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

На правах рукописи

Червонцева Зоя Сергеевна

Влияние вторичной структуры мРНК на экспрессию генов

1.5.8 — математическая биология, биоинформатика

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель:
доктор биологических наук, профессор
Михаил Сергеевич Гельфанд

Москва — 2023

Оглавление

Стр.

Введение

Актуальность темы исследования

Из всех типов биологических молекул рибонуклеиновые кислоты (РНК) выделяются многообразием функций, которые они способны выполнять. В дополнение к кодированию информации о последовательности белка молекулы РНК могут участвовать в регуляции транскрипции, трансляции и деградации себя и других РНК, а также могут, подобно белкам, катализировать химические реакции. Исходя из этого, РНК может по праву считаться самой универсальной составляющей живой клетки. Эта универсальность в свое время породила гипотезу РНК-мира, согласно которой первые живые системы на Земле использовали именно РНК во всех ключевых клеточных процессах [4].

Важным свойством РНК, обеспечивающим значительную часть ее функций, является способность формировать стабильные структуры за счет комплементарных взаимодействий. Большая часть исследований структуры РНК изначально была сконцентрирована на анализе некодирующих РНК или же некодирующих элементов матричных РНК, однако в последние годы появляется все больше свидетельств, что структура кодирующих областей матричных РНК также может играть роль в различных клеточных процессах. В частности, модификация, трансляция, локализация и деградация матричных РНК может зависеть от их структуры.

Появление новых экспериментальных методов полногеномного анализа структур РНК сделало возможным выявление новых общих закономерностей формирования структур в кодирующих областях РНК и изучение их функций. Вместе с тем, дороговизна и сложность этих методов пока что не позволяет полностью отказаться от чисто вычислительных подходов к предсказанию структуры РНК. В нашей работе мы использовали данные из множества доступных источников для установления роли вторичной структуры в различных процессах, происходящих с мРНК бактерии *Escherichia coli*, а также в

редактировании мРНК у головоногих моллюсков.

Цели и задачи исследования

Целями данной работы были изучение роли вторичной структуры в трансляции, деградации и эволюции мРНК у *Escherichia coli*, а также описание закономеростей, связывающих вторичную структуру мРНК мягкотелых головоногих моллюсков с частотой гидролитического дезаминирования аденинов (редактирования).

Были поставлены следующие задачи.

1. Проанализировать экспериментальные данные по экспрессии флуоресцентного белка со случайными вставками в начале гена в клетках *Escherichia coli* и выявить свойства случайных вставок, определяющие эффективность трансляции.
2. Сравнить структуры матричных РНК генов, кодирующих эквимолярные субъединицы одного белкового комплекса и закодированных в одном опероне.
3. Изучить связь стабильности вторичной структуры матричных РНК *Escherichia coli* со скоростью их деградации и паттернами их эволюции.
4. Изучить связь стабильности вторичной структуры со степенью редактирования аденинов в матричных РНК мягкотелых головоногих моллюсков.
5. Оценить вклад вторичной структуры матричных РНК в скоррелированное редактирование аденинов у мягкотелых головоногих моллюсков.

Научная новизна

Впервые исследовано влияние случайных вставок в начале гена на эффективность инициации трансляции мРНК *Escherichia coli*. Впервые исследованы структурные факторы, влияющие на согласованную трансляцию генов эквимолярных субъединиц белковых комплексов у *Escherichia coli*. Впервые исследовано влияние вторичной структуры на массовое, консервативное, скоординированное редактирование мРНК у колеоидов.

Практическая значимость

Результаты, показывающие влияние различных аспектов структуры мРНК на эффективность инициации трансляции и стабильность мРНК, могут иметь значение для оптимизации экспрессии генов, особенно в гетерологичных системах. Разработанные методы и подходы могут быть применены в медицинской генетике для оценки влияния мутаций в некодирующих областях и синонимичных мутаций. Эволюционные аспекты исследования могут использоваться как источник задач для самостоятельной работы в курсах молекулярной эволюции.

Положения, выносимые на защиту

1. Последовательность в 5'-области гена оказывает существенное влияние на эффективность трансляции *Escherichia coli*:
    (a) выраженная вторичная структура и участки, схожие с последовательностью Шайна-Дальгарно, снижают эффективность трансляции;
    (b) вопреки существующим представлениям, редкие кодоны в начале гена не оказывают влияния на эффективность трансляции;

   (c) в бедной среде кодоны аминокислот, синтез которых требует существенных энергетических затрат, уменьшают эффективность трансляции.

2. Равная эффективность трансляции генов *Escherichia coli*, кодирующих эквимолярные субъединицы белковых комплексов в составе одного оперона, обеспечивается, в том числе, близкой по стабильности структурированностью мРНК.
3. Средняя степень структурированности мРНК *Escherichia coli* не коррелирует со скоростью ее деградации.
4. Вторичная структура существенна для редактирования мРНК колеоидов:
   (a) аденины в структурированных областях редактируются чаще, чем в неструктурированных;
   (b) уровень редактирования увеличивается с уровнем стуктурированности;
   (c) указанные эффекты сильнее проявляются, если редактирование консервативно;
   (d) вторичная структура способствует коррелированному редактированию сближенных аденинов.

Степень достоверности и апробация результатов

Достоверность результатов подтверждается статистическим анализом с использованием различных критериев, поправками на множественное тестирование, применением специально разработанных процедур рандомизации. По материалам диссертации опубликовано три статьи в рецензируемых научных журналах. Результаты были доложены на конференциях «Belgrade BioInformatics Conference» в 2018 году (Белград, Сербия), «5th Meeting of Regulation with RNA in Bacteria and Archaea» в 2018 году (Севилья, Испания), «Информационные технологии и системы» в 2017 и 2018 годах (ИТиС 2017 –

Уфа, ИТиС 2018 – Казань).

Объем и структура работы

Полный объём диссертации составляет 100 страниц, включая 28 рисунков и 3 таблицы. Список литературы содержит 119 наименований.

Список публикаций по теме диссертации

По теме диссертации опубликовано три статьи в рецензируемых международных научных журналах, входящих в основные библиометрические базы данных (PubMed, WoS и Scopus):

1. Translation at first sight: the influence of leading codons / I. A. Osterman*, Z. S. Chervontseva*, S. A. Evfratov, A. V. Sorokina, V. A. Rodin, M. P. Rubtsova, E. S. Komarova, T. S. Zatsepin, M. R. Kabilov, A. A. Bogdanov, M. S. Gelfand, O. A. Dontsova, P. V. Sergiev // Nucleic Acids Research. — 2020. — Vol. 48, no. 12. — 6931 *Joint first authors.
2. Adaptive evolution at mRNA editing sites in soft-bodied cephalopods / M. Moldovan, Z. Chervontseva, G. Bazykin, M. S. Gelfand // PeerJ. — 2020. — Vol. 8: e10456.
3. A hierarchy in clusters of cephalopod mRNA editing sites / M. A. Moldovan, Z. S. Chervontseva, D. S. Nogina, M. S. Gelfand // Scientific Reports. — 2022. — Vol. 12, 1: 3447.

Кроме того, результаты работы опубликованы в сборниках тезисов международных и российских конференций:

1. The role of mRNA secondary structure in the control of translation and mRNA degradation in *E. coli* / Z. Chervontseva, E. Khodzhaeva, I. Ponamareva, A. Mironov, M. Gelfand // Proceedings of the 2nd Belgrade BioInformatics Conference (BelBi 2018), June 18 –22, 2018, Belgrade, Serbia.

2. Role of mRNA structure in the control of protein synthesis in *E. coli* / E. Khodzhaeva, M. Gelfand, A. Mironov, Z. Chervontseva // Proceedings of the 5th Meeting of Regulation with RNA in Bacteria and Archaea, March 19 – 22, 2018, Seville, Spain.

3. Редактирование мРНК головоногих моллюсков как пример преадаптации / М. Молдован, З. Червонцева, М. Гельфанд // Труды конференции молодых ученых и специалистов ИППИ РАН (ИТиС), Сентябрь 25 – 30, 2018, Казань, Россия.

Список сокращений и условных обозначений

РНК	Рибонуклеиновая кислота
мРНК	Матричная РНК
нкРНК	Некодирующая РНК
тРНК	Транспортная РНК
рРНК	Рибосомальная РНК
A, C, G, T	Нуклеотиды аденин, цитозин, гуанин и тимин
ГТФ	Гуанозинтрифосфат
ШД	Последовательность Шайна –Дальгарно
LB	Lysogeny broth, питательно богатая среда для выращивания культур бактерий
M9	Мининальная среда для выращивания культур бактерий
IRES	Internal Ribosome Entry Site, участок внутренней посадки рибосомы
SHAPE-seq	Selective 2'-hydroxyl acylation analyzed by primer extension sequencing, селективное ацилирование 2'-гидроксила РНК с последующим анализом удлинения затравки
DMS-seq	Dimethyl sulfate sequencing, секвенирование с использованием диметилсульфата
CER	Cerulean, лазурный флуоресцентный белок
RFP	Red fluorescent protein, красный флюоресцентный белок
TEF, ФЭТ	Translation efficiency fraction, фракция эффективности трансляции

Глава 1. Обзор литературы

1.1 Понятие степени структурированности РНК

Множество важных биологических функций РНК выполняют, сворачиваясь в определенные трехмерные структуры. Главными силами, стабилизирующими такие структуры, являются водородные связи между комплементарными нуклеотидами и стекинг-взаимодействия между соседними комплементарными парами [5—7], поэтому структуры РНК принято приближенно описывать в терминах того, какой участок с каким спарен (вторичная структура РНК).

Исходя из информации о спаренных участках и известных оценках энергий, выделяющихся при формировании отдельных элементов, для каждой структуры РНК можно оценить свободную энергию ее сворачивания. Известно, что в клетке чаще реализуются структуры, суммарная свободная энергия которых близка к минимальной для этой последовательности [8]. Поэтому из-за дороговизны и сложности экспериментов возможные структуры РНК часто предсказывают вычислительно, оптимизируя свободную энергию сворачивания. При этом, чтобы уменьшить пространство поиска, вводится ряд ограничений. Главным ограничением самого широко используемого алгоритма предсказания структуры РНК (алгоритм Зукера [9]) является то, что он не учитывает псевдоузлы, — взаимодействия между петлями, — которые часто встречаются в реальных структурах [10; 11].

Чтобы сравнивать между собой разные РНК по их способности образовывать стабильные структуры, в этой работе мы будем использовать понятие структурированности. Степень структурированности последовательности РНК – это мера того, насколько стабильная структура в принципе может быть сформирована этой последовательностью. В случае, если было произведено вычислительное предсказание, степень структурированности в этой работе оценивалась как предсказанная свободная энергия сворачивания последовательности. Если же происходила обработка экспериментальных данных, то степень структури-

рованности оценивалась соответственно методу получения этих данных (см. 1.4 «Экспериментальные методы определения структуры РНК»).

1.2 Свойства вторичной структуры мРНК

В отличие от некодирующих РНК (нкРНК), матричным РНК (мРНК) менее свойственно формировать стабильные структуры. Рибосомы расплетают комплементарные участки при трансляции, поэтому спаривание далеких участков последовательности для мРНК менее вероятно, чем для нкРНК, и большинство функциональных комплементарных взаимодействий происходят локально. Известно множество консервативных элементов мРНК с определенной структурой, некоторые из которых будут упомянуты далее. Однако известны и случаи, когда конкретной консервативной структуры не наблюдается, но структурированность транскрипта или определенных его участков следует некой общей закономерности. В этой главе также будут представлены некоторые из таких закономерностей. Некоторые из них эволюционно консервативны, что указывает на их возможную функциональную важность.

1.2.1 Распределение вторичной структуры мРНК вдоль транскрипта

Большая часть известных цис-регуляторных РНК-элементов прокариот (рибопереключатели, термосенсоры, аттенюаторы и др.) участвуют в регуляции транскрипции или трансляции и располагаются в 5'-нетранслируемой области гена, нередко захватывая начало кодирующей области [12—16]. Для эукариот известно значительно меньше консервативных регуляторных РНК-элементов, а те, что известны, как правило, располагаются в интронах или 5' и 3'-нетранслируемых областях и осуществляют регуляцию инициации трансляции (IRES-элементы) [17] или сплайсинга [18—21].

Систематический массовый анализ, проведенный на основе вычислительно предсказанных энергий сворачивания для участков кодирующих областей транскриптов, показал, что кодирующие участки мРНК у большинства видов всех трех надцарств (бактерий, архей, эукариот) устроены следующим образом: начало и конец кодирующей области структурированы слабее, чем можно было бы ожидать, исходя из закодированных аминокислот и общего нуклеотидного состава, а участок, находящийся на расстоянии $\sim 30 - 70$ нуклеотидов вглубь гена от старта трансляции, структурирован сильнее ожидаемого [22].

Избегание структуры вокруг старта трансляции у бактерий было независимо показано во множестве работ, посвященных как изучению эволюции природных транскриптов [23—25], так и дизайну синтетических последовательностей, обеспечивающих высокую экспрессию целевого белка [26; 27]. Стабильная вторичная структура, затрагивающая место посадки рибосомы, может затруднять инициацию трансляции и поэтому избегается.

Несмотря на то, что механизм инициации трансляции у эукариот значительно отличается от механизма инициации трансляции у бактерий и опосредован 5'-кэпом, у 36 из 78 видов эукариот, рассмотренных в исследовании [22], участок $\sim 0 - 20$ нуклеотидов вглубь гена от старта трансляции тоже структурирован слабее, чем ожидается. Это наблюдение согласуется с результатами других работ [23; 28; 29], однако функциональная роль этого избегания структурированности в начале кодирующей области на данный момент не ясна.

Как уже было упомянуто ранее, за участком пониженной структурированности часто следует участок повышенной структурированности, располагающийся, в среднем, на расстоянии $\sim 30 - 70$ нуклеотидов вглубь гена от старта трансляции. Предполагают, что этот элемент стабильной структуры служит для повышения эффективности трансляции [30—32]. Один из возможных механизмов заключается в том, что расплетание структуры замедляет продвижение недавно инициировавшей рибосомы и таким образом предотвращает ее столкновение с предыдущими рибосомами, уже осуществляющими трансляцию гена. Столкновения рибосом считаются фактором, отрицательно влияющим на

точность трансляции [33; 34]. Кроме того, для ряда транскриптов дрожжей *Saccharomyces cerevisiae*, кодирующих мембранные и секретируемые белки, было экспериментально показано, что прохождение рибосомы через такие структурированные элементы на участке $\sim 30 - 70$ нуклеотидов от старта трансляции обеспечивается активностью РНК-хеликазы Dhh1 [35]. Таким образом, этот элемент структуры мРНК может служить дополнительной контрольной точкой для принятия решения о трансляции и тем самым обеспечивать дополнительный уровень регуляции.

Что же касается пониженного уровня структурированности перед стоп-кодоном, то предполагается, что это может увеличивать точность распознавания стоп-кодона [22]. Повышает ли наличие стабильной вторичной структуры в этой области частоту ошибочного игнорирования стоп-кодона (stop codon readthrough) не известно, однако единичные наблюдения над вирусными мРНК [36] позволяют такую связь предположить.

1.2.2 Специфичность сворачивания мРНК

В отличие от некодирующих РНК, матричные РНК регулярно расплетаются рибосомами и потому не могут иметь постоянной структуры. После прохождения рибосомы затронутый участок мРНК может свернуться либо так же, как был свернут до прохождения рибосомы, либо по-новому. В случае, если элемент структуры воспроизводится после событий трансляции, можно считать его сворачивание специфичным.

Специфичность сворачивания можно оценить на основании экспериментальных полногеномных *in vivo* данных лигирования близких участков РНК (RNA proximity ligation, RPL). В работе [37] был проведен анализ специфичности сворачивания мРНК для дрожжей *Saccharomyces cerevisiae* и мыши *Mus musculus*. Авторы показали, что специфичность сворачивания не связана напрямую со стабильностью полученной структуры (ΔG), а у высоко экспрессирующихся и высоко консервативных генов специфичность сворачивания выше, чем

у слабо экспрессирующихся и низко консервативных, соответственно. Кроме того, внутри отдельных генов консервативные позиции оказываются в составе более специфично сворачивающихся участков мРНК, чем неконсервативные.

Отдельно интересно, что обсуждавшийся выше участок повышенной структурированности вокруг позиции 70 после старт-кодона оказался более высоко-специфично сворачивающимся у тех генов, трансляция которых, согласно работе [35], активируется РНК-хеликазой Dhh1.

1.3 Значение структуры мРНК для биологических процессов

1.3.1 Связь структуры и стабильности мРНК

Вторичная структура способна влиять на время жизни мРНК несколькими способами. Известно, что слабо транслирующиеся мРНК прокариот быстрее деградируют [38; 39], а наличие вторичной структуры в начале кодирующей области может затруднять инициацию трансляции, приводя таким образом к ускоренной деградации транскрипта [24]. Высокий уровень структурированности в кодирующей части гена связан с низкой трансляционной эффективностью у бактерий [40], что тоже может приводить к ускоренной деградации. Вместе с тем, у дрожжей *Saccharomyces cerevisiae*, напротив, была показана повышенная по сравнению с ожидаемой средняя структурированность мРНК кодирующих областей и связанное с этим повышение эффективности трансляции и времени жизни транскрипта [41]. Но как у прокариот, так и у эукариот, транскрипты, содержащие длинные двуцепочечные участки, могут ошибочно приниматься клеткой за вирусные РНК и уничтожаться соответствующими защитными системами [42; 43].

Структурированность отдельных участков транскрипта может повышать время его жизни. У некоторых бактерий деградация транскриптов осуществляется ферментами двух типов: эндо- и экзонуклеазами, причем экзоноклеазы расщепляют транскрипт, начиная с 3'-конца [44]. Эти экзонуклеазы обладают

слабой хеликазной активностью, поэтому стабильная структура их останавливает, или, по крайней мере, замедляет. Таким образом, стабильные шпильки, закодированные в концах транскриптов бактерий (ро-независимые терминаторы транскрипции) и в концах отдельных генов могут защищать кодирующую область гена от 3'-экзонуклеаз [45]. Вместе с тем, от эндонуклеаз, разрезающих РНК в длинных АТ-богатых одноцепочечных участках (например, RNase E), транскрипты могут быть защищены транслирующими рибосомами или вторичными структурами в кодирующей части гена [44].

Таким образом, степень и направление влияния вторичной структуры мРНК на время жизни транскрипта отличается у разных видов, и отдельные структурированные элементы потенциально могут участвовать в регуляции, результатом которой может являться изменение времени жизни транскрипта в зависимости от внешних условий.

1.3.2 Связь структуры и локализации мРНК в клетке

Локализация мРНК в клетке может определять локализацию закодированных в ней белков, а через это и их функциональность. Транспорт белков через внутренние мембраны клетки (например, в эндоплазматический ретикулум), во внешнюю среду или в периплазматическое пространство (у грамотрицательных бактерий) осуществляется у всех организмов одной или несколькими эволюционно консервативными транспортными системами, такими как SEC, SRP, TAT и другие. В большинстве случаев этот транспорт направляется сигнальными пептидами, закодированными на 5'-концах транспортируемых белков [46]. Однако, в части случаев локализация мРНК определяется связанными с ней белками, и это связывание может зависеть от структуры РНК. Так, например, человеческий белок STAU1, участвующий в транспорте мРНК к поверхности гранулярного эндоплазматического ретикулума, связывает исключительно двуцепочечные участки [47], а белок RBM15, участвующий в транспорте

ряда транскриптов из ядра в цитоплазму – напротив, связывает только одноцепочечные участки [48].

1.3.3 Как происходит расплетение структур мРНК при трансляции

Для того, чтобы декодировать белок, записанный в мРНК, рибосоме необходимо получить доступ ко всем кодонам. Это означает, что в случае, если нуклеотиды какого-нибудь кодона связаны с белками или другими нуклеотидами той же мРНК, эта связь должна быть нарушена, причем заблаговременно. Когда бактериальная рибосома декодирует кодон под номером i и собирается транслоцироваться, ей нужно расплести структуру, затрагивающую кодон под номером $i+4$, потому что как раз примерно 4 кодона (13 ± 2 нуклеотида) укладываются во входной канал рибосомы от места входа мРНК до пептидил-сайта, где происходит декодирование [49]. Считается, что у прокариот хеликазную активность осуществляет рибосомный белок S1, который может действовать как в составе малой единицы рибосомы, так и в свободной форме [50]. У *Escherichia coli* S1 необходим для трансляции большинства белок-кодирующих генов [51; 52], однако в геномах некоторых других видов бактерий не закодировано его гомологов [53], что позволяет предположить существование альтернативных механизмов расплетения структуры мРНК.

У эукариот известно несколько хеликаз, которые участвуют в расплетении структур мРНК, по большей части, в 5'-нетранслируемых областях транскрипта и влияют на инициацию трансляции (*DHX36*, *DDX5*, *DDX17*, *DDX2A*, *DDX2B* и др.). Элонгирующая же рибосома, как считается, по большей части сама расплетает мРНК за счет энергии гидролиза ГТФ одновременно с транслокацией [54].

1.3.4 Связь структуры мРНК и эффективности трансляции

Общее количество белка, полученного с одной мРНК гена, может регулироваться через множество параметров, таких как время жизни этой мРНК (ее стабильность), частота инициации трансляции, скорость элонгации трансляции. Отдельно может регулироваться качество сворачивания полученного в итоге белка, в том числе, за счет неравномерной скорости элонгации [55; 56]. Так как в этой работе нас интересовала эффективность трансляции у *Escherichia coli*, дальнейшее изложение в этом разделе будет касаться исключительно бактерий.

Известно несколько факторов, влияющих на инициацию трансляции у бактерий. Сворачивание 5'-конца мРНК может затруднять связывание рибосомы за счет уменьшения доступности самого старт-кодона или сайта инициации трансляции (последовательности Шайна–Дальгарно) [57]. В случае полицистронных транскриптов, сильная вторичная структура в начале последующих генов может затруднять ре-инициацию трансляции рибосомами, закончившими трансляцию предыдущего гена на той же мРНК [58]. Наличие или отсутствие самой последовательности Шайна–Дальгарно, а также степень ее аффинности к соответствующему участку 16S рРНК (анти-Шайна–Дальгарно) также может влиять на эффективность инициации трансляции [59].

В отличие от доказанного влияния на инициацию трансляции, роль вторичной структуры мРНК в элонгации трансляции остается непроясненной. В исследовании [40] было показано, что *in vivo* стабильность структуры мРНК всего гена, а не только его 5'-части, значительно анти-скорелирована с эффективностью его трансляции. Причем этот эффект сохраняется и при ингибировании трансляции касугамицином, а значит, не объясняется исключительно хеликазной активностью рибосом. Вместе с тем, в аналогичном исследовании на том же организме (*Escherichia coli*), но с использованием другого экспериментального метода определения структуры РНК, напротив, было показано, что эффективность трансляции связана только со структурой первых $\sim 100 - 150$

нуклеотидов от начала гена [60]. Работы, сопоставляющие плотности рибосом и стабильность вторичной структуры на небольших участках мРНК в пределах одного гена, также репортируют отсутствие статистически значимой корреляции между скоростью трансляции и вторичной структурой РНК вдали от начала гена [61; 62].

Еще одним фактором, возможно, также связанным со вторичной структурой РНК, является оптимальность используемых кодонов. Высоко экспрессируемые гены, в среднем, закодированы кодонами, соответствующими более частым транспортным РНК, чем низко экспрессируеммые. Однако значительное число высоко экспрессируемых генов начинаются с участка, закодированного редкими кодонами, так называемой «кодонной рампы». Считается, что медленная трансляция этого участка может уменьшать число столкновений рибосом дальше по последовательности гена, что в свою очередь может увеличивать точность и эффективность трансляции [63; 64]. Однако альтернативное объяснение заключается в том, что использование редких кодонов в начале гена обусловлено избеганием на этом участке сильной вторичной структуры, так как редкие кодоны, в среднем, более AT-богатые [65].

Подобная же неоднозначность связана и с влиянием на скорость элонгации фрагментов, похожих на последовательность Шайна–Дальгарно. Наблюдение о повышенной плотности рибосом в таких участках мРНК [66] было впоследствии признано артефактом использовавшегося протокола рибосомного профилирования [67; 68]. Однако, несколько независимых работ, использовавших другие экспериментальные методы, также указывали на более низкую эффективность трансляции генов, содержащих Шайна–Дальгарно-подобные последовательности близко к началу гена [59; 69].

В нашей работе мы постарались оценить влияние всех этих факторов и их вклад в эффективность трансляции *Escherichia coli* как при помощи уже опубликованных данных (Глава 3), так и на основе новых данных об эффективности трансляции десятков тысяч рандомизированных последовательностей (Глава 2).

1.3.5 Регуляция редактирования мРНК

Еще один аспект биологии мРНК, который мы затрагиваем в своих работах, связан с гидролитическим дезаминированием аденинов (далее – редактированием). Это самая часто встречающаяся модификация РНК у эукариот, при которой аденин (A) теряет C6-аминогруппу и превращается в инозин (I). Если модификация происходит в матричной РНК, то при последующем сплайсинге или трансляции такое основание распознается как гуанин, что может значительно изменять свойства закодированного белка [70—72]. Этот механизм диверсификации транскриптома широко распространен в разных группах организмов, отдельные случаи редактирования аденинов известны и в мРНК бактерий [73; 74].

У эукариот гидролитическое дезаминирование аденинов, как правило, осуществляется ферментами семейства ADAR. Чаще всего редактируемые аденины находятся в определенном контексте как по последовательности, так и по структуре РНК. Редактирование $A \to I$ значимо чаще происходит в определенных контекстах (± 1 нуклеотид), однако эти предпочтения слабы и ощутимо различаются для разных ферментов ADAR-семейства [75]. Вместе с тем, закономерости, связанные со структурой РНК вокруг редактируемого аденина, как предполагают, являются более универсальными. Сайты редактирования, как правило, располагаются в длинных (~ 50 нуклеотидов) двуцепочечных участках [75], в то время как сам редактируемый аденин, как правило, не спарен [76; 77].

Редактирование аденинов отдельных генов может быть функционально важным (например, такие случаи известны у дрозофилы [78], нематоды *Caenorhabditis elegans* [79] и человека [80; 81]), однако чаще всего сайты редактирования располагаются в Alu-повторах в некодирующих областях транскрипта, и функция таких событий неизвестна. В большинстве изученных видов редактирование происходит не более чем в 5% транскриптов. При этом подавляющая часть сайтов редактирования у млекопитающих не консервативна [80], что сви-

детельствует о низкой важности редактирования в большинстве сайтов. Совсем иначе обстоит дело в случае мягкотелых головоногих моллюсков (колеоидов – осьминогов, кальмаров и каракатиц). У них в той или иной степени редактируется большая часть транскриптов [82], и сайты редактирования консервативны между видами [83].

В нашей работе [2] мы показали, что такое массовое редактирование у колеоидов может являться эпигенетическим механизмом, который создает функциональный эквивалент мутациям в геноме. В этой и последующей работе [3] мы также описали закономерности вторичной структуры РНК вокруг редактируемых сайтов и изменения этой структуры между видами, что легло в основу Главы 4 настоящей диссертационной работы.

1.4 Экспериментальные методы определения структуры РНК

Современные методы, как правило, не позволяют точно определить структуру большой молекулы РНК. Однако, они позволяют установить, какие участки этой РНК скорее всего спарены. Более точно, данные современных массовых экспериментов позволяют сделать три типа наблюдений: оценить степень доступности отдельных нуклеотидов (PARS [84], DMS-seq [85], SHAPE [86], FragSeq [87]); детектировать пары участков РНК, которые спарены между собой (PARIS [88]); и детектировать пары участков, которые сближены в пространстве (в силу спаривания или сближенности в третичной структуре) (RPL [89], SPLASH [90]). В этой работе (см. Главу 3) мы использовали данные методов DMS-seq и SHAPE.

При DMS-seq (он же Structure-seq) происходит модификация доступных аденинов и цитозинов диметилсульфатом (ДМС). Доступными оказываются нуклеотиды, не спаренные в структуре и не закрытые рибосомами или РНК-связывающими белками. Дальше происходит обратная транскрипция, и осуществляющий ее фермент (ревертаза) прекращает работу, встретив модифицированный нуклеотид. Полученные фрагменты ДНК амплифицируют полимеразной

цепной реакцией и секвенируют, после чего полученные последовательности ридов картируют на транскриптом. Ту же самую процедуру повторяют с контрольным образцом, в который не добавляют диметилсульфат. По относительной плотности концов ридов в эксперименте с ДМС относительно контроля можно оценить, насколько часто какие участки РНК оказываются доступны для модификации (т. н. реактивность) [85].

При обработке данных DMS-seq в качестве меры уровня стуктурированности часто используют индекс Джини, посчитанный для реактивностей аденинов и цитозинов, находящихся в интересующем участке транскрипта [91]. Индекс Джини отражает степень неоднородности интересующего множества по какому-либо признаку. В данном случае, участки РНК, обладающие стабильной вторичной структурой, содержат как нуклеотиды с высокой реактивностью (открытые), так и нуклеотиды с низкой реактивностью (закрытые), что соответствует высоким значениям индекса Джини. Неструктурированные же участки, напротив, демонстрируют более равномерное распределение реактивности, что соответствует более низким значениям индекса.

Метод SHAPE имеет два основных отличия от DMS-seq: во-первых, в нем, как правило, используют реагенты, модифицирующие все четыре типа нуклеотидов, а во-вторых, получаемые модификации не останавливают ревертазу, а заставляют ее вносить мутации в позициях ДНК, соответствующих модифицированным основаниям. Таким образом, в итоговых ридах доступные нуклеотиды определяют по отличию от референсного транскриптома. При обработке данных SHAPE в качестве меры уровня структурированности часто используют медианное значение реактивностей всех нуклеотидов, находящихся в интересующем участке транскрипта. Чем меньше это значение, тем большая часть нуклеотидов находится в спаренном состоянии, а следовательно, тем выше структурированность участка [60].

Глава 2. Роль первых кодонов в трансляции у *Escherichia coli*

Давно известно, что употребление синонимичных кодов специфично для генома [92; 93] и даже для групп генов в одном геноме [94]; более того, оно может различаться даже вдоль одного гена [63; 95]. Предполагают, что использование определенных кодонов в определенных позициях гена может влиять на эффективность его трансляции, однако, несмотря на множество исследований, механизм этого влияния не вполне ясен.

В этой главе описан анализ данных об эффективности трансляции в клетках *E. coli* для генов, содержащих рандомизированные участки. Эти данные были получены в лаборатории П.В.Сергиева. Вся экспериментальная работа была проделана коллегами, все расчеты и обработка данных, кроме непосредственно распределения вариантов по фракциям на основе данных секвенирования, а также интерпретации результатов, полученных на проверочных последовательностях, была проделана автором. Изложение следует нашей работе [1], оттуда же взяты рисунки.

2.1 Материалы и методы

2.1.1 Получение плазмид, содержащих ген белка CER с рандомизированными вставками

Лабораторией П.В. Сергиева в НИИ ФХБ МГУ была разработана методика исследования свойств кодирующей последовательности, влияющих на эффективность трансляции [96]. Эта методика основана на порождении больших библиотек плазмид, которые содержат ген флуоресцентного белка со случайными вставками в интересующей области. Так как считается, что эти вставки не нарушают флюоресценцию модифицируемого белка, уровень флюоресценции полагается соответствующим количеству произведенного белка, то есть эффективности его трансляции. Точнее, эффективность трансляции определяется

как соотношение уровня флюоресценции белка CER, закодированного модифицированным геном, к уровню флюоресценции закодированного в той же плазмиде белка RFP, чей ген не подвергается модификациям и имеет тот же уровень транскрипции, что и ген белка CER. Клетки, трансфицированные такими плазмидами со случайными вставками, сортируют по соотношению свечения модифицированного CER и контрольного RFP, после чего производят массовое секвенирование вставок из клеток с равным отношением уровней свечения двух белков. В результате для нескольких классов эффективности трансляции (т. наз. фракций эффективности трансляции, ФЭТ, TEF) получаются большие наборы вставок, обеспечивающих этот уровень эффективности. В дальнейшем можно изучать и сопоставлять наборы вставок из разных фракций. Схема используемого метода приведена на Рисунке 2.1.

Плазмиды со случайными вставками были получены из плазмиды pRFPCER, описанной в работе [96], путем вставки 30-нуклеотидной случайной последовательности ДНК сразу после старт-кодона гена флюоресцентного белка CER.

2.1.2 Штаммы *Escherichia coli*, используемые в работе

Плазмидами, содержащими ген с рандомизированным участком, трансформировали клетки штамма *Escherichia coli* BW25113, а также клетки модифицированного штамма (Δ Arg), в котором были вырезаны гены, кодирующие три аргининовых тРНК: *argY*, *argZ* и *argQ*. На место вырезанных генов была вставлена кассета устойчивости к канамицину.

2.1.3 Оценка свойств последовательностей

Последовательности случайных вставок, прошедшие предварительную очистку и фильтрацию и расклассифицированные по пяти фракциям, были

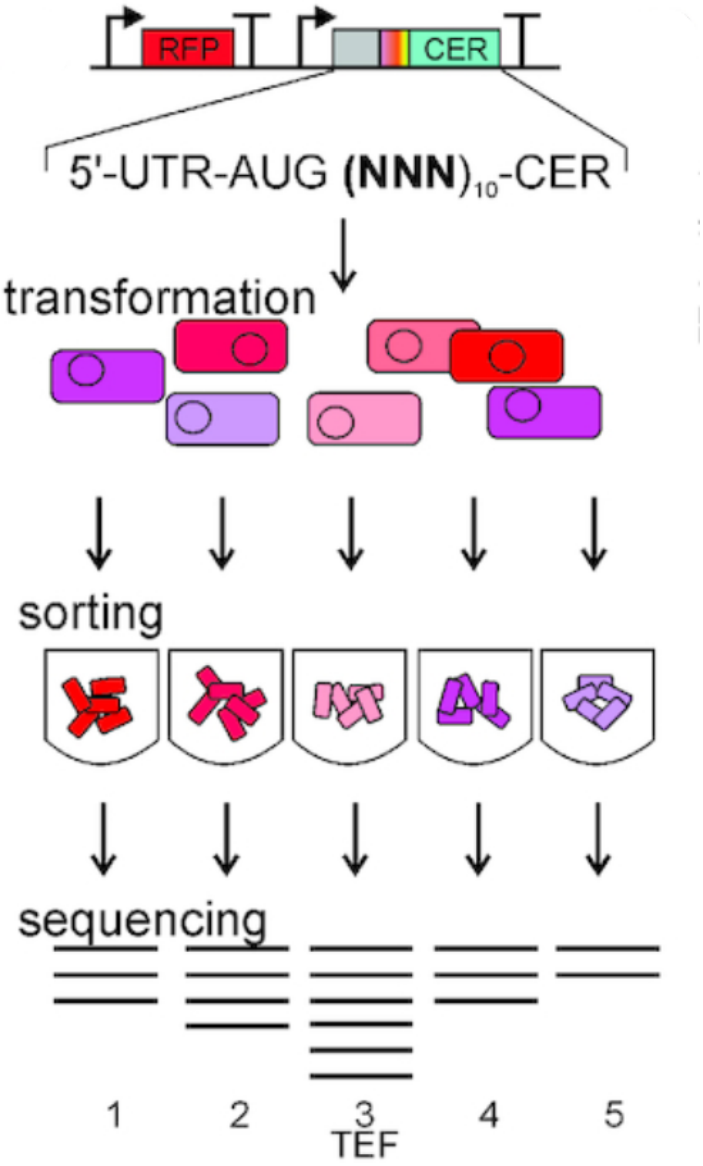

Рисунок 2.1 — Схема эксперимента Flow-seq. Сверху вниз: создание библиотеки, трансформация клеток *E. coli* плазмидной библиотекой, сортировка по соотношению флюоресценции CER/RFP на пять фракций (TEF), секвенирование.

предоставлены П.В. Сергиевым и С.А. Евфратовым. Для анализа свойств этих последовательностей, мы применяли следующие методы:

– Энергию вторичной структуры РНК вычисляли при помощи программы RNAfold ver. 2.1.7 из пакета ViennaRNA с параметрами по умолчанию [97]. При этом вычислялась энергия следующей последовательности: AGAAGGAGAUAUCAU + AUG + <случайная вставка размером 30 нуклеотидов> + GGAUCCCUGAAAGAGACGGACGAGAGCGGCCUGGUGAGCAAGG GCGAGGA, где AGAAGGAGAUAUCAU – стандартная 5'-нетранс-

лируемая лидерная последовательность, AUG – стартовый кодон, GGAUCCCUGAAAGAGACGGACGAGAGCGGCCUGGUGAGCAAGG GCGAGGA – константная часть гена.

– Сходство с участком Шайна–Дальгарно (ШД) оценивалось при помощи RNAfold как свободная энергия гибридизации с анти-Шайн–Дальгарно участком (анти-ШД) в 16S рРНК (CACCUCCU).
– Метаболическая стоимость аминокислот была взята из работы [98].

2.1.4 Контроль на нуклеотидный состав фракций

Чтобы оценить, насколько описываемые свойства обусловлены разницей в нуклеотидном составе, мы всюду в качестве контроля использовали 10 000 наборов случайно перемешанных последовательностей, таких что частоты нуклеотидов отдельно в каждой позиции каждой фракции оставались такими же, как были в исходной выборке.

2.1.5 Оценка значимости различия распределений

Для некоторых исследуемых признаков (энергия вторичной структуры РНК, сходство с участком Шайн-Дальгарно) были построены распределения значений признаков для последовательностей каждого из пяти классов эффективности. Чтобы оценить значимость различия этих распределений между классами, была проделана следующая процедура:

1. Значение нижнего квартиля обобщенного распределения для всех пяти классов было взято в качестве порогового.
2. Распределение каждой фракции было разделено на «голову» (выше порога) и «хвост» (ниже порога).
3. Количество последовательностей, оказавшихся в голове или хвосте каждого из распределений, было записано в таблицу размером 5x2, и для нее было посчитано значение статистики $\tilde{\chi}^2$.

4. Та же процедура была применена к значениям признака для 10 000 перемешанных выборок, и для каждой из них было получено свое значение статистики $\tilde{\chi}^2$.
5. Считалась доля случайно перемешанных выборок, для которых значение $\tilde{\chi}^2$ было больше либо равно значения для настоящей выборки. Таким образом оценивалась общая значимость наблюдаемого различия между классами по исследуемому признаку, контролируемая на различия нуклеотидного состава между классами.

2.1.6 Оценка скоррелированности частот отдельных кодонов и аминокислот с распределением последовательностей по фракциям

Для каждого кодона в каждой позиции вставленной последовательности (2-11) были посчитаны его частоты в каждой из фракций. Из всех пар кодон + позиция были оставлены те, для которых изменение частоты в зависимости от номера фракции было монотонно, что в свою очередь определялось через коэффициент корреляции Спирмена, — от него требовалось быть по модулю больше 0.8. Такие пары кодон + позиция мы будем назвать монотонными.

Сила эффекта кодона на распределение по фракциям была определена как коэффициент линейной регрессии позиционной частоты кодона в зависимости от номера фракции; для немонотонных позиций сила эффекта была приравнена к нулю. *p-value* и *z-score* для этих значений были посчитаны относительно соответствующих распределений коэффициентов регрессий для перемешанных выборок (см. 2.1.4. «Контроль на нуклеотидный состав фракций»).

Сила эффекта аминокислоты была посчитана как взвешенная сумма эффектов кодонов, соответствующих этой аминокислоте, с весами, соответствующими частотам этих кодонов в генах *E. coli*. *Сила эффекта позиции* была посчитана как среднее абсолютное значение эффектов для всех монотонных кодонов в этой позиции.

2.1.7 Дизайн проверочных последовательностей

Для того, чтобы экспериментально подтвердить эффекты, обнаруженные при анализе данных массового эксперимента, для каждого изучаемого признака мы составили такой набор последовательностей, чтобы в нем этот признак значительно варьировал, в то время как по остальным признакам этот набор был бы максимально однородным.

Для подтверждения позиционных эффектов отдельных кодонов мы использовали набор последовательностей, содержащих один, два или три кодона, для которых был предсказан отрицательный эффект на продуктивность трансляции. Последовательности не содержали дополнительных старт-кодонов, участков, похожих на участок Шайна-Дальгарно (ШД-подобных участков), а их энергия сворачивания находилась в интервале 15.8 ± 1.7 Ккал/моль. Аналогичным образом была составлена выборка для подтверждения влияния положения ШД-подобных участков и дополнительных старт-кодонов. И, наконец, для подтверждения эффекта метаболической стоимости аминокислот были протестированы два набора последовательностей, имеющие почти одинаковый кодонный состав и энергию сворачивания, но кодирующие аминокислоты разной метаболической стоимости. Полный список проверочных последовательностей представлен в Приложении.

2.2 Результаты и обсуждение

2.2.1 Распределение последовательностей по фракциям в значительной степени совпадает между репликами

Число случайных последовательностей длины 30 значительно превышает количество вариантов, способных поместиться в одну библиотеку, поэтому в этой работе мы имели дело лишь с малой частью всех возможных вариантов. Размер полученной библиотеки составлял порядка 10^6 независимых клонов, по-

сле секвенирования и фильтрации было получено 154 927 различных вариантов вставки, и из них 51 078 были найдены в обеих репликах эксперимента. Один и тот же фрагмент чаще всего попадал в ту же фракцию, а если и оказывался в другой, то чаще всего в соседней (Рисунок 2.2).

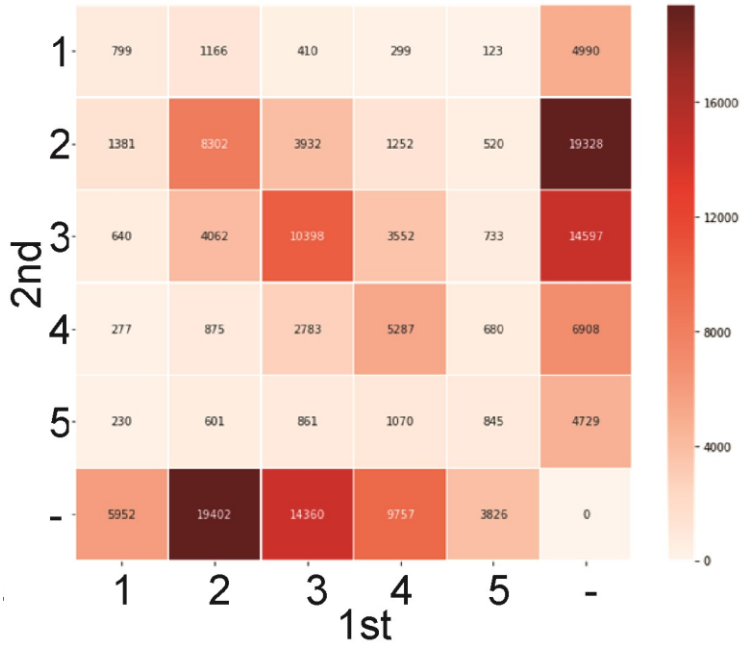

Рисунок 2.2 — Матрица воспроизводимости для двух независимых экспериментов (трансформация и сортировка). Клетка с координатами (i, j) соответствует вариантам плазмид, которые в первой реплике были обнаружены в i-ой фракции, а во второй — в j-ой. Строка без индекса и колонка без индекса соответствуют вариантам, обнаруженным только в одной из реплик.

Среди случайных последовательностей могли оказаться последовательности, содержащие стоп-кодон в рамке считывания. Естественно было бы ожидать, что такие последовательности практически не транслируются (если только не происходит ошибок трансляции или после стоп-кодона так же случайно не сформировался дополнительный старт-кодон). Мы проверили распределение таких последовательностей по фракциям, и, действительно, оказалось, что в ФЭТ с высокой эффективностью трансляции последовательностей, содержащих стоп-кодоны, значительно меньше, чем в ФЭТ с низкой эффективностью (см. Рисунок 2.3)

Для дальнейшего анализа мы отобрали фрагменты, которые не содержали стоп-кодонов в рамке, которые присутствовали в обоих репликах, и номе-

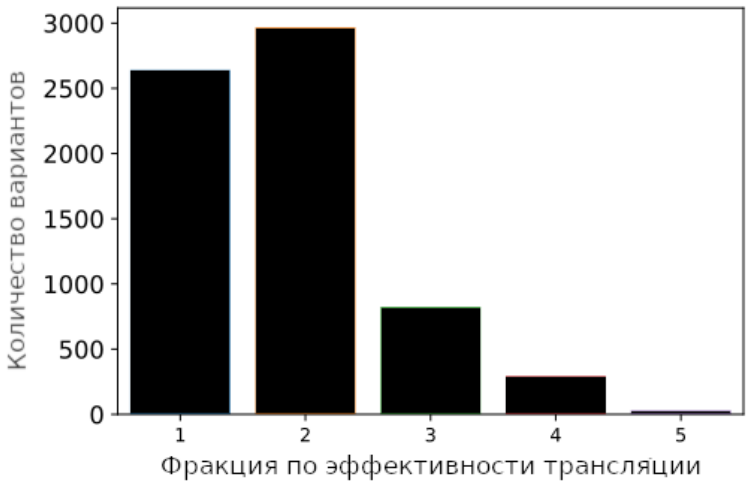

Рисунок 2.3 — Распределение по фракциям вариантов последовательностей, содержащих стоп-кодон в рамке считывания.

ра фракций которых в этих репликах совпадали. Таких вариантов оказалось 25 631.

2.2.2 Стабильная вторичная структура избегается в 5'-областях высокотранслируемых последовательностей

В соответствии с предыдущими наблюдениями [27; 57; 99—102], вторичная структура в 5'-области гена мешает эффективной трансляции: распределение энергии сворачивания РНК в эффективных ФЭТ сдвинуто в сторону малых по абсолютной величине значений (Рисунок 2.4). Поскольку эффективные ФЭТ имеют в среднем более низкий GC-состав (Рисунок 2.5), для оценки статистической значимости был использован контроль с рандомизацией последовательностей (см. 2.1.4. «Контроль на нуклеотидный состав фракций»). Значимость избегания сохраняется и после такого контроля ($p\text{-}value < 10^{-4}$).

Как энергии сворачивания, так и нуклеотидный состав эффективно транслируемых вставок близки к соответствующим значениям для 5'-участков настоящих генов *E. coli*.

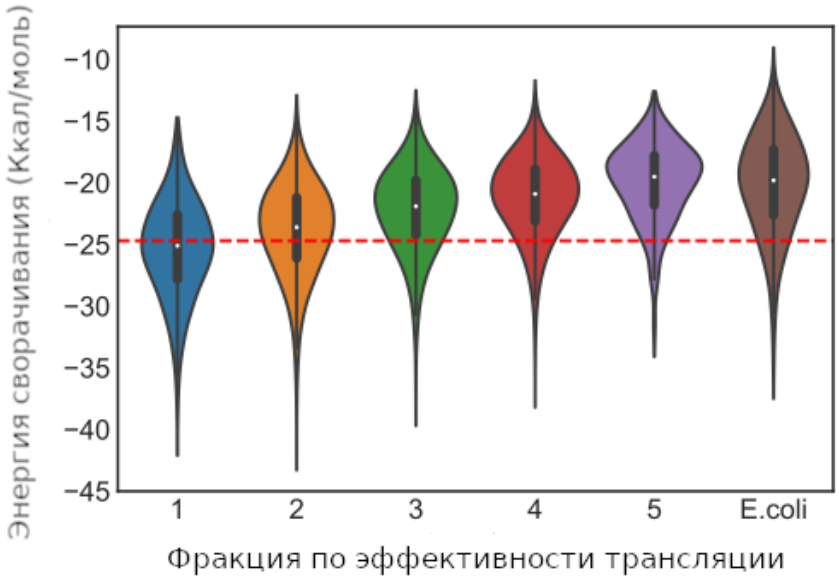

Рисунок 2.4 — Распределения значений энергии сворачивания 5'-конца гена со случайной вставкой. Крайнее правое распределение соответствует энергиям сворачивания 5'-концов генов *E.coli*. Пунктиром показано значение нижнего квартиля для объединенной выборки экспериментальных последовательностей.

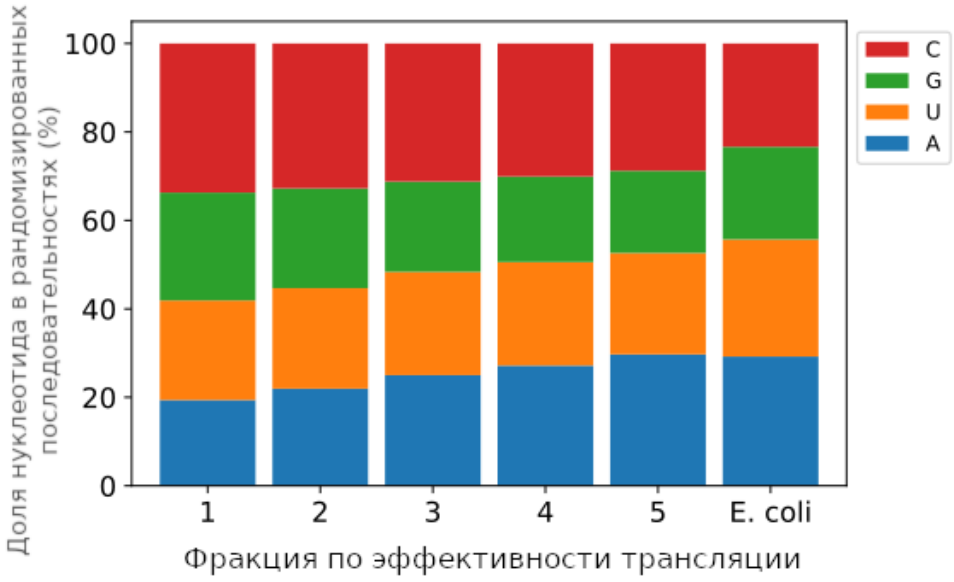

Рисунок 2.5 — Частоты нуклеотидов в разных фракциях и в начальных фрагментах настоящих генов *E.coli* (для каждого кодирующего гена брались первые 30 нуклеотидов).

2.2.3 Участки, похожие на последовательность Шайна–Дальгарно, также избегаются в 5'-областях высокотранслируемых последовательностей

Избегание ШД-подобных участков в белок кодирующих генах было вычислительно показано для многих бактерий [103; 104]. В экспериментах по рибосомному профайлингу в *E. coli* была обнаружена повышенная плотность рибосом на ШД-подобных участках. Была высказана гипотеза, что рибосомы притормаживают на таких участках, и это понижает эффективность трансляции [66]. Однако более тщательный анализ дал основания полагать, что повышенная плотность рибосом на таких участках являлась артефактом используемого метода [67], поэтому вопрос о влиянии внутригенных ШД-подобных участков оставался не до конца проясненным.

Анализ сродства наших случайных последовательностей к анти-Шайн–Дальгарно участку 16S рРНК (см. 2.1.3. «Оценка свойств последовательностей») показал избегание возможности такого взаимодействия во вставках из эффективно транслируемых фракций (Рисунок 2.6, $p\text{-}value < 10^{-4}$). Это подтверждает предыдущие наблюдения о негативном влиянии внутригенных ШД-подобных участков на эффективность трансляции [59; 66].

Для подтверждения наблюдаемого эффекта был синтезирован набор проверочных последовательностей, и для каждой из них точно так же, как и в массовом эксперименте, была измерена эффективность трансляции. Каждая из последовательностей содержала ШД-подобный участок в одном из возможных положений, в то время как нуклеотидный состав и энергии сворачивания этих последовательностей значительно не различались (-13 ± 0.2 Ккал/моль, сами последовательности представлены в Приложении. Наблюдаемые значения эффективности трансляции для этих последовательностей согласуются с нашим наблюдением, полученным на массовых данных: трансляция последовательностей, содержащих ШД-мотив практически в любой из позиций, менее эффективна, чем трансляция контрольной последовательности, не содержащей

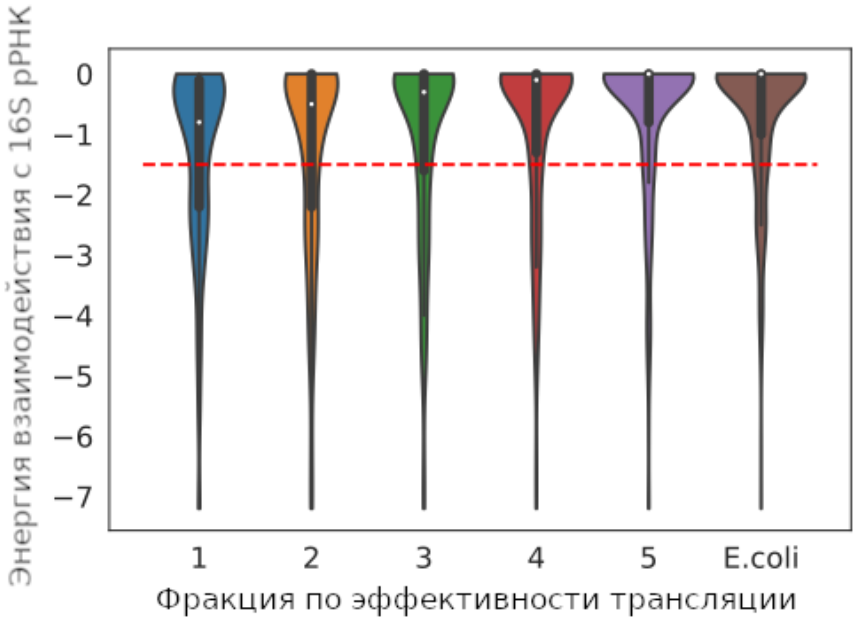

Рисунок 2.6 — Распределение энергий взаимодействия с анти-ШД-участком 16S рРНК (Ккал/моль) в разных ФЭТ и в начальных фрагментах настоящих генов *E.coli*. Пунктиром показан нижний квартиль значений всей выборки.

ШД-подобных участков (Рисунок 2.7). Кроме того, чем ближе ШД-подобный участок находится к старт-кодону, тем больший отрицательный эффект он оказывает. Один из возможных механизмов, обеспечивающих этот эффект, заключается в следующем: рибосома тормозит на ШД-подобном участке, закрывает старт-кодон, и это мешает новым рибосомам начать трансляцию.

2.2.4 Влияние конкретных кодонов на эффективность трансляции

Для оценки влияния конкретных кодонов на эффективность трансляции мы определили позиционные частоты кодонов в каждой фракции и сравнили их с таковыми для рандомизированных последовательностей (Рисунок 2.8а). Сравнение левой и правой панели рисунка показывает, что значительную часть эффектов можно было бы с хорошей вероятностью получить и на случайной выборке, сохраняющей только позиционные частоты нуклеотидов в разных фракциях (*z-score* таких наблюдений близок к нулю, и соответствующие ячейки тепловой карты в правой панели нейтрально-желтые). Однако количество ячеек

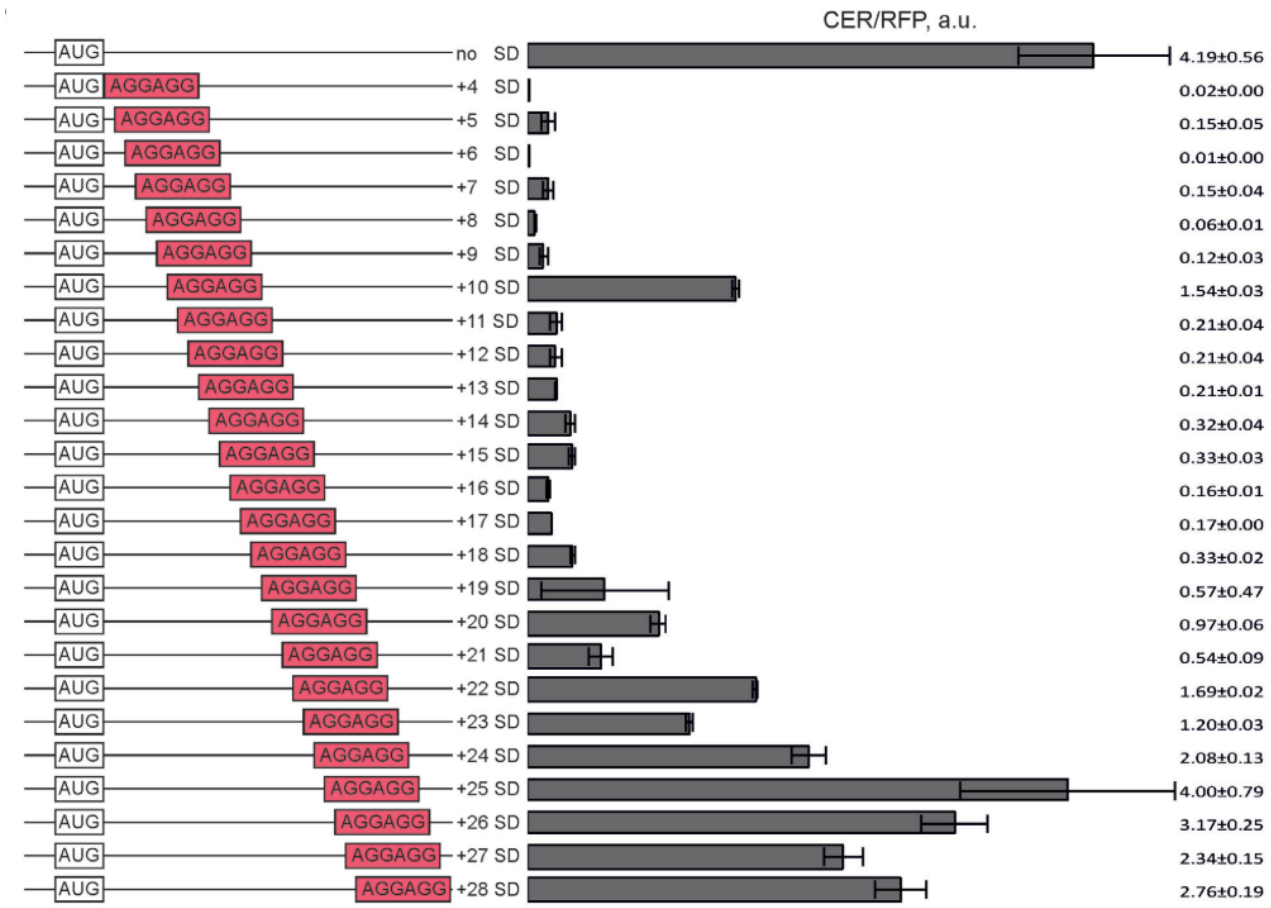

Рисунок 2.7 — Эффективность трансляции для проверочных последовательностей, содержащих ШД-подобные участки. Схематичное представление положения ШД-подобного участка в последовательности (слева) и значения эффективности трансляции (справа).

(пар (*кодон, позиция*)), для которых наблюдается хоть какой-то эффект (цвет ячейки отличен от желтого), существенно выше в настоящей выборке, чем в перемешанных контролях (Рисунок 2.8б). Это свидетельствует о том, что, хотя для отдельных кодонов наблюдаемые эффекты и могут объясняться различием позиционного нуклеотидного состава в разных фракциях, всю совокупность наблюдаемых эффектов нуклеотидный состав не объясняет. Кроме того, влияние кодонов уменьшается с отдалением от точки старта трансляции (Рисунок 2.8в), что согласуется с предыдущими наблюдениями [63; 95] и косвенно валидирует наш способ статистической обработки.

Оцененное таким образом влияние отдельных кодонов на трансляцию не противоречит нашим предыдущим наблюдением об избегании ШД-подобных участков. Так, например, кодоны GGA, GGG, AGG, последовательности которых похожи на часть участка Шайна–Дальгарно (консенсус в *E. coli* AGGAGG),

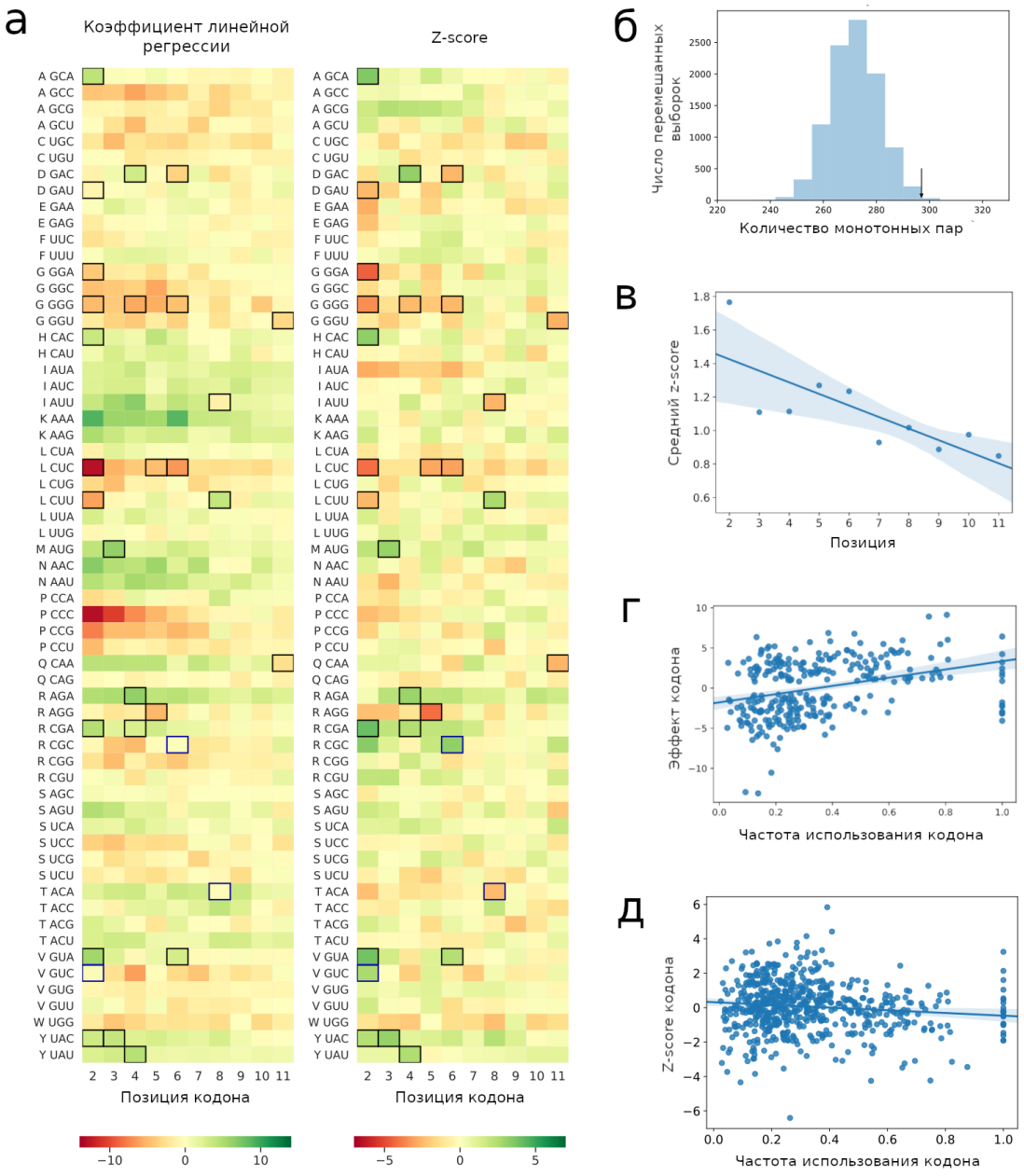

Рисунок 2.8 — **Влияние кодонов (а)** В каждой клетке показан коэффициент линейной регрессии частоты кодона в данной позиции в зависимости от номера ФЭТ (слева) и *z-score* наблюдаемого коэффициента регрессии при сравнении с перемешанными последовательностями. Зеленый: положительные значения, т.е. кодоны, усиливающие уровень трансляции; красный: отрицательные значения; желтый: кодоны с немонотонной зависимостью. Клетки с $p\text{-value} < 0.01$ обведены рамкой, по случайным причинам в каждой тепловой карте ожидается 6 таких клеток. **(б)** Распределение числа пар (*кодон, позиция*), для которых в перемешанных последовательностях наблюдается монотонная зависимость частоты от ФЭТ. Стрелкой показано число пар для настоящей выборки, $p\text{-value} < 10^{-4}$. **(в)** Средний нормированный эффект кодона (*z-score*) в зависимости от позиции. Коэффициент корреляции Пирсона $r = -0.8$, $p\text{-value} = 0.05$ **(г)** Эффект кодона, не нормированный на нуклеотидный состав (коэффициент корреляции из п. а), в зависимости от частоты кодона в генах *E. coli*. $r = 0.33$, $p\text{-value} = 9.6 \times 10^{-9}$ **(д)** То же, что и в п. (г), после нормировки на нуклеотидный состав. $r = -0.12$, $p\text{-value} = 0.004$

.

оказывают отрицательный эффект на трансляцию. Кроме того, общие отрицательные эффекты конформационно гибкого глицина (G) и жесткого пролина (P) согласуются с наблюдениями о том, что рибосомы чаще застревают в окрестностях этих аминокислот [67; 105; 106].

Для прямой экспериментальной проверки наблюденных эффектов был использован набор последовательностей, содержащих 1, 2 или 3 кодона с предсказанным ингибиторным эффектом, в то время как остальные кодоны были обладали предсказанным положительным эффектом. Энергии сворачивания этих последовательностей значительно не различались (-15.8 ± 1.7 Ккал/моль), и сродство к участку анти-Шайна–Дальгарно было минимально. Сами последовательности представлены в Приложении. Как видно на Рисунке 2.9а, в случае одного ингибиторного кодона, отрицательное влияние на трансляцию, в целом, тем менее значительно, чем дальше этот кодон находится от старта трансляции. Эта закономерность соответствует нашему наблюдению, сделанному на массовых данных (Рисунок 2.8в). Кроме того, добавление дополнительных кодонов с отрицательным эффектом усиливает ингибирование трансляции (Рисунок 2.9б).

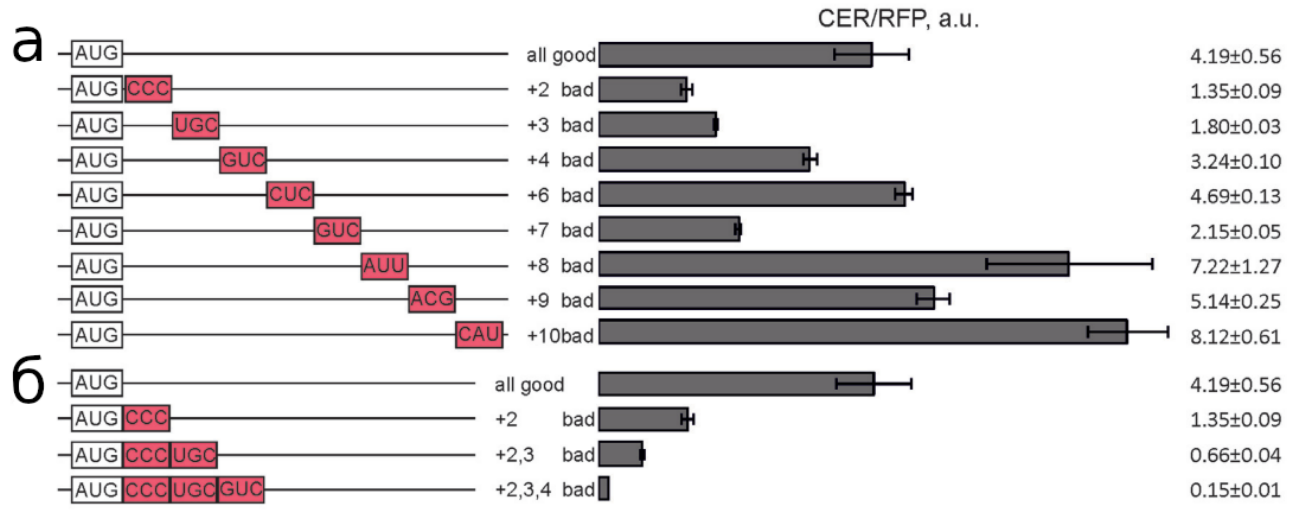

Рисунок 2.9 — Эффективность трансляции для проверочных последовательностей с ингибирующими кодонами. (а), (б) Схематичное представление положений ингибирующих кодонов (слева) и значений эффективности трансляции (справа).

Отдельно мы проверили, действительно ли дополнительный старт-кодон AUG оказывает положительное влияние на трансляцию (Рисунок 2.8а). Для

этого были использованы проверочные последовательности, состоящие, в основном, из слабо ингибирующих трансляцию кодонов и содержащие AUG в каждом из возможных положений. В то же время энергия сворачивания для всех последовательностей находилась в пределах -24 ± 0.9 Ккал/моль, а их сродство к анти-Шайна–Дальгарно участку 16S рРНК оставалось низким. Результаты показаны на Рисунке 2.10. Суммарное распределение эффективности трансляции для всех последовательностей, содержащих AUG в рамке считывания, оказалось сдвинуто в сторону больших значений относительно суммарного распределения значений для последовательностей, содержащих этот кодон не в рамке ($p\text{-}value = 0.018$, тест Манна-Уитни).

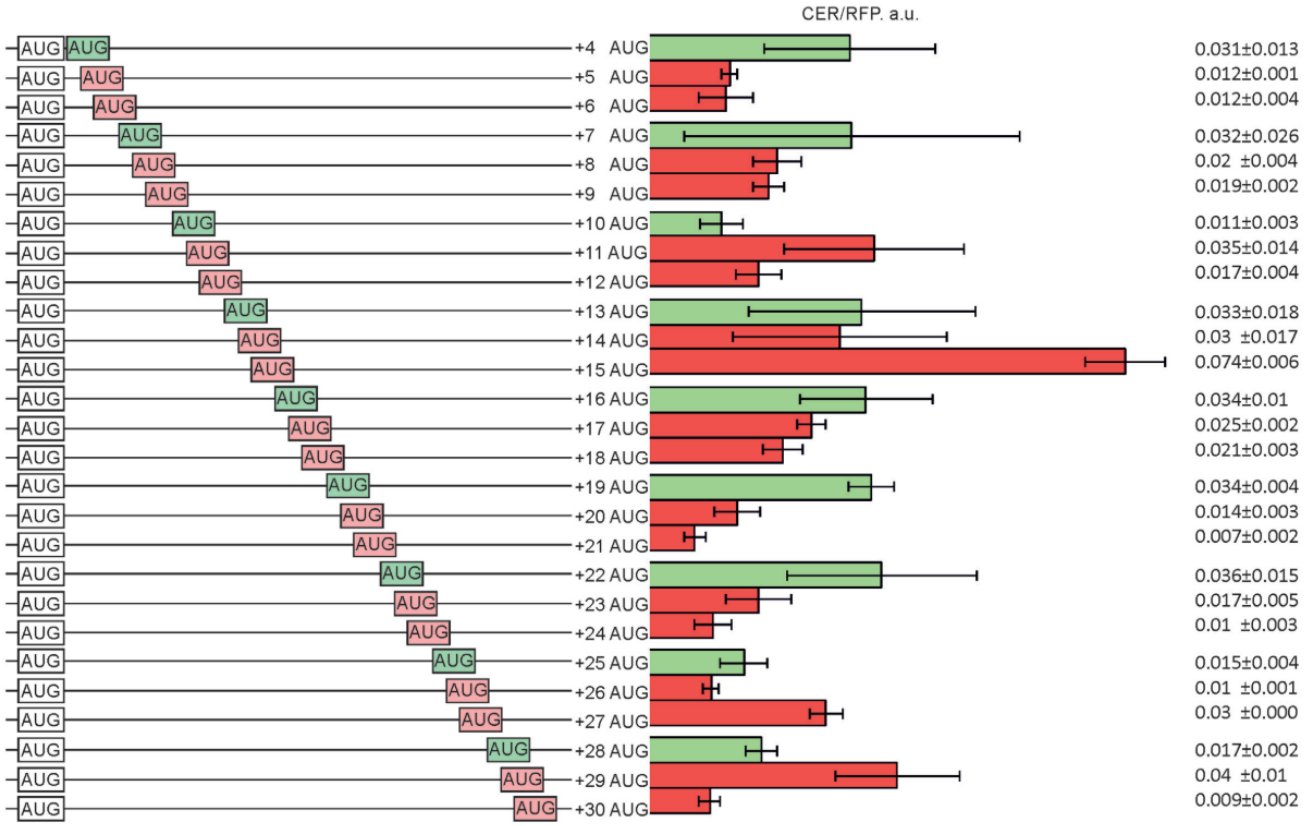

Рисунок 2.10 — Влияние дополнительных AUG-кодонов во вставке на эффективность трансляции. Схематичное представление (слева) и значения эффективности трансляции (справа). Красные столбики соответствуют положению кодона не в рамке, в то время как положения в рамке обозначены зеленым цветом. Полные последовательности вставок и соответствующие им экспериментальные значения представлены в Приложении.

2.2.5 Редкие кодоны в начале гена слабо влияют на эффективность трансляции

До настоящего времени вопрос о том, увеличивают ли редкие кодоны в начале гена эффективность его трансляции [63; 101], оставался непроясненным. Известно, что редкие кодоны встречаются в начале генов чаще, чем ожидается, однако это может быть следствием отбора против сильной вторичной структуры [100; 107]. В наших данных эффект кодона на трансляцию, напротив, слабо положительно скоррелирован с частотой использования кодона в генах *E. coli* (Рисунок 2.9г), но эта корреляция пропадает при контроле на позиционный нуклеотидный состав (Рисунок 2.9д). Использование индекса оптимальности tAI, основанного не на наблюдаемых в генах частотах кодонов, а на концентрациях тРНК, также не показывает значимых зависимостей (Рисунок 2.11).

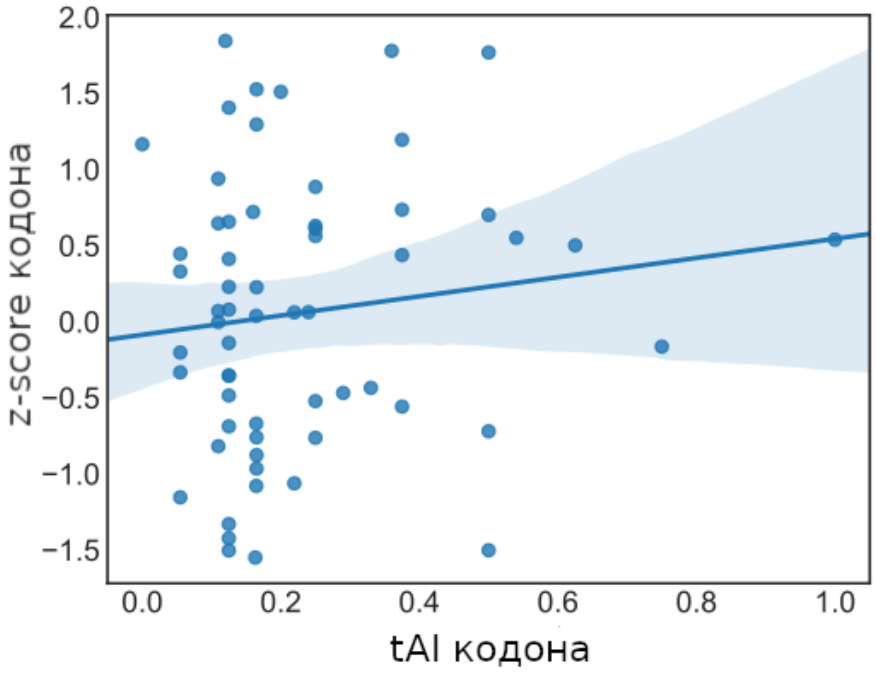

Рисунок 2.11 — Норминованный на нуклеотидный состав эффект кодона в зависимости от концентрации соответствующей ему тРНК (tAI индекс). Корреляция не значима.

Чтобы подробнее исследовать этот вопрос, были использованы аналогичные данные для штамма, в котором были удалены три из четырех генов, ко-

дирующих $\mathrm{tRNA}^{\mathrm{Arg}}_{\mathrm{ACG}}$. (Здесь ACG — антикодон. У *E. coli* четыре тРНК с этим антикодоном неспецифическим образом связываются с кодонами CGU, CGC и CGA в процессе трансляции. Для остальных трех аргининовых кодонов есть отдельные три тРНК с точно соответствующими им антикодонами.) Благодаря такой делеции концентрация $\mathrm{tRNA}^{\mathrm{Arg}}_{\mathrm{ACG}}$ уменьшилась в 5 раз, а следовательно, и значения tAI для кодонов CGU, CGC и CGA тоже уменьшились примерно в той же степени (tAI линейно зависит от концентраций). Сравнение влияния кодонов в штамме с делецией и в диком типе не выявило каких-либо ярких различий в степени влияния этих кодонов (Рисунок 2.12).

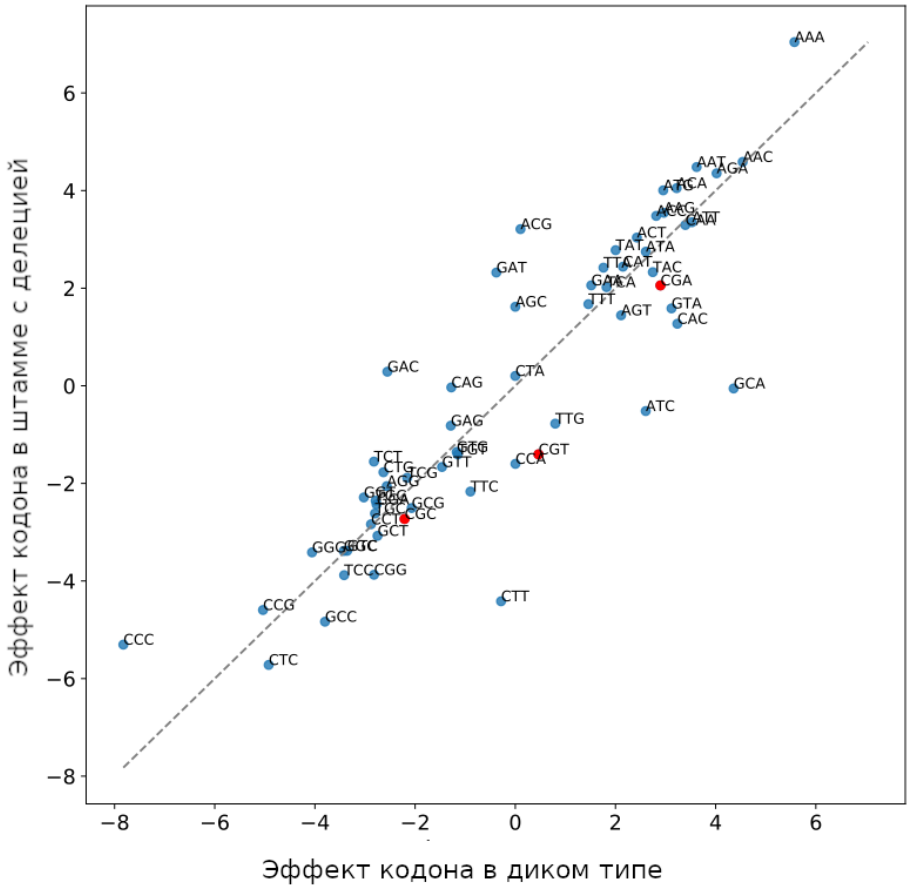

Рисунок 2.12 — Диаграмма рассеяния коэффициентов линейной регрессии частот кодонов в зависимости от ФЭТ (эффект кодона) для штамма с делетированными генами аргининовой тРНК и в диком типе. Красным отмечены кодоны, для которых уменьшилась концентрация тРНК.

2.2.6 При культивации в бедной среде эффективнее транслируются те последовательности, в которых закодированы более метаболически дешевые аминокислоты

Описанные ранее эксперименты проводились на бактериях, культивируемых в питательно богатой среде LB. Вместе с тем, в реальных условиях ресурсы среды могут быть значительно ограничены. Можно ожидать, что при дефиците аминокислот их влияние на эффективность трансляции будет расти. Чтобы проверить это предположение, массовый эксперимент был повторен в питательно бедной среде M9. Наблюдения, связанные со структурой РНК (избегание вторичной структуры и участков, похожих на мотив Шайна–Дальгарно, воспроизвелись полностью) (данные не приведены). Результаты по влиянию отдельных кодонов оказались, в целом, похожими на те, что наблюдались на богатой среде (Рисунок 2.13). Чтобы проанализировать наблюдаемые различия, мы использовали описанную далее дополнительную информацию о метаболизме *E. coli*.

Известно, что синтез разных аминокислот требует разного количества молекул АТФ и окислительных эквивалентов, которые, в свою очередь, тоже могут конвертироваться в АТФ [98]. Согласно упоминаемой статье, метаболическая стоимость каждой аминокислоты может быть выражена через число высокоэнергетических фосфатных связей, разрыв которых необходим для синтеза данной аминокислоты. Используя эти значения, для каждого наблюденного варианта вставки мы посчитали суммарную метаболическую стоимость закодированных аминокислот. При этом оказалось, что во вставках мРНК, эффективно транслируемых на бедной среде (ФЭТ с бóльшими номерами), чаще (относительно богатой среды) встречаются аминокислоты с низкой метаболической стоимостью (Рисунок 2.14а).

Для проверки этого наблюдения была измерена эффективность трансляции для двух пар вставок, сильно различающихся по метаболической стоимости и при этом максимально совпадающих по остальным важным признакам: значениям стабильности вторичной структуры, афинности к участку анти-Шайн-

Дальгарно, — и не содержащие дополнительных старт-кодонов. Клетки, трансформированные плазмидами с рассматриваемыми вставками, культивировали на пяти разных средах, из которых одна была богатая (LB), а остальные в разной степени бедные. Уровень трансляции двух последовательностей, кодирующих метаболически дорогие аминокислоты (дорогих последовательностей), был значительно ниже в бедных средах, в то время как для последовательностей, кодирующих метаболически дешевые аминокислоты (дешевых последовательностей), этот тренд в одном случае менее выражен, а в другом и вовсе обратный (Рисунок 2.14б). Таким образом, наблюдается эффект в ту же сторону, что и было предсказано по массовым данным, однако, из-за малого числа протестированных последовательностей (четыре) оценить значимость этого результата не представляется возможным.

Вероятный механизм образования дефицита метаболически дорогих аминокислот в высокотранслируемых классах может заключаться в следующем: если закодированной аминокислоты мало, рибосома может долго простаивать в ожидании заряженной тРНК, а значит, за то же время будет произведено меньше белка, чем если бы закодированной аминокислоты было много.

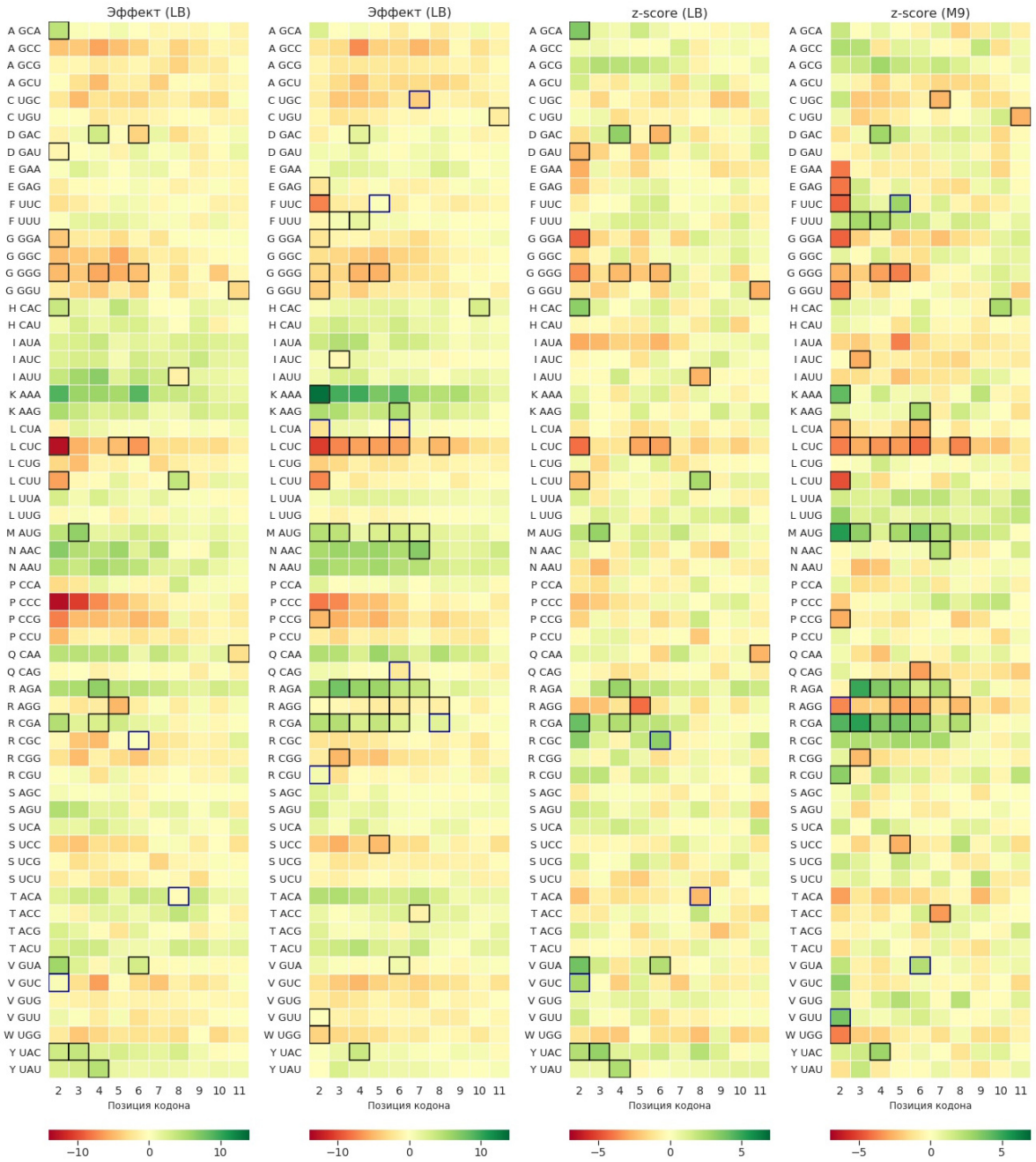

Рисунок 2.13 — Влияние кодонов для богатой (LB) и бедной (M9) сред. В каждой клетке показан коэффициент регрессии частоты кодона в данной позиции от номера ФЭТ (слева) и *z-score* наблюдаемого коэффициента регрессии при сравнении с перемешанными последовательностями (справа). Обозначения такие же, как на Рисунке 2.8а.

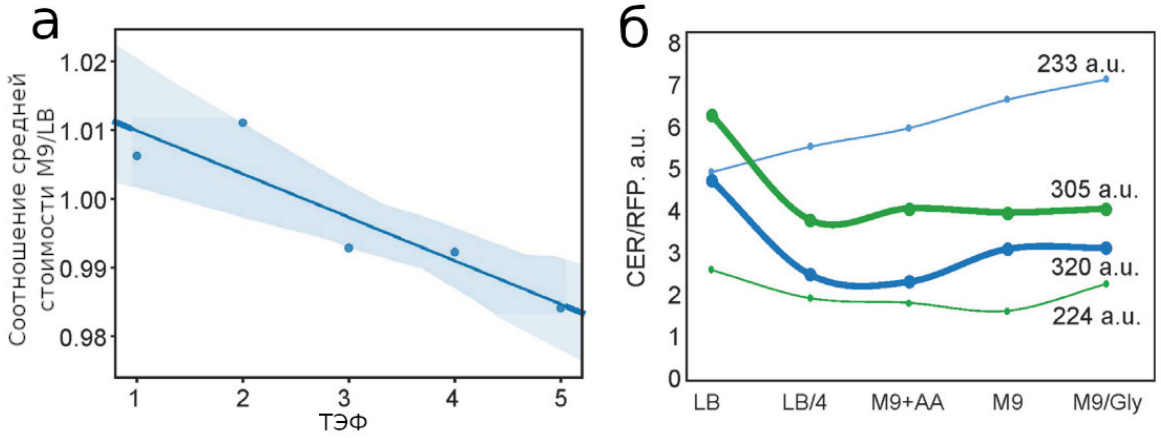

Рисунок 2.14 — Влияние метаболической стоимости аминокислот на эффективность трансляции в зависимости от среды. **(а)** Отношение средних метаболических стоимостей аминокислот, закодированных кодонами вставленной последовательности, попавших в соответствующую ФЭТ, для бедной (M9) и богатой (LB) сред. Коэффициент корреляции Пирсона $r = -0.91$, $p\text{-}value = 0.018$. Затененная область соответствует 95% доверительным интервалам для коэффициента линейной регрессии. **(б)** Эффективность трансляции в разных средах. Жирные линии соответствуют последовательностям, кодирующим дорогие аминокислоты, тонкие — кодирующим дешевые. Были использованы следующие среды: LB, разведенная в 4 раза LB (LB/4), M9 с глюкозой и аминокислотами (M9+AA), M9 с глюкозой (M9), M9 без глюкозы, но с глицеролом вместо нее (M9/Gly).

Глава 3. Роль вторичной структуры в функционировании мРНК у *Escherichia coli*

Вторичная структура матричной РНК может оказывать влияние на такие процессы, как трансляция, деградация и эволюция мРНК. Предполают, что элементы вторичной структуры, находящиеся перед (downstream) рибосомой, могут замедлять ее движение [108], что, в свою очередь, может ограничивать эффективность трансляции, а также регулировать ко-транскрипционное сворачивание белка [109]. Однако, в литературе нет консенсуса относительно того, важна ли для эффективности трансляции структура мРНК всего гена или только его начала. Чтобы глубже исследовать этот вопрос, мы рассмотрели пары генов, наиболее скоординированных по эффективности трансляции: эквимолярные субъединицы одного белкового комплекса. Кроме того, мы исследовали закономерности деградации мРНК, объединив открытые данные из разных источников.

3.1 Материалы и методы

Для исследования влияния вторичной структуры на жизненный цикл мРНК были использованы полногеномные данные о стабильности вторичной структуры, позициях рибосом и времени жизни мРНК. Стабильность вторичной структуры была оценена стандартным образом (см. 1.4. «Экспериментальные методы определения структуры РНК») из опубликованных данных о доступности РНК-нуклеотидов, полученных методами диметилсульфат-секвенирования (DMS-seq) [40] и избирательного 2'-гидроксил ацетилирования с продлением праймера (SHAPE-seq) [60]. Позиции рибосом были получены из опубликованных данных рибосомного профилирования [68], а времена полураспада мРНК – из данных РНК-секвенирования после добавления ингибитора инициации транскрипции [110].

При проверке гипотезы о скоррелированной структурированности субъединиц для каждого комплекса мы случайным образом выбирали только одну пару субъединиц, чтобы избежать зависимости между наблюдениями (например, если комплекс состоит из белков А, Б и В, наблюдения для пар генов (а,б) и (а, в) будут зависимы). При построении распределения коэффициентов корреляции для участков длины 100 нуклеотидов мы повторяли случайный выбор независимых пар 5000 раз.

Для изучения полиморфизмов в спаренных позициях генов мы выровняли геномы трехсот пяти репрезентативных штаммов *Escherichia coli* при помощи программы npge. Список репрезентативных штаммов представлен в Приложении, он был получен при помощи R-пакета GGRaSP [111], который кластеризует геномные последовательности с использованием гаусовских смешанных моделей. Мы воспроизвели анализ, проведенный в статье [111], используя обновленную базу последовательностей полных геномов *Escherichia coli*, доступных на февраль 2019 года.

В качестве внешней группы (outgroup) для поляризации замен был использован геном *Escherichia fergusonii*.

3.2 Результаты и обсуждение

3.2.1 мРНК генов, кодирующих эквимолярные субъединицы одного белкового комплекса, имеют сходную степень структурированности

Мы рассмотрели мРНК генов, кодирующих эквимолярные субъединицы одного белкового комплекса и находящихся в одном опероне. Ранее было показано, что эффективности трансляции таких пар генов с высокой точностью равны [112], но механизма предложено не было. В более поздней работе на данных диметилсульфат-секвенирования [40] была показана сильная отрицательная корреляция между трансляционной эффективностью гена и уровнем

стабильности его вторичной структуры. Тем самым, степени структурированности генов, кодирующих субъединицы, могут быть близки, и это может являться частью искомого механизма, обеспечивающего равенство эффективности трансляции.

Для проверки этого предположения мы сравнили корреляцию в степени структурированности кодирующих областей для пар генов из комплексов и для контрольных оперонов, гены в которых комплексов не образуют. Данные о генах, кодирующих субъединицы белковых комплексов, были взяты из [112]. Данные о составе оперонов были взяты из [113] и дополнительно отфильтрованы исходя из покрытий общего РНК-секвенирования: рассматривались только те пары генов, общее количество мРНК которых значимо не отличалось. Степень стуктурированности была оценена двумя способами: с использованием данных DMS-seq [40] и SHAPE-seq [60].

Оказалось, что степень структурированности в оперонах, кодирующих субъединицы, коррелирует, в отличие от контрольных оперонов (Рисунок 3.1). Мы проверили, что это наблюдение не связано с гомологичностью субъединиц: распределение уровеней сходства генов субъединиц не отличается от такового для контрольных пар (Рисунок 3.2).

В литературе встречаются утверждения, что для определения уровня трансляции важен не столько уровень структурированности всей кодирующей области, сколько уровень структурированности 5'-области и первых кодонов [60; 114]. Чтобы это проверить, аналогичные вычисления были проделаны для первых 100 нуклеотидов транскриптов. Оказалось, что корреляция, вычисленная для таких фрагментов, в среднем не значительно отличается от корреляции, вычисленной для случайных фрагментов по 100 нуклеотидов, взятых из тех же генов (Рисунок 3.3). Это наблюдение свидетельствует в пользу того, что структура всей мРНК гена важна для эффективности трансляции.

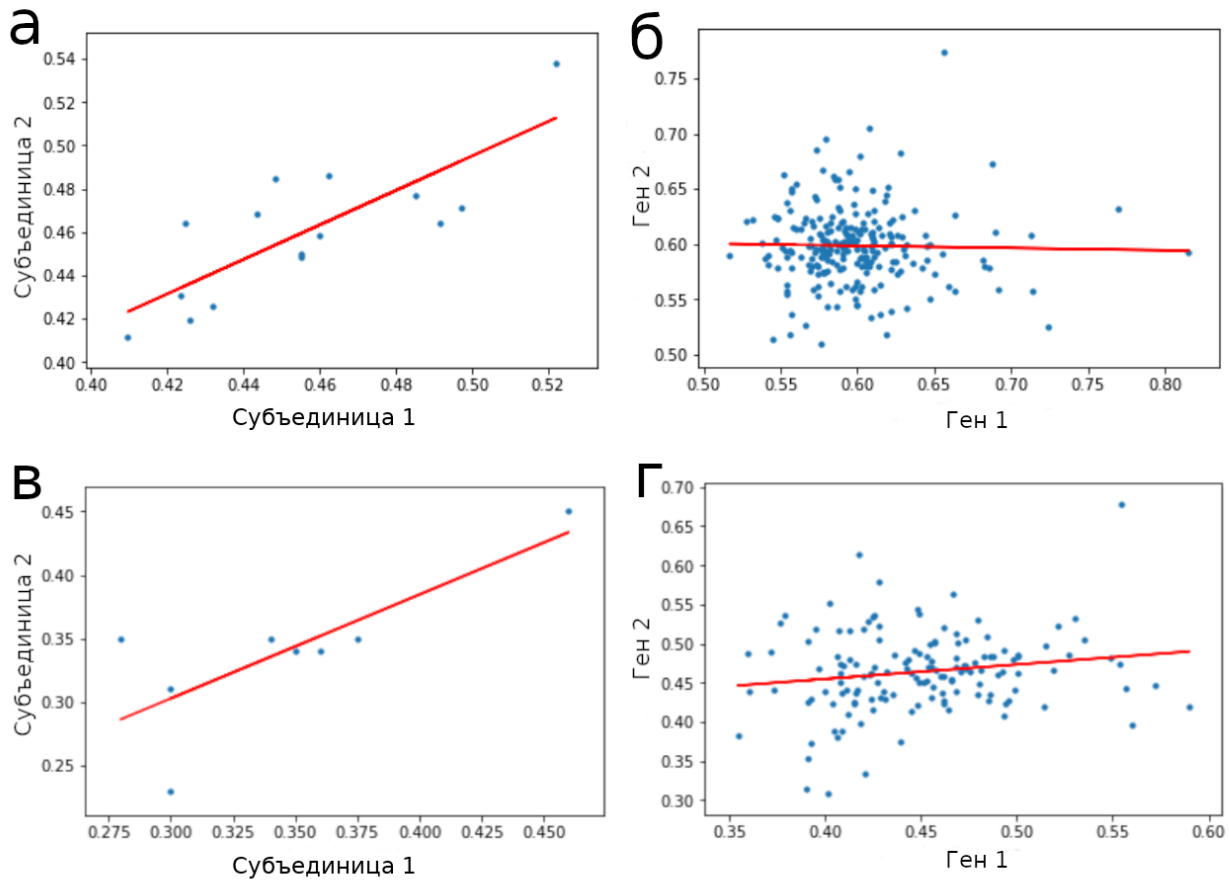

Рисунок 3.1 — Корреляция между уровнем структурированности мРНК для пар генов, кодирующих субъединицы белковых комплексов, и контрольных пар генов. Каждая точка соответствует одной паре генов, по горизонтальной оси отложена структурированность 5'-гена, по вертикальной — 3'-гена. Графики (а) и (б) построены согласно данным DMS-seq, значения по осям соответствуют индексу Джини значений покрытия соответствующих генов. Корреляция Пирсона для графика **(а)** $r = 0.7$, $p\text{-}value = 0.0024$, в то время как для графика **(б)** корреляция не значима $r = 0.058$, $p\text{-}value = 0.49$. Графики **(в)** и **(г)** построены согласно данным SHAPE, значения по осям соответствуют медианам значений покрытия соответствующих генов. Корреляция Пирсона для графика **(в)** $r = 0.86$, $p\text{-}value = 0.013$, в то время как для графика **г)** $r = 0.16$, $p\text{-}value = 0.047$.

3.2.2 Общая степень структурированности мРНК не коррелирует со скоростью ее деградации

Далее мы исследовали, как вторичная структура влияет на деградацию мРНК. Эффективно транслируемые гены деградируют немного медленнее (Рисунок 3.4а), однако сравнение времени жизни со структурированностью на уровне целых генов не выявило значимой корреляции (Рисунок 3.4б). Анализ на уровне фрагментов отдельных генов показал наличие слабой связи: более

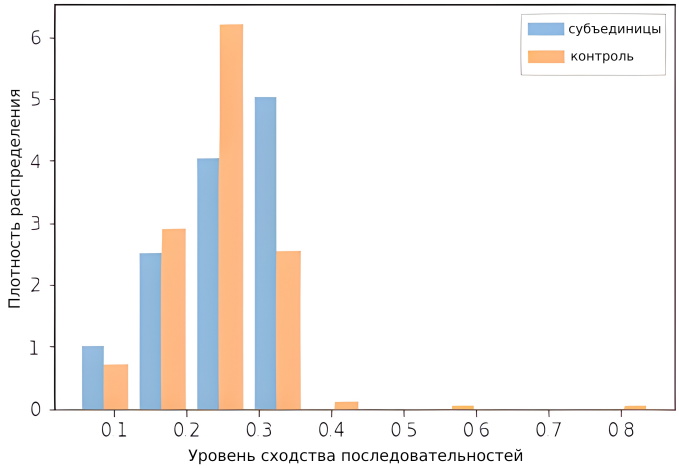

Рисунок 3.2 — Распределение уровней сходства пар генов, кодирующих субъединицы (голубой) и контрольных пар генов (оранжевый). Сходство последовательностей определено как вес локального выравнивания по Смиту-Ватерману, нормированный на сумму длин последовательностей. Выравнивания построены с штрафом за открытие разрыва -10 и штрафом за продолжение -1.

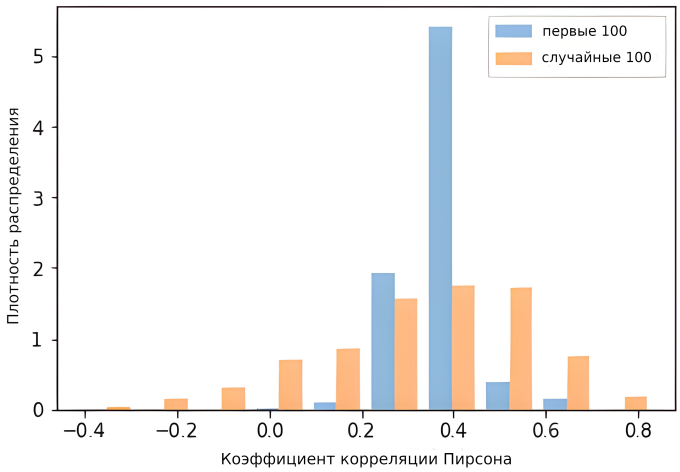

Рисунок 3.3 — Распределение коэффициентов корреляции структурированности для пар генов, кодирующих субъединицы одного белкового комплекса. Структурированность в данном случае посчитана по данным DMS-seq (индекс Джини). Среди всех возможных пар субъединиц выбирался случайным образом такой набор пар, чтобы каждому белковому комплексу соответствовала ровно одна пара, и для этого набора пар генов считалась корреляция индексов Джини покрытий либо для первых 100 нуклеотидов, либо для случайных 100. Процедура выбора пар повторялась 5000 раз.

структурированные участки одной мРНК деградируют медленне, чем менее структурированные. Это было показано следующим образом: каждый ген длины более 400 нуклеотидов был поделен на фрагменты по 100 нуклеотидов и была посчитана корреляция между уровнем структурированности и временем жизни фрагмента. Из-за малого числа точек большинство таких корреляций были незначимы, однако оказалось, что для 211 генов она была положительной, а для 161 гена – отрицательной; различие статистически значимо на уровне $p\text{-}value = 0.02$ (биномиальный тест).

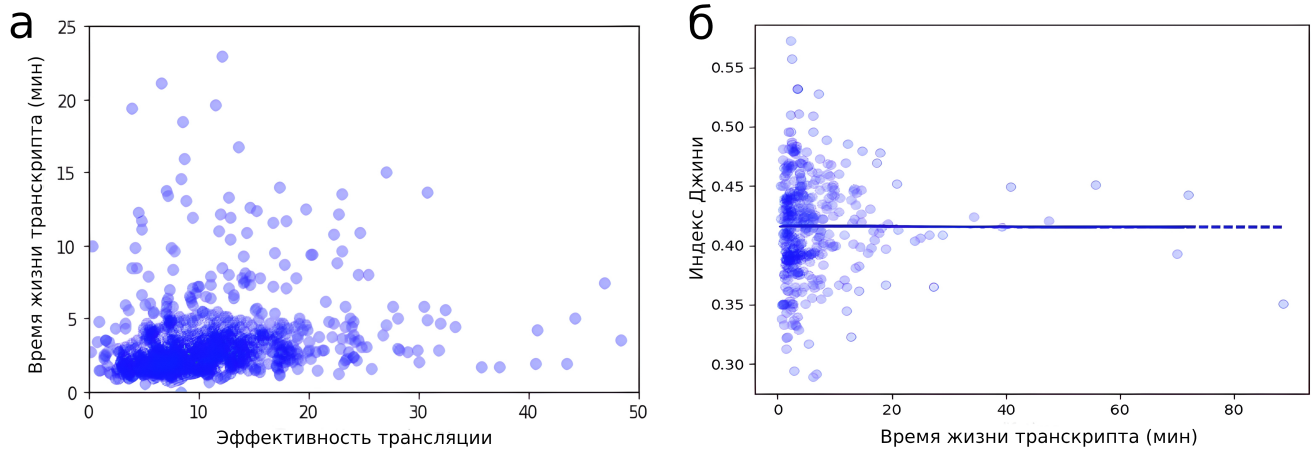

Рисунок 3.4 — (а) Время жизни транскрипта слабоположительно коррелирует с эффективностью его трансляции. Коэффициент корреляции Пирсона $r = 0.41$, $p\text{-}value = 0.014$. **(б) Время жизни транскрипта не связано с общей степенью его структурированности.** Структурированность гена была определена как индекс Джини покрытия гена по данным DMS-seq.

В исследовании [45] были показаны консервативные структуры, защищающие некоторые из генов в составе сложных полицистронных транскриптов от деградации. Вместе с тем известно, что структуры определенного типа (длинные шпильки), напротив, привлекают эндонуклеазы и необходимы для разрезания и последующей деградации [42; 115]. В нашем подходе эти два типа структур никак не различались, что может отчасти быть причиной слабости наблюдаемого эффекта.

Тем самым, в части случаев структура, по-видимому, может локально защищать мРНК от деградации, возможно, являясь препятствием для нуклеаз. Однако среднее время жизни транскрипта, судя по всему, определяется другими факторами, такими как его локализация в клетке [116; 117].

3.2.3 Частоты замен спаренных нуклеотидов

Далее мы посмотрели, как связана структура мРНК и эволюция последовательности. Для этого были рассмотрены полиморфизмы в 300 штаммах кишечной палочки. Оказалось, что полиморфизмы более вероятны в структурированных областях (Рисунок 3.5а). При этом с увеличением вероятности структурированности позиции растет число замен C=>T и, в меньшей степени, C=>A, и уменьшается число замен A=>G (Рисунок 3.5б). Тем самым, GC-состав структурированных областей со временем должен уменьшаться, что согласуется со старыми представлениями об избегании выраженной структуры в транслируемых областях. В то же время, следует отметить, что эти результаты не воспроизводятся на данных SHAPE (данные не показаны). Это может означать, что мутации происходят в определенных контекстах (т. наз. мутационные подписи; С. Гарушянц, частное сообщение), и похожие контексты способствуют химическим модификациям в методе DMS-Seq.

Таким образом, наблюдение о большей частоте мутаций в спаренных областях может являться артефактом используемого метода и требует дальнейших независимых подтверждений.

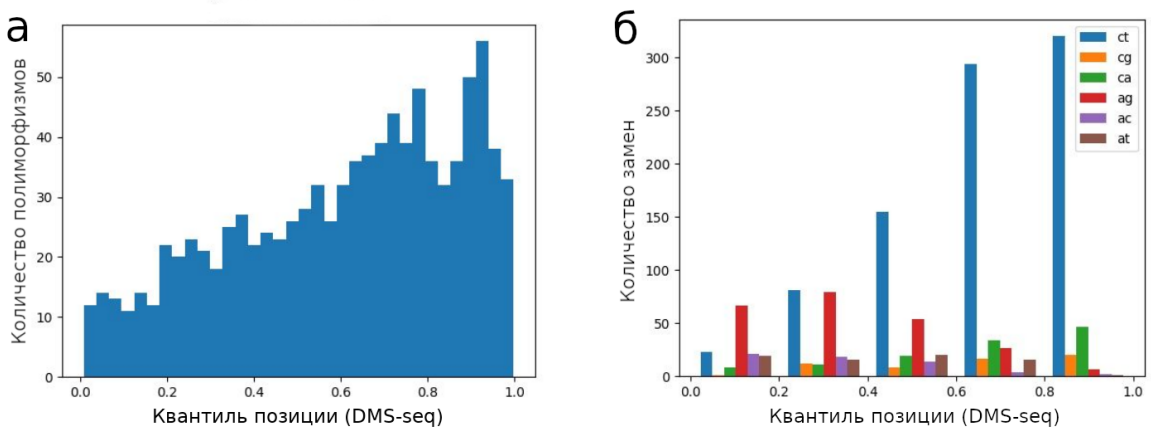

Рисунок 3.5 — (а) Количество полиморфизмов и вероятность спаривания позиции. Горизонтальная ось соответствует квантилю позиции по структурированности внутри своего гена по данным DMS-seq у *Escherichia coli* str. K12 substr. MG1655. Чем ближе квантиль к единице, тем с большей вероятностью эта позиция спарена. Вертикальная ось соответствует количеству полиморфизмов, наблюдаемых у 300 штаммов *Escherichia coli* в позициях с такой степенью структурированности. В случае, если бы полиморфизмы были распределены независимо от структуры мРНК, мы бы ожидали увидеть на этом графике равномерное распределение. **(б) Вероятности замен разных типов в позициях с различной вероятностью спаривания.** Цветовые обозначения типов замен указаны в легенде на рисунке. Замены поляризованы сопоставлением с *Escherichia fergusonii*

Глава 4. Роль вторичной структуры в редактировании мРНК у головоногих моллюсков

Самый часто встречающийся тип редактирования мРНК – это дезаминирование аденина. При этом аденин превращается в инозин, который впоследствии воспринимается аппаратом трансляции как гуанин, и таким образом, редактирование мРНК может изменять последовательность закодированного белка. При этом редактирование редко бывает абсолютным, поэтому чаще всего в клетке встречаются как отредактированные, так и неотредактированные транскрипты.

Известно, что редактирование мРНК осуществляется аденозин деаминазами ADAR, для работы которых необходимо образование двойной спирали РНК [2; 83]. Поэтому представляло интерес изучить особенности вторичной структуры мРНК в окрестности редактируемых аденинов. Кроме того, сайты редактирования часто кластеризуются, и было интересно изучить, какую роль в этом может играть вторичная структура мРНК. Дальнейшее изложение следует нашим работам [2; 3], оттуда же взяты рисунки.

4.1 Материалы и методы

4.1.1 Данные

Для исследования были использованы секвенированные транскриптомы и геномы двух осьминогов *Octopus vulgaris* и *Octopus bimaculoides*, кальмара *Loligo pealei* и каракатицы *Sepia esculenta*. В качестве внешних видов (outgroups) были взяты головоногий моллюск (не колеоид) *Nautilus pompilius* и брюхоногий моллюск морской заяц *Aplysia californica* [83]. Предварительный анализ сырых данных: картирование, определение сайтов и уровня их редактирования (УР), выравнивание гомологичных последовательностей – был прове-

ден М.А. Молдованом. Уровень редактирования был определен как доля ридов с G, картируемых на (редактируемый) A.

4.1.2 Сопоставление сайтов между видами

Позициям в транскриптоме каждого вида колеоидов путем выравнивания были сопоставлены позиции в других транскриптомах. Для каждой пары видов проделывалась следующая процедура: один из видов выбирался в качестве опорного, и для каждого транскрипта из опорного вида запускался BLASTn с порогом на *E-value* в 10^{-15}. Полученные выравнивания использовались для сопоставления позиций. Выбор тех или иных опорных видов оказывал пренебрежимо малое влияние на представленные далее результаты (данные не приведены).

4.1.3 Оценка стабильности вторичной структуры

Для оценки выраженности вторичной структуры использовали программу RNASurface [118; 119]. При этом *z-score* фрагмента последовательности определялся как $z = (E - \mu)/\sigma$, где E – минимальная свободная энергия фрагмента, μ и σ – соответственно, среднее и среднеквадратичное отклонение распределения энергии для случайных последовательностей той же длины и с тем же динуклеотидным составом. Для анализа использовались фрагменты минимальной длины 20 нт и максимальной – 350 нт. Каждой позиции транскрипта сопоставлялся минимальный *z-score* всех фрагментов, содержащих эту позицию, если эта величина составляла –2 или менее, в противном случае позиции сопоставлялось нулевое значение. Тем самым, в каждом транскрипте были выделены структурированные и неструктурированные области, причем в структурированных областях каждой позиции был приписан *z-score*. Для оценки различий структурного потенциала при заменах A на G учитывали только позиции, в которых модуль разницы *z-score* превосходил 2.

4.1.4 Оценка сближенности пар редактируемых сайтов в структуре мРНК

Для каждого транскрипта при помощи программы plfold из пакета ViennaRNA [97] были посчитаны вероятности спаривания нуклеотидов. Эта программа дает усредненные по всем возможным структурам (с учетом их энергий) вероятности для пар нуклеотидов образовать водородные связи. Значениями параметров $-W$ и $-L$ программы plfold были установленных равными длине анализируемого транскрипта, а значение параметра $-$cutoff было равно 0.0.

Далее для каждой последовательности был построен граф, в котором вершинами были нуклеотиды, а ребра проводились между теми парами вершин, которые либо являлись соседними в последовательности (ребра первого типа), либо, согласно предсказанию plfold, были спарены с вероятностью большей либо равной 0.8 (ребра второго типа). Такие графы использовались для определения расстояния между нуклеотидами по структуре. *Расстояние по структуре* между двумя нуклеотидами определялось как минимальная длина пути в графе между вершинами, соответствующими этим нуклеотидам. *Расстояние по последовательности* определялось как разность номеров позиций.

При анализе сайтов редактирования, сближенных в пространстве благодаря вторичной структуре РНК, мы рассматривали для каждого транскрипта все возможные пары редактируемых сайтов, и каждая такая пара была отнесена в одну из трех групп: «сближенные благодаря структуре», «далекие, неструктурированные» или «промежуточные». Сближенными считались такие пары, для которых расстояние по структуре было меньше, чем расстояние по последовательности, и к тому же расстояние по структуре составляло не более восьми ребер. Далекими считались пары нуклеотидов, для которых расстояние по структуре было равно расстоянию по последовательности и при этом расстояние по последовательности было больше сорока ребер. Пары сайтов, не попадающие в описываемые два класса, были отнесены в группу промежуточных.

4.2 Результаты и обсуждение

4.2.1 Доля аденинов в структурированных областях одинакова для всех колеоидов

В этом исследовании мы использовали данные для шести разных видов, у четырех из которых высокий общий уровень редактирования (Рисунок 4.1).

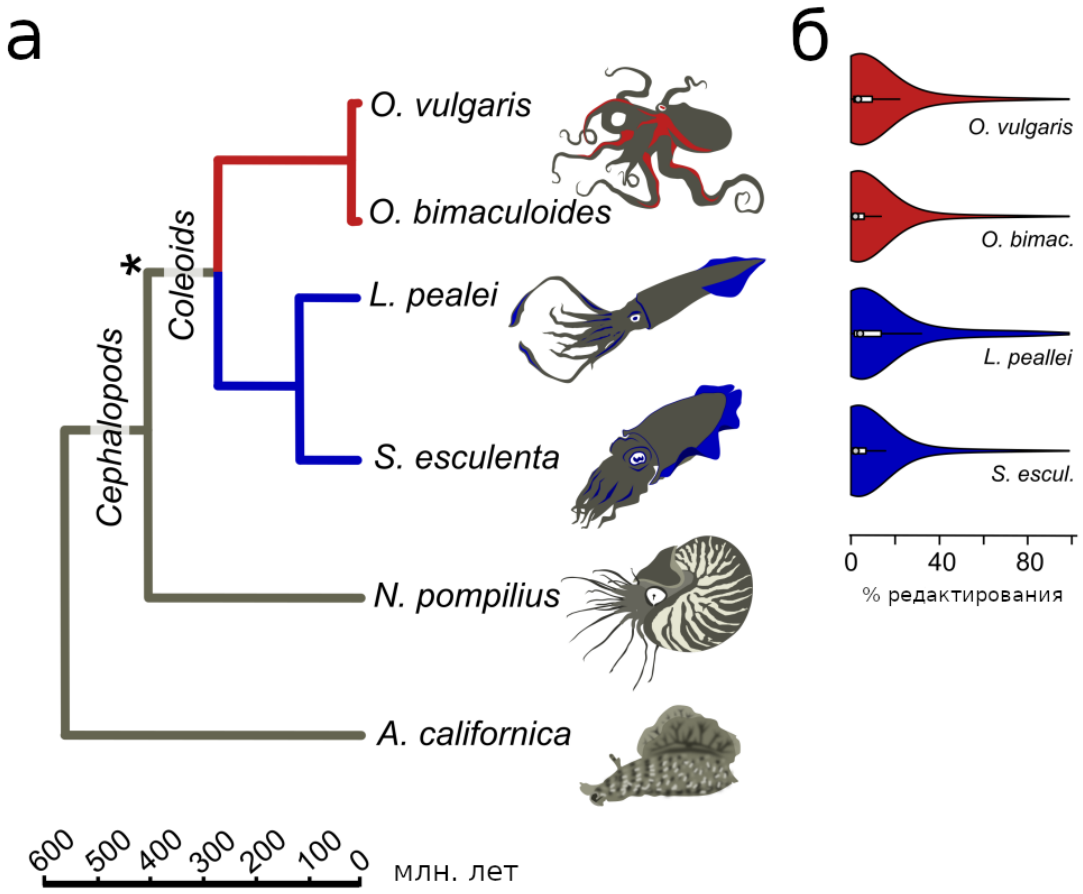

Рисунок 4.1 — Уровень редактирования мРНК у колеоидов. (а) Филогенетическое дерево видов. Звездочкой обозначен предполагаемый момент, когда редактирование транскриптов стало массовым. **(б)** Расределения уровней редактирования для отдельных редактируемых аденинов (% I в позиции редактируемого аденина при картировании фрагментов транскриптома на геномную ДНК). При интерпретации распределений следует иметь в виду, что слаборедактируемые сайты слабоэспрессирующихся генов могут быть не детектированы из-за недостаточного покрытия.

Для всех шести рассмотренных видов были построены структурированные сегменты (см. 3.1.3. «Оценка стабильности вторичной структуры»). Как видно на Рисунке 4.2, общая доля аденинов, находящихся в структурирован-

ных участках, одинакова у всех пяти исследованных головоногих моллюсков, при том что транскриптом не-колеоида *N. pompilius* слабо подвержен редактированию. Это доказывает, что редактирование мРНК у колеоидов не вызвано общим повышением структурированности вследствие сторонних факторов (например, изменения нуклеотидного состава).

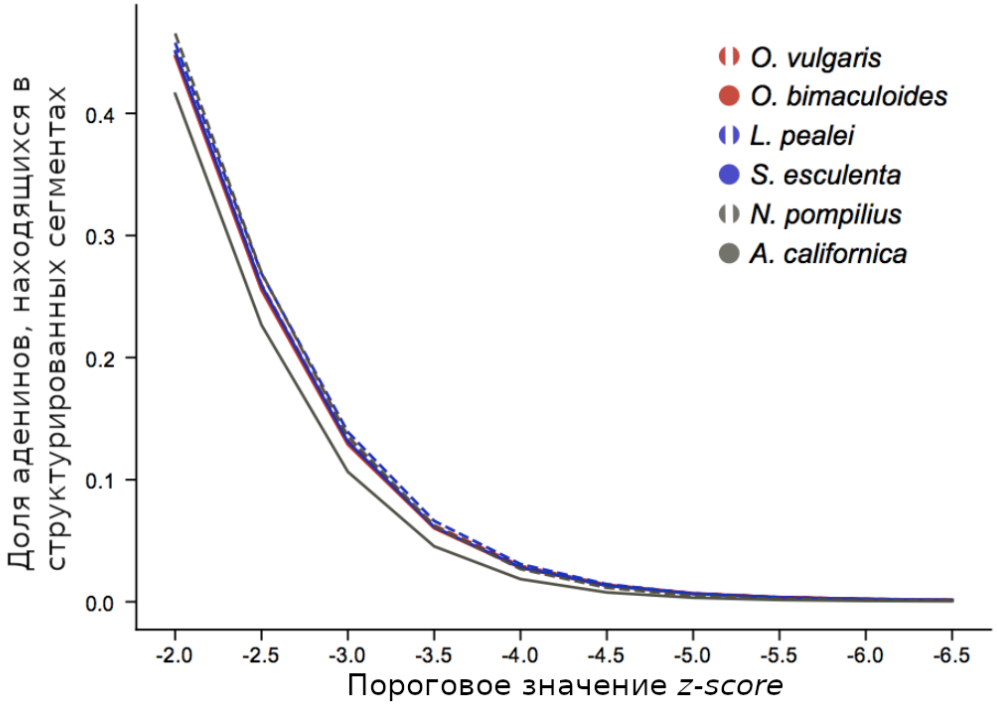

Рисунок 4.2 — Доля аденинов, находящихся в структурированных сегментах, в зависимости от порога на *z-score*. У морского зайца (*A. californica*) эта доля меньше, чем у других видов.

4.2.2 Аденины в структурированных областях редактируются чаще и более интенсивно, чем в неструктурированных

Чтобы оценить вклад вторичной структуры РНК в редактирование, мы сравнили доли аденинов, находящихся в структурированных участках, для выборок с разным уровнем редактирования. Как видно из Рисунка 4.3, редактируемые аденины чаще, чем нередактируемые, находятся в структурированных сегментах. Более того, доля аденинов в структурированных участках растет с увеличением уровня редактирования.

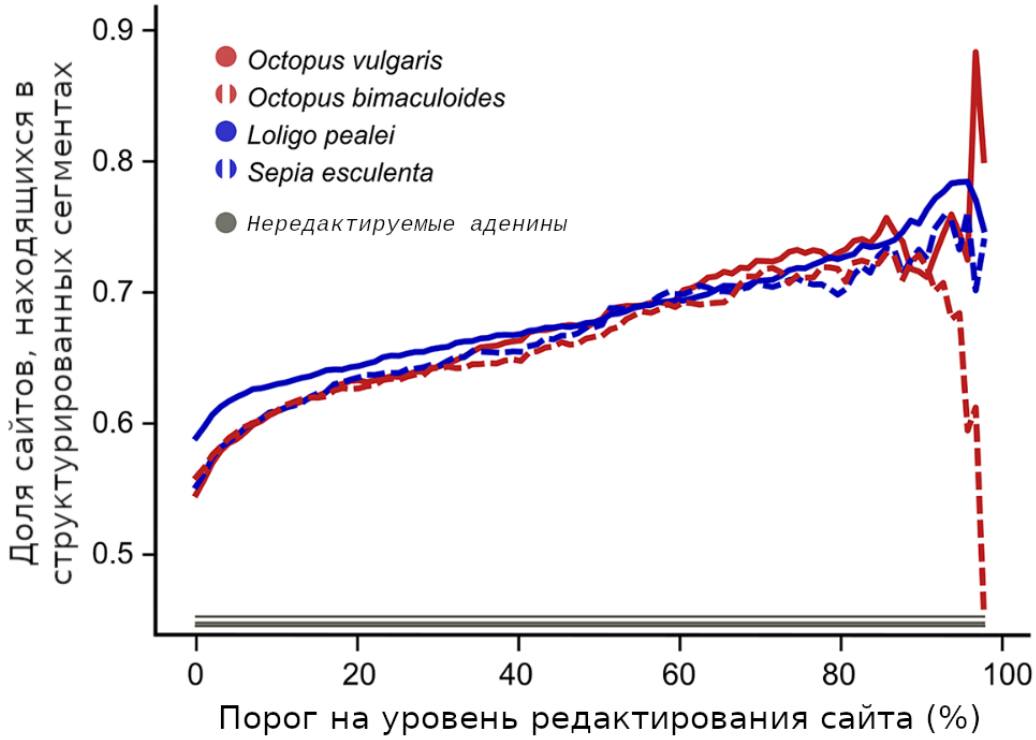

Рисунок 4.3 — Доля редактируемых аденинов, лежащих в структурированных сегментах, растет с увеличением уровня редактирования (горизонтальная ось). Доля нередактируемых аденинов, лежащих в структурированных областях, составляет $\sim$ 0,45. Колебания в области больших значений уровня редактирования связаны с малым числом сильно (>90%) редактируемых аденинов.

4.2.3 Консервативно редактируемые аденины чаще находятся в структурированных областях, чем неконсервативно редактируемые

Как уже было упомянуто ранее, многие сайты редактирования консервативны между колеоидами. Мы собрали три набора сайтов: консервативные между двумя близкими видами осьминогов, консервативные между более далекой парой кальмар –каракатица, и консервативные между всеми четырьмя видами. Каждый набор сайтов мы разбили на категории по наименьшему в рассматриваемых видах уровню их редактирования и для каждой категории каждого набора посчитали долю сайтов, находящихся в структурированных участках (Рисунок 4.4). Оказалось, что более консервативные сайты (редактирование которых сохраняется между более далекими видами), в четырех из пяти категорий чаще находятся в структурированных участках, чем менее кон-

сервативные. Различие статистически значимо в двух категориях с наименьшим уровнем редактирования.

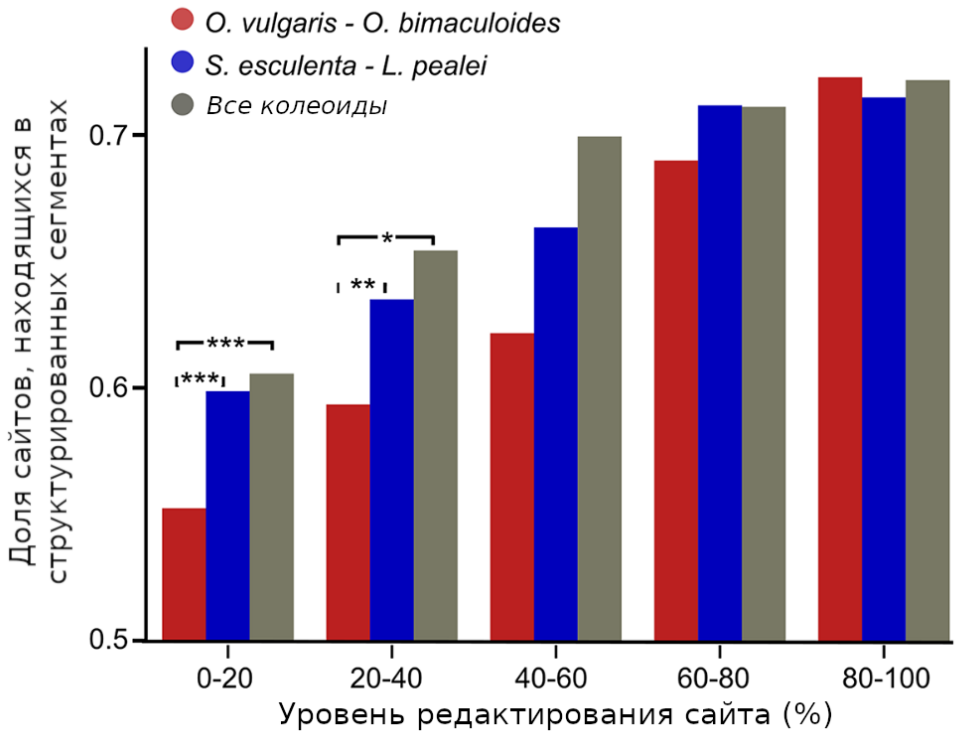

Рисунок 4.4 — Доля редактируемых аденинов, лежащих в структурированных сегментах, выше для консервативно редактируемых аденинов. Горизонтальная ось – порог на минимальный уровень редактирования. Красные столбцы: аденины, консервативно редактируемые у осьминогов, синие – у кальмара и каракатицы, серые – у всех четырех колеоидов. Статистическая значимость была оценена путем генерации пар распределений долей по подвыборкам из исходных выборок (bootstrap). *** $p\text{-}value < 0.001$, * $p\text{-}value < 0.05$

4.2.4 Разница в уровнях редактирования между гомологичными сайтами близких видов связана с разной степенью структурированности этих сайтов

В процессе работы с данными об уровнях редактирования мы обнаружили, что гомологичные сайты даже в близких видах осьминогов могут иметь существенно разные уровни редактирования. Чтобы изучить связь этого различия со структурой, для каждого значения порога мы взяли все такие пары гомологичных сайтов, что уровень их редактирования различается больше, чем на пороговое значение, и посчитали по таким парам сайтов корреляцию

между разницей в уровне их редактирования и разницей в степени их структурированности (*z-score*) (Рисунок 4.5, сплошная линия). При низком пороге на разницу в уровне редактирования (5%) корреляции практически нет ($r = 0.1$, $p\text{-}value < 0.05$), в то время как если рассматривать только пары сайтов со значительной разницей в уровне редактирования (>50%), корреляция с разницей в структурированности оказывается значительной ($r = 0.7$, $p\text{-}value < 0.05$). Возможное объяснение заключается в том, что большие изменения уровня редактирования могут быть действительно связаны с изменением структуры, в то время как незначительные изменения могут быть следствием случайного шума.

Мы также посмотрели, как соотносятся уровень редактирования и различия в степени структурированности в парах гомологичных сайтов, из которых один редактируется, а другой нет. Для этого мы повторили для этой группы пар сайтов анализ, описанный ранее, с тем отличием, что уровень редактирования одного из сайтов в каждой паре теперь был равен нулю. В этом случае тенденция оказалась более слабой (Рисунок 4.5, пунктирная линия). Возможно, причина этого в том, что для превращения нередактируемого сайта в редактируемый одной только стабильной вторичной структуры недостаточно.

Мы также посмотрели распределение разницы уровня структурированности для всех таких пар сайтов (редактируемый аденин и нередактируемый гомологичный аденин) независимо от силы редактирования первого сайта (Рисунок 4.6). Редактируемый аденин значимо чаще оказывается в структурированной области, чем его нередактируемый гомолог.

Теоретически, редактирование может увеличивать структурный потенциал просто за счет увеличения локального GC-состава. Мы сопоставили структурный потенциал редактируемых аденинов в состояниях A и G, в качестве контроля служили нередактируемые аденины, вычислительно замененные на G. Как ожидалось, при любой замене A на G локальная структурированность увеличивается, однако этот эффект сильнее для редактируемых аденинов (Рисунок 4.7). Что особенно интересно, для пары осьминогов, т.е. на небольшом эволюционном расстоянии, этот эффект значимо сильнее для редактируемых

аденинов, гомологами которых является гуанин (самый правый боксплот, тест Вилкоксона, $p\text{-}value < 10^{-5}$). Это может быть свидетельством преадаптации (другие свидетельства обсуждаются в [2]).

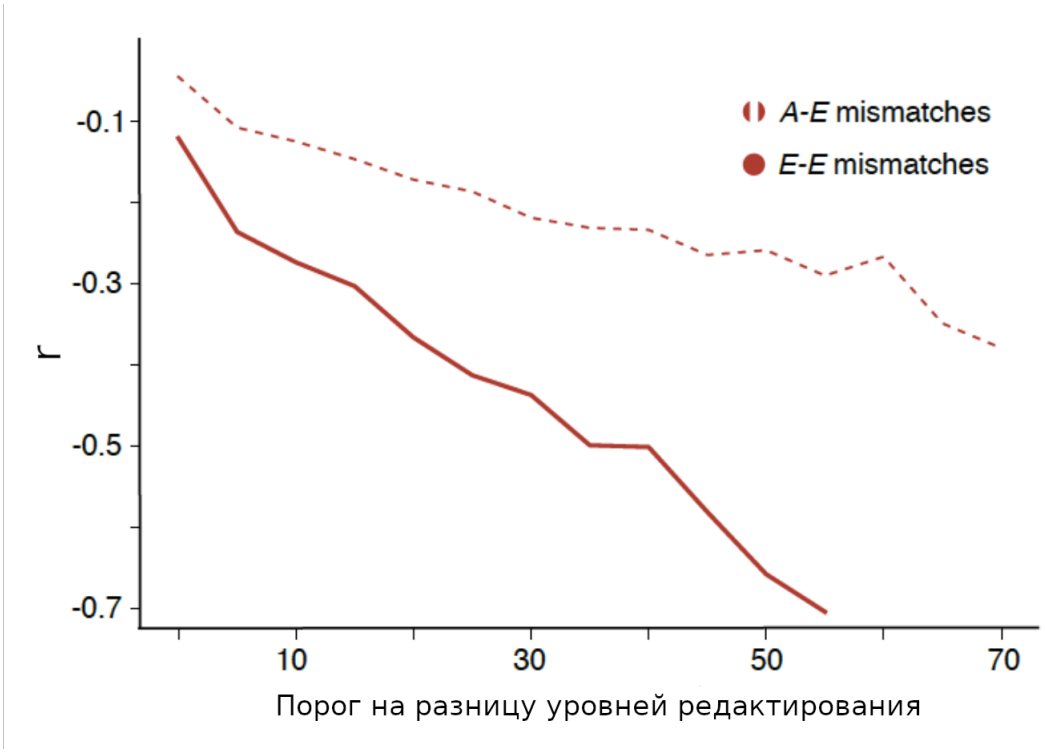

Рисунок 4.5 — Коэффициент корреляции Пирсона между разницей в степени структурированности (*z-score*) и разницей в уровне редактирования при разных значениях нижнего порога на разницу в уровне редактирования. Сплошной линией показаны значения для пар гомологичных сайтов, редактируемых в обоих видах осьминогов. Пунктирной линией показаны значения для пар гомологичных сайтов, из которых один редактируется, а другой нет. Коэффициенты корреляции показаны только для тех значений порога, где *p-value* для этого значения коэффициента корреляции не превышает 0.05.

4.2.5 Длина структурированного участка вокруг сайта, в среднем, не превышает 50 нуклеотидов

Чтобы оценить размер структуры мРНК вокруг редактируемых сайтов, мы проанализировали вероятности спаривания окружающих их нуклеотидов. Для этого было использована программа plfold из пакета ViennaRNA [97], которая для каждой пары нуклеотидов в последовательности выдает усредненную по всем возможным структурам (с учетом их энергий) вероятность быть спаренными. Каждому нуклеотиду мы сопоставили его суммарную вероятность быть

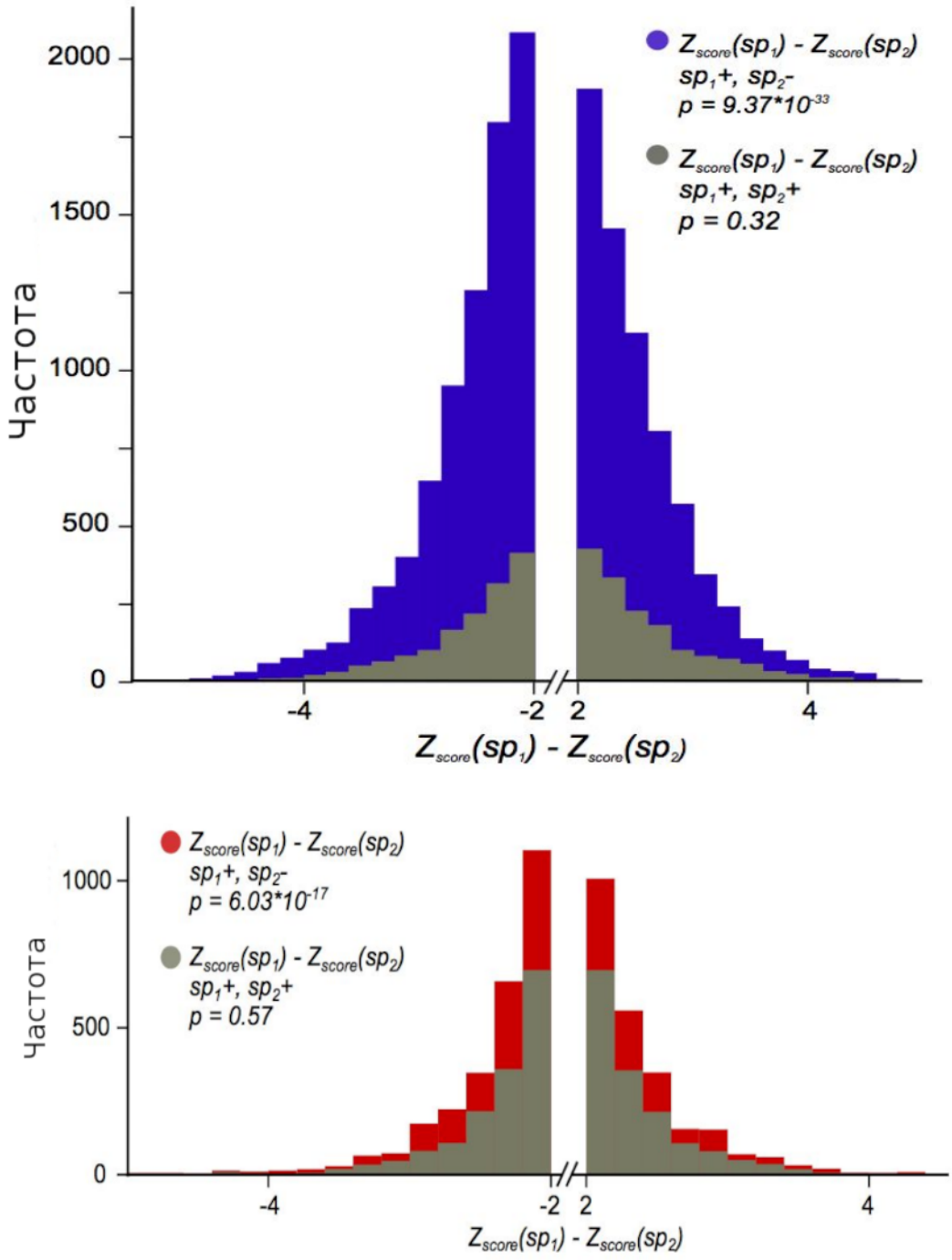

Рисунок 4.6 — Разница структурных потенциалов редактируемых и гомологичных им нередактируемых аденинов для пары осьминогов (синий, сверху) и пары кальмар–каракатица (красный, снизу). Серым показаны аналогичные гистограммы для пар, в которых оба аденина редактируются. В обоих случаях левый хвост значимо тяжелее правого, что свидетельствует о большей склонности редактируемых аденинов находиться в структурированных областях по сравнению с гомологичными им нередактируемыми аденинами.

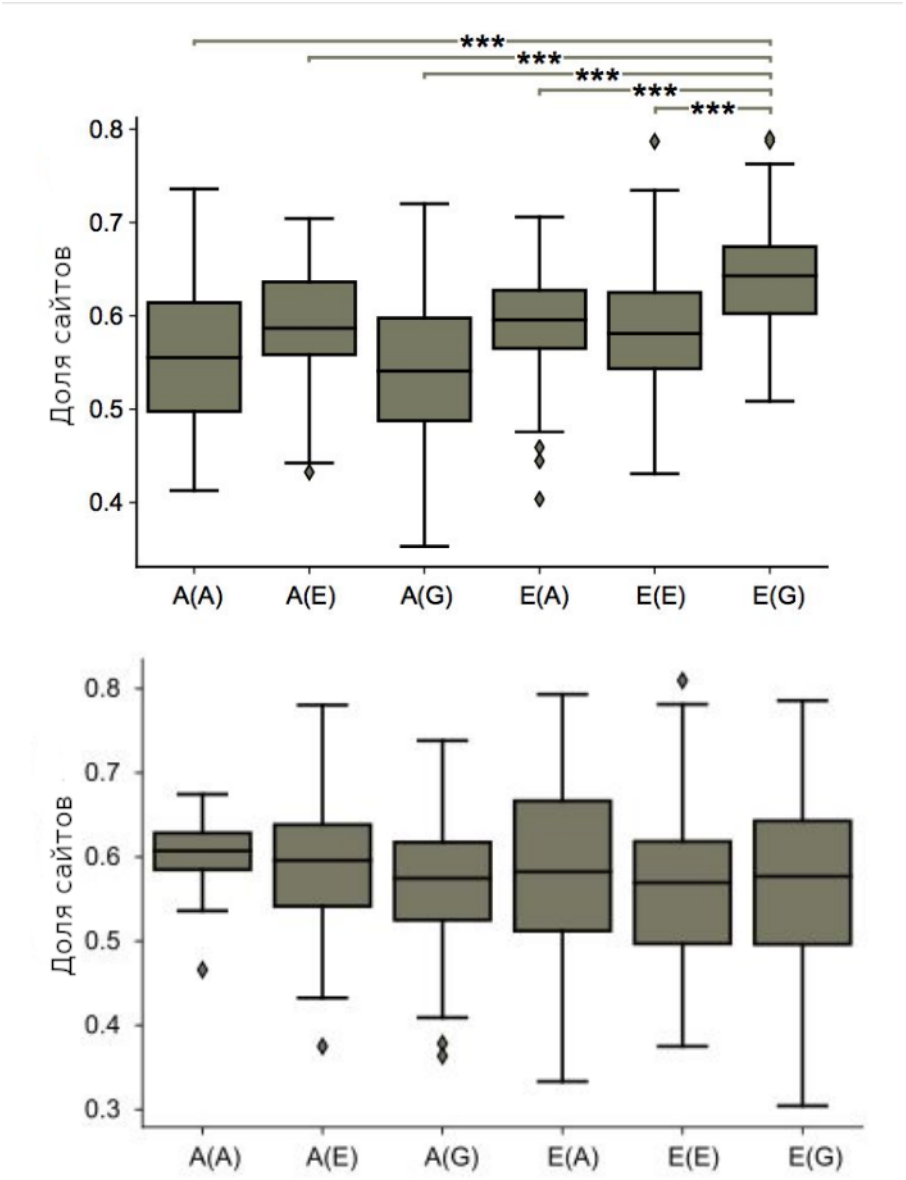

Рисунок 4.7 — Локальное увеличение структурного потенциала при заменах аденина на гуанин для редактируемых (E, три боксплота справа) и нередактируемых (A, три панели слева) аденинов. В скобках – гомологичный нуклеотид (A – нередактируемый аденин, E – редактируемый аденин, G – гуанин). Значимые эффекты наблюдаются только для пары осьминогов (верхняя панель), но не для пары кальмар–каракатица (нижняя панель).

вовлеченным в какие-либо структурные взаимодействия, сложив вероятности для всех пар, включающих данный нуклеотид. Результат этого анализа показан на Рисунке 4.8). Видно, что по мере удаления от сайта редактирования, средняя вероятность находиться в спаренном состоянии падает, причем вне участка ±25 нуклеотидов вокруг сайта колебания вероятности не отличимы от шума.

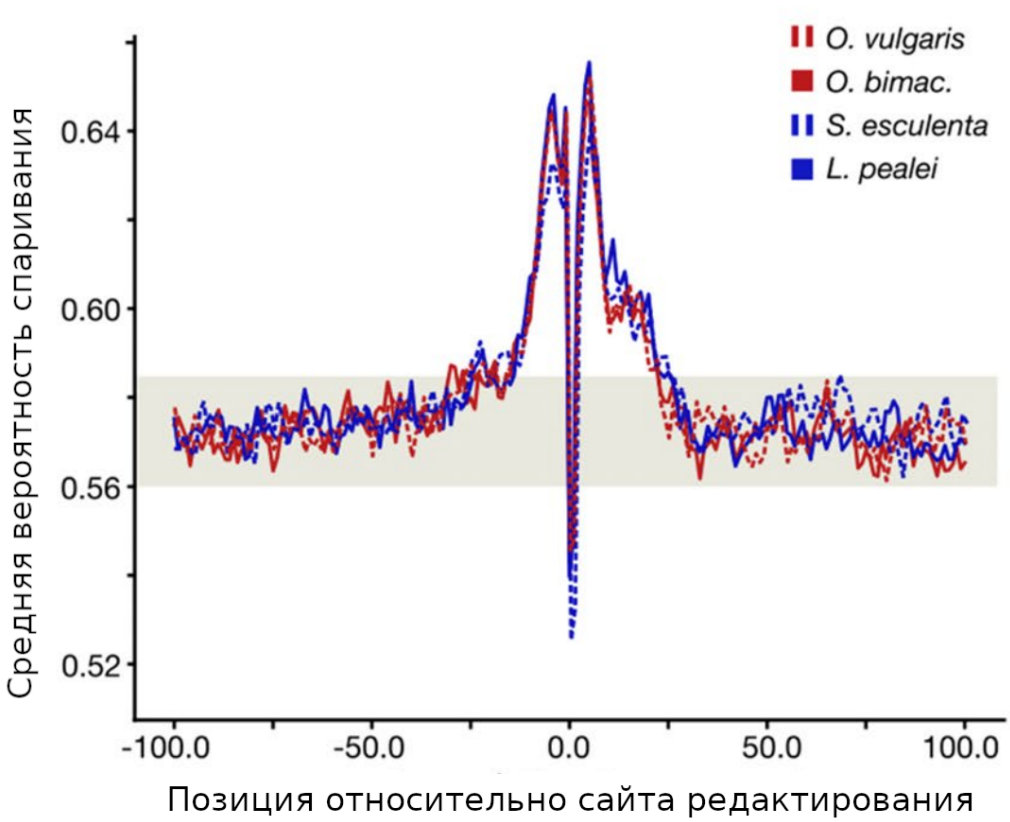

Рисунок 4.8 — Средняя вероятность спаривания в участках мРНК, окружающих сайты редактирования, для четырех видов колеоидов. Серая полоса маркирует диапазон значений, принимаемых вероятностями в значительно отдаленных позициях (>200 нт), эти значения предположительно соответствуют шуму. Значения выше серой полосы (по центру графика) соответствуют усредненной вторичной структуре мРНК вокруг сайта редактирования, ширина центрального пика соответствует предполагаемому размеру структуры. Провал по центру графика, вероятно, связан с низкой вероятностью самого редактируемого аденина быть спаренным в структуре.

4.2.6 Сайты, сближенные в структуре, чаще редактируются одновременно

Аденины, находящиеся близко в последовательности, часто редактируются одновременно [3]. Однако благодаря вторичной структуре могут оказаться сближены и сайты, находящиеся на значительном расстоянии друг от друга. Чтобы проверить, связана ли пространственная близость со скоординированностью редактирования, мы рассмотрели две группы пар сайтов: сближенные благодаря структуре и далекие друг от друга (см. 3.1.4. «Оценка сближенности пар редактируемых сайтов в структуре мРНК»).

В качестве меры скоординированности редактирования в паре сайтов (A_i, A_j) мы использовали значение r', определяемое как $r'(A_i, A_j) = f_{i,j}/(f_i f_j)$, где $f_{i,j}$ — частота редактирования обоих сайтов одновременно, а f_i и f_j — общие частоты редактирования для каждого сайта по отдельности. Оказалось, что сайты редактирования, сближенные благодаря структуре, редактируются одновременно значимо чаще, чем контрольные сайты, находящиеся на том же расстоянии по последовательности ($p\text{-}value = 7.8 \times 10^{-7}$, тест Манна-Уитни) (Рисунок 4.9). Этот результат показывает, что несмотря на в среднем низкую стабильность (вероятность спаривания на расстояниях более 25 нуклеотидов не отличима от шума, см. Рисунок 4.8), вторичная структура не только определяет возможность и силу редактирования, но и позволяет скоординировать далекие события редактирования между собой.

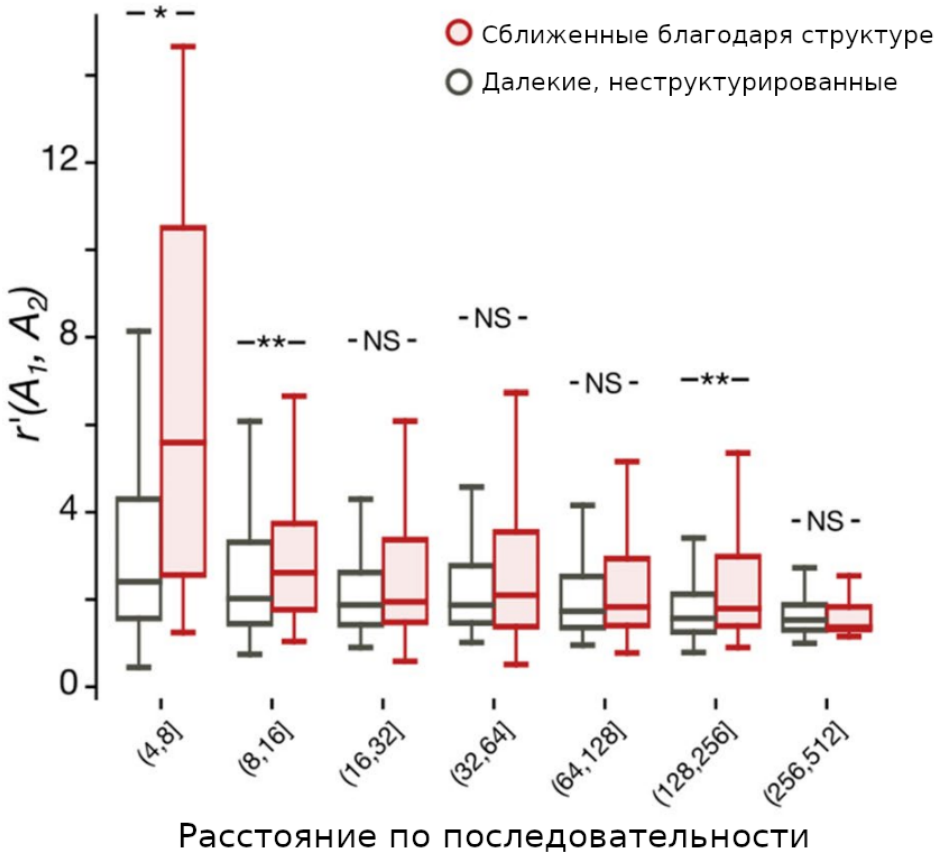

Рисунок 4.9 — Распределения значений r' для пар сайтов, сближенных благодаря вторичной структуре РНК (красные боксы), и для пар сайтов, далеких друг от друга и структурой не сближенных (серые боксы). По горизонтальной оси отложено расстояние между парами сайтами по последовательности. Звездочками помечена статистическая значимость различия согласно тесту Манна-Уитни с поправкой Бонферрони на множественное тестирование. ** $p\text{-}value < 0.01$, * $p\text{-}value < 0.05$

Заключение

Вторичная структура мРНК влияет на множество процессов, которые с этой мРНК происходят. В этой работе мы наблюдали влияние структуры на эффективность трансляции у *Exchericia coli*, а также связь структуры со степенью редактирования мРНК у головоногих моллюсков. Вторичная структура может играть регуляторную функцию, способствуя более тонкой подстройке уровня экспрессии гена или же и вовсе изменяя конечную последовательность экспрессируемого белка. В случае головоногих моллюсков элементы вторичной структуры могут также обеспечивать согласованное редактирование мРНК в нескольких отдаленных друг от друга участках.

Эти функции вторичной структуры могут иметь значение для приспособления организма к среде и сохраняться в процессе эволюции, что может накладывать дополнительные ограничения на эволюцию кодирующих последовательностей. Поэтому анализ вторичной структуры мРНК может расширить понимание эффектов, вызываемых различными мутациями, и способствовать развитию как синтетической микробиологии, так и медицинской генетики.

Выводы

1. Высокая эффективность трансляции синтетических конструкций в клетках *Escherichia coli* требует слабой вторичной структуры в 5'-области и начале гена, а также отсутствия в начале гена последовательностей, афинных к анти-Шайна-Дальгарно участку 16S рРНК.
2. Гипотеза о положительном влиянии редких кодонов в начале гена на эффективность трансляции у бактерий не подтверждается при анализе данных эффективности трансляции в клетках *Escherichia coli* флуоресцентного белка со случайными вставками в начале гена.
3. Метаболическая стоимость закодированных аминокислот может влиять на эффективность трансляции в клетках *Escherichia coli* в среде с низким содержанием питательных веществ.
4. Не наблюдается корреляция между степенью структурированности мРНК и скоростью ее деградации.
5. Повышенная частота и типы полиморфизмов в структурированных областях мРНК зависят от того, каким методом определялась структурированность, и требуют дальнейшего исследования.
6. Равная эффективность трансляции эквимолярных субъединиц белковых комплексов в клетках *Escherichia coli* может достигаться за счет сравнимой стабильности вторичной структуры мРНК генов, кодирующих эти субъединицы.
7. Аденины в структурированных областях транскриптов колеоидов чаще подвергаются гидролитическому дезаминированию, чем аденины в неструктурированных областях. При этом консервативно редактируемые аденины также чаще находятся в структурированных областях, чем неконсервативно редактируемые.
8. Вторичная структура может участвовать в координации редактирования аденинов в транскриптах колеоидов. Сайты, сближенные благодаря структуре, чаще редактируются одновременно.

Благодарности

Автор благодарит Михаила Сергеевича Гельфанда за научное руководство и помощь в работе над диссертацией, Петра Владимировича Сергиева и Михаила Молдована за формулировку задач и предоставленные данные, Андрея Александровича Миронова за терпение, Асю Менделевич и Валентину Бурскую за психологическую поддержку при работе над текстом, а также свою семью за мотивацию и помощь в организации процесса.

Список литературы

1. Translation at first sight: the influence of leading codons / I. A. Osterman*, Z. S. Chervontseva*, S. A. Evfratov, A. V. Sorokina, V. A. Rodin, M. P. Rubtsova, E. S. Komarova, T. S. Zatsepin, M. R. Kabilov, A. A. Bogdanov, M. S. Gelfand, O. A. Dontsova, P. V. Sergiev // Nucleic Acids Research. — 2020. — Vol. 48, no. 12. — 6931 *Joint first authors.
2. Adaptive evolution at mRNA editing sites in soft-bodied cephalopods / M. Moldovan, Z. Chervontseva, G. Bazykin, M. S. Gelfand // PeerJ. — 2020. — Vol. 8: e10456.
3. A hierarchy in clusters of cephalopod mRNA editing sites / M. A. Moldovan, Z. S. Chervontseva, D. S. Nogina, M. S. Gelfand // Scientific Reports. — 2022. — Vol. 12, 1: 3447.
4. *Neveu M.*, *Kim H. J.*, *Benner S. A.* The "strong" RNA world hypothesis: fifty years old // Astrobiology. — 2013. — Vol. 13, no. 4. — P. 391–403.
5. *Gorodkin J.* RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods. Vol. 1097. Issue 800. — 2014. — P. 33–43.
6. *Zuker M.*, *Jaeger J. A.*, *Turner D. H.* A comparison of optimal and sub-optimal RNA secondary structures predicted by free energy minimization with structures determined by phylogenetic comparison. // Nucleic Acids Research. — 1991. — May. — Vol. 19, issue 10. — P. 2707.
7. *Yakovchuk P.*, *Protozanova E.*, *Frank-Kamenetskii M. D.* Base-stacking and base-pairing contributions into thermal stability of the DNA double helix // Nucleic Acids Research. — 2006. — Feb. — Vol. 34, issue 2. — P. 564.
8. *Mathews D. H.*, *Moss W. N.*, *Turner D. H.* Folding and Finding RNA Secondary Structure // Cold Spring Harbor Perspectives in Biology. — 2010. — Vol. 2, issue 12.

9. *Zuker M.*, *Stiegler P.* Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. // Nucleic Acids Research. — 1981. — Jan. — Vol. 9, issue 1. — P. 133.

10. *Batenburg F. H. V.*, *Gultyaev A. P.*, *Pleij C. W.* PseudoBase: structural information on RNA pseudoknots // Nucleic Acids Research. — 2001. — Jan. — Vol. 29, issue 1. — P. 194.

11. *Bellaousov S.*, *Mathews D. H.* ProbKnot: Fast prediction of RNA secondary structure including pseudoknots // RNA. — 2010. — Oct. — Vol. 16, issue 10. — P. 1870.

12. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes / M. S. Gelfand, A. A. Mironov, J. Jomantas, Y. I. Kozlov, D. A. Perumov // Trends in genetics : TIG. — 1999. — Nov. — Vol. 15, no. 11. — P. 439–442.

13. Comparative genomic analysis of T-box regulatory systems in bacteria / A. G. Vitreschak, A. A. Mironov, V. A. Lyubetsky, M. S. Gelfand // RNA. — 2008. — Vol. 14, no. 4. — P. 717.

14. RNA thermometers are common in alpha- and gamma-proteobacteria / T. Waldminghaus, A. Fippinger, J. Alfsmann, F. Narberhaus // Biological chemistry. — 2005. — Dec. — Vol. 386, no. 12. — P. 1279–1286.

15. *Righetti F.*, *Narberhaus F.* How to find RNA thermometers // Frontiers in Cellular and Infection Microbiology. — 2014. — Vol. 4, SEP.

16. *Tucker B. J.*, *Breaker R. R.* Riboswitches as versatile gene control elements // Current opinion in structural biology. — 2005. — Vol. 15, no. 3. — P. 342–348.

17. *Holcik M.*, *Sonenberg N.* Translational control in stress and apoptosis // Nature reviews. Molecular cell biology. — 2005. — Apr. — Vol. 6, no. 4. — P. 318–327.

18. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches / M. T. Cheah, A. Wachter, N. Sudarsan, R. R. Breaker // Nature. — 2007. — May. — Vol. 447, no. 7143. — P. 497–500.

19. Thiamine biosynthesis in algae is regulated by riboswitches / M. T. Croft, M. Moulin, M. E. Webb, A. G. Smith // Proceedings of the National Academy of Sciences of the United States of America. — 2007. — Dec. — Vol. 104, no. 52. — P. 20770–20775.

20. Riboswitch-dependent gene regulation and its evolution in the plant kingdom / S. Bocobza, A. Adato, T. Mandel, M. Shapira, E. Nudler, A. Aharoni // Genes & development. — 2007. — Nov. — Vol. 21, no. 22. — P. 2874–2879.

21. Conserved long-range base pairings are associated with pre-mRNA processing of human genes / S. Kalmykova, M. Kalinina, S. Denisov, A. Mironov, D. Skvortsov, R. Guigó, D. Pervouchine // Nature Communications 2021 12:1. — 2021. — Apr. — Vol. 12, no. 1. — P. 1–17.

22. *Peeri M.*, *Tuller T.* High-resolution modeling of the selection on local mRNA folding strength in coding sequences across the tree of life // Genome Biology. — 2020. — Vol. 21, no. 1. — P. 1–20.

23. *Itzkovitz S.*, *Hodis E.*, *Segal E.* Overlapping codes within protein-coding sequences // Genome research. — 2010. — Nov. — Vol. 20, no. 11. — P. 1582–1589.

24. Local absence of secondary structure permits translation of mRNAs that lack ribosome-binding sites / L. B. Scharff, L. Childs, D. Walther, R. Bock // PLoS genetics. — 2011. — June. — Vol. 7, no. 6.

25. *Studer S. M.*, *Joseph S.* Unfolding of mRNA secondary structure by the bacterial translation initiation complex // Molecular cell. — 2006. — Apr. — Vol. 22, no. 1. — P. 105–115.

26. Analysis of 11,430 recombinant protein production experiments reveals that protein yield is tunable by synonymous codon changes of translation initiation sites / B. K. Bhandari, C. S. Lim, D. M. Remus, A. Chen, C. van Dolleweerd, P. P. Gardner // PLoS Computational Biology. — 2021. — Oct. — Vol. 17, no. 10.

27. Coding-sequence determinants of gene expression in Escherichia coli / G. Kudla, A. W. Murray, D. Tollervey, J. B. Plotkin // Science (New York, N.Y.) — 2009. — Apr. — Vol. 324, no. 5924. — P. 255–258.

28. *Shabalina S. A.*, *Ogurtsov A. Y.*, *Spiridonov N. A.* A periodic pattern of mRNA secondary structure created by the genetic code // Nucleic acids research. — 2006. — Vol. 34, no. 8. — P. 2428–2437.

29. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast / S. Dvir, L. Velten, E. Sharon, D. Zeevi, L. B. Carey, A. Weinberger, E. Segal // Proceedings of the National Academy of Sciences of the United States of America. — 2013. — July. — Vol. 110, issue 30.

30. Composite effects of gene determinants on the translation speed and density of ribosomes / T. Tuller, I. Veksler-Lublinsky, N. Gazit, M. Kupiec, E. Ruppin, M. Ziv-Ukelson // Genome biology. — 2011. — Nov. — Vol. 12, no. 11.

31. Universally increased mRNA stability downstream of the translation initiation site in eukaryotes and prokaryotes / Y. Mao, W. Wang, N. Cheng, Q. Li, S. Tao // Gene. — 2013. — Apr. — Vol. 517, no. 2. — P. 230–235.

32. Stem-Loop Structures within mRNA Coding Sequences Activate Translation Initiation and Mediate Control by Small Regulatory RNAs // Molecular cell. — 2017. — Oct. — Vol. 68, no. 1. — 158–170.e3.

33. *Mitarai N.*, *Sneppen K.*, *Pedersen S.* Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization // Journal of molecular biology. — 2008. — Sept. — Vol. 382, no. 1. — P. 236–245.

34. Ribosome collisions trigger cis-acting feedback inhibition of translation initiation / S. Juszkiewicz, G. Slodkowicz, Z. Lin, P. Freire-Pritchett, S. Y. Peak-Chew, R. S. Hegde // eLife. — 2020. — July. — Vol. 9. — P. 1–29.

35. A novel translational control mechanism involving RNA structures within coding sequences / J. Jungfleisch, D. D. Nedialkova, I. Dotu, K. E. Sloan, N. Martinez-Bosch, L. Brüning, E. Raineri, P. Navarro, M. T. Bohnsack, S. A. Leidel, J. Díez // Genome Research. — 2017. — Jan. — Vol. 27, no. 1. — P. 95–106.

36. *Palma M.*, *Lejeune F.* Deciphering the molecular mechanism of stop codon readthrough // Biological reviews of the Cambridge Philosophical Society. — 2021. — Feb. — Vol. 96, no. 1. — P. 310–329.

37. Specificity of mRNA Folding and Its Association with Evolutionarily Adaptive mRNA Secondary Structures / G. Yu, H. Zhu, X. Chen, J. R. Yang // Genomics, Proteomics & Bioinformatics. — 2021. — Dec. — Vol. 19, no. 6. — P. 882–900.

38. *Deana A.*, *Belasco J. G.* Lost in translation: the influence of ribosomes on bacterial mRNA decay // Genes & development. — 2005. — Nov. — Vol. 19, no. 21. — P. 2526–2533.

39. Effect of ribosome shielding on mRNA stability / B. L. Bailey, K. Visscher, J. Watkins, N. Mitarai, S. Pedersen, C. Deneke, R. Lipowsky, A. Valleriani // Physical Biology. — 2013. — July. — Vol. 10, no. 4. — P. 046008.

40. Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. / D. H. Burkhardt, S. Rouskin, Y. Zhang, G.-W. Li, J. S. Weissman, C. A. Gross // eLife. — 2017. — Jan. — Vol. 6.

41. *Zur H.*, *Tuller T.* Strong association between mRNA folding strength and protein abundance in S. cerevisiae // EMBO Reports. — 2012. — Mar. — Vol. 13, no. 3. — P. 272.

42. The molecular mechanism of dsRNA processing by a bacterial Dicer / L. Jin, H. Song, J. E. Tropea, D. Needle, D. S. Waugh, S. Gu, X. Xinhua // Nucleic Acids Research. — 2019. — May. — Vol. 47, no. 9. — P. 4707–4720.

43. *Hur S.* Double-Stranded RNA Sensors and Modulators in Innate Immunity // Annual review of immunology. — 2019. — Apr. — Vol. 37. — P. 349–375.

44. *Hui M. P.*, *Foley P. L.*, *Belasco J. G.* Messenger RNA Degradation in Bacterial Cells // Annual Review of Genetics. — 2014. — Nov. — Vol. 48, no. 1. — P. 537–559.

45. *Dar D.*, *Sorek R.* Extensive reshaping of bacterial operons by programmed mRNA decay // PLOS Genetics / ed. by C. Buchrieser. — 2018. — Apr. — Vol. 14, no. 4. — e1007354.

46. A comprehensive review of signal peptides: Structure, roles, and applications / H. Owji, N. Nezafat, M. Negahdaripour, A. Hajiebrahimi, Y. Ghasemi // European Journal of Cell Biology. — 2018. — Vol. 97, no. 6. — P. 422–441.

47. *Saunders L. R.*, *Barber G. N.* The dsRNA binding protein family: critical roles, diverse cellular functions // The FASEB Journal. — 2003. — June. — Vol. 17, issue 9. — P. 961–983.

48. Transcriptome-wide identification of single-stranded RNA binding proteins / R. Zhao, X. Fang, Z. Mai, X. Chen, J. Mo, Y. Lin, R. Xiao, X. Bao, X. Weng, X. Zhou // Chemical Science. — 2023. — Apr. — Vol. 14, issue 15. — P. 4038–4047.

49. The Ribosome Uses Two Active Mechanisms to Unwind mRNA During Translation / X. Qu, J. D. Wen, L. Lancaster, H. F. Noller, C. Bustamante, I. Tinoco // Nature. — 2011. — July. — Vol. 475, no. 7354. — P. 118.

50. Ribosomal protein S1 unwinds double-stranded RNA in multiple steps / X. Qu, L. Lancaster, H. F. Noller, C. Bustamante, I. Tinoco // Proceedings of the National Academy of Sciences of the United States of America. — 2012. — Vol. 109, no. 36. — P. 14458–14463.

51. *Sørensen M. A.*, *Fricke J.*, *Pedersen S.* Ribosomal protein S1 is required for translation of most, if not all, natural mRNAs in Escherichia coli in vivo // Journal of molecular biology. — 1998. — July. — Vol. 280, no. 4. — P. 561–569.

52. *Milón P.*, *Rodnina M. V.* Kinetic control of translation initiation in bacteria // Critical reviews in biochemistry and molecular biology. — 2012. — July. — Vol. 47, issue 4. — P. 334–348.

53. *Nikolaeva D. D.*, *Gelfand M. S.*, *Garushyants S. K.* Simplification of Ribosomes in Bacteria with Tiny Genomes // Molecular Biology and Evolution. — 2021. — Jan. — Vol. 38, no. 1. — P. 58.

54. *Georgakopoulos-Soares I.*, *Parada G. E.*, *Hemberg M.* Secondary structures in RNA synthesis, splicing and translation // Computational and Structural Biotechnology Journal. — 2022. — Jan. — Vol. 20. — P. 2871–2884.

55. *Waudby C. A.*, *Dobson C. M.*, *Christodoulou J.* Nature and Regulation of Protein Folding on the Ribosome // Trends in Biochemical Sciences. — 2019. — Nov. — Vol. 44, issue 11. — P. 914–926.

56. *Caniparoli L.*, *O'Brien E. P.* Modeling the effect of codon translation rates on co-translational protein folding mechanisms of arbitrary complexity // The Journal of chemical physics. — 2015. — Apr. — Vol. 142, issue 14.

57. Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences / A. E. Borujeni, D. Cetnar, I. Farasat, A. Smith, N. Lundgren, H. M. Salis // Nucleic acids research. — 2017. — May. — Vol. 45, no. 9. — P. 5437–5448.

58. A possible universal role for mRNA secondary structure in bacterial translation revealed using a synthetic operon / Y. Chemla, M. Peeri, M. L. Heltberg, J. Eichler, M. H. Jensen, T. Tuller, L. Alfonta // Nature Communications. — 2020. — Dec. — Vol. 11, issue 1.

59. Influences on gene expression in vivo by a Shine-Dalgarno sequence / H. Jin, Q. Zhao, E. I. Gonzalez De Valdivia, D. H. Ardell, M. Stenström, L. A. Isaksson // Molecular microbiology. — 2006. — Apr. — Vol. 60, no. 2. — P. 480–492.

60. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing / A. M. Mustoe, S. Busan, G. M. Rice, C. E. Hajdin, B. K. Peterson, V. M. Ruda, N. Kubica, R. Nutiu, J. L. Baryza, K. M. Weeks // Cell. — 2018. — Vol. 173, issue 1. — 181–195.e18.

61. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate / T. E. Gorochowski, Z. Ignatova, R. A. L. Bovenberg, J. A. Roubos // Nucleic Acids Research. — 2015. — Vol. 43, issue 6. — P. 3022–3032.

62. Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function / C. D. Campo, A. Bartholomäus, I. Fedyunin, Z. Ignatova // PLoS Genetics. — 2015. — Vol. 11, issue 10. — P. 1–23.

63. An evolutionarily conserved mechanism for controlling the efficiency of protein translation / T. Tuller, A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman, Y. Pilpel // Cell. — 2010. — Vol. 141, no. 2. — P. 344–354.

64. A short translational ramp determines the efficiency of protein synthesis / M. Verma, J. Choi, K. A. Cottrell, Z. Lavagnino, E. N. Thomas, S. Pavlovic-Djuranovic, P. Szczesny, D. W. Piston, H. S. Zaher, J. D. Puglisi, S. Djuranovic // Nature Communications 2019 10:1. — 2019. — Dec. — Vol. 10, issue 1. — P. 1–15.

65. Translational Control by Ribosome Pausing in Bacteria: How a Non-uniform Pace of Translation Affects Protein Production and Folding / E. Samatova, J. Daberger, M. Liutkute, M. V. Rodnina // Frontiers in Microbiology. — 2021. — Vol. 1.

66. *Li G. W.*, *Oh E.*, *Weissman J. S.* The anti-Shine–Dalgarno sequence drives translational pausing and codon choice in bacteria // Nature 2012 484:7395. — 2012. — Mar. — Vol. 484, no. 7395. — P. 538–541.

67. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling / F. Mohammad, C. J. Woolstenhulme, R. Green, A. R. Buskirk // Cell reports. — 2016. — Feb. — Vol. 14, no. 4. — P. 686–694.

68. *Mohammad F.*, *Green R.*, *Buskirk A. R.* A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution // eLife. — 2019. — Feb. — Vol. 8.

69. Following translation by single ribosomes one codon at a time / J. D. Wen, L. Lancaster, C. Hodges, A. C. Zeri, S. H. Yoshimura, H. F. Noller, C. Bustamante, I. Tinoco // Nature. — 2008. — Apr. — Vol. 452, issue 7187. — P. 598.

70. RNA editing and alternative splicing: The importance of co-transcriptional coordination / J. Laurencikiene, A. M. Källman, N. Fong, D. L. Bentley, M. Öhman // EMBO Reports. — 2006. — Mar. — Vol. 7, issue 3. — P. 303–307.

71. *Slotkin W.*, *Nishikura K.* Adenosine-to-inosine RNA editing and human disease // Genome Medicine. — 2013. — Nov. — Vol. 5, issue 11. — P. 1–13.

72. *Walkley C. R.*, *Li J. B.* Rewriting the transcriptome: adenosine-to-inosine RNA editing by ADARs // Genome Biology 2017 18:1. — 2017. — Oct. — Vol. 18, issue 1. — P. 1–13.

73. *Bar-Yaacov D.*, *Pilpel Y.*, *Dahan O.* RNA editing in bacteria: occurrence, regulation and significance // RNA biology. — 2018. — July. — Vol. 15, issue 7. — P. 863–867.

74. A-to-I RNA editing in bacteria increases pathogenicity and tolerance to oxidative stress / W. Nie, S. Wang, R. He, Q. Xu, P. Wang, Y. Wu, F. Tian, J. Yuan, B. Zhu, G. Chen // PLoS Pathogens. — 2020. — Aug. — Vol. 16, issue 8.

75. *Eggington J. M.*, *Greene T.*, *Bass B. L.* Predicting sites of ADAR editing in double-stranded RNA // Nature communications. — 2011. — Vol. 2, issue 1.

76. *Wong S. K.*, *Sato S.*, *Lazinski D. W.* Substrate recognition by ADAR1 and ADAR2. // RNA. — 2001. — Vol. 7, issue 6. — P. 846.

77. *Morse D. P.*, *Aruscavage P. J.*, *Bass B. L.* RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA // Proceedings of the National Academy of Sciences of the United States of America. — 2002. — June. — Vol. 99, issue 12. — P. 7906.

78. The Landscape of A-to-I RNA Editome Is Shaped by Both Positive and Purifying Selection / Y. Yu, H. Zhou, Y. Kong, B. Pan, L. Chen, H. Wang, P. Hao, X. Li // PLoS genetics. — 2016. — July. — Vol. 12, issue 7.

79. A-to-I RNA editing promotes developmental stage-specific gene and lncRNA expression / B. Goldstein, L. Agranat-Tamir, D. Light, O. B. N. Zgayer, A. Fishman, A. T. Lamm // Genome research. — 2017. — Mar. — Vol. 27, issue 3. — P. 462–470.

80. *Pinto Y.*, *Cohen H. Y.*, *Levanon E. Y.* Mammalian conserved ADAR targets comprise only a small fragment of the human editosome // Genome Biology. — 2014. — Jan. — Vol. 15, issue 1. — P. 1–15.

81. Landscape of adenosine-to-inosine RNA recoding across human tissues // Nature Communications 2022 13:1. — 2022. — Mar. — Vol. 13, issue 1. — P. 1–17.

82. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing / S. Alon, S. C. Garrett, E. Y. Levanon, S. Olson, B. R. Graveley, J. J. Rosenthal, E. Eisenberg // eLife. — 2015. — Jan. — Vol. 2015, issue 4.

83. Trade-off between Transcriptome Plasticity and Genome Evolution in Cephalopods / N. Liscovitch-Brauer, S. Alon, H. T. Porath, B. Elstein, R. Unger, T. Ziv, A. Admon, E. Y. Levanon, J. J. Rosenthal, E. Eisenberg // Cell. — 2017. — Apr. — Vol. 169, issue 2. — 191–202.e11.

84. Genome-wide measurement of RNA secondary structure in yeast / M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, E. Segal // Nature. — 2010. — Sept. — Vol. 467, issue 7311. — P. 103–107.

85. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features / Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, S. M. Assmann // Nature. — 2014. — Vol. 505, issue 7485. — P. 696–700.

86. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. / A. M. Mustoe, S. Busan, G. M. Rice, C. E. Hajdin, B. K. Peterson, V. M. Ruda, N. Kubica, R. Nutiu, J. L. Baryza, K. M. Weeks // Cell. — 2018. — Vol. 173, issue 1. — 181–195.e18.

87. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing / J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, D. Haussler // Nature methods. — 2010. — Dec. — Vol. 7, issue 12. — P. 995.

88. RNA duplex map in living cells reveals higher order transcriptome structure / Z. Lu, Q. C. Zhang, B. Lee, R. A. Flynn, M. A. Smith, J. T. Robinson, C. Davidovich, A. R. Gooding, K. J. Goodrich, J. S. Mattick, J. P. Mesirov, T. R. Cech, H. Y. Chang // Cell. — 2016. — May. — Vol. 165, issue 5. — P. 1267.

89. *Ramani V.*, *Qiu R.*, *Shendure J.* High-throughput determination of rNA structure by proximity ligation // Nature Biotechnology. — 2015.

90. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation Molecular Cell Article In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation / J. Ghut, A. Aw, Y. Shen, A. Wilm, Z. Fu, N. Nagarajan, Y. W. Correspondence, M. Sun, X. N. Lim, K.-L. Boon, S. Tapsin, Y.-S. Chan, C.-P. Tan, A. Y. L. Sim, T. Zhang, T. T. Susanto, Y. Wan // Molecular Cell. — 2016. — Vol. 62. — P. 603–617.

91. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo / S. Rouskin, M. Zubradt, S. Washietl, M. Kellis, J. S. Weissman // Nature 2013 505:7485. — 2013. — Dec. — Vol. 505, issue 7485. — P. 701–705.

92. Codon catalog usage and the genome hypothesis. / R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Pavé // Nucleic Acids Research. — 1980. — Jan. — Vol. 8, no. 1. — r49.

93. *Sharp P. M.*, *Li W. H.* The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. // Nucleic Acids Research. — 1987. — Feb. — Vol. 15, no. 3. — P. 1281.

94. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels / M. Frenkel-Morgenstern, T. Danon, T. Christian, T. Igarashi, L. Cohen, Y. M. Hou, L. J. Jensen // Molecular Systems Biology. — 2012. — Vol. 8. — P. 572.

95. Quantifying Position-Dependent Codon Usage Bias / A. J. Hockenberry, M. I. Sirer, L. A. Amaral, M. C. Jewett // Molecular Biology and Evolution. — 2014. — July. — Vol. 31, no. 7. — P. 1880–1893.

96. Application of sorting and next generation sequencing to study 5'-UTR influence on translation efficiency in Escherichia coli / S. A. Evfratov, I. A. Osterman, E. S. Komarova, A. M. Pogorelskaya, M. P. Rubtsova, T. S. Zatsepin, T. A. Semashko, E. S. Kostryukova, A. A. Mironov, E. Burnaev, E. Krymova, M. S. Gelfand, V. M. Govorun, A. A. Bogdanov, P. V. Sergiev, O. A. Dontsova // Nucleic Acids Research. — 2017. — Apr. — Vol. 45, no. 6. — P. 3487–3502.

97. ViennaRNA Package 2.0 / R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I. L. Hofacker // Algorithms for Molecular Biology : AMB. — 2011. — Nov. — Vol. 6, no. 1. — P. 26.

98. *Akashi H.*, *Gojobori T.* Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis // Proceedings of the National Academy of Sciences of the United States of America. — 2002. — Mar. — Vol. 99, no. 6. — P. 3695.

99. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation / C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, D. Koller // Molecular Systems Biology. — 2014. — Dec. — Vol. 10, no. 12. — P. 770.

100. Efficient translation initiation dictates codon usage at gene start / K. Bentele, P. Saffert, R. Rauscher, Z. Ignatova, N. Blüthgen // Molecular systems biology. — 2013. — Vol. 9.

101. *Goodman D. B.*, *Church G. M.*, *Kosuri S.* Causes and effects of N-terminal codon bias in bacterial genes // Science (New York, N.Y.) — 2013. — Vol. 342, no. 6157. — P. 475–479.

102. *De Smit M. H.*, *Van Duin J.* Control of translation by mRNA secondary structure in Escherichia coli. A quantitative analysis of literature data // Journal of molecular biology. — 1994. — Nov. — Vol. 244, no. 2. — P. 144–150.

103. Depletion of Shine-Dalgarno sequences within bacterial coding regions is expression dependent / C. Yang, A. J. Hockenberry, M. C. Jewett, L. A. Amaralv // G3: Genes, Genomes, Genetics. — 2016. — Vol. 6, no. 11. — P. 3467–3474.

104. Within-Gene Shine–Dalgarno Sequences Are Not Selected for Function / A. J. Hockenberry, M. C. Jewett, L. A. Amaral, C. O. Wilke // Molecular Biology and Evolution. — 2018. — Oct. — Vol. 35, issue 10. — P. 2487–2498.

105. *Mankin A. S.* Nascent peptide in the "birth canal" of the ribosome // Trends in biochemical sciences. — 2006. — Vol. 31, issue 1. — P. 11–13.

106. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP / C. J. Woolstenhulme, N. R. Guydosh, R. Green, A. R. Buskirk // Cell reports. — 2015. — Apr. — Vol. 11, issue 1. — P. 13–21.

107. *Allert M.*, *Cox J. C.*, *Hellinga H. W.* Multifactorial determinants of protein expression in prokaryotic open reading frames // Journal of molecular biology. — 2010. — Oct. — Vol. 402, issue 5. — P. 905–918.

108. *Rodnina M. V.* The ribosome in action: Tuning of translational efficiency and protein folding // Protein Science : A Publication of the Protein Society. — 2016. — Aug. — Vol. 25, issue 8. — P. 1390.

109. Role of mRNA structure in the control of protein folding / G. Faure, A. Y. Ogurtsov, S. A. Shabalina, E. V. Koonin // Nucleic Acids Research. — 2016. — Dec. — Vol. 44, issue 22. — P. 10898.

110. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli / H. Chen, K. Shiroguchi, H. Ge, X. S. Xie // Molecular Systems Biology. — 2015. — Jan. — Vol. 11, issue 1. — P. 781–781.

111. GGRaSP: a R-package for selecting representative genomes using Gaussian mixture models / T. H. Clarke, L. M. Brinkac, G. Sutton, D. E. Fouts // Bioinformatics. — 2018. — Sept. — Vol. 34, issue 17. — P. 3032–3034.

112. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources / G. W. Li, D. Burkhardt, C. Gross, J. S. Weissman // Cell. — 2014. — Apr. — Vol. 157, no. 3. — P. 624–635.

113. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in Saccharomyces cerevisiae // Nucleic Acids Research. — 2014. — Vol. 42, issue 8. — P. 4813.

114. *Tietze L.*, *Lale R.* Importance of the 5' regulatory region to bacterial synthetic biology applications // Microbial Biotechnology. — 2021. — Nov. — Vol. 14, issue 6. — P. 2291–2315.

115. In vivo cleavage rules and target repertoire of RNase III in Escherichia coli / Y. Altuvia, A. Bar, N. Reiss, E. Karavani, L. Argaman, H. Margalit // Nucleic Acids Research. — 2018. — Aug. — Vol. 46, issue 19. — P. 10380–10394.

116. *Fei J.*, *Sharma C. M.* RNA localization in bacteria // Microbiology spectrum. — 2018. — Sept. — Vol. 6, issue 5.

117. *Buskilay A. A. A.*, *Kannaiahy S.*, *Amster-Choder O.* RNA localization in bacteria // RNA Biology. — 2014. — Aug. — Vol. 11, issue 8. — P. 1051.

118. *Soldatov R. A.*, *Vinogradova S. V.*, *Mironov A. A.* RNASurface: Fast and accurate detection of locally optimal potentially structured RNA segments // Bioinformatics. — 2014. — Vol. 30, issue 4. — P. 457–463.

119. Probing-directed identification of novel structured RNAs / S. V. Vinogradova, R. A. Sutormin, A. A. Mironov, R. A. Soldatov // RNA biology. — 2016. — Feb. — Vol. 13, issue 2. — P. 232–242.

Список рисунков

Список таблиц

Приложение А

Описание данных к главе 2

Таблица 2 — Проверочные последовательности и соответствующие им экспериментальные данные.

Обозначение	Последовательность	CER/RFP	Ст. откл.	РНК/ДНК	Ст. откл.
control	CACAGAAGAAACAAACAACTTATCAGACGC	4.19	0.56	nd	nd
+2 BAD	cccAGAAGAAACAGACAACTTAAGAGACGC	1.35	0.09	0.49	0.15
+3 BAD	CACtgcATTAACAAACAAACCATCAGACGC	1.8	0.03	0.5	0.23
+4 BAD	CACAGAgtcAACAAACAACTTATCAGACGC	3.24	0.1	0.36	0.34
+6 BAD	CACAGAATTAACctcCAACTTAAGAGACGC	4.69	0.13	0.51	0.13
+7 BAD	CACAGAAGAAGAAAAgtcCTTAAGAGACGC	2.15	0.05	1.63	0.54
+8 BAD	CACAGAATTAACAAACAAattAAGTCGCGC	7.22	1.27	0.92	1.4
+9 BAD	CACAGAAGAAACAAACAACTTacgTCGCGC	5.14	0.25	0.55	0.41
+10 BAD	CACAGAAGAAACAAACAACTTAAGcatCGC	8.12	0.61	0.69	0.89
+2,3 BAD	ccccccATTAGAAGACAACTTAAGAGACGC	0.66	0.04	0.47	0.17
+2,4 BAD	ctcAGAgtcAACAAAAGAACCATCAGACGC	0.12	0	0.35	0.17
+2,6 BAD	ctcAGAATTAACctcAGACTTAAGAGACGC	0.26	0.02	0.35	0.16
+2,7 BAD	cccAGAAGAAGAAAAgtcCTTAAGAGACGC	0.46	0.01	0.55	0.25
+2,8 BAD	ctcAGAATTAACAAACAAattAAGAGACGC	0.47	0.02	0.52	0.47
+2,9 BAD	ctcAGAAGAAACAAACAACTTacgAGACGC	0.36	0.02	0.26	0.08
+2,10 BAD	ctcTACAGAAACAAACAACTTATCcatCGC	0.52	0.06	0.63	0.19
+3,4 BAD	CACcccgtcAACAAACAACTTAAGAGACGC	0.41	0.07	0.43	0.15
+3,6 BAD	CACcccAGAAGActcAGAACCATCAGACGC	2.09	0.34	0.46	0.19
+3,7 BAD	CACcccAGAAACAAAgctCTTAAGAGACGC	1.33	0.04	0.55	0.23
+3,8 BAD	CACtgcAGAAACAGACAAacaAAGTCGCGC	3.23	0.66	0.55	0.21
+3,9 BAD	CACcccAGAAACAAACAACTTacgAGACGC	1.42	0.02	0.59	0.27
+3,10BAD	CACtgcATTAGAAAACAACTTATCcatCGC	5.69	1.13	0.58	0.43
+4,6 BAD	CACAGAgtcAACctcCAACTTAAGAGACGC	4.42	1.68	0.61	0.2
+4,7 BAD	CACAGAgtcAGAAAAgtcCTTAAGAGACGC	1.9	0.06	0.57	0.29
+4,8 BAD	CACAGAgtcAACAGACAAattATCAGACGC	5.6	6	0.44	0.14
+4,9 BAD	CACTACgtcAACAGAAGAACCaacAGACGC	3.85	0.12	0.77	0.22
+4,10 BAD	CACAGAgtcAACAAAAGAACCAAGcatCGC	3.86	0.28	0.99	0.5
+5,6 BAD	CACTACATTaggctcAGACTTAAGAGACGC	0.86	0.03	0.48	0.17
+6,7 BAD	CACAGAAGAAGActcgtcACCAAGAGACGC	0.02	0	0.57	0.96
+6,8 BAD	CACAGAATTAACctcCAAacaATCAGACGC	3.77	0.11	0.63	0.27
+6,9 BAD	CACTACAGAAACgggCAACTTaacAGACGC	1.86	0.03	0.75	0.32
+6,10 BAD	CACAGAAGAAACctcAGACTTAAGcatCGC	4.04	0.09	1.28	1.16
+7,8 BAD	CACTACATTAGAAAAgctacaAAGTCGCGC	3.57	0.39	0.63	0.42
+7,9 BAD	CACAGAAGAAGAAAAgtcCTTaacAGACGC	2.51	0.07	1.11	0.25
+7,10 BAD	CACTACATTAACAGAgctACCATCcatCGC	7.62	0.32	0.67	0.38
+8,9 BAD	CACAGAATTAACAAACAAattacgTCGCGC	7.47	0.63	0.2	0.15
+8,10 BAD	CACAGAATTAACAAACAAattAAGcatCGC	11.85	1.98	nd	nd
+9,10 BAD	CACAGAAGAAACAAACAACTTacgcatCGC	10.24	0.22	0.48	0.33
+2,3,4 BAD	ccccccgtcAACAAACAACTTAAGAGACGC	0.15	0.01	0.31	0.1
+2,3,6 BAD	ccccccATTAACctcCAACTTAAGAGACGC	0.33	0.04	0.47	0.2
+2,3,7 BAD	ccccccAGAAACAAAgtcCTTAAGAGACGC	0.26	0.02	0.43	0.12
+2,3,8 BAD	ccccccAGAAACAAACAAacaATCAGACGC	0.35	0	0.37	0.14
+2,3,9 BAD	ctctgcAGAAACAGACAACTTaacAGACGC	0.03	0	0.29	0.11
+2,3,10 BAD	ctctgcATTAGAAAACAACTTATCcatCGC	0.11	0.06	0.61	0.22
+2,4,6 BAD	cccAGAgtcAACctcCAACTTAAGAGACGC	0.63	0.01	0.4	0.13
+2,4,7 BAD	ctcAGAgtcAACAAAgctACCAAGAGACGC	0.09	0.01	0.46	0.18
+2,4,8 BAD	ctcAGAgtcAACAGACAAacaAAGAGACGC	0.32	0.01	0.31	0.24
+2,4,9 BAD	ctcAGAgtcAACAAACAAACCacgAGACGC	0.22	0.04	0.29	0.18

+2,4,10 BAD	cccAGAgtcAACAAACAAACCATCcatCGC	1.41	0.02	1.05	1.31
+2,5,6 BAD	cccAGAAGAaggctcCAACTTAAGAGACGC	0.07	0.01	0.4	0.41
+2,6,7 BAD	ctcAGAAGAAGActcgtcACCAAGAGACGC	0.09	0	0.55	0.66
+2,6,8 BAD	ctcAGAATTAACctcCAAattAAGAGACGC	0.31	0	0.37	0.2
+2,6,9 BAD	cccAGAATTAGActcCAAACCaacAGACGC	0.55	0.16	0.4	0.41
+2,6,10 BAD	ctcAGAAGAAGActcAGAACCAAGcatCGC	0.22	0.01	0.36	0.17
+2,7,8 BAD	ctcAGAAGAAACAAAgtcattATCAGACGC	0.23	0.02	0.58	0.96
+2,7,10 BAD	ctcAGAATTAACAAAgctACCATCcatCGC	0.31	0.04	0.58	0.96
+2,8,9 BAD	cccTACAGAAACAAACAAacaacgAGACGC	1.73	0.01	0.44	0.06
+2,8,10 BAD	cccAGAATTAACAGACAAacaATCcatCGC	2.65	0.11	0.6	0.17
+2,9,10 BAD	cccAGAAGAAGAAGACAACTTaaccatCGC	1.41	0.09	0.64	0.16
+3,4,6 BAD	CACcccgtcAGActcCAACTTAAGAGACGC	0.1	0.01	0.57	0.16
+3,4,7 BAD	CACcccgtcAACAAAgtcCTTAAGAGACGC	0.12	0.01	0.3	0.32
+3,4,8 BAD	CACcccgtcAACAAACAAacaATCAGACGC	0.55	0.02	0.47	0.25
+3,4,9 BAD	CACcccgtcAACAAACAAACCaacAGACGC	0.42	0.01	0.86	1.14
+3,4,10 BAD	CACtgcgtcAACAAACAACTTAAGcatCGC	1.98	0.02	0.5	0.38
+3,5,6 BAD	CACtgcAGAaggctcAGACTTAAGAGACGC	0.2	0.01	0.43	0.24
+3,6,7 BAD	CACcccAGAAGActcgtcACCAAGAGACGC	0.02	0	nd	nd
+3,6,8 BAD	CACtgcAGAAACctcAGAacaATCAGACGC	1.66	0.06	0.66	0.38
+3,6,9 BAD	CACtgcATTAACctcAGAACCaacAGACGC	1.74	0.08	1.51	1.55
+3,6,10 BAD	CACcccAGAAACctcAGACTTAAGcatCGC	1.17	0.02	0.58	0.21
+3,7,8 BAD	CACtgcAGAAACAAAgtcacaATCAGACGC	1.74	0.11	1.17	0.76
+3,7,9 BAD	CACtgcAGAAACAGAgctACCaacAGACGC	1.8	0.05	1.51	1.44
+3,7,10 BAD	CACtgcAGAAACAAAgctACCATCcatCGC	1.21	0.38	0.77	0.29
+3,8,9 BAD	CACtgcAGAAACAGACAAacaacgTCGCGC	2.03	0.02	0.79	0.39
+3,8,10 BAD	CACtgcATTAACAAACAAattAAGcatCGC	1.93	0.05	nd	nd
+3,9,10 BAD	CACcccATTAACAGACAACTTacgcatCGC	2.29	0.2	1.2	1.02
+4,5,6 BAD	CACAGAgtcaggctcCAACTTAAGAGACGC	4	0.21	0.83	0.33
+4,6,7 BAD	CACAGAgtcAGActcgtcACCAAGAGACGC	4.19	0.91	0.38	0.16
+4,6,8 BAD	CACAGAgtcAACctcCAAacaATCAGACGC	0.1	0.02	0.11	0.19
+4,6,9 BAD	CACAGAgtcAACctcCAAACCaacAGACGC	2.96	1.29	0.46	0.29
+4,6,10 BAD	CACAGAgtcAGActcAGAACCAAGcatCGC	1.77	0.77	0.3	0.08
+4,7,8 BAD	CACAGAgtcAACAGAgtcacaATCAGACGC	3.57	0.63	0.83	0.35
+4,7,9 BAD	CACAGAgtcAGAAAAgtcCTTaacAGACGC	1.6	0.03	0.48	0.2
+4,7,10 BAD	CACAGAgtcAACAAAgtcCTTAAGcatCGC	2.64	0.33	0.81	0.41
+4,8,9 BAD	CACAGAgtcAACAGAAGAacaacgTCGCGC	2.97	0.09	0.69	0.22
+4,8,10 BAD	CACAGAgtcAACAAAAGAacaAAGcatCGC	3.44	0.76	nd	nd
+4,9,10 BAD	CACAGAgtcAACAAAAGAACCacgcatCGC	0.53	0.05	0.73	0.21
+5,6,7 BAD	CACAGAAGAaggctcgtcACCAAGAGACGC	0.53	0.06	0.48	0.17
+5,6,8 BAD	CACAGAATTaggctcCAAacaAAGAGACGC	0.18	0.02	0.19	0.32
+5,6,9 BAD	CACTACATTgggctcAGAACCaacAGACGC	0.63	0.07	nd	nd
+5,6,10 BAD	CACTACAGAaggctcAGACTTAAGcatCGC	0.98	0.03	0.4	0.53
+6,7,8 BAD	CACAGAAGAAGActcgctacaAAGAGACGC	2.26	0.61	1.83	1.45
+6,7,9 BAD	CACAGAAGAAGActcgtcACCaacAGACGC	1.77	0.56	0.63	0.2
+6,7,10 BAD	CACAGAAGAAGActcgctACCATCcatCGC	3.29	1.19	1.86	2.11
+6,8,9 BAD	CACAGAAGAAGActcAGAacaacgTCGCGC	3.17	0.2	1.13	0.26
+6,8,10 BAD	CACAGAAGAAACctcAGAacaAAGcatCGC	2.61	1.8	nd	nd
+6,9,10 BAD	CACAGAATTAACctcCAACTTaaccatCGC	3.65	0.21	0.67	0.35
+7,8,9 BAD	CACAGAAGAAACAAAgctacaacgTCGCGC	11.02	7.89	0.94	0.19
+7,8,10 BAD	CACTACATTAACAGAgctacaATCcatCGC	5.26	0.63	nd	nd
+7,9,10 BAD	CACTACATTAGAAAAgtcACCaaccatCGC	9.11	5.37	0.82	0.31
+8,9,10 BAD	CACAGAATTAACAAACAAattacgcatCGC	8.3	0.24	nd	nd
+4 AUG	ATGctggtcggcctcctgacatgctgcacc	0.03	0.01	0.42	0.1
+5 AUG	cATGtgcccggcccggctatttgccatacc	0.01	0	0.33	0.14
+6 AUG	ccATGgcccggcccgctgacaaaccatacc	0.01	0	0.37	0.12
+7 AUG	cccATGgtcgggctcctgacaaactgcacc	0.03	0.03	1.09	0.39
+8 AUG	ccccATGccggcctcctgacaacgtgcacc	0.02	0	0.35	0.1
+9 AUG	cccctATGcggcccgctgattaactgcacc	0.02	0	0.33	0.11
+10 AUG	cccctgATGggcccggtcacaacgcatacc	0.01	0	0.48	0.19
+11 AUG	ccctgcgATGgcctcgctacaaaccatacc	0.04	0.01	0.31	0.09
+12 AUG	cccctgccATGcctcctgacaacgtgcacc	0.02	0	0.32	0.11

+13 AUG	ccccccgtcATGctcctgacatgccatacc	0.03	0.02	0.38	0.12
+14 AUG	cccctgcccgATGcgctgacatgccatacc	0.03	0.02	0.57	0.17
+15 AUG	cccctgcccagATGcctgacatgccatacc	0.07	0.01	0.47	0.2
+16 AUG	ccccccgtcggcATGctgattaactgcacc	0.03	0.01	0.57	0.19
+17 AUG	cccctggtcaggcATGtcacaaactgcacc	0.03	0	0.82	0.31
+18 AUG	cccctggtcggcccATGcacatgccatacc	0.02	0	0.56	0.21
+19 AUG	ccctgcgtcggcccgATGacaacgtgcacc	0.03	0	0.38	0.12
+20 AUG	ccctgcgtcggcctccATGcaacgcatacc	0.01	0	0.36	0.1
+21 AUG	cccctggtcggcctcgcATGaaaccatacc	0.01	0	0.44	0.15
+22 AUG	ccccccgtcggcctcgctATGaaccatacc	0.04	0.01	0.38	0.08
+23 AUG	cccctggtcaggctcctgaATGaccatacc	0.02	0.01	0.87	0.43
+24 AUG	cccctggtcggcctcctgatATGgcatacc	0.01	0	0.48	0.22
+25 AUG	ccccccccggcccggctacaATGtgcacc	0.01	0	0.35	0.09
+26 AUG	ccctgcgtcggcctcgctattaATGatacc	0.01	0	0.27	0.08
+27 AUG	cccctggtcaggctcgtcacaaaATGcacc	0.03	0	0.84	0.36
+28 AUG	ccccccgtcggcccggctattaacATGacc	0.02	0	0.61	0.22
+29 AUG	ccccccgtcggcctcctgacaacgcATGcc	0.04	0.01	0.38	0.15
+30 AUG	cccctgcccggcccggctacaaaccaATGc	0.01	0	nd	nd
+4 SD	aggaggAGAAACAAACAAACCAAGAGACGT	0.04	0	2.07	1.14
+5 SD	CaggaggTTAACAAACAACTTAAGAGACGC	0.15	0.05	1.12	0.78
+6 SD	CAaggaggAAACAAACAAACCAAGAGACGT	0.01	0	nd	nd
+7 SD	CACaggaggAACAAACAAACCATCAGACGT	0.15	0.04	1.63	0.48
+8SD	GTATaggaggACAAACAAACCAAGAGACGC	0.06	0.01	1.41	0.63
+9 SD	CACAGaggaggCAAACAACTTAAGAGACGC	0.12	0.03	1.36	0.56
+10 SD	CACTACaggaggAAACAAACCATCAGACGT	1.54	0.03	1.76	0.37
+11 SD	CACAGAAaggaggGACAAACCAAGAGACGC	0.21	0.04	nd	nd
+12 SD	CACAGAAGaggaggAAGACTTAAGAGACGT	0.21	0.04	1.79	0.49
+13 SD	CACTACATTaggaggAGAACCAAGAGACGC	0.21	0.01	1.14	0.3
+14 SD	GTATACAGAAaggaggAACTTAAGAGACGT	0.32	0.04	2.36	0.76
+15 SD	GTAAGAATTAAaggaggAACCAAGAGACGC	0.33	0.03	1.73	0.33
+16 SD	CACTACAGAAGAaggaggCTTAAGAGACGC	0.16	0.01	1.25	0.4
+17 SD	CACAGAAGAAACAaggaggTTATCAGACGT	0.17	0	0.91	0.2
+18 SD	CACTACATTAACAAaggaggTAAGAGACGT	0.33	0.02	2.22	0.68
+19 SD	GTAAGAAGAAACAAAaggaggAAGTCGCGT	0.57	0.47	1.43	0.37
+20 SD	GTATACAGAAACAAACaggaggAGAGACGC	0.97	0.06	1.73	0.45
+21 SD	CACAGAAGAAACAAACAaggaggCAGACGC	0.54	0.09	0.82	0.33
+22 SD	GTATACAGAAACAAACAAaggaggAGACGC	1.69	0.02	2.06	0.59
+23 SD	CACAGAAGAAACAAACAAAaggaggCGCGC	1.2	0.03	1.1	0.41
+24 SD	CACTACAGAAACAAACAAACaggaggACGT	2.08	0.13	1.95	0.95
+25 SD	GTATACAGAAACAAACAACTTaggaggCGT	4	0.79	1.89	0.89
+26 SD	GTAAGAAGAAACAAACAAACCAaggaggGC	3.17	0.25	1.42	0.33
+27 SD	CACTACAGAAACAAACAACTTATaggaggT	2.34	0.15	1.71	0.72
+28 SD	GTAAGAAGAAACAAACAAACCAAGaggagg	2.76	0.19	1.88	0.8
Expensive 1, 320 au	CACTACATTAGAAAAAGACTTATCAGACGC	3.84	0.24	0.73	0.47
Cheap1, 233 au	GTAAGAAGAAACAGACAACTTAAGTCGCGT	5.33	0.29	0.77	0.18
Cheap2, 224 au	GTAAGAAGAAACAGACAAACCAAGTCGCGC	2.46	0.01	0.83	0.33
Expensive 2, 305 au	GTATACATTAGAAAAAGACTTATCAGACGC	5.55	0.18	nd	nd
3 CGA + AGA	cgaAGAcgacgaAAACAACTTAAGAGACGC	1.79	0.05	0.95	0.26
4 AGA	agaAGAagaagaAAACAACTTAAGAGACGC	0.02	0	1.21	0.41
3 CGC + AGA	cgcAGAcgccgcAAACAACTTAAGAGACGC	0.38	0.02	0.44	0.15
3 CGU + AGA	cgtAGAcgtcgtAAACAACTTAAGAGACGC	1.31	0.02	0.7	0.24

Приложение Б

Репрезентативные штаммы *Escherichia coli*

Таблица 3 — Штаммы *Escherichia coli*, используемые для изучения полиморфизмов в спаренных позициях, и идентификаторы соответствующих геномных сборок в базе данных RefSeq.

Идентификатор RefSeq	**Название штамма**
GCF_001286185.1	Escherichia coli strain 102366_aEPEC
GCF_000019645.1	Escherichia coli SMS-3-5
GCF_000692855.1	Escherichia coli UCI
GCF_001900455.1	Escherichia coli strain H7
GCF_000778785.1	Escherichia coli strain upec-213
GCF_001447405.1	Escherichia coli str. Deng
GCF_000408645.1	Escherichia coli KTE108
GCF_001749705.1	Escherichia coli strain AF7607
GCF_001644725.1	Escherichia coli strain 2011C-3911
GCF_001264175.1	Escherichia coli strain IH53473
GCF_001616415.1	Escherichia coli strain swine71
GCF_001645225.1	Escherichia coli strain 2011C-3198
GCF_001265975.1	Escherichia coli strain 402310
GCF_001262905.1	Escherichia coli strain CFSAN026806
GCF_001703435.1	Escherichia coli strain LV13072
GCF_001893775.1	Escherichia coli strain 711
GCF_001893125.1	Escherichia coli strain 486
GCF_900092915.1	Escherichia coli strain F1-8-ERB4
GCF_000350685.1	Escherichia coli KTE10
GCF_001900535.1	Escherichia coli strain C5
GCF_001902735.1	Escherichia coli strain 256
GCF_000703725.1	Escherichia coli 2-156-04_S3_C2
GCF_001892765.1	Escherichia coli strain 508
GCF_000351425.1	Escherichia coli KTE47
GCF_001266375.1	Escherichia coli strain 401675
GCF_000408565.1	Escherichia coli KTE100
GCF_001900675.1	Escherichia coli strain D5
GCF_000704045.1	Escherichia coli 2-316-03_S3_C1
GCF_000601195.1	Escherichia coli 1-176-05_S3_C2
GCF_001614635.1	Escherichia coli strain cattle13
GCF_003721655.1	Escherichia coli strain C85
GCF_000352425.1	Escherichia coli KTE78
GCF_001936315.1	Escherichia coli SLK172
GCF_003721775.1	Escherichia coli strain C43
GCF_001309535.1	Escherichia coli strain DM18-3
GCF_000352665.1	Escherichia coli KTE144

GCF_001467005.1	Escherichia coli NCCP
GCF_003721715.1	Escherichia coli strain C56
GCF_000408605.1	Escherichia coli KTE103
GCF_000227625.1	Escherichia coli O7:K1 str. CE10
GCF_000812765.1	Escherichia coli strain EC2_2
GCF_001419945.1	Escherichia coli strain 199
GCF_000326365.1	Escherichia coli KTE148
GCF_001679985.1	Escherichia coli strain 210205630
GCF_001894245.1	Escherichia coli strain 495
GCF_001901405.1	Escherichia coli strain S50
GCF_002190715.1	Escherichia coli strain ECOR31
GCF_001615985.1	Escherichia coli strain swine49
GCF_001942165.1	Escherichia coli strain 62D3
GCF_000355255.2	Escherichia coli Jurua
GCF_001607765.1	Escherichia coli strain STEC
GCF_000351625.1	Escherichia coli KTE66
GCF_000355155.2	Escherichia coli 2719100
GCF_000627435.1	Escherichia coli 1-250-04_S3_C1
GCF_001893615.1	Escherichia coli strain 709
GCF_003721795.1	Escherichia coli strain C22
GCF_000805835.1	Escherichia coli strain 4608-58
GCF_002769915.1	Escherichia coli O26:H11 strain 2243
GCF_000482065.1	Escherichia coli SCD2
GCF_000968515.1	Escherichia coli VR50
GCF_000352625.1	Escherichia coli KTE140
GCF_000326705.1	Escherichia coli KTE145
GCF_003722035.1	Escherichia coli strain C101
GCF_001886575.1	Escherichia coli strain MRSN346355
GCF_001900595.1	Escherichia coli strain D1
GCF_000797775.1	Escherichia coli strain CVM
GCF_001900395.1	Escherichia coli strain D8
GCF_001900695.1	Escherichia coli strain D6
GCF_000194645.1	Escherichia coli 3.2303
GCF_001860505.1	Escherichia coli strain Y5
GCF_000326185.1	Escherichia coli KTE109
GCF_000351925.1	Escherichia coli KTE142
GCF_001900335.1	Escherichia coli strain C10
GCF_001938625.2	Escherichia coli strain 2016C-3936C1
GCF_001265755.1	Escherichia coli strain 300075
GCF_000804365.1	Escherichia coli strain EC14
GCF_001900985.1	Escherichia coli strain S43
GCF_000492655.1	Escherichia coli BWH
GCF_000700365.1	Escherichia coli 3-267-03_S1_C2
GCF_001886935.1	Escherichia coli strain FORC_041
GCF_001262785.1	Escherichia coli strain CFSAN026784
GCF_000461475.1	Escherichia coli UMEA
GCF_001186315.1	Escherichia coli M114
GCF_000827105.1	Escherichia coli O157:H16 strain Santai

GCF_000353185.1	Escherichia coli KTE104
GCF_000460075.1	Escherichia coli UMEA
GCF_000576655.1	Escherichia coli MP1
GCF_000354795.2	Escherichia coli MP021561.2
GCF_000408625.1	Escherichia coli KTE107
GCF_000459915.1	Escherichia coli KOEGE
GCF_000447125.2	Escherichia coli C9_92
GCF_000352505.1	Escherichia coli KTE101
GCF_001012335.1	Escherichia coli strain CFSAN026802
GCF_000398885.1	Escherichia coli KTE33
GCF_001575545.1	Escherichia coli strain G213
GCF_001901315.1	Escherichia coli strain S1
GCF_001677475.1	Escherichia coli strain 06-00048
GCF_001893845.1	Escherichia coli strain 720
GCF_000010245.2	Escherichia coli str. K-12
GCF_000948905.1	Escherichia coli strain OLC-975
GCF_000731355.1	Escherichia coli D6-113.11
GCF_000460635.1	Escherichia coli UMEA
GCF_001652805.1	Escherichia coli strain E-78
GCF_001749365.1	Escherichia coli strain VL119
GCF_002319705.1	Escherichia coli strain C104
GCF_000687125.1	Escherichia coli 2-011-08_S1_C1
GCF_001266225.1	Escherichia coli strain 302137
GCF_000458705.1	Escherichia coli HVH
GCF_001891655.1	Escherichia coli strain 481
GCF_000736735.1	Escherichia coli strain 48
GCF_000211395.1	Escherichia coli AA86
GCF_003721675.1	Escherichia coli strain C59
GCF_001900885.1	Escherichia coli strain H14
GCF_001682305.2	Escherichia coli strain EC590
GCF_001521075.1	Escherichia coli strain GN02446
GCF_001902655.1	Escherichia coli strain N11-1317
GCF_001890365.1	Escherichia coli strain MRSN346647
GCF_000461015.1	Escherichia coli UMEA
GCF_000703345.1	Escherichia coli O128:H2 str. 2011C-3317
GCF_000299455.1	Escherichia coli O104:H4 str. 2011C-3493
GCF_000507685.1	Escherichia coli UMEA
GCF_000340255.1	Escherichia coli S17
GCF_000194665.1	Escherichia coli 3003
GCF_000249715.1	Escherichia coli DEC7B
GCF_000460975.1	Escherichia coli UMEA
GCF_002193095.1	Escherichia coli strain H105
GCF_001892865.1	Escherichia coli strain 655
GCF_001419925.1	Escherichia coli strain 166
GCF_000163195.1	Escherichia coli B354
GCF_000356165.2	Escherichia coli P0298942.1
GCF_001616515.1	Escherichia coli strain sheep19
GCF_001893015.1	Escherichia coli strain 483

GCF_001912385.1	Escherichia coli strain 102928
GCF_001748905.1	Escherichia coli strain AF7709
GCF_000784925.1	Escherichia coli strain ECONIH1
GCF_000352925.1	Escherichia coli KTE196
GCF_001900495.1	Escherichia coli strain C2
GCF_000468515.1	Escherichia coli LY180
GCF_000713745.1	Escherichia coli 2-177-06_S3_C3
GCF_000819015.1	Escherichia coli strain CVM
GCF_002810235.1	Escherichia coli strain 6383
GCF_001284165.1	Escherichia coli strain 402078
GCF_001283865.1	Escherichia coli strain 401363
GCF_000407805.1	Escherichia coli KTE14
GCF_001748765.1	Escherichia coli strain AF85
GCF_000460375.1	Escherichia coli UMEA
GCF_001562635.1	Escherichia coli strain SYAU2
GCF_002319355.1	Escherichia coli strain 65-57
GCF_001910985.1	Escherichia coli strain 300769
GCF_001266175.1	Escherichia coli strain 200077
GCF_900015995.1	Escherichia coli strain M858
GCF_000163175.1	Escherichia coli B185
GCF_000833145.1	Escherichia coli strain BL21
GCF_000179115.1	Escherichia coli MS
GCF_000026545.1	Escherichia coli 0127:H6
GCF_001266125.1	Escherichia coli strain 103338
GCF_000798035.1	Escherichia coli strain CVM
GCF_000352465.1	Escherichia coli KTE84
GCF_001608245.1	Escherichia coli strain STEC
GCF_001614645.1	Escherichia coli strain sheep8
GCF_001266405.1	Escherichia coli strain 402981
GCF_001901445.1	Escherichia coli strain S56
GCF_001576035.1	Escherichia coli strain G27
GCF_003591595.1	Escherichia coli ATCC
GCF_001616045.1	Escherichia coli strain swine50
GCF_000242055.1	Escherichia coli TA124
GCF_000352705.1	Escherichia coli KTE147
GCF_000460155.1	Escherichia coli UMEA
GCF_001901125.1	Escherichia coli strain M10
GCF_000351945.1	Escherichia coli KTE143
GCF_000212715.2	Escherichia coli UMNK88
GCF_001265465.1	Escherichia coli 401140
GCF_001616755.1	Escherichia coli strain sheep31
GCF_000785795.1	Escherichia coli strain UMNturkey5
GCF_001900735.1	Escherichia coli strain D10
GCF_000326165.1	Escherichia coli KTE106
GCF_003721725.1	Escherichia coli strain C25
GCF_001901005.1	Escherichia coli strain H15
GCF_000303635.2	Escherichia coli N1
GCF_000284495.1	Escherichia coli LF82

GCF_000459015.1	Escherichia coli HVH
GCF_001191055.1	Escherichia coli strain CFSAN026776
GCF_000194415.1	Escherichia coli 1.2264
GCF_001893035.1	Escherichia coli strain 487
GCF_001675145.1	Escherichia coli strain ECONIH2
GCF_000812725.1	Escherichia coli strain EC2_1
GCF_001721225.1	Escherichia coli strain CFSAN004176
GCF_001566675.1	Escherichia coli strain ZH193
GCF_001950775.1	Escherichia coli strain 16309
GCF_001901365.1	Escherichia coli strain S30
GCF_001893165.1	Escherichia coli strain 489
GCF_000225205.1	Escherichia coli STEC_MHI813
GCF_001700405.1	Escherichia coli strain ECO37
GCF_000219515.2	Escherichia coli PCN033
GCF_000781295.1	Escherichia coli strain upec-112
GCF_000461615.1	Escherichia coli UMEA
GCF_001645245.1	Escherichia coli strain 2014C-3250
GCF_001683595.1	Escherichia coli strain NGF2
GCF_002769495.1	Escherichia coli O26:H11 strain 1657
GCF_000010385.1	Escherichia coli SE11
GCF_000241995.1	Escherichia coli E101
GCF_001420935.1	Escherichia coli strain 2012C-4227
GCF_001900375.1	Escherichia coli strain D7
GCF_000801165.1	Escherichia coli strain RM9387
GCF_000408545.1	Escherichia coli KTE98
GCF_000335195.2	Escherichia coli 3.4880
GCF_001615835.1	Escherichia coli strain swine41
GCF_001660565.1	Escherichia coli strain S51
GCF_001894235.1	Escherichia coli strain 493
GCF_000091005.1	Escherichia coli O26:H11 str. 11368
GCF_001891535.1	Escherichia coli strain 458
GCF_000356725.1	Escherichia coli BCE008_MS-13
GCF_001616635.1	Escherichia coli strain sheep25
GCF_001900555.1	Escherichia coli strain C8
GCF_001641925.1	Escherichia coli strain TN1
GCF_001286025.1	Escherichia coli strain 402924
GCF_001007915.1	Escherichia coli strain CFSAN029787
GCF_000356985.2	Escherichia coli P0299438.11
GCF_000714975.1	Escherichia coli strain UCD_JA65_pb
GCF_001012505.1	Escherichia coli strain CFSAN026836
GCF_000194575.1	Escherichia coli STEC_7v
GCF_001484475.1	Escherichia coli strain AZ72
GCF_001309485.1	Escherichia coli strain AW1.3
GCF_001030515.1	Escherichia coli strain BIDMC102
GCF_000010745.1	Escherichia coli O103:H2 str. 12009
GCF_001900315.1	Escherichia coli strain C4
GCF_001420075.1	Escherichia coli strain 224
GCF_000194685.1	Escherichia coli TW07793

GCF_000461855.1	Escherichia coli UMEA
GCF_000352725.1	Escherichia coli KTE154
GCF_001607815.1	Escherichia coli strain STEC
GCF_000671295.1	Escherichia coli O145:H28 str. RM12581
GCF_000458035.1	Escherichia coli HVH
GCF_000714295.1	Escherichia coli 5-366-08_S3_C2
GCF_001266275.1	Escherichia coli strain 303139
GCF_000797975.1	Escherichia coli strain CVM
GCF_001893625.1	Escherichia coli strain 675
GCF_001284705.1	Escherichia coli strain 500989
GCF_000408085.1	Escherichia coli KTE185
GCF_000220005.1	Escherichia coli UMNF18
GCF_000297235.3	Escherichia coli EC302/04
GCF_000355575.2	Escherichia coli 2785200
GCF_001677515.1	Escherichia coli strain 08-00022
GCF_001900945.1	Escherichia coli strain S40
GCF_002917695.1	Escherichia coli strain 241
GCF_000457655.1	Escherichia coli HVH
GCF_000780315.1	Escherichia coli strain upec-154
GCF_000356605.2	Escherichia coli p0305293.1
GCF_000461515.1	Escherichia coli UMEA
GCF_000352685.1	Escherichia coli KTE146
GCF_000408585.1	Escherichia coli KTE102
GCF_000801185.2	Escherichia coli strain 94-3024
GCF_001750845.1	Escherichia coli strain FORC_031
GCF_000703365.1	Escherichia coli O121:H19 str. 2011C-3609
GCF_000690815.1	Escherichia coli DSM 30083 = JCM 1649 = ATCC 11775 strain DSM
GCF_000167915.2	Escherichia coli 53638
GCF_001900475.1	Escherichia coli strain H10
GCF_001651695.1	Escherichia coli strain UPEC_014
GCF_000027125.1	Escherichia coli 042
GCF_001912425.1	Escherichia coli strain 402286
GCF_000460435.1	Escherichia coli UMEA
GCF_000350845.1	Escherichia coli KTE28
GCF_000304255.1	Escherichia coli AD30
GCF_001900815.1	Escherichia coli strain H3
GCF_001901045.1	Escherichia coli strain M3
GCF_000352645.1	Escherichia coli KTE141
GCF_001442495.1	Escherichia coli strain YD786
GCF_001614495.1	Escherichia coli strain sheep6
GCF_001901465.1	Escherichia coli strain S10
GCF_001419895.1	Escherichia coli strain 165
GCF_000351325.1	Escherichia coli KTE233
GCF_001747365.1	Escherichia coli strain ICBEC7P
GCF_001463455.1	Escherichia coli strain 50870281
GCF_001893085.1	Escherichia coli strain 638
GCF_001284265.1	Escherichia coli strain 401250_aEPEC
GCF_000017745.1	Escherichia coli E24377A

GCF_001900575.1	Escherichia coli strain C9
GCF_000026305.1	Escherichia coli ED1a
GCF_001891705.1	Escherichia coli strain 480
GCF_001191215.1	Escherichia coli strain CFSAN026817
GCF_000458935.1	Escherichia coli HVH
GCF_000215165.1	Escherichia coli H30
GCF_000647475.1	Escherichia coli STEC O174:H46 str. I-151
GCF_000798515.1	Escherichia coli strain CVM
GCF_001519525.1	Escherichia coli strain GN02235
GCF_000326145.1	Escherichia coli KTE105
GCF_000461675.1	Escherichia coli UMEA
GCF_001865895.1	Escherichia coli strain Tob1
GCF_000492235.1	Escherichia coli BIDMC
GCF_001607515.1	Escherichia coli strain STEC
GCF_000233895.1	Escherichia coli str. 'clone
GCF_001443095.1	Escherichia coli strain MNCRE22
GCF_001606525.1	Escherichia coli strain STEC
GCF_000458915.1	Escherichia coli HVH
GCF_001650605.1	Escherichia coli strain EC1636
GCF_000987875.1	Escherichia coli strain SEC470
GCF_001893105.1	Escherichia coli strain 484
GCF_001911335.1	Escherichia coli strain 703420
GCF_001607335.1	Escherichia coli strain STEC
GCF_001900795.1	Escherichia coli strain H2
GCF_003721835.1	Escherichia coli strain C21
GCF_003722015.1	Escherichia coli strain C102
GCF_001521285.1	Escherichia coli strain GN02461
GCF_000459855.1	Escherichia coli KOEGE
GCF_000482045.1	Escherichia coli SCD1