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

На правах рукописи

Драненко Наталия Олеговна

Эволюция семейств бактериальных белков, участвующих во взаимодействии патогена с хозяином

1.5.8 – математическая биология, биоинформатика

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель:
доктор биологических наук, профессор
Михаил Сергеевич Гельфанд

Москва, 2025

Оглавление

Введение

Актуальность темы исследования

Белки играют центральную роль в биологических процессах. Белки выполняют разнообразные функции в живых организмах, включая каталитическую, структурную, транспортную, сигнальную и регуляторную. Понимание структуры и функции белков помогает раскрыть механизмы, лежащие в основе жизнедеятельности клеток и тканей.

Бактерии являются крайне изменчивыми организмами и мутации в генах, кодирующих белки, ведут к изменению также и в последовательностях самих белков. При этом точечные мутации могут повлиять на приспособленность бактерии как положительно, так и отрицательно, а также практически не оказывать влияния. В присутствии отбора такие изменения приводят к адаптации колонии к новой окружающей среде и изменениям в ней. Отслеживание такого рода изменений и разделение их на способствующие возникновению или поддержанию той или иной функции и нейтральные позволяет раскрыть механизмы, лежащие в основе жизнедеятельности бактерий. Однако эта задача до сих пор не решена полностью в силу колоссального разнообразия бактериальных сообществ и способов их адаптации к различным экологическим нишам.

Определить влияние тех или иных изменений на функцию белка возможно либо экспериментально, путём целенаправленного внесения изменений в последовательность кодирующего исследуемый белок гена, либо биоинформатически, на основе сравнения большого количества последовательностей белков одного семейства, то есть белков, произошедших от общего предка, и сходных по функции, последовательности и структуре [1]. Сравнение последовательностей белков одного семейства позволяет определить, какие именно функциональные изменения происходили в исследуемых белках и распространять информацию из ограниченного набора экспериментальных данных на более широкие группы объектов.

Хотя белки сами по себе являются носителями биологической функции, не все эти изменения происходят только в последовательности белка. Не менее важными являются также изменения, происходящие на уровне регуляции соответствующих генов. В одном геноме могут быть гены, кодирующие белки с очень близкой или даже одинаковой функцией. Зачастую дополнительная копия просто увеличивает предел экспрессии конкретного гена, однако во многих случаях такая избыточность ведет к разделению белков, используемых бактерией в разных условиях, и, соответственно, строгой регуляции соответствующих генов в этих условиях [2].

Накопление последовательностей полностью собранных геномов бактерий и экспериментальных данных о структурах и специфичности белков в последние годы позволило проводить комплексный разносторонний анализ белковых семейств в контексте эволюционной истории видов и штаммов для того, чтобы точно описать функции этих семейств и подсемейств и выявить основные закономерности их развития и влияния на жизнедеятельность соответствующих бактерий.

Степень разработанности темы

Несмотря на значительный прогресс в качестве аннотации по гомологии последних лет, функция многих белков остается неизвестной. По состоянию на март 2020 года около 32% белков в базе Pfam содержали в описании ключевые слова «неизвестная функция» [3]. Особенно ярко выражена эта проблема у прокариот, где неизвестна может оказаться функция около половины генов в геноме [4]. Взаимодействия бактерий с окружающей средой крайне разнообразны, поэтому бактерии имеют массу приспособлений к различным условиям, многие из которых скрываются за генами с неизвестной функцией: метаболической, транспортной, сигнальной и другими.

Использование геномного контекста и структурных данных позволяет сгруппировать белки с неизвестной функцией в семейства и во многих

случаях предсказать их функциональность [5]. Однако, для такого типа анализа требуется значительное количество данных. И хотя количество последовательностей в публичных базах данных, в том числе данных полногеномного секвенирования, неуклонно растёт, количество последовательностей, доступных для разных видов, возрастает неоднородно. Например, для кишечной палочки в базе данных Genbank по состоянию на август 2024 года доступен 4731 полный геном и 281 544 геномов разного уровня сборки, в то время как для рода *Wolbachia* доступна 1661 сборка, из которых 294 полные, а для большинства бактериальных родов доступен и вовсе только один геном. Таким образом, хотя основные подходы к изучению белковых семейств разработаны, остаётся широкое поле для их применения.

Такого рода недостаточность данных наблюдается в том числе в семействах, где отдельные представители описаны и, возможно, даже имеют разрешённые пространственные структуры, однако ключевые моменты их участия в метаболизме остаются неясными. Например, в случае семейства IpaH из *Shigella* разрешена структура нескольких представителей [6] и даже известны молекулярные функции отдельных представителей семейства [7], однако подробно не исследовалась структура семейства, устойчивость геномного контекста и распространённость соответствующих генов в разных хозяевах. При этом этот белок является мишенью для вакцин против шигеллёза [8]. Ещё одним показательным примером семейства белков, являющихся мишенью для вакцин, является семейство белков гистидиновых триад у *Streptococcus*, которые подробно изучались у *S. pneumoniae* [9], однако они представлены также у многих других видов *Streptococcus*. Важным моментом здесь является связь между представленностью этих факторов патогенности *Streptococcus* и степенью агрессивности соответствующего вида, а также охват генов белков этого семейства фазовой вариацией [10].

Особый интерес представляют также семейства, представленные во

множестве копий у широко распространённых видов, такие как ANK-белки у *Wolbachia*. Хотя было показано, что эти белки могут быть секретирующимися эффекторами для взаимодействия с организмом хозяина, были исследованы в основном взаимодействия с конкретными хозяевами на малом числе доступных штаммов [11–13]. Кроме того, важным объектом для исследования остаются семейства белков, представленные в большом разнообразии видов и родов, такие как CorA. Несмотря на то, что белки этого семейства играют основную роль в транспорте магния у прокариот и эукариотических митохондрий, вопрос о строгости их субстратной специфичности и роли фрагмента белка, который считается селективным фильтром, остаётся открытым [14].

Цели и задачи

Целью работы был комплексный анализ эволюционной истории нескольких белковых семейств, играющих важную роль во взаимодействии между бактерией и хозяином, как характерных для отдельных родов бактерий, так и широко представленных в разных бактериях.

Для достижения поставленной цели были поставлены следующие задачи:

1. Установить состав семейства IpaH и разнообразие в геномах эффекторов, отвечающих за взаимодействие с иммунным ответом хозяина у кишечных патогенов человека *Shigella*.
2. Исследовать разнообразие семейства поверхностных антигенов Htp и их изменчивость под действием различных генетических механизмов у *Streptococcus*.
3. Определить специфичность транспортёров металлов семейства CorA и описать их эволюцию у всех доступных бактерий.
4. Проанализировать связь между белками полигенного семейства ANK-белков и мобильными элементами в контексте их влияния на структуру генома *Wolbachia*.

Методология и методы исследования

В силу высокого разнообразия способов адаптации бактерий к экологическим нишам, для изучения каждого семейства белков была разработана методология, учитывающая уникальные особенности семейства и поставленного вопроса. В ней использовался широкий спектр современных методов биоинформатики и сравнительной геномики, включая методы множественного выравнивания последовательностей и анализа их консервативности, построения филогенетических деревьев, методы анализа функциональных сайтов в аминокислотных и нуклеотидных последовательностях, выравнивания белковых и РНКовых структур. Для предсказания регуляторных элементов были использованы методы на основе построения позиционно-весовых матриц. Все эти методы применялись к наиболее широким доступным на время проведения исследования наборам данных. Для компьютерной обработки данных и визуализации результатов использовались языки программирования Python и R. Для оценки статистической значимости полученных результатов использовались соответствующие задаче методы статистики.

Научная новизна

В работе был получен ряд новых результатов, касающихся особенностей набора бактериальных белковых семейств, связанных с адаптацией патогенов к хозяевам и агрессивным концентрациям ионов металлов во внешней среде. В частности впервые исследована структура семейства эффекторов IpaH у бактерий родов *Escherichia* и *Shigella* из человека и животных в сочетании с особенностями регуляции соответствующих генов. Впервые исследовано влияние последовательности селективного фильтра, то есть части белка, расположенной в устье канала и обеспечивающей прохождение через канал строго определённых ионов, на специфичность белков семейства CorA на наборе данных из широкого круга

бактерий. Впервые описана структура семейства белков гистидиновых триад у *Streptococcus* разных видов и выявлены представители семейства с регуляцией, зависящей от концентрации ионов меди. Впервые изучено влияние генов, кодирующих ANK-белки, на структуру геномов *Wolbachia* и их связь с генами, кодирующими мобильные элементы.

Практическая значимость

Полученные в этой работе результаты имеют в первую очередь теоретическую ценность и расширяют фундаментальные знания о бактериальных белковых семействах. Тем не менее, в силу медицинской значимости большинства исследуемых организмов и роли изучаемых белков в патогенезе, эти результаты также имеют практические приложения в медицине.

Такого рода анализ важен для отбора потенциальных целей для вакцин, так как гены белков, обладающих высокой иммуногенностью, таких как белки гистидиновых триад стрептококковмогут подвергаться быстрым обратимым изменениям и, тем самым, уходить от действия иммунитета. Кроме того, важным фактором является распространённость той или иной вариации белка в популяции бактерий, так как эффективной мишенью для вакцины может являться не весь белок целиком, а его наиболее консервативная часть.

Важное медицинское значение имеет комплексный анализ семейства эффекторов IpaH у патогенов, способных заражать разные виды хозяев. Такой анализ позволяет предсказывать возможные горизонтальные переносы генов между бактериями от разных хозяев и прогнозировать возможные новые зоонозы.

Немаловажным является также и анализ обширного семейства белков, присутствующего почти повсеместно в геномах бактерий, CorA. Понимание особенностей работы транспортных каналов клетки может позволить рассматривать эти каналы в качестве мишеней новых классов антибиотиков

или способов их доставки в бактериальную клетку.

Основные положения, выносимые на защиту

1. Семейство эффекторов IpaH является характеристической особенностью бактерий *Shigella* и энтероинвазивных кишечных палочек и состоит из девяти классов эффекторов, имеющих общий С-концевой домен, отвечающий за убиквитин-лигазную активность, и различающихся эффекторным N-концевым доменом, распознающим белок-мишень. В одном из классов происходит расхождение паралогов на две группы, что может привести к формированию двух новых классов эффекторов. Белки этого семейства также были обнаружены у патогенов крыс, сурков и овец, причём у бактерий этих хозяев набор эффекторных доментов отличается от доменов патогенов человека.
2. Белки гистидиновых триад у *Streptococcus* крайне разнообразны, однако структура филогенетического дерева соответствует вертикальному наследованию генов этих белков, а не горизонтальным переносам. Большинство генов этих белков контролируются цинковыми репрессорами, однако присутствуют две группы без такой регуляции. Гены белки одной из этих групп контролируются медными репрессорами. Фазовой вариации подвергаются только гены белков гистидиновых триад из *S. pneumoniae*, но не других видов *Streptococcus*.
3. Роль характеристической последовательности GxN семейства белков CorA в определении специфичности транспортёра была ранее переоценена. Белки с отличными от канонической последовательностями в этом мотиве претерпевают частые горизонтальные переносы, располагаются на филогенетическом дереве в основном на одной ветви и способны к транспорту тех же катионов. Если последовательность GxN и оказывает влияние на специфичность, то слабое.

4. Между числом копий генов ANK-белков и размером генома *Wolbachia* присутствует значимая положительная корреляция. Среди соседей этих генов мобильные элементы присутствуют статистически чаще, чем в среднем по геному. Мобильные элементы могут являться драйверами эволюции этих генов.

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 результаты, представленные в настоящей работе, были получены автором лично за исключением данных о координатах локально коллинеарных блоков в геномах *Wolbachia*.

Структура и объем работы

Работа изложена на 117 страницах. Она состоит из одиннадцати разделов: введение, главы 1-5, заключение, выводы, благодарности, список литературы и приложение. В главе 1 представлен обзор литературы по теме работы. В главах 2-5 представлены описания собственных исследований. Работа содержит 21 рисунок и две таблицы. Список литературы содержит 193 наименования. Приложение содержит три рисунка и девять таблиц.

Апробация работы и публикации по теме

По материалам работы опубликованы три статьи в международных рецензируемых журналах. Результаты работы были представлены на Московской международной конференции по вычислительной молекулярной биологии (Moscow Conference on Computational Molecular Biology – MCCMB’21), на конференции «Информационные технологии и системы» (ИТиС'23, Огниково), на шестом Международном симпозиуме по системной биологии микробных инфекций (6th International Symposium on Systems Biology of Microbial Infections, 2021, онлайн).

Глава 1. Обзор литературы

1.1 Эволюция белка как последовательности аминокислот

Ранние исследования, посвященные эволюции белков, привели к возникновению двух теорий, которые легли в основу молекулярной биологии и сравнительной геномики. Ещё в 1958 году Криком было описано возникновение в будущем «таксономии белков», то есть изучения эволюции видов путем сопоставления аминокислотных последовательностей белков из этих видов [15]. А уже в 1965 году появилась теория молекулярных часов, выдвинутая Полингом и Цукеркандлем. Согласно этой теории, эволюционно значимые замены в биологических последовательностях происходят с практически постоянной скоростью [16,17]. Это позволило датировать события прошлого, которые не оставили доступных для анализа окаменелостей, но оставили биологические следы. Основополагающим для молекулярной биологии стало также наблюдение Кимуры, что скорость молекулярной эволюции слишком высока, чтобы такое можно было объяснить только положительным отбором. Сочетание этого наблюдения с другими фактами привело к созданию теории нейтральной эволюции: большая часть эволюционных событий происходит на молекулярном уровне, а большая часть различий между и внутри видов обусловлена дрейфом генов мутантных аллелей, которые являются нейтральными или почти нейтральными, а не положительными, в терминах отбора [18].

Изменения последовательности белок-кодирующего гена могут быть связаны с положительным отбором, нейтральным дрейфом или недостаточно сильным очищающим отбором. В случае дрейфа накапливаются нейтральные мутации, не приводящие к изменению структуры и функции белка. Слабый очищающий отбор может привести к закреплению мутаций, незначительно ухудшающих свойства организма; это характерно для видов с малым эффективным размером популяции, в частности, у прокариот – для облигатных эндосимбионтов [19]. В случае положительного отбора

происходит адаптация, то есть изменение свойств белка, включая приобретение новых биохимических активностей [20]. Существуют значительные различия в скорости эволюции разных белок-кодирующих генов. Одним из главных определяющих факторов является строгость структурных или функциональных ограничений. Белки, на которые действует строгий отрицательный отбор, накапливают значительно меньше мутаций, чем остальные белки. В качестве количественной меры силы отбора используется отношение частоты несинонимичных мутаций к частоте синонимичных в генах, кодирующих исследуемые белки. Различные эволюционные модели, позволяющие оценивать такие соотношения, в применении к бактериям показали, что необходимые для выживания гены бактерий эволюционируют медленнее прочих, а гены домашнего хозяйства подвергаются более строгому очищающему отбору, чем гены, кодирующие секретируемые белки [21].

В случае несинонимичных мутаций происходит изменение аминокислоты в белке, что может либо сохранять исходную структуру, либо нарушать сворачивание настолько, что образование некоторой стабильной и функциональной структуры становится невозможным, либо, реже, приводить к формированию новой структуры [22]. В большинстве случаев точечные аминокислотные замены сохраняют структуру и функцию белка: около 50-80% аминокислот могут быть изменены и при этом структура не претерпит значительных изменений [23,24]. Для современных программ предсказания структуры по последовательности методом моделирования по гомологии достаточно около 30% аминокислотного сходства [25]. Однако также было показано, что в некоторых случаях точечной мутации, приводящей к изменению одной аминокислоты, достаточно для формирования значительно отличающейся стабильной структуры [26,27].

Хотя, как правило, замена одной аминокислоты сохраняет структуру и функцию белка, это часто снижает стабильность этой структуры. В случае

таких замен в генах, кодирующих жизненно важные белки, могут возникать так называемые компенсаторные мутации, работающие в противовес вредной мутации [28]. Компенсаторные мутации могут либо возникать в других генах организма, уменьшая потребность в белке с нарушенной функцией, либо восстанавливать нарушенную молекулярную функцию [29]. Таким образом, эволюция белковой последовательности происходит под действием противонаправленных сил, одна из которых препятствует накоплению изменений, а другая ведёт к их накоплению.

1.2 Белковые семейства

Семейство белков это такая группа белков, которая происходит от общего предка, что отражается в сходстве последовательности, структуры и функции [1]. Традиционный подход к описанию белковых семейств является иерархическим, хотя и эта иерархия весьма условна и зависит от рассматриваемой задачи. Наивысшим элементом этой классификации является суперсемейство, группа, включающая потомков одного максимально удалённого общего предка, причём в такой группе функции и доменная структура белков могут заметно различаться. В рамках же семейства различия в функции и доменной структуре уже отсутствуют, однако может различаться специфичность. С другой стороны, семейства белков могут быть разделены на подсемейства, в рамках подсемейства уровень сходства последовательностей и близость функции выше, чем в рамках семейства, и специфичности уже не различаются [1,30].

Ещё одним важным аспектом в установлении связей уровня семейства для белков является сходство структур. Известно, что структура белка изменяется гораздо медленнее, чем последовательность [31]. Однако очевидное сходство структур белков при отсутствии видимого сходства последовательностей может проистекать из двух принципиально разных ситуаций: либо структурное сходство следует из общего происхождения белков, однако в процессе эволюции последовательности разошлись очень

далеко, либо структурное сходство возникло вследствие конвергентной эволюции, то есть белки не имеют общего происхождения [1]. При этом в обоих случаях функция может быть либо одинаковой на некотором уровне специфичности, либо различаться, однако с большей вероятностью общая функция будет наблюдаться у гомологов.

Биологически формирование белковых семейств вызвано вертикальным наследованием соответствующих генов и дупликациями генов или их фрагментов. Гены и белки, разошедшиеся в ходе эволюции у потомков одного предка в процессе видообразования, называются ортологами, а гены или белки, появившиеся в результате дупликации, называются паралогами. В случае ортологов функция белка как правило сохраняется. В случае дупликации обе копии сохраняются, если их функции будут различаться [1].

1.3 Расхождение паралогов с формированием новой функциональной специфичности

Дупликация гена является одним из возможных эволюционных событий, в результате которого в геноме образуются две тождественные копии какого-то гена. Изначально предполагалось, что в таком случае одна из копий освобождается от давления очищающего отбора и начинает активно изменяться [32]. В большинстве случаев такие изменения приводят к полной потери функциональности и вторая копия псевдогенизируется. Однако возможен исход, когда эта копия закрепляется эволюцией как полезное приобретение [32]. Впоследствии было показано, что этот путь реализуется, хотя и редко. Более вероятным же вариантом является ситуация, когда на короткое время ослабевает давление отбора на обе копии, в результате чего обе они быстро эволюционируют. Несмотря на такое ускорение, было показано, что средняя скорость эволюции для паралогов ниже, чем для генов, представленных в единственной копии. Возможным объяснением такого поведения является большая вероятность возникновения или закрепления паралогов у генов домашнего хозяйства, которые эволюционируют медленнее

других генов [32].

Полученные в результате дупликации копии в ходе эволюции могут следовать разными путями. Возможны варианты гипофункционализации, субфункционализации, неофункционализации, баланса дозы, компенсаторного дрейфа и нейтральной вариации [33]. В случае гипофункционализации обе копии гена сохраняют свою функцию, однако уровень их экспрессии снижается, таким образом для нормального функционирования организма требуются обе копии, и обе они сохраняются. В случае, если исходный ген нес более чем одну функцию, возможна субфункционализация, то есть распределение этих функций по копиям. Такие копии сохраняются, чтобы сохранился полный набор необходимых функций. В случае неофункционализации одна из копий сохраняет свою исходную функцию, а другая приобретает некоторую новую, что повышает адаптированность организма. Ещё один вариант сохранения нескольких копий гена с сохранением функции связан с участием гена в многокомпонентных взаимодействиях, чувствительных к дозе гена, как, например, транскрипционные факторы. Также при компенсаторном дрейфе под воздействием эффекта баланса дозы может повышаться вероятность того, что вторая копия получит новую функцию, а не псевдогенизируется. По мере уменьшения экспрессии одной из копий гена, сила отбора, направленного на сохранение первоначальной функции снижается. В какой-то момент мутации, псевдогенизирующие этот ген, все ещё сильно вредны, однако мутации, приводящие к неофункционализации, полезны, что способствует формированию у второй копии новой функции. Также объектом отбора может являться не функция как таковая, а регуляторные особенности паралогов. В таком случае гены могут иметь разные уровни экспрессии в разных условиях среды, разных тканях и стадиях развития, и т.п. [33]. При этом с течением времени и биохимическая специфичность паралогичных белков начинает расходиться.

Важную роль в формировании новой функциональной специфичности может играть изначальная малая специфичность белка к субстрату: некоторые белки способны участвовать в более чем одной реакции и, таким образом, иметь сразу несколько функций [34]. Новая функциональная специфичность часто появляется при отборе вариантов белков, у которых возникли мутации, ослабляющие функциональную специфичность [35]. Если в результате мутации возникает возможность участвовать в реакции, в которой исходный белок участвовать не мог, и это участие даёт преимущество в приспособленности, то такая мутация закрепится и будет поддержана последующими дополнительными мутациями [35]. Белки, способные участвовать в более чем одной реакции, легче претерпевают такого рода изменения. Важно отметить, что возможность неспецифичных взаимодействий зачастую сопровождается некоторым снижением приспособленности. Эволюционный путь в таком случае должен проходить на ландшафте приспособленности таким образом, чтобы в каждый момент времени потеря стабильности или других свойств была не слишком неблагоприятна [35].

Детали процесса расхождения паралогов чрезвычайно трудно наблюдать. Многие паралоги разошлись эволюционно довольно далеко друг от друга. Например, в случае модельного организма *E. coli* около 80% паралогов идентичны менее чем на 50% [36]. В случае большого количества замен, учитывая возможную нейтральность замен и эпистатическое взаимодействие, чрезвычайно трудно определить, какие именно замены повлияли на функцию [37]. Зачастую сравнение большого количества доступных последовательностей при доступной структуре белка позволяет определить важные замены в активном сайте, однако возможные функциональные замены вдалеке от этой области трудно определить [37].

1.4 Регуляция экспрессии генов в связи со специфичностью кодируемого белка

Бактериальная клетка использует обширный набор регуляторных механизмов для успешного существования при постоянно изменяющихся условиях среды. Бактерии способны контролировать каждую ступень экспрессии генов, начиная от инициации транскрипции и заканчивая посттрансляционными взаимодействиями. Экспрессия генов может регулироваться в соответствии с температурой, концентрацией ионов металлов, присутствием антибиотиков, доступностью тех или иных питательных веществ и другими факторами среды, а также зависеть от внутренних потребностей клетки [38]. Зачастую контроль экспрессии генов является многоступенчатым, то есть один регулятор контролирует один или несколько других, образуя регуляторные сети [38].

Активация и подавление экспрессии у бактерий, как правило, реализуются действием регуляторных белков. Такие белки способны напрямую связывать ДНК по сайтам, содержащим конкретный мотив, и контролировать инициацию транскрипции. Бактериальные регуляторные белки можно разделить на четыре типа по сочетаниям положительного и отрицательного контроля у репрессоров и активаторов [38]. В случае отрицательного контроля индуцируемых генов репрессор связывается с операторной областью гена и блокирует транскрипцию (Рисунок 1а). При наличии кофактора, репрессор связывает кофактор, что препятствует связыванию репрессора с ДНК, и транскрипция происходит (Рисунок 1б). В случае отрицательного контроля подавляемых генов репрессор не способен связывать ДНК без кофактора, поэтому в отсутствие кофактора транскрипция происходит (Рисунок 1в), а в присутствии кофактора транскрипция подавляется (Рисунок 1г). В случае положительного контроля индуцируемых генов активатор может связывать ДНК только в комплексе с кофактором, поэтому без кофактора транскрипция не происходит (Рисунок 1д, е). В случае

положительного контроля подавляемых генов кофактор связывается с активатором и препятствует связыванию активатора с ДНК, что подавляет транскрипцию (Рисунок 1ж, з) [38].

Регуляция экспрессии на уровне транскрипции в бактериальной клетке возможна не только с помощью регуляторных белков, но также и регуляторных РНК. Малые регуляторные РНК могут связывать комплиментарные участки мРНК, что ведет к деградации этой мРНК [39]. Ещё одним способом регуляции экспрессии являются рибопереключатели. Рибопереключатель это элемент нетранслируемой области мРНК обычно на 5’ конце, способный специфично связывать маленькие молекулы. Этот элемент обеспечивает способность мРНК образовывать две стабильные структуры, в присутствии и отсутствии лиганда. На уровне транскрипции это конформационное изменение влияет на то, будет ли продолжаться синтез соответствующей мРНК. На уровне трансляции этот механизм обеспечивает доступность мРНК для связывания рибосомой и инициации трансляции [38].

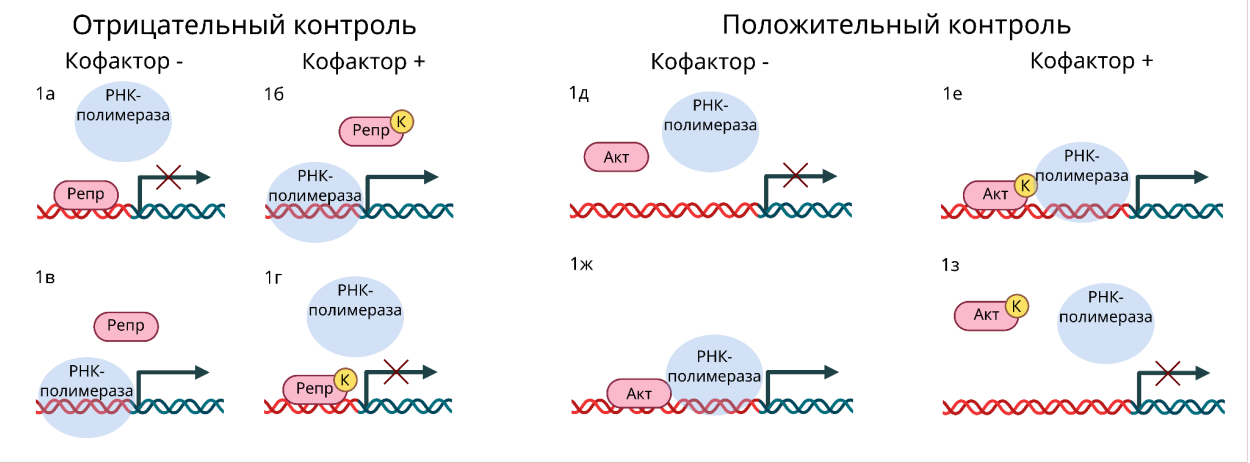


Рисунок 1. Возможные стратегии контроля транскрипции генов репрессорами (а-г) и активаторами (д-з) в зависимости от наличия кофактора.

1.5 Фазовая вариация

Фазовая вариация определяется как быстрое и обратимое изменение экспрессии, которое может происходить либо по принципу включения/выключения, либо переключения между несколькими аллельными

вариантами [40]. Гены, которые подвергаются фазовой вариации, как правило кодируют в бактериях элементы, которые располагаются на поверхности клетки. Это позволяет иметь в бактериальной популяции группы с разным фенотипом и быстро реагировать на изменяющиеся условия среды [40]. Наличие таких генов позволяет бактериальной популяции поддерживать разнообразие вариантов фенотипа, которые могут быть полезны при разных условиях. Например, один из вариантов может способствовать более эффективной начальной колонизации хозяина, однако быть крайне иммуногенным. В таком случае преимущество получают те бактерии, у которых этот ген активен только на начальной стадии инфекции, а впоследствии работает вариант, который позволяет более эффективно избегать иммунного ответа хозяина [40].

Существует несколько генетических механизмов, обеспечивающих возможность фазовой вариации. Одним из таких механизмов является изменение длины участков простых повторов. На таких участках может происходить проскальзывание полимеразы во время репликации, поэтому возможно быстрое изменение длины такого участка. Если такой участок расположен внутри открытой рамки считывания (ОРС), то с гена может считываться полноразмерный белок, его фрагмент или совсем никакого продукта [40]. Участки с простыми повторами могут также располагаться в промоторной области гена. В таком случае существует градиент экспрессии гена, так как участки с простыми повторами между сайтами –10 и –35 влияют на связывание сигма-фактора РНК-полимеразы, а расположенные дальше от старта гена, чем сайт –35, влияют на связывание сопутствующих факторов (как правило, активаторов) и их взаимодействие с РНК-полимеразой [41].

Ещё одним возможным механизмом фазовой вариации является инверсия фрагмента ДНК. Рекомбинация между инвертированными повторами может происходить между парой гомологичных локусов, причем один из них может быть даже неэкспрессирующимся. За счёт такой инверсии

происходит перестановка фрагментов гомологичных генов, в результате чего расширяется разнообразие белков в популяции. Инверсии может также подвергаться промоторная область, что приводит к возможности попеременного включения и выключения экспрессии [40]. Кроме того, возможно эпигенетическое влияние на экспрессию, когда метилирование участка промотора приводит к регуляции экспрессии через эпигенетические механизмы [40].

В силу быстрых и обратимых изменений, продукты генов, подвергающихся фазовой вариации, обычно не рассматриваются как потенциальные мишени для разработки вакцин. Хотя этот подход в целом рационален, в некоторых ситуациях он не оправдан. Если гены, подверженные фазовой вариации, особенным образом полезны патогену в некоторой конкретной фазе заболевания, то вакцинация может помочь против этой фазы заболевания, возможно ключевой [40].

1.6 Методы исследования эволюции белковых семейств

В большинстве подходов к исследованиям возможно наблюдать состояние белка в конкретный момент времени, однако эволюция это процесс, поэтому целью эволюционных исследований является отслеживание состояний системы в течение продолжительного периода и определение закономерностей её развития во времени. Так как в большинстве случаев эволюционный процесс невозможно изучать в контролируемых лабораторных условиях, значительным становится выбор методов анализа имеющихся данных для реконструкции эволюционной истории [42].

Одним из способов наблюдать все доступное разнообразие белков является построение множественного выравнивания. Выравниванием называется расположение друг под другом последовательностей ДНК, РНК или белков таким образом, чтобы сходные участки располагались друг под другом, что позволяет увидеть сходства и различия таких последовательностей. Выравнивание аминокислотных последовательностей

белков несёт больше биологического смысла, чем выравнивание нуклеотидных последовательностей генов, так как именно белок является ключевой функциональной биологической молекулой и несёт функциональную и структурную информацию [43]. Именно выравнивание, как правило, является стартовой точкой в исследованиях белковых семейств. В то же время, анализ нуклеотидных последовательностей полезен при анализе сайтов связывания транскрипционных факторов и вторичных структур РНК.

Первый алгоритм попарного выравнивания использовался для сравнения нуклеотидных последовательностей и начислял положительный вес выравнивания в случае соответствия символов, отрицательный вес за несоответствие символов, а также предполагал использование штрафа за разрыв как дополнительное условие [44]. Современные алгоритмы построения таких выравниваний учитывают сходство или различие символов в каждой позиции, событие открытия разрыва, длину разрыва и кратность длины разрыва трем для нуклеотидных последовательностей генов [45]. В случае же аминокислотных последовательностей важным параметром также является используемая матрица замен аминокислот. Двумя наиболее часто используемыми матрицами замен являются матрица закрепившихся точечных мутаций (point accepted mutation, PAM) [46] и матрица замен блоков (BLOcks SUbstitution Matrix, BLOSUM) [47].

Существуют две формы выравнивания: локальное и глобальное. В случае локального выравнивания целью является поиск наиболее схожего участка внутри более длинной последовательности, а в случае глобального выравнивания происходит поиск сходства на всей длине обеих последовательностей [43]. Глобальное выравнивание часто используется для анализа сходства белков в рамках одного семейства, однако в случае перемешивания доменов или наличия только локального, но высокого сходства, глобальное выравнивание не показывает в полной мере этих

свойств последовательностей. В таких случаях или большой разнице в длине выравниваемых последовательностей используется именно локальное выравнивание [43].

Для построения множественного выравнивания наиболее популярными являются три подхода: точный, итеративный и прогрессивный [43]. Точные методы выравнивания опираются на динамическое программирование и пытаются оптимизировать целевую функцию для выравнивания трех или более последовательностей. Так как сложность вычислений есть показательная функция от количества последовательностей, предпочтительными становятся эвристические методы [43].

При прогрессивном подходе ко множественному выравниванию последовательности собираются последовательно, начиная с самой похожей пары. Для установления порядка добавления последовательностей к выравниванию используется направляющее дерево. Самым распространенным методом, использующим такой подход, является ClustalW [45]. Этот метод строит попарные глобальные выравнивания всех последовательностей, после чего на основе этих выравниваний строит матрицу попарных расстояний и затем строит направляющее дерево. По направляющему дереву строится прогрессивное выравнивание, начиная с самой сходной пары и с постепенным добавлением следующих последовательностей [45]. Основным недостатком такого рода методов является неизменность ранних выравниваний, используемых как базовые для дальнейших построений. При таком подходе ошибки ранних этапов наследуются на всех последующих и могут накапливаться [43].

Итеративные методы на некоторых шагах используют прогрессивные методы, но также выполняют постобработку полученных этими методами результатов. Такие методы изменяют структуру направляющего дерева. На каждом шаге они перерасчитывают матрицу расстояний на основе полученного прогрессивными методами множественного выравнивания, по

новой матрице строится новое направляющее дерево. На основе нового направляющего дерева строится новое множественное выравнивание и процедура повторяется итеративно. В некоторых итеративных методах весь набор данных многократно разделяется на две подгруппы, каждая из которые перевыравнивается, пока процесс не сойдется к некоторому итоговому результату [43].

Кроме выравниваний, важным элементом работы с белковыми семействами является филогенетический анализ. Филогенетический анализ позволяет установить эволюционные отношения между организмами или другими биологическими единицами. Филогенетический анализ белкового семейства предполагает построение некоторой модели эволюции, отражающей реальные эволюционные отношения в семействе. Графическим представлением такого рода отношений является филогенетическое дерево [48].

Методы построения филогенетических деревьев могут быть разделены на две крупные группы: зависящие и не зависящие от выравнивания. Наиболее распространёнными и применимыми на практике на данных момент являются методы, зависящие от выравнивания. Их можно разделить на методы, основанные на оценке расстояний между последовательностями, и методы, сравнивающие все последовательности в одной позиции [49]. Эволюционные расстояния во всех возможных парах белковых последовательностей могут быть вычислены путём применения эволюционной модели к исследуемому набору данных для объяснения замен, наблюдаемых в выравнивании [50]. Это расстояние отражает оценку среднего числа изменений на участке длины последовательности с момента их расхождения от общего предка и используется группой филогенетических методов, называемых дистанционными методами построения деревьев [49]. К методам, основанным на оценке расстояний между последовательностями относятся метод невзвешенной группировки с арифметическим средним

(Unweighted Pair Group Method with Arithmetic Mean, UPGMA) и метод присоединения соседей (Neighbor joining, NJ) [49]. В методе UPGMA строится матрица попарных расстояний между анализируемыми последовательностями, после чего выбираются две максимально похожие последовательности. Эти последовательности объединяются в узел, расположенный на равном расстоянии до исходных последовательностей, так как метод предполагает для всех последовательностей равные темпы эволюции. Этот узел заменяет две исходные последовательности при расчёте новой матрицы расстояний. Такая процедура повторяется, пока дерево не будет полностью разрешено [51]. Метод присоединения соседей использует аналогичную процедуру, однако предполагает возможность разных темпов эволюции для разных последовательностей и использует другой способ перерасчета матрицы расстояний. Принцип этого метода заключается в поиске пар соседей, которые минимизируют общую длину ветви на каждом этапе кластеризации, начиная со звездообразного дерева [52]. В настоящее время методы, основанные на кластеризации, особенно NJ, уже редко используются самостоятельно, однако часто применяются для построения предварительного дерева, которое впоследствии дорабатывается другими методами.

К методам, основанным на сравнении всех последовательностей в одной позиции, относятся филогенетические методы максимального правдоподобия и максимальной экономии. В случае метода максимального правдоподобия производится оценка вероятности того, что некоторое дерево с заданным набором параметров в предположении конкретной эволюционной модели произведет рассматриваемый набор последовательностей на листьях и наиболее. Соответственно, цель таких методов найти дерево, для которого эта вероятность максимальна [50]. В случае большого исходного набора данных вычисление всех возможных деревьев становится крайне ресурсоёмкой задачей, поэтому практические реализации этого подхода

используют эвристики, позволяющие производить расчёт в разумное время [50]. Метод максимальной экономии минимизирует число мутаций для получения современных последовательностей при развитии согласно дереву. Поскольку этот метод минимизирует число событий, он не может учитывать параллельные события, потому более других страдает от притяжения длинных ветвей [53]. Поскольку метод максимальной экономии не учитывает вероятность происхождения события, а только их число, он даёт оценку снизу необходимого числа эволюционных изменений, но не отражает наиболее вероятного эволюционного пути [54].

Помимо отслеживания эволюции белка как последовательности, важно учитывать также структурную изменчивость белка, так как биологическую функцию белок несёт именно как физический объект. Структура белка, как свойство определяющее специфичность, как правило, гораздо более консервативна, чем последовательность, поэтому сопоставление структур очень важно при анализе далёких белков одного семейства, когда сходство последовательностей уже практически на случайном уровне [31]. Кроме того, долгое время считалось, что соответствие между структурой белка и его последовательностью взаимно однозначное, однако в редких случаях были обнаружены различные структуры, соответствующие одной и той же последовательности [55]. Таким образом, при анализе белков одного семейства, важно обращать внимание также на консервативность структуры, так как иная структура, как правило, ассоциирована с другой функцией. Отсюда возникает потребность в сравнении структур и формулировании некоторого численного выражения такого сходства. Идеальный критерий сходства структур должен отражать природу свёртывания белка, быть устойчивым к мелким ошибкам и неточностям эксперимента, основываться на небольшом количестве понятных параметров и иметь доступную интерпретацию [55]. В силу сложности задачи такой универсальный критерий пока не был сформулирован, однако были разработаны некоторые

практически применимые варианты [55]. Большая часть этих методов была разработана для анализа сходства экспериментально полученной структуры и теоретической модели, однако они применимы и в случае сравнения экспериментальных или теоретических структур между собой.

Методы сравнения структур также можно разделить на зависящие и не зависящие от последовательности. Методы, зависящие от последовательности, предполагают, что задано соответствие между аминокислотными остатками целевого и сравниваемого белка, а независящие от последовательности методы выполняют геометрическое сравнение структур с последующим построением выравнивания последовательностей на основе структурного наложения. Применимость не зависящих от последовательности методов ограничена случаями, когда структуры достаточно схожи, однако возникает несоответствие последовательности из-за сдвигов витков альфа-спирали или различий в числе структурных повторов [55]. Тем не менее, за исключением случаев очень далёких гомологов, зависящие и независящие от выравнивания методы показывают одинаковые результаты. Оба типа методов могут оценивать как глобальное, так и локальное сходство. Неупорядоченные фрагменты белка или многодоменная структура могут влиять на уровень глобального сходства структур, однако такие структуры могут демонстрировать высокое локальное сходство. Возможна также и обратная ситуация, когда глобально схожие белки локально демонстрируют довольно низкое сходство [55].

Большинство методов сравнения структур основываются на измерении расстояний между некоторыми опорными точками. Такого рода расстояния в свою очередь зависят от способа наложения. Наиболее популярным методом измерения расстояния и соответственно наложения структур является метод оптимизации общего среднеквадратичного отклонения. Главной проблемой такого метода является высокая чувствительность к большим отклонениям малых фрагментов [55]. Такого рода ошибкам не подвержен метод, при

котором веса наиболее отклоняющихся фрагментов итеративно уменьшаются при наложении, что позволяет определить ядро наложения максимального размера. У таких методов главным недостатком является произвольность выбора ядра. Таким образом, разные способы наложения оптимизируют различные параметры, давая различные итоговые оптимальные положения [55].

Не зависящие от наложения методы сравнения структур лишены типичных недостатков зависящих от наложения методов. К таким относится класс методов, определяющих сходство структур через контакты между остатками. В таком случае вместо наложения структур друг на друга сравниваются взаимодействия или расстояния, то есть положения фрагментов, в структурах. В зависимости от принятого определения контакта изменяется и итог сравнения. Например, изменение расстояния контакта в определении позволяет сделать меру локальной или глобальной, то есть определяет глубину сравнения [55].

Конкретные примеры описанных выше подходов и особенности изучаемых в рамках этой работы семейств будут рассмотрены в соответствующих главах.

Глава 2. Эволюция семейства эффекторов IpaH

2.1 Введение

Шигеллез — широко распространенное кишечное инфекционное заболевание человека. Шигеллёзная инфекция передается фекально-оральным путем и распространена по всему миру, как в развитых, так и развивающихся странах [7]. Его возбудитель, *Shigella*, является самой распространённой причиной бактериальной диареи по всему миру и, например, в 2016 году была в второй по значимости причиной смерти от диареи у людей всех возрастов, доля таких случаев среди всех смертей от диареи составила около 13% [56].

Бактерии *Shigella* были выделены Киёси Шигой из образцов стула больных дизентерией в 1898 году и названы *Bacillus dysenteriae*. Позднее они были переименованы в честь первооткрывателя в *Shigella dysenteriae*. Кроме того, Шига определил, что открытая им бактерия является грамотрицательной палочкой. Основываясь на симптомах, молекулярных особенностях и отчасти географии распространения инфекции, род *Shigella* был классифицирован на четыре вида [7]. *S. flexneri* и *S. sonnei* встречаются по всему миру и являются причиной наибольшего числа случаев заболевания шигеллёзом, *S. dysenteriae* вызывает эпидемии бациллярной дизентерии и считается самым вирулентным вариантом, а *S. boydii* самый редкий вид, ареал распространения которого преимущественно ограничивается Индийским субконтинентом.

Тем не менее, эти виды *Shigella*, и тем более сам род, не являются монофилетическими. Основываясь на генетических, а не фенотипических, особенностях, было установлено, что все они являются патогенными вариантами кишечной палочки, однако название рода до сих пор сохраняется из-за его медицинской важности [57,58]. Сначала секвенирование 16S рРНК, а потом и подробный филогенетический анализ позволили установить, что *Shigella* spp. возникали независимо из различных непатогенных *Escherichia*

coli путем приобретения большой плазмиды *pINV* и двух геномных островков SHI-1 и SHI-2 [59].

Плазмида *pINV* содержит локус *mix-spa*, кодирующий систему секреции третьего типа (CC3T) и значительное количество других генов инвазивности. Геномный островок SHI-1 несёт гены нескольких токсинов, а геномный островок SHI-2 кодирует аэробактин, и содержит гены, кодирующие антибиотики и белки уклонения от иммунитета хозяина [60]. Кроме того, в отличие от большинства вариантов кишечной палочки, все шигеллы — это неподвижные бактерии, так как у них произошла инактивация жгутиков. Также они не имеют генов фимбриальных адгезинов, используемых кишечными палочками для колонизации кишечника [60]. Известно, что у шигелл нарушены или полностью отсутствуют четыре важных гена, имеющихся у других кишечных палочек: протеазы внешней мембраны *ompT*, биосинтеза никотиновой кислоты *nadA/nadB*, лизиндекарбоксилазы *cadA* и *speG*, превращающего спермидин в нереактивный ацетилспермидин [61].

В процессе адаптации к внутриклеточному стилю жизни значительную роль играет накопление мобильных элементов и процесс псевдогенизации. Как правило, гены и другие функциональные элементы в бактериальном геноме плотно упакованы, а между приобретением и потерей генов поддерживается баланс. С одной стороны, бактерия получает внешнюю ДНК и приобретает признаки, позволяющие лучше адаптироваться к внешней среде, с другой стороны геном подвергается сокращению за счёт псевдогенизации некоторых генов или делеции частей генома [62]. При переходе от свободного к внутриклеточному стилю жизни этот баланс может быть нарушен. В некоторых случаях, например, у *Mycobacterium leprae*, это приводит к тому, что большое количество генов становится инактивировано [63], в других, как у *Mycoplasma genitalium*, происходит избавление от всех не необходимых для выживания генов и сокращение генома [64]. Кроме того, экспрессия некоторых генов при внутриклеточном стиле жизни может

снижать приспособленность. Уже упомянутый *cadA* у шигелл подавлял бы активность энтеротоксина, тем самым снижая вирулентность. Инактивация генов может происходить в том числе за счёт встраивания мобильных элементов. Таким образом накопление мобильных элементов и большое число псевдогенов является признаком относительно недавней адаптации шигелл к внутриклеточному стилю жизни [62].

Существуют и другие агрессивно патогенные штаммы кишечной палочки, из которых наиболее близкой к шигеллам является энтероинвазивная кишечная палочка (ЭИКП) [65]. До обнаружения первого штамма ЭИКП в 1944 обычные кишечные палочки и шигеллы легко разделялись, однако ЭИКП разделяют множество характеристик кишечной палочки, в то же время демонстрируя патогенное поведение, подобное шигеллам [65]. Как и *Shigella*, ЭИКП произошли от непатогенных кишечных палочек путём приобретения плазмиды *pINV* и геномных островков SHI-1 и SHI-2 на хромосоме. Поскольку ЭИКП сохраняют способность жить вне клеток-хозяев, а их геномы содержат значительно меньше мобильных элементов и псевдогенов, считается, что ЭИКП являются предшественниками линий *Shigella* [65].

Механизм развития инфекции также сходен для шигелл и ЭИКП, хотя инфицирующие дозы сильно отличаются, составляя 10^1–10^4 и 10^6–10^{10} клеток соответственно [66]. Процесс заражения происходит в несколько этапов. На начальном этапе бактерии проникают в организм из просвета кишки через М-клетки кишечного эпителия [67]. При проникновении через М-клетки происходит взъерошивание мембраны, проникновение бактерии в базолатеральный карман М-клетки, после чего бактерия поглощается макрофагами. Внутри макрофагов шигеллы быстро запускают апоптоз. При гибели макрофагов выделяются провоспалительные цитокины IL-1β и IL-18, медиаторы острой воспалительной реакции [7]. IL-1β запускает сильное воспаление кишечника, а IL-18 участвует в формировании

антибактериального ответа путём активации NK-клеток, способствуя выработке гамма интерферона и привлечению полиморфноядерных нейтрофильных лейкоцитов, что нарушает целостность слизистой оболочки и способствует дальнейшему проникновению шигелл [67].

Как только шигеллы высвобождаются из погибшего макрофага в подслизистую, они становятся способны проникнуть в эпителиальные клетки с внутренней стороны слоя эпителия и реплицироваться в цитоплазме. Бактерии в цитоплазме перемещаются путём направленной полимеризации актина, что позволяет им распространяться в соседние клетки эпителия, избегая внеклеточной иммунной защиты организма-хозяина [67]. Само по себе такое вторжение вызывает активную воспалительную реакцию, инициированную внутриклеточными системами. Nod1-опосредованная система распознаёт фрагменты бактериального пептидогликана, выделяемого шигеллами, и активирует ядерный фактор NF-κB, запускающий секрецию цитокина IL-8. IL-8 массово рекрутирует нейтрофилы, что дестабилизирует связи между соседними клетками и даёт возможность для дальнейшей инвазии. Таким образом, гибель макрофагов, разрушение эпителиального слоя и массовый приток нейтрофилов усугубляют бактериальную инфекцию и поражение тканей [67].

С помощью белков, кодируемых на *pINV* и в островках патогенности (ОП) на хромосоме, шигеллы взаимодействуют с клетками хозяина для развития воспаления. Большая часть генов, отвечающих за проникновение в клетку-хозяина расположена в области размером около 30 т.п.о. на *pINV*, этот участок называется входной областью [7]. Входная область кодирует CC3T в локусе *mxi-spa*, а также содержит гены *ipa* и *ipg*, продукты которых необходимы для инвазии в клетки кишечного эпителия. Белки CC3T выполняют ряд разнообразных функций, являясь структурными белками, шаперонами, которые защищают белки вирулентности шигелл и ЭИКП от агрегации и деградации, кроме того, через CC3T происходит экспорт

эффекторных белков, которые избирательно связывают определенные белки хозяина для регулирования биологической активности этих белков [68].

Одним из способов регулирования биологической активности является убиквитинилирование, которое используется многими патогенами для модуляции иммунного ответа хозяина на заражение. У *Shigella* убиквитинилирование производят так называемые *новые E3 убиквитин лигазы* (NEL, Novel E3 Ligases [6]), кодируемые генами *ipaH*, что приводит к деградации белков хозяина [69]. Белки IpaH состоят из двух доменов: высоко консервативного С-концевого домена новой E3 убиквитин лигазы (NEL), который связывает убиквитин, и N-концевого домена, содержащего вариабельный домен богатых лейцином повторов (БЛП), который связывает различные белки человека, тем самым обеспечивая специфичность к субстрату [70]. Считается, что белки IpaH запускают клеточную смерть и модулируют сигналы организма-хозяина, связанные с воспалением, во время бактериальной инфекции; однако субстратная специфичность многих белков IpaH остается неизвестной [6,71].

Экспрессия генов *ipaH* может регулироваться несколькими факторами транскрипции. MxiE, активатор транскрипции, кодируемый во «входной» области *pINV*, регулирует внутриклеточную экспрессию генов, кодирующих многочисленные факторы, секретируемые CC3T, в том числе, OspB, OspC1, OspE2, OspF, VirA и IpaH [72]. Известно, что два кодируемых плазмидой транскрипционных фактора вирулентности, VirF и VirB, включают вирулентность шигелл путем активации основных факторов вирулентности, что каскадно активирует экспрессию и генов *ipaH* [73]. Как у *virF*, так и у *virB* есть сайты связывания глобального сайленсера транскрипции H-NS, причём это связывание чувствительно к температуре. При 30°C оба гена *virF* и *virB* подавляются [73]. При проникновении шигелл в организм хозяина белки H-NS отделяются от ДНК, что запускает экспрессию каскадов вирулентности. H-NS обычно связывает AT-богатые последовательности,

образуя мостики или петли ДНК, которые влияют на транскрипцию с промоторов-мишеней [74,75]. Регуляторные области не только *virF* и *virB*, но и многих других генов вирулентности у шигелл содержат АТ-богатые последовательности, которые могут связываться с H-NS. Такие последовательности часто являются следами недавнего горизонтального переноса гена [76].

Хотя считалось, что шигеллы являются специфичными для приматов патогенами, эксперименты показали, что они могут заражать других животных, хотя и с меньшей эффективностью [77,78]. Недавно системы секреции, аналогичные CC3T шигелл, и связанные с ними эффекторы были обнаружены у *Escherichia marmotae*, потенциально инвазивного патогена сурков, который, как также было показано, способен проникать в клетки человека [79]. Маркерные гены шигелл были также обнаружены в изолятах, полученных из экскрементов телят с диареей, хотя данные по всему геному отсутствовали [80].

Целью этой работы было определение репертуара IpaH у *Shigella* spp. и энтероинвазивных *E. coli* и выявление различий у наборов доменов этих белков, отвечающих за распознавание и связь с белками-мишенями.

2.2 Материалы и методы

2.2.1 Геномные данные

В работе использовались 130 полных геномов *Shigella* spp., доступных в GenBank [81] по состоянию на ноябрь 2020 года, и три полных генома ЭИКП (Дополнительная таблица 1). Дополнительно были загружены все сборки *Escherichia* spp., хозяева которых были животными, содержавшие рамки считывания, продукты которых, согласно BLAST [82], имели сходство с NEL-доменом IpaH *Shigella* spp. (Дополнительная таблица 2).

2.2.2 Идентификация генов *ipaH*

С использованием pBLAST-поиска NEL-домена (PDB: Эффектор

Shigella flexneri IpaH1880 5KH1 https://www.rcsb.org/structure/5KH1) были обнаружены 445 белковых последовательностей, принадлежащих к семейству E3 убиквитин лигаз. Затем последовательности были сгруппированы с помощью CD-hit [83] с порогом в 90% на идентичность аминокислотных последовательностей, после чего был проведён дополнительный поиск tBLASTn для репрезентативных последовательностей из каждого кластера. Это позволило добавить 419 последовательностей, включая неаннотированные гены и псевдогены. В общей сложности обнаружены и классифицированы 864 последовательности *ipaH* (Дополнительная таблица 3). Гены *ipaH* у *Escherichia* spp. из животных хозяев были обнаружены с использованием той же схемы и представлены в дополнительной таблице 4.

2.2.3 Тепловые карты

Тепловые карты для сходства последовательностей были составлены с использованием R-пакетов seqinr, RColorBrewer и gplots.

2.2.4 Филогенетическое дерево

Для построения дерева видов *Shigella* spp. использовался инструмент PanACoTA [84]. Он аннотирует кодирующие области, находит ортологичные группы и строит филогенетическое дерево для конкатенированного выравнивания универсальных однокопийных генов. Ортологичные ряды были сконструированы с порогом 80% на идентичность в аминокислотных последовательностях, филогенетическое древо было построено с использованием модуля IQ-TREE 2 [85]. Дерево было визуализировано онлайн с помощью iTOL [86].

2.2.5 Аннотация регуляторных элементов

Выравнивания регуляторных областей *ipaH* генов были построены с помощью инструмента Pro-Coffee [87], дополнительные промоторы были предсказаны с помощью алгоритма platprom [88]. Сайты связывания VirF были предсказаны вручную на основе филогенетического футпринтинга

известных сайтов связывания. Последовательности нуклеотидов A и T классифицировались как политреки, если они содержали не менее шести одинаковых нуклеотидов A или T подряд.

2.2.6 Моделирование и визуализация белковых структур

Трехмерные структуры белков IpaH из *Escherichia marmotae* были смоделированы с использованием программы Swiss-Model [89] и структуры PDB: 5KH1.1 в качестве шаблона. В качестве инструмента визуализации использовалось программное обеспечение UCSF Chimera [90].

2.3 Результаты

2.3.1 Валидация геномных сборок

Было проанализировано 130 полных геномов *Shigella* spp., включая 46 *S. flexneri*, 25 *S. dysenteriae*, 19 *S. boydii*, 39 *S. sonnei* и один неклассифицированный штамм шигелл (Дополнительная таблица 1). Два критерия были введены для подтверждения аннотации *Shigella*: наличие генов *ipaH* и других компонентов CC3T (Таблица 1). В качестве маркеров T3SS использовались гены *mxiC*, *mxiE*, *mxiG*, *virB*, *virF*, *spa15*, *spa32*, *spa40*, *ipgA*, *ipgB*, *ipgD*, *apaA*, *ipaB*, *ipaC*, *ipad*, *mxiH*, *icsB*. В трех сборках не было обнаружено ни *ipaH*, ни генов T3SS. Эти образцы были взяты из почвы, речных отложений и с антарктического лишайника, поэтому было классифицированы как неинвазивная кишечная палочка и исключены из анализа. Кроме того, была произведена проверка, что неинвазивные штаммы *E. coli* не имеют ни одной из этих детерминант вирулентности, для чего был использован набор из 414 геномов *E. coli* и *Shigella* spp. из материалов [91]. В 17 сборках плазмиды отсутствовали, но присутствовали хромосомные гены *ipaH*. 37 сборок содержали плазмиды, но ни одна из них не содержала компонентов CC3T. Эти результаты могут быть объяснены элиминацией плазмид во время культивирования [92]. Только 64 сборки содержали все необходимые элементы вирулентности.

Кроме того, были охарактеризованы гены *ipaH* в трех доступных геномах ЭИКП из [60]. Один штамм (*E. coli* NCTC 9031) не содержал генов *ipaH* или CC3T, поэтому был исключён из анализа. Два других штамма (*E. coli* CFSAN029787 и *E. coli* 8-3-Ti3) имели плазмиду инвазивности с генами CC3T и гены *ipaH* на хромосомах и плазмидах (Дополнительная таблица 1).

Таблица 1. Статистика по геномным сборкам *Shigella* spp.

количество сборок	**наличие плазмид**	**наличие компонент CC3T**	**наличие *ipaH***	
			на хромосоме	**на плазмиде**
64	да	да	да	да
8	да	нет	да	да
1	да	да	да	нет
37	да	нет	да	нет
17	нет	нет	да	нет
2*	нет	нет	нет	нет
1*	да	нет	нет	нет

* Эти штаммы были классифицированы как неинвазивные *E. coli.*

2.3.2 Классификация генов *ipaH*

Не существует согласованной номенклатуры генов *ipaH* у разных видов рода *Shigella*, и количество таких генов в штаммах варьируется (см. таблицу 1 в [93]), поэтому была предложена объединяющая классификация всех генов семейства *ipaH*. В 127 сборках шигелл были обнаружены 864 гена, кодировавших белки из семейства убиквитин лигаз E3 (см. «Материалы и методы», Дополнительная таблица 3). На основании сходства последовательностей распознающих доменов и состава регуляторных элементов в регуляторных областях, все гены *ipaH* были разделены на девять классов (Рисунок 2). Эту классификацию подтверждает то, что белки из разных классов различаются по длине, количеству богатых лейцином

повторов (БЛП) и длине консервативных участков в регуляторной области (Таблица 2). С учетом высокого сходства ортологичных генов у всех видов *Shigella*, для аннотации использовались консенсусные последовательности *ipaH* (Дополнительная таблица 5) из каждого класса.

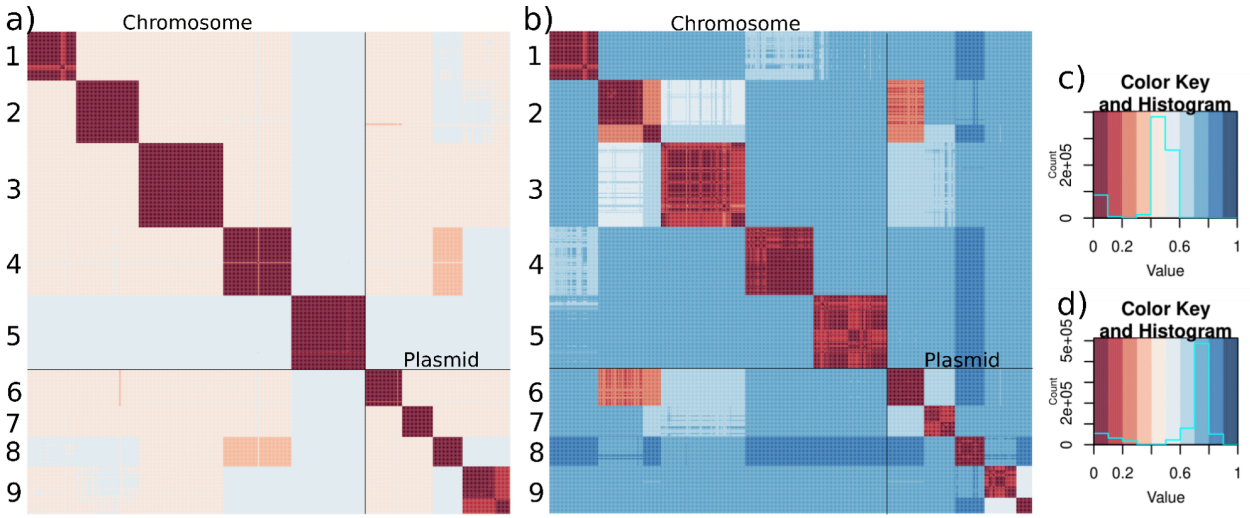


Рисунок 2. Тепловые карты попарных расстояний и соответствующие цветовые ключи для (a, c) генов *ipaH*; (b, d) 5'- последовательностей в *Shigella* spp. Попарные расстояния были рассчитаны как $\sqrt{1 - identity}$.

Интересно, что гены *ipaH* из классов 1-5 присутствовали только на хромосомах, в то время как гены из классов 6-9 были обнаружены только на плазмидах. Единственным исключением был дуплицированный ген *ipaH* из класса 5 в штамме 0228 *Shigella flexneri* 1a, где одна копия находилась на хромосоме, а другая на плазмиде. Поскольку эта сборка не содержала плазмиду инвазивности с генами CC3T, наблюдение могло быть вызвано неправильной сборкой. Белки, кодируемые генами из классов 4 и 8, были наиболее сходными внутри классов, в то время как регуляторные области были наиболее схожи для генов из классов 2 и 6.

Только 45% геномов шигелл содержат полный набор хромосомных генов *ipaH*, и 20% геномов содержат полный набор плазмидных генов *ipaH* (для плазмид это оценка по нижней границе, поскольку во многих сборках плазмидные последовательности отсутствуют). Более того, во многих геномах классы *ipaH* 3, 5 и 9 были представлены более чем одной копией.

Большинство копий *ipaH* были идентичны, исключение составляют два подкласса (9a и 9b), которые были различимы как по белок-кодирующим генам, так и по регуляторным областям (Рис. Ш1). Подкласс 9b был обнаружен почти во всех геномах *Shigella flexneri*, поэтому возникло предположение, что ген *ipaH* 9b был приобретен общим предком ветви *S. flexneri*.

Таблица 2. Классификация генов *ipaH Shigella*: кодирующие последовательности и регуляторные области.

	на хромосоме					**на плазмиде**			
класс ipaH	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9***
другие часто используемые названия *ipaH [93,94]*	*ipaH 1880*	*ipaH 1383*	*ipaH 2202*	*ipaH 0722*	*ipaH 2610*	*ipaH9.8*	*ipaH7.8*	*ipaH4.5*	*ipaH1.4*
	ipaHd	*ipaHc*	*ipaHe*	*ipaHa*	*ipaHb*	*ipaH9.8*	*ipaH7.8*	*ipaH4.5*	-
длина белка, а.к.	585	571	547	587	609	545	565	574	575
% полностью консервативных позиций в белке	95%	95%	99%	91%	96%	91%	99%	99%	94%
% полностью консервативных позиций в 5’-области гена	96%	99%	98%	93%	96%	94%	98%	98%	98%
количество БЛП	8	6	6	8	6	4	6	6	7
длина регуляторной области, п.о.	618: 343-928	339: 201-859	315: 170-731	580: 108-943	491: 489-881	393: 213-869	943: 524-951	428: 425-452	389:355 -1500 (335: 217-337)*

наличие сайта связывания MxiE	+	+	+	+	+	+	+	+	– (–)*

* Число в скобках — значения для паралогов.

2.3.3 Регуляторные паттерны генов *ipaH*

В дополнение к высокому уровню сходства белок-кодирующих последовательностей в каждом классе *ipaH*, регуляторные области генов также были высоко консервативными. Действительно, регуляторные области различных генов *ipaH* содержали 300-900 п.о. с идентичностью более 90% в каждом классе, за исключением класса 9 (см. ниже). Интересно, что сходство было высоким, начиная от стартового кодона трансляции до предполагаемых сайтов связывания фактора транскрипции MxiE (и включая их), особенно в классах 2 и 6, что указывает на ключевую роль MxiE в регуляции транскрипции *ipaH*. Ранее относительное расположение сайтов связывания MxiE и начала транскрипции, а также последовательности сайта связывания MxiE, бокса −10 и спейсера между ними использовались для классификации генов *ipaH* на восемь регуляторных классов [93]. Каждый класс, определенный с помощью предложенного подхода по подобию последовательностей, за исключением класса 9, соответствует одному из регуляторных классов (Рисунок 3B). Действительно, каждый класс имеет свой уникальный регуляторный паттерн, характеризующийся не только положением сайта связывания MxiE и последовательностью спейсеров, но и наличием участков, богатых A и T, в качестве возможных мишеней для взаимодействия с VirF и H-NS. В частности, классы 4, 5 и 7 обладают как A-, так и T-богатыми последовательностями (Рисунок 3A как пример), классы 1 и 2 имеют в основном полиT-последовательности, а класс 3 имеет в основном полиA-последовательности.

Плазмидные гены *ipaH* класса 9 были разделены на две группы. У генов из класса 9b разрушены регуляторные области из-за вставки профага и,

по-видимому, не имеют регуляторных элементов, типичных для других классов *ipaH* (Рисунок 3C). Кроме того, не удалось идентифицировать ни одного промотора-кандидата для *ipaH* класса 9b, что позволяет предположить, что эти гены могут не транскрибироваться. Гены класса 9a также не имели сайта связывания MxiE в регуляторной области, однако они могли бы транскрибироваться полицистронно с геном *ospE* (Рисунок 3C), используя его регуляторные элементы. Гены *ipaH* из класса 9a были окружены множеством треков, богатых A и T, типичных для мобильных элементов или профагов [76].

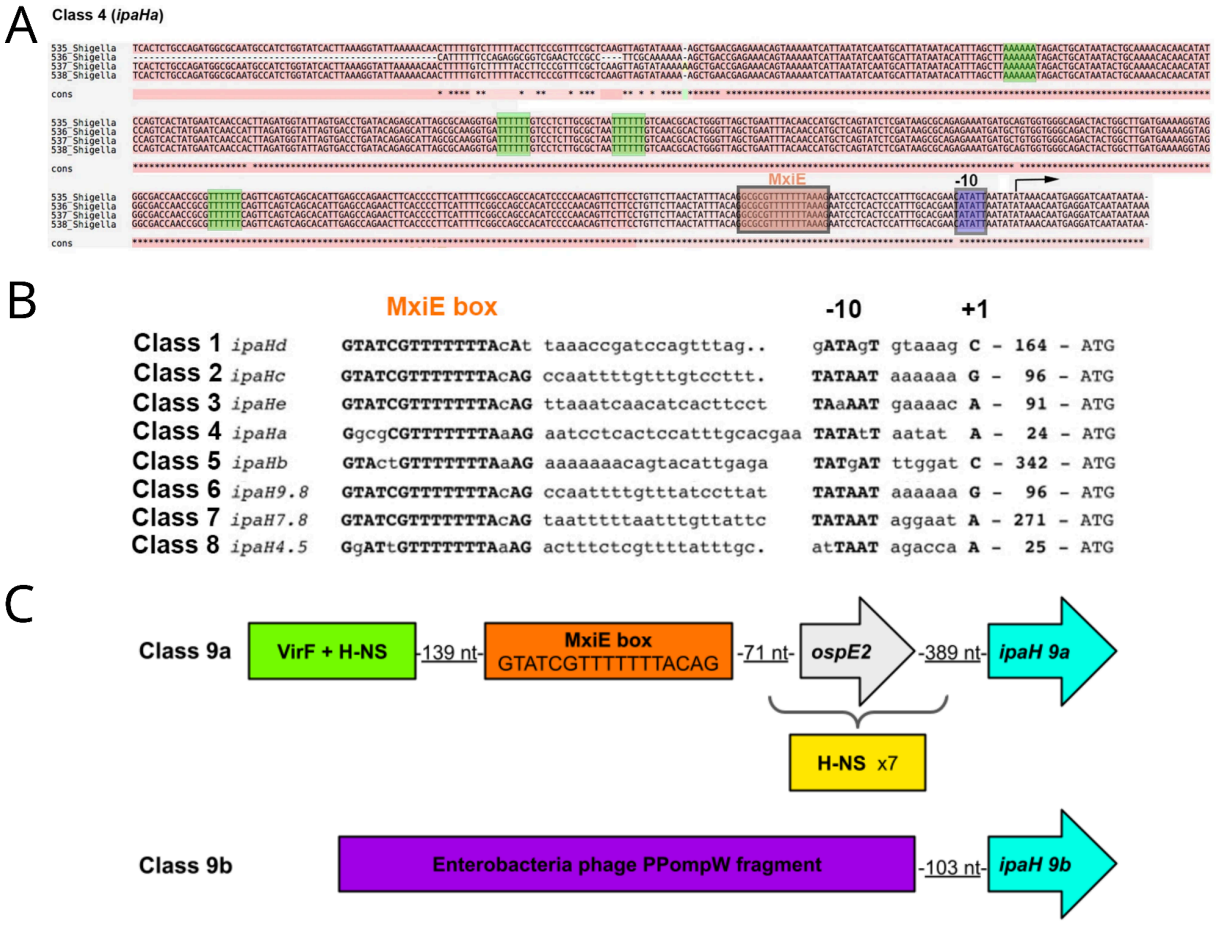


Рисунок 3. Регуляторные элементы, контролирующие гены *ipaH*. A) Выравнивание регуляторной области для отдельных представителей *ipaH* класса 4. Представители были отобраны на основе их последовательностей таким образом, чтобы были представлены все варианты последовательностей. Предполагаемый блок MxiE обозначен оранжевым прямоугольником, полиA/T участки — зелеными прямоугольниками. Начало транскрипции обозначено черной стрелкой. B) Сравнение классификации

ipaH, основанной на последовательности, с классификацией, основанной на расположении блока MxiE, элемента –10 и последовательности спейсера. Адаптировано из [93]. C) Схема регуляторных областей для классов *ipaH* 9a и 9b.

Регуляторные области различных классов *ipaH* не похожи, единственными исключениями являются классы 2 (хромосомный) и 6 (плазмидный), которые имеют очень похожий фрагмент регуляторной области длиной 150 п.о. между сайтом связывания MxiE и кодоном старта трансляции (Рисунок 3B).

2.3.4 Филетические паттерны генов *ipaH*

Были проанализированы филетические паттерны *ipaH* у штаммов шигелл и ЭИКП (Рисунок 4, Дополнительная таблица 3, Дополнительный рисунок 1). Реконструированное филогенетическое дерево в целом соответствовало предыдущим реконструкциям [60] и имело пять основных клад *Shigella* spp., причем названия видов не соответствуют топологии дерева. В этом наборе данных *S. sonnei* и *S. flexneri* были монофилетическими (отмечены желтым и фиолетовым на Рисунке 3 соответственно), *S. boydii* и *S. dysenteriae* были смешаны в двух удаленных кладах (отмечены оранжевым и красным на Рисунке 3 соответственно), а набор штаммов *S. dysenteriae* образовал пятую кладу (зеленая клада на Рисунке 3). Филетические паттерны генов *ipaH* были в высшей степени мозаичными. Тем не менее, наблюдались некоторые специфичные для клад закономерности. В частности, класс 1 был редок в оранжевой кладе, в то время как класс 3 отсутствовал в зеленой кладе.

Штаммы ЭИКП не группировались с основными кладами *Shigella* или друг с другом (Рисунок 4). Контекст генов и их геномное распределение также отражали полифилетическое происхождение штаммов ЭИКП. В частности, *E. coli* 8-3-Ti3 имела полный набор *ipaH*, в то время как у *E. coli* CFSAN029787 отсутствовали два хромосомных гена *ipaH*. Эти гены не

отличались от эффекторов шигелл, и их расположение на хромосомах и плазмидах соответствовало таковому у шигелл. У *E.* CFSAN029787 в *ipaH* 1 и *ipaH* 3 произошел сдвиг рамки, что, вероятно, привело к псевдогенизации.

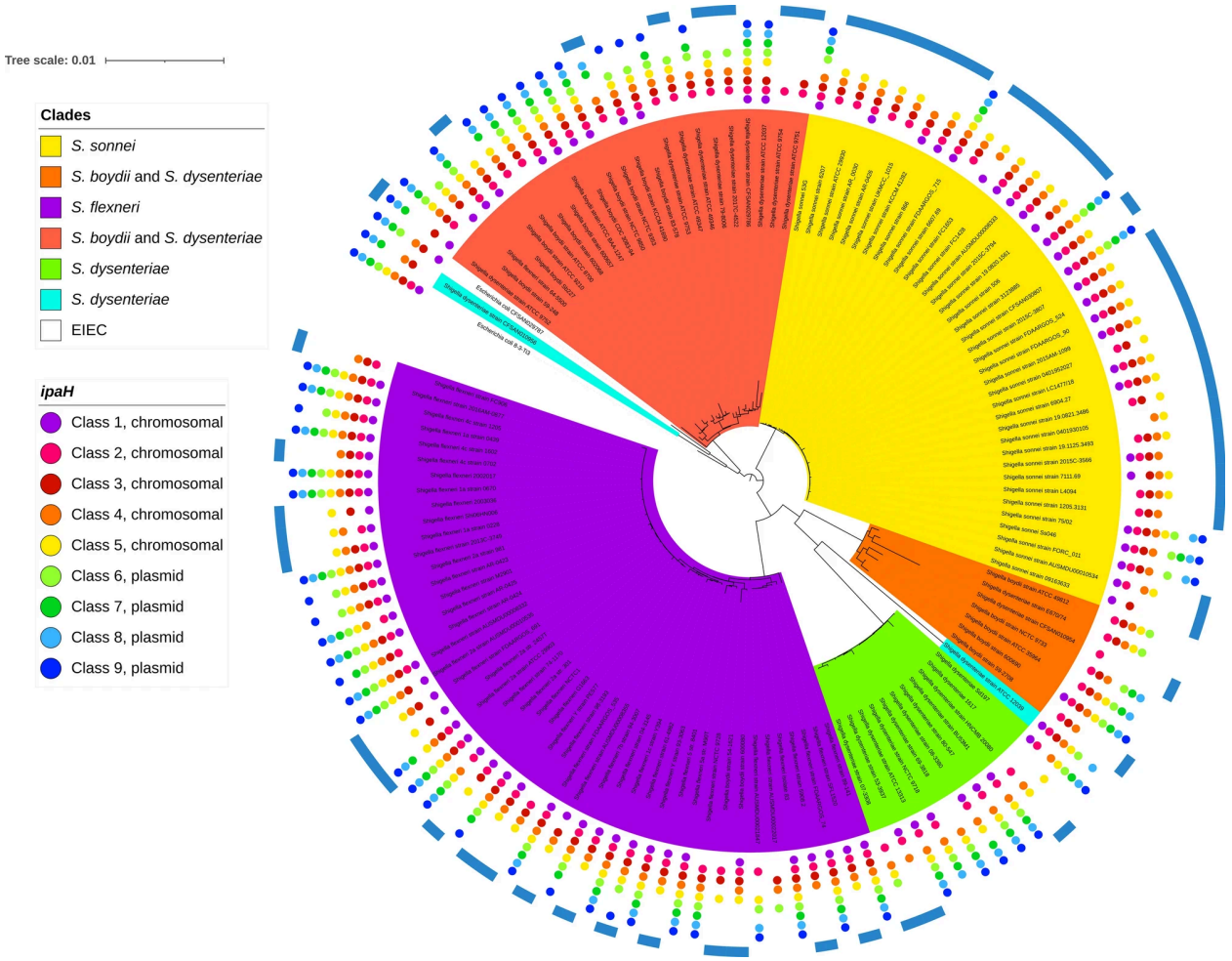


Рисунок 4. Филетические паттерны генов *ipaH* у шигелл и ЭИКП. Покраска неукоренённого дерева отражает основные клады шигелл, которые предположительно произошли от различных непатогенных *E. coli*; два отдаленных штамма шигелл показаны голубым цветом, штаммы ЭИКП показаны белым. Наличие генов *ipaH* показано точками, цвет которых отражает класс *ipaH* (см. условные обозначения). Гены классов 1-5 расположены на хромосомах; гены классов 6-9 — на плазмидах. Геномы, отмеченные внешними синими дугами, не содержат генов CC3T.

Интересно, что копии генов *ipaH* были обнаружены во многих геномах шигелл как на хромосомах, так и на плазмидах (Дополнительная таблица 3). Паралоги *ipaH* 2, 4, 5 наблюдаются в оранжевой (*S. boydii* и *S. dysenteriae*) кладе, паралоги только *ipaH* 4 присутствуют в зеленой (*S. dysenteriae*) кладе и паралоги только *ipaH* 3 были обнаружены в красной (*S. boydii* и *S. dysenteriae*)

кладе (Дополнительный рисунок 2A). Геномы фиолетовой клады (*S. flexneri*) имели паралоги *ipaH* 3, 4, 5, 7 и 9 (Дополнительный рисунок 2B), в то время как геномы в желтой кладе (*S. sonnei*) не имели дуплицированных *ipaH* (Дополнительный рисунок 2C). Удивительно, но ни один из паралогов *ipaH* не был тандемным повтором; напротив, копии располагались на некотором расстоянии друг от друга и часто были окружены профагами и псевдогенами.

2.3.5 Репертуар *ipaH* у *Escherichia* spp. из животных хозяев

Был проведен поиск и сравнение генов *ipaH* у патогенных штаммов *Escherichia* spp., выделенных у животных хозяев (Дополнительная таблица 4, Рисунок 5). Ранее девять генов *ipaH* и две коротких ОРС, содержащих фрагменты генов *ipaH*, были обнаружены в геноме *Escherichia marmotae* HT073016, выделенном из фекалий *Marmota himalayana* [79]. Авторы сообщили об автоматическом аннотировании одиннадцати генов как *ipaH*: четырех на плазмиде *pEM148*, пяти на плазмиде *pEM76* и двух на хромосоме. В соответствии с используемой в работе процедурой идентификации *ipaH* (см. «Материалы и методы») было подтверждено восемь из них и обнаружен один дополнительный хромосомный ген. Из анализа были исключены короткие ОРС, которые не содержали N-концевой домен, поскольку представляется вероятным, что это результат неправильной аннотации или остатки псевдогенов.

Кроме того, *E. coli*, выделенная из животных хозяев, содержала гены CC3T, а также *ipaH*. В частности, два штамма, выделенные из фекалий крыс, *E. coli* CFSAN092688 и *E. coli* CFSAN085900, имели шесть и три гена *ipaH* соответственно, а штамм из фекалий овец, *E. coli* RHB04-C17, имел три гена *ipaH*.

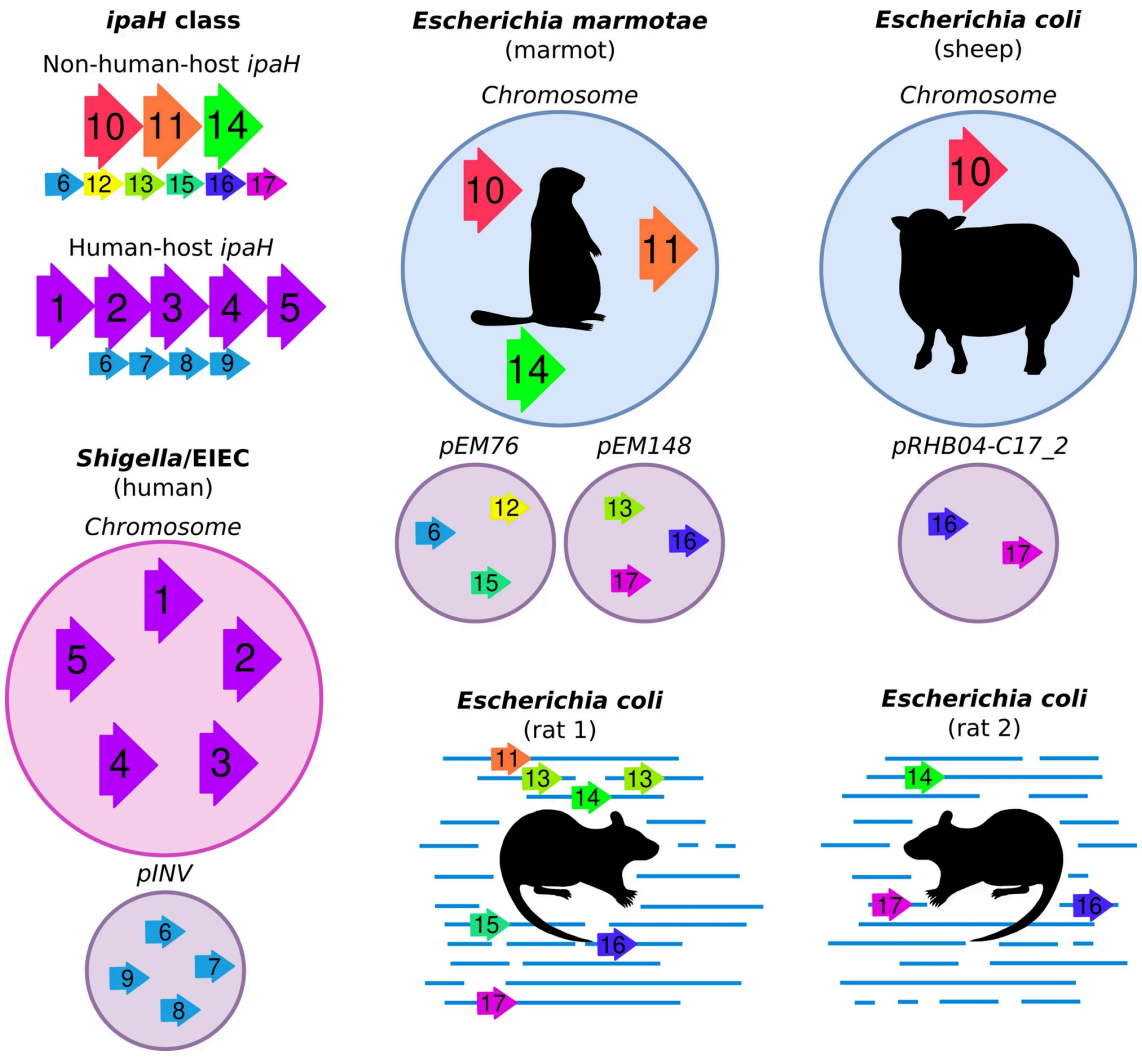


Рисунок 5. Состав генов *ipaH* в геномах *Escherichia* spp. от разных хозяев. Сборки *Escherichia* из сурка и овцы полные, геномы *Escherichia* spp. из крысы собраны в виде контигов. Для шигелл показан геном с полным набором генов *ipaH*.

Основываясь на сходстве последовательностей распознающих доменов, белки IpaH из *E. coli* с хозяевами-животными были разделены на девять классов (Рисунок 6A,C). Уровень сходства последовательностей между IpaH из животных, принадлежащих к одному и тому же классу, меньше, чем у *Shigella* spp.. Основываясь на расположении генов *ipaH* в полностью собранных геномах *Escherichia* spp., то есть из сурков и овец, можно предположить, что они сохраняют свое местоположение на репликонах.

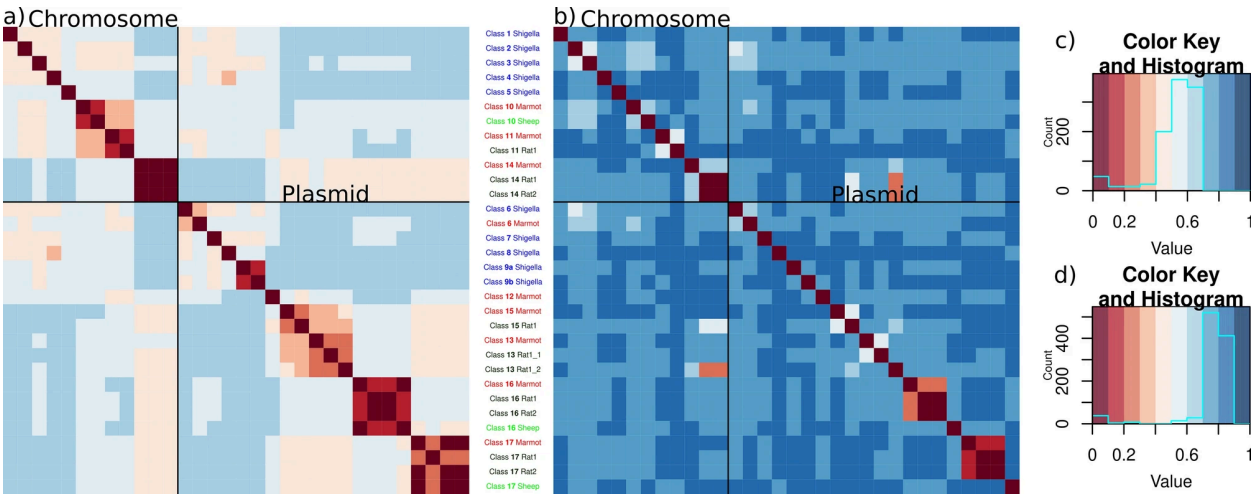


Рисунок 6. Тепловая карта попарных расстояний и соответствующие цветовые ключи (A, C) генов *ipaH*; (B, D) их регуляторных последовательностей у штаммов *Escherichia* spp. из животных. Хозяева помечены по следующему принципу: сурок отмечен красным, крыса отмечена темно-зеленым, овца отмечена светло-зеленым. Репрезентативные последовательности генов *ipaH Shigella* spp. также были включены в сравнение, они отмечены синим цветом. Попарные расстояния были рассчитаны как $\sqrt{1 - identity}$.

Два класса *ipaH* 16 и 17, предположительно плазмидные, были обнаружены у всех видов *Escherichia* spp. из животных хозяев. Предположительно хромосомный класс *ipaH* 14 присутствовал у штаммов *Escherichia* spp. из сурка и крысы; в свою очередь, геномы *Escherichia* spp. из сурка и овцы имеют общий класс *ipaH* — 10. Только один из генов *ipaH* у *Escherichia* spp. из животных (класс 6), присутствовал в *Shigella* spp., однако регуляторные области *ipaH* класса 6 у *Shigella* spp. и *E. marmotae* значительно отличались (Рисунок 6B,D). В частности, регуляторные области *ipaH* у штаммов *Escherichia* spp. из животных хозяев не содержит ни сайтов связывания MxiE, ни полиA/полиT-последовательностей.

Регуляторные области большинства классов *ipaH* были сходны у *E. coli* из крыс и сурков, в то время как регуляторные области в генах *ipaH* у *E. coli* из овец были уникальны. Классы 13 (на плазмиде) и 14 (на хромосоме) демонстрировали сходство последовательности генов и регуляторной области, но имели разное количество БЛП, что указывает на их эволюцию

путем дупликации генов и последующих делеций или, скорее, тандемной дупликации коротких сегментов генома.

В отличие от шигелл и кишечных палочек человека у штаммов *Escherichia* spp. из животных хозяев С-концевой домен белков IpaH не сохранялся (Дополнительная таблица 6). У IpaH классов 10, 11, 12 из животных хозяев был С-концевой домен, аналогичный таковому у *Shigella* spp. (92% идентичности по аминокислотам), в то время как С-концевые домены IpaH классов 13-17 отличались сильнее (75% идентичности по аминокислотам) (Рисунок 7). Это наблюдение также объясняет результаты [79], согласно которым только часть обнаруженных генов *ipaH* были гомологичны *ipaH* из *Shigella* spp.. Примечательно, что оба варианта являются специфичными для *E. coli* и отличными от убиквитин-лигазных доменов других патогенов, таких как *Salmonella*, *Yersinia* и другие.

Аминокислотные замены между консенсусными последовательностями С-концевых доменов IpaH из *Shigella* spp. и *Escherichia* spp. из животных хозяев были отмечены на трехмерной структуре эффектора *Shigella flexneri* IpaH1880 (PDB: 5KH1) (Рисунок 7, Дополнительный рисунок 3). Эти различия не были сгруппированы и не влияли на активный сайт белка.

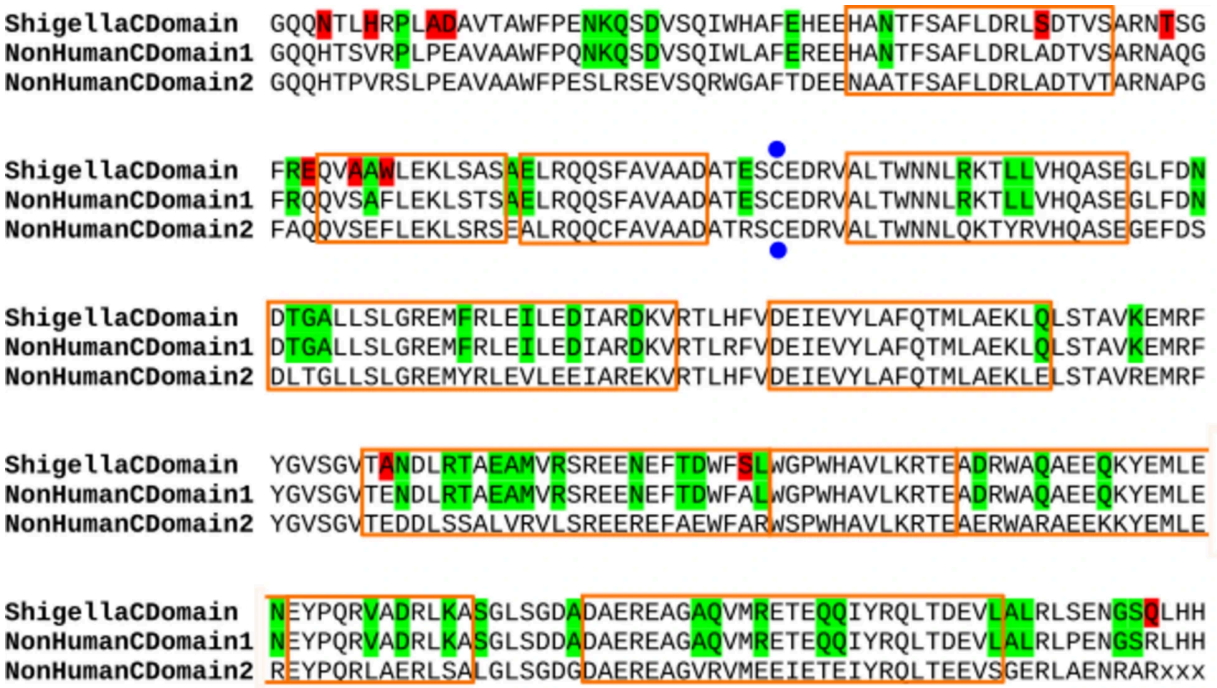


Рисунок 7. Выравнивание С-концевых доменов IpaH у *Shigella* spp. и *E. marmotae*. Показаны консенсусные последовательности. Активный сайт отмечен синими точками, альфа-спирали показаны оранжевыми рамками. Различия между белками из *Shigella* spp. и обоими белками *E. marmotae* отмечены красным цветом, различия между двумя типами С-концевых доменов у *E. marmotae* отмечены зеленым.

2.4 Обсуждение

Штаммы шигелл и ЭИКП обладают широким спектром эффекторов IpaH, которые играют значительную роль в инвазии, модуляции воспаления и ответа хозяина [7]. Ранее в нескольких исследованиях предпринимались попытки описать семейство генов *ipaH* у шигелл, направленные на сравнение репрезентативных штаммов из разных видов шигелл [93,94]. Однако виды шигелл и известные линии ЭИКП являются парафилетическими с сильно изменчивыми геномами. Поэтому для получения общей картины состава и эволюции семейства генов потребовался всесторонний сравнительный анализ, объединяющий все доступные геномные данные.

В настоящей работе был собран большой набор генов *ipaH*, которые были классифицированы на основе сходства последовательностей, их номенклатура была унифицирована, с сохранением ссылок на ранее

использовавшиеся названия генов [93,94]. Хотя последовательности большинства генов *ipaH* высоко консервативны у разных штаммов, в классе 9 (*ipaH1.4*) была замечена диверсификацая паралогов, которая может указывать на формирование нового класса *ipaH*. Учитывая важную роль этого семейства в вирулентности шигелл, последовательная аннотация генов имеет прямое медицинское значение. Полученные результаты показывают, что использование консенсусных последовательностей *ipaH* из каждого класса для аннотации генов уменьшает количество ошибок в аннотациях и может быть полезно для будущих исследований этого семейства генов.

Наличие эффекторов IpaH является одним из маркеров, используемых для серотипирования *Shigella* [95], однако менее половины секвенированных геномов содержали весь набор генов *ipaH*. Вопреки предыдущим наблюдениям на небольших наборах данных [91], ни один из генов *ipaH* не является общим для всех штаммов *Shigella*, а филетические паттерны генов *ipaH* предполагают многочисленные независимые потери генов. Мишени многих белков IpaH неизвестны, однако было показано, что некоторые белки влияют на один и тот же путь на разных стадиях, работая вместе, чтобы вызвать заболевание [7]. В этом случае полный набор IpaH был бы функционально избыточным и не обязательно должен быть сохранен. Также стоит отметить, что в случае бактериальных изолятов элиминация плазмид и факторов вирулентности в процессе культивирования, возможно, привела к потере плазмид до секвенирования генома [92].

Наличие нетандемных копий генов *ipaH* с консервативными регуляторными участками у многих штаммов шигелл указывает на получение фрагментов ДНК с *ipaH* из того же источника, а также на функциональность и специфичность регуляторных областей *ipaH*.

Не было обнаружено каких-либо последовательных различий в репертуаре генов *ipaH* у патотипов шигелл и ЭИКП. Более того, регуляторные элементы в 5’-областях генов *ipaH* были одинаковыми.

Поскольку обозначения патотипов шигелл и ЭИКП четко не определены, до сих пор неясно, ответственны ли эти факторы за различия в инфекционной дозе и тяжести заболевания при патотипах шигелл/ЭИКП.

Интересно, что состав *ipaH* и 5’-области генов у штаммов эшерихий из животных существенно отличались от штаммов, полученных из человека. В общей сложности было обнаружено восемь новых классов эффекторов IpaH у эшерихий из животных хозяев. Как и в кишечных палочках из человека, гены, кодирующие эффекторы, сохраняют свое расположение на репликонах. Стоит отметить, что, хотя *E. marmotae* является внешней группой для клады *E. coli* [79], *E. coli* из крысы и овцы содержат эффекторы IpaH, сходные с таковыми у *E. marmotae*, в то время как эффекторы *E. coli* из человека уникальны; единственный класс *ipaH* (6, *ipaH9.8*) присутствовал как на плазмидах шигелл, так и на плазмидах *E. marmotae*. Несоответствия между филогенией штаммов и составом эффекторов указывают на горизонтальный перенос генов между *E. coli*, адаптированными к разным хозяевам. В отличие от *ipaH* шигелл, регуляторные области *ipaH* у штаммов эшерихий из животных хозяев не содержат ни сайтов связывания MxiE, ни множества полиA/T-последовательностей: единственным примером с двумя такими треками является *ipaH* класса 16 у сурка и крысы.

В белках IpaH, кодируемых в геномах эшерихий из животных, были обнаружены два различных С-концевых домена. Это наблюдение может быть объяснено горизонтальным переносом эффекторов, а также дифференциацией их функциональных ролей у эшерихий из различных хозяев.

Белки IpaH рассматриваются в качестве потенциальных мишеней для антибиотиков из-за их специфичности для шигелл. Первая стратегия заключается в нацеливании на С-концевой домен, поскольку он консервативен среди эффекторов IpaH у шигелл [96]. Однако IpaH может влиять на антимикробную активность белков хозяина даже в отсутствие

каталитической активности [97]. Таким образом, нацеливание на N-концевые домены может быть более эффективным, но эта стратегия требует понимания репертуара *ipaH* у конкретных штаммов. На сегодняшний день подходы, используемые для тестирования наличия факторов вирулентности шигелл, не позволяют различать представителей семейства *ipaH* [98], поэтому требуется разработка геноспецифичных праймеров.

Варианты *ipaH* инвазивных эшерихий из диких и домашних животных потребуют дополнительного изучения, поскольку они могут способствовать эволюции патогена человека. Примечательно, что аннотация источника образцов кишечной палочки может вводить в заблуждение. В частности, геном *E. coli*, выделенный из овечьего кала, собранного с пола фермы (биопроба: SAMN15147991), может быть загрязнен бактериями другого хозяина, например крысы, живущей на ферме. Если бы это было так, то новый вариант С-концевого домена белков IpaH мог бы быть специфичным для грызунов. Обширный отбор проб и последующее секвенирование геномов эшерихий из разных хозяев могут пролить свет на специфичность системы инвазивности и эффекторов IpaH к хозяевам патогенов.

Глава 3. Эволюция белков гистидиновых триад у *Streptococcus*

3.1 Введение

Бактерии рода *Streptococcus* могут быть как компонентами нормальной флоры, так и возбудителями различных инфекций у животных и человека; кроме того, в этот род входят виды, важные для пищевой промышленности [99]. В этом роду присутствует сразу несколько медицински значимых видов, такие как *S. pneumoniae, S. pyogenes, S. agalactiae* [100]. Как правило эти бактерии вызывают болезни дыхательных путей, однако иногда могут вызвать сепсис и заболевания любых органов и систем. Наиболее изученным представителем этого рода является *S. pneumoniae*, патоген, являющийся причиной наибольшего числа случаев развития пневмонии по всему миру [100].

В силу высокого разнообразия свойств видов исследуемого рода, в разное время предлагались различные классификации видов внутри рода. Часто виды этого рода разделяются по способности к гемолизу: вызывающие частичный, полный и никакого гемолиза [101]. Кроме того, в медицине для β-гемолитических штаммов продолжает использоваться классификация по Лэнсфилд, то есть разделение бактерий по группам по углеводному составу белков на клеточной стенке [102]. Однако в силу развития методов сравнительной геномики и после выделения новых родов из семейства *Streptococcaceae*, было предложено ещё множество различных классификаций [99]. На текущий момент в роде *Streptococcus* на основе филогенетических отношений и молекулярных маркеров предлагается разделение на две крупные клады, *Mitis–Suis* и *Pyogenes–Equinus–Mutans,* которые в свою очередь разделяются на 14 более мелких групп [99].

В силу того, что патогенные штаммы представлены в таксономически удалённых подгруппах *Streptococcus* spp., выделение общего механизма и детерминант патогенности для всех стрептококков представляется крайне

затруднительным [103], однако для отдельных клад такого рода анализ оказывается результативен. Например, в случае *S. pneumoniae* было показано, что ключевыми аспектами колонизации организма этими бактериями являются установление контакта с эпителием, расщепление слизи, взаимодействие с системой комплемента, контроль связывания металлов и провоспалительные эффекты пневмолизина [104]. Одним из факторов патогенности, участвующих во взаимодействии с системой комплемента, являются белки гистидиновых триад [104].

Белки гистидиновых триад являются поверхностными белками стрептококков и несут по несколько копий характерного мотива HxxHxH, где H — гистидин, а x — любая другая аминокислота [105]. Изначально эти белки были обнаружены в процессе поиска потенциальной мишени для вакцины от пневмококковой инфекции [106]. Белок PhtA был одной из шести мишеней, для которой был показан протективный эффект от сепсиса на мышах [106]. Всего у *S. pneumoniae* было обнаружено четыре варианта таких белков: PhtA, PhtB, PhtD и PhtE [107]. Для этих белков было показано связывание фактора комплемента H, что помогает бактерии ускользать от опосредованных комплементом иммунных реакций хозяина [108]. Впоследствие аналогичные белки были обнаружены в других заражающих людей стрептококках *S. pyogenes* [109] и *S. agalactiae* [110], а также в зоонозном *S. suis* [111].

Ранее предполагалось, что белки гистидиновых триад встречаются исключительно у бактерий рода *Streptococcus* [112], однако позже они были обнаружены в таких родах как *Granulicatella, Gemella, Slackia, Catonella, Aerococcus* и *Facklamia* [103]. При этом род *Catonella* принадлежит к классу *Clostridia*, *Slackia* — к классу *Coriobacteriia*, а остальные относятся к тому же классу *Bacilli*, что и *Streptococcus*. Было выделено две крупные группы белков гистидиновых триад, содержащие и не содержащие домен, богатый лейциновыми повторами [103]. По структуре дерева этих белков было

определено, что разделение белков гистидиновых триад на эти две группы произошло ещё до формирования рода *Streptococcus* и, кроме того, гены белков гистидиновых триад подвергались горизонтальным переносам [103]. Скорость эволюции вдоль белка также не постоянна: сам триадный мотив и домен, богатый лейциновыми повторами, консервативны, а остальные части белка находятся под сильным положительным отбором [103].

Для генов, кодирующих белки гистидиновых триад в *S. pneumoniae*, была показана парная локализация на хромосоме, *phtD* в паре с *phtE* и *phtA* в паре с *phtB* [105]. Экспрессия генов белков гистидиновых триад, как и многих других факторов патогенности у стрептококков, регулируется глобальным стрептококковым регулятором AdcR, чувствительном к концентрации цинка [113]. Кроме того, белки гистидиновых триад сами участвуют в гомеостазе цинка совместно с AdcAII [112,114]. Было показано, что первый из пяти мотивов гистидиновой триады у PhtD играет важную роль в гомеостазе цинка у *S. pneumoniae* [114]. Примечательной особенностью стрептокков также является перекрестное взаимодействие регуляторов, чувствительных к цинку, и регуляторов, чувствительных к меди, что обеспечивает большую устойчивость бактерий к цинковой интоксикации [115]. У стрептококков группы B был показан ответ медного регулятора CopY и глобального регулятора вирулентности CovR на цинковый стресс [115].

Ещё одной важной особенностью генов, кодирующих белки гистидиновых триад у пневмококков, является их подверженность фазовой вариации [10]. Несмотря на то, что изначально эти белки были идентифицированы как потенциальная мишень для противопневмококковых вакцин и показывали в исследованиях значительную иммуногенность у мышей [105,116], было обнаружено, что гены этих белков могут изменяться путём обратимой инверсии крупного участка генома [10,117]. Эта инверсия захватывает 5’-концевые фрагменты генов белков гистидиновой триады стаким образом, что после инверсии получаются два полнофункциональных

гена, регулируемых цинковым репрессором, с которых считываются два новых белка. Путем такой перестановки бактерия получает иную комбинацию поверхностных белков, что позволяет ей избегать иммунного ответа хозяина [117].

Целью этого проекта было описать функции белков гистидиновых триад у *Streptococcus* spp. на основе анализа регуляции, геномного контекста и разнообразия их генов.

3.2 Материалы и методы

3.2.1 Геномные данные

В работе использовались 819 полных геномов *Streptococcus* spp., доступных в GenBank [81] по состоянию на апрель 2023 года (Дополнительная таблица 7).

3.2.2 Идентификация генов, кодирующих белки гистидиновых триад

Для начальной аннотации геномов *Streptococcus* spp. был использован модуль annotate инструмента PanACoTA [84]. В рамках этого модуля осуществляется предсказание открытых рамок считывания и предсказание функций полученных генов с помощью prokka [118] и prodigal [119].

Так как гены известных из экспериментов белков гистидиновых триад, как правило, определяются автоматическими аннотаторами как «hypothetical protein», требовались дополнительные способы поиска соответствующих генов.

Первым таким способом стало использование базы известных последовательностей белков гистидиновых триад из GenBank [81] в качестве материала для составления HMM-профиля (Hidden Markov Model) и дальнейшего поиска соответствующих белков с помощью HMMer [120]. Было обнаружено 2555 белков. Значительным недостатком такого подхода стал тот факт, что обнаруженные таким способом белки и соответствующие им гены не представляли собой полные ортологические группы. Кроме того, у части

обнаруженных белков не было мотивов гистидиновых триад. Такой способ поиска был признан неподходящим.

Вторым способом поиска белков гистидиновых триад выступил прямой поиск белков, содержащих по меньшей мере два мотива гистидиновых триад HxxHxH. В результате такого поиска было обнаружено 1802 белка, образующих при пороге сходства в 50% 47 ортогрупп, состоящих только из белков гистидиновых триад, в числе которых 14 синглтонов, и 7 ортогрупп, в которых присутствуют и другие белки. Каждая ортогруппа, содержащая не только белки гистидиновых триад, была проверена вручную. Во всех таких группах число мотивов триад уменьшалось в результате делеции, затрагивающей исследуемый мотив.

3.2.3 Поиск регуляторных мотивов в регуляторных областях генов, кодирующих белки гистидиновых триад

Для предсказания регуляторных мотивов была использована программа, разработанная в нашей лабораторией И. Жаровым на основе библиотеки MOODS [121]. В качестве исходных мотивов для составления позиционно-весовых матриц были использованы данные, приведённые в публикациях о CsoR [122], CopY [123], MtsR [124], AdcR [113] (Дополнительная таблица 8). В качестве порогового веса для обнаружения сайта в случае CopY, MtsR и AdsR было выбрано значение 3.5, а для CsoR значение 4 в силу большей длины сайта.

3.3.4 Предсказание генов, предположительно участвующих в фазовой вариации

В настоящей работе рассматривался случай фазовой вариации, при котором инверсия захватывает части кодирующих последовательностей пары генов (Рисунок 8A,B). Для предсказания такого типа фазовой вариации использовалась следующая процедура. Для всех возможных попарных сочетаний генов, кодирующих белки гистидиновых триад, из каждой

возможной пары геномов строилось кодонное выравнивание. Каждое выравнивание было разбито на 20 окон, для каждого такого окна вычислялся процент несовпадений. Две пары генов считались потенциальными кандидатами на участие в фазовой вариации, если выполнялись следующие условия: наличие 14 окон, уровень различий в которых не превосходит 20%, наличие не менее двух окон, уровень различий в которых превосходит 40%, отсутствие длинных инделов, наличие резкой границы, то есть скачка на расстоянии не более одного окна, между участками высокого сходства и различия, расположение генов на разных цепях. После этого производился поиск таких двух геномов, у которых у одной пары генов совпадают последовательности в 5’ области гена, а у другой — в 3’ области гена (Рисунок 8C). Таким образом отбирались геномы, в которых предположительно происходит фазовая вариация.

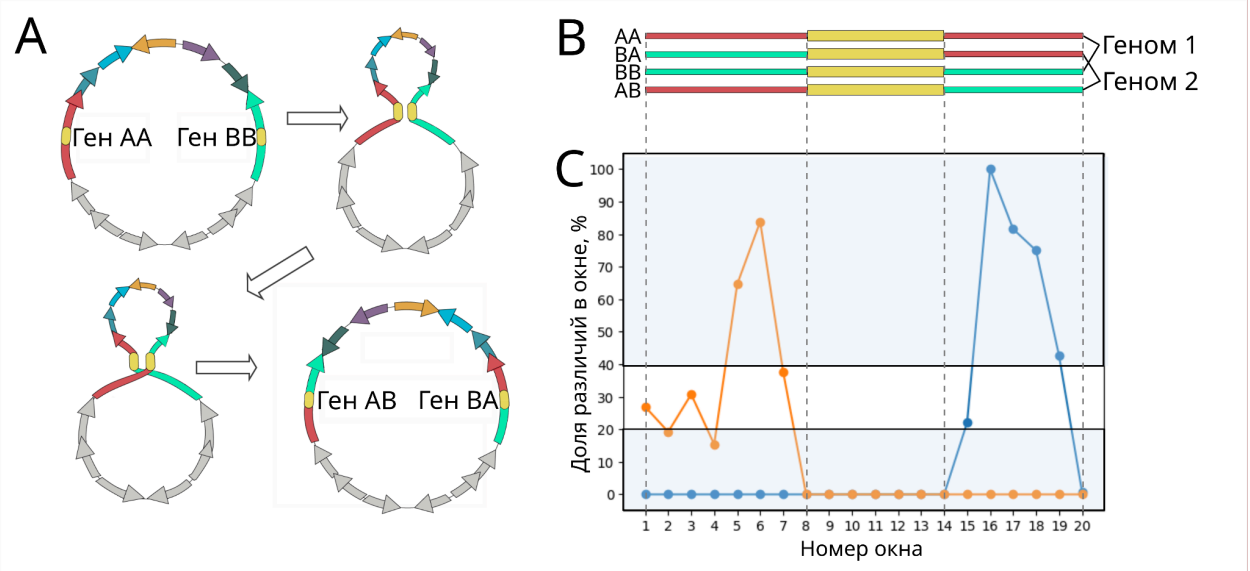


Рисунок 8. A) Схема исследуемого типа фазовой вариации. B) Схематичное изображение выравнивания генов, подверженных фазовой вариации. C) Пример выравнивания генов, подверженных фазовой вариации по типу инверсии. Оранжевая линия соответствует выравниванию генов AA-BA и BB-AB из примера с панели B, а голубая линия выравниванию генов AA-AB и BA-BB.

3.3 Результаты

3.3.1 Разнообразие белков гистидиновых триад

Было проанализировано 819 полных геномов *Streptococcus* spp., принадлежащих к 64 различным известным видам (Дополнительная таблица 7). В 696 геномах из 54 видов встретились гены, кодирующие белки гистидиновых триад, в количестве от одного до шести генов на геном.

Всего в исследуемом наборе данных присутствовало 1802 гена, кодирующих белки гистидиновых триад, из которых 14 синглтоны, то есть гены, не входящие ни с какими другими в одну ортогруппу. На дереве соответствующих последовательностей эти белки не образуют клады по видам (Рисунок 9). В большинстве случаев автоматические аннотаторы не предсказывают функий исследуемых белков, однако представители одной из ортологических групп стабильно аннотируются как CopB, белки, участвующие в гомеостазе меди [125]. Интересно, что в штаммах стрептококков гены, аннотируемые как *copB*, принадлежат шести независимым ортогруппам, из которых одна крупная группа полностью состоит из генов, кодирующих белки, не содержащие мотивов гистидиновых триад. Остальные пять состоят из генов, кодирующих белки гистидиновых триад.

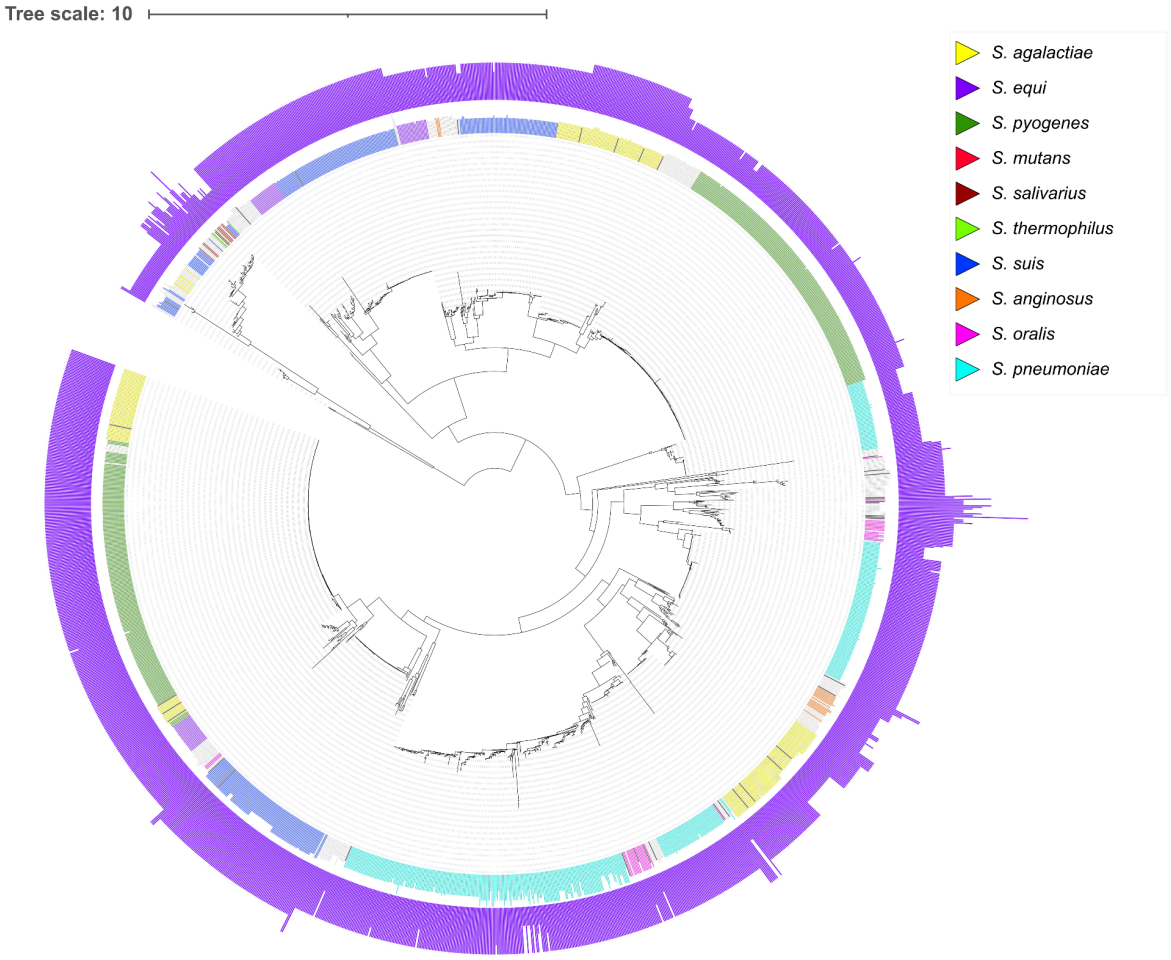


Рисунок 9. Филогенетическое дерево белков гистидиновых триад. Листья окрашены в соответствии с видом *Streptococcus*, из которого был взят белок. Высота столбца во внешнем фиолетовом кольце указывает на число гистидиновых триад в белке.

В полных геномах стрептококков были найдены гены белков гистидиновых триад, у которых присутствовало от 2 до 10 и только один раз 14 мотивов гистидиновых триад (Рисунок 10).

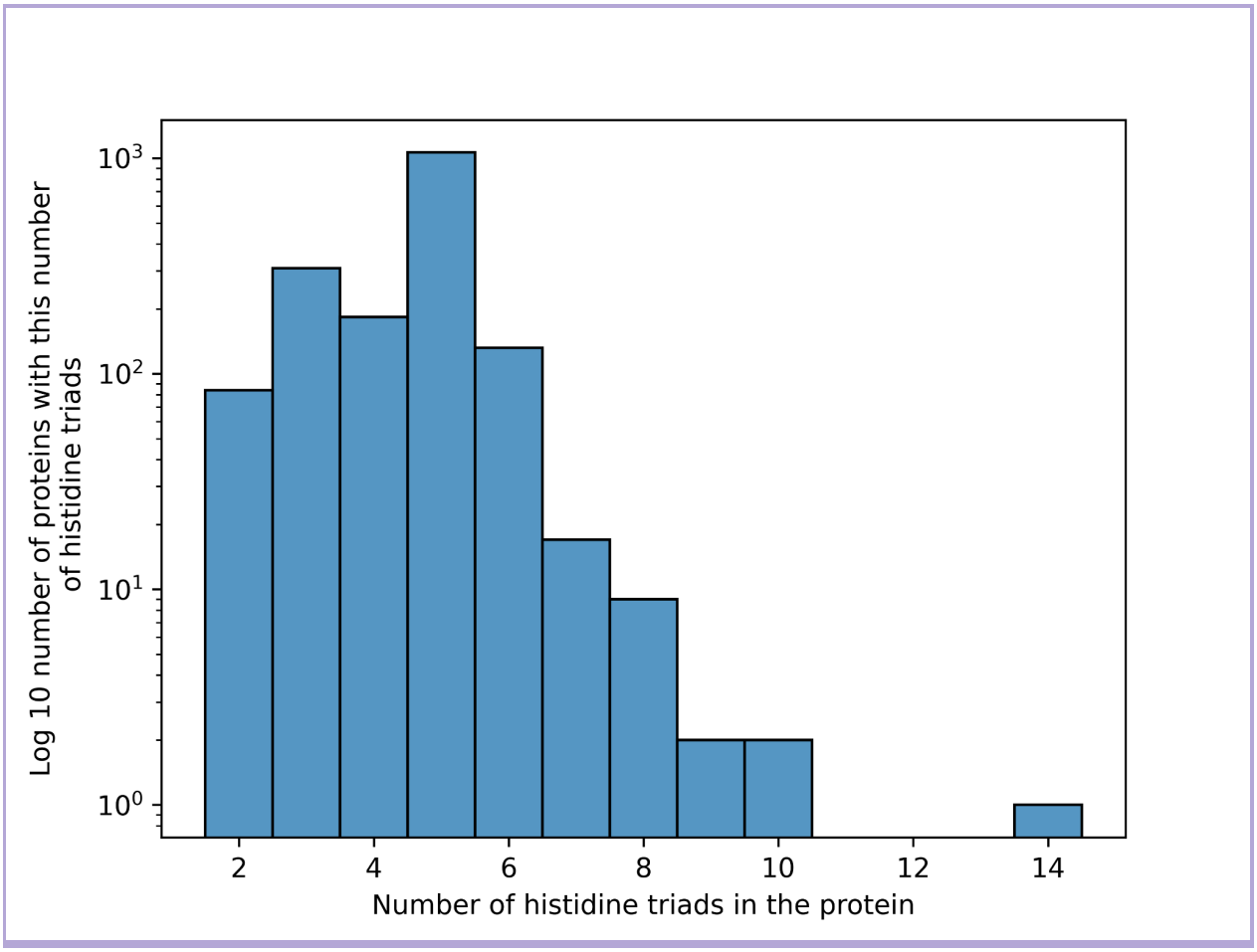


Рисунок 10. Гистограмма числа исследуемых белков с соответствующим числом гистидиновых триад в них.

При анализе мотивов гистидиновых триад мы сформулировали две гипотезы. Первая гипотеза состояла в том, что соседние вдоль белка мотивы похожи больше, чем удалённые; она не подтвердилась. Вторая гипотеза состояла в том, что мотивы группируются по сходству по позициям вдоль белка, то есть мотивы с одинаковым номером внутри белка более похожи, чем белки с разными номерами. Далее в рамках этой гипотезы возникло дополнительное предположение, что самый близкий к N-концу белка мотив гистидиновой триады должен быть самым консервативным, так как именно для него для одного из белков гистидиновых триад у *S. pneumoniae* была показана принципиальная важность в гомеостазе цинка [114]. И действительно, в случае белков, у которых пять, как у PhtD из *S. pneumoniae*, или шесть гистидиновых триад, первый мотив является самым

консервативным по последовательности. Тем не менее, для белков с меньшим количеством гистидиновых триад такая консервативность не характерна (Рисунок 11). Более того, консенсусы первых двух мотивов для белков с пятью и шестью триадами различаются: HGDHYH и HGDHEH, соответственно; любопытно также, что третья триада у обеих групп имеет консенсус HGDHYH, а четвертая — HGDHEH. В первой триаде белков с меньшим числом триад и в ряде других случаев консенсус, скорее, HXXHSH, где X = G, D, S.

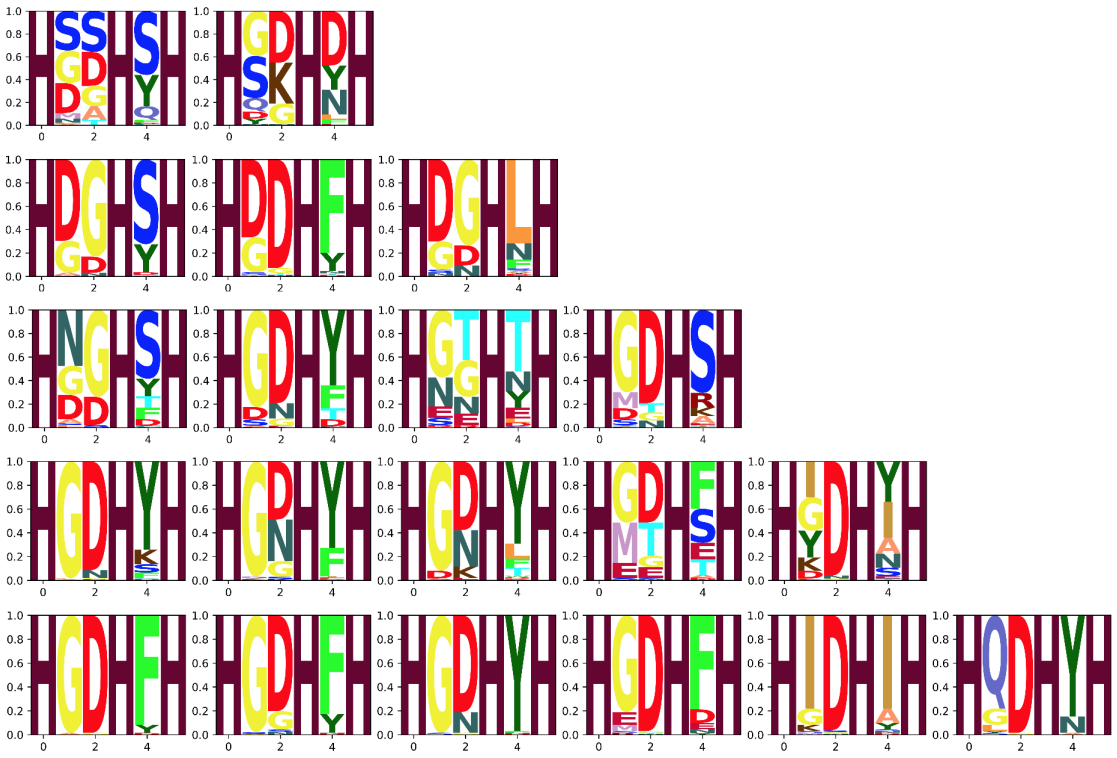


Рисунок 11. Частоты встречаемости аминокислот в гистидиновых триадах, сгруппированные по белкам с разным количеством гистидиновых триад и представленные в виде лого. Рассмотрены только группы, содержащие 20 или более белков.

3.3.2 Представленность разных вариантов в разных видах *Streptococcus*

В 10 видах исследуемого набора данных не обнаружилось ни одного гена, кодирующего белок гистидиновых триад: *S. mutans*, *S. equinus*, *S. sobrinus*, *S. ruminicola*, *S. vestibularis*, *S. troglodytae*, *S. ratti*, *S. lactarius*, *S. ferus*, *S. alactolyticus*. Почти все из них, кроме *S. alactolyticus*, описаны как представители нормальной оральной или кишечной флоры, разве что

вызывающие в некоторых случаях кариес у хозяина [126].

Наибольшее же число генов, кодирующих белки гистидиновых триад, в среднем четыре, были обнаружены у *S. ruminantium*, *S. iniae* и *S. pneumoniae*. Все эти виды патогенны и вызывают серьёзные инфекции у жвачных животных, рыб и людей соответственно [127–129]. При этом максимальное число таких генов, шесть, было обнаружено у представителей вида *S. pneumoniae* (Рисунок 12).

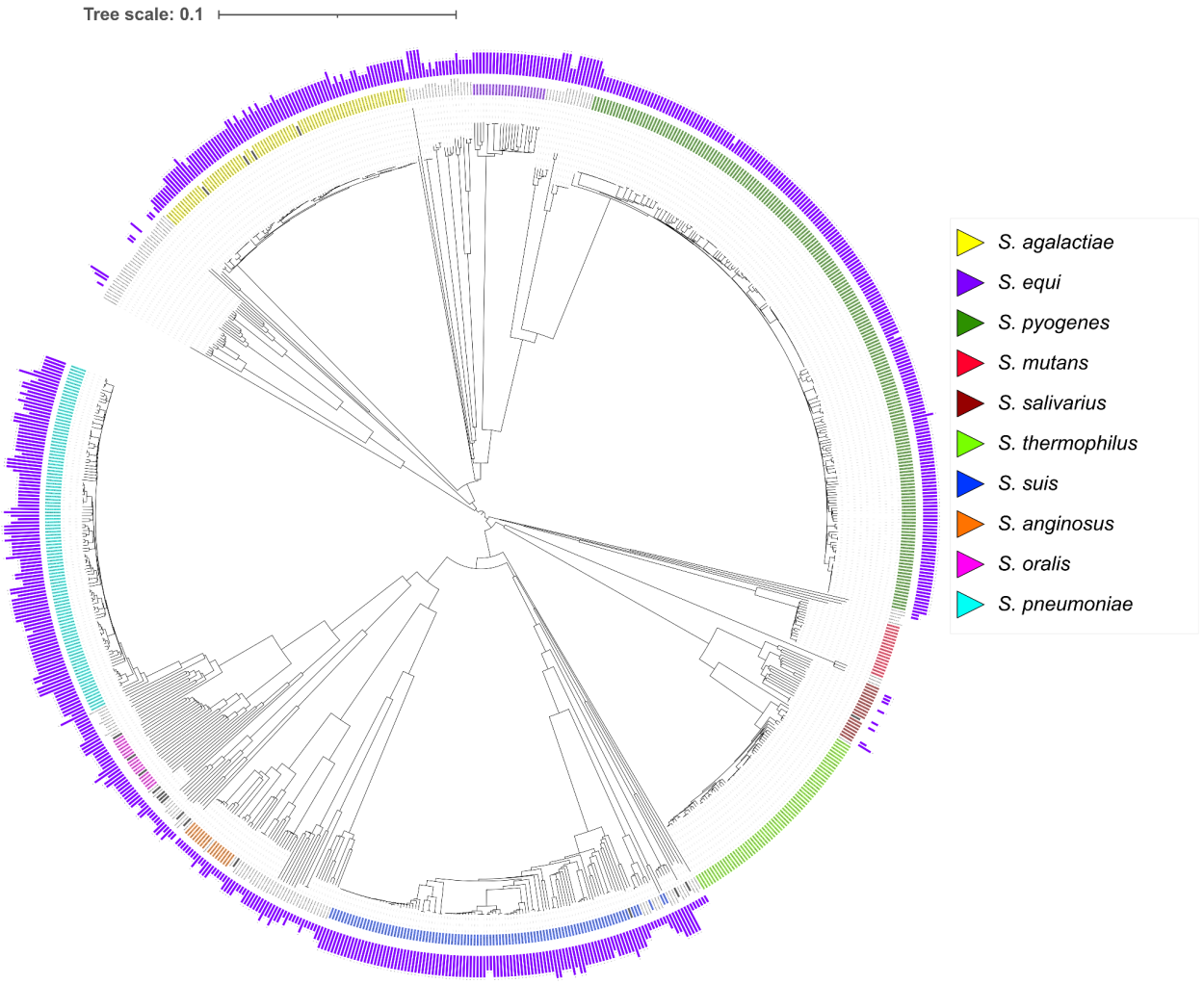


Рисунок 12. Филогенетическое дерево штаммов *Streptococcus*, построенное по универсальным однокопийным генам. Листья окрашены в соответствии с видом *Streptococcus*. Ширина полосы внешнего фиолетового кольца отображает количество исследуемых генов в соответствующем геноме.

3.3.3 Регуляция генов, кодирующих белки гистидиновых триад

Так как из литературы известно, что гены, кодирующие белки

гистидиновых триад, регулируются стрептококковым цинковым репрессором AdcR [130], для всех исследуемых генов был проведён поиск сайтов связывания AdcR.

Такие сайты были обнаружены у всех исследуемых генов за исключением двух групп, локализованных на дереве на двух удалённых ветвях, и нескольких единичных генов в других местах дерева (Рисунок 13). На одной из этих ветвей представлены только пневмококковые гены, причём ни один из них не регулируется цинковым репрессором. Вторая же ветвь содержит гены из разных видов *Streptococcus*, и часть этих генов регулируется цинковым репрессором.

Чтобы понять, чем могут регулироваться гены с этих ветвей и в каких процессах участвовать, для каждого гена, кодирующего белок гистидиновых триад, были определены гены-соседи с обеих сторон [131]. На ветвях с цинковыми репрессорами исследуемые гены часто ко-локализованы с *znuA* (из 1671 гена на этих ветвях 769 имеют в соседях *znuA,* 46%) и лежат с ним предположительно в одном опероне, так как сонаправлены и имеют только один сайт связывания репрессора. Белок ZnuA — это субъединица транспортёра ZnuABC, который отвечает за гомеостаз цинка [132,133]. На ветви, где ни один из генов не регулируется цинковым репрессором, гены белков гистидиновых триад всегда соседствуют с открытыми рамками считывания, для которых не удалось предсказать функцию кодируемых белков. На второй же ветви исследуемые гены часто соседствуют с генами гомеостаза меди, иногда встречаются транспортёры калия.

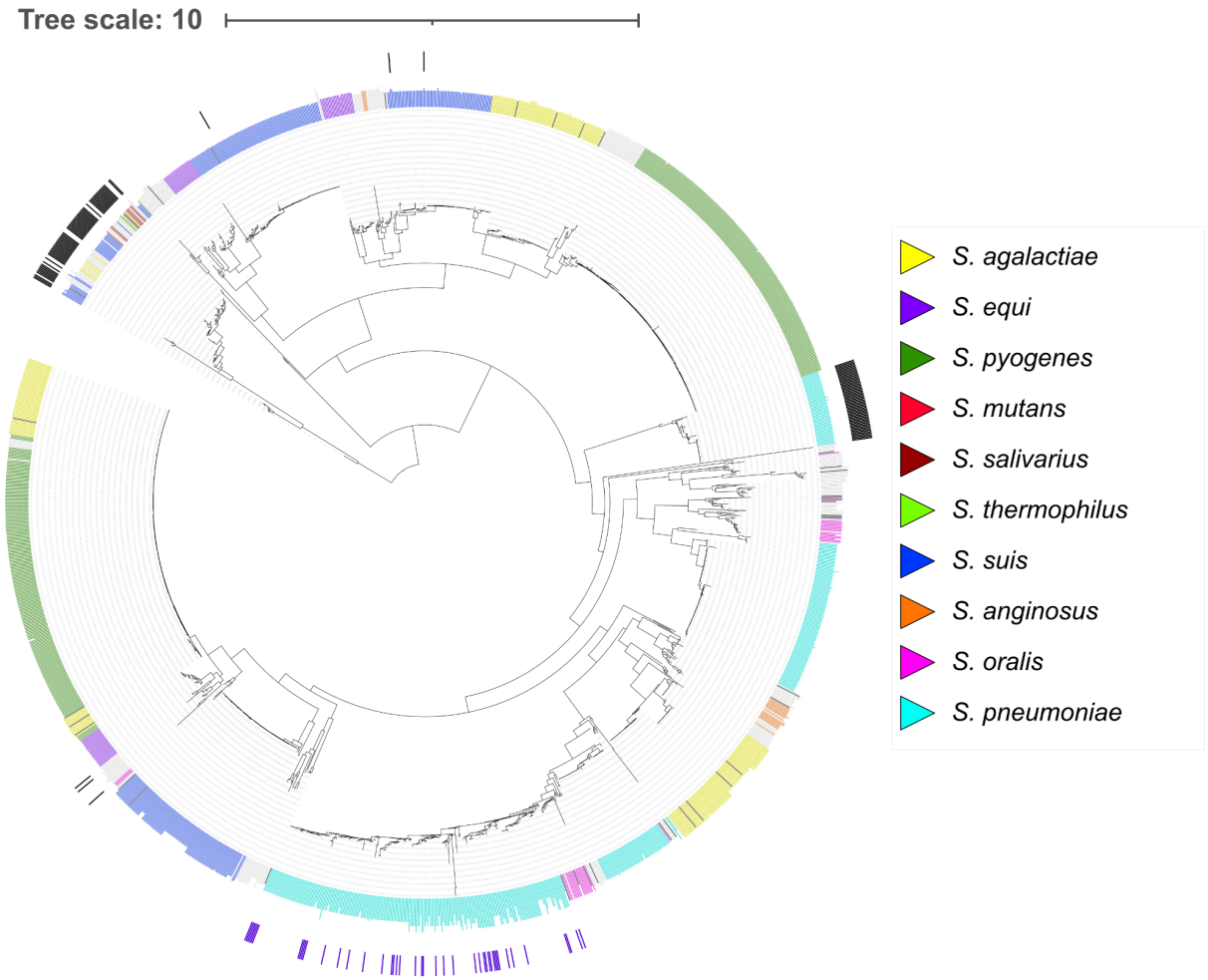


Рисунок 13. Филогенетическое дерево белков гистидиновых триад с рисунка 9. Цвета листьев соответствуют виду *Streptococcus*. Черными шрихами обозначены гены, для которых не найдены потенциальные сайты связывания цинковых репрессоров. Фиолетовыми штрихами (внешнее кольцо) обозначены гены, для которых возможна фазовая вариация по механизму инверсии (детали см. в тексте).

Чтобы проверить, действительно ли гены белков гистидиновых триад регулируются медными репрессорами, был произведён поиск мотивов медных регуляторов CsoR, CopY и MtsR (см. Материалы и методы). И действительно, на ветви, где исследуемые гены аннотируются как *copB* и соседствуют с генами гомеостаза меди, часть генов белков гистидиновых триад контролируется медными репрессорами (Рисунок 14). Примечательно, что для регулятора CopY был показан ответ не только на изменение

концентрации меди, но и на изменение концентрации цинка [115]. Ещё одним отличием белков этой ветви от остальных белков дерева является высокая доля серина в последовательностях гистидиновых триад и меньшее количество самих триад. На этой ветви доля серина составляет 32%, в то время как для остальных белков дерева доля серина только 4%, z-тест для долей с двусторонней альтернативой показывает, что это отличие является значимым на высоком уровне (z-stat = 34.55, p-value = 1.45×10^{-261}).

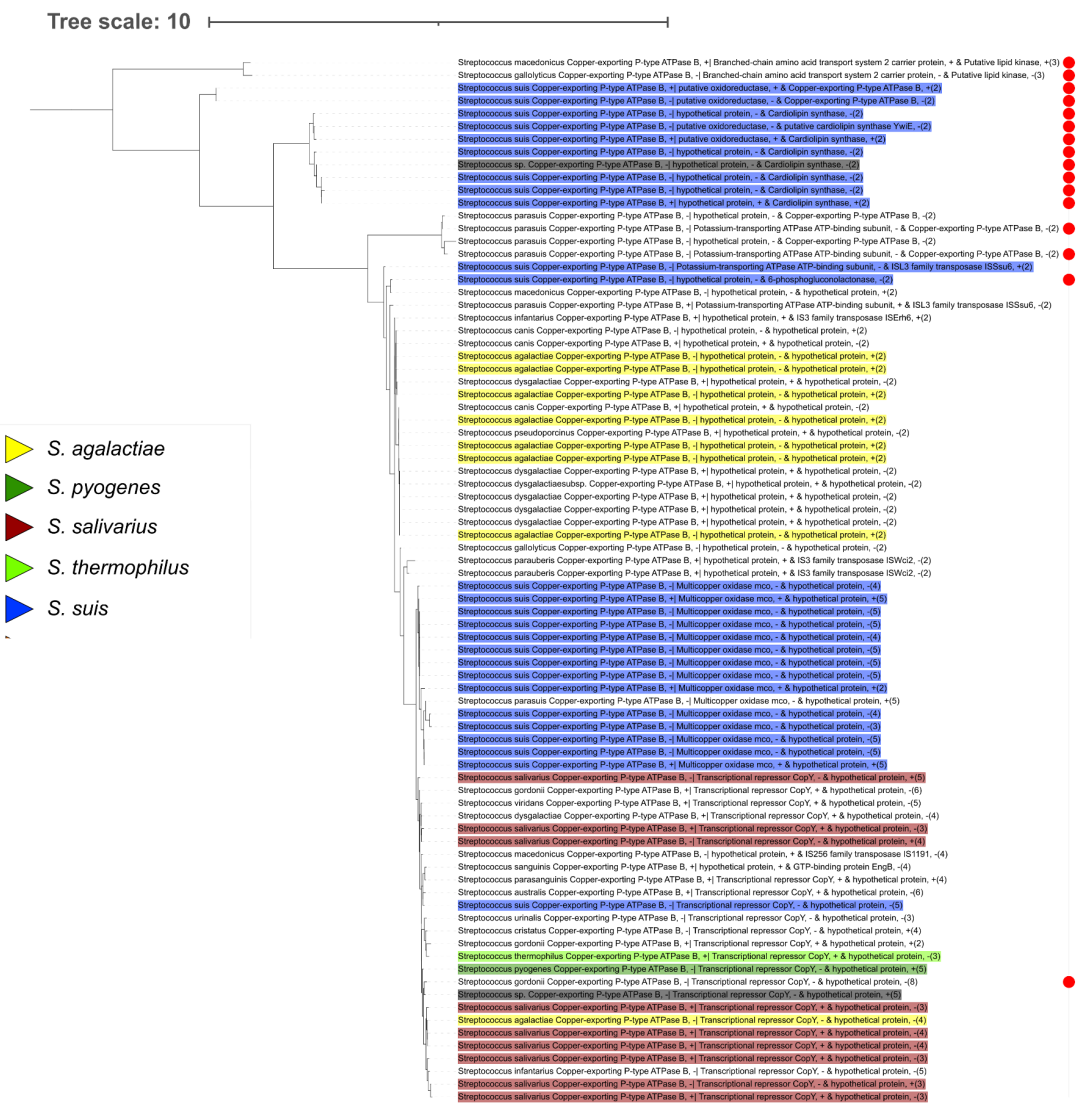


Рисунок 14. Детализация ветви дерева с рисунка 11, гены на которой не регулируются цинковыми репрессорами, но регулируются медными репрессорами. Красными точками отмечены гены, предположительно контролируемые медными репрессорами.

Вторая ветвь, не регулируемая цинковыми репрессорами, содержит белки, которые с помощью BLAST аннотируются как PhtE. Несмотря на такую аннотацию, средняя длина белка на этой ветви в пять раз меньше описанной в литературе. Кроме того, отсутствуют характерная для генов *phtE* регуляция цинковыми репрессорами, и частая для гена *phtE* ко-локализация с *znuA*.

3.3.4 Фазовая вариация как способ избегания иммунитета

Для некоторых генов стрептококков было показано наличие фазовой вариации. Для генов систем рестрикции-модификации третьего типа показана фазовая вариация по числу простых повторов [134]. Такого рода вариации в регуляторных областях генов белков гистидиновых триад обнаружить не удалось. Ещё одним известным механизмом фазовой вариации для генов *Streptococcus* является инверсия. Для двух пневмококковых белков гистидиновых триад было показано наличие фазовой вариации [10], при которой инверсия захватывает фрагменты двух генов, начиная со стопа. Чтобы проверить, есть ли у каких-либо ещё видов стрептококков пары генов белков гистидиновых триад, для которых возможна фазовая вариация по тому же механизму, был предложен метод поиска таких последовательностей (см. Материалы и методы).

В исследуемых геномах гены в подходящем положении были обнаружены в 21 геноме, только у вида *S. pneumoniae*, причём во всех случаях потенциальная инверсия может происходить между двумя конкретными локусами, расположенными на расстоянии около 150 т.п.о. (Рисунок 13). При этом два варианта последовательности N-конца белка комбинируются с двумя вариантами последовательности C-конца белка, что даёт четыре (две пары) возможных поверхностных белка, кодируемых такой системой генов. Кроме варьирования последовательности белка такого рода инверсия может контролировать экспрессию, если у пары генов только один промотор. Тогда в зависимости от положения инвертируемого фрагмента экспрессируется

только один ген из пары. В случае пневмококков в одном из участвующих в вариации локусов расстояние от старта гена, кодирующего белок гистидиновых триад, до старта предыдущего гена составляет около 250 п.о. и на расстоянии 83 п.о. от этого старта располагается потенциальный сайт связывания цинкового репрессора. Во втором локусе старт гена, кодирующего белок гистидиновых триад, располагается на расстоянии всего 8 п.о. от предыдущего гена *znuA*. Ген *znuA*, расположенный перед геном, кодирующим белок гистидиновых триад, находится под контролем цинкового репрессора, потенциальный сайт связывания которого располагается на расстоянии в 64 п.о. от старта *znuA*. В силу крайне малого расстояния между генами во втором локусе, участвующем в фазовой вариации, можно предположить, что исследуемый ген экспрессируется и контролируется цинковыми репрессором вместе со *znuA*.

3.4 Обсуждение

Белки гистидиновых триад представляют собой один из факторов патогенности бактерий в роде *Streptococcus*. Они широко представлены на дереве штаммов этого рода [103]. Ранее в основном описывали пневмококковые белки гистидиновых триад [130,135], а анализ других представителей семейства проводился на выборках гораздо меньшего размера [103]. В настоящем исследовании был проведён анализ белков гистидиновых триад из всех доступных полных геномов стрептококков, что позволило подробно изучить также регуляторные особенности, геномный контекст и участие в фазовой вариации соответствующих генов. В качестве определяющего фактора для идентификации белков гистидиновых триад был выбран не профиль HMM, а наличие не менее двух триадных мотивов.

Исследуемые белки на дереве располагаются группами, соответствующими виду, что согласуется с предыдущим наблюдением о том, что гены, кодирующие белки гистидиновых триад, возникли в бактериях ещё до выделения рода *Streptococcus* [103], и независимо дуплицировались в ходе

эволюции отдельных видов. Только одна из ветвей представляет исключение из этого правила: белки из разных видов локализованы плотными группами. Примечательно, что именно у этих белков относительно длинные ветви.

Ранее было показано, что гены, кодирующие белки гистидиновых триад, контролируются у стрептококков цинковым репрессором AdcR и участвуют в гомеостазе цинка [112–114]. Хотя цинковая регуляция имеет место у большинства генов, кодирующих белки гистидиновых триад, на дереве обнаружились две ветви, являющиеся исключениями. Одна из них содержит гены, для которых предсказана регуляция медными репрессорами, а вторая представляет собой набор белков, которые аннотированы как PhtE и содержат два или три мотива гистидиновых триад, однако эта версия PhtE гораздо короче описанного и исследованного в литературе PhtE, характерной особенностью которого является наличие шести мотивов гистидиновых триад [107,112]. Вопрос о том, что представляют собой белки этой ветви, остаётся открытым. В то же время наличие ветви, регулируемой медными репрессорами, может быть закономерным следствием участия регулятора CopY в борьбе не только с медным, но и цинковым стрессом, так как для этого регулятора был показан ответ на оба этих типа стресса [115]. В то же время нехарактерная для других ветвей дерева высокая доля серина в мотивах гистидиновых триад на этой ветви может быть знаком именно смены специфичности, так как, например, для PhtD из *S. pneumoniae* было показано связывание цинка именно в гистидиновой триаде [136]. Таким образом можно предположить формирование нового класса гистидиновых триад, взаимодействующих именно с ионами меди.

Примечательным является тот факт, что фазовая вариация обнаружена только в случае *S. pneumoniae*, хотя этот вид далеко не единственный патоген в исследуемом наборе данных: довольно агрессивные патогены *S. agalactiae* и *S. pyogenes* также присутствуют в исследуемой выборке.

Глава 4. Эволюция специфичности транспортёров металлов семейства CorA

4.1 Введение

Металлы являются важной частью жизненных процессов клетки, около 25-30% белков нуждаются в металлах для нормального функционирования [137]. Ионы металлов могут входить в состав крупных биомолекул, элементов регуляторных систем, играть роль каталитических кофакторов в различных реакциях, участвовать в процессах переноса электронов [138]. Например, цинк важен для процессов регуляции синтеза ДНК, дифференцировки, пролиферации и митоза, но избыток цинка токсичен для клетки и токсичная концентрация в этом случае относительно низка [139]. Кроме того, это важный элемент взаимодействия хозяин-патоген, необходимый для заражения и колонизации патогеном хозяина [140]. В процессе такого взаимодействия, чтобы снизить вирулентность патогена, хозяин пытается секвестрировать весь свободный цинк как внутри клеток, так и во внеклеточном пространстве [141].

Одним из самых важных двухвалентных ионов металлов является магний. Этот катион участвует в основных клеточных процессах, таких как репликация ДНК, транскрипция, трансляция, передача сигналов в клетке и энергетический обмен [142]. Кроме того, катион магния обладает уникальными свойствами. Он сильно отличается от других биологически значимых двухвалентных катионов по размеру, структуре в водном растворе и плотности заряда. Например, размеры гидратированного и негидратированного катиона магния различаются в 400 раз, а активность катиона магния сильно зависит от геометрии связанных молекул воды [143]. Особые свойства ионов магния определяют уникальность химии взаимодействия с ним и, как следствие, уникальность магниевых транспортных систем. Как правило, транспортёру сначала требуется распознать ион в его крупной гидратированной форме, потом удалить с него

гидратную оболочку и траспортировать уже маленький ион без оболочки в клетку. В случае с магнием из-за существенной разницы в размерах гидратированных и негидратированных ионов это совершенно особая задача. Для её решения многие бактерии используют транспортёр CorA (Cobalt Resistance protein A) [144].

Этот белок был описан в *Escherichia coli* при изучении устойчивости к кобальту [145]. Было показано, что воздействие высоких концентраций кобальта подавляет рост бактерий, однако этот эффект не проявляется в присутствии больших концентраций магния, а в случаях, когда бактерии были устойчивы к большим концентрациям кобальта, они также демонстрировали сниженный уровень транспорта магния. Такое поведение позволяет сделать заключение о совместном транспорте этих катионов [146]. Примечательно, что CorA также допускает экспорт в зависимости от концентрации магния вне клетки [146].

Как правило, CorA находится в клетке в состоянии гомопентамера, каждый мономер которого состоит из большого растворимого сильно заряженного гидрофильного домена на N-конце и небольшого охватывающего мембрану гидрофобного домена на С-конце [147,148]. У большинства бактерий длина белка CorA составляет от 300 до 400 аминокислот. Сходство белков семейства CorA между собой возрастает от N-конца к С-концу; в то же время трехмерная структура различных представителей семейства очень консервативна.

В семействе CorA и его функциональных аналогах Mrs2 и Alr1 присутствует высококонсервативная последовательность, считающаяся сигнатурным мотивом всего семейства и частью селективного фильтра транспортера, последовательность которого определяет возможность транспорта катионов металлов через эти белки. Однако литературные данные противоречивы и нет общего мнения относительно связи последовательности в мотиве и транспорта конкретных катионов. В работе [149] упоминались два

консервативных мотива: YGMNF и MPEL, но также говорится, что первый из них может быть преобразован в GMN у эукариотических гомологов. Также было высказано предположение, что замены в мотиве GMN, за исключением GVN и GIN, предотвращают любой перенос по этому каналу. В [150] было показано, что мотив GMN работает как внешний координационный лиганд для селективного фильтра. В работе [151] также упоминается, что существует подкласс белков CorA, называемый CorA-II, который утратил мотив MPEL и, как предполагается, является экспортером магния, а не импортером. Авторы также предположили, что отделение CorA-II произошло довольно рано. В семействе CorA традиционно выделяют две подгруппы: подгруппа A, близкие гомологи CorA из *Termotoga marititma*, и подгруппа B, близкие гомологи CorA из *Escherichia coli*, однако различий в сигнатурном мотиве между группами не выявлялось.

Отдельным подтверждением возможности одновременного транспорта как магния, так и кобальта, является химическое сходство молекул, которые эти ионы образуют в природе. Катион магния почти всегда связывает шесть молекул воды в октаэдрической форме. Эта вода образует гидратную оболочку толщиной около 5 мкм, которая, как уже упоминалось, является самой крупной гидратированной формой среди других двухвалентных катионов. Такое же лигандное состояние характерно для катионов переходных металлов, например кобальта, никеля и рутения. Эти катионы могут ковалентно связываться с различными лигандами, которые образуют альтернативную оболочку. Например, катион кобальта, ковалентно связанный с шестью молекулами аммиака, идентичен гидратированному катиону магния с точки зрения размера и геометрии получаемого соединения [152]. Такие катионы имитируют гидратированный катион магния. Присутствие катионов кобальта не оказывает влияния на другие транспортёры магния и магний-зависимые белки, но ингибирует поступление магния через CorA. Другие катионные гексаамины не оказывают ингибирующего влияния на

CorA. Кроме того, данные о соединениях, препятствующих транспорту магния через CorA, показывают, что начальный участок связывания магния имеет диаметр около 5 мм, приспособленный для связывания недегидратированного катиона, как у многих других транспортёров, а полностью гидратированного, и не различает двух- и трехвалентные катионы.

Хотя у многих бактерий белок CorA экспрессируется конститутивно, в некоторых случаях экспрессия гена *corA* может контролироваться M-box рибопереключателем [153] (другое название этого рибопереключателя YkoK-leader [154]). Он может регулировать экспрессию генов трех основных классов бактериальных транспортёров магния CorA, MgtE, MgtA/B, но часто встречается только перед генами семейств MgtE и MgtA/B. В недавней статье [155] предсказан кандидат на роль рибопереключателя, специфичного для семейства CorA.

Несмотря на то, что ионы цинка, как правило, имеют отличную от ионов магния химическую конформацию, для некоторых представителей семейства CorA (ZntB/CmaX) была показана возможность транспорта цинка [156,157]. По-видимому, транспортные механизмы у CorA и ZntB разные, что было показано в эксперименте о чувствительности к мембранному потенциалу, к которому CorA чувствителен, а ZntB — нет [14]. Было также показано, что ZntB переносит протон и зависит от протонного градиента, а CorA от него не зависит [14]. В других экспериментах было установлено, что градиент pH увеличивает поглощение цинка через транспортер ZntB, а обратный градиент подавляет поглощение [157]. Структура транспортёра ZntB похожа на структуры других представителей семейства CorA. Это гомопентамер, каждый мономер которого состоит из двух независимых доменов: большого цитоплазматического N-концевого домена и небольшого охватывающего мембрану C-концевого домена, состоящего из двух сегментов. Для ZntB были показаны иные, чем у CorA, сигнатурные мотивы селективного фильтра: либо GVN [157], либо GIN [149] вместо GMN, однако

эти данные противоречивы.

Как правило, в клетке должна поддерживаться конкретная концентрация ионов металлов. Таким образом, транспорт металлов имеет важное значение для биологических систем. В ситуации, когда клетка нуждается в определенном металле для выполнения конкретной задачи, а его избыток может быть токсичным, используются различные системы для выборочной транспортировки его в клетку. Для накопления определенного уровня требуемого иона металла поддерживается баланс между активностью импортёров и экспортёров [158]. Это позволяет избежать ситуации, когда из-за доступности различных металлов в клетке происходит замещение менее активных металлов более активными. Активность транспортёров управляется через контроль уровня экспрессии их генов чувствительными к концентрации металлов транскрипционными факторами или рибопереключатель [159]. Такие сенсоры могут контролировать определенный ген или целый регулон [160]. Конкретные механизмы специфичны для разных металлов и сенсоров.

Таким образом, признаком специфичности транспортёра может служить наличие в 5’-области его гена регуляторных элементов, зависящих от концентрации транспортируемых ионов. Аналогично контролю транспорта магния с помощью магниевого рибоперелючателя M-box, транспорт кобальта в клетках также может контролировать кобаламиновый рибопереключатель, так как синтез кобаламина требует наличия ионов кобальта [161]. В свою очередь импорт цинка обычно регулируется глобальными факторами, такими как ZUR (zinc uptake regulator) или AdcR у стрептококков, связывающими участки ДНК в регуляторной области гена [162].

Целью этой работы было предсказание специфичности белков семейства CorA и анализ связи последовательности сигнатурного мотива GxN с их специфичностью для дальнейшей экспериментальной верификации.

4.2 Материалы и методы

4.2.1 Данные

Последовательности белков семейства CorA в настоящей работе представляют собой один ортологический ряд COG0598 из базы данных EggNOG, содержащий 2545 последовательностей [163]. Из этого ортологического ряда были удалены последовательности из не бактериальных организмов. Из выборки также были удалены фрагментарные последовательности и последовательности, содержащие более одного гена, то есть опероны, для которых не был предсказан стоп-кодон. Итоговая выборка была образована 2102 белками длиной от 233 до 461 аминокислот.

Данные о доступных структурах белков семейства CorA были взяты из базы данных RCSB PDB [164]. HMM-профиль магниевого и кобаламинового рибопереключателей был взят из базы данных Rfam (магниевый: RF00380, кобаламиновые: RF01482, RF01689) [165].

4.2.2 Построение выравниваний

Множественные выравнивания строились с помощью MUSCLE с параметрами по умолчанию [166].

4.2.3 Филогенетическое дерево

Для построения филогенетического дерева всех исследуемых белковых последовательностей был использован пакет phyML [167] с моделью LG и дискретным гамма-распределением с 4 категориями на 100 бутстрэпов. Визуализация дерева проводилась онлайн с помощью iTOL [86].

4.2.4 Аннотация регуляторных элементов

Для предсказания рибопереключателей, контролирующих гены исследуемых белков, была использована программа Infernal [168]. Для предсказания регуляторных мотивов была использована программа, разработанная в нашей лабораторией И. Жаровым на основе библиотеки

MOODS [121]. В качестве потенциальной позиции для расположения сайта рассматривалась область перед геном длиной 300 пар оснований, а порог веса сайта определялся по появлению потенциальных сайтов перед известными генами без цинковой регуляции. Данные об известных сайтах связывания цинковых репрессоров ZUR и AdcR были взяты из [113]. Известные сайты использовались в качестве основы для построения позиционной весовой матрицы, после чего новые сайты с наибольшими весами, расположенные перед генами, для которых из литературы известна регуляция цинковыми репрессорами, добавлялись в выравнивание для построения итоговой матрицы.

4.2.5 Предсказание позиций, определяющих функциональную специфичность

Для предсказания последовательностей, определяющих функциональную специфичность белка, использовался инструмент SDPpred [169].

4.2.6 Визуализация белковых структур

В качестве инструмента визуализации использовалось программное обеспечение UCSF Chimera [90].

4.3 Результаты

4.3.1 Разнообразие мотивов в семействе

Как указывалось выше, из литературы следует, что мотив GxN является определяющим специфичность транспортёра в семействе CorA. В исследуемом наборе белков самым распространённым оказался мотив GMN, который ассоциирован с магниевым транспортом и описывается как каноничный мотив семейства; таких белков оказалось 1773. Неканоничных, но также встречающихся в литературе, вариантов оказалось на порядок меньше: 114 GVN и 87 GIN. Эти три мотива в совокупности имеются в 1974

из 2102 белков, то есть около 94%.

4.3.2 Мотивы и филогенетика

Филогенетическое дерево всех имеющихся последовательностей в целом согласуется с известной таксономией с точностью до нескольких горизонтальных переносов. Для того, чтобы выявить связь между мотивом предположительно селективного фильтра и специфичностью белка, на дерево была добавлена разметка различных сигнатурных мотивов (Рисунок 15).

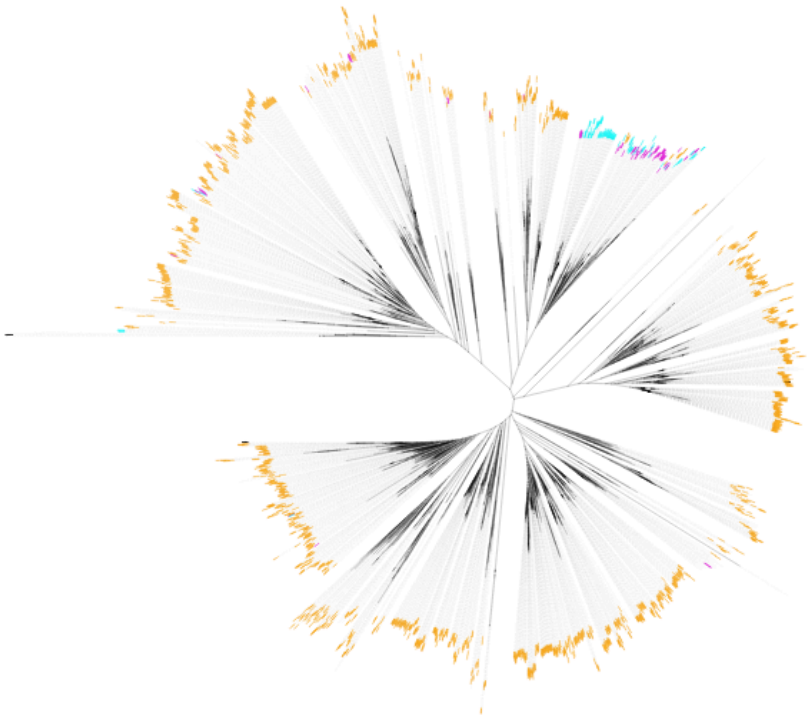


Рисунок 15. Филогенетическое дерево последовательностей транспортёров металлов семейства CorA. Цвета листьев соответствуют сигнатурному мотиву соответствующего белка: GMN — оранжевый, GVN — бирюзовый, GIN — розовый, минорные варианты мотива отмечены черным цветом.

Ветвь, на которой располагается большинство неканоничных мотивов селективного фильтра, представлена в основном гамма-протеобактериями и альфа-протеобактериями, с редкими вкраплениями других протеобактерий (Рисунок 16). На этой ветви можно видеть, что гены этого семейства активно передаются горизонтально: белки из *Salmonella* обнаружились на ветви белков *Citrobacter*, белок из *Cronobacter* оказался на ветви белков

Enterobacter, белок из *Grimontia* располагается на ветви *Vibrio*, а целая группа белков из разных штаммов *Vibrio* располагаются на ветви, близкой к альфа-протеобактериям, что лучше всего объясняется переносом генов от альфа-протеобактерий к общему предку *Vibrio*. Белки из дельтапротеобактерий не образуют монофилетической группы и образуют включения на ветвях как гамма- так и альфа-протеобактерий. Основываясь на структуре дерева, можно предположить, что горизонтальный перенос сопровождался появлением белков с мотивами GVN и GIN в родах *Vibrio*, *Oceanicola* и *Pseudoalteromonas*.

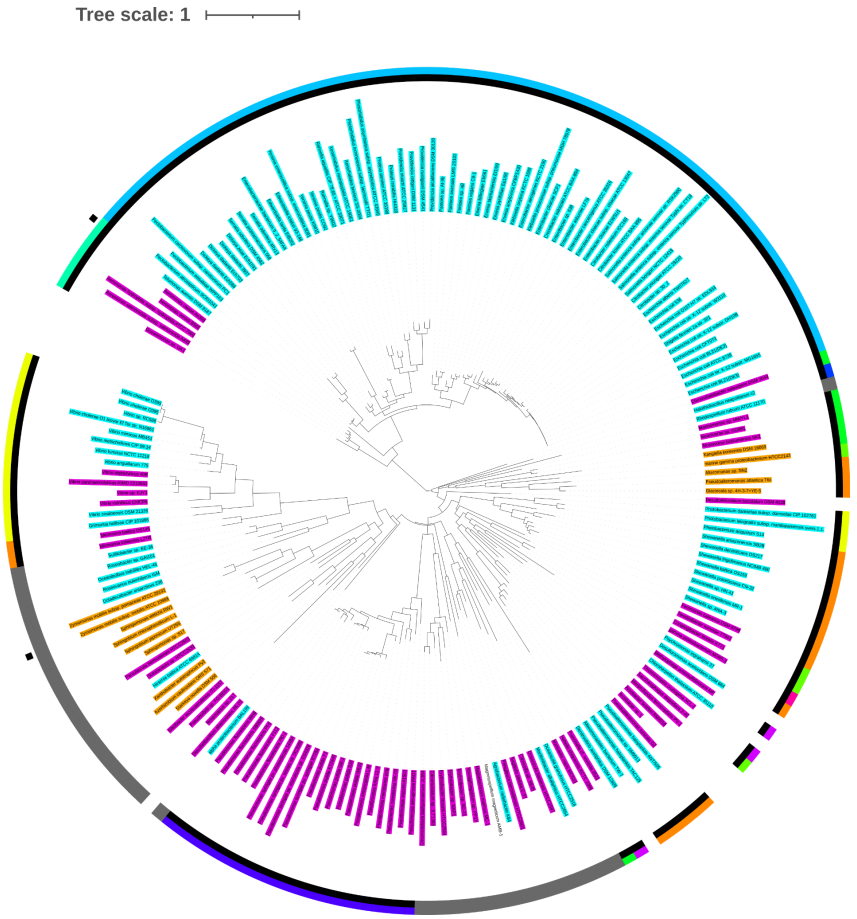


Рисунок 16. Детализация ветви дерева с Рисунка 13, которая включает большинство белков семейства CorA с неканоническими мотивами в селективном фильтре. Цвета листьев соответствуют сигнатурному мотиву соответствующего белка: GMN — оранжевый, GVN — бирюзовый, GIN — розовый, лист с другим мотивом не окрашен. Цветовой код колец: внутреннее кольцо — гамма-протеобактерии отмечены черным, альфа–протеобактерии серым, все остальные белым; внешнее кольцо — девять цветов соответствуют девяти отрядам гамма-протеобактерий, альфа-протеобактерии отмечены серым, все остальные отряды белым. Два черных квадрата — гены,

которые имеют мотивы связывания с цинковыми репрессорами в регуляторных областях.

Для проверки того, как функциональная специфичность распределяется по дереву и соответствует ли она сигнатурному мотиву, на дереве были также размечены известные из экспериментальных проверок результаты. Оказалось, что наличие мотива GMN не определяет однозначно магниевую специфичность, впрочем как и цинковую.

4.3.3 Предсказание регуляции

Ещё одним свидетельством в пользу той или иной специфичности белка является способ регуляции его гена. В случае магниевой специфичности в качестве регуляторного элемента рассматривался магниевый рибопереключатель Ykok-leader, а в случае цинковой регуляции целью поиска были сайты связывания цинкового репрессора ZUR или AdcR в случае *Streptococcus*, также исследуемые гены были проверены на регуляцию кобаламиновым рибопереключателем.

Потенциальные сайты связывания цинковых репрессоров были обнаружены у разных таксономических групп в небольшом количестве: два у альфа-протеобактерий, шесть у гамма-протеобактерий, шесть у *Streptococcus*, четыре у других фирмикут и ещё два у актинобактерий.

У девяти бактерий перед рассматриваемыми генами был обнаружен регуляторный элемент, предположительно являющийся магниевым рибопереключателем Ykok-leader. Рибопереключателей других типов перед рассматриваемыми генами найдено не было.

Исходя из расположения по дереву генов, контролируемых магниевыми рибопереключателями и цинковыми репрессорами, а также распределения по нему мотивов селективного фильтра, можно сделать вывод о том, что последовательность в мотиве GxN не является определяющей в вопросе специфичности белка. Более того, предположительно цинковые и магниевые

транспортёры не образуют на дереве монофилетических групп, что подтверждает идею, что изменение специфичности возможно за счёт малого числа мутаций без значительных изменений последовательности белка. Таким образом, встаёт вопрос предсказания позиций белка, которые в действительности определяют эту специфичность.

4.3.4 Определение позиций белка, отвечающих за специфичность

На основе предсказания специфичности исследуемых белков, последовательности были распределены по двум группам: магниевой и цинковой. На основе сравнения этих групп были определены позиции в последовательности, которые могут быть определяющими в вопросе специфичности белка. Эти позиции были размечены на известной структуре CorA 4EED (Рисунок 17).

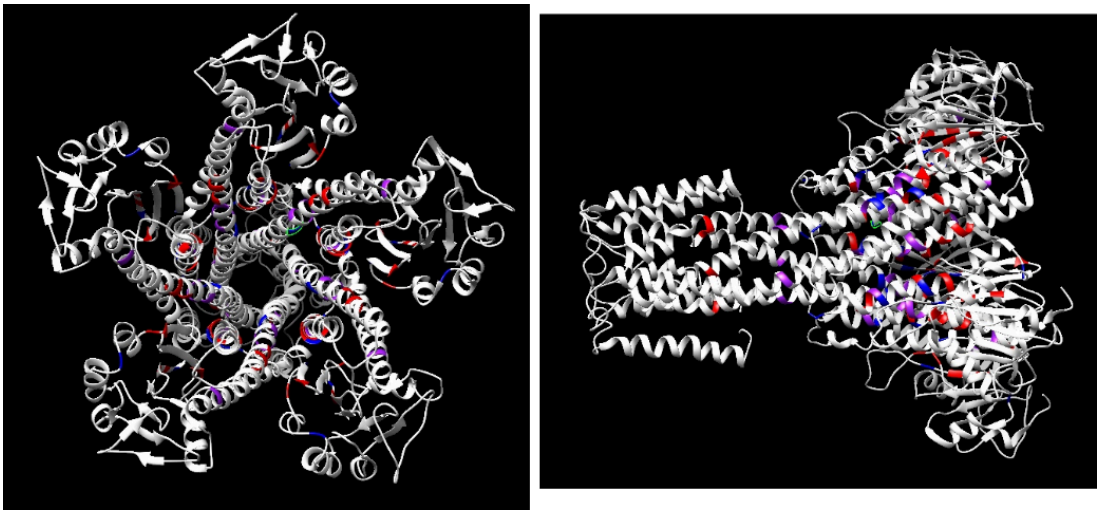


Рисунок 17. Структура пентамера CorA из *Thermotoga maritima* с разметкой предсказанных позиций, определяющих специфичность. Цвет разметки соответствует вероятности влияния позиции, наиболее вероятные позиции помечены красным цветом, далее идут синие и фиолетовые позиции.

Примечательно, что все позиции, потенциально определяющие специфичность, располагаются глубоко в цитоплазматической части белка, за исключением двух позиций на спиралях, образующих белковый канал, при этом позиция в последовательности, которая рассматривается как селективный фильтр, по-видимому не играет роли в специфичности.

4.4 Обсуждение

Для проверки этого предсказания нашими коллегами А. Стеценко, П. Стеханцевым и А. Гуськовым из университета Гронингена в Нидерландах была получена структура CmaX (ZntB) семейства CorA из *Pseudomonas aeruginosa* и в эксперименте определена его специфичность. Для того, чтобы охарактеризовать транспорт катионов через исследуемый белок, в эксперименте был проведён анализ флуоресцентного поглощения цинка через этот белок, вставленный в липосому. Кроме того, аналогичный анализ был проведён для других бивалентных катионов металлов, таких как кадмий, кобальт и никель. Кроме того, для анализа специфичности были измерены константы диссоциации комплекса исследуемого белка и катионов тех же металлов. Эта работа подтвердила, что несмотря на то, что этот белок имеет сигнатурный мотив GIN, он способен транспортировать не только цинк, но и кадмий, кобальт и никель на сопоставимом с другими белками семейства уровне [170]. В сочетании с данными о том, что большая часть белков с отличными от GMN мотивами располагается на одной ветви филогенетического дерева и внутри неё активно происходят горизонтальные переносы, такие результаты позволяют предположить, что если последовательность мотива GxN и оказывает влияние на специфичность, то незначительное.

Глава 5. Эволюция семейства белков анкириновых повторов у *Wolbachia*

5.1 Введение

Бактерии рода *Wolbachia* — облигатные внутриклеточные симбионты членистоногих и нематод, являющиеся самой распространённой группой внутриклеточных альфа-протеобактерий [171].

Считается, что вольбахии у нематод наследуются вертикально, а у вольбахий из членистоногих замечены множественные случае смены хозяина [172]. При этом способы взаимодействия вольбахий с хозяевами разнообразны и включают направление развития яиц по женскому типу и смещение доли самок в популяции, феминизацию и уничтожение самцов, цитоплазматическую несовместимость [171]. Некоторые вольбахии, заражающие членистоногих, обладают полным биотиновым опероном, что полезно для хозяина [173]. Вольбахии клопов и блох могут обеспечивать хозяев отсутствующими в их рационе витаминами группы B, необходимыми хозяевам для роста и размножения [173]. В то же время во многих видах членистоногих вольбахии негативно влияют на размножение хозяина. У членистоногих вольбахии, как правило, факультативные симбионты, у нематод же это облигатные симбионты, необходимые для эмбриогенеза, линьки и роста нематод [174]. Несмотря на то, что установлено множество способов взаимодействия вольбахий с хозяевами, строгого соответствия между фенотипом хозяина и геномными особенностями бактерий обнаружено не было [175].

На текущий момент весь род *Wolbachia* определяется как один вид, который дальше делится на супергруппы. Общепринятая классификация предполагает разделение всех штаммов вольбахий на 17 групп супергрупп: A-F, H-Q и S [172,176,177]. Большинство вольбахий, заражающих членистоногих, относятся к супергруппам A и B; к супергруппам C и D относятся бактерии, заражающие нематод. В группах E и F встречаются вольбахии как членистоногих, так и нематод. Вольбахии остальных

супергрупп заражают отдельные отряды членистоногих [178]. Несмотря на возможность заражения одних и тех же клеток разными штаммами вольбахий, супергруппы остаются филогенетически различными кладами [179]. Тем не менее, между супергруппами в некоторых генах, таких как ген поверхностного белка вольбахий *wsp*, происходит активная рекомбинация, поэтому до сих пор не ясно, является ли классификация по супергруппам показательной для всех геномов вольбахий в целом [179]. Кроме того, не было выявлено каких-то фенотипических особенностей штаммов, заражающих членистоногих, которые бы поддерживали их разделение на супергруппы [179].

Значительные различия в составе генов определяются даже среди штаммов одной супергруппы. В супергруппах C, D и F наблюдается больше приобретений и потерь генов, чем в супергруппах A и B [178]. При этом частоты мутаций в этих двух группах одинаковы, то есть в супергруппах C, D, F очень гибкие штаммы, в которых адаптация к условиям среды происходит за счёт активных геномных перестроек [180].

Геномы вольбахий относительно невелики. Размеры геномов у штаммов, заражающих членистоногих, составляют от 1.2 до 1.8 м.п.о., а для мутуалистов нематод диапазон размеров составляет от 0.96 до 1.1 м.п.о. [180]. Тем не менее, во многих геномах вольбахий высока доля повторяющихся элементов, таких как инсерционные последовательности (IS), интроны II типа, фрагменты профагов, а также мультигенных семейств, таких как гены анкириновых повторов (ANK). Такого рода элементы составляют до 10% генома и могут играть важную роль в их адаптации за счёт, например, влияния на экспрессию соседних генов, псевдогенизации, увеличения частоты геномных перестроек, а могут быть следствием ослабления отбора [11,19,181]. Анкириновый повтор — это белковый мотив из 30-34 аминокислот, который опосредует белок-белковые взаимодействия [11]. Как правило, белки с ANK-доменами встречаются у эукариот и вирусов, однако в

редких случаях встречаются у бактерий и архей [182]. ANK-белки участвуют в широком спектре взаимодействий, таких как регуляция транскрипции, регуляция клеточного цикла, передача сигналов и во многих других [183]. У бактерий ANK-белки образуют семейство бактериальных рецепторов IV типа и играют важную роль во взаимодействии с хозяином и развитии инфекции, служа, в том числе, как эффекторы [184,185]. У вольбахий ANK-гены представляют собой самое крупное семейство, склонное к паралогизации. Число ANK-генов может составлять до 5% генов в геноме, например, в штамме wMel их 23 из примерно 1100 генов, а в штамме wPip — 60 из 1300 [11].

Накопление повторов в геноме вызывает существенные изменения в количестве генов и порядке их следования в молодых внутриклеточных патогенах [91,186]. Большинство такого рода изменений, геномных перестроек, отрицательно влияют на приспособленность, однако перестройки, происходящие параллельно на разных ветвях филогенетического дерева в процессе адаптации к одной и той же экологической нише, могут быть результатом положительного отбора [91], [187,188]. Кроме того, геномные перестройки могут приводить к обратимому изменению фенотипа в патогенах и симбионтах с помощью фазовой вариации. Фазовая вариация влияет на множество бактериальных мультигенных семейств, белки которых участвуют во взаимодействии бактерии с хозяином [10,189]. В случае вольбахий фазовой вариации за счёт инверсии могут предположительно подвергаться ANK-гены.

Целью этой главы было установить связь между репертуаром ANK-генов и адаптацией вольбахий к хозяину, а также выявить генетические механизмы, влияющие на накопление ANK-генов в геномах.

5.2 Материалы и методы

5.2.1 Геномные данные

В работе рассмотрено 159 полных геномов *Wolbachia* spp., доступных в базе данных RefSeq [81] по состоянию на март 2023 года (Дополнительная таблица 9). Информация о хозяевах была получена из исходных метаданных сборок.

5.2.2 Базовый анализ геномов

Для аннотации геномов, формирования ортологических рядов и построения дерева видов использовался инструмент PanACoTA [84]. Для аннотации было использовано интегрированное в этот модуль программное обеспечение prokka [118] в сочетании с prodigal [119]. Построение ортогрупп было выполнено с помощью OrthoFinder [190]. Дерево штаммов было построено с помощью модуля iqtree2 [85].

5.2.3 Аннотация мобильных элементов

Для аннотации мобильных элементов использовалась программа ISFinder [191] как модуль prokka [118].

5.2.4 Предсказание и выравнивание ANK-генов

Для поиска генов, кодирующих представители семейства белков анкириновых повторов (ANK), сначала был проведён поиск доменов анкириновых повторов с помощью PfamScan.pl [192] по PFAM записям PF00023.33, PF12796.10, PF13606.9, PF13637.9 и PF13857.9, а затем был проведён дополнительный поиск с помощью BLASTP [82] для всех открытых рамок считывания, включая короткие, с порогом 90% на идентичность и покрытие. Множественное выравнивание ANK-белков было выполнено с использованием muscle с параметрами по умолчанию [166]. Кластеризация последовательностей ANK-белков была выполнена с использованием CD-hit с порогом 70% на сходство. [193].

5.3 Результаты

5.3.1 *Wolbachia* и хозяева

Был проведён анализ 159 полных геномов вольбахий из трёх отрядов насекомых (двукрылые — 64 генома, чешуекрылые — 47 геномов, перепончатокрылые — 19 геномов), нематод (10 геномов) и по одному геному из коллембол и паукообразных. На дереве вольбахии из членистоногих распределены в несколько крупных клад, в то время как вольбахии из нематод образуют монофилетическую группу, в которую, однако, входит несколько штаммов из членистоногих (Рисунок 18).

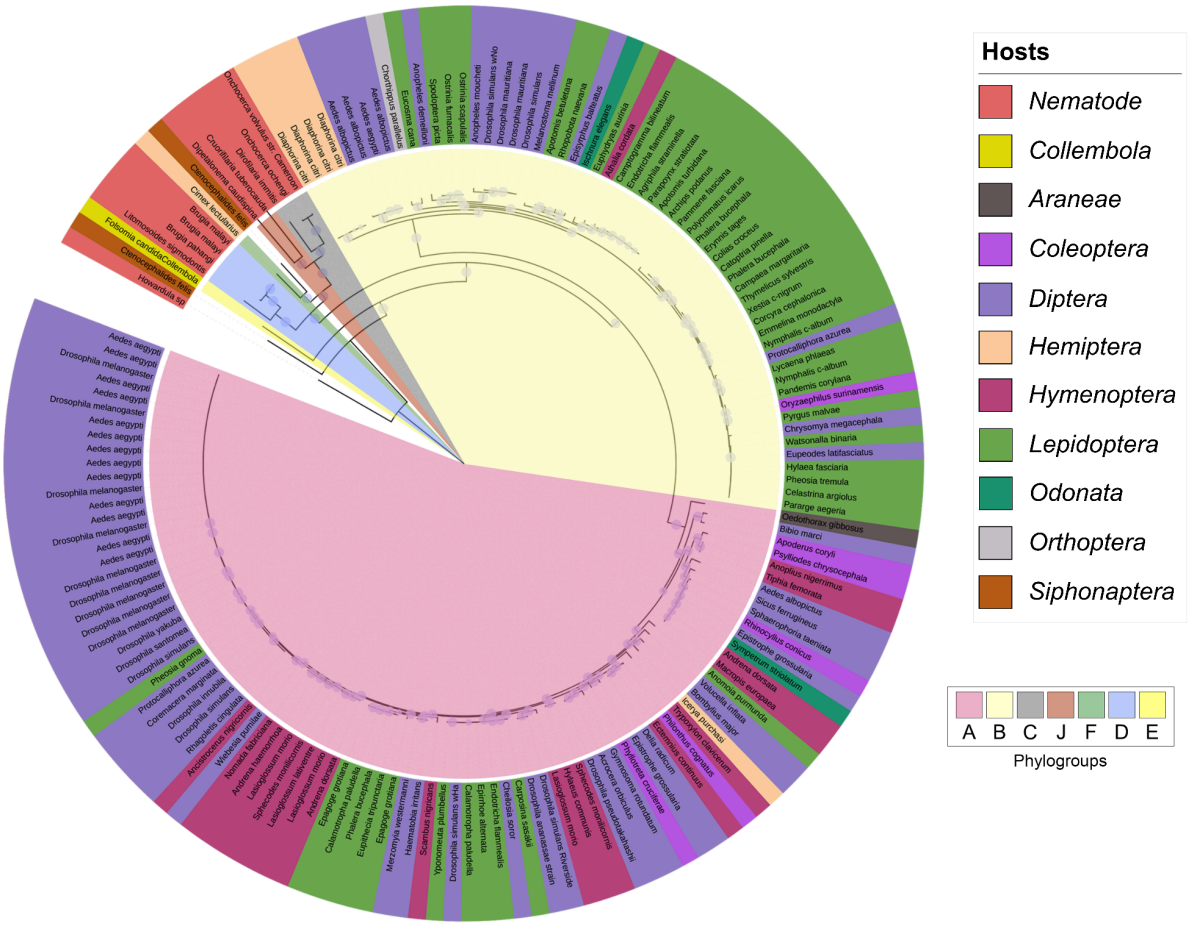


Рисунок 18. Филогенетическое дерево штаммов *Wolbachia*. Цвета листьев обозначают хозяина, цвета секторов — принадлежность к филогруппе.

5.3.2 Разнообразие генов ANK-белков и мобильных элементов

Вольбахии, заражающие членистоногих, содержат инсерционные

последовательности из 13 семейств, которые покрывают до 18% генома, а также до 97 генов ANK. Содержание ANK-генов и инсерционных последовательностей в геномах вольбахий, заражающих нематод, ниже (Рисунок 19).

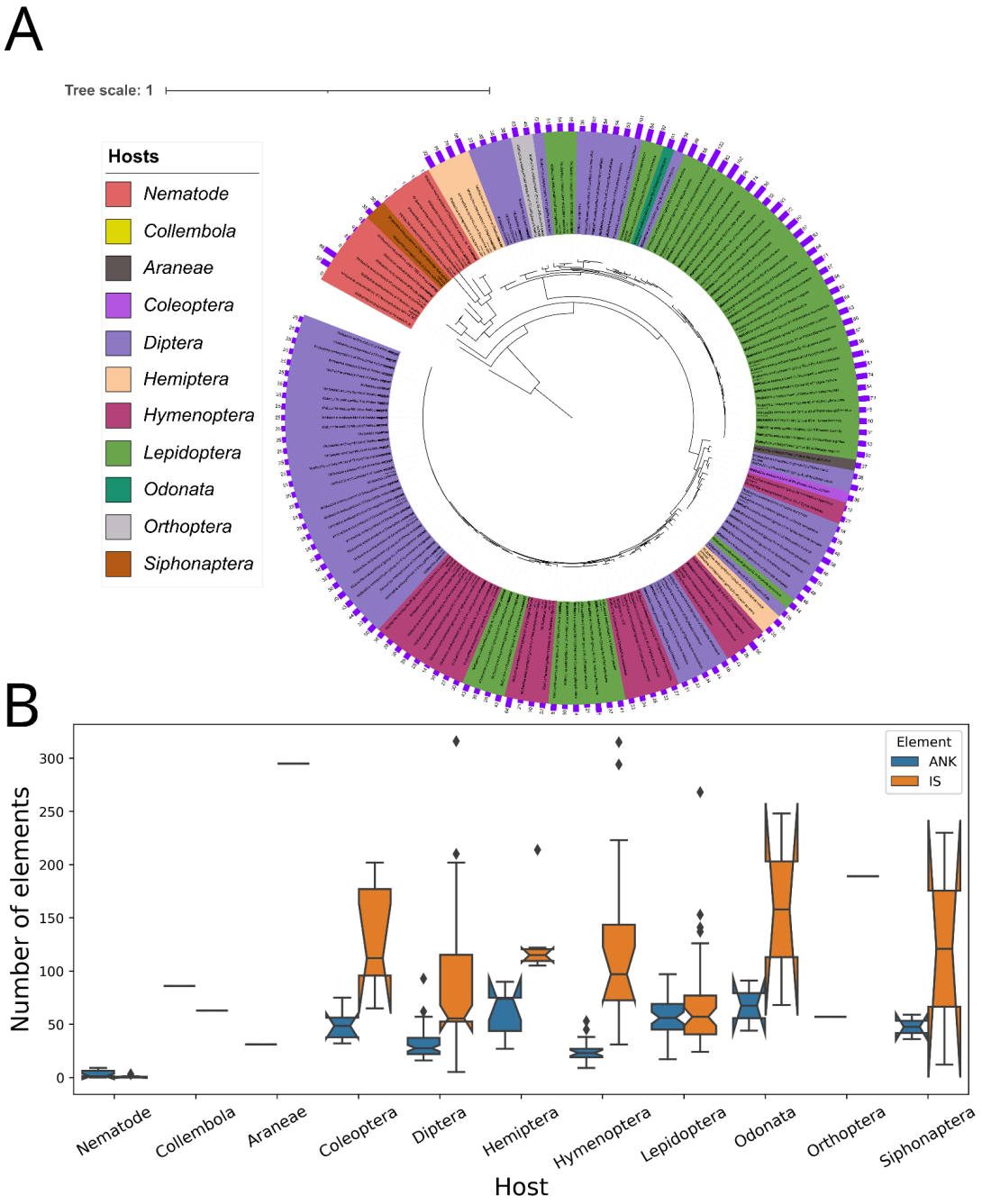


Рисунок 19. A) Филогенетическое дерево штаммов *Wolbachia*. Цвета листьев соответствуют отрядам хозяев, высота столбца внешнего кольца — числу ANK-генов в геноме. B) Диаграмма размаха числа ANK-генов и инсерционных последовательностей.

Примечательно, что при этом штаммы из нематод имеют меньший размер генома (0.9-1.1 м.п.о.), чем штаммы из членистоногих (1.2-1.8 м.п.о.), даже те из последних, которые находятся на дереве на ветви нематод. Более того, была обнаружена сильная корреляция между числом ANK-генов в геноме вольбахии с его размером (коэффициент корреляции Пирсона R=0,78, $p=3{,}2\times10^{-33}$), а для числа мобильных элементов такой корреляции не обнаружено (Рисунок 20). В то же время транспозоны наблюдались по соседству с ANK-генами статистически чаще, чем в среднем по геному. Нулевая гипотеза в этой проверке предполагала, что транспозоны среди пар ближийших к ANK-генам генов встречаются так же часто, как и для остальных генов. 95% доверительные интервалы для средних значений составили (0,15; 0,17) и (0,05; 0,06) соответственно, тест Вилкоксона с двусторонней альтернативной гипотезой дал $p=1.21\times10^{-26}$. Таким образом, мобильные элементы могут способствовать амплификации ANK-генов.

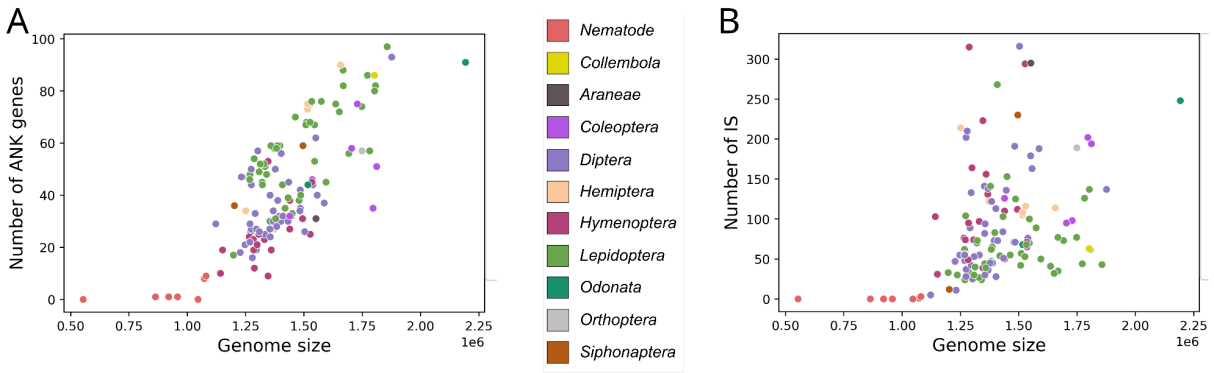


Рисунок 20. Взаимосвязь между количеством A) ANK-генов и B) инсерционных последовательностей в геноме (вертикальные оси) и размером генома (горизонтальная ось).

Сами по себе ANK-гены крайне разнообразны и образуют 624 ортогруппы, причём длины ANK-белков сильно разнятся в диапазоне от 102 до 4793 аминокислот. Ни одна из ортогрупп ANK-генов не имеет представителей во всех рассмотренных геномах, при этом в самую крупную из них входят гены 96% штаммов *Wolbachia*, заражающих членистоногих. Кроме того, около половины ANK-генов — это синглтоны, некоторые из

которых являются фрагментами более длинных ANK-генов, разрушенных мобильными элементами, так как эти фрагменты с двух сторон от мобильного элемента выравниваются на ANK-гены из близких штаммов. Однако прямая кластеризация последовательностей генов чувствительна к изменению длины гена или перестановке его фрагментов. Чтобы избежать ошибок такого рода, было также учтено, с какими генами ко-локализуются ANK-гены в геномах *Wolbachia*.

5.3.3 Связь ANK-генов, мобильных элементов и геномного контекста

Для анализа устойчивости геномного контекста вокруг ANK-генов было произведено сравнение положений ANK-генов относительно локально коллинеарных блоков (ЛКБ) генома, то есть таких фрагментов генома, которые не подвержены геномным перестройкам в исследуемом наборе штаммов, предоставленных Екатериной Востоковой. Это позволило определить, что 98% мобильных элементов и 93,5% ANK-генов были расположены частично или полностью за пределами ЛКБ (Рисунок 21A).

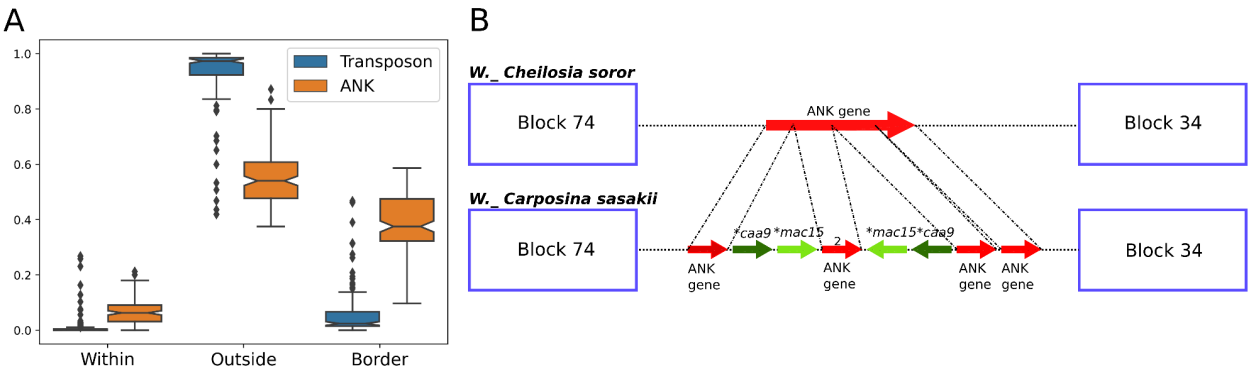


Рисунок 21. A) Положение ANK-генов и мобильных элементов относительно ЛКБ: внутри, вне или частично вне. Все попарные сравнения статистически значимы. B) Пример разрушения ANK-гена инвертированными повторами транспозаз, приводящими к образованию четырех коротких ОРС у *Wolbachia* из *Carposina sasakii*. Цвета стрелок обозначают гены из разных ортогрупп.

Чтобы выявить влияние инсерционных последовательностей на доменную структуру кодируемых белков, были выравнены ANK-гены из гомологичных локусов, определенных как области между двумя

консервативными ЛКБ у близкородственных вольбахий. И действительно, наблюдались многочисленные случаи интеграции мобильных элементов в ANK-гены, влияющие на последовательности и предполагаемые продукты последних. Примером такой интеграции является разрушение ANK-гена длиной ~1 т.п.о., расположенного в локусе между генами, кодирующими глутамил-тРНК(Gln)-амидотрансферазу GatB и цистеин-тРНК-лигазу, транспозазами, наблюдаемыми у вольбахии-эндосимбионта *Carposina sasakii* (Рисунок 21B). Это привело к образованию четырех коротких открытых рамок считывания с ANK-генами длиной 171 п.о., 252 п.о., 240 п.о. и 351 п.о..

Генов, предположительно подвергающихся фазовой вариации, производящей обмен двух функциональных фрагментов генов, в геномах вольбахий не обнаружено.

5.4 Обсуждение

Вольбахии были впервые описаны век назад, однако геномные факторы адаптации бактерий к хозяину до сих пор остаются неизвестными. Одной из таких адаптаций может служить в том числе сочетание ANK-генов и мобильных элементов, активно распространяющихся по геномам [11].

Переход к внутриклеточному образу жизни часто сопровождается накоплением мобильных элементов, вызывающих существенные изменения в составе генома и порядке расположения генов из-за более слабого давления отбора [19,91]. Действительно, такого рода нестабильность у вольбахий ассоциирована с мобильными элементами на границах перестроек. Мобильные элементы составляют у этих бактерий до 18% генома и часто встречаются на границах ЛКБ как и ANK-гены. Более того, мы обнаружили сильную корреляцию между количеством ANK-генов и размером генома, а также значительную перепредставленность транспозонов рядом с ANK-генами, что может указывать на амплификацию ANK-генов, стимулируемую мобильными элементами.

Ранее исследование отдельных штаммов супергрупп A и B показало,

что ANK-гены вариабельны по последовательности и подвержены влиянию внутригеномной рекомбинации и горизонтального переноса между разными штаммами вольбахий [11]. Последовательности ANK-генов изменяются не только за счёт мутаций, но и за счёт рекомбинации между разными ANK-генами, а также потери или приобретения фрагментов ANK-генов, так как в них присутствуют прямые повторы [11]. В настоящем исследовании подтвердилось, что состав ANK-генов даже в близкородственных штаммах крайне разнообразен, чему способствует интеграция в ANK-гены мобильных элементов, а также по-видимому, события гомологичной рекомбинации, что значительно расширяет репертуар ANK-белков у вольбахий.

Заключение

Разнообразные стратегии адаптации бактерий формируются за счет широкого круга молекулярных механизмов. Сравнительно-геномный анализ белковых семейств позволяет получать новые знания о функциональных особенностях этих белков и их вовлечённости в метаболизм бактерий. Такие знания могут быть впоследствии использованы для разработки новых стратегий лечения, а также применяться в биотехнологической промышленности, использующей генно-модифицированные штаммы бактерий.

Например, из наблюдения про возможность обмена эффекторами IpaH между патогенами из разных хозяев, можно сделать вывод о потенциальной опасности новых зоонозов для общественного здравоохранения. Большее количество полных сборок бактерий кишечной микробиоты разных животных позволит уточнить набор IpaH у бактерий из животных и установить возможные последствия заражения человека этими бактериями или влияние горизонтального переноса этих генов в шигелл на протекание болезни у человека.

В случае белков гистидиновых триад было показано, что хотя у *Streptococcus pneumoniae* гены двух из них подвергаются фазовой вариации, остальные варианты белков могут рассматриваться в качестве мишени для разработки вакцин, особенно в случае других патогенных стрептококков.

В свою очередь понимание субстратной специфичности транспортёров металлов семейства CorA, в силу практически повсеместного распространения этой системы транспорта у бактерий, является важной задачей из-за малой изученности этих систем.

Демонстрация связи между распространённостью генов, кодирующих белки анкириновых повторов, и мобильными элементами у вольбахий позволяет предположить, что именно мобильные элементы являются драйверами эволюции этих генов.

Анализ четырёх семейств белков в рамках этой работы демонстрирует,

что различия в специализациях отдельных семейств, а также разнообразие механизмов их эволюции не позволяет применять стандартные биоинформатические схемы. Поэтому задача описания эволюции белковых семейств требует понимания конкретных научных задач, построения гипотез и разработки специализированных подходов.

Выводы

1. Показано, что семейство эффекторов IpaH, являющееся характеристической особенностью шигелл и энтероинвазивных кишечных палочек, состоит из девяти классов эффекторов, имеющих общий С-концевой домен, отвечающий за убиквитин-лигазную активность, и отличающихся N-концевым доменом, распознающим белок-мишень. В одном из классов происходит расхождение паралогов на две группы, что может привести к формированию двух новых классов эффекторов. Белки этого семейства также были обнаружены у патогенов крыс, сурков и овец, причём у бактерий этих хозяев набор эффекторов отличается от эффекторов патогенов человека.
2. Установлено, что, хотя белки гистидиновых триад у *Streptococcus* spp. крайне разнообразны, структура дерева белков соответствует вертикальному наследованию генов этих белков, а не горизонтальным переносам извне. Большинство генов этих белков контролируются цинковыми репрессорами, однако присутствуют две группы без такой регуляции. Гены белков одной из этих групп контролируются медными репрессорами, у белков другой группы не было обнаружено признаков общей для ветви регуляции. Фазовой вариации подвергаются только гены двух типов белков гистидиновых триад из *S. pneumoniae*, но не других *Streptococcus* spp..
3. Роль характеристической последовательности GxN семейства белков CorA в определении специфичности транспортёра была ранее переоценена. Показано, что белки с отличными от канонической последовательностями в этом мотиве располагаются на филогенетическом дереве в основном на одной ветви, в рамках этой ветви подвергаются горизонтальным переносам и способны к транспорту тех же катионов, что и белки с каноническим мотивом. Если последовательность GxN и оказывает влияние на специфичность, то слабое.

4. Показано, что число копий генов, кодирующих белки с ANK-повторами значимо коррелирует с размером генома *Wolbachia* spp., в отличие от числа мобильных элементов, для которых такой корреляции не наблюдается. Среди соседей этих генов на хромосоме мобильные элементы присутствуют статистически чаще, чем в среднем по геному. Мобильные элементы могут являться драйверами эволюции генов ANK-белков.

Благодарности

Я признательна своему научному руководителю Михаилу Сергеевичу Гельфанду, который все эти годы направлял и поддерживал меня на научном пути. Отдельно хочу поблагодарить Ольгу Бочкарёву за всестороннюю помощь в процессе работы и моральную поддержку. Кроме того, я благодарна своим соавторам Марии Тутукиной, Альберту Гуськову, Артёму Стеценко. Я благодарна за помощь и поддержку при написании диссертации также своей семье и друзьям, в особенности Маргарите Барановой, Атаю Добрынину и Елизавете Григорашвили.

Список литературы

1. Orengo CA, Thornton JM. PROTEIN FAMILIES AND THEIR EVOLUTION—A STRUCTURAL PERSPECTIVE. Annu Rev Biochem. 2005;74: 867–900.

2. Kafri R, Levy M, Pilpel Y. The regulatory utilization of genetic redundancy through responsive backup circuits. Proceedings of the National Academy of Sciences. 2006;103: 11653–11658.

3. Poudel S, Cope AL, O'Dell KB, Guss AM, Seo H, Trinh CT, et al. Identification and characterization of proteins of unknown function (PUFs) in Clostridium thermocellum DSM 1313 strains as potential genetic engineering targets. Biotechnol Biofuels. 2021;14: 1–19.

4. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. Elife. 2022;11. doi:10.7554/eLife.67667

5. Rodríguez del Río Á, Giner-Lamia J, Cantalapiedra CP, Botas J, Deng Z, Hernández-Plaza A, et al. Functional and evolutionary significance of unknown genes from uncultivated taxa. Nature. 2023;626: 377–384.

6. Singer AU, Rohde JR, Lam R, Skarina T, Kagan O, Dileo R, et al. Structure of the Shigella T3SS effector IpaH defines a new class of E3 ubiquitin ligases. Nat Struct Mol Biol. 2008;15. doi:10.1038/nsmb.1511

7. Mattock E, Blocker AJ. How Do the Virulence Factors of Shigella Work Together to Cause Disease? Front Cell Infect Microbiol. 2017;7. doi:10.3389/fcimb.2017.00064

8. Kapoor N, Ndungo E, Pill L, Desalegn G, Berges A, Oaks EV, et al. Efficient production of immunologically active Shigella invasion plasmid antigens IpaB and IpaH using a cell-free expression system. Appl Microbiol Biotechnol. 2022;106: 401.

9. Aceil J, Avci FY. Pneumococcal Surface Proteins as Virulence Factors, Immunogens, and Conserved Vaccine Targets. Front Cell Infect Microbiol. 2022;12: 832254.

10. Shelyakin PV, Bochkareva OO, Karan AA, Gelfand MS. Micro-evolution of three Streptococcus species: selection, antigenic variation, and horizontal gene inflow. BMC Evol Biol. 2019;19. doi:10.1186/s12862-019-1403-6

11. Siozios S, Ioannidis P, Klasson L, Andersson SGE, Braig HR, Bourtzis K.

The Diversity and Evolution of Wolbachia Ankyrin Repeat Domain Genes. PLoS One. 2013;8. doi:10.1371/journal.pone.0055390

12. Rice DW, Sheehan KB, Newton ILG. Large-Scale Identification of Wolbachia pipientis Effectors. Genome Biol Evol. 2017;9: 1925–1937.

13. Duron O, Boureux A, Echaubard P, Berthomieu A, Berticat C, Fort P, et al. Variability and Expression of Ankyrin Domain Genes in Wolbachia Variants Infecting the Mosquito Culex pipiens. J Bacteriol. 2007;189: 4442.

14. Stetsenko A, Guskov A. Cation permeability in CorA family of proteins. Sci Rep. 2020;10. doi:10.1038/s41598-020-57869-z

15. Fh C. On protein synthesis. Symposia of the Society for Experimental Biology. 1958;12. Available: https://pubmed.ncbi.nlm.nih.gov/13580867/

16. Zhang J, Yang J-R. Determinants of the rate of protein sequence evolution. Nat Rev Genet. 2015;16: 409–420.

17. Zuckerkandl E. On the molecular evolutionary clock. Journal of molecular evolution. 1987;26. doi:10.1007/BF02111280

18. Kimura M. The Number of Heterozygous Nucleotide Sites Maintained in a Finite Population Due to Steady Flux of Mutations. Genetics. 1969;61: 893.

19. Mamirova L, Popadin K, Gelfand MS. Purifying selection in mitochondria, free-living and obligate intracellular proteobacteria. BMC Evol Biol. 2007;7. doi:10.1186/1471-2148-7-17

20. Jayaraman V, Toledo‐Patiño S, Noda‐García L, Laurino P. Mechanisms of protein evolution. Protein Sci. 2022;31. doi:10.1002/pro.4362

21. Liu J, Zhang Y, Lei X, Zhang Z. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol. 2008;9: R69.

22. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. The origin, evolution and structure of the protein world. Biochem J. 2009;417. doi:10.1042/BJ20082063

23. Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. BMC Genomics. 2012;13: 1–10.

24. Rost B. Twilight zone of protein sequence alignments. Protein Eng. 1999;12. doi:10.1093/protein/12.2.85

25. Casadio R, Bartoli L, Fariselli P, Tasco G, Martelli PL. Protein structure prediction in the genomic era: Annotation-facilitated remote homology detection. Medicinal Chemistry in Drug Discovery Design, Synthesis and Screening. IND; 2014. pp. 197–218.

26. Khan SJ. Global Conformational Change Induced By Single Amino Acid Residue of Photoactive Yellow Protein in Time Domain. Biophys J. 2010;98: 26a.

27. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. A minimal sequence code for switching protein structure and function. Proc Natl Acad Sci U S A. 2009;106. doi:10.1073/pnas.0906408106

28. Davis BH, Poon AFY, Whitlock MC. Compensatory mutations are repeatable and clustered within proteins. Proceedings of the Royal Society B: Biological Sciences. 2009;276: 1823.

29. Szamecz B, Boross G, Kalapis D, Kovács K, Fekete G, Farkas Z, et al. The Genomic Landscape of Compensatory Evolution. PLoS Biol. 2014;12: e1001935.

30. EMBL-EBI. What are protein families? [cited 29 Aug 2024]. Available: https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/what-are-protein-families/

31. Illergård K, Ardell DH, Elofsson A. Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. Proteins. 2009;77. doi:10.1002/prot.22458

32. Koonin EV. Orthologs, Paralogs, and Evolutionary Genomics1. Annu Rev Genet. 2005;39: 309–338.

33. Birchler JA, Yang H. The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. Plant Cell. 2022;34: 2466.

34. Espinosa-Cantú A, Cruz-Bonilla E, Noda-Garcia L, DeLuna A. Multiple Forms of Multifunctional Proteins in Health and Disease. Front Cell Dev Biol. 2020;8: 517062.

35. Atkins WM. Biological messiness vs. biological genius: Mechanistic aspects and roles of protein promiscuity. J Steroid Biochem Mol Biol. 2015;151: 3.

36. Pushker R, Mira A, Rodríguez-Valera F. Comparative genomics of gene-family size in closely related bacteria. Genome Biol. 2004;5.

doi:10.1186/gb-2004-5-4-r27

37. Copley SD. Evolution of new enzymes by gene duplication and divergence. FEBS J. 2020;287: 1262–1283.

38. Lee J-W. Bacterial Regulatory Mechanisms for the Control of Cellular Processes: Simple Organisms' Complex Regulation. J Microbiol. 2023;61: 273–276.

39. Kim T, Kim T-K. Regulatory RNA: from molecular insights to therapeutic frontiers. Exp Mol Med. 2024;56: 1233–1234.

40. Phillips ZN, Tram G, Seib KL, Atack JM. Phase-variable bacterial loci: how bacteria gamble to maximise fitness in changing environments. Biochem Soc Trans. 2019;47. doi:10.1042/BST20180633

41. Åberg A, Gideonsson P, Vallström A, Olofsson A, Öhman C, Rakhimova L, et al. A Repetitive DNA Element Regulates Expression of the Helicobacter pylori Sialic Acid Binding Adhesin by a Rheostat-like Mechanism. PLoS Pathogens. 2014;10: e1004234.

42. Thorne JL. Models of protein sequence evolution and their applications. Curr Opin Genet Dev. 2000;10. doi:10.1016/s0959-437x(00)00142-8

43. Chowdhury B, Garai G. A review on multiple sequence alignment from the perspective of genetic algorithm. Genomics. 2017;109. doi:10.1016/j.ygeno.2017.06.007

44. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48. doi:10.1016/0022-2836(70)90057-4

45. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22: 4673.

46. Dayhoff MO. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. 1972;5: 89–99.

47. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89. doi:10.1073/pnas.89.22.10915

48. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020;21: 428–444.

49. Munjal G, Hanmandlu M, Srivastava S. Phylogenetics Algorithms and Applications. Ambient Communications and Computer Systems. 2019;904: 187.

50. Gabaldón T. Evolution of proteins and proteomes: a phylogenetics approach. Evol Bioinform Online. 2005;1: 51.

51. A Statistical Method for Evaluating Systematic Relationships. In: Google Books [Internet]. [cited 29 Aug 2024]. Available: https://books.google.com/books/about/A_Statistical_Method_for_Evaluating_Syst.html?hl=ru&id=o1BlHAAACAAJ

52. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4. doi:10.1093/oxfordjournals.molbev.a040454

53. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol. 1996;266. doi:10.1016/s0076-6879(96)66026-1

54. Brower AVZ. Statistical consistency and phylogenetic inference: a brief review. Cladistics. 2018;34: 562–567.

55. Kufareva I, Abagyan R. Methods of protein structure comparison. Methods Mol Biol. 2012;857: 231.

56. Khalil IA, Troeger C, Blacker BF, Rao PC, Brown A, Atherly DE, et al. Morbidity and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990–2016. Lancet Infect Dis. 2018;18: 1229.

57. Lampel KA, Formal SB, Maurelli AT. A Brief History of Shigella. EcoSal Plus. 2018;8. doi:10.1128/ecosalplus.ESP-0006-2017

58. Ranjbar R, Farahani A. Shigella: Antibiotic-Resistance Mechanisms And New Horizons For Treatment. Infect Drug Resist. 2019;12: 3137.

59. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, et al. The Intriguing Evolutionary Journey of Enteroinvasive E. coli (EIEC) toward Pathogenicity. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.02390

60. Hawkey J, Monk JM, Billman-Jacobe H, Palsson B, Holt KE. Impact of insertion sequences on convergent evolution of Shigella species. PLoS Genet. 2020;16. doi:10.1371/journal.pgen.1008931

61. Prosseda G, Di Martino ML, Campilongo R, Fioravanti R, Micheli G, Casalino M, et al. Shedding of genes that interfere with the pathogenic lifestyle: the Shigella model. Res Microbiol. 2012;163. doi:10.1016/j.resmic.2012.07.004

62. Feng Y, Chen Z, Liu S-L. Gene Decay in Shigella as an Incipient Stage of Host-Adaptation. PLoS One. 2011;6. doi:10.1371/journal.pone.0027754

63. Gómez-Valero L, Rocha EPC, Latorre A, Silva FJ. Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. Genome Res. 2007;17: 1178–1185.

64. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. Proc Natl Acad Sci U S A. 2006;103: 425–430.

65. van den Beld MJ, Reubsaet FA. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli. Eur J Clin Microbiol Infect Dis. 2012;31. doi:10.1007/s10096-011-1395-7

66. Hsu BM, Wu SF, Huang SW, Tseng YJ, Ji DD, Chen JS, et al. Differentiation and identification of Shigella spp. and enteroinvasive Escherichia coli in environmental waters by a molecular method and biochemical test. Water Res. 2010;44. doi:10.1016/j.watres.2009.10.004

67. Schroeder GN, Hilbi H. Molecular Pathogenesis of Shigella spp.: Controlling Host Cell Signaling, Invasion, and Death by Type III Secretion. Clin Microbiol Rev. 2008;21: 134.

68. Wagner S, Grin I, Malmsheimer S, Singh N, Torres-Vargas CE, Westerhausen S. Bacterial type III secretion systems: a complex device for the delivery of bacterial effector proteins into eukaryotic host cells. FEMS Microbiol Lett. 2018;365. doi:10.1093/femsle/fny201

69. Perrett CA, Lin DY-W, Zhou D. Interactions of Bacterial Proteins with Host Eukaryotic Ubiquitin Pathways. Front Microbiol. 2011;2. doi:10.3389/fmicb.2011.00143

70. Keszei AFA, Sicheri F. Mechanism of catalysis, E2 recognition, and autoinhibition for the IpaH family of bacterial E3 ubiquitin ligases. Proc Natl Acad Sci U S A. 2017;114: 1311.

71. Maculins T, Fiskin E, Bhogaraju S, Dikic I. Bacteria-host relationship: ubiquitin ligases as weapons of invasion. Cell Res. 2016;26: 499.

72. Kane CD, Schuch R, Day WA Jr, Maurelli AT. MxiE Regulates Intracellular Expression of Factors Secreted by the Shigella flexneri 2a Type III Secretion System. J Bacteriol. 2002;184: 4409.

73. Dorman MJ, Dorman CJ. Regulatory Hierarchies Controlling Virulence Gene Expression in Shigella flexneri and Vibrio cholerae. Front Microbiol. 2018;9. doi:10.3389/fmicb.2018.02686

74. Landick R, Wade JT, Grainger DC. H-NS and RNA polymerase: a love-hate relationship? Curr Opin Microbiol. 2015;24. doi:10.1016/j.mib.2015.01.009

75. Grainger DC. Structure and function of bacterial H-NS protein. Biochem Soc Trans. 2016;44. doi:10.1042/BST20160190

76. Dorman CJ. H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. Plasmid. 2014;75. doi:10.1016/j.plasmid.2014.06.004

77. Shi R, Yang X, Chen L, Chang H-T, Liu H-Y, Zhao J, et al. Pathogenicity of Shigella in Chickens. PLoS One. 2014;9. doi:10.1371/journal.pone.0100264

78. Elkenany R, Eltaysh R, Elsayed M, Abdel-Daim M, Shata R. Characterization of multi-resistant Shigella species isolated from raw cow milk and milk products. J Vet Med Sci. 2022;84: 890.

79. Liu S, Feng J, Pu J, Xu X, Lu S, Yang J, et al. Genomic and molecular characterisation of Escherichia marmotae from wild rodents in Qinghai-Tibet plateau as a potential pathogen. Sci Rep. 2019;9. doi:10.1038/s41598-019-46831-3

80. Zhu Z, Wang W, Cao M, Zhu Q, Ma T, Zhang Y, et al. Virulence factors and molecular characteristics of Shigella flexneri isolated from calves with diarrhea. BMC Microbiol. 2021;21. doi:10.1186/s12866-021-02277-0

81. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, et al. GenBank. Nucleic Acids Res. 2013;41: D36.

82. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10. doi:10.1186/1471-2105-10-421

83. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150.

84. Perrin A, Rocha EPC. PanACoTA: a modular tool for massive microbial

comparative genomics. NAR Genomics and Bioinformatics. 2021;3. doi:10.1093/nargab/lqaa106

85. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37: 1530.

86. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44: W242.

87. Magis C, Taly JF, Bussotti G, Chang JM, Di Tommaso P, Erb I, et al. T-Coffee: Tree-based consistency objective function for alignment evaluation. Methods Mol Biol. 2014;1079. doi:10.1007/978-1-62703-646-7_7

88. Shavkunov KS, Masulis IS, Tutukina MN, Deev AA, Ozoline ON. Gains and unexpected lessons from genome-scale promoter mapping. Nucleic Acids Res. 2009;37: 4919.

89. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46: W296.

90. Meng EC, Pettersen EF, Couch GS, Huang CC, Ferrin TE. Tools for integrated sequence-structure analysis with UCSF Chimera. BMC Bioinformatics. 2006;7: 1–10.

91. Seferbekova Z, Zabelkin A, Yakovleva Y, Afasizhev R, Dranenko NO, Alexeev N, et al. High Rates of Genome Rearrangements and Pathogenicity of Shigella spp. Front Microbiol. 2021;12. doi:10.3389/fmicb.2021.628622

92. Sansonetti PJ, Kopecko DJ, Formal SB. Shigella sonnei plasmids: evidence that a large plasmid is necessary for virulence. Infect Immun. 1981;34: 75.

93. Bongrand C, Sansonetti PJ, Parsot C. Characterization of the Promoter, MxiE Box and 5′ UTR of Genes Controlled by the Activity of the Type III Secretion Apparatus in Shigella flexneri. PLoS One. 2012;7. doi:10.1371/journal.pone.0032862

94. Ashida H, Toyotome T, Nagai T, Sasakawa C. Shigella chromosomal IpaH proteins are secreted via the type III secretion system and act as effectors. Mol Microbiol. 2007;63. doi:10.1111/j.1365-2958.2006.05547.x

95. Wu Y, Lau HK, Lee T, Lau DK, Payne J. In Silico Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of Shigella Identification.

Appl Environ Microbiol. 2019;85. doi:10.1128/AEM.00165-19

96. Ashida H, Sasakawa C. Shigella IpaH Family Effectors as a Versatile Model for Studying Pathogenic Bacteria. Front Cell Infect Microbiol. 2015;5. doi:10.3389/fcimb.2015.00100

97. Ye Y, Xiong Y, Huang H. Substrate-binding destabilizes the hydrophobic cluster to relieve the autoinhibition of bacterial ubiquitin ligase IpaH9.8. Communications Biology. 2020;3. doi:10.1038/s42003-020-01492-1

98. Aranda KRS, Fagundes-Neto U, Scaletsky ICA. Evaluation of Multiplex PCRs for Diagnosis of Infection with Diarrheagenic Escherichia coli and Shigella spp. J Clin Microbiol. 2004;42: 5849.

99. Patel S, Gupta RS. Robust demarcation of fourteen different species groups within the genus Streptococcus based on genome-based phylogenies and molecular signatures. Infect Genet Evol. 2018;66. doi:10.1016/j.meegid.2018.09.020

100. Krzyściak W, Pluskwa KK, Jurczak A, Kościelniak D. The pathogenicity of the Streptococcus genus. Eur J Clin Microbiol Infect Dis. 2013;32: 1361.

101. Kanwal S, Vaitla P. Streptococcus Pyogenes. StatPearls [Internet]. StatPearls Publishing; 2023.

102. Facklam R. What Happened to the Streptococci: Overview of Taxonomic and Nomenclature Changes. Clin Microbiol Rev. 2002;15: 613.

103. Shao Z-Q, Zhang Y-M, Pan X-Z, Wang B, Chen J-Q. Insight into the Evolution of the Histidine Triad Protein (HTP) Family in Streptococcus. PLoS One. 2013;8: e60116.

104. Weiser JN, Ferreira DM, Paton JC. Streptococcus pneumoniae: transmission, colonization and invasion. Nat Rev Microbiol. 2018;16: 355.

105. Plumptre CD, Ogunniyi AD, Paton JC. Polyhistidine triad proteins of pathogenic streptococci. Trends Microbiol. 2012;20. doi:10.1016/j.tim.2012.06.004

106. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, et al. Use of a Whole Genome Approach To Identify Vaccine Molecules Affording Protection against Streptococcus pneumoniae Infection. Infect Immun. 2001;69: 1593.

107. Adamou JE, Heinrichs JH, Erwin AL, Walsh W, Gayle T, Dormitzer M, et

al. Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis. Infect Immun. 2001;69. doi:10.1128/IAI.69.2.949-958.2001

108. Maruvada R, Prasadarao NV, Rubens CE. Acquisition of factor H by a novel surface protein on group B Streptococcus promotes complement degradation. The FASEB Journal. 2009;23: 3967.

109. Kunitomo E, Terao Y, Okamoto S, Rikimaru T, Hamada S, Kawabata S. Molecular and biological characterization of histidine triad protein in group A streptococci. Microbes Infect. 2008;10. doi:10.1016/j.micinf.2008.01.003

110. Waldemarsson J, Areschoug T, Lindahl G, Johnsson E. The Streptococcal Blr and Slr Proteins Define a Family of Surface Proteins with Leucine-Rich Repeats: Camouflaging by Other Surface Structures. J Bacteriol. 2006;188: 378.

111. Shao Z, Pan X, Li X, Liu W, Han M, Wang C, et al. HtpS, a novel immunogenic cell surface-exposed protein of Streptococcus suis, confers protection in mice. FEMS Microbiol Lett. 2011;314. doi:10.1111/j.1574-6968.2010.02162.x

112. Rioux S, Neyt C, Di Paolo E, Turpin L, Charland N, Labbé S, et al. Transcriptional regulation, occurrence and putative role of the Pht family of Streptococcus pneumoniae. Microbiology. 2011;157: 336–348.

113. Panina EM, Mironov AA, Gelfand MS. Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. Proc Natl Acad Sci U S A. 2003;100: 9912.

114. Eijkelkamp BA, Pederick VG, Plumptre CD, Harvey RM, Hughes CE, Paton JC, et al. The First Histidine Triad Motif of PhtD Is Critical for Zinc Homeostasis in Streptococcus pneumoniae. Infect Immun. 2016;84: 407.

115. Sullivan MJ, Goh KGK, Ulett GC. Regulatory cross-talk supports resistance to Zn intoxication in Streptococcus. PLoS Pathog. 2022;18. doi:10.1371/journal.ppat.1010607

116. Ochs MM, Williams K, Sheung A, Lheritier P, Visan L, Rouleau N, et al. A bivalent pneumococcal histidine triad protein D-choline-binding protein A vaccine elicits functional antibodies that passively protect mice from Streptococcus pneumoniae challenge. Hum Vaccin Immunother. 2016;12: 2946.

117. Slager J, Aprianto R, Veening JW. Deep genome annotation of the

opportunistic human pathogen Streptococcus pneumoniae D39. Nucleic Acids Res. 2018;46. doi:10.1093/nar/gky725

118. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30: 2068–2069.

119. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11: 1–11.

120. Eddy SR. Accelerated Profile HMM Searches. PLoS Comput Biol. 2011;7: e1002195.

121. GitHub - jhkorhonen/MOODS: MOODS: Motif Occurrence Detection Suite. In: GitHub [Internet]. [cited 29 Aug 2024]. Available: https://github.com/jhkorhonen/MOODS

122. Smaldone GT, Helmann JD. CsoR regulates the copper efflux operon copZA in Bacillus subtilis. Microbiology. 2007;153: 4123.

123. Reyes A, Leiva A, Cambiazo V, Méndez MA, González M. Cop-like operon: Structure and organization in species of the Lactobacillale order. Biol Res. 2006;39: 87–93.

124. Toukoki C, Gold KM, McIver KS, Eichenbaum Z. MtsR is a dual regulator that controls virulence genes and metabolic functions in addition to metal homeostasis in GAS. Mol Microbiol. 2010;76: 971.

125. Singh K, Senadheera DB, Lévesque CM, Cvitkovitch DG. The copYAZ Operon Functions in Copper Efflux, Biofilm Formation, Genetic Transformation, and Stress Tolerance in Streptococcus mutans. J Bacteriol. 2015;197. doi:10.1128/JB.02433-14

126. Bloch S, Hager-Mair FF, Andrukhov O, Schäffer C. Oral streptococci: modulators of health and disease. Front Cell Infect Microbiol. 2024;14: 1357631.

127. Okura M, Maruyama F, Ota A, Tanaka T, Matoba Y, Osawa A, et al. Genotypic diversity of Streptococcus suis and the S. suis-like bacterium Streptococcus ruminantium in ruminants. Vet Res. 2019;50: 1–16.

128. Agnew W, Barnes AC. Streptococcus iniae: an aquatic pathogen of global veterinary significance and a challenging candidate for reliable vaccination. Vet Microbiol. 2007;122. doi:10.1016/j.vetmic.2007.03.002

129. Brooks LRK, Mias GI. Streptococcus pneumoniae's Virulence and Host Immunity: Aging, Diagnostics, and Prevention. Front Immunol. 2018;9. doi:10.3389/fimmu.2018.01366

130. Ogunniyi AD, Grabowicz M, Mahdi LK, Cook J, Gordon DL, Sadlon TA, et al. Pneumococcal histidine triad proteins are regulated by the Zn2+-dependent repressor AdcR and inhibit complement deposition through the recruitment of complement factor H. FASEB J. 2009;23. doi:10.1096/fj.08-119537

131. Osterman A, Overbeek R. Missing genes in metabolic pathways: a comparative genomics approach. Curr Opin Chem Biol. 2003;7: 238–251.

132. Patzer SI, Hantke K. The ZnuABC high-affinity zinc uptake system and its regulator Zur in Escherichia coli. Mol Microbiol. 1998;28. doi:10.1046/j.1365-2958.1998.00883.x

133. Pederick VG, Eijkelkamp BA, Begg SL, Ween MP, McAllister LJ, Paton JC, et al. ZnuA and zinc homeostasis in Pseudomonas aeruginosa. Sci Rep. 2015;5: 1–14.

134. Atack JM, Yang Y, Seib KL, Zhou Y, Jennings MP. A survey of Type III restriction-modification systems reveals numerous, novel epigenetic regulators controlling phase-variable regulons; phasevarions. Nucleic Acids Res. 2018;46: 3532–3542.

135. Melin M, Di Paolo E, Tikkanen L, Jarva H, Neyt C, Käyhty H, et al. Interaction of pneumococcal histidine triad proteins with human complement. Infect Immun. 2010;78. doi:10.1128/IAI.00811-09

136. Luo Z, Pederick VG, Paton JC, McDevitt CA, Kobe B. Structural characterisation of the HT3 motif of the polyhistidine triad protein D from Streptococcus pneumoniae. FEBS Lett. 2018;592: 2341–2350.

137. Waldron KJ, Robinson NJ. How do bacterial cells ensure that metalloproteins get the correct metal? Nat Rev Microbiol. 2009;7. doi:10.1038/nrmicro2057

138. Principles of Bioinorganic Chemistry. In: Google Books [Internet]. [cited 29 Aug 2024]. Available: https://books.google.com/books/about/Principles_of_Bioinorganic_Chemistry.html?hl=ru&id=zGJtXzPINAUC

139. MacDonald RS. The role of zinc in growth and cell proliferation. J Nutr. 2000;130. doi:10.1093/jn/130.5.1500S

140. Wang S, Cheng J, Niu Y, Li P, Zhang X, Lin J. Strategies for Zinc Uptake in Pseudomonas aeruginosa at the Host–Pathogen Interface. Front Microbiol. 2021;12: 741873.

141. Kehl-Fie TE, Skaar EP. Nutritional immunity beyond iron: a role for manganese and zinc. Curr Opin Chem Biol. 2010;14. doi:10.1016/j.cbpa.2009.11.008

142. Maguire ME, Cowan JA. Magnesium chemistry and biochemistry. Biometals. 2002;15. doi:10.1023/a:1016058229972

143. Cowan JA. Structural and catalytic chemistry of magnesium-dependent enzymes. Biometals. 2002;15. doi:10.1023/a:1016022730880

144. Maguire ME. Magnesium transporters: properties, regulation and structure. Front Biosci. 2006;11. doi:10.2741/2039

145. Silver S. ACTIVE TRANSPORT OF MAGNESIUM IN Escherichia coli. Proc Natl Acad Sci U S A. 1969;62: 764.

146. Franken GAC, Huynen MA, Martínez-Cruz LA, Bindels RJM, de Baaij JHF. Structural and functional comparison of magnesium transporters throughout evolution. Cell Mol Life Sci. 2022;79: 1–17.

147. Eshaghi S, Niegowski D, Kohl A, Martinez MD, Lesley SA, Nordlund P. Crystal structure of a divalent metal ion transporter CorA at 2.9 angstrom resolution. Science. 2006;313. doi:10.1126/science.1127121

148. Cleverley RM, Kean J, Shintre CA, Baldock C, Derrick JP, Ford RC, et al. The Cryo-EM structure of the CorA channel from Methanocaldococcus jannaschii in low magnesium conditions. Biochim Biophys Acta. 2015;1848. doi:10.1016/j.bbamem.2015.06.002

149. Niegowski D, Eshaghi S. The CorA family: structure and function revisited. Cell Mol Life Sci. 2007;64. doi:10.1007/s00018-007-7174-z

150. Kitjaruwankul S, Wapeesittipan P, Boonamnaj P, Sompornpisut P. Inner and Outer Coordination Shells of Mg2+ in CorA Selectivity Filter from Molecular Dynamics Simulations. 2016 [cited 29 Aug 2024]. doi:10.1021/acs.jpcb.5b10925

151. Kehres DG, Maguire ME. Structure, properties and regulation of magnesium transport proteins. Biometals. 2002;15. doi:10.1023/a:1016078832697

152. Kent Murmann R. Mechanisms of inorganic reactions - A study of metal

complexes in solution (Basolo, Fred; Pearson, Ralph G.). 1968 [cited 29 Aug 2024]. doi:10.1021/ed045pA146

153. Ramesh A, Winkler WC. Magnesium-sensing riboswitches in bacteria. RNA Biol. 2010;7. doi:10.4161/rna.7.1.10490

154. Dann CE 3rd, Wakeman CA, Sieling CL, Baker SC, Irnov I, Winkler WC. Structure and mechanism of a metal-sensing regulatory RNA. Cell. 2007;130: 878–892.

155. Stav S, Atilho RM, Mirihana Arachchilage G, Nguyen G, Higgs G, Breaker RR. Genome-wide discovery of structured noncoding RNAs in bacteria. BMC Microbiol. 2019;19: 1–18.

156. Worlock AJ, Smith RL. ZntB is a novel Zn2+ transporter in Salmonella enterica serovar Typhimurium. J Bacteriol. 2002;184. doi:10.1128/JB.184.16.4369-4373.2002

157. Gati C, Stetsenko A, Slotboom DJ, Scheres SHW, Guskov A. The structural basis of proton driven zinc transport by ZntB. Nat Commun. 2017;8: 1–8.

158. Nies DH. Biochemistry. How cells control zinc homeostasis. Science. 2007;317. doi:10.1126/science.1149048

159. Giedroc DP, Arunkumar AI. Metal sensor proteins: nature's metalloregulated allosteric switches. Dalton Trans. 2007; 3107–3120.

160. Maciag A, Dainese E, Rodriguez GM, Milano A, Provvedi R, Pasca MR, et al. Global analysis of the Mycobacterium tuberculosis Zur (FurB) regulon. J Bacteriol. 2007;189. doi:10.1128/JB.01190-06

161. The requirement for cobalt in vitamin B12: A paradigm for protein metalation. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research. 2021;1868: 118896.

162. Choi S-H, Lee K-L, Shin J-H, Cho Y-B, Cha S-S, Roe J-H. Zinc-dependent regulation of zinc import and export genes by Zur. Nature Communications. 2017;8: 1–11.

163. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2018;47: D309–D314.

164. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al.

The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242.

165. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. Nucleic Acids Res. 2020;49: D192–D200.

166. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792.

167. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 2010;59: 307–321.

168. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29. doi:10.1093/bioinformatics/btt509

169. Kalinina OV, Mironov AA, Gelfand MS, Rakhmaninova AB. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. Protein Sci. 2004;13: 443.

170. Stetsenko A, Stehantsev P, Dranenko NO, Gelfand MS, Guskov A. Structural and biochemical characterization of a novel ZntB (CmaX) transporter protein from Pseudomonas aeruginosa. Int J Biol Macromol. 2021;184. doi:10.1016/j.ijbiomac.2021.06.130

171. Bi J, Wang Y. The effect of the endosymbiont Wolbachia on the behavior of insect hosts. Insect Sci. 2020;27: 846.

172. Lo N, Casiraghi M, Salati E, Bazzocchi C, Bandi C. How many wolbachia supergroups exist? Mol Biol Evol. 2002;19. doi:10.1093/oxfordjournals.molbev.a004087

173. Balvín O, Roth S, Talbot B, Reinhardt K. Co-speciation in bedbug Wolbachia parallel the pattern in nematode hosts. Sci Rep. 2018;8: 1–9.

174. Manoj RRS, Latrofa MS, Epis S, Otranto D. Wolbachia: endosymbiont of onchocercid nematodes and their vectors. Parasites & Vectors. 2021;14: 1–24.

175. Kaur R, Shropshire JD, Cross KL, Leigh B, Mansueto AJ, Stewart V, et al. Living in the endosymbiotic world of Wolbachia: A centennial review. Cell Host Microbe. 2021;29. doi:10.1016/j.chom.2021.03.006

176. Laidoudi Y, Marie JL, Tahir D, Watier-Grillot S, Mediannikov O, Davoust

B. Detection of Canine Vector-Borne Filariasis and Their Wolbachia Endosymbionts in French Guiana. Microorganisms. 2020;8. doi:10.3390/microorganisms8050770

177. Lefoulon E, Clark T, Borveto F, Perriat-Sanguinet M, Moulia C, Slatko BE, et al. Pseudoscorpion Wolbachia symbionts: diversity and evidence for a new supergroup S. BMC Microbiol. 2020;20: 1–15.

178. Liu B, Ren YS, Su CY, Abe Y, Zhu DH. Pangenomic analysis of Wolbachia provides insight into the evolution of host adaptation and cytoplasmic incompatibility factor genes. Front Microbiol. 2023;14. doi:10.3389/fmicb.2023.1084839

179. Ellegaard KM, Klasson L, Näslund K, Bourtzis K, Andersson SGE. Comparative Genomics of Wolbachia and the Bacterial Species Concept. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003381

180. Scholz M, Albanese D, Tuohy K, Donati C, Segata N, Rota-Stabelli O. Large scale genome reconstructions illuminate Wolbachia evolution. Nat Commun. 2020;11. doi:10.1038/s41467-020-19016-0

181. Chafee ME, Funk DJ, Harrison RG, Bordenstein SR. Lateral phage transfer in obligate intracellular bacteria (wolbachia): verification from natural populations. Mol Biol Evol. 2010;27. doi:10.1093/molbev/msp275

182. Al-Khodor S, Price CT, Kalia A, Kwaik YA. Ankyrin-repeat containing proteins of microbes: a conserved structure with functional diversity. Trends Microbiol. 2010;18: 132.

183. Sedgwick SG, Smerdon SJ. The ankyrin repeat: a diversity of interactions on a common structural framework. Trends Biochem Sci. 1999;24. doi:10.1016/s0968-0004(99)01426-7

184. Habyarimana F, Al-Khodor S, Kalia A, Graham JE, Price CT, Garcia MT, et al. Role for the Ankyrin eukaryotic-like genes of Legionella pneumophila in parasitism of protozoan hosts and human macrophages. Environ Microbiol. 2008;10. doi:10.1111/j.1462-2920.2007.01560.x

185. Pan X, Lührmann A, Satoh A, Laskowski-Arce MA, Roy CR. Ankyrin repeat proteins comprise a diverse family of bacterial type IV effectors. Science. 2008;320. doi:10.1126/science.1158160

186. Bochkareva OO, Moroz EV, Davydov II, Gelfand MS. Genome rearrangements and selection in multi-chromosome bacteria Burkholderia spp. BMC Genomics. 2018;19. doi:10.1186/s12864-018-5245-1

187. Brandis G, Hughes D. The SNAP hypothesis: Chromosomal rearrangements could emerge from positive Selection during Niche Adaptation. PLoS Genet. 2020;16. doi:10.1371/journal.pgen.1008615

188. Dobrindt U, Chowdary MG, Krumbholz G, Hacker J. Genome dynamics and its impact on evolution of Escherichia coli. Med Microbiol Immunol. 2010;199. doi:10.1007/s00430-010-0161-2

189. Sadarangani M, Hoe CJ, Makepeace K, van der Ley P, Pollard AJ. Phase variation of Opa proteins of Neisseria meningitidis and the effects of bacterial transformation. J Biosci. 2016;41. doi:10.1007/s12038-016-9588-y

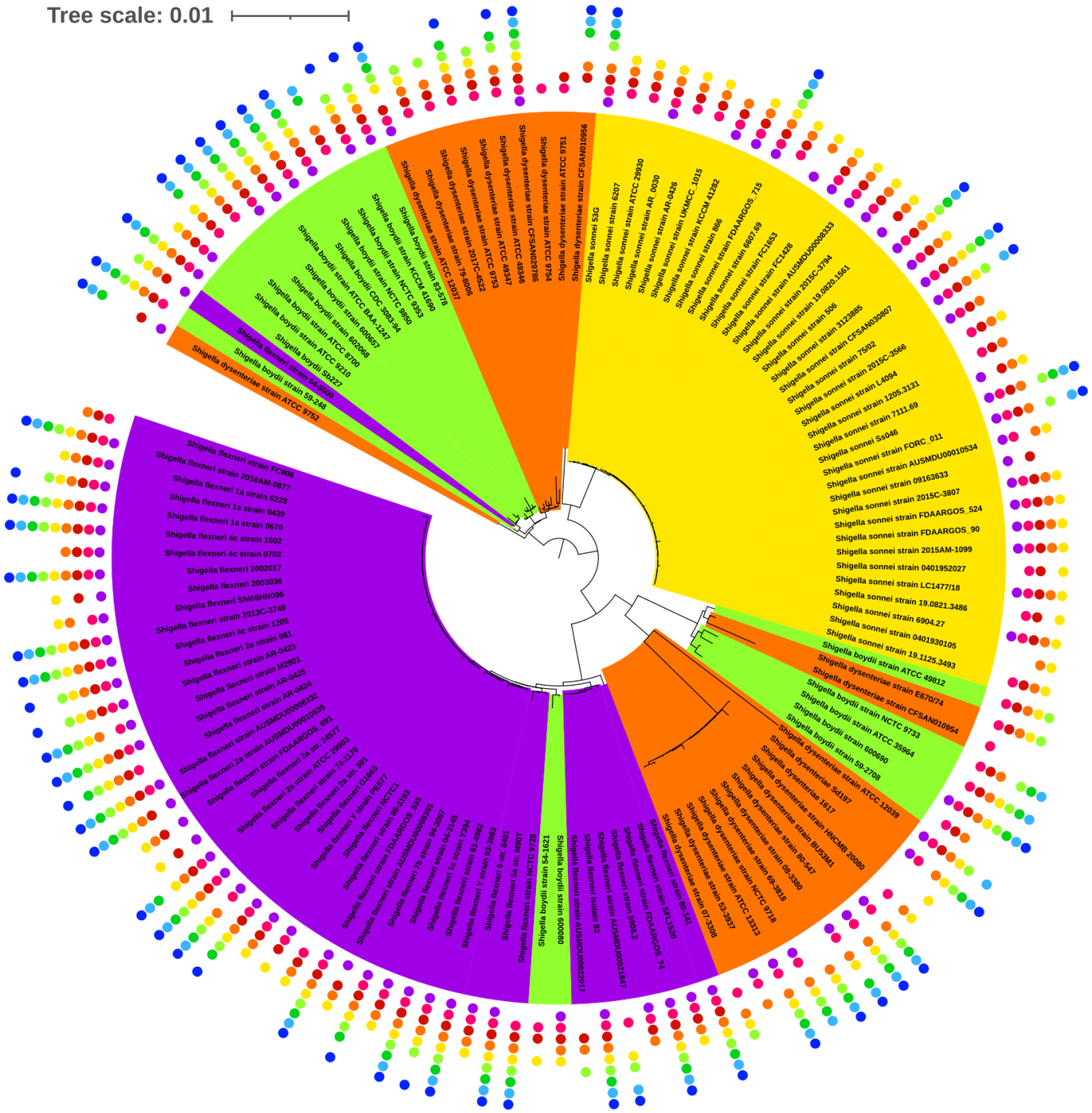
190. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20: 238.

191. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34. doi:10.1093/nar/gkj014

192. [No title]. [cited 29 Aug 2024]. Available: https://www.ebi.ac.uk/Tools/pfa/pfamscan/

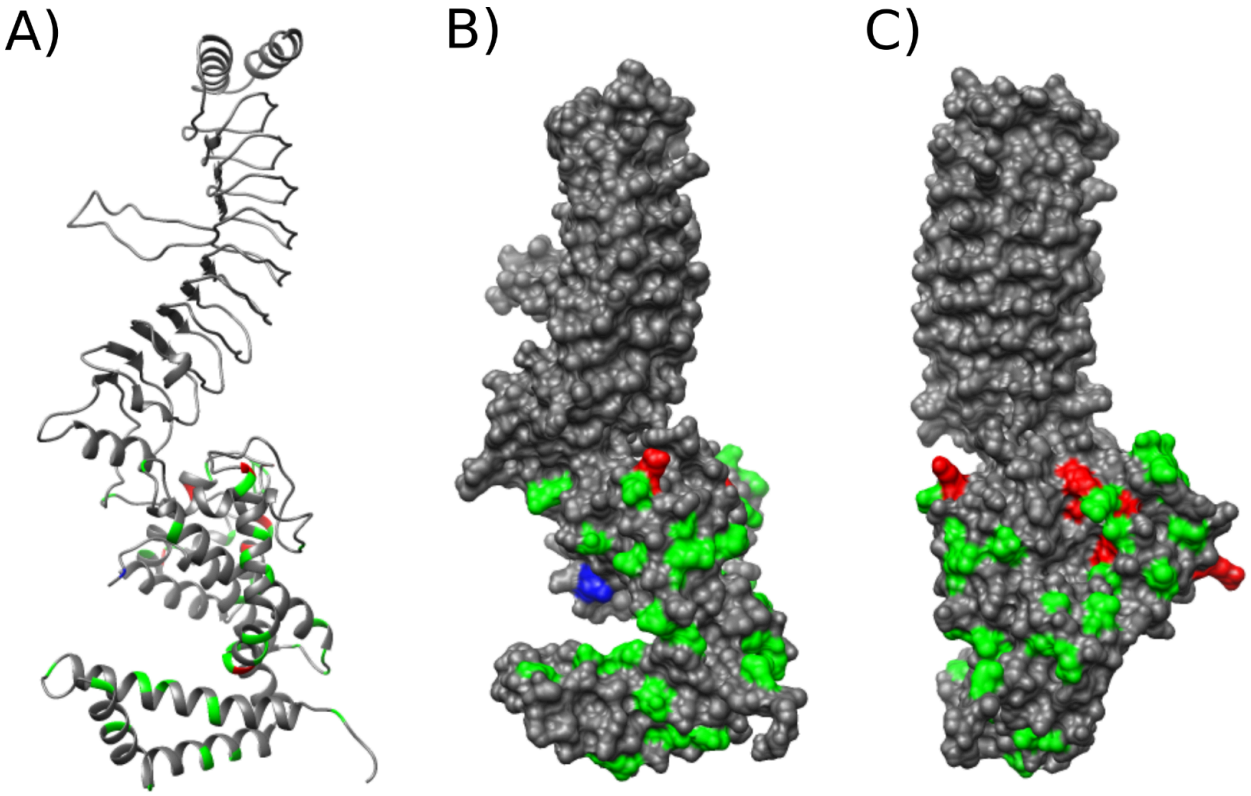
193. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–1659.

Приложение



Дополнительный рисунок 1. Филогенетическое дерево штаммов *Shigella*, окрашенное по видовым названиям.

Дополнительный рисунок 2. Разнообразие представителей семейства *ipaH* в кладах *Shigella* для трёх клад дерева. Обозначения соответствуют рисунку 4. Чёрными рамками обозначены дуплицированные гены. Панели a),b),c) соответствуют трём крупным кладам дерева с рисунка 4.



D)

```
ShigellaCdomain    GPQIYFSMSDGQQNTLHRPLADAVTAWFPENKQSDVSQIWHAFEHEEHANTFSAFLDRLS
Non-humanCdomain1 GPQIHFSMSDGQQHTSVRPLPEAVAAWFPQNKQSDVSQIWLAFEREEHANTFSAFLDRLA
Non-humanCdomain2 GPQIHFSMSDGQQHTPVRSLPEAVAAWFPESLRSEVSQRWAAFTDEENAATFSAFLDRLA

ShigellaCdomain    DTVSARNTSGFREQVAAWLEKLSTSAELRQQSFAVAADATESCEDRVALTWNNLRKTLLV
Non-humanCdomain1 DTVSARNAQGFRQQVSAFLEKLSTSAELRQQSFAVAADATESCEDRVALTWNNLQKSLLV
Non-humanCdomain2 DTVTARNAPGFAQQVSEFLEKLSMSEALRQQCFAVAADATRSCEDRVALTWNNLQKTYRV

ShigellaCdomain    HQASEGLFDNDTGALLSLGREMFRLEILEDIARDKVRTLHFVDEIEVYLAFQTMLAEKLQ
Non-humanCdomain1 HQASEGLFDNDTGALLSLGREMFRLEILEDIARDKVRTLRFVDEIEVYLAFQTMLAEKLQ
Non-humanCdomain2 HQASEGEFDSDLTGLLSLGREMYRLEVLEEIAREKVRTLHFVDEIEVSLAFQTMLAEKLE

ShigellaCdomain    LSTAVKEMRFYGVSGVTANDLRTAEAMVRSREENEFTDWFSLWGPWHAVLKRTEADRWAQ
Non-humanCdomain1 LSTAVKEMRFYGVSGVTENDLRTAEAMVRSREENEFTDWFALWGPWHAVLKRTEADRWAQ
Non-humanCdomain2 LSTAVREMRFYGVSGVTEDDLSSALVRVLSREEREFAEWFARWSPWHAVLKRTEAERWAR

ShigellaCdomain    AEEQKYEMLENEYPQRVADRLKASGLSGDADAEREAGAQVMRETEQQIYRQLTDEVLALR
Non-humanCdomain1 AEEQKYEMLENEYPQRVADRLKASGLSGDADAEREAGAQVMRETEQQIYRQLTDEVLALR
Non-humanCdomain2 AEEKKYEMLERKYPQRLAERLSALDLSGDGDAEREAGVRVMEEIETEIYRQLTEEVSGER

ShigellaCdomain    RLPENGSRLHHS
Non-humanCdomain1 RLPENGSQLHHS
Non-humanCdomain2 RLAENRAR----
```

Дополнительный рисунок 3. Моделирование структуры эффектора класса 1 из *Escherichia marmotae* по гомологии с эффектором из *Shigella* A) в ленточном представлении, B) с поверхностью белка со стороны сайта связывания, C) с поверхностью белка со стороны, противоположной сайту связывания. D) Выравнивание последовательностей трех вариантов С-концевых доменов у *Shigella* spp. и *E. marmotae*. Активный сайт отмечен синим цветом. Позиции, одинаковые у двух доменов *E. marmotae*, но отличные у *Shigella* spp. отмечены красным цветом; позиции, одинаковые у

Shigella spp. и доменом шигелльгого типа у *E. marmotae*, отмечены зеленым цветом.

Дополнительная таблица 1. Таблица геномов шигелл и энтероинвазивных кишечных палочек из человека.
https://github.com/zaryanichka/DissertationTables/blob/main/1HumanHostGenomesTable

Дополнительная таблица 2. Таблица геномов *Escherichia* spp. из животных.
https://github.com/zaryanichka/DissertationTables/blob/main/2NonHumanHostGenomesTable

Дополнительная таблица 3. Таблица IpaH из шигелл и энтероинвазивных кишечных палочек из человека.
https://github.com/zaryanichka/DissertationTables/blob/main/3HumanHostAllLigasesTable

Дополнительная таблица 4. Таблица IpaH из *Escherichia* spp. из животных.
https://github.com/zaryanichka/DissertationTables/blob/main/4NonHumanHostAllLigasesTable

Дополнительная таблица 5. Консенсусы IpaH всех классов из шигелл и энтероинвазивных кишечных палочек из человека.
https://github.com/zaryanichka/DissertationTables/blob/main/5HumanHostConsensusAllClassesTable

Дополнительная таблица 6. IpaH всех классов из *Escherichia* spp. из животных с указанием кодов генов и свойств убиквитин-лигазного домена.
https://github.com/zaryanichka/DissertationTables/blob/main/6NonHumanLigasesNames

Дополнительная таблица 7. Таблица геномов *Streptococcus* spp. по видам.
https://github.com/zaryanichka/DissertationTables/blob/main/7StreptococcusSpeciesTable

Дополнительная таблица 8. Мотивы транскрипционных факторов CsoR, CopY, MtsR, AdcR.
https://github.com/zaryanichka/DissertationTables/blob/main/8TFMotifsStreptococcus

Дополнительная таблица 9. Таблица геномов *Wolbachia* spp. с указанием хозяев.
https://github.com/zaryanichka/DissertationTables/blob/main/9WolbachiaHostTable