Усольцев Дмитрий Андреевич

Методы вероятностной кластеризации для интерпретации матричной и графовой информации в разнородных базах данных

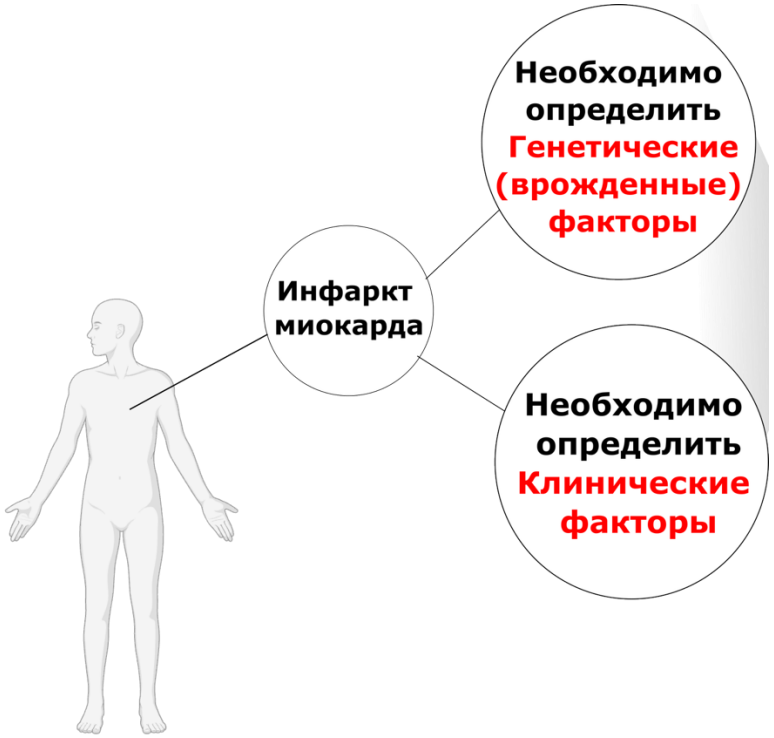
Диссертация на соискание учёной степени кандидата технических наук
Специальность 2.3.8.
«Информатика и информационные процессы (технические науки)»

Научный руководитель: доктор техн. наук, профессор Шалыто Анатолий Абрамович

Санкт-Петербург – 2025

Актуальность 1/10

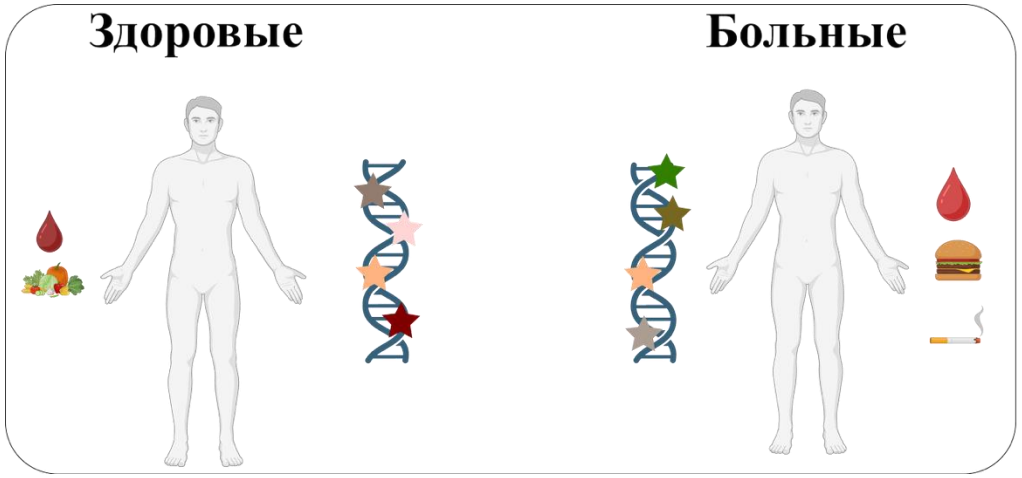
Проблема: ранняя диагностика рисков хронических заболеваний



Анализ крупных массивов информации

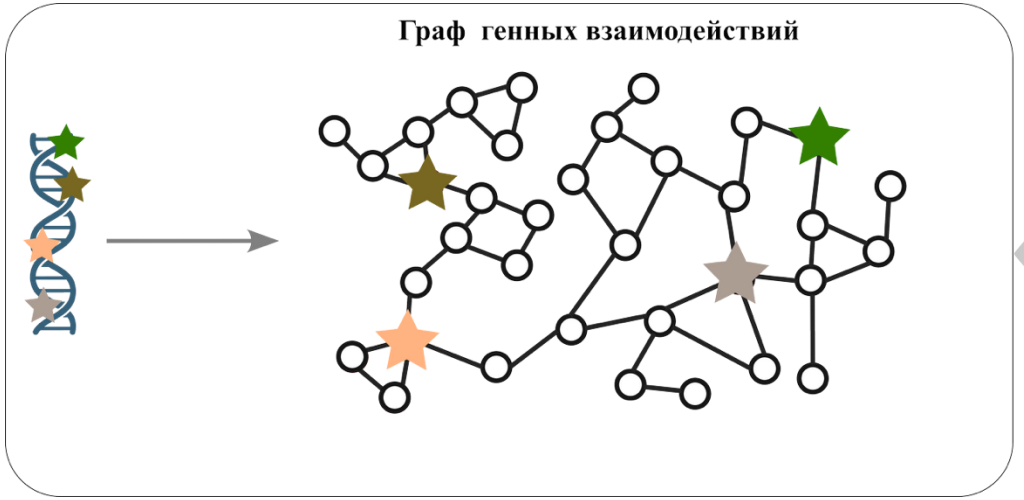
Для определения этих факторов хронических заболеваний необходимо создание больших баз данных с помощью специальных методов.

Актуальность 2/10
Медико-генетические базы данных



Матричная информация

Популяционная выборка большого числа индивидов
-Генотипирование ДНК
-демография
-анкеты
-лабораторные показатели



Графовая информация

-взаимосвязи между генами или белками

Исходная информация для формирования медико-генетической базы данных

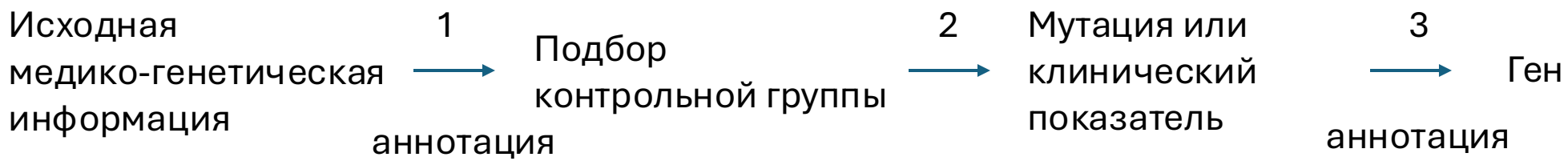
Современные медико-генетические базы данных представляют собой гибридные хранилища, объединяющие матричные и графовые форматы данных.

Актуальность 3/10
Аннотация информации

Аннотация (интерпретация) является важным процессом при формировании медико-генетической базы данных.

Аннотация – процесс присвоения объектам базы данных стандартизированных функциональных значений, обладающих смыслом. Например, если в строке указан идентификатор пациента, аннотация заключается в сопоставлении ему группы (контрольной или тестовой). Обычно аннотация выполняется вручную, однако в диссертации предлагаются методы, автоматизирующие этот процесс. Одна из трудностей аннотации состоит в том, что база данных может содержать десятки тысяч строк.

Этапы формирования медико-генетической базы данных



1. Отбор контрольной группы для проведения сравнения между больными и здоровыми индивидами.
2. Статистический тест между контрольной и тестовой группами по частотам мутаций или уровню клинических показателей.
3. Присвоение генов мутациям и биомаркерам, выявленным в ходе статистического теста.

Актуальность 4/10
Аннотация информации. Примеры

Примеры аннотирования информации

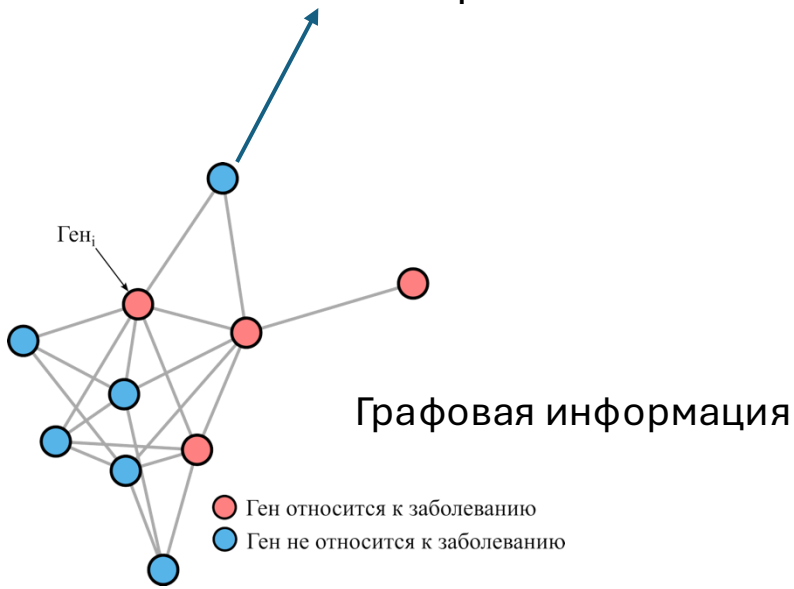
Исходная медико-генетическая информация → (1) Подбор контрольной группы → (2) Мутация или клинический показатель → (3) Ген

аннотация

Матричная информация

Индивиды	Группа
$Индивид_1$	Контрольная
$Индивид_2$	Тестовая
$Индивид_3$	Контрольная
…	…

аннотация



Графовая информация

Актуальность 5/10

Методы кластеризации

Аннотация может быть выполнена методами кластеризации.

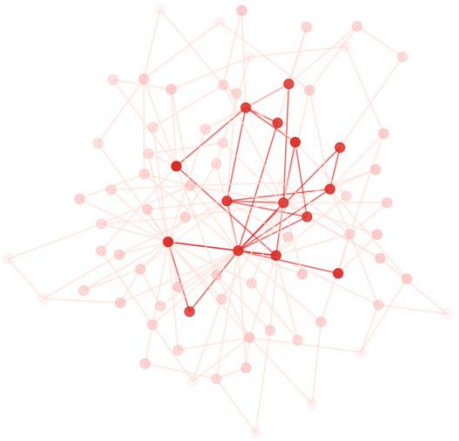
Актуальна разработка и улучшение точности **методов вероятностной кластеризации** для аннотирования матричной и графовой информации в медико-генетических базах данных.

Виды методов кластеризации

Дискретная кластеризация

Вершины однозначно распределены по кластерам: каждая вершина принадлежит только одному кластеру

Вероятностная кластеризация



Каждой вершине соответствует распределение вероятностей принадлежности к кластерам

Актуальность 6/10

Требования к методам кластеризации

1. Масштабируемость – эффективная работа с большими объемами данных.
2. Наличие вероятностной оценки – наличие вероятностной интерпретации полученных результатов.
3. Устойчивость к шумным данным – корректная работа при наличии выбросов и аномальных наблюдений в реальных данных.
4. Возможность интеграции априорных знаний – возможность учитывать экспертную информацию и внешние источники знаний.

Актуальность 7/10

Недостатки существующих методов кластеризации

Подходы	Масштабируемость	Вероятностная оценка	Устойчивость к шуму	Возможность интеграции априорных знаний
k-means[1]	Да	Нет	Нет	Нет
DBSCAN[2]	Да	Нет	Да	Нет
метод Монте-Карло по схеме марковских цепей (МЦМК) (Markov Chain Monte Carlo – MCMC)[3]	Да	Да	Да	Частично (начальное распределение вероятностей на вершинах)
Подход на основе поиска подграфа максимального веса (Maximum Weight Connected Subgraph – MWCS)[4]	Да	Нет	Нет	Нет
Метод графовой кластеризации для совместного анализа результатов генотипирования и экспрессии генов (диссертация **Александра Лободы**)[5]	Да	Нет	Нет	Нет

1. MacQueen J. Some methods for classification and analysis of multivariate observations // In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. V. 1. N 14. P. 281–297.
2. Ester M., Kriegel H.P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise // In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. 1996. P. 226–231.
3. Alexeev N., Isomurodov J., Sukhov V., et al. Markov chain Monte Carlo for active module identification problem // BMC Bioinformatics. 2020. V. 21. Article number: 261.
4. Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T., Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach // Bioinformatics. 2008. V. 24. P. 223–231.
5. Лобода А.А. Метод графовой кластеризации для совместного анализа данных генотипирования и экспрессии генов Диссертация на соискание ученой степени кандидата технических наук / Лобода А.А. Университет ИТМО. 2022.

Актуальность 8/10

Цель и задачи диссертационной работы

Повышение качества кластеризации в задачах матричного и графового анализа разнородных баз данных за счет использования априорных знаний, что позволит в таблице, приведенной на предыдущем слайде во всех столбцах ответить “Да”.

Задачи, решаемые в диссертации:

– обзор современных методов вероятностной кластеризации матричных и графовых данных, а также статистических методов, применяемых для разработки медико-генетических баз данных с разнородной информацией;
– разработка вероятностного метода матричной кластеризации на основе априорного распределения признаков для формирования несмещенного набора данных;
– разработка вероятностного метода графовой кластеризации на основе методов Монте-Карло по схеме марковских цепей и априорной информации, интегрируемой с помощью машинного обучения, для выявления кластеров, обладающих целевыми признаками;
– апробация и внедрение результатов исследования на основе методики, использующей предложенные вероятностные методы для учёта априорных знаний с целью повышения качества кластеризации при анализе баз данных с разнородной информацией.

Положения, выносимые на защиту

1. **Вероятностный метод кластеризации для формирования набора данных**, использующий сингулярное разложение матриц, отличающийся тем, что **с целью снижения вероятности ложноположительных результатов при сравнении двух наборов данных**, в нём применяются расстояние Махаланобиса и априорное распределение признаков.
2. **Вероятностный метод графовой кластеризации**, использующий метод Монте-Карло по схеме марковских цепей, отличающийся тем, что **с целью повышения точности кластеризации**, в нем используется априорная информация, учитываемая с помощью машинного обучения, и несколько источников априорных знаний.
3. **Методика анализа данных в разнородных базах**, основанная на сравнительном анализе и последующей графовой кластеризации его результатов, отличающаяся тем, что **с целью повышения качества кластеризации**, она включает предложенные вероятностные методы матричной и графовой кластеризации.

Научной новизной обладают первые два положения, а третье – имеет важное практическое значение.

Соответствие паспорту специальности 2.3.8

Тринадцатый пункт паспорта специальности: «Разработка и применение методов распознавания образов, **кластерного анализа**, нейро-сетевых и нечетких технологий, решающих правил, мягких вычислений при анализе разнородной информации в базах данных».

Предложен **вероятностный метод матричной кластеризации** с априорным распределением признаков, сокращающий систематические смещения при формировании контрольных выборок.

Разработан **вероятностный метод графовой кластеризации** с использованием методов Монте-Карло по схеме марковских цепей и машинного обучения. Разработанные методы были применены автором, в том числе к разработанной в ходе диссертационного исследования медико-генетической базе данных Российской популяции (МГБД-Р), и показали более высокое качество кластеризации в сравнении с аналогами.

Первое положение, выносимое на защиту 1/10

Вероятностный метод кластеризации для формирования набора данных, использующий сингулярное разложение матриц, отличающийся тем, что **с целью снижения вероятности ложноположительных результатов при сравнении двух наборов данных**, в нём применяются расстояние Махаланобиса и априорное распределение признаков.

Постановка задачи

Задача

Предположим, имеется два набора данных – тестовый и контрольный. В тестовом наборе собрана информация о пациентах, у которых наблюдается определённое заболевание или особый фенотип, а в контрольном наборе данные большой популяции людей без данного диагноза. Возникает задача найти в обширном контрольном наборе подгруппу, которая будет максимально похожа на тестовую группу по ключевым факторам, чтобы избежать систематических искажений при сравнительном анализе.

Аналоги: метод ближайшего соседа[1], DBSCAN[2]

Прототип: метод ближайшего соседа

Недостатки прототипа: отсутствует вероятностная оценка, не учитывается априорное предположение о распределении признаков в тестовом наборе данных

1. Ikotun A. M., Ezugwu A. E., Abualigah L., Abuhaija B. and Heming J. K-means clustering algorithms: a comprehensive review, variants analysis, and advances in the era of big data // Inf. Sci. 2023. V. 622. P. 178–210.
2. Jianwei Li, Anna Zheng, Wei Guo, Nairwita Bandyopadhyay, Yanji Zhang & Qianfeng Wang Urban flood risk assessment based on DBSCAN and K-means clustering algorithm // Geomatics, Natural Hazards and Risk. 2023. V. 14. N. 1. Article Number: 2250527.

Предлагаемый метод

Метод состоит из двух этапов:

1. Снижение размерности тестового и контрольного наборов данных.

2. Вероятностная кластеризация в пространстве признаков главных компонент.

Первое положение, выносимое на защиту 4/10

Снижение размерности тестового и контрольного наборов данных (1/2)

Пусть A_1 и A_2 – матрицы признаков; $\dim(A_1) = n_1 \times m$, $\dim(A_2) = n_2 \times m$, где m – число различных характеристик (например, возраст, пол, различные лабораторные показатели, социальные факторы), n_1 – число участников в тестовой группе, n_2 – в контрольной.

1.1. В тестовом наборе $A_1 \in \mathbb{R}^{n_1 \times m}$ вычисляются средние значения по каждому столбцу. Эти средние значения затем вычитаются из обеих матриц A_1 и $A_2 \in \mathbb{R}^{n_2 \times m}$, что обеспечивает выравнивание выборок по среднему значению признаков и устраняет глобальные линейные смещения.

1.2. К матрице A_1 применяется сингулярное разложение[1]: $A_1 = U \Sigma V^T$, где $U \in \mathbb{R}^{n_1 \times n_1}$ – матрица левых сингулярных векторов, Σ – диагональная матрица сингулярных чисел, $V \in \mathbb{R}^{m \times m}$ – матрица правых сингулярных векторов. Из матрицы U выбираются первые k столбцов, соответствующих наибольшим сингулярным числам. Полученная матрица $\hat{U} \in \mathbb{R}^{n_1 \times k}$ определяет проекцию в подпространство наибольшей дисперсии.

1.3. Проекции обоих наборов данных вычисляются следующим образом: $\hat{A}_1 = \hat{U}A_1$ и $\hat{A}_2 = \hat{U}A_2$, соответственно. Каждая строка матрицы $\hat{A}_1$ и $\hat{A}_2$ теперь представляет наблюдение в новом пространстве признаков размерности k общем для обеих выборок. Это обеспечивает сопоставимость обоих наборов данных.

1. Brunton S.L., Kutz J.N. Singular Value Decomposition (SVD) // Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge: Cambridge University Press. 2019. P. 3–46.

Первое положение, выносимое на защиту 5/10

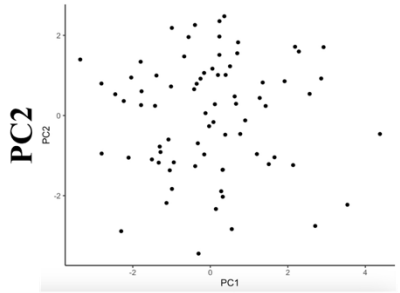
Снижение размерности тестового и контрольного наборов данных (2/2)

Тестовый набор данных 1.1

Фенотипы	Показатель$_1$	Показатель$_2$	…
Индивид$_1$	**10**	**20**	…
Индивид$_2$	**0**	**1**	…
Индивид$_3$	**0,2**	**0,9**	…
…	…	…	…

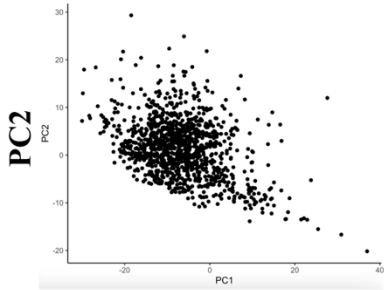
Сингулярное разложение → 1.2

Двумерная проекция тестового набора данных в пространство главных компонент



1.3

Двумерная проекция контрольного набора данных в пространство главных компонент тестового набора данных



PC1 – первая главная компонента
PC2 – вторая главная компонента

1.1. Вычитание средних значений из тестового набора данных.
1.2. Сингулярное разложение тестового набора данных.
1.3. Проекции тестового и контрольного наборов данных в пространство главных компонент.

Вероятностная кластеризация в пространстве признаков главных компонент (1/2)

2.1. **Оценка априорного распределения признаков в тестовой выборке**. После проекции на пространство главных компонент $\hat{A}_1$, компоненты тестовой выборки рассматриваются как реализация случайной величины, подчиняющейся многомерному нормальному распределению. Оцениваются вектор математических ожиданий $\mu \in \mathbb{R}^k$ и ковариационная матрица $\Sigma \in \mathbb{R}^{k \times k}$, задающие априорное распределение признаков: $\hat{A}_1 \sim N(\mu, \Sigma)$.

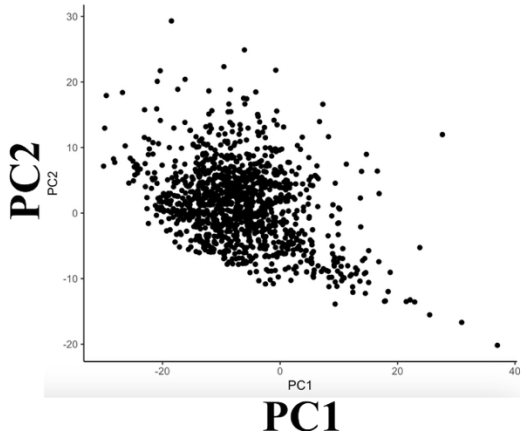
2.2. **Формулировка вероятностного критерия принадлежности**. Для каждого объекта $x \in \hat{A}_2$ вычисляется расстояние Махаланобиса[1] как мера правдоподобия его принадлежности к распределению тестовой выборки: $D_M(x) = \sqrt{(x-\mu)^T \Sigma^{-1}(x-\mu)}$. Таким образом, каждый объект из контрольной выборки получает количественную оценку степени принадлежности к априорному распределению.

2.3. **Выделение вероятностного кластера**. В отличие от детерминированного отнесения объектов к кластерам, используется вероятностный критерий: объект включается в кластер $\hat{S}_2$, если значение $D_M(x)$ не превышает порог, соответствующий 95 %-квантилю распределения хи-квадрат с k степенями свободы. Это соответствует включению в доверительную область уровня 0.95 для многомерного нормального распределения. Таким образом, кластер $\hat{S}_2 \subseteq A_2$ формируется как $\hat{S}_2 = \{x \in A_2 \mid D_M(\hat{U}^T x) \leq \chi^2_{k,1-\alpha}\}$.

1. Ghosh A., Ghosh A.K., SahaRay R., Sarkar S. Classification using global and local Mahalanobis distances // Journal of Multivariate Analysis. 2025. V. 207. Article number: 105417.

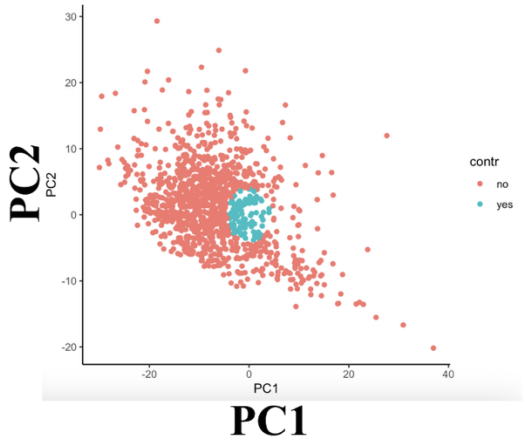
Вероятностная кластеризация в пространстве признаков главных компонент (2/2)

2.1 Вычисление средних значений и ковариационной матрицы первых k компонент



2.2 Формулировка вероятностного критерия принадлежности

2.3 Выделение вероятностного кластера



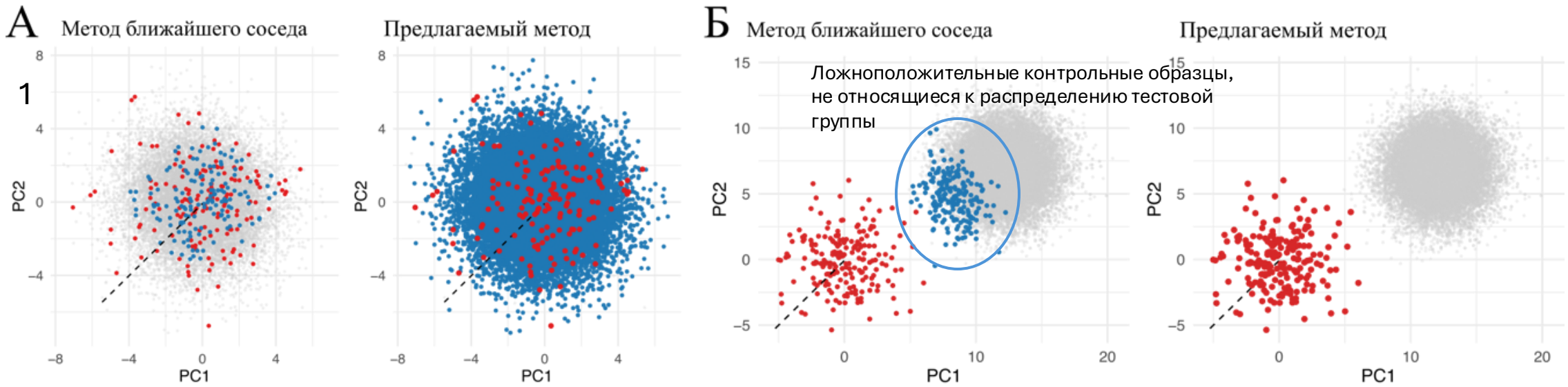
PC1 – первая главная компонента
PC2 – вторая главная компонента

2.1. Оценка априорного распределения признаков в тестовой выборке.
2.2. Формулировка вероятностного критерия принадлежности.
2.3. Выделение вероятностного кластера.

Экспериментальная валидация метода

Контрольная и тестовая группы симулируются из одного нормального распределения

Контрольная и тестовая группы симулируются из разных нормальных распределений

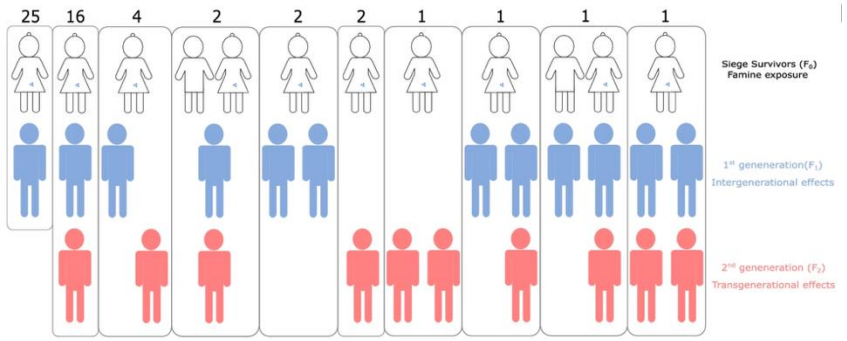


А. Метод ближайшего соседа отбирает заданное число контрольных образцов, в то время как предлагаемый метод отбирает все контрольные образцы из распределения, соответствующего тестовой выборке.

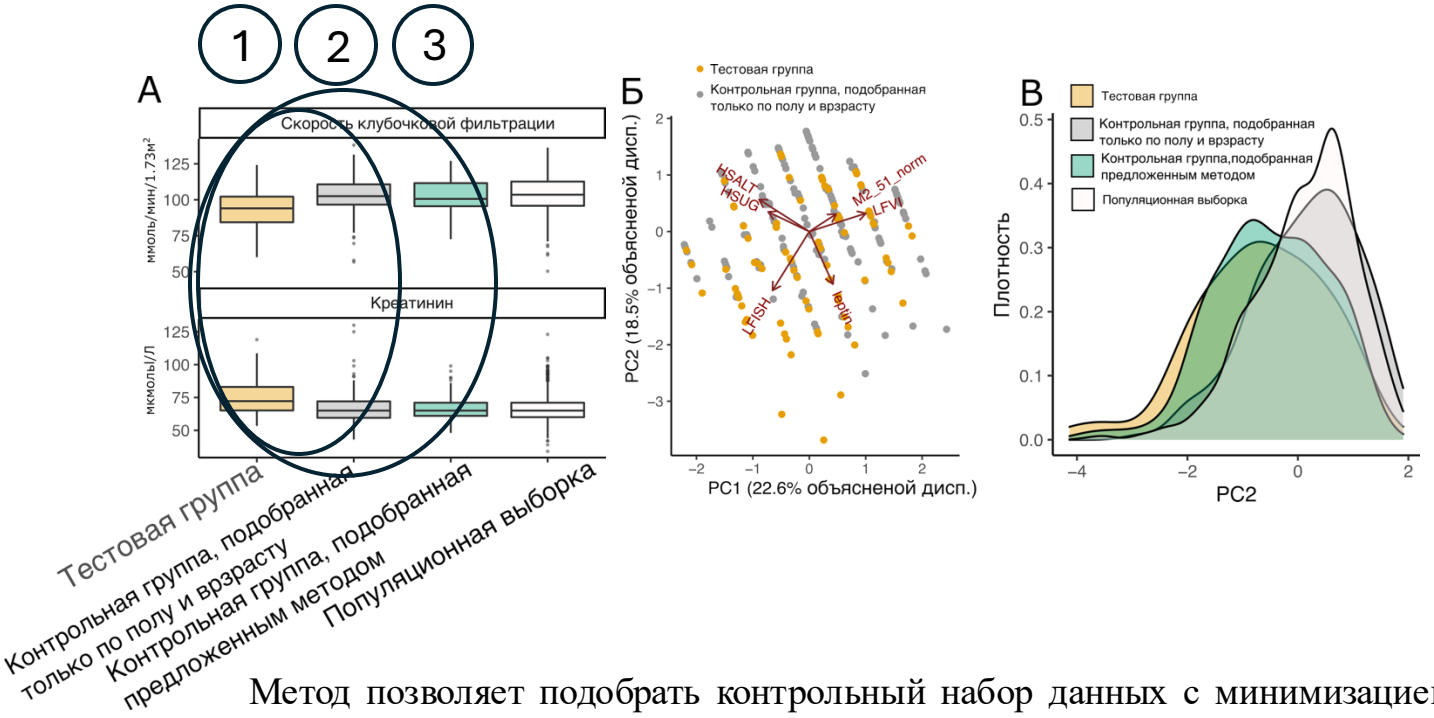
PC1 – первая главная компонента
PC2 – вторая главная компонента

Б. Метод ближайшего соседа отбирает ложно-положительные контрольные образцы из распределения, не соответствующего тестовому распределению, в то время как предложенный метод не отбирает такие контрольные образцы.

Практическое применение метода



Исследование потомков жителей блокадного Ленинграда



Клиническое исследование:

1. Тестовая группа (желтая) – потомки жителей блокадного Ленинграда с особым паттерном питания (Рисунок А).
2. Контрольная группа (серая), подобранная без предложенного метода (отличалась от тестовой группы по распределению факторов питания) (Рисунок А)
3. Контрольная группа (зеленая) – выборка из популяции с помощью предложенного метода (схожая с тестовой группой по распределению факторов питания) (Рисунок Б).
4. **Предложенный метод позволил подтвердить отсутствие ложно-положительных результатов в сравнительном анализе работы почек между потомками жителей блокадного Ленинграда и контрольной популяцией.**

Метод позволяет подобрать контрольный набор данных с минимизацией систематических отличий по ключевым признакам относительно тестового набора данных. При сравнении этих наборов данных результирующая статистическая значимость (*p*-значение) является более достоверной.

Выводы по первому положению, выносимому на защиту

1. Предлагаемый метод позволяет подобрать контрольный набор данных с минимизацией систематических отличий по выбранным признакам относительно тестового набора данных за счет учета априорных знаний о распределении признаков в тестовом наборе данных и вероятностной оценки принадлежности наблюдений к кластерам. Это особенно важно в медицинских исследованиях (например, в эпидемиологических), где точность статистических выводов имеет первостепенное значение.
2. Предлагаемый метод был использован при исследовании межпоколенческих эффектов потомков жителей блокадного Ленинграда на особенности их физиологии. Исследование проведено в Национальном медицинском исследовательском центре имени В.А. Алмазова.

Научные результаты опубликованы в:

1. **Усольцев Д.А.** Вероятностный метод матричной кластеризации с априорным распределением признаков для формирования несмещенной контрольной группы // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 999–1001.
2. **Usoltsev D**., Tolkunova K., Moguchaia E. *et al.* Transgenerational and intergenerational effects of early childhood famine exposure in the cohort of offspring of Leningrad Siege survivors // Scientific Reports. 2023. V. 13. Article number: 11188.

Второе положение, выносимое на защиту 1/14

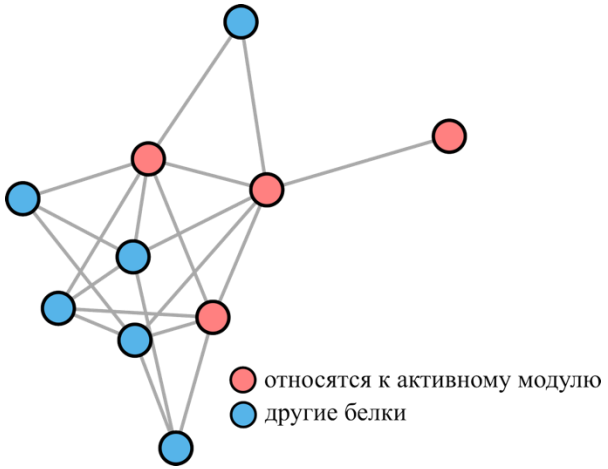
Вероятностный метод графовой кластеризации, использующий метод Монте-Карло по схеме марковских цепей, отличающийся тем, что **с целью повышения точности кластеризации**, в нем используется априорная информация, учитываемая с помощью машинного обучения, и несколько источников априорных знаний.

Задача графовой кластеризации

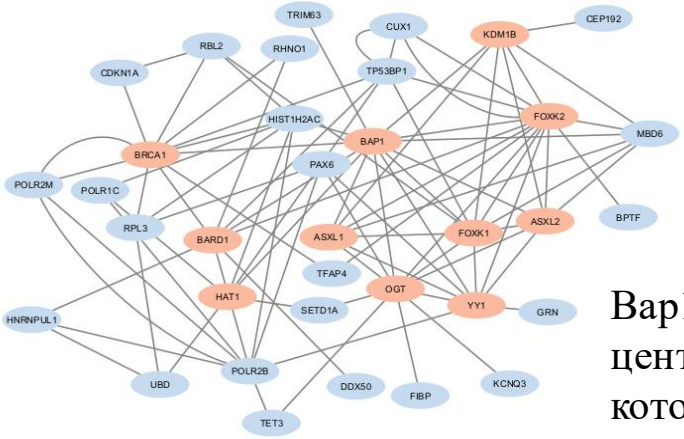
Графовая кластеризации состоит в том, чтобы по заданному множеству взаимосвязанных объектов выделить группы (кластеры) вершин, которые максимально похожи друг на друга и существенно отличаются от вершин других групп.

Задача графовой кластеризации сводится к задаче поиска в графе функционального подграфа (кластера) – активного модуля, вершины которого выполняют конкретную совместную функцию. Важно повысить точность нахождения этого модуля.

Схема активного модуля внутри белковой сети



Пример активного модуля. Bap1-комплекс[1], участвующий в восстановлении ДНК. Красным показаны белки, образующие Bap1-комплекс.



Bap1 – белковый комплекс, центральным компонентом которого является белок Bap1

Как правило, заранее известны некоторые вершины функционального кластера

1. Carbone M., Yang H., Pass H.I., Krausz T., Testa J.R., Gaudino G. BAP1 and cancer // Nat Rev Cancer. 2013. V. 13. N 3. P. 153–159.

Текущие решения задачи графовой кластеризации

Аналоги: метод Монте-Карло по схеме марковских цепей[1], метод подграфа наибольшего веса[2,3], метод Фишера[4]

Прототип: метод Монте-Карло по схеме марковских цепей

Недостаток прототипа: прототип не в полной мере учитывает априорную информацию об известных вершинах

1. Alexeev N., Isomurodov J., Sukhov V., Korotkevich G., Sergushichev A. Markov chain Monte Carlo for active module identification problem // BMC Bioinformatics. 2020. V. 21. Article number: 261.
2. Лобода А.А. Метод графовой кластеризации для совместного анализа данных генотипирования и экспрессии генов Диссертация на соискание ученой степени кандидата технических наук / Лобода А.А. Университет ИТМО. 2022.
3. Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T., Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach // Bioinformatics. 2008. V. 24. P. 223–231.
4. Ham H., Park T. Combining p-values from various statistical methods for microbiome data // Front Microbiol. 2022. V. 13. Article number: 990870.

Описание метода (1/6)

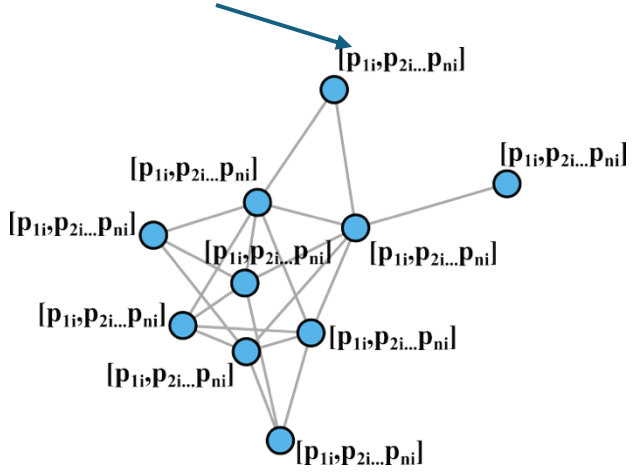
1. Учет нескольких источников экспериментов

1.1. Для каждого набора *p*-значений, соответствующих вершинам графа, применяется метод Фишера: вычисляется $X^2 = -2\sum_{i=1}^{n} \ln(p_i),\ 0 < p_i \le 1,$ и результирующие *p*-значения по формуле $p_{comb} = P(\chi^2 \ge X^2, df = 2*n),$

n – число экспериментов

Статистическая значимость из одного эксперимента

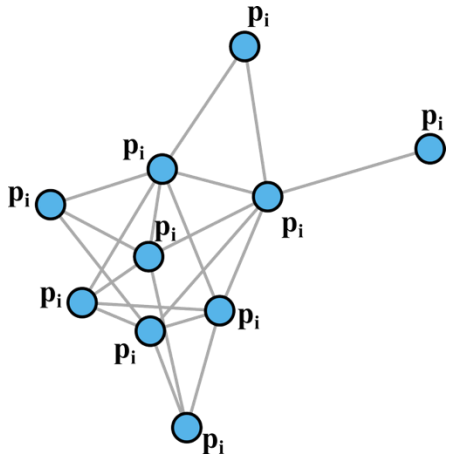
Результирующая статистическая значимость



Метод Фишера[1]

$$X^2 = -2\sum_{i=1}^{n} \ln(p_i),\ 0 < p_i \le 1,$$

$$p_{comb} = P(\chi^2 \ge X^2, df = 2n)$$



p_i – p-значение i-го эксперимента;
X^2 – статистика хи-квадрат;
χ^2 – распределение хи-квадрат;
p_{comb} – результирующие p-значение;
P – вероятность;
df – степень свободы.

1. Ham H., Park T. Combining p-values from various statistical methods for microbiome data // Front. Microbiol. 2022. V. 13. Article number: 990870.

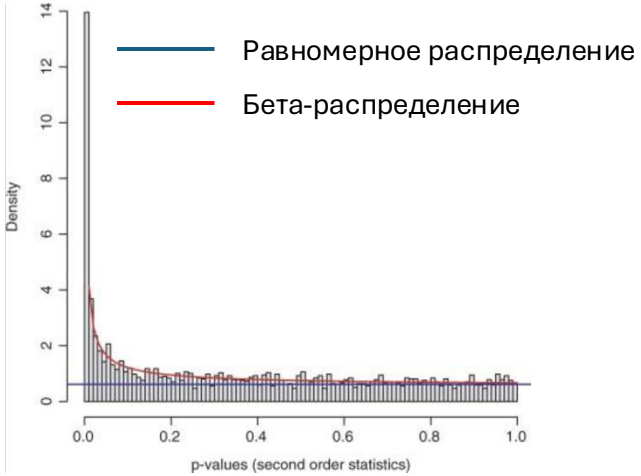
Описание метода (2/6). Вычисление признаков каждой вершины (1/4). Первый признак (1/3)

2.1. **Первый признак** – вероятность вхождения вершины в функциональный кластер, определяется по *p*-значениям. Для каждой вершины вероятность определяется посредством аппроксимации *p*-значений с помощью бета-равномерного распределения. Далее вероятность того, что подграф является функциональным кластером, рассчитывается как произведение вероятностей того, что каждая вершина принадлежит этому кластеру. Для получившегося вероятностного пространства на множестве связных подграфов набирается выборка размера 100 с помощью алгоритма Монте-Карло по схеме марковских цепей (МЦМК)[1], реализованного в R-библиотеке mcmcRanking (v0.1.0)[2], используя 10 000 итераций МЦМК. После этого для каждой вершины определяется эмпирическая вероятность вхождения в функциональный кластер – доля подграфов из выборки, которые включают в себя эту вершину.

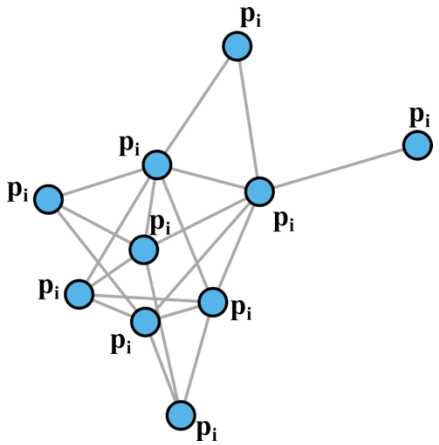
1. Alexeev N., Isomurodov J., Sukhov V., Korotkevich G., Sergushichev A. Markov chain Monte Carlo for active module identification problem // BMC Bioinformatics. 2020. V. 21. Article number: 261.
2. GitHub – ctlab/mcmcRanking: Tool To Solve The Active Module Problem. [Электронный ресурс] – Режим доступа: https://github.com/ctlab/mcmcRanking, свободный. Яз. англ. (дата обращения 15.12.2024).

Описание метода (3/6). Вычисление признаков каждой вершины (2/4). Первый признак (2/3)

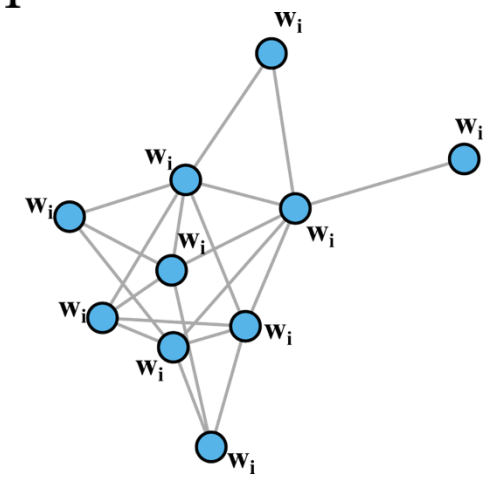
Эмпирическое распределение p-значений аппроксимируется бета-равномерным распределением[1]



1. $f(p) = \lambda + (1-\lambda)\alpha p^{\alpha-1},\ 0 \le \lambda,\ \alpha \le 1,\ 0 < p \le 1,$
2. $\mathrm{pFDR} = \left(\frac{\lambda + (1-\lambda)\alpha - \mathrm{FDR}^{*}\lambda}{\mathrm{FDR}(1-\lambda)}\right)^{\left(\frac{1}{\alpha-1}\right)},\ 0 \le \lambda,\ \alpha \le 1,$
3. $w(p) = \left(\frac{p}{\mathrm{pFDR}}\right)^{\alpha-1},\ 0 \le \alpha \le 1,\ 0 < p \le 1$



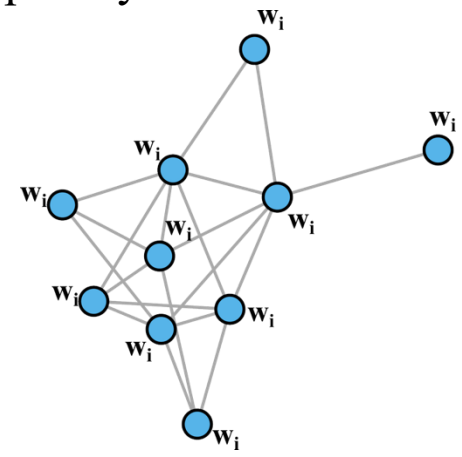
Аппроксимация p-значений (p) с помощью бета-равномерного распределения[1] с параметрами:
λ – вес равномерной компоненты;
α – параметр формы бета-компоненты для заданного уровня ложноположительных результатов (FDR).
pFDR – вероятность того, что p-значение принадлежит равномерному распределению;
w_i – вес i-ой вершины



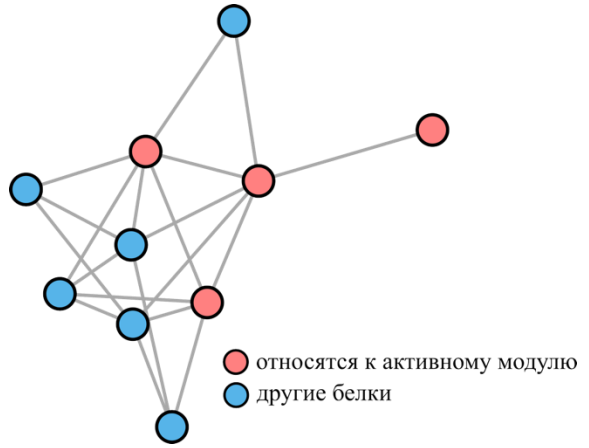
1. Dittrich M.T., Klau G.W., Rosenwald A., Dandekar T., Müller T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach // Bioinformatics. 2008. V. 24. P. 223–231.

Описание метода (4/6). Вычисление признаков каждой вершины (3/4). Первый признак (3/3)

Вероятность того, что подграф является функциональным кластером, рассчитывается как произведение вероятностей того, что каждая вершина принадлежит этому кластеру. Для получившегося вероятностного пространства на множестве связных подграфов набирается выборка размера 100 с помощью алгоритма МЦМК, реализованного в R-библиотеке mcmcRanking (v0.1.0), используя 10 000 итераций МЦМК. После этого для каждой вершины определяется эмпирическая вероятность вхождения в функциональный кластер – доля подграфов из выборки, которые включают в себя эту вершину.



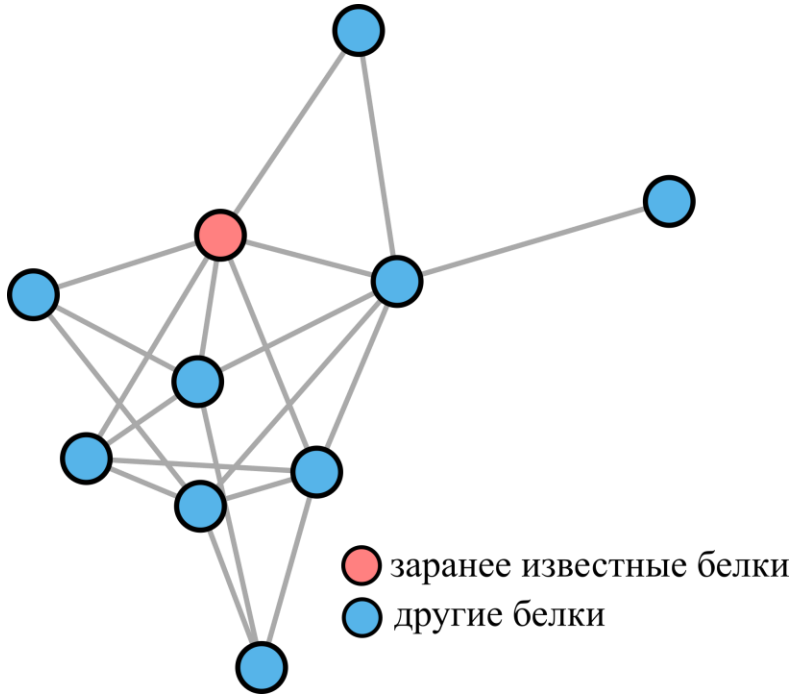
Метод Монте-Карло по схеме марковских цепей (МЦМК)



Вывод: применение метода МЦМК позволяет определить вероятность принадлежности вершины функциональному кластеру, однако метод не учитывает уже известные вершины. Поэтому на следующих слайдах этот недостаток устранен.

Описание метода (5/6). Вычисление признаков каждой вершины (4/4)

2.2. **Три признака**, описывающие каждую вершину, определяются как расстояния до трех ближайших известных вершин функционального кластера.



Вершины	Расстояние$_1$	…
Вершина$_1$	**2**	…
Вершина$_2$	**3**	…
Вершина$_3$	**1**	…
…	…	…

Описание метода (6/6)

3. Обучение модели градиентного бустинга для предсказания принадлежности вершины функциональному кластеру

Используя полученные признаки, обучается модель градиентного бустинга – xgboost[1]. Она реализована в R-библиотеке xgboost (v1.5.0.2)[2]. Максимальная глубина каждого дерева в модели равняется трем. Параметр скорости обучения – 0.1. Число итераций обучения равно трем. Известные вершины в функциональных кластерах исключаются. При тренировке модели вершинам, включенным в функциональный кластер, присваивается единица, а остальным вершинам – ноль.

Вершины	Расстояние$_1$	МЦМК	…
Вершина$_1$	**2**	**0.4**	…
Вершина$_2$	**3**	**0.2**	…
Вершина$_3$	**1**	**0.6**	…
…	…		…

Модель градиентного бустинга

→

Вершины	**Скор**
Вершина$_1$	**0.8**
Вершина$_2$	**0.7**
Вершина$_3$	**0.9**
…	…

1. Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York. NY. USA. ACM. 2016. P. 785–794.
2. Chen T., He T., Benesty M., Khotilovich V., et al. xgboost: Extreme Gradient Boosting. R package version 1.5.0.2. [Электронный ресурс] – Режим доступа: https://CRAN.R-project.org/package=xgboost, свободный. Яз. англ. (дата обращения 17.08.2024)

Валидация метода (1/4)

Валидация метода на симулированных данных

1. Для получения графов с топологией, встречающейся в реальных данных, выбираются 100 случайных подграфов из графа белок-белковых взаимодействий InWeb_InBioMap (InWebIM)[1] с числом вершин равным 1000.

2. В каждом подграфе InWebIM из пункта 1.1 равновероятно среди всех возможных связных подграфов заданного размера выбирается активный модуль с использованием метода МЦМК из R-библиотеки mcmcRanking (v0.1.0) при условии, что все вершины имеют одинаковый вес. Модули выбираются после 1000 итераций МЦМК, чтобы гарантировать их независимость от подграфов, использованных для инициализации МЦМК.

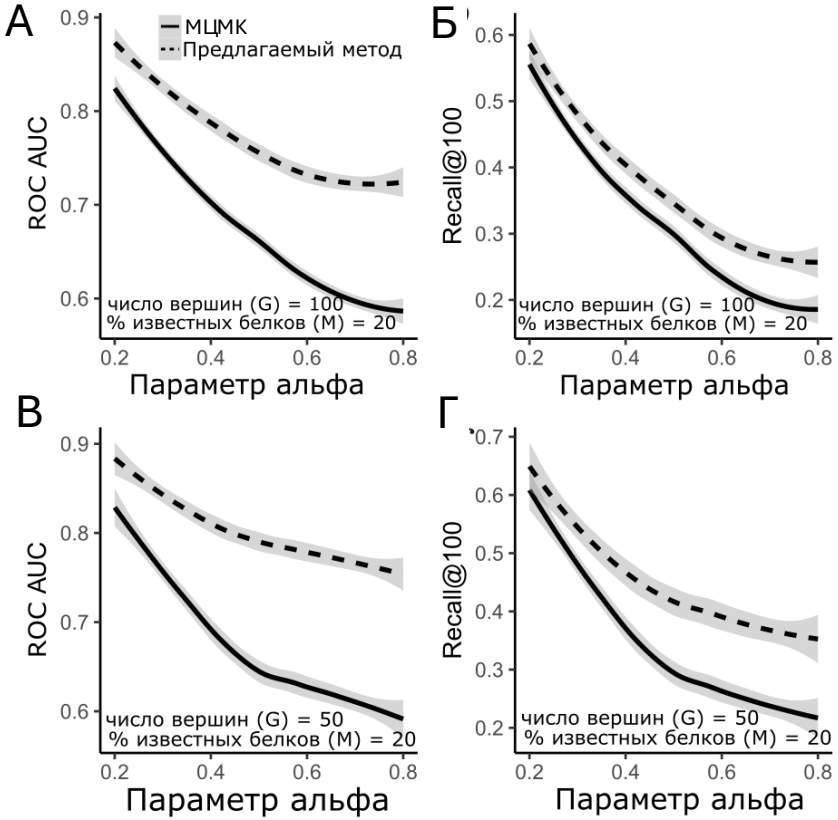
3. Для каждой вершины симулируются *p*-значения эксперимента, проверяющего принадлежность вершины активному модулю. Для вершин вне активного модуля *p*-значения выбираются из равномерного распределения, для вершин из активного модуля – из бета-распределения.

4. Так как общее число функциональных модулей равняется 100, то 50 используются для тренировки модели, а оставшиеся 50 – для тестирования.

1. Li T., Wernersson R., Hansen R.B., Horn H., et al. A scored human protein-protein interaction network to catalyze genomic interpretation // Nat Methods. 2017. V. 14. N 1. P. 61–64.

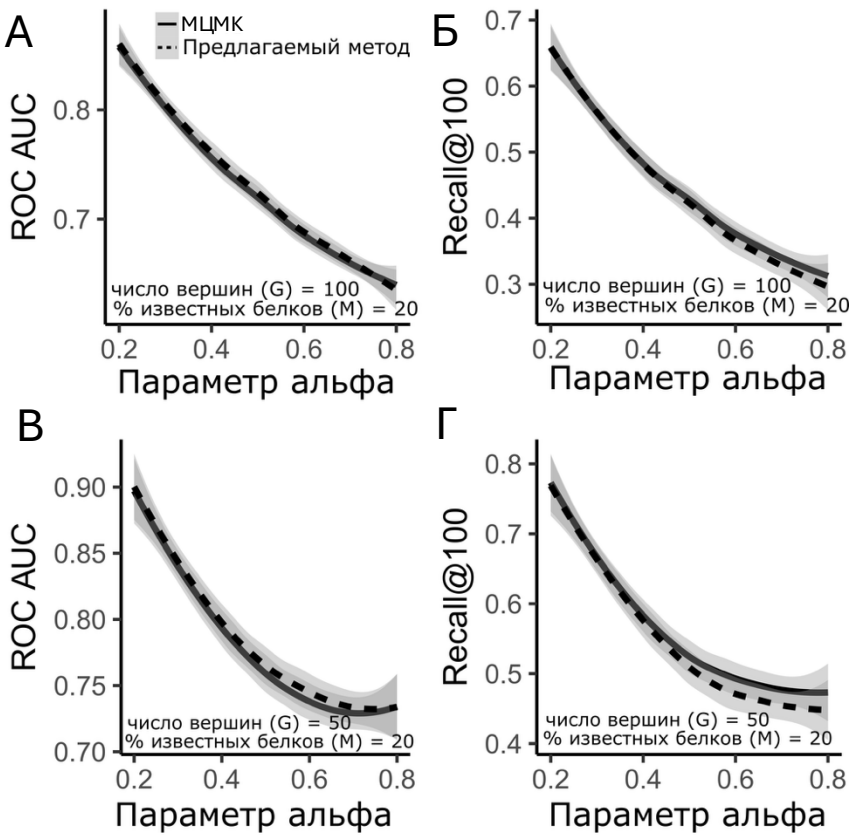
Валидация метода (2/4)

Активные модули симулированы поиском в ширину



А-Б – активные модули размера 100; В-Г – активные модули размера 50

Активные модули симулированы равномерно

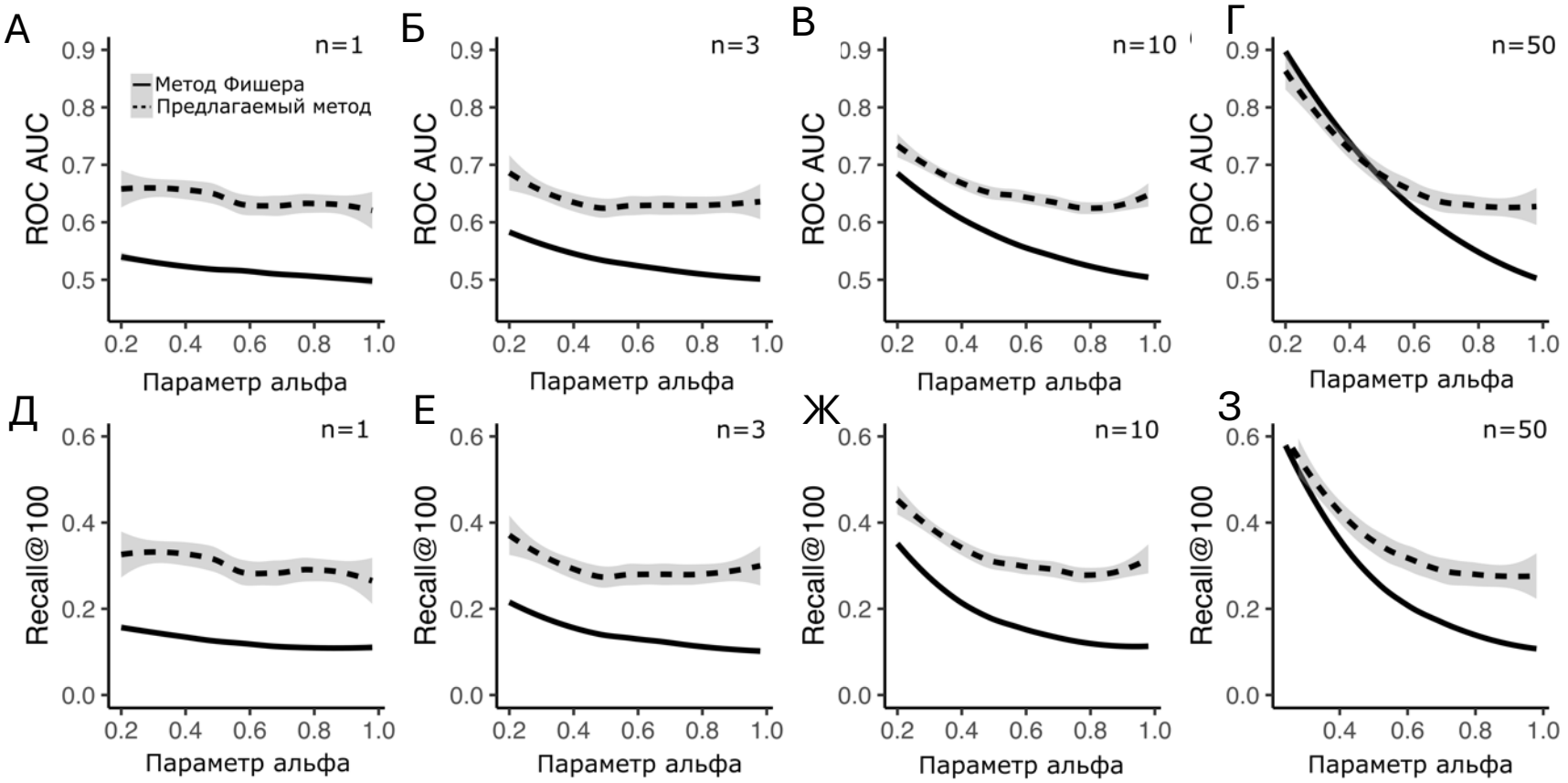


А-Б – активные модули размера 100; В-Г – активные модули размера 50

Предлагаемый метод **позволяет повысить точность графовой кластеризации по сравнению с методом только на основе МЦМК** за счет учета априорных знаний с помощью машинного обучения в случае симметричных кластеров от пяти до 20% ROC AUC (от шести до 30% Recall@100) и сохраняет точность в случае кластеров со случайной структурой. (ROC AUC – это стандартная метрика качества бинарных классификаторов в машинном обучении, где значение варьируется от нуля до одного, где один – идеальное разделение классов, а 0,5 – случайное угадывание. Recall@100 – доля правильно определенных вершин активного модуля в первой сотне вершин, ранжированных по уменьшению предсказанной с помощью модели вероятности вхождения в активный модуль).

Валидация метода (3/4)

Активные модули симулированы случайным поиском в белок-белковом графе InWebIM.

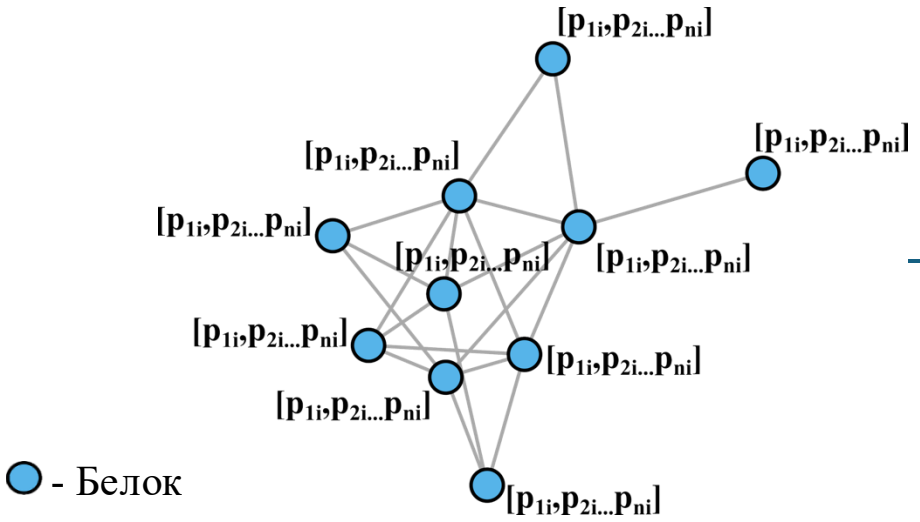


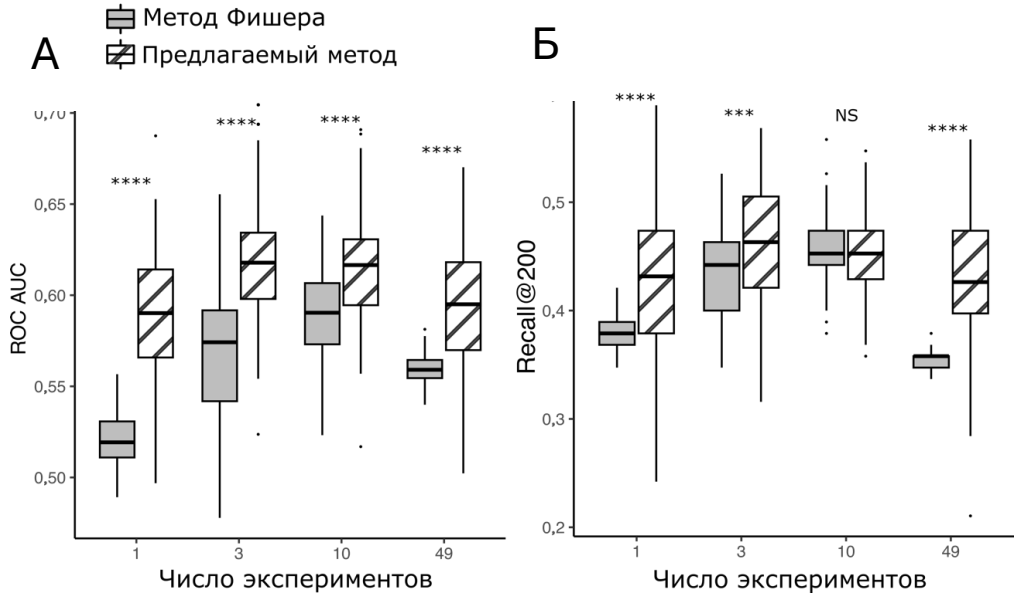
А-Г – активные модули размера 100;
Д-З – активные модули размера 50

Предложенный метод работает также или лучше, чем метод Фишера без учета графовой структуры (значения критерия ROC AUC выше от одного до 30%, значения критерия Recall@100 выше от одного до 100% на симулированных данных).

Валидация метода (4/4)

Валидация метода на реальных данных исследования шизофрении[1]





Предлагаемый метод в среднем на 13 % точнее (согласно метрике качества ROC AUC для одного эксперимента) определил причинные гены шизофрении, чем метод Фишера без учета графовой структуры.

1. Pardiñas A.F., Holmans P., Pocklington A.J. *et. al*. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection // Nat Genet. 2018. V. 50. N 3. P. 381–389.

Выводы по второму положению, выносимому на защиту

1. Метод позволяет повысить точность графовой кластеризации за счет учета априорных знаний с помощью машинного обучения в случае симметричных кластеров и сохраняет точность в случае кластеров со случайной структурой. Это обеспечивает его устойчивость по сравнению с использованием только метода на основе МЦМК или классических методов, таких как метод Фишера, не учитывающих графовые зависимости.
2. Метод позволяет интегрировать результаты нескольких экспериментов. Это позволяет значительно повысить точность определения кластеров.

Научные результаты опубликованы в:

1. **Усольцев Д.А.**, Молотков И.И., Артемов Н.Н., Сергушичев А.А., Шалыто А.А. Применение марковских цепей Монте-Карло и машинного обучения для нахождения активного модуля в биологических графах // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 962–971.
2. **Усольцев Д.А.**, Молотков И.И., Артемов Н.Н., Сергушичев А.А., Шалыто А.А. Метод определения активного модуля в биологических графах с многокомпонентными весами вершин // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 487–497.

Заключение по первому и второму положениям, выносимым на защиту

Подходы	Масштабируемость	Вероятностная оценка	Устойчивость к шуму	Возможность интеграции априорных знаний
k-means	Да	Нет	Нет	Нет
DBSCAN	Да	Нет	Да	Нет
метод Монте-Карло по схеме марковских цепей (МЦМК) (Markov Chain Monte Carlo – MCMC)	Да	Да	Да	Частично (начальное распределение вероятностей на вершинах)
Подход на основе поиска подграфа максимального веса (Maximum Weight Connected Subgraph – MWCS)	Да	Нет	Нет	Нет
Метод графовой кластеризации для совместного анализа результатов генотипирования и экспрессии генов (диссертация Александра Лободы)	Да	Нет	Нет	Нет
Предложенный метод вероятностной кластеризации матричных данных	Да	Да	Да	Да
Предложенный метод вероятностной кластеризации графовых данных	Да	Да	Да	Да

Третье положение, выносимое на защиту 1/14

Методика анализа данных в разнородных базах, основанная на сравнительном анализе и последующей графовой кластеризации его результатов, отличающаяся тем, **что с целью повышения качества кластеризации**, она включает предложенные вероятностные методы матричной и графовой кластеризации.

Третье положение, выносимое на защиту 2/14

Задача

Создание Российской медико-генетической базы данных и повышение качества анализа медико-генетической информации в ней, а также интегрирование России в международные генетические исследования для того, чтобы повысить вероятность обнаружения новых генетических рисков хронических заболеваний. Российская медико-генетическая база отсутствовала.

Описание методики

Методика включает в себя три пункта:

- разработка базы данных с разнородной информацией – медико-генетической базы данных российской популяции (МГБД-Р), состоящей из двух модулей (МГБД-Р-1 и МГБД-Р-2);
- проведение сравнительного анализа между двумя наборами данных – контрольным и тестовым, причем контрольный набор данных, находящийся в модуле МГБД-Р-1, сформирован **первым методом, выносимым на защиту**;
- аннотация результатов сравнительного анализа, хранящихся в модуле МГБД-Р-2, выполнена с использованием **второго метода, выносимого на защиту**.

Модуль МГБД-Р-1 (1/6)

Суверенная база данных с уникальной медико-генетической информацией Российской популяции

А Географическая карта сбора информации

Состав информации

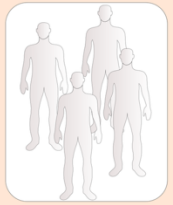
Б

В Структура первого модуля МГБД-Р

Закрытый вычислительный сервер с персональной информацией

Клиническая информация

Фенотипы	Индивид$_1$	Индивид$_2$	…
Показатель$_1$	**10**	**20**	…
Показатель$_2$	**0**	**1**	…
Показатель$_3$	**0,2**	**0,9**	…
…	…	…	…



Генетическая информация

Ген.вариант	Индивид$_1$	Индивид$_2$	…
Вариат$_1$	**0|0**	**0|0**	…
Вариат$_2$	**0|0**	**1|0**	…
Вариат$_3$	**1|0**	**0|0**	…
…	…	…	…

Матрица клинических признаков

556 признаков
4 800 индивидов

Матрица генетических признаков

11 077 392 генетических ДНК-варианта
4 800 индивидов

Матрица является основой для определения генетической структуры, проведения статистических тестов и хранится в форматах vcf[1], hail[2], plink[3]

1. Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G. T., Sherry S. T., et al. The Variant Call Format and VCFtools // Bioinformatics. 2011. V.27. N 15. P. 2156–2158.
2. Chang C.C., Chow C.C., Tellier L.C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets // GigaScience. 2015. V. 4. Article number: 7.
3. Hail Team. Hail: an open-source framework for scalable genetic data analysis // Genome Biology. 2017. V. 18. Article number: 99.

Модуль МГБД-Р-1 (2/6)

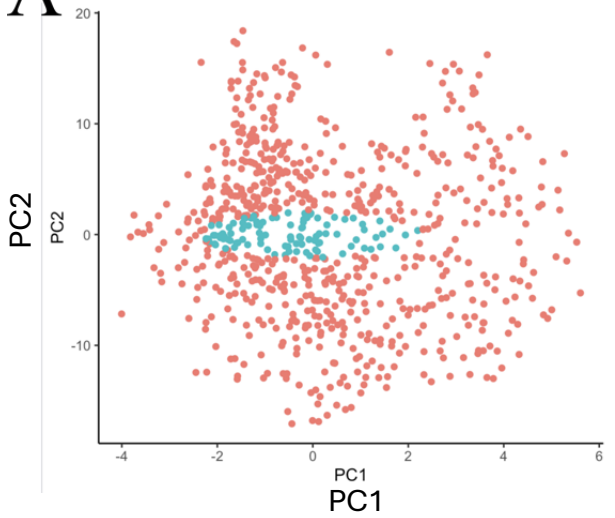
Применение первого метода, выносимого на защиту

Для построения модуля МГБД-Р-1 был применен **первый метод, выносимый на защиту**, позволяющий подобрать несмещенную контрольную группу для полногеномного ассоциативного исследования[1] (статистического теста).

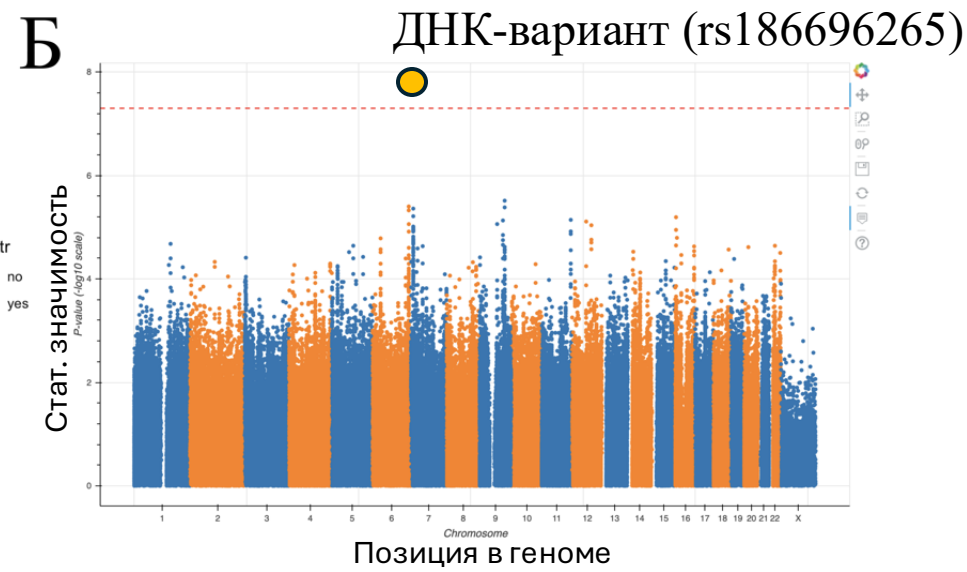
Клинико-генетическое исследование:

1. Тестовая группа – индивиды с высоким уровнем липопротеина А.
2. Контрольная группа (синяя) – индивиды со схожим липидным профилем (Рисунок А).
3. Полногеномное сравнительное исследование выявило независимое влияние ДНК-варианта (rs186696265) на уровень липопротеина А (Рисунок Б).

А



Б

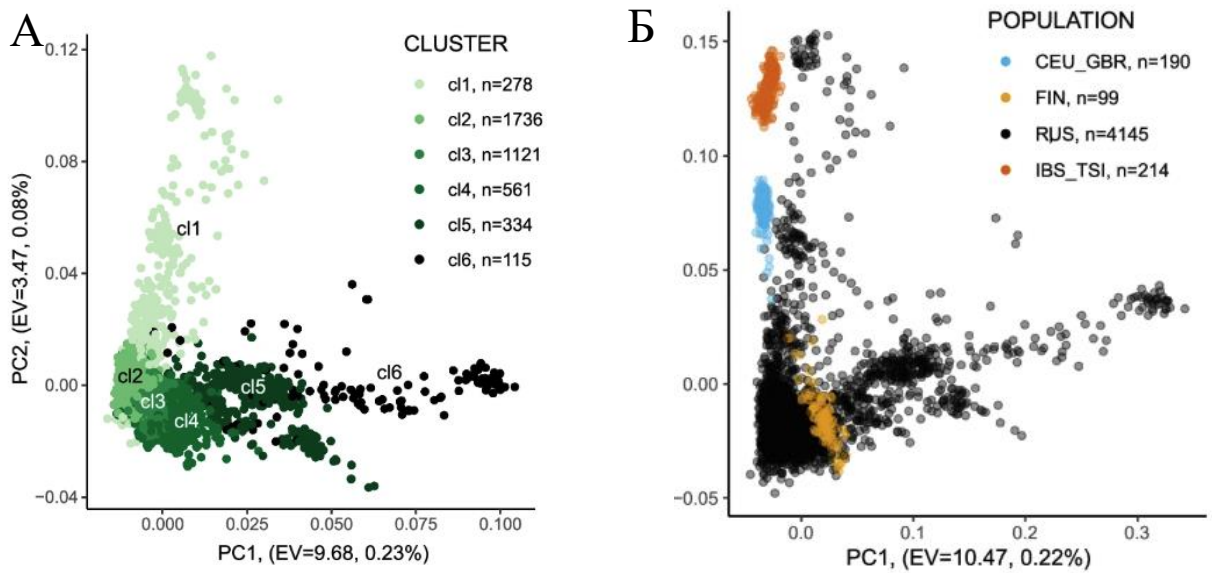


1. Uffelmann E., Huang Q.Q., Munung N.S., et al. Genome-wide association studies // Nat Rev Methods Primers. 2021. V. 1. Article number: 59.

Модуль МГБД-Р-1 (3/6)

Новизна модуля МГБД-Р-1

Впервые для научного сообщества представлена генетическая структура Российской популяции



Автором впервые была проанализирована генетическая структура Российской популяции, показано ее генетическое родство с другими популяциями. Выявленное высокое сродство Российской популяции к финской позволяет сделать **вывод о возможности использования Российской популяции в валидационных генетических исследованиях**, проводимых в крупнейшем финском биобанке – Финнгене.

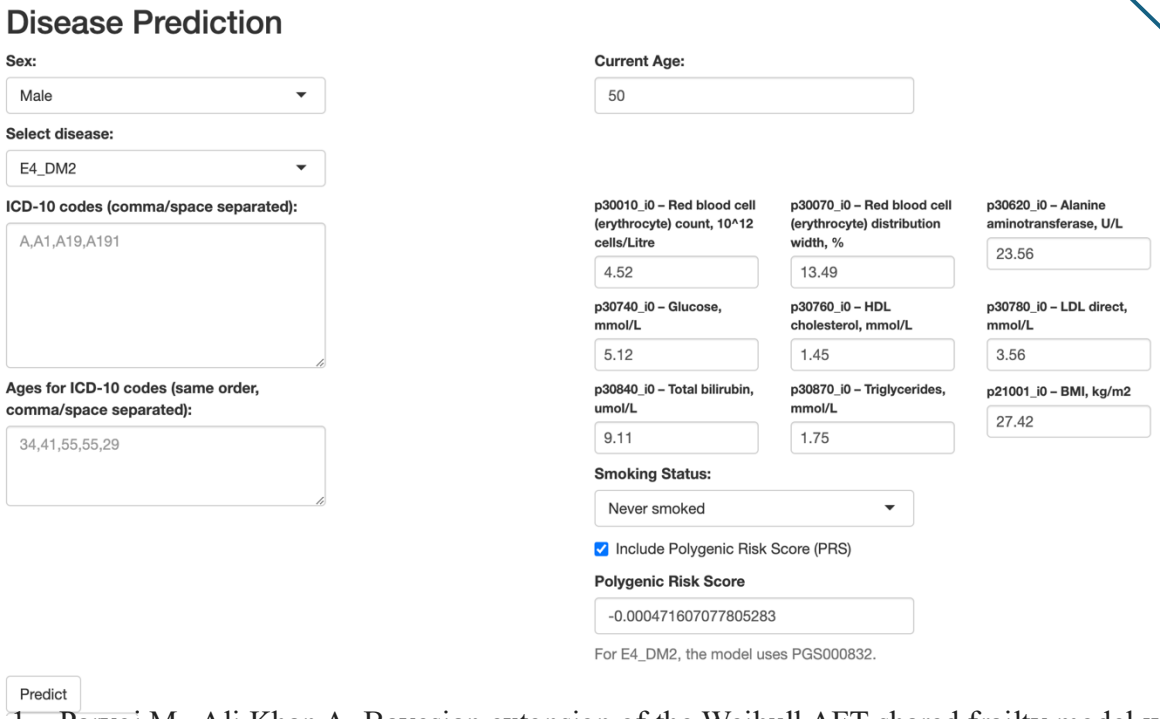
Модуль МГБД-Р-1 (4/6)

Модуль МГБД-Р-1 совместно с зарубежными МГБД лег в основу построения моделей для предсказания риска хронических заболеваний.

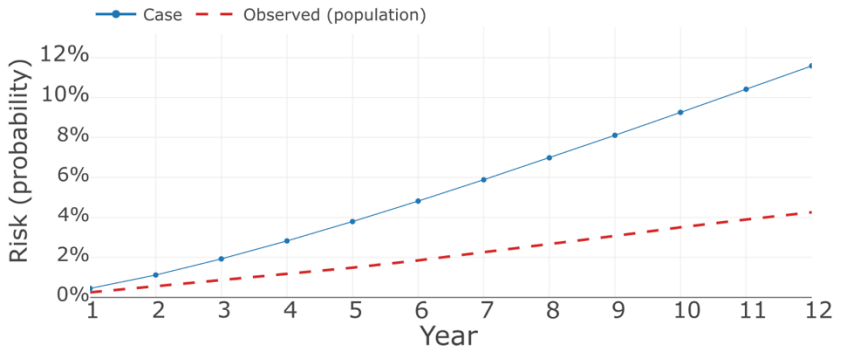
Входные клинико-генетические данные
История болезни, биохимия крови, генетические риски

Обученная AFT (Accelerated Failure Time) модель[1]

Матрицы клинической и генетической информации служат основой для построения клинико-генетических моделей предсказания времени возникновения заболеваний.



Выходные данные: риски получения заболевания (например, диабета второго типа) в следующие 12 лет



1. Parvej M., Ali Khan A. Bayesian extension of the Weibull AFT shared frailty model with generalized family of distributions for enhanced survival analysis using censored data // J Appl Stat. 2024. V. 51. N. 15. P. 3125–3153.

Модуль МГБД-Р-1 (5/6)

На основе модуля МГБД-Р-1 были построены калькуляторы рисков сердечно-сосудистых инцидентов и диабета второго типа для диагностики пациентов в ФГБУ «НМИЦ им. В. А. Алмазова»

Было получено два свидетельства о регистрации результатов интеллектуальной деятельности

Калькуляторы основаны на модели градиентного бустинга

Предикторы модели:

- общий холестерин;
- холестерин высокой плотности;
- систолическое артериальное давление;
- статус диабета второго типа;
- статус курения;
- статус приема антигипертензивных препаратов;
- статус приема статинов;
- скорость гломерулярной фильтрации;
- Chinese Visceral Adiposity Index (CVAI);
- значение NT-proBNP – мозгового натрийуретического гормона;
- частота употребления алкоголя;
- частота употребления сыра.

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024691771

Электронный калькулятор риска сердечно-сосудистых инцидентов

Правообладатель: *Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр имени В.А. Алмазова» Министерства здравоохранения Российской Федерации (RU)*

Авторы: *Ротарь Оксана Петровна (RU), Бабенко Алина Юрьевна (RU), Колесова Екатерина Павловна (RU), Бояринова Мария Анатольевна (RU), Могучая Екатерина Викторовна (RU), Ерина Анастасия Максимовна (RU), Толкунова Кристина Михайловна (RU), Кван Валерий Владимирович (RU), Усольцев Дмитрий Андреевич (RU), Конради Александра Олеговна (RU), Шляхто Евгений Владимирович (RU)*

Заявка № 2024689689

Дата поступления 02 декабря 2024 г.

Дата государственной регистрации в Реестре программ для ЭВМ *24 декабря 2024 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2024690421

Электронный калькулятор риска диабета второго типа

Правообладатель: *Федеральное государственное бюджетное учреждение «Национальный медицинский исследовательский центр имени В.А. Алмазова» Министерства здравоохранения Российской Федерации (RU)*

Авторы: *Ротарь Оксана Петровна (RU), Бабенко Алина Юрьевна (RU), Колесова Екатерина Павловна (RU), Бояринова Мария Анатольевна (RU), Могучая Екатерина Викторовна (RU), Ерина Анастасия Максимовна (RU), Толкунова Кристина Михайловна (RU), Кван Валерий Владимирович (RU), Усольцев Дмитрий Андреевич (RU), Конради Александра Олеговна (RU), Шляхто Евгений Владимирович (RU)*

Заявка № 2024689622

Дата поступления 02 декабря 2024 г.

Дата государственной регистрации в Реестре программ для ЭВМ *16 декабря 2024 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Ю.С. Зубов

Предикторы модели:

- глюкоза;
- LAP (индекс накопления липидов);
- Cw2 (ожирение в соответствии с объемом талии ATPIII);
- Adn (Адипонектин);
- IAGT (Статус применения препаратов для снижения арт. давл.);
- ген FTO;
- TAG (те, кто принимает антигипертензивную терапию или арт. давл. выше 140/90 мм рт ст);
- статус нарушения сердечного ритма;
- COM (кто имеет арт. гиперт. и принимает терапию).

Модуль МГБД-Р-1 (6/6)

Модуль МГБД-Р-1 лег в основу клинико-генетических исследований в России

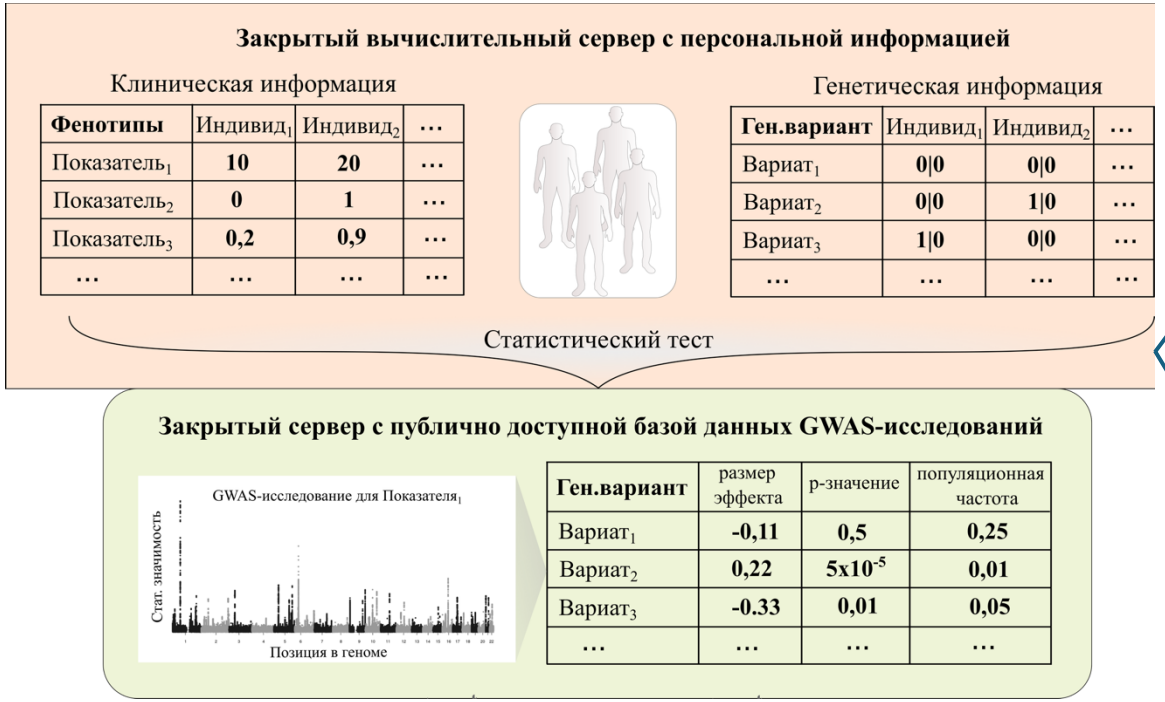
Примеры клинико-генетических исследований с участием автора:

1. Описан фенотип сосудистой жёсткости в Российской популяции, отсутствующий в крупных мировых МГБД.
2. Найдены новые патогенные ДНК-варианты для целиакии.
3. Написаны медицинские рекомендации для использования зарубежной клинической шкалы сердечно-сосудистых заболеваний в Российской популяции.
4. Найдены новые патогенные ДНК-варианты для аортального стеноза.

По всем этим исследованиям имеются публикации с участием автора.

Модуль МГБД-Р-2 (1/4)

Модуль МГБД-Р-1 суверенный



Модуль МГБД-Р-2 Структура модуля МГБД-Р-2

Для каждого из 556 клинических признаков, хранящихся в модуле МГБД-Р-1 был проведен полногеномный поиск ассоциаций (GWAS)[1]:

Для непрерывных признаков (Y) построена линейная модель для каждого из ~8млн ДНК-вариантов (G), с учетом ковариат (C_i): пол, возраст, 10 первых принципиальных компонент генетической структуры

$$Y = \beta_0 + \beta_1 G + \beta_2 C_1 + \beta_3 C_2 + \cdots + \beta_p C_p + \varepsilon$$

Для бинарных признаков (Y) построена логистическая модель для каждого из ~8млн ДНК-вариантов (G), с учетом ковариат (C_i): пол, возраст, 10 первых принципиальных компонент генетической структуры

$$P(Y = 1|G, C) = \frac{1}{1 - e^{-(\beta_0 + \beta_1 G + \beta_2 C_1 + \beta_3 C_2 + \cdots + \beta_p C_p)}}$$

Из каждой модели были сохранены:

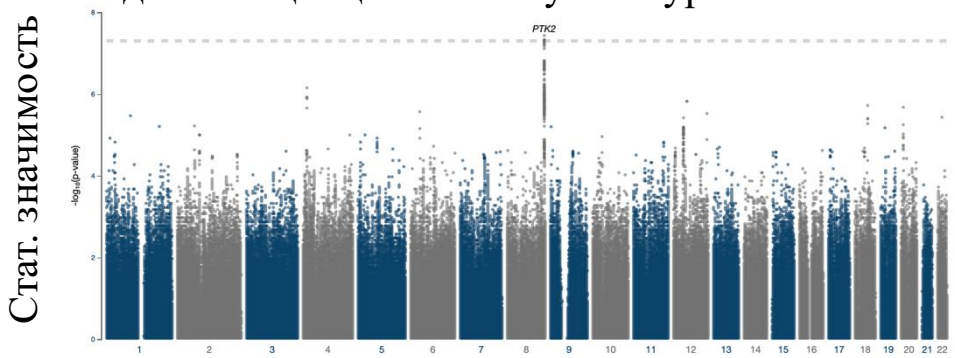
- размер эффекта,
- статистическая значимость (p-значение)
- z-статистика

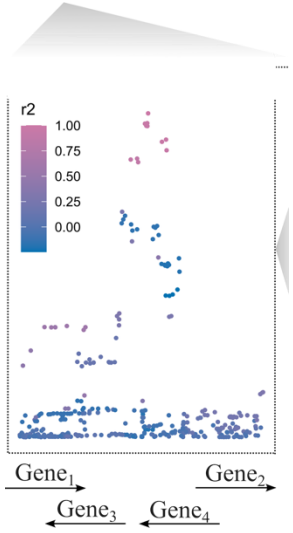
1. Uffelmann E., Huang Q.Q., Munung N.S., et al. Genome-wide association studies // Nat Rev Methods Primers. 2021. V. 1. Article number: 59.

Модуль МГБД-Р-2 (2/4)

Для модуля МГБД-Р-2 был применен **второй метод, выносимый на защиту**

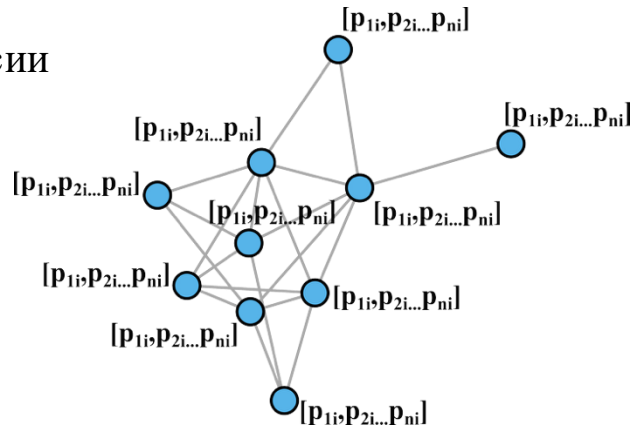
Статистическая значимость ДНК-вариантов
для ассоциации со статусом курения

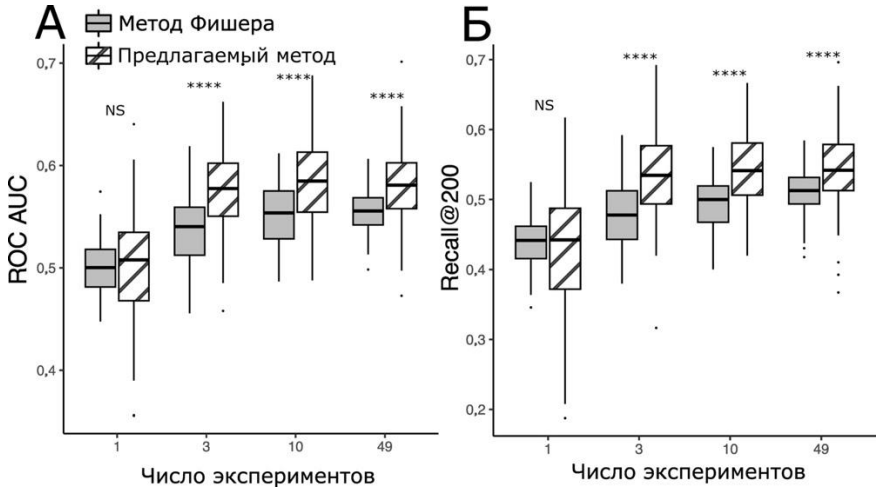




Данные экспрессии

InWebIM граф





Предлагаемый метод в среднем на 7 % точнее (согласно метрике качества ROC AUC для трех экспериментов) определил причинные гены пристрастия к курению, чем метод Фишера без учета графовой структуры.

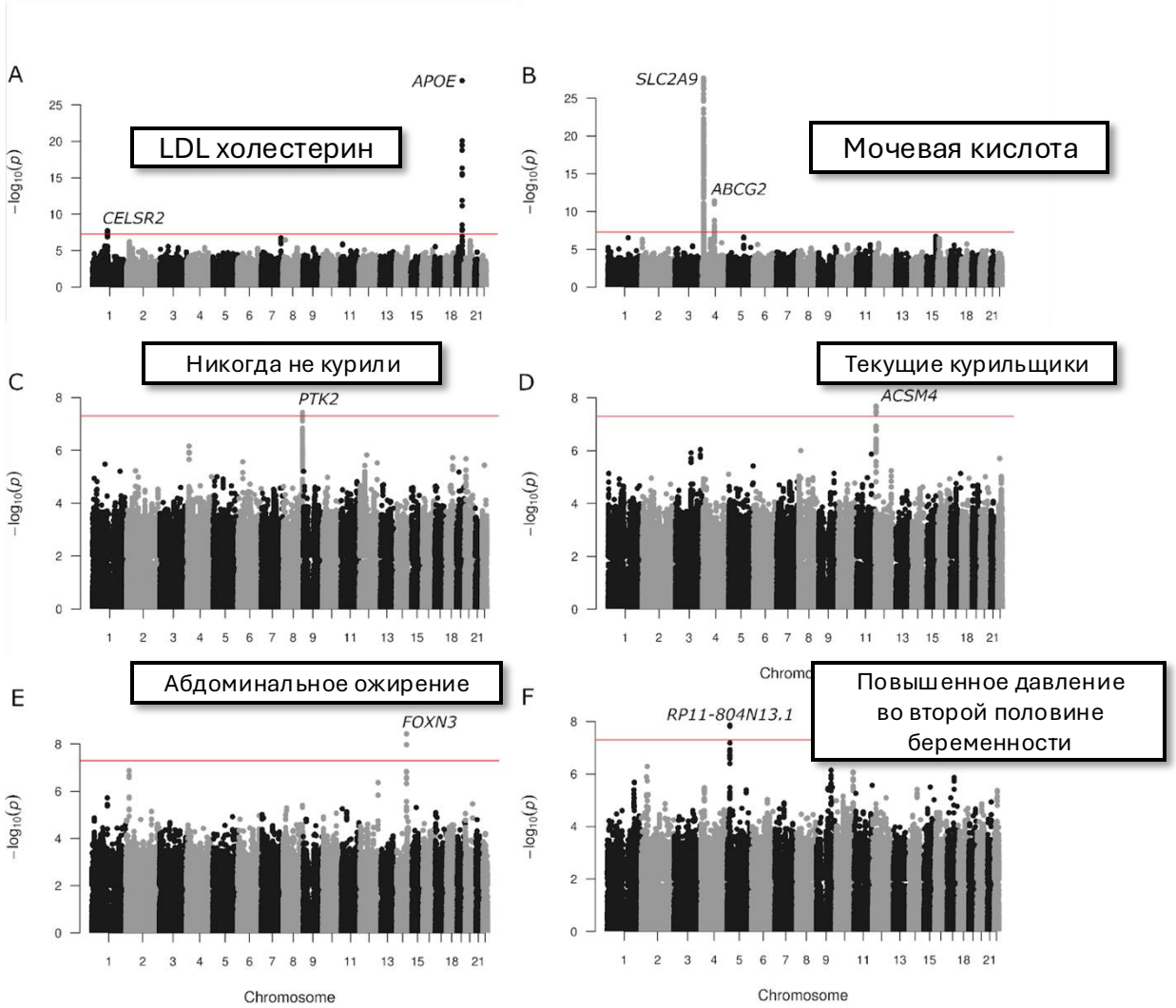
Модуль МГБД-Р-2 (3/4)

Новизна модуля МГБД-Р-2

Репликация Биобанка Великобритании

Значимые ассоциации в модуле МГБД-Р-2 (могут быть реплицированы в Биобанке Великобритании)

Специфичные значимые ассоциации в модуле МГБД-Р-2



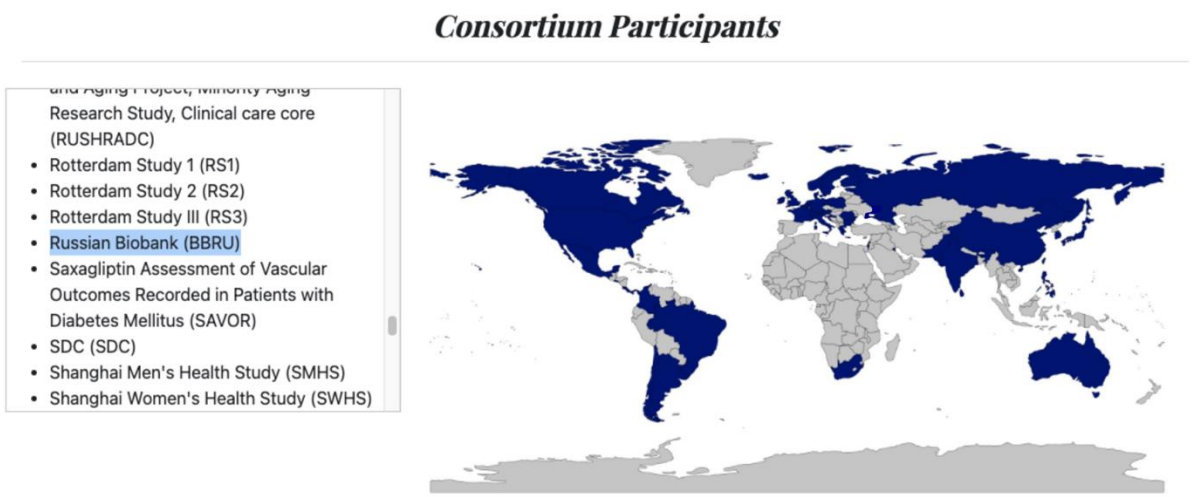
Полногеномные клинико-генетические модели ранее не строились в Российской популяции. Автором впервые были построены такие модели. Модели позволили выявить новые ДНК-варианты для рисков заболеваний.

Модуль МГБД-Р-2 (4/4)

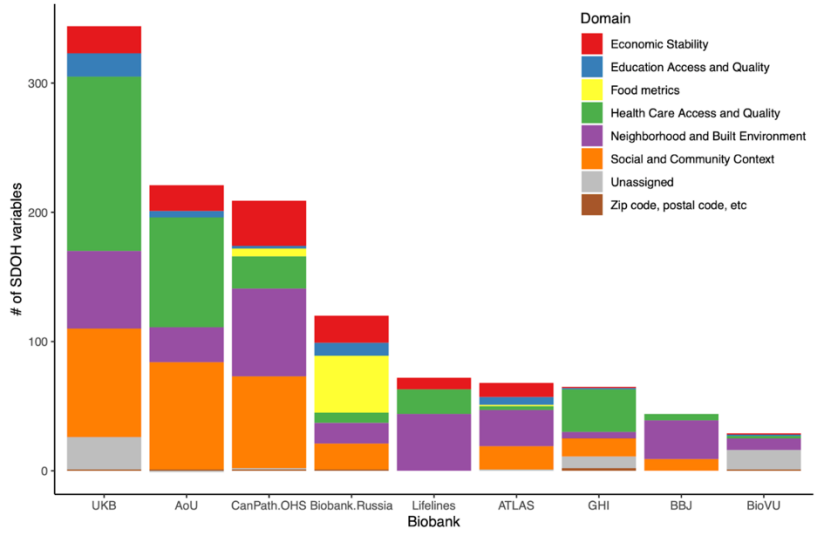
Модуль МГБД-Р-2 доступен онлайн по адресу https://biobankrus.almazovcentre.ru/.

Модуль МГБД-Р-2 включен в крупнейшие мировые консорциумы по изучению генетических заболеваний.

The Global Lipids Genetics Consortium:
https://www.lipidgenetics.org/



Global Biobank Meta-analysis Initiative:
без публичного доступа



Выводы по третьему положению, выносимому на защиту

1. Разработана методика анализа данных в разнородных базах, которая включает создание такой базы на примере медико-генетической базы Российской популяции (МГБД-Р), состоящей из двух модулей, а также аннотацию данных модуля МГБД-Р-2. Модуль МГБД-Р-2 доступен по адресу https://biobank.almazovcentre.ru/.
2. Разработанная методика реализована в виде программного обеспечения на языке R: https://github.com/11Dmitriy11/mgc. Пакет позволяет проводить отбор контрольной группы из модуля МГБД-Р-1 с использованием предлагаемого вероятностного метода матричной кластеризации и интерпретацию (аннотацию) результатов сравнительного анализа с использованием предлагаемого вероятностного метода графовой кластеризации.
3. Применение методики повышает качество аннотирования медико-генетической информации в МГБД-Р.
4. Россия интегрирована в международные генетические исследования хронических заболеваний.

Научные результаты опубликованы в:

1. **Usoltsev D.**, Kolosov N., Rotar O., Sergushichev A. O., Artomov N.N. *et al.* Complex trait susceptibilities and population diversity in a sample of 4,145 Russians // Nature Communications. 2024. V. 15. Article number: 6212.
2. Zlotina A., Barashkova S., Zhuk S., Skitchenko R., **Usoltsev D.**, Sokolnikova P., Artomov M., Alekseenko S., Simanova T., Goloborodko M., Berleva O., Kostareva A. Characterization of pathogenic genetic variants in Russian patients with primary ciliary dyskinesia using gene panel sequencing and transcript analysis // Orphanet Journal of Rare Diseases. 2024. V. 19. Article number: 310.

Заключение 1/5

1. Разработан вероятностный метод кластеризации для подбора статистически сопоставимого контрольного набора данных, основанного на априорном распределении признаков и сингулярном разложении матриц. Метод был применен автором к модулю МГБД-Р-1 с целью формирования контрольной группы для проведения полногеномного ассоциативного исследования.
2. Разработан вероятностный метод кластеризации в графовых структурах, учитывающий топологию сети и интеграцию нескольких источников экспериментальных данных через многокомпонентные веса. Метод реализует комбинированный подход на основе МЦМК и машинного обучения для учета априорной информации. Метод был применен автором к модулю МГБД-Р-2 с целью повышения качества кластеризации графовой информации.
3. Разработана методика для аннотации разнородной информации в базах данных. Методика использует предложенные вероятностный метод кластеризации для подбора несмещенного набора данных и вероятностный метод графовой кластеризации, учитывающий априорные знания для повышения качества аннотации данных, которая заключается в поиске подграфа с целевыми признаками. В среднем качество нахождения такого подграфа повышено на 13 % относительно метода Фишера, который не использует графовую кластеризацию.

Опубликованные работы

1. **Усольцев Д.А.**, Молотков И.И., Артемов Н.Н., Сергушичев А.А., Шалыто А.А. Метод определения активного модуля в биологических графах с многокомпонентными весами вершин // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 3. С. 487–497.
2. **Усольцев Д.А.**, Молотков И.И., Артемов Н.Н., Сергушичев А.А., Шалыто А.А. Применение марковских цепей Монте-Карло и машинного обучения для нахождения активного модуля в биологических графах // Научно-технический вестник информационных технологий, механики и оптики. 2024. Т. 24, № 6. С. 962–971.
3. **Усольцев Д.А.** Вероятностный метод матричной кластеризации с априорным распределением признаков для формирования несмещенной контрольной группы // Научно-технический вестник информационных технологий, механики и оптики. 2025. Т. 25, № 5. С. 999–1001.
4. **Usoltsev D**., Tolkunova K., Moguchaia E. *et al.* Transgenerational and intergenerational effects of early childhood famine exposure in the cohort of offspring of Leningrad Siege survivors // Scientific Reports. 2023. V. 13. Article number: 11188.
5. **Usoltsev D.**, Kolosov N., Rotar O., Sergushichev A. O., Artomov N.N. et al. Complex trait susceptibilities and population diversity in a sample of 4,145 Russians // Nature Communications. 2024. V. 15. Article number: 6212.
6. Zlotina A., Barashkova S., Zhuk S., Skitchenko R., **Usoltsev D.**, Sokolnikova P., Artomov M., Alekseenko S., Simanova T., Goloborodko M., Berleva O., Kostareva A. Characterization of pathogenic genetic variants in Russian patients with primary ciliary dyskinesia using gene panel sequencing and transcript analysis // Orphanet Journal of Rare Diseases. 2024. Т. 19, № 310. Open access.

Заключение 3/5

Акты внедрения

Медико-генетическая база данных российской популяции (https://biobank.almazovcentre.ru/), а также предложенные вероятностные методы матричной и графовой кластеризации применяются в качестве примеров анализа больших биомедицинских данных при подготовке магистров по специальности «Bioinformatics and systems biology» на курсе «Introduction to Linux».

ИТМО

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО» (Университет ИТМО)

Кронверкский пр-т, д. 49, лит. А, Санкт-Петербург, Россия, 197101
Тел.: (812) 480-00-00
od@itmo.ru | itmo.ru

№_____

АКТ
об использовании в учебном процессе результатов диссертационного исследования Усольцева Дмитрия Андреевича на тему «Методы вероятностной кластеризации для интерпретации матричной и графовой информации в разнородных базах данных»

Настоящий акт подтверждает, что результаты научно-исследовательской работы, выполненной Усольцевым Дмитрием Андреевичем в рамках диссертационного исследования на тему «Методы вероятностной кластеризации для интерпретации матричной и графовой информации в разнородных базах данных», используются в учебном процессе Университета ИТМО на магистерской программе «Биоинформатика и системная биология» института прикладных компьютерных наук.

Разработанная Усольцевым Д. А. медико-генетическая база данных российской популяции (https://biobank.almazovcentre.ru/), а также предложенные вероятностные методы матричной и графовой кластеризации применяются в качестве примеров анализа больших биомедицинских данных при подготовке магистров. Использование базы данных и реализованных алгоритмов позволяет решать практические задачи по интеграции и аннотированию медико-генетической информации с использованием инструментов операционной системы Linux и специализированных пакетов анализа данных.

Указанные результаты включены в:
курс «Введение в Linux» в 2022–2024 годах.

29.08.2025

Жук Анна Сергеевна
к.б.н., доцент ИПКН,
руководитель образовательной программы «Биоинформатика и системная биология»,
Университет ИТМО

МЕНЕДЖЕР ОВС

Разработанный автором ресурс (https://biobank.almazovcentre.ru/), используется при чтении лекции по курсу «Биоинформатические методы анализа геномных данных» в научном центре мирового уровня «Центр персонализированной медицины» Федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр имени В.А. Алмазова» Министерства здравоохранения Российской Федерации.

МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ «НАЦИОНАЛЬНЫЙ МЕДИЦИНСКИЙ ИССЛЕДОВАТЕЛЬСКИЙ ЦЕНТР ИМЕНИ В. А. АЛМАЗОВА»

197341, Россия, Санкт-Петербург, ул. Аккуратова, 2
Тел. +7 (812) 705-37-70, Факс +7 (812) 705-37-01
e-mail: 9mail@almazovcentre.ru
ОГРН 1037800031011 ИНН 7802034429 КПП 781401001

№ _____
на № _____ от _____

«УТВЕРЖДАЮ»
Директор института медицинского образования
ФГБУ НМИЦ им. В. А. Алмазова
Минздрава России
к.м.н., доцент
Е.В. Пармон
«27» 09 2025 г.

АКТ О ВНЕДРЕНИИ РЕЗУЛЬТАТОВ В ПРАКТИКУ УЧЕБНОЙ РАБОТЫ

1. **Наименование предложения:** Медико-генетическая база данных российской популяции (Биобанк России, https://biobank.almazovcentre.ru/).
2. **Краткая аннотация:** Разработанная база данных предназначена для систематизированного хранения и анализа клинико-генетической информации жителей России. Ресурс обеспечивает возможность проведения ассоциативных исследований (GWAS), интеграции с международными консорциумами, а также создания статистически сопоставимых контрольных выборок с учётом генетического разнообразия российской популяции. База состоит из двух модулей: (1) хранения персональных клинико-генетических данных и (2) агрегированного статистического блока, поддерживающего распределённый обмен информацией.
3. **Эффект от внедрения:** Использование базы данных позволяет формировать навыки работы с большими биомедицинскими данными, обучать методам обработки и анализа клинико-генетической информации, осваивать современные подходы к интеграции и интерпретации данных. Внедрение базы в образовательный процесс позволит сформировать у обучающихся компетенции в области биоинформатики, анализа данных и персонализированной медицины.
4. **Место и время использования предложения:** Предложение внедрено в рамках программы Дополнительного профессионального образования, программа повышения квалификации «Современные представления о патогенезисе сосудистого воспаления и подходах к его терапии» продолжительностью 144 академических часа", Научный центр мирового уровня «Центр персонализированной медицины» Федерального государственного бюджетного учреждения «Национальный медицинский исследовательский центр имени В.А. Алмазова» Министерства здравоохранения Российской Федерации.
5. **Форма внедрения:** предложение внедрено в форме материалов семинарских и лекционных занятий
6. **Название темы диссертационной работы и шифр специальности:** «Методы вероятностной кластеризации для интерпретации матричной и графовой информации в разнородных базах данных», специальность 2.3.8 – «Информатика и информационные процессы».

Автор: Усольцев Дмитрий Андреевич, аспирант Университета ИТМО.

Руководитель подразделения – базы внедрения:
Заведующий департаментом непрерывного профессионального развития ИМО ФГБУ «НМИЦ им В. А. Алмазова» Минздрава России — Голубева И.С.

Заведующий центром развития образовательной среды ИМО ФГБУ «НМИЦ им В. А. Алмазова» Минздрава России — Петрова Н.Н.

Заместитель директора ИМО по учебной и методической работе ФГБУ «НМИЦ им В. А. Алмазова» Минздрава России — Кухарчик Г.А.

м.п.

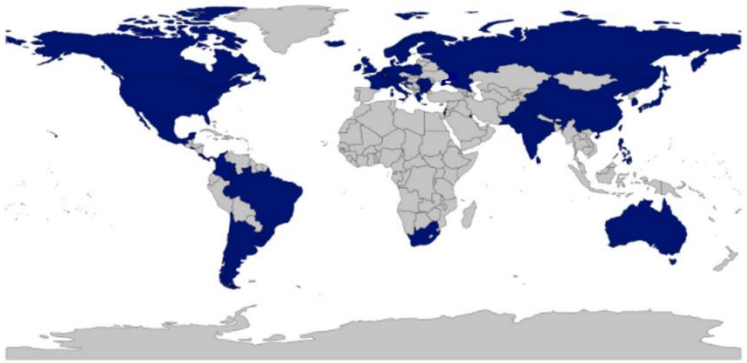
Внедрение МГБД-Р в международные исследования

Россия стала участником крупнейших международных генетических исследований

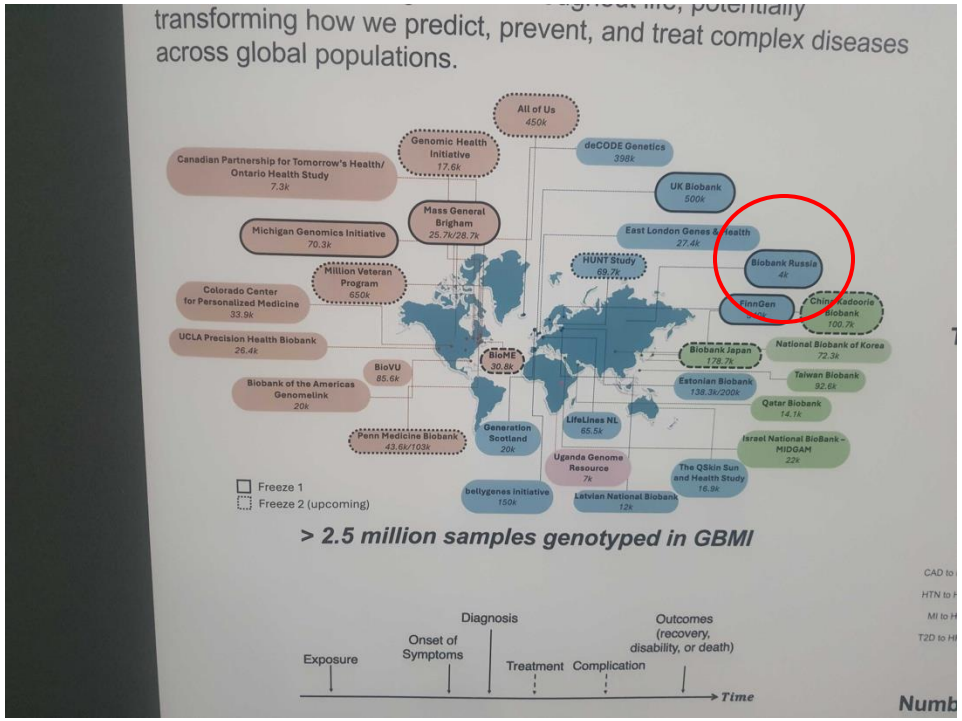
The Global Lipids Genetics Consortium:
https://www.lipidgenetics.org/

Consortium Participants

Research Study, Clinical care core (RUSHRADC)
- Rotterdam Study 1 (RS1)
- Rotterdam Study 2 (RS2)
- Rotterdam Study III (RS3)
- Russian Biobank (BBRU)
- Saxagliptin Assessment of Vascular Outcomes Recorded in Patients with Diabetes Mellitus (SAVOR)
- SDC (SDC)
- Shanghai Men's Health Study (SMHS)
- Shanghai Women's Health Study (SWHS)



Global Biobank Meta-analysis Initiative:
без публичного доступа



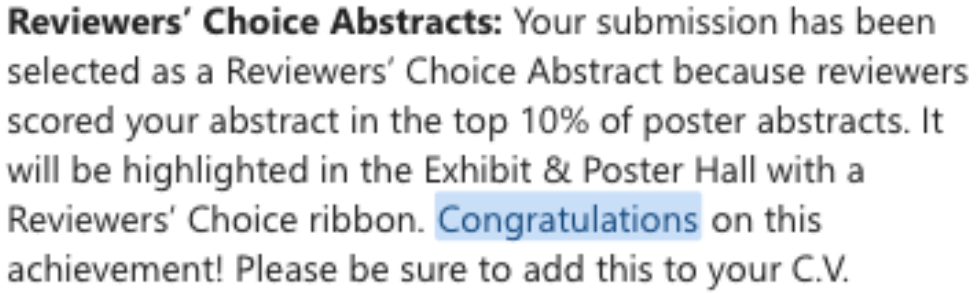
Апробация результатов

Конференции

- XIV Конгресс молодых ученых, 2025, Университет ИТМО, Санкт-Петербург, Россия // Биобанк России – онлайн-платформа для хранения и анализа результатов GWAS-исследований полигенных заболеваний, а также разработка фенотипических предсказательных моделей на базе когорты Биобанка России.
- American Society of Human Genetics, 2024, Denver, CO, USA // A meta-predictor for causal gene identification in GWAS overcomes limitations of existing computational approaches.
- Cold Spring Harbor Laboratory Conference, 2024, Laurel Hollow, NY, USA // Understanding complex trait susceptibilities and ethnic diversity in a sample of 4,145 Russians through analysis of clinical and genetic data.
- American Society of Human Genetics, 2023, Washington, D.C., USA // Understanding complex trait susceptibilities and ethnical diversity in a sample of 4,145 Russians through analysis of clinical and genetic data.

Награда

- **Reviewers' Choice Abstract** (10 % лучших работ на American Society of Human Genetics, 2024, Denver, CO, USA). Постерный доклад: A meta-predictor for causal gene identification in GWAS overcomes limitations of existing computational approaches


Reviewers' Choice Abstracts: Your submission has been selected as a Reviewers' Choice Abstract because reviewers scored your abstract in the top 10% of poster abstracts. It will be highlighted in the Exhibit & Poster Hall with a Reviewers' Choice ribbon. Congratulations on this achievement! Please be sure to add this to your C.V.

СПАСИБО ЗА ВНИМАНИЕ!

Усольцев Дмитрий Андреевич

dusoltsev.27@gmail.com