coalescent simulations, we determined that DRC150 is highly sensitive in detecting signals of positive selection within the past approximately 20,000 years compared to other methods38,45 (Methods and Supplementary Fig. 6).

Analyzing 63,103 windows of length 0.05 cM in the UK Biobank dataset, we detected 12 genome-wide significant loci (P < 0.05/63,103 = 7.9 × 10−5; Fig. 3a and Table 1). The loci that we detected exhibited strong enrichment of recent coalescence events spanning up to the past 20,000 years (Fig. 3b,c and Supplementary Fig. 7), consistent with our simulations (Supplementary Fig. 6). Of the 12 loci, 6 are loci known to be under recent positive selection, harboring genes linked to nutrition (LCT29), immune response (HLA39, TLR40 and IGH41), eye color (GRM542) and skin pigmentation (MC1R43). We also detected 6 new loci, harboring genes involved in immune response (STAT444, associated with autoimmunem disease43-45), mucus production (MUCSB46) within a cluster of mucin genes, involved in protection against infectious disease41, associated with several types of cancer47 and lung disease48), taste reception (PKD1L349, associated with kidney disease50,51), cardiac and fetal muscle (MYL4, associated with atrial fibrillation52), blood coagulation (ANXA353, associated with cancer44 and immune disease54) and brain-specific expression and immune response (FAM19A55). We note that suggestive loci implicated by the DRC150 statistic (P < 10−4; Supplementary Table 6) include known targets of selection linked to eye color (HERC256), retinal and cochlear function (PCDH1557), celiac disease (SLC22A458,59) and skin pigmentation (SLC45A260).

Heritability enrichment in regions with background selection. We next sought to detect signals of natural selection at deeper time scales. To accomplish this, we used ASMC to estimate locus-specific TMRCA for all ~0.5 million pairs of haploid genomes from unrelated individuals in the Genome of the Netherlands (GoNL) WGS dataset (498 samples and 19,730,834 variants after quality control; Methods); we note that WGS data are required to achieve accurate resolution at deeper time scales (Fig. 1a). Because natural selection modulates the effective population size along the genome61, we estimated its strength by measuring the average pairwise TMRCA at each site, which is proportional to the effective population size61. We refer to this annotation as ASMCavg. Forward-in-time
simulations confirmed that the ASMC\textsubscript{org} annotation captures the presence of unusual TMRC\textsubscript{A} variation due to background and positive selection, which leads to lower values of ASMC\textsubscript{org} (Methods and Supplementary Fig. 8). We expect much or most of the variation in the ASMC\textsubscript{org} annotation to be driven by deleterious effects, as supported by several recent studies\textsuperscript{60,62–66}, and thus interpret ASMC\textsubscript{org} as an annotation of background selection. We note, however, that in general ASMC\textsubscript{org} can be affected by several types of selection that have an effect on effective population size\textsuperscript{35,60}, including background, positive and balancing selection, and that some authors have suggested that positive selection has an important role in shaping genomic diversity\textsuperscript{37,67}. The genome-wide average of ASMC\textsubscript{org} in the GoNL data was 17,399 generations (s.d. = 3,030), with an effective size of approximately 10,000 commonly assumed in the literature\textsuperscript{68,69} (we note, however, that our analysis is limited to obtaining posterior TMRC\textsubscript{A} estimates, which are driven by the demographic model provided as input). We thus expect the ASMC\textsubscript{org} annotation to capture background selection occurring within the past several hundred thousand years. As expected, ASMC\textsubscript{org} was highly correlated with other measures of background selection, including nucleotide diversity ($r = 0.50$), the McVicker B-statistic\textsuperscript{60} ($r = -0.28$), and allele age predicted by ARGWeaver\textsuperscript{6}, quantile-normalized within 10 minor allele frequency (MAF) bins\textsuperscript{6} ($r = 0.26$, Supplementary Table 7).

Analyses using stratified linkage disequilibrium (LD) score regression (S-LDSC)\textsuperscript{70} have shown that regions under background selection are enriched for disease and complex trait heritability\textsuperscript{7}; enrichment was observed for the nucleotide diversity, McVicker B-statistic, and ARGWeaver-predicted allele age annotations, as well as three other annotations linked to LD and recombination. We evaluated the ASMC\textsubscript{org} background selection annotation for heritability enrichment by applying S-LDSC to summary association statistics from 20 independent diseases and complex traits (Supplementary Table 8, average $n = 86,000$). We performed both an unconditioned analysis using only the ASMC\textsubscript{org} annotation and a joint analysis conditioned on the 75 annotations from the baselineLD model\textsuperscript{65} (which includes a broad set of functional annotations, in addition to the six annotations linked to background selection and LD), in order to specifically assess whether our annotation provides additional signal. Focusing on the ASMC\textsubscript{org} annotation, we computed the $r^*$ metric\textsuperscript{65}, defined as the proportionate change in per-SNP heritability resulting from a 1-s.d. increase in the value of the annotation, conditional on other annotations included in the model.

In the unconditioned analysis, a lower ASMC\textsubscript{org} was associated with higher per-SNP heritability for all 20 analyzed traits (Fig. 4a), confirming that regions under background selection are enriched for disease heritability. Meta-analyzed across the 20 traits, the $r^*$ for ASMC\textsubscript{org} had a value of $-0.81$ (s.e. = 0.01; $Z$-test $P < 10^{-10}$). After conditioning on the baselineLD model, the $r^*$ for ASMC\textsubscript{org} remained strongly significant at $-0.25$ (s.e. = 0.01; $Z$-test $P = 7 	imes 10^{-13}$), suggesting that ASMC\textsubscript{org} remains informative for disease heritability after conditioning on other annotations that are linked to background selection as well as a broad set of functional annotations. Furthermore, ASMC\textsubscript{org} attained a larger value of $r^*$ than each of the other six annotations linked to background selection (Fig. 4b), indicating that it was the most disease-informative background selection annotation in this analysis; we note that adding ASMC\textsubscript{org} to the baselineLD model reduced the $|r^*|$ of the nucleotide diversity annotation (from 0.13 to 0.00 and reduced the $|r^*|$ of the ARGWeaver\textsuperscript{6}-predicted allele age annotation from 0.25 to 0.13, indicating that ASMC\textsubscript{org} includes signals from these annotations. We computed the proportion of heritability explained by each quintile of the ASMC\textsubscript{org} annotation, which provides a more intuitive interpretation of the strength of the annotation’s effect (Fig. 4c). We observed that SNPs in the smallest quintile of the annotation explained 33.1% (s.e. = 0.5%) of heritability, 3.8x more than SNPs in the highest quintile (8.7%, s.e. = 0.5%), the largest ratio among annotations linked to background selection (Supplementary Table 9) (tied to the nucleotide diversity annotation, the effect of which was however included in the ASMC\textsubscript{org} annotation; Fig. 4b). Annotations constructed on the basis of average pairwise TMRC\textsubscript{A} that are conditional on the allele present on each chromosome were further informative for disease heritability (Supplementary Figs. 9, 10 and Methods).

**Discussion**

We introduce a method for inferring pairwise coalescence times, ASMC, that can be applied to either SNP array or WGS data and is highly computationally efficient. Exploiting the speed of ASMC, we analyzed around 2.2 billion pairs of haploid chromosomes from 113,851 British samples from the UK Biobank dataset, and detected strong evidence of recent positive selection at six known loci and six new loci linked to immune response and other biological functions. We further used ASMC to detect background selection at deeper

---

**Table 1 | Genome-wide significant signals of recent positive selection**

| Chromosome | Location (Mb) | Min. P value | SNP | Candidate gene(s) |
|------------|--------------|-------------|-----|-------------------|
| 2          | 134.44–139.01 | $<10^{-16}$ | rs10206673 | LCT\textsuperscript{68} |
| 11         | 38.44–38.97   | $<10^{-16}$ | rs7660745 | TLR gene family |
| 4          | 42.64–45.18   | $2.87 	imes 10^{-7}$ | rs75229873 | MYL4 |
| 22         | 48.98–49.08   | $4.94 	imes 10^{-7}$ | rs78014641 | FAM19A5 |

We report genomic locations, minimum (Min.) P values across 0.05-cM windows (not adjusted for multiple testing and capped at $10^{-10}$), SNPs corresponding to signal peak, and candidate gene for the 12 genome-wide significant signals of recent positive selection (adjusting for multiple testing, $P < 0.05/6.3103 = 7.9 	imes 10^{-9}$). Novel loci are denoted in bold. The DRC\textsubscript{org} statistic of recent positive selection was computed using all individuals of British ancestry from the UK Biobank ($n = 113,851$, divided in batches of around 10,000 samples; see Methods for details on how P values were computed).
time scales, estimating the average TMRCA at each position along the genome of 498 WGS-phased samples from the Netherlands. Using this annotation in a stratified LD score regression analysis of 20 diseases and complex traits, we detected a strong enrichment for heritability in regions predicted to be under background selection; our annotation had the largest effect among available annotations for quantifying background selection.

High-throughput inference of ancestral relationships has a number of applications beyond those related to recent positive selection and disease heritability that we have pursued in this work. Genotype calling and imputation methods25–28, for instance, infer unobserved genotypes relying on ancestral relationships, which are usually estimated using computationally efficient approximations of the coalescent model (for example, the copying model34). Related ideas have been applied to detect phenotypic associations22–24. The processing speed achieved by the ASMC approach, on the other hand, enables making minimal simplifications to the full coalescent process, while retaining high computational scalability. In addition, accurate detection of very recent common ancestors (IBD regions) across samples is a key component of several other types of analysis, including long-range phasing47–50, estimation of recombination rates using haplotype boundaries4–7,61, haplotype-based association studies77, estimation of mutation and gene conversion rates79. In addition, the linear-time forward–backward algorithm of ASMC can be leveraged to scale up demographic inference in both WGS and SNP array data. The use of this approach in large SNP datasets, in particular, will enable the accurate inference of fine-scale demographic history within the past tens of generations, improving on methods that focus on summary statistics of shared long-range haplotypes79–81, rather than directly estimating recent coalescence rates.

Although the ASMC offers new opportunities for inference of pairwise coalescence times, we note several limitations. First, the ASMC can operate either on pairs of unphased chromosomes within a single diploid individual, or on pairs of phased chromosomes across individuals. To prevent biases1, the latter application requires haplotypes phased with extremely high accuracy, which may be difficult to obtain. In this work, extremely accurate phasing was possible in the UK Biobank dataset because of the very large sample size paired with the Eagle phasing algorithm24 (in the order of one switch error every ~10 cM; see also refs 71,81), and also possible in the GoNL dataset by leveraging trio information. Second, ASMC assumes a demographic model that includes a single panmictic population, and does not allow for the presence of samples from multiple ethnic backgrounds. Analyses of multi-ethnic
samples will require extending the current approach so that it can accommodate demographic models involving multiple populations. Furthermore, ancestry-specific SNP ascertainment may lead to depletion of high-frequency markers, creating an upward bias (Supplementary Fig. 3). Third, ASMC is not currently applicable to imputed data. We have shown that higher genotyping density leads to higher accuracy in the inference of coalescence times. However, our preliminary tests involving the use of ASMC on imputed data for which only markers with high-quality imputation accuracy are retained (for example, imputation $r^2 > 0.99$) resulted in substantial upward biases of the inferred coalescence times, which are due to spurious genotype calls. Effectively extending the ASMC to handle imputed data will thus require additional modeling of imputation accuracy. Fourth, our approach to assess the statistical significance of loci that are under recent positive selection is based on approximate $P$-value calculations. The use of approximate $P$ values has previously been used to detect signals of positive selection\(^5\), and is more conservative than the commonly used approach of simply ranking top loci\(^5\); nonetheless, the construction of an improved null model is a desirable direction of future development\(^5\). Finally, we note that although the speed of ASMC enables the analysis of large datasets, the computational costs of inferring pairwise coalescence times scale quadratically with the number of analyzed individuals. It may be possible to improve on this quadratic scaling given that at each location in the genome the ancestral relationships of a set of $n$ samples can be efficiently represented using a tree-shaped genealogy containing $n - 1$ nodes. However, the task of efficiently reconstructing a samples’ ancestral recombination graph\(^2\) is substantially more complex than that of estimating pairwise TMRCAs, and remains an exciting direction of future research. Despite these limitations and avenues for further improvement, we expect that ASMC will be a valuable tool for computationally efficient inference of pairwise coalescence times using SNP array or WGS data.

**URLs.** ASMC software, http://www.palamaralab.org/software/ASMC, https://github.com/pierpal/ASMC; UK Biobank, http://www.ukbiobank.ac.uk/; Genome of the Netherlands, www.nlgenome.nl; UK Biobank Genotyping and quality control, http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank-genotyping_QC_documentation-web.pdf; human genetic maps, http://www.shapeit.fr/files/genetic_map_b37.tar.gz; The dbPSH database of positive selection, ftp://jiwanglab.org/dbPSH/curation/dbPSH_20131001.tab; Python’s Scipy library, http://www.scipy.org/; 1000 Genomes Project Phase 3 data, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502; SMCC++ program, https://github.com/popgenmethods/smcpp; ARGON simulator, https://github.com/pierpal/ARGON; Simupop software, http://simupop.sourceforge.net/; Selscan software, https://github.com/szpiech/selscan; SLiM simulator, https://messerlab.org.slim/

**Methods**

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0177-x.

Received: 21 August 2017; Accepted: 21 June 2018; Published online: 13 August 2018

**References**

1. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496 (2011).
2. Schifflers, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. Nat. Genet. 46, 919–925 (2014).
3. 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).
4. Hobolth, A., Christensen, O. F., Mailund, T. & Schierup, M. H. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3, e7 (2007).
5. Sheehan, S., Harris, K. & Song, Y. S. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. Genetics 194, 647–662 (2013).
6. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10, e1004342 (2014).
7. Hudson, R. R. & Kaplan, N. L. The coalescent process in models with selection and recombination. Genetics 120, 831–840 (1988).
8. Wuf, C. & Hein, J. Recombination as a point process along sequences. Theor. Popul. Biol. 55, 248–259 (1999).
9. McVean, G. A. & Cardin, N. J. Approximating the coalescent with recombination. Philos. Trans. R. Soc. Lond. B 360, 1387–1393 (2005).
10. Marjoram, P. & Wall, J. D. Fast “coalescent” simulation. BMC Genet. 7, 16 (2006).
11. Hobolth, A. & Jensen, J. L. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theor. Popul. Biol. 85, 48–58 (2014).
12. 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015).
13. Skoglund, P. et al. Genetic evidence for two founding populations of the Americas. Nature 525, 104–108 (2015).
14. Raghavan, M. et al. Genomewide evidence for the Pleistocene and recent population history of Native Americans. Science 349, aaab884 (2015).
15. Green, R. E. et al. A draft sequence of the Neanderthal genome. Science 328, 710–722 (2010).
16. Prüfer, K. et al. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature 505, 43–49 (2014).
17. Sankararaman, S. et al. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507, 354–357 (2014).
18. Vernot, B. & Akey, J. M. Resurrecting surviving Neandertal lineages from modern human genomes. Science 334, 1017–1021 (2014).
19. Tennessen, J. A. et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337, 64–69 (2012).
20. Stewart, J. R. & Stringer, C. B. Human evolution out of Africa: the role of refuge and climate change. Science 335, 1317–1321 (2012).
21. Hunter-Zincz, H. & Clark, A. G. Aberrant time to most recent common ancestor as a signature of natural selection. Mol. Biol. Evol. 32, 2784–2797 (2015).
22. Morris, A. P., Whittaker, J. C. & Balding, D. J. Fine-scale mapping of disease loci via shuffled coalescent modeling of genealogies. Am. J. Hum. Genet. 70, 686–707 (2002).
23. Zollner, S. & Pritchard, J. K. Coalescent-based association mapping and fine mapping of complex trait loci. Genetics 169, 1071–1092 (2005).
24. Minichiello, M. J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. Am. J. Hum. Genet. 79, 910–922 (2006).
25. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529 (2009).
26. Fuchsberger, C., Abecasis, G. R. & Hinds, D. A. minimac2: faster genotype imputation. Bioinformatics 31, 782–784 (2015).
27. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multithreaded method for genotype-wide association studies by imputation of genotypes. Nat. Genet. 39, 906–913 (2007).
28. Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Res. 21, 952–960 (2011).
29. Sudlow, C. et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. 12, e1001779 (2015).
30. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat. Genet. 46, 818–825 (2014).
31. Palamara, P. F. ARGON: fast, whole-genome simulation of the discrete time Wright–Fisher process. Bioinformatics 32, 3032–3034 (2016).
32. Wakeley, J. & Wilson, P. In Encyclopedia of Evolutionary Biology Vol. 1 (ed. Kliman, R. M.) 287–292 (Oxford Academic, Oxford, 2016).
33. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194, 459–471 (2013).
34. Loh, P. R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. 48, 811–816 (2016).
35. Banshod, M. & Wooding, S. P. Signatures of natural selection in the human genome. Nat. Rev. Genet. 4, 99–111 (2003).
36. 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).
37. Field, Y. et al. Detection of human adaptation during the past 2000 years. Science 354, 760–764 (2016).
38. Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. Am. J. Hum. Genet. 74, 1111–1120 (2004).
39. Barreiro, L. B. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. 11, 17–30 (2010).
40. The Welcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007).
41. Sabeti, P. C. et al. Genome-wide detection and characterization of positive selection in human populations. Nature 449, 913–918 (2007).
42. Thierfelder, W. E. et al. Requirement for Stat4 in interleukin-12-mediated responses of natural killer and T cells. Nature 382, 171–174 (1996).
43. Li, Y. L. et al. Association of STAT4 rs7574865 polymorphism with autoimmune diseases: a meta-analysis. Mol. Biol. Rep. 39, 8873–8882 (2012).
44. Kobayashi, S. et al. Association of STAT4 with susceptibility to rheumatoid arthritis and systemic lupus erythematosus in the Japanese population. Arthritis Rheum. 58, 1940–1946 (2008).
45. Konno, H. D. S. L., Gregersen, P. K. & Remmers, E. F. STAT4: genetics, mechanisms, and implications for autoimmunity. Curr. Allergy Asthma Rep. 8, 398–403 (2008).
46. Gendler, S. J. & Spencer, A. P. Epithelial mucin genes. Annu. Rev. Physiol. 57, 607–634 (1995).
47. Kufe, D. W. Mucins in cancer: function, prognosis and therapy. Nat. Rev. Cancer 9, 874–885 (2009).
48. Seibold, M. A. et al. A common MUC5B promoter polymorphism and pulmonary fibrosis. N. Engl. J. Med. 364, 1503–1512 (2011).
49. Ishimaru, Y. et al. Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. Proc. Natl Acad. Sci. USA 105, 12569–12574 (2008).
50. Li, A., Tian, X., Sung, S. W. & Somlo, S. Identification of two novel polycystic kidney disease-1-like genes in human and mouse genomes. Genomics 81, 596–608 (2003).
51. Ishimaru, Y. et al. Interaction between PKD1L3 and PKD2L1 through their transmembrane domains is required for localization of PKD2L1 at taste pores in cells of circumvallate and foliate papillae. PASEB J. 24, 4058–4067 (2010).
52. Gudbjartsson, D. F. et al. Large-scale whole-genome sequencing of the Icelandic population. Nat. Genet. 47, 435–444 (2015).
53. Raynal, P. & Pollard, H. B. Annexins: the problem of assessing the biological role for a gene family of multifunctional calcium- and phospholipid-binding proteins. Biochim. Biophys. Acta 1197, 63–93 (1994).
54. Wu, N., Liu, S., Guo, C., Hou, Z. & Sun, M. Z. The role of annexin A3 mutation and gene-conversion rates. Genomics 83, 727–734 (2004).
55. Sturm, R. A. et al. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue–brown eye color. Am. J. Hum. Genet. 82, 424–431 (2008).
56. Mathieson, I. et al. Genome-wide patterns of selection in 230 ancient populations and individuals. Nature 467, 1099–1103 (2010).
57. Hinch, A. G. et al. The landscape of recombination in African Americans. Nature 476, 170–175 (2011).
58. Wegmann, D. et al. Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. 43, 847–853 (2011).
59. Gusev, A. et al. DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. Am. J. Hum. Genet. 88, 706–711 (2011).
60. Palamara, P. F. et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. Am. J. Hum. Genet. 97, 775–789 (2015).
61. Palamara, P. F., Lencz, T., Darvasi, A. & Pê’er, I. Length distributions of identity by descent reveal fine-scale demographic history. Am. J. Hum. Genet. 91, 809–822 (2012).
62. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. PLoS Biol. 11, e1001555 (2013).
63. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. Am. J. Hum. Genet. 97, 404–418 (2015).
64. Nei, M., Suzuki, Y. & Nozawa, M. The neutral theory of molecular evolution in the genomic era. Annu. Rev. Genom. Hum. Genet. 11, 265–289 (2010).
65. Griffiths, R. C. & Marjoram, P. An ancestral recombination graph. In Progress in Population Genetics and Human Evolution 257–270 (Springer, New York, 1997).

Acknowledgements
We thank P. K. Loh for suggesting several coding improvements for the ASMC software, and for support with the phasing and processing of the UK Biobank data; S. Gazal for support with the S-LDSC analysis and the baselineLD model; I. Shykhet for support with the COSI2 simulator; Y. Field for support with the simulation setup in the analysis of recent positive selection; D. Reich for providing computational resources; H. Finucane, Y. Reshef and A. Gusev for helpful discussions. This research was conducted using publicly available datasets (see URLs): the UK Biobank Resource under Application #16549, and the Genome of the Netherlands resource under Application #2017149. We thank the participants of the UK Biobank and the Genome of the Netherlands projects. P.E. and A.L.P. were supported by NIH grants R01 MH101244, R01 HG06399 and R01 GM105857; J.T. and Y.S.S. were supported in part by NIH grant R01 GM094402 and a Packard Fellowship for Science and Engineering; Y.S.S. is a Chan Zuckerberg Biohub investigator.

Author contributions
P.E. and A.L.P. conceived the study and analyzed results. P.E. developed the ASMC algorithm, performed simulations and data analysis. J.T. and Y.S.S. developed the CSFS algorithm, P.F.P. and A.L.P. wrote the manuscript with comments from J.T. and Y.S.S.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0177-z.
Reprints and permissions information is available at www.nature.com/reprints.

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

We provide an overview of the main components of the ASMC approach. An extended description can be found in the Supplementary Note.

ASMC model overview. The ASMC is a coalescent-based HMM (see the Supplementary Note for background on related methods). At each site along the genome, hidden states represent the time at which a pair of analyzed haploid individuals coalesce, which we also refer to as TMRCAs. In this model, time is discretized using a set of user-specified time intervals, each representing a particular state. The TMRCAs can change between adjacent states whenever a recombination event occurs along the lineages connecting the two individuals to their MRCA. The transition probability between states is modeled using a Markovian approximation of the full coalescent process. Observations are estimated using the general approach of the colored HMM, with details provided in the Supplementary Note.

Emission model. The emission model calculations of ASMC rely on the genome-wide allele frequency spectrum (AFS), which is estimated using either WGS or SNP array platforms. The AFS run on available WGS data for the analyzed population) and provided as input. In addition to this, we have to estimate the frequency and its age, which is modeled using the set of undistinguished samples and used to improve the inference of TMRCAs for the distinguished pairs. Because the set of undistinguished samples is solely used to obtain allele frequencies, the ancestral relationships need not be tracked, leading to a substantially simplified and tractable model. In the ASMC, this approach is extended to accommodate the fact that the observed sites may not be a randomly ascertained subset of polymorphic variants in the sample. To this end, we write the emission probability as \( P(o|d+u) \times P(d|u) \), where the additional term \( P(o|d+u) \) represents the probability that a site with \( d+u \) is observed in the ascertained data. In the ASMC, this probability is estimated in the same way as the empirical allele frequency spectrum obtained from the analyzed data and the allele frequency spectrum that is expected under neutrality for the demographic model. To achieve increased resolution, we also used a larger number of time intervals provides increased resolution, the choice of time discretization should take into account that a larger number of time intervals typically results in noisier MAP estimates of TMRCAs (Supplementary Table 5).

The transition model. The transition model describes the probability of transitioning along the genome between any pair of the \( s \) possible time intervals for the TMRCAs of the two analyzed samples (which we refer to as distinguished individuals in the emission model). These transition probabilities are computed using the conditional posterior Church model (CSC) [23]. In contrast to previously proposed Markovian approximations of the coalescent process, such as the SMC [24] and the SMC* [25], the CSC model remains accurate even if the observed genotypes are distant from one another. This is an important requirement for the analysis of SNP array data, as markers in this type of data are separated by substantially larger genetic distances than is the case for WGS data. Details for the calculation of transition probabilities can be found in the Supplementary Note.

Inference. The standard HMM forward–backward algorithm to perform posterior inference [15] has computational cost \( O(s^m) \) for analysis using \( s \) hidden states in a sequence of length \( m \). Current analyses that use coalescent HMMs to infer demographic histories utilize a number of hidden states in the order of \( 10^4 \). When human WGS datasets are analyzed, the number of observed sites is in the order of \( 10^6 \). Thus, the computational cost of applying the standard HMM approach is very high and a number of solutions to speed up the inference have been proposed (see the Supplementary Note for an overview). Here, we devise a new approach that uses dynamic programming to reduce the computational dependence on the number \( s \) of hidden states from quadratic to linear, resulting in a gain of two orders of magnitude for the average analysis compared to the standard algorithm. A related procedure exists for the SMC transition model [26], but cannot be applied to the more accurate and more complex CSC approach used in this work. The speedup in the HMM forward algorithm is obtained by computing an updated vector of emission probabilities using the current forward vector, \( \alpha \), and the transition matrix, \( T \), which is obtained from the CDC model. Computing the \( n \)th entry of this vector normally requires performing the summation \( \alpha_n = \sum_{l=1}^{s} \alpha_l T_{ln} \), which has computational cost \( O(s) \). This operation, however, can be rewritten as a linear combination of three terms, each of which can be recursively computed in time \( O(1) \), reducing the cost of computing the entire forward vector from \( O(s^2) \) to \( O(s) \) (see the Supplementary Note for a detailed derivation). An equivalent speedup can be obtained for the backward algorithm. Furthermore, to reduce the dependence of ASMC’s running time on sequence length when WGS data are analyzed, we use a more accurate approximation (see below). This in turn introduces a larger number of time intervals. Details are provided in the Supplementary Note.

ASMC simulations. We performed extensive coalescent simulations to assess the accuracy of the ASMC method. Unless otherwise specified, all simulations use the setup described in this section (standard setup). We used the ARGON simulator ([4,46], incorporating a human recombination rate map (see URL) and a recent demographic model for European individuals). We simulated 300 haploid individuals and a region of 30 Mb. To simulate SNP array data, we subsampled polymorphic variants to match the genotype density and allele frequency spectrum observed in the UK Biobank dataset (described below). We used recombination rates from the first 30 Mb of chromosome 2, for which the average rate of 1.66 cM per Mb accurately represents the recombination rates observed along the genome (mean = 1.45 cM per Mb, s.d. = 0.33 cM per Mb across autosomes). The demographic model and genetic map used to simulate the data were used when running ASMC, unless otherwise specified.

Time discretization. We ran ASMC using different numbers of discrete time intervals, which were chosen to correspond to quantiles of the pairwise coalescence distribution induced by the demographic model. To achieve increased resolution into the recent past, some simulations utilized 160 discretization intervals chosen as follows: 40 intervals of length 10 between generations 0 and 400, 80 intervals of length 20 between generations 400 and 2,000, and 40 intervals corresponding to every 10 generations to the coalescence distribution, starting at generation 2,000. Although using a larger number of time intervals provides increased resolution, the choice of time discretization should take into account that a larger number of time intervals typically results in noisier MAP estimates of TMRCAs (Supplementary Table 5).

Accuracy evaluation. The accuracy inference of ASMC was evaluated using two metrics. For a given region, and for all pairs of samples in a simulated dataset, we computed the squared correlation \( r^2 \) between the true and inferred sum of TMRCAs at each site within the region. This metric captures the accuracy of inferred genetic kinship, but is unchanged by potential scaling factors and possible systematic biases in the TMRCAs estimates. We also computed the difference between true and inferred TMRCAs at individual sites, which we usually report as a percentage difference compared to analysis of WGS data for improved readability. For our analysis of IBD detection accuracy, we defined as true IBD regions all sites for which pairwise TMRCAs were lower than a specified time threshold (note that several definitions exist for IBD sharing among unrelated individuals) and that IBD is sometimes also defined as the set of sufficiently long genomic regions for which two chromosomes share a common ancestor uninterrupted by recombination [27]. We ran Beagle [28] (v.4.1) providing the true genetic map and using default parameters, and used threshold values for the output LDF score (LDF score: base 10 log of the likelihood ratio, “ibd/autosomal program”) to infer the set of inferred IBD sites. To detect IBD using ASMC, we obtained MAP estimates of TMRCAs at all sites using 160 discretization intervals (see “Time discretization”), and used thresholds on the inferred TMRCAs values to select the set of inferred IBD sites. For both methods, we computed accuracy using the precision-recall curve. Neither Beagle nor ASMC enable obtaining recall values in the full \([0,1]\) range, due to the presence of a lower bound for Beagle's
admissible LOD threshold values, and ASMC’s time discretization. To compare the accuracies of the two methods for each simulation, we computed the area under the precision-recall curve only within the range in which the accuracy of both methods could be measured, and reported the percentage difference between the area under the precision-recall curve for the two methods (Supplementary Fig. 11). The PRC curve between observed points was interpolated using a previously published method64.

Model misclassifications. To mimic inaccuracies in the genetic map, we simulated data using a human recombination map for the simulated region, but with ASMC using a map with added noise. To introduce noise, the recombination rate between each pair of contiguous markers in the map was altered by randomly adding or subtracting a fraction of its true value (Supplementary Table 4). To test whether deviations from the assumption of frequency-based ascertainment introduce significant biases, we repeated the simulation over-ascertainment of rare variants in genetic regions of the genome. To this end, we randomly sampled around 25% of the markers from 10-kb-long genes placed every 200 kb, while the remaining variants were sampled to match the UK Biobank frequency spectrum as for the standard simulations, and compared the distribution of coalescent times within over-ascertained regions and the rest of the genome (Supplementary Fig. 2). To test the robustness of ASMC to an accurate demographic model, we simulated data under a European demographic history, but ran ASMC assuming a constant population size of 10,000 diploid individuals (Supplementary Table 2). To test the effects of ancestry-specific SNP ascertainment, we simulated an analysis for which a group of individuals is genotyped using an array that has been designed using a different, highly diverged population. To this end, we simulated two populations that split 2,000 generations (or approximately 60,000 years) in the past. The two populations have identical, European-like effective population size histories after the split, and a symmetric migration rate of 0.0, \(3 \times 10^{-5}\) or \(1 \times 10^{-7}\) per generation. We simulated ancestry-specific marker ascertainment by selecting SNPs based on frequencies from only one of the two populations, matching the spectrum observed in the UK Biobank. We then inferred coalescence times in both populations independently as described in the experiments involving a single population. Results are reported in Supplementary Fig. 3.

UK Biobank dataset. The UK Biobank interim release data comprise 152,729 samples, from which we extracted 113,851 individuals of British ancestry (as described in ref. 37). In addition, 95 trio parents were excluded and used to assess phasing quality with the Eagle software, leaving a total of 113,756 samples. From these, we created 11 batches with 10,000 samples and 1 batch with the remaining 3,756 samples, which we analyzed using ASMC. Out of the original around 800,000 variants (for basic quality control details see URLs), we analyzed a total of 678,956 SNPs that were autosomal, polymorphic in the set of analyzed samples, biallelic, with \(\leq 10\%\) missing values, and not included in a set of 65 variants with significantly different allele frequencies between the UK BiLEVE array and the UK Biobank array. We divided the genome in 39 autosomal regions from different chromosomes or separated by centromeres.

Detection of recent positive selection. To detect the occurrence of recent positive selection, we computed a statistic related to the density of recent coalescence events within the past \(T\) generations (DRC, statistic). The DRC statistic was measured as follows. At a given site along the genome, we first averaged all posterior TMRCA estimates obtained from all analyzed pairs of samples and renormalized these averages to obtain an average pairwise coalescence distribution times \(T\). The DRC statistic was then obtained by integrating this distribution between generations 0 and \(T\). The statistic was measured in windows of 0.05 cm, reporting an average of all SNPs within each window. We tested the sensitivity of the DRC statistic in detecting recent positive selection using extensive simulation. Details for these simulations can be found in the Supplementary Note.

Null model calibration. Given \(n\) samples from a population of recent effective size \(N\), the DRC, statistic is approximately Gamma-distributed under the null for sufficiently small values of \(T\) and \(n < N\). The rationale of this approximation is that for \(n < N\), a small number of coalescence events will have occurred within the short time span of \(T\) generations. In this regime, the coalescence time of each pair of lineages may be modeled as independent and exponentially distributed, which enables the approximation of the total number of early coalescence events as a Gamma-distributed random variable. Similar approximations have been recently used elsewhere64–66. We thus computed approximate \(P\) values for our selection scan in the UK Biobank dataset using the following approach. We first extracted a subset of the autosomal genome regions that span 18% of the genome and defined as any genotyped site at a distance greater than 500 kb from regions contained in a recent database of positive selection31 (see URLs). We then built an empirical null model by fitting a Gamma distribution (using Python’s Scipy library, see URLs) to these putatively neutral regions, and used this model to obtain approximate \(P\) values for the ASMC-estimated TMRCA at 100 windows, using a Bonferroni significance threshold of \(0.05/63,103\). We thus iterated this procedure, excluding this locus from the set of putatively neutral loci.

GoNL dataset. The GoNL dataset consists of 748 individuals who passed quality control and were sequenced at an average of approximately 13x (quality control details for the release 4 data are described elsewhere67). We analyzed 19,730,834 sequenced variants for 498 trio-phased unrelated parents, excluding centromeres and dividing the genome into the same 39 autosomal regions used for analysis of the UK Biobank dataset.

The ASMC\(_\text{sim}\) annotation. We estimated the strength of background selection by measuring variation in local effective population size along the genome68. We used ASMC to estimate the posterior mean TMRCA at all sites and for all pairs of haploid individuals in the GoNL dataset. We averaged these estimates at each site to obtain an annotation of background selection, which we refer to as ASMC\(_\text{sim}\). We similarly computed other annotations, conditioning on whether either or both individuals at a site carried a mutated allele. The ASMC\(_\text{sim}\) annotation (Supplementary Fig. 9), was obtained by averaging at each site the posterior mean TMRCA estimates for all pairs of individuals that were found to be heterozygous at each site. Other annotations were similarly computed using only pairs carrying, for example, minor/major alleles at each site (Supplementary Fig. 10). We verified that the ASMC\(_\text{sim}\) annotation captures the effects of natural selection using forward simulation. Details for these simulations can be found in the Supplementary Note.

S-LDSC analysis. We investigated whether large values of our annotations related to background selection corresponded to an enrichment in heritability for 20 complex traits and diseases listed in Supplementary Table 8. The S-LDSC analysis was run on datasets containing European individuals using standard guidelines69. The sets of LD-score, regression and heritability SNPs were defined as follows. LD-score SNPs were set to be 9,997,231 biallelic SNPs with at least 5 minor alleles observed in 489 European samples from the 1000 Genomes Phase 3 dataset (see URLs); regression SNPs were set as 1,217,312 HapMap Project Phase 3 SNPs; and heritability SNPs, used to compute trait heritability, were chosen as the 5,961,159 reference SNPs with MAF \(\geq 0.05\). The MHC region (2Mb 25–34 on chromosome 6) and SNPs with \(P > 80\) or 0.0001 were excluded from the analysis. Annotations contained in the baselineLD model, which we included in our joint analyses, can be found in supplementary table 9 of ref. 68. To avoid MAF-mediated effects, all ASMC-related annotations used in the S-LDSC analysis were quantile-normalized with respect to the MAF of regression SNPs. Specifically, we used 10 MAF ranges specified in the baselineLD model, corresponding to 10 frequency quantiles for the regression SNPs. For each range, we ranked values of an annotation for the corresponding SNPs, and mapped them to quantiles of a standard normal distribution. Annotation effects, \(\tau^2\), were obtained from the output of S-LDSC, as described previously68. Independent traits were selected on the basis of low genetic correlation, as previously described69. Meta-analysis of \(\tau^2\) values across independent traits was performed computing a weighted average of individual estimates of \(\tau^2\), weighted using 1/(\(\tau^2_i\)), where \(\tau^2_i\) represents heritability for the ith trait, and \(\tau\) represents the standard error of the trait’s \(\tau^2\) estimate.

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

Code availability. The ASMC program and source code can be downloaded at http://www.palamaralab.org/software/ASMC and https://github.com/pierpal/ASMC.

Data availability. Genomic annotations of positive and background selection can be downloaded at http://www.palamaralab.org/software/ASMC.

References
84. Simonsen, K. L. & Churchill, G. A. A Markov chain model of coalescence with recombination. Theor. Popul. Biol. 52, 43–59 (1997).
85. Rabiner, L. R. & Juang, B.-H. An introduction to hidden Markov models. IEEE ASSP Mag. 3, 4–16 (1986).
86. Harris, K., Sheehan, S., Kamm, J. A. & Song, Y. Decoding coalescent hidden Markov models in linear time. Res. Comput. Mol. Biol. 8394, 100–114 (2014).
87. Browning, B. L. & Browning, S. R. Detecting identity by descent and estimating genotype error rates in sequence data. Am. J. Hum. Genet. 93, 840–851 (2013).
88. Davis, J. & Goodrich, M. The relationship between precision-recall and ROC curves. In Proc. 23rd Int. Conf. on Machine Learning 233–240 (ACM, 2006).
89. Galinsky, K. J., Loh, P. R., Mallick, S., Patterson, N. J. & Price, A. L. Population structure of UK Biobank and ancient Eurasians reveals adaptation at genes influencing blood pressure. Am. J. Hum. Genet. 99, 1130–1139 (2016).
90. Mathieson, I. & McVean, G. Demography and the age of rare variants. PLoS Genet. 10, e1004528 (2014).
91. Li, M. J. et al. dbPSHP: a database of recent positive selection across human populations. Nucleic Acids Res. 42, D910–D916 (2014).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

### Experimental design

1. **Sample size**
   - Describe how sample size was determined.
   - Sample size was determined based on the number of genotyped individuals that were analyzed. Sample size of specific experiments was determined based on number of independent replicates, as reported.

2. **Data exclusions**
   - Describe any data exclusions.
   - Samples were excluded using commonly adopted criteria on genetic ancestry and relatedness. No data was excluded following this preliminary analysis.

3. **Replication**
   - Describe whether the experimental findings were reliably reproduced.
   - Experimental findings were obtained using two large publicly available data sets and were not reproduced on additional data sets.

4. **Randomization**
   - Describe how samples/organisms/participants were allocated into experimental groups.
   - Samples were randomly assigned to subgroups in some analyses.

5. **Blinding**
   - Describe whether the investigators were blinded to group allocation during data collection and/or analysis.
   - No. Grouping was performed by random shuffling of a list of samples.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. **Statistical parameters**
   - For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).
   - **n/a**
   - **Confirmed**

   - The exact sample size \( (n) \) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
   - A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
   - A statement indicating how many times each experiment was replicated
   - The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
   - A description of any assumptions or corrections, such as an adjustment for multiple comparisons
   - The test results (e.g. \( P \) values) given as exact values whenever possible and with confidence intervals noted
   - A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
   - Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.
Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.

Custom program ASMC will be made publicly available as an open source software package. We also utilized ARGON (v0.1.160415), COSI2 (v2.0), SimuPOP (v1.1.8.3), SMC++ (pre-release), SLIM (v1.8), Selscan (v1.1.0b), Python (v3.6).

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No restrictions.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

No antibodies were used.

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

No eukaryotic cell lines were used.

b. Describe the method of cell line authentication used.

No eukaryotic cell lines were used.

c. Report whether the cell lines were tested for mycoplasma contamination.

No eukaryotic cell lines were used.

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

No eukaryotic cell lines were used.

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

No animals were used in the study.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

The study involved publicly available data sets (Genome of the Netherlands Project data and UK Biobank project data). No phenotypic data was directly analyzed in the study. Summary association statistics from other studies were utilized in the LDSC analysis.