Inference and Applications of Genetic Relatedness in Human Populations



Juba Nait Saada
Balliol College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2022

To my parents,
for their sacrifices
and unlimited support.

Abstract

Exploring the human genome holds benefits for many fields, which include molecular medicine, human evolution, population genetics and genetic epidemiology. The human genome contains around 3 billion base pairs and approximately 99.9 percent of these base pairs are identical between any two individuals. Differences in the remaining 0.1 percent could explain the astonishing variety between individuals - not only in appearance, but also in health.

Building upon this observation, large-scale genomic collections, through efforts like the NIH All of Us research program or the UK BioBank, have yielded data sets of hundreds of thousands of individuals and are expected to grow to millions in the coming years. Leveraging such data sets to understand disease and phenotypic variation requires understanding the fine-scale genetic relationships between individuals.

The work presented in this thesis is spread across three main chapters including two published articles and one preprint, and they are summarised as:

Chapter 2 Nait Saada et al. (2020) focuses on detecting genetic relationships using "identical-by-descent" genomic regions that are inherited from a common ancestor between purportedly "unrelated" pairs of individuals in a data set. IBD segments originating within the past 1,500 years are inferred between 487,409 UK Biobank individuals, revealing fine scale population structure, signals of recent selection and ultra-rare variant associations.

Chapter 3 Nait Saada et al. (2021) explores the use of deep neural networks for accurately predicting the time to the most recent common ancestor (TMRCA), or coalescence time, between individuals at a specific genomic location. Pairwise TMRCAs can be leveraged to estimate the age of observed genomic variants in a data set.

Chapter 4 Nait Saada et al. (2022) infers ages of ~ 80 million mutations across 26 diverse populations in the $1,000$ Genomes Project data set using deep learning. By leveraging allele age estimates, signatures of selection for individual variants and complex traits are identified.

Contents

1 Background **1**
1.1 Statistical genetics . . . 1
1.1.1 Basic definitions . . . 1
1.1.2 Quantitative genetic model . . . 6
1.1.3 Trait association testing . . . 8
1.1.4 Population structure . . . 11
1.1.5 Partitioning heritability . . . 13
1.2 Population genetics . . . 14
1.2.1 Coalescent theory . . . 14
1.2.2 Coalescent Hidden Markov Models . . . 18
1.2.3 Identity By Descent . . . 21
1.2.4 Natural selection . . . 22
1.3 Deep learning . . . 24
1.3.1 Basic definitions . . . 24
1.3.2 Architectures . . . 30
1.3.3 Explainable deep neural networks . . . 33
1.3.4 Recent advances for population genetics . . . 34

2 Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations **49**

3 Inference of pairwise coalescence times and allele ages using deep neural networks **103**

4 Inference of genealogical relatedness and variant age using deep neural networks identifies signatures of selection **118**

5 Conclusion **177**
5.1 Summary . . . 177
5.2 Future work . . . 179

Acknowledgments

I would first like to thank my supervisor Pier Palamara for having given me the opportunity to conduct research in the amazing, stimulating and beautiful place that is Oxford. I have learned so much by his side in the past five years. Thanks to all the members of the lab, to my collaborators and to the Department of Statistics for creating a vibrant environment for pursuing research. Thank you to Alexander Gusev and Robert Davies who have been such a great source of inspiration and who have provided me support throughout my DPhil. I am very grateful to my thesis committee members for the time dedicated to my work. Thanks to Balliol College for having awarded me the Jowett Scholarship and to the Medical Research Council for their financial support under the Oxford – MRC Doctoral Training Partnership.

I would specially like to thank my office mates and friends Brian Zhang and Petya Kindalova, I am so grateful that the three of us embarked on the DPhil adventure together. I would also like to thank Zoi Tsangalidou who not only helped me cross the finish line, but also reminded me to laugh when I did not even want to smile.

I am particularly grateful to the Balliol MCR, to the Holywell Manor community and to my amazing Balliol friends who made my college experience truly one of a kind. Thanks to all the friends I was fortunate to make during my time in Oxford. Special thanks go to Ashna Patel, Deborah Sulem, Eloise Hamilton, Lois Ogunlana, Saad Labyad and Stefania Vietti-Michelina for making my time in the city of dreaming spires so enjoyable.

I am extremely thankful to my friends and confidants Anthony Hu and Oumayma Bounou, who both did a PhD at the same time as me. Sharing with them my anxiety, excitement, research ideas, happiness and disappointments really helped me go through this journey. My sincere thanks also go to Aymen El Hadidy, Cecile Gressard, Clement Ngassam and Victor Chatelais for always being there for me. This journey would not have been possible without Marie Ver Eecke, who encouraged me to strive for excellence. She has changed my life forever and I will always be immensely grateful to her.

To my family, Antinia, Belkacem, Karima, Macicilia and Thiziri, for their unlimited love, kindness and devotion. Thanks to my wonderful sisters for always believing in me, I am so blessed to have you in my life and I don't know what I would do without you. Finally, I want to specifically express my gratitude to my mom for her endless support. She taught me essential values and she will always be my role model. Thank you mams, for everything.

Chapter 1

Background

1.1 Statistical genetics

1.1.1 Basic definitions

The genetic instructions used in the functioning and reproduction of all known living organisms and many viruses are contained in a molecule called deoxyribonucleic acid, or **DNA**, first isolated by Friedrich Miescher in 1869[1]. In eukaryotic organisms (whose cells have a nucleus enclosed within a nuclear envelope), the hereditary material is believed to be contained in both nuclear and mitochondrial DNA, and regulated by epigenetic markers. In what follows, we will only focus on nuclear DNA in humans, and we will refer to it as, simply, DNA.

DNA is organized into long structures called **chromosomes**. Humans have 23 pairs of chromosomes, giving a total of 46 per cell and they can be divided in two groups:

one pair of sex chromosomes (or allosomes), and 22 pairs of non sex chromosomes (or **autosomes**). Because two copies of each chromosome are stored in cells, humans are called **diploid** organisms, as opposed to **haploid** organisms who only have one copy of each chromosome. During the process of **meiosis** (first described by the biologist Oscar Hertwig in 1876[2]), diploid individuals produce **gametes** (egg cells for female individuals and sperm cells for male individuals) for sexual reproduction. The process of meiosis is a type of cell division that reduces the chromosome number by half. The gametes are haploid cells: they contain only one copy of each chromosome and are formed by mixing the two original existing copies. Two gametes, one from the maternal parent and one from the paternal parent, fuse during fertilization, creating a diploid cell with a complete set of paired chromosomes. Two chromosomes in a pair are called **homologous** chromosomes.

DNA is composed of two strands, known as polynucleotides, each of them composed of simpler monomeric units called nucleotides. Its molecular double helix structure was identified in 1953 by Watson and Crick[3]. Each nucleotide is composed of one of four nitrogen-containing **nucleobases** (adenine [A], cytosine [C], guanine [G] or thymine [T]), a sugar called deoxyribose, and a phosphate group. Each base couples with its complement (A with T, and C with G) to form a **base pair (bp)**. The order of succession of the nucleobases in each strand constitutes the **sequence** of the strand.

A specific location in the sequence may be referred to as a **genomic site**, or simply as a site. A **gene** is region of DNA (i.e. a set of sites) encoding a **protein**. Proteins are made of amino acids and are essential to life, as they are responsible for most of a cell's functionality. A **locus** (plural **loci**) is the physical position of a specific gene on a chromosome. The haploid human genome contains around 3 billion base pairs and about 30,000 genes. The **exome** refers to all coding regions (or exons) and constitutes about 1.1% of the total genome[4,5].

Due to the occurrence of **mutations** (i.e. a change in the DNA sequence) and other rare genomic events, different versions of the DNA sequence (a single base or a segment of bases) at a given genomic location may exist in a population, referred to as **alleles**. **Recessive** alleles only manifest if present in both homologous chromosomes, while **dominant** alleles show their effect even if the individual only has one copy. Two haplotypes are said to be **homozygous** at a given genomic site if they both carry the same allele, and they are said to be **heterozygous** if they carry different versions.

There are different classes of mutations, including point mutations, that occur at the nucleotide level, and chromosomal mutations, that alter a whole region of a chromosome. We will only focus on point mutations where a single nucleotide base is changed (substitution), inserted (insertion) or deleted (deletion) from a DNA sequence. Single base pair alterations that are common in the population (e.g. frequency $> 1\%$) are

referred to as **"single nucleotide polymorphisms" (SNPs**).

In preparation for meiosis, the DNA is duplicated, resulting in 23 pairs of chromosomes where each chromosome is made of 2 identical sister **chromatids**. During prophase I of meiosis, homologous maternal and paternal chromosomes pair and exchange genetic information in different ways, a process called **recombination**. For instance, homologous chromosomes' non-sister chromatids can exchange different segments of genetic material, referred to as crossing over, which results in recombinant chromosomes. Recombination leads to new combinations of alleles (i.e. offspring have different combinations of genes from those of their parents) and increases genetic variation. Recombination is believed to play an important role to correct damages to the DNA sequences (**DNA repair**)[6]. A **centimorgan** (abbreviated cM) is a unit of measure for the frequency of genetic recombination such that if two sites are seperated by one centimorgan there is on average 0.01 recombination events occurring between them in one generation. On average in the human genome, 1 cM approximately corresponds to 1 million bp.

The growing interest in better understanding diseases and health outcomes has resulted in a worldwide effort to collect genetic data. Research programs, such as the NIH All of Us initiative[7], the UK BioBank[8], and Genomics England[9], have already produced data sets comprising hundreds of thousands of individuals. Genotyping allows

us to explore genetic variation between individuals by determining their DNA sequences at specific positions within the human genome, called **genotypes**. A DNA microarray chip can be used to measure the genetic information in some regions of the genome. For instance, **SNP array** is a tool for targeting and measuring SNPs specifically. The set of genotyped SNPs (or tagged SNPs) captures the most significant genetic variation between individuals, but fails to identify rare and de novo (not inherited from a parent) mutations. On the other hand, **whole-genome sequencing (WGS)** is the process of determining (nearly) the entire DNA sequence of an organism. In 2008, the advent of "next-generation sequencing" technologies replaced previous methods by leveraging massively parallel processing. From 2001 to 2021, the price of sequencing a single human genome dropped from 100 million dollars to 1,000 dollars and is still expected to decrease[10]. Another approach, called **whole-exome sequencing (WES)**, consists of sequencing only the protein-coding regions at a currently lower cost than WGS. However, studies have shown that non-coding regions of the human genome play an important role in regulating complex traits and diseases[11].

Genotypes capture the allelic variation at each site on both homologous chromosomes and do not record the origin of each allele (maternal or paternal). **Phasing** is the process of separating maternally and paternally inherited copies of each chromosome into **haplotypes** (or haploid genotypes)[12]. A haplotype is a group of alleles that

are inherited together from a single parent. Finally, note that unobserved genotypes can be inferred by using a set of known haplotypes in a population called a **reference panel** (e.g. from the 1,000 Genomes Project[13]), through a process named **imputation**[14,15].

1.1.2 Quantitative genetic model

The field of Statistical Genetics builds the theory of the statistical relationship between genotypic variation and phenotypic variation. A **phenotype** is any observable trait and can be a **case-control trait** (or binary trait, with two possible values, e.g. presence or absence of the disease), or a **quantitative trait** (continuous range of values, e.g. height). A **complex trait** is the result of multiple genetic and environmental factors, while a **Mendelian trait** is controlled by a single locus. Originally developed in 1918 by Ronald Fisher, the **infinitesimal model** supposes that variation in a quantitative trait is influenced by an infinite number of genetic markers, each of which with an infinitely small contribution, as well as by environmental factors. The model provides the following relationship between the phenotypic value Y, the genotypic value G (i.e. all genetic loci that influence the phenotype), the environmental deviation E (i.e. all non-genetic factors that influence the phenotype) and the interaction between genotype

and environment $G * E$:

$$Y = G + E + G * E$$

The phenotypic variation σ_Y^2 in a population can be partitioned as:

$$\sigma_Y^2 = \sigma_G^2 + \sigma_E^2 + 2 * \sigma_{G,E}^2 + \sigma_{G*E}^2$$

where σ_G^2 is the genotypic variation, σ_E^2 is the environmental variation, $\sigma_{G,E}^2$ is the G and E covariation and σ_{G*E}^2 is the $G*E$ interaction variation. Both G and E covariation and $G * E$ interaction are usually difficult to estimate and often ignored[16], in which case σ_Y^2 is partitioned as:

$$\sigma_Y^2 = \sigma_G^2 + \sigma_E^2$$

According to the model, the genotypic value G at a given locus can be partitioned into an **additive** component A (mean deviation caused by an allelic substitution with no interactions between alleles of the same locus or between different loci) and deviations from additivity as a result of **dominance** D (interactions between alleles at the same locus) and **epistasis** I (interactions between alleles at different loci):

$$G = A + D + I$$

The genotypic variation can be further subdivided into an additive genetic variance associated with the average additive effects of alleles σ_A^2, a deviation due to dominance interactions σ_D^2 and an epistatic deviation due to interactions across genes σ_I^2:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

The **narrow-sense heritability**, or simply **heritability**, h^2 is the proportion of the total phenotypic variance due to additive effects, defined as:

$$h^2 = \frac{\sigma_A^2}{\sigma_Y^2}$$

The **broad-sense heritability** H^2 is the proportion of the total phenotypic variance due to all genetic effects (additive and deviations from additivity), defined as:

$$H^2 = \frac{\sigma_G^2}{\sigma_Y^2}$$

1.1.3 Trait association testing

In the past two decades, the scientific community has been focused on identifying loci that are associated with specific phenotypes, including diseases. A **genome-wide association study (GWAS)** is a research approach that screens the genome for as-

sociations between millions of genetic variants and a complex trait. In 2002, the first GWAS successfully identified a susceptibility gene for myocardial infarction[17]. Nearly twenty years later, more than 55,000 unique loci associated with approximately 5,000 diseases and traits have been reported by more than 4,500 GWASs[18].

When the phenotype is a case-control trait, **allelic association tests** allow us to test for independence between the trait and each allele. Alleles are treated as the sampling units under the Hardy–Weinberg principle assumption (allele frequencies remain constant from generation to generation in the absence of disturbing factors). The classical Pearson's χ^2 test is often used for allelic association testing, where the null hypothesis is that there is no association between the genetic marker and the trait, and the alternative hypothesis is that there is an association between the two.

On the other hand, when the trait is quantitative, most GWASs perform (in their simplest form) a univariate linear regression per SNP. It is common to restrict association tests to common genetic markers (frequency $\geq 1\%$) with two allelic types (**biallelic** markers). This approach is called **SNP association testing** with linear regression assuming an additive model. It can also be extended to binary phenotypes by simply considering them as continuous random variables. For a given individual n, let X_{nm} denote their genotype at a fixed SNP m and Y_n their phenotype. Both variables can be normalised, such that the mean and the variance across all samples are 0 and 1

respectively. Let $\tilde{X}_{nm}$ and $\tilde{Y}_n$ be the normalised variables. The model is thus given by:

$$\tilde{Y_n} = \beta_m \tilde{X}_{nm} + \epsilon_{nm} \quad \text{where} \quad \epsilon_{nm} \overset{\text{iid}}{\sim} N(0, \sigma_m^2) \tag{1.1}$$

The maximum likelihood estimate (MLE) of β_m is called the **effect size** of SNP m, denoted as $\hat{\beta}_m$. We estimate its standard error by $\hat{se}(\hat{\beta}_m)$ and compute the z-score z_m for hypothesis testing as follows:

$$z_m = \frac{\hat{\beta}_m}{\hat{se}(\hat{\beta}_m)}$$

The SNP association testing is then performed by carrying out a one-sided test of z_m^2 against a χ_1^2 distribution, where the null hypothesis is that there is no association between the phenotype and SNP m, while the alternative is that there is association between the two.

It is important to note that, generally, causal variants will not be genotyped in the SNP array used for obtaining the data set. **Linkage Disequilibrium (LD)** refers to correlations between genotypes of nearby markers. In GWAS, genetic markers that tag the actual causal variants are sometimes detected instead (we say that they are in LD with the causal variants). **Fine-mapping** seeks to determine the exact causal genetic variant(s) in a trait-associated region from a GWAS.

For most traits, the heritability due to additive effects of statistically significant SNPs in GWAS does not recover all of the estimated heritability ($h^2_{GWAS} < h^2$). This problem is referred to as **missing heritability**. A few leading possible explanations for this observation include the following: complex traits could be affected by many rare variants with possibly larger effects; GWAS is potentially not sufficiently powerful to detect the many common variants with extremely weak effects; or heritability estimates could have been overestimated in the first place.

1.1.4 Population structure

Additional variables (or **covariates**) can be added to the regression model in (1.1) in order to account for confounders and decrease the amount of false positives (i.e. non-associated variants incorrectly identified as associated). Covariates usually include age, sex or **population structure** (i.e. genetic differences between populations due to non-random mating).

Population stratification refers to ancestry differences between cases and controls in a genetic association study. **Principal component analysis (PCA)** is the most widely used approach for identifying and correcting for population stratification in genome-wide association studies[19,20]. PCA is a linear dimensionality reduction algorithm. When applied to genetic data, the goal is to reduce high dimensional genotypes

by calculating **principal components (PCs)** that best explain differences among individuals. The top PCs correspond to continuous axes of variation that reflect genetic variation due to ancestry and they can be added to (1.1) as covariates. Individuals with similar values in the top PCs typically originate from similar ancestry.

The **LD score** of some variant j over M total variants is defined as:

$$l_j = \sum_{k=1}^{M} r_{jk}^2$$

where r_{jk}^2 is the correlation between variants j and k. It was shown by Bulik-Sullivan et al.[21] that regressing the LD score of SNPs on their χ^2-GWAS association statistic could help distinguishing between true polygenic association signal and confounding, such as population stratification. This approach is called **LD score regression (LDSC)** and provides a relationship between GWAS summary statistics and LD:

$$\mathbb{E}[\chi_j^2 | l_j] = N\frac{h_{SNP}^2}{M} l_j + Na + 1 \tag{1.2}$$

where χ_j^2 is the χ^2-statistic of variant j, N is the sample size, $\frac{h_{SNP}^2}{M}$ represents the average heritability explained per SNP and a measures the contribution of confounding biases. The intuition behind this model is the following: the more genetic markers a variant tags (i.e. the higher the LD score), the more likely this variant will tag a causal variant. On

the other hand, the impact of **cryptic relatedness** (i.e. unknown genetic relatedness between samples) or population stratification on the GWAS summary statistics does not correlate with LD score. Consequently, the intercept in (1.2) is an estimator of the mean contribution of confounding bias to the inflation in the test statistics.

1.1.5 Partitioning heritability

Although LDSC was originally developed as a tool to distinguish confounding from polygenicity in GWAS, the slope of the regression in (1.2) can be used to estimate genome-wide SNP-heritability h^2_{SNP}[21]. LDSC offers an alternative method for estimating heritability without requiring individual genotypes, as opposed to linear mixed models[22,23,24], by using only GWAS statistics and a reference LD panel.

Further work has shown that the contribution to heritability for complex traits is not uniform across the genome, and that regions with certain functional annotations contribute disproportionately to the overall heritability[25,26]. Finucane et al.[27] extended the LDSC model to **stratified LD score regression (S-LDSC)** by partitioning SNPs into binary functional categories with different contributions to heritability. By doing so, for a given SNP, LD to a category that is enriched for heritability will increase the χ^2-statistic more than LD to a category that does not contribute to heritability.

Re-using notations from (1.2), the S-LDSC model is defined as follows[27]:

$$\mathbb{E}[\chi_j^2|l_j] = N\sum_{c\in C}\tau_c l_j(c) + Na + 1 \tag{1.3}$$

where C indexes categories, $l_j(c)$ is the LD score of SNP j with respect to category c (defined as $l_j(c) = \sum_{k\in c} r_{jk}^2$) and τ_c corresponds to the per-SNP heritability of category c (also called effect size).

Gazal et al.[28] extended S-LDSC to continuous-valued annotations by redefining the LD score as $l_j(c) = \sum_{k\in C} a_c(k) r_{jk}^2$, where $a_c(k)$ is the continuous value of annotation c at SNP k. The per-standardized-annotation effect sizes τ^* is defined as the proportionate change in per-SNP heritability associated with a 1 standard deviation increase in the value of the annotation and can be compared across annotations and across traits. τ^* is obtained by standardizing the estimated effect sizes $\hat{\tau}$.

1.2 Population genetics

1.2.1 Coalescent theory

A widely used model in population genetics describing the genealogical relationship among alleles, called the **Wright-Fisher model**[29,30], was introduced by Sewall Wright (1931) and Ronald A. Fisher (1930). It describes the sampling of alleles in a popula-

tion under multiple assumptions: discrete and non-overlapping generations, constant population size, equally fit individuals, no geographical or social structure and no recombination. For a diploid population of constant size N (which can be described by the haploid model with size $2N$ haplotypes), reproduction under the Wright-Fisher model works as follows: first randomly select a gamete from one haplotype, and then randomly select another gamete from a second haplotype (which may be from the same diploid sample) to make one offspring. This process is repeated N times, until an offspring population of $2N$ haplotypes has been generated. Some lineages do not leave any offspring and disappear over time, while others leave many descendants. If we trace the history of all present haplotypes back in time, we will eventually find an ancestor from which all alleles are descended, called the **most recent common ancestor (MRCA)**.

Since many lineages eventually die out and leave no trace on the alleles in the present day, we would like to avoid including them in the model. Indeed, we can only consider the events that are ancestral to the samples in the population. The genealogical process underlying the history of a given locus from samples in a population is called the **coalescent** and was first introduced in 1982 by Kingman[31]. The coalescent traces lineages backwards in time for a given sample and a given locus. When the lineages of two samples meet, they are said to have coalesced. The distance, in generations, that separates these individuals and their most recent common ancestor at a specific genomic

location is referred to as **time to most recent common ancestor (TMRCA)**, or **coalescence time**[31,32]. In order to gain some intuition, consider the simplest case of two samples and no recombination, population structure, or other complicating factors. Let N_e be the size of an idealised population (unit is haploid samples) such that the rate of inbreeding and the rate of gene frequencies variation coincide with the ones actually observed in the population, also called the **effective population size**[33]. The probability that two alleles came from the same parental gamete in the previous generation is $\frac{1}{N_e}$, and the probability that they came from different parental gametes is $1 - \frac{1}{N_e}$. Hence, the pairwise coalescence time follows a geometric distribution with parameter $\frac{1}{N_e}$ (the average coalescence time is given by $\mathbb{E}[t] = N_e$).

The coalescent usually incorporates mutations under the **infinite-sites model**, proposed by Kimura[34] in 1969. The model assumes an infinite number of sites where mutations can occur so that sites are at most biallelic. This assumption is justified by the fact that the total number of mutated sites in human DNA is extremely small compared to the total number of sites (any two human beings are 99.4% identical[13]). Each genomic site in the sequence that is passed on between two generations is subject to a mutation event with probability μ, called the **mutation rate**. For two haplotypes with TMRCA t on a chromosomal region of length n sites, the number of heterozygous sites (i.e. sites with different allele versions) follows a Binomial distribution with parameters

$(n, 2t\mu)$ since mutations can occur on both lineages independently. Note that it is very common to find the scaled mutation rate $\theta = 4N_e\mu$ in the literature.

Since the original model was published, coalescent theory has been largely developed and extended to allow for increasingly realistic representations of genealogical inheritance[35]. In 1983, Hudson developed an algorithm to simulate the coalescent with recombination under the infinite-site model, called the **coalescent with recombination process**[32]. The model allows crossing-over to happen at any of an infinite number of sites distributed along the simulated genomic region. Each genomic site that is passed on between two generations is subject to a recombination event with probability r, called the **recombination rate**. Similar to the mutation rate, the scaled recombination rate is defined as $\rho = 4N_e r$. Shortly after, Griffiths and Marjoram[36] embedded the model within **ancestral recombination graph (ARG)**, a three-dimensional object capturing coalescence times, recombination events and mutations.

Probabilistic modeling of several quantities of interest under the coalescent is computationally challenging. In 2005, McVean and Cardin introduced a simplification of the standard coalescent process called the **sequentially Markov coalescent (SMC)** that makes inference more tractable by only allowing coalescence events for lineages with overlapping ancestral material[37]. The SMC has a Markovian structure which assumes that each marginal tree only depends on its immediate predecessor as we move

along the genome. In 2006, Marjoram and Wall[38] slightly modified and improved the approximation of the SMC model.

In 2002, Hudson[39] published `ms`, the first simulation program based on the coalescent with recombination. Several coalescent simulators and extensions have been developed since then, including Wiuf and Hein's algorithm[40], `msms`[41], and `cosi2`[42]. In 2016, Kelleher et al.[43] developed an efficient simulator for the coalescent with recombination process based on the ARG representation, called `msprime`, resulting in significant improvements in simulation speed and storage efficiency.

1.2.2 Coalescent Hidden Markov Models

Accurate prediction of pairwise coalescence times is a key element in several genomic analyses, such as the reconstruction of a population's demographic history[44]. Many methods that infer pairwise coalescence times rely on a **hidden Markov model (HMM)**[45,46,47]. A HMM is an inferential procedure based on dynamic programming in which we assume the system to be governed by a Markov process, defined by $\{S, \Pi, T, E\}$ where:

- S is the set of all possible (hidden) s states;
- Π is a vector of s elements, where π_i is the probability for S_i to be the initial state;

- T is a $s \times s$ transition matrix, where T_{ij} is the transition probability from S_i to S_j;

- E is a $s \times e$ emission matrix, where E_{ij} is the probability to emit the j^{th} observation while being in the hidden state S_i (where e is the total number of possible observations).

with: $\sum_{i=1}^{s} \pi_i = 1$ $\sum_{j=1}^{s} T_{ij} = 1$ $\sum_{j=1}^{e} E_{ij} = 1$

HMMs allow us to solve three main problems: evaluation (estimating the probability of a sequence of observations for a given model), decoding (determining the most likely hidden states that generated a given sequence of observations), and learning (adjusting the model's hyperparameters to maximize the likelihood of the observations)[48]. Different inference algorithms may be used for decoding. The forward-backward algorithm computes the most likely marginal state at every position along the sequence, while the Viterbi algorithm is used to find the most probable overall sequence of states.

In 2007, Hobolth et al.[49] developed one of the first coalescent HMMs, called CoalHMM, that infers speciation times of Human, Chimpanzee and Gorilla. More recent work focused on population genetic time scales, such as the **pairwise coalescent Markovian coalescent (PSMC)**[44]. For a haplotype of length L, observations x_i, $i \in \{1, ..., L\}$, have value 1 if the two haplotypes are heterozygous or value 0 if they are homozygous at the site. Hidden states $t_i \in S$, $i \in \{1, ..., L\}$ correspond to the TMRCA between the

two haplotypes at every genomic site. Time is measured in generations and needs to be discretised for computational tractability such that S is a set of possible time intervals. Initial state probabilities are computed from the coalescent distribution under a probabilistic prior distribution based on a specific demographic history. Transition probabilities between consecutive hidden states are calculated from the SMC model. Given a mutation rate μ, the probability of observing a heterozygous site is given by $P(x_i = 1|t, \mu) = 1 - e^{-2t\mu}$, and the probability of observing a homozygous site is given by $P(x_i = 0|t, \mu) = e^{-2t\mu}$. The PSMC model can be used to estimate locus-specific coalescence times for pairs of chromosomes (HMM decoding problem), or to learn hyperparameters such as effective population size, mutation and recombination rates (HMM learning problem). However, the PSMC model suffer from two important caveats: first, it can only accurately handle WGS data; second, the standard computation of posterior probabilities via the forward-backward algorithm has cost $\mathcal{O}(d^2 l)$ for d hidden states and an observation sequence of length l[48]. In 2018, Palamara et al.[50] addressed these issues by developing the **ascertained sequentially Markovian coalescent (ASMC)** model that can efficiently estimate the coalescence time of two haplotypes at each position in the genome using SNP array data (with cost $\mathcal{O}(dl_p)$, where $l_p \ll l$ is the number of loci that are polymorphic in the analyzed samples).

1.2.3 Identity By Descent

The genomes of two individuals from a population are connected through genealogical relationships that lead to common ancestors. Since the expected TMRCA for a randomly selected genomic site and pair of haplotypes is in the order of several thousands of generations, individuals may not be aware of their genetic relationships, being therefore purportedly unrelated. Two homologous DNA sequences (from the same or different individual(s)) are **identical by descent (IBD)** if one of them is descended from the other or if both are descended from a third ancestral segment. "Descent" refers to DNA replication and transmission from parent to offspring, while "identical" refers to both sequences being almost identical since copied from the transmitting common ancestor, although a small number of mutations and other rare genomic events may occur on each segment during the transmission process. It is common to use a time-based definition of IBD segments when inferring them (although other definitions have been proposed[51]): for a given time threshold t (in generations), two genomic sites are considered to be IBD if their TMRCA is lower than t. IBD segments may be used to characterize fine-scale genetic relationships between individuals and are informative for several genomic analyses. Because there is a chance that the recombination process breaks haplotypes during meiosis in every generation, long IBD sharing is usually a sign of recent genetic relatedness across individuals, while short co-inherited segments

are likely to reflect deep ancestry.

One of the main challenges in IBD detection is to distinguish IBD segments from **identical by state (IBS) segments**, which are haplotypes with identical nucleotide sequences. If a time-based IBD definition is used, and a low time threshold is selected, IBD segments are also expected to be IBS, with the exception of de-novo mutations and other variants introduced by e.g. non-crossover gene conversion events in the generations leading to their recent common ancestor[52]. In addition, individuals can share IBS regions that are not IBD, which are inherited from common ancestors that are older than the time threshold used in the time-based definition of IBD segments, but where no observed mutations occurred. Detection of IBD segments in millions of individuals within modern biobanks poses a number of computational challenges. Several IBD detection methods have been published and can be classified in two types: model-based methods rely on complex probabilistic computations[53], whereas string-matching methods solely depend on genotypic similarity to detect shared ancestry[54,55,56].

1.2.4 Natural selection

Some regions in the human genome are known to have a direct effect on phenotypic variability, while others may be considered to be **neutral**. **Natural selection** is the differential survival and reproduction of individuals due to differences in phenotype.

Mutations that increase the **fitness** of individuals in their environment might increase in frequency because these individuals have higher survival and reproduction probabilities, a process called **positive selection**. On the other hand, deleterious variants tend to disappear over time in a population due to **negative selection**. Negative selection can result in the loss of genetic diversity at non-deleterious loci due to LD, referred to as **background selection**. Finally, **balancing selection** corresponds to the set of adaptive forces by which multiple alleles are maintained in the population at frequencies larger than expected by chance.

Signals of natural selection have been identified in the human genome and have helped us shape our understanding of the genetic basis of disease[57]. It has been shown that natural selection distorts local genealogies in distinct ways[58]. These distortions can be identified by estimating the pairwise TMRCA distributions at a given locus without necessarily targeting a specific type of selection[59]. Natural selection influences the effective population size along the genome, which in turn, impacts the average coalescence time[60]. Leveraging this principle, locus specific TMRCA estimates have been used to detect signatures of positive and background selection[50].

1.3 Deep learning

1.3.1 Basic definitions

The study of algorithms that learn and improve automatically through experience (data) is called **machine learning (ML)**[61]. **Supervised learning** is the machine learning task that consists of learning a function that maps an input to an output (or label) $f : X \to Y$ based on observed input-output pairs. The most common supervised learning tasks are **regression** (where $Y \subseteq \mathbb{R}^n$) and **classification** (where Y is a set of discrete indices). On the other hand, **unsupervised learning** is the machine learning task of learning underlying features (or patterns) from observed unlabeled data. Unsupervised learning tasks include **clustering** (where the objective is to assign every data point to a group) and anomaly detection (where the aim is to identify outliers in a data set). Machine learning models are often fitted on a **training data set**, then evaluated on a **validation data set** while tuning hyperparameters, and finally applied to a **test data set** to obtain an unbiased evaluation of the final model.

In classical machine learning, the data is first pre-processed to manually extract relevant features, which requires both careful engineering and domain expertise. **Representation learning** is a set of methods that allows an algorithm to be fed with raw data and to automatically discover the suitable internal representation. **Deep learning**

(DL) methods are representation learning techniques where multiple non-linear modules (or **hidden layers**) are used to extract progressively higher level features from data. Each layer consists of multiple hidden units (or **neurons**) and builds upon the previous one to form a complex computational graph of interconnected nodes. Non-linear functions, also called **activation functions**, are applied to the output of each of these nodes. The rectified linear unit (ReLU), defined as $f(x) = \max(x, 0)$, is currently the most commonly used activation function. In the past, neural networks used smoother non-linearities, such as the sigmoid function ($f(x) = \frac{1}{1+e^{-x}}$) or the hyperbolic tangent function ($f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$). By combining enough such non-linear transformations, extremely complex functions can be learned. These functions are referred to as **(deep) neural networks**[62].

Feedforward neural networks (FNN) are neural networks where connections between nodes in the computational graph do not form a cycle, as opposed to **recurrent neural networks (RNN)**. In 1989, Hornik et al.[63] showed that for any continuous function f on a compact set K, there exists a feedforward neural network, having only a single hidden layer, which uniformly approximates f on K to an arbitrary precision. At the same time, Cybenko[64] independently reached the same conclusion when using sigmoid functions as activation functions. More formally, Cybenko[64] demonstrated that the set of feedforward neural networks was dense in the space of continuous functions in

$[0,1]^n$ with respect to the uniform norm, where n refers to the output dimension and the uniform norm for $f : X \to Y$ is defined as $\|f\| = \sup_{x \in X} |f(x)|$. We typically refer to these results as the **universal approximation theorem for neural networks**. Note that the theorem does not provide a way of constructing the neural network parameters, but only states that such a construction is possible. Moreover, there is no limitation imposed on the number of hidden units required for the approximation to hold.

Parameters (or **weights**) in neural networks are trained by a procedure called **gradient descent**[65,66] that aims to minimise a given loss function $\mathcal{L}$, which is a proxy for the distance between the current and the expected (ground truth) output. Every weight w is updated by iteratively moving in the opposite direction of the first derivative (or gradient) of $\mathcal{L}$ with respect to w as follows:

$$w \leftarrow w - \alpha \frac{\partial \mathcal{L}}{\partial w}$$

where $\alpha > 0$ is the **learning rate**. When the true gradient is approximated by the gradient of $\mathcal{L}$ at a single input observation, the procedure is referred to as **stochastic gradient descent (SGD)** and weights are updated after each observation[67]. On the other hand, **batch gradient descent** consists of evaluating $\mathcal{L}$ on the whole training data set, and the true gradient is approximated using every single input data simultaneously. In this case, weights are updated after processing all the inputs. However,

looping through all of the training data for every update is computationally inefficient. In practice, neural networks process a subset of the entire training data set at a time in **mini-batches**. The loss function is computed across the whole input batch and weights are updated after each observed batch of data, a procedure called **mini-batch gradient descent**.

In 1964, Polyak[68] extended stochastic gradient descent by introducing **gradient descent with momentum**. The idea is to compute an exponentially weighted average of the gradient and use it to update the weights instead. As previously done, the gradient $\frac{\partial \mathcal{L}}{\partial w}$ is first computed (on a single data point, a batch or the whole data set), and the moving average $V_{\partial w}$ is updated as follows:

$$\begin{cases} V_{\partial w} \leftarrow \beta V_{\partial w} + (1-\beta)\frac{\partial \mathcal{L}}{\partial w} \\ \\ w \leftarrow w - \alpha V_{\partial w} \end{cases}$$

where $0 \leq \beta \leq 1$. This strategy was shown to result in faster convergence by building inertia in a specific direction in the parameter search space and by averaging out to 0 the oscillations of noisy gradients. Using a similar approach, Hinton et al.[69] developed the **Root Mean Squared Propagation (RMSprop)** method that computes a moving

average of the squared gradient $S_{\partial w}$ with the following update rules:

$$\begin{cases} S_{\partial w} \leftarrow \beta S_{\partial w} + (1-\beta)(\frac{\partial \mathcal{L}}{\partial w})^2 \\ w \leftarrow w - \frac{\alpha}{\sqrt{S_{\partial w}}} \frac{\partial \mathcal{L}}{\partial w} \end{cases}$$

In 2014, Kingma and Ba[70] combined both gradient descent with momentum and RMSprop in the **adaptive moment estimation (Adam)** optimisation algorithm. The neural network parameters are updated as follows:

$$w \leftarrow w - \frac{\alpha}{\sqrt{S_{\partial w}}} V_{\partial w}$$

At present, Adam is the most popular optimiser for gradient descent, although many other algorithms and extensions have been developed[71,72,73,74,75].

In order to compute gradients, deep learning uses **automatic differentiation**[76] which decomposes the very complex functions learned by neural networks as sequences of elementary functions and arithmetic operations. Automatic differentiation makes use of the **chain rule** which, for $h = f \circ g$, is expressed as $h' = (f' \circ g) \cdot g'$. By applying the chain rule repeatedly to these elementary operations, derivatives can be easily computed. Deep neural networks are trained by leveraging the **backpropagation algorithm**[77], which is a standard automatic differentiation procedure that computes

the gradient of a loss function with respect to the weights of hidden layers to optimise the model parameters through gradient descent. The backpropagation algorithm is made of two steps. For a given input data point, a forward pass is first performed by feeding the input to the neural network to obtain the error term (or loss). The gradient of the loss with respect to some parameter in a given layer can then be computed by working backwards from the derivative of the loss with respect to the output of the preceding module. By repeatedly applying this strategy for all parameters, the algorithm propagates gradients backwards through all layers, starting from the last one (where the error term is computed) and all the way to the first one (to which the input data is fed).

The major performance gains offered by deep learning in many applications have been enabled by both the explosion of data and the increase in computational capacity. Indeed, one of the main revolutionary steps for deep learning happened in 1999 when NVIDIA launched the first commercial **Graphics Process Unit (GPU)** GeForce 256. Although GPUs were originally developed for accelerating graphics processing, they dramatically speed up computational processes for deep learning due to their ability of parallel processing. In 2007, NVIDIA released the Compute Unified Device Architecture (CUDA) framework for software programmers building GPU-accelerated applications[78].

1.3.2 Architectures

Deep neural networks use many hidden layers, or modules. There exists a variety of layers. A layer is said to be **fully connected (FC)** when each hidden unit receives inputs from all hidden units of the preceding layer. A **convolutional layer** is a special form of FC layer in which the same FC layer is applied locally, a property referred to as local connectivity. By sharing the same parameters across positions, the total number of parameters is drastically reduced, making the training process much faster. Convolutional layers are designed to process data in the form of multiple arrays (e.g. audio sequences, text, images or videos). They scan the input data with several filters (or kernels) and produce a scalar value at every position, which reflects the match between the filter and the local input information. Neural networks that make use of convolutional layers are called **Convolutional Neural Networks (CNNs)**.

In 1989, LeCun et al.[79] combined CNNs with backpropagation for the first time and developed LeNet-5, a 7-layer CNN that classifies 32x32 pixel greyscale images of hand-written digits (MNIST data set[80]). Since then, CNNs have achieved state-of-the-art results in many fundamental tasks in computer vision. The ImageNet project is a large database containing more than 14 million images that have been hand-annotated for object recognition[81]. In 2012, Krizhevsky et al.[82] entered the annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) and won the competition by reducing

the top-5 error from 26% to 15.3% with a CNN called AlexNet. Shortly after that, Szegedy et al.[83] implemented GoogLeNet, a 22-layer deep CNN, and achieved a top-5 error of 6.67% by using **batch normalisation**[84], a process that makes training more stable by re-centering and re-scaling the layers' inputs. In 2016, He et al.[85] introduced a novel architecture called the **residual neural network (ResNet)** with "skip connections" that are used to jump over some layers in the computational graph. Residual connections allow gradients to propagate in deep neural networks without getting too small, preventing the so-called **vanishing gradient problem**[86]. By leveraging this new architecture, they efficiently trained a neural network with 152 layers. ResNet achieved a top-5 error of 3.57% and finally outperformed human-level performance on the ImageNet data set, estimated to be around 5.1%[87]. Recent CNN architectures use the ReLu activation function, have hundreds of layers, hundreds of millions of weights and usually only require a few hours to train. CNNs have successfully been applied to object recognition (e.g. object detection[88,89,90,91] or segmentation[92]). Major recent practical successes of CNNs also include cancer detection[93], X-ray image analysis[94], face recognition[95], self-driving cars[96,97,98] or speech recognition[99].

In the context of deep learning, the **receptive field** of a layer is defined as the size of the region in the layer's input that produces the layer's output. The receptive field of a convolutional layer is intrinsically limited by the size of its kernel[100]. **Dilated**

convolution[101] is a technique that expands the receptive field by making the kernel skip consecutive elements. In order to capture long-range dependencies with CNNs, one would still need to stack many hidden layers. An alternative attention-based architecture, called the **Transformer**[102], has recently been developed for natural language processing (NLP) applications. The attention mechanism allows to selectively focus on valuable parts of the input data (i.e. to pay more attention to them) while diminishing other parts. For a given token i in the input sequence, the **self-attention layer** proceeds as follows: for every other token j, a non-negative weight w_{ij} is computed by multiplying a **query vector** q_i and a **key vector** k_j, such that $\sum_j w_{ij} = 1$, and the output is finally obtained by computing a weighted average of the **value vectors** v_j, defined as $\sum_j w_{ij} \cdot v_j$. The query, key and value vectors are functions of the input data and generated using learned parameters. All input tokens are processed simultaneously and calculations are done in matrix form for faster processing by introducing the query, key and value matrices (Q, K and V). The matrix product QK^T is called the **attention matrix** and contains all the attention weights of a given input sequence. By definition, a Transformer is invariant with respect to reordering of the input and a positional encoding of the input sequence would be necessary for specific tasks where the ordering matters (e.g. NLP). In this case, the positional information can be included in the raw data directly[102] or added to the query and key vectors[103]. By capturing

long-range dependencies, attention-based models achieve state-of-the-art performance in NLP applications[104,105,106] (e.g. text generation, translation or question answering), and are now becoming an efficient tool in computer vision[107,108].

1.3.3 Explainable deep neural networks

Due to the over-parameterized black-box nature of deep neural networks, it is difficult to understand their predictions. In recent years, efforts have been made to produce explainable machine learning methods and enhance data insights. **Explainable Artificial Intelligence (XAI)** or **explainable Machine Learning (XML)** focuses on understanding and interpreting predictions made by ML algorithms[109].

Most methods used to interpret neural networks consist of highlighting specific regions or features that strongly influence the output, referred to as visualisation methods[110,111,112]. The most common way to achieve so in computer vision tasks is to compute the gradient of the prediction with respect to the input data and to visualise values on a heatmap, also called **saliency map** or **attribution map**. A standard approach is to multiply the gradient with the input itself, a method named "gradient*input", so that the attribution score for a given pixel is high if and only if (1) the region of the image is relevant to the prediction and (2) the pixel value along a specific channel is high[113]. This method can be used for **feature importance** which identifies the most

relevant features in the data. Another related technique, called **input perturbation**, consists of manually perturbing the input data (in a chosen direction, as opposed to gradient based methods) and measure the prediction change. When using attention-based models (e.g. in NLP tasks), the attention matrix can directly be visualised to identify relevant context.

Visualisation can also be done by **optimisation**[114]. Since neural networks are differentiable with respect to their inputs, it is possible to maximise a specific prediction by starting from a random input and by iteratively applying gradient descent on it directly. This allows to find out what kind of input data would cause a certain prediction. Other interpretability methods include ablation studies that remove certain components of the input data or the neural network with the aim to better understand their individual contribution.

1.3.4 Recent advances for population genetics

Deep learning algorithms have transformed fields such as computer vision[115,85,92] and natural language processing[102,104,105], and they are now emerging as an effective tool in several genomic applications[116,117]. With the advancement of genomic research and the availability of large-scale data sets, genomics is increasingly becoming data-driven, and therefore amenable to machine learning[118]. Some preliminary work has demonstrated

the use of convolutional neural networks to discover genomic sequence motifs[119,120,121]. Applications of neural networks in genomics include predicting functional effects of non-coding variants[119,122,123,124], basecalling of nanopore data[125], identifying the sequence specificities of DNA- and RNA-binding proteins[120], inferring demographic history and population structure[126,127,128], and generating synthetic data[129,130,131]. Recent studies have investigated new ways of extracting biological meaning from long-range dependencies in DNA sequences with transformers[132,133,134,135].

Processing multiple genome sequences simultaneously with deep learning requires architectures that are invariant to sample permutations. **Exchangeable neural networks**, also called permutation invariant neural networks, have recently been developed to allow for exchangeable input data[136,137,138]. However, solving reference panel based problems (e.g. phasing or imputation) with neural networks still remains a challenging task.

Chapter 1 - References

[1] Ralf Dahm. Discovering DNA: Friedrich miescher and the early years of nucleic acid research. *Human genetics*, 122(6):565–581, 2008.

[2] Katherine Brind'Amour and Benjamin Garcia. Wilhelm August Oscar Hertwig (1849-1922). *Embryo Project Encyclopedia*, 2012.

[3] James D Watson and Francis HC Crick. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[4] Meena Kishore Sakharkar, Vincent TK Chow, and Pandjassarame Kangueane. Distributions of exons and introns in the human genome. *In silico biology*, 4(4): 387–393, 2004.

[5] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[6] Larry H Thompson and David Schild. Recombinational DNA repair and human disease. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 509(1-2):49–78, 2002.

[7] All of Us Research Program Investigators. The "All of Us" research program. *New England Journal of Medicine*, 381(7):668–676, 2019. doi: https://doi.org/10.1056/NEJMsr1809937.

[8] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018. doi: https://doi.org/10.1038/s41586-018-0579-z.

[9] Vivien Marx. The DNA of a nation. *Nature*, 524:503–505, 2015. doi: https://doi.org/10.1038/524503a.

[10] Wetterstrand KA. DNA sequencing costs: Data from the nhgri genome sequencing program (GSP). Available at `https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data`. Accessed in 2022.

[11] Feng Zhang and James R Lupski. Non-coding genetic variants in human disease. *Human molecular genetics*, 24(R1):R102–R110, 2015.

[12] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443–1448, 2016.

[13] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[14] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018.

[15] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS genetics*, 16(11): e1009049, 2020.

[16] Peter M Visscher, William G Hill, and Naomi R Wray. Heritability in the genomics era—concepts and misconceptions. *Nature reviews genetics*, 9(4):255–266, 2008.

[17] Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650–654, 2002.

[18] Ruth JF Loos. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications*, 11(1):1–3, 2020.

[19] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[20] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[21] Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. Ld score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

[22] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[23] Noah Zaitlen and Peter Kraft. Heritability in the genome-wide association era. *Human genetics*, 131(10):1655–1664, 2012.

[24] David Heckerman, Deepti Gurdasani, Carl Kadie, Cristina Pomilla, Tommy Carstensen, Hilary Martin, Kenneth Ekoru, Rebecca N Nsubuga, Gerald Ssenyomo, Anatoli Kamali, et al. Linear mixed model for heritability estimation that explicitly addresses environmental variation. *Proceedings of the National Academy of Sciences*, 113(27):7377–7382, 2016.

[25] Lea K Davis, Dongmei Yu, Clare L Keenan, Eric R Gamazon, Anuar I Konkashbaev, Eske M Derks, Benjamin M Neale, Jian Yang, S Hong Lee, Patrick Evans, et al. Partitioning the heritability of tourette syndrome and obsessive compulsive disorder reveals differences in genetic architecture. *PLoS genetics*, 9(10):e1003864, 2013.

[26] Alexander Gusev, S Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, Brendan Bulik-Sullivan, Eli Stahl, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *The American Journal of Human Genetics*, 95(5): 535–552, 2014.

[27] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015.

[28] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10): 1421–1427, 2017.

[29] Ronald Aylmer Fisher. *The genetical theory of natural selection*. Рипол Классик, 1958.

[30] Nicholas H Barton. Sewall wright on evolution in mendelian populations and the "shifting balance". *Genetics*, 202(1):3–4, 2016.

[31] John FC Kingman. On the genealogy of large populations. *Journal of applied probability*, 19(A):27–43, 1982.

[32] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[33] James Franklin Crow. *An introduction to population genetics theory*. Scientific Publishers, 2017.

[34] Motoo Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893, 1969.

[35] John FC Kingman. Origins of the coalescent: 1974-1982. *Genetics*, 156(4):1461–1463, 2000.

[36] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

[37] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[38] Paul Marjoram and Jeff D Wall. Fast "coalescent" simulation. *BMC genetics*, 7 (1):1–9, 2006.

[39] Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[40] Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical population biology*, 55(3):248–259, 1999.

[41] Gregory Ewing and Joachim Hermisson. Msms: a coalescent simulation program including recombination, demographic structure and selection at a single locus. *Bioinformatics*, 26(16):2064–2065, 2010.

[42] Ilya Shlyakhter, Pardis C Sabeti, and Stephen F Schaffner. Cosi2: an efficient simulator of exact and approximate coalescent with selection. *Bioinformatics*, 30 (23):3427–3429, 2014.

[43] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[44] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493, 2011.

[45] Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[46] Leonard E Baum and John Alonzo Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73(3):360–363, 1967.

[47] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[48] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[49] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS genetics*, 3(2):e7, 2007.

[50] Pier Francesco Palamara, Jonathan Terhorst, Yun S Song, and Alkes L Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50(9):1311–1317, 2018.

[51] John Wakeley and Peter R Wilton. Coalescent and models of identity by descent. 2016.

[52] Pier Francesco Palamara, Laurent C Francioli, Peter R Wilton, Giulio Genovese, Alexander Gusev, Hilary K Finucane, Sriram Sankararaman, Shamil R Sunyaev, Paul IW de Bakker, John Wakeley, et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *The American Journal of Human Genetics*, 97(6):775–789, 2015.

[53] Brian L Browning and Sharon R Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[54] Alexander Gusev, Jennifer K Lowe, Markus Stoffel, Mark J Daly, David Altshuler, Jan L Breslow, Jeffrey M Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome research*, 19(2):318–326, 2009.

[55] Ardalan Naseri, Xiaoming Liu, Kecong Tang, Shaojie Zhang, and Degui Zhi. Rapid: ultra-fast, powerful, and accurate detection of segments identical by descent (ibd) in biobank-scale cohorts. *Genome biology*, 20(1):1–15, 2019.

[56] Ruhollah Shemirani, Gillian M Belbin, Christy L Avery, Eimear E Kenny, Christopher R Gignoux, and José Luis Ambite. Rapid detection of identity-by-descent tracts for mega-scale datasets. *Nature communications*, 12(1):1–13, 2021.

[57] Rasmus Nielsen, Ines Hellmann, Melissa Hubisz, Carlos Bustamante, and Andrew G Clark. Recent and ongoing selection in the human genome. *Nature Reviews Genetics*, 8(11):857–868, 2007.

[58] Michael Bamshad and Stephen P Wooding. Signatures of natural selection in the human genome. *Nature Reviews Genetics*, 4(2):99–110, 2003.

[59] Haley Hunter-Zinck and Andrew G Clark. Aberrant time to most recent common ancestor as a signature of natural selection. *Molecular biology and evolution*, 32 (10):2784–2797, 2015.

[60] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics*, 5 (5):e1000471, 2009.

[61] Tom M Mitchell et al. Machine learning, 1997.

[62] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521 (7553):436–444, 2015.

[63] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[64] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[65] Claude Lemaréchal. Cauchy and the gradient method. *Doc Math Extra*, 251(254): 10, 2012.

[66] Jacques Hadamard. *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, volume 33. Imprimerie nationale, 1908.

[67] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[68] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[69] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8):2, 2012.

[70] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[71] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[72] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[73] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[74] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[75] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[76] Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

[77] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[78] NVIDIA, Péter Vingelmann, and Frank H.P. Fitzek. Cuda, release: 10.2.89, 2020. URL `https://developer.nvidia.com/cuda-toolkit`.

[79] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[80] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[81] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[82] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[83] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[84] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[85] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[86] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[87] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[88] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[89] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[90] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[91] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[92] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[93] Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):1–12, 2019.

[94] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12, 2020.

[95] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

[96] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[97] Jeffrey Hawke, Richard Shen, Corina Gurau, Siddharth Sharma, Daniele Reda, Nikolay Nikolov, Przemysław Mazur, Sean Micklethwaite, Nicolas Griffiths, Amar Shah, et al. Urban driving with conditional imitation learning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 251–257. IEEE, 2020.

[98] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15273–15282, 2021.

[99] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *SSW*, 125:2, 2016.

[100] André Araujo, Wade Norris, and Jack Sim. Computing receptive fields of convolutional neural networks. *Distill*, 4(11):e21, 2019.

[101] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

[102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[103] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*, 2019.

[104] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[105] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[106] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[107] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[108] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[109] Ning Xie, Gabrielle Ras, Marcel van Gerven, and Derek Doran. Explainable deep learning: A field guide for the uninitiated. 2020.

[110] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[111] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

[112] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[113] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[114] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[115] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[116] Tianwei Yue and Haohan Wang. Deep learning for genomics: A concise overview. *arXiv preprint arXiv:1802.00810*, 2018.

[117] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature genetics*, 51 (1):12–18, 2019.

[118] Michael KK Leung, Andrew Delong, Babak Alipanahi, and Brendan J Frey. Machine learning in genomic medicine: a review of computational problems and data sets. *Proceedings of the IEEE*, 104(1):176–197, 2015.

[119] Jian Zhou and Olga G Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature methods*, 12(10):931–934, 2015.

[120] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[121] Jack Lanchantin, Ritambhara Singh, Zeming Lin, and Yanjun Qi. Deep motif: Visualizing genomic sequence classifications. *arXiv preprint arXiv:1605.01133*, 2016.

[122] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[123] David R Kelley, Yakir A Reshef, Maxwell Bileschi, David Belanger, Cory Y McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome research*, 28(5):739–750, 2018.

[124] Jian Zhou, Chandra L Theesfeld, Kevin Yao, Kathleen M Chen, Aaron K Wong, and Olga G Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature genetics*, 50(8):1171–1179, 2018.

[125] Haotian Teng, Minh Duc Cao, Michael B Hall, Tania Duarte, Sheng Wang, and Lachlan JM Coin. Chiron: translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7(5):giy037, 2018.

[126] Sara Sheehan and Yun S Song. Deep learning for population genetic inference. *PLoS computational biology*, 12(3):e1004845, 2016.

[127] Jonas Meisner and Anders Albrechtsen. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *bioRxiv*, 2020.

[128] Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with

approximate bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2021.

[129] Nathan Killoran, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan J Frey. Generating and designing dna with deep generative models. *arXiv preprint arXiv:1712.06148*, 2017.

[130] Sam Sinai, Eric Kelsic, George M Church, and Martin A Nowak. Variational auto-encoding of protein sequences. *arXiv preprint arXiv:1712.03346*, 2017.

[131] Daniel Mas Montserrat, Carlos Bustamante, and Alexander Ioannidis. Class-conditional vae-gan for local-ancestry simulation. *arXiv preprint arXiv:1911.13220*, 2019.

[132] Jim Clauwaert, Gerben Menschaert, and Willem Waegeman. Explainability in transformer models for functional genomics. *Briefings in bioinformatics*, 22(5): bbab060, 2021.

[133] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[134] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature methods*, 18(10):1196–1203, 2021.

[135] Sheikh Jubair, James R Tucker, Nathan Henderson, Colin W Hiebert, Ana Badea, Michael Domaratzki, and WG Dilantha Fernando. Gptransformer: A transformer-based deep learning method for predicting fusarium related traits in barley. *Frontiers in plant science*, 12:761402–761402, 2021.

[136] Benjamin Bloem-Reddy and Yee Whye Teh. Neural network models of exchangeable sequences. In *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[137] Jeffrey Chan, Valerio Perrone, Jeffrey Spence, Paul Jenkins, Sara Mathieson, and Yun Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems*, 31, 2018.

[138] Yuki Saito, Takuma Nakamura, Hirotaka Hachiya, and Kenji Fukumizu. Exchangeable deep neural networks for set-to-set matching and learning. In *European Conference on Computer Vision*, pages 626–646. Springer, 2020.

Chapter 2

Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations

The attached paper is published in *Nature Communications* (Nait Saada et al., 2020).

A statement of contribution is published as part of the paper.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19588-x OPEN

Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations

Juba Nait Saada [1✉], Georgios Kalantzis [1], Derek Shyr [2], Fergus Cooper [3], Martin Robinson [3], Alexander Gusev [4,5,7] & Pier Francesco Palamara [1,6,7✉]

Detection of Identical-By-Descent (IBD) segments provides a fundamental measure of genetic relatedness and plays a key role in a wide range of analyses. We develop FastSMC, an IBD detection algorithm that combines a fast heuristic search with accurate coalescent-based likelihood calculations. FastSMC enables biobank-scale detection and dating of IBD segments within several thousands of years in the past. We apply FastSMC to 487,409 UK Biobank samples and detect ~214 billion IBD segments transmitted by shared ancestors within the past 1500 years, obtaining a fine-grained picture of genetic relatedness in the UK. Sharing of common ancestors strongly correlates with geographic distance, enabling the use of genomic data to localize a sample's birth coordinates with a median error of 45 km. We seek evidence of recent positive selection by identifying loci with unusually strong shared ancestry and detect 12 genome-wide significant signals. We devise an IBD-based test for association between phenotype and ultra-rare loss-of-function variation, identifying 29 association signals in 7 blood-related traits.

[1] Department of Statistics, University of Oxford, Oxford, UK. [2] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA. [3] Department of Computer Science, University of Oxford, Oxford, UK. [4] Brigham & Women's Hospital, Division of Genetics, Boston, MA 02215, USA. [5] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [6] Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [7] These authors jointly supervised this work: Alexander Gusev, Pier Francesco Palamara. ✉email: juba.naitsaada@stats.ox.ac.uk; palamara@stats.ox.ac.uk

Large-scale genomic collection, through efforts like the NIH All of Us research program[1], the UK BioBank[2], Genomics England[3], and the Million Veteran Program[4], is expected to yield datasets of hundreds of thousands of individuals and to grow to millions in the coming years. Utilizing such datasets to understand disease and health outcomes requires understanding the fine-scale genetic relationships between individuals. These relationships can be characterized using short segments (less than 10 centimorgans [cM] in length) that are inherited identical by descent (IBD) from a common ancestor between purportedly unrelated pairs of individuals in a dataset[5]. Accurate detection of shared IBD segments has a number of downstream applications, which include reconstructing the fine-scale demographic history of a population[6–9], detecting signatures of recent adaptation[10,11], discovering phenotypic association[12,13], estimating haplotype phase[5,14,15], and imputing missing genotype data[16,17], a key step in genome-wide association studies (GWAS)[18]. Detection of IBD segments in millions of individuals within modern biobanks poses a number of computational challenges. Although several IBD detection methods have been published[19–21], few scale to analyses comprising more than several thousand samples. However, scalable methods that do exist trade modeling accuracy for computational speed. As a result, current IBD detection algorithms are either scalable but heuristic, solely relying on genotypic similarity to detect shared ancestry and not providing calibrated estimates of uncertainty, or too slow to be applied to modern biobanks. Here, we introduce an IBD detection algorithm, called fast sequentially Markovian coalescent (FastSMC), which is both fast, enabling IBD analysis of modern biobank datasets, and accurate, relying on coalescence modeling to detect short IBD segments (down to 0.1 cM). FastSMC quantifies uncertainty and estimates the time to most recent common ancestor (TMRCA) for individuals that share IBD segments. It does so by efficiently leveraging information provided by allele sharing, genotype frequencies, and demographic history, which results in a cost-effective boost in accuracy.

We validate the scalability, accuracy, and robustness of the FastSMC algorithm in detecting IBD sharing within recent millennia using extensive coalescent simulation. Leveraging the speed and accuracy of FastSMC, we analyze IBD sharing in 487, 409 phased individuals from the UK Biobank dataset, identifying and characterizing ~214 billion IBD segments transmitted by shared ancestors within the past 50 generations. This network of shared ancestry enables us to reconstruct a fine-grained picture of time-dependent genomic relatedness in the UK. Analysis of the distribution of recent sharing within specific genomic regions reveals evidence of recent positive selection at 12 loci. We find the sharing of IBD to be highly correlated with geographic distance and the sharing of rare variants. Leveraging this correlation, we detect 20 associations to genomic loci harboring loss-of-function (LoF) variants with seven blood-related phenotypes. These results underscore the importance of modeling distant relatedness to reveal subtle population structure, recent evolutionary history, and rare pathogenic variation.

Results

Overview of the FastSMC method. The algorithm we developed, called FastSMC, detects IBD segments using a two-step procedure. In the first step (identification), FastSMC uses genotype hashing to rapidly identify IBD candidate segments, which enables us to scale to very large datasets. In the second step (verification), each candidate segment is tested using a coalescent hidden Markov model (HMM), which enables us to improve accuracy, compute the posterior probability that the segment is IBD (IBD quality score), and provide an estimate for the TMRCA in the genomic region. The identification step leverages the GERMLINE2 algorithm that we developed and present here, which improves over GERMLINE's[19] speed and memory requirements and thus enables us to very efficiently detect IBD candidate regions in millions of genotyped samples. GERMLINE2 utilizes hash functions to identify pairs of individuals whose genomes are identical in small genomic regions, thus being Identical-By-State (IBS). The presence of these short identical segments triggers a local search for longer segments that are likely to reflect recent TMRCA and thus IBD sharing in the region. Although the original GERMLINE algorithm utilizes a similar strategy, GERMLINE2 offers two key improvements, which result in faster computation and lower memory consumption. First, the GERMLINE algorithm can become inefficient in regions where certain short haplotypes can be extremely common in the population (e.g. due to high linkage disequilibrium), which results in hash collisions across a large fraction of samples, effectively reverting back to a nearly all-pairs analysis and monopolizing computation time. GERMLINE2 avoids this issue by introducing recursive hash tables, which require haplotypes to be sufficiently diverse before they are explored for pairwise analysis and significantly decrease downstream computation (Supplementary Fig. 1). Second, the GERMLINE local search (extension) step requires storing the entire genotype dataset in memory, which is prohibitive for biobank-scale analyses. Instead, GERMLINE2 uses an on-line strategy, reading a polymorphic site at a time without storing complete genotype information in memory, which enables scaling this analysis to millions of individuals. While long IBS regions are often co-inherited from recent common ancestors, thus being IBD, this need not always be the case[22]. In its verification step, FastSMC thus leverages coalescence modeling to filter out candidate segments that are IBS, but not IBD. To achieve this, FastSMC analyzes every detected candidate segment using the ASMC algorithm[23], a recently proposed coalescent-based HMM that builds on recent advances in population genetics inference[24–27] to enable efficient estimation of the posterior of the TMRCA for a pair of individuals at each site along the genome. A key advantage of the ASMC algorithm over previous coalescent-based models is that it enables estimating the TMRCA in SNP array data in addition to sequencing data. FastSMC can thus be tuned to be applied to both types of data. FastSMC produces a list of pairwise IBD segments with each segment associated to an IBD quality score - i.e the average probability of the TMRCA being between present time and the user-specified time threshold – and an age estimate – i.e. the average maximum a posteriori (MAP) TMRCA along the segment. FastSMC can be parallelized to efficiently scale up to biobank datasets. The FastSMC software implements both the GERMLINE2 and ASMC algorithms, and is freely available.

Comparison to existing methods. We measured FastSMC's accuracy using extensive realistic coalescent simulations that mimic data from the UK Biobank[2]. We benchmarked IBD detection for FastSMC in addition to three other widely used or recently published IBD detection methods: GERMLINE[19], RefinedIBD[20], and RaPID[21]. Parameters for all methods were optimized to maximize accuracy and evaluated on the detection of IBD segments within the past 25, 50, 100, 150, or 200 generations on simulated populations with a European ancestry (Supplementary Table 1). We found that FastSMC outperforms the accuracy of all the other methods at all time scales (Fig. 1a, c, Supplementary Fig. 2, and Supplementary Table 2). As expected, model-based algorithms such as FastSMC and RefinedIBD tend to achieve better results in detecting older (shorter) IBD segments than genotype-matching methods, which cannot reliably exclude

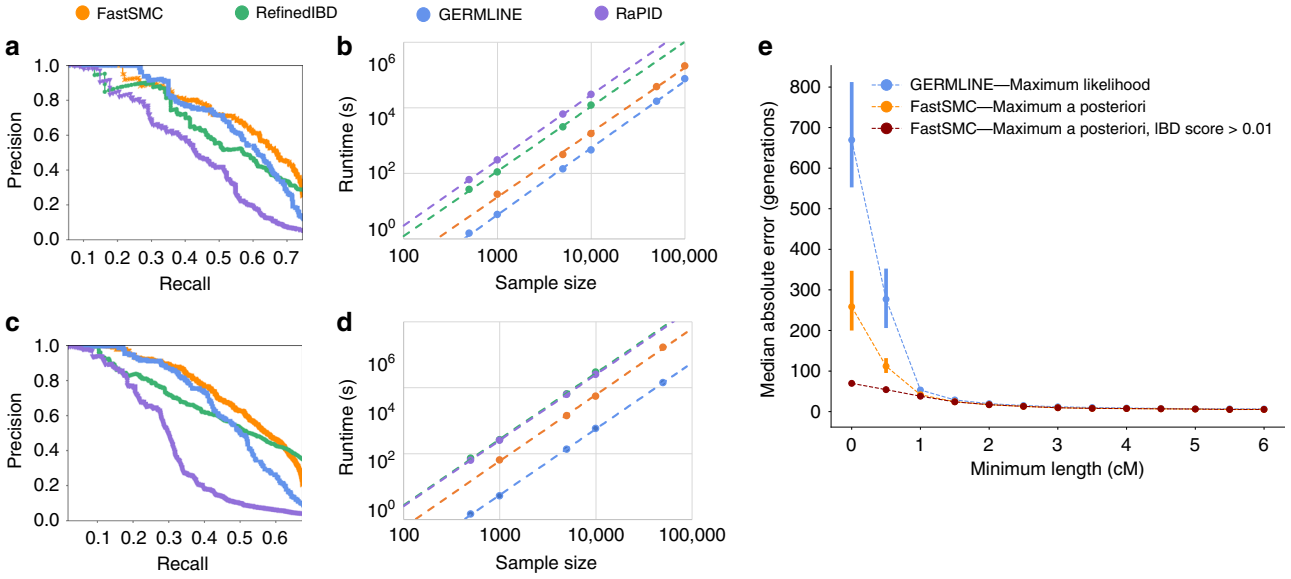


Fig. 1 FastSMC in coalescent simulations. a (respectively **c**). Precision-recall curve randomly sampled from 10 realistic European simulated datasets with 300 haploid samples for IBD segments detection within the past 50 generations (respectively 100 generations), within the recall range where all methods are able to provide predictions. **b** (respectively **d**). Running time (CPU seconds) using chromosome 20 of the UK Biobank for IBD segments detection within the past 50 generations (respectively 100 generations). The entire cohort of 487, 409 samples (across 7913 SNPs) was randomly downsampled into smaller batches. Only one thread was used for each method and running time trend lines in logarithmic scale are shown, reflecting differences in the quadratic components of each algorithm. Parameters for all methods were optimized to maximize accuracy and used for both accuracy and running time benchmarking (details in Methods section). **e** Median absolute error of the length-based maximum likelihood age estimate (MLE) on segments detected by GERMLINE in blue and our MAP age estimate on segments detected by FastSMC (no filtering on the IBD quality score in orange, and a minimum IBD quality score of 0.01 in dark red) for 300 haploid simulated samples with European ancestry. As a reference, the average TMRCA at a random site for a pair is in the order of several thousands of generations[23]. Only detected IBD segments longer than the minimum length represented on the *X*-axis were considered. We ran both algorithms with the same minimum length of 0.001 cM and other parameters from grid search results for a time threshold of 50 generations. Data are represented as mean values ± SEM over 10 simulations.

short segments where genotypes are IBS but not IBD (Supplementary Table 2). FastSMC relies on the ASMC algorithm in its validation step, which was shown to be robust to the use of an inaccurate recombination rate map or violations of assumptions on allele frequencies in SNP ascertainment[23]. We thus expect it to be similarly robust to several types of model misspecification. In particular, we tested the effects of using a misspecified demographic model on FastSMC's accuracy, and observed that while this results in biased estimates of segment age (Supplementary Fig. 3), a wrong demographic model does not affect accuracy (Supplementary Table 3).

Next, we evaluated the computational efficiency of FastSMC and other methods using phased data from chromosome 20 of the UK BioBank (Fig. 1b, d and Supplementary Fig. 4). As expected, the improvement in accuracy achieved leveraging FastSMC's validation step leads to slightly increased computing time compared to only using GERMLINE, the most scalable method. FastSMC is faster than RefinedIBD, the closest method in terms of accuracy for short segments. For instance, detecting IBD segments within the past 50 generations on 10,000 samples takes 27 min for FastSMC, 9 min for GERMLINE, 3 h and 17 min for RefinedIBD, and 6 h and 58 min for RaPID. We finally assessed the memory cost of FastSMC and other methods (Supplementary Fig. 5). FastSMC does not store the genotype hashing or IBD segments, resulting in a very low memory footprint, whereas the memory requirements of other methods become prohibitive for large sample sizes such as those required to analyze the entire UK Biobank cohort. For example, analyzing chromosome 20 for a time threshold of 50 generations and 10,000 random diploid individuals from the UK Biobank dataset, FastSMC requires 1.4 GB of RAM compared to 3.8 GB for GERMLINE, 11.5 GB for RefinedIBD, and 62.9 GB for RaPID.

Downstream analysis of IBD sharing such as demographic inference or the study of natural selection often involves estimating the age of IBD segments. Because current approaches do not explicitly model the TMRCA between IBD individuals, segment age is typically estimated through the length of the IBD segment. FastSMC, on the other hand, explicitly models TMRCA across individuals, leveraging additional prior information (such as demography and allele frequencies) to produce an improved estimate of IBD segment age. We found FastSMC's segment age estimates to be more accurate than a length-based estimator (Supplementary Fig. 6), with significant gains for short segments as a result of the additional modeling in the validation step of the algorithm (e.g. the median error from FastSMC's segments age estimate decreased by ~60% for segments ≥0.5 cM compared to the current approach, Fig. 1e). FastSMC's increased accuracy in estimating coalescence times in IBD segments will translate in improved resolution for downstream applications that leverage this type of information.

IBD sharing and population structure in the United Kingdom. We leveraged the scalability and accuracy of FastSMC to analyze 487,409 phased British samples from the UK Biobank, obtaining a fine-grained picture of the genetic structure of the United Kingdom. We detected ~214 billion IBD segments shared within the past 1500 years, with ~75% of all pairs of individuals sharing at least one common ancestor (Supplementary Fig. 7). Analyzing the fraction of genome covered by IBD segments, we observed that

~93% of individuals in the cohort have >90% of their genome covered by at least one IBD segment in the past 50 generations, compared to only ~4% of individuals for the past 10 generations (Supplementary Fig. 8A). Looking for geographic patterns, we noticed that, despite the large sample size of the UK Biobank cohort, the average fraction of genome covered by at least one IBD segment is substantially heterogeneous across UK postcodes for recent time scales – ranging from 53.4% in London Eastern Central (EC) to 76.2% in Stockport (SK) for 10 generations – but more uniform at deeper time scales – ranging from 96.7% in London EC to 99.2% in Stockport for 50 generations (Supplementary Fig. 8B, C, D). The observation of a non-uniform IBD coverage has implications for downstream methods that rely on distant relatedness at each genomic site, such as variant discovery, phasing, and imputation.

Next, we analyzed the network of recent genetic relatedness for 432,968 samples for whom birth coordinates are available. As observed in previous studies of fine-scale genetic structure[28] with a considerably smaller sample size[29], genetic clusters within the UK tend to be localized within geographic regions (Fig. 2). Leveraging FastSMC's accuracy and the large sample size of the UK Biobank dataset we were able to zoom into increasingly smaller regions, finding that such clusters extend beyond broad geographic clines. Smaller geographic regions revealed increasingly fine-grained clusters of individuals born within a few tens of kilometers from each other, likely reflecting the presence of extended families which experienced limited migration during recent centuries. We found that individuals throughout the UK, including cosmopolitan regions, find overwhelmingly more recent genetic ancestors within their own postcode than in other regions of the country, reflecting isolation-by-distance due to limited migration across the country in recent generations (Supplementary Fig. 9 and https://ukancestrymap.github.io/ for an interactive website displaying these results). For instance, within the past ~300 years, two individuals born in North London share on average 0.0092 common ancestors and two individuals born in Birmingham share on average 0.0043 common ancestors. In contrast, and despite the relative geographic proximity, an individual born in North London shares on average a substantially lower 0.00059 ancestors with one born in Birmingham. We observed similar trends when we restricted this analysis to samples of self-reported and inferred non-White British ancestry[2], as shown in Supplementary Fig. 10. We further visualized the strong link between genetic and physical distances (Supplementary Fig. 11) by building a low-dimensional planar representation of pairwise genetic distances across postcodes within the past 600 years (Supplementary Fig. 12B), which

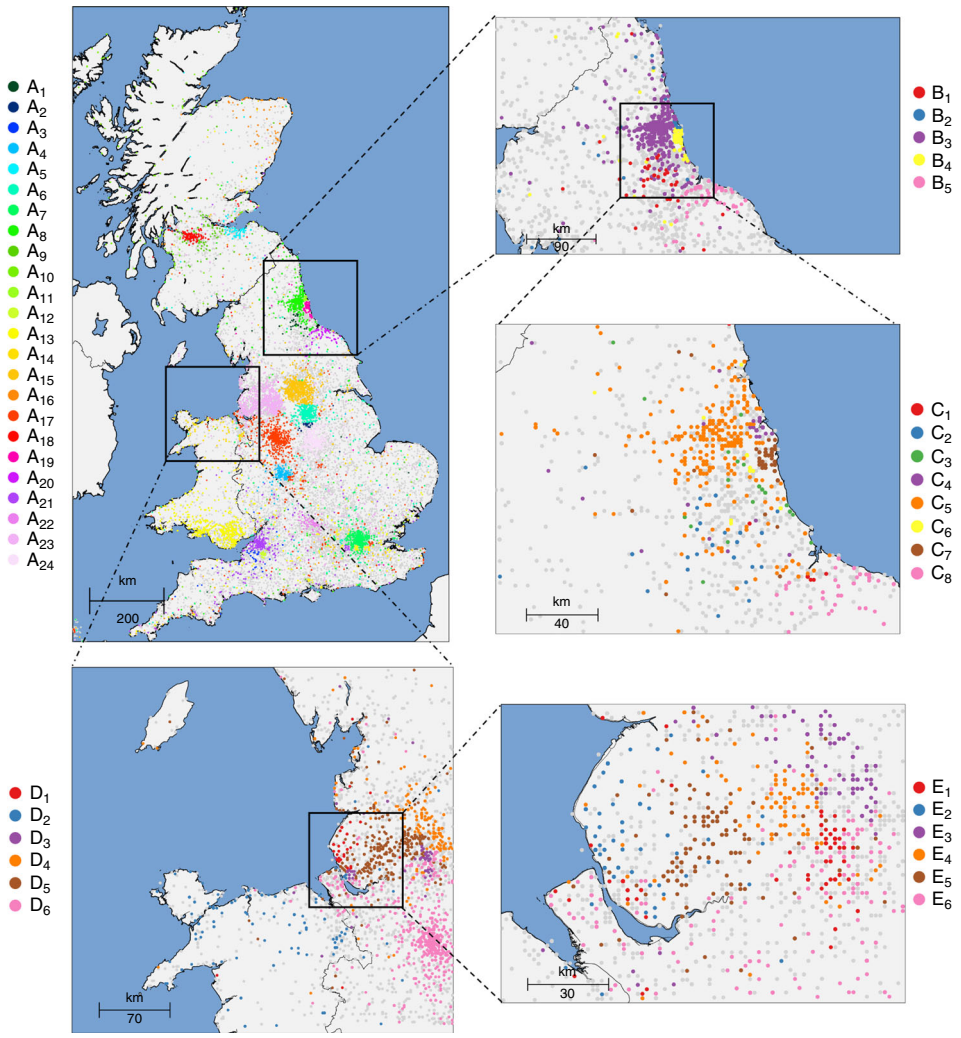


Fig. 2 Fine-scale population structure in the UK. Hierarchical clustering of 432,866 individuals from the UK Biobank dataset based on the sharing of IBD segments within the past 10 generations. Individuals in clusters with <500 samples are shown in light gray. We observed 24 main clusters across the country (top left) and we refined two regions, corresponding to Newcastle (NE) (top right) and Liverpool (L) (bottom), revealing fine scale population structure. No relationship between clusters is implied by the colors or cluster labeling across different plots (details are provided in Methods section).

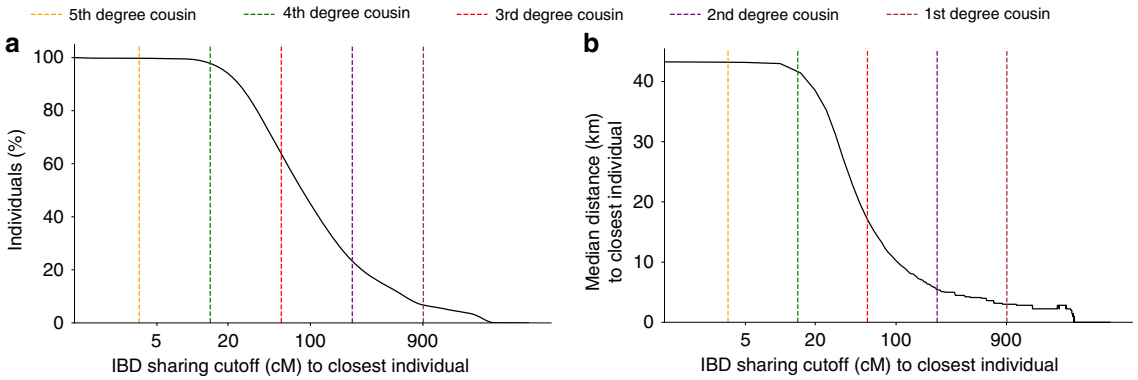


Fig. 3 Genetic relatedness and geographic distances in the UK Biobank dataset. For each of the 432,968 UK Biobank samples with available geographic data, we detected the individual sharing the largest total amount (in cM) of genome IBD within the past 10 generations (referred to as closest individual). **a** For each value x of total shared genome (in cM) on the X-axis, we report the percentage of UK Biobank samples (Y-axis) that share x or more with their closest individual. **b** For each value x of total shared genome (in cM) on the X-axis, we report the median distance (km, computed every 10 cM) for all pairs of (sample, closest individual) who shared at least x. Vertical dashed lines indicate the expected value of the total IBD sharing for kth degree cousins, computed using $2G(1/2)^{2(k+1)}$, where $G = 7247.14$ is the total diploid genome size (in cM) and k represents the degree of cousin relationship (e.g. $k = 2$ for second degree cousins, separated by $2(k+1)$ generations)[11]. The value of 45 observed when no sharing cutoff is considered (i.e. when the x value approaches 0) reflects the median prediction error for a random individual, regardless of how much IBD they share with the closest individual.

we found to closely reflect geographic distance across these regions.

We hypothesized that the presence of such a fine-scale genetic structure in the UK may be used to effectively predict the birth location of an individual. This would imply that FastSMC may be used to predict other subtle environmental covariates, which may be causing confounding in GWAS[30]. We computed the fraction of individuals who find at least one close genetic relative within the UK Biobank dataset (Fig. 3a). We observed that almost all individuals (99.8%) find a genetic relative with IBD sharing equivalent to a 5th degree cousin (3.5 cM) or closer relationship, with 64.6% of samples finding a putative 3rd degree cousin (56.6 cM) or closer relative. We found that, indeed, stronger genetic ties translate into greater proximity of birth locations (Fig. 3b). For instance, for individuals sharing a fraction of genome equivalent to 3rd degree cousins or closer, the median distance between birth locations is 17 km. Very close genetic relationships are also pervasive in the dataset: about one in four individuals (23.4%) has a relative with genetic sharing equivalent to a 2nd degree cousin (226.5 cM) or closer; for these samples the median distance between birth locations is only 5 km. We then sought to quantify how accurate IBD sharing is at predicting the birth coordinate of a random UK Biobank sample. We found that the birth coordinate of the closest genetic relative for a random UK Biobank individual can be used to predict the individual's birth location with a median error of 45 km. In contrast, taking a random individual to predict the birth location would result in a median error of 200.6 km (95% bootstrap CI = [200.3, 200.8]). Using this approach to predict a random individual's current place of living, rather than birth location, resulted in a higher median error of 75 km (Supplementary Fig. 13). Additional details are shown in Supplementary Fig. 14. Prediction in non-White British samples resulted in decreased accuracy due to the smaller sample size (Supplementary Fig. 15B). Conversely, restricting this analysis to only samples of White British ancestry resulted in a slightly improved median error of 42 km (Supplementary Fig. 15A). These findings provide empirical support to recent hypotheses that extensive segment sharing within genealogical databases may be used to recover the genotypes of target individuals[31], or to re-identify individuals through long-range familial searches[32].

As expected, entries of the IBD matrix within the past 10 generations are highly correlated to corresponding allele sharing entries of the genetic relatedness matrix (GRM) that is widely used in applications such as principal component analysis[33] and linear mixed models[34] ($r = 0.66$, SE = 0.02 computing the GRM for 50,000 random samples using Plink[35]). However, by conditioning on the presence of recent shared ancestors, the IBD matrix is better suited to capture genomic relationships that are recent but would not result in large outlier entries in the GRM. To compare the performance of our IBD-based approach with an approach based on allele sharing in the absence of extremely close relatives, we removed individuals with very recent genetic ties (≤3rd degree relatives, e.g. first degree cousins[2]) and used a simple machine learning approach (K Nearest Neighbors) to perform the prediction of birth coordinates. Prediction of birth coordinates using the IBD matrix was 44% more accurate than prediction based on allele sharing (Supplementary Fig. 16). We also found birth coordinates predicted using IBD sharing to be strongly correlated to true coordinates ($r = 0.74$, 95% CI = [0.73, 0.75], for Y-coordinates and $r = 0.6$, 95% CI = [0.59, 0.62], for X-coordinates), substantially higher than the correlation achieved using allele sharing ($r = 0.43$, 95% CI = [0.41, 0.45], for Y-coordinates and $r = 0.31$, 95% CI = [0.30, 0.33], for X-coordinates).

Analyzing broader patterns of IBD sharing, we found that individuals living in the North of the country share more common ancestry than in the South, and that more generally regions within Scotland, England, and Wales tend to cluster with other regions within the same country. We estimated the effective population size from 300 years ago within each postcode and detected significant correlation ($r = 0.28$, 95% CI = [0.09, 0.47] by bootstrap using postcodes) with present-day population density (Supplementary Fig. 12A). As we look deeper in time, IBD sharing patterns tend to shift and reflect historical migration events within the country. Notably, we find that individuals throughout England share deep genealogical connections with other individuals currently living in the North-West and the North of Wales (Supplementary Fig. 17). These regions correspond to the unromanised regions of the UK and Britons living there are believed to have experienced limited admixture during the Anglo Saxon settlement of Britain occurring at the end of the Roman rule in the 5th century[36]. Elevated IBD

sharing between these regions may thus reflect deep genealogical connections within the ancient Briton component of modern day individuals, which is overrepresented in the North-West of the country. As expected, large cosmopolitan regions display substantially more uniform ancestry across the UK. London, in particular, has a uniform distribution of ancestry at deep time scales (Supplementary Fig. 9), suggesting that it has attracted substantial migration for extended periods of time.

Signals of recent positive selection in the UK Biobank. We analyzed locus-specific patterns of recent shared ancestry, seeking evidence for recent positive selection by identifying loci with unusually high density of recent coalescence times in the UK Biobank dataset. We computed the DRC_{50} (Density of Recent Coalescence) statistic[23], capturing the density of recent coalescence events along the genome within the past 50 generations (Supplementary Fig. 18). Large values of the DRC_{50} statistic are found at loci where a large number of individuals descend from a small number of recent common ancestors, a pattern that is likely to reflect the rapid increase in frequency of a beneficial allele due to recent positive selection. Although, as expected, the DRC_{50} statistic computed in this analysis is strongly correlated ($r = 0.67$) with the DRC_{150} statistic that was computed using fewer samples from a previous UK Biobank data release by Palamara et al.[23], the DRC_{50} statistic reflects more recent coalescence events than the DRC_{150} statistic, and thus more specifically reflects natural selection occurring within recent centuries.

Analyzing the distribution of the 52,003 windows in the UK Biobank dataset, we detected 12 genome-wide significant loci (at an approximate DRC_{50} $p < 0.05/52,003 = 9.6 \times 10^{-7}$; Fig. 4 and Supplementary Table 4). Five of these loci are known to be under recent positive selection, harboring genes involved in immune response (*NBPF1*[37], *HLA*[38]), nutrition (*LCT*[39], *LDLR*[40]), and mucus production (*MUC2*[23]). We also identified 7 other loci, harboring genes related to immune response (*MRC1*, playing a role in both the innate and adaptive immune systems[41], and *BCAM*, encoding the Lutheran antigen system, also associated with low density lipoprotein cholesterol measurement[42]), mucus production (*CAPN8*, involved in gastric mucosal defense[43]), tumor growth (*CHD1L*, associated with tumor progression and chemotherapy resistance in human hepatocellular carcinoma[44], and *BANP*, encoding a tumor suppressor and cell cycle regulator protein[45]), as well as genetic disorders (*HYDIN*, causing primary ciliary dyskinesia[46], and *EFTUD2*, causing mandibulofacial dysostosis with microcephaly[47]). We checked that these regions are not extreme in recombination rate or marker density (Supplementary Table 5).

IBD sharing enables association of ultra-rare variants. Individuals who co-inherit a genomic region IBD from a recent common ancestor are also expected to have identical genomic sequences within that region, with the exception of de-novo mutations and other variants introduced by e.g. non-crossover gene conversion events in the generations leading to their recent common ancestor[48]. We thus expect the sharing of IBD segments to be strongly correlated to the sharing of ultra-rare genomic variants (MAF < 0.0001), which tend to be very recent in origin and are usually co-inherited through recent ancestors who carried such variants. We verified this by testing for correlation between the sharing of IBD segments at various time scales and the sharing of rare variants for the ~50k individuals included in the UK Biobank's initial exome sequencing data release[49] (Supplementary Fig. 19A). Specifically, we analyzed mutations that are carried by N out of 99,920 exome-sequenced haploid genomes (for $2 < N < 500$), which we refer to as F_N mutations[50,51]. We compared the sharing of F_N mutations to the sharing of IBD segments in the past 10 generations within all postcodes in the UK (Fig. 5a). We found that there is indeed a strong correlation between the per-postcode sharing of ultra-rare variants and the per-postcode sharing of recent ancestors (e.g. $r = 0.3$, 95% CI = [0.22,0.37] for F_3 mutations). The correlation between IBD sharing and F_N variant sharing decreases as N increases, with slightly higher correlation for more recent IBD segments, while deeper IBD sharing (within 50 generations) tends to provide better tagging of ultra-rare variants of slightly higher frequency (e.g. bootstrap $p < 0.05/50 = 0.001$ for $N = 20$; Supplementary Fig. 19B).

Based on this correlation between sharing of mutations and sharing of IBD segments, we hypothesized that IBD sharing of disease causing mutations would be predictive of disease. In particular, individuals who share an IBD segment with a pathogenic rare variant carrier in a known gene have a higher probability of carrying the pathogenic variant (by inheriting it from the shared ancestor) than the general population, and would thus be at increased risk for the phenotypic effect. The UK Biobank exome pilot[49] identified multiple rare coding variant burden associations with complex phenotypes, some of which were recently replicated[52,53]. We set out to test whether our IBD

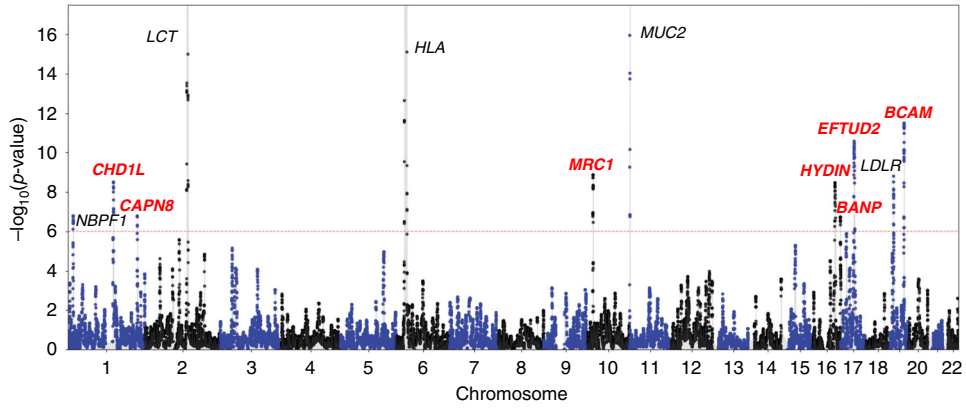


Fig. 4 Genome-wide scan for recent positive selection in the UK Biobank dataset. Manhattan plot with candidate gene labels for 12 loci detected at genome-wide significance (adjusting for multiple testing, approximate 1-sided DRC_{50} $p < 0.05/52,003 = 9.6 \times 10^{-7}$; dashed red line). The DRC_{50} statistic of shared recent ancestry within the past 50 generations was computed using 487,409 samples within the UK Biobank cohort. FastSMC detected 5 loci known to be under recent positive natural selection (gene labels in black) and 7 novel loci (in red). The corresponding p-values are reported in Supplementary Table 4.

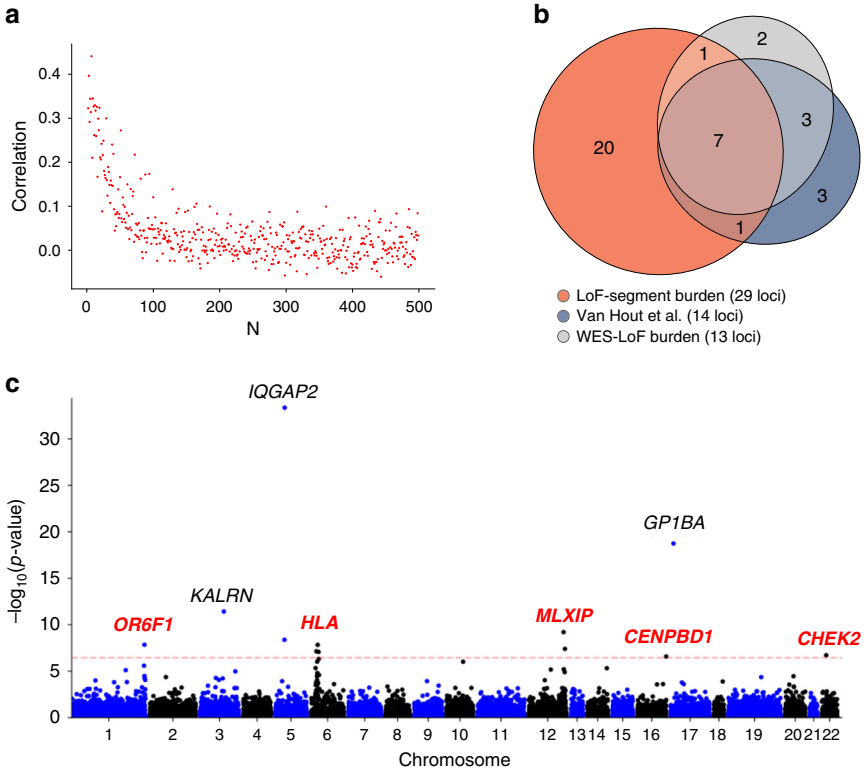


Fig. 5 IBD sharing and rare variant associations. a Correlation between IBD sharing (average number of IBD segments per pair across UK postcodes in the past 10 generations in the UK Biobank's 487,409 samples) and ultra-rare variants sharing (average number of F_N mutations per pair across UK postcodes in the UK Biobank 50k Exome Sequencing Data Release for increasing values of N). **b** Venn diagram representing the sets of exome-wide significant associated loci for 7 blood-related traits using three methods: the WES-based LoF burden test reported by Van Hout et al.[49], a WES-based LoF burden test we performed (WES-LoF burden), and the IBD-based LoF burden test we performed (LoF-segment burden). The corresponding p-values were computed using two-sided t-tests and are reported in Tables 1, 2 and Supplementary Table 8. **c** Exome-wide Manhattan plot for mean platelet (thrombocyte) volume, after SNP-correction, using 303, 125 unrelated UK Biobank samples not included in the exome sequencing cohort. Labeled genes are exome-wide significant after adjusting for multiple testing: $p < 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$; dashed red line. Black labels indicate genes that were previously reported by Van Hout et al.[49] (*KALRN, GP1BA*, and *IQGAP2*), while red labels indicate novel associations detected by our LoF-segment burden analysis. The corresponding p-values were computed using two-sided t-tests and are reported in Table 2.

inference could empower us to replicate and refine these associations using the larger non-sequenced cohort. For each previously reported gene-phenotype association, we identified all sequenced individuals with a rare LoF variant (mirroring the definition of LoF from Van Hout et al.[49]) and any IBD segments they shared with individuals in the non-sequenced cohort (we refer to these as putative LoF-segments, though they will also include sharing of the non-LoF haplotypes because the phase of the LoF is unknown). As expected due to the effects of natural selection[54], the LoF variants belonging to a small subset for which an age estimate has recently been computed[55] are slightly younger than other exome-sequenced variants (Supplementary Table 6). Note that the majority of these variants are singletons or doubletons[49] and would be excluded from imputation by most current algorithms[56]. Then, in the (independent) non-sequenced cohort, we tested for association between carrying a LoF-segment (the LoF-segment burden) and the phenotype known to be associated with that gene. This approach would be optimal when IBD individuals carry a LoF variant that arose prior to the TMRCA of their shared segment. We thus tested 10 transformations of the LoF-segment metric for association with phenotype to model uncertainty about the age distribution of the underlying causal variants.

Using our LoF-segment burden, we replicated 11 out of 14 previously reported[49] associations with seven blood-related traits at $p < 0.05/10 = 0.005$ (adjusted for testing of 10 transformations; Table 1). Strikingly, we found eight of these associations to be exome-wide significant in the non-sequenced cohort ($p < 0.05/(10 \times 14{,}249)$, reflecting 14,249 genes tested using 10 transformations; Fig. 5b). We next aimed at quantifying how effective IBD sharing is at detecting associations, compared to testing directly based on exome sequencing data. We computed the phenotypic variance explained by the indirect IBD-based test and the direct exome-based test (after subtracting the effect of covariates from both), focusing on the 14 loci reported by Van Hout et al.[49]. The ratio of these variances was 19.64% on average, corresponding to the decrease in effect-size (in units of variance) due to estimation error and inclusion of segments sharing the non-LoF haplotype. We note that, due to phase uncertainty, we expect LoF-segment burden to explain at most 50% of the variance explained by direct sequencing. Assuming the ratio of variances corresponds to the squared correlation between the LoF-segment burden estimate and the true exome burden, the LoF-segment burden estimator has statistical power equivalent to a direct exome sequencing study of 19.64% of the 303, 125, or ~60k samples[57] – effectively doubling the size of the exome study. This demonstrates FastSMC's accuracy and, more broadly, the utility of leveraging distant relatedness in identifying disease associations.

Motivated by these results, we next expanded our study to all sequenced genes for this same set of primarily blood-related traits. We identified a total of 186 exome-wide significant gene associations ($p < 0.05/(10 \times 14{,}249)$) spanning 33 genomic loci in

Table 1 Comparison between association analyses.

	Gene	Trait	Van Hout et al.	WES LoF burden	LoF-segment burden	R^2_{prop}
1	*IL33*	Eosinophil count	3.30E-10	2.01E-03	8.64E-15	72.26
2	*GP1BA*	Mean platelet (thrombocyte) volume	6.40E-08	8.84E-08	1.82E-19	32.57
3	*TUBB1*	Platelet distribution width	2.50E-23	7.34E-18	7.38E-12	07.25
4	*TUBB1*	Mean platelet (thrombocyte) volume	2.40E-08	3.01E-07	2.15E-03	04.11
5	*TUBB1*	Platelet count	2.10E-09	7.45E-07	4.21E-05	07.84
6	*HBB*	Red blood cell distribution width	5.80E-08	3.49E-02	2.25E-03	23.99
7	*HBB*	Red blood cell count	1.70E-09	7.95E-02	2.68E-02	18.23
8	*KLF1*	Red blood cell distribution width	1.50E-13	6.95E-13	3.49E-34	32.99
9	*KLF1*	Mean corpuscular hemoglobin	1.70E-16	9.11E-15	6.79E-21	16.76
10	*ASXL1*	Platelet distribution width	4.70E-09	1.44E-06	0.16E-00	00.98
11	*ASXL1*	Red blood cell distribution width	2.40E-11	8.23E-04	0.32E-00	01.03
12	*KALRN*	Mean platelet (thrombocyte) volume	2.70E-23	3.85E-18	3.79E-12	07.33
13	*IQGAP2*	Mean platelet (thrombocyte) volume	1.10E-19	3.72E-15	4.40E-34	27.43
14	*GMPR*	Mean corpuscular hemoglobin	1.10E-08	2.94E-06	7.60E-11	22.18

We report association statistics for 14 loci and 7 traits as detected by Van Hout et al.[49] (obtained using a linear mixed model), our whole-exome sequencing burden analysis (two-sided *t*-test; labeled as WES LoF burden); and the LoF-segment burden (two-sided *t*-test). The Bonferroni-corrected exome-wide significance threshold for the first two approaches is 3.4×10^{-6}, after correcting for multiple testing with ~15k genes, and 3.51×10^{-7} for the LoF-segment burden, after adjusting for 14,249 genes and 10 time transformations. We identify 10 genes at exome-wide significance with the WES-LoF burden test, and we replicate 11/14 at $p < 0.05/10 = 0.005$ (adjusted for testing of 10 transformation) using the LoF-segment association in non-sequenced samples (8 at exome-wide significance). The last column estimates the proportion of the phenotypic variation (R^2_{prop}, in %; Supplementary Table 9) of the sequenced samples that can be explained by the non-sequenced cohort; on average that is 19.64% for all the 14 reported associations, or 27.35% if focusing on the exome-wide significant signals.

the non-sequenced cohort by only leveraging LoF-segments (Table 1); these genes were not significant in the exome burden analysis due to insufficient statistical power. We noticed that some loci included multiple significant associations, suggestive of correlation between associated features as is often seen for GWAS signals. We hypothesized that, in some cases, the true underlying causal variant may be better tagged by known high-frequency SNPs, which are likely to have been detected in previous GWAS analyses. Repeating this analysis including previously associated common variants as covariates reduced the number of significant associations to 111 across 29 loci, suggesting that inclusion of significant common associations in rare variant burden tests may lead to improved interpretability and fewer false-positives due to tagging (Table 2, Fig. 5c and Supplementary Figs. 20–25).

Our exome-wide significant signals include the association between platelet count and *MPL* (two-sided *t*-test $p = 1.99 \times 10^{-7}$), which encodes the thrombopoietin receptor that acts as a primary regulator of megakaryopoiesis and platelet production and has not been previously implicated by genome-wide scans of either rare or common variants. We also detect several associations in genes that were not detected using exome sequencing but have been previously implicated in genome-wide scans for common variants, including associations between eosinophil count, *GFI1B* (two-sided *t*-test $p = 1.92 \times 10^{-7}$) and *RPH3A* (two-sided *t*-test $p = 7.63 \times 10^{-14}$)[58,59], and associations between platelet count, *IQGAP2* (two-sided *t*-test $p = 6.52 \times 10^{-8}$) and *GP1BA* (two-sided *t*-test $p = 1.43 \times 10^{-7}$)[59,60]. We also identify genes that were previously associated with other blood-related phenotypes in other populations, including the association between platelet distribution width and *APOA5* (two-sided *t*-test $p = 1.94 \times 10^{-8}$), a gene that encodes proteins regulating the plasma triglyceride levels; common variants linked to this gene have been linked to platelet count in individuals of Japanese descent[61]. We detect association between red blood cell distribution width and *APOC3* (two-sided *t*-test $p = 3.67 \times 10^{-11}$), a gene encoding a protein that interacts with proteins encoded by other genes (*APOA1, APOA4*) associated with the same trait. The association between *APOC3* and platelet count was also detected with our WES-LoF burden analysis (two-sided *t*-test $p = 2.13 \times 10^{-7}$) and by previous studies based on common SNPs[59]. We also found an association between *CHEK2* and both mean corpuscular hemoglobin (two-sided *t*-test $p = 1.43 \times 10^{-7}$) and mean platelet volume (two-sided *t*-test $p = 1.93 \times 10^{-7}$). This gene plays an important role in tumor suppression and was found to be associated with other blood traits, such as platelet crit, using both exome sequencing[49] and common GWAS SNPs[59], and red blood cell distribution width[58]. Overall, this analysis highlights the utility of applying FastSMC on a hybrid sequenced/genotyped cohort to identify novel, rare variant associations and/or characterize known signals in larger cohorts.

Discussion

We developed FastSMC, an algorithm for IBD detection that scales well in analyses of very large biobank datasets, is more accurate than existing methods, enables estimating the time to most recent common ancestor for IBD individuals, and provides an estimate of uncertainty for detected IBD regions. We leveraged FastSMC to analyze 487,409 British samples from the UK Biobank dataset, detecting ~214 billion IBD segments transmitted by shared common ancestors in the past ~1500 years. This enabled us to obtain high-resolution insight into recent population structure and natural selection in British genomes. Lastly, we used IBD sharing between exome sequenced and non-sequenced samples to infer the presence of LoF variants, successfully replicating known burden associations with seven blood-related complex traits and revealing novel gene-based associations.

The level of geographic granularity that could be captured by the IBD networks emerging from our analysis underscores the importance of modeling distant relationships in genetic studies. Indeed, detecting IBD segments among close and distant relatives is a key step in many analyses, such as genotype imputation or haplotype phase inference[5,14,15], haplotype-based associations[12,13], as well as in the estimation of evolutionary parameters such as recombination rates[62,63], or mutation and gene conversion rates[48,64]. More broadly, the observed geographic heterogeneity in IBD sharing and fine-scale structure is a reminder that human populations substantially deviate from random mating, even within small geographic regions. In particular, efforts to generate optimal sequenced reference panels for imputation[65] may be greatly improved by directly sampling based on distant relatedness. Our findings also provide empirical support to recent hypotheses that relatedness within available genomic databases is sufficiently pervasive to enable recovering the genotypes of target individuals[31], or to re-identify individuals

Table 2 Associations detected using LoF-segment burden.

	Trait	Chr	Region (Mb)	Min. *p*-value	Candidate gene(s)
1	Eosinophil count	chr6	26.01:31.10	1.21E-26	*HIST1H1A, HIST1H1C, HIST1H1T, HIST1H2BF, HIST1H3E, HIST1H4F, BTN3A2, BTN2A2, BTN3A3, BTN2A1, BTN1A1, ABT1, HIST1H2AG, HIST1H2AH, PRSS16, POM121L2, ZNF391, HIST1H2BM, PGBD1, HIST1H2AK, HIST1H2BO, OR2B2, OR2B6, ZNF165, ZSCAN16, ZKSCAN8, ZSCAN9, ZKSCAN4, NKAPL,* **ZSCAN31**, *ZSCAN12, ZSCAN23, GPX6, PSORS1C2*
2	Eosinophil count	chr9	6.21:6.25	8.64E-15	**IL33**[49,59]
3	Eosinophil count	chr9	135.82:135.86	1.92E-07	**GFI1B**[59]
4	Eosinophil count	chr12	113.01:113.41	7.63E-14	**RPH3A**, OAS3[58]
5	Mean corpuscular hemoglobin	chr6	16.23:16.29	7.60E-11	**GMPR**[49,58,59]
6	Mean corpuscular hemoglobin	chr6	25.72:31.10	3.82E-69	*HIST1H2BA, SLC17A2, HIST1H2AB, HFE*[59], *HIST1H4C, HIST1H2BD,* **HIST1H4D**, *HIST1H2BG, HIST1H2AE, HIST1H1D, BTN2A1, HIST1H2AG, HIST1H4I, ZNF184, PSORS1C2*
7	Mean corpuscular hemoglobin	chr19	12.98:12.99	6.79E-21	**KLF1**[49,58,59]
8	Mean corpuscular hemoglobin	chr22	29.08:29.13	1.43E-07	**CHEK2**
9	Mean platelet thrombocyte volume	chr1	247.87:247.88	1.44E-08	**OR6F1**
10	Mean platelet thrombocyte volume	chr3	123.81:124.44	3.79E-12	**KALRN**[49,59,60]
11	Mean platelet thrombocyte volume	chr5	74.80:76.00	4.40E-34	*POLK,* **IQGAP2**[49,59]
12	Mean platelet thrombocyte volume	chr6	26.18:27.92	1.46E-08	*HIST1H4D,* **POM121L2**, *OR2B6*
13	Mean platelet thrombocyte volume	chr12	122.51:124.49	6.29E-10	**MLXIP**, *ZNF664*
14	Mean platelet thrombocyte volume	chr16	90.03:90.03	2.61E-07	**CENPBD1**
15	Mean platelet thrombocyte volume	chr17	4.83:4.83	1.82E-19	**GP1BA**[49,59]
16	Mean platelet thrombocyte volume	chr22	29.08:29.13	1.93E-07	**CHEK2**
17	Platelet count	chr1	43.80:43.82	1.99E-07	**MPL**
18	Platelet count	chr5	75.69:76.00	6.52E-08	**IQGAP2**[59]
19	Platelet count	chr6	26.59:26.60	1.1E-07	**ABT1** (within *HLA* region)
20	Platelet count	chr12	109.88:113.33	4.82E-13	*KCTD10, TCHP,* **RPH3A**[60]
21	Platelet count	chr17	4.83:4.83	1.43E-07	**GP1BA**[59,60]
22	Platelet distr. width	chr11	116.66:116.66	1.94E-08	**APOA5**
23	Platelet distr. width	chr17	4.83:4.83	4.26E-09	**GP1BA**[59]
24	Platelet distr. width	chr20	57.59:57.60	7.38E-12	**TUBB1**[49,59]
25	Red blood cell count	chr6	26.45:31.10	1.39E-10	*BTN2A1, POM121L2, HIST1H2BM, HIST1H2BO, ZNF165,* **ZSCAN9**, *ZKSCAN4, PGBD1, ZSCAN31, GPX6, PSORS1C2*
26	Red blood cell distr. width	chr6	26.01:28.48	3.03E-15	*HIST1H1A, HIST1H3A, HIST1H1C, HIST1H1T, HIST1H4D, HIST1H4F, BTN3A2, HIST1H2AG, HIST1H2AH,* **ZNF391**, *HIST1H2BM, HIST1H2AM, ZNF165, ZSCAN16, NKAPL, PGBD1, ZSCAN31, ZSCAN12, GPX6*
27	Red blood cell distr. width	chr9	135.82:135.86	8.1E-08	**GFI1B**[58,59]
28	Red blood cell distr. width	chr11	116.69:116.70	3.67E-11	**APOC3**
29	Red blood cell distr. width	chr19	12.98:12.99	3.49E-34	**KLF1**[49]

Exome-wide significant associations (after adjusting for multiple testing $p < 0.05/(14{,}249 \times 10) = 3.51 \times 10^{-7}$) detected using LoF-segment burden (SNP-adjusted). Associated genes are clustered in 29 loci. For each locus we report the set of associated genes and minimum *p*-value. The gene corresponding to the minimum *p*-value is highlighted in bold.

through familial searches[32]. As demonstrated by our analysis, leveraging IBD to impute ultra-rare variant burden from small sequenced reference panels can be an effective approach to reveal gene associations in complex traits and diseases. Although our analysis only considered association to individual genes, in principle genome-wide IBD sharing could also be leveraged to enhance common SNP-based risk prediction, which is particularly relevant for non-European cohorts where sequencing is limited.

FastSMC inherits some of the limitations and caveats of the GERMLINE and ASMC algorithms. First, as with most IBD detection methods, FastSMC requires the availability of phased data. We note, however, that when very large cohorts are analyzed, long-range computational phasing results in high-quality haplotype estimation with switch errors rates as low as one every several tens of centimorgans[5,14]. This is particularly true in regions that harbor IBD segments, which are of interest in our analysis. Nevertheless, our analysis has shown the presence of substantial regional heterogeneity in the extent to which individual genomes are spanned by IBD segments, which is likely reflected in a heterogeneous quality of phasing, as well as other downstream analyses, such as imputation, that rely on the presence of IBD segments. Second, FastSMC requires the input of a demographic model and allele frequencies as a prior in order to accurately estimate the age of IBD segments, and is thus subject to biases whenever the demographic model is misspecified. Although we have verified that these biases are not substantial, future work may enable us to simultaneously estimate IBD sharing and demographic history. Like ASMC, FastSMC may tolerate reasonable levels of model misspecification, but a user should be aware that issues such as substantial inaccuracies in the genetic map or strong heterogeneity of genotyping density or quality may lead to biases. Third, FastSMC currently does not enable analysis of imputed data, a limitation that is shared by other IBD detection methods. Finally, the accurate identification of extremely short IBD segments (<1 cM) spanning hundreds of generations remains a challenge, both computationally (as the number of such segments increases very rapidly) and methodologically (as fewer variants are available to provide signal for distinguishing IBS from IBD). This underscores the importance of quantifying uncertainty through estimates such as IBD segment quality, precision, and recall, which are affected by parameter choices for FastSMC and other IBD detection methods. Our analysis detected an average of 1.8 IBD segments per pair in the UK Biobank dataset within the past 50 generations. This is consistent with a previous study focusing on longer and more recent segments (average of 0.1 segments >2.9 cM per pair[66]), but less than another recent study in a similar length range (average 1.96 segments >2 cM per pair[67]). Taking uncertainty of the detected IBD segments into account may reconcile these estimates.

In addition to algorithmic improvements to address the limitations above, we believe there are a number of interesting future extensions and interactions with other existing methods in this area. FastSMC's identification step currently relies on the GERMLINE2 genotype hashing strategy. It will be interesting to test other heuristic strategies for rapidly identifying identical segments, such as the locality-sensitive hashing strategy recently implemented in the iLASH algorithm (exhibiting 95% concordance with GERMLINE in application to real multi-ethnic data[66]), or methods that rely on the positional Burrows-Wheeler transform (PBWT) data structure[17,67,68]. Several methods now exist to reconstruct gene genealogies in large samples[69–72]. Two recent methods substantially improved the scalability of this type of analysis, but they either focus on data compression, relying on fast heuristics to achieve scalability at the cost of deteriorating accuracy in sparse array data[71], or employ further modeling that requires sequencing data[72], with a computational cost that is quadratic in sample size (the same computational complexity required to run the full ASMC algorithm on all sample pairs). It is of continued interest to identify synergies between fast heuristic and accurate probabilistic approaches, which can lead to computationally efficient methods that remain robust to real world data heterogeneity. Although in this work we focused on large modern biobanks comprising SNP array data, sequencing datasets are quickly becoming available. FastSMC may be tuned to enable the analysis of sequencing data as well. Finally, looking at downstream applications, a direction of future work will be to leverage FastSMC to better control for subtle population stratification for both rare and common variants in association studies. Our results show that geographic coordinates can be effectively inferred from recent IBD sharing, and suggest that this may be a path towards capturing subtle environmental covariates[30] that are missed by genome-wide IBS-based approaches. FastSMC's output could thus account for subtle stratification even when the non-genetic confounder has a small and sharp distribution, where methods such as genomic control, PCA, or mixed models have limited efficacy[73].

Methods

FastSMC identification step. FastSMC's identification step leverages genotype hashing, a strategy that was introduced by the GERMLINE algorithm[19] to obtain substantial gains in computational scalability in the detection of pairwise shared IBD segments. This approach restricts the search space of IBD pairs to those that have small, identical shared segments, which are then extended to long segments with some tolerance. This in turn reduces the IBD search from all pairs of individuals to the subset of pairs that produce a hash collision at a given segment plus the cost of hashing the genotype data, which is linear in sample size and genome length, thus dramatically reducing the cost of IBD detection. However, a limitation of this strategy is that certain short haplotypes can be extremely common in the population and result in hash collisions across a large fraction of samples, effectively reverting back to a nearly all-pairs analysis and monopolizing computation time (Supplementary Fig. 1, gray). These common haplotypes are likely due to recombination cold-spots and typically contain little variation to classify shared segments. The majority of computation is thus spent processing regions with the least information content. The GERMLINE2 algorithm, which we developed in this work, proposes an adaptive hashing approach that adjusts to local haplotype complexity to dramatically reduce computational and memory requirements. GERMLINE2 proceeds as follows: (1) the input haplotype data is divided into small windows containing 16 or 32 SNPs each (depending on memory architecture); for a given window w, (2) all haplotypes are converted to binary sequences and efficiently hashed into bins of identical segments; (3) for each bin that contains more individuals than a fixed threshold (i.e. a low complexity bin) step 2 is recursively performed for window $w+1$ until no more low complexity bins are found; (4) all pairs of individuals sharing within a bin are then recorded in a separate hash table that stores putative segments; (5) pairs of individuals sharing contiguous windows that are sufficiently long are reported for validation. The primary computation speed-up comes from the recursive hashing step, which requires haplotypes to be sufficiently diverse before they are explored for pairwise analysis and stored (Supplementary Fig. 1, green). To allow for possible phasing errors, a putative shared segment is maintained through a parameterized number of non-identical windows, and the total number of non-identical windows within the segment is also reported for filtering. Most phasing errors either appear as blips, where a phase switch is immediately followed by a switch back, or by single switches followed by long stretches of accurate phase[14,15] – both of which are permitted by allowing periodic non-matching along the putative IBD segment. This permissive treatment of phasing is further filtered in the validation step (below). GERMLINE2 thus does not require any backtracking and only a small number of physical windows need to be stored in memory at any time (only enough to perform the recursion), allowing the method to run on input data of unlimited length.

FastSMC validation step. Every segment detected by GERMLINE2 in the identification step is added to a buffer of candidate segments. These segments are immediately decoded by the ascertained sequentially Markovian coalescent (ASMC) algorithm[23] once the buffer is full. ASMC is a coalescent-based HMM[24,25,27,74] that estimates the posterior of the coalescence time, or TMRCA, for a pair of individuals at each site along the genome using either sequencing or SNP

array platforms. It leverages a demographic model as prior on the TMRCA, which increases accuracy in detecting regions of low TMRCA[26], but would be infeasible to apply to the analysis of all pairs of genomes, and is thus only applied with the goal of validating previously identified candidate IBD regions. The hidden states of the HMM are discretized intervals, corresponding to a user-specified set of TMRCA intervals. The HMM emissions probabilities correspond to the probabilities of observing both the genotypes of the pair of analyzed individuals and the frequencies of mutations along the sequence, given the pair's TMRCA at each site, and the frequency of the allele[27]. The HMM transitions between hidden states correspond to changes in TMRCA along the genome due to recombination events, based on the conditional Simonsen-Churchill model[75,76]. The demographic history of the analyzed haplotypes is first estimated using other methods[25,27] and provided in input, so that the initial state distribution, the transition, and the emission probabilities can be computed. The most likely posterior sequence of TMRCAs along the genome is inferred using a dynamic programming approach that requires computing time linear in the number of hidden states, leading to a substantial speed-up over an HMM's standard forward-backward algorithm, which scales quadratically in the number of hidden states. When decoding the buffer of candidate segments, ASMC computes the posterior of the coalescence time for each candidate segment and each site from the minimum starting position to the maximum ending position in the buffer. At each site, if the posterior of coalescence time being between present time and the user-specified time threshold is higher than its prior, the site is considered to be IBD and the IBD segment is extended to the next site if the same condition is still satisfied, obtaining multiple IBD segments, all shorter than the original IBD candidate segments. The average probability of the TMRCA being between present time and the user-specified time threshold is computed over all sites until the segment breaks. This average probability corresponds to an IBD quality score: the higher it is, the more likely the segment is IBD. Each segment is also associated with an age estimate corresponding to the average MAP along the segment. FastSMC finally outputs each IBD segment with its corresponding IBD quality score and age estimate.

FastSMC parallelization. We divided the genome in 39 autosomal regions from different chromosomes or separated by centromeres[23]. This enabled us to efficiently parallelize analyses on the UK Biobank dataset, and prevented issues due to low marker density in centromeres. In addition, FastSMC enables a user to parallelize the analysis by specifying the total number of computing jobs to be run in parallel. If the user requests to run K independent parallel jobs, the total number of $\binom{N}{2}$ sample pairs to be analyzed are subdivided into K groups of approximately equal number of pairs. More in detail, the samples are divided into $S = \frac{1+\sqrt{1+8K}}{2}$ disjoint sets of approximately the same size, where K is such that S is an integer. Each parallel job then processes a unique pair of sets $\{i, j\}$ for $1 \le i \le j$ and $1 \le j \le S$. Note that only samples in sets i and j are loaded in memory, which leads to substantially lower memory footprint when large data sets are processed. IBD detection is then only performed for the $N_i \times N_j$ pairs across sets i and j if $i \neq j$, or for the $\binom{N_i}{2}$ pairs within set i if $i = j$. We leveraged this approach to extensively parallelize the analysis of the UK Biobank dataset.

Simulations. Unless otherwise specified, all simulations use the setup of Palamara et al.[23], which is described in this section. We used the ARGON simulator (v.0.1.160615)[77], incorporating recombination rates from a human chromosome 2 and a recent demographic model for European individuals (Northern European [CEU] population[27]). For each dataset, we simulated 300 haploid individuals and a region of 30 Mb. To simulate SNP array data, we subsampled polymorphic variants to match the genotype density and allele frequency spectrum observed in the UK Biobank dataset. We used recombination rates from the first 30 Mb of chromosome 2 (average rate of 1.66 cM per Mb). No genotyping or phasing error was introduced in our simulations. We simulated one dataset following this setup to fine-tune parameters, and 10 other datasets (all with different seeds) for accuracy benchmarking. The demographic model and genetic map used to simulate the data were used when running FastSMC, unless otherwise specified. When testing FastSMC's robustness to demographic model misspecification, we simulated data under a constant population size of 10,000 diploid individuals, but ran FastSMC assuming a European demographic model.

Accuracy evaluation. We compared FastSMC to the most recent published software version available for existing methods at the time we conducted this analysis: germline-1-5-2, refined-ibd.23Apr18.249 and RaPID_v.1.2. Throughout this work, we define a genomic site to be shared IBD by a pair of phased haploid individuals if their TMRCA at the site is lower than a specified time threshold (e.g. 50 generations). This is a natural definition for IBD sharing, as it is closely related to several other quantities that are of interest in downstream analyses, such as genealogical relatedness or the probability of sharing rare genomic variants. We note, however, that a number of other definitions can be found in the literature[22]. This is often due to the fact that current IBD detection algorithms cannot effectively estimate the TMRCA of a putative IBD segment. Downstream analyses of shared segments (e.g. refs. [6,9]) thus often resort to using the length of detected segments as a proxy for its age, since a segment's length is expected to be inversely proportional to its TMRCA. We benchmarked all methods using several such time thresholds (25, 50, 100, 150, and 200 generations), testing all polymorphic sites for all pairs of genomes in the simulated data, across 10 coalescent simulations. Accuracy was quantified using the area under the precision-recall curve (auPRC), which effectively addresses issues with class imbalance that are expected in this analysis due to the low prevalence of IBD sites compared to non-IBD sites. Precision represents the fraction of identified sites that are indeed IBD (following the TMRCA-based definition of IBD), and recall represents the fraction of true IBD sites that are successfully identified. Particularly, for a given IBD time threshold, a site inferred to be IBD by one of the methods was considered correct (true positive) if the true TRMCA at this site was indeed below the specified IBD time threshold, and incorrect (false positive) if the true TRMCA at this site was above the IBD time threshold. Similarly, a site that was not reported to be IBD by the tested method was considered correct/incorrect (true/false negative) if the true TMRCA at the site was found to be below/above the IBD time threshold. We used these definitions to compute the precision and recall values for all methods. Each method presents different parameters, which can be used to tune precision and recall, e.g. by allowing a more or less permissive detection of IBD segments. String-matching methods (GERMLINE and RaPID) report long, approximately IBS regions and do not produce calibrated estimates of segment quality. A commonly used proxy for the likelihood of a detected IBS segment being IBD is its length, with longer IBS segments being more likely to be IBD than shorter ones. In lack of an interpretable measure of accuracy, precision and recall for the output of GERMLINE and RaPID was thus tuned by using different segment length cutoffs. RefinedIBD and FastSMC, on the other hand, both provide an explicit quantification of segment quality. RefinedIBD outputs a LOD score for each segment, while FastSMC computes a segment's IBD quality score, which is the posterior of the TMRCA being between present time and the user-specified time threshold. We thus used LOD score and IBD quality score to tune precision and recall of RefinedIBD and FastSMC, respectively. Each method presents a number of additional parameters, which we further optimized using a grid-search, so that each method can be run with a set of parameters that is as close to optimal as possible. Despite the extensive tuning, not all accuracy values could be explored by all methods. Namely, some recall values cannot be achieved using realistic parameters, due to factors such as the minimum allowed LOD parameter for RefinedIBD, the time discretisation introduced in FastSMC and the minimum length parameter for all methods. We thus evaluated all algorithms by restricting the comparison to the range of recall values that could be achieved by all methods, which we refer to as common recall range. Furthermore, some of these parameters affect the speed of each algorithm. The parameters we chose for comparing methods are optimized for maximum accuracy, although we avoided parameters that would result in degeneracies (e.g. the minimum length in FastSMC's identification step could be set to values below 0.1 cM, effectively disabling this step and reverting to a pairwise ASMC analysis, which would lead to a higher accuracy and larger recall range, at the cost of unreasonable computation).

Fine-tuning of methods. For each method (FastSMC, GERMLINE[19], RefinedIBD[20], and RaPID[21]) and each IBD time threshold (25, 50, 100, 150, and 200 generations), we performed a grid-search over possible parameter values to optimize the accuracy on one simulated dataset and select the best set of parameters. For each method, we then explored the obtained set of parameters to make the algorithm faster while negligibly compromising the accuracy (in most cases, this resulted in a slightly smaller recall range but with a substantial gain in speed). We finally used 10 independent simulated datasets to validate the accuracy and the UK Biobank dataset to measure running time and memory usage (Supplementary Figs. 4–5). Unless otherwise specified, the parameters presented in Supplementary Table 1 were used in all analysis.

Computing confidence intervals. Unless otherwise indicated, confidence intervals (CIs) were computed by bootstrap using 39 genomic regions as resampling unit. The use of genomic regions as resampling unit, rather than individuals, ensures that approximately independent bootstrap replicates are utilized. These 39 genomic regions were obtained by dividing the genome (autosomal chromosomes only) in regions from different chromosomes or separated by centromeres.

Estimating the age of an IBD segment. A common way to estimate the age of an IBD segment is to use its length to obtain a maximum likelihood estimator (MLE). When the time in generations g to the most recent common ancestor is known, the total length of a randomly chosen shared IBD segment follows an exponential distribution with rate $1/2g$ per Morgan[6]. The likelihood function is thus given by $\mathcal{L}(g) = (2g)e^{-2lg}$, where l denotes the segments length in Morgans. As $\frac{d\mathcal{L}}{dg}(g) = 2e^{-2lg}(1 - 2lg)$, the MLE is given by $\hat{g} = \frac{1}{2l}$. FastSMC does additional modeling and provides a different age estimate, which consists in the average MAP estimate of the TMRCA along the segment. We sometimes report segment age estimates in years, rather than generations, assuming 30 years per generation.

Effective population size estimate. Assuming a constant effective size N_e, the probability of finding a common ancestor at a given site for a pair of individuals is exponentially distributed with mean N_e. The probability of finding a common ancestor before any time threshold T is thus given by $\mathrm{P}(t_{\mathrm{TMRCA}} \le T|N_e) = \int_0^T \frac{1}{N_e} e^{\frac{-t}{N_e}} \mathrm{d}t = 1 - e^{\frac{-T}{N_e}} \simeq \frac{T}{N_e}$ for $N_e \to \infty$ i.e $N_e = \frac{T}{\mathrm{P}(t_{\mathrm{TMRCA}} \le T|N_e)}$ for $T > 0$. We estimate this probability using the fraction of genome shared by IBD segments denoted by f_T. Let Γ denote the set of sites along the genome and θ the demographic model. $\mathbb{E}(f_T|\theta) = \mathbb{E}\left(\frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma} 1_\theta(t_{\mathrm{TMRCA}} \text{ at site } \gamma)\right) = \frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\mathbb{E}(1_\theta(t_{\mathrm{TMRCA}} \text{ at site } \gamma)) = \frac{1}{|\Gamma|}\sum_{\gamma\in\Gamma}\mathrm{P}(t_{\mathrm{TMRCA}} \le T \text{ at site } \gamma|\theta) = \mathrm{P}(t_{\mathrm{TMRCA}} \le T|\theta)$. We finally estimate the effective population size N_e using $\hat{N}_e = \frac{T}{\overline{f_T}}$ for any $T > 0$, where $\overline{f_T}$ is the sample mean for the fraction of genome shared f_T. Note that, for simplicity, we infer a single aggregate effective population size across the past T generations rather than comparing more complex demographic models.

UK Biobank dataset and definition of postcodes. The UK Biobank cohort[2] contains 487,409 samples, which were phased at a total of 678,956 autosomal biallelic SNPs using Eagle2[15]. In all, 49,960 of these individuals were also exome-sequenced resulting in ~4 million polymorphic variants, 98.4% of which have frequency <1%[49]. We used the same sets of unrelated individuals ($N = 407{,}219$) and individuals of self-reported and inferred White British ancestry ($N = 408{,}974$) as defined by Bycroft et al.[2]. Related individuals refer here to ≤3rd degree relatives, e.g. first degree cousins, estimated using the software KING[78]. The UK Biobank cohort predominantly contains samples of White British ancestry[2] and the inclusion of non-White British samples did not result in substantial biases for analyses of population structure and natural selection. We thus decided not to exclude samples based on ancestry, unless otherwise specified, so that results are representative of the average UK Biobank participant. FastSMC was run on the UK Biobank data set using the optimal set of parameters for IBD detection within the past 50 generations, using the demographic model for CEU individuals inferred by Terhorst et al.[27]. We note that several candidate demographic models could be adopted[9] and that simultaneous estimation of IBD sharing and demographic model are an attractive direction of future investigation.

Birth coordinates (all within the UK) were available for 432,968 individuals in the cohort in the Ordnance Survey Great Britain 1936 (OSGB36) Eastings and Northings system. We refer to these coordinates as X and Y coordinates, which can be converted into longitude and latitude. Home addresses at assessment were available for 482,832 samples in the OSGB36 Eastings and Northings system. We refer to these coordinates as home addresses. We analyzed population structure using 120 postcodes in the UK, only looking at the first one or two letters indicating the city or region. Postcodes BT, BF, BN, and CR were not included due to lack of samples.

Hierarchical clustering. We constructed a similarity matrix for all individuals with birth coordinates in the UK Biobank dataset (432,968 samples) using the sharing of IBD segments within the past 10 generations. This resulted in a sparse 432,968 × 432,968 matrix, where entry (i, j) corresponds to the fraction of genome shared by common ancestry in the past 10 generations between individuals i and j. We computed the largest connected component of this matrix, which comprised all but 102 individuals, which we excluded from further analysis. We then applied an Agglomerative Hierarchical Clustering algorithm for Sparse Similarity Matrices using average linkage with the sparseAHC library. We obtained a dendrogram with 432,866 leaves (one for each sample) and a single root. Each node is annotated with a distance from the root, ranging from 0 (the root itself) to 1 (the leaves). A node's distance from the root corresponds to the fraction of genome shared by individuals whose TMRCA is such a node. To visualize clustering of individuals in Fig. 2, we cut the tree at increasingly large distances from the root, corresponding to increasingly fine-grained clusters in terms of both genetic and geographic proximity of the samples. In each case, we only highlight large clusters containing at least 500 individuals. To plot results, we divided each submap into 10,080 grid cells (80 lines along the X-axis and 126 lines along the Y-axis). In each cell, we computed the most represented cluster (i.e the cluster with the largest number of individuals in that cell) and individuals from that cluster are shown in the corresponding color. The transparency of all points within a cell (ranging from 0 to 1) was set to the fraction of individuals from that cell corresponding to the most represented cluster. All light gray dots correspond to individuals that are either in clusters containing <500 samples or part of a cluster different from the one represented in the grid cell. Including/excluding individuals of non-White British ancestry has a negligible impact on this analysis, as most of these samples belong to smaller clusters not represented in Fig. 2 (in light gray).

K-NN prediction of birth location. We randomly sampled two subsets of 10,000 individuals each from the UK Biobank cohort and used the sharing of recent common ancestors in the past 600 years with the remaining 412,866 individuals in the UK Biobank dataset to predict their birth locations, applying the K Nearest Neighbors algorithm. One dataset was used to find the optimal value for the parameter K, while the second one was used to validate the results (details shown in Supplementary Fig. 16). We computed pairwise genetic similarity across individuals using either FastSMC-estimated pairwise IBD sharing, or using a standard estimate of kinship based on genome-wide allele sharing. This kinship estimator was obtained by computing the product $\mathbf{X}\mathbf{X}^\mathrm{T}$, where $\mathbf{X}$ is the $N \times S$ genotype matrix ($N = 432{,}866$ samples and $S = 716{,}175$ autosomal SNPs), standardized to have mean 0 and variance 1 for each column. Sharing of very close relatives is highly informative of geographic proximity. We thus excluded ≤3rd degree relatives[2] from the dataset to bypass this source of information and test the generality of this approach. Prior to the exclusion of close relatives, the IBD-based predictor obtained an average error of 86 km (95% CI = [83,88], optimal $K = 1$), while the allele sharing distance predictor obtained an average mean error of 118 km (95% CI = [115,121], optimal $K = 1$). After removing close relatives (which brought sample size down to 8226), the IBD-based predictor obtained an average error of 95 km, 95% CI = [93,97], optimal $K = 5$ compared to 137 km, 95% CI = [135,139], optimal $K = 5$ for allele sharing. Note that this analysis focused on mean error, which is larger than median error (e.g. as in Fig. 3) due to the presence of outliers.

We regressed the true X (resp. Y) birth coordinates on the predicted X (resp. Y) birth coordinates using either IBD sharing in the past 20 generations allele sharing, for the set of 8226 random samples we obtained after excluding 3rd degree relatives[2]. The estimated coefficient for the IBD-based predictor was 0.91, 95% CI = [0.88, 0.94], (resp. 0.96, 95% CI = [0.94, 0.98]), substantially larger than the estimated coefficient for the allele sharing predictor (0.12, 95% CI = [0.09,0.16]; resp. 0.12, 95% CI = [0.09, 0.15]). Finally, after excluding close relatives, we computed the correlation between true and predicted coordinates for both methods, obtaining a stronger correlation when using IBD sharing ($r = 0.6$ for X coordinate and $r = 0.74$ for Y coordinate) than when using allele sharing ($r = 0.31$ for X coordinate and $r = 0.43$ for Y coordinate, respectively). Correlation for IBD sharing was also stronger without removing close relatives ($r = 0.63$ for X coordinate and $r = 0.77$ for Y coordinate, compared to $r = 0.40$ and $r = 0.42$ respectively for allele sharing). We did not exclude individuals from these analyses based on their inferred or reported ancestry, so that these results are reflective of the average UK Biobank participant. Prediction accuracy for non-White British individuals will be lower due to a smaller sample size. We thus expect that restricting these analyses to White British samples will lead to slightly improved accuracy.

Detection of recent positive selection. In order to identify genomic regions with an usually high density of coalescence times, we computed the Density of Recent Coalescence (DRC_T) statistic within the past T generations[23]. FastSMC does not output the posterior of the TMRCA but provides an IBD quality score, corresponding to the sum of posterior probabilities between generations 0 and T, where T is the user-specified threshold. As the UK Biobank dataset was analyzed for $T = 50$, the DRC_{50} statistic at a given site along the genome was estimated by averaging all IBD quality scores obtained from all analyzed pairs of samples (assuming a score of 0 if no segment is present for a pair). The DRC_{50} statistic reflects the probability that a random pair of individuals coalesced at a given genomic site during the past 50 generations. We averaged it within windows of 0.05 cM along the genome. Results are presented in Fig. 4 and Supplementary Table 4.

Given n samples from a population of recent effective size N, the DRC_{50} statistic is approximately Gamma-distributed under the null for $n \ll N$[23]. We then built an empirical null model using the database of regions under positive selection used by Palamara et al.[23]. We fitted a Gamma distribution (using the Scipy library[79]) to the estimated DRC_{50} values within putative neutral regions (after excluding the regions of known positive selection and 500 kb windows around them), and used this model to obtain approximate one-sided p-values throughout the genome. We analyzed 52,003 windows, using a Bonferroni significance threshold of $0.05/52{,}003 = 9.6 \times 10^{-7}$. When multiple candidate genes were found, we only retained the one nearest to the top SNP (i.e with smallest p-value). Three of the genome-wide significant signals we detected (*MRC1* locus, chr10:17.43-18.10 Mb; *HYDIN* locus, chr16:70.10-72.69 Mb, and *EFTUD2* locus, chr17:41.84-44.95 Mb) fell within the putative neutral regions of the genome. We thus iterated this procedure, excluding these loci from the set of putative neutral loci. Once again, one of the genome-wide significant loci (*BANP* locus, chr16:88.25-88.48 Mb) overlapped with the putative neutral regions. We excluded this locus and iterated the procedure again. Results from the empirical null model fitting are presented in Supplementary Fig. 18. Finally, we verified that the genome-wide significant peaks detected using this approach are not found in regions of extremely high recombination rate or low marker density, which may introduce systematic biases in IBD detection Supplementary Table 5. This analysis was replicated after excluding samples of non-White British ancestry but resulted in the same set of genome-wide significant loci.

We note that computing significance for the DRC_{50} statistic using this approach relies on several simplifying assumptions. Specifically, we assume that all pairs of individuals coalesce independently within the past 50 generations, which is not conservative and may be violated in practice. On the other hand, we fit the null model using real data, which likely contains regions under weak selection, and neglect the correlation across DRC estimates in different bins, applying a Bonferroni correction; both are conservative choices. Overall, using this approach to assess significance for the DRC statistic within small time windows was observed

to not lead to significant inflation[23] (Supplementary Fig. 18), although the resulting approximate p-values should be interpreted with care.

Association analyses. We used each IBD segment between exome-sequenced LoF carrying individuals and non-sequenced individuals as a surrogate for the latter carrying an untyped LoF mutation, which we then tested for association with phenotype. For a given gene and a given non-sequenced individual, we define a LoF-segment as any IBD segment shared with an exome-sequenced LoF mutation carrier. We then compute a LoF-segment burden for each individual as the sum of probabilities (IBD quality scores) of all LoF-segments involving that individual, under the assumption that increased IBD probability and incidence corresponds to increased probability of sharing the LoF variant. Finally, this burden is tested for association with each target phenotype (rank-based inverse normal transformed) in a linear regression with covariates for age, sex, BMI, smoking status, and four principal components (two-sided test), similarly to the study from Van Hout et al.[49].

Although this test captures uncertainty about the sharing of IBD segments through the use of IBD quality scores, it makes use of all LoF-segments, regardless of their age. As a result, it may be suboptimal in cases where the LoF arose after the TMRCA, for which a LoF-segment is independent of underlying LoF sharing, and thus do not contribute signal to the burden test. We thus augmented the LoF burden test by separately considering only LoF-segments older than a specified threshold. For each gene, we divided all LoF-segments into deciles based on IBD quality score. For instance, segments with IBD quality scores in the tenth decile (which corresponds to the IBD quality score interval [0.47, 1]), strongly suggest the sharing of common ancestors that lived recently and have therefore transmitted extremely recent variation. We then constructed 10 separate LoF-segment burdens, with increasingly more stringent quality score cutoffs (referred as time transformations in our analysis), and performed 10 association tests for each gene, taking the test that resulted in the lowest p-value after adjusting significance thresholds by conservatively assuming independence for all tests. Not all genes contained shared LoF-segments for testing, which reduced the total number of tested genes to 14,249. This resulted in a Bonferroni-corrected exome-wide significance threshold of $0.05/(10 \times 14{,}249)$ for our LoF-segment burden analysis. We note that this approach does not rely on phasing information and that the IBD segment shared with a carrier is equally likely to involve or not the haplotype that harbors the LoF variant. In principle it may be possible to leverage phasing information to increase the accuracy of this approach. Phasing of rare exome sequencing variants is however challenging[80] and we do not attempt that in this work.

Although gene-based burden tests are meant to implicate specific genes with a known directional effect on the trait, the observed signal may not always be driven by a causal variant, and instead be due to tagging of causal variants in nearby genes. In this case, it is possible that the underlying rare causal variant is tagged by a common variant, which may have been detected in a previous GWAS, leading to synthetic associations – an effect similar to that hypothesized in the context of common variant association[81]. In particular, these common variants may provide better tagging of the underlying true causal variation than our LoF-segment burden score, and would thus remove or significantly reduce the association signal if included as covariates in the test. Based on this principle, for each gene and each trait, we selected up to three genotyped SNPs that were in proximity (±1 Mb from the gene), which were significantly ($p < 1 \times 10^{-8}$) associated by Loh et al.[82], and used them as covariates. We observed that this approach often improves the association signal (Supplementary Fig. 20), removing signals that were likely caused by tagging common variants. We refer to analyses that include top associated SNPs as covariates as SNP-adjusted, for either the LoF-segment or WES-LoF burden test; results without the SNP-adjustment are shown in Supplementary Fig. 20 and Supplementary Table 7.

We validated our approaches, both LoF-segment and not SNP-adjusted LoF-segment, which seek to implicitly impute ultra-rare LoF variation between sequenced and non-sequenced individuals, by testing for association between rare variation and seven blood-related phenotypes recently analyzed by Van Hout et al.[49], and comparing to the results of that same study. However, because summary association statistics for this analysis are not available, we performed our own exome-wide burden testing. Specifically, we used the same testing framework we used in our LoF-segment burden analysis to test for association between phenotypes and burden of LoF variants within a gene in exome-sequenced individuals, adjusting for the same covariates and using the same rank-based inverse normal transformation for the phenotype. We refer to this analysis as WES-LoF burden analysis.

Both LoF-segment burden and WES LoF burden analyses were restricted to unrelated individuals of White British ancestry, as defined by Bycroft et al.[2], and the LoF-segment burden analysis was further restricted to individuals for which exome sequencing data is not available. This resulted in 303,125 individuals for the two LoF-segment burden tests and 34,422 individuals for the WES LoF. Finally, we note that the UK Biobank has recently released a statement regarding incorrectly mapped variants in the 50k WES Functionally Equivalent (FE) dataset (available online at https://www.ukbiobank.ac.uk/wp-content/uploads/2019/12/Description-of-the-alt-aware-issue-with-UKB-50k-WES-FE-data.pdf.), which we however believe did not introduce any significant biases in our analyses.

Applying our WES-LoF burden analysis, we detected 10 out of 14 exome-wide significant associations also reported by Van Hout et al.[49]. We also detected three additional associations that were not reported by Van Hout et al.[49]: *MAPK8* and *APOC3* with red blood cell distribution width (two-sided t-test $p = 1.33 \times 10^{-6}$ and $p = 2.13 \times 10^{-7}$ respectively), and *TET2* with eosinophil count (two-sided t-test $p = 1.79 \times 10^{-8}$). Results are summarized in Fig. 5b. These differences are likely ascribed to the slightly different testing strategy we adopted, e.g. the use of a linear model, rather than a linear mixed model, and the exclusion of related samples. Detailed results for these analyses are reported in Supplementary Tables 7 and 8. A QQ-plot verifying the calibration of our test for the SNP-adjusted LoF-segment burden analysis is shown in Supplementary Fig. 26 and, as explained at that point, rare variant stratification is likely to be included in our results[73] but addressing this issue goes beyond the scope of the current study.

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

Data availability

The main data items presented in this manuscript (simulation data, UK relatedness matrices, DRC selection annotations, association summary statistics) are available on Zenodo (DOI: 10.5281/zenodo.4012676)[83]. The interactive map with results on population structure is publicly available online at https://ukancestrymap.github.io/. Genomic data sets and annotations analysed in this study include: UK Biobank http://www.ukbiobank.ac.uk/, genetic maps ftp://1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110106_recombination_hotspots/, gene coordinates http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz, natural selection annotations ftp://jjwanglab.org/dbPSHP/curation/dbPSHP_20131001.tab, association summary statistics https://data.broadinstitute.org/alkesgroup/UKBB/, allele age estimates http://human.genome.dating, population sizes in UK regions https://www.nomisweb.co.uk/census/2011/qs102ew and https://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html. Data analyses are based on open-source libraries and software programs that are available online: Scipy[79], Matplotlib[84], NumPy[85,86], Pandas[87,88], Seaborn[89], BCFtools (http://samtools.github.io/bcftools/), Basemap (https://matplotlib.org/basemap/index.html) and sparseAHC (https://github.com/khabbazian/sparseAHC/).

Code availability

The FastSMC software is freely available at https://github.com/PalamaraLab/FastSMC.

Received: 2 April 2020; Accepted: 2 October 2020;
Published online: 30 November 2020

References

1. All of Us Research Program Investigators. The All of Us research program. *N. Engl. J. Med.* **381**, 668–676 (2019).
2. Bycroft, C. et al. The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
3. Marx, V. The DNA of a nation. *Nature* **524**, 503–505 (2015).
4. Gaziano, J. M. et al. Million veteran program: a mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
5. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
6. 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**, 809–822 (2012).
7. Palamara, P. F. & Pe'er, I. Inference of historical migration rates via haplotype sharing. *Bioinformatics* **29**, i180–i188 (2013).
8. Ralph, P. & Coop, G. The geography of recent genetic ancestry across europe. *PLoS Biol.* **11**, e1001555 (2013).
9. Browning, S. R. & Browning, B. L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *Am. J. Hum. Genet.* **97**, 404–418 (2015).
10. Albrechtsen, A., Moltke, I. & Nielsen, R. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* **186**, 295–308 (2010).
11. Gusev, A. et al. The architecture of long-range haplotypes shared within and across populations. *Mol. Biol. Evol.* **29**, 473–486 (2011).
12. Browning, S. R. & Thompson, E. A. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics* **190**, 1521–1531 (2012).
13. 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).
14. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
15. Loh, P.-R. et al. Reference-based phasing using the haplotype reference consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

16. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
17. Rubinacci, S., Delaneau, O. & Marchini, J. Genotype imputation using the positional burrows wheeler transform. Preprint at https://www.biorxiv.org/content/10.1101/797944v2 (2019).
18. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010).
19. Gusev, A. et al. Whole population, genomewide mapping of hidden relatedness. *Genome Res.* **19**, 318–26 (2008).
20. Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013).
21. Naseri, A., Liu, X., Tang, K., Zhang, S. & Zhi, D. Rapid: ultra-fast, powerful, and accurate detection of segments identical by descent (ibd) in biobank-scale cohorts. *Genome Biol.* **20**, 143 (2019).
22. Wakeley, J. & Wilton, P. R. *Coalescent and Models of Identity By Descent*. Vol. 1, 287–292 (Academic Press, Oxford, 2016).
23. Palamara, P. F., Terhorst, J., Song, S. & Price, A. L. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nat. Genet.* **50**, 1311–1317 (2018).
24. McVean, G. A. & Cardin, N. J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. B Biol. Sci.* **360**, 1387–1393 (2005).
25. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
26. Tataru, P., Nirody, J. A. & Song, Y. S. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics* **30**, 3430–3431 (2014).
27. 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).
28. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* 8, e1002453 (2012).
29. Leslie, S. et al. The fine-scale genetic structure of the british population. *Nature* **519**, 309–314 (2015).
30. Haworth, S. et al. Apparent latent structure within the uk biobank sample has implications for epidemiological analysis. *Nat. Commun.* **10**, 1–9 (2019).
31. Edge, M. D. & Coop, G. Attacks on genetic privacy via uploads to genealogical databases. *eLife* **9**, e51810 (2020).
32. Erlich, Y., Shor, T., Pe'er, I. & Carmi, S. Identity inference of genomic data using long-range familial searches. *Science* **362**, 690–694 (2018).
33. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
34. Yang, J. et al. Common snps explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
35. Chang, C. C. et al. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742–015 (2015).
36. Jones, B. & Mattingly, D. *An Atlas of Roman Britain: An Atlas of Roman Britain* (Oxbow Books, 1990).
37. Vandepoele, K., Van Roy, N., Staes, K., Speleman, F. & van Roy, F. A novel gene family NBPF: intricate structure generated by gene duplications during primate evolution. *Mol. Biol. Evol.* **22**, 2265–2274 (2005).
38. Barreiro, B. L. & Quintana-Murci, L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat. Rev. Genet.* **11**, 17–30 (2009).
39. Bersaglieri, T. et al. Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–20 (2004).
40. Fagundes, N., M Salzano, F., Batzer, M., Deininger, P. & Bonatto, S. Worldwide genetic variation at the 3'-utr region of the ldlr gene: possible influence of natural selection. *Ann. Hum. Genet.* **69**, 389–400 (2005).
41. Stahl, P. D. & Ezekowitz, R. A. B. The mannose receptor is a pattern recognition receptor involved in host defense. *Curr. Opin. Immunol.* **10**, 50–55 (1998).
42. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
43. Hata, S. et al. Calpain 8/nCL-2 and Calpain 9/nCL-4 constitute an active protease complex, g-calpain, involved in gastric mucosal defense. *PLoS Genet.* **6**, 1–14 (2010).
44. Li, Y. et al. Chd1l contributes to cisplatin resistance by upregulating the abcb1-nf-κb axis in human non-small-cell lung cancer. *Cell Death Dis.* **10**, 99 (2019).
45. Birot, A.-M. et al. Identification and molecular analysis of banp. *Gene* **253**, 189–196 (2000).
46. Raidt, J. et al. Recessive hydin mutations cause primary ciliary dyskinesia without situs abnomalities. *Eur. Respir. J.* **40**, P4808 (2012).
47. Lines, M. et al. Haploinsufficiency of a spliceosomal gtpase encoded by eftud2 causes mandibulofacial dysostosis with microcephaly. *Am. J. Hum. Genet.* **90**, 369–77 (2012).
48. Palamara, P. et al. Leveraging distant relatedness to quantify human mutation and gene-conversion rates. *Am. J. Hum. Genet.* **97**, 775–789 (2015).
49. Van Hout, C. V., Tachmazidou, I., Backman, J. D. et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature.* **586**, 749–756 (2020).
50. Consortium, G. P. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
51. Mathieson, I. & McVean, G. Demography and the age of rare variants. *PLoS Genet.* **10**, e1004528 (2014).
52. Cirulli, E. T. et al. Genome-wide rare variant analysis for thousands of phenotypes in over 70,000 exomes from two cohorts. *Nat. Commun.* **11**, 1–10 (2020).
53. Zhao, Z. et al. Uk biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. *Am. J. Hum. Genet.* **106**, 3–12 (2020).
54. Kiezun, A. et al. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS Genet* **9**, e1003301 (2013).
55. Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS Biol.* **18**, e3000586 (2020).
56. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
57. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
58. Kichaev, G. et al. Leveraging polygenic functional enrichment to improve gwas power. *Am. J. Hum. Genet.* **104**, 65–75 (2019).
59. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
60. Gieger, C. et al. New gene functions in megakaryopoiesis and platelet formation. *Nature* **480**, 201–208 (2011).
61. Kanai, M. et al. Genetic analysis of quantitative traits in the japanese population links cell types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
62. Wegmann, D. et al. Recombination rates in admixed individuals identified by ancestry-based inference. *Nat. Genet.* **43**, 847–853 (2011).
63. Hinch, A. G. et al. The landscape of recombination in african americans. *Nat. Genet.* **476**, 170–175 (2011).
64. Tian, X., Browning, B. L. & Browning, S. R. Estimating the genome-wide mutation rate with three-way identity by descent. *Am. J. Hum. Genet.* **105**, 883–893 (2019).
65. Gusev, A. et al. Low-pass genome-wide sequencing and variant inference using identity-by-descent in an isolated human population. *Genetics* **190**, 679–689 (2012).
66. Shemirani, R. et al. Rapid detection of identity-by-descent tracts for mega-scale datasets. Preprint at https://www.biorxiv.org/content/10.1101/749507v1 (2019).
67. Zhou, Y., Browning, S. R. & Browning, B. L. A fast and simple method for detecting identity by descent segments in large-scale data. *Am. J. Hum. Genet.* **106**, 426–437 (2020).
68. Durbin, R. Efficient haplotype matching and storage using the positional burrows–wheeler transform (pbwt). *Bioinformatics* **30**, 1266–1272 (2014).
69. Minichiello, M. J. & Durbin, R. Mapping trait loci by use of inferred ancestral recombination graphs. *Am. J. Hum. Genet.* **79**, 910–922 (2006).
70. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-wide inference of ancestral recombination graphs. *PLoS Genet.* **10**, e1004342 (2014).
71. Kelleher, J. et al. Inferring whole-genome histories in large population datasets. *Nat. Genet.* **51**, 1330–1338 (2019).
72. Speidel, L., Forest, M., Shi, S. & Myers, S. R. A method for genome-wide genealogy estimation for thousands of samples. *Nat. Genet.* **51**, 1321–1329 (2019).
73. Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.* **44**, 243 (2012).
74. 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).
75. Simonsen, K. & Churchill, G. A markov chain model of coalescence with recombination. *Theor. Popul. Biol.* **52**, 43–59 (1997).
76. Hobolth, A. & Jensen, J. L. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theor. Popul. Biol.* **98**, 48–58 (2014).
77. Palamara, P. F. ARGON: fast, whole-genome simulation of the discrete time wright-fisher process. *Bioinformatics* **32**, 3032–4 (2016).
78. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).
79. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272 (2020).
80. Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLoS Genet.* **14**, e1007308 (2018).

81. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
82. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
83. Nait Saada, J. et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations: data related to publication (2020).
84. Hunter, J. D. Matplotlib: a 2d graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
85. Oliphant, T. E. *A guide to NumPy*, Vol. 1 (Trelgol Publishing USA, 2006).
86. Van Der Walt, S., Colbert, S. C. & Varoquaux, G. The numpy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
87. The pandas development team. pandas-dev/pandas: Pandas (2020).
88. Wes, M. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, 56–61 (2010).
89. Waskom, M. et al. mwaskom/seaborn: v0.8.1 (September, 2017).

Acknowledgements

We thank Po-Ru Loh, Augustine Kong, Robert Davies, Simon Myers, Geoff Nicholls, Kevin Sharp, and Brian Zhang for helpful discussions on various aspects of this work; John Watts and David Russell for discussions on historical interpretation of our results. This work was conducted using the UK Biobank resource (Application #43206). We thank the participants of the UK Biobank project. This work was supported by the MRC grant MR/S502509/1 and the Balliol Jowett Scholarship (to J.N.S.); EPSRC and MRC grant EP/L016044/1 (to G.K.); NIH grant T32-GM74897 and T32-GM135117 (to D.S.); EPSRC grant D4D00010 (to F.C. and M.R.); NIH grant R21-HG010748-01 (to P.F.P. and A.G.); Wellcome Trust ISSF grant 204826/Z/16/Z (to P.F.P. and M.R.); and by ERC Starting Grant 850869 (to P.F.P.). We thank the Biomedical Research Computing team at the University of Oxford and the Research Computing Group at Harvard Medical School for support with the ResComp and O2 compute clusters. The research was partly supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z with additional support from the NIHR Oxford BR (the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health).

Author contributions

P.F.P. planned the study; P.F.P. and A.G. designed and supervised analyses; J.N.S., A.G., and P.F.P implemented the FastSMC algorithm; J.N.S. performed simulations and data analysis on population structure and natural selection; G.K. and D.S. performed data analysis on trait associations; F.C. and M.R. developed software tools for the FastSMC package and data visualization; J.N.S., G.K., A.G., and P.F.P. analyzed results and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19588-x.

Correspondence and requests for materials should be addressed to J.N.S. or P.F.P.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contributions to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

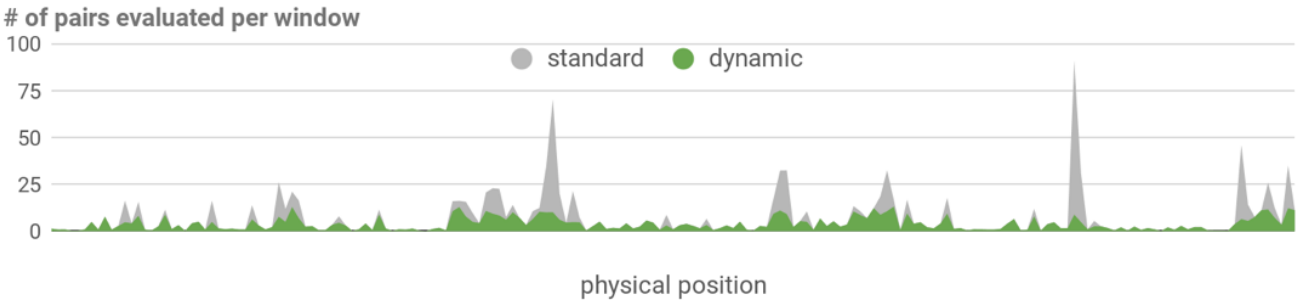
© The Author(s) 2020

Supplementary Information

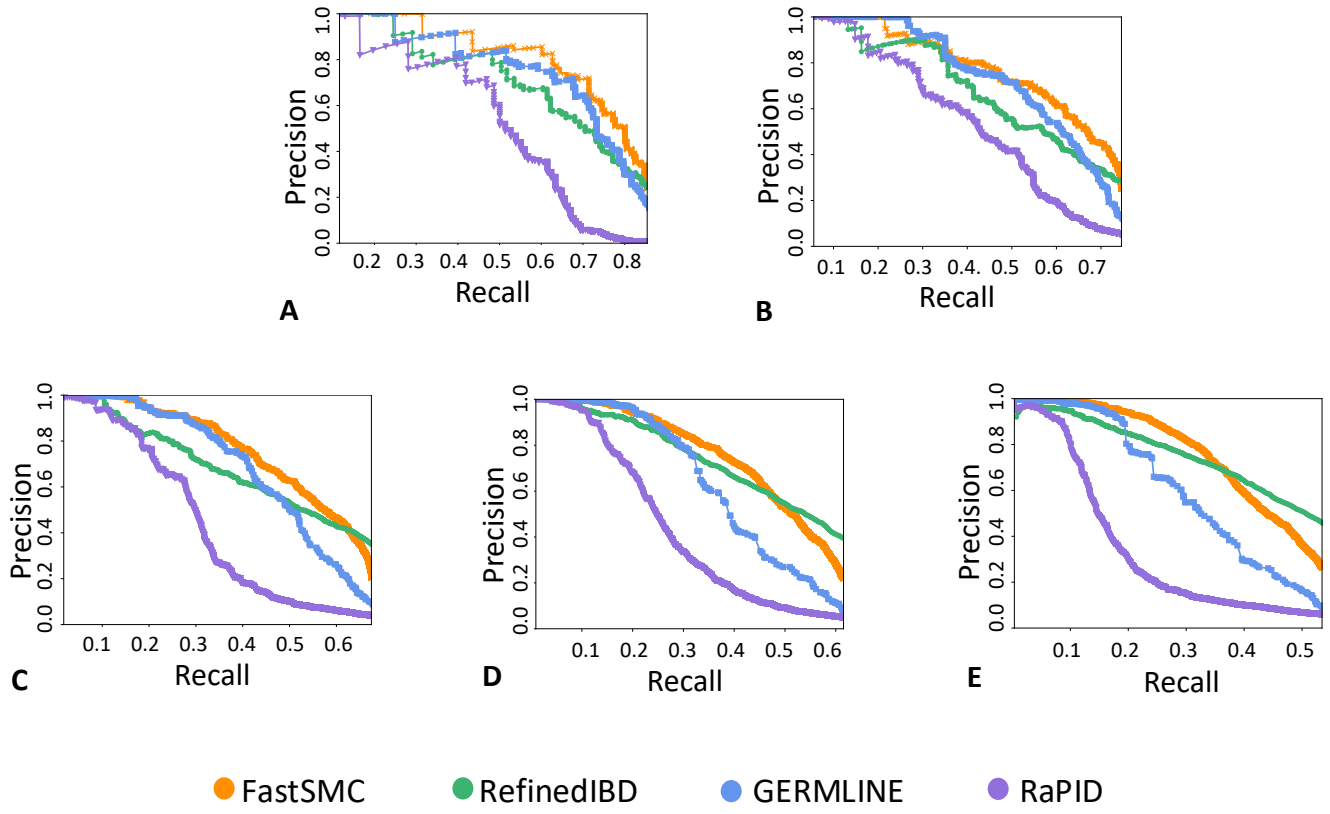
Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations

Nait Saada et al.

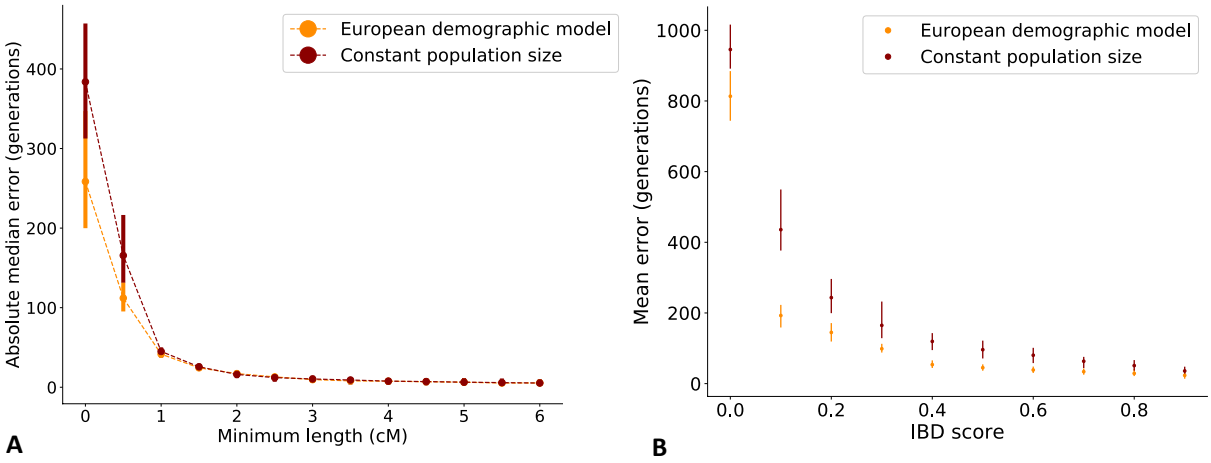
Supplementary Figures



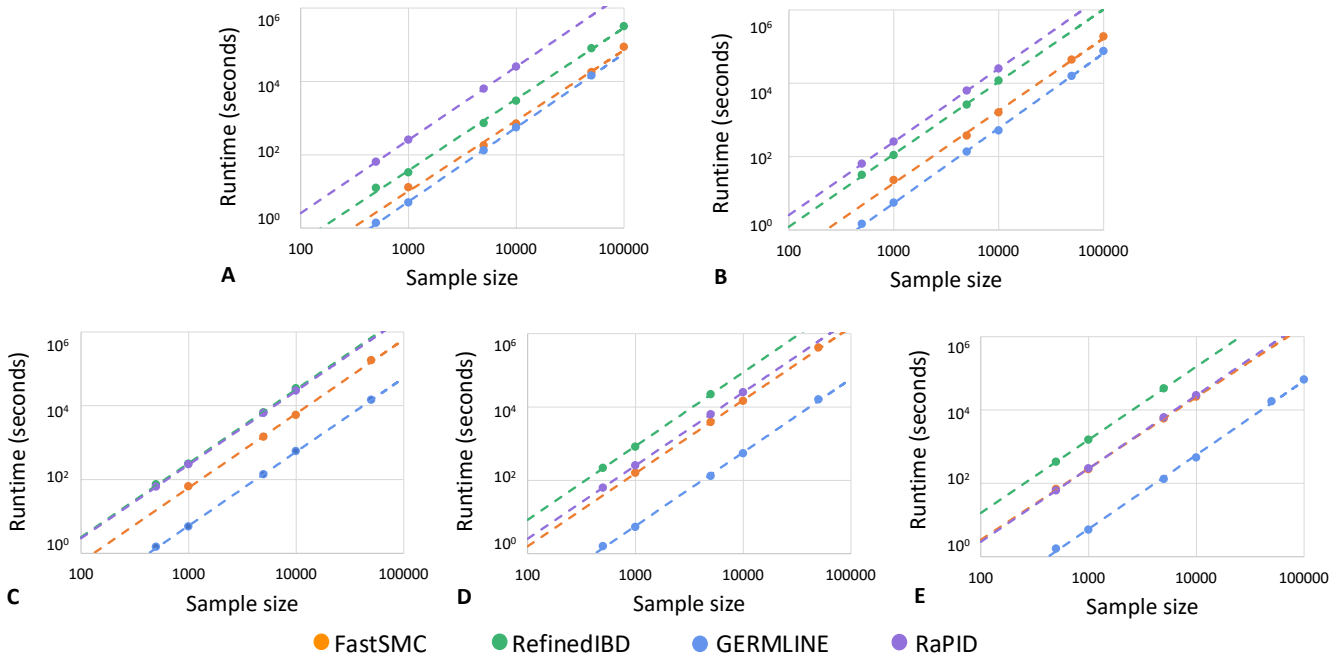
Supplementary Fig. 1. GERMLINE and GERMLINE2 comparison. Search complexity for standard vs dynamic hashing. Number of pairs of individuals (in millions) with identical haplotypes in each genome window using GERMLINE standard hash (gray) and the proposed GERMLINE2 dynamic hash (green). Analysis of 16,000 random UK Biobank samples from chromosome 22. Multiple regions of the genome exhibit sharing between more than 10% of all pairs, requiring extensive follow-up analysis in standard mode.



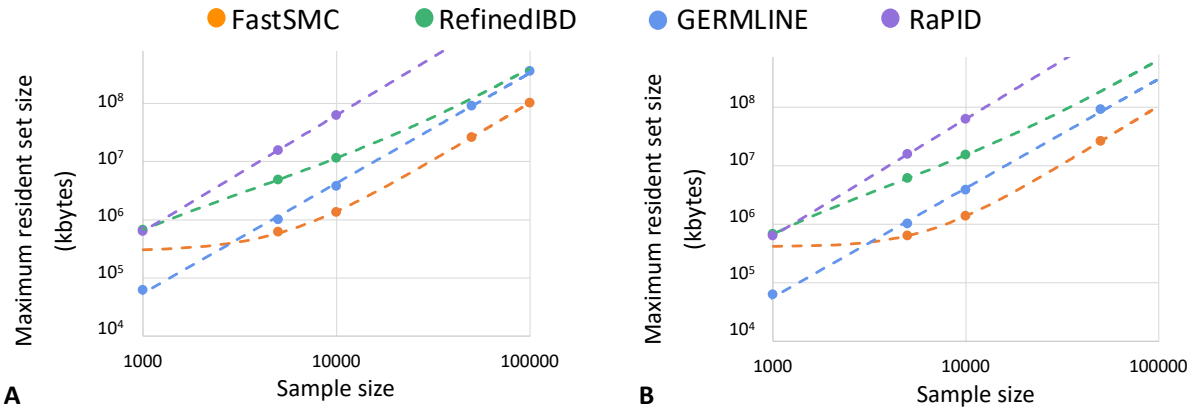
Supplementary Fig. 2. Accuracy evaluation for IBD detection at different time scales. Precision-recall curves within the common recall range (i.e where all methods are able to provide predictions) in the past 25 (A), 50 (B), 100 (C), 150 (D) and 200 (E) generations, randomly sampled from 10 realistic simulated datasets. Each dataset consists of a 30Mb chromosome under European demographic history model, recombination rates from a human chromosome 2 and SNP ascertainment matching UKBB allele frequencies. Precision refers to the fraction of true IBD segments among the retrieved segments while recall measures the proportion of actual IBD segments that are correctly identified as such. For each method (FastSMC, RefinedIBD [1], GERMLINE [2] and RaPID [3]), optimal parameters from the grid search were used.



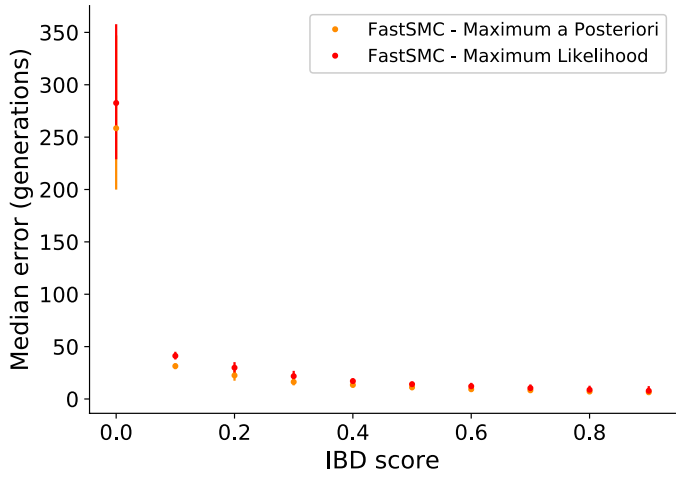
Supplementary Fig. 3. Demographic model misspecification. Effects of demographic model misspecification in age estimate of IBD segments. We simulated 10 batches of 300 haploid samples from the first 30Mb of a human chromosome 2 and a constant population size of $10,000$ diploid individuals, and 10 realistic batches according to the setup described in Methods. We ran FastSMC on both datasets, assuming a European demographic model, with a time threshold of 50 generations and a minimum length of 0.001 cM. We report (A) the absolute median error between the true age and the MAP age estimate while varying the minimum length of the IBD segments, and (B) the mean error between the true age and the MAP estimate while varying the minimum IBD score. Vertical lines represent the mean values +/- SEM over all 10 batches in both (A) and (B). Misspecifying the demographic model results in biased age estimates for very short ($<$ 1cM) or intermediate IBD score segments.



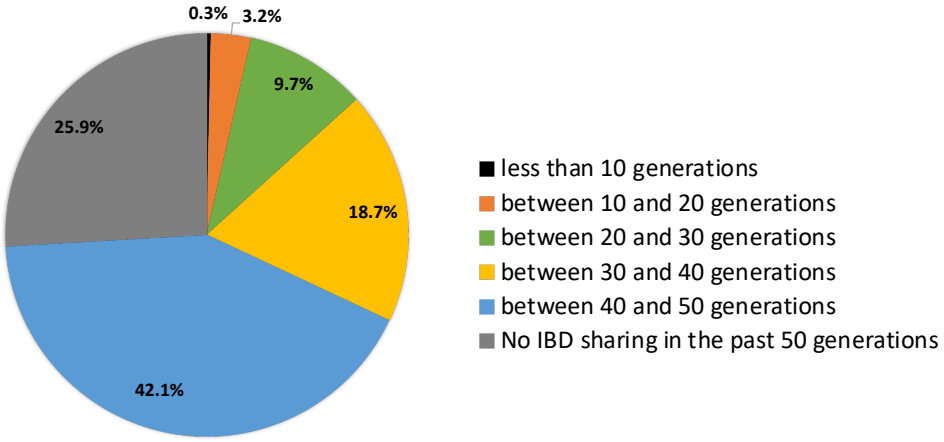
Supplementary Fig. 4. Running time evaluation for IBD detection at different time scales. Running time (CPU seconds) using chromosome 20 of the UKBB across 7,913 SNPs, for IBD segments detection within the past 25 (A), 50 (B), 100 (C), 150 (D) and 200 (E) generations. The complete cohort of 487,409 samples from the UKBB was randomly downsampled into batches. Only one thread was used for each method (FastSMC, RefinedIBD, GERMLINE and RaPID) and parameters from grid search were used. Trend lines in logarithmic scale reflect differences in the quadratic components of each algorithm.



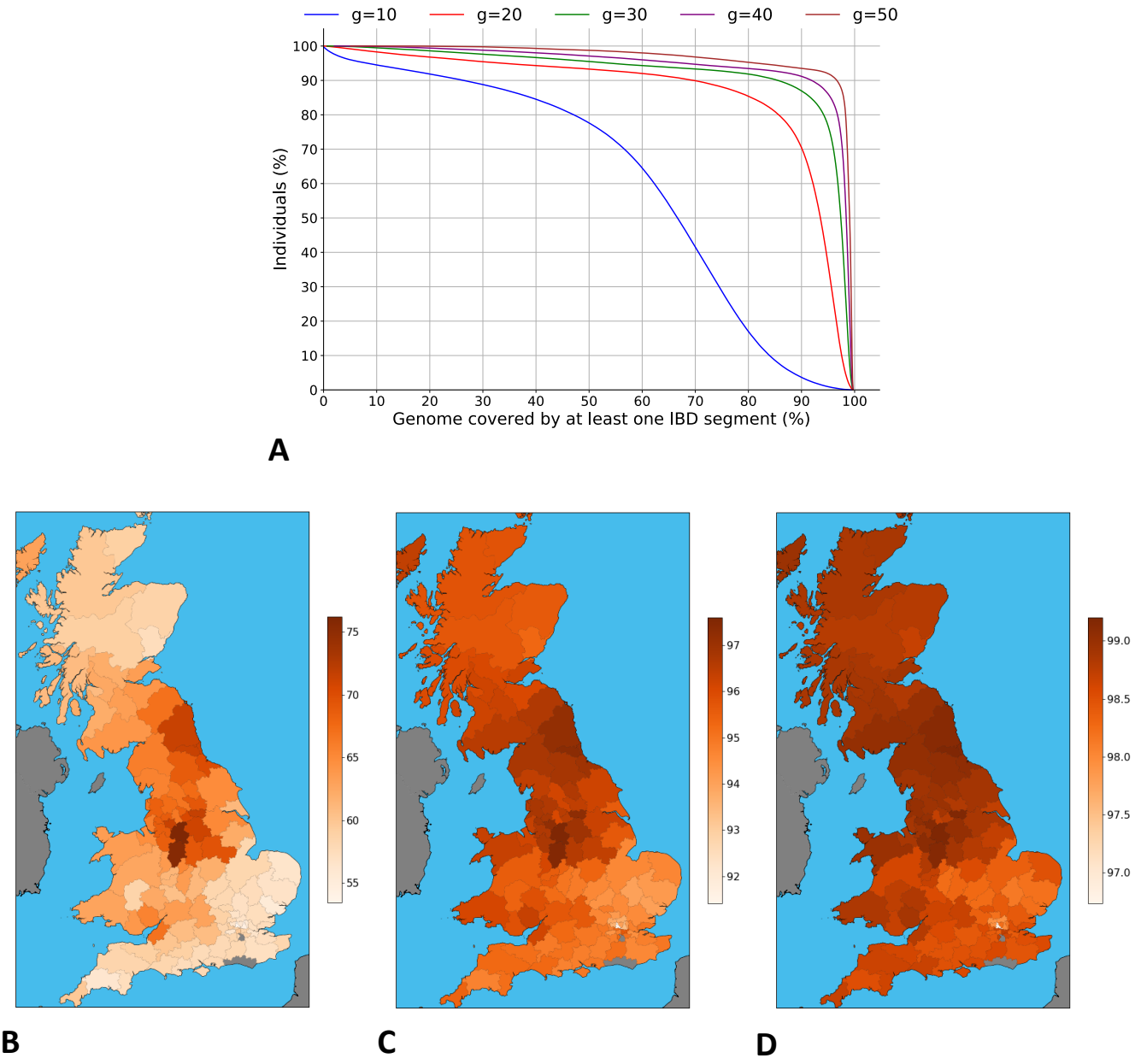
Supplementary Fig. 5. Memory usage for IBD detection at different time scales. Memory usage in kilobytes of FastSMC, RefinedIBD, GERMLINE and RaPID for IBD detection within the past 50 (A) and 100 (B) generations, with optimal parameter values from the grid search across 7,913 SNPs (chromosome 20). The complete cohort of 487,409 samples from the UKBB was randomly downsampled into batches.



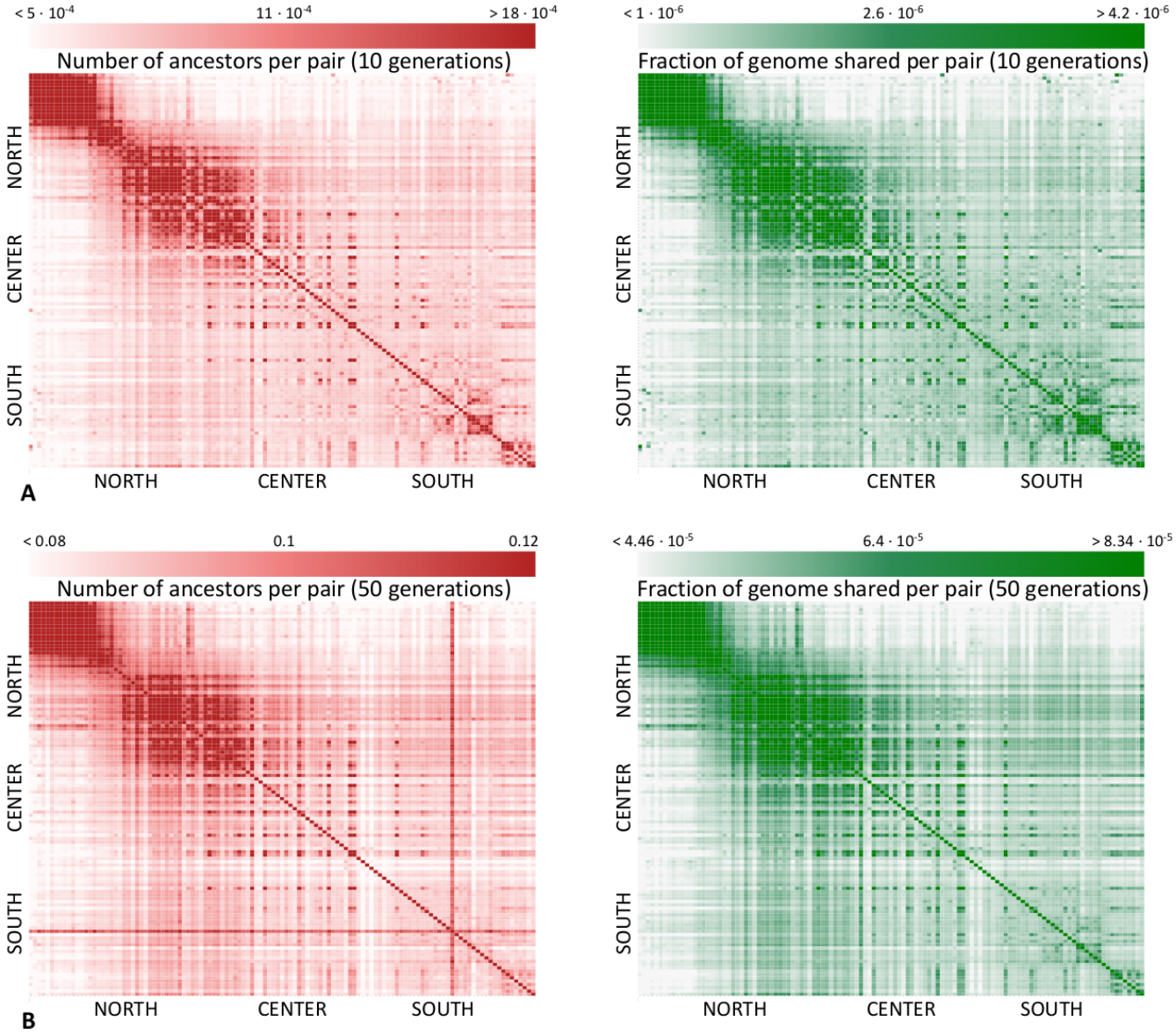
Supplementary Fig. 6. MAP and MLE age estimates comparison. Absolute median error in age estimation in FastSMC for both the MAP and the MLE estimates, at different IBD score thresholds. Data are presented as mean values with 95% confidence intervals over 10 different simulations (each of them consisting of a 30Mb chromosome under European demographic history model, recombination rates from a human chromosome 2 and SNP ascertainment matching UKBB allele frequencies).



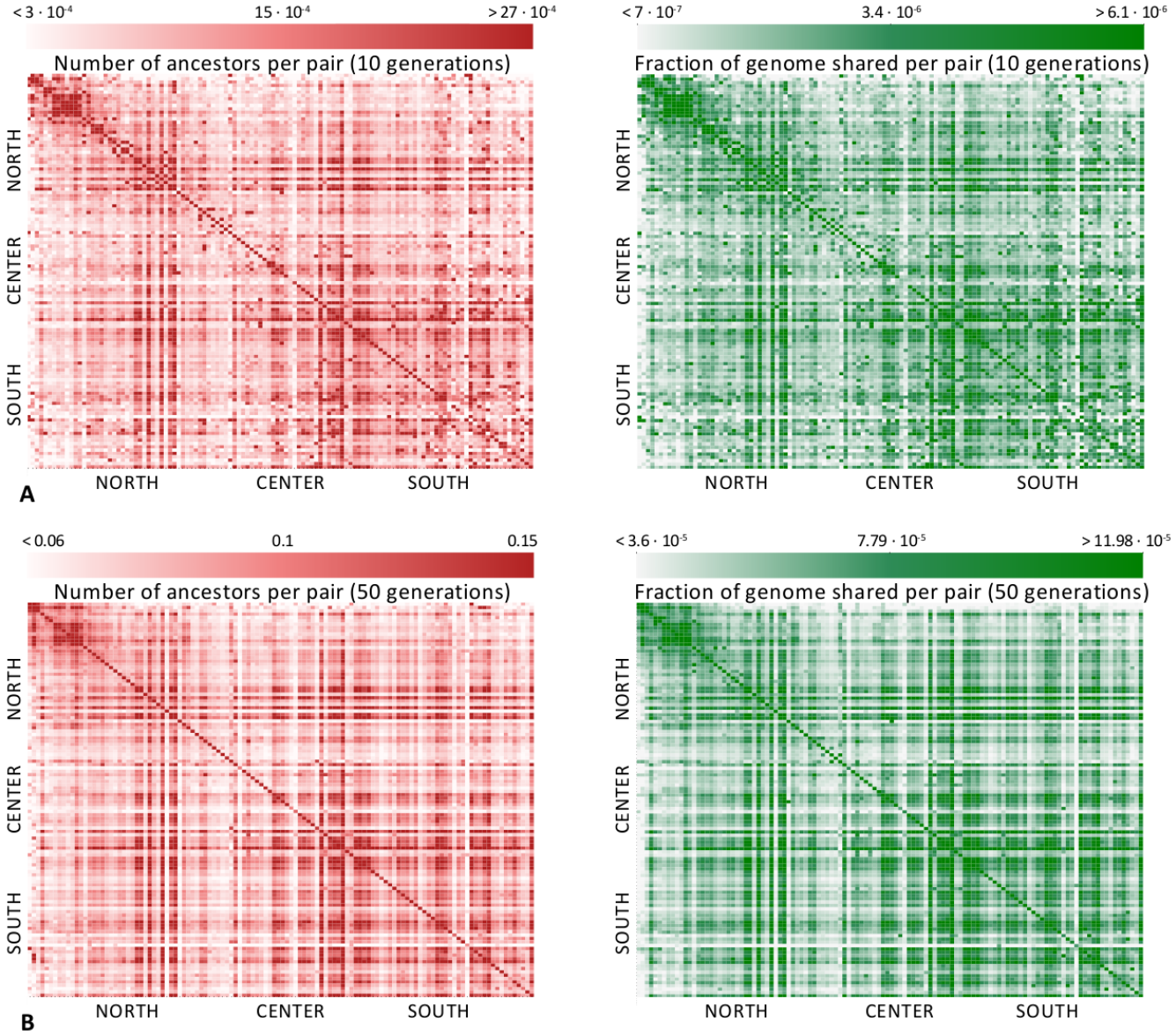
Supplementary Fig. 7. IBD sharing among pairs of samples in the UK Biobank dataset. For each pair of individuals (out of 118, 783, 522, 936 total pairs) in the UK Biobank cohort (487, 409 diploid samples) we determined the most recent estimated shared segment, which provides an upper bound for the pair's genealogical relationship. 74% of all pairs of British individuals were estimated to share common ancestry at some point within the past 50 generations.



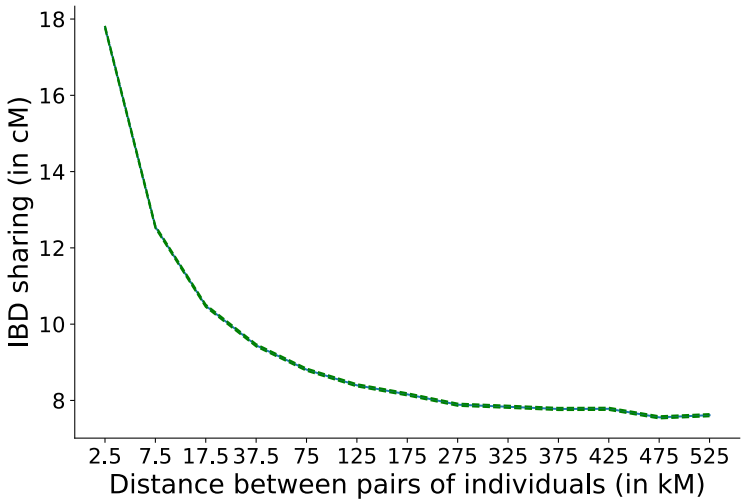
Supplementary Fig. 8. Fraction of genome covered by IBD segments in the UK Biobank dataset. A. Fraction of genome covered by at least one IBD segment (%) for 487,409 samples from the UK Biobank cohort within the past 10, 20, 30, 40 and 50 generations. B, C, D. Average fraction of genome covered by at least one IBD segment (%) within UK postcodes for 432,968 samples with known geographic coordinats from the UK Biobank cohort in the past 10 (B), 30 (C) and 50 (D) generations. Regions with no data available are coloured in gray.



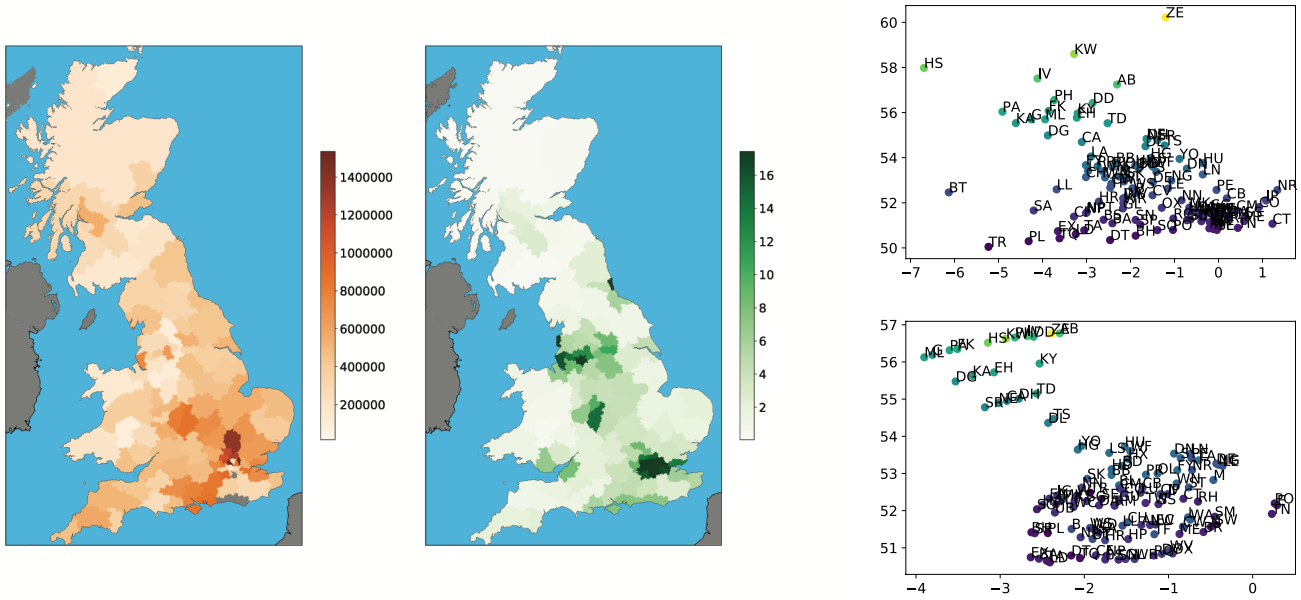
Supplementary Fig. 9. IBD sharing in the past 10 and 50 generations across UK postcodes in the UK Biobank dataset. Average number of IBD segments (left) and fraction of genome (right) shared per pair of individuals across all pair of postcodes in the UK (excluding Northern Ireland) in the past 10 (A) and 50 (B) generations. Segments with IBD score smaller than 0.4 were excluded, resulting approximately in recall of 0.6 and precision of 0.8 based on simulations. The red cross appearing in the South of the UK in B corresponds to Sutton postcode area (covering south-west London), a cosmopolitan region where samples share many ancestors but small fraction of genome with the rest of the country, revealing deep ancestral ties (i.e short IBD segments) throughout the UK.



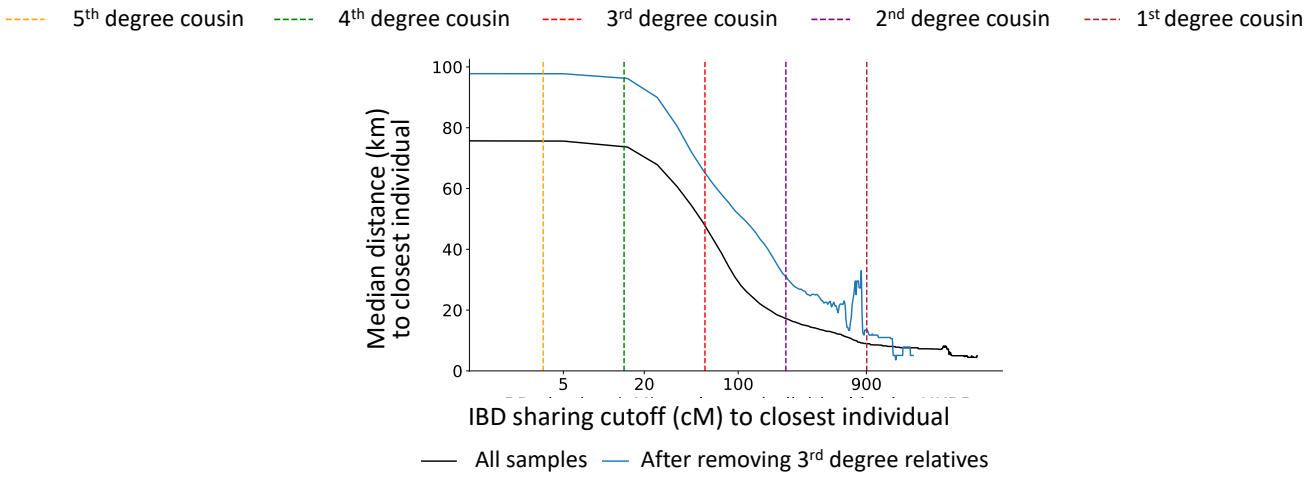
Supplementary Fig. 10. IBD sharing in the past 10 and 50 generations across UK postcodes for non-White British samples in the UK Biobank dataset. Average number of IBD segments (left) and fraction of genome (right) shared per pair of individuals across all pair of postcodes in the UK (excluding Northern Ireland) in the past 10 (A) and 50 (B) generations. Segments with IBD score smaller than 0.4 were excluded, resulting approximately in recall of 0.6 and precision of 0.8 based on simulations. White British samples as defined by Bycroft et al. [4] were excluded, resulting in 38,801 individuals with available birth location. Results reveal similar patterns with the ones observed for the whole UK Biobank cohort.



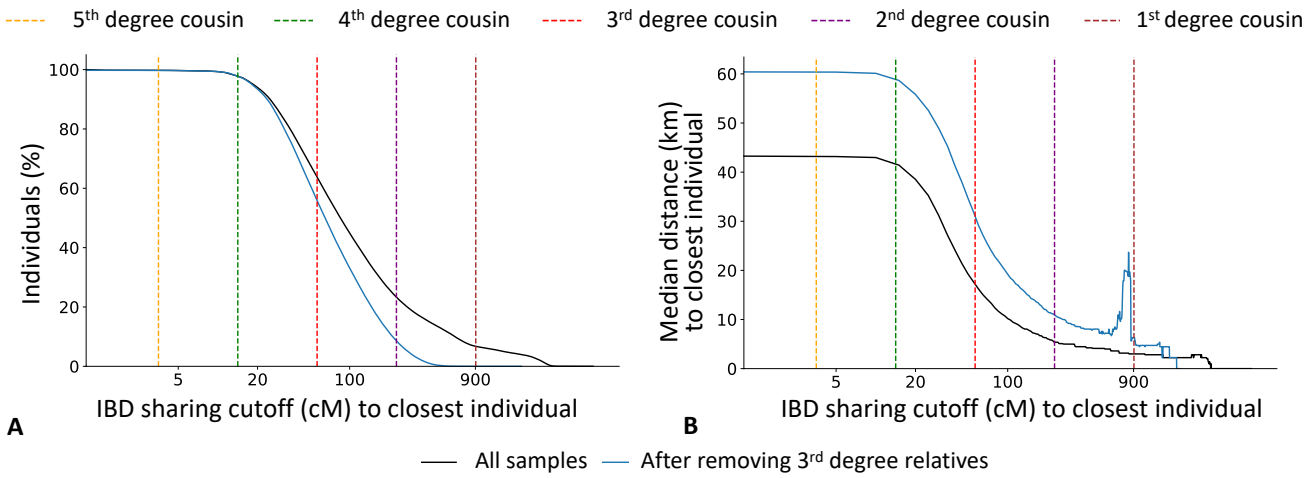
Supplementary Fig. 11. IBD sharing and geographic distance. We randomly sampled $5,000,000$ pairs from the UKBB cohort and computed the IBD sharing in cM (i.e sum of IBD segments lengths along the genome within the past 10 generations) for each of them. We partitioned these pairs depending on the distance in kilometers (kM) in the birth locations of the two individuals (less than 5kM, between 5 and 10kM, between 10 and 25kM, between 25 and 50kM, and then every 50kM up to 500kM). The blue line represents the average IBD sharing across all pairs of random samples, and the green trend lines correspond to the standard error of the mean across all pairs (assuming pairs of individuals are independent, which is approximately true when looking at sharing of very recent segments). We observe strong correlation between genetic and geographic distance, demonstrating that close relatives tend to be geographically clustered.



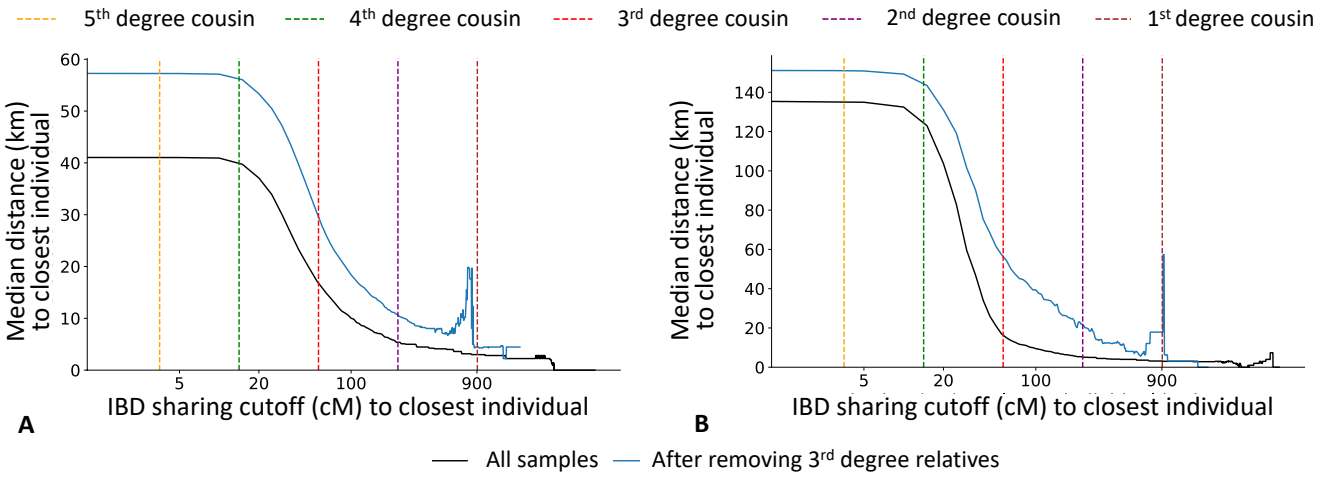
Supplementary Fig. 12. Estimation of effective size and reconstruction of physical distances with IBD sharing. A. Recent effective population size estimation based on IBD sharing in the past 10 generations (left) and 2011 census population size (number of persons per hectare) within postcodes (right). Census population data comes from Nomis web-based dataset for England and Wales, and from the Data Warehouse for Scotland (data only available on the area level, we computed estimates for postcodes). Regions with no data available are coloured in gray. B. Real physical distances between postcodes (top) and isomap projections using IBD sharing within the past 600 years (bottom), x-axis is longitude and y-axis is latitude. The isomap projection was obtained on the IBD sharing postcode dissimilarity matrix and by considering 8 nearest neighbours. We then applied affine transformations (rotation, translation and scaling) to minimize the root mean squared of the reconstruction error (RMSE) [5]. The RMSE obtained is 179 km, 95% CI=[163,196].



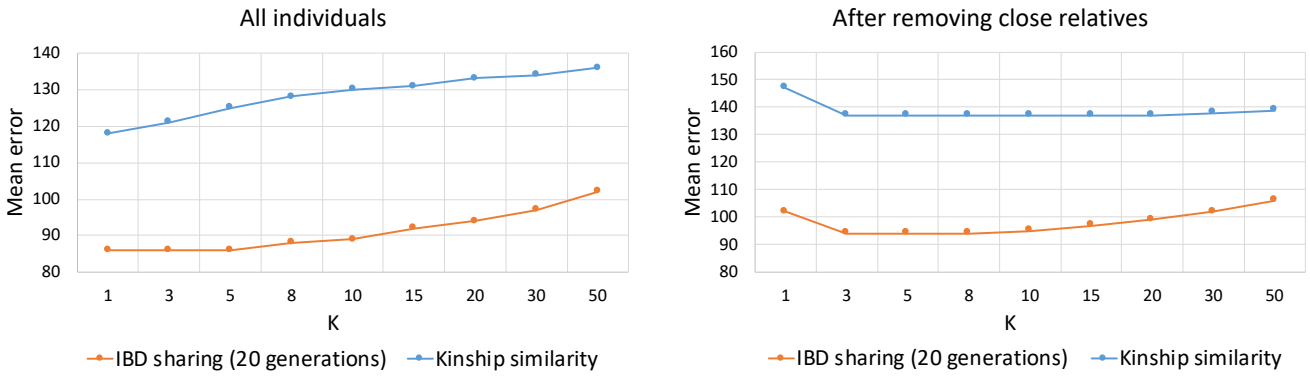
Supplementary Fig. 13. Genetic relatedness and geographic distances (home address) in the UK Biobank dataset, before and after removing related individuals. For each UK Biobank sample with available home address, we detected the individual sharing the largest total amount (in cM) of genome IBD within the past 10 generations (referred to as closest individual). For each value x of total shared genome (in cM) on the X-axis, we report the median distance (km, computed every 10 cM) for all pairs of (sample, closest individual) who shared at least x. Vertical dashed lines indicate the expected value of the total IBD sharing for k-th degree cousins, computed using $2G(1/2)^{2(k+1)}$, where $G = 7247.14$ is the total diploid genome size (in cM) and k represents the degree of cousin relationship (e.g. $k = 2$ for second degree cousins, separated by $2(k+1)$ generations) [6]. We show results obtained using either all individuals with available home address data ($N = 482,832$; black line), or all individuals with available home address data after removing $\leq$3rd degree relatives (e.g. first degree cousins), detected by Bycroft et al. [4] using the KING software [7] ($N = 403,356$; blue line). After excluding $\leq$3rd degree relatives, the sample size for high IBD sharing ($\geq$700 cM) becomes small ($\leq$30, compared to $\leq$30,000 without excluding any sample), making the estimate noisy. The peak observed for large values of IBD sharing is thus likely due to a small number of related groups.



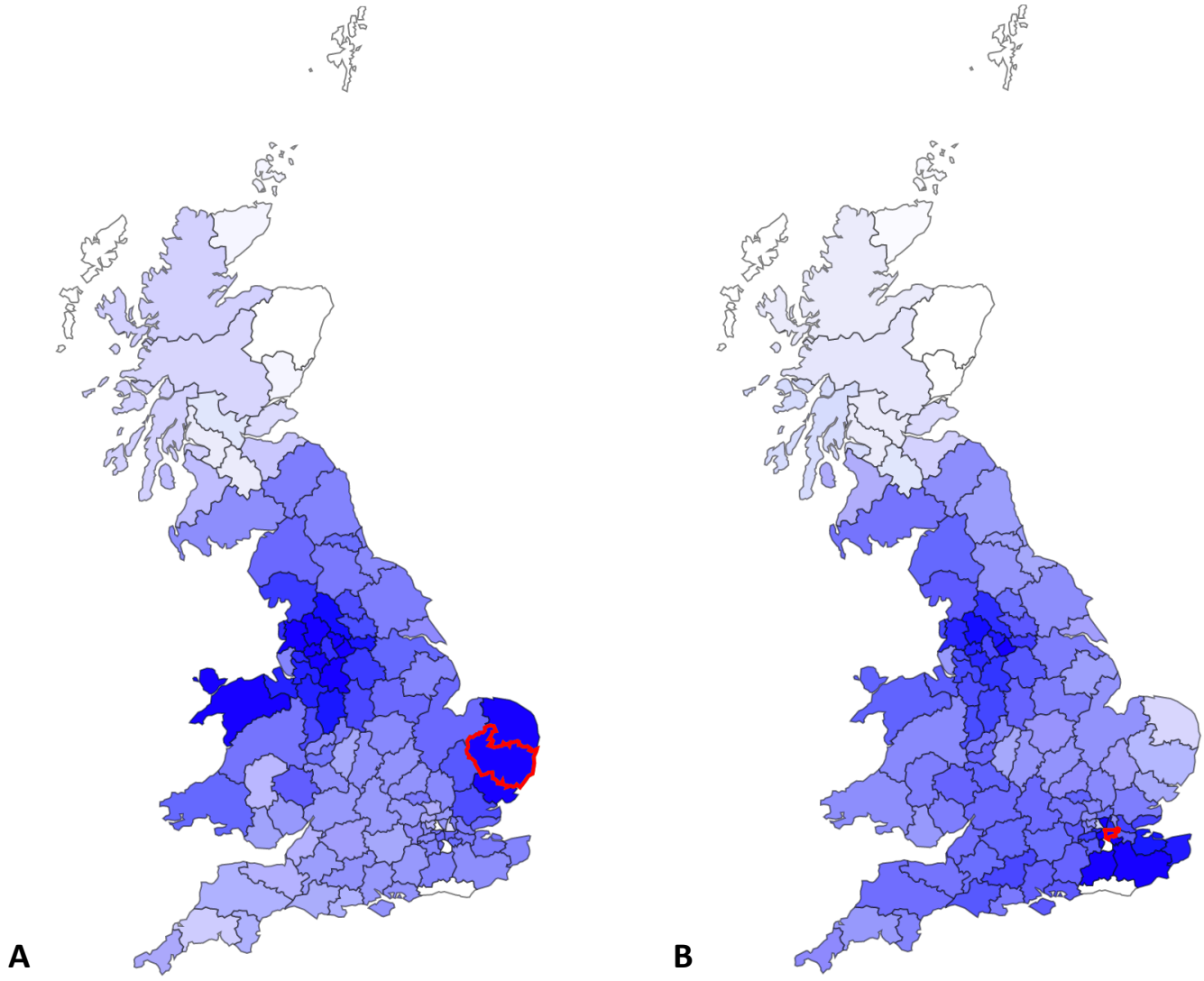
Supplementary Fig. 14. Genetic relatedness and geographic distances (birth locations) in the UK Biobank dataset, before and after removing related individuals. For each UK Biobank sample with available geographic data, we detected the individual sharing the largest total amount (in cM) of genome IBD within the past 10 generations (referred to as closest individual). A. For each value x of total shared genome (in cM) on the X-axis, we report the percentage of UK Biobank samples (Y-axis) that share x or more with their closest individual. B. For each value x of total shared genome (in cM) on the X-axis, we report the median distance (km, computed every 10 cM) for all pairs of (sample, closest individual) who shared at least x. Vertical dashed lines indicate the expected value of the total IBD sharing for k-th degree cousins, computed using $2G(1/2)^{2(k+1)}$, where $G = 7247.14$ is the total diploid genome size (in cM) and k represents the degree of cousin relationship (e.g. $k = 2$ for second degree cousins, separated by $2(k+1)$ generations) [6]. We show results obtained using either all White British (A) and non-White British (B) individuals with available birth location data ($N = 432,968$; black lines), or all White British (A) and non-White British (B) individuals with available geographic data after removing $\leq$3rd degree relatives (e.g. first degree cousins), detected by Bycroft et al. [4] using the KING software [7] ($N = 357,588$; blue lines). After excluding $\leq$3rd degree relatives, the sample size for high IBD sharing ($\geq$700 cM) becomes small ($\leq$30, compared to $\leq$30,000 without excluding any sample), making the estimate noisy. The peak observed for large values of IBD sharing is thus likely due to a small number of related groups.



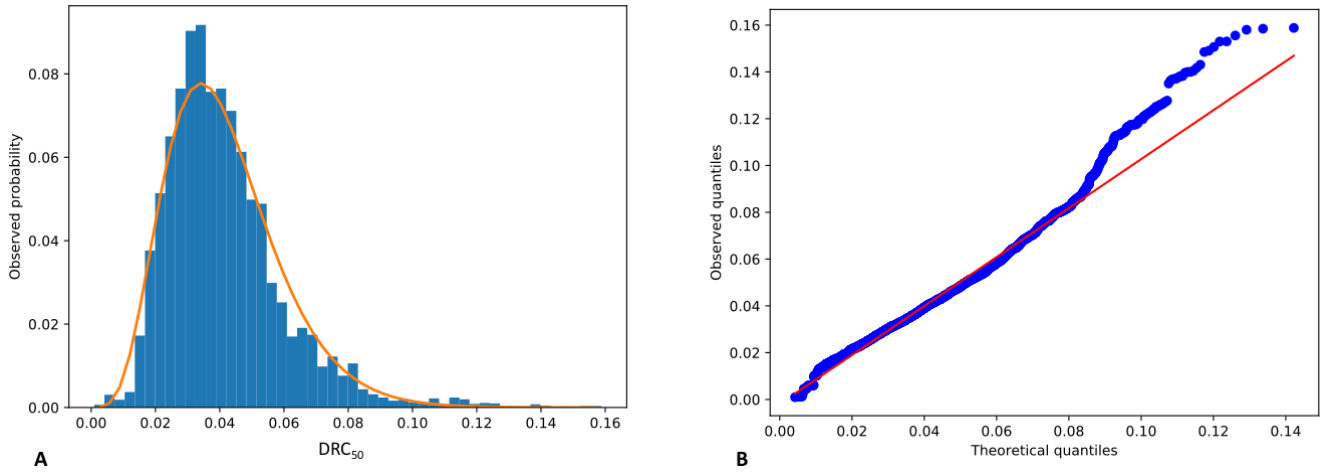
Supplementary Fig. 15. Genetic relatedness and geographic distances (birth location) in the UK Biobank dataset, before and after removing related individuals, for White British and non-White British samples. For each UK Biobank sample with available birth location among White British individuals (A) or non-White British individuals (B), we detected the individual sharing the largest total amount (in cM) of genome IBD within the past 10 generations (referred to as closest individual). For each value x of total shared genome (in cM) on the X-axis, we report the median distance (km, computed every 10 cM) for all pairs of (sample, closest individual) who shared at least x. Vertical dashed lines indicate the expected value of the total IBD sharing for k-th degree cousins, computed using $2G(1/2)^{2(k+1)}$, where $G = 7247.14$ is the total diploid genome size (in cM) and k represents the degree of cousin relationship (e.g. $k = 2$ for second degree cousins, separated by $2(k + 1)$ generations) [6]. We show results obtained using either all White British (A) and non-White British (B) individuals with available birth location data ($N = 394,167$ in A, $N = 38,801$ in B; black lines), or all White British (A) and non-White British (B) individuals with available geographic data after removing ≤3rd degree relatives (e.g. first degree cousins), detected by Bycroft et al. [4] using the KING software [7] ($N = 324,767$ in A, $N = 32,821$ in B; blue lines). A. Only considering White British samples results in a median error of 42km for any random White British sample (compared to 45km when considering all samples) while the median error when taking random pairs of White British samples (irrespective of IBD sharing) is 199.2km (95% CI=[198.9,199.4]). B. Only considering non-White British samples results in a median error for any random non-White British sample of 135km, while the median error when taking random pairs of non-White British samples (irrespective of IBD sharing) is 211.9km (95% CI=[211.2, 212.2]). After excluding ≤3rd degree relatives, the sample size for high IBD sharing (≥700 cM) becomes small, making the estimate noisy. The peak observed for large values of IBD sharing is thus likely due to a small number of related groups.



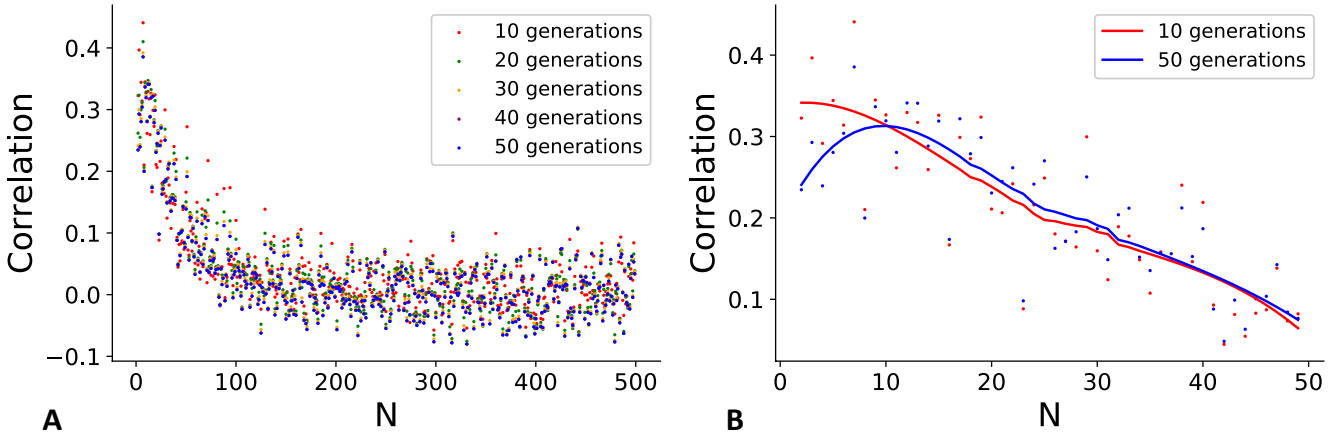
Supplementary Fig. 16. Fine-tuning of the K-nearest neighbours algorithm. (left) Average error when predicting birth location of 10,000 random samples from the UKBB, applying the K-nearest-neighbours algorithm while varying the value of the parameter K, using both IBD sharing among individuals within the past 20 generations and kinship similarity. (right) Average error of 8,226 random samples from the UKBB, after excluding close relatives ($\leq$3rd degree relatives), using the same procedure.



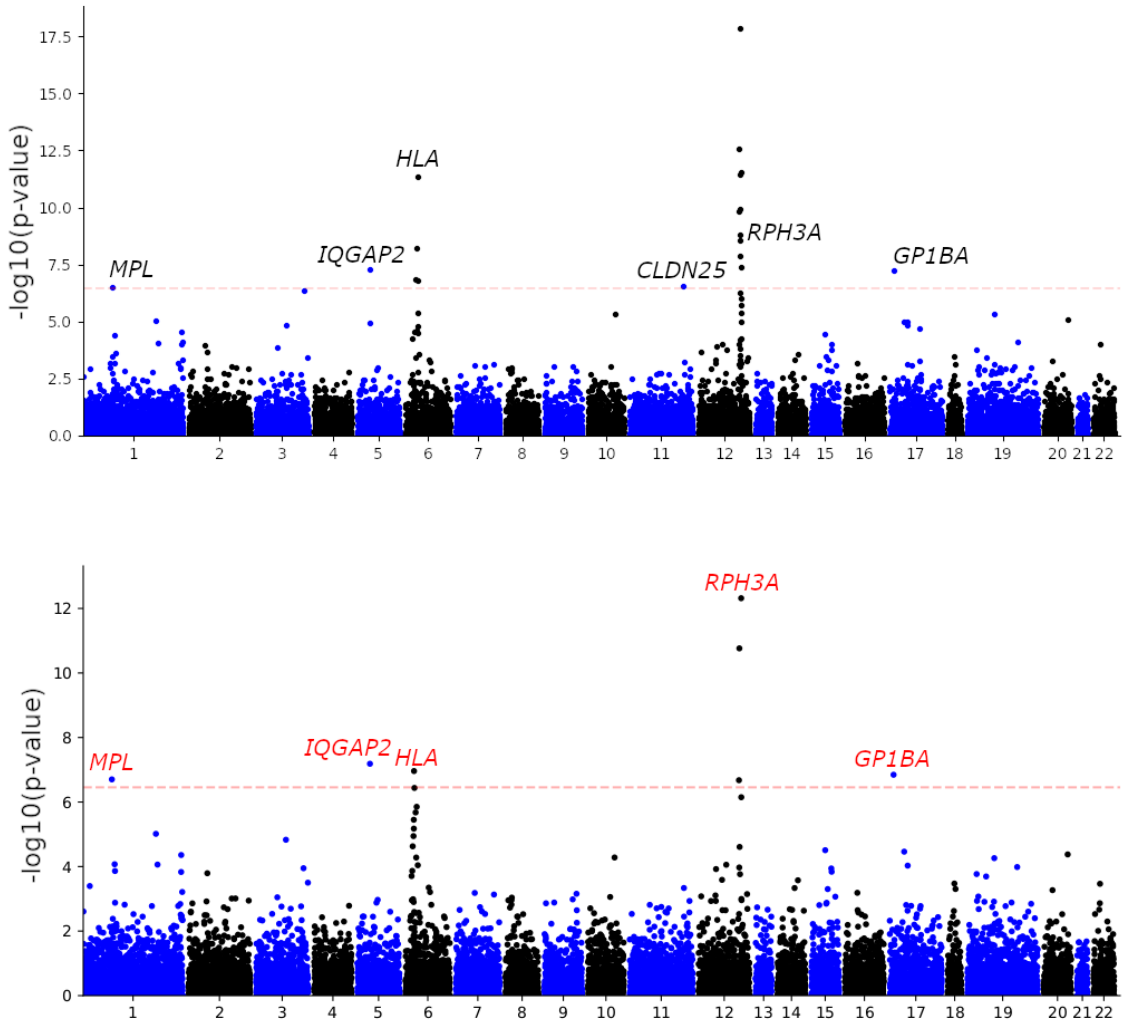
Supplementary Fig. 17. Pervasive IBD sharing with North-West England. Individuals throughout the UK, including individuals from cosmopolitan regions such as London, have deep genetic relationship with modern-day individuals from North-West England, in addition to nearby regions. This figure displays the average fraction of genome shared through IBD segments in the past $1,500$ years per pair of individuals between the IP postcode (corresponding to Ipswich, in red) and other UK regions (A), and between the SE postcode (corresponding to South East London, in red) and other UK regions (B). The darker the color is, the higher the average fraction of genome shared is. More details and an interactive map can be found at https://ukancestrymap.github.io/



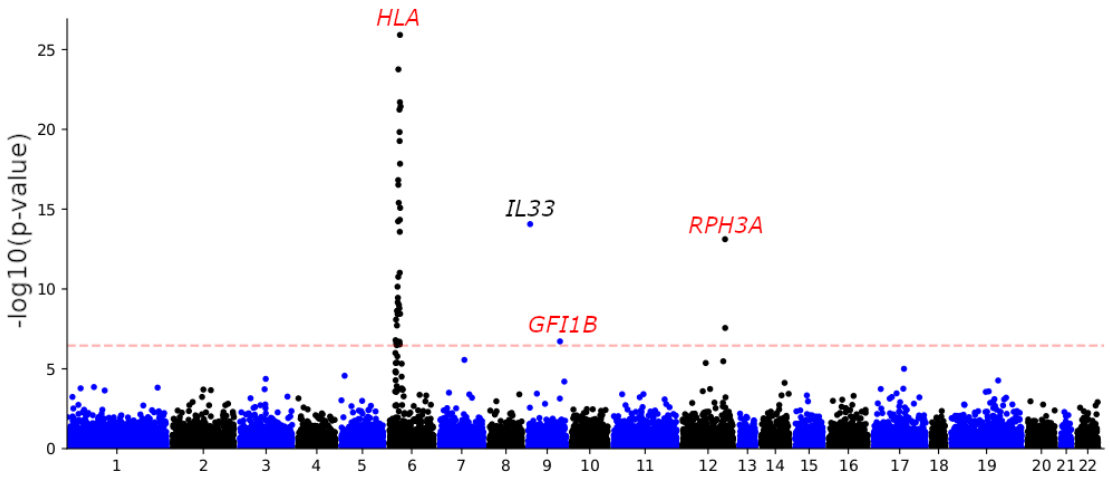
Supplementary Fig. 18. Empirical null model for the DRC_{50} statistic. Empirical null model for detection of recent positive selection, fitted using a Gamma distribution with shape, location and scale parameters. A. Empirical distribution (in blue) and Gamma-fit (orange curve) for the DRC_{50} statistic in the putative neutral regions of the genome in the UKBB, after excluding significant loci falling within these putative neutral regions (9, 165 observations from 0.05 cM windows). B. Quantile-quantile plot for the DRC_{50} statistic in the putative neutral regions of the genome in the UKBB, after excluding significant loci falling within these putative neutral regions (9, 165 observations from 0.05 cM windows).



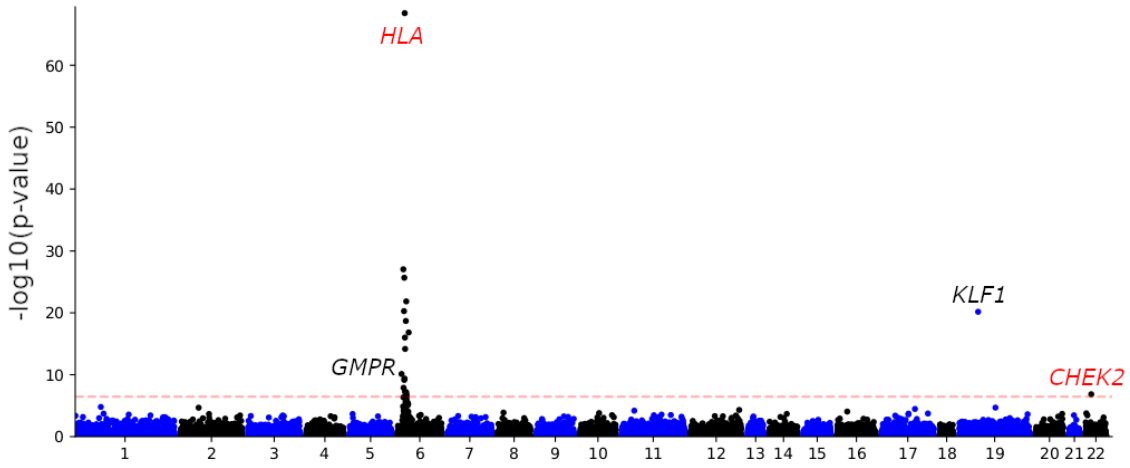
Supplementary Fig. 19. Correlation between IBD sharing and sharing of rare variants. A. Correlation between IBD sharing (average number of IBD segments per pair within UK regions in the past 10, 20, 30, 40 and 50 generations in the UK Biobank cohort) and ultra-rare variants sharing (average number of F_N mutations per pair within UK regions in the UK Biobank 50k Exome Sequencing Data Release, for any N between 2 and 499). B. Non-parametric regression of the correlation between IBD sharing (average number of IBD segments per pair within UK regions in the past 10 and 50 generations in $487,409$ samples in the UK Biobank dataset) and ultra-rare variants sharing (average number of F_N mutations per pair within UK regions in the UK Biobank 50k Exome Sequencing Data Release, for values of N smaller than 50).



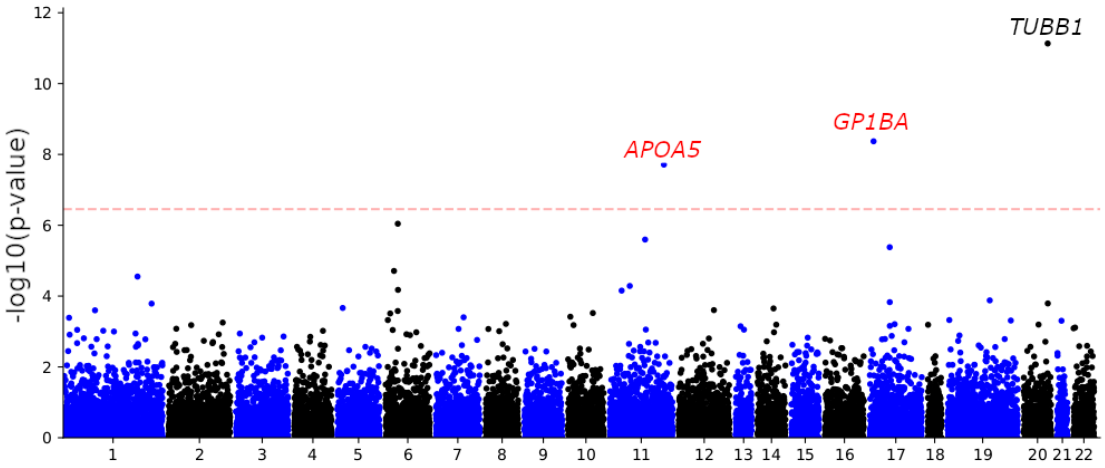
Supplementary Fig. 20. LoF-segment burden exome-wide Manhattan plots for platelet count with and without SNP-adjustment. Labelled genes are exome-wide significant after adjusting for multiple testing ($p < 0.05/(14,249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). We compare results before (top) and after (bottom) adjusting for common SNP associations. Both LoF-segment burden analyses used 303,125 UK Biobank samples not included in the exome sequencing cohort. The cluster of genes in chromosome 12 labeled as *RPH3A* (the top association) contains *KCTD10, TCHP* and *RPH3A* and the signal with *CLDN25* was cleared after SNP-adjustment. Red labels in the lower plot indicate associations that were not detected in our WES-based LoF burden analysis or reported by Van Hout et al. [8]. Exact p-values are reported in Supplementary Table 7.



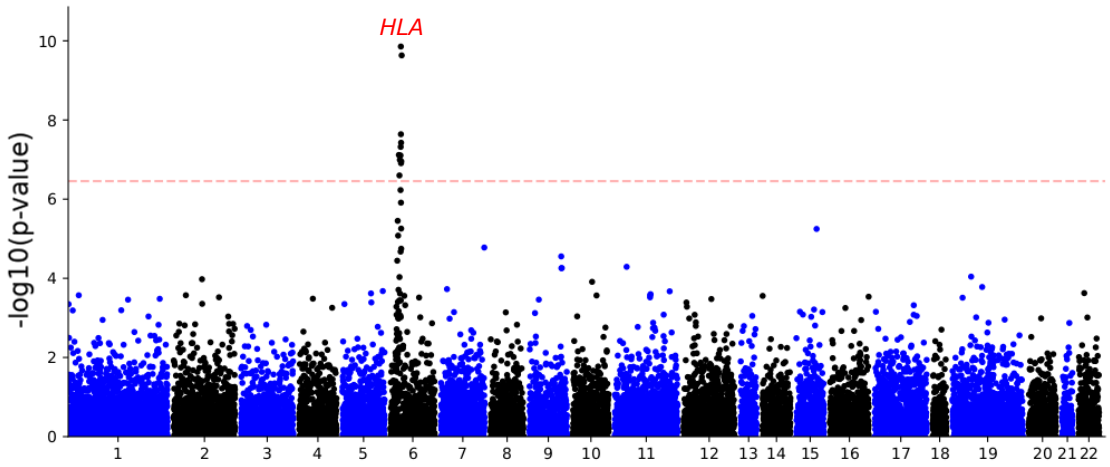
Supplementary Fig. 21. LoF-segment burden exome-wide Manhattan plot for eosinophill count. Labelled genes are exome-wide significant (after adjusting for multiple testing, $p < 0.05/(14,249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified one locus previously reported by Van Hout et al. [8] (black label), and additional loci on chromosomes 6,9 and 12 (labels in red). Table 2 reports the list of genes within the gene cluster labeled as HLA. Exact p-values are reported in Supplementary Table 7.



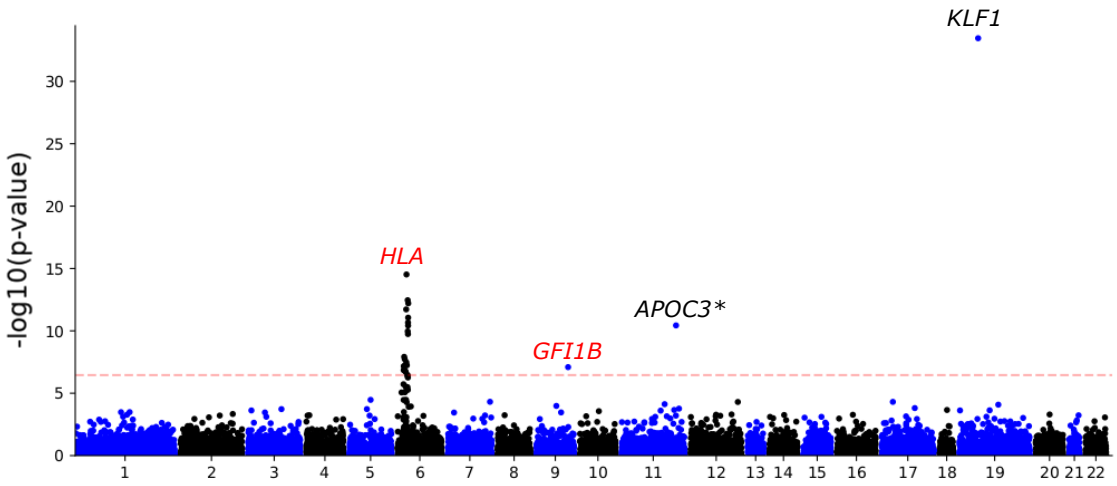
Supplementary Fig. 22. LoF-segment burden exome-wide Manhattan plot for mean corpuscular haemoglobin. Labelled genes are exome-wide significant (after adjusting for multiple testing, $p < 0.05/(14,249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified two loci previously reported by Van Hout et al. [8], *KLF1* and *GMPR* (gene labels in black), and two novel associations at HLA and *CHEK2* (labels in red). Exact p-values are reported in Supplementary Table 7.



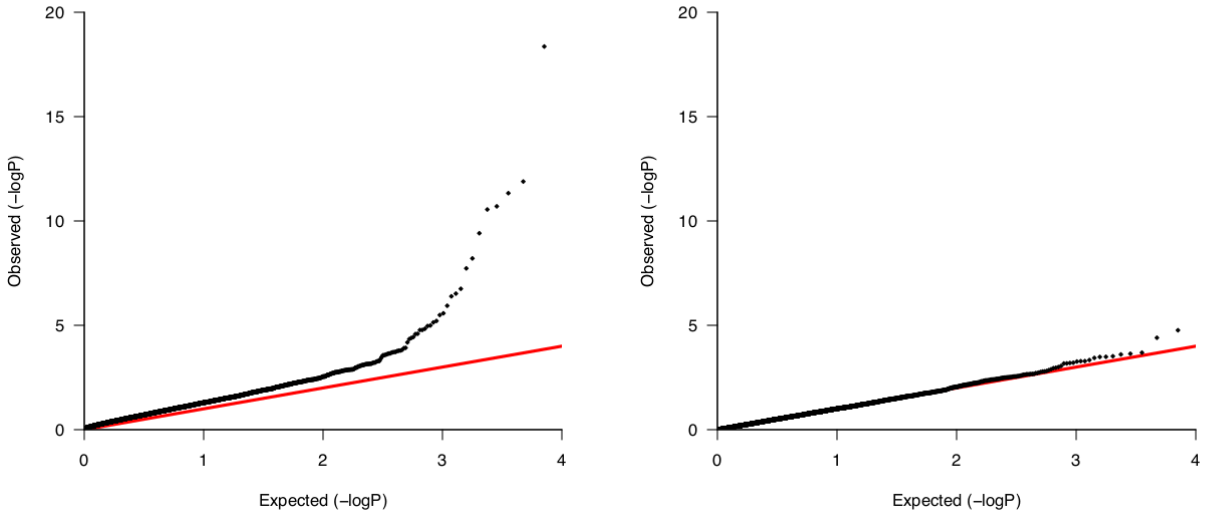
Supplementary Fig. 23. LoF-segment burden exome-wide Manhattan plot for platelet distribution width. Labelled genes are exome-wide significant (after adjusting for multiple testing, $p < 0.05/(14,249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified one locus previously reported by Van Hout et al. [8], *TUBB1* (black label), and two additional genes, *APOA5* and *GP1BA* (labels in red). Exact p-values are reported in Supplementary Table 7.



Supplementary Fig. 24. LoF-segment burden exome-wide Manhattan plot for red blood cell count. Labelled genes are exome-wide significant (after adjusting for multiple testing, $p < 0.05/(14,249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We detected one novel association at the HLA locus which was not detected by either of the WES-burden tests. Exact p-values are reported in Supplementary Table 7.



Supplementary Fig. 25. LoF-segment burden exome-wide Manhattan plot for red blood cell distribution width. Labelled genes are exome-wide significant (after adjusting for multiple testing, $p < 0.05/(14,249 \times 10) = 3.51 \times 10^{-7}$; dashed red line). The LoF-segment burden analysis (with SNP adjustment) used 303,125 UK Biobank samples not included in the exome sequencing cohort. We identified two previously-reported loci (black labels), *KLF1* (detected by Van Hout et al. [8]) and *APOC3* (detected by our WES burden analysis), along with two additional (labels in red). Exact p-values are reported in Supplementary Table 7.



Supplementary Fig. 26. Quantile-quantile plots for LoF-segment burden association. Quantile-quantile plots for mean platelet (thrombocyte) volume (left) and for the same trait, but with randomly permuted phenotype values (right). We observe no signal in the permuted phenotype analysis, suggesting a well-calibrated test. The genomic inflation factor for the (SNP-adjusted) LoF-segment burden test, calculated as the the ratio between the observed median chi-squared association statistic and the median chi-squared association statistic expected under the null, is 1.982 (similar values were observed for the other traits). Values larger than 1 may be caused by pervasive polygenicity [9] and are also observed in analyses of common variants (1.492 for this trait using summary statistics from Loh et al. [10]) and by population stratification remaining after correcting for principal components [11].

Supplementary Tables

Method \ Time		25	50	100	150	200
FastSMC	t	25	50	100	150	200
	length	1.5	1	0.5	0.25	0.1
GERMLINE	min	0.001	0.001	0.001	0.001	0.001
	bits	64	64	64	64	64
	err	0	0	0	0	0
RaPID	l	0.5	0.5	0.5	0.5	0.5
	r	10	10	10	10	10
	s	8	8	8	5	5
	w	3	3	3	3	3
RefinedIBD	length	1.5	1	0.75	0.5	0.1
	lod	1	1	1	1	1

Supplementary Table 1. Optimal parameters from the grid search for IBD detection methods. The accuracy of each method (FastSMC, GERMLINE, RaPID and RefinedIBD) was optimised at different time scale (25, 50, 100, 150 and 200 generations) on a simulated dataset of 300 haploid individuals from a European demographic model and a region of 30 Mb from chromosome 2. FastSMC has two main parameters that we tuned: a time threshold (t) to indicate how deep in time the user wants to detect common ancestry, and a minimum length parameter (in cM) (length) for the IBD segments. The time threshold parameter was always set to be the same as the time threshold used for the benchmarking. We tuned three parameters in GERMLINE: the minimum length of IBD segments (in cM) (min, optimized over $[0.001, 0.01, 0.1, 0.5, 0.75, 1, 1.5, 3, 5]$), the numbers of bits (SNPs) in each window (bits, optimized over $[32, 64, 128, 256]$), the minimum number of mismatches allowed in each of them (err, optimized over $[0, 2, 5, 10, 20]$). RaPID has four major parameters: the minimum IBD segments length (in cM) (l, optimized over $[0.001, 0.01, 0.1, 0.5, 1, 1.5, 3, 5]$), the number of iterations for the PBWT algorithm (r, optimized over $[1, 5, 10, 40, 70]$), the minimum number of successes (s, optimized over $[1, 5, 8, 10, 20, 35, 40, 50, 70]$) and the window size (in SNPs) (w, optimized over $[1, 3, 5, 10, 15, 20, 50, 85, 110, 150]$). RefinedIBD's parameters include a minimum length (in cM) (length, optimized over $[0.001, 0.01, 0.1, 0.5, 0.75, 1, 1.5, 3, 5]$) and a minimum LOD score (proxy for quality) (lod, optimized over $[0.01, 0.1, 1, 3, 5]$). We used default values for all parameters not mentioned here.

time threshold (generations)	average common recall range	method	auPRC	average percent auPRC improvement: 100 ($\text{auPRC}_{\text{FastSMC}}/\text{auPRC}_{\text{X}}$ - 1)
25	[0.07, 0.86]	FastSMC	0.62 (0.06)	
		RefinedIBD	0.54 (0.07)	13.98 (5.70)
		GERMLINE	0.59 (0.06)	4.71 (2.52)
		RaPID	0.46 (0.07)	35.61 (13.23)
50	[0.03, 0.77]	FastSMC	0.59 (0.03)	
		RefinedIBD	0.52 (0.04)	14.61 (4.45)
		GERMLINE	0.57 (0.03)	4.30 (0.88)
		RaPID	0.41 (0.05)	46.27 (12.37))
100	[0.01, 0.66]	FastSMC	0.51 (0.02)	
		RefinedIBD	0.46 (0.03)	10.05 (2.27)
		GERMLINE	0.46 (0.02)	11.34 (1.51)
		RaPID	0.29 (0.04)	80.15 (19.62)
150	[0.01, 0.60]	FastSMC	0.46 (0.01)	
		RefinedIBD	0.44 (0.02)	4.91 (1.54)
		GERMLINE	0.38 (0.01)	19.04 (1.96)
		RaPID	0.22 (0.03)	114.36 (28.98)
200	[0.00, 0.53]	FastSMC	0.41 (0.01)	
		RefinedIBD	0.40 (0.01)	1.04 (1.15)
		GERMLINE	0.33 (0.01)	24.11 (2.36)
		RaPID	0.18 (0.02)	133.95 (32.67)

Supplementary Table 2. Accuracy measurement. Difference in accuracy between FastSMC and other IBD detection methods within the past 25, 50, 100, 150 and 200 generations. We report the percent improvement for the area under the precision-recall curve (auPRC) of FastSMC over other methods. For all methods, precision can only be estimated within a limited recall range (due mostly to the minimum length parameter) and we only report accuracy measurement on the common recall range. Optimal parameters from fine-tuning were used (Supplementary Table 1). Accuracy was measured on 10 realistic simulated datasets, all different from the one used for parameters fine-tuning and all consisting of a 30Mb chromosome under European demographic history model for 300 samples, recombination rates from a human chromosome 2 and SNP ascertainment matching UKBB allele frequencies. Numbers in round brackets represent standard errors over 10 simulations.

time threshold (generations)	average recall range	auPRC
25	[0.07, 0.84]	0.62 (0.09)
50	[0.03, 0.72]	0.57 (0.05)
100	[0.01, 0.65]	0.51 (0.03)
150	[0.0, 0.59]	0.45 (0.02)
200	[0.0, 0.55]	0.4 (0.01)

Supplementary Table 3. Effects of demographic model misspecification. We simulated 10 batches of 300 haploid samples from the first 30Mb of a human Chromosome 2 and a constant population size of $10,000$ diploid individuals. We ran FastSMC at different time scales, assuming a European demographic model and we reported the auPRC and the average recall range. Numbers in round brackets represent standard errors. Values are very similar to results obtained on samples simulated with a European demographic model (Supplementary Table 2). Demographic model misspecification does not impact auPRC.

Chr	Region (Mb)	Min. p-value	Top SNP	Candidate gene(s)	Top gene
1	16.56-17.35	1.59e-07	rs142953444	*RSG1, SDHB, CROCC, MFAP2, NBPF1, SZRD1, FBXO42, NECAP2, ATP13A2, FAM231A, FAM231B, FAM231C, SPATA21, RSG1, SDHB, EPHA2, HSPB7, PADI1, PADI2, PADI3, CLCNKA, CLCNKB, C1orf64, FAM131C, ARHGEF19, C1orf134*	*NBPF1*
1	146.79-147.42	3.23e-09	rs35618507	*ACP6, BCL9, GJA5, GJA8, GPR89B, FMO5, CHD1L, GPR89B, NBPF24, PRKAB2, BX842679.1*	***CHD1L***
1	223.71-224.09	1.65e-07	rs73128182	*APN2, CAPN8, TP53BP2, SUSD4, FBXO28, C1orf65, AC138393.1*	***CAPN8***
2	134.57-138.75	1.00e-20	rs80188644	*LCT, DARS, HNMT, MCM6, ACMSD, CCNT2, CXCR4, MGAT5, UBXN4, R3HDM1, THSD7B, ZRANB3, MAP3K19, TMEM163, RAB3GAP1, HNMT*	*LCT*
6	25.15-33.65	1.00e-20	rs116460689	*HLA* region	*HLA* region
10	17.43-18.10	1.37e-09	rs61842124	*MRC1, STAM, PTPLA, MRC1L1, ST8SIA6, TMEM236, VIM, MRC1, TRDMT1, ST8SIA6, SLC39A12*	***MRC1***
11	0.84-1.71	1.00e-20	rs9704919	*MUC* gene family	*MUC2*
16	70.10-72.69	3.40e-09	rs1833931	*HP, FUK, HPR, TAT, AARS, COG4, IL34, IST1, PDPR, AP1G1, WWP2, ZFHX3, CLEC18A CALB2, CHST4, CMTR2, DHODH, DHX38, HYDIN, SF3B3, VAC14, ZNF19, ZNF23, ATXN1L, DDX19A, DDX19B, EXOSC6, FKSG63, MTSS1L, PHLPP2, PMFBP1, TXNL4B, ZNF821, CLEC18C, ST3GAL2, FLJ00418, MARVELD3, AC009060.1, AC010547.9, RP11-529K1.3*	***HYDIN***
16	88.25-88.48	1.83e-07	rs9925058	*MVD, BANP, CYBA, IL17C, ZFPM1, ZC3H18, ZNF469*	***BANP***
17	41.84-44.95	2.67e-11	rs1230396	*GRN, NSF, PPY, PYY, STH, FZD2, GFAP, GJC1, MAPT, MPP2, MPP3, NAGS, NMT1, UBTF, WNT3, ACBD4, ASB16, C1QL1, CRHR1, DBF4B, DCAKD, DUSP3, FMNL1, G6PC3, HDAC5, LSM12, PLCD3, TMUB2, WNT9B, ADAM11, ARL17A, ARL17B, CCDC43, EFTUD2, HEXIM1, HEXIM2, HIGD1B, ITGA2B, KANSL1, KIF18B, SLC4A1, SPPL2C, ATXN7L3, CCDC103, CD300LG, FAM187A, FAM215A, GPATCH8, LRRC37A, PLEKHM1, RUNDC3A, SPATA32, TMEM101, ARHGAP27, C17orf53, FAM171A2, LRRC37A2, SLC25A39, C17orf104, C17orf105, AC003043.1, AC003102.1, RP11-527L4.2, DHX8, ETV4, SOST, GOSR2, MEOX1, RPRML, WNT9B, RP11-156P1.2*	***EFTUD2***
19	11.19-11.56	1.60e-09	rs36005514	*EPOR, LDLR, RGL3, DOCK6, KANK2, RAB3D, SPC24, PRKCSH, SWSAP1, CCDC151, CCDC159, TMEM205, TSPAN16, C19orf80, DKFZP761J1410, ACP5, CNN1, DNM2, CARM1, ECSIT, ELOF1, TMED1, YIPF2, ELAVL3, ZNF627, ZNF653, SMARCA4, C19orf38, C19orf52, CTC-398G3.6*	*LDLR*
19	44.60-45.45	3.23e-12	rs72480795	*PVR, APOE, BCAM, BCL3, CBLC, APOC1, PVRL2, IGSF23, TOMM40, ZNF112, ZNF180, ZNF224, ZNF225, ZNF226, ZNF227, ZNF229, ZNF233, ZNF234, ZNF235, ZNF285, CEACAM16, CEACAM19, CTC-512J12.6 RELB, APOC2, APOC4, NKPD1, ZNF45, CLASRP, CLPTM1, GEMIN7, ZNF155, ZNF221, ZNF222, ZNF223, ZNF230, ZNF283, ZNF284, ZNF296, ZNF404, BLOC1S3, PPP1R37, TRAPPC6A, AC005779.2, APOC4-APOC2 CTB-129P6.11*	***BCAM***

Supplementary Table 4. Genome-wide significant selection loci. We report loci with elevated values of the DRC_{50} statistic (in the past 50 generations) in the UKBB (1-sided test, after adjusting for multiple testing $p < 0.05/52,003 = 9.6 \times 10^{-7}$). The DRC_{50} statistic of recent positive selection was computed using all 487,409 individuals from the UKBB. When multiple candidate genes were found, we only retained the one nearest to the top SNP, referred to as the top gene (i.e with the smallest p-value). Novel genes are denoted in bold.

Chr	Region (Mb)	Min. p-value	Top SNP	Recombination rate at peak	Percentile recombination rate	Marker density at peak	Percentile marker density
1	16.56-17.35	1.59e-07	rs142953444	0.54	0.9804	12	0.9343
1	146.79-147.42	3.23e-09	rs35618507	0.17	0.687	36	0.2345
1	223.71-224.09	1.65e-07	rs73128182	0.19	0.7286	39	0.1857
2	134.57-138.75	1.00e-20	rs80188644	0.1	0.5096	22	0.5945
6	25.15-33.65	1.00e-20	rs116460689	0.25	0.8159	47	0.0966
10	17.43-18.10	1.37e-09	rs61842124	0.1	0.5327	8	0.9853
11	0.84-1.71	1.00e-20	rs9704919	0.3	0.8655	133	0.003
16	70.10-72.69	3.40e-09	rs1833931	0.45	0.9577	19	0.7041
16	88.25-88.48	1.83e-07	rs9925058	0.19	0.7293	21	0.6344
17	41.84-44.95	2.67e-11	rs1230396	0.04	0.2594	21	0.6338
19	11.19-11.56	1.60e-09	rs36005514	0.44	0.9542	577	0.0001
19	44.60-45.45	3.23e-12	rs72480795	0.45	0.9581	58	0.042

Supplementary Table 5. Marker density and recombination rate percentiles for genome-wide significant selection loci under selection. We divided the genome into 0.1 Mb windows and computed recombination rate and marker density within each window. We then ranked the windows by marker density (from high to low values) and by recombination rate (from low to high values) to get percentiles for each window. We associated each of the selection peaks reported in Supplementary Table 4 to the window they fell in. We report the marker density and recombination rate percentiles for each selection peak.

	LoF	Non-LoF	All
Total number of variants	214,753	10,233,971	10,448,724
Number of age-estimated variants	3,445	159,614	163,059
Mean age estimate	6,377.95	6,487.54	6,485.22
Median age estimate	879.4	904.86	904.25
95% age estimates CI	$[106.57; 47,794.92]$	$[116.62; 48,002.80]$	$[116.51; 48,002.80]$

Supplementary Table 6. Age estimates for exome sequencing alleles. We report a summary of the allele age estimates (in generations) for a subset of exome sequencing variants analyzed by Albers et al. [12]. All corresponds to the complete set of exome-sequenced variants; LoF refers to loss-of-function mutations we analyzed and non-LoF refers to its complement.

Trait	Gene	LoF-segment burden not SNP-adjusted	LoF-segment burden SNP-adjusted	WES LoF burden not SNP-adjusted	WES LoF burden SNP-adjusted
Eosinophil count	***IL33***	2.19E-18	8.64E-15	2.01E-03	6.85E-03
	HIST1H1A	4.17E-16	1.62E-07	6.78E-01	6.81E-01
	HIST1H1C	4.67E-19	8.25E-09	2.94E-01	2.80E-01
	HIST1H1T	5.17E-15	2.67E-07	9.48E-01	9.58E-01
	HIST1H2BF	1.05E-13	3.31E-07	2.65E-03	2.77E-03
	HIST1H3E	1.76E-17	1.94E-08	1.46E-01	1.48E-01
	HIST1H4F	1.39E-18	2.35E-09	7.66E-01	7.71E-01
	BTN3A2	1.23E-17	2.32E-09	3.56E-01	3.56E-01
	BTN2A2	8.02E-21	7.20E-11	5.59E-01	5.63E-01
	BTN3A3	5.36E-17	3.80E-09	9.55E-01	9.65E-01
	BTN2A1	1.65E-17	6.95E-10	9.75E-01	1.00E+00
	BTN1A1	4.33E-13	2.06E-07	1.00E+00	9.92E-01
	ABT1	4.12E-18	3.55E-10	2.15E-01	2.18E-01
	HIST1H2AG	2.71E-16	5.81E-15	4.09E-02	4.13E-02
	HIST1H2AH	3.72E-19	1.51E-17	2.75E-01	2.73E-01
	PRSS16	1.66E-12	1.75E-11	6.60E-02	6.62E-02
	POM121L2	1.28E-18	2.98E-17	6.73E-01	6.76E-01
	ZNF391	4.30E-26	1.74E-24	4.71E-01	4.69E-01
	HIST1H2BM	4.32E-17	4.01E-16	3.29E-02	3.25E-02
	HIST1H2AK	8.31E-11	9.96E-10	1.73E-01	1.71E-01
	HIST1H2BO	2.90E-10	1.61E-09	7.40E-01	7.42E-01
	OR2B2	1.36E-23	5.80E-22	4.20E-02	4.27E-02
	OR2B6	2.21E-08	2.01E-07	6.87E-02	6.84E-02
	ZNF165	3.50E-22	1.48E-20	9.55E-01	9.36E-01
	ZSCAN16	2.52E-21	5.39E-20	2.41E-01	2.41E-01
	ZKSCAN8	2.26E-07	2.96E-07	6.54E-01	6.40E-01
	ZSCAN9	2.89E-13	9.59E-12	3.44E-01	3.40E-01
	ZKSCAN4	1.14E-15	2.63E-14	4.15E-01	4.18E-01
	NKAPL	9.19E-16	4.63E-15	5.71E-01	5.54E-01
	PGBD1	5.04E-24	1.99E-22	1.06E-01	1.06E-01
	ZSCAN31	1.12E-28	1.21E-26	4.16E-01	4.31E-01
	ZSCAN12	5.12E-20	1.43E-18	4.51E-01	4.55E-01
	ZSCAN23	7.33E-10	3.59E-09	5.07E-01	5.07E-01
	GPX6	5.40E-17	8.29E-16	3.86E-01	3.89E-01
	PSORS1C2	8.36E-26	3.71E-22	5.32E-01	5.43E-01
	RPH3A	2.27E-07	7.63E-14	3.16E-01	2.74E-01
	OAS3	5.49E-04	2.76E-08	2.07E-02	1.63E-02
	GFI1B	1.93E-07	1.92E-07	3.96E-01	3.99E-01
Mean corpuscular haemogliobin	***KLF1***	1.88E-20	6.79E-21	9.11E-15	1.51E-14
	GMPR	6.60E-12	7.60E-11	2.94E-06	7.73E-06
	HIST1H2BA	6.40E-31	9.26E-28	5.50E-01	4.65E-01
	SLC17A2	2.58E-03	1.29E-08	5.34E-01	5.29E-01
	HIST1H2AB	2.13E-20	5.30E-21	9.81E-01	9.13E-01
	HFE	1.39E-02	6.90E-10	5.68E-01	5.44E-01
	HIST1H4C	1.60E-16	2.13E-26	3.33E-01	3.47E-01
	HIST1H2BD	2.86E-05	4.24E-10	7.62E-01	8.38E-01
	HIST1H4D	4.02E-69	3.82E-69	7.34E-01	7.21E-01
	HIST1H2BG	2.89E-19	9.86E-17	5.73E-01	5.13E-01
	HIST1H2AE	1.68E-12	5.62E-08	6.05E-01	7.18E-01
	HIST1H1D	1.28E-05	6.79E-15	6.73E-01	6.17E-01
	BTN2A1	1.98E-07	2.07E-19	3.18E-01	2.97E-01
	HIST1H2AG	6.61E-11	6.77E-08	3.86E-01	2.63E-01
	HIST1H4I	2.01E-17	1.42E-22	4.39E-02	3.64E-02
	ZNF184	3.95E-33	1.74E-07	1.48E-01	1.66E-01
	CHEK2	5.58E-08	1.43E-07	1.44E-04	2.19E-04
	PSORS1C2	2.49E-17	1.47E-17	9.16E-01	9.23E-01
Mean platelet thrombocyte volume	***GP1BA***	3.48E-20	1.82E-19	8.84E-08	1.03E-07
	HIST1H4D	2.54E-07	7.26E-08	1.84E-01	2.10E-01
	POM121L2	2.39E-06	1.46E-08	2.73E-01	2.84E-01
	OR2B6	6.48E-08	8.39E-08	3.69E-01	3.64E-01
	CHEK2	5.98E-08	1.93E-07	1.21E-01	1.38E-01
	POLK	2.01E-11	4.17E-09	5.93E-01	6.37E-01
	IQGAP2	4.49E-36	4.40E-34	3.72E-15	2.55E-15
	CENPBD1	1.64E-07	2.61E-07	2.67E-01	2.74E-01
	MLXIP	4.23E-10	6.29E-10	2.31E-03	1.12E-03
	KALRN	2.31E-14	3.79E-12	3.85E-18	3.01E-17
	ZNF664	2.11E-09	3.98E-08	7.75E-01	7.94E-01
	OR6F1	7.01E-12	1.44E-08	4.78E-01	7.77E-01
Platelet count	*GP1BA*	5.97E-08	1.43E-07	3.59E-05	4.47E-05
	ABT1	7.91E-02	1.10E-07	4.78E-03	5.28E-03
	MPL	3.08E-07	1.99E-07	1.86E-04	1.54E-04
	IQGAP2	5.26E-08	6.52E-08	1.72E-04	1.13E-04
	KCTD10	1.49E-10	2.11E-07	1.30E-02	1.28E-02
	TCHP	2.76E-13	1.73E-11	3.98E-01	3.73E-01
	RPH3A	4.06E-08	4.82E-13	9.14E-01	9.57E-01

Trait	Gene	LoF-segment burden not SNP-adjusted	LoF-segment burden SNP-adjusted	WES LoF burden not SNP-adjusted	WES LoF burden SNP-adjusted
Platelet distribution width	*GP1BA*	7.51E-10	4.26E-09	7.37E-06	9.78E-06
	TUBB1	6.10E-12	7.38E-12	7.34E-18	1.23E-17
	APOA5	1.12E-08	1.94E-08	2.14E-01	2.39E-01
Red blood cell count	*BTN2A1*	7.79E-11	7.65E-08	5.42E-01	5.39E-01
	POM121L2	6.22E-08	2.52E-07	2.56E-01	2.53E-01
	HIST1H2BM	8.71E-08	1.02E-07	4.18E-01	4.17E-01
	HIST1H2BO	2.79E-08	7.86E-08	1.70E-01	1.65E-01
	ZNF165	3.54E-08	4.79E-08	8.95E-01	9.59E-01
	ZSCAN9	4.04E-11	1.39E-10	5.86E-01	5.75E-01
	ZKSCAN4	9.64E-09	2.29E-08	5.24E-01	5.04E-01
	PGBD1	1.05E-07	1.24E-07	6.66E-01	7.48E-01
	ZSCAN31	3.36E-08	3.75E-08	7.18E-01	7.53E-01
	GPX6	7.71E-07	1.11E-07	4.62E-01	4.64E-01
	PSORS1C2	3.16E-11	2.33E-10	1.61E-01	1.55E-01
Red blood cell distribution width	***KLF1***	4.60E-33	3.49E-34	6.95E-13	5.69E-13
	HIST1H1A	1.48E-27	1.42E-07	9.81E-01	9.87E-01
	HIST1H3A	7.40E-28	6.91E-08	2.60E-01	2.90E-01
	HIST1H1C	2.01E-30	1.21E-08	9.70E-01	9.66E-01
	HIST1H1T	1.29E-25	6.54E-08	3.33E-01	4.61E-01
	HIST1H4D	1.04E-33	1.86E-08	8.85E-01	8.07E-01
	HIST1H4F	9.32E-25	1.60E-07	1.54E-01	1.87E-01
	BTN3A2	4.12E-23	1.95E-07	7.95E-01	7.46E-01
	HIST1H2AG	3.48E-17	1.90E-12	4.03E-01	3.08E-01
	HIST1H2AH	1.07E-14	3.34E-08	8.55E-01	7.89E-01
	ZNF391	1.44E-20	3.03E-15	2.88E-01	3.13E-01
	HIST1H2BM	2.06E-13	5.75E-08	7.22E-01	7.06E-01
	HIST1H2AM	1.79E-09	3.36E-07	3.34E-01	3.26E-01
	ZNF165	1.35E-14	3.51E-13	7.91E-01	7.54E-01
	ZSCAN16	7.53E-12	1.09E-10	5.50E-01	5.76E-01
	NKAPL	1.11E-13	1.81E-10	8.64E-01	8.35E-01
	PGBD1	1.58E-12	2.18E-11	4.90E-01	4.88E-01
	ZSCAN31	1.06E-12	3.77E-11	6.55E-01	5.65E-01
	ZSCAN12	1.22E-17	8.79E-12	4.83E-01	4.22E-01
	GPX6	8.37E-19	6.20E-13	4.76E-01	4.89E-01
	APOC3	1.37E-12	3.67E-11	2.13E-07	2.57E-07
	GFI1B	3.93E-07	8.10E-08	6.32E-01	6.40E-01

Supplementary Table 7. Exome-wide significant genes by LoF-segment burden, and comparison to other tests. A detailed comparison between the four association tests we performed. We report all genes found significant using SNP-adjusted LoF-segment burden (using 303,125 non-sequenced samples; 111 associations in total), as well corresponding p-values for LoF-segment burden without SNP-adjustment, and WES-LoF association (using 34,422 sequenced samples) with or without SNP-adjustment. P-values are computed using two-sided t-tests. Genes in bold correspond to hits previously reported by Van Hout et al. [8], for a total of eight replicated signals at exome-wide significance in the non-sequenced cohort. SNP-adjustments resulted in decreased significance for the LoF-segment burden test, but had a smaller effect on the WES-LoF-based approach; in total WES-LoF with SNP-adjustment achieved similar p-values in all but one (*GMPR*) of the exome-wide significant hits obtained by the simple WES-LoF test.

	Gene	Trait	WES LoF burden
1	*TET2*	Eosinophil count	1.79E-08
2	*GMPR*	Mean corpuscular haemoglobin	2.94E-06
3	*KLF1*	Mean corpuscular haemoglobin	9.11E-15
4	*TUBB1*	Mean platelet (thrombocyte) volume	3.01E-07
5	*GP1BA*	Mean platelet (thrombocyte) volume	8.84E-08
6	*IQGAP2*	Mean platelet (thrombocyte) volume	3.72E-15
7	*KALRN*	Mean platelet (thrombocyte) volume	3.85E-18
8	*TUBB1*	Platelet count	7.45E-07
9	*ASXL1*	Platelet distribution width	1.44E-06
10	*TUBB1*	Platelet distribution width	7.34E-18
11	*MAPK8*	Red blood cell distribution width	1.33E-06
12	*APOC3*	Red blood cell distribution width	2.13E-07
13	*KLF1*	Red blood cell distribution width	6.95E-13

Supplementary Table 8. Exome-wide significant genes by our WES-LoF Burden. Association statistics for the 13 significant loci detected by our WES-LoF burden test (not SNP-adjusted), similar to Van Hout et al. [8] but using a simple linear model and $34,422$ exome-sequenced samples. P-values are computed using two-sided t-tests.

	Gene	Trait	WES LoF burden p	WES LoF burden R^2	LoF-segment burden p	LoF-segment burden R^2	R^2_{prop} (%)
1	*IL33*	Eosinophil count	2.01E-03	2.85E-04	8.64E-15	2.06E-04	72.26
2	*GP1BA*	Mean platelet thrombocyte volume	8.84E-08	8.51E-04	1.82E-19	2.77E-04	32.57
3	*TUBB1*	Platelet distribution width	7.34E-18	2.20E-03	7.38E-12	1.60E-04	7.25
4	*TUBB1*	Mean platelet thrombocyte volume	3.01E-07	7.81E-04	2.15E-03	3.21E-05	4.11
5	*TUBB1*	Platelet count	7.45E-07	7.29E-04	4.21E-05	5.71E-05	7.84
6	*HBB*	Red blood cell distribution width	3.49E-02	1.32E-04	2.25E-03	3.18E-05	23.99
7	*HBB*	Red blood cell count	7.95E-02	9.15E-05	2.68E-02	1.67E-05	18.23
8	*KLF1*	Red blood cell distribution width	6.95E-13	1.53E-03	3.49E-34	5.06E-04	32.99
9	*KLF1*	Mean corpuscular haemoglobin	9.11E-15	1.79E-03	6.79E-21	2.99E-04	16.76
10	*ASXL1*	Platelet distribution width	1.44E-06	6.91E-04	1.58E-01	6.78E-06	0.98
11	*ASXL1*	Red blood cell distribution width	8.23E-04	3.33E-04	3.15E-01	3.43E-06	1.03
12	*KALRN*	Mean platelet thrombocyte volume	3.85E-18	2.24E-03	3.79E-12	1.64E-04	7.33
13	*IQGAP2*	Mean platelet thrombocyte volume	3.72E-15	1.84E-03	4.40E-34	5.04E-04	27.43
14	*GMPR*	Mean corpuscular haemoglobin	2.94E-06	6.50E-04	7.60E-11	1.44E-04	22.18

Supplementary Table 9. Comparison between association analyses. Association statistics for the 14 significant loci detected by Van Hout et al. [8] reported in Table 1, here including the p-value and the phenotypic variance explained (R^2) by each approach. The last column reports the estimated proportion of the phenotypic variation (R^2_{prop}, in %) of the sequenced samples that can be explained by the non-sequenced cohort; on average that is 19.64% for all the 14 reported associations, or 27.35% when restricting to the exome-wide significant signals.

Supplementary References

[1] Browning, B. L. & Browning, S. R. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics* **194**, 459–471 (2013). URL https://www.genetics.org/content/194/2/459.

[2] Gusev, A. *et al.* Whole population, genomewide mapping of hidden relatedness. *Genome research* **19**, 318–26 (2008).

[3] Naseri, A., Liu, X., Tang, K., Zhang, S. & Zhi, D. RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome Biology* **20** (2019).

[4] Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203 (2018).

[5] Umeyama, S. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13**, 376–380 (1991).

[6] Gusev, A. *et al.* The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution* **29**, 473–486 (2011).

[7] Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

[8] Van Hout, C. V. *et al.* Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at https://www.biorxiv.org/content/10.1101/572347v1 (2019).

[9] Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19**, 807–812 (2011).

[10] Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics* **50**, 906 (2018).

[11] Mathieson, I. & McVean, G. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics* **44**, 243 (2012).

[12] Albers, P. K. & McVean, G. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS biology* **18**, e3000586 (2020).

Chapter 3

Inference of pairwise coalescence times and allele ages using deep neural networks

The attached paper is published in the 2021 *NeurIPS LMRL Workshop* (Nait Saada et al., 2021) and has overlapping material with Chapter 4. A statement of contribution is attached at the end of the chapter.

Inference of pairwise coalescence times and allele ages using deep neural networks

Juba Nait Saada
Department of Statistics
University of Oxford

Anthony Hu
Department of Engineering
University of Cambridge

Pier Francesco Palamara
Department of Statistics
University of Oxford

Abstract

Accurate inference of the time to most recent common ancestor (TMRCA) between pairs of individuals using genetic variation plays a key role in a number of population genetic analyses. We developed CoalNN, an algorithm that leverages deep neural networks to predict pairwise TMRCAs, recombination breakpoints, and the age of genomic variants using raw genomic data. CoalNN can be used to analyze sequencing, SNP array, or imputed genotype data and can be adapted to varying population parameters such as demographic history or recombination rate maps. CoalNN is scalable to large data sets and improves upon available methods in TMRCA and allele age prediction in a variety of simulated conditions. These results underscore the efficacy of deep neural networks in the analysis of genealogical relationships in large genomic data sets.

1 Introduction

The genomes of two individuals from a population are connected through genealogical relationships that lead to common ancestors. The distance, in generations, that separates these individuals and their common ancestor at a specific genomic location is referred to as time to most recent common ancestor (TMRCA), or coalescence time [1, 2]. Accurate prediction of pairwise coalescence times is a key element in several genomic analyses, such as the reconstruction of a population's demographic history [3] or the inference of identical-by-descent (IBD) segments [4], which in turn may be applied to identify signatures of natural selection [5, 6] or detect association of rare genomic variants to traits and diseases [7, 8].

We developed a model, CoalNN, that leverages deep neural networks to infer pairwise TMRCAs. Deep learning algorithms achieve state-of-the-art performance in fields such as computer vision [9, 10, 11] and natural language processing [12, 13, 14], and are now emerging as an effective tool in several genomic applications. These include predicting functional effects of noncoding variants [15, 16, 17, 18], basecalling of nanopore data [19], identifying the sequence specificities of DNA- and RNA-binding proteins [20], inferring demographic history and population structure [21, 22, 23], and generating synthetic data [24, 25, 26]. However, deep neural networks have not yet been applied to perform explicit inference of genealogical relationships. We build a convolutional neural network (CNN) that enables predicting locus specific pairwise coalescence times, as well as the location of recombination breakpoints that result in a change in coalescence time. Simulation-based training of deep neural networks has been shown to offer several advantages over other likelihood-free inference strategies such as Approximate Bayesian Computation (ABC) [27]. We use a "simulation-on-the-fly" training procedure that prevents CoalNN from overfitting and accounts for imbalanced distributions in training data.

Using extensive simulations of sequencing, SNP array, and imputed data (see Figure 1) and comparison to available methods, we find that CoalNN offers improved performance for TMRCA prediction under several evaluation metrics. By predicting recombination breakpoints that result in TMRCA

Learning Meaningful Representations of Life (NeurIPS 2021).

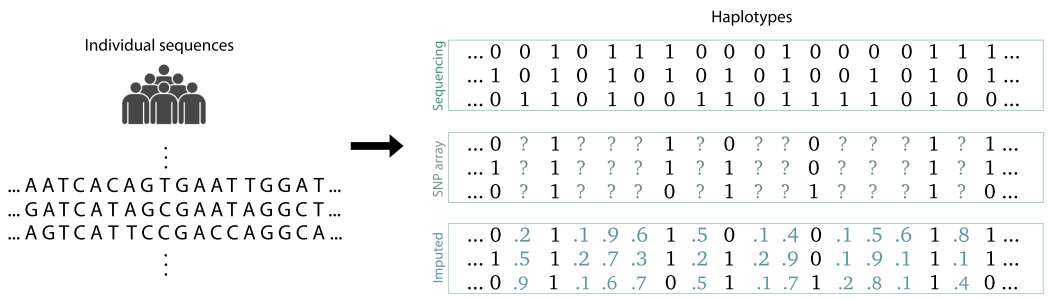


Figure 1: **Input types.** CoalNN can be used to analyze phased sequencing, SNP array, or imputed genotype data. The input haplotypes (i.e. sets of genomic variants located on the same chromosome) are obtained from individual DNA sequences consisting of nucleotides (A, C, G or T). Observations have value 0 if the individual carries the ancestral allele, and value 1 otherwise (derived state). Sequencing data comprises all polymorphic sites, while SNP array data only includes specific variants. Imputed data consists of estimated probabilities for inferred missing genotypes.

changes, CoalNN allows the predicted TMRCAs to have a piecewise constant structure that mimicks the underlying coalescent process and leads to improved compression of the output. Through transfer learning we show that CoalNN can be efficiently retrained to different evolutionary scenarios, such as different demographic models. Finally, we use CoalNN to estimate the age of observed genomic variants. Dating variants has a number of downstream applications such as quantifying the strength of natural selection acting on heritable traits and diseases [28]. We compare our results to leading methods for variant dating [29, 30, 31] and observe improved accuracy across several metrics in simulation.

2 Related Work

Current leading approaches for the inference of TMRCAs rely on probabilistic modeling based on stochastic processes such as the coalescent with recombination [32, 33] and inference is performed using algorithms such as Hidden Markov Models (HMMs) [3, 34, 35, 36, 37, 38]. Because accurate probabilistic modeling is often intractable, coalescent-based HMMs rely on a number of model simplifications, including Markovian approximations [39, 40] and the discretization of TMRCA values within user-specified time intervals.

A related class of methods enables inferring the ancestral recombination graph (ARG) for a set of individuals [41, 29, 42, 31, 43]. The ARG compactly represents the evolutionary history of a set of samples, and may be used to extract pairwise TMRCAs. These approaches generally rely on various levels of probabilistic modeling under simplified versions of the coalescent (e.g. the Li and Stephens model [44]) combined with scalable heuristics. Methods that more accurately model the coalescent process are generally less scalable and do not model biases found in common data modalities such as SNP array or imputed data. Highly scalable methods, on the other hand, rely on fast approximations that do not allow modeling of TMRCAs or evolutionary features such as non-homogenous demographic history.

3 Methods

3.1 Model architecture

For a given pair of haploid individuals and a chromosomal region, our goal is to jointly predict a positive value for the TMRCA (a regression task on $\mathbb{R}^+$) and the probability of observing a recombination event that affects TMRCA (a binary classification task) at each genomic site. We apply a convolutional neural network on $L = \sim 10$ centimorgans (cM) windows to which we add some context of fixed length on either side to avoid any padding, resulting in an input sequence of length $L_1 > L$. Given the approximately Markovian nature of the recombination process [39, 40], the local connectivity of convolutional modules is well suited for our goals. Consistent with this, we have tried several more complex architectures that can capture long-range dependencies such as transformers [12] or residual neural networks [45], but they did not perform better than convolutions.

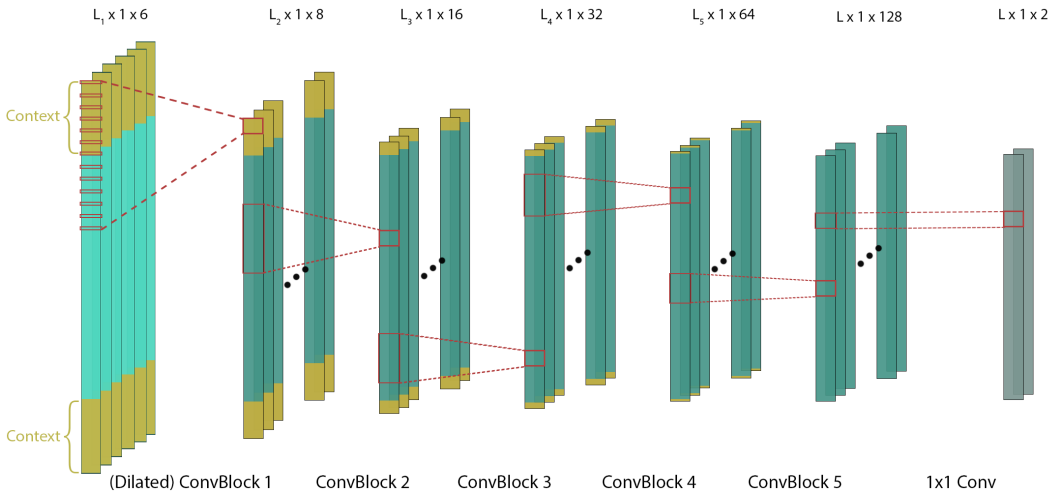


Figure 2: **CoalNN architecture.** CoalNN comprises a batch normalisation layer followed by five convolution blocks (convolution layer + batch normalisation + ReLU) and a 1×1 convolution layer. The first ConvBlock uses dilated convolution with a large kernel size to capture long-range dependencies, while subsequent layers focus on smaller and smaller windows (kernel sizes of $701, 201, 51, 7, 3$ respectively). The input sequence includes additional contextual data (in yellow).

For a given genotype, observations x_i, $i \in \{1, ..., L_1\}$, have value 1 if the individual carries a mutation (derived state) and 0 otherwise (ancestral state) in sequencing and SNP array data, or floating-point numbers in $[0, 1]$ (reflecting the inferred probability of carrying a mutation) in imputed data (see Figure 1). We construct a $L_1 \times 1 \times 6$ input tensor by computing the AND and the XOR logic gates from the two observed genotypes (or use the raw floating-point haplotypes as input for imputed data), the physical (in base pairs, bp) and the genetic (in centimorgans, cM) distances between consecutive sites, the minor allele frequencies (MAF) and the length of identical-by-state (IBS) segments (number of consecutive identical sites in both directions).

We first apply a batch normalisation layer and we stack five convolution blocks, each of them consisting of a convolution layer, a batch normalisation layer, and a ReLU activation function. Since recent TMRCAs are characterized by long (> 1 cM) and nearly identical haplotypes, we increase the receptive field of our network without affecting the computation and memory costs by utilizing a dilated convolution in the first block. The dilated factor is adaptive and is computed so that the receptive field approximately corresponds to 5 cM. After each convolution block, the sequence length is reduced until ultimately reaching L ($L_1 > L_2 > L_3 > L_4 > L_5 > L$). We finally apply a 1×1 convolution layer to reduce the channel dimension. The resulting output is a $L \times 1 \times 2$ tensor, containing the TMRCA estimates and the unscaled probabilities (logits) of observing a recombination breakpoint for each of the L genomic loci. The network architecture, which resulted in 130K trainable parameters, is illustrated in Figure 2. When deploying CoalNN, we apply an additional scaling layer on both estimates: the predicted TMRCAs that exceed the maximum coalescence time observed while training are clipped and the logits are fed to a softmax.

Because the TMRCA prediction regression task and the breakpoint prediction classification task are directly linked, we opted for a multi-task learning approach, using a shared representation that would enable leveraging commonalities across the two tasks [46]. We adopted the huber loss function to train the TMRCA regression task and the cross entropy loss for the breakpoint classification task. We implemented the method introduced in [47] to simultaneously learn both objectives using homoscedastic uncertainty.

3.2 Training procedure

We divided the genome in 39 autosomal regions from different chromosomes or separated by centromeres. This enabled us to parallelly train and test CoalNN across the entire genome and prevented issues due to low marker density in centromeres. We trained three different models, one for each type of data (sequencing, SNP array, and imputed). For each of them, we used the msprime simulator (v.1.0) [48] to generate training and validation sets. Samples within each simulation are

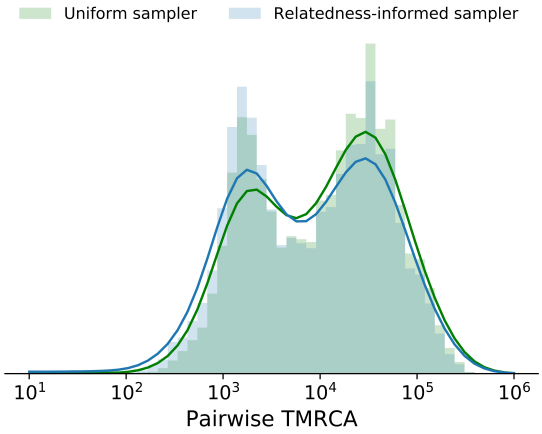


Figure 3: **Pairwise TMRCA sampling distribution.** Distributions of pairwise TMRCA in generations (histograms) obtained from a uniform sampler (in green) and a relatedness-informed sampler (in blue). Solid curves show kernel density estimations using Gaussian kernels (bandwidth $= 0.25$).

not independent due to shared underlying genealogical relationships, so that relying on a single simulation may lead overfitting. To prevent this from happening, at the start of each training epoch CoalNN performs the following steps, which may be parallelized over multiple cores:

1. n_{sim} pairs of random seeds and chromosomal regions are uniformly sampled in $\{2, ..., 2^8\} \times \Gamma$, where Γ denotes the set of 39 chromosomal regions.

2. For each (seed, chromosomal region) pair, a training dataset of n diploid individuals is generated under a user-specified demographic model with constant mutation rate $\mu = 1.65 \times 10^{-8}$ per bp per generation and human recombination rates [49, 50]. For imputed data, the diploid reference panel size n_{ref} is randomly sampled in $\{150, ..., 1000\}$ and we generate $n + n_{ref}$ samples.

3. Variants are downsampled following a data-specific procedure: for sequencing data, only polymorphic variants of n_d random individuals are retained, so that the distance between consecutive variants remains constant regardless of sample size; for SNP array data, polymorphic variants are subsampled to match the genotype density and allele frequency spectrum observed in the UK Biobank (UKBB) dataset [51]; for imputed data, variants are first downsampled following the SNP array data procedure, and then imputed from a simulated reference panel using Beagle 5.1 [52].

Using this approach, each batch then consists of N data points from different simulations, preventing CoalNN from overfitting a single set of parameters without resorting to explicit regularization. On the other hand, the validation set is generated only once, using the same procedure as the one described for the training set but with a fixed random seed of 1 and recombination rates from the first 30 mega base pairs (Mbp) of chromosome 2 (with an average recombination rate of 1.66 cM per Mbp).

Our analysis used $N = 64$, $n_{sim} = 64$, $n = 150$, $n_d = 150$ and the Adam optimizer with a learning rate of 0.001.

3.3 Sampling procedure for imbalanced data

We identified two sources of imbalanced data in our simulations: recombination events and recent TMRCAs ($< 1,000$ generations) are both extremely rare (the expected TMRCA for a randomly selected genomic site is in the order of several thousands of generations, Figure 3). The former was resolved by introducing weights to the cross entropy component of the loss. For the latter, we introduced a sampling procedure to construct input batches, which we refer to as "relatedness-informed sampling". For a given simulation consisting of $2n$ haploid individuals in our training set, the first pair of haplotypes is randomly sampled. The relatedness-informed sampler then builds a pair by uniformly picking one of the two processed haplotypes and then selecting the haplotype with smallest average TMRCA with them among the $2n - 2$ remaining samples. The same procedure is repeated until all haplotypes have been paired. This approach enables oversampling recent TMRCAs. Figure 3 shows how TMRCAs from the relatedness-informed sampler compare to those from the

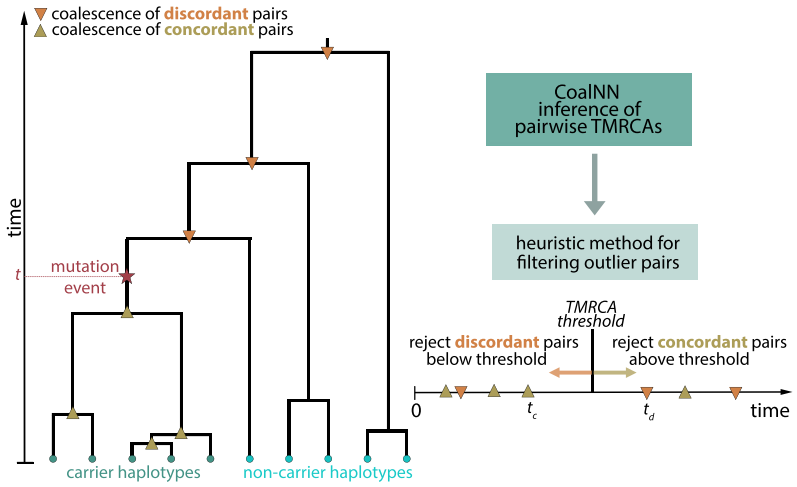


Figure 4: **Allele age inference.** At a given genomic site, individuals are connected through the sharing of underlying genealogical relationships (left). We aim to infer the time t at which a mutation arose (in red) and resulted in carrier haplotypes (in green) and non-carrier haplotypes (in cyan). We first infer TMRCAs across all concordant (two carriers) and discordant (one carrier and one non-carrier) pairs of haplotypes using CoalNN and we then reject outlier pairs using a heuristic method [30] (right). The TMRCA threshold is computed by minimising the total number of rejected pairs. The predicted age estimate is obtained by averaging the maximum coalescence time across concordant pairs t_c and the minimum coalescence time across discordant pairs t_d.

uniform sampler. One downside of this approach is that the relatedness-informed sampler trains the network to expect more recent common ancestry, leading to downward biased TMRCA predictions. To minimize this issue, we alternate between uniform and relatedness-informed samplers every other epoch.

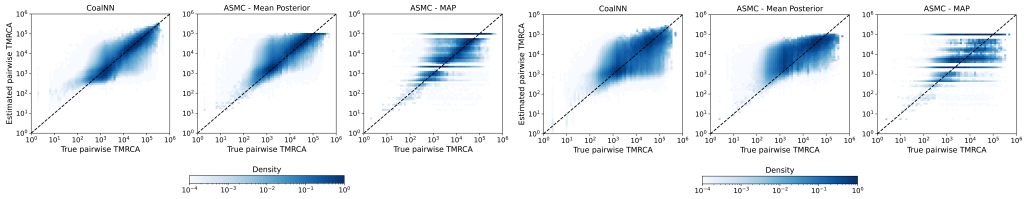
We also applied a log-transformation to the TMRCA ground truth. By predicting the log-TMRCA instead of the TMRCA itself, the training procedure penalizes the ratio between the ground truth and the prediction rather than the distance, preventing the loss from being dominated by large TMRCA values. Finally, we designed a validation procedure that helps focusing on recent TMRCAs. First, we only validated on the top 5 percent closest pairs of haplotypes. Second, we implemented a weighted huber loss as a validation score, giving more importance to under-represented TMRCA regions of the genome (i.e. both recent and extremely old TMRCAs). The model parameters were only saved if the validation score improved.

3.4 Dating genomic variants

We extended CoalNN to infer the origin (in generations before present) for observed genomic variants by utilizing the inferred pairwise coalescence times. For a given mutation, the maximum TMRCA across all concordant pairs (i.e. those for which the mutation is carried by both haplotypes) gives a lower bound on the mutation age, while the minimum TMRCA across all discordant pairs (i.e. the mutation is carried by only one haplotype) gives an upper bound, as shown in Figure 4 (left). However, in practice, because of the noise in our estimates, the lower bound can be bigger than the upper bound. To address this potential issue, we identify and filter out outlier pairs in TMRCA distributions using a heuristic method described in [30], which consists in rejecting all concordant (resp. discordant) pairs above (resp. below) a certain TMRCA threshold. The threshold is computed by minimising the total number of rejected pairs. The age estimate is finally obtained by averaging (arithmetic mean) the lower and upper bounds. The heuristic method for filtering outlier pairs is illustrated in Figure 4 (right). We compared our approach to several leading methods to date mutations. We observed a high bias in age estimates inferred by all methods for singletons and very recent mutations, which is explained by the absence of concordant pairs. We thus decided to exclude singletons from further analyses for all methods.

(a) Percent performance improvement of CoalNN over ASMC.

			ASMC mean posterior	ASMC MAP
Sequencing		MAE	17.94 (1.13)	70.57 (1.51)
		RMSE	20.1 (1.23)	62.67 (1.24)
Array		MAE	3.33 (0.67)	116 (1.44)
		RMSE	0.34 (0.79)	71.42 (1.16)
Imputed	panel size $= 300$	MAE	5.11 (0.88)	35.58 (0.99)
		RMSE	10.62 (1.13)	33.64 (1.14)
	panel size $= 1000$	MAE	5.49 (1.15)	43.71 (1.35)
		RMSE	12.08 (1.54)	42.63 (1.36)
	panel size $= 2000$	MAE	5.0 (0.9)	46.16 (1.02)
		RMSE	11.23 (1.2)	44.95 (1.08)



(b) sequencing data. (c) Array data.

Figure 5: **Pairwise TMRCAs prediction.** We report in (a) the average percent performance improvement of CoalNN over ASMC for the mean absolute error (MAE) and the root mean squared error (RMSE) across 20 simulations. Numbers in round brackets represent standard errors. For imputed data, reference panel sizes are in haploid units. We visualise true pairwise TMRCAs (x axis) versus those estimated by CoalNN and ASMC (y axis) on (b) sequencing data and (c) array data for one simulation.

4 Results

Experimental settings. We measured CoalNN's accuracy, running time, and robustness using extensive realistic coalescent simulations. Unless otherwise specified, all simulations use the following setup: we used the msprime simulator (v.1.0) [48] to simulate 150 diploid individuals and a chromosomal region of 30 Mbp, incorporating recombination rates from a human chromosome 2 and a recent demographic model for European individuals (Northern European CEU) [50], with random seeds not used in training ($> 2^8$). We simulate array data with SNP ascertainment matching UKBB allele frequencies. To simulate imputed data, array data was imputed using a simulated reference panel with Beagle 5.1 [52].

Performance on coalescence time prediction. We compared the pairwise TMRCA estimates inferred by CoalNN to both the Maximum-A-Posteriori (MAP) and the posterior mean estimates provided by ASMC (v.1.0), a recent coalescent HMM that enables efficiently estimating the posterior of the pairwise coalescence times [37]. We found that CoalNN outperforms ASMC on sequencing, array, and imputed data, as shown in Figure 5a. Pairwise TMRCA estimates for both methods on sequencing and array data are illustrated in Figure 5b and Figure 5c.

Running time. Training CoalNN took approximately 52, 9 and 23 hours on sequencing, array and imputed data respectively, using an Nvidia A100 GPU card and 6 CPUs. We then evaluated the computational efficiency of CoalNN compared to ASMC. ASMC is optimized to compute the posterior TMRCA probabilities across all the pairs of genomes in a specific order, reducing cache misses. It may also be run in a non-optimized version that decodes all pairs (or a given subset of pairs) using a user-specified order. The non-optimized version also saves the MAP and mean posterior age estimates, which causes additional computation and I/O. When using a GPU, we observed that CoalNN is around one order of magnitude faster than ASMC, as shown in Figure 6.

Transfer learning. We finally tested CoalNN's robustness to demographic model misspecification by simulating data under a constant population size of $10,000$ diploid individuals but loading model parameters trained on a European demographic model. Similarly, ASMC was optimized for a

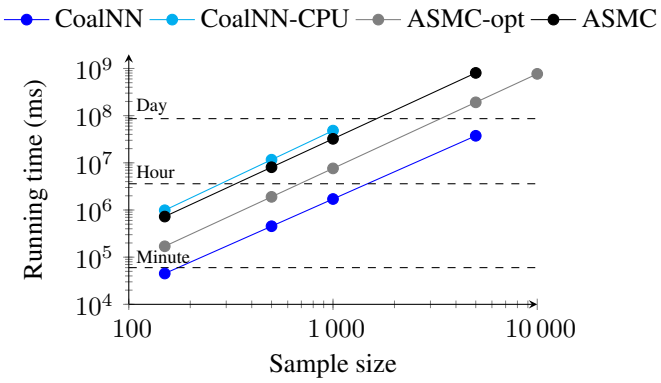


Figure 6: **Running time evaluation.** Running time (in milliseconds) of CoalNN (on a single A100 GPU card and a single CPU in blue, and on a single CPU only in cyan) and ASMC (on a single CPU, optimized version in grey and non-optimized version in black) on array data using the first 30 Mbp of chromosome 2 across $6,749$ SNPs. The batch size for both methods is 64.

European model. Although our simulations suggest that CoalNN generalizes better than ASMC on sequencing data (14.27% (SE=0.52) and 11.12% (SE=0.68) performance improvement over the ASMC mean posterior estimates on MAE and RMSE respectively across 20 simulations), ASMC is more robust than CoalNN on array data (-11.34% (SE=0.38) and -19.34% (SE=0.5)). We used transfer learning to re-train CoalNN on constant population size by initializing the neural network with the weights trained on a European demographic model and by fine-tuning all layers. Re-training CoalNN took approximately 30 and 5 hours on sequencing and array data respectively, using an Nvidia A100 GPU card and 6 CPUs. Using the European pre-trained model as a starting point enables to considerably reduce training time by reusing the prior knowledge gained by CoalNN (~ 1.5 and 5 times faster than training from scratch on sequencing and array data respectively). As expected based on the robustness of the model for sequencing data, the performance did not substantially improve (4.02% (SE=0.53) and 2.25% (SE=0.5) performance improvement over the previous model on MAE and RMSE respectively). On array data, on the other hand, this approach resulted in a substantial performance gain (16.44% (SE=0.87) and 16.18% (SE=0.93)).

Piecewise constant coalescence times. Due to the properties of the coalescent with recombination process [1, 2], TMRCAs along the genome are expected to be piecewise constant. However, the posterior mean estimate output by ASMC is not piecewise constant, while the MAP estimate is piecewise constant but can only take a small number of values and is less accurate. By using the probabilities of observing a recombination event at every locus, CoalNN enables producing piecewise

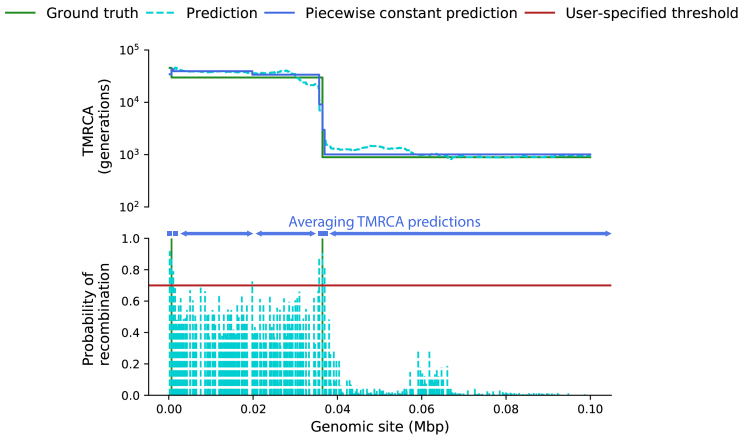
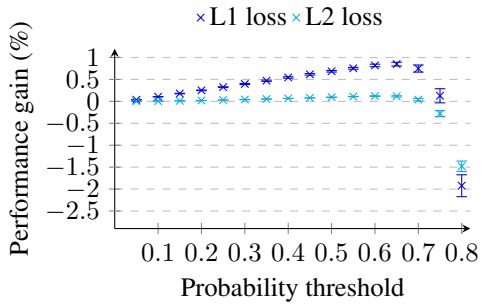
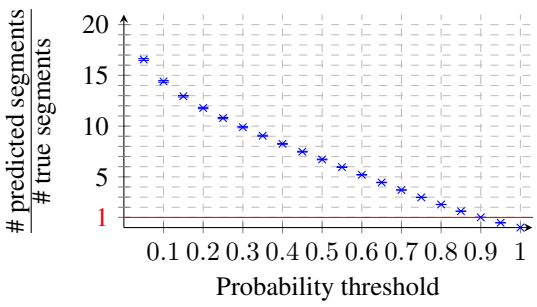


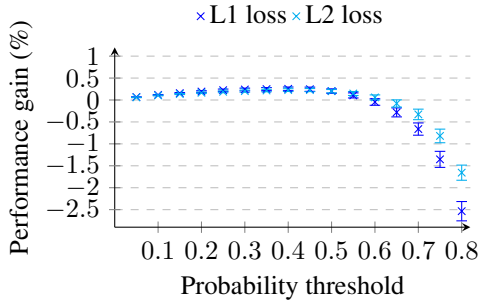
Figure 7: **Computing the piecewise constant coalescence times.** The inferred TMRCAs are averaged between every consecutive genomic site with an estimated probability of recombination breakpoint that exceeds a user-specified threshold.



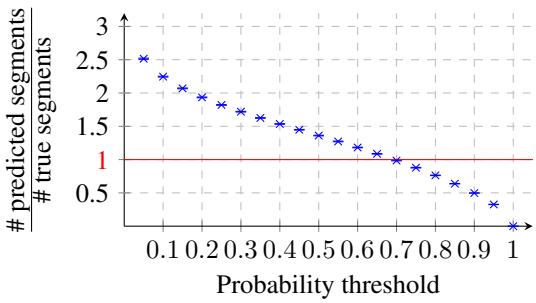
(a) Percent performance gain on sequencing data.



(b) Level of compression on sequencing data.



(c) Percent performance gain on array data.



(d) Level of compression on array data.

Figure 8: **Piecewise constant prediction**. We report in (a) (resp. in c) the average percent improvement of the piecewise constant TMRCA prediction over the raw output of CoalNN on the L1 and L2 losses for sequencing (resp. array) data, for increasing threshold values. We report in (b) (resp. in d) the average ratio between the total number of predicted segments and the total number of true segments for increasing threshold values, where a segment refers to a piece of the genome with constant TMRCA value. Error bars represent standard errors across 10 simulations.

constant predictions. To this end, we average the inferred coalescence times between genomic sites where the probability of observing a recombination event exceeds a user-specified threshold. This strategy is illustrated in Figure 7. In addition to better mimicking the true underlying TMRCAs, this approach improves the accuracy in predicting the TMRCA estimates for both sequencing and array data, with optimal probability thresholds around 0.7 and 0.55 for sequencing and array data respectively, as shown in Figure 8. Producing piecewise constant output also results in significant compression of the inferred TMRCAs when a run-length encoding is used.

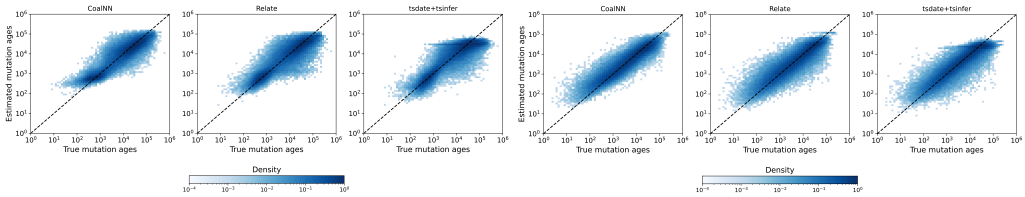
Performance on allele age prediction. We inferred mutation ages under a European demographic model and a constant population size ($N_e = 10,000$) using CoalNN, as well as Relate [29] and tsinfer+tsdate [42, 31] (tsinfer and tsdate run successively), two recent genealogical inference approaches that enable inferring ARGs. Both ARG methods infer lower and upper age estimates, which we averaged (arithmetic mean) to obtain a point estimate. CoalNN and Relate used the simulated demographic model, while tsinfer+tsdate assumes a constant population size of $10,000$ diploid individuals. In these simulations, CoalNN achieved highest accuracy on several metrics, as shown in Figure 9. We also tested the GEVA method [30], which however only dated 50.45% (SE=0.23) of the non-singletons variants due to multiple computational issues and had lower accuracy than the other methods we considered - e.g using the mode estimate of the composite posterior distribution obtained by GEVA under the joint clock model on a 10Mbp chromosome with recombination rates from chromosome 2 and a constant population size, CoalNN's accuracy was 67% (SE=3) and 89% (SE=3) higher on the RMSE and MAE respectively.

5 Conclusion

We developed CoalNN, a method that uses a deep neural network to predict pairwise coalescence times and recombination breakpoints from sequencing, array, or imputed genotype data. Using

(a) Accuracy evaluation.

		RMSE	MAE	MAD	r^2
CEU	**Relate**	24753 (215)	12355 (93)	4271 (46)	0.53 (0.0)
	tsinfer+tsdate	29685 (280)	14363 (118)	4695 (48)	0.37 (0.0)
	CoalNN	**23325 (154)**	**11689 (79)**	**4001 (43)**	**0.58 (0.0)**
Constant	**Relate**	11029 (117)	5341 (44)	1647 (12)	0.63 (0.0)
	tsinfer+tsdate	12909 (258)	5864 (68)	1557 (12)	0.53 (0.0)
	CoalNN	**10901 (125)**	**5108 (46)**	**1487 (11)**	**0.65 (0.0)**



(b) European demographic model.

(c) Constant population size.

Figure 9: **Dating variants.** We report in (a) the average performance (in generations) of Relate, tsinfer+tsdate and CoalNN on non-singleton variants under European demographic history model (CEU) and a constant population size ($N_e = 10,000$) across 10 simulations, with numbers in round brackets representing standard errors. We report the root mean squared error (RMSE), the mean absolute error (MAE), the median absolute deviation (MAD) and the square of the Pearson correlation coefficient (r^2). Methods in bold obtained statistically significantly lower (or higher) scores than non-bold methods (one-sided paired t-test $p \leq 0.05$). We visualise true non-singleton variants ages (x axis) versus those estimated by each method (y axis) under a European demographic model in (b) and under a constant population size ($N_e = 10,000$) in (c).

extensive simulations, we found that CoalNN improves upon several current approaches for the inference of TMRCAs and the dating of genomic variants. These results demonstrate that deep learning approaches can be effectively applied in population genetic analyses that involve inferring the underlying genealogical structure of a set sequenced or genotyped samples.

We highlight a few limitations of our approach, as well as directions of future and ongoing work. Compared to other methods, CoalNN requires a GPU card for optimal performance, but a trained model may also be used using standard CPU hardware. Although our simulations modeled a number of realistic evolutionary parameters, future work will explore robustness to additional features, such as phasing and genotyping errors, in addition to imputation errors. A desirable future improvement is to allow CoalNN to generalize to varying demographic models without the need for additional training. Furthermore, although we focused on sequencing, SNP array, and imputed data, it will be interesting to extend CoalNN to support analysis of low coverage sequencing and ancient DNA data. Finally, our application of CoalNN has been limited to the inference of pairwise TMRCAs. A direction of future work will be to explore extensions of this approach that simultaneously consider multiple samples, which may be achieved by leveraging permutation invariant or exchangeable neural networks [27], with applications that include haplotype phasing [53] and imputation [52, 54].

6 Acknowledgments

This work was supported by MRC grant MR/S502509/1 and Balliol Jowett Scholarship (to J.N.S.); Toshiba Europe grant G100453 (to A.H); Wellcome Trust ISSF grant 204826/Z/16/Z, NIH grant R21-HG010748-01, and ERC Starting Grant 850869 (to P.F.P.). We thank the Biomedical Research Computing team at the University of Oxford for support with the ResComp compute clusters. The research was partly supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z with additional support from the NIHR Oxford BR (the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health).

References

[1] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13 (3):235–248, 1982.

[2] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[3] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493, 2011.

[4] Juba Nait Saada, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev, and Pier Francesco Palamara. Identity-by-descent detection across 487,409 British samples reveals fine-scale evolutionary history and trait associations. *Nature Communications*, 11(1):6130, 2020. doi: 10.1038/s41467-020-19588-x.

[5] Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, 2010. doi: https://doi.org/10.1534/genetics.110.113977.

[6] Alexander Gusev, Pier Francesco Palamara, Gregory Aponte, Zhong Zhuang, Ariel Darvasi, Peter Gregersen, and Itsik Pe'er. The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution*, 29(2):473–486, 2011. doi: https://doi.org/10.1093/molbev/msr133.

[7] Alexander Gusev, Eimear E Kenny, Jennifer K Lowe, Jaqueline Salit, Richa Saxena, Sekar Kathiresan, David M Altshuler, Jeffrey M Friedman, Jan L Breslow, and Itsik Pe'er. Dash: a method for identical-by-descent haplotype mapping uncovers association with recent variation. *The American Journal of Human Genetics*, 88(6):706–717, 2011. doi: 10.1016/j.ajhg.2011.04.023.

[8] Sharon R Browning and Elizabeth A Thompson. Detecting rare variant associations by identity-by-descent mapping in case-control studies. *Genetics*, 190(4):1521–1531, 2012. doi: https://doi.org/10.1534/genetics.111.136937.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. URL http://arxiv.org/abs/1411.4038.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. URL http://arxiv.org/abs/1512.03385.

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017. URL http://arxiv.org/abs/1703.06870.

[12] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL http://arxiv.org/abs/1706.03762.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[14] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec

Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2005.14165.

[15] Jian Zhou and O. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12:931–934, 2015.

[16] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7):990–999, 2016.

[17] David Kelley, Yakir Reshef, Maxwell Bileschi, David Belanger, Cory McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28:gr.227819.117, 03 2018. doi: 10.1101/gr.227819.117.

[18] Jian Zhou, Chandra Theesfeld, Kevin Yao, Kathleen Chen, Aaron Wong, and Olga Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50, 08 2018. doi: 10.1038/s41588-018-0160-6.

[19] Haotian Teng, Minh Duc, Michael Hall, Tania Duarte, Sheng Wang, and Lachlan Coin. Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7, 12 2017. doi: 10.1093/gigascience/giy037.

[20] Babak Alipanahi, Andrew Delong, Matthew Weirauch, and Brendan Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature biotechnology*, 33, 07 2015. doi: 10.1038/nbt.3300.

[21] Sara Sheehan and Yun S. Song. Deep learning for population genetic inference. *PLOS Computational Biology*, 12(3):1–28, 03 2016. doi: 10.1371/journal.pcbi.1004845. URL https://doi.org/10.1371/journal.pcbi.1004845.

[22] Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, 2020. doi: https://doi.org/10.1111/1755-0998.13224. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13224.

[23] Jonas Meisner and Anders Albrechtsen. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *bioRxiv*, 2020. doi: 10.1101/2020.12.28.424587. URL https://www.biorxiv.org/content/early/2020/12/28/2020.12.28.424587.

[24] Nathan Killoran, Leo Lee, Andrew Delong, David Duvenaud, and Brendan Frey. Generating and designing dna with deep generative models. *Advances in Neural Information Processing Systems, Computational Biology Workshop (NeurIPS)*, 2017.

[25] Sam Sinai, Eric Kelsic, George Church, and Martin Nowak. Variational auto-encoding of protein sequences. *Advances in Neural Information Processing Systems, Computational Biology Workshop (NeurIPS)*, 2017.

[26] Daniel Mas Montserrat, Carlos Bustamante, and Alexander Ioannidis. Class-conditional vae-gan for local-ancestry simulation. *Advances in Neural Information Processing Systems, Computational Biology Workshop (NeurIPS)*, 2019.

[27] Jeffrey Chan, Valerio Perrone, Jeffrey P. Spence, Paul A. Jenkins, Sara Mathieson, and Yun S. Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. doi: 10.1101/267211.

[28] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421–1427, 2017.

[29] Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, 51:1321–1329, 2019. doi: https://doi.org/10.1038/s41588-019-0484-x.

[30] Patrick K Albers and Gil McVean. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS biology*, 18(1):e3000586, 2020.

[31] Anthony Wilder Wohns, Yan Wong, Ben Jeffery, Ali Akbari, Swapan Mallick, Ron Pinhasi, Nick Patterson, David Reich, Jerome Kelleher, and Gil McVean. A unified genealogy of modern and ancient genomes. *bioRxiv*, 2021. doi: 10.1101/2021.02.16.431497. URL https://www.biorxiv.org/content/early/2021/04/15/2021.02.16.431497.

[32] Richard R Hudson and Norman L Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 120(3):831–840, 1988. ISSN 0016-6731. URL https://www.genetics.org/content/120/3/831.

[33] Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical Population Biology*, 55(3):248–259, 1999. ISSN 0040-5809. doi: https://doi.org/10.1006/tpbi.1998.1403. URL https://www.sciencedirect.com/science/article/pii/S0040580998914034.

[34] Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

[35] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, 2014.

[36] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, 2017.

[37] Pier Francesco Palamara, Jonathan Terhorst, Yun S. Song, and Alkes L. Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50, 09 2018. doi: https://doi.org/10.1038/s41588-018-0177-x.

[38] Jeffrey P Spence, Matthias Steinrücken, Jonathan Terhorst, and Yun S Song. Inference of population history using coalescent hmms: review and outlook. *Current opinion in genetics & development*, 53:70–76, 2018.

[39] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[40] Asger Hobolth and Jens Ledet Jensen. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical population biology*, 98:48–58, 2014.

[41] Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014.

[42] Jerome Kelleher, Yan Wong, Anthony Wohns, Chaimaa Fadil, Patrick Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51:1330–1338, 09 2019. doi: 10.1038/s41588-019-0483-y.

[43] Brian C Zhang, Arjun Biddanda, and Pier Francesco Palamara. Biobank-scale inference of ancestral recombination graphs enables genealogy-based mixed model association of complex traits. *bioRxiv*, 2021. doi: 10.1101/2021.11.03.466843. URL https://www.biorxiv.org/content/10.1101/2021.11.03.466843v1.

[44] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[46] Rich Caruana. Multitask learning. *Machine Learning*, 28, 07 1997. doi: 10.1023/A:1007379606734.

[47] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, abs/1705.07115, 2017. URL http://arxiv.org/abs/1705.07115.

[48] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[49] John A. Kamm, Jeffrey P. Spence, Jeffrey Chan, and Yun S. Song. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics*, 203 (3):1381–1399, 2016. ISSN 0016-6731. doi: 10.1534/genetics.115.184820. URL https://www.genetics.org/content/203/3/1381.

[50] Jeffrey P. Spence and Yun S. Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10), 2019. doi: 10.1126/sciadv.aaw9206. URL https://advances.sciencemag.org/content/5/10/eaaw9206.

[51] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018. doi: https://doi.org/10.1038/s41586-018-0579-z.

[52] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3):338–348, 2018. doi: https://doi.org/10.1016/j.ajhg.2018.07.015.

[53] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443, 2016. doi: https://doi.org/10.1038/ng.3679.

[54] Simone Rubinacci, Olivier Delaneau, and Jonathan Marchini. Genotype imputation using the positional burrows wheeler transform. *PLoS Genetics*, 16(11):e1009049, 2020.

Statement of Authorship

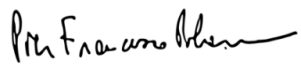
Title of Paper	Inference of pairwise coalescence times and allele ages using deep neural networks
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and unsubmitted work written in a manuscript style
Publication Details	Juba Nait Saada, Anthony Hu, and Pier Palamara. Inference of pairwise coalescence times and allele ages using deep neural networks. *Advances in Neural Information Processing Systems, LMRL Workshop (NeurIPS)*, 2021. URL `https://www.lmrl.org/papers`.

Student Confirmation

Student Name:	Juba Nait Saada		
Contribution to the Paper	Juba Nait Saada (J.N.S.) developed and implemented the model architecture and performed all analyses. Anthony Hu (A.H.) provided guidance on deep learning. Pier Palamara (P.P.) and J.N.S. planned the project. P.P. supervised the project. J.N.S. and P.P. wrote the manuscript.		
Signature 		Date	21/05/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. Pier Palamara		
Supervisor comments: Juba had a leading role in this project, as detailed above.		
Signature 	Date	26/06/2022

Chapter 4

Inference of genealogical relatedness and variant age using deep neural networks identifies signatures of selection

This chapter is an unpublished work written in a manuscript style and has overlapping material with Chapter 3. A statement of contribution is attached at the end of the chapter.

Inference of genealogical relatedness and variant age using deep neural networks identifies signatures of selection

Juba Nait Saada[1,*,†], Zoi Tsangalidou[1,*], Miriam Stricker[1], and Pier Francesco Palamara[1,2,†]

[1] *Department of Statistics, University of Oxford, Oxford, UK.*
[2] *Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK.*

[*] *These authors contributed equally to this work.*
[†] *correspondence: juba.naitsaada@stats.ox.ac.uk, palamara@stats.ox.ac.uk*

Abstract

Accurate inference of genealogical properties such as the time to the most recent common ancestor (TMRCA) between pairs of individuals and of the age of genomic variants is key in several population genetic analyses. We developed a likelihood-free approach, called CoalNN, which uses a deep neural network to predict pairwise TMRCAs and allele ages from sequencing, SNP array, or imputed genotype data. CoalNN is trained through simulation and can be adapted to varying parameters such as demographic history using transfer learning. CoalNN obtained improved accuracy compared to other available methods for TMRCA and allele age prediction in a range of simulated scenarios. We applied CoalNN to 2,504 samples from 26 populations in the 1,000 Genome Project data set and inferred ages for ∼80 million variants. We observed substantial inter-population variation, reflecting heterogeneous demographic histories, as well as variation across different levels of predicted variant pathogenicity, consistent with the action of negative selection. We used CoalNN's predicted allele ages to construct genome-wide annotations capturing the signature of past negative selection. We performed LD-score regression analysis of heritability using summary association statistics from 63 independent complex traits and diseases (average $N = 314$k), observing increased annotation-specific effects on heritability compared to a previous allele age annotation. These results underscore the efficacy of deep neural networks in analyses involving genealogical relationships in large genomic data sets.

Introduction

The genomes of two individuals from a population are connected through genealogical relationships that lead to common ancestors. The distance, in generations, that separates these individuals and their common ancestor at a specific genomic location is referred to as time to most recent common ancestor (TMRCA), or coalescence time [1, 2]. Accurate prediction of pairwise coalescence times is a key element in several genomic analyses, such as inference of the age of genomic variants [3], the reconstruction of a population's demographic history [4, 5], detection of identical-by-descent (IBD) segments [6], and the study of natural selection [7, 8, 9].

Current approaches for the inference of TMRCAs rely on probabilistic modeling based on stochastic processes such as the coalescent with recombination [10, 11]. Among these, coalescent hidden Markov models [4, 12, 13, 14, 15] (or "coalescent HMMs", reviewed in [16]) have received considerable interest during recent years. Other recently developed methods enable inferring

the ancestral recombination graph (ARG) for a set of individuals [17, 18, 19, 20, 21], which compactly represents the evolutionary history of a set of samples, including their TMRCAs. Despite considerable progress in the development of probabilistic inference algorithms, accurate modeling under the coalescent with recombination remains intractable. For this reason, all available methods for TMRCA inference resort to simplifying assumptions that trade inference accuracy for computational efficiency, such as the discretization of TMRCA values within time intervals, or the use of approximate genealogical models [22, 23, 24].

Difficulties in dealing with intractable likelihoods has motivated the development of "likelihood-free" inference strategies such as approximate Bayesian computation (ABC, [25, 26, 27]), where simulation is used to replace analytical modeling. These methods have been widely applied in population genetics and other fields [28]. Recent advances in probabilistic machine learning provided further momentum for the development of simulation-based inference [29], and simulation-based training of deep neural networks (DNNs) has been shown to offer advantages compared to approaches such as ABC [30]. More broadly, deep learning algorithms have achieved state-of-the-art performance in several domains [31, 32, 33, 34, 35, 36], and are now emerging as an effective tool in genomic applications. These include predicting functional effects of noncoding variants [37, 38, 39, 40], basecalling of nanopore data [41], identifying the sequence specificities of DNA- and RNA-binding proteins [42], inferring demographic history and population structure [43, 44, 45], detecting selective sweeps [46, 47], and generating synthetic data [48, 49, 50]. However, DNNs have not yet been applied to infer genealogical relationships across individuals from a population.

We developed an algorithm, called CoalNN, that uses a simulation-trained deep convolutional neural network (CNN) to jointly predict pairwise TMRCAs and recombination breakpoints, and further leverages these predictions to estimate the age of genomic variants. While not requiring explicit probabilistic modeling, CoalNN achieves higher inference accuracy compared to existing methodology in a variety of simulated scenarios and can be efficiently retrained to work with sequencing, imputed, and SNP array data, as well as different demographic parameters. Interpretability analyses of CoalNN's TMRCA predictions suggest that the network learns to rely on combinations of genomic features, such as allele frequency, recombination rates, and haplotype sharing, that require complex modeling in coalescent HMM approaches. We apply CoalNN to infer the age of ∼80 million variants identified in 26 populations from the 1,000 Genomes Project (1kGP) [5], and observe allele age variation consistent with population-specific demographic histories and the action of negative selection. Finally, we leverage CoalNN's predicted allele ages to construct genome-wide annotations capturing the signature of past negative selection. We use stratified LD-score regression (S-LDSC [51, 52]) and summary association statistics from 63 independent complex traits and diseases (average N=314k) to analyze trait heritability, detecting heritability effects that complement existing evolutionary annotations.

Material and methods

Overview of the CoalNN method

For a given pair of haploid individuals and a chromosomal region, we aim to predict a TMRCA value at each polymorphic genomic site. Rather than relying on probabilistic modeling, CoalNN is trained using realistic coalescent simulation and learns to map the genomic data provided in input to the target TMRCA value. To this end, CoalNN uses a deep convolutional neural network, whose input consists of the raw genotype values for each genomic variant, as well as a set of basic features. For each site, these include the minor allele frequency (MAF), physical and genetic distances (in base pairs [bp] and centimorgan [cM], respectively), and the number of consecutive identical-by-state (IBS) sites between the analyzed pair of individuals, in either direction.

For a pair of haploid individuals, TMRCAs along the genome may be represented as a piecewise constant function taking positive real values, where each TMRCA interval is delimited by past recombination events. Therefore, in order to facilitate the output of piecewise constant TMRCA values, CoalNN jointly predicts the presence of a recombination breakpoint at each site. These estimated breakpoint positions are then utilized to produce piecewise constant TMRCA predictions, as shown in Supplementary Fig. 1a. The overall network architecture, which comprises 130K trainable parameters, is illustrated in Fig. 1a.

Unlike coalescent HMMs, which rely on explicitly modeling of evolutionary parameters such as demographic history [4] and of the type of input data (e.g. sequencing, SNP array, or imputed data [15]), CoalNN training only requires data for each model variation to be efficiently simulated. A trained CoalNN model for a given configuration may then be adapted through transfer learning by fine-tuning its parameters across each network layer. While this approach introduces additional computation, it circumvents the need for complex probabilistic modeling, which often leads to working with simplified evolutionary models.

Finally, CoalNN implements a strategy described in [3] to process the set of predicted pairwise coalescence times and further produce an estimate for the time of origin of each observed genomic variant (Fig. 1b). More in detail, given a genomic variant, an age estimate is computed by averaging the maximum TMRCA across all concordant pairs (i.e. those for which the variant is carried by both haplotypes) and the minimum TMRCA across all discordant pairs (those for which the variant is carried by only one haplotype), as shown in Fig. 1c.

Neural network architecture

Since changes in the TMRCA value are linked to the presence of recombination breakpoints, CoalNN aims to jointly infer these target values using a multi-task learning approach that relies on an underlying shared representation [53]. We adopt the Huber loss function for the regression

task of TMRCA prediction and the cross entropy loss for the classification task of breakpoint prediction, implementing the method introduced by Kendall et al. [54] to simultaneously learn both objectives using homoscedastic uncertainty. We apply a convolutional neural network on genomic windows of fixed length L, where L is the number of sites necessary to cover the first 10 centimorgans (cM) of chromosome 2. Rather than using padding, we add a fixed amount of genomic data as context on either side of the input sequence, resulting in a region of length $L_1 > L$. Due to the approximately Markovian nature of the coalescent with recombination [23, 24] the local connectivity of convolutional blocks provides a suitable model for these prediction tasks. Consistent with this, we did not observe improved performance when using more complex architectures that allow capturing long-range dependencies, such as transformers [36] or residual neural networks [55].

The input genotype data is encoded as x_i, $i \in \{1, ..., L_1\}$, where x_i takes floating point values. For simulated SNP array and sequencing data $x_i \in \{0, 1\}$ indicates whether the haploid individual carries a minor or a derived allele, while in experiments involving imputed data x_i takes continuous dosage values in $[0, 1]$, reflecting imputation uncertainty. When encoded using floating-point values, raw genotype data is provided as input to the network. For binary data, we instead compute the AND and the XOR logic gates between the two haplotypes, which reflect the presence of mismatching or shared alleles, respectively.

We first apply a batch normalisation layer and stack five convolution blocks (ConvBlocks), each consisting of a convolution layer, a batch normalisation layer, and a ReLU activation function. The first convolution block aims to capture long range dependencies by using a large kernel size while subsequent layers focus on smaller and smaller windows (kernel sizes of 701, 201, 51, 7, 3 respectively). Since small TMRCAs correspond to shared ancestry in the recent past, they are characterized by long (> 1 cM) and nearly identical haplotype segments. To capture these segments, we increased the receptive field of the network by utilizing a dilated convolution in the first block. The dilation factor is automatically tuned so that the receptive field approximately corresponds to the first 5 cM of Chromosome 2. After each convolution block, the sequence length is reduced until ultimately reaching L ($L_1 > L_2 > L_3 > L_4 > L_5 > L$). We finally apply a 1×1 convolution layer to reduce the channel dimension. The resulting output is a $L \times 1 \times 2$ tensor, containing the TMRCA estimates and the unscaled estimated probabilities (logits) of observing a recombination breakpoint for each of the L sites. When deploying CoalNN for inference, we apply an additional scaling layer on both estimates: predicted TMRCAs that exceed the maximum coalescence time observed while training are clipped and a softmax function is applied to the logits.

Training procedure and simulation parameters

Given a set of evolutionary parameters, e.g. demographic model and mutation rate, a data modality, e.g. sequencing, SNP array, or imputed data, and any additional parameters, e.g. phasing and genotyping error rates, we use simulation to generate synthetic genotype and genealogical data for training and validation. Because all samples from a simulation are related through underlying genealogical relationships, relying on a single simulation during training would lead to overfitting. To circumvent this issue, at the start of each training epoch CoalNN generates data from 64 independent coalescent simulations, each providing a single pair of individuals for the current training batch. If a non-constant recombination rate is specified, each independent simulation also randomly samples a different genomic region, so that several subsets of the genetic map are observed during training.

For all validations and method comparisons, we simulated an input sequence of 30 cM for 150 individuals from a European (CEU [5]) demographic model, using the genetic map inferred in [56] and a constant mutation rate of 1.65×10^{-8}. When performing training and inference in sequencing data, we only retained variants that were polymorphic in 150 random individuals, so that the distance between consecutive variants remained approximately constant regardless of sample size. For SNP array data, polymorphic variants were subsampled to match the genotype density and allele frequency spectrum observed in the UK Biobank (UKBB) data set [57], as done in [6]. For imputed data, variants were first downsampled to create array-like data and then imputed using Beagle 5.1 [58] and a simulated diploid reference panel of size n_{ref}, where n_{ref} was randomly chosen to be between 150 and 1,000. When deploying the model to analysed data from the 1,000 Genomes Project (1kGP described below) we trained CoalNN with simulated sequencing data, using demographic models and genetic maps inferred for each population in [56]. We used the msprime simulator (v.1.0) [59] for all simulations and used the Adam optimizer with a learning rate of 0.001. We parallelized several steps of these procedures, such as the running of independent simulations, using multiple cores.

Sampling procedure for imbalanced data

Because both recombination breakpoints along the sequence and very recent TMRCAs are observed in only a minority of data points during training, they create possible issues due to imbalanced data. To address the low frequency of recombination breakpoints, we introduce weights to the cross entropy component of the training loss. In addition, we oversample recent TMRCAs by adopting a relatedness-informed sampling procedure during the construction of input batches. For a given training set simulation consisting of $2n$ haploid individuals, the first pair of haplotypes is randomly sampled. We then choose the next pair by uniformly sampling one of the haplotypes that have been previously processed and then selecting the haplotype with smallest average TMRCA (across the entire simulated chromosomal region) among the $2n - 2$

remaining samples. We repeat this procedure until all haplotypes have been paired. Although this approach increases the accuracy for recent TMRCA prediction, it introduces biases due to a deviation between observed and expected TMRCA distributions (see Supplementary Fig. 2). To mitigate this issue, we alternate between uniform and relatedness-informed sampling at every other training epoch.

To further address the scarcity of recent TMRCAs in the training data, we predict log-TMRCA rather than TMRCA values. This penalizes the ratio between ground truth and prediction, rather than their absolute distance, preventing the loss from being dominated by large TMRCA values. Finally, we designed a validation procedure that focuses on recent TMRCAs after each training epoch. To this end, we only used the top 5 percent closest pairs of haplotypes for validation and adopted a weighted huber loss as validation score, which assigns more importance to under-represented recent and extremely old TMRCA regions of the genome.

Transfer learning

A trained CoalNN model may be adapted to varying parameters, such as demographic history, through transfer learning. For our experiments, we trained a baseline CoalNN model for a European (CEU) demographic model, and later fine-tuned all layers of this model to work in other settings, such as a demographic model of constant size ($N_e = 10,000$) used for benchmarking and method comparison. Re-training CoalNN using an Nvidia A100 GPU card and 6 CPUs took approximately 30 and 5 hours on sequencing and array data, respectively, while training new models took $\sim$40 and $\sim$20 hours, respectively. Similarly, when analyzing the 1kGP data set, we loaded the weights of the baseline model and retrained CoalNN using population specific demographic models and genetic maps.

Forming piecewise constant TMRCA estimates

Ancestors shared along the genome by pairs of haploid individuals change as a result of recombination events, so that pairwise TMRCAs are expected to take the form of a piecewise constant function. However, raw CoalNN predictions tend to take different values at each site. CoalNN therefore implements a post-processing step that leverages the recombinaiton breakpoints predicted by the network to refine TMRCA predictions and allow producing piecewise constant estimates. More in detail, we average the TMRCA values inferred between genomic sites at which the estimated probability of observing a recombination event exceeds a specified threshold, as illustrated in Supplementary Fig. 1a. We observed this approach to lead to accuracy improvements for both sequencing and array data, with optimal probability thresholds around 0.7 and 0.55 respectively, as shown in Supplementary Fig. 1b-e.

Dating of genomic variants

We used the pairwise coalescence times inferred by CoalNN to predict the time of origin (in generations before present) of genomic variants. For a given variant, the maximum TMRCA across all concordant pairs (those which are homozygous for the derived allele) provides a lower bound on the age of the variant, while the minimum TMRCA across all discordant pairs (those which are heterozygous) provides an upper bound (Fig. 1b). However, due to noise in the TMRCA estimates, in practice the lower bound may be larger than the upper bound. To address this potential issue, we use the heuristic approach described by Albers and McVean [3] and illustrated in Fig. 1c to filter out outlier TMRCAs, removing all concordant (resp. discordant) pairs above (resp. below) a TMRCA threshold. This threshold is chosen by minimizing the total number of excluded values. The fraction of TMRCA estimates that are filtered out is provided in output and may be used as a proxy for the quality of the inferred variant age.

We compared allele age estimates for CoalNN, Relate (v1.1.6) [18], and tsinfer+tsdate (tsinfer v0.2.1 [19] and tsdate v0.1.3 [21] run successively). Both Relate and tsinfer+tsdate output allele age estimates as a range reflecting the lower and upper ends of the genealogical branch the mutation is estimated to be on. To obtain a point estimate for the age of an allele, we used the average of lower and upper ends of this range. CoalNN and Relate were provided with the simulated demographic model, while tsdate assumes a constant population size of $10,000$ diploid individuals. All methods produced highly biased age estimates for singleton variants, for which no concordant pairs are available. We therefore excluded singletons from further analyses. We also considered GEVA (v1beta) [3] in our benchmarks. However, GEVA produced estimates for $\sim$50% of non-singletons variants; for these variants we observed a lower accuracy compared to the other methods we considered, so we restricted our benchmarks to CoalNN, Relate, and tsinfer+tsdate.

Interpretability analyses

CoalNN's input contains 6 basic genomic features, which are propagated through multiple ConvBlocks to compute predictions. In order to explore how the network relies on the input features to calculate TMRCAs, we obtained saliency maps [60, 61] on sequencing data for simulated European samples. For a given pair of haplotypes and a given genomic region, we used the following procedure:

1. Perform a forward pass on the input of shape $L_1 \times 1 \times 6$, and obtain a TMRCA prediction per site of shape $L \times 1 \times 1$. Multiply the predicted TMRCAs by 10^6 to rescale and avoid vanishing gradients.

2. Calculate the absolute value of the gradient (gradient norm) of the rescaled TMRCA on a target site with respect to the normalized input (i.e. the input after going through the first batch

normalisation layer), resulting in $6 \times L_1$ gradient values. Note that the gradient is computed with respect to the normalized version of the input and not the raw input, so that all features are on the same standardized scale.

3. Multiply the gradient norm by the normalized input norm element-wise (known as the input $\times$ gradient method, [62]) and visualise results on an $6 \times L_1$ grid. Note that as we do not consider gradient directions, we only visualise absolute values in saliency maps.

In addition to using this procedure for producing and visualizing saliency maps, we further used the gradients to explore the relevance of the various input features in specific informative settings. Feature combinations for which we computed gradients (reporting average and SE across simulations) included heterozygous sites (XOR=1) with high MAF ($> 30\%$); rare (MAF $< 5\%$) homozygous (AND=1) variants; and regions of high ($\geq$ 90-th percentile) recombination rates. We often focused on sites where pairs of individuals have recent or deep coalescence times, which we defined as TMRCA < 200 and TMRCA $> 100,000$ generations, respectively.

We also performed perturbation analyses, in which we applied small changes to specific inputs (while keeping all other inputs fixed) and measured variation in the network's prediction. More in detail, given a pair of haplotypes, we either subtracted 0.05 from MAFs at all homozygous sites (AND=1) or added 0.05 to MAFs at all heterozygous sites (XOR=1), and calculated the difference in TMRCA predictions before and after the perturbation. If the MAF value at a site was smaller (resp. higher) than 0.05 (resp. 0.45), we set the MAF to singleton frequency (resp. 0.5).

We finally tested whether the network learns to model standard features used in coalescent HMMs, such as the sharing of alleles between the two input haplotypes. To this end, we loaded the weights of the CoalNN model trained on imputed data, which takes raw genotypes as inputs rather than the XOR / AND logic gates, and applied it to sequencing data. To test whether the model automatically learns these features, we measured Pearson correlation between each of the 8 channels in the output of the first ConvBlock and the XOR / AND logic gates. Note that the output of the first ConvBlock has shape $\mathrm{L}_2 \times 1 \times 8$, while each logic gate has shape $\mathrm{L}_1 \times 1 \times 1$ (see Fig. 1a). In order to compute correlations, we only considered the L_2 central sites in the logic gates (discarding $\frac{\mathrm{L}_1 - \mathrm{L}_2}{2}$ sites on both ends), and we assumed a one-to-one mapping between the output of the ConvBlock and the logic gates. For all of these analyses we averaged results across all genomic sites and pairs in a simulation.

Computation of standard errors

Unless otherwise indicated, standard errors (SE) were computed over 20 independent simulations (using random seeds not seen in training) for analyses performed on simulated data. For real data analyses, we applied bootstrapping with 30 genomic regions as resampling units. These 30

genomic regions were obtained by dividing the autosomal chromosomes in contiguous regions with equal number of variants.

1,000 Genomes Project data set preprocessing

We analysed the GRCh38-build 1kGP Phase 3 data set [5], which was lifted using CrossMap (version 0.5.4) [63] from GRCh37 (1kGP release 20130502). The data set comprises 2,504 unrelated samples from 26 populations. To estimate allele ages, CoalNN uses knowledge of whether an allele is ancestral or derived. In oder to recode the genotype data from reference/alternate to ancestral/derived, we used ancestral allele annotations provided with the data set. These were inferred from Ensembl multiple alignments (human assembly GRCh37) using Ortheus in the EPO (Enredo-Pecan-Ortheus) pipeline [64], resulting in either a high-confidence call, a low-confidence call, or no call for each variant. We used these annotations to enforce ancestral/derived encoding for all variants for which a call was available. We retrained CoalNN on each of the 26 groups by transfer learning, using population-specific demographic models and GRCh38 recombination maps [56]. This resulted in 26 trained CoalNN models, one per 1kGP population.

Allele age estimation in the 1,000 Genomes Project data set

We inferred allele ages for all polymorphic variants identified in the 1kGP data set within each population separately. We aggregated age estimates for the following 5 super-groups (see Supplementary Table 1), using population assignments provided with the 1kGP data set: African (AFR), Admixed American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS). For each super-group and each variant, we computed a weighted average of the lower age estimates across all constituent populations. We used the number of concordant pairs retained after filtering as weights for each population's contribution to the average. Similarly, we computed a weighted average of the upper age estimates using the number of discordant pairs after filtering as weights. Age estimates were then obtained by taking the mean of these upper and lower weighted averages. We dated ∼42 million variants in AFR, ∼28 million in AMR, ∼24 million in EAS, ∼24 million in EUR, ∼27 million in SAS, and ∼80 million overall. When performing S-LDSC analyses (described below) we also combined age estimates from all 26 populations, using the same procedure. We have made all age estimates publicly available (see URLs). Unless otherwise specified, we only retained variants with high-confidence ancestral state annotations and excluded singleton variants in all downstream analyses.

VEP variant annotation

We extracted filtered pathogenicity annotations from the Ensembl Variant Effect Predictor (VEP) [65] generated by SIFT [66] (annotated as "deleterious" or "tolerated") and by PolyPhen-2 [67] (annotated as "possibly damaging", "probably damaging" or "benign") for all analysed

variants, and lifted their coordinates from GRCh37 to GRCh38. After excluding variants with low-confidence or missing ancestral state, we obtained 115,288 variants in AFR, 77,672 in AMR, 70,049 in EAS, 80,782 in EUR, 48,358 in SAS for Polyphen-2, and 138,573 variants in AFR, 93,034 in AMR, 84,523 in EAS, 97,753 in EUR, 57,605 in SAS for SIFT. We computed the cumulative distribution functions of each annotated group of variants in each super-group, stratified by derived allele frequency.

S-LDSC analyses

We built annotations to perform stratified LD-score regression analysis [51], using the procedure described by Gazal et al. [52] to analyse the effect of an annotation based on allele ages on complex trait heritability. More in detail, we constructed 26 different MAF-adjusted annotations using allele age predictions obtained by CoalNN for each 1kGP population. We partitioned variants using the same 10 MAF bins (all with MAF ≥ 0.05) used by Gazal et al. [52]; within each MAF bin, we quantile normalized the allele ages to a standard normal distribution. We used the same approach to construct 6 additional MAF-adjusted and quantile-normalized annotations: one for each population super-group and one encompassing all 26 populations. Overall, we produced 32 allele age annotations and we evaluated them by applying S-LDSC to summary association statistics from 63 independent diseases and complex traits, described in Supplementary Table 2. S-LDSC analyses were performed using the S-LDSC software [51] and the baselineLD v2.2 model.

Results

Simulation results

We assessed the accuracy of CoalNN in inferring locus-specific pairwise TMRCAs and allele ages through extensive simulations of sequencing, SNP array, and imputed data. We compared the pairwise TMRCA estimates inferred by CoalNN to both the MAP and the posterior mean estimates provided by ASMC (v.1.0 [15]). Because ASMC does not support a floating point input representation of genotype data, when analyzing imputed data with ASMC we binarized the input by rounding genotype dosages. In these experiments, CoalNN had a higher accuracy compared to ASMC in all data modalities, as shown in Table 1, Fig. 2a, and Supplementary Fig. 3a. We also tested CoalNN's accuracy in estimating allele ages using its TMRCA predictions (see Methods). We simulated data from a European demographic model and from a constant population size of $N_e = 10,000$ and compared to allele age predictions obtained using Relate [18] and tsinfer+tsdate [19, 21]. In these simulations, CoalNN achieved the highest allele age estimation accuracy across several metrics, shown in Fig. 2b, Supplementary Fig. 3b and Table 2.

Next, we validated the robustness of CoalNN to various types of model misspecification, including phasing errors, genotyping errors, and an inaccurate demographic model. We first measured the robustness of CoalNN and ASMC to the presence of genotyping and phasing errors, using simulated sequencing data under a European demographic model. In these simulations, CoalNN was more robust than ASMC across a range of simulated error rates, with the performance gap increasing for larger error rates (Supplementary Fig. 4). For a switch (resp. genotyping) error rate of 0.1%, CoalNN had a 17.41% (SE=0.93) lower mean absolute error (MAE, resp. 18.10%, SE=0.79) compared to ASMC's posterior mean estimates, and a 71.14% (SE=1.18) lower MAE error (resp. 71.26%, SE=1.01) compared to ASMC's MAP estimates. We tested CoalNN's robustness to demographic model misspecification by simulating data under a constant population size of $10,000$ diploid individuals, using model parameters trained on a European demographic model. For comparison, we also ran ASMC while incorrectly assuming a European demographic model. CoalNN generalized better than ASMC on sequencing data (14.27%, SE=0.52 and 11.12%, SE=0.68 performance improvement compared to ASMC mean posterior estimates on MAE and root mean squared error, RMSE, respectively). ASMC, however, was more robust when using array data (-11.34%, SE=0.38 and -19.34%, SE=0.5).

Finally, we evaluated the computational efficiency of CoalNN. When inferring TMRCAs for random pairs of individuals from a data set, CoalNN run on a NVIDIA A100 GPU was $\sim 10.5\times$ faster than ASMC run on an Intel Skylake 2.6 GHz CPU, as shown in Fig. 3. ASMC may be run in batch-optimized mode, where memory locality and single instruction/multiple data (SIMD) processing are leveraged for faster analysis of several pairs of contiguous individuals from the input genotype matrix. Under the same hardware configuration, CoalNN remained $\sim 2.5\times$ faster than batch-optimized ASMC. When run on a CPU architecture, CoalNN was $\sim 15.6\times$ slower compared to its GPU performance, but remained comparable to ASMC for TMRCA inference in random individual pairs, being only $\sim 1.5\times$ slower. Training CoalNN on a CEU demographic model required ~ 52 hours for sequencing data, ~ 23 hours for imputed data, and ~ 9 hours for SNP array data, using an Nvidia A100 GPU card and 6 CPUs. Transfer learning, where a previously trained CoalNN model is fine-tuned for a new setting (e.g. a different demographic parameters, see Methods), led to considerable computational improvements, being $\sim 1.5\times$ faster for sequencing data and $\sim 5\times$ faster for array data compared to the training of randomly initialized models.

Interpretation of CoalNN's predictions

While being accurate, CoalNN's trained model may be hard to interpret, as it automatically learns a map between input data and inferred TMRCA values rather than relying on an explicitly designed model. We therefore investigated whether CoalNN implicitly infers and leverages genomic features that have been described and utilized in other probabilistic approaches, such as coalescent HMMs.

To this end, we first examined saliency maps, which use the gradient of the output with respect to the input as a way of quantifying the importance of different input regions in forming the model's prediction (see Methods and Supplementary Fig. 5 for two representative examples). Examining regions where pairs of individuals shared recent common ancestors (TMRCA$<$ 200), we observed that the XOR feature, which reflects whether individuals are identical-by-state and is a standard input of coalescent HMMs [4], remained informative for long stretches of the input region, indicating the presence of long shared haplotypes. Gradients for the AND feature, which allows quantifying whether the individuals are both carrier of a derived allele and was adopted in more recent coalescent HMMs [14], were more localized around the focal site. Larger gradients for the genetic distance feature, which is informative for the detection of recombination breakpoints, tended to be localized in regions of high recombination rates. For instance, when examining regions where TMRCA values are large ($> 100,000$ generations), we observed an average gradient of 46.2 (SE=0.03) for regions with recombination rate in the 90-th percentile or above, compared to 16.26 (SE=0.02) for lower recombination rates.

Next, we used input perturbations to test whether CoalNN learns to utilize more complex combinations of these basic features, such as MAF and allele sharing, that require complex probabilistic modeling in coalescent HMMs [14]. Assuming no recurrent mutation, individuals sharing rare alleles are expected to co-inherit these alleles from a recent common ancestor, since rare variant are on average younger than higher frequency variants [68, 69]. Similarly, individuals that are heterozygous for high frequency variants are more likely to share a distant common ancestor. To verify that CoalNN learns to rely on combinations of MAF and allele sharing to form its predictions, we perturbed MAF values for individuals based on their allele sharing status. We observed that increasing the input MAF value by 5% at sites for which individuals are heterozygous for high frequency variants (XOR=1, MAF$>$ 45%) resulted in an average increase in predicted TMRCA of 181.5 (SE=3.7) generations, while decreasing the MAF at homozygous sites (AND=1, MAF$<$ 5%) resulted in an average decrease of predicted TMRCA of 27.6 (SE=1.7) generations (see Methods, Supplementary Fig. 6).

Finally, we sought to verify that CoalNN would allow independently recovering basic engineered features, such as the XOR and AND logic gates, using raw haplotype data. To this end, rather than providing XOR and AND features in input to increase training efficiency, we provided CoalNN with raw genotype data, and observed that the output channels of the first hidden layer were significantly correlated with the XOR and AND logic gates ($r = -0.172$, SE=0.001 and $r = -0.272$, SE=0.001, respectively, see Supplementary Table 3).

Allele age prediction in 1,000 Genomes Project populations

We applied CoalNN to 2,504 samples from the 1kGP Phase 3 data set and we inferred allele ages for $\sim$80 million variants (see Methods). We analysed each of the 26 populations separately and

aggregated estimates for each population group (labeled as AFR, AMR, EAS, EUR and SAS, see Methods, Supplementary Table 1). Allele ages inferred using CoalNN were highly correlated with previously published age estimates. For variants with high-confidence ancestral states, the average correlation across all 26 populations was $r = 0.3$ (SE=0.07) for ages predicted by GEVA [3], and $r = 0.67$ (SE=0.11) for ages inferred by Relate [18].

We analyzed the genome-wide distribution of frequency-stratified allele ages (see Fig. 4a) and observed significant differences between populations, reflecting population-specific histories of migration and population size variation [5, 3]. For instance, for variants with derived allele frequency between 1% and 2.5% we observed a median age of $4,904$ generations (SE=53) in AFR, $4,852$ generations (SE=64) in AMR, $2,634$ generations (SE=27) in EAS, $1,593$ generations (SE=20) in EUR, and $1,517$ generations (SE=29) in SAS (see Supplementary Table 4). We replicated this experiment using CoalNN weights trained by assuming a constant population size of $N_e = 20,000$ for three representative populations (CEU, CHS, YRI). Using this approach, which is likely to result in age estimates that are slightly less accurate (see Supplementary Fig. 7), we observed similar patterns of allele age variation (see Supplementary Fig. 8 and Supplementary Table 5), suggesting that the observed variation is not only due to the use of CoalNN models trained on population-speific demographic priors.

Next, we analyzed variation in the age distribution of frequency-stratified alleles predicted to have different pathogenic effects predicted using SIFT [66] and PolyPhen-2 [67], reflecting the action of negative selection [3] (see Methods). To control for the relationship between allele frequency and allele age, we compared sets of neutral and potential pathogenic variants observed to have the same derived allele frequency (DAF). Allele age distributions of rare variants (1% ¡ DAF¡ 2.5%) annotated by PolyPhen-2 and SIFT are shown in Fig. 4b. Results for variants at other frequencies are illustrated in Supplementary Fig. 9-10 and summarised in Supplementary Tables 6-7. We observed deleterious alleles to be younger than neutral alleles of the same frequency in every population group, consistent with the action of negative selection [70, 3] (e.g median age of 900 generations, SE=19, for AFR for DAF$<$ 1% variants identified as deleterious by SIFT, compared to $1,061$ generations (SE=23) for variants annotated as tolerated).

Stratified LD-Score regression analysis

Stratified LD-score regression (S-LDSC [51]) has been used to test whether genome-wide annotations built using evolutionary features are predictive of heritability enrichments across complex traits and diseases [52]. In particular, an annotation based on allele ages inferred using the ARGweaver algorithm [17] has been observed to have significant effects on complex trait heritability [52], likely reflecting the effects of natural selection on allele age variation. These observed enrichments persisted when conditioning on the per-allele effects of several other evolutionary annotations, including nucleotide diversity, a background selection statistic [71] (B-statistic),

average pairwise TMRCA estimated using ASMC [15] (ASMC_{avg}), recombination rate, the level of LD in African populations [52] (LLD-AFR), and CpG content [72].

In order to test the informativeness of an annotation built using allele age estimates obtained through CoalNN, we constructed 26 MAF-adjusted annotations, one for each 1kGP population (see Methods). We found these annotations to be highly correlated with other MAF-adjusted evolutionary annotations present in the Baseline-LD model [52] (see Fig. 5a), including $r = 0.56$, $SE = 0.01$ (average across populations) for the ARGweaver allele age annotation; $r = 0.3$, $SE = 0.01$ for LLD-AFR, $r = 0.28$, $SE = 0.004$for ASMC_{avg}; and $r = 0.16$, $SE = 0.004$ for nucleotide diversity. We tested the CoalNN allele age annotations by applying S-LDSC to summary statistics from 63 independent complex traits and diseases (see Supplementary Table 2). We quantified the informativeness of an annotation by performing a meta-analysis of the heritability effect τ^* across these traits. The heritability effect τ^* is defined as the proportionate change in per-SNP heritability associated with a 1 s.d. increase in the value of the annotation, conditional on other annotations included in the baseline LD model [52].

We first investigated the informativeness of the CoalNN annotations when conditioned on the full set of 97 annotations within the baseline-LD model, including the ARGweaver allele age annotation. In these conditional analyses, we observed the CoalNN annotations for several populations to have a significant effect (Supplementary Fig. 11b). However, these effects were heterogeneous and occasionally in opposite directions (e.g. $\tau^*_{\text{ACB}} = -0,12$, SE=0.02, and $\tau^*_{\text{CHS}} = 0.11$, SE=0.02). This heterogeneity may be linked to the high correlation between the ARGweaver and CoalNN MAF-adjusted alelle age annotations (see Supplementary Fig. 11a). Repeating these analyses after removing the ARGweaver annotation from the baseline model resulted in consistent effect size directionality across populations (see Supplementary Fig. 11c). We thus opted to evaluate the informativeness of the CoalNN annotations using analyses where the ARGweaver annotation is first removed from the baseline model, conditioning on the remaining set of 96 functional and evolutionary annotations.

Next, we aggregated allele age estimates across all 26 populations to obtain a single CoalNN annotation (see Methods). We assessed its heritability effect when no other evolutionary annotation is included in the model, by only conditioning on the 90 remaining annotations within the Baseline-LD model. In this scenario, the CoalNN annotation obtained a meta-analysed effect size τ^* of -1.01 (SE=0.02), while performing the same analysis with the ARGweaver annotation resulted in a τ^* of -0.88 (SE=0.02). We repeated this analysis, this time conditioning on other evolutionary annotations (LLD-AFR, recombination rate, nucleotide diversity, B-statistic, CpG content, and ASMC_{avg}), observing a τ^* of -0.22 (SE=0.02) for CoalNN and of -0.15 (SE=0.01) for ARGweaver (Fig. 5b, see Supplementary Fig. 12 for individual trait results). We observed larger effect sizes for evolutionary annotations when including ages inferred by CoalNN as a new annotation compared to when including ages predicted by ARGweaver, suggesting a larger

amount of overlapping information between ARGWeaver and the other evolutionary annotations. Finally, we repeated these conditional analyses using population specific annotations (Supplementary Fig. 13). We observed heterogeneous effect sizes across groups (e.g. $\tau^*_{AFR} = -0.219$, SE=0.019; $\tau^*_{EUR} = -0.148$, SE=0.018, see Supplementary Table 8), which may be a consequence of differences in the demographic history of these groups, or reflect a reduced overlap with information already captured using other annotations built using data sets of predominantly European ancestry.

Overall, these results suggest that the CoalNN annotation captures orthogonal heritability effects that are not captured by the ARGWeaver annotation or other evolutionary annotations contained in the baseline-LD model. The negative sign of the heritability effect size indicates that, after conditioning on allele frequency, alleles with a younger estimated age correspond to larger phenotypic effects, consistent with the action of negative selection [70, 52].

Discussion

We developed CoalNN, a likelihood-free method that uses a deep convolutional neural network to predict pairwise coalescence times and recombination breakpoints from sequencing, array, or imputed genotype data. Using extensive simulations, we found that CoalNN improves upon several current approaches for the inference of TMRCAs and the dating of genomic variants. We applied CoalNN to the 1kGP data set and estimated the age of ∼80 million variants across 26 human populations. We observed differences in allele age distributions between populations, reflecting diverse demographic histories, as well as between predicted pathogenic and neutral alleles, reflecting the action of natural selection. We built genome-wide alelle age annotations to capture the effects of selection and used stratified LD-score regression to analyze the genetic architecture of 63 diseases and complex traits. Using this approach, we showed that annotations built using CoalNN capture significant heritability effects. These effects were larger than those of a previous annotation based on allele ages and remained significant after conditioning on several other evolutionary annotations.

Overall, these results demonstrate that deep learning algorithms trained using simulation provide an effective route to inferring genealogical relationships for a set of sequenced or genotyped samples. This strategy circumvents difficulties linked to intractable likelihood calculations and overcomes limitations of current HMM-based methodology, such as the need to rely on discretized time intervals. This approach may also be used to incorporate additional evolutionary features that are not as easily approximated as a Markovian process along the genome, such as the occurrence of non-crossover gene conversion events [73]. Although our analyses have focused on the inference of coalescence times and allele ages, other applications of this framework are possible. In particular, although CoalNN infers genealogical relationships for pairs of individuals, it may be leveraged as a building block to infer the full genealogy for larger sets of samples. Pairwise

coalescence times inferred using CoalNN may be used to sequentially thread new samples in an existing genealogy, as recently done in the ARG-Needle algorithm [20] using ASMC [15]. Similarly, improved allele age estimates can be utilized to more accurately infer the age of ancestors used to assemble genealogies in the tsinfer and tsdate algorithms [19, 21].

We note a few areas of future improvement, as well as limitations of this work. Although a trained model may also be applied to data using standard CPU hardware, CoalNN relies on GPU hardware for optimal computational performance. Moreover, although we observed that CoalNN trained on a constant demographic prior performs well under different scenarios, properly generalizing to varying evolutionary settings, including demographic models, requires additional training through transfer learning. Building models that can be more easily adapted to varying evolutionary parameters is a desirable area of future improvement. A further limitation is linked to our choice of architecture, which poses limits to CoalNN's receptive field and may affect its performance in inferring recent TMRCAs that involve long shared haplotypes. This was highlighted using saliency maps, where we observed non-zero gradients on the edges of the receptive field for sites with recent coalescence. Although we experimented with neural network architectures without observing significant performance gains (e.g attention-based models [36]), several improvements are likely possible. Lastly, our approach for dating variants requires pairwise haplotype comparisons, resulting in computation time quadratic in sample size. In addition to using CoalNN as a building block for other algorithms that efficiently infer full genealogies, as previously discussed, the development of extensions that jointly infer TMRCAs across several samples is an interesting direction of future work. Despite these limitations and areas of future development, we believe that the CoalNN model provides a valuable tool for the inference of coalescence times and allele ages, and demonstrates the potential of using deep neural networks in the analysis of genealogical relationships.

Data and code availability

Software, annotations, and data sets used in this study include: demographic models and genetic maps https://github.com/popgenmethods/pyrho; ASMC software https://github.com/PalamaraLab/ASMC; Relate software https://myersgroup.github.io/relate/; tsinfer software https://tsinfer.readthedocs.io/; tsdate software https://tsdate.readthedocs.io; GEVA software https://github.com/pkalbers/geva; Relate allele ages https://zenodo.org/record/3234689.; GEVA allele ages https://human.genome.dating; AMSC_{avg} annotation https://palamaralab.github.io/software/asmc/data/; 1kGP Phase 3 data set (release 2013050) http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/phase3_liftover_nygc_dir/; pathogenicity annotations http://ftp.1000genomes.ebi.ac.uk//vol1/ftp/release/20130502/supporting/functional_annotation/filtered/; LDSC baseline model https://alkesgroup.broadinstitute.org/LDSCORE/; LDSC

software https://github.com/bulik/ldsc.

Data analyses are based on open-source libraries and software programs that are available online: Scipy [74], Matplotlib [75], NumPy [76, 77], Pandas [78, 79], PyTorch [80].

The CoalNN software and pre-trained neural network weights will be made available on GitHub upon publication.

Acknowledgments

We thank Steven Gazal for helpful discussions and for curating the list of the 63 independent disease and complex trait summary statistics used in S-LDSC analyses, Alkes Price and Hilary Finucane for helpful discussions on S-LDSC analyses. This work was supported by MRC grant MR/S502509/1 and Balliol Jowett Scholarship (to J.N.S.); EPSRC grant EP/S023151/1 (to Z.T.); EPSRC and MRC grant EP/L016044/1 (to M.S.); Wellcome Trust ISSF grant 204826/Z/16/Z and ERC Starting Grant 850869 (to P.F.P.). We thank the Biomedical Research Computing team at the University of Oxford for support with the ResComp compute clusters. The research was partly supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z with additional support from the NIHR Oxford BR (the views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health).

References

[1] John Frank Charles Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

[2] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[3] Patrick K Albers and Gil McVean. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS biology*, 18(1):e3000586, 2020.

[4] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493, 2011.

[5] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[6] Juba Nait Saada, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev, and Pier Francesco Palamara. Identity-by-descent detection across 487,409 British samples reveals fine-scale evolutionary history and trait associations. *Nature Communications*, 11(1):6130, 2020. doi: 10.1038/s41467-020-19588-x.

[7] Anders Albrechtsen, Ida Moltke, and Rasmus Nielsen. Natural selection and the distribution of identity-by-descent in the human genome. *Genetics*, 186(1):295–308, 2010. doi: https://doi.org/10.1534/genetics.110.113977.

[8] Alexander Gusev, Pier Francesco Palamara, Gregory Aponte, Zhong Zhuang, Ariel Darvasi, Peter Gregersen, and Itsik Pe'er. The architecture of long-range haplotypes shared within and across populations. *Molecular biology and evolution*, 29(2):473–486, 2011. doi: https://doi.org/10.1093/molbev/msr133.

[9] Haley Hunter-Zinck and Andrew G Clark. Aberrant time to most recent common ancestor as a signature of natural selection. *Molecular biology and evolution*, 32(10):2784–2797, 2015.

[10] Richard R Hudson and Norman L Kaplan. The coalescent process in models with selection and recombination. *Genetics*, 120(3):831–840, 1988.

[11] Carsten Wiuf and Jotun Hein. Recombination as a point process along sequences. *Theoretical population biology*, 55(3):248–259, 1999.

[12] Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

[13] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, 2014.

[14] Jonathan Terhorst, John A Kamm, and Yun S Song. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, 49(2): 303–309, 2017.

[15] Pier Francesco Palamara, Jonathan Terhorst, Yun S. Song, and Alkes L. Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50, 09 2018. doi: https://doi.org/10.1038/s41588-018-0177-x.

[16] Jeffrey P Spence, Matthias Steinrücken, Jonathan Terhorst, and Yun S Song. Inference of population history using coalescent hmms: review and outlook. *Current opinion in genetics & development*, 53:70–76, 2018.

[17] Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014.

[18] Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, 51:1321–1329, 2019. doi: https://doi.org/10.1038/s41588-019-0484-x.

[19] Jerome Kelleher, Yan Wong, Anthony Wohns, Chaimaa Fadil, Patrick Albers, and Gil McVean. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51: 1330–1338, 09 2019. doi: 10.1038/s41588-019-0483-y.

[20] Brian C Zhang, Arjun Biddanda, and Pier Francesco Palamara. Biobank-scale inference of ancestral recombination graphs enables genealogy-based mixed model association of complex traits. *bioRxiv*, 2021.

[21] Anthony Wilder Wohns, Yan Wong, Ben Jeffery, Ali Akbari, Swapan Mallick, Ron Pinhasi, Nick Patterson, David Reich, Jerome Kelleher, and Gil McVean. A unified genealogy of modern and ancient genomes. *Science*, 375(6583):eabi8264, 2022.

[22] Na Li and Matthew Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[23] Gilean AT McVean and Niall J Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1387–1393, 2005.

[24] Asger Hobolth and Jens Ledet Jensen. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. *Theoretical population biology*, 98:48–58, 2014.

[25] Simon Tavaré, David J Balding, Robert C Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[26] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[27] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[28] Mark A Beaumont. Approximate bayesian computation in evolution and ecology. *Annual review of ecology, evolution, and systematics*, pages 379–406, 2010.

[29] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[30] Jeffrey Chan, Valerio Perrone, Jeffrey P. Spence, Paul A. Jenkins, Sara Mathieson, and Yun S. Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. doi: 10.1101/267211.

[31] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[34] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[37] Jian Zhou and O. Troyanskaya. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12:931–934, 2015.

[38] David R Kelley, Jasper Snoek, and John L Rinn. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research*, 26(7): 990–999, 2016.

[39] David Kelley, Yakir Reshef, Maxwell Bileschi, David Belanger, Cory McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28:gr.227819.117, 03 2018. doi: 10.1101/gr.227819.117.

[40] Jian Zhou, Chandra Theesfeld, Kevin Yao, Kathleen Chen, Aaron Wong, and Olga Troyanskaya. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nature Genetics*, 50, 08 2018. doi: 10.1038/s41588-018-0160-6.

[41] Haotian Teng, Minh Duc, Michael Hall, Tania Duarte, Sheng Wang, and Lachlan Coin. Chiron: Translating nanopore raw signal directly into nucleotide sequence using deep learning. *GigaScience*, 7, 12 2017. doi: 10.1093/gigascience/giy037.

[42] Babak Alipanahi, Andrew Delong, Matthew Weirauch, and Brendan Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature biotechnology*, 33, 07 2015. doi: 10.1038/nbt.3300.

[43] Sara Sheehan and Yun S Song. Deep learning for population genetic inference. *PLoS computational biology*, 12(3):e1004845, 2016.

[44] Jonas Meisner and Anders Albrechtsen. Haplotype and population structure inference using neural networks in whole-genome sequencing data. *Genome Research*, 32(8):1542–1552, 2022.

[45] Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. *Molecular Ecology Resources*, 21(8):2645–2660, 2021.

[46] Alexander T Xue, Daniel R Schrider, and Andrew D Kern. Discovery of ongoing selective sweeps within anopheles mosquito populations using deep learning. *Molecular biology and evolution*, 38(3):1168–1183, 2021.

[47] Hussein A Hejase, Ziyi Mo, Leonardo Campagna, and Adam Siepel. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *Molecular biology and evolution*, 39(1):msab332, 2022.

[48] Nathan Killoran, Leo Lee, Andrew Delong, David Duvenaud, and Brendan Frey. Generating and designing dna with deep generative models. *Advances in Neural Information Processing Systems, Computational Biology Workshop (NeurIPS)*, 2017.

[49] Sam Sinai, Eric Kelsic, George Church, and Martin Nowak. Variational auto-encoding of protein sequences. *Advances in Neural Information Processing Systems, Computational Biology Workshop (NeurIPS)*, 2017.

[50] Daniel Mas Montserrat, Carlos Bustamante, and Alexander Ioannidis. Class-conditional vae-gan for local-ancestry simulation. *Advances in Neural Information Processing Systems, Computational Biology Workshop (NeurIPS)*, 2019.

[51] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, Gosia Trynka, Yakir Reshef, Po-Ru Loh, Verneri Anttila, Han Xu, Chongzhi Zang, Kyle Farh, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics*, 47(11):1228–1235, 2015.

[52] Steven Gazal, Hilary K Finucane, Nicholas A Furlotte, Po-Ru Loh, Pier Francesco Palamara, Xuanyao Liu, Armin Schoech, Brendan Bulik-Sullivan, Benjamin M Neale, Alexander Gusev, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics*, 49(10):1421–1427, 2017.

[53] Rich Caruana. Multitask learning. *Machine Learning*, 28, 07 1997. doi: 10.1023/A:1007379606734.

[54] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491, 2018.

[55] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[56] Jeffrey P Spence and Yun S Song. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances*, 5(10):eaaw9206, 2019.

[57] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018. doi: https://doi.org/10.1038/s41586-018-0579-z.

[58] Brian L Browning, Ying Zhou, and Sharon R Browning. A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103(3): 338–348, 2018. doi: https://doi.org/10.1016/j.ajhg.2018.07.015.

[59] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[60] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations*, 2014.

[61] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

[62] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

[63] Hao Zhao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Liguo Wang. Crossmap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, 30(7):1006–1007, 2014.

[64] Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research*, 18(11):1814–1828, 2008.

[65] William McLaren, Laurent Gil, Sarah E Hunt, Harpreet Singh Riat, Graham RS Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome biology*, 17(1):1–14, 2016.

[66] Ngak-Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C Ng. Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research*, 40(W1):W452–W457, 2012.

[67] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nature methods*, 7(4):248–249, 2010.

[68] Motoo Kimura and Tomoko Ohta. The age of a neutral mutant persisting in a finite population. *Genetics*, 75(1):199–212, 1973.

[69] Robert C Griffiths and Simon Tavaré. The ages of mutations in gene trees. *The Annals of Applied Probability*, 9(3):567–590, 1999.

[70] Adam Kiezun, Sara L Pulit, Laurent C Francioli, Freerk van Dijk, Morris Swertz, Dorret I Boomsma, Cornelia M van Duijn, P Eline Slagboom, GJB van Ommen, Cisca Wijmenga, et al. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS genetics*, 9(2):e1003301, 2013.

[71] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics*, 5(5):e1000471, 2009.

[72] Le Zhang, Zichun Dai, Jun Yu, and Ming Xiao. Cpg-island-based annotation and analysis of human housekeeping genes. *Briefings in bioinformatics*, 22(1):515–525, 2021.

[73] Carsten Wiuf and Jotun Hein. The coalescent with gene conversion. *Genetics*, 155(1): 451–462, 2000.

[74] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17 (3):261–272, 2020.

[75] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007.

[76] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[77] Stefan Van Der Walt, S Chris Colbert, and Gael Varoquaux. The numpy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2):22–30, 2011.

[78] The pandas development team. *pandas-dev/pandas: Pandas*, February 2020. URL https://doi.org/10.5281/zenodo.3509134.

[79] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: https://doi.org/10.25080/Majora-92bf1922-00a.

[80] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

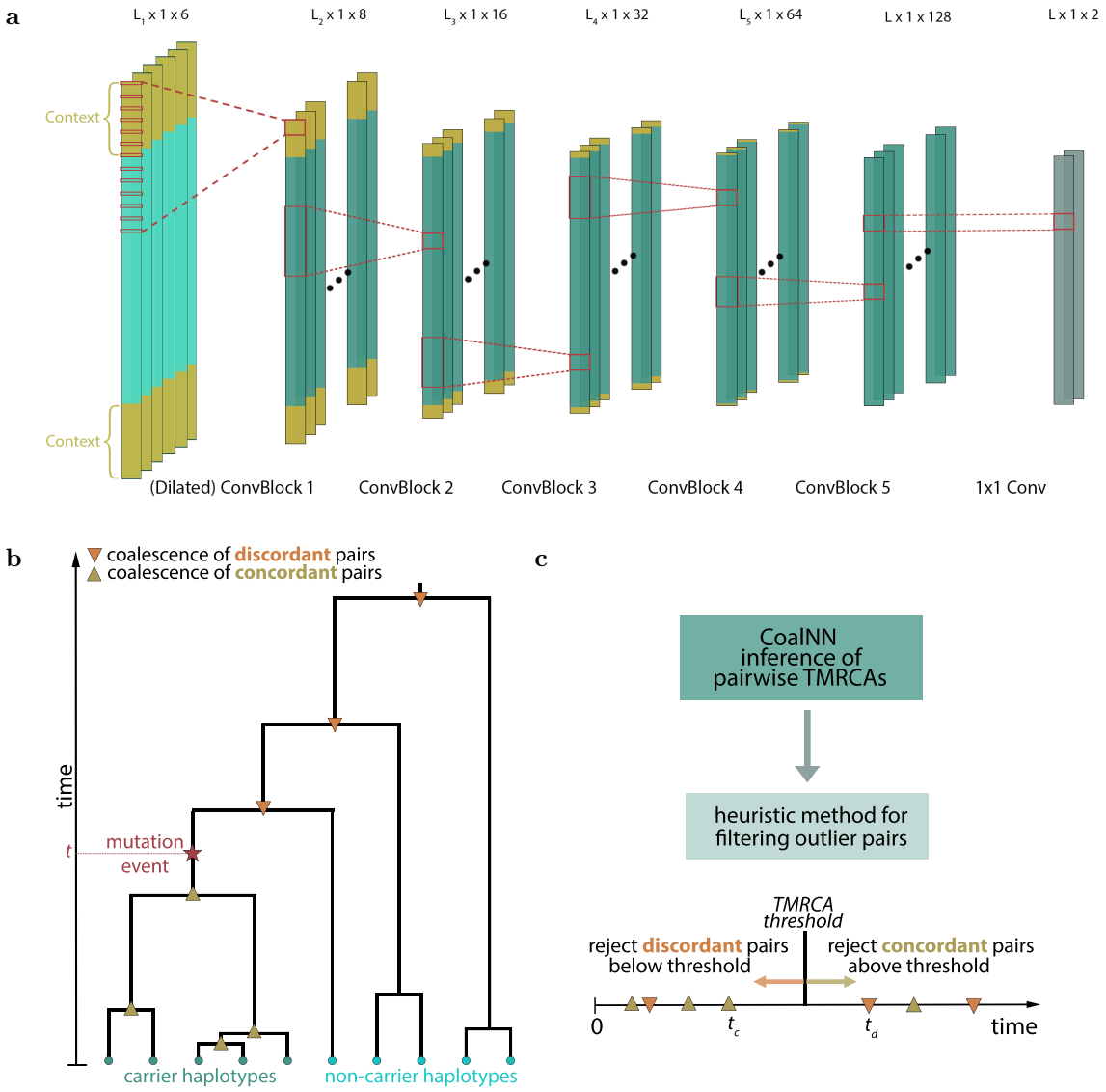


Fig. 1. Overview of the CoalNN model. a. CoalNN comprises a batch normalisation layer followed by five convolution blocks (convolution layer + batch normalisation + ReLU) and a final 1×1 convolution layer. The input sequence includes additional contextual data (in yellow) in lieu of using padding. **b.** At a given genomic site, individuals are connected through the sharing of underlying genealogical relationships. We aim to infer the time t at which a mutation arose (in red) and resulted in carrier haplotypes (in green) and non-carrier haplotypes (in cyan). **c.** When dating variants, CoalNN first infers TMRCAs across all concordant (two carriers) and discordant (one carrier and one non-carrier) pairs of haplotypes and then rejects outlier pairs using a heuristic method [3]. The TMRCA rejection threshold is computed by minimising the total number of rejected pairs. The predicted age estimate for the variant is obtained by averaging the maximum coalescence time across concordant pairs t_c and the minimum coalescence time across discordant pairs t_d after filtering.

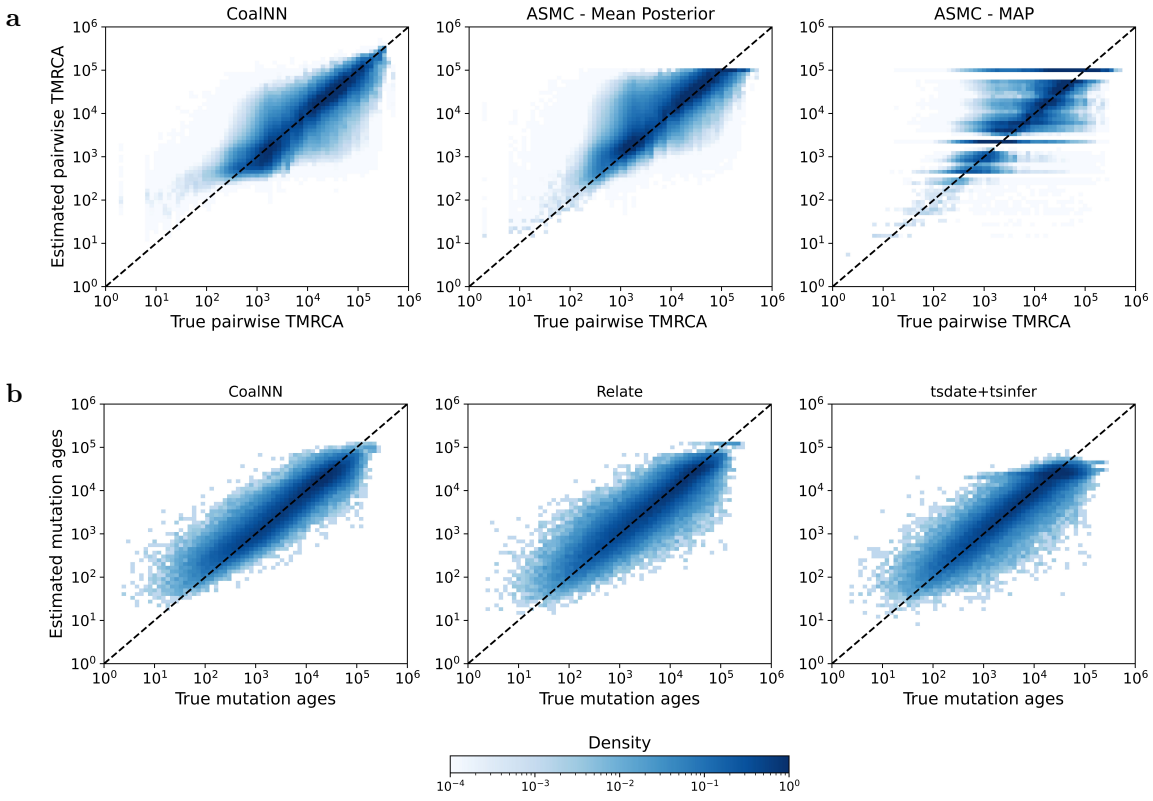


Fig. 2. Pairwise TMRCA and allele age prediction on sequencing data. a. True pairwise TMRCAs (x axis) versus those estimated by CoalNN and ASMC (y axis) under a European demographic model for one simulation. **b.** True non-singleton variant ages (x axis) versus those estimated by CoalNN, Relate and tsdate+tsinfer (y axis) under a constant diploid population size $N_e = 10,000$.

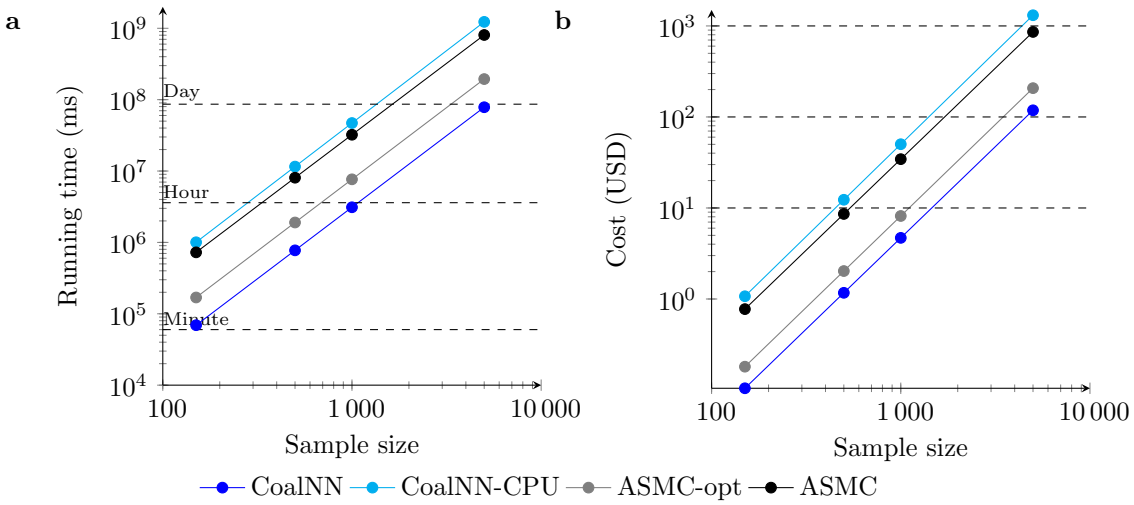


Fig. 3. Running time and cost evaluation. Running time (in milliseconds) of CoalNN (on a single A100 GPU card and a single CPU in blue, and on a single CPU only in cyan) and ASMC (on a single CPU, optimized version in grey and non-optimized version in black) in **a**, and their cost (in USD) estimated by Amazon Web Services (AWS) using the ml.g4dn.16xlarge instance (5.44 USD per hour instance cost) for CoalNN (GPU) and the ml.m4.16xlarge instance (3.84 USD per hour instance cost) for CoalNN (CPU) and ASMC (both optimized and non-optimized) in **b**. Simulations are based on array data using the first 30 Mbp of chromosome 2 across 6,749 SNPs. The batch size used for both methods is 64.

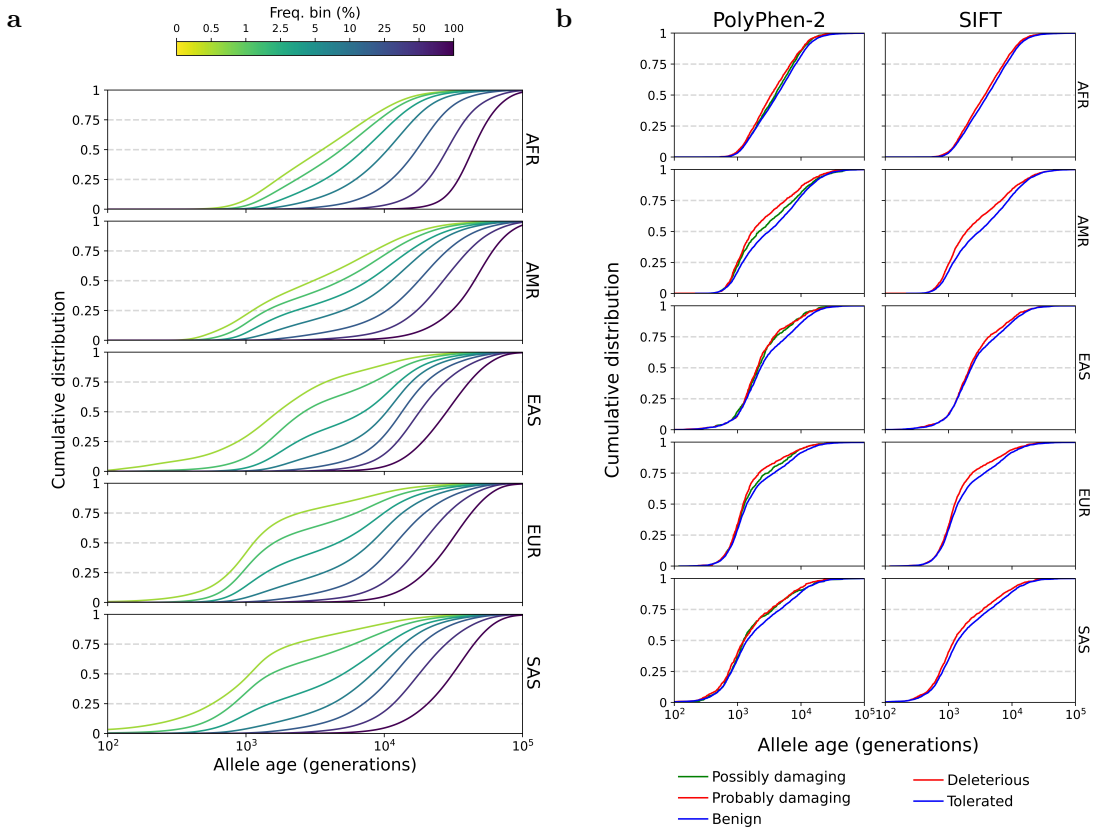


Fig. 4. Age distribution of dated variants among different population groups. a. Cumulative age distribution function of all dated variants across the human genome per population group. For each line, only non-singleton polymorphic variants present in that population within a given derived allele frequency bin were considered. **b.** Differences in allele age distribution between pathogenic mutations (annotated as such by PolyPhen-2 and by SIFT) and neutral variants for a derived allele frequency between 1% and 2.5% within each population group.

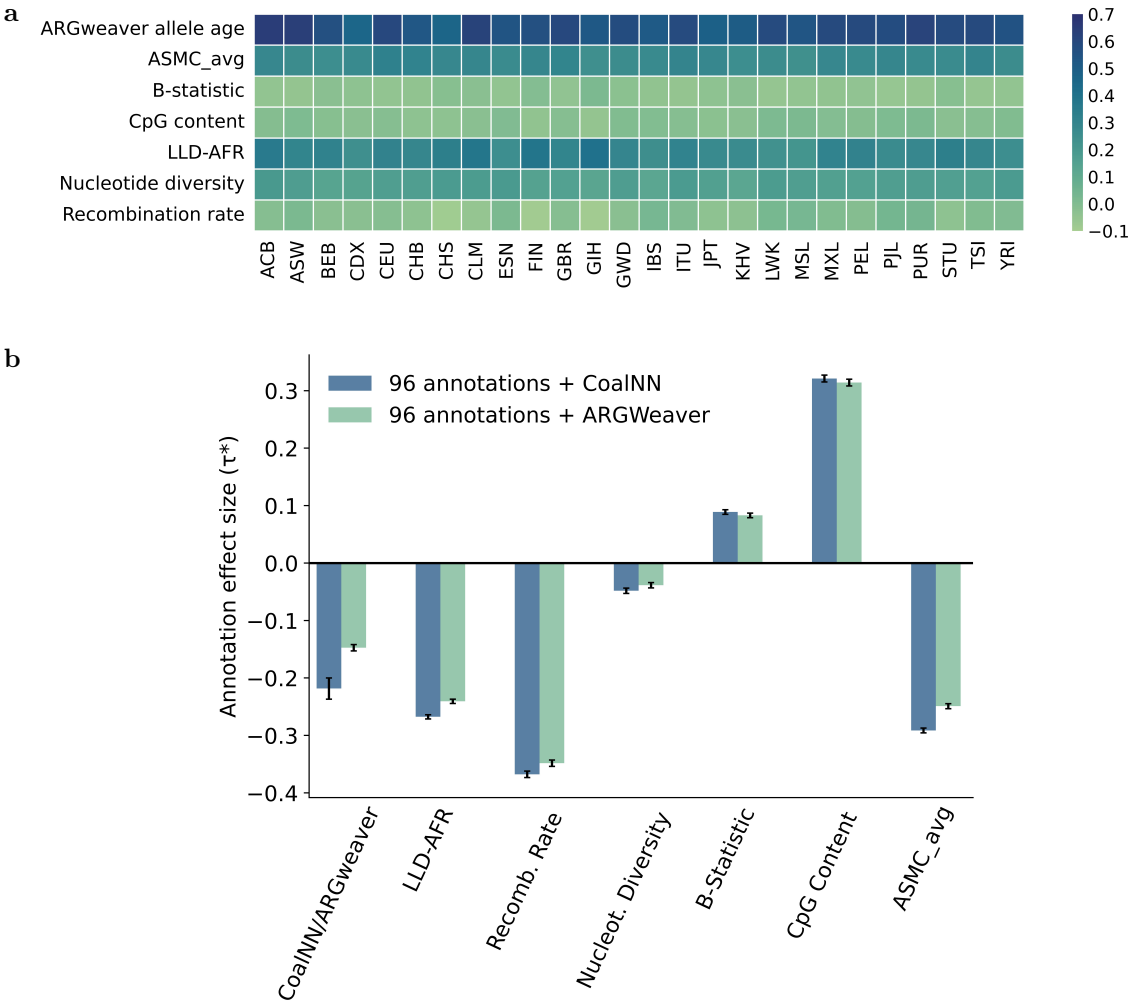


Fig. 5. S-LDSC analysis of CoalNN MAF-adjusted allele age annotations. a. We report correlations computed on common SNPs (MAF $\geq 5\%$) between each of the 26 population specific MAF-adjusted CoalNN annotations and evolutionary annotations from the baseline model. ARGweaver allele age, ASMC_{avg}, and LLD-AFR annotations are also adjusted for MAF. Numerical results are reported in Supplementary Table 9. **b.** Effect size τ^* estimates (meta-analysed across 63 independent diseases and complex traits listed in Supplementary Table 2) of CoalNN MAF-adjusted allele age annotation on all 26 populations (in blue) and of ARGweaver MAF-adjusted allele age annotation [17] (in green), in marginal S-LDSC analysis conditioned on 96 baseline annotations (the full baseline model except for ARGweaver) [52]. We also report effect sizes of baselineLD evolutionary annotations (level of LD measured in African populations LLD-AFR, recombination rate, nucleotide diversity, *B*-statistic [71], CpG content [72] and average pairwise TMRCA ASMC_{avg} [15]) after the introduction of either the CoalNN or ARGWeaver allele age annotation. Error bars represent standard errors of the meta-analysed τ^* estimates. See Supplementary Table 10 for numerical results.

			ASMC mean posterior	ASMC MAP
Sequencing		MAE	17.94 (1.13)	70.57 (1.51)
		RMSE	20.1 (1.23)	62.67 (1.24)
Array		MAE	3.33 (0.67)	116 (1.44)
		RMSE	0.34 (0.79)	71.42 (1.16)
Imputed	panel size = 300	MAE	5.11 (0.88)	35.58 (0.99)
		RMSE	10.62 (1.13)	33.64 (1.14)
	panel size = 1000	MAE	5.49 (1.15)	43.71 (1.35)
		RMSE	12.08 (1.54)	42.63 (1.36)
	panel size = 2000	MAE	5.0 (0.9)	46.16 (1.02)
		RMSE	11.23 (1.2)	44.95 (1.08)

Table 1. Accuracy of pairwise TMRCA prediction. We report the average percent performance improvement of CoalNN over ASMC for the mean absolute error (MAE) and the root mean squared error (RMSE) across 20 simulations. Numbers in round brackets represent standard errors. For imputed data, reference panel sizes are in haploid units.

		RMSE	MAE	MAD	r^2
CEU	**Relate**	24753 (215)	12355 (93)	4271 (46)	0.53 (0.0)
	tsinfer+tsdate	29685 (280)	14363 (118)	4695 (48)	0.37 (0.0)
	CoalNN	**23325 (154)**	**11689 (79)**	**4001 (43)**	**0.58 (0.0)**
Constant	**Relate**	11029 (117)	5341 (44)	1647 (12)	0.63 (0.0)
	tsinfer+tsdate	12909 (258)	5864 (68)	1557 (12)	0.53 (0.0)
	CoalNN	**10901 (125)**	**5108 (46)**	**1487 (11)**	**0.65 (0.0)**

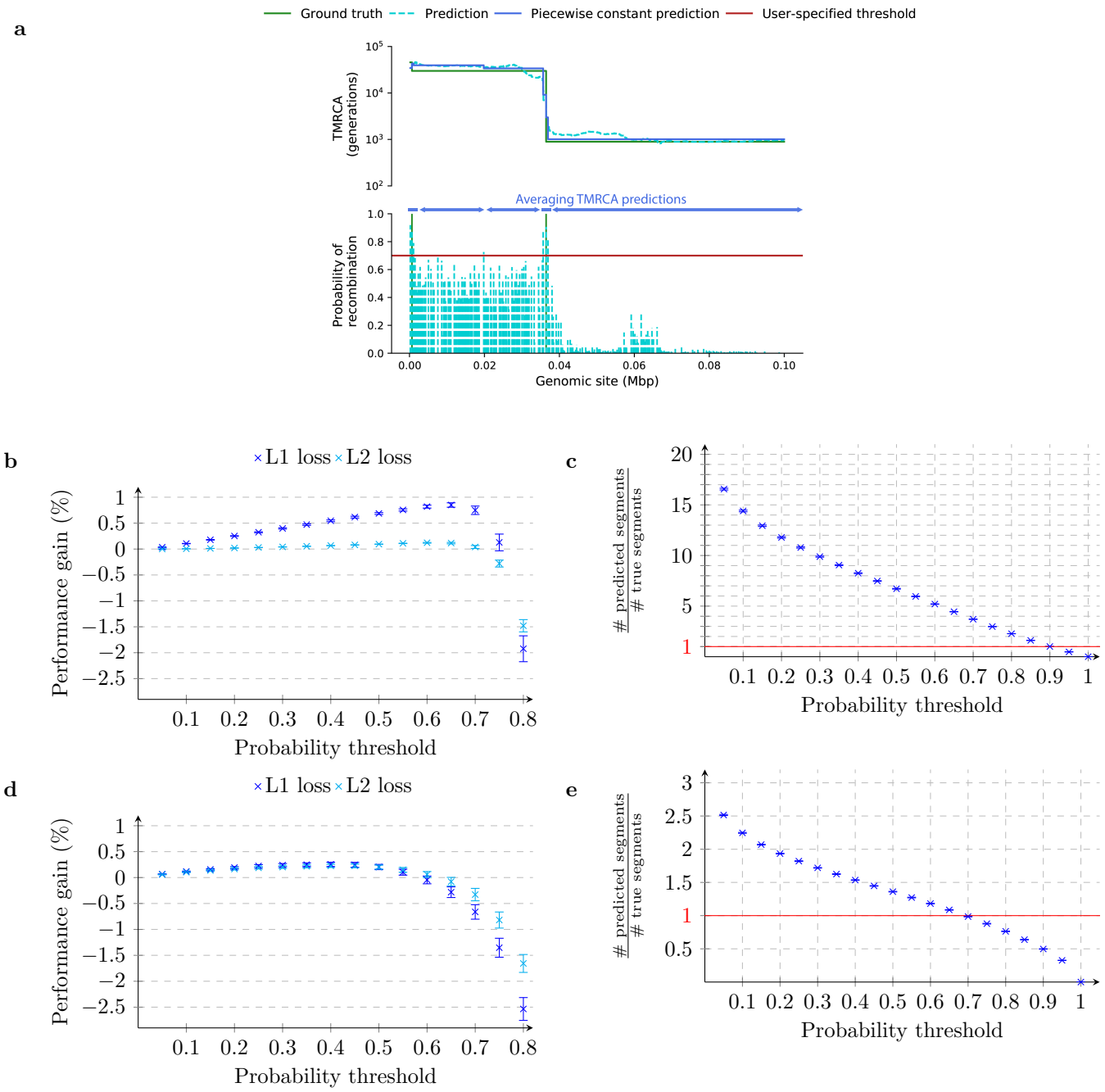
Table 2. Accuracy of estimated allele ages. We report the average performance (in units of generations) of Relate, tsinfer+tsdate and CoalNN on non-singleton variants under a European demographic history model (CEU) and a constant population size ($N_e = 10,000$) model across 10 simulations, with numbers in round brackets representing standard errors. We report the root mean squared error (RMSE), the mean absolute error (MAE), the median absolute deviation (MAD) and the square of the Pearson correlation coefficient (r^2). Methods in bold obtained statistically significantly lower (or higher) scores than non-bold methods (one-sided paired t-test $p \leq 0.05$).

Supplementary Information

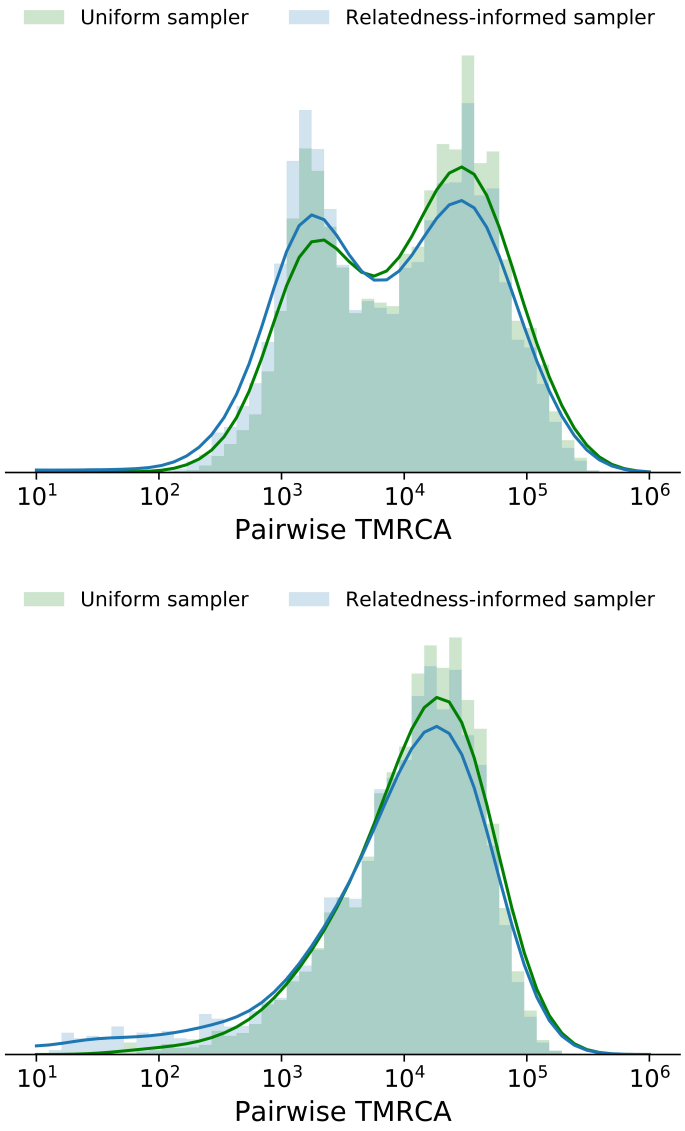
Inference of genealogical relatedness and variant age using deep neural networks identifies signatures of selection

Juba Nait Saada, Zoi Tsangalidou, Miriam Stricker, Pier Francesco Palamara

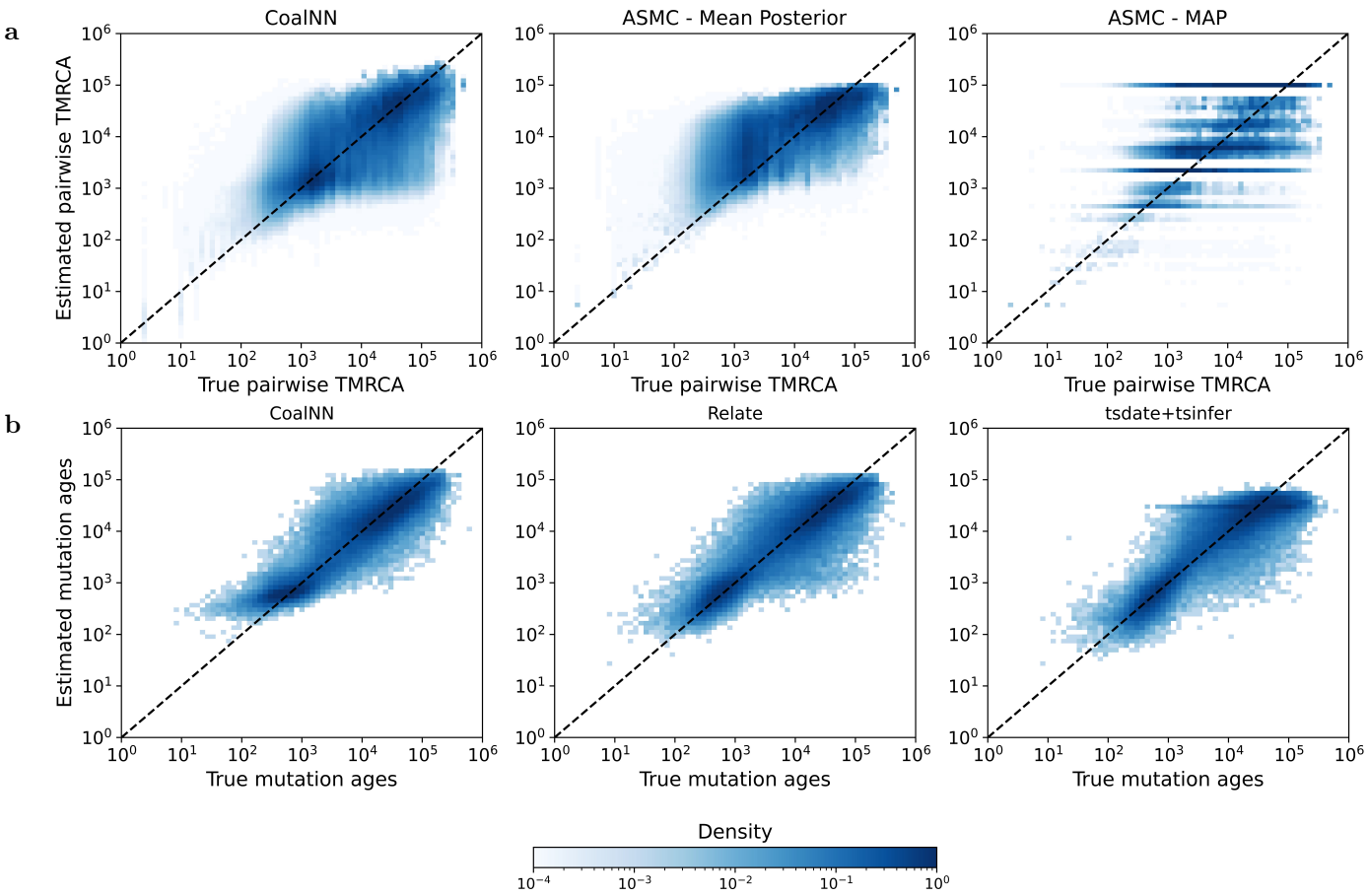
Supplementary Figures



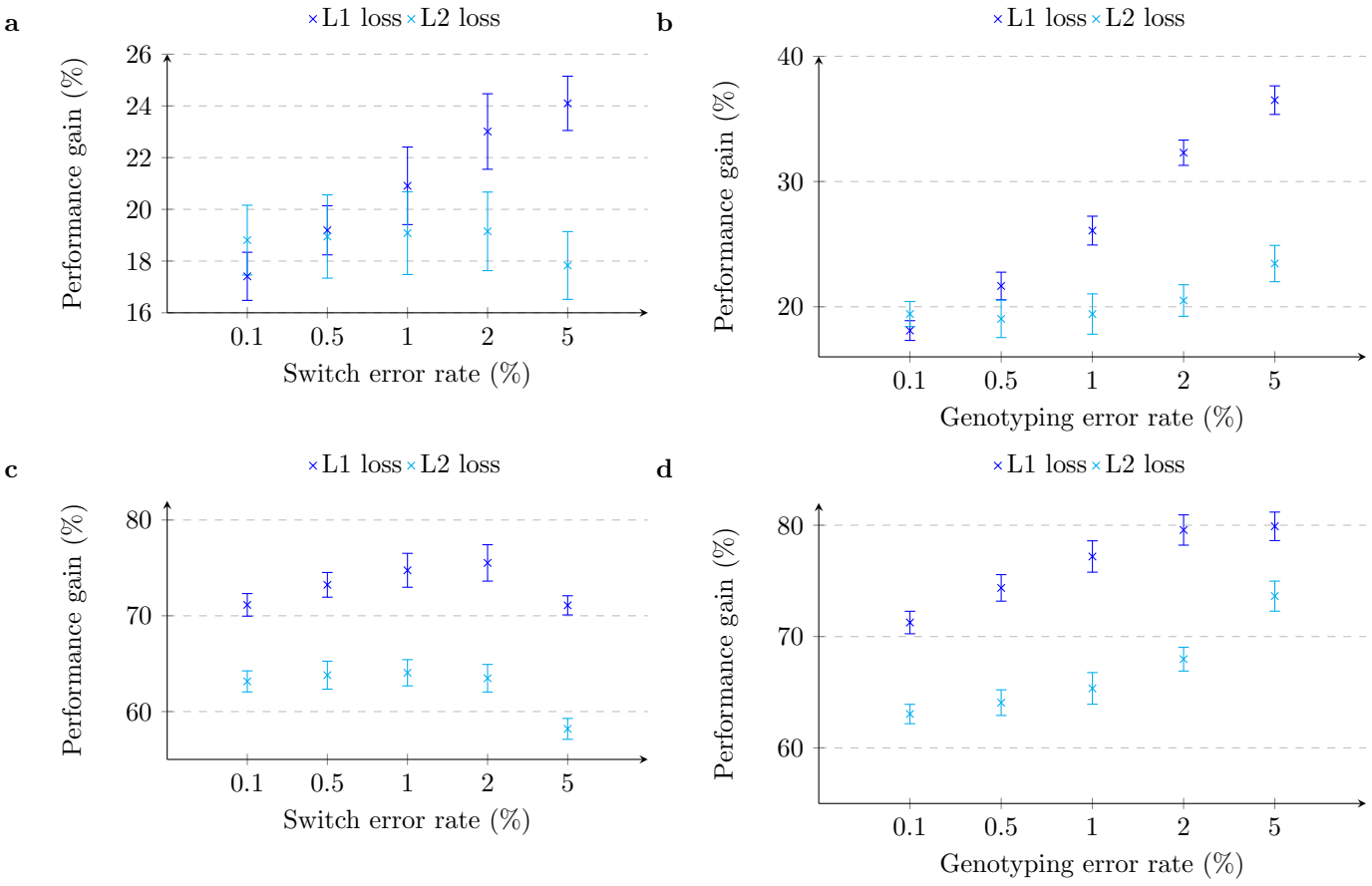
Supplementary Fig. 1. Forming piecewise constant TMRCA estimates. a. When making the output piecewise constant, CoalNN averages all inferred TMRCAs between consecutive genomic sites with an estimated probability of recombination that exceeds a user-specified threshold. **b, d.** In **b** (resp in **d**) we report the average percent improvement of piecewise constant TMRCA prediction over the raw output of CoalNN using L1 and L2 losses for sequencing (resp. array) data, for increasing threshold values. **c, e.** In **c** (resp. in **e**) we report the average ratio between the total number of predicted segments and the total number of true segments for increasing threshold values, where a segment refers to a piece of the genome with constant TMRCA value. Error bars represent standard errors across 10 simulations.



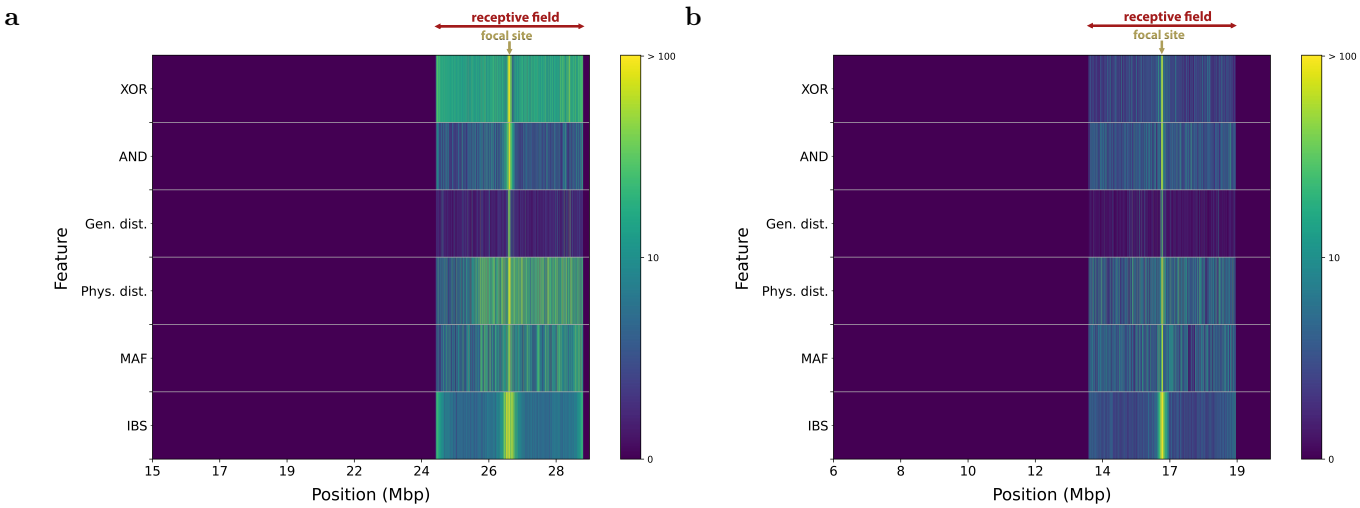
Supplementary Fig. 2. Pairwise TMRCA sampling distribution. Distributions of pairwise TMRCA in generations (empirical histograms) obtained from a uniform sampler (in green) and a relatedness-informed sampler (in blue) from a single simulation under a European demographic model (top) and a constant population size $N_e = 10,000$ (below). Solid curves show kernel density estimations using Gaussian kernels (bandwidth = 0.25).



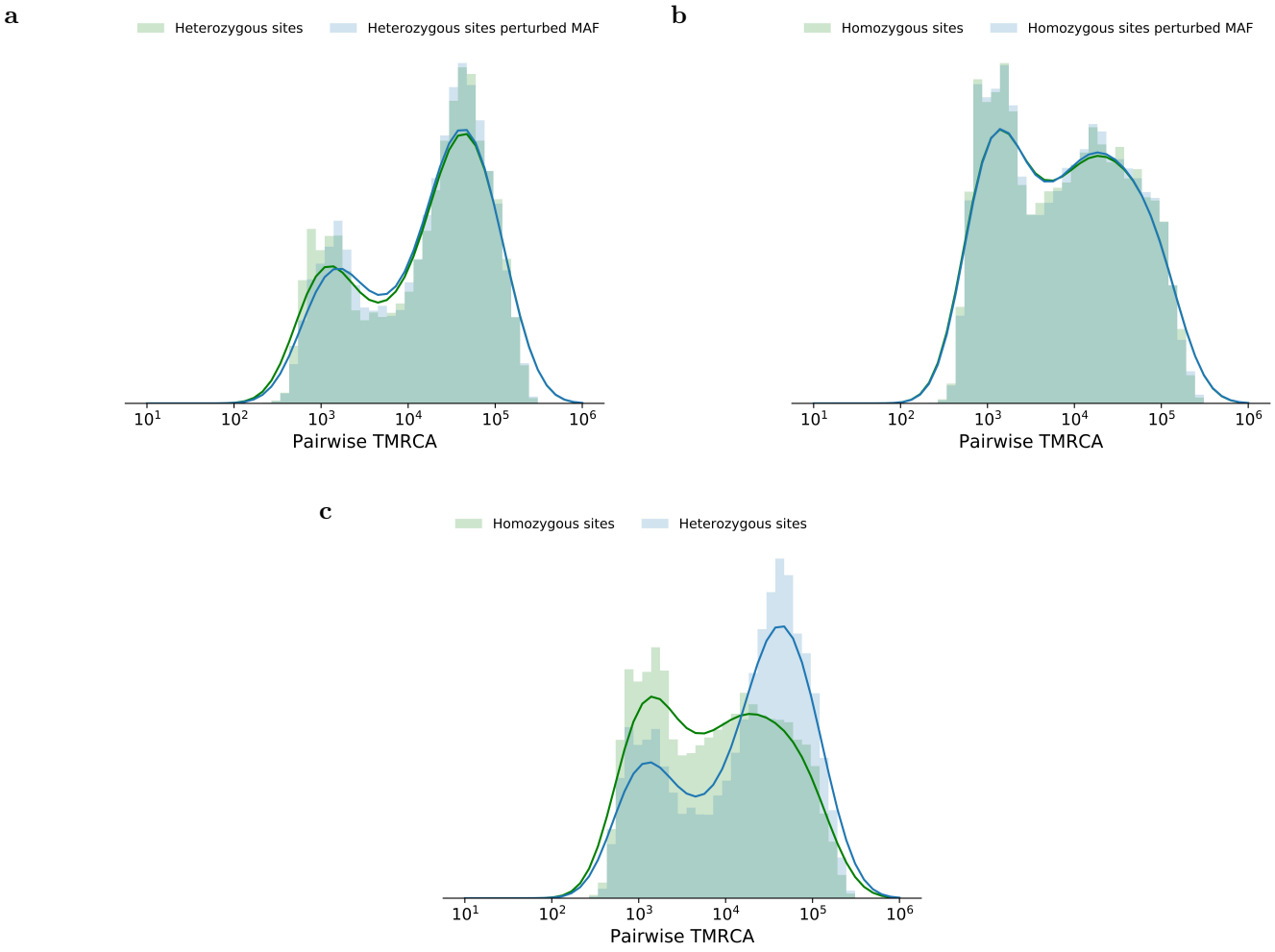
Supplementary Fig. 3. Pairwise TMRCA and allele age prediction under a European demographic model. a. True pairwise TMRCAs (x axis) versus those estimated by CoalNN and ASMC (y axis) on array data for one simulation. **b.** True non-singleton variant ages (x axis) versus those estimated by CoalNN, Relate and tsinfer+tsdate (y axis) on sequencing data for one simulation.



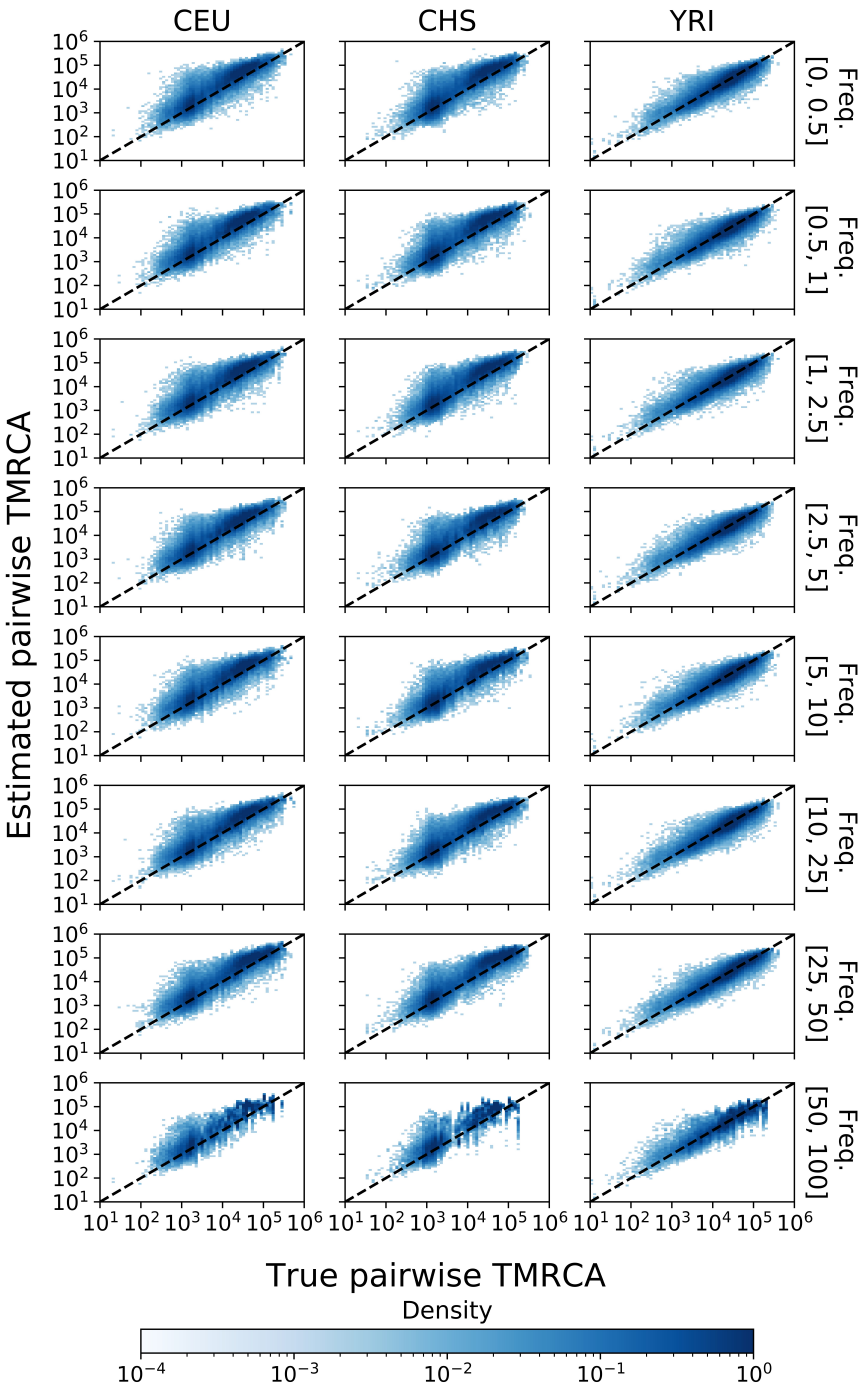
Supplementary Fig. 4. Robustness to genotyping and phasing perrors. Average percent performance improvement of CoalNN over ASMC mean posterior in **a, b** and over ASMC MAP in **c, d** for the L1 and L2 losses for increasing switch and genotyping error rates respectively. Error bars represent standard errors across 20 simulations.



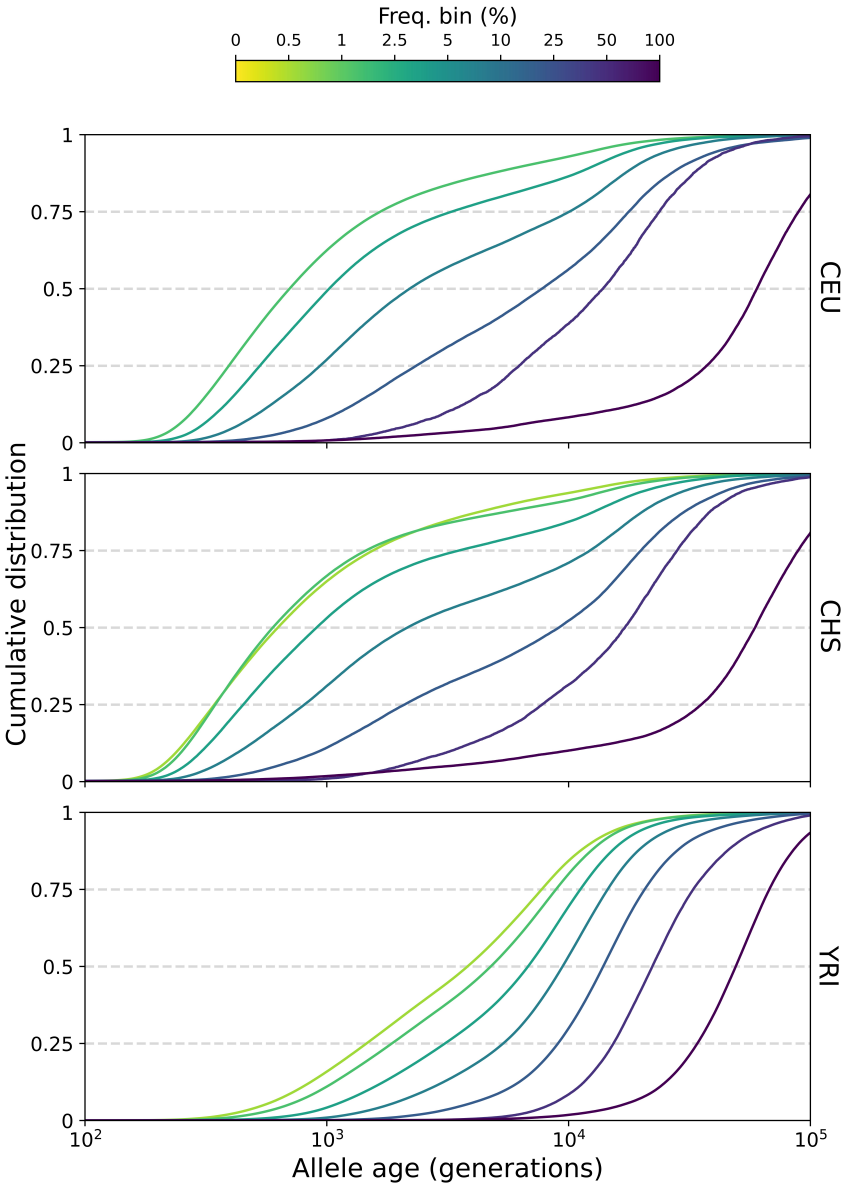
Supplementary Fig. 5. Feature importance analysis. We obtained saliency maps [1, 2] at a genomic locus with recent coalescence time in **a** (TMRCA = 37 generations) and with deep coalescence time in **b** (TMRCA = 237,079 generations). Values are shown in log scale and correspond to |input × gradient| [3], where input refers to the batch normalised input (see Methods). Higher values (closer to yellow) indicate input positions with large effects on the TMRCA estimate. Positions with value 0 have no impact on the prediction and are typically outside of the receptive field.



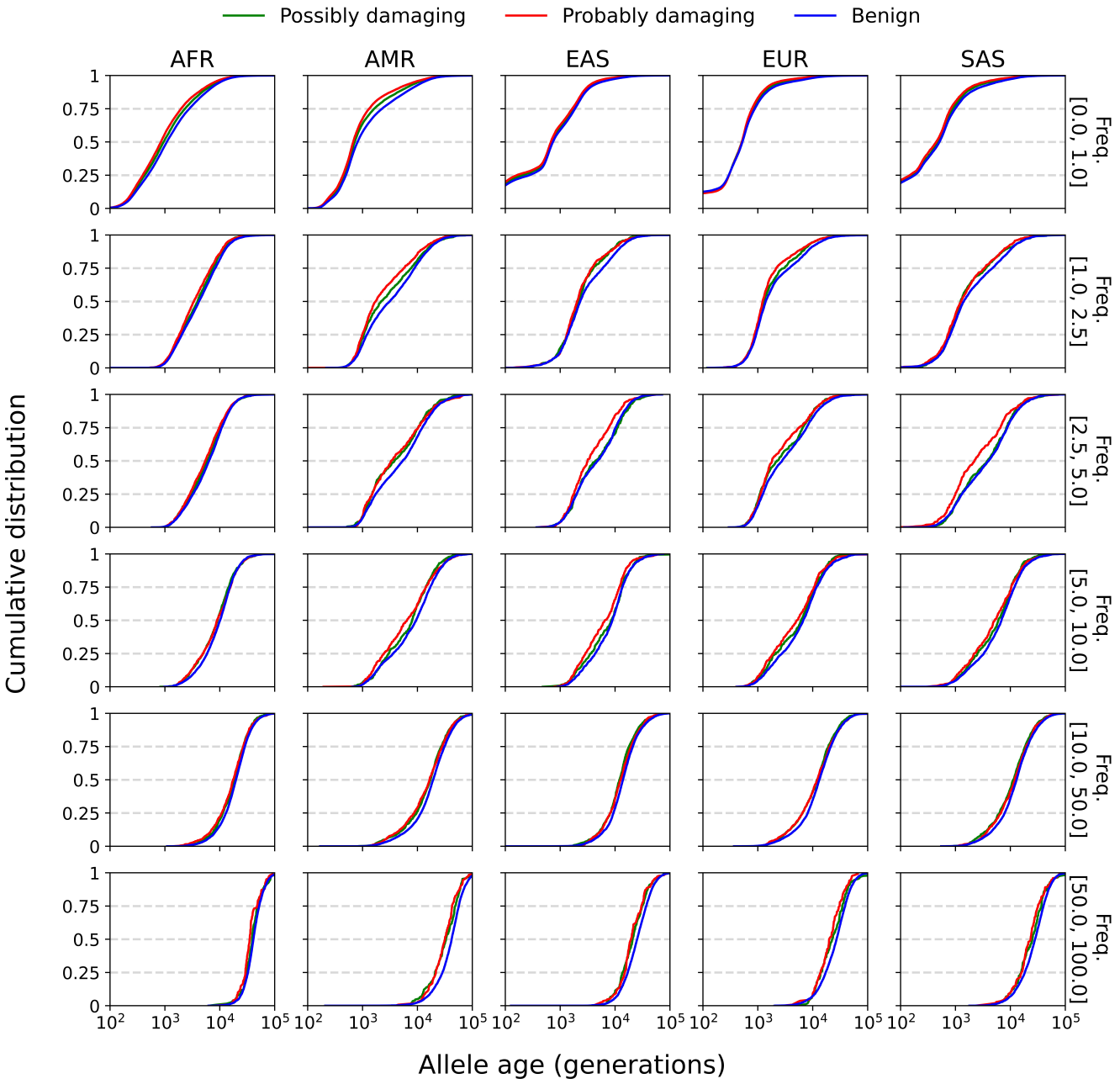
Supplementary Fig. 6. Predicted TMRCA distributions under MAF perturbation. Empirical histograms showing the predicted TMRCAs at heterozygous (in **a**) or at homozygous derived (in **b**) genomic sites before and after the MAF perturbation, for one simulation. Increasing the input MAF value by 5% at sites for which individuals are heterozygous resulted in an average increase in predicted TMRCA of 181.5 (SE=3.7) generations. Decreasing the MAF at homozygous sites resulted in an average decrease of predicted TMRCA of 27.6 (SE=1.7) generations. **c** shows TMRCA values predicted by CoalNN at heterozygous and homozygous derived sites, without any MAF perturbation. Solid curves show kernel density estimations using Gaussian kernels (bandwidth = 0.25).



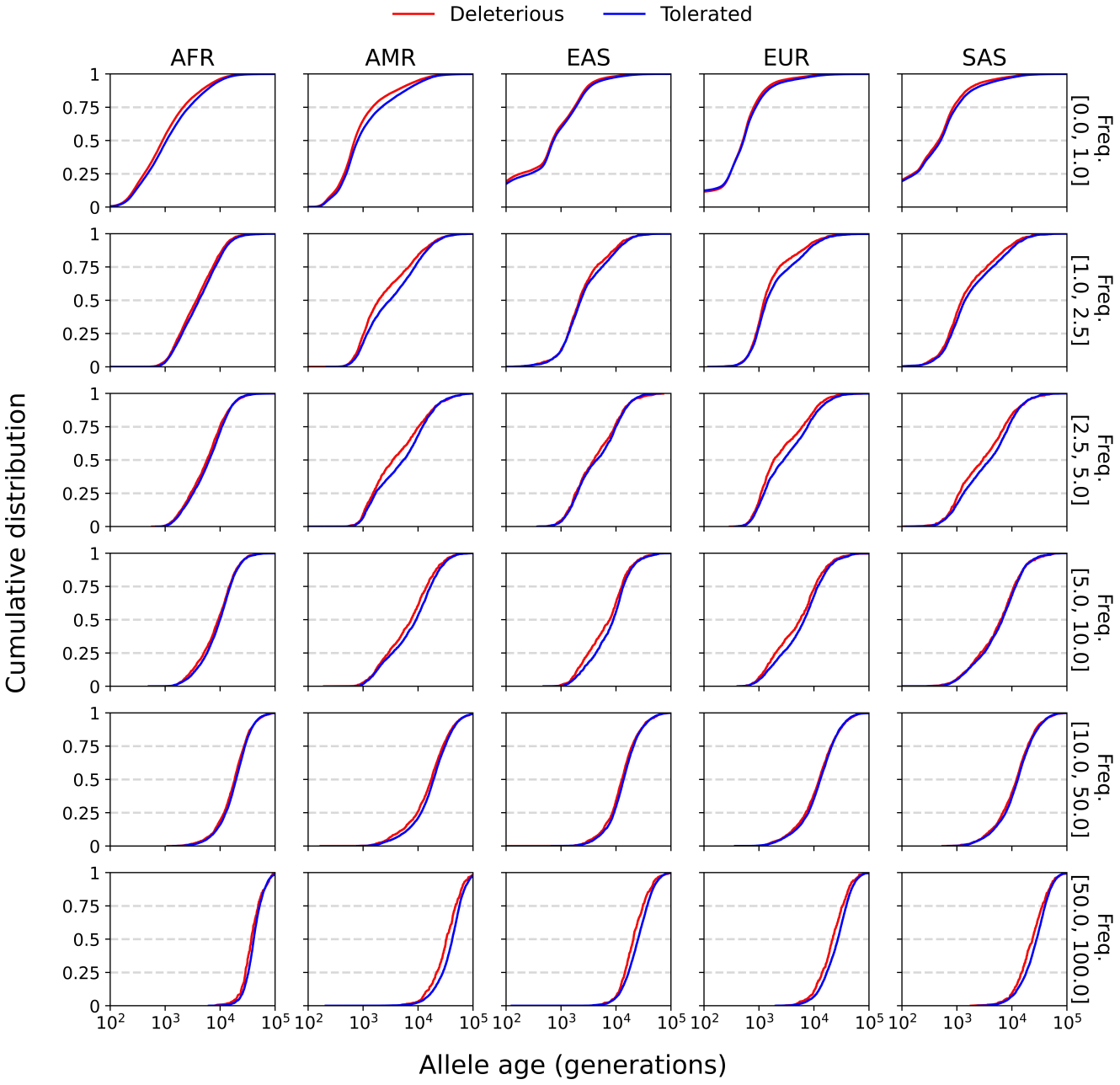
Supplementary Fig. 7. Performance of CoalNN trained on a constant demographic prior. We visualise true pairwise TMRCAs (x axis) versus those estimated by CoalNN trained on a constant demographic population size ($N_e = 20,000$) (y axis), stratified by derived allele frequency, on simulations using recombination rates and demographic history from CEU, CHS and YRI [4, 5] (sequencing data for 150 diploid individuals and a 30 Mbp region from chromosome 2).



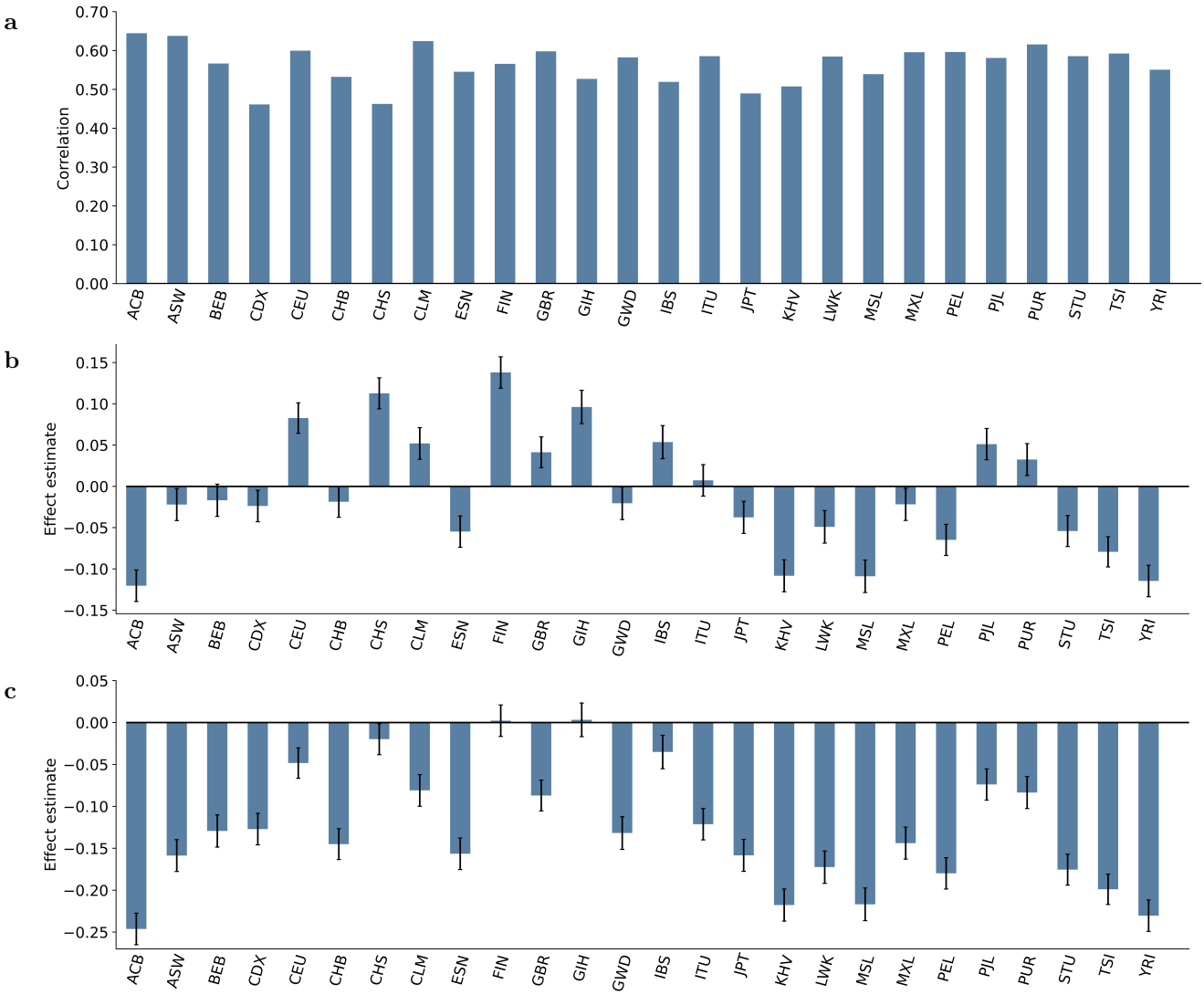
Supplementary Fig. 8. Age distribution of variants dated by CoalNN trained on a constant demographic prior. Cumulative age distribution of genome-wide variants dated by CoalNN trained on a constant demographic population size ($N_e = 20,000$), stratified by derived allele frequency as observed within a given population group. See Supplementary Table 5 for additional information.



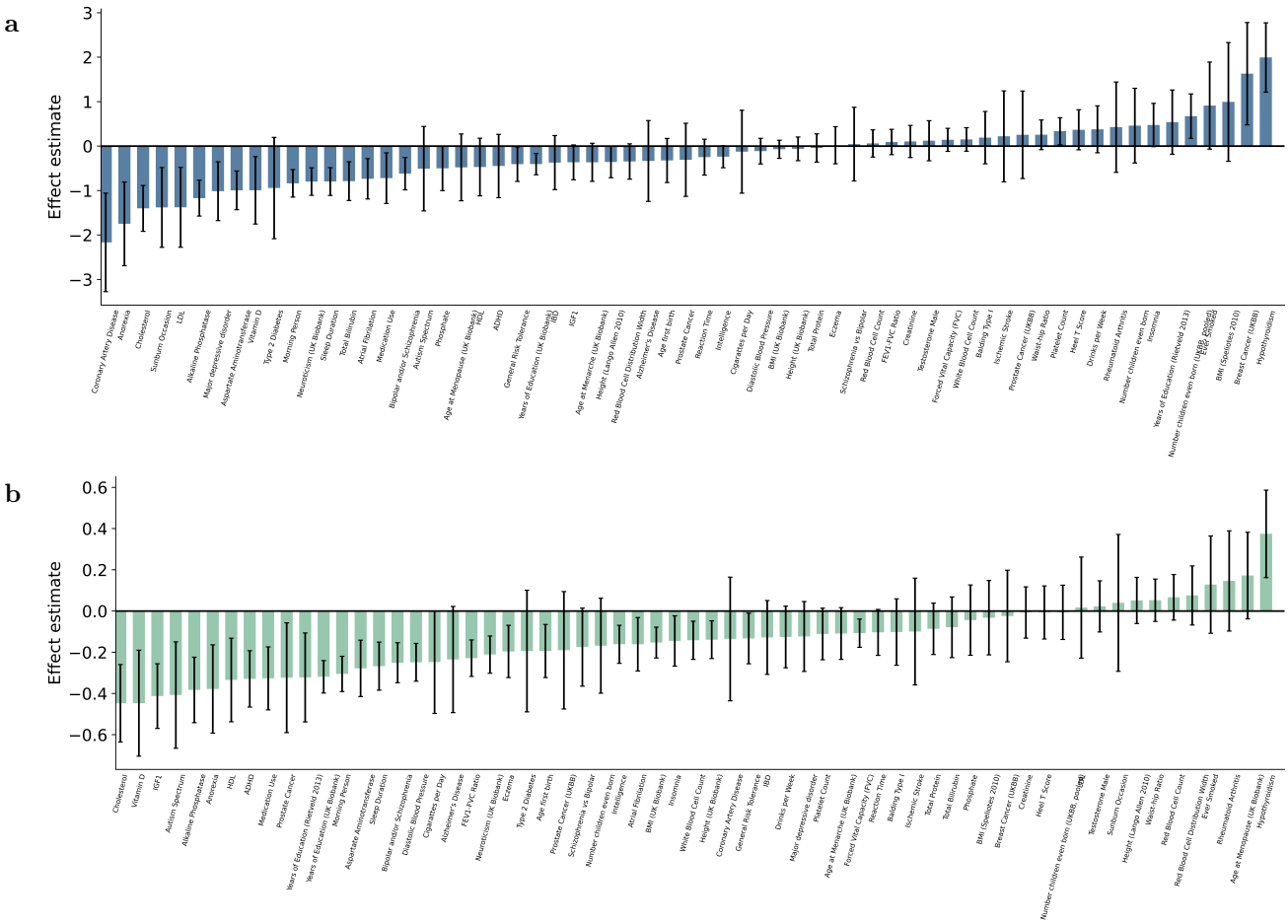
Supplementary Fig. 9. Allele age distribution of variants annotated by PolyPhen-2. Cumulative distribution function of allele ages inferred by CoalNN for different variant pathogenicity levels predicted using PolyPhen-2 ("possibly damaging", "probably damaging" or "benign") [6], stratified by derived allele frequency within each population group.



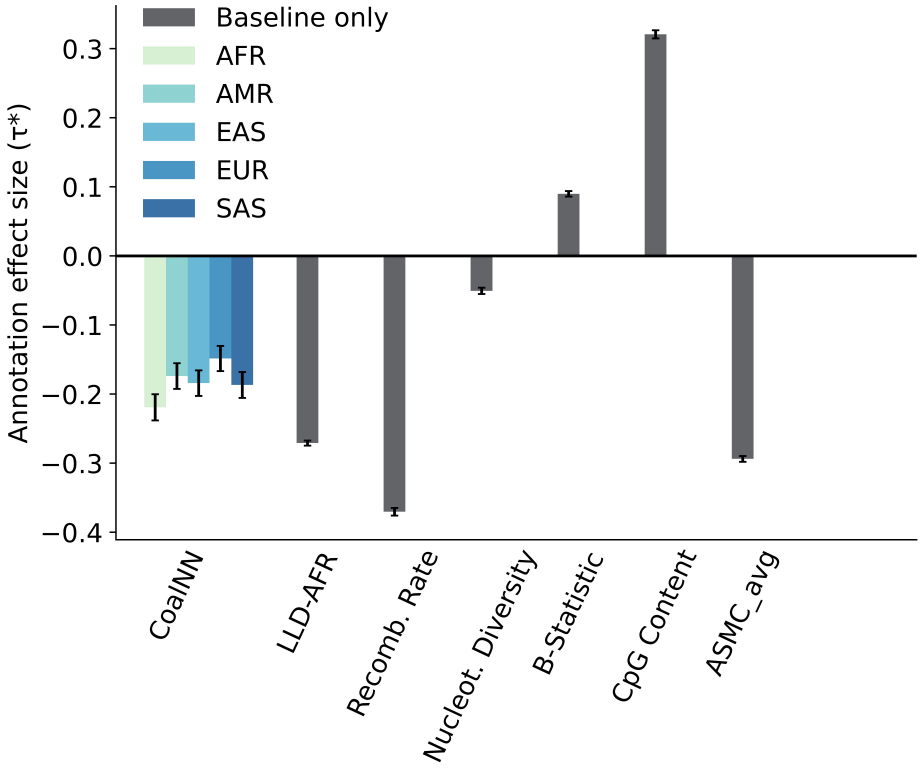
Supplementary Fig. 10. Allele age distribution of variants annotated by SIFT. Cumulative distribution function of allele ages inferred by CoalNN for different variant pathogenicity levels predicted using SIFT ("deleterious" or "tolerated") [7], stratified by derived allele frequency within each population group.



Supplementary Fig. 11. S-LDSC analyses including both ARGweaver and CoalNN. a. Correlations between each of the 26 population-specific MAF-adjusted CoalNN annotations and the MAF-adjusted ARGweaver annotation. **b, c.** τ^* estimates (meta-analysed across 63 independent diseases and complex traits) of CoalNN MAF-adjusted allele age annotation on each of the 26 populations in marginal independent S-LDSC analyses conditioned on the full 97 Baseline-LD model in **b** and conditioned on 96 Baseline-LD annotations (all but ARGweaver) in **c**.



Supplementary Fig. 12. CoalNN heritability effect for individual traits. S-LDSC τ^* value of the CoalNN MAF-adjusted annotation (averaged across 26 populations) in **a** and of the ARGweaver MAF-adjusted annotation in **b** for 63 independent diseases and traits (listed in Supplementary Table 2), conditioned on 96 baselineLD annotations. Error bars represent s.e. of the τ^* estimate.



Supplementary Fig. 13. S-LDSC analysis of CoalNN annotations built using population groups. S-LDSC τ^* estimates of CoalNN annotations built using each population group (AFR, AMR, EAS, EUR and SAS) in independent S-LDSC analyses conditioned on 96 baselineLD annotations (all but ARGweaver), meta-analyzed across 63 independent disease and traits (listed in Supplementary Table 2). For comparison, we also report τ^* effects of other Baseline-LD evolutionary annotations (level of LD measured in African populations, recombination rate, nucleotide diversity, McVicker B-statistic and average pairwise TMRCA ASMC_{avg}) before the introduction of the CoalNN annotations (in dark grey). Adding the CoalNN annotations did not result in a significant change of τ^* for other annotations. Error bars represent standard errors of the meta-analyzed τ^* estimates. Numerical results are reported in Supplementary Table 8.

Supplementary Tables

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

Supplementary Table 1. 1kG populations. Population labels and supergroup assignments. (Retrieved from www.internationalgenome.org)

Phenotype	Reference	h_g^2	Z-score	N
BMI (UK Biobank)	UK Biobank	0.278	41.478	457,824
Years of Education (UK Biobank)	UK Biobank	0.136	35.789	454,813
Forced Vital Capacity (FVC)	UK Biobank	0.234	32.444	371,949
Intelligence	Savage et al. 2018 Nat Genet	0.186	29.047	264,193
Neuroticism (UK Biobank)	UK Biobank	0.114	28.475	372,066
Diastolic Blood Pressure	UK Biobank	0.233	27.738	422,771
White Blood Cell Count	UK Biobank	0.227	27.386	444,502
Height (UK Biobank)	UK Biobank	0.674	25.824	458,303
Morning Person	UK Biobank	0.106	24.674	410,520
Age at Menarche (UK Biobank)	UK Biobank	0.253	24.563	242,278
TotalProtein	UK Biobank	0.181	24.147	397,652
Waist-hip Ratio	UK Biobank	0.171	23.681	458,417
FEV1-FVC Ratio	UK Biobank	0.276	23.615	371,949
Reaction Time	Davies et al. 2018 Nat Comm	0.081	22.556	300,486
Creatinine	UK Biobank	0.221	21.911	434,158
Red Blood Cell Count	UK Biobank	0.261	21.775	445,174
Bipolar and/or Schizophrenia	Ruderfer et al. 2018 Cell	0.334	21.656	107,620
Platelet Count	UK Biobank	0.351	21.242	444,382
Heel T Score	UK Biobank	0.355	20.514	445,921
Sleep Duration	Dashti et al. 2019 Nat Comm	0.071	19.667	446,118
IGF1	UK Biobank	0.288	18.682	432,292
General Risk Tolerance	Karlsson Linner et al. 2019 Nat Genet	0.051	18.357	466,571
AspartateAminotransferase	UK Biobank	0.114	18.306	430,982
Red Blood Cell Distribution Width	UK Biobank	0.217	17.803	442,700
Insomnia	Jansen et al. 2019 Nat Genet	0.044	17.600	383,948
Height (Lango Allen 2010)	Lango Allen et al. 2010 Nature	0.222	17.512	131,547
Number children even born (UKBB, pooled)	UK Biobank	0.039	16.783	456,500
Atrial Fibrilation	Nielsen et al. 2018 AJHG	0.022	16.769	1,030,836
TestosteroneMale	UK Biobank	0.184	16.577	196,813
BMI (Speliotes 2010)	Speliotes et al., 2010 Nat Genet	0.146	16.233	122,033
Hypothyroidism	UK Biobank	0.055	15.167	459,324
Age first birth	Barban et al., 2016 Nat Genet	0.062	15.073	222,037
Balding Type I	UK Biobank	0.223	14.697	208,336
AlkalinePhosphatase	UK Biobank	0.235	14.675	433,862
Eczema	UK Biobank	0.085	14.603	458,699
Drinks per Week	Liu et al. 2019 Nat Genet	0.046	14.438	527,299
Phosphate	UK Biobank	0.129	14.198	397,561
Medication Use	Wu et al. 2019 Nat Comm	0.157	14.036	78,808
IBD	de Lange et al. 2017 Nat Genet	0.321	13.193	59,957
VitaminD	UK Biobank	0.086	12.464	415,700
Major depressive disorder	Wray et al. 2018 Nat Genet	0.073	12.339	170,229
Cholesterol	UK Biobank	0.131	11.670	435,137
ADHD	Demontis et al. 2019 Nat Genet	0.248	11.634	55,374
TotalBilirubin	UK Biobank	0.084	11.066	429,423
Age at Menopause (UK Biobank)	UK Biobank	0.109	11.040	143,025
Sunburn Occasion	UK Biobank	0.073	9.618	344,229
Number children even born	Barban et al., 2016 Nat Genet	0.023	9.542	318,863
Schizophrenia vs Bipolar	Ruderfer et al. 2018 Cell	0.265	9.495	38,855
Years of Education (Rietveld 2013)	Rietveld et al., 2013 Science	0.077	9.354	126,559
HDL	Teslovich et al., 2010 Nature	0.125	9.259	97,749
LDL	Teslovich et al., 2010 Nature	0.103	8.821	93,354
Cigarattes per Day	Liu et al. 2019 Nat Genet	0.058	8.657	257,073
Ischemic Stroke	Malik et al. 2018 Nat Genet	0.017	8.350	440,328
Anorexia	Boraska et al., 2014 Mol Psych	0.236	8.274	32,143
Alzheimer's Disease	Jansen et al. 2019 Nat Genet	0.014	7.833	433,886
Rheumatoid Arthritis	Okada et al., 2014 Nature	0.187	7.510	37,681
Ever Smoked	TAG Consortium, 2010 Nat Genet	0.085	7.482	74,035
Prostate Cancer	Schumacher et al. 2018 Nat Genet	0.131	7.016	72,729
Prostate Cancer (UKBB)	UK Biobank	0.017	7.000	459,324
Breast Cancer (UKBB)	UK Biobank	0.018	6.846	459,324
Type 2 Diabetes	Morris et al., 2012 Nat Genet	0.095	6.463	60,786
Autism Spectrum	PGC Cross-Disorder Group, 2013 Lancet	0.622	6.191	10,263
Coronary Artery Disease	Schunkert et al., 2011 Nat Genet	0.075	6.016	77,210

Supplementary Table 2. Traits used in S-LDSC analyses. We report phenotype name, study reference, SNP heritability h_g^2, heritability Z-score, and number of samples N for the 63 independent diseases and complex traits [8] used in S-LDSC analyses.

	AND gate	XOR gate
Channel 1	0.035 (0.001)	-0.009 (0.001)
Channel 2	0.024 (0.000)	-0.056 (0.001)
Channel 3	-0.008 (0.001)	0.017 (0.001)
Channel 4	0.014 (0.000)	0.002 (0.000)
Channel 5	-0.232 (0.001)	-0.016 (0.001)
Channel 6	-0.051 (0.000)	0.08 (0.001)
Channel 7	-0.272 (0.001)	-0.172 (0.001)
Channel 8	-0.057 (0.000)	0.101 (0.001)

Supplementary Table 3. Recovery of the AND/XOR features from raw haplotypes. Pearson correlation coefficient between the XOR/AND logic gates calculated on the raw haplotypes and the 8 channels of output of the first ConvBlock (see Methods). The numbers in brackets are standard errors calculated over 20 simulations.

DAF bin (%)	Number of dated variants					Median age (generations)				
	AFR	AMR	EAS	EUR	SAS	AFR	AMR	EAS	EUR	SAS
[0, 0.5]	5,156,014	2,248,935	1,819,869	1,648,497	2,159,787	2,160 (34)	1,570 (16)	1,045 (38)	960 (16)	927 (61)
[0.5, 1]	2,762,406	1,884,075	1,191,727	1,358,230	1,493,628	3,564 (53)	3,071 (53)	1,703 (25)	1,137 (9)	1,066 (16)
[1, 2.5]	3,082,203	1,923,613	913,021	1,181,402	1,384,546	4,904 (53)	4,852 (64)	2,634 (27)	1,593 (20)	1,517 (29)
[2.5, 5]	2,333,277	918,075	646,257	848,596	846,087	7,138 (65)	7,620 (67)	7,089 (93)	4,593 (62)	5,041 (47)
[5, 10]	2,028,081	958,302	736,868	904,633	930,768	1,0581 (71)	11,303 (82)	10,395 (79)	8,046 (48)	8,187 (48)
[10, 25]	2,350,274	1,536,898	1,321,044	1,479,233	1,547,460	17,193 (84)	17,737 (67)	13,441 (56)	12,138 (40)	11,802 (40)
[25, 50]	1,523,362	1,404,734	1,312,117	1,377,671	1,420,932	28,476 (95)	2,6709 (123)	17,970 (93)	18,847 (83)	17,873 (83)
[50, 100]	2,057,347	2,184,241	1,954,993	2,015,680	1,979,944	43,636 (153)	44,824 (201)	28,261 (198)	29,741 (150)	31,261 (172)

Supplementary Table 4. Median age of dated variants among different population groups. We report the number of variants dated by CoalNN and their median age in generations stratified by derived allele frequency (DAF) as observed within a given population group. We only considered non-singleton polymorphic variants for which high-confidence ancestral information was available [9]. Numbers in round brackets represent standard errors obtained by bootstrapping (see Methods).

Derived allele frequency bin (%)	Number of dated variants			Median age (generations)		
	CEU	CHS	YRI	CEU	CHS	YRI
[0.5, 1]	N.A	516,114	1,477,617	N.A	628 (17)	3,818 (34)
[1, 2.5]	1,101,465	510,236	2,243,382	701 (18)	599 (17)	4,762 (33)
[2.5, 5]	44,8342	253,321	1,638,788	1,023 (20)	907 (30)	6,795 (34)
[5, 10]	225,484	147,593	1,374,079	2,203 (73)	2,165 (138)	9,438 (39)
[10, 25]	94,380	83,756	833,036	7,761 (314)	9,092 (418)	13,901 (57)
[25, 50]	7,934	12,096	118,731	13,850 (574)	17,104 (727)	22,201 (175)
[50, 100]	58,060	53,894	423,917	60,023 (719)	59,019 (493)	49,115 (215)

Supplementary Table 5. Median age of variants among different populations dated by CoalNN trained on a demographic prior of constant size. We report the number of variants dated by CoalNN trained on a constant population size ($N_e = 20,000$) and their median age in generations, stratified by derived allele frequency as observed within CEU, CHS or YRI. Only variants that remained polymorphic within each population were considered. Singletons were not included. Numbers in round brackets represent standard errors obtained by bootstrapping (see Methods). N.A denotes the absence of dated variant within the derived allele frequency bin for the corresponding population.

	DAF bin (%)	Number of dated variants					Median age (generations)				
		AFR	AMR	EAS	EUR	SAS	AFR	AMR	EAS	EUR	SAS
Benign	[0, 1]	60243	40505	36714	41811	21027	1083 (24)	812 (10)	741 (36)	524 (10)	510 (19)
	[1, 2.5]	7573	4742	2655	3384	3181	4187 (107)	3115 (142)	2311 (44)	1391 (21)	1416 (56)
	[2.5, 5]	5277	2217	1590	2185	2005	6364 (80)	5385 (313)	4807 (390)	2743 (126)	4004 (176)
	[5, 10]	4215	2098	1632	2028	2053	10116 (181)	9433 (312)	8917 (261)	6730 (242)	6868 (182)
	[10, 50]	7264	5876	5233	5724	5837	19695 (243)	19293 (329)	13908 (190)	13168 (188)	12956 (189)
	[50, 100]	3949	4367	3802	3997	3807	40953 (301)	41976 (444)	25168 (423)	27154 (441)	28143 (402)
Poss. damaging	[0, 1]	17193	11322	11085	13094	5926	949 (24)	720 (11)	709 (32)	510 (11)	480 (23)
	[1, 2.5]	1725	1100	685	829	794	3908 (116)	2204 (178)	2123 (70)	1284 (39)	1243 (34)
	[2.5, 5]	1042	480	351	510	439	5952 (244)	3812 (435)	4394 (505)	2255 (224)	3787 (457)
	[5, 10]	760	440	331	381	400	9055 (305)	7948 (412)	8509 (609)	6161 (390)	6096 (358)
	[10, 50]	1018	816	777	804	823	18243 (365)	17221 (578)	12037 (431)	11993 (421)	11295 (447)
	[50, 100]	130	174	211	172	174	38587 (970)	33168 (2346)	20898 (925)	22610 (1776)	24821 (1798)
Prob. damaging	[0, 1]	22284	15063	16221	19044	7941	846 (18)	675 (10)	677 (30)	511 (11)	445 (22)
	[1, 2.5]	1803	1145	696	975	878	3362 (135)	1719 (108)	2036 (65)	1226 (38)	1262 (49)
	[2.5, 5]	1067	425	375	440	419	5433 (179)	3390 (396)	3193 (230)	1816 (171)	2211 (289)
	[5, 10]	703	409	283	364	410	8963 (373)	6391 (538)	6673 (794)	5148 (468)	5190 (389)
	[10, 50]	810	694	678	694	654	17500 (444)	16570 (692)	12561 (420)	12095 (510)	11993 (369)
	[50, 100]	100	131	170	136	146	34742 (1125)	31530 (2180)	19270 (1157)	20549 (1538)	22579 (1657)

Supplementary Table 6. Median age of variants among different population groups annotated by PolyPhen-2. We report the number of annotated variants by PolyPhen-2 dated by CoalNN and their median age in generations, stratified by derived allele frequency (DAF) as observed within a given population group.

	DAF bin (%)	Number of dated variants					Median age (generations)				
		AFR	AMR	EAS	EUR	SAS	AFR	AMR	EAS	EUR	SAS
Tolerated	[0, 1]	63965	42999	39319	45297	22214	1059 (22)	797 (9)	733 (32)	523 (11)	505 (16)
	[1, 2.5]	7838	4929	2754	3578	3366	4113 (91)	3058 (140)	2284 (45)	1388 (26)	1407 (49)
	[2.5, 5]	5352	2275	1673	2263	2030	6335 (115)	5443 (323)	4453 (364)	2807 (123)	3952 (201)
	[5, 10]	4262	2169	1637	2049	2151	9946 (159)	9236 (264)	8912 (271)	6677 (269)	6689 (212)
	[10, 50]	7290	5859	5314	5723	5870	19505 (197)	19150 (328)	13741 (157)	13042 (169)	12804 (167)
	[50, 100]	3900	4346	3772	3958	3777	40922 (324)	41934 (463)	25187 (489)	27256 (442)	28157 (402)
Deleterious	[0, 1]	36868	24710	25536	29613	13149	900 (18)	698 (9)	698 (34)	510 (10)	467 (22)
	[1, 2.5]	3361	2121	1341	1685	1559	3716 (112)	1925 (91)	2154 (47)	1247 (29)	1222 (33)
	[2.5, 5]	2127	870	678	912	859	5761 (155)	3489 (259)	3954 (325)	1890 (128)	2923 (184)
	[5, 10]	1438	823	634	759	758	9243 (253)	7474 (425)	7638 (525)	5570 (360)	6267 (333)
	[10, 50]	1878	1563	1414	1532	1481	18326 (332)	17349 (428)	12800 (318)	12449 (337)	12143 (262)
	[50, 100]	294	370	451	384	391	36896 (1064)	32819 (1539)	20601 (507)	21681 (889)	23579 (1261)

Supplementary Table 7. Median age of variants among different population groups annotated by SIFT. We report the number of variants annotated by SIFT dated by CoalNN and their median age in generations, stratified by derived allele frequency (DAF) as observed within a given population group.

Annotation	Effect size τ^* not including CoalNN	Effect size τ^* including CoalNN				
		AFR	AMR	EAS	EUR	SAS
CoalNN	N.A	-0.219 (0.019)	-0.174 (0.019)	-0.184 (0.019)	-0.148 (0.018)	-0.187 (0.019)
LLD-AFR	-0.271 (0.004)	-0.267 (0.004)	-0.269 (0.004)	-0.268 (0.004)	-0.269 (0.004)	-0.268 (0.004)
Recombination rate	-0.370 (0.006)	-0.367 (0.006)	-0.369 (0.006)	-0.369 (0.006)	-0.369 (0.006)	-0.368 (0.006)
Nucleotide diversity	-0.051 (0.005)	-0.046 (0.005)	-0.050 (0.005)	-0.048 (0.005)	-0.050 (0.005)	-0.049 (0.005)
B-statistic	0.090 (0.004)	0.089 (0.004)	0.089 (0.004)	0.089 (0.004)	0.089 (0.004)	0.089 (0.004)
CpG content	0.321 (0.006)	0.321 (0.006)	0.321 (0.006)	0.321 (0.006)	0.321 (0.006)	0.321 (0.006)
$ASMC_{avg}$	-0.294 (0.004)	-0.292 (0.004)	-0.292 (0.004)	-0.291 (0.004)	-0.293 (0.004)	-0.292 (0.004)

Supplementary Table 8. S-LDSC analysis of CoalNN annotation on population groups. τ^* estimates (meta-analysed across 63 independent diseases and complex traits) of CoalNN MAF-adjusted allele age annotation on each population group (AFR, AMR, EAS, EUR and SAS) in independent marginal S-LDSC analyses conditioned on 96 baseline annotations (the full baseline model except for ARGweaver). We also report effect sizes of Baseline-LD evolutionary annotations (level of LD measured in African populations LLD-AFR, recombination rate, nucleotide diversity, B-statistic [10], CpG content [11] and average pairwise TMRCA $ASMC_{avg}$ [12]), before and after the introduction of the allele age annotation.

	ARGweaver allele age	$ASMC_{avg}$	B-statistic	CpG content	LLD-AFR	Nucleotide diversity	Recombination rate
ACB	0.64462	0.28373	-0.0449627	-0.0068927	0.357883	0.193833	-0.0115892
ASW	0.637648	0.262087	-0.0511699	0.0073261	0.308689	0.175725	0.0188068
BEB	0.566617	0.25335	-0.0249983	-0.0177649	0.315825	0.139136	-0.0245286
CDX	0.461265	0.273944	-0.0317672	-0.0211324	0.248805	0.143154	-0.0212694
CEU	0.599494	0.321749	-0.0459838	-0.0102032	0.314191	0.168587	-0.0112538
CHB	0.532193	0.31198	-0.0411218	-0.0324496	0.291585	0.166983	-0.0367068
CHS	0.462519	0.28222	-0.0184473	-0.0334043	0.331941	0.183432	-0.0848482
CLM	0.624238	0.29408	-0.0219222	-0.0232567	0.379901	0.181055	-0.0551782
ESN	0.545247	0.268409	-0.0438037	0.00498154	0.253668	0.187784	0.0165792
FIN	0.565649	0.294069	-0.00533878	-0.0418206	0.38372	0.143829	-0.0882545
GBR	0.597827	0.302148	-0.0403849	-0.00995346	0.308844	0.151967	-0.0144756
GIH	0.527073	0.244278	0.0185592	-0.0524937	0.405602	0.129283	-0.0885153
GWD	0.58232	0.260541	-0.0263054	0.000694733	0.2974	0.174894	-0.0206123
IBS	0.519219	0.276006	-0.0430249	-0.000972771	0.252918	0.12374	0.0289131
ITU	0.585557	0.305872	-0.0515937	-0.0102396	0.311108	0.182995	0.00317832
JPT	0.489656	0.282032	-0.0326494	-0.026941	0.275463	0.158177	-0.0352778
KHV	0.507532	0.253314	-0.0239953	-0.0244129	0.267331	0.125592	-0.0289965
LWK	0.584643	0.260102	-0.0550596	0.0119386	0.235916	0.178538	0.0261056
MSL	0.539203	0.243146	-0.0455326	0.0163128	0.214835	0.16732	0.0288875
MXL	0.59563	0.2825	-0.0410389	-0.0051117	0.312468	0.158115	0.000506131
PEL	0.59611	0.275924	-0.0391656	-0.00776521	0.314023	0.15596	-0.00922475
PJL	0.580811	0.293767	-0.057732	0.00708515	0.268913	0.167449	0.0398443
PUR	0.615749	0.274605	-0.0522987	0.0173573	0.283352	0.172801	0.0367664
STU	0.585506	0.265688	-0.0169196	-0.0237441	0.334048	0.159808	-0.0349042
TSI	0.592384	0.303182	-0.0535334	-0.0155446	0.296478	0.154585	0.00063352
YRI	0.550575	0.262261	-0.0448539	0.00265146	0.254752	0.185644	0.00491168

Supplementary Table 9. Correlation between MAF-stratified CoalNN allele age annotation and other evolutionary annotations. We report correlations computed on common SNPs (MAF $\geq$ 5%) between each of the 26 population specific MAF-adjusted CoalNN annotations and evolutionary annotations from the Baseline-LD model. ARGweaver allele age, $ASMC_{avg}$, and LLD-AFR annotations were also MAF-stratified.

a

Annotation	Effect size τ^* not including CoalNN	Effect size τ^* including CoalNN
CoalNN	N.A	-0,218 (0,018)
LLD-AFR	-0,271 (0,004)	-0,268 (0,004)
Recombination rate	-0,370 (0,006)	-0,368 (0,006)
Nucleotide diversity	-0,051 (0,005)	-0,048 (0,005)
B-statistic	0,090 (0,004)	0,089 (0,004)
CpG content	0,321 (0,006)	0,321 (0,006)
ASMC_{avg}	-0,294 (0,004)	-0,291 (0,004)

b

Annotation	Effect size τ^* not including ARGweaver	Effect size τ^* including ARGweaver
ARGweaver	N.A	-0.147 (0.005)
LLD-AFR	-0.271 (0.004)	-0.241 (0.004)
Recombination rate	-0.370 (0.006)	-0.348 (0.006)
Nucleotide diversity	-0.051 (0.005)	-0.039 (0.005)
B-statistic	0.090 (0.004)	0.083 (0.004)
CpG content	0.321 (0.006)	0.314 (0.006)
ASMC_{avg}	-0.294 (0.004)	-0.249 (0.004)

Supplementary Table 10. S-LDSC analysis of predicted allele age annotations. Numerical values from Fig. 6b. Effect size τ^* estimates (meta-analysed across 63 independent diseases and traits) of CoalNN MAF-adjusted allele age annotation on all 26 populations (in **a**) and of ARGweaver MAF-adjusted allele age annotation (in **b**), in marginal S-LDSC analysis conditioned on 96 baseline annotations (the full Baseline-LD model except for ARGweaver). We also report effect sizes of Baseline-LD evolutionary annotations (level of LD measured in African populations LLD-AFR, recombination rate, nucleotide diversity, B-statistic [10], CpG content [11] and average pairwise TMRCA ASMC_{avg} [12]), before and after the introduction of the allele age annotation.

Supplementary References

[1] Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations* (2014).

[2] Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T. (eds.) *Computer Vision – ECCV 2014*, 818–833 (Springer International Publishing, Cham, 2014).

[3] Shrikumar, A., Greenside, P., Shcherbina, A. & Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713* (2016).

[4] Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics* **49**, 303–309 (2017).

[5] Spence, J. P. & Song, Y. S. Inference and analysis of population-specific fine-scale recombination maps across 26 diverse human populations. *Science Advances* **5**, eaaw9206 (2019).

[6] Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–249 (2010).

[7] Sim, N.-L. *et al.* Sift web server: predicting effects of amino acid substitutions on proteins. *Nucleic acids research* **40**, W452–W457 (2012).

[8] Gazal, S. *et al.* Combining snp-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity. *medRxiv* (2021).

[9] Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome research* **18**, 1814–1828 (2008).

[10] McVicker, G., Gordon, D., Davis, C. & Green, P. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics* **5**, e1000471 (2009).

[11] Zhang, L., Dai, Z., Yu, J. & Xiao, M. Cpg-island-based annotation and analysis of human housekeeping genes. *Briefings in bioinformatics* **22**, 515–525 (2021).

[12] Palamara, P. F., Terhorst, J., S. Song, Y. & L. Price, A. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics* **50** (2018).

Statement of Authorship

Title of Paper	Inference of genealogical relatedness and variant age using deep neural networks identifies signatures of selection
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and unsubmitted work written in a manuscript style
Publication Details	This chapter is a preprint paper and will soon be submitted for peer review. Juba Nait Saada, Zoi Tsangalidou, Miriam Stricker, and Pier Palamara. “Inference of genealogical relatedness and variant age using deep neural networks identifies signatures of selection”.

Student Confirmation

Student Name:	Juba Nait Saada		
Contribution to the Paper	Juba Nait Saada (J.N.S.) developed and implemented the model architecture and applied it to simulated and real data. J.N.S and Zoi Tsangalidou (Z.T.) performed analyses of interpretability and natural selection. J.N.S., Z.T. and Miriam Stricker (M.S.) performed analyses of heritability. Pier Palamara (P.P.) and J.N.S. planned the project. P.P. supervised the project. J.N.S., Z.T. and P.P analyzed results and wrote the manuscript.		
Signature Juba N.S.		Date	21/05/2022

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof. Pier Palamara		
Supervisor comments: Juba had a leading role in this project, as detailed above.		
Signature Pier Francesco Palamara	Date	26/06/2022

Chapter 5

Conclusion

5.1 Summary

Large volumes of genetic data are now available and are expected to grow at an unprecedented rate. By accurately and efficiently measuring genetic relatedness between individuals, these data sets can be used to gain a better understanding of demographic history, natural selection and disease. In this thesis, we presented previous approaches for genetic relatedness inference. We also developed novel methods and applied them to large data sets to investigate some applications of genetic relatedness in modern human populations.

In Chapter 1, we provided relevant background for understanding the contributions made in the remainder of the thesis. This includes a brief overview of the field of statistical genetics, an introduction to population genetics and coalescent theory, and

a review of deep learning.

In Chapter 2, we developed FastSMC, an algorithm for Identity-By-Descent (IBD) detection that scales well in analyses of very large biobank data sets and achieves better performance than existing methods[1]. FastSMC provides age estimates of the time to most recent common ancestor for IBD individuals. We applied FastSMC to 487,409 UK Biobank[2] samples and detected around 214 billion IBD segments transmitted by shared ancestors within the past 1,500 years. We thus obtained a fine-grained picture of genetic relatedness in the UK and a high-resolution insight into the actions of recent positive natural selection. Lastly, we devised an IBD-based test for association between phonotype and loss-of-function variations by using IBD sharing between exome sequenced and non-sequenced samples. This allowed us to successfully replicate known associations with 7 blood-related traits[3] and to identify novel gene-based associations.

Then, in Chapter 3, we introduced CoalNN, a likelihood-free approach which uses deep learning to predict pairwise times to the most recent common ancestors, recombination breakpoints, and allele ages[4]. By leveraging convolutional neural networks, CoalNN improves upon available state-of-the-art methods in TMRCA and allele age prediction in a variety of simulated conditions. These results demonstrate that deep learning approaches can be effectively applied to explicit inference of genealogical relationships.

Finally in Chapter 4, we applied our deep learning algorithm described in Chapter 3 to $2,504$ samples from 26 different populations in the 1,000 Genome Project data set[5]. We used CoalNN to date approximately 80 million variants in the human genome. We obtained evidence for the Out-of-Africa bottleneck theory for the origin of anatomically modern humans[6,7] and for selection against deleterious mutations[8,9]. Using the predicted allele ages, we built a genome-wide annotation capturing the efffects of past natural selection, detecting significant effects on trait and disease heritability. Overall, these results pave the way for the development and application of deep neural networks for analyses involving genealogical relationships in large genomic data sets.

5.2 Future work

Both CoalNN and FastSMC share common limitations and caveats with existing methods for pairwise coalescence time inference and IBD detection. First, as with most algorithms based on the coalescent process with recombination, both require the availability of phased data. We note, however, that when very large cohorts are available, long-range computational phasing results in high-quality haplotype estimation with switch errors rates as low as one every several tens of centimorgans[10,11]. An important direction of future work will be to extend our methods to support unphased data as well as other data modalities, such as low coverage sequencing data or ancient DNA.

Studying genetic information from ancient samples or extinct species is a difficult task due to post-mortem DNA damage and degradation[12]. Since deep learning allows to perform likelihood-free inference, CoalNN could be leveraged to overcome the obstacle of intractable analytic calculations encountered by most coalescent-based methods. Similarly, CoalNN could be extended to biological processes not currently modeled by coalescent HMMs such as gene conversion (which results in homologous haplotypes with identical sequences).

Second, both of our methods necessitate additional information for accurate predictions. In the case of FastSMC, the demographic model and the allele frequencies are required in order to accurately estimate the age of IBD segments, while CoalNN needs additional training through transfer learning to generalise to various evolutionary scenarios. Although we verified the robustness of our methods, failure to accurately provide such prior knowledge would result in biased estimates. A desirable area of future improvement is to dispose of this limitation by creating a generalisable model, e.g. by automatically learning the evolutionary parameters from the observed data directly.

Finally, our inference of genetic relatedness has been limited to the inference of pairwise genetic relations by only considering two haplotypes at a time. It would be interesting to explore extensions of our approaches that simultaneously consider multiple samples, with applications that include haplotype phasing or imputation. Additionally,

jointly inferring coalescence times in a large data set would allow to reconstruct the network of shared ancestry among individuals which can be represented using Ancestral Recombination Graphs[13,14,15]. Novel deep learning architectures and algorithms such as exchangeable neural networks[16,17,18] or Graph Neural Networks[19] can be extended to support joint inference of genetic relatedness.

Chapter 5 - References

[1] Juba Nait Saada, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev, and Pier Francesco Palamara. Identity-by-descent detection across 487,409 british samples reveals fine scale population structure and ultra-rare variant associations. *Nature communications*, 11(1):1–15, 2020.

[2] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203, 2018. doi: https://doi.org/10.1038/s41586-018-0579-z.

[3] Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua D Hoffman, Daren Liu, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Bin Ye, Nilanjana Banerjee, et al. Exome sequencing and characterization of 49,960 individuals in the uk biobank. *Nature*, 586(7831):749–756, 2020.

[4] Juba Nait Saada, Anthony Hu, and Pier Palamara. Inference of pairwise coalescence times and allele ages using deep neural networks. 2021. URL `https://www.lmrl.org/papers`.

[5] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[6] William Amos and Joe I Hoffman. Evidence that two main bottleneck events shaped modern human genetic diversity. *Proceedings of the Royal Society B: Biological Sciences*, 277(1678):131–137, 2010.

[7] Mark Lipson and David Reich. A working model of the deep relationships of diverse modern human genetic lineages outside of Africa. *Molecular biology and evolution*, 34(4):889–902, 2017.

[8] Adam Kiezun, Sara L Pulit, Laurent C Francioli, Freerk van Dijk, Morris Swertz, Dorret I Boomsma, Cornelia M van Duijn, P Eline Slagboom, GJB van Ommen,

Cisca Wijmenga, et al. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency. *PLoS genetics*, 9(2):e1003301, 2013.

[9] Patrick K Albers and Gil McVean. Dating genomic variants and shared ancestry in population-scale sequencing data. *PLoS biology*, 18(1):e3000586, 2020.

[10] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.

[11] Po-Ru Loh, Pier Francesco Palamara, and Alkes L Price. Fast and accurate long-range phasing in a uk biobank cohort. *Nature genetics*, 48(7):811–816, 2016.

[12] Ludovic Orlando, Robin Allaby, Pontus Skoglund, Clio Der Sarkissian, Philipp W Stockhammer, María C Ávila-Arcos, Qiaomei Fu, Johannes Krause, Eske Willerslev, Anne C Stone, et al. Ancient dna analysis. *Nature Reviews Methods Primers*, 1(1):1–26, 2021.

[13] Richard R Hudson et al. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

[14] Robert C Griffiths and Paul Marjoram. Ancestral inference from samples of dna sequences with recombination. *Journal of Computational biology*, 3(4):479–502, 1996.

[15] Robert C Griffiths and Paul Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

[16] Benjamin Bloem-Reddy and Yee Whye Teh. Neural network models of exchangeable sequences. In *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[17] Jeffrey Chan, Valerio Perrone, Jeffrey Spence, Paul Jenkins, Sara Mathieson, and Yun Song. A likelihood-free inference framework for population genetic data using exchangeable neural networks. *Advances in neural information processing systems*, 31, 2018.

[18] Yuki Saito, Takuma Nakamura, Hirotaka Hachiya, and Kenji Fukumizu. Exchangeable deep neural networks for set-to-set matching and learning. In *European Conference on Computer Vision*, pages 626–646. Springer, 2020.

[19] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020.