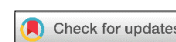


БИОИНФОРМАТИКА BIOINFORMATICS



УДК 57.087.1
DOI: 10.37661/1816-0301-2025-22-2-81-94

Оригинальная статья
Original Article

Программный комплекс для имитационного моделирования сайтов однонуклеотидного генетического полиморфизма

Н. Н. Яцков[✉], Д. Д. Сарнацкий, В. В. Скакун, В. В. Гринев

Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь
[✉]E-mail: yatskou@bsu.by

Аннотация

Цели. В настоящее время высокопроизводительные методы секвенирования широко используются в фундаментальных и прикладных исследованиях различных заболеваний человека. Секвенирование функционально значимых регионов генома человека позволяет одновременно идентифицировать множество сайтов генетического полиморфизма, имеющих диагностическую и (или) прогностическую значимость в отношении генетических заболеваний человека. В числе приоритетных целей в этой области стоит разработка эффективных программных инструментов обработки геномных данных и идентификации сайтов однонуклеотидного полиморфизма с использованием методов компьютерного моделирования и анализа больших данных.

Методы. Разработан программный комплекс для имитационного моделирования и идентификации сайтов однонуклеотидного полиморфизма с использованием методов машинного обучения. Реализована методика подхода имитационного моделирования и анализа сайтов однонуклеотидного полиморфизма в молекулах ДНК на основе бета-распределения или нормального закона распределения, параметры которых определяются по имеющимся экспериментальным данным, и методов интеллектуального анализа, обученных на смоделированных данных и применяемых для точной идентификации сайтов однонуклеотидного полиморфизма. Комплекс включает R-пакет, веб-приложение и вспомогательные программные средства для обработки экспериментальных данных геномного секвенирования.

Результаты. Проверка работоспособности представленного программного комплекса проведена на наборах смоделированных и экспериментальных данных геномного секвенирования клеток человека. Выполнен сравнительный анализ наиболее эффективных алгоритмов идентификации сайтов однонуклеотидных полиморфизмов. Наилучшие результаты получены для моделей машинного обучения.

Заключение. Применение программного комплекса повышает точность определения сайтов генетического полиморфизма в ходе анализа больших данных геномного секвенирования. Комплекс может использоваться для моделирования синтетических данных по экспериментальным данным или самостоятельно с целью всестороннего тестирования и выбора наилучших алгоритмов идентификации однонуклеотидных полиморфизмов, а также для генеративного моделирования данных, используемых при обучении алгоритмов идентификации на основе методов интеллектуального анализа.

Ключевые слова: однонуклеотидный генетический полиморфизм, программный комплекс, имитационное моделирование, машинное обучение, интеллектуальный анализ данных, R-пакет, веб-приложение

Благодарности. Работа выполнена в рамках государственной программы научных исследований «Конвергенция-2025» (грант № 3.04.3.1, № гос. регистрации 20211918).

Для цитирования. Программный комплекс для имитационного моделирования сайтов однонуклеотидного генетического полиморфизма / Н. Н. Яцков, Д. Д. Сарнацкий, В. В. Скакун, В. В. Гринев // Информатика. – 2025. – Т. 22, № 2. – С. 81–94. – DOI: 10.37661/1816-0301-2025-22-2-81-94.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 11.04.2025

Подписана в печать | Accepted 05.05.2025

Опубликована | Published 30.06.2025

Software complex for simulation modelling of single nucleotide genetic polymorphism sites

Mikalai M. Yatskou[✉], Dzianis D. Sarnatski, Victor V. Skakun, Vasily V. Grinev

Belarusian State University,
av. Nezavisimosti, 4, Minsk, 220030, Belarus

[✉]E-mail: yatskou@bsu.by

Abstract

Objectives. High-throughput sequencing methods have recently become widely used in the fundamental and applied research of various human diseases. Sequencing of functionally significant regions of the human genome enables the simultaneous identification of multiple genetic polymorphism sites that have diagnostic and/or prognostic significance for human genetic diseases. One of the key goals in this area is to develop efficient software tools for processing genomic data and identifying single nucleotide polymorphism sites using computer modelling and big data analysis methods.

Methods. A software complex has been developed for simulation modelling and identification of single nucleotide polymorphism sites using machine learning methods. The methods for the approach to simulation modelling and analysis of single nucleotide polymorphism sites in DNA molecules are implemented based on the beta or normal distributions, the parameters of which are determined from the available experimental data, and machine learning models trained on simulated data and used to accurately identify single nucleotide polymorphism sites. The software complex includes an R package, a web application, and auxiliary computational tools for processing experimental genomic sequencing data.

Results. The performance of the developed software complex was tested on sets of simulated and experimental data from human cell genomic sequencing. A comparative analysis of the most effective algorithms for identifying single nucleotide polymorphism sites was performed. The best results were obtained for machine learning models.

Conclusion. The use of the software complex increases the accuracy of identifying genetic polymorphism sites during the analysis of big genomic sequencing data. The software can be used for modelling synthetic data, based on experimental data or independently, for the purpose of comprehensive testing and selection of the best algorithms for identifying single nucleotide polymorphisms, as well as for generative data modelling used in training identification algorithms based on machine learning methods.

Keywords: single nucleotide genetic polymorphism, software complex, simulation modelling, machine learning, data mining, R package, web application

Acknowledgments. This work was carried out in the framework of the state programme of scientific research "Convergence-2025" (grant No. 3.04.3.1, state registration No. 20211918).

For citation. Yatskou M. M., Sarnatski D. D., Skakun V. V., Grinev V. V. *Software complex for simulation modelling of single nucleotide genetic polymorphism sites*. Informatika [Informatics], 2025, vol. 22, no. 2, pp. 81–94 (In Russ.). DOI: 10.37661/1816-0301-2025-22-2-81-94.

Conflict of interest. The authors declare of no conflict of interest.

Введение. Генетический полиморфизм влияет на фенотип человека и других живых организмов [1]. Одним из наиболее распространенных типов генетических вариаций в геноме человека является однонуклеотидный полиморфизм (от англ. single nucleotide polymorphism), когда индивидуальные геномы отличаются по отдельным нуклеотидам. При этом идентификация таких мест (сайтов) в геноме, по которым наблюдаются однонуклеотидные различия, имеет не только фундаментальное, но и сугубо прикладное значение, в частности для диагностики и прогнозирования течения онкологических заболеваний человека [2].

Среди существующих способов определения сайтов однонуклеотидного полиморфизма (СОП) следует отметить статистические методы точного теста Фишера, биномиального распределения, на основе энтропии и машинного обучения [1, 3, 4]. Методы достаточно универсальны и просты для программной реализации, однако вычислительно затратны и трудно применимы при анализе экспериментальных данных с высоким уровнем шума и различными экспериментальными искажениями, являющимися источниками пропусков, повторов и прочих аномальных явлений [5]. Поэтому для выбора наиболее эффективного алгоритма идентификации сайтов, проверки конкурирующих методик анализа и оценки производительности конкретных экспериментальных планов исследования биомолекулярных систем зачастую используется имитационное моделирование [6, 7].

Так, имитационное моделирование применяется при генерации синтетических данных для методов машинного обучения (или интеллектуального анализа данных) с целью прямой идентификации СОП различных организмов [4]. В этом случае формирование смоделированных обучающих данных может иметь преимущества по точности и эффективности при анализе экспериментальных данных как с невысоким числом покрытий, так и с наличием пропусков, обусловленных экспериментальными искажениями. Существующие подходы к математическому моделированию СОП основаны на учете параметров экспериментального оборудования, применении вероятностных моделей и статистических подходов, использовании вспомогательной биологической информации [8, 9]. Однако из-за различий в типах генетических данных, методах моделирования, эволюционных характеристиках, форматах данных, терминологии и допущениях, принятых в существующих программных приложениях, выбор надежного инструмента для конкретного исследования является весьма сложным и субъективным процессом [10]. Следует отметить, что немногие методы моделирования используют экспериментальные результаты или наблюдаемые характеристики для моделирования, включающие сложные схемы моделирования с учетом структуры и особенностей экспериментального шума.

На данный момент разработаны различные подходы и программные средства к математическому моделированию сайтов генетического полиморфизма [6, 7, 11–15]. Основными ограничениями существующих программных решений являются: отсутствие вычислительных процедур анализа и моделирования распределений измеряемых характеристик нуклеотидных сайтов, учитывающих особенности конкретных экспериментов; необходимость дополнительной реализации процедур генерации выборок нуклеотидных последовательностей для обучения классификационных моделей с целью последующей идентификации однонуклеотидных полиморфизмов; отсутствие специализированного R-пакета, интегрирующего основные алгоритмы обработки экспериментальных данных, моделирования и идентификации СОП, для широкого использования биоинформатическим сообществом. Следовательно, разработка программной платформы имитационного моделирования и интеллектуального анализа генетических полиморфизмов по данным геномного секвенирования по-прежнему является актуальной задачей.

Целью настоящей работы является создание программного комплекса для имитационного моделирования и идентификации СОП с использованием моделей машинного обучения, параметры которых оцениваются на базе имеющихся экспериментальных данных. Комплекс включает R-пакет, веб-приложение и вспомогательные программные средства для обработки экспериментальных данных геномного секвенирования. Он позволяет смоделировать особенности конкретных экспериментов и сформировать эталонные выборки для обучения классификационных моделей, что обязательно в алгоритмах машинного обучения (интеллектуального анализа). Работоспособность разработанных программных средств подтверждена в ходе сравнитель-

ного анализа наиболее эффективных алгоритмов идентификации СОП на примерах смоделированных и экспериментальных данных секвенирования генома человека.

Программный комплекс. В программном комплексе реализована методика подхода имитационного моделирования и анализа СОП в молекулах ДНК на основе законов распределений, параметры которых определяются по имеющимся экспериментальным данным, и методов интеллектуального анализа, обученных на смоделированных данных и применяемых для точной идентификации СОП [16]. Схема методики подхода представлена на рис. 1, а и включает пять этапов: построения и аппроксимации гистограмм числа нуклеотидных покрытий с помощью выбранных функций законов распределений; имитационного моделирования нуклеотидных покрытий в каналах регистрации нуклеотидных прочтений сайтов; генерации информативных признаков сайтов, характеризующих покрытия нуклеотидов А, С, G и Т; обучения моделей идентификации (интеллектуального анализа); идентификации СОП в смоделированных или экспериментальных данных с использованием настроенных моделей машинного обучения.

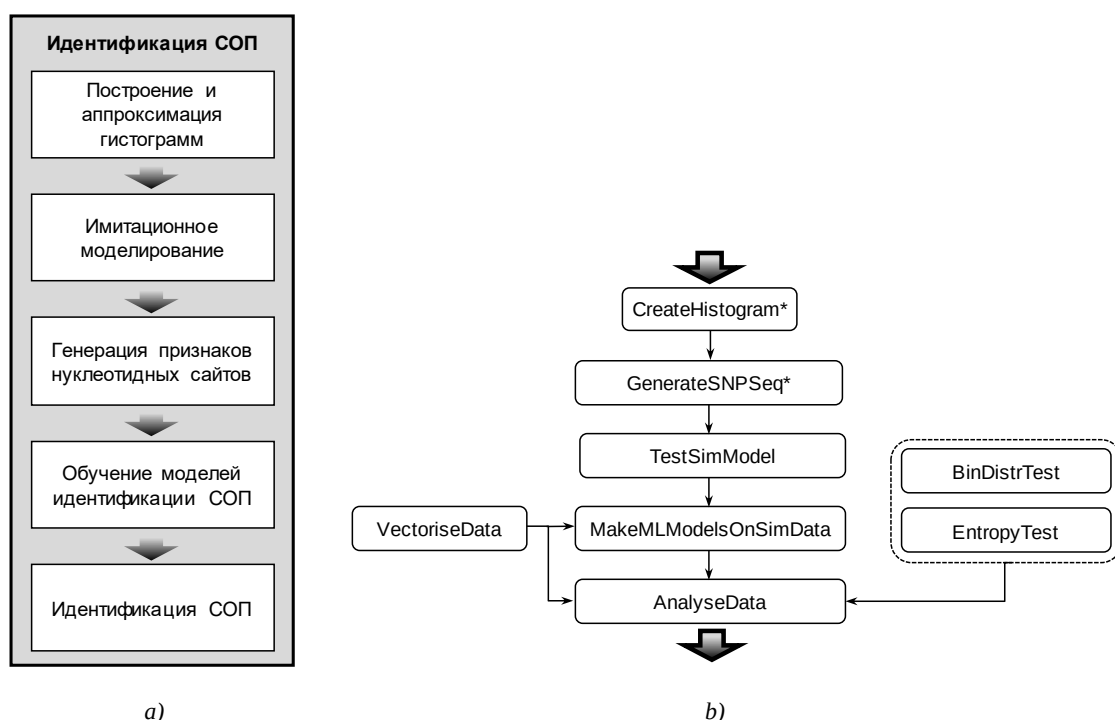


Рис. 1. Методика подхода имитационного моделирования и идентификации СОП нуклеотидных последовательностей (а) и схема организации взаимодействия функций в R-пакете (б).

*CreateHistogram** – обозначение функций *CreateHistogram1*, *CreateHistogram2*, *CreateHistogram3* и *CreateHistogramSNVs*; *GenerateSNPSeq** – обозначение функций *GenerateSNPSeqBeta*, *GenerateSNPSeqBetaGaussNoise*, *GenerateSNPSeqGauss* и *GenerateSNPSeqGaussGaussNoise*

Fig. 1. The methods of the approach to simulation modelling and identification of single nucleotide polymorphism (SNP) in nucleotide sequences (a) and the functional scheme in the R package (b).

*CreateHistogram** is the designation of the functions *CreateHistogram1*, *CreateHistogram2*, *CreateHistogram3* and *CreateHistogramSNVs*; *GenerateSNPSeq** is the designation of the functions *GenerateSNPSeqBeta*, *GenerateSNPSeqBetaGaussNoise*, *GenerateSNPSeqGauss* and *GenerateSNPSeqGaussGaussNoise*

В ходе работы выполняется анализ экспериментальных характеристик выбранного набора данных геномного секвенирования с целью определения закона распределения и оценки неизвестных параметров. Для аппроксимации чисел покрытий нуклеотидов сайтов предлагается использовать бета- или нормальное распределения. Результаты исследований позволяют сделать вывод о том, что бета-распределение удовлетворительно подходит для анализируемых интегральных характеристик экспериментального набора сайтов в геноме человека [17]. Нор-

мальное распределение – менее точное при исследовании генома человека [17], однако применимо к экспериментальным характеристикам, демонстрирующим гауссовость данных. Следует отметить возможность применения других типов вероятностных распределений или функций полиномиального сглаживания. Полученные экспериментальные оценки параметров распределений используются в имитационных моделях. Имитационное моделирование включает случайную генерацию позиций СОП в последовательности рассматриваемой молекулы, состоящей из фиксированного числа нуклеотидных сайтов, для каждого из которых воспроизводятся числа покрытий нуклеотидов по бета-распределению или нормальному распределению в заданном диапазоне. Экспериментальные гистограммы рассматриваются в качестве законов распределений. Для нуклеотидного сайта моделируется общее число покрытий четырех нуклеотидов, по которому затем генерируются числа покрытий для эталонного и альтернативного нуклеотидов. Аналогично моделируются покрытия для СОП. Принимается допущение о том, что каждый сайт может быть представлен не более чем двумя вариантами покрытий нуклеотидов – эталонным или альтернативным. Предложенный алгоритм позволяет воспроизводить наборы данных, максимально приближенные к эмпирическим условиям, задаваемым числами покрытий и законами их распределений, а также количеством полиморфных сайтов. Особенности подхода являются применение алгоритмов имитационного моделирования для воспроизведения регистрируемых характеристик конкретных экспериментов и генерация «реалистичных» выборок для обучения классификационных моделей.

Для реализации программного обеспечения используются различные вычислительные платформы и технологии программирования [18]. Основным преимуществом среды статистического программирования R является возможность использования огромного набора биоинформационных алгоритмов, алгоритмов интеллектуального анализа данных и разнообразных статистических вычислительных ресурсов научного сообщества¹ [19]. Для разработки программного комплекса в работе выбрана вычислительная среда R.

Отдельное направление в разработке R-приложений связано с созданием «реактивных» веб-интерфейсов с помощью пакета Shiny и размещением программной реализации на ресурсе shinyapps.io, предоставляемом разработчиками открытого программного обеспечения RStudio². Достоинством данного подхода является возможность удаленной работы с приложением широкой научной аудиторией пользователей в режиме онлайн через сеть Интернет. Для разработки веб-интерфейса приложения выбран пакет Shiny.

Программный комплекс включает R-пакет, веб-приложение и набор вспомогательных функций.

R-пакет. Разработанные R-функции, реализующие различные этапы анализа данных и построения имитационных моделей, интегрированы в отдельный R-пакет [20]. Назначение пакета – моделирование синтетических данных по экспериментальным данным или самостоятельно с целью всестороннего тестирования и выбора наилучших алгоритмов идентификации СОП, а также генеративного моделирования данных для обучения алгоритмов идентификации на основе методов машинного обучения. Функциональная схема и описание функций пакета представлены на рис. 1, *b* и в табл. 1.

Пакет включает функции: *CreateHistogram** – построения гистограмм чисел покрытий нуклеотидных сайтов, их аппроксимации с помощью бета- и нормального распределений (R-функции *dbeta* и *dnorm*) и оценки параметров распределений (R-функция *nls*); *GenerateSNPSeq** – моделирования нуклеотидных покрытий по бета- или нормальному распределениям с добавлением аддитивного гауссового шума; *TestSimModel* – проверки разработанных имитационных моделей; *MakeMLModelsOnSimData* – обучения моделей интеллектуального анализа на смоделированных данных; *AnalyseData* – идентификации СОП с использованием предобученных моделей или классических методов тестов биномиального распределения и на основе вычисления

¹R Core Team. R: A language and Environment for Statistical Computing / R Foundation for Statistical Computing. – 2024. – URL: <http://www.R-project.org/> (date of access: 10.04.2025).

²Posit team. RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. – 2024. – URL: <http://www.posit.co/> (date of access: 10.04.2025).

энтропии; *VectoriseData* – формирования четырех информативных признаков нуклеотидных сайтов для применения алгоритмов машинного обучения: X_1 – число покрытий эталонного нуклеотида, X_2 – X_4 – отсортированные в порядке убывания числа покрытий для альтернативных нуклеотидов (данные нормируются к общему числу покрытий сайта n). В качестве методов машинного обучения выбраны алгоритмы на основе деревьев условного вывода (функция *ctree* пакета *party*), классификации и регрессии построением дерева решений CART (функция *rpart* пакета *rpart*), машины опорных векторов (функция *svm* пакета *e1071*) и ансамблевого алгоритма XGBoost (функция *xgboost* пакета *xgboost*). В пакет включены модели *Classif_ctree_SimData.rds*, *Classif_rpart_SimData.rds*, *Classif_SVM_SimData.rds*, *Classif_xgboost_SimData.rds*, обученные на смоделированных данных по бета-закону распределения на выборке из 40 000 сайтов, из которых 20 000 являются СОП. Параметры моделей оценены на наборе данных геномного секвенирования молекул ДНК человека [21]. Эффективность алгоритмов оценивается с помощью мер точности *Precision*, чувствительности *Recall* и score F_1 , характеризующих свойства алгоритмов не включать ложноположительные события (*Precision*, неверно классифицированные сайты как СОП) и истинно положительные события (*Recall*, верно классифицированные как СОП), и их комбинированного вклада (F_1) [22].

Таблица 1
Функции R-пакета
Table 1
Functions of the R-package

R-функция <i>R function</i>	Описание <i>Description</i>	Результат <i>Value</i>
<i>AnalyseData</i>	Идентификация СОП в смоделированных или экспериментальных данных	Списки СОП
<i>BinRatioTest</i>	Программная реализация теста биномиального распределения	Список СОП, p -величины
<i>EntropyTest</i>	Программная реализация теста на основе энтропии	Список СОП, оценки энтропии E и p -величины
<i>CreateHistogram1</i> <i>CreateHistogram2</i> <i>CreateHistogram3</i> <i>CreateHistogramSNVs</i>	Построение гистограмм общего числа покрытий (*1), максимального числа покрытий (*2), разностей общего и максимального чисел покрытий (*3) нуклеотидных сайтов и СОП (*SNVs). Аппроксимация гистограмм производится с помощью бета- и нормального законов распределений, выполняется оценка параметров распределений	Оцененные параметры бета- и нормального распределений
<i>GenerateSNPSeqBeta</i> <i>GenerateSNPSeqBetaGaussNoise</i> <i>GenerateSNPSeqGauss</i> <i>GenerateSNPSeqGaussGaussNoise</i>	Имитационное моделирование сайтов и чисел их покрытий в нуклеотидной последовательности с использованием бета-распределения (*Beta), нормального распределения (*Gauss) и с добавлением гауссового шума (*BetaGaussNoise и *GaussGaussNoise)	Смоделированные данные, представленные символами референсного нуклеотида и числами покрытий сайта в каналах нуклеотидов А, С, G и Т
<i>MakeMLModelsOnSimData</i>	Создание и обучение моделей идентификации СОП на смоделированных данных	Модели машинного обучения
<i>TestSimModel</i>	Построение гистограмм чисел покрытий нуклеотидных сайтов по смоделированным данным. Аппроксимация гистограмм с помощью бета- и нормального распределений. Оценка параметров распределений	Оцененные параметры бета- и нормального распределений
<i>VectoriseData</i>	Генерация информативных признаков сайтов, характеризующих покрытия нуклеотидов А, С, G и Т, для обучения моделей классификации	Преобразованные к информативным признакам наборы данных

В ходе работы создан R-пакет SNPSimulatoR. Архив для установки – файл SNPSimulatoR_0.1.0.tar.gz – доступен по ссылке https://sstcenter.com/download/SNPSimulatoR_test_data. Примеры скриншотов подключения пакета SNPSimulatoR в RStudio показаны на рис. 2.

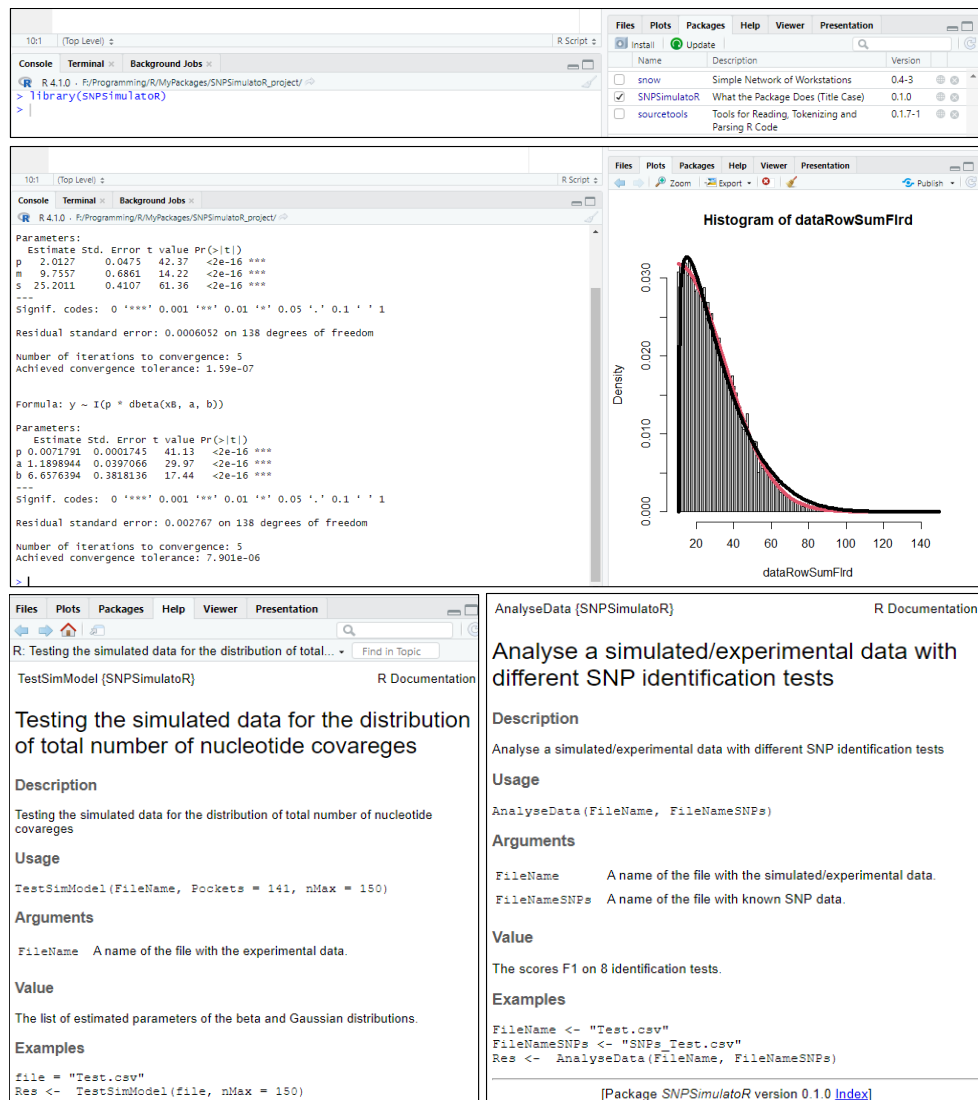


Рис. 2. Примеры подключения пакета, вывода результатов функции *CreateHistogram1* и представления информации о функциях *TestSimModel* и *AnalyseData*

Fig. 2. Examples of connecting the package, outputting the results of the function *CreateHistogram1*, and presenting information about the functions *TestSimModel* and *AnalyseData*

Веб-приложение. Разработанное веб-приложение SNPSimulatoR доступно по ссылке <https://dsa-cm.shinyapps.io/SNPSimulatoR>. Оно интегрирует реализованные алгоритмы вычислительного подхода. Примеры окон интерфейса веб-приложения представлены на рис. 3.

Главное окно интерфейса состоит из девяти панелей, соответствующих пяти этапам анализа: загрузки и моделирования данных; аппроксимации экспериментальных характеристик по бета-или нормальному распределением; обучения моделей на генеративных данных; идентификации однонуклеотидных полиморфизмов с помощью моделей машинного обучения; визуализации и интерпретации результатов. На каждом этапе анализа пользователь должен установить системные параметры алгоритмов, при необходимости выбрать соответствующий файл предобученной модели идентификации СОП. Результаты анализа сохраняются в отдельный csv-файл. Приложение предназначено для анализа экспериментальных и смоделированных данных.

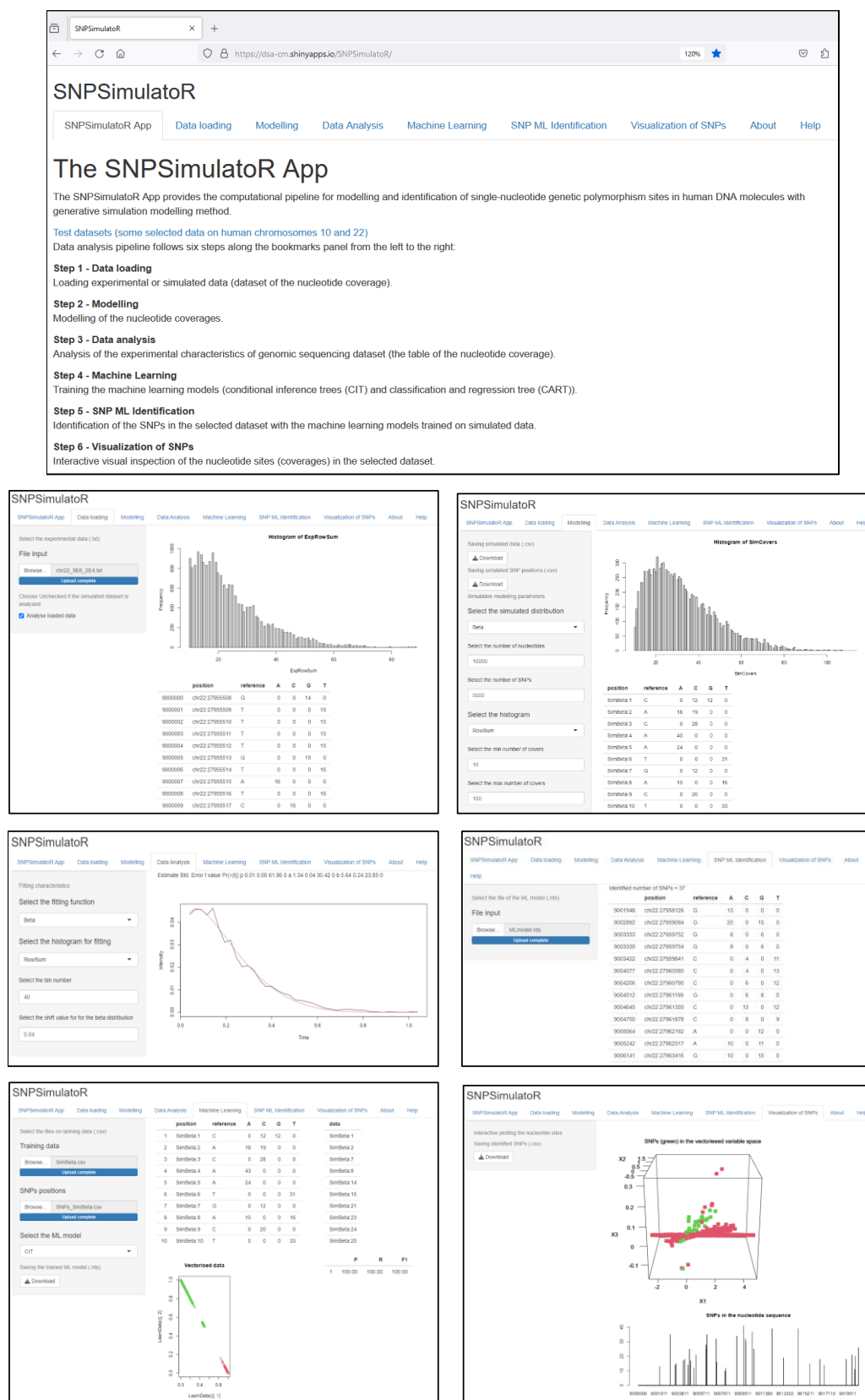


Рис. 3. Основные этапы исследования экспериментальных данных геномного секвенирования с использованием подхода на основе имитационного моделирования и интеллектуальной идентификации СОП
 Fig. 3. Main stages of the study of experimental genomic sequencing data using the approach based on simulation modelling and intelligent identification of single nucleotide polymorphism (SNP)

Вспомогательные функции. Реализованы дополнительные функции для преобразования BAM(Binary Alignment Map)-файлов экспериментальных данных геномного секвенирования в частотные таблицы для непосредственной обработки в программном комплексе, фильтрации данных и сохранения результатов анализа в специализированном VCF(Variant Call Format)-формате.

Проверка работоспособности программного комплекса. Работоспособность разработанных средств программного комплекса подтверждена в ходе сравнительного анализа наиболее эффективных алгоритмов идентификации СОП на примерах смоделированных и экспериментальных данных геномного секвенирования [23, 24]. Для всестороннего исследования алгоритмов идентификации СОП смоделированы данные с учетом добавления гауссового шума с параметрами $\mu = 0$, $\sigma_l = q_l \cdot b_l$, $l = 1-4$ (индексы соответствуют нуклеотидам А, С, G и Т), где $q_l > 0$, $\{b_1, b_2, b_3, b_4\}$ – набор из n прочтений нуклеотидных оснований А, С, G или Т, покрывающих сайт j . Варьирование параметром σ_l изменяет уровень экспериментального шума, а именно регулирует информативность полезного сигнала, что позволяет всесторонне исследовать эффективность разрабатываемых или выбранных алгоритмов идентификации СОП и воссоздавать специальные экспериментальные условия. В качестве экспериментального набора рассмотрены эталонные данные геномного секвенирования молекул ДНК человека, предоставленные консорциумом GIAB (Genome in a Bottle Consortium) [21]. Выбор данных GIAB обусловлен тем, что на сегодняшний день это наиболее надежные бенчмарк-данные для решения задач, связанных с изучением геномного полиморфизма у человека (от разработки новых инструментальных методов «мокрой» биологии до сравнения алгоритмов обнаружения полиморфных сайтов). Набор данных содержит характеристики 29 633 768 нуклеотидных сайтов, из которых 36 150 являются истинными СОП. Фрагмент набора данных представлен в табл. 2. Проведено исследование наиболее эффективных алгоритмов идентификации сайтов: тестов на основе биномиального распределения, энтропии и модифицированного точного теста Фишера, базовых методов машинного обучения – деревьев условного вывода, классификации и регрессии построением дерева решений CART, машины опорных векторов с линейной разделяющей функцией, обученных на имитационно-смоделированных данных. Эффективность алгоритмов оценена с помощью меры точности F_1 .

Таблица 2
Фрагмент экспериментального набора данных

Table 2
Fragment of the experimental dataset

Хромосома : позиция <i>Chromosome : position</i>	Эталонный нуклеотид <i>Reference nucleotide</i>	Покрытие <i>Coverage</i>			
		A	C	G	T
chr22 : 47891620	T	0	0	0	27
chr22 : 47891621	G	0	0	28	0
chr22 : 47891622	T	0	0	0	30

Исследование смоделированных данных в условиях возрастающего шума позволяет сделать вывод о том, что для незашумленных данных предпочтительнее использовать алгоритмы машинного обучения, а в условиях существенного зашумления – статистический тест на основе энтропии и модель машинного обучения CART [23]. При исследовании экспериментальных данных точность идентификации СОП по мере F_1 на 2–5 % выше у метода на основе дерева решений условного вывода, чем у сравниваемых методов [24]. Прочие модели машинного обучения и классические тесты идентификации имеют сопоставимую точность. Следует отметить, что точность моделей идентификации СОП, обученных на экспериментальных данных, не превышала 60–70 %. Невысокая точность может быть обусловлена двумя факторами: ограниченностью обучающей экспериментальной выборки или тем, что имитационная модель действительно лучше формирует обучающие данные, фокусируясь на воспроизведении важных источников информации в данных, и не учитывает второстепенные сигналы, присутствующие в реальных данных.

Ограничения и дальнейшее развитие программного комплекса. Программный комплекс имеет ряд ограничений, таких как зависимость от сторонних R-библиотек, медленная инициализация, неоптимальная обработка больших файлов данных. Эти недостатки будут устранены путем построения специализированных библиотек C++, оптимизации и распараллеливания кодов промежуточной обработки данных и вычислительных алгоритмов для работы с большими данными [25], а также включения автоматического обновления моделей машинного обучения. В дальнейшем планируется разработка усовершенствованных имитационных моделей на основе дополнительных законов распределений (например, гамма-распределения и распределения Вейбулла) для воспроизведения чисел покрытий в четырех каналах регистрации прочтений нуклеотидных сайтов с учетом статистических взаимосвязей между каналами [26], способов генерации дополнительных информативных признаков сайтов нуклеотидных последовательностей [27]. Кроме того, будет существенно расширен список предобученных моделей идентификации СОП, в том числе включением моделей нейронных сетей глубокого обучения [28].

Закключение. Разработан программный комплекс для моделирования и анализа сайтов нуклеотидных последовательностей по экспериментальным наборам данных, основанный на генерации случайных событий бета-распределения или нормального закона распределения, параметры которых оцениваются по имеющимся экспериментальным данным, и методов интеллектуального анализа, обученных на смоделированных данных и применяемых для точной идентификации сайтов однонуклеотидных полиморфизмов. Комплекс имеет следующие преимущества в сравнении с существующими программными решениями: обеспечивает моделирование данных, близко воспроизводящих реальные экспериментальные условия, с целью изучения надежности конкретных экспериментов и оценки точности результатов, полученных в наблюдаемых экспериментальных условиях; генерирует синтетические данные для обучения методов интеллектуального анализа и последующего создания моделей идентификации однонуклеотидных полиморфизмов в конкретных экспериментальных наборах данных; моделирует наборы данных для тестирования и сравнения доступных методов идентификации однонуклеотидных полиморфизмов при анализе экспериментальных данных; предоставляет возможность удаленной работы с приложением широкой научной аудитории пользователей в режиме онлайн через глобальную сеть Интернет.

Проверка работоспособности разработанных моделей и методов программного комплекса проведена на примерах наборов смоделированных и экспериментальных данных геномного секвенирования молекул ДНК человека. Выполнен сравнительный анализ наиболее эффективных алгоритмов идентификации сайтов однонуклеотидных полиморфизмов. Наилучшие результаты получены для моделей машинного обучения – точность идентификации сайтов по мере F_1 на 2–5 % выше у обученных на смоделированных данных методов деревьев решений, чем у традиционных методов идентификации.

Комплекс может использоваться для моделирования синтетических данных по экспериментальным данным или самостоятельно с целью всестороннего тестирования и выбора наилучших алгоритмов идентификации СОП, а также для генеративного моделирования данных с целью обучения алгоритмов идентификации на основе методов машинного обучения (нейронных и байесовских сетей, ансамблевых алгоритмов и пр.).

Вклад авторов. *Н. Н. Яцков* предложил методику и разработал программный комплекс для имитационного моделирования и идентификации сайтов однонуклеотидного полиморфизма с использованием методов машинного обучения, подготовил текст статьи. *Д. Д. Сарнацкий* разработал R-функции программного комплекса. *В. В. Скакун* разработал веб-сайт, базу данных и репозиторий для сохранения результатов анализа. *В. В. Гринев* сформулировал задачу об идентификации сайтов однонуклеотидного полиморфизма и разработал R-функции программного комплекса. Все авторы приняли участие в обобщении, анализе и оформлении полученных результатов.

Список использованных источников

1. Sung, W. K. Algorithms for Next Generation Sequencing / W. K. Sung. – 1st ed. – N. Y. : Chapman & Hall / CRC, 2017. – 364 p.
2. Kappelmann-Fenzl, M. Next Generation Sequencing and Data Analysis / ed. M. Kappelmann-Fenzl. – 1st ed. – Cham : Springer, 2021. – 218 p.
3. Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications / X. L. Wu, J. Xu, G. Feng [et al.] // PLoS ONE. – Sept. 2016. – Vol. 11, no 9. – P. e0161719. – DOI: 10.1371/journal.pone.0161719.
4. Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants / W. Korani, J. P. Clevenger, Y. Chu, P. Ozias-Akins // Plant Genome. – Mar. 2019. – Vol. 12, iss. 1. – P. 180023. – DOI: 10.3835/plantgenome2018.05.0023.
5. Masoudi-Nejad, A. Next Generation Sequencing and Sequence Assembly. Methodologies and Algorithms / A. Masoudi-Nejad, Z. Narimani, N. Hosseinkhan. – 1st ed. – N. Y. : Springer, 2013. – 86 p.
6. Su, Z. HAPGEN2: simulations of multiple disease SNPs / Z. Su, J. Marchini, P. Donnelly // Bioinformatics. – 2011. – Vol. 27, iss. 16. – P. 2304–2305.
7. Oh, J. H. SITDEM: a simulation tool for disease/endpoint models of association studies based on single nucleotide polymorphism genotypes / J. H. Oh, J. O. Deasy // Computers in Biology and Medicine. – 2014. – Vol. 45. – P. 136–142.
8. A comparison of gene region simulation methods / A. E. Hendricks, J. Dupuis, M. Gupta [et al.] // PLoS ONE. – 2012. – Vol. 7, no 7. – P. e40925. – DOI: 10.1371/journal.pone.0040925.
9. Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators / B. Peng, H. S. Chen, L. E. Mechanic [et al.] // Bioinformatics. – 2013. – Vol. 29, iss. 8. – P. 1101–1102.
10. Genetic data simulators and their applications: an overview / B. Peng, H. S. Chen, L. E. Mechanic [et al.] // Genetic Epidemiology. – 2015. – Vol. 39, iss. 1. – P. 2–10.
11. Tahmasbi, R. GeneEvolve: a fast and memory efficient forward-time simulator of realistic whole-genome sequence and SNP data / R. Tahmasbi, M. C. Keller // Bioinformatics. – 2017. – Vol. 33, iss. 2. – P. 294–296.
12. Posada, D. Simulating haplotype blocks in the human genome / D. Posada, C. Wiuf // Bioinformatics. – 2003. – Vol. 19, iss. 2. – P. 289–290.
13. DHOEM: a statistical simulation software for simulating new markers in real SNP marker data / L. Jacquin, T. V. Cao, C. Grenier, N. Ahmadi // BMC Bioinformatics. – Dec. 2015. – Vol. 16. – P. 404. – DOI: 10.1186/s12859-015-0830-7.
14. Meyer, H. V. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships / H. V. Meyer, E. Birney // Bioinformatics. – 2018. – Vol. 34, iss. 17. – P. 2951–2956.
15. sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs / A. Dimitromanolakis, J. Xu, A. Krol, L. Briollais // BMC Bioinformatics. – Jan. 2019. – Vol. 20, no 1. – P. 26. – DOI: 10.1186/s12859-019-2611-1.
16. Яцков, Н. Н. Генеративное имитационное моделирование сложных биофизических систем / Н. Н. Яцков, В. В. Апанасович, В. Н. Яцков // Компьютерные технологии и анализ данных (CTDA'2024) : материалы IV Междунар. науч.-практ. конф., Минск, 25–26 апр. 2024 г. – Минск : БГУ, 2024. – С. 211–214.
17. Simulation modelling for machine learning identification of single nucleotide polymorphisms in human genomes / M. M. Yatskou, E. V. Smolyakova, V. V. Skakun, V. V. Grinev // Pattern Recognition and Information Processing (PRIP'2023) : Proc. of the 16th Intern. Conf., Minsk, 17–19 Oct. 2023. – Minsk : BSU, 2023. – P. 49–53.
18. Яцков, Н. Н. Вычислительный подход и программный пакет RNAexploreR для группировки молекул РНК генов человека по их экзонным признакам / Н. Н. Яцков, В. В. Скакун, В. В. Гринев // Информатика. – 2019. – Т. 16, № 4. – С. 7–24.
19. Gentleman, R. Bioconductor: Open software development for computational biology and bioinformatics / R. Gentleman, V. J. Carey, D. M. Bates // Genome Biology. – 2004. – Vol. 5, no. 10, art. R80. – URL: <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-10-r80> (date of access: 10.04.2025). – DOI: 10.1186/GB-2004-5-10-R80.
20. Программный пакет SNPSimulatoR для моделирования сайтов однонуклеотидного генетического полиморфизма / Н. Н. Яцков, Е. В. Смолякова, В. В. Скакун, В. В. Гринев // Квантовая электроника : материалы XIV Междунар. науч.-техн. конф., Минск, 21–23 нояб. 2023 г. – Минск : БГУ, 2023. – С. 510–515.
21. An open resource for accurately benchmarking small variant and reference calls / J. M. Zook, J. McDaniel, N. D. Olson [et al.] // Nature Biotechnology. – 2019. – Vol. 37, no. 5. – P. 561–566.

22. Идентификация сайтов однонуклеотидного генетического полиморфизма с использованием методов машинного обучения / Н. Н. Яцков, Е. В. Смолякова, К. И. Грудовик [и др.] // Квантовая электроника : материалы XIV Междунар. науч.-техн. конф., Минск, 21–23 нояб. 2023 г. – Минск : БГУ, 2023. – С. 504–509.
23. Identification of single nucleotide genetic polymorphism sites using machine learning methods / M. M. Yatskou, E. V. Smolyakova, V. V. Skakun, V. V. Grinev // *Advances in Transdisciplinary Engineering*. – 2023. – Vol. 42. – P. 1031–1037.
24. Yatskou, M. M. Simulation modelling of single nucleotide genetic polymorphisms / M. M. Yatskou, V. V. Apanasovich, V. V. Grinev // *Journal of the Belarusian State University. Mathematics and Informatics*. – 2024. – No. 2. – P. 104–112.
25. Яцков, Н. Н. Вычислительная платформа FluorSimStudio для обработки кинетических кривых затухания флуоресценции с использованием алгоритмов имитационного моделирования и интеллектуального анализа данных / Н. Н. Яцков, В. В. Апанасович // *Журнал прикладной спектроскопии*. – 2021. – Т. 88, № 3. – С. 452–461.
26. Сарнацкий, Д. Д. Имитационная модель генерации сайтов однонуклеотидного полиморфизма в молекулах ДНК человека / Д. Д. Сарнацкий, Н. Н. Яцков, В. В. Гринева // *Компьютерные технологии и анализ данных (СТДА'2024) : материалы IV Междунар. науч.-практ. конф., Минск, 25–26 апр. 2024 г.* – Минск : БГУ, 2024. – С. 265–268.
27. Сарнацкий, Д. Д. Исследование информативности признаков нуклеотидных сайтов при определении генетических полиморфизмов с использованием методов машинного обучения / Д. Д. Сарнацкий, Н. Н. Яцков, В. В. Гринева // *Информационные технологии и системы (ИТС 2024) : материалы Междунар. науч. конф., Минск, 20 нояб. 2024 г.* – Минск : БГУИР, 2024. – С. 69–70.
28. Яцков, Н. Н. Нейросетевое имитационное моделирование при анализе экспериментальных данных флуоресцентной спектроскопии / Н. Н. Яцков, В. В. Апанасович // *Компьютерные технологии и анализ данных (СТДА'2024) : материалы IV Междунар. науч.-практ. конф., Минск, 25–26 апр. 2024 г.* – Минск : БГУ, 2024. – С. 215–218.

References

1. Sung W. K. *Algorithms for Next Generation Sequencing*, 1st ed. New York, Chapman & Hall / CRC, 2017, 364 p.
2. Kappelmann-Fenzl M. (ed.). *Next Generation Sequencing and Data Analysis*, 1st ed. Cham, Springer, 2021, 218 p.
3. Wu X. L., Xu J., Feng G., Wiggans G. R., Taylor J. F., ..., Bauck S. Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications. *PLoS ONE*, September 2016, vol. 11, no 9, p. e0161719. DOI: 10.1371/journal.pone.0161719.
4. Korani W., Clevenger J. P., Chu Y., Ozias-Akins P. Machine learning as an effective method for identifying true single nucleotide polymorphisms in polyploid plants. *Plant Genome*, March 2019, vol. 12, iss. 1, p. 180023. DOI: 10.3835/plantgenome2018.05.0023.
5. Masoudi-Nejad A., Narimani Z., Hosseinkhan N. *Next Generation Sequencing and Sequence Assembly. Methodologies and Algorithms*, 1st ed. New York, Springer, 2013, 86 p.
6. Su Z., Marchini J., Donnelly P. HAPGEN2: simulations of multiple disease SNPs. *Bioinformatics*, 2011, vol. 27, iss. 16, p. 2304–2305.
7. Oh J. H., Deasy J. O. SITDEM: a simulation tool for disease/endpoint models of association studies based on single nucleotide polymorphism genotypes. *Computers in Biology and Medicine*, 2014, vol. 45, pp. 136–142.
8. Hendricks A. E., Dupuis J., Gupta M., Logue M. W., Lunetta K. L. A comparison of gene region simulation methods. *PLoS ONE*, 2012, vol. 7, no 7, p. e40925. DOI: 10.1371/journal.pone.0040925.
9. Peng B., Chen H. S., Mechanic L. E., Racine B., Clarke J., ..., Feuer E. J. Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators. *Bioinformatics*, 2013, vol. 29, iss. 8, pp. 1101–1102.
10. Peng B., Chen H. S., Mechanic L. E., Racine B., Clarke J., ..., Feuer E. J. Genetic data simulators and their applications: an overview. *Genetic Epidemiology*, 2015, vol. 39, iss. 1, pp. 2–10.
11. Tahmasbi R., Keller M. C. GeneEvolve: a fast and memory efficient forward-time simulator of realistic whole-genome sequence and SNP data. *Bioinformatics*, 2017, vol. 33, iss. 2, pp. 294–296.
12. Posada D., Wiuf C. Simulating haplotype blocks in the human genome. *Bioinformatics*, 2003, vol. 19, iss. 2, pp. 289–290.

13. Jacquin L., Cao T. V., Grenier C., Ahmadi N. DHOEM: a statistical simulation software for simulating new markers in real SNP marker data. *BMC Bioinformatics*, December 2015, vol. 16, p. 404. DOI: 10.1186/s12859-015-0830-7.
14. Meyer H. V., Birney E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*, 2018, vol. 34, iss. 17, pp. 2951–2956.
15. Dimitromanolakis A., Xu J., Krol A., Briollais L. sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, January 2019, vol. 20, no. 1, p. 26. DOI: 10.1186/s12859-019-2611-1.
16. Yatskou M. M., Apanasovich V. V., Yatskou U. M. *Generative simulation modelling of complex biophysical systems*. Kompyuternye tehnologii i analiz dannyh (CTDA'2024): materialy IV Mezhdunarodnoj nauchno-prakticheskoy konferencii, Minsk, 25–26 aprelya 2024 g. [Computer Technologies and Data Analysis (CTDA'2024): Proceedings of the IV International Scientific Conference, Minsk, 25–26 April 2024]. Minsk, Belorusskij gosudarstvennyj universitet, 2024, pp. 211–214 (In Russ.).
17. Yatskou M. M., Smolyakova E. V., Skakun V. V., Grinev V. V. Simulation modelling for machine learning identification of single nucleotide polymorphisms in human Genomes. *Pattern Recognition and Information Processing (PRIP'2023) : Proceedings of the 16th International Conference, Minsk, 17–19 October 2023*. Minsk, Belarusian State University, 2023, pp. 49–53.
18. Yatskou M. M., Skakun V. V., Grinev V. V. A computational approach and software package RNAexploreR for grouping RNA molecules of human genes by exon features. *Informatika [Informatics]*, 2019, vol. 16, no. 4, pp. 7–24 (In Russ.).
19. Gentleman R., Carey V. J., Bates D. M. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 2004, vol. 5, no. 10, art. R80. Available at: <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2004-5-10-r80> (accessed 10.04.2025). DOI: 10.1186/GB-2004-5-10-R80.
20. Yatskou M. M., Smolyakova E. V., Skakun V. V., Grinev V. V. *R-package SNPSimulator for modelling single nucleotide genetic polymorphism sites*. Kvantovaya elektronika: materialy XIV Mezhdunarodnoj nauchno-tehnicheskoy konferencii, Minsk, 21–23 noyabrya 2023 g. [Quantum Electronics: Proceedings of the XIV International Scientific and Practical Conference, Minsk, 21–23 November 2023]. Minsk, Belorusskij gosudarstvennyj universitet, 2023, pp. 510–515 (In Russ.).
21. Zook J. M., McDaniel J., Olson N. D., Wagner J., Parikh H., ..., Salit M. An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 2019, vol. 37, no 5, pp. 561–566.
22. Yatskou M. M., Smolyakova E. V., Grudovik K. I., Skakun V. V., Grinev V. V. *Identification of single nucleotide genetic polymorphisms using machine learning methods*. Kvantovaya elektronika: materialy XIV Mezhdunarodnoj nauchno-tehnicheskoy konferencii, Minsk, 21–23 noyabrya 2023 g. [Quantum Electronics: Proceedings of the XIV International Scientific and Practical Conference, Minsk, 21–23 November 2023]. Minsk, Belorusskij gosudarstvennyj universitet, 2023, pp. 504–509 (In Russ.).
23. Yatskou M. M., Smolyakova E. V., Skakun V. V., Grinev V. V. Identification of Single nucleotide genetic polymorphism sites using machine learning methods. *Advances in Transdisciplinary Engineering*, 2023, vol. 42, pp. 1031–1037.
24. Yatskou M. M., Grinev V. V., Apanasovich V. V. Simulation modelling of single nucleotide genetic polymorphisms. *Journal of the Belarusian State University. Mathematics and Informatics*, 2024, no. 2, pp. 104–112.
25. Yatskou M. M., Apanasovich V. V. Computational platform FluorSimStudio for processing kinetic curves of fluorescence decay using simulation modeling and data mining algorithms. *Journal of Applied Spectroscopy*, 2021, vol. 88, no. 3, pp. 452–461.
26. Sarnatski D. D., Yatskou M. M., Grinev V. V. *Simulation model of generation of single nucleotide polymorphism sites in human DNA molecules*. Kompyuternye tehnologii i analiz dannyh (CTDA'2024): materialy IV Mezhdunarodnoj nauchno-prakticheskoy konferencii, Minsk, 25–26 aprelya 2024 g. [Computer Technologies and Data Analysis (CTDA'2024): Proceedings of the IV International Scientific Conference, Minsk, 25–26 April 2024]. Minsk, Belorusskij gosudarstvennyj universitet, 2024, pp. 265–268 (In Russ.).
27. Sarnatski D. D., Yatskou M. M., Grinev V. V. *Study of the informativeness of nucleotide site features in determining genetic polymorphisms using machine learning methods*. Informacionnye tehnologii i sistemy (ITS 2024): materialy Mezhdunarodnoj nauchnoj konferencii, Minsk, 20 noyabrya 2024 g. [Information Technologies and Systems 2024 (ITS 2024): Proceedings of the International Scientific Conference, Minsk, 20 November 2024]. Minsk, Belorusskij gosudarstvennyj universitet informatiki i radioelektroniki, 2024, pp. 69–70 (In Russ.).
28. Yatskou M. M., Apanasovich V. V. *Neural network simulation modelling when analyzing experimental fluorescence spectroscopy data*. Kompyuternye tehnologii i analiz dannyh (CTDA'2024): materialy IV Mezhdunarodnoj nauchno-prakticheskoy konferencii, Minsk, 25–26 aprelya 2024 g. [Computer Technologies and Data Analysis (CTDA'2024): Proceedings of the IV International Scientific Conference, Minsk, 25–26 April 2024]. Minsk, Belorusskij gosudarstvennyj universitet, 2024, pp. 215–218 (In Russ.).

Информация об авторах

Яцков Николай Николаевич, кандидат физико-математических наук, доцент, доцент кафедры системного анализа и компьютерного моделирования, факультет радиоп физики и компьютерных технологий, Белорусский государственный университет.
E-mail: yatskou@bsu.by

Сарнацкий Денис Дмитриевич, студент кафедры системного анализа и компьютерного моделирования, факультет радиоп физики и компьютерных технологий, Белорусский государственный университет.
E-mail: denisiussarnatski@gmail.com

Скакун Виктор Васильевич, кандидат физико-математических наук, доцент, заведующий кафедрой системного анализа и компьютерного моделирования, факультет радиоп физики и компьютерных технологий, Белорусский государственный университет.
E-mail: skakun@bsu.by

Гринеv Василий Викторович, кандидат биологических наук, доцент, доцент кафедры генетики, биологический факультет, Белорусский государственный университет.
E-mail: grinev_vv@bsu.by

Information about the authors

Mikalai M. Yatskou, Ph. D. (Phys.-Math.), Assoc. Prof., Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University.
E-mail: yatskou@bsu.by

Dzianis D. Sarnatski, Student, Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University.
E-mail: denisiussarnatski@gmail.com

Victor V. Skakun, Ph. D. (Phys.-Math.), Assoc. Prof., Head of Department of Systems Analysis and Computer Modelling, Faculty of Radiophysics and Computer Technologies, Belarusian State University.
E-mail: skakun@bsu.by

Vasily V. Grinev, Ph. D. (Biol.), Assoc. Prof., Department of Genetics, Faculty of Biology, Belarusian State University.
E-mail: grinev_vv@bsu.by