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

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата биологических наук

Москва — 2025

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научный руководитель:	**Гельфанд Михаил Сергеевич** доктор биологических наук, профессор
Официальный оппоненты:	**Фишман Вениамин Семенович** доктор биологических наук, заведующий сектором геномных механизмов онтогенеза Института цитологии и генетики СО РАН, ведущий научный сотрудник лаборатории структурно-функциональный организации генома Новосибирского государственного университета **Спирин Сергей Александрович** кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского института физико-химической биологии имени А.Н. Белозерского Московского государственного университета имени М.В. Ломоносова
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

Защита состоится 9 февраля 2026 г. в 17:00 на заседании диссертационного совета 24.1.101.01 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации имени А.А. Харкевича Российской академии наук (ИППИ РАН) по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН, а также на сайте ИППИ РАН по адресу: http://iitp.ru/upload/content/1744/DNO%20dissertation.pdf

Автореферат разослан «__» декабря 2025 г.

Учёный секретарь
диссертационного совета 24.1.101.01,
доктор биологических наук Казенников Олег Васильевич

Общая характеристика работы

Актуальность темы исследования

Белки играют центральную роль в биологических процессах. Мутации в генах, кодирующих белки, ведут к изменению также и в последовательностях самих белков; эти мутации могут быть полезными, вредными или нейтральными. Бактерии являются крайне изменчивыми организмами, и в присутствии отбора мутации приводят к адаптации колонии к новой или меняющейся окружающей среде. Отслеживание такого рода изменений позволяет раскрыть механизмы, лежащие в основе жизнедеятельности бактерий.

Определить влияние тех или иных изменений на функцию белка возможно либо экспериментально, либо биоинформатически. 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. Мутации могут влиять не только на последовательность белка, но также и на регуляцию соответствующих генов. Например, дупликация гена может просто увеличить производство белка, а может привести к специализации копий для работы в разных условиях под действием разной регуляции.

С появлением большого количества данных о геномах и структурах белков стал возможен комплексный анализ эволюции белковых семейств, что помогает точнее описать их функции и роль в жизни бактерий.

Степень разработанности темы

Функция многих белков, несмотря на прогресс в качестве аннотации, остается неизвестной, например, около половины генов в бактериальном геноме могут быть не аннотированы. Предсказать функцию белков с неизвестным назначением позволяет анализ геномного контекста и структурных данных. Однако этот метод требует большого количества данных, а их доступность сильно варьируется между видами бактерий (например, для *Escherichia coli* доступны сотни тысяч геномов, а для большинства родов — всего один). Это создаёт основную сложность для применения современных подходов.

Проблема недостатка данных актуальна даже для хорошо изученных семейств. Например, для семейства белков IpaH шигелл известны структуры и молекулярные функции отдельных представителей, но не изучена эволюция семейства и его распространённость в разных хозяевах. Белки гистидиновых триад стрептококков детально изучены у пневмококка, но их роль у других видов и связь с патогенностью требуют исследования. Особый интерес представляют также многокопийные семейства, такие как ANK-белки вольбахий, так как их роль как

эффекторов у разных штаммов и хозяев изучена слабо, и широко распространённые семейства, такие как транспортёры семейства CorA, для которых важно понимание их субстратной специфичности и механизма работы.

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*.

Научная новизна

В работе был получен ряд новых результатов. 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 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 работы

Результаты данного исследования имеют фундаментальное теоретическое значение, расширяя знания о бактериальных белковых семействах, а также практическую медицинскую ценность благодаря изучению белков, играющих ключевую роль в патогенезе.

Белки гистидиновых триад высоко иммуногенны, изучение их особенностей критично для создания эффективных вакцин, устойчивых к ускользанию патогена от иммунного ответа. Изучение семейства эффекторов IpaH у патогенов позволяет прогнозировать горизонтальный перенос генов и потенциальное возникновение новых зоонозных инфекций. Анализ широко распространённой транспортной системы CorA открывает перспективы для разработки новых классов антибиотиков или методов доставки лекарств, нацеленных на эти бактериальные каналы.

Методология и методы исследования

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.

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 автора 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).

Содержание работы

В первой главе приведён обзор литературы по теме исследования, состоящий из шести разделов.

В разделе 1.1 описана эволюция белка как последовательности аминокислот, изложены основные факторы, влияющие на изменение последовательностей белок-кодирующих генов.

В разделе 1.2 рассматривается понятие белкового семейства и особенности входящих в одно семейство белков, сходство их последовательностей и структур.

Раздел 1.3 посвящён формированию новой функциональной специфичности у паралогов, описаны возможные эволюционные пути, которыми следуют гены после дупликации, рассмотрено влияние субстратной специфичности на вероятность приобретения новой функции.

В разделе 1.4 описаны различные способы регуляции экспрессии генов в связи со специфичностью кодируемого белка, такие как регуляция с помощью белков, малых РНК и рибопереключателей.

В разделе 1.5 речь идёт о понятии фазовой вариации, различных её механизмах и преимуществах, которые фазовая вариации даёт бактериям.

В разделе 1.6 приводится обзор методов исследования эволюции белковых семейств, рассмотрены методы построения выравниваний последовательностей, основные принципы построения филогенетических деревьев, методы сравнения белковых структур.

Вторая глава посвящена эволюции семейства эффекторов IpaH у шигелл. Она состоит из четырёх разделов.

Раздел 2.1 содержит обзор литературы на тему непосредственно шигелл и эффекторов IpaH. Описана немонофилетичность дерева шигелл, их принадлежность кишечным палочкам и близость по механизму инвазии к энтероинвазивным кишечным палочкам. Представлен механизм развития инфекции для шигелл и роль IpaH в нём. Описана структура входной области плазмиды инвазивности и особенности регуляции *ipaH* в ней. Показана доменная структура IpaH.

Раздел 2.2 описывает методы, используемые для изучения IpaH из шигелл.

В работе использовались 130 полных геномов *Shigella* spp., доступных в GenBank по состоянию на ноябрь 2020 года, три полных генома ЭИКП и все сборки *Escherichia* spp., хозяева которых были животными, содержавшие рамки считывания, продукты которых, согласно BLAST, имели сходство с NEL-доменом IpaH *Shigella* spp. С использованием pBLAST-поиска NEL-домена были обнаружены 445

белковых последовательностей, принадлежащих к семейству E3 убиквитин лигаз, они были сгруппированы с помощью CD-hit, после чего был проведён дополнительный поиск tBLASTn для репрезентативных последовательностей из каждого кластера. В общей сложности обнаружены и классифицированы 864 последовательности *ipaH*.

Тепловые карты для сходства последовательностей были составлены с использованием R-пакетов seqinr, RColorBrewer и gplots. Для построения дерева видов *Shigella* spp. использовался инструмент PanACoTA. Выравнивания регуляторных областей генов *ipaH* были построены с помощью инструмента Pro-Coffee, дополнительные промоторы были предсказаны с помощью алгоритма platprom. Сайты связывания VirF были предсказаны вручную на основе филогенетического футпринтинга известных сайтов связывания. Трехмерные структуры белков IpaH из *Escherichia marmotae* были смоделированы с использованием программы Swiss-Model и визуализированы с помощью UCSF Chimera.

Раздел 2.3 описывает результаты анализа семейства белков IpaH у шигелл и состоит из пяти подразделов.

Подраздел 2.3.1 посвящён валидации 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. Два критерия были введены для подтверждения аннотации *Shigella*: наличие генов *ipaH* и компонентов CC3T. Только 64 сборки содержали все необходимые элементы вирулентности. В 55 сборках присутствовали только хромосомные, но не плазмидные *ipaH*. В 8 сборках содержались как хромосомные, так и плазмидные *ipaH*, но не было обнаружено компонент CC3T. Геномов с CC3T, но без *ipaH*, обнаружено не было., при этом в трёх сборках не было ни *ipaH*, ни компонент CC3T. Из трёх охарактеризованных полных геномах ЭИКП два штамма имели плазмиду инвазивности с генами CC3T и гены *ipaH* на хромосомах и плазмидах.

В подразделе 2.3.2 приводится подход к классификации генов *ipaH*.

Не существует согласованной номенклатуры генов *ipaH* у разных видов рода *Shigella*, и количество таких генов в штаммах варьируется, поэтому была предложена объединяющая классификация всех генов семейства *ipaH*. В 127 сборках шигелл были обнаружены 864 гена, кодировавших белки из семейства убиквитин лигаз E3. На основании сходства последовательностей распознающих доменов и состава регуляторных элементов в регуляторных областях, все гены *ipaH* были разделены на девять классов.

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. 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. Подкласс 9b был обнаружен почти во всех геномах *Shigella flexneri*, поэтому возникло предположение, что ген *ipaH* 9b был приобретен общим предком ветви *S. flexneri*.

В подразделе 2.3.3 описаны 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. 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*. Каждый класс, определенный с помощью предложенного подхода по подобию последовательностей, за исключением класса 9, соответствует одному из регуляторных классов (Рис. 1A,B). Кроме того, регуляторные участки содержат участки, богатые A и T, в качестве возможных мишеней для взаимодействия с VirF и H-NS.

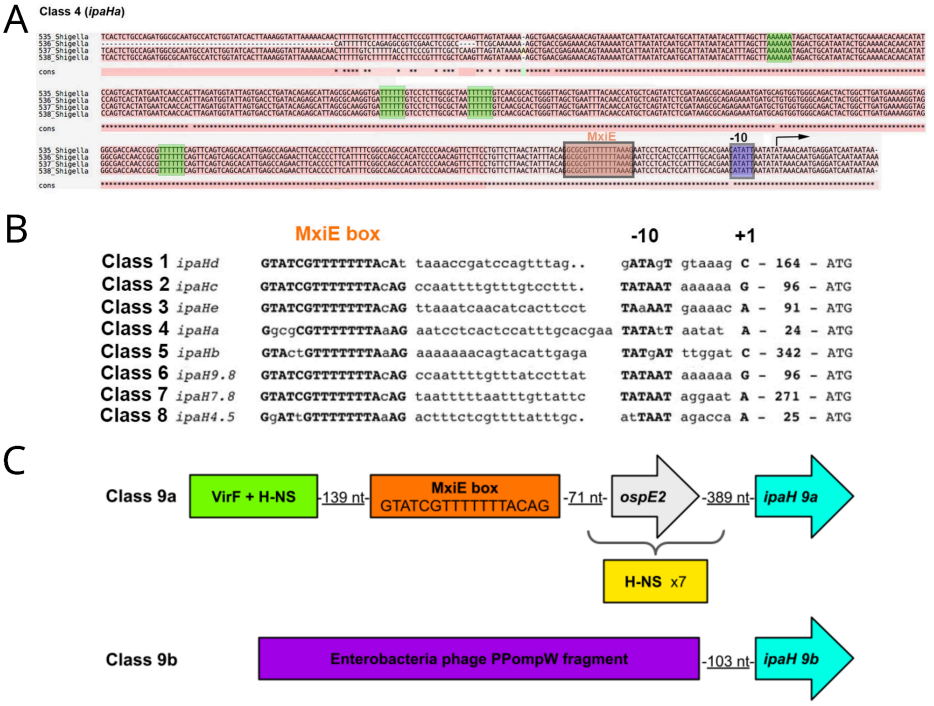

Рисунок 1. Регуляторные элементы, контролирующие гены *ipaH*. A) Выравнивание регуляторной области для отдельных представителей *ipaH* класса 4. Представители

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 (Bongrand et. al. 2012). C) Схема регуляторных областей для классов *ipaH* 9a и 9b.

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* (Рис. 1C). 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* (Рис. 1C), используя его регуляторные элементы.

Подраздел 2.3.4 описывает филетические паттерны генов *ipaH*.

Реконструированное филогенетическое дерево в целом соответствовало предыдущим реконструкциям и имело пять основных клад *Shigella* spp., причем названия видов не соответствуют топологии дерева. В этом наборе данных *S. sonnei* (желтые) и *S. flexneri* (фиолетовые) были монофилетическими, *S. boydii* и *S. dysenteriae* были смешаны в двух удаленных кладах (оранжевые и красные), а набор штаммов *S. dysenteriae* образовал пятую кладу (зеленые) (Рис. 2). Филетические паттерны генов *ipaH* были мозаичными, но наблюдались некоторые специфичные для клад закономерности. Штаммы ЭИКП не группировались с основными кладами *Shigella* или друг с другом.

Интересно, что копии генов *ipaH* были обнаружены во многих геномах шигелл как на хромосомах, так и на плазмидах. Паралоги *ipaH* 2, 4, 5 наблюдаются в оранжевой кладе, паралоги только *ipaH* 4 присутствуют в зеленой кладе и паралоги только *ipaH* 3 были обнаружены в красной кладе. Геномы фиолетовой клады имели паралоги *ipaH* 3, 4, 5, 7 и 9, в то время как геномы в желтой кладе не имели дуплицированных *ipaH*, причём ни один из дуплицировавшихся генов не был тандемным повтором.

Подраздел 2.3.5 посвящён репертуару *ipaH* у *Escherichia* spp. из животных хозяев.

Был проведен поиск и сравнение генов *ipaH* у патогенных штаммов *Escherichia* spp., выделенных у животных хозяев (Рис. 3). Ранее девять генов *ipaH* и две коротких ОРС, содержащих фрагменты генов *ipaH*, были обнаружены в геноме *Escherichia marmotae* HT073016, выделенном из фекалий *Marmota himalayana*: четыре на плазмиде *pEM148*, пять на плазмиде *pEM76* и два на хромосоме. В соответствии с используемой

процедурой идентификации *ipaH* было подтверждено восемь из них и обнаружен один дополнительный хромосомный ген.

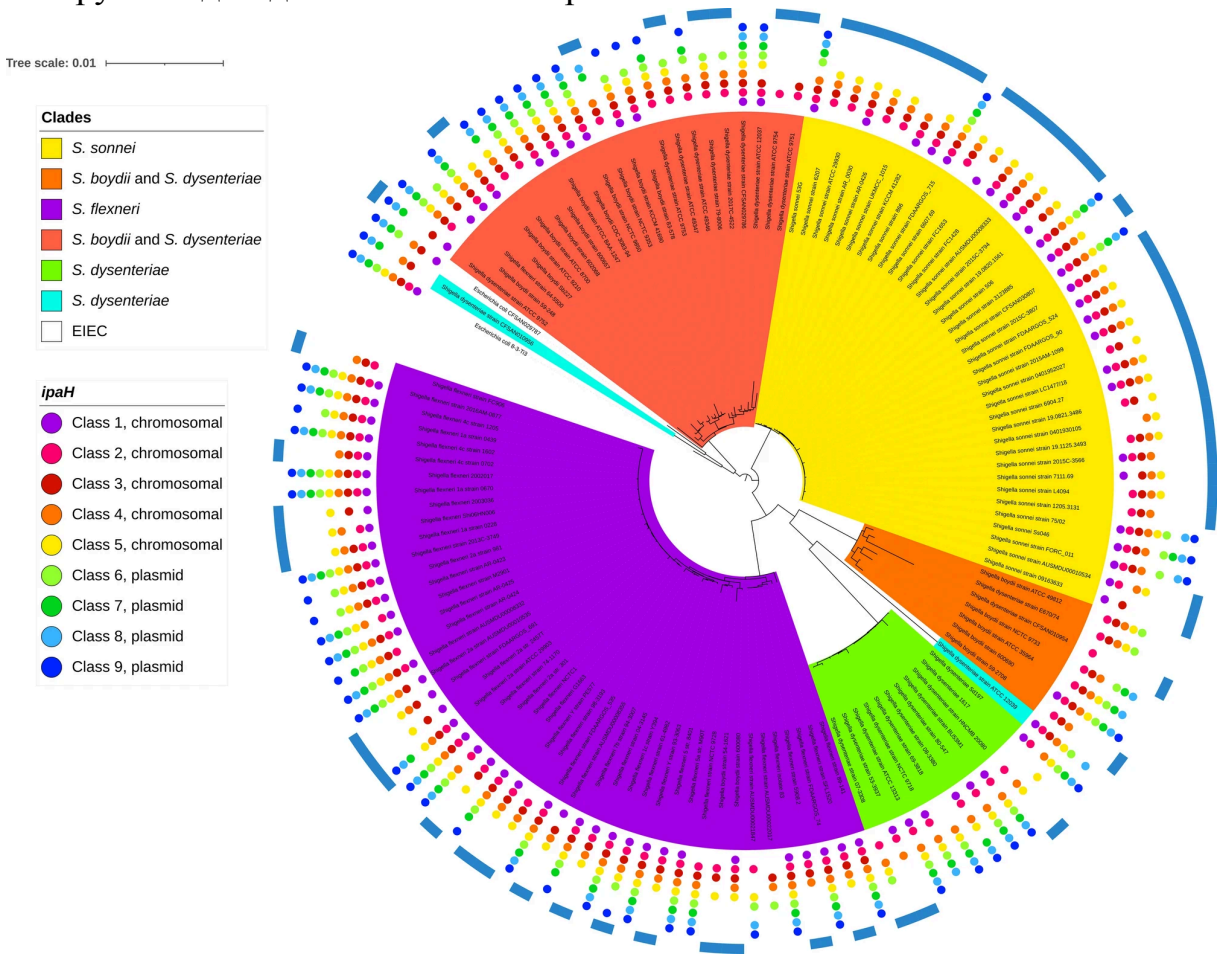

Рисунок 2. Филетические паттерны генов *ipaH* у шигелл и ЭИКП. Покраска неукоренённого дерева отражает основные клады шигелл, которые предположительно произошли от различных непатогенных *E. coli*; два отдаленных штамма шигелл показаны голубым цветом, штаммы ЭИКП показаны белым. Наличие генов *ipaH* 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.

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*.

Основываясь на сходстве последовательностей распознающих доменов, белки IpaH из *E. coli* с хозяевами-животными были разделены на девять классов. 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.

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. В частности, регуляторные области *ipaH* у штаммов *Escherichia* spp. из животных хозяев не содержит ни сайтов связывания MxiE, ни полиA/полиT-последовательностей.

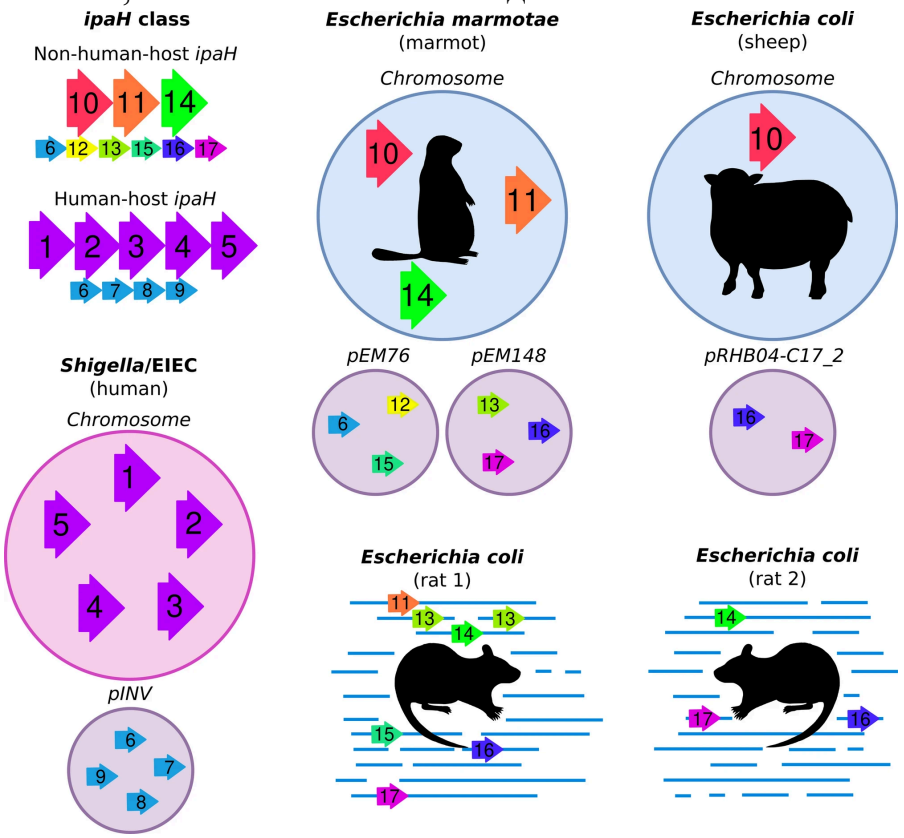

Рисунок 3. Состав генов *ipaH* в геномах *Escherichia* spp. от разных хозяев. Сборки *Escherichia* из сурка и овцы полные, геномы *Escherichia* spp. из крысы собраны в виде контигов. Для шигелл показан геном с полным набором генов *ipaH*.

В отличие от шигелл и кишечных палочек человека у штаммов *Escherichia* spp. из животных хозяев С-концевой домен белков IpaH не сохранялся. 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). Оба варианта являются специфичными для *E. coli* и отличными от убиквитин-лигазных доменов других патогенов, таких как *Salmonella*, *Yersinia* и другие. Аминокислотные различия между консенсусными последовательностями С-концевых доменов IpaH из *Shigella* spp. и *Escherichia* spp. из животных хозяев не были сгруппированы и не влияли на активный сайт белка.

Раздел 2.4 представляет собой обсуждение результатов анализа IpaH из шигелл, приведённого в предыдущем разделе.

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. 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.

Хотя наличие эффекторов IpaH является одним из маркеров, используемых для серотипирования *Shigella*, ни один ген *ipaH* не является универсальным для всех штаммов *Shigella*. Наблюдаются многочисленные независимые потери генов, что может быть связано с функциональной избыточностью самих эффекторов.

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.

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. 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. Кроме того, в белках IpaH, кодируемых в геномах эшерихий из животных, были обнаружены два различных С-концевых домена.

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. Однако IpaH может влиять на антимикробную активность белков хозяина даже в отсутствие

каталитической активности. Таким образом, нацеливание на N-концевые домены может быть более эффективным, но эта стратегия требует понимания репертуара *ipaH* у конкретных штаммов.

Третья глава посвящена анализу белков гистидиновых триад у стрептококков и состоит из четырёх разделов.

Раздел 3.1 содержит обзор литературы о белках гистидиновых триад у стрептококков. Описаны патогенные и непатогенные представители рода *Streptococcus*. Представлено понятие гистидиновой триады, дано определение рассматриваемых белков, указана их медицинская значимость. Описана регуляция генов, кодирующих белки гистидиновых триад. Показано, что для генов белков гистидиновых триад известен пример фазовой вариации по типу инверсии.

В разделе 3.2 представлены методы, использованные при анализе белков гистидиновых триад у стрептококков.

В работе использовались 819 полных геномов *Streptococcus* spp., доступных в GenBank по состоянию на апрель 2023 года. Для начальной аннотации геномов *Streptococcus* spp. был использован модуль annotate инструмента PanACoTA. Для поиска генов белков гистидиновых триад использовалось два способа: использование базы известных последовательностей белков гистидиновых триад из GenBank в качестве материала для составления HMM-профиля и дальнейшего поиска соответствующих белков с помощью HMMer, результаты такого подхода были признаны неудовлетворительными, и прямой поиск белков, содержащих по меньшей мере два мотива гистидиновых триад HxxHxH, такой поиск дал 1802 белка.

Для предсказания сайтов связывания CsoR, CopY, MtsR и AdcR была использована программа на основе библиотеки MOODS. Для предсказания генов, предположительно подверженным фазовой вариации путём инверсии, использовался специально разработанный алгоритм.

В разделе 3.3 описаны результаты анализа белков гистидиновых триад у стрептококков, он состоит из четырёх подразделов.

В подразделе 3.3.1 содержится описание разнообразия белков гистидиновых триад.

Было проанализировано 819 полных геномов *Streptococcus* spp., принадлежащих к 64 различным известным видам. 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. На дереве соответствующих последовательностей эти белки не образуют клады по видам (Рис. 4). Для большинства белков автоматические

аннотаторы не предсказывают функции. Представители одной из ортологических групп, в которые входят белки гистидиновых триад, стабильно аннотируются как CopB, белки, участвующие в гомеостазе меди.

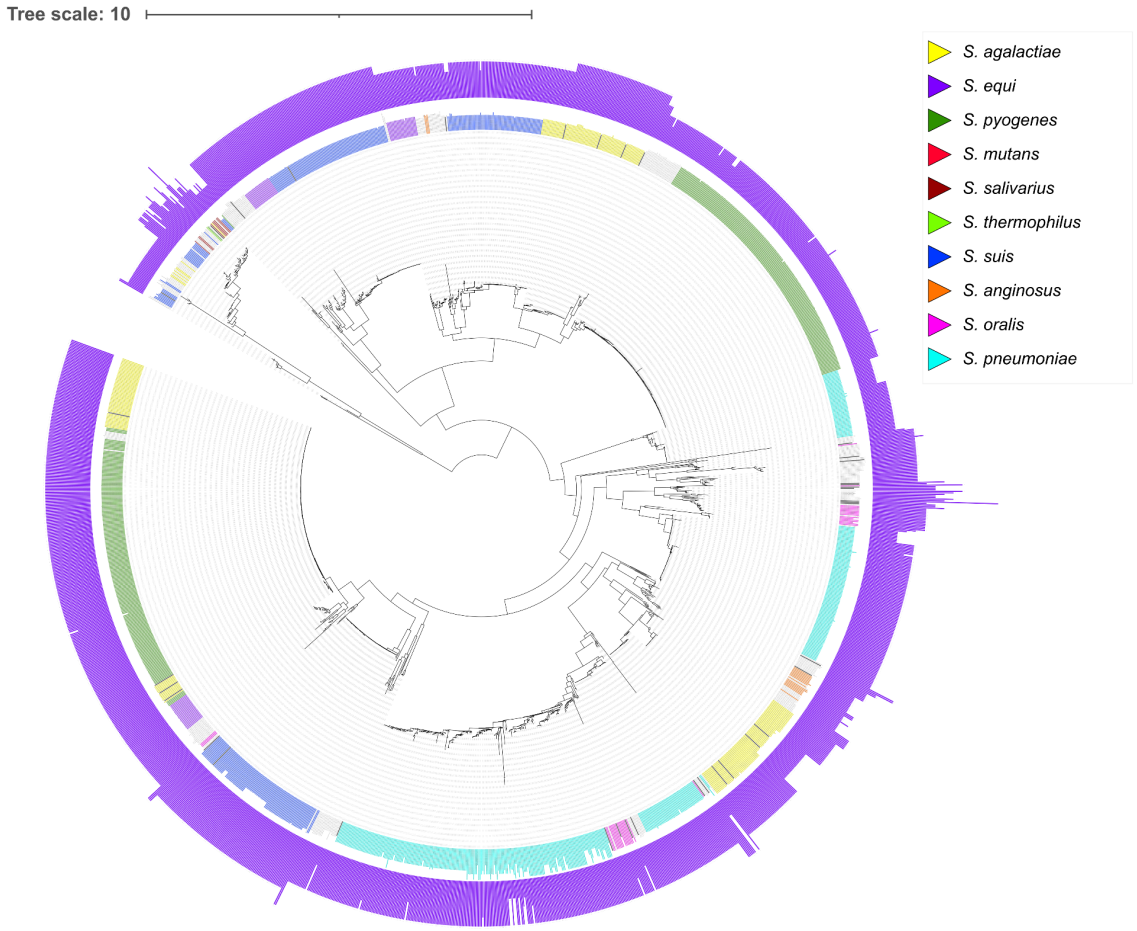

Рисунок 4. Филогенетическое дерево белков гистидиновых триад. Листья окрашены в соответствии с видом *Streptococcus*, из которого был взят белок. Высота столбца во внешнем фиолетовом кольце указывает на число гистидиновых триад в белке.

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.

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 (Рис. 5). Более того, консенсусы первых двух мотивов для белков с пятью и шестью триадами различаются: HGDHYH и HGDHEH, соответственно; при этом 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.

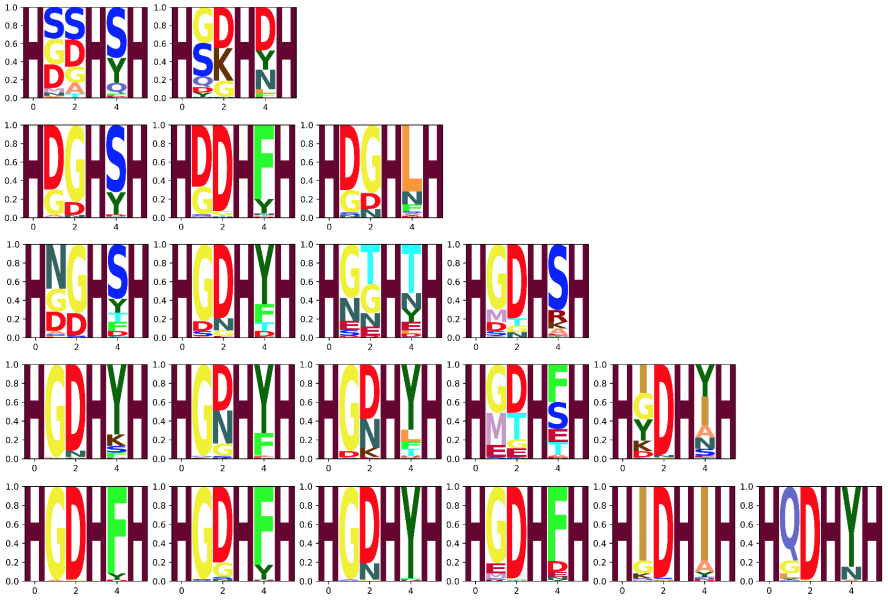

Рисунок 5. Частоты встречаемости аминокислот в гистидиновых триадах, сгруппированные по белкам с разным количеством гистидиновых триад и представленные в виде лого. Рассмотрены только группы, содержащие 20 или более белков.

Подраздел 3.3.2 посвящён представленности разных вариантов белков гистидиновых триад у представителей разных видов *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.

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. При этом максимальное число таких генов, шесть, было обнаружено у представителей вида *S.*

pneumoniae.

В подразделе 3.3.3 речь идёт о регуляции генов, кодирующих белки гистидиновых триад у стрептококков.

Для всех исследуемых генов был проведён поиск сайтов связывания цинкового репрессора 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 (Рис. 6).

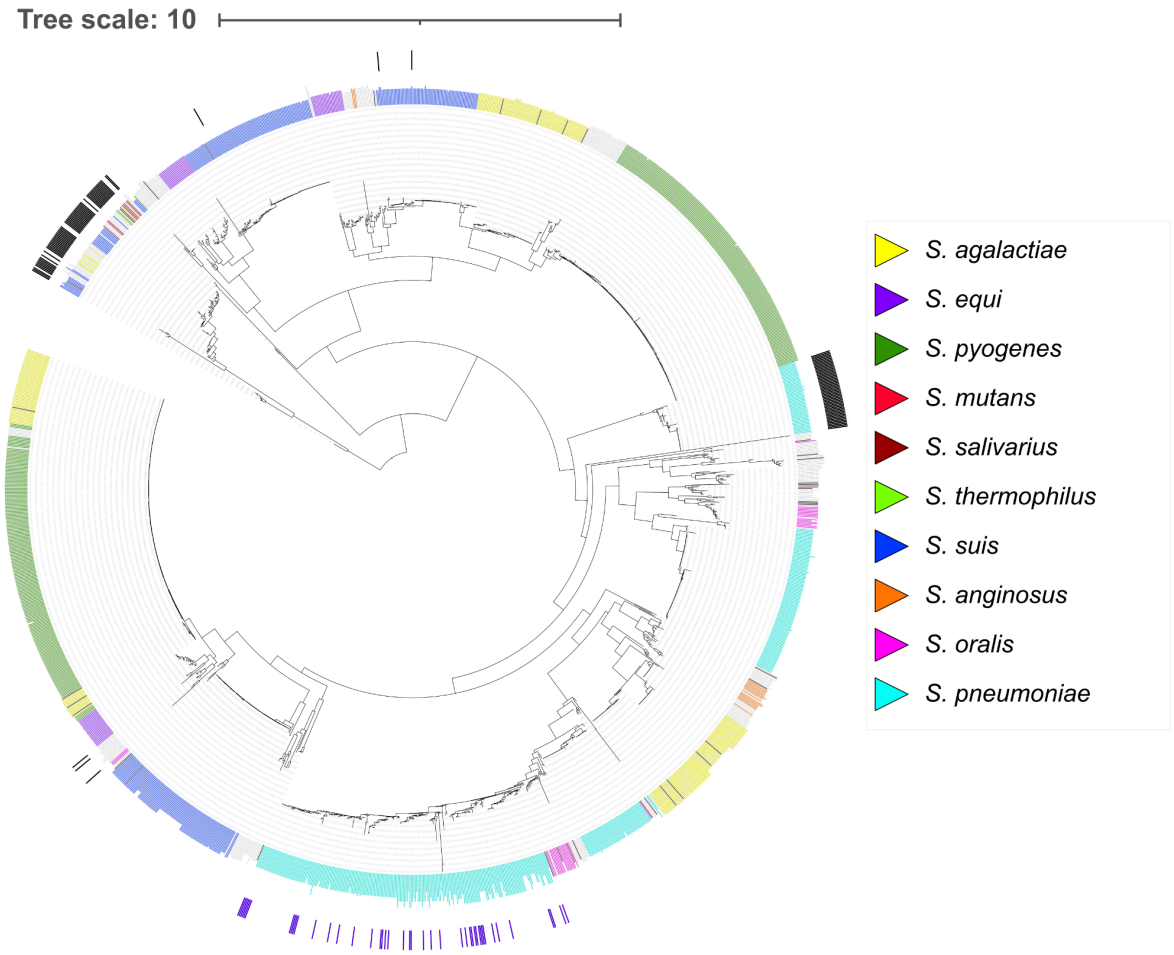

Рисунок 6. Филогенетическое дерево белков гистидиновых триад. Цвета листьев соответствуют виду *Streptococcus*. Черными шрихами обозначены гены, для которых не найдены потенциальные сайты связывания цинковых репрессоров. Фиолетовыми штрихами (внешнее кольцо) обозначены гены, для которых возможна фазовая вариация по механизму инверсии (детали см. в тексте).

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. 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.

Чтобы проверить, действительно ли гены белков гистидиновых триад регулируются медными репрессорами, был произведён поиск мотивов медных регуляторов CsoR, CopY и MtsR. И действительно, на ветви, где исследуемые гены аннотируются как *copB* и соседствуют с генами гомеостаза меди, часть генов белков гистидиновых триад контролируется медными репрессорами. Примечательно, что для регулятора CopY был показан ответ не только на изменение концентрации меди, но и на изменение концентрации цинка. Ещё одним отличием белков этой ветви от остальных белков дерева является высокая доля серина в последовательностях гистидиновых триад и меньшее количество самих триад.

Вторая ветвь, в которой гены не регулируются 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. Помимо отсутствия характерной для генов *phtE* регуляции цинковыми репрессорами, не наблюдается частая для гена *phtE* ко-локализация с *znuA*.

В подразделе 3.3.4 представлены результаты поиска среди генов белков гистидиновых триад генов, предположительно подверженных фазовой вариации по механизму инверсии.

Фазовая вариация генов у стрептококков может происходить через инверсию участков ДНК. Ранее она была показана для двух генов белков гистидиновых триад у *S. pneumoniae*, где инверсия затрагивает 3′-фрагменты двух генов. В данном исследовании гены, потенциально способные к такой фазовой вариации, были обнаружены только у *S. pneumoniae* (в 21 геноме). Инверсия происходит между двумя конкретными локусами, расположенными на расстоянии ~150 тыс. пар оснований, что позволяет комбинировать разные N- и C-концы белков и формировать четыре варианта поверхностных белков.

Кроме варьирования последовательности белка такого рода инверсия может контролировать экспрессию, если у пары генов только один промотор. 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*.

Раздел 3.4 содержит обсуждение результатов анализа белков гистидиновых триад у стрептококков.

Белки гистидиновых триад представляют собой один из факторов патогенности бактерий рода *Streptococcus*. Они широко представлены среди штаммов этого рода. 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.

Белки гистидиновых триад формируют на эволюционном дереве видовые кластеры, что подтверждает их происхождение до выделения рода *Streptococcus* и независимые дупликации в разных видах. Однако одна ветвь является исключением — она объединяет белки из разных видов с длинными ветвями.

Ранее было показано, что гены, кодирующие белки гистидиновых триад, контролируются у стрептококков цинковым репрессором AdcR и участвуют в гомеостазе цинка. 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. Вопрос о том, что представляют собой белки этой ветви, остаётся открытым. В то же время наличие ветви, регулируемой медными репрессорами, может быть следствием участия регулятора CopY в ответе не только на медный, но и на цинковый стресс, так как для этого регулятора был показан ответ на оба этих типа стресса. 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. Таким образом можно предположить формирование нового класса гистидиновых триад, взаимодействующих именно с ионами меди.

Фазовая вариация обнаружена только в случае *S. pneumoniae*, хотя

этот вид далеко не единственный патоген в исследуемом наборе данных.

Четвёртая глава посвящена исследованию специфичности транспортёров металлов семейства CorA. Она состоит из четырёх разделов.

В разделе 4.1 приведён обзор литературы, посвящённый семейству белков CorA. Описана роль катионов дивалентных металлов, в частности магния и цинка, в клетке, химические особенности этих катионов, особенности транспортных систем, способных к импорту магния в клетку. Представлен белок CorA, как один из транспортёров магния, описана его структура, характеристические мотивы, результаты исследований о специфичности этого транспортёра. Приведены способы регуляции экспресии генов транспортёров, рибопереключатели и транскрипционные факторы. Описаны различия в транспорте ионов кобальта, магния и цинка.

Раздел 4.2 содержит описание методов, использованных для анализа семейства белков CorA.

Последовательности белков семейства CorA в настоящей работе представляют собой один ортологический ряд COG0598 из базы данных EggNOG, содержащий 2545 последовательностей. Из этого ортологического ряда были удалены последовательности из архей и эукариот. Из выборки также были удалены фрагментарные последовательности. 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. HMM-профиль магниевого и кобаламинового рибопереключателей был взят из базы данных Rfam. Множественные выравнивания строились с помощью MUSCLE. Для построения филогенетического дерева всех исследуемых белковых последовательностей был использован пакет phyML. Визуализация дерева проводилась онлайн с помощью iTOL. Для предсказания рибопереключателей, контролирующих гены исследуемых белков, была использована программа Infernal. Для предсказания регуляторных мотивов была использована программа на основе библиотеки MOODS. Для предсказания последовательностей, определяющих функциональную специфичность белка, использовался инструмент SDPpred. В качестве инструмента визуализации использовалось программное обеспечение UCSF Chimera.

Раздел 4.3 содержит результаты анализа семейства белков CorA и состоит из четырёх подразделов.

В подразделе 4.3.1 описано разнообразие мотивов в семействе CorA.

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%.

В подразделе 4.3.2 рассказывается о представленности мотива GxM на филогенетическом дереве белков CorA.

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 (Рис. 7).

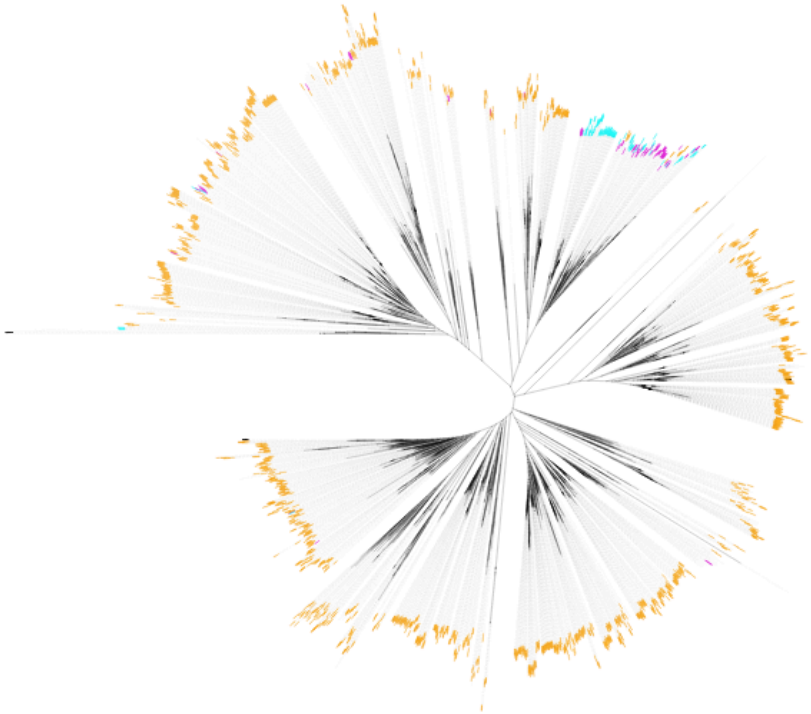

Рисунок 7. Филогенетическое дерево последовательностей транспортёров металлов семейства CorA. Цвета листьев соответствуют сигнатурному мотиву соответствующего белка: GMN — оранжевый, GVN — бирюзовый, GIN — розовый, минорные варианты мотива отмечены черным цветом.

Ветвь, на которой располагается большинство неканонических мотивов селективного фильтра, представлена в основном гамма-протеобактериями и альфа-протеобактериями, с редкими вкраплениями других протеобактерий (Рис. 8). На этой ветви можно видеть, что гены этого семейства активно передаются горизонтально. Основываясь на структуре дерева, можно предположить, что горизонтальный перенос сопровождался появлением белков с мотивами GVN и GIN в родах *Vibrio*, *Oceanicola* и *Pseudoalteromonas*.

Анализ доступных экспериментальных данных для семейства показал, что наличие мотива GMN не определяет однозначно магниевую специфичность, впрочем как и цинковую.

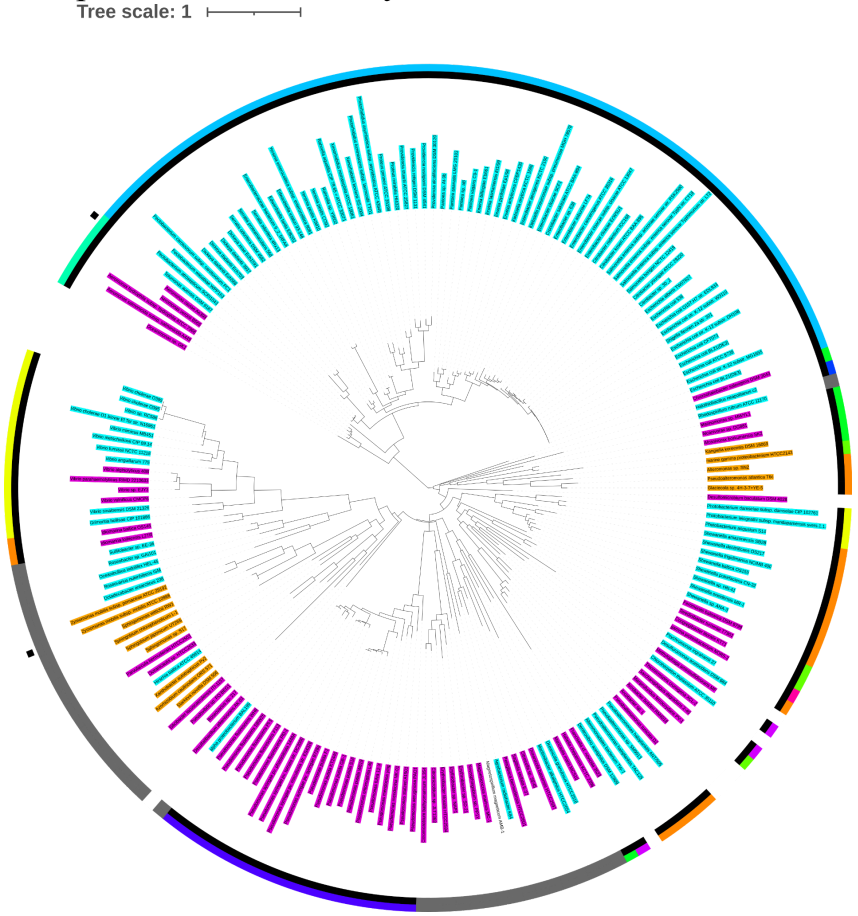

Рисунок 8. Детализация ветви дерева с Рисунка 13, которая включает большинство белков семейства CorA с неканоническими мотивами в селективном фильтре. Цвета листьев соответствуют сигнатурному мотиву соответствующего белка: GMN — оранжевый, GVN — бирюзовый, GIN — розовый, лист с другим мотивом не окрашен. 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.

В подразделе 4.3.3 рассматривается регуляция генов, кодирующих белки семейства CorA.

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, в случае кобальтовой - кобаламиновый, а в случае цинковой регуляции целью поиска были сайты связывания цинкового репрессора ZUR или AdcR у *Streptococcus*.

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. Таким образом, встаёт вопрос предсказания позиций белка, которые в действительности определяют эту специфичность.

Подраздел 4.3.4 посвящён предсказанию позиций, определяющих специфичность в белках семейства CorA.

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 (Рис. 9).

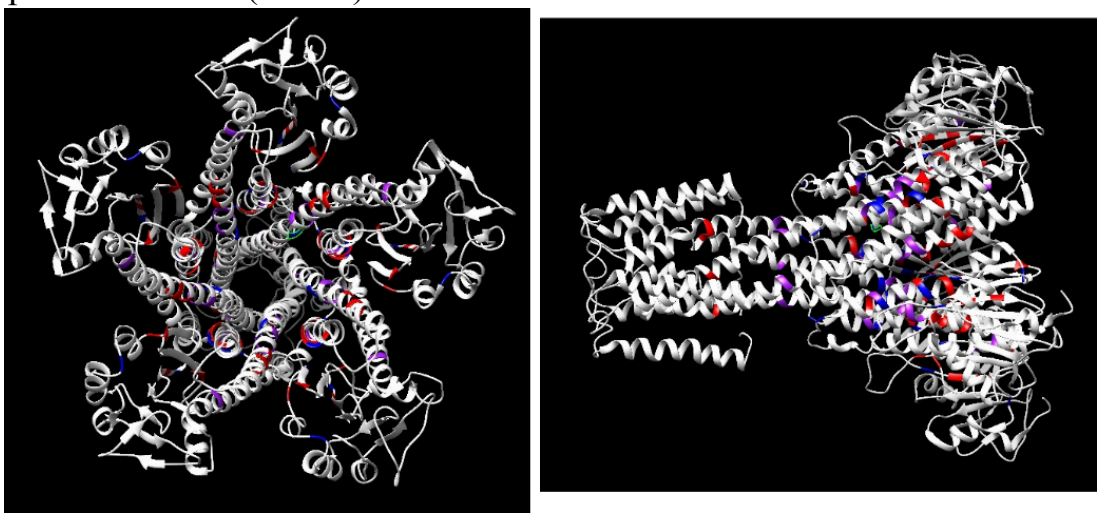

Рисунок 9. Структура пентамера CorA из *Thermotoga maritima* с разметкой предсказанных позиций, определяющих специфичность. Цвет разметки соответствует вероятности влияния позиции, наиболее вероятные позиции помечены красным цветом, далее идут синие и фиолетовые позиции.

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.

Раздел 4.4 содержит обсуждение результатов анализа семейства белков CorA.

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. 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.

Пятая глава посвящена анализу семейства белков анкириновых повторов у вольбахий. Она состоит из четырёх разделов.

В разделе 5.1 представлен обзор литературы о белках анкириновых повторов у вольбахий. Описан репертур хозяев вольбахий, а также особенности взаимодействия вольбахий с хозяевами. Дана классификация вольбахий по супергруппам. Указаны особенности геномов вольбахий, такие как малый размер, насыщенность мобильными элементами и ANK-генами. Описаны свойства последовательностей ANK-генов и известные функции их белков. Дана перспектива эволюции вольбахий в контексте адаптации к хозяину.

В разделе 5.2 представлены методы, использованные при анализе ANK-белков вольбахий.

В работе рассмотрено 159 полных геномов *Wolbachia* spp., доступных в базе данных RefSeq по состоянию на март 2023 года. Информация о хозяевах была получена из исходных метаданных сборок.

Для аннотации геномов, формирования ортологических рядов и построения дерева видов использовался инструмент PanACoTA. Для

аннотации мобильных элементов использовалась программа ISFinder. Для поиска генов, кодирующих представители семейства белков анкириновых повторов (ANK), сначала был проведён поиск доменов анкириновых повторов с помощью PfamScan.pl по PFAM записям PF00023.33, PF12796.10, PF13606.9, PF13637.9 и PF13857.9, а затем был проведён дополнительный поиск с помощью BLASTP. Множественное выравнивание ANK-белков было выполнено с использованием muscle. Кластеризация последовательностей ANK-белков была выполнена с использованием CD-hit.

В разделе 5.3 описаны результаты анализа семейства ANK-белков у вольбахий. Этот раздел состоит из трёх подразделов.

Подраздел 5.3.1 посвящён вольбахиям из разных хозяев.

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 (Рис. 10).

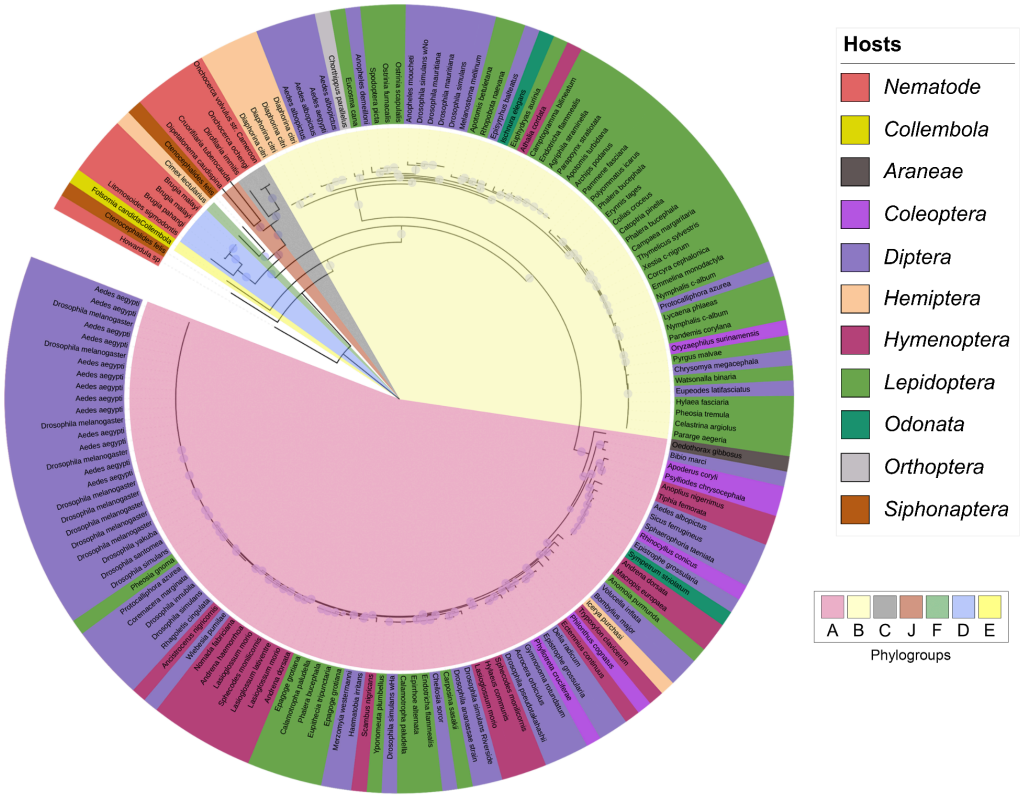

Рисунок 10. Филогенетическое дерево штаммов *Wolbachia*. Цвета листьев обозначают хозяина, цвета секторов — принадлежность к филогруппе.

В подразделе 5.3.2 показано разнообразие ANK-генов и мобильных элементов у вольбахий.

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.

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 (Рис. 11). В то же время транспозоны наблюдались по соседству с ANK-генами статистически чаще, чем в среднем по геному. Таким образом, мобильные элементы могут способствовать амплификации ANK-генов.

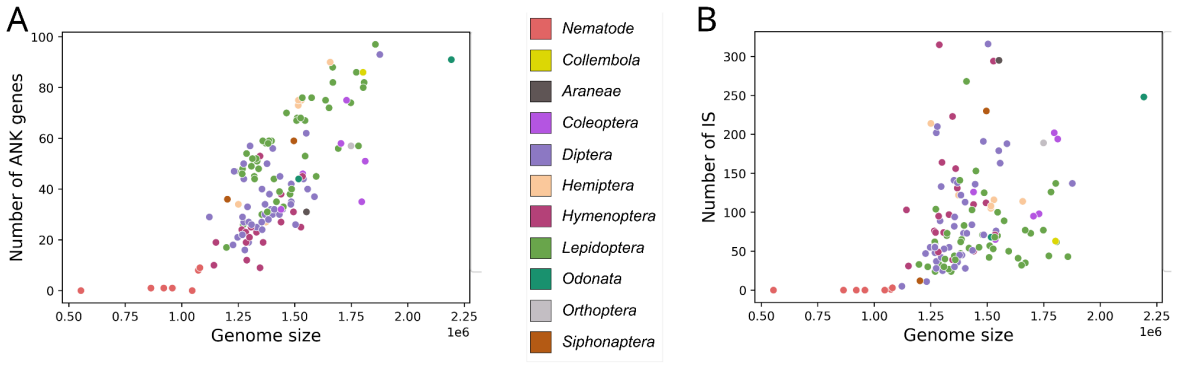

Рисунок 11. Взаимосвязь между количеством A) ANK-генов и B) инсерционных последовательностей в геноме (вертикальные оси) и размером генома (горизонтальная ось).

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. 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*.

В подразделе 5.3.3 описана связь между ANK-генами, мобильными элементами и геномным контекстом.

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. Это позволило определить, что 98% мобильных элементов и 93,5% ANK-генов были расположены частично или полностью за пределами ЛКБ (Рисунок 12А).

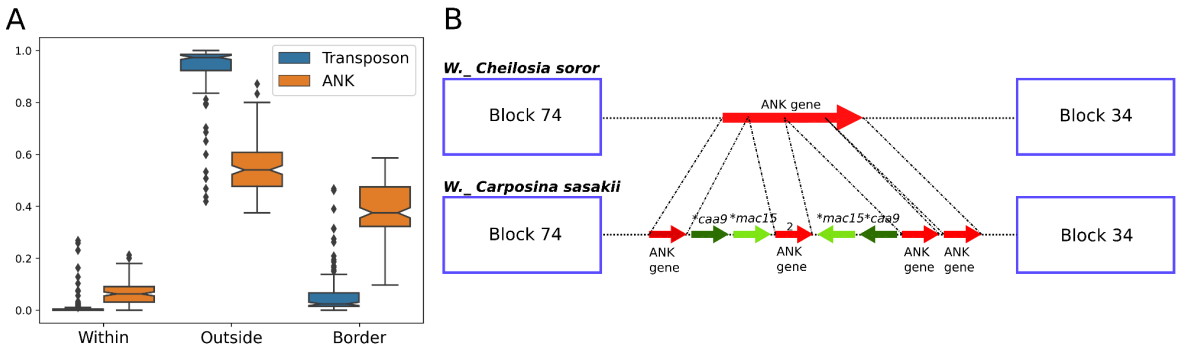

Рисунок 12. А) Положение ANK-генов и мобильных элементов относительно ЛКБ: внутри, вне или частично вне. Все попарные сравнения статистически значимы. В) Пример разрушения ANK-гена инвертированными повторами транспозаз, приводящими к образованию четырех коротких ОРС у *Wolbachia* из *Carposina sasakii*. 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 12В). 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..

В разделе 5.4 приводится обсуждение результатов анализа семейства ANK-белков у вольбахий.

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.

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. 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. 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.

Список публикаций по теме диссертации

По материалам конференции опубликованы статьи в рецензируемых научных журналах:

1. Dranenko, N., Tutukina, M., Gelfand, M., Kondrashov, F. and Bochkareva, O., 2022. Chromosome-encoded IpaH ubiquitin ligases indicate non-human pathogenic *Escherichia*. Scientific Reports, 12(1), 6868.
2. Stetsenko, A., Stehantsev, P., Dranenko, N., Gelfand, M. and Guskov, A., 2021. Structural and biochemical characterization of a novel ZntB (CmaX) transporter protein from *Pseudomonas aeruginosa*. International Journal of Biological Macromolecules, 184, pp.760-767.
3. Seferbekova, Z., Zabelkin, A., Yakovleva, Y., Afasizhev, R., Dranenko, N., Alexeev, N., Gelfand, M. and Bochkareva, O., 2021. High Rates of Genome Rearrangements and Pathogenicity of *Shigella* spp. Frontiers in Microbiology, 12, 628622.

Публикации в сборниках конференций и препринты:

1. Vostokova, Ekaterina V., Natalia O. Dranenko, Mikhail S. Gelfand, and Olga O. Bochkareva. 2023. Genome Rearrangements Drive Evolution of ANK Genes in *Wolbachia*. bioRxiv. https://doi.org/10.1101/2023.10.25.563763
2. Драненко Н.О., Бочкарёва О.О., Гельфанд М.С. 2023. Разнообразие белков высоковариабельных семейств в *Streptococcus* и *Wolbachia*. Сборник трудов 47-й междисциплинарной школы-конференции ИППИ РАН “Информационные технологии и системы 2023”.
3. Dranenko, Natalia O., Maria Tutukina, Olga Bochkareva. 2021. The Classification of *ipaH* Genes in *Shigella* and Enteroinvasive *Escherichia*. Proceedings of 10th Moscow Conference on Computational Molecular Biology MCCMB'21.
4. Dranenko, Natalia O., Maria Tutukina, Olga Bochkareva. 2021. Evolution of *ipaH* family proteins in human and non-human hosts *Escherichia*. 6th International Symposium on Systems Biology of Microbial Infections.