Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский университет
«Высшая школа экономики»

На правах рукописи

Щур Владимир Львович

Математические модели и анализ данных в популяционной геномике

РЕЗЮМЕ ДИССЕРТАЦИИ

на соискание ученой степени доктора наук
по прикладной математике

Москва – 2023

Диссертационная работа выполнена в международной лаборатории статистической и вычислительной геномики в Национальном исследовательском университете «Высшая школа экономики».

1 Введение

1.1 Актуальность

Геномика - новая междисциплинарная наука, возникшая на стыке генетики, математики и компьютерных наук. Популяционная и эволюционная генетика является одним из важных разделов в этой области. В связи с быстрым удешевлением технологий секвенирования и генотипирования появляется всё больше генетических данных, доступных для анализа, содержащих информацию о процессах развития популяций. Геномы содержат в себе информацию об истории и структуре популяции, об эволюционных факторах и механизмах естественного отбора. Например, за последние 15 лет из геномных данных получено множество новых данных об истории расселения человечества по Земле, примешивании древних людей (неандертальцев и денисовцев) к предкам современного человека, его адаптации к различным климатическим условиям и географическим территориям. С другой стороны, пандемия коронавируса SARS-CoV-2 показала важность геномного эпидемиологического надзора в реальном времени. К ноябрю 2022 года в базе данных GISAID уже доступно около 13.5 миллионов образцов коронавируса. Эти данные позволяют отслеживать пути распространения, обнаруживать новые варианты вируса, изучать его эволюцию. Таким образом, разработка новых математических моделей и методов анализа генетических данных является актуальной и востребованной проблемой.

В диссертационной работе представлены теоретические результаты в области популяционной и эволюционной генетики, новые математические модели, разработаны методы анализа генетических данных, а также получены результаты анализа экспериментальных данных. Результаты работы расширяют арсенал методов для исследований в области популяционной и эволюционной генетики, позволяя выяснять более детальную и точную картину истории развития популяций, получать новые знания об эволюционных процессах и адаптации различных видов животных и вирусов. Разработанные модели и методы дают возможность оценивать из генетических данных такие основополагающие параметры как скорость миграции, пропорции примешивания (при однократных, пульсовых, миграциях), время разделения и примешивания популяций, силу естественного отбора. Новые алгоритмы и программное обеспечение отвечает требованиям современных и перспективных задач геномики,

требующих обработки большого объема данных. Например, разработанный нами метод `MiSTI` [18*][1] позволяет одновременно оценивать время разделения популяций и скорости миграции, основываясь на оценках исторического эффективного размера популяций, полученных, например, методом `PSMC` [1]. Используя наш метод, мы оспорили результат [2] о глубоком разделении (260-350 тысяч лет назад) между африканскими популяциями бушменов и динка, получив новую оценку в ≈ 107 тысяч лет назад и односторонней миграцией из динка в бушменов, то есть примерно в $2.5 - 3.5$ раза меньше, чем сообщалось ранее. Наши результаты также подтверждены проведенным нами компьютерным моделированием.

1.2 Степень разработанности

Представленные в работе теоретические и методологические результаты позволяют получать новые знания в области популяционной и эволюционной генетики и геномной эпидемиологии, которые невозможно было получить ранее существовавшими методами, а также закладывают основу для разработки новых, ещё более точных и эффективных методов.

Получены следующие *теоретические результаты*: изучено число p-сестер в большой выборке из диплоидной популяции Райта-Фишера, изучено распределение длин участков хромосом при адаптивной интрогрессии, исследована точность приближения моделью SMC' структурированной коалесценции с рекомбинацией, формализовано и изучено понятие локального эффективного размера популяции, построена модель неравновесного сцепления трех локусов при примешивании двух популяций, исследована количественная задача квазиизометричности гиперболических пространств.

Благодаря этим теоретическим результатам *разработаны методы*, открывающие новые возможности для анализа экспериментальных генетических данных. Так, метод для предсказания адаптивной интрогрессии позволяет обнаружить один из важнейших механизмов адаптации популяций. Метод вычисления локального эффективного размера популяции и оценки скоростей миграции и времени разделения популяций и ме-

[1]Здесь и далее звездочка означает публикации из списка, представленного соискателем к защите в разделе 1.9.

тод оценки времен множественного примешивания позволяют уточнить данные об истории развития популяций, в частности изучить процессы миграции в разных масштабах времени. Метод предсказания исторического размера популяции при помощи глубинного обучения имеет важное методологическое значение, открывая возможность в перспективе использовать глубинное обучение для полногеномного анализа ДНК. Это особенно актуально для задачах, в которых вероятностные методы вычислительно неэффективны. Новый метод компьютерного моделирования вирусных генеалогий позволяет генерировать датасеты размерами равными и превосходящими текущие экспериментальные данные геномов коронавируса SARS-CoV-2. Разработанное ПО необходимо для валидации существующих и разрабатываемых методов обработки и анализа в области геномной эпидемиологии. Предложено примененять вариационные автокодировщики с евклидовым и гиперболическим латентными пространствами в качестве метода кластеризации и визуализации данных для популяционного анализа.

Проведен *анализ экспериментальных данных*: изучен естественный отбор у чилийцев после примешивания коренной, европейской и африканской популяции; оценено время разделения человеческих популяций, оспорен результат о глубоком времени разделения популяций бушменов и динка; изучена филодинамика коронавируса в НИИ им. Вредена (март-апрель 2020), клады AY.122+ORF7a:P45L дельта в Москве (апрель-сентябрь 2021).

1.3 Цели и задачи исследования

Целью исследования является разработка новых математических моделей, методов, алгоритмов и программного обеспечения для изучения популяционной и эволюционной истории из геномных последовательностей, а именно для изучения процессов разделения и примешивания популяций, определение адаптивных участков генома и силы естественного отбора, оценки изменений исторического размера популяции.

Задачами исследования являются:

- Изучить число особей с p-сестрами в большой выборке из популяции.

- Построить математическую модель адаптивной интрогрессии для точного и эффективного вычисления распределения длин участков хромосом.
- Разработать метод для предсказания адаптивной интрогрессии на основе построенной математической модели.
- Разработать симулятор вирусных генеалогий, масштабируемый на реалистичные размеры выборок, полученных в течение пандемии.
- Исследовать точность приближения структурированной коалесценции с рекомбинацией моделью SMC'.
- Разделить эффекты генетического дрейфа (локального эффективного размера популяции) и миграции, разработать метод для вычисления локального эффективного размера популяции из исторического эффективного размера популяции для представителей двух популяций, а также для оценки скоростей миграции и времени разделения популяций.
- Построить математическую теорию неравновесного сцепления трех локусов, разработать метод на основе этой теории для оценки времен множественного примешивания.
- Изучить естественный отбор в чилийской популяции после постколумбова примешивания.
- Разработать метод для предсказания изменений эффективного размера популяции при помощи глубинного обучения.
- Изучить филодинамику коронавируса SARS-CoV-2 в России.
- Разработать теорию и применить гиперболическую геометрию для анализа генетических данных в популяционной генетике.

1.4 Методы исследования

Методы исследования включают в себя использование и развитие популяционных моделей (модель Райта-Фишера и её обобщения, коалесцентная модель, секвенциальная марковская коалесценция, компартментные эпидемиологические модели), вероятностные подходы, скрытая марковская модель, алгоритм Гиллеспи (включая приближенный алгоритм τ-leaping), глубинное обучение и геометрические методы анализа данных. Программное обеспечение реализовано на языках Python и C/C++ (включая технологию cython). Также использовались существующие методы популяционной и эволюционной геномики: PSMC, Admixture, BEAST2. Для компьютерных вычислений использовался высокопроизводительный кластер НИУ ВШЭ.

1.5 Теоретическая и практическая значимость

Теоретическая значимость состоит в разработке математической теории в области популяционной и эволюционной генетики, в частности получены новые результаты для моделей Райта-Фишера и коалесцентной модели. Также было разработано несколько методов анализа данных и алгоритмов, использующих эти новые теоретические результаты. Были проанализированы экспериментальные данные при помощи новых и существующих методов, в частности получены новые знания об истории человеческой популяции, распространения коронавируса SARS-CoV-2 в России. Практическая значимость состоит в разработке программного обеспечения, реализующего новые методы и алгоритмы анализа генетических данных и компьютерного моделирования популяций. Все разработанное программное обеспечение доступно в открытом доступе в репозиториях GitHub.

1.6 Результаты, выносимые на защиту

- *О числе p-сестер в большой выборке из популяции [13*].* Выведена асимптотическая формула для математического ожидания доли особей в выборке размера K из популяции размера N, не имеющих p-сестер в этой выборке, при $N \to \infty$ and $K/N = const$. Формулы

получены для моногамной и для немоногамной диплоидных моделей Райта-Фишера. Показано, что для больших выборок, размер которых сопоставим с некоторыми масштабными исследованиями в области генетики, нельзя пренебрегать близкородственными связями. Результат важен при планировании, например, GWAS (полногеномный поиск ассоциаций) проектов с большими когортами.

- *Математическая модель адаптивной интрогрессии [10*].* Разработана математическая модель адаптивной интрогрессии, где траектория частоты аллеля, находящегося под естественным отбором, моделируется при помощи детерминированной логистической кривой. Модель является вычислительно эффективной, при этом точна в широком диапазоне параметров адаптивной интрогрессии. Также показано, что этот диапазон можно расширить на случаи, где логистическое приближение неточно из-за генетического дрейфа, численно оценивая среднюю траекторию адаптивного аллеля. Модель позволила разработать два метода (метод вычисления распределения длин участков хромосом при адаптивной интрогрессии и метод предсказания адаптивной интрогрессии), что в свою очередь открывает новые возможности для изучения адаптации у различных видов животных, включая человека.

- *Метод предсказания адаптивной интрогрессии [6*].* Теоретическая модель адаптивной интрогрессии легла в основу метода на основе скрытой марковской модели для поиска и оценки параметров адаптивной интрогрессии. При помощи компьютерного моделирования показано, что метод точен для многих реалистичных сценариев на датасетах среднего размера. Метод аппробирован на Drosophila melanogaster из Южной Африки, найдено 17 локусов со значимым уровнем отбора, из которых 4 локуса ранее были ассоциированы с устойчивостью к пестицидам. Ожидается, что разработанный метод будет широко востребован научным сообществом для обнаружения адаптационных процессов в популяциях различных видов животных.

- *Симулятор вирусных генеалогий [2*].* Разработан программный пакет `VGsim` для моделирования эпидемий и возникающих при этом вирусных генеалогий. Функционал программного пакета включает

в себя моделирование эпидемиологических, эволюционных и популяционных аспектов. Симуляция развития эпидемии основано на алгоритме Гиллеспи, симуляция генеалогий - на структурированной коалесценции, обусловленной эпидемиологической динамикой. Программный пакет является самым быстрым известным нам решением в области геномной эпидемиологии. Он позволяет симулировать генеалогии десятков миллионов образцов в сложных эпидемиологических сценариях, что превосходит текущие размеры базы GISAID. Это делает `VGsim` перспективным решением для валидации результатов анализа данных и новых методов анализа данных в геномной эпидемиологии.

- *Точность приближения структурированной коалесценции с рекомбинацией моделью SMC' [1*]* Исследована точность приближения коалесценции с рекомбинацией моделью SMC' в случае двух популяций с миграцией. Проанализирована полная вариация разности между совместными распределениями времен до общего предка двух локусов в моделях коалесценции с рекомбинацией и SMC' как функция от генетического расстояния между этими локусами. Показано, что для двух популяций с миграцией полная вариация убывает существенно медленнее, чем в случае однородной популяции. Это показывает, что при наличии структуры популяции, методы анализа данных, основанные на модели SMC', могут приводить к неверным результатам.

- *Эффективный размер популяции и миграция [18*].* Формализовано понятие локального эффективного размера популяции для сценария с двумя популяциями и миграцией между ними. Изучено влияние миграции на оценку размера популяции методом PSMC. На основе разработанной математической теории разработан метод для вычисления локального эффективного размера популяции и для оценки времени разделения популяций и скоростей миграции между ними. Работа имеет важное методологическое значение для теории структурированной коалесценции, а также позволяет точно реконструировать историю потока генов между популяциями.

- *Множественное примешивание и неравновесное сцепление трех локусов [8*].* Построена математическая теория неравновесного сцепления трех генетических локусов при примешивании популяций. На

его основе были разработаны метод и программное обеспечение для оценки времен примешивания между двумя популяциями при двух пульсах миграции. Разработанный метод позволяет точно исследовать недавнюю (в пределах нескольких десятков поколений) историю примешивания популяций в сложных сценариях, для которых существовавшие ранее методы были неприменимы или неточны.

- *Естественный отбор в чилийской популяции после пост-колумбова примешивания [11*].* При помощи компьютерного моделирования проверены результаты поиска естественного отбора после перемешивания коренного населения, европейцев и африканцев в Чили на основе предсказания локального происхождения. Показана состоятельность выбранного статистического критерия для поиска участков генома, находящихся под естественным отбором. Это позволило успешно и достоверно исследовать адаптационные процессы в современной чилийской популяции.

- *Глубинное обучение для демографического анализа [4*, 3*].* Разработан метод на основе глубинного обучения для предсказания локальных времен до общего предка вдоль диплоидного генома. Метод может быть использован также для предсказания траектории эффективного размера популяций аналогично методу PSMC [1]. Работа имеет важное методологическое значение для дальнейшего развития методов глубинного обучения для анализа полногеномных последовательностей.

- *Филодинамика коронавируса SARS-CoV-2 в России [9*, 7*].* Проведен байесовский филодинамический анализ вспышки Covid-19 в НИИ травматологии им. Вредена (Санкт Петербург) в марте-апреле 2020 года, а также клады (AY.122+ORF7a:P45L) варианта дельта в апреле-сентябре 2021 года в Москве с использованием программного пакета BEAST2. В первом исследовании показано, что внутрибольничная вспышка явилась результатом не менее двух, вероятно, трех заносов коронавируса в больницу. Во втором исследовании независимо от эпидемиологических данных подтверждено, что основная клада (AY.122+ORF7a:P45L) ответственна за летнюю эпидемическую волну в 2021 году, и, вероятно, за последовавшую осеннюю волну. Результаты дают объективную картину распространения коронавируса SARS-CoV-2 в России, что важно

при анализе принимаемых эпидемиологических мер при борьбе с пандемий.

- *Гиперболическая геометрия и анализ генетических данных [14*-17*].* Поставлена и исследована численная задача квазиизометричности гиперболических пространств. Рассмотрено применение вариационных автокодировщиков с гиперболическим латентным пространством для задачи визуализации генетического разнообразия популяций (аналогично методу главных компонент). Эти результаты имеют как фундаментальную математическую значимость, так и открывают возможность для разработки и применения принципиально новых подходов в популяционной генетике.

1.7 Новизна и достоверность

Все научные результаты, выносимые на защиту, являются новыми. Предложена новая математическая модель для распределения длин участков хромосом при адаптивной интрогрессии. Предложены новые методы для предсказания адаптивной интрогрессии, компьютерного моделирования вирусных генеалогий, оценке времени разделения и скоростей миграции между популяциями, оценке времен множественного примешивания из неравновесного сцепления трех локусов, оценке исторического эффективного размера популяции при помощи глубинного обучения. При помощи этих и существующих методов решены следующие задачи: оценены времена разделения между человеческими популяциями и оспорен результатат о глубоком времени разделения между африканскими популяциями бушменов и динка, оценены времена примешивания при формировании современных популяций мексиканцев и колумбийцев, изучена адаптация у чилийцев после пост-колумбова примешивания, изучена филодинамика коронавируса SARS-CoV-2 в России.

Достоверность результатов обосновывается тем, что все результаты, выносимые на защиту, опубликованы в ведущих рецензируемых научных журналах, индексируемых в научных базах Web of Sciences и Scopus с квартилями Q1 - 13 статей, Q3 - 2 статьи, из них 3 статьи опубликованы в журналах из списка Nature Index. Программные коды опубликованы в репозиториях GitHub, как 7 программных комплексов.

1.8 Апробация полученных результатов

Основные результаты диссертации докладывались на следующих международных конференциях и семинарах:

- *Estimating the timing of multiple admixture events using 3-locus Linkage Disequilibrium*, конференция Moscow Conference on Computational Molecular Biology (MCCMB'21), июль 2021, Москва, Россия.
- *Deep learning for demographic inference from whole-genome sequences*, конференция Moscow Conference on Computational Molecular Biology (MCCMB'21), июль 2021, Москва, Россия.
- *ngsPSMC: genotype likelihood-based PSMC for analysis of low coverage NGS data*, конференция Probabilistic Modeling in Genomics, октябрь 2019, Оссуа, Франция.
- *ngsPSMC: genotype likelihood-based PSMC for analysis of low coverage NGS data*, конференция Moscow Conference on Computational Molecular Biology (MCCMB'19), июль 2019, Москва, Россия.
- *Estimation of population split times and migration rates with variable population sizes*, конференция Probabilistic Modeling In Genomics, октябрь 2018, Колд Спринг, США.
- *ngsPSMC: modifying PSMC to work with NGS data"*, UCCGC workshop, 15–18 августа 2017, Blue Oak Ranch Reserve, США.
- *Tree consistent PBWTs and their application to reconstructing ancestral recombination graphs and demographic inference*, конференция Probabilistic Modeling in Genomics, октябрь 12–17 2015, Колд Спринг, США.
- *Tree consistent PBWT and their application to reconstructing Ancestral Recombination Graphs and demographic inference*, Recomb 2015, Варшава, Польша. Best poster award.
- *On modern problems and methods for data analysis in human genomics*, Computer Simulation in Physics and beyond 2015, Москва, Россия, пленарный доклад

- *Tree consistent PBWT and their application to reconstructing ancestral recombination graphs and population structure inference*, Biology of Genomes, 10–14 мая 2015, Колд Спринг, США
- *Extension of PBWT and its connection with ARG*, конференция International meeting on genomics, апрель 2014, Доха, Катар.

1.9 Список статей, представленных к защите по теме диссертации (с указанием личного вклада соискателя)

Работы, опубликованные автором в рецензируемых научных изданиях, входящих в международную систему цитирования Scopus

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

Исследована модель SMC', аппроксимирующая коалесценцию с рекомбинацией для случая двух популяций с миграцией. Показано, что полная вариация между совместным распределением времени до ближайшего общего предка в двух локусах с ростом генетического расстояния между локусами убывает существенно медленнее, чем в случае однородной популяции.

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`

Разработан симулятор вирусных генеалогий `VGsim`, являющийся самым быстрым решением в своей области. Разработаны математическая модель, программно реализован алгоритм, а также руководство всеми стадиями исследования и разработки остальных частей ПО.

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

Предложен и исследован метод предсказания локальных времен до общего предка вдоль генома, а также их частного вероятностного распределения, при помощи глубинного обучения.

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

Предложен прототип метода предсказания локальных времен до общего предка вдоль генома.

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

Выполнен пермутационный анализ для изучения параллельной высотной адаптации у ящериц Phrynocephalus Тибетского плато.

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

Разработана скрытая марковская модель для адаптивной интрогрессии. Предложен подход для приближенного быстрого вычисления переходных вероятностей вблизи адаптивного локуса.

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

Проведен филодинамический анализ клады Y.122ORF7a:P45L варианта коронавируса в Москве в апреле-сентябре 2021 года.

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

Разработана математическая модель неравновесного сцепления примешивания трех локусов для двух популяций.

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

Проведен филодинамический анализ внутрибольничной вспышки коронавируса SARS-CoV-2 в НИИ травматологии им. Вредена в марте-апреля 2020 года.

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

Построена математическая модель для примешанных участков генома при адаптивной интрогрессии на основе коалесцентной теории и приближения траектории частоты аллеля под отбором детерминированной логистической кривой.

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

При помощи компьютерного моделирования проведена верификация статистического метода для поиска генов под отбором в чилийской популяции после примешивания.

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

Предложена архитектура скрытой марковской модели для определения сегментов древнего происхождения при помощи внешней популяции без примешивания. Эта архитектура позволила обнаружить сегменты неандертальской и денисовской ДНК в представителях Папуа Новая Гвинея с высокой точностью.

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

Выведены формулы для математического ожидания числа особей без p-сестер в выборке из популяций в моногамной и немоногамной моделях Райта-Фишера, а также найдено асимптотическое поведение в зависимости от доли выборки от размера N популяции при $N \to \infty$.

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

Исправлено доказательство количественной версии леммы Морса о расстоянии от квазигеодезического до геодезическго сегмента в гиперболическом пространстве.

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

Исследованы количественные свойства квазиизометрий: рассмотрен перенос неравенств Пуанкаре, получены точные оценки роста квазиизометрического искажения для некоторого класса гиперболических метрических пространств. Также доказана линейность роста квазиизометрического искажения между гиперболическим пространством $\mathbb{H}^n$ и бинарным деревом.

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

Представлена количественная версия леммы Морса, двойственная лемма анти-Морса, исследованы квазиизометрии, фиксирующие границу на бесконечности.

Публикации и препринты автора в других изданиях

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

Предложен метод популяционной кластеризации генетических данных на основе вариационных автокодировщиков с евклидовым и гиперболическим латентными пространствами.

18.* Препринт 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

Разделены понятия исторического и локального эффективных размеров популяции. Разработан метод для оценки времени разделения и скоростей миграций совместно с вычислением локального эффективного размера популяции из исторических эффективных размеров популяция представителей двух популяций.

2 Результаты

В этом разделе мы последовательно излагаем детали полученных в диссертации основных результатов проведенного научного исследования.

2.1 О числе p-сестер в большой выборке из популяции

По мере удешевления методов геномного секвенирования и генотипирования увеличиваются объёмы анализируемых данных. В исследованиях по полногеномному поиску ассоциаций (GWAS) родственные особи обычно удаляются из выборки, но существуют и другие стратегии использования родства в качестве ковариаты в статистическом анализе (например, [3]). В связи с этими наблюдениями возникает следующий вопрос. Сколько близких родственников мы ожидаем найти в выборке при данном эффективном размере популяции? Ответ на этот вопрос может помочь определить структуру исследований и стратегии для решения проблемы родства в популяционных выборках, в частности для GWAS. Особый интерес представляет число лиц в выборке без родственников, т.е. число лиц, остающихся в выборке, если удалить лиц с родственниками.

В настоящем разделе мы представляем результаты о числе близких родственников в двух диплоидных моделях Райта-Фишера - моногамной и полностью немоногамной из работы [13*]. Мы будем использовать эти модели для получения распределения и математического ожиданий числа особей, имеющих родных, двоюродных, троюродных и т.д. сестер и братьев в выборке.

Для моделирования популяции с двумя родителями мы используем два обобщения модели Райта-Фишера. Первое обобщение - моногамная

модель Райта-Фишера, в которой пары родителей фиксируются. Второе обобщение - немоногамная модель Райта-Фишера, в которой для каждой особи каждый из двух родителей выбирается независимо из множеств мужских и женских особей соответственно.

Сестрами и братьями (далее для краткости будем писать просто сестрами) будем называть особей с общими родителями.

Далее, будем считать, что в каждом поколении популяции ровно по N мужских и женских особей. Будем обозначать через G_0 наблюдаемое (современное) поколение и нумеровать поколения в обратном по времени порядке, то есть особи поколения G_i являются родителями для поколения G_{i-1}.

Рассмотрим случайную выборку S из поколения G_0. Обозначим через U_T (для моногамной модели) и V_T (для немоногамной модели) число особей в выборке S, у которых нет T-сестер и T-полусестер в S и в генеалогии которых нет циклов (то есть нет инбридинга). Вероятность возникновения циклов мала, если 2^T (число предков в поколении T без инбридинга) много меньше эффективного размера популяций. Таким образом, в интересующих нас случаях инбридингом можно пренебречь.

Ниже мы дадим формулы для распределений и математических ожиданий чисел сестер U_1 в моногамной модели и полусестер V_1 в немоногамной модели. Также мы выведем асимптотическую формулу для математического ожидания U_T и V_T при фиксированной доле выборки и большом размере популяции.

Напомним, что число Стирлинга второго рода $S(n,k)$ равно числу разбиений множества из n элементов на k непустых подмножеств. r-ассоциированное число Стирлинга второго рода $S_r(n,k)$ [4] - это число разбиений множества из n элементов на k непустых подмножеств размера не менее r.

2.1.1 Результаты для моногамной модели Райта-Фишера

Рассмотрим особей в случайной выборке S размера K, таких что в эту же выборку S не попали их сестринские особи. Нас интересует распределение числа таких особей, его математическое ожидание и асимптотика при большом размере популяции и фиксированной доле выборки S от совокупной популяции (то есть при фиксированном отношении K/N и пределе $N \to \infty$).

Теорема 1 *Пусть U_1 - случайная величина, обозначающая число особей*

в случайной выборке S размера K, не имеющих сестер в той же выборке S, в моногамной модели Райта-Фишера. Тогда

- *распределение вероятностей для U_1 имеет следующий вид*

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

- *математическое ожидание U_1 равно*

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

- *более того, если $K/N = \alpha$, то*

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

Для числа U_2 особей из выборки, не имеющих в этой же выборке $2-$сестер (то есть без двоюродных сестер) мы нашли математическое ожидание и асимптотику.

Теорема 2 *Пусть U_2 - случайная величина, равная числу особей в выборке S без $2-$сестер, также попавших в выборку S, в моногамной модели Райта-Фишера. Тогда математическое ожидание U_2 равно*

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

Более того, если $K/N = \alpha$, то

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

Наши результаты обобщаются для произвольной степени родства, то есть для числа U_p особей в выборке S, таких что ни одна их p-сестра не попала в S.

Теорема 3 • *Для любого натурального числа $p \geq 1$ математическое ожидание числа U_p равно*

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

где использованы следующие обозначения $2m_0 := K$,

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

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

и

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

• *Если $K/N = \alpha$ $(i = 1, 2, \ldots, p)$, то*

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

2.1.2 Результаты для немоногамной модели Райта-Фишера

Мы получили аналогичные результаты для случая немоногамной модели Райта-Фишера. Однако в отличие от случая моногамной модели, вероятность того, что две особи являются полными p-сестрами мала по сравнению с вероятностью быть p-полусестрами. Итак, в этом случае нас будет интересовать число V_p особей в выборке S чьи p-полусестры и полные p-сестры не попали в эту выборку.

Теорема 4 • *Для любого натурального $p \geq 1$, математическое ожидание от V_p равно*

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

где мы полагаем $m_0 = K$ и

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

и

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

• *Если $K/N = \alpha$, то*

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

В частности, для V_1 имеем

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

Как следствие, заключаем, что поведение U_p и V_p связано следующим образом

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

2.2 Моделирование адаптивной интрогрессии

2.2.1 Адаптивная интрогрессия

В работе [10*] мы разработали математическую модель, позволяющую эффективно и точно численно оценивать распределение длин предковых участков хромосом вокруг генетического локуса, находящегося под

действием естественного отбора. Далее, в работе [6*] мы использовали нашу теоретическую модель для разработки метода на основе скрытой марковской модели для поиска адаптивной интрогрессии и оценке её параметров.

2.2.2 Описание модели

Перейдем к изложению нашего подхода. Рассмотрим случайный процесс вдоль хромосомы с двумя состояниями, соответствующими двум предковым популяциям. Переходы между предковыми состояниями возникают вследствие рекомбинаций, при этом участки хромосомы слева и справа от точки рекомбинации происходят из разных популяций. Мы рассматриваем модель с тремя генетическими локусами. Таким образом, мы вычисляем вероятности переходов между предковыми состояниями в двух локусах в зависимости от расстояния до третьего локуса, находящегося под отбором.

Сформулируем более детально рассматриваемую популяционную модель. Пусть α - генетический локус под отбором с двумя возможными аллелями: аллель A под отбором и нейтральный аллель a. Мы рассматриваем сценарий, при котором аллель A попадает в популяцию вследствие адаптивной интрогрессии. То есть особи из одной популяции единократно заменяют определенную долю особей из второй популяции. Такую интрогрессию часто называют предковым пульсом [5, 6]. Также мы делаем допущение, что в момент интрогрессии все особи донорской популяции являются носителями аллеля A, а все особи популяции-получателя имеют аллель a.

Мы описываем частоту аллеля детерменированной траекторией. В зависимости от значений параметров адаптивной интрогрессией используется либо логистическая кривая (см., например, [7]), либо численно оцененная кривая, получаемая усреднением по большому числу случайных реализаций траектории при условии, что не происходит фиксации одного из двух аллелей. Мы использовали второй подход для анализа Денисовской интрогрессии в предков Тибетцев, где доля примешивания оценивается всего лишь в 0.06% [8].

2.2.3 Построение детерминированной приближенной модели

Итак, пусть нам дана наблюдаемая хромосома. Наша цель - описать переходы (вдоль генома) между предковыми состояниями локусов. В подходе коалесценции с рекомбинацией (с направлением времени из настоящего в прошлое) наша модель является марковской при условии фиксированной траектории частоты аллеля. Для описания динамики адаптивной интрогрессии с тремя локусами нам необходимо сначала перечислить все возможные состояния, соответствующие предковым конфигурациям трех локусов. Далее, мы приближаем этот процесс марковским процессом, направленным уже вдоль генома (модели SMC/SMC' [9, 10]), где состояния в локусах наблюдаемой хромосомы будут соответствовать одной из двух предковых популяций.

Модель имеет шесть возможных состояний. Каждое состояние соответствует одной из предковых конфигураций хромосомы, состоящей из трех локусов, которые обозначим α, β и γ. В локусе α мы отслеживаем аллель, A или a, которая также определяет происхождение этого локуса. В локусах β и γ нам необходимо лишь знать, является ли хромосома предковой для наблюдаемой хромосомы или нет. Будем использовать обозначение β^a и γ^a, если ДНК в локусах β и γ соответственно является предковой для наблюдаемой хромосомы. β^n и γ^n используются для обозначения того, что в локусах β и γ предковой хромосомы ДНК не является предковой для наблюдаемой хромосомы. Таким образом, в модели есть следующие шесть состояний: $X_1 = (A-\beta^a-\gamma^a)$, $X_2 = (A-\beta^a-\gamma^n, A-\beta^n-\gamma^a)$, $X_3 = (A-\beta^a-\gamma^n, a-\beta^n-\gamma^a)$, $X_4 = (a-\beta^a-\gamma^n, A-\beta^n-\gamma^a)$, $X_5 = (a-\beta^a-\gamma^n, a-\beta^n-\gamma^a)$, $X_6 = (a-\beta^a-\gamma^a)$.

Обозначим частоту аллеля под естественным отбором во время t через $\omega(t)$. Как мы ранее указали, мы полагаем, что частота $\omega(t)$ детерминированно следует логистической траектории (с точностью до симметрии относительно вертикальной оси и сдвига вдоль горизонтальной):

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

с учетом направления времени из настоящего в прошлое.

Рекомбинации действуют с частотой, пропорциональной генетическому расстоянию между локусами, иначе говоря, расстояние измеряется в Морганах. Обозначим частоту рекомбинаций между локусам α и β через r_1, а частоту рекомбинаций между локусами β и γ через r_2.

Положим $\lambda = 1/2N_e$, где $2N_e$ - гаплоидный эффективный размер популяции. Переходы в марковском процессе соответствуют двум типам: коалесценции и рекомбинации. Коалесценции могут происходить только между хромосомами с одним аллелем в локусе α. Таким образом, матрица переходов марковского процесса задается следующим образом

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

где $\bar{\omega}(t) = 1 - \omega(t)$ является частотой аллеля a. Итак, наша система описывается уравнением Колмогорова

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

Начальным условием для этого уравнения, соответствующим динамике интрогрессировавшего участка (то есть с аллелем A), является

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

а для участка из популяции-получателя (с аллелем a) начальным условием является

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

2.2.4 Частоты переходов между предковыми состояниями вдоль хромосомы

В нашей модели вероятность того, что локус имеет происхождение типа 1 (донорская популяция) или типа 0 (популяция-получатель) равна вероятности того, что предковая хромосома несет аллель A или a соответственно в момент интрогрессии.

Рассмотрим новый марковский процесс, который описывает происхождение локуса при движении вдоль хромосомы при удалении от адаптивного локуса. Это является лишь приближением (см. модель SMC/SMC' [9, 10]), так как на самом деле указанный процесс не является марковским. Состояния этого процесса - происхождение типа 0 и 1. По определению, частоты переходов между состояниями s_1 и s_2 в позиции r для этого марковского процесса

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

Так, частота перехода $\tau_{10}(r)$ происхождения типа 1 в тип 0, соответствующего окончанию интрогрессировавшего участка, равна

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

Числитель $P(S(r + dr) = 0, S(r) = 1)$ является вероятностьи $P(X_3)$, а знаменатель $P(S(r) = 1)$ равен $P(X_1) + P(X_2) + P(X_3)$. Выражение (5) можно легко оценить численно для достаточно малых значений r_2.

2.2.5 Численные результаты

Мы продемонстрировали, что наша модель точно моделирует распределение длин трактов, в отличие от экспоненциального распределения. Так, на рисунке 1 показано распределение расстояния от адаптивного локуса до одного из концов интрогрессировавшего участка. Это распределение было оценено при помощи компьютерного моделирования, с помощью нашей детерминированной приближенной модели и экспоненциального распределения с параметром, обратно пропорциональным среднему численно смоделированного распределения (т.е. оценкой математического ожидания). На рисунке также представлены QQ-графики для всех трех пар распределений. Причина, почему экспоненциальное распределение неточно моделирует распределение длин интрогрессировавших участков, заключается в том, что оно не учитывает возможность обратной коалесценции после рекомбинации.

Далее, с помощью нашего метода мы продемонстрировали, возможно, контринтуитивный факт того, что при условии наблюдаемой частоты адаптивного аллеля и фиксированном времени интрогрессии, более сильный отбор приводит к более коротким интрогрессировавшим участкам хромосом (рис. 2).

2.2.6 Выводы

Адаптивная интрогрессия - важное и распространенное явление в эволюционной генетике [11]. Мы разработали приближенную математическую модель для численного вычисления распределения длины предковых участков хромосом при адаптивной интрогрессии вблизи локуса под

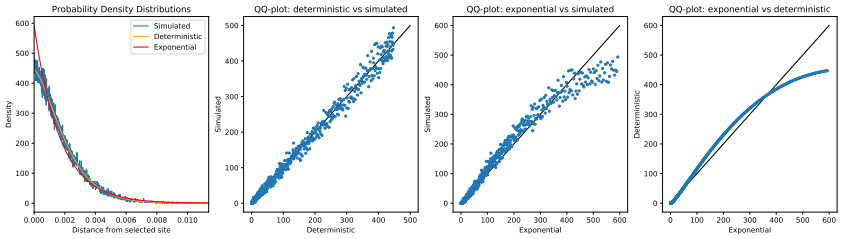


Рис. 1: Распределение расстояния от адаптивного локуса до одного из концов интрогрессировавшего участка. Коэффициент отбора равен $s = 0.01$, доля примешивания равна $\omega_1 = 0.1$ и время с момента интрогрессии равно $T = 1000$ поколениям. Наблюдаемая частота аллеля $\omega_0 = 0.94$. На первой панели находятся плотности вероятности для эмпирического распределения, полученного при помощи компьютерного моделирования, для детеминированной приближенной модели и для экспоненциального распределения. Три оставшиеся панели показывают QQ-графики трех пар указанных распределений.

действием отбора в случае одного пульса примешивания. Этот подход позволяет эффективно и быстро вычислять длины таких участков для широкого диапазона реалистичных сценариев адаптивной интрогрессии.

2.3 Метод предсказания адаптивной интрогрессии

Разработанная в предыдущем разделе теория адаптивной интрогрессии легла в основу нового метода AHMM-S. Этот метод является модификацией метода Ancestry_HMM [6] для оценки локального происхождения хромосом при перемешивании двух популяций, основанном на скрытой марковской модели. Так, мы предполагаем одно дискретное событие примешивания (интрогрессии). При этом вероятности наблюдений в модели с естественным отбором остаются теми же, что и без отбора, то есть совпадают с ними для метода Ancestry_HMM. Важное отличие заключается в том, что естественный отбор влияет на вероятности переходов между состояниями. Эти вероятности могут быть вычислены в детерминистической модели адаптивной интрогрессии, представленной нами в предыдущем разделе. Такая модель оптимизируется на равных интервалах вдоль хромосомы, а результат оптимизации сравнивается с результатом для нейтральной модели (без естественного отбора). Это позволяет

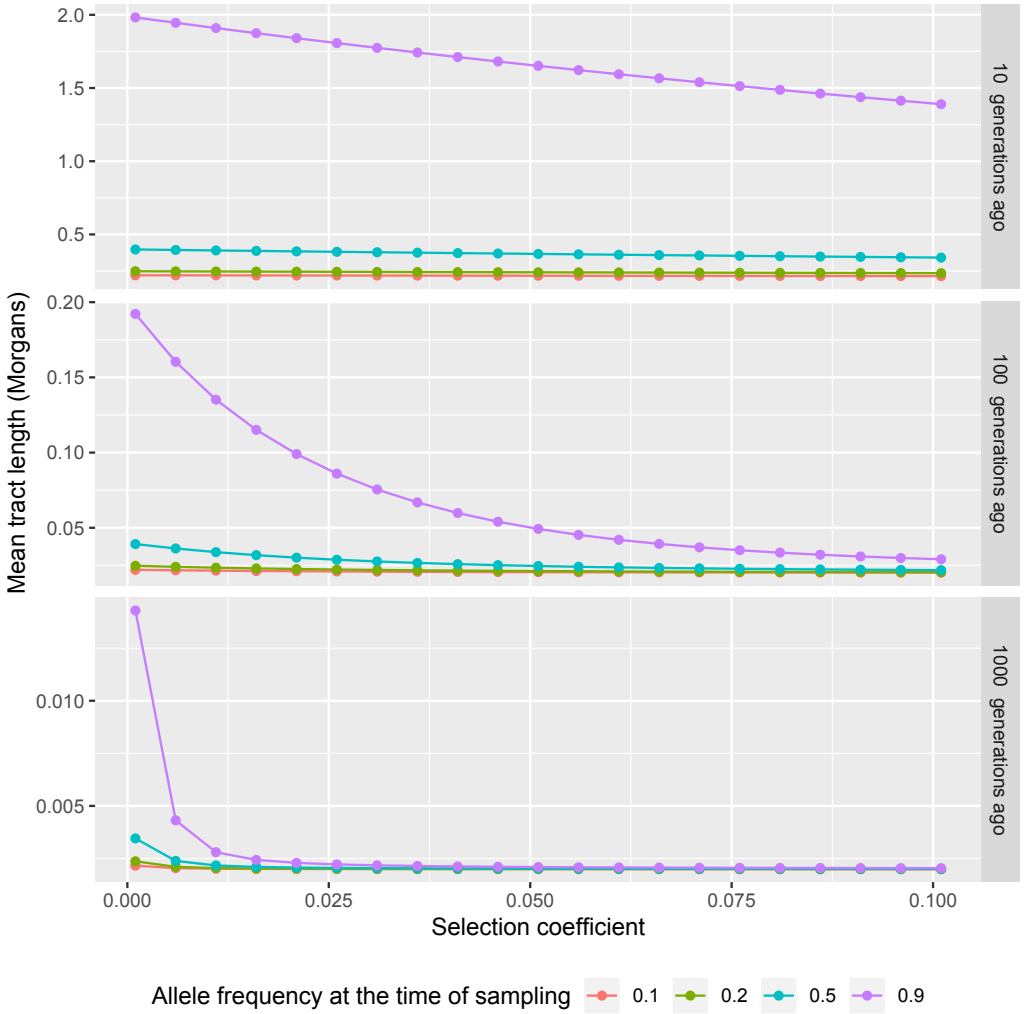


Рис. 2: Зависимость ожидаемой длины интрогрессировавшего участка при условии фиксированной частоты аллеля в момент наблюдения. Разные панели соотвествуют разным временам интрогрессии (10, 100 и 1000 поколений назад соответственно). Разные цвета линий отвечают разным частотам аллеля в момент наблюдения (0.1, 0.2, 0.5, 0.9).

находить локусы под отбором, а также оценивать силу отбора, действующую на эти участки.

Для эффективного вычисления матрицы переходных вероятностей, мы предложили 4-х точечное приближение частот переходов $f_{10}(r)$ и $f_{01}(r)$. Например, для частоты перехода $f_{10}(r)$ из предкового состояния 1 (интрогрессировавшая популяция) в состояние 0 на генетическом расстоянии r (в Морганах) от локуса под отбором мы ищем приближение в виде функции

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

где L, k, α и p - численно вычисляемые параметры. Это приближение было проверено численно для широкого диапазона параметров. Оно существенно ускоряет время работы разработанного программного обеспечения.

Разработанный метод был проверен на широком диапазоне параметров адаптивной интрогрессии (время интрогрессии, сила отбора, доля примешивания). Он демонстрирует высокую чувствительность и точность оценки параметров. Программное обеспечение опубликована в открытом доступе на GitHub `https://github.com/jesvedberg/Ancestry_HMM-S/`.

2.4 Симулятор вирусных генеалогий `VGsim`

Беспрецедентная общемировая работа по сбору и публикации геномных данных вируса SARS-CoV-2 предоставляет возможность проследить распространение и эволюцию вируса в реальном времени. Одновременно появилась очевидная необходимость в развитии вычислительных методов для изучения вирусной эволюции. При этом необходимо иметь средства для точного моделирования вирусных эволюционных процессов, так как научному сообществу необходимо, во-первых, валидировать новые разработанные методы анализа, а во-вторых, понимать влияние различных факторов и сложностей на результаты исследований.

Датасеты размеров, сравнимых с полученными во время пандемии данными, требовательны к вычислительной масштабируемости и работе с памятью используемых методов. Симулятор вирусных генеалогий `VGsim`, основанный на комбинации обобщенной SIS-модели и подхода структурированной коалесценции, эффективно масштабируется для подобных задач. На первом этапе (прямой проход) моделируется эволю-

ционная динамика вируса при помощи алгоритма Гиллеспи, учитывая многие реалистичные факторы. На втором этапе (обратный проход) используется подход структурированной коалесценции, который по смоделированной динамике моделирует генеалогическое дерево образцов патогенов, попавших в моделируемый датасет.

2.4.1 Модель и реализация

Наша эпидемиологическая модель основана на компартментальных моделях [12]. Случайные траектории реализованы при помощи алгоритма Гиллеспи [13] (реализованы логарифмический прямой метод [13] и приближенный τ-leaping алгоритм [14]). Различные компартменты в нашей модели определяются несколькими взаимодействующими между собой факторами: (1) структура популяции хозяев, (2) различные группы инфицированных в зависимости от штамма и (3) различные группы уязвимых особей-хозяев.

Как было сказано ранее, моделирование состоит из двух частей: прямого прохода, генерирующего эпидемиологическую динамику, и обратного прохода, генерирующего генеалогию образцов по этой динамике. Эпидемиологическая динамика представлена в виде цепи событий (Рис. 3). Эта цепь событий определяют вероятностное пространство для генеалогии, которая генерируется во время обратного прохода. Для этого используется коалесцентный подход, обусловленный цепью событий.

`VGsim` предоставляет удобный и компактный пользовательский интерфейс на Python. Критические вычислительные части реализованы на C++ посредством Cython [15].

2.4.2 Результаты

Мы сравнили производительность `VGsim` с производительностью популярного в эпидемиологических исследованиях [16–18] симулятора `MASTER` [19], реализующиего алгоритм Гиллеспи. Масштабируемость и используемая память `VGsim` значительно превосходят показатели `MASTER` (см. Рис. 4).

Мы также провели сравнение с эпидемиологическим симулятором `TiPS` [20], который также использует комбинацию обобщенной SIS модели и структурированной коалесценции и генерирует генеалогии с учетом эпидемиологических траекторий. Для простой SIR модели реализация

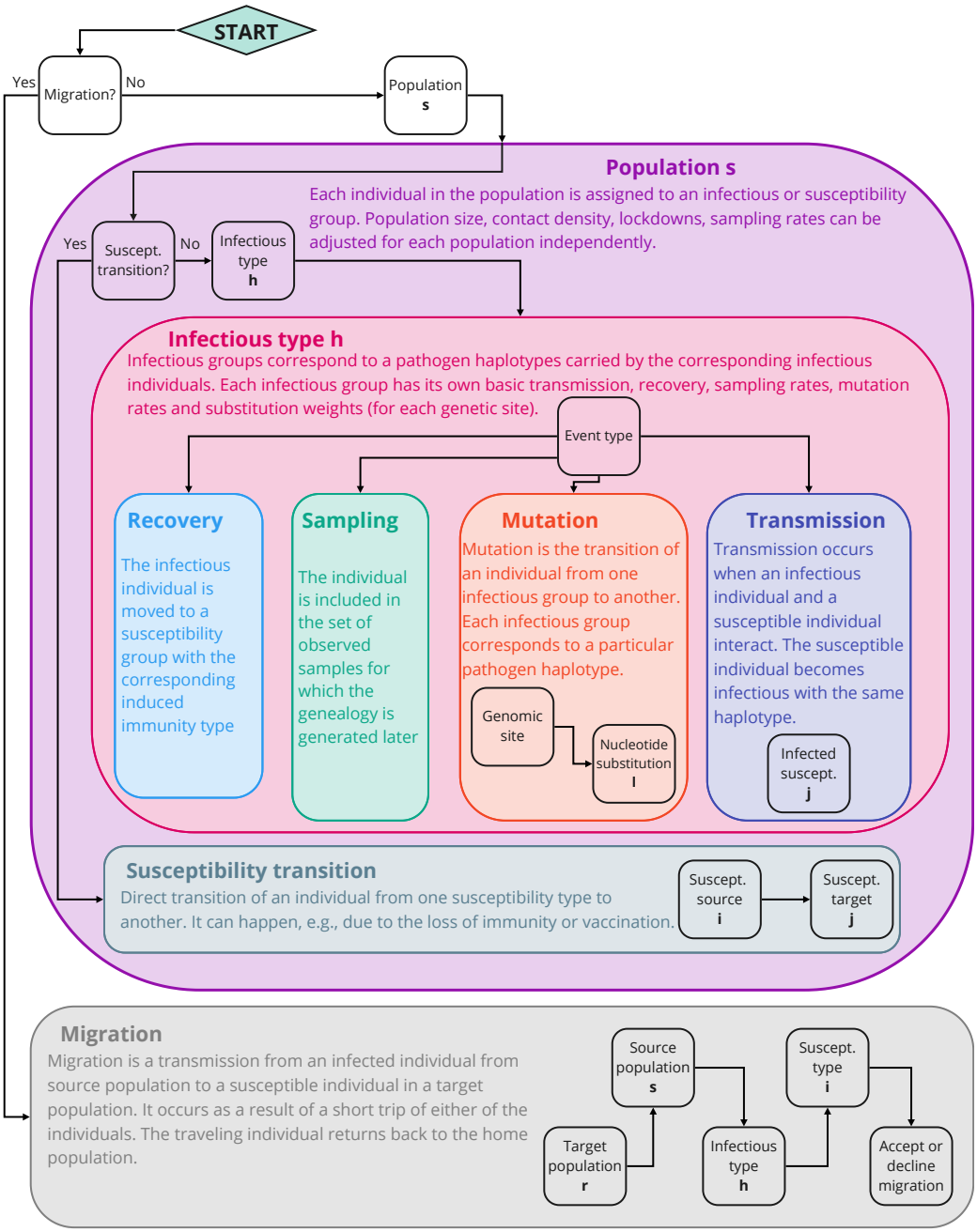


Рис. 3: Схема алгоритма Гиллеспи, используемая при генерации события в прямом проходе. Черные квадраты соответсвуют последовательным шагам, где подгруппы событий выбираются согласно их весами, или склонностями, зависящим от параметризации модели и текущего состояния эпидемиологического процесса. Склонности для каждого шага кэшируются и обновляются, на основе графа зависимостей, только если их значения изменяются вследствие выбранного события.

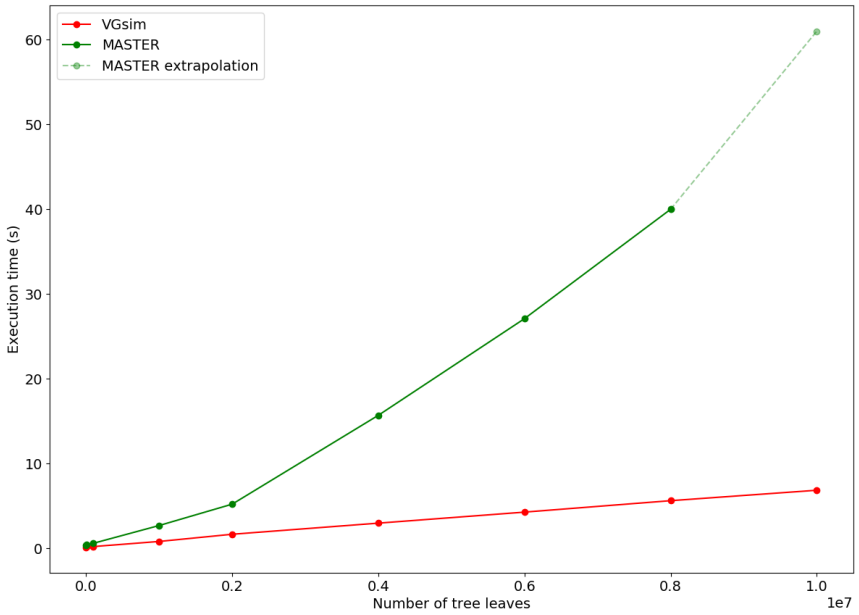


Рис. 4: Сравнение производительности симуляторов `VGsim` и `MASTER`: время для симуляции генеалогии с фиксированным числом листьев.

точного алгоритма Гиллеспи в `VGsim` примерно в два раза быстрее, чем в `TiPS` (см. Таблицу 1). Для обратного прохода (генерации генеалогий по эпидемиологическим траекториям) `VGsim` масштабируется гораздо лучше, чем `TiPS` (Таблица 2).

2.5 Точность приближения структурированной коалесценции с рекомбинацией моделью SMC'

Популяционная модель SMC и её модификация SMC' используются во многих методах анализа популяционной структуры из полногеномных данных. В работе [21] было показано, что для однородной популяции SMC' является очень точным приближением коалесценции с рекомбинацией. В то же время в реальности часто присутствует сложная структура популяции. Мы исследовали точность приближения структурированной

Число итераций	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

Таблица 1: Время прямого прохода в секундах для различного числа итераций в SIR модели. Частота выздоровления установлено равным 1, частота заражения равна 2.5. Тесты проведены на компьютере MacBookPro с процессором Quad-Core Intel Core i5 2 GHz и 16GB ОЗУ.

Число листьев	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

Таблица 2: Время обратного прохода в секундах для различного числа образцов в SIR модели. Частота выздоровления установлено равным 1, частота заражения равна 2.5. Размер популяции установлен равным 10^7. Тесты проведены на компьютере MacBookPro с процессором Quad-Core Intel Core i5 2 GHz и 16GB ОЗУ.

коалесценции с рекомбинацией с помощью модели SMC', рассмотрев случай двух популяций с миграцией между ними. Мы показали, что в этом случае точность приближения существенно снижается по сравнению с однородным случаем, что необходимо учитывать при применении таких методов как PSMC [22], MSMC [23] и др.

Рассмотрим совместное распределение времен до ближайшего общего предка двух хромосом в двух локусах на генетическом расстоянии ρ (в данном случае ρ является частотой рекомбинации между двумя локусами) в моделях коалесценции с рекомбинацией и SMC'. Обозначим эти распределения, соответственно, через $p_{\rho,CR}(t,s)$ и $p_{\rho,SMC'}(t,s)$.

Для вычисления $p_{\rho,CR}(t,s)$ мы рассматриваем марковский процесс с

непрерывным временем (время течет из настоящего в прошлое). Частоты коалесценция в первой и второй популяциях обозначим через λ_1 и λ_2, частоты миграций через m_{12} и m_{21}.

Состояниями этого процесса являются различные конфигурации предковых хромосом. Каждая предковая хромосома состоит из двух локусов и находится в одной из двух рассматриваемых популяций.

Предковые линии хромосом могут сливаться (то есть происходить от общего предка), если они находятся в одной популяции. Всего есть 40 состояний процесса, а также два поглощающих состояния, соответсвующих ближайшему общему предку. Всего есть три типа переходов между состояниями:

- рекомбинации с частотой ρ,
- коалесценции с частотами λ_1 и λ_2,
- миграции с частотами m_{12} и m_{21}.

Модель SMC' приближает коалесценцию с рекомбинацией ещё одним марковским процессом, направленным вдоль генома. Состояниями этого процесса являются генеалогические деревья в локусе. В случае двух хромосом форма дерева тривиальна, а состояниями фактически являются времена до общего предка.

Во-первых, мы представляем разность между совместными распределениями $p_{\rho,CR}(t,s)$ и $p_{\rho,SMC'}(t,s)$ для однородного и структурированного случаев. Как показано на рисунке 5, есть очевидное качественное отличие между ними.

Далее, мы вычислили полную вариацию между этими совместными распределениями в зависимости от генетического расстояния ρ. Следуя определению из [21], полная вариация - это L_1-норма разности между двумя совместными распределениями $p_{\rho,CR}(t,s)$ и $p_{\rho,SMC'}(t,s)$

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

Как продемонстрировано на Рис. 6, для структурированного случая полная вариация больше, чем для однородного. Важно, что полная вариация убывает существенно медленнее для структурированного случая. Таким образом, SMC' не является точным приближением коалесценции с рекомбинацией.

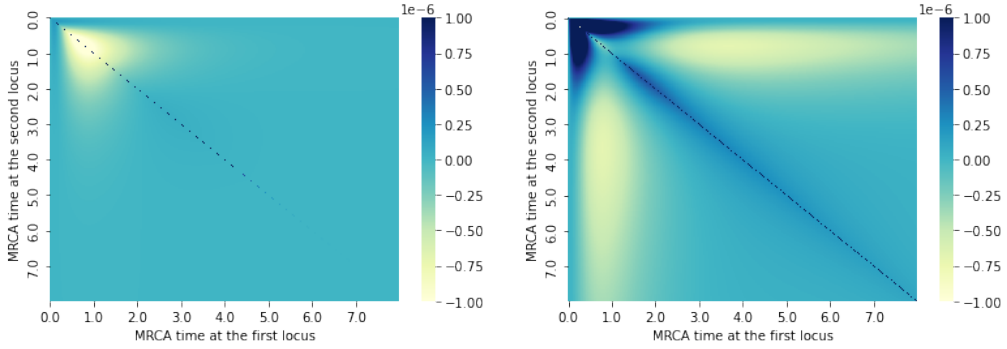


Рис. 5: Разность между совместным распределением времен до ближайшего общего предка в двух локусах на генетическом расстоянии $\rho = 2$ для моделей коалесценции с рекомбинацией и SMC'. Левая панель показывает разность для однородной популяции. Правая панель показывает разность для двух популяций с миграцией.

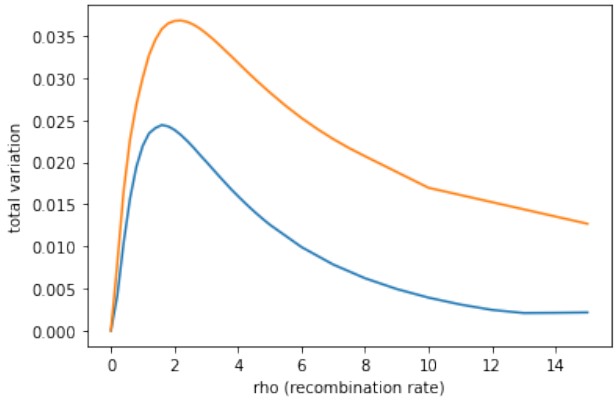


Рис. 6: Полная вариация между совместными распределениями времен до ближайшего общего предка в двух локусах в зависимости от генетического расстояния ρ. Синяя (нижняя) линия соответствует случаю с однородной популяцией. Оранжевая (верхняя) линия соответствует случаю двух популяций с миграцией.

2.6 Эффективный размер популяции и миграция

Эффективный размер популяции может быть определен как среднее время до коалесценции (до ближайшего общего предка) двух предковых линий, измеряемое в количестве поколений [24]. В стандартной коалесцент-

ной модели с одной однородной популяцией [25] имеет место следующая простая интерпретация. При размере популяции $N >> 1$ частота коалесценций равна $1/N$, и ожидаемое время коалесценции равно $\lambda^{-1} = N$. Это определение можно естественным образом обобщить для эффективного размера перемешанной популяции. Для некоторой популяции мы рассматриваем частоту коалесценций в момент времени t между парами предковых линий. Время рассматривается в обратном направлении из настоящего ($t = 0$) в прошлое. Частота коалесценций $\lambda(t)$ определяет марковский неоднородный по времени процесс, который описывает распределение коалесцентных времен. Следовательно, вероятностное распределение коалесцентных времен T_c задается законом

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

Мы определяем *исторический эффективный размер популяции* в момент времени t как обратное к частоте коалесценции $\lambda(t)$. Эта величина зависит от структуры популяции и демографии. Она позволяет приближать популяции со сложной историей одной идеализированной популяцией (например, популяцией Райта-Фишера). Такой подход полезен при интерпретации результатов таких методов как PSMC, позволяющих оценивать изменение размера популяции во времени. В некоторых случаях он также приближает величину, оцениваемую PSMC [1, 26]. Мы показали, что несмотря на то, что в некоторых сценариях PSMC действительно дает хорошую оценку коалесцентных времен, в других сценариях PSMC может приводить к существенно отличающимися результатам. Мы качественно показали, что PSMC, неформально, ищет наилучшее приближение для матрицы переходов секвенциальной марковской модели. Это объясняется тем, что функция правдоподобия, оптимизируемая PSMC, непосредственно зависит от переходных вероятностей. Тем не менее, матрица переходов может иметь иное предельное распределение, определяющее эффективный размер популяции.

Рассмотрим случай двух перемешивающихся популяций - это может происходить как через непрерывную миграцию, так и через пульсовую миграцию. Обозначим предковые популяции через $S_1(t)$ и $S_2(t)$. В любой момент времени предковая линия наблюдаемой популяции S_m находится либо в популяции $S_1(t)$, либо в популяции $S_2(t)$ в следствие миграции. Внутри популяций $S_1(t)$ и $S_2(t)$ предковые линии неотличимы, что означает, что любая пара линий в пределах популяции имеет одну и ту же

вероятность коалесценции. Обозначим эффективные размеры популяций $S_1(t)$ и $S_2(t)$ через $N_{L1}(t)$ и $N_{L2}(t)$ соответственно. Будем называть $N_{L1}(t)$ и $N_{L2}(t)$ локальными эффективными размерами популяций, то есть эти величины представляют только эффективное число особей в популяции в момент времени t, тем самым отделяется эффект от миграции на генетическое разнообразие популяций.

Если две предковые линии находятся в одной популяции $S_i(t)$ ($i = 1, 2$) в момент времени t, между ними может произойти коалесценция с частотой $1/N_{Li}(t)$. Если же они находятся в разных популяциях, то коалесценция между ними невозможна. При условии, что две линии не слились к моменту времени t, пусть $P_1(t)$ и $P_2(t)$ - вероятности того, что две линии находятся в популяции S_1 и популяции S_2 соответственно. Пусть $P_0(t) = 1 - P_1(t) - P_2(t)$ - вероятность того, что две линии находятся в разных популяциях. Тогда частота коалесценции между парой предковых линий в момент времени t равна

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

и исторический эффективный размер популяции равен

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

Условие того, что наблюдаемая популяция S_m - это популяция $S_1(0)$, эквивалентно начальным условиям на вероятности

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

Таким образом, мы видим очевидную разницу между локальными размерами популяций ($N_{L1}(t)$ и $N_{L2}(t)$) и историческим размером популяции ($N(t)$). Во многих случаях оценки эффективного размера популяции, полученные PSMC и схожими методами, являются оценками исторического эффективного размера популяции, а не локальных эффективных размеров.

2.6.1 Отделение эффекта миграции из эффективного размера популяции

Положим, что мы наблюдаем две популяции $S_m^{(1)} = S_1(0)$ и $S_m^{(2)} = S_2(0)$, между которыми происходила миграция. Записывая уравнение 7 для

обеих популяций, мы получаем систему уравнений, связывающую исторический эффективный размер популяций $S_m^{(1)}$ и $S_m^{(2)}$ (N_1 и N_2) с локальным эффективным размером предковых популяций (N_{L1} и N_{L2}).

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

Таким образом, при известных исторических эффективных размерах популяций N_1 и N_2 (например, оцененных методом PSMC) и частотах миграции m_{12} и m_{21} можно численно определить локальные эффективные размеры популяций N_{L1} и N_{L2}.

2.6.2 Оценка параметров

Мы применили разработанный метод для проверки гипотезы, выдвинутой в работе [2], о глубоком разделении популяций внутри Африки между San и Dinka. А именно, оценка времени разделения, полученная TT методом, превышает 8500 поколений назад. Модель с наибольшим правдоподобием, полученная нашим методом MiSTI, оценивает время разделения в ≈ 3700 поколений назад (то есть около 107,000 лет назад, полагая продолжительность поколения 29 лет) с практически односторонней миграцией из Dinka в San (Таблица 3). Эти выводы согласуются с оценками, полученными из симуляций с аналогичными параметрами.

m1 Dinka в San	m2 San в Dinka	Время разделения MiSTI (в поколениях)	log(lik)	Время разделения TT (в поколениях)
2.5	2.03×10^{-9}	3729	-4381	-
2.5	-	3729	-4381	-
-	1.49	3210	-4582	-
-	-	3001	-4607	T1 = 8582, T2 = 8527

Таблица 3: Оценки времени разделения и скоростей миграции методами MiSTI и TT для африканских популяций San и Dinka для моделей с двусторонней, односторонней и без миграции.

2.7 Множественное примешивание и неравновесное сцепление трех локусов

Существует много методов для предсказания наличия примешивания между популяциями [27–32]. Также проведено существенное число исследований по развитию теории и методов для оценки времени примешивания. Один подход основан на предсказании длин предковых участков хромосом [6, 33–36] и [6*].

Другой подход, которым мы пользуемся в этой работе, основан на убывании неравновесного сцепления примешивания (ALD). Неравновесное сцепление присутствует в любой популяции из-за мутаций и генетического дрейфа. В однородной и генетически изолированной популяции с рекомбинацией оно быстро убывает в масштабе генома. Однако, предковые участки, попадающие в популяции при примешивании, приводят к появлению ALD на значительно больших расстояниях. После единственного примешивания неравновесное сцепление в перемешанной популяции начинает постепенно убывать в последующих поколений из-за рекомбинаций. Эта идея была использована в методах ROLLOFF [37] и ALDER [38], где оценивается ALD для двух генетических локусов. ROLLOFF и ALDER хорошо подходят для оценки времени примешивания в том случае, если примешивание может быть приближено единственным пульсом миграции. Тем не менее, во многих реалистичных сценариях примешивание происходило через несколько пульсов миграции. Например, известный пример такого примешивания - это примешивания коренных американцев к коренному населению острова Пасхи [39], а также перемешанных популяций Америки [40]. В таких случаях ожидаемое убывание ALD становится смесью экспоненциальных законов. Существующие методы датировки примешивания, основанные на ALD, на данное время могут либо оценивать время последнего примешивания [37], либо использоваться для отвергания гипотезы о единственном пульсе [38].

В нашей работе мы используем определение Беннетта и Слаткина [41, 42] для неравновесного сцепления трех локусов с тем, чтобы изучить убывание ALD как функцию расстояний между этими локусами. Мы вывели аналитическое уравнение, описывающее убывание ALD при множественных пульсах миграции, а также разработали метод для оценки времен двух пульсов миграции при перемешивании двух популяций. Результаты провалидированы на данных, полученных при помощи компьютерного моделирования и в качестве примера применены к экспери-

ментальным данным - образцам мексиканцев и колумбийцев из проекта 1000 геномов.

2.7.1 Неравновесное сцепление и локальное происхождение

Обозначим через x, y, z три последовательных генетических локуса с расстояниями между ними d и d'. $H_{i,x}, H_{i,y}, H_{i,z}$ - гаплотипы ($\{0, 1\}$) или генотипы ($\{0, 1/2, 1\}$) в соответствующих локусах i-ого генома. Неравновесное сцепление трех локусов определяется как ковариация H_x, H_y, H_z

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

Неравновесное сцепление в перемешанной популяции зависит от генетической дифференциации между исходными популяциями и их историей примешивания. Пусть A_x представляет локальное происхождение локуса x, причем $A_x = 0$ в случае, если x унаследован из популяции, которая примешивается два раза, и $A_x = 1$, если локус унаследован из популяции, которая примешивается один раз. Тогда D_3 можно представить через частоты аллелей и ковариацию локального происхождения A_x, A_y, A_z. Рассмотрим условное математическое ожидание $\mathbb{E}(H_x|A_x) = g_x + \delta_x A_x$, где g_x - частота аллеля в локусе x в популяции 0 и $\delta_x = f_x - g_x$ - разность частот аллеля в локусе x в двух исходных популяциях. Мы предполагаем, что частоты аллеля в исходной популяции известны и фиксированы. Тогда

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

Более того, в общем случае имеет место равенство

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

2.8 Глубинное обучение для демографического анализа

В работах [4*, 3*] представлен метод на основе глубинного обучения для предсказания локальных времен ближайшего общего предка из диплоидного генома.

Предсказание демографии, то есть оценка исторического эффективного размера популяции, является одной из важнейших задач популяционной генетики. Это один из ключевых факторов генетического разнообразия популяции. Например, в [1] показано, что все неафриканские популяции прошли через бутылочное горлышко между приблизительно 30 и 100 тысячами лет назад. В африканских популяциях это явление отсутствует. Этот факт поддерживает гипотезу об африканском происхождении современного человека.

Глубинное обучение демонстрирует высокую точность для решения многих задач, включая анализ различных последовательностей. Мы разработали архитектуру на основе рекуррентных нейронных сетей для предсказания локальных времен до общего предка из диплоидной последовательности (аналогично PSMC). Эта задача имеет две ключевые сложности: большая длина геномной последовательности ($3.2 \cdot 10^9$ для человека) и отсутствие размеченных данных для обучения.

На данный момент глубинное обучение постепенно начинает использоваться в популяционной генетике, хотя и не является популярным подходом. Первый метод для предсказания детальной популяционной истории, используя глубинное обучение, предложен в работе [43] и доказывает, что подходы с нейронными сетями могут быть мощным инструментов в популяционной генетике. Тем не менее, предстоит приложить еще немало усилий в этой области, включая изучение преимуществ, ограничений и недостатков глубинного обучения в рассматриваемых задачах, прежде чем эти методы найдут свое широкое применение в анализе экспериментальных данных.

Мы использовали программный симулятор `msprime` [44] для генерации подходящих выборок для обучения. На вход нейронной сети подается последовательность 0 (гомозиготные сайты) и 1 (гетерозиготные сайты). Целью предсказания является один из временных интервалов, где находится локальный общий предок. Задача решалась как задача классификации. Программный код находится в открытом доступе на GitHub `https://github.com/Genomics-HSE/deepgen`. Пример предсказания времени для общего предка вдоль генома продемонстрирован на Рисунке 7. Ось x соответствует позициям вдоль генома, ось y - временным интервалам, куда попадает локальный общий предок. Цвета соответствуют вероятностям попадания локального общего предка в определенный временной интервал. Таким образом, разработанный метод глубинного обучения достаточно точно предсказывает время до локального ближай-

шего общего предка.

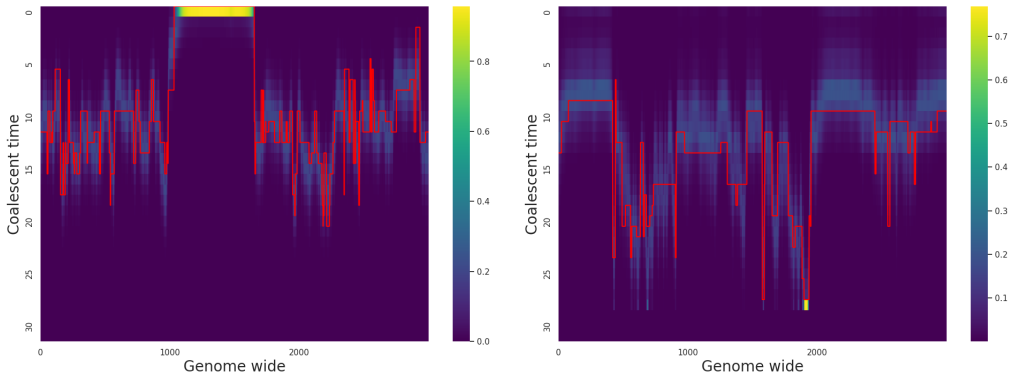


Рис. 7: Время до ближайшего общего предка (LCA) вдоль хромосомы. Тепловая карта показывает вероятность того, что время LCA попала в данный временной интервал (ось y) в рассматриваемой позиции на хромосоме (ось x). На левой панели показан пример для случая популяции с постоянным размером и нейросетью, натренированной на примерах из популяций постоянного размера. На правой панели показан пример для популяции с бутылочным горлышком, нейросеть тренировалась на примерах из случайных демографических историй. Красные линии показывают истинные времена LCA, известные из соответствующих симуляций.

2.9 Естественный отбор в чилийской популяции после пост-колумбова примешивания

В статье [11*] исследован вопрос о наличии естественного отбора после примешивания (адаптивная интрогрессия) в чилийской популяции. Современное население Чили появилось в следствие перемешивания коренного населения Южной Америки, европейских колонизаторов и африканцев. Для решения задачи мы провели полногеномный поиск отклонения доли европейского локального происхождения от среднего значения по всему геному.

После предсказания локального происхождения методом LAMP-LD, мы вычислили долю европейского локального происхождения для каждого SNP. Далее мы использовали односторонний t-критерий Стьюдента, чтобы определить отклонения от среднего европейского происхождения.

В каждом SNP мы сравнивали долю европейского происхождения со средним значением по геному $p_0 = 0.52$. Мы провели статистический тест с нулевой гипотезой $H_0 : p_i = p_0$ и альтернативной гипотезой $H_1 : p_i > p_0$ для каждого SNP i (здесь p_i - доля европейского происхождения в SNP i). Варианты (SNP), достигшие значимого уровня с $p < 10^{-5}$, рассматривались как находящиеся под отбором после примешивания [45].

Для того чтобы убедиться в состоятельности такого подхода и выборе уровня значимости, мы провели компьютерное моделирование распределения p_i. Поскольку отклонения p_i от среднего значения часто обусловлены генетическим дрейфом, мы оценили эффективный размер популяции для наших чилийских образцов приравнивая эмпирическое и теоретическое значения дисперсии p_i. Согласно [46], дисперсия

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

где p_0 - доля примешивания, T - время, прошедшее с момента примешивания, $2N_e$ - гаплоидный эффективный размер популяции.

Мы использовали полученные таким образом оценки эффективного размера популяции $2N_e$ для моделирования локального происхождения в пакете SELAM [47]. Мы рассмотрели сценарий с одним единовременным перемешиванием трех популяций T поколений назад (для разных реалистичных значений T). Доли примешивания соответствовали долям европейской, коренной американской и африканской компонентам, оцененным при помощи LAMP-LD. Для исключения влияния неравновесного сцепления генов мы отобразили позиции SNP из нашего эмпирического датасета на смоделированные последовательности. Для всех комбинаций параметров мы посчитали t-критерий Стьюдента. Ни в одном случае p-значения не оказались ниже критического порога $p = 10^{-5}$. Таким образом, выбранный уровень статистической значимости действительно можно считать индикатором наличия отбора после примешивания.

Наш численный подход позволил проверить нашу гипотезу о том, что после примешивания чилийцы подверглись действию естественного отбора по генетической вариации европейского происхождения. Поскольку предковые вариации могут содержать одни и те же генетические варианты, для каждого SNP мы определили его происхождение в конкретном геноме. Затем мы использовали отклонение доли локального предкового происхождения [45,48] (а не непосредственно частоты аллелей) от среднего геномного происхождения (оценка 0,52 для европейцев) в качестве сиг-

нала отбора. Мы построили t-тест с нулевой гипотезой $H_0 : \mu_{EUR,i} = 0,52$ и конкурирующей гипотезой $H_1 : \mu_{EUR,i} > 0,52$ для каждого варианта i.

Мы обнаружили 85 SNP, которые достигают порога статистической значимости $P < 10^{-5}$, рекомендованного для недавно смешавшихся популяций [45]. Мы обосновали выбор такого уровня статистической значимости при помощи компьютерного моделирования (см. выше). 85 SNPs соответсвуют пику европейского происхождения на хромосоме 12. Этот участок связан с несколькими регуляторными регионами, включая две lncRNA (*RP11-13A1.1 и RP11-13A1.3*) и один псевдоген (*RP11-13A1.2*).

2.10 Филодинамика коронавируса SARS-CoV-2 в России

Пандемия Covid-19 поставила много вызовов научному сообществу. В частности, большие усилия были сосредоточены на секвенировании образцов коронавируса SARS-CoV-2 в большинстве регионов мира. Это в свою очередь позволило проводить геномный эпидемиологический анализ для изучения распространения различных вариантов (штаммов) коронавируса. В этом разделе мы представляем результаты филодинамического анализа вспышки Covid-19 в НИИ травматологии имени Вредена в марте-апреле 2020 года [9*] и варианта коронавируса дельта в Москве в апреле-сентябре 2021 года [7*].

2.10.1 Вспышка Covid-19 в НИИ травматологии имени Вредена

Мы исследовали крупный трансмиссионный кластер - внутрибольничную вспышку Covid-19 в НИИ травматологии им. Вредена (далее - больница Вредена) в Санкт Петербурге в начале пандемии. Согласно внутреннему расследованию, предположительного нулевого пациента прооперировали 27 марта 2020 года. Несмотря на то, что регулярное тестирование на Covid-19 в больнице Вредена началось 18 марта 2020 года, первый положительный образец был получен 3 апреля 2020 года. После этого поэтапно были введены карантинные меры между 7 и 9 апреля 2020 года, которые включали в себя полное закрытие больницы, изоляцию отделений, отключение общебольничной вентиляционной системы. 474 пациента и 270 человек персонала оставались в больнице не протяжении 35 дней.

Наш датасет состоит из геномов вируса SARS-CoV-2 от 52 пациентов и работников больницы Вредена. Филогенетический анализ показал, что эти образцы образуют три различные группы со своим уникальным набором мутаций. В самую большую группу (далее группа 1) входит 41 последовательность, полученная между 3 и 22 апреля 2020 года. Группа 2 состоит из 7 последовательностей, а соответствующая ей клада на мировом филогенетическом дереве включает в себя также одну последовательность из Англии. И, наконец, третья группа состоит из 4 последовательностей. Образцы группы 1 происходят из разных отделений, находящихся на разных этажах, в то время как группы 2 и 3 - каждая из своего собственного отделения.

Группы 1 и 2 являются филогенетически отдаленными от группы 3. Ближайшего общего предка групп 1 и 2 отделяет шесть мутаций от группы 3. Группы 1 и 2 принадлежат линии B.1.1, определенной тремя мутациями в позициях 28881, 28882 и 28883, и далее определяются мутациями в позициях 26750 и 1191 соответственно. В то же время группа 3 принадлежит линии B.1.5, а также дополняется мутацией в позиции 20268, которая на тот момент была распространена по всему миру и появилась на ранних этапах филогенетической истории, а также ещё двумя дополнительными мутациями. Таким образом, мы получили сильное свидетельство того, что группа 3 появилась в результате независимого заноса инфекции относительно групп 1 и 2.

Для более детального изучения распространения вспышки этой внутрибольничной инфекции мы провели Байесовский филогенетический анализ в горизонтной модели рождения-смерти (birth-death skyline) [49] в пакете BEAST2 [50]. Из-за высокой вероятности нескольких интродукций инфекции, мы провели анализ как всего вреденского датасета, состоящего из групп 1, 2 и 3, а также двух его подмножеств: одного, состоящего из групп 1 и 2, и второго, состоящего из группы 1. Результаты нашего анализа представлены на рисунках 8 и 9.

Мы обнаружили, что Байесовский анализ поддерживает по крайней мере две различные интродукции коронавируса SARS-CoV-2 в больницу Вредена. Это подтверждается глубоким разделением между группами 1-2 и группой 3. Ближайший общий предок этого датасета датируется 21 февраля 2020 года (95% апостериорный достоверный интервал 20 января-21 марта). Это более чем на месяц раньше предполагаемой даты первой интродукии (27 марта), что снова подтверждает, что группа 3 и все остальные образцы были занесены в больницу независимо.

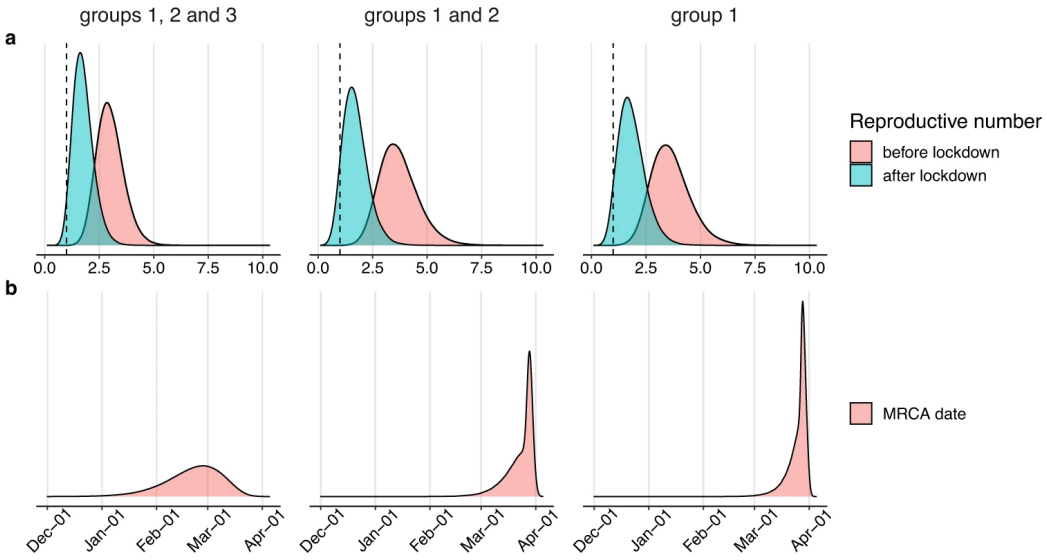


Рис. 8: Оценки филодинамических параметров вспышки Covid-19 в больнице Вредена в горизонтной модели рождения-смерти в пакете BEAST2. Верхняя панель демонстрирует апостериорные распределения эффективного репродуктивного числе R_e (верхняя панель) с пунктирной линией, показывающей критическое значение $R_e = 1$. Нижняя панель демонстрирует апостериорные распределения даты ближайшего общего предка. На каждой панели приведены оценки по трем датасетам из вреденских образцов: группы 1, 2 и 3 (левая колонка), группы 1 и 2 (средняя колонка) и группа 1 (правая колонка).

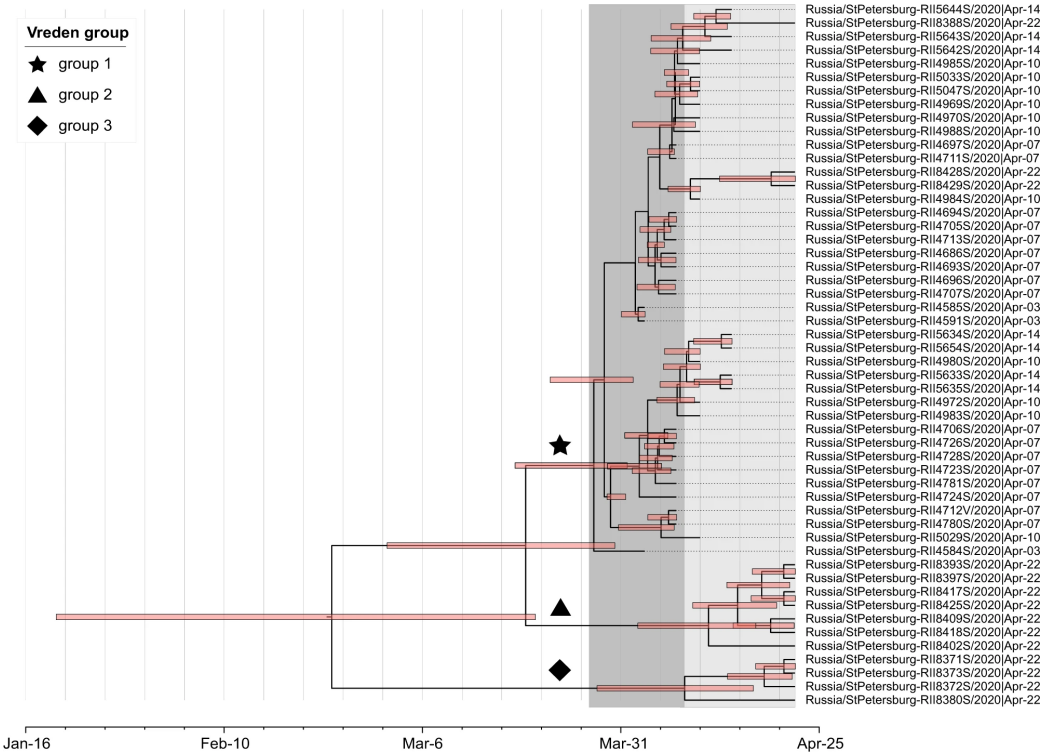


Рис. 9: Дерево максимально достоверных клад (maximum clade credibility) для вспышки Covid-19 в больнице Вредена. Группы 1, 2 и 3 отмечены звездочкой, треугольником и ромбом соответственно. Розовые полоски показывают 95% достоверные интервалы для времен вершин. Период вспышки отмечен серым фоном, причем время от предполагаемого нулевого пациента (27 марта) до введения карантина (8 апреля) выделено темным цветом.

Третья интродукция в больницу Вредена также весьма вероятна. В самом деле, ближайший общий предок групп 1 и 2 датируется 24 марта 2020 года (95% достоверный интервал 6 марта-1 апреля). Учитывая отсутствие явных признаков инфекции в больнице в конце марта, весьма вероятно, что группы 1 и 2 происходят от двух независимых интродукций. Корень (ближайший общий предок) группы 1 датируется 26 марта (95% достоверный интервал 13 марта-2 апреля), что согласуется с периодом болезни предполагаемого нулевого пациента. Дополнительным подтверждением независимого происхождения групп 1 и 2 является наличие нероссийской (английской) последовательности в одной кладе.

Мы оценили филодинамические параметры до и после введения карантинных мер. В анализе всех трех датасетов, оценки эффективного репродуктивного числа остаются стабильными и согласуются друг с другом. Исходя из анализа всех трех групп, мы обнаружили, что эффективное репродуктивное число R_e было равно 3.0 (95% достоверный интервал $1.85 - 4.25$) до 8 апреля и снизилось до 1.76 (95% достоверный интервал $0.91 - 2.71$) после 8 апреля (Рисунок 8). Аналогичные оценки для эффективного репродуктивного числе R_e для группы 1 равны 3.64 (95% достоверный интервал $2.01 - 5.43$) до карантина и 1.85 (95% достоверный интервал $0.77 - 3.06$) после карантина соответственно. Эти оценки согласуются друг с другом, и возможные эффекты структурированной популяции (из-за независимых интродукций) не создают значительных смещений оценок.

2.10.2 Выводы

Детальный анализ локализованных трансмиссонных кластеров помогает лучше понять процесс распространения вируса. Хорошо изученные на тот момент случаи включали в себя круизный лайнер “Diamond Princess” [51–54], круизный лайнер Grand Princess [55], международную конференцию в Бостоне [56], общежитие в окрестностях Бостона [56] и внутрибольничную вспышку в госпитале Netcare St. Augustine (ЮАР). Во всех случаях, кроме одного, вспышки были генетически гомогенны, что означает, что каждая из них развилась вследствие одного случая заражения. В случае с общежитием произошло несколько интродукций, но тем не менее там была основная клада, которая включала в себя практически все образцы, в то время как остальные клады были редкими [56]. В то же время в случае вспышки в больнице Вредена, мы наблюдаем

несколько (вероятно, 2 или 3) интродукции, каждая из которых привела к появлению полноценной клады. Это может означать, что эта вспышка случилась из-за нескольких случаев супер-распространения.

Далее, наша оценка начального эффективного репродуктивного числа R_e (в период до карантина) около ~ 3.00, что является достаточно высоким значением. Несколько случаев супер-распространения и высокое значение R_e могут являться следствием того, что в силу специфики (травматология) этот госпиталь не был оборудован для инфекционного контроля, в частности, тесные контакты (например, распространение среди сотрудников), отсутствие защитных мер и отсутствие осведомленности. Во второй фазе вспышки мы наблюдаем существенное снижение эффективного репродуктивного числа до ~ 1.76. Это изменение может быть объяснено двумя факторами (или их сочетанием). Во-первых, это может быть следствием повышения осведомленности и введением карантинных мер, начиная с 7 апреля. Во-вторых, это может быть следствием того, что много людей уже заболело к тому времени, что в свою очередь предотвращало дальнейшее распространение инфекции. В самом деле, около 30% людей в госпитале были заражены к 22 апреля. У нас нет возможности оценить вклад каждого из этих факторов в замедление распространения инфекции при помощи доступных данных и методов.

2.10.3 Филодинамика варианта дельта коронавируса SARS-CoV-2 в Москве

К середине 2021 году вариант дельта коронавируса SARS-CoV-2 вытеснил все остальные варианты по всему миру. Этот вариант отличался повышенными трансмиссивностью и смертностью. В России, в отличие от большинства других стран, распространилась одна трансмиссионная линия *nsp2:K81N + ORF7a:P45L* (более 90% случаев), при этом эта линия редко встречается за пределами России. Мы исследовали распространение этой линии в стране, в частности оценили филодинамику этой линии в Москве (наиболее хорошо представленном регионе в нашей генетической выборке).

Далее все даты, указанные в этом разделе, относятся к 2021 году.

Для того чтобы оценить скорость распространения самой большой подлинии варианта Дельта, мы провели филодинамический анализ с использованием пакета BEAST2 [50]. Эпидемия Covid-19 в разных регионах России протекает по-разному и несинхронно. Так, например, време-

на эпидемических волн различаются в разных регионах. Для того чтобы минимизировать эффекты географической неоднородности, в этом анализе мы сфокусировались на одном регионе. Мы выбрали 333 образца, собранных в Москве, поскольку, как было сказано ранее, это российский регион с наибольшим количеством данных.

Филодинамические оценки эффективного репродуктивного числа R_e для указанной основной клады равны 1.82 (95% CI [1.49 − 2.16]) в мае, 1.24 (95% CI [1.07−1.41]) в июне. В июле, значение R_e упало до 0.58 (95% CI [0.40 − 0.77]), и затем снова выросло до 0.99 (95% CI [0.79 − 1.20]) в августе и до 1.27 (95% CI [0.62 − 1.94]) в сентябре, последнем месяце, вошедшем в наш генетический анализ (Рисунок 10).

В целом, указанная динамика согласуется с эпидемиологическими данными: повышенные значения R_e предваряют подъемы в числах случаев в день и соответствуют оценкам R_e, полученным методом EpiEstim из числа регистрируемых случаев. Важно отметить, что число случаев до июня включает в себя большую долю не-дельтовских случаев. Подъем общего числа случаев в мае был медленнее, чем это предсказывает соответствующее R_e, что можно объяснить уменьшением числа недельтовских случаев. Тем не менее, высокие значения R_e в мае и июне согласуются с летней волной, которая достигла пика 25 июня, а низкое значение R_e в июле согласуется с уменьшением числа случаев в этот период (Рисунок 10). Эти данные подтверждают, что основная клада (AY.122+ORF7a:P45L) ответственна за летнюю эпидемическую волну, и, вероятно, за последовавшую осеннюю волну. Такая бимодальная динамика похожа на многие другие страны северного полушария, где приход лета замедлил распространение инфекции, как, например, в Великобритании, Франции и США.

2.11 Гиперболическая геометрия и анализ генетических данных

Неевклидова, и, в частности, гиперболическая геометрия находит всё больше применений в анализе данных. Поскольку в основе генеалогий лежат деревья, мы предположили, что гиперболическая геометрия является перспективным инструментом анализа генетических данных, и провели исследования в этом направлении. Нами заложен теоретический фундамент для такого анализа в работах о численных аспектах геомет-

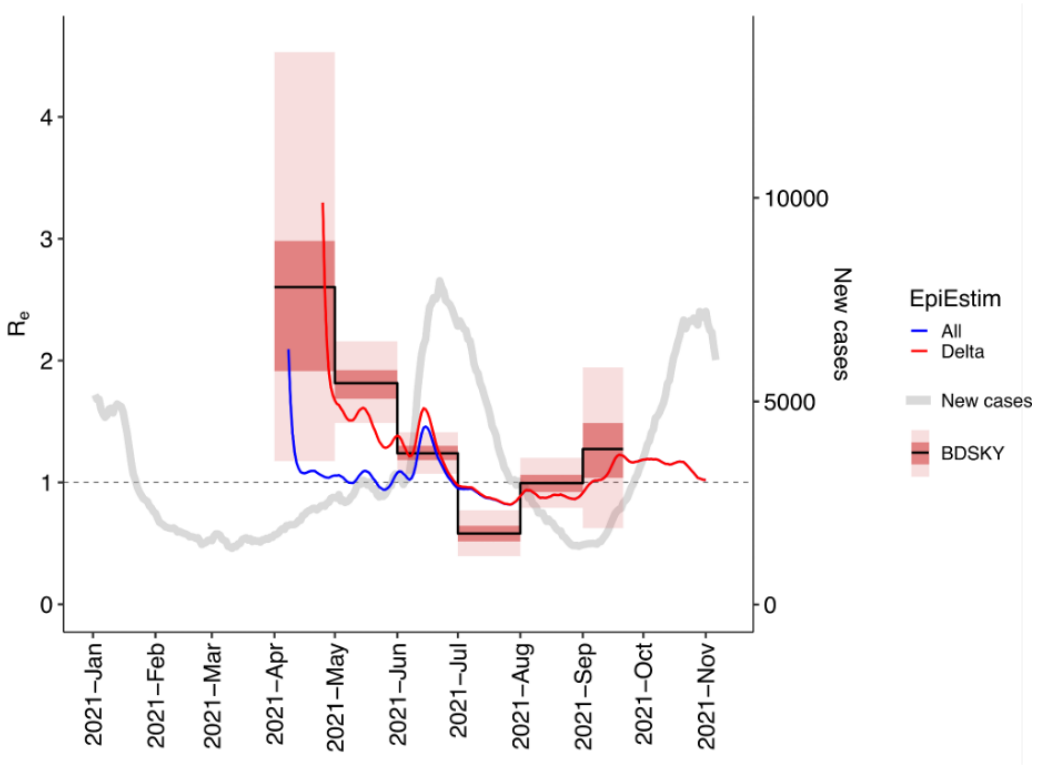


Рис. 10: Динамика эффективного репродуктивного числа R_e для основной клады варианта коронавируса дельта в Москве, оцененная в горизонтной модели рождения-смерти (черная линия; красные и розовые полосы показывают 50% и 95% достоверные интервалы соответственно), а также оцененная пакетом EpiEstim для всех (синяя линяя) или только для варианта дельта (красная линия) случаев SARS-CoV-2 в Москве. Серая линия показывает семидневное скользящее среднее ежедневного числа новых случаев в Москве независимо от генотипа.

рии гиперболических пространств. Получены следующие основные результаты:

- получена оптимальная (с точностью до мультипликативной константы) оценка для леммы Морса, утверждающей, что в гиперболическом по Громову пространстве λ-квазигеодезическая γ лежит в λ^2 окрестности геодезической σ с теми же концами. Более того, эта геодизическая σ находится в $\log \lambda$-окрестности квазигеодизической γ. Эта оценка также является оптимальной [16*, 14*].

- формализована численная задача проблемы квазиизометричности, указаны несколько важных факторов, позволяющих получить различные результаты и оценки для квазиизометрического искажения. В частности, исследовано поведение объемов и связности. Затем исследован перенос неравенства Пуанкаре при квазиизометрических отображениях, а также даны точные оценки сверху и снизу для гомотопического роста искажений для нескольких классов гиперболических пространств. Исследованы свойства квазиизометрических вложений деревьев в гиперболическую плоскость [15*].

Более подробно, доказана следующая теорема:

Теорема 5 *Пусть γ является (λ, c)-квазигеодезической в δ-гиперболическом пространстве E и пусть σ является геодезическим сегментом, соединяющем ее концы. Тогда γ лежит в H-окрестности σ, где*

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

и A_1 - некоторая универсальная константа.

Этот результат является оптимальным. Иными словами, нами был найден пример геодезической, самая удаленная точка которой лежит на расстоянии $\lambda^2 c/4$ от соответствующего геодезического сегмента. Далее, была доказана следующая теорема, являющаяся в некотором смысле двойственной к теореме 5.

Теорема 6 *Пусть γ является (λ, c)-квазигеодезической в δ-гиперболическом пространстве E и пусть σ является геодезическим сегментом, соединяющем ее концы. Пусть также $4\delta << \ln \lambda$. Тогда σ лежит в H_{am}-окрестности γ, где*

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

где A_2 некоторая универсальная константа.

Теоремы 5 и 6 позволили получить нетривиальные оценки квазиизометрического искажения для максимального смещения точек пространства X ауто-квазиизометриями $X \to X$, фиксирующими его границу. Далее, было предложено три подхода к интерпретации количественной задачи квазиизометрического искажения в масштабе R. Пусть X и Y - два метрических пространства с выделенными точками x_0 и y_0 соответственно. Для данного $R > 0$ рассматриваются три семейства отображения

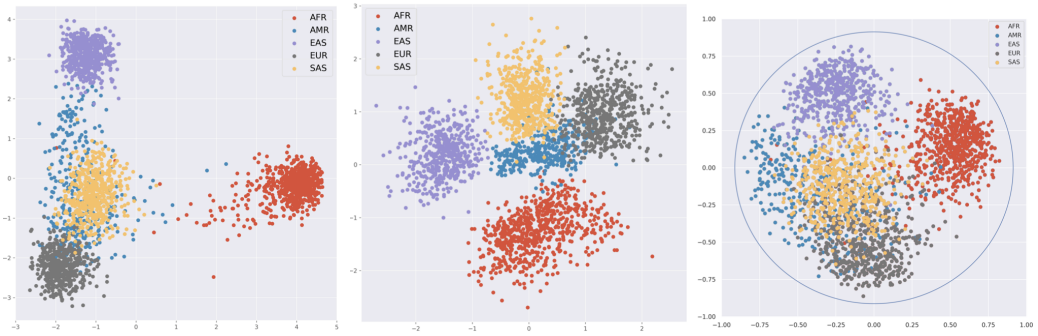
- квазиизометрии из шара $B_X(x_0, R)$ на шар $B_X(x_0, R)$,
- квазиизометрии из шара $B_X(x_0, R)$ на шар $B_X(x_0, \rho(R))$ для некоторой функции $\rho : \mathbb{R}_+ \to \mathbb{R}_+$,
- квазиизометрическое вложение шара $B_X(x_0, R)$ в Y.

Изучение переноса неравенств Пуанкаре квазиизометриями позволило получить нижнюю оценку для (λ, c)-квазиизометрического искажения между шарами радиуса R в локально однородных пространств отрицательной кривизны вида $Z_\mu = \mathbb{T}^n \times \mathbb{R}$ с метрикой $dt^2 + \sum_i e^{2\mu_i t} dx_i^2$ $(0 \leq \mu_1 \leq \ldots \leq \mu_n)$. Следующую теорему приведем здесь без технических деталей, точную формулировку можно найти в статье [15*].

Теорема 7 *Любое (λ, c)-квазиизометрическое вложение шара радиуса R из Z_μ в $Z_{\mu'}$ удовлетворяет неравенству*

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

Мы применили гиперболическую геометрию и глубинное обучение при анализе генетичеких данных в работе [17*]. А именно мы применили вариационные автокодировщики (VAE) с евклидовым и гиперболическим латентными пространствами для кластеризации геномов представителей различных современных человеческих популяций. Обычно для этой задачи используется метод главных компонент. Вариационные автокодировщики (VAE) позволяют проводить нелинейную кластеризацию данных в отличие от метода главных компонент (PCA), широко используемого в популяционном анализе. Сравнение результатов применения VAE к пяти популяциям из проекта 1000 геномов [57] представлены на Рис. 11.



(a) PCA (b) VAE (c) HVAE

Рис. 11: Результаты применения PCA, VAE с евклидовым латентным пространством и VAE с гиперболическим латентным пространством (HVAE) к представителям пяти популяций из проекта 1000 геномов.

Список литературы

[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] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[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] 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.

[21] 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.

[22] Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011 Jul;475(7357):493-6. Available from: https://doi.org/10.1038/nature10231.

[23] 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.

[24] 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.

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

[26] 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.

[27] Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. Nature. 2009;461(7263):489-94.

[28] Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012 Nov;192(3):1065-93.

[29] Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011 Aug;28(8):2239-52.

[30] Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945-59.

[31] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Research. 2009;19(9):1655-64.

[32] Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. The American Journal of Human Genetics. 2013;93(2):278-88.

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

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

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

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

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

[43] 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`.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] Sekizuka T, Itokawa K, Kageyama T, Saito S, Takayama I, Asanuma H, et al. Haplotype networks of SARS-CoV-2 infections in the Diamond Princess 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.

[55] 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.

[56] 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.

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