National Research University Higher School of Economics

As a manuscript

Vladimir L. Shchur

Mathematical models and data analysis in population genomics

DISSERTATION SUMMARY

for the purpose of obtaining academic degree
Doctor of Science in Applied Mathematics

Moscow – 2023

The dissertation work was carried out at the International laboratory of statistical and computational genomics at the National Research University Higher School of Economics.

1 Introduction

1.1 Relevance

Genomics is a new interdisciplinary scientific field that emerged at the intersection of genetics, mathematics, and computer science. Population and evolutionary genetics is one of the important sections in this field. Due to the rapid decrease in the costs of sequencing and genotyping technologies, more and more genetic data are available for analysis, providing information about the processes of population development. Genomes contain information about the history and structure of populations, evolutionary factors and mechanisms of natural selection. For example, over the past 15 years, a lot of new insights has been obtained from genomic data about the history of peopling of Earth, the admixture of ancient humans (Neanderthals and Denisovans) with the ancestors of modern humans, their adaptation to different climatic conditions and geographic territories. On the other hand, the SARS-CoV-2 coronavirus pandemic has shown the importance of real-time genomic epidemiological surveillance. By November 2022, about 13.5 million samples of the coronavirus have been already available in the GISAID database. This data makes it possible to trace transmission paths, detect new variants of the virus, and study its evolution. Thus, the development of new mathematical models and methods for analyzing genetic data is an important and timely problem.

The dissertation presents theoretical results in population and evolutionary genetics, new mathematical models, methods of genetic data analysis, as well as the results of the analysis of experimental data. The results of the work extend the arsenal of methods for research in the field of population and evolutionary genetics, allowing to clarify a more detailed and accurate picture of the history of population development, to obtain new knowledge about the evolutionary processes and adaptation of different species of animals and viruses. The developed models and methods make it possible to use genetic data for estimation of such fundamental parameters as migration rates, proportions of admixture (for single, pulse, migrations), time of separation and admixture of populations, and the strength of natural selection. New algorithms and software meet the requirements of modern and future genomics problems that require processing of large datasets. For example, our new method `MiSTI` [18*][1] allows simultaneous estimation of population split

[1] Here an asterisk denotes publications from the list submitted by the applicant for

times and migration rates based on estimates of historical effective population size obtained, for example, by PSMC [1]. Using our method, we challenged the result [2] of a deep separation (260-350 thousand years ago) between African Bushmen and Dinka populations, obtaining a new estimate of ≈ 107 thousand years ago and one-way migration from Dinka to Bushmen, that is about $2.5 - 3.5$ times smaller than previously reported. Our results are also confirmed by our computer simulations.

1.2 State-of-the-art

The theoretical and methodological results presented in this work allow us to obtain new knowledge in the field of population and evolutionary genetics and genomic epidemiology, which could not be obtained by previously existing methods, as well as lay the foundation for the development of new, even more accurate and effective methods.

The following *theoretical results* have been obtained: the number of p-cousins in a large sample from a diecious Wright-Fisher population was studied, the distribution of chromosome tract lengths under adaptive introgression was studied, the accuracy of the SMC' approximation of structured coalescence with recombination was investigated, the concept of local effective population size was formalized and studied, a three-locus admixture linkage disequilibtium model was constructed for two populations, and a quantitative quasiisometric hyperbolic space problem was studied.

Based on these theoretical results we *developed methods* that provide new possibilities for the analysis of real genetic data. Thus, a method for inferring adaptive introgression allows to study one of the most important population adaptation mechanisms. The method for calculating the local effective population size and estimation of migration rates and split times and the method for estimation of multiple admixture times allow us to clarify the history of population development, in particular, to study migration processes on different time scales. The method for predicting the historical population size using deep learning has an important methodological significance, opening up the possibility to use deep learning for genome-wide analysis in the future. This is especially relevant for tasks in which probabilistic methods are computationally intractable. A new method for computer simulation of viral

defense listed in section 1.9.

genealogies allows generating datasets of sizes equal to and larger than the existing experimental datasets of SARS-CoV-2 coronavirus genomes. The developed software is required for validation of existing and new methods in the field of genomic epidemiology. Variational autocoders with Euclidean and hyperbolic latent spaces are proposed to be used as a method for clustering and visualizing data for population analysis.

Real data analysis was performed: we studied natural selection in Chileans after admixture of indigenous, European and African populations; we estimated the time of separation of human populations, challenged the result about deep time of separation of Bushmen and Dinka populations; we studied coronavirus philodynamics in Vreden Research Institute (March-April 2020), AY.122+ORF7a:P45L delta clades in Moscow (April-September 2021).

1.3 Aims and tasks of the study

The aim of the study is to develop new mathematical models, methods, algorithms and software to study population and evolutionary history from genomic sequences, namely to study the processes of population separation and admixture, determine the adaptive loci of the genome and the strength of natural selection, and estimate changes in the historical size of the population.

The objectives of the study are:

- Study the number of individuals with p-cousins in a large sample from a population.

- Develop a mathematical model of adaptive introgression to accurately and efficiently calculate the distribution of chromosome tract lengths.

- Develop a method for inferring adaptive introgression based on the developed mathematical model.

- Develop a viral genealogy simulator scaling to realistic sample sizes collected during the SARS-CoV-2 pandemic.

- Investigate the accuracy of the approximation of structured coalescence with recombination by the SMC' model.

- Separate the effects of genetic drift (local effective population size) and migration, develop a method to calculate the local effective population size from the historical effective population size for the two populations, and estimate migration rates and separation times.
- Develop a mathematical theory of admixture linkage disequilibrium of the three loci, and develop a method based on this theory to estimate the timings of multiple admixture.
- Study natural selection in Chilean population following post-Columbian admixture.
- Develop a method for predicting changes in effective population size using deep learning.
- Study the phylodynamics of SARS-CoV-2 coronavirus in Russia.
- Develop a theory and apply hyperbolic geometry to analyze genetic data in population genetics.

1.4 Research methods

Research methods include the use and development of population models (Wright-Fisher model and its generalizations, coalescent model, sequential Markov coalescence, compartmental epidemiological models), probabilistic approaches, Hidden Markov Model, Gillespie algorithm (including approximate τ-leaping algorithm), deep learning and geometric data analysis methods. The software is implemented in Python and C/C++ (including cython technology). Existing methods of population and evolutionary genomics were also used: PSMC, Admixture, BEAST2. For computer calculations, a high-performance cluster of the National Research University Higher School of Economics was used.

1.5 Theoretical and practical significance

The theoretical significance lies in the development of mathematical theory in population and evolutionary genetics; in particular, new results were obtained for the Wright-Fisher and coalescent models. Several data analysis methods and algorithms have also been developed using these new theoretical results. Experimental data were analyzed using new and existing methods, in particular new knowledge was gained about the history of the human population, the spread of the SARS-CoV-2 coronavirus in Russia. The practical significance lies in the development of software that implements new methods and algorithms for genetic data analysis and computer modeling of populations. All the developed software packages are publicly available in the corresponding GitHub repositories.

1.6 Provisions for the defense

- *On the number of p-cousins in a large sample from a population [13*].* An asymptotic formula for the expectation of the fraction of individuals in a sample of size K from a population of size N without p-cousins in that sample is derived for the limit of $N \to \infty$ and $K/N = const$. The formulas were obtained for monogamous and for non-monogamous Wright-Fisher diecious models. It is shown that for large samples, whose size is comparable to some large-scale studies in genetics, close relatedness cannot be neglected. The result is important when planning, for example, GWAS (genome-wide association search) projects with large cohorts.

- *Mathematical model of adaptive introgression [10*].* A mathematical model of adaptive introgression has been developed. The allele frequency trajectory under natural selection is modeled using a deterministic logistic curve. The model is computationally efficient while being accurate over a wide range of adaptive introgression parameters. It is also shown that this range can be extended to cases where the logistic approximation is inaccurate due to genetic drift by numerically estimating the average trajectory of the adaptive allele. The model allowed the development of two methods (a method for calculating the distribution of tract lengths under adaptive introgression and a method for inferring adaptive introgression), which in turn opens up new possibilities for studying adaptation in various animal species, including humans.

- *Viral genealogy simulator [2*].* A software package `VGsim` for modeling epidemics and the resulting viral genealogies has been developed.

The functionality of the software package includes simulation of epidemiological, evolutionary and population complexities. The simulation of epidemic development is based on the Gillespie algorithm, the simulation of genealogies is based on structured coalescence driven by epidemiological dynamics. The software package is the fastest genomic epidemiology solution we know of. It allows us to simulate genealogies of tens of millions of samples under complex epidemiological scenarios, which exceeds the current size of the GISAID database. This makes `VGsim` a promising solution for validating data analysis results and new data analysis methods in genomic epidemiology.

- *Accuracy of the SMC' approximation of structured coalescent [1*]* The accuracy of the SMC' approximation of coalescence with recombination in the case of two populations with migration was investigated. We analyzed the total variation in the difference between the joint distributions of times to a common ancestor of two loci in the coalescence with recombination and SMC' models as a function of the genetic distance between these loci. It is shown that for two populations with migration, the total variation decreases significantly slower than in the case of a homogeneous population. This shows that in the presence of population structure, data analysis methods based on the SMC' model may lead to inaccurate results.

- *Effective population size and migration [18*].* The notion of local effective population size for the scenario with two populations and migration was formalized. The effect of migration on population size estimation by PSMC method was studied. Based on this mathematical theory, we developed a method for calculating the local effective population size and for estimating the time of population separation and migration rates between them. The work has important methodological significance for the theory of structured coalescence, and also allows us to accurately reconstruct the history of gene flow between populations.

- *Multiple admixture and three loci linkage disequilibrium [8*].* The mathematical theory of three genetic loci linkage disequilibrium in admixed population was developed. Based on this theory, a method and software were developed to estimate the timings of admixture between two populations and two pulses of migration. The developed method makes it possible to accurately investigate the recent (within several tens of generations) history of population admixture in complex scenarios for which previously existing methods were inapplicable or inaccurate.

- *Selection in Chilean population due to post-Columbian admixture [11*].* Using computer simulations, the results of the scan for natural selection after admixture of indigenous, European, and African populations in Chile based on the prediction of local ancestry are verified. The consistence of the chosen statistics for the selection genome-wide scan was confirmed. This supported the reliability of the study of adaptation processes in the modern Chilean population.

- *Deep learning for demographic analysis [4*, 3*].* A method based on deep learning was developed to predict local times to the last common ancestor along the diploid genome. The method can also be used to infer the trajectory of effective population size similar to the PSMC method PSMC [1]. The work has important methodological significance for the further development of deep learning methods for the analysis of whole genomes.

- *Phylodynamics of the SARS-CoV-2 coronavirus in Russia [9*, 7*].* A Bayesian phylodynamic analysis of the Covid-19 outbreak in the Vreden Hospital (St. Petersburg) in March-April 2020, as well as the (AY.122+ORF7a:P45L) clade of the delta variant in April-September 2021 in Moscow using software package BEAST2. The first study shows that the nosocomial outbreak was the result of at least two, probably three, introductions of the coronavirus into the hospital. The second study confirmed independently of the epidemiological data that the main clade (AY.122+ORF7a:P45L) was responsible for the summer epidemic wave in 2021, and probably for the subsequent fall wave. The results provide an objective picture of SARS-CoV-2 coronavirus spread in Russia, which is important in the analysis of epidemiological measures for pandemic control.

- *Hyperbolic geometry and genetic data analysis [14*-17*].* The numerical problem of quasiisometric hyperbolic spaces is stated and studied. The application of variational autocoders with hyperbolic latent space to the problem of population genetic diversity visualization (similar to the principal component method) was considered. These results have both fundamental mathematical significance and open up the possibility to develop and apply novel approaches in population genetics.

1.7 Novelty and reliability

All of the scientific results presented for the defense are novel. New mathematical model for the distribution of chromosome lengths during adaptive introgression has been proposed. New methods for inferring adaptive introgression, computer modeling of viral genealogies, estimation of split times and migration rates between populations, estimation of timings of multiple admixture from linkage disequilibrium of three loci, and estimation of historical effective population size using deep learning were proposed. Using these and existing methods, the following problems were solved: split times between human populations were estimated and the result of deep split time between African populations of San and Dinka was challenged, admixture times for modern populations of Mexicans and Colombians were estimated, adaptation in Chileans after post-Columbian admixture was studied, philodynamics of coronavirus SARS-CoV-2 in Russia was studied.

The reliability of the results is justified by the fact that all the results presented for the defense, were published in leading peer-reviewed scientific journals indexed in the scientific databases Web of Sciences and Scopus with quartiles Q1 - 13 articles, Q3 - 2 articles, of which 3 articles were published in journals from the Nature Index. Software code are published in open GitHub repositories as 7 software packages.

1.8 Approbation of the obtained results

The main results of the dissertation work were reported at the following international conferences and seminars:

- *Estimating the timing of multiple admixture events using 3-locus Linkage Disequilibrium*, conference Moscow Conference on Computational Molecular Biology (MCCMB'21), July 2021, Moscow, Russia.

- *Deep learning for demographic inference from whole-genome sequences*, conference Moscow Conference on Computational Molecular Biology (MCCMB'21), July 2021, Moscow, Russia.

- *ngsPSMC: genotype likelihood-based PSMC for analysis of low coverage NGS data*, conference Probabilistic Modeling in Genomics, October 2019, Aussois, France.

- *ngsPSMC: genotype likelihood-based PSMC for analysis of low coverage NGS data*, conference Moscow Conference on Computational Molecular Biology (MCCMB'19), July 2019, Moscow, Russia.

- *Estimation of population split times and migration rates with variable population sizes*, conference Probabilistic Modeling In Genomics, October 2018, Cold Spring, USA.

- *ngsPSMC: modifying PSMC to work with NGS data"*, UCCGC workshop, 15–18 August 2017, Blue Oak Ranch Reserve, USA.

- *Tree consistent PBWTs and their application to reconstructing ancestral recombination graphs and demographic inference*, conference Probabilistic Modeling in Genomics, October 12–17 2015, Cold Spring, USA.

- *Tree consistent PBWT and their application to reconstructing Ancestral Recombination Graphs and demographic inference*, Recomb 2015, Warsaw, Poland. Best poster award.

- *On modern problems and methods for data analysis in human genomics*, Computer Simulation in Physics and beyond 2015, Moscow, Russia, plenary talk

- *Tree consistent PBWT and their application to reconstructing ancestral recombination graphs and population structure inference*, Biology of Genomes, 10–14 May 2015, Cold Spring, USA

- *Extension of PBWT and its connection with ARG*, conference International meeting on genomics, April 2014, Doha, Qatar.

1.9 List of papers on the topic of the dissertation work presented for the defense (with the personal contribution of the candidate)

Author's publications in peer-reviewed scientific journals listed in the international citation system Scopus

1.* Shchur V. *Accuracy of the SMC' approximation of structured coalescent* Lobachevskii journal of mathematics **43(12)** (2022), pp. 3626–3630

   The accuracy of the SMC' approximates of coalescence with recombination for the case of two populations with migration was evaluated. It is shown that the total variation between the joint distribution of time to the last common ancestor in two loci decreases significantly slower with increasing genetic distance between loci than in the case of a homogeneous population.

2.* Shchur V., Spirin V., Burovski E., De Maio N., Corbett-Detig R. *VGsim: scalable viral genealogy simulator for global pandemic* // PLoS Computational Biology. **18(8)** (2022), e1010409.

   `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010409`

   A viral genealogy simulator `VGsim` has been developed, which is the fastest solution in its field to the best of our knowledge. Mathematical models and algorithms were developed, the core part of the software was implemented, and all the other stages of research and development were supervised.

3.* Arzymatov K., Khomutov E., Shchur V. *Deep learning for inferring distribution of time to the last common ancestor from a diploid genome* // Lobachevskii Journal of Mathematics **43(8)** (2022) pp. 2092–2098.

   `https://doi.org/10.1134/S1995080222110075`

   A deep learning based method for predicting local times to the last common ancestor along the genome, as well as their marginal probability distribution, was proposed and evaluated.

4.* Khomutov E., Arzymatov K., Shchur V. *Deep learning based methods for estimating distribution of coalescence rates from genome-wide data* // Journal of Physics: Conference Series **1740** (2021). 012031.

`https://iopscience.iop.org/article/10.1088/1742-6596/1740/1/012031`

A prototype method for predicting local times to the last common ancestor along the genome is proposed.

5.* Jin Y., Brandt D. Y., Li J., Wo Y., Tong H., Shchur V. *Elevation as a selective force on mitochondrial respiratory chain complexes of the Phrynocephalus lizards in the Tibetan plateau* // Current Zoology **67(2)** (2021), pp. 191–199.

`https://academic.oup.com/cz/article/67/2/191/5909995`

A permutation analysis was performed to study parallel altitude adaptation in Phrynocephalus lizards in the Tibetan Plateau.

6.* Svedberg J., Shchur V., Reinman S., Nielsen R., Corbett-Detig R. *Inferring Adaptive Introgression Using Hidden Markov Models* // Molecular Biology and Evolution **38(5)** (2021), pp. 2152–2165.

`https://academic.oup.com/mbe/article/38/5/2152/6120794`

A Hidden Markov Model for adaptive introgression was developed. An approach for approximate fast computation of transition probabilities near the adaptive locus was proposed.

7.* Klink G. V., Safina K. R., Nabieva E., Shvyrev N., Garushyants S., Alekseeva E., Komissarov A. B., Danilenko D. M., Pochtovyi A. A., Divisenko E. V., Vasilchenko L. A., Shidlovskaya E. V., Kuznetsova N. A., Speranskaya A. S., Samoilov A. E., Neverov A. D., Popova A. V., Fedonin G. G., Akimkin V. G., Lioznov D., Gushchin V. A., Shchur V., Bazykin G. A. *The rise and spread of the SARS-CoV-2 AY.122 lineage in Russia* // Virus Evolution **8** (2022), pp. 1–11.

`https://academic.oup.com/ve/article/8/1/veac017/6542789`

A phylodynamic analysis of the Y.122ORF7a:P45L coronavirus clade in Moscow in April-September 2021 was performed.

8.* Liang M., Shishkin M., Mikhailova A., Shchur V., Nielsen R. *Estimating the timing of multiple admixture events using 3-locus Linkage Disequilibrium* // PLOS Genetics **18(7)** (2022), e1010281.

`https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1010281`

A mathematical model of the three loci linkage disequilibrium was developed for two populations.

9.* Komissarov A. B., Safina K. R., Garushyants S. K., Fadeev A. V., Sergeeva M. V., Ivanova A. A., Danilenko D. M., Lioznov D., Shneider O. V., Shvyrev N., Spirin V., Glyzin D., Shchur V., Bazykin G. A. *Genomic epidemiology of the early stages of the SARS-CoV-2 outbreak in Russia* // Nature Communications **12** (2021), pp. 1–13.

`https://www.nature.com/articles/s41467-020-20880-z`

A phylodynamic analysis of a nosocomial SARS-CoV-2 outbreak at the Vreden Hospital in March-April 2020 was performed.

10.* Shchur V., Svedberg J., Medina P., Corbett-Detig R., Nielsen R. *On the Distribution of Tract Lengths During Adaptive Introgression* // G3: Genes, Genomes, Genetics **10(10)** (2020), pp. 3663–3673.

`https://academic.oup.com/g3journal/article/10/10/3663/6053540`

We constructed a mathematical model for introgressed genome tracts under adaptive introgression based on coalescent theory and an approximation of the selected allele frequency trajectory with a deterministic logistic curve.

11.* Vicuña L., Klimenkova O., Norambuena T., Martinez F. I., Fernandez M. I., Shchur V., Eyheramendy S. *Post-Admixture Selection on Chileans Targets Haplotype Involved in Pigmentation, Thermogenesis and Immune Defense Against Pathogens* // Genome Biology and Evolution **12(8)** (2020), pp. 1459–1470.

`https://academic.oup.com/gbe/article/12/8/1459/5866553`

The choice of a statistics for LAI selection scan in the admixed Chilean population was verified using simulations.

12.* Skov L., Hui R., Shchur V., Hobolth A., Scally A., Schierup M. H., Durbin R. *Detecting archaic introgression using an unadmixed outgroup* // PLoS Genetics **14** (2018), pp. 1–15.

`https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007641`

A Hidden Markov Model architecture was proposed to detect segments of archaic origin using a non-admixed outgroup population. This architecture made it possible to detect Neanderthal and Denisovian DNA segments in individuals from Papua New Guinea with high accuracy.

13.* Shchur V., Nielsen R. *On the number of siblings and p-th cousins in a large population sample* // Journal of Mathematical Biology **77(5)** (2018), pp. 1279–1298.

`https://link.springer.com/article/10.1007/s00285-018-1252-8`

We derive formulas for the mathematical expectation of the number of individuals without p-cousing in a sample of population in monogamous and non-monogamous diecious Wright-Fisher models, and find asymptotic behavior as a function of the sample fraction with the population size N going to infinity.

14.* Gouezel S., Shchur V. *A corrected quantitative version of the Morse lemma* // Journal of Functional Analysis. **277(4)** (2019), pp. 1258-1268. `https://www.sciencedirect.com/science/article/pii/S0022123619300801`

The proof of the quantitative version of Morse's lemma about the distance from a quasi-geodesic to a geodesic segment in hyperbolic space was corrected.

15.* Shchur V. *On the quantitative quasi-isometry problem: Transport of Poincaré inequalities and different types of quasi-isometric distortion growth* // Journal of Functional Analysis. **269(10)** (2015), pp. 3147–3194.

`https://www.sciencedirect.com/science/article/pii/S0022123615003699`

Quantitative properties of quasi-isometries are investigated: the transport of Poincaré inequalities is considered, exact estimates of the growth of quasi-isometric distortion for some class of hyperbolic metric spaces are obtained. The linearity of quasi-isometric distortion growth between the hyperbolic space $\mathbb{H}^n$ and the binary tree is also proved.

16.* Shchur V. *A quantitative version of the Morse lemma and quasi-isometries fixing the ideal boundary* // Journal of Functional Analysis. **264(3)** (2013), pp. 815–836.

`https://www.sciencedirect.com/science/article/pii/S002212361200434X`

A quantitative version of Morse lemma, the dual anti-Morse lemma, are proven, and quasi-isometries fixing the ideal boundary are investigated.

Other publications and preprints:

17.* I. Bogdanov and V. Shchur, *Variational Autoencoders with Euclidean and Hyperbolic Latent Spaces for Population Genetics* // 2021 XVII International Symposium "Problems of Redundancy in Information and Control Systems" (REDUNDANCY), 2021, pp. 91–94

`https://ieeexplore.ieee.org/abstract/document/9606448`

A method for population clustering of genetic data based on variational autoencoders with Euclidean and hyperbolic latent spaces is proposed.

18.* Preprint Shchur V., Brandt D. Y., Ilina A., Nielsen R. Estimating population split times and migration rates from historical effective population sizes / Cold Spring Harbor Laboratory. Series 005140 "Biorxiv". 2022

`https://www.biorxiv.org/content/10.1101/2022.06.17.496540v1`

The concepts of historical and local effective population sizes are introduced. A method for split time and migration rates estimation was developed from the historical effective population sizes of individuals from two populations. This method simultaneously infers the local effective population size.

2 Results

In this section, we outline the main research results presented in the dissertation.

2.1 On the number of siblings and p-cousins in a large population sample

As genomic sequencing and genotyping techniques are becoming cheaper, the data sets analysed in genomic studies are becoming larger. With an increase in the proportion of individuals in the population sampled, we might also expect an increase in the proportion of related individuals in the sample. In Genome Wide Association Studies (GWAS), related individuals are routinely removed from the sample, but other strategies also exist for using relatedness as a covariate in the statistical analyses (e.g., [3]). These observations raise the following question: given a particular effective population size, how many close relatives would we expect to find in a sample? The answer to this question may help guide study designs and strategies for addressing relatedness in population samples and improve design for GWAS. Of particular interest is the number of individuals in the sample without relatives, i.e. the number of individuals remaining in the sample if individuals with relatives are removed.

In this section, we present results on the number of close relatives in two diecious Wright-Fisher models, the monogamous model and the random mating model from [13*]. We will use these models to obtain the distribution and expectation of the number of individuals with siblings, and expectation of the number of cousins, third cousins, etc. in the sample.

We use two generalizations of the Wright-Fisher model to model the diecious population. The first generalization is the monogamous Wright-Fisher model, in which pairs of parents are fixed. The second generalization is the random mating Wright-Fisher model, in which for each individual, each of the two parents is chosen independently from the sets of male and female individuals, respectively.

Further, we assume that each generation contains exactly N male and N female individuals. Denote by G_0 the observed generation, and the ancestral generations will be numbered backward in time, i.e. the generation G_i is parental to the generation G_{i-1}.

Consider a random sample S from generation G_0. Denote by U_T (for monogamous model) and V_T (for random mating model) the number of individuals in the sample S who have no $(T+1)$-cousins and $(T+1)$-semicousins in S and whose genealogy has no cycles (that is, there is no inbreeding in their genealogy). The probability of cycles is small if 2^T (the number of ancestors in the T-th generation without inbreeding) is much smaller than the effective population size. Thus, inbreeding can be neglected. We will denote $p = T + 1$.

Below we give formulas for the distributions and expectations of the numbers of U_1 sisters in the monogamous model and V_1 half-sisters in the non-monogamous model. We will also derive an asymptotic formula for the mathematical expectation of U_T and V_T when the sample fraction is fixed and the population size is large.

Recall that the Stirling number of the second kind $S(n, k)$ is equal to the number of partitions of a set of n elements into k non-empty subsets. The r-associated Stirling number of the second kind $S_r(n, k)$ [4] is the number of partitions of the set of n elements into k non-empty subsets of size at least r.

2.1.1 Results for the monogamous Wright-Fisher model

Consider individuals in a random sample S of size K such that the same sample S does not include their siblings. We are interested in the distribution of the number of such individuals, its expectation, and asymptotic behavior of the expectation for a large population size and a fixed sample S fraction of the total population (that is, for a fixed ratio K/N and a limit $N \to \infty$).

Theorem 1 *Let U_1 be a random variable, denoting the number of individuals in a random sample S of size K without siblings (0-cousins) in the same sample S, under monogamous Wright-Fisher model. Then*

- *the probability distribution of U_1 is given by*

$$\mathbb{P}(U_1 = u) = \frac{\binom{K}{u} \sum_{t=1}^{\lfloor \frac{K-u}{2} \rfloor} S_2(K-u, t) \binom{N}{u+t} (u+t)!}{\sum_{t=1}^{m} S(K, t) \binom{N}{t} t!};$$

- *expectation of U_1 is*

$$\mathbb{E}(U_1) = K(1 - 1/N)^{K-1};$$

- *moreover, if $K/N = \alpha$, then*

$$\lim_{K \to \infty} \frac{\mathbb{E}(U_1)}{K} = e^{-\alpha}.$$

For the number U_2 of individuals in a sample, without cousins we found the expectation and its asymptotic behavior.

Theorem 2 *Let U_2 be a random variable, denoting the number of individuals in a random sample S without cousins ($1-$cousins) in the same sample S, under monogamous Wright-Fisher model. Then the expectation of U_2 is equal to*

$$\frac{\mathbb{E}(U_2)}{K} = K\frac{\sum_{m=1}^{K} S(K,m)\binom{N}{m}m!N(N-1)(N-2)^{2m-2}}{\sum_{m=1}^{K} S(K,m)\binom{N}{m}m!N^{2m}}.$$

Moreover, if $K/N = \alpha$, then

$$\lim_{K\to\infty} \mathbb{E}(U_2) = e^{-4\alpha}.$$

Our results can be generalized to any degree of kinshop, that is for the number U_p of individuals in a sample S, without p-cousins in S.

Theorem 3

- *For any natural $p \geq 1$, the expectation of U_p is*

$$\mathbb{E}(U_p) = K\frac{\underbrace{\sum_{m_1=1}^{K} R_1 \sum_{m_2=2}^{2m_1} R_2 \dots \sum_{m_{p-1}=4}^{2m_{p-2}} R_{p-1}}_{(p-1)\ nested\ sums} N^{2m_{p-1}}W(p)}{\underbrace{\sum_{m_1=1}^{K} R_1 \sum_{m_2=2}^{2m_1} R_2 \dots \sum_{m_{p-1}=4}^{2m_{p-2}} R_{p-1}}_{(p-1)\ nested\ sums} N^{2m_{p-1}}}, \qquad (1)$$

where $2m_0 := K$,

$$Q_p(N,M) = \sum_{t=0}^{p} \binom{p}{t} S(N-p, M-t)\binom{M-t}{p-t}(k-t)!,$$

$$R(j) = Q_{2^{j-1}}(2m_{j-1}, m_j)\binom{N}{m_j}m_j!,$$

and

$$W(p) = \left(1-\frac{2^{p-1}}{N}\right)^{2m_{p-1}-2^{p-1}} \prod_{s=1}^{2^{p-1}} \left(1-\frac{s}{N}\right).$$

- *If $K/N = \alpha$ ($i = 1, 2, \dots, p$), then*

$$\lim_{K\to\infty} \frac{\mathbb{E}(U_p)}{K} = \lim_{K\to\infty} \left(1-\frac{2^{p-1}\alpha}{K}\right)^{2^{p-1}K} = e^{-(2^{2p-2})\alpha}. \qquad (2)$$

2.1.2 Results for random mating Wright-Fisher model

We obtained similar results for the case of the random mating Wright-Fisher model. However, unlike in the case of the monogamous model, the probability that two individuals are full p-cousins is small compared to the probability of being p-semicousins. So, in this case we will be interested in the number V_p of individuals in the sample S whose p-semicousins and p-cousins are not in the sample.

Theorem 4 - *For any natural $p \geq 1$, the expectation of V_p is*

$$\mathbb{E}(V_p) = K \frac{\underbrace{\sum_{m_1=1}^{K} P_1 \sum_{m_2=2}^{2m_1} P_2 \ldots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} P_{p-1}}_{(p-1)\ \textit{nested sums}} N^{2m_{p-1}} W^2(p)}{\underbrace{\sum_{m_1=1}^{K} P_1 \sum_{m_2=2}^{2m_1} P_2 \ldots \sum_{m_{p-1}=2^{p-2}}^{2m_{p-2}} P_{p-1}}_{(p-1)\ \textit{nested sums}} N^{2m_{p-1}}}, \qquad (3)$$

where we assume $m_0 = K$ and

$$P_j := \sum_{n=2^{j-1}}^{m_j - 2^{j-1}} Q_{2^{j-1}}(m_{j-1}, n) Q_{2^{j-1}}(m_{j-1}, m_j - n) \binom{N}{n} \binom{N}{m_j - n} n!(m_j - n)!$$

and

$$W(p) = \left(1 - \frac{2^{p-1}}{N}\right)^{m_{p-1} - 2^{p-1}} \prod_{s=1}^{2^{p-1}-1} \left(1 - \frac{s}{N}\right).$$

- *If $K/N = \alpha$, then*

$$\lim_{K\to\infty} \frac{\mathbb{E}(V_p)}{K} = e^{-(2^{2p-1})/\alpha}.$$

In particular, for V_1 we have

$$\mathbb{E}(V_1) = K(1 - 1/N)^{2(K-1)}.$$

Finally, we conclude that there is the following relation between U_p and V_p:

$$\lim_{K\to\infty} \frac{\mathbb{E}(V_p)}{K} = \left(\lim_{K\to\infty} \frac{\mathbb{E}(U_p)}{K}\right)^2.$$

2.2 Modeling adaptive introgression

2.2.1 Adaptive introgression

In [10*], we developed a mathematical model to efficiently and accurately numerically estimate the distribution of ancestral chromosome lengths around a genetic locus under the influence of natural selection. Further, in [6*] we used our theoretical model to develop a method based on a Hidden Markov Model to infer adaptive introgression and estimate its parameters.

2.2.2 Model overview

Let us now proceed to the details of our approach. Consider a random process along a chromosome with two states corresponding to two ancestral populations. Transitions between ancestral states occur due to recombinations, with chromosome tracts to the left and right of the recombination point coming from different populations. We consider a model with three genetic loci. Thus, we calculate the transition probabilities between ancestral states at two loci depending on the distance to the third locus under selection.

Let us formulate in more detail the population model under consideration. Let α be a genetic locus under selection with two possible alleles: selected allele A and neutral allele a. We consider the scenario in which allele A enters the population due to adaptive introgression. That is, at a certain time individuals from one population replace a certain proportion of individuals from a second population [5, 6]. We also make the assumption that at the time of introgression all individuals of the donor population are carriers of allele A and all individuals of the recipient population have allele a.

The expected trajectory of the selected allele frequency is accurately described by a logistic curve (see, e.g., [7]) if selection is strong enough and allele frequency is not too close to 0 or to 1 [8,9]. Approximating the random trajectory of the allele frequency under the logistic curve selection avoids integration due to uncertainty due to genetic drift. Using computer simulations, we have shown that within a certain range of parameters, our approximation allows us to estimate very accurately the expected length of the introgressed genetic sites.

Outside this parameter range (e.g., when the admixture fraction is small), the genetic drift is strong and the logistic approximation is inaccurate. This observation coincides with the results of [10]. In this case, the expected trajectory can be efficiently estimated numerically by averaging over a large number of random realizations, provided that one of the two alleles is not fixed. We used this approach to analyze Denisovan introgression in Tibetan ancestors. where the admixture fraction is estimated at only 0.06% [11].

2.2.3 Derivation of the approximate deterministic model

So, we consider a sample chromosome. Our goal is to describe the transitions (along the genome) between ancestral states of loci. In the coalescence with recombination model (with time direction from the present to the past), our model is Markovian, provided that the allele frequency trajectory is fixed. To describe the dynamics of adaptive introgression with three loci, we first need to enumerate all the possible states corresponding to the ancestral configurations of the three loci. Then, we approximate this process by another Markov process along the genome (SMC/SMC' [12, 13]), where the states in the loci of the observed chromosome will correspond to one of the two ancestral populations.

The model has 6 possible states. Each state represents an ancestral configuration for an observed chromosome with three loci: α, β and γ. At α we track the allelic state, A or a, which also indicates ancestry. In β and γ we only need to know if the given chromosome is ancestral to the observed chromosome or not. We use the notation β^a and γ^a to indicate DNA in β or γ that is ancestral to the observed chromosome. β^n and γ^n is used to indicate DNA that is not ancestral to the observed chromosome. The six states are then: $X_1 = (A - \beta^a - \gamma^a), X_2 = (A - \beta^a - \gamma^n, A - \beta^n - \gamma^a), X_3 = (A - \beta^a - \gamma^n, a - \beta^n - \gamma^a), X_4 = (a - \beta^a - \gamma^n, A - \beta^n - \gamma^a), X_5 = (a - \beta^a - \gamma^n, a - \beta^n - \gamma^a), X_6 = (a - \beta^a - \gamma^a)$.

We denote the frequency of the selected allele at time t by $\omega(t)$. As we indicated previously, we assume that $\omega(t)$ deterministically follows a logistic function:

$$\omega(t) = 1 - \frac{1}{1 + e^{-st/2}} = \frac{1}{1 + e^{st/2}},$$

because we are working in backward time.

Recombination acts at a rate proportional to the recombination distances between loci. We assume that recombination between loci α and β occurs at rate r_1 and recombination occurs between loci β and γ at rate r_2.

Let $\lambda = 1/2N_e$, where $2N_e$ - haploid effective population size. Transitions in the Markov process correspond to two types: coalescences and recombinations. Coalescences are possible only between chromosomes with the same allele at locus α. So, the transition matrix of the Markov process is

$$\mathbb{M}(t) = \begin{pmatrix} -r_1\bar{\omega}(t) - r_2 & r_2\omega(t) & r_2\bar{\omega}(t) & 0 & 0 & r_1\bar{\omega}(t) \\ \lambda/\omega(t) & -\lambda/\omega(t) - (2r_1 + r_2)\bar{\omega}(t) & (r_1 + r_2)\bar{\omega}(t) & r_1\bar{\omega}(t) & 0 & 0 \\ 0 & (r_1 + r_2)\omega(t) & -r_1 - r_2\omega(t) & 0 & r_1\bar{\omega}(t) & 0 \\ 0 & r_1\omega(t) & 0 & -r_1 - r_2\bar{\omega}(t) & (r_1 + r_2)\bar{\omega}(t) & 0 \\ 0 & 0 & r_1\omega(t) & (r_1 + r_2)\omega(t) & -\lambda/\bar{\omega}(t) - (2r_1 + r_2)\omega(t) & \lambda/\bar{\omega}(t) \\ r_1\omega(t) & 0 & 0 & r_2\omega(t) & r_2\bar{\omega}(t) & -r_1\omega(t) - r_2 \end{pmatrix},$$

where $\bar{\omega}(t) = 1 - \omega(t)$ is the allele frequency of a. So, our system is described by the Kolmogorov equation

$$\mathbb{P}'(t) = \mathbb{P}(t)\mathbb{M}(t). \tag{4}$$

The initial condition for this equation, corresponding to the dynamics of the introgressed site (that is, with the A allele), is

$$\mathbb{P}(t_0) = (1, 0, 0, 0, 0, 0),$$

and for a site from the recipient population (with allele a) the initial condition is

$$\mathbb{P}(t_0) = (0, 0, 0, 0, 0, 1).$$

2.2.4 Transition rates between ancestral states along the chromosome

In our model, the probability that the locus has an ancestry of type 1 (donor population) or type 0 (recipient population) is equal to the probability that the ancestral chromosome carries the A or a allele, respectively, at the time of introgression.

Now we will consider a new Markov process that describes the ancestries of the locus as it moves along the chromosome away from the adaptive locus. This is only an approximation (see SMC/SMC' [12, 13] model), since the indicated process is not actually Markovian. The states of this process are ancestries of type 0 and 1. By definition, the transition rates between states s_1 and s_2 at position r for this Markov process

$$\tau_{s_1,s_2}(r) = \lim_{dr \to 0} \frac{P(S(r + dr) = s_2 | S(r) = s_1)}{dr}.$$

Thus, the transition rate $\tau_{10}(r)$ of type 1 ancestry to type 0, corresponding to the end of the introgressed tract is

$$\tau_{10}(r) = \lim_{dr \to 0} \frac{P(S(r+dr)=0|S(r)=1)}{dr} = \lim_{dr \to 0} \frac{1}{dr} \frac{P(S(r+dr)=0, S(r)=1)}{P(S(r)=1)}. \quad (5)$$

The numerator $P(S(r+dr)=0, S(r)=1)$ is the probability $P(X_3)$, and the denominator $P(S(r)=1)$ is equal to $P(X_1)+P(X_2)+P(X_3)$. The expression (5) can be easily evaluated numerically for sufficiently small values of r_2.

2.2.5 Numerical results

We have demonstrated that our model accurately models the distribution of tract lengths, as opposed to the exponential distribution. Thus, the figure 1 shows the distribution of the distance from the adaptive locus to one end of the introgressed section. This distribution was estimated by computer simulation, using our deterministic approximation model and an exponential distribution with a parameter inversely proportional to the mean of the numerically modeled distribution (i.e., an estimate of the mathematical expectation). The figure also shows QQ plots for all three pairs of distributions. The reason why the exponential distribution does not accurately model the distribution of introgressed site lengths is that it does not account for the possibility of reverse coalescence after recombination.

Further, using our method, we have demonstrated the perhaps counterintuitive fact that, given the observed frequency of the adaptive allele and a fixed introgression time, stronger selection leads to shorter introgressed tracts (Fig. 2).

2.2.6 Discussion

Adaptive introgression is an important and common phenomenon in evolutionary genetics [14]. We have developed an approximate mathematical model to numerically calculate the distribution of ancestral tract lengths during adaptive introgression near the locus under selection in the case of a single admixture pulse. This approach allows efficient and rapid calculation of such site lengths for a wide range of realistic adaptive introgression scenarios. However, our model does not extend to scenarios such as complex demographics, continuous migration, and multiple mixing pulses, which requires further work.

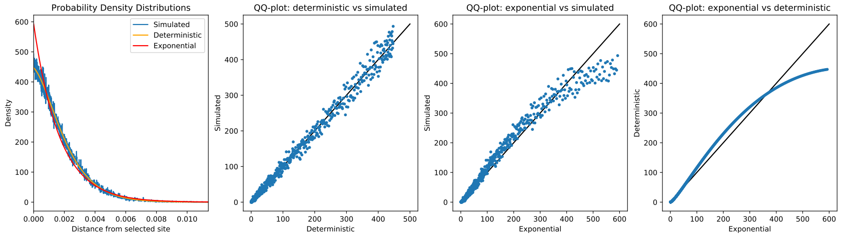


Figure 1: Distribution of the distance from the selected locus to one end of the introgressed tract. Selection coefficient $s = 0.01$, admixture fraction $\omega_1 = 0.1$ and time since introgression $T = 1000$ generations. The observed allele frequency is $\omega_0 = 0.94$. The first panel shows the probability density functions for the empirical distribution obtained by simulations, the distribution calculated under the deterministic approximation and the exponential distribution with the mean set equal to the simulated mean. Three other panels are qq-plots showing all three pairs of the presented distributions.

2.3 Method for inferring adaptive introgression

The adaptive introgression theory developed in the previous section was the basis of the new AHMM-S method. This method is a modification of the Ancestry_HMM [6] method for the local ancestry inference for admixture of two populations based on a Hidden Markov Model. Thus, we assume a single discrete admixture (introgression) event. The emission probabilities in the model with natural selection remain the same as those without selection, that is, they are the same as for the Ancestry_HMM method. The important difference is that natural selection affects the probabilities of transitions between states. These probabilities can be calculated in the deterministic adaptive introgression model presented in the previous section. Such a model is optimized at equal intervals along the chromosome, and the optimization result is compared with the result for the neutral model (without natural selection). This makes it possible to find loci under selection, as well as to estimate the selection coefficient at these sites.

To efficiently compute the transition probability matrix, we proposed a 4-point approximation of the transition rates $f_{10}(r)$ and $f_{01}(r)$. For example, for the transition rate $f_{10}(r)$ from the ancestral state 1 (introgressed population) to the state 0 at genetic distance r (in Morgans) from the selected

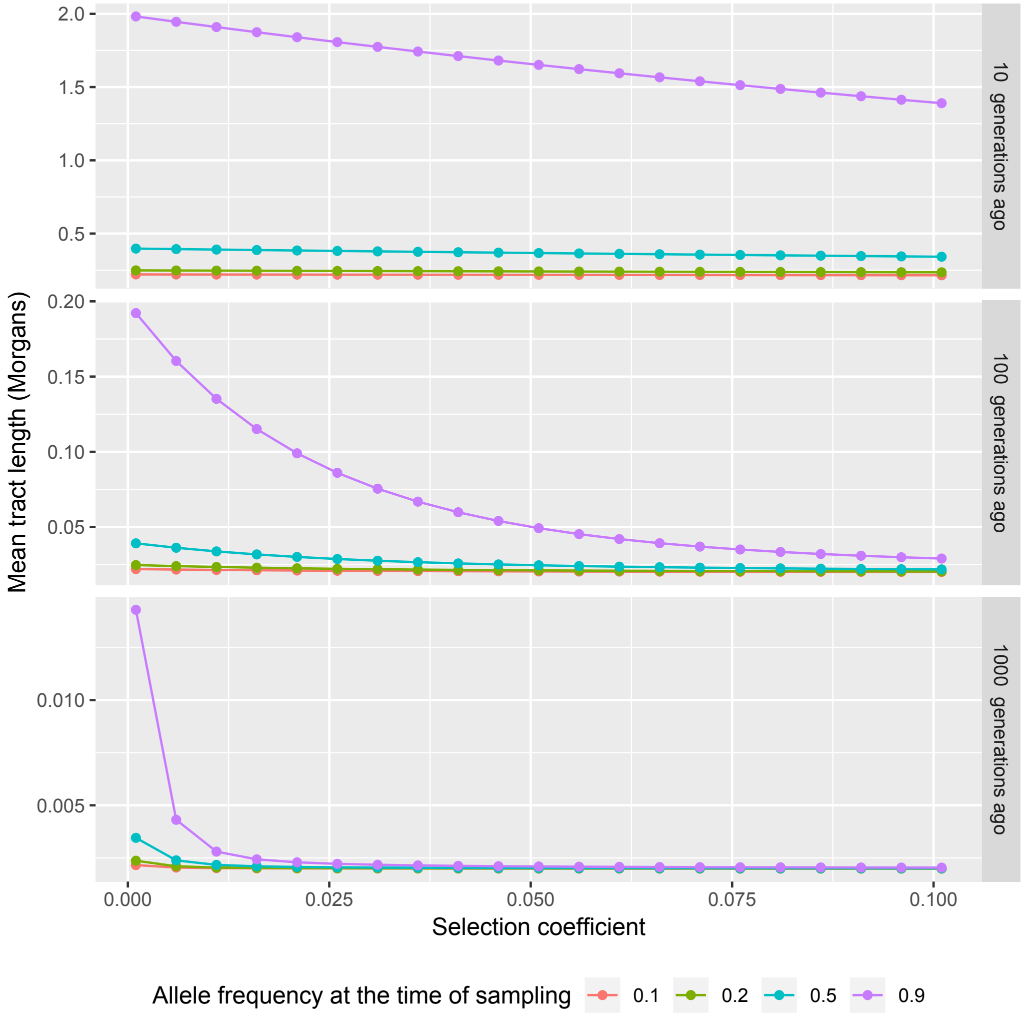


Figure 2: Dependence of the expected length of the introgressed tracts conditional on the allele frequency at the time of observation. Different panels correspond to different introgression times (10, 100, and 1000 generations ago, respectively). Different colors of the lines correspond to different allele frequencies at the time of observation (0.1, 0.2, 0.5, 0.9).

locus we use an approximation in the form

$$\hat{f}_{10}(r) = L - ke^{-\alpha r^p},$$

where L, k, α and p are estimated numerically. This approximation was verified in a wide range of parameters. It substantially improves the performance on the developed software.

The developed method was tested on a wide range of parameters (introgression time, selection strength, admixture fraction). It demonstrates high sensitivity and accuracy of parameter estimation. The software is publicly available on GitHub `https://github.com/jesvedberg/Ancestry_HMM-S/`.

2.4 Viral genealogy simulator `VGsim`

The unprecedented world-wide effort to produce and share viral genomic data for the ongoing SARS-CoV-2 pandemic allows us to trace the spread and the evolution of the virus in real time, and has made apparent the need for improved computational methods to study viral evolution. It is essential that we also have tools to accurately simulate viral evolutionary processes so that the research community can validate inference methods and develop novel insights into the effects of such complexities.

Pandemic-scale datasets impose technical problems associated with the scalability and memory usage of computational methods. The viral genealogy simulator `VGsim`, based on a combination of the generalized SIS model and the structured coalescent approach, efficiently scales for such problems. In the first step (forward pass), the evolutionary dynamics of the virus is simulated using the Gillespie algorithm taking into account many realcistic complexities. The second step (backward pass) uses a structured coalescence approach, which simulates the genealogical tree of pathogen samples conditional on the simulated dynamics obtained during the forward pass.

2.4.1 Model and implementation

Our epidemiological model is based on compartmental models [15]. Random trajectories are implemented using the Gillespie algorithm [16] (the logarithmic direct method [16] and the approximate τ-leaping algorithm [17] are implemented). The different compartments in our model are defined by several interacting factors: (1) host population structure, (2) different groups of infected depending on the strain, and (3) different groups of susceptible host individuals.

As stated earlier, the simulation consists of two parts: a forward pass generating epidemiological dynamics, and a backward pass generating a genealogy of samples based on these dynamics. The epidemiological dynamics are represented as a chain of events (Figure 3). This chain of events define the probability space for the genealogy that is generated during the backward pass. A coalescent approach conditional by the chain of events is used for this purpose.

`VGsim` provides a convenient and compact user interface in Python. The critical computational parts are implemented in C++ using Cython [18].

2.4.2 Results

We compared the performance of `VGsim` with that of the simulator `MASTER` [19] which is popular in epidemiological studies [20–22] and also implements the Gillespie algorithm. The scalability and memory usage of `VGsim` is significantly higher than that of `MASTER` (see Fig. 4).

We also made a comparison with the epidemiological simulator `TiPS` [23], which also uses a combination of the generalized SIS model and structured coalescence and generates genealogies with epidemiological trajectories. For a simple SIR model, the implementation of the exact Gillespie algorithm in `VGsim` is about twice as fast as in `TiPS` (see Table 1). For backward run (generating genealogies using epidemiological trajectories), `VGsim` scales much better than `TiPS`

Number of iterations	VGsim	TiPS
10^6	0.19	0.31
$5 \cdot 10^6$	0.96	1.72
10^7	1.84	3.44
$5 \cdot 10^7$	9.87	17.57
10^8	19.06	38.94

Table 1: Forward run time in seconds for different number of iterations under the SIR model. The recovery rate is set to 1 and the transmission rate to 2.5. The tests were run on a MacBookPro with Quad-Core Intel Core i5 2 GHz processor and 16GB of memory.

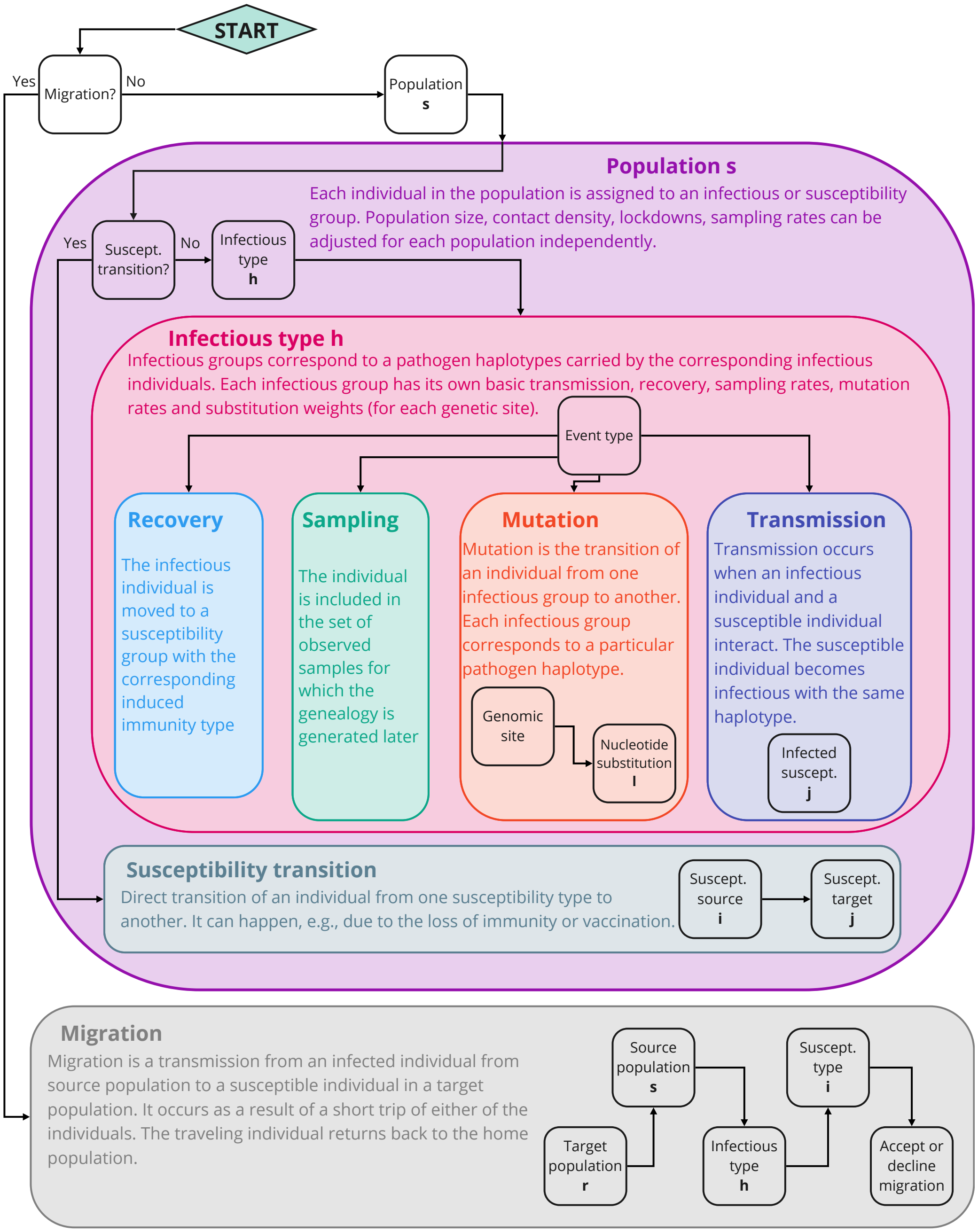

START
Migration?
Yes
No
Population s
Population s
Each individual in the population is assigned to an infectious or susceptibility group. Population size, contact density, lockdowns, sampling rates can be adjusted for each population independently.
Suscept. transition?
Yes
No
Infectious type h
Infectious type h
Infectious groups correspond to a pathogen haplotypes carried by the corresponding infectious individuals. Each infectious group has its own basic transmission, recovery, sampling rates, mutation rates and substitution weights (for each genetic site).
Event type
Recovery
The infectious individual is moved to a susceptibility group with the corresponding induced immunity type
Sampling
The individual is included in the set of observed samples for which the genealogy is generated later
Mutation
Mutation is the transition of an individual from one infectious group to another. Each infectious group corresponds to a particular pathogen haplotype.
Genomic site
Nucleotide substitution l
Transmission
Transmission occurs when an infectious individual and a susceptible individual interact. The susceptible individual becomes infectious with the same haplotype.
Infected suscept. j
Susceptibility transition
Direct transition of an individual from one susceptibility type to another. It can happen, e.g., due to the loss of immunity or vaccination.
Suscept. source i
Suscept. target j
Migration
Migration is a transmission from an infected individual from source population to a susceptible individual in a target population. It occurs as a result of a short trip of either of the individuals. The traveling individual returns back to the home population.
Source population s
Suscept. type i
Target population r
Infectious type h
Accept or decline migration

Figure 3: Scheme of the Gillespie algorithm used in the forward pass used for the chain of events generation. The black squares correspond to sequential steps, where subsets of events are selected according to their weights, or propensities, depending on the model parameterization and the current state of the epidemiological process. The propensities for each step are cached and updated, based on the dependency graph, only if their values change due to the generated event.

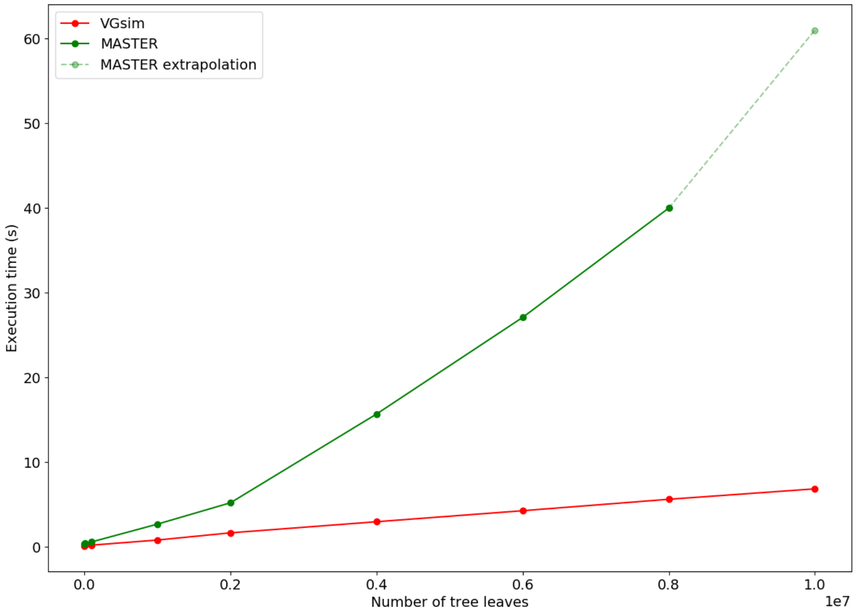


Figure 4: The comparison of `VGsim` and `MASTER` performance: the time to simulate a tree with a given number of leaves.

2.5 Accuracy of the SMC' approximation of structured coalescent

Sequential Markovian Coalescent (shortly SMC) [12] and its modification SMC' [13] are two of the most important models in population genetics which underlie many algorithms and methods for genetic data analysis, e.g. diCal [24], PSMC [1] and MSMC [25]. These models approximate the full coalescent with recombination [**?**] by considering a Markovian process along a chromosome. In our recent work [**?**] we showed that under panmictic SMC model applied to a sample with structured population history, leads to biased and inaccurate estimates of the distribution of times to the most recent common ancestor.

Consider the joint distribution of times to the last common ancestor of two chromosomes at two loci at the genetic distance ρ (in this case, ρ is the recombination rate between the two loci) in the coalescence models with

Number of samples	VGsim	TiPS
10^4	0.059	242.4
$2 \cdot 10^4$	0.11	808.2
$3 \cdot 10^4$	0.18	1377.6
$4 \cdot 10^4$	0.22	1921.2
$5 \cdot 10^4$	0.27	3157.2

Table 2: Backward run time in seconds to generate genealogies for different sample sizes under the SIR model. The population size is 10^7 in all runs. The recovery rate is 1 while the transmission rate is 2.5. The tests were run on a MacBookPro with Quad-Core Intel Core i5 2 GHz processor and 16GB of memory.

recombination and SMC'. We denote these distributions by $p_{\rho,CR}(t,s)$ and $p_{\rho,SMC'}(t,s)$, respectively.

To compute $p_{\rho,CR}(t,s)$, we consider a Markov process with continuous time (time flows from present to past). We denote the coalescence rates in the first and second populations by λ_1 and λ_2, and the migration rates by m_{12} and m_{21}.

The states of backward in time Markov process (structured coalescent with recombination) correspond to different configurations of ancestral chromosomes. Each chromosome consists of two loci and is found in one of the two ancestral populations.

Chromosomes might coalesce when they are in the same population. In total there are 40 states and additionally two more absorbing states corresponding to the LCA. There are three types of transitions in this model:

- recombinations with rate ρ,
- coalescences with rates λ_1 and λ_2 per pair depending on a population,
- migrations with rates m_{12} and m_{21}.

The SMC' model approximates coalescence with recombination by another Markov process directed along the genome. The states of this process are genealogical trees at the locus. In the case of two chromosomes, the shape of the tree is trivial, and the states are actually times to the last commin ancestor.

Firstly, we show the difference between joint distribution of the TMRCA times of two chromosome with two loci at recombination distance ρ. As shown at Fig. 5, there is a clear difference in the qualitative behavior between a single population model and two population model.

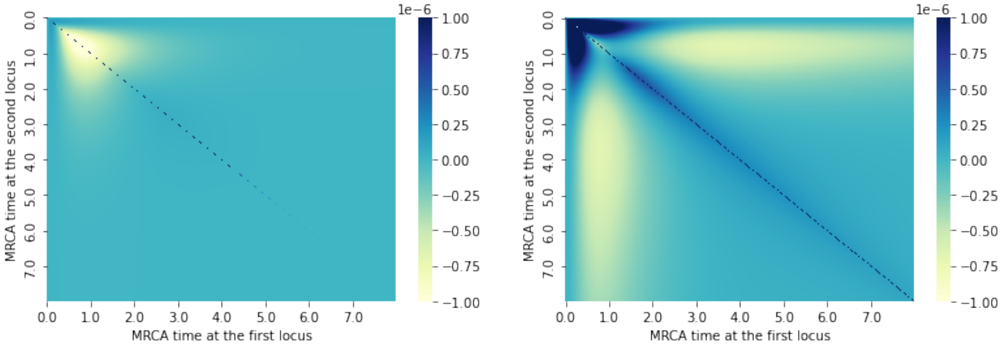


Figure 5: Differences between joint probability distributions of the MRCA times at two loci for recombination distance $\rho = 2$ under full coalescent with recombination and SMC' models. Left panel shows the difference for a single population scenario. Right panel shows the difference for two populations with migration scenario.

Secondly, we calculate the total variation between these joint probabilities in function of the recombination distance ρ. As defined in [26], total variation is the L_1-norm of the difference between two joint distributions divided by two. For our analysis we calculate the total variation between the joint MRCA times distributions $p_{\rho,CR}(t,s)$ (under coalescent with recombination) and $p_{\rho,SMC'}(t,s)$ (under SMC')

$$TV(\rho) = \frac{1}{2}\int_0^\infty \int_0^\infty |p_{\rho,CR}(t,s) - p_{\rho,SMC'}(t,s)|\, dtds.$$

From Fig. 6, one can notice that for the structured model total variation is larger than for the panmictic case. Importantly, the decay of variation distance is much slower for two populations with migrations.

2.6 Effective population size and migration

The effective population size can be defined as the average time to coalescence (to the last common ancestor) of two ancestral linages, measured in the

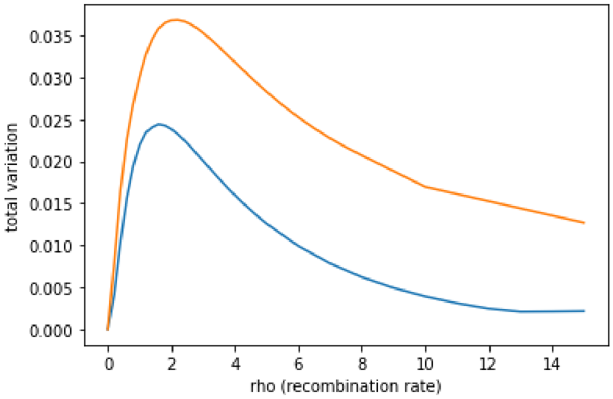


Figure 6: Total variation distance between joint probability distribution of the MRCA times at two loci under full coalescent with recombination and SMC' models as a function of recombination distance ρ. Blue (lower) line corresponds to a single population scenario, orange (upper) line corresponds to the two population with migration scenario.

number of generations [27]. In the standard coalescent model with one homogeneous population [28], the following simple interpretation takes place. If the population size is $N >> 1$, the coalescence rate is $1/N$ and the expected coalescence time is $\lambda^{-1} = N$. This definition can be naturally generalized for the effective size of a mixed population. For some population, we consider the coalescence rate at time t between pairs of ancestral linages. Time is considered in the reverse direction from the present ($t = 0$) to the past. The coalescence rate $\lambda(t)$ defines a Markovian time-inhomogeneous process that describes the distribution of coalescence times. Hence, the probability distribution of coalescence times T_c is given by the law

$$P(T_c = t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}.$$

We define *historical effective population size* at time t as the inverse of the coalescence rate $\lambda(t)$. This value depends on population structure and demography. It allows us to approximate populations with complex histories by a single idealized population (e.g., the Wright-Fisher population). This approach is useful in interpreting the results of methods such as PSMC, which allow us to estimate the change in population size over time. In some cases, it also approximates the value estimated by PSMC [1, 29]. We have shown that although PSMC does give a good estimate of coalescent times in some scenarios, in other scenarios PSMC can lead to significantly different results. We have shown qualitatively that PSMC, informally, searches for the best approximation for the transition matrix of the sequential Markov model. This is because the likelihood function optimized by PSMC depends directly on the transition probabilities. Nevertheless, the transition matrix may have a different marginal distribution that determines the effective population size.

Consider the case of two populations with migration between them - this can occur either through continuous migration or pulse migration. Let us denote the ancestral populations by $S_1(t)$ and $S_2(t)$. At any point in time, the ancestral linage of the observed population S_m is either in population $S_1(t)$ or in population $S_2(t)$ due to migration. Within populations $S_1(t)$ and $S_2(t)$, ancestral lineages are indistinguishable, which means that any pair of lineages within a population has the same probability of coalescence. Let us denote the effective population sizes $S_1(t)$ and $S_2(t)$ by $N_{L1}(t)$ and $N_{L2}(t)$, respectively. We will call $N_{L1}(t)$ and $N_{L2}(t)$ the local effective population sizes, that is, these values represent only the effective number of individuals in the population at time t, thus separating the effect of migration and the genetic diversity of populations.

If two ancestral lineages are in the same population $S_i(t)$ $(i = 1, 2)$ at time t, coalescence can occur between them with a rate of $1/N_{Li}(t)$. If they are in different populations, coalescence between them is not possible. Assuming that the two lineages have not merged by the time t, let $P_1(t)$ and $P_2(t)$ be the probabilities that the two lineages are in population S_1 and population S_2, respectively. Let $P_0(t) = 1 - P_1(t) - P_2(t)$ be the probability that the two lineages are in different populations. Then the coalescence rate between a pair of ancestral lines at time t is

$$\lambda(t) = P_1(t)\frac{1}{N_{L1}(t)} + P_2(t)\frac{1}{N_{L2}(t)} + P_0(t) \cdot 0, \tag{6}$$

and historical effective population size is

$$N(t) = \frac{1}{\lambda(t)} = \frac{1}{P_1(t)\frac{1}{N_{L1}(t)} + P_2(t)\frac{1}{N_{L2}(t)}}. \tag{7}$$

The condition that the observed population S_m is the population $S_1(0)$ is equivalent to the initial conditions on the probabilities

$$P_1(0) = 1, P_2(0) = P_0(0) = 0.$$

Thus, we see an obvious difference between the local population sizes ($N_{L1}(t)$ and $N_{L2}(t)$) and the historical population size ($N(t)$). In many cases, estimates of effective population size obtained by PSMC and similar methods are estimates of historical effective population size rather than local effective size.

2.6.1 Disentangling the effect of migration on effective population size

Assume that we observe two populations $S_m^{(1)} = S_1(0)$ and $S_m^{(2)} = S_2(0)$, which had ancestral admixture with each other. Writing Equation 7 for samples from both populations, we get the system of equations relating the *ordinary effective population size* of $S_m^{(1)}$ and $S_m^{(2)}$ (N_1 and N_2) with the *local effective population size* of each of the two parental populations (N_{L1} and N_{L2}).

$$\begin{cases} N_1(t) = \dfrac{1}{P_1^{(1)}(t)\frac{1}{N_{L1}(t)} + P_2^{(1)}(t)\frac{1}{N_{L2}(t)}}, \\ N_2(t) = \dfrac{1}{P_1^{(2)}(t)\frac{1}{N_{L1}(t)} + P_2^{(2)}(t)\frac{1}{N_{L2}(t)}}. \end{cases} \tag{8}$$

Thus, with known historical effective population sizes N_1 and N_2 (e.g., estimated by the PSMC method) and migration rates m_{12} and m_{21}, the local effective population sizes N_{L1} and N_{L2} can be calculated numerically.

2.6.2 Parameter estimation

We applied the developed method to test the hypothesis of [2] about the deep split of populations within Africa between San and Dinka. Namely, the estimate of the split time obtained by the TT method exceeds 8,500 generations ago. The model with the highest likelihood obtained by our MiSTI method estimates the split time of $\approx$ 3700 generations ago (that is, about 107,000 years ago, assuming a generation length of 29 years) with almost unilateral migration from Dinka to San (Table 3). These conclusions are consistent with estimates obtained from simulations with similar parameters.

m1	m2	MiSTI split time		TT split time
Dinka to San	San to Dinka	(generations)	log(lik)	(generations)
2.5	2.03×10^{-9}	3729	-4381	-
2.5	-	3729	-4381	-
-	1.49	3210	-4582	-

-	-	3001	-4607	T1 = 8582,
				T2 = 8527

Table 3: MiSTI estimates of split times and migration rates between the San and Dinka populations in models with bidirectional migration (top row), unidirectional migration, or no migration (bottom row).

2.7 Multiple admixture and three loci linkage disequilibrium

There are many methods for predicting the presence of admixture between populationsreich2009reconstructing, Patterson:2012aa, Durand:2011aa, pritchard2000inference, alexander2009fast, maples2013rfmix. There has also been a substantial amount of research on the development of theory and methods for estimating admixture times. One approach is based on predicting ancestral chromosome lengths (originating from different ancestral populations) [6, 30–33] and [6*].

Another approach we use in this work is based on the admixture linkage disequilibrium (ALD) decay. Linkage disequilibrium is present in any population because of mutations and genetic drift. In a homogeneous and genetically isolated population with recombination, it decreases rapidly on a genome-wide scale. However, ancestral sites entering populations through admixture result in a noticable ALD at much greater distances. After a single admixture, the linkage disequilibrium in the admixed population begins to gradually decrease in subsequent generations due to recombinations. This idea was used in the methods ROLLOFF [34] and ALDER [35], where ALD is estimated for two genetic loci. ROLLOFF and ALDER are well suited for estimating the admixture time if admixture can be approximated by a single pulse of migration. However, in many realistic scenarios, the admixture occurred through several migration pulses. For example, a well-known example of such admixture is the admixture of the Native American population of Easter Island [36] as well as the admixed American populations [37]. In such cases, the expected ALD decrease becomes a mixture of exponential laws. Existing ALD-based admixture dating methods at this time can either estimate the time of the last admixture [34] or be used to reject the single pulse hypothesis [35].

In our work, we use Bennett and Slatkin's definition [38,39] for the linkage disequilibrium of three loci in order to study the ALD decay as a function of distances between these loci. We derived an analytical equation describing the ALD decay with multiple pulses of migration and also developed a method to estimate the times of two pulses of migration. The results are validated through computer simulations and applied to real data, samples of Mexicans and Colombians from the 1000 Genomes Project, as an example.

2.7.1 Linkage disequilibrium and local ancestry

Denote by x, y, z three successive genetic loci with the distances d and d' between them. $H_{i,x}, H_{i,y}, H_{i,z}$ - haplotypes ($\{0,1\}$) or genotypes ($\{0,1/2,1\}$) in the corresponding loci of i-the genome. Three loci linkage disequilibriumis defined as the covariance of H_x, H_y, H_z

$$D_3(d,d') = \mathrm{cov}(H_x, H_y, H_z) = \mathbb{E}[(H_x - \mathbb{E}H_x)(H_y - \mathbb{E}H_y)(H_z - \mathbb{E}H_z)]. \quad (9)$$

The linkage disequilibrium in an admixed population depends on the genetic differentiation between the original populations and their history of admixture. Let A_x represent the local origin of locus x, with $A_x = 0$ if x is inherited from a population that is admixed twice, and $A_x = 1$ if the locus is inherited from a population that is admixed once. Then D_3 can be represented through allele frequencies and local-origin covariance A_x, A_y, A_z. Consider the conditional expectation $\mathbb{E}(H_x|A_x) = g_x + \delta_x A_x$, where g_x is the allele frequency at locus x in population 0 and $\delta_x = f_x - g_x$ is the difference in allele frequencies at locus x in two source populations. We assume that the allele frequencies in the initial population are known and fixed. Then

$$D_3(d,d') = \mathrm{cov}(H_x, H_y, H_z) = \delta_x \delta_y \delta_z \mathrm{cov}(A_x, A_y, A_z). \quad (10)$$

Moreover, the following equality holds

$$\begin{aligned} \mathrm{cov}(H_{S_1}, \ldots, H_{S_N}) &= \mathrm{cov}\left(g_{S_1} + \delta_{S_1} A_{S_1}, \ldots, g_{S_N} + \delta_{S_N} A_{S_N}\right) \\ &= \mathrm{cov}(A_{S_1}, \ldots, A_{S_N}) \prod_{i=1}^{N} \delta_{S_i}. \quad (11) \end{aligned}$$

2.8 Deep learning for demographic analysis

In [4*, 3*] a method based on deep learning to predict the local times of the last common ancestor from the diploid genome is presented.

Predicting demography, that is, estimating the historical effective population size, is one of the most important tasks of population genetics. It is one of the key factors of population genetic diversity. For example, [1] shows that all non-African populations went through a bottleneck between approximately 30 and 100 thousand years ago. African populations do not have this phenomenon. This fact supports the hypothesis of the African origin of modern humans.

Deep learning demonstrates high accuracy for many problems, including analysis of different sequences. We have developed an architecture based on recurrent neural networks to predict local times to the last common ancestor from a diploid sequence (similar to PSMC). This task has two key challenges: the large length of the genomic sequence ($3.2 \cdot 10^9$ for humans) and the lack of labeled data for training.

At the moment, deep learning is gradually beginning to be used in population genetics, although it is not a popular approach. The first method for predicting detailed population history using deep learning is proposed in [40] and proves that neural network approaches can be powerful tools in population genetics. Nevertheless, much work remains to be done in this area, including the advantages, limitations, and drawbacks of deep learning in the tasks at hand, before these methods find widespread application in the analysis of experimental data.

We used the software simulator `msprime` [41] to generate suitable samples for training. A sequence of 0 (homozygous sites) and 1 (heterozygous sites) is fed to the input of the neural network. The prediction target is one of the time intervals where the local last common ancestor is located. The problem was solved as a classification problem. The program code is publicly available on GitHub `https://github.com/Genomics-HSE/deepgen`. An example of time prediction for a common ancestor along the genome is demonstrated in Figure 7. The x axis corresponds to positions along the genome, the y axis to the time intervals where the local common ancestor falls. The colors correspond to the probabilities of the local last common ancestor falling into a certain time interval. Thus, the developed deep learning method predicts quite accurately the time to the local last common ancestor.

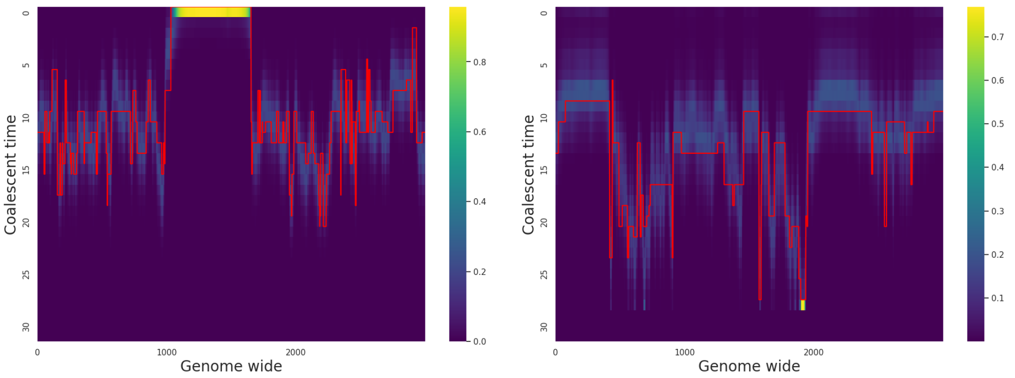


Figure 7: The time to the last common ancestor (LCA) along the chromosome. The heat map shows the probability that the LCA time fell within a given time interval (y-axis) at the considered position on the chromosome (x-axis). The left panel shows an example for the case of a population of constant size and a neural network trained on examples from constant size populations. The right panel shows an example for a population with a bottleneck, with the neural network trained on examples from random demographic histories. The red lines show the true LCA times known from the corresponding simulations.

2.9 Selection in Chilean population adter post-Columbian adxmiture

The article [11*] investigates the presence of natural selection after admixture (adaptive introgression) in the Chilean population. The modern Chilean population emerged as a consequence of the admixture of the indigenous South American population, European colonizers and Africans. To solve the problem, we performed a genome-wide search for the deviation in the proportion of European local ancestry from the genome-wide average.

After predicting local ancestry using the LAMP-LD method, we calculated the proportion of European local ancestry for each SNP. Next, we used a one-way Student's t-test to determine the deviations from the mean European ancestry. In each SNP, we compared the proportion of European ancestry with the genome-wide mean $p_0 = 0.52$. We performed a statistical test with the null hypothesis $H_0 : p_i = p_0$ and the alternative hypothesis $H_1 : p_i > p_0$ for each SNP i (here p_i is the proportion of European origin in SNP i). Variants (SNPs) that reached a significant level with $p < 10^{-5}$ were considered to be under selection after admixture [42].

To verify the validity of this approach and the choice of the significance level, we performed computer simulations of the p_i distribution. Since deviations of p_i from the mean are often due to genetic drift, we estimated the effective population size for our Chilean samples by equating the empirical and theoretical variance of p_i. According to [43], the variance

$$\mathbb{V}p_i = p_0(1-p_0)\left(1-e^{-\frac{T}{2N_e}}\right),$$

where p_0 is the proportion of admixture, T is the time since admixture, $2N_e$ is the haploid effective population size.

The distribution p_i can be approximated by a beta distribution with appropriate expectation and variance. Let us also consider that the fraction of European origin observed in the empirical data depends on the sample size. In each SNP with number i, the number k_i of sequences of European origin is a realization of a binomial random variable with sample size K (in our case $K = 370$) and probability p_1. Hence, the distribution for k_i is

$$k_i \sim \int_0^1 P(Binom(p_i, K))d(p_i).$$

From here, we can estimate the value of $2N_e$ such that the variance p_i coincides with the variance estimate of European local ancestry from the empirical data (for all SNPs in our Chilean dataset with $N = 370$ haplotypes).

We used the $2N_e$ effective population size estimates thus obtained to model local origin in the SELAM package [44]. We considered a scenario with a single simultaneous mixing of three populations T generations ago (for different realistic values of T). The admixture fractions corresponded to the fractions of the European, Native American, and African components estimated using LAMP-LD. We then mapped the SNP positions from our empirical dataset to the modeled sequences to obtain the same correlation structure (resulting from recombinations). We calculated Student's t-test for all parameter combinations. In no case were the p-values below the critical threshold $p = 10^{-5}$. Thus, the chosen level of statistical significance can indeed be considered an indicator of the presence of selection after admixture.

Our numerical approach allowed us to test our hypothesis that, after admixture, Chileans were subjected to natural selection by genetic variation of European origin. Because ancestral variation can contain the same genetic variants, for each SNP we determined its genome-specific origin. We then used the deviation of the local ancestry proportion [42,45] (rather than directly the allele frequency) from the genome-wide average ancestry (estimate 0.52 for Europeans) as a selection signal. We constructed a t-test with the null hypothesis $H_0 : \mu_{EUR,i} = 0.52$ and the competing hypothesis $H_1 : \mu_{EUR,i} > 0.52$ for each variant i.

We found 85 SNPs that reach the statistical significance threshold $P < 10^{-5}$ recommended for recently mixed populations [42]. We justified the choice of this level of statistical significance using computer modeling (see above). 85 SNPs correspond to a peak of European origin on chromosome 12. This site is associated with several regulatory regions, including two lncRNAs (*RP11-13A1.1 and RP11-13A1.3*) and one pseudogene (*RP11-13A1.2*).

2.10 Phylodynamics of coronavirus SARS-CoV-2 in Russia

The Covid-19 pandemic raised many challenges to the scientific community. In particular, great efforts have been focused on the sequencing of SARS-CoV-2 coronavirus samples in most regions of the world. This, in turn, has enabled genomic epidemiological analysis to study the distribution of different variants (strains) of the coronavirus. In this section, we present the results of phylodynamic analysis of the Covid-19 outbreak at the Vreden hospital (Research Institute of Traumatology) in March-April 2020 [9*] and the coronavirus delta variant in Moscow in April-September 2021 [7*].

2.10.1 Covid-19 nosocomial outbreak in the Vreden hospital

We investigated a large transmission cluster, the nosocomial outbreak of Covid-19 at the Vreden Hospital. Vreden Hospital in St. Petersburg at the beginning of the pandemic. According to an internal investigation, the presumptive patient zero was operated on March 27, 2020. Although regular testing for Covid-19 at Vreden Hospital began on March 18, 2020, the first positive sample was obtained on April 3, 2020. Quarantine measures were then phased in between April 7 and April 9, 2020, which included complete hospital closure, isolation of departments, and shutdown of the hospital-wide ventilation system. 474 patients and 270 staff remained in the hospital for 35 days.

Our dataset consists of SARS-CoV-2 virus genomes from 52 patients and staff at Vreden Hospital. Phylogenetic analysis showed that these samples form three different groups with their own unique set of mutations. The largest group (hereafter group 1) includes 41 sequences obtained between April 3 and April 22, 2020. Group 2 consists of 7 sequences, and its corresponding clade on the world phylogenetic tree also includes one sequence from England. Finally, Group 3 consists of 4 sequences. Group 1 specimens come from different compartments on different floors, while groups 2 and 3 are each from their own compartment.

Groups 1 and 2 are phylogenetically distant from group 3. The closest common ancestor of groups 1 and 2 is separated by six mutations from group 3. Groups 1 and 2 belong to lineage B.1.1, defined by three mutations at positions 28881, 28882, and 28883, and further defined by mutations at positions 26750 and 1191, respectively. At the same time, group 3 belongs to lineage B.1.5, and is also complemented by a mutation at position 20268, which at that time was distributed throughout the world and appeared early in the phylogenetic history, as well as two additional mutations. Thus, we obtained strong evidence that group 3 appeared as a result of independent introduction of infection relative to groups 1 and 2.

To examine the spread of this outbreak of nosocomial infection in more detail, we performed a Bayesian phylogenetic analysis in the [46] birth-death skyline model in the BEAST2 [47] package. Because of the high probability of multiple introductions of infection, we analyzed both the entire dataset consisting of groups 1, 2, and 3 as well as its two subsets: one consisting of groups 1 and 2 and the other consisting of group 1. The results of our analysis are shown in figures 8 and 9.

We found that Bayesian analysis supports at least two different introductions of SARS-CoV-2 coronavirus into Vreden Hospital. This is supported by a deep split between groups 1-2 and group 3. The last common ancestor of this dataset is dated February 21, 2020 (95% posterior credible interval January 20-March 21). This is more than a month earlier than the estimated date of the first introduction (March 27), which again confirms that group 3 and all other specimens were introduced into the hospital independently.

A third introduction to Vreden Hospital is also highly probable. Indeed, the last common ancestor of groups 1 and 2 is dated March 24, 2020 (95% credible interval March 6-April 1). Given the lack of obvious signs of infection at the hospital in late March, it is very likely that Groups 1 and 2 are derived from two independent introductions. The root (last common ancestor) of group 1 is dated March 26 (95% credible interval March 13-April 2), which is consistent with the period of illness of putative patient zero. Additional

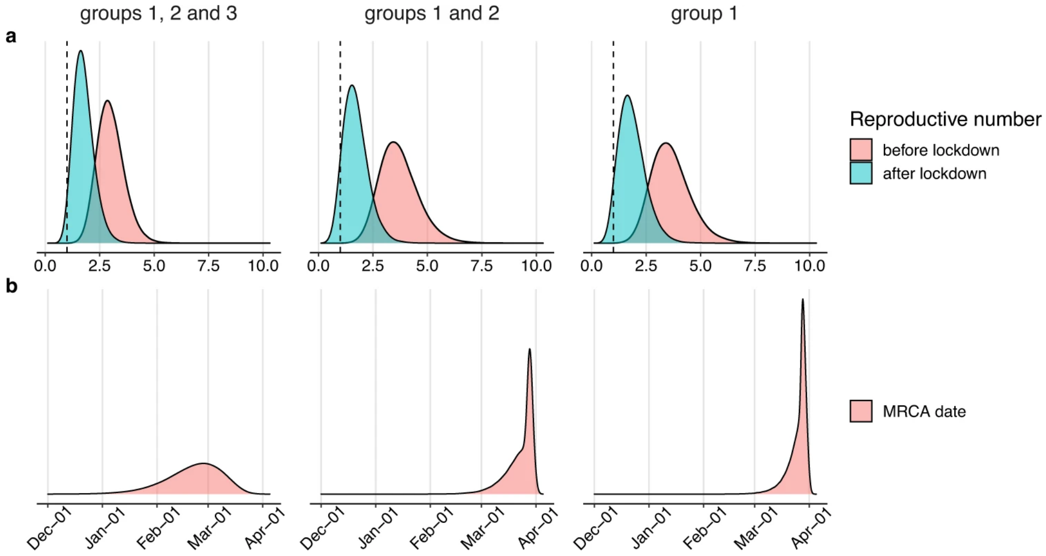


Figure 8: Estimates of the phylodynamic parameters of the Covid-19 outbreak at Vreden Hospital in the horizonal birth-death model in the BEAST2 package. The upper panel shows posterior distributions of the effective reproductive number R_e (upper panel) with a dashed line showing the critical value $R_e = 1$. The lower panel shows posterior distributions of the date of the last common ancestor. Each panel shows the estimates for three datasets from pest samples: groups 1, 2, and 3 (left column), groups 1 and 2 (middle column), and group 1 (right column).

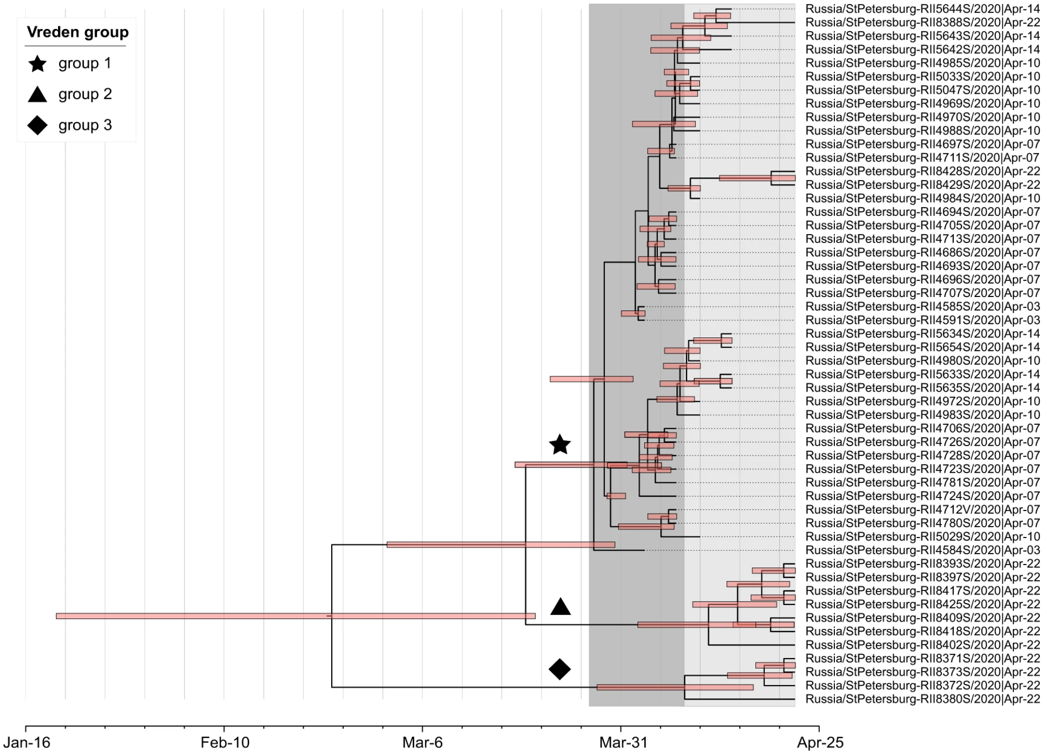


Figure 9: Maximum clade credibility tree for the Covid-19 outbreak at Vreden Hospital. Groups 1, 2, and 3 are marked with an asterisk, a triangle, and a diamond, respectively. Pink bars show 95% credible intervals for node times. The outbreak period is marked by a gray background, with the time from estimated patient zero (March 27) to quarantine (April 8) highlighted in dark color.

confirmation of the independent ancestry of groups 1 and 2 is the presence of a non-Russian (English) sequence in the same clade.

We evaluated phylodynamic parameters before and after the introduction of quarantine measures. In the analysis of all three datasets, estimates of the effective reproductive number remain stable and consistent with each other. Based on analysis of all three groups, we found that effective reproductive number R_e was 3.0 (95% credible interval $1.85 - 4.25$) before April 8 and decreased to 1.76 (95% credible interval $0.91 - 2.71$) after April 8 (Figure 8). Similar estimates for the effective reproductive number R_e for group 1 are 3.64 (95% credible interval $2.01 - 5.43$) before quarantine and 1.85 (95% credible interval $0.77 - 3.06$) after quarantine, respectively. These estimates are consistent with each other, and the possible effects of a structured population (due to independent introductions) do not create significant biases in the estimates.

2.10.2 Discussion

A detailed analysis of localized transmisson clusters helps to better understand the process of virus spreading. Well-studied cases at the time included the cruise ship “Diamond Princess” [48–51], the cruise ship “Grand Princess” [52], an international conference in Boston [53], a hostel near Boston [53] and an outbreak in a hospital at the Netcare St. Augustine Hospital in South Africa. Augustine, South Africa. In all but one case, the outbreaks were genetically homogeneous, meaning that each developed from a single case of infection. In the case of the hostel, several introductions occurred, but nevertheless there was a major clade that included almost all samples, while the other clades were rare [53]. At the same time, in the case of the outbreak at Vreden Hospital, we observe several (probably 2 or 3) introductions, each of which resulted in a separate clade. This could mean that this outbreak happened because of several instances of super-spreading.

Further, our estimate of the initial effective reproductive number R_e (during the period before quarantine) is about ~ 3.00, which is a fairly high value. The few cases of super-spreading and the high R_e value may be a consequence of the specificity (traumatology) of this hospital not being equipped for infection control, particularly close contact (e.g., spread among staff), lack of protective measures, and lack of awareness. In the second phase of the outbreak, we observe a significant decrease in the effective reproductive number to ~ 1.76. This change can be explained by two factors (or a combination of them). First, it may be a consequence of increased awareness and the introduction of quarantine measures starting on April 7. Second, it could be a consequence of the fact that many people had already gotten sick by then, which in turn prevented further spread of the infection. In fact, about 30% of the people in the hospital were infected by April 22. We have no way of assessing the contribution of each of these factors in slowing the spread of infection with the available data and methods.

2.10.3 Phylodynamics of delta variant of SARS-CoV-2 coronavirus in Moscow

By mid-2021, the delta variant of the SARS-CoV-2 coronavirus had displaced all other variants worldwide. This variant was characterized by increased transmissibility and lethality. In Russia, unlike most other countries, one transmission lineage *nsp2:K81N + ORF7a:P45L* spread (over 90% of cases), and this lineage is rarely seen outside of Russia. We investigated the distribution of this lineage in the country, in particular we evaluated the phylodynamics of this lineage in Moscow (the most well represented region in our genetic sample).

Further, all the dates in this section refer to the year 2021.

In order to estimate the transmission rate of the largest Delta variant sublineage, we performed a phylodynamic analysis using the BEAST2 [47] package. The Covid-19 epidemic runs differently and non-synchronously in different regions of Russia. For example, the timing of epidemic waves differs in different regions. In order to minimize the effects of geographic heterogeneity, we focused on a single region in this analysis. We chose 333 samples collected in Moscow, since, as mentioned earlier, this is the Russian region with the largest amount of data.

Phylodynamic estimates of the effective reproductive number R_e for the indicated major clade are 1.82 (95% CI $[1.49 - 2.16]$) in May, 1.24 (95% CI $[1.07 - 1.41]$) in June. In July, the R_e value fell to 0.58 (95% CI $[0.40 - 0.77]$) and then rose again to 0.99 (95% CI $[0.79 - 1.20]$) in August and to 1.27 (95% CI $[0.62 - 1.94]$) in September, the last month included in our genetic analysis (Figure 10).

In general, the above dynamics are consistent with epidemiological data: elevated R_e values precede surges in the number of cases per day and are consistent with estimates of R_e derived by the EpiEstim method from the number of cases reported. It is important to note that the number of cases through June includes a large proportion of non-Delta cases. The rise in the total number of cases in May was slower than the corresponding R_e predicts, which can be explained by the decrease in the number of non-Delta cases. However, the high R_e values in May and June are consistent with a summer wave that peaked on June 25, and the low R_e value in July is consistent with a decrease in cases during this period (Figure 10). These data confirm that the major clade (AY.122+ORF7a:P45L) is responsible for the summer epidemic wave, and probably for the subsequent fall wave. This bimodal dynamic is similar to many other countries in the northern hemisphere, where the arrival of summer has slowed the spread of infection, such as in the United Kingdom, France, and the United States.

2.11 Hyperbolic geometry and genetic data analysis

Non-Euclidean, and hyperbolic geometry in particular, is finding more and more applications in data analysis. Since genealogies are based on trees, we conjectured that hyperbolic geometry is a promising tool for the analysis of genetic data and conducted research in this direction. We have laid theoretical foundation for such analysis in works on numerical aspects of hyperbolic geometry. The following main results have been obtained:

- an optimal (up to the multiplicative constant) estimate is obtained for the Morse lemma stating that in a hyperbolic Gromov space λ-quasi-geodesic γ lies in the λ^2 neighborhood of a geodesic σ with the same ends. Moreover, this geodesic σ lies in the $\log \lambda$-neighborhood of the quasi-geodesic γ. This estimate is also optimal [16*, 14*].

- formalized the numerical problem of the quasiisometric problem, indicating several important factors that allow to obtain different results and estimates for the quasiisometric distortion. In particular, the behavior of volumes and connectivity are investigated. Then the transfer of Poincaré inequality under quasiisometric mappings is investigated, and precise upper and lower estimates for homotopic distortion growth for several classes of hyperbolic spaces are given. The properties of quasiisometric tree embeddings in the hyperbolic plane [15*] are investigated.

The following theorem is proved:

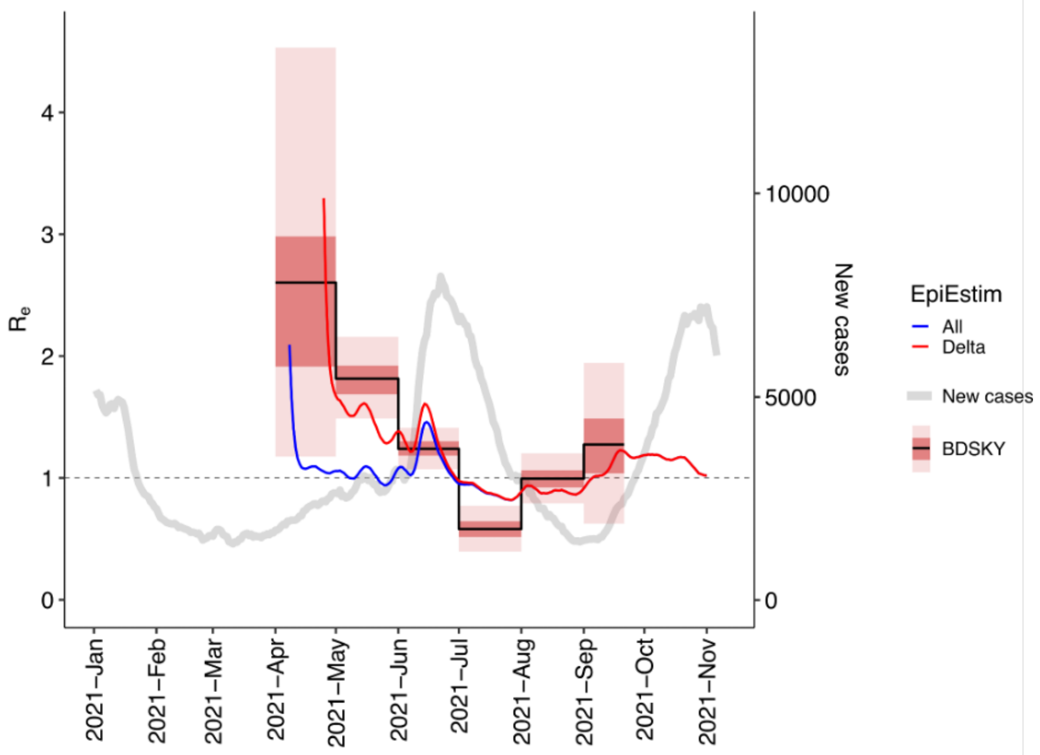


Figure 10: Dynamics of effective reproductive number R_e for the major clade of delta variant coronavirus in Moscow, estimated in the birth-death skyline model (black line; red and pink bars show 50% and 95% credible intervals, respectively) and estimated by EpiEstim package for all (blue line) or only delta variant (red line) SARS-CoV-2 cases in Moscow. The gray line shows the seven-day moving average of the daily number of new cases in Moscow regardless of genotype.

Theorem 5 *Let γ be (λ, c)-quasi-geodesic in δ-hyperbolic space E and let σ be the geodesic segment connecting its ends. Then γ lies in the H-neighborhood of σ, where*

$$H = A_1\lambda^2(c + \delta + 1),$$

and A_1 is some universal constant.

This result is optimal. In other words, we found an example of a geodesic whose furthest point lies at a distance $\lambda^2 c/4$ from the corresponding geodesic segment. Further, the following theorem, which is in some sense dual to the theorem 5, was proved.

Theorem 6 *Let γ be (λ, c)-quasi-geodesic in δ-hyperbolic space E and let σ be the geodesic segment connecting its ends. Let also $4\delta << \ln \lambda$. Then σ lies in the H_{am}-neighborhood of γ, where*

$$H_{am} = A_2(\delta \ln \lambda + \delta + c),$$

and A_2 is some universal constant.

The theorems 5 and 6 allowed us to obtain nontrivial estimates of the quasiisometric distortion for the maximum displacement of points of space X by auto-quasiisometries $X \to X$ fixing its boundary. Further, three approaches to interpreting the quantitative quasi-isometric distortion problem at the R scale have been proposed. Let X and Y be two metric spaces with base points x_0 and y_0, respectively. For a given $R > 0$, three families of mappings are considered

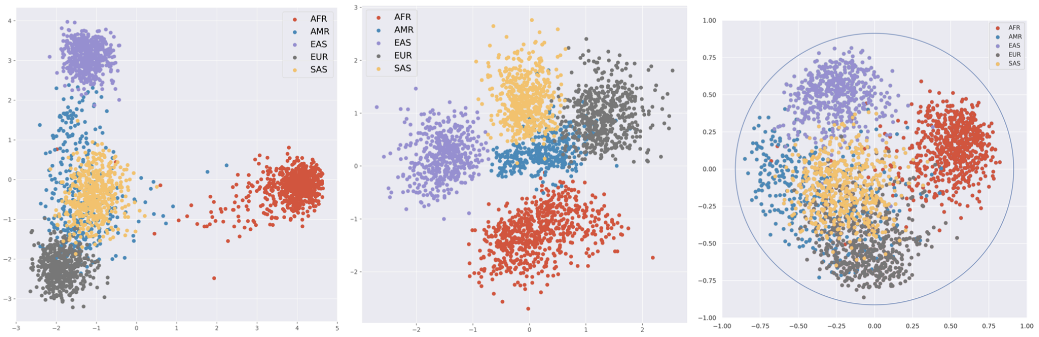
- quasi-isometries from ball $B_X(x_0, R)$ to ball $B_X(x_0, R)$,
- quasi-isometries from the ball $B_X(x_0, R)$ on the ball $B_X(x_0, \rho(R))$ for some function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$,
- quasi-isometric embedding of the ball $B_X(x_0, R)$ in Y.

The study of the transport of Poincaré inequalities by quasiisometries allowed us to obtain a lower estimate for (λ, c)-quasiisometric distortion between balls of radius R in locally homogeneous spaces of negative curvature of the form $Z_\mu = \mathbb{T}^n \times \mathbb{R}$ with metric $dt^2 + \sum_i e^{2\mu_i t} dx_i^2$ $(0 \leq \mu_1 \leq \ldots \leq mu_n)$. The following theorem is given here without technical details, the exact formulation can be found in [15*].

Theorem 7 *Any (λ, c)-quasiisometric embedding of a ball of radius R from Z_μ into $Z_{\mu'}$ satisfies the inequality*

$$\lambda + c \geq \left(\frac{\sum \mu_i}{\mu_n} - \frac{\sum \mu_i'}{\mu_n'} \right) R.$$

We applied hyperbolic geometry and deep learning to the analysis of genetic data in [17*]. Namely, we applied variational autocoders (VAE) with Euclidean and hyperbolic latent spaces to cluster the genomes of various modern human populations. Typically, the principal component method is used for this task. Variational autoencoders (VAEs) allow non-linear clustering of data, in contrast to the principal component method (PCA) widely used in population analysis. A comparison of the results of applying VAE to five populations from the 1000 Genome Project [54] is presented in Fig. 11.



(a) PCA

(b) VAE

(c) HVAE

Figure 11: Results of applying PCA, VAE with Euclidean latent space, and VAE with hyperbolic latent space (HVAE) to individuals from five macro populations from the 1000 Genomes Project.

References

[1] Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011 7;475(7357):493-6. Available from: http://www.nature.com/articles/nature10231.

[2] Schlebusch CM, Malmström H, Günter T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017;358:652-5.

[3] Visscher PM, Andrew T, Nyholt DR. Genome-wide association studies of quantitative traits with related individuals: little (power) lost but much to be gained. European Journal of Human Genetics. 2008;16(3):387-90. Available from: https://doi.org/10.1038/sj.ejhg.5201990.

[4] Comtet L. Advanced Combinatorics. Springer Netherlands; 1974. Available from: https://doi.org/10.1007%2F978-94-010-2196-8.

[5] Gravel S. Population Genetics Models of Local Ancestry. Genetics. 2012;191(2):607-19. Available from: https://www.genetics.org/content/191/2/607.

[6] Corbett-Detig R, Nielsen R. A Hidden Markov Model Approach for Simultaneously Estimating Local Ancestry and Admixture Time Using Next Generation Sequence Data in Samples of Arbitrary Ploidy. PLOS Genetics. 2017 01;13(1):1-40. Available from: https://doi.org/10.1371/journal.pgen.1006529.

[7] Kaplan NL, Hudson RR, Langley CH. The "hitchhiking effect" revisited. Genetics. 1989;123(4):887-99. Available from: https://www.genetics.org/content/123/4/887.

[8] Smith JM. What use is sex? Journal of Theoretical Biology. 1971;30(2):319 335. Available from: http://www.sciencedirect.com/science/article/pii/0022519371900580.

[9] Messer PW, Neher RA. Estimating the Strength of Selective Sweeps from Deep Population Diversity Data. Genetics. 2012;191(2):593-605. Available from: https://www.genetics.org/content/191/2/593.

[10] Durrett R, Schweinsberg J. Approximating selective sweeps. Theoretical Population Biology. 2004;66(2):129 138. Available from: http://www.sciencedirect.com/science/article/pii/S0040580904000607.

[11] Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of

Denisovan-like DNA. Nature. 2014;512(7513):194-7. Available from: https://doi.org/10.1038/nature13408.

[12] McVean GAT, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005;360(1459):1387-93. Available from: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2005.1673.

[13] Marjoram P, Wall JD. Fast "coalescent" simulation. BMC Genetics. 2006;7(1):16. Available from: https://doi.org/10.1186/1471-2156-7-16.

[14] Hedrick PW. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Molecular Ecology. 2013;22(18):4606-18. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.12415.

[15] Kermack William Ogilvy MAG, Thomas WG. Thomas A contribution to the mathematical theory of epidemics. Proceedings of Royal Society A. 1927;115:700 721.

[16] Gillespie DT. Stochastic Simulation of Chemical Kinetics. Annual Review of Physical Chemistry. 2007;58(1):35-55. PMID: 17037977. Available from: https://doi.org/10.1146/annurev.physchem.58.032806.104637.

[17] Cao Y, Gillespie DT, Petzold LR. Efficient step size selection for the tau-leaping simulation method. The Journal of Chemical Physics. 2006;124(4):044109. Available from: https://doi.org/10.1063/1.2159468.

[18] Behnel S, Bradshaw R, Citro C, Dalcin L, Seljebotn DS, Smith K. Cython: The best of both worlds. Computing in Science & Engineering. 2011;13(2):31-9.

[19] Vaughan TG, Drummond AJ. A Stochastic Simulator of Birth–Death Master Equations with Application to Phylodynamics. Molecular Biology and Evolution. 2013 03;30(6):1480-93. Available from: https://doi.org/10.1093/molbev/mst057.

[20] Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Phylodynamics with Migration: A Computational Framework to Quantify Population Structure from Genomic Data. Molecular Biology and Evolution. 2016 04;33(8):2102-16. Available from: https://doi.org/10.1093/molbev/msw064.

[21] Poon AFY. Impacts and shortcomings of genetic clustering methods for infectious disease outbreaks. Virus Evolution. 2016 12;2(2). Vew031. Available from: https://doi.org/10.1093/ve/vew031.

[22] Volz EM, Didelot X. Modeling the Growth and Decline of Pathogen Effective Population Size Provides Insight into Epidemic Dynamics and Drivers of Antimicrobial Resistance. Systematic Biology. 2018 02;67(4):719-28. Available from: https://doi.org/10.1093/sysbio/syy007.

[23] Danesh G, Saulnier E, Gascuel O, Choisy M, Alizon S. Simulating trajectories and phylogenies from population dynamics models with TiPS. bioRxiv. 2020. Available from: https://www.biorxiv.org/content/early/2020/11/09/2020.11.09.373795.

[24] Sheehan S, Harris K, Song YS. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. Genetics. 2013 7;194(3):647-62. Available from: http://www.ncbi.nlm.nih.gov/pubmed/6628982http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1202167http://www.ncbi.nlm.nih.gov/pubmed/23608192http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3697970.

[25] Schiffels S, Durbin R. Inferring human population size and separation history from multiple genome sequences. Nature genetics. 2014;46(8):919-25. Available from: http://dx.doi.org/10.1038/ng.3015.

[26] Wilton PR, Carmi S, Hobolth A. The SMC' Is a Highly Accurate Approximation to the Ancestral Recombination Graph. Genetics. 2015 03;200(1):343-55. Available from: https://doi.org/10.1534/genetics.114.173898.

[27] Wakeley J, Sargsyan O. Extensions of the coalescent effective population size. Genetics. 2009 1;181(1):341-5. Available from: `http://www.ncbi.nlm.nih.gov/pubmed/19001293http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2621185`.

[28] Kingman JFC. On the genealogy of large populations. Journal of Applied Probability. 1982;19(A):27–43.

[29] Spence JP, Steinrücken M, Terhorst J, Song YS. Inference of population history using coalescent HMMs: review and outlook. Current Opinion in Genetics and Development. 2018;53:70-6.

[30] Pool JE, Nielsen R. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics. 2009;181(2):711-9.

[31] Gravel S. Population genetics models of local ancestry. Genetics. 2012;191(2):607-19.

[32] Liang M, Nielsen R. The Lengths of Admixture Tracts. Genetics. 2014:genetics-114.

[33] Ni X, Yuan K, Yang X, Feng Q, Guo W, Ma Z, et al. Inference of multiple-wave admixtures by length distribution of ancestral tracks. Heredity. 2018;121(1):52-63.

[34] Moorjani P, Patterson N, Hirschhorn JN, Keinan A, Hao L, Atzmon G, et al. The history of African gene flow into Southern Europeans, Levantines, and Jews. PLoS genetics. 2011;7(4):e1001373.

[35] Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics. 2013;193(4):1233-54.

[36] Moreno-Mayar JV, Rasmussen S, Seguin-Orlando A, Rasmussen M, Liang M, Flåm ST, et al. Genome-wide Ancestry Patterns in Rapanui Suggest Pre-European Admixture with Native Americans. Current Biology. 2014.

[37] Gravel S, Zakharia F, Moreno-Estrada A, Byrnes JK, Muzzio M, Rodriguez-Flores JL, et al. Reconstructing native American migrations from whole-genome and whole-exome data. PLoS genetics. 2013;9(12):e1004023.

[38] Bennett J. On the theory of random mating. Annals of Eugenics. 1952;17(1):311-7.

[39] Slatkin M. On treating the chromosome as the unit of selection. Genetics. 1972;72(1):157-68.

[40] Sanchez T, Cury J, Charpiat G, Jay F. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. Molecular Ecology Resources. 2021;21(8):2645-60. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13224.

[41] Kelleher J, Etheridge AM, McVean G. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. PLOS Computational Biology. 2016 05;12(5):1-22. Available from: https://doi.org/10.1371/journal.pcbi.1004842.

[42] Bhatia G, Tandon A, Patterson N, Aldrich MC, Ambrosone CB, Amos C, et al. Genome-wide Scan of 29,141 African Americans Finds No Evidence of Directional Selection since Admixture. The American Journal of Human Genetics. 2014;95(4):437-44. Available from: https://www.sciencedirect.com/science/article/pii/S0002929714003553.

[43] Tataru P, Simonsen M, Bataillon T, Hobolth A. Statistical Inference in the Wright–Fisher Model Using Allele Frequency Data. Systematic Biology. 2016 08;66(1):e30-46. Available from: https://doi.org/10.1093/sysbio/syw056.

[44] Corbett-Detig R, Jones M. SELAM: simulation of epistasis and local adaptation during admixture with mate choice. Bioinformatics. 2016 06;32(19):3035-7. Available from: https://doi.org/10.1093/bioinformatics/btw365.

[45] Jeong C, Alkorta-Aranburu G, Basnyat B, Neupane M, Witonsky DB, Pritchard JK, et al. Admixture facilitates genetic adaptations to high altitude in Tibet. Nature Communications. 2014;5(1):3281. Available from: https://doi.org/10.1038/ncomms4281.

[46] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences.

2013;110(1):228-33. Available from: https://www.pnas.org/content/110/1/228.

[47] Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLOS Computational Biology. 2019 04;15(4):1-28. Available from: https://doi.org/10.1371/journal.pcbi.1006650.

[48] Vaughan TG, Nadeau SA, Sciré J, Stadler T. Phylodynamic Analyses of outbreaks in China, Italy, Washington State (USA), and the Diamond Princess. Virological. 2020 03. Available from: https://virological.org/t/phylodynamic-analyses-of-outbreaks-in-china-italy-washington-state-usa-and-the-diamond-princess/439.

[49] Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020. Eurosurveillance. 2020;25(10). Available from: https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2020.25.10.2000180.

[50] Zhang S, Diao M, Yu W, Pei L, Lin Z, Chen D. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. International Journal of Infectious Diseases. 2020;93:201-4. Available from: https://www.sciencedirect.com/science/article/pii/S1201971220300916.

[51] Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, et al. Haplotype networks of SARS-CoV-2 infections in the <i>Diamond Princess</i> cruise ship outbreak. Proceedings of the National Academy of Sciences. 2020;117(33):20198-201. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.2006824117.

[52] Deng X, Gu W, Federman S, du Plessis L, Pybus OG, Faria NR, et al. Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California. Science. 2020;369(6503):582-7. Available from: https://www.science.org/doi/abs/10.1126/science.abb9263.

[53] Lemieux JE, Siddle KJ, Shaw BM, Loreth C, Schaffner SF, Gladden-Young A, et al. Phylogenetic analysis of SARS-CoV-2 in the Boston area highlights the role of recurrent importation and superspreading events. medRxiv. 2020. Available from: `https://www.medrxiv.org/content/early/2020/08/25/2020.08.23.20178236`.

[54] Consortium GP, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68.