Московский физико-технический институт (государственный университет)
Факультет биологической и медицинской физики
Кафедра Молекулярной медицины

Диссертация допущена к защите
зав. кафедрой

_____ Лазарев В.Н.

«____» _____ 2016 г.

Выпускная квалификационная работа на соискание степени МАГИСТРА

Тема: **Сравнительный анализ ..**

Направление: 010900 – Прикладные математика и физика
Магистерская программа: 010982 – Физико-химическая биология и биотехнология

Выполнил студент гр. 0112 _____ Шашкова Т.И.

Научный руководитель,
к. б. н. _____ Лазарев В.Н.

Рецензент,
_____

Работа выполнена в ФГБУ ФНКЦ ФХМ ФМБА России
Москва – 2016

Содержание

Список принятых сокращений

UTR - untranslatate rigion (нетранслируемый регион) нт - нуклеотид ТСС - траснкрипционнй старт-сайт ТФ - траснкрипционный фактор ББВ - белок-белковые взаимодействия ТРС - трасниприционная регуляторная сеть ПТМ - посттраснляционные модификации

Введение

Актуальность работы. Организмы, даже самые простые, являются сложными системами. Актуальным вопросом является понимание, как устроены такие системы и как им удается поддерживать свою работоспособность в случае внешних и внутренних возмущений. Совокупность количественных «омикс»-ных данных дает возможность делать шаги к пониманию механизмов работы биологических систем в целом, а не только на уровне отдельных клеточных процессов.

В процессе эволюционных изменений бактерии мутируют, образуя новые виды. Как можно увидеть из анализа геномных данных, в результате мутаций выживают только некоторые организмы (образующие кластеры), которые, по-видимому, представляют собой некоторые локальные энергетические минимумы — оптимальное соотношение клеточных компонент и процессов регуляции. В течение жизни бактерии постоянно приходиться реагировать на множество внешних сигналов, в соответствии с которыми она должна моментально выносить решения и регулировать процессы своей жизнедеятельности. Таким образом, бактерии как системе можно приписать следующие свойства: устойчивость — способность адаптироваться к изменениям окружающей среды, стабильность — поддержание состояния при незначительных внутренних и внешних изменениях.

Согласно центральной догме биологии, сформулированной Френсисом Криком в 1956 году, генетическая информация в клетке реализуется по схеме ДНК $\rightarrow$ РНК $\rightarrow$ белок. В дальнейшем к этим переходам были добавлены репликация ДНК (ДНК $\rightarrow$ ДНК), репликация РНК (РНК $\rightarrow$ РНК) и обратная транскрипция (РНК $\rightarrow$ ДНК). Каждый из этих переходов является много стадийным процессом, где каждая стадия в свою очередь находиться под контролем различных факторов (белков, макромолекул). Регуляция

всех этих механизмов на каждом этапе и есть результат обработки внешних сигналов — выбранная стратегия поведения клетки/организма. Количественными показателями, по которым можно проводить наблюдения и строить поведенческие модели, будут являться уровни представленности клеточных компонент: транскриптов, белков, метаболитов. По данным показателям можно сравнивать стратегии поведения организмов в тех или иных условиях, а также сравнивать близкородственные организмы между собой.

В текущей работе в качестве объекта исследования была выбрана бактерия *H. pylori*. Данная бактерия пользуется большим интересом среди ученых за счет своих особенностей таких как, высокая гетерогенность, патогенность, вирулентность и обилие систем рестрикции-модификации, кроме того данная бактерия ассоциирована с воспалительными заболеваниями желудка. На примере данной бактерии, проводится анализ соотношения протеомных и транскриптомных данных у штамма *H. pylori* A45 и трех его изогенных мутантов по генам систем рестрикции-модификации, а также штаммов 26695 и J99, с целью выявления возможных соотношений между белками и транскриптами внутри организма и попыткой объяснить почему в каких-то случаях происходить регуляция на уровне транскриптома, а в других — на протеомном.

По результатам работы, можно сказать, что ...

Цель работы Определить возможные количественные комбинации сочетания представленности транскриптов и белков характерных для бактерии вида *H. pylori*.

Задачи работы

- Анализ транскриптомных профилей
- Анализ протеомных данных
- Сопоставление траснкриптомных и протеомных данных

- Комбинация результатов с картой белок-белковых взаимодействий

Глава 1

Обзор литературы

1.1. Бактерия *Helicobacter pylori* — общая информация

Helicobacter pylori это грам-отрицательная микроаерофильная бактерия, которая колонизирует желудок, нетипичную высоко кислотную среда для микроорганизмов. В 1983 году Робин Уоррен (Robin Warren) и Барри Маршалл (Barry Marshall) нашли, что инфицирование этой бактерией связано с воспалением слизистой оболочки желудка человека и болезней двенадцатиперстной кишки [**?**], позже бактерия была признана в качестве одного из основных факторов риска развития рака желудка . Всемирная организация здравоохранения классифицирует хеликобактер как канцероген I класса. Бактерия представлена примерно у 50% населения, но болезни желудка развиваются меньше, чем у 10% носителей [**?**].

Геном эталонного штамма *H. pylori* 26695 был полностью секвенирован в 1997 году [**?**] и кодирует 1587 белков, около 950 (61%) которых имеют определенные функции (за исключением "гипотетических белков"; UniProt, CMR [**?**]). Эти цифры показывают, что значительная часть белков хеликобактера не охарактеризована функционально.

Интенсивно изучается патогенность бактерии *H. pylori*, особенно факторы вирулентности: эффектор CagA, цитотоксин VacA, IceA и BabA, а также ее адгезины и уреазы [**? ? ?**]. Последние позволяют бактерии нейтрализовать кислую среду за счет производства аммониа. Выживаемости в желудке зависит от согласованной экспрессии факторов вирулентности и генов «домашнего хозяйства», которые позволяют *H. pylori* выдерживать нагрузки, налагаемые суровой кислой средой и противодействовать ответу хозяина. Ограничен-

ность обитания *H. pylori* в желудке связана с пониженной функциональной избыточностью ее небольшого генома (∼1,6 Мб), характеризующегося ограниченным количеством транскрипционных регуляторов [**?**]. На сегодняшний день получен большой объем метаданных о бактерии *H. pylori*, что позволяет провести комплексный биоинформатический анализ и разобраться в механизмах регуляции.

1.2. Транскрипция

Процесс транскрипции является первым этапом на пути реализации генетической информации. У бактерий транскрипционная единица (TE) представляет собой упорядоченный набор генетических структур: регуляторный регион (upstream), траснкрипционный старт-сайт (TCC), 5'-нетранслируемая область (5'UTR), одна или более открытых рамок считывания, 3'-нетранслируемая область (3'UTR), и сайт терминации транскрипции (CTT). Регуляторные регионы содержат в себе промоторные последовательности, а также некоторые из них имеют сайты связывания с транскрипционными факторами. Промоторные последовательности распознаются ДНК-зависимая РНК-полимеразой и, как правило, состоят из двух консервативных доменов, располагающихся на -10 и -35 позиции относительно TCC.

Комплекс ДНК-зависимая РНК-полимераза был охарактеризован в 1960 году Стененом (Stenes), Баргессом (Burgess) и коллегами [**?**]. Основная часть РНК-полимеразы состоит из субъединиц $\beta\beta'\alpha_2\omega$, и такая форма способна синтезировать РНК, но не в состоянии найти промоторы. Для инициации транскрипции необходима σ-субъединца (σ-фактор), которая несет в себе основные детерминанты для распознавания промоторов. Каждому σ-фактору соответствует своя консервативная промоторная последовательность. Форма полимеразы $\beta\beta'\alpha_2\omega\sigma$ (Рис. 1.1) известна как голоэнзим и является компетент-

ной для инициации транскрипции.

Рис. 1.1. Модель, основанная на кристаллографическом исследовании связывания РНК-полимеразы(голоэнзима) с промотором [**?**].

У бактерий σ-факторы разделены на две основные филогенетические группы: σ^{70} и σ^{54}. Основной является σ^{70}, которая отвечает за транскрипцию большинства генов «домашнего хозяйства» в нормальных условиях. У *H. pylori* также представлена дополнительная σ-субъединица - σ^{28}. При этом данная субъединица имеет анти-σ^{28} фактор, который ингибирует активность белка, поэтому активность самого σ-фактора, в частности, зависит от соотношения σ/анти-σ и механизмов диссоциации этого комплекса [**?**].

1.3. Транскриптом

Совокупность всех транскриптов в организме называется *транскриптом*. Транскриптомы бактерий различаются по представленности в них транскриптов в зависимости от организма и от условий окружающей среды. Сравнение представленности транскрипотов проводят на основе данных RNA-seq и при помощи биоинформатический методов анализа. Прежде чем сравнивать уровень представленности транскриптов, необходимо определить их структуру. Как правило, для большинства бактерий известна только структура гена, позиция старт- и стоп-кадонов, и нет информации о TCC и TTC.

В своей работе Синтия Шарма (Cynthia Sharma) с коллегами [**?**] исследовала транскриптом штамма *H. pylori* 26695. В работе представлен новый метод поиска TCC, который заключается в сравнении транскрипционных профилей в двух библеотеках: обработанной и необработанной экзонуклеазой. Экзонуклеаза отрезает 5'-фосфаты (5'P) и не трогает 5'-трифосфаты (5'PPP). Т.е. рРНК и тРНК не подвержены разрушения, а у остальных остается по

несколько нт с 5'-конца и при сравнении покрытия будет наблюдаться пик в 5'-областях. Для того чтобы найти наибольшее количество возможных ТСС, бактерии выращивались в пяти различных средах (кислотный стресс, с клетками эпителия желудка, с клетками печени, просто клеточная культура, клетки в средне логарифмической фазе роста). В качестве проверки достоверности ТСС проводился поиск промоторных последовательностей, которые должны располагаться на соответствующих позициях относительно ТСС. Найденные ТСС классифицировались в зависимости от их расстояния до начала гены:

- первичные ($<$ 500 нт от старт-кодона)
- вторичные/альтернативные ($<$ 500 нт от старт-кодона, но уровень покрытие ниже, чем у первичных)
- внутренние (находятся в генах)
- антисмысловые (на противоположенной цепи от гена и $>$ 100 нт от конца гена)
- орфаны (на расстоянии $<$ 500 нт гены не обнаружены)

Таким образом, для штамма *H. pylori* 26695 составлен каталог из 1907 различных ТСС для 717 генов. На основе найденных ТСС исследователи сделали переаннотацию некоторых генов, а также определили новые и альтернативные опероны, найдены новые гены и короткие РНК.

Еще один метод обнаружения ТСС представлен в работе Павла Мазина (Pavel Mazin) с коллегами [**?**] по исследованию транскриптома *M. gallisepticum*. Метод основывается на статистическом анализе транскрипционных профилей, результатом которого является набор предполагаемых ТСС. Затем ТСС отбираются и корректируются так же, как и в предыдущей работе, на основе наличия и расположения промоторных последовательностей. Помимо этого, описанные в работе метод дает возможность определить ТТС и структуру оперонов.

ТСС дает нам информацию о том, где располагаются промоторные последовательности, границы 5'UTR и где искать ТФСС. В следующем разделе описаны механизмы регуляции транскрипции и как нуклеотидные замены в регуляторных областях могут повлиять на уровень экспрессии генов.

1.4. Регуляция транскрипции

Регуляция транскрипции происходит осуществляется специальными белками и макромолекулами. Минимальная регуляторная система транскрипции представляет собой взаимодействие транс-факторов, которыми являются в частности транскрипционные факторы (ТФ), с цис-регуляторными элементами. Цис-элементы закодированы в плазмиде или хромосоме по соседству с геном, в то время как транс-факторы - это диффундирующие молекулы, которые могут связываться с ДНК. Изменения в ТФ, цис-регуляторных элементах и целевых генах дает бактерии естественную способность адаптироваться к изменениям окружающей среды.

ТФ классифицируются по группам в зависимости от их структуры. Классификация базируется хотя бы на двух доменах, которые позволяют ТФ функционировать как регуляторные элементы (свитчи). Один домен - это сигнальный сенсор, который осуществляет связь с белками и/или лигандами. Другой домен ответственный за переключение, направленно взаимодействующий с целевой ДНК или сайтом связывания транскрипционного фактора (ТФСС).

Эффективность регуляции зависит от концентрации ТФ и аффиности к ТФСС: слабые ТФ требуют высокой концентрации и наоборот. Когда промотор полностью соответствует консервативной последовательности сайта связывания (*сильный промотор*) и ТФ в изобилии, тогда скорость транскрипции четко определена [**?**]. В противном случае, транскрипция приобретает некий

“шум”. Есть два источника шума: внутренний и внешний. В гипотетической клетке с двумя идентичными генами внутренний шум будет вызывать различный уровень их транскрипции. Внешний шум возникает из-за вариаций клеточных компонент (cell-to-cell), например, количества полимераз[**?**].

Замечено, что локальные ТФ (располагаются рядом с целевым геном), как правило, имеют высокую аффиность. Глобальные ТФ менее специфичны и могут связываться с большим набором ТФСС и должны иметь высокий уровень экспрессии.

В общем случае, негативные регуляторы связываются с промотором, в то время как положительные регуляторы "садятся"на upstream-регином промотора.Некоторые ТФ одновременно считаются и активаторами и репрессорами. К примеру, ТФ связывается в межгенном регионе между двумя ТЕ, которые транскрибируются в разном направлении, и регуляция имеет разный эффект для каждой (пример: sugar catabolism loci). Альтернативная схема, при которой работает двойная регуляция, когда есть два ТФСС для одного ТФ, слабый отрицательный сайт внутри промотора и сильной положительной сайта рядом с ним, тогда ответ зависит от концентрации данного ТФ.

Количество генов, кодирующих ТФ, возрастет с общим количеством генов. В бактериальных геномах это квадратичная зависимость [**? ?**]. Также, в маленьких геномах гены чаще кластеризованы в опероны [**?**]. Хотя, есть доказательства, что среднее количество ТФСС на регуляторных регион не зависит от размера генома [**?**].

1.4.1. Регуляция инициации транскрипции

Простые типы активации транскрипции. К этому типу относятся варианты активации для которых необходима только одна активирующая молекула. Существует два механизма, по которым может быть осуществлена

активация: активатор может изменить конформацию промотора ДНК для улучшения качества промотора или взаимодействовать непосредственно с РНК-полимеразой для компенсации дефектов промотора.

Изменение конформации. Лучше всего для понимания этого механизма подходит рассмотрение ТФ семейства MerR [**?**]. В большинстве случаев они садятся на ДНК между -10 и -35 позицией целевого промотора. Целевые промоторы для этих ТФ диффектны, так как имеют неоптимальное расстояние между -35 и -10 элементами. После посадки ТФ, ДНК дифформируется и расстояние между элементами изменятся так, что теперь полимераза может сесть на ДНК и начать транксрипцию.

Взаимодействие с РНК-полимеразой. В большинстве случаев когда промоторам для активации необходим единичный ТФ, ТФ связывается с ДНК либо перед либо перекрываясь с -35 элементом так, что связанный активатор может напрямую соединиться с РНК-полимерзой, когда она сцепляется с промотором. Примечательно, что, независимо от того, является ли полимеразное связывание зависимо или независимо от активатора, конечный организация открытого комплекса по-видимому, аналогична.

Доказательством существования данного факта, является наличие положительных мутаций, которые приводят к единичным ак заменам, в следствии чего снижается или полностью пропадает способность активатора связываться с РНК-полимерзой, при этом не затрагивая других функций, таких как связывания с ДНК или регуляция. Положительные мутации были использованы для идентификации аминокислотных боковых цепей в активаторах, которые необходимы для непосредственного взаимодействия с полимеразой [**? ?**].

Промотры зависимые от σ^{54}: парадигма вторичной активации. Транскрипция с участием фактора σ^{54} не может происходить без участие ТФ-активатора. Около 60% бактерий имеют гены, кодирующие σ^{54}-схожие белки. Ключевые элементы таких промоторов располагаются на позициях -12 и - 24. РНК-полимераза распознает эти промоторы, но не может сформировать устойчивые открытые комплексы без вмешательства активатора, который содержит домен AAA+, который в свою очередь взаимодействует непосредственно с σ^{54} (Рис. 1.2) [**?**]. Этот тип активаторов известен как enhancer-binding protein и, как правило, содержит 3 домена: регуляторный - отвечает на специфические метаболические сигналы, AAA+ - активирующий домен, ДНК-связывающий домен.

Рис. 1.2. Активация транскрипции σ^{54}-зависимого промотора [**?**].

На данный момент модель взаимодействия описывается следующим образом: АТФ гидролиз приводит в движение петли (surface-exposed loops) в домене активатора, чтобы она взаимодействовала с σ^{54}, и образовался нормальный открытый комплекс [**?**]. Это было названо *второй парадигмой активации* транскрипции, так как домен-активатор нацелен непосредственно "нажать необходимую кнопку" для удаления препятствия.

Уровень σ^{54} не сильно флуктуирует во время смены фаз роста, поэтому регуляция уровня транскрипции происходит за счет белков-активаторов, чья активность модулируется различными механизмами в ответ на изменение метаболизма или условий окружающей среды [**?**]. Это находится в резком контрасте с большинством других альтернативных σ-факторов, чья деятельность регулируется либо их количеством, либо степенью доступности, без вмешательства транскрипционных факторов.

Регуляция, связанная со структурой РНК-полимеразы Суть данных механизмов заключается не в изменении промотора, а в перепрограммировании предпочтения холоэнзима полимеразы к промотору. Например, когда в ответ на конкретный сигнал, фактор "домашнего хозяйства" меняется на альтернативный σ-фактор, тем самым изменяется специфичность к промотроной последовательности пропорционально количеству клеточных полимераз.

Механизмы, которые фокусируются на РНК-полимеразе, больше чем на промоторе, являются регулируемыми способами выбора, когда бактерия должны реагировать эффективно. Внезапный тепловой шок требует мгновенного реагирования, чтобы избежать гибели, и это может быть сделано при помощи смены σ-фактора РНК-полимеразы, в то время как выбор между лактозой и арабинозой не так критичен и может регулироваться ТФ.

Маленькие лиганды. Маленькие лиганды осуществляют механизмы, по которым РНК-полимераза быстро и эффективно реагирует на сигналы окружающей среды. Хорошим примером является гуанозин-3'5'-бифосфат (ppGpp), который синтезируется когда доступность ак ограничена при условии, что трансляция также ограничена. ppGpp работает на дистабилизацию открытого комплекса у промотора. По факту, хотя взаимодействие ppGpp с РНК-полимеразой не специфично к промотору, ppGpp зависимое ингибирование происходит только на тех промоторах, которые формируют нестабильные открытые комплексы. Такие промоторы, как правило, имеют GC-богатые участки рядом с $+1$ позицией и нужны для контроля большинства генов, которые кодируют продукты, необходимые для трансляции. Такие промоторы слабо функциональны при низкой концентрации инициирующих нуклеотидов, обычно АТФ. Поэтому, можно предположить, что ppGpp контролируют экспрессию трансляционной машинерии в ответ на внезапное "голодание", тогда

как АТФ доступность контролирует экспрессию в ответ на темпы роста.

1.4.2. Система рестрикции-модификации

Первоначально, системы рестрикции-модификации (РМ) были описаны, как механизм борьбы с чужеродной ДНК бактериофагов. РМ система состоит из двух ферментов: метилтрансфераза — катализирует добавление метильной группы от донора S-аденозин-метионина к аденину или цитозину, и эндонуклеза рестрикции — разрывает фосфодиэфирную связь в ДНК. Оба энзима узнают специфические сайты в нуклеотидной цепи, и метилирование такого сайта предотвращает разрезание ДНК. РМ системы классифицируются по трем группам: тип I, II и III, в соответствии с их составом, требованиям к кофакторам, структуре их сайтов распознавания и способом действия. Позднее был добавлен IV класс, который кодируется одним или двумя генами, которые в свою очередь являются метил-зависимыми рестриктазами.

Остановимся на первой активности данных систем — метилировании, так как с недавних пор, метилирование относят также к механизмам регуляции транскрипции. В случае когда ДНК заметилирована, полимераза не может сесть на ДНК и/или ее продвижение по ДНК затрудняется, в таком случае экспрессия заметилированных генов падает. Геном хеликобактера известен своим большим числом РМ систем и их разнообразием и штаммовой специфичностью. Так, например, в первом отсеквенированном штамме 26695 было идентифицировано 26 предположительных метилтрансфераз [**?**], в штамме J99 было также найдено 26 возможных РМ систем [1], в штамме HPAG1 — 30 РМ систем [**?**].

Исследования по сравнению метилома штаммов 26695 и J99 показывают, что ДНК *H. pylori* являются высоко метилированными, и паттерны метилирования сильно различаются между штаммами[**?**]. Анализ 50 штаммов *H.*

pylori показал, что разнообразие метилтрансфераз является достаточно высоким, для того чтобы использовать статус метилирования генома в качестве инструмента типирования штаммов[1].

Метилирование также привлекает к себе внимание за счет того, что является одной из причин возникновения мутаций. Во-первых, из-за метилирования в ходе рекомбинации ДНК возникают ошибки копирования. Во-вторых, 5mC (метилированный цитозин в 5 положении) подвержен реакции спонтанного дезаминирования, в результате которой цитозин превращается в урацил. Соединение урацила с аденином во время репликации приводит к итоговой замене цитозина на тимин. Данный процесс контролируется специальным ферментом — урацил-ДНК-гликозилазой, который предотвращает такие замены, но несмотря на это, замены C→T являются доминирующими у бактерий. Взяв во внимание тот факт, что РМ системы передаются в процессе эволюции между бактериями, количество мутаций в сайтах рестрикции-модификации зависит от времени присутствия соответствующей РМ системы в бактерии.

1.4.3. Транскрипционные регуляторные сети

Транскрипционные регуляторные сети (ТРС) отображают взаимодействия между ТФ и целевыми генами. Сети показывают общую организацию транскрипционной регуляции: модульные и иерархические структуры. Сети состоят из структурных единиц - *регулонов*: набор целевых генов совместно регулирующихся некоторым набором ТФ. Внутри или между регулонами, ТФ и целевые гены связаны в специфичные локальные структуры, называемые мотивами. Они позволяют системе отвечать со специфической динамикой, соответствующей природе полученного сигнала. Интересно, что некоторые ортологичные гены могут быть включены в разные типы мотивов. Регуло-

ны делятся на простые и сложные в соответствии с тем, включают они в себя один или несколько ТФ. Большинство регулонов в бактериях являются сложными. Исследования показывают, что ТФ эволюционируют намного быстрее, чем их целевые гены. Предполагается, что ТРС у бактерий пластичны, динамичны и быстро адаптируются к изменениям окружающей среды [**?**]. Данные явление получило название *regulatory rewiring* [**?**]. На изменение ТРС могут повлиять такие эволюционные процессы, как дупликация и горизонтальный перенос генов. Например, потеря и дубликация ТФ и ТФСС могут быть причиной расширения, сужения, слияния, деления и даже создания и разрушения регулона[**?**]. Большинство ТФ *E. coli* имеют паралоги, которые с большой вероятностью являются результатом горизонтального переноса (ГП) [**?**]. Более того, кажется, что в случаях гор переноса на близких филогинетических расстояниях, трансфер локальных ТФ происходит легче, чем глобальных ТФ [**? ?**]. Уровень экспрессии недавно перенесенных генов низкий, вероятно это сигнал медленной интеграции перенесенного гена в существующую регуляторную сеть [**?**].

Helicobacter pylori является очень привлекательной системой для исследования того, как устроены бактериальные ТРС и как бактерии удается поддерживать инфекции в организме хозяина. Идентификация определенных мотивов и организация ТРС в *H. pylori* может значительно помочь в понимании регуляции у других человеческих микробных патогенах.

В исследовании Альберто Даниелли (Alberto Danielli) [**?**] с коллегами было найдено 224 направленных взаимодействий, разделенных соответственно по ТФ-сенсорам. Но в итоге для *H. pylori* зарегистрировано только 17 достоверных ТФ (Таблица **??**), охватывающих четыре основных регулона, которые связаны с ключевыми физиологическими реакциями, необходимыми для колонизации желудка: 1) тепловой стресс; 2) подвижность и хемотаксис; 3) кислотный стресс; и 4) гомеостаза металла.

Таблица 1.1. Таблица транскрипционных факторов, идентифицированных у *H. pylori* [**?**]
.

H. pylori не имеет четко разделенных модулей регуляции, вместо этого все ТФ переплетены между собой. Данные показывают, что ТРС хеликобактера однозначно построена для поддержания гомеостаза. Они не приспособлены для адаптации к большинству сигналов окружающей среды, и, по-видимому, не достаточно гибкие, чтобы реагировать на метаболические сигналы, возникающих за пределами желудка.

Тепловой шок и стресс-ответ Регулон теплового шока - наиболее понятный регуляторный модуль у *H.pylori*. В то время как грам-отрицательные бактерии используют дополнительную субъединицу сигма 32, *H.pylori* использует другую стратегию, которая включает в себя два реперссора - HspR и HrcA. Два данных ТФ направленно блокируют три основных оперона, включая groESL ген-шаперон. Все три оперона отвечают за тепловой шок и актируются в присутствии неправильно свернутого белка или стресс-сигнала. HspR в одиночку блокирует транскрипцию cbpA - оперона, тем самым отрицательно авторегулирует свой синтез. Напротив, оба ТФа нужны для двойной блокировки опреонов groESL и hrcA (Рисунок А). Активность обоих репрессоров зависит от производства целевого гена groESL, предполгая что ТФы связаны с GroE системой шаперонов при помощи регуляторной петли ОС.

Несмотря на то, что оба ТФ необходимы для полной регуляции, эксперименты in vivo показывают, что их связвание с ДНК происходит не зависимо. Последовательная посадка двух независимых, но биохимически функционально взаимосвязанных регуляторов может быть объяснена при помощи логической схемы- FFL. По факту, модуль теплового шока - это пример некогерентной FFL (dissruppted FFL). Некогерентные FFLs значительно ускоряют кинетику реагирования регулирующего каскада на том же стационарном

уровне

Модуль биосинтеза жгутиков. Гены кодирующие жгутики, хемотаксис, и моторные белки - основные вирулентные факторы в *H. pylori*. Их делеция приводит к ослаблению (attenuated) инфекции на модельных животных, возможно из-за неспособности двигаться в ответ на благоприятные или вредные градиенты веществ.

Как и у других бактерий, гены жгутиков имеют положительную регуляцию и иерархически организованы в три основных класса в соответствии с необходимым для них сигма-фактором:

- класс I - охватывает генны транскрибируемые вегетативной sigms70-содержащий РНК-полимеразе (РНКП), и включает в себя в основном жгутик-регуляторные гены (flgR, flgS, rpoN, flhA);
- класс II - гены, регулируемые sigma54 (RpoN) и кодирующие компоненты жгутиков и крючки;
- класс III - гены, кодирующие последнию структуру (late flagellar structure) жгутиков (сигма28 — fliA).

У *H. pylori* отсутствует единый руглятор, схожий у всех энетробактерий - FlhDC, вместо этого у нее есть каскад, регулируемый сигма-единицами, инициированный сигма 70, и где каждый сигма фактор активирует соответствующей целевой ген по схеме одиночного входящего мотива (OBM). Таким образом, регуляция цепи биосинтеза жгутиков гарантирует корректную последовательную экспрессию генов для ранних, средних и поздний компонент. Помимо того, у H.Pylori есть анти-сигма единица (aniti-sigma28) - flgM, саматранскрибирующийся, (,?) как класс II, вовлечена в ингибирование FliA. Кроме того, компоненты базального тела, FlhA и FlhF, по-видимому, модулируют rpoN- и fliA- зависимую транскрипцию средних и поздних жгутиковых генов [32]. Кроме того, HP0958 белок, описанный как RpoN шаперон, действу-

ет как фактор, способствующий распаду комплекса при траснкрипции мРНК FliA [43]. Таким образом, хотя точные молекулярные механизмы полностью не изучены, ясно, что, по аналогии с регулоном теплового шока, контролирующего уровень GroESL, белок-белковые взаимодействия вспомогательных факторов с ТФ являются центральными для обратной связи в модуляции регулирующей биосинтез жгутиков. В добавок, выход модуля HspR-HrcA опосредованно оказывает влияние на моторику, скорее всего, через изменения флагеллярного биосинтеза, происходящие в штаммах, где внутриклеточный уровень шаперонов являются дерегулированы [+47]. Это связывает цепь жгутикообразования с ответом на стресс.

Кислотный стресс. Способность расти в суровых кислых условиях, встречающихся в желудке является отличительной особенностью HP и связана с вирулентностью. Соответственно, регулируемая экспрессия специального набора, так называемых кислотных генов акклиматизации (уреазы оперон ure, aliphatiic Амидазы AMIE и amiF, аргиназы ROC, и т.д.), позволяет *H.pylori* сохранить кислотность бактериальной периплазмы близкой к нейтральной, и поддерживает физиологический уровень pH в цитоплазме в присутствии мочевины и активности уреазы. Транскрипции опреона кислотного стресса находится под контролем сигма80 и регулируется спецальным фактором необходимым для ответа на стресс - ArsR. ArsR саморгеулируется и закодирован в опреоне, который также охватывает родственную трансмембранную гистидинкиназу ArsS. Было предложено, что сигнал воспринимаемый ArsS - подкисление периплазмы, трансдуцирует путем внесения изменений в протонирования нескольких остатков гистидина (pKa 6,0), включенныя в сигнальную область внешненей цитозоли[55]. Этот стимул вызывает фосфорилирование ArsR, способствуя тем самым его ДНК-связывающей активности к определенному набору промотеров. Несмотря на то, что ArsR необходимый

ген, мутантные штаммы, где этот ген отсутсвет или теряет активность, тоже жизнеспособны. Несмотря на некоторые расхождения в экспериментальных наборах данных из нескольких систематических исследований, возможно, вытекающих из использования различных штаммов [53,55-57,59,60], то новые данные указывают на существование различных целевых нормативов, которые контролируются в соответствии со статусом фосфорилирования ArsR: Первый кластер: гены P ArsR- зависимого оперона, регулируемого ArsR в зависиомости от статуса фосфорилиорвания, по сигналу ArsS (omp11, carbonic,anhydrase, hypA, ureAB) гены целевого оперона, который регулируется более жесткими кислотными условиями, приводящими к подкислению цитоплазме. Их регуляция P ArsR-зависимая и фосфориолирование также обеспечивает высокое сродство к промтору ДНК. Несмотря на то, что их регуляция не нарвуашется при отстутвии гена ArsS, и может зависит от трансдукции сигнала, приводящего к фосфорилировнию, необходимому для активации. Эта группа содержит главные гены для ответа на стресс(amiE, amiF и др.) гены, зависящие от нефософрилировного ArsR (вклюая сам ген) и их регуляция не обязательно зависит от pH среды. Функции генов неизвестны, но они важны для жизнедеятельности.

На данный момент не выялена консенсусная последоватльность связвания ArsR с ДНК, но он может быть охарактеризован шириной сайта и нт последоватленостью. Это обосновывает гипотезу о двусторонность ArsR регулона, который управляет через различные Sims транскрипцией различных наборов генов, в зависимости от состояния фосфорилирования регулятора. Еще одной сложностью является метосположение посадки ArsR на ДнК, в зависиомости от того, до или после прсотора садится белок, он может быть активатором или репрессором транскрипции. “Очень интересно, недавняя работа определила FlgS как гистидинкиназу принадлежащую к модулю биосинтеза жгутиков, как также необходимсу для выживания *H.pylori*, при низ-

ких значениях pH [61]. Хотя это и не известно может ли FlgS вызвать ArsR фосфорилирование при подкислении цитоплазмы,но он может быть хорошим кандидатом в качестве датчика цитозольной кислотности для ArsR регулона.

Регулон гомеостаза метала. Во многих бактериальных патогенах, в том числе хеликобактере, металл-голодание вызывает экспрессию факторов вирулентности, что позволяет им конкурировать с хозяином за эти важные питательные вещества. С другой стороны, металлы - это токсины, если находятся вне клеток в высокой концентрации. Таки образом, их гомеостаз должен быть под тщательным контролем. У HP их регуляция происходит при помощи двух компонентной системы CrdRS: fur (ferric intake regulon = регулон усовения металла) вовлечен в гомеостаз металла и гомолога Ni-responosive регулятора NikR E.coli. При этом только определенные целевые гены CrdRS системы участвуют в устойчиовсти к меди, Fur и NikR описаны как плеотропическике регуляторы (т.е. влияют на несколько процессов).

Fur регулирует гены, участвующие в обоих процесах: поглащение Fe2+ и детоксификация. Кофактор иона металла может действовть как ко-репрессор (holoFur) или как индуктор (apoFur). Таким образом, информация о присутствии или отсутствии ионов металла может быть траслирована двумя разными способами. Эксперемнетально показа многофункциональность Fur, это высоко предатсвленный белок и может связываться еще с 200 генами, которые находятся под действием других регуляторов (rpoN, flgR, flgS, cheA, nikR) [74, 75]. Соответнно, примерно 100 генов разрегулированы Fur, которые не являются его целевыми генами.

С другой стороны, NikR медиатор регуляции Ni2+ гомеостаза в клетке, главный в активности фермента никеля уреазы. apoNikR не может связываться с ДНК, Ni2+ имеет высокое сродство с металлическим сайтом связывания и приводит к аллостерическим изменениям активности связывания holoNikR

с ДНК [76-78]. NikR также может быть активатором и репрессором в зависимости от места посадки на ДНК.

Вместе эти белки образуют bifan motif, который связан с метаболизмом мтелла. горизонатльной и симметричной логикой (Рис D). Логика гомеостаза метала в дальнейшем дополняется многокомпонентной петлей (Рис B). По этому сценарию, регуляция общих целей переходит в схему FFL, когда концентрация кофактора, регулирующего связывания с ДНК обоих факторов, достигает порогового уровня. Эти ТФ также связаны с регулоном ответа на тепловой стресс, так как в отсутствии mikR этот оперон дерегулирован. Сайт связывания NikR обнаружены в регуляторных областях генов hspR и hspA, закодированных в опероне, предположительно регуляция ответа на стресс находится под контролем транскрипции NikR. Fur оказывает положительное влияние на набор генов [74], включающих гены важные для хемотаксиса и мобильности, взаимодействия в хозяином (fibfibBP) и redox равновесия (oor). Это говорит о интеграции прямой связи между модулем отклика металла, вирулентности и схемы жгутиков, которые, безусловно, заслуживают дальнейшего изучения.

1.5. Трансляция

Процесс трансляции также как и транскрипция проходит в три этапа и помимо основных компонентов — рибосом, тРНК и мРНК, включает в себя многочисленные лиганды и ионы, нуклеотиды и белки, такие как тРНК-модифицирующие ферменты, аминоцил-тРНК-синтетаы, а также белков, связанных с рибосомами — факторы инициации, элонгации и терминации.

Процесс трансляции начинается с формирования рибосомального комплекса на на 5'-конец мРНК. Сначала, факторы инициации связываются с малой субъединицой рибосомы. Этот комплекс и метионин-тРНК связывают-

ся с мРНК в специфических сайтах (последовальность Шайна — Дальгарно) рядом со старт-кодоном (AUG) и формируют комплекс инициации. После этого с комплексом связывается большая рибосомальная субъединица. Далее начинается стадия элонгация, на которой в соответствии с генетическим кодом происходит последовательная встройка аминокислот, которые доставляются тРНК и присоединяются к синтезируемой цепи при помощи рибосомального комплекса и факторов элонгации. Таким образом рибосома продвигается по мРНК в направлении 5' $\rightarrow$ 3' до тех пор, пока не дойдет до стоп-кодона. Триплету стоп-кодон не соответствует ни одна тРНК, а вместо этого присоединяется фактор терминации и синтезированная цепь высвобождается из рибосомы. *H. pylori* содержит два набора из 23S-5S-16S рРНК генов в дополнение к одному 5S рРНК гену и один структурный РНК ген (ssrD). Всего 36 тРНК генов были идентифицированы в бактерии. Зрелая 70S рибосомальная единица состоит из двух отдельных субъединиц: 50S и 30S. В общей сложности 21 рибосомальный белок соответствует 30S субъединице, и для 50S в свою очередь был найден 31 ортолог.

Большая часть трансляционного механизма *H. pylori* аналогична другим бактериям, за исключением отсутствия двух тРНК синтетаз.

1.6. Синтез и деградация белков

В отличии от транскриптов, чья представленность в бактериях по большей части зависит от скорости их синтеза, так как транскрипты являются короткоживущими молекулами, концентрация белков зависит не только от скорости синтеза, но и от скорости деградации, а также непосредственно от представленности необходимых мРНК.

Разберем подробнее эти вопросы. В первую очередь для синтеза белка необходима мРНК. Но не всякой мРНК суждено стать белком, помешать это-

му процессу могут регуляторные РНК. Такие РНК имеют участки комплиментарные мРНК, и в итоге РНК «скливаются» в этих местах друг с другом. Результатом может быть полная блокировка дальнейшей трансляции за счет закрытия сайта связывания с рибосомой, активация трансляции в результате разрушения вторичной структуры, закрывающей сайт связывания рибосом, и повышение стабильности целевой мРНК. По типу расположения относительно гена не кодирующие РНК делятся на цис и транс. Цис РНК располагаются в том же локусе, что и целевая мРНК, только на противоположенной цепи, что обеспечивает им полную комплементарность. Транс РНК располагаются в участках генома, удаленных от местоположения регулируемого гена, и как правило синтезируются в условиях стресса. При этом транс РНК только частично комплементарные к целевой РНК и очень чувствительны к нуклеотидным заменам. Для стабилизации связи транс РНК-мРНК в большинстве случаев необходимы специализированные шапероны. Среди бактерий распространен белок Hfq, который связывается с 40% малых РНК в *E. coli*. Но у *H. pylori* и у всех (эпсилон)-протеобактерий данный белок отсутствует, что оставляет открытым вопрос понимания механизмов регуляции малыми некодирующими РНК у данных бактерий.

Синтез белковой цепи происходит в соответствии с генетическим кодом, где каждому нуклеотидному триплету, приписана единственная аминокислота, при этом обратное неверно. Помимо того, что генетический код вырожден, триплеты, кодирующие одну и ту же аминокислоту, неравнозначны в скорости чтения, так как концентрации тРНК с соответствующими антикодонами разные. В таком случае даже синонимические замены в генах могут повлиять на представленность данного белка в организме. Немалую роль во влиянии на скорость синтеза также оказывает структура и нуклеотидный состав 5'UTR. Наличие петель требует время на их рас плетение. Присутствие сайтов, схожих с мотивом Шайна-Дальгарно, также задерживает рибосому, так

как имеет высокий потенциал связи. Помимо этого, также как и для процесса транскрипции, существуют белковое факторы, которые связываются с мРНК и регулируют процесс инициции, в итоге повышая или снижая/блокировая весь процесс трансляции.

Время жизни белков в клетке сильно различается в зависимости от их функции и необходимости существования в данных условиях. Одним из механизмов, отвечающим за деградацию белков у прокариотах, является пупилироваие (система подобна убиквинтилированию в эукариотах). Подобно убиквитину, Pup-белок (prokaryotic ubiquitin-like protein) прикрепляется к конкретным остаткам лизина у белка-субстрата путем формирования изопептидных связей. Затем протеосомная АТФаза доставляет Pup-субстрат в протеасому для деградации.

1.7. Белок-белковые взаимодействия *H. pylori*

Белки контролируют и исполняют большинство процессов в клетке. Их свойства модифицируются и контролируются через их взаимодействия друг с другом или другими молекулами. Совокупность белок-белковых взаимодействия (ББВ) называется интерактом. Они могут быть постоянными или временными, посттранслиционно модифицированы или нет и могут формировать гомо и гетеро-белковые комплексы. Посттрансляционные модификации (ПТМ) могут содействовать ББВ. ПТМ чаще всего обратимы и более того могут действовать как молекулярные “переключатели”. Часто ББВ требуют определенной ПТМ, которая необходимы для взаимодействия с партнером. Фосфорилирование белков наиболее широко распространенная модификация. Оно может индуцировать изменение конформации белков и как следствие их активность. Фосфорилирование может функционировать как сайт связывания с взаимодействующей молекулой и более того регулировать фор-

мирование комплексов. Оно важно для активности, например, SH2 домена или 14-3-3 семейства белков, которые могут связываться только с фосфорилированными сайтами. Фосфорилирование белков киназами - это один из центральных механизмов контроля и регуляции клеточных функций, влияющие более чем на треть всех белков в клетке и межклеточные взаимодействия. Многие сигнальные пути регулируются через активация/инактивацию белков, чья активность зависит от фосфорилирования. Другие ПТМ включают в себя метилирование и ацетилирование, гликозилирование или пренилировнные. Многие пренилировнные белки такие как Rho, Rac или Ras включены в базовые клеточные механизмы.

Бактерия *H. pylori* первая бактерия, для который была построена сеть белок-белковых взаимодействий (ББВ). ББВ необходимы практически для всех биологических процессов. Совокупность ББВ называется интерактом. Интерактомы полезны для понимания функций белков и/или путей и как они связаны с плохо или не охарактеризованными белками.

В 2001 году, Рэйн и его коллеги установили частичный интерактом хеликобактера, первая опубликованная сеть взаимодействий белков бактерий (20). В этом исследовании проводился скрининг 261 белковой конструкции против случайной библиотеки, в результате было обнаружено более 1500 ББВ. Хотя эта сеть, вероятно, представляет собой небольшую часть всех ББВ, которые происходят в H. pylori, многие последующий исследования были продиктованы этими результатами. В 2014 году Хоузер с коллегами идентифицировали 1515 ББВ, из которых 1461 являлись новыми. Интеграции обоих исследований дает информацию о 3004 уникальных взаимодействий, которые связывают около 70% протеома. Без учета взаимодействия разнородных белков в исследовании получена базовая сеть, состоящая из 908 взаимодействий. Исследователи сравнили полученные данные с несколькими другими бактериальными интерактомами и экспериментально протестировали сохранение

взаимодействий с использованием 365 белковых пар (interologs) кишечной палочки, из которых одна треть оказалась консервативна у обоих видов.

1.8. Связь между транскриптами и белками: биоинформатические методы в помощь биологии

На основе вышеизложенной информации, можно рассуждать почему же в разных штаммах или в разных условиях жизни уровни транскрипции и белков различаются между организмами и более того, отсутствует явная корреляция между представленностью белков и транскриптов даже внутри одного организма.

Алгоритмы выравнивания последовательностей, помимо поиска замен, позволяют определять комплементарные взаимодействия мРНК и антисмысловых некодирующих РНК, не проводя при этом сложных биологический экспериментов.

Глава 2

Материалы и методы

2.1. Экспериментальные данные

Штаммы *H. pylori*. В работе использовались следующие штаммы: 26695, J99 и A45. Данные штаммы были выращены на среде, содержащий колумбийский агар и инактивированную лошадиную сыворотку.

Также в работе использовались изогенные штаммы A45. В этих штаммах была произведена вставка гена устойчивости к канамицину в соответствующие гены, относящиеся к системе рестрикции-модификации.

Секвенирование ридов RNA-seq. Из выращенных культур в логарифмической стадии роста была выделена и отсеквенврована тотальная РНК. Секвенирование производилось на приборе SOLID 5500.

Масс-спектрометический анализ. В двух временных точках в логарифмической стадии роста из всех штаммов был взят протеом. Анализ белков производился на масс-спектрометре Tripple-TOF с методом ионизации электроспреем ESI.

Эксперименты, секвенировние и масс-спектрометрический анализ производились лабораторией Системной биологии МФТИ и лабораторией протеомного анализа ФГБУ ФНКЦ ФХМ ФМБА России.

2.2. Обработка данных

2.2.1. Данные RNA-seq

Картирование. Риды, полученные после секвенирования, были картированы на соответствующие референсные геномы, взяты из базы данных NCBI (http://ncbi.nlm.nih.gov/). Картировние ридов осуществлялось при помощи программы bowtie. Параметры выбраны в соответствии с рекомендациями разработчиков, а также добавлен параметр -trim3 N, обрезающий N нуклеотидов с конца рида. Данный параметр выбран исходя из анализа качества ридов.

Расчет дифференциальной экспрессии генов. Для расчета количество ридов на ген, была использована программа coverageBed (http://bedtools.read
из программного пакета bedtools

В результате получены таблицы, где указано название генов и количество уникально выравнявшихся на них ридов. Для того чтобы сделать расчет дифференциальной экспрессии (ДЭ) генов, необходим сопоставить гены по группам ортологии. Поиск ортологов между штаммами был произведен при помощи программы OrthoFinder (https://github.com/davidemms/OrthoFinder). Данная программа определяет ортологов по аминокислотными последовательностям соответствующим генам и объединяет гены в группы, если схожесть аминокислотных последовательностей выше %.

На выходе программы получаем таблицу с генам, разделенными по группам ортологии. Для расчета ДЭ были выбраны только те группы, которые содержат по одному гену из каждого штамма.

Расчет ДЭ был сделан в R-скрипте с использованием библиотеки DESeq (http://bioconductor.org/packages/release/bioc/html/DESeq.html).

Поиск транскрипционных старт-сайтов (TCC). Для поиска TCC по траснкрипционному профилю был применен следующий алгоритм:

1. По данным покрытия рассчитываем количество ридов, начинающихся на данной позиции в геноме(по обратной и прямой цепи) ⇒ получаем профиль покрытия.
2. По профилю покрытия определяем локальные максимумы ⇒ каталог локальных пиков.
3. Определяем статистически-значимые пики: сравниваем на сколько совпадает транскрипционный профиль вблизи пика с теоретическим профилем для TCC (f(x) = 1, x > TCC; f(x) = 0, x < TCC) (Рис. 2.1) ⇒ каталог статистически значимых пиков (p-value < 0.05).

Рис. 2.1. Пример: Пик перед геном HP0014

4. Проводим аннотация отобранных пиков: приписываем пики к генам, исходя из расстояния от них до гена. Классификация TCC использовалась такая же как и в статье Синтии Шарм [].

Алгоритм реализован на языке R.

2.2.2. Выравнивание последовательностей и консервативность TCC

Для выравнивания последовательностей использовалась программа nucmmer. На вход программе подаются последовательности, которые хотим сравнить, - qwery, и, с которыми хотим сравнить - reference. В общем случае qwery представляет собой набор из подпоследовательностей, a reference - собранный геном. В качестве критериев оценки схожести последовательностей рассчитываются следующие показатели:

MLength/QLength - % выравненной области;

NSNP - количество полиморфизмов;

FSNP - частота полиморфизмов;

Для расчета консервативности TCC и 5'UTR в качестве qwery берем набор подпоследовательностей, составленный из участков генома NC_000915.1, вырезанных по позициям от TSS до конца гена, а reference - полные геномы хеликобактера, имеющиеся в базе NCBI. Дополнительно по результатам выравнивания считаем показатели:

DSatrt - расстояние от 1-го совпавшего нуклеотида до старта гена;

Olaps - отношение ширины 5'UTR в другом штамме к 26695 (%)

По полученным показателям отбираем TCC и 5'UTR, которые мы будем считать совпавшими с TCC из штамма 26695. Критерии отбора были разделены на первичные и вторичные. Первичные используются для фильтрации неверно выравненных последовательностей, вторичные оценивают показатели совпадения 5'UTR:

- DStart $>$ 0 & DSatrt $<$ 500 - первичные
- 0.8 $<$ Mlength/QLength $<$ 1.1 - первичные
- 0.9 $<$ Olaps $<$ 1.1- вторичные
- QStart $==$ 1 - вторичные

Итоговая величина консервативности TCC равняется отношению количества штаммов с совпавшими TCC к количеству штаммов с гомологичным геном, к которому приписан TCC:

Консервативность TSS = (количество штаммов с совпавшими TCC / количество штаммов с гомологичным геном)

2.2.3. Поиск промоторных последоватльностей.

Промоторные последовательности можно представить в виде матриц, содержащих информацию о вероятность нахождения данного нуклеотида в

данной позиции. Такие матрицы были составлены при помощи программы MEME. Далее в регуляторных последовательностях производился поиск по матрицам при помощи программы RSAT.

2.2.4. Данные масс-спектрометрического анализа

Для пептидных идентификация использовалась программы Mascot v2.2.07. При идентификации программой Mascot были выставлены следующие параметры: точность 20 мд для родительский ионов, 0.5 Да для дочерних ионов, модификации — окисление и карбомидометил. Квантификация пептидов производилась при помощи программы Progenesis [], которая на одном из этапов анализа использует результаты идентификации Mascot. В ходе работы программы Progenesis отбирались пептиды с рангом меньше 15 и зарядом не более 4. Также квантификация производилась независимо по результатам работы программы Mascot отдельно по каждому набору спектров для образцов.

2.2.5. Статистический анализ и визуализация

Для статистического анализа использовался язык программирования R [], в качестве среды разработки использовался RStudio []. Для визуализации данных использовались библиотеки R (qplots, ggplot2) и Cytoscape [] для отображения сетей взаимодействий.

Глава 3

Результаты

3.1. Покрытие генов по данным RNA-seq для штаммов 26695, J99 и A45

[1] Полученные риды для каждого штамма с разделением на биологические повторы были картированы на соответствующие геномы этих штаммов. По результату картирования был произведен поиск *регионов*. Под регионами мы понимаем участки с непрерывным покрытием ридами (хотя бы один рид на каждом участке региона). В среднем по образцам получаем около 4500 регионов на цепь. Будем работать только с теми из них, длина которых больше чем 25нт (25 нт - длина одного рида после обрезки по качеству). Таких регионов 2500 на цепь. Для дальнешей работы проведем поиск общих регионов по биоповторам. Это можно сделать по следующим схемам: объединять если

Способ 1: хотя бы в 2х повторах;

Способ 2: хотя бы в 1м повторе;

Способ 3: хотя бы в 2х повторах или покрытие в одном из повторов составляет больше 5 ридов.

На примере J99 были протестированы все три способа. При использовании способа 1, получается много коротких регионов (по сравнению если искать регионы только в одном повторе). При этом, большинство регионов меньше длины гена (lg(Длина гена) $\geqslant$ 1.5). Такие данные будет не удобно использовать при анализе TCC, так как UTR области разбиты на участки, и тяжело определить, где они начинаются. Способ 2 дает длинные регионы, но при этом мы получаем множество регионов с плохим покрытием. Поэтому воспользуемся Способом 3 и введем порог на глубину покрытия, если оно

присутствует только в одном повторе, 5 ридов. Гистограмма распределения длин регионов представлена на Рис. 3.1,. Длинные регионы соответствуют оперонам.

Рис. 3.1. Плотность распределения длин общих регионов при поиске регионов способом 3 (26695, J99, A45 без регионов длиной 25 нт)

Используя информацию по аннотации генома, позиции найденных регионов были пересечены с позициями генов, в результате:

- отобраны *аннотированные регионы*, т.е. такие регионы в которых находятся гены, либо их участки (регион лежит на гене, регион частично перекрывается геном);
- получена информация о количестве генов, имеющих покрытие данными ридами, т.е. экспрессирующихся в данных штаммах (Таблица 3.1);
- отобраны кандидаты в антисмысловые РНК и новые гены — *не аннотированные регионы*.

Как мы видим на Рис. 3.1, большинство регионов имеют длину меньше длины гена, т.е. много генов с кусочным покрытием. Чтобы посмотреть процент покрытия генов, посчитаем суммарную долю от длины гена, которая перекрывается с общими регионами.

Таблица 3.1. Количество генов, покрытых ридами в длину больше заданного порога

Штамм	Генов по аннотации	$\geqslant 100\%$	$> 90\%$	$> 70\%$	$> 50\%$	$> 25\%$	$< 10\%$
26695	1554	1221	1380	1441	1466	1495	20
J99	1528	1264	1403	1435	1457	1480	15
A45	1619	1057	1326	1438	1492	1542	17

В дальнейший анализ берем гены с покрытием более 90%.

3.2. Каталог TCC

3.2.1. Консервативность TCC и 5'UTR

Штаммы *H. pylori* с собранным геномом относительно штамма 26695. Мы имеем список TCC только для штамма 26695. Чтобы понять, можем ли мы использовать эти же TCC, с поправкой на позиции гомологичных генов, для других штаммов проведем анализ консервативности TCC и советующей 5'UTR. Данный анализ имеет место только для первичных TCC.

Для анализа из списка всех TCC были отобраны первичные TCC, у которых TSS не совпадает с началом гена. Таких TCC оказалось 663 штуки (отброшенные первичные TCC: 11 TCC стоят позже старта гена, - эти гены были переаннотированны в статье Синтии и 36 совпадают со стартом - Leaderless genes). Консервативность TCC будем определять по схожести 5'UTR в разных штаммах относительно штамма 26695 (Материалы и методы). Результаты представлены в Таблице 3.2

Таблица 3.2. Количество TCC со степенью консервативности выше данной

Консервативность	100%	$> 90\%$	$> 70\%$	$> 50\%$	$> 25\%$
Кол-во TCC	280	455	571	609	653

Посмотрим теперь на количество полиморфизмов в 5'UTR. Сразу обратим внимание на тот факт, что помимо единичных нуклеотидных замен, в данных областях также наблюдаются вставки/делеции. 5'UTR со вставкой/делецией более 10% от длины области мы исключим из данного анализа, так как будем считать их не консервативными.

Рис. 3.2. Гистограмма распределения частота встречаемости полиморфизмов в 5'UTR для консервативных и не консервативных TCC.

Из литературных данных известно, что гетерогенность штаммов 26695 и J99 составляет 6%, т.о. если предположить что полиморфизмы равномерно распределены по геному,то в среднем их частота будет 0.06.

Проведенный анализ 5'UTR и первичных TCC показывает, что по выбранным критериям 69% областей имеют уровень консервативности более 0.9 (80% более 0.8). По предложенным параметрам, таким как количество полиморфизмов и их встречаемость, сложно разделить консервативные от не консервативных по данному критерию. Но в свою очередь, данные параметры можно использовать для более жесткого отбора консервативных TCC.

Штаммы A45 и J99 относительно штамма 26695. Для исследуемых трех штаммов посмотрим более подробно на схожесть не только последовательностей 5'UTR, но и регуляторных последовательностей перед TCC (200 нт). Так же как и в предыдущем параграфе за референс будем брать штамм 26695 и смотреть все показатели относительно него.

Для сравнения частоты мутаций было посчитано количество мутаций для трех областей: внутри гена, в 5'UTR и регуляторных регионах (не учитываем вставки/дилеции) (Рис. 3.3.)

Рис. 3.3. Гистограмма распределения частоты встречаемости полиморфизмов в областях: 5'UTR, upstream и ген. A) штамм J99 относительно штамма 26695, B) штамм A45 относительно штамма 26695

В результате данного анализа, мы заметили, что большинство последовательностей 5'UTR и upstream совпадают (частота замен составляет менее 5%). Исходя из этого, можно предположить, что TCC в штаммах J99 и A45 будут располагаться на таком же расстоянии относительно гена, что и в штамме 26695. Составим каталог таких позиций для каждого из трех штаммов и назовем его *"кандидаты в TCC"*.

Помимо этого, частота полиморфизмов в 5'UTR меньше, чем в генах. В большинстве случаев 5'UTR полностью сходятся (нет нуклеотидных замен). Данные результаты приводят к выводу, что регуляторные последовательности жестко определяют экспрессию генов и поэтому бактерии с мутациями в данных областях не выживают.

3.2.2. Проверка TCC по покрытию

Анализ первичных TCC. Имея данные RNA-seq и каталог "кандидатов в TCC мы можем проверить, совпадают ли позиции TCC с началом наших найденных регионов. Будет проверять покрытие соответствующей позиции и расстояние между данной позицией и началом ближайшего региона. В проверку включаем как и аннотированные регионы, так и не аннотированные. В ходе анализа искомые TCC мы приписываем к региону, в котором он был обнаружен и далее классифицируем их по следующим группам (Таблице 3.3):

- найденные TCC - позиции исходных TCC имеют покрытие ридами в наших данных;
- измененные TCC - позиции исходных TCC не покрыты ридами ;
- TCC находятся в одном регионе с соответствующим геном;

Таблица 3.3. Информация по количеству найденных TCC

Штамм	Искомые TCC	Найденные TCC	Измененные TCC	В одном регионе
26695	495	481	14	477
J99	508	483	25	477
A45	474	429	35	439

Анализ расстояния от TCC до начала региона показал, что в основном данное расстояние не превышает 10 нт. TCC которые располагаются даль-

ше чем 1000 нт принадлежат к генам,входящим в опероны, или к перекрывающимся генам. Для измененных TCC расстояния от их предполагаемой позиции составляет менее 10 нт для TCC со степенью консерватинвности больше 0.9 и $\sim$ 30 нт для остальных TCC, при этом некоторые из новых TCC совпадают с началом гена. Посмотрев на расстояние между регионами с предполагаем TCC и с геном было обнаружено, что это расстояние сравнимо с длиной 5'UTR в случае, когда сам ген имеет низкую глубину покрытия ($\sim$ 5 ридов), в других случаях это расстояние составляет около 7 нт. Отсутствие покрытия (полное или частичное) в 5'UTR области можно объяснить его нуклеотидным составом и способностью секвенатора к чтению конкретных нт.

По итогам поиска TCC в покрытии наших данных RNA-seq можно сказать, что большая часть (97%) TCC совпадает с началом регионов, как в штамме 26695 для которого был сделан каталог, так и в двух других штаммах. Данный факт указывает на консервативность 5'UTR между штаммами, а также воспроизводимость транскриптомного анализа.

3.2.3. Поиск новых TCC по покрытию

На предыдущих этапах мы анализировали консервативность 5'UTR и искали первичные TCC для генов на основе гомологии. По итогам составили каталог TCC для каждого их трех штаммов, но он не является полным: во-первых, исходный каталог составлен не для всех генов; во-вторых, не все последовательности 5'UTR сходны между штаммами. Поэтому проведем независимый поиск TCC по транскрипционному профилю по нашим данным.

Используя алгоритм, описанной в главе Материалы и методы, мы составили каталог *"новых TCC"* для каждого штамма. Мы искали только первичные TCC, поэтому отбирали только те пики, которые располагаются не

дальше чем 500 нт от старт-кодона и не дальше чем 10 нт после старт-кодона. Высота пика должна быть не ниже 10 ридов. По такому алгоритму была найдено примерно по 3400 - 4300 пиков в каждом штамме, в среднем получается по 3 пика на ген (Таблица 3.4). Такая перепредставленность пиков связана с неровным профилем покрытия из-за коротких ридов.

Таблица 3.4. Информация о количестве найденных пиков

Штамм	26695	J99	A45
Кол-во пиков	3405	4202	4338
Кол-во генов	1075	1094	1169

3.2.4. Корректировка TCC и поиск промоторных последовательностей

Каталог "новых TCC". В качестве проверки является ли найденные позиции верные, был произведен поиск мотивов связывания с РНК-полимеразой. *H. pylori* имеет три σ-единицы — 28, 54 и 70, каждая из которых узнает свои специфические последовательности. На Рис. 3.4 представлены лого, составленные по таким матрицам (Материалы и Методы). Полученные результаты согласуются с ранними исследованиями[].

а) б) в)

Рис. 3.4. Логотипы промоторных последовательностей:а) σ^{70} б) σ^{54} в) σ^{28}

Мотив для субъединицы σ^{70} составлен только для промторной последовательности на -10 позиции, так как характерный мотив на -35 позиции у *H. pylori* заменен на чередующийся АТ-контент, расположенный от -14 позиции. Мотивы для сигма-факторов σ^{54} и σ^{28} построены на основе литературных

данных, и состоят из двух консервативных частей. При этом поиск для мотива σ^{28} будет производиться сначала по последовательности на -10, так расстояние между консервативными участками варьируется 12-13 нт, что налагает некоторые ограничения на работу программы.

По матрицам был произведен поиск позиций, на которых располагаются промоторные последовательности в регуляторной области, относительно ТСС. Сделаем корректировку выбора пиков на основе найденных промоторных последовательностей. Т.е. если для одного гены идентифицировано несколько пиков, но при этом в upstream найден только один мотив, то считаем, что все пики, кроме одного, "ложные". Выбираем соответственно тот пик, расстояние до которого от мотива ближе всего к правильному, при этом он не всегда самый высокий. На Рис. 3.5 отображено, как располагаются мотивы относительно пиков, после сортировки.

а) б) в)

Рис. 3.5. Плотность распределения мотивов относительно старт-кодона после сортировки: а) штамм 26695; б) штамм J99; в) штамм A45.

Для всех σ-субъединиц мы видим пик в распределении на позициях 8-12 нт от найденного нами пика в покрытии. Позиции с таким расположением мотивов будем считать верно идентифицированными. Наиболее явно он выражен для σ^{70}, так как большая часть генов регулируется данными σ-фактором, и матрица по которой производился поиск является оптимальной для данного мотива, в отличии от матриц для других субъединиц. Наличие мотивов для σ^{70} на позициях отличных от предполагаемой, также связано с тем, что для некоторых ТСС было идентифицировано по нескольку мотивов, так как в чередующимся АТ-контенте встречаются участки подходящие под описание промоторной последовательности на -10 позиции.

Сделаем корректировку позиций ТСС из каталога "новые ТСС но основе найденных к ним мотивов, заменив предположительные ТСС на позиции, которая будет соответствовать найденным мотивам.

Сравнение промоторных последовательностей в разных штаммах. Поиск промоторных последовательностей был также сделан и для каталога "гомологичных ТСС"(Таблица 3.5). Ранее мы выяснили, что регуляторные последовательности у гомологичных генов схожи, здесь мы решили сравнить на сколько совпадают непосредственно промоторные последовательности в исследуемых штаммах.

Таблица 3.5. Информация о промоторных последовательностях

Штамм	σ^{70}	σ^{54}	σ^{28}	совпадают 26695/J99/A45
26695	368	78	19	*/296/314
J99	342	90	24	296/*/309
A45	346	76	20	314/309/*

У 286 генов совпадают мотивы во всех трех штаммах, остальные гены имеют нуклеотидные замены в промоторах. С учетом, что гомологичных генов с ТСС осталось 536 штук, а мотивы были определены идентифицированы у 400 генов, консервативными являются 70% промоторов. А если брать только мотивы для σ^{70}-единиц, которые в свою очередь определены более достоверно, то 90% генов имеют одинаковые промоторы.

Итоговый каталог ТСС. По итогам выше указанных результатов, был собран итоговый каталог ТСС, состоящий из гомологичных и новых ТСС. В Таблице 3.6 представлены количественный показатели по данному каталогу. Используя два подхода поиска, мы нашли ТСС для 63% генов. Как известно,

бактерий большая часть генов (?%) объединена в опероны и транскрибируются совместно, т.е. с одного ТСС. У штамма *H. pylori* 26695 по базе DOOR найдено 311 оперонов, у J99 - 318, по штамму A45 информации об оперонах нет. Среди найденных ТСС, половина принадлежат к генам, находящимся внутри оперонной структуры. Данный факт говорит о возможной альтернативности прочтения оперона, когда клетки нуждается только в части генов из оперона. (можно сравнить различаются эти гены в гомологичных оперонах или нет)

Таблица 3.6. Количественные показатели каталога ТСС

Штамм	26695	J99	A45
Гены с новыми ТСС	352	460	444
Гены с гомологичными ТСС	649	482	430
Всего генов с ТСС	1001	996	980
Внутри оперона	516	521	нет инф.

3.3. Дифференциальная экспрессия генов

Штаммы 26695, J99 и A45. На основе гомологии, мы сравнили уровень экспрессии генов между штаммами. В качестве референсного штамма был выбран штамм A45. Мы считали гены дифференциально экспрессирующимися (ДЭ) если уровень их представленности изменялся более чем в 4 раза, при статистической значимости p-value < 0.01 (Таблица 3.7.

% найденных ДЭ генов находиться в оперонах, но при этом не все из этих оперонов являются ДЭ. Одним из объяснений данного факта, может быть то, что опероны имеют альтернативные внутренние ТСС и не всегда

Таблица 3.7. Дифференицильно экспрессирующиеся гены в штаммах 26695 и J99 относительно A45.

	всего	снижается	растет
26695	86	44	42
J99	80	46	34
общие	26		

считываются целиком. (надо посмотреть есть ли TCC у таких оперонных ДЭ генов). Как было замечено раннее, некоторые гены имеют нуклеотидные замены в промоторных областях, что влияет на уровень их экспрессии. У найденных ДЭ генов промотороные последовательности являются консервытивнами/неконсверватинами.

Изогенные штаммы и штамм A45. Мы сравнили уровень представленности генов в штамме A45 с его изогенными штаммами (Таблица 3.8 (критерии отбора были такие же как и при сравнении A45 с двумя другими штаммами). Количество ДЭ генов в данном сравнении меньше, чем при сравнении A45 с 26695 и J99. Наибольшая разница в представленности генов среди изогенных штаммов наблюдается у Hpy. Главное отличие данного мутанта от остальных, в том, что мутации произведена по гену метилазы, отвечающего за регуляцию процессов.

На данном этапе исследования, мы хотели понять, как влияет профиль метилирования на уровень экспрессии генов. Мы посмотрели, где располагаются сайты метилирования для "выключенных"метилаз и провели сравнения зависимости ДЭ генов от представленности и локации сайтов в этих генах. В результате, ...

Таблица 3.8. Дифференицильно экспрессирующиеся гены в изогенны штаммах 1352, 9192 и Hpy относительно A45.

	всего	снижается	растет
1352	12	9	3
9192	26	14	12
Hpy	55	46	9
общие	8		

3.4. Представленность белков

3.4.1. Квантификация белков

Мы работали с двумя наборами данных для двух временных точек. Квантификация белков проводилась при помощи двух программ: Progenesis и Mascot (Материалы и методы) (Таблица). Данные Mascot были рассчитаны для каждого биоповтора и временной точки отдельно, в то время как программа Progenesis обсчитывала данные разделяя их только по образцам (штаммам). Таким образом, Macsot идентифицирует больше белков, так как делает это независимо для образов. В Progenesis первым шагом является выравнивания спектров MC1, и идентифицируется совпавшие пики. Прежде всего мы посмотрели кластеризуются ли протеомные данные по образцам. Для сравнения методов кластеризация проводилась для обоих подходов. Кластаризация проводилась на основе корреляции представленности белков в разнцх штаммах, соответственно для этого были выбраны только гены с однозначной гомологией: 416 - квантификация Mascot, 362 - квантификация Progenesis. Кластеризация по Progenesis более наглядная, так как коэфиценты корреляции выше, чем в случае идентификаций Mascot (данные не представлены), но в целом картина схожая. На Рис. 3.6 представлены результаты идентификации Progenesis и мы видим, что явно выделяются кластеры для

штаммов 26695 и J99 и большой кластер - штамм A45 и его изогенные штаммы. Внутри большого кластера мы наблюдаем, что образы в первую очередь разделение по временным точкам, а потом по штаммам, за исключением самого штамма A45.

Рис. 3.6. Корреляция протеомных данных по идентификации Progenesis.

В дальнейшем будет работать с результатами программы Progenesis.

3.4.2. Сравнение представленности белков в разных временных точках

Посмотрим, насколько отличается представленность белков в разных фазах роста, т.е. сравним результаты квантификаций в перовой и второй временной точке. Несмотря на то, что в данному случае анализ представленности проводиться внутри одного штамма, будем брать только белки от гомологичных генов (362 штуки). Делаем это с целью упрощения процедуры дальнейшего сравнения отличий во времени с межштаммовыми. С учетом того, что методы протеомного анализа менее чувствительные к концентрациям вещества, чем методы транскрипционного анализа, мы считали, что представленность белков достаточно сильно отличается между штаммами, если это отличие более чем в 2 раза.

В Таблице 3.9 представлены результаты сравнения. Мы заметили, что количество белков, изменяющих свою концентрация во времени более чем в 2 раза, больше всего у изогенных штаммов. В работе (Куват) доказано, что скорость роста изогенных штаммов различается мжеду собой и от штаммам A45. Данный факт был учтен при выращивании культур и сбора образов. Тем не менее, обилие отличных по уровню представленности белков скорей всего связано с особенностями их скорости роста.

Таблица 3.9. Количество отличающихся белков по уровню представленности в двух временных точках

	отличаются	$\geq 1.5x$	$\geq 2x$	$\geq 10x$
A45	168	78	24	0
26695	301	88	42	4
J99	234	57	19	0
1352	231	121	75	10
9192	224	140	80	7
Hpy	225	145	94	10

3.4.3. Сравнение представленности белков в разных штаммах

Для начала делаем нормализацию полученных данных, так как квантификация была сделана отдельно для каждого образца. Все штаммы мы также как и в случае экспрессии генов сравнивали с штаммом A45. Результаты представлены в Таблице 3.10 и можно заметить, что протеом между штаммами 26695 и J99 имеет более явные отличие, от штамма A45, чем его изогенные штаммы, что ожидаемо.

Таблица 3.10. Количество отличающихся белков по уровню представленности в штаммах 26695, J99, изогенны штаммах 1352, 9192 и Hpy относительно A45.

	точка 1			точка 2		
	отличаются	≥ 1.5	≥ 2	отличаются	≥ 1.5	≥ 2
26695	289	212	141	249	214	159
J99	241	213	164	264	234	170
1352	241	182	109	231	166	85
9192	218	186	120	262	190	115
Hpy	217	189	120	277	181	100

3.5. Белковые комплексы

Как известно, белки для совместного функционирования объединяются в комплексы. Представленность белкового комплекса, предположительно, будет ограничена наименее представленной компонентой. Мы взяли карту белок-белковых взаимодействий для штамма 26695 и оставили в ней только те связи, для которых однозначно определены ортологи в штамме A45 и для которых идентифицированы белки. В результате мы получали карту взаимодействий, состоящих из 65 белков и содержащую 55 взаимодействий. Карта представляют собой 3 кластера (более 3 белков), 1 подсеть из 3 белков, 9 бинарных взаимодействий и несколько отдельных белков образующих связь сами с собой (не будем их рассматривать).

Рассмотрим более подробно кластеры и назовем некоторые из них для удобства дальнейшего обращения:

Уреазный кластер : белки-субъединицы уреазного комплекса и белки, участвующие в процессе производства уреазы;

Рибосомальный кластер : рибосомальные-субъединицы и белки, сопутствующие процессу связывания рибосомы с РНК;

Кластер хемотаксиса : белки, участвующие в процессе хемотаксиса и биосинтеза жгутиков.

Остальные взаимодействия : рибосомальные-субъединицы; урезазный белок и белок хемотаксиса, субъединицы-ДНК-гиразы и некоторые другие.

Интересный вопрос, как меняется представленность комплексов и их компонент в разных штаммах: является ли это согласованное изменение всех компонент или же регуляция происходит за счет одного участника? Взяв значения отношения представленности белков в штаммах относительно штамма A45, мы посмотрели, являются ли эти изменения сонаправленными или же

нет.

3.6. Сравнение представленности белков и транскриптов по штаммам

3.6.1. Корреляция представленности белков и транскриптов

Рассмотрим как зависит представленность белков от представленности транскрипотов. Для начала, посмотрим на общую корреляцию Рис. 3.7. Значение корреляции колеблется около 0.5, что соответствует другим исследованиям по данному вопросу. На тепловой карте, мы видим, что образцы (и протемоные, и транскриптомные) по прежнему кластеризуются по 3 группам: штаммам 26695, штамм J99 и штамм A45 вместе с изогенными. Напомним, что транскриптомные данные сняты по первой временной точке, поэтому, как и предполагается, корреляция представленности белков с соответствующими транскриптами в первой точке выше, чем во второй. Однако, группировку по временным точкам мы видим для штаммов J99 и 26695. Изогенные штаммы A45 в общем тоже группируются по временным точкам, за исключение того, что образца от самого штамма из второй точки попадают в группу с изогенными штаммами по первой точки. С одной стороны, это связано с особенностями метода кластеризации, но также можно предположить, что доля отличий в представленности белков в разных фазах роста схожа с отличиями, вызванными отключением гена PM системы.

Рис. 3.7. Корреляция протеомных и транскриптомных данных.

3.6.2. Корреляция представленности белков и транскриптов отдельно для генов

После того, как мы исследовали корреляцию представленности белков и транскриптов в общем по штаммам, мы посмотрели, как меняются представленность этих компонент в частности для каждого гена.

Литература

[1] Alm R. A., Ling L.-S. L., Moir D. T. et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori // Nature. 1999. Vol. 397, no. 6715. P. 176–180.