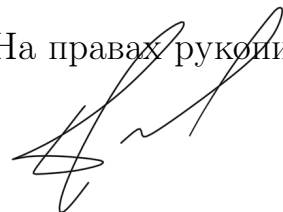


Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

На правах рукописи



Бочкарёва Ольга Олеговна

Реконструкция эволюционной истории геномных перестроек в бактериях

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени
кандидата биологических наук

Москва – 2019

Работа выполнена в *Федеральном государственном бюджетном учреждении науки*
«Институт проблем передачи информации имени А.А. Харкевича
Российской академии наук».

Научный руководитель: **Гельфанд Михаил Сергеевич**

кандидат физико-математических наук, доктор биологических наук, зам. директора по научной работе в Федеральном государственном бюджетном учреждении науки «Институт проблем передачи информации имени А.А. Харкевича Российской академии наук»

Официальные оппоненты: **Карягина-Жулина Анна Станиславовна**

доктор биологических наук, главный научный сотрудник лаборатории биологически активных наноструктур Федерального государственного бюджетного учреждения «Научно-исследовательский институт эпидемиологии и микробиологии им. почетного академика Н.Ф. Гамалеи» Министерства здравоохранения РФ.

Райко Михаил Петрович

кандидат биологических наук, старший научный сотрудник центра алгоритмической биотехнологии Санкт-Петербургского государственного университета

Ведущая организация:

Институт биофизики клетки Российской академии наук – обособленное подразделение Федерального государственного бюджетного учреждения науки «Федеральный исследовательский центр «Пущинский научный центр биологических исследований Российской академии наук».

Защита состоится «_____» _____ 2019 г. в _____ часов на заседании диссертационного совета Д.002.077.04 при *Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук*, расположенном по адресу: *127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.* С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук, а также на сайте <http://www.iitp.ru>. Автореферат разослан «_____» _____ 2019 г.

Ученый секретарь диссертационного совета Д.002.077.04,
доктор биологических наук, профессор,

Рожкова Г.И.

Общая характеристика работы

Актуальность работы

Расшифровка большого количества прокариотических геномов позволила проводить сравнения геномов близких видов и штаммов бактерий. Такой анализ необходим для уточнения видовой классификации бактерий, понимания механизмов вирулентности патогенов, описания эволюции устойчивости к антибиотикам, понимания метаболических различий между штаммами. Кроме того, восстановление истории эволюционных отношений близких геномов позволяет выявлять механизмы отбора и видообразования.

Большая часть исследований молекулярной эволюции бактерий сконцентрирована на анализе мутаций в белок-кодирующих областях или изменении генного состава. Однако в ходе эволюции структура самой хромосомы существенно меняется за счет внутригеномной рекомбинации. Как следствие, возникают геномные перестройки, такие как крупные делеции, инверсии, транслокации и дупликации.

Геномные перестройки играют большую роль в эволюции бактерий, поскольку могут разрушать существующие гены и опероны, создавать новые гены, менять расстояние от гена до точки начала репликации и тем самым менять уровень его экспрессии, менять цепь (лидирующую или запаздывающую), на которой находится ген, что также влияет на экспрессию.

Геномные перестройки способствуют быстрой эволюции патогенных бактерий, в том числе позволяя им быстро приобретать устойчивость к антибиотикам. В частности, именно геномные перестройки сыграли ключевую роль в видообразовании и патогенизации *Yersinia pestis* (возбудителя чумы). В видообразовании внутриклеточных патогенов, таких как *Burkholderia mallei*, уменьшение размера генома благодаря крупным делециям дает существенное эволюционное преимущество.

Хотя на момент начала работы над диссертацией было выявлено и описано

большое количество перестроек на основе парных сравнений штаммов, общий ландшафт геномных перестроек оставался слабо изученным. Для понимания динамики бактериальной эволюции необходимо изучать геномные перестройки в контексте бактериальной эволюции, что стало возможным благодаря накоплению достаточного количества бактериальных данных и разработке соответствующих математических моделей и алгоритмов.

Более того, восстановление филогенетических отношений для близких штаммов бактерий часто остается трудновыполнимой задачей, поскольку традиционные методы филогении строятся на выравнивании последовательностей отдельных генов с последующим объединением полученных выравниваний. Такие методы, с одной стороны, чувствительны к действию гомологичной рекомбинации, а с другой – имеют малую разрешающую способность на малых филогенетических расстояниях (штаммы одного вида) и могут давать недостоверные результаты. Информация об изменении порядка генов на хромосомах может быть использована для верификации топологии, полученной в рамках традиционной модели, и уточнения узлов с низким уровнем бутстреп-поддержки.

Таким образом, анализ геномных перестроек в эволюции бактерий из разных экологических ниш, выполненный в данной работе, позволил описать эволюционные процессы в геномах и выявить основные закономерности в действии эволюционного отбора на разные типы событий.

Цели и задачи диссертационной работы

Целью работы были реконструкция и анализ эволюционной истории геномных перестроек для нескольких бактериальных родов и видов и выявление действия эволюционного отбора на разные типы событий.

В работе были решены следующие задачи.

1. Изучить существующие алгоритмы для работы с порядком генов в бактериальных геномах и разработать общую методику восстановления и ана-

лиза эволюционной истории перестроек.

2. Реконструировать и описать эволюционную историю геномных перестроек в нескольких родах и видах бактерий с большим числом отсеквенированных полностью собранных геномов (*Yersinia pestis*, *Burkholderia* spp., *Streptococcus* spp., *Chlamydia* spp.).
3. Реконструировать филогенетические деревья на основе расположения синтенных блоков и сравнить результаты реконструкции с филогенетическими деревьями, полученными традиционными методами; предложить гипотезы, объясняющие выявленные противоречия.
4. На основе паттернов перестроек выявить действие эволюционного отбора на разные типы событий.
5. Выявить перестройки, произошедшие независимо на разных ветках эволюционных деревьев штаммов, и описать их геномный контекст, поскольку такие события могут повышать приспособленность вида и потому поддерживаться эволюционным отбором.
6. Разработать методику оценки потока генов, горизонтально переносящихся в популяцию, по данным о взаимной расположении генов на хромосомах и протестировать ее на нескольких бактериальных видах.

Научная новизна. В диссертационной работе описан интегральный анализ геномных перестроек в контексте эволюционной истории с использованием различных подходов и инструментов. Разработанные методики протестированы на геномах представителей нескольких бактериальных родов из различных сред обитания.

Показано, что филогенетическая реконструкция на основе расположения генов позволяет уточнить филогению молодых патогенов с высокой частотой внутригеномной рекомбинации. Впервые показано действие отбора на сохране-

ние положения гена относительно лидирующей/запаздывающей цепи при транс-локациях между хромосомами в мультихромосомных бактериях. Описаны новые случаи фазовой вариации и распространения геномных островков в патогенах человека *Streptococcus pneumoniae* и *Burkholderia pseudomallei*.

Разработана методика оценки потока генов, распространяющихся в популяции путем горизонтальных переносов, на основе анализа порядка генов в геномах. Методика протестирована на двух бактериальных видах из разных экологических ниш.

Практическая ценность. Работа имеет преимущественно теоретический характер, однако полученные знания могут быть использованы в генной инженерии и медицине. Тщательное изучение всех механизмов геномной эволюции и факторов, влияющих на пластичность бактериальных геномов, позволит улучшить методики получения промышленных штаммов.

Факторы вирулентности патогенов, такие как избегание иммунной системы и выживание в организме хозяина, также определяются архитектурой бактериальных геномов. Обнаружение новых случаев переноса геномных островков и антигенной вариации за счет инверсий необходимо для успешной разработки лекарств и вакцин в силу все новых адаптационных изменений патогенных бактерий.

Положения, выносимые на защиту:

1. Для штаммов с высокой частотой внутригеномной рекомбинации продемонстрировано, что интегральный анализ филогенетических деревьев, построенных на основе выравнивания последовательностей общих генов и на основе расположения синтенных блоков на хромосомах, позволяет уточнить эволюционную историю изучаемых организмов.
2. Анализ эволюционной истории перестроек в контексте филогении позволил выявить новые геномные островки в штаммах мульти-хромосомных бактерий *Burkholderia* и новые случаи антигенной вариации за счет

инверсий в патогенах человека *Streptococcus pneumoniae* и *Burkholderia pseudomallei*. Предложенная в диссертации методика биоинформатического поиска генов-кандидатов позволяет быстро протестировать большое количество геномов и выявить случаи, требующие дальнейшего экспериментального подтверждения.

3. Показано, что инверсии, лежащие внутри одной реплихоры, значимо короче инверсий, концы которых лежат на разных реплихорах, а также такие инверсии встречаются реже, чем это ожидается из случайной модели. Выявлен отбор на сохранение положения генов в транслоцированных сегментах относительно лидирующей / запаздывающей цепи при межхромосомных транслокациях.
4. Разработана методика оценки потока генов, распространяющихся в популяции путем горизонтальных переносов, построенная на сравнении двух характеристик: согласованности топологии дерева гена с деревом видов и положении гена в геномах относительно стабильных локусов. Методика протестирована на двух родах бактерий с различной геномной архитектурой: *Streptococcus* spp. и *Chlamydia* spp.

Степень достоверности и апробация результатов. Результаты диссертации докладывались на российских и международных конференциях:

устные доклады на “Хромосома” (20 – 24 августа 2018, Новосибирск), “Bioinformatics: from Algorithms to Applications” (BiATA) (16 – 19 июля 2018, Санкт-Петербург), 23th European Meeting of PhD Students in Evolutionary Biology (EMPSEB) (10 – 15 сентября 2017, Краков, Польша), 21ой Международной пушинской школе-конференции молодых ученых “Биология наука 21 века” (17 – 21 апреля 2017, Пущино), III Пушинской школе-конференции “Биохимия, физиология и биосферная роль микроорганизмов” (5 – 9 декабря 2016, Пущино);

постерные доклады на “Bioinformatics of Genome Regulation and Structure / Systems Biology” BGRS SB (20 – 25 августа 2018, Новосибирск), 8th Moscow

Conference of Computational Molecular Biology (MCCMB) (27 – 30 июля 2017, Москва), Информационные технологии и системы (ИТиС) (6 – 11 сентября 2015, Сочи), 7th Moscow Conference on Computational Molecular Biology (MCCMB) (16 – 19 июля 2015, Москва), 6th Moscow Conference on Computational Molecular Biology (MCCMB) (25 – 28 июля 2013, Москва).

Личный вклад автора. Личный вклад соискателя состоит в непосредственном планировании исследований, формулировке гипотез, теоретической разработке и практической реализации методов, формулировании результатов и выводов, подготовке и публикации научных статей. Все представленные в диссертации результаты получены лично автором.

Структура и объем диссертации. Диссертация изложена на 111 страницах. Она состоит из девяти разделов: введение, главы 1-5, заключение, список литературы и приложение. Глава 1 содержит обзор литературы по теме диссертации. Глава 2 содержит описание используемых методов и программ. Главы с 3 по 5 содержат описание собственных исследований. Работа содержит 37 рисунков и 5 таблиц. Список литературы содержит 112 наименований. Приложение содержит 5 таблиц.

Содержание работы

Во Введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

В первой главе приведен обзор литературы по теме диссертации, состоящий из трех разделов. Обзор начинается с описания методов изучения геномных перестроек. Приводится описание разных типов геномных перестроек и математических моделей, используемых для реконструкции эволюции геномов. Далее приводится обзор биоинформатических алгоритмов, используемых для

разбиения геномов на синтенные блоки и реконструкции геномных перестроек, указаны достоинства и недостатки описанных подходов.

Во втором разделе приводится описание различных механизмов эволюции бактериальных геномов, приводящих к изменению порядка генов на хромосомах. Описаны примеры изменения бактериального фенотипа за счет сайт-специфической рекомбинации и горизонтального переноса генов. Описана архитектура бактериальной кольцевой хромосомы и обсуждаются гипотезы о действии отбора на закрепление геномных перестроек в зависимости от положения на хромосоме сегментов, затронутых перестройкой. Заканчивается раздел описанием структуры и эволюции пан-генома бактериальной популяции.

Последний раздел посвящен описанию характеристик бактериальных видов, использованных в работе. Приведена информация об их таксономии, патогенности, приспособленности к условиям среды и особенностях геномной архитектуры.

Во второй главе описаны программы и методы, использованные в диссертации для восстановления геномных перестроек в бактериальных геномах.

В работе были использованы пять наборов полностью собранных бактериальных геномов, доступных в базе данных биологических последовательностей NCBI: 12 штаммов *Yersinia pestis*, 103 штамма *Streptococcus* spp., 127 штаммов *Burkholderia* spp., 127 штаммов *Chlamydia* spp., 41 штамм *Escherichia coli* и *Salmonella enterica*. Ортологические ряды были построены с помощью Protein-Ortho со следующими параметрами: минимальный процент совпадений в аминокислотном выравнивании не менее 50% и покрытие не менее 67%.

При работе с штаммами одного вида или близкими штаммами одного рода для построения синтенных блоков было использовано полногеномное выравнивание с помощью алгоритма Sibelia, разработанного специально для анализа бактериальных геномов. Алгоритм позволяет работать с геномами, состоящими из нескольких хромосом, а также находить как общие для всех изучаемых геномов синтенные области, так и области, присутствующие только в части дан-

ных. Разбиение геномов на синтенные области, присутствующие во всех изучаемых штаммах, позволяет выявлять инверсии и транслокации, в то время как информация о синтенных областях, присутствующих только в части геномов, используется для анализа эволюционной истории крупных вставок и делеций.

Для далеких видов с низким уровнем сходства генного состава была использована модель построения синтенных блоков на основе порядка универсальных (общих и не имеющих паралогов) генов, реализованная в алгоритме DRIMM-Synteny. В синтенные блоки объединялись два и более общих гена, сохраняющие свое взаимное расположение во всех рассматриваемых геномах.

Для восстановления топологии филогенетического дерева по данным расположения синтенных блоков в геномах были использованы метод максимального правдоподобия, реализованный в алгоритме MLGO, и байесовский подход, реализованный в алгоритме Badger. Реконструированные филогенетические деревья сравнивались с результатами, полученными с помощью традиционных моделей восстановления филогении штаммов на основе объединения выравниваний генов из универсальных ортологических рядов.

Для реконструкции расположения синтенных блоков в предковых штаммах был использован алгоритм MGRA. На основе заданных топологии филогенетического дерева и расположения синтенных блоков в геномах на листьях этого дерева алгоритм позволяет реконструировать расположение синтений в узлах дерева и цепочку эволюционных событий на ветвях, таких как инверсии, потери, приобретения и транслокации.

Точки начала и конца репликации для определения последовательностей, принадлежащих лидирующей и запаздывающей цепи, были определены с помощью вычисления функции $GC-skew = ([G] - [C]) / ([G] + [C])$, где $[G]$ и $[C]$ – количество нуклеотидов G и C в подпоследовательностях ДНК с шагом 1000 п.н. Этот подход основывается на наблюдении, что для бактериальных хромосом характерна перепредставленность гуанина на лидирующей и цитозина на запаздывающей цепи. Таким образом, точки максимума и минимума этой функ-

ции соответствуют точкам терминации и начала репликации.

Статистическая значимость перепредставленности инверсий с концами на разных реплихорах вычислялась, исходя из вероятности получить такое количество инверсий этого класса при случайном накидывании инверсий заданной длины на хромосому заданной длины. Вероятность отдельно взятой инверсии пройти через точку начала репликации вычислялась как отношение длины инверсии к длине реплихоры, что соответствует случайной модели.

В третьей главе представлены результаты применения алгоритмов реконструкции филогении по данным о расположении порядка генов к геномам *Y. pestis*. Референсное филогенетическое дерево для штаммов *Y. pseudotuberculosis* и *Y. pestis* было построено по выравниваниям 2408 ортологических рядов, имеющих ровно по одному представителю в каждом виде (Рис. 1А).

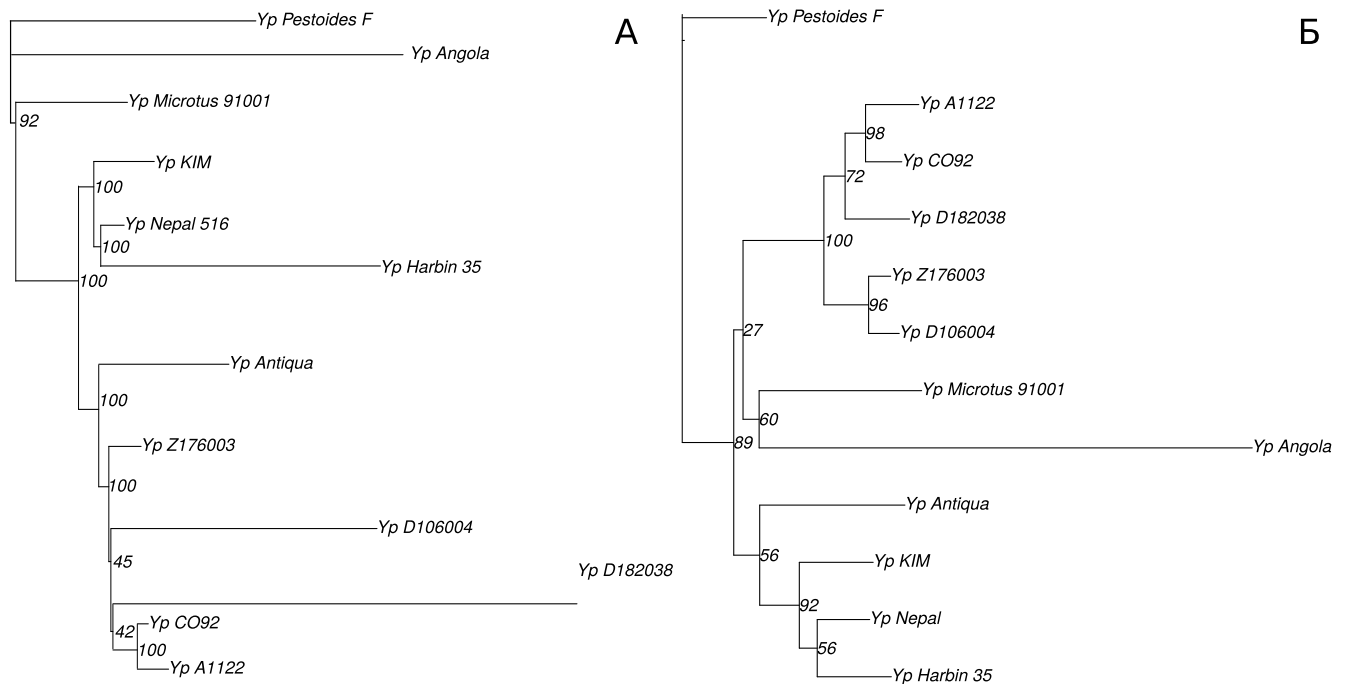


Рис. 1. Филогенетические деревья *Yersinia* spp., А) дерево построено по конкатенату нуклеотидных выравниваний однокопийных универсальных генов; Б) дерево построено на основе положения синтенных блоков.

Для того, чтобы разрешить узлы с низким уровнем бутстреп-поддержки, было реконструировано дерево на основе порядка генов (синтений). С помо-

щью полногеномных выравниваний геномы *Y. pestis* были разбиты на 123 общих синтенных блока длиной более 5000 п.н., общих для всех геномов. Для точной филогенетической реконструкции была применена байесовская модель, далее полученные филогенетические деревья с наибольшими апостериорными вероятностями были визуализированы в виде консенсусной сети (Рис. 2).

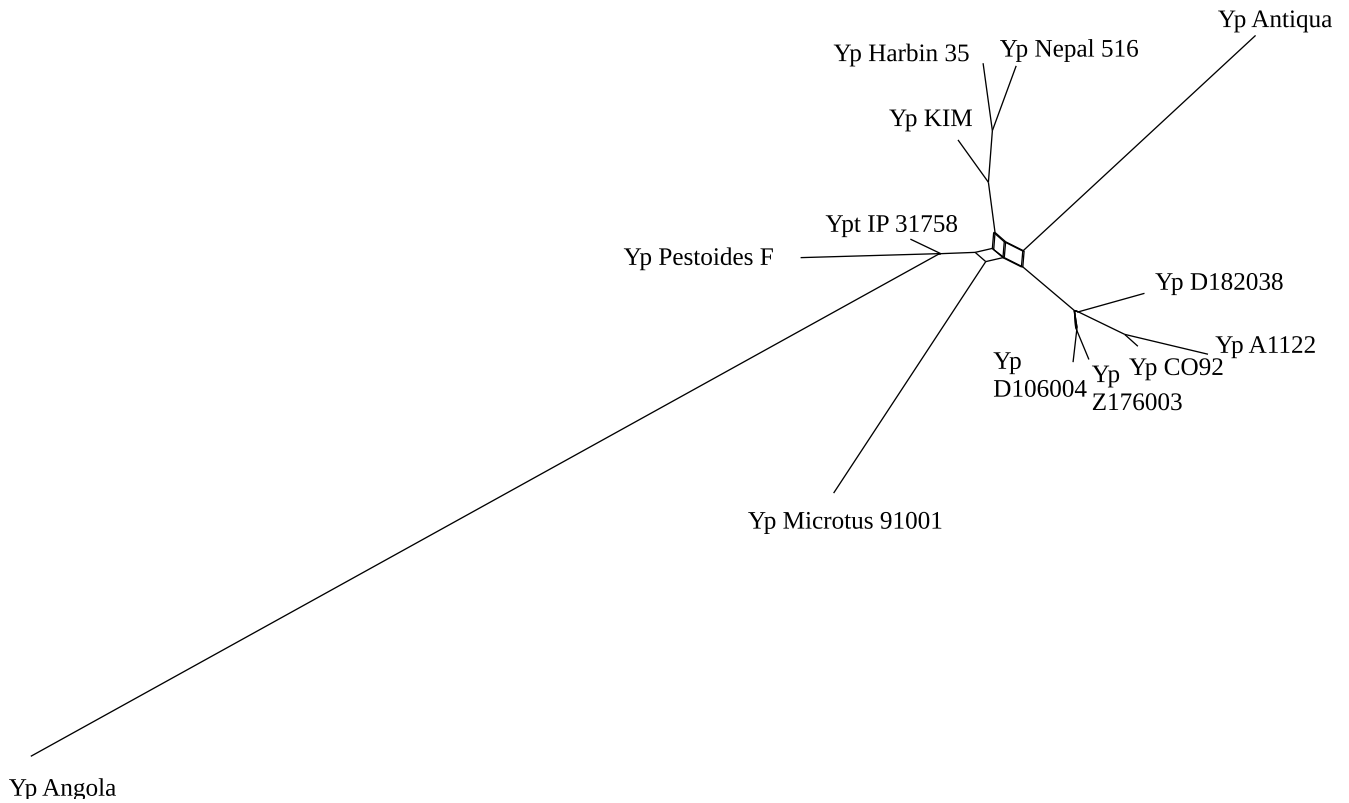


Рис. 2. Филогенетическая сеть *Yersinia pestis*, построенная с пороговым значением апостериорной вероятности 0.1.

Филогенетическая сеть имеет сложную, разветвленную структуру, не позволяющую однозначно определить положение длинных ветвей на итоговом дереве, однако на нем хорошо разрешены клады с близкими штаммами, в том числе те узлы, которые имели низкую бутстреп-поддержку на дереве, построенном по выравниваниям генов.

Применение метода максимального правдоподобия для построения филогении по расположению общих синтенных блоков позволило получить оптимальную топологию филогенетического дерева (Рис. 1Б). На этом дереве штаммы *Y. pestis* D106004 и *Y. pestis* Z176003 формируют отдельную ветку, штамм *Y. pestis*

D182038 в данной топологии является внешним. Это произошло вследствие общей инверсии длиной порядка 150 т.п.н. в штаммах *Y. pestis* D106004 и *Y. pestis* Z176003. При этом на дереве, построенном по выравниваниям генов, внешним является *Y. pestis* Z176003, но поскольку этот узел имеет низкий уровень бутстреп-поддержки, это может быть вызвано недостаточным количеством замен для восстановления филогении молодых штаммов (Рис. 1А).

Еще одна общая (параллельная) инверсия длиной порядка 350 т.п.н. найдена у штаммов *Y. pestis* A1122 и *Y. pestis* D182038, что приводит к низкому уровню бутстреп-поддержки в этом узле на дереве, отражающем историю инверсий. Границы всех найденных инверсий сформированы генами, кодирующими транспозазы.

Поскольку количество выявленных перестроек чувствительно к выбранному пороговому значению длины синтенного блока, была выполнена проверка устойчивости результата филогенетической реконструкции к этому параметру.

Результаты реконструкции подтверждают, что алгоритмы построения филогении по данным о расположении синтенных блоков позволяют уточнить филогению молодых штаммов *Y. pestis*, но могут давать ложное описание положения длинных ветвей, близких к корню дерева. Таким образом, оптимальной исследовательской стратегией является сочетание двух моделей и анализ расхождений топологий.

В четвертой главе описываются результаты восстановления истории геномных перестроек в штаммах *Streptococcus* spp. и *Burkholderia* spp.

Выравнивания хромосом стрептококков трех видов (*S. pneumoniae*, *S. suis* и *S. pyogenes*) показали сильное сходство порядка генов в штаммах. Количество перестроек оказалось недостаточным для реконструкции топологии филогенетического дерева по данным о расположении синтенных блоков, поэтому история перестроек была восстановлена по топологии, полученной на основе объединения выравниваний общих однокопийных ортологических рядов. Было реконструировано 13 инверсий в *S. pneumoniae*, 21 инверсия в *S. suis* и 26 ин-

версий в *S. pyogenes*, в эволюционной истории двух видов стрептококков были детектированы параллельные события.

Границы всех инверсий в стрептококках сформированы повторяющимися элементами генома, такими как мобильные элементы (фаговые белки, транспозазы и т. д.) и/или кластерами, состоящими из генов 16S и 23S рРНК. Единственное исключение — инверсия в *S. pneumoniae* длиной около 15 т.п.н., детектированная на четырех различных ветках дерева, границы которой проходят по паралогам, кодирующим поверхностные антигены PhtB и PhtD. Чтобы точно определить положение точки рекомбинации внутри этих паралогов, для четырех штаммов, где была детектирована эта инверсия, и их внешних таксонов было выполнено выравнивание нуклеотидных последовательностей паралогов, далее процент несовпадений в этих выравниваниях с шагом 10 п.н. был нанесен на график (Рис. 4). Этот анализ показал, что граница инверсии проходит внутри генов, кодирующих PhtB и PhtD, и инверсия приводит к замене их сегментов.

По результатам картирования инверсий на хромосомы 57 инверсий из 62 проходят через точку старта репликации, что подтверждает действие сильного отрицательного отбора на инверсии, лежащие внутри одной реплихоры. Вероятность такого исхода в случайной модели составляет $p = 9 \cdot 10^{-14}$.

В мультихромосомных штаммах *Burkholderia* на основе распределения 1024 универсальных генов по хромосомам была выполнена реконструкция межхромосомных транслокаций, включившая в себя 210 событий на филогенетическом дереве. Анализ положения генов на цепях показал действие отбора на сохранение положения транслоцированных генов относительно лидирующей/запаздывающей цепи (двусторонний биномиальный тест, $p = 0.03$), и эта тенденция сильнее выражена для более длинных блоков (Рис. 5А).

В штаммах *B. mallei* реконструкция истории инверсий согласно топологии дерева, полученной по выравниваниям нуклеотидных последовательностей общих генов, выявила 88 инверсий на первых и 27 инверсий на вторых хромо-

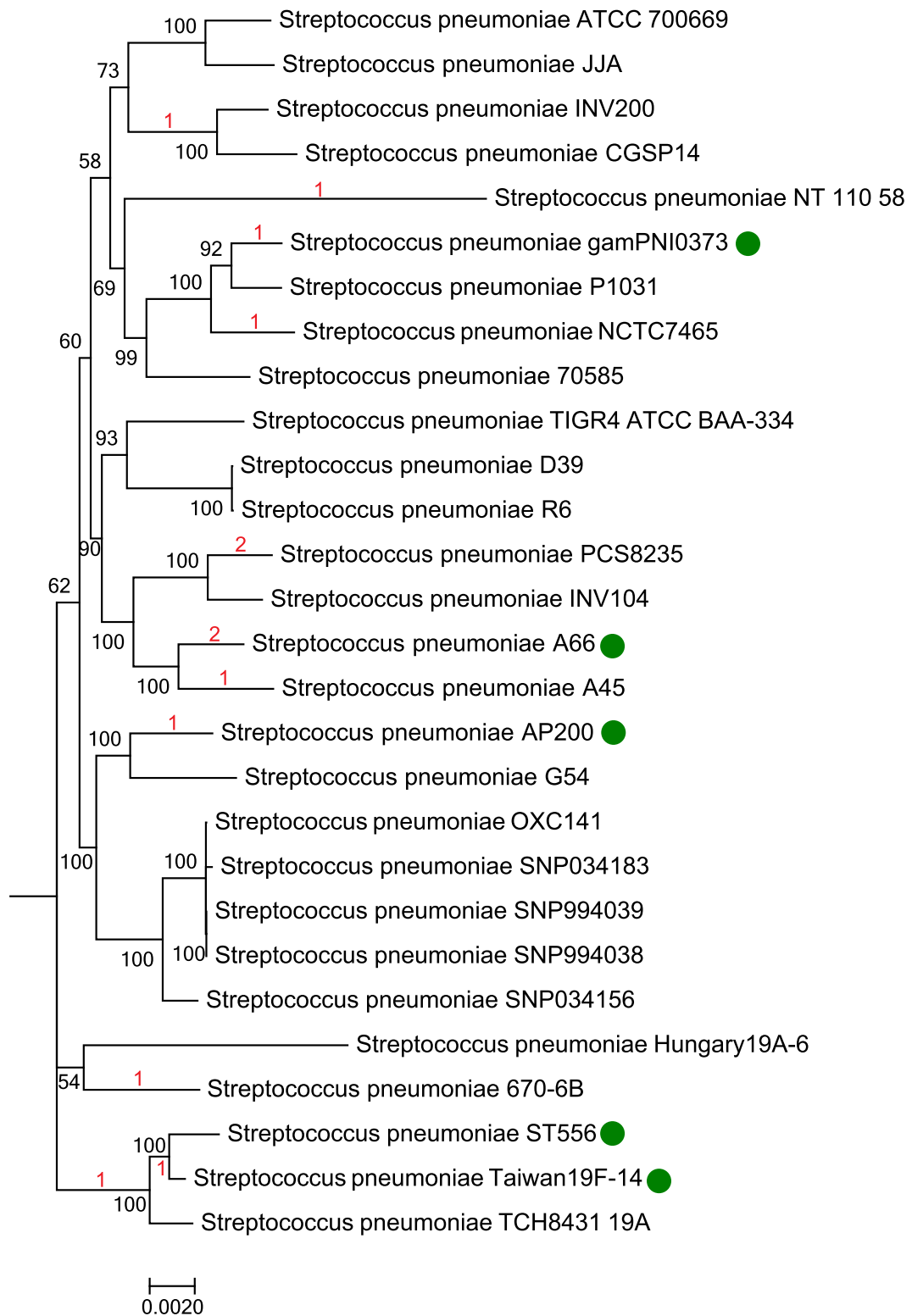


Рис. 3. Филогенетическое дерево *S. pneumoniae*, построенное по конкатенату нуклеотидных выравниваний однокопийных универсальных генов. Красные числа над ветками дерева показывают количество инверсий, реконструированных на разных ветках; цветными кругами помечены штаммы, в которых детектированы одинаковые инверсии.

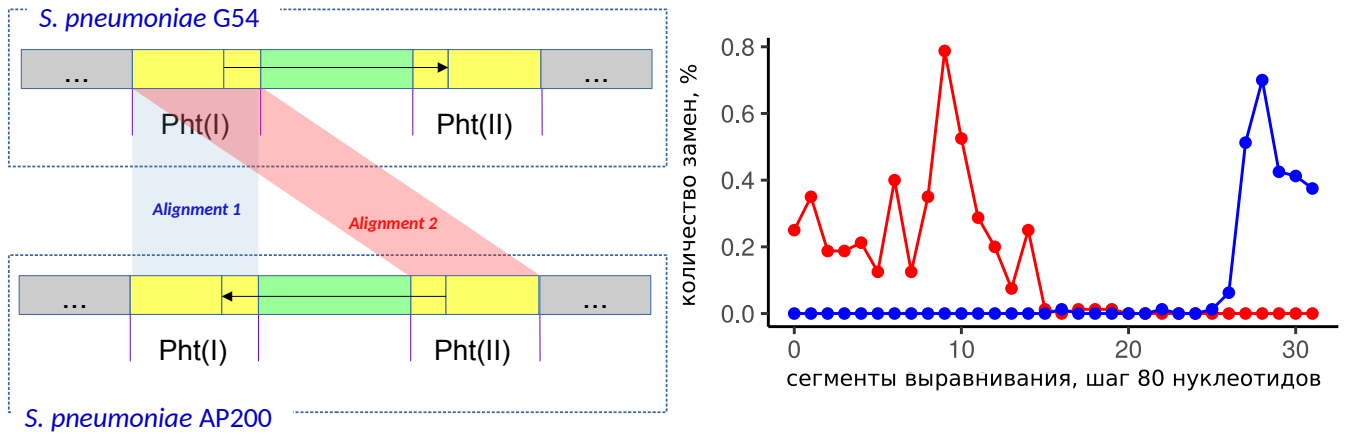


Рис. 4. Поиск точки рекомбинации в брекпоинтах параллельной инверсии в стрептококках. Слева: схема выравнивания паралогов на границах инверсий. Справа: количество несовпадений в выравниваниях белков гистидиновой триады в штаммах *S. pneumoniae* ST556 и *S. pneumoniae* TCH8431/19A. Точки показывают процент несоответствий на участках при выравнивании PhtB из *S. pneumoniae* ST556 и PhtB из *S. pneumoniae* TCH8431/19A (красные) и PhtD из *S. pneumoniae* TCH8431/19A (синие).

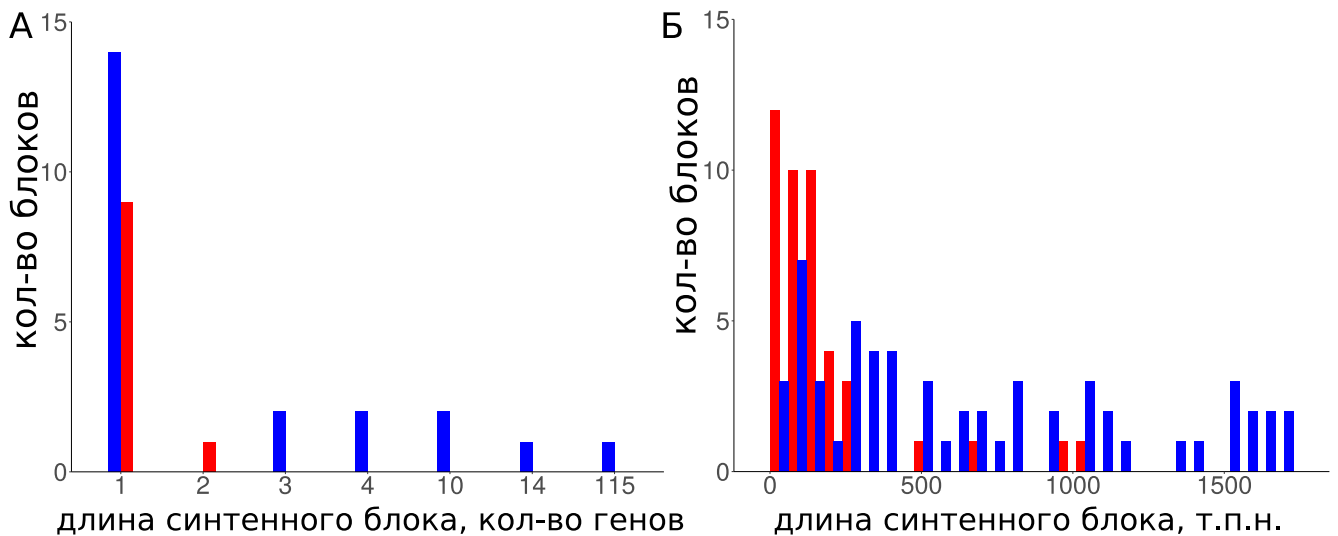


Рис. 5. Распределение длин синтенных блоков. А) транслоцированных между хромосомами в ходе эволюции *Burkholderia* spp.; Б) инвертированных в *Burkholderia mallei*. Красный цвет соответствует блокам, гены в которых поменяли свое положение относительно лидирующей/запаздывающей цепи, синий — блокам, в которых гены сохранили свое положение относительно цепи.

COMaX.

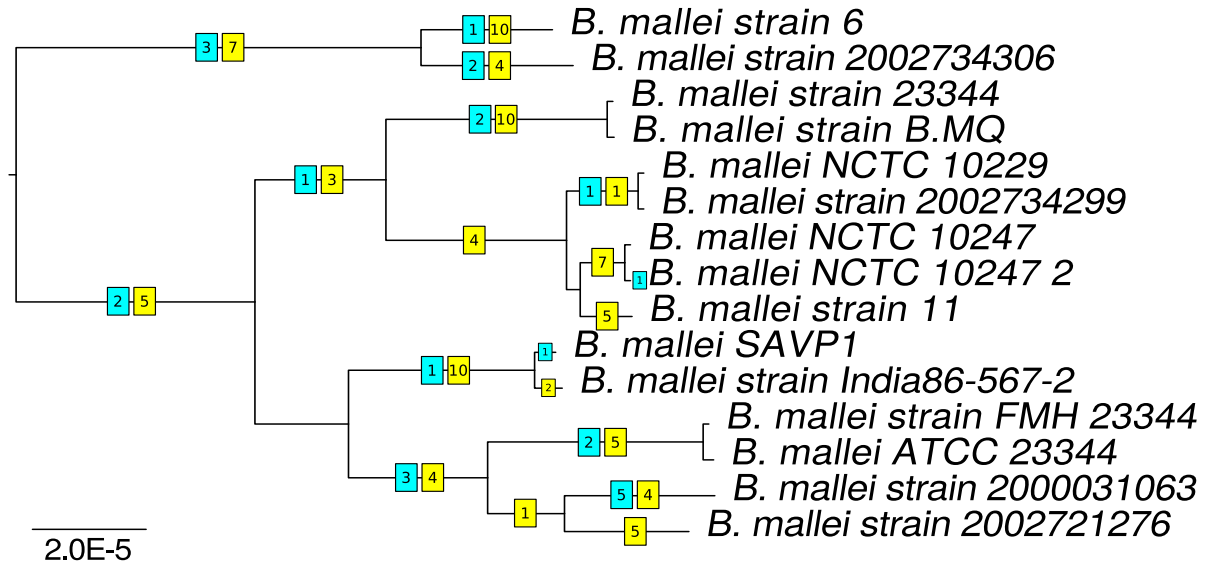


Рис. 6. Филогенетическое дерево *B. mallei*, построенное по конкатенату выравниваний одно-копийных универсальных генов. Числа на ветках показывают количество реконструированных на них инверсий: желтые соответствуют инверсиям на первых хромосомах, синие — на вторых.

Инверсии, проходящие через точку начала репликации, перепредставлены и на первых ($p < 10^{-33}$), и на вторых ($p < 10^{-30}$) хромосомах. Длины инверсий, проходящих через точку репликации, имеют широкое распределение, достигая практически длины реплихоры (Рис. 5Б), в то время как длины инверсий, лежащих внутри реплихоры, не превышают 15% от ее длины. Только две из наблюдаемых инверсий превысили этот порог, однако поскольку обе произошли на ветке штамма *B. mallei* FMH23344 и частично перекрываются, можно предположить, что эти два события описывают одну транслокацию небольшого участка. В геномах *B. mallei* была обнаружена корреляция между частотой инверсий в геномах и частотой замен в общих генах (тест Спирмана, $\rho = 0.8$, $p = 10^{-7}$).

Для 51 штамма *B. pseudomallei* было реконструировано три инверсии на первых и пять на вторых хромосомах (Рис. 7). Инверсии на вторых хромосомах длиной около 1.3 м.п.н. имеют одинаковые границы в семи штаммах, расположенных на удаленных ветках филогенетического дерева. Границы этих

инверсий проходят по повтору, состоящему из шести генов, кодирующих (1,2) субъединицы А и В рамнозилтрансферазы первого типа; (3) транспортер устойчивости к антибиотикам, относящийся к *mrB/QacA* подсемейству; (4) рамнозилтрансферазу второго типа и (5,6) компоненты RND системы эффлюкса, липопротеины внешней мембраны *nodT* and *emrA*. Поскольку мембранные белки, лежащие на границах этих инверсий, распознаются иммунной системой хозяина, эти инверсии также могут быть результатом действия фазовой вариации. Однако, в данном случае из-за точного совпадения последовательностей белков в повторах выравнивание границ не позволяет установить границу рекомбинации.

В штаммах *B. serasia* анализ расположения синтенных блоков позволил выявить два геномных островка, предположительно распространяющихся горизонтально и закрепляющихся на разных хромосомах в штаммах группы. Первый фрагмент длиной около 8.5 т.п.н. найден только в двух штаммах и расположен на первой хромосоме в геноме *B. seposerasia* и на второй хромосоме в геноме *B. serasia* ATCC 25416. Фрагмент содержит пять генов, продукты которых участвуют в метаболизме железа, и регулятор транскрипции семейства *AgaC*. Другой блок длиной около 5.5 т.п.н. был найден в 17 из 30 штаммов группы и также имеет мозаичный паттерн расположения на хромосомах. Кассета состоит из четырех генов, кодирующих ацетил-КоА карбоксилазный комплекс, гена, кодирующего гликозид гидролазу (углеводный метаболизм), и гена, кодирующего регулятор транскрипции семейства *LysR*.

В пятой главе описана разработанная автором модель для оценки горизонтально потока генов в популяцию и представлены результаты применения этой модели к нескольким наборам бактериальных данных.

Для каждого ортологического ряда, не содержащего паралогов и имеющего не менее трех представителей в каждом изучаемом виде, выполнялся анализ двух характеристик: консервативность расположения в геноме относительно стабильного окружения и согласованность дерева гена с деревом видов. Модель

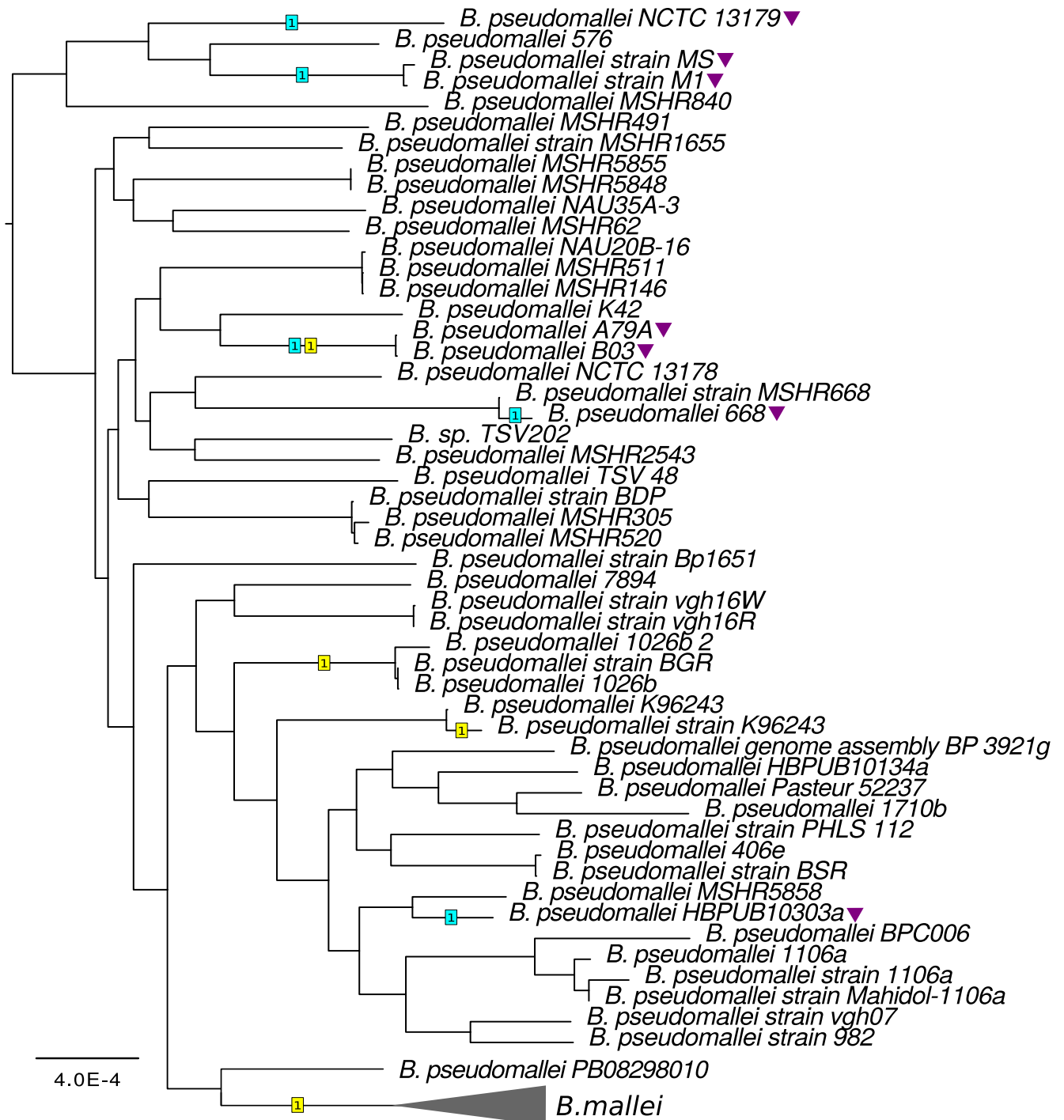


Рис. 7. Филогенетическое дерево *B. pseudomallei*, построенное по конкатенату выравниваний однокопийных универсальных генов. Числа на ветках показывают количество реконструированных на них инверсий: желтые соответствуют инверсиям на первых хромосомах, синие — на вторых.

основана на следующих наблюдениях. Для универсальных однокопийных генов наиболее вероятным эволюционным сценарием является вертикальное наследо-

вание, поскольку как сами эти гены, так и их порядок часто сохраняется во всех изучаемых геномах. Тогда и для периферийных генов, найденных внутри этих стабильных связок в геномах, наиболее вероятным эволюционным сценарием является вертикальное наследование и далее потери в некоторых штаммах. В свою очередь, расположение периферийных генов внутри различных стабильных связок общих генов может являться индикатором горизонтального переноса.

Традиционным методом идентификации горизонтального переноса в истории гена является анализ его филогенетического дерева, поскольку в этом случае положение штаммов, между которыми произошел перенос, на дереве гена отличается от их положения на дереве, построенном по объединению выравниваний общих генов. Поскольку построение точной филогении штаммов одного вида остается трудноразрешимой задачей, анализировались периферийные гены, присутствующие более чем в двух видах рода и более чем в трех штаммах в каждом. Для таких генов топология дерева проверялась следующим образом. В случае, если удастся найти такое положение корня, при котором образуются монофилетические ветки, соответствующие разделению видов, будем говорить, что топология дерева гена согласована с деревом видов, в противном случае — не согласована.

Для 103 штаммов *Streptococcus* spp. было протестировано 2858 периферийных ортологичных групп (Таблица 1). 48 из них имеют филогенетические деревья с мозаичной структурой и их представители найдены в различных стабильных связках общих генов, что можно рассматривать как свидетельство в пользу горизонтального переноса этих генов. Более половины из этих генов оказались редкими для всех трех видов *Streptococcus*, то есть представители рядов были найдены менее чем в трети штаммов вида; остальные были редки хотя бы для одного из изучаемых видов (Рис. 8).

Для 127 штаммов *Chlamydia* spp. было протестировано 1046 периферийных ортологичных групп; представители только 12 из них были найдены в различ-

Таблица 1. Характеристики периферийных генов в стрептококках

Дерево	Расположение		
	консервативное	неконсервативное	всего
согласованное	1181	128	1309
несогласованное	44	48	92
всего	1225	176	1401

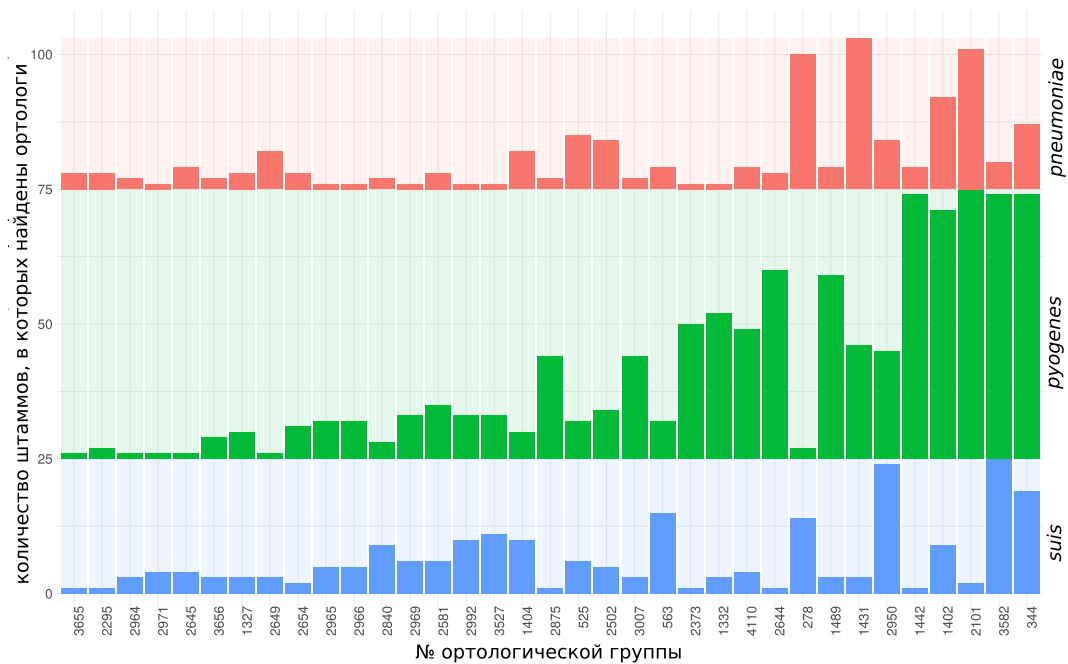


Рис. 8. Частота встречаемости представителей ортологических рядов, найденных в разных геномах в различных синтенных областях и имеющих деревья с мозаичной структурой, в *Streptococcus* spp. Каждый столбец соответствует одному ортологическому ряду. Цветные сегменты столбца показывают долю штаммов каждого вида, в которых найдены представители ряда.

ных стабильных связках общих генов. Восемь генов, относящихся к биосинтезу и метаболизму триптофана, формируют триптофановый оперон, который расположен в зоне пластичности в штаммах *C.aviae* и *C.felis*, но лежит внутри стабильной синтении в штаммах *C.pescorum* (Рис. 9А). Кроме того, три гена из этого оперона были найдены в зоне пластичности *C.trachomatis*, в остальных штаммах гены этого оперона отсутствуют (Рис. 9Б).

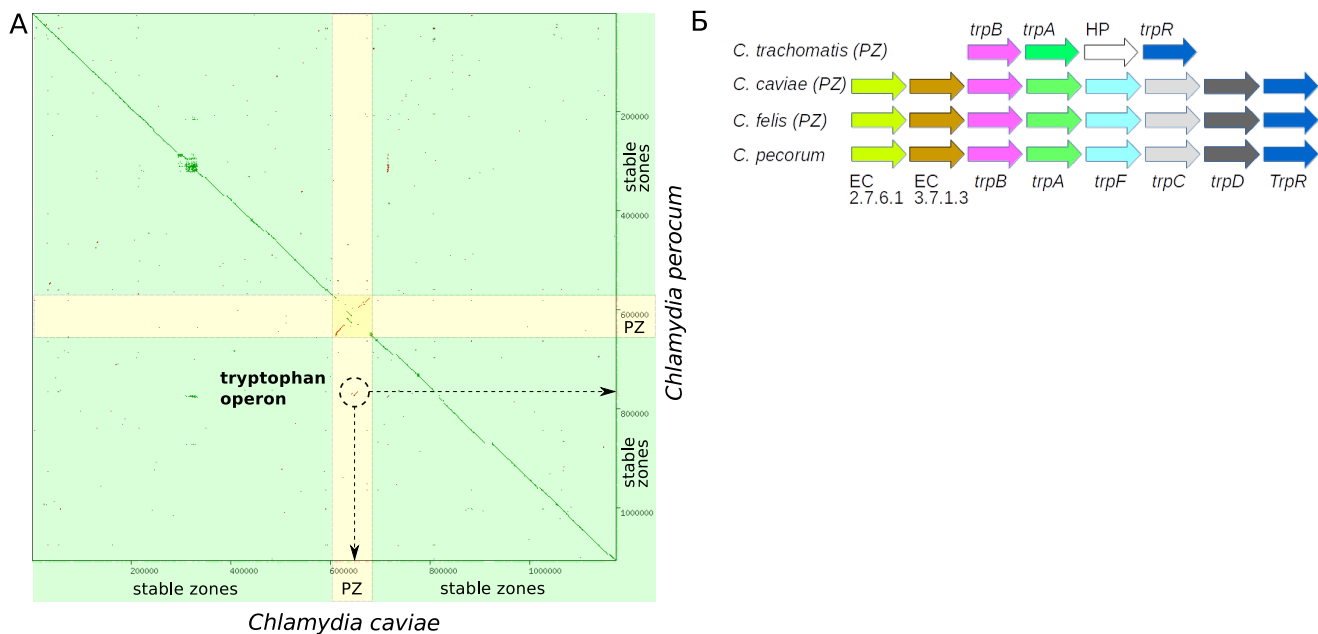


Рис. 9. (А) Положение триптофанового оперона в геномах *C. rescouri* и *C. saviae*. (Б) Схема триптофанового оперона и колокализованных с ним генов в штаммах разных видов хламидий.

Применение модели к нескольким наборам бактериальных данных показало, что разные виды имеют разные частоты потоков генов. В частности, штаммы *Streptococcus* характеризуются открытым пангеномом с большой периферией, каждый новый штамм добавляет в пангеном десятки новых генов. По результатам предложенной модели, более 10% периферийных генов распространяются в популяции горизонтально, что может объясняться высокой пластичностью геномов стрептококков за счет системы компетенции. В свою очередь, считается, что большинство новых генов в геномах хламидий возникает за счет паралогаизации и функционального расхождения генов, принадлежащим всего нескольким крупным семействам, что может быть вызвано главным образом изолированным внутриклеточным стилем жизни. Анализ периферийных генов хламидий выявил всего несколько случаев горизонтального переноса, что согласуется с этой гипотезой и отражает отличия хламидий от большинства бактериальных родов, для которых основной движущей силой геномной эволюции является горизонтальный перенос.

Выводы

1. Филогенетические деревья, построенные по расположению синтенных блоков, позволяют уточнить эволюционную историю штаммов и разрешить узлы в кладах молодых штаммов при изучении эволюции бактериальных видов с высокой частотой перестроек.
2. Инверсии, лежащие внутри одной реплихоры, значимо короче инверсий, концы которых лежат на разных реплихорах, а также такие инверсии встречаются реже, чем это ожидается из случайной модели.
3. При межхромосомных транслокациях действует отбор на сохранение положения генов в транслоцированных сегментах относительно лидирующей / запаздывающей цепи.
4. В штаммах мульти-хромосомных бактерий *Burkholderia* выявлены новые геномные островки; в патогенах человека *Streptococcus pneumoniae* и *Burkholderia pseudomallei* предсказаны новые случаи антигенной вариации за счет инверсий. В *Streptococcus pneumoniae* установлено расположение точки рекомбинации внутри генов, кодирующих антигены PhtB и PhtD.
5. Сравнение двух характеристик периферийных генов: согласованности топологии дерева гена с деревом видов и положении гена в геномах относительно стабильных локусов — позволяет выявить гены, распространяющиеся в популяции горизонтально.
6. В геномах хламидий выявлено всего несколько случаев горизонтального переноса, в то время как в стрептококках более 10% периферийных генов были перенесены в популяцию горизонтально. Эти различия коррелируют со структурой пангенома родов и их приспособленностью к разным экологическим нишам.

Список публикаций по теме диссертации. По материалам диссертации опубликованы статьи в рецензируемых научных журналах:

1. Genome rearrangements and phylogeny reconstruction in *Yersinia pestis* / O. Bochkareva, N. Dranenko, E. Ocheredko, G. Kanevsky, Y. Lozinsky, V. Khalaycheva, I. Artamonova, M. Gelfand // PeerJ. 2018. Т. 6. e4545
2. Genome rearrangements and selection in multi-chromosome bacteria *Burkholderia* spp. / O. Bochkareva, E. Moroz, I. Davydov, M. Gelfand // BMC Genomics. 2018. Т. 19, № 1. С. 965

Публикации в сборниках конференций и препринты:

1. Bochkareva O.O., Shelyakin P.V., Gelfand M.S. Evolution of bacterial chromosomes. Материалы международной конференции “Chromosome’18” // Издательство Новосибирского национального исследовательского государственного университета (Новосибирск). 2018. С. 10-11
2. Bochkareva O.O., Gelfand M.S. Genome rearrangements in bacterial genomes // Труды международной конференции “Bioinformatics of Genome Regulation and Structure/Systems Biology (BGRS SB’18)”, Новосибирск. 2018. С. 30
3. Bochkareva O.O., Shelyakin P.V., Gelfand M.S. Genome rearrangements in bacteria. // Proceedings of the conference “Bioinformatics: from Algorithms to Applications” (BiATA), Санкт-Петербург. 2018. С. 69.
4. Bochkareva O.O., Moroz E.V., Davydov I.I. Evolution of *Burkholderia* spp. // Proceedings of the 8th Moscow Conference of Computational Molecular Biology (MCCMB), Москва. 2017. Постер 119.
5. Bochkareva O.O., Moroz E.V., Davydov I.I. Evolution of *Burkholderia* spp. // Proceedings of the 23th European Meeting of PhD Students in Evolutionary Biology (EMPSEB), Краков. 2017.

6. Бочкарева О.О., Мороз Е.В. Эволюция геномов *Burkholderia* spp. // Труды 21ой Международной пушинской школы-конференции молодых ученых “Биология наука 21 века”, Пущино. 2017. С. 58.
7. Бочкарева О.О. Восстановление эволюционной истории геномных перестроек в бактериях. // Труды III Пушинской школы-конференции “Биохимия, физиология и биосферная роль микроорганизмов”, Пущино. 2016.
8. Moroz (Lopatina) E.V., Bochkaryova O.O., Kazanov M.D., Kalinina A.S. Insights into evolution history of *Burkholderia* spp. // Труды конференции молодых ученых и специалистов ИППИ РАН (ИТиС), Сочи. 2015. С. 651-655.
9. Bochkareva O.O., Kazanov M.D. Evolutionary history of rearrangements in *Burkholderia* spp. // Proceedings of the 7th Moscow Conference on Computational Molecular Biology (MCCMB), Москва. 2015. Постер 155.
10. Bochkareva O.O., Kazanov M.D., Gordienko E. Evolutionary history of recombination events in *E.coli*, *Shigella* and *Salmonella* genomes. // Proceedings of the 6th Moscow Conference on Computational Molecular Biology (MCCMB), Москва. 2013. Постер 210.
11. Бочкарева О.О. Анализ порядка генов в геномах бактерий и реконструкция геномных перестроек // Труды конференции молодых ученых и специалистов ИППИ РАН (ИТиС), Петрозаводск. 2012. С. 280-281.
12. *Chlamydia* pan-genomic analysis reveals balance between host adaptation and selective pressure to genome reduction. / O. Sigalova, A. Chaplin, O. Bochkareva, P. Shelyakin, V. Filaretov, E. Akkuratov, V. Burskaya, M. Gelfand // BioRxiv. 2018. С. 506121. DOI: [10.1101/506121](https://doi.org/10.1101/506121). Preprint
13. Micro-evolution of three *Streptococcus* species: selection, antigenic variation, and horizontal gene inflow / P. Shelyakin, O. Bochkareva, A. Karan, M.

Научное издание

Бочкарёва Ольга Олеговна

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук на тему:

Реконструкция эволюционной истории геномных перестроек в бактериях