

На правах рукописи

**Червонцева Зоя Сергеевна**

**Влияние вторичной структуры мРНК  
на экспрессию генов**

Специальность 1.5.8 —  
«математическая биология, биоинформатика»

Автореферат  
диссертации на соискание учёной степени  
кандидата биологических наук

Москва — 2023

Работа выполнена в Учебно-научном центре «Биоинформатика» Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук.

Научный руководитель: **Гельфанд Михаил Сергеевич**  
кандидат физико-математических наук, доктор биологических наук, профессор  
(Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации имени А.А. Харкевича Российской академии наук)

Официальные оппоненты: **Самсонова Мария Георгиевна**,  
доктор биологических наук, зав. лабораторией  
(Федеральное государственное автономное образовательное учреждение высшего образования «Санкт-Петербургский политехнический университет Петра Великого»),

**Дмитриев Сергей Евгеньевич**  
кандидат биологических наук, доцент  
(Факультет биоинженерии и биоинформатики Московского государственного университета имени М.В. Ломоносова)

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Московский физико-технический институт (национальный исследовательский университет)»

Защита состоится 25 сентября 2023 года в 15 –00 на заседании диссертационного совета 24.1.101.01 при Институте проблем передачи информации имени А.А. Харкевича Российской академии наук (ИППИ РАН) по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке ИППИ РАН, а также на сайте ИППИ РАН по адресу: <http://iitp.ru/ru/dissertation/1624.htm>

Автореферат разослан «\_\_\_\_\_» \_\_\_\_\_ 2023 года

Ученый секретарь  
диссертационного совета  
24.1.101.01,  
доктор биологических наук

Казенников Олег Васильевич

## Общая характеристика работы

**Актуальность темы.** 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.

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.

**Целями** данной работы были изучение роли вторичной структуры в трансляции, деградации и эволюции мРНК у *Escherichia coli*, а также описание закономерностей, 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 эквимоллярные субъединицы одного белкового комплекса и закодированных в одном опероне.
3. Изучить связь стабильности вторичной структуры матричных РНК *Escherichia coli* со скоростью их дегградации 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.

**Основные 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 эквимоллярные 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 структурированности;
  - (c) указанные эффекты сильнее проявляются, если редактирование консервативно;
  - (d) вторичная структура способствует коррелированному редактированию сближенных аденинов.

**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 эквимольярных 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.

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

## Содержание работы

**Глава 1** содержит обзор литературы по теме диссертации, включающий четыре раздела.

В разделе 1.1 вводится понятие структурированности РНК.

В разделе 1.2 обсуждаются свойства вторичной структуры матричных РНК у про- и эукариот. Приведен обзор работ, описывающих характер распределения элементов стабильной вторичной структуры вдоль транскрипта и специфичность их сворачивания.

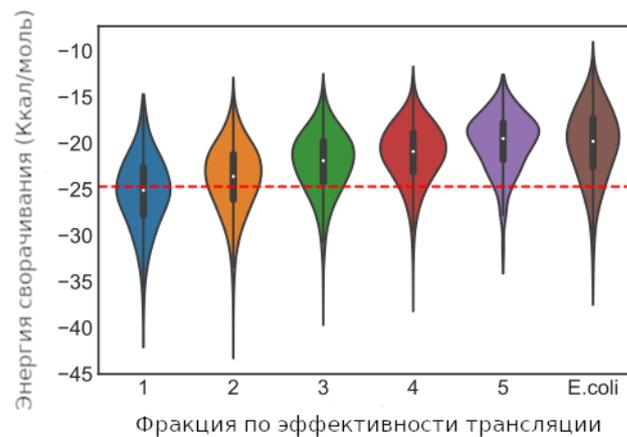
Раздел 1.3 посвящен роли вторичной структуры мРНК в различных биологических процессах: деградации, локализации, трансляции и редактировании мРНК. Упомянуты противоречивые данные о факторах, определяющих эффективность трансляции у бактерий. Приведено обоснование выбора мягкотелых головоногих моллюсков (колеоидов) в качестве объекта для изучения редактирования мРНК.

Раздел 1.4 содержит описание экспериментальных методов определения структуры РНК.

В **главе 2** описаны результаты анализа экспериментальных данных по экспрессии флуоресцентного белка со случайными вставками в начале гена в клетках *Escherichia coli*.

Эксперименты были произведены лабораторией П.В. Сергиева в НИИ ФХБ МГУ, где была разработана методика исследования свойств кодирующей последовательности, влияющих на эффективность трансляции. Эта методика основана на порождении больших библиотек плазмид, которые содержат ген флуоресцентного белка со случайными вставками в начале гена. Уровень флуоресценции полагается соответствующим количеству произведенного белка, то есть эффективности его трансляции. Точнее, эффективность трансляции определяется как соотношение уровня флуоресценции бел-

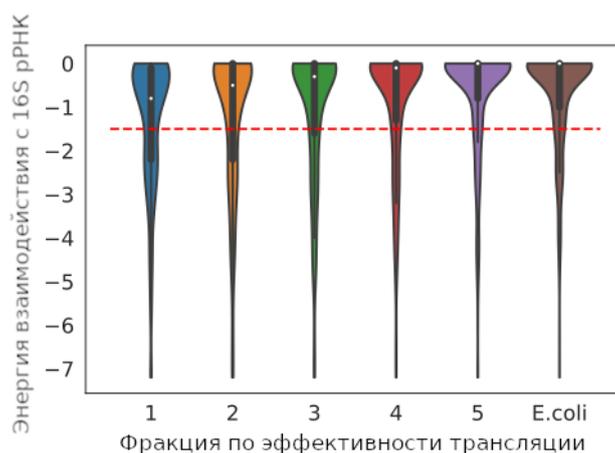
ка CER, закодированного модифицированным геном, к уровню флюоресценции закодированного в той же плазмиде белка RFP, чей ген не подвергается модификациям и имеет тот же уровень транскрипции, что и ген белка CER. Клетки, трансфицированные такими плазмидами со случайными вставками, сортируют по соотношению 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) получают большие наборы вставок, обеспечивающих этот уровень эффективности.



**Рис. 1** — 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.

Мы подтвердили предыдущие наблюдения, что 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 1). Поскольку эффективные ФЭТ имеют в среднем более низкий GC-состав, для оценки статистической значимости был использован контроль с рандомизацией последовательностей. 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*.

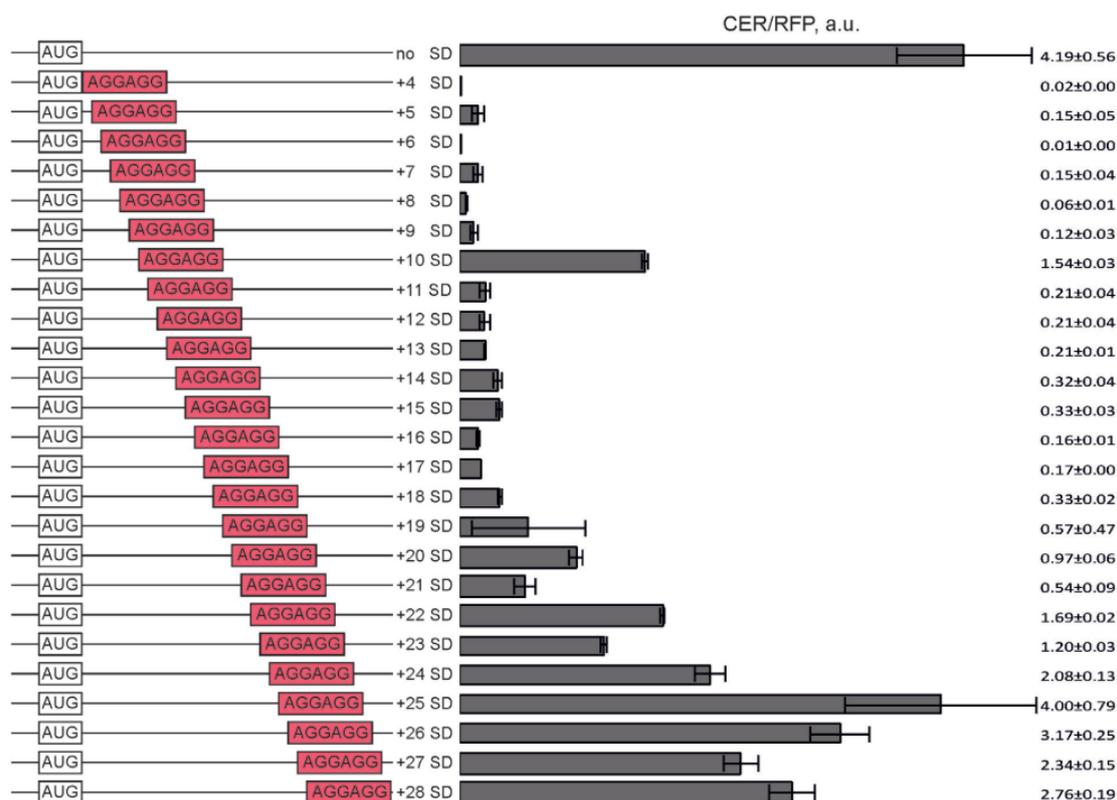
Анализ сродства изучаемых случайных последовательностей к участку анти-Шайна–Дальгарно 16S рРНК показал избегание такого взаимодействия во вставках из эффективно транслируемых фракций (Рисунок 2,  $p\text{-value} < 10^{-4}$ ). Это наблюдение также согласуется с предыдущими исследованиями о негативном влиянии внутригенных ШД-подобных участков на эффективность трансляции.



**Рис. 2** — 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.

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. Наблюдаемые значения эффективности трансляции для этих последовательностей согласуются с наблюдением, полученным на массовых данных: трансляция последовательностей, содержащих ШД-мотив практически в любой из позиций, менее эффективна, чем трансляция контрольной последовательности, не содержащей ШД-подобных участков (Рисунок 3). Кроме того, чем ближе ШД-подобный участок находится к старт-кодону, тем больший отрицательный эффект он оказывает.

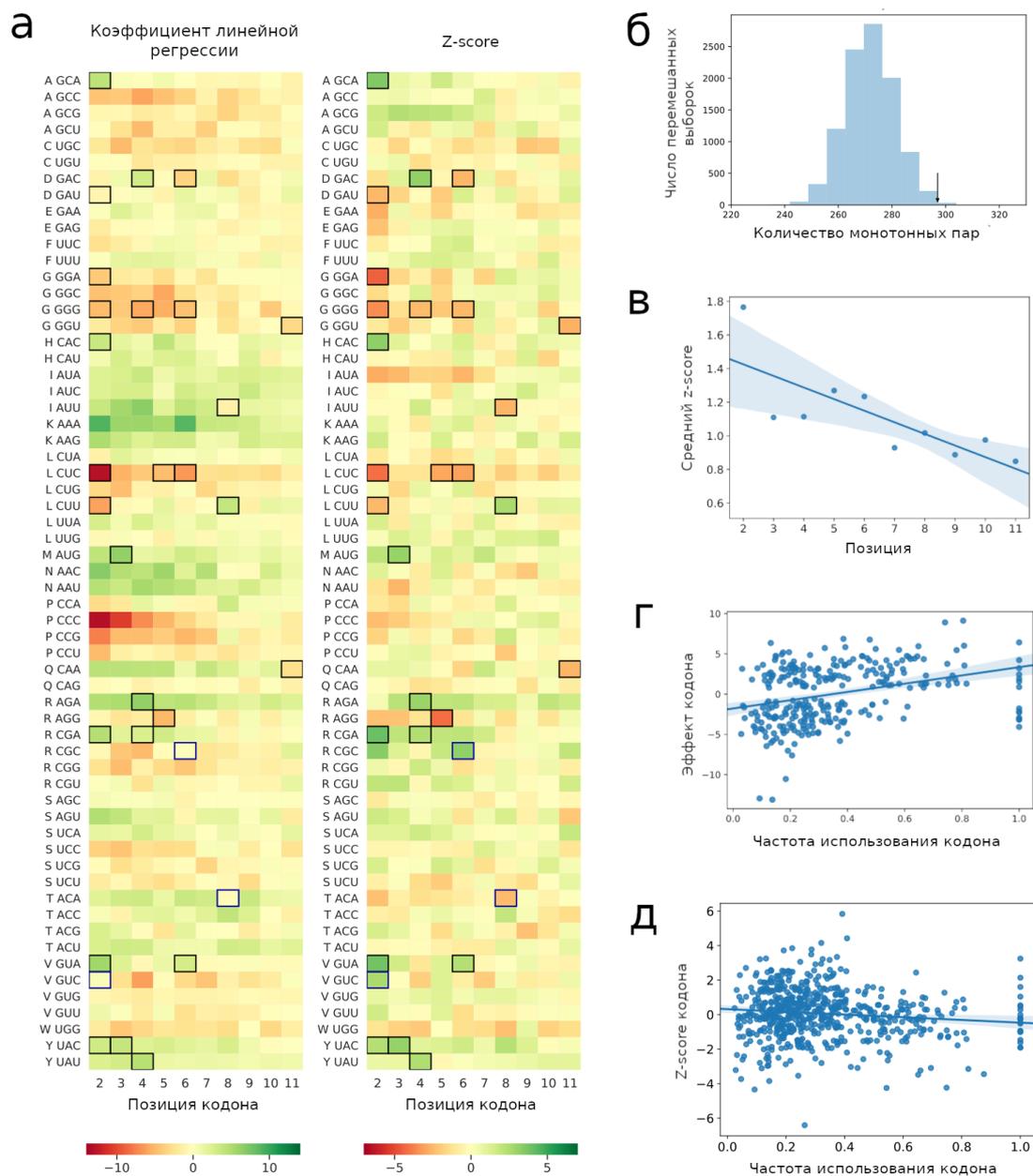
Для оценки влияния конкретных кодонов на эффективность трансляции мы определили позиционные частоты кодонов в каждой фракции и сравнили их с таковыми для рандомизированных последовательностей (Рисунок



**Рис. 3** — Эффективность трансляции для проверочных последовательностей, содержащих ШД-подобные участки. Схематичное представление положения ШД-подобного участка в последовательности (слева) и значения эффективности трансляции (справа).

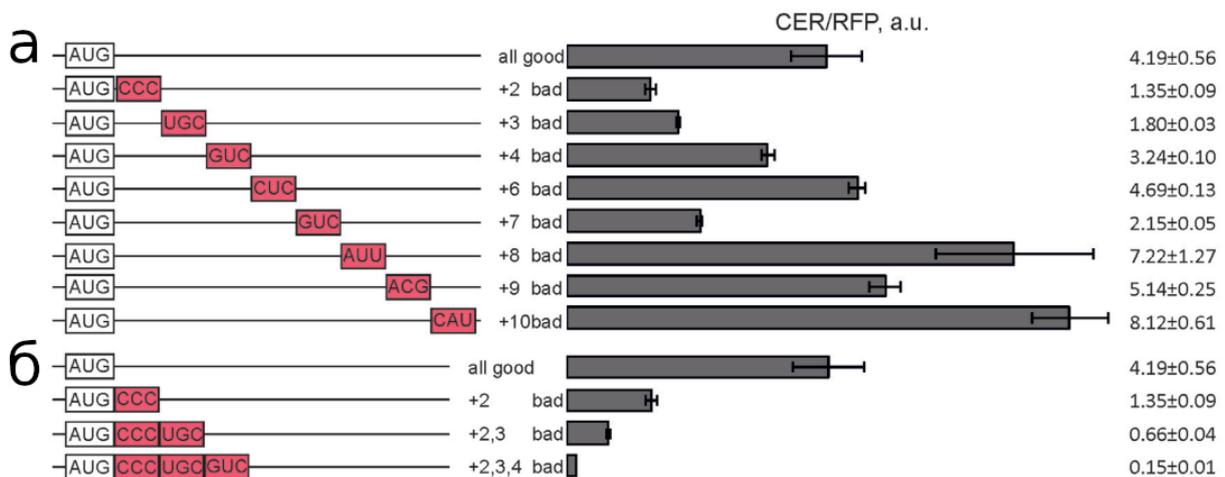
4). Сравнение левой и правой панели рисунка показывает, что значительную часть эффектов можно было бы с хорошей вероятностью получить и на случайной выборке, сохраняющей только позиционные частоты нуклеотидов в разных фракциях (z-score таких наблюдений близок к нулю, и соответствующие ячейки тепловой карты в правой панели нейтрально-желтые). Однако количество ячеек (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 4б). 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 4в), что согласуется с предыдущими наблюдениями и косвенно валидирует наш способ статистической обработки.

Для прямой экспериментальной проверки наблюдаемых эффектов был использован набор последовательностей, содержащих 1, 2 или 3 кодона с



**Рис. 4 — 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  $p$ -value  $< 0.01$  обведены рамкой. (б) Распределение числа пар (кодон, позиция), для которых в перемешанных последовательностях наблюдается монотонная зависимость частоты от ФЭТ. Стрелкой показано число пар для настоящей выборки,  $p$ -value  $< 10^{-4}$ . (в) Средний нормированный эффект кодона ( $z$ -score) в зависимости от позиции. Коэффициент корреляции Пирсона  $r = -0.8$ ,  $p$ -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*.  $r = 0.33$ ,  $p$ -value =  $9.6 \times 10^{-9}$  (д) То же, что и в п. (г), после нормировки на нуклеотидный состав.  $r = -0.12$ ,  $p$ -value = 0.004

предсказанным ингибиторным эффектом, в то время как остальные кодоны обладали предсказанным положительным эффектом. Энергии сворачивания этих последовательностей значительно не различались, и сродство к участку анти-Шайна–Дальгарно было минимально. По результатам этого дополнительного эксперимента в случае одного ингибиторного кодона отрицательное влияние на трансляцию оказалось тем менее значительным, чем дальше этот кодон находится от старта трансляции (Рисунок 5). Эта закономерность соответствует нашему наблюдению, сделанному на массовых данных (Рисунок 4в). Кроме того, добавление дополнительных кодонов с отрицательным эффектом усиливало ингибирование трансляции, что опять-таки подтверждает наши наблюдения.

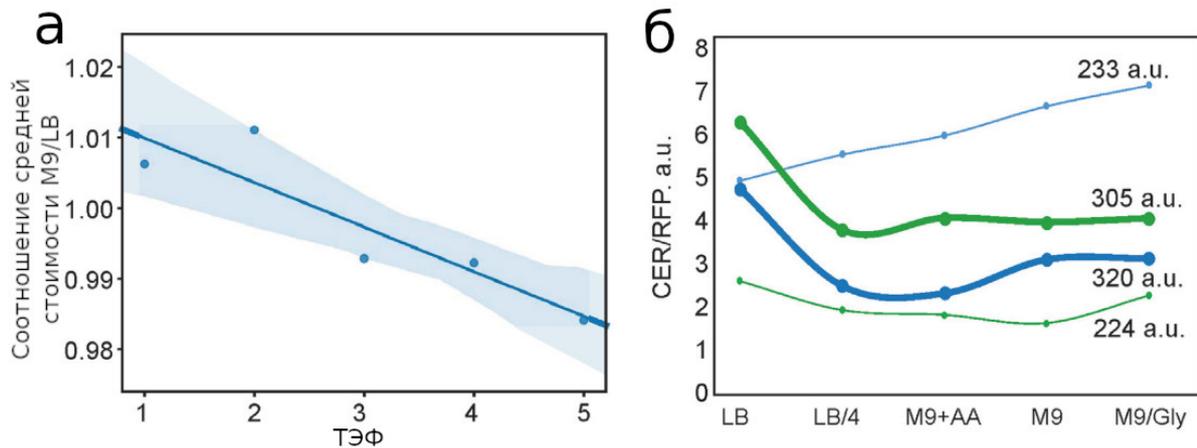


**Рис. 5** — Эффективность трансляции для проверочных последовательностей с ингибирующими кодонами. (а), (б) Схематичное представление положений ингибирующих кодонов (слева) и значений эффективности трансляции (справа).

До настоящего времени вопрос о том, увеличивают ли редкие кодоны в начале гена эффективность его трансляции, оставался непроясненным. Известно, что редкие кодоны встречаются в начале генов чаще, чем ожидается, однако это может быть следствием отбора против сильной вторичной структуры. В наших данных эффект кодона на трансляцию, напротив, слабо положительно скоррелирован с частотой использования кодона в генах *E. coli* (Рисунок 4г), но эта корреляция пропадает при контроле на позиционный нуклеотидный состав (Рисунок 4д). Таким образом, утверждение о положительном влиянии редких кодонов на эффективность трансляции не подтверждается.

Описанные выше 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 бедной среде M9. Чтобы проанализировать наблюдаемые различия, мы использовали описанную далее дополнительную информацию о метаболизме *E. coli*.



**Рис. 6** — 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).

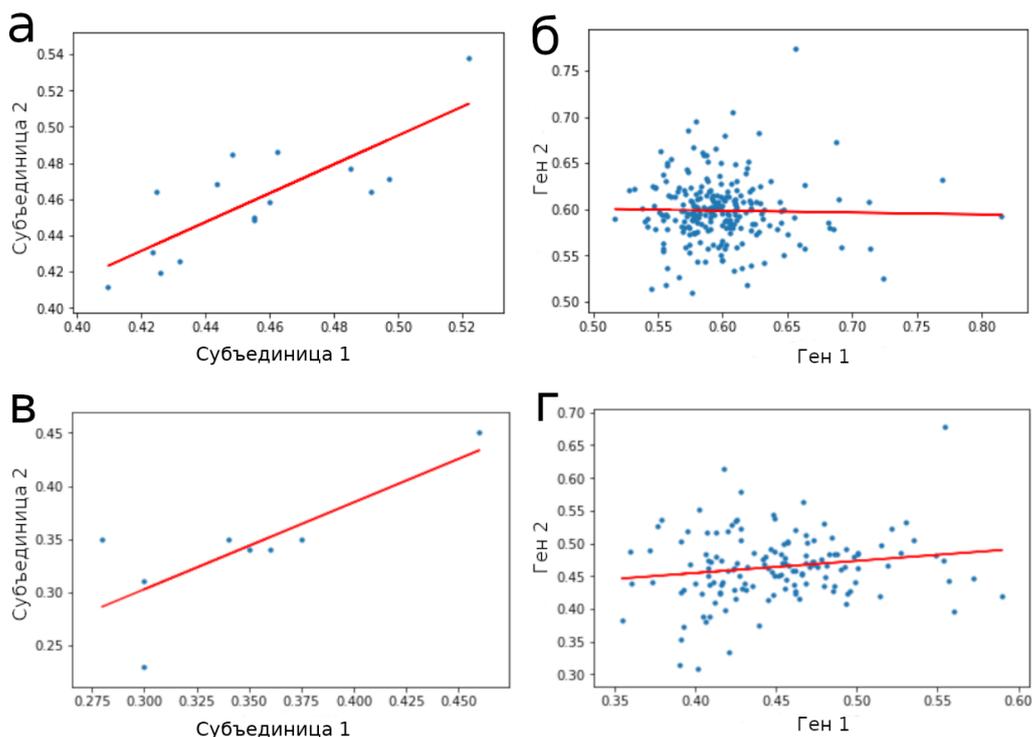
Известно, что синтез разных аминокислот требует разного количества молекул АТФ и окислительных эквивалентов, которые, в свою очередь, тоже могут конвертироваться в АТФ. Метаболическая стоимость каждой аминокислоты может быть выражена через число высокоэнергетических фосфатных связей, разрыв которых необходим 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 большими номерами), чаще (относительно богатой среды) встречаются аминокислоты с низкой метаболической стоимостью (Рисунок 6а).

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-Шайна-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, был значительно ниже в бедных средах, в то время как для последовательностей, кодирующих метаболически дешевые аминокислоты, этот тренд в одном случае менее выражен, а в другом и вовсе обратный (Рисунок 6б). 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.

В главе 3 было дополнительно исследовано влияние вторичной структуры на трансляцию, деградацию и эволюцию мРНК у *Escherichia coli*.

Мы рассмотрели мРНК генов, кодирующих эквимоллярные субъединицы одного белкового комплекса и находящихся в одном опероне. Ранее было показано, что эффективности трансляции таких пар генов с высокой точностью равны, но механизма предложено не было. В более поздней работе на данных диметилсульфат-секвенирования 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, кодирующих субъединицы белковых комплексов, и для контрольных оперонов, кодирующих белки, комплексов не образующие.

Оказалось, что степень структурированности в оперонах, кодирующих субъединицы, коррелирует, в отличие от контрольных оперонов (Рисунок 7).



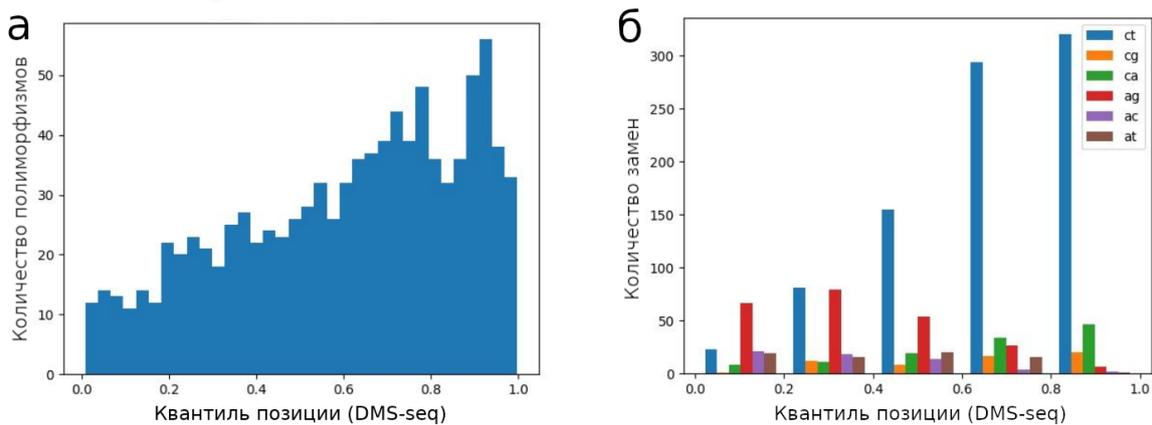
**Рис. 7** — Корреляция между уровнем структурированности мРНК для пар генов, кодирующих субъединицы белковых комплексов, и контрольных пар генов. Каждая точка соответствует одной паре генов, по горизонтальной оси отложена структурированность 5'-гена, по вертикальной — 3'-гена. Графики (а) и (б) построены 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)  $r = 0.16$ ,  $p\text{-value} = 0.047$ .

Мы проверили, что это наблюдение не связано с гомологичностью субъединиц: распределение уровней сходства генов субъединиц не отличается от такового для контрольных пар. Кроме того, мы показали, что структура всей мРНК гена, а не только ее 5'-области, важна для эффективности трансляции.

Далее мы исследовали, как вторичная структура влияет на деградацию мРНК. Эффективно транскрибируемые гены деградируют немного медленнее, однако сравнение времени жизни со структурированностью на уровне целых генов не выявило значимой корреляции.

Мы посмотрели также, как связана структура мРНК и эволюция последовательности. Для этого были рассмотрены полиморфизмы в 300 штам-

мах кишечной палочки. Оказалось, что если определять структурированные области согласно данным, полученным методом DMS-seq, то полиморфизмы значительно чаще в структурированных областях, чем в неструктурированных (Рисунок 8). Однако, это наблюдение не подтверждается при анализе данных о вторичной структуре, полученных другим методом (SHAPE). Таким образом, наблюдение о большей частоте мутаций в спаренных областях может являться артефактом используемого метода и требует дальнейших независимых подтверждений.

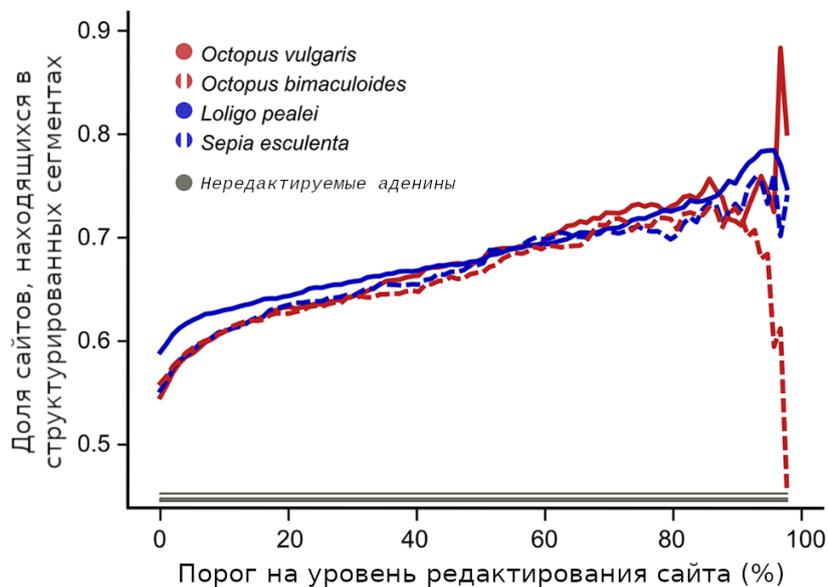


**Рис. 8 — (а) Количество полиморфизмов и вероятность спаривания позиции.** Горизонтальная ось соответствует квантилю позиции по структурированности внутри своего гена по данным DMS-seq у *Escherichia coli* str. K12 substr. MG1655. Чем ближе квантиль к единице, тем с большей вероятностью эта позиция 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*

В главе 4 приведен анализ роли вторичной структуры мРНК в редактировании аденинов у мягкотелых головоногих моллюсков (колеоидов). Дезаминирование аденина (редактирование) – это самый часто встречающийся тип модификации мРНК. 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. Есть основания полагать, что массовое редактирование мРНК у колеоидов может являться эпигенетическим механизмом, который создает функциональный эквивалент мутациям в геноме. Поэтому

изучение закономерностей редактирования мРНК у колеоидов представляется интересной задачей, имеющей важное теоретическое значение.

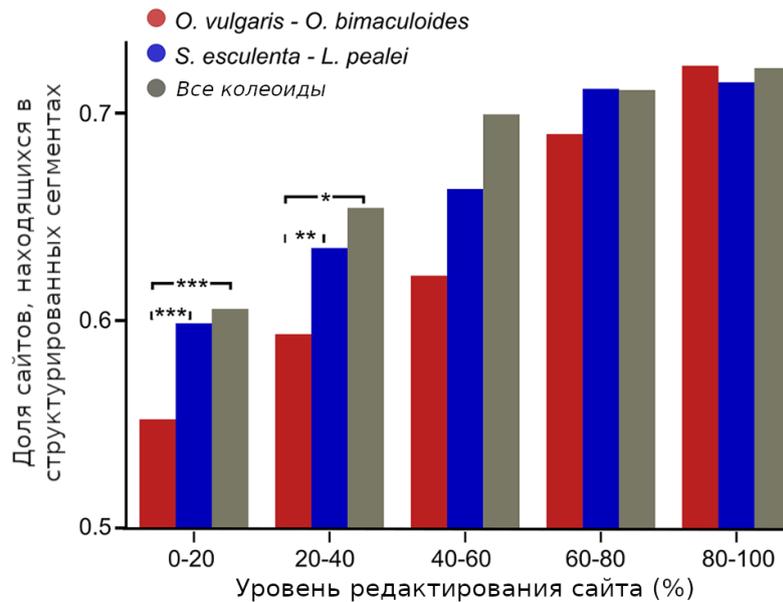
Известно, что редактирование мРНК осуществляется аденозин деаминазами семейства ADAR, для работы которых необходимо образование двойной спирали РНК. В своей работе мы описали закономерности вторичной структуры РНК вокруг редактируемых сайтов и изменения этой структуры между видами. Кроме того, мы продемонстрировали роль вторичной структуры в скоординированном редактировании аденинов.



**Рис. 9** — 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.

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 9, редактируемые аденины чаще, чем нередактируемые, находятся в структурированных сегментах. Более того, доля аденинов в структурированных участках растет с увеличением уровня редактирования.

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

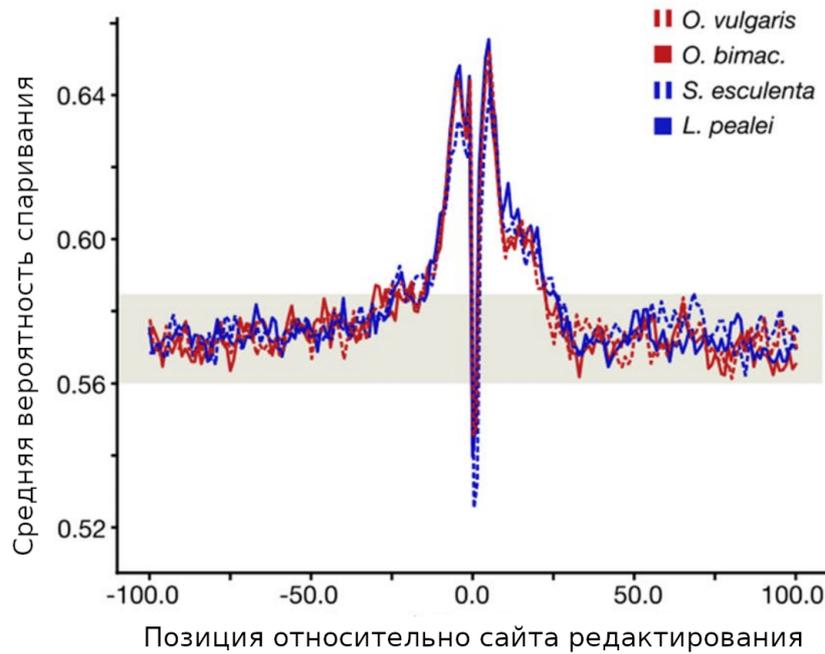


**Рис. 10** — Доля редактируемых аденинов, лежащих в структурированных сегментах, выше для консервативно редактируемых аденинов. Горизонтальная ось – порог на минимальный уровень редактирования. Красные столбцы: аденины, консервативно редактируемые у осьминогов, синие – у кальмара и каракатицы, серые – у всех четырех колеоидов. Статистическая значимость была оценена путем генерации пар распределений долей по подвыборкам из исходных выборок (bootstrap). \*\*\*  $p$ -value < 0.001, \*  $p$ -value < 0.05

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 10). Оказалось, что более консервативные сайты (редактирование которых сохраняется между более далекими видами), в четырех из пяти категорий чаще находятся в структурированных участках, чем менее консервативные. Различие статистически значимо в двух категориях с наименьшим уровнем редактирования.

Чтобы оценить размер структуры мРНК вокруг редактируемых сайтов, мы проанализировали вероятности спаривания окружающих их нуклеотидов. Для этого было использована программа *plfold* из пакета *ViennaRNA*, 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 вовлеченным в какие-либо структурные взаимодействия, сложив вероятности для всех пар, включающих данный нуклеотид. Результат этого анализа показан на Рисунке 11). Видно, что по мере удаления от сай-

та редактирования средняя вероятность находиться в спаренном состоянии падает, причем вне участка  $\pm 25$  нуклеотидов вокруг сайта колебания вероятности не отличимы от шума.



**Рис. 11** — 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.

Аденины, находящиеся близко в последовательности, часто редактируются одновременно. Однако благодаря вторичной структуре могут оказаться пространственно сближены и сайты, находящиеся на значительном линейном 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.

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, ре-

дактируются 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 12). Этот результат показывает, что несмотря на в среднем низкую стабильность (вероятность спаривания на расстояниях более 25 нуклеотидов не отличима от шума, см. Рисунок 11), вторичная структура не только определяет возможность и силу редактирования, но и позволяет скоординировать далекие события редактирования между собой.

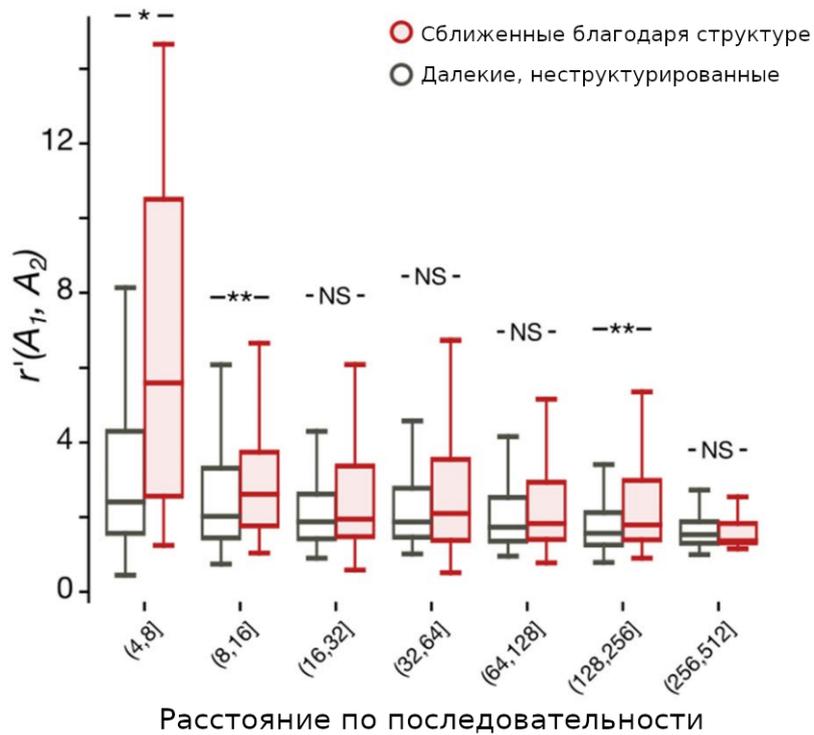


Рис. 12 — Распределения значений  $r'$  для пар сайтов, сближенных благодаря вторичной структуре РНК (красные боксы), и для пар сайтов, далеких друг от друга и структурой не сближенных (серые боксы). По горизонтальной оси отложено расстояние между парами сайтами по последовательности. Звездочками помечена 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 *Escherichia 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 эквимоллярных 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.

## 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. Ponomareva, 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.