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

Supplementary Note

Pier Francesco Palamara[1,2,3], Jonathan Terhorst[4], 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, 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 $P(x = 1|t, \mu) = 1 - e^{-2\mu t_i}$, 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, $P(x = 0|t, \mu) = e^{-2\mu t_i}$. 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.* , 2017), extends the PSMC approach by incorporating knowledge of the frequency of the analyzed genetic polymorphisms in the emission model the of 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.* , 2017) for details). As in the MSMC approach, the transion 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.* , 2017). 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.* , 2017), 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 $\mathcal{O}(d\ell_p)$ computation under the conditional Simonsen-Churchil transition model (Hobolth & Jensen, 2014).

These two advances enable performing high-troughput 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 randonly 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.* , 2017). 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 probabily 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.* , 2017). 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 and 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 $ACSFS(\tau)_{d,u} = CSFS(\tau)_{d,u} \times P(obs|d+u) \times norm$, where $norm$ is a normalizing constant such that $\sum_{d=0}^{2} \sum_{u=0}^{n} ACSFS(\tau)_{d,u} = 1$. 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 $SFS_a(x)$ 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 ACSFS 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.* , 2017). 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),

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 $R_u = [T_u, T_{u+1})$

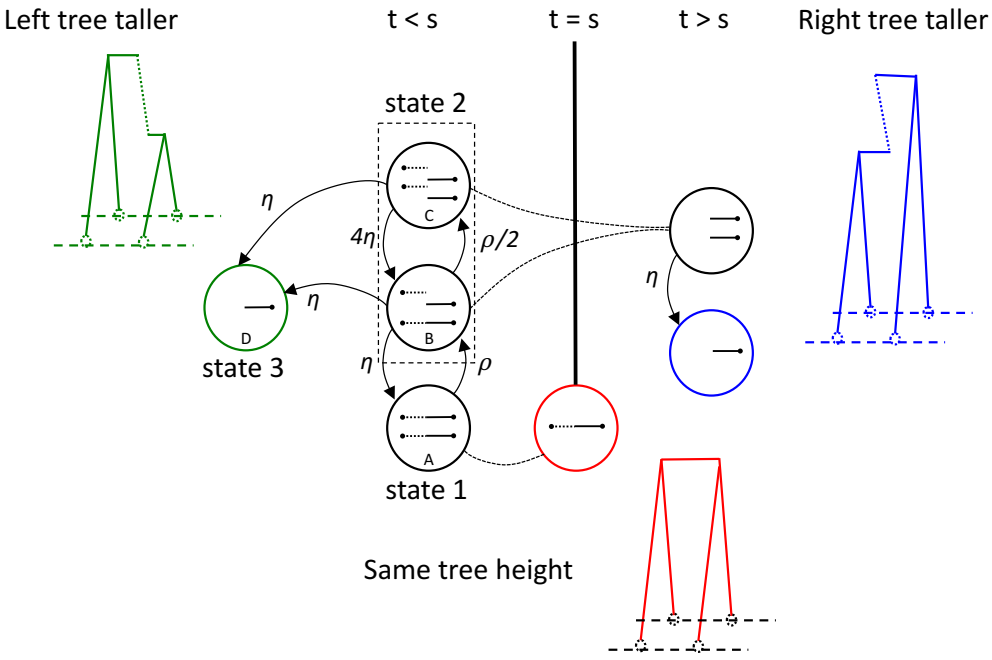


Figure 1: The conditional Simonsen-Churchil model (modified from Figure 1b of (Hobolth & Jensen, 2014)). Four relevant states from the full CSC model are labeled using letters within each circle.

the Markov chain of Figure 1 was used to descibe the distribution of ancestry at one locus conditional on the ancestry at the other locus. The transition matrix for this model is

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

where each row and colum 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 $C(t)_{i,j} = e^{tM}_{i,j}$ 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

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

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

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. $[\Omega(t)]_{12}$ respresents 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

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

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

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

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

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

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

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

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

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

where $M(...)$ and $m(...)$ indicate maximum and minimum, respectively. The rate $\int_{T_i}^{T_{i+1}} \eta(t)t\ dt = \int_{T_i}^{T_{i+1}} \eta_i t\ dt = \eta_i \left(T_{i+1} - T_i\right)$ 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

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

and the cumulative coalescent probability with

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

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

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

2.3.1.2 Discretization

Using Equation 5, the cumulative transition probability is

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

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

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

(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,

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

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 $i \in [1, d]$.
- 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 probablity 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

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

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. Inhomegeneous (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 $\alpha' = \alpha(TE_s)^n TE_p$, 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 $(TE_s)^n$ as $T^n E_s^n$. Having done that, we can now rely on the previosly described linear time algorithm to compute the product $\alpha(TE_s)^n TE_p \sim \alpha T^n E_s^n TE_p$. 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.

Appendix

Ratio of columns in the ASMC transition matrix

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

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

□

Ratio of rows in the ASMC transition matrix

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

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

□

Above diagonal elements

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

and

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

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. *Nature genetics*, **48**(7), 811.

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. *Nature genetics*, **48**(7), 817.

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*, **194**(3), 647–662.

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. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature genetics*, **49**(2), 303.

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.

Supplementary Information

This section provides additional details on simulations involving natural selection.

DRC_T simulations

We used the simulation setup recently adopted by Field et al.[1] 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 model[2], using the COSI2 coalescent simulator[3]. 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 simuPOP[4] 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. [5]), SDS (ref. [1]), or DRC_{150} (**Supplementary Figure 6a**). The iHS statistic was computed using the Selscan software[6] 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 12a**.

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 6b**). 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 shown[1]. We also report DRC_{20} results in **Supplementary Figure 12b**.

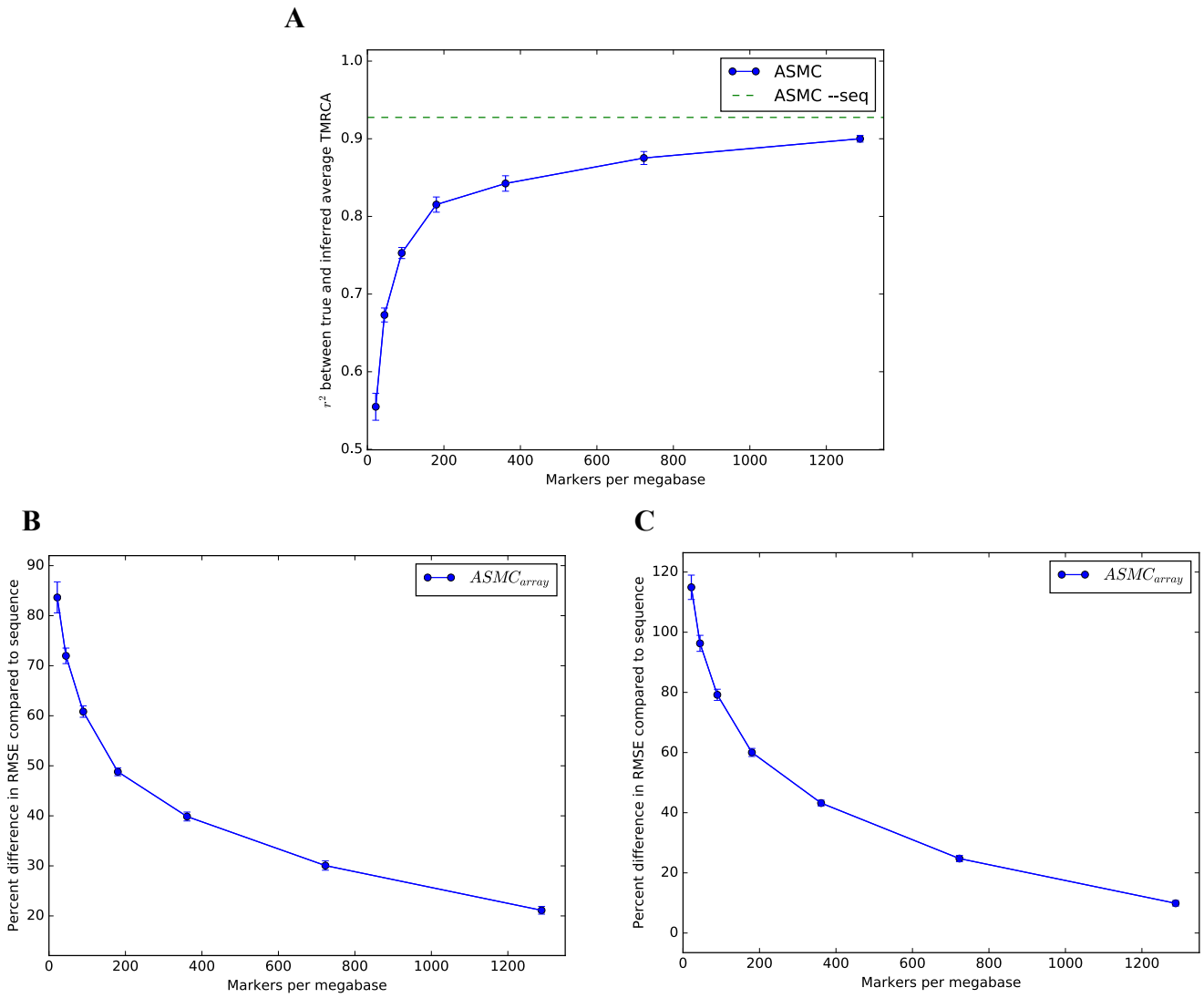
We performed additional simulation to evaluate the calibration of the null model. We observed an excellent fit for the DRC_{20} statistic (**Supplementary Figure 13a**), and only moderate inflation for the DRC_{150} statistic (**Supplementary Figure 13b**). The amount of inflation observed in the empirical null model obtained using the DRC_{150} statistic within the UKBB data set was consistent with our coalescent simulations (**Supplementary Figure 13c,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 14** 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.

$ASMC_{avg}$ simulations

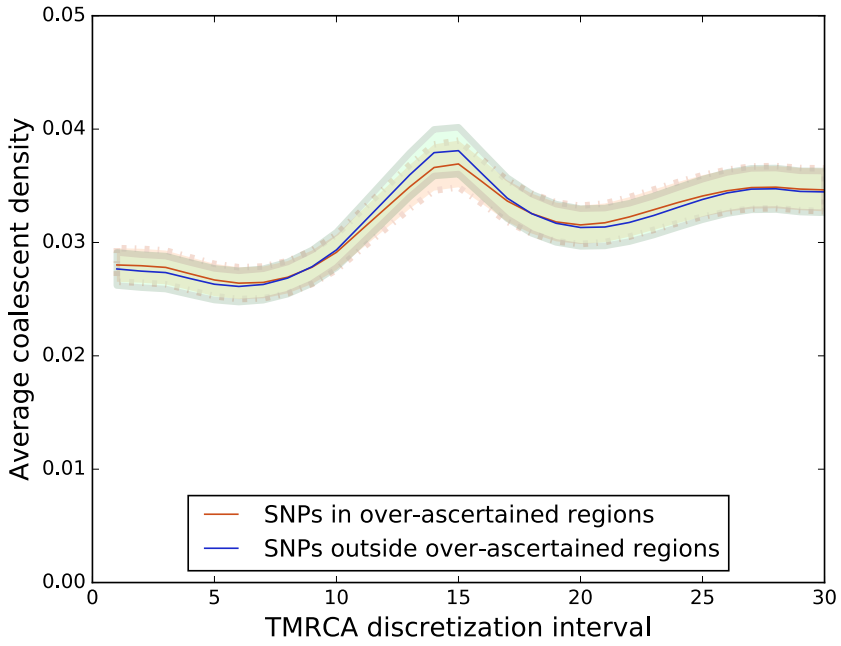
We performed forward-in-time simulations using the SLiM software[7] (v1.8) to test the effects of negative (background) or positive selection on the $ASMC_{avg}$ annotation. We simulated 3 Mb for a population of 10,000 diploid individuals, with recombination rate 1×10^{-8} and mutation rate 1.65×10^{-8} per base pair, per generation. Simulations were run forward in time for 200,000 generations. Within each genome, we simulated 3 equidistant 100 Kb-long regions undergoing either positive or negative selection. Selection coefficients for new mutations in regions undergoing negative selection were sampled from a gamma distribution with shape 0.2 and mean -5×10^{-4}, while positively selected regions had selection coefficients sampled from an exponential distribution with mean 10^{-4}. Within these regions, new mutations were neutral with probability

0.4, or under selection with probability 0.6. The dominance coefficient was set to 0.5 in all cases. We computed the $ASMC_{avg}$ annotation as previously described, using 300 haploid samples in each simulation. We simulated 50 independent replicates for positive and negative selection, as well as 100 neutral regions. Results are shown in **Supplementary Figure 8**.

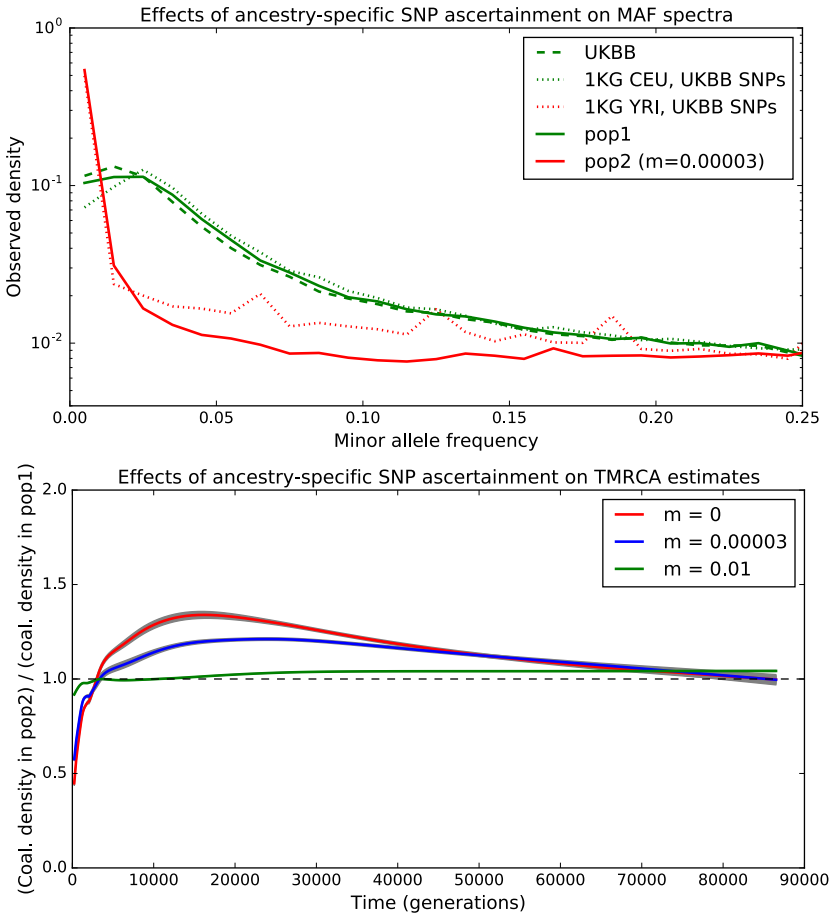
Supplementary Figures and Tables



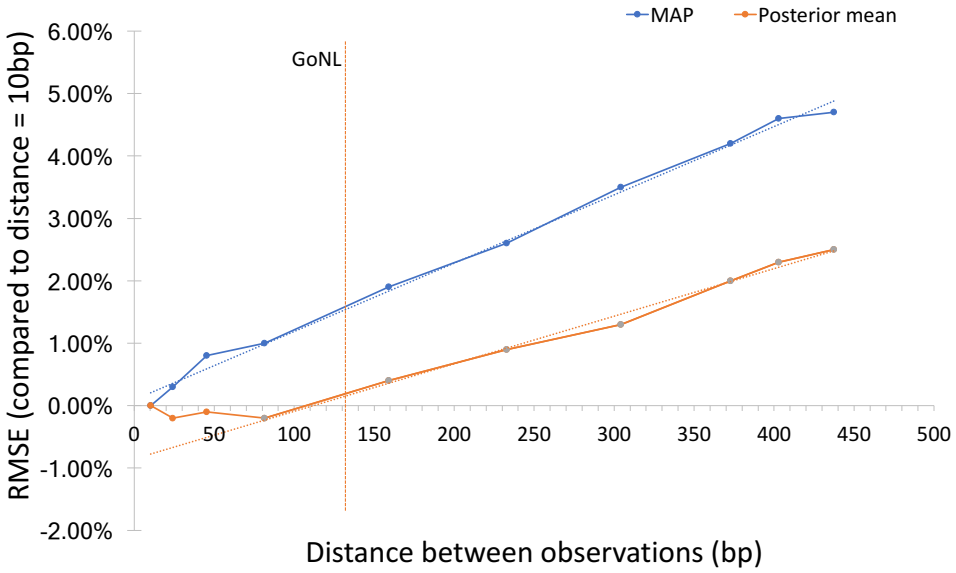
Supplementary Figure 1 – ASMC performance in simulations as a function of marker density. (A) r^2, as a function of marker density, between true average TMRCA within the simulated region and average TMRCA inferred using the maximum-a-posteriori (MAP) of the posterior distribution computed by ASMC. ASMC-seq represents the accuracy obtained using ASMC on WGS data. (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. In all panels dots and error bars represent average and SE from 10 independent simulations.



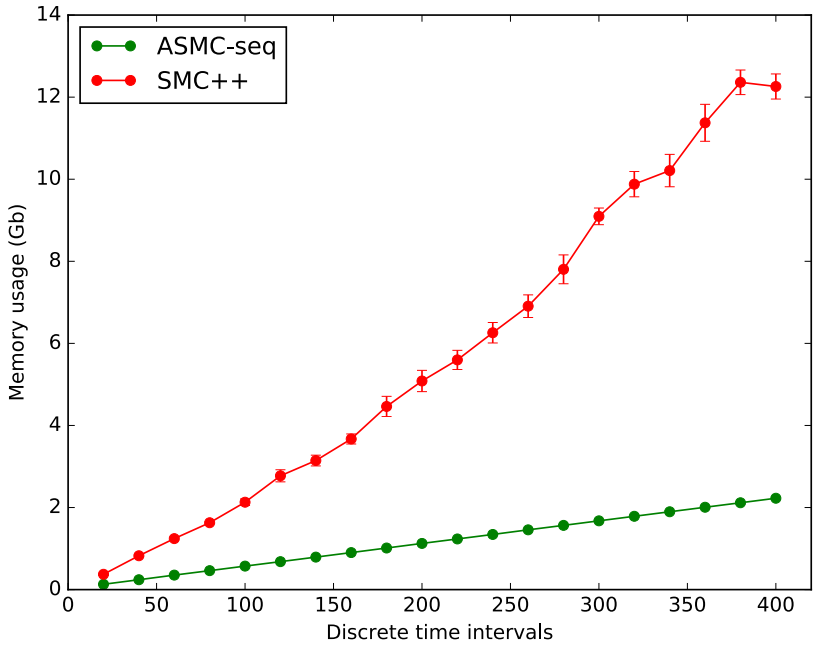
Supplementary Figure 2 - Robustness to deviations from frequency-based ascertainment. Approximately 25% of the variants found on the UK Biobank Axiom Array were selected based on their functional relevance, particularly in coding regions, while the remaining ~75% were ascertained based on frequency. 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. Lines represent average values from 20 independent simulations, shaded regions indicate 95% confidence intervals . We observed minimal deviation between coalescence densities inferred within and outside the simulated gene regions.



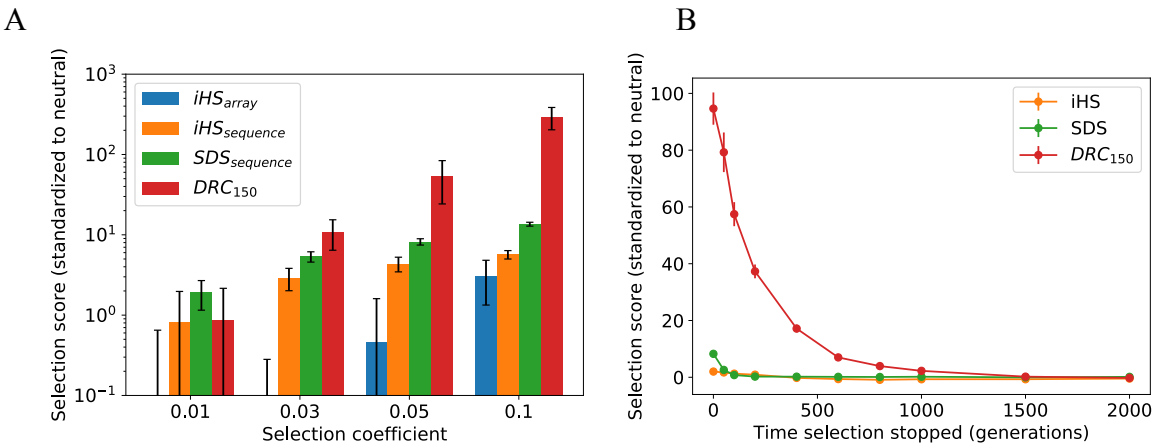
Supplementary Figure 3 – Effects of ancestry-specific SNP ascertainment. We simulated two populations that split 2,000 generations in the past. The two populations have identical, European-like effective size histories after the split, and a symmetric migration rate of 0.0, 0.00003, or 0.01. To simulate the effects of population-specific ascertainment of variants on array data, we selected SNPs from one of the two populations (pop1), matching the frequency spectrum observed in the UK Biobank dataset. SNPs in this set are expected to have drifted to different frequencies in the other population (pop2). The top plot shows allele frequency spectra for several real and simulated populations. We report frequency spectra for the UK Biobank (UKBB); 1,000 Genomes Project (1KG) European (CEU) and Yoruba (YRI) populations, for which only UKBB SNPs on Chromosome 2 are considered; simulated populations pop1 and pop2, with SNPs sampled as previously described. As expected, the simulated population pop2 exhibits a depletion of informative markers similar to what would be observed as a result of ancestry-specific SNP ascertainment for different continental populations. In the bottom plot, we used ASMC to infer coalescence times in both populations independently. We report a comparison (pop2/pop1 ratio) of the inferred average genome-wide coalescence density as a function of time for the two populations. Because the two populations have identical demographic history, the true expected ratio is 1. Ancestry-specific ascertainment, however, introduces a substantial depletion of informative markers for pop2, which leads to an upward bias in coalescence times inferred in pop2. The magnitude of the bias is mediated by the amount of post-split migration across the two groups. Lines represent averages from 25 independent simulations, gray bands represent one SD.



Supplementary Figure 4 – Effects of ASMC-seq transition approximation. When computing forward-backward probabilities in WGS data, ASMC-seq makes the approximation $(\mathrm{TE}_0)^n \cong \mathrm{T}^n\mathrm{E}_0^n$, 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 ~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 ~0.008% per base pair, and increased error below 0.5% at GoNL genotyping density.

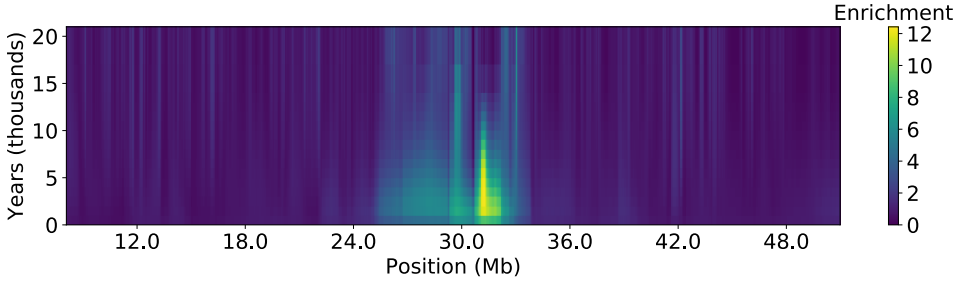


Supplementary Figure 5 – Memory use of ASMC-seq and SMC++. Memory usage for the analysis of coalescence times in a 5Mb region using WGS data from 100 haploid individuals. Dots represent averages from 10 independent simulations, bars represent SE.

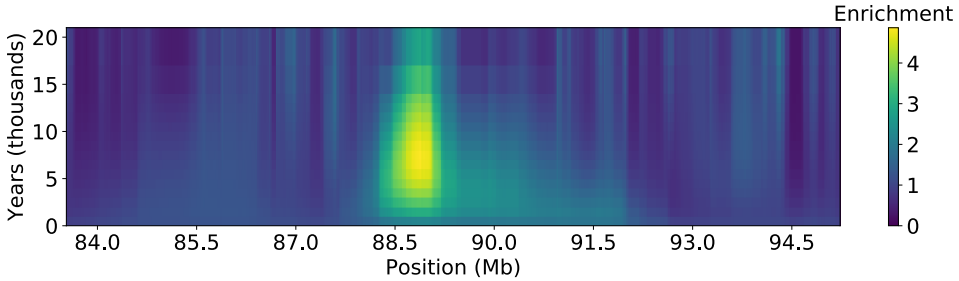


Supplementary Figure 6 – Selection simulations for DRC_{150}. (A) Simulation of different strengths of recent positive selection starting 200 generations in the past: iHS score [Voight et al. PLoS Biol. 2006] run on array data (iHS_{array}); iHS score on causal variant from sequencing data ($iHS_{sequence}$); SDS score [Field et al. Science 2016] 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. Reported values represent averages from 50 independent simulations, error bars represent SE. (B) Specificity to recent past for iHS and SDS run son sequencing data, and for DRC_{150}. Simulation of selection starting at time -∞ 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 ~20,000 years). Dots represent averages from 50 independent simulations, error bars represent SE. Numerical results are reported in **Supplementary Table 14.**

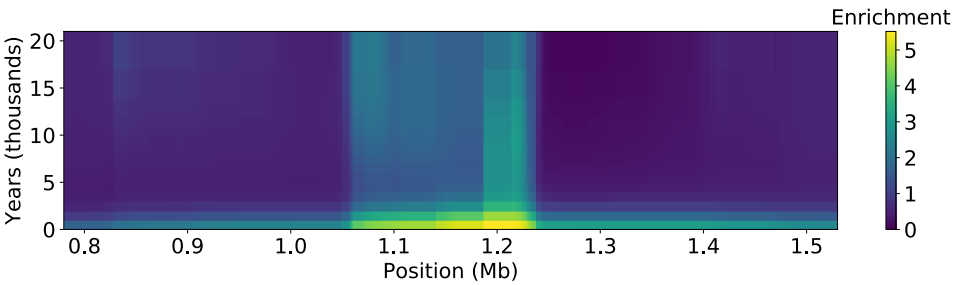
Chr 6



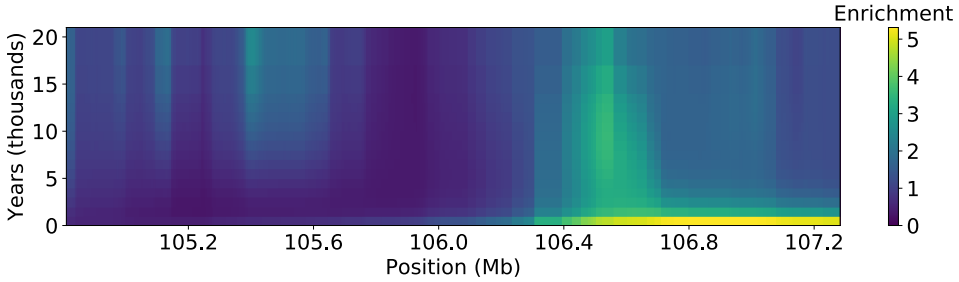
Chr 11



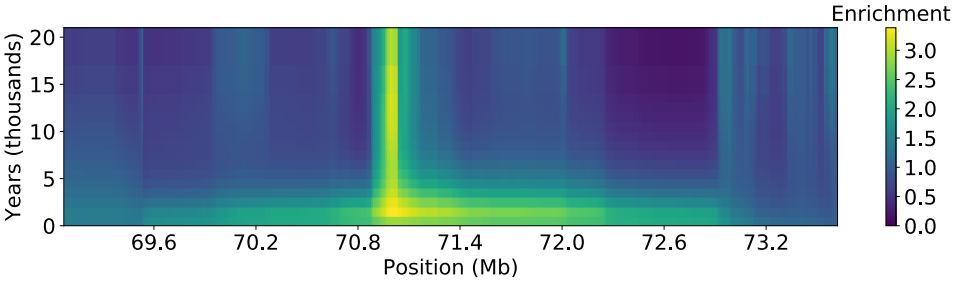
Chr 11



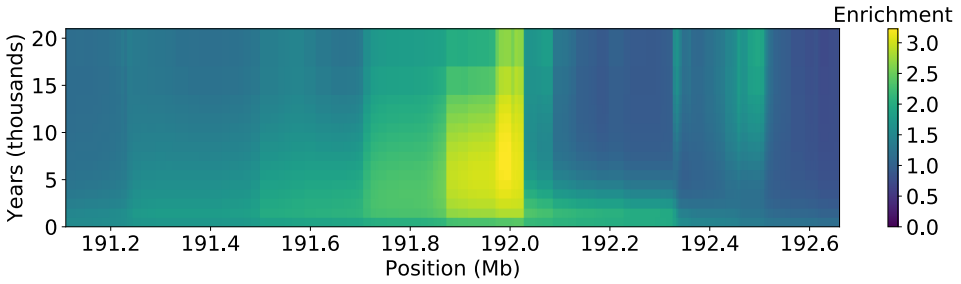
Chr 14



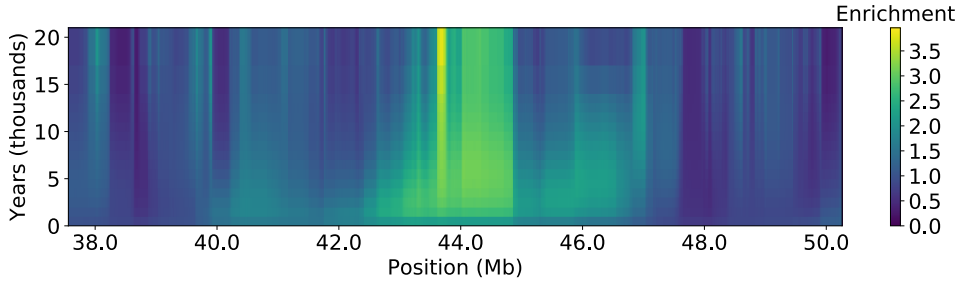
Chr 16



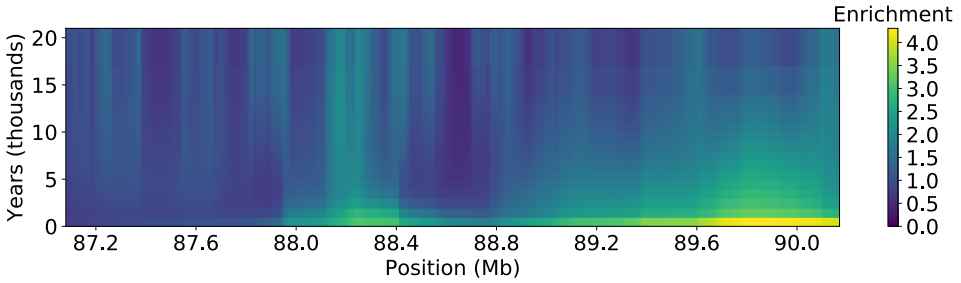
Chr 2



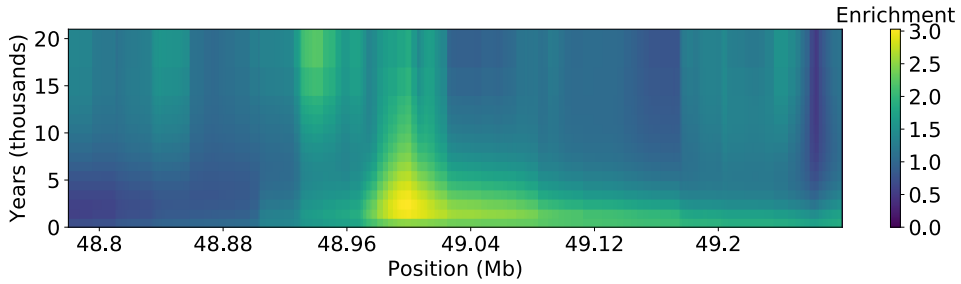
Chr 17



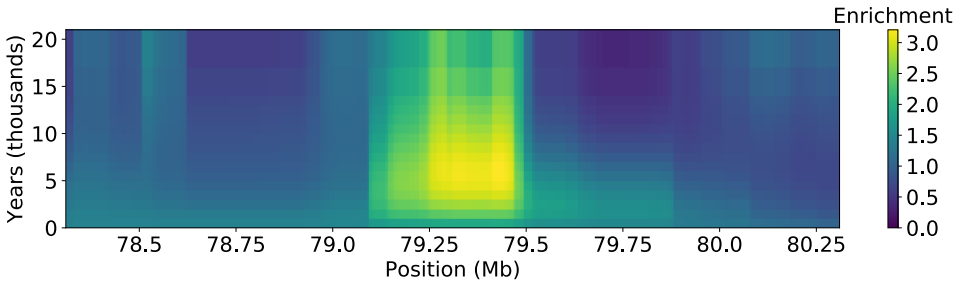
Chr 16



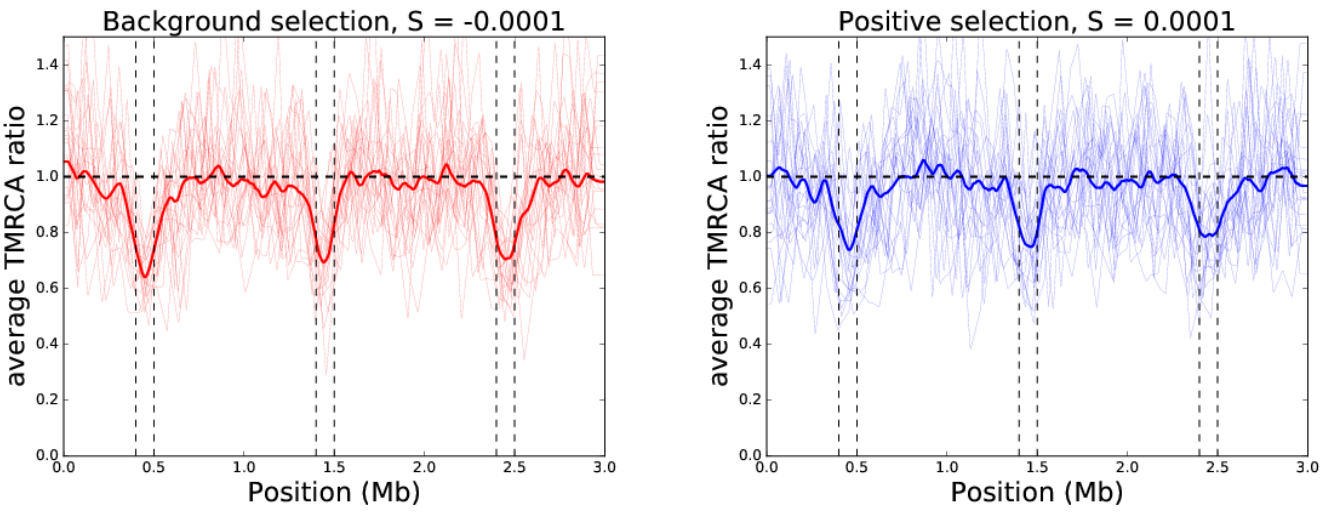
Chr 22



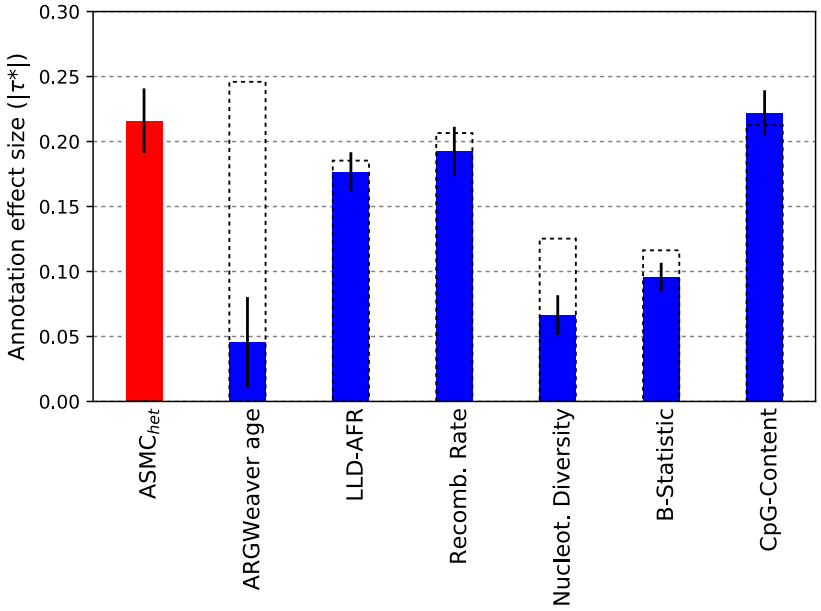
Chr 4



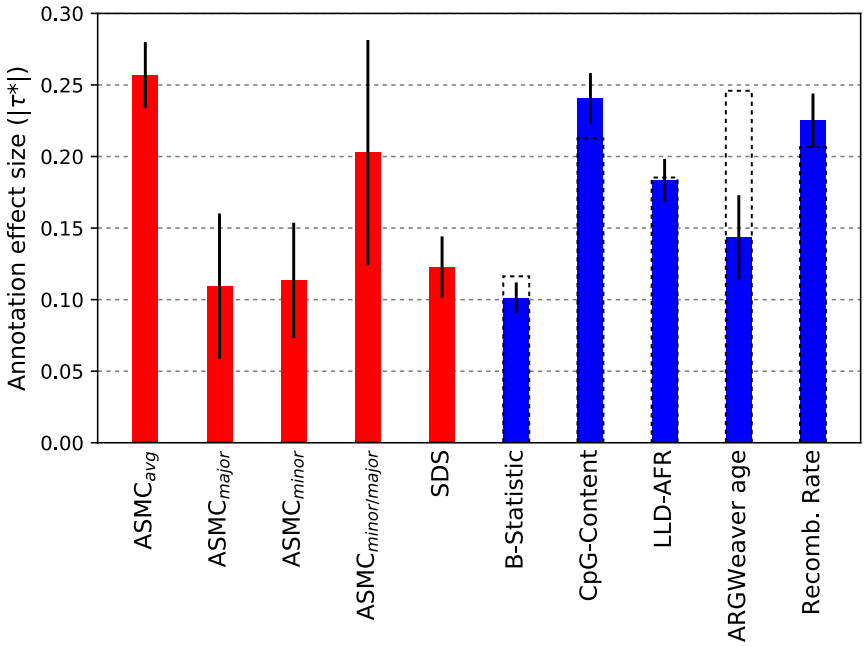
Supplementary Figure 7 – Enrichment of coalescence density in the past 20,000 years. At each site along the regions (horizontal axis) we plot the enrichment for the density of coalescence events in the past ~20,000 years, computed as $\frac{\text{posterior}_{site,time}}{\text{posterior}_{genomewide,time}}$. Time axes assumes a 30-year generation.



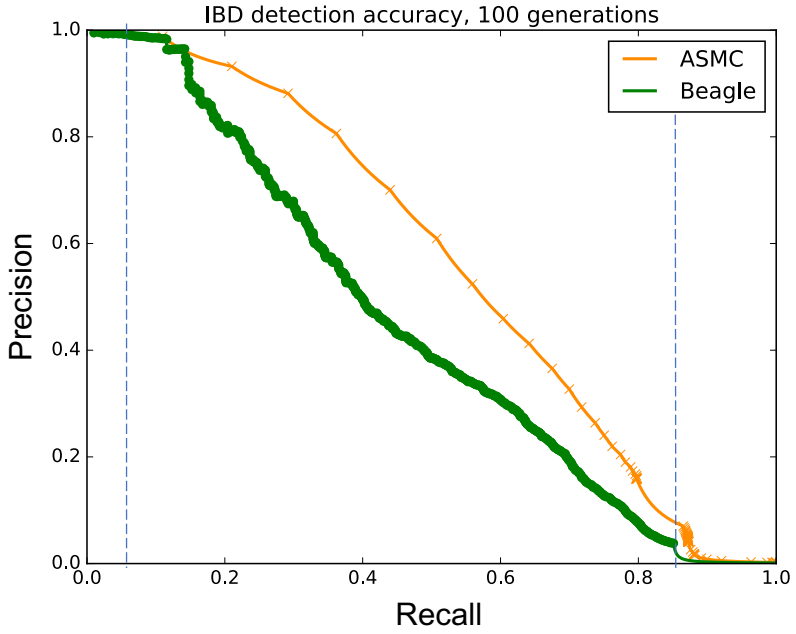
Supplementary Figure 8 – Effects of background and positive selection on the $ASMC_{avg}$ annotation. We used the SLiM software (v1.8) to simulate a 3 Mb genome for a population of 10,000 diploid individuals, with recombination rate 1×10^{-8} and mutation rate 1.65×10^{-8} per base pair, per generation. Simulations were run forward in time for 200,000 generations. Within each genome, we simulated background selection (red lines, left plot) or positive selection (blue lines, right plot) in four 100-Kb long regions (delimited by vertical dashed lines in each plot, coordinates in Mb: 0.4 to 0.5, 1.4 to 1.5, and 2.4 to 2.5). Selection coefficients for new mutations in regions of background selection were sampled from a gamma distribution with shape 0.2 and mean -5×10^{-4}, resulting in average selection coefficients $S=-10^{-4}$. Selection coefficients for new mutations in positively selected regions were sampled from an exponential distribution with mean 10^{-4}, resulting in average selection coefficients $S=10^{-4}$. Within these regions, new mutations were neutral (S=0) with probability 0.4, and under selection with probability 0.6. Dominance coefficients were set to 0.5 in all cases. We performed 50 independent replicates for background and positive selection, and 100 additional neutral simulations (S=0 for all mutations). We computed the $ASMC_{avg}$ annotation as previously described, using 300 haploid samples in each simulation. Lines in the figures represent the value of the ratio T_{sel}/T_{neut} along the genome, where T_{sel} is the value of the $ASMC_{avg}$ annotation for a simulation involving background or positive selection, and T_{neut} is the value of the $ASMC_{avg}$ annotation, averaged across all neutral simulations. Within each plot, thin lines represent the results of 20 randomly selected individual simulations, thick lines represent the average across all 50 replicates involving selection. In all simulations, $ASMC_{avg}$ decreases in regions undergoing selection compared to neutral regions, with an average reduction of 32% for background selection (z-test z = -30.9), and a 22% reduction for positive selection (z = -16.0). Variance of the $ASMC_{avg}$ annotation within regions under selection is also lower (-33% on average for background selection, z = -19.3; -21% on average for positive selection, z = -9.2). Considering regions as a whole, simulations involving selection had a lower mean value of the $ASMC_{avg}$ annotation compared to neutral simulations (-6%, on average for background selection, z = -11.6; -5% on average for positive selection, z = -8.5); and slightly lower variance (-3%, on average for background selection, z = -2.9; -2%, for positive selection, z = -2.6).



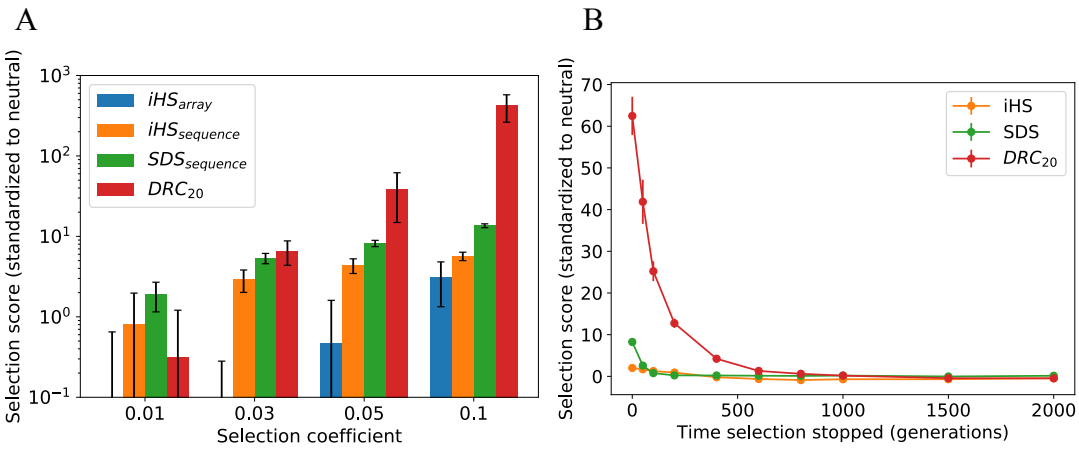
Supplementary Figure 9 - S-LDSC analysis of ASMC$_{het}$ background selection annotation and disease heritability. We built an annotation, ASMC$_{het}$, reflecting the average coalescence time for heterozygous individuals (i.e. chromosomes carrying discordant alleles) at each site. As for the ASMC$_{avg}$ 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 (SE 0.012) to -0.046 (SE 0.018). We report τ^* value of the ASMC$_{avg}$ annotation for 20 independent diseases and complex traits (sample sizes in Supplementary Table 8). Error bars represent SE of the τ^* estimate. Dashed bars reflect values for six baselineLD annotations linked to background selection before the introduction of the ASMC$_{het}$ annotation.



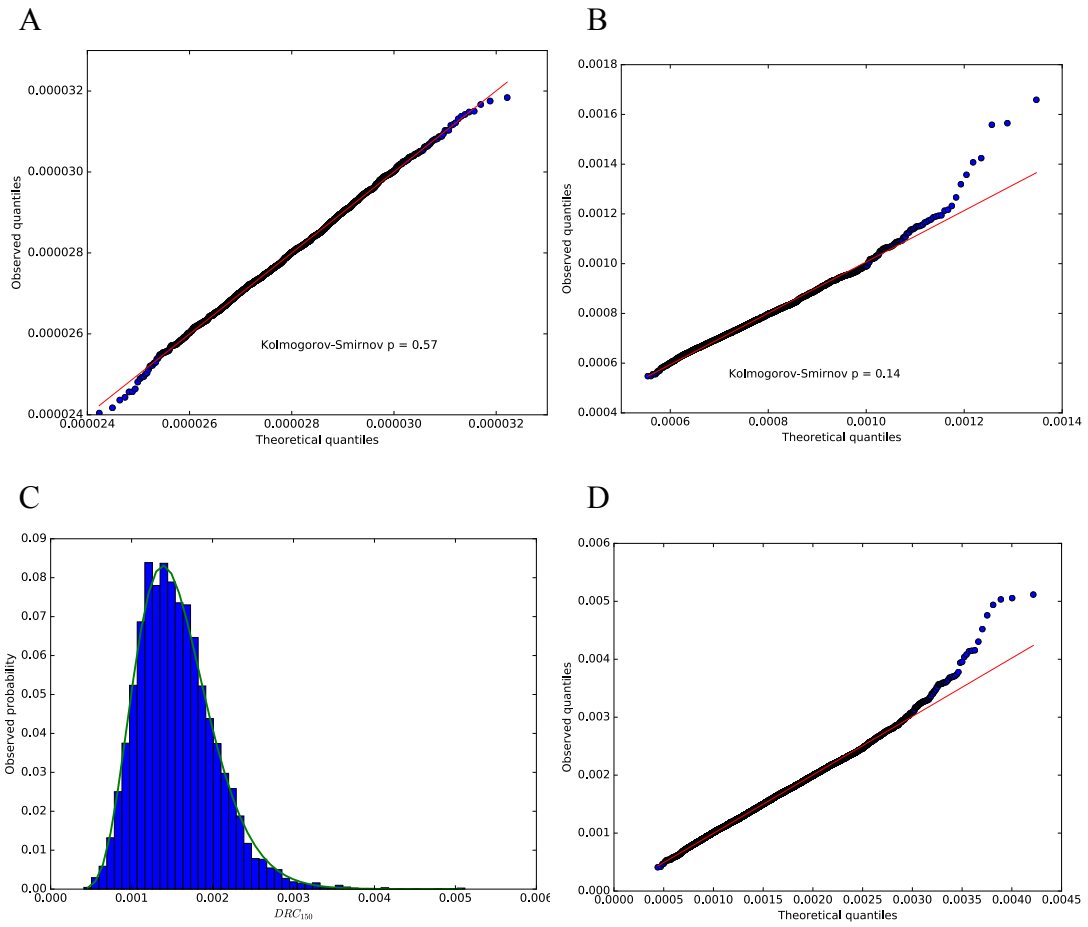
Supplementary Figure 10 - S-LDSC analysis of several annotations related to background selection. We built several annotations related to average coalescence time at each site, conditioning on the allele present on each analyzed chromosome from the GoNL data set. In addition to the $ASMC_{avg}$ annotation (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 $\log(T_{minor}/T_{major})$ 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 [Field et al. Science 2016], 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 (sample sizes in Suppleentary Table 8). Error bars represent SE of the τ^* estimate. 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.



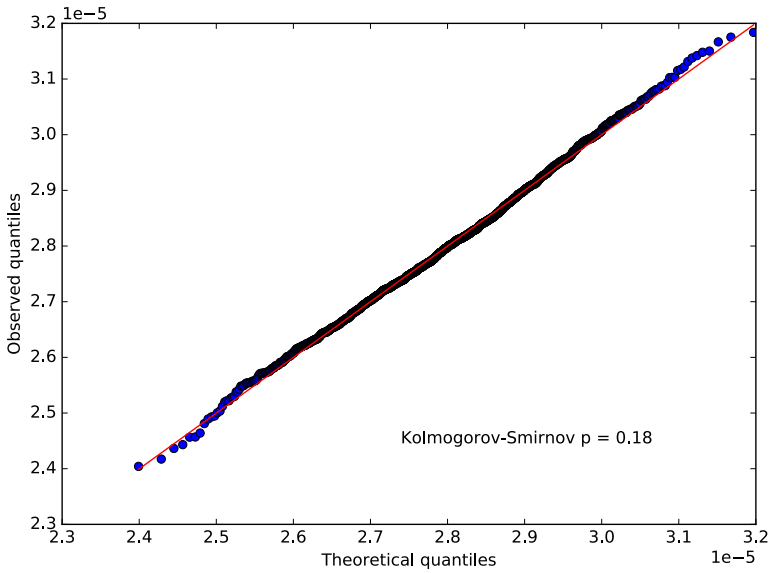
Supplementary Figure 11 – Illustration of auPRC measure for IBD detection accuracy. We measured accuracy of IBD detection for ASMC and Beagle using the area under the precision-recall curve (auPRC) for both programs. For both methods, recall can only be estimated within a limited precision range, due 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 [Davis and Goadrich, ICML 2006]. Results are from a single simulation of 300 samples and 30Mb. Averages across multiple simulations are reported in Supplementary Table 1.



Supplementary Figure 12 – Selection simulations for DRC_{20}. (A) Simulation of different strengths of recent positive selection starting 200 generations in the past: iHS score [Voight et al. PLoS Biol. 2006] run on array data (iHS_{array}); iHS score on causal variant from sequencing data ($iHS_{sequence}$); SDS score [Field et al. Science 2016] 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. Reported values represent averages from 50 independent simulations, error bars represent SE. (B) Specificity to recent past for iHS and SDS run son sequencing data, and for DRC_{20}. Simulation of selection starting at time $-\infty$ stopping at the specified generation, followed by neutral drift. Dots represent averages from 50 independent simulations, error bars represent SE.



Supplementary Figure 13 – Empirical null model. (A) QQ plot for the DRC_{20} statistic in 2,000 independent neutral coalescent simulations using the European demographic model of [Tennessen et al. Science 2012]. (B) QQ plot for the DRC_{150} statistic in 2,000 independent neutral coalescent simulations using the European demographic model of [Tennessen et al. Science 2012]. (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 14 – 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 [Tennessen et al. Science 2012], 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 11**). Averages and SE in brackets were computed from 30 independent simulations.

IBD time threshold	Average recall range	Avg auPRC within recall range		Average 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. Averages (SE) were obtained 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 11**). 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. Averages (SE) were computed from 30 independent simulations.

IBD time threshold	Average recall range	Avg auPRC within recall range		Average percent AuPRC improvement: $100\times\left(\text{auPRC}_{ASMC}/\text{auPRC}_{Beagle}-1\right)$
		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. 300 haploid samples from the first 30Mb of a human Chromosome 2 and a European demographic model were simulated for each map type. 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 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**). 300 haploid samples from the first 30Mb of a human Chromosome 2 and a European demographic model were simulated for each number of discretization intervals. 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 ($10^{-4} < p < 10^{-3}$). The DRC_{150} statistic of recent positive selection was computed using all individuals of British ancestry from the UK Biobank (n=113,851, divided in batches of ~10,000 samples; see Online Methods for details on how p-values were computed).

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, QRICH1, 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[8]
5	129.56-131.81	4.69×10^{-6}	rs739718	SLC22A4[8]
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[9]

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, TINF2, 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[8]
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, CHRNB1, 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, SCRN2, 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 – Genome-wide c orrelation between $ASMC_{avg}$ and other annotations from the baselineLD model.

baselineLD annotation	Correlation with $ASMC_{avg}$ (r)
B-statistic[10]	-0.28
CpG content	0.03
Recombination rate	0.07
LLD-Africa[11]	0.08
ARGWeaver allele age[12]	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[13]	32,143	9.58	http://www.med.unc.edu/pgc/downloads/
Autism spectrum[14]	10,263	8.72	http://www.med.unc.edu/pgc/files/resultfiles/pgcasdeuro.gz
Blood pressure, diastolic (UKBB)	134,011	24.12	.
BMI[15]	122,033	17.45	http://www.broadinstitute.org/collaboration/giant/index.php/GIANT_consortium_data_files
Coronary artery disease[16]	77,210	8.47	http://www.cardiogramplusc4d.org/
Crohn's disease[17]	20,883	10.34	http://www.ibdgenetics.org/downloads.html
Eczema (UKBB)	145,416	11.42	.
Height (UKBB)	145,368	29.29	.
LDL[18]	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[19]	170,911	9.54	http://ssgac.org/documents/
Putamen volume[20]	12,924	7.08	http://enigma.ini.usc.edu/wp-content/uploads/E2_EVIS
Rheumatoid arthritis[21]	37,681	10.20	http://plaza.umin.ac.jp/yokada/datasource/software.html
Schizophrenia[22]	70,100	21.82	http://www.med.unc.edu/pgc/downloads/
Smoking status (UKBB)	145,227	19.70	.
Systemic lupus erythematosus[23]	14,267	6.37	https://www.immunobase.org/downloads/protected_data/GWAS_Data/
Years of education[24]	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 $ASMC_{TMRCA}$ 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**
	1st	**2nd**	**3rd**	**4th**	**5th**	
$ASMC_{avg}$	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[11]	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[12]	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[10]	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. Averages (SE) were 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. Averages (SE) were computed using 10 independent simulations.

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 5**. 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. Averages (SE) were computed using 10 independent simulations.

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 (SE) of the $ASMC_{avg}$ annotation for 20 independent diseases and complex traits (sample sizes in Supplementary Table 8). (B) Absolute values of τ^* (SE), meta-analyzed across 20 independent diseases and complex traits (sample sizes in Supplementary Table 8). τ^* values were computed in a joint analysis conditioned on baselineLD annotations. Numerical values for **Figure 4c** can be found in **Supplementary Table 9**.

A

Trait	τ^* (SE)
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[12]	-0.133 (0.013)	-0.246 (0.012)
LLD-Africa[11]	-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[10]	0.102 (0.006)	0.116 (0.006)
CpG content	0.220 (0.009)	0.213 (0.009)

References

1. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760-764 (2016).
2. Tennessen, J.A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337**, 64-9 (2012).
3. Shlyakhter, I., Sabeti, P.C. & Schaffner, S.F. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics* **30**, 3427-9 (2014).
4. Peng, B. & Amos, C.I. Forward-time simulations of non-random mating populations using simuPOP. *Bioinformatics* **24**, 1408-9 (2008).
5. 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).
6. 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).
7. Messer, P.W. SLiM: simulating evolution with selection and linkage. *Genetics* **194**, 1037-9 (2013).
8. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499-503 (2015).
9. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-8 (2007).
10. McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet* **5**, e1000471 (2009).
11. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* (2017).
12. Rasmussen, M.D., Hubisz, M.J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet* **10**, e1004342 (2014).
13. Boraska, V. *et al.* A genome-wide association study of anorexia nervosa. *Mol Psychiatry* **19**, 1085-94 (2014).
14. 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).
15. 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).
16. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet* **43**, 333-8 (2011).
17. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119-24 (2012).
18. Teslovich, T.M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707-13 (2010).
19. 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).

20. Hibar, D.P. *et al.* Common genetic variants influence human subcortical brain structures. *Nature* **520**, 224-9 (2015).
21. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376-81 (2014).
22. Schizophrenia Working Group of the Psychiatric Genomics, C. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421-7 (2014).
23. 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).
24. Rietveld, C.A. *et al.* GWAS of 126,559 individuals identifies genetic variants associated with educational attainment. *Science* **340**, 1467-71 (2013).