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).

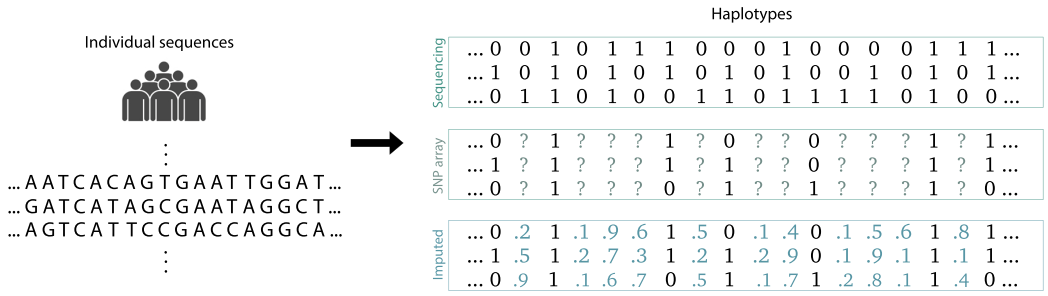


Figure 1: **Input types.** CoalNN can be used to analyze phased sequencing, SNP array, or imputed genotype data. The input haplotypes (i.e. sets of genomic variants located on the same chromosome) are obtained from individual DNA sequences consisting of nucleotides (A, C, G or T). Observations have value 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.

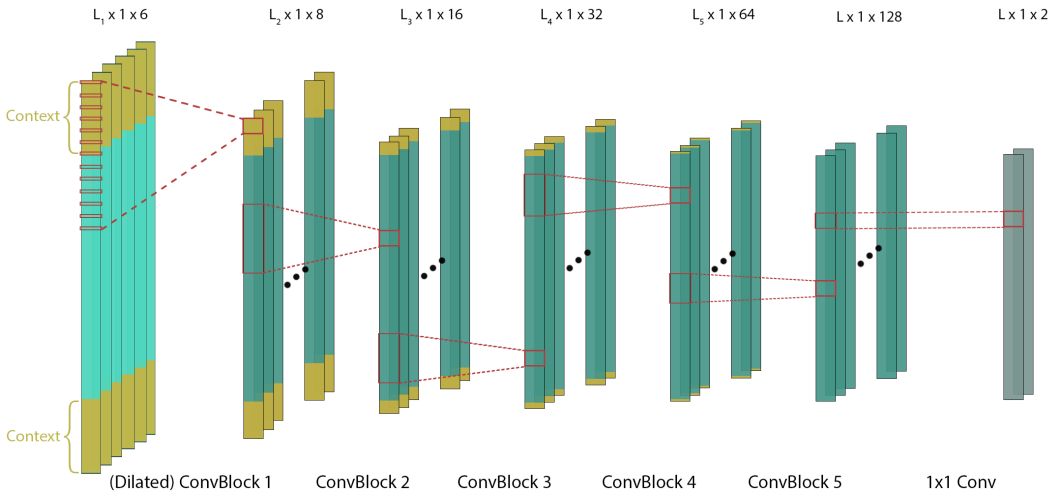


Figure 2: **CoalNN architecture.** CoalNN comprises a batch normalisation layer followed by five convolution blocks (convolution layer + batch normalisation + ReLU) and a 1×1 convolution layer. The first ConvBlock uses dilated convolution with a large kernel size to capture long-range dependencies, while 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

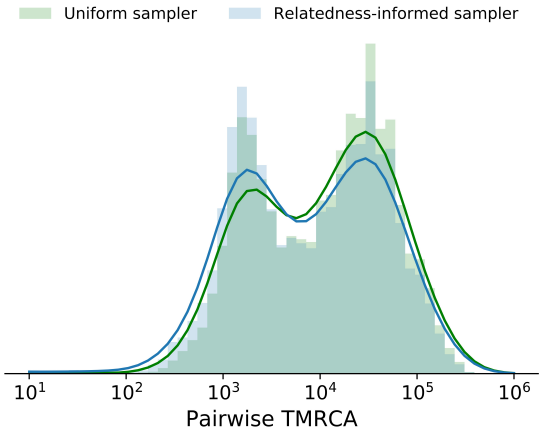


Figure 3: **Pairwise TMRCA sampling distribution.** Distributions of pairwise TMRCA in generations (histograms) obtained from a uniform sampler (in green) and a relatedness-informed sampler (in blue). Solid curves show kernel density estimations using Gaussian kernels (bandwidth $= 0.25$).

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

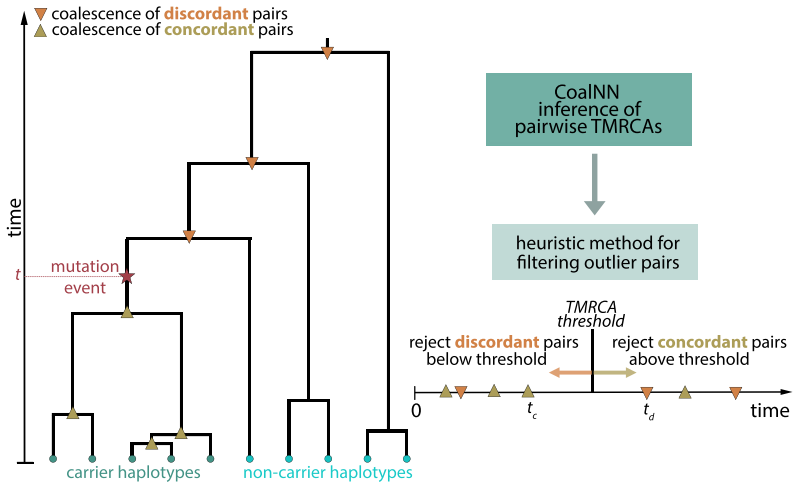


Figure 4: **Allele age inference.** At a given genomic site, individuals are connected through the sharing of underlying genealogical relationships (left). We aim to infer the time t at which a mutation arose (in red) and resulted in carrier haplotypes (in green) and non-carrier haplotypes (in cyan). 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.

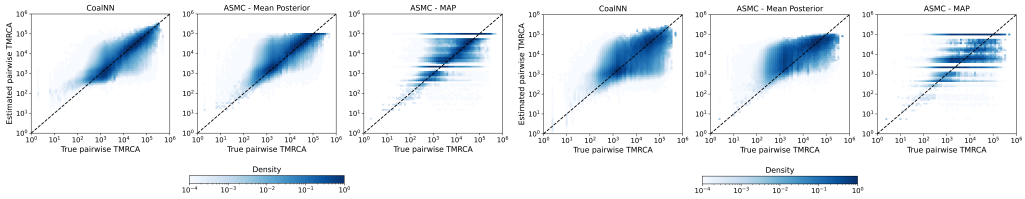
We also applied a log-transformation to the TMRCA ground truth. By predicting the log-TMRCA instead of the TMRCA itself, the training procedure penalizes the ratio between the ground truth and the prediction rather than the distance, preventing the loss from being dominated by large TMRCA values. Finally, 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

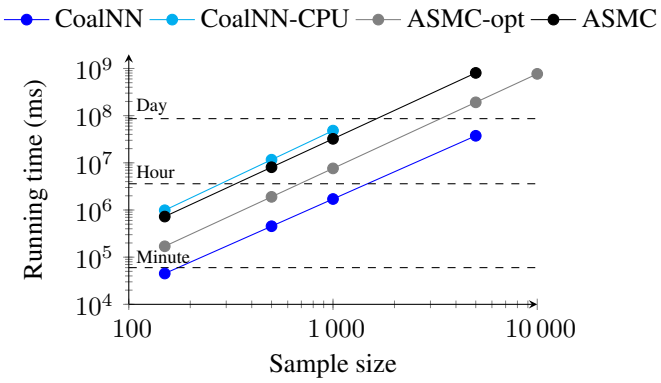


Figure 6: **Running time evaluation.** Running time (in milliseconds) of CoalNN (on a single A100 GPU card and a single CPU in blue, and on a single CPU only in cyan) and ASMC (on a single CPU, optimized version in grey and non-optimized version in black) on array data using the first 30 Mbp of chromosome 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

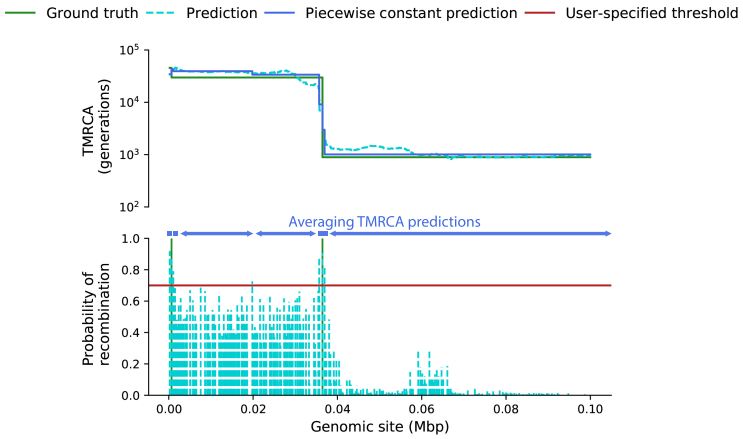
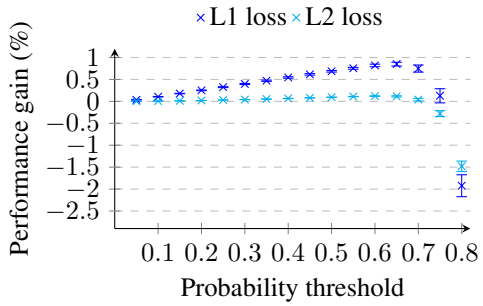
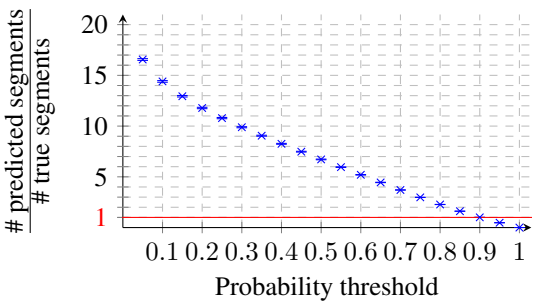


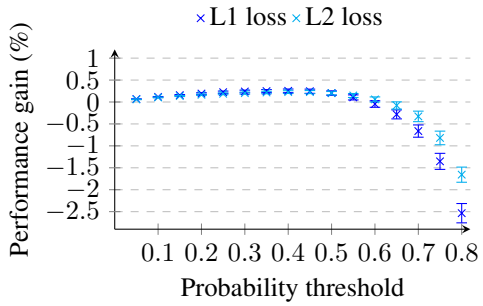
Figure 7: **Computing the piecewise constant coalescence times.** The inferred TMRCAs are averaged between every consecutive genomic site with an estimated probability of recombination breakpoint that exceeds a user-specified threshold.



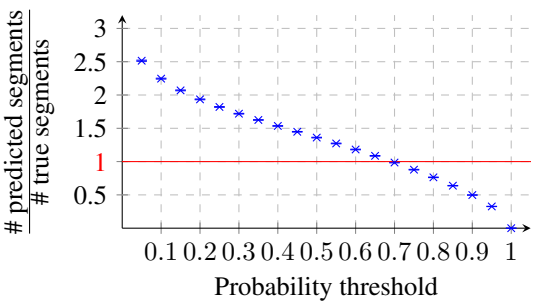
(a) Percent performance gain on sequencing data.



(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.

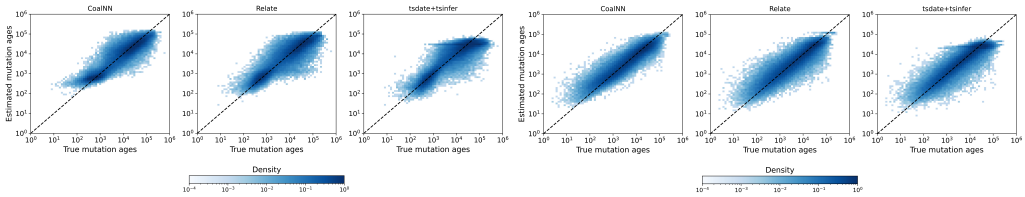
Performance on allele age prediction. We inferred mutation ages under a European demographic model and a constant population size ($N_e = 10,000$) using CoalNN, as well as Relate [29] and tsinfer+tsdate [42, 31] (tsinfer and tsdate run successively), two recent genealogical inference approaches that 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.