На правах рукописи



# Манолов Александр Иванович

Биоинформатический анализ изменчивости генного состава прокариот, в том числе в ассоциации с патогенностью

Специальность 1.5.8 ––

«Математическая биология, биоинформатика»

Автореферат

диссертации на соискание учёной степени кандидата биологических наук

Москва — 2021

Работа выполнена в Федеральном государственном бюджетном учреждении ”Федеральный научно­клинический центр физико­химической медицины Феде­ рального медико­биологического агентства”.

Научный руководитель: доктор биологических наук, член­корреспондент

РАН

## Ильина Елена Николаевна

Официальные оппоненты: **Фамилия Имя Отчество,**

доктор физико­математических наук, профессор, Не очень длинное название для места работы,

старший научный сотрудник

**Фамилия Имя Отчество,**

кандидат физико­математических наук,

Основное место работы c длинным длинным длин­ ным длинным названием,

старший научный сотрудник

Ведущая организация: Федеральное государственное бюджетное образо­

вательное учреждение высшего профессиональ­ ного образования с длинным длинным длинным длинным названием

Защита состоится DD mmmmmmmm YYYY г. в XX часов на заседании диссер­ тационного совета Д 123.456.78 при Название учреждения по адресу: Адрес.

С диссертацией можно ознакомиться в библиотеке Название библиотеки.

Отзывы на автореферат в двух экземплярах, заверенные печатью учреждения, просьба направлять по адресу: Адрес, ученому секретарю диссертационного со­ вета Д 123.456.78.

Автореферат разослан DD mmmmmmmm2021 года. Телефон для справок: +7 (0000) 00­00­00.

Ученый секретарь

диссертационного совета Д 123.456.78,

д­р физ.­мат. наук Фамилия Имя Отчество

# Общая характеристика работы

**Актуальность темы.** Геном прокариот представляет собой сложно органи­ зованную структуру. Помимо кодирующих и регуляторных областей, в нем имеется ряд элементов, необходимых для взаимодействия ДНК с молекуляр­ ными комплексами, осуществляющими процессы транскрипции, репликации и репарации. Пространственная укладка генетического материала в клетке не слу­ чайна и выполняет ряд регуляторных функций. Подобные наблюдения меняют представление о геноме, как о простом хранилище последовательностей генов расположенных в случайном порядке, и позволяют говорить об архитектуре генома — закономерностях, которые необходимы для успешного функциони­ рования живой клетки.

К настоящему времени известен ряд элементов геномной организации. Гены, продукты которых необходимы клетке в больших количествах, располо­ жены рядом с сайтом начала репликации, поскольку в быстро делящихся клетках такое расположение позволяет повысить уровень их экспрессии за счет уве­ личения копийности матричной ДНК. Пространственная укладка ДНК может сближать гены, расположенные в разных областях линейной последовательно­ сти, что оказывается полезно для генов, кодирующих регулятор и его мишени. Экспериментально было установлено, что действие глобальных регуляторов, таких как гистоноподобный белок H­NS, зависит от местоположения генов ми­ шеней. Склонность к транскрипции (уровень экспрессии генов, не зависящий от их последовательности) значительно меняется в зависимости от положения гена в хромосоме. Взаимодействие РНК­полимераз, возникающее за счет изменения уровня суперскрученности ДНК, может играть роль в регуляции транскрипции соседних генов.

Геномные перестройки и горизонтальный перенос генов могут приводить к изменению оптимального расположения генов и других элементов генома, что может приводить к снижению жизнеспособности организма. Известно, что изменения в геномах преимущественно локализуются в отдельных местах — ”горячих” точках. Возможно, эти участки свободны от ”архитектурных” ограни­ чений, и таким образом более толерантны к изменениям. Возможно, эти участки имеют некоторые признаки, способствующие более высокой частоте происходя­ щих изменений. Нельзя исключить возможность, что “горячие” точки возникли в результате генетического дрейфа, а их расположение случайно и не обладает функциональным значением. Какие из этих вариантов, и в какой степени, реа­ лизуются в действительности к настоящему моменту неизвестно: локализация ”тихих” консервативных участков и ”горячих” высокоизменчивых областей не имеет общепринятых объяснений. Для проведения исследований в данной об­ ласти необходим инструмент, позволяющий находить и анализировать области

генома с повышенной и пониженной изменчивостью. Разработка и применение подобного инструментария и стала основной темой данной работы.

**Целью** данной работы является разработка программного конвейера для выявления высокоизменчивых областей геномов прокариот и применение его для анализа изменчивости в локусах генома *Escherichia coli*, ассоциированных с болезнью Крона.

Для достижения поставленной цели необходимо было решить следующие

**задачи**:

1. Разработать алгоритм оценки уровня изменчивости геномов, основан­ ный на их графовом представлении.
2. Сравнить профили изменчивости геномов, принадлежащих различным родам, видам и подвидовым структурам прокариот.
3. Оценить вклад различных факторов геномной организации в уровень изменчивости генома.
4. Разработать алгоритм визуализации подграфов, соответствующих от­ дельным локусам генома.
5. Разработать алгоритм поиска и выявить в геноме *E. coli* опероны, ко­ торые значимо чаще встречаются в изолятах от пациентов с болезнью Крона, чем в изолятах от здоровых людей.

**Научная новизна:** Предложенный в нашей работе подход, насколько нам известно, является первым предложенным и реализованным методом для коли­ чественной оценки изменчивости генома.

Насколько нам известно, мы впервые провели сравнительный анализ рас­ положения областей повышенной изменчивости. Мы обнаружили, что некото­ рые высокоизменчивые локусы генома могут сохранять свое расположение у представителей близкородственных видов.

## Практическая значимость

Изменчивость генома — важный фактор в возникновении патогенных штаммов бактерий и приобретении устойчивости к антибиотикам. Знание зако­ номерностей подобных изменений важно для разработки оптимальных методов контроля над появлением штаммов бактерий, угрожающих жизни и здоровью людей. Возможно, полученные знания о закономерностях изменчивости и кон­ сервативности различных областей генома окажутся полезными при создании новых последовательностей геномов в области синтетической биологии.

## Основные положения, выносимые на защиту:

1. Графовое представление геномов позволяет эффективно проводить по­ иск областей генома с повышенной изменчивостью.
2. Визуализация в виде графа позволяет компактно представлять сравне­ ние больших выборок геномов (порядка сотен и тысяч геномов).
3. Геномы представителей различных филогрупп и филогенетически близ­ ких видов имеют консервативно расположенные области повышенной изменчивости (расположенные в местах генома с одинаковым генным контекстом).
4. В геномах изолятов *E. coli* от пациентов с болезнью Крона значимо ча­ ще выявляются опероны захвата сорбозы, захвата гемина, утилизации глиоксилата, утилизации пропандиола, синтеза и экспорта капсульных полисахаридов.

**Достоверность** предложенного метода обосновывается результатами компьютерного моделирования. Результаты находятся в соответствии с ре­ зультатами, полученными другими авторами. Основные результаты работы были доложены на конференциях: ”Итоговая научно­практическая конференция ФГБУ ФНКЦ ФХМ ФМБА России” (18­19 декабря 2019 года, Москва), “ПО­ СТГЕНОМ 2018” (29 октября ­ 2 ноября 2018 года, Казань),”Биотехнология: состояние и перспективы развития” (20–22 февраля 2017 года, Москва, ), ”Вы­ сокопроизводительное секвенирование в геномике” (Новосибирск, 18–23 июня 2017 года), ”4th World Congress on Targeting Microbiota” (17­19 октября 2016, Париж).

**Личный вклад.** Автором были предложены подходы графового представ­ ления набора генов в геномах и оценки геномной вариабельности на основе выбора подграфа. Написан код на языках R, perl и Snakemake для графового представления набора геномов и автоматизации анализа геномных последова­ тельностей (исправлении ошибок в гомополимерных областях, поиска контами­ наций в наборе прочтений, построения ортогрупп, филогенетического анализа). Проведена сборка последовательностей геномов изолятов *E. coli*, полученных от пациентов с болезнью Крона, и проведено сравнение их с геномами ком­ менсальных штаммов. Проведен анализ расположения областей повышенной изменчивости у различных родов, видов и внутривидовых структур прокариот.

**Публикации.** Основные результаты по теме диссертации изложены в 12 печатных изданиях, 6 из которых изданы в периодических научных журна­ лах, индексируемых Web of Science и Scopus, 6 — в тезисах докладов.

**Объем и структура работы.** Диссертация состоит из введения и 6 глав. Полный объём диссертации составляет 131 страницу, включая 45 рисунков и 2 таблицы. Список литературы содержит 231 наименование.

Диссертационная работа была выполнена при поддержке гранта россий­ ского научного фонда №16­15­00258 “*E. coli* как мишень терапии при болезни Крона” (руководитель Побегуц О.В.).

# Методология и методы исследования

**Анализ изменчивости геномов отдельных видов прокариот**

## Отбор последовательностей геномов для анализа геномной вариабельности

Для анализа геномной изменчивости *E. coli* мы использовали 327 геномов данно­ го организма доступные в базе RefSeq на момент ноября 2017 года и собранных до уровня репликонов (”финишированная” сборка). Для анализа изменчивости

в различных филогруппах *E. coli* нами были отобраны пять геномов — предста­ вителей наиболее крупных филогрупп данного организма (подбор проводился на основе литературных данных); затем для каждого представителя были вы­ браны 100 наиболее близких по нуклеотидному составу геномов, доступных в базе RefSeq. Для поиска наиболее близких последовательностей геномов мы проводили выравнивание программой nucmer и суммарную длину выровненных участков использовали в качестве меры сходства последовательностей.

Для анализа внутривидовых структур у *Pseudomonas aeruginosa*, *Pseudomonas fluorescens* и *Neisseria gonorrhoeae* нами были выгружены все полногеномные последовательности, доступные в RefSeq. Для каждого вида в отдельности было построено филогенетическое дерево при помощи утилиты ParSNP v1.2. На основании полученных филогенетических деревьев мы выбрали (визуальным анализом, основываясь на количестве геномов и изолированности от иных клад) от двух до четырех клад дерева.

Для анализа других видов прокариот мы собрали набор последова­ тельностей геномов всех видов, для которых было доступно не менее 50 последовательностей геномов в базе данных RefSeq. При наличии более 100 последовательностей геномов, в анализ включали 100 случайно выбранных последовательностей. Таким образом была сформирована выборка из 143 видов прокариот, включая два вида архей.

## Анализ геномной вариабельности

Белок кодирующие последовательности во всех загруженных геномах были ан­ нотированы с помощью программы Prokka ver 1.11. Гены были отнесены к ортогруппам с помощью OrthoFinder ver. 2.2.6.

Скрипты на языке Python, содержащиеся в разработанном нами прило­ жении GCB, использовали для оценки уровня изменчивости генома и создания подграфов вокруг интересующих областей генома. Принципы их работы опи­ саны в соответствующих разделах главы Результаты. В разработке приложения приняли участие: Конанов Д.Н., Федоров Д.Е., Верещагин Р.И.

Визуализация подграфов проводилась в программе Cytoscape. Для фор­ мализации определения областей генома с повышенной изменчивостью мы использовали критерий Тьюки, основанный на межквартильном расстоянии.

Статистическую обработку и визуализацию данных мы проводили на язы­ ке R. Для определения коэффициентов корреляции Спирмена использовали функцию *cor*. Статистическая значимость корреляций определялась при помо­ щи функции *cor.test*. Индексы согласованности признаков с филогенетическим деревом (retention index) мы рассчитывали с использованием функции RI из библиотеки phangorn для языка R. Для построения линейных моделей исполь­ зовалась функция *lm* языка R.

Для построения филогенетического дерева различных видов рода *Bacillus* мы выровняли транслированные последовательности всех ортоло­ гичных однокопийных генов при помощи программы muscle, преобразовали их в выравнивания кодонов с помощью pal2nal и построили дерево с помощью iqtree v1.6 с опцией ModelFinder Plus (оптимальный подбор эво­ люционной модели); конвейер snakemake для этих шагов доступен по адресу https://github.com/paraslonic/orthosnake/blob/tree/Snakefile\_tree.

Поиск областей синтении мы проводили с помощью программы nucmer (при сравнении последовательностей геномов различных штаммов одного вида), либо программы Mauve (при сравнении последовательностей геномов принад­ лежащих различным видам).

Для определения профагов в геномах мы использовали онлайн сервис Phaster.

Для нормировки матрицы частот хромосомных контактов использовалась функция *normalizeCore.performIterativeCorrection* из библиотеки gcMapExplorer.

# Сборка и анализ геномов *E. coli* от пациентов с болезнью Крона

## Группа пациентов и клинический материал

Пациенты были отобраны из двух клинических центров (ЦНИИ гастроэнтеро­ логии и Государственного научного центра колопроктологии) в Москве, Рос­ сийская Федерация, с 2012 по 2014 год. В исследование были включены десять пациентов. Критерии включения были следующими: возраст старше 18 лет, болезнь Крона была диагностирована эндоскопически и гистологически под­ тверждена. Критериями исключения были признаки неопределенного колита, инфекционные заболевания, недавнее лечение антибиотиками. Для исследова­ ния были собраны три типа образцов: 1) образцы кала; 2) биопсийный материал, полученный в ходе эндоскопического исследования; 3) жидкое содержимое подвздошной кишки. В подборе пациентов и организации сбора материала при­ нимали участие: Щербаков П.Л., Маев И.В., Павленко А.В., Андреев Д.Н., Халиф И.Л.

**Выделение изолятов *E. coli***

Выделение *E. coli* выполняли следующим образом. Приблизительно 0,05 мл объема фекалий помещали в 0,5 мл стерильного буфера (PBS), перемешива­ ли на вортексе до гомогенности, аликвоту разбавляли примерно в 106 раз. Затем 0,1 мл полученной жидкости наносили на чашки со средой LB. После инкубации в течение ночи при 37*◦*C изолированные колонии идентифициро­ вали с помощью программного обеспечения Matrix Assisted Laser Desorbtion / Ionization (MALDI) Biotyper (Bruker Daltonics, Германия) с использованием масс­ спектрометра Microflex LT (Bruker Daltonics, Германия). Для экстракции ДНК

все штаммы *E. coli* выращивали в бульоне LB при 37*◦*C при встряхивании (200 об/мин) в течение ночи и собирали центрифугированием.

## Экстракция ДНК и геномное секвенирование

Геномную ДНК из отдельных культур экстрагировали с помощью набора QIAamp DNA Mini (Qiagen) в соответствии с протоколом производителя. Экс­ трагированная ДНК (100 нг для каждого образца) была разрушена на фрагменты размером 200 – 300 пар нуклеотидов с помощью системы Covaris S220 (Covaris, Woburn, Massachusetts, USA). Эмульсию ПЦР проводили с помощью набора Ion PGM Template OT2 200 (Life Technologies). Секвенирование ДНК выполняли с помощью Ion Torrent PGM (Life Technologies) с чипом Ion 318 и набором Ion PGM Sequencing 200 v2 (Life Technologies).

Получение культур и секвенирование проводилось в геномном центре ФНКЦ ФХМ при участии Кострюковой Е.С., Бабенко В.В., Карповой И.Ю., Ли­ сициной Е.С.

Всего было получено 28 геномных последовательностей *E. coli* от 10 па­ циентов с болезнью Крона.

## Сборка генома, исправление ошибок в гомополимерных областях

Для сборки последовательностей геномов мы применяли программы Mira 4.0 и SPADES 3.10.0 с настройками, соответствующими типу секвенирования (Ion Torrent).

Каждая сборка проверялась на наличие возможных контаминаций (по­ следовательностей нецелевого организма) при помощи скрипта, написанного на языке R и доступного по адресу: [https://github.com/paraslonic/](https://github.com/paraslonic/BacPortrait/blob/master/portrait_spades.r) [BacPortrait/blob/master/portrait\_spades.r](https://github.com/paraslonic/BacPortrait/blob/master/portrait_spades.r).

Данный скрипт отображает каждый контиг (отдельный фрагмент сбор­

ки) на диаграмме с ГЦ­составом и глубиной покрытия контига; дополнительно отображается информация о таксономической аннотации выбранного числа кон­ тигов.

Для технологии секвенирования Ion Torrent характерно наличие значитель­ ного количества ошибок в определении копийности нуклеотидов, особенно в гомополимерных областях. Для исправления данного типа ошибок, которые мо­ гут приводить к ошибкам сборки и искусственному сдвигу рамки считывания в кодирующих последовательностях (CDS), нами был разработан следующий метод. Проводилось картирование прочтений на сборку; поиск позиций со вставками либо делециями в картированных прочтениях при помощи утилиты VarScan; выравнивание областей сборки вокруг найденных позиций при помо­ щи программы BLAST на базу nt (NCBI); выбор варианта последовательности, который соответствует лучшему выравниванию BLAST и представлен в прочте­ ниях с частотой не ниже 25%. Этот метод уменьшает количество артефактных мутаций в сборке примерно в 2,5 раза (оценка основана на сравнении сборок

считываний Ion Torrent до и после исправления со считываниями более точных технологий секвенирования, таких как Illumina, SOLID и Sanger). Вычислитель­ ный конвейер для данной процедуры доступен по адресу: [www.github.com/](http://www.github.com/paraslonic/HomoHomo) [paraslonic/HomoHomo](http://www.github.com/paraslonic/HomoHomo).

## Сбор внешних данных

Подбор последовательностей геномов *E. coli* с охарактеризованным источником выделения проводился на основе анализа литературных источников. Было ото­ брано 24 последовательности генома комменсальных штаммов *E. coli* (выделен­ ных от здоровых людей), 17 последовательности генома изолятов, полученных от пациентов с болезнью Крона, 29 последовательностей геномов изолятов по­ лученных от пациентов с иными заболеваниями, для которых кишечная палочка была описана как инфекционный агент.

## Поиск ортогрупп

Полученные последовательности геномов аннотировали с помощью программы PROKKA 1.7. Информация об оперонной структуре была получена из базы дан­ ных DOOR. Ортогруппы (группы гомологий включающие как ортологичные, так и паралогичные гены) были получены с помощью программы OrthoFinder v1.0.8. Статистический анализ представленности генов и оперонов в последова­ тельностях генома комменсальных штаммов *E. coli*, либо штаммов, полученных от пациентов с болезнью Крона, проводился при помощи скриптов на языке R,

принцип работы которых описан ниже.

В организации работы и интерпретации полученных результатов прини­ мали участие: Говорун В.М., Ракитина Д.В., Гарушянц С.К.

# Результаты

**Графовое представление геномов**

Нами был предложен и реализован метод представления набора геномов в виде графа. Гены, относящиеся к одной ортогруппе, представляются в виде узла гра­ фа, а ребра отражают взаимное расположение генов в геномах. Ребро проводится между двумя узлами в том случае, если соответствующие гены расположены по соседству хотя бы в одном геноме; чем в большем количестве геномов наблюда­ ется такое расположение генов, тем выше вес ребра (рисунок [1](#_bookmark0)).

Предложенная схема требует уточнения в случае, когда несколько ге­ нов из одной геномной последовательности принадлежат к одной ортогруппе (мы будем называть такие гены паралогами). Мы предложили два подхода представления паралогичных генов в графе. В первом случае мы игнорируем паралогичные гены и не добавляем их в граф. Во втором — мы проводим про­ цедуру ”ортологизации” паралогичных генов: гены из ортогруппы добавляем в

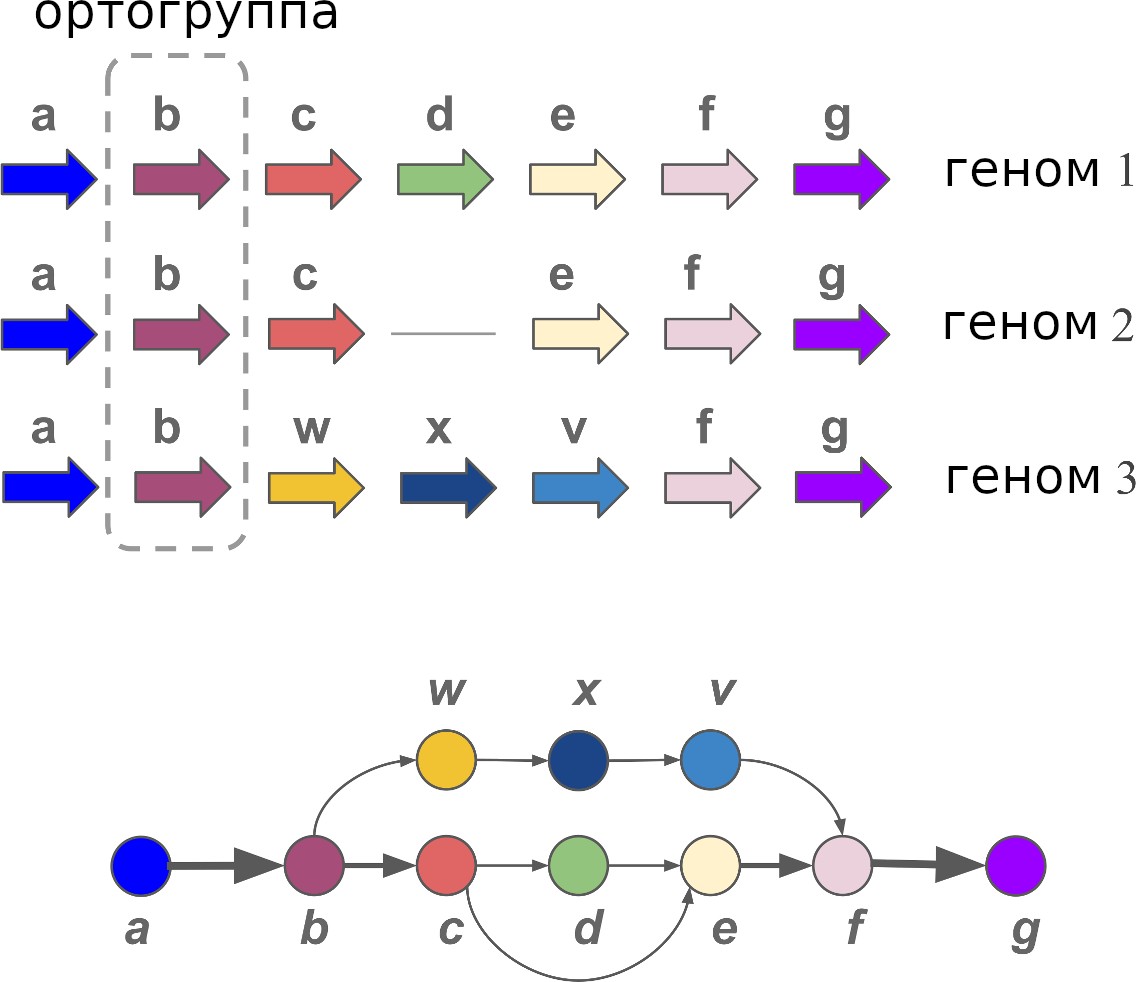


Рис. 1 –– Представление контекста генов в виде графа. Рассматриваем три гипо­ тетических генома, состоящих из 6­7 генов. Гены показаны стрелками, цветом и буквами обозначены гены, относящиеся к одной ортогруппе. В нижней части рисунка показано графовое представление для данного набора геномов.

граф с некоторым суффиксом, при этом гены, имеющие одинаковый контекст, добавляются с одним и тем же суффиксом. В нашей программной реализации пользователь может выбирать применяемый подход.

# Оценка изменчивости геномов при помощи графового представления

Для исследования изменений генного состава в отдельном локусе генома, мы разработали процедуру построения подграфа, включающего только опреде­ ленную область интереса референсного генома. Мы предложили использовать подсчет количества путей в соответствующем подграфе в качестве меры локаль­ ной изменчивости. В случае, если в некотором регионе не происходит изменений генного состава, то подграф представляет из себя простую структуру с одним возможным путем (способом обхода узлов). В случае, если в данной обла­ сти происходят изменения, то в подграфе появляются новые пути. Чем больше различных вариантов сочетаний генов наблюдается в геномных последователь­ ностях, тем больше путей в соответствующем подграфе.

Оценка локальной вариабельности производится для одного выбранного гена референсного генома с учетом окна (смежных генов), размер которого зада­ ется пользователем. Алгоритм расчета состоит в следующем. Вначале строится референсная цепочка узлов, после чего происходит подсчет путей, которые оги­ бают узел (то есть начинаются по одну сторону, а заканчиваются – по другую). Для подсчета числа путей в графе мы применили вероятностный подход, что поз­ волило увеличить скорость вычислений. Поиск одного пути происходит, начиная с первого узла цепочки, и затем случайным образом выбираются последующие вершины, пока путь не вернется в референсную цепочку, либо пока не будет достигнуто максимальное количество переходов. Совершается фиксированное количество итераций поиска путей, каждый новый уникальный путь запомина­ ется. Количество найденных различающихся путей мы считаем мерой локальной изменчивости.

# Верификация метода оценки локальной изменчивости

Для верификации предложенного алгоритма и его программной реализации мы использовали компьютерное моделирование изменений геномных последова­ тельностей. Для каждого локуса моделируемого генома мы задавали опреде­ ленный уровень изменчивости, затем вероятностным образом изменяли генный состав, получая множество различающихся геномов, и оценивали наблюдаемую вариабельность при помощи предложенного нами подхода. Изменения генно­ го состава происходили за счет вставки, удаления либо перемещения генов, а также инверсий фрагментов генома. Вероятность инверсий была в 100 раз мень­ ше вероятностей остальных событий, а размер области для инверсии выбирался случайно из экспоненциального распределения. Локализация событий измене­ ний генного состава выбиралась на каждом шаге случайно, в соответствии с распределением уровня вариабельности. Значения локальной вариабельности мы задавали в соответствии с одним из типов профилей: пилообразным, сту­ пенчатым либо синусоидальным. Значения коэффициента *R*2 между исходно заданным уровнем изменчивости и тем, что был оценен при помощи графо­ вого подхода, составили 0.95, 0.77 и 0.8 для ступенчатого, синусоидального и пилообразного профиля, соответственно, корреляции имели статистическую значимость с *p − value <* 10*−*10.

# Анализ расположения областей повышенной изменчивости в геномных последовательностях кишечной палочки

Для проведения данного анализа мы собрали коллекцию из 327 геномов бакте­ рии *E. coli*, доступных в базе данных RefSeq, построили группы гомологии при помощи программы Orthofinder и применили разработанной нами метод оценки уровня геномной вариабельности; в качестве референсного генома был выбран геном штамма *E. coli LF82*, изолированного от пациента с болезнью Крона. На

рисунке [2](#_bookmark1) показан профиль изменчивости и области расположения профагов, островов патогенности, генов рекомбиназ (для выявления следов мобильных элементов генома), жизненно необходимых генов (мутации в которых являют­ ся летальными) и генов транспортных и рибосомных РНК.

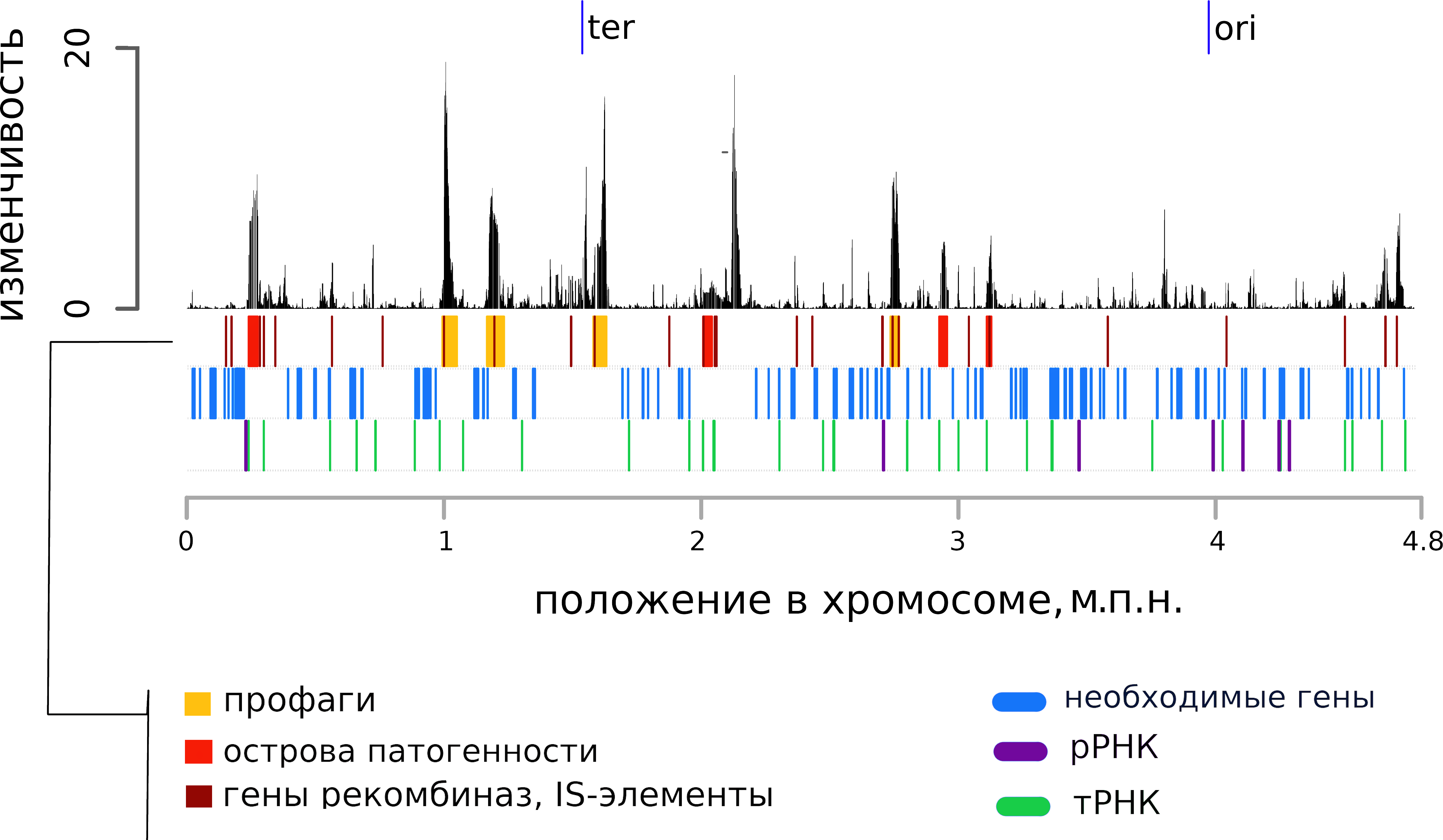


Рис. 2 –– Профиль вариабельности генома *Escherichia coli LF82*. Цветом обо­ значены острова патогенности, профаги, гены, ассоциированные с мобильными элементами генома, жизненно необходимые гены, гены транспортных и рибо­ сомных РНК. Области повышенной изменчивости содержат меньше жизненно необходимых генов. Профаги и острова патогенности обладают повышенным уровнем изменчивости по сравнению с остальной частью генома. Вертикальные линии с подписями ter и ori показывают места начала и окончания репликации, соответственно.

Можно заметить, что участки генома, содержащие жизненно необходимые гены, как правило, мало изменчивы, а к высокоизминчивым областям генома относятся профаги и острова патогенности. При этом, наблюдаются также вы­ сокоизменчивые области генома, в которых отсутствуют признаки мобильных элементов; причины их высокой изменчивости остаются неизвестными. Гены транспортных РНК не проявляют явной ассоциации с профилем изменчивости. Гены рибосомной РНК находятся преимущественно в мало изменчивых обла­ стях генома.

Время существования вида *E. coli* оценивается в несколько десятков мил­ лионов лет; возникает вопрос об устойчивости расположения ”горячих” обла­ стей генома с течением времени. Мы провели сравнение профилей изменчивости

подвидовых структур кишечной палочки. Для создания набора геномов мы вы­ брали по одному представителю из наиболее крупных филогенетических клад (филогрупп A, B1, B2, E) и подобрали для каждого из них по сто ближайших геномов из базы NCBI RefSeq. Затем мы рассчитали профили изменчивости для каждой филогруппы в отдельности и сравнили их между собой. Результат сравнения показан на рисунке [3](#_bookmark2), серым цветом обозначены блоки синтении, оранжевым – области нахождения профагов.

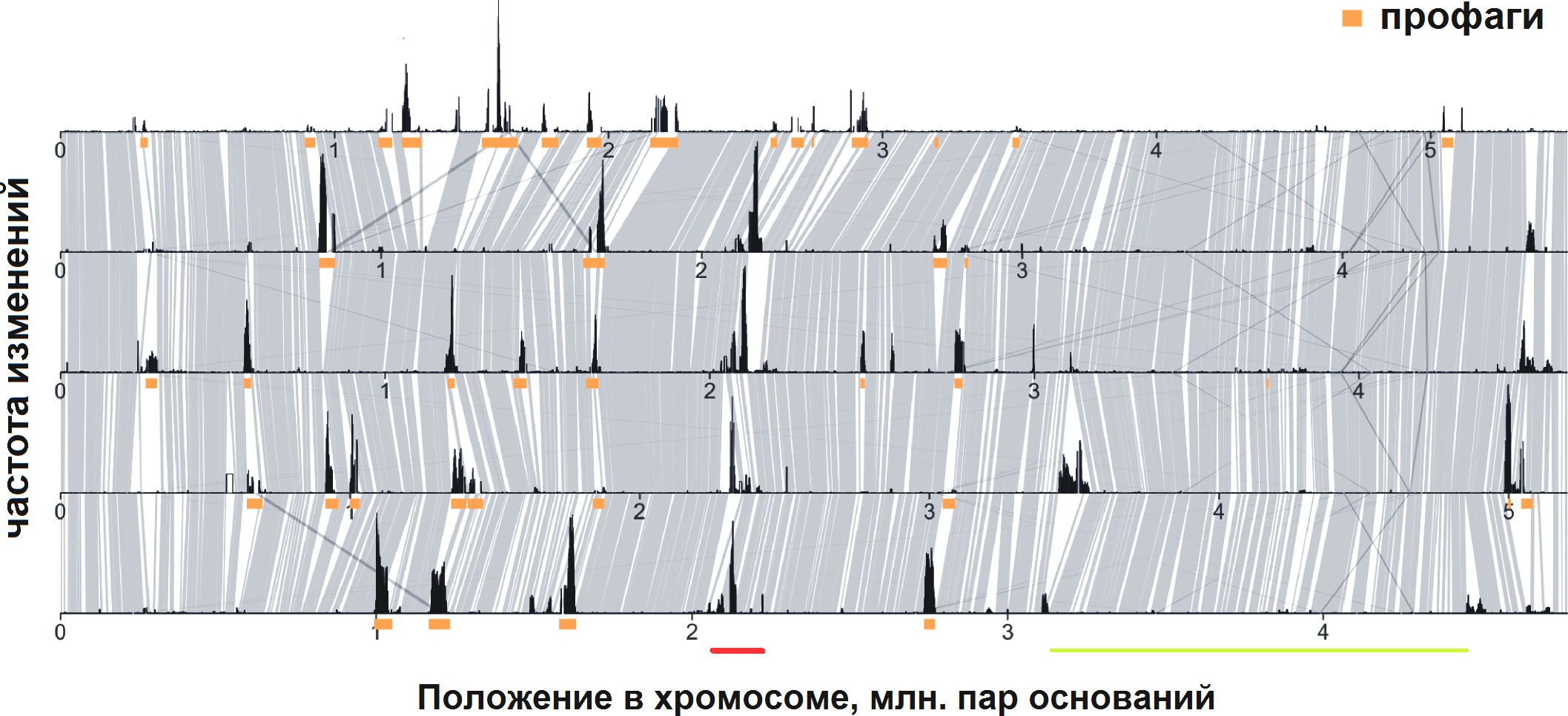


Рис. 3 –– Сравнение профилей вариабельности представителей пяти филогрупп

*E. coli LF82*. Оранжевым цветом выделены области, определенные как профаго­ вые. Блоками серого цвета показаны области синтении. Показаны профили для штаммов: O157:H7 (филогруппа E), IAI1 (B1), K12 (A), UMN026 (D), LF82 (B2).

Сравнение показывает, что ряд областей генома с повышенной изменчи­ востью присутствуют во всех или в большинстве филогрупп, что, по­видимому, свидетельствует об их существовании на протяжении значительной части вре­ мени жизни вида. Часть из этих областей соответствуют местам расположения профагов, и их устойчивость можно объяснить сайт специфичным характером встраивания данных элементов. Одна из устойчивых ”горячих” точек (выделена линией красного цвета) не имеет выявленных признаков мобильных элемен­ тов генома. Значительная часть ”горячих” областей геномов является местами встройки фагов, что особенно заметно на примере филогруппы E (верхний про­ филь на рисунке [3](#_bookmark2)), для которой была описана вирусная экспансия, значительно удлинившая геномную последовательность (на рисунке длины геномов приведе­ ны к одинаковой ширине, но о размере генома можно судить по координатным осям ниже профилей изменчивости).

Имеются также и ”холодные” области генома с низкой вариабельностью во всех филогруппах. Длина этих участков может значительно превышать длины,

характерные для оперонов, и достигать величин порядка миллиона пар основа­ ний (например, область, выделенная зеленой линией на рисунке [3](#_bookmark2)).

# Алгоритмизация подхода выявления оперонов, наличие которых ассоциировано с определенным признаком.

Предположим, что у нас есть некоторый признак, по которому мы можем раз­ бить набор геномов на группы и наша задача — установить, какие гены значимо чаще (либо реже) встречаются в одной из групп. Простым и распространенным методом анализа, в подобном случае, является вычисление статистики и оценка значимости по каждому отдельному гену. Затем необходимо применить по­ правку на множественное сравнение, так как иначе следует ожидать множество ложно положительных результатов. В случае, если работа ведется на данных о полной последовательности генома, имеется большое количество анализируе­ мых генов (порядка 103 104), а размеры групп, как правило, незначительны (порядка 101 102), что приводит к тому, что после поправки на множественное сравнение, ни один из анализируемых генов не проходит даже низкие пороги на значимость. В качестве одного из способов преодоления описанной выше проблемы мы предложили использовать информацию об организации генов в опероны для поиска значимых ассоциаций. Рассматривая оперон как структур­ ную единицу, мы значительно сокращаем количество анализируемых признаков.

*−*

*−*

**Алгоритм поиска генетических ассоциаций** На входе необходимо иметь два или более набора геномных последовательностей и выбрать один рефе­ ренсный геном, для которого известна оперонная структура. Первым шагом выполняется построение групп гомологий и оценка статистической значимо­ сти их неравной представленности в сравниваемых выборках. Затем ищутся опероны, в которых количество найденных ассоциированных генов выше, чем ожидалось бы при случайном распределении генов по оперонам. Для оценки ожидаемого количества генов при их случайном распределении мы предложили использовать два подхода. Первый основан на пермутациях таблицы соответ­ ствий генов и оперонов. Второй предполагает расчет ожидаемого значения, исходя из распределения Пуассона, параметр которого рассчитывается как до­ ля значимо ассоциированных генов среди общего числа генов в референсном геноме. Финальным шагом является проведение поправки на множественное сравнение, но уже не для генов, а для оперонов, количество которых, как пра­ вило, значительно ниже общего количества генов.

## Поиск оперонов, значимо чаще встречающихся у изолятов бактерий

***E. coli*, изолированных от людей с болезнью Крона** В анализе были исполь­ зованы геномные последовательности 51 изолята *E. coli*, 27 из которых были получены от пациентов с болезнью Крона, 24 — от здоровых людей. При по­ мощи программы OrthoFinder мы получили 11885 групп гомологии. Далее, при помощи точного теста Фишера оценили статистическую значимость их нерав­ номерной распространенности между группами.

Следующим шагом мы осуществили описанный выше тип анализа, для поиска значимо дифференциально представленных оперонов. Информация об оперонах была взята из базы данных DOORS. В качестве референсного гено­ ма мы использовали геном *E. coli LF82* ­ данный штамм был изолирован из пациента с болезнью Крона и является модельным в исследованиях адгезивно­ инвазивного фенотипа у кишечной палочки. Затем, мы провели 10000 случайных перестановок соответствий между генами и оперонами. Для каждой переста­ новки мы вычисляли зависимость количества генов с p­value < 0.05 от длины оперона. Визуализация сравнения наблюдаемых и полученных при случайных перестановках результатов показана на рисунке [4](#_bookmark3). Опероны, для которых на­ блюдаемое число генов было выше, чем максимальное количество генов при случайных перестановках, мы считали статистически значимо пере­ либо недо­ представленными (в пермутационном тесте p­value < 0.0001).

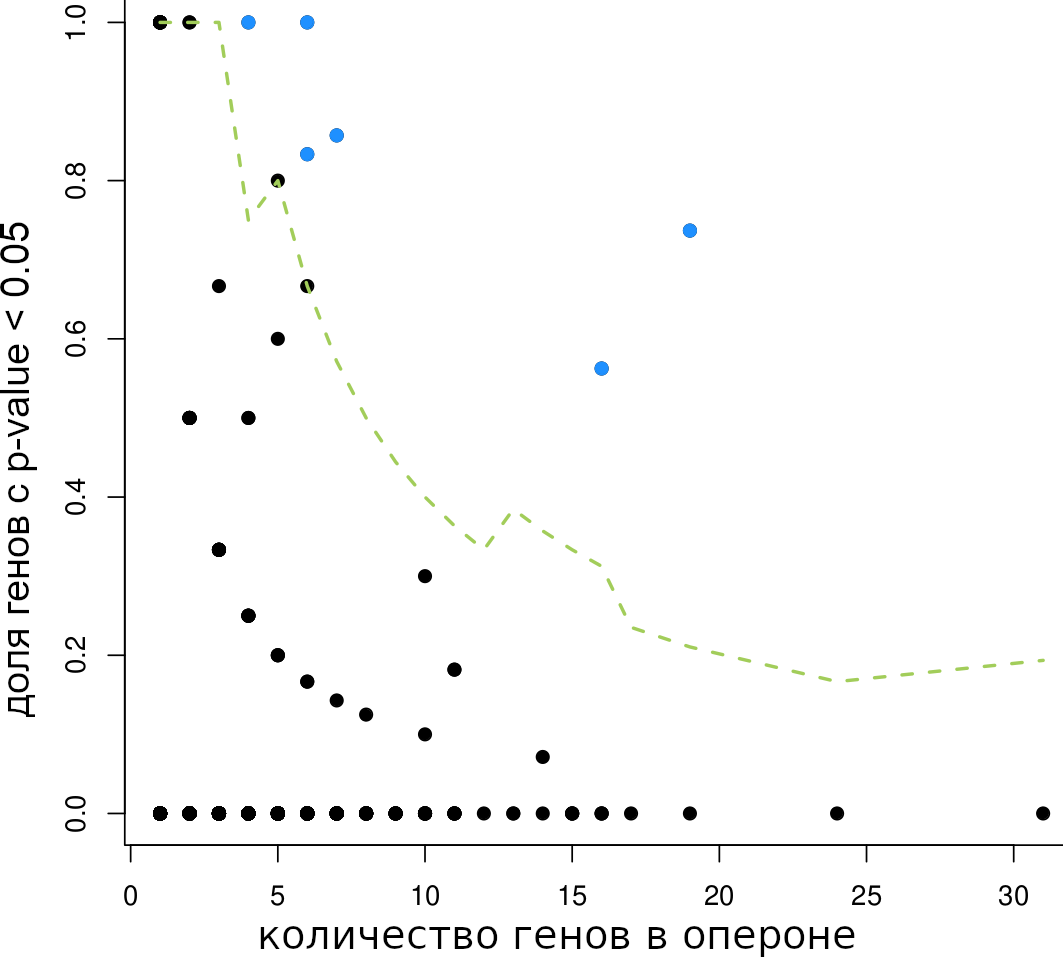


Рис. 4 –– Зависимость доли генов в оперонах, с уровнем значимости p­value < 0.05, от количества генов, входящих в оперон. Пунктирная линия показывает максимальные значения, полученные при проведении 10000 случайных соот­ несений генов и оперонов. Синим цветом выделены точки, соответствующие оперонам, которые неравномерно представлены между группами согласно пер­ мутационному тесту.

Так, например, оперон утилизации пропандиола состоит из 19 генов, из которых 14 генов (74%) имеют p­value < 0.05 в точном тесте Фишера. При про­ ведении 10000 случайных пермутаций уровней значимости по генам в данном опероне в среднем наблюдалось 3% перепредставленных генов, а максималь­ ная доля составила 21%. Таким образом, можно сделать вывод, что повышенная представленность данного оперона в изолятах из пациентов, но не из здоровых людей, не является случайным наблюдением. Полный список оперонов, опреде­ ленных как значимо чаще встречающихся у изолятов из пациентов с болезнью Крона приведен в таблице [1](#_bookmark4).

Таблица 1 –– Список оперонов статистически значимо перепредставленных в группе штаммов *E. coli* изолированных от пациентов с болезнью Крона. N ­ ко­ личество генов, Pobs ­ наблюдаемая доля перепредставленных генов в опероне, Pmean ­ средняя доля перепредставленных генов при случайных пермутациях, Pmax ­ максимальная доля перепредставленных генов при случайных пермута­ циях.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **N** | **Pobs** | **Pmean** | **Pmax** | **функция** |
| 4 | 1 | 0.03 | 0.75 | утилизации глиоксилата |
| 6 | 0.83 | 0.02 | 0.67 | синтез и экспорт капсулы |
| 6 | 1 | 0.02 | 0.67 | захват гемина |
| 7 | 0.86 | 0.03 | 0.57 | утилизации сорбозы |
| 19 | 0.74 | 0.03 | 0.21 | утилизации пропандиола |

# Разработка и применение графового подхода для визуализации локальной вариабельности геномов

Множество сочетаний генов, встречающихся в наборе геномов, можно предста­ вить в виде графа и использовать его для визуализации сравнений геномных последовательностей. Преимуществом данного подхода является компактность визуализации, так что становится возможным сравнение генных контекстов в сотнях геномных последовательностей. Компактность достигается за счет того, что сочетание генов, одинаковое для всех геномов, представляется при помощи одного ребра, вне зависимости от количества геномов. Новые узлы и ребра добав­ ляются в граф только когда соответствующие сочетания ранее не наблюдались.

Визуализация полного графа для набора геномов возможна в случае небольших вирусных геномов; в случае бактериальных геномов имеет смысл проводить визуализацию и анализ подграфа — части полного графа, соответ­ ствующей некоторому региону интереса (например, оперону).

Алгоритм построения подграфа схож с описанным выше. Первым шагом является построение референсной цепочки узлов — узлов графа, которые со­ ответствуют генам референсного генома расположенным в выбранном регионе.

Далее, в подграф включаются все пути, которые начинаются или заканчиваются на референсной цепочке. Мы предусмотрели несколько фильтров, позволяющих снизить количество отображаемых узлов и ребер. Фильтр минимального веса ребра позволяет отсечь ребра, которые представляют редкие сочетания генов (встречающихся в небольшом количестве геномов). Фильтр длинных путей за­ меняет слишком протяженные цепочки узлов (они могут появиться в результате геномных перестроек) на короткие фрагменты (”хвосты” ) некоторой заданной длины.

Визуализация подграфов обладает практической ценностью, поскольку позволяет создавать наглядную визуализацию возможных сочетаний генов в определенной области генома. Подобный тип визуализации можно использовать для определения того, представлен ли некоторый оперон в одном и том же генном контексте, или в разных. Под генным контекстом мы понимаем гены, располо­ женные непосредственно перед и после оперона. Мы говорим об одинаковом контексте в случае, если гены перед опероном относятся к одной ортогруппе, и гены после оперона относятся к некоторой другой ортогруппе. В графовом пред­ ставлении при этом мы будем наблюдать, что цепочка узлов, соответствующая генам оперона, будет соединена с одним узлом с одной стороны и с другим уз­ лом — с другой.

# Примеры применения графового подхода для визуализации сравнения геномов

Рассмотрим, как выглядят подграфы, соответствующие некоторым из оперонов статистически значимо чаще встречающихся у изолятов *E. coli*, полученных от пациентов с болезнью Крона, по отношению к изолятам от здоровых людей. На этих примерах будут проиллюстрированы основные моменты анализа графового представления фрагментов генома.

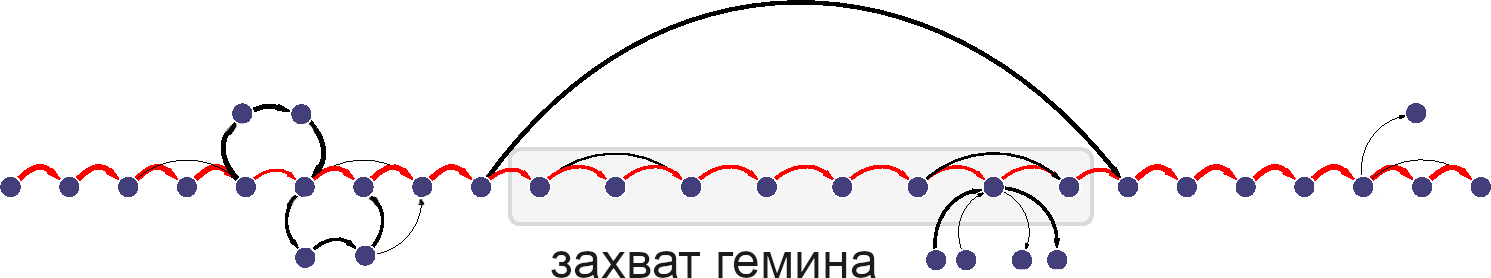


Рис. 5 –– Граф представляющий окрестность оперона утилизации гемина (hemin uptake, hmu). В ряде геномов оперон отсутствует, что видно из наличия ребра в графе, которое начинается перед и заканчивается непосредственно после генов оперона.

На рисунке [5](#_bookmark5) показан граф, построенный в окрестности оперона захвата гемина (hemin uptake, hmu). Как видно из графа, данный оперон расположен

в консервативном генном контексте. Дуговое ребро обходящее оперон сверху говорит о том, что в некотором наборе геномов данный оперон отсутствует и других последовательностей генов в этом локусе не наблюдается.

На рисунке [6](#_bookmark6) показан граф, построенный в окрестности оперона ути­ лизации пропандиола (propanediol utilization operon, pdu). Оперон утилизации пропандиола также имеет одинаковый контекст в разных геномах. Ребро, обхо­ дящее оперон (дуга ниже оперона), говорит о том, что в ряде штаммов в данном контексте нет иных вариантов последовательностей генов. Наблюдается неко­ торая вариабельность внутри оперона, соответствующая нескольким вариантам данного оперона. Помимо pdu оперона, в том же контексте у ряда штаммов пред­ ставлен альтернативный набор генов. В этот альтернативный набор входят гены транспорта железа (FepC, FcuA, HmuU), гены мобильных элементов (retroviral integrase core domain, transposase DDE Tnp ISL3) и множество гипотетических генов с неизвестной функцией. Примечательна вариабельность этого альтерна­ тивного набора генов. Вероятно, данный участок генома часто служит местом рекомбинационных событий, приводящих к изменению генов, причем эти изме­ нения часто накладываются друг на друга (вероятно, не сайт специфичны).

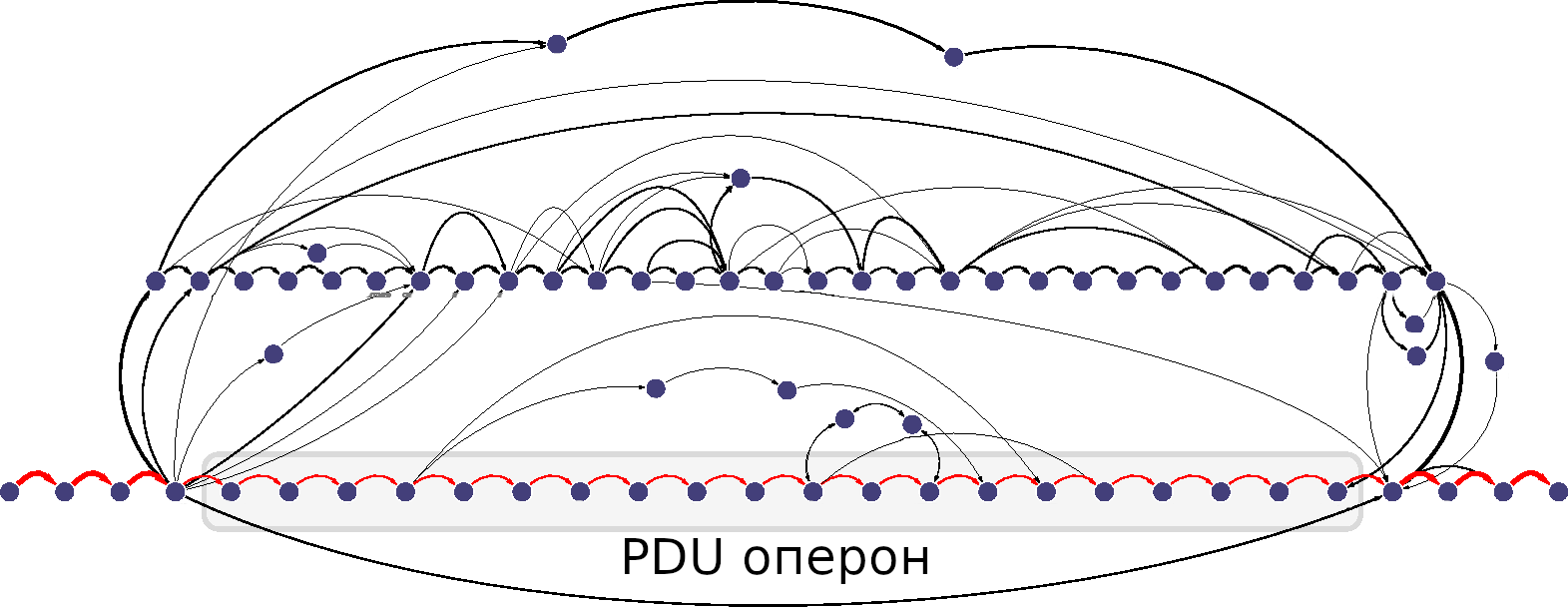


Рис. 6 –– Граф представляющий окрестность оперона утилизации пропандиола (propanediol utilization operon, pdu). Видна значительная вариабельность генного состава в данном регионе, что отражено во множестве путей на графе обходящих область генов оперона.

На рисунке [7](#_bookmark7) показано ближайшее окружение кластера генов, отвечающего за синтез и транспорт полисахаридов бактериальной капсулы группы 2 (описа­ на как остров патогенности капсульной сборки IV у *E. coli LF82*). Видно, что оперон состоит из консервативных фрагментов, окружающих вариабельный уча­ сток. Вариабельная часть оперона соответствует генам, отвечающим за синтез серотип­специфичного набора полимеров капсулы; гены консервативной части кодируют белки, участвующие в транспорте синтезированных веществ через клеточную стенку.

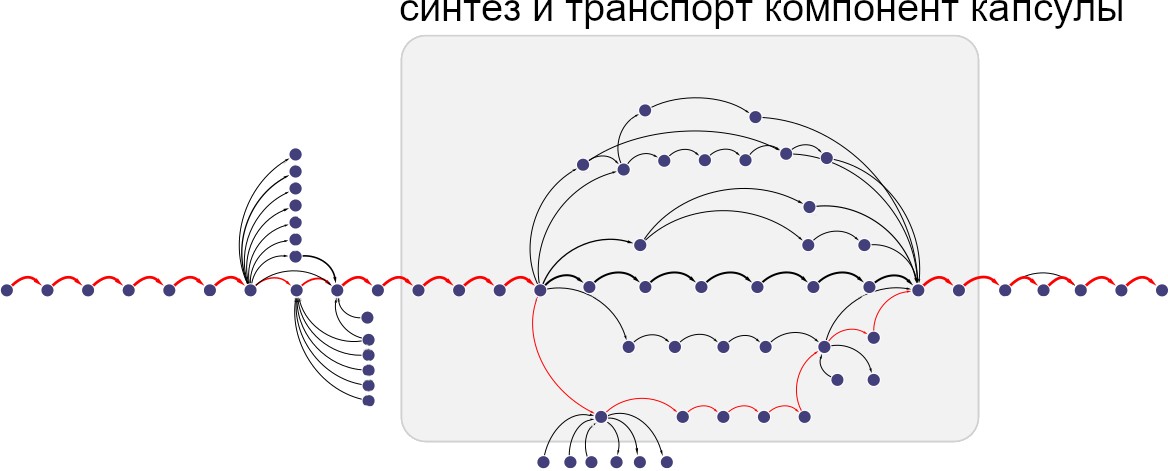


Рис. 7 –– Граф представляющий окрестность кластера генов синтеза и транспор­ тировки компонент бактериальной капсулы.

Таким образом, предложенный нами способ визуализации позволяет опре­ делить общую вариабельность генного состава в определенном локусе генома, выявить встречающиеся варианты взаимного расположения генов внутри оперо­ на и найти вариабельные и консервативные области оперонов при их наличии.

# Разработка компьютерного приложения для анализа вариабельности геномов

Оценка профиля изменчивости и визуализация подграфов — это варианты ана­ лиза геномной изменчивости на различных уровнях: на уровне хромосомы или плазмиды в первом случае и на уровне небольших геномных локусов (например, оперонов) — во втором. Для проведения анализа на двух уровнях одновременно мы разработали приложение Genome Complexity Browser (GCB). Данное прило­ жение доступно по адресу [https://gcb.rcpcm.org](https://gcb.rcpcm.org/), и может быть запущено на локальном компьютере пользователя. Веб­версия содержит данные о геном­ ной изменчивости для 143 видов прокариот. Использование локальной версии необходимо для анализа групп геномов не представленных на веб­сервере.

Основные сценарии использования программы следующие:

1. Если интерес представляет некоторый оперон или группа генов.

Пользователь выбирает организм, геном и задает координаты области интереса. Происходит построение подграфа для выбранной области. При необ­ ходимости, пользователь меняет параметры визуализации (например, если по­ лучаемый граф слишком сложен для анализа можно увеличить минимальный отображаемый вес ребра для исключения редко встречающихся комбинаций ге­ нов). Доступен экспорт визуализации подграфа в графическом формате, либо в формате XML для последующей визуализации его в программе Cytoscape или иных редакторах графов.

1. Если интерес представляют области повышенной либо пониженной из­ менчивости генома.

Пользователь выбирает организм и геном. Происходит визуализация про­ филя изменчивости выбранного генома. Пользователь может выбрать инте­ ресующий его регион генома (например, область с максимальным уровнем изменчивости) и выполнить визуализацию подграфа в данной области ­ таким образом можно установить, какие гены содержатся в данном локусе у различ­ ных геномов, и каков паттерн изменений. Пользователь может экспортировать профиль изменчивости в виде текстового файла для последующей визуализации (например, для сравнения профилей изменчивости разных организмов), либо со­ хранить области с повышенной изменчивостью в файл в формате BED.

Документация и видеоматериалы по использованию программы (как веб­версии, так и консольных утилит) доступны по адресу [https://gcb.](https://gcb.readthedocs.io/) [readthedocs.io/](https://gcb.readthedocs.io/).

# Заключение

В настоящей работе реализовано представление набора геномов в виде графа для количественной оценки уровня изменчивости в отдельных локусах генома и визуализации подграфов соответствующих отдельным областям генома.

Визуализация подграфов позволяет дать ответы на ряд вопросов о контек­ сте генов интереса. Например, находится ли ген или гены интереса в одинаковом окружении во всех рассматриваемых геномах? Какие альтернативные генные контексты существуют и в каких геномах они представлены? Какие части набора генов (например, оперона или генного острова) являются консервативными, а ка­ кие вариабельными? Какие геномы содержат определенную комбинацию генов? При помощи графового представления геномов мы реализовали метод количественной оценки локальной изменчивости, основанный на поиске уни­ кальных путей в подграфе. Под изменчивостью в данном случае мы понимаем изменение состава либо взаимного расположения генов в геноме. Под ло­ кальностью — то, что изменения затрагивают небольшую область генома, не превышающую размер выбранного окна анализа (выбирается пользователем, обычно составляет около 20­40 генов). Насколько нам известно, разработанный нами вычислительный конвейер (Genome Complexity Browser, GCB) является первым доступным инструментом, позволяющим количественно определять из­ менчивость генома на основе заданного пользователем набора геномов. GCB предоставляет способ оценки профиля изменчивости вдоль репликонов, что поз­ воляет находить ”горячие” области генома, в которых уровень изменчивости значительно выше, чем в остальной части генома, и его ”тихие” области. Про­ веденный анализ показал, что значительная часть высокоизменчивых областей генома соответствует местам встройки профагов и островов патогенности. В то же время, для кишечной палочки мы также наблюдали существование протяжен­ ной выосокоизменчивой области генома, в которой нет признаков мобильных

элементов. При этом данная область обладает высоким уровнем изменчивости у всех крупных филогрупп данного вида.

Мы провели поиск оперонов, которые чаще встречаются у кишечных палочек, выделенных из образцов фекалий и кишечных смывов пациентов с болезнью Крона — тяжелого воспалительного заболевания кишечника. Функ­ ция большинства найденных оперонов ясна, они позволяют захватывать железо, утилизировать пропандиол (продукт переработки слизистого слоя), менять анти­ генные свойства, тем самым убегая от иммунного ответа. Интересно, что анализ графов, представляющих контекст этих генов, показал очень разные картины. Два оперона: утилизации пропандиола и производства капсулы находятся в вы­ сокоизменчивых –”горячих” – областях генома. Опероны утилизации гемина, утилизации глиоксилата, захвата сорбозы напротив находятся в ”тихих” обла­ стях. Роль генетических факторов, находящихся в областях генома с разным уровнем изменчивости, в формировании генотипа и фенотипа бактерий — пред­ мет дальнейших исследований. В случае оперона синтеза и экспорта капсульных полисахаридов, можно предположить, что нахождение данного оперона в ”горя­ чей” области генома может способствовать более высокой изменчивости состава оперона (у него есть высоко вариативная часть, отвечающая за синтез капсу­ лы), что в свою очередь выгодно для эффективного избегания иммунного ответа организма­хозяина.

# Выводы

* 1. Графовое представление геномов позволяет эффективно проводить по­ иск областей генома с повышенной изменчивостью.
  2. Геномы представителей различных филогрупп и филогенетически близ­ ких видов прокариот имеют консервативно расположенные области повышенной изменчивости (расположенные в местах генома с одина­ ковым генным контекстом).
  3. Уровень геномной изменчивости ассоциирован с плотностью хромо­ сомных контактов (коэффициент корреляции составил ­0.36) и плот­ ностью расположения сайтов Chi (коэффициент корреляции составил

­0.25).

* 1. Следующие опероны значимо чаще (p­value < 0.0001) встречаются в изолятах *E. coli* от пациентов с болезнью Крона: захват сорбозы, захват гемина, утилизации глиоксилата, утилизации пропандиола, синтеза и экспорта капсульных полисахаридов.
  2. Оперон утилизации пропандиола и оперон синтеза и экспорта кап­ сульных полисахаридов расположены в высокоизменчивых областях, а опероны захвата сорбозы, захвата гемина и утилизации глиоксилата — в консервативных участках генома *E. coli*.

# Публикации автора по теме диссертации

## В изданиях, входящих в международную базу цитирования Web of Science

1. *Manolov*, *A.* Genome Complexity Browser: Visualization and quantification of genome variability / A. Manolov, D. Konanov, D. Fedorov, I. Osmolovsky,

R. Vereshchagin, E. Ilina // PLoS computational biology. –– 2020. –– Т. 16,

№ 10. –– e1008222.

1. *Tyakht*, *A. V.* Genetic diversity of Escherichia coli in gut microbiota of patients with Crohn’s disease discovered using metagenomic and genomic analyses /
   1. V. Tyakht, A. I. Manolov, A. V. Kanygina, D. S. Ischenko, B. A. Kovarsky,

A. S. Popenko, A. V. Pavlenko, A. V. Elizarova, D. V. Rakitina, J. P. Baikova [и др.] // BMC genomics. –– 2018. –– Т. 19, № 1. –– С. 1––14.

1. *Terekhov*, *S. S.* Ultrahigh­throughput functional profiling of microbiota communities / S. S. Terekhov, I. V. Smirnov, M. V. Malakhova, A. E. Samoilov,
   1. I. Manolov, A. S. Nazarov, D. V. Danilov, S. A. Dubiley, I. A. Osterman,

M. P. Rubtsova [и др.] // Proceedings of the National Academy of Sciences. –– 2018. –– Т. 115, № 38. –– С. 9551––9556.

1. *Rakitina*, *D. V.* Genome analysis of E. coli isolated from Crohn’s disease patients /

D. V. Rakitina, A. I. Manolov, A. V. Kanygina, S. K. Garushyants, J. P. Baikova,

D. G. Alexeev, V. G. Ladygina, E. S. Kostryukova, A. K. Larin, T. A. Semashko [и др.] // BMC genomics. –– 2017. –– Т. 18, № 1. –– С. 1––17.

1. *Bulaev*, *A.* Genome analysis of Acidiplasma sp. MBA­1, a polyextremophilic archaeon predominant in the microbial community of a bioleaching reactor /
   1. Bulaev, A. Kanygina, A. Manolov // Microbiology. –– 2017. –– Т. 86, № 1. –– С. 89––95.
2. *Zakharzhevskaya*, *N. B.* Outer membrane vesicles secreted by pathogenic and nonpathogenic Bacteroides fragilis represent different metabolic activities /

N. B. Zakharzhevskaya, A. A. Vanyushkina, I. A. Altukhov, A. L. Shavarda,

I. O. Butenko, D. V. Rakitina, A. S. Nikitina, A. I. Manolov, A. N. Egorova,

E. E. Kulikov [и др.] // Scientific reports. –– 2017. –– Т. 7, № 1. –– С. 1––16.

## В сборниках трудов конференций

1. *Манолов*, *А.* Сравнительный анализ частоты геномных перестроек у прока­ риот / А. Манолов, Д. Конанов, Д. Федоров, И. Осмоловский, Е. Ильина // “VI Съезд биохимиков России.” –– 2019. –– С. 144.
2. *Манолов*, *А.* Метод анализа контекста генов и интенсивности геномных пе­ рестроек в бактериальных геномах / А. Манолов, Д. Конанов, Д. Федоров, И. Осмоловский, Е. Ильина // “ПОСТГЕНОМ’2018”, Казань. –– 2018. –– С. 219.
3. *Манолов*, *А.* Поиск генов бактерий вида Escherichia coli ассоциированных с болезнью Крона / А. Манолов, О. Побегуц, Е. Кострюкова, А. Ларин, Т. Семашко, В. Бабенко, Р. Городничев, Е. Лисицина, П. Щербаков, Е. Ильи­ на, В. Говорун // ВЫСОКОПРОИЗВОДИТЕЛЬНОЕ СЕКВЕНИРОВАНИЕ В ГЕНОМИКЕ II Всероссийская конференция с международным участи­ ем. –– 2017. –– 46a.
4. *Манолов*, *А.* Поиск генов бактерий вида Escherichia coli ассоциированных с болезнью Крона / А. Манолов, О. Побегуц, Е. Кострюкова, А. Ларин, Т. Се­ машко, В. Бабенко, Р. Городничев, Е. Лисицина, П. Щербаков, Е. Ильина, В. Говорун // БИОТЕХНОЛОГИЯ: СОСТОЯНИЕ И ПЕРСПЕКТИВЫ РАЗ­ ВИТИЯ Москва, 20–22 февраля 2017 года. –– 2017. –– С. 487––489.
5. *Манолов*, *А.* Сборка de­novo и сравнительный анализ генома бактерии P. stutzeri KOS6, извлеченной из нефтяного шламма / А. Манолов, А. Каныги­ на, Т. Григорьева, Д. Алексеев // “HIGH­THROUGHPUT SEQUENCING IN GENOMICS” Новосибирск, 21–25 июля 2013 года. –– 2013. –– С. 53.
6. *Манолов*, *А.* Поиск генетических маркеров бактерий вида Escherichia coli, ассоциированных с болезнью Крона / А. Манолов, О. Побегуц, Е. Ко­ стрюкова, А. Ларин, Т. Семашко, В. Бабенко, Р. Городничев, Е. Лисицина, П. Щербаков, Е. Ильина, В. Говорун // “V СЪЕЗД БИОХИМИКОВ РОС­ СИИ”, Дагомыс. –– 2016. –– С. 114.

*Манолов Александр Иванович*

Биоинформатический анализ изменчивости генного состава прокариот, в том числе в ассоциации с патогенностью

Автореф. дис. на соискание ученой степени канд. биол. наук

Подписано в печать Заказ №

Формат 60 90/16. Усл. печ. л. 1. Тираж 100 экз.

*×*

Типография