

Федеральное государственное бюджетное учреждение ”Федеральный научно-клинический центр физико-химической медицины Федерального медико-биологического агентства”

На правах рукописи

Манолов Александр Иванович

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

Специальность 1.5.8 —  
«Математическая биология, биоинформатика»

Диссертация на соискание учёной степени  
кандидата биологических наук

Научный руководитель:  
доктор биологических наук, член-корреспондент РАН  
Ильина Елена Николаевна

Москва — 2021

## Оглавление

	Стр.
<b>Введение . . . . .</b>	<b>5</b>
<b>Глава 1. Обзор литературы . . . . .</b>	<b>9</b>
1.1 Архитектура генома прокариот . . . . .	9
1.1.1 Гены неравномерно расположены на (+) и (-) цепях ДНК . . . . .	10
1.1.2 ГЦ-состав генома . . . . .	11
1.1.3 ГЦ смещение генома . . . . .	12
1.1.4 Chi сайты неравномерно расположены в геноме . . . . .	14
1.1.5 Пространственная укладка генетического материала . . . . .	16
1.2 Горизонтальный перенос генов . . . . .	20
1.2.1 Трансформация . . . . .	21
1.2.2 Мобильные элементы генома . . . . .	22
1.2.3 Факторы, влияющие на горизонтальный перенос генов . . . . .	32
1.3 Биоинформационные методы исследования изменчивости генома . . . . .	34
1.3.1 Методы поиска ортологии . . . . .	34
1.3.2 Методы поиска горизонтально переносимых генов . . . . .	37
1.3.3 Методы визуализации отличий в геномах . . . . .	39
1.4 Применение графов для анализа геномных данных . . . . .	40
1.5 Ассоциативная связь болезни Крона с колонизацией <i>E. coli</i> . . . . .	44
<b>Глава 2. Материалы и методы . . . . .</b>	<b>46</b>
2.0.1 Отбор последовательностей геномов . . . . .	46
2.0.2 Анализ геномной вариабельности . . . . .	47
2.1 Сборка и анализ геномов <i>E. coli</i> от пациентов с болезнью Крона . . . . .	48
2.1.1 Группа пациентов и клинический материал . . . . .	48
2.1.2 Выделение изолятов <i>E. coli</i> . . . . .	48
2.1.3 Экстракция ДНК и геномное секвенирование . . . . .	49
2.1.4 Сборка генома, исправление ошибок в гомополимерных областях . . . . .	49
2.1.5 Сбор внешних данных . . . . .	50
2.1.6 Поиск ортогрупп . . . . .	51
<b>Глава 3. Результаты . . . . .</b>	<b>52</b>

3.1	Разработка и верификация метода оценки изменчивости генома на основе графового представления расположения генов . . . . .	52
3.1.1	Разработка способа представления расположения генов в виде графа . . . . .	52
3.1.2	Разработка алгоритма оценки изменчивости генома на основе графового представления . . . . .	55
3.1.3	Верификация предложенного метода оценки профиля изменчивости генома . . . . .	56
3.2	Исследование применимости метода оценки локальной вариабельности генома . . . . .	60
3.2.1	Зависимость результатов от размера выборки . . . . .	60
3.2.2	Анализ зависимости результатов от качества сборки генома	62
3.3	Применение метода оценки локальной вариабельности генома . . . . .	62
3.3.1	Профиль вариабельности генома <i>E. coli</i> . . . . .	62
3.3.2	Сравнение профилей вариабельности филогрупп <i>E. coli</i> . . . . .	64
3.3.3	Сравнение профилей вариабельности филогрупп у других видов . . . . .	67
3.3.4	Сравнение профилей вариабельности между близкородственными видами . . . . .	71
3.4	Связь между уровнем изменчивости и характеристиками генома . . . . .	74
3.4.1	Связь с распределением сайтов Chi. . . . .	74
3.4.2	Связь с плотностью хромосомных контактов . . . . .	76
3.5	Разработка и применение метода анализа локальной вариабельности при помощи построения подграфов . . . . .	79
3.5.1	Проблема сравнительного анализа расположения генов в большом наборе геномов. . . . .	79
3.5.2	Алгоритм поиска подграфа для анализа участка генома . . . . .	79
3.6	Примеры применения представления порядка чередования генов в виде графа. . . . .	82
3.7	Разработка компьютерного приложения для анализа вариабельности геномов . . . . .	88
3.8	Алгоритмизация подхода выявления оперонов, наличие которых ассоциировано с определенным признаком. . . . .	91

3.8.1	Проблема поиска генетических ассоциаций для субпопуляций бактерий . . . . .	91
3.8.2	Алгоритм поиска генетических ассоциаций . . . . .	91
<b>Глава 4. Обсуждение.</b>	. . . . .	<b>96</b>
<b>Глава 5. Выводы</b>	. . . . .	<b>103</b>
<b>Список литературы</b>	. . . . .	<b>104</b>
<b>Глава 6. Благодарности</b>	. . . . .	<b>126</b>
<b>Глава 7. Финансирование</b>	. . . . .	<b>127</b>

## Введение

Геном прокариот представляет собой сложно организованную структуру. Помимо кодирующих и регуляторных областей, в нем имеется ряд элементов, необходимых для взаимодействия ДНК с молекулярными комплексами, осуществляющими процессы транскрипции, репликации и репарации. Пространственная укладка генетического материала в клетке не случайна и выполняет ряд регуляторных функций. Подобные наблюдения меняют представление о геноме, как о простом хранилище последовательностей генов расположенных в случайному порядке, и позволяют говорить об архитектуре генома — закономерностях, которые необходимы для успешного функционирования живой клетки.

К настоящему времени известен ряд элементов геномной организации. Гены, продукты которых необходимы клетке в больших количествах, расположены рядом с сайтом начала репликации, поскольку в быстро делящихся клетках такое расположение позволяет повысить уровень их экспрессии за счет увеличения копийности матричной ДНК. Пространственная укладка ДНК может сближать гены, расположенные в разных областях линейной последовательности, что оказывается полезно для генов, кодирующих регулятор и его мишени. Экспериментально было установлено, что действие глобальных регуляторов, таких как гистоноподобный белок H-NS, зависит от местоположения генов мишеней. Склонность к транскрипции (уровень экспрессии генов, не зависящий от их последовательности) значительно меняется в зависимости от положения гена в хромосоме. Взаимодействие РНК-полимераз, 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" ограничений, и таким образом более толерантны к изменениям. Возможно, эти участки имеют некоторые признаки, способствующие более высокой частоте происходящих 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, ре-

лизуются в действительности, к настоящему моменту, неизвестно: локализация "тихих" консервативных участков и "горячих" высокоизменчивых областей не имеет общепринятых объяснений. Для проведения исследований в данной области 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 прокариотических геномов 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 генома, основанный 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 отдельным локусам генома.
5. Разработать алгоритм поиска и выявить в геноме *E. coli* опероны, которые 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, был первым предложенным и реализованным методом для количественной оценки изменчивости генома.

Насколько нам известно, мы впервые провели сравнительный анализ расположения областей повышенной изменчивости. Мы обнаружили, что некоторые локусы генома остаются высокоизменчивыми у всех представителей вида, в то время, как ряд других локусов являются таковыми лишь у некоторых видов и филогрупп.

### **Практическая значимость**

Изменчивость генома — важный фактор в возникновении патогенных штаммов бактерий и приобретении устойчивости к антибиотикам. Знание закономерностей подобных изменений важно для разработки оптимальных методов

контроля над появлением штаммов бактерий, угрожающих жизни и здоровью людей. Возможно, полученные знания о закономерностях изменчивости и консервативности 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 поиск областей генома с повышенной изменчивостью.
2. Визуализация графового представления позволяет компактно представлять сравнение больших выборок геномов (порядка сотен и тысяч геномов).
3. Геномы представителей различных филогрупп и филогенетически близких 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 чаще выявляются опероны захвата сорбозы, захвата гемина, утилизации глиоксилата, утилизации пропандиола, синтеза и экспорта капсулных полисахаридов.

**Достоверность** предложенного метода обосновывается результатами компьютерного моделирования. Результаты находятся в соответствии с результатами, полученными другими авторами. Основные результаты работы докладывались на конференциях: "Итоговая научно-практическая конференция ФГБУ ФНКЦ ФХМ ФМБА России" (18-19 декабря 2019 года, Москва), "ПОСТГЕНОМ 2018" (29 октября - 2 ноября, Казань), "Биотехнология: состояние и перспективы развития" (20–22 февраля 2017 года, Москва, ), "Высокопроизводительное секвенирование в геномике" (Новосибирск, 18–23 июня 2017 года), "4th World Congress on Targeting Microbiota" (17-19 октября 2016, Париж).

**Личный вклад.** Автором были предложены подходы графового представления 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 последовательностей (исправлении ошибок в гомополимерных областях, поиска контаминаций 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 комменсальными

штаммами. Проведен анализ расположения областей повышенной изменчивости у различных видов и внутривидовых структурах прокариот.

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

**Объем и структура работы.** Диссертация состоит из введения, 7 глав, заключения и 0 приложен. Полный объём диссертации составляет 127 страниц, включая 45 рисунков и 2 таблицы. Список литературы содержит 231 наименование.

## Глава 1. Обзор литературы

### 1.1 Архитектура генома прокариот

Понятие геном было введено немецким ботаником Гансом Винклером в 1920-м году для обозначения гаплоидного набора хромосом в ядре эукариотической клетки [1]. Считается, что слово было образовано за счет объединения терминов "ген" и "хромосома". Сейчас, у прокариот под геномом понимают генетическую информацию, которая находится в фрагментах ДНК, способных к репликации — репликонах — хромосоме (одной или нескольких), и в плазмидах (при их наличии).

Как правило, геном прокариот представлен одной кольцевой хромосомой; также в клетке может находиться одна либо несколько плазмид. Есть и исключения из этого правила. Существуют бактериальные виды с несколькими хромосомами, например, у холерного вибриона (*Vibrio cholerae*) есть две хромосомы [2]. Необычно организован геном бактерии *Borrelia burgdorferi* — возбудителя болезни Лайма — в ее клетках присутствует множество (до одиннадцати) копий линейной хромосомы, диффузно распределенных по клетке [3].

В данной работе под архитектурой (либо закономерностями) организации генома мы будем понимать всё, что отличает геном от простого набора генов.

Некоторые закономерности организации бактериальных геномов были известны еще до появления методов секвенирования [4]. Было известно, что размеры геномов могут значительно различаться в пределах одного вида [5], что ГЦ-состав (доля гуанина (Г) и цитозина (Ц) в нуклеотидной последовательности генома) хромосомы, но может значительно различаться между видами [6], что гены расположены почти вплотную друг к другу, а межгенные области занимают незначительную часть генома [7], что порядок генов у близкородственных видов в значительной степени сохраняется [8]. Также было известно, что геномы подвержены изменению за счет вставок, дупликаций, инверсий и транслокаций, частично обусловленных мобильными элементами генома [9]. Имелись сведения об организации бактериальной хромосомы в макродомены, ассоциированные со стартом и концом репликации [10]. Все же, основная часть информации и значительная часть подробных сведений об организации геномов стала известна лишь

с ростом числа прочитанных последовательностей геномов различных микроорганизмов.

Ниже приводится краткий обзор ряда элементов архитектурных элементов генома.

## **Оперон как функциональная единица архитектуры генома**

Образующаяся при транскрипции молекула мРНК может содержать не один, но несколько генов. Группы генов, имеющих единый промотор и попадающих на одну мРНК, называют опероном. Часто гены из одного оперона кодируют метаболически связанные ферменты [11] либо белки составляющие один белковый комплекс [12].

Концепция оперонов — совместно экспрессируемых генов - была предложена Жакобом и Моно в 1960 году [13]. По их исходному предположению, основная роль оперонов — обеспечение одновременной экспрессии функционально связанных генов [14]. Если бы у отдельных генов, которые должны экспрессироваться совместно, были отдельные регуляторы — это было бы избыточно и ненадежно, случайные мутации в одной из регуляторных областей могли бы привести к рассогласованию их экспрессии. Иная гипотеза была предложена Лоренцом и Ротом [15], согласно которой близкое расположение генов в оперонах необходимо для их совместного переноса в другой организм, при участии данного фрагмента ДНК в горизонтальном переносе генов. Данное предположение может объяснить наличие супер-оперонов, устойчивых (встречающихся во множестве организмов) комбинаций оперонов [16].

### **1.1.1 Гены неравномерно расположены на (+) и (-) цепях ДНК**

Процесс репликации ДНК у прокариот требует наличия ряда нуклеотидных мотивов, неравномерно представленных вдоль репликона [17]. У *Escherichia coli*, как и у большинства бактерий, репликация хромосомы начинается в специфичном локусе (*ori*). С этим участком связывается белок DnaA, который участвует в

сборке крупной молекулярной машины — реплисомы, осуществляющей репликацию ДНК. Репликация происходит одновременно в двух направлениях: вдоль левой и правой реплихоры — частях хромосомы, расположенных по разные стороны от ori и заканчивающихся в месте окончания репликации — ter регионе. Для расцепления двух хромосом, после репликации, необходимо участие белков FtsK, которые связываются с dif сайтами, расположенными поблизости от ter региона [18].

Репликация бактериальной хромосомы может происходить достаточно быстро, примерно 42 минуты в случае *E. coli*. В случае благоприятных условий она происходит достаточно часто, у кишечной палочки новые раунды могут запускаться каждые 20 минут. Во время репликации в клетке продолжаются процессы жизнедеятельности, в том числе — транскрипция генов. При этом реплисома и РНК полимераза могут сталкиваться между собой. Подобные коллизии могут приводить к замораживанию процесса репликации, преждевременного обрыва транскрипции, образованию разрывов в ДНК, появлению мутаций [19]. Вероятность этих нежелательных последствий выше в случае "лобовых" столкновений, и ниже — в случае столкновений при сонаправленном движении [19]. Этим можно объяснить значительно большее количество генов, расположенных на лидирующей цепи ДНК, поскольку при такой ориентации генов, процессы репликации и транскрипции оказываются сонаправленными [20]. В исследовании 725 геномов [21] было обнаружено, что неравномерное расположение генов по цепям ДНК более выражено у более быстрорастущих организмов. Также были обнаружены некоторые функциональные различия, так рибосомальные гены чаще располагаются на лидирующей цепи, в то время как транскрипционные факторы имели обратную тенденцию.

### 1.1.2 ГЦ-состав генома

Различные виды прокариот обладают характерным ГЦ-составом. ГЦ состав использовался в таксономическом анализе еще до развития технологий секвенирования. Причины постоянства ГЦ-состава вдоль хромосомы не совсем ясны. Существует зависимость между ГЦ-составом участка ДНК и температурой его

плавления, но в большинстве исследований не удается обнаружить связь ГЦ-состав генома с оптимальной температурой роста микроорганизмов [22; 23].

Обнаружена связь между ГЦ-составом и наличием кислородного метаболизма: у аэробов более высокий ГЦ-состав [24]. В недавней работе был проведен анализ зависимости ГЦ-состава от наличия кислородного метаболизма с учетом поправки на филогенетическую близость сравниваемых организмов (что не учитывалось ранее). Используя метод филогенетического регрессионного анализа, авторы показали, что ГЦ-состав, действительно, значимо повышен у obligатных аэробных организмов [25].

По версии, предложенной Жаном Лобри [26], такая зависимость связана с тем, что аминокислоты, на синтез которых в аэробных условиях необходимо затрачивать меньше энергии, кодируются кодонами с более высоким содержанием Г и Ц (рис. 1.1).

Анализ лабораторных экспериментов по накоплению мутаций выявил значительное мутационное смещение в сторону увеличения ГЦ-состава [27]. Причем оно наблюдалось как в четырежды вырожденных сайтах (все замены являются синонимичными), так и в ноль-вырожденных сайтах (все замены являются несинонимичными), хотя в последних и в значительно меньшей степени.

### 1.1.3 ГЦ смещение генома

Неравномерная представленность нуклеотидов в лидирующей и отстающей цепи была обнаружена в 60-х годах двадцатого века [28]. Анализ первых трех секвенированных геномов показал, что лидирующая цепь почти равномерно обогащена нуклеотидами Г и Т, в то время как в отстающей чаще встречаются А и Ц [29]. Позднее, данная закономерность, названная ГЦ смещением (GC skew) была выявлена у эубактерий и архей, геномов митохондрий, хлоропластов и вирусов, но не для эукариот со множественными сайтами инициации репликации [30]. Существуют бактерии без выраженного ГЦ смещения, например, некоторые цианобактерии: *Gloeobacter violaceus* и *Synechococcus elongatus* [31]. Основную роль в возникновении ГЦ смещения играет, по-видимому, различная мутационная нагрузка на две цепи ДНК, действующая во время репликации. При репликации, лидирующая и отстающая (достраивающаяся за счет фрагментов Оказаки) цепи

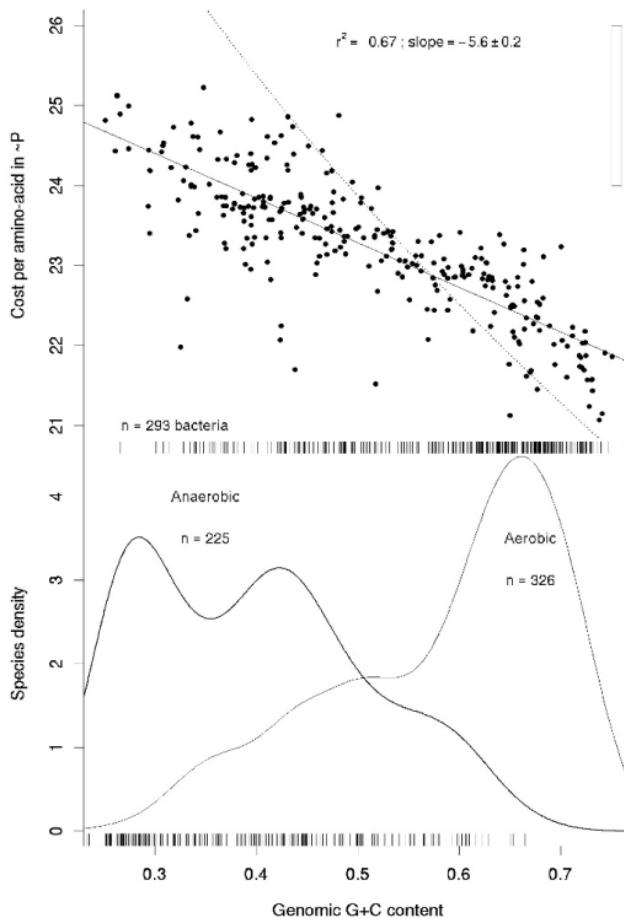


Рисунок 1.1 — Аэробные бактерии имеют более высокий ГЦ-состав и используют аминокислоты, требующие меньше энергозатрат для синтеза в аэробных условиях. График сверху: зависимость средней энергии, которую нужно затратить на синтез аминокислот представленных в протеоме бактерий, от ГЦ-состава. График снизу: плотность распределения ГЦ-состава у аэробных и анаэробных микроорганизмов. Источник изображения: [26].

проводят неравное время в одноцепочечном состоянии и в различной степени подвержены дезаминированию, с последующей заменой цитозиновых нуклеотидов на тиминовые. Данные в пользу этого предположения были получены в лабораторных экспериментах по ускоренной эволюции, в которых было показано влияние уровня дезаминирования цитозина (которое, контролировалось в эксперименте) на итоговый уровень ГЦ смешения [32].

### 1.1.4 Chi сайты неравномерно расположены в геноме

Для сохранения жизни необходима стабильность генома. Для колонизации новых экологических ниш, борьбы с конкурентами, создания более высокоорганизованных организмов, наоборот, нужна динамичность генома - возможность появления и сохранения изменений. Как в обеспечении постоянства генома, так и в его изменчивости, важную роль играет процесс рекомбинации - обмен молекул ДНК своими фрагментами.

Хорошо изучен процесс гомологичной рекомбинации, под которой понимают процесс образования контакта между идентичными или схожими по последовательности фрагментами ДНК с последующим обменом генетическим материалом. Данный процесс позволяет клеткам восстанавливать поврежденные участки ДНК, при условии наличия неповрежденной копии [33]. Гомологичная рекомбинация также задействована в возобновлении работы остановившихся вилок репликации (например, в следствии недостаточно эффективной работы хеликазы, слишком высокой скрученности ДНК в области репликации, значительного количества белков связанных с ДНК, столкновении репликазы и РНК полимеразы [19; 34]).

У прокариот гомологичная рекомбинация играет большую роль в пластичности генома. Именно благодаря этому процессу осуществляются дупликации либо делеции участков ДНК при репликации либо транскрипции, а также замена фрагментов генома на чужеродный материал при горизонтальном перенос генов [35].

В случае, когда в рекомбинации участвуют чужеродные фрагменты ДНК, выделяют два возможных исхода (рисунок 1.2). В одном случае, у двух фрагментов имеются схожие последовательности на краях, и в результате один фрагмент заменяется на другой (рисунок 1.2). Если же привходящая ДНК находится в кольцевой форме и обладает одним схожим по последовательности фрагментом, то происходит вставка без вырезания.

У бактерий процесс рекомбинации хорошо изучен на примере кишечной палочки *E. coli*. Важную роль в рекомбинации у данного организма играет комплекс белков RecBCD, обладающий хеликазной, экзо- и эндонуклеазной активностями. Работа этих белков создает ДНК с "липким" концом - краевым одноцепочечным

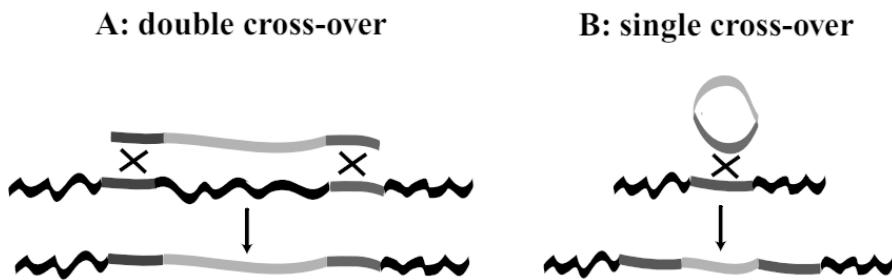


Рисунок 1.2 — Варианты рекомбинации при участии чужеродного фрагмента ДНК. Схожие по последовательности фрагменты обозначены темно-серым цветом. А: в случае наличия схожих фрагментов на краях молекул, одна последовательность заменяется на другую; В: в случае, когда привходящая ДНК находится в кольцевой форме и обладает одним участком с высоким сходством последовательности, происходит вставка нового материала без потери исходного варианта. Источник изображения: [36].

фрагментом, что является необходимым условием дальнейших этапов рекомбинации.

Для образования липких концов необходимо наличие несимметричного разрыва двухцепочечной ДНК. При этом липкие концы образуются на некотором расстоянии от места разрыва. Это расстояние определяется локализацией специфичных последовательностей - Chi сайтов. Данные сайты распознаются белками комплекса RecBCD и, в зависимости от соотношения концентрации АТФ и ионов магния, происходит либо надрез цепи в зоне Chi сайта, либо завершается экзонуклеазная активность RecBCD комплекса (рисунок 1.3) [37; 38]. В дальнейшем, на одноцепочечном фрагменте происходит кластеризация белка RecA, образование синапса, переброс цепей, образование и разрешение структуры Холидея.

Название сайта Chi является сокращением от crossover hot spot instigator (инициатор горячих точек перекреста). Данные сайты были обнаружены как последовательность, необходимая для осуществления RecBCD зависимой рекомбинации фага лямбда [40]. В геноме *E. coli*, находится около тысячи Chi сайтов (примерно по одному на 5 тысяч п.н.) [41].

Chi сайты распределены в геноме не равномерно. Они представлены преимущественно в общей для вида части генома, в кор-геноме (core-genome), что вероятно способствует поддержанию его стабильности [42]. В то же время, рибосомные опероны не содержат в себе Chi сайтов (но содержали бы при их случайном расположении) [43]. Наблюдается более редкая встречаемость Chi сай-

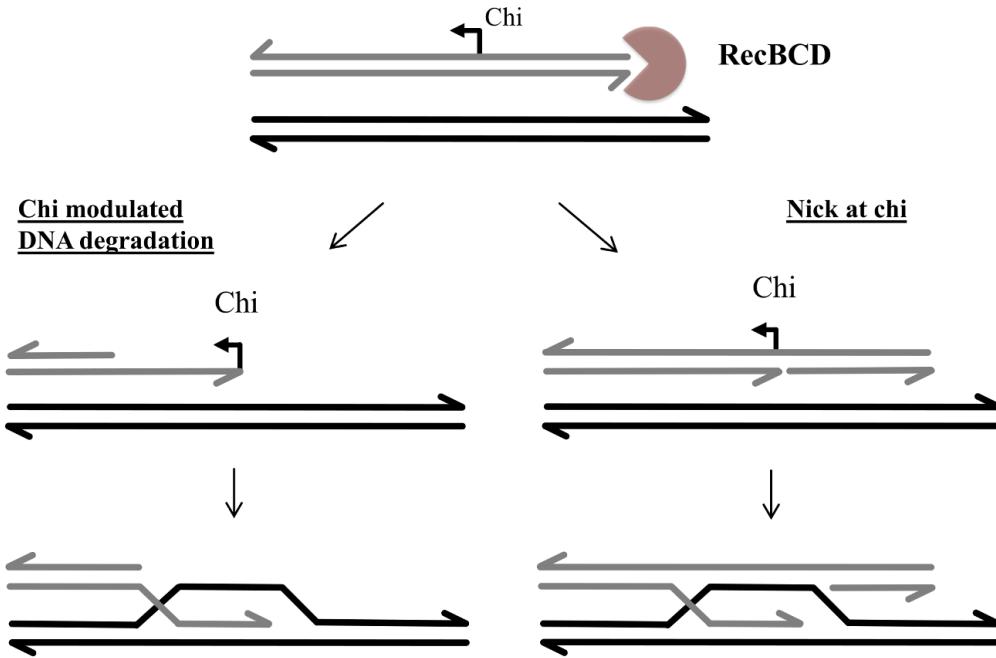


Рисунок 1.3 — Участие Chi сайтов в процессе рекомбинации. Данная последовательность распознается белками RecBCD комплекса, что приводит к прекращению продвижения комплекса вдоль цепи ДНК и задает локализацию двухцепочечного разрыва. Источник изображения: [39].

тов в повторяющихся участках генома, что, по-видимому, защищает геном от лишних перестроек, которые бы возникали в случае близкорасположенных повторов [44].

### 1.1.5 Пространственная укладка генетического материала

У всех организмов ДНК находится в свернутом состоянии, что необходимо для ее размещения внутри клетки (длина хромосомы на несколько порядков превышает длину клетки) [45]. При этом, пространственная конфигурация молекулы ДНК не случайна, она отражает и регулирует функциональное состояние генома [46]. Так, для эукариот было показано, что петли хроматина, которые размещают промоторы и удаленные энхансеры в непосредственной пространственной близости, играют важную роль в регуляции транскрипции [47]. У прокариот, укладка хромосомы также, по-видимому, неслучайна и выполняет

ряд функций [48]. Подходы, основанные на флуоресцентной микроскопии, позволяют определять субклеточное положение отдельных хромосомных локусов, а высокопроизводительные подходы на основе методов 3C и Hi-C позволяют количественно определять частоты взаимодействия между локусами, которые впоследствии могут использоваться для определения средних трехмерных расстояний между ними. Автоматизация этих методов в начале 2000-х годов позволила проводить исследования пространственной укладки хромосом в масштабе всего генома [49].

Укладка хромосом у прокариот имеет иерархический характер: от крупномасштабных макродоменов до более мелких структур. Она контролируется при помощи нуклеоид-ассоциированных белков [45], одним из которых является гистоноподобный белок H-NS. Он прикрепляется к ДНК преимущественно в локусах с повышенным АТ составом и высокой частотой АрТ динуклеотида, что характерно для горизонтально перенесенных фрагментов генома [50]. Склонность белка H-NS к образованию цепочек стимулирует образование протяженных нуклеопротеиновых филаментов вдоль одной либо между двух спиралей ДНК [51]. Считается, что у *E. coli* основными функциями данного белка являются компактизация ДНК и снижение уровня экспрессии горизонтально перенесенных генов [45]. Множество бактерий имеют гомологи, либо аналоги белка H-NS.

Другим важным способом поддержания пространственной укладки ДНК являются комплексы структурного поддержания хромосом (structural maintenance of chromosomes, SMC). Эти комплексы способствуют образованию петель ДНК и поддерживают их устойчивость, формируя кольцеобразную структуру вокруг петель [52]. Комплексы SMC также участвуют в сегрегации вновь реплицированных сестринских хромосом [45]. Мутанты *E. coli* по гену *muC* (один из основных белков SMC комплекса у данного организма) имеют нарушенное разделение ДНК по дочерним клеткам и часто производят клетки, лишенные хромосомы [53]. Ряд других белков (IHF, HU, Fis) участвуют в формировании пространственной укладки за счет изгибаания ДНК и поддержания её отрицательной суперскрученности [45].

В масштабе от десятков до сотен тысяч оснований, бактериальная хромосома разделена на домены взаимодействия хромосом (chromosome interaction domains, CID), которые аналогичны топологически ассоциированным доменам (topologically associating domains, TAD) у эукариот. ДНК расположена значитель-

но ближе и чаще контактирует внутри доменов, чем между ними (рисунок 1.4) [45].

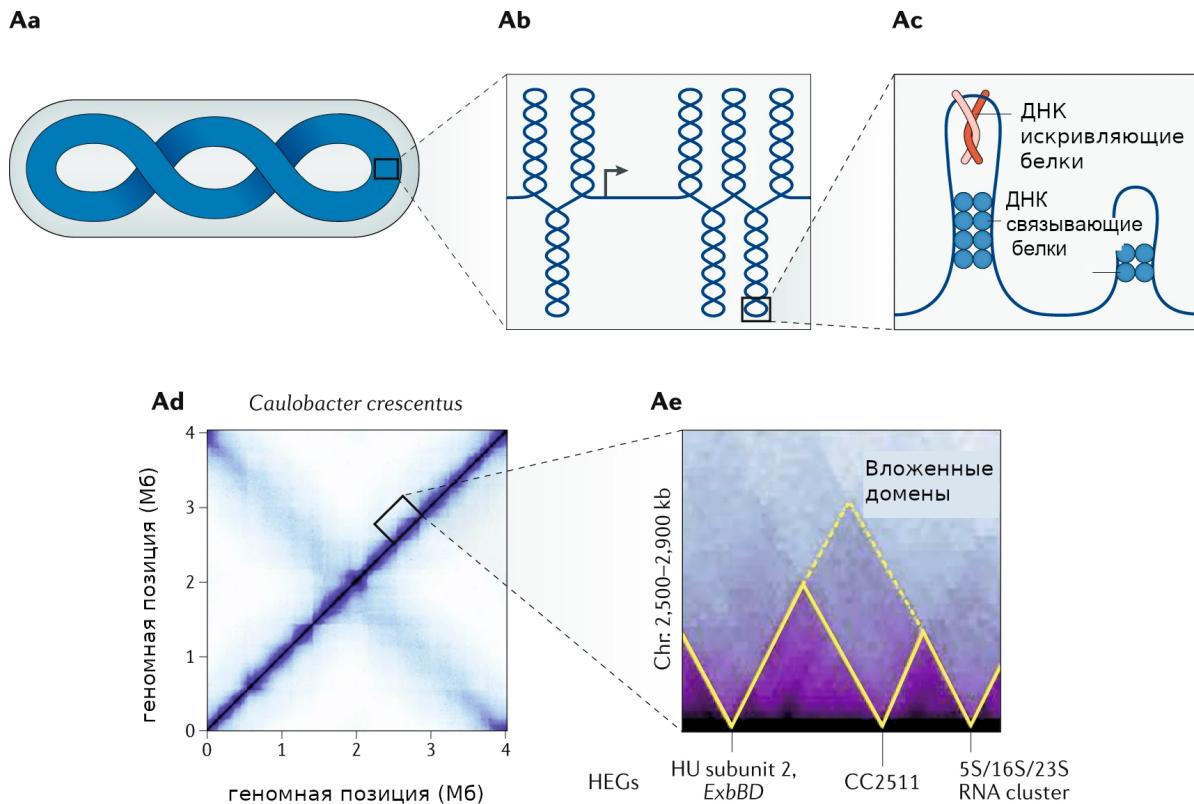


Рисунок 1.4 — Уровни организации пространственной укладки ДНК у прокариот. Пространственная укладка хромосомы многих бактерий имеет форму спирали (Аа), что выражается в наличии двух диагоналей на карте хромосомных контактов (Ад). В масштабе от десятков до сотен тысяч оснований хромосома подразделяется на домены взаимодействия хромосом (Аб). Эти структуры выглядят как квадраты вдоль главной диагонали карты хромосомных контактов (Ад) или как треугольники при наблюдении одной половины симметричной карты (Ае). Домены взаимодействия часто являются вложенными: более крупные домены (прерывистая желтая линия) организованы в более мелкие субдомены (сплошная желтая линия) (Ае). Границы между доменами обычно образованы высокоэкспрессируемыми генами длиной более 2 т.п.н., которые физически разделяют flankирующий хроматин (часть Аб). Изображение адаптировано из [45].

Количество доменов может зависеть от состояния клетки, так у *Caulobacter crescentus* хромосома организована в 23 домена взаимодействия во время экспоненциального роста в богатой среде и в 29 доменов в условиях голодаания; изменение количества доменов вероятно связано с изменением уровня транскрипции генов [45]. В хромосоме *E. coli* можно выделить 31 домен взаимодействия

размером от 40 до 300 тысяч п.н. Двадцать две границы домена соответствуют положениям высоко экспрессируемых генов, а девять границ совпадают с положениями генов, кодирующих белки с сигнальной последовательностью экспорта [54]. Выделенное положение генов, кодирующих экспортируемые из клетки белки, может объясняться необходимостью сопряжения процессов транскрипции, трансляции и транслокации [55]. Бактериальные домены взаимодействия имеют вложенный (иерархический) характер, каждый домен состоит из меньших субдоменов (рисунок 1.4 Ae), самые мелкие единицы этой организации могут соответствовать отдельным оперонам [45].

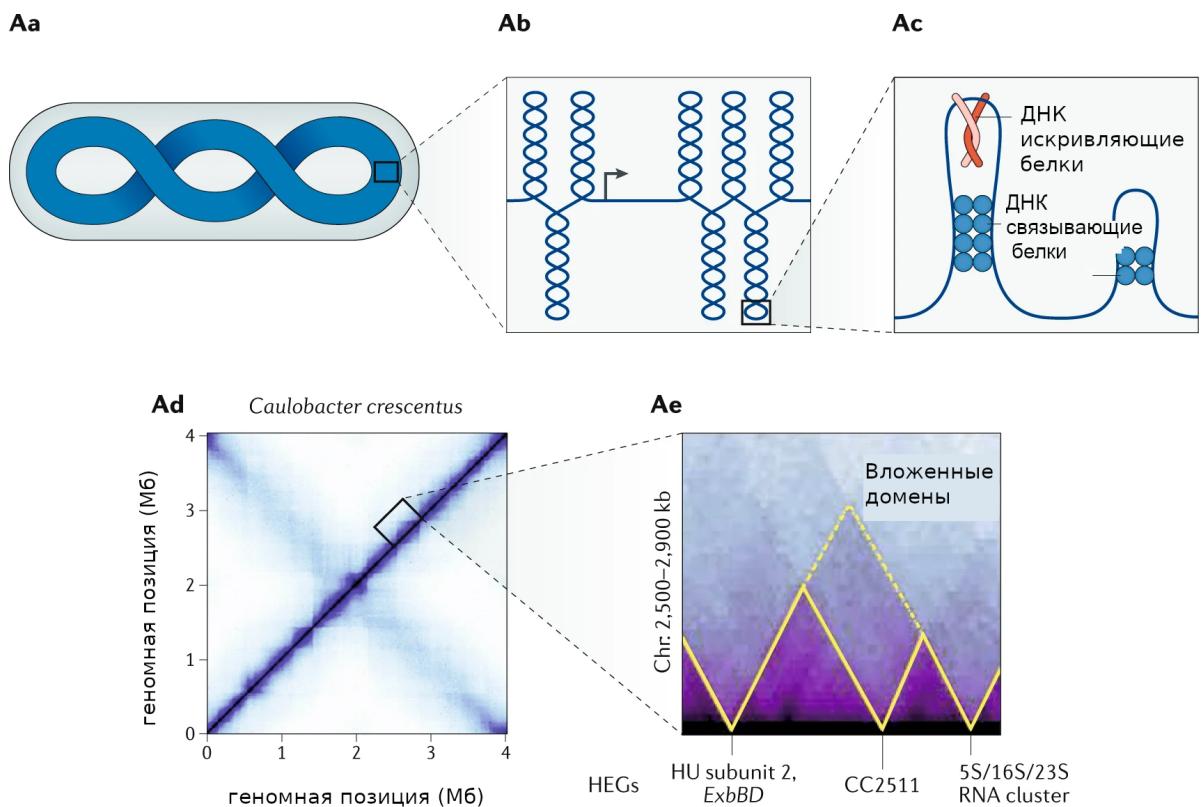


Рисунок 1.5 — Пространственная укладка хромосомы *E. coli*. А) Схематическое изображение пространственной укладки хромосомы *E. coli*, в которой выделяют следующие макродомены: левый (Left), правый (Right), макродомен около сайта начала репликации (Ori) и места окончания репликации (Ter). Б) карта межхромосомных контактов в экспоненциальной фазе роста. В) карта межхромосомных контактов в стационарной фазе роста. Изображение адаптировано из [45].

Наиболее крупным уровнем пространственной организации прокариотической хромосомы являются макродомены. У *E. coli* хромосома разделена на четыре макродомена и две неструктурированные области (рисунок 1.5). Все макродомены обладают пониженной подвижностью внутри клетки по сравнению с

неструктуризованными хромосомными участками. Таким образом, макродомены имеют тенденцию взаимодействовать с неструктурными областями, но не с другими макродоменами [56]. У *E. coli* выделяют макродомен около сайта начала репликации (Ori домен). Ограниченнная подвижность Ori домена требует активности белка MaoP (белок макродомена Ori) и мотива из 17 нуклеотидов в вышележащей межгенной области, названной *taoS* (последовательность макродомена Ori) [57]. Механизм, благодаря которому белок MaoP ограничивает подвижность ДНК остается не известен [45]. Макродомен Ter содержит место окончания репликации, для него описан белок MatP (macrodomain Ter protein), необходимый для существования макродомена (при деактивации данного белка ДНК становится неструктурной). Белок MatP распознает мотив из 12 нуклеотидов, встречающийся преимущественно в Ter регионе.

Макродоменная организация хромосомы более выражена во время экспоненциальной фазы роста бактерий. Переход к стационарной фазе связан со снижением уровня доменной организации ДНК, что наблюдается как "размытие" квадратов, соответствующих доменам, на карте межхромосомных контактов, получаемой в экспериментах на основе метода HI-C [54].

## 1.2 Горизонтальный перенос генов

Бактерии и археи размножаются в основном за счет бинарного деления, которому предшествует дупликация генома в родительской клетке. Дочерние клетки получают геном от родительской клетки, что называется вертикальным наследованием. Существует процесс горизонтального переноса генов (ГПГ), при котором клетки приобретают чужеродный генетический материал. На основе анализа геномов было установлено, что горизонтальный перенос часто встречается у прокариот и вероятно сыграл важную роль в их эволюции. Данный процесс позволяет микробам приобретать новые метаболические возможности, занимать новые экологические ниши, приобретать резистентность к воздействию антибиотиков, фагов, атак со стороны эукариот [35; 58; 59]. Горизонтальный перенос генов внес наибольший вклад в расширение генных семейств (появления различных вариантов гомологичных белков) [60]. В настоящее время интерес к этому процессу во

многом продиктован ростом числа резистентных к антибиотикам бактерий, в том числе, бактерий устойчивых ко всем известным антибиотикам [61].

Горизонтальный перенос генов происходит наиболее часто между близкородственными организмами, что объясняется наличием барьеров для переноса между филогенетически-далекими организмами и зачастую низкой функциональностью генов, приобретенных от несхожих организмов [62]. Тем не менее, описаны случаи переноса крупных фрагментов генома между дальнородственными организмами [63]. Например, у архей часто находят гены, горизонтально перенесенные от бактерий [64]. В недавней работе, был сделан вывод о происхождении гетеротрофных аэробных архей *haloarchaeae* от архей, являющихся аутотрофными анаэробами, в результате переноса большого фрагмента бактериальной ДНК, содержащей около 1000 генов [64]. Описаны отдельные случаи переноса генов между про- и эукариотами [65].

Основные способы горизонтального переноса генов у бактерий и архей таковы: трансдукция — перенос генов при помощи фагов, трансформация — захват ДНК из окружающей среды, конъюгация — проникновение ДНК от клетки-донора при контакте с клеткой-реципиентом [64; 66]. Описаны и более экзотические способы, при помощи ДНК содержащих мембранных везикул [67], передача ДНК при контакте бактерий с помощью синтезируемых ими нанотрубок [68], вирусоподобные агенты горизонтального переноса (virus-like gene transfer agents, GTA) [69]. В большинстве типов переноса, переносимая ДНК должна находиться в одноцепочечной форме, у *E. coli* также описаны механизмы переноса двухцепочечной ДНК [70].

### 1.2.1 Трансформация

Трансформация — это процесс захвата и интеграции внеклеточной ДНК характерный для прокариот. Поглощение ДНК требует, чтобы клетка находилась в физиологическом состоянии, известном как компетентность. Для поддержания состояния компетентности требуется активность нескольких десятков (20–50) белков, которые как правило высоко-консервативны [66; 70]. Интенсивность трансформации зависит от ряда факторов: от концентрации внеклеточной ДНК,

от количества соседних компетентных клеток, степени нехватки ресурсов (стресса) [71].

При трансформации, перемещаемая ДНК становится одноцепочечной при прохождении через мембрану, и затем может подвергаться гомологичной рекомбинации или использоваться в качестве источника питательных веществ [72]. Интересно, что в то время как белки участвующие в транслокации ДНК обладают высокой консервативностью (почти у всех видов эту функцию выполняют гомологичные белки), пути, регулирующие переключение в состояние компетентности, значительно различаются - у разных видов в этом участвуют различные белки, реагирующие на различные сигналы [71]. Это значительно усложняет определение того, является ли некоторый вид естественно-компетентным. Так, *Vibrio cholerae* не считался компетентным до обнаружения роли продуктов распада хитина в индукции состояния компетентности [73]. У некоторых видов (например, различных видов *Neisseria* и *Helicobacter*) клетки находятся в компетентном состоянии почти все время жизни [74].

Захват ДНК в бактериальных клетках осуществляется при помощи пилей IV типа либо иными подобными структурами [75]. Этот процесс наиболее хорошо изучен у грам-отрицательных бактерий, в частности у *Vibrio cholerae* и *Neisseria gonorrhoeae*. У данных бактерий после захвата ДНК, пиль сокращается, ДНК проникает внутрь клетки, после чего связывается с периплазматическими белками для предотвращения обратной диффузии, одна цепь ДНК затем переносится через цитоплазматическую мембрану, а другая цепь деградирует [74]. Подобный механизм, вероятно, действует и у грамположительных бактерий, но эти системы изучены в меньшей степени [74].

### 1.2.2 Мобильные элементы генома

Горизонтальный перенос генов часто осуществляется при участии мобильных элементов генома. Под мобильными элементами генома понимают фрагменты ДНК, которые способны перемещаться внутри генома или между геномами различных клеток и кодируют все либо часть белков, необходимых для своего перемещения [76]. Наиболее известными и хорошо изученными мобильными элементами являются плазмиды, транспозоны и бактериофаги. Ниже

приведено короткое описание этих и некоторых других типов мобильных элементов.

## Плазмиды

Плазмиды — это кольцевые или линейные внехромосомные репликоны, которые присутствуют во многих бактериях и археях и встречаются у эукариот [77]. Они являются одним из основных средств обмена генетической информацией у микробов благодаря способности эффективно перемещаться из одного организма в другой при помощи конъюгации [78].

Длина плазмид находится в диапазоне от единиц до сотен тысяч пар нуклеотидов. Их репликация может осуществляться различными способами. Для кольцевых плазмид характерна тета-репликация, репликация по типу "катящегося кольца" репликация с вытеснением цепи [79]. На краях линейных плазмид, как правило, присутствуют короткие повторы по функциям схожие с теломерами, которые играют важную роль в репликации данных плазмид [80].

Конъюгативные плазмиды несут в себе генетически сложные системы для горизонтальной передачи плазмид, включая белки для образования пор спаривания, а также белки репликации и передачи ДНК. Существуют мобилизуемые плазмиды, которые кодируют только часть функций, необходимых для переноса; их горизонтальная передача может происходить только при наличии в клетки других плазмид, несущих недостающие белки. Соотношение подвижных плазмид (конъюгативных и мобилизуемых) к плазмидам, для которых не обнаружены факторы передачи (немобильные), равно примерно 2:1 [81]. Немобильные плазмиды также могут горизонтально передаваться при помощи процесса трансформации [82].

Недавний анализ более чем десяти тысяч плазмид, показал, что для них можно выделить кластеры ("таксономические единицы") с высоким сходством последовательностей внутри кластеров (выше 90%) и низким, как правило, ниже 70% сходства в остальных случаях [81].

Плазмиды могут встраиваться в хромосому своих хозяев и таким образом обеспечивать изменчивость хромосом [83; 84]. Такие интеграционные события

были обнаружены в геномах *Enterococcus faecalis*, *Shigella flexneri*, *Yersinia pestis*, *E. coli* и ряда других организмов [85].

Встраивание плазмид может сильно влиять на фенотип организма. Для *S. flexneri* и *E. coli* описана плазмида размером 220 т.п.н., встраивание которой в хромосому приводит к тому, что факторы вирулентности, кодируемые данной плазмидой, перестают экспрессироваться и бактерии теряют способность к инвазии. Затем может произойти точное вырезание плазмиды, с восстановлением вирулентности бактерии. Также может происходить неточное (частичное) вырезание, что может приводить к тому, что фрагменты плазмиды остаются в хромосоме [85]. Описан ретро-транспорт плазмид — их передача в клетку-реципиент с последующим возвратом в донорскую клетку. Данный процесс может служить способом приобретения новых генов клеткой-донором, за счет их встраивания в плазмиду, при ее нахождении в клетке-реципиенте [86]. Плазмиды обладают высоким уровнем изменчивости и часто содержат в себе различные мобильные элементы. Конъюгативные и мобилизуемые плазмиды таким образом могут служить средством транспорта мобильных элементов генома от одних клеток к другим. Интересной и сравнительно новой темой исследований является взаимосвязь мобильных элементов генома. Так, для холерного вибриона описана мобилизация геномного острова конъюгативной плазмодой [87]. Согласно предложенной модели, мобилизация обеспечивается тем, что на плазмиде кодируется транскрипционный фактор AcaCD, который активирует в том числе и гены, находящиеся в геномных островах, в частности ген фактора направленности рекомбинации *xis*, продукт которого увеличивает частоту вырезания геномного острова из хромосомы [88].

### Интегративные конъюгативные элементы

Впервые, процесс конъюгации был открыт на плазмidaх [89]. Позднее были обнаружены хромосомные элементы генома способные к вырезанию и конъюгативной передаче — интегративные конъюгативные элементы (ИКЭ)[90]. Вырезание и интеграция ИКЭ осуществляется за счет интеграз либо транспозаз. На рисунке 1.6 показано схематичное изображение структуры и принципа функционирования интегративных конъюгативных элементов.

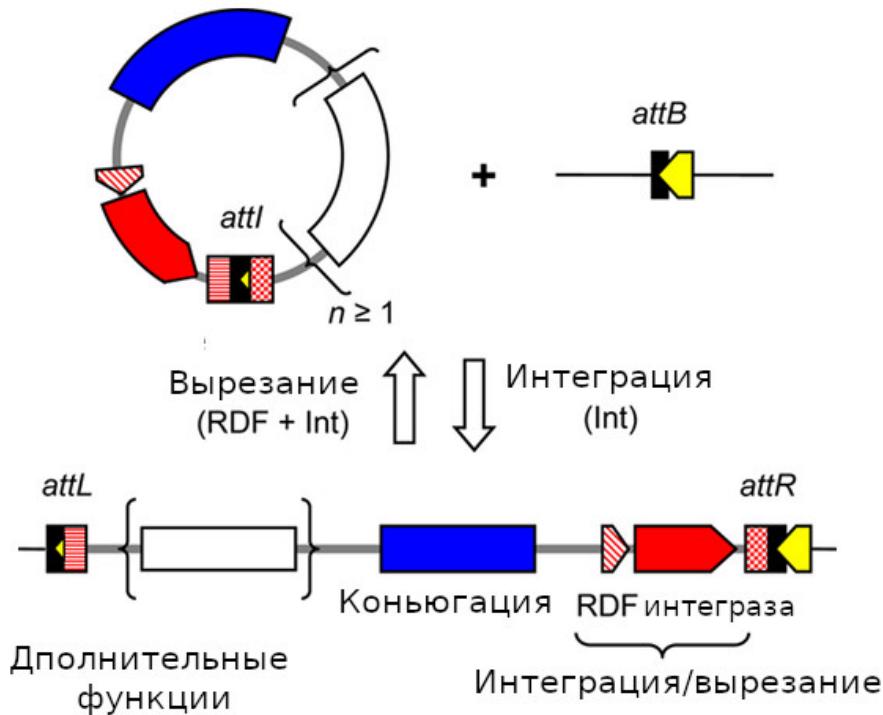


Рисунок 1.6 — Устройство и принцип работы интегративных конъюгативных элементов (ИКЭ). Сайт-специфическая интеграция и вырезание катализируются интегразой. RDF является кофактором, необходимым для процесса вырезания (у ряда интеграз). Большие прямоугольники обозначают модули. Тонкие черные линии и толстые серые линии обозначают геном хозяина и ИКЭ, соответственно. Стрелками обозначены гены, а меленькими прямоугольниками - сайты интеграции. Изображение адаптировано из [76].

Разные типы интеграз и транспозаз отличаются сайтами интеграции и уровнем сайт-специфичности. Наиболее хорошо изучены ИКЭ, кодирующие тирозиновые рекомбиназы. Они часто обладают высокой сайт специфичностью, а сайты интеграции широко представлены в генах тРНК и ряде других генов домашнего хозяйства. Описаны также ИКЭ с низкой сайт-специфичностью интеграции. Например, элемент CTnDOT, встречающийся у бактерий рода *Bacteroides* и встраивается в низко-специфичные (но не случайные) сайты [91].

Интегративные конъюгативные элементы, наряду с плазмидами, считаются одними из наиболее эффективных способов распространения генов устойчивости к антибиотикам [92]. ИКЭ семейства SXT были признаны основными факторами распространения генов устойчивости к антибиотикам среди нескольких видов семейств Enterobacteriaceae и Vibrionaceae, включая экологические и клинические изоляты *V. cholerae* [93].

Интегративные конъюгативные элементы значительно отличаются по длине: от десятков т.п.н. (например, элемент pSAM2 у *Streptomyces ambofaciens* [94]) до сотен т.п.н. (у *Streptomyces turgidiscabies* описан элемент PAISt длиной в 674 т.п.н.[95]). Эта разница в размере в значительной степени зависит от карго-генов, количество которых варьирует от одного гена устойчивости к тетрациклину у Tn916 (одного из наиболее хорошо изученных ИКЭ), до значительного числа генов (в том числе с неизвестными функциями) у PAISt. Состав карго-генов может значительно между близкими ИКЭ, несущими одинаковые либо схожие по последовательности модули конъюгации и рекомбинации.

## Фаги

Фаги - вирусы бактерий и архей - являются важным фактором изменчивости геномов своих хозяев [96]. Первые проекты по секвенированию бактериальных геномов показали значительный вклад профаговых последовательностей в наблюдаемые межштаммовые различия. Так, в последовательности генома *E. coli O157 Sakai* было идентифицировано 18 профагов (или остатков профагов), они составили примерно половину всех различий в генном составе между данным штаммом и лабораторным *K-12* [97].

Размеры фаговых геномов значительно варьируют, наименьший описанный фаговый геном состоит всего из 2,4 т.п.н., а размеры наиболее длинных геномов превышают 400 т.п.н. Подобно остальным мобильным элементам, фаговые геномы могут нести карго-гены, среди которых встречаются гены патогенности (включая, токсины) [98]. Так, некоторые фаги семейства лямбда содержат в себе шига-токсин, который является мощным фактором патогенности у шига-продуцирующих *E. coli*; у *Vibrio cholerae* описаны нитчатые фаги CTXf, несущие ген, кодирующий токсин холеры (CTX); известны фаги, наличие которых необходимо для патогенности бактерий *Corynebacterium diphtheriae* и *Clostridium botulinum* [99]. Ряд факторов патогенности присутствует в фагах золотистого стафилококка, сальмонелл, стрептококков [99]. Экспериментально было показано, что некоторые гены вирулентности, содержащиеся в фагах, могут экспрессироваться при нахождении его в виде профага [96], ряд других генов (например, шига-

токсин фагов семейства лямбда) экспрессируются только когда фаг находится в липитической фазе цикла [100].

Фаги могут участвовать в горизонтальном переносе генов у своих хозяев при помощи процессов специальной и общей (генерализованной) трансдукции [101; 102]. Умеренные фаги могут встраиваться в геном хозяина - становиться профагами. При последующей индукции, происходит вырезание профага из генома хозяина и наработка вирусных частиц, с последующим их выходом из клетки. Вырезание профага может не совпадать с границами вирусной последовательности и содержать соседние области генома. Тогда части ДНК клетки-хозяина попадут в капсид и будут перенесены в новую клетку, заражаемую фагом. Данный процесс называется специализированной трансдукцией. В соответствии с этой моделью, в ряде геномов в непосредственной близости от сайта интеграции фагов обнаруживаются горизонтально перенесенные гены [101].

Хорошо изучен механизм интеграции у фага лямбда. Интеграза данного фага представляет собой сайт-специфическую тирозиновую рекомбиназу, обеспечивающую рекомбинацию между двумя комплементарными последовательностями ДНК: сайтом attP (250 п.н.), расположенный в геноме фага, и сайте attB (21 п.н.), расположенном внутри бактериального генома [103]. У многих других фагов, использующих сайт-специфическую рекомбинацию, для успешной интеграции необходимо наличие вспомогательных белков — факторов интеграции, кодируемых бактериями [103].

Горизонтальный перенос генов может происходить также за счет общей (генерализованной) трансдукции, при которой ДНК хозяина упаковывается в капсид вместо генетического материала фага. После заражения другой клетки, такая ДНК может участвовать в процессе рекомбинации и стать частью генома зараженной клетки. У некоторых бактерий, фаги являются механизмом передачи определенных, неслучайных, фрагментов ДНК. Так, у *Staphylococcus aureus* описан остров патогенности SaPI1 длиной 15 т.п.н., кодирующий токсин Tst участвующий в токсическом шоке. В клетках стафилококка, инфицированных фагом 80a, этот остров вырезается из хромосомы, автономно реплицируется и попадает в капсид собираемых фагов. При проникновении в организм-реципиент, SaPI1 интегрируется с использованием собственных интеграз [104].

Фаговые геномы обладают высоким уровнем изменчивости, а их отдельные фрагменты часто имеют различную эволюционную историю (мозаицизм) [105]. Вероятно, основным фактором такой изменчивости, является незаконная

рекомбинация или рекомбинация между короткими консервативными последовательностями; значительный вклад в эти процессы могут вносить фаговые рекомбиназы [105; 106]. Особенностью горизонтальной передачи генов при помощи фагов является их устойчивость: фаги могут сохранять свою способность к инфицированию на протяжении многих лет, даже в агрессивной внешней среде, в которой свободная ДНК деградирует. Фаги, как правило, обладают более узким спектром хозяев (по сравнению с плазмидами), что вероятно объясняется их зависимостью от наличия специфических рецепторов на инфицируемых клетках [85].

## Интегроны

Интегроны — это специализированные системы приобретения генов в бактериальных геномах [107]. Они достаточно широко распространены и встречаются примерно в 10%-20% прочитанных геномах [108; 109]. Как и ряд других мобильных элементов, они могут участвовать в приобретении, экспрессии и распространении генов устойчивости к антибиотикам [107] и факторов вирулентности [110]. Интегроны могут быть закодированы в хромосоме, либо находиться в составе плазмид и транспозонов [108].

Интегроны содержат ген интегразы (*intI*), сайт рекомбинации (*attI*) и промоторные области, необходимые для экспрессии гена интегразы и транскрипции перенесенных в интегрон генов (рисунок 1.7). Эта структура служит платформой для сайт-специфической интеграции нового генетического материала — генных кассет. Кассеты - кольцевые фрагменты ДНК, как правило состоящие из одного или нескольких генов, лишенных промоторной области. Кассеты также содержат сайт рекомбинации (*attC*) [107]. После встраивания, генная кассета становится функциональным геном. Уровень экспрессии встроенных кассет тем выше, чем ближе они расположены к промотору *Pc* (рисунок 1.7). Интеграза также может случайным образом вырезать генные кассеты и реинтегрировать их на прежнее, либо новое место в интегроне. Таким образом может происходить перестановка генных кассет, с изменением уровня экспрессии встроенных генов. Такая перестановка может служить механизмом подобным простому типу памяти. Кассеты, которые были полезны ранее могут перемещаться в более отдаленные от промотора участки интегрона (например, за счет встраивания новых кассет) что приводит

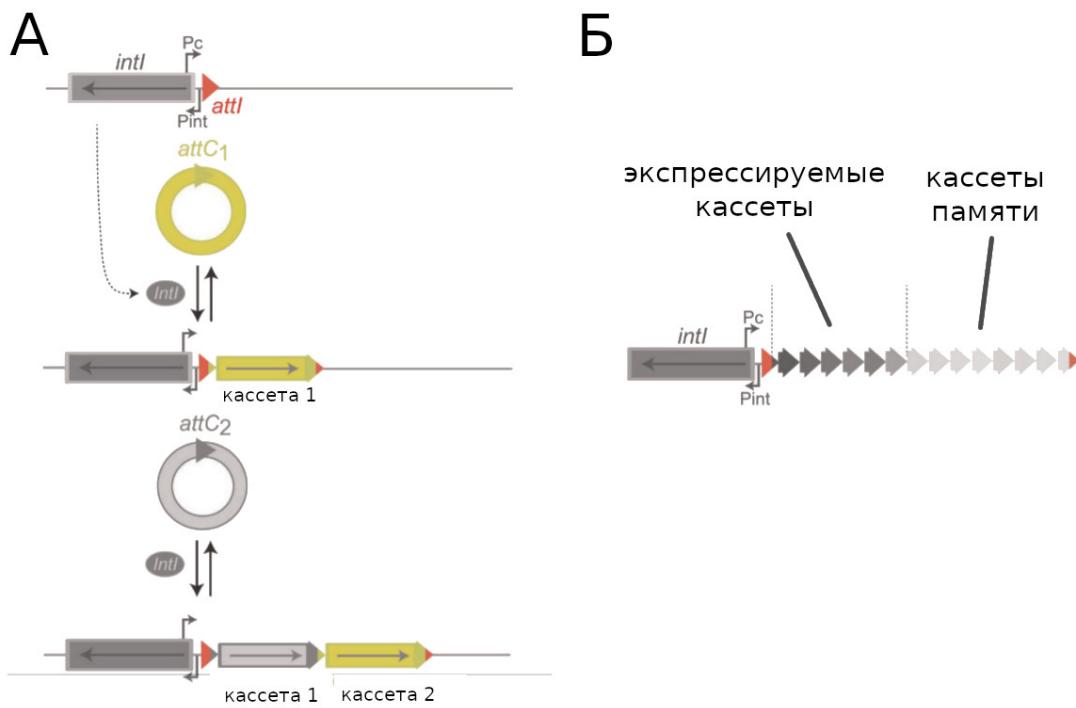


Рисунок 1.7 — Устройство и принцип работы оперонов. А) Вставка и вырезание кассет. Интегрон содержит функциональную платформу, состоящую из гена интегразы (*intI1*), промотора интегразы (*Pint*), промотора кассеты (*Pc*) и сайта рекомбинации *attI*. Б) Уровень экспрессии кассет (показан градациями серого цвета) падает по мере удаленности от промотора, только первые из кассет экспрессируются, остальные можно отнести к кассетам “памяти”. Изображение адаптировано из [111].

к снижению или прекращению их экспрессии, но не удалению из генома. При последующем возврате к прежним условиям внешней среды, эти кассеты могут подвергнуться процессу перестановки и оказаться вблизи промотора с последующей наработкой с них белков.

### Геномные острова

Понятие геномного острова не определено однозначно. Островами могут называть любые фрагменты ДНК, которые попали в геном в результате горизонтального переноса [112; 113]. В других работах, под ними понимают такие горизонтально переносимые фрагменты генома, которые содержат в себе ”ре-

комбинационный модуль включающий в себя интегразу и иные вспомогательные факторы (рисунок 1.8) [111]. Филогенетический анализ показал, что интегразы, встречающиеся в геномных островах, значительно отличаются от интеграз иных известных мобильных элементов (фагов, интегронов, IS-элементов, интегративных конъюгативных элементов), что может говорить о том, что подобные геномные острова - отдельный вид мобильных элементов генома [111]. Для ряда геномных островов наблюдалось их вырезание из хромосомы и существование в виде отдельной кольцевой ДНК [114; 115]. После вырезания, острова могут быть реинтегрированы в хромосому, деградировать либо попасть в новую клетку при помощи общей трансдукции, трансформации либо за счет встраивания в другие мобильные элементы [111; 116; 117]. На частоту вырезания влияет активность экспрессии рекомбиназ и – для ряда организмов – вспомогательных факторов рекомбинации [117].

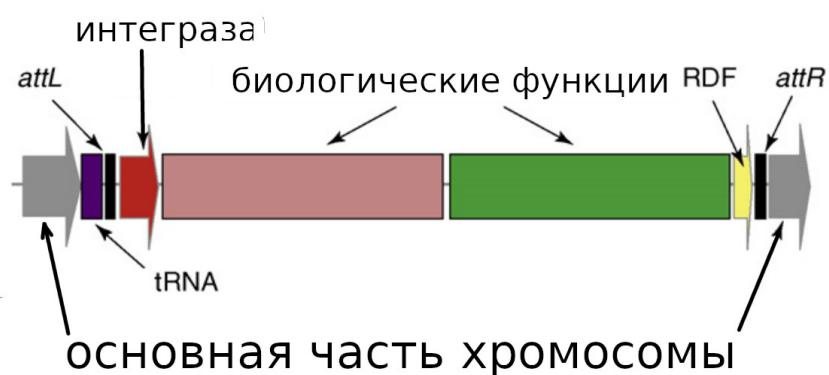


Рисунок 1.8 — Схематическое изображение основных компонентов геномного острова. Серые стрелки - основные (коровые) хромосомные гены, фиолетовый прямоугольник представляет - ген тРНК, черные прямоугольники - сайты прикрепления (attL и attR), красная стрелка - ген интегразы, желтая стрелка - фактор направленности рекомбинации (RDF, recombination directionality factor, обнаружены у ряда островов), другие цветные прямоугольники представляют - гены с различными биологическими функциями. Изображение адаптировано из [111].

Большинство геномных островов, исследованных в работе [111], были встроены в гены тРНК, что по мнению авторов обусловлено сайт-специфичностью интеграз и нахождением соответствующих сайтов в генах тРНК. Некоторые геномные острова из рода *Vibrio*, были интегрированы в гены транспортно-матричной РНК (тмРНК).

Геномные острова отличаются от интегративных конъюгативных элементов тем, что не несут в себе последовательности генов, необходимых для конъюгации, и могут не содержать в своем составе интеграз. Предполагается, что некоторые геномные острова могут передаваться горизонтально за счет использования белков, закодированных в других мобильных элементах (плазмидах, фагах, интегративных конъюгативных элементов) [111].

## IS-элементы и транспозоны

Последовательности вставки или IS-элементы (insertion sequences, IS) являются одними из самых простых мобильных генетических элементов [118]. Они представляют из себя короткие (около 1-3 т.п.н.) сегменты ДНК, кодирующие один-два гена и flankированные инвертированными повторами (рисунок 1.9А). IS элементы способны к вставке во множество разных мест в. IS-элементы содержат гены транспозазы - белка катализирующего разрезание ДНК и обмен цепями; в некоторых элементах присутствуют регуляторные белки, влияющие на активность транспозазы [119]. IS-элементы могут перемещаться как внутри генома, так и передаваться между организмами, находясь в составе фагов, плазмид, интегративных конъюгативных элементов [118]. Описаны частично деградированные IS-элементы, которые способны к перемещению по геному за счет активности других, интактных, IS-элементов. Такие элементы называются MITE-элементами (miniature inverted repeat transposable elements, MITE) [119; 120].

IS-элементы играют существенную роль в вариабельности генома [121]. Они могут накапливаться в значительном количестве в геноме, а события рекомбинации между ними часто приводят к удалению геномных фрагментов, что, по-видимому, играет важную роль в уменьшении размера генома видов, перешедших к паразитическому образу жизни [119]. Встраивание данных элементов в гены, может приводить к потере их функциональности, что в ряде случаев может увеличивать приспособленность организма, например в следствии изменения антигенных детерминант у патогенных микроорганизмов [122]. IS-элементы могут увеличивать экспрессию генов за счет образования гибридных промоторов, частично состоящих из фрагментов последовательности IS-элемента [123].

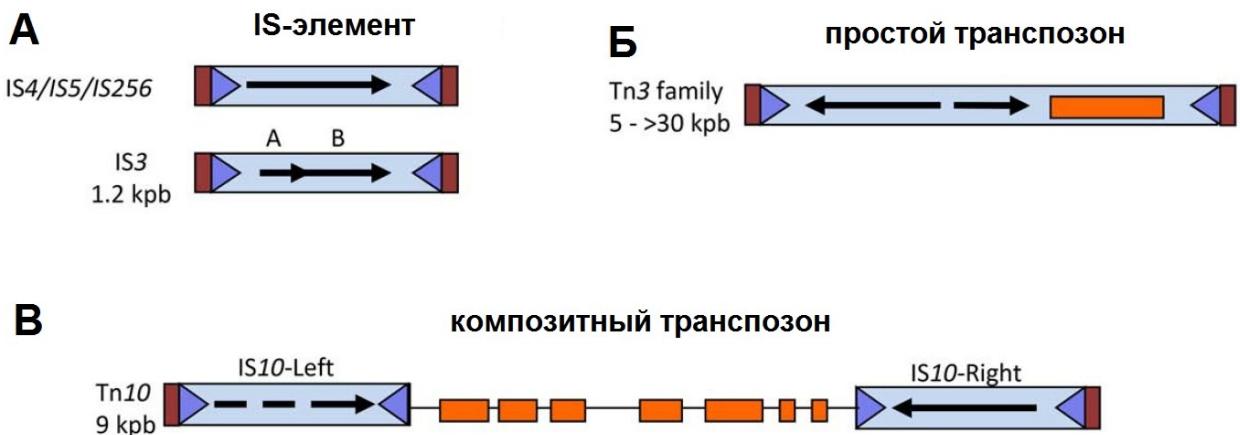


Рисунок 1.9 — Схематическое изображение основных типов мобильных элементов, содержащих в себе IS-элементы. IS-элементы показаны голубыми прямоугольниками, концевые инвертированные повторы показаны синими треугольниками, фланкирующие прямые повторы (возникают при интеграции элемента) показаны красными прямоугольниками. Слева указаны семейства IS-элементов и транспозонов, являющиеся их типичными представителями. А) Типичный IS-элемент содержит ген транспозазы и окружен инвертированными и прямыми повторами. Б) Простой транспозон содержит один или несколько генов между повторами. В) Композитный транспозон — это один или несколько генов, окруженных IS-элементами. Изображение адаптировано из [119].

IS элементы могут участвовать в передаче генов, находящихся между ними; такие элементы генома называются составными (композитными или сложными) транспозонами (рисунок 1.9В) [124]. Если IS-элемент содержит в себе один либо несколько карго-генов, то его называют простым транспозоном (рисунок 1.9Б). Специфичность встраивания различных транспозонов и IS-элементов значительно различается [125].

### 1.2.3 Факторы, влияющие на горизонтальный перенос генов

Известно, что горизонтальному переносу подвергаются гены всех функциональных категорий, включая и наиболее базовые функции - например, рибосомальные опероны [126]. При этом, переноситься могут гены, которые уже

присутствуют в геноме реципиента; в таком случае результатом будет не появление нового гена, но замена одно варианта другим. Данный процесс важен для поддержания стабильности генома. Например, для *E. coli* описан сценарий появления склонных к мутациям штаммов за счет поломки генов системы репарации. Затем может происходить исходного уровня мутабельности, за счет горизонтального переноса генов дикого типа [127]. Гипермутабельность может быть полезна в условиях стресса, поскольку может привести к нахождению лучшего в данных условиях варианта генома, а восстановление исходного уровня мутабельности важно для стабилизации нового варианта [128].

Для случая переноса новых генов, высказано предположение, что на вероятность того, что ген будет передаваться горизонтально, влияет количество взаимодействий его продукта с другими белками внутри организма [129]. Согласно данному предположению, чем более тесно связан ген различными взаимодействиями, тем меньше вероятность того, что он будет успешно функционировать в непохожем организме, из-за отсутствия необходимых ему белков-партнеров [130].

У ряда естественно-компетентных организмов описаны короткие последовательности (около 12 нуклеотидов), наличие которых увеличивает вероятность горизонтального переноса содержащих их генов. Такие последовательности получили название “последовательности захвата” (“uptake sequences”), поскольку они увеличивают вероятность захвата и переноса ДНК. Подобные последовательности описаны у различных представителей семейств Neisseriaceae и Pasteurellaceae [73; 131], вида *Haemophilus influenzae* [132]. Различные виды нейсерий обладают немного различными последовательностями захвата [133]; различия в последовательностях захвата связаны со снижением уровня межвидового переноса ДНК. Чаще всего последовательности захвата представлены в генах ”домашнего хозяйства” и кор-геноме в целом [134], что подчеркивает важную роль трансформации в поддержании стабильности геномов за счет исправления мутаций при гомологичной рекомбинации [135].

Геномные перестройки, обусловленные наличием повторов в геноме могут играть регуляторную роль [136]. Так, для кишечной палочки описана инверсия фрагмента генома, окруженного IS-элементами, и дающая выгоды при голодаании бактерий [137].

## Области повышенной частоты горизонтального переноса генов

Области генома с повышенной частотой событий горизонтального переноса генов — ”горячие точки” — были описаны у ряда бактерий [138]. Они могут возникать за счет сайт-специфичной интеграции мобильных элементов и находиться в генах тРНК, либо иных генах. У сальмонелл описана область внутри гена *gyrA*, служившая сайтом интеграции для чужеродной ДНК [139]. Многочисленные изменения генного состава наблюдаются в составе геномных островов, для которых описано наличие консервативных и вариабельных фрагментов [140] (механизм вариабельности при этом остается не известным [85]). Хорошо изученными точками повышенной изменчивости являются интегроны [108].

Наиболее масштабное исследование ”горячих точек” горизонтального переноса, которое нам известно, было опубликовано в 2017 году, группой Эдуардо Роча [138]. Авторы провели исследование уровня изменчивости в 80 бактериальных видах; по их оценкам перенесенные гены сконцентрированы только в 1% хромосомных областей (горячих точек). Они наблюдали, что большинство мобильных элементов генома и генов устойчивости к антибиотикам находятся в ”горячих точках” но, при этом, во многих высокоизменчивых областях генома мобильные элементы отсутствуют.

Также в работе [138] описано наблюдение, что в окрестности ”горячих точек” генома наблюдается повышенная частота событий гомологичной рекомбинации. По мнению авторов, это объясняется значительной ролью данного процесса в изменении генного состава в высоковариабельных областях генома.

### 1.3 Биоинформационические методы исследования изменчивости генома

#### 1.3.1 Методы поиска ортологии

Ортологичные гены — это гены, происходящие от общего предка (гомологи), связанные с общим предком лишь событиями специализации, без участия событий горизонтального переноса либо дупликаций [141]. Определение ор-

тологиченых генов является важным шагом для филогенетического анализа, предсказания функции генов, сравнительных исследований геномов [142]. Поиск ортологов вычислительными методами - сложная задача. Для наиболее точного решения требуется реконструкция генного состава предковых геномов и расчет эволюционных сценариев для каждого гена с реконструкцией событий дупликации, потери и горизонтального переноса генов [142]. Помимо значительной вычислительной сложности, полное решение подобной задачи, требует наличия большого количества геномов, которые были бы репрезентативными для исследуемой группы организмов [143]. Зачастую, задача поиска ортологов решается приближенными методами.

Наиболее вычислительно простыми являются методы, основанные на оценке сходства последовательностей. К ним относятся, в частности, методы, основанные на поиске лучших двунаправленных совпадений (best bidirectional hit). В основе лежит предположение, что последовательности ортологичных генов более похожи друг на друга, чем на любые другие последовательности в соответствующих геномах [144]. Данный подход был реализован в таких программах как Hieranoid [145], ранних версиях и публичной базы данных ОМА [146]. Данный подход может учитывать только взаимно-однозначные отношения генов. Если в любой из двух сравниваемых линий произошли дупликации генов, для правильного описания потребуется отношение “один ко многим” или “многие ко многим”.<sup>□</sup> В таких случаях подход, основанный на поиске лучших двунаправленных совпадений, упускает из виду многие истинные ортологи [147; 148].

Широкое распространение получила классификация генов по КОГам - кластерам ортологичных групп (Clusters of Orthologous Groups, COG). КОГи рассчитываются на основе поиска троек наилучших совпадений генов и объединении троек, имеющих общие ребра [149]; данный подход не столь консервативен как поиск лучших двунаправленных совпадений и не склонен ”пропускать” гены, ошибочно не соотнося их с их ортологами [150]. Конструирование базы COG было произведено с задействованием курирования полученных кластеров человеком, что было возможно во времена, когда количество прочитанных геномов исчислялось десятками [151], но не реализуемо в настоящее время с учетом огромного количества прочитанных геномов. Существуют подходы, основаны на кластеризации генов на основе рассчитанных попарных расстояний между ними, в частности, при помощи метода кластеризации на графах MCL (Markov Cluster Algorithm) [152; 153]. В отличие от методов, основанных на поиске лучших

дву направленных совпадений, подход MCL является более инклюзивным (меньшее количество генов ошибочно не соотносятся с группами ортологии). В тоже время он может объединять не ортологичные гены в одну группу, за счет сходства последовательностей [150]. Примером реализации данного подхода является программа OrthoFinder [153]. На вход ей подаются аминокислотные последовательности белок-кодирующих генов во всех рассматриваемых геномах, после чего происходит запуск попарного выравнивания BLAST (опционально можно использовать diamond [154]), после чего идет этап кластеризация последовательностей при помощи алгоритма кластеризации на графах MCL [155], полученные в результате кластеры называются ортогруппами. В новой реализации данной программы возможен также дополнительный шаг поиска ортологов на основе филогенетического подхода [156].

Существуют методы поиска ортологов, в которых филогенетический анализ является начальным этапом анализа [147]. Проводимое вручную сравнение филогенетического дерева, построенного на основе выравнивания гомологичных последовательностей, с деревом рассматриваемых видов служило "классическим" методом определения эволюционных сценариев; с появлением большого количества прочитанных геномов подобный анализ стал слишком трудозатратным. Были разработаны автоматические алгоритмы сопоставления генного дерева с деревом видов для получения минимального набора событий дупликации и потери генов, позволяющие объяснить наблюдаемые данные [157]. Это позволило проводить поиск ортологов на основе филогенетического подхода в масштабе полных геномов [158]. Основным ограничением ранних реализаций филогенетического подхода является предположение о том, что филогенетические деревья генов и видов не содержат ошибок, что часто не соответствует реальной ситуации, особенно для деревьев отдельных генов в которых уровень филогенетической информации низок и часто присутствуют одинаковые или очень схожие последовательности [159]. Для борьбы с данным недостатком был предложен подход, при котором фрагменты деревьев с низким уровнем достоверности сначала превращались в полиномии (то есть "схлопывались" так, что множество ветвей выходило из одного узла) и затем происходило согласование деревьев для гена и для организмов на основе минимизации получаемых событий дупликаций и потери генов [160].

### 1.3.2 Методы поиска горизонтально перенесенных генов

Для выделения горизонтально перенесенных фрагментов генома, используют различные типы методов [161]. В первом типе, определяют фрагменты генома, значительно выделяющиеся по некоторой характеристике (сигнатуре) на фоне остального генома. В качестве сигнатуры может выступать ГЦ-состав, индекс использования кодонов, k-мерный спектр (встречаемость комбинаций из k нуклеотидов) [162]. Такой подход применялся с начала 1990 годов (например, [163] и является наиболее простым в вычислительном плане. Особенно эффективен данный метод для определения недавних событий переноса между филогенетически далекими организмами - за счет дальности организмов ожидаются значительные отличия в сигнатаурах, и эти отличия еще не успели затереться из-за процесса "одомашнивания" перенесенного фрагмента ДНК [164]. Соответственно, данный тип методов малоприменим для определения переноса между близкими организмами (например, различными штаммами одного вида), поскольку при этом ГЦ состав и иные геномные характеристики ожидаются иметь очень схожие значения. Ограничения данного подхода проистекают также из наличия в геномах вариаций в сигнатаурах, не являющихся следствием горизонтального переноса генов [165; 166]; в частности, отличными от остального генома характеристиками обладают гены мобильных элементов, вирусов и плазмид, а также гены, находящиеся в области конца репликации у ряда бактерий [167].

Данный тип методов не способен определить донора и реципиента, участвовавших в горизонтальном переносе, задача этих методов — поиск геномных островов (фрагментов генома, являющихся результатом горизонтального переноса). Примерами реализации данного подхода являются такие программы как Alien hunter [168], GI hunter [169], GIPSY [170]. Alien Hunter основан на методе обнаружения атипичных областей в геноме при помощи подхода интерполированных мотивов переменного порядка (IVOM, Interpolated Variable Order Motifs) при анализе содержания G + C, присутствия динуклеотидов и частоте кодонов. Прогнозы оптимизируются с использованием скрытых марковских моделей (HMM) для определения точки входа в атипичные и нетипичные области генома (то есть для уточнения границ горизонтально перенесенного фрагмента). Alien Hunter может делать прогнозы без предварительной аннотации генома. GI hunter тоже основан на методе IVOM, но дополнительно учитывает расположение генов тРНК,

наличие интеграз и транспозаз, информацию о генах с высоким уровнем экспрессии, межгенном расстоянии. Все эти данные поступают на вход дерева решений, построенного на обучающей выборке [169]. Программа GIPSY также учитывает расположение генов тРНК и генов ассоциированных с мобильными элементами генома, ее отличает функция классификации горизонтально перенесенных фрагментов генома на различные типы: острова патогенности, острова устойчивости к антибиотикам и острова, обеспечивающие симбиоз (*symbiotic islands*) [170]. В недавнем сравнении эффективности методов данного типа, наибольшую точность показал Alien Hunter [171].

Второй поход основан на построении филогенетических деревьев отдельно по генам и сравнении их с филогенетическим деревом рассматриваемых организмов (построенном, как правило, на основании общей части геномов (коргенома) либо последовательности 16S рРНК). Гены, для которых эволюционная история плохо совпадает с историей организмов, считаются унаследованными не вертикально (то есть с участием горизонтального переноса либо дубликации) [172]. Такой подход также начал применяться еще в 1990 годах (например, [173]). Данный подход часто применяется в настоящее время. Например, методом определения горизонтального переноса в базе данных горизонтально-перенесенных генов HGTree [174] является метод Ranger-DTL, основанный на филогенетическом подходе [175]. Ranger-DTL — это программный пакет для определения эволюции генного семейства с учетом процесса специализации, дупликации генов, горизонтальному переносу генов и потере генов. Данная программа принимает в качестве входных данных дерево генов и дерево видов и согласовывает их за счет введения предполагаемых событий видеообразования, дублирования, передачи и потери генов. Сложность применения данного подхода проистекает от его зависимости от задач, которые сложны и сами по себе — филогенетическая реконструкция и лежащие в ее основе поиск ортологов и множественное выравнивание генов [161]. Также, проблему представляет слабость филогенетического сигнала, содержащегося в последовательностях отдельных генов при рассмотрении близкородственных организмов (например, штаммов одного вида). В таком случае деревья, построенные по отдельным генам, содержат много случайных разветвлений с низким уровнем поддержки (например, низкими значениями bootstrap). Для борьбы с возникающими неоднозначностями при реконструкции филогенетического дерева для отдельных генов, в методе GeneRax [176] предложено проводить филогенетическую реконструкцию по генам на основе филогенетического де-

рева для рассматриваемых организмов, за счет этого разрешая возникающие неоднозначности. В качестве входных данных, GeneRax принимает множественное выравнивание для гена и укорененное дерево рассматриваемых организмов (например, построенное на основе общей части геномов). Затем он проводит реконструкцию филогенетического дерева для рассматриваемого гена методом максимального правдоподобия, с учетом различных эволюционных сценариев — событий переноса, дубликации либо потери гена.

Существуют также программы, реализующие "неявный" филогенетический подход. Например, HGTeator основан на сравнении сходств белков рассматриваемого генома по отношению к филогенетически близким и филогенетически более далеким организмам. Ряд методов основаны на поиске отличий в генном составе близкородственных геномов: участки которые есть в определенном геноме и при этом отсутствует в геномах большинства филогенетически близких организмов считаются результатами горизонтального переноса генов. Появляются также гибридные методы, использующие в качестве критериев как различие в сигнатаурах, так и филогенетическую информацию [177].

Ожидаемо, разнообразие подходов и методов определения горизонтального переноса приводят к разнообразию в получаемых результатах [178].

### 1.3.3 Методы визуализации отличий в геномах

В данном разделе мы кратко опишем основные методы визуализации, используемые при сравнении геномов.

Одним из первых способов сравнения последовательностей геномов был метод точечных диаграмм сходства, при котором два сравниваемых генома располагаются по двум осям графика, а точками либо линиями отображаются области сходства между геномами (в англоязычной литературе они получили название dot plot alignment). Пример показан на рисунке 1.10А. Такой способ применим только для пар геномов и сравнительно редко используется в настоящее время.

Для сравнения нескольких небольших геномов (например, вирусов) или фрагментов больших геномов применяются графики, в которых отдельными горизонтальными линиями показаны геномы, стрелками обозначены гены, а вертикальные линии соединяют гомологичные гены либо области синтезии —

схожих по последовательности участков геномов (рисунок 1.10Г). Несомненным преимуществом такого способа визуализации является наглядность и удобство прослеживания изменения в расположении отдельных генов; метод применим для небольшого числа сравниваемых геномов (порядка 10-20) и небольшом числе генов.

Схожим образом, можно показывать множественное выравнивание полных последовательностей геномов бактерий, такие графики часто строят в программе Mauve [179] (рисунок 1.10Б). Сначала происходит поиск блоков синтезии. Блоки синтезии отображаются прямоугольниками (как правило, цветными) и соответствующие блоки соединяются линиями. Данный подход также ограничен в применимости сравнительно небольшим количеством геномов (порядка 10), после чего полученные визуализации становятся трудны для восприятия.

Еще одним подходом к визуализации отличий геномов является представление их на круговой диаграмме (рисунок 1.10В). При этом, один геном выбирается в качестве референса и отображается на внутреннем круге, остальные геномы располагаются на внешних кругах. Для построения данного типа графиков часто применяется программа BRIG (Blast Ring Image Generator) [180]; она автоматически производит поиск блоков сходства при помощи алгоритма BLAST и позволяет добавить на график различную метаинформацию (ГЦ-состав, добавленные пользователем аннотированные области). Такой подход позволяет наглядно отобразить области, уникальные для референсного генома (например, геномные острова), но не позволяет показать альтернативные варианты генного состава. Количество сравниваемых геномов может быть достаточно большим (многие десятки) без потери информативности визуализации.

## 1.4 Применение графов для анализа геномных данных

Представление расположения генов в различных геномах в виде графа было применено в работах группы Евгения Кунина в начале 2000х годов [16; 183]. Узлами графа были кластеры ортологичных генов (COG, Clusters of Orthologous Groups of proteins), а ребрами соединялись консервативные пары генов: закодированные на одной цепи, разделенные не более чем двумя генами и представленные в трех и более геномах (из 31 генома, прочтенных к тому моменту). Целью ана-

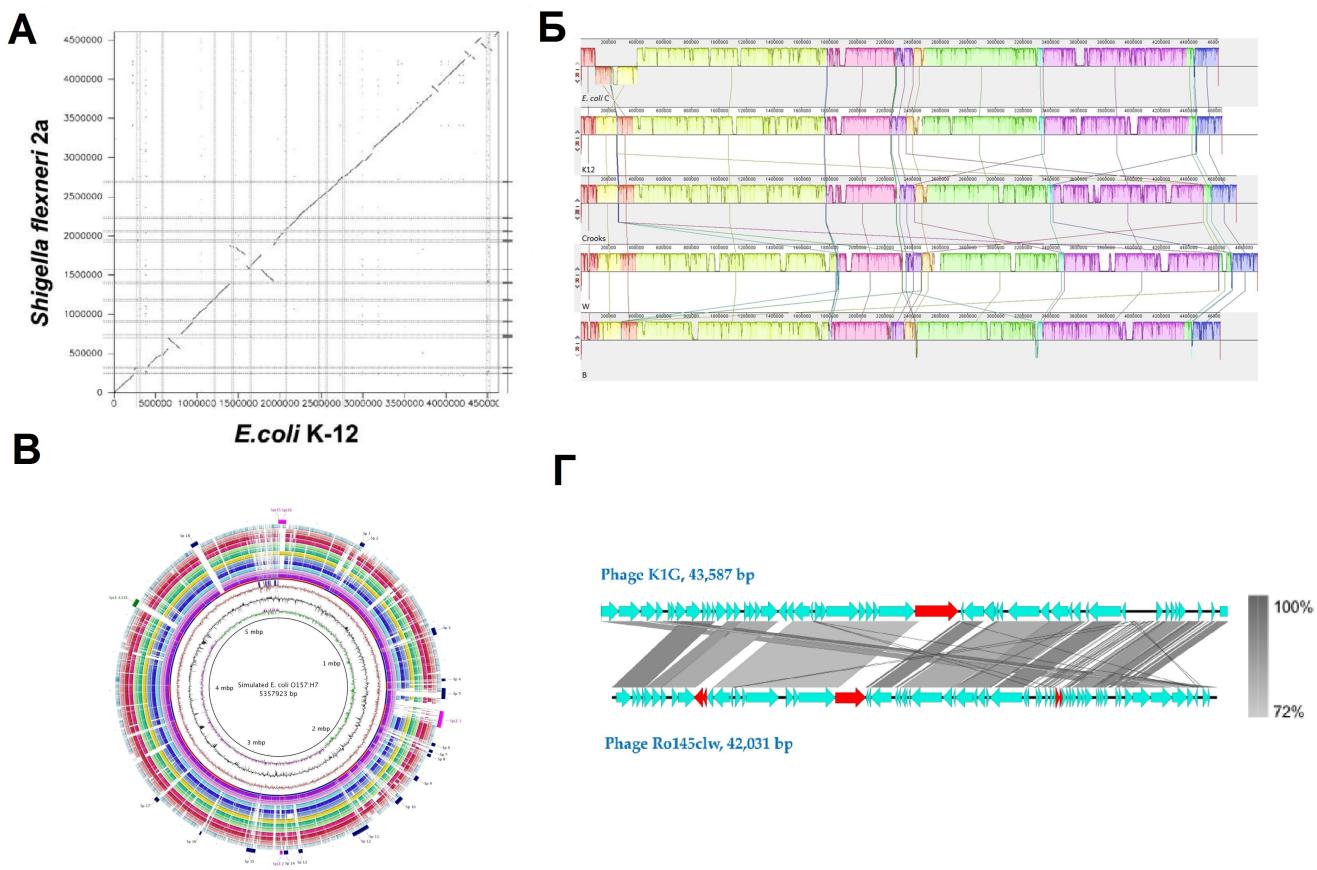


Рисунок 1.10 — Различные способы визуализации сравнений геномов. А) Сравнение различных штаммов *E. coli*. Координаты геномов сравниваемых штаммов отложены по осям абсцисс и ординат, на пересечении координат ставится точка или проводится линия, при условии совпадения последовательностей по этим координатам. Горизонтальными и вертикальными линиями показано расположение бактериофагов. Изображение адаптировано из [99]. Б) Сравнение различных штаммов *E. coli* при помощи программы Mauve [179]. Показаны крупные блоки синтезии; соответствующие блоки соединены линиями. Источник изображения: [181]. В) Сравнение геномов различных штаммов *E. coli* при помощи программы BRIG [180]. Геномы предоставлены в виде колец, внутреннее кольцо соответствует референсному геному. Источник изображения: [180]. Г) Сравнение состава генов и их расположения в двух фагах. Стрелками обозначены гены, полосами серого цвета показаны схожие последовательности (степень сходства закодирована градиентом серого цвета). Источник изображения: [182].

лиза было нахождение эволюционно устойчивых комбинаций генов. Всего было обнаружено 1505 консервативных пар генов. Большинство пар были представлены лишь в небольшом количестве геномов и только 21 пара генов присутствовала во всех сравниваемых геномах и включала гены рибосомных белков и субъ-

единиц РНК-полимеразы. Также, авторы искали устойчивые кластеры генов, представляющие из себя комбинации устойчивых пар. На рисунке 1.11 показан граф, представляющий один из обнаруженных кластеров [16]. Информация об устойчивых кластерах была использована для предсказания функций ранее не охарактеризованных генов архей [183]. В целом, анализ показал, что во многих случаях гены, входящие в кластер, не имели очевидных функциональных связей. По предположению авторов, возможны две альтернативные интерпретации этого результата: 1) гены по соседству только кажутся функционально несвязанными, тогда как в действительности они имеют дополнительные, еще не обнаруженные функции; 2) хотя функциональной связи не существует, продукты этих генов требуются примерно в равных количествах и при тех же условиях, что объясняет преимущество совместного регулирования, обеспечиваемого близким расположением данных генов [16].

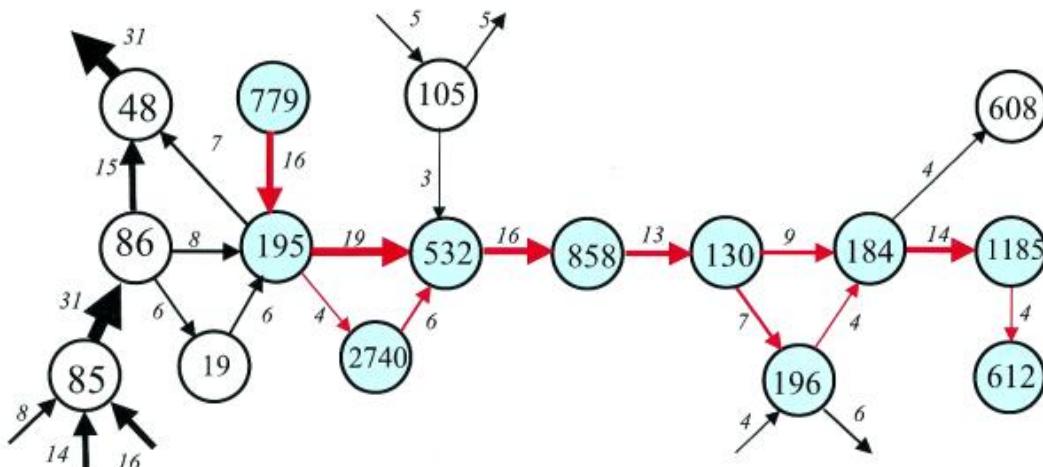


Рисунок 1.11 — Набор генов, представленный в виде ориентированного графа. Узлы соответствуют кластерам ортологичных генов (COG), номера COG указаны внутри кружков. Иллюстрация взята из работы [16].

В утилите PPanGGOLiN графы, построенные на основе расположения генов (подход, применяемый и в нашей работе), были использованы для классификации генов из пангенома на три категории: устойчивых генов (есть у всех), генов из оболочки (есть у многих) и генов из облака (встречаются у небольшого числа представителей) [184]. Авторы наблюдали, что доля неконсервативных генов не коррелирует с размером генома.

Для визуализации графа, который строится в программе PPanGGOLiN, можно использовать графовые редакторы, такие как Gephi (рис 1.12). В декабре

2020 года вышла статья, описывающая модуль программы PPanGGOLiN, предназначенный для поиска геномных островов [113].

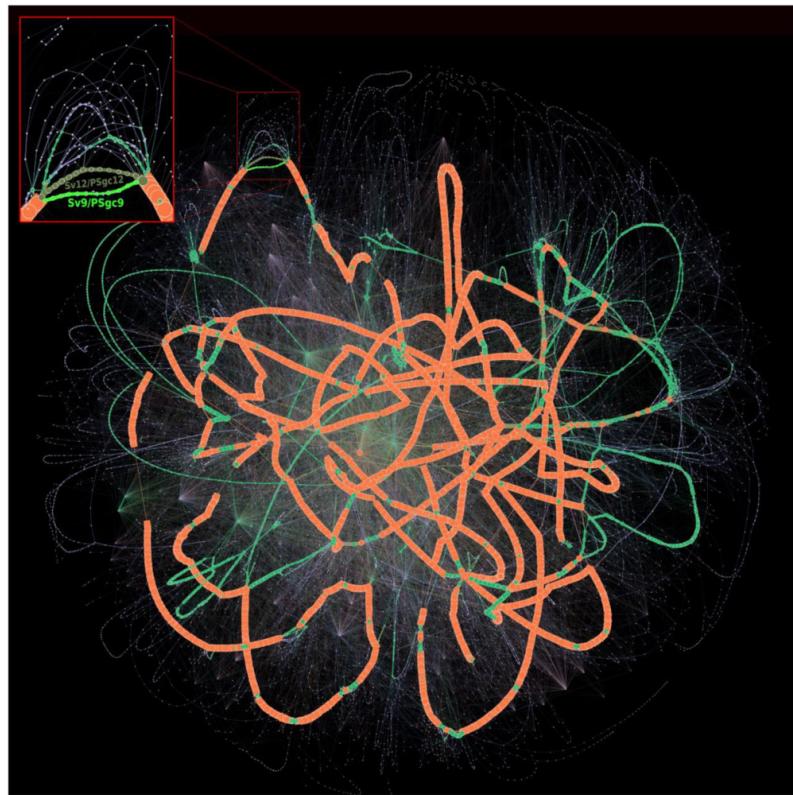


Рисунок 1.12 — Пангеномный граф, построенный программой PPanGGOLiN на основе 3117 геномов вида *Acinetobacter baumannii*. Узлы соответствуют семействам генов, ребра соответствуют солокализации генов в геномах. Ребра между генами из устойчивой части генома, оболочки и облака окрашены в оранжевый, зеленый и синий цвета, соответственно. Изображение из публикации [184]

Также, построение данного типа графов реализовано в пакете FindMyFriends (<https://github.com/thomasp85/FindMyFriends>) для языка R, предназначенном для проведения пангеномного анализа (поиска групп гомологий, анализа их представленности в геномах). На рис 1.13 приведен пример визуализации сорасположения генов из различных групп гомологий, выполненный при помощи пакета FindMyFriends.

Графы геномной вариабельности, построенные на основе множественного выравнивания, могут быть использованы в качестве референса при картировании прочтений, что улучшает результат картирования (снижает эффект более низкой глубины покрытия в вариабельных участках генома) [185; 186].

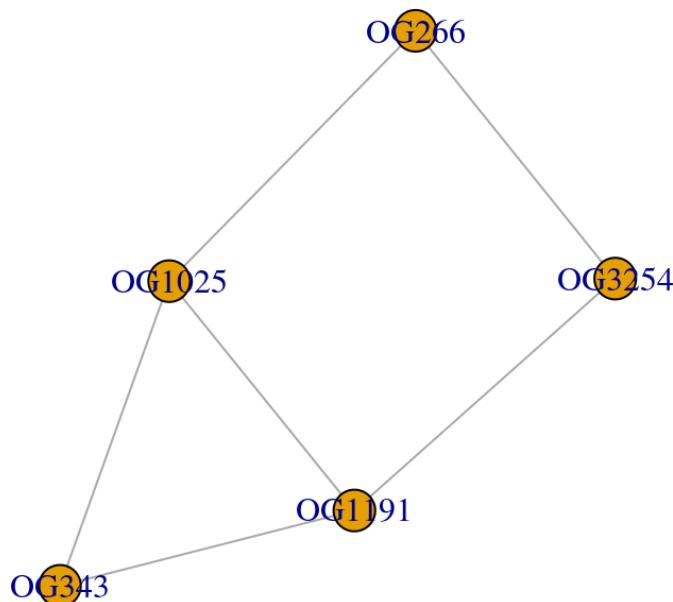


Рисунок 1.13 — Граф сорасположения генов из различных групп гомологий, выполненный при помощи пакета FindMyFriends. Пример из руководства к данному пакету.

## 1.5 Ассоциативная связь болезни Крона с колонизацией *E. coli*

Болезнь Крона (БК) — это рецидивирующее воспалительное заболевание кишечника неизвестной этиологии, характеризующееся наличием трансмурального гранулематозного воспаления, диареей, болью в животе, потерей веса, иногда - осложнениями, проявляющимися в иных органах (суставы, печень, глаза) [187]. Была обнаружена ассоциативная связь между повышенной представленностью *E. coli* в кишечном микробиоме с наличием и тяжестью болезни Крона (как, впрочем, и другими воспалительными заболеваниями кишечника); у пациентов доля кишечной палочки в микробиоте повышена в 10-100 раз (по разным оценкам) по сравнению с ее содержанием у здоровых людей [188—191]. Обнаружены мутации в геноме человека, которые увеличивают предрасположенность к болезни Крона. Ряд этих мутаций находится в генах, продукты которых задействованы во взаимодействии человеческого организма и микробиоты [192]. Наиболее вероятным кажется предположение, что в развитии данного заболевания играют роль как факторы внешней среды (включая, нарушение состава микробиоты), так и предрасположенность организма [192].

Филогенетическое положение изолятов *E. coli* из пациентов с БК значительно варьирует и затрагивает все мажорные филогруппы (A, B1, B2, D)

данного организма [193]. Экспериментально было обнаружено, что ряд изолятов кишечных палочек, полученных от пациентов с БК, оказались способными проникать в эпителиальные клетки кишечника и выживать в макрофагах, что было названо адгезивно-инвазивным фенотипом [190; 194; 195]. Согласно мета-анализу, проведенному в 2020 году, кишечная палочка с адгезивно-инвазивным фенотипом встречается у 29% пациентов с болезнью Крона, что значимо чаще, чем у здоровых людей (у них она встречается в 9% случаев) [196]. Установить с определенностью генетические детерминанты, позволяющие бактериям приобретать адгезивно-инвазивный фенотип, не удается [197; 198]. Среди генов, чаще встречающихся у изолятов кишечной палочки из пациентов, по сравнению с изолятами из здоровых людей, в нашем и других исследованиях, были определены гены оперона утилизации 1,2-пропандиола [193; 199; 200]. Данное вещество является продуктом переработки фукозы комменсальными микробами. У сальмонелл (филогенетически близких к кишечным палочкам патогенных микроорганизмов), утилизация 1,2-пропандиола происходит наиболее эффективно в условиях протекания в кишечнике воспалительной реакции, поскольку при воспалении в просвет выделяются вещества, которые сальмонелла использует как акцепторы электронов. При этом, она становится бенефициаром патологического процесса, а значительная часть иной микрофлоры погибает [201]. Схожий сценарий можно предположить и для кишечной палочки. Вероятно, для нее выгодно провоцировать воспалительную реакцию, поскольку это позволяет эффективно утилизировать доступный источник питания и получать доминирующее положение в микробиоме [193]. Оперон утилизации пропандиола представлен у филогенетически далеких организмов и часто рассматривается как продукт горизонтального переноса генов [202; 203].

К другим генам кишечных палочек ассоциированными с наличием у их носителя болезни Крона, в разных работах относят гены, вовлеченные в захват железа [193; 204], синтез капсулы [193; 205], синтез пилей первого типа [206], ген кодирующий белок внешней мембранны OmpC [207] и ряд других.

## Глава 2. Материалы и методы

### 2.0.1 Отбор последовательностей геномов

Для анализа геномной изменчивости *E. coli* мы использовали 327 геномов данного 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 от представителей наиболее крупных филогрупп (подбор проводился на основе литературных данных). Затем для каждого представителя были выбраны 100 наиболее близких по геномам, доступных в базе RefSeq. В качестве меры сходства последовательностей мы использовали суммарную ширину выравниваний фрагментов генома, выравнивание проводилось программой pusteg [208].

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 [209]. 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 последовательностей геномов всех видов, для которых было доступно не менее 50 последовательностей геномов в базе данных RefSeq. При наличии более 100 последовательностей геномов, в анализ включались 100 случайно выбранных последовательностей. Таким образом была сформирована выборка из 143 видов прокариот, включая два вида архей.

## 2.0.2 Анализ геномной вариабельности

Белок-кодирующие последовательности во всех загруженных геномах были аннотированы с помощью программы Prokka ver 1.11 [210]. Гены были отнесены к ортогруппам с помощью OrthoFinder ver. 2.2.6 [153].

Скрипты на языке Python, содержащиеся в приложении GCB (<https://gcb.rcpm.org/>) использовались для оценки уровня изменчивости генома и создания подграфов вокруг интересующих областей генома. Принципы их работы описаны 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 [211]. Для формализации определения областей генома с повышенной изменчивостью мы использовали критерий Тьюки, основанный на межквартильном расстоянии.

Статистическая обработка и визуализация данных мы проводили на языке R. Для определения коэффициентов корреляции Спирмена использовали функцию *cor*. Статистическая значимость корреляций определялась при помощи функции *cor.test*. Индексы согласованности признаков с филогенетическим деревом (retention index) рассчитывались с использованием функции RI из библиотеки phangorn для языка R. Для построения линейных моделей использовалась функция *lm* языка R.

Для построения филогенетического дерева различных видов рода *Bacillus* мы выровняли транслированные последовательности всех ортологичных однокопийных генов при помощи программы muscle [212], преобразовали их в выравнивания кодонов с помощью pal2nal и построили дерево с помощью iqtree v1.6 [213] с опцией ModelFinder Plus (оптимальный подбор эволюционной модели); конвейер snakemake для этих шагов доступен по адресу [https://github.com/paraslonic/orthosnake/blob/tree/Snakefile\\_tree](https://github.com/paraslonic/orthosnake/blob/tree/Snakefile_tree).

Поиск областей синтезии мы проводили с помощью программы pustmer [208] (при сравнении последовательностей геномов различных штаммов одного вида), либо программы Mauve [179] (при сравнении последовательностей геномов принадлежащих различным видам).

Для определения профагов в геномах мы использовали онлайн сервис Phaster [214].

Для *E. coli* мы использовали нормированную матрицу контактов из работы [54] и доступную в репозитории: [https://github.com/koszullab/E\\_coli\\_analysis/tree/master/data](https://github.com/koszullab/E_coli_analysis/tree/master/data). Для *B. subtilis* мы использовали матрицу контактов из статьи [215]. Для нормировки матрицы хромосомных контактов использовалась функция *normalizeCore.performIterativeCorrection* из библиотеки *gcMapExplorer* <https://github.com/rjdkmr/gcMapExplorer>, количество итераций равнялось 1000, значение шага составило 0.00001. Для построения линейных моделей использовалась функция *lm* языка R.

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

### 2.1.1 Группа пациентов и клинический материал

Пациенты были отобраны из двух клинических центров (ЦНИИ гастроэнтерологии и Государственного научного центра колопроктологии) в Москве, Российской Федерации, 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 подтверждена. Критериями исключения были признаки неопределенного колита, инфекционные заболевания, недавнее лечение антибиотиками. Для исследования 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 принимали участие: Щербаков П.Л., Маев И.В., Павленко А.В., Андреев Д.Н., Халиф И.Л.

### 2.1.2 Выделение изолятов *E. coli*

Выделение *E. coli* выполняли следующим образом. Приблизительно 0,05 мл объема фекалий помещали в 0,5 мл стерильного буфера (PBS), перемешивали

вали на вортексе до гомогенности, аликовту разбавляли примерно в  $10^6$  раз. Затем 0,1 мл полученной жидкости наносили на чашки со средой LB. После инкубации в течение ночи при 37°C изолированные колонии идентифицировали с помощью программного обеспечения Matrix Assisted Laser Desorption / Ionization (MALDI) Biotype (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.

### **2.1.3 Экстракция ДНК и геномное секвенирование**

Геномную ДНК из отдельных культур экстрагировали с помощью набора QIAamp DNA Mini (Qiagen) в соответствии с протоколом производителя. Экстрагированная 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., Лисицкой Е.С.

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

### **2.1.4 Сборка генома, исправление ошибок в гомополимерных областях**

Последовательности генома были собраны с использованием программ Mira 4.0 [216] со стандартными параметрами и SPADES 3.10.0 [217].

Каждая сборка проверялась на наличие возможных контаминаций (последовательностей нецелевого организма) при помощи скрипта, написанного

на языке R и доступного по адресу: [https://github.com/paraslonic/BacPortrait/blob/master/portrait\\_spades.r](https://github.com/paraslonic/BacPortrait/blob/master/portrait_spades.r). Данный скрипт отображает каждый контиг (отдельный фрагмент сборки) на диаграмме с ГЦ-составом и глубиной покрытия контига; дополнительно отображается информация о таксономической аннотации заданного числа контигов.

Для технологии секвенирования Ion Torrent характерно наличие значительного количества ошибок в определении копийности нуклеотидов, особенно в гомополимерных областях. Для исправления данного типа ошибок, которые могут приводить к ошибкам сборки и искусциальному сдвигу рамки считывания в кодирующих последовательностях (CDS), нами был разработан следующий метод. Проводилось картирование прочтений на сборку; поиск позиций с вставками либо делециями в картированных прочтениях при помощи утилиты VarScan; выравнивание областей сборки вокруг найденных позиций при помощи программы BLAST на базу nt (NCBI); выбор варианта последовательности, который соответствует лучшему выравниванию BLAST и представлен в прочтениях 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 с считываниями более точных технологий секвенирования, таких как Illumina, SOLID и Sanger). Вычислительный конвейер для данной процедуры доступен по адресу: [www.github.com/paraslonic/HomoHomo](http://www.github.com/paraslonic/HomoHomo).

Последовательности геномов доступны в GenBank со следующими номерами доступа: RCE01 (JUDV00000000), RCE02 (JUDW00000000), RCE03 (JUDX00000000), RCE04 (JUDY00000000), RCE05 (JWJZ00000000), RCE06 (JWKA00000000), RKA00000000 (JWKA00000000), RKA000006 (JWKA00000000), RKA00000000 JWKB00000000), RCE08 (LAXB00000000), RCE10 (LAXA00000000), RCE11 (LAWZ00000000).

## 2.1.5 Сбор внешних данных

Для проведения анализа по поиску оперонов, чаще встречающихся в кишечных палочках, изолированных от пациентов с болезнью Крона по сравнению с изолятами, полученными от здоровых людей, мы также использовали внешние данные, собранные на основании анализа литературы. Геномы *E. coli* изолиро-

ванные от пациентов с болезнью Крона были взяты из работ [190; 204; 218]. В группу контроля входили геномы, описанные как комменсальные или лабораторно культивируемые непатогенные штаммы (полный список доступен в публикации [193]).

### 2.1.6 Поиск ортогрупп

Полученные последовательности генома аннотировали с помощью программы PROKKA 1.7 [210]. Информация об оперонной структуре была получена из базы данных DOOR [219]. Ортогруппы (группы гомологий включающие как ортологичные, так и паралогичные гены) были получены с помощью программы OrthoFinder v1.0.8 [153] с параметрами по умолчанию.

Статистический анализ представленности генов и оперонов в последовательностях 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 анализа и интерпретации полученных результатов принимали участие: Говорун В.М., Ракитина Д.В., Гарушянц С.К.

## Глава 3. Результаты

### 3.1 Разработка и верификация метода оценки изменчивости генома на основе графового представления расположения генов

#### 3.1.1 Разработка способа представления расположения генов в виде графа

Рассмотрим три условные генома, показанные в верхней части рисунка 3.1. Предположим, что мы установили группы гомологии и стрелками одного цвета показаны гомологичные гены. Для представления этого набора геномов в виде графа, будем ставить в соответствие каждому множеству гомологичных генов узел графа. Ребрами соединим узлы, для которых гены из соответствующих групп гомологии расположены последовательно хотя бы в одном геноме. Вес ребра установим в количество геномов, подтверждающих данную связь (в этих геномах соответствующие гены расположены последовательно). Таким образом мы закодировали информацию о расположении генов в наборе геномов в виде графа (нижняя часть рисунка 3.1).

В рассмотренном примере, мы представили три генома, содержащих суммарно 21 ген, в виде графа из 10 узлов. Чем большее количество геномов будет использовано в анализе, тем выше будет "экономия" — разница между количеством генов и количеством узлов в построенном графе.

Мы выбрали способ представления набора геномов в виде направленного графа: ребра имеют направление, соответствующие обходу генома от начала к концу. Это было сделано для того, чтобы сохранить всю информацию о порядке расположения генов, в частности о геномных инверсиях. Рассматриваемые геномы при этом должны быть ориентированы единообразно. В публичных базах геномов (RefSeq, GenBank, Patric и другие) не предусмотрено требований к ориентации последовательностей: они начинаются с произвольных позиций и могут быть ориентированы в произвольном направлении. Поэтому перед построением графа, мы производим согласование ориентаций геномов друг с другом. Для этого выбирается референсный геном (произвольно, из финишированных геномов), затем для каждого генома устанавливаются области синтезии (относи-

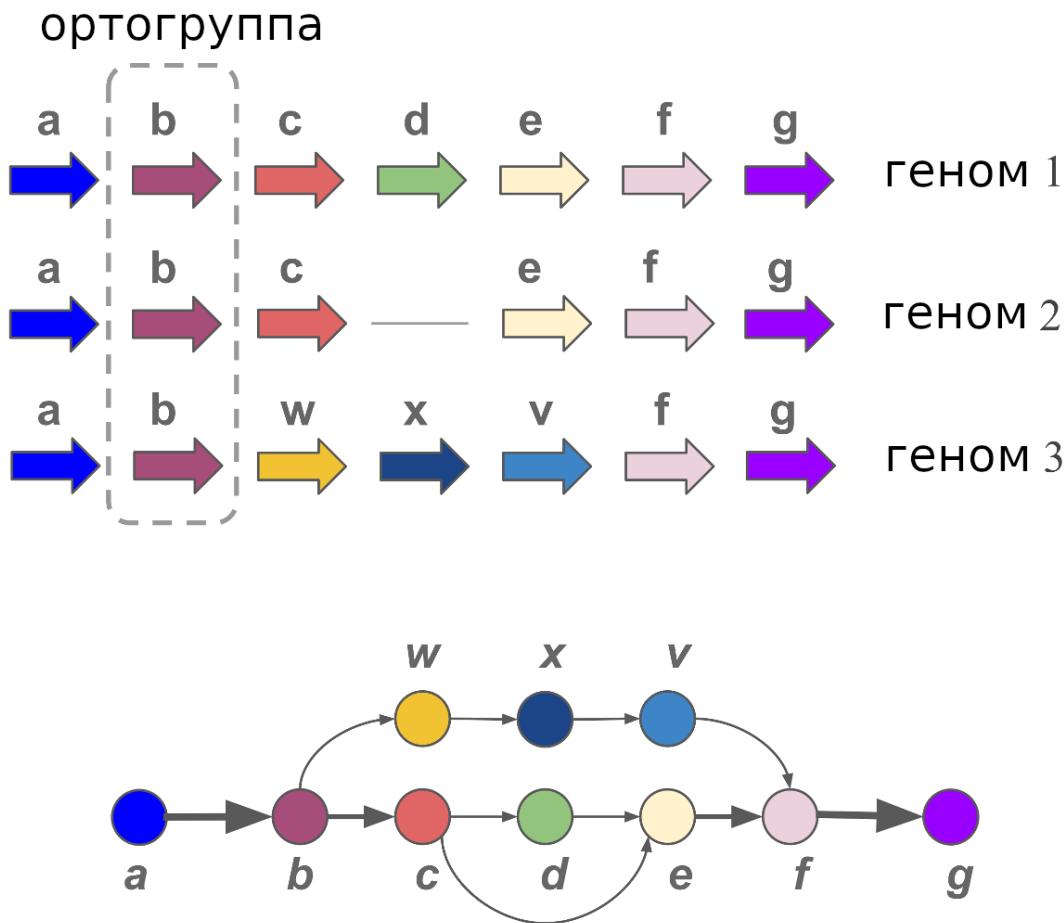


Рисунок 3.1 — Представление контекста генов в виде графа. Рассматриваем три гипотетических генома, состоящих из 6-7 генов. Гены показаны стрелками, цветом и буквами обозначены гены, относящиеся к одной ортогруппе. В нижней части рисунка показано графовое представление для данного набора геномов.

тельно референсного генома); подсчитывается сумма произведений длин блоков на нуклеотидное сходство для всех блоков в прямой и обратной ориентации; в зависимости от того, какая из сумм оказалась выше, оставляется либо исходная ориентация, либо она меняется на противоположную. Если рассматриваемый геном состоит из нескольких контигов, то согласование ориентации с референсом производится для каждого контига в отдельности.

Скрипт на языке R, выполняющий согласование ориентаций генома и построение графа доступен по адресу [https://github.com/paraslonic/graph\\_complexity/scripts/OG2graph.r](https://github.com/paraslonic/graph_complexity/scripts/OG2graph.r). На вход ему подается таблица групп гомологии в формате выдачи программы OrthoFinder - текстовый файл, каждая строка которого соответствует одной группе гомологии и имеет формат: <идентификатор группы>: <ген1> <ген2> ... <ген n>. На выходе создается файл

paths.sif, содержащий граф в формате sif: <идентификатор группы 1> <идентификатор группы 2> <геном>.

Для автоматизации представления набора геномов в виде графа мы реализовали вычислительный конвейер, схема которого представлена на рисунке 3.2. В нем осуществляются следующие шаги: 1) аннотация геномов (выделение белок-кодирующих областей), 2) форматирование файла в формате FASTA с аминокислотными последовательностями генов (в заголовках последовательностей мы указываем информацию о расположении генов), 3) построение групп гомологий и 4) формирование графа. На вход подается набор нуклеотидных последовательностей геномов в формате FASTA, выходом является граф в формате sif. Данный конвейер реализован в системе управления рабочим процессом snakemake и доступен по адресу [https://github.com/paraslonic/graph\\_complexity](https://github.com/paraslonic/graph_complexity).

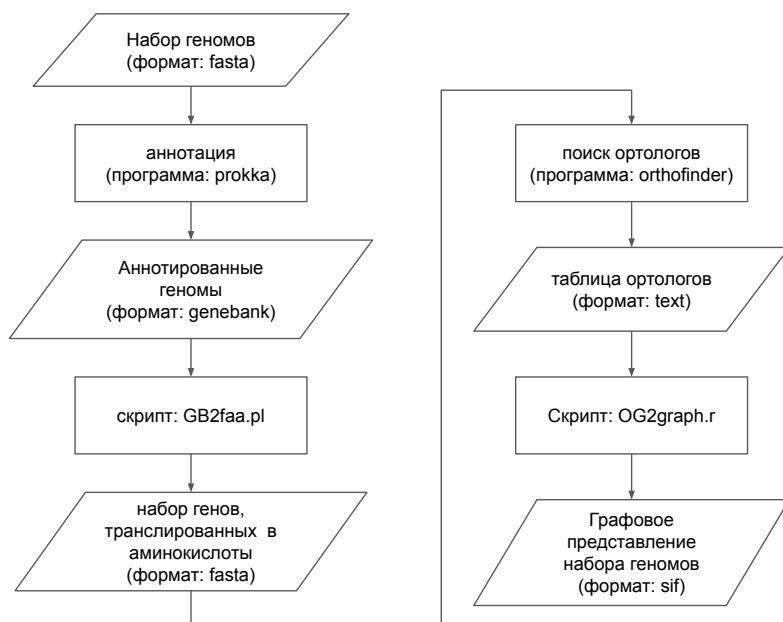


Рисунок 3.2 — Вычислительный конвейер для представления контекста генов в наборе геномов в виде графа.

### 3.1.2 Разработка алгоритма оценки изменчивости генома на основе графового представления

Предположим, что в ходе эволюции, в рассматриваемой группе геномов, наблюдались лишь изменения микро-масштаба – однонуклеотидные замены. В таком случае, набор и взаимное расположение генов не изменялись и соответствующий граф будет содержать лишь один путь. Любые изменения в составе, либо расположении, генов (делеции, вставки, перемещения) будут приводить к появлению новых ребер и новых узлов (в случае вставки нового гена). Чем больше узлов и ребер содержит график — тем больше возможных путей его обхода. Это позволяет использовать количество путей в графике в качестве меры вариабельности генома.

Разные фрагменты генома могут существенно различаться по уровню изменчивости. Ниже мы опишем процедуру построения профиля изменчивости генома. Данный профиль отражает как уровень изменчивости меняется вдоль хромосомы (либо иного репликона). Для подсчета уровня локальной изменчивости в отдельном фрагменте генома происходит подсчет количества путей в подмножестве графа (подграфе), соответствующему данному фрагменту. Локальная изменчивость подсчитывается во всех возможных локусах генома и таким образом получается профиль изменчивости. Иначе говоря, мы проходим по геному скользящим окном, выбирая области фиксированного размера (ширина окна задается пользователем) и оцениваем количество зафиксированных изменений в этих областях.

Принцип подсчета локальной вариабельности генома проиллюстрирован на рисунке 3.3. Из множества геномов выбирается один геном, служащий референсом (для него будет производиться построение профиля вариабельности). Путь, соответствующий чередованию генов в референсном геноме будем называть базовым путем (узлы этого пути обозначены зеленым цветом). Рассмотрим некоторую окрестность гена X, ограниченную тремя генами слева и справа от гена X. Произведем поиск путей, которые начинаются (исходят из базового пути) и заканчиваются (возвращаются в базовый путь) внутри рассматриваемой окрестности и при этом не проходят через узел X — такие пути будем называть обходными путями. Суммарное количество обходных путей принимается в качестве меры локальной вариабельности генома окрестности гена X. На нижней части рисунка 3.3 показаны примеры: слева — участка с низкой вариабельностью

и двумя обходными путями, справа — с более высокой вариабельностью и шестью обходными путями. Для получения профиля вариабельности, мы рассчитываем локальную вариабельность для каждого гена из референсного генома.

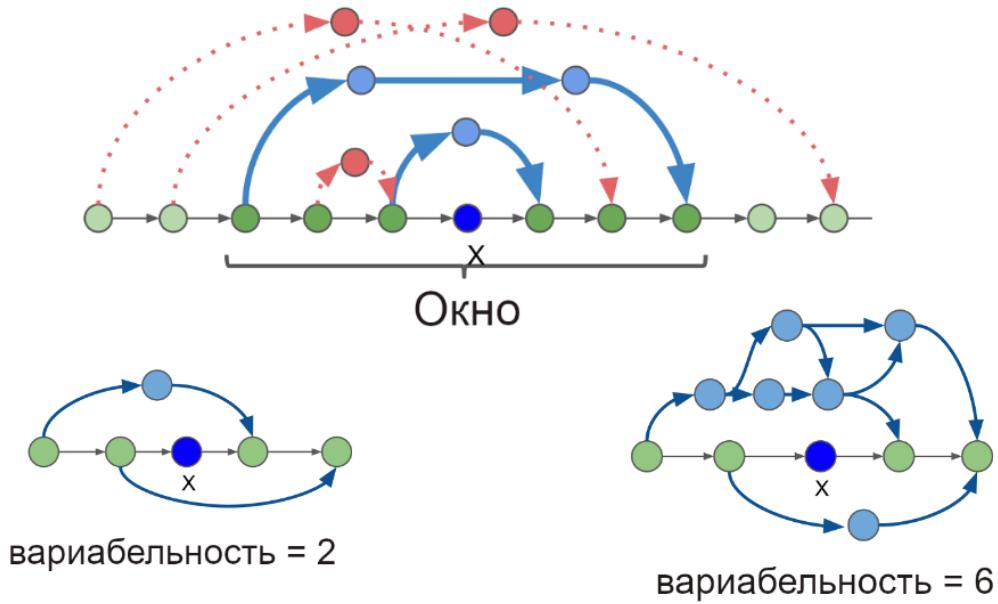


Рисунок 3.3 — Подсчет локальной вариабельности генома на основе количества путей в графе. Вверху показан подграф, представляющий окрестность гена X. Ребра черного цвета являются частью базового пути, который соответствует порядку генов в референсном геноме, ребра синего цвета — обходным путям, а ребра красного цвета — путям, которые не вносят вклад в значение вариабельности. На нижней части рисунка приведен пример графа с низкой (слева) и высокой (справа) вариабельностью, указаны соответствующие количества обходных путей.

### 3.1.3 Верификация предложенного метода оценки профиля изменчивости генома

Мы провели верификацию предложенного подхода оценки профиля изменчивости и его программной реализации при помощи: 1) симуляции эволюции геномов; 2) сравнения результатов нашего метода с ранее опубликованными оценками других авторов.

## Симуляция эволюции геномов

Мы проводили моделирование эволюции геномов на основе задаваемого профиля изменчивости. Данный профиль определял вероятности изменений генного состава в зависимости от положения гена в геноме. Мы использовали следующие варианты изменений: вставки либо удаления генов, перемещения генов, инверсии фрагментов моделируемого генома. Мы использовали несколько профилей изменчивости, для каждого из которых проводили моделирование геномных изменений, оценивали профиль изменчивости предложенным нами методом и сравнивали полученный результат с изначально заданным профилем. Нашей целью была проверка применимости предложенного метода в том случае, когда в геноме присутствуют области пониженной и повышенной изменчивости.

Модельный геном представлял из себя последовательность целых чисел ( $n=5000$ ). Вначале, создавались одинаковые геномы. Далее, проводилось 3000 итераций, в каждой из которых происходило изменение генома. Локализация изменений определялась на основе профиля изменчивости. Вероятности вставки и удаления генов были выбраны равными друг другу (для сохранения длины геномов), вероятности геномных инверсий были в 100 раз меньше (оценка получена исходя из литературных данных). Длины инверсий, вставок и делеций выбиралась случайно, на основе экспоненциального распределения (изменения малой длины были более вероятны, чем изменения большой длины). Вставка генов производилась из набора, размер которого в два раза превышал исходный набор генов (исходно в геноме были числа от 0 до 5000, вставлялись числа от 5000 до 15000).

Профили изменчивости были следующих типов: 1) ступенчатый, 2) синусоидальный, 3) пилообразный. Для каждого профиля изменчивости проводилось 10 симуляций. На рисунке 3.4 приведен пример сравнений исходного профиля изменчивости (слева) и профиля, полученного в результате анализа предложенным нами способом (справа). В таблице 1 приведены средние значения  $R^2$  и коэффициента корреляции исходного профиля изменчивости и профиля, полученного в результате анализа и усредненного по 10 повторам.

Отметим, что значительное влияние на точность оценки профиля изменчивости оказывает частота геномных инверсий: чем она выше - тем была ниже точность. Таким образом, частые перестройки генома служат существенным ограничением применимости предложенного метода.

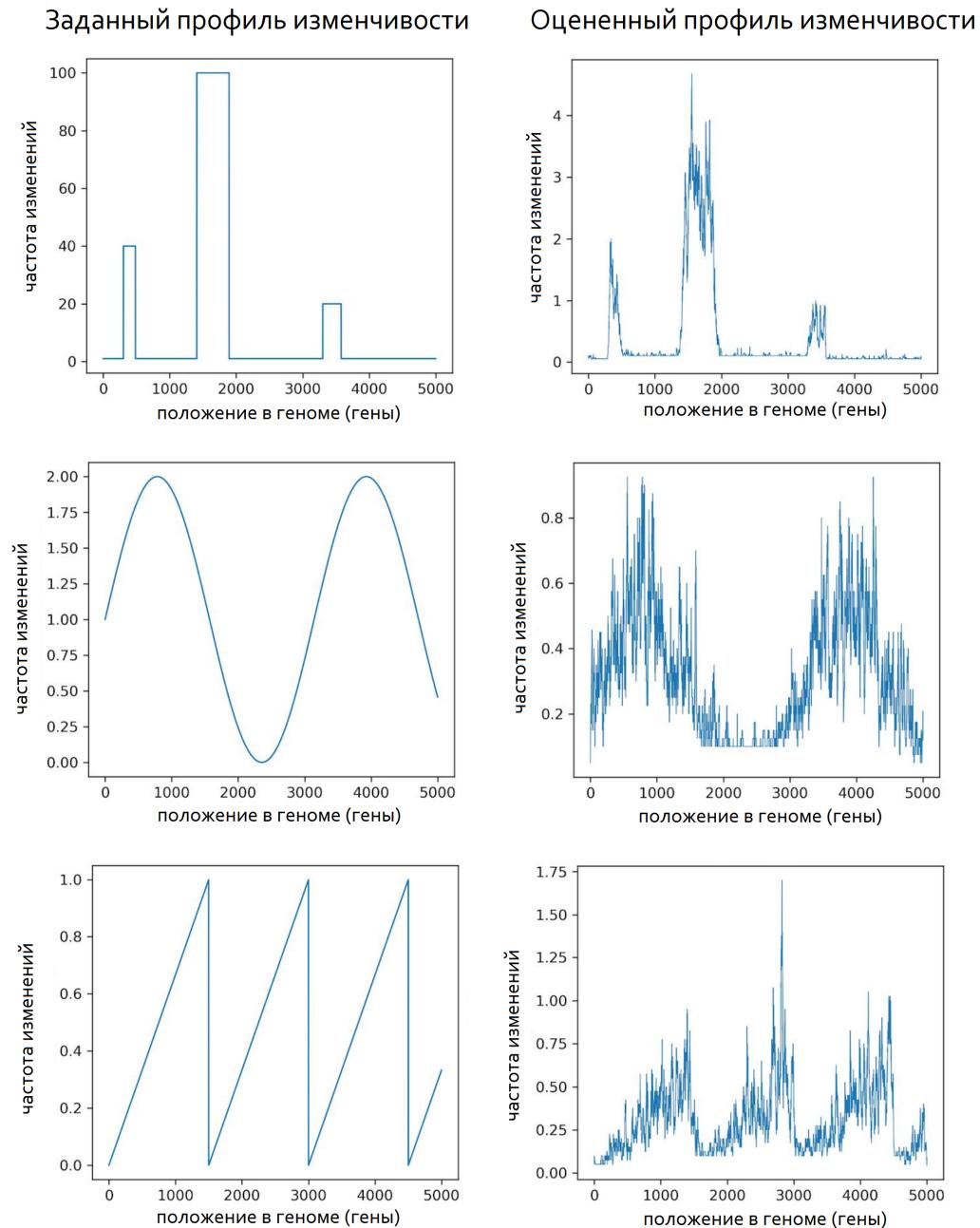


Рисунок 3.4 — Сравнение профиля изменчивости, задаваемого для моделирования геномных изменений (слева), и профиля, получаемого в результате анализа предложенным нами методом (справа).

### Сравнения результатов предложенного метода с ранее опубликованными оценками других авторов

В работе [138] авторы проводили поиск горячих точек горизонтального переноса генов (ГПГ) для 80 бактериальных видов. Мы провели сравнение найденных ими горячих точек с профилем изменчивости генома, оцененным опи-

Таблица 1 — Результаты сравнения задаваемого при моделировании и оцененного профиля изменчивости при разной форме профиля.

Тип профиля	Значение $R^2$	Коэффициент корреляции Спирмена
Ступенчатый	0.95	0.69
Синусоидальный	0.77	0.82
Пилообразный	0.80	0.85

санным выше способом. На рисунке 3.5 приведен пример сравнения результатов двух методов для генома *E. coli*. Видно, что, в данном случае, оба метода хорошо согласуются между собой.

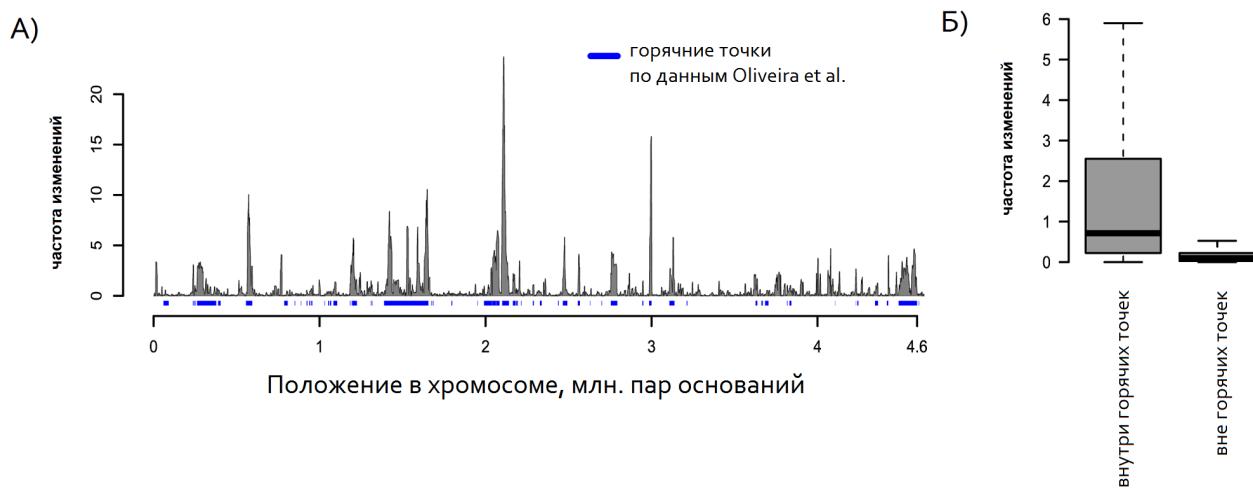


Рисунок 3.5 — Сравнение профиля изменчивости *E. coli* K12 MG1655, установленного нашим методом, с горячими точками горизонтального переноса генов в работе Oliveira et al. [138].

Для других видов бактерий, мы наблюдали как хорошо согласующиеся между собой, так и весьма непохожие результаты (например, в случае *Brucella melitensis* и *Burkholderia pseudomallei*). Расхождение может быть следствием: 1) малого количества геномов используемого для анализа в работе [138] (для многих видов оно было меньше 20, в среднем 12 геномов для одного вида), что значительно ухудшает соотношение сигнал/шум как нашего, так вероятно и используемого в работе [138] метода, 2) различием в подходах для оценки изменчивости (мы берем в рассмотрения все события, затрагивающие изменение порядка генов, не только горизонтальный перенос генов). Выяснение точной причины расхождений затруднено, поскольку авторами работы [138] не предоставляется программная реализация метода.

## 3.2 Исследование применимости метода оценки локальной вариабельности генома

### 3.2.1 Зависимость результатов от размера выборки

Чем больше геномов будет использовано для анализа, тем полнее будет получено представление о возможных вариантах со-расположения генов. Слишком маленькое количество геномов приведет к снижению чувствительности анализа и снизит соотношение сигнал/шум. В тоже время, слишком большое количество геномов потребует значительных вычислительных ресурсов и приведет к увеличению времени анализа. Поэтому, актуален вопрос о зависимости результатов анализа от размера выборки и определение минимально допустимого количества геномов.

На рисунке 3.6 показан график зависимости коэффициента корреляции профиля вариабельности, рассчитанного на основании 100 геномов, либо на основании выборки меньшего размера. Для анализа использованы финишированные геномы *Escherichia coli*. Для каждого размера выборки геномов проводилось 10 испытаний, для чего геномы выбирались случайным образом из полного набора.

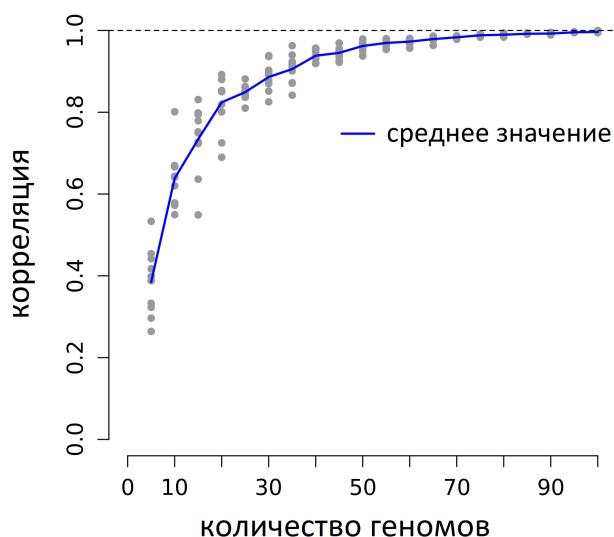


Рисунок 3.6 — Значение коэффициента корреляции профиля вариабельности, оцененной на основе 100 геномов с профилями оцененными на основе меньшего количества геномов. Синяя линия проведена по значениям корреляции, усредненным по 10 испытаниям.

При количестве геномов большем чем 50, значение коэффициента корреляции превышало уровень 0.9. В дальнейшем, мы ориентировались на эту оценку и брали в рассмотрение лишь те виды, для которых доступное количество геномов составляло не менее 50.

На рисунке 3.7 показана зависимость времени, необходимого для построения графа и анализа профиля вариабельности от количества геномов. Анализ проводился на основании финишированных геномов *E. coli*. Видно, что зависимость носит линейный характер. Анализ проводился в один поток на персональном компьютере с процессором i7 9750h. Наиболее продолжительным шагом общего анализа является этап построения групп гомологии. Так, для тысячи геномов, время анализа составило пять суток, анализ проводился в 50 потоков на вычислительном кластере на основе процессоров AMD EPYC 7502.

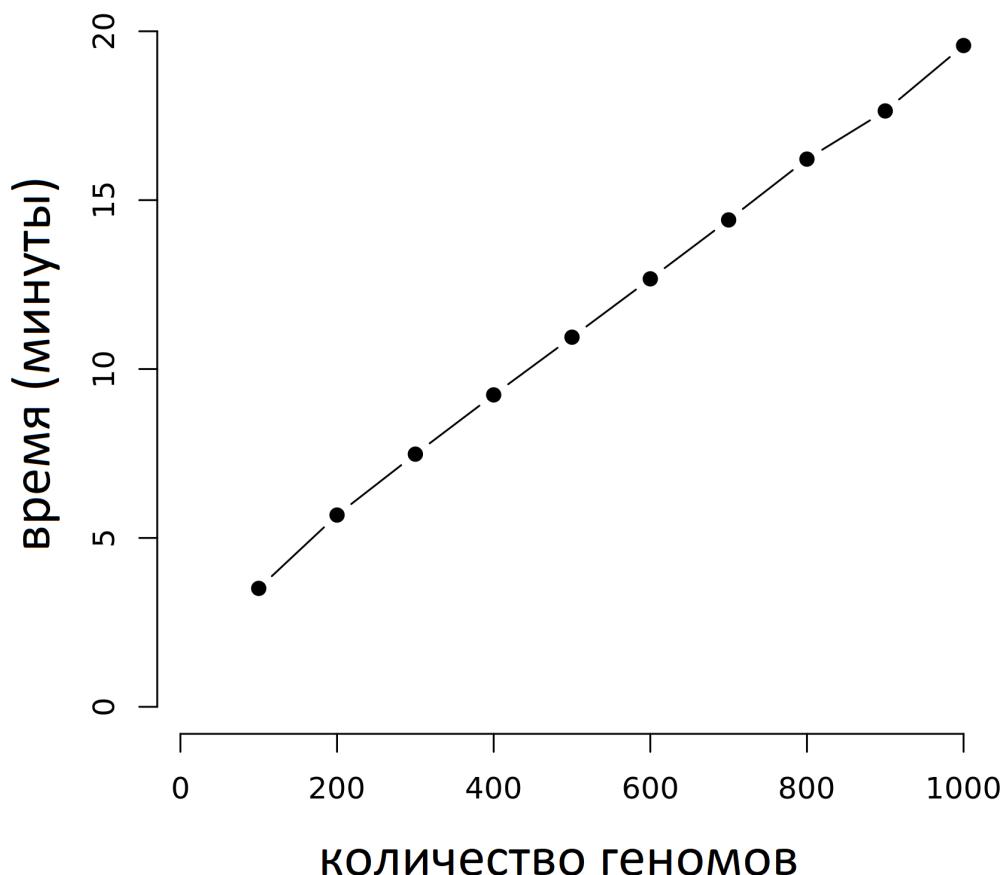


Рисунок 3.7 — Зависимость времени, необходимого для построения графа и расчета профиля изменчивости, от количества геномов. Для анализа использованы финишированные геномы *E. coli*.

### 3.2.2 Анализ зависимости результатов от качества сборки генома

Многие геномы прокариот собраны не до уровня репликонов, но состоят из контигов - фрагментов генома, которые удается собрать из коротких прочтений, без применения дополнительных экспериментальных методов. Для анализа чувствительности алгоритма к типу входных данных: финишированные (собранные до уровня репликонов), либо фрагментированные геномы (состоят из контигов), мы выбирали случайным образом по 100 геномов одного либо другого типа и оценивали профили вариабельности по одному и тому же референсному геному. Сравнение профилей по финишированным либо фрагментированным геномам для двух видов бактерий показано на рисунке 3.8. Видно, что профили обладают значительной степенью сходства (коэффициент корреляции Пирсона составил 0.87 и 0.81 для *E. coli* и *Pseudomonas aeruginosa*, соответственно).

Заметим, что значения профиля изменчивости зависят от набора геномов, используемых для сравнения. Мы выбирали случайные геномы в наборы финишированных и фрагментированных геномов, что также вносило вклад в наблюдавшее различия профилей. Вклад собственно качества сборки геномов еще меньше наблюдаемого выше уровня различий. Основываясь на этой оценке, мы считали, что финишированность геномов не является обязательным для включения геномов в анализ.

## 3.3 Применение метода оценки локальной вариабельности генома

### 3.3.1 Профиль вариабельности генома *E. coli*

На рисунке 3.9 показан профиль изменчивости генома *Escherichia coli* LF82, рассчитанный на основании графового представления 326 геномов других штаммов данного вида. Цветом обозначены жизненно-необходимые гены (мутации в которых летальны), гены транспортных и рибосомных РНК, острова патогенности и места встройки профагов, описанные для данного генома. Видно, что большая часть локусов с повышенной вариабельностью ("горячие точки" из-

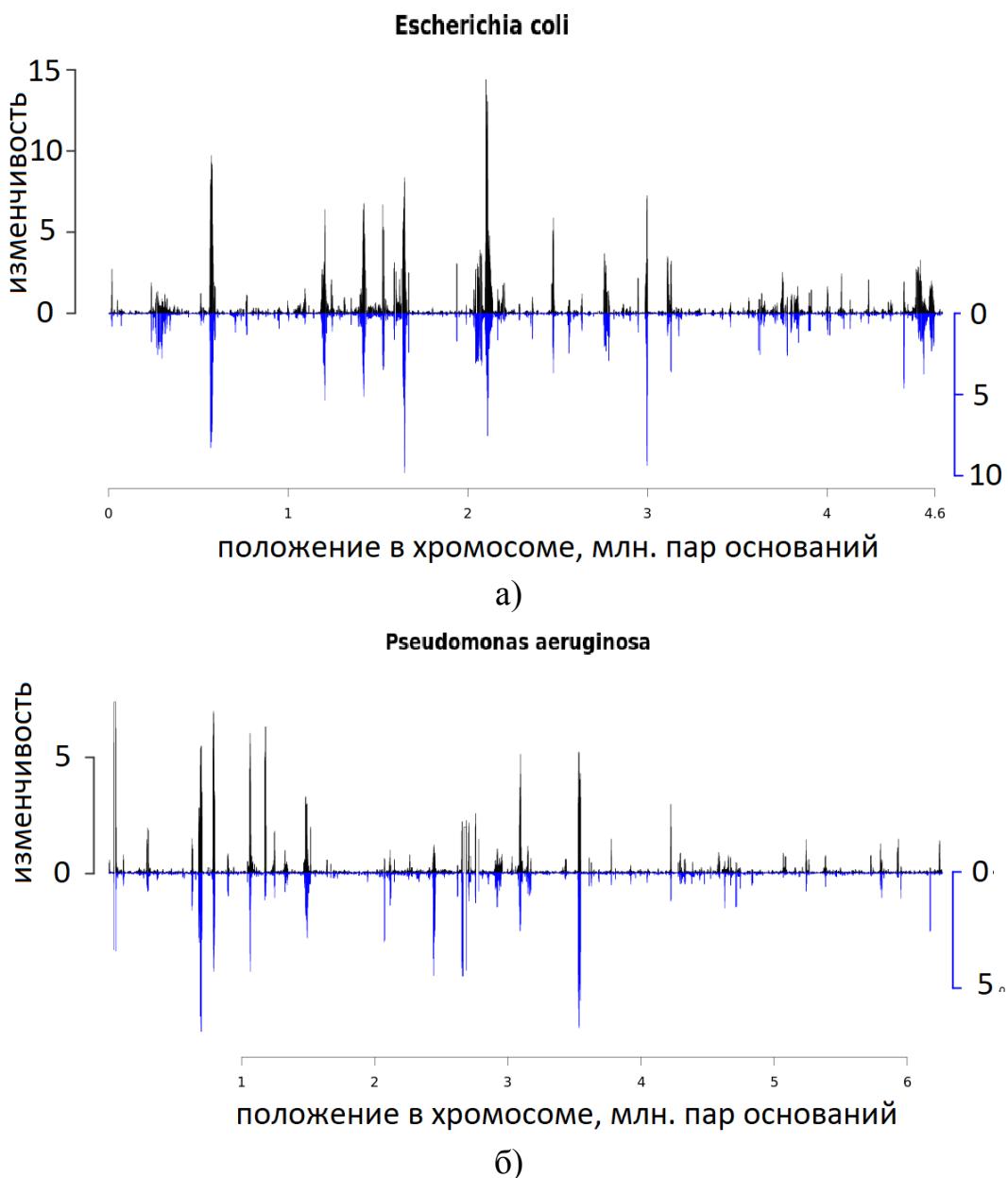


Рисунок 3.8 — Сравнение профилей вариабельности 100 финишированных и 100 фрагментарных геномов видов: а) *Escherichia coli*, б) *Pseudomonas aeruginosa*. Черным цветом показан профиль вариабельности на основе финишированных геномов, синим - на основе собранных до уровня контигов.

менчивости) находятся именно в тех местах, где у данного штамма находятся профаги, либо острова патогенности. Гены транспортных и рибосомных РНК не имеют очевидной связи с профилем вариабельности.

Отметим, что фрагмент генома с наибольшим уровнем вариабельности (расположен в области 2,115,791-2,164,382 п.о.) не содержит описанных детерминант мобильности ДНК ни в референсном геноме, ни в геномах других штаммов используемых для анализа. Наиболее консервативный набор генов в этом региона

следующий: имидазол-глицерин фосфат синтаза, 1-(5-фосфорибозил)-5-[(5-фосфорибозиламино) метилиденамино] имидазол-4-карбоксамид изомераза, имидазол-глицерин фосфатсинтаза, фосфорибозил-АМФ циклогидролаза, определяющий длину цепи белок (*chain length determinant protein*), UDP-глюкозо 6-дегидрогеназа, 6-фосфоглюконат дегидрогеназа, фосфатидил-мио-инозитол маннозил трансфераза, фосфатидил-мио-инозитол диманнозидсинтаза, фосфоманномутаза / фосфоглюкомутаза, маннозо-1-фосфатгуанилтрансфераза, альфа-D-канозаминилтрансфераза, ГДФ-манноза маннозилгидролаза, ГДФ-L-фукозо синтаза, ГДФ-манноза-4,6-дегидратаза, серин ацетилтрансфераза, белок биосинтеза липополисахаридов *WzxC*, N-ацетил-альфа-D-глюкозаминилдифосфо-дитранс октацис-ундекапренол-4-эпимераза, глюказилтрансфераза *WfgD*. Для его получения мы использовали алгоритм построения подграфа (будет описан ниже) и рассматривали только сочетания генов представленные в не менее двадцати геномах. Данные ферменты принимают участие в синтезе компонентов клеточной стенки. Причем внутри этого региона также содержится вариабельный участок, содержащий следующие гены: ацетилтрансфераза *EpsM*, гликозилтрансфераза *EpsE*, рамнозилтрансфераза *WbbL*, переносчик O-антитела и ряд гипотетических белков.

### 3.3.2 Сравнение профилей вариабельности филогрупп *E. coli*

Рассмотрим, как соотносятся профили изменчивости геномов из различных внутривидовых структур данного вида. Ценность подобного анализа состоит в возможности установить временную динамику "горячих" точек изменчивости (их устойчивость во времени). Также, подобное сравнение является еще одной проверкой предложенного метода: если у геномов из разных подвидовых структур не будет наблюдаться схожести в профиле вариабельности, то это ставит под сомнение ценность анализа для вида в целом. С другой стороны, воспроизводимость — схожесть профилей — будет свидетельством того, что получаемый при анализе профиль изменчивости отражает некоторый реальный биологический эффект.

Для проведения данного анализа мы использовали геномы из пяти наиболее крупных филогрупп (A, B1, B2, D, E) *E. coli*. Мы выбрали по одному референсному геному из каждой филогруппы; для каждого из референсных геномов

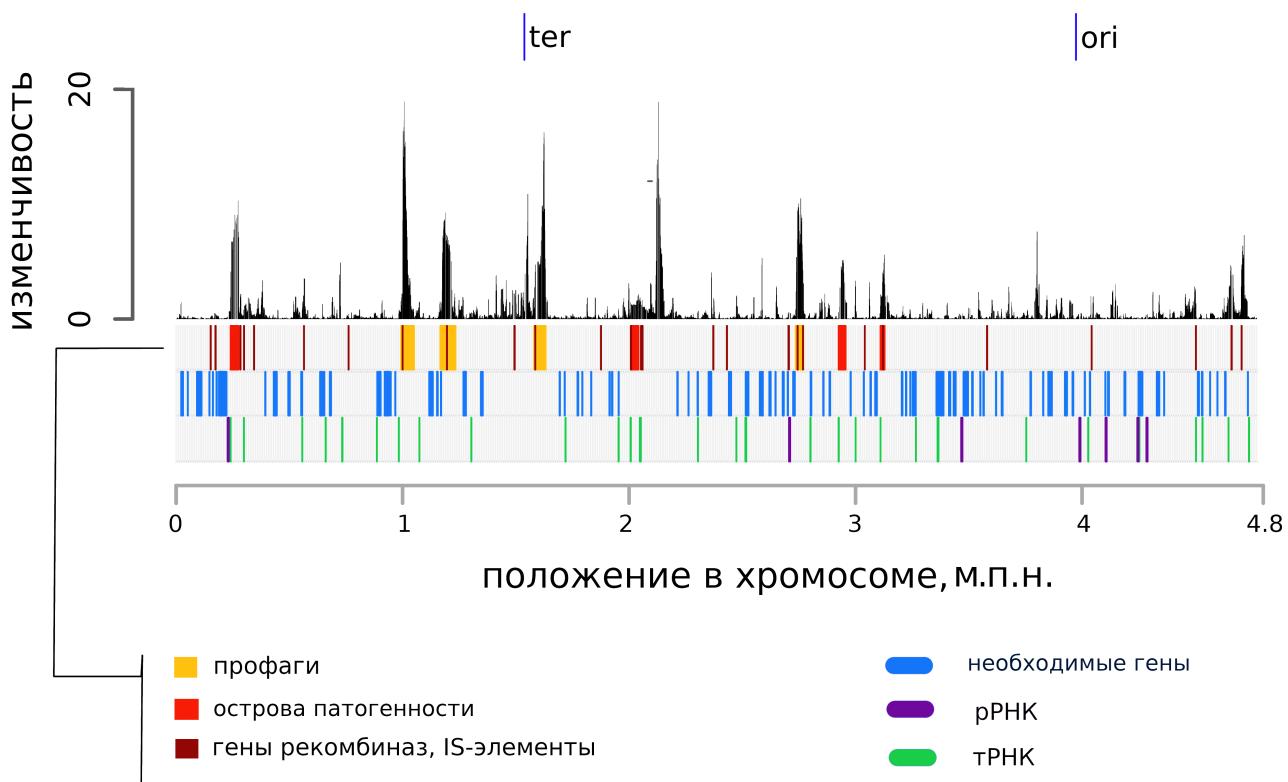


Рисунок 3.9 — Профиль вариабельности генома *Escherichia coli* LF82. Цветом обозначены острова патогенности, профаги, гены, ассоциированные с мобильными элементами генома, жизненно-необходимые гены, гены транспортных и рибосомных РНК. Области повышенной изменчивости содержат меньше жизненно-необходимых генов. Профаги и острова патогенности обладают повышенным уровнем изменчивости по сравнению с остальной частью генома.

подобрали 100 наиболее близких (по суммарной ширине выравнивания) геномов, доступных в базе RefSeq, в которой на момент проведения анализа содержалось 5466 геномов кишечной палочки; оценили профиль изменчивости внутри каждой из филогрупп независимо, используя только геномы данной филогруппы.

На рисунке 3.10 показано филогенетическое дерево подобранного таким образом набора геномов. На рисунке 3.11 показаны профили вариабельности по пяти геномам из различных филогрупп, серым цветом обозначены блоки синтезии, оранжевым - области нахождения профагов. Длины геномов значительно различаются: от 4.6 млн. п.о. в случае K12 до 5.5 м.п.н. у O157:H7 Sakai. Профили изменчивости геномов масштабированы таким образом, чтобы они совпадали по длине на рисунке; снизу от профилей изменчивости показаны координатные оси.

Как и в вышеописанном случае анализа 327 геномов кишечной палочки, анализ по отдельным филогруппам показывает, что большая часть областей с повышенной вариабельностью соответствует местам встройки фагов. Особенно это

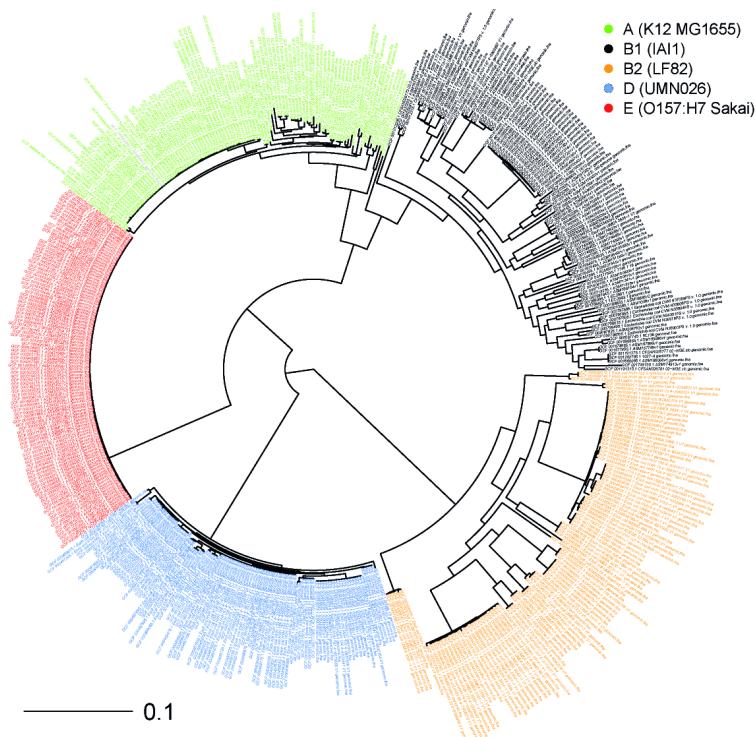


Рисунок 3.10 — Филогенетическое дерево выборки геномов *E. coli*, состоящей из представителей пяти крупных филогрупп: А, В1, В2, Д, Е (обозначены зеленым, черным, оранжевым, синим, красным цветом соответственно).

заметно в случае филогруппы Е (штамм *O157:H7 Sakai*), экспансия фагов в которой привела к значительному увеличению размера генома.

В целом, профили обладают высокой долей сходства: значительная часть областей повышенной вариабельности сохраняет свое расположение у части, либо во всех филогруппах. Это наблюдается как для профаговых областей (рис. 3.11 Б), так и для областей без признаков профагов и других мобильных элементов (рис. 3.11 В). Горячие точки могут иметь различную степень изменчивости в разных филогруппах (рис. 3.11 Б, Г, Д).

Заметно существование "холодных" областей генома, обладающих низкой вариабельностью во всех филогруппах. Длина этих низко-вариабельных участков может значительно превышать характерные длины оперонов, и достигать величин порядка миллиона пар оснований (например, область в окрестности 4 млн. п.о. на рисунке 3.11 А).

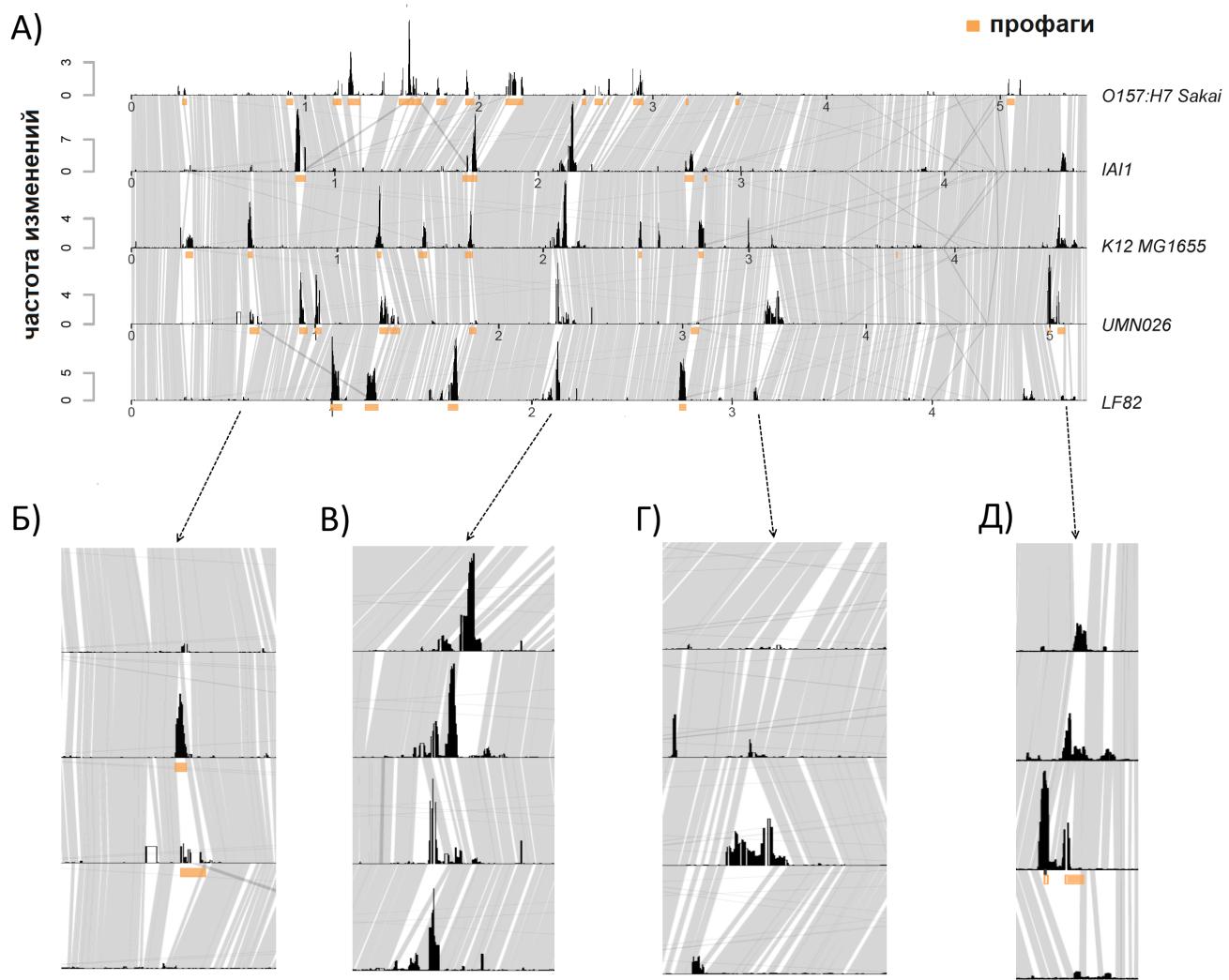


Рисунок 3.11 — Сравнение профилей вариабельности представителей пяти филогрупп *E. coli* LF82. Оранжевым цветом выделены области, определенные как профаговые. Блоками серого цвета показаны блоки синтезии.

### 3.3.3 Сравнение профилей вариабельности филогрупп у других видов

В данном разделе мы приводим результаты сравнения профилей изменчивости для филогрупп у других видов бактерий. Мы рассмотрели филогруппы двух видов рода *Pseudomonas* (*P. aeruginosa* и *P. fluorescens*), один из этих видов является естественно-компетентным, а также *Neisseria gonorrhoeae* - также естественно компетентного вида, для которого характерна высокая частота рекомбинационных событий [220].

*Pseudomonas* - род грамотрицательных бактерий, отдельные виды которого значительно различаются по метаболическому потенциалу и занимаемым

экологическим нишам. *P. fluorescens* обитают преимущественно в почве и являются естественно компетентными. *P. aeruginosa* - оппортунистические патогены человека, для которых компетентность наблюдалась лишь в условиях жизни в биопленке [221]. На рисунке 3.12 показаны филогенетические деревья для двух видов, цветом выделены клады дерева, отобранные для оценки и сравнения профилей изменчивости. Для *P. fluorescens* в первой кладе было 95 генома, и 115 геномов — во второй; для *P. fluorescens* в первой кладе было 73 генома и 143 генома — во второй.

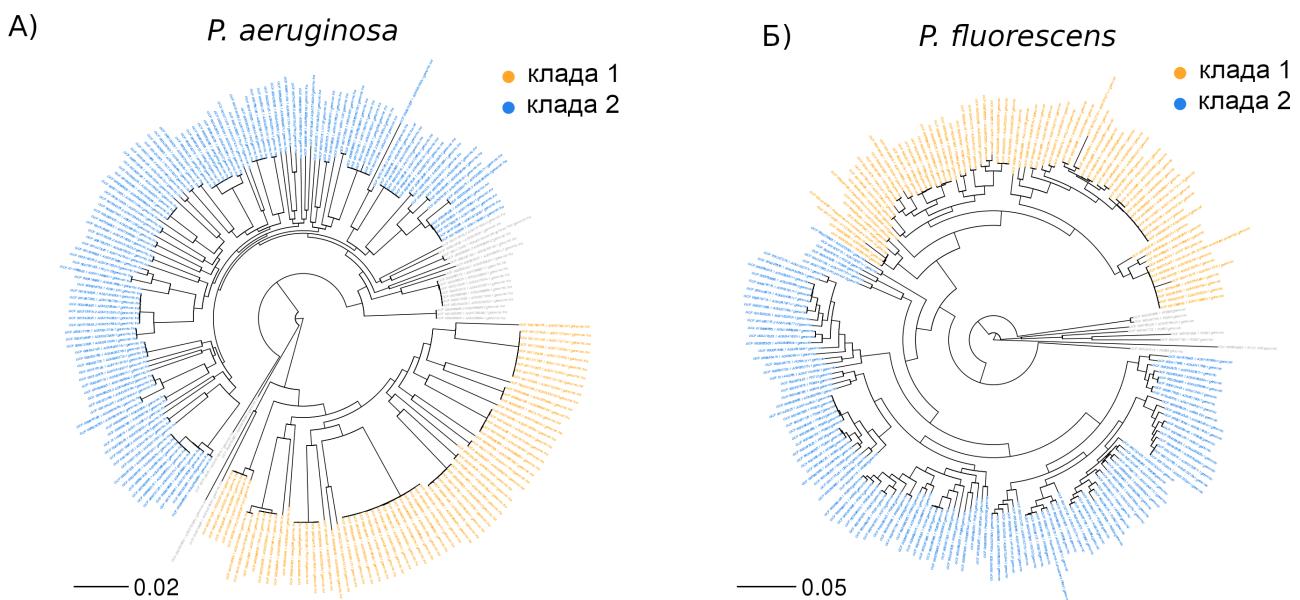
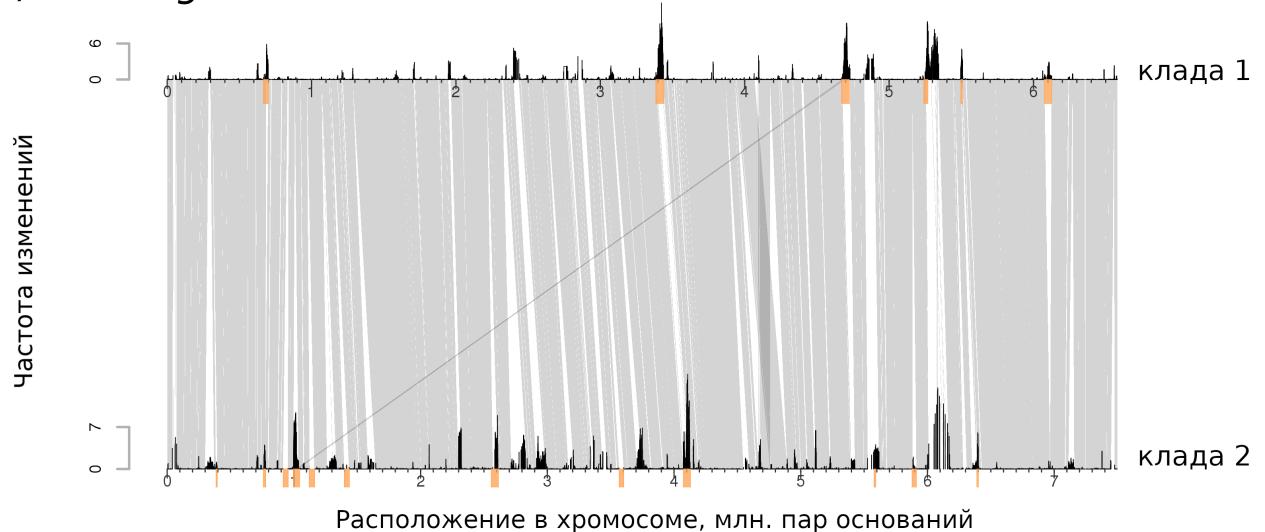


Рисунок 3.12 — Филогенетическое дерево геномов видов: А) *P. aeruginosa* и Б) *P. fluorescens*. Оранжевым и синим цветом выделены клады дерева, отобранные для анализа профилей изменчивости.

На рисунке 3.13 показано сравнение профилей изменчивости, оцененных независимо для геномов из различных клад филогенетического дерева. В случае *P. aeruginosa*, подобно *E. coli*, наблюдается значительное сходство профилей изменчивости геномов в двух филогенетических кладах (рис. 3.13 А). В случае *P. fluorescens* подобное сходство выражено слабо и заметно только в небольшом фрагменте ближе к концу хромосомы, после 6 млн. пар оснований, а сами профили вариабельности более равномерны и обладают меньшим количеством выраженных горячих точек (рис. 3.13 Б). Вероятная причина наблюдаемой разницы - высокая частота крупных хромосомных перестроек у *P. fluorescens* (заметно, в том числе, по визуализации блоков синтезии). Частые хромосомные перестройки делают анализ локальной изменчивости мало информативным, для применимости метода необходимо: во-первых, наличие локальных участков

повышенной изменчивости, и во-вторых, чтобы глобальные перестройки хромосомы происходили сравнительно редко относительно локальных изменений (вставки, делеции и транслокации отдельных генов).

### А) *P. aeruginosa*



### Б) *P. fluorescens*

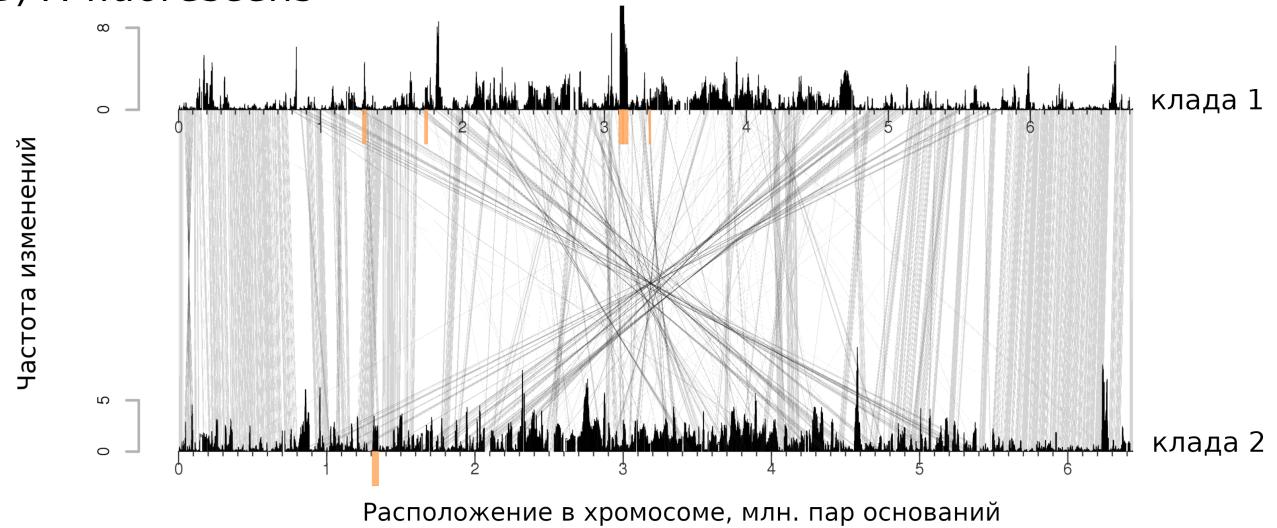


Рисунок 3.13 — Сравнение профилей вариабельности двух клад вида А) *P. aeruginosa* и Б) *P. fluorescens*. Оранжевым цветом выделены области, определенные как профаговые.

Для *N. gonorrhoeae* мы рассмотрели 4 клады. Количество геномов составило 49, 51, 47 и 75 геномов для первой, второй, третьей и четвертой клады, соответственно. Профили геномов из различных клад филогенетического дерева у обладают высоким уровнем сходства (рис 3.14).

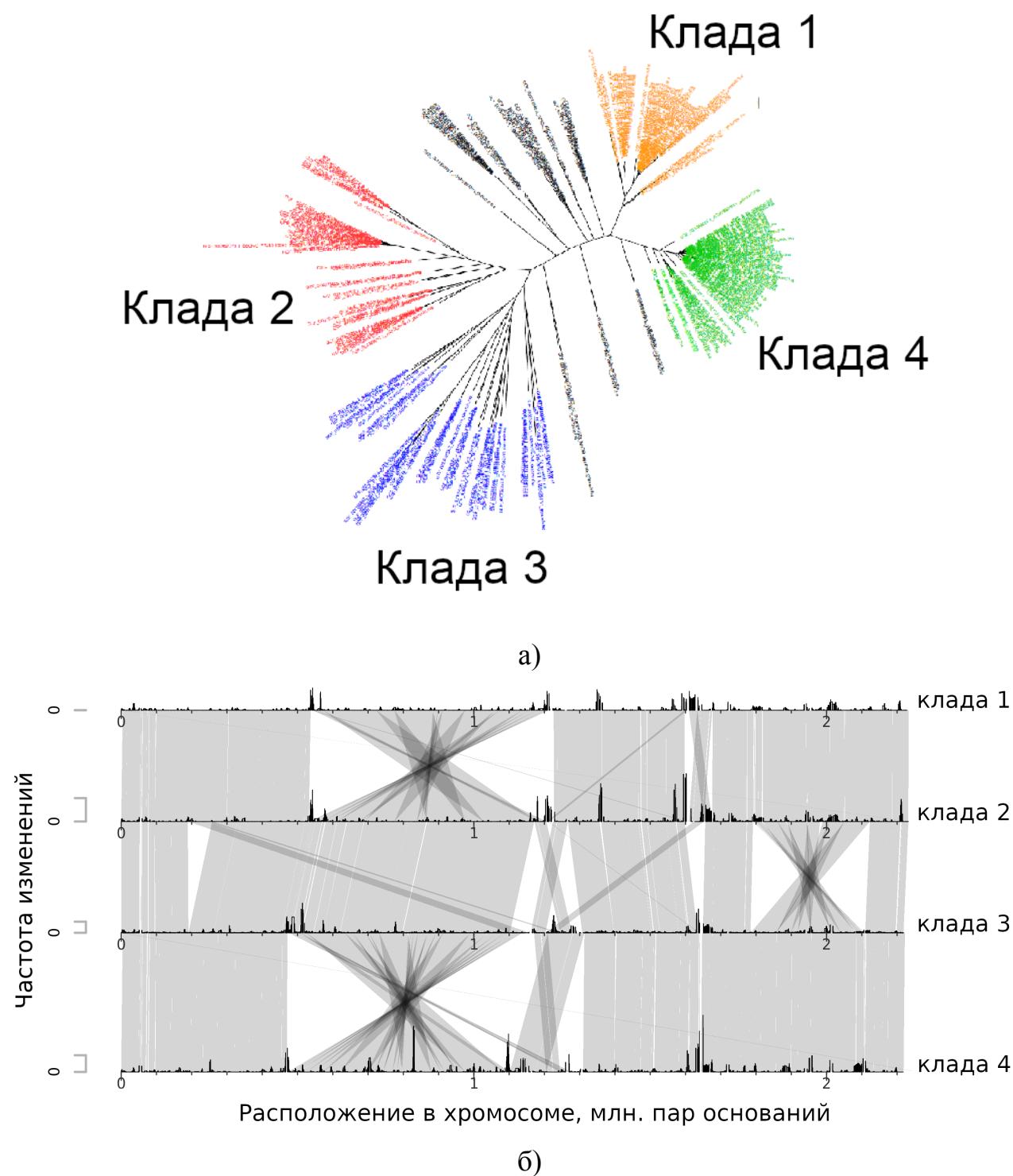


Рисунок 3.14 — Сравнение профилей вариабельности четырех филогенетических клад вида *N. gonorrhoeae*. А)

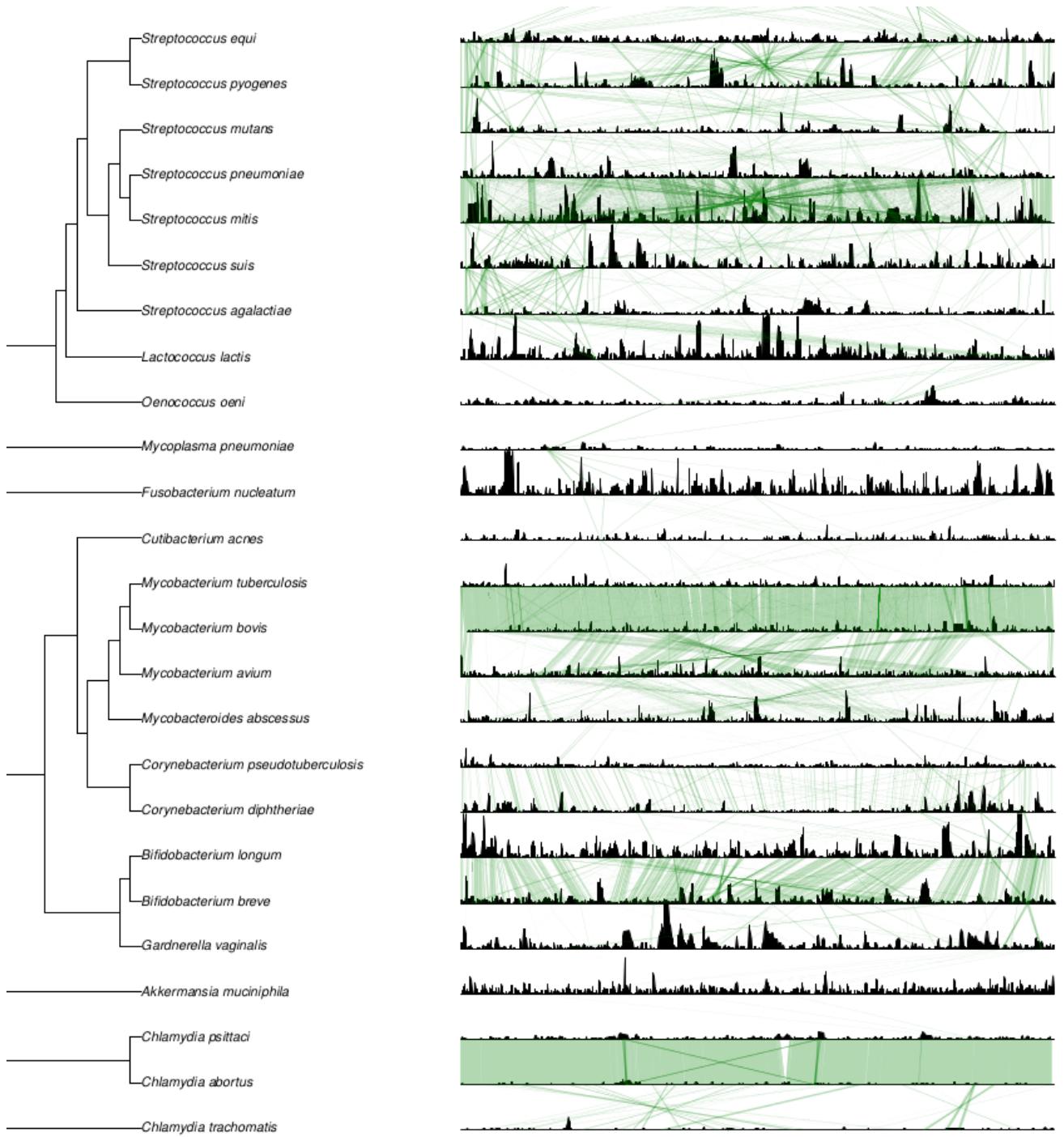


Рисунок 3.15 — Фрагмент визуализации профилей изменчивости геномов различных видов. Профили расположены в соответствии с филогенетическим деревом видов, блоки синтезии показаны для смежных по дереву видов.

### 3.3.4 Сравнение профилей вариабельности между близкородственными видами

Внутри вида профили изменчивости обладают высокой степенью сходства. Для выяснения степени сходства профилей изменчивости для геномов

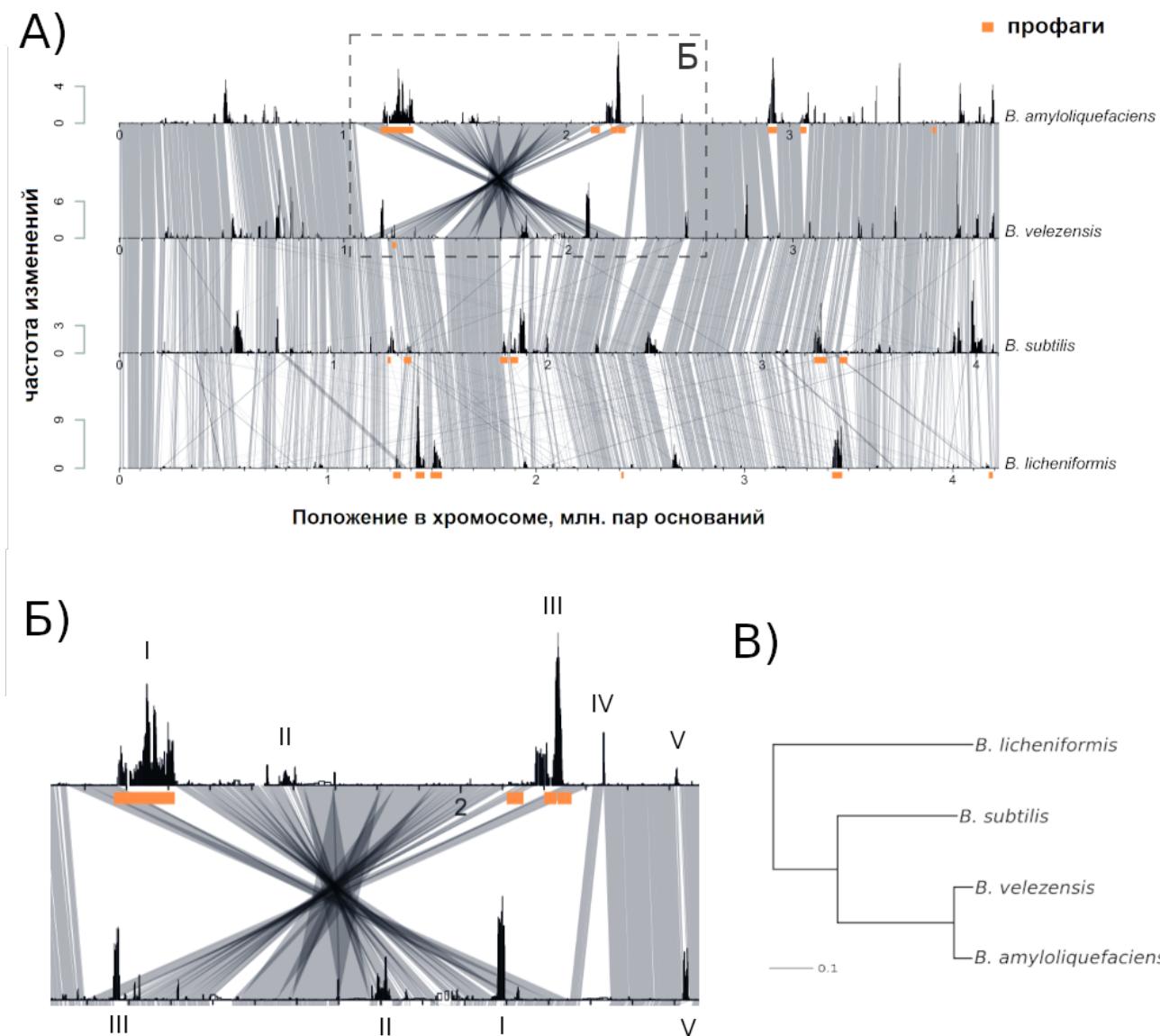


Рисунок 3.16 — Сравнение профилей вариабельности четырех различных видов рода *Bacillus*. А) Профили изменчивости и блоки синтезии, оранжевым цветом выделены области, определенные как профаговые, рамка из штрихованных линий показывает фрагмент, представленный на Б). В) Филогенетическое дерево рассматриваемых видов.

из разных видов, мы провели анализ изменчивости у 143 видов бактерий и архей. В рассмотрение мы брали те виды, для которых было доступно как минимум 50 геномов на момент начала анализа (2016 год). Визуализация профилей доступна на веб-сервере [gcb.rcpcm.org](http://gcb.rcpcm.org), о котором будет подробнее рассказано ниже. Все профили изменчивости были расположены в соответствии с филогенетическим деревом, блоки синтезии рассчитывались между всеми парами видов, расположеннымми последовательно на дереве. Фрагмент итоговой визуализации показан на рисунке 3.15, полная визуализация доступна по адресу

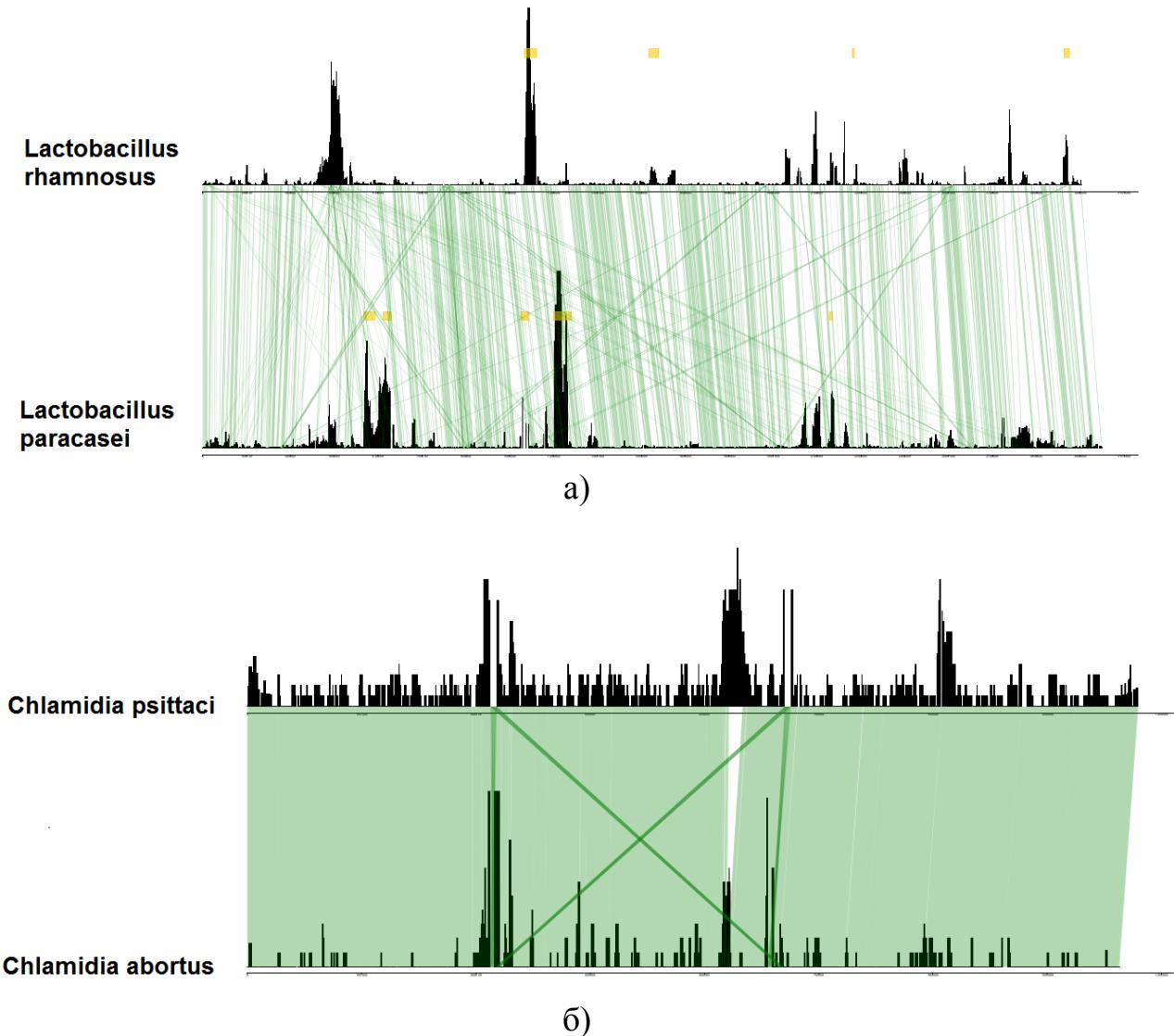


Рисунок 3.17 — Сравнение профилей вариабельности для: А) двух видов рода *Lactobacillus*, Б) двух видов рода *Chlamidia*. Оранжевым цветом выделены области, определенный как профаговые.

сy [https://github.com/paraslonic/GCB\\_revision/blob/master/figures/S3\\_Fig.pdf](https://github.com/paraslonic/GCB_revision/blob/master/figures/S3_Fig.pdf).

В большинстве случаев, когда между видами наблюдается малое количество геномных перестроек, профили изменчивости похожи (горячие точки находятся в схожем окружении). Ниже, для примера, мы приводим отдельные сравнения профилей изменчивости.

На рисунке 3.16 А показано сравнение профилей изменчивости четырех видов рода *Bacillus*, филогенетическое дерево для рассмотренных видов показано на рисунке 3.16 В. У геномов видов *B. amyloliquefaciens* и *B. velezensis* заметна крупная инверсия (рис. 3.16 А, выделена рамкой), при этом, области высокой

изменчивости сохранили свою активность (рис 3.16 Б, пики изменчивости обозначены римскими цифрами).

На рисунке 3.17 показано сравнение профилей изменчивости для двух видов лактобацилл и двух видов хламидий.

### 3.4 Связь между уровнем изменчивости и характеристиками генома

Профили изменчивости для геномов из различных филогрупп и близкородственных видов показали значительную степень сходства: наблюдается множество "горячих" точек изменчивости со стабильным положением в хромосоме. Открытым остается вопрос о причинах этого постоянства положения зон вариабельности, а также факторов определяющих их возникновение, рост либо снижение активности. Ниже мы приводим результаты сравнения между профилем изменчивости и расположением сайтов Chi, а также частотой межхромосомных контактов.

#### 3.4.1 Связь с распределением сайтов Chi.

Процесс горизонтального переноса генов подразумевает проведение одного из типов рекомбинации. Как было отмечено в обзоре литературы, у бактерии *E. coli* описаны сайты Chi, играющие важную роль в инициации процесса гомологичной рекомбинации. На рисунке 3.18 А показан уровень локальной вариабельности и локализация Chi-сайтов для хромосомы штамма *E. coli* LF82. Из приведенного рисунка видно неравномерное распределение сайтов Chi по реплихорам хромосомы и их сниженная представленность в местах локализации профагов/островов патогенности, а также в областях с высокой вариабельностью. На рисунке 3.18 б) показано сравнение уровня вариабельности и расстояний до ближайшего сайта Chi (на соответствующей цепи, лидирующей для правой реплихоры, и отстающей - для левой). Корреляция между уровнем вариабельности и расстоянием до ближайшего Chi-сайта значима ( $p\text{-value} < 1e-16$ ), но значение коэффициента корреляции невелико и составляет 0.25.

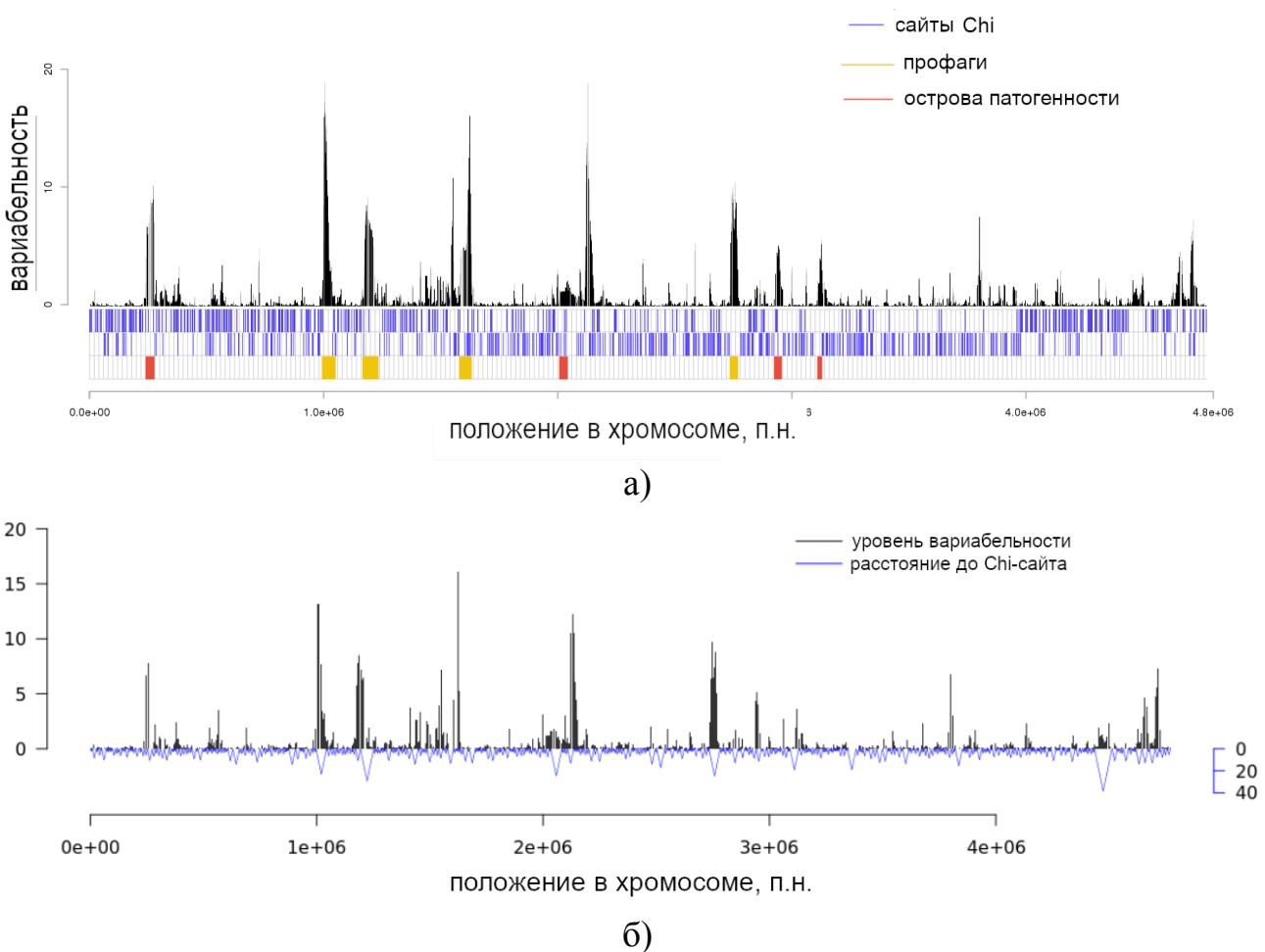


Рисунок 3.18 — Уровень вариабельности и локализация сайтов Chi на двух цепях хромосомы *E. coli* LF82. А) Синие вертикальные линии показывают расположение сайтов Chi на двух цепях хромосомы. Вертикальные пунктирные линии показывают границы реплихор (места начала и конца репликации). Желтым и красным цветом обозначены области расположения профагов и островов патогенности соответственно. Б) Черной линией показан уровень вариабельности, синей - расстояние до ближайшего сайта Chi (в масштабе тысяч пар оснований).

Эффект более низкой представленности сайтов Chi в горизонтально перенесенных фрагментах был описан ранее[42]. Вероятно, связь между вариабельностью и расстоянием до ближайшего сайта Chi объясняется тем, что вариабельные участки содержат много горизонтально перенесенных генов.

### 3.4.2 Связь с плотностью хромосомных контактов

На рисунке 3.19 показано сопоставление профилей вариабельности с матрицами частот межхромосомных контактов для бактерий *E. coli* и *B. subtilis*. Очевидной связи между профилем вариабельности и частотой хромосомных контактов, с нашей точки зрения, не наблюдается.

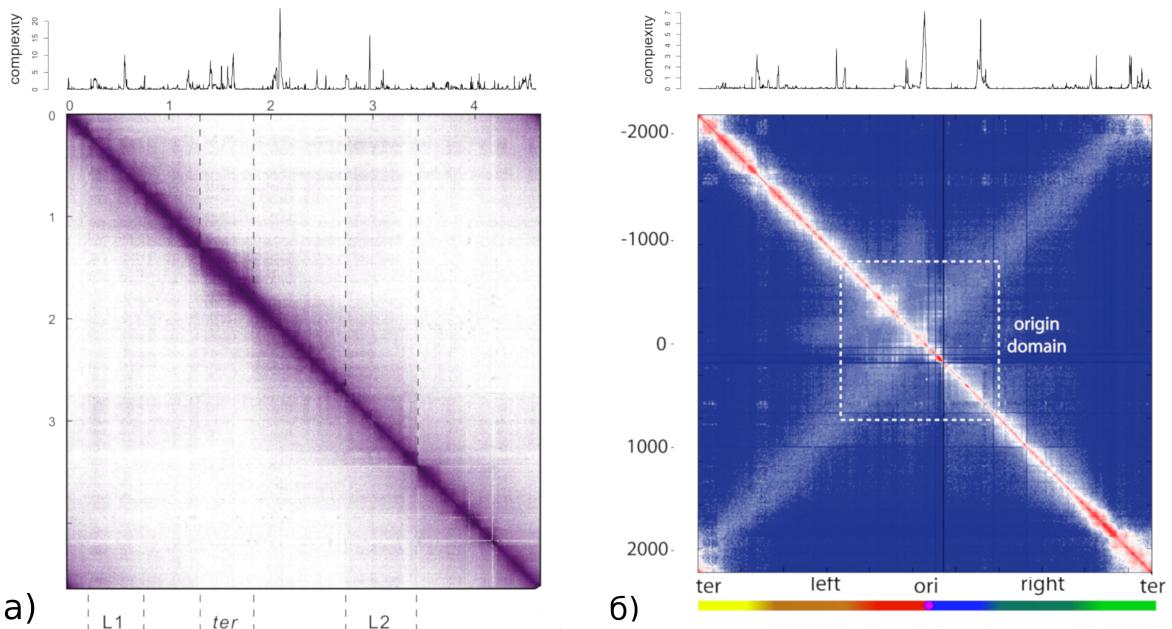


Рисунок 3.19 — Сравнение профилей вариабельности с нормированными матрицами хромосомных контактов для *E. coli* и *B. subtilis*

В тоже время, признаки существования подобной связи имеются, при сравнении профиля вариабельности с шкалограммой, построенной на основе матрицы контактов (метод построения шкалограмм описан в [54]), см. рисунок 3.20.

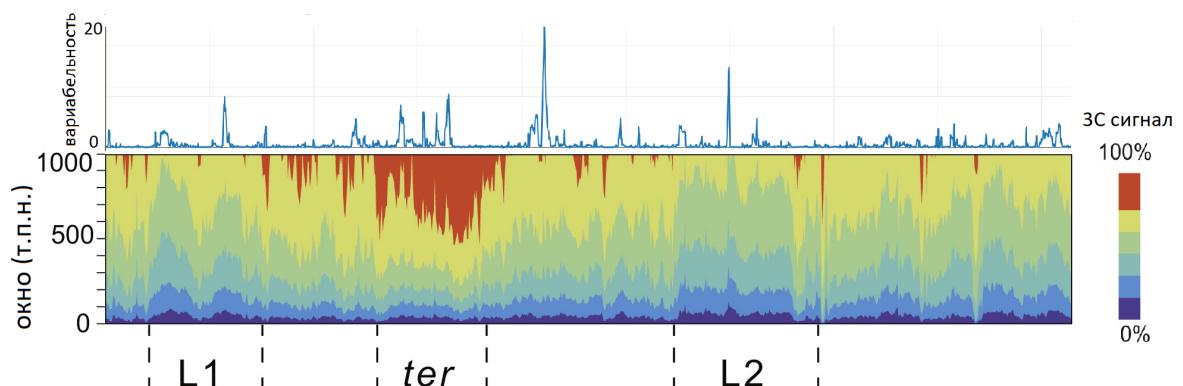


Рисунок 3.20 — Сравнение профиля вариабельности с шкаллограммой хромосомных контактов из публикации [54].

Для оценки статистической значимости связи мы использовали линейные модели, связывающие плотность контактов в фиксированном диапазоне длин и значение вариабельности. Плотность контактов  $S$  для гена с индексом  $i$  считалась как сумма значений нормированной матрицы контактов  $M$ :  $S_i = \sum_{j=1}^J M[i, i+j]$ . Далее, мы строили линейные модели для связи  $V_i \sim S_i$ , где  $V$  - это уровень изменчивости. Нами была выявлена статистически значимая зависимость между уровнем изменчивости генома и суммой межхромосомных контактов, как для бактерий *E. coli*, так и *B. subtilis*.

На рисунке 3.21 приведено значение p-value для линейных моделей, связывающих значение между уровнем вариабельности генома (рассчитанным с размером окна в 20 генов) и суммой контактов, рассчитанных в окнах различного размера. В случае *E. coli* наибольшее соответствие достигается при размере окна суммирования равному 15 тысяч п.о., а в случае *B. subtilis* — 10 тысяч п.о.. Это различие согласуется с различием в средней длине генов: для *B. subtilis* она равна 867 нуклеотидам, а для *E. coli* — 929 нуклеотидам.

Статистическая значимость сохраняется также и при удалении главной и второстепенной диагоналей в матрице контактов (рекомендации сотрудников MIT, <https://www.biostars.org/p/208512/>). Так, для *E. coli* значимость связи между вариабельностью и плотностью контактов в пределах 20 тысяч п.о. составляет  $p - value = 1.23 * 10^{-7}$ ). Значимость ранговой корреляции при этом составила  $p - value < 2.2 * 10^{-16}$ . Хотя связь между уровнем вариабельности и плотностью хромосомных контактов статистически значима, она объясняет лишь малую долю вариабельности вариабельности. Так, значение  $R^2$  для *E. coli* составляет 0.04, а для *B. subtilis*  $R^2 = 0.06$ .

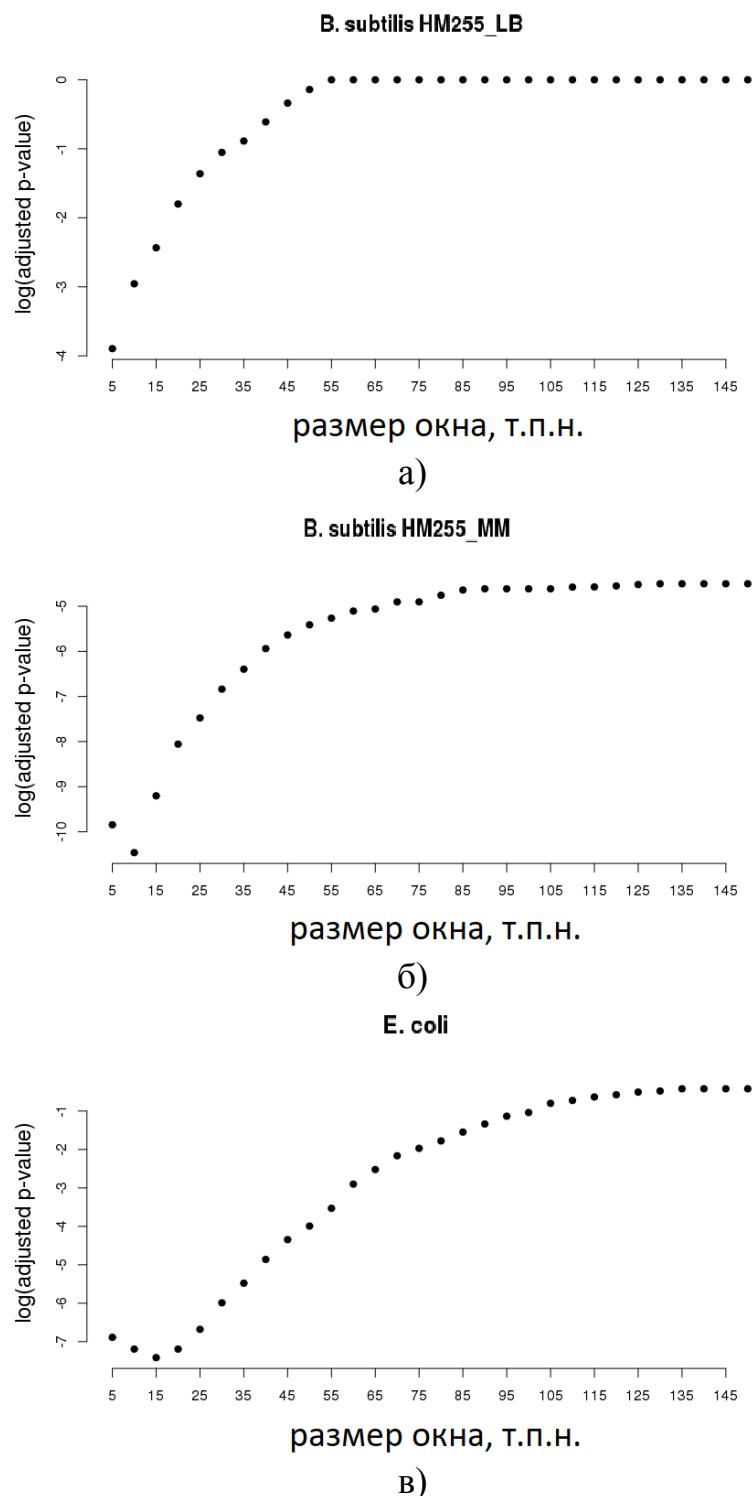


Рисунок 3.21 — Зависимость уровня статистической значимости (значение  $p$  — value) связи между уровнем изменчивости генома и плотностью хромосомных контактов. Размер окна (в тысячах пар нуклеотидов), по которому производилось суммирование значений нормированной матрицы хромосомных контактов, отложен по оси абсцисс. Десятичный логарифм значения  $p$ -value линейной модели отложен на оси ординат. На А и Б приведены значения для *B. subtilis* в условиях роста на богатой среде и бедной среде, соответственно; на В приведены значения для *E. coli*.

### **3.5 Разработка и применение метода анализа локальной вариабельности при помощи построения подграфов**

Работа выполнена совместно с Конановым Дмитрием Николаевичем.

#### **3.5.1 Проблема сравнительного анализа расположения генов в большом наборе геномов.**

Визуализация блоков синтезии часто применяется для сравнения генного состава оперонов, островов патогенности либо других областей генома, интересующих исследователя. Такой способ визуализации нагляден и хорошо подходит для небольшого количества геномов. Но сравнение нескольких десятков последовательностей затруднительно, а нескольких сотен — практически невозможно. При этом, количество прочитанных геномов быстро растет, для многих видов их уже доступно несколько сотни и даже тысяч. Анализ геномов в подобных масштабах требует новых подходов. Мы предлагаем в качестве такого подхода графовое представление порядка расположения генов в геномах. Выше мы описали применение графового представления для численной оценки локальной вариабельности геномов. Ниже мы опишем применение графового представления для визуализации изменений в небольших фрагментах генома.

#### **3.5.2 Алгоритм поиска подграфа для анализа участка генома**

Визуализация полного графа набора геномов допустима в случае небольших геномов, характерных для вирусов, но не в случае бактерий — полный граф будет слишком велик для эффективной визуализации. Для бактериальных геномов имеет смысл проводить анализ подграфа - части полного графа, соответствующей некоторому региону интереса (например, оперону).

Для построения подграфа мы реализовали следующий набор действий. Вначале, выбирается референсный геном и указывается начало и конец анализируемой области. Затем, строится граф, содержащий цепочку узлов референсного генома и к нему добавляются узлы, представленные в других геномах и связанные с референсной цепочкой. Добавление узлов происходит пока путь снова не вернется в референсную цепочку, либо пока не будет достигнут предел на длину (параметр *tails*). Алгоритм для построения подграфа можно представить следующим псевдокодом:

1. Вход: graph (граф группы геномов)
2. Параметры: *reference* (референсный геном), *start\_node* (начало анализируемого фрагмента), *end\_node* (конец анализируемого фрагмента), *max\_depth* (ограничение на длину обходных путей), *tails* (ограничение на длину свободный путей), *minimal\_edge\_weight* (минимальный вес ребра).
3. Результат: subgraph (подграф интересующего фрагмента генома)
4.  $\text{subgraph} \leftarrow$  пустой граф
5.  $\text{target\_chain} \leftarrow$  цепочка узлов из референсного генома между *start\_node* и *end\_node*
6. добавить *target\_chain* в *subgraph*
7.  $\text{deviating\_paths} \leftarrow$  найти все пути начинающиеся и заканчивающиеся в *target\_chain*
8. для каждого пути *path* из *deviating\_paths* выполнить
9. если длина *path* < *max\_depth* то
10. добавить *path* в *subgraph*
11. иначе
12.  $\text{path\_tails} \leftarrow$  начальный и конечный фрагмент *path* длиной *tails*
13. добавить *path\_tails* в *subgraph*
14. для каждого ребра *edge* из *subgraph* выполнить
15. если вес *edge* < *minimal\_edge\_weight* то
16. удалить *edge* из *subgraph*
17. удалить области связности *subgraph* не связанные с *target\_chain*
18. вернуть *subgraph*

Данный алгоритм, реализованный на языке Python3, доступен в репозитории: [https://github.com/DNKonanov/gene\\_graph\\_lib](https://github.com/DNKonanov/gene_graph_lib)

В приведенном выше алгоритме предусмотрено два фильтра, которые помогают уменьшать размер подграфа, если он оказывается слишком велик. Даже при

рассмотрении небольшого участка генома, соответствующий ему подграф может содержать длинные пути, например, из-за крупных хромосомных перестроек в рассматриваемых геномах. Для того, чтобы эти длинные пути не затрудняли визуализацию и последующий анализ, в алгоритме предусмотрен фильтр, который заменяет длинный путь его начальным и концевым фрагментом ("усами"). Максимальная длина пути, который будет показан полностью, задается параметром *max\_depth*, размер фрагментов, которые сохраняются вместо длинного пути (длина "усов") задается параметром *tails* (рисунок 3.22). Пути, которые начинаются либо заканчиваются, но не начинаются и заканчиваются, в рассматриваемой области ("уходящие" за пределы рассматриваемой области) также сокращаются до фрагментов длиной *tails*.

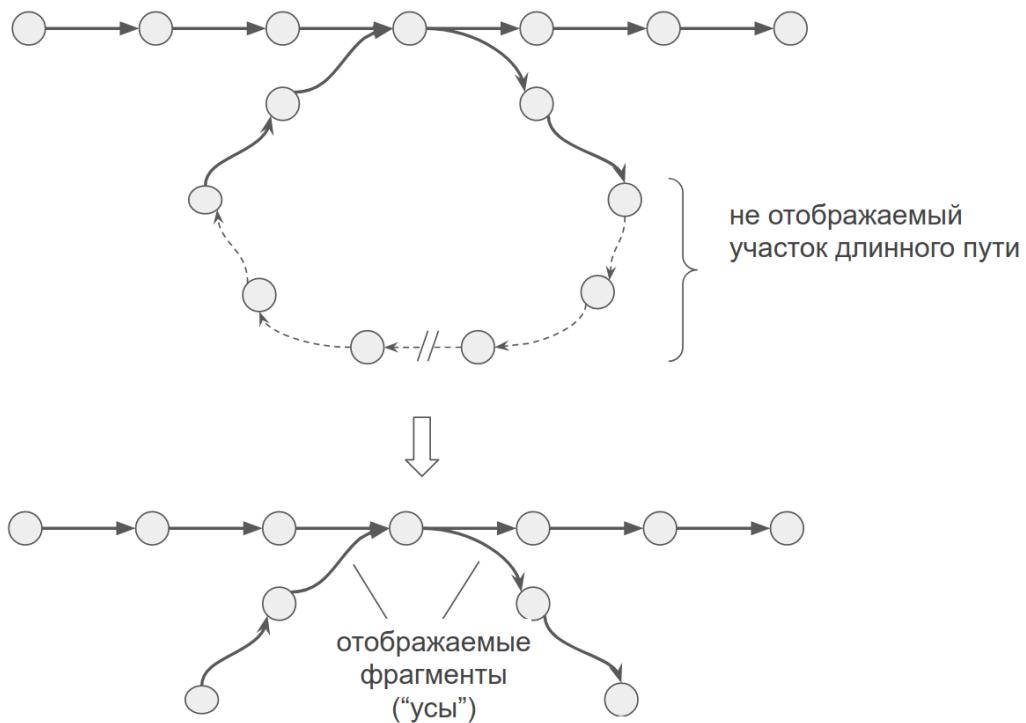


Рисунок 3.22 — Схематическая иллюстрация работы фильтра длинных путей.

Еще один фильтр позволяет не включать в подграф ребра с маленьким весом, то есть те, которые встречаются в малом количестве геномов. За это отвечает параметр *minimal\_edge\_weight*. Применение фильтров особенно важно при анализе "горячих" точек изменчивости.

### 3.6 Примеры применения представления порядка чередования генов в виде графа.

В работе [193], были установлены опероны, которые статистически значимо чаще (метод определения будет описан ниже) встречались у изолятов *E. coli* полученных от пациентов с воспалительным заболеванием кишечника — болезнью Крона, по отношению к изолятам от здоровых людей. Носительство данных оперонов, вероятно, выгодно при нахождении бактерий в условиях воспалительной реакции со стороны организма хозяина, а может и провоцирует воспаление. Метод для поиска оперонов, различающих группы бактерий, будет описан ниже.

Рассмотрим, как выглядят подграфы, соответствующие оперонам, чья встречаемость оказалась выше в изолятах, полученных от пациентов с болезнью Крона, по сравнению с комменсальными штаммами. На этих примерах будут проиллюстрированы основные моменты анализа графового представления фрагментов генома.

#### Опероны утилизации гемина и пропандиола

На рисунках 3.23 и 3.24 показаны графы, построенные в окрестностях оперонов захвата гемина (hemin uptake, hmu) и утилизации пропандиола (propanediol utilization operon, pdu), соответственно. В качестве референсного генома нами был взят геном *Escherichia coli* LF82, в анализ были включены 327 финишированных геномов доступных в базе RefSeq. Для построения данных графов мы использовали следующие параметры:  $tails = 1$ ,  $max\_depth = 30$ ,  $minimal\_edge\_weight = 5$ .

Рассмотрим для начала более простой случай оперона захвата гемина. Данный оперон состоит из следующих генов: транспортный белок (Hemin transport protein HemS), АТФ-связывающий белок захвата гемина (Hemin import ATP-binding protein HmuV), гемин-связывающий периплазматический белок (hemin-binding periplasmic protein HmuT), кислороднезависимый копропорфирина-III оксидазоподобный белок (oxygen-independent coproporphyrinogen-III oxidase-like protein) и два гипотетических белка (здесь и далее мы приводим аннотации последовательностей, полученные при помощи утилиты prokka [210]).

Как видно из графа, представленного на рисунке 3.23, данный оперон расположен в консервативном генном контексте: слева расположен регулятор транскрипции НTH-типа, справа — гипотетический белок. Дуговое ребро обходя-

щее оперон сверху говорит о том, что в некотором наборе геномов данный оперон отсутствует и других вариантов генов у них в этом локусе не наблюдается.

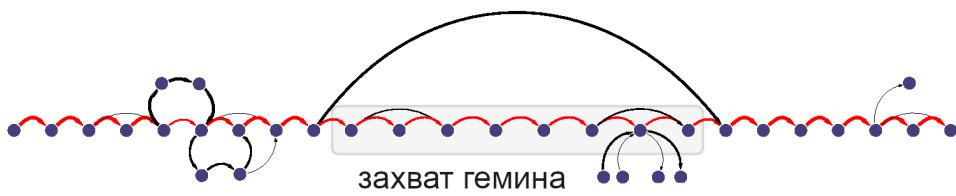


Рисунок 3.23 — Граф представляющий окрестность оперона утилизации гемина (hemin uptake, hmu).

В оперон утилизации пропандиола (propanediol utilization operon, pdu) входят гены, кодирующие: белок утилизации пропандиола PduU, малая субъединица пропандиол-дегидратазы, белок утилизации пропандиола PduB, белок с механизмом концентрации углекислого газа CcmK, субъединица альфа-фактора реактивации диолдегидратазы, белок утилизации пропандиола PduV, альдегид-алкоголь дегидрогеназа, С-диамид аденоцил трансфераза ириновой кислоты Cob I, большая субъединица пропандиол дегидратазы, фосфат-пропаноил трансфераза, альдегид-алкоголь дегидрогеназа. Как видно из рисунка 3.24 наблюдается консервативность контекста расположения pdu оперона у тех штаммов, в которых он представлен: слева от него расположен ген, кодирующий белок CobU (участвует в синтезе витамина B12), справа — гипотетический белок. Ребро обходящее оперон (дуга ниже оперона) говорит о том, что в ряде штаммов в данном контексте нет иных вариантов генов. Наблюдаются некоторая вариабельность внутри оперона, соответствующая некоторым вариантам данного оперона [193]. Помимо pdu оперона, в том же контексте, у ряда штаммов наблюдается альтернативный набор генов (узлы и ребра, расположенные выше оперона), в который 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 альтернативного набора генов. Вероятно, данный участок генома часто служит местом рекомбинационных событий, приводящих к изменению набора генов, причем эти изменения не имеют строгого начала и конца (не сайт специфичны), но часто накладываются друг на друга.

Если рассмотреть более длинный фрагмент генома вблизи pdu оперона (рисунок 3.25), видно расположенный рядом еще более вариабельный участок.

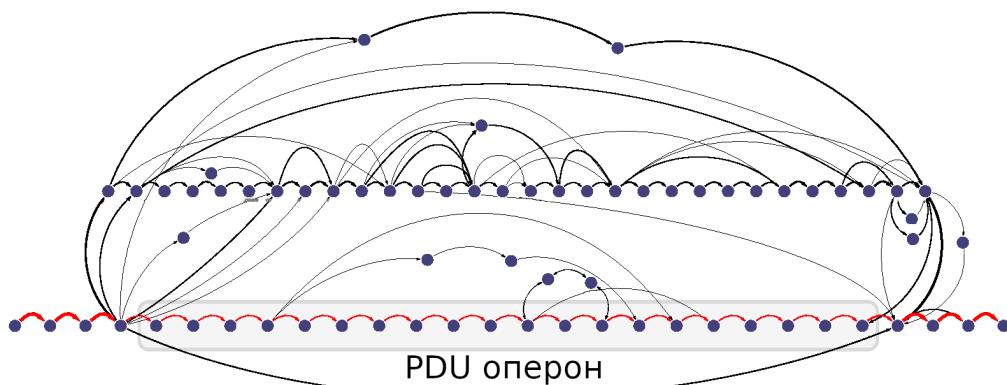


Рисунок 3.24 — Граф представляющий окрестность оперона утилизации пропандиола (propanediol utilization operon, pdu).

Примечательно, что в нем не содержится аннотированных генов связанных с мобильностью ДНК, вместо этого присутствуют гены синтеза капсулы и многочисленные гипотетические гены с неизвестной функцией.

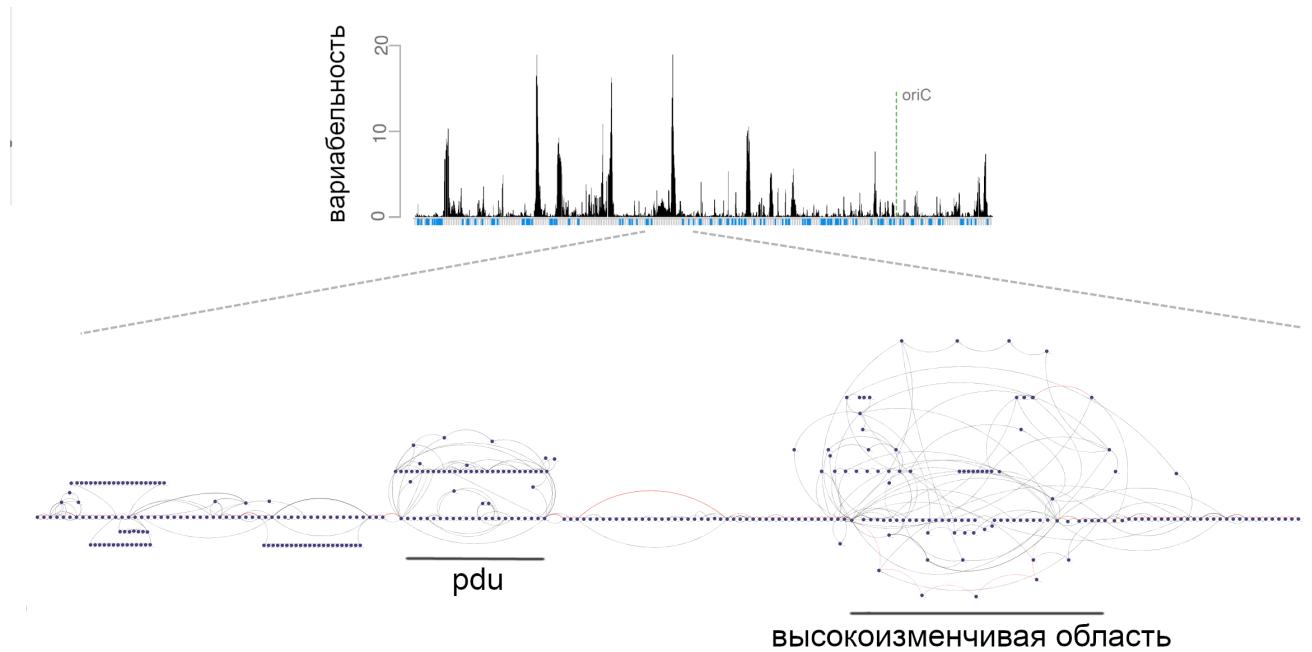


Рисунок 3.25 — Граф представляющий окрестность оперона утилизации пропандиола (propanediol utilization operon, pdu).

### Кластер генов синтеза бактериальной капсулы

На рисунке 3.26 показано графовое представление окрестности оперона синтеза бактериальной капсулы. Видно, что оперон состоит из консервативных фрагментов, flankирующих вариабельный участок. Вариабельная часть оперона соответствует генам, отвечающим за синтез серотип-специфичного набора полимеров капсулы и кодирующими: глицерин-3-фосфат цитидилтрансферазу,

### синтез и транспорт компонент капсулы

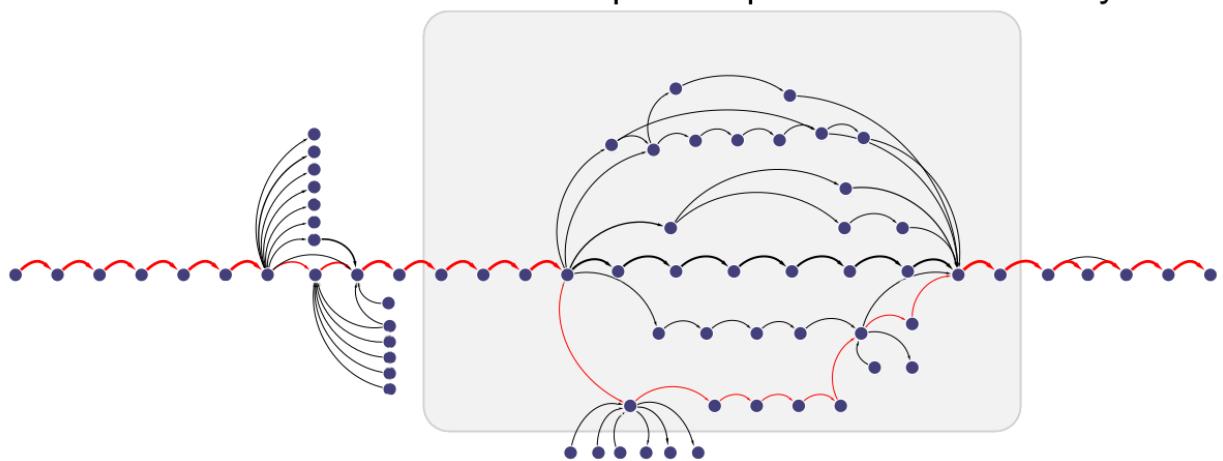


Рисунок 3.26 — Граф представляющий окрестность кластера генов синтеза бактериальной капсулы.

синтазу N,N'-диацетиллегионаминовой кислоты, N-ацилнейраминат цитидилилтрансферазу, белок биосинтеза полисиаловой кислоты P7, O-ацетилтрансферазу полисиаловой кислоты, ацетилтрансферазу EpsM, гликозилтрансферазу EpsJ, глицерол-фосфат трансферазу тейхоевой кислоты, поли (глицеролфосфат) полимеразу тейхоевой кислоты, поли (рибитол-фосфат) полимеразу тейхоевой кислоты, рибитол-фосфат-полимеразу тейхоевой кислоты TarL, UDP-галактопиранозную мутазу, UDP-Glc альфа-D-GlcNAc-дифосфоундекапренол бета-1,3-глюкозилтрансферазу WfgD, UDP-глюкозо-4-эпимеразу, вирджинамицин A ацетилтрансферазу. Гены консервативной части кодируют белки, участвующие в транспорте синтезированных веществ через клеточную стенку: транспортный белок полисиаловой кислоты KpsD, 3-дезокси-манно-октулосонат цитидилилтрансферазу, транспортный белок полисиаловой кислоты KpsM, транспортный АТФ-связывающий белок полисиаловой кислоты KpsT [222]. В референсном штамме *LF82* в вариабельную часть входят гены синтеза тейхоевых кислот.

Фрагменты единичной длины, расположенные слева от оперона, говорят о том, что есть много путей, которые не попадают в построенный подграф поскольку либо слишком длинные, либо заканчиваются вне отображаемой области и поэтому были сокращены до коротких фрагментов (параметр *tails* при анализе был равен 1). Если расширить рассматриваемую область, то многие из этих путей станут видны (рисунок 3.27, данный граф получен с параметрами: *tails* =

1, *minimal\_edge* = 2). Вид данного графа говорит о значительной вариабельности участка генома, содержащего гены синтеза капсулы.

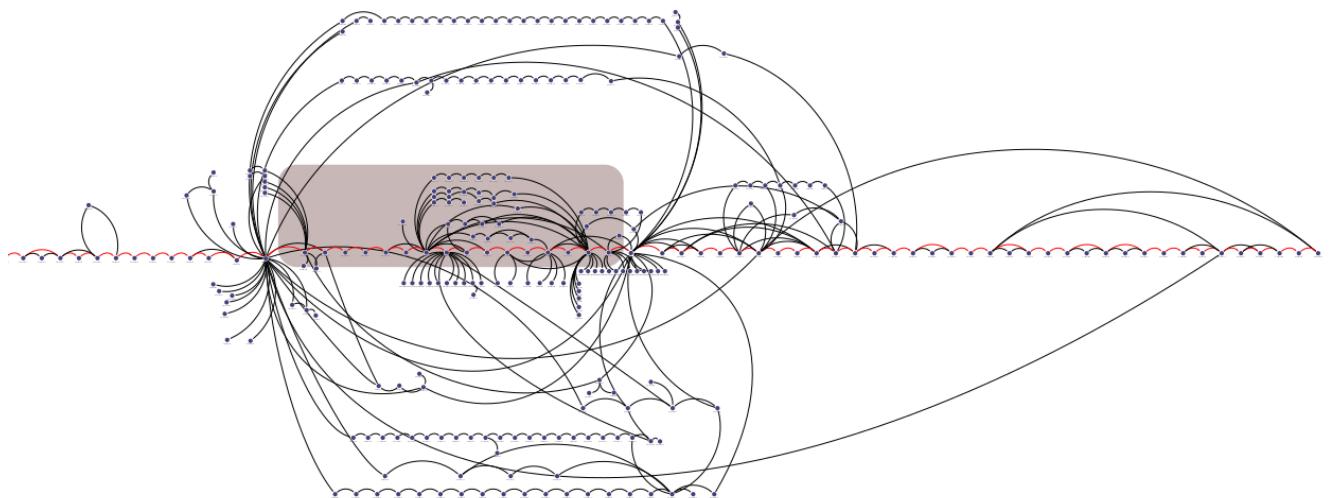


Рисунок 3.27 — Граф представляющий расширенную окрестность генов синтеза бактериальной капсулы (отмечены областью красного цвета).

Вариабельность состава капсулы адаптивна (это помогает бактериям избегать распознавание иммунной системой организма-хозяина либо бактериофагами [223; 224]. Можно предположить, что для бактерии выгодно такая локализация этих генов, где изменения могут происходить чаще. Можно выдвинуть гипотезу, что данный участок генома обладает некоторыми (еще не установленными) свойствами, которые обеспечивают значительную вариабельность генного состава, и в том числе - вариабельность состава генов капсулы.

### Оперон метаболизма глиоксилата

На рисунке 3.28 показан график отражающий область генома в окрестности оперона метаболизма глиоксилата. Данный оперон является частью геномного острова, включающего в себя четыре оперона: три оперона метаболизма углеводов и глицерина (*ptn*, *cgl* и *gcs*) и один оперон инвазии *ibe*. Данный остров описан у вызывающего менингит штамма *Escherichia coli K1*, и, методом экспериментального мутагенеза, показан как функционально значимый для проявления ее патогенных свойств [225].

Интересно, что хотя мы наблюдали значительную вариабельность геномных островов у *E. coli* и других бактериальных видов, в данном случае наблюдается весьма низкая вариабельность генного состава: остров либо представлен

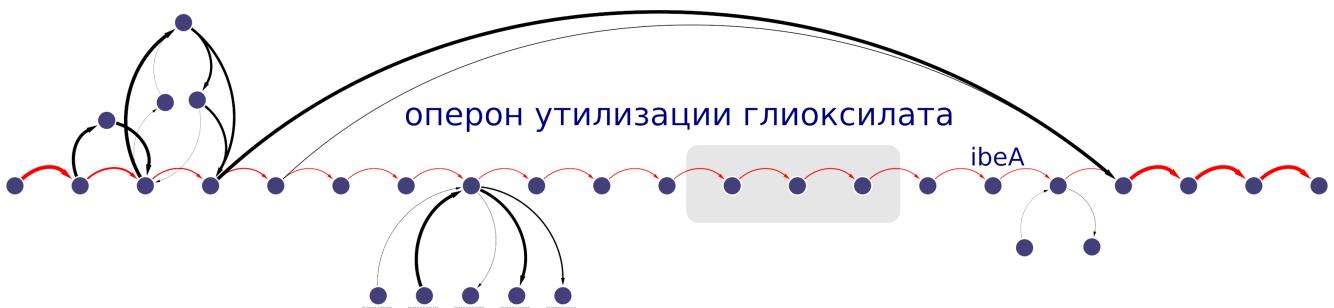


Рисунок 3.28 – Граф представляющий окрестность генов оперона метаболизма глиоксилата.

у некоторых организмов, либо - нет, и в последнем случае, на его месте не обнаруживаются какие-либо вставки.

## **Оперон захвата и утилизации сорбозы**

Данный оперон содержит следующие гены, кодирующие: транскрипционный регулятор оперона, сорбитол дегидрогеназу, фосфоенолпириват-зависимую сахарную фосфотрансферазу ЕII и дигидроантиказин-7-дегидрогеназу.

В случае данного оперона наблюдается картина, схожая с опероном захвата гемина и глиоксилата: оперон либо присутствует у части штаммов, либо нет, без выраженной изменчивости геномов в данном регионе (рисунок 3.29).

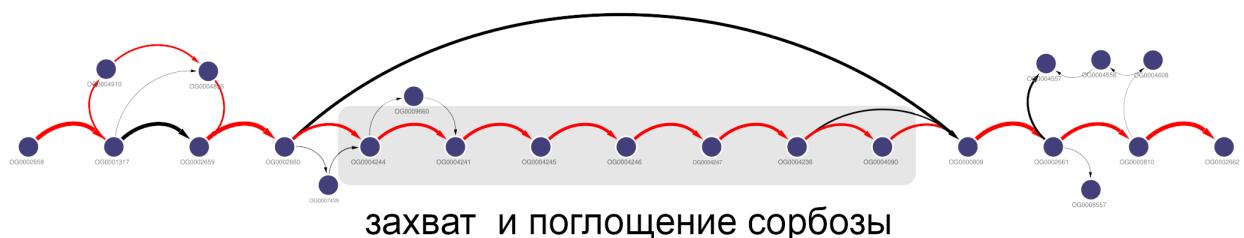


Рисунок 3.29 – Граф представляющий окрестность генов оперона захвата и утилизации сорбозы.

### 3.7 Разработка компьютерного приложения для анализа вариабельности геномов

Выполнено совместно с Конановым Дмитрием Николаевичем.

Оценка профиля изменчивости вдоль хромосомы и построение подграфов для фрагментов геномов - варианты анализа геномной изменчивости. Они позволяют проводить анализ на уровне отдельных репликонов (например, хромосомы) 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 приложение доступно на веб-сервере по адресу [gcb.rcpcm.org](http://gcb.rcpcm.org), и может быть звущено на локальном компьютере пользователя. Веб-версия содержит пред просчитанные данные для 143 видов. Использование локальной версии необходимо для анализа групп геномов не представленных на веб-сервере. Интерфейс программы показан на рисунке 3.30 и состоит из трех основных частей: области визуализации профиля изменчивости, области визуализации подграфов для выбранных областей генома и боковой панели с настройками и элементами управления (поиск генов, экспорт/импорт файлов).

Основные сценарии использования программы следующие:

I. Интерес представляет некоторый оперон или группа генов.

Пользователь выбирает организм, геном и задает координаты области интереса. Происходит построение подграфа для выбранной области. При необходимости, пользователь меняет параметры визуализации (например, увеличивает минимальный отображаемый вес ребра для исключения редко встречающихся комбинаций генов, в случае если получаемый граф слишком сложен для анализа). Экспорт полученной визуализации в графическом формате, либо в формате XML для последующей визуализации подграфа в программе Cytoscape (графовый редактор).

II. Интерес представляют области повышенной либо пониженной изменчивости генома.

Пользователь выбирает организм и геном. Происходит визуализация профиля изменчивости выбранного генома. Пользователь может выбрать интересующий его регион генома (например, область с максимальным уровнем изменчивости) и выполнить визуализацию подграфа в данной области - таким

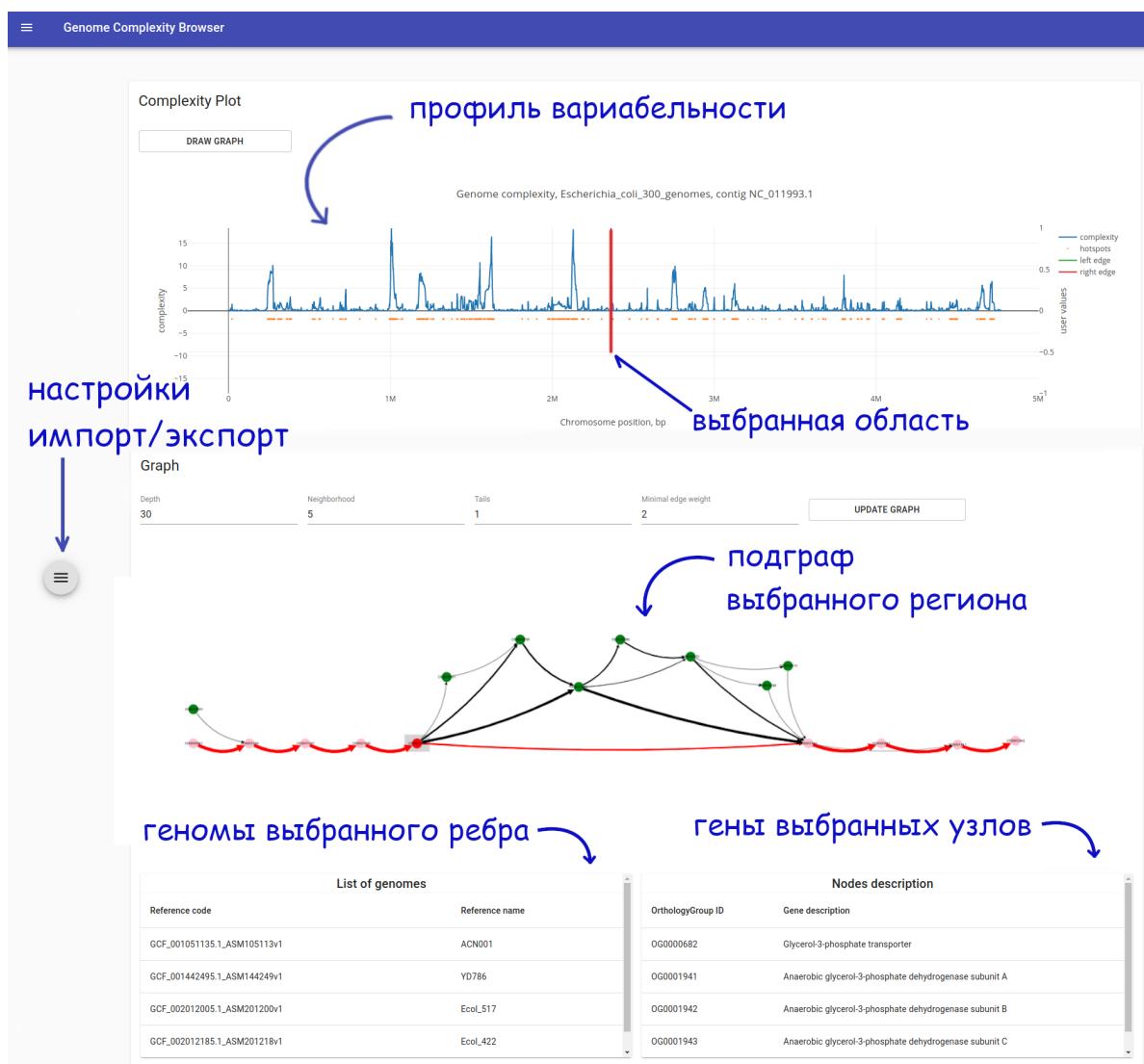


Рисунок 3.30 — Графический интерфейс программы Genome Complexity Browser.

образом можно установить, какие гены содержатся в данном локусе у различных геномов, и каков паттерн зафиксированных в ходе эволюции изменений. Пользователь может экспорттировать профиль изменчивости в виде текстового файла, для последующей визуализации (например, сравнение профилей изменчивости разных организмов), либо сохранить области с повышенной изменчивостью в файл в формате BED.

### Возможности программы

В программе предусмотрено автоматическое выделение "горячих точек" изменчивости. Они отображаются на профиле изменчивости (точки оранжевого цвета, расположенные ниже профиля), и их положение можно скачать в виде файла в формате BED. Для выделения областей повышенной изменчивости, мы использовали критерий Тьюки: искали значения уровня вариабельности, которые превышают 75 перцентиль на полтора межквартильных расстояния [226].

Для анализа геномов, недоступных на веб-сервере [gcb.rcpcm.org](https://gcb.rcpcm.org), необходимо использование локальной версии программы. Вначале подготовить (расположить в одной папке) интересующие пользователя геномы, затем запустить конвейер orthoSnake (доступен по адресу: <https://github.com/paraslonic/orthosnake>) и консольное приложение gg.py (доступно по адресу <https://github.com/DNKonanov/geneGraph>) - в итоге, будет создан файл с базой данных, содержащей графовое представление геномов и значения изменчивости. Файл с базой данных необходимо скопировать в папку, куда была установлена локальная версия сервера GCB, после этого можно работать с интерфейсом программы в веб-браузере. Схема анализа показана на рисунке 3.31, полный пример (с установкой всех зависимостей и проведением анализа) доступен по адресу: <https://gcb.readthedocs.io/en/latest/standalone.html#complete-step-by-step-example>.

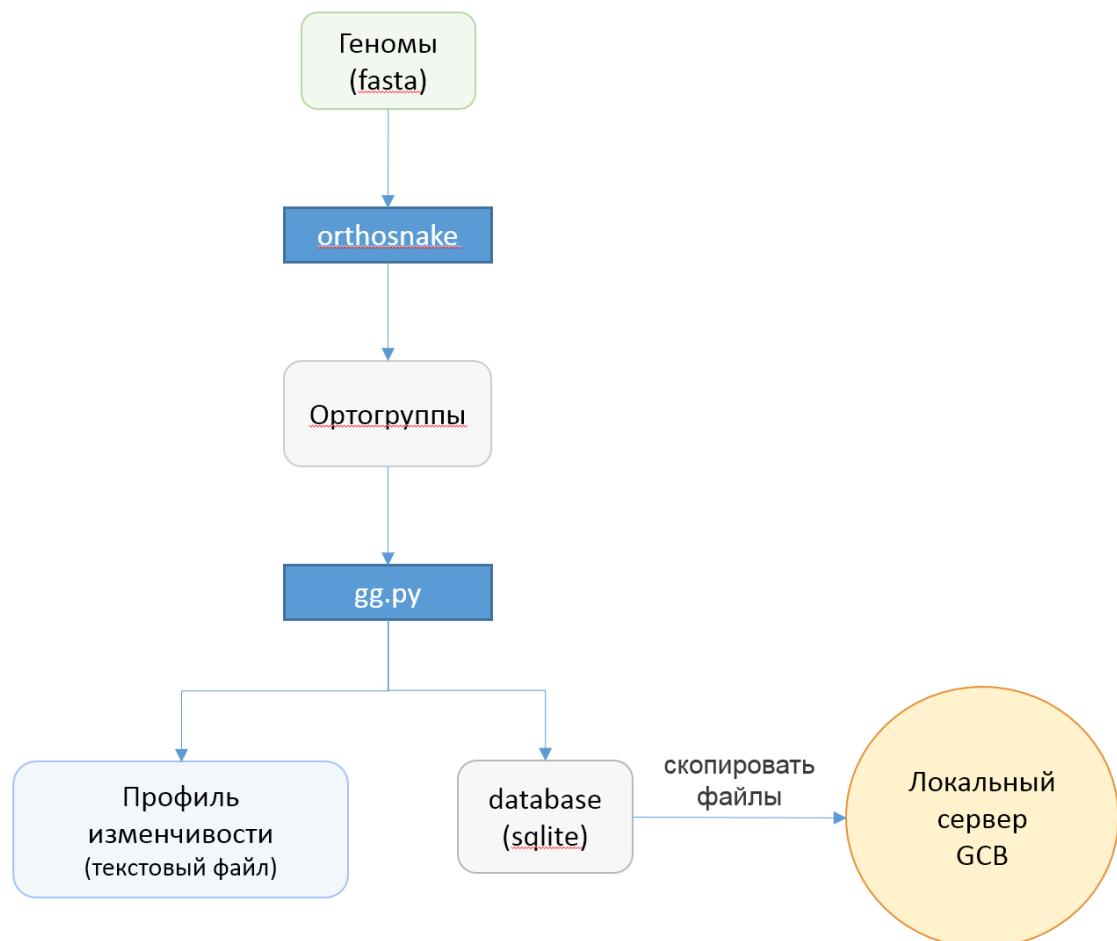


Рисунок 3.31 — Графический интерфейс программы Genome Complexity Browser.

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

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

#### **3.8.1 Проблема поиска генетических ассоциаций для субпопуляций бактерий**

Предположим, что у нас есть некоторый признак, по которому мы можем разбить набор геномов на группы, и наша задача - установить, какие гены значимо чаще (либо реже) встречаются в одной из групп. Распространенным подходом в подобном случае является оценка значимости по каждому отдельному гену. После проведения этих тестов необходимо применить поправку на множественное сравнение, так как иначе следует ожидать множество должно положительных результатов. В случае, если работа ведется на данных о полной последовательности генома, имеется большое количество анализируемых генов (порядка  $10^3 - 10^4$ ), а размеры групп, как правило, незначительны (порядка  $10^1 - 10^2$ ), 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 [193]. Рассматривая оперон как структурную единицу, мы значительно сокращаем количество анализируемых признаков, так, что оно становится сравнимым с размерами групп.

#### **3.8.2 Алгоритм поиска генетических ассоциаций**

Первым шагом выполняется оценка значимости в точном teste Фишера по каждому гену независимо. Затем ищутся опероны, в которых количество генов с заданным уровнем значимости выше, чем ожидается при случайному распределении генов по оперонам. Для оценки количества, ожидаемого при случайному распределении, применяется метод случайных перестановок либо оценка, основанная на распределении Пуассона. В первом случае, перестановки производятся в таблице соответствий генов и оперонов и подсчитывается максимальная доля

значимых генов в оперонах данной длины. Во втором случае, на основе наблюдаемого количества генов с уровнем значимости превышающим пороговый, рассчитывается их средняя плотность на единицу длины и затем для каждого оперона оценивается вероятность наблюдать данное либо большее количество подобных генов в соответствие с распределением Пуассона. И в первом, и во втором случае, финальным шагом является проведение поправки на множественное сравнение, но уже не для генов, а для оперонов, количество которых, как правило, значительно ниже общего количества генов.

Алгоритм первого подхода можно записать следующим образом:

Пусть  $I$  - набор оперонов, определенный на референсном геноме  $R$ .

1. Определить группы ортологии  $O$ .
2. Для каждой группы ортологии  $o \in O$  имеющей хотя один ген, представленный в  $R$ , посчитать  $P_o = p\text{-value}$  в точном тесте Фишера.
3. Для каждого оперона  $i \in I$  посчитать  $F_{obs_i} =$ доля входящих в него генов, для которых  $P_o < 0.05$ .
4. Для каждого  $j \in [1..10000]$ 
  - а) Провести случайную пермутацию значений  $P$ .
  - б) Для каждого оперона  $i \in I$  посчитать  $Frnd_{i,j} =$ доля генов, для которых  $P_o < 0.05$ .
5. Для каждого оперона  $i \in I$  найти  $F_{maxrnd_i} = \max_j(Frnd_{i,j})$  - максимальную долю генов в опероне при случайных перестановках значений значимостей для генов.
6. Для каждого оперона  $i \in I$ , если  $F_{obs_i} > F_{maxrnd_i}$ , считать его значимым с уровнем значимости  $p$ .

Перейдем к описанию результатов применения данного подхода.

### **Поиск оперонов, значимо чаще встречающихся у изолятов бактерий *E. coli*, изолированных от людей с болезнью Крона.**

В анализе был использован 51 геном *E. coli*, из которых 27 были геномы бактерий, изолированных от больных с болезнью Крона, а 24 генома принадлежали изолятам, полученным от здоровых людей. При помощи программы OrthoFinder [153] мы получили 11885 ортогрупп. Далее, при помощи

точного теста Фишера, для каждой ортогруппы мы оценили статистическую значимость ее неравномерной представленности в группах. Минимальное значение p-value, при этом, было меньше, чем 0.00037, а медианное - 0.78. После поправки на множественные сравнения (метод Бенджамини-Хохберга [227]) минимальное значение pvalue составляет 1 (такой же результат дают методы Бенджамини-Йекутили [228] и Холма [229])). Код для проведения данного анализа доступен в репозитории [https://github.com/paraslonic/Rakitina\\_etal\\_Crohn\\_paper/blob/master/ogEnrichment/calcFDR.r](https://github.com/paraslonic/Rakitina_etal_Crohn_paper/blob/master/ogEnrichment/calcFDR.r).

Следующим шагом мы осуществили описанный выше тип анализа, для поиска значимо дифференциально представленных оперонов. Информация об оперонах была взята из базы данных DOORS [219]. В качестве референсного генома мы использовали геном *Escherichia coli* LF82 - данный штамм был изолирован из пациента с болезнью Крона и является модельным в исследованиях адгезивно-инвазивного фенотипа у кишечной палочки [190]. Затем, мы провели 10000 случайных перестановок соответствий между генами и оперонами. Для каждой перестановки мы вычисляли зависимость количества генов в оперонах с p-value < 0.05 от длины оперона. Визуализация сравнения наблюдаемых и полученных при случайных перестановках результатов показана на рисунке 3.32. Опероны, для которых наблюдаемое число генов было выше, чем максимальное количество генов при случайных перестановках, считались статистически значимо пере- либо недо-представленными, поскольку для них можно считать, что в пермутационном teste p-value < 0.0001.

Так, например, оперон утилизации пропандиола состоит из 19 генов, из которых 14 генов (74%) имеют p-value < 0.05 в точном teste Фишера. При проведении 10000 случайных пермутаций уровней значимости по генам, в данном опероне в среднем наблюдалось 3% генов, а максимальная доля составила 21%. Таким образом, можно сделать вывод, что повышенная представленность данного оперона в изолятах из пациентов, но не здоровых людей, не является случайнм наблюдением. Полный список оперонов, определенных как значимо чаще встречающихся у изолятов из пациентов с болезнью Крона приведен в таблице 2.

Исходный код программ, определяющих опероны, статистически значимо чаще встречающихся у изолятов, полученных от пациентов с болезнью Крона, доступен в репозитории [https://github.com/paraslonic/Rakitina\\_etal\\_Crohn\\_paper/blob/master/operonPval/](https://github.com/paraslonic/Rakitina_etal_Crohn_paper/blob/master/operonPval/).

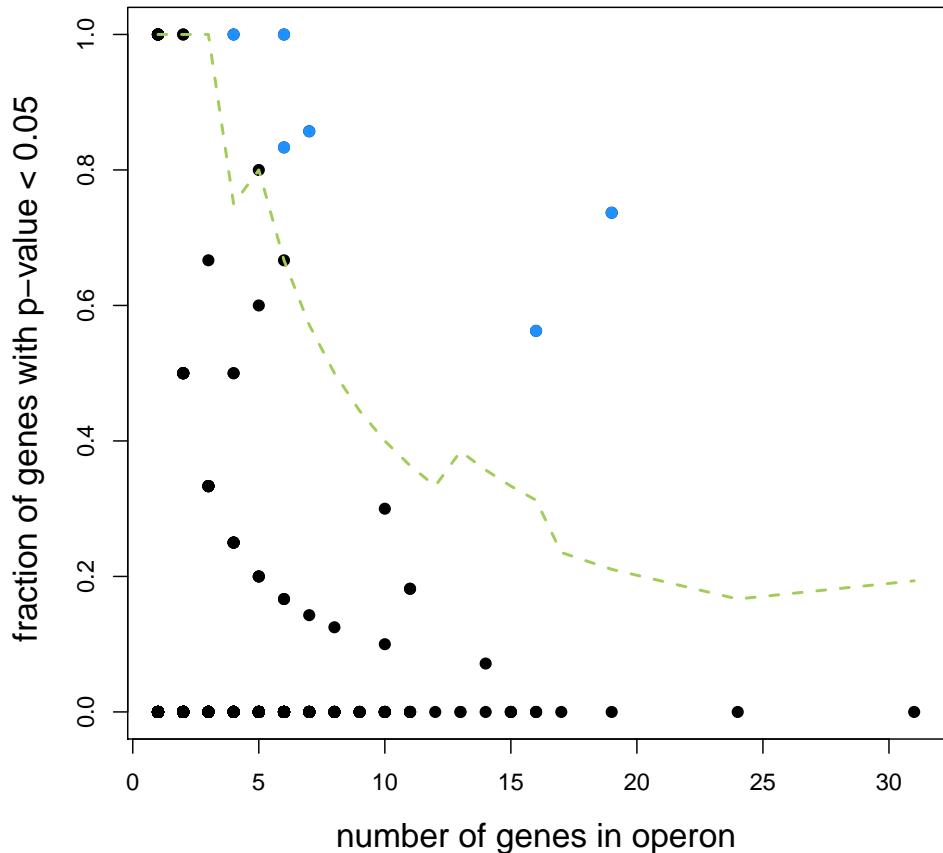


Рисунок 3.32 — Зависимость доли генов в оперонах, с уровнем значимости  $p\text{-value} < 0.05$ , от количества генов входящих в оперон. Пунктирная линия показывает максимальные значения, полученные при проведении 10000 случайных соотнесений генов и оперонов.

Таблица 2 — Список оперонов статистически значимо пере-представленных в группе штаммов *E. coli* изолированных от пациентов с болезнью Крона. N - количество генов, Pobs - наблюдаемая доля перепредставленных генов в опероне, Pmean - средняя доля перепредставленных генов при случайных пермутациях, Pmax - максимальная доля перепредставленных генов при случайных пермутациях.

<b>N</b>	<b>Pobs</b>	<b>Pmean</b>	<b>Pmax</b>	<b>функция</b>
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

## Глава 4. Обсуждение.

В настоящее время, для некоторых видов прокариот доступны сотни и даже тысячи геномов. Постоянно пополняющуюся коллекцию геномных последовательностей можно использовать для получения информации о вариабельности генома, его архитектуре, устройства различных оперонов, геномных островов и иных мобильных элементов.

Инструменты визуализации синтезии (например, Mauve, BRIG, genePlotR) часто используются для сравнительных исследований геномов прокариот и вирусов. Они позволяют визуально определить наличие больших и малых перестроек генома, но применимы для сравнительно небольших количеств сравниваемых геномов (10-20), а при большем размере выборки визуальный анализ становится затруднителен. В настоящее время существует недостаток методов визуализации сравнения большого (превышающих десятки) количества геномов и численной оценки изменчивости геномов. Подобного рода инструменты необходимы для исследования структуры генома, отдельных геномных элементов и изучения факторов, влияющих на возникновении либо исчезновение высокоЗизменчивых областей генома и областей с пониженной изменчивостью.

Для эффективного анализа и визуализаций больших наборов геномов мы предложили подход на основе графов, в котором гены представлены в виде узлов, узлы связываются между собой ребром, если соответствующие гены расположены последовательно в одном из геномов. Такое представление позволяет создавать компактные визуализации изменчивости определенного локуса в десятках и сотнях геномов. Мы также предложили использовать подсчет путей на графике для оценки изменчивости геномных последовательностей.

Графы имеют давнюю историю применения в анализе геномов. Они давно применяются для сборки геномов из множества коротких прочтений. При этом узлами графа служат либо сами прочтения, либо отдельные подстроки фиксированной длины ( $k$ -меры). Поиск оптимального обхода на графике дает оптимальное выравнивание прочтений (либо  $k$ -меров), что позволяет воссоздать последовательность генома. Применяются также методы картирования прочтений на набор референсных геномов, представленных в виде графа. Такой подход позволяет учсть уже известные варианты геномов для более полного картирования прочтений.

Графы применялись также для анализа изменчивости генома. В нескольких работах группы Е. Кунина использовалось графовое представление расположения кластеров ортологичных групп (COG) во множестве геномов. Значительное распространение получило представление перестроек генома в виде графов разрыва (breakpoints graph), удобное для реконструкции предковых состояний генома, но не для задач визуализации.

В нашей работе, мы использовали представление набора геномов в виде графа для двух задач: численной оценки уровня изменчивости в отдельных локусах генома и визуализации подграфов, соответствующих отдельным областям генома. Визуализация подграфов позволяет дать ответы на ряд вопросов о контексте 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 какие вариабельными? Какие геномы содержат определенную комбинацию генов?

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. Функция большинства найденных оперонов ясна, они позволяют микробу захватывать железо, утилизировать пропандиол (продукт переработки слизистого слоя), менять антигенные 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 капсул находятся в высоко-изменчивых –”горячих” – областях генома. Опероны утилизации гиммина, утилизации глиоксилата, захвата сорбозы напротив находятся в ”тихих” областях. Роль генетических факторов, находящихся в областях генома с разным уровнем изменчивости, в формировании генотипа и фенотипа бактерий — предмет дальнейших исследований. В случае оперона синтеза и экспорта капсулных полисахаридов, можно предположить, что нахождение данного оперона в ”горячей” области генома может способствовать более высокой изменчивости состава оперона (у него есть высоко вариативная часть, отвечающая за синтез капсулы), что в свою очередь выгодно для эффективного избегания иммунного ответа организма-хозяина.

При помощи графового представления геномов мы реализовали метод количественной оценки локальной изменчивости, основанный на поиске уникальных

путей в подграфе. Под изменчивостью в данном случае мы понимаем изменение состава либо взаимного расположения генов в геноме. Под локальностью — то, что изменения затрагивают небольшую область генома, не превышающую размер выбранного окна анализа (выбирается пользователем, обычно, составляет около 20-40 генов). Насколько нам известно, разработанный нами вычислительный конвейер (Genome Complexity Browser, GCB) является первым доступным инструментом, который позволяет количественно определять изменчивость генома на основе заданного пользователем набора геномов. GCB предоставляет способ оценки профиля изменчивости вдоль репликонов, что позволяет находить "горячие точки" генома, в которых уровень изменчивости значительно выше, чем в остальной части генома, и его "тихие" области.

Нам представляется перспективным направление исследований, включающее поиск изменений в интенсивности и местоположении "горячих" и "тихих" областей. Такой анализ можно проводить при помощи сравнения профилей изменчивости на разных уровнях филогенетического сходства: между геномами из различных внутривидовых структур (например, клад филогенетического дерева или экотипов) или между геномами близких видов. Результатом подобных исследований должно стать знание факторов, влияющих на уровень изменчивости генома в отдельных его областях. GCB позволяет оценивать изменчивость в определенном локусе генома (на основании задаваемых пользователем близкородственных геномов). Размер локуса может задаваться параметром окна, указываемом при запуске программы. Анализ геномов при помощи скользящего окна позволяет получить профиль уровня изменчивости вдоль генома. Мы предусмотрели метод автоматического определения областей повышенной изменчивости на основе критерия Тьюки (критерий учитывает медианное значение и межквартильный размах).

Для проверки применимости предложенного метода мы провели вычислительные эксперименты по моделированию эволюции бактериальных геномов. Мы допустили возможность вставки, удаления, перемещения генов, а также геномные инверсии. Вероятность этих событий была не равномерна вдоль модельной хромосомы, но менялась в соответствии с задаваемым профилем. Таким образом, в основе моделирования лежало предположение, что в геноме существуют области повышенной и пониженной изменчивости, положение которых устойчиво во времени. Затем мы оценивали профиль изменчивости на основе полученных геномов и предложенного нами подхода. Мы наблюдали высокий

уровень сходства исходных и вычисленных профилей. Вероятность геномных инверсий в наших экспериментах была на два порядка ниже, чем остальные события, что согласуется с наблюдаемой частотой у многих бактериальных видов. Этот параметр крайне важен для применимости метода: для эффективной оценки локальной изменчивости, глобальные перестройки генома не должны нарушать понятие локальности.

В применимости нашего подхода нас также убеждает анализ областей генома, содержащих известные "горячие точки" изменчивости: интегроны и геномные острова. Данные области обладают высоким уровнем изменчивости при анализе предложенным методом, резко выделяясь на фоне прилегающей части генома.

Профиль изменчивости генома *E. coli* обладает ярко выраженной контрастностью. На нем видны протяженные участки с низкой изменчивостью и области, в которых изменчивость очень высока. Большинство областей с высокой изменчивостью совпадает с локализацией профагов, либо описанных ранее островов патогенности. Существуют также "горячие" точки, включая и наиболее изменчивые области, внутри которых нам не удалось обнаружить следов мобильных элементов (такое же наблюдение описано в исследовании [138]). Протяженность низко-изменчивых областей оказалась выше (наиболее протяженная область составляет порядка миллиона пар нуклеотидов), высоко-изменчивые области как правило короче (наиболее длинные области достигают длины в 200-300 т.п.н.).

Мы сравнили профили изменчивости геномов *E. coli* из пяти основных филогрупп: A, B1, B2, D и E. Для каждой филогруппы был выбран один референсный геном и сто ближайших к нему геномов из базы RefSeq, для каждого набора из 101 геномы мы провели анализ изменчивости независимо. Референсные геномы обладали достаточно высоким уровнем сохранности структуры генома (без крупных перестроек), что позволило сравнивать расположение областей повышенной и пониженной изменчивости. Сравнение показало, что множество областей повышенной изменчивости расположено в близких по контексту областях генома. Значительная часть этих областей соответствует профаговым областям, так что их консервативное расположение можно объяснить сайт-специфичным характером их встраивания и изменчивостью самих фаговых геномов. Особенно заметна роль профагов в вариабельности геномов из филогруппы E (референсный геном: *O157:H7 Sakai*). Интересно, что области интеграции разных фагов значительно различаются по уровню своей изменчивости. Ряд областей повышенной изменчивости геномов, не содержащих профагов, также показал консервативность

расположения, в том числе - наиболее изменчивые фрагменты геномов оказались расположены в схожих областях генома.

Мы провели сравнение профилей изменчивости для представителей различных филогрупп у трех других видов бактерий. Для *Pseudomonas aeruginosa* и *Neisseria gonorrhoeae* мы наблюдали, что области повышенной изменчивости расположены в сходном геномном контексте. Иная картина наблюдалась в случае *Pseudomonas fluorescens*, для которой сходство профилей было незначительным, а сами профили отличались равномерностью (отсутствием ярко выраженных областей с повышенной и пониженной изменчивостью). На наш взгляд, это можно объяснить высокой частотой геномных перестроек, в частности - инверсий, зафиксированных в геномах данного вида. Как уже отмечалось выше, высокая частота крупных геномных перестроек делает наш подход неприменимым к анализу локальной изменчивости геномов.

Мы также провели сравнение профилей изменчивости для геномов из филогенетически близких видов и наблюдали ряд "горячих точек" расположенных в близких местах генома и ряд областей повышенной изменчивости, специфичных для отдельных видов.

Предложенный в нашей работе подход к оценке геномной изменчивости не универсален. Например, он не подходит для обнаружения крупных геномных перестроек (больше, чем параметр окна, который обычно составляет несколько десятков генов) или изменений в некодирующих частях генома. Точность и применимость нашего подхода зависит от точности поиска ортологичных генов — сложной вычислительной задачи, часто не имеющей однозначного решения. В нашем исследовании мы использовали инструмент OrthoFinder, который использует алгоритм кластеризации графов MCL для значений попарного сходства белков, нормированных на длину выравнивания. Мы считаем этот инструмент оптимальным на настоящий момент с точки зрения эффективности и точности. Недостатком данного подхода является частое попадание паралогичных генов в одну ортогруппу, что затрудняет дальнейший анализ. Мы наблюдали, что в среднем 0,5% всех ортогрупп на геном содержат по крайней мере один паралогичный ген; среди всех ортогрупп, предполагаемых для вида, доля ортогрупп с паралогами составляет почти 16%.

Мы реализовали два подхода к анализу ортогрупп, содержащих паралогичные гены. Подход, реализованный как метод по умолчанию в GCB, заключается в игнорировании таких ортогрупп. Другой подход заключается в искусственной

ортологизации паралогов (каждый паралогичный ген с уникальным левым и правым контекстом генов добавляется в граф с уникальным суффиксом). Исходя из нашего опыта, оптимальной стратегией является использование режима по умолчанию для первоначального анализа с последующей проверкой всех выводов в режиме ортологизации.

В GCB предусмотрено использование двух алгоритмов автоматической компоновки (layout) узлов графов: *Dagre* и *Graphviz*. Все же, для обеспечения наиболее ясной компоновки часто требуется ручные выравнивание. Новые алгоритмы автоматической компоновки, возможно, позволят снизить необходимость "ручного" выравнивания.

Несмотря на вышеупомянутые недостатки, мы считаем, что предложенный метод анализа изменчивости геномов уже информативен и применим в достаточной степени, для проведения дальнейшего анализа факторов, влияющих на локализацию областей повышенной и пониженной изменчивости у прокариот.

Горячие точки изменчивости генома были описаны для нескольких видов бактерий. В работе [138] авторы проанализировали области генома, в которых часто наблюдается последствия горизонтального переноса генов. Они провели анализ для 80 видов бактерий и пришли к выводу, что во многих "горячих точках" отсутствуют мобильные генетические элементы, и предположили, что гомологичная рекомбинация в первую очередь ответственна за изменчивость этих локусов. Аналогичный вывод о ведущей роли в горизонтальном переносе генов процесса гомологичной рекомбинации без участия мобильных элементов был сделан в работе [230]. Факторы, определяющие расположение высокоизменчивых точек, влияющие на их появление и устраниние, остаются открытыми вопросами. Являются ли эти области повышенной изменчивости теми местами, где действительно чаще происходят изменения? Альтернативным является предположение, что изменения происходят значительно равномернее вдоль генома, но организмы, в которых изменения произошли в неподходящих местах не выживают и мы их не наблюдаем.

Одной из исходных гипотез данного исследования являлась связь между уровнем изменчивости и пространственной укладкой хромосомы. Нам удалось найти лишь слабую связь между профилем межхромосомных контактов и уровнем изменчивости. Попытка сопоставить уровень изменчивости с реконструированной пространственной укладкой хромосомы кишечной палочки ([231]) также не дала значимого результата (что также может быть связано с несовершен-

ством моделей пространственной укладки). Весьма вероятно, что на уровень изменчивости разных локусов генома влияет не один, но значительное количество различных факторов. Ряд этих факторов могут быть связаны с архитектурой генома. С меньшей вероятностью будут выживать организмы, в которых изменения в геномах привели к нарушениям в архитектуре - то есть к нарушению клеточных процессов с участием хромосомы и иных элементов генома. В таком случае, дальнейшие исследования изменчивости геномов могут пролить свет на еще не известные элементы геномной архитектуры и способствовать более эффективному конструированию генов в рамках программы синтетической биологии. С другой стороны, можно предположить, что распределение областей высокой изменчивости носит случайный характер, например в следствии случайного распределения различных сайтов интеграции мобильных элементов, часть которых может быть не известна. Выяснение вклада различных факторов в расположение более и менее изменчивых областей - дело будущих исследований.

## Глава 5. Выводы

1. Графовое представление геномов позволяет эффективно проводить поиск областей генома с повышенной изменчивостью.
2. Геномы представителей различных филогрупп и филогенетически близких видов прокариот имеют консервативно расположенные области повышенной изменчивости (расположенные в местах генома с одинаковым генным контекстом).
3. Уровень геномной изменчивости ассоциирован с плотностью хромосомных контактов (коэффициент корреляции составил -0.36) и плотностью расположения сайтов Chi (коэффициент корреляции составил -0.25).
4. Следующие опероны значимо чаще ( $p\text{-value} < 0.00001$ ) встречаются в изолятах *E. coli* от пациентов с болезнью Крона: захват сорбозы, захват гемина, утилизации глиоксилата, утилизации пропандиола, синтеза и экспорта капсулных полисахаридов.
5. Опероны значимо чаще встречающиеся в изолятах *E. coli* от пациентов с болезнью Крона расположены в высокоизменчивых областях (опероны утилизации пропандиола, синтеза и экспорта капсулных полисахаридов) и в консервативных участках генома (захват сорбозы, захват гемина, утилизации глиоксилата).

## Список литературы

1. *Noguera-Solano, R.* Genome: twisting stories with DNA / R. Noguera-Solano, R. Ruiz-Gutierrez, J. M. Rodriguez-Caso // Endeavour. — 2013. — Т. 37, № 4. — С. 213—219.
2. The *Vibrio cholerae* genome contains two unique circular chromosomes / M. Trucksis [и др.] // Proceedings of the National Academy of Sciences. — 1998. — Т. 95, № 24. — С. 14464—14469.
3. *Hinnebusch, B. J.* The bacterial nucleoid visualized by fluorescence microscopy of cells lysed within agarose: comparison of *Escherichia coli* and spirochetes of the genus *Borrelia*. / B. J. Hinnebusch, A. J. Bendich // Journal of bacteriology. — 1997. — Т. 179, № 7. — С. 2228—2237.
4. *Bobay, L.-M.* The evolution of bacterial genome architecture / L.-M. Bobay, H. Ochman // Frontiers in genetics. — 2017. — Т. 8. — С. 72.
5. *Herdman, M.* The evolution of bacterial genomes / M. Herdman // The Evolution of Genome Size. — 1985. — С. 37—68.
6. The mosaic genome of *Anaeromyxobacter dehalogenans* strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria / S. H. Thomas [и др.] // PloS one. — 2008. — Т. 3, № 5.
7. *Mira, A.* Deletional bias and the evolution of bacterial genomes / A. Mira, H. Ochman, N. A. Moran // Trends in Genetics. — 2001. — Т. 17, № 10. — С. 589—596.
8. *Rocha, E. P.* The organization of the bacterial genome / E. P. Rocha // Annual review of genetics. — 2008. — Т. 42. — С. 211—233.
9. Evidence for symmetric chromosomal inversions around the replication origin in bacteria / J. A. Eisen [и др.] // Genome biology. — 2000. — Т. 1, № 6. — research0011—1.
10. *Boccard, F.* Spatial arrangement and macrodomain organization of bacterial chromosomes / F. Boccard, E. Esnault, M. Valens // Molecular microbiology. — 2005. — Т. 57, № 1. — С. 9—16.
11. Computational identification of operons in microbial genomes / Y. Zheng [и др.] // Genome research. — 2002. — Т. 12, № 8. — С. 1221—1230.

12. *Wells, J. N.* Operon gene order is optimized for ordered protein complex assembly / J. N. Wells, L. T. Bergendahl, J. A. Marsh // Cell reports. — 2016. — Т. 14, № 4. — С. 679—685.
13. LOPERON-GROUPE DE GENES A EXPRESSION COORDONNEE PAR UN OPERATEUR / F. Jacob [и др.] // COMPTES RENDUS HEBDOMADAIRE DES SEANCES DE L ACADEMIE DES SCIENCES. — 1960. — Т. 250, № 9. — С. 1727—1729.
14. *Jacob, F.* Genetic regulatory mechanisms in the synthesis of proteins / F. Jacob, J. Monod // Journal of molecular biology. — 1961. — Т. 3, № 3. — С. 318—356.
15. *Lawrence, J. G.* Selfish operons: horizontal transfer may drive the evolution of gene clusters / J. G. Lawrence, J. R. Roth // Genetics. — 1996. — Т. 143, № 4. — С. 1843—1860.
16. Connected gene neighborhoods in prokaryotic genomes / I. B. Rogozin [и др.] // Nucleic acids research. — 2002. — Т. 30, № 10. — С. 2212—2223.
17. DNA motifs that sculpt the bacterial chromosome / F. Touzain [и др.] // Nature Reviews Microbiology. — 2011. — Т. 9, № 1. — С. 15—26.
18. FtsK, a literate chromosome segregation machine / S. Bigot [и др.] // Molecular microbiology. — 2007. — Т. 64, № 6. — С. 1434—1441.
19. The nature of mutations induced by replication-transcription collisions / T. S. Sankar [и др.] // Nature. — 2016. — Т. 535, № 7610. — С. 178—181.
20. *Brewer, B. J.* When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome / B. J. Brewer // Cell. — 1988. — Т. 53, № 5. — С. 679—686.
21. The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces / X. Mao [и др.] // Nucleic acids research. — 2012. — Т. 40, № 17. — С. 8210—8218.
22. *Galtier, N.* Relationships between genomic G+ C content, RNA secondary structures, and optimal growth temperature in prokaryotes / N. Galtier, J. Lobry // Journal of molecular evolution. — 1997. — Т. 44, № 6. — С. 632—636.

23. *Wang, H.-C.* On the correlation between genomic G+ C content and optimal growth temperature in prokaryotes: data quality and confounding factors / H.-C. Wang, E. Susko, A. J. Roger // Biochemical and biophysical research communications. — 2006. — Т. 342, № 3. — С. 681—684.
24. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes / H. Naya [и др.] // Journal of molecular evolution. — 2002. — Т. 55, № 3. — С. 260—264.
25. Aerobic prokaryotes do not have higher GC contents than anaerobic prokaryotes, but obligate aerobic prokaryotes have / S. Aslam [и др.] // BMC evolutionary biology. — 2019. — Т. 19, № 1. — С. 35.
26. *Lobry, J. R.* Life history traits and genome structure: aerobiosis and G+ C content in bacteria / J. R. Lobry // International Conference on Computational Science. — Springer. 2004. — С. 679—686.
27. Evolutionary determinants of genome-wide nucleotide composition / H. Long [и др.] // Nature ecology & evolution. — 2018. — Т. 2, № 2. — С. 237.
28. *Szybalski, W.* Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis / W. Szybalski, H. Kubinski, P. Sheldrick // Cold Spring Harbor symposia on quantitative biology. Т. 31. — Cold Spring Harbor Laboratory Press. 1966. — С. 123—127.
29. *Lobry, J. R.* Asymmetric substitution patterns in the two DNA strands of bacteria. / J. R. Lobry // Molecular biology and evolution. — 1996. — Т. 13, № 5. — С. 660—665.
30. *Lobry, J. R.* Genomic landscapes / J. R. Lobry // Microbiology Today. — 1999. — Т. 26. — С. 164—165.
31. *Arakawa, K.* The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection / K. Arakawa, M. Tomita // Evolutionary Bioinformatics. — 2007. — Т. 3. — С. 117693430700300006.
32. *Kono, N.* Accelerated laboratory evolution reveals the influence of replication on the GC skew in *Escherichia coli* / N. Kono, M. Tomita, K. Arakawa // Genome biology and evolution. — 2018. — Т. 10, № 11. — С. 3110—3117.
33. *Kuzminov, A.* Recombinational Repair of DNA Damage in *Escherichia coli* and Bacteriophage  $\lambda$  / A. Kuzminov // Microbiol. Mol. Biol. Rev. — 1999. — Т. 63, № 4. — С. 751—813.

34. Resolution of Holliday junctions by RuvABC prevents dimer formation in rep mutants and UV-irradiated cells / B. Michel [и др.] // Molecular microbiology. — 2000. — Т. 37, № 1. — С. 180—191.
35. *Ochman, H.* Lateral gene transfer and the nature of bacterial innovation / H. Ochman, J. G. Lawrence, E. A. Groisman // nature. — 2000. — Т. 405, № 6784. — С. 299—304.
36. *Mullany, P.* The dynamic bacterial genome. Т. 8 / P. Mullany. — Cambridge University Press, 2005.
37. *Smith, G. R.* How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view / G. R. Smith // Microbiol. Mol. Biol. Rev. — 2012. — Т. 76, № 2. — С. 217—228.
38. Crystal structure of RecBCD enzyme reveals a machine for processing DNA breaks / M. R. Singleton [и др.] // Nature. — 2004. — Т. 432, № 7014. — С. 187—193.
39. RecG directs DNA synthesis during double-strand break repair / B. Azeroglu [и др.] // PLoS genetics. — 2016. — Т. 12, № 2.
40. Rec-mediated recombinational hot spot activity in bacteriophage lambda II. A mutation which causes hot spot activity / S. T. Lam [и др.] // Genetics. — 1974. — Т. 77, № 3. — С. 425—433.
41. Hotspots for generalized recombination in the Escherichia coli chromosome / R. E. Malone [и др.] // Journal of molecular biology. — 1978. — Т. 121, № 4. — С. 473—491.
42. Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling / D. Halpern [и др.] // PLoS genetics. — 2007. — Т. 3, № 9.
43. Recombination and annealing pathways compete for substrates in making rrn duplications in *Salmonella enterica* / A. B. Reams [и др.] // Genetics. — 2014. — Т. 196, № 1. — С. 119—135.
44. The positioning of Chi sites allows the RecBCD pathway to suppress some genomic rearrangements / C. Li [и др.] // Nucleic acids research. — 2019. — Т. 47, № 4. — С. 1836—1846.

45. *Dame, R. T.* Chromosome organization in bacteria: mechanistic insights into genome structure and function / R. T. Dame, F.-Z. M. Rashid, D. C. Grainger // Nature Reviews Genetics. — 2020. — Т. 21, № 4. — С. 227—242.
46. *Dekker, J.* Gene regulation in the third dimension / J. Dekker // Science. — 2008. — Т. 319, № 5871. — С. 1793—1794.
47. Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression / D. Vernimmen [и др.] // The EMBO journal. — 2007. — Т. 26, № 8. — С. 2041—2051.
48. *Hofmann, A.* The role of loops on the order of eukaryotes and prokaryotes / A. Hofmann, D. W. Heermann // FEBS letters. — 2015. — Т. 589, № 20. — С. 2958—2965.
49. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation / M. A. Umbarger [и др.] // Molecular cell. — 2011. — Т. 44, № 2. — С. 252—264.
50. H-NS mediates the silencing of laterally acquired genes in bacteria / S. Lucchini [и др.] // PLoS Pathog. — 2006. — Т. 2, № 8. — e81.
51. *Dame, R. T.* Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation / R. T. Dame, M. C. Noom, G. J. Wuite // Nature. — 2006. — Т. 444, № 7117. — С. 387—390.
52. *Nolivos, S.* The bacterial chromosome: architecture and action of bacterial SMC and SMC-like complexes / S. Nolivos, D. Sherratt // FEMS microbiology reviews. — 2014. — Т. 38, № 3. — С. 380—392.
53. Chromosome partitioning in *Escherichia coli*: novel mutants producing anucleate cells. / S. Hiraga [и др.] // Journal of Bacteriology. — 1989. — Т. 171, № 3. — С. 1496—1505.
54. Multiscale structuring of the *E. coli* chromosome by nucleoid-associated and condensin proteins / V. S. Lioy [и др.] // Cell. — 2018. — Т. 172, № 4. — С. 771—783.
55. *Woldringh, C. L.* The role of co-transcriptional translation and protein translocation (transsertion) in bacterial chromosome segregation / C. L. Woldringh // Molecular microbiology. — 2002. — Т. 45, № 1. — С. 17—29.

56. *Espeli, O.* DNA dynamics vary according to macrodomain topography in the *E. coli* chromosome / O. Espeli, R. Mercier, F. Boccard // Molecular microbiology. — 2008. — Т. 68, № 6. — С. 1418—1427.
57. *Valens, M.* The MaoP/maoS site-specific system organizes the Ori region of the *E. coli* chromosome into a macrodomain / M. Valens, A. Thiel, F. Boccard // PLoS genetics. — 2016. — Т. 12, № 9. — e1006309.
58. *Rodriguez-Valera, F.* Flexible genomic islands as drivers of genome evolution / F. Rodriguez-Valera, A.-B. Martin-Cuadrado, M. López-Pérez // Current opinion in microbiology. — 2016. — Т. 31. — С. 154—160.
59. Migration and horizontal gene transfer divide microbial genomes into multiple niches / R. Niehus [и др.] // Nature communications. — 2015. — Т. 6. — С. 8924.
60. *Treangen, T. J.* Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes / T. J. Treangen, E. P. Rocha // PLoS Genet. — 2011. — Т. 7, № 1. — e1001284.
61. Horizontal gene transfer mediated bacterial antibiotic resistance / D. Sun [и др.] // Frontiers in microbiology. — 2019. — Т. 10. — С. 1933.
62. *González-Candelas, F.* Barriers to Horizontal Gene Transfer: Fuzzy and Evolvable Boundaries / F. González-Candelas, M. P. Francino // Horizontal Gene Transfer in Microorganisms. — 2012. — Т. 47.
63. *Caro-Quintero, A.* Inter-phylum HGT has shaped the metabolism of many mesophilic and anaerobic bacteria / A. Caro-Quintero, K. T. Konstantinidis // The ISME journal. — 2015. — Т. 9, № 4. — С. 958—967.
64. Mechanisms of gene flow in archaea / A. Wagner [и др.] // Nature Reviews Microbiology. — 2017. — Т. 15, № 8. — С. 492—501.
65. *Lacroix, B.* Transfer of DNA from bacteria to eukaryotes / B. Lacroix, V. Citovsky // MBio. — 2016. — Т. 7, № 4.
66. *Thomas, C. M.* Mechanisms of, and barriers to, horizontal gene transfer between bacteria / C. M. Thomas, K. M. Nielsen // Nature reviews microbiology. — 2005. — Т. 3, № 9. — С. 711—721.
67. Bacterial vesicles in marine ecosystems / S. J. Biller [и др.] // science. — 2014. — Т. 343, № 6167. — С. 183—186.

68. *Dubey, G. P.* Intercellular nanotubes mediate bacterial communication / G. P. Dubey, S. Ben-Yehuda // Cell. — 2011. — Т. 144, № 4. — С. 590—600.
69. Homologues of genetic transformation DNA import genes are required for *Rhodobacter capsulatus* gene transfer agent recipient capability regulated by the response regulator CtrA / C. A. Brimacombe [и др.] // Journal of bacteriology. — 2015. — Т. 197, № 16. — С. 2653—2663.
70. *Sun, D.* Pull in and push out: mechanisms of horizontal gene transfer in bacteria / D. Sun // Frontiers in microbiology. — 2018. — Т. 9. — С. 2154.
71. Bacterial transformation: distribution, shared mechanisms and divergent control / C. Johnston [и др.] // Nature Reviews Microbiology. — 2014. — Т. 12, № 3. — С. 181—196.
72. *Finkel, S. E.* DNA as a nutrient: novel role for bacterial competence gene homologs / S. E. Finkel, R. Kolter // Journal of bacteriology. — 2001. — Т. 183, № 21. — С. 6288—6293.
73. *Mell, J. C.* Natural competence and the evolution of DNA uptake specificity / J. C. Mell, R. J. Redfield // Journal of bacteriology. — 2014. — Т. 196, № 8. — С. 1471—1483.
74. *Dubnau, D.* Mechanisms of DNA uptake by naturally competent bacteria / D. Dubnau, M. Blokesch // Annual review of genetics. — 2019. — Т. 53. — С. 217—237.
75. *Piepenbrink, K. H.* DNA uptake by type IV filaments / K. H. Piepenbrink // Frontiers in molecular biosciences. — 2019. — Т. 6. — С. 1.
76. Conjugative and mobilizable genomic islands in bacteria: evolution and diversity / X. Bellanger [и др.] // FEMS Microbiology Reviews. — 2014. — Т. 38, № 4. — С. 720—760.
77. *Shintani, M.* Genomics of microbial plasmids: classification and identification based on replication and transfer systems and host taxonomy / M. Shintani, Z. K. Sanchez, K. Kimbara // Frontiers in microbiology. — 2015. — Т. 6. — С. 242.
78. The repertoire of ICE in prokaryotes underscores the unity, diversity, and ubiquity of conjugation / J. Guglielmini [и др.] // PLoS genet. — 2011. — Т. 7, № 8. — e1002222.

79. Replication and control of circular bacterial plasmids / G. Del Solar [и др.] // Microbiology and molecular biology reviews. — 1998. — Т. 62, № 2. — С. 434—464.
80. *Ravin, N. V.* N15: The linear phage–plasmid / N. V. *Ravin* // Plasmid. — 2011. — Т. 65, № 2. — С. 102—109.
81. Pathways for horizontal gene transfer in bacteria revealed by a global map of their plasmids / S. Redondo-Salvo [и др.] // Nature communications. — 2020. — Т. 11, № 1. — С. 1—13.
82. An evolutionary perspective on plasmid lifestyle modes / N. Hülter [и др.] // Current opinion in microbiology. — 2017. — Т. 38. — С. 74—80.
83. Conjugative-DNA transfer processes / E. Zechner [и др.] // The horizontal gene pool: bacterial plasmids and gene spread. — 2000. — Т. 23. — С. 419.
84. *Brunder, W.* Genome plasticity in Enterobacteriaceae / W. Brunder, H. Karch // International journal of medical microbiology. — 2000. — Т. 290, № 2. — С. 153—165.
85. *Myers, G.* The role of mobile DNA in the evolution of prokaryotic genomes / G. Myers, I. Paulsen, C. Fraser // The implicit genome, OUP, Oxford. — 2006. — С. 133—137.
86. *Ronchel, M. C.* Retrotransfer of DNA in the rhizosphere / M. C. Ronchel, M. A. Ramos-Díaz, J. L. Ramos // Environmental microbiology. — 2000. — Т. 2, № 3. — С. 319—323.
87. IncA/C conjugative plasmids mobilize a new family of multidrug resistance islands in clinical *Vibrio cholerae* non-O1/non-O139 isolates from Haiti / N. Carraro [и др.] // MBio. — 2016. — Т. 7, № 4.
88. Mobilizable genomic islands, different strategies for the dissemination of multidrug resistance and other adaptive traits / N. Carraro [и др.] // Mobile genetic elements. — 2017. — Т. 7, № 2. — С. 1—6.
89. *Lederberg, J.* Gene recombination in *Escherichia coli* / J. Lederberg, E. L. Tatum // Nature. — 1946. — Т. 158, № 4016. — С. 558—558.

90. The ICESt1 element of *Streptococcus thermophilus* belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration / V. Burrus [и др.] // Plasmid. — 2002. — Т. 48, № 2. — С. 77—97.
91. Integration and Excision of *aBacteroides* Conjugative Transposon, CTnDOT / Q. Cheng [и др.] // Journal of bacteriology. — 2000. — Т. 182, № 14. — С. 4035—4043.
92. *Botelho, J.* The role of integrative and conjugative elements in antibiotic resistance evolution / J. Botelho, H. Schulenburg // Trends in Microbiology. — 2020.
93. *Carraro, N.* Biology of three ICE families: SXT/R391, ICEBs1, and ICESt1/ICESt3 / N. Carraro, V. Burrus // Mobile DNA III. — 2015. — С. 289—309.
94. *Pernodet, J.-L.* Plasmids in different strains of *Streptomyces ambofaciens*: free and integrated form of plasmid pSAM2 / J.-L. Pernodet, J.-M. Simonet, M. Guérineau // Molecular and General Genetics MGG. — 1984. — Т. 198, № 1. — С. 35—41.
95. A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species / J. A. Kers [и др.] // Molecular microbiology. — 2005. — Т. 55, № 4. — С. 1025—1033.
96. Transcription analysis of *Streptococcus thermophilus* phages in the lysogenic state / M. Ventura [и др.] // Virology. — 2002. — Т. 302, № 1. — С. 21—32.
97. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157: H7 and genomic comparison with a laboratory strain K-12 / T. Hayashi [и др.] // DNA research. — 2001. — Т. 8, № 1. — С. 11—22.
98. *Schroven, K.* Bacteriophages as drivers of bacterial virulence and their potential for biotechnological exploitation / K. Schroven, A. Aertsen, R. Lavigne // FEMS Microbiology Reviews. — 2021. — Т. 45, № 1. — fuaa041.
99. *Brüssow, H.* Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion / H. Brüssow, C. Canchaya, W.-D. Hardt // Microbiology and molecular biology reviews. — 2004. — Т. 68, № 3. — С. 560—602.

100. Carriage of Shiga toxin phage profoundly affects *Escherichia coli* gene expression and carbon source utilization / P. Berger [и др.] // BMC genomics. — 2019. — Т. 20, № 1. — С. 1—14.
101. Phage as agents of lateral gene transfer / C. Canchaya [и др.] // Current opinion in microbiology. — 2003. — Т. 6, № 4. — С. 417—424.
102. *Touchon, M.* Embracing the enemy: the diversification of microbial gene repertoires by phage-mediated horizontal gene transfer / M. Touchon, J. A. M. De Sousa, E. P. Rocha // Current opinion in microbiology. — 2017. — Т. 38. — С. 66—73.
103. The Site-Specific Recombination System of the *Escherichia coli* Bacteriophage Φ24B / M. R. Mohaisen [и др.] // Frontiers in microbiology. — 2020. — Т. 11. — С. 2467.
104. *Ruzin, A.* Molecular genetics of SaPI1—a mobile pathogenicity island in *Staphylococcus aureus* / A. Ruzin, J. Lindsay, R. P. Novick // Molecular microbiology. — 2001. — Т. 41, № 2. — С. 365—377.
105. *Hatfull, G. F.* Bacteriophages and their genomes / G. F. Hatfull, R. W. Hendrix // Current opinion in virology. — 2011. — Т. 1, № 4. — С. 298—303.
106. The origins and ongoing evolution of viruses / R. W. Hendrix [и др.] // Trends in microbiology. — 2000. — Т. 8, № 11. — С. 504—508.
107. *Gillings, M. R.* Integrons: past, present, and future / M. R. Gillings // Microbiology and Molecular Biology Reviews. — 2014. — Т. 78, № 2. — С. 257—277.
108. Integrons: mobilizable platforms that promote genetic diversity in bacteria / Y. Boucher [и др.] // Trends in microbiology. — 2007. — Т. 15, № 7. — С. 301—309.
109. *Cambray, G.* Integrons / G. Cambray, A.-M. Guerout, D. Mazel // Annual review of genetics. — 2010. — Т. 44. — С. 141—166.
110. *Kovach, M. E.* A putative integrase gene defines the distal end of a large cluster of ToxR-regulated colonization genes in *Vibrio cholerae* / M. E. Kovach, M. D. Shaffer, K. M. Peterson // Microbiology. — 1996. — Т. 142, № 8. — С. 2165—2174.

111. *Boyd, E. F.* Genomic islands are dynamic, ancient integrative elements in bacterial evolution / E. F. Boyd, S. Almagro-Moreno, M. A. Parent // Trends in microbiology. — 2009. — Т. 17, № 2. — С. 47—53.
112. *Langille, M. G.* Detecting genomic islands using bioinformatics approaches / M. G. Langille, W. W. Hsiao, F. S. Brinkman // Nature Reviews Microbiology. — 2010. — Т. 8, № 5. — С. 373—382.
113. panRGP: a pangenome-based method to predict genomic islands and explore their diversity / A. Bazin [и др.] // Bioinformatics. — 2020. — Т. 36, Supplement\_2. — С. i651—i658.
114. Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an *Escherichia coli* wild-type pathogen. / G. Blum [и др.] // Infection and immunity. — 1994. — Т. 62, № 2. — С. 606—614.
115. The vibrio pathogenicity island of epidemic *Vibrio cholerae* forms precise extrachromosomal circular excision products / C. Rajanna [и др.] // Journal of bacteriology. — 2003. — Т. 185, № 23. — С. 6893—6901.
116. *Carpenter, M. R.* Pathogenicity island cross talk mediated by recombination directionality factors facilitates excision from the chromosome / M. R. Carpenter, S. Rozovsky, E. F. Boyd // Journal of bacteriology. — 2016. — Т. 198, № 5. — С. 766—776.
117. Molecular insights into the genome dynamics and interactions between core and acquired genomes of *Vibrio cholerae* / A. Pant [и др.] // Proceedings of the National Academy of Sciences. — 2020. — Т. 117, № 38. — С. 23762—23773.
118. The impact of insertion sequences on bacterial genome plasticity and adaptability / J. Vandecraen [и др.] // Critical reviews in microbiology. — 2017. — Т. 43, № 6. — С. 709—730.
119. Everyman's guide to bacterial insertion sequences / P. Siguier [и др.] // Mobile DNA III. — 2015. — С. 555—590.
120. *Oggioni, M. R.* Repeated extragenic sequences in prokaryotic genomes: a proposal for the origin and dynamics of the RUP element in *Streptococcus pneumoniae* / M. R. Oggioni, J.-P. Claverys // Microbiology. — 1999. — Т. 145, № 10. — С. 2647—2653.

121. *Siguier, P.* Bacterial insertion sequences: their genomic impact and diversity / P. Siguier, E. Gourbeyre, M. Chandler // FEMS microbiology reviews. — 2014. — T. 38, № 5. — C. 865—891.
122. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica* / J. Parkhill [и др.] // Nature genetics. — 2003. — T. 35, № 1. — C. 32—40.
123. *Glansdorff, N.* Activation of gene expression by IS2 and IS3 / N. Glansdorff, D. Charlier, M. Zafarullah // Cold Spring Harbor symposia on quantitative biology. T. 45. — Cold Spring Harbor Laboratory Press. 1981. — C. 153—156.
124. *Alton, N. K.* Nucleotide sequence analysis of the chloramphenicol resistance transposon Tn 9 / N. K. Alton, D. Vapnek // Nature. — 1979. — T. 282, № 5741. — C. 864—869.
125. Transposases are responsible for the target specificity of IS 1397 and ISKpn 1 for two different types of palindromic units (PUs) / C. Wilde [и др.] // Nucleic acids research. — 2003. — T. 31, № 15. — C. 4345—4353.
126. *Gogarten, J. P.* Prokaryotic evolution in light of gene transfer / J. P. Gogarten, W. F. Doolittle, J. G. Lawrence // Molecular biology and evolution. — 2002. — T. 19, № 12. — C. 2226—2238.
127. Phylogenetic Evidence for Horizontal Transfer of *mutS* Alleles among Naturally Occurring *Escherichia coli* Strains / E. W. Brown [и др.] // Journal of Bacteriology. — 2001. — T. 183, № 5. — C. 1631—1644.
128. *Caporale, L. H.* The implicit genome / L. H. Caporale. — Oxford University Press, 2006. — C. 112—117.
129. *Cohen, O.* The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer / O. Cohen, U. Gophna, T. Pupko // Molecular biology and evolution. — 2011. — T. 28, № 4. — C. 1481—1489.
130. *Novick, A.* Horizontal persistence and the complexity hypothesis / A. Novick, W. F. Doolittle // Biology & Philosophy. — 2020. — T. 35, № 1. — C. 2.
131. DNA uptake sequences in *Neisseria gonorrhoeae* as intrinsic transcriptional terminators and markers of horizontal gene transfer / R. Spencer-Smith [и др.] // Microbial genomics. — 2016. — T. 2, № 8.

132. Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome / H. O. Smith [и др.] // Science. — 1995. — Т. 269, № 5223. — С. 538—540.
133. Dialects of the DNA uptake sequence in Neisseriaceae / S. A. Frye [и др.] // PLoS Genet. — 2013. — Т. 9, № 4. — e1003458.
134. Biased distribution of DNA uptake sequences towards genome maintenance genes / T. Davidsen [и др.] // Nucleic acids research. — 2004. — Т. 32, № 3. — С. 1050—1058.
135. The impact of the neisserial DNA uptake sequences on genome evolution and stability / T. J. Treangen [и др.] // Genome biology. — 2008. — Т. 9, № 3. — С. 1—17.
136. Смирнов, Г. Механизмы приобретения и потери генетической информации бактериальными геномами / Г. Смирнов // Успехи современной биологии. — 2008. — Т. 128, № 1. — С. 52—76.
137. Bacterial evolution through the selective loss of beneficial genes: trade-offs in expression involving two loci / E. R. Zinser [и др.] // Genetics. — 2003. — Т. 164, № 4. — С. 1271—1277.
138. The chromosomal organization of horizontal gene transfer in bacteria / P. H. Oliveira [и др.] // Nature communications. — 2017. — Т. 8, № 1. — С. 1—11.
139. Insertion hot spot for horizontally acquired DNA within a bidirectional small-RNA locus in *Salmonella enterica* / R. Balbontín [и др.] // Journal of bacteriology. — 2008. — Т. 190, № 11. — С. 4075—4078.
140. The aerobactin iron transport system genes in *Shigella flexneri* are present within a pathogenicity island / S. A. Vokes [и др.] // Molecular microbiology. — 1999. — Т. 33, № 1. — С. 63—73.
141. Fitch, W. M. Distinguishing homologous from analogous proteins / W. M. Fitch // Systematic zoology. — 1970. — Т. 19, № 2. — С. 99—113.
142. Advances and Applications in the Quest for Orthologs / N. Glover [и др.] // Molecular biology and evolution. — 2019. — Т. 36, № 10. — С. 2157—2164.
143. Theißen, G. Orthology: secret life of genes / G. Theißen // Nature. — 2002. — Т. 415, № 6873. — С. 741—741.

144. *Wolf, Y. I.* A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes / Y. I. Wolf, E. V. Koonin // Genome biology and evolution. — 2012. — Т. 4, № 12. — С. 1286—1294.
145. *Schreiber, F.* Hieranoid: hierarchical orthology inference / F. Schreiber, E. L. Sonnhammer // Journal of molecular biology. — 2013. — Т. 425, № 11. — С. 2072—2081.
146. *Roth, A. C.* Algorithm of OMA for large-scale orthology inference / A. C. Roth, G. H. Gonnet, C. Dessimoz // BMC bioinformatics. — 2008. — Т. 9, № 1. — С. 1—10.
147. *Gabaldón, T.* Large-scale assignment of orthology: back to phylogenetics? / T. Gabaldón // Genome biology. — 2008. — Т. 9, № 10. — С. 1—6.
148. *Dalquen, D. A.* Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals / D. A. Dalquen, C. Dessimoz // Genome biology and evolution. — 2013. — Т. 5, № 10. — С. 1800—1806.
149. *Tatusov, R. L.* A genomic perspective on protein families / R. L. Tatusov, E. V. Koonin, D. J. Lipman // Science. — 1997. — Т. 278, № 5338. — С. 631—637.
150. *Derelle, R.* Broccoli: combining phylogenetic and network analyses for orthology assignment / R. Derelle, H. Philippe, J. K. Colbourne // Molecular Biology and Evolution. — 2020. — Т. 37, № 11. — С. 3389—3396.
151. The COG database: a tool for genome-scale analysis of protein functions and evolution / R. L. Tatusov [и др.] // Nucleic acids research. — 2000. — Т. 28, № 1. — С. 33—36.
152. *Li, L.* OrthoMCL: identification of ortholog groups for eukaryotic genomes / L. Li, C. J. Stoeckert, D. S. Roos // Genome research. — 2003. — Т. 13, № 9. — С. 2178—2189.
153. *Emms, D. M.* OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy / D. M. Emms, S. Kelly // Genome biology. — 2015. — Т. 16, № 1. — С. 157.
154. *Buchfink, B.* Fast and sensitive protein alignment using DIAMOND / B. Buchfink, C. Xie, D. H. Huson // Nature methods. — 2015. — Т. 12, № 1. — С. 59—60.

155. *vanDongen, S.* A cluster algorithm for graphs / S. vanDongen // Information Systems [INS]. — 2000. — R 0010.
156. *Emms, D. M.* OrthoFinder: phylogenetic orthology inference for comparative genomics / D. M. Emms, S. Kelly // Genome biology. — 2019. — T. 20, № 1. — C. 1—14.
157. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences / M. Goodman [и др.] // Systematic Biology. — 1979. — T. 28, № 2. — C. 132—163.
158. Automatic genome-wide reconstruction of phylogenetic gene trees / I. Wapinski [и др.] // Bioinformatics. — 2007. — T. 23, № 13. — C. i549—i558.
159. *Rasmussen, M. D.* Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes / M. D. Rasmussen, M. Kellis // Genome research. — 2007. — T. 17, № 12. — C. 1932—1942.
160. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony / A.-C. Berglund-Sonnhammer [и др.] // Journal of molecular evolution. — 2006. — T. 63, № 2. — C. 240—250.
161. *Sevillya, G.* Detecting horizontal gene transfer: a probabilistic approach / G. Sevillya, O. Adato, S. Snir // BMC genomics. — 2020. — T. 21, № 1. — C. 1—11.
162. *Garcia-Vallvé, S.* Horizontal gene transfer in bacterial and archaeal complete genomes / S. Garcia-Vallvé, A. Romeu, J. Palau // Genome Research. — 2000. — T. 10, № 11. — C. 1719—1725.
163. Evidence for horizontal gene transfer in *Escherichia coli* speciation / C. Médigue [и др.] // Journal of molecular biology. — 1991. — T. 222, № 4. — C. 851—856.
164. *Marri, P. R.* Gene amelioration demonstrated: the journey of nascent genes in bacteria / P. R. Marri, G. B. Golding // Genome. — 2008. — T. 51, № 2. — C. 164—168.
165. *Koski, L. B.* Codon bias and base composition are poor indicators of horizontally transferred genes / L. B. Koski, R. A. Morton, G. B. Golding // Molecular biology and evolution. — 2001. — T. 18, № 3. — C. 404—412.

166. *Gogarten, J. P.* Horizontal gene transfer, genome innovation and evolution / J. P. Gogarten, J. P. Townsend // Nature Reviews Microbiology. — 2005. — T. 3, № 9. — C. 679—687.
167. *Daubin, V.* The source of laterally transferred genes in bacterial genomes / V. Daubin, E. Lerat, G. Perrière // Genome biology. — 2003. — T. 4, № 9. — C. 1—12.
168. *Vernikos, G. S.* Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands / G. S. Vernikos, J. Parkhill // Bioinformatics. — 2006. — T. 22, № 18. — C. 2196—2203.
169. An accurate genomic island prediction method for sequenced bacterial and archaeal genomes / D. Che [и др.] // Journal of Proteomics & Bioinformatics. — 2014. — T. 7, № 8. — C. 214.
170. GIPSY: genomic island prediction software / S. C. Soares [и др.] // Journal of biotechnology. — 2016. — T. 232. — C. 2—11.
171. Comparative Analysis of Genomic Island Prediction Tools / A. C. da Silva Filho [и др.] // Frontiers in genetics. — 2018. — T. 9. — C. 619.
172. *Tofigh, A.* Simultaneous identification of duplications and lateral gene transfers / A. Tofigh, M. Hallett, J. Lagergren // IEEE/ACM Transactions on Computational Biology and Bioinformatics. — 2010. — T. 8, № 2. — C. 517—535.
173. *Smith, M. W.* Evolution by acquisition: the case for horizontal gene transfers / M. W. Smith, D.-F. Feng, R. F. Doolittle // Trends in biochemical sciences. — 1992. — T. 17, № 12. — C. 489—493.
174. HGTree: database of horizontally transferred genes determined by tree reconciliation / H. Jeong [и др.] // Nucleic acids research. — 2016. — T. 44, № D1. — C. D610—D619.
175. *Bansal, M. S.* Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss / M. S. Bansal, E. J. Alm, M. Kellis // Bioinformatics. — 2012. — T. 28, № 12. — C. i283—i291.
176. GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss / B. Morel [и др.] // Molecular biology and evolution. — 2020. — T. 37, № 9. — C. 2763—2774.

177. ShadowCaster: Compositional Methods under the Shadow of Phylogenetic Models to Detect Horizontal Gene Transfers in Prokaryotes / D. Sánchez-Soto [и др.] // Genes. — 2020. — Т. 11, № 7. — С. 756.
178. *Ragan, M. A.* Do different surrogate methods detect lateral genetic transfer events of different relative ages? / M. A. Ragan, T. J. Harlow, R. G. Beiko // Trends in microbiology. — 2006. — Т. 14, № 1. — С. 4—8.
179. Mauve: multiple alignment of conserved genomic sequence with rearrangements / A. C. Darling [и др.] // Genome research. — 2004. — Т. 14, № 7. — С. 1394—1403.
180. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons / N.-F. Alikhan [и др.] // BMC genomics. — 2011. — Т. 12, № 1. — С. 402.
181. Genome rearrangements induce biofilm formation in *Escherichia coli* C—an old model organism with a new application in biofilm research / J. E. Król [и др.] // BMC genomics. — 2019. — Т. 20, № 1. — С. 1—18.
182. Characterization of a lytic bacteriophage as an antimicrobial agent for biocontrol of shiga toxin-producing *Escherichia coli* O145 strains / Y.-T. Liao [и др.] // Antibiotics. — 2019. — Т. 8, № 2. — С. 74.
183. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis / K. S. Makarova [и др.] // Nucleic acids research. — 2002. — Т. 30, № 2. — С. 482—496.
184. PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph / G. Gautreau [и др.] // PLoS computational biology. — 2020. — Т. 16, № 3. — e1007732.
185. Cactus: Algorithms for genome multiple sequence alignment / B. Paten [и др.] // Genome research. — 2011. — Т. 21, № 9. — С. 1512—1528.
186. Variation graph toolkit improves read mapping by representing genetic variation in the reference / E. Garrison [и др.] // Nature biotechnology. — 2018. — Т. 36, № 9. — С. 875—879.
187. *Chervy, M.* Adherent-Invasive *E. coli*: Update on the Lifestyle of a Troublemaker in Crohn's Disease / M. Chervy, N. Barnich, J. Denizot // International Journal of Molecular Sciences. — 2020. — Т. 21, № 10. — С. 3734.

188. Mucosal flora in inflammatory bowel disease / A. Swidsinski [и др.] // Gastroenterology. — 2002. — Т. 122, № 1. — С. 44—54.
189. Changes in the bacterial flora of the neoterminal ileum after ileocolonic resection for Crohn's disease / C. Neut [и др.] // The American journal of gastroenterology. — 2002. — Т. 97, № 4. — С. 939—946.
190. Complete genome sequence of Crohn's disease-associated adherent-invasive *E. coli* strain LF82 / S. Miquel [и др.] // PloS one. — 2010. — Т. 5, № 9.
191. The treatment-naive microbiome in new-onset Crohn's disease / D. Gevers [и др.] // Cell host & microbe. — 2014. — Т. 15, № 3. — С. 382—392.
192. Younis, N. Inflammatory bowel disease: between genetics and microbiota / N. Younis, R. Zarif, R. Mahfouz // Molecular biology reports. — 2020. — Т. 47, № 4. — С. 3053—3063.
193. Genome analysis of *E. coli* isolated from Crohn's disease patients / D. V. Rakitina [и др.] // BMC genomics. — 2017. — Т. 18, № 1. — С. 1—17.
194. Invasive ability of an *Escherichia coli* strain isolated from the ileal mucosa of a patient with Crohn's disease / J. Boudeau [и др.] // Infection and immunity. — 1999. — Т. 67, № 9. — С. 4499—4509.
195. Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease / M. Martinez-Medina [и др.] // Inflammatory bowel diseases. — 2009. — Т. 15, № 6. — С. 872—882.
196. Prevalence of the pathobiont adherent-invasive *Escherichia coli* and inflammatory bowel disease: a systematic review and meta-analysis / B. Nadalian [и др.] // Journal of Gastroenterology and Hepatology. — 2020.
197. Shaler, C. R. The unique lifestyle of Crohn's disease-associated adherent-invasive *Escherichia coli* / C. R. Shaler, W. Elhenawy, B. K. Coombes // Journal of molecular biology. — 2019. — Т. 431, № 16. — С. 2970—2981.
198. Comparative genomics reveals new single-nucleotide polymorphisms that can assist in identification of adherent-invasive *Escherichia coli* / C. Camprubí-Font [и др.] // Scientific reports. — 2018. — Т. 8, № 1. — С. 1—11.

199. Adherent-invasive *E. coli* metabolism of propanediol in Crohn's disease regulates phagocytes to drive intestinal inflammation / M. Viladomiu [и др.] // Cell Host & Microbe. — 2021.
200. Inflammation-associated adherent-invasive *Escherichia coli* are enriched in pathways for use of propanediol and iron and M-cell translocation / B. Dogan [и др.] // Inflammatory bowel diseases. — 2014. — Т. 20, № 11. — С. 1919—1932.
201. Respiration of microbiota-derived 1, 2-propanediol drives *Salmonella* expansion during colitis / F. Faber [и др.] // PLoS pathogens. — 2017. — Т. 13, № 1. — e1006129.
202. *Anast, J. M.* The cobalamin-dependent gene cluster of *Listeria monocytogenes*: Implications for virulence, stress response, and food safety / J. M. Anast, T. A. Bobik, S. Schmitz-Egger // Frontiers in microbiology. — 2020. — Т. 11.
203. Propanediol utilization genes (pdu) of *Salmonella typhimurium*: three genes for the propanediol dehydratase. / T. A. Bobik [и др.] // Journal of Bacteriology. — 1997. — Т. 179, № 21. — С. 6633—6639.
204. Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes / J. H. Nash [и др.] // BMC genomics. — 2010. — Т. 11, № 1. — С. 1—15.
205. Adherent-invasive *Escherichia coli* (AIEC) in pediatric Crohn's disease patients: phenotypic and genetic pathogenic features / M. P. Conte [и др.] // BMC research notes. — 2014. — Т. 7, № 1. — С. 1—12.
206. Genetic diversity and virulence determinants of *Escherichia coli* strains isolated from patients with Crohn's disease in Spain and Chile / S. Céspedes [и др.] // Frontiers in microbiology. — 2017. — Т. 8. — С. 639.
207. *Rolhion, N.* OmpC and the σE regulatory pathway are involved in adhesion and invasion of the Crohn's disease-associated *Escherichia coli* strain LF82 / N. Rolhion, F. A. Carvalho, A. Darfeuille-Michaud // Molecular microbiology. — 2007. — Т. 63, № 6. — С. 1684—1700.
208. MUMmer4: A fast and versatile genome alignment system / G. Marçais [и др.] // PLoS computational biology. — 2018. — Т. 14, № 1. — e1005944.

209. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes / T. J. Treangen [и др.] // Genome biology. — 2014. — Т. 15, № 11. — С. 524.
210. Seemann, T. Prokka: rapid prokaryotic genome annotation / T. Seemann // Bioinformatics. — 2014. — Т. 30, № 14. — С. 2068—2069.
211. Cytoscape 2.8: new features for data integration and network visualization / M. E. Smoot [и др.] // Bioinformatics. — 2011. — Т. 27, № 3. — С. 431—432.
212. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput / R. C. Edgar // Nucleic acids research. — 2004. — Т. 32, № 5. — С. 1792—1797.
213. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies / L.-T. Nguyen [и др.] // Molecular biology and evolution. — 2015. — Т. 32, № 1. — С. 268—274.
214. PHASTER: a better, faster version of the PHAST phage search tool / D. Arndt [и др.] // Nucleic acids research. — 2016. — Т. 44, W1. — W16—W21.
215. Condensin-and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging / M. Marbouty [и др.] // Molecular cell. — 2015. — Т. 59, № 4. — С. 588—602.
216. Genome sequence assembly using trace signals and additional sequence information. / B. Chevreux, T. Wetter, S. Suhai [и др.] // German conference on bioinformatics. Т. 99. — Citeseer. 1999. — С. 45—56.
217. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing / A. Bankevich [и др.] // Journal of computational biology. — 2012. — Т. 19, № 5. — С. 455—477.
218. Complete genome sequence of adherent invasive Escherichia coli UM146 isolated from Ileal Crohn's disease biopsy tissue / D. O. Krause [и др.] // Journal of bacteriology. — 2011. — Т. 193, № 2. — С. 583—583.
219. DOOR 2.0: presenting operons and their functions through dynamic and integrated views / X. Mao [и др.] // Nucleic acids research. — 2014. — Т. 42, № D1. — С. D654—D659.

220. A genome-wide identification of genes undergoing recombination and positive selection in *Neisseria* / D. Yu [и др.] // BioMed research international. — 2014. — Т. 2014.
221. *Pseudomonas aeruginosa* is capable of natural transformation in biofilms / L. M. Nolan [и др.] // bioRxiv. — 2019. — C. 859553.
222. Clarke, B. R. Genetic organization of the *Escherichia coli* K10 capsule gene cluster: identification and characterization of two conserved regions in group III capsule gene clusters encoding polysaccharide transport functions / B. R. Clarke, R. Pearce, I. S. Roberts // Journal of bacteriology. — 1999. — Т. 181, № 7. — C. 2279—2285.
223. Masquerading microbial pathogens: capsular polysaccharides mimic host-tissue molecules / B. F. Cress [и др.] // FEMS microbiology reviews. — 2014. — Т. 38, № 4. — C. 660—697.
224. Lukáčová, M. Role of structural variations of polysaccharide antigens in the pathogenicity of Gram-negative bacteria / M. Lukáčová, I. Barak, J. Kazar // Clinical microbiology and infection. — 2008. — Т. 14, № 3. — C. 200—206.
225. A novel genetic island of meningitic *Escherichia coli* K1 containing the ibeA invasion gene (GimA): functional annotation and carbon-source-regulated invasion of human brain microvascular endothelial cells / S.-H. Huang [и др.] // Functional & integrative genomics. — 2001. — Т. 1, № 5. — C. 312—322.
226. Tukey, J. W. Exploratory data analysis. Т. 2 / J. W. Tukey. — Reading, Mass., 1977.
227. Benjamini, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing / Y. Benjamini, Y. Hochberg // Journal of the Royal statistical society: series B (Methodological). — 1995. — Т. 57, № 1. — C. 289—300.
228. The control of the false discovery rate in multiple testing under dependency / Y. Benjamini, D. Yekutieli [и др.] // The annals of statistics. — 2001. — Т. 29, № 4. — C. 1165—1188.
229. Holm, S. A simple sequentially rejective multiple test procedure / S. Holm // Scandinavian journal of statistics. — 1979. — C. 65—70.
230. Ely, B. Recombination and gene loss occur simultaneously during bacterial horizontal gene transfer / B. Ely // PloS one. — 2020. — Т. 15, № 1. — e0227987.

231. *Hacker, W. C.* Features of genomic organization in a nucleotide-resolution molecular model of the *Escherichia coli* chromosome / W. C. Hacker, S. Li, A. H. Elcock // Nucleic acids research. — 2017. — T. 45, № 13. — C. 7541—7554.

**Глава 6. Благодарности**

Автор выражает глубокую благодарность своему научному руководителю и коллегам, принимавшим участие в проведении работы (их участие отмечено при описании материалов и методов.).

**Глава 7. Финансирование**

Работа выполнена при поддержке гранта российского научного фонда №16-15-00258 “*E. coli* как мишень терапии при болезни Крона” (руководитель Побегуц О.В.).