

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

Pier Francesco Palamara^{1,2,3}, Jonathan Terhorst⁴, Yun S. Song^{5,6}, Alkes L. Price^{2,3}

1 - Department of Statistics, University of Oxford, Oxford, UK

2 - Department of Epidemiology, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

3 - Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

4 - Department of Statistics, University of Michigan, Ann Arbor, MI, USA

5 - Department of Statistics, Computer Science Division, University of California, Berkeley, Berkeley, CA, USA

6 - Chan Zuckerberg Biohub, San Francisco, CA 94158, USA

Correspondence: palamara@stats.ox.ac.uk, aprice@hsph.harvard.edu

Abstract

1 Interest in reconstructing demographic histories has motivated the development of methods to
2 estimate locus-specific pairwise coalescence times from whole-genome sequence data. We
3 developed a new method, ASMC, that can estimate coalescence times using only SNP array data,
4 and is 2-4 orders of magnitude faster than previous methods when sequencing data are available.
5 We were thus able to apply ASMC to 113,851 phased British samples from the UK Biobank,
6 aiming to detect recent positive selection by identifying loci with unusually high density of very
7 recent coalescence times. We detected 12 genome-wide significant signals, including 6 loci with
8 previous evidence of positive selection and 6 novel loci, consistent with coalescent simulations
9 showing that our approach is well-powered to detect recent positive selection. We also applied
10 ASMC to sequencing data from 498 Dutch individuals (Genome of the Netherlands data set) to
11 detect background selection at deeper time scales. We observed highly significant correlations
12 between average coalescence time inferred by ASMC and other measures of background
13 selection. We investigated whether this signal translated into an enrichment in disease and
14 complex trait heritability by analyzing summary association statistics from 20 independent
15 diseases and complex traits (average $N=86k$) using stratified LD score regression. Our
16 background selection annotation based on average coalescence time was strongly enriched for
17 heritability ($p = 7 \times 10^{-153}$) in a joint analysis conditioned on a broad set of functional annotations
18 (including other background selection annotations), meta-analyzed across traits; SNPs in the top

20% of our annotation were 3.8x enriched for heritability compared to the bottom 20%. These results underscore the widespread effects of background selection on disease and complex trait heritability.

Introduction

Recently developed methods such as the Pairwise Sequentially Markovian Coalescent (PSMC)¹ utilize Hidden Markov Models (HMM) to estimate the coalescence time of two homologous chromosomes at each position in the genome¹⁻⁶, leveraging previous advances in coalescent theory⁷⁻¹¹. These methods have been broadly applied to reconstructing demographic histories of human populations¹²⁻²⁰. More generally, methods for inferring ancestral relationships among individuals have potential applications to detecting signatures of natural selection²¹, genome-wide association studies²²⁻²⁴, and genotype calling and imputation²⁵⁻²⁸. However, all currently available methods for inferring pairwise coalescence times require whole genome sequencing (WGS) data, and can only be applied to small data sets 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 samples²⁹. We verified ASMC's accuracy using coalescent simulations, and determined that it is orders of magnitude faster than other methods when WGS data are available. Leveraging ASMC's speed, we analyzed SNP array and WGS data sets 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 data set²⁹, detecting 12 loci with unusually high density of very recent coalescence times as a result of recent positive selection at these sites. These include 6 known loci linked to nutrition, immune response, and pigmentation, as well as 6 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 data set³⁰ 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 $N=86k$), even when conditioning on a broad set of functional annotations (including other background selection annotations).

Results

Overview of 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 Hidden Markov Model (HMM), which is built using the coalescent with recombination process⁷⁻¹¹; 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 PSMC¹, the MSMC², and, in particular, the recently developed SMC++³. In contrast with these methods, however, ASMC's main objective 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 data sets. To this end, the ASMC improves over 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 **Online Methods** section; we have released open-source software implementing the method (see URLs).

Simulations

We assessed ASMC's accuracy in inferring locus-specific pairwise TMRCA from SNP array and WGS data via coalescent simulations using the ARGON software³¹. Briefly, we measured the correlation between true and inferred average TMRCA for all pairs of 300 individuals simulated using a European demographic model³, for a 30 Mb region with SNP density and allele frequencies matching those of the UK Biobank data set (**Figure 1**; see **Online 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 (e.g. $r^2=0.87$ at UK Biobank SNP array density), 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 Figure 1**). Inferring locus-specific TMRCA is closely related to the task of detecting genomic regions that are identical-by-descent (IBD), i.e. regions for which the true TMRCA is lower than a specified cut-off; ASMC attained higher IBD detection accuracy (area under the precision-recall curve) than the widely used Beagle IBD detection method³² (**Supplementary Table 1**).

We evaluated the robustness of ASMC to various types of model misspecification, including an inaccurate demographic model, inaccurate recombination rate map, and violations of the assumption of frequency-based SNP ascertainment. To evaluate the impact of using an inaccurate demographic model, we simulated data under a European demographic history, but assumed a constant effective population size when inferring TMRCA (see **Online Methods**). As expected, this introduced biases, decreasing the accuracy of inferred TMRCA as measured by the RMSE, but had a negligible effect on the correlation between true and inferred TMRCA (**Supplementary Table 2**). An inaccurate demographic model is thus likely to result in biased TMRCA estimates, but has little effect on the relative ranking of TMRCA along the genome. Consistent with this observation, IBD detection remained accurate when an incorrect demographic model was used (**Supplementary Table 3**). We used a similar approach to evaluate the impact of using an inaccurate recombination rate map (see **Online Methods**), observing only negligible effects on the accuracy of inferred TMRCA (**Supplementary Table 4**). We next tested the robustness of ASMC to violations of the assumption that observed polymorphisms are ascertained solely based on their frequency, by instead ascertaining more rare variants in certain regions (mimicking genic regions; see **Online Methods**). We found that the distribution of inferred TMRCA in these “genic” regions did not deviate substantially from other regions (**Supplementary Figure 2**). Finally, we evaluated the impact of varying the number s of discrete TMRCA intervals (i.e. states of the HMM); we observed that increasing s had only a minor

impact on posterior mean estimates of TMRCA, although the higher resolution led to noisier MAP estimates (**Supplementary Table 5**).

We next evaluated the running time and memory cost of ASMC. Letting s be the number of discrete TMRCA intervals (i.e. states of the HMM) and m be the number of observed polymorphic sites, ASMC has asymptotic running time $O(sm)$. In comparison, the SMC++ method, which was shown to be more computationally efficient than other coalescent-based methods³, has asymptotic running time $O(s^3m)$. Accordingly, we observed that the running time of ASMC was 2 to 4 orders of magnitude faster than SMC++ when applied to simulated WGS data, depending on the number of discrete TMRCA intervals (**Figure 2**). For example, analysis of a pair of simulated genomes using 100 discrete time intervals required 7.4 seconds on a single processor for ASMC, compared to 3.3 hours for SMC++. This speedup does not involve a significant loss in accuracy (**Supplementary Figure 3**). The memory cost of ASMC was also efficient compared to SMC++, scaling linearly with s (**Supplementary Figure 4**).

Application to 113,732 samples from the UK Biobank reveals signals of recent positive selection
ASMC's computational efficiency enables its application to analyses of TMRCA in large data sets. 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 QC and phased using Eagle³³ (see **Online Methods**); we note that phasing accuracy in this data set is very high, with average switch error rate on the order of 0.3% (one switch error every ~ 10 cM³³). 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 sought to identify genomic regions with an unusually high density of very recent inferred TMRCA events (i.e. within the past several thousand years). Such signals are expected at sites undergoing recent positive selection, since 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 expectation³⁴; approaches to detect selection based on distortions in inferred coalescence times have been recently applied at different time scales²¹. We thus computed a statistic, DRC_T , 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 DRC_T statistic under the null is approximately Gamma-distributed. We thus obtained approximate p-values for the DRC_T statistic by fitting a Gamma distribution to the null 18% of the genome obtained by conservatively excluding 500Kb windows around regions previously implicated in scans for positive selection (see **Online Methods**). Using coalescent simulations, we determined that DRC_{150} is highly sensitive in detecting signals of positive selection within the past ~20,000 years, as compared to other methods^{35,36} (see **Online Methods, Supplementary Figure 5**).

Analyzing 63,103 windows of length 0.05cM in the UK Biobank data set, we detected 12 genome-wide significant loci ($p < 0.05 / 63,103 = 7.9 \times 10^{-7}$; see **Figure 3A** and **Table 1**). The loci that we detected exhibited strong enrichment of recent coalescent events spanning up to the past 20,000 years (**Figure 3B, 3C** and **Supplementary Figure 6**), consistent with our simulations (**Supplementary Figure 5**). Of the 12 loci, 6 are loci known to be under recent positive selection, harboring genes linked to nutrition (LCT^{37}), immune response (HLA^{38} , TLR^{39} , $IGHG^{40}$), eye color ($GRM5^{40}$), and skin pigmentation ($MC1R^{40}$). We also detected 6 novel loci, harboring genes involved in immune response ($STAT4^{41}$, associated with autoimmune disease⁴²⁻⁴⁴); mucus production ($MUC5B^{45}$ within cluster of mucin genes, involved in protection against infectious disease⁴³, associated with several types of cancer⁴⁶ and lung disease⁴⁷); taste reception ($PKD1L3^{48}$, associated with kidney disease^{49,50}); cardiac and fetal muscle ($MYL4$, associated with atrial fibrillation⁵¹); blood coagulation ($ANXA3^{52}$, associated with cancer⁵³ and immune disease⁵⁴); and brain-specific expression and immune response ($FAM19A5^{55}$). We note that suggestive loci implicated by the DRC_{150} statistic ($p < 10^{-4}$; **Supplementary Table 6**) include known targets of selection linked to eye color ($HERC2^{56,57}$), retinal and cochlear function ($PCDH15^{40}$), celiac disease ($SLC22A4^{57,58}$) and skin pigmentation ($SLC45A2^{57}$).

Background selection annotation and heritability enrichment for complex traits

We next sought to detect signals of background 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 data set (498 samples and 19,730,834 variants after QC; see **Online Methods**); we note that WGS data are required to achieve accurate resolution at deeper time scales (**Figure 1A**). Motivated by the fact that background selection reduces the effective population size at affected regions^{34,59}, we

estimated its strength by measuring the average pairwise TMRCA at each site, which is proportional to effective population size⁶⁰. We refer to this background selection annotation as $ASMC_{avg}$. The genome-wide average of $ASMC_{avg}$ in the GoNL data was 17,399 generations (s.d. = 9,957 generations); we thus expect the $ASMC_{avg}$ annotation to reflect background selection occurring within the past several hundred thousand years. As expected, $ASMC_{avg}$ was highly correlated with other measures of background selection, including nucleotide diversity ($r=0.50$), the McVicker B-statistic⁵⁹ ($r=-0.28$), and allele age predicted by ARGWeaver⁶, quantile-normalized within 10 minor allele frequency bins⁶¹ ($r=0.26$, see **Supplementary Table 7**).

Analyses using stratified LD score regression (S-LDSC)⁶² have shown that regions under background selection are enriched for disease and complex trait heritability⁶¹; 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 $ASMC_{avg}$ 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 $N=86k$). We performed both an unconditioned analysis using only the $ASMC_{avg}$ annotation, and a joint analysis conditioned on the 75 annotations from the baselineLD model⁶¹ (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 $ASMC_{avg}$ annotation, we computed the τ^* metric⁶¹, defined as the proportionate change in per-SNP heritability resulting from a 1 standard deviation increase in the value of the annotation, conditional on other annotations included in the model.

In the unconditioned analysis, lower $ASMC_{avg}$ was associated with higher per-SNP heritability for all 20 traits analyzed (**Figure 4A**), confirming that regions under background selection are enriched for disease heritability. Meta-analyzed across the 20 traits, the τ^* for $ASMC_{avg}$ had a value of -0.81 (s.e. = 0.01; Z-test $p < 10^{-300}$). After conditioning on the baselineLD model, the τ^* for $ASMC_{avg}$ remained strongly significant at -0.25 (s.e. = 0.01; Z-test $p = 7 \times 10^{-153}$), implying that $ASMC_{avg}$ remains informative for disease heritability after conditioning on other annotations linked to background selection as well as a broad set of functional annotations. Furthermore, $ASMC_{avg}$ attained a larger value of τ^* than each of the other six annotations linked to

background selection (**Figure 4B**), implying that it was the most disease-informative background selection annotation in this analysis; we note that adding $ASMC_{avg}$ to the baselineLD model reduced the $|\tau^*|$ of the nucleotide diversity annotation from 0.13 to 0.00 and reduced the $|\tau^*|$ of the ARGWeaver⁶ predicted allele age annotation from 0.25 to 0.13, indicating that $ASMC_{avg}$ subsumes signals from these annotations. We computed the proportion of heritability explained by each quintile of the $ASMC_{avg}$ annotation, which provides a more intuitive interpretation of the strength of the annotation's effect (**Figure 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 with the nucleotide diversity annotation, whose effect was however subsumed by the $ASMC_{avg}$ annotation; **Figure 4B**). Annotations constructed based on average pairwise TMRCA conditional on the allele present on each chromosome were further informative for disease heritability (**Supplementary Figure 7** and **Supplementary Figure 8**; see **Online Methods**).

Discussion

We have introduced a new method for inferring pairwise coalescence times, ASMC, that can be applied to either SNP array or WGS data and is highly computationally efficient. Exploiting ASMC's speed, we analyzed ~2.2 billion pairs of haploid chromosomes from 113,851 British samples within the UK Biobank data set, and detected strong evidence of recent positive selection at 6 known loci and 6 novel loci linked to immune response and other biological functions. We further used ASMC to detect background selection at deeper 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 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 (e.g. the *copying model*⁶³). 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 phasing^{33,64,65}, estimation of recombination rates using haplotype boundaries⁶⁶⁻⁶⁸, haplotype-based association studies⁶⁹, estimation of mutation and gene conversion rates⁷⁰, and inference of fine-scale demographic history within the past tens of generations⁷¹⁻⁷³.

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. The latter application requires the availability of phased haplotypes, which may not be easy to obtain. We note, however, that recently developed methods for computational phasing attain sufficiently low switch error rates as to be considered negligible in most applications^{33,64,74}. 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. 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. It may be possible to extend ASMC to incorporate information on imputation accuracy, enabling its application to imputed data. Fourth, our approach to assess the statistical significance of loci under recent positive selection is based on approximate p-value calculations. The use of approximate p-values has previously been adopted in detecting signals of positive selection³⁶, and is more conservative than the widespread approach of simply ranking top loci³⁵; nonetheless, the construction of an improved null model is a desirable direction of future development⁷⁵. Finally, we note that although ASMC's speed enables the analysis of large data sets, 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. The task of efficiently reconstructing a samples' ancestral recombination graph (ARG)^{6,24,76}, however, is substantially more complex than that of estimating pairwise TMRCA, 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.

Online 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^{1-3,5} (see **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 their time to most recent common ancestor (TMRCA). In this model, time is discretized using a set of s user-specified time intervals, each representing a possible hidden state. The TMRCA may change between adjacent sites 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 approximation¹¹ of the full coalescent process. Observations are obtained using genotypes of the pair of analyzed samples, as well as a set of additional samples, as detailed below, and emission probabilities reflect the chance of observing a specific genotypic configuration, conditional on the pair's TMRCA at a site. Calculations of the initial state distribution, the transition, and the emission probabilities consider the demographic history of the analyzed sample, which is separately estimated (e.g. using other coalescent HMMs run on available WGS data for the analyzed population) and provided as input. The main goal of the ASMC is to perform high-throughput inference of posterior TMRCA probabilities along the genome for many pairs of haploid individuals genotyped using either WGS or SNP array platforms.

Emission model. ASMC's emission probability calculations rely on the recently developed Conditioned Sample Frequency Spectrum (CSFS)³, which is extended to handle non-randomly ascertained genotype observations (e.g. SNP array data). Consider a sample of n individuals, and define 2 of them as *distinguished*, $(n-2)$ of them as *undistinguished*. We are interested in estimating posterior TMRCA probabilities at a set of observed sites in the genome of the 2 distinguished samples. At each site, the CSFS model³ allows computing the HMM emission probability $P(d, u | \tau)$, i.e. the probability that

$d \in \{0, 1, 2\}$ derived alleles are carried by the distinguished pair of samples, while $u \in [0, n-2]$ derived alleles are observed in the $(n-2)$ undistinguished samples, conditioned on the fact that the distinguished pair's TMRCA (the HMM's hidden state) is τ . Intuitively, this approach enables exploiting the relationship between an allele's frequency and its age, which is modeled using the set of undistinguished samples and used to improve the inference of TMRCA for the distinguished pairs³. Because the set of undistinguished samples is solely used to obtain allele frequencies, their 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 $P(\text{obs}|d+u) \times P(d, u|\tau)$, where the additional term $P(\text{obs}|d+u)$ represents the probability that a site with $(d+u) \in [0, n]$ carriers of the derived allele is observed in the ascertained data. In the ASMC, this probability is estimated using the ratio between the empirical allele frequency spectrum obtained from the analyzed data and the allele frequency spectrum that is expected under neutrality for the demographic model provided in input. Details are provided in the **Supplementary Note**. The emission model enables handling both major/minor and ancestral/derived genotype data encoding. We verified using coalescent simulation (see **Simulations**), that the number of individuals used when computing the CSFS model does not have a substantial impact on accuracy (**Supplementary Table 10**).

Transition model. The transition model describes the probability of transitioning along the genome between any pair of the s possible time intervals for the TMRCA of the two analyzed samples (which we referred to as *distinguished* individuals in the emission model). These transition probabilities are computed using the conditional Simonsen-Churchill model (CSC)^{11,77}. In contrast to previously proposed Markovian approximations of the coalescent process, such as the SMC⁹ and the SMC'¹⁰, the CSC model remains accurate even if the observed genotypes are distant from one another¹¹. This is an important requirement in the analysis of SNP array data, as markers in this type of data are separated by substantially larger genetic distances than in the case of WGS data. Details on the calculation of transition probabilities can be found in the

Supplementary Note. ASMC supports variable recombination rates along the genome through a genetic map provided in input.

Inference. The standard HMM forward-backward algorithm to perform posterior inference has computational cost $O(s^2m)$ for analysis using s hidden states in a sequence of length m ⁷⁸. Current analyses making use of coalescent HMMs to infer demographic histories utilize a number of hidden states in the order of 10^2 . When human WGS data is analyzed, the number of observed sites is in the order of 10^9 . 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 **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 2 orders of magnitude for the average analysis compared to the standard algorithm. A related procedure exists for the SMC transition model⁷⁹, but cannot be applied to the more accurate and more complex CSC approach used in this work. The speed-up in the HMM forward algorithm is obtained by simplifying the key operation of computing an updated α' vector of forward probabilities using the current forward vector, α , and the transition matrix, T , which is obtained from the CDC model. Computing the i -th entry of this vector normally requires performing the summation $\alpha'_i = \sum_{k=1}^s \alpha_k T_{k,i}$, 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 speed-up 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 make the following approximation. Consider two polymorphic sites separated by a stretch of n monomorphic sites. Computing an updated forward probability vector α' using the standard approach would require performing the operation $\alpha' = \alpha(TE_0)^n TE_p$, where E_0 is a diagonal matrix with emission probabilities for a monomorphic site in its diagonal entries and E_p is the equivalent matrix for the emission at the next polymorphic site in the sequence. For short genetic distances that are observed between polymorphic sites, the matrix T is close to

diagonal, and we can thus effectively approximate this product as $\alpha T^n E^{n_0} T E_p$ (see **Supplementary Figure 3**). Using the previously described dynamic programming approach, this operation can be computed in time $O(s)$, and only needs to be performed at a subset of polymorphic sites, resulting in a further speedup of 2-3 orders of magnitude compared to the standard forward/backward approach operating on all sites. This approximation is not needed when SNP array data are analyzed, as we need not integrate over large stretches of monomorphic sites, treating instead all sites between a pair of genotyped SNPs as unobserved. In addition to this, most quantities involved in the $O(ms)$ forward/backward operations can be precomputed and stored in a cache, substantially reducing constant terms in the computation.

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 simulator³¹, incorporating a human recombination rate map (see **URLs**) and a recent demographic model for European individuals³. We simulated 300 haploid individuals and a region of 30Mb. To simulate SNP array data, we subsampled polymorphic variants to match the genotype density and allele frequency spectrum observed in the UK Biobank data set (described below). We used recombination rates from the first 30Mb of Chromosome 2, whose average rate of 1.66 cM/Mb well represents the recombination rates observed along the genome (mean 1.45 cM/Mb, s.d. 0.33 cM/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 quantiles of the coalescence distribution, starting at generation 2,000. While 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 TMRCA (see **Supplementary Table 5**).

Accuracy evaluation. ASMC's inference accuracy was evaluated using two metrics. For a given region, and for all pairs of samples in a simulated data set, we computed the squared correlation (r^2) between the true and inferred sum of TMRCA 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 TMRCA estimates. We thus also measured the root mean square error (RMSE) between true and inferred TMRCA at individuals sites, which we usually report as a percent 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 TMRCA were lower than a specified time threshold. We ran Beagle³² (v4.1) providing the true genetic map and using default parameters, and used threshold values for the output LOD score (*ibd lod*) to select the set of inferred IBD sites. To detect IBD using ASMC, we obtained MAP estimates of TMRCA at all sites using 160 discretization intervals (see **Time discretization**), and used thresholds on the inferred TMRCA 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 two methods' accuracies in each simulation, we computed the area under the precision-recall curve (auPRC) only within the range in which the accuracy of both methods could be measured, and reported the percent difference between the two methods' auPRC (see **Supplementary Figure 9** for an illustration). The PRC curve between observed points was interpolated using the method of ref⁸⁰.

Model misspecifications. To test ASMC's robustness 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 (see **Supplementary Table 2**). To

mimic inaccuracies in the genetic map we simulated data using a human recombination map for the simulated region, but run 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 (see **Supplementary Table 4**). To test whether deviations from the assumption of frequency-based ascertainment introduce significant biases, we mimicked over-ascertainment of rare variants in genic regions of the genome. To this end, we randomly sampled ~25% of the markers from 10Kb-long genes placed every 200Kb, while the remaining variants were sampled to match the UK Biobank frequency spectrum as in standard simulations, and compared the distribution of coalescent times within over-ascertained regions and the rest of the genome (**Supplementary Figure 2**).

UK Biobank (UKBB) data set. The UK Biobank interim release data comprise 152,729 samples, from which we extracted 113,851 individuals of British ancestry (as described in ref. ⁸¹). 95 trio parents were excluded and used to assess phasing quality with the Eagle³³ 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 ~800k variants (for basic quality control details see **URLs: UK Biobank Genotyping and QC**), we analyzed a total of 678,956 SNPs that were autosomal, polymorphic in the set of analyzed samples, biallelic, with missingness $\leq 10\%$, 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_T statistic). The DRC_T 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 at the site. The DRC_T 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.

Null model calibration. Given n samples from a population of recent effective size N , the DRC_T statistic is approximately Gamma-distributed under the null for sufficiently small values of T and $n \ll N$. The rationale of this approximation is that for $n \ll N$, 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 allows approximating the total number of *early* coalescence events as a Gamma-distributed random variable. Similar approximations have been recently used elsewhere^{36,82}. We thus computed approximate p-values for our selection scan in the UKBB data set using the following approach. We first extracted a subset of “neutral” genomic regions, spanning a total of 18% of the genome, and defined as any genotyped site at a distance greater than 500Kb from regions contained in a recent database of positive selection⁸³ (see **URLs**: database of positive selection). 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 throughout the genome. We analyzed 63,103 windows, using a Bonferroni significance threshold of $0.05 / 63,103 = 7.9 \times 10^{-7}$. One of the genome-wide significant signals that we detected (PKD1L3 locus, chr16:70.89-71.80Mb) fell within the putatively neutral portion of the genome. We thus iterated this procedure, excluding this locus from the set of putatively neutral loci.

Selection simulations. We used the simulation setup recently adopted by Field et al.³⁶ to test the sensitivity of the DRC_T statistic in detecting recent positive selection, and its specificity to recent time scales. We simulated several replicates for a region of 10Mb and 6,000 haploid individuals from a European demographic model¹⁹, using the COSI2 coalescent simulator⁸⁴. An allele at the center of the region was simulated to undergo recent positive selection, reaching a high present-day frequency of 0.7. We used the simuPOP⁸⁵ software to obtain allele frequency trajectories under additive selection models, for several values of the selection coefficient. To test for specificity to recent

time scales, we varied the period during which selection was active, posing no constraints on whether selection acted on a novel allele or on standing variation.

To assess power, we simulated 50 independent replicates for positive selection occurring in the past 200 generations (or ~6,000 years), using selection coefficients $S=0.01, 0.03, 0.05, 0.1$. We detected positive selection using either iHS (ref. ³⁵), SDS (ref. ³⁶), or DRC_{150} (**Supplementary Figure 5a**). The iHS statistic was computed using the Selscan software⁸⁶ with default parameters. We computed the iHS statistic at either the sequenced causal variant ($iHS_{sequence}$), or averaged at SNPs within a 0.05 cM window around the causal variant in simulated SNP array data (iHS_{array}), which we obtained from simulated sequencing data as detailed above for neutral simulations. The DRC_{150} statistic was similarly computed by averaging within a 0.05 cM window on SNP array data. The SDS statistic was computed at the sequenced causal variant ($SDS_{sequence}$). We found the DRC_{150} statistic computed on SNP array data to be highly sensitive to recent positive selection starting at $S=0.03$. Similar results for DRC_{20} are also reported in **Supplementary Figure 10a**.

To assess the specificity of DRC_{150} to recent time scale, we simulated selection starting at time $-\infty$ and ending at a generation in 0, 50, 100, 200, 400, 600, 800, 1000, 1500, 2000 (**Supplementary Figure 5b**). We observed the DRC_{150} statistic to be mostly sensitive to selection acting during the past ~700 generations (or ~20,000 years), a similar time-span compared to the iHS statistic computed at the sequenced causal variant, which was however generally less sensitive, while the SDS statistic computed at the sequenced causal variant was only sensitive to extremely recent positive selection, as previously shown³⁶. We also report DRC_{20} results in **Supplementary Figure 10b**.

We performed additional simulation to evaluate the calibration of the null model. We observed an excellent fit for the DRC_{20} statistic (**Supplementary Figure 11a**), and only moderate inflation for the DRC_{150} statistic (**Supplementary Figure 11b**). The amount of inflation observed in the empirical null model obtained using the DRC_{150} statistic within the UKBB data set was consistent with our coalescent simulations (**Supplementary**

Figure 11c,d). We note that for very small values of T the independence assumption is more accurately met, so that the DRC_T statistic is well approximated using a Normal distribution (see **Supplementary Figure 12** for DRC_{20}). We expect the moderate amount of inflation observed in neutral simulations for the DRC_{150} statistic to be counterbalanced in real data analysis by the conservative use of a Bonferroni significance threshold and the fitting of null model parameters using an empirical distribution of test statistics, which is likely to result in over-dispersion of the null model due to signals of positive selection that are too weak to be detected. Consistent with this hypothesis, genome-wide significant loci (**Table 1**) and suggestive loci (**Supplementary Table 6**) contain several regions of known recent adaptation.

Genome of the Netherlands (GoNL) data set. The data set consists of 748 individuals who passed quality control and were sequenced at an average of $\sim 13\times$ (quality control details for the Release 4 data are described elsewhere³⁰). We analyzed 19,730,834 sequenced variants for 498 trio-phased unrelated parents, excluding centromeres and dividing the genome in the same 39 autosomal regions used for analysis of the UKBB data set.

ASMC_{avg} annotation. We set out to estimate the strength of background selection by measuring variation in local effective population size along the genome⁵⁹. We used ASMC to estimate the posterior mean TMRCA at all sites and for all pairs of haploid individuals in the GoNL data set. We averaged these estimates at each site to obtain an annotation of background selection, which we refer to as ASMC_{avg}. We similarly computed other annotations, conditioning on whether either or both individuals at a site carried a mutated allele. The ASMC_{het} annotation (**Supplementary Figure 7**), 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 e.g. minor/major alleles at each site (see **Supplementary Figure 8**).

Stratified LD Score (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 data sets containing European individuals using standard guidelines⁶². 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 data set¹² (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 $\chi^2 > 80$ or $0.0001N$ were excluded from the analysis. Annotations contained in the baselineLD model, which we included in our joint analyses, can be found in **Supplementary Table S9 of ref.** ⁶¹. To avoid minor allele frequency (MAF)-mediated effects, all ASMC-related annotations used in the S-LDSC analysis were quantile-normalized with respect to 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, τ^* , were obtained from the output of S-LDSC, as described in ref. ⁶¹. Independent traits were selected on the basis of low genetic correlation, as previously described⁶². Meta-analysis of τ^* values across independent traits was performed computing a weighted average of individual estimates of τ^* , weighted using $1/(h_i^2 \epsilon_i^2)$, where h_i^2 represents heritability for the i -th trait, and ϵ_i represents the standard error of the trait's τ^* estimate.

Acknowledgements

We thank Po-Ru Loh for suggesting several coding improvements for the ASMC software, and for support with the phasing and processing of the UK Biobank data; Steven Gazal for support with the S-LDSC analysis and the baselineLD model; Ilya Shlyakhter for support with the COSI2 simulator; Yair Field for support with the simulation setup in the analysis of recent positive selection; David Reich for providing computational resources; Hilary Finucane, Yakir Reshef, and Sasha Gusev for helpful discussions. This research was conducted using publicly available data sets (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.P. and A.L.P. were supported by NIH grant R01 MH101244; J.T. and Y.S.S. were supported in part by an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering; Y.S.S. is a Chan Zuckerberg Biohub investigator.

URLs

- The ASMC program will be released prior to publication as a publicly available, open source software package at <http://www.palamaralab.org/software/> and <https://github.com/pierpal/ASMC>
- UK Biobank website: <http://www.ukbiobank.ac.uk/>
- Genome of the Netherlands website: www.nlgenome.nl
- UK Biobank Genotyping and QC: 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 dbPSHP database of positive selection: ftp://jjwanglab.org/dbPSHP/curation/dbPSHP_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>

References

1. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493-6 (2011).
2. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-25 (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**, 647-62 (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**, 831-40 (1988).
8. Wiuf, C. & Hein, J. Recombination as a point process along sequences. *Theor Popul Biol* **55**, 248-59 (1999).
9. McVean, G.A. & Cardin, N.J. Approximating the coalescent with recombination. *Philos Trans R Soc Lond B Biol Sci* **360**, 1387-93 (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* **98**, 48-58 (2014).
12. Genomes Project, C. *et al.* 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**, 104-8 (2015).
14. Raghavan, M. *et al.* POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
15. Green, R.E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710-22 (2010).
16. Prufer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43-9 (2014).
17. Sankararaman, S. *et al.* The genomic landscape of Neanderthal ancestry in present-day humans. *Nature* **507**, 354-7 (2014).
18. Vernot, B. & Akey, J.M. Resurrecting surviving Neandertal lineages from modern human genomes. *Science* **343**, 1017-21 (2014).
19. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
20. Stewart, J.R. & Stringer, C.B. Human evolution out of Africa: the role of refugia and climate change. *Science* **335**, 1317-21 (2012).

21. Hunter-Zinck, H. & Clark, A.G. Aberrant Time to Most Recent Common Ancestor as a Signature of Natural Selection. *Mol Biol Evol* **32**, 2784-97 (2015).
22. Morris, A.P., Whittaker, J.C. & Balding, D.J. Fine-scale mapping of disease loci via shattered 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**, 1071-92 (2005).
24. Minichiello, M.J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am J Hum Genet* **79**, 910-22 (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**, 782-4 (2015).
27. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906-13 (2007).
28. Le, S.Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* **21**, 952-60 (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, C. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818-25 (2014).
31. Palamara, P.F. ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinformatics* **32**, 3032-4 (2016).
32. Browning, B.L. & Browning, S.R. A fast, powerful method for detecting identity by descent. *Am J Hum Genet* **88**, 173-82 (2011).
33. Loh, P.R., Palamara, P.F. & Price, A.L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat Genet* **48**, 811-6 (2016).
34. Bamshad, M. & Wooding, S.P. Signatures of natural selection in the human genome. *Nat Rev Genet* **4**, 99-111 (2003).
35. 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).
36. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 (2016).
37. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* **74**, 1111-20 (2004).
38. 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).
39. Wellcome Trust Case Control, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-78 (2007).
40. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-8 (2007).
41. Thierfelder, W.E. *et al.* Requirement for Stat4 in interleukin-12-mediated responses of natural killer and T cells. *Nature* **382**, 171-4 (1996).

42. Liang, Y.L. *et al.* Association of STAT4 rs7574865 polymorphism with autoimmune diseases: a meta-analysis. *Mol Biol Rep* **39**, 8873-82 (2012).
43. Kobayashi, S. *et al.* Association of STAT4 with susceptibility to rheumatoid arthritis and systemic lupus erythematosus in the Japanese population. *Arthritis Rheum* **58**, 1940-6 (2008).
44. Korman, B.D., Kastner, D.L., Gregersen, P.K. & Remmers, E.F. STAT4: genetics, mechanisms, and implications for autoimmunity. *Curr Allergy Asthma Rep* **8**, 398-403 (2008).
45. Gendler, S.J. & Spicer, A.P. Epithelial mucin genes. *Annu Rev Physiol* **57**, 607-34 (1995).
46. Kufe, D.W. Mucins in cancer: function, prognosis and therapy. *Nat Rev Cancer* **9**, 874-85 (2009).
47. Seibold, M.A. *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis. *N Engl J Med* **364**, 1503-12 (2011).
48. Ishimaru, Y. *et al.* Transient receptor potential family members PKD1L3 and PKD2L1 form a candidate sour taste receptor. *Proc Natl Acad Sci U S A* **103**, 12569-74 (2006).
49. 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).
50. Ishimaru, Y. *et al.* Interaction between PKD1L3 and PKD2L1 through their transmembrane domains is required for localization of PKD2L1 at taste pores in taste cells of circumvallate and foliate papillae. *FASEB J* **24**, 4058-67 (2010).
51. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* **47**, 435-44 (2015).
52. 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).
53. Wu, N., Liu, S., Guo, C., Hou, Z. & Sun, M.Z. The role of annexin A3 playing in cancers. *Clin Transl Oncol* **15**, 106-10 (2013).
54. Okada, Y. *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat Genet* **44**, 511-6 (2012).
55. Tom Tang, Y. *et al.* TAFA: a novel secreted family with conserved cysteine residues and restricted expression in the brain. *Genomics* **83**, 727-34 (2004).
56. 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**, 424-31 (2008).
57. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499-503 (2015).
58. Huff, C.D. *et al.* Crohn's disease and genetic hitchhiking at IBD5. *Mol Biol Evol* **29**, 101-11 (2012).
59. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5**, e1000471 (2009).
60. Wakeley, J. *Coalescent theory : an introduction*, xii, 326 p. (Roberts & Co. Publishers, Greenwood Village, Colo., 2009).

61. Gazal, S. *et al.* Linkage disequilibrium dependent architecture of human complex traits reveals action of negative selection. *Nature Genetics* **In press**, preprint: <http://biorxiv.org/content/early/2017/04/03/082024> (2016).
62. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).
63. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-33 (2003).
64. Loh, P.R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet* **48**, 1443-1448 (2016).
65. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* **40**, 1068-75 (2008).
66. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-103 (2010).
67. Hinch, A.G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170-5 (2011).
68. Wegmann, D. *et al.* Recombination rates in admixed individuals identified by ancestry-based inference. *Nat Genet* **43**, 847-53 (2011).
69. Gusev, A. *et al.* DASH: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *Am J Hum Genet* **88**, 706-717 (2011).
70. Palamara, P.F. *et al.* Leveraging Distant Relatedness to Quantify Human Mutation and Gene-Conversion Rates. *Am J Hum Genet* **97**, 775-89 (2015).
71. Palamara, P.F., Lencz, T., Darvasi, A. & Pe'er, I. Length distributions of identity by descent reveal fine-scale demographic history. *Am J Hum Genet* **91**, 809-22 (2012).
72. Ralph, P. & Coop, G. The geography of recent genetic ancestry across Europe. *PLoS Biol* **11**, e1001555 (2013).
73. 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**, 404-18 (2015).
74. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**, 817-20 (2016).
75. Nei, M., Suzuki, Y. & Nozawa, M. The neutral theory of molecular evolution in the genomic era. *Annu Rev Genomics Hum Genet* **11**, 265-89 (2010).
76. Griffiths, R.C. & Marjoram, P. An ancestral recombination graph. *Institute for Mathematics and its Applications* **87**, 257 (1997).
77. Simonsen, K.L. & Churchill, G.A. A Markov Chain Model of Coalescence with Recombination. *Theor Popul Biol* **52**, 43-59 (1997).
78. Rabiner, L.R. & Juang, B.-H. An introduction to hidden Markov models. *ASSP Magazine, IEEE* **3**, 4-16 %@ 0740-7467 (1986).
79. Harris, K., Sheehan, S., Kamm, J.A. & Song, Y.S. Decoding coalescent hidden Markov models in linear time. *Res Comput Mol Biol* **8394**, 100-114 (2014).
80. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. 233-240 %@ 1595933832 (ACM, 2006).

81. 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).
82. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet* **10**, e1004528 (2014).
83. Li, M.J. *et al.* dbPSHP: a database of recent positive selection across human populations. *Nucleic Acids Res* **42**, D910-6 (2014).
84. Shlyakhter, I., Sabeti, P.C. & Schaffner, S.F. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30**, 3427-9 (2014).
85. Peng, B. & Amos, C.I. Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics* **24**, 1408-9 (2008).
86. Szpiech, Z.A. & Hernandez, R.D. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol Biol Evol* **31**, 2824-7 (2014).
87. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol Psychiatry* **19**, 1085-94 (2014).
88. Cross-Disorder Group of the Psychiatric Genomics, C. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet* **381**, 1371-9 (2013).
89. Speliotes, E.K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* **42**, 937-48 (2010).
90. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43**, 333-8 (2011).
91. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
92. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
93. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat Genet* **48**, 624-33 (2016).
94. Hibar, D.P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224-9 (2015).
95. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
96. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
97. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet* **47**, 1457-1464 (2015).
98. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467-71 (2013).

Tables

Table 1. Genome-wide significant signals of recent positive selection. We report genomic locations, minimum p-value (capped at 10^{-16}) across 0.05cM windows, SNP corresponding to signal peak, and candidate gene for the 12 genome-wide significant signals of recent positive selection ($p < 0.05 / 63,103 = 7.9 \times 10^{-7}$). Novel loci are denoted in bold font.

| Chromosome | Region (Mb) | Min. p-value | SNP | Candidate gene(s) |
|------------|----------------------|--|--------------------|-------------------------------|
| 2 | 134.44-139.01 | $<10^{-16}$ | rs10206673 | LCT ³⁷ |
| 2 | 191.73-192.07 | 1.81×10^{-7} | rs7556924 | STAT4 |
| 4 | 38.44-38.97 | $<10^{-16}$ | rs7660745 | TLR gene family ³⁹ |
| 4 | 79.11-79.51 | 5.90×10^{-7} | rs2867461 | ANXA3 |
| 6 | 25.18-33.82 | $<10^{-16}$ | rs2104362 | HLA ³⁸ |
| 11 | 1.08-1.23 | 4.21×10^{-9} | rs11019228 | GRM5 ⁴⁰ |
| 11 | 88.21-90.55 | 1.20×10^{-10} | rs72636988 | MUC gene family |
| 14 | 106.35-107.12 | 9.49×10^{-9} | rs10142951 | IGHG ⁴⁰ |
| 16 | 70.89-71.80 | 7.73×10^{-8} | rs141399030 | PKD1L3 |
| 16 | 89.12-90.14 | 3.78×10^{-7} | rs62052682 | MC1R ⁴⁰ |
| 17 | 42.64-45.18 | 2.87×10^{-7} | rs75229873 | MYL4 |
| 22 | 48.98-49.08 | 4.94×10^{-7} | rs78014641 | FAM19A5 |

Figures

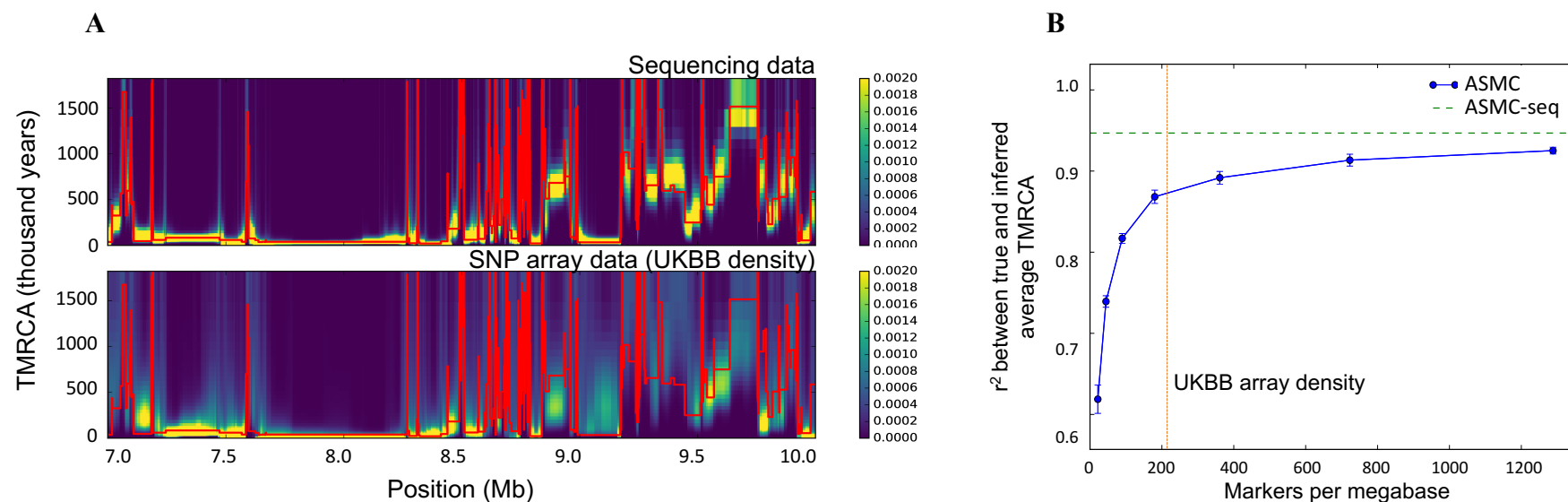


Figure 1. ASMC accuracy in coalescent simulations. (A) Sample posterior decoding of TMRCA along a 3 Mb segment for a pair of simulated individuals with ASMC run on WGS data (top) and on SNP array data (bottom). Red lines represent the true TMRCA, while the heat map represents the inferred posterior distribution. Posterior density tends to concentrate more tightly around the true TMRCA when WGS data are analyzed, due to the higher density of polymorphic variants. Posterior estimates using SNP array data are more dispersed for distant TMRCA, but remain highly concentrated for recent TMRCA. (B) Accuracy (r^2 between true and inferred average TMRCA) as a function of marker density. TMRCA are inferred using the posterior mean obtained by ASMC at each site. ASMC-seq represents the accuracy obtained using ASMC on WGS data. The red vertical line indicates marker density in the UK Biobank data set. Errors bars represent standard errors. Numerical results are reported in **Supplementary Table 11**.

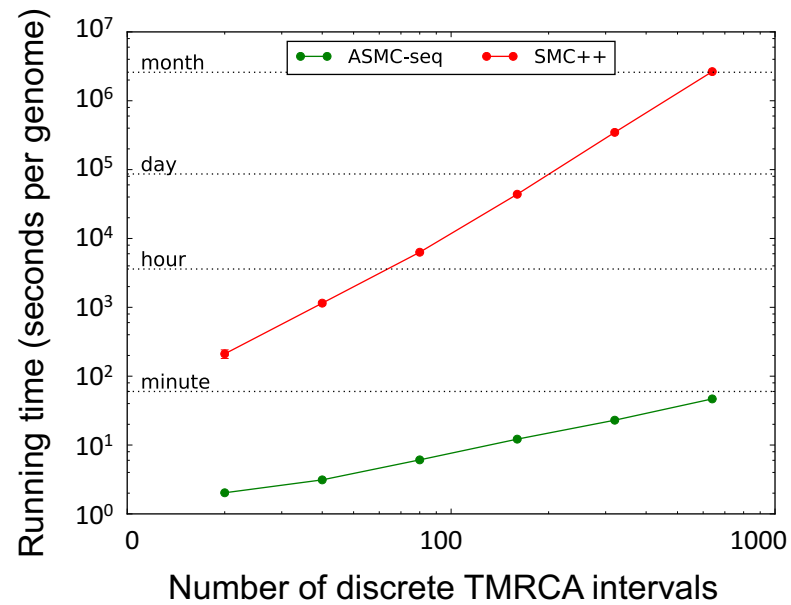


Figure 2. Running time of ASMC. We report the running time required to analyze a pair of simulated haploid genomes (extrapolated from running times in 5Mb regions) as a function of the number of discrete TMRCA intervals. Both SMC++ and ASMC-seq were run on WGS data. Numerical results are reported in **Supplementary Table 12**.

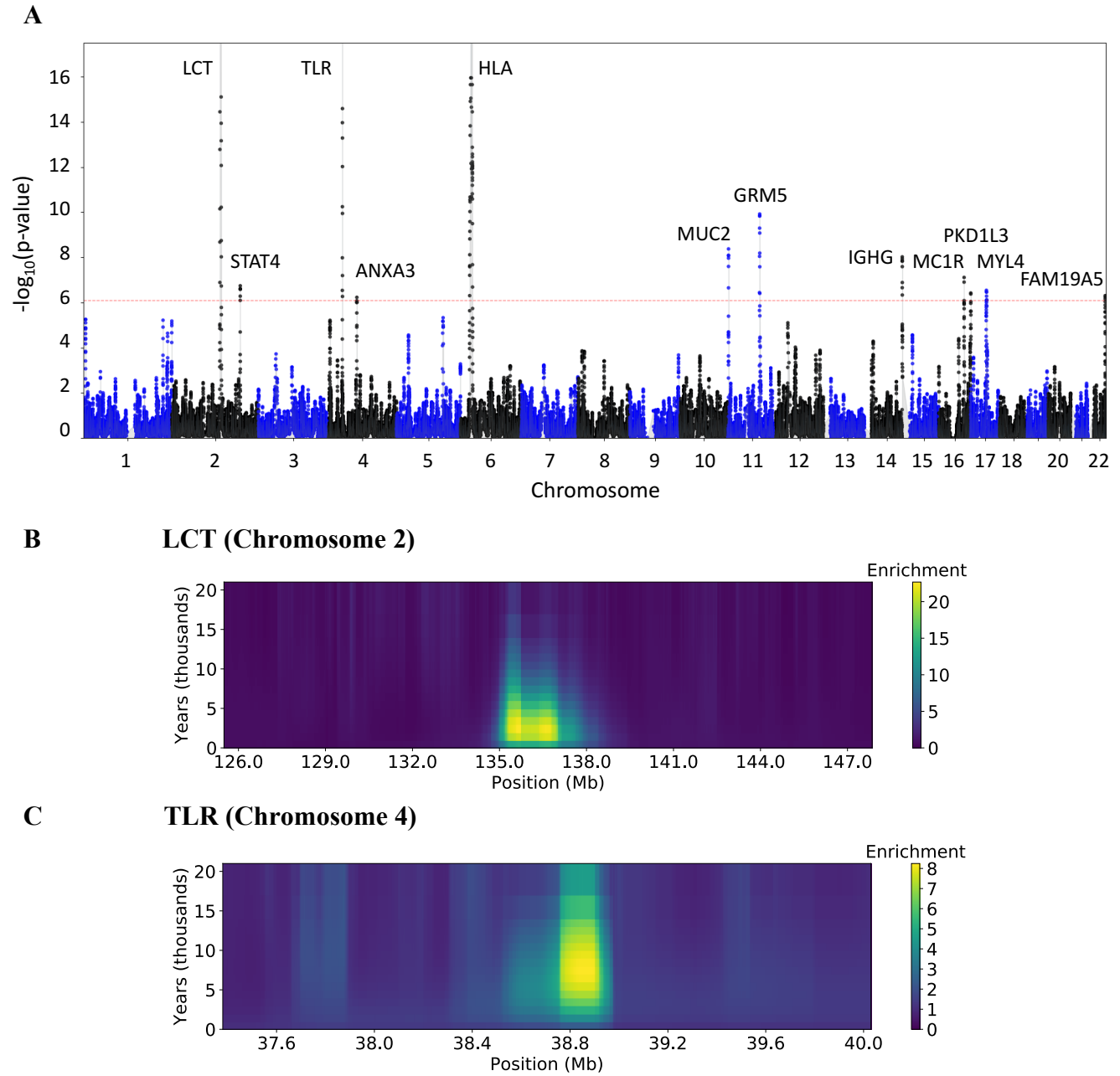


Figure 3. Genome-wide scan for recent positive selection in the UK Biobank data set. (A) Manhattan plot with candidate gene labels for 12 loci detected at genome-wide significance ($p < 0.05 / 63,103 = 7.9 \times 10^{-7}$; dashed red line). Numerical results for top loci are reported in **Table 1**; additional suggestive loci are reported in **Supplementary Table 6**. (B) Enrichment for recent coalescence events at the LCT locus (Chromosome 2). (C) Enrichment for recent coalescence events at the TLR locus (Chromosome 4). y-axis labels assume a 30-year generation time. Analogous plots for other top loci are provided in **Supplementary Figure 6**.

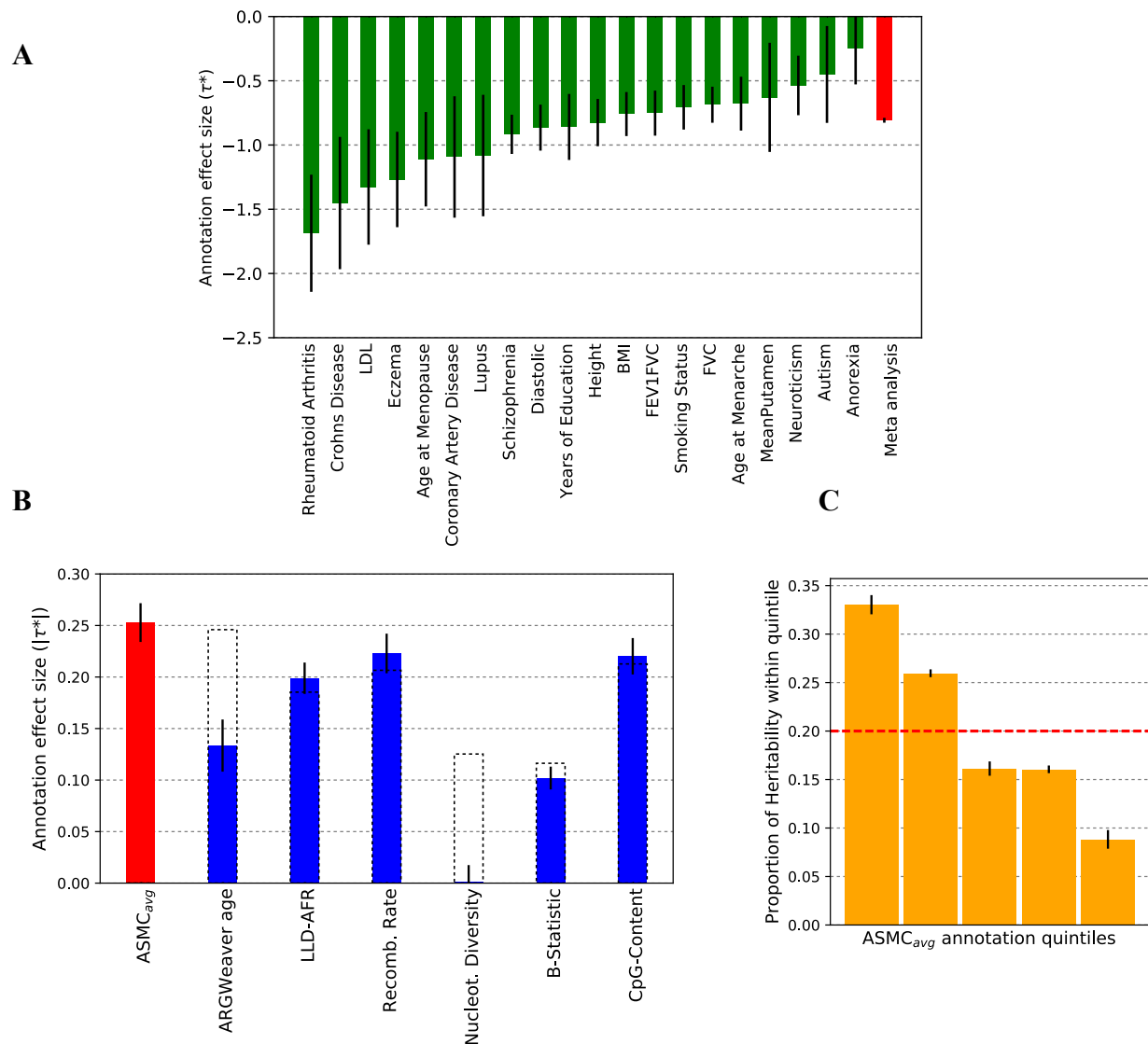
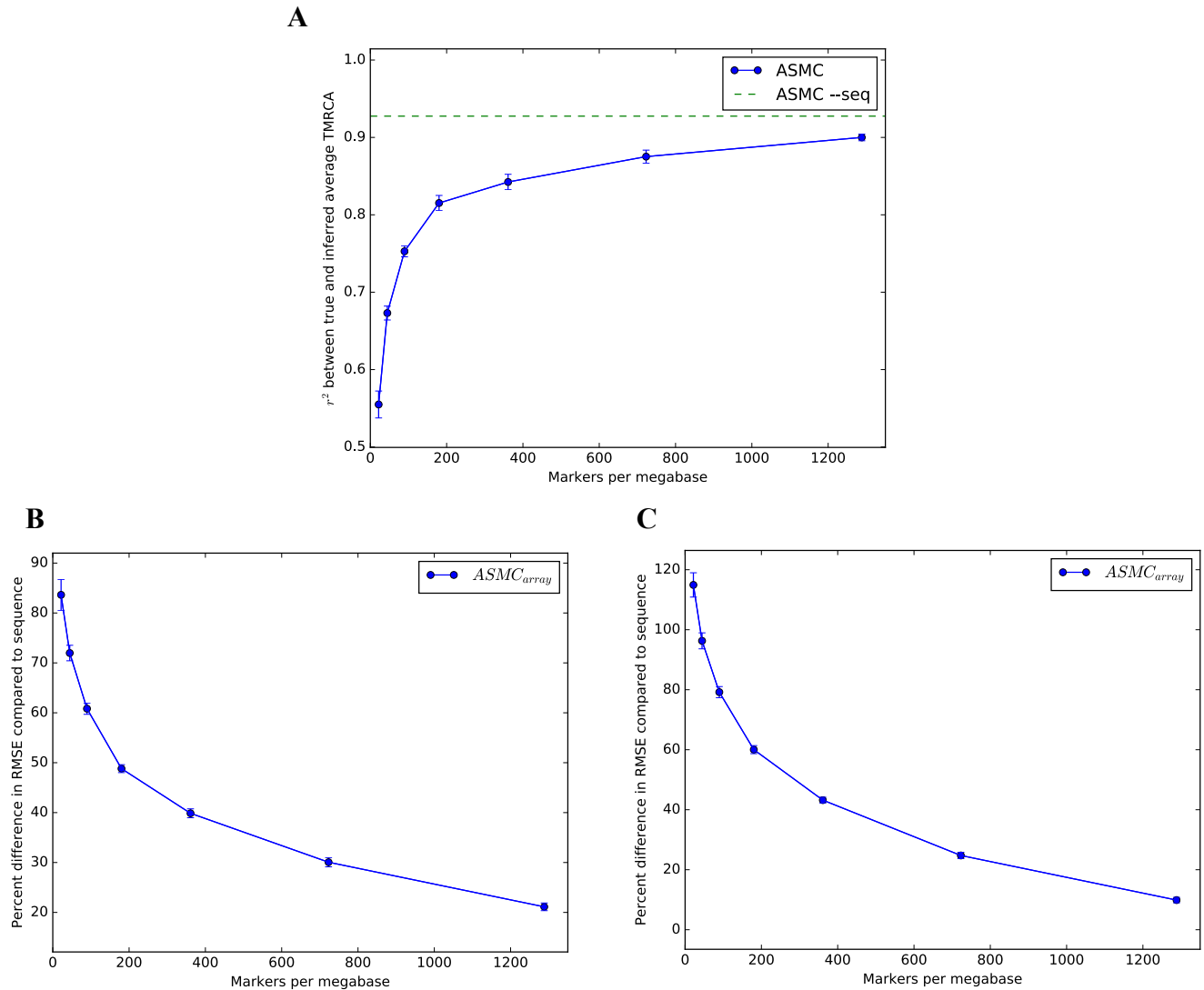
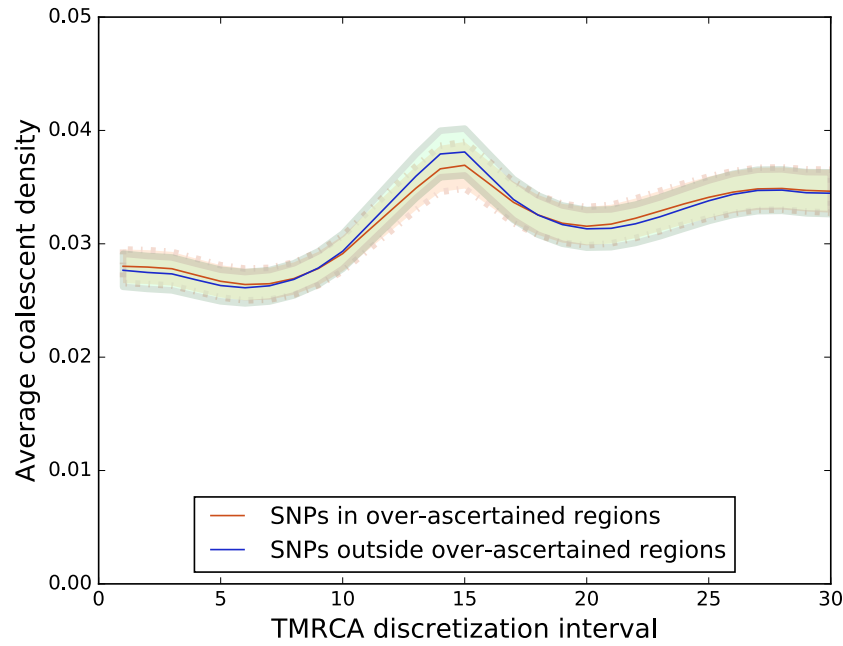


Figure 4. S-LDSC analysis of ASMC_{avg} background selection annotation and disease heritability. (A) τ^* value of the ASMC_{avg} annotation for 20 independent diseases and complex traits. (B) Absolute values of τ^* values (meta-analyzed across 20 independent diseases and complex traits) in joint analysis conditioned on baselineLD annotations. Dashed bars reflect values for six baselineLD annotations linked to background selection before the introduction of the ASMC_{avg} annotation. (C) Proportion of heritability explained by SNPs within different quintiles of ASMC_{avg} annotation (in joint analysis conditioned on baselineLD annotations). Error bars represent standard errors. Numerical results are reported in **Supplementary Table 13**.

Supplementary Figures and Tables

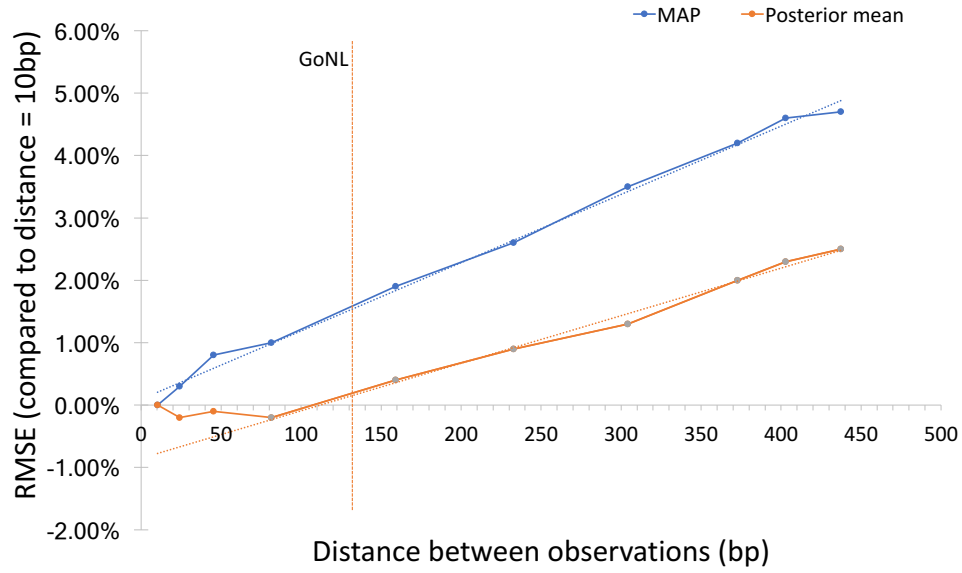


Supplementary Figure 1 – ASMC performance in simulations as a function of marker density. (A) r^2 , as a function of marker density, between true average TMRCA within the simulated region and average TMRCA inferred using the maximum-a-posteriori (MAP) of the posterior distribution computed by ASMC. ASMC-seq represents the accuracy obtained using ASMC on WGS data. Error bars represent standard errors. (B) We measure RMSE between true TMRCA at each site, and the TMRCA inferred by ASMC using the posterior mean on either SNP array or WGS data. We report the percent difference in per-site RMSE between analysis of SNP array data and WGS data (C) We measure RMSE between true TMRCA at each site, and the TMRCA inferred by ASMC using the MAP on either SNP array or WGS data. We report the percent difference in per-site RMSE between analysis of SNP array data and WGS data.

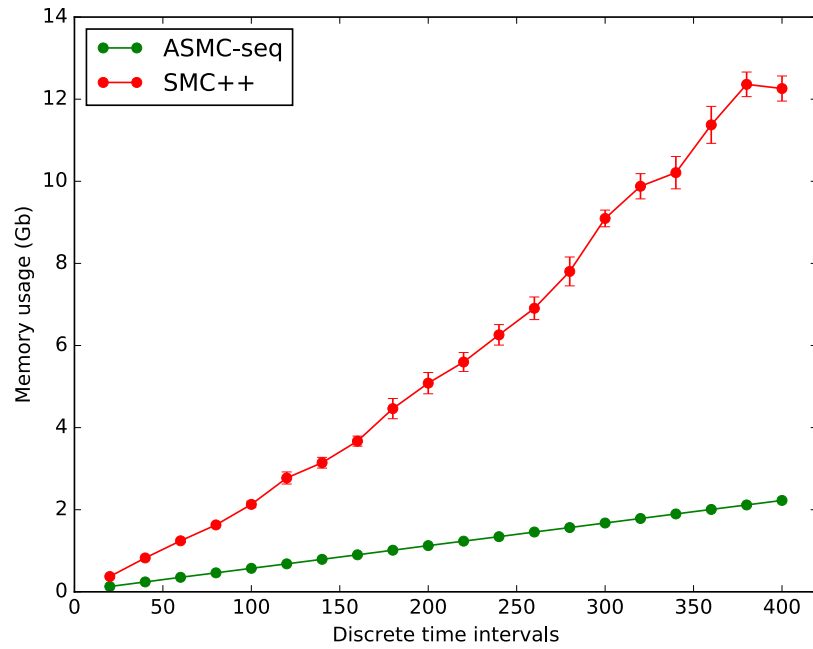


Supplementary Figure 2 - Robustness to deviations from frequency-based ascertainment.

Approximately 25% of the variants found on the UK Biobank Axiom Array were selected based on their functional relevance, particularly in coding regions, while the remaining ~75% were ascertained based on frequency. To mimic this ascertainment scheme in our simulations, we randomly sampled ~25% of the markers from 10Kb-long genes placed every 200Kb, while the remaining variants were sampled to match the UK Biobank frequency spectrum as in standard simulations. Simulations were performed using the standard setup and 30 discretization intervals for TMRCA inference. Shaded regions indicate 95% confidence intervals. We observed minimal deviation between coalescence densities inferred within and outside the simulated gene regions.

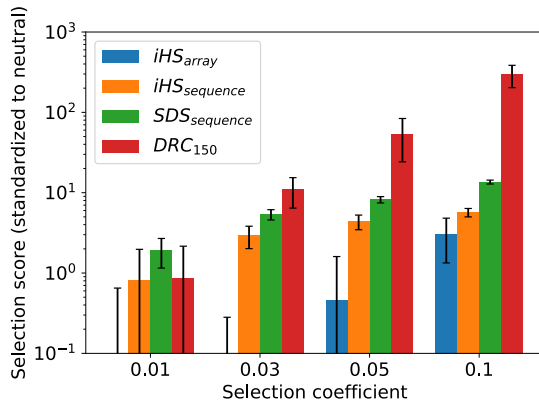


Supplementary Figure 3 – Effects of ASMC-seq transition approximation. When computing forward-backward probabilities in WGS data, ASMC-seq makes the approximation $(TE_0)^n \cong T^n E_0^n$, where n is the number of sites between two consecutive observations in the sequencing data, E_0 is a diagonal matrix reflecting emission probabilities for n monomorphic sites, and T is the transition matrix between two sites at distance n , which is close to diagonal with off-diagonal entries growing with n . Exact calculations are obtained for $n=1$ (or when ASMC is run on SNP array data), while an approximation is made for $n>1$. To measure the extent to which this approximation affects inference accuracy, we measured the per-site RMSE between true TMRCA and TMRCA inferred using either maximum-a-posteriori (MAP) or posterior mean. We simulated 100 European samples in a 10 Mb region at the beginning of Chromosome 2 (with recombination rate 2.18 cM per Mb), and randomly inserted monomorphic sites along the genome to measure accuracy at different values of n , running ASMC-seq with 30 time intervals. We report RMSE for different values of n , as a percentage of the RMSE measured for $n=10$. The red vertical bar represents the genotyping density observed for the GoNL data set ($n=136$). For MAP inference, the error linearly increased at a rate of $\sim 0.01\%$ per base pair, remaining below 2% for a genotyping density similar to the GoNL data set. For posterior mean inference, a negligible difference in accuracy was observed for $n<100$, followed by a linear increase at an approximate rate of $\sim 0.008\%$ per base pair, and increased error below 0.5% at GoNL genotyping density.

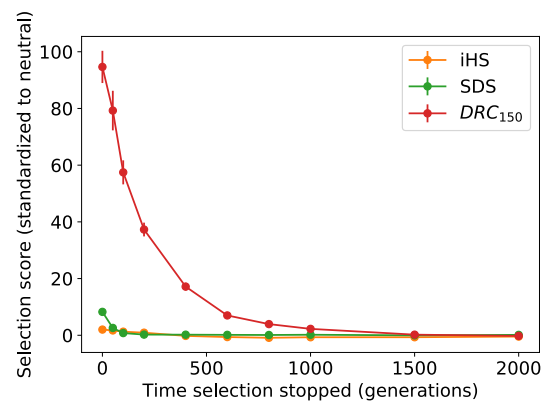


Supplementary Figure 4 – Memory use of ASMC-seq and SMC++. Memory usage for the analysis of coalescence times in a 5Mb region using WGS data from 100 haploid individuals. Bars represent standard errors.

A

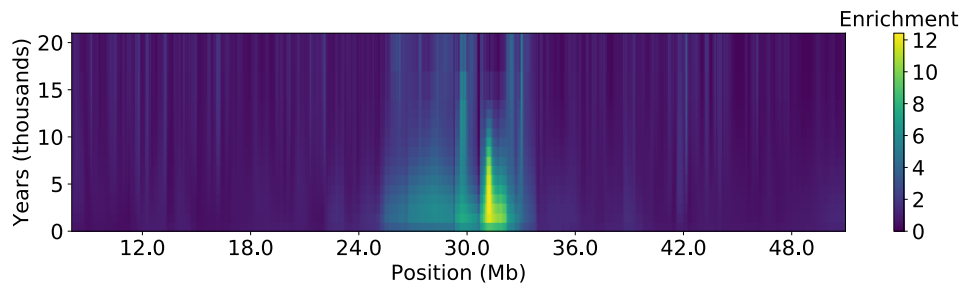


B

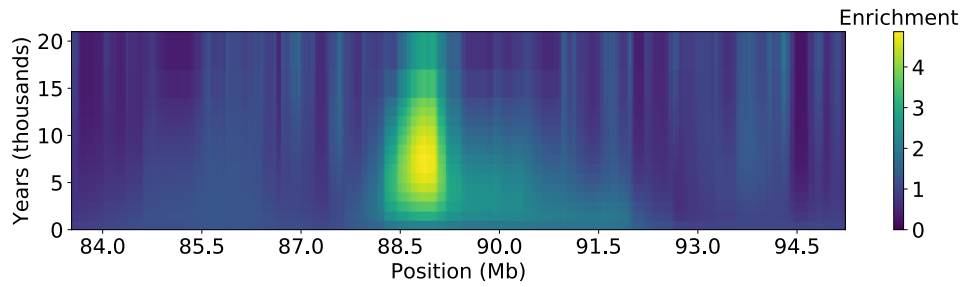


Supplementary Figure 5 – Selection simulations for DRC_{150} . (A) Simulation of different strengths of recent positive selection starting 200 generations in the past: iHS score³⁵ run on array data (iHS_{array}); iHS score³⁵ on causal variant from sequencing data ($iHS_{sequence}$); SDS score³⁶ on causal variant from sequencing data ($SDS_{sequence}$); DRC_{150} score on array data. Scores of each method are standardized with respect to corresponding scores obtained in neutral simulations. Bars indicate standard deviations. (B) Specificity to recent past for iHS and SDS run on sequencing data, and for DRC_{150} . Simulation of selection starting at time $-\infty$ stopping at the specified generation, followed by neutral drift. The DRC_{150} statistic is mostly sensitive to selection that has been active within the past ~ 700 generations (or $\sim 20,000$ years). Bars represent standard errors.

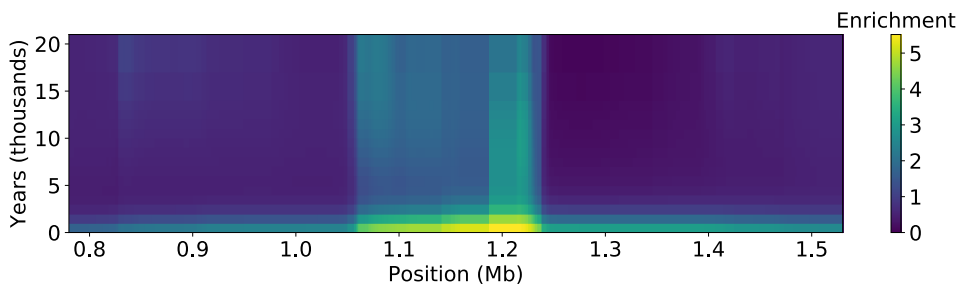
Chr 6



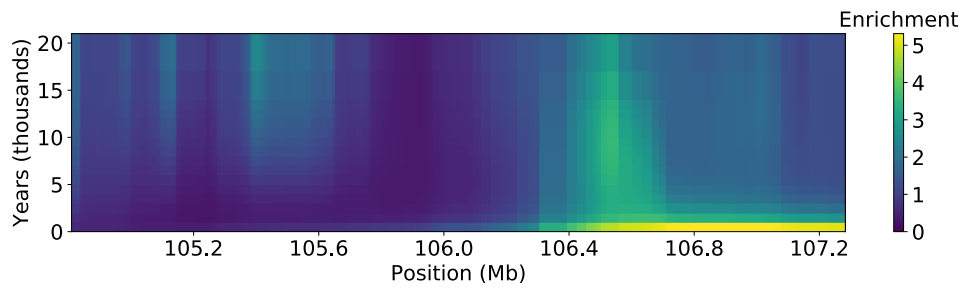
Chr 11



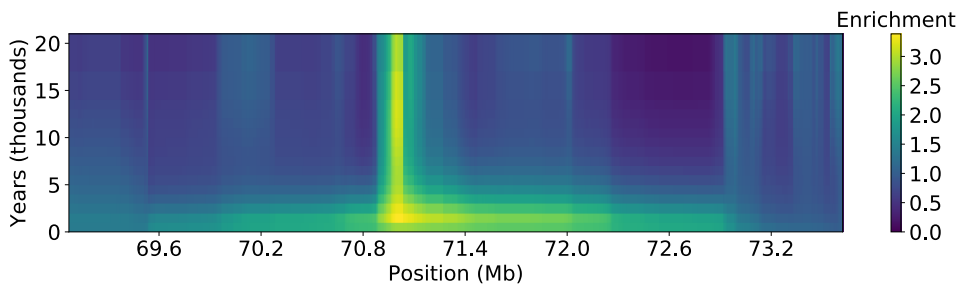
Chr 11



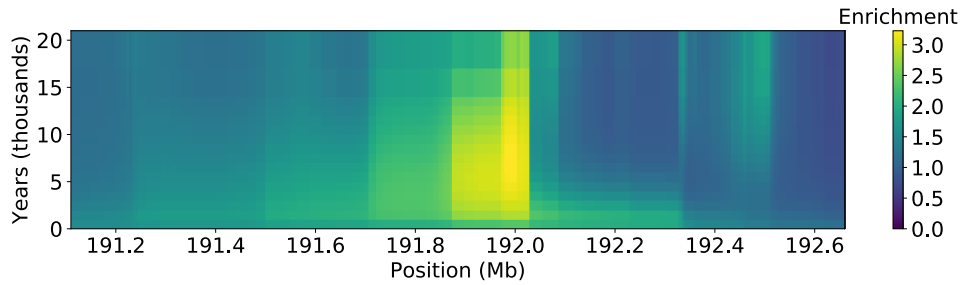
Chr 14



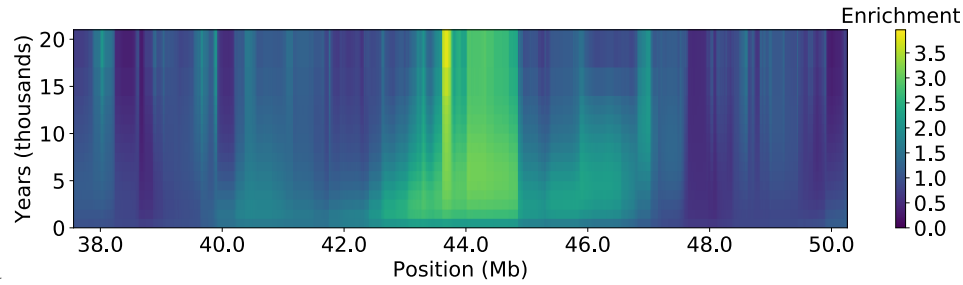
Chr 16



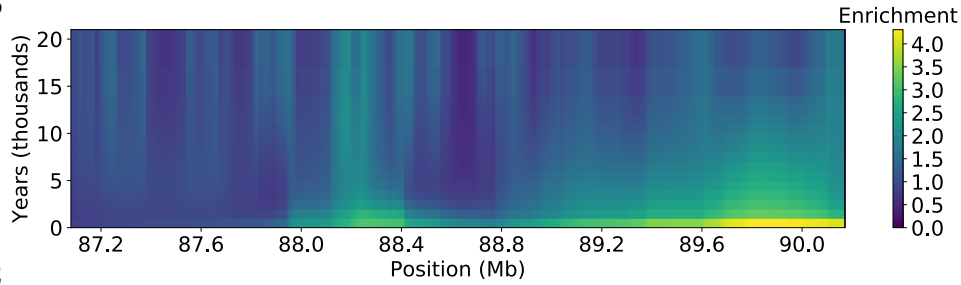
Chr 2



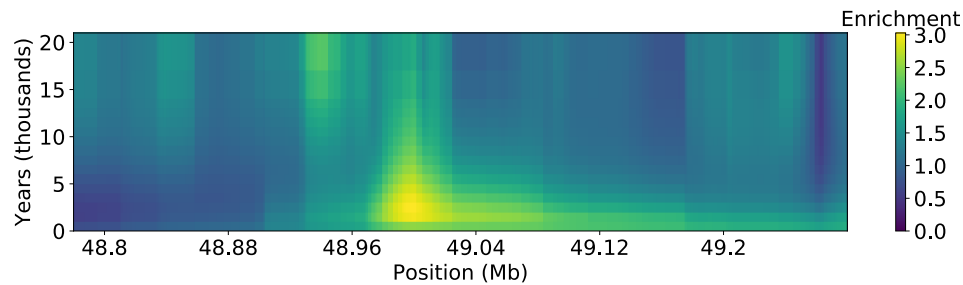
Chr 17



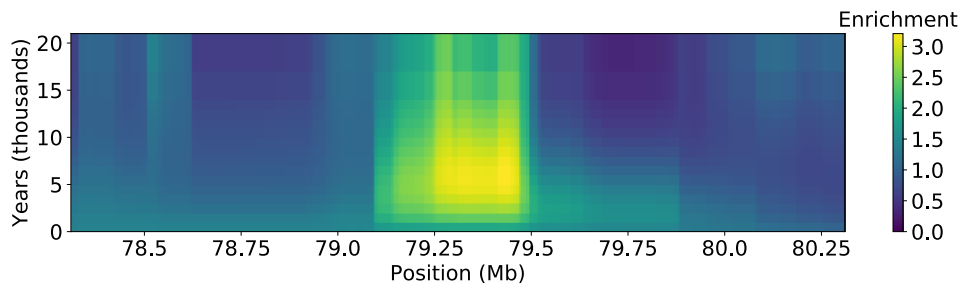
Chr 16



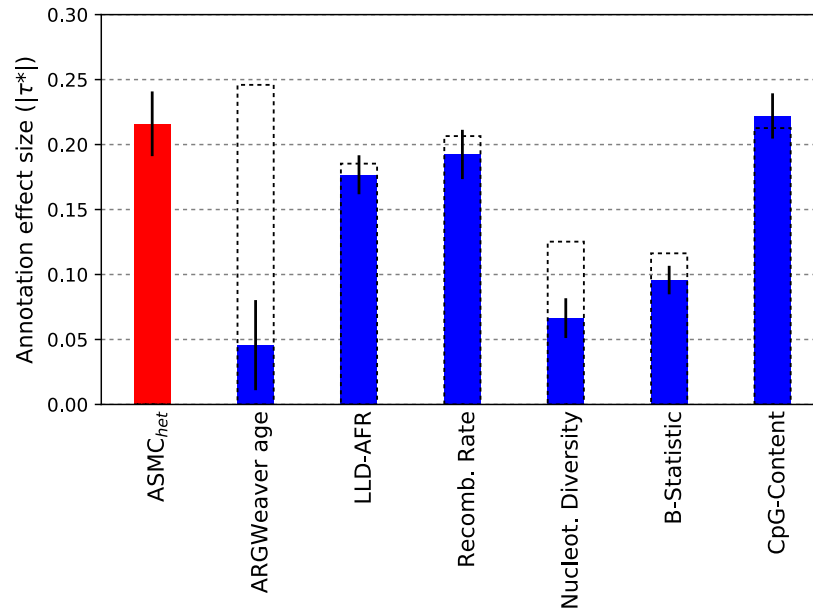
Chr 22



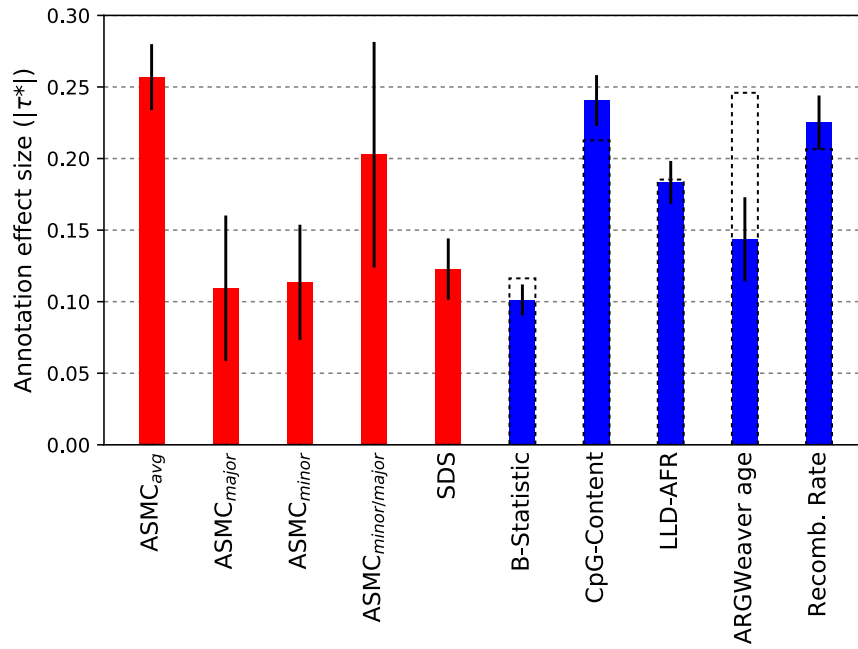
Chr 4



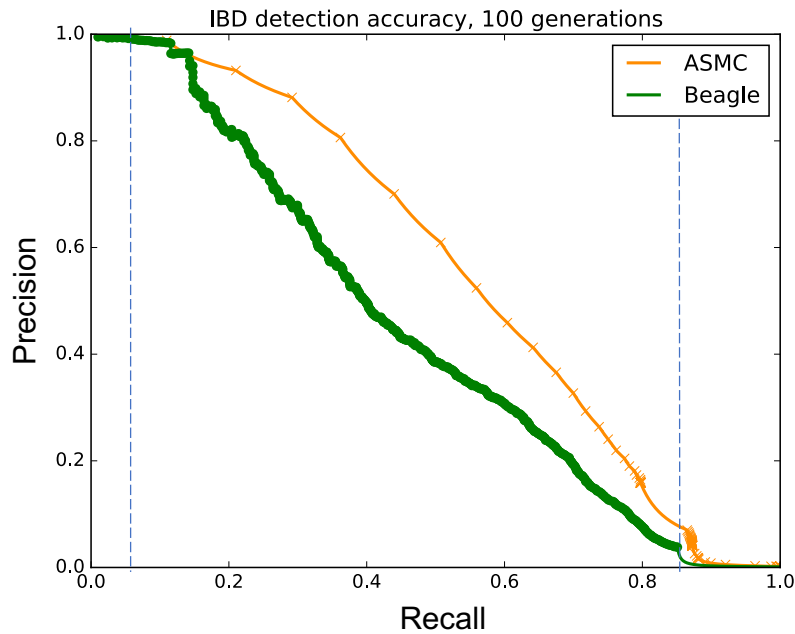
Supplementary Figure 6 – Enrichment of coalescence density in the past 20,000 years. At each site along the regions (horizontal axis) we plot the enrichment for the density of coalescence events in the past ~20,000 years, computed as $\frac{\text{posterior}_{\text{site,time}}}{\text{posterior}_{\text{genomewide,time}}}$. Time axes assumes a 30-year generation.



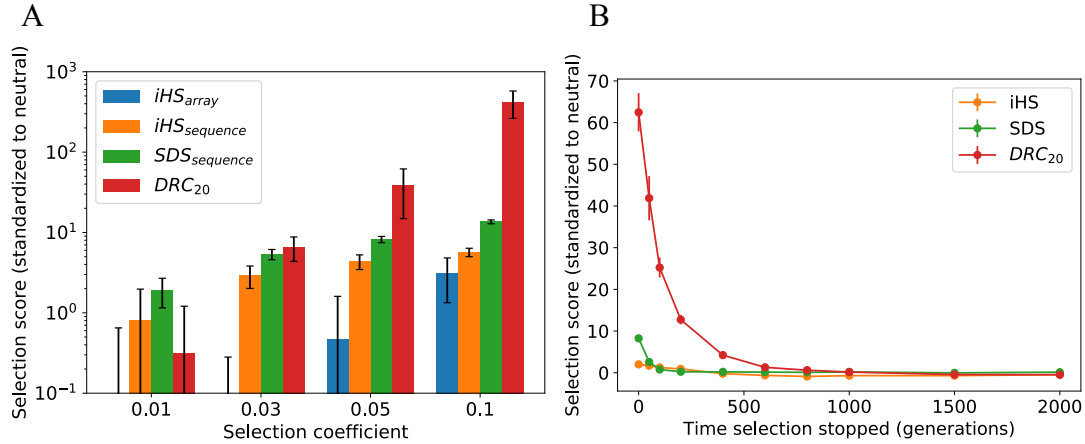
Supplementary Figure 7 - S-LDSC analysis of ASMC_{het} background selection annotation and disease heritability. We built an annotation, ASMC_{het}, reflecting the average coalescence time for heterozygous individuals (i.e. chromosomes carrying discordant alleles) at each site. As for the ASMC_{avg} annotation, ASMC_{het} is quantile normalized using 10 MAF bins. ASMC_{het} is expected to be proportional to the age of polymorphic alleles in the sample. Consistent with this expectation, in a joint S-LDSC analysis using the ASMC_{het} annotation and the baselineLD model, we observed that the meta-analyzed τ^* for the quantile normalized ARGWeaver allele age annotation was reduced from 0.250 (s.e. 0.012) to -0.046 (s.e. 0.018).



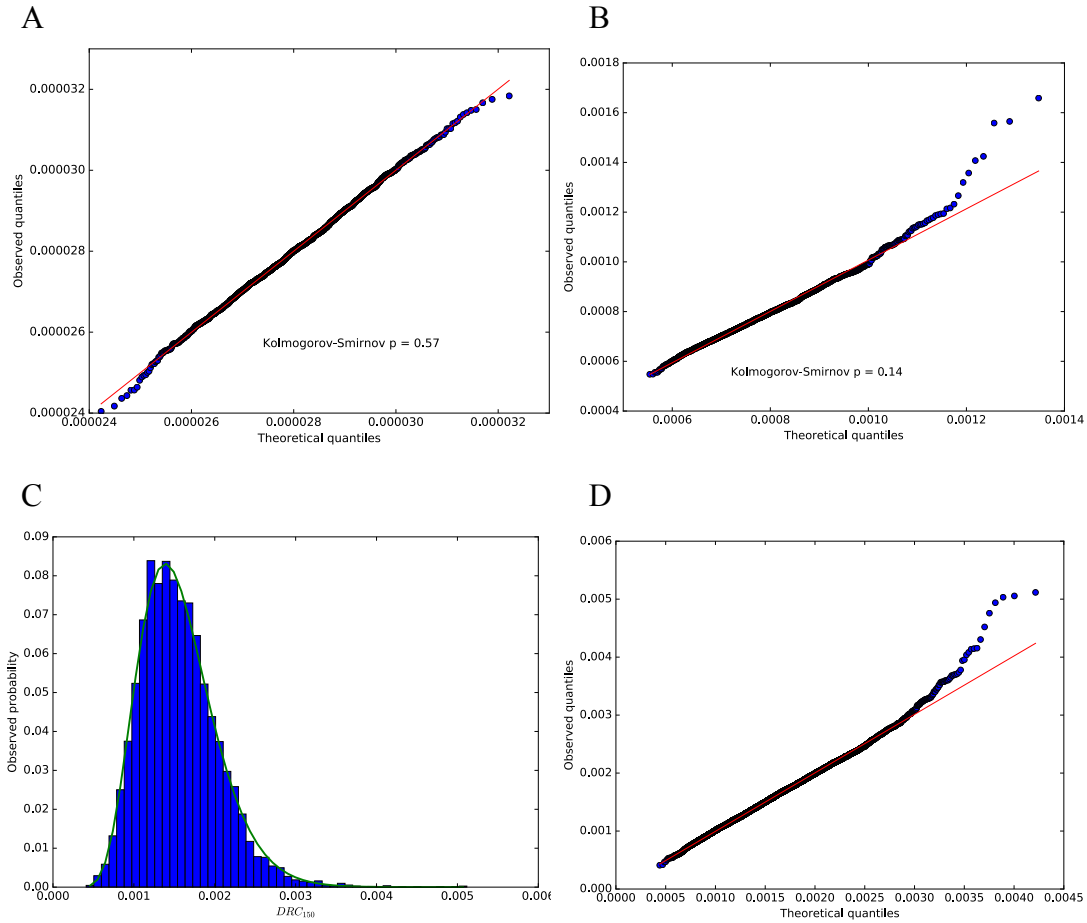
Supplementary Figure 8 - S-LDSC analysis of several annotations related to background selection. We built several annotations related to average coalescent time at each site, conditioning on the allele present on each analyzed chromosome from the GoNL data set. In addition to the ASMC_{avg} annotation (see **Online Methods**), we computed average coalescence time for carriers of a minor allele (ASMC_{minor}), carriers of a major allele (ASMC_{major}), and an annotation containing the value of $\log(T_{\text{minor}}/T_{\text{major}})$ at each site, i.e. the logarithm of the ratio of average coalescence time for individuals carrying a minor allele and individuals carrying a major allele (ASMC_{minor/major}). All annotations were quantile normalized with respect to 10 MAF bins, as done for the ASMC_{avg} annotation. We performed a joint S-LDSC analysis including these annotations, the SDS annotation from ref. ³⁶, and all annotations from the baselineLD model, excluding the nucleotide diversity annotation, whose effects are subsumed by the ASMC_{avg} annotation (see **Figure 4**). ASMC_{het} was also excluded, as it was subsumed by ASMC_{avg}. We report $|\tau^*|$ values meta-analyzed across 20 independent traits. Dashed lines for the baselineLD annotations represent meta-analyzed $|\tau^*|$ values in a joint S-LDSC analysis that does not include annotations represented in red.



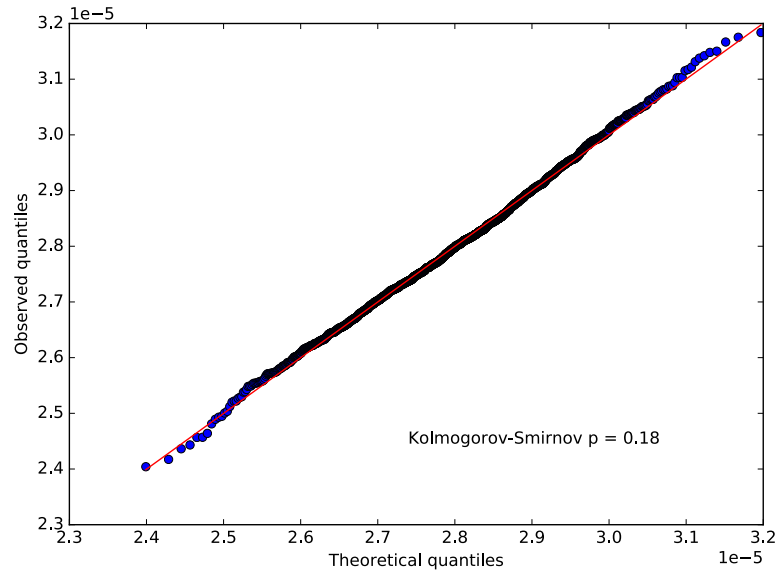
Supplementary Figure 9 – Illustration of auPRC measure for IBD detection accuracy. We measured accuracy of IBD detection for ASMC and Beagle using the area under the precision-recall curve (auPRC) for both programs. For both methods, recall can only be estimated within a limited precision range, due to the time-discretization used by ASMC, and the limited range of LOD-score thresholds allowed by Beagle. We thus compare the auPRC within the region where the precision and recall of both methods can be measured. In this example, ASMC's recall can be measured for values greater than 0.05, while Beagle's recall can be measured for values smaller than 0.85. We thus compare the auPRC for the two methods in the range [0.05, 0.85] (blue vertical lines). Interpolation between pairs of observed precision/recall values was obtained using the method of ref. ⁸⁰.



Supplementary Figure 10 – Selection simulations for DRC_{20} . (A) Simulation of different strengths of recent positive selection starting 200 generations in the past: iHS score³⁵ run on array data (iHS_{array}); iHS score³⁵ on causal variant from sequencing data ($iHS_{sequence}$); SDS score³⁶ on causal variant from sequencing data ($SDS_{sequence}$); DRC_{20} score on array data. Scores of each method are standardized with respect to corresponding scores obtained in neutral simulations. Bars indicate standard deviations. (B) Specificity to recent past for iHS and SDS run on sequencing data, and for DRC_{20} . Simulation of selection starting at time $-\infty$ stopping at the specified generation, followed by neutral drift. Error bars represent standard errors.



Supplementary Figure 11 – Empirical null model. (A) QQ plot for the DRC_{20} statistic in 2,000 independent neutral coalescent simulations using the European demographic model of Tennesen et al.¹⁹. (B) QQ plot for the DRC_{150} statistic in 2,000 independent neutral coalescent simulations using the European demographic model of Tennesen et al.¹⁹. (C) Empirical distribution and Gamma-fit for the DRC_{150} statistic in the putatively neutral portion of the genome in the UKBB data set (11,221 observations from 0.05 cM windows). (D) QQ plot for the DRC_{150} statistic in the putatively neutral portion of the genome in the UKBB data set (11,221 observations from 0.05 cM windows). All models are fit using a Gamma distribution with shape, location and scale parameters, using Python's Scipy library (see URLs).



Supplementary Figure 12 – Empirical null model for the DRC_{20} statistic. QQ plot for the DRC_{20} statistic in 2,000 independent neutral coalescent simulations using the European demographic model of Tennesen et al.¹⁹, fit using a Normal distribution.

Supplementary Table 1 – IBD detection. We report the difference in accuracy between ASMC- and Beagle-based IBD detection. IBD regions are defined using several time thresholds. We report the percent improvement for the area under the precision-recall curve (auPRC) of ASMC over Beagle. For both methods, precision can only be estimated within a limited recall range, due to the time-discretization used by ASMC, and the limited range of LOD-score thresholds allowed by Beagle. We thus compare the auPRC within the region where the precision and recall of both methods can be measured (“Average precision range” column, also see **Supplementary Figure 9**). Numbers in round brackets represent standard errors.

| IBD time threshold | Average recall range | auPRC within recall range | | Percent auPRC improvement: $100 \times (\text{auPRC}_{ASMC} / \text{auPRC}_{Beagle} - 1)$ |
|--------------------|----------------------|---------------------------|-------------|--|
| | | ASMC | Beagle | |
| 25 | [0.26, 0.98] | 0.44 (0.01) | 0.38 (0.01) | 17.79 (2.18) |
| 50 | [0.12, 0.95] | 0.54 (0.01) | 0.44 (0.01) | 21.71 (1.22) |
| 75 | [0.07, 0.90] | 0.54 (0.01) | 0.44 (0.01) | 22.66 (1.11) |
| 100 | [0.04, 0.86] | 0.52 (0.00) | 0.43 (0.00) | 21.47 (0.95) |
| 150 | [0.02, 0.77] | 0.49 (0.00) | 0.41 (0.00) | 19.04 (0.81) |
| 200 | [0.01, 0.70] | 0.46 (0.00) | 0.39 (0.00) | 17.58 (0.53) |
| 400 | [0.00, 0.49] | 0.37 (0.00) | 0.31 (0.00) | 17.78 (0.39) |
| 600 | [0.00, 0.36] | 0.29 (0.00) | 0.24 (0.00) | 19.99 (0.37) |

Supplementary Table 2 – Effects of demographic model misspecification. We simulated batches of 300 haploid samples from the first 30Mb of a human Chromosome 2 and a European demographic model, and ran ASMC using 160 discretization intervals (see **Online Methods**, Discretization Intervals). ASMC was ran assuming a constant effective population size of 10,000 diploid individuals, rather than the European model used to generate the data. We report percent difference in accuracy (RMSE and r^2), compared to using the appropriate demographic model. We observed an increase in RMSE error compared to ASMC analysis using the correct demographic model, and no significant difference in r^2 . Numbers in round brackets indicate standard errors from 10 independent simulations.

| | % difference when using wrong demographic model |
|---|--|
| RMSE of posterior mean estimate of TMRCA | +29.18 (1.21) |
| RMSE of maximum-a-posteriori estimate of TMRCA | +82.29 (0.99) |
| r^2 of posterior mean estimate of TMRCA | -0.43 (0.49) |
| r^2 of maximum-a-posteriori estimate of TMRCA | +1.28 (1.29) |

Supplementary Table 3 – IBD detection when ASMC demographic model is incorrect. We report the difference in accuracy between ASMC- and Beagle-based IBD detection. IBD regions are defined using several time thresholds. We report the percent improvement for the area under the precision-recall curve (auPRC) of ASMC over Beagle. For both methods, precision can only be estimated within a limited recall range, due to the time-discretization used by ASMC, and the limited range of LOD-score thresholds allowed by Beagle. We thus compare the auPRC within the region where the precision and recall of both methods can be measured (“Average precision range” column, also see **Supplementary Figure 9**). Numbers in parenthesis represent standard errors. Data were simulated under a European demographic model, but ASMC was run assuming a constant effective population size of 10,000 diploid individuals. This had negligible effects on accuracy, although the TMRCA bias introduced by this model misspecification slightly shifted the average precision range where ASMC’s AuPRC could be measured.

| IBD time threshold | Average recall range | auPRC within recall range | | Percent AuPRC improvement: $100 \times (\text{auPRC}_{ASMC} / \text{auPRC}_{Beagle} - 1)$ |
|--------------------|----------------------|---------------------------|-------------|--|
| | | ASMC | Beagle | |
| 25 | [0.32, 0.98] | 0.38 (0.01) | 0.32 (0.01) | 19.83 (2.92) |
| 50 | [0.15, 0.94] | 0.51 (0.01) | 0.41 (0.01) | 24.80 (1.40) |
| 75 | [0.08, 0.90] | 0.53 (0.01) | 0.42 (0.01) | 26.72 (1.55) |
| 100 | [0.05, 0.85] | 0.53 (0.01) | 0.42 (0.01) | 26.94 (1.09) |
| 150 | [0.03, 0.77] | 0.51 (0.00) | 0.40 (0.00) | 25.49 (0.80) |
| 200 | [0.02, 0.70] | 0.48 (0.00) | 0.39 (0.00) | 23.79 (0.44) |
| 400 | [0.01, 0.49] | 0.37 (0.00) | 0.31 (0.00) | 18.88 (0.49) |
| 600 | [0.00, 0.36] | 0.28 (0.00) | 0.24 (0.00) | 17.68 (0.58) |

Supplementary Table 4 – Effects of noise in the recombination rate map. To mimic inaccuracies in the genetic map we simulated data using a human recombination map, and ran ASMC using a map with added noise. The recombination rate between each pair of contiguous markers in the map was altered by randomly adding or subtracting a specified percentage of its true value (% noise). We report accuracy using RMSE and r^2 . RMSE is measured between true and inferred TMRCA at each site, and “RMSE %” refers to the percent difference in RMSE between TMRCA inferred in SNP array data (UKBB density) using the indicated genetic map and TMRCA inferred in WGS data using the correct genetic map. Error attained using the true map is reported at the top for comparison. r^2 indicates squared correlation between true and inferred average TMRCA in the simulated region. Each simulation involved a single batch of 300 haploid samples from the first 30Mb of a human Chromosome 2 and a European demographic model. ASMC was run using 160 discretization intervals (see **Online Methods**, Discretization Intervals).

| Map type | MAP RMSE % | Post. mean RMSE % | MAP r^2 | Post. Mean r^2 |
|-----------------|-------------------|--------------------------|-----------------------------|------------------------------------|
| True map | +49.46 | +45.15 | 0.85 | 0.90 |
| 10% noise | +51.74 | +48.95 | 0.85 | 0.88 |
| 20% noise | +50.62 | +48.08 | 0.86 | 0.89 |
| 30% noise | +51.16 | +46.94 | 0.84 | 0.90 |
| 40% noise | +51.74 | +45.55 | 0.85 | 0.89 |
| 50% noise | +53.89 | +50.61 | 0.85 | 0.89 |
| 60% noise | +53.95 | +50.55 | 0.85 | 0.88 |
| 70% noise | +50.09 | +49.36 | 0.84 | 0.89 |
| 80% noise | +53.70 | +52.18 | 0.84 | 0.87 |
| 90% noise | +55.72 | +57.21 | 0.83 | 0.87 |
| 100% noise | +53.25 | +52.76 | 0.83 | 0.88 |

Supplementary Table 5 – Effects of the number of time discretization intervals. We estimated coalescence times at each locus within 30Mb regions simulated using the standard setup using either the maximum-a-posteriori (MAP) or the posterior mean of the inferred coalescence distributions. In each simulation, we run ASMC using a different number of discretization intervals, which are chosen such that the coalescence distribution is expected to be uniform in all intervals (see **Online Methods**, Discretization intervals). For each value, we report the percent difference in RMSE accuracy between coalescence times inferred in SNP array data and WGS data, and the r^2 between true and inferred average TMRCA in the regions.

| Discretization intervals | MAP RMSE % | Post. mean RMSE % | MAP r^2 | Post. mean r^2 |
|---------------------------------|-------------------|--------------------------|-----------------------------|------------------------------------|
| 25 | +11.47 | +0.40 | 0.85 | 0.89 |
| 50 | +19.80 | -1.42 | 0.85 | 0.90 |
| 100 | +33.63 | +0.04 | 0.85 | 0.90 |
| 200 | +46.86 | +0.79 | 0.83 | 0.89 |
| 400 | +58.29 | +1.63 | 0.81 | 0.88 |

Supplementary Table 6 – Suggestive selection loci. We report loci under suggestive selection ($p < 10^{-4}$), as well as additional loci with elevated values of the DRC_{150} statistic in the UK Biobank data set ($10^{-4} < p < 10^{-3}$).

| Chromosome | Region (Mb) | Min. p-value | Top SNP | Candidate gene(s) |
|------------|---------------|-----------------------|-------------|--|
| 1 | 5.76-5.91 | 5.55×10^{-6} | rs12144662 | NPHP4, KCNAB2, CDH5, RPL22 |
| 1 | 223.71-223.86 | 6.07×10^{-6} | rs7525446 | CAPN2, CAPN8 |
| 1 | 235.15-235.17 | 6.47×10^{-4} | rs35894003 | RBM34, ARID4B, TOMM20 |
| 1 | 236.69-236.91 | 2.30×10^{-5} | rs2297860 | ACTN2, HEATR1, LGALS8 |
| 1 | 248.12-248.62 | 6.59×10^{-6} | rs28625479 | OR2L2, OR2L3, OR2L5, OR2M2, OR2M3, OR2M4, OR2M5, OR2M7, OR2T1, OR2T2, OR2T4, OR2T6, OR2T7, OR2AK2, OR2L13, OR2T12, OR2T33, OR14C36, LOC105373279 |
| 3 | 48.48-50.34 | 1.89×10^{-4} | rs146587089 | CYB561D2, CACNA2D2, AMT, BSN, APEH, DAG1, GPX1, MST1, NAT6, QARS, RBM5, RBM6, RHOA, TCTA, TMA7, UBA7, UCN2, USP4, WDR6, ARIH2, ATRIP, CAMKV, CDHR4, GMPPB, GNAI2, GNAT1, HYAL1, HYAL3, IFRD2, IP6K1, IP6K2, LAMB2, MON1A, MST1R, NICN1, P4HTM, TRAIP, TREX1, USP19, AMIGO3, CCDC36, CCDC71, CELSR3, COL7A1, DALRD3, IMPDH2, LSMEM2, PFKFB4, QRIH1, RNF123, SEMA3F, SHISA5, TMEM89, UQCRC1, ARIH2OS, C3orf62, C3orf84, FAM212A, KLHDC8B, NCKIPSD, NDUFAF3, PRKAR2A, SLC26A6, SLC25A20, RP11-3B7.1, CTD-2330K9.3 |
| 3 | 94.28-94.65 | 7.20×10^{-4} | rs114565822 | |
| 4 | 3.80-3.88 | 6.13×10^{-6} | rs28615087 | ADRA2C, LINC00955 |
| 4 | 24.96-24.98 | 3.58×10^{-4} | rs74870548 | LGI2, LOC102723675, CCDC149 |
| 5 | 33.68-34.36 | 2.76×10^{-5} | rs114118675 | SLC45A2 ⁵⁷ |
| 5 | 129.56-131.81 | 4.69×10^{-6} | rs739718 | SLC22A4 ⁵⁷ |
| 5 | 180.02-180.10 | 5.24×10^{-4} | rs6601131 | FLT4 |
| 6 | 139.41-139.55 | 6.37×10^{-4} | rs76157938 | HECA |
| 7 | 62.90-63.74 | 5.94×10^{-4} | rs118009401 | ZNF679, ZNF727 |
| 8 | 11.70-11.87 | 1.37×10^{-4} | rs4841682 | CTSB, DEFB134, DEFB135, DEFB136, RP11-481A20.11 |
| 8 | 17.95-18.22 | 1.48×10^{-4} | rs28556847 | NAT1 |
| 8 | 73.99-74.03 | 3.89×10^{-4} | rs6472748 | SBSPON |
| 9 | 136.99-137.02 | 2.11×10^{-4} | rs28650068 | WDR5 |
| 10 | 55.92-56.32 | 2.32×10^{-4} | rs12762168 | PCDH15 ⁴⁰ |
| 11 | 120.16-120.17 | 7.95×10^{-4} | rs2282537 | POU2F3 |
| 12 | 33.07-36.36 | 7.91×10^{-6} | rs4579984 | ALG10, SYT10 |

| | | | | |
|----|---------------|-----------------------|---------------|--|
| 12 | 53.17-53.17 | 9.71×10^{-4} | rs1873647 | |
| 12 | 54.35-54.58 | 9.49×10^{-5} | rs111779723 | HOXC4, HOXC5, HOXC6, HOXC8, HOXC9, SMUG1, HOXC10, HOXC11, HOXC12, RP11-834C11.12 |
| 12 | 55.44-55.97 | 1.17×10^{-4} | rs61411633 | OR6C1, OR6C2, OR6C3, OR6C4, OR6C6, OR9K2, OR10A7, OR2AP1, OR6C65, OR6C68, OR6C70, OR6C74, OR6C75, OR6C76 |
| 12 | 111.72-113.21 | 2.16×10^{-4} | rs10492023 | ATXN2, SH2B3 |
| 12 | 123.40-124.01 | 1.30×10^{-4} | rs61742326 | ABCB9, SBNO1, SETD8, OGFOD2, RILPL1, RILPL2, ARL6IP4, CDK2AP1, PITPNM2, SNRNP35, C12orf65, MPHOSPH9 |
| 14 | 20.47-20.52 | 3.63×10^{-4} | rs11158599 | OR4Q2, OR4K13, OR4K14 |
| 14 | 24.62-24.90 | 5.18×10^{-5} | rs4982912 | IPO4, IRF9, MDP1, NOP9, REC8, TGM1, ADCY4, CBLN3, CIDEB, DHRS1, GMPR2, LTB4R, NEDD8, PSME2, RIPK3, RNF31, TNF2, TSSK4, CHMP4A, LTB4R2, NFATC4, NYNRIN, TM9SF1, RABGGTA, NEDD8-MDP1, RP11-468E2.2, RP11-468E2.4, RP11-934B9.3 |
| 15 | 27.83-28.26 | 2.74×10^{-5} | rs145242923 | HERC2, OCA2 ⁵⁷ |
| 16 | 0.20-0.32 | 6.89×10^{-4} | Affx-80252323 | HBM, HBZ, HBA1, HBA2, HBQ1, ITFG3, LUC7L, RGS11, ARHGDIG |
| 16 | 55.84-55.88 | 4.01×10^{-5} | rs4784598 | CES1, CES5A |
| 16 | 88.15-88.30 | 6.25×10^{-5} | rs80193813 | |
| 17 | 7.33-7.61 | 2.69×10^{-4} | rs62062590 | CD68, FXR2, SAT2, SHBG, TP53, FGF11, MPDU1, SENP3, SOX15, ZBTB4, ATP1B2, CHRNA1, EIF4A1, POLR2A, WRAP53, SLC35G6, TMEM102, TNFSF12, TNFSF13, C17orf74, AC007421.1, TNFSF12-TNFSF13 |
| 17 | 45.35-46.31 | 1.27×10^{-4} | rs16957364 | SP2, SP6, CBX1, PNPO, COPZ2, ITGB3, KPNB1, SCRIN2, SKAP1, SNX11, TBX21, LRRC46, MRPL10, NFE2L1, NPEPPS, OSBPL7, PRR15L, TBKBP1, EFCAB13, CDK5RAP3 |
| 20 | 17.76-17.84 | 7.02×10^{-4} | rs2328224 | |

Supplementary Table 7 – Correlation between $ASMC_{avg}$ and other annotations from the baselineLD model.

| baselineLD annotation | Correlation with $ASMC_{avg}$ (r) |
|-----------------------------------|---|
| B-statistic ⁵⁹ | -0.28 |
| CpG content | 0.03 |
| Recombination rate | 0.07 |
| LLD-Africa ⁶¹ | 0.08 |
| ARGWeaver allele age ⁶ | 0.26 |
| Nucleotide diversity | 0.50 |

Supplementary Table 8 – Traits analyzed in S-LDSC analysis. We report phenotype name (and reference), number of samples in the study, Z-score for the trait's heritability, and URL (if summary statistics are publicly available).

| Phenotype | N | h ² Z-score | URL |
|--|---------|------------------------|---|
| Age at menarche (UKBB) | 74,944 | 18.27 | . |
| Age at menopause (UKBB) | 44,410 | 9.87 | . |
| Anorexia ⁸⁷ | 32,143 | 9.58 | http://www.med.unc.edu/pgc/downloads/ |
| Autism spectrum ⁸⁸ | 10,263 | 8.72 | http://www.med.unc.edu/pgc/files/resultfiles/pgcasdeuro.gz |
| Blood pressure, diastolic (UKBB) | 134,011 | 24.12 | . |
| BMI ⁸⁹ | 122,033 | 17.45 | http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files |
| Coronary artery disease ⁹⁰ | 77,210 | 8.47 | http://www.cardiogramplusc4d.org/ |
| Crohn's disease ⁹¹ | 20,883 | 10.34 | http://www.ibdgenetics.org/downloads.html |
| Eczema (UKBB) | 145,416 | 11.42 | . |
| Height (UKBB) | 145,368 | 29.29 | . |
| LDL ⁹² | 93,354 | 9.49 | http://www.broadinstitute.org/mpg/pubs/lipids2010/ |
| Lung FEV1/FVC ratio (UKBB) | 123,935 | 22.04 | . |
| Lung forced expiratory volume (UKBB) | 123,935 | 27.99 | . |
| Neuroticism ⁹³ | 170,911 | 9.54 | http://ssgac.org/documents/ |
| Putamen volume ⁹⁴ | 12,924 | 7.08 | http://enigma.ini.usc.edu/wp-content/uploads/E2_EVIS |
| Rheumatoid arthritis ⁹⁵ | 37,681 | 10.20 | http://plaza.umin.ac.jp/yokada/datasource/software.html |
| Schizophrenia ⁹⁶ | 70,100 | 21.82 | http://www.med.unc.edu/pgc/downloads/ |
| Smoking status (UKBB) | 145,227 | 19.70 | . |
| Systemic lupus erythematosus ⁹⁷ | 14,267 | 6.37 | https://www.immunobase.org/downloads/protected_data/GWAS_Data/ |
| Years of education ⁹⁸ | 126,559 | 11.97 | http://www.ssgac.org/ |

Supplementary Table 9 – Percent heritability explained by SNPs within annotation quintiles. We performed a joint analysis of $ASMC_{TMRC}$ and other annotations in the baselineLD model using S-LDSC, and estimated the fraction of heritability explained by SNPs in each quintile of an annotation. The highest ratio between largest and smallest mean quintile effects was observed for the $ASMC_{avg}$ and nucleotide diversity annotations. The effect (measured using τ^* , see **Figure 4B**) of the nucleotide diversity annotation, however, is subsumed by the $ASMC_{avg}$ annotation.

| Annotation | % of heritability for SNPs in each quintile | | | | | largest/smallest |
|-----------------------------------|---|-----------------|-----------------|-----------------|-----------------|------------------|
| | 1 st | 2 nd | 3 rd | 4 th | 5 th | |
| $ASMC_{avg}$ | 33.11 (0.53) | 26.05 (0.21) | 16.00 (0.39) | 16.05 (0.22) | 8.73 (0.51) | 3.79 |
| Nucleotide diversity | 31.53 (0.36) | 24.37 (0.15) | 20.40 (0.08) | 15.44 (0.16) | 8.31 (0.38) | 3.79 |
| LLD-Africa ⁶¹ | 29.12 (0.38) | 23.89 (0.13) | 20.45 (0.07) | 16.58 (0.16) | 9.94 (0.33) | 2.93 |
| ARGWeaver allele age ⁶ | 29.25 (0.68) | 25.43 (0.23) | 14.42 (0.31) | 20.16 (0.27) | 10.69 (0.63) | 2.74 |
| CpG content | 11.94 (0.21) | 16.76 (0.16) | 20.11 (0.13) | 22.09 (0.12) | 28.49 (0.40) | 2.39 |
| B-statistic ⁵⁹ | 13.17 (0.30) | 15.91 (0.14) | 19.57 (0.09) | 22.77 (0.12) | 28.35 (0.38) | 2.15 |
| Recombination rate | 19.08 (0.29) | 19.87 (0.20) | 20.71 (0.19) | 22.02 (0.15) | 18.27 (0.62) | 1.04 |

Supplementary Table 10 – Effects of the number of samples used in the emission model. We simulated data using standard parameters, and measured accuracy of ASMC-inferred coalescence times using RMSE and r^2 , for either WGS and SNP array data. We estimated coalescence times at each locus using either the maximum-a-posteriori (MAP) or the posterior mean of the inferred coalescence distributions. In each simulation, we ran ASMC using 100 discretization intervals and a different number of samples to compute the CSFS in the emission model. For RMSE, we report the percent difference in accuracy between coalescence times inferred in SNP array data and WGS data. Better RMSE or r^2 performance results for better use of allele frequency information via the CSFS emission model. We observed that the performance plateaus when using more than 100 samples in the CSFS. Approximate standard errors indicated in round brackets are computed using 5 independent simulations.

| Individuals in the emission model | MAP RMSE % | Post. mean RMSE % | MAP r^2 | Post. mean r^2 |
|--|-------------------|--------------------------|-----------------------------|------------------------------------|
| 50 | +62.92 (2.91) | +52.24 (2.14) | 0.750 (0.013) | 0.818 (0.013) |
| 100 | +51.54 (2.02) | +45.74 (0.86) | 0.829 (0.012) | 0.888 (0.005) |
| 150 | +54.75 (0.77) | +49.61 (1.33) | 0.824 (0.011) | 0.879 (0.007) |
| 200 | +51.59 (0.52) | +43.45 (0.61) | 0.814 (0.010) | 0.874 (0.005) |
| 250 | +55.21 (0.57) | +47.46 (1.25) | 0.834 (0.005) | 0.885 (0.005) |
| 300 | +55.92 (1.64) | +48.17 (0.92) | 0.830 (0.009) | 0.887 (0.008) |

Supplementary Table 11 - ASMC accuracy in coalescent simulations. Numerical values from Figure 1. Numbers in round brackets represent standard errors. The r^2 attained by ASMC-seq using WGS data is 0.946 (0.017). Average SNP density observed in the UK Biobank data set was 225. TMRCA was inferred using the ASMC posterior mean coalescence time at each site within the simulated region.

| Density (SNPS/Mb) | r^2 between true and inferred average TMRCA |
|-------------------|---|
| 21.7 | 0.619 (0.017) |
| 44.4 | 0.739 (0.007) |
| 89.6 | 0.817 (0.006) |
| 180.1 | 0.868 (0.008) |
| 361.0 | 0.892 (0.008) |
| 722.9 | 0.913 (0.007) |
| 1288.2 | 0.925 (0.004) |
| 1637.7 | 0.935 (0.005) |
| 1812.6 | 0.938 (0.005) |

Supplementary Table 12 – Computational cost of ASMC. Numerical values from **Figure 2** and **Supplementary Figure 4**. Running times are extrapolated from those obtained in 5Mb long regions of WGS data, assuming a 3,235 Mb genome. Memory usage reflects analysis of a 5Mb region using WGS data from 100 haploid individuals. Numbers in round brackets represent standard errors.

| TMRCA intervals | Running time (seconds per genome) | | Memory usage (Gb) | |
|-----------------|-----------------------------------|--------------------|-------------------|--------------|
| | ASMC-seq | SMC++ | ASMC-seq | SMC++ |
| 20 | 2.03 (0.05) | 211 (30) | 0.13 (0.0) | 0.37 (0.03) |
| 40 | 3.12 (0.06) | 1,150 (68) | 0.24 (0.0) | 0.83 (0.02) |
| 80 | 6.09 (0.21) | 6,310 (115) | 0.46 (0.01) | 1.63 (0.02) |
| 160 | 12.18 (0.41) | 43,879 (869) | 0.9 (0.01) | 3.67 (0.12) |
| 320 | 22.94 (0.97) | 347,385 (13,505) | 1.79 (0.03) | 9.88 (0.31) |
| 640 | 46.77 (0.77) | 2,649,817 (62,670) | 3.55 (0.06) | 21.67 (1.25) |

Supplementary Table 13 - S-LDSC analysis of $ASMC_{avg}$ background selection annotation and disease heritability. Numerical values from **Figure 4**. (A) τ^* value of the $ASMC_{avg}$ annotation for 20 independent diseases and complex traits. (B) Absolute values of τ^* values (meta-analyzed across 20 independent diseases and complex traits) in joint analysis conditioned on baselineLD annotations. Values in round brackets represent standard errors. Numerical values for **Figure 4c** can be found in **Supplementary Table 9**.

A

| Trait | τ^* |
|-------------------------|----------------|
| Rheumatoid Arthritis | -1.687 (0.233) |
| Crohns Disease | -1.452 (0.263) |
| LDL | -1.327 (0.229) |
| Eczema | -1.268 (0.189) |
| Age at Menopause | -1.111 (0.188) |
| Coronary Artery Disease | -1.093 (0.241) |
| Lupus | -1.082 (0.241) |
| Schizophrenia | -0.917 (0.078) |
| Diastolic | -0.865 (0.091) |
| Years of Education | -0.86 (0.131) |
| Height | -0.825 (0.094) |
| BMI | -0.759 (0.087) |
| FEV1FVC | -0.752 (0.089) |
| Smoking Status | -0.707 (0.088) |
| FVC | -0.686 (0.072) |
| Age at Menarche | -0.678 (0.107) |
| MeanPutamen | -0.629 (0.217) |
| Neuroticism | -0.536 (0.118) |
| Autism | -0.451 (0.192) |
| Anorexia | -0.25 (0.142) |
| Meta analysis | -0.807 (0.01) |

B

| Annotation | Annotation τ^* (meta-analysis) | Annotation τ^* (meta-analysis) not including $ASMC_{avg}$ in the model |
|-----------------------------------|-------------------------------------|---|
| $ASMC_{avg}$ | -0.253 (0.010) | N/A |
| ARGWeaver allele age ⁶ | -0.133 (0.013) | -0.246 (0.012) |
| LLD-Africa ⁶¹ | -0.199 (0.008) | -0.185 (0.008) |
| Recombination rate | -0.223 (0.010) | -0.207 (0.010) |
| Nucleotide diversity | -0.001 (0.008) | -0.125 (0.008) |
| B-statistic ⁵⁹ | 0.102 (0.006) | 0.116 (0.006) |
| CpG content | 0.220 (0.009) | 0.213 (0.009) |

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

Supplementary Note

Pier Francesco Palamara^{1,2,3}, Jonathan Terhorst⁴, Yun S. Song^{5,6}, Alkes L. Price^{2,3}

¹Department of Statistics, University of Oxford, Oxford, UK

²Department of Epidemiology, Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

³Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA

⁴Department of Statistics, University of Michigan, Ann Arbor, MI, USA

⁵Department of Statistics, Computer Science Division, University of California, Berkeley, Berkeley, CA, USA

⁶Chan Zuckerberg Biohub, San Francisco, CA, USA

Correspondence: palamara@stats.ox.ac.uk, aprice@hsph.harvard.edu

1 Background

1.1 The pairwise sequentially Markovian coalescent (PSMC)

The pairwise sequentially Markovian coalescent (PSMC (Li & Durbin, 2011)) is a widely adopted coalescent-based hidden Markov model (HMM) that describes the ancestral relationship of a pair of haploid individuals at all sites along their genome. We provide a high-level description of this approach, upon which our model and several recent extensions have been built.

The vector of observations in the HMM is obtained from the genotypes of a pair of haploid individuals that are randomly sampled from a population. For a sequence of length ℓ , observations $x_i, i \in \{1 \dots \ell\}$, have value 1 if the two individuals have discordant genotypes (they are heterozygous at the site) or 0 if they have identical genotypes (they are homozygous at the site). At each site along the genome, the hidden state $t_i \in \{1 \dots d\}$ of the Markov chain represents the time to most recent common ancestor (TMRCA) of the pair of haploid individuals at site i . Time is measured in generations (or in coalescent units) and is discretized into a predefined set of d possible time intervals.

The probability of observing a heterozygous site for the pair of individuals given their TMRCA is t is expressed as $P(x = 1|t, \mu) = 1 - e^{-2\mu t}$, where μ is the per generation, per base pair mutation rate, which is assumed to be constant along the genome and throughout time. Conversely, for a homozygous site, $P(x = 0|t, \mu) = e^{-2\mu t}$. The initial state probabilities for the HMM are obtained from the coalescent distribution induced by the effective size history of the population from which the two individuals were sampled. Transition probabilities between discrete TMRCA states along the genome are obtained using the sequentially Markovian coalescent (SMC) model, which provides a Markovian approximation to the coalescent process (McVean & Cardin, 2005) described as a sequence of recombination and coalescent events along the genome (Wiuf & Hein, 1999). Details of the transition model can be found in (Li & Durbin, 2011). The PSMC enables all usual applications of HMMs (Rabiner, 1989), including inferring the posterior probability of TMRCA at each site in the genome (posterior decoding), and learning the model’s hyperparameters, namely the population’s size history, mutation, and recombination rates.

1.2 Related work on coalescent HMMs

The CoalHMM model (Hobolth *et al.*, 2007) is one of the earliest coalescent HMMs, although its fundamental difference compared to the PSMC and derived approaches is that it operates at phylogenetic time scales, rather than population genetic time scales. The MSMC approach (Schiffels & Durbin, 2014), extended the PSMC to analysis of multiple haploid individuals. The hidden states of the MSMC model represent the time of the earliest coalescent event in the set of analyzed individuals, a modification that leads to increased insight into recent time scales. Another improvement of the MSMC over the PSMC is the use of the SMC’ model (Marjoram & Wall, 2006) in computing transition probabilities, which leads to increased accuracy compared to the SMC model (Hobolth & Jensen, 2014; Wilton *et al.*, 2015). When two individuals are analyzed, the MSMC approach reduces to the PSMC approach, though with the improved SMC’ transition model. The DiCal model (Sheehan *et al.*, 2013; Tataru *et al.*, 2014; Steinrücken *et al.*, 2015) is another coalescent HMM approach that enables simultaneous analysis of multiple samples, and explicit modeling of complex demographic scenarios. This approach relies on the conditional sampling distribution (CSD, (Paul *et al.*, 2011)), which approximates the full coalescent process by focusing on the conditional distribution of the n -th haploid individual given $(n - 1)$ individuals have been observed. When two individuals are analyzed, the DiCal approach reduces to the PSMC model. The computational burden of both the MSMC and the DiCal approach limits their use to no more than ~ 10 haploid individuals. The recently developed SMC++ method (Terhorst *et al.*, 2016), extends the PSMC approach by incorporating knowledge of the frequency of the analyzed genetic polymorphisms in the emission model of the HMM, effectively

utilizing genotype information from multiple samples while computing posterior coalescent probabilities for a single pair of haploid individuals. To achieve this, the SMC++ approach crucially relies on the notion of “conditioned sample frequency spectrum” (CSFS, see section 2.1 for an overview, and (Terhorst *et al.*, 2016) for details). As in the MSMC approach, the transition model of the SMC++ provides an improvement over the PSMC’s approximation of the full coalescent process. The SMC++ adopts the conditional Simonsen-Churchil model (CSC) proposed in (Hobolth & Jensen, 2014), which is superior to the SMC’ approach, as it considers the possibility of multiple recombination events occurring between two sites without affecting the TMRCA for a pair of analyzed individuals.

1.3 Computational cost and phasing requirements of other methods

Standard computation of posterior probabilities via the forward-backward algorithm, which we will simply refer to as “posterior decoding” in the remainder of this note, has cost $\mathcal{O}(d^2\ell)$ for d hidden states and an observation sequence of length ℓ (Rabiner, 1989). The standard forward-backward calculations adopted in the PSMC and MSMC methods therefore lead to $\mathcal{O}(d^2\ell)$ computational cost to estimate posterior coalescent probabilities for a set of d discretized TMRCA intervals and a sequence of length ℓ base pairs. PSMC reduces computational costs by pooling sites in blocks of 100 base pairs, while MSMC uses precomputation and caching to improve run time. The DiCal method (Steinrücken *et al.*, 2015) uses a “locus-skipping” approach (Paul & Song, 2012), which enables running the forward-backward algorithm in time $\mathcal{O}(d^2\ell_p)$, where ℓ_p is the set of loci that are polymorphic in the analyzed samples. This leads to substantial speed-ups, since usually $\ell \gg \ell_p$. A previous version of DiCal utilizes properties of the SMC model to reduce the run-time complexity of the forward-backward algorithm to $\mathcal{O}(d\ell)$ (Harris *et al.*, 2014). These approaches, however, are limited to use within the CSD model, which reduces to the SMC model when two haplotypes are analyzed. Compared to the SMC’ and the CSC model, the SMC provides a less accurate Markovian approximation of the coalescent (Hobolth & Jensen, 2014; Wilton *et al.*, 2015). The SMC++ approach, which utilizes the more accurate CSC model, implements a novel “locus-skipping” approach that enables computing the forward-backward dynamics in time $\mathcal{O}(d^3\ell_p)$.

The coalescent HMM approaches discussed thus far require the availability of accurate phasing information in order to perform TMRCA posterior decoding for haplotypes from distinct diploid individuals. Accurate computational phasing, however, cannot be achieved in modern sequencing data sets, particularly for rare variants. This often limits the application of coalescent HMM approaches to the maternal and paternal haplotypes within unphased diploid individuals, or results in noisy estimates of TMRCA distributions in the presence of pervasive phasing errors (Terhorst *et al.*, 2016). Although the

SMC++ approach provides an effective way of pooling information from the genotype of multiple unphased individuals from a sample, TMRCA decoding for pairs of haplotypes sampled across different diploid individuals still requires access to phasing information.

2 The ascertained sequentially Markovian coalescent

Here, we develop the Ascertained Sequentially Markovian Coalescent (ASMC). The ASMC is most closely related to the SMC++ (Terhorst *et al.*, 2016), and makes the following methodological innovations:

- The ability to perform posterior decoding using a non-random subset of genomic variants, such as the subset of common variants that are genotyped using SNP array technologies.
- A new formulation of the forward-backward algorithm that requires $\mathcal{O}(dl_p)$ computation under the conditional Simonsen-Churchil transition model (Hobolth & Jensen, 2014).

These two advances enable performing high-throughput coalescent-based analysis of relatedness in large SNP array data sets, which are now widely available and often comprise several tens or hundreds of thousand samples. Furthermore, owing to recent advances in computational phasing algorithms (Loh *et al.*, 2016a; Loh *et al.*, 2016b; O’Connell *et al.*, 2016), large cohorts such as the UK Biobank can now be computationally phased with very high accuracy, with switch error rates in the order of 0.3% (one every ~ 10 cM). This creates the possibility of analyzing coalescent times for potentially all pairs of haploid individuals in the sample, with negligible effects of phasing errors. The dramatic speedup achieved by ASMC also makes analysis of all pairs of available haploid genomes feasible in sequencing data sets, whenever high-quality phasing information is available.

2.1 ASMC emission

The emission model of a coalescent HMM approach for the inference of TMRCA in non-randomly ascertained genotype data, such as SNP array data, needs to tackle two key technical challenges, namely

1. The information content of the observed genotype data with respect to the coalescent time of the analyzed individuals is greatly reduced, as the vast majority of genotype variants are unobserved.
2. The set of observed variants are not randomly ascertained from the underlying sequencing variants. This ascertainment leads to significant bias in TMRCA inference if not accounted for.

To address these challenges, the ASMC adopts and extends the “conditioned sample frequency spectrum” (CSFS) model (Terhorst *et al.*, 2016). In addition to modeling allele sharing at each genomic site along the genome of the analyzed pair of individuals, as done in the PSMC approach, the CSFS enables taking into account the total number of individuals carrying each derived allele in a population sample. Modeling of allele frequencies using the CSFS allows to (1) increase the informativeness of the observations, enabling inference of TMRCA despite a substantial reduction in genotyped variants (2) remove biases due to frequency-based ascertainment, by explicitly modeling the probability of observing a variant in the data provided it is polymorphic at a given frequency in the analyzed sample.

The CSFS model can be briefly described as follows. Having obtained a set of $(n+2)$ haploid samples from a panmictic population with known demographic history, we denote 2 of these samples as “distinguished”, and the remaining n as “undistinguished”. Given that the pair of distinguished lineages coalesce at time τ at a site along the genome, the CSFS expresses the probability that exactly d out of the two distinguished individuals and u out of the n undistinguished individuals carry a mutated allele. We denote this probability as $CSFS(\tau)_{d,u}$, so that a $CSFS(\tau)$ is a $2 \times n$ table where entry $\{d, u\}$ corresponds to the probability that d derived alleles are observed in the two distinguished samples, and u derived alleles are found in the n undistinguished samples (the value of n is dropped to simplify the notation). Details on the derivation of the CSFS for a given demographic model can be found in (Terhorst *et al.*, 2016). We note that in this paper we are mainly concerned with the task of decoding TMRCA along the genome of a pair of haploid individuals, and we will adopt a demographic model inferred from previous analysis of whole-genome sequencing data.

Assume now that variants in the observed data set have been genotyped based on their frequency in a population sample, in other words, that the probability of observing a variant in the data can be expressed as $P(obs|d+u)$. The *ascertained* conditioned site frequency spectrum is then obtained as $ACFS(\tau)_{d,u} = CSFS(\tau)_{d,u} \times P(obs|d+u) \times norm$, where $norm$ is a normalizing constant such that $\sum_{d=0}^2 \sum_{u=0}^n ACFS(\tau)_{d,u} = 1$. In practice, we need to estimate $P(obs|d+u)$, and we do so by computing $\hat{P}(obs|d+u) = SFS_a(d+u)/SFS_s(d+u)$, where $SFS_a(x)$ and $SFS_s(x)$ represent counts for the number of sites polymorphic in x individuals for a sample of size $n+2$. Note that the normalization of $SFS_a(\cdot)$ and $SFS_s(\cdot)$, which should take into account terms related to e.g. the population mutation rate, is irrelevant, as these scaling constants vanish when the ACFS is renormalized. To estimate the ascertained $SFS_a(\cdot)$, we compute the sample frequency spectrum in the analyzed data. The sequence-level site frequency spectrum, $SFS_s(\cdot)$ is obtained using the population demographic model, which is known and provided in input. The unconditioned site frequency spectrum may be obtained from the CSFS as $SFS_s(x) = \int_{\tau=0}^{\infty} \eta(\tau) \sum_{d,u|d+u=x} CSFS(\tau)_{d,u}$ where $\eta(\tau)$ is the coalescent

probability for the known demographic model at time τ .

2.2 ASMC transition

The transition model of a coalescent HMM dealing with sparsely ascertained data needs to account for the increased distance between observed markers. Observed variants in common SNP array data sets, for instance, are separated by several kilobases on average. The SMC transition model (McVean & Cardin, 2005) originally adopted in the PSMC approach (Li & Durbin, 2011) becomes particularly inaccurate in this setting, as it postulates that at most one recombination event may occur between two contiguous sites. Furthermore, the SMC assumes that any recombination event leads to a change in the value of the TMRCA, whereas the full coalescent model admits the possibility that a recombination event between two loci is followed by a coalescent event to the same lineage such that the TMRCA remains unchanged. This modeling limitation is mitigated in the improved SMC’ model (Marjoram & Wall, 2006), which allows for multiple recombination and coalescent events between two loci, and is adopted (though allowing for at most one recombination event) in the MSMC approach (Schiffels & Durbin, 2014). The ASMC transition model adopts the “conditional Simonsen-Churchil” model (CSC) described in (Hobolth & Jensen, 2014), also implemented in the SMC++ approach (Terhorst *et al.*, 2016). The CSC further improves modeling of recurring recombination and coalescent events between a pair of sites that are separated by large genetic distances, such as markers in SNP array data.

2.3 A general linear time forward-backward algorithm

Although several computational improvements have been proposed in previous coalescent HMM methods (see Section 1.3), further speed-ups are required for the analysis of all pairs of haploid samples in large data sets under the CSC model. We thus devise a new algorithm that enables performing forward-backward posterior calculations using the CSC transition model in time $\mathcal{O}(d\ell_p)$, where ℓ_p is a set of observed loci for which we want to estimate TMRCA, and d is the number of discrete hidden TMRCA states. We start by introducing the Conditional Simonsen-Churchill model (Hobolth & Jensen, 2014), making use of the notation reported in Table 1.

2.3.1 The conditional Simonsen-Churchill model

Consider two loci at recombination distance $\rho/2$ in a population of constant size N , corresponding to a per-generation coalescent rate of η . In (Hobolth & Jensen, 2014), the Markov chain of Figure 1 was used to describe the distribution of ancestry at one locus conditional on the ancestry at the other locus. The transition matrix for this

Table 1: Table of notation for current section

| | | |
|-----------------------|--------------|---|
| ρ | \triangleq | Recombination rate |
| η_t or $\eta(t)$ | \triangleq | Coalescent rate at time t |
| N_t | \triangleq | Population size at time t |
| M_{SMC} | \triangleq | Transition rate for the SMC model |
| $M_{SMC'}$ | \triangleq | Transition rate for the SMC' model |
| M_{CSC} | \triangleq | Transition rate for the conditional Simonsen-Churchil model |
| $e_{i,j}^{tM}$ | \triangleq | Entry $\{i, j\}$ for the matrix exponential of tM |
| $\Omega(t)$ | \triangleq | Cumulative transition probability after compressing to 3 states |
| $C(t)$ | \triangleq | Cumulative transition probability before compressing to 3 states |
| $[\Omega(t)]_{i,j}$ | \triangleq | Entry $\{i, j\}$ for the cumulative transition probability |
| $q(t s)$ | \triangleq | Transition probability for locus 1 at time s and locus 2 at time t |
| $\pi(s, t)$ | \triangleq | Coalescent probability between time s and t |
| $\tilde{\pi}(s, t)$ | \triangleq | Probability of not having coalesced between time s and t |
| $\Pi(s, t)$ | \triangleq | Cumulative coalescent probability between time s and t |
| $Q(t s)$ | \triangleq | Cumulative transition probability for locus 1 at time s and locus 2 at time t |
| R_u | \triangleq | Time range $R_u = [T_u, T_{u+1})$ |

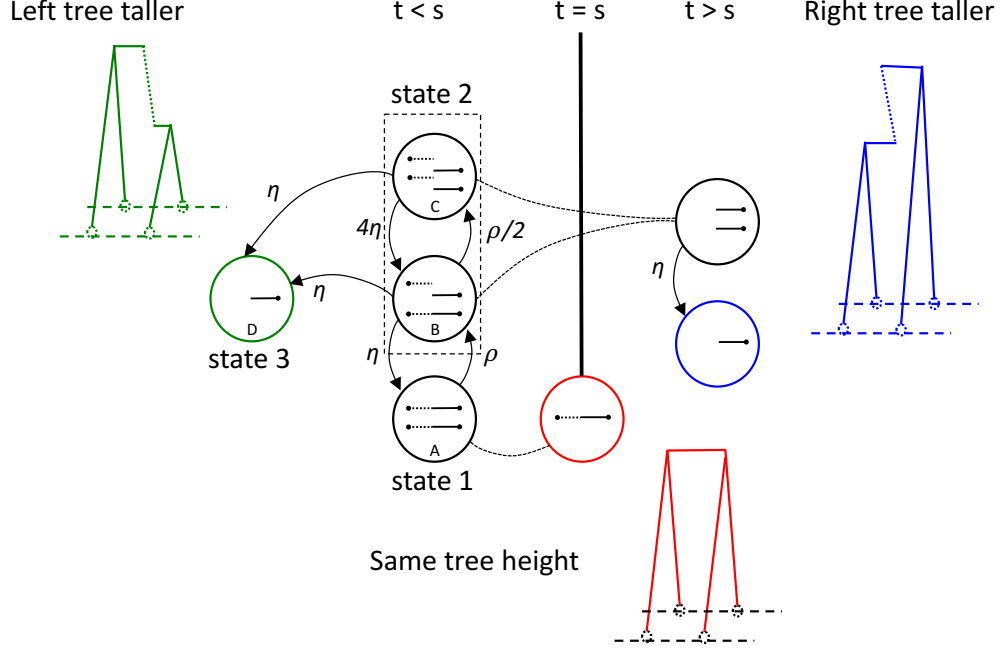


Figure 1: The conditional Simonsen-Churchil model (modified from Figure 1b of (Hobolth & Jensen, 2014)). Four relevant states from the full CSC model are labeled using letters within each circle.

model is

$$M_{CSC} = \begin{bmatrix} -\rho & \rho & 0 & 0 \\ \eta & -(2\eta + \rho/2) & \rho/2 & \eta \\ 0 & 4\eta & -5\eta & \eta \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad (1)$$

where each row and column of the matrix represents one of the four states for which $t < s$ (circles labeled with letters to the left of the vertical bar in Figure 1). Although the CSC model has four states, we will be mostly concerned with the probability that the Markov chain is in one of the three numbered states in Figure 1, that is, it will be irrelevant for our calculations whether at a given point in time the exact state of the chain is either state B or C within the dashed box. We thus define the matrix $\Omega(t)$, whose first row is $[\Omega(t)]_{1\bullet} = [C(t)_{AA}, C(t)_{AB} + C(t)_{AC}, C(t)_{AD}]$, where $C(t)_{i,j} = e_{i,j}^{tM}$ is the cumulative probability of transitioning from state i to state j after time t , for $i, j \in \{A, B, C, D\}$.

For completion, we note that although we are mostly concerned with the CSC model, the discussion below also applies to the SMC and SMC' models, which may be seen as

special cases of the CSC where states B and C have been collapsed, with updated rate matrices

$$M_{SMC} = \begin{bmatrix} -\rho & \rho & 0 \\ 0 & -\eta & \eta \\ 0 & 0 & 0 \end{bmatrix} \quad (2)$$

$$M_{SMC'} = \begin{bmatrix} -\rho & \rho & 0 \\ \eta & -2\eta & \eta \\ 0 & 0 & 0 \end{bmatrix}. \quad (3)$$

Note that $M_{SMC'}$ actually represents a process that is similar, but slightly different from the SMC', as discussed in (Wilton *et al.*, 2015). Thus, $[\Omega(t)]_{11}$ will hold the probability that no recombination occurred from time 0 to time t or, for the SMC' and CSC models, that at least one recombination occurred, but the lineages colasced back to state 1. $[\Omega(t)]_{12}$ resrepresents the probability that recombination occurred after time 0, but the lineages have not recoalesced back to state 1 or to a state such that the right tree has coalesced (state 3). $[\Omega(t)]_{13}$ represents the probability that the right tree is lower than the left tree, i.e. the two lineages coalesced at time $t < s$. Using these quantities, we can write the transition distribution for the height of the right tree, t , conditional on knowing the height of the left tree, s as

$$q(t|s) = \begin{cases} \eta [\Omega(t)]_{12} & \text{if } t < s, \\ [\Omega(s)]_{11} & \text{if } t = s, \\ \pi(s, t) [\Omega(s)]_{12} & \text{if } t > s, \end{cases} \quad (4)$$

where $\pi(s, t)$ is the coalescent probability between time s and t . This probability is computed as $\int_s^t \eta(t) t \, dt = \eta e^{-(t-s)\eta}$ for a constant population size with coalescent rate η . Equation 4 is normalized, since

$$\int_0^s \eta [\Omega(t)]_{12} \, dt = [\Omega(s)]_{13}, \quad (5)$$

$$\int_s^\infty \pi(s, t) [\Omega(s)]_{12} \, dt = [\Omega(s)]_{12}, \quad (6)$$

and $[\Omega(s)]_{11} + [\Omega(s)]_{12} + [\Omega(s)]_{13} = 1$.

2.3.1.1 Piecewise constant demographic model

If the population size is piecewise constant, for each time period k ranging in $R_k \in [T_k, T_{k+1})$, there is a different transition rate matrix M_k . If t is contained in the interval

R_k , then the state matrix at time t can be computed as

$$C(t) = \left[\prod_{i=1}^{k-1} e^{(T_{i+1}-T_i)M_i} \right] e^{(t-T_k)M_k}. \quad (7)$$

For a piece-wise constant model, the coalescent probability after time s can be similarly computed as

$$\begin{aligned} \pi(s, t) &= \eta_t \prod_{i=u|s \in R_u}^{v|t \in R_v} \exp \{ -\eta_i [m(t, T_{i+1}) - M(s, T_i)] \} \\ &= \eta_t \exp \left\{ \sum_{i=u|s \in R_u}^{v|t \in R_v} \eta_i [M(s, T_i) - m(t, T_{i+1})] \right\}, \end{aligned} \quad (8)$$

where $M(\dots)$ and $m(\dots)$ indicate maximum and minimum, respectively. The rate $\int_{T_i}^{T_{i+1}} \eta(t) t \, dt = \int_{T_i}^{T_{i+1}} \eta_i t \, dt = \eta_i (T_{i+1} - T_i)$ in the argument of the exponential should be substituted with the appropriate rate for inhomogeneous (e.g. exponential) models. We indicate the probability of not having coalesced at time t with

$$\begin{aligned} \tilde{\pi}(s, t) &= \eta_t^{-1} \pi(s, t) \\ &= \exp \left\{ \sum_{i=u|s \in R_u}^{v|t \in R_v} \eta_i [M(s, T_i) - m(t, T_{i+1})] \right\}. \end{aligned} \quad (9)$$

and the cumulative coalescent probability with

$$\Pi(s, t) = 1 - \tilde{\pi}(s, t). \quad (10)$$

Using the quantities above, the transition probability for tree heights is still given by

$$q(t|s) = \begin{cases} \eta_t [\Omega(t)]_{12} & \text{if } t < s, \\ [\Omega(s)]_{11} & \text{if } t = s, \\ \pi(s, t) [\Omega(s)]_{12} & \text{if } t > s. \end{cases} \quad (11)$$

2.3.1.2 Discretization

Using Equation [5](#), the cumulative transition probability is

$$Q(t|s) = \begin{cases} [\Omega(t)]_{13} & \text{if } t < s, \\ [\Omega(s)]_{11} + [\Omega(s)]_{13} & \text{if } t = s, \\ [\Omega(s)]_{11} + \Pi(s, t) [\Omega(s)]_{12} + [\Omega(s)]_{13} & \text{if } t > s. \end{cases} \quad (12)$$

The probability of transitioning between time s and the time range R_u is then obtained as $Q(T_{u+1}|s) - Q(T_u|s)$. The same approach can be used to further partition time in discrete states that do not necessarily correspond to population size changes. If we assume time has been discretized into d intervals, then we can obtain a transition matrix T such that entry $T_{i,j}$ corresponds to the probability of transitioning from time interval i to time interval j . Each entry of the transition matrix is then obtained as $T_{i,j} = Q(T_{j+1}|s_i) - Q(T_j|s_i)$, where we indicated the expected coalescent time within interval R_i as s_i .

2.3.2 Linear time computation of posterior coalescent times

We now describe a forward-backward algorithm to compute posterior coalescent probabilities in time $\mathcal{O}(d\ell_p)$, where d is the number of discrete coalescent time intervals, and ℓ_p is the number of sites for which we wish to obtain TMRCA estimates (e.g. the set of observed sites). We use the notation reported in Table 2

2.3.2.1 Forward probabilities

We want to compute α'_i , the forward probability at position p for state i , given a vector of forward probabilities for position $p - 1$ (which we denote as α_k , dropping the position index to simplify notation). Using standard considerations from hidden Markov models, this can be obtained as $\alpha'_i = \xi_i \sum_{k=1}^d \alpha_k T_{k,i} = \xi_i A_i$, where ξ_i represents the emission probability for the observation at position p (dropped to simplify the notation) given state i . Because this operation involves a vector-matrix multiplication, the cost of computing $A_i = \sum_{k=1}^d \alpha_k T_{k,i}$ is linear in d , and because d forward probabilities need to be computed, the overall cost will be quadratic in d . However, we note that the entries below the diagonal in T are all identical, since $Q(t|s)$ in Eq. 12 does not depend on s for $t < s$. Furthermore, the ratio of subsequent columns in the transition matrix can be computed as

$$T_{i,j+1}/T_{i,j} = \frac{\tilde{\pi}(T_j, T_{j+1}) [1 - \tilde{\pi}(T_{j+1}, T_{j+2})]}{[1 - \tilde{\pi}(T_j, T_{j+1})]} \quad (13)$$

(see Appendix). This ratio does not depend on i , so that it will be the same for all rows of the T matrix, as long as the entries are above the diagonal. Taken together, these observations imply that the sum $A_i = \sum_{k=1}^d \alpha_k T_{k,i}$ can be computed recursively in constant time. We assume the following quantities have been precomputed (in time linear in d), and are available for the computation of A_i :

- The diagonal entries of the transition matrix $D_i = T_{i,i}$ for $i \in [1, d]$.

Table 2: Table of notation for current section

| | | |
|------------------|--------------|--|
| p | \triangleq | Positions along the sequence |
| α'_k | \triangleq | Forward probability for state k at position p |
| α_k | \triangleq | Forward probability for state k at position $p - 1$ |
| ξ_k | \triangleq | Emission probability for state k at position p (for forward calculations) and at $p + 1$ (for backward calculations) |
| $T_{i,j}$ | \triangleq | HMM transition probability from discrete time i to discrete time j |
| d | \triangleq | Number of discrete states (time intervals) in the HMM |
| A_i | \triangleq | $\sum_{k=1}^d \alpha_k T_{k,i}$ |
| T_i | \triangleq | Start time for discrete interval i |
| T_{i+1} | \triangleq | End time for discrete interval i |
| D_i | \triangleq | Diagonal entry of the transition matrix |
| U_i | \triangleq | Entries above the diagonal for the transition matrix |
| B_i | \triangleq | Entries below the diagonal for the transition matrix |
| A_i^\downarrow | \triangleq | $\sum_{k=1}^{i-1} \alpha_k T_{k,i}$ in forward calculations |
| A_i^\uparrow | \triangleq | $\sum_{k=i+1}^d \alpha_k T_{k,i}$ in forward calculations |
| $\hat{\alpha}_i$ | \triangleq | $\sum_{k=i+1}^d \alpha_k$ |
| β_k | \triangleq | Backward probability for state k at position $p + 1$ |
| β'_k | \triangleq | Backward probability for state k at position p |
| v_i | \triangleq | $\xi_i \beta_i$ in backward calculations |
| B_i^\downarrow | \triangleq | $\sum_{k=1}^{i-1} v_k T_{i,k}$ in backward calculations |
| B_i^\uparrow | \triangleq | $\sum_{k=i+1}^d v_k T_{i,k}$ in backward calculations |

- The elements right above the diagonal $U_i = T_{i-1,i}$, for $i \in [2, d]$.
- The elements right below the diagonal $B_i = T_{i+1,i}$ for $i \in [1, d-1]$.
- The cumulative sum of the α vector of forward probabilities from the previous position, $\hat{\alpha}_i = \sum_{k=i+1}^d \alpha_k$.

We now rewrite the previous sum as

$$\begin{aligned}
A_i &= \sum_{k=1}^d \alpha_k T_{k,i} \\
&= \sum_{k=1}^{i-1} \alpha_k T_{k,i} + \alpha_i T_{i,i} + \sum_{k=i+1}^d \alpha_k T_{k,i} \\
&= A_i^\uparrow + \alpha_i D_i + A_i^\downarrow
\end{aligned} \tag{14}$$

Then, the quantities A_i^\uparrow and A_i^\downarrow can be computed in constant time as follows:

- $A_i^\uparrow = B_i \hat{\alpha}_i$
- $A_{i+1}^\downarrow = \alpha_i U_i + \frac{T_{i,j+1}}{T_{i,j}} A_i^\downarrow$, for $i \in [2, d]$, after having set $A_1^\downarrow = 0$.

Having computed the above quantities (in time linear in d), all entries $A_i = A_i^\uparrow + \alpha_i D_i + A_i^\downarrow$ can be computed in linear time. The final forward vector is obtained multiplying the emission probabilities to obtain $\alpha'_i = \xi_i A_i$.

2.3.2.2 Backward probabilities

The linear-time backward calculations can be obtained in a similar way. In this case, given ξ , the emission probability vector at sequence position $p+1$, and β , the backward probability vector for position $p+1$, we want to compute $\beta'_i = \sum_{k=1}^d T_{i,k} \xi_k \beta_k$, the backward probability at state i , position p . We again use observations (1) and (2) from the previous section to efficiently compute this sum. It is convenient to define the vector v such that $v_i = \xi_i \beta_i$. As in the previous case, we rewrite the above sum as

$$\begin{aligned}
\beta'_i &= \sum_{k=1}^d v_k T_{i,k} \\
&= \sum_{k=1}^{i-1} v_k T_{i,k} + v_i T_{i,i} + \sum_{k=i+1}^d v_k T_{i,k} \\
&= B_i^\downarrow + v_i D_i + B_i^\uparrow
\end{aligned} \tag{15}$$

We have previously noted that the ratio of subsequent columns above the diagonal is constant (see Appendix). We now note that the same holds for the ratio of columns. In particular, it can be shown (see Appendix), that

$$T_{i,j}/T_{i+1,j} = \frac{[\Omega(s_{i+1})]_{12}}{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, s_{i+1})} \quad \forall i > j. \quad (16)$$

Using this result, the quantities B_i^\downarrow and B_i^\uparrow can be efficiently computed as

- $B_i^\downarrow = \sum_{k=0}^{i-1} B_{k-1} v_{k-1}$, having set $B_1^\downarrow = 0$.
- $B_i^\uparrow = v_{i+1} U_i + \frac{T_{i,j}}{T_{i+1,j}} B_{i+1}^\uparrow$, having set $B_d^\uparrow = 0$.

From these quantities, we can then obtain $\beta'_i = B_i^\downarrow + v_i D_i + B_i^\uparrow$. Note that these calculations hold for the SMC, SMC' and CSC models, provided the corresponding transition matrices are used to compute entries of the Ω vector. Inhomogeneous (e.g. exponential) models can be handled by computing the corresponding coalescent quantities in the above calculations.

2.3.2.3 Approximate decoding for stretches of identical observations

When ascertained data is analyzed and no information on the sequence content between observed ℓ_p markers is available, the linear time algorithm described above yields exact posterior TMRCA probabilities. Using a locus-skipping approximation, it is also possible to use the same linear-time forward-backward algorithm for the analysis of sequencing data, where we wish to obtain TMRCA estimates for ℓ_p loci (e.g. polymorphic loci), while accounting for the fact that all sites between any other two contiguous observations share the same emission probabilities (e.g. they are all monomorphic in the analyzed sample, or homozygous if frequency information is not used in the emission model). To this end we note that the forward step of the forward-backward algorithm between two sites separated by a stretch of n identical observations requires computing the product $\alpha' = \alpha(T E_s)^n T E_p$, where T is the transition matrix between two sites in the region, E_s is a diagonal matrix with the emission probability for a given emission character (e.g. homozygous/monomorphic site), and E_p is a diagonal matrix with emission for the site at position p in the sequence. We observe that, for relatively small genetic distances between the two observed sites, and for realistic demographic models, the matrix T is close to diagonal. Thus, we can use the commutative property of diagonal matrices to approximate the product $(T E_s)^n$ as $T^n E_s^n$. Having done that, we can now rely on the previously described linear time algorithm to compute the product $\alpha(T E_s)^n T E_p \sim \alpha T^n E_s^n T E_p$. In the ASMC program, the matrices T^n and E_s^n are precomputed (in linear time) and stored so that these need not be computed for each

analyzed haploid pair. Note that the ASMC uses genetic distances from a human recombination map, rather than assuming a constant recombination rate along the genome, so that the matrix T^n will actually depend on genomic position, while the emission matrix E_s^n will only depend on the number of loci between a pair of sites.

3 Appendix

3.1 Ratio of columns in the transition matrix

$T_{i,j+1}/T_{i,j} = \frac{\tilde{\pi}(T_j, T_{j+1})[1 - \tilde{\pi}(T_{j+1}, T_{j+2})]}{[1 - \tilde{\pi}(T_j, T_{j+1})]} \quad \forall j > i$. *Proof:*

$$\begin{aligned}
T_{i,j} &= Q(T_{j+1}|s_i) - Q(T_j|s_i) \\
&= ([\Omega(s_i)]_{11} + \Pi(s_i, T_{j+1}) [\Omega(s_i)]_{12} + [\Omega(s_i)]_{13}) - ([\Omega(s_i)]_{11} + \Pi(s_i, T_j) [\Omega(s_i)]_{12} + [\Omega(s_i)]_{13}) \\
&= \Pi(s_i, T_{j+1}) [\Omega(s_i)]_{12} - \Pi(s_i, T_j) [\Omega(s_i)]_{12} \\
&= [\Omega(s_i)]_{12} (\Pi(s_i, T_{j+1}) - \Pi(s_i, T_j)) \\
&= [\Omega(s_i)]_{12} [(1 - \tilde{\pi}(s_i, T_{j+1})) - (1 - \tilde{\pi}(s_i, T_j))] \\
&= [\Omega(s_i)]_{12} [\tilde{\pi}(s_i, T_j) - \tilde{\pi}(s_i, T_{j+1})] \\
&= [\Omega(s_i)]_{12} [\tilde{\pi}(s_i, T_j) - \tilde{\pi}(s_i, T_j)\tilde{\pi}(T_j, T_{j+1})] \\
&= [\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_j) [1 - \tilde{\pi}(T_j, T_{j+1})],
\end{aligned} \tag{17}$$

which implies

$$\begin{aligned}
\frac{T_{i,j+1}}{T_{i,j}} &= \frac{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_{j+1}) [1 - \tilde{\pi}(T_{j+1}, T_{j+2})]}{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_j) [1 - \tilde{\pi}(T_j, T_{j+1})]} \\
&= \frac{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_j) \tilde{\pi}(T_j, T_{j+1}) [1 - \tilde{\pi}(T_{j+1}, T_{j+2})]}{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_j) [1 - \tilde{\pi}(T_j, T_{j+1})]} \\
&= \frac{\tilde{\pi}(T_j, T_{j+1}) [1 - \tilde{\pi}(T_{j+1}, T_{j+2})]}{[1 - \tilde{\pi}(T_j, T_{j+1})]}
\end{aligned} \tag{18}$$

□

3.2 Ratio of rows in the transition matrix

$T_{i+1,j}/T_{i,j} = \frac{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, s_{i+1})}{[\Omega(s_{i+1})]_{12}} \quad \forall i > j$. Again, using

$$T_{i,j} = [\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_j) [1 - \tilde{\pi}(T_j, T_{j+1})], \tag{19}$$

we have

$$\begin{aligned}
\frac{T_{i+1,j}}{T_{i,j}} &= \frac{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, s_{i+1}) \tilde{\pi}(s_{i+1}, T_j) [1 - \tilde{\pi}(T_j, T_{j+1})]}{[\Omega(s_{i+1})]_{12} \tilde{\pi}(s_{i+1}, T_j) [1 - \tilde{\pi}(T_j, T_{j+1})]} \\
&= \frac{[\Omega(s_i)]_{12} \tilde{\pi}(s_i, s_{i+1})}{[\Omega(s_{i+1})]_{12}}
\end{aligned} \tag{20}$$

□

3.3 Above diagonal elements

$$\begin{aligned} T_{i,i} &= Q(T_{i+1}|s_i) - Q(T_i|s_i) \\ &= [\Omega(s_i)]_{11} + \Pi(s_i, T_{i+1}) [\Omega(s_i)]_{12} + [\Omega(s_i)]_{13} - [\Omega(T_i)]_{13} \end{aligned} \quad (21)$$

and

$$\begin{aligned} T_{i,i+1} &= Q(T_{i+2}|s_i) - Q(T_{i+1}|s_i) \\ &= [\Omega(s_i)]_{12} \tilde{\pi}(s_i, T_{i+1}) [1 - \tilde{\pi}(T_{i+1}, T_{i+2})], \end{aligned} \quad (22)$$

References

- Harris, Kelley, Sheehan, Sara, Kamm, John A, & Song, Yun S. 2014. Decoding coalescent hidden Markov models in linear time. *Pages 100–114 of: Research in Computational Molecular Biology*. Springer.
- Hobolth, Asger, & Jensen, Jens Ledet. 2014. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical population biology*, **98**, 48–58.
- Hobolth, Asger, Christensen, Ole F, Mailund, Thomas, & Schierup, Mikkel H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet*, **3**(2), e7.
- Li, Heng, & Durbin, Richard. 2011. Inference of human population history from individual whole-genome sequences. *Nature*, **475**(7357), 493–496.
- Loh, Po-Ru, Palamara, Pier Francesco, & Price, Alkes L. 2016a. *Fast and accurate long-range phasing in a UK Biobank cohort*. Tech. rept. Nature Publishing Group.
- Loh, Po-Ru, Danecek, Petr, Palamara, Pier Francesco, Fuchsberger, Christian, Reshef, Yakir A, Finucane, Hilary K, Schoenherr, Sebastian, Forer, Lukas, McCarthy, Shane, Abecasis, Goncalo R, *et al.* . 2016b. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, **48**(11), 1443–1448.
- Marjoram, Paul, & Wall, Jeff D. 2006. Fast "coalescent" simulation. *BMC Genetics*, **7**(1), 16.
- McVean, Gilean AT, & Cardin, Niall J. 2005. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1459), 1387–1393.

- O’Connell, Jared, Sharp, Kevin, Shrine, Nick, Wain, Louise, Hall, Ian, Tobin, Martin, Zagury, Jean-Francois, Delaneau, Olivier, & Marchini, Jonathan. 2016. *Haplotype estimation for biobank-scale data sets*. Tech. rept. Nature Publishing Group.
- Paul, Joshua S, & Song, Yun S. 2012. Blockwise HMM computation for large-scale population genomic inference. *Bioinformatics*, **28**(15), 2008–2015.
- Paul, Joshua S, Steinrücken, Matthias, & Song, Yun S. 2011. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics*, **187**(4), 1115–1128.
- Rabiner, Lawrence R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.
- Schiffels, Stephan, & Durbin, Richard. 2014. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, **46**(8), 919–925.
- Sheehan, Sara, Harris, Kelley, & Song, Yun S. 2013. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics*.
- Steinrücken, Matthias, Kamm, John A, & Song, Yun S. 2015. Inference of complex population histories using whole-genome sequences from multiple populations. *bioRxiv*, 026591.
- Tataru, Paula, Nirody, Jasmine A, & Song, Yun S. 2014. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics*, **30**(23), 3430–3431.
- Terhorst, Jonathan, Kamm, John A, & Song, Yun S. 2016. *Robust and scalable inference of population history from hundreds of unphased whole genomes*. Tech. rept. Nature Research.
- Wilton, Peter R, Carmi, Shai, & Hobolth, Asger. 2015. The SMC is a highly accurate approximation to the ancestral recombination graph. *Genetics*, **200**(1), 343–355.
- Wiuf, Carsten, & Hein, Jotun. 1999. Recombination as a point process along sequences. *Theoretical population biology*, **55**(3), 248–259.