Федеральное государственное бюджетное учреждение науки

Институт проблем передачи информации им. А.А. Харкевича

Российской академии наук

На правах рукописи

Казнадзей Анна Денисовна

Геномная ко-локализация генов углеводного метаболизма бактерий

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание учёной степени

кандидата биологических наук

Научный руководитель:

доктор биологических наук, профессор

М.С. Гельфанд

Москва – 2019

Оглавление

Актуальность работы

С развитием технологий секвенирования в последние годы количество данных о последовательностях ДНК растет с огромной скоростью. При этом задачи, связанные со сравнением нуклеотидных последовательностей, не характеризующихся очень высоким уровнем сходства, по-прежнему решаются либо с помощью чувствительных и медленных, либо с помощью быстрых и малочувствительных алгоритмов. В результате либо время работы инструмента оказывается неприемлемо долгим, либо в ходе поиска теряется значительная часть результатов. Таким образом, актуальной на данный момент является разработка быстрых, но при этом точных и чувствительных методов сравнения неблизких последовательностей ДНК. Первый этап настоящей работы был посвящен разработке такого инструмента.

Одним из важнейших объектов современных исследований являются бактериальные геномы. Бактерии способны приспосабливаться к самым разным условиям среды и, в частности, катаболизировать широкий спектр углеводов. Белки, участвующие в соответствующих процессах, закодированы в бактериальных генах. Исследования, касающиеся структуры, функций и регуляции работы таких генов, а также их сочетаний, ведутся уже несколько десятков лет. Так, лактозный оперон кишечной палочки, состоящий из трех генов, стал первым описанным опероном прокариот. До сих пор, однако, не было проведено масштабных исследований, касающихся общих тенденций взаиморасположения генов углеводного метаболизма в бактериальных геномах и факторов, влияющих на эти тенденции. Второй этап данной работы был посвящен проведению такого анализа, в том числе, с применением инструмента для сравнения нуклеотидных последовательностей, разработанного на предыдущем этапе.

Известно, что консервативность сочетаний генов на хромосомах может позволять делать успешные предсказания о свойствах этих генов. Экспериментальная проверка подобных предсказаний важна с точки зрения соотношения теоретических и практических знаний и вносит существенный вклад в понимание эволюционного значения геномного окружения генов. Третьим этапом данной работы стало предсказание связи кассеты генов *Escherichia coli*, участвующей в сульфогликолизе, с метаболизмом лактозы, которое было сделано на основе анализа консервативных ко-локализационных тенденций генов углеводного метаболизма. Предсказание было подтверждено экспериментально; в частности, была показана выраженная активация экспрессии генов кассеты *Escherichia coli* при росте на лактозе, что свидетельствовало об их вовлеченности в процесс ее утилизации. Положительный результат данного эксперимента подтвердил актуальность подобных предсказаний и позволил затронуть, в свою очередь, малоизученный вопрос о мультифункциональных свойствах бактериальных белков.

Цели и задачи исследования

Целью работы было выяснить, как организованы геномные локусы бактерий, содержащие гены углеводного метаболизма, какие факторы влияют на эту организацию, какие эволюционные механизмы стоят в ее основе, и как можно использовать данные о ко-локализации этих генов для предсказания их функций.

Были поставлены следующие задачи.

1. Оценить, как часто гены углеводного метаболизма располагаются на бактериальных хромосомах рядом, т.е. формируют в геномах кассеты, и как часто они располагаются по отдельности, а также описать разнообразие кассет.

2. Выяснить, как функциональные и структурные характеристики кодируемого белка влияют на склонность соответствующего гена к формированию кассет, а также как склонность к формированию кассет варьирует среди разных таксонов бактерий.

3. Оценить тенденции к ко-локализации генов разных функций и тенденции к ко-локализации генов сходных функций.

4. Разработать инструмент, позволяющий эффективно оценивать уровень сходства нуклеотидных последовательностей, различающихся на 10% и более, и применить этот инструмент для оценки вклада событий локальной дупликации в ко-локализацию генов сходных функций.

5. Применить анализ тенденций ко-локализации генов углеводного метаболизма для конкретного случая предсказания функций генов с последующей проверкой.

Научная новизна и практическая ценность

В работе рассмотрены актуальные вопросы и решен ряд задач современной сравнительной геномики.

Разработан и программно реализован биоинформатический инструмент, позволяющий проводить поиск заданных нуклеотидных последовательностей удаленного сходства в больших базах данных ДНК, который по совокупности таких параметров, как чувствительность, точность и скорость превосходит инструменты, считающиеся индустриальным стандартом.

Впервые проведен масштабный и детальный анализ ко-локализационных особенностей генов углеводного метаболизма бактерий. Выявлены основные факторы, влияющие на формирование кассет таких генов. Исследованы тенденции попарных сочетаний генов разных функциональных классов и разных

ортологических кластеров, а также тенденции ко-локализации генов сходных функций. Выявлен вклад в такие случаи событий локальной дупликации генов.

Выдвинута гипотеза о том, что сравнительный анализ сочетаний функциональных классов генов углеводного метаболизма внутри кассет может позволять предсказывать общую функцию кассеты и ее участие в соответствующем метаболическом пути. Гипотеза подтверждена для кассеты генов кишечной палочки, участвующей в сульфогликолизе и совпадающей по общему функциональному составу с консервативной кассетой, участвующей в катаболизме лактозы у бактерий класса Bacilli. Впервые, таким образом, описан альтернативный путь катаболизма лактозы у кишечной палочки, а также предсказаны мультифункциональные характеристики соответствующих белков. Также впервые были картированы промоторы генов данной кассеты и описан механизм переключения регуляции их экспрессии.

Основные результаты и положения, выносимые на защиту

Разработан инструмент NSimScan для поиска нуклеотидных последовательностей удаленного сходства; наилучшим образом он подходит для поиска последовательностей, различающихся на 60-90%. По совокупности таких параметров как чувствительность, точность и скорость он превосходит все стандартные инструменты в своей области.

Описана сеть эволюционных связей 148 тысяч генов углеводного метаболизма 665 видов бактерий, выраженная в форме их ко-локализационных тенденций. 53% таких генов находятся в составе кассет, то есть ко-локализованы, остальные располагаются на бактериальных геномах по отдельности.

Склонность к формированию кассет различается у разных генов; ключевыми факторами, влияющими на их ко-локализационные тенденции, являются

функциональные и структурные характеристики гена и филогенетические свойства соответствующей бактерии. Склонность к формированию кассет у разных функциональных классов составляет от 23 до 93%; у разных кластеров ортологических групп генов – 0 до 100%, у разных бактериальных классов – от 40 до 76%.

Функциональные классы могут формировать консервативные и, по всей видимости, эволюционно значимые ко-локализационные связи; всего описано 45 таких связей для 19 исследуемых классов. Количество связей для каждого класса сильно варьирует, что указывает на существенное различие в предпочтениях к непосредственному геномному окружению у генов разных функций. Гены 11 функциональных классов демонстрируют выраженное предпочтение к внутриклассовой ко-локализации, причем большинство таких случаев, по-видимому, не являются результатом событий локальных дупликаций.

Исследование консервативных комбинаций внутри кассет генов углеводного метаболизма позволяет успешно предсказывать их функции. На основании сходства консервативной кассеты генов семейства Enterobacteriaceae, отвечающей за катаболизм серосодержащих сахаров, с консервативной кассетой бактерий класса Bacilli, участвующей в катаболизме лактозы, предсказано и экспериментально подтверждена роль кассеты *Escherichia coli* в утилизации лактозы. Описан, таким образом, ранее неизвестный путь катаболизма лактозы у кишечной палочки и предсказаны мультифункциональные характеристики соответствующих белков. В переключении механизмов экспрессии генов этой кассеты при смене источника углерода в среде участвуют локальный регулятор YihW и глобальный регулятор CRP.

Структура и объем диссертации.

Диссертация изложена на 145 страницах. Она состоит из 4 глав: "Литературный обзор", "Инструмент NSimScan для поиска удаленных сходств последовательностей ДНК", "Организация генов углеводного метаболизма бактерий", и "Участие *yih*-кассеты *Escherichia coli* в катаболизме лактозы". Работа содержит 21 рисунок и 3 таблицы. Приложение содержит 4 таблицы.

Список публикаций по теме диссертации

По материалам диссертации опубликовано три статьи в рецензируемых научных журналах, входящих в Web of Science:

1. V. Novichkov, A. Kaznadzey, N. Alexandrova, D. Kaznadzey (2016) NSimScan: DNA comparison tool with increased speed, sensitivity and accuracy. Bioinformatics 32(15):2380-1.

2. A. Kaznadzey, P. Shelyakin, M. Gelfand (2017) Sugar Lego: gene composition of bacterial carbohydrate metabolism genomic loci. Biology Direct 12(1):28.

3. A. Kaznadzey, P. Shelyakin, E. Belousova, A. Eremina, U. Shvyreva, D. Bykova, V. Emelianenko, A. Korosteleva, M. Tutukina, M. Gelfand (2018) The genes of the sulphoquinovose catabolism in Escherichia coli are also associated with a previously unknown pathway of lactose degradation. Scientific Reports 8(1):3177.

Результаты работы были представлены на международных и российских конференциях:

1. A. Kaznadzey (2010) Evolutional study of carbohydrate metabolism loci in bacterial genomes, Interdisciplinary School and Conference of Information Technology and Systems (ITaS'10), Геленджик.

2. A. Kaznadzey, P. Shelyakin (2011) Study of evolution and classification of genome loci of carbohydrate metabolism of bacteria. Interdisciplinary School and Conference of Information Technology and Systems (ITaS'11), Геленджик.

3. A. Kaznadzey, P. Shelyakin (2011) Evolution study and classification of carbohydrate metabolism genome loci in bacteria. International Moscow Conference on Computational Molecular Biology (MCCMB'11), Москва.

4. A. Kaznadzey, P. Shelyakin (2012) Diversity of genome loci and co-localization patterns study of the protein families from different functional classes of the bacterial carbohydrate metabolism. 8th International Conference on the Bioinformatics of Genome Regulation and Structure – Systems Biology (BGRS\SB-2012), Новосибирск.

5. A. Kaznadzey, P. Shelyakin (2012) Diversity of genome loci and co-localization patterns study of the protein families from different functional classes of the bacterial carbohydrate metabolism. Interdisciplinary School and Conference of Information Technology and Systems (ITaS'12), Петрозаводск.

6. A. Kaznadzey, P. Shelyakin (2013) Structure, classification, evolution and phylogenetics of carbohydrate metabolism gene loci in bacteria. Moscow Conference on Computational Molecular Biology (MCCMB'13), Москва.

7. A. Kaznadzey, P. Shelyakin (2015) Co-evolution of carbohydrate metabolism genes of same and different functional classes in bacteria' (ITaS'15), Сочи.

8. A. Kaznadzey, M. Tutukina, A. Eremina, E. Belousova, P. Shelyakin, M. Gelfand (2016) Escherichia coli gene cassette previously described as an operon responsible for sulphoglycolipide degradation: not an operon and has other functions as well. Interdisciplinary School and Conference of Information Technology and Systems (ITaS'16), Санкт-Петербург.

Список используемых обозначений

COG – Cluster of Orthologous Gene groups, кластер групп ортологических генов

IMG – Integrated Microbial Genomes & Microbiomes, обобщенная база данных геномов микробов института Joint Genome Institute

ДНК – дезоксирибонуклеиновая кислота

РНК – рибонуклеиновая кислота

ORF – open reading frame, открытая рамка считывания

CRP – цАМФ-зависимый катаболит-активируемый белок

цАМФ (cAMP) – циклический аденозинмонофосфат

PEP – фосфоенолпируват-фосфотрансферазная система

УНР – усредненное нуклеотидное расстояние

ПЦР - полимеразная цепная рекция

NGS – next generationsequencing, технологии секвенирования "нового поколения"

HSP – high scoring segment pair, пара последовательностей с высоким сходством

п.н. – пары нуклеотидов

Глава 1. Литературный обзор

Настоящая работа состоит из трех основных частей. Первая часть посвящена разработке биоинформатического инструмента для поиска нуклеотидных последовательностей с удаленным сходством. Вторая часть посвящена анализу ко-локализационных тенденций генов углеводного метаболизма бактерий; инструмент, полученный на первом этапе, применялся для оценки вклада в них событий локальной дупликации генов. Третья часть посвящена предсказанию функций генов на основании результатов второго этапа работы и проверке гипотезы об эволюционной значимости консервативных сочетаний генов углеводного метаболизма; в данном случае предсказание касалось участия сульфогликолитической кассеты *Escherichia coli* в катаболизме лактозы. Глава "Литературный обзор" поделена, таким образом, на три соответствующих раздела.

1.1. Сравнение нуклеотидных последовательностей

Недавняя революция в технологиях секвенирования нуклеиновых кислот возвела требования к сравнению их последовательностей на новый уровень. Для успешного анализа соответствующих данных (в том числе, в рамках клинических тестирований) были разработаны эффективные методы картирования коротких фрагментов ДНК (прочтений, sequencing reads), полученных непосредственно в результате секвенирования. Под картированием в данном случае подразумевается определение местоположения и выравнивание таких прочтений с уже известной последовательностью ДНК, т.н. референсным геномом, с которым сравнивают новые фрагменты. Последовательности, которые подвергают картированию, как правило, несущественно отличаются от референсных, поэтому алгоритмы соответствующих инструментов (например, BWA [1] или Bowtie2 [2]) направлены на поиск близких совпадений между целевыми и референсными фрагментами.

Результат работы этих инструментов позволяет анализировать точечные мутации в геномах разных представителей известного вида. Например, их успешно применяют для поиска однонуклеотидных замен, а также небольших вставок и делеций в человеческих геномах; при этом стоит отметить, что различие между нуклеотидными последовательностями геномных локусов у людей составляет в среднем не более 0,1% [3] (не учитывая микросателлитные последовательности, которые характеризуются более высокой скоростью накопления эволюционных изменений по сравнению с остальным геномом [4]).

BowTie2 и BWA предназначены, таким образом, для работы с короткими (как правило, длиной до 1000 нуклеотидов), много раз повторяющимися прочтениями. Их получают в результате применения современных технологий секвенирования, таких, как NGS ("секвенирование нового поколения"). В основе алгоритмов этих инструментов лежит специализированное представление нуклеотидной последовательности референсного генома в виде суффиксного массива ("FM-index") на основе преобразования Барроуза–Уилера [5] и поиск оптимального совпадения прочтения с референсным геномом. Здесь используется жадный эвристический метод, в общем случае не гарантирующий обнаружение наилучшего выравнивания. В данном случае, однако, такой подход является оптимальным, именно из-за того, что на референсную последовательность картируют прочтения, которые должны соответствовать ей или несущественно от нее отличаться. Соответствующие инструменты характеризуются высокой скоростью работы и требуют относительно небольших затрат памяти.

Среди других инструментов, используемых для поиска почти идентичных нуклеотидных последовательностей, можно назвать также более ранние инструменты SSAHA [6] и BLAT [7]. Инструмент SSAHA, созданный в 2001 году, предназначен для работы с большими базами данных; в основе его алгоритма

лежит составление таблицы местоположений k-меров нуклеотидных последовательностей базы данных (длина k-мера по умолчанию составляет 10 нуклеотидов), что позволяет быстро отыскивать точные совпадения и совпадения с относительно редкими однонуклеотидными заменами в искомых последовательностях; для поиска последовательностей с более существенными расхождениями такой инструмент не подходит. Программа BLAT, также разработанная в начале 2000-ых годов для сборки и аннотирования человеческого генома, была ориентирована на повышение скорости именно этих процессов, и оказалась приблизительно в 500 раз быстрее аналогов своего времени, используемых для работы с геномами позвоночных животных. Как и в случае SSAHA, алгоритм BLAT использует таблицу вхождений k-меров (длина k-мера в ней по умолчанию составляет 11 нуклеотидов), созданную на основе последовательностей базы данных; он позволяет находить последовательности с 95% сходством на длине от 40 нуклеотидов. Один из вариантов его применения, более медленный, также позволяет искать k-меры с однонуклеотидными заменами.

Задачи поиска нуклеотидных последовательностей удаленного сходства (последовательностей, совпадающих менее, чем на 90%) по-прежнему решаются либо с помощью чувствительных и медленных инструментов, разработанных тогда, когда приток новых геномных данных был небольшим, либо с помощью новых и быстрых, но малочувствительных алгоритмов. В первом случае критическим фактором оказывается время работы инструмента, а во втором теряется значительная часть искомых результатов. При этом благодаря быстро развивающимся технологиям секвенирования количество новых данных по последовательностям нуклеиновых кислот, требующих дальнейшего анализа, растет экспоненциально. Самым распространенным видом такого анализа является сравнение полученных последовательностей друг с другом и с большими базами

данных уже известных нуклеиновых и белковых последовательностей для выявления всевозможных структурных и эволюционных связей между ними. Инструменты, которые сейчас чаще всего применяют для поиска последовательностей удаленного сходства, это BLAST [8], SSearch [9], MegaBLAST [10] и USEARCH [11].

Наиболее чувствительный поиск сходств последовательностей возможен с помощью алгоритма Смита–Ватермана, разработанного Т. Смитом и М. Ватерманом в 1981 году [12]. Он позволяет проводить локальное выравнивание последовательностей, осуществляя выравнивание отрезков всех возможных длин и затем оптимизируя меру сходства по всем полученным выравниваниям. Здесь используется принцип динамического программирования, то есть представление сложной задачи в виде рекурсивной последовательности более простых подзадач [13]. При составлении выравниваний применяется матрица замен и система штрафов за пропуски (вставки и делеции). Один из первых инструментов, использующий данный алгоритм в исходном виде и получивший широкое распространение для сравнения нуклеотидных последовательностей ДНК (а также для сравнения "переведенных" в нуклеотидную последовательность белковых последовательностей с другими нуклеотидными последовательностями), стал FASTA [14], разработанный еще в 1987 году. Алгоритм Смита–Ватермана в нем применяется после того, как составляется словарь потенциальных кандидатов для выравнивания на основе поиска коротких совпадающих k-меров (длиной 4 или 6 нуклеотидов) для каждой пары сравниваемых последовательностей и определяется штраф за пропуски между найденными совпадениями.

Алгоритм Смита–Ватермана позволяет строить любые выравнивания, в том числе для неблизких или даже случайных последовательностей. В сравнении с инструментами, в ходе работы которых вначале осуществляется отбор

последовательностей базы данных с совпадающими k-мерами, сам по себе алгоритм Смита–Ватермана позволял бы проводить гораздо более чувствительных поиск. Лимитирующим фактором, однако, является время его работы: при поиске в современных крупных базах данных с нуклеотидными последовательностями оно становится практически бесконечным, возрастая пропорционально произведению длины искомой последовательности и суммарной длины последовательностей базы данных.

Поэтому многие последующие алгоритмы были созданы таким образом, чтобы полностью или частично отказаться от применения алгоритма Смита-Ватермана. В том числе, эта задача стояла при разработке широко применяемого инструмента BLAST.

В ходе работы BLAST вначале составляется словарь k-меров искомой последовательности. Длина нуклеотидного k-мера для BLAST составляет по умолчанию 11 нуклеотидов. Затем проводится поиск точных вхождений всех таких k-меров в заранее подготовленной базе данных, представленной в бинарном виде. В исходной версии BLAST найденные таким образом точные соответствия затем продлеваются в обе стороны до тех пор, пока доля сходства полученного локуса ("зародыша" или High Scoring Segment Pair, HSP) с исходной последовательностью не опускается ниже определенного порога. Доля сходства определяется из количества совпадений продлеваемой последовательности с использованием системы весов Смита-Ватермана. В современной версии BLAST для увеличения чувствительности поиска используется метод "gapped BLAST", в котором статистическая значимость HSP, располагающихся по соседству, оценивается совместно. Для оценки значимости (e-value) HSP используется экстремальное распределение Гумбеля [15]

Несмотря на то, что параметры поиска BLAST можно менять (назначая разные штрафы за пропуски, меняя длину k-мера и т.п.), обеспечить чувствительность BLAST на уровне исходного алгоритма Смита–Ватермана невозможно, однако в данном случае важен очень существенный выигрыш в скорости и возможность работы с большими базами данных.

Инструмент MegaBLAST работает с кратными четырем k-мерами длиной от 16 нуклеотидов и больше (часто используемая длина для быстрого поиска с низкой чувствительностью – 28) и также ищет вначале их точные вхождения. Он удобен для быстрого, масштабного и не очень чувствительного поиска. При поиске нескольких последовательностей он сливает их в одну, причем таким образом он может обрабатывать более пятнадцати тысяч искомых последовательностей за один запуск. Метод работы этого инструмента характеризуется, в частности, очень низкими штрафами за пропуски. Последние версии MegaBLAST используют двухуровневый индексированный словарь из нуклеотидных последовательностей базы данных, так, чтобы для большинства возможных искомых последовательностей было достаточно одного прохождения по базе. В среднем MegaBLAST работает в 10 раз быстрее, чем BLAST, и способен относительно быстро обрабатывать крупные базы данных и последовательности очень большой длины, для чего и был создан (одним из типичных вариантов его применения является работа с метагеномами [16]).

Инструмент SSearch [17] работает на основе алгоритма выравнивания Смита–Ватермана, без дополнительных ускоряющих этапов. Для оценки значимости полученных результатов он учитывает веса выравниваний и логарифм их длины. SSearch не подразумевает необходимости наличия между искомой последовательностью и базой данных точных совпадений определенной длины,

поэтому он значительно более чувствительный, чем BLAST, но и гораздо более медленный.

Инструмент USEARCH работает примерно в 10 раз быстрее, чем BLAST. Его алгоритм основан на отборе одного или нескольких результатов с наибольшим количеством коротких точных вхождений и игнорировании всех остальных результатов (порог задается специальным параметром). Разработчики сообщают о хороших результатах работы инструмента при поиске сходств нуклеотидных последовательностей от 65% и выше, однако из-за отсечения значительной части результатов после обнаружения нескольких первых совпадений существенно повышается риск потери совпадений с равной или даже более высокой значимостью. Таким образом существенно снижается чувствительность поиска, множество подходящих последовательностей остается ненайденным.

Чувствительностью поиска называется количество истинно-положительных результатов относительно суммы истинно-положительных результатов с истинно-отрицательными. Точностью поиска называется количество ошибок, т.е. доля ложно-положительных результатов среди всех найденных. Одной из целей данной работы была разработка быстрого и при этом точного и чувствительного алгоритма для полноценного поиска сходств между нуклеотидными последовательностями, отличающимися друг от друга более, чем на 10%. Такие условия поиска необходимы, например, в рамках проведения филогенетических исследований и других методах сравнительного анализа, а также для осуществления функциональных предсказаний. В данной работе этот алгоритм использовался, в частности, для выявления событий локальной дупликации генов углеводного метаболизма.

1.2. Организация генов углеводного метаболизма бактерий

Углеводный метаболизм бактерий отличается большим разнообразием, поскольку самые разные углеводы служат бактериям источниками энергии. Углеводы также участвуют во множестве ключевых клеточных процессов и являются важным структурным элементом бактериальной клетки; в частности, они входят в состав клеточной стенки [18]. Метаболизм моносахаридов, олигосахаридов и полисахаридов осуществляется у разных бактерий с помощью десятков различных метаболических путей [19–23]. Ферменты, отвечающие за разные этапы таких путей, транспортные белки, обеспечивающие доставку углеводов в клетку и из клетки, и соответствующие факторы транскрипции закодированы в бактериальных генах. В данной работе мы исследовали бактериальные геномные локусы, содержащие эти гены.

Метаболические пути – это наборы последовательных химических реакций, происходящих в клетке. Промежуточные и итоговые продукты этих реакций называются метаболитами. На метаболиты воздействуют ферменты, катализируя соответствующие реакции. Метаболической картой называют совокупность всех известных метаболических путей конкретного организма или группы организмов, представленных в форме единой сети взаимосвязанных реакций [24]. Метаболические пути подразделяют на катаболические, участвующие в деградации (распаде) химических веществ, и анаболические, участвующие в их синтезе, а соответствующие процессы называют катаболизмом и анаболизмом.

Известно, что гены, кодирующие белки, относящиеся к одному и тому же метаболическому пути, часто располагаются на бактериальной хромосоме вблизи друг от друга [25–27]. В случае, когда гены, расположенные подряд на одной хромосоме, объединены общим или несколькими общими промоторами, и РНК-полимераза может считывать с них единый транскрипт, такой набор генов

называют опероном [28]. Огромное количество исследований посвящено конкретным оперонам и их эволюции среди близкородственных видов.

Одной из задач сравнительной геномики является выяснение общих причин и тенденций ко-локализации генов. Известно, в частности, что белки, физически взаимодействующие друг с другом, могут иметь тенденцию к тому, чтобы быть закодированными на хромосоме рядом и в определенном порядке [29]. Однако ко-локализация генов функционально связанных белков не является обязательным правилом и не всегда оказывается закрепленным с эволюционной точки зрения событием.

Было показано, тем не менее, что кластеры генов, сформированные по признаку совместного присутствия в геноме, сохранения относительного расстояния или непосредственной хромосомной ко-локализации в выборке бактериальных видов, насыщены функционально связанными элементами, т.е. гены из одного такого кластера часто кодируют белки, функции которых относятся к родственным биологическим процессам [25,30,31]. На основании анализа корреляции метаболических и ко-локализационных характеристик генов, построенных с помощью известной на тот момент карты метаболических путей *Escherichia coli*, проведенного в 2006 году, выяснилось, что у ферментов, участвующих в соседних реакциях метаболической карты (или расположенных на метаболической карте не далее, чем за три реакции друг от друга), примерно в 16 раз больше вероятность быть закодированными в генах, обладающих ко-локализационной связью, по сравнению с ожидаемыми значениями случайным образом перемешанной выборки [32]. Под ко-локализационной связью в данном случае подразумевается совместное присутствие генов в одних и тех же штаммах бактерий, их ко-локализация на хромосомах или наличие известных случаев слияния таких генов в один. В случае длинных и неразветвленных метаболических путей эта корреляция

распространяется даже на ферменты, находящиеся на расстоянии вплоть до 7 реакций друг от друга на метаболической карте.

Было также показано, однако, что при детальном сравнении эволюционных модулей генов (консервативных сочетания, сохраняющихся от вида к виду) с функциональными гены, кодирующие белки одного метаболического пути, не всегда демонстрируют склонность к консервативному окружению, т.е. эволюционные модули не обязательно тождественны функциональным [32,33]. Более того, в состав одного эволюционного модуля часто входят неполные элементы разных метаболических путей. Таким образом, ко-локализационные свойства генов не всегда определяются их непосредственной функциональной взаимосвязью.

О консервативности оперонов или иных комбинации генов говорят, когда один и тот же набор структурных или функциональных свойств генов, закодированных поблизости на хромосомах, полностью или частично совпадает у разных видов или даже более высоких классификационных таксонов бактерий. Консервативность указывает на определенные эволюционные преимущества такой организации, поскольку демонстрирует действие отбора против потока случайных транслокаций, приводящих к нарушению порядка генов в бактериальных хромосомах.

В случае, когда консервативность касается нуклеотидной последовательности генов, это часто говорит об их родстве, т.е. общем происхождении. Аналогичным образом, существование сходных комбинаций генов в разных геномах может быть связано с наличием такой комбинации у общего предка исследуемой группы бактерий. Оно также может быть связано с событиями горизонтального переноса, при котором генетический материал передается между самыми разными бактериями вне процесса клеточного деления [34,35].

Анализ распространенных комбинаций может быть важным шагом предсказания функций генов. Например, если гены, кодирующие определенные известные функции, располагаются на хромосомах рядом друг с другом, то третий ген с неизвестной функцией, часто обнаруживающийся по соседству с ними, может относится к тому же метаболическому пути [31,36–38] Первым успешным предсказанием такого рода стало определение функции гена, кодирующего шикиматкиназу у археи *Methanococcus jannaschii* [39]. Все гены, относящиеся к пути синтеза хоризмовой кислоты, у данного вида были гомологичны генам, кодирующим соответствующие белки у бактерий. Гена, похожего по последовательности на ген шикиматкиназы бактерий, у архей, однако, не наблюдалось. С помощью сравнительного анализа кластеров генов, относящихся к пути синтеза хоризмовой кислоты у разных архей, был выявлен другой кандидат на данную роль, который часто располагался рядом с остальными генами данного пути. Ферментативная активность продукта этого гена была подтверждена экспериментально, он действительно кодировал шикимат-киназу.

Одной из основных задач данной работы являлось исследование общих тенденций ко-локализации генов углеводного метаболизма у широкого спектра видов бактерий. Такие гены кодируют ферменты, участвующие в реакциях преобразования углеводов, т.е. в их расщеплении и синтезе – гидролазы, киназы, фосфорилазы, дегидратазы, ацетилазы и т.д.; также мы рассматривали транскрипционные факторы, регулирующие работу соответствующих оперонов, и транспортные белки, участвующие в процессах переноса углеводов через бактериальные мембраны.

Комбинации, которые такие гены формируют на бактериальных хромосомах, располагаясь на них подряд, будут в дальнейшем называться кассетами. В данном случае в определении кассеты не учитывается оперонная структура генов, их

порядок и расположение на нитях ДНК. В недавнем исследовании [40] было показано, что из 4,5 миллиона белок-кодирующих генов в большой выборке прокариотических геномов 68,7% формируют консервативные кассеты. В указанной работе под консервативностью авторы подразумевали, что кассеты, совпадающие по комбинации COG (кластеров групп ортологических генов [41,42]), встречались среди исследуемых геномов хотя бы дважды, то есть данный критерий был очень нестрогим, и тем не менее, почти треть генов выборки не прошла его порог. Таким образом, около трети всех известных прокариотических генов, по-видимому, не обладают эволюционно закрепленными связями со своими ближайшими соседями.

Это наблюдение соответствует транскрипционным данным для хорошо изученных организмов, таких, как *Escherichia coli* и *Bacillus subtilis;* известно, что около трети генов в их геномах формируют моноцистронные опероны (состоящие из одного гена) [43], и, таким образом, их транскрипция может регулироваться отдельно от остальных генов и обеспечивать некоторую их эволюционную независимость [44,45].

В настоящей работе мы хотели выяснить, в частности, каким образом приведенные выше утверждения соотносятся со структурой геномных локусов, связанных с углеводным метаболизмом бактерий.

Мы воспользовались базой данных Integrated Microbial Genomics [46] для получения выборки из 148 тысяч генов углеводного метаболизма, относящихся к 264 различным кластерам COG и принадлежащих 665 геномам бактериальных видов 30 разных классов. Нашей задачей было установить, как часто такие гены располагаются на бактериальных хромосомах в кассетах и проанализировать два фактора, которые могли бы влиять на тенденцию к формированию кассет – функцию генов и филогенетические характеристики бактерий. Мы также оценили

разнообразие кассет по размеру и составу, определили наиболее распространенные комбинации генов и сравнили их с комбинациями функций хорошо изученных метаболических путей. Кроме того, мы выявили тенденции к попарным сочетаниям генов разных функций и разных ортологических кластеров, и определили наиболее консервативные попарные связи, попадающие под действие положительного отбора. Наконец, мы изучили случаи ко-локализации генов сходных функций и оценили, насколько велик в них вклад событий локальной дупликации, используя разработанный нами ранее инструмент NSimScan для поиска нуклеотидных последовательностей удаленного сходства.

Общей целью данного этапа работы было получить полномасштабную картину, детально описывающую ко-локализационные тенденции для генов, относящихся к углеводному метаболизму бактерий, и факторы, влияющие на них.

1.3. Экспериментальная проверка предсказаний функций кассет генов

Мы предположили, что сравнительный анализ состава кассет генов углеводного метаболизма, позволяющий выявить консервативные сочетания, дает возможность предсказывать общую функцию кассеты и ее участие в соответствующих метаболических путях. В результате экспериментальной проверки одного из таких предсказаний мы подтвердили наличие связи кассеты генов, кодирующих ферменты сульфогликолиза у *Escherichia coli* с катаболизмом лактозы.

1.3.1 Выбор источника углевода у бактерий и регуляция работы соответствующих генов

Как уже говорилось выше, многие бактерии способны усваивать широкий спектр углеводных субстратов. Бактерии эффективно подстраиваются под изменения в окружающей среде за счет быстрого переключения между разными метаболическими путями. В случае, когда в среде присутствует множество

источников углерода, бактерия выбирает оптимальный путь, который позволяет получить наибольшее количество энергии с наименьшими затратами – в первую очередь, это касается утилизации глюкозы [47,48].

Разные белки, как правило, отвечают за разные функции и могут требоваться организму в разные моменты времени. Известно, однако, что белки могут быть мультифункциональны, участвуя в разных химических реакциях с разными метаболитами; в том числе это касается и некоторых белков углеводного метаболизма бактерий. Например, термофильная гидролаза CoGH1A, принадлежащая бактерии *Caldicellulosiruptor owensensis*, обладает широким спектром действия, и способна катализировать гидролиз самых разных углеводов [49]. Она обладает активностью гликозидазы, экзоглюканазы, ксилозидазы, галактозидазы, а также способна к трансгалактозилированию. Полисахаридная лиаза Smlt1473, принадлежащая бактерии *Stenotrophomonas maltophilia*, катализирует реакции неокислительных разрывов в разных молекулах полисахаридов, в зависимости от уровня pH среды воздействуя на альгиновую кислоту, полиглюкуроновую кислоту или гиалуроновую кислоту [50]. Гексокиназа Glk/TM1469, принадлежащая бактерии *Thermotoga maritima*, катализирует ряд процессов углеводного фосфорилирования, причем механизм ее действия зависит от температурных условий [51]. Однако до сих пор вопрос распространенности мультифункциональных свойств среди бактериальных белков, в целом, остается открытым. Чаще всего об альтернативных функциях белков узнают в ходе экспериментальной работы с конкретными бактериями и ферментами, а исследований с масштабными предсказаниями такого рода до сих пор не проводилось.

Метаболизм большинства известных углеводов у бактерий контролируется с помощью механизма катаболитной репрессии. Принцип его действия заключается

в подавлении экспрессии генов, кодирующих ферменты метаболических путей, относящихся к отсутствующему в данный момент в среде типу углевода [52,53]. В первую очередь он нацелен на использование глюкозы, если она есть в среде, независимо от наличия других источников углерода (Рис. 1). Благодаря этому в клетке поддерживается нормальный энергетический баланс, поскольку ее ресурсы не расходуются без необходимости [54].

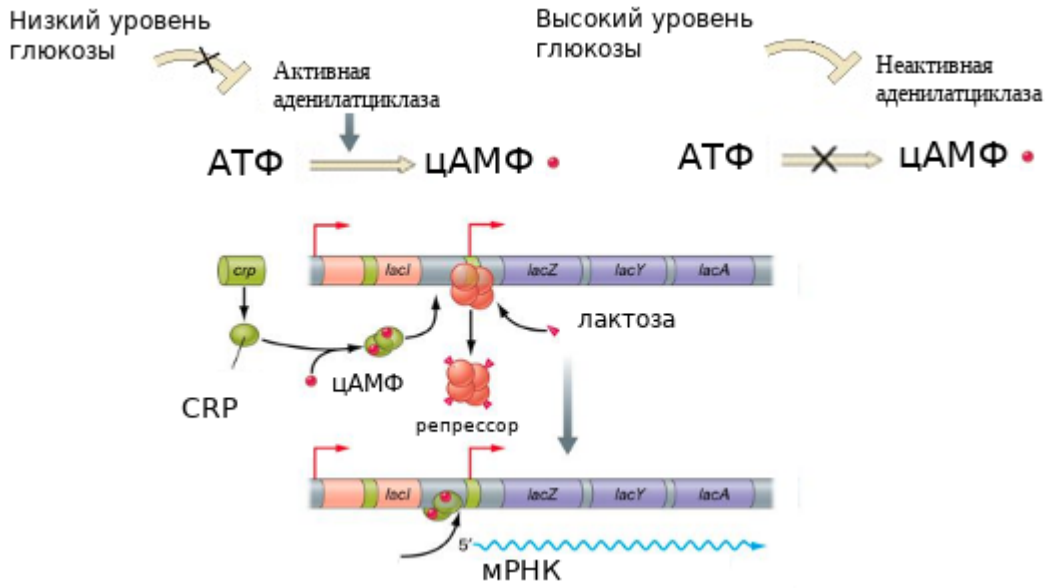

Рис. 1. Механизм катаболитной репрессии на примере регуляции лактозного оперона *Escherichia coli*. При низком уровне глюкозы фермент аденилатциклаза активен и вырабатывает цАМФ (cAMP), который связывается с димеризующимся белком CRP (цАМФ-зависимым катаболит-активируемым белком), и в комплексе с ним позволяет идти активной экспрессии генов *lac*-оперона, кодирующих ферменты катаболизма лактозы (отдельно стоит отметить, что экспрессия происходит только при наличии лактозы в клетке: молекула лактозы связывается с репрессором, в результате чего он перестает блокировать связывание РНК-полимеразы с промотором). При высоком уровне глюкозы, вне зависимости от наличия лактозы в клетке, аденилатциклаза неактивна, синтеза cAMP не происходит, комплекса CRP-cAMP не образуется и экспрессия генов *lac*-оперона значительно понижается.

При отсутствии основного источника углевода или в условиях стресса клетка начинает использовать альтернативные источники энергии [55]. В оперонах

Escherichia coli и родственных ей бактериях этот механизм переключения обычно контролируется с помощью глобального регулятора cAMP-CRP и локального регулятора, который закодирован, как правило, либо внутри оперона вместе с генами ферментов, участвующих в соответствующем метаболическом пути, либо поблизости от них на обратной нити ДНК [56–58].

Регулятор cAMP-CRP, он же CRP (цАМФ-зависимый катаболит-активируемый белок) – один из важнейших факторов транскрипции кишечной палочки и родственных ей бактерий, который регулирует инициацию транскрипции более чем сотни генов [59,60]. Это гены, отвечающие за катаболизм лактозы, галактозы, мальтозы, рибозы, лимонной и других карбоновых кислот, гены, кодирующие фосфотрансферазную систему транспорта (PEP), различные пермеазы и другие имеющие отношение к трансмембранному транспорту белки, а также множество белков других функций. CRP работает в форме димера, как правило, в комплексе с цАМФ, и может как активировать, так и подавлять экспрессию генов. В большинстве случаев, когда его сайт связывания расположен выше относительно старта транскрипции, CRP подавляет экспрессию гена [61,62]. Активация транскрипции с помощью CRP происходит, обычно, в том случае, когда комплекс взаимодействует с сайтом, расположенным выше промотора, и непосредственно взаимодействует с РНК-полимеразой [63]. CRP также способен связываться с ДНК, не образуя комплекс с цАМФ, и это, как правило, приводит к слабому подавлению транскрипции соответствующих генов [61].

Субстрат-зависимая регуляция осуществляется за счет локального регулятора, который принадлежит к таким семействам, как LacI, RpiR, ROK, DeoR, AraC и GntR [64]. Сайты связывания локального регулятора располагаются вблизи от сайтов связывания глобального регулятора (например, CRP или CcpA у бактерий типа Firmicutes), иногда перекрываясь с ними. Часто локальные регуляторы

играют роль, противоположную роли глобальных. Локальные регуляторы, как правило, работают в форме димеров, при этом их димеризация происходит после связывания с углеводом-лигандом и приводит к переключению метаболизма бактерии. Стоит отметить, что функция зависимости активности экспрессии локальных регуляторов от наличия в среде цАМФ и углевода-лиганда может быть нелинейной и подвержена влиянию разных факторов, в том числе, здесь может играть сложную роль структура промотора локального регулятора [65].

В целом работа систем регуляции экспрессии генов углеводного метаболизма нацелена на то, чтобы бактерия использовала именно тот сахар, который в данный момент имеется в окружающей среде или который оказывается оптимален в качестве источника энергии при наличии нескольких вариантов.

1.3.2. Способы утилизации лактозы у бактерии *Escherichia coli*

В единственном известном пути катаболизма лактозы *Escherichia coli* участвуют ферменты, закодированные в хорошо изученном лактозном опероне (*lac*-опероне), впервые описанном Ф. Жакобом и Ж. Моно еще в 1961 году [66]. В 1965 году по результатам этой работы исследователи получили Нобелевскую премию по физиологии и медицине "за открытия, касающиеся генетического контроля синтеза ферментов и вирусов". Описанный ими оперон состоит из трех генов, кодирующих β-галактозидазу (*lacZ*), β-галактозидпермеазу (*lacY*) и β-галактозидтрансацетилазу (*lacA*). β-галактозидпермеаза – мембранный транспортный белок, вторичный транспортер, который переносит лактозу через клеточную мембрану внутрь бактериальной клетки; при этом происходит симпорт лактозы и протона. β-галактозидаза расщепляет лактозу на глюкозу и галактозу (Рис. 2).

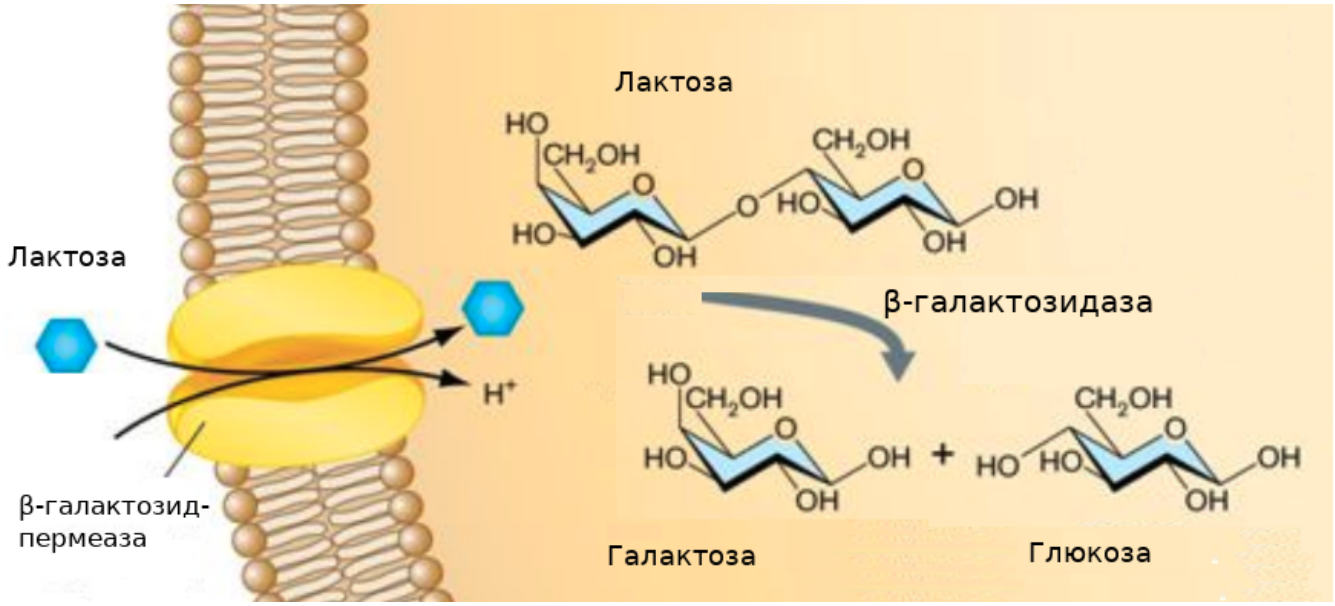

Рис. 2. Лактоза переносится через клеточную мембрану *E. coli* β-галактозидпермеазой одновременно с протоном, после чего расщепляется на глюкозу и галактозу с помощью β-галактозидазы.

β-галактозидтрансацетилаза способна переносить ацетильную группу от ацетил-коА на бета-галактозиды, и ее роль до конца неизвестна, возможно, она нужна для детоксикации неметаболизируемых галактозидов, чтобы они не могли вернуться в клетку после выведения [67].

После гидролиза галактоза и глюкоза, по-видимому, сначала выводятся из клетки, а затем доставляются с помощью других транспортных систем обратно внутрь нее (было показано, что клетки с нарушенным трансмембранным транспортом глюкозы и галактозы не способны эффективно расти на лактозе; гипотеза о выведении и повторном введении метаболитов лактозы была дополнительно подтверждена радиоизотопными методами). Глюкоза подвергается фосфорилированию за счет использования PTS (PEP-транспортера, или фосфоенолпируват-фосфотрансферазной системы), и превращается в глюкозу-1-фосфат. Затем с помощью фосфоглюкомутазы она превращается в глюкозу-6-

фосфат и включается в гликолиз. Галактоза доставляется в клетку путем пассивного транспорта или за счет активных транспортеров, таких, как GalP или Mgl. Галактоза также может превращаться в глюкозу-1-фосфат [68,69].

На момент проведения работы у кишечной палочки не было описано никаких альтернативных способов утилизации лактозы.

1.3.3. Путь утилизации лактозы у бактерий класса *Bacilli*

Для бактерий класса Bacilli, таких как *Streptococcus*, *Staphylococcus* и *Lactococcus* spp, характерен механизм утилизации лактозы с помощью особого метаболического пути, который отличается от распространенного у большинства других видов, в том числе кишечной палочки, где основной стадией является гидролитическое расщепление лактозы на глюкозу и галактозу с помощью β-галактозидазы. В случае класса Bacilli лактоза доставляется в клетку с помощью PTS, в результате чего она сразу фосфорилируется и превращается в D-лактозо-6-фосфат [70]. Затем фермент фосфо-β-галактозидаза осуществляет реакцию гидролиза, в результате чего образуется глюкоза и D-галактозо-6-фосфат [71,72]. D-галактозо-6-фосфат превращается с помощью D-галактозо-6-фосфат-изомеразы в D-тагатозо-6-фосфат [72], это соединение фосфорилирует D-тагатозо-6-фосфат-киназа, образуя D-тагатозо-1,6-бисфосфат [73]. После этого фермент D-тагатозо-1,6-бисфосфат-альдолаза расщепляет это соединение на дигидроксиацетонфосфат и глицеральдегид-3-фосфат [74], которые далее включаются в соответствующие этапы гликолиза (Рис. 3).

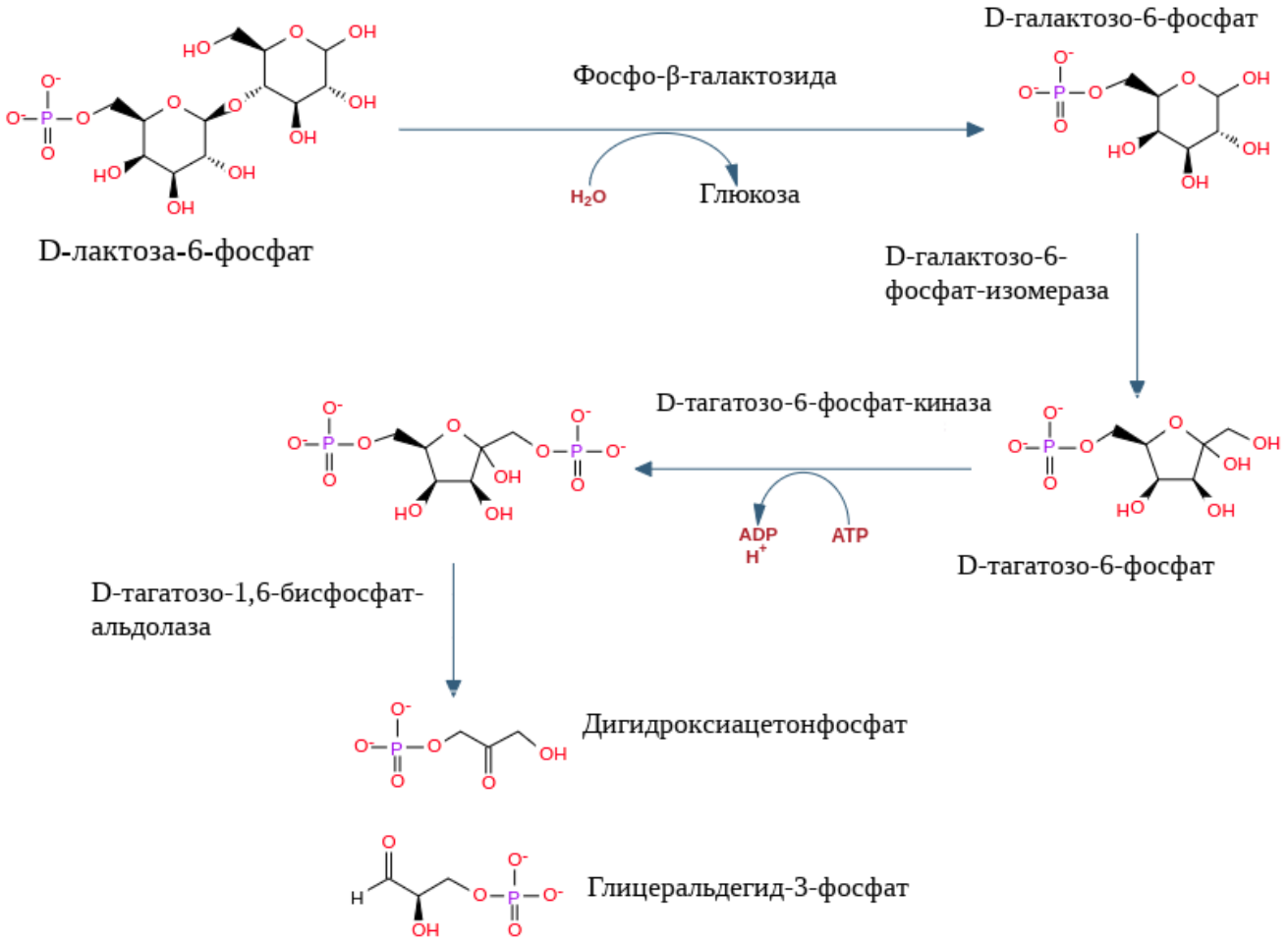

Рис. 3. Схема пути катаболизма лактозы бактерий класса Bacilli

Соответствующие гены закодированы в кассете генов *lacGEFDCBAR* (см. Рис. 4, a). Ген *lacR* кодирует локальный регулятор, от которого зависит экспрессия генов кассеты.

1.3.4. Функции *yih*-кассеты Escherichia coli

К. Денгер и коллеги в недавнем исследовании описали одну из функций кассеты *E. coli ompL-yihOPQRSTUVW* [75]. Эта кассета состоит из 10 генов, названия девяти из которых начинаются с "*yih*"; в дальнейшем мы будем называть ее *yih*-кассетой (Рис. 4, b). Согласно их предположению, работа этих генов завершает

биогеохимический цикл серы в природе, представляя, таким образом, ранее неизвестное его звено. В ходе исследования с помощью анализа ферментативной активности исследователями было показано участие четырех закодированных в *yih*-кассете белков в четырех реакциях сульфогликолитического метаболизма (YihS, YihT, YihU, YihV).

Было выдвинуто предположение, что остальные гены этой кассеты также участвуют в этом пути, а соответствующие белки осуществляют реакции гидролиза и транспорта серосодержащих соединений углеводов, а также регуляцию транскрипции кассеты. При этом механизмы регуляции транскрипции в работе не были рассмотрены, и кассета была представлена как единый оперон, несмотря на ее значительную длину; не были описаны промоторы, сайты связывания транскрипционных факторов и транскрипционный профиль генов кассеты.

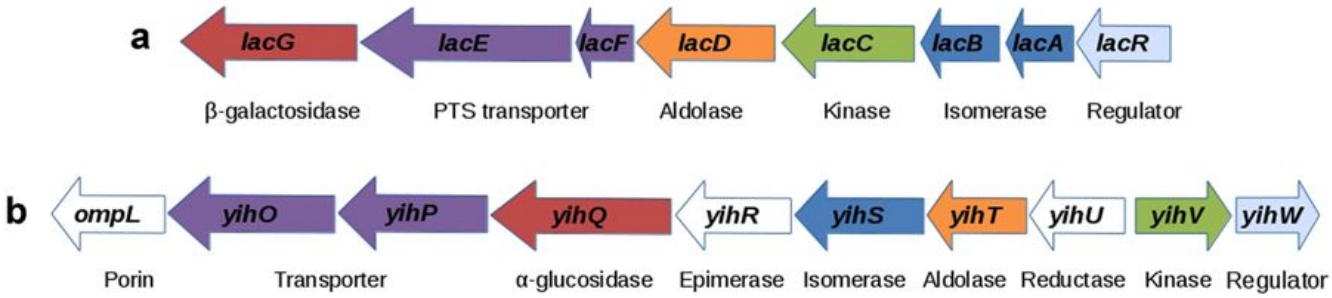

Рис. 4. Кассета бактерий класса Bacilli, участвующая в катаболизме лактозы (a) и кассета семейства Enterobacteriaceae, участвующая в сульфогликолизе (b). Одинаковые цвета генов обозначают пересечение функций кодируемых белков. Белым отмечены гены, кодирующие функции, не представленные в другой кассете.

Ранее никогда не предполагалось, что гены *yih*-кассеты могут участвовать в катаболизме лактозы; наша гипотеза, касающегося такой ее функции, была

основана на результатах общего анализа консервативных кассет катаболизма углеводов, а конкретнее, на сходстве набора функций генов этой кассеты и кассеты Bacilli, участвующей в катаболизме лактозы (Рис. 4). Для того, чтобы выяснить, играют ли эти гены роль в утилизации лактозы, мы провели анализ их экспрессии при росте культуры кишечной палочки *E. coli* K-12 MG1655 на этом субстрате, сравнивая ее с экспрессией при росте на глюкозе. Кроме этого, мы описали промоторы кассеты и сайты связывания транскрипционных факторов, участвующих в переключении работы ее генов.

Глава 2. Инструмент NSimScan для поиска удаленных сходств последовательностей ДНК

2.1. Описание и область применения NSimScan

Инструмент NSimScan (Nucleotide Similarity Scanner) был разработан нами для поиска сходных нуклеотидных последовательностей в больших базах данных ДНК. Он предназначается, в том числе, для проведения филогенетического анализа, предсказания функций генов и для других сравнительных исследований, а также для исследования некодирующих последовательностей и детекции загрязнения образцов ДНК. Производительность NSimScan превосходит инструменты, считающиеся индустриальным стандартом, по совокупности таких параметров, как чувствительность, точность и скорость. По чувствительности NSimScan сравним с BLAST [8] и USEARCH [11], по точности – с SSearch [9], а по скорости – с MegaBLAST [10]. Наилучшим образом он подходит для поиска последовательностей, отличающихся друг от друга на 10-40 процентов.

2.2. Алгоритм работы NSimScan

Инструмент NSimScan представляет собой генератор предполагаемых участков сходства (первичных совпадений), объединенных с серией фильтров с увеличивающейся вычислительной нагрузкой, принцип организации которой соответствует ранее разработанному нами принципу для инструмента поиска сходств в белковых последовательностях (PSimScan) [76].

Последовательные этапы алгоритма приведены ниже.

1) При запуске NSimScan сначала прочитывает все искомые последовательности (queries) и составляет индексную таблицу, в которой хранит координаты всех k-

меров каждой последовательности. Величина k-мера задается в качестве вводного параметра. По умолчанию этот параметр составляет 11, оптимальный диапазон его составляет от 8 до 12. Таблица адресуется непосредственно двоично упакованным представлением последовательности k-мера. Это весьма существенный момент, отличающий инструмент NSimScan от остальных, и позволяющий значительно ускорять процедуру поиска в таблице. В случае, если включен параметр "--approximate", в таблицу также вносятся неточные совпадения k-меров, с возможностью замены одного нуклеотида в любой позиции k-мера.

2) Если при составлении индексной таблицы в искомых последовательностях встречаются последовательно расположенные повторяющиеся k-меры, то в таблицу они не записываются. Рамки данного фильтра контролируются с помощью параметра "--kred". Он определяет минимальное расстояние между позициями одинаковых k-меров в искомой последовательности, при котором они еще вносятся в таблицу.

3) Как вариант, алгоритм может учитывать список частот k-меров; он читается из внешнего файла, заданного с помощью параметра "--kdistr". На основании этого списка вычисляются относительные веса k-меров. Если список не предоставлен, то веса k-меров считаются одинаковыми и составляют 100 каждый.

4) Выделяется место в памяти для списка диагоналей матрицы совпадений. Число диагоналей соответствует сумме длин искомых последовательностей и длины последовательности из базы данных, с которой проводится сравнение. Если поиск проводится по обеим нитям ДНК, число искомых позиций удваивается (по умолчанию этот параметр включен). Объект "диагональ" содержит данные по суммарному количеству совпавших k-меров на данной диагонали и нескольких соседних. Максимальная удаленность диагонали в такой группе определяется параметром "--mxshift", по умолчанию он составляет 3.

5) Последовательно прочитываются все нуклеотидные последовательности ("записи") из базы данных, в которой осуществляется поиск.

6) Для каждой позиции из каждой записи проверяется наличие соответствующего k-мера в индексной таблице. С целью ускорения работы проверяться может не каждая позиция, а каждая вторая, третья, четвертая и т.д. Шаг выборки k-меров контролируется параметром «-q» (или «--step»), по умолчанию он равен 1. При поиске последовательностей с высокой степенью сходства использование шага большего чем 1 может существенно (обратно пропорционально размеру шага) сокращать время поиска без уменьшения точности.

7) Первичные вхождения используются для обновления значений весов диагоналей матрицы сходства у каждой позиции из записи. Процедура вычисления веса диагонали использует вес k-мера, расстояние до ближайших вхождений (перекрывающихся или отдельных) на текущей и соседних диагоналях, а также может учитывать статистическую значимость вхождения и степень вырожденности окрестности вхождения. Если вхождение изолированно, то вес соответствующего k-мера добавляется к весу диагонали. Если вхождения на предыдущей позиции данной диагонали уже зафиксированы, к ним добавляется вес не перекрывающейся с ними части текущего вхождения. Если предыдущее, не перекрывающиеся с текущим, вхождение было зафиксировано на соседней диагонали, расположенной не дальше, чем значение параметра "--mxshift", то вес переносится на текущую диагональ, с вычитанием штрафа за пропуск по количеству нуклеотидов между вхождениями ("gap cost").

8) Когда вес диагонали превышает порог, заданный параметром "--kthresh" (по умолчанию он составляет 250, что приблизительно соответствует трем

перекрывающимся вхождениям k-меров с весом 100), такая диагональ успешно проходит фильтр оценки выравнивания.

9) Для прошедших первичный фильтр диагоналей далее строится субоптимальное выравнивание. Для этого используется жадный эвристический алгоритм, который составляет выравнивание за одно прохождение по диагонали путем последовательного расширения зоны сходства по текущей и нескольким соседним диагоналям в обоих направлениях, пока вес выравнивания остается положительным. Это очень быстрая процедура, поскольку она является линейной – скорость зависит только от длины выравнивания. Высокая эффективность процедуры также достигается благодаря тому, что выравнивание осуществляется с помощью битовых операций над упакованными последовательностями (искомые последовательности и записи из базы данных представлены в бинарном виде). Во время построения выравниваний величина пропусков не превышает значение задаваемого в командной строке параметра "--mxshift".

10) Полученные выравнивания пропускаются через фильтр соотношения длины и процента сходства. Данный фильтр контролируется с помощью трех параметров: минимальная длина выравнивания ("--minlen"), процент сходства последовательностей на минимальной длине ("--minthr") и процент сходства на полной длине выравнивания ("--maxthr"). Длина выравнивания прошедших фильтр должна превышать минимальную, а процент сходства должен составлять более (minlen x minthr + (alignment_length – minlen) x maxthr).

11) Выравнивания могут также проходить проверку на наличие тандемных повторов/перепредставленности каких-либо фрагментов. В этом случае вычисляется вес выравниваний, последовательно полученных при передвижении вперед и назад по одной из нитей последовательности на число позиций, ограниченных параметром "--replen". Их вес сравнивается с весом исходного

выравнивания. Если новый вес оказывается больше, и соотношение нового и старого веса не меньше параметра "--replev", последовательность считается перепредставленной, и выравнивание не соответствует требованиям фильтра. По умолчанию эта проверка включена, параметр --replen составляет 4, а --replev составляет 50. Выключить проверку можно, указав в командной строке --replen 0.

12) Поскольку оценка параметров проводится каждый раз, когда вес диагонали оказывается выше порога --kthresh, для некоторых позиций выстраивается серия длинных и относительно хороших выравниваний, которые могут несколько различаться в силу эвристического алгоритма их построения. Из таких перекрывающихся выравниваний выбирается одно, обладающее наибольшим весом.

13) Выравнивания не могут содержать внутри себя вставки или делеции ("gaps") длиннее, чем величина параметра --mxshift. Поэтому выравнивания часто представляют собой серию коротких "доменов". Как вариант, они могут по окончанию процесса поиска сходств сливаться в единое выравнивание. Это достигается за счет включения параметра "--mdom" в командной строке; он запускает отдельный алгоритм, основанный на динамическом программировании, отыскивающий оптимальную комбинацию коротких выравниваний для слияния.

14) Последовательности из базы данных, содержащие большое количество повторов, приводят к появлению множества сходных выравниваний. Если интерес представляет только один, лучший представитель из базы данных, можно отфильтровать остальные путем включения параметра "--mrep" в командной строке.

15) Количество выравниваний на одну искомую запись ограничено параметром "--rpq". По умолчанию он составляет 500. Если количество найденных

соответствий будет его превышать, в результатах будет представлено только 500 лучших выравниваний. При необходимости его можно увеличить.

16) Количество лучших выравниваний, которые записываются и хранятся для каждой записи из базы данных, можно ограничивать, задав нужное число с помощью параметра "--rps". По умолчанию этот параметр отключен.

Таким образом, основными задаваемыми параметрами программы являются размер k-мера, весовой порог диагонали (первичный фильтр) и параметры вторичного фильтра для выравнивания: минимальная длина выравнивания, минимальная доля сходства на минимальной длине и минимальная доля сходства на полной длине. Отношение сигнала к шуму можно дополнительно улучшать, предоставив таблицу частот k-меров или объединяя k-меры, обладающие определенным количеством различий.

Инструмент NSimScan доступен для скачивания на сайте https://github.com/abadona/qsimscan. Подробное руководство по применению и примеры параметров для командной строки находятся по адресу https://github.com/abadona/qsimscan/blob/master/nsimscan_users_guide.txt.

2.3. Методы оценки эффективности работы NSimScan

Для того, чтобы оценить эффективность работы инструмента NSimScan, в качестве стандартной выборки мы использовали бактериальные гены, кодирующие рибосомные белки. Семейства таких белков достаточно консервативны и хорошо изучены [77]. Всего мы использовали 1244 бактериальных генома разных видов, для которых мы выбрали 53 семейства генов, кодирующих рибосомные белки, так, чтобы каждое из этих семейств имело более 600 аннотированных представителей. Для каждого семейства мы случайным образом выбрали по 200 представителей, получив таким образом 10600 последовательностей.

Каждую их них мы сравнивали со всем остальным набором. Совпадение с представителями своего семейства считалось истинно положительными результатом (TP), совпадение с представителями чужих семейств – ложно положительным (FP), отсутствие совпадения между членами одного семейства – ложно-отрицательным (FN), а отсутствие совпадения между членами разных семейств – истинно-отрицательным (TN).

Для полученных совпадений, отсортированных по ожидаемым значениям (e-value – этот параметр был получен для всех возможных пар рибосомных генов с помощью инструмента SSearch [9]), мы вычислили точность, т.е. количество ошибок на каждый поиск – FP/(общее число последовательностей) и чувствительность, которая также называется покрытием – TP/(TP + FN). Соответствующие данные представлены в виде графика (Рис. 5).

Мы также протестировали скорость работы NSimScan на другой модельной задаче – в рамках филогенетического анализа большого набора метагеномных данных. Для этого мы провели поиск репрезентативных последовательностей 16S РНК для 749 таксонов из базы данных Silva версии 123 [78] против образца метагенома корней огурца SRR908208 из базы данных NCBI Short Read Archive [79], в котором содержалось 67 миллионов 200-нуклеотидных парно-концевых последовательностей.

Все тесты проводились на компьютере с процессором Intel Core i7-3820, работающем на 3.60 GHz, с 64 гигабайтами оперативной памяти DDR3 и 2-террабайтным жестким диском SATA3, с операционной системой Fedora 21 Linux OS.

2.4. Результаты сравнения производительности NSimScan с другими инструментами

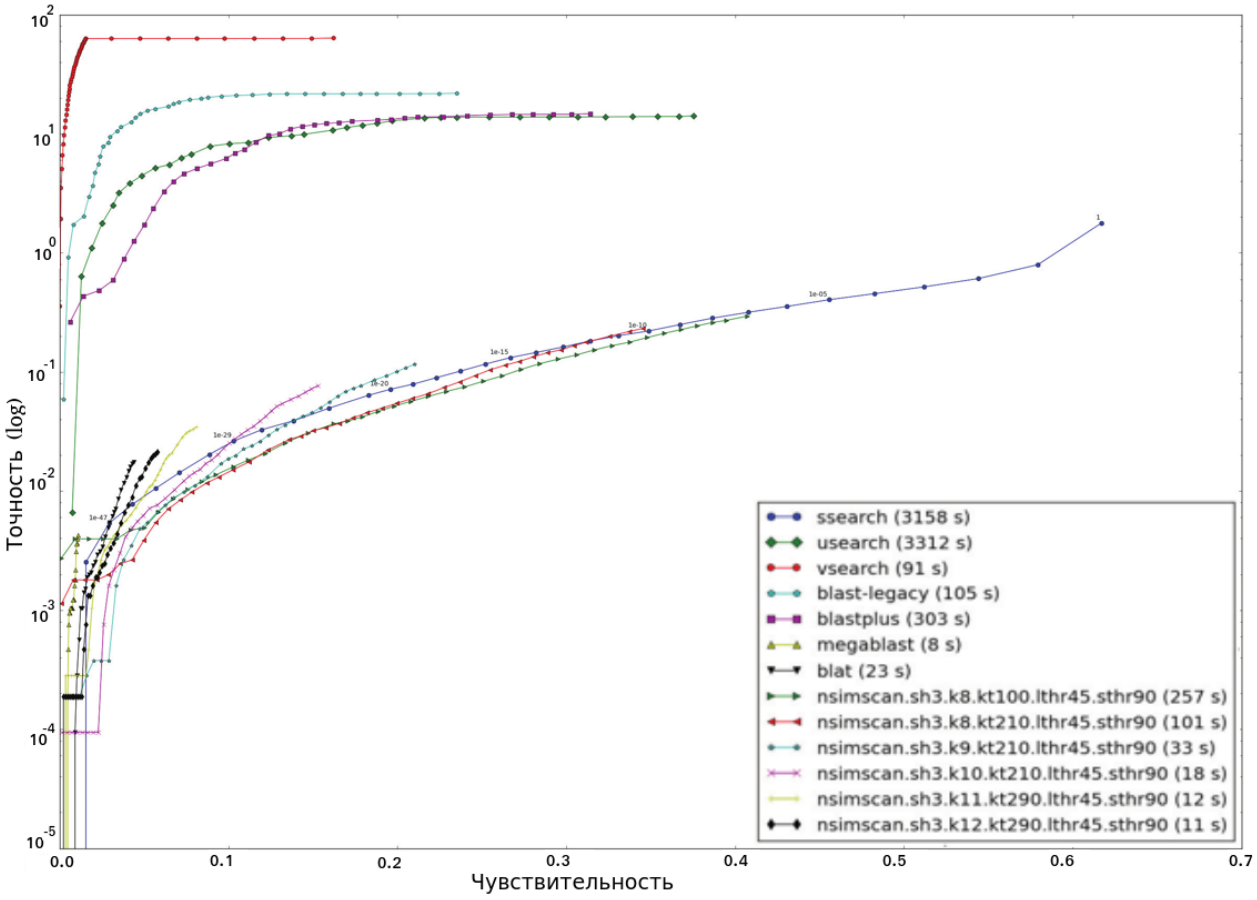

Рис. 5. Эффективность работы инструмента NSimScan, представленная в виде точности поиска (по вертикали, логарифмическая шкала) относительно его чувствительности (по горизонтали) при разных собственных параметрах (с возрастающей строгостью первичного отбора вхождений), и в сравнении с другими, стандартными в области инструментами: SSearch, USEARCH, BLAT, BLAST (современная версия BLAST plus и исходная версия Legacy BLAST) и MegaBLAST. Параметры NSimScan, указанные в легенде: *sh* – максимальный сдвиг по диагонали; *k* – размер k-мера; *kt* – порог веса диагонали; *lthr* – наименьшее совпадение на полной длине выравнивания; (*sthr* – наименьшее совпадение на минимальной длине выравнивания, составляющей 40 нуклеотидов). В скобках указано время работы каждого инструмента в секундах.

На Рис. 5 показаны результатов шести запусков инструмента NSimScan с разными параметрами первичных фильтров, а также результаты работы

стандартных инструментов, которые часто применяют в исследованиях в области сравнительной геномики. Стоит отметить, что дополнительное ужесточение вторичных параметров сдвигает часть графика с более высокой чувствительностью в сторону большей точности, а ослабление их добавляет сегмент справа, соответствующий меньшей точности, т.е. более высокому проценту ошибок.

Во всех вариантах заданных условий NSimScan по соотношению чувствительности и точности оказался на уровне SSearch (который считается самым чувствительным инструментом среди указанных), обогнав при этом все остальные инструменты на два порядка. Даже с наименее жесткими первичными параметрами поиска (большой длиной k-мера и низким порогом веса диагонали) NSimScan работает с несколько большей чувствительностью, чем USEARCH (который демонстрирует наилучшие показатели по чувствительности среди остальных инструментов, кроме SSearch), при этом точность у NSimScan почти на два порядка выше, а скорость выше на порядок (в ходе данного сравнения USEARCH запускался без включения параметра "--usort", который увеличивал бы скорость его работы в сто раз, но приводил бы к стократной потере в чувствительности). По скорости NSimScan сопоставим с MegaBLAST при средних порогах чувствительности и с BLAST при высоких порогах. Он не уступает уровню BLAST по чувствительности, при этом скорость работы NSimScan при соответствующих параметрах оказывается в три раза выше.

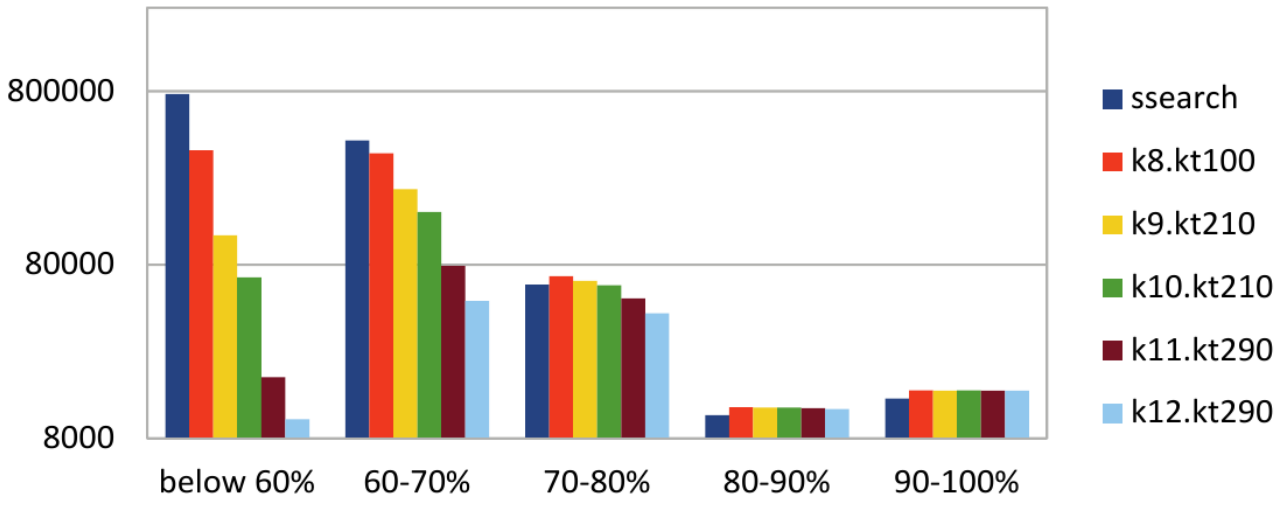

Рис. 6. Распределение чувствительности поиска для последовательностей с разной долей сходства. Справа указаны использованные инструменты: SSearch (синий) и NSimScan с разными параметрами: k (величина k-мера) и kt (порог веса диагонали). По оси ординат указано количество найденных последовательностей. По оси абсцисс – доля сходства итоговых выравниваний.

На Рис. 6 представлено сравнение результатов работы инструментов NSimScan и SSearch при разной заданной доле сходства последовательностей. Выяснилось, что при доле сходства последовательностей более 70% SSearch и NSimScan показывают почти одинаковую чувствительность. Ниже 70% чувствительность NSimScan гораздо сильнее зависит от параметров, связанных со скоростью (величины k-мера и порога оценки выравнивания) и, соответственно, существенно падает при высоких скоростях.

Tool	Time	MemUse	Detected	Tx#	MissTx#	ExtraTx#
MegaBLAST	5 h 18 min	24.6 Gb	252956	310	n/a	n/a
NSimScan	26 min	5.7 Gb	240934	360	4.7%	21.6%

Таблица 1. Сравнение работы NSimScan и MegaBLAST в рамках филогенетического анализа. В колонках: Tool – инструмент, Time – время работы инструмента; MemUse – количество использованной оперативной памяти; Detected – количество обнаруженных фрагментов 16S РНК; Tx# – количество выявленных бактериальных таксонов, MissTx# – количество таксонов, которые обнаружены с помощью MegaBLAST, но не обнаружены с помощью NSimScan, ExtraTx# – количество таксонов, которые обнаружены с помощью NSimScan, но не обнаружены с помощью MegaBLAST.

Результаты работы NSimScan в сравнении с инструментом MegaBLAST в эксперименте по филогенетическому анализу большого набора метагеномных данных представлены в Таблице 1. NSimScan работает в 10 раз быстрее, чем MegaBLAST, использует в 4 раза меньше оперативной памяти, и находит практически все искомые фрагменты – более 95% таксонов, которые находит MegaBLAST и существенное количество таксонов (21,6%), которые MegaBLAST не находит.

2.5. Применение NSimScan в научных исследованиях

NsimScan был успешно использован исследовательской группой из калифорнийского института Joint Genomic Institute для вычисления УНР (усредненного нуклеотидного расстояния) в широкомасштабном филогенетическом исследовании, включающем данные геномов 3032 видов прокариот [80]. Мы рассчитываем, что данный инструмент будет полезен и в других, самых разнообразных проектах, требующих эффективного обнаружения нуклеотидных последовательностей удаленного сходства.

В ходе данной работы инструмент NsimScan был использован для оценки количества событий дупликаций генов углеводного метаболизма в бактериальных геномах (см. Главу 3), а также для оценки сходства ортологов генов *yihT*/*lacD* бактерий семейства Enterobacteriaceae (см. Главу 4).

2.6. Заключение

Мы провели сравнение производительности разработанного нами инструмента NSimScan для поиска нуклеотидных последовательностей удаленного сходства в больших базах данных с инструментами, являющимися промышленным стандартом в области сравнительной геномики, и продемонстрировали значительные преимущества NSimScan по совокупности таких параметров, как точность, скорость и чувствительность. NSimScan работает со скоростью, соответствующей самым быстрым инструментам из всех представленных в области, в том числе, MegaBLAST. Чувствительность поиска NSimScan сравнима с показателями самого чувствительного инструмента, SSearch. Работа NSimScan характеризуется при этом высокой степенью точности; уровень ошибок NSimScan соответствует или оказывается ниже, чем у SSearch, и всегда ниже, чем у всех остальных протестированных инструментов. Поиск с использованием NSimScan оказывается также полноценным – инструмент отыскивает все последовательности, имеющие указанную долю сходства, не теряя никаких результатов.

Наибольшее преимущество NSimScan представляет при поиске относительно далеких последовательностей (60-90% идентичности) на больших наборах данных в рамках широкомасштабных проектов анализа последовательностей ДНК.

Глава 3. Организация генов углеводного метаболизма бактерий

3.1. Материалы и методы

3.1.1. Геномы и гены

Всего было изучено 665 бактериальных геномов разных видов, штамм каждого вида выбирался случайным образом (см. Приложение А). Общее количество исследованных генов углеводного метаболизма бактерий составило 148 тысяч.

Данные по аннотации генов были получены из базы данных IMG [46,81]; большая часть была принадлежала категории "G" – углеводного метаболизма. Дополнительные гены были взяты из других категорий (например, категории генов с неизвестной функцией или категории генов, участвующих в построении клеточной стенки) на основании наблюдения событий слияний таких генов в некоторых бактериях с генами категории "G" (см. раздел 3.1.2). Аннотация каждого гена содержала его подтвержденные или предсказанные функции и его координаты на бактериальной хромосоме, а также указывала на его принадлежность к определенным кластерам COG.

Последовательности генов были взяты из базы данных GenBank [82].

3.1.2. Классификация генов углеводного метаболизма бактерий

Мы использовали двухуровневую классификацию генов. Первый ее уровень, классы, соответствовал глобальной функции гена и учитывал реакционную и субстратную специфичность соответствующих ферментов. Гены, кодирующие транспортные белки и транскрипционные факторы, были вынесены в два отдельных класса. Принадлежность ферментов к определенному классу определялась с помощью международной иерархической классификации Enzyme

Nomenclature, созданный Комиссией по ферментам при Международном союзе биохимии и молекулярной биологии IUBMB [83].

Каждый полный классификационный номер этой системы содержит последовательность из четырёх чисел, разделённых точкой. Каждое число представляет собой всё более уточняющую классификацию фермента. Первое число соответствует одному из семи главных типов ферментов – оксидоредуктазы (1), трансферазы (2), гидролазы (3), лиазы (4), изомеразы (5), лигазы (6) и транслоказы (7). Второе число характеризует основной тип субстрата. Например, у трансфераз вторая цифра указывает на природу той группы, которая подвергается переносу, у гидролаз – на тип гидролизуемой связи и т. д. Третье число более конкретно уточняет природу химических соединений доноров или акцепторов, участвующих в данной реакции. Четвертое число, как правило, определяет конкретную специфичность фермента, например, то, что он взаимодействует конкретно с альбумином или фруктозой. Так, фосфофруктокиназа имеет номер 2.7.1.56, где число 2 соответствует трансферазам, 7 – трансферазам, переносящим фосфатный остаток (фосфотрансферазам, они же киназы), 1 – фосфотрансферазам, акцептором для которых является гидроксильная группа, а 56 – киназам, переносящим фосфатный остаток на молекулы фруктозы.

Всего мы определили 19 классов функций генов, относящихся к углеводному метаболизму, в том числе гликозидазы, киназы, изомеразы и т.п. (см. Таблицу 1).

Функциональный класс	Количество генов	Склонность к образованию кассет	Идентификатор Enzyme Nomenclature
транскрипционные факторы (transcriptional)	39136	35,29%	-
транспортные белки (transport)	29701	70,83%	-
гликозилтрансферазы (glycosyltransferase)	14579	62,30%	2.4.1.
гликозидазы (glycosidase)	11475	64,74%	3.2.1.
киназы (kinase)	9250	57,95%	2.7.1.; 2.7.9
изомеразы (isomerase)	6458	55,20%	5.3.1.
дегидрогеназы-OH (dehydrogenase-OH)	5518	57,67%	1.1.
декарбоксилазы (decarboxylase)	2788	58,97%	4.1.
нуклеотидилтрансферазы (nucleotydiltransferase)	2125	70,96%	2.7.7.; 2.7.8
дегидратазы (dehydratase)	2091	52,75%	4.2.
фосфотазы (phosphotase)	2036	37,77%	3.1.3.
эпимеразы (epimerase)	1753	61,78%	5.1.3.
деацетилазы (deacetylase)	1525	51,02%	3.5.1.
трансальдолазы/транскетолазы (transaldolase/transketolase)	1514	70,54%	2.2.1.
мутазы (mutase)	1502	40,35%	5.4.2.
карбоксил-эстеразы (carboxylic-esterase)	1153	63,49%	3.1.1.
дегидрогеназы-O (dehydrogenase-O)	781	69,78%	1.2.
нуклеозидазы (nucleosidase)	597	23,28%	3.2.2.
мальто-олигозилтрегалоз-синтазы (malto-oligosyltrehalose-synthase)	100	93,00%	5.4.99

Таблица 2. Функциональные классы генов углеводного метаболизма.

Второй уровень классификации соответствовал структурно-эволюционным характеристикам гена, отраженным в его принадлежности к определенному COG (кластеру групп ортологических генов) [41,42]. В базе данных IMG гены распределяются по кластерам с помощью автоматизированной процедуры, в ходе которой осуществляется поиск нуклеотидной последовательности гена с помощью инструмента RPS-BLAST против позиционных весовых матриц PSSM (COG position-specific scoring matrices), составленной на основе базы данных консервативных доменов CDD (conserved domains database) [84]. Из этой базы

данных мы взяли 239 бактериальных кластера COG из категории "G", которые встречались среди выбранных нами штаммов 665 видов бактерий.

Около 2% генов, относящихся к данным кластерам, имели также дополнительные идентификационные номера COG; такой результат автоматизированной аннотации может указывать на события слияния генов [85]. В этом случае последовательности двух разных генов, кодирующих разные белки и представленных в одних геномах по отдельности, в других геномах оказываются входящими в состав одного гена, и кодируют один белок, но с несколькими доменами. Согласно недавним исследованиям, около 6% генов бактерий и архей, по-видимому, являются результатом события слияния двух и более генов [40]. Чаще всего такие события являются свидетельством тесной функциональной связи соответствующих белков.

Поскольку нашей задачей было, в частности, изучение подобных связей, в рамках данного исследования случаи потенциальных событий слияний генов рассматривались так же, как случаи отдельных ко-локализованных генов. Анализ всех потенциальных событий слияния генов углеводного метаболизма с другими генами выявил 34 дополнительных кластера COG, аннотации которых указывали на их возможную принадлежность к углеводному метаболизму. Большинство из них принадлежало, согласно данным базы данных IMG, к категории "M" (биосинтез клеточной стенки/мембраны), "R" (гены с предсказанной общей функцией) и "K" (транскрипция). Мы включили эти 34 кластера в исследование.

Примером такого дополнительного кластера является COG4158 из категории "R", в аннотации базы данных IMG которого предсказано, что входящие в него гены, в частности, кодируют "белки из семейства CUT2 ABC-транспортеров моносахаридов" и "ABC-транспортеры рибозы, пермеазы". Данный кластер был отнесен к в рамках нашей классификации к классу транспортеров.

В результате мы получили набор из 264 кластеров COG (см. Приложение Б).

3.1.3. Определение кассет генов и их анализ

Кассеты были определены на основании ко-локализации генов на бактериальных хромосомах. Считалось, что гены формировали кассеты, если они были включены в составленную нами классификации генов углеводного метаболизма и располагались на хромосоме подряд, причем расстояние между каждой парой не превышало 200 нуклеотидов. Данный критерий по порогу ко-локализации генов внутри оперонов был получен из OperonDB [86], крупной базы данных, содержащей предсказанные и подтвержденные оперонные структуры прокариот. Порядок генов в кассетах и их расположение на нитях ДНК не учитывались.

В кассете был разрешен один длинный интервал длиной 1500 нуклеотидов, что приблизительно соответствует длине одного бактериального гена и двух межгенных интервалов, его окружающих. Это позволяло включать в кассету один дополнительный ген, для которого еще не было показано участие в углеводном метаболизме, например, ген с неизвестными функциями. Исходя из окружения такого гена, можно предположить, что он тоже может иметь отношение к углеводному метаболизму [38], и такое допущение позволяло не нарушать структуру целой кассеты, не существенно увеличивая при этом количество и длину кассет в целом.

С помощью языка программирования Python мы разработали инструменты, которые позволили проводить дальнейшие исследования ко-локализационных тенденций генов, в частности, проанализировать полученные кассеты по размеру (количеству входящих в них генов) и составу в целом и в разных бактериальных таксонах, исследовать их разнообразие на уровне кластеров COG и на уровне

функциональных классов и выявить наиболее консервативные комбинации. Эти комбинации мы также сравнивали с наборами участников известных метаболических путей, взятых из баз данных Metacyc [20] и KEGG [19].

3.1.4. Анализ ко-локализационных особенностей функциональных классов

Одной из целей нашего исследования был анализ ко-локализационных тенденций генов, принадлежащих к разным функциональным классам. Чтобы выявить статистическую значимость таких событий, мы сравнивали их со случайной моделью. Для этого мы случайным образом перемешали исследуемые гены 10000 раз по их позициям на бактериальных геномах (отдельно в каждом геноме) и вычислили, как часто пары генов из разных функциональных классов встречаются друг с другом в такой модели. При этом событием ко-локализации считался случай, в которой в одной кассете оба соответствующих функциональных класса были представлены хотя бы один раз.

Полученное распределение мы использовали для того, чтобы рассчитать вероятность ошибки при отклонении нулевой гипотезы (p-value или P-значения) для настоящих событий ко-локализации. P-значение в данном случае соответствует вероятности того, что случайная величина с данным распределением примет значение, не меньшее, чем фактическое значение. Если функциональные классы вообще не встречались в случайной модели, P-значение для данной пары приравнивалось 1/10001.

После этого к общему числу проанализированных пар мы применили поправку для множественных гипотез (поправку Бонферрони) [87] с уровнем значимости $\alpha = 0,05$.

Такой же статистический анализ проводился для случаев ко-локализации генов одного и того же функционального класса. В данном случае отдельно считались

события встречи в одной кассете ровно двух представителей класса, ровно трех представителей и т.п.

3.1.5. Анализ ко-локализационных особенностей кластеров COG

Мы проанализировали ко-локализационные тенденции представителей разных кластеров COG внутри каждой пары функциональных классов. Чтобы учесть количество генов в разных COG, мы сравнивали наблюдаемые количества событий встреч с ожидаемыми, которые зависели от размеров соответствующих COG. Для статистической проверки с помощью критерия хи-квадрат (где согласно нулевой гипотезе количество событий ко-локализации кластеров зависит только от размера кластеров) для каждой пары было получено значение квадрата разницы между наблюдаемым и ожидаемым значениями.

Для того, чтобы разделить встречи между разными COG на частые, редкие и промежуточные варианты, события встреч COG в рамках каждой пары функциональных классов были кластеризованы с помощью алгоритма *k*-средних [88], реализованной на языке программирования Perl. Этот алгоритм позволяет кластеризовать числовые данные в многомерных пространствах (минимизируя суммарное квадратичное отклонение точек кластеров от центров этих кластеров).

Процесс кластеризации был повторен несколько раз с возрастающим числом кластеров и параллельно возрастающим штрафом, зависящим от квадрата количества кластеров. В результате для каждой пары функциональных классов было найдено оптимальное количество типов частоты встречаемости пар COG.

Подобная же процедура кластеризации была проведена для описанных выше значений нулевой гипотезы. Если частота встречаемости пар оказывалась существенно выше ожидаемых значений, это означало, что она определялась не только размером соответствующих COG. Были выявлены пары COG, которые

оказались включены в кластеры с самыми высокими значениями как в рамках первой, так и второй кластеризации, т.е. такие пары, которые встречались в одних и тех же кассетах чаще всего, и эти встречи не были случайными.

3.1.6. Сравнение последовательностей генов

Чтобы выяснить, в каких случаях расположенные рядом гены, относящиеся к одному и тому же кластеру COG, являются результатом события локальной дупликации, мы сравнивали последовательности этих генов друг с другом и со всеми остальными генами того же кластера из нашей базы данных. Мы выбирали двухсторонние лучшие совпадения (bi-directional best hits), то есть пары, в которых первый ген был больше всего похож на второй, а второй больше всего похож на первый.

Для этого мы использовали разработанный нами инструмент NSimScan (см. Главу 2) со следующими параметрами: -k (размер k-мера) 7; -t (порог оценки по диагонали) 80; --it (минимальный процент сходства на минимальной допущенной длине выравнивания) 50; --xt (минимальный процент сходства на максимальной возможной длине выравнивания) 50; –mrep (данный параметр включает режим, при котором из группы найденных результатов в рамках одного и того же генома выбирается только один, самый лучший).

3.2. Результаты и обсуждение

3.2.1. Склонность генов к ко-локализации и разнообразие кассет генов

Только 53% из 148 тысяч бактериальных генов углеводного метаболизма формировали кассеты, то есть располагались рядом друг с другом на бактериальных хромосомах. Изначально мы ожидали увидеть более сильную тенденцию к ко-локализации у генов, белки которых потенциально выполняют

взаимосвязанные функции [25,30,89,90]. Известно, однако, что эволюционные модули, состоящие из групп генов, всегда одновременно присутствующих или отсутствующих в геномах или даже непосредственно рядом друг с другом, не обязательно тождественны функциональным модулям [32,33]. Кроме того, как уже говорилось выше, в исследовании большой выборки прокариотических генов всевозможных функций было показано, что менее двух третей из них формируют консервативные кассеты, т.е. имеют хоть сколько-нибудь заметную склонность к эволюционной устойчивости своего окружения [40].

Всего исследуемые гены вошли в состав 26 тысяч кассет. Большая часть этих кассет были короткими; 55% состояли из двух генов, 20% – из трех (Рис. 7).

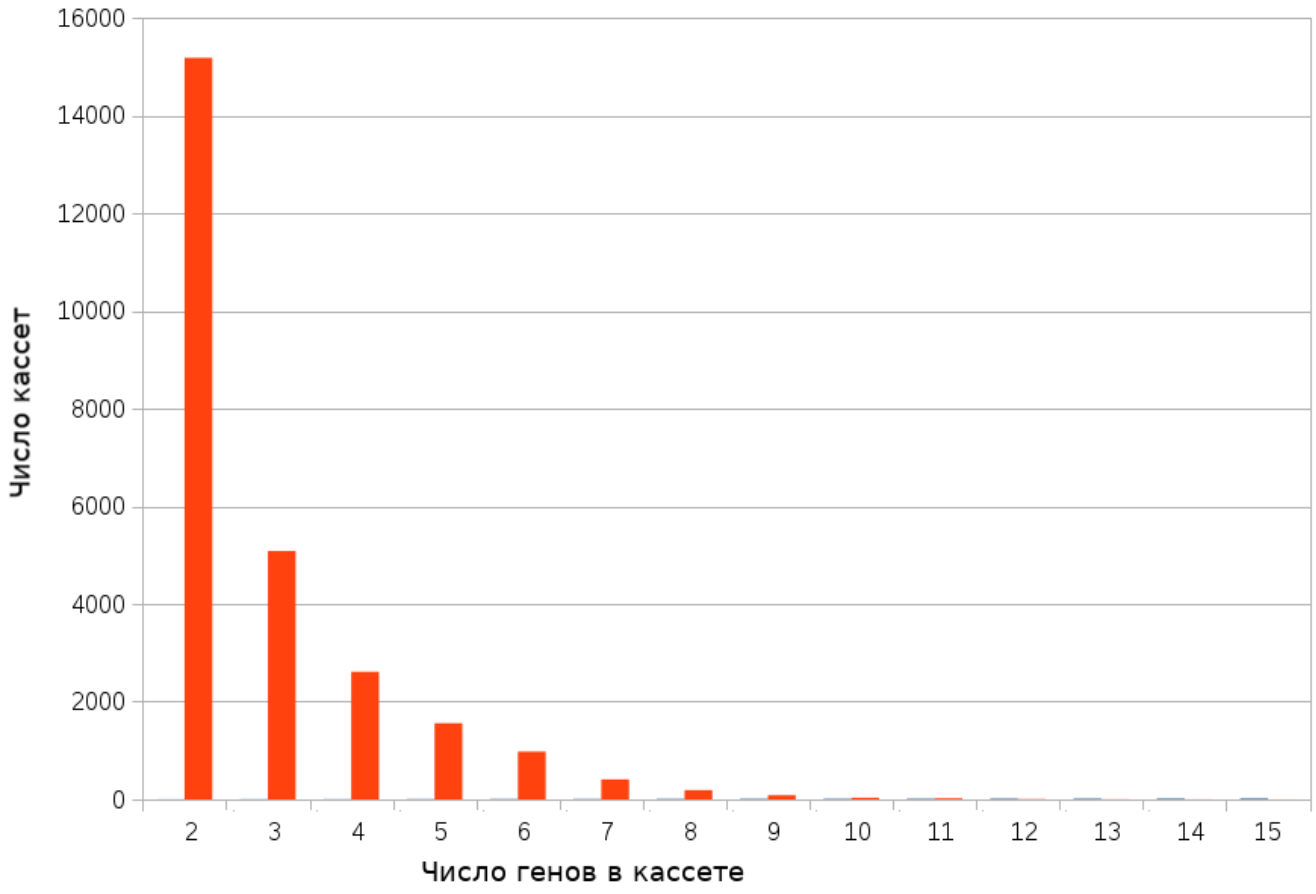

Рис. 7. Распределение количества кассет в зависимости от их размера (числа генов в кассете)

Распределение кассет по размеру среди разных классов бактерий, а также распределение функциональных классов генов в кассетах разных размеров показаны на Рис. 8 и Рис. 9, соответственно. Большинство представленных на этих графиках кривых соответствует тенденции, отраженной на Рис. 7, однако есть несколько исключений. Так, гены, кодирующие транспортные белки, встречаются в 2-генных кассетах почти так же часто, как и в 3-генных, что можно объяснить широким распространением крупных белковых транспортных комплексов, таких как ABC-транспортеры, которые состоят не менее, чем из 3 субъединиц [91]. У Fusobacteria, Thermotogae и Firmicutes 5- и 6-генные кассеты встречаются практически не реже, чем 4-генные.

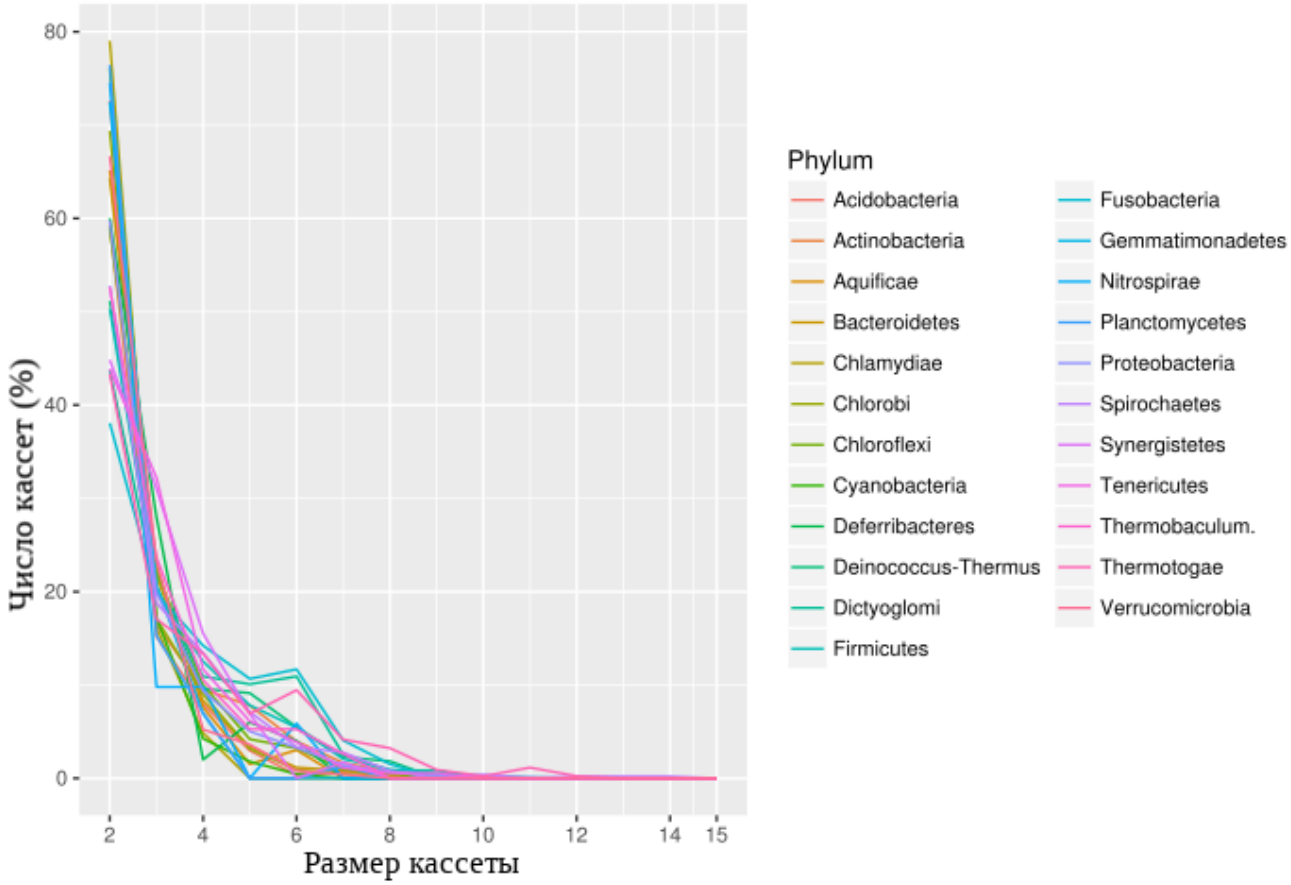

Рис. 8. Распределение кассет по размеру среди разных типов бактерий.

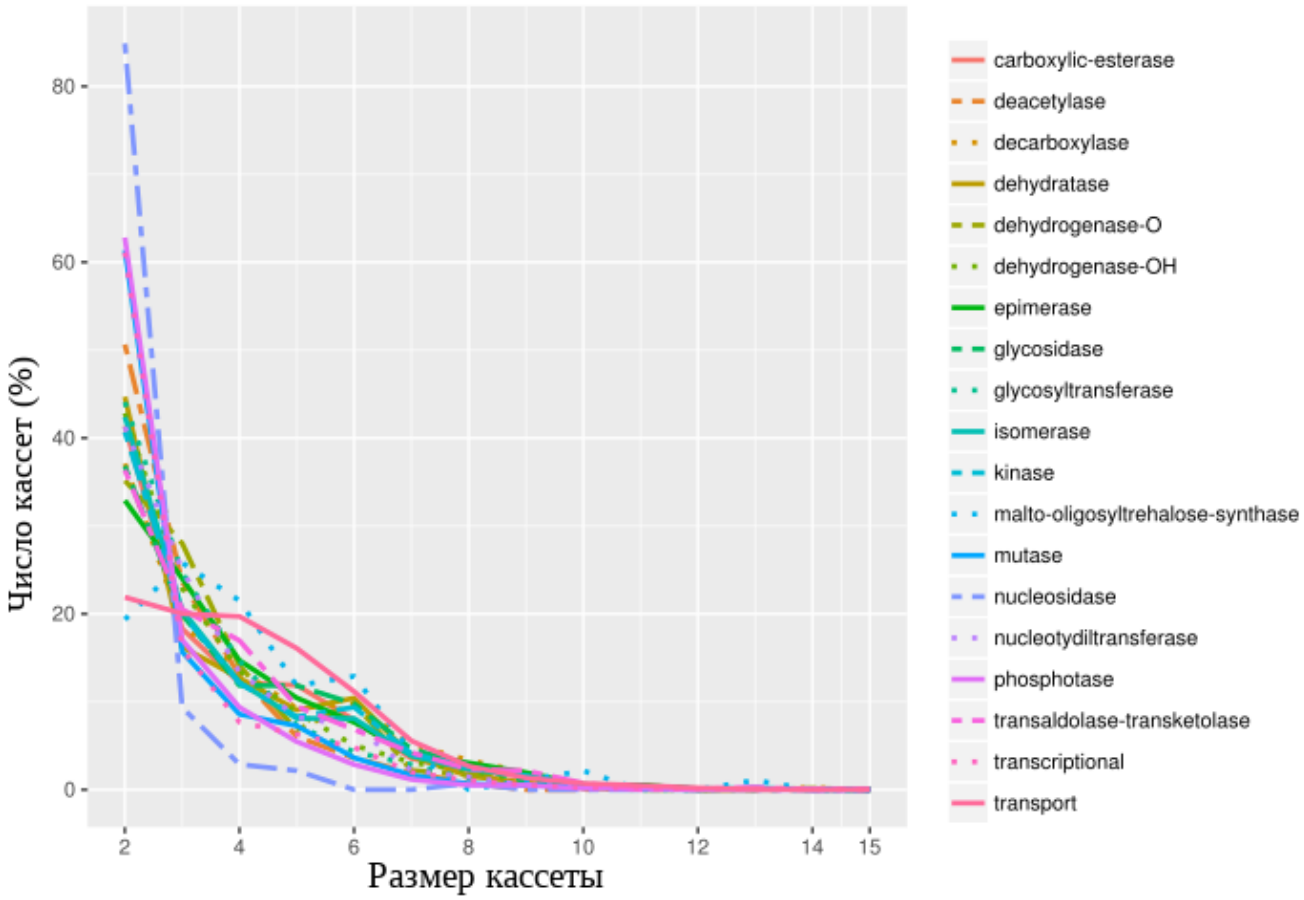

Рис. 9. Распределение по размеру кассет, содержащих гены разных функциональных классов.

Всего в кассетах встречалось около 10,4 тысяч разных комбинаций кластеров COG и около 2,5 тысяч разных комбинаций функциональных классов генов. По своему функциональному составу 45% кассет были уникальными, то есть встречались в исследуемых геномах только один раз.

Более того, только 43% всех исследованных нами генов входили в состав консервативных по составу кластеров COG кассет (кассета считалась консервативной, если встречалась в исследованных геномах по крайней мере дважды), тогда как в упомянутых выше исследованиях для всех бактериальных

белок-кодирующих генов эта доля составляла 69%. Такие наблюдения подтверждают гипотезу о том, что значительная часть прокариотических генов не формирует эволюционно-устойчивых комбинаций на бактериальных хромосомах, причем оказывается, что внутри сегмента углеводного метаболизма их доля еще больше.

Последний эффект можно объяснить возможным формированием эволюционно-устойчивых связей между аннотированными генами углеводного метаболизма и генами других функций. Последние могут кодировать ферменты, относящиеся к метаболизму нуклеотидов или иных соединений, содержащих углеводные остатки, например, гликолипидов или гликопротеинов; при этом они не взаимодействуют напрямую с углеводными остатками. В их аннотации углеводный метаболизм чаще всего не фигурирует. В нашей выборке присутствовали гены, кодирующие нуклеотидилтрансферазы, которые катализируют, в том числе, реакции переноса и присоединения углеводных остатков к нуклеотидам, и нуклеозидазы, катализирующие их отщепление. Закодированные в соседних с ними генах ферменты могут входить в состав одних с ними метаболических путей, но аннотация относит их к другим сегментам метаболизма, и в нашей выборке их нет; поэтому мы не наблюдаем соответствующих кассет.

3.2.2. Склонность генов разных функциональных классов и кластеров COG к формированию кассет

Долю генов, входящую в состав кассет, мы будем дальше называть склонностью к образованию кассет для данной группы генов. Функциональные классы значительно различались по этому параметру – он варьировал от 23% до 93% (см. Таблицу 2). Наименьшей склонностью к образованию кассет обладали нуклеозидазы, фосфатазы и мутазы (она составляла для них 23%, 38% и 42%, соответственно). Это можно объяснить, как уже было сказано выше, участием

продуктов таких генов в других типах метаболических путей, традиционно не относящихся к углеводному метаболизму. В ходе работы нуклеозидаз нуклеотиды подвергаются гидролизу с получением моносахаридов, поэтому нуклеозидазы имеют отношение как к углеводному метаболизму, так и к путям катаболизма и синтеза нуклеотидов, и, возможно, формируют устойчивые эволюционные комбинации только с последними.

Наибольшая склонность к образованию кассет – 93%, наблюдалась у небольшого класса мальтоолигозилтрегалозсинтаз, на втором месте оказались трансальдолазы и транскетолазы, а на третьем – транспортеры. Последнее, как уже обсуждалось выше, связано с тем, многие транспортные комплексы (такие как системы ABC и PTS) в бактериальной клетке состоят из нескольких субъединиц, гены которых часто закодированы рядом в составе единых оперонов [91,92].

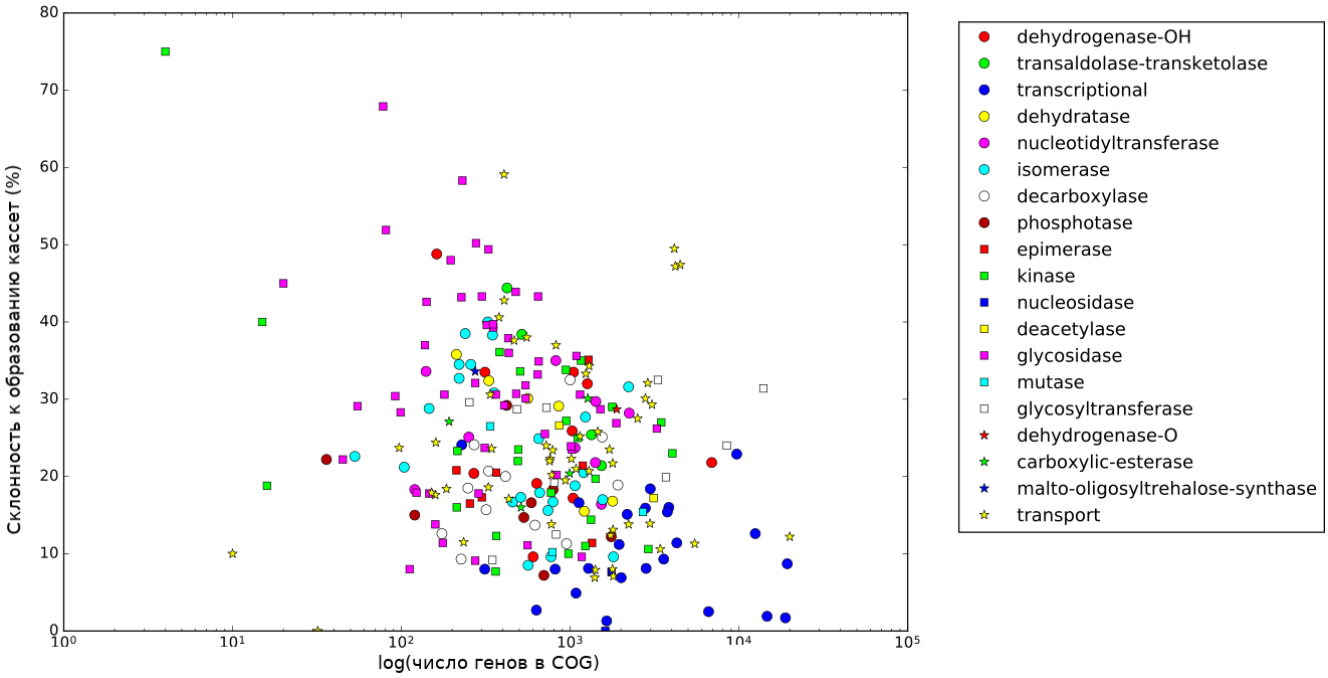

Рис. 10. Склонность к образованию кассет (по вертикали) у разных кластеров COG (размер кластера, т.е. число генов в COG, указан по горизонтали, логарифмическая шкала). Форма и цвет значка каждого кластера указывают на функциональный класс, к которому он принадлежит (расшифровано справа).

Склонность к образованию кассет у разных кластеров COG различалась еще сильнее, чем у функциональных классов, варьируя между 0% и 100% (см. Приложение Б, Рис. 10 и Рис. 11). Большинство крупных кластеров, содержащих более 4 тысяч генов, имели большую долю генов без соседей, относящихся к углеводному метаболизму, и склонность к образованию кассет для таких кластеров, в том числе для вторичных транспортеров суперсемейства MFS и многих транскрипционных регуляторов, составляла менее 40%. Исключением оказался большой кластер гликозилтрансфераз COG0438, включающий 6,5 тысяч генов, кодирующих белки, участвующие в синтезе клеточной стенки, склонность к образованию кассет которого оказалась весьма значительной и составила 66%.

Склонность к образованию кассет у некоторых кластеров среднего размера, включающих от двух до четырех тысяч генов, составляла более 90% (здесь представлены, в том числе, транспортеры систем ABC).

Самые маленькие кластеры, включающих менее двух тысяч генов, с наиболее высокой склонностью к образованию кассет, принадлежали к классам дегидрогеназ, изомераз, киназ, эпимераз и трансальдолаз/транскетолаз.

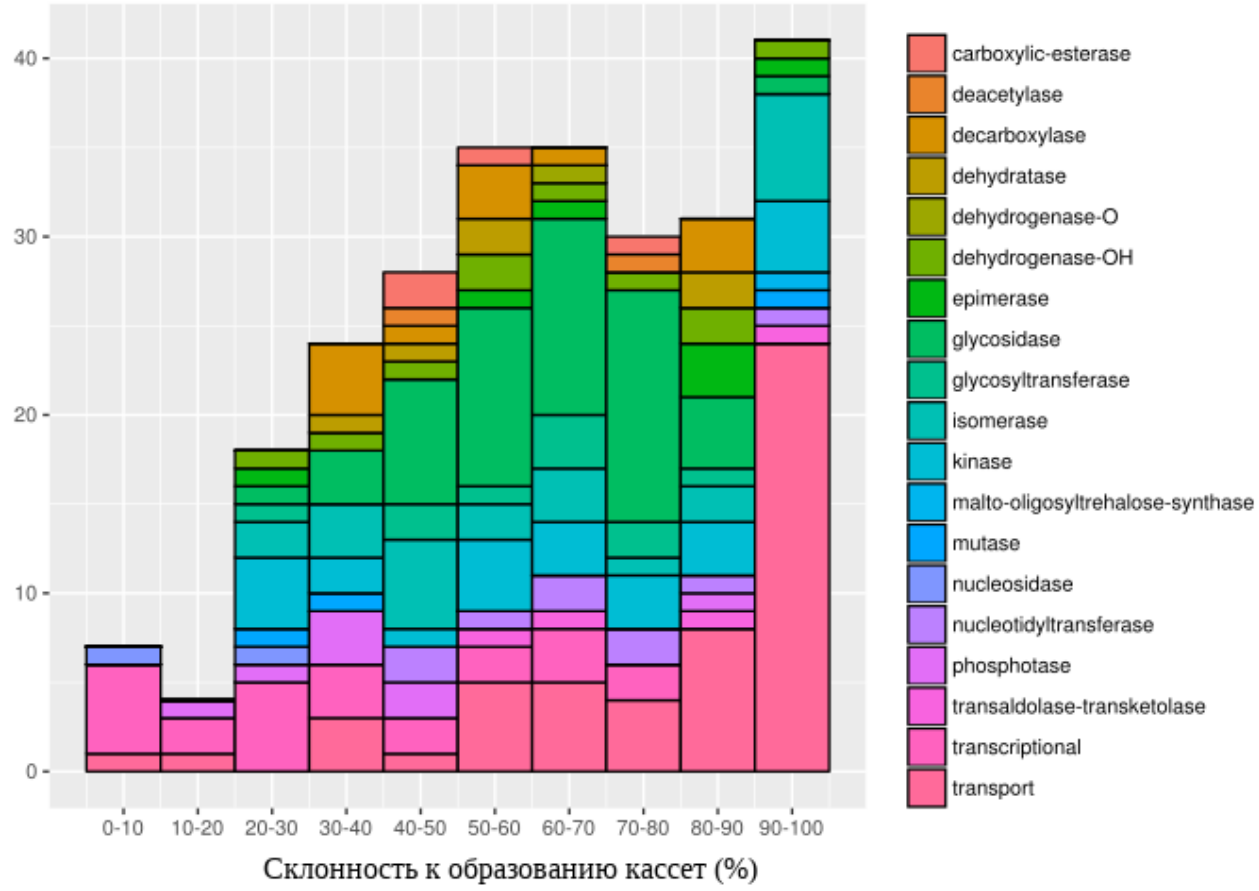

Рис. 11. Распределение склонности к образованию кассет у кластеров COG; цветом отмечено, к каким функциональным классам принадлежат кластеры.

3.2.3. Склонность генов разных бактериальных классов к формированию кассет

Филогенетические факторы также играли важную роль в склонности генов к формированию кассет в геномах. Для разных бактериальных классов эта склонность варьировала между 37% и 76% (Рис. 12). В данном случае для анализа мы выбрали классы, в которых было представлено не менее двух геномов нашей выборки, в каждом из которых было не меньше ста аннотированных генов углеводного метаболизма.

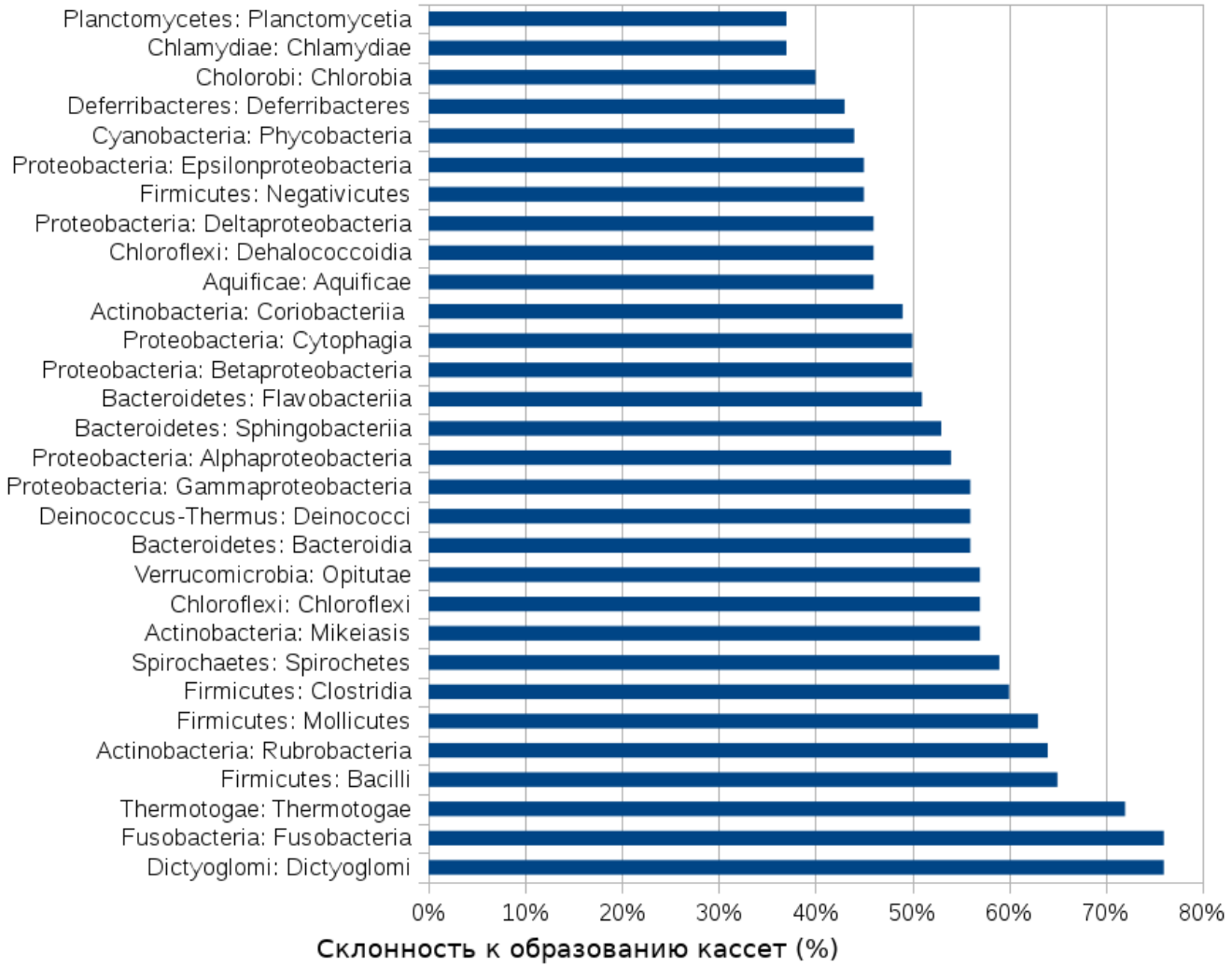

Рис. 12. Склонность к образованию кассет у генов, принадлежащих геномам бактерий разных классов. По вертикали указаны тип и класс бактерий. По горизонтали - склонность к образованию кассет в процентах.

Наибольшей склонностью к образованию кассет обладали представители классов Dictyoglomi и Fusobacteria (76%), Thermotogae (72%) и Bacilli (65%). Это соответствует опубликованным данным о том, что гены представителей этих классов (например, рода *Streptococcus*) часто лежат в составе длинных оперонов [44,45]. Наименьшей склонностью к образованию кассет обладали представители

классов Planctomycetia (37%), Chlamydiae (37%), Chlorobia (40%), Deferribacteres (42%) и Cyanobacteria (43%).

Среди крупных классов, в каждом из которых было аннотировано не менее восьми тысяч генов углеводного метаболизма, представители класса Deltaproteobacteria обладали наименьшей склонностью к образованию кассет (46%), средней склонностью (50%) обладали Betaproteobacteria, тогда как у Alphaproteobacteria, Gammaproteobacteria и Actinobacteria она была несколько выше, и составляла 54%, 56% и 57%, соответственно, а наибольшая склонность к образованию кассет оказалась у классов Clostridia (60%) и Bacilli (64%).

3.2.4. Функциональный состав кассет генов углеводного метаболизма

Наиболее распространенным участником кассет среди функциональных классов оказались гены, кодирующие транспортеры, гликозидазы и гликозилтрансферазы. Самая длинная кассета, обнаруженная в геноме *Stackebrandtia nassauensis* DSM 44728, включала 15 генов, среди которых было 11 транспортеров, 2 изомеразы, одна гликозидаза и одна гликозилтрансфераза.

Транспортеры встречались в 18% кассет, причем 10% кассет содержали не меньше двух транспортеров. Гликозидазы встречались в 19% кассет, причем 5,8% кассет имели не меньше двух гликозидаз, а 1,7% кассет имели не меньше трех. В геномах *Prevotella ruminicola* 23 и *Bifidobacterium dentium* Bd1 обнаружились кассеты с рекордным количеством гликозидаз – по семь представителей класса в каждой. Гликозилтрансферазы также встречались в 19% кассет, причем в 9,4% кассет они встречались не менее двух раз, а в 3,3% кассет – не менее трех. Наибольшее число гликозилтрансфераз обнаружилось в кассетах *Pedobacter saltans* DSM 12145 и *Bacillus weihenstephanensis* KBAB4 и составило девять генов на кассету.

Ни один из функциональных классов не оказался представлен одновременно более, чем в пятой части изученных кассет, что подчеркивает существенное разнообразие ко-локализационных тенденций генов, относящихся к углеводному метаболизму бактерий.

3.2.5. Попарные ко-локализационные тенденции представителей разных функциональных классов

Для того, чтобы оценить значимость событий попарной ко-локализации генов разных функциональных классов, мы сравнивали соответствующие события ко-локализации в кассетах с событиями случайной модели так, как описано в разделе Методы. Полученные данные представлены в Таблице 3. Ожидалось, что разнообразие эволюционно значимых связей между классами будет достаточно высоким, поскольку оно могло бы соответствовать значительному разнообразию комбинаций функций в распространенных метаболических путях. Однако из 190 возможных пар функциональных классов только у 45 (24%) число событий ко-локализации оказалось значительно выше, чем в случайной модели (P-значение их составляло менее 0,0001).

Функциональный класс		События ко-локализации	События ко-локализации сл. модели
transport	transport	5542	4196,87
glycosyltransferase	glycosyltransferase	2458	821,40
glycosidase	glycosidase	1533	813,05
isomerase	kinase	921	594,38
dehydrogenase-OH	glycosyltransferase	809	558,85
kinase	kinase	668	466,42
decarboxylase	kinase	653	265,67
nucleotydiltransferase	glycosyltransferase	508	230,81
nucleotydiltransferase	dehydrogenase-OH	416	95,08
isomerase	isomerase	402	224,11
dehydrogenase-OH	dehydrogenase-OH	371	139,10
dehydrogenase-O	kinase	354	66,86
dehydrogenase-OH	epimerase	334	82,18
isomerase	decarboxylase	333	190,63
nucleotydiltransferase	epimerase	331	35,44
dehydratase	dehydrogenase-OH	304	101,85
transaldolase-transketolase	transaldolase-transketolase	288	12,52
glycosyltransferase	deacetylase	244	155,98
carboxylic-esterase	dehydrogenase-OH	239	60,85
epimerase	kinase	229	162,00
dehydratase	isomerase	198	140,25
isomerase	dehydrogenase-O	195	49,77
carboxylic-esterase	kinase	175	104,64
isomerase	mutase	164	101,42
dehydratase	decarboxylase	162	62,68
isomerase	transaldolase-transketolase	146	103,53
decarboxylase	transaldolase-transketolase	131	47,84
dehydrogenase-O	transaldolase-transketolase	127	12,81
dehydrogenase-OH	transaldolase-transketolase	121	73,01
decarboxylase	decarboxylase	116	45,05
decarboxylase	epimerase	107	52,45
nucleotydiltransferase	nucleotydiltransferase	101	18,77
malto-oligosyltrehalose-synthase	glycosidase	91	10,40
nucleotydiltransferase	mutase	90	32,98
decarboxylase	dehydrogenase-O	89	22,05
carboxylic-esterase	deacetylase	87	16,90
carboxylic-esterase	dehydratase	83	24,88
carboxylic-esterase	decarboxylase	72	33,86
dehydratase	dehydrogenase-O	71	17,32
carboxylic-esterase	transaldolase-transketolase	66	16,46
dehydratase	dehydratase	62	21,56
epimerase	dehydrogenase-O	35	13,86
deacetylase	deacetylase	35	13,43
malto-oligosyltrehalose-synthase	mutase	18	1,73
dehydrogenase-O	dehydrogenase-O	16	1,91

Таблица 3. События попарной ко-локализации представителей разных функциональных классов и средние значения такой ко-локализации в случайной модели. В данной таблице приведены пары классов с P-значением не ниже 0.00001 (см. Методы)

Количество связей варьировало для каждого класса от 0 до 8 (Рис. 13). Размер класса, то есть число входящих в него генов, напрямую не влиял на это значение. Так, несмотря на большие размеры класса транспортеров, включающего более 21 тысячи генов в составе кассет, он не имел ни одной значимой связи с другими

классами. Класс трансальдолаз/транскетолаз, включающий около тысячи генов, входящих в кассеты, продемонстрировал шесть значимых связей с другими классами, тогда как сходный по размеру класс деацетилаз обладал всего тремя. Склонность к формированию кассет представителей класса, как таковая, по-видимому, также не влияла напрямую на число ко-локализационных связей этого класса с другими. Класс декарбоксилаз, склонность к формированию кассет которого составляла 60%, участвовал в восьми связях, а класс гликозилтрансфераз с аналогичной склонностью участвовал всего в четырех.

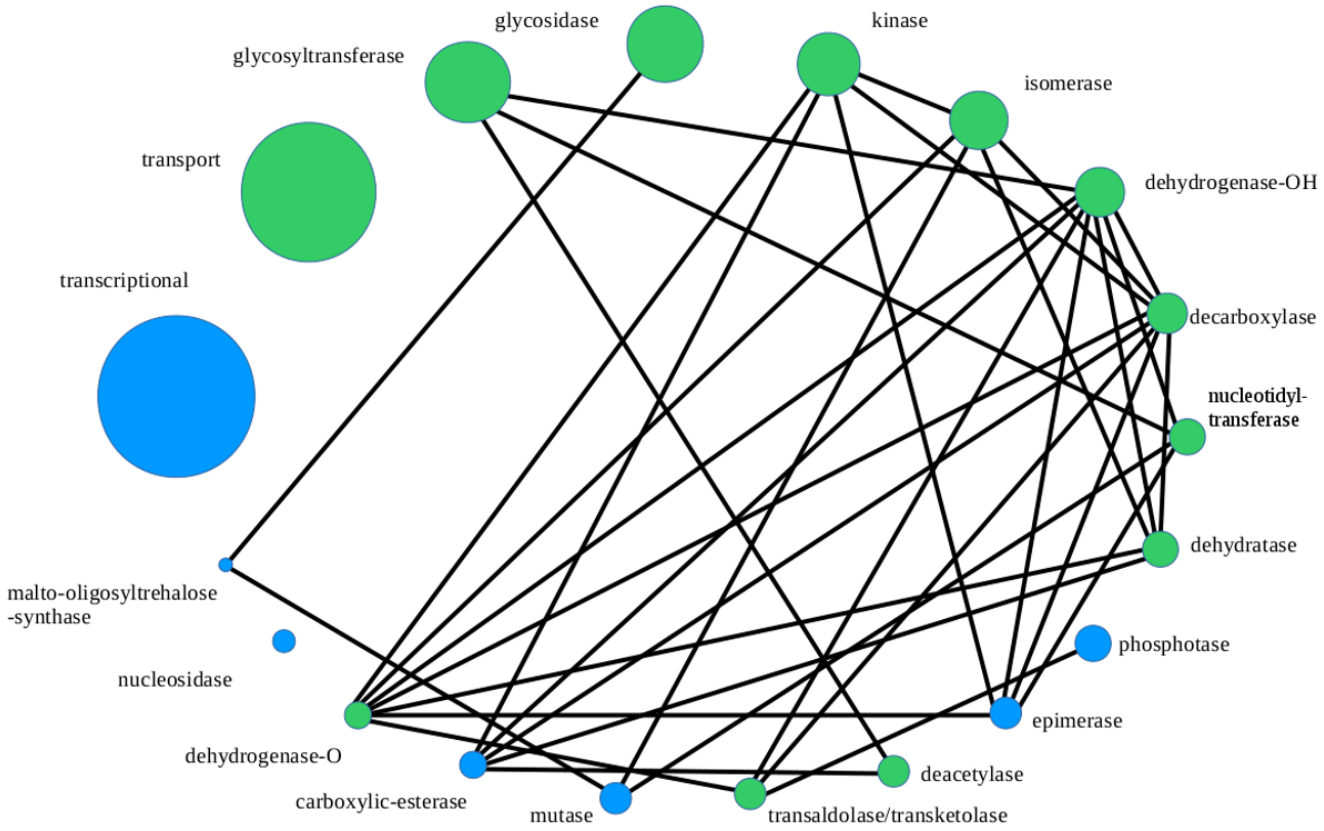

Рис. 13. Ко-локализационные связи между функциональными классами генов углеводного метаболизма. Кругами представлены разные классы, размер круга соответствует относительному размеру класса. Линиями соединены классы, имеющие значимую ко-локализационную связь. Зеленым цветом отмечены классы, представители которых имеют тенденцию к ко-локализации друг с другом.

Значительная часть связей была сформирована классами декарбоксилаз, дегидрогеназ-ОН и дегидрогеназ-О (у них оказалось 8, 8 и 7 связей, соответственно). Таким образом, именно эти классы обладали наиболее разнообразными и при этом неслучайными предпочтениями по отношению к своему геномному окружению.

Большая часть связей была образована парами функций, встречающихся в распространенных и хорошо изученных метаболических путях. Так, например, изомераза и киназа одновременно присутствуют в путях, связанных с деградацией лактозы, галактозы, хитина и арабинозы. Декарбоксилаза и киназа присутствуют во всех вариациях пути Энтнера-Дудорова (путь катаболизма глюкозы, отличный от гликолиза и пентозофосфатного пути) [93]. Эпимераза и мутаза встречаются в путях гликолиза [94] и глюконеогенеза [95], а также, например, в пути деградации маннана [96]. Дегидрогеназа и карбоксил-эстераза участвуют в путях деградации галактозы [97,98].

Это наблюдение соответствует представлениям о том, что белки, участвующие в одном и том же метаболическом пути, часто закодированы в ко-локализованных генах или даже расположены в составе единого оперона [27,30,66,99]. Однако ко-локализационные события многих пар функций, присутствующих в известных метаболических путях, в рамках данного анализа не преодолели порога значимости – например, гликозидаза и киназа, совместно участвующие в гликолизе и других метаболических путях. Гены гликозидаз и киназ недостаточно часто встречались в кассетах вместе, чтобы эти события можно было отличить от случайных событий ко-локализации. Такой результат подтверждает гипотезу о том, что ко-локализация не является обязательным условием для генов, кодирующих белки с взаимосвязанными функциями.

3.2.6. Попарные ко-локализационные тенденции кластеров COG

Попарные комбинации функций были сформированы, в свою очередь, попарными комбинациями представителей разных кластеров COG. Чтобы изучить соответствующие ко-локализационные тенденции мы выявили кластеры, представители которых наиболее часто встречались рядом на бактериальных геномах. В Приложении Б, содержащем список исследованных в данной работе кластеров COG, в качестве примера для каждого кластера в седьмой колонке таблицы мы привели список наиболее распространенных его соседей из трех наиболее часто встречающихся кассет.

После этого мы выбрали из всех наиболее часто встречающихся пар значимые с помощью критерия хи-квадрат так, как это описано в Методах. В каждой паре классов такой порог преодолевало от 0 до 2 пар кластеров COG (Приложение В, преодолевшие порог кластеры отмечены жирным шрифтом). Такой критерий позволял отличать события неслучайной ко-локализации кластеров от событий, зависящих только от размера кластера.

Функциональные классы образовывали также и тройственные связи. Для некоторых из таких случаев, например, для киназы, изомеразы и дегидрогеназы-О, три наиболее часто встречающиеся попарные комбинации кластеров COG были представлены тремя кластерами (дегидрогеназа – COG0057, изомераза – COG0149, киназа – COG0126). Таким образом, все три эти кластера обладали выраженным предпочтением к геномной локализации друг с другом, что указывало на возможную их эволюционную связь. Такая картина наблюдалась, однако, не для все тройственных связей. Три наиболее часто и неслучайно встречающиеся пары для классов нуклеотидилтрансфераз, гликозилтрансфераз и

дегидрогеназ-ОН оказались представлены шестью разными кластерами COG, то есть пересечений в этих парах не было. Все перекрестные комбинации, с повторяющимися участниками, встречались на два порядка реже. (см. Приложение В).

Анализ подобных эволюционных связей между функциональными классами или между кластерами COG может предоставлять новые данные для более точной аннотации их представителей, поскольку ко-локализационные предпочтения групп генов часто отражают роль соответствующих белков в метаболических путях бактерий [29,38,39].

Так, наиболее часто встречающаяся пара нуклеотидилтрансфераз и гликозилтрансфераз оказалась представлена кластерами COG0448, аннотированным как гены глюкозо-1-фосфат аденидилтрансферазы (EC 2.7.7.27) и COG0297, аннотированным как гены гликогенсинтазы (EC:2.4.1.21), взаимодействующей с АДФ-глюкозой; два соответствующих фермента задействованы, например, в последовательных этапах путей метаболизма крахмала. Самая распространенная пара кластеров нуклеотидилтрансфераз и дегидрогеназ-ОН была представлена COG1091, аннотированным как дТДФ-4-дегидрорамнозоредуктаза (EC 1.1.1.133) и COG1209, аннотированным как глюкозо-1-фосфаттимидилтрансфераза (EC 2.7.7.24) – два соответствующих фермента являются, в частности, участниками путей биосинтеза дТДФ-6-деоксигексоз. Связь между участниками наиболее часто встречающейся пары кластеров из классов гликозилтрансфераз и дегидрогеназ-ОН (COG0451 и COG0438) оказалась не так очевидна. Оба кластера содержали большое число генов – более шести тысяч и более тринадцати тысяч, соответственно. Согласно представленным в базе данных IMG аннотациям, эти гены кодируют белки с множеством разных предсказанных функций, в том числе, например, дТДФ-

дегидрорамнозоредуктазу для некоторых представителей COG0451 и гликогенсинтазу для некоторых представителей COG0438. Гены этих кластеров оказались ко-локализованы в различных бактериальных геномах более тысячи раз. Такое выраженное предпочтение к ко-локализации, вероятно, указывает на значимую функциональную и эволюционную связь между ними, и является поводом к дальнейшему исследованию функциональных особенностей соответствующих белков и биологической роли их взаимосвязи и, соответственно, к уточнению их аннотации.

3.2.7. Попарные ко-локализационные тенденции представителей одних и тех же функциональных классов

Из 45 выявленных ко-локализационных связей 12 были сформированы благодаря ко-локализации представителей одного и того же класса. Это означает, что в составе общих кассет часто присутствовали два или несколько генов, принадлежащих к одному и тому же функциональному классу, и такие события оказались неслучайны. Из 19 изученных функциональных классов, таким образом, почти две трети продемонстрировали склонность к подобной ко-локализации (Таблица 3).

Больше всего ко-локализованных генов одного класса оказались среди транспортеров, гликозидаз, транскетолаз/трансальдолаз и гликозилтрансфераз. Стоит отметить, что класс гликозилтрансфераз и класс трансальдолаз/транскетолаз были представлены в кассетах несколькими генами чаще, чем одним.

Гены одного и того же класса, ко-локализованные в кассетах, делились на две группы – гены, кодирующие разные субъединицы белковых комплексов, и гены, кодирующие отдельные белки. Наиболее распространенным примером участников первой группы являлись гены, кодирующие субъединицы транспортных

комплексов. Более того, выяснилось, что устойчивые ко-локализационные связи формируют гены транспортеров, лежащие в кассетах не менее, чем по три гена. События ко-локализации не более двух генов транспортеров при этом не прошли порога отличия от случайной ко-локализации, описанного в Методах. Это явление, вероятнее всего, объясняется мультидоменной структурой транспортных комплексов, таких, как ABC-системы, которым требуется, по меньшей мере, три гена, кодирующих три основные ее субъединицы [91].

Остальные ко-локализованные гены одного класса чаще всего кодировали самостоятельные белки, не являющиеся субъединицами белковых комплексов; в части случаев они оказывались участниками последовательных этапов метаболических путей.

Так, например, известно, что несколько гликозидаз могут участвовать в последовательных этапах деградации сложных полисахаридов. Такие гликозидазы могут быть закодированы в составе одного оперона или близлежащих оперонов. Например, утилизация ламинарина в *Gramella forsetii* осуществляется с помощью оперона, содержащего три гена, кодирующих гликозидазы. Утилизация альфа-1,4-гликанов тоже осуществляется в этой бактерии с помощью белков, закодированных в двух соседних оперонах с четырьмя генами гликозидаз [100].

Несколько гликозилтрансфераз могут участвовать в последовательных этапах путей биосинтеза клеточной стенки бактерии. Так, в геномах *Lactococcus lactis* и других лактобактерий в соответствующих оперонах встречаются иногда одновременно более семи гликозилтрансфераз, необходимых для этого процесса [101]. Трансальдолазы и транскетолазы – участники пентозофосфатного пути, и соответствующие гены могут также быть закодированы рядом, как, например, это происходит в случае генов *Escherichia coli talA* и *tktB* [45]. Две или три киназы также могут одновременно участвовать в последовательных этапах одного

метаболического пути – например, гликолиза или деградации лактозы [19]. Для других случаев, например, для декарбоксилаз, причины ко-локализации генов одного и того же функционального класса не столь очевидны.

3.2.8. Роль событий локальной дупликации и образования ксенологов и псевдопаралогов в ко-локализации генов сходных функций

Одним из известных механизмов, лежащих в основе ко-локализации генов сходных функций, является локальная дупликация генов [102]. Локальная дупликация – это удвоение определенного участка хромосомы, в результате которого рядом на хромосоме оказываются две изначально одинаковые или очень сходные нуклеотидные последовательности (это один из вариантов возникновения паралогов). Локальная дупликация может происходить за счет гомологической рекомбинации, например, при наличии повторов в нуклеотидных последовательностях и при нарушении работы топоизомераз. Гены, получившиеся в результате таких событий, впоследствии могут приобретать разные мутации, и их нуклеотидные последовательности со временем начинают различаться все сильнее. Нашей задачей было выяснить, как часто ко-локализованные гены, имеющие сходные функции, являются результатом событий локальной дупликации, поскольку в противном случае их ко-локализация не имеет очевидного объяснения и имеет смысл обсуждать более глубокие эволюционные или функциональные ее причины.

В 44% случаев ко-локализованные гены одного и того же функционального класса также принадлежали к одному и тому же кластеру COG, а следовательно, обладали определенным структурным сходством. Среди 264 кластеров COG нашей базы данных 189 кластеров были ко-локализованы в исследуемых геномах хотя бы однажды. Для того, чтобы оценить, как часто такие случаи были результатом ранее произошедших локальных дупликаций, мы использовали разработанный нами

инструмент NSimScan, алгоритм действия которого описан в Главе 2, с параметрами, указанными в Методах настоящей главы. Для каждого гена из ко-локализованной пары одного и того же кластера COG мы отобрали лучшие совпадения среди всех представителей данного кластера в исследуемой выборке генов, а также сравнили последовательности представителей пары друг с другом. Только в 3,6% случаев гены в паре продемонстрировали наибольшее сходство друг с другом (т.е. ни для одного из участников пары не нашлось более похожего на него кандидата среди остальной выборки). Во всех остальных случаях для одного или обоих участников пары среди других представителей COG отыскивалось более близкое совпадение, причем в 62% случаев соответствующий ген располагался не только в другой кассете, но и в другом геноме.

Известно, однако, что паралоги подвержены действию положительного естественного отбора в меньшей степени, чем ортологи [103], то есть в ходе эволюции два паралога внутри одного генома быстрее накапливают различия, чем два ортолога в разных геномах. Поэтому данные результаты не позволяют полностью отбраковывать случаи, когда два ко-локализованных гена оказываются более похожи на родственный ген в другом геноме, чем друг на друга, но при этом исходно являются результатом давнего события локальной дупликации. С другой стороны, такая ко-локализация может являться, например, результатом события горизонтального переноса (см. ниже).

Одним из альтернативных объяснений ко-локализации генов общего функционального класса и кластера COG могут быть события ксенологической замены генов [104]. В этом случае после события локальной дупликации происходит замена одного из пары генов на другой, похожий ген, путем горизонтального переноса из другого генома. В результате ко-локализованными оказываются гены, которые называют псевдоортологами или ксенологами. Новый

ген при этом больше похож по нуклеотидной последовательности на исходный ген из другого генома, чем на тот, который оказывается с ним рядом.

Другой причиной ко-локализации генов, принадлежащих к одному кластеру COG, может быть процесс, приводящий к возникновению генов, которые называются псевдопаралоги. При этом новый ген из другого генома переносится в локус рядом с гомологичным геном, не замещая никаких его исходных гомологичных соседей. Два оказавшихся в результате рядом гена будут сходны между собой, но на исходный ген из своего генома один из них будет похож больше.

В нашей выборке, однако, менее 10% пар ко-локализованных генов из одного COG оказывались более похожими на один и тот же ген в другом геноме, чем друг на друга; в большинстве случаев они были похожи на два разных гена. Таким образом, мы можем утверждать, что ко-локализация генов близких функций в преобладающем большинстве случаев, по-видимому, не является результатом события локальной дупликации.

3.2.9. Эволюционное значение попарной ко-локализации представителей одного функционального класса

Мы предполагаем, что гены сходных функций, расположенные на бактериальной хромосоме рядом друг с другом (особенно многочисленными группами, как это происходит, например, у гликозилтрансфераз и гликозидаз), могут применяться одновременно в ситуациях определенного типа. Такой набор может иметь общий механизм регуляции транскрипции, и в определенных условиях его гены могут экспрессироваться одновременно, например, когда клетке необходимо включение целого ряда ферментов, участвующих в деградации или биосинтезе углеводов. Это происходит, например, при утилизации или биосинтезе

сложных полисахаридов, где разные гликозилтрансферазы или гликозидазы задействованы в рамках общих или тесно переплетенных метаболических путей [19].

Кроме того, известно, что, оказавшись в неоптимальных для роста условиях среды, клетка может активировать экспрессию сразу целой группы генов. Так, виды рода *Bacillus* в условиях голодания одновременно активируют экспрессию множества генов, ответственных за катаболизм и транспорт альтернативных источников углерода [105]. Это касается, в частности, транспортеров и гидролаз, относящихся к углеводному метаболизму. Ко-локализация генов, активирующихся в подобных стрессовых условиях, может объясняться удобством одновременной регуляции их транскрипции. Кроме того, ко-локализованные гены будет чаще совместно передаваться в другие геномы при событиях горизонтального переноса, и соответствующие комбинации могут эволюционно закрепляться как в родственных, так и в других видах бактерий.

3.3. Заключение

Мы провели детальный анализ хромосомной организации локусов, относящихся к углеводному метаболизму бактерий и описали сложную сеть эволюционных связей соответствующих генов, выраженную в форме их ко-локализационных тенденций.

Из 148 тысяч проанализированных генов углеводного метаболизма в 665 бактериальных геномах только 53% оказались склонны к ко-локализации. Остальные не имели непосредственных соседей, относящихся к углеводному метаболизму. Кассеты, образованные ко-локализованными генами, варьировали по размеру и составу и включали от двух до пятнадцати генов. Большинство кассет были короткими, двух- и трех-генными. Двумя существенными факторами,

влияющих на склонность гена к формированию кассет, оказались функция гена и филогенетические характеристики вида.

В целом, мы получили полную картину тенденций к ко-локализации генов 19 основных функциональных классов, более двухсот кластеров ортологических групп генов и тридцати классов бактерий. Тенденция к формированию кассет составляла от 23 до 93% для генов разных функциональных классов, причем наибольшую склонность к ко-локализации с другими генами углеводного метаболизма продемонстрировали гены мальтоолигозилтрегалозсинтаз, трансальдолаз/транскетолаз и транспортеров. У разных бактериальных классов тенденция к формированию кассет варьировала от 40 до 76%, наибольшее число генов углеводного метаболизма находилось в составе кассет у классов Fusobacteria, Dictyoglomia и Thermotogae.

Анализ попарно встречающихся генов выявил наличие 45 эволюционно значимых связей между 19 функциональными классами генов. Количество таких связей для каждого класса варьировало от нуля до восьми, что указывает на существенную разницу в предпочтениях к хромосомному окружению у генов разных функций. Наибольшим числом связей обладали классы декарбоксилаз и дегидрогеназ – соответствующие гены продемонстрировали наиболее выраженные предпочтения по отношению к специфике своего окружения. Классы транспортеров и гликозидаз, несмотря на большой размер и участие во многих кассетах, не продемонстрировали значимых специфических предпочтений к ко-локализации с генами другим классов. При этом в случае этих двух и девяти других классов кассеты чаще ожидаемого содержали одновременно несколько генов одного и того же класса. Большинство таких случаев, по-видимому, не являлось результатом событий локальной дупликации.

Многие из подобных выявленных эволюционных связей объяснялись участием соответствующих генов в известных метаболических путях; другие оставляли простор для дальнейших исследований, в том числе экспериментальных, касающихся функций и взаимодействий соответствующих белков.

Мы предположили, что консервативные сочетания представителей определенных функциональных классов могут указывать на сходные функции соответствующих кассет. Для того, чтобы оценить силу подобных предсказаний, мы экспериментально проверили одно из них для случая, когда комбинация функций в консервативной кассете бактерий семейства Enterobacteriaceae, участвующей в сульфогликолизе, совпала с комбинацией консервативной кассеты класса Bacilli, участвующих в катаболизме лактозы (см. Главу 4).

Глава 4. Участие *yih*-кассеты *Escherichia coli* в катаболизме лактозы

4.1. Сравнительный анализ консервативных кассет и экспериментальная задача для проверки функционального предсказания

Мы предположили, что сравнительный анализ комбинаций полученных нами кассет может позволить предсказывать их функции. Как уже говорилось выше, в некоторых случаях геномное окружение может позволять успешно выявлять роль генов в тех или иных процессах [31,36–38]. Для составления предсказаний мы использовали наиболее распространенные комбинации функциональных классов внутри кассет, исходя из предположения, что консерватизм будет указывать на эволюционные преимущества такой ко-локализации.

Мы сравнивали между собой не только целые кассеты, но и все возможные сочетания внутри них. Это позволило нам выявить все наиболее часто встречающиеся комбинации функциональных классов или кластеров COG. Большинство таких консервативных сочетаний оказались, как и большинство самих кассет, короткими (двух- или трех-генными). Среди более длинных сочетаний функциональных классов в качестве кандидата для экспериментальной проверки мы выбрали комбинацию из шести функциональных классов – транспортера, регулятора транскрипции, гликозидазы, альдолазы, киназы и изомеразы. Она встречалась, в том числе, в кассетах не близкородственных бактерий, что дополнительно указывало на неслучайность подобной ко-локализации. Кассета с таким составом оказалась распространена как у гаммапротеобактерий семейства Enterobacteriaceae, среди которых наиболее известным представителем является кишечная палочка *Escherichia coli*, так и у бактерий класса Bacilli, среди родов *Streptococcus* и *Staphylococcus*.

В исследовании 2014 года было описано участие белков, закодированных в генах кассеты *Escherichia coli*, в метаболизме серосодержащих углеводных соединений [75]. У представителей класса Bacilli кассета с таким же набором функциональных классов кодирует белки, участвующие в катаболизме лактозы [45]. Мы предположили, что помимо редкой функции катаболизма серосодержащих углеводов и более сложных серосодержащих соединений (например, сульфогликолипидов), кассета кишечной палочки может также участвовать в утилизации лактозы, а ее гены могут, таким образом, кодировать мультифункциональные белки.

Предположение было подтверждено с помощью экспериментального исследования. Эта часть работы выполнялась в лаборатории функциональной геномики и клеточного стресса Института биофизики клетки РАН г. Пущино под руководством М.Н. Тутукиной. Выяснилось, что экспрессия генов, кодирующих в кишечной палочке альдолазу (*yihT*), изомеразу (*yihS*) и киназу (*yihV*) значительно повышалась во время роста клеток на лактозе. После этого были идентифицированы точки начала (старты) транскрипции *in silico*, *in vitro* and *in vivo* и показано, что из трех промоторов гена альдолазы один активировался именно при росте клеток на лактозе.

Кроме этого мы проанализировали механизм переключения регуляции транскрипции данной кассеты. Было показано, что в этом процессе участвует локальный регулятор YihW, принадлежащего к семейству регуляторов DeoR. Он служит двойным переключателем, механизм действия которого зависит от доступного клетке источника углерода. В частности, оказалось, что он поддерживает рост клеток кишечной палочки на лактозе при выключенном *lac*-опероне, то есть при инактивации известных и хорошо изученных генов, участвующих в утилизации лактозы. Выяснилось также, что YihW действует в

паре с глобальным регулятором углеводного метаболизма cAMP-CRP, и в зависимости от условий среды, они могут оказывать либо комплементарное, либо противоположное воздействие на экспрессию генов кассеты.

4.2. Методы

4.2.1. Штаммы, плазмиды и выращивание культур

Все штаммы и плазмиды, которые использовались в данной работе, перечислены в Приложении Д. Гены *yihW* и *crp* были удалены с помощью лямбда-ред рекомбиназы в штамме *E. coli* BW25113 [106]. Затем мутации были перенесены в штаммы *E. coli* K-12 MG1655 и *E. coli* K-12 M182 с помощью P1-трансдукции [107]. Штамм M182 получен из штамма K-12 MG1655 путем выключения *lac*-оперона [108].

Клетки растили на минимальной среде Minimal Salts (MS) с 5% или 10% LB (объемная доля, v/v) и 0,2% (доля массы к объему, w/v) исследуемого источника углерода – D-глюкозы, D-галактозы, лактозы или глицерина. Культуры растили в аэробных условиях при 37 °C и постоянном перемешивании. Клетки собирали спустя 4,5 часа роста (средняя экспоненциальная фаза роста, с оптической плотностью OD = ~0.2–0.4, варьирующей у разных штаммов при росте на разных источниках углерода).

Для получения клеток с суперпродукцией (существенным повышением экспрессии) белка cAMP-CRP клетки *E. coli* BL21* (DE3) трансформировали плазмидой pET_CRP, которая была разработана на основе плазмиды pET28b (Invitrogen) со вставкой гена *crp* между сайтами NdeI и Bpu1102. Клетки растили на средах LB или Terrific broth (TB) при 37 °C до достижения оптической плотности OD_{650} = 0,3 и индуцировали синтез целевого белка путем добавления IPTG (изопропилтиогалактозида) очень низкой концентрации (20 мкм), чтобы

избежать токсических эффектов. Спустя 3 часа после начала индукции содержание CRP в клетке составляло ~70% от общего количества белка, а спустя 16 часов - ~80% (дорожки 1 и 4 на Рис. 14)

4.2.2. Выделение белка cAMP-CRP

Чтобы выделить CRP, клетки BL21*(DE3) с pET-CRP растили в 200 мл среды TB до OD_{650}=0.3, индуцировали 20 мкм IPTG и затем собирали с помощью центрифугирования на 10,000 об/мин при +4 °C. Клетки отмывали от среды 1xPBS и лизировали с помощью 2 мл реактива для экстрагирования белков BugBuster (Novagen) в течение 20 минут при +4 °C. Ввиду крайне высокого уровня индукции, все белки оказывались в нерастворимой фракции даже при очень низкой концентрации IPTG (Рис. 14, дорожка 3). Чтобы выделить CRP из телец включения, выполнялась двухступенчатая процедура лизирования: после первичной обработки реактивом BugBuster, к белку добавляли 6 мл ледяного лизирующего буфера Lysis [59,63] (50 мМ K-фосфатны буфер pH 7,5; 2 мМ ЭДТА; 0,2 М NaCl; 5% (w/v) глицерин; 2 мМ ДТТ; 50 мг/мл ФМСФ), после чего клетки подвергали дополнительному разрушению ультразвуком на льду 3 раза по 30 секунд. Затем лизат очищали от остатков клеток путем центрифугирования в течение 20 минут при 15,000 об/мин и +4 °C и наносили на колонку с цАМФ-агарозой (Sigma). Белок элюировали с помощью 5 мМ цАМФ. В результате из 200 мл исходной культуры было получено ~20 мг очищенного CRP. Все шаги выделения CRP показаны на Рис. 14.

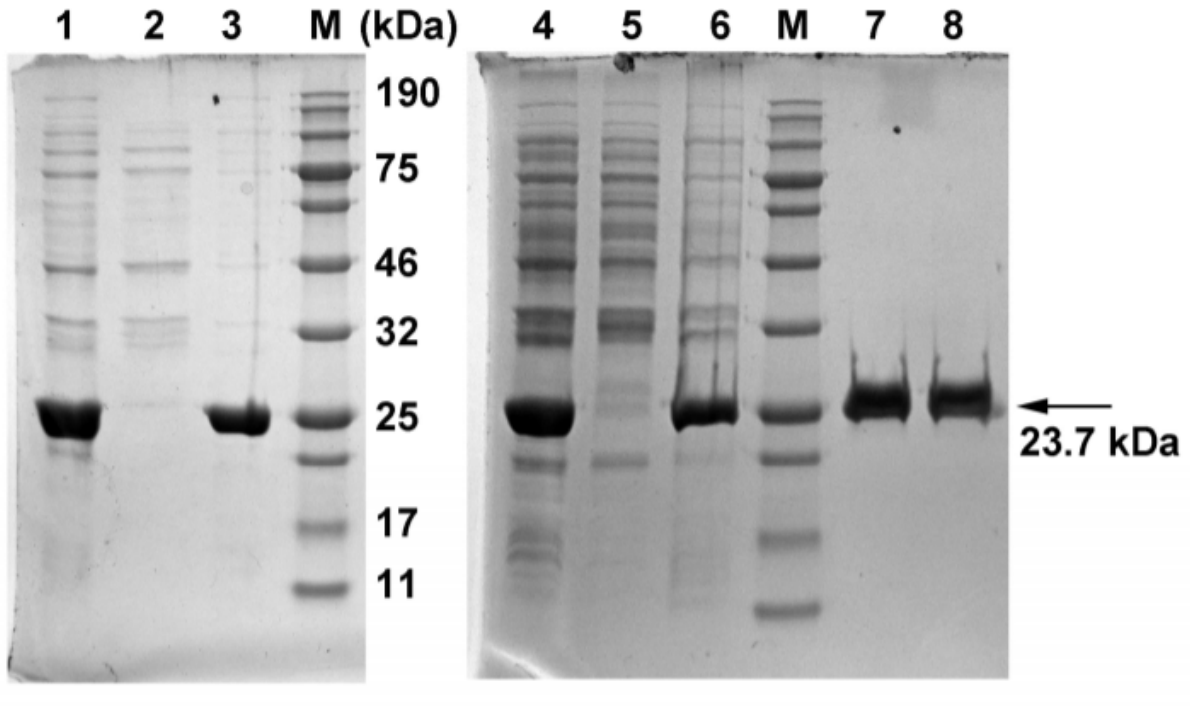

Рис. 14. Этапы индукции и очистки фактора CRP. Дорожки 1 и 4 – тотальный лизат спустя 16-часовой индукции с 20мкМ IPTG, 2 – растворимая фракция, 3 – нерастворимая фракция, 5 – раствор после пропускания лизата через колонку с цАМФ-содержащей агарозой, 6 – очищенный лизат после обработки реактивом BugBuster и разрушения ультразвуком, 7 и 8 – очищенный CRP, элюированный 5 мМ цАМФ. В качестве маркеров молекулярного веса были использованы предокрашенные маркеры 11-190 кДа, New England Biolabs).

4.2.3. Картирование промоторов

Для картирования промоторов и сайтов связывания факторов транскрипции были выбраны четыре межгенных участка генов *yih*-кассеты (*yihR/S*, *yihT/U*, *yihU/V* и *yihV/W*), принадлежащих *Escherichia coli*, *Enterobacter cloacae*, *Salmonella enterica*, *Cronobacter turicensis* и *Pantoea anantis*. Их длины составляли около 40-50 нуклеотидов. Пользуясь базой данных GenBank [82], мы получили их последовательности и добавили к каждой из них 100 нуклеотидов с обоих концов, захватывающие участки аннотированных открытых рамок считывания, чтобы учесть возможные ошибки этой аннотации. Последовательности затем были подвергнуты процедуре множественного выравнивания с помощью инструмента

T-Coffee [109]. Потенциальные промоторы для *E. coli* K-12 MG1655 (геном в базе данных GenBank U00096.2) были предсказаны с помощью инструмента PlatProm и его унифицированной версии PlatPromU, в которой не учитывалось влияние сигма-фактора на узнавание РНК-полимеразой исследуемой промоторной последовательности [110]. Связывание σ^{70}-РНК-полимеразы с исследуемыми областями было подтверждено с помощью электрофореза с задержкой в геле, а точки старта транскрипции были определены с помощью однократной инициации транскрипции *in vitro* и измерения длин продуктов Primer extension *in vivo*. Для Primer extension клетки выращивали в среде MS с добавлением 10% LB и 0,2% глюкозы, лактозы или глицерина и собирали их при оптической плотности OD_{650} = 0,4. РНК выделяли из 10 мл культуры клеток с помощью реактива TRIzol (Invitrogen, США) и обрабатывали ДНКазой I (New England Biolabs, USA) по стандартному протоколу. 10 мг тотальной РНК инкубировали с 4 пкмоль [γ-^{32}P]АТФ-мечеными праймерами (Приложение Г) и обратной транскриптазой SuperScript II (Invitrogen, США). Полученные образцы центрифугировали, растворяли в 7 мкл формамидного буфера (98% формамид, 8 мМ NaOH, 4 мМ EDTA), прогревали 5 минут при 90°C, охлаждали до 0°C и наносили на 6% денатурирующий полиакриламидный гель. В качестве контроля использовали [γ-^{32}P]АТФ-меченые маркеры ДНК 50 п.н. (New England Biolabs, США). Затем гели визуализировали на Molecular Imager (Bio-Rad, США).

4.2.4. Поиск сайтов связывания факторов транскрипции

Потенциальные сайты связывания транскрипции были выявлены в межгенных участках с помощью метода филогенетического футпринта и базы данных Virtual Footprint [111]. Сайты искали в области -250/+50 от аннотированных точек начала транскрипции. Полученные сайты сортировали по их показателю качества (score) и степени консервативности среди видов семейства Enterobacteriaceae. Для

дальнейшего анализа отбирали самые консервативные сайты с лучшими показателями качества.

4.2.5. Электрофорез с задержкой в геле

Для оценки эффективности связывания РНК-полимеразы с регуляторными областями генов *yih*-кассеты мы использовали метод электрофореза с задержкой в геле (EMSA). Фрагменты ДНК, содержащие потенциальные промоторные последовательности *yih*-генов, амплифицировали с помощью ПЦР (метода полимеразной цепной реакции). Для этого для каждого гена мы подобрали последовательности прямых и обратных праймеров (отмечены в Приложении Г как R и F, соответственно). 1 пмоль каждого фрагмента ДНК, содержащего потенциальные промоторы, инкубировали при 37 °C в транскрипционном буфере (Transcription buffer) [112] в течение 10 минут, после чего к нему добавляли очищенную σ^{70}-РНК-полимеразу в разных соотношениях с ДНК, от 1:1 до 4:1 (М:М). Спустя еще 30 минут инкубации при 37 °C полученные комплексы загружали в 5% полиакриламидный гель, предварительно разогретый до 37 °C, и подвергали электрофорезу в буфере 1xTBE, в результате чего белковые комплексы с ДНК отделялись от остальной ДНК. В качестве положительного контроля использовали фрагмент ДНК с известным σ^{70}-промотором гена *hns*, который амплифицировали с праймерами hns_Bgl_263 и hns_Xba (Приложение Г). Участок гена hns, в котором нет промоторных областей, использовали в качестве отрицательного контроля (с праймерами hns_RT и hns_PCR).

Для того, чтобы оценить эффективность связывания комплекса цАМФ-CRP с регуляторными последовательностями *yih*-генов мы применяли такой же метод, но с лизатом клеткок с суперпродукцией белка CRP. Благодаря этому нам удалось избежать необходимости добавлять внеклеточный CRP к образцу *in vitro*. Белок CRP в избыточном количестве синтезировался в клетках BL21*(DE3). Мы

отбирали 10 мл этой культуры, отмывали и ресуспендировали ее в 3мл 0.5× транскрипционного буфера (Transcription buffer) с 10 мМ фенилметилсульфонил фторидом (PMSF). Лизат готовили с помощью стандартного метода трех 30-секундных повторов обработки культуры ультразвуком. Итоговую концентрацию белка определяли с помощью реактива Bradford (Sigma, США). После оценки приблизительной молярной концентрации CRP в лизате для проведения электрофореза использовались концентрации в соотношениях с ДНК от 1:1 до 8:1.

Для того, чтобы оценивать эффективность образования комплексов при проведении обоих типов анализа, свободные фрагменты ДНК и белки также наносили на отдельные дорожки геля.

После проведения электрофореза гели окрашивали бромистым этидием, который позволял увидеть полоски, содержащие ДНК. Для подтверждения наличия CRP в соответствующих комплексах использовался вестерн-блоттинг. Для этого все дорожки в данном электрофорезе были дублированы: один набор красили бромистым этидием, другой переносили на поливинилиденфторидовые (PVDF) мембраны (Immobilon, Sigma-Aldrich, США) с использованием системы мини транс-блот (Bio-rad, США). Мембрану затем блокировали с помощью 0,5% обезжиренного молока (Oxoid, США) и инкубировали с антителами к CRP (T14,Cell Signalling,США) в буфере TBS (20 мМ Tric-HCl, pH 7,4, 150 мМ NaCl) с 0,5% обезжиренном молоком и 0,1% Tween-20 в течение 2 часов при 37 °C. После этого ее отмывали три раза по 10 минут в TBS-T, добавляли вторичные антитела (A3687, антитела к IgG кролика, Sigma, США) и подвергали инкубации в течение 2 часов при комнатной температуре. Затем мембраны красили стабилизированным субстратом для щелочной фосфатазы (Western-Blue stabilized substrate for alkaline phosphatase, Promega, США) и проводили сканирование.

Для того, чтобы оценить вклад цАМФ в эффективность связывания CRP с ДНК, мы провели такой же эксперимент с очищенным белком CRP, с добавлением 200 мкМ цАМФ к образцам, к гелю и к буферу 1xTBE и без него; для предотвращения неспецифического связывания использовался гепарин, который добавляли к образцам в концентрации 20 мкг/мл.

4.2.6. Количественная ПЦР

Для проведения количественной ПЦР использовался амплификатор DT-322 (DNA-Technology, Россия), в качестве флуоресцентного интеркалирующего красителя применялся SYBR Green I (Invitrogen, США). Праймеры, которые использовали для обратной транскрипции (с окончанием "-RT" в названии) и амплификации (с окончанием "-PCR" в названии), указаны в Приложении Г. В отсутствии обратной транскриптазы в отрицательном контроле не наблюдалось продуктов ПЦР. Чтобы избежать влияния изменения роста культуры на транскрипцию генов в целом, в качестве дополнительного контроля также использовался ген *hns* и антисмысловая РНК *ysaA* [113].

Данные, полученные не менее чем от трех образцов в трех повторностях подвергали статистическому анализу с помощью стандартного метода для оценки экспрессии генов ΔΔCt [114]. Планки погрешностей на соответствующих графиках (Рис. 17 и Рис. 18) отражают стандартное отклонение от средних значений.

4.3. Результаты и обсуждение

4.3.1. Сходство кассет Enterobacteriaceae и Bacilli

Как уже говорилось выше, одна из консервативных длинных комбинаций функций в собранной нами базе данных кассет углеводного метаболизма обнаружилась как у ряда представителей семейства Enterobacteriaceae, так и в

геномах бактерий класса Bacilli. Эта комбинация включала шесть функциональных классов, и у *E. coli* была представлена в составе кассеты *ompL-yihOPQRSTUVW*, кодирующей ферменты сульфогликолиза, а у Bacilli – в составе кассеты *lacGEFDCBAR*, кодирующей ферменты пути катаболизма лактозы (Рис. 15).

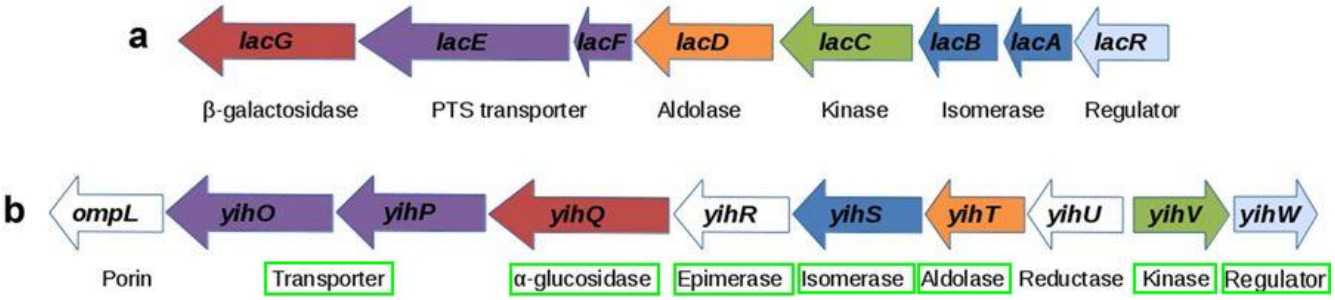

Рис. 15. Кассеты бактерий класса Bacilli (a) и семейства Enterobacteriaceae(b). Одинаковые цвета генов указывают на сходство функций кодируемых ими белков. Белым отмечены гены, кодирующие функции, не представленные в другой кассете. Зеленым в кассете Bacilli обведены функции белков, которые могут быть задействованы в катаболизме лактозы.

Среди Enterobacteriaceae аннотированная кассета с таким составом встречалась у большей части штаммов *E. coli*, а также у *Enterobacter cloacae*, *Salmonella enterica*, *Cronobacter turicensis* и *Pantoea anantis.* У Bacilli соответствующая аннотированная кассета встречалась в полном виде у видов *Streptococcus* (*S. gallolyticus*, *S. suis*, *S. pyogenes*, *S. agalactiae*, *S. uberis*, *S. equi*, *S. mutans* и *S. sanguinis*) и видов *Staphylococcus* (*S. aureus*, *S. epidermidis*, *S. haemolyticus* и *S. lugdunensis*).

Гены, принадлежащие к одним и тем же функциональным классам, обладали, соответственно, общими функциональными характеристиками (см. Главу 3). Согласно номенклатуре ферментов Enzyme Nomenclature, кодируемые ими белки представляли собой гликозидазу (3.2.1), дегидрогеназу (EC 4.1.2), киназу (EC 2.7.1) и изомеразу (2.3.1). Пятым и шестым пересечением кассет были гены,

кодирующие транспортер и транскрипционный фактор. Гены, кодирующие ферменты из класса дегидрогеназ (альдолазы), принадлежали, кроме того, к одному и тому же кластеру COG (COG3684).

В кассете Enterobacteriaceae в катаболизме лактозы (см. Рис. 3) могли участвовать белки, кодируемые генами *yihO*, *yihP* (транспортеры), *yihQ* (гидролаза), *yihS* (изомераза), *yihT* (альдолаза), *yihV* (киназа) и *yihW* (транскрипционный фактор). Эти функции присутствуют в обоих кассетах. Кроме того, в катаболизме лактозы потенциально мог участвовать ген *yihV* (эпимераза), поскольку эпимераза присутствует в пути катаболизма галактозы Лелуара [115],взаимодействуя с молекулой галактозы после реакции изначального гидролиза лактозы (превращая β-D-галактозу в α-D-галактозу).

Нашей задачей было проверить экспериментально, участвует ли кассета *E. coli* в катаболизме лактозы. Подтвержденных данных о позициях сайтов инициации транскрипции (промоторов) для генов *yih*-кассеты на момент начала работы не было, поэтому первым шагом стало картирование этих сайтов.

4.3.2. Промоторные области *yih*-кассеты Escherichia coli

Анализ межгенных участков с помощью методов филогенетического футпринта и алгоритма для поиска промоторных последовательностей PlatProm позволил нам предсказать потенциальные промоторные участки для генов *E. coli*, кодирующих альдолазу (*yihT*), киназу (*yihV*), изомеразу (*yihS*) и альфа-гликозидазу (*yihQ*). Чтобы получить полную картину транскрипционных особенностей данного геномного локуса, мы также включили в эту часть исследования гены *yihU* и *yihR*, кодирующие редуктазу и эпимеразу, соответственно (Рис. 16, а).

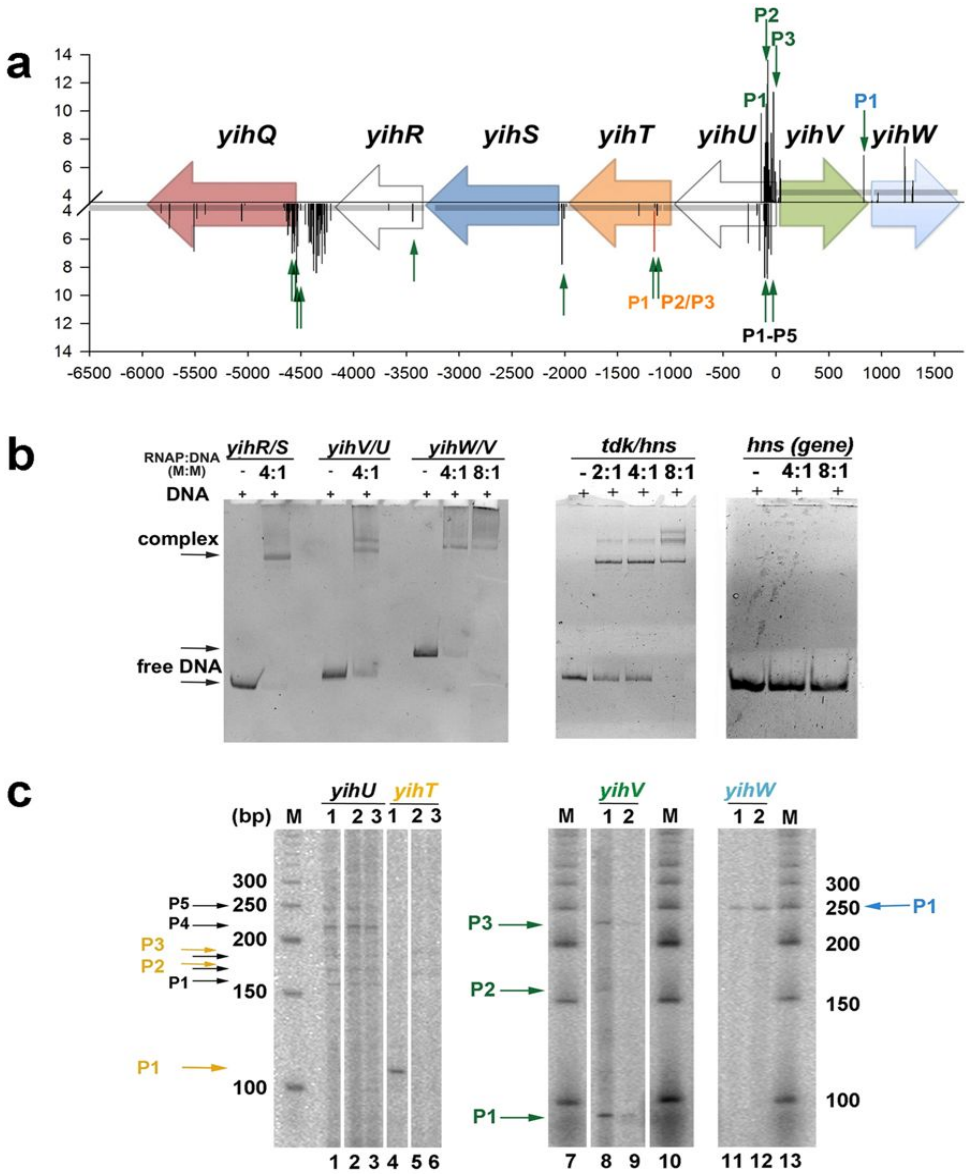

Рис. 16. Картирование промоторов в *yih*-кассете *Escherichia coli*. (a) Схема расположения генов в кассете с отмеченными промоторами, предсказанными *in silico*. Горизонтальными стрелками отмечены позиции генов. Столбиками отмечены точки начала транскрипции, предсказанные с помощью алгоритма PlatProm на обеих нитях ДНК, а соответствующие скоры PlatProm указаны на оси Y. Скор промотора yihTP1 отмечен красным. По оси абсцисс указана позиция относительно аннотированного старт-кодона гена *yihV*. Зеленые вертикальные стрелки указывают на 5'-концы РНК, выявленные с помощью специфического 5'-концевого РНК-секвенирования. (b) Результаты электрофореза с задержкой в геле с σ^{70}-РНК полимеразой указывает на наличие единичных промоторов в межгенных участках yihV/W и yihS/R и двух промоторов в участке yihU/V. Молярное соотношение РНК-полимеразы и ДНК-фрагментов указано над дорожками. Все образцы были подвергнуты электрофорезу на одном и том же геле. Положительным контролем служил известный σ^{70}-зависимый промотор гена *hns*. Соседний участок его внутригенной области, не имеющий промоторов, использовался как отрицательный контроль. (c) Реакция удлинения праймера выявила точку начала транскрипции для гена *yihW* и несколько точек начала транскрипции для генов *yihU*, *yihT* и *yihV*. Некоторые из них активировались при росте культуры на лактозе: (1) рост на лактозе, (2) рост на глюкозе, (3) рост на глицерине. Образцы с *yihU* и *yihT* были подвергнуты электрофорезу на одном геле, образцы с *yihV* и *yihW* на другом.

Предсказания, полученные таким образом, мы сравнили с точками начала транскрипции, полученными с помощью метода специфического 5'-концевого РНК-секвенирования [116]. Полученные потенциальные точки начала транскрипции были подтверждены с помощью реакции удлинения праймера (Рис. 16, с). Метод задержки в геле показал наличие промоторов, способных успешно с связываться с σ^{70}-РНК-полимеразой, в межгенных участках *yihU/V*, *yihV/W* и *yihS/R*. В межгенном участке *yihU/V*, более того, практически с одинаковой эффективностью произошло формирование двух комплексов, что указывает на наличие по крайней мере двух одинаково сильных промоторов между этими генами. Специфичность связывания межгенных участков с полимеразой была подтверждена с помощью положительного контроля, в котором было продемонстрировано эффективное связывание полимеразы с известным σ^{70}-зависимым промотором гена *hns*, и отрицательного контроля, который показал отсутствие образования комплексов в соседнем участке этого же гена, на котором не имелось σ^{70}-зависимых промоторов.

Чтобы выяснить, насколько работа соответствующих промоторов зависит от присутствия лактозы в среде, мы провели реакцию удлиннения праймера с РНК, выделенной из клеток, которые росли в течение 6 часов на средах M9 + 10% LB в присутствии 0,2% лактозы, глюкозы или глицерина как единственного источника углерода.

Обратная транскрипция с праймером yihU_RT выявила наличие нескольких продуктов (отмечены черными стрелками на Рис. 16, с), соответствующих точкам начала транскрипции в позициях −15 (скор PlatProm составил 6.83), −25 (5.04), −35 (3.52), −62/−63 (8.75), −82/−92 (крупный кластер со скорами 8.75–13.61) относительно аннотированного старт-кодона ATG. Они были обозначены как

yihUP1-yihUP5. Наличие транскрипции РНК с промоторов yihUP1, yihUP2, yihUP4 и yihUP5 также было подтверждено с праймером yihU/yihV_F (см. Приложение Г) и с помощью методики 5′-концевого специфического РНК-секвенирования (зеленые стрелки на Рис 14, а). Стоит отметить, однако, что транскрипционная активность ни одного из этих промоторов не изменялась при добавлении разных источников углерода, что, по-видимому, означает, что *yihU* (кодирующий редуктазу) не участвует в катаболизме лактозы. Это соответствовало нашим представлениям о потенциальном лактозном пути, закодированном в данной кассете, поскольку редуктазы в нем не имеется (см Рис. 3).

Для гена *yihV* (киназы) мы также картировали несколько стартов транскрипции, основной из которых располагался в позиции –25/–27 относительно аннотированного старт-кодона ATG (крупный кластер предсказанных промоторов с максимальным скором 11.35, в позиции 4071737 в геноме кишечной палочки U00096.2). Нам удалось показать, что этот промотор активировался во время роста культуры бактерий на лактозе (Рис. 16, с).

Для гена *yihT* мы выявили три возможных точки начала транскрипции (оранжевые стрелки на Рис. 16, а). Две из них, соответствующие промоторам yihTP2 и yihTP3, расположенные в позиции +35/+45 относительно предполагаемого старт-кодона ATG, обладали относительно низкими скорами, 3.67–4.33, что объясняло их низкую транскрипционную активность. Во время роста на лактозе, однако, их активность полностью исчезала (дорожка 1 на Фиг.), в то же время активировался другой промотор (yihTP1), соответствующая точка начала транскрипции которого была расположена в позиции +93/+94 относительно старт-кодона (дорожка 4 на Рис. 16, с). Этот промотор был предсказан с помощью унифицированного алгоритма PlatPromU, а не PlatProm, что позволяет предположить его связывание с альтернативными сигма-факторами, что, однако,

трудно проверить путем непосредственного эксперимента из-за близкого расположения промоторов. Присутствие транскрипционного переключателя в данном участке также подтверждается наличием в нем мотива GCGC между точкой начала транскрипции и –10 элементом промотора yihTP1. Известно, что такой мотив может быть связан с переключением транскрипционной активности у бактерий при голодании [47,48,105].

Все картированные промоторы для гена *yihT* располагались в пределах ORF (открытой рамки считывания), что указывает либо на неправильную аннотацию точки старта трансляции белка (мы выявили по крайней мере четыре потенциальных альтернативных старт-кодона, расположенных в позиции +51, +78, +84 и +87 относительно аннотированного), либо на их регуляторную роль в транскрипции данного гена [117].

Для гена *yihW* мы картировали один промотор, он располагался в позиции −27/−28 относительно аннотированного старт-кодона ATG (Рис. 16, а).

4.3.3. Экспрессия генов во время роста культуры на разных источниках углерода

Чтобы детально выяснить, как наличие лактозы в среде меняет характер экспрессии генов *yih*-кассеты, мы провели сравнительный анализ уровней соответствующих мРНК при росте культуры на разных источниках углерода. Клетки росли в тех же условиях, которые мы ранее использовали для сравнительного анализа в эксперименте с Primer Extension, то есть на глюкозе, лактозе и глицерине. Кроме того, мы проанализировали экспрессию *yih*-генов при росте на галактозе, возможном промежуточном соединении пути катаболизма лактозы. В данном случае среда с глюкозой представляла собой стандартную

углеводную среду, а среда с глицерином служила контрольной средой без углеводов (соответственно, представляя собой бедный источник углерода).

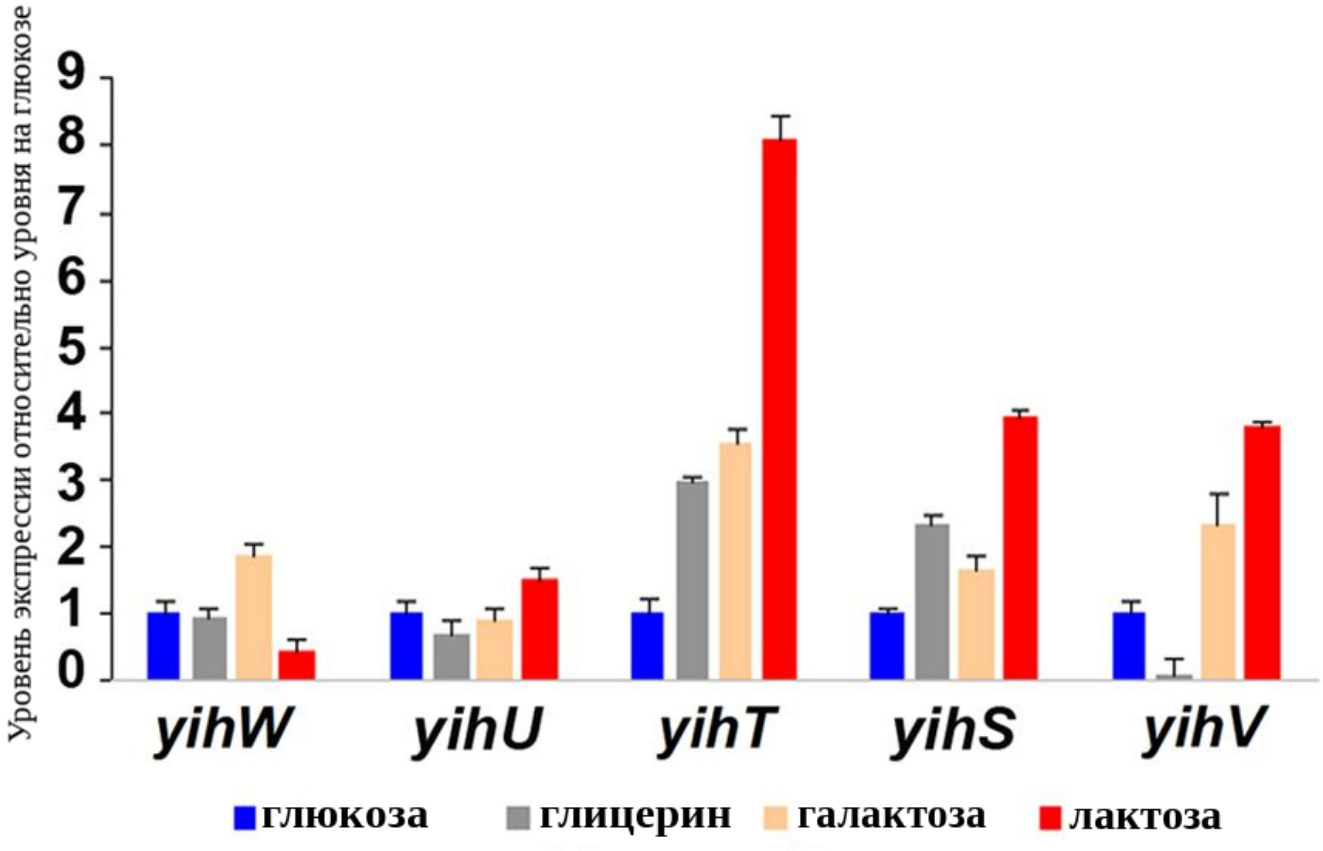

Рис. 17. Уровни мРНК сравнивали с помощью метода количественной ПЦР с детекцией в реальном времени (qRT-PCR). Условия роста культур указаны снизу – глюкоза, глицерин, галактоза и лактоза. В качестве контроля использовали мРНК гена *hns* и аРНК гена *ysaA*, их уровень экспрессии не менялся. Уровень экспрессии генов при росте на глюкозе взят за единицу. Стандартное отклонение было посчитано с помощью трех биологических и трех технических повторов.

Результаты количественного ПЦР-анализа с детекцией в реальном времени (qRT-PCR) для культур, росших в течение 6 часов в разных условиях, представлены на Рис. 17. Выяснилось, что гены, кодирующие киназу (*yihV*), альдолазу (*yihT*) и изомеразу (*yihS*), активировались при росте культуры на лактозе и галактозе, причем на лактозе эта активация была наиболее существенно выражена. Экспрессия гена *yihW*, кодирующего фактор транскрипции, напротив,

на лактозе была подавлена почти вдвое, что может указывать на его роль в данной временной точке в качестве репрессора *yih*-генов и/или собственного гена. Экспрессия гена *yihU*, кодирующего редуктазу, не зависела от источника углерода, что еще раз подтверждает отсутствие роли этого гена в потенциальном пути катаболизма лактоза. Это, как уже говорилось выше, соответствует изначальному предположению о составляющих ферментах данного пути.

Экспрессия генов *yihR* и *yihQ*, кодирующих эпимеразу и гликозидазу, не зависела (в случае *yihR*) и практически не зависела (в случае *yihQ*) от источника углерода в среде. Это соответствовало нашему предположению об их независимой экспрессии и позволяло предположить, что эти гены, по-видимому, не участвуют в катаболизме лактозы.

Стоит отметить, что многие представители семейства Enterobacteriaceae, включая штаммы видов *Escherichia albertii* и *Citrobacter koseri*, обладают ортологами генов *yihSTUVW*, в то время как остальная часть *yih*-кассеты, в том числе гены *yihR* и *yihQ*, отсутствует. Это подтверждает наше предположение о том, что разные части исходной десятигенной кассеты функционируют по-разному и могут участвовать в разных метаболических путях независимо друг от друга.

4.3.4. Роль транскрипционных факторов cAMP-CRP и YihW в регуляции транскрипции *yih*-кассеты

Снижение уровня мРНК гена *yihW* в присутствии лактозы позволило нам предсказать участие транскрипционного фактора YihW в регуляции генов *yih*-кассеты в качестве локального регулятора. Мы предположили, что YihW может работать в паре с глобальным регулятором углеводного метаболизма, cAMP-CRP. Для того, чтобы проверить это предположение, мы провели поиск потенциальных

сайтов связывания CRP в области картированных промоторов и выяснили, являются ли соответствующие промоторы CRP-зависимыми.

Потенциальные сайты связывания CRP были обнаружены в межгенных участках *yihT/U*, *yihU/V* и *yihV/W*. Метод филогенетического футпринтинга, в ходе которого мы выравнивали последовательности этих участков у разных представителей семейства Enterobacteriaceae, показал достаточно высокую их консервативность. Так, в межгенном участке *yihV/W* в позиции −41.5 относительно начала промотора *yihW* имеется высококонсервативный мотив, расположение которого типично для CRP-зависимых промоторов II класса (Рис. 18, d)

Нам удалось подтвердить высокую эффективность связывания CRP с указанными участками экспериментальным путем с помощью метода задержки в геле (Рис. 18, b). Для этого вначале мы провели эксперимент с клеточным лизатом, содержащим суперпродуцированный CRP (уровень экспрессии CRP показан на Рис. 18, a). Присутствие CRP в комплексах с соответствующими межгенными участками было подтверждено с помощью вестерн-блоттинга. Сильное специфическое связывание CRP с участками *yihV/U* и *yihV/W* было также подтверждено в отдельном эксперимента с использованием очищенного белка (см. Методы). Взаимодействие с обоими участками увеличивалось на ~30% в присутствии цАМФ (Рис. 18, c).

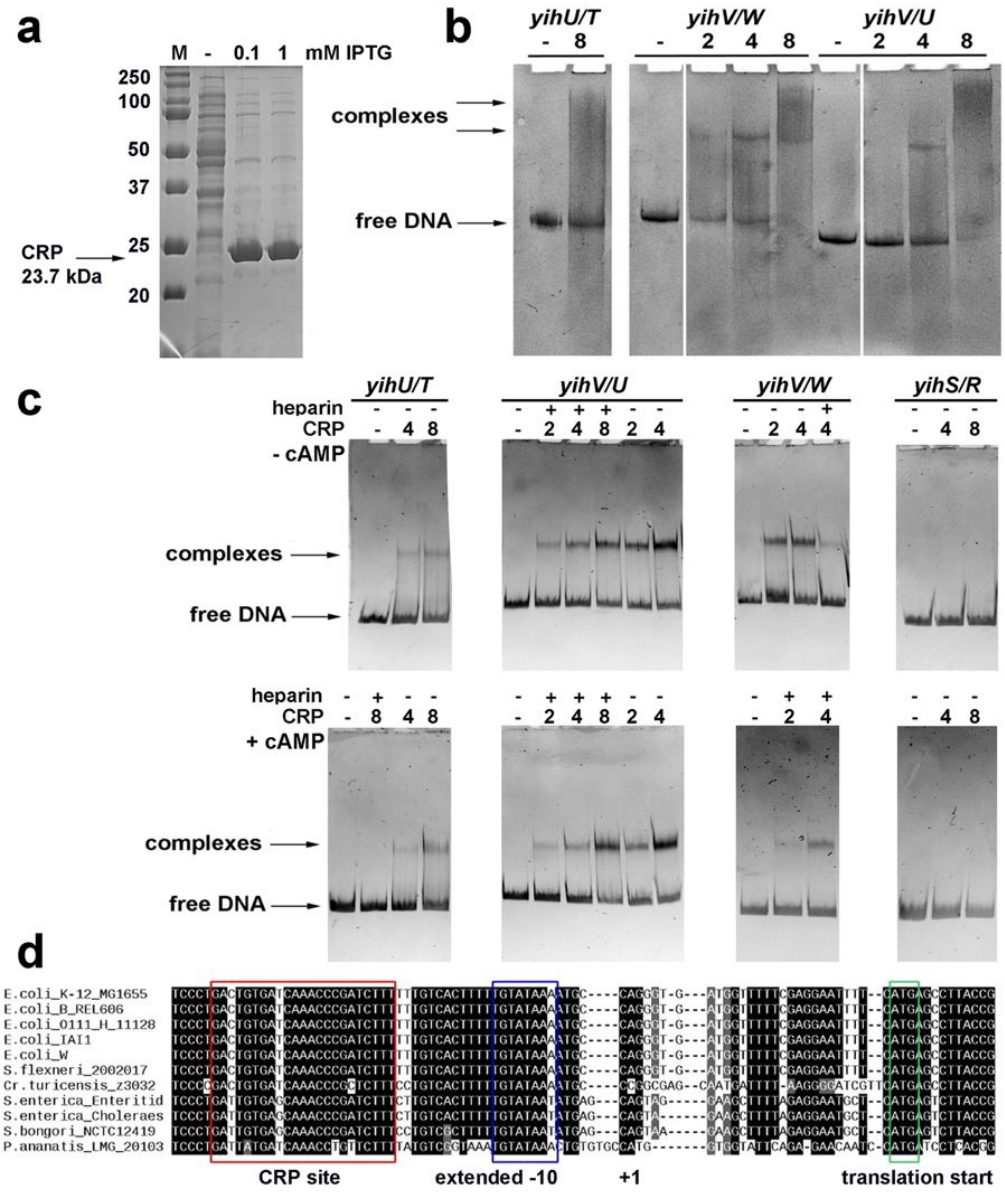

Рис. 18. (a) Уровень cAMP-CRP, синтезированного в клетках из pET_CRP после индукции IPTG. (b) cAMP-CRP взаимодействует с межгенными участками *yihU/T*, *yihV/W* и *yihV/U*; наиболее эффективное связывание показано для участка *yihV/W*. Молярное соотношение белка и ДНК указано над дорожками. Контрольная дорожка для участка *yihV/W* находится на отдельном геле вместе с дорожкой с соотношением белка и ДНК 1:1 (для данного варианта существенного уровня связывания показано не было). (c) Связывание очищенного CRP с исследуемыми участками в присутствии и в отсутствии cAMP (отмечено как -cAMP в верхнем ряду и +cAMP в нижнем ряду). Соотношение белка и ДНК, а также наличие гепарина в образце указано над дорожками. (d) Множественное выравнивание межгенных областей ортологов yihV/W для нескольких представителей бактерий Enterobacteriaceae, которое показало высокую консервативность картированного промотора гена *yihW* (extended -10, отмечен синим) и потенциального сайта связывания CRP (CRP site, отмечен красным). Предполагаемый старт-кодон ATG отмечен зеленым.

Мы предположили, что фактор CRP работает как глобальный регулятор для всех интересующих нас *yih*-генов, а фактор YihW – как локальный. Для того, чтобы проверить это предположение, мы исследовали уровень роста клеток K-12 MG1655 с выключенными генами *yihW* и *crp* на лактозе и глюкозе.

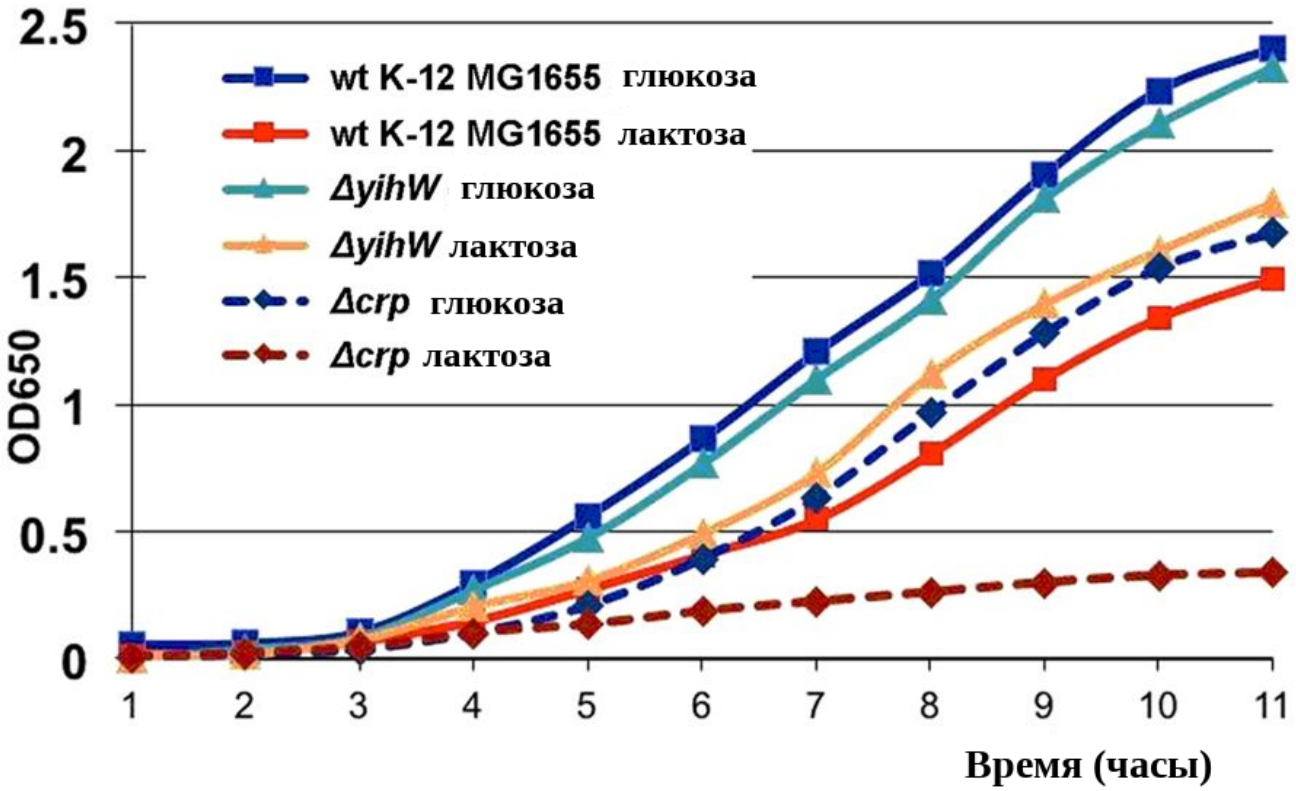

Рис. 19. Влияние выключения генов *yihW* и *crp* на рост клеточной культуры в присутствии 0,2% глюкозы или лактозы в течение 11 часов. Линией с квадратами представлен рост исходной культуры, с треугольниками – мутанта по *yihW*, прерывистой линией – мутанта по *crp*. Каждая кривая построена исходя из средних значений трех независимых измерений оптической плотности.

На среде с глюкозой клетки дикого типа и мутанты по *yihW (*Δ *yihW)* росли с одинаковой скоростью, в то время как скорость роста мутантов по *crp* (*Δcrp)* была снижена, что объясняется ключевой ролью данного транскрипционного фактора в общей регуляции сахарного метаболизма кишечной палочки. На лактозе клетки дикого типа росли медленнее, чем на глюкозе, а рост Δ*yihW* оказался, напротив,

значительно быстрее (Рис. 19). При этом роста *Δcrp* на лактозе после поглощения базовых питательных веществ среды LB практически не наблюдалось. Эти наблюдения указывали на то, что YihW, по-видимому, действительно участвует в регуляции лактозного метаболизма, причем его роль может быть противоположна роли CRP. Следующий этап работы состоял в детальном исследовании роли данных белков в транскрипции *yih*-генов.

Предположение о том, что YihW, действительно, регулирует работу генов *yih*-кассеты, было подтверждено с помощью количественного ПЦР-анализа с детекцией в реальном времени (Рис. 20). Для этого мы использовали штамм *E. coli* M182 с выключенным лактозным опероном (*lac*-опероном) [108]. Его клетки не могли катаболизировать лактозу с помощью своего стандартного, хорошо известного пути. Всего работа проводилась с тремя типами культур M182 – диким типом (wt), мутантом по *yihW* и мутантом по *crp* [106]. Культуры росли на так называемой "минимальной среде" – с уменьшенной вдвое концентрацией LB (5%), чтобы можно было наиболее отчетливо наблюдать эффекты, вызванные сменой основного источника углерода.

Выяснилось, что экспрессия гена *yihT* как на глюкозе, так и на лактозе контролируется фактором YihW, который выполняет роль углевод-зависимого двойного переключателя (Рис. 20, a). Во время роста на глюкозе экспрессия гена *yihT* подавляется фактором CRP. Экспрессия самого *yihW* активируется с помощью CRP на лактозе и подавляется на глюкозе (Рис. 20, b). Наконец, оба фактора YihW и CRP работают как репрессоры транскрипции гена *yihV* (Рис. 20, a).

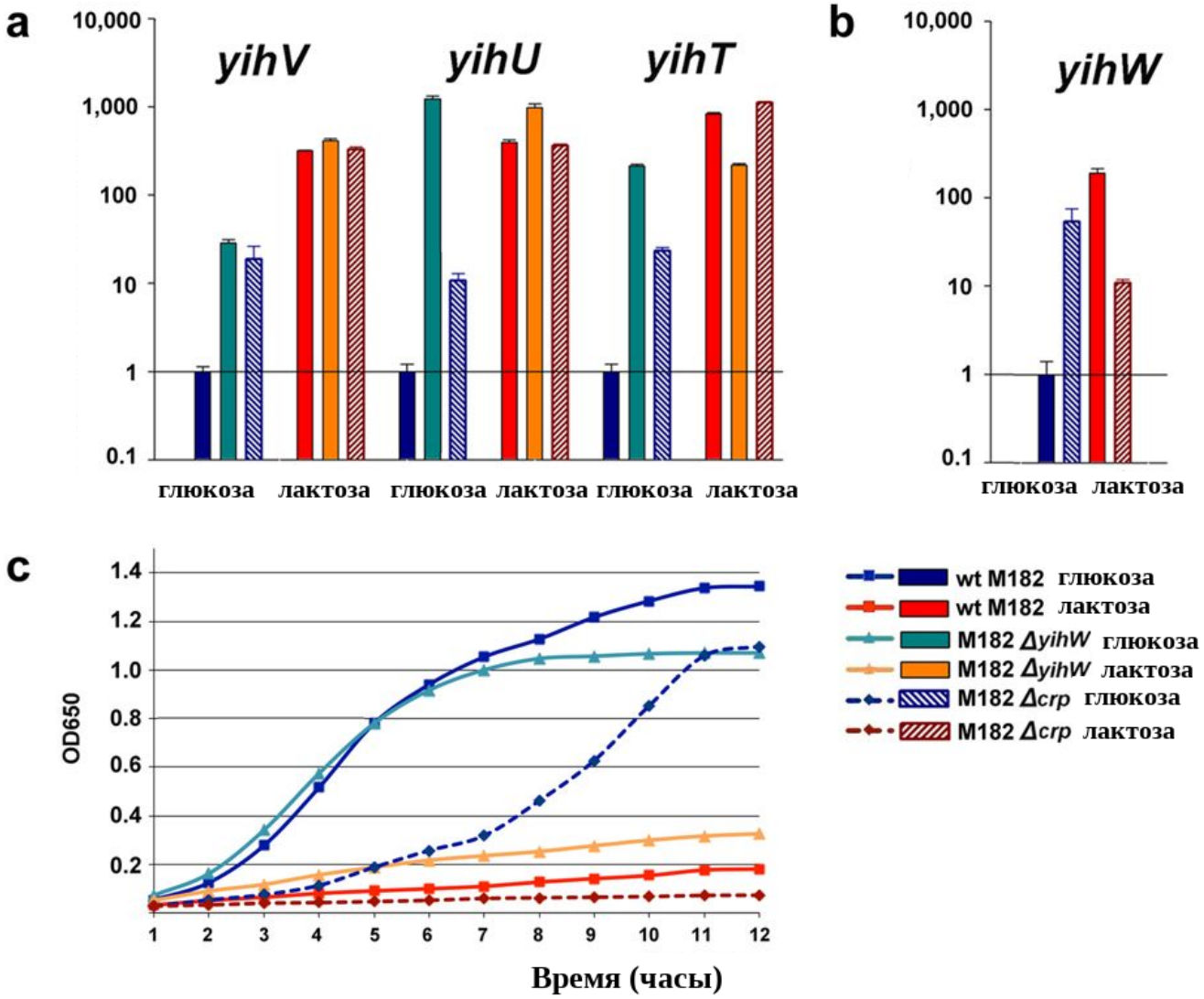

Рис. 20. Влияние делеции генов *yihW* и *crp* на уровень мРНК в генах *yih*-кассеты (a,b) и рост клеток на глюкозе и лактозе (c). Условия роста культур указаны снизу справа. Уровни мРНК указаны относительно уровня в родительском штамме при росте культуры на глюкозе. В качестве контроля использовались мРНК *hns* и аРНК *ysaA*, их уровни не менялись. Стандартные отклонения вычисляли на основе трех биологических и трех технических повторов. Кривая роста (c) построена на основании трех независимых измерений. Квадратами обозначен родительский штамм, треугольниками – мутант по гену *yihW*. Прерывистыми линиями обозначен мутант по *crp*.

4.3.5. Общая схема работы *yih*-кассеты *Escherichia coli*

На основании сходства основных функций белков, закодированных в *yih*-кассете генов семейства Enterobacteriaceae и в кассете класса Bacilli, участвующей в лактозном катаболизме, мы предположили, что *yih*-кассета может также

участвовать в утилизации этого сахара. Чтобы подтвердить эту гипотезу, мы провели детальное исследование экспрессии генов данной кассеты в разных условиях роста культуры.

Если брать за основу путь катаболизма лактозы у Bacilli (Рис. 3), то после первичной реакции гидролиза производные лактозы должны последовательно взаимодействовать с киназой, альдолазой и изомеразой, которые в данном случае представлены ферментами YihV, YihT и YihS, соответственно (Рис. 3). Первичная реакция гидролиза, возможно, происходит с участием β-галактозидазы из *lac*-оперона, однако даже штаммы с выключенным *lac*-опероном могли расти на средах с лактозой или галактозой (кроме штамма Δ*crp*), что указывает на существование альтернативных способов расщепления этого дисахарида, пока неизвестных (Рис. 19).

С помощью метода количественного ПЦР-анализа с детекцией в реальном времени мы выяснили, что экспрессия всех трех этих генов активируется как на лактозе, так и на галактозе, возможном промежуточном соединении пути катаболизма лактозы. Ген *yihW,* кодирующий потенциальный транскрипционный фактор семейства DeoR, активировался во время фазы экспоненциального роста культуры на лактозе, но был репрессирован после поглощения культурой основной части субстрата и выхода роста на плато.

Выраженный рост культур с выключенным *lac*-опероном в отсутствии YihW на лактозе, и, в еще большей степени, на галактозе, представляет отдельный интерес. Эти наблюдения поддерживают гипотезу о том, что фактор YihW и остальные гены *yih*-кассеты играют существенную роль в процессе роста кишечной палочки на лактозе в отсутствии *lac*-оперона, участвуя в катаболизме этого дисахарида и заменяя, таким образом, обычные функции *lac*-оперона.

Влияние на транскрипцию кассет генов локальных регуляторов, таких, как YihW, обычно либо дополняет, либо противопоставлено действию глобальных регуляторов, в зависимости от условий среды. В данном случае мы предположили и подтвердили с помощью экспериментальных методов, что глобальным регулятором для *yih*-кассеты служит фактор cAMP-CRP. Активация гена *yihW* во время экспоненциального роста культуры на среде с лактозой является CRP-зависимой. В отсутствии лактозы в среде CRP подавлял экспрессию гена *yihW*.

Работа YihW необходима для сбалансированной регуляции гена *yihT*, кодирующего альдолазу – этот фактор отвечает за активацию экспрессии *yihT* в присутствии лактозы и за ее подавление в отсутствии лактозы. По-видимому, то же касается и гена *yihS*, кодирующего изомеразу, поскольку профиль его транскрипции практически идентичен *yihT*. В случае, когда ген *yihW* выключен, экспрессия *yihT* полностью перестает зависеть от источника углерода (Рис. 17). CRP в этом случае играет роль репрессора, работа которого не зависит от типа углеводов в среде, по-видимому, за счет образования слабой связи с промоторными участками без образования комплекса с цАМФ.

Отдельно стоит отметить, что белок YihT, предсказанной функцией которого является 6-дезокси-6-сульфофруктозо-1-фосфат альдолаза, является гомологом тагатозо-1,6-дифосфат альдолазы LacD (оба соответствующих гена принадлежат к кластеру ортологических групп генов COG3684). Ген, кодирующий тагатозо-1,6-дифосфат альдолазу, встречается как у многих видов *Salmonella* и *Shigella*, так и в ряде штаммов кишечных палочек, например, у *Escherichia coli* APEC O1 и O157:H7.

Мы использовали инструмент NSimScan (см. Главу 2), для того, чтобы оценить меру сходства между этими генами в разных геномах семейства Enterobacteriaceae, и выяснили, что сходство их последовательностей, как правило, превышает 80%,

причем большинство замен встречается в третьей позиции кодонов, т.е. не влияет на закодированные в них аминокислоты. При этом состав соседних генов различается у разных видов и штаммов семейства Enterobacteriaceae, однако в большинстве случаев поблизости от них располагается ген, гомологичный гену транскрипционного фактора YihW, расположенный к *lacD*/*yihT* дивергентным образом (так же, как в *yih*-кассете *E. coli*).

Экспрессия гена *yihV* также была активирована во время роста культуры на лактозе и подавлена во время роста на глюкозе. Регуляция обоих этих процессов также осуществлялась с помощью YihW и CRP (Рис. 21).

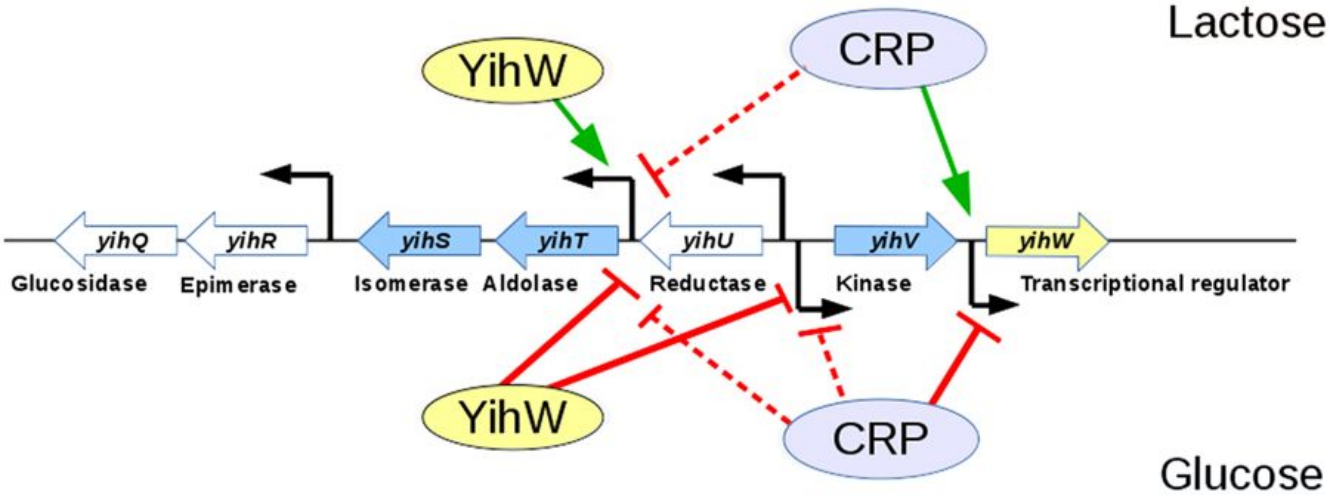

Рис. 21. Регуляция генов *yih*-кассеты при росте культуры на лактозе (сверху) и глюкозе (снизу), осуществляемая факторами транскрипции CRP и YihW. Зелеными стрелками отмечена активация транскрипции, красными линиями – подавление транскрипции. Прерывистыми линиями отмечены процессы, где происходит лишь умеренное подавление, которое, возможно, осуществляется не напрямую, а через дополнительные транскрипционные факторы.

В целом, YihW, по-видимому, играет в *yih*-кассете роль двойного переключателя, активируя некоторые ее гены (*yihT*, *yihS*) во время фазы экспоненциального роста культуры на лактозе, и репрессируя некоторые ее гены (*yihT*, *yihS*, *yihV*) во время

роста на глюкозе (Рис. 21). При этом, как уже говорилось выше, при полном выключении гена *yihW* культура на лактозе растет быстрее по сравнению с контролем; тем самым, механизм действия данного транскрипционного фактора оказался сложным и подлежит дальнейшему анализу. Фактор CRP при росте на лактозе активирует транскрипцию гена *yihW*, то есть выполняет роль, комплементарную YihW, а при росте на глюкозе репрессирует его транскрипцию. При выключении гена *crp* культура на лактозе растет гораздо медленнее по сравнению с контролем, что, вероятно, также связано с другими функциями этого глобального регулятора.

4.3.6. Заключение

С тех пор, как Жакоб и Моно показали схему работы лактозного оперона кишечной палочки в 1961 году, никаких альтернативных способов утилизации лактозы для *Escherichia coli* описано не было. В ходе данной работы мы сумели показать, что кассета генов этой бактерии, для которой ранее было известно только участие в деградации серосодержащих соединений глюкозы, также связана с катаболизмом лактозы. Таким образом мы можем говорить о наличии у кишечной палочки альтернативного пути утилизации лактозы, включающего в себя все этапы после первичного гидролиза.

Описанный случай является примером успешного предсказания функций генов на основе их ко-локализационных особенностей. Мы подтвердили, что консервативность комбинаций функциональных классов генов может служить поводом для предсказания функций соответствующих генов.

В исследовании *yih*-кассеты *Escherichia coli* мы, по-видимому, столкнулись с генами, кодирующими мультифункциональные ферменты. Выбранный нами

способ предсказания функций на основании сравнения кассет может позволять выявлять не только неизвестные, но и альтернативные известным функции генов.

Механизм регуляции транскрипции генов *yih*-кассеты оказался достаточно сложным и зависимым от условий среды. Пользуясь тонко налаженной системой регуляции транскрипции, бактерия, по всей видимости, может использовать один и тот же набор белков для разных задач. Эта работа поднимает ряд вопросов по поводу биохимических характеристик соответствующих ферментов, в частности, о механизмах их специфичности, взаимодействии с лактозой и серосодержащими молекулами, а также о регуляции экспрессии их генов во время роста на серосодержащих субстратах.

Выводы

1. Разработан инструмент для поиска нуклеотидных последовательностей удаленного сходства NSimScan; по совокупности таких параметров как чувствительность, точность и скорость он превосходит все стандартные инструменты в своей области. Наилучшим образом он подходит для поиска последовательностей, различающихся на 60-90%.

2. Описана сеть эволюционных связей 148 тысяч генов углеводного метаболизма 665 видов бактерий, выраженная в форме их ко-локализационных тенденций. 53% таких генов оказались ко-локализованы, остальные располагаются на бактериальных геномах по отдельности.

3. Склонность к ко-локализации, т.е. к формированию кассет различается у разных генов; ключевыми ее факторами являются функциональные и структурные характеристики гена и филогенетические свойства бактерии. Склонность к формированию кассет у разных функциональных классов составляет от 23 до 93%; у разных кластеров ортологических групп генов – 0 до 100%, у разных бактериальных классов – от 40 до 76%.

4. Среди 19 исследуемых функциональных классов 45 пар формируют консервативные и, по всей видимости, эволюционно значимые ко-локализационные связи. Количество таких связей для каждого класса сильно варьирует, что подчеркивает существенное различие в предпочтениях к геномному окружению у генов разных функций. Гены 11 функциональных классов демонстрируют выраженное предпочтение к внутриклассовой ко-локализации, причем большинство таких случаев, по-видимому, не является результатом событий локальных дупликаций.

5. Анализ консервативных сочетаний внутри кассет генов позволяет успешно предсказывать их функции. С его помощью предложено и экспериментально подтверждено участие *yih*-кассеты *Escherichia coli* в катаболизме лактозы; описан, таким образом, новый путь утилизации лактозы у кишечной палочки и предсказаны мультифункциональные характеристики соответствующих белков. В переключении механизмов экспрессии генов этой кассеты при росте бактерий в разных условиях среды участвуют локальный регулятор YihW и глобальный регулятор CRP.

Список литературы

1. Li H., Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform // Bioinformatics. 2009. Vol. 25, № 14. P. 1754–1760.
2. Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2 // Nature Methods. 2012. Vol. 9, № 4. P. 357–359.
3. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes // Nature. 2012. Vol. 491, № 7422. P. 56–65.
4. Pumpernik D., Oblak B., Borstnik B. Replication slippage versus point mutation rates in short tandem repeats of the human genome // Mol. Genet. Genomics. 2008. Vol. 279, № 1. P. 53–61.
5. Simpson J.T., Durbin R. Efficient construction of an assembly string graph using the FM-index // Bioinformatics. 2010. Vol. 26, № 12. P. i367–i373.
6. Ning Z., Cox A.J., Mullikin J.C. SSAHA: a fast search method for large DNA databases // Genome Res. 2001. Vol. 11, № 10. P. 1725–1729.
7. Kent W.J. BLAT--the BLAST-like alignment tool // Genome Res. 2002. Vol. 12, № 4. P. 656–664.
8. Camacho C. et al. BLAST+: architecture and applications // BMC Bioinformatics. 2009. Vol. 10. P. 421.
9. Pearson W.R. Flexible sequence similarity searching with the FASTA3 program package // Methods Mol. Biol. 2000. Vol. 132. P. 185–219.
10. Morgulis A. et al. Database indexing for production MegaBLAST searches // Bioinformatics. 2008. Vol. 24, № 16. P. 1757–1764.
11. Edgar R.C. Search and clustering orders of magnitude faster than BLAST // Bioinformatics. 2010. Vol. 26, № 19. P. 2460–2461.
12. Smith T.F., Waterman M.S. Identification of common molecular subsequences // J. Mol. Biol. 1981. Vol. 147, № 1. P. 195–197.
13. Giegerich R., Meyer C., Steffen P. A discipline of dynamic programming over sequence data // Science of Computer Programming. 2004. Vol. 51, № 3. P. 215–263.
14. Pearson W.R., Lipman D.J. Improved tools for biological sequence comparison // Proc. Natl. Acad. Sci. U.S.A. 1988. Vol. 85, № 8. P. 2444–2448.

15. Gumbel E.J. Les valeurs extrêmes des distributions statistiques // Annales de l’Institut Henri Poincaré. 1935. Vol. 5, № 2. P. 115–158.
16. Randle-Boggis R.J. et al. Evaluating techniques for metagenome annotation using simulated sequence data // FEMS Microbiol. Ecol. 2016. Vol. 92, № 7.
17. Pearson W.R. Comparison of methods for searching protein sequence databases // Protein Sci. 1995. Vol. 4, № 6. P. 1145–1160.
18. Campbell N. et al. Biology. 8th ed. 2008. P. 118.
19. Kanehisa M., Goto S. KEGG: kyoto encyclopedia of genes and genomes // Nucleic Acids Res. 2000. Vol. 28, № 1. P. 27–30.
20. Caspi R. et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases // Nucleic Acids Res. 2016. Vol. 44, № D1. P. D471-480.
21. Keseler I.M. et al. The EcoCyc database: reflecting new knowledge about Escherichia coli K-12 // Nucleic Acids Res. 2017. Vol. 45, № D1. P. D543–D550.
22. Kenyon J.J., Hall R.M. Variation in the complex carbohydrate biosynthesis loci of Acinetobacter baumannii genomes // PLoS ONE. 2013. Vol. 8, № 4. P. e62160.
23. Grondin J.M. et al. Polysaccharide Utilization Loci: Fueling Microbial Communities // J. Bacteriol. 2017. Vol. 199, № 15.
24. Voet D., Voet J., Pratt C. Fundamentals of Biochemistry: Life at the Molecular Level. 4th ed. John Wiley & Sons.
25. Ogata H. et al. A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters // Nucleic Acids Res. 2000. Vol. 28, № 20. P. 4021–4028.
26. Rodionov D.A. Comparative genomic reconstruction of transcriptional regulatory networks in bacteria // Chem. Rev. 2007. Vol. 107, № 8. P. 3467–3497.
27. Overbeek R. et al. The use of gene clusters to infer functional coupling // Proc. Natl. Acad. Sci. U.S.A. 1999. Vol. 96, № 6. P. 2896–2901.
28. Lodish H. et al. Molecular Cell Biology. 6th ed. W. H. Freeman, 2007.
29. Dandekar T. et al. Conservation of gene order: a fingerprint of proteins that physically interact // Trends Biochem. Sci. 1998. Vol. 23, № 9. P. 324–328.
30. Glazko G.V., Mushegian A.R. Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns // Genome Biol. 2004. Vol. 5, № 5. P. R32.

31. von Mering C. et al. Genome evolution reveals biochemical networks and functional modules // Proc. Natl. Acad. Sci. U.S.A. 2003. Vol. 100, № 26. P. 15428–15433.
32. Spirin V. et al. A metabolic network in the evolutionary context: multiscale structure and modularity // Proc. Natl. Acad. Sci. U.S.A. 2006. Vol. 103, № 23. P. 8774–8779.
33. Snel B., Huynen M.A. Quantifying modularity in the evolution of biomolecular systems // Genome Res. 2004. Vol. 14, № 3. P. 391–397.
34. Lawrence J. Selfish operons: the evolutionary impact of gene clustering in prokaryotes and eukaryotes // Curr. Opin. Genet. Dev. 1999. Vol. 9, № 6. P. 642–648.
35. Lawrence J.G., Roth J.R. Selfish operons: horizontal transfer may drive the evolution of gene clusters // Genetics. 1996. Vol. 143, № 4. P. 1843–1860.
36. Pellegrini M. et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles // Proc. Natl. Acad. Sci. U.S.A. 1999. Vol. 96, № 8. P. 4285–4288.
37. Li H., Pellegrini M., Eisenberg D. Detection of parallel functional modules by comparative analysis of genome sequences // Nat. Biotechnol. 2005. Vol. 23, № 2. P. 253–260.
38. Chen L., Vitkup D. Predicting genes for orphan metabolic activities using phylogenetic profiles // Genome Biol. 2006. Vol. 7, № 2. P. R17.
39. Daugherty M. et al. Archaeal shikimate kinase, a new member of the GHMP-kinase family // J. Bacteriol. 2001. Vol. 183, № 1. P. 292–300.
40. Mavromatis K. et al. Gene context analysis in the Integrated Microbial Genomes (IMG) data management system // PLoS ONE. 2009. Vol. 4, № 11. P. e7979.
41. Tatusov R.L. et al. The COG database: a tool for genome-scale analysis of protein functions and evolution // Nucleic Acids Res. 2000. Vol. 28, № 1. P. 33–36.
42. Galperin M.Y. et al. Expanded microbial genome coverage and improved protein family annotation in the COG database // Nucleic Acids Res. 2015. Vol. 43, № Database issue. P. D261-269.
43. Hartl D., Jones E.W. Genetics. 6th ed. Jones and Bartlett, 2005.
44. Dehal P.S. et al. MicrobesOnline: an integrated portal for comparative and functional genomics // Nucleic Acids Res. 2010. Vol. 38, № Database issue. P. D396-400.
45. Gama-Castro S. et al. RegulonDB version 9.0: high-level integration of gene regulation, coexpression, motif clustering and beyond // Nucleic Acids Res. 2016. Vol. 44, № D1. P. D133-143.
46. Chen I.-M.A. et al. IMG/M: integrated genome and metagenome comparative data analysis system // Nucleic Acids Res. 2017. Vol. 45, № D1. P. D507–D516.

47. Stülke J., Hillen W. Coupling physiology and gene regulation in bacteria: the phosphotransferase sugar uptake system delivers the signals // Naturwissenschaften. 1998. Vol. 85, № 12. P. 583–592.
48. Titgemeyer F., Hillen W. Global control of sugar metabolism: a gram-positive solution // Antonie Van Leeuwenhoek. 2002. Vol. 82, № 1–4. P. 59–71.
49. Peng X. et al. A multifunctional thermophilic glycoside hydrolase from Caldicellulosiruptor owensensis with potential applications in production of biofuels and biochemicals // Biotechnol Biofuels. 2016. Vol. 9. P. 98.
50. MacDonald L.C., Berger B.W. Insight into the role of substrate-binding residues in conferring substrate specificity for the multifunctional polysaccharide lyase Smlt1473 // J. Biol. Chem. 2014. Vol. 289, № 26. P. 18022–18032.
51. Rodionova I.A. et al. Diversity and versatility of the Thermotoga maritima sugar kinome // J. Bacteriol. 2012. Vol. 194, № 20. P. 5552–5563.
52. Carvalho S.M. et al. CcpA ensures optimal metabolic fitness of Streptococcus pneumoniae // PLoS ONE. 2011. Vol. 6, № 10. P. e26707.
53. Lulko A.T. et al. Transcriptome analysis of temporal regulation of carbon metabolism by CcpA in Bacillus subtilis reveals additional target genes // J. Mol. Microbiol. Biotechnol. 2007. Vol. 12, № 1–2. P. 82–95.
54. Chang D.-E. et al. Carbon nutrition of Escherichia coli in the mouse intestine // Proc. Natl. Acad. Sci. U.S.A. 2004. Vol. 101, № 19. P. 7427–7432.
55. Görke B., Stülke J. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients // Nat. Rev. Microbiol. 2008. Vol. 6, № 8. P. 613–624.
56. Mironov A.A. et al. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes // Nucleic Acids Res. 1999. Vol. 27, № 14. P. 2981–2989.
57. Aidelberg G. et al. Hierarchy of non-glucose sugars in Escherichia coli // BMC Syst Biol. 2014. Vol. 8. P. 133.
58. Bren A. et al. Glucose becomes one of the worst carbon sources for E.coli on poor nitrogen sources due to suboptimal levels of cAMP // Sci Rep. 2016. Vol. 6. P. 24834.
59. Kolb A. et al. Transcriptional regulation by cAMP and its receptor protein // Annu. Rev. Biochem. 1993. Vol. 62. P. 749–795.

60. Zheng D. et al. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling // Nucleic Acids Res. 2004. Vol. 32, № 19. P. 5874–5893.
61. Lee D.J., Busby S.J.W. Repression by cyclic AMP receptor protein at a distance // MBio. 2012. Vol. 3, № 5. P. e00289-00212.
62. Nakano M. et al. Involvement of cAMP-CRP in transcription activation and repression of the pck gene encoding PEP carboxykinase, the key enzyme of gluconeogenesis // FEMS Microbiol. Lett. 2014. Vol. 355, № 2. P. 93–99.
63. Busby S., Ebright R.H. Transcription activation by catabolite activator protein (CAP) // J. Mol. Biol. 1999. Vol. 293, № 2. P. 199–213.
64. Khoroshkin M.S. et al. Transcriptional Regulation of Carbohydrate Utilization Pathways in the Bifidobacterium Genus // Front Microbiol. 2016. Vol. 7. P. 120.
65. Kaplan S. et al. Diverse two-dimensional input functions control bacterial sugar genes // Mol. Cell. 2008. Vol. 29, № 6. P. 786–792.
66. Jacob F. et al. [Operon: a group of genes with the expression coordinated by an operator] // C. R. Hebd. Seances Acad. Sci. 1960. Vol. 250. P. 1727–1729.
67. Wang X.-G., Olsen L.R., Roderick S.L. Structure of the lac Operon Galactoside Acetyltransferase // Structure. 2002. Vol. 10, № 4. P. 581–588.
68. Huber R.E., Hurlburt K.L. Escherichia coli growth on lactose requires cycling of beta-galactosidase products into the medium // Can. J. Microbiol. 1984. Vol. 30, № 3. P. 411–415.
69. Huber R.E., Lytton J., Fung E.B. Efflux of beta-galactosidase products from Escherichia coli // J. Bacteriol. 1980. Vol. 141, № 2. P. 528–533.
70. Hengstenberg W., Penberthy W.K., Morse M.L. Purification of the staphylococcal 6-phospho-beta-D-- galactosidase // Eur. J. Biochem. 1970. Vol. 14, № 1. P. 27–32.
71. Hengstenberg W., Egan J.B., Morse M.L. Carbohydrate transport in Staphylococcus aureus. V. The accumulation of phosphorylated carbohydrate derivatives, and evidence for a new enzyme-splitting lactose phosphate // Proc. Natl. Acad. Sci. U.S.A. 1967. Vol. 58, № 1. P. 274–279.
72. Bissett D.L., Wenger W.C., Anderson R.L. Lactose and D-galactose metabolism in Staphylococcus aureus. II. Isomerization of D-galactose 6-phosphate to D-tagatose 6-phosphate by a specific D-galactose-6-phosphate isomerase // J. Biol. Chem. 1980. Vol. 255, № 18. P. 8740–8744.

73. Bissett D.L., Anderson R.L. Lactose and D-galactose metabolism in Staphylococcus aureus. III. Purification and properties of D-tagatose-6-phosphate kinase // J. Biol. Chem. 1980. Vol. 255, № 18. P. 8745–8749.
74. Bissett D.L., Anderson R.L. Lactose and D-galactose metabolism in Staphylococcus aureus. IV. Isolation and properties of a class I D-ketohexose-1,6-diphosphate aldolase that catalyzes the cleavage of D-tagatose 1,6-diphosphate // J. Biol. Chem. 1980. Vol. 255, № 18. P. 8750–8755.
75. Denger K. et al. Sulphoglycolysis in Escherichia coli K-12 closes a gap in the biogeochemical sulphur cycle // Nature. 2014. Vol. 507, № 7490. P. 114–117.
76. Kaznadzey A. et al. PSimScan: algorithm and utility for fast protein similarity search // PLoS ONE. 2013. Vol. 8, № 3. P. e58505.
77. Korobeinikova A.V., Garber M.B., Gongadze G.M. Ribosomal proteins: Structure, function, and evolution // Biochemistry Moscow. 2012. Vol. 77, № 6. P. 562–574.
78. Quast C. et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools // Nucleic Acids Res. 2013. Vol. 41, № Database issue. P. D590-596.
79. Wheeler D.L. et al. Database resources of the National Center for Biotechnology Information // Nucleic Acids Res. 2008. Vol. 36, № Database issue. P. D13-21.
80. Varghese N.J. et al. Microbial species delineation using whole genome sequences // Nucleic Acids Res. 2015. Vol. 43, № 14. P. 6761–6771.
81. Chen I.-M.A. et al. IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes // Nucleic Acids Res. 2019. Vol. 47, № D1. P. D666–D677.
82. Benson D.A. et al. GenBank // Nucleic Acids Res. 2013. Vol. 41, № Database issue. P. D36-42.
83. Bairoch A. The ENZYME database in 2000 // Nucleic Acids Res. 2000. Vol. 28, № 1. P. 304–305.
84. Marchler-Bauer A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures // Nucleic Acids Res. 2017. Vol. 45, № D1. P. D200–D203.
85. Marcotte C.J.V., Marcotte E.M. Predicting functional linkages from gene fusions with confidence // Appl. Bioinformatics. 2002. Vol. 1, № 2. P. 93–100.
86. Pertea M. et al. OperonDB: a comprehensive database of predicted operons in microbial genomes // Nucleic Acids Res. 2009. Vol. 37, № Database issue. P. D479–D482.

87. Westfall P., Young S. Resampling-based multiple testing : examples and methods for p-value adjustment. SERBIULA (sistema Librum 2.0), 2019.
88. Ding C., He X. K-means Clustering via Principal Component Analysis // Proceedings of the Twenty-first International Conference on Machine Learning. New York, NY, USA: ACM, 2004. P. 29–.
89. Eisen M.B. et al. Cluster analysis and display of genome-wide expression patterns // Proc. Natl. Acad. Sci. U.S.A. 1998. Vol. 95, № 25. P. 14863–14868.
90. Pál C., Hurst L.D. Evidence against the selfish operon theory // Trends Genet. 2004. Vol. 20, № 6. P. 232–234.
91. Davidson A.L. et al. Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems // Microbiol Mol Biol Rev. 2008. Vol. 72, № 2. P. 317–364.
92. Kotrba P., Inui M., Yukawa H. Bacterial phosphotransferase system (PTS) in carbohydrate uptake and control of carbon metabolism // Journal of Bioscience and Bioengineering. 2001. Vol. 92, № 6. P. 502–517.
93. Peekhaus N., Conway T. What's for dinner?: Entner-Doudoroff metabolism in Escherichia coli // J. Bacteriol. 1998. Vol. 180, № 14. P. 3495–3502.
94. Bloxham D.P. et al. A model study of the fructose diphosphatase-phosphofructokinase substrate cycle // Biochem. J. 1973. Vol. 134, № 2. P. 581–586.
95. Eisenstein A.B. Current concepts of gluconeogenesis // Am. J. Clin. Nutr. 1967. Vol. 20, № 3. P. 282–289.
96. Senoura T. et al. New microbial mannan catabolic pathway that involves a novel mannosylglucose phosphorylase // Biochem. Biophys. Res. Commun. 2011. Vol. 408, № 4. P. 701–706.
97. Maier E., Kurz G. D-Galactose dehydrogenase from Pseudomonas fluorescens // Meth. Enzymol. 1982. Vol. 89 Pt D. P. 176–181.
98. Wong T.Y., Yao X.T. The DeLey-Doudoroff Pathway of Galactose Metabolism in Azotobacter vinelandii // Appl. Environ. Microbiol. 1994. Vol. 60, № 6. P. 2065–2068.
99. Ermolaeva M.D., White O., Salzberg S.L. Prediction of operons in microbial genomes // Nucleic Acids Res. 2001. Vol. 29, № 5. P. 1216–1221.
100. Kabisch A. et al. Functional characterization of polysaccharide utilization loci in the marine Bacteroidetes "Gramella forsetii" KT0803 // ISME J. 2014. Vol. 8, № 7. P. 1492–1502.

101. Lamothe G.T. et al. Genetic and biochemical characterization of exopolysaccharide biosynthesis by Lactobacillus delbrueckii subsp. bulgaricus // Arch. Microbiol. 2002. Vol. 178, № 3. P. 218–228.
102. Reams A.B., Roth J.R. Mechanisms of Gene Duplication and Amplification // Cold Spring Harb Perspect Biol. 2015. Vol. 7, № 2.
103. Kondrashov F.A. et al. Selection in the evolution of gene duplications // Genome Biol. 2002. Vol. 3, № 2. P. RESEARCH0008.
104. Makarova K.S. et al. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell // Nucleic Acids Res. 2005. Vol. 33, № 14. P. 4626–4638.
105. Voigt B. et al. The glucose and nitrogen starvation response of Bacillus licheniformis // Proteomics. 2007. Vol. 7, № 3. P. 413–423.
106. Datsenko K.A., Wanner B.L. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products // Proc. Natl. Acad. Sci. U.S.A. 2000. Vol. 97, № 12. P. 6640–6645.
107. Studier F.W. Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system // J. Mol. Biol. 1991. Vol. 219, № 1. P. 37–44.
108. Casadaban M.J., Cohen S.N. Analysis of gene control signals by DNA fusion and cloning in Escherichia coli // J. Mol. Biol. 1980. Vol. 138, № 2. P. 179–207.
109. Notredame C., Higgins D.G., Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment // J. Mol. Biol. 2000. Vol. 302, № 1. P. 205–217.
110. Shavkunov K.S. et al. Gains and unexpected lessons from genome-scale promoter mapping // Nucleic Acids Res. 2009. Vol. 37, № 15. P. 4919–4931.
111. Münch R. et al. Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes // Bioinformatics. 2005. Vol. 21, № 22. P. 4187–4189.
112. Ozoline O.N., Fujita N., Ishihama A. Mode of DNA-protein interaction between the C-terminal domain of Escherichia coli RNA polymerase alpha subunit and T7D promoter UP element // Nucleic Acids Res. 2001. Vol. 29, № 24. P. 4909–4919.
113. Purtov Y.A. et al. Promoter islands as a platform for interaction with nucleoid proteins and transcription factors // J. Bioinform. Comput. Biol. 2014. Vol. 12, № 02. P. 1441006.
114. Schmittgen T.D., Livak K.J. Analyzing real-time PCR data by the comparative C_T method // Nature Protocols. 2008. Vol. 3, № 6. P. 1101–1108.

115. Frey P.A. The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose // FASEB J. 1996. Vol. 10, № 4. P. 461–470.
116. Dornenburg J.E. et al. Widespread antisense transcription in Escherichia coli // MBio. 2010. Vol. 1, № 1.
117. Wade J.T., Grainger D.C. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes // Nat. Rev. Microbiol. 2014. Vol. 12, № 9. P. 647–653.

Приложения

Приложение А

Список исследуемых геномов

Вид и штамм бактерии	Класс
Acaryochloris marina MBIC11017	Cyanobacteria
Acetohalobium arabaticum DSM 5501	Firmicutes
Acholeplasma laidlawii PG-8A	Tenericutes
Achromobacter xylosoxidans A8	Proteobacteria
Acidaminococcus fermentans DSM 20731	Firmicutes
Acidimicrobium ferrooxidans DSM 10331	Actinobacteria
Acidithiobacillus caldus SM-1	Proteobacteria
Acidithiobacillus ferrooxidans ATCC 23270	Proteobacteria
Acidobacterium capsulatum ATCC 51196	Acidobacteria
Acidothermus cellulolyticus 11B	Actinobacteria
Acidovorax avenae subsp avenae ATCC 19860	Proteobacteria
Acidovorax citrulli AAC00-1	Proteobacteria
Acidovorax ebreus TPSY	Proteobacteria
Acinetobacter oleivorans DR1	Proteobacteria
Acinetobacter sp ADP1	Proteobacteria
Actinobacillus pleuropneumoniae serovar 3 str JL03	Proteobacteria
Actinobacillus succinogenes 130Z	Proteobacteria
Actinoplanes missouriensis 431	Actinobacteria
Actinosynnema mirum DSM 43827	Actinobacteria
Aerococcus urinae ACS-120-V-Col10a	Firmicutes
Aeromonas hydrophila subsp hydrophila ATCC 7966	Proteobacteria
Aeromonas veronii B565	Proteobacteria
Aggregatibacter aphrophilus NJ8700	Proteobacteria
Agrobacterium radiobacter K84	Proteobacteria
Agrobacterium sp H13-3	Proteobacteria
Agrobacterium tumefaciens str C58	Proteobacteria
Agrobacterium vitis S4	Proteobacteria
Akkermansia muciniphila ATCC BAA-835	Verrucomicrobia
Alcanivorax borkumensis SK2	Proteobacteria
Alicyclobacillus acidocaldarius subsp acidocaldarius DSM 446 NC_0132051 GI:258510020	Firmicutes
Alkalilimnicola ehrlichii MLHE-1	Proteobacteria
Alkaliphilus metalliredigens QYMF	Firmicutes
Alkaliphilus oremlandii OhILAs	Firmicutes
Allochromatium vinosum DSM 180	Proteobacteria
Aminobacterium colombiense DSM 12261	Synergistetes
Ammonifex degensii KC4	Firmicutes

Вид и штамм бактерии	Класс
Anaerolinea thermophila UNI-1	Chloroflexi
Anaeromyxobacter dehalogenans 2CP-1	Proteobacteria
Anaeromyxobacter sp Fw109-5	Proteobacteria
Anaplasma centrale str Israel	Proteobacteria
Anaplasma marginale str Florida	Proteobacteria
Anoxybacillus flavithermus WK1	Firmicutes
Aquifex aeolicus VF5	Aquificae
Arcanobacterium haemolyticum DSM 20595	Actinobacteria
Arcobacter nitrofigilis DSM 7299	Proteobacteria
Aromatoleum aromaticum EbN1	Proteobacteria
Arthrobacter aurescens TC1	Actinobacteria
Aster yellows witches'-broom phytoplasma AYWB	Tenericutes
Asticcacaulis excentricus CB 48	Proteobacteria
Atopobium parvulum DSM 20469	Actinobacteria
Azoarcus sp BH72	Proteobacteria
Azorhizobium caulinodans ORS 571	Proteobacteria
Azospirillum sp B510	Proteobacteria
Bacillus amyloliquefaciens DSM 7	Firmicutes
Bacillus atrophaeus 1942	Firmicutes
Bacillus cellulosilyticus DSM 2522	Firmicutes
Bacillus cereus 03BB102	Firmicutes
Bacillus clausii KSM-K16	Firmicutes
Bacillus coagulans 2-6	Firmicutes
Bacillus halodurans C-125	Firmicutes
Bacillus megaterium DSM 319	Firmicutes
Bacillus pseudofirmus OF4	Firmicutes
Bacillus pumilus SAFR-032	Firmicutes
Bacillus selenitireducens MLS10	Firmicutes
Bacillus subtilis subsp subtilis str 168	Firmicutes
Bacillus thuringiensis str Al Hakam	Firmicutes
Bacteroides helcogenes P 36-108	Bacteroidetes
Bacteroides salanitronis DSM 18170	Bacteroidetes
Bacteroides thetaiotaomicron VPI-5482	Bacteroidetes
Bacteroides vulgatus ATCC 8482	Bacteroidetes
Bartonella bacilliformis KC583	Proteobacteria
Bartonella clarridgeiae 73	Proteobacteria
Bartonella grahamii as4aup	Proteobacteria
Bartonella henselae str Houston-1	Proteobacteria
Baumannia cicadellinicola str Hc (Homalodisca coagulata)	Proteobacteria
Bdellovibrio bacteriovorus HD100	Proteobacteria
Beutenbergia cavernae DSM 12333	Actinobacteria
Bifidobacterium adolescentis ATCC 15703	Actinobacteria
Bifidobacterium dentium Bd1	Actinobacteria
Bifidobacterium longum subsp longum BBMN68	Actinobacteria
Blastococcus saxobsidens DD2	Actinobacteria
Bordetella avium 197N	Proteobacteria

Вид и штамм бактерии	Класс
Bordetella parapertussis 12822	Proteobacteria
Bordetella petrii DSM 12804	Proteobacteria
Borrelia hermsii DAH	Spirochaetes
Borrelia recurrentis A1	Spirochaetes
Borrelia turicatae 91E135	Spirochaetes
Brachybacterium faecium DSM 4810	Actinobacteria
Brachyspira hyodysenteriae WA1	Spirochaetes
Brachyspira murdochii DSM 12563	Spirochaetes
Brachyspira pilosicoli 95/1000	Spirochaetes
Bradyrhizobium japonicum USDA 110	Proteobacteria
Bradyrhizobium sp BTAi1	Proteobacteria
Brevibacillus brevis NBRC 100599	Firmicutes
Brevundimonas subvibrioides ATCC 15264	Proteobacteria
Brucella canis ATCC 23365	Proteobacteria
Brucella melitensis ATCC 23457	Proteobacteria
Brucella microti CCM 4915	Proteobacteria
Brucella ovis ATCC 25840	Proteobacteria
Brucella pinnipedialis B2/94	Proteobacteria
Buchnera aphidicola str 5A (Acyrthosiphon pisum)	Proteobacteria
Burkholderia cenocepacia AU 1054	Proteobacteria
Burkholderia gladioli BSR3	Proteobacteria
Burkholderia glumae BGR1	Proteobacteria
Burkholderia mallei ATCC 23344	Proteobacteria
Burkholderia phymatum STM815	Proteobacteria
Burkholderia phytofirmans PsJN	Proteobacteria
Burkholderia rhizoxinica HKI 454	Proteobacteria
Burkholderia sp CCGE1001	Proteobacteria
Burkholderia thailandensis E264	Proteobacteria
Burkholderia xenovorans LB400	Proteobacteria
Butyrivibrio proteoclasticus B316	Firmicutes
Caldicellulosiruptor bescii DSM 6725	Firmicutes
Caldicellulosiruptor hydrothermalis 108	Firmicutes
Caldicellulosiruptor kronotskyensis 2002	Firmicutes
Caldicellulosiruptor obsidiansis OB47	Firmicutes
Caldicellulosiruptor owensensis OL	Firmicutes
Caldicellulosiruptor saccharolyticus DSM 8903	Firmicutes
Campylobacter curvus 52592	Proteobacteria
Campylobacter fetus subsp fetus 82-40	Proteobacteria
Campylobacter jejuni subsp jejuni 81116	Proteobacteria
Campylobacter lari RM2100	Proteobacteria
Candidatus Amoebophilus asiaticus 5a2	Bacteroidetes
Candidatus Desulforudis audaxviator MP104C	Firmicutes
Candidatus Hamiltonella defensa 5AT (Acyrthosiphon pisum)	Proteobacteria
Candidatus Koribacter versatilis Ellin345	Acidobacteria
Candidatus Nitrospira defluvii	Nitrospirae
Candidatus Pelagibacter sp IMCC9063	Proteobacteria

Вид и штамм бактерии	Класс
Candidatus Phytoplasma australiense	Tenericutes
Candidatus Protochlamydia amoebophila UWE25	Chlamydiae
Candidatus Puniceispirillum marinum IMCC1322	Proteobacteria
Candidatus Ruthia magnifica str Cm (Calyptogena magnifica)	Proteobacteria
Candidatus Solibacter usitatus Ellin6076	Acidobacteria
Candidatus Sulcia muelleri CARI	Bacteroidetes
Candidatus Vesicomyosocius okutanii HA	Proteobacteria
Capnocytophaga canimorsus Cc5	Bacteroidetes
Capnocytophaga ochracea DSM 7271	Bacteroidetes
Carboxydothermus hydrogenoformans Z-2901	Firmicutes
Carnobacterium sp 17-4	Firmicutes
Catenulispora acidiphila DSM 44928	Actinobacteria
Caulobacter crescentus CB15	Proteobacteria
Caulobacter segnis ATCC 21756	Proteobacteria
Caulobacter sp K31	Proteobacteria
Cellulomonas flavigena DSM 20109	Actinobacteria
Cellulophaga algicola DSM 14237	Bacteroidetes
Cellulophaga lytica DSM 7489	Bacteroidetes
Cellvibrio japonicus Ueda107	Proteobacteria
Chitinophaga pinensis DSM 2588	Bacteroidetes
Chlamydia trachomatis 434/Bu	Chlamydiae
Chlamydophila abortus S26/3	Chlamydiae
Chlamydophila caviae GPIC	Chlamydiae
Chlamydophila felis Fe/C-56	Chlamydiae
Chlamydophila pneumoniae AR39	Chlamydiae
Chlorobaculum parvum NCIB 8327	Chlorobi
Chlorobium chlorochromatii CaD3	Chlorobi
Chlorobium limicola DSM 245	Chlorobi
Chlorobium luteolum DSM 273	Chlorobi
Chlorobium phaeobacteroides BS1	Chlorobi
Chlorobium phaeovibrioides DSM 265	Chlorobi
Chlorobium tepidum TLS	Chlorobi
Chloroflexus aggregans DSM 9485	Chloroflexi
Chloroflexus aurantiacus J-10-fl	Chloroflexi
Chloroflexus sp Y-400-fl	Chloroflexi
Chloroherpeton thalassium ATCC 35110	Chlorobi
Chromobacterium violaceum ATCC 12472	Proteobacteria
Chromohalobacter salexigens DSM 3043	Proteobacteria
Citrobacter koseri ATCC BAA-895	Proteobacteria
Citrobacter rodentium ICC168	Proteobacteria
Clostridiales genomosp BVAB3 str UPII9-5	Firmicutes
Clostridium acetobutylicum ATCC 824	Firmicutes
Clostridium beijerinckii NCIMB 8052	Firmicutes
Clostridium botulinum A2 str Kyoto	Firmicutes
Clostridium cellulolyticum H10	Firmicutes
Clostridium cellulovorans 743B	Firmicutes

Вид и штамм бактерии	Класс
Clostridium lentocellum DSM 5427	Firmicutes
Clostridium ljungdahlii DSM 13528	Firmicutes
Clostridium novyi NT	Firmicutes
Clostridium phytofermentans ISDg	Firmicutes
Clostridium saccharolyticum WM1	Firmicutes
Clostridium sticklandii DSM 519	Firmicutes
Clostridium tetani E88	Firmicutes
Clostridium thermocellum ATCC 27405	Firmicutes
Collimonas fungivorans Ter331	Proteobacteria
Colwellia psychrerythraea 34H	Proteobacteria
Conexibacter woesei DSM 14684	Actinobacteria
Coprothermobacter proteolyticus DSM 5265	Firmicutes
Coraliomargarita akajimensis DSM 45221	Verrucomicrobia
Corynebacterium kroppenstedtii DSM 44385	Actinobacteria
Corynebacterium resistens DSM 45100	Actinobacteria
Corynebacterium urealyticum DSM 7109	Actinobacteria
Coxiella burnetii CbuG_Q212	Proteobacteria
Croceibacter atlanticus HTCC2559	Bacteroidetes
Cronobacter sakazakii ATCC BAA-894	Proteobacteria
Cronobacter turicensis z3032	Proteobacteria
Cryptobacterium curtum DSM 15641	Actinobacteria
Cupriavidus necator N-1	Proteobacteria
Cupriavidus taiwanensis LMG 19424	Proteobacteria
Cyanobacterium UCYN-A	Cyanobacteria
Cytophaga hutchinsonii ATCC 33406	Bacteroidetes
Dechloromonas aromatica RCB	Proteobacteria
Deferribacter desulfuricans SSM1	Deferribacteres
Dehalococcoides sp BAV1	Chloroflexi
Dehalogenimonas lykanthroporepellens BL-DC-9	Chloroflexi
Deinococcus deserti VCD115	Deinococcus-Thermus
Deinococcus geothermalis DSM 11300	Deinococcus-Thermus
Deinococcus maricopensis DSM 21211	Deinococcus-Thermus
Deinococcus proteolyticus MRP	Deinococcus-Thermus
Deinococcus radiodurans R1	Deinococcus-Thermus
Delftia acidovorans SPH-1	Proteobacteria
Delftia sp Cs1-4	Proteobacteria
Denitrovibrio acetiphilus DSM 12809	Deferribacteres
Desulfarculus baarsii DSM 2075	Proteobacteria
Desulfatibacillum alkenivorans AK-01	Proteobacteria
Desulfitobacterium hafniense DCB-2	Firmicutes
Desulfobacterium autotrophicum HRM2	Proteobacteria
Desulfobulbus propionicus DSM 2032	Proteobacteria
Desulfococcus oleovorans Hxd3	Proteobacteria
Desulfohalobium retbaense DSM 5692	Proteobacteria
Desulfomicrobium baculatum DSM 4028	Proteobacteria
Desulfotalea psychrophila LSv54	Proteobacteria

Вид и штамм бактерии	Класс
Desulfotomaculum acetoxidans DSM 771	Firmicutes
Desulfotomaculum reducens MI-1	Firmicutes
Desulfotomaculum ruminis DSM 2154	Firmicutes
Desulfovibrio aespoeensis Aspo-2	Proteobacteria
Desulfovibrio alaskensis G20	Proteobacteria
Desulfovibrio desulfuricans subsp desulfuricans str ATCC 27774 NC_0118831 GI:220903286	Proteobacteria
Desulfovibrio salexigens DSM 2638	Proteobacteria
Desulfovibrio vulgaris str 'Miyazaki F'	Proteobacteria
Desulfurispirillum indicum S5	Chrysiogenetes
Desulfurivibrio alkaliphilus AHT2	Proteobacteria
Desulfurobacterium thermolithotrophum DSM 11699	Aquificae
Dichelobacter nodosus VCS1703A	Proteobacteria
Dickeya dadantii 3937	Proteobacteria
Dickeya zeae Ech1591	Proteobacteria
Dictyoglomus thermophilum H-6-12	Dictyoglomi
Dictyoglomus turgidum DSM 6724	Dictyoglomi
Dinoroseobacter shibae DFL 12	Proteobacteria
Dyadobacter fermentans DSM 18053	Bacteroidetes
Edwardsiella ictaluri 93-146	Proteobacteria
Edwardsiella tarda EIB202	Proteobacteria
Eggerthella lenta DSM 2243	Actinobacteria
Eggerthella sp YY7918	Actinobacteria
Ehrlichia canis str Jake	Proteobacteria
Ehrlichia chaffeensis str Arkansas	Proteobacteria
Ehrlichia ruminantium str Gardel	Proteobacteria
Elusimicrobium minutum Pei191	Elusimicrobia
Enterobacter cloacae SCF1	Proteobacteria
Enterobacter sp 638	Proteobacteria
Erwinia amylovora ATCC 49946	Proteobacteria
Erysipelothrix rhusiopathiae str Fujisawa	Firmicutes
Erythrobacter litoralis HTCC2594	Proteobacteria
Escherichia coli str K-12 substr MG1655	Proteobacteria
Escherichia fergusonii ATCC 35469	Proteobacteria
Ethanoligenens harbinense YUAN-3	Firmicutes
Eubacterium eligens ATCC 27750	Firmicutes
Eubacterium limosum KIST612	Firmicutes
Eubacterium rectale ATCC 33656	Firmicutes
Exiguobacterium sp AT1b	Firmicutes
Ferrimonas balearica DSM 9799	Proteobacteria
Fervidobacterium nodosum Rt17-B1	Thermotogae
Flavobacteriaceae bacterium 3519-10	Bacteroidetes
Flavobacterium johnsoniae UW101	Bacteroidetes
Flavobacterium psychrophilum JIP02/86	Bacteroidetes
Francisella novicida U112	Proteobacteria
Francisella sp TX077308	Proteobacteria
Frankia alni ACN14a	Actinobacteria

Вид и штамм бактерии	Класс
Frankia sp CcI3	Actinobacteria
Frankia symbiont of Datisca glomerata	Actinobacteria
Gallibacterium anatis UMN179	Proteobacteria
Gallionella capsiferriformans ES-2	Proteobacteria
Gamma proteobacterium HdN1	Proteobacteria
Gardnerella vaginalis 409-05	Actinobacteria
Gemmatimonas aurantiaca T-27	Gemmatimonadetes
Geobacillus sp C56-T3	Firmicutes
Geobacillus thermodenitrificans NG80-2	Firmicutes
Geobacillus thermoglucosidasius C56-YS93	Firmicutes
Geobacter bemidjiensis Bem	Proteobacteria
Geobacter lovleyi SZ	Proteobacteria
Geobacter sp FRC-32	Proteobacteria
Geobacter sp M18	Proteobacteria
Geobacter uraniireducens Rf4	Proteobacteria
Geodermatophilus obscurus DSM 43160	Actinobacteria
Glaciecola nitratireducens FR1064	Proteobacteria
Glaciecola sp 4H-3-7+YE-5	Proteobacteria
Gloeobacter violaceus PCC 7421	Cyanobacteria
Gluconacetobacter diazotrophicus PAl 5	Proteobacteria
Gordonia bronchialis DSM 43247	Actinobacteria
Gramella forsetii KT0803	Bacteroidetes
Granulibacter bethesdensis CGDNIH1	Proteobacteria
Granulicella tundricola	Acidobacteria
Haemophilus ducreyi 35000HP	Proteobacteria
Haemophilus parasuis SH0165	Proteobacteria
Hahella chejuensis KCTC 2396	Proteobacteria
Halanaerobium hydrogeniformans	Firmicutes
Haliangium ochraceum DSM 14365	Proteobacteria
Halobacillus halophilus DSM 2266	Firmicutes
Halomonas elongata DSM 2581	Proteobacteria
Halorhodospira halophila SL1	Proteobacteria
Halothermothrix orenii H 168	Firmicutes
Halothiobacillus neapolitanus c2	Proteobacteria
Helicobacter acinonychis str Sheeba	Proteobacteria
Helicobacter felis ATCC 49179	Proteobacteria
Helicobacter hepaticus ATCC 51449	Proteobacteria
Helicobacter mustelae 12198	Proteobacteria
Heliobacterium modesticaldum Ice1	Firmicutes
Herbaspirillum seropedicae SmR1	Proteobacteria
Herminiimonas arsenicoxydans	Proteobacteria
Herpetosiphon aurantiacus DSM 785	Chloroflexi
Hirschia baltica ATCC 49814	Proteobacteria
Hyphomicrobium sp MC1	Proteobacteria
Hyphomonas neptunium ATCC 15444	Proteobacteria
Ignavibacterium album JCM 16511	Ignavibacteria

Вид и штамм бактерии	Класс
Ilyobacter polytropus DSM 2926	Fusobacteria
Intrasporangium calvum DSM 43043	Actinobacteria
Jannaschia sp CCS1	Proteobacteria
Janthinobacterium sp Marseille	Proteobacteria
Jonesia denitrificans DSM 20603	Actinobacteria
Kangiella koreensis DSM 16069	Proteobacteria
Kitasatospora setae KM-6054	Actinobacteria
Klebsiella variicola At-22	Proteobacteria
Kocuria rhizophila DC2201	Actinobacteria
Kosmotoga olearia TBF 1951	Thermotogae
Kribbella flavida DSM 17836	Actinobacteria
Krokinobacter sp 4H-3-7-5	Bacteroidetes
Kyrpidia tusciae DSM 2912	Firmicutes
Kytococcus sedentarius DSM 20547	Actinobacteria
Lacinutrix sp 5H-3-7-4	Bacteroidetes
Lactobacillus brevis ATCC 367	Firmicutes
Lactobacillus casei ATCC 334	Firmicutes
Lactobacillus crispatus ST1	Firmicutes
Lactobacillus gasseri ATCC 33323	Firmicutes
Lactobacillus reuteri DSM 20016	Firmicutes
Lactobacillus sakei subsp sakei 23K	Firmicutes
Laribacter hongkongensis HLHK9	Proteobacteria
Leadbetterella byssophila DSM 17132	Bacteroidetes
Legionella longbeachae NSW150	Proteobacteria
Legionella pneumophila 2300/99 Alcoy	Proteobacteria
Leifsonia xyli subsp xyli str CTCB07	Actinobacteria
Leptospira biflexa serovar Patoc strain 'Patoc 1 (Ames)'	Spirochaetes
Leptospira borgpetersenii serovar Hardjo-bovis JB197	Spirochaetes
Leptospira interrogans serovar Copenhageni str Fiocruz L1-130 NC_0058231 GI:45655914	Spirochaetes
Leptothrix cholodnii SP-6	Proteobacteria
Leptotrichia buccalis C-1013-b	Fusobacteria
Leuconostoc citreum KM20	Firmicutes
Leuconostoc gasicomitatum LMG 18811	Firmicutes
Leuconostoc sp C2	Firmicutes
Listeria innocua Clip11262	Firmicutes
Listeria seeligeri serovar 1/2b str SLCC3954	Firmicutes
Listeria welshimeri serovar 6b str SLCC5334	Firmicutes
Magnetococcus marinus MC-1	Proteobacteria
Magnetospirillum magneticum AMB-1	Proteobacteria
Mannheimia succiniciproducens MBEL55E	Proteobacteria
Maribacter sp HTCC2170	Bacteroidetes
Maricaulis maris MCS10	Proteobacteria
Marinomonas sp MWYL1	Proteobacteria
Meiothermus silvanus DSM 9946	Deinococcus-Thermus
Melissococcus plutonius ATCC 35311	Firmicutes
Mesoplasma florum L1	Tenericutes

Вид и штамм бактерии	Класс
Mesorhizobium ciceri biovar biserrulae WSM1271	Proteobacteria
Mesorhizobium loti MAFF303099	Proteobacteria
Methylacidiphilum infernorum V4	Verrucomicrobia
Methylibium petroleiphilum PM1	Proteobacteria
Methylobacillus flagellatus KT	Proteobacteria
Methylobacterium chloromethanicum CM4	Proteobacteria
Methylobacterium nodulans ORS 2060	Proteobacteria
Methylocella silvestris BL2	Proteobacteria
Methylococcus capsulatus str Bath	Proteobacteria
Methylotenera mobilis JLW8	Proteobacteria
Methylotenera versatilis 301	Proteobacteria
Methylovorus glucosetrophus SIP3-4	Proteobacteria
Methylovorus sp MP688	Proteobacteria
Micavibrio aeruginosavorus ARL-13	Proteobacteria
Microbacterium testaceum StLB037	Actinobacteria
Micrococcus luteus NCTC 2665	Actinobacteria
Microcystis aeruginosa NIES-843	Cyanobacteria
Microlunatus phosphovorus NM-1	Actinobacteria
Micromonospora aurantiaca ATCC 27029	Actinobacteria
Micromonospora sp L5	Actinobacteria
Mobiluncus curtisii ATCC 43063	Actinobacteria
Moorella thermoacetica ATCC 39073	Firmicutes
Moraxella catarrhalis RH4	Proteobacteria
Mycobacterium africanum GM041182	Actinobacteria
Mycobacterium avium 104	Actinobacteria
Mycobacterium bovis AF2122/97	Actinobacteria
Mycobacterium canettii CIPT 140010059	Actinobacteria
Mycobacterium gilvum PYR-GCK	Actinobacteria
Mycobacterium leprae Br4923	Actinobacteria
Mycobacterium sp JDM601	Actinobacteria
Mycobacterium vanbaalenii PYR-1	Actinobacteria
Mycoplasma agalactiae PG2	Tenericutes
Mycoplasma arthritidis 158L3-1	Tenericutes
Mycoplasma capricolum subsp capricolum ATCC 27343	Tenericutes
Mycoplasma conjunctivae HRC/581	Tenericutes
Mycoplasma crocodyli MP145	Tenericutes
Mycoplasma fermentans JER	Tenericutes
Mycoplasma genitalium G37	Tenericutes
Mycoplasma hominis ATCC 23114	Tenericutes
Mycoplasma mobile 163K	Tenericutes
Mycoplasma penetrans HF-2	Tenericutes
Mycoplasma pulmonis UAB CTIP	Tenericutes
Mycoplasma synoviae 53	Tenericutes
Myxococcus fulvus HW-1	Proteobacteria
Myxococcus xanthus DK 1622	Proteobacteria
Nakamurella multipartita DSM 44233	Actinobacteria

Вид и штамм бактерии	Класс
Nautilia profundicola AmH	Proteobacteria
Neisseria gonorrhoeae FA 1090	Proteobacteria
Neisseria lactamica 020-06	Proteobacteria
Neisseria meningitidis 053442	Proteobacteria
Neorickettsia risticii str Illinois	Proteobacteria
Neorickettsia sennetsu str Miyayama	Proteobacteria
Nitratifractor salsuginis DSM 16511	Proteobacteria
Nitratiruptor sp SB155-2	Proteobacteria
Nitrobacter hamburgensis X14	Proteobacteria
Nitrobacter winogradskyi Nb-255	Proteobacteria
Nitrosococcus watsonii C-113	Proteobacteria
Nitrosomonas europaea ATCC 19718	Proteobacteria
Nitrosospira multiformis ATCC 25196	Proteobacteria
Nocardia farcinica IFM 10152	Actinobacteria
Nocardioides sp JS614	Actinobacteria
Nocardiopsis dassonvillei subsp dassonvillei DSM 43111	Actinobacteria
'Nostoc azollae' 0708	Cyanobacteria
Nostoc punctiforme PCC 73102	Cyanobacteria
Novosphingobium aromaticivorans DSM 12444	Proteobacteria
Novosphingobium sp PP1Y	Proteobacteria
Oceanimonas sp GK1	Proteobacteria
Oceanobacillus iheyensis HTE831	Firmicutes
Ochrobactrum anthropi ATCC 49188	Proteobacteria
Odoribacter splanchnicus DSM 20712	Bacteroidetes
Oenococcus oeni PSU-1	Firmicutes
Olsenella uli DSM 7084	Actinobacteria
Onion yellows phytoplasma OY-M	Tenericutes
Opitutus terrae PB90-1	Verrucomicrobia
Orientia tsutsugamushi str Boryong	Proteobacteria
Paenibacillus mucilaginosus 3016	Firmicutes
Paenibacillus polymyxa E681	Firmicutes
Paenibacillus sp JDR-2	Firmicutes
Paludibacter propionicigenes WB4	Bacteroidetes
Pantoea sp At-9b	Proteobacteria
Pantoea vagans C9-1	Proteobacteria
Parabacteroides distasonis ATCC 8503	Bacteroidetes
Parachlamydia acanthamoebae UV-7	Chlamydiae
Paracoccus denitrificans PD1222	Proteobacteria
Parvibaculum lavamentivorans DS-1	Proteobacteria
Parvularcula bermudensis HTCC2503	Proteobacteria
Pectobacterium atrosepticum SCRI1043	Proteobacteria
Pectobacterium carotovorum subsp carotovorum PC1	Proteobacteria
Pectobacterium wasabiae WPP163	Proteobacteria
Pediococcus pentosaceus ATCC 25745	Firmicutes
Pedobacter heparinus DSM 2366	Bacteroidetes
Pedobacter saltans DSM 12145	Bacteroidetes

Вид и штамм бактерии	Класс
Pelagibacterium halotolerans B2	Proteobacteria
Pelobacter carbinolicus DSM 2380	Proteobacteria
Pelodictyon phaeoclathratiforme BU-1	Chlorobi
Pelotomaculum thermopropionicum SI	Firmicutes
Petrotoga mobilis SJ95	Thermotogae
Phenylobacterium zucineum HLK1	Proteobacteria
Photorhabdus luminescens subsp laumondii TTO1	Proteobacteria
Pirellula staleyi DSM 6068	Planctomycetes
Planctomyces brasiliensis DSM 5305	Planctomycetes
Planctomyces limnophilus DSM 3776	Planctomycetes
Polaromonas sp JS666	Proteobacteria
Polynucleobacter necessarius subsp asymbioticus QLW-P1DMWA-1 NC_0093791 GI:145588189	Proteobacteria
Porphyromonas gingivalis ATCC 33277	Bacteroidetes
Prevotella denticola F0289	Bacteroidetes
Prevotella melaninogenica ATCC 25845	Bacteroidetes
Prevotella ruminicola 23	Bacteroidetes
Prochlorococcus marinus str AS9601	Cyanobacteria
Propionibacterium freudenreichii subsp shermanii CIRM-BIA1 NC_0142151 GI:297625198	Actinobacteria
Prosthecochloris aestuarii DSM 271	Chlorobi
Pseudoalteromonas atlantica T6c	Proteobacteria
Pseudoalteromonas haloplanktis TAC125	Proteobacteria
Pseudoalteromonas sp SM9913	Proteobacteria
Pseudogulbenkiania sp NH8B	Proteobacteria
Pseudomonas brassicacearum subsp brassicacearum NFM421	Proteobacteria
Pseudomonas entomophila L48	Proteobacteria
Pseudomonas mendocina NK-01	Proteobacteria
Pseudomonas stutzeri A1501	Proteobacteria
Pseudomonas syringae pv syringae B728a	Proteobacteria
Pseudovibrio sp FO-BEG1	Proteobacteria
Pseudoxanthomonas spadix BD-a59	Proteobacteria
Pseudoxanthomonas suwonensis 11-1	Proteobacteria
Psychrobacter arcticus 273-4	Proteobacteria
Psychromonas ingrahamii 37	Proteobacteria
Pusillimonas sp T7-7	Proteobacteria
Rahnella sp Y9602	Proteobacteria
Ralstonia eutropha H16 mega	Proteobacteria
Ramlibacter tataouinensis TTB310	Proteobacteria
Renibacterium salmoninarum ATCC 33209	Actinobacteria
Rhizobium leguminosarum bv trifolii WSM1325	Proteobacteria
Rhodobacter capsulatus SB 1003	Proteobacteria
Rhodobacter sphaeroides 241	Proteobacteria
Rhodococcus equi 103S	Actinobacteria
Rhodococcus jostii RHA1	Actinobacteria
Rhodomicrobium vannielii ATCC 17100	Proteobacteria
Rhodopirellula baltica SH 1	Planctomycetes

Вид и штамм бактерии	Класс
Rhodopseudomonas palustris BisA53	Proteobacteria
Rhodospirillum centenum SW	Proteobacteria
Rhodospirillum photometricum DSM 122	Proteobacteria
Rhodospirillum rubrum ATCC 11170	Proteobacteria
Rhodothermus marinus DSM 4252	Bacteroidetes
Rickettsia africae ESF-5	Proteobacteria
Rickettsia akari str Hartford	Proteobacteria
Rickettsia bellii OSU 85-389	Proteobacteria
Rickettsia conorii str Malish 7	Proteobacteria
Rickettsia felis URRWXCal2	Proteobacteria
Rickettsia japonica YH	Proteobacteria
Rickettsia peacockii str Rustic	Proteobacteria
Rickettsia rickettsii str 'Sheila Smith'	Proteobacteria
Robiginitalea biformata HTCC2501	Bacteroidetes
Roseburia hominis A2-183	Firmicutes
Roseiflexus castenholzii DSM 13941	Chloroflexi
Roseiflexus sp RS-1	Chloroflexi
Roseobacter denitrificans OCh 114	Proteobacteria
Rothia dentocariosa ATCC 17931	Actinobacteria
Rothia mucilaginosa DY-18	Actinobacteria
Rubrivivax gelatinosus IL144	Proteobacteria
Rubrobacter xylanophilus DSM 9941	Actinobacteria
Ruegeria pomeroyi DSS-3	Proteobacteria
Ruegeria sp TM1040	Proteobacteria
Ruminococcus albus 7	Firmicutes
Saccharomonospora viridis DSM 43017	Actinobacteria
Saccharophagus degradans 2-40	Proteobacteria
Saccharopolyspora erythraea NRRL 2338	Actinobacteria
Salinibacter ruber DSM 13855	Bacteroidetes
Salinispora arenicola CNS-205	Actinobacteria
Salinispora tropica CNB-440	Actinobacteria
Salmonella bongori NCTC 12419	Proteobacteria
Salmonella enterica subsp arizonae serovar 62:z4	Proteobacteria
Sanguibacter keddieii DSM 10542	Actinobacteria
Sebaldella termitidis ATCC 33386	Fusobacteria
Segniliparus rotundus DSM 44985	Actinobacteria
Shewanella amazonensis SB2B	Proteobacteria
Shewanella denitrificans OS217	Proteobacteria
Shewanella frigidimarina NCIMB 400	Proteobacteria
Shewanella halifaxensis HAW-EB4	Proteobacteria
Shewanella loihica PV-4	Proteobacteria
Shewanella oneidensis MR-1	Proteobacteria
Shewanella pealeana ATCC 700345	Proteobacteria
Shewanella piezotolerans WP3	Proteobacteria
Shewanella sediminis HAW-EB3	Proteobacteria
Shewanella violacea DSS12	Proteobacteria

Вид и штамм бактерии	Класс
Shewanella woodyi ATCC 51908	Proteobacteria
Sideroxydans lithotrophicus ES-1	Proteobacteria
Sinorhizobium fredii NGR234	Proteobacteria
Sinorhizobium medicae WSM419	Proteobacteria
Sinorhizobium meliloti 1021	Proteobacteria
Slackia heliotrinireducens DSM 20476	Actinobacteria
Sodalis glossinidius str 'morsitans'	Proteobacteria
Sorangium cellulosum 'So ce 56'	Proteobacteria
Sphaerobacter thermophilus DSM 20745	Chloroflexi
Sphingobium sp SYK-6	Proteobacteria
Spirochaeta smaragdinae DSM 11293	Spirochaetes
Spirochaeta thermophila DSM 6192	Spirochaetes
Spirosoma linguale DSM 74	Bacteroidetes
Stackebrandtia nassauensis DSM 44728	Actinobacteria
Staphylococcus carnosus subsp carnosus TM300	Firmicutes
Staphylococcus epidermidis ATCC 12228	Firmicutes
Staphylococcus haemolyticus JCSC1435	Firmicutes
Staphylococcus lugdunensis HKU09-01	Firmicutes
Staphylococcus saprophyticus subsp saprophyticus ATCC 15305	Firmicutes
Starkeya novella DSM 506	Proteobacteria
Stenotrophomonas maltophilia D457	Proteobacteria
Stigmatella aurantiaca DW4/3-1	Proteobacteria
Streptobacillus moniliformis DSM 12112	Fusobacteria
Streptococcus equi subsp equi 4047	Firmicutes
Streptococcus gordonii str Challis substr CH1	Firmicutes
Streptococcus mitis B6	Firmicutes
Streptococcus oralis Uo5	Firmicutes
Streptococcus parasanguinis ATCC 15912	Firmicutes
Streptococcus parauberis KCTC 11537	Firmicutes
Streptococcus pasteurianus ATCC 43144	Firmicutes
Streptococcus pneumoniae 670-6B	Firmicutes
Streptococcus sanguinis SK36	Firmicutes
Streptococcus suis 05ZYH33	Firmicutes
Streptococcus thermophilus CNRZ1066	Firmicutes
Streptococcus uberis 0140J	Firmicutes
Streptomyces avermitilis MA-4680	Actinobacteria
Streptomyces coelicolor A3(2)	Actinobacteria
Streptomyces griseus subsp griseus NBRC 13350	Actinobacteria
Streptomyces scabiei 8722	Actinobacteria
Streptosporangium roseum DSM 43021	Actinobacteria
Sulfurihydrogenibium azorense Az-Fu1	Aquificae
Sulfurihydrogenibium sp YO3AOP1	Aquificae
Sulfurimonas autotrophica DSM 16294	Proteobacteria
Sulfurimonas denitrificans DSM 1251	Proteobacteria
Sulfurospirillum deleyianum DSM 6946	Proteobacteria
Sulfurovum sp NBC37-1	Proteobacteria

Вид и штамм бактерии	Класс
Symbiobacterium thermophilum IAM 14863	Firmicutes
Synechococcus elongatus PCC 6301	Cyanobacteria
Synechococcus sp CC9311	Cyanobacteria
Syntrophobacter fumaroxidans MPOB	Proteobacteria
Syntrophobotulus glycolicus DSM 8271	Firmicutes
Syntrophomonas wolfei subsp wolfei str Goettingen	Firmicutes
Syntrophothermus lipocalidus DSM 12680	Firmicutes
Syntrophus aciditrophicus SB	Proteobacteria
Teredinibacter turnerae T7901	Proteobacteria
Terriglobus saanensis SP1PR4	Acidobacteria
Thauera sp MZ1T	Proteobacteria
Thermaerobacter marianensis DSM 12885	Firmicutes
Thermanaerovibrio acidaminovorans DSM 6589	Synergistetes
Thermincola potens JR	Firmicutes
Thermoanaerobacter brockii subsp finnii Ako-1	Firmicutes
Thermoanaerobacter italicus Ab9	Firmicutes
Thermoanaerobacterium thermosaccharolyticum DSM 571	Firmicutes
Thermoanaerobacterium xylanolyticum LX-11	Firmicutes
Thermoanaerobacter mathranii subsp mathranii str A3	Firmicutes
Thermoanaerobacter pseudethanolicus ATCC 33223	Firmicutes
Thermoanaerobacter sp X513	Firmicutes
Thermoanaerobacter tengcongensis MB4	Firmicutes
Thermobaculum terrenum ATCC BAA-798	Thermobaculum
Thermobifida fusca YX	Actinobacteria
Thermobispora bispora DSM 43833	Actinobacteria
Thermocrinis albus DSM 14484	Aquificae
Thermodesulfatator indicus DSM 15286	Thermodesulfobacteria
Thermodesulfovibrio yellowstonii DSM 11347	Nitrospirae
Thermomicrobium roseum DSM 5159	Chloroflexi
Thermomonospora curvata DSM 43183	Actinobacteria
Thermosediminibacter oceani DSM 16646	Firmicutes
Thermosipho africanus TCF52B	Thermotogae
Thermosipho melanesiensis BI429	Thermotogae
Thermosynechococcus elongatus BP-1	Cyanobacteria
Thermotoga lettingae TMO	Thermotogae
Thermotoga naphthophila RKU-10	Thermotogae
Thermotoga neapolitana DSM 4359	Thermotogae
Thermotoga petrophila RKU-1	Thermotogae
Thermotoga sp RQ2	Thermotogae
Thermus scotoductus SA-01	Deinococcus-Thermus
Thermus thermophilus HB27	Deinococcus-Thermus
Thioalkalivibrio sp K90mix	Proteobacteria
Thioalkalivibrio sulfidophilus HL-EbGr7	Proteobacteria
Thiobacillus denitrificans ATCC 25259	Proteobacteria
Thiomicrospira crunogena XCL-2	Proteobacteria
Thiomonas intermedia K12	Proteobacteria

Вид и штамм бактерии	Класс
Tolumonas auensis DSM 9187	Proteobacteria
Treponema azotonutricium ZAS-9	Spirochaetes
Treponema brennaborense DSM 12168	Spirochaetes
Treponema denticola ATCC 35405	Spirochaetes
Treponema primitia ZAS-2	Spirochaetes
Trichodesmium erythraeum IMS101	Cyanobacteria
Tropheryma whipplei TW08/27	Actinobacteria
Truepera radiovictrix DSM 17093	Deinococcus-Thermus
Tsukamurella paurometabola DSM 20162	Actinobacteria
Ureaplasma parvum serovar 3 str ATCC 27815	Tenericutes
Ureaplasma urealyticum serovar 10 str ATCC 33699	Tenericutes
Veillonella parvula DSM 2008	Firmicutes
Vibrio anguillarum 775	Proteobacteria
Vibrio cholerae IEC224	Proteobacteria
Vibrio fischeri ES114	Proteobacteria
Vibrio sp EJY3	Proteobacteria
Vibrio splendidus LGP32	Proteobacteria
Vibrio vulnificus CMCP6	Proteobacteria
Waddlia chondrophila WSU 86-1044	Chlamydiae
Wolbachia endosymbiont of Culex quinquefasciatus Pel	Proteobacteria
Wolbachia sp wRi	Proteobacteria
Wolinella succinogenes DSM 1740	Proteobacteria
Xanthobacter autotrophicus Py2	Proteobacteria
Xanthomonas albilineans GPE PC73	Proteobacteria
Xanthomonas axonopodis pv citri str 306	Proteobacteria
Xanthomonas campestris pv campestris str 8004	Proteobacteria
Xanthomonas oryzae pv oryzae KACC 10331	Proteobacteria
Xenorhabdus bovienii SS-2004	Proteobacteria
Xenorhabdus nematophila ATCC 19061	Proteobacteria
Xylanimonas cellulosilytica DSM 15894	Actinobacteria
Xylella fastidiosa 9a5c	Proteobacteria
Zunongwangia profunda SM-A87	Bacteroidetes

Приложение Б

Исследованные кластеры COG

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
carboxylic-esterase	COG0363	479	80,00%	28,41%	6-phosphogluconolactonase/Glucosamine-6-phosphate isomerase/deaminase
carboxylic-esterase	COG4677	88	59,10%	54,55%	Pectin methylesterase and related acyl-CoA thioesterases
carboxylic-esterase	COG3386	407	49,90%	67,42%	Sugar lactone lactonase YvrE
carboxylic-esterase	COG2706	179	45,80%	72,35%	6-phosphogluconolactonase, cycloisomerase 2 family
deacetylase	COG1820	296	77,40%	31,82%	N-acetylglucosamine-6-phosphate deacetylase
deacetylase	COG0726	1229	43,80%	75,38%	Peptidoglycan/xylan/chitin deacetylase, PgdA/CDA1 family
decarboxylase	COG0800	364	89,30%	16,29%	2-keto-3-deoxy-6-phosphogluconate aldolase
decarboxylase	COG0269	81	84,00%	22,35%	3-keto-L-gulonate-6-phosphate decarboxylase
decarboxylase	COG3684	55	83,60%	22,73%	Tagatose-1,6-bisphosphate aldolase
decarboxylase	COG0191	563	69,10%	40,53%	Fructose/tagatose bisphosphate aldolase
decarboxylase	COG0235	649	55,90%	59,47%	Ribulose-5-phosphate 4-epimerase/Fuculose-1-phosphate aldolase
decarboxylase	COG1850	120	54,20%	61,74%	Ribulose 1,5-bisphosphate carboxylase, large subunit, or a RuBisCO-like protein
decarboxylase	COG1830	159	52,20%	64,39%	Fructose-bisphosphate aldolase class Ia, DhnA family
decarboxylase	COG3836	174	48,90%	68,18%	2-keto-3-deoxy-L-rhamnonate aldolase RhmA
decarboxylase	COG3957	136	36,80%	79,55%	Phosphoketolase
decarboxylase	COG2140	68	32,40%	85,98%	Oxalate decarboxylase/archaeal phosphoglucose isomerase, cupin superfamily
decarboxylase	COG3961	65	32,30%	86,36%	TPP-dependent 2-oxoacid decarboxylase, includes indolepyruvate decarboxylase
decarboxylase	COG2301	354	30,50%	87,12%	Citrate lyase beta subunit
dehydratase	COG2721	194	87,10%	19,70%	Altronate dehydratase
dehydratase	COG1312	133	80,50%	27,65%	D-mannonate dehydratase
dehydratase	COG1086	423	58,90%	54,92%	NDP-sugar epimerase, includes UDP-GlcNAc-inverting 4,6-dehydratase FlaA1 and capsular polysaccharide biosynthesis protein EpsC
dehydratase	COG3866	137	55,50%	60,23%	Pectate lyase
dehydratase	COG0129	688	43,50%	75,76%	Dihydroxyacid dehydratase/phosphogluconate dehydratase
dehydratase	COG0148	516	36,40%	80,68%	Enolase
dehydrogenase-O	COG0057	781	69,30%	40,15%	Glyceraldehyde-3-phosphate

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
					dehydrogenase/erythrose-4-phosphate dehydrogenase
dehydrogenase-OH	COG3429	81	97,50%	6,44%	Glucose-6-phosphate dehydrogenase assembly protein OpcA, contains a peptidoglycan-binding domain
dehydrogenase-OH	COG0246	296	89,90%	15,53%	Mannitol-1-phosphate/altronate dehydrogenases
dehydrogenase-OH	COG0364	399	88,00%	18,56%	Glucose-6-phosphate 1-dehydrogenase
dehydrogenase-OH	COG1091	531	76,30%	32,58%	dTDP-4-dehydrorhamnose reductase
dehydrogenase-OH	COG1023	155	67,70%	41,29%	6-phosphogluconate dehydrogenase (decarboxylating)
dehydrogenase-OH	COG0362	226	53,50%	62,50%	6-phosphogluconate dehydrogenase
dehydrogenase-OH	COG0451	2954	50,80%	66,29%	Nucleoside-diphosphate-sugar epimerase or dehydrogenase
dehydrogenase-OH	COG2379	128	43,00%	76,14%	Glycerate-2-kinase
dehydrogenase-OH	COG2133	554	32,30%	86,74%	Glucose/arabinose dehydrogenase, beta-propeller fold
dehydrogenase-OH	COG4993	194	29,90%	88,64%	Glucose dehydrogenase
epimerase	COG3623	42	100,00%	1,14%	L-ribulose-5-phosphate 3-epimerase UlaE
epimerase	COG1898	506	89,30%	16,67%	dTDP-4-dehydrorhamnose 3,5-epimerase or related enzyme
epimerase	COG3010	87	86,20%	20,45%	Putative N-acetylmannosamine-6-phosphate epimerase
epimerase	COG4154	54	81,50%	26,14%	L-fucose mutarotase/ribose pyranase, RbsD/FucU family
epimerase	COG2017	419	60,60%	51,89%	Galactose mutarotase or related enzyme
epimerase	COG0676	101	51,50%	65,15%	D-hexose-6-phosphate mutarotase
epimerase	COG0036	544	28,30%	90,53%	Pentose-5-phosphate-3-epimerase
glycosidase	COG4724	17	94,10%	9,47%	Endo-beta-N-acetylglucosaminidase D
glycosidase	COG3661	59	89,80%	15,91%	Alpha-glucuronidase
glycosidase	COG1486	256	88,70%	17,80%	Alpha-galactosidase/6-phospho-beta-glucosidase, family 4 of glycosyl hydrolase
glycosidase	COG1621	202	85,60%	21,59%	Sucrose-6-phosphate hydrolase SacC, GH32 family
glycosidase	COG1874	252	82,90%	23,48%	Beta-galactosidase GanA
glycosidase	COG0383	139	79,90%	28,79%	Alpha-mannosidase
glycosidase	COG2723	548	79,00%	29,17%	Beta-glucosidase/6-phospho-beta-glucosidase/ beta-galactosidase
glycosidase	COG3867	76	78,90%	29,55%	Arabinogalactan endo-1,4-beta-galactosidase
glycosidase	COG3664	65	78,50%	29,92%	Beta-xylosidase
glycosidase	COG0296	449	78,00%	31,06%	1,4-alpha-glucan branching enzyme
glycosidase	COG2152	176	76,10%	32,95%	Predicted glycosyl hydrolase, GH43/DUF377 family
glycosidase	COG1501	282	75,50%	34,09%	Alpha-glucosidase, glycosyl hydrolase family

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
					GH31
glycosidase	COG4945	12	75,00%	34,47%	Carbohydrate-binding DOMON domain
glycosidase	COG3669	176	73,90%	35,98%	Alpha-L-fucosidase
glycosidase	COG3934	57	73,70%	36,74%	Endo-1,4-beta-mannosidase
glycosidase	COG3534	191	72,80%	37,50%	Alpha-L-arabinofuranosidase
glycosidase	COG3345	226	72,10%	38,26%	Alpha-galactosidase
glycosidase	COG1449	133	70,70%	39,02%	Alpha-amylase/alpha-mannosidase, GH57 family
glycosidase	COG3537	183	69,40%	39,77%	Putative alpha-1,2-mannosidase
glycosidase	COG3250	569	68,20%	40,91%	Beta-galactosidase/beta-glucuronidase
glycosidase	COG3507	419	66,80%	42,80%	Beta-xylosidase
glycosidase	COG3693	243	66,70%	43,18%	Endo-1,4-beta-xylanase, GH35 family
glycosidase	COG1640	272	66,50%	43,94%	4-alpha-glucanotransferase
glycosidase	COG0366	1298	65,90%	44,70%	Glycosidase
glycosidase	COG1554	183	65,00%	46,59%	Trehalose and maltose hydrolase (possible phosphorylase)
glycosidase	COG1523	383	62,90%	48,48%	Pullulanase/glycogen debranching enzyme
glycosidase	COG3408	251	62,20%	49,24%	Glycogen debranching enzyme (alpha-1,6-glucosidase)
glycosidase	COG5309	45	62,20%	49,62%	Exo-beta-1,3-glucanase, GH17 family
glycosidase	COG3459	142	62,00%	50,38%	Cellobiose phosphorylase
glycosidase	COG2730	231	59,70%	54,17%	Aryl-phospho-beta-D-glucosidase BglC, GH1 family
glycosidase	COG0058	412	58,70%	55,30%	Glucan phosphorylase
glycosidase	COG3405	87	58,60%	55,68%	Endo-1,4-beta-D-glucanase Y
glycosidase	COG2273	238	58,40%	56,06%	Beta-glucanase, GH16 family
glycosidase	COG3525	254	57,90%	57,20%	N-acetyl-beta-hexosaminidase
glycosidase	COG5297	175	56,00%	59,09%	Cellulase/cellobiase CelA1
glycosidase	COG3387	296	55,40%	60,98%	Glucoamylase (glucan-1,4-alpha-glucosidase), GH15 family
glycosidase	COG1472	954	53,00%	63,26%	Periplasmic beta-glucosidase and related glycosidases
glycosidase	COG4833	53	52,80%	64,02%	Predicted alpha-1,6-mannanase, GH76 family
glycosidase	COG2731	119	52,10%	64,77%	Beta-galactosidase, beta subunit
glycosidase	COG4124	110	50,00%	66,67%	Beta-mannanase
glycosidase	COG4692	52	50,00%	67,05%	Predicted neuraminidase (sialidase)
glycosidase	COG4409	45	48,90%	68,56%	Neuraminidase (sialidase)
glycosidase	COG1363	346	48,60%	69,70%	Putative aminopeptidase FrvX
glycosidase	COG4193	42	47,60%	70,83%	Beta- N-acetylglucosaminidase
glycosidase	COG4678	19	47,40%	71,21%	Muramidase (phage lambda lysozyme)
glycosidase	COG3325	158	46,80%	71,59%	Chitinase, GH18 family

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
glycosidase	COG1626	64	39,10%	78,41%	Neutral trehalase
glycosidase	COG3469	63	34,90%	82,95%	Chitinase
glycosidase	COG4305	29	34,50%	83,71%	Peptidoglycan-binding domain, expansin
glycosidase	COG2951	424	26,40%	91,67%	Membrane-bound lytic murein transglycosylase B
glycosyltransferase	COG3754	86	87,20%	19,32%	Lipopolysaccharide biosynthesis protein
glycosyltransferase	COG0380	183	76,00%	33,71%	Trehalose-6-phosphate synthase
glycosyltransferase	COG1216	1472	73,20%	37,12%	Glycosyltransferase, GT2 family
glycosyltransferase	COG0438	6587	66,60%	43,56%	Glycosyltransferase involved in cell wall bisynthesis
glycosyltransferase	COG0297	316	66,10%	44,32%	Glycogen synthase
glycosyltransferase	COG1442	157	65,60%	45,08%	Lipopolysaccharide biosynthesis protein, LPS:glycosyltransferase
glycosyltransferase	COG0463	3748	54,20%	62,12%	Glycosyltransferase involved in cell wall bisynthesis
glycosyltransferase	COG1215	1502	48,90%	68,94%	Glycosyltransferase, catalytic subunit of cellulose synthase and poly-beta-1,6-N-acetylglucosamine synthase
glycosyltransferase	COG1819	384	40,40%	77,27%	UDP:flavonoid glycosyltransferase YjiC, YdhE family
glycosyltransferase	COG2236	144	22,20%	95,08%	Hypoxanthine phosphoribosyltransferase
isomerase	COG4806	43	97,70%	6,06%	L-rhamnose isomerase
isomerase	COG2160	97	94,80%	8,33%	L-arabinose isomerase
isomerase	COG1904	138	94,20%	9,09%	Glucuronate isomerase
isomerase	COG3718	96	92,70%	12,12%	5-deoxy-D-glucuronate isomerase
isomerase	COG2407	83	91,60%	13,64%	L-fucose isomerase or related protein
isomerase	COG2115	146	91,10%	14,77%	Xylose isomerase
isomerase	COG3717	82	87,80%	18,94%	5-keto 4-deoxyuronate isomerase
isomerase	COG4130	14	85,70%	21,21%	Predicted sugar epimerase, xylose isomerase-like family
isomerase	COG2942	139	78,40%	30,30%	Mannose or cellobiose epimerase, N-acyl-D-glucosamine 2-epimerase family
isomerase	COG0149	528	64,60%	46,97%	Triosephosphate isomerase
isomerase	COG1082	1103	63,60%	48,11%	Sugar phosphate isomerase/epimerase
isomerase	COG1482	264	61,40%	50,76%	Mannose-6-phosphate isomerase, class I
isomerase	COG0836	472	56,10%	58,33%	Mannose-1-phosphate guanylyltransferase
isomerase	COG3622	157	56,10%	58,71%	Hydroxypyruvate isomerase
isomerase	COG0166	505	48,90%	69,32%	Glucose-6-phosphate isomerase
isomerase	COG0033	169	45,00%	73,11%	Phosphoglucomutase
isomerase	COG0698	454	44,30%	73,86%	Ribose 5-phosphate isomerase RpiB
isomerase	COG0588	261	44,10%	74,24%	Phosphoglycerate mutase (BPG-dependent)

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
isomerase	COG0279	327	40,40%	77,65%	Phosphoheptose isomerase
isomerase	COG0696	295	40,00%	78,03%	Phosphoglycerate mutase (BPG-independent, AlkP superfamily)
isomerase	COG0120	245	30,20%	87,50%	Ribose 5-phosphate isomerase
isomerase	COG0662	575	30,10%	88,26%	Mannose-6-phosphate isomerase, cupin superfamily
isomerase	COG3635	74	29,70%	89,02%	2,3-bisphosphoglycerate-independent phosphoglycerate mutase, archeal type
isomerase	COG1015	191	25,10%	92,80%	Phosphopentomutase
kinase	COG4809	3	100,00%	1,14%	Archaeal ADP-dependent phosphofructokinase/glucokinase
kinase	COG3734	51	98,00%	4,92%	2-keto-3-deoxy-galactonokinase
kinase	COG4573	36	94,40%	8,71%	Tagatose-1,6-bisphosphate aldolase non-catalytic subunit AgaZ/GatZ
kinase	COG2376	344	92,70%	12,12%	Dihydroxyacetone kinase
kinase	COG3265	141	82,30%	24,62%	Gluconate kinase
kinase	COG0126	498	81,50%	26,14%	3-phosphoglycerate kinase
kinase	COG1105	318	81,10%	27,27%	Fructose-1-phosphate kinase or kinase (PfkB)
kinase	COG1070	659	78,10%	30,68%	Sugar (pentulose or hexulose) kinase
kinase	COG0153	229	74,20%	34,85%	Galactokinase
kinase	COG2971	187	73,80%	36,36%	BadF-type ATPase, related to human N-acetylglucosamine kinase
kinase	COG0837	161	67,10%	42,05%	Glucokinase
kinase	COG1940	1438	65,10%	46,21%	Sugar kinase of the NBD/HSP70 family, may contain an N-terminal HTH domain
kinase	COG0524	1515	61,40%	51,14%	Sugar or nucleoside kinase, ribokinase family
kinase	COG5026	10	60,00%	53,03%	Hexokinase
kinase	COG3892	48	58,30%	56,44%	Myo-inositol catabolism protein IolC
kinase	COG1929	249	55,40%	60,98%	Glycerate kinase
kinase	COG0469	544	51,10%	65,53%	Pyruvate kinase
kinase	COG0205	563	49,20%	67,80%	6-phosphofructokinase
kinase	COG2074	9	33,30%	84,09%	2-phosphoglycerate kinase
kinase	COG3001	136	33,10%	84,85%	Fructosamine-3-kinase
kinase	COG0574	643	29,70%	89,39%	Phosphoenolpyruvate synthase/pyruvate phosphate dikinase
kinase	COG0406	1154	26,60%	90,91%	Broad specificity phosphatase PhoE
kinase	COG0061	523	26,00%	92,05%	NAD kinase
kinase	COG0063	434	22,60%	94,32%	NAD(P)H-hydrate repair enzyme Nnr, NAD(P)H-hydrate dehydratase domain
malto-oligosyltrehalose-synthase	COG3280	100	92,00%	13,26%	Maltooligosyltrehalose synthase

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
mutase	COG3281	95	93,70%	10,23%	Predicted trehalose synthase
mutase	COG1109	1134	36,70%	80,68%	Phosphomannomutase
mutase	COG2513	273	29,30%	89,77%	2-Methylisocitrate lyase and related enzymes, PEP mutase family
nucleosidase	COG0775	597	22,30%	94,70%	Nucleoside phosphorylase
nucleosidase	COG1957	1	0,00%	99,24%	Inosine-uridine nucleoside N-ribohydrolase
nucleotidyltransferase	COG4468	52	90,40%	15,15%	Galactose-1-phosphate uridylyltransferase
nucleotidyltransferase	COG0448	348	82,80%	23,86%	ADP-glucose pyrophosphorylase
nucleotidyltransferase	COG1080	415	74,20%	35,23%	Phosphoenolpyruvate-protein kinase (PTS system EI component in bacteria)
nucleotidyltransferase	COG1209	570	74,00%	35,61%	dTDP-glucose pyrophosphorylase
nucleotidyltransferase	COG2148	960	65,60%	45,45%	Sugar transferase involved in LPS biosynthesis (colanic, teichoic acid)
nucleotidyltransferase	COG1213	104	60,60%	52,27%	Choline kinase
nucleotidyltransferase	COG1208	478	52,90%	63,64%	NDP-sugar pyrophosphorylase, includes eIF-2Bgamma, eIF-2Bepsilon, and LPS biosynthesis proteins
nucleotidyltransferase	COG1210	523	48,00%	70,45%	UTP-glucose-1-phosphate uridylyltransferase
nucleotidyltransferase	COG4284	50	44,00%	75,00%	UDP-N-acetylglucosamine pyrophosphorylase
phosphotase	COG1877	150	82,00%	25,00%	Trehalose-6-phosphatase
phosphotase	COG1494	222	44,10%	74,62%	Fructose-1,6-bisphosphatase/sedoheptulose 1,7-bisphosphatase or related protein
phosphotase	COG0647	343	42,30%	76,89%	Ribonucleotide monophosphatase NagD, HAD superfamily
phosphotase	COG3855	47	38,30%	78,79%	Spore germination protein YaaH
phosphotase	COG0158	214	36,40%	80,68%	Fructose-1,6-bisphosphatase
phosphotase	COG1980	23	34,80%	83,33%	Archaeal fructose 1,6-bisphosphatase
phosphotase	COG0483	737	28,90%	90,15%	Archaeal fructose-1,6-bisphosphatase or related enzyme of inositol monophosphatase family
phosphotase	COG1778	300	16,30%	96,97%	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase KdsC and related HAD superfamily phosphatases
transaldolase-transketolase	COG3959	206	91,30%	14,02%	Transketolase, N-terminal subunit
transaldolase-transketolase	COG3958	237	84,00%	22,35%	Transketolase, C-terminal subunit
transaldolase-transketolase	COG0021	504	67,50%	41,67%	Transketolase
transaldolase-	COG0176	567	58,00%	56,82%	Transaldolase

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
transketolase					
transcriptional	COG3933	71	77,50%	31,44%	Transcriptional regulatory protein LevR, contains PRD, AAA+ and EIIA domains
transcriptional	COG3711	458	71,60%	38,64%	Transcriptional antiterminator
transcriptional	COG1349	841	65,30%	45,83%	DNA-binding transcriptional regulator of sugar metabolism, DeoR/GlpR family
transcriptional	COG1609	3473	64,00%	47,35%	DNA-binding transcriptional regulator, LacI/PurR family
transcriptional	COG2390	301	62,10%	50,00%	DNA-binding transcriptional regulator LsrR, DeoR family
transcriptional	COG1737	783	56,60%	57,95%	DNA-binding transcriptional regulator, MurR/RpiR family, contains HTH and SIS domains
transcriptional	COG3449	122	53,30%	62,88%	DNA gyrase inhibitor GyrI
transcriptional	COG2188	1279	48,10%	70,08%	DNA-binding transcriptional regulator, GntR family
transcriptional	COG4977	1355	42,40%	76,52%	Transcriptional regulator GlxA family, contains an amidase domain and an AraC-type DNA-binding HTH domain
transcriptional	COG2207	4349	36,30%	81,82%	AraC-type DNA-binding domain and AraC-containing proteins
transcriptional	COG1414	1475	33,00%	85,23%	DNA-binding transcriptional regulator, IclR family
transcriptional	COG1221	77	32,50%	85,61%	Transcriptional regulators containing an AAA-type ATPase domain and a DNA-binding domain
transcriptional	COG1476	863	26,40%	91,67%	DNA-binding transcriptional regulator, XRE-family HTH domain
transcriptional	COG1396	1300	25,70%	92,42%	Transcriptional regulator, contains XRE-family HTH domain
transcriptional	COG3708	221	24,00%	93,56%	Predicted transcriptional regulator YdeE, contains AraC-type DNA-binding domain
transcriptional	COG3829	894	23,70%	93,94%	Transcriptional regulator containing PAS, AAA-type ATPase, and DNA-binding Fis domains
transcriptional	COG0745	7649	22,00%	95,45%	DNA-binding response regulator, OmpR family, contains REC and winged-helix (wHTH) domain
transcriptional	COG1555	519	20,00%	95,45%	DNA uptake protein ComE and related DNA-binding proteins
transcriptional	COG1974	710	19,40%	96,21%	SOS-response transcriptional repressor LexA (RecA-mediated autopeptidase)
transcriptional	COG2771	254	8,30%	97,35%	DNA-binding transcriptional regulator, CsgD family
transcriptional	COG3706	1986	8,20%	97,73%	Two-component response regulator, PleD family, consists of two REC domains and a diguanylate cyclase (GGDEF) domain
transcriptional	COG2524	204	6,90%	98,11%	Predicted transcriptional regulator, contains C-terminal CBS domains

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
transcriptional	COG2197	4929	6,50%	98,48%	DNA-binding response regulator, NarL/FixJ family, contains REC and HTH domains
transcriptional	COG2204	5023	5,40%	98,86%	DNA-binding transcriptional response regulator, NtrC family, contains REC, AAA-type ATPase, and a Fis-type DNA-binding domains
transport	COG3730	27	100,00%	1,14%	Phosphotransferase system sorbitol-specific component IIC
transport	COG3732	28	100,00%	1,14%	Phosphotransferase system sorbitol-specific component IIBC
transport	COG3833	175	100,00%	1,14%	ABC-type maltose transport system, permease component
transport	COG4214	154	100,00%	1,14%	ABC-type xylose transport system, permease component
transport	COG3444	184	99,50%	2,27%	Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IIB
transport	COG3716	185	99,50%	2,27%	Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IID
transport	COG1869	104	99,00%	3,03%	D-ribose pyranose/furanose isomerase RbsD
transport	COG1175	2004	98,90%	3,41%	ABC-type sugar transport system, permease component
transport	COG0395	2077	98,80%	3,79%	ABC-type glycerol-3-phosphate transport system, permease component
transport	COG4209	243	98,80%	3,79%	ABC-type polysaccharide transport system, permease component
transport	COG3775	62	98,40%	4,55%	Phosphotransferase system, galactitol-specific IIC component
transport	COG1172	859	97,90%	5,30%	Ribose/xylose/arabinose/galactoside ABC-type transport system, permease component
transport	COG3715	174	97,70%	6,06%	Phosphotransferase system, mannose/fructose/N-acetylgalactosamine-specific component IIC
transport	COG4211	40	97,50%	6,82%	ABC-type glucose/galactose transport system, permease component
transport	COG1129	950	97,30%	7,20%	ABC-type sugar transport system, ATPase component
transport	COG3731	35	97,10%	7,58%	Phosphotransferase system sorbitol-specific component IIA
transport	COG1447	180	96,10%	7,95%	Phosphotransferase system cellobiose-specific component IIA
transport	COG2182	224	93,80%	9,85%	Maltose-binding periplasmic protein MalE
transport	COG1653	2280	93,20%	10,61%	ABC-type glycerol-3-phosphate transport system, periplasmic component
transport	COG4580	87	93,10%	10,98%	Maltoporin (phage lambda and maltose receptor)
transport	COG4213	187	93,00%	11,36%	ABC-type xylose transport system, periplasmic component
transport	COG4668	115	93,00%	11,74%	Mannitol/fructose-specific phosphotransferase system, IIA domain

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
transport	COG1134	329	92,70%	12,12%	ABC-type polysaccharide/polyol phosphate transport system, ATPase component
transport	COG3414	172	91,30%	14,02%	Phosphotransferase system, galactitol-specific IIB component
transport	COG1440	187	88,80%	17,05%	Phosphotransferase system cellobiose-specific component IIB
transport	COG1455	256	88,70%	17,80%	Phosphotransferase system cellobiose-specific component IIC
transport	COG4158	26	88,50%	18,18%	Predicted ABC-type sugar transport system, permease component
transport	COG2893	263	86,70%	20,08%	Phosphotransferase system, mannose/fructose-specific component IIA
transport	COG1879	1039	86,00%	20,83%	ABC-type sugar transport system, periplasmic component, contains N-terminal xre family HTH domain
transport	COG3839	834	82,60%	24,24%	ABC-type sugar transport system, ATPase component
transport	COG2213	91	81,30%	26,89%	Phosphotransferase system, mannitol-specific IIBC component
transport	COG2190	269	80,30%	28,03%	Phosphotransferase system IIA component
transport	COG1593	535	77,20%	32,20%	TRAP-type C4-dicarboxylate transport system, large permease component
transport	COG1925	528	76,10%	33,33%	Phosphotransferase system, HPr and related phosphotransfer proteins
transport	COG2610	369	72,60%	37,88%	H+/gluconate symporter or related permease
transport	COG1638	629	70,60%	39,39%	TRAP-type C4-dicarboxylate transport system, periplasmic component
transport	COG1762	615	67,00%	42,42%	Phosphotransferase system mannitol/fructose-specific IIA domain (Ntr-type)
transport	COG1264	175	64,00%	47,73%	Phosphotransferase system IIB components
transport	COG1445	153	62,70%	48,86%	Phosphotransferase system fructose-specific component IIB
transport	COG0738	471	61,10%	51,52%	Fucose permease
transport	COG1263	600	60,20%	52,65%	Phosphotransferase system IIC components, glucose/maltose/N-acetylglucosamine-specific
transport	COG1682	628	60,00%	53,41%	ABC-type polysaccharide/polyol phosphate export permease
transport	COG4975	45	60,00%	53,79%	Glucose uptake protein GlcU
transport	COG2211	682	56,70%	57,58%	Na+/melibiose symporter or related transporter
transport	COG1299	257	55,60%	59,85%	Phosphotransferase system, fructose-specific IIC component
transport	COG3090	603	50,90%	65,91%	TRAP-type C4-dicarboxylate transport system, small permease component
transport	COG2271	1386	44,40%	73,48%	Sugar phosphate permease
transport	COG0697	660	35,80%	82,20%	Permease of the drug/metabolite transporter

Функциональный класс	COG	Кол-во генов	Склонность к формированию кассет	Процентиль	Аннотация COG (NCBI)
					(DMT) superfamily
transport	COG2814	6827	35,60%	82,58%	Predicted arabinose efflux permease, MFS family
transport	COG0477	3	33,30%	84,47%	MFS family permease
transport	COG4608	680	18,70%	96,59%	ABC-type oligopeptide transport system, ATPase component
transport	COG1548	15	0,00%	99,24%	Uncharacterized protein, hydantoinase/oxoprolinase family
uncertain-class	COG3936	12	83,30%	23,11%	Membrane-bound acyltransferase YfiQ, involved in biofilm formation
uncertain-class	COG4813	38	81,60%	25,38%	Trehalose utilization protein
uncertain-class	COG5039	27	81,50%	26,14%	Exopolysaccharide biosynthesis protein EpsI, predicted pyruvyl transferase
uncertain-class	COG4354	20	55,00%	61,36%	Uncharacterized protein, contains GBA2_N and DUF608 domains
uncertain-class	COG4421	47	46,80%	71,97%	Capsular polysaccharide biosynthesis protein
uncertain-class	COG3594	99	45,50%	72,73%	Fucose 4-O-acetylase or related acetyltransferase
uncertain-class	COG0702	900	37,60%	79,17%	Uncharacterized conserved protein YbjT, contains NAD(P)-binding and DUF2867 domains
uncertain-class	COG4282	19	36,80%	79,92%	Cell wall assembly regulator SMI1
uncertain-class	COG4632	140	36,40%	81,44%	Exopolysaccharide biosynthesis protein related to N-acetylglucosamine-1-phosphodiester alpha-N-acety,,,
uncertain-class	COG3858	242	30,20%	87,88%	Spore germination protein YaaH
uncertain-class	COG4101	37	24,30%	93,18%	Uncharacterized protein, RmlC-like cupin domain

Приложение В

Кластеры COG разных функциональных классов, наиболее часто встречающиеся друг с другом. Жирным отмечены пары, преодолевшие порог статистической значимости (см. Методы Главы 3)

Функциональный класс	**carboxylic esterase**	**deacetylase**	**decarboxylase**	**dehydratase**	**dehydrogenase-O**	**dehydrogenase-OH**
carboxylic esterase	COG0363-COG0363		COG0363-COG0800	**COG0363-COG0129**		COG0363-COG0364
	COG2706-COG2706		COG3386-COG0800	COG3386-COG0129		COG0363-COG3429
	COG4677-COG4677					
	COG3386-COG3386					
	COG2706-COG4677					
deacetylase		COG0726-COG0726	COG0363-COG0800	**COG0363-COG0129**		COG0363-COG0364
			COG1820-COG0191	COG1820-COG2721		COG0726-COG0451
decarboxylase			COG0235-COG1850	COG0800-COG0129	COG0191-COG0057	COG0800-COG0364
			COG0269-COG0235	COG0800-COG1312	COG3957-COG0057	**COG0235-COG0451**
					COG1850-COG0057	
					COG0800-COG0057	
dehydratase				**COG2721-COG2721**	**COG0148-COG0057**	COG1086-COG0451
				COG3866-COG3866	**COG0129-COG0057**	COG0129-COG0364
				COG1086-COG1086		COG1312-COG0246
						COG2721-COG0246
dehydrogenase-O						COG0057-COG0364
						COG0057-COG3429
						COG0057-COG0451
dehydrogenase-OH						**COG0451-COG0451**

Функциональный класс	**glycosidase**	**glycosyltransferase**	**isomerase**	**kinase**	**Malto-oligosyltrehalose synthase**
glycosidase	COG0296-COG1523	COG0296-COG0297	COG2723-COG0698	COG1621-COG0524	COG0296-COG3280
	COG0366-COG0366	COG0058-COG0297	COG1501-COG2942	COG2723-COG1940	COG1523-COG3280
	COG0366-COG0296			COG2731-COG1940	
	COG0296-COG0058			COG0383-COG1940	
	COG1640-COG0058				
	COG0296-COG1640				
	COG1523-COG0058				
glycosyltransferase		COG0438-COG0438	COG0438-COG0836	**COG0438-COG0524**	COG0297-COG3280
		COG0463-COG0438	COG1216-COG0836	**COG0438-COG1940**	**COG0438-COG3280**
		COG0438-COG0463		COG0438-COG0406	
		COG1216-COG0438		COG0297-COG3265	
		COG0438-COG1216		COG0463-COG1940	
isomerase			**COG0149-COG0696**	**COG0149-COG0126**	
			COG1082-COG3718	COG2115-COG1070	
			COG0662-COG0836		
			COG0149-COG0698		
			COG1082-COG1082		
			COG0698-COG0698		
			COG1082-COG0698		
kinase				**COG2376-COG2376**	
				COG0469-COG0205	
malto-oligosyltrehalose synthase					COG3280-COG3280

Функциональный класс	epimerase	glycosidase	glycosyltransferase	isomerase	kinase
carboxylic esterase	COG3386-COG0451	COG2706-COG3387	**COG0363-COG0438**	COG0363-COG0176	COG0363-COG1940
	COG3386-COG2017	COG0363-COG3387	COG0363-COG1442	COG0363-COG0166	COG0363-COG0837
	COG0363-COG3010	**COG0363-COG2723**		COG0363-COG0149	
	COG0363-COG0036	**COG0363-COG0366**			
deacetylase	COG0726-COG0451	COG0363-COG2723	COG0726-COG0438	COG0363-COG0176	COG1820-COG1940
	COG1820-COG3010	COG0363-COG3387	COG0726-COG0463	COG0363-COG0166	COG0363-COG1940
		COG1820-COG3525	COG0726-COG1215	COG0363-COG0149	COG0363-COG0837
		COG0726-COG3693			
		COG0726-COG0366			
decarboxylase	COG0235-COG3623	COG3684-COG2723	COG0269-COG0438	COG0191-COG0126	COG0235-COG1070
	COG0235-COG4154	COG3684-COG1501	**COG0191-COG0438**	COG3684-COG0698	COG0191-COG0126
	COG0269-COG3623	COG0269-COG2731	COG0191-COG0463		COG0800-COG0524
	COG0269-COG0235	COG0235-COG2731	COG0800-COG1442		COG0800-COG3734
		COG0235-COG3534			
		COG0235-COG1363			
		COG0191-COG4833			
dehydratase	COG1086-COG0451	COG0148-COG2951	COG1086-COG0438	COG0148-COG0149	COG0148-COG0126
	COG1086-COG1898	COG1312-COG1472	COG1086-COG0463	COG0148-COG0126	COG0129-COG0837
				COG0148-COG0696	COG0129-COG3265
				COG2721-COG1904	
dehydrogenase-O	COG0057-COG0676			**COG0057-COG0126**	**COG0057-COG0126**
	COG0057-COG2017			COG0057-COG0149	COG0057-COG0469
dehydrogenase-OH	COG1091-COG1898	COG0364-COG3387	**COG0451-COG0438**	COG0451-COG0836	COG0246-COG0524
		COG0364-COG0366	COG0451-COG0463	COG0451-COG0279	COG0246-COG1070
		COG0364-COG1640		COG0364-COG0176	COG0364-COG0837
		COG0451-COG2951		COG0246-COG1904	COG0364-COG0469
		COG0451-COG0296			COG1023-COG1070
		COG1023-COG3387			COG2379-COG0469
		COG0246-COG3250			COG0362-COG1070
epimerase	COG1898-COG1898	**COG3010-COG2731**	**COG0451-COG0438**	COG0451-COG0836	COG0235-COG1070
	COG0036-COG0036	COG2942-COG1501	COG0451-COG0463	COG0451-COG0279	**COG2017-COG0153**
	COG1898-COG0676	COG2017-COG1501		COG0235-COG3622	
	COG2017-COG2017	COG3623-COG2731		COG0235-COG2160	
				COG0235-COG2407	

Функциональный класс	mutase	nucleotydiltransferase	phosphatase	Transaldolase/transketo	transcriptional	transport
carboxylic esterase			COG0363-COG0647	**COG0363-COG0176**	COG0363-COG1737	COG0363-COG2190
			COG0363-COG1494	**COG0363-COG0021**	COG0363-COG2188	COG3386-COG1172
				COG0363-COG0166	COG3386-COG1414	COG3386-COG1879
				COG2706-COG0176		COG3386-COG1129
						COG3386-COG2271
deacetylase			COG1820-COG0647	COG0363-COG0176	COG1820-COG2188	**COG1820-COG2190**
			COG0363-COG0647	**COG0363-COG0021**	COG0363-COG1737	COG1820-COG3444
			COG0726-COG1494	COG0363-COG0166		COG1820-COG3716
				COG4193-COG0176		COG1820-COG3715
						COG0363-COG2190
						COG0726-COG1682
decarboxylase			COG0191-COG1494	COG0191-COG0021	COG0235-COG1349	COG0235-COG1762
			COG0191-COG0158	COG0191-COG0176	COG0235-COG2207	COG0235-COG3414
			COG0191-COG0647		**COG0235-COG1609**	COG0191-COG1762
			COG1850-COG0158		COG0191-COG1349	COG0191-COG3414
			COG0235-COG0647		COG0800-COG1414	COG1830-COG1172
					COG3684-COG1349	COG0269-COG1762
						COG0269-COG3414
						COG3836-COG2271
dehydratase		**COG1086-COG1210**		COG0129-COG3959	COG0148-COG2390	**COG2721-COG2271**
		COG0129-COG0448		COG0129-COG0166	COG0129-COG1737	COG0129-COG2814
				COG0129-COG3958		
				COG0148-COG0166		
				COG2721-COG0176		
dehydrogenase-O				COG0057-COG0021	COG0057-COG2390	COG0057-COG2814
				COG0057-COG0176	COG0057-COG0745	COG0057-COG2271
dehydrogenase-OH	**COG0451-COG1109**	COG1091-COG1209	COG0364-COG0158	COG0364-COG0176	COG0364-COG1737	**COG0246-COG4668**
	COG0362-COG1109	COG0451-COG1208	COG0362-COG1494	COG0364-COG0021	COG0451-COG2207	**COG0246-COG2213**
			COG3429-COG0158	COG3429-COG0176	COG0451-COG0745	COG0246-COG2271
			COG0451-COG1494		COG0451-COG1555	COG0451-COG2814
			COG0451-COG0483		COG0246-COG3711	COG0451-COG2271
			COG4993-COG1877		COG0246-COG2390	
					COG0246-COG1609	
epimerase	**COG0451-COG1109**	**COG1898-COG1209**	COG0036-COG1494	COG0036-COG0021	COG0235-COG1349	COG0235-COG1762
	COG1898-COG1109	COG0451-COG1208	COG0036-COG0158	COG0036-COG0176	COG0235-COG2207	COG0235-COG3414
			COG0451-COG1494	COG0451-COG3959	**COG0235-COG1609**	COG0451-COG2814
			COG0451-COG0483	COG0451-COG3958	COG0235-COG3711	**COG3623-COG3414**
				COG0235-COG0021	COG0451-COG2207	
				COG0235-COG0176	COG0451-COG0745	
				COG3623-COG0021	COG3010-COG1737	
					COG2017-COG1609	
					COG4154-COG1349	

Функциональный класс	mutase	nucleotydiltransferase	phosphatase	Transaldolase/transketolase	transcriptional	transport
glycosidase	COG0366-COG3281	COG0296-COG0448	**COG3387-COG1877**	COG0296-COG0166	COG0366-COG1609	COG0366-COG1175
	COG0296-COG3281	**COG0058-COG0448**	COG0058-COG0158	COG0058-COG0166	COG1621-COG1609	**COG2723-COG1455**
	COG1523-COG3281	COG1523-COG0448		COG3387-COG0176	**COG2723-COG1609**	COG2723-COG1447
				COG3387-COG0021	COG2723-COG3711	
				COG1449-COG0166	COG3250-COG1609	
				COG1640-COG0166	COG1486-COG2207	
				COG1640-COG0176		
				COG0366-COG0166		
glycosyltransferase	**COG0438-COG1109**	**COG0297-COG0448**	**COG0380-COG1877**	COG0438-COG0166	COG0438-COG0745	COG0438-COG1134
	COG0463-COG1109	COG0438-COG0836	COG0438-COG0483	COG0297-COG0166	COG0463-COG0745	COG0438-COG1682
				COG0463-COG3958	COG1215-COG0745	COG0438-COG2814
				COG0463-COG3959		COG1216-COG1134
				COG0463-COG0021		COG1216-COG1682
						COG0463-COG1134
						COG0463-COG1682
						COG0463-COG2814
isomerase	COG0836-COG1109	COG0836-COG1209	COG0149-COG1494	COG0176-COG0021	COG1082-COG1609	COG1082-COG1879
	COG0149-COG1109	COG0279-COG1208	COG0176-COG1494	COG0176-COG0166	COG0149-COG2390	COG1082-COG1172
		COG0166-COG1210	**COG1082-COG0483**	COG0126-COG0021		COG1082-COG1129
		COG0166-COG0448		COG0166-COG0176		COG1082-COG0395
		COG0033-COG0448				COG1082-COG1175
		COG1482-COG0836				COG1082-COG1653
						COG1082-COG2814
						COG0698-COG1447
kinase	COG2971-COG1109	**COG0153-COG4468**	COG1940-COG0647	COG0126-COG0021	COG0524-COG1609	COG0524-COG1879
	COG0063-COG1109	COG3265-COG0448	COG1940-COG0483	COG0469-COG0021	COG1105-COG1349	COG0524-COG1129
	COG0524-COG1109		**COG0061-COG0647**			COG0524-COG1172
	COG1940-COG1109		COG1070-COG0647			COG1070-COG1172
	COG0153-COG1109					
mutase	COG1109-COG1109	COG1109-COG0836		COG1109-COG0166	**COG2513-COG3829**	COG2513-COG2814
	COG2513-COG2513	COG1109-COG1210		COG1109-COG3958	COG1109-COG0745	**COG1109-COG2814**
						COG1109-COG1762
nucleosidase					COG0775-COG0745	COG0775-COG2814
					COG0775-COG1609	COG0775-COG2271
nucleotydiltransferase		**COG0448-COG0448**		COG1210-COG0166	COG4468-COG1609	COG1209-COG1134
				COG0448-COG0166	COG0662-COG2207	COG1209-COG1682
				COG0836-COG0176	**COG0448-COG1609**	COG0836-COG1682
phosphatase				COG1494-COG0176	COG0483-COG3711	COG0158-COG1172
				COG0158-COG0176	COG0483-COG1609	COG0483-COG2814
				COG0158-COG0021	COG0483-COG0745	COG0647-COG2190
					COG0483-COG1396	
					COG1877-COG1609	
transaldolase/transketolase				COG3959-COG3958	COG0176-COG3711	COG0176-COG1762
					COG0176-COG1609	COG0176-COG2814
					COG0176-COG2390	COG0176-COG1445
					COG0021-COG3711	COG0176-COG1299
					COG0021-COG1974	COG0021-COG1762
transcriptional					COG0745-COG0745	COG1609-COG1175
					COG2197-COG2197	COG1609-COG1653
					COG2204-COG0745	COG1609-COG0395
					COG2204-COG2204	COG1609-COG1129
transport						COG1653_COG0395
						COG0395_COG3839
						COG1653_COG1175
						COG1593_COG1638
						COG0395_COG1175
						COG1172_COG1879
						COG1638_COG3090
						COG1593_COG3090
						COG1879_COG1129
						COG1175_COG3839
						COG3839_COG1653
						COG1172_COG1129

Приложение Г

Использованные в работе праймеры

Праймер	**Последовательность**
yihU/V_F	CGTTCACATCAAAGACGCGA
yihU/V_R	GTCGGTAACCCTTCCACGTA
yihV/W_F	TCAACCGGCCTTCAAAGTTG
yihV/W_R	GCGATCAGCATGAGGAGTTG
yihS/R_F	AGCTGGATGCGGACAATAAG
yihS/R_R	GGCATCTCTTCGGGTTTGTG
yihW_RT	CCGTATTAACGACGCTGGAA
yihW_PCR	GCCGAGCGTGGGTATATGAA
yihV_RT	TCATCACCTACGCGACCAAT
yihV_PCR	TTCGTGTTGCTTGTGTAGGT
yihU_RT	GGAGTCGCACCTTTGTCTAC
yihU_PCR	CGCGTTTATCGGTTTAGGAC
yihT_RT	ATTGTTGATCTACCAGAATCG
yihT_PCR	CGAAGCCATGCGCATGATGT
yihS_RT	CCGTGATCAACCAACGAGTA
yihS_PCR	GGTTTTGGCTGGTTAGGCAA
hns_RT	ATTTAACGGCAGCAAGGCTATT
hns_PCR	GAAGTTGAAGAGCGCACTCG
hns_Bgl_263	AGGGAGATCTCGTAAACACAACTA
hns_Xba	GTTGTCTAGAATTTTAAGTGCTTCG
CRP_NdeI	ACCGCATATGGTGCTTGGCAAACCGCAA
CRP_Bpu1102	CCACGCTGAGCGGATTAACGAGTGCCGTA