

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ  
имени М.В. ЛОМОНОСОВА  
ИНСТИТУТ ФУНКЦИОНАЛЬНОЙ ГЕНОМИКИ

*На правах рукописи*

**Комарова Екатерина Сергеевна**

**Изучение особенностей 5'-нетранслируемой области  
бактериальных мРНК, влияющих на эффективность трансляции,  
с помощью библиотек репортёрных конструкций**

Специальность:

1.5.3. Молекулярная биология

**ДИССЕРТАЦИЯ**

на соискание ученой степени  
кандидата химических наук

Научный руководитель:

д.х.н., член-корр. РАН Сергиев Пётр Владимирович

Москва – 2023

## ОГЛАВЛЕНИЕ

<b>Список сокращений</b> .....	<b>6</b>
<b>Введение</b> .....	<b>8</b>
Актуальность проблемы .....	8
Цели и задачи исследования .....	9
Объект и предмет исследования .....	9
Научная новизна исследования .....	10
Теоретическая и практическая значимость исследования .....	11
Методология диссертационного исследования .....	11
Основные положения, выносимые на защиту .....	12
Степень достоверности результатов .....	12
Апробация работы .....	12
Публикации по теме диссертации .....	13
Личный вклад автора .....	14
Структура и объём диссертации .....	14
<b>Обзор литературы</b> .....	<b>15</b>
1. Роль 5'-нетранслируемой области в эффективности трансляции в бактериальных клетках .	15
1.1. Особенности организации мРНК у бактерий .....	15
1.2. Бактериальная трансляция: краткий обзор .....	17
1.3. Разнообразие 5'-UTR у бактерий .....	23
1.3.1. Мотивы 5'-UTR генов рибосомных белков и 3'-конца 16S рРНК у бактерий .....	23
1.3.2. Полногеномный анализ 5'-UTR у бактерий .....	26
1.4. Бактериальная трансляция: детальный обзор инициации .....	30
1.4.1. Инициация трансляции канонических мРНК .....	30
1.4.2. Инициация трансляции неканонических мРНК .....	43
2. Подходы к изучению роли 5'-UTR в эффективности трансляции .....	47
2.1. 5'-UTR – мишень для регуляции эффективности трансляции .....	47
2.2. Разнообразие подходов к изучению 5'-UTR в эффективности трансляции .....	48
2.2.1. Широкомасштабный анализ генов .....	48
2.2.2. Использование репортёрных конструкций .....	50
2.2.3. Сайт-направленный мутагенез .....	51
2.2.4. Аффинное выделение рибосом с мРНК .....	52
2.2.5. Оценка стабильности вторичных структур .....	53
2.2.6. Тоупринтинг .....	56

2.2.7. Секвенирование нового поколения (NGS) .....	57
2.2.8. Рибосомный профайлинг (Ribo-seq) .....	60
2.2.9. Термодинамическое моделирование .....	62
2.2.10. Flow-seq метод .....	63
2.2.10.1. История появления и развития метода Flow-seq .....	63
2.2.10.2. Применение метода Flow-seq для оценки эффективности трансляции .....	68
2.2.10.3. Вклад метода Flow-seq в синтетическую биологию .....	73
<b>Результаты и обсуждение .....</b>	<b>77</b>
1. Изучение влияния спейсерного участка в 5'-UTR на эффективность трансляции .....	77
1.1. Создание репортёрных конструкций с 4N-фрагментом в спейсерном участке 5'-UTR между последовательностью SD и старт-кодоном гена <i>cer</i> .....	77
1.2. Клонирование полученной библиотеки плазмидных ДНК .....	82
1.3. Валидация метода Flow-seq для широкомасштабного анализа эффективности трансляции .....	83
1.4. Применение метода Flow-seq для разделения клеток <i>E. coli</i> с разными репортёрными конструкциями и последующего анализа их последовательностей .....	87
1.5. Оценка влияния особенностей спейсерного участка на эффективность трансляции .....	89
2. Изучение влияния рандомизированных 5'-UTR на эффективность трансляции .....	94
2.1. Создание репортёрных конструкций с рандомизированной 5'-UTR перед стартовым кодоном гена <i>cer</i> .....	94
2.2. Клонирование полученной библиотеки плазмидных ДНК .....	97
2.3. Применение метода Flow-seq для разделения клеток <i>E. coli</i> с разными репортёрными конструкциями, содержащими рандомизированную 5'-UTR, и последующего анализа их последовательностей .....	98
2.4. Оценка влияния особенностей рандомизированных 5'-UTR на эффективность трансляции .....	101
2.5. Нестандартные 5'-UTR с высокой эффективностью трансляции .....	108
2.6. Природные 5'-UTR <i>E.coli</i> перед геном <i>cer</i> с высокой эффективностью трансляции .....	117
3. Изучение влияния природных 5'-UTR на эффективность трансляции .....	120
3.1. Отбор природных 5'-UTR и создание репортёрных конструкций с ними перед стартовым кодоном гена <i>cer</i> .....	120
3.2. Применение метода Flow-seq для разделения клеток <i>E. coli</i> с разными репортёрными конструкциями, содержащими природные 5'-UTR, и последующего анализа их последовательностей .....	124
3.3. Оценка влияния вторичной структуры и состава природных 5'-UTR на эффективность	

трансляции .....	125
3.4. Сравнение результатов, полученных с помощью методов Flow-seq и Ribo-seq .....	128
4. Особенности применения метода Flow-seq для оценки эффективности трансляции .....	130
<b>Заключение .....</b>	<b>133</b>
<b>Выводы .....</b>	<b>135</b>
<b>Материалы и методы .....</b>	<b>136</b>
1. Материалы .....	136
1.1. Используемые приборы .....	136
1.2. Реактивы и биопрепараты .....	136
1.3. Состав буферов и сред .....	137
1.4. Штаммы, плазмиды и олигонуклеотиды .....	138
2. Методы .....	141
2.1. Манипуляции с ДНК: клонирование .....	141
2.1.1. Определение концентрации ДНК в растворе .....	141
2.1.2. Разделение в агарозном геле ДНК-фрагментов.....	141
2.1.3. Выделение ДНК-фрагментов из агарозного геля.....	142
2.1.4. Приготовление векторов и вставок .....	142
2.1.5. Очистка фенолом и хлороформом и осаждение этанолом.....	142
2.1.6. Очистка ПЦР-продуктов из реакционной смеси.....	143
2.1.7. Лигирование.....	143
2.1.8. Выделение плазмидной ДНК .....	143
2.1.9. Полимеразная цепная реакция (ПЦР) .....	144
2.1.10. Создание репортёрной конструкции pRFPCER BamHI .....	145
2.1.11. Создание отдельных конструкций с нестандартными и их мутированными вариантами или природными 5'-UTR перед генами <i>cer</i> или <i>rfp</i> .....	145
2.1.12. Секвенирование ДНК.....	145
2.1.12.1. Секвенирование плазмидной ДНК или ПЦР-продуктов .....	145
2.1.12.2. Высокопроизводительное секвенирование (NGS) .....	146
2.2. Манипуляции с клетками .....	146
2.2.1. Приготовление компетентных клеток <i>Escherichia coli</i> .....	146
2.2.2. Трансформация и электропорация компетентных клеток <i>Escherichia coli</i> .....	147
2.2.3. Сортировка клеток .....	147
2.2.4. Измерение флуоресценции с помощью спектрофлуориметра .....	148
2.3. Биоинформатический анализ результатов NGS .....	148
2.3.1. Анализ результатов для контрольного набора конструкций .....	148

2.3.2. Анализ результатов для 4N библиотеки .....	149
2.3.2.1. Обработка полученных ридов и их распределение по фракциям .....	149
2.3.2.2. Оценка нуклеотидного состава последовательностей 5'-UTR.....	149
2.3.2.3. Оценка стабильности вторичной структуры мРНК.....	150
2.3.2.4. Оценка сходства последовательности 5'-UTR с SD.....	151
2.3.3. Анализ результатов для 20N и 30N библиотек .....	151
2.3.3.1. Обработка полученных ридов и их распределение по фракциям .....	151
2.3.3.2. Оценка нуклеотидного состава последовательностей 5'-UTR.....	151
2.3.3.3. Оценка стабильности вторичной структуры мРНК.....	152
2.3.3.4. Анализ SD-подобных последовательностей в 5'-UTR.....	152
2.3.4. Анализ результатов для библиотеки природных 5'-UTR.....	153
2.3.4.1. Обработка полученных ридов и их распределение по фракциям .....	153
2.3.4.2. Оценка стабильности вторичной структуры мРНК и анализ SD-подобных последовательностей в 5'-UTR.....	153
<b>Благодарности .....</b>	<b>154</b>
<b>Список литературы .....</b>	<b>156</b>
<b>Приложение 1 .....</b>	<b>172</b>
<b>Приложение 2 .....</b>	<b>174</b>
<b>Приложение 3 .....</b>	<b>177</b>

## СПИСОК СОКРАЩЕНИЙ

ДНК	дезоксирибонуклеиновая кислота
кДНК	комплементарная ДНК
РНК	рибонуклеиновая кислота
мРНК	матричная РНК
рРНК	рибосомная РНК
тРНК	транспортная РНК
TIR/TIS	сайт инициации трансляции
IR	участок инициации
5'-UTR/3'-UTR	5'/3'-нетранслируемая область мРНК
ORF	открытая рамка считывания
CDS	белок-кодирующая последовательность
RBS	сайт связывания рибосом
SD	последовательность Шайна-Дальгарно
анти-SD/anti-SD/aSD	последовательность анти-Шайна-Дальгарно
70S	прокариотическая рибосома
30S/SSU и 50S/LSU	малая и большая субчастицы прокариотической рибосомы
30S PIC/70S PIC	30S/70S пре-инициаторный комплекс
P/I	пептидил/инициаторное положение
30S IC/70S IC	30S/70S инициаторный комплекс
ПТЦ/PTC	пептидил-трансферазный центр
ДЦ/CD/CDR	декодирующий центр
СРП/SRL	сарцин-рициновая петля
IF1, IF2, IF3	факторы инициации 1, 2, 3
EF-Tu, EF-G	факторы элонгации Tu, G
RF1, RF2, RF3, RRF	факторы терминации 1, 2, 3, фактор рециклирования
aa-тРНК	аминоацил-тРНК
пептидил-тРНК	тРНК, связанная с полипептидом
fMet/fMet-тРНК	формилметионин/формилметионин-тРНК
fMet-тРНК <sup>fMet</sup>	актуально для всех тРНК: Met – метионин – аминокислота, присоединенная к тРНК, тРНК <sup>Met</sup> – тРНК, специфичная по отношению к аминокислоте Met – метионину
АТФ/АТР	аденозин-5'-трифосфат
ГТФ/GTP, GDP	гуанозин-5'-трифосфат, гуанозин-5'-дифосфат

dNTP	дезоксирибонуклеозидтрифосфат
поли(A/C)	последовательность нуклеотидов, состоящая только из аденинов/цитозинон
нт	нуклеотиды
ПЦР	полимеразная цепная реакция
FRET	флуоресцентный резонансный перенос энергии
NGS	секвенирование нового поколения
DNA-seq/RNA-seq	ДНК-секвенирование/РНК-секвенирование
Ribo-seq	рибосомный профайлинг и секвенирование клеточного транскриптома
ChIP-seq	иммунопреципитация хроматина и секвенирование ДНК, связанной с белками
Flow-seq/FACS-seq	проточная цитофлюориметрия с последующим NGS
ПААГ	полиакриламидный гель
РСА	рентгеноструктурный анализ
ЯМР	ядерный магнитный резонанс
крио-ЭМ	криоэлектронная микроскопия
<i>E. coli</i>	<i>Escherichia coli</i>
<i>T. thermophiles</i>	<i>Thermus thermophiles</i>
<i>B. subtilis</i>	<i>Bacillus subtilis</i>
<i>L. lactis</i>	<i>Lactococcus lactis</i>
<i>G. stearothermophilus</i>	<i>Geobacillus stearothermophilus</i>
MFE	минимальная энергия укладки
об./мин	обороты в минуту
t °C	температура, вместо t может стоять любое число
ц/ф	центрифугирование
X/ Y"	количество минут/секунд, вместо X/Y может стоять любое число
ед.	единица(-ы) (в отношении фермента)
п.л.	печатный(-е) лист(-ы)

Аминокислотные остатки, нуклеотиды, олиго- и полинуклеотиды обозначены символами, рекомендованными Комиссией по номенклатуре союза биохимиков (IUB). При обозначении нуклеозидов или дуплексов нуклеиновых кислот префикс «d» (дезоксид) может быть опущен. Для латинских названий генов используется курсив и строчные буквы, а для белков – прямой шрифт и заглавные буквы.

# ВВЕДЕНИЕ

## Актуальность проблемы

Трансляция – фундаментальный процесс, лежащий в основе жизнедеятельности всех организмов на Земле и заключающийся в биосинтезе белков. Её ключевым участником выступает такой сложноустроенный макромолекулярный рибонуклеопротеидный комплекс, как рибосома, осуществляющая декодирование информации, записанной в матричной РНК (мРНК), и переводящая её в последовательность аминокислот, формирующую белок [1-2]. мРНК участвует в данном процессе не только в качестве пассивного посредника. Она способна предопределять эффективность трансляции благодаря наличию определённых функциональных элементов [3-10].

Эффективность бактериальной трансляции в первую очередь зависит от такого элемента, как 5'-нетранслируемая область (5'-UTR), предшествующая последовательности, кодирующей белок [3, 6]. Компьютерный и экспериментальный анализ разнообразия бактериальных 5'-UTR выявил в них ряд особенностей, влияющих на уровень биосинтеза белка. Наиболее важной и лучше всего изученной из них является последовательность Шайна-Дальгарно (SD), которая расположена в сайте посадки рибосомы (RBS) в 5'-UTR мРНК и комплементарна 3'-концевому участку 16S рибосомной РНК (рРНК) в малой 30S субчастице рибосомы [9-13]. Для обеспечения высокой эффективности трансляции необходимо расположение последовательности SD на оптимальном расстоянии от инициаторного кодона (5-9 нуклеотидов (нт)), при этом её длина может изменяться в пределах 4-8 нт. А иногда можно встретить и несколько SD в одной 5'-UTR. Дополнительным сайтом узнавания рибосомы может выступать AU-богатая область 5'-UTR, где преобладают основания аденина и урацила, и с которой взаимодействует белок bS1 малой субчастицы рибосомы [6, 14-15].

Для увеличения эффективности белкового синтеза необходимо, чтобы участок инициации трансляции (TIR) был полностью одноцепочечным или имел укладку вторичной структуры, которую легко разрушить. Ещё одной особенностью мРНК, влияющей на уровень биосинтеза белка в зависимости от его нуклеотидного состава и стабильности вторичной структуры, является начальный «разгонный» участок кодирующей области, следующий за старт-кодоном.

Однако не для всех мРНК наблюдаемые *in vitro/in vivo* уровни эффективности трансляции можно объяснить, только исходя из перечисленных выше особенностей. Для решения задач микробной биотехнологии важно уметь настраивать определённые уровни биосинтеза белков. С этой целью разработали разные способы предсказаний *in silico* скорости инициации трансляции конкретной мРНК, в основе которых лежит термодинамическая модель, оценивающая силу молекулярных взаимодействий 30S комплекса с мРНК-транскриптом. Валидация результатов предсказаний *in vivo* не всегда свидетельствует о надёжности данных, полученных *in silico*.

Поэтому остаётся актуальным вопрос о том, что именно позволяет той или иной мРНК обеспечить быструю инициацию и эффективную трансляцию. Такая информация представляла бы интерес с фундаментальной точки зрения, более подробно раскрывая особенности такого сложного процесса, как трансляция, и пролила бы новый свет на возможности регуляции уровня биосинтеза белка с практической точки зрения, позволяя решить биотехнологические задачи. Данная научно-исследовательская работа посвящена изучению особенностей различных синтетических и природных 5'-нетранслируемых областей бактериальных мРНК, влияющих на эффективность трансляции, с помощью библиотек репортёрных конструкций.

Разнообразные подходы, разработанные на сегодняшний день, позволяют активно изучать функциональную значимость отдельных участков мРНК для биосинтеза белка. Комбинирование различных методов даёт возможность работать одновременно с библиотеками из десятков тысяч репортёрных конструкций. Успешным примером такой комбинации выступает метод Flow-seq, сочетающий проточную цитометрию с последующим секвенированием нового поколения (NGS). Он нашёл активное применение при определении уровня биосинтеза белка сразу для набора вариантов мРНК, кодирующих этот белок. В данной работе метод Flow-seq является основным, для него приведён сравнительный анализ с другими подходами для оценки эффективности трансляции, представлены особенности его применения и получаемые с его помощью результаты.

### **Цели и задачи исследования**

Целью данной работы является изучение особенностей 5'-UTR бактериальных мРНК, влияющих на эффективность трансляции, с помощью библиотек репортёрных конструкций.

Для достижения поставленной цели определены следующие задачи:

1. Создание библиотек репортёрных конструкций с различными синтетическими и природными 5'-UTR перед геном флуоресцентного белка.
2. Адаптирование метода Flow-seq для более подробного изучения влияния особенностей 5'-UTR бактериальных мРНК на эффективность трансляции.
3. Анализ полученных данных и сравнение их с результатами других работ с применением схожих или отличных методов.
4. Поиск новых эффективных сайтов посадки рибосомы в 5'-UTR бактериальных мРНК и определение основных элементов таких сайтов, обуславливающих высокий уровень биосинтеза белка.

### **Объект и предмет исследования**

В данном исследовании объект для изучения представлен 5'-нетранслируемой областью бактериальных мРНК, являющейся одним из регуляторных элементов в обеспечении разного

уровня эффективности трансляции.

Предметом настоящего исследования является длина и нуклеотидный состав 5'-UTR бактериальных мРНК наряду со стабильностью и особенностями их вторичной структуры, способностью взаимодействовать с 3'-концевым участком 16S рРНК малой 30S субчастицы рибосомы.

### **Научная новизна исследования**

В данной научно-исследовательской работе впервые проводится подробный анализ *in vivo* более 24,5 тысяч различных последовательностей 5'-нетранслируемой области бактериальных мРНК, представленных как синтетическими, так и природными вариантами в репортёрных конструкциях перед сенсорными генами флуоресцентных белков CER (голубого) или RFP (красного). Такое разнообразие мРНК позволило обеспечить изменение эффективности трансляции в пределах четырёх порядков между самым высоким и самым низким уровнем синтезированного белка.

В то же время в представленной работе продемонстрированы успешные валидация и применение метода Flow-seq для разделения клеток с библиотеками репортёрных конструкций с помощью FACS, флуоресцентной проточной цитометрии, и последующего анализа участков их последовательностей путём высокопроизводительного секвенирования. Полученные результаты свидетельствуют о возможности использования комплексного метода Flow-seq для надёжной и точной широкомасштабной оценки эффективности трансляции. Также в работе, опираясь на опыт предыдущих и нынешнего исследований, выявлены особенности его применения.

Для полученных библиотек 5'-UTR различной длины провели анализ их нуклеотидного состава, стабильности и особенностей их вторичной структуры, их способности образовывать комплементарные взаимодействия с 3'-концевым участком 16S рРНК 30S субчастицы рибосомы. В результате выделили несколько различных особенностей, свойственных набору вариантов последовательностей 5'-UTR, наиболее эффективных в трансляции. Среди них определены все известные детерминанты, обеспечивающие увеличение уровня синтезируемого белка и представленные последовательностью Шайна-Дальгарно оптимальной длины и на оптимальном расстоянии от старт-кодона, AU-богатыми участками, предшествующими SD. Также выявлено отрицательное влияние на инициацию трансляции маскировки сайта связывания рибосомы за счёт образования вторичной структуры мРНК, известное и прежде. В дополнение к известным найдено несколько особенностей, способствующих увеличению эффективности трансляции и ранее не описанных, таких как низкая доля остатков цитидина, AG-повторы и множественные последовательности SD. AG-повторы можно идентифицировать как новый энхансер трансляции, хотя и менее эффективный, чем последовательность SD, для нескольких природных 5'-UTR.

Подробный мутационный анализ некоторых вариантов 5'-UTR, обеспечивающих высокий уровень биосинтеза белка, позволил отметить возможный аддитивный эффект нескольких последовательностей SD, дополнительного старт-кодона, приводящий к увеличению эффективности трансляции. Опираясь на полученные результаты представлены некоторые рекомендации по выбору последовательностей 5'-UTR для бактериальных систем экспрессии.

Для библиотеки природных 5'-UTR впервые провели сравнение результатов, полученных с помощью комбинированных методов Flow-seq и Ribo-seq (рибосомного профилирования) для широкомасштабной оценки эффективности трансляции, и выявили хорошее соответствие между ними. Это подтверждает возможность успешного использования такого метода, как Flow-seq, для анализа библиотек репортёрных конструкций с 5'-UTR различной длины и состава перед геном сенсорного флуоресцентного белка (например, CER), а также для оценки влияния 5'-UTR на эффективность трансляции в системе *Escherichia coli*.

### **Теоретическая и практическая значимость исследования**

Полученные результаты исследования представляют интерес с фундаментальной точки зрения, поскольку раскрывают новые аспекты такого сложного процесса, как биосинтез белка, более подробно позволяют взглянуть на систему его регуляции и лучше понять, как особенности 5'-нетранслируемой области бактериальных мРНК влияют на эффективность трансляции.

В то же время с практической точки зрения результаты данной научно-исследовательской работы проливают новый свет на возможности регуляции уровня биосинтеза белка, требуемой для решения биотехнологических задач. Они могут быть использованы на практике для создания систем регулируемой экспрессии генов в клетках *Escherichia coli*, в том числе при необходимости тонкой настройки уровней экспрессии экзогенных генов белков, входящих, например, в состав гетеромультимерного комплекса в заданном стехиометрическом соотношении.

### **Методология диссертационного исследования**

Для достижения цели и решения поставленных задач в представленной работе применили комбинацию классических давно используемых и современных методов и подходов, активно применяемых в мире для проведения такого рода исследований в области биоорганической химии, молекулярной биологии, микробиологии, биоинформатики и биоинженерии, статистики. Классические методы представлены манипуляциями с ДНК для проведения клонирования, с клеточными культурами *E. coli*, основами статистики. Среди использованных современных методов можно выделить Flow-seq, сочетающий в себе сортировку клеток, основанную на флуоресценции, и высокопроизводительное секвенирование. Благодаря Flow-seq появилась возможность с высокой точностью разделять на фракции библиотеки клеток, несущих

определённые варианты репортёрных конструкций, и идентифицировать последовательности представленных вариантов в каждой фракции. Дальнейший широкомасштабный анализ с применением актуальных биоинформатических подходов позволяет выделить закономерности между определёнными последовательностями и уровнями эффективности трансляции, оценить значимость их влияния на биосинтез белка.

### **Основные положения, выносимые на защиту**

1. Метод Flow-seq успешно адаптирован для высокопроизводительного определения эффективности трансляции синтетических и природных 5'-нетранслируемых областей, а следовательно и участков посадки рибосомы, многократно превышающих их природное разнообразие у *E. coli*.
2. Высокая эффективность трансляции некоторых мРНК не может быть объяснена только наиболее изученными особенностями: взаимодействием рибосомы с последовательностью Шайна-Дальгарно либо наличием AU-богатого энхансера на 5'-конце мРНК, или отсутствием ингибирующего влияния вторичной структуры мРНК.
3. Эффективность трансляции уменьшается с увеличением доли остатков цитозина в последовательности 5'-нетранслируемой области.
4. Дополнительный старт-кодон и множественные последовательности Шайна-Дальгарно имеют аддитивное влияние на эффективность трансляции.
5. AG-повторы в 5'-нетранслируемой области представляют собой энхансер трансляции.
6. мРНК с природными 5'-нетранслируемыми областями менее вариабельны по эффективности трансляции, чем рандомизированные последовательности. В более длинных, 30-звенных 5'-нетранслируемых областях чаще встречаются благоприятные для трансляции варианты из случайных последовательностей, чем в 20-звенных.

### **Степень достоверности результатов**

Достоверность результатов представленного научного исследования подтверждена их воспроизводимостью, статистической обработкой полученных данных. Все проведённые экспериментальные процедуры соответствуют цели работы и поставленным задачам. При их проведении использовали современное оборудование и реактивы от ведущих мировых компаний. Итоговые результаты опубликованы в рецензируемых международных и отечественных журналах с индексированием в базах данных Web of Science и Scopus.

### **Апробация работы**

Результаты диссертационной работы представлены на заседании Учёного совета

института функциональной геномики МГУ имени М.В. Ломоносова и на заседании кафедры химии природных соединений Химического факультета МГУ имени М.В. Ломоносова, и на следующих четырёх международных конференциях:

1. XXII Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2015», Москва, Россия, 13-17 апреля 2015.
2. RNA 2014, Квебек, Канада, 2014.
3. XXI Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2014», Москва, Россия, 7-11 апреля 2014.
4. XX Международная научная конференция студентов, аспирантов и молодых учёных «Ломоносов-2013», Москва, Россия, 8-13 апреля 2013.

### Публикации по теме диссертации

По теме диссертационной работы опубликовано 4 научных труда, все из них статьи в рецензируемых научных журналах с индексированием в базах данных Web of Science и Scopus.

Статьи в рецензируемых научных изданиях, индексированных в базах данных Web of Science и Scopus, рекомендованных для защиты в диссертационном совете МГУ.014.2 по специальности 1.5.3. Молекулярная биология (химические науки):

1. **Komarova Ekaterina S.**, Dontsova Olga A., Pyshnyi Dmitry V., Kabilov Marsel R., Sergiev Petr V. Flow-Seq Method: Features and Application in Bacterial Translation Studies // *Acta Naturae* (англоязычная версия), 2022. Vol. 14. № 4. P. 20-37. **IF 2,0 (Web of Science)**. [16].  
**doi: 10.32607/actanaturae.11820**
2. **Komarova Ekaterina S.**, Slesarchuk Anna N., Rubtsova Maria P., Osterman Ilya A., Tupikin Alexey E., Pyshnyi Dmitry V., Dontsova Olga A., Kabilov Marsel R., Sergiev Petr V. Flow-Seq Evaluation of Translation Driven by a Set of Natural *Escherichia coli* 5'-UTR of Variable Length // *International Journal of Molecular Sciences*, 2022. Vol. 23. № 20. P. 12293. **IF 5,6 (Web of Science)**. [17].  
**doi: 10.3390/ijms232012293**
3. **Komarova Ekaterina S.**, Chervontseva Zoya S., Osterman Ilya A., Evfratov Sergey A., Rubtsova Maria P., Zatsepin Timofei S., Semashko Tatiana A., Kostryukova Elena S., Bogdanov Alexey A., Gelfand Mikhail S., Dontsova Olga A., Sergiev Petr V. Influence of the spacer region between the Shine-Dalgarno box and the start codon for fine-tuning of the translation efficiency in *Escherichia coli* // *Microbial Biotechnology*, 2020. Vol. 13. № 4. P. 1254-1261. **IF 5,7 (Web of Science)**. [18].  
**doi: 10.1111/1751-7915.13561**
4. Evfratov S.A.<sup>1</sup>, Osterman Ilya A.<sup>1</sup>, **Komarova Ekaterina S.<sup>1</sup>**, Pogorelskaya A.M., Rubtsova M.P., Zatsepin T.S., Semashko T.A., Kostryukova E.S., Mironov A.A., Burnaev E., Krymova E., Gelfand M.S., Govorun V.M., Bogdanov A.A., Sergiev P.V., Dontsova O.A. Application of sorting and next

generation sequencing to study 5'-UTR influence on translation efficiency in *Escherichia coli* // *Nucleic Acids Research*, 2017. Vol. 45. № 6. P. 3487-3502. (<sup>1</sup>-равный вклад авторов). **IF 14,9 (Web of Science)**. [19]. **doi: 10.1093/nar/gkw1141**

### **Личный вклад автора**

Поиск и анализ литературных данных, постановка задач, планирование и проведение экспериментов, анализ и представление полученных результатов, апробация их на научных конференциях, подготовка и написание научных статей осуществлялись автором лично или при его непосредственном участии.

В обзорной статье [16] анализ литературных данных, особенностей и применения метода Flow-seq в изучении бактериальной трансляции выполнен полностью автором лично. В статье [17] с природными вариантами – при определяющем вкладе автора в рамках дипломной работы выполняла Слесарчук А. Н. В статьях [18-19] с рандомизированными и природными библиотеками 5'-UTR автором лично проведены все экспериментальные процедуры. Исключение составляют анализ и сортировка клеток на проточном цитометре [17-19], выполненные совместно с профессором Рубцовой М.П., и секвенирования нового поколения, произведённые совместно с руководителем ЦКП «Геномика» СО РАН Кабиловым М. Р. [17] и научной группой академика РАН Говоруна В.М. [18-19]. Биоинформатический анализ полученных данных проводился коллегами: руководителем ЦКП «Геномика» СО РАН Кабиловым М. Р. [17], Червонцевой З. С. [18], Евфратовым С. А. и научной группой профессора Гельфанда М.С. [19], при активном участии автора.

### **Структура и объем диссертации**

Данная кандидатская диссертация изложена на 177 страницах и состоит из следующих основных разделов: «Список сокращений», «Введение», «Обзор литературы», «Результаты и обсуждение», «Заключение», «Выводы», «Материалы и методы», «Список литературы». Также в работу включены 3 приложения, благодарности. В работе представлены 72 рисунка, 8 таблиц, отдельно раздел «Материалы и методы» содержит 1 рисунок и 1 таблицу, а в приложении данные приведены в 5 таблицах. Библиография включает список из 261 источников литературы.

## ОБЗОР ЛИТЕРАТУРЫ

### 1. Роль 5'-нетранслируемой области в эффективности трансляции в бактериальных клетках

#### 1.1. Особенности организации мРНК у бактерий

Процесс биосинтеза белка является ключевым в жизнедеятельности всех организмов, включая бактерии. Одним из основных участников этого процесса выступает матричная РНК (мРНК), не только как пассивный носитель обеспечивающая передачу наследственной информации с ДНК на рибосому [1-2], но и предопределяющая эффективность трансляции.

Участок инициации трансляции (TIR) на мРНК характеризуется наличием стартового кодона и добавочных элементов, необходимых для правильной инициации. Большинство мРНК содержат 5'-нетранслируемую область (5'-UTR), предшествующую старт-кодону [3] (Рис. 1). В качестве стартовых кодонов могут выступать AUG, GUG, UUG с частотой встречаемости 90, 8 и 1% соответственно, исключительным является AUU триплет, используемый как стартовый кодон всего в двух генах: *infC*, кодирующем IF3, и *pcnB*, кодирующем поли(А) полимеразу *E. coli* [4].

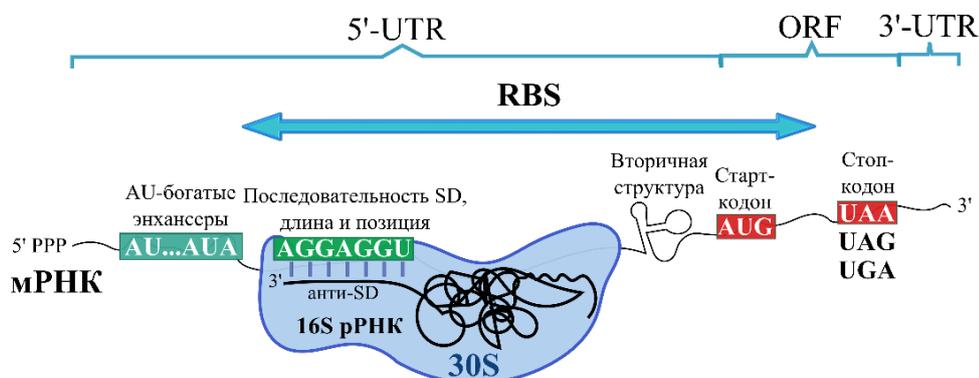


Рисунок 1. Строение мРНК у бактерий. 5'-UTR и 3'-UTR – 5'- и 3'-нетранслируемые области мРНК соответственно, ORF – открытая рамка считывания, RBS – сайт связывания рибосомы, SD и анти-SD – последовательности Шайна-Дальгарно и анти-Шайна-Дальгарно, 30S – малая субчастица рибосомы. Рисунок выполнен в программе Inkscape.

В ходе инициации трансляции происходит специфичное взаимодействие мРНК как с транспортной РНК (тРНК), так и с малой 30S субчастицей рибосомы. Для связывания с малой субчастицей мРНК содержит особый участок, называемый сайтом посадки рибосомы (RBS) длиной около 30 нуклеотидов (Рис. 1) [5-8]. Следует отметить, что в литературе встречается и другая интерпретация понятия сайта связывания рибосомы, определяющая его как участок, предшествующий стартовому кодону и ограничивающийся только консенсусом, называемым последовательностью Шайна-Дальгарно (SD) [9-11]. В то же время в литературе можно найти и

другие аббревиатуры со схожим значением: RDS, сайт докинга рибосомы длиной около 30 нуклеотидов вокруг стартового кодона, RRS, сайт узнавания рибосомой длиной около 10 пар оснований с консенсусной последовательностью SD [12]. В данной работе под сайтом посадки рибосомы (RBS) будем понимать именно участок мРНК, покрываемый рибосомой при связывании с ней во время инициации трансляции, длиной около 30 нуклеотидов.

Последовательность Шайна-Дальгарно, входящая в состав сайта посадки рибосомы в канонических мРНК, в свою очередь взаимодействует с малой субчастицей рибосомы за счет образования дуплекса с комплементарной последовательностью на 3'-конце 16S рРНК 30S субчастицы, называемой последовательностью анти-Шайна-Дальгарно (анти-SD/anti-SD/aSD) [13] (Рис. 2). В качестве дополнительного сайта узнавания рибосомы может выступать область 5'-UTR мРНК, богатая основаниями аденина и урацила (AU-богатая) и взаимодействующая с bS1 белком малой субчастицы рибосомы [6, 14-15].

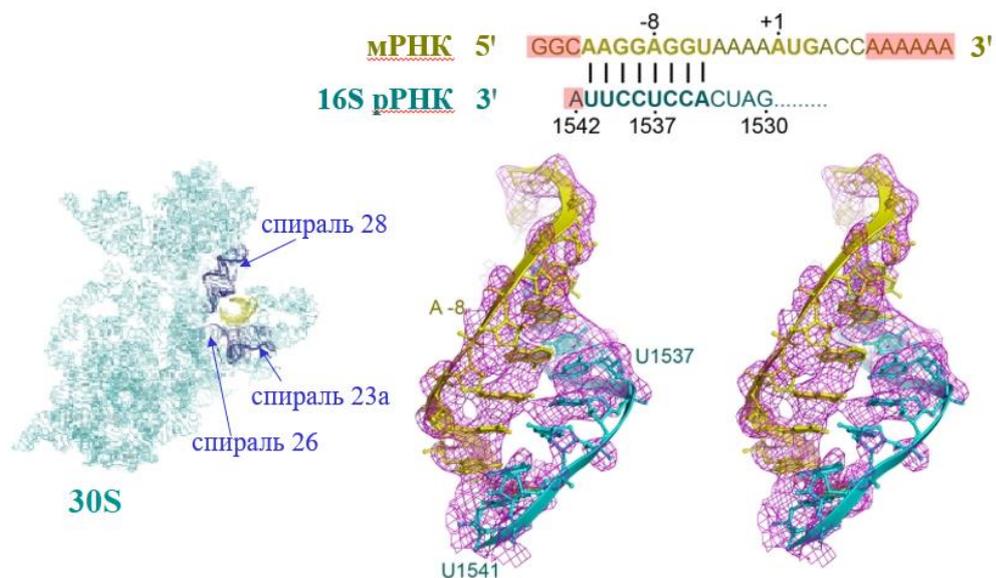


Рисунок 2. Взаимодействие последовательности Шайна-Дальгарно с рибосомой. Слева показано расположение спирали SD и 3'-концевого участка 16S рРНК относительно структуры 30S субчастицы. Спирали 23а, 26 и 28, окружающие дуплекс 16S рРНК и SD, представлены тёмно-синими цепочками, спираль 16S рРНК и SD выделена жёлтым цветом. Справа вверху приведены нуклеотидные последовательности и образование пар оснований между участками мРНК и 16S рРНК в инициаторном комплексе. Последовательности SD и анти-SD, старт-кодон отмечены жирным шрифтом, нуклеотиды с выделенным фоном разупорядочены в разрешённой структуре, полученной с помощью PCA. Стереорезультат, приведённый справа внизу, карты плотности при взаимодействии спирали SD с анти-SD 16S рРНК. Адаптировано на основе источника [20].

За 5'-UTR в мРНК следует открытая рамка считывания (ORF), включающая стартовый кодон, белок-кодирующую последовательность (CDS) и терминаторный кодон UAA, UAG или

UGA [6]. За стоп-кодоном располагается 3'-нетранслируемая область (3'-UTR), или в случае полицистронных мРНК может находиться следующая кодирующая область (Рис. 3) [21].

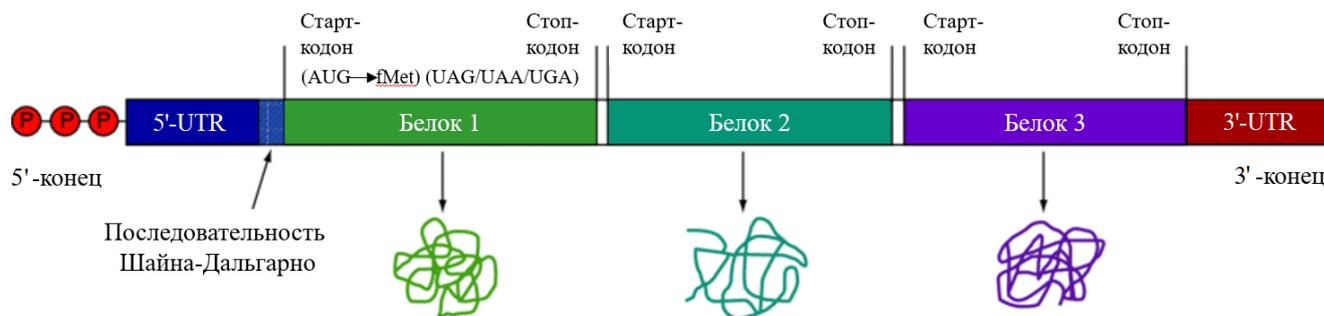


Рисунок 3. Схема бактериальной полицистронной мРНК. Адаптировано на основе работы [3].

У бактерий одна мРНК может кодировать несколько белков, в таком случае её называют полицистронной. При этом каждая белок-кодирующая последовательность окружена своими инициаторным и терминаторным кодонами. Такая организация на одной мРНК позволяет одновременно транслировать белки, часто функционально связанные и называемые также «продуктами генов» (Рис. 3) [3].

## 1.2. Бактериальная трансляция: краткий обзор

Биосинтез белка осуществляется большим макромолекулярным рибонуклеопротеидным комплексом, рибосомой, в ходе процесса, называемого трансляцией. Это очень сложный многостадийный энергозатратный процесс с очень большим количеством участников, пространственных взаимодействий и конформационных перестроек, до сих пор до конца не изученных и привлекающих особое внимание исследователей. Далее будут представлены лишь ключевые моменты столь важного процесса в упрощённом виде.

Во время трансляции происходит своего рода перевод информации, записанной на языке нуклеиновых кислот в виде цепочки из четырёх вариантов нуклеотидов, на язык белковых последовательностей из 20 вариантов аминокислот в соответствии с генетическим кодом (Рис. 4) [22]. Выделяют несколько его свойств: триплетность, однозначность, вырожденность и непрерывность. Первое заключается в соответствии одной аминокислоте последовательности из трёх нуклеотидов, называемой триплетом или кодоном, например, метионину соответствует триплет AUG в последовательности РНК, в свою очередь в последовательности ДНК это будет ATG. Под вторым свойством понимается однозначный перевод последовательности кодона в аминокислоту, например, кодон GGC соответствует только глицину и никакой другой аминокислоте. Вырожденность генетического кода заключается в возможности закодировать одну аминокислоту несколькими вариантами триплетов, так фенилаланин может быть представлен триплетами UUU или UUC. Последнее свойство сводится к тому, что в кодирующей

области кодоны следуют строго друг за другом без перекрывания и вставок [6, 22].

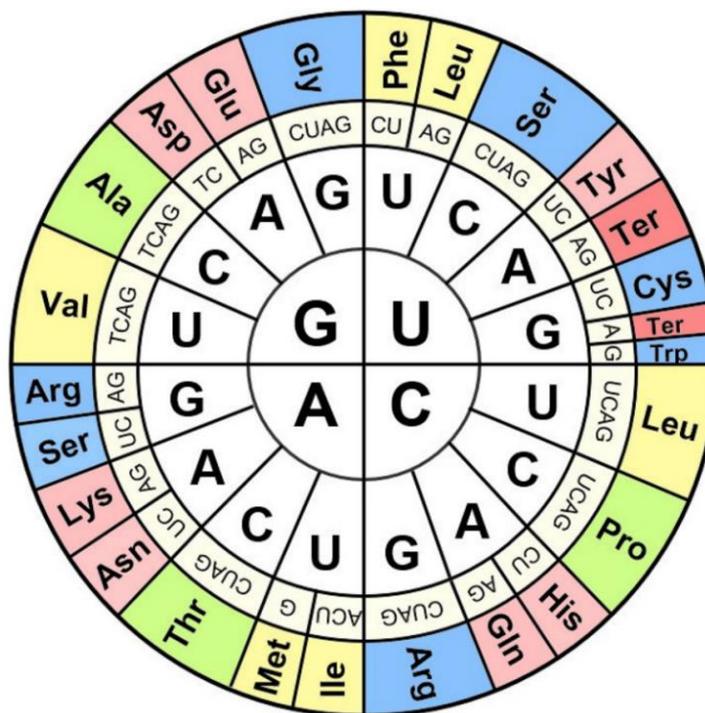


Рисунок 4. Классическое изображение генетического кода в виде колеса соответствия всех вариантов кодонов РНК (для ДНК вместо U будет T) аминокислотам или стоп-кодонам. Первая позиция кодона находится в центре, вторая позиция – в середине, а третья – вблизи периферии колеса рядом с трёхбуквенным обозначением аминокислот, расположенным во внешнем кольце колеса. Ter, терминационный кодон полипептидной цепи, стоп-кодон [22].

Бактериальные рибосомы массой 2,4 МДа имеют относительную скорость седиментации 70S и состоят из двух субчастиц: большой (50S), ответственной за синтез белка, и малой (30S) с основной функцией – декодирование информации, записанной в мРНК (Рис. 5). Субчастицы формируются из сложной сети рРНК (16S в составе малой субчастицы, 5S и 23S в составе большой) и около 50 различных белков (у *E. coli* 21 белок (S1-S21) входит в состав малой субчастицы и 33 (L1-L36) в состав большой) [23-25].

Сторону 30S субчастицы, обращённую к 50S, принято называть передней, фронтальной или внутренней, в то время как сторону, контактирующую с растворителем и обращённую к цитозолу, задней, спинной или внешней.

У рибосомы выделяют следующие сайты связывания тРНК: А, Р и Е, по которым она перемещается последовательно в процессе синтеза белковой молекулы. В А-сайте происходит связывание так называемой аминоацил-тРНК (aa-тРНК), а именно тРНК с присоединённой на 3'-ССА-конце аминокислотой, согласно генетическому коду соответствующей текущему триплету на мРНК, в свою очередь комплементарному антикодону на самой тРНК. При декодировании

мРНК в декодирующем центре (ДЦ) рибосомы аминокислота в составе aa-тРНК добавляется к растущей пептидной цепи (Рис. 6). В ходе каталитической реакции, осуществляемой рибосомой в ПТЦ, пептидил-трансферазном центре, происходит удлинение пептида путём его переноса с тРНК, находящейся в Р-сайте, на эту аминокислоту, в итоге он оказывается присоединённым к тРНК, расположенной в А-сайте (А-тРНК). При транслокации связанная с полипептидом тРНК – пептидил-тРНК (Р-тРНК) – перемещается из А- в Р-сайт, в то время как деацилированная тРНК без аминокислот переходит из Р- в Е-сайт (Е-тРНК) и в дальнейшем покидает рибосому [26-28].

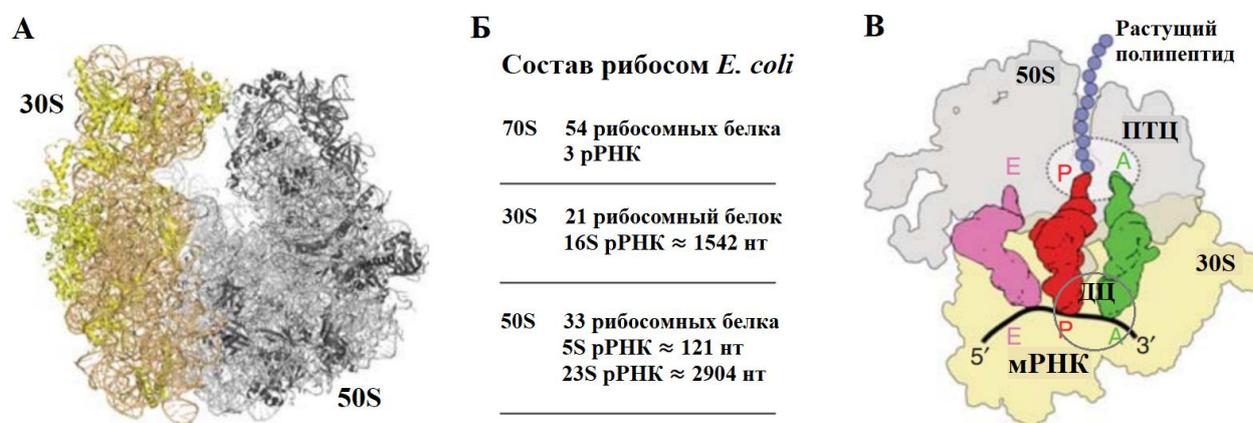


Рисунок 5. Структура и функциональные центры бактериальной рибосомы. А. Обзор 70S рибосомы *Escherichia coli* с 30S и 50S субчастицами, показанными жёлтым и серым цветами соответственно. Б. Таблица сборки компонентов 70S рибосомы и её субчастиц 30S и 50S. В. Схематичное представление 70S рибосомы, связанной в соответствующих сайтах А, Р и Е с тремя тРНК, показанными разными цветами: зелёным – А-тРНК, красным – Р-тРНК, розовым – Е-тРНК, наряду с 30S субчастицей – жёлтым, 50S субчастицей – серым, растущей полипептидной цепью – фиолетовым, и мРНК – чёрным. Пептидил-трансферазный центр (ПТЦ) на большой 50S субчастице выделен пунктирной сферой. Декодирующий центр (ДЦ) отмечен серой сферой. Адаптировано из источника [25].

Структура тРНК является консервативной не только среди бактерий, но и среди всего разнообразия живых организмов [26-36]. В её составе можно выделить несколько функционально значимых элементов: 5'- и 3'-концы вместе образуют акцепторный стебель, несущий сайты взаимодействия с аминоацил-тРНК синтетазами, с большой субчастицей рибосомы, с фактором элонгации EF-Tu (в случае бактерий), причём на 3'-конце располагается последовательность ССА, с которой связывается конкретная аминокислота, Т- и D-петли и соответствующие им стебли, содержащие сайты взаимодействия с большой субчастицей рибосомы, переменная V-петля, антикодоновая петля, комплементарная определённому кодону на мРНК, и соответствующий ей стебель, содержащие сайты взаимодействия с аминоацил-тРНК

синтетазами, А-, Р-, Е-сайтами рибосомы. Подробный анализ позволил выделить два типа тРНК: эволюционные тРНК I (большинство тРНК) и II (как правило, тРНК<sup>Ser</sup>, тРНК<sup>Leu</sup>, тРНК<sup>Tyr</sup>) типа, отличающиеся длиной вариательной петли (V-петли) (Рис. 7) [31, 33].

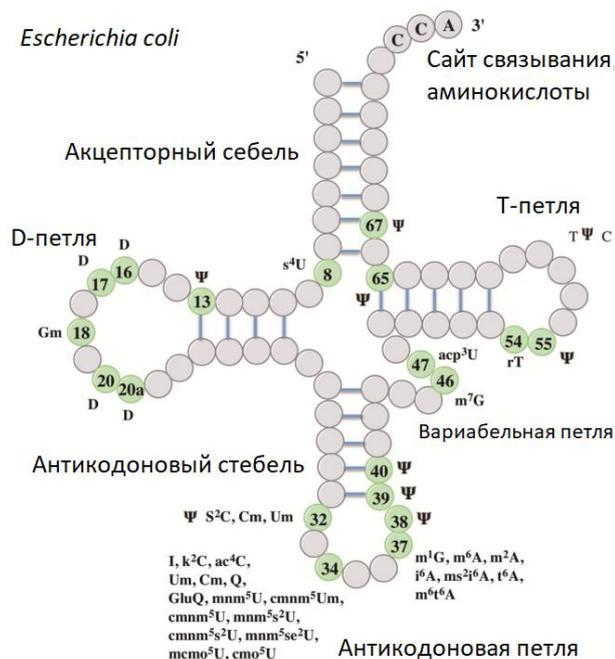


Рисунок 6. Известные модификации и их позиции в тРНК *E. coli*. Аббревиатуры конкретных групп включают m-, c-, n-, o-, t-, i-, k-, r- и s- для обозначения метильной, углеродной, амино, окси, треониновой, изопентильной, лизиновой, рибозной и серной (тио) групп соответственно. Другие аббревиатуры включают asp: 3-амино-3-карбоксыпропил, D: дигидроуридин, I: инозин, Ψ: псевдоуридин, Q: квеуозин (оригинал “queuosine”). Остальные заглавные буквы обозначают соответствующие основания: урацил, цитозин, гуанин и аденин. Индекс и показатель степени указывают номер и позицию замены, например, 6-диметиладенозин написан как m<sup>6</sup><sub>2</sub>A. Адаптировано на основе источника [28].

В процессе трансляции можно выделить следующие четыре стадии: инициацию, элонгацию, терминацию и возврат рибосомы для нового цикла, рециклирование (ресайклинг), каждая из которых соответственно регулируется своим набором различных факторов (Рис. 8) [6, 37-43]. В то же время рибосома является мишенью для действия большого разнообразия антибиотиков, относящихся к разным классам и действующих на разных этапах её функционального цикла [38].

На первой стадии малая 30S субчастица рибосомы связывается с сайтом посадки рибосомы (RBS) на мРНК с образованием дуплекса между, как правило, последовательностью Шайна-Дальгарно (SD) канонической мРНК и анти-SD на 3'-конце 16S рРНК. Далее происходит ассоциация инициаторных факторов IF1 и IF3 с 30S субчастицей, в комплексе с которыми

осуществляется присоединение ещё одного фактора IF2, GTP и fMet-тРНК<sup>Met</sup>. Затем IF1 и IF3 уходят, на следующем этапе, при связывании большой 50S субчастицы, GTP гидролизуется до GDP и свободной фосфатной группы, что приводит к диссоциации комплекса IF2 с рибосомой и переходу к следующей стадии – элонгации (Рис. 8) [6, 42].

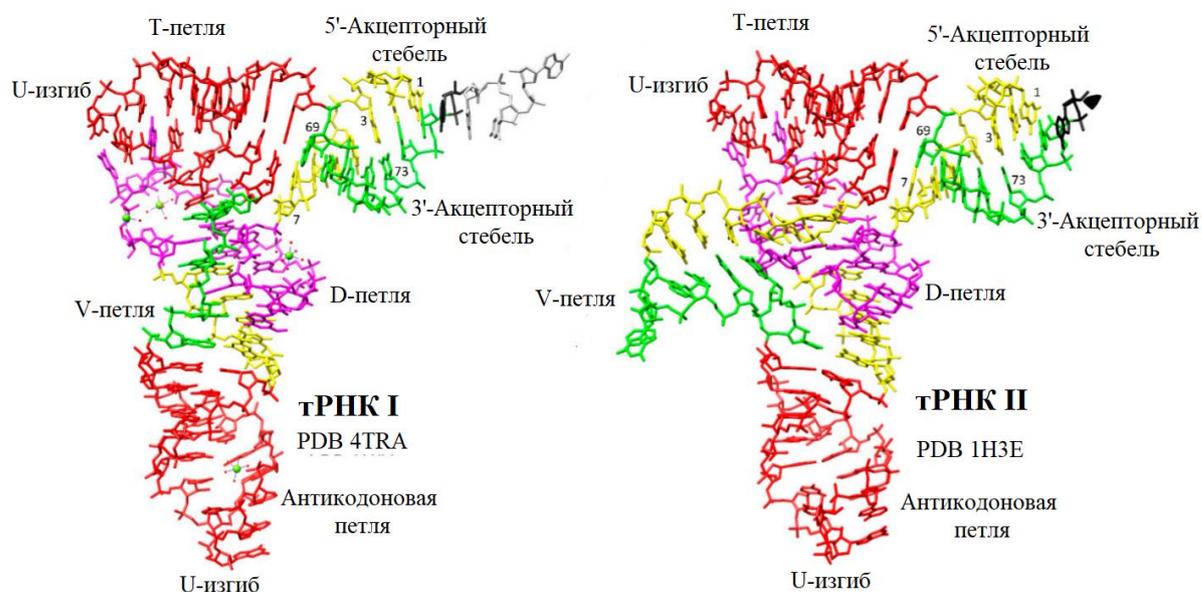


Рисунок 7. Эволюционные тРНК I (большинство тРНК) и II (тРНК<sup>Ser</sup>, тРНК<sup>Leu</sup>, тРНК<sup>Tyr</sup>) типа, отличающиеся длиной вариабельной петли (V-петли). Представленные тРНК I и II типа принадлежат дрожжам вида *Saccharomyces cerevisiae* и бактериям вида *Thermus thermophilus* соответственно. Гомологичные участки показаны одинаковым цветом. Антикодоновая петля (антипараллельна и комплементарна определённому кодону на мРНК, содержит сайты взаимодействия с аминоацил-тРНК синтетазами, А-, Р-, Е-сайтами рибосомы) и Т-петля, и соответствующие им стебли отмечены красным цветом. 5'-конец акцепторного стебля и гомологичные участки выделены жёлтым, а его 3'-конец (с ССА на конце, где присоединена аминокислота) и гомологичные участки – зелёным цветом, акцепторный стебель включает сайты взаимодействия с аминоацил-тРНК синтетазами, с большой субчастицей рибосомы, с фактором элонгации IF-Tu. V-петля является вариабельной и отличается по размеру у тРНК I и II типов. D-петля в виде микроспирали продемонстрирована розовым цветом. Т- и D-петли содержат сайты взаимодействия с большой субчастицей рибосомы. Ионы магния показаны в виде зелёных сфер. 4TRA, 1HZE – идентификаторы обеих структур из базы данных Protein Data Base (PDB). Адаптировано на основе источников [31, 33].

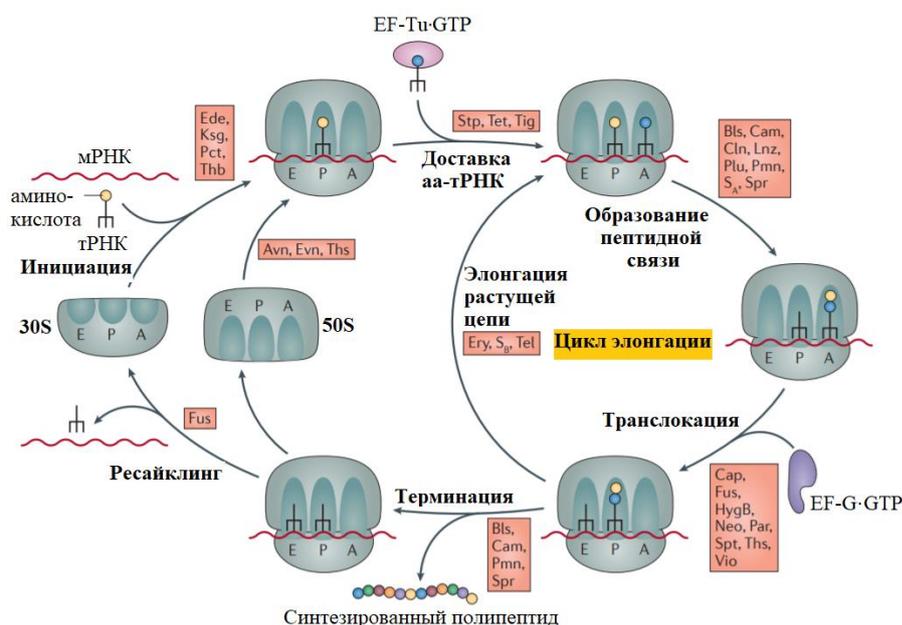


Рисунок 8. Стадии биосинтеза белка с указанием сайтов действия антибиотиков. Инициация белкового синтеза включает формирование 70S рибосомы (состоящей из 30S и 50S субчастиц) с инициаторной тРНК и стартовым кодоном на мРНК, располагающимися в Р-сайте. Этот процесс ингибируется следующими антибиотиками: эдеином (Ede), касугамицином (Ksg), пактамицином (Pct) и терморубином (Thb) на 30S субчастице и с помощью ортосомицинов: авиламицина (Avn) и эвернимидина (Evn), наряду с тиострептоном (ThS) на 50S субчастице. В цикле элонгации аминокислотированная тРНК (aa-тРНК) доставляется в А-сайт рибосомы элонгационным фактором Tu (EF-Tu), что блокируется стрептомицином (Str), тетрациклинами (Tet) и глицилциклинами (тигедиклином (Tig)). Формирование пептидной связи между тРНК, находящимися в А- и Р-сайтах, подавляется бластицидином С (Bls), хлорамфениколом (Cam), линкозамидами (клиндамицином (Cln)), оксазолидинонами (линезолидом (Lnz)), плевромутилинами (Plu), пуромицином (Pmn), стрептограмином А (SA) и спарсомицином (Spr). Транслокация тРНК катализируется фактором EF-G, этот процесс могут тормозить туберактиномицины капреомицин (Cap) и виомицин (Vio), аминогликозиды гигромицин В (HygB), неомицин (Neo) и паромомицин (Par), наряду с фузидовой кислотой (Fus), спектиномицином (Spt) и тиострептоном (ThS). Элонгацию растущей пептидной цепи могут тормозить макролиды (эритромицин (Ery)), стрептограмин В (Sb) и кетолиды (телитромицин (Tel)). Финальные стадии терминация и ресайклинг приводят к освобождению полипептидной цепи и последующей диссоциации 70S рибосомы, сопровождаемой подготовкой компонентов к следующему раунду инициации. Терминации препятствуют такие пептидил-трансферазные ингибиторы, как бластицидин С (Bls), хлорамфеникол (Cam), пуромицин (Pmn) и спарсомицин (Spr), в то время как ресайклинг блокируется ингибиторами транслокации, в частности фузидовой кислотой (Fus). Адаптировано из источника [38].

По мере движения рибосомы по мРНК от 5'- к 3'-концу происходит удлинение полипептидной цепочки путём последовательного переноса растущего пептида на аминокислоты, присоединённых к тРНК, согласно генетическому коду (Рис. 4). На этапе элонгации с рибосомой последовательно взаимодействуют два фактора: EF-Tu, обеспечивающий доставку aa-тРНК на рибосому, и EF-G, перемещающий рибосому по мРНК. Оба белка в ходе своей работы гидролизуют GTP [6, 40-41]. Процесс элонгации заканчивается при встрече рибосомой стоп-кодона в цепи мРНК (Рис. 8).

Терминация трансляции осуществляется при помощи нескольких факторов. RF1 и RF2 узнают стоп-кодоны: причём оба фактора узнают UAA, однако UAG распознаётся фактором RF1, а UGA – фактором RF2 [42], и отделяют синтезированный белок от тРНК. Впоследствии, RF3 вытесняет RF1 или RF2 из комплекса с рибосомой, сопровождая этот процесс гидролизом GTP. На следующем этапе RRF вместе с EF-G способствуют диссоциации рибосомы на субчастицы и отсоединению от неё мРНК. Диссоциированные большая и малая субчастицы рибосомы готовы для нового цикла трансляции, для рециклирования (Рис. 8) [38].

### 1.3. Разнообразие 5'-UTR у бактерий

#### 1.3.1. Мотивы 5'-UTR генов рибосомных белков и 3'-конца 16S рРНК у бактерий

Последовательность Шайна-Дальгарно предшествует аутентичным стартовым кодонам прокариотических мРНК, отличая их от случайно встречающихся в РНК AUG триплетов. Она обычно длиной 4-5 нуклеотидов, расположена на расстоянии 5-8 нуклеотидов перед стартовым кодоном и образует молекулярное взаимодействие с комплементарной ей последовательностью 3'-конца 16S рРНК 30S субчастицы рибосомы. Сравнительный анализ с помощью выравнивания 3'-концов 16S рРНК из 6-ти геномов класса грамотрицательных бактерий *Bacteroidetes*: *Bacteroides thetaiotaomicron* VPI 5482, *Bacteroides vulgatus* ATCC 8482, *Bacteroides ovatus* V975, *Bacteroides fragilis* NCTC 9343, *Prevotella denticola* F0289, *Prevotella bryantii* B14, и 4-х отсеквенированных 3'-концов 16S рРНК из *Escherichia coli*, *Bacillus subtilis*, *Geobacillus stearothermophilus* и *Lactococcus lactis* демонстрирует высокую консервативность последовательностей 3'-концов 16S рРНК (Рис. 9) [44].

Похожий подход для оценки консервативности 3'-конца 16S рРНК был использован, но уже для большей выборки из 277 геномов прокариот (249 представителей *Eubacteria* и 28 *Archaeobacteria*), при этом брались последние 13 нуклеотидов с 3'-конца. Было обнаружено, что 98 из 277 аннотированных последовательностей не содержат на 3'-концах мотив CCUCC – последовательность анти-SD из *E. coli*, что, скорее всего, указывает на ошибочность полученных аннотаций [45]. Поэтому для проверки было построено множественное выравнивание более широких терминальных последовательностей 16S рРНК. В результате определили высоко

консервативный мотив для каждого вида, соответствующий 3'-концу 16S рРНК у *Escherichia coli* (Рис. 10). В частности, последовательность анти-SD полностью сохранялась у всех исследованных видов, за исключением трёх *Archaeobacteria*, у которых цитозин в последнем положении замещён на аденин [46].

```

B_thetaiota_VPI5482 :AAAACUGGUAAUUGGGGCuaagucguaacaagguagccguaccggaaggugcggcuggaacaccuccuuucu.
B_vulgatus_ATCC8482 :AAAACUGGUAGCUGGGGCUAAGUCGUAACAAGGUAGCCGUACCGAAGGUGCGGCUGGAACACCUCUUUCU.
B_ovatus_V975       :AAAACUGGUAAUUGGGGCUAAGUCGUAACAAGGUAGCCGUACCGAAGGUGCUUCUGGAACACCUCUUUCU.
B_fragilis_NCTC9343 :AAAACUGGUAGCUGGGGCUAAGUCGUAACAAGGUAGCCGUACCGAAGGUGCGGCUGGAACACCUCUUUCU.
Prev_denticola_F0289 :AAAACUGGUAAUUGGGGCuaagucguaacaagguagccguaccggaaggugcggcuggaacaccuccuuucu.
Prev_bryantii       :AAAACUGGUAAUUGGGGCuaagucguaacaagguagccguaccggaaggugcggcuggaacaccuccuuucu.
L_lactis_IL1403     :.....CCUCCUUUCU.
B_subtilis          :.....CCUCCUUUCU.
B_stearothermophilus :.....CCUCCUUUCUA
E_coli              :.....CCUCCUUA...

```

Рисунок 9. ClustalW выравнивание генетических последовательностей 16S рРНК, найденных в 6 полных геномах представителей класса *Bacteroidetes* с отсековенными 3'-концами 16S рРНК из *E. coli*, *B. subtilis*, *G. stearothermophilus* и *L. lactis* IL1403 [44].



Рисунок 10. Высоко консервативная последовательность 3'-конца 16S рРНК. Logo последовательностей было получено из множественного выравнивания 16S рРНК 277 видов. В каждой позиции содержится стек из нуклеотидов (A, U, G, C), высота которого указывает на консервативность последовательности в определённом положении, тогда как высота каждого символа в стеке показывает относительную частоту определённого нуклеотида в данном положении. Звёздочка – индикатор положения, соответствующего 3'-концу 16S рРНК *E. coli* [46].

Дальнейший анализ проводился для 5'-UTR генов рибосомных белков, взятых с позиции -30 до +5 (первое основание стартового кодона имеет позицию 0), из 6 упомянутых ранее геномов *Bacteroidetes* и геномных последовательностей, принадлежащих видам *E. coli*, *B. subtilis* и *L. lactis*. Рибосомные белки большинства бактерий по биоинформатическим оценкам хорошо экспрессируются, и помимо оптимизированного использования кодонов это достигается за счет наличия последовательности SD, высоко аффинной к 3'-концу 16S рРНК и расположенной на расстоянии, более близком к оптимальному, чем последовательности SD перед генами с умеренной экспрессией согласно предсказаниям [47]. Количество 5'-UTR рибосомных генов с такими SD у многих видов бактерий часто достигает 80% [47] в то время, как соответствующие

5'-UTR анализированных генов представителей класса *Bacteroidetes* содержали подобные последовательности SD значительно реже – в 18-28% случаев [44].

Для определения консенсуса оптимального сайта посадки рибосомы для всех описанных выше организмов были построены так называемые logo последовательностей для 5'-UTR рибосомных генов с помощью программы WebLogo 3.3 (Рис. 11). В результате logo последовательностей хорошо соответствуют 3'-концу 16S рРНК для *E. coli*, *B. subtilis* и *L. lactis* за счет присутствия SD (5'-GGAGG-3'). При этом можно отметить, что *E. coli*, использующая рибосомный белок bS1 для помощи в инициации трансляции [48], имеет менее жёсткие требования к последовательности SD по длине и частоте встречаемости, чем *B. subtilis* и *L. lactis*. Что касается представителей *Bacteroidetes*, то такого консенсуса, как у первых трёх, не наблюдается, а видно обогащение RBS аденозином и в меньшей степени тимидином [44].

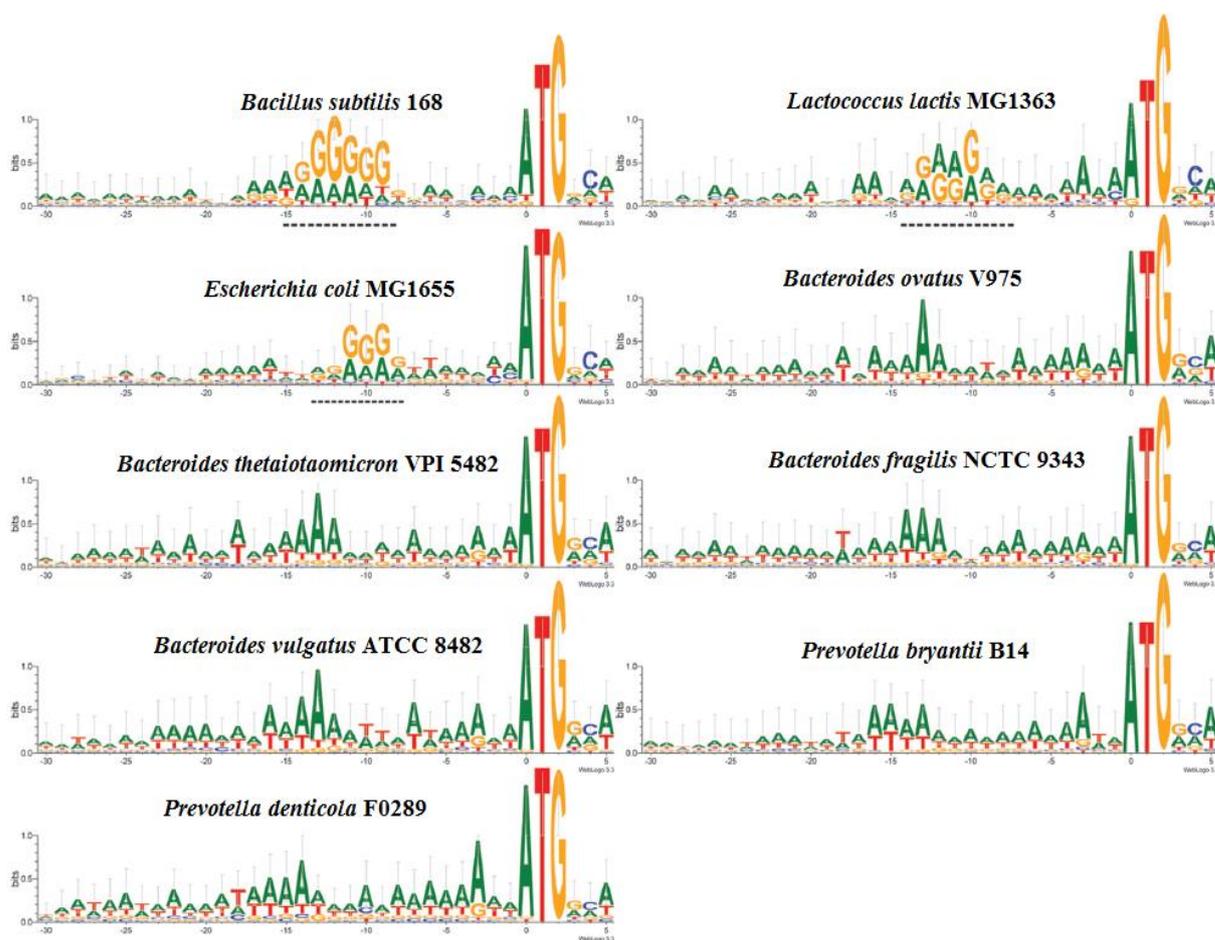


Рисунок 11. Logo последовательностей 5'-UTR генов рибосомных белков, охватывающих участок -30 до +5. Число использованных генов для построения logo для *E. coli*, *B. subtilis* и *L. lactis*, *B. thetaiotaomicron*, *B. vulgatus*, *B. ovatus*, *B. fragilis*, *P. denticola*, *P. bryantii* составляет 55, 60, 55, 53, 51, 45, 52, 53 и 48 соответственно. Первый нуклеотид стартового кодона находится в позиции 0, пунктирная линия выделяет последовательность SD (5'-GGAGG-3') [44].

### 1.3.2. Полногеномный анализ 5'-UTR у бактерий

Присутствие последовательности Шайна-Дальгарно по всему геному обуславливает обогащение аденином и гуанином logo последовательностей на расстоянии -13 до -5 нуклеотидов, предшествующих стартовому кодону гена. Такое обогащение можно найти в logo последовательностей большинства бактериальных филумов: *Proteobacteria*, *Firmicutes*, *Actinobacteria*, *Thermotogae*, *Chloroflexi* и *Aquificae* (Рис. 12). Заметным исключением среди *Proteobacteria* являются паразитические и симбиотические формы с низким процентным содержанием GC из семейства *Rickettsiales* и *Tenericutes* из рода *Mycoplasma* лишь с небольшим увеличением частоты встречаемости гуанина в участке последовательности SD. Logo последовательностей разных видов, содержащих SD и принадлежащих одному и тому же филуму, не единообразны. Такая неоднородность иногда обусловлена различиями в GC составе видов, однако резкие изменения в участке последовательности SD отсутствуют, что указывает на консервативность именно этого участка. *Moorella thermoacetica* и *Clostridium acetobutylicum* с геномным содержанием GC 55,8% и 30,9% соответственно имеют лучше выраженную SD на карте logo, чем *Bifidobacterium longum* и *Arthrobacter aureescens*. У представителей таких филумов, как *Spirochetes*, *Acidobacteria*, *Chlamydiae*, *Deinococcus-Thermus*, *Fibrobacteres* и некоторых *Cyanobacteria* SD менее очевидна, но все ещё обнаруживается при построении logo [49].

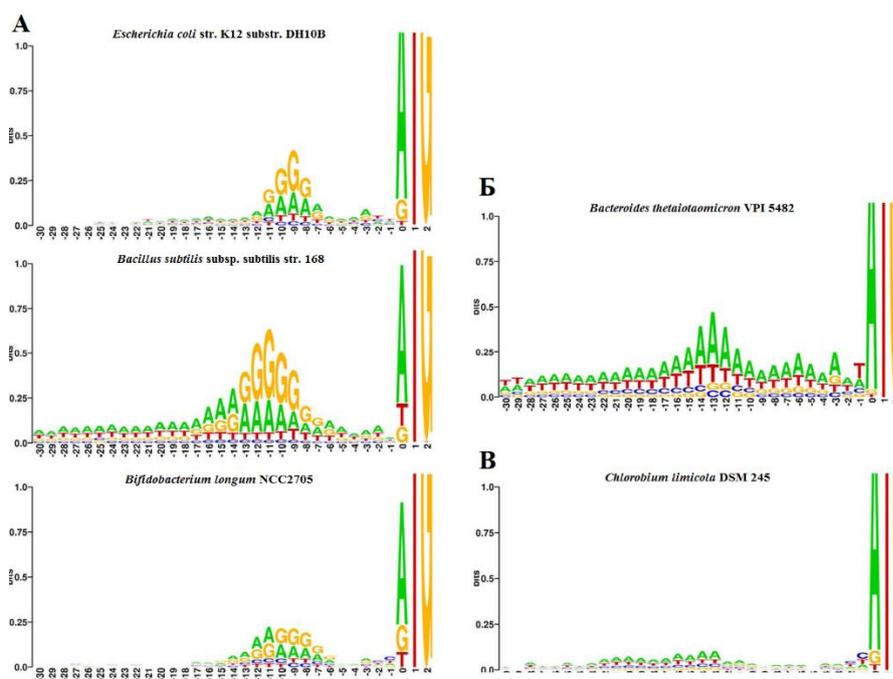


Рисунок 12. Logo анализ последовательностей, предшествующих стартовому кодону. А. Logo последовательностей, показывающий присутствие SD у хорошо известных бактерий. Logo последовательностей *Bacteroidetes* (Б) и *Chlorobi* (В) демонстрирует отсутствие SD. Нулевая позиция соответствует первому нуклеотиду старт-кодона. Адаптировано из источника [49].



Как известно, у прокариотических организмов 5'-UTR часто содержит некоторую вариацию консенсуса SD (5'-GGAGGU-3'), необходимого для инициации трансляции, однако, основываясь на анализе 162 полных прокариотических геномов гены с 5'-UTR без такой последовательности встречаются не реже (Рис. 14) [51]. Наибольшей консервативностью по последовательности SD обладают представители таксономических групп *Bacillales*, *Clostridia*, *Lactobacillales* в то время, как у большей части представленных *Proteobacteria* содержание последовательности SD можно оценить в 40-70%. Для остальных анализированных организмов высокой консервативности не наблюдается. Из полученных результатов сразу вытекает вопрос: как происходит узнавание правильного инициаторного кодона, если мРНК не содержит в качестве RBS последовательность SD?

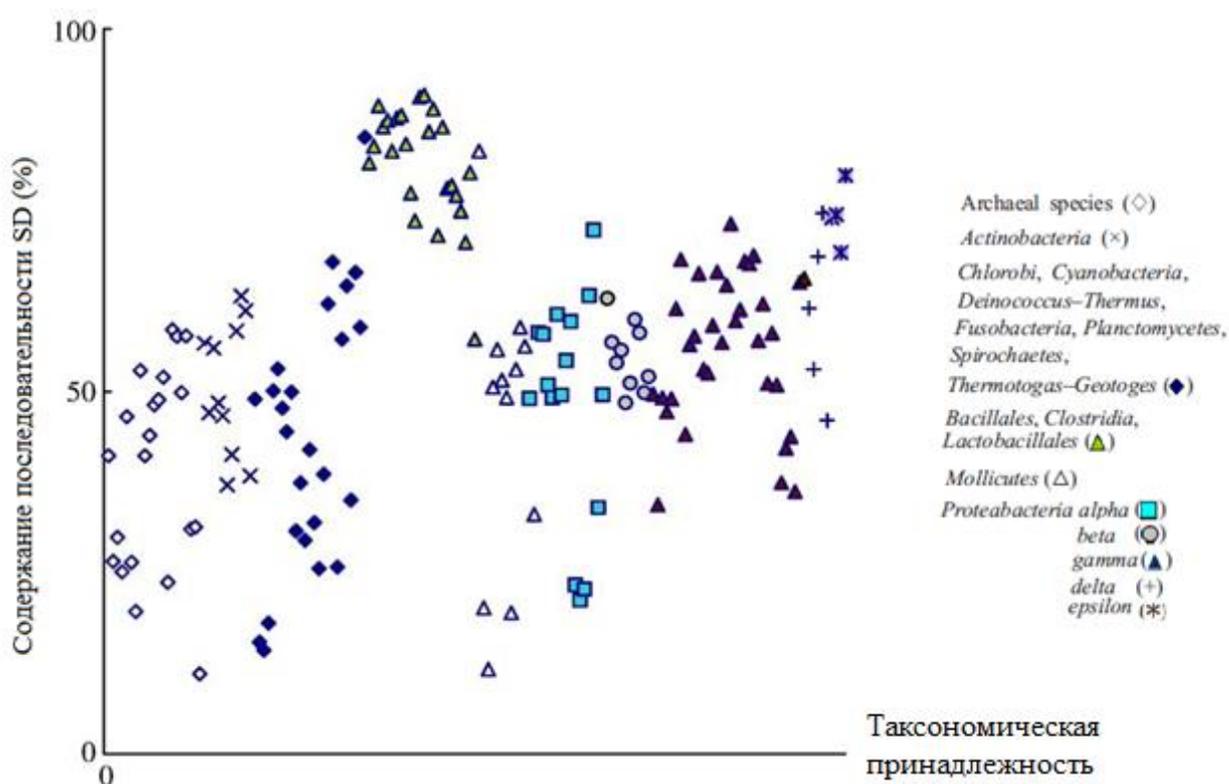


Рисунок 14. Содержание последовательности SD в 162 геномах прокариот. Адаптировано на основе источника [51].

Для ответа на поставленный вопрос был проведён поиск по всему геному в различных группах прокариот, чтобы определить элементы последовательности и/или мотивы вторичных структур РНК, которые могли бы участвовать в инициации трансляции в случае мРНК без SD. В результате биоинформатического и экспериментального анализа было обнаружено, что такие мРНК, как правило, менее структурированы в участке инициации трансляции и имеют минимальную укладку на стартовом кодоне (Рис. 15). Также было показано, что AUG кодон,

обозначающий начало для синтеза белка, находится в одноцепочечных участках последовательности, тогда как внутренние AUG кодоны в гене обычно находятся в структурированных участках мРНК [52]. Таким образом, возможно, что локальное отсутствие вторичной структуры необходимо и достаточно для инициации трансляции в отсутствие RBS, содержащего последовательность SD для анализируемых организмов.

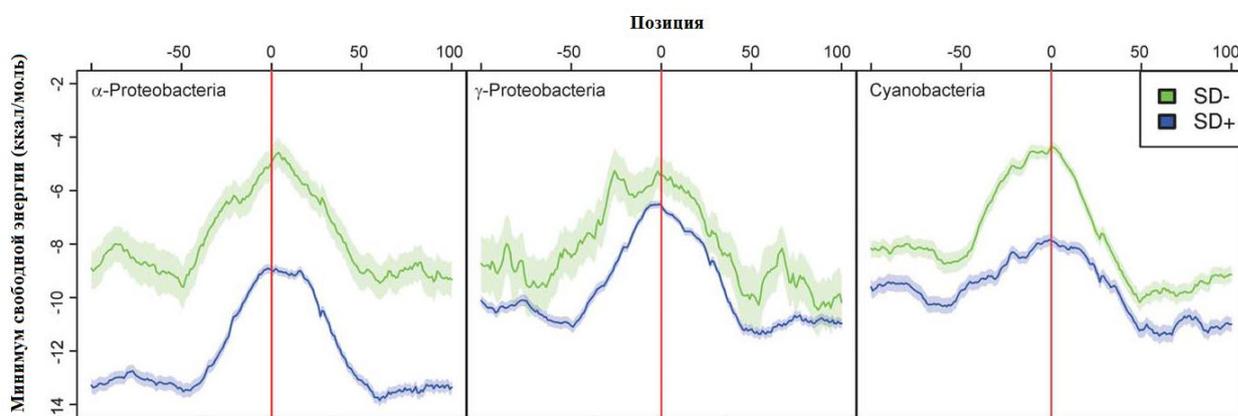


Рисунок 15. Графики предсказанной стабильности вторичной структуры РНК вблизи стартового кодона у  $\alpha$  – *Proteobacteria*,  $\gamma$  – *Proteobacteria* и *Cyanobacteria*. Нулевая позиция соответствует первому нуклеотиду стартового кодона. Гены без SD показаны зелёными линиями, с SD – синими. Линии идут согласно среднему значению минимума свободной энергии для 5000 генов, тень вокруг неё определяет стандартное отклонение от среднего. Минимум свободной энергии (MFE) определялся с использованием «скользящего окна» длиной в 50 нт. Разница в MFE между зелёными и синими линиями до и после инициаторного кодона в трёх бактериальных группах в основном из-за различий в АТ-содержании между геномами в пределах индивидуальных групп организмов и корреляции между обогащением АТ нуклеотидами и преобладанием независимой от последовательности SD трансляцией. Адаптировано на основе источника [52].

Известно, что существует много прокариотических генов без 5'-UTR, а соответственно и без SD, так называемых «генов без лидера». Однако *in silico* анализ макроэволюции демонстрирует уменьшение числа таких генов у бактерий в ходе эволюции (Рис. 16). При этом были рассмотрены участки инициации трансляции всех генов 953 бактериальных и 72 архейных геномов, распределённых по группам, согласно их расстояниям до корня (между бактериями и археями) на филогенетическом древе 16S рРНК. Была рассчитана средняя доля безлидерных генов в каждой группе: сначала она быстро падает, а затем колеблется на низком уровне [53].

Среди всего разнообразия прокариотических мРНК выделяют следующие виды: канонические мРНК [54], содержащие сайт посадки рибосомы (RBS) с последовательностью SD в 5'-UTR, в то же время известные как “SD-led”, и неканонические мРНК: не имеющие в своём

составе последовательность SD в 5'-UTR, также известные как “non-SD-led”, и безлидерные мРНК (так называемые “leaderless”) [6, 37].

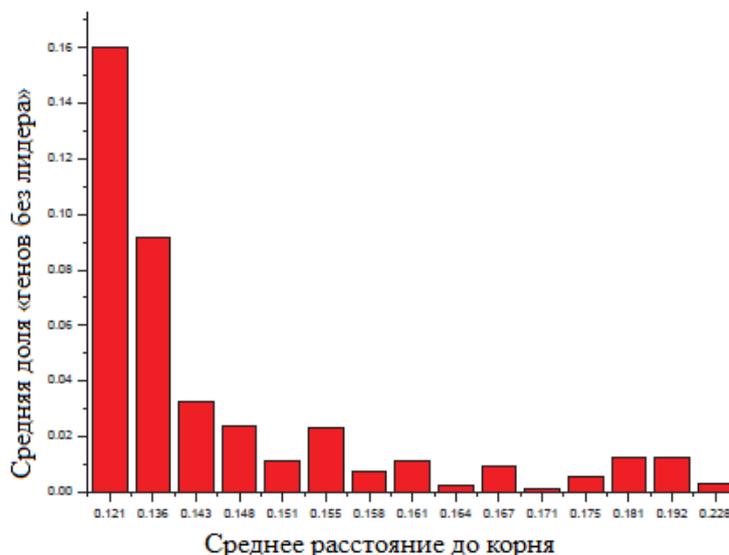


Рисунок 16. Тренд в эволюции «генов без лидера». 903 бактериальных генома распределены по 15 группам согласно их расстояниям до корня (между бактериями и археями) на филогенетическом дереве 16S рРНК. Рассчитана средняя доля «генов без лидера» в каждой группе. Адаптировано из источника [53].

#### 1.4. Бактериальная трансляция: детальный обзор инициации

##### 1.4.1. Инициация трансляции канонических мРНК

Бактериальные транскрипция и трансляция – тесно сопряжённые клеточные процессы, так как рибосома приступает к инициации трансляции с мРНК, ещё синтезируемой во время транскрипции. Инициация трансляции является скоростью-лимитирующей и наиболее строго регулируемой стадией из четырёх в биосинтезе белка [6, 55]. Скорость сборки рибосом на мРНК составляет порядка секунд, хотя и специфична для каждой мРНК. Далее рибосомы транслируют мРНК со скоростью примерно 12 аминокислот в секунду [6].

На начальном этапе инициации биосинтеза белка 30S и 50S субчастицы рибосомы диссоциированы [31] (Рис. 17). Малая субчастица ответственна за обеспечение правильного взаимодействия мРНК с тРНК, за сохранение соответствующей рамки считывания, а также за проверку, мониторинг и подтверждение распознавания кодона и соответствующего ему антикодона. Большая субчастица катализирует образование пептидной связи между уже синтезированной полипептидной цепью и вновь пришедшей аминокислотой, организует продвижение целого субкомплекса, сформированного из мРНК и тРНК, с шагом в длину одного кодона, так называемую транслокацию, с помощью специального элонгационного фактора G

(EF-G) и обеспечивает предварительную укладку и прохождение нативной цепи через рибосомный туннель. Две субчастицы, 30S и 50S, имеющие разные роли в работе 70S рибосомы, как становится сразу ясно, должны быстро передавать сигналы друг другу о текущем состоянии соответствующих функциональных центров [21], поэтому так важно, чтобы процесс их ассоциации проходил правильно.

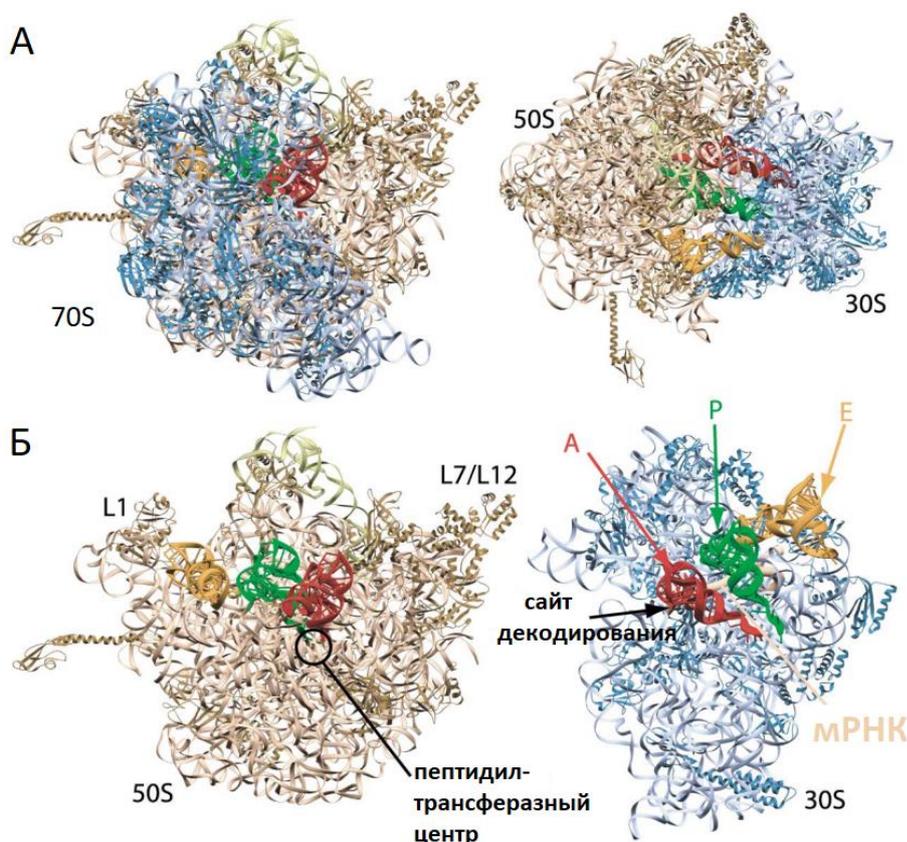


Рисунок 17. Кристаллические структуры рибосомы. А. Два вида на 70S рибосому в комплексе с мРНК и тРНК [56], при этом слева вид «сверху», а справа вид со стороны 30S. Б. Покомпонентное изображение 50S (слева) и 30S (справа) субчастиц в 70S рибосоме с отмеченным расположением А-, Р- и Е-сайтов связывания тРНК. Адаптировано на основе источника [42].

Инициация прокариотической трансляции требует формирования 30S пре-инициаторного комплекса (30S PIC), который у бактерий включает малую субчастицу рибосомы (SSU), три фактора инициации (IFs): IF1, IF2 и IF3, а также формируемую инициаторную fMet-тРНК<sup>fMet</sup> и каноническую мРНК, содержащую в 5'-UTR последовательность SD, со своей вторичной структурой (“SD-led” мРНК). Для таких канонических мРНК механизм инициации трансляции наиболее изучен [54]. При этом факторы IF1 и IF3 обеспечивают точность трансляции, тогда как фактор IF2 доставляет fMet-тРНК<sup>fMet</sup> [54]. Благодаря формилрованию метионил-тРНК, опосредованному трансформилазой, происходит отличие инициаторной fMet-тРНК<sup>fMet</sup> от элонгационной Met-тРНК<sup>Met</sup>, обеспечивающей встраивание метионина, встречающегося внутри

кодирующей области, в синтезируемый полипептид (Рис. 18) [54]. Главной целью формирования 30S пре-инициаторного комплекса является правильное позиционирование инициаторной fMet-тРНК<sup>fMet</sup> в пептидил/инициаторном положении (P/I), при котором она связана со стартовым кодоном на мРНК в Р-сайте малой субчастицы рибосомы [54].

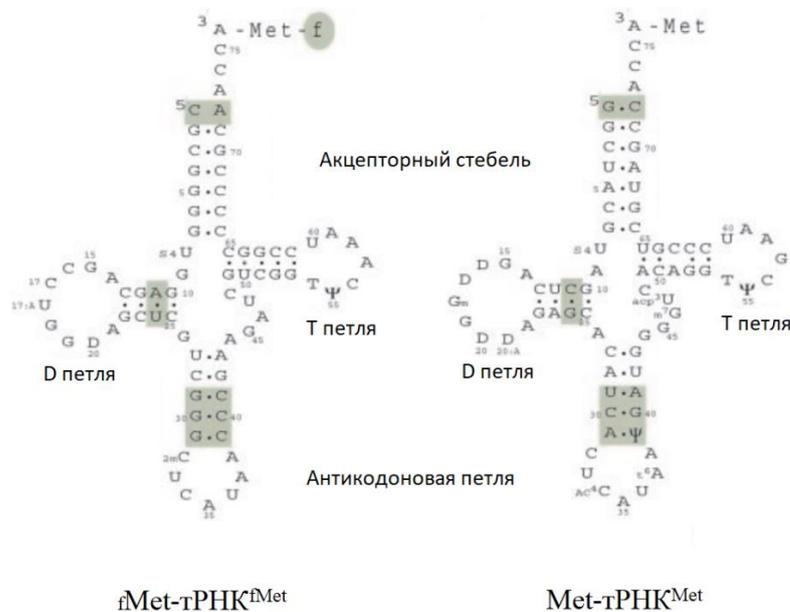


Рисунок 18. Инициаторная и элонгационная метионин-акцепторные тРНК. Представление тРНК, принимающих метионин, в виде «листьев клевера»: слева показана инициаторная тРНК, fMet-тРНК<sup>fMet</sup>, а справа элонгационная тРНК, Met-тРНК<sup>Met</sup>. Участки, важные для идентификации инициаторной тРНК, выделены цветными рамками. Адаптировано на основе источника [6].

Первая стадия биосинтеза белка начинается на уровне малой 30S субчастицы, которая при присоединении инициаторного фактора IF3 диссоциирует от большой 50S субчастицы из состава 70S рибосомы (Рис. 19) [55]. При этом фактор IF3 связывается с Е-сайтом своим N-концевым доменом рядом с белками uS7 и uS11 и С-концевым доменом на поверхности платформы 30S субчастицы, совпадая с расположением спирали H69 23S рРНК 50S субчастицы, образуя с 30S комплекс [43, 54]. Таким образом предотвращается преждевременное взаимодействие с большой 50S субчастицей до ассоциации с IF1, IF2, мРНК и инициаторной тРНК с 30S субчастицей [57]. Более того, IF3 влияет на скорости ассоциации тРНК с Р-сайтом и диссоциации этого комплекса, проводит мониторинг трёх уникальных пар оснований G:C в fMet-тРНК<sup>fMet</sup> (Рис. 18), его присутствие необходимо для обеспечения точности кодон-антикодонового взаимодействия в Р-сайте 30S субчастицы. Особенно важно, что IF3 имеет решающее значение для определения участка инициации трансляции (TIR) на мРНК. В отсутствие данного фактора рибосома в значительной степени неспособна различать непредпочтительные TIR или неправильные старт-кодоны [58].

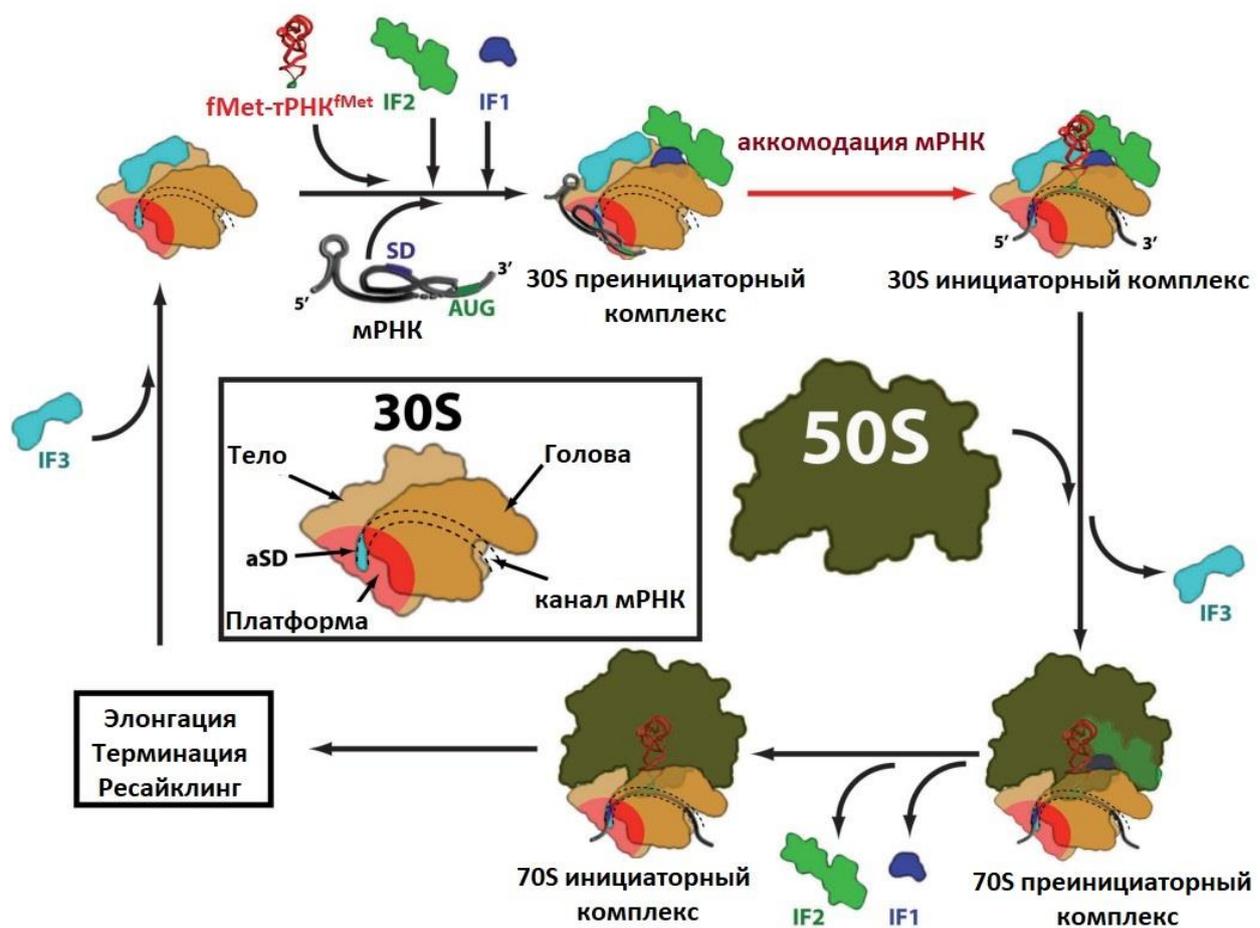


Рисунок 19. Схематическое изображение процесса инициации трансляции. Формирование 30S и 70S инициаторных комплексов, содержащих малую и большую субчастицы рибосомы (30S и 50S субчастицы показаны оранжевым и темно-зелёным цветами соответственно), инициаторную  $fMet-tRNA^{fMet}$ , мРНК и инициаторные факторы IF1 (в синем цвете), IF2 (в зелёном цвете), IF3 (в голубом цвете). Представлен вид 30S субчастицы рибосомы и рибосомы сверху. Платформа 30S субчастицы показана красным с последовательностью анти-Шайна-Дальгарно (aSD) голубого цвета. Структурированная мРНК связывается с 30S субчастицей в два отдельных этапа: расположение (докинг) мРНК на платформе 30S субчастицы приводит к формированию преинициаторного комплекса с последующей аккомодацией мРНК, способствующей правильному кодон-антикодону взаимодействию в Р-сайте [59]. Полученный 30S инициаторный комплекс задействует большую 50S субчастицу для формирования 70S инициаторного комплекса, из которого вытесняются факторы инициации, и синтез закодированного белка может приступать к последующим стадиям: элонгации, терминации и рециклирования рибосомы. Адаптировано из источников [55, 60].

Далее другие два инициаторных фактора IF2, связанный с GTP и являющийся ГТФазой, и IF1 связываются с 30S субчастицей, способствуя привлечению структурированной мРНК и инициаторной  $fMet-tRNA^{fMet}$  [55]. IF1 связывается в А-сайте 30S субчастицы, в «расщелине»,

образованной спиралью h44, 530 петлёй 16S рРНК и белком uS12, прикрывая, но не закрывая полностью канал в основании А-сайта, через который должна идти мРНК [61]. В то же время IF1 стабилизирует присоединение IF2 в Р-сайт 30S [58, 62] и ускоряет IF2-зависимое привлечение инициаторной тРНК с формилированным метионином – fMet-тРНК<sup>fMet</sup> [6]. Более того IF1 участвует в диссоциации субчастиц, предшествующей инициации, стимулируя формирование 30S комплекса, высвобождение IF2 из 70S рибосомы и блокирование А-сайта от преждевременного присоединения тРНК [61]. IF2 содержит три домена: N-концевой с двумя субдоменами, контактирующий с IF1 и uS12, GTP-связывающий с тремя субдоменами и С-концевой с двумя субдоменами [37]. В совокупности три фактора инициации кинетически контролируют процесс инициации и обеспечивают его правильность и точность [55, 63].

Следует отметить, что последовательность присоединения факторов инициации отличается в представленных схемах из разных источников. Фактор IF3 отдельно [54-55] или совместно с IF1 [6] или IF2 [64] связывается с 30S субчастицей на самой ранней стадии. Далее происходит присоединение факторов IF1 и IF2 в первом случае либо отдельно фактора IF2 или IF1 в последующих двух случаях соответственно. На втором этапе также образуются взаимодействия с инициаторной тРНК и мРНК [6, 54-55] или же они последовательно присоединяются на следующих двух этапах [64]. В результате образуется 30S пре-инициаторный комплекс, включающий малую 30S субчастицу рибосомы, структурированную мРНК, инициаторную тРНК с формилированным метионином – fMet-тРНК<sup>fMet</sup> – наряду с тремя инициаторными факторами IF1, IF2 и IF3 [6, 54-55, 64].

Далее происходит процесс аккомодации мРНК на 30S субчастице рибосомы, в результате которого формируется активный 30S инициаторный комплекс (30S IC) (Рис. 19). В данном комплексе первый кодон мРНК, обычно AUG, физически взаимодействует с антикодоном инициаторной тРНК в Р-сайте 30S субчастицы [55]. Для многих бактериальных мРНК выбор подходящего инициаторного кодона и трансляционной рамки считывания зависит во многом от формирования короткого дуплекса между последовательностью Шайна-Дальгарно, часто представленной GGAGG, расположенной в 5'-нетранслируемой области мРНК, и консервативной последовательностью анти-Шайна-Дальгарно, присутствующей на 3'-конце 16S рРНК (Рис. 2) [13, 55, 65-66].

Спираль SD плотно прилегает к полости, примыкающей ко входу в канал для размещения в нём мРНК в малой 30S субчастице рибосомы. При этом две стенки полости преимущественно состоят из спиралей 16S рРНК: h28 и h37 из головного домена, h23a и h26 из центрального домена, формирующего платформу. В то же время рибосомные белки uS7 и uS11, формирующие голову и платформу 30S соответственно, контактируют друг с другом, тем самым окружая вход для мРНК. На противоположной стороне полости uS2 соединяет два домена для формирования

дистальной стенки, расположенной близко к одному концу спирали SD, а белок bS18 находится на большем отдалении (Рис. 20) [65].

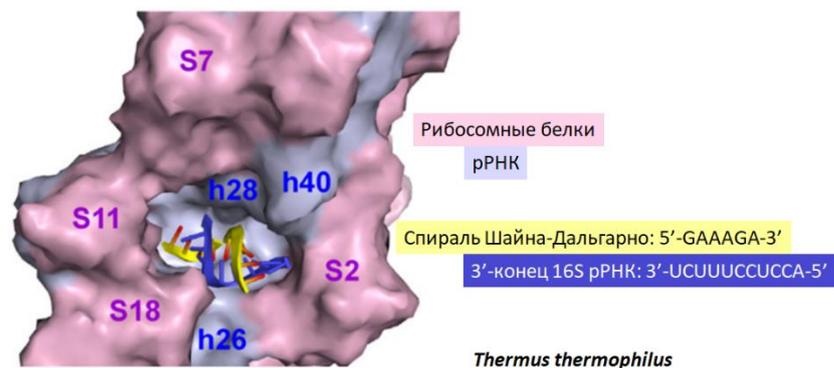


Рисунок 20. Расположение спирали SD на 30S рибосомной субчастице в виде стереоизображения молекулярной поверхности. рРНК и рибосомные белки окрашены в светло-сиреневый и бледно-розовый цвета соответственно. Олигонуклеотид (5'-GAAAGA-3') представлен жёлтым цветом, а 3'-концевые нуклеотиды 16S рРНК от A1534 – тёмно-синим. Адаптировано из источника [65].

Согласно результатам анализа бактериальных геномов, последовательность SD присутствует в 5'-UTR более, чем у 40% всех генов, в 22 из 30 геномов [47]. У *Bacillus subtilis* и *Thermotoga maritima* зарегистрировали самый высокий процент встречаемости SD: 89.4% и 90.1% соответственно. Самый низкий процент SD был характерен для *Rickettsia prowazekii*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Halobacterium sp.* штамм NRC-1, *Thermoplasma acidophilum*, *Sulfolobus solfataricus* и *Pseudomonas aeruginosa*, у каждого вида примерно по 20% [47]. Среди 162 полных аннотированных прокариотических геномов число генов с SD в 5'-UTR варьирует от ~ 12% до 90%, предполагая значительное число мРНК без SD в 5'-UTR или безлидерных мРНК, не имеющих вовсе 5'-UTR [51]. В целом быстрорастущие бактерии, грамотрицательные термофилы, спирохеты, метаногены и гипертермофильные археи имеют относительно высокую долю генов, содержащих SD в 5'-UTR (до 90%), при этом у облигатных внутриклеточных паразитов, поверхностных паразитов и цианобактерий эта доля уменьшается (всего до 10-20%) [37, 47, 51].

Спейсерный участок (известный в англоязычной литературе как “spacer/spacing region”) между последовательностью SD и инициаторным кодоном варьирует в пределах от 5 до 13 оснований, при этом оптимальная его длина является 7-10 оснований в мРНК *E. coli* [37]. Длина самой SD различна у разных генов *E. coli* и варьирует в пределах 4-8 нуклеотидов.

Взаимодействие SD с анти-SD приводит к расположению старт-кодона в Р-сайте 30S субчастицы, где он распознаётся fMet-тРНК<sup>fMet</sup> [37]. Важность дуплекса SD с анти-SD продемонстрирована экспериментами, показывающими, что трансляция мРНК может быть

замедлена мутациями в последовательности SD, а затем восстановлена компенсаторными мутациями в анти-SD [67-68]. Однако в целом сила SD лишь умеренно коррелирует с эффективностью трансляции [47, 69-71]. При экспрессии на умеренном уровне (когда задействованы 30-40% от общего количества рибосом) рибосомы с изменённой анти-SD всего в два раза менее активны, чем рибосомы в норме при трансляции набора клеточных мРНК [72]. Рибосомы без анти-SD на 3'-конце 16S рРНК правильно выбирают сайт начала трансляции в некоторых природных мРНК, предполагая, что последовательность SD не единственная детерминанта в инициации трансляции [70]. В совокупности эти данные свидетельствуют о том, что взаимодействие SD с анти-SD играет важную, но не определяющую роль в биосинтезе основной массы клеточных белков [37].

Малая рибосомная 30S субчастица связывается с одноцепочечными участками мРНК в участке инициации трансляции [73]. Образование дуплекса между последовательностями SD в 5'-UTR на мРНК и анти-SD на 3'-конце 16S рРНК происходит в «расщелине» между головой и платформой на внешней стороне 30S субчастицы, тогда как около 30 нуклеотидов мРНК заворачивают в «бороздку», опоясывающую шею 30S субчастицы, образованную спиралью h28 [20, 58, 74-75]. Причём последовательность SD на мРНК может связываться с рибосомой независимо от инициаторных факторов и fMet-тРНК<sup>fMet</sup> [6, 74].

Другие элементы последовательности мРНК, присутствующие в TIR, как было показано, влияют на образование инициаторного комплекса: старт-кодон, спейсерный участок между инициаторным кодоном и последовательностью SD, неслучайное распределение нуклеотидов, предшествующих SD и следующих за старт-кодоном в начале кодирующей области, и элементы вторичной структуры мРНК, присутствующие во всём TIR [55, 59-60, 76-77]. К примеру, у *E. coli* мРНК со слабой SD обычно содержат участок, богатый нуклеотидами AU/U, на 5'-конце перед SD, который выступает как мотив для распознавания рибосомным белком bS1 [14, 78]. Таким образом данный белок преодолевает потенциальную проблему закрепления мРНК на 30S субчастице.

Показана способность белка bS1, подвижного и высоко консервативного у прокариот, эффективно иницировать трансляцию независимо от наличия SD в 5'-UTR мРНК. Структурное понимание этого процесса затруднено ввиду того, что традиционный метод исследования структуры, такой как рентгеновская кристаллография, этого не позволяет сделать. Благодаря применению белкового кросс-сшивания и масс-спектрометрии высокого разрешения определён сайт связывания белка bS1 и подтверждена его динамика. В результате пришли к выводу о действии данного белка как захватывающей руки, доставляющей мРНК в рибосому, тем самым облегчая инициацию трансляции (Рис. 21) [15].

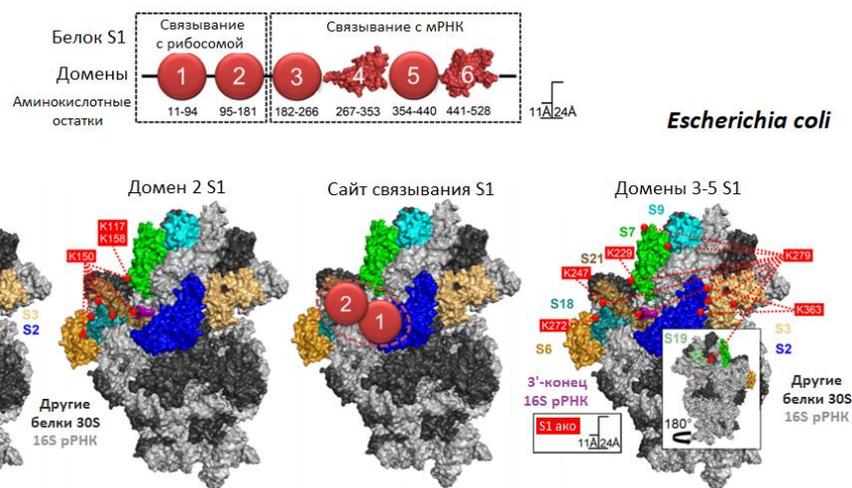


Рисунок 21. Сверху представлена доменная архитектура белка bS1, для доменов 4 и 6 показаны структуры, полученные с помощью ЯМР. Снизу продемонстрировано кросс-сшивание N-концевого участка связывания белка bS1 с 30S рибосомной субчастицей. Слева снизу следуют структуры с кросс-сшиванием доменов 1 и 2, включающие остатки, локализованные вблизи 3'-конца 16S рРНК. Остатки на 30S субчастице (в базе PDB структура 2AVY), которые перекрестно связаны с bS1, отмечены красными кружками. Треугольник на белке bS6 показывает приблизительное положение остатка, отсутствующего в кристаллической структуре. Остатки белка bS1, участвующие в этих перекрестных взаимодействиях, выделены красным цветом. Справа внизу представлен сайт связывания N-концевого участка белка bS1 с рибосомой, согласующийся с анализом кросс-сшиваний. Адаптировано из источника [15].

Структурированные элементы в сайте посадки рибосомы на мРНК играют важную роль в контроле инициации трансляции [55]. Несмотря на то, что мРНК почти не структурирована в пределах специального канала для мРНК на 30S субчастице, есть работы, свидетельствующие о способности платформы 30S распознавать и связывать структурированные мРНК [55, 59, 77]. Как было показано, такие мРНК связываются с 30S в двухступенчатом процессе [73, 77].

Первый этап, обычно называемый пре-инициаторным, является кратковременным и состоит в расположении (докинге/на английском языке “docking”) мРНК на платформе 30S субчастицы, причём сама платформа хорошо подходит для этого этапа (Рис. 19) [59]. Стадия связывания может включать последовательность Шайна-Дальгарно, когда она доступна для формирования спирали между SD и анти-SD [59], или может потребоваться неспаренная последовательность перед структурой в виде шпильки, изолирующей SD на мРНК [73, 77]. В такой момент, несмотря на одновременное присутствие мРНК и инициаторной тРНК на 30S субчастице, первое кодон-антикодонное взаимодействие ещё не сформировано. Полученный пре-инициаторный комплекс поэтому неактивный, и в нём легко могут происходить перемещения.

Далее следует более медленный этап, называемый аккомодацией, на котором мРНК адаптируется с помощью инициаторных факторов и fMet-тРНК<sup>fMet</sup> в предназначенном для неё канале и в результате которого формируется активный инициаторный комплекс на 30S (30S IC) (Рис. 19) [55].

При получении структуры крио-ЭМ 30S пре-инициаторного комплекса был учтён тот факт, что структурированные мРНК могут быть стабилизированы на 30S, если стадия аккомодации заблокирована [59]. Этот механизм, известный как захват (на английском языке “entrapment”), используется рибосомными белками uS15 и uS4 в *E. coli* для подавления трансляции своих собственных мРНК, содержащих в структуре оперона много кодирующих областей других рибосомных белков [79-80]. Структура 30S PIC, например, в присутствии репрессорного белка uS15, показывает связывание неадаптированной мРНК в свёрнутом состоянии на платформе малой субчастицы. Несколько рибосомных белков, располагающихся на платформе (uS2, uS7, uS11 и bS18) либо контактируют со скрученной мРНК, либо находятся близко к ней.

Систематический анализ структуры и последовательности выявил образование участков с положительными зарядами консервативных остатков этих белков на поверхности платформы вблизи захваченной мРНК [81]. Довольно неожиданно, что 5'-поли(А)- или поли(С)-богатые удлинения перед последовательностью SD также могут образовывать стабильную структуру со стеблем и петлёй на платформе 30S субчастицы [66]. Это предполагает наличие общего сайта докинга на платформе для структурированных мРНК во время пре-инициации с последующей адаптацией мРНК в специальном для неё канале. Время задержки между докингом и адаптацией отражает стабильность структур мРНК, которые в конечном счёте должны быть расплавлены, чтобы способствовать кодон-антикодонному взаимодействию. Таким образом, лиганды (белки, метаболиты, не кодирующие РНК и т.д.), стабилизирующие свёрнутое состояние мРНК, могут блокировать рибосому на стадии пре-инициации, мешая стартовому кодону достигнуть декодирующего центра внутри рибосомы [59]. Более того, платформа является широким и доступным открытым пространством, дающим возможность связаться регуляторным лигандам непосредственно на ней (например, белку-репрессору uS15 с *rpsO* мРНК), подчёркивая важность этого сайта связывания с рибосомой для продолжительности и регуляции трансляции [55].

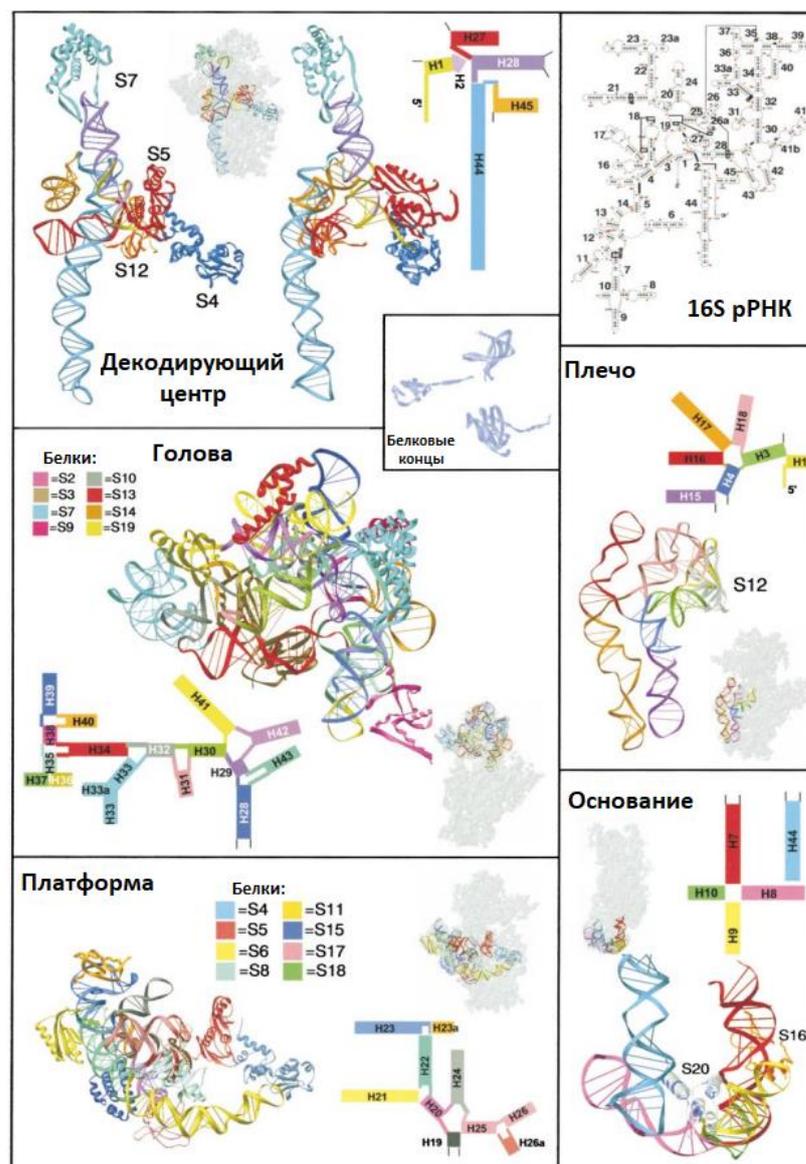
На платформе 30S мРНК разворачивается и занимает классическое расположение в своём канале, приводя к формированию активного комплекса инициации трансляции (Рис. 19). Процесс адаптации контролировался экспериментами с помощью сайт-направленного кросс-сшивания [82] и флуоресцентного резонансного переноса энергии (FRET) [73], демонстрируя, что одновременное присутствие факторов инициации и fMet-тРНК<sup>fMet</sup> способствует окончательной корректировке в ДЦ. Некоторые из рибосомных белков платформы могут облегчать расплетание

структуры мРНК, способствуя адаптации. В связи с этим рибосомные белки uS2, uS7, uS11, bS18 и bS21, а также bS1 располагаются вблизи места связывания с платформой. Поскольку известно, что последний белок обладает хеликазными свойствами, возможно, что его роль в инициации трансляции касается процесса адаптации мРНК [55].

Структуры, полученные с помощью PCA, 70S-мРНК-тРНК инициаторных комплексов показывают, что на стадии аккомодации около 30 нуклеотидов, охватывающих SD и AUG кодон, лежат в канале, опоясывающем шею 30S субчастицы, подтверждая результаты предыдущих экспериментов, биохимических и с перекрёстными сшиваниями [83-84]. После места взаимодействия SD с анти-SD мРНК проходит через короткий туннель, ограниченный С-концевой  $\alpha$ -спиралью рибосомного белка uS7, петлёй, соединяющей два его уникальных  $\beta$ -тяжа, и спиралью h23 и h28 16S рРНК. В декодирующем туннеле, в межсубъединичном пространстве, взаимодействия между мРНК и 30S устанавливаются почти полностью через 16S рРНК, за исключением нескольких контактов с белком uS12 в А-сайте (Рис. 22). Фосфатная группа нуклеотида 1401 на вершине спирали h44 располагается непосредственно на пути мРНК между кодонами в А- и Р-сайтах 30S субчастицы (А- и Р-кодонами), тем самым заставляя остов мРНК делать перегиб между двумя кодонами. Далее после А-кодона мРНК проходит через туннель большего размера, сформированный между головой и плечом малой субчастицы и включающий белки uS3, uS4 и uS5. Перечисленные рибосомные белки образуют кольцо, способное раскручивать структурированные участки мРНК во время элонгации [55, 74, 85].

Несмотря на отдалённое расположение, белок bS20 вносит значительный вклад в процесс декодирования: мутанты, лишённые его, демонстрируют неправильное прочтение мРНК, чувствительность к температуре, сниженную эффективность к ассоциации субчастиц и нарушенное взаимодействие кодона с антикодоном в Р-сайте, тем самым указывая на участие белка bS20 в правильной ориентации спирали h44 [86-87].

Представленная универсальная номенклатура рибосомных белков позволяет их классифицировать по принадлежности к трём доменам филогенетического древа жизни (бактериям, археям, эукариотам) [88-93]. Согласно принятой номенклатуре перед каждым белком появился один из префиксов: “u” = “universal”, обозначающий универсальность рибосомного белка среди прокариот и эукариот (uS2), при этом нумерация может отличаться у эукариот, “b” = “bacterial”, указывающий на исключительно бактериальный белок (bS6), “e” = “eukaryotic”, говорящий о специфичности белка для эукариот или для эукариот и архей одновременно (eS6), при этом белок не имеет гомолога у прокариот [93]. Универсальные белки (15 в 30S и 19 в 50S [88]) представляют собой эволюционную основу белковой составляющей рибосом, которая, должно быть, являлась значимой и существовала ещё до разделения доменов.



***Thermus thermophilus***

Рисунок 22. Отдельные структурные элементы малой 30S субчастицы рибосомы. Для каждого элемента показаны цепочки фрагментов РНК и входящие в его состав белки вместе с позиционированием элемента на контуре 30S, показанной серым, а также выделенные разными цветами элементы вторичной структуры рРНК. Вид на декодирующий центр и платформу представлен с внешней стороны, обращенной к цитозолу. В рамке декодирующего центра мРНК будет двигаться по направлению справа (издалека) налево (приближаясь). В рамке платформы спирали h21-h23, образующие каркас платформы, видны со стороны тела, обращенной к растворителю. Вид на голову и плечо показан с внутренней стороны, обращенной к 50S субчастице в собранной рибосоме, при этом «нос» указывает налево. Вид на основание приведен при повороте на 90°, при этом 50S субчастица была бы слева в собранной рибосоме, а растворитель – справа. Справа сверху представлена схема вторичной структуры 16S рРНК для *Thermus thermophilus* [94] с нумерацией спиралей согласно [95]. Адаптировано из источника [96].

Последующий этап процесса инициации включает в себя присоединение большой 50S субчастицы рибосомы к комплексу 30S IC и образование 70S пре-инициаторного комплекса (70S PIC) (Рис. 19) [54]. Образование нового комплекса активирует гидролиз молекулы GTP, связанной с IF2, до GDP и свободной фосфатной группы фактором IF2, что делает выбор сайта инициации необратимым и приводит к диссоциации ложных комплексов [97]. В свою очередь заключительное формирование 70S инициаторного комплекса (70S IC) сопровождается диссоциацией комплекса факторов инициации IF1-IF3 и рибосомы, причём очередность ухода факторов отличается в разных источниках [6, 54-55, 64]. При этом fMet-тРНК<sup>Met</sup> остаётся в классическом положении Р/Р, когда Р-сайты большой и малой субчастицы расположены друг под другом, а инициаторная тРНК располагается в этих Р-сайтах. Готовый 70S IC переходит к элонгации, во время которой происходит быстрое считывание мРНК [55, 64, 81].

На стадии инициации выделяют четыре основные кинетические контрольные точки, так называемые чекпоинты (на английском языке “checkpoints”): первая – связывание мРНК с 30S, также известное как рекрутирование, вторая – разворачивание вторичной структуры мРНК, третья – распознавание старт-кодона, четвёртая – образование комплекса малой 30S и большой 50S субчастиц рибосомы (Рис. 23) [64]. При этом скорость процесса на первом чекпоинте сильно зависит от последовательности 5'-UTR мРНК ввиду разной способности разных мРНК связываться с малой субчастицей, что указывает на важность влияния нуклеотидной последовательности мРНК на эффективность трансляции.

#### Кинетические чекпоинты при выборе мРНК

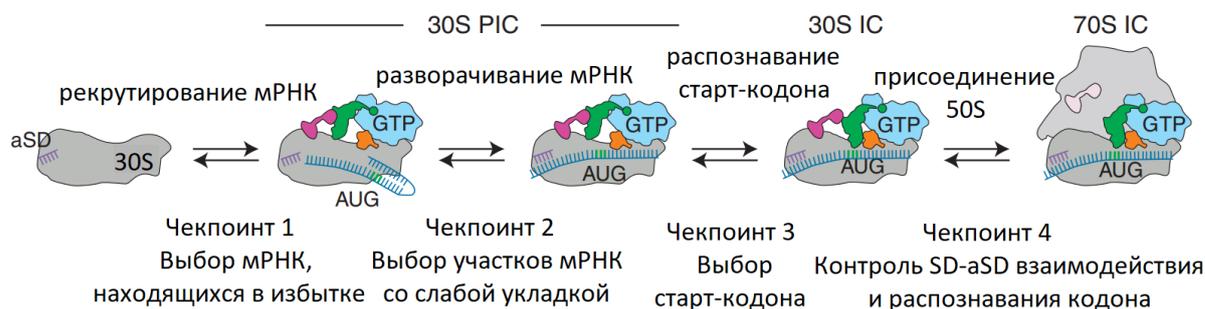


Рисунок 23. Кинетические чекпоинты при выборе мРНК в ходе инициации бактериальной трансляции. Структурированные мРНК могут рекрутироваться, т.е. взаимодействовать с платформой 30S субчастицы, разворачиваться, а затем претерпевать аккомодацию в мРНК-связывающем канале 30S субчастицы. Адаптировано на основе источников [64, 98].

Далее происходят декодирование и аккомодация стадии элонгации трансляции (Рис. 24). После инициации А-сайт остаётся открытым для прихода следующей за инициаторной аминоксил-тРНК в составе тройного комплекса наряду с элонгационным фактором Tu (EF-Tu) и

GTP [54]. Этот комплекс изначально связывается в такой конформации, при которой тРНК взаимодействует своей антикодоновой петлёй с текущим кодоном на мРНК в декодирующем центре, тогда как её аминокислотированный 3'-акцепторный стебель остаётся ассоциированным с EF-Tu, такое состояние принято называть A/T положением [54].

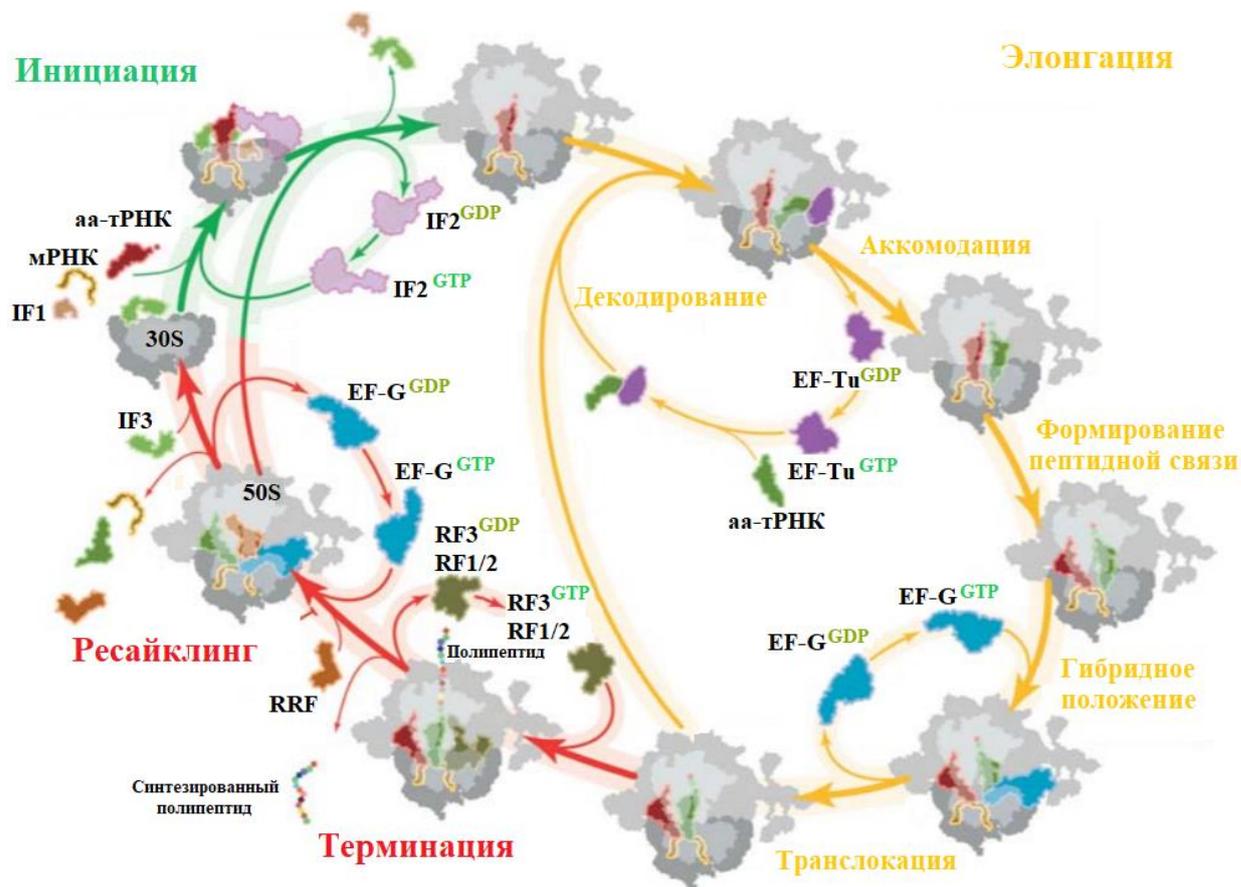


Рисунок 24. Цикл прокариотической трансляции. Стадии инициации, элонгации, терминации и ресайклинга выделены зелёным, жёлтым и красным цветами соответственно. тРНК, транспортная РНК; мРНК, матричная РНК; IF1-IF3, инициаторные факторы 1-3; EF-Tu, элонгационный фактор Tu; GDP, гуанозиндифосфат; GTP, гуанозинтрифосфат; EF-G, элонгационный фактор G; RRF, рибосомный фактор рециклирования. Адаптировано на основе источников [54, 99].

Для обеспечения точности трансляции рибосоме необходимо отличать присоединение соответствующей (“cognate”) от несоответствующей (“non-cognate”) тРНК путём мониторинга взаимодействия между кодоном на мРНК в А-сайте и антикодомом тРНК [100]. Во время распознавания кодона нуклеотиды A1492 и A1493 (согласно нумерации в *E. coli*) принимают конформацию вывернутой спирали 44 (h44) 16S рРНК и вместе с G530 проверяют правильность Уотсон-Криковской геометрии первых двух пар оснований кодон-антикодонного взаимодействия [100]. В третьем положении в кодоне допускаются пары нуклеотидов G·U. Это

позволяет одной тРНК декодировать несколько кодонов, отличающихся только по третьему положению в антикодоне, таким образом обеспечивая вырожденность генетического кода. Такое распознавание запускает много конформационных изменений в рибосоме, затрагивая EF-Tu, что в конечном итоге приводит к стимулированию гидролиза GTP [54].

Активность EF-Tu гидролизовать GTP контролируется правильным позиционированием каталитического гистидина 84 (H84/His84) в EF-Tu вблизи фосфата A2662 сарцин-рициновой петли (СПП/SRL) в спирали H95 23S рРНК большой субчастицы рибосомы. Это позволяет спирали H84 23S рРНК координировать молекулу воды для нуклеофильной атаки по  $\gamma$ -фосфату GTP, который затем гидролизуется. Гидролиз GTP и высвобождение неорганического фосфата ( $P_i$ ) вызывают структурные перестройки в EF-Tu, приводящие к диссоциации фактора от рибосомы, и таким образом, позволяя тРНК переходить из положения A/T в A/A. Во время аккомодации акцепторный стебель тРНК перемещается в A-сайт ПТЦ на большой 50S субчастице.

Следуя по циклу прокариотической трансляции, далее происходят повторяющиеся стадии элонгации: формирование пептидной связи и транслокация по мере синтеза полипептида [54].

#### **1.4.2. Инициация трансляции неканонических мРНК**

В зависимости от особенностей строения 5'-UTR выделяют разные механизмы инициации трансляции у прокариот (Рис. 25) [6, 37, 64, 101]. Помимо хорошо изученного вышеописанного процесса инициации для мРНК, содержащих последовательность Шайна-Дальгарно в 5'-UTR, существуют и другие механизмы начала трансляции.

Далеко не все мРНК содержат в 5'-нетранслируемой области последовательность SD. Такие мРНК встречаются у большинства бактерий и архей [47, 51, 102-103]. Гены со стартовым AUG кодоном с большей вероятностью имеют последовательность SD в 5'-UTR, чем гены либо с GUG, либо с UUG в качестве инициаторных кодонов [47]. У архей и некоторых бактерий у внутренних генов оперонов гораздо более высокая частота встречаемости SD в 5'-UTR, чем у лидирующих генов [47, 51].

Очень мало известно о механизме инициации трансляции с мРНК без SD в 5'-UTR (“non-SD-led” мРНК). Как правило, такие мРНК менее структурированы в участке инициации трансляции. Фактически отсутствие локальной вторичной структуры РНК вокруг старт-кодона имеет решающее значение для экспрессии в бактериях конструкций репортёрных генов, в 5'-UTR мРНК которых отсутствует SD [52, 104]. Более того, этот участок в таких мРНК демонстрирует изменение частот нуклеотидов, создающее неблагоприятные условия для образования вторичной структуры РНК. В отличие от стартового кодона, для фланкирующих последовательностей внутренних AUG кодонов эти особенности не наблюдаются [52, 105], что свидетельствует об

отсутствии вторичной структуры как универсальной особенности в управлении инициацией трансляции мРНК без SD в 5'-UTR на правильном старт-кодоне. Данная особенность по-видимому облегчает формирование инициаторного комплекса, тем самым компенсируя отсутствие взаимодействий между последовательностями SD и анти-SD в 5'-UTR мРНК.

мРНК без последовательности SD в 5'-UTR быстро связываются с 30S субчастицей и образуют устойчивый комплекс, приступающий к трансляции [98]. Рибосомный белок bS1 и фактор IF3, как показано, поддерживают трансляцию таких мРНК в *E. coli* [106]. Белок bS1 облегчает связывание 30S субчастицы с мРНК благодаря высокой аффинности к U-богатой или AU-богатой последовательности, обычно присутствующей в 5'-UTR без SD [14]. IF3 в данном случае потенциально участвует в выборе инициаторного кодона в Р-сайте [107].

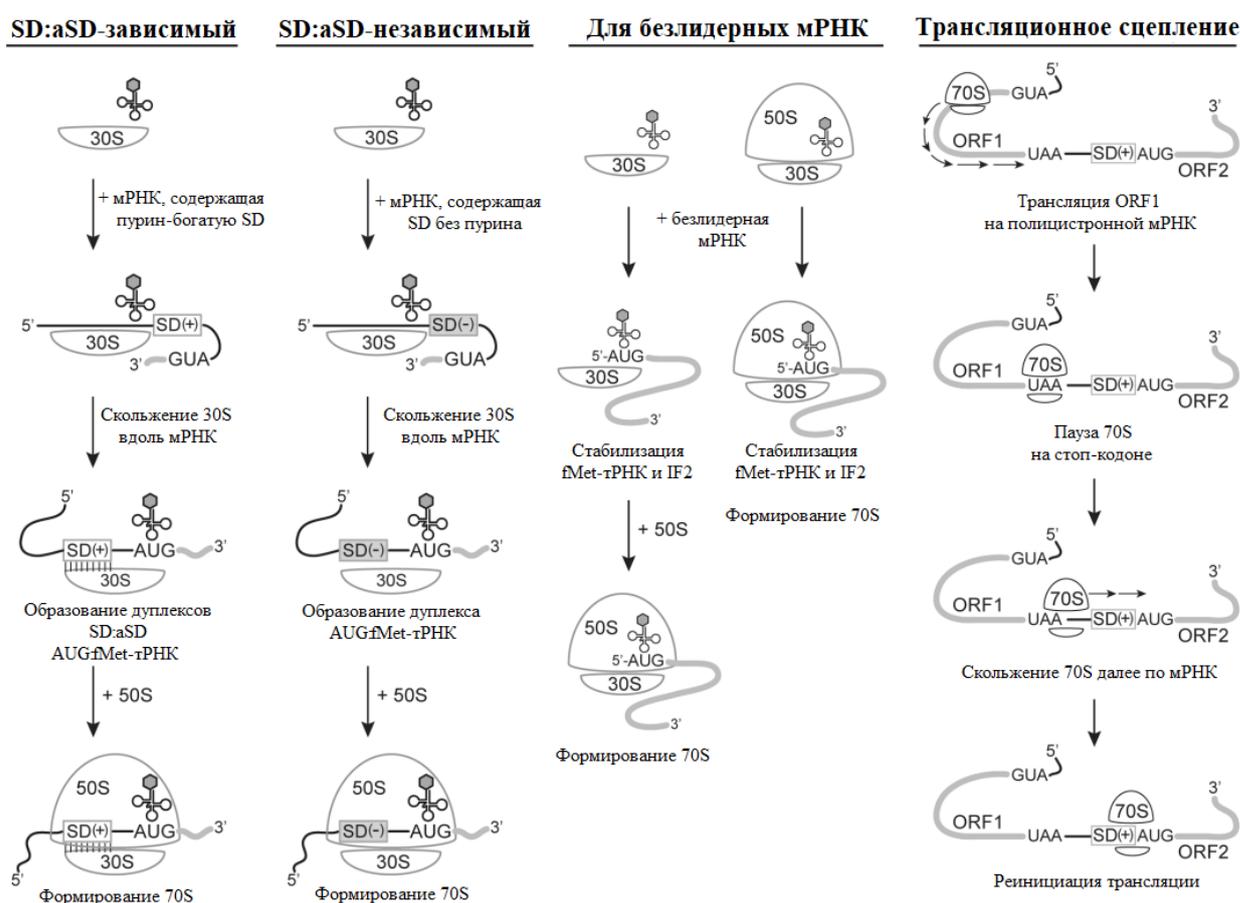


Рисунок 25. Механизмы инициации трансляции у прокариот. Последовательности SD на мРНК, способные или неспособные формировать стабильный дуплекс с анти-SD (aSD) на 16S рРНК, обозначены как SD(+) и SD(-) соответственно. Открытые рамки считывания (ORF) на мРНК показаны серым цветом. Процесс сканирования 70S представлен как репрезентативный для трансляционного сцепления. Адаптировано из источника [101].

Тем не менее, механизмы, по которым выбирается правильный инициаторный кодон в

таких мРНК, до конца не ясны. Единственное предположение пока основано на наблюдаемом отсутствии вторичной структуры вокруг старт-кодона [105]. При этом AUG кодон сам по себе является важным элементом для распознавания. Другим потенциальным медиатором является белок bS1, который может узнавать AU-богатый энхансер в мРНК как с SD, так и без неё.

Один из самых любопытных примеров мРНК без SD в 5'-UTR в *E. coli* представляет собой мРНК гена *rpsA*, кодирующая рибосомный белок bS1 малой субчастицы, распознающий AU/U-богатые элементы на мРНК. Участок инициации трансляции bS1 мРНК демонстрирует ряд необычных черт. Он простирается далеко перед инициаторным AUG кодоном (до -91 позиции относительно старт-кодона), у него отсутствует каноническая SD в 5'-UTR, и он может укладываться в три следующие друг за другом шпильки, которые необходимы для его высокой трансляционной активности [14, 37, 72]. Интересно, что эффективность трансляции *rpsA* мРНК сильно репрессируется свободным белком bS1, что указывает на механизм ауторегуляции [14].

В мРНК прокариотических «генов без лидера» 5'-UTR очень короткая или полностью отсутствует («leaderless» мРНК) вместе с канонической SD перед кодирующей областью. Длина такой 5'-UTR составляет менее 8 [101, 108-109] или 6 [110-111] нуклеотидов (нт), при этом в случае первой длины предполагают, что это минимальный размер 5'-UTR, необходимый для образования взаимодействия SD-анти-SD. «Гены без лидера» широко распространены в различных археях и некоторых бактериях [101, 109]. Например, у *Actinobacteria*, *Deinococcus* и *Thermus* более 20% генов являются безлидерными [53], в отличие от *E. coli*, где обнаружено менее 8% таких генов [51, 110, 112-114]. При этом данные оценок разнятся от порядка 50 [110] до 170 генов из базы RegulonDB [114] без лидерной последовательности 5'-UTR, причём пересечение между данными наблюдается для около 30 таких генов, преимущественно они без 5'-UTR и зачастую кодируют малые РНК.

Представление о механизме инициации трансляции безлидерных мРНК (Рис. 25) остаётся весьма ограниченным [64, 115]. Ввиду того, что в случае разных мРНК (с SD и без неё в 5'-UTR, без 5'-UTR) 5'-кодирующая последовательность не демонстрирует отличительных черт [53], старт-кодон, расположенный на 5'-конце безлидерной мРНК, по-видимому является наиболее важным сигналом для инициации трансляции [116-117]. Белки bS1 и uS2 малой 30S субчастицы, модулирующие привлечение мРНК с 5'-UTR, являются несущественными для трансляции безлидерных мРНК [116]. Более того, ряд других рибосомных белков также не обязателен: bS6, uS12, bS18 и bS21 [118].

Экспериментальные исследования показали, что удаление стартового кодона или хотя бы 5'-терминального фосфата в безлидерных мРНК значительно нарушало их трансляцию [119-120]. Вставка дополнительной последовательности перед старт-кодоном в мРНК, которая исходно была без 5'-UTR, также приводила к снижению эффективности инициации трансляции [121-122].

В лаборатории И.Н. Шатского впервые, ещё в 1992 году, было показано, что 70S рибосома *E. coli* способна связываться с 5'-концом безлидерной мРНК *in vitro* и инициировать трансляцию [123]. В дальнейшем это открытие было подтверждено другими исследователями [122, 124]. Естественно, что образование пар оснований между fMet-тРНК<sup>fMet</sup> и старт-кодоном играло важную роль в стабилизации комплекса рибосомы с мРНК. Кроме того, было показано, что фактор IF2 облегчает распознавание стартового кодона в безлидерной мРНК [115].

Дальнейшие эксперименты *in vivo* также показали, что безлидерные мРНК напрямую связываются с 70S рибосомами, инициаторная fMet-тРНК<sup>fMet</sup> направляется к AUG кодону при участии факторов IF2 и IF3 [125-126]. В то время как IF2 способен связываться аналогичным образом как с 30S субчастицей, так и с 70S рибосомой [127]. Связывание IF3 с 70S рибосомой способствует её диссоциации на субчастицы. Следовательно, возникает вопрос о том, как IF3 может стимулировать инициацию на 70S рибосоме без разделения её на субчастицы. По-видимому, IF3 должен перемещаться со своего сайта связывания на 30S субчастице при присоединении 50S субчастицы. Полученные результаты показывают, что после диссоциации от своего сайта связывания на 30S при присоединении 50S IF3 может оставаться связанным с неканоническим сайтом на 50S субчастице, что позволило бы фактору действовать в случае инициации с безлидерной мРНК, не приводя к диссоциации 70S рибосомы на субчастицы [127].

После трансляции первой открытой рамки считывания полицистронной мРНК при трансляционном сцеплении рибосома может приступить к реинициации далее перед второй рамкой, используя механизм сканирования 70S по матрице РНК, требующий присутствия fMet-тРНК<sup>fMet</sup> и IF3 (Рис. 25) [126]. Непосредственная близость инициаторного кодона к 5'-концу является сильной детерминантой как для связывания рибосомы, так и для экспрессии белка [128].

Примечательно, что безлидерные мРНК могут иметь важную роль в адаптации к стрессовым условиям у бактерий. Индуцированная в условиях стресса эндорибонуклеазная активность токсина MazF у *E. coli* разрезает одноцепочечные мРНК по сайту АСА непосредственно перед AUG старт-кодоном некоторых специфичных мРНК, тем самым образуя безлидерные мРНК [129]. Более того, мишенью MazF также является 16S рРНК, у которой нуклеаза удаляет последовательность анти-SD с 3'-конца. Это приводит к формированию субпопуляции рибосом, избирательно транслирующих безлидерные мРНК в условиях стресса [129]. В то же время встречалась информация о том, что трансляция мРНК с или без лидерной последовательности перед геном по-разному реагирует на изменения температуры [130], демонстрируя связь типа используемой мРНК с адаптацией к изменениям окружающей среды.

В ходе такого ключевого процесса в жизнедеятельности всех организмов, как трансляция, происходит биосинтез белков с помощью рибосомы – сложного макромолекулярного рибонуклеопротеидного комплекса. При этом рРНК и рибосомные белки, входящие в состав

рибосомы, демонстрируют сложную сеть структур и взаимодействий между собой, нарушение которых может приводить к неправильному синтезу белков или остановке трансляции. Рибосома декодирует информацию, в виде последовательности нуклеотидов записанную в мРНК, и переводит её в цепочку аминокислот, формирующих белок [22]. Природные 5'-UTR мРНК представлены широким разнообразием у прокариот. В зависимости от строения 5'-UTR процесс инициации трансляции, в свою очередь ограничивающий скорость биосинтеза белка, идёт по разным механизмам. При этом механизмы инициации трансляции не для всех видов 5'-UTR мРНК хорошо изучены.

## **2. Подходы к изучению роли 5'-UTR в эффективности трансляции**

### **2.1. 5'-UTR – мишень для регуляции эффективности трансляции**

Регуляция трансляции на стадии инициации используется клетками чаще, чем регуляция на других стадиях, так как это позволяет минимизировать непродуктивную трату клеточных ресурсов. Исследования этого процесса проводятся уже давно. Известно, что контроль инициации трансляции дифференцирован, так как осуществляется благодаря различным участникам регуляции. Одним из них является 5'-нетранслируемая область мРНК (Рис. 1) [131], поэтому её можно назвать мишенью для регуляции эффективности трансляции. Известные элементы, обеспечивающие эффективное связывание рибосомы с мРНК, кодирующей тот или иной белок, локализуются, прежде всего, в 5'-UTR.

В 5'-UTR располагается сайт посадки рибосомы, содержащий в канонических мРНК последовательность Шайна-Дальгарно [9, 13, 47, 50, 54, 65, 96, 132-134], в свою очередь комплементарную 3'-концу 16S рРНК [44, 135] и являющуюся самым изученным регуляторным элементом. Для обеспечения высокой эффективности трансляции SD должна быть оптимальной длины и находиться на оптимальном расстоянии от старт-кодона [136-138]. В одной 5'-UTR можно встретить несколько SD [137]. Для начала эффективной трансляции участок инициации трансляции должен быть либо целиком одноцепочечным, либо принимать укладку вторичной структуры, которую можно легко разрушить [139-142]. Другими известными элементами, способными влиять на эффективность трансляции, являются богатый остатками аденина и урацила (AU-богатый) участок мРНК, с которым взаимодействует белок малой 30S субчастицы рибосомы bS1 [6, 15, 93], и начальный участок кодирующей области, напрямую следующий за инициаторным кодоном [143-145].

На сегодняшний день известны различные методы, позволяющие изучать функциональную значимость отдельных участков мРНК для трансляции. Среди них выделяют сайт-направленный мутагенез [146] или рандомизацию участков в 5'-UTR [147-148], зачастую предшествующей гену флуоресцентного белка, и оценку интенсивности его свечения *in vitro* или

*in vivo*, которая отражает эффективность трансляции.

Следующий метод представлен термодинамическим моделированием *in silico* [138, 149-153], основанным на оценке силы молекулярных взаимодействий комплекса 30S с мРНК-транскриптом и предсказывающим скорость инициации трансляции. Его применение позволяет определить величины, обеспечивающие связь с эффективностью трансляции. Результаты моделирования можно выборочно валидировать экспериментально, используя репортёрные конструкции.

С появлением такого метода, как проточная цитометрия, стала возможной одновременная оценка различных параметров большого количества клеток *in vivo* и сортировка их на отдельные фракции по сходству отдельных параметров, как например, по уровню экспрессии гена флуоресцентного белка [154].

Активное развитие высокопроизводительного секвенирования (NGS) благоприятно сказалось на разработке новых усложнённых подходов к исследованию геномов, поиску корреляции между фенотипом и генотипом, например, представленных полногеномным секвенированием *de novo* (DNA-seq), РНК-секвенированием (RNA-seq) для профайлинга клеточного транскриптома и выделения эффективно транслирующихся мРНК (Ribo-seq), секвенированием библиотек плазмидных ДНК (DNA-seq), ChIP-секвенированием (ChIP-seq) с целью определения сайтов связывания белков, ассоциированных с ДНК [155].

## **2.2. Разнообразие подходов к изучению 5'-UTR в эффективности трансляции**

### **2.2.1. Широкомасштабный анализ генов**

Широкомасштабный анализ генов *Escherichia coli* показал наличие SD у большинства мРНК (Рис. 26), при этом сама SD была открыта в 1970-х в нескольких бактериальных мРНК [9] и является важной для эффективной инициации трансляции [136-138]. Располагаясь на расстоянии 5-8 нуклеотидов перед стартовым кодоном, SD служит местом связывания бактериальной 30S субчастицы, в отличие от эукариотической рибосомы, которая связывается на 5'-конце мРНК для начала сканирования [132]. Длина самой SD различна у разных генов *E. coli* и варьирует в пределах 4-8 нуклеотидов [133].

Разработка методов поиска SD упрощает аннотацию генома и предсказание эффективности экспрессии генов [133]. Для моделирования сайтов посадки рибосомы были использованы две модели: жёсткая, в которой оба параметра SD и участок инициации (IR) зафиксированы относительно нулевой координаты и гибкая, состоящая из двух жёстких участков SD и IR, расстояние между которыми менялось. Полученные результаты при анализе 4122 генов из EcoGene12 (брали все возможные открытые рамки считывания) [134] и 569 генов из «проверенного набора» на основе данных о стартовых участках трансляции белков с высоким

уровнем экспрессии согласно результатам применения двухмерного электрофореза указывают на наличие консервативной последовательности SD перед стартовым кодоном (Рис. 27). С помощью количественного математического описания можно выявить такой важный параметр, как расстояние между SD и IR. Варьирование длины этого участка в жёсткой и гибкой моделях демонстрирует среднее удаление центрального G в последовательности SD от IR на расстоянии 8-11 нт от старт-кодона (-8 – -11) [133], и такое расстояние необходимо для оптимальной экспрессии генов. Наиболее вероятный состав SD на интервале -4 – -18 – agGa.

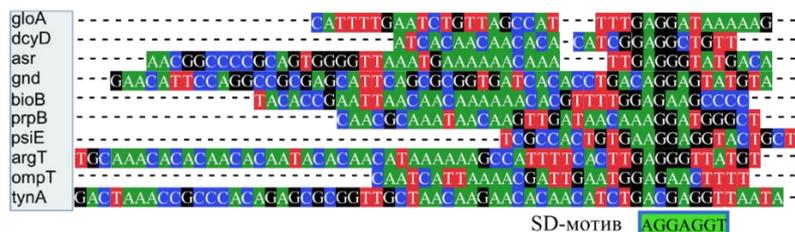


Рисунок 26. Пример выравнивания последовательностей 5'-UTR мРНК, используемого при широкомасштабном анализе этих нетранслируемых областей генов с выделением SD-мотива. Рисунок выполнен в программе Inkscare.

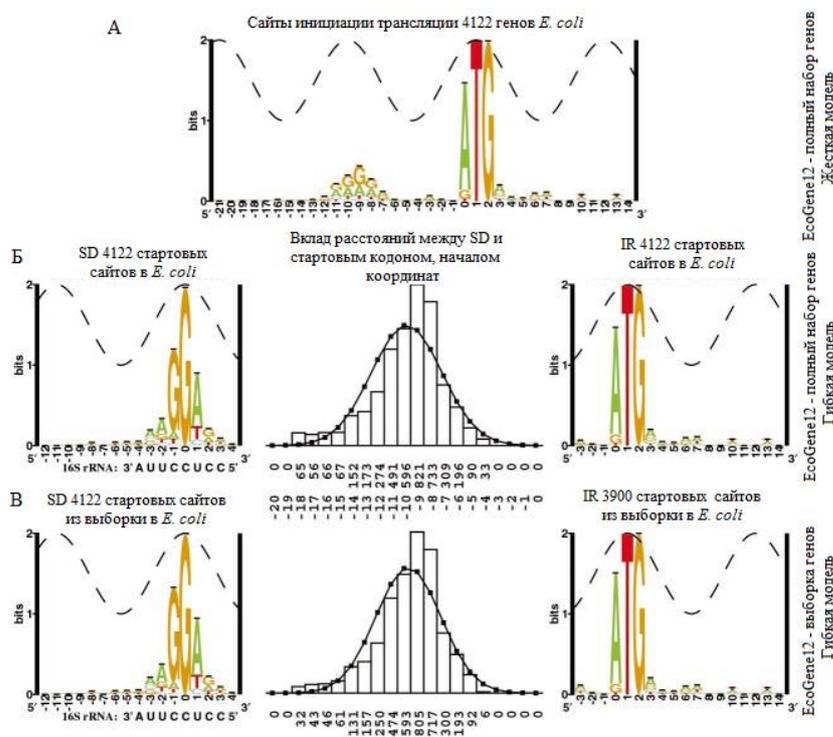


Рисунок 27. Консервативные последовательности, предшествующие IR. А. Анализ всех 4122 ORF в жёсткой модели. Б. Анализ всех 4122 ORF в гибкой модели. В. Анализ всех ORF из «проверенного набора» в гибкой модели. Для подвижной модели приведено распределение расстояний между SD и IR. Высота букв отражает частоту встречаемости нуклеотида в данной позиции. Адаптировано на основе источника [133].

## 2.2.2. Использование репортёрных конструкций

Зависимость эффективности трансляции от длины последовательности Шайна-Дальгарно и её расстояния от стартового кодона исследовали с помощью различных подходов. Одним из самых распространённых является использование двойной генно-инженерной конструкции, в которой закодированы гены двух флуоресцентных белков: например, красного RFP, служащего внутренним контролем, и голубого CER, выступающего в качестве сенсорного в ответ на изменения в элементах 5'-UTR мРНК (Рис. 28) [137]. Оценка отношения измеренных интенсивностей флуоресценции двух белков CER/RFP *in vivo* позволяет нивелировать эффекты, связанные с размером клетки бактерий и флуктуациями копийности репортёрной конструкции.



Рисунок 28. Схема основной части плазмиды pRFPCER, 5'-UTR – 5'-нетранслируемая область, T5-1, T5-2 – промоторы, стрелками отмечены места начала транскрипции, прямоугольниками – места окончания. Расположение сайтов узнавания эндонуклеаз рестрикции, используемых для модификации плазмиды, также отмечено. Адаптировано на основе источников [137, 156].

Вышеописанный подход был применён для постановки эксперимента с целью изучить влияние длины SD, расстояния от SD до инициаторного кодона и их комбинаций на эффективность трансляции белка CER. Этот эксперимент основан на молекулярном клонировании с использованием 16 репортёрных плазмид с SD длиной 2, 4, 6 и 8 нт, находящейся на разном расстоянии (7, 10, 13 и 16 нт) от старт-кодона гена белка CER, и одной контрольной плазмиды без участков комплементарности к анти-SD 16S рРНК. В результате были обнаружены следующие наблюдения: при увеличении расстояния от SD до инициаторного кодона уменьшается эффективность трансляции мРНК с 8-нуклеотидной последовательностью SD, для 6-нуклеотидной SD оптимальным расстоянием является 10 нт, схожая с 8-нуклеотидной SD зависимость найдена и для 4-нуклеотидной SD, для короткой SD длиной 2 нт существенного влияния расстояния не видно, несмотря на обеспечение этой SD на порядок большей эффективности трансляции по сравнению с контрольной плазмидой без SD. Изменение таких параметров позволяет варьировать уровнем биосинтеза белка до 4-х порядков, свидетельствуя об их значимости при определении уровня большинства белков в клетке [137].

### 2.2.3. Сайт-направленный мутагенез

Быстрый и достаточно простой количественный анализ экспрессии генов *in vitro* возможен при получении сайт-направленным мутагенезом на основе полимеразной цепной реакции (ПЦР) многочисленных вариаций последовательности SD и AU-богатого трансляционного энхансера. При этом продукт ПЦР, содержащий T7-промотор, тестируемый вариант 5'-UTR и ген флуоресцентного белка eGFP, напрямую используется в системе *in vitro*, полученной из клеток *E. coli*, где сопряжены транскрипция и трансляция [146]. Эффективность трансляции можно оценить по наблюдаемой интенсивности флуоресценции белка eGFP в данной системе (Рис. 29). Полученные таким образом 54 последовательности 5'-UTR: 18 и 36 вариантов последовательностей SD и AU-богатых соответственно, обеспечили диапазон относительных уровней экспрессии от 0,1 до 2,0, демонстрируя влияние различных RBS на эффективность трансляции [146]. Однако у такого точечного подхода есть сильное ограничение – небольшой набор тестируемых вариантов, далёкий от всего разнообразия природных 5'-UTR перед генами, число которых составляет около  $4 \times 10^3$  у *E. coli* [134].

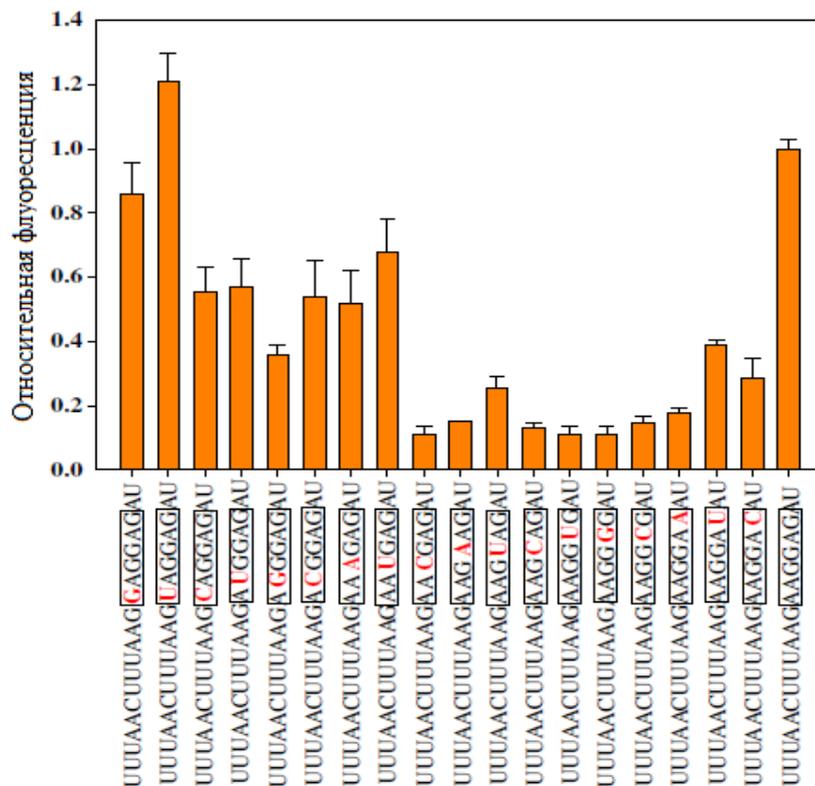


Рисунок 29. Относительные уровни экспрессии флуоресцентного белка eGFP, нормализованные относительно интенсивности флуоресценции белка eGFP с контрольным RBS (последний вариант), после *in vitro* трансляции с использованием точечных мутаций (красные буквы) в каждой позиции SD в RBS. Адаптировано на основе источника [146].

## 2.2.4. Аффинное выделение рибосом с мРНК

Немного позже разработана экспериментальная система (Рис. 30) [147], в основе которой *in vitro* трансляция и которая позволила провести селекцию наиболее эффективно транслируемых мРНК из большой выборки искусственных последовательностей. Для этого использовалась модельная мРНК, содержащая 81 нуклеотидную 5'-UTR, в которой 18 оснований перед стартовым кодоном были полностью рандомизированы, что позволило получить библиотеку из  $\sim 6.9 \times 10^{10}$  различных последовательностей.

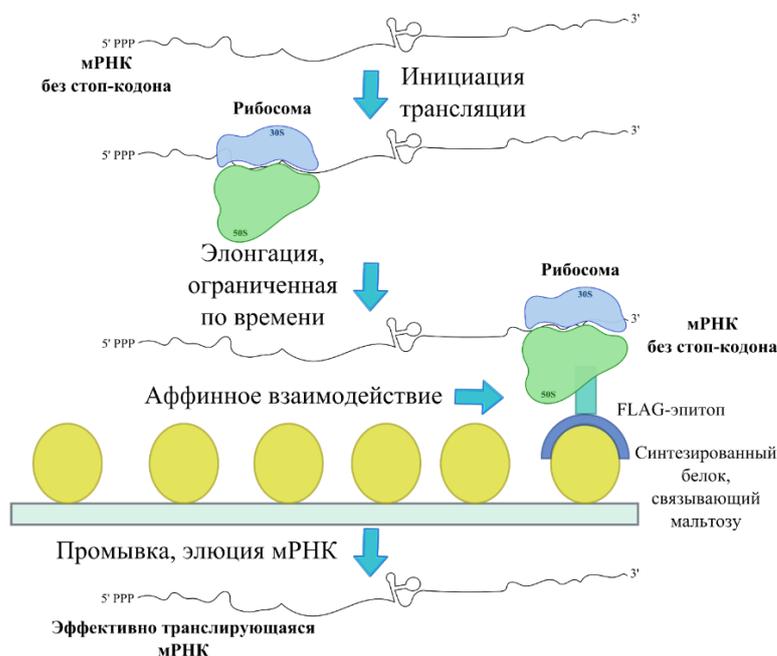


Рисунок 30. Схема аффинного выделения рибосом с эффективно транслирующейся мРНК. Отбор проводился лимитированием времени *in vitro* трансляции. мРНК содержит 5'-UTR, кодирующую область, включающую участок, кодирующий FLAG-эпитоп, который взаимодействует с синтезируемым белком, связывающим мальтозу, и TolA, позволяющий эпитопу выйти из рибосомного туннеля и правильно свернуться. В конструкции нет стоп-кодона, поэтому рибосома остаётся на мРНК. Рисунок выполнен в программе Inkscape.

В модельной мРНК закодирован гибридный белок, содержащий примерно в середине домен, связывающий мальтозу, и FLAG-эпитоп, позволяющий проводить аффинную очистку рибосом, синтезирующих этот гибридный белок. После домена, используемого для аффинной очистки, закодирован фрагмент белка TolA, служащий исключительно в качестве спейсера, достаточного для того, чтобы по окончании синтеза полноразмерного гибридного белка был экспонирован аффинный домен из пептидного туннеля. В данной мРНК не закодирован стоп-кодон, поэтому она оставалась связанной с рибосомой и по окончании синтеза белка. Таким образом, мРНК могла быть экстрагирована из аффинно-выделенных рибосом и далее

амплифицирована. Ключевым параметром селекции мРНК служило ограниченное время трансляции. При этом только мРНК, транслируемые быстро, могли быть аффинно очищены и использованы в следующем раунде отбора (Рис. 30) [147].

Удивительно, что среди отобранных последовательностей, обеспечивающих наиболее быструю трансляцию в *in vitro* системе, 76% не содержали SD и имели С-богатые короткие участки, комплементарные 16S рРНК. Однако для мРНК с такими С-богатыми последовательностями не наблюдалась высокая экспрессия *in vivo*, возможно, из-за разного среднестатистического соотношения рибосом и мРНК, конкурирующих, наряду с С-богатыми RBS, за связывание с рибосомами, в *in vitro* и *in vivo* системах [147].

Тот же самый эксперимент был проведён с библиотекой более коротких мРНК с 5'-UTR длиной 40 нуклеотидов [148], близкой к наиболее часто встречающимся у мРНК *E.coli* [110, 157]. Секвенирование нового поколения и применение статистических методов позволили определить мотивы связывания мРНК и рибосомы. Библиотека мРНК, полученная при отборе с ограниченным временем трансляции, содержала 46% последовательностей с SD (против 24% в предыдущей работе [147]), а остальные были преимущественно G/U-богатыми [148].

Обе системы в последней [147] и более ранней [148] работах отличаются только длиной 5'-UTR, что может указывать на преобладание определённых нуклеотидов в зависимости от контекста: С-богатые участки не обеспечивают эффективную инициацию трансляции, находясь в более коротких 5'-UTR, в то время как G/U-богатые последовательности – в более длинных.

### 2.2.5. Оценка стабильности вторичных структур

Последовательность 5'-UTR мРНК может определять различную укладку участка перед инициаторным кодоном. Связь между стабильностью вторичных структур в участке инициации и эффективностью трансляции подтвердили крупномасштабным вычислительным анализом [139], указывающим на тенденцию генов прокариот и эукариот, особенно с высоким уровнем экспрессии, к дестабилизации вторичной структуры мРНК вблизи старт-кодона [140].

Стабильность ( $< -12$  ккал/моль) шпильчатой структуры, содержащей сайт посадки рибосомы, изменяли с помощью сайт-направленного мутагенеза и анализировали выход белка *in vivo*. Оказалось, что чем выше стабильность спирали РНК, содержащей участок связывания рибосомы, тем ниже эффективность биосинтеза белка среди 23 вариантов TIR. Возрастание её  $\Delta G^0$  на  $-1,4$  ккал/моль (т.е. меньше, чем разница между А·U и G·C парами) соответствует падению скорости инициации в 10 раз [141]. Поэтому замена одного нуклеотида может приводить к 500-кратному снижению экспрессии, если образуется комплементарная пара оснований и стабилизируется вторичная структура мРНК. Таким образом, была показана полная зависимость начала трансляции от разворачивания всего участка инициации мРНК [141]. Однако эта

спонтанность обусловлена присутствием всех необходимых элементов комплекса инициации (Рис. 31) [142].

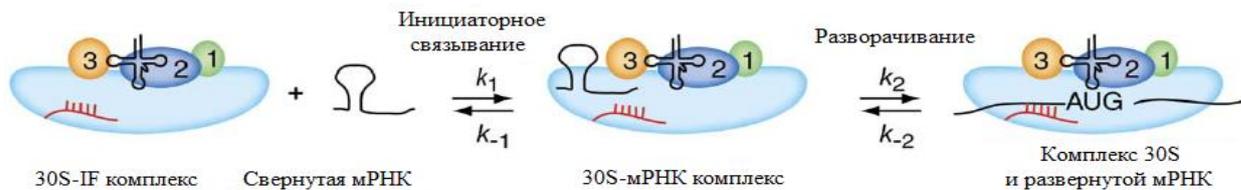


Рисунок 31. Связывание структурированной мРНК с бактериальным инициаторным комплексом. На первом этапе свёрнутая мРНК связывается с 30S субчастицей в комплексе с fMet-тРНК<sup>fMet</sup>, IF1, IF2 и IF3. На следующем шаге 30S-комплекс разворачивает мРНК и формирует Уотсон-Криковские пары с последовательностью анти-SD 16S рРНК (показана красным цветом), таким образом получается стабильный 30S инициаторный комплекс. Адаптировано из источника [142].

Как показано при анализе 12-ти мРНК, формирующих вторичные структуры с различной стабильностью и содержащих SD разной длины либо без SD, разворачивание вторичных структур мРНК требует наличия последовательности SD, стартового кодона, fMet-тРНК<sup>fMet</sup>, GTP-связанной формы фактора IF2 в комплексе с 30S субчастицей рибосомы. Вклад каждого такого индивидуального элемента в нарушение укладки инициаторного участка мРНК оценен с помощью константы диссоциации фрагмента мРНК с 6-нуклеотидной SD (Таблица 1) [142].

Таблица 1. Кинетические параметры диссоциации мРНК,  $k_{dis}$  ( $c^{-1}$ ) в присутствии различных комбинаций компонентов инициации трансляции.

Комплекс	мРНК с SD-6
30S	$0,8 \pm 0,08$
30S + IFs**	$0,7 \pm 0,1$
30S + fMet-тРНК <sup>fMet</sup>	$0,7 \pm 0,1$
30S + IF1 + fMet-тРНК <sup>fMet</sup>	$0,7 \pm 0,1$
30S + IF3 + fMet-тРНК <sup>fMet</sup>	$0,7 \pm 0,1$
30S + IF1 + IF3 + fMet-тРНК <sup>fMet</sup>	$0,7 \pm 0,1$
30S + IF2 + IF3 + fMet-тРНК <sup>fMet</sup>	$0,7 \pm 0,1$
30S + IFs** + fMet-тРНК <sup>fMet</sup>	$0,006 \pm 0,002$
30S + IF2 + fMet-тРНК <sup>fMet</sup>	$0,0002 \pm 0,00005$
30S + IF1 + IF2 + fMet-тРНК <sup>fMet</sup>	$0,0001 \pm 0,00004$

Кинетические константы представляют среднее значение как минимум из трёх экспериментов с учётом стандартных отклонений. “IFs” - совокупность всех трёх типов факторов IF1, IF2, IF3.

Далее проведён анализ с помощью метода FRET этого же фрагмента мРНК ( $T_{пл} = 58^{\circ}\text{C}$ ,  $\Delta G^{\circ} = -2,5$  ккал/моль) с мечеными Cy3 и Cy5 5'- и 3'-концами соответственно в чистом виде, при добавлении 30S субчастицы вместе со всеми элементами, необходимыми для инициации трансляции. В качестве контрольной использовали мРНК без SD, но с близкой по стабильности вторичной структурой ( $T_{пл} = 51^{\circ}\text{C}$ ,  $\Delta G^{\circ} = -1,7$  ккал/моль). В результате проведённых экспериментов был сделан вывод о значимости последовательности SD в разворачивании вторичной структуры мРНК (Рис. 32) [142]. При этом показана высокая эффективность FRET для свёрнутой мРНК, концы которой образовывали комплементарное взаимодействие SD с анти-SD, и слабая для развёрнутой формы мРНК [142].

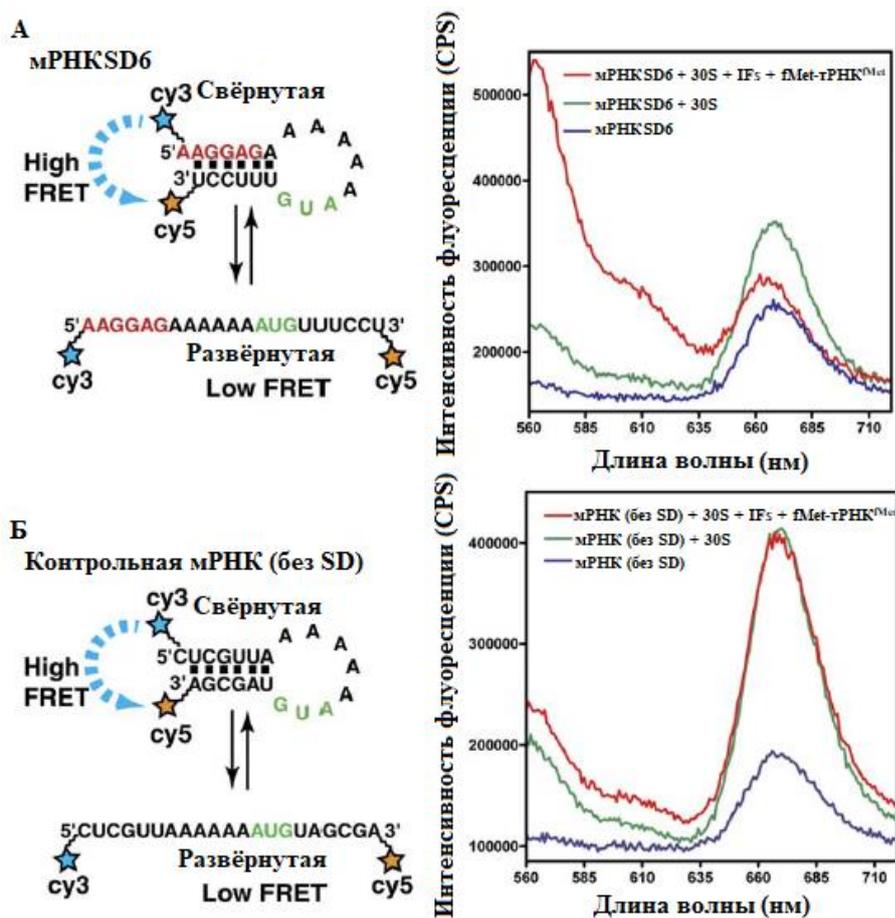


Рисунок 32. Разворачивание мРНК с SD длиной 6 нт. А. Вторичная структура мРНК с SD длиной 6 нт с Cy3 (синяя звёздочка) и Cy5 (оранжевая звёздочка) на 5' и 3' концах соответственно (слева). В свёрнутом состоянии метки Cy3 (эмиссия при 570 нм) и Cy5 (эмиссия при 660 нм) близки друг к другу, и наблюдается высокая эффективность FRET (“High FRET”), в развёрнутом – низкая эффективность FRET (“Low FRET”). Б. Вторичная структура контрольной мРНК без SD (слева). Зависимость интенсивности флуоресценции от длины волны для разных мРНК из А и Б без всего и при добавлении 30S субчастицы либо полного комплекса для инициации трансляции (справа). IFs, совокупность всех трёх типов факторов IF1, IF2, IF3. Адаптировано на основе работы [142].

### 2.2.6. Тоупринтинг

Для оценки эффективности связывания субчастиц рибосомы с определённой последовательностью 5'-UTR мРНК существует метод, называемый тоупринтингом/тоупринтом (на английском языке “toeprinting”) (Рис. 33). Он основан на использовании изотопно или флуоресцентно меченных праймеров, которые комплементарны 3'-концу мРНК. После сборки инициаторного комплекса на мРНК проводится реакция обратной транскрипции с последующим электрофоретическим анализом реакционной смеси. При электрофоретическом разделении можно визуализировать кДНК-продукты удлинения праймеров: до 5'-конца мРНК в случае, если обратная транскриптаза (также известная как ревертаза) не встретила рибосому, и более короткие продукты при ингибировании удлинения праймера, вызванном тем, что ревертаза остановилась в месте связывания рибосомы с мРНК. По соотношению длинных и коротких кДНК-продуктов (тоупринтов) можно судить о проценте мРНК, которые связались с рибосомой в ходе трансляции (Рис. 34) [158-159].

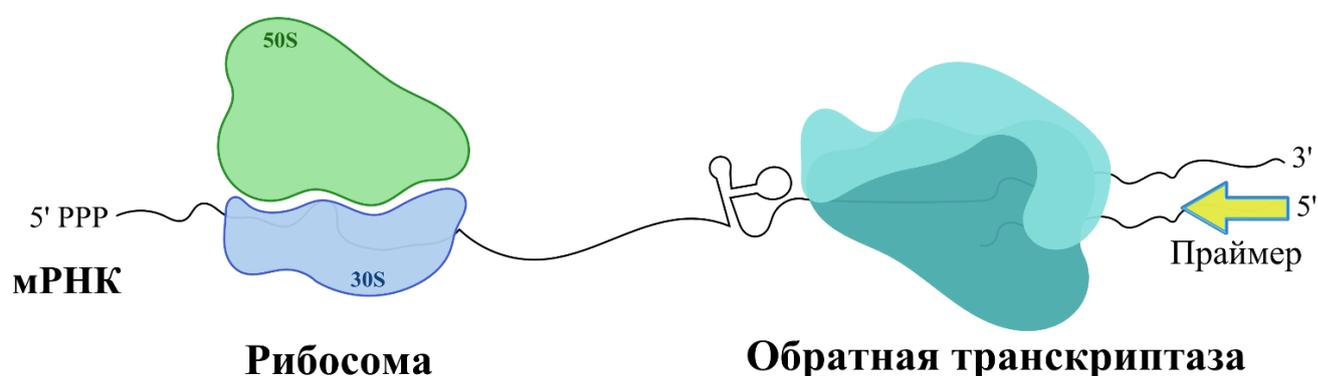


Рисунок 33. Принцип метода тоупринтинга. Стабильные рибосомные комплексы останавливают ревертазу (обратную транскриптазу) в конкретной позиции на мРНК, образуя короткие продукты кДНК определённой длины. Для обратной транскриптазы в качестве праймеров могут служить как радиоактивно, так и флуоресцентно меченные олигонуклеотиды. Рисунок создан в программе Inkscape.

Данный метод нашёл активное применение для изучения локализации остановок рибосомы в присутствии различных антибиотиков, ингибирующих биосинтез белка. Результаты тоупринтинга при добавлении ингибитора трансляции могут позволить пролить новый свет на механизм его действия. Следует также отметить, что использование различных мРНК даёт возможность детектировать сиквенс-специфичность остановок, вызванных действием определённого антибиотика, связанного с рибосомой [160].

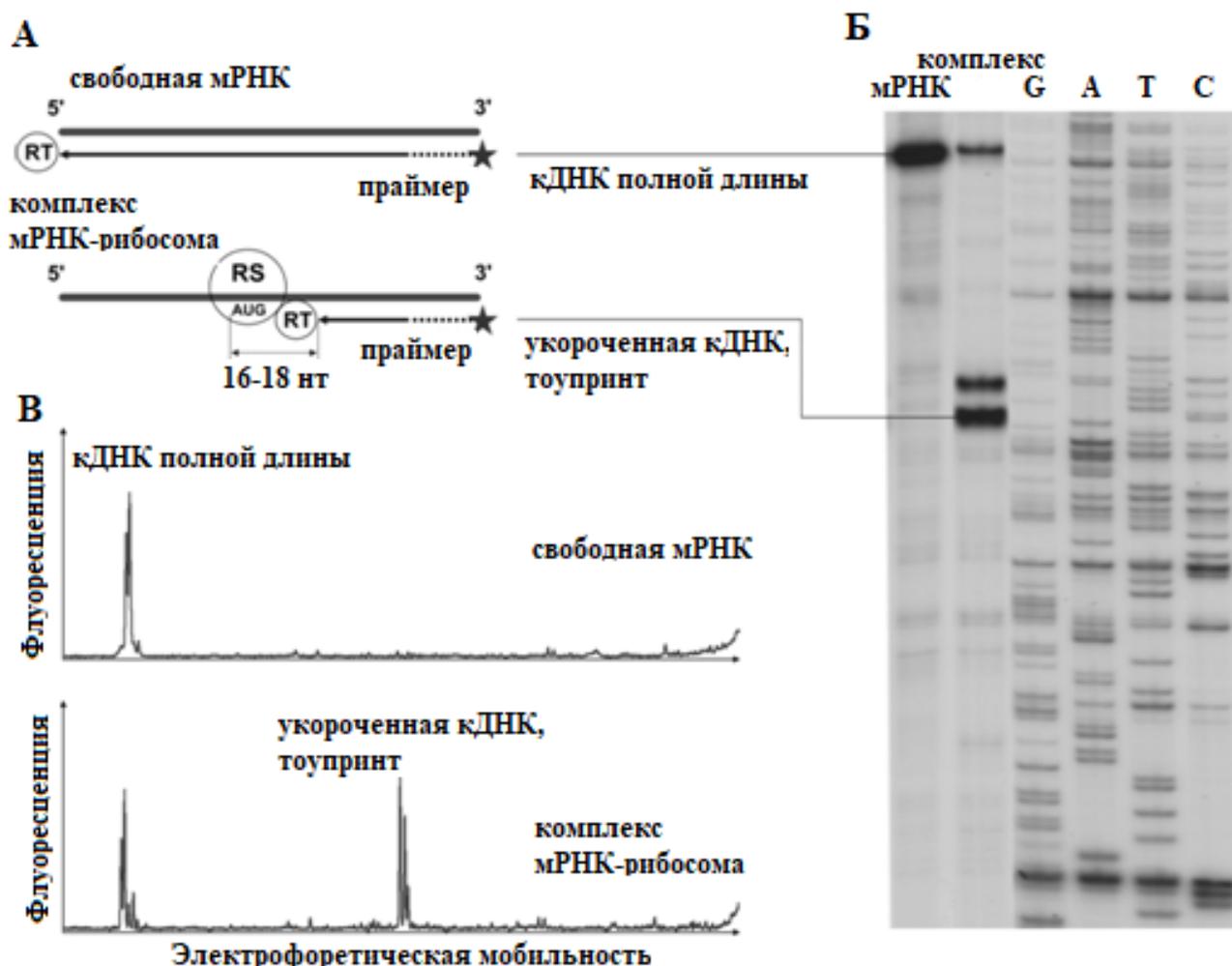


Рисунок 34. Анализ формирования комплекса рибосома-мРНК с помощью тоупринтинга. А. Стабильные рибосомные комплексы (RS) останавливают обратную транскриптазу (RT) в конкретной позиции на мРНК, генерируя короткие кДНК-продукты определённой длины. Сравнение Б – классического эксперимента тоупринтинга, электрофоретического анализа радиоактивно меченных продуктов реакции ингибирования удлинения праймера и В – альтернативного подхода с использованием флуоресцентно меченных праймеров и электрофореза для определения кДНК-продуктов. Адаптировано на основе источника [159].

### 2.2.7. Секвенирование нового поколения (NGS)

По мере накопления экспериментальных результатов и развития разных подходов для их анализа всё большую роль начинают играть биоинформатические методы, позволяющие работать с большими массивами данных.

Способы исследования молекулярных механизмов функционирования биологических систем коренным образом изменились с начала применения экономически эффективных методов с высокой пропускной способностью, таких как секвенирование нового поколения (NGS). Развитие NGS способствовало разработке новых подходов, связанных с изучением геномов, для

определения корреляции между генотипом и фенотипом (Рис. 35). Уменьшение стоимости и времени для получения данных о последовательности ДНК благодаря NGS привело к целому ряду новых методов, таких как полногеномное секвенирование (DNA-seq), РНК-секвенирование (RNA-seq) для профайлинга клеточного транскриптома, рибосомный профайлинг (Ribo-seq) для определения эффективности трансляции всех клеточных мРНК, ChIP-секвенирование (ChIP-seq) для определения сайтов связывания ДНК-ассоциированных белков [155].



Рисунок 35. Трёхуровневые вклады NGS в картирование связей генотип-фенотип. Рисунок создан в программе Inkscape на базе источника [155].

В настоящее время существуют различные подходы к определению последовательности нуклеиновых кислот, отличающихся от классического секвенирования по Сенгеру возможностью одновременного «прочтения» многих участков генома, что позволяет сэкономить время и уменьшить затраты. Рассмотрим несколько таких подходов.

Первый метод – пиросеквенирование (454 Life Science) – заключается в амплификации с помощью ПЦР с иммобилизованным на микросфере праймером одноцепочечной матрицы ДНК, прикрепленной за счёт линкера на микросферу, находящуюся в капле воды в масляной эмульсии. Затем каждая такая микросфера помещается в «микрочип» и добавляются необходимые ферменты для секвенирования: ДНК-полимераза, при синтезе комплементарной цепи ДНК отщепляющая от dNTP пирофосфат, который использует АТФ-сульфурилаза для синтеза АТР, который в свою очередь используется люциферазой для окисления субстрата с последующим испусканием кванта света, таким образом, число квантов света пропорционально количеству использованных молекул АТФ. Возможное применение: геномное секвенирование *de novo*, RNA-seq, re-секвенирование [155].

Следующий метод – Illumina/Solexa – использует подложку с «пришитыми» олигонуклеотидами, с которыми гибридизуются необходимые для секвенирования последовательности ДНК небольшой длины. Синтез комплементарной цепи происходит при наличии ДНК-полимеразы с добавлением dNTPs, имеющих свои для каждого типа флуоресцентные метки, различные сигналы от которых фиксируются камерой, затем по ним определяется последовательность. Известно четыре метода создания библиотек 5'-UTR РНК с применением секвенирования с помощью Illumina [161]. Области использования: геномное секвенирование *de novo*, RNA-seq, ре-секвенирование, метагеномика, ChIP-seq [155].

Третий метод – Ion Torrent Sequencing – придуман для секвенирования большого числа образцов на основе детектирования полученных протонов при синтезе цепи ДНК, приводящих к изменению рН раствора в каждой отдельной реакции, с помощью ион-чувствительных микросенсоров [162]. Отличительной особенностью данного способа от других технологий является то, что в нём не используются оптические методы детекции и модифицированные dNTPs. Метод Ion Torrent применим к исследованию транскриптомов, секвенированию библиотек фрагментов ДНК, проведению ChIP-seq.

Основные сравнительные характеристики включают в себя следующее: принцип работы, длину единичного «прочтения», количество «прочтений» за цикл, время работы цикла и недостатки, и для трёх рассмотренных методов NGS представлены в Таблице 2 [155, 162-163]. Каждый из них имеет свои преимущества и недостатки, которые стоит учитывать при выборе того или иного вида секвенирования для решения определённой задачи исследования.

Таблица 2. Основные характеристики методов NGS.

Метод	454 Life Sciences	Illumina-SOLEXA	Ion Torrent
Принцип работы	Пиросеквенирование	Секвенирование путём синтеза комплементарной цепи	Ионный полупроводник
Длина одного «прочтения» (в парах оснований)	400-700	$75/2 * 100^a$ ("a" зависит от платформы)	200
Количество «прочтений» за цикл	$10^6$	$10^7-10^9$	$10^5-10^6$
Время работы одного цикла	7-10 часов	3-10 дней	2-3 часа
Недостатки	Ошибки секвенирования, асинхронный синтез, стоимость	Ошибки секвенирования, стоимость	Ошибки секвенирования

### 2.2.8. Рибосомный профайлинг (Ribo-seq)

Быстрое развитие методов секвенирования нового поколения (NGS) и накопленные навыки работы с системой трансляции позволили разработать такой метод, как рибосомное профилирование/рибосомный профайлинг (Ribo-seq). Он основан на глубоком секвенировании фрагментов мРНК, которые защищены рибосомой во время трансляции (Рис. 36) [164]. Данный подход оказался эффективным для изучения экспрессии генов на уровне транскрипции и трансляции одновременно, в том числе в ответ на различные воздействия [164-167]. Ribo-seq позволяет получить информацию о локализации рибосом на мРНК с однонуклеотидным разрешением. Такая точность даёт возможность выявлять трансляцию участков мРНК вне аннотированной рамки считывания, а также трансляцию перекрывающихся рамок считывания и смысловое декодирование стоп-кодонов. С применением рибосомного профайлинга удалось выявить транслируемые рамки считывания в РНК, ранее считавшихся некодирующими. В то же время появилась возможность для оценки влияния разных факторов и изменений условий на трансляцию мРНК в клетках, к примеру, разных используемых сред, изучаемых антибиотиков или модификаций исследуемых белков [168-173].

Активное использование рибосомного профилирования дало возможность выявить ряд артефактов и сложностей, связанных с обработкой результатов и методологией эксперимента [174-176]. Для многообещающей технологии Ribo-seq, используемой для изучения скорости декодирования рибосом, характерно наличие редких высоких пиков плотности рибосомного футпринта (защищённого рибосомой фрагмента РНК) и длинных интервалов при выравнивании соответствующих последовательностей мРНК. С целью уменьшить влияние неоднородности данных разработали способ нормализации, который эффективен в случае наличия гетерогенного шума. Представленный способ позволил выявить значимые различия в распределении по мРНК ридов («прочитанных» последовательностей) и контекстных детерминантах частот рибосомного футпринта в 30 общедоступных наборах данных Ribo-seq, таким образом вызывая сомнение в надёжности представленного подхода как точного определения локального распределения плотности рибосом, не имея предварительного контроля качества [174]. Данное наблюдение обращает внимание на далеко не полное представление о том, как параметры используемого протокола эксперимента оказывают влияние на плотность распределения футпринтов.

Наиболее известная причина представленного наблюдения возможно кроется в неравномерной эффективности образования библиотек для секвенирования в зависимости от последовательностей мРНК, которая возникает при создании библиотеки рибосомных футпринтов и её переводе в кДНК, и последующем секвенировании [175]. В перечисленных стадиях проводится ряд реакций, вовлекающих ферменты, которые обладают сиквенс-

специфичностью, к примеру, нуклеазы [177]. Также для обработки рибосом перед профайлингом могут использоваться и антибиотики, тоже обладающие специфичностью к последовательности [178-180]. Все выше указанные особенности необходимо принимать во внимание во время постановки эксперимента.

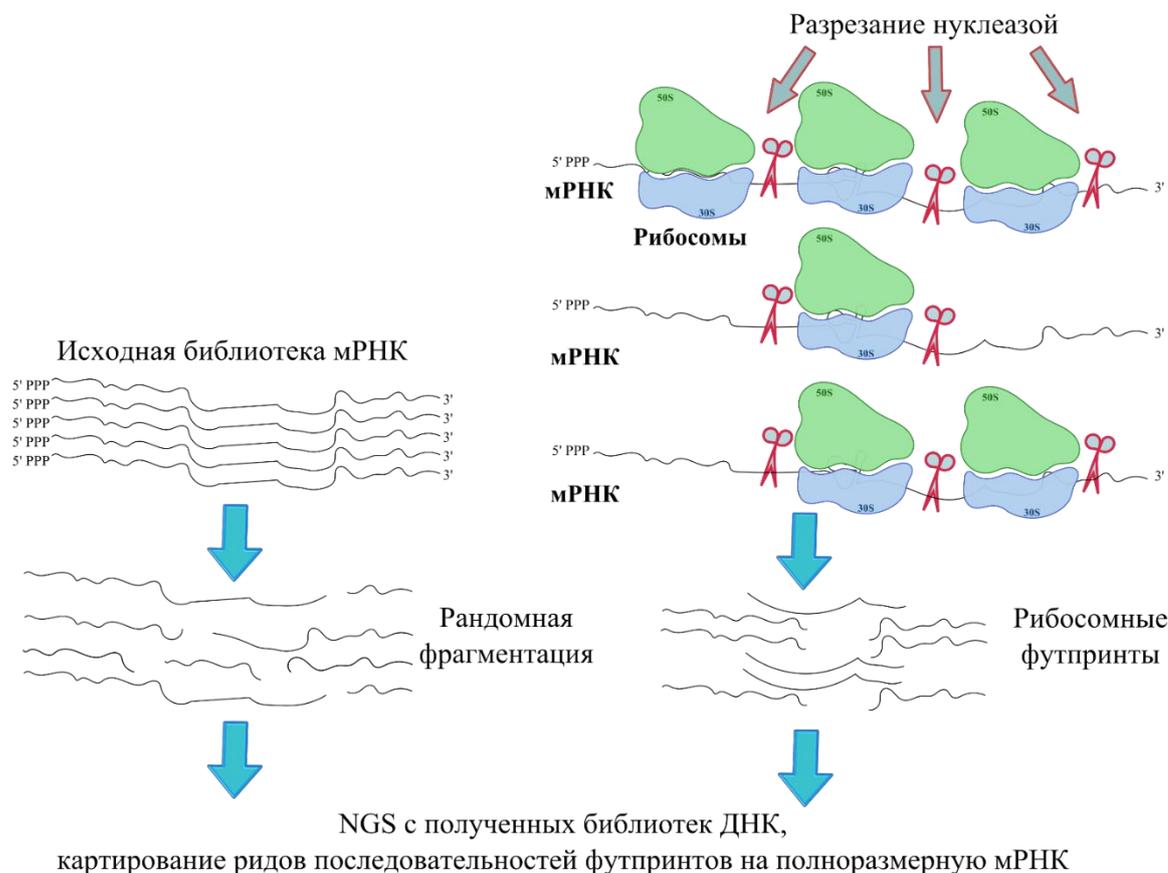


Рисунок 36. Схема метода рибосомного профилирования (Ribo-seq). После инициации трансляции проводят разрезание мРНК в местах, не защищаемых рибосомами, специфичной нуклеазой. В параллель путём рандомной фрагментации готовят библиотеку исходной мРНК для NGS. Она будет использоваться как референсная последовательность. Все выделенные рибосомные футпринты отправляют на приготовление библиотеки ДНК для последующего глубокого секвенирования. По итогу NGS делают картирование ридов последовательностей футпринтов на полноразмерную мРНК. Рисунок выполнен в программе Inkscape.

Благодаря применению метода Ribo-seq на бактериях продемонстрировано значимое увеличение покрытия рибосомами участков с последовательностями SD, которые случайно встречаются в кодирующей области [181]. Наряду с этим SD, предшествующие старт-кодону, имеют хорошо известную функцию в инициации трансляции. Полученные данные говорят о том, что стадия элонгации замедляется краткосрочным образованием дуплекса между SD-мотивами внутри ORF и последовательностью анти-SD в 16S рРНК, при этом подобные паузы определяли

по всему геному более 70% сильных остановок. Представленное наблюдение свидетельствует о таких паузах, как об основной детерминанте остановок бактериальной трансляции в ORF [181].

Позже с применением более усовершенствованного варианта метода рибосомного профилирования при улучшенном разрешении обнаружили, что показанное ранее обогащение покрытия рибосомами SD-мотивов, по всей вероятности, объясняется наличием пауз на глициновых кодонах и отсутствием возможности выделения всей популяции фрагментов мРНК, защищённых рибосомой. Сделан вывод о том, что SD-мотивы, вероятно, не являются основной причиной большого количества остановок трансляции *in vivo* [182].

## 2.2.9. Термодинамическое моделирование

Биофизические модели предоставляют оценку эффективности взаимодействия биомолекул. Примером такой модели служит термодинамическая модель [149], одновременно оценивающая силу молекулярных взаимодействий 30S комплекса с мРНК-транскриптом, рассчитывающая свободную энергию Гиббса  $\Delta G_{tot}$  для каждого элемента в составе конкретной мРНК и предсказывающая скорость инициации трансляции ( $r$ ). Причём  $r$  связывают следующим образом с изменением свободной энергии Гиббса,  $\Delta G_{tot}$ :

$$r \propto \exp(-\beta\Delta G_{tot}) \quad (1)$$

где  $\beta$  – фактор Больцмана ( $\beta=1/(k_B T)$ ),  $k_B$  – константа Больцмана,  $T$  – температура).  $\Delta G_{tot}$ , в свою очередь, зависящая от последовательности мРНК вокруг стартового кодона, определяется как:

$$\Delta G_{tot} = \Delta G_{\text{мРНК:рРНК}} + \Delta G_{\text{start}} + \Delta G_{\text{spacing}} - \Delta G_{\text{standby}} - \Delta G_{\text{мРНК}},$$

где  $\Delta G_{\text{мРНК:рРНК}}$  – энергия, которая высвобождается при гибридизации последних 9-ти нуклеотидов (3'-AUUCCUCCA-5') 16S рРНК *E. coli* и участка мРНК ( $\Delta G_{\text{мРНК:рРНК}} < 0$ );  $\Delta G_{\text{start}}$  – свободная энергия при комплементарном взаимодействии стартового кодона и антикодонной петли (3'-UAC-5') тРНК;  $\Delta G_{\text{spacing}}$  – штраф за неоптимальное расстояние между сайтом связывания 16S рРНК и стартовым кодоном ( $\Delta G_{\text{spacing}} > 0$ ), при отклонении от оптимума – 5 нуклеотидов ( $\sim 17 \text{ \AA}$ ) [138] – 30S комплекс изменяется так, что это приводит к снижению скорости инициации;  $\Delta G_{\text{мРНК}}$  – работа, необходимая для разворачивания взаимодействующего с рибосомой участка мРНК, когда он сложен в самую стабильную вторичную структуру, соответствующую минимуму свободной энергии ( $\Delta G_{\text{мРНК}} < 0$ );  $\Delta G_{\text{standby}}$  – работа, необходимая для разворачивания любой вторичной структуры standby-сайта, предшествующего участку мРНК, комплементарному последним 9-ти нуклеотидам 16S рРНК ( $\Delta G_{\text{standby}} < 0$ ). Структура standby-сайта может изолировать «резервный сайт» после сборки 30S комплекса, «резервный сайт» – 4 нуклеотида перед сайтом связывания 16S рРНК, которые могут быть важны для

правильной посадки рибосомы на RBS [138].  $\Delta G_{tot}$  уменьшается при образовании комплементарных пар и увеличивается, когда присутствуют взаимоисключающие вторичные структуры;  $\Delta G_{исх} = 0$  для полностью развёрнутой последовательности мРНК вокруг старт-кодона (Рис. 37). Таким образом, получается, что чем большую энергию надо затратить для разворачивания элементов мРНК, тем ниже скорость инициации трансляции.

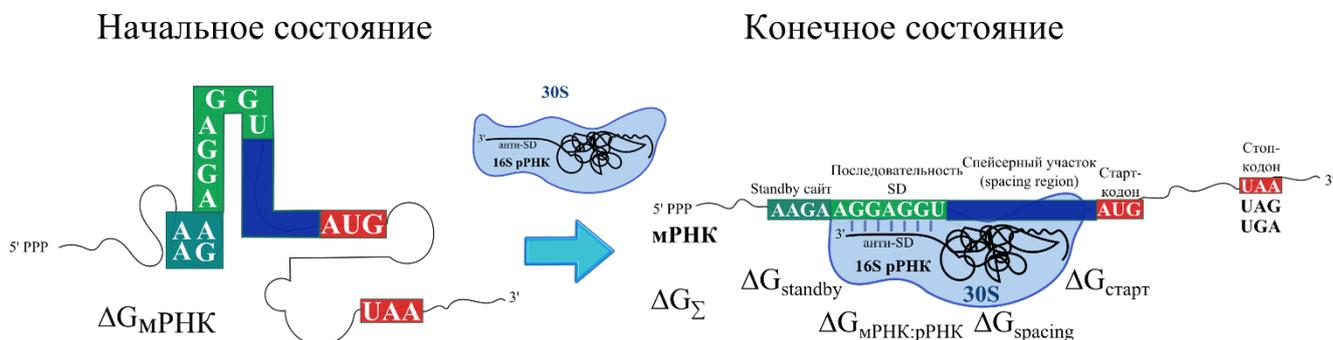


Рисунок 37. Термодинамическая модель инициации трансляции у бактерий. Изменение свободной энергии во время стадии инициации определяется пятью типами молекулярных взаимодействий, устанавливающих начальное и конечное положения системы. Рисунок создан в программе Inkscare.

Представленную модель можно использовать в двух случаях [149]. Во-первых, для предсказания относительной скорости инициации трансляции существующего сайта посадки рибосомы в 5'-UTR (т. н. «обратный инжиниринг»). Во-вторых, для дизайна последовательности RBS, обеспечивающей требуемую скорость инициации трансляции (т. н. «прямой инжиниринг»). Для генерируемой последовательности *de novo* путём произвольных вставок, удалений или замен нуклеотидов в соответствии с заданными требованиями производится подсчёт  $\Delta G_{tot}$  на каждом цикле до тех пор, пока её  $\Delta G_{tot}$  не окажется в пределах 0,25 ккал/моль от целевого значения. Дальнейшую проверку предсказаний можно провести *in vivo* путём измерений флуоресценции белков, полученных в ходе трансляции последовательностей, содержащих тестируемый вариант RBS.

## 2.2.10. Flow-seq метод

### 2.2.10.1. История появления и развития метода Flow-seq

Метод Flow-seq сочетает в себе проточную цитометрию, при которой можно измерить светорассеяние и интенсивность флуоресценции клеток с помощью луча лазера и разделить схожие по характеристикам клетки на отдельные фракции, и высокопроизводительное секвенирование выделенных библиотек, например, плазмидных ДНК, из клеток полученных фракций.

Первое упоминание о таком методе, как проточная цитометрия, датируется 1976 годом, когда на конференции во Флориде спустя 8 лет после представления первого проточного цитометра, основанного на флуоресценции, (в 1968 году) [183] было принято решение о таком названии этого метода. Несмотря на оригинальное название данной технологии «импульсная цитофотометрия», основанное на первой патентной заявке на неё, в научном сообществе быстро стало популярным именно название «проточная цитометрия». Первое же устройство проточной цитометрии на основе электрического сопротивления раскрыто в патенте ещё в 1953 году [184-185]. Предшествовавшие самой разработке исследования в этой области шаг за шагом накапливали знания и умения, приведшие в конечном итоге к изобретению проточного цитометра [186-189].

Постепенно проточная цитометрия (Flow cytometry/FCM) стала мощным методом анализа множества индивидуальных химических и физических параметров гетерогенных популяций клеток или частиц. Десятки тысяч клеток в секунду проходят одна за другой через один или несколько лазерных лучей в проточном цитометре, который измеряет рассеянный свет под разными углами и эмиссию флуоресценции (Рис. 38) [190].

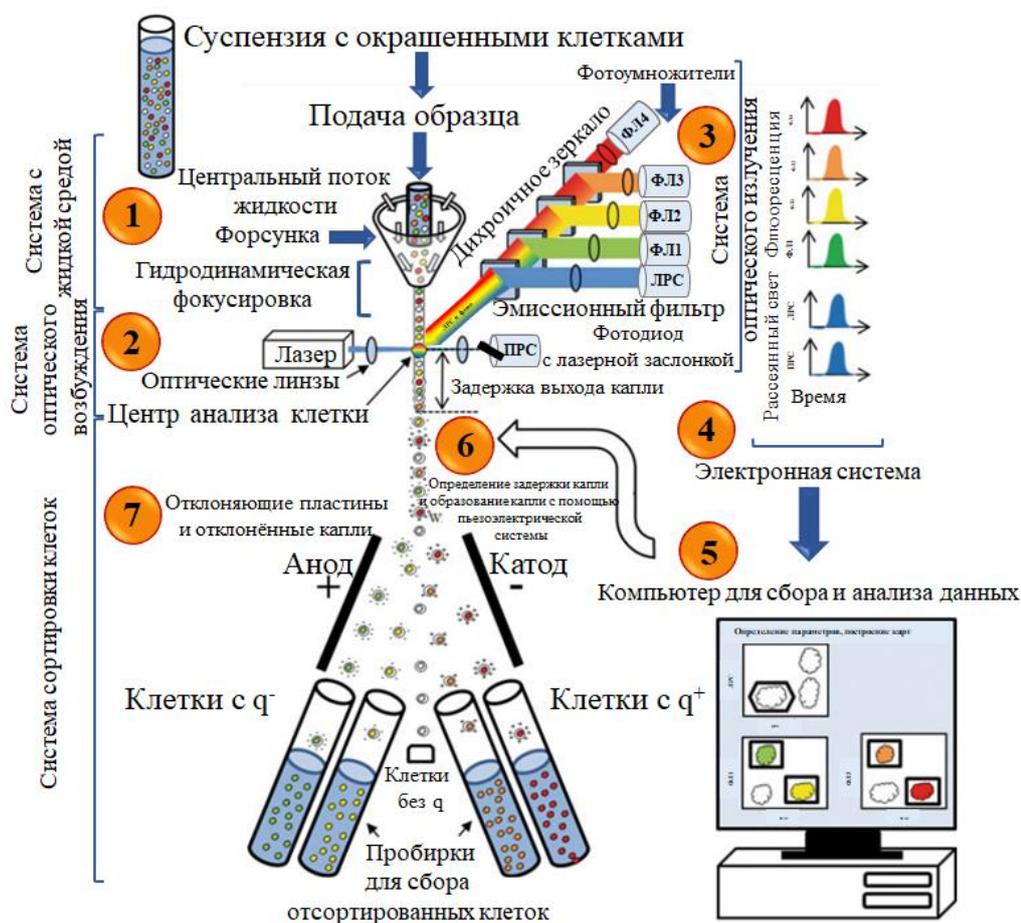


Рисунок 38. Основной принцип работы проточной флуориметрии. ЛРС, латеральный/боковой рассеивающий свет. ПРС, прямой рассеивающий свет. Адаптировано на основе источника [190].

Три части составляют данный прибор: первая представлена системой с жидкой средой, позволяющей осуществлять гидродинамическую фокусировку; вторая состоит из источника возбуждения и системы оптического излучения от фильтров по длинам волн к детекторам, образующих оптическую часть; и последняя – электронная система оцифровывает сигнал, чтобы впоследствии его можно было проанализировать с помощью специального программного обеспечения на компьютере [191]. Некоторые системы оснащены температурным контролем, другие предлагают возможность добавления реагентов или лекарственных средств при анализе.

Для обеспечения хорошего качества гидродинамической фокусировки жидкостная среда должна быть очень стабильной. Образец в виде суспензии сначала сжимают под давлением, а затем вводят в центр в окружение латерального потока жидкости перед прохождением через небольшое отверстие на выходе из форсунки (Рис. 38). Более детально происходит следующее: быстрое перемещение клеточной суспензии при подаче воздуха и под его давлением в центр, вокруг которого подается большой поток струящейся жидкости, проходящей через проточную ячейку. Тем самым создаётся своего рода оболочка вокруг суспензии и удерживаются частицы образца в небольшом пространстве в центральной части потока, проходящего через фокус таким образом, чтобы клетки проходили через лазерный луч одна за другой, обеспечивая так называемую гидродинамическую фокусировку [188]. Она формируется за счёт разницы в давлении между внешней окружающей струей и пробой, причём у первой оно больше.

Скорость прохождения клеточной суспензии зависит от давления окружающей жидкости, фиксированное для анализатора и регулируемое для клеточного сортера, способного разделять клетки по фракциям по набору определённых параметров. Чем выше давление в образце, тем больше клеток имеют возможность двигаться латерально в потоке, что приводит к снижению точности гидродинамической фокусировки и, как следствие, падению качества анализа образца в связи с увеличением коэффициента вариации пиковых значений. Дизайн проходного отверстия существенно важен для получения так называемых ламинарных потоков для перемещения клеток к центру основного потока: оно должно быть узким и идеально откалибровано (в зависимости от применения в диапазоне 50-2000 микрон) для обеспечения точности регистрации сигналов [190]. Инструмент позволяет детектировать объекты порядка 0,5 микрометров в диаметре в образце [191]. Однако существуют и другие концепции фокусировки, такие как микрокапиллярная технология или акустическая фокусировка, позволяющая выравнивать клетки с помощью звуковых волн без повреждения самих клеток [192-193].

Оптическая система состоит из возбуждающей оптики (лазеров) и собирающей оптики (фотоумножителей и фотодиодов), генерирующих сигналы падающего и флуоресцентного света, используемые для анализа образца (Рис. 38) [194]. Под действием лазерного луча при прохождении через него клетки или частицы рассеивают свет под разными углами:

прямой/малоугловой до  $10^\circ$  (PPC/FSC), латеральный/боковой (LPC/SSC). Последний далее направляется к собирающим линзам, в конечном итоге образуя электрические сигналы о клетках и/или частицах на выходе фотодетектора [195]. Интенсивность рассеяния в зависимости от морфологии, ориентации в потоке клетки, от поляризации падающего излучения может отличаться, при этом регистрируются сигналы о форме, размере, гранулированности клеток, позволяя в дальнейшем различить разные их типы. Например, выделить из общей массы отдельно эритроциты, лимфоциты, нейтрофилы, моноциты.

Клетки могут быть мечены флуоресцентными маркерами, поэтому свет поглощается, а затем излучается в определённом диапазоне длин волн [196]. Данные об интенсивности детектируемой флуоресценции по нескольким каналам со своими эмиссионными фильтрами также фиксируются и подаются в электронную систему (Рис. 38) [191]. Интенсивность флуоресценции определяется двумя вкладками. Первый из которых связан с излучением молекул флуорохрома, который специфически взаимодействует с конкретными клеточными компонентами: ДНК, РНК, белками, рецепторами. А второй – с излучением собственных молекул клетки: нуклеиновых кислот, белков. При этом интенсивность будет зависеть от оптических свойств флуоресцирующих объектов: их количества, длины волны и спектра испускания и поглощения, квантового выхода, числовой апертуры оптической системы для сбора флуоресценции и других особенностей [197]. Далее накопленные результаты анализируются и визуализируются на компьютере в виде одно-, чаще двумерных диаграмм, после чего принимается решение об их сортировке и сборе определённых клеток, образующих отдельную субпопуляцию или группу со схожими параметрами.

В системе сортировки клеток имеются пьезокристалл, разбивающий струю с последовательно идущими клетками на капли, система подачи заряда на капли и отклоняющие пластины с разными зарядами, что в свою очередь позволяет клеткам, несущим заряд, отклоняться при прохождении между катодом и анодом и попадать в одну из нескольких собирающих пробирок (Рис. 38). Таким образом, удаётся разделить и отсортировать клетки с разными параметрами в разные группы/фракции [198-199]. Существует два типа клеточных сортеров: кварцевая кювета и так называемая «струя в воздухе», отличающиеся тем, где располагается центр анализа клеток под действием лазера. Первый тип имеет фиксированное лазерное выравнивание, и их легче приготовить для сортировки. Второй требует ежедневного выравнивания лазеров, и их сложнее настроить, однако они лучше подходят для обнаружения частиц малого размера [194].

Есть и другие вариации наряду со стандартной системой детекции, использующей фотоумножители для проточной цитометрии, обладающей высокой чувствительностью и низким уровнем фона, что делает её полезной в сфере флуоресцентной технологии. Одним из последних

является визуализирующий проточный цитометр, он комбинируют традиционную проточную цитометрию с флуоресцентной микроскопией. Ещё можно выделить масс-цитометры, комбинирующие масс-спектрометрию с детекцией времени пролёта частиц и проточную цитометрию, при этом клетки метятся антителами с ионами тяжёлых металлов вместо флуоресцентно-меченных антител. Также существуют и цитометры для анализа «на специальных шарах», ставшие популярными для больших количеств аналитов в небольших объёмах образца. Кратко, при этом используют шары с известной интенсивностью флуоресценции в определённом канале и репортёрной молекулой, обнаруженной отдельным лазером для количественного подсчёта захваченного аналита конкретным шаром. По сути, это эквивалентно 100 анализам с помощью метода ELISA, использующего связанный с антителом фермент для детекции и иммуносорбент для захвата целевого антигена. В то же время есть и спектральные анализаторы, направленные на решение одной из проблем многопараметрической проточной цитометрии, заключающейся в компенсации или стирании перекрытия спектров между флуорохромами [199]. Такие анализаторы измеряют полный спектр флуоресцентного излучения для каждого флуорохрома отдельно в многоцветном образце для создания спектрального отпечатка («отпечатка пальца») [197]. Затем при анализе каждый спектр остаётся несмешанным, чтобы обеспечить чистый сигнал для каждого флуорохрома. Такой метод обнаружения начинает замещать традиционные фотоумножители при крупномасштабной проточной цитометрии [194].

Проточная цитометрия нашла большое применение в фундаментальных исследованиях, клинических испытаниях и практике [197, 200-204]. Её используют для биологических и функциональных диагностик, а также в исследовательских приложениях, включая сортировку и подсчёт клеток, иммунофенотипирование, многопараметрический анализ ДНК, анализ пролиферации клеток, различные методы с использованием флуоресцентных белков [197]. Широкое применение данный метод особенно нашёл в медицинских исследованиях и практике в качестве лабораторной тест-системы для оценки состава костного мозга, периферической крови, ликвора, суставной, плевральной и асцитической жидкостей в организме, с его помощью также возможен анализ суспендированных клеток тканей, к примеру, для выявления опухолевых.

Особое развитие метод проточной цитометрии получил в сочетании с NGS. Для данного комплексного метода в литературе встречаются следующие названия: Flow-seq/Flowseq или FACS-seq, или Sort-seq. Первые статьи с их упоминанием по базе данных PubMed датируются 2013 годом, и с этого момента насчитывается около 50 таких статей, 5 из которых представлены нашей научной группой. Однако сам подход, совмещающий флуоресцентную проточную цитометрию и секвенирование нового поколения, применялся и раньше. В PubMed сочетания “flow cytometry” и “next-generation sequencing” по отдельности находятся в статьях, начиная с 1974 и 1994 годов, и насчитывают уже около 249 и 87 тысяч публикаций соответственно.

### 2.2.10.2. Применение метода Flow-seq для оценки эффективности трансляции

Метод Flow-seq нашёл активное применение при оценке эффективности биосинтеза белка для множества вариантов мРНК одновременно. Для анализа влияния определённых элементов мРНК или их набора на эффективность экспрессии конкретного гена часто используют библиотеку из тысяч репортёрных конструкций, позволяющих одновременно или по отдельности перебрать различные варианты таких элементов, как промоторы, 5'-UTR, или же отдельные участки в них: сайты посадки рибосомы (RBS), последовательности SD, предшествующие им области (standby-сайты) или следующие за ними спейсерные участки (spacer regions), а также «разгонные» участки в начале кодирующей области (ramp region) (Приложение 1). Такие репортёрные конструкции, как правило, включают гены двух флуоресцентных белков. Причём первый из них используется в качестве сенсорного – его экспрессия чувствительна к изменяющимся элементам. В то время как второй выступает в роли внутреннего контроля, при этом область, изменяемая для первого, у него остаётся консервативной, что даёт возможность нивелировать эффекты внутри клеток, которые возникают как ответ на биосинтез двух перечисленных посторонних для клетки белков. Далее полученными библиотеками конструкций трансформируют модельный бактериальный штамм, подходящий для клонирования. Затем проводят селекцию трансформированных клеток по наличию резистентности к определённому антибиотику, ген устойчивости к которому также включён в состав плазмиды. Зачастую из отобранных клеток выделяют библиотеку плазмид и повторно трансформируют в другой штамм, подходящий для последующих экспрессии и сортировки. Затем с использованием проточной цитометрии для популяции полученных клеток проводится оценка интенсивностей флуоресценции двух белков, и они сортируются на фракции приблизительно с одинаковым соотношением измеренных уровней флуоресценций этих белков. После наращивания собранных клеток, из них выделяют плазмиды, чью вариабельную область амплифицируют и используют для высокопроизводительного секвенирования с целью определения последовательностей ДНК/РНК, которые попали в отдельную фракцию и обеспечивают наблюдаемые уровни экспрессии репортёрных генов.

Данный метод использовали при создании библиотеки конструкций, в которых были представлены различные сочетания промоторов и сайтов связывания рибосомы одновременно (Рис. 39). При этом для каждого варианта конструкции, трансформированной в клетки, определяли эффективность и трансляции, и транскрипции путём измерения синтезированных клетками количеств РНК и зелёного флуоресцентного белка (GFP), перед геном которого меняли тип промотора и RBS, и соотношения их к количеству соответствующей ДНК. Одновременно в плазмиде содержался ген другого флуоресцентного белка mCherry, который использовали как

внутренний контроль с консервативными промотором и RBS [205]. В итоге получили 12653 плазмиды с различными комбинациями 114 промоторов и 111 вариаций RBS. Селекцию клеток, трансформированных библиотекой конструкций, проводили по наличию резистентности к антибиотику канамицину, ген устойчивости к которому включен в состав плазмиды. Для оценки уровней ДНК и РНК в стационарной фазе проводили глубокое секвенирование ДНК (DNA-seq) и РНК (RNA-seq), выделенных из клеток, выращенных до этой фазы. Уровни флуоресцентных белков GFP и mCherry оценивали путём сортировки клеток на отдельные фракции со сходным соотношением интенсивностей флуоресценции GFP/mCherry. Далее из собранных фракций экстрагировали плазмидную ДНК, проводили секвенирование нового поколения её вариативных участков. Причём на выделенную ДНК добавляли баркодную последовательность, специфичную для каждой отдельной группы, для удобства дальнейшего поиска и распределения по ранее заданным фракциям последовательностей при анализе ридов, полученных в результате NGS.

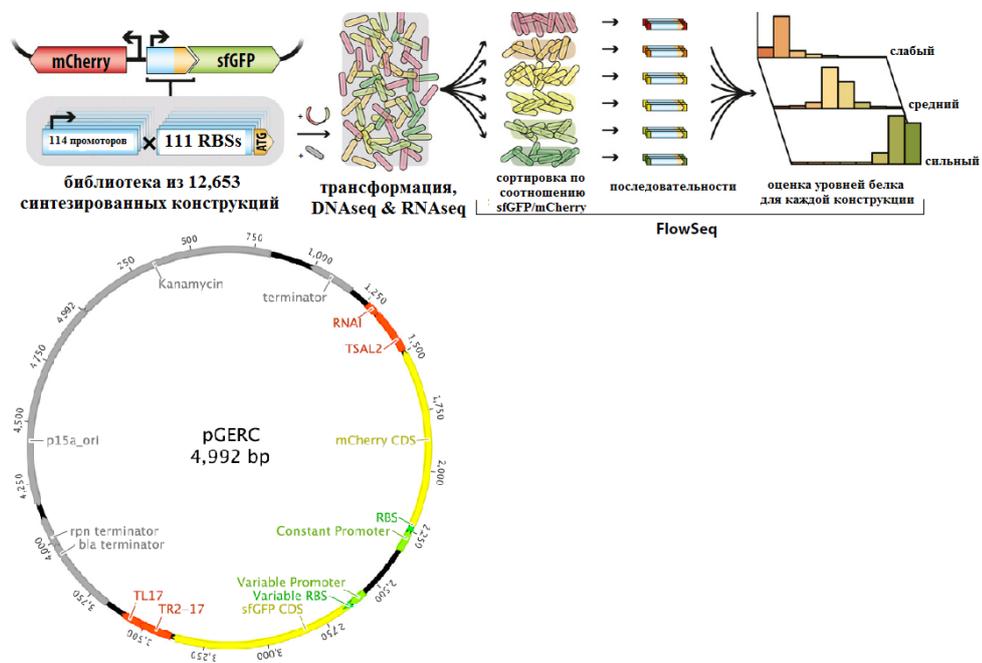


Рисунок 39. Схема работы с библиотекой, содержащей 12,653 конструкции при комбинации 114 промоторов и 111 вариантов RBS. Полученную библиотеку клонировали в плазмиду для экспрессии GFP, mCherry независимо экспрессировался с конститутивного промотора как внутриклеточный контроль. Для библиотеки клеток с набором конструкций проводили анализ с помощью DNA-seq, RNA-seq, Flow-seq для оценки уровней ДНК, РНК, белков для каждой индивидуальной плазмиды. Карта pGERC включает основу исходного вектора (показана серым) с геном устойчивости к канамицину, ориджином репликации и терминаторами транскрипции и изменённую часть с двумя генами флуоресцентных белков (показаны жёлтым), вариативными последовательностями промоторов и RBS (показаны зелёным), терминаторами транскрипции генов флуоресцентных белков (показаны красным). Адаптировано на основе источника [205].

После оценки уровней двух флуоресцентных белков и расчёта соотношения GFP/mCherry в качестве меры эффективности трансляции клетки разделили в зависимости от величины данного соотношения на три вида: сильные, средние и слабые, а затем соотносили с соответствующими им последовательностями. Как и ожидалось, уровень контрольного белка mCherry в клетках почти не отличался, а его интенсивность флуоресценции имела нормальное распределение и изменялась в пределах одного порядка. При этом уровень экспрессии белка GFP различался в пределах четырёх порядков. Благодаря секвенированию провели верификацию 282 отдельных колоний: 55% из них имели правильные константные области и ожидаемые варианты промоторов и сайтов посадки рибосомы без мутаций, для большего количества из них измерили уровни флуоресценции белков и далее использовали их в качестве контрольного набора.

На основе результатов NGS для ДНК и РНК и измерения соотношений уровней интенсивностей флуоресценции белков были построены репрезентативные карты, учитывающие уровень транскрипции и трансляции для каждой отдельной конструкции, содержащей определённую комбинацию промотора и сайта посадки рибосомы из всей библиотеки вариантов (Рис. 40).

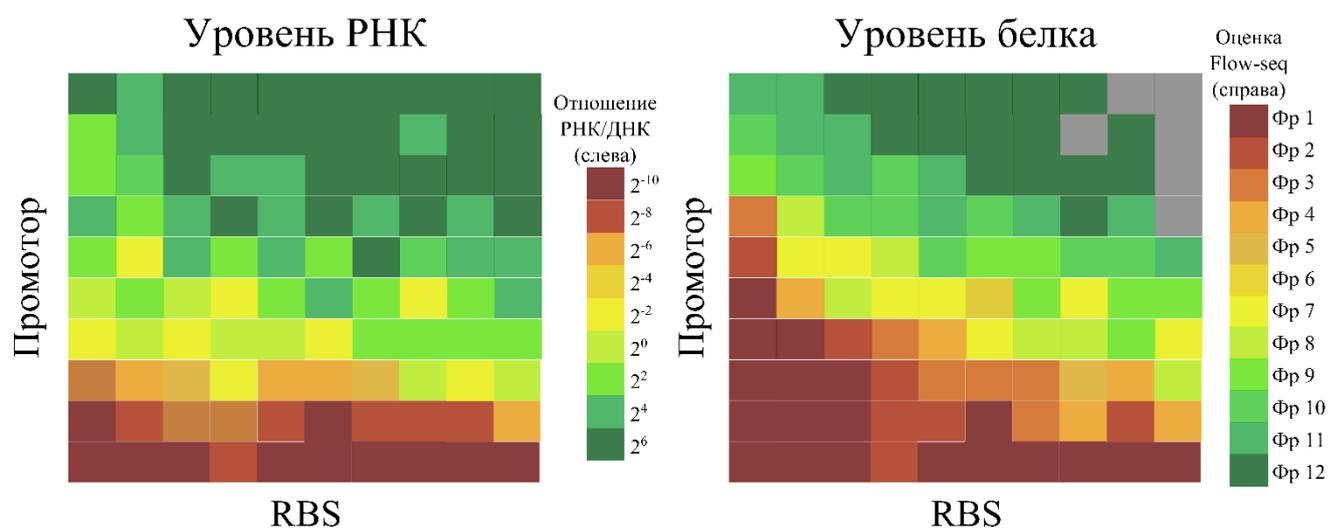


Рисунок 40. Схема репрезентативных карт с уровнями эффективности биосинтеза РНК и белка. Уровни РНК (слева) и белка (справа) для нескольких конструкций расположены на сетке по идентичности промоторов (ось Y) и RBS (ось X). Промоторы и RBS представлены по возрастанию средних значений эффективности биосинтеза РНК и белка соответственно. Ячейки отмечены серым цветом для конструкций, соответствующих уровням, находящимся ниже порога, заданного эмпирически. Шкалы для уровня РНК (отношения РНК к ДНК) и уровня белка (отношения флуоресценции белков GFP (зелёного) к RFP (красному)) приведены справа от их карт соответственно. Рисунок создан на базе источника [205] в программе Inkscape.

Подробный дисперсионный анализ (ANOVA) [206] проведён для уровней РНК и белка, заданных независимо типом промотора и RBS. Также он позволил учесть эффекты, которые свидетельствуют о зависимости уровня РНК от скорости биосинтеза белка. Для того, чтобы визуализировать полученные результаты, использовали программы, написанные на языках Python [207] и R [208] и адаптированные для работы с большими массивами полученных данных. В результате применения дисперсионного анализа удалось объяснить различия в уровнях РНК для 92% случаев выбором типа промотора, для 4% выбором типа RBS, а для остальных 4% нельзя соотнести видимые эффекты с выбором только одного из вариативных параметров. Наблюдаемые отличия в уровне экспрессии белка GFP можно объяснить в 54% случаев выбором промотора, в 30% - выбором сайта посадки рибосомы, остальные случаи нельзя определить этими двумя изменяемыми элементами. В итоге получается, что выбор промотора в большей степени влияет на уровень РНК и в меньшей – на уровень белка, ввиду того, что на биосинтез белка также оказывает влияние выбор RBS наряду с возможными другими факторами [205].

Изучение влияния последовательностей 5'-UTR разной длины и их отдельных элементов на эффективность трансляции сенсорного флуоресцентного белка описано в нескольких работах, в которых использовали метод Flow-seq [209-213].

Даже незначительные изменения в составе регуляторных элементов 5'-UTR, в частности в последовательности SD, могут повлиять на уровень экспрессии гена, что сильно усложняет изучение данного влияния [151, 214]. Для создания моделей, способных надёжно предсказывать такие изменения, возможно использование зависимости эффективности биосинтеза белка от нуклеотидного состава 5'-UTR в многосоставной генной инженерии [211]. На основе результатов применения метода Flow-seq для оценки уровня экспрессии белка GFP в ответ на присутствие определённой SD создана база данных EMOPEC, дающая возможность проектировать уровень экспрессии генов при решении биоинженерных задач [215].

Похожий подход использовался для изучения влияния на эффективность трансляции редких кодонов, располагающихся в начале кодирующей области мРНК [216]. Согласно предположениям, редкие кодоны чаще встречаются на N-конце природных генов, особенно экспрессирующихся на высоком уровне, и, может быть, играют важную роль в обеспечении высокого уровня синтезируемого белка [181, 217-220]. Однако опираясь на другие данные, можно наблюдать, что главной движущей силой отбора служит минимальная вторичная структура начала кодирующей области мРНК, а не редкость кодонов, являющаяся лишь следствием этого отбора [139, 216, 220]. Дискуссия о причинах и следствиях появления группы редких кодонов в начале кодирующих участков генов и об их влиянии на эффективность биосинтеза белка ещё продолжается в научных публикациях. Среди возможных причин расхождения мнений могут выступать особенности получения результатов, служащих

основанием для этих мнений. Например, разные научные группы в своей работе опирались на природные [217-222] или синтетические [218, 223-228] последовательности, и в то же время использовали отличные подходы для анализа данных [217-228].

С целью определения причины увеличения частоты встречаемости в начале кодирующей области редких кодонов и возможной функциональной роли данного наблюдения создали библиотеку из 14232 вариантов последовательностей на основе набора олигонуклеотидов. Каждый вариант представлял собой комбинацию одного из двух промоторов (слабого и сильного), одного из четырёх RBS (природного, слабого, среднего или сильного) и одной из вариативных последовательностей начальных 13 кодонов из 137 генов *Escherichia coli*. Изменяющиеся регуляторные участки предшествовали варианту гена белка GFP, способному эффективно сворачиваться (sfGFP), в конструкциях, с которых конститутивно коэкспрессируется белок mCherry [216]. Благодаря применению методов DNA-seq, RNA-seq и Flow-seq для всей полученной библиотеки оценили уровни синтезированных ДНК, РНК и белков соответственно.

Согласно гипотезе о “codon ramp”, замедленная трансляция начальных кодонов на N-конце в кодирующей области далее в ходе синтеза белка способствует снижению частоты столкновений рибосом [217, 226-227]. Изменение вторичной структуры мРНК может объяснить некоторое увеличение в эффективности биосинтеза белка в случае наличия редких триплетов на N-конце кодирующей области, а не сам факт присутствия на нём таких кодонов [216]. Методом Ribo-seq получили профили распределения рибосом по мРНК и показали отсутствие корреляции между скоростью трансляции и концентрацией тРНК, по сути определяющей частоту использования кодонов. Возможно создание специфическими редкими триплетами внутренних мотивов, напоминающих последовательность SD и возможно, в свою очередь, оказывающих влияние на эффективность биосинтеза белка в клетках *E. coli* [181]. Слабая, но статистически значимая взаимосвязь обнаружена между изменением экспрессии и внутренними SD-подобными мотивами.

Следующий вывод был сделан при изучении влияния на эффективность трансляции синонимических мутаций, не приводящих к замене аминокислотного состава белка, но играющих роль в изменении уровня экспрессии гена [223]. Наличие редких кодонов в *E. coli*, зачастую содержащих А или Т в третьей позиции, более часто коррелирует с увеличением экспрессии, чем присутствие синонимичных триплетов, оканчивающиеся на G/C, тем самым указывая на взаимосвязь с вторичной структурой у мРНК [223]. Наряду с этим продемонстрирована корреляция между уменьшением содержания GC нуклеотидов и увеличением эффективности трансляции [216]. Предсказание вторичной структуры мРНК первых 120 нт всех транскриптов с применением специальной программы NUPACK [229] позволило обнаружить корреляцию между увеличением стабильности вторичной структуры

мРНК и снижением уровня экспрессии, объяснив тем самым большую вариацию, чем любое другое ранее оцененное изменение [216].

В более поздней работе представили анализ библиотеки из более 30-ти тысяч различных вариантов кодонов, располагающихся в позициях с 2 по 11 в начале кодирующей области флуоресцентного белка. Набор последовательностей создан путём рандомизации первых 30-ти нт, следующих за инициаторным AUG кодоном. В качестве внутреннего контроля использовали константный ген ещё одного флуоресцентного белка. Полученные конструкции исследовали с применением метода Flow-seq [145]. В результате удалось подтвердить то, что вторичная структура мРНК отрицательно влияет на эффективность трансляции. При этом для редких кодонов начала кодирующей области не наблюдалось положительной роли в экспрессии гена.

В ходе анализа найдены и ещё некоторые закономерности. Отметим положительное влияние на уровень экспрессии ряда кодонов, располагающихся в начале кодирующего участка, AUG, AGA, GUA, GCA, CAC, CGA, UAC, AAA, которые кодируют дополнительный метионин кроме стартового, аминокислоты Arg, Lys, His, имеющие положительный заряд, гидрофобные алифатические аминокислоты Ala, Val и ароматический тирозин. В то же время на экспрессию наблюдали отрицательное влияние таких кодонов, как CUC, CCC, CCG, CUG, GGA, GGG, GGC, GCC, которые кодируют гидрофобные алифатические аминокислоты и аминокислоты, обладающие большей или меньшей свободой по конформации, если сравнивать с другими аминокислотами Leu, Pro, Gly, Ala. Причём отмечалась зависимость влияния соответствующего кодона от удаления его от стартового: чем ближе он к нему, тем сильнее его влияние. Наличие дополнительных инициаторных кодонов, находящихся в рамке считывания, способствует трансляции. Если синтез аминокислот, присутствующих на N-конце кодирующей области белка, требует больших затрат ресурсов клетки, то в бедной среде белки с такими аминокислотами транслируются не эффективно.

Область применения метода Flow-seq не ограничена описанными случаями. Данный метод нашёл также применение и для оценки эффектов, оказывающих влияние на процессы гликолиза, при этом используются биосенсорные репортёрные конструкции в разных бактериальных штаммах, включая нокаутные [230]. Flow-seq применяли и для изучения вариантов последовательностей терминатора [231], и для определения генов, которые вовлечены в изменения определённого метаболического пути [232], и для изучения сплайсинга [233].

### **2.2.10.3. Вклад метода Flow-seq в синтетическую биологию**

Синтетическая биология представляет собой относительно недавно появившуюся область науки, связанную с созданием и дизайном сложных процессов живых систем, представленных в природных организмах [12, 234-235]. Возникновение и развитие данной дисциплины произошло

при сочетании различных технологий генной инженерии, применения рекомбинантных ДНК с моделированием *in silico* (Рис. 41). Синтетическая биология направлена на определение поведения организмов и отдельных процессов в них с целью дальнейшей модификации и объединения их в определённых комплексных задачах. Точный контроль над уровнями экспрессии белков, участвующих в отдельных процессах системы, является важным требованием для надёжного функционирования мультиплексных синтетических систем [12].

Для оценки эффективности трансляции на базе последовательности 5'-UTR мРНК были созданы три программы-калькуляторы, опирающиеся на предположение о том, что суммарная скорость биосинтеза белка пропорциональна скорости его инициации, и дающие близкие к реальным оценки количества синтезируемого белка.

Один из таких калькуляторов представлен RBS Calculator [150, 236], основанный на ранее описанной термодинамической модели и используемый для предсказательного дизайна инициации трансляции и биосинтеза белка в целом у бактерий [149-150]. Благодаря применению данного подхода удалось изменить в пределах 5-ти порядков уровень экспрессии белка [150-151].



Рисунок 41. Иерархия синтетической биологии, напоминающей компьютерную инженерию. Адаптировано на базе источника [234].

Следующим методом моделирования, предсказывающим эффективность биосинтеза белка по последовательности мРНК, содержащей определённую 5'-UTR, является UTR Library Designer/UTR Designer [237-238]. Он похож на предыдущий и использует термодинамический параметр, представленный разницей в свободной энергии Гиббса, рассчитанной до и после сборки 30S IC на мРНК, и учитывает аффинность связывания рибосомы и доступность рибосомы и мРНК. UTR Designer сходно с RBS Calculator может быть использован для двух вариантов дизайна: в случае прямого генерируются 5'-UTR с необходимым уровнем экспрессии для

заданной белковой последовательности, а в случае обратного – происходит предсказание уровня трансляции белка для определённой последовательности мРНК, включающей 5'-UTR и первые 35 нт белок-кодирующей области. Принцип работы UTR Designer при создании набора мРНК с вариативными 5'-UTR основан на генерировании последовательностей 5'-UTR путём поочерёдной рандомизации нуклеотидов, комбинаторного перебора модификаций конструкций с отбором тех, которые способны обеспечить необходимый уровень биосинтеза белка. Также есть возможность зафиксировать нужную константную область в 5'-UTR, при этом комбинаторный перебор будет происходить только для её окружения. Для валидации этого метода создали две библиотеки 5'-UTR с 16 последовательностями, обеспечивающими разные уровни экспрессии в пределах определённого диапазона, на основе флуоресцентного репортёра. Полученные результаты измерений *in vivo* хорошо совпадали с предсказаниями *in silico* [237].

Последняя программа, рассчитывающая свободные энергии подобно двум предыдущим и отличающаяся методом оценки скорости биосинтеза белка, представлена RBS Designer. Она основана на стационарной кинетической модели и рассчитывает вероятность взаимодействия определённой мРНК с рибосомой, оценивая вероятность доступности заданной области мРНК с RBS и аффинности связывания с рибосомой. Все программы продемонстрировали сходную точность предсказания [12].

Благодаря большой накопленной базе данных в результате NGS, анализу разнообразных библиотек конструкций и результатам применения других методов геной инженерии удалось построить несколько моделей предсказания. Примером может служить предсказание сайтов инициации биосинтеза белка, удобное для определения расположения белок-кодирующих областей генов во время аннотации геномов архей и бактерий [239], наряду с предсказанием геномных последовательностей, которые могут соответствовать функциональным мотивам РНК [240], или же уровня экспрессии генов для новых комбинаций генетических элементов [214].

Экспериментальная валидация эффективности биосинтеза белка, заданной той или иной областью связывания мРНК в модельной системе, не всегда может гарантировать сохранение той же эффективности трансляции при замене последовательности кодирующей области. Причиной расхождений может служить образование вторичных структур в случае комплементарного взаимодействия 5'-UTR и кодирующей области. Для обеспечения более предсказуемого уровня экспрессии любого из генов в гетерологичной системе создали библиотеку бицистронных конструкций, в которых стандартная короткая рамка считывания находилась перед кодирующей областью репортёра, чья эффективность трансляции определялась с помощью цитофлуориметра. Участок реинициации трансляции был рандомизирован и обеспечивал перекрытие рамок. В результате использование реинициации позволяет исключить зависимость эффективности трансляции от кодирующей области гена второго белка. При этом искусственный оперон

содержал *gfp* или *rfp* в качестве второго гена. Наблюдалась хорошая корреляция между полученными уровнями экспрессии разных перечисленных генов [214].

Использование методов цитофлуориметрии или Flow-seq позволяет надёжно и напрямую экспериментально определять эффективность трансляции для созданных экспрессионных конструкций в области синтетической биологии.

Рассмотренные методы имеют свои достоинства и недостатки и ограничены в применении своими особенностями. Комбинирование различных подходов даёт возможность расширить границы их отдельного применения и получить данные, проливающие свет на фундаментальные основы экспрессии генов и открывающие новые возможности в биотехнологии. Успешными примерами такой комбинации являются методы Ribo-seq и Flow-seq, сочетающие в себе, во-первых, рибосомное профилирование и проточную цитометрию соответственно, и, во-вторых, глубокое секвенирование выделенных библиотек ДНК. Дальнейший анализ полученных с их помощью данных значительно позволил расширить представления об особенностях мРНК, влияющих на эффективность трансляции, и по-новому взглянуть на исследуемый процесс в целом.

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Анализ различных методов изучения роли 5'-UTR в регуляции биосинтеза белка позволил выявить некоторые недостатки и учесть их при разработке многостороннего подхода для изучения влияния на эффективность трансляции одного из элементов 5'-UTR – спейсерного участка (спейсера) между последовательностью Шайна-Дальгарно (SD) и старт-кодоном – или длины и состава целой 5'-нетранслируемой области, представленной конкретными, природными или рандомизированными вариантами последовательностей.

### 1. Изучение влияния спейсерного участка в 5'-UTR на эффективность трансляции

При подготовке данного раздела диссертации использована следующая публикация, выполненная соискателем в соавторстве, в которой, согласно Положению о присуждении учёных степеней в МГУ, отражены основной результат, положения, выводы исследования: [18], общий объём статьи – 0,5 печатного листа (п.л.). Автором лично проведены все экспериментальные процедуры, причём FACS, NGS и биоинформатический анализ полученных данных проводились совместно с автором, при активном его участии.

#### 1.1. Создание репортёрных конструкций с 4N-фрагментом в спейсерном участке 5'-UTR между последовательностью SD и старт-кодоном гена *cer*

Для определения влияния последовательности спейсерного участка мРНК, находящегося между SD и инициаторным кодоном, на эффективность трансляции создали набор репортёрных конструкций с рандомизированной вставкой длиной 4 нуклеотида (4N) в спейсере по следующей схеме (Рис. 42), опираясь на опубликованную ранее методику [241].

В качестве исходных фрагментов для последующей вставки в вектор взяли синтетически синтезированные одноцепочечные олигонуклеотиды длиной 33 нт (оц олиги) с фиксированной последовательностью SD (5'-GGAG-3'), двумя палиндромными сайтами узнавания эндонуклеаз рестрикции SacII и NdeI и рандомизированным участком (4N):

5' – А С Т G **CCGCGG** С А С А С А С С GGAG С **NNNN** **CATATG** – 3',

которые вносят следующие разрывы: SacII: 5' – **CCGC** **GG** – 3'      NdeI: 5' – **CA** **TATG** – 3'

3' – **GG** **CGCC** – 5'      3' – **GTA** **TAC** – 5',

причём в сайте узнавания NdeI содержится стартовый кодон ATG. Дальнейшая подготовка проводилась путём отжига в течение 15' при 80°C этих праймеров друг на друга с образованием дуплексов, с последующей достройкой фрагментом Клёнова ДНК-полимеразы I *E. coli* при добавлении dNTP до двухцепочечных (дц олигов, 60 нт). На следующем этапе была проведена рестрикция сначала одной эндонуклеазой SacII, потом другой – NdeI по соответствующим

сайтам. Проверка эффективности работы с олигонуклеотидами на каждой стадии была проведена путём электрофореза нуклеиновых кислот в 15% ПААГ в денатурирующих условиях (Рис. 43). Полученный целевой дуплекс длиной ~ 25 нт очищался от низкомолекулярных примесей (солей, нуклеотидов) с помощью гель-фильтрации либо экстракции комбинацией растворителей: фенола и хлороформа, и последующего осаждения этанолом.

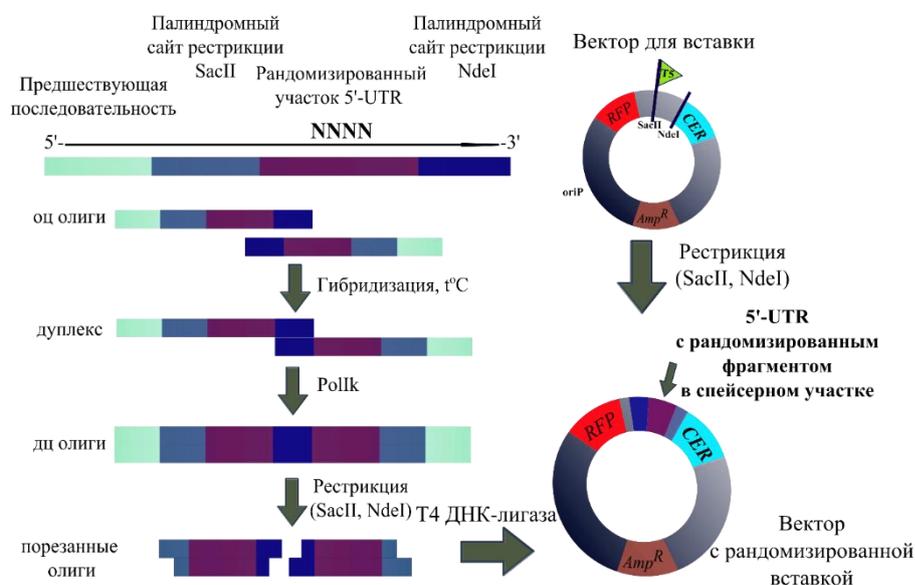


Рисунок 42. Схема создания библиотеки плазмид с фрагментом рандомизации (4N) в спейсерном участке 5'-UTR мРНК. Оц олиги – одноцепочечные олигонуклеотиды длиной 33 нт. Дуплекс – отождённые друг на друга одноцепочечные олигонуклеотиды. Дц олиги – дуплексы, достроенные до двухцепочечных фрагментом Клёнова ДНК-полимеразы I *E. coli* (Polk). Порезанные олиги – дц олиги после рестрикции эндонуклеазами SacII и NdeI, используемые как целевой дуплекс для дальнейшего встраивания в плазмиду, порезанную по тем же сайтам рестрикции. Amp<sup>R</sup> – белок устойчивости к ампициллину. oriP – ориджин репликации плазмиды. Рисунок выполнен в программе Inkscapе.

На следующем этапе проводилась подготовка вектора для вставки. Использованный вектор pRFP-CER включает двойную репортёрную конструкцию (как на Рис. 44, только без начала кодирующей области гена *cer* с сайтом рестрикции BamHI) [156], содержащую гены двух флуоресцентных белков RFP (красного) и CER (голубого). Ранее её брали в основу для изучения влияния длины SD, расстояния от SD до инициаторного кодона и их сочетаний на эффективность трансляции белка CER [137]. В данной конструкции 5'-UTR перед геном белка RFP не меняется, а синтез RFP служит внутренним контролем. Экспрессия CER отражает эффективность трансляции мРНК с варибельной 5'-UTR перед его геном. Значения длин волн в нанометрах (нм) на максимумах пиков возбуждения/испускания для белка CER – 430/486 нм, а для белка RFP – 531/595 нм. Одинаковые промоторы и терминаторы транскрипции были отдельными для каждого

гена флуоресцентного белка в конструкции, и их работа не зависела от процессов внутри клеток, а синтез самих белков происходил независимо друг от друга и от других клеточных белков.

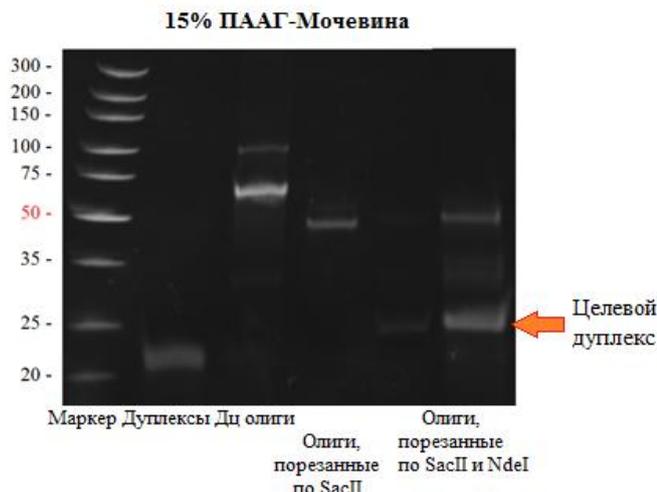


Рисунок 43. Результат электрофореза обработанных праймеров в 15% ПААГ с 7М мочевиной. Маркер – Gene Ruler™ Ultra Low Range DNA Ladder. Дуплексы – отождённые друг на друга одноцепочечные (33 нт) олигонуклеотиды. Дц олиги – достроенные до двухцепочечных дуплексы фрагментом Клёнова ДНК-полимеразы I *E. coli* (60 нт). Олиги, порезанные по *SacII* – дц олиги после рестрикции эндонуклеазой *SacII*. Олиги, порезанные по *SacII* и *NdeI* – дц олиги после рестрикции эндонуклеазами *SacII* и *NdeI*. Целевой дуплекс (~ 25 нт), использованный для дальнейшего встраивания в вектор, указан стрелкой.

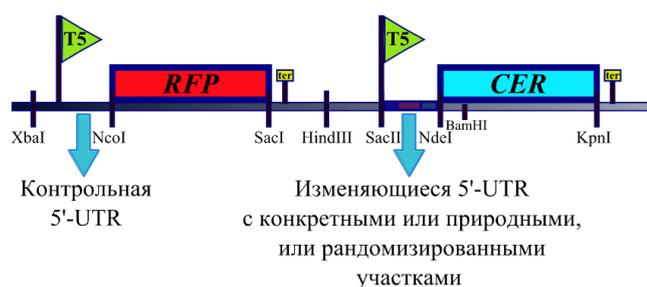


Рисунок 44. Схема двойной репортёрной конструкции с двумя генами флуоресцентных белков RFP и CER. T5 и ter – промоторы и терминаторы транскрипции соответственно, отдельные для каждого гена. 5'-UTR – 5'-нетранслируемая область, причём она константная перед геном красного флуоресцентного белка RFP, а перед голубым флуоресцентным белком CER она переменна, последняя может быть представлена разными вариантами последовательностей с конкретными, природными или рандомизированными участками для оценки влияния их длины и состава на эффективность трансляции. Расположения сайтов эндонуклеаз рестрикции, которые можно использовать для модификации плазмиды, также отмечены. Рисунок выполнен в программе Inkscaper.

Встраивание порезанных целевых дуплексов в вектор проводилось по двум известным сайтам эндонуклеаз рестрикции – SacII и NdeI, находящимся в 5'-UTR перед геном, кодирующим белок CER, и с последующей сшивкой T4 ДНК-лигазой. Перед добавлением смеси вставок порезанный вектор дефосфорилировали фосфатазой CIAP во избежание сшивки его с самим собой и образования «пустых колец» (без вставок) исходной плазмиды. В случае, если они всё-таки образовались, дополнительный контроль в виде рестрикции с помощью эндонуклеазы EcoRV по следующему сайту:

EcoRV: 5' – **GAT**ATC – 3'

3' – CTA**TAG** – 5',

исключал их при обработке ею лигазной смеси. Так как этот участок имеется в исходной плазмиде между участками узнавания эндонуклеаз SacII и NdeI, то соответственно порез по нему позволяет определить, есть ли вставка или нет (Рис. 45).

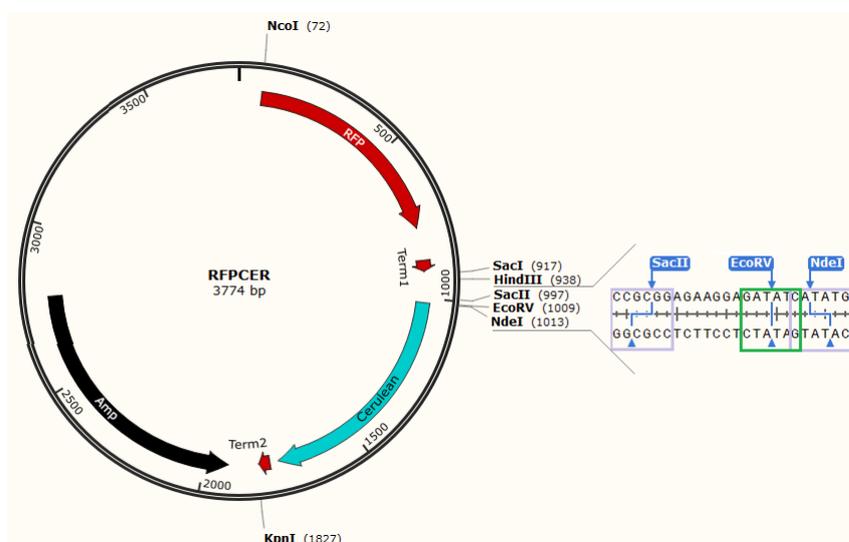


Рисунок 45. Схема плазмиды rRFPCER с выделенными использованными сайтами рестрикции эндонуклеаз SacII, EcoRV и NdeI. Участки узнавания выделены рамками, для EcoRV и NdeI они перекрываются. В векторе закодирован также белок устойчивости к ампициллину (Amp<sup>R</sup>), по которому в дальнейшем проводилась селекция. Схема создана в программе Clone Manager.

Таким образом, были получены плазмиды с различными вставками в спейсерном участке длиной 8 нт, расположенном между последовательностью SD и старт-кодоном в 5'-UTR длиной 22 нт перед геном *cer*. Данные вставки содержали рандомизированный фрагмент длиной 4 нт – 4N (Рис. 42), представленный всеми возможными  $4^4=256$  вариантами последовательностей.

Данный метод имеет значительное преимущество перед ПЦР-амплификацией. Так как при использовании в ПЦР Taq-полимеразы, очень процессивной и устойчивой к высокой температуре, но не имеющей корректирующей 3'-5' экзонуклеазной активности, происходит

накопление нежелательных мутаций в ДНК, к тому же возможно нарушение равновероятного соотношения всех различных вариантов вставок в ПЦР-смеси, даже если использовать другую полимеразу с такой активностью. Однако наличие такой активности у фрагмента Клёнова ДНК-полимеразы I *E. coli* могло привести к отрезанию лишнего, а именно концевой участка гибридизации, и нарушению образования двухцепочечных олигов. Поэтому были добавлены dNTP в небольшом избытке, которого бы хватило на достройку и корректировку. По результатам электрофореза видно, что произошло образование большей части двухцепочечных фрагментов (Рис. 43).

Наличие рандомизированного участка уже в составе вектора pRFPCER проверялось секвенированием по Сэнгеру (Рис. 46). Его результат свидетельствует о создании нужной библиотеки 5'-UTR с 4N-фрагментом в спейсерном участке, которой можно трансформировать клетки *E. coli*, и для которой можно проводить дальнейший анализ.

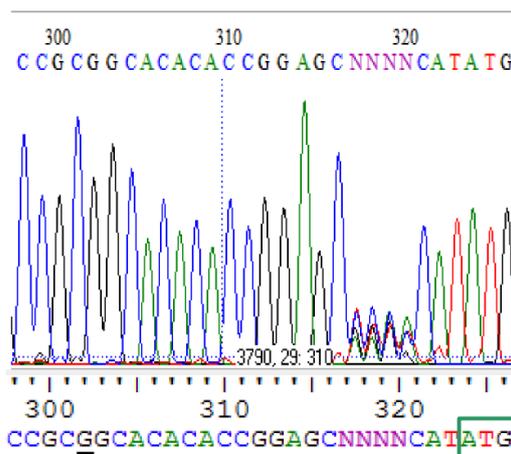


Рисунок 46. Результаты секвенирования плазмид с рандомизированной 4N-вставкой. Первый нуклеотид вставки, с которого пойдет транскрипция, подчеркнут. Старт-кодон гена белка CER для дальнейшей трансляции его мРНК выделен зелёной рамкой.

Следует отметить, что в данном эксперименте использовали перед репортёрным сенсорным геном *cer* 5'-UTR длиной 22 нт, близкой к наиболее часто встречающимся у мРНК *E. coli* (представлены в диапазоне длин 20-40 нт) [110, 157]. В то же время длина спейсерного участка в наших 5'-UTR составляет 8 нт и лежит в оптимальном диапазоне, определённом как 7-10 оснований, для мРНК *E. coli* [37]. Также в данном эксперименте в состав 5'-UTR входит последовательность SD длиной 4 нт, расположенная на расстоянии 8 нт от старт-кодона (при расчёте удаления центрального G в SD от IR получается 11 нт). При этом длина и расстояние от инициаторного кодона являются оптимальными согласно литературным данным [133, 136-138]: у разных генов *E. coli* длина SD варьирует в пределах 4-8 нуклеотидов и находится на расстоянии 5-13 нт [49], для оптимальной экспрессии генов оно составляет 8-11 нт от старт-кодона (при

расчёте удаления центрального G в SD от IR). Перед контрольным репортёрным геном *rfp* и перед геном *cer* в контрольной плазмиде находилась консервативная 5'-UTR длиной 17 нт, также близкой к самым часто встречающимся у мРНК *E.coli*, с оптимальными SD длиной 6 нт на расстоянии 9 нт от старт-кодона (при расчёте удаления центрального G в SD от IR) и спейсерным участком длиной 7 нт (Рис. 45). Таким образом, опираясь на литературные данные, была создана библиотека конструкций с учётом оптимальных характеристик основных элементов 5'-UTR мРНК для экспрессии репортёрных генов в клетках *E. coli*.

## 1.2. Клонирование полученной библиотеки плазмидных ДНК

На первом этапе химически и электрокомпетентные клетки штамма JM109 *E.coli* трансформировали контрольной плазмидой с одинаковыми консервативными 5'-UTR перед генами флуоресцентных белков RFP и CER. В случае электропорации трансформация прошла эффективнее (на 2 порядка больше выросших колоний), чем с помощью теплового шока (Рис. 47), поэтому для очень большой библиотеки плазмидных ДНК разумнее использовать именно первый метод. Например, при наличии участков рандомизации в 5'-UTR длиной до 6 нт (6N) получаем размер библиотеки со всеми возможными комбинациями 6 нт из четырёх типов нуклеотидов, насчитывающий  $4^6 = 4096$  различных вариантов последовательностей, в таком случае трансформация с помощью теплового шока будет оптимальной. Тогда как для библиотек плазмидных ДНК с участками рандомизации в 5'-UTR длиной 7 нт (7N,  $4^7 = 16384$  различных варианта последовательностей) и лучше использовать электропорацию. Для полученного набора плазмид с участком рандомизации 4 нт (4N), представленного всего  $4^4 = 256$  вариантами, достаточно было провести трансформацию с помощью теплового шока клеток штамма JM109 (для клонирования библиотеки плазмид) и штамма BW25113 (для сортировки клеток, трансформированных библиотекой плазмид, выделенной уже из клеток штамма JM109) *E.coli*.

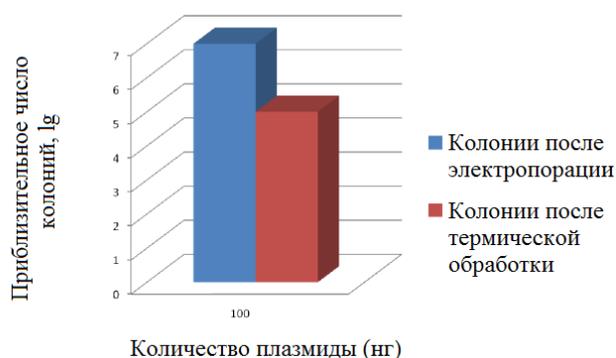


Рисунок 47. Результат проверочной трансформации компетентных клеток JM109 *E. coli* контрольной плазмидой с помощью теплового шока и электропорации. Рисунок выполнен в программе Microsoft Excel.

На следующей стадии к трансформированным клеткам добавляли жидкую среду SOC с глюкозой и перемещали в термостат на 37°C для наработки белка устойчивости к ампициллину, по которому после высевания клеток на большую чашку Петри проводилась селекция колоний. В параллель с помощью титрования, последовательного разведения клеточной культуры и высевания её на чашку Петри, проводили подсчёт колоний (Рис. 48).

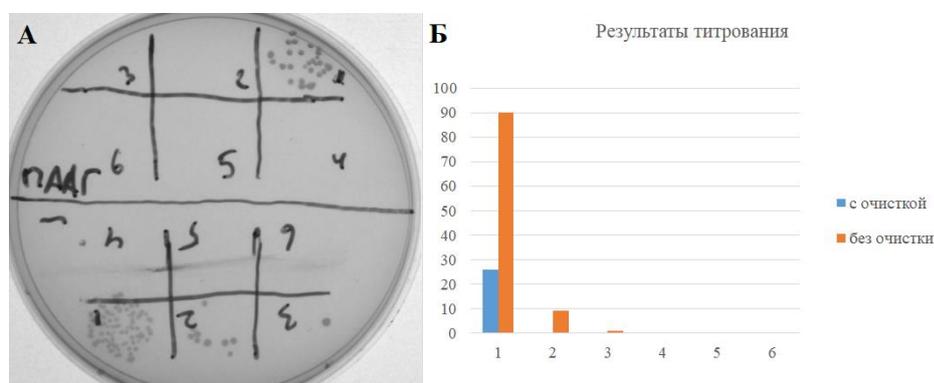


Рисунок 48. Результат титрования клеточной культуры после трансформации лигазной смесью, очищенной из геля после электрофореза в 15% ПААГ с 7М мочевиной (А, сверху) и без очистки через гель (А, снизу). Числа означают номер разведения в 10 раз, начиная с 10 мкл клеточной культуры для 1-го номера. Б. То же, что и в А, только в виде гистограммы, выполненной в программе Microsoft Excel.

В результате взяли необходимое количество клеточной культуры, чтобы в несколько раз покрыть разнообразие библиотеки, и наносили его на большую чашку Петри со средой LB с ампициллином. С этой чашки после инкубирования при 37°C в течение 16 ч смывали колонии раствором PBS и подготавливали клетки к дальнейшей сортировке. Параллельно контрольную плазмиду с идентичными 5'-UTR перед старт-кодонами обоих флуоресцентных белков и, как следствие, с приблизительно одинаковым соотношением белков CER/RFP, трансформировали в клетки JM109 и BW25113 *E.coli*, а затем готовили к сортировке.

### 1.3. Валидация метода Flow-seq для широкомасштабного анализа эффективности трансляции

Оценка эффективности трансляции проводилась по измерению *in vivo* соотношения интенсивностей флуоресценции двух белков: сенсорного CER, перед геном которого находилась переменная 5'-нетранслируемая область, и контрольного RFP с константной 5'-UTR. Такая оценка позволяет нивелировать эффекты, оказываемые экспрессией двух посторонних белков на жизнедеятельность клетки и обусловленные флуктуациями копияности репортёрной плазмиды и размером бактериальной клетки. Если на уровень экспрессии репортёрных генов будет влиять

какой-то неучтённый нами фактор (состав среды, интенсивность перемешивания, температура и т.д.), то он будет одинаково воздействовать на оба белка, и соотношение их флуоресценций будет зависеть только от влияния изучаемых нами особенностей мРНК, а не от внешних условий. В то же время флуоресцентные белки CER и RFP обладают легко различимыми спектральными свойствами, позволяющими одновременно их детектировать в одной бактериальной культуре как при измерении совокупности клеток с применением флуориметра, так и на уровне отдельных клеток с помощью проточной цитометрии [137, 145, 156].

Использование клеточного сортера позволяет в зависимости от эффективности трансляции – соотношения интенсивностей флуоресценции белков CER/RFP – с высокой точностью разделять клетки с различными плазмидами. Применение метода Flow-seq, сочетающего сортировку флуоресцентных клеток и высокопроизводительное секвенирование (NGS), для анализа комбинаторных репортёрных библиотек конструкций продемонстрировало превосходную производительность в ранее опубликованных исследованиях [205, 216, 242-244].

В первую очередь мы провели сортировку клеток с контрольной плазмидой pRFP-CER. Из миллиона клеток культуры сначала выделялись светящиеся, их число составляло 98,6% от общего числа просмотренных прибором клеток, и несветящиеся, составившие всего 1,4% (фракция P1). Затем происходила сортировка первых в зависимости от соотношения двух флуоресцентных белков. Получилось, что из светящихся клеток примерно 91% приходится на фракцию F5, которой соответствует примерно одинаковое соотношение интенсивностей флуоресценции белков RFP и CER, принятое в качестве контроля, на его значение далее будут нормированы измерения флуоресценции для других конструкций (Рис. 49). Важно, что клетки, трансформированные одной конструкцией, преимущественно располагаются в одной фракции.

С целью сравнить соотношения CER/RFP, измеренные с помощью спектрофлуориметра и клеточного сортера, и оценить эффективность сортировки сначала был проведён контрольный эксперимент (Рис. 50А, Б). Для этого смешали равное количество клеток, трансформированных репортёрными конструкциями, содержащими 8/7, 6/10, 4/7, 4/7AU, 2/7 и 0 (первое число – длина SD, второе – расстояние от SD до старт-кодона; расстояние рассчитывалось от большого G, начиная с a, к примеру, в последовательности aagGagg, и до первого нуклеотида в старт-кодоне) 5'-UTR перед геном, кодирующим белок CER. Конструкция 4/7AU отличается от варианта 4/7 наличием добавочного AU-богатого трансляционного энхансера, а вариант 0 вовсе не содержит SD. Для всех плазмид ранее были определены значения соотношений CER/RFP (Таблица 3) [137]. Затем клетки с этими конструкциями рассортировали на 8 фракций с различным соотношением CER/RFP (Рис. 50А). Общее число светящихся клеток в этот раз составило около 98,2% от общего числа просмотренных прибором клеток.

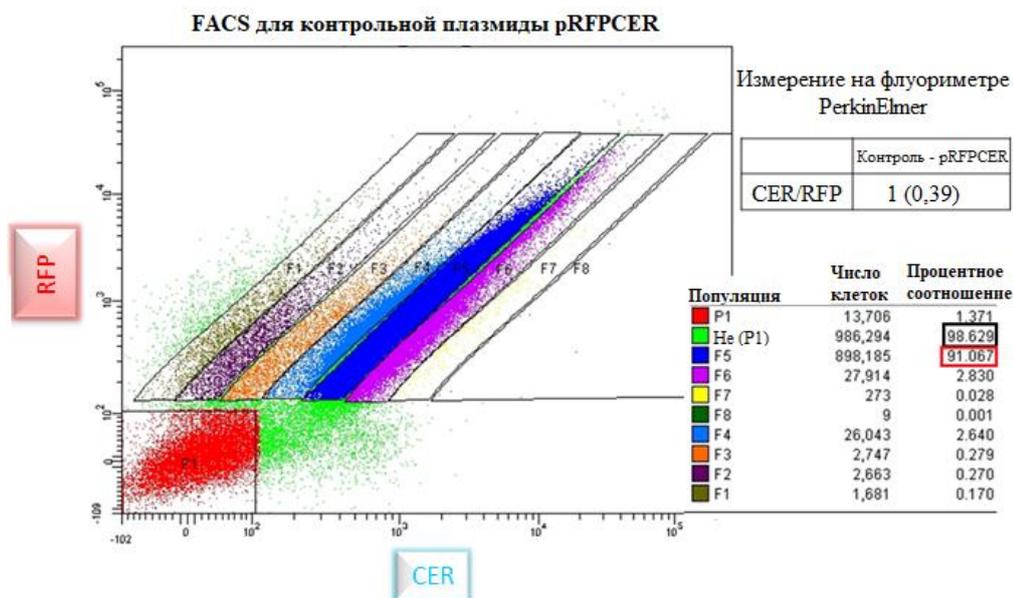


Рисунок 49. Распределение интенсивностей флуоресценции CER и RFP для совокупности клеток, трансформированных контрольной плазмидой pRFPCER, и числа клеток в каждой выделенной фракции (справа, чёрным прямоугольником выделен процент светящихся клеток, красным – процент клеток из фракции F5, соответствующей контролю), выданные программой для работы с клеточным сортером BD FACS Aria™ III. Каждая точка на графике соответствует отдельной клетке. Интенсивности флуоресценции CER и RFP увеличиваются вдоль осей X и Y соответственно. Каждая фракция клеток F1-F8, отличающаяся по соотношению CER/RFP, ограничена наклонной рамкой. Фракции F1-F8 отсортированы по увеличению эффективности трансляции. Справа вверху приведено значение соотношения CER/RFP (в скобках) контрольной плазмиды, использованное для нормировки остальных конструкций, принятое за «1».

После сортировки измерили общее соотношение флуоресценции CER/RFP во всех фракциях с помощью спектрофлуориметра (Рис. 50В) и определили пропорции клеток, которые содержат каждую репортёрную конструкцию, во всех фракциях (Рис. 50Б). Клетки, несущие плазмиды, которые кодируют ген белка CER, и показавшие более эффективную трансляцию этого белка согласно измерениям общей флуоресценции с помощью флуориметра, обнаружены во фракциях, демонстрирующих более высокое соотношение CER/RFP по данным клеточной сортировки. Большинство клеток, содержащих каждую контрольную репортёрную конструкцию, концентрируется преимущественно в одной или двух фракциях (Таблица 3).

Результаты контрольного эксперимента показали, что метод Flow-seq можно использовать для надёжной и точной оценки соотношения CER/RFP на уровне индивидуальной клетки, причём полученные с его помощью измерения совпадают с измерениями общей флуоресценции для клеточной культуры с применением спектрофлуориметра.

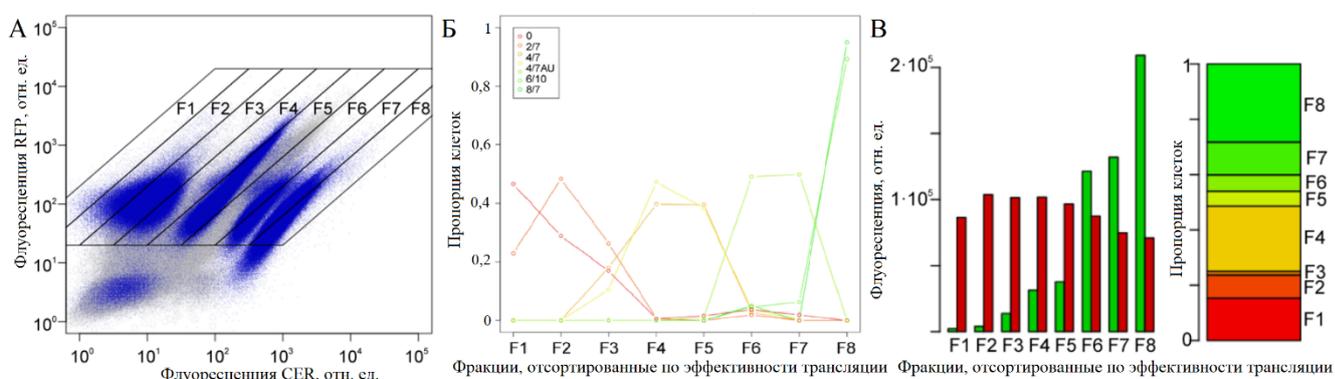


Рисунок 50. Оценка эффективности клеточной сортировки для контрольного набора конструкций. А. Распределение интенсивностей флуоресценции CER и RFP для клеток, трансформированных контрольным набором конструкций 8/7, 6/10, 4/7AU, 4/7, 2/7 и 0. Интенсивности флуоресценции CER и RFP увеличиваются вдоль осей X и Y соответственно. Каждая фракция клеток F1-F8, отличающаяся по соотношению CER/RFP, ограничена наклонной рамкой. Б. Распределение в каждой фракции пропорций клеток, трансформированных отдельной конструкцией: 0 (красные), 2/7 (оранжевые), 4/7 (светло-оранжевые и жёлтые, так как 2 реплики для данной конструкции), 4/7AU (светло-зелёные), 6/10 (зелёные) и 8/7 (тёмно-зелёные). В. Средние пропорции клеток и значения интенсивностей флуоресценции CER и RFP в каждой фракции после сортировки. Интенсивности общей флуоресценции клеток, измеренные спектрофлуориметром, в каждой фракции показаны слева зелёными для CER и красными для RFP столбцами. Пропорция клеток, соответствующая каждой фракции, в исходной библиотеке с контрольными конструкциями показана справа. Фракции F1-F8 отсортированы по увеличению эффективности трансляции.

Таблица 3. Репортёрные конструкции с известным уровнем экспрессии CER к RFP.

Название конструкции	Длина SD (нт)	Расстояние от SD до старт-кодона (нт)	Эффективность трансляции*	Номер максимальной фракции FACS**
8/7	8	7	17	F8
6/10	6	10	13	F8
4/7AU	4	7	3,7	F6-F7
4/7	4	7	0,8	F4-F5
2/7	2	7	0,03	F2
0	0	0	0,001	F1

\* Эффективность трансляции оценивалась ранее измеренным соотношением флуоресценции двух белков CER/RFP [137]

\*\* Номер максимальной фракции FACS соответствует наибольшей пропорции клеток с определённой конструкцией

#### 1.4. Применение метода Flow-seq для разделения клеток *E. coli* с разными репортёрными конструкциями и последующего анализа их последовательностей

На следующем этапе для разделения клеток *E. coli*, трансформированных библиотекой плазмид с 4N-вставкой в спейсерном участке 5'-UTR (Рис. 51А-Б), и для последующего анализа соответствия определённого соотношения белков CER/RFP и переменных последовательностей 5'-UTR, задающих это соотношение, перед геном *cer* (Рис. 51В-Д) применили метод Flow-seq.

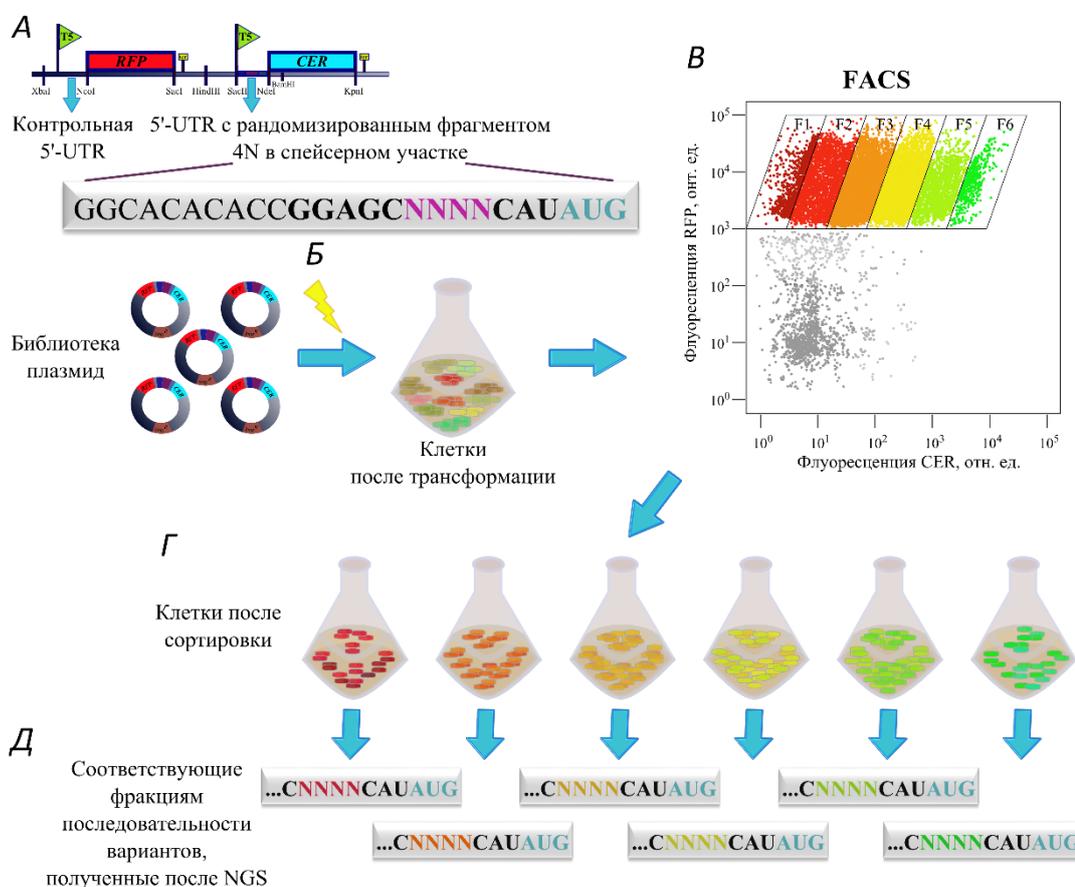


Рисунок 51. Схема метода Flow-seq. А. Схема библиотеки репортёрных конструкций и её создания. Отмечены промоторы, терминаторы, 5'-UTR, кодирующие области флуоресцентных белков CER и RFP. Ниже схемы приведена последовательность 5'-UTR перед геном белка CER: SD выделена красным, старт-кодон – голубым, 4N-фрагмент – фиолетовым. Б. Электропорация клеток *E. coli* библиотекой плазмид. В. Сортировка клеток по соотношению интенсивностей флуоресценции белков CER/RFP с помощью FACS – флуоресцентной проточной цитометрии. Представлено реальное графическое изображение, полученное с помощью FACS, для клеток, трансформированных библиотекой плазмид с 4N-фрагментом в спейсерном участке 5'-UTR. Г. Сбор фракций клеток с конкретным соотношением CER/RFP (в данном случае их 6: F1-F6). Д. Глубокое секвенирование (NGS) ДНК-фрагментов с 4N-участком, полученных после выделения плазмид из клеток каждой фракции и ПЦР-амплификации. Рисунок создан в программе Inkscape.

В ходе сортировки такие клетки были разделены на 6 фракций согласно соотношению белков CER/RFP, измеренному как отношение интенсивностей флуоресценции на длинах волн 405/530 нм для CER и 561/582 нм для RFP (Рис. 51В-Г). Клетки, соответствующие определённой фракции, выделены наклонными рамками на реальной карте, выдаваемой FACS (Рис. 51В).

Далее разделённые на фракции клетки собирали в отдельные пробирки. Затем из клеток выделяли плазмидную ДНК и проводили ПЦР-амплификацию её фрагмента, содержащего 4N-участок в 5'-UTR. Полученные ампликоны длиной 138 нт проверяли с помощью электрофореза в 2% агарозном геле (Рис. 52). Далее для них проводили высокопроизводительное секвенирование (NGS), которое позволило установить соответствие между клетками, несущими определённые варианты вставок в плаزمиде, и их распределением по фракциям, разделённым по соотношению белков CER/RFP, и, как следствие, присвоить определённое значение эффективности трансляции каждой отдельной конструкции (Приложение 2).

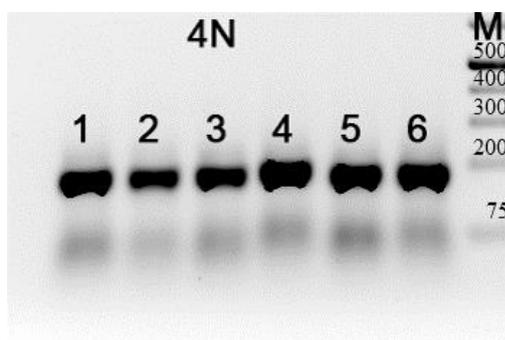


Рисунок 52. Электрофорез в 2% агарозном геле ампликонов, содержащих 4N-фрагмент рандомизации и использованных для NGS. Плазмиды, изолированные из клеток после сортировки на фракции 1-6, использовали для ПЦР-амплификации с праймерами, фланкирующими с обеих сторон область с рандомизированным и константными участками. М – маркер со смесью ДНК-фрагментов определённой длины (1kb Plus DNA Ladder GeneRuler™).

Благодаря секвенированию участков 5'-UTR гена *cer* из отсортированных клеток в библиотеке обнаружили 249 уникальных вариантов из теоретически возможных 256 для участка рандомизации длиной 4 нт (Приложение 2). При этом репортёрные конструкции, содержащие такой участок между последовательностью SD и стартовым кодоном, продемонстрировали не более, чем 100-кратную разницу между самым высоким и самым низким уровнем белка CER на выходе (Рис. 53). Данный диапазон эффективности обеспечивает умеренный, но существенный вклад последовательности спейсерного участка в общую эффективность трансляции.

Принимая во внимание относительно узкое распределение наблюдаемых эффективностей трансляции и неравномерное распределение вариантов 5'-UTR среди шести первоначально полученных фракций (Рис. 53), мы объединили эти фракции в две группы, представляющие

высокие и низкие уровни эффективности трансляции таким образом, чтобы обе группы содержали примерно равное количество вариантов последовательностей. Для этих двух групп и провели дальнейший компьютерный анализ.

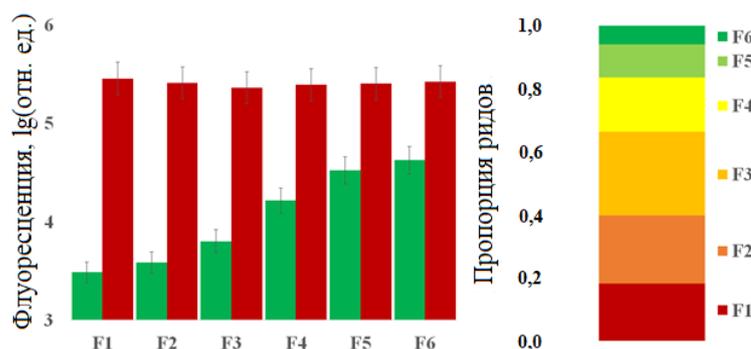
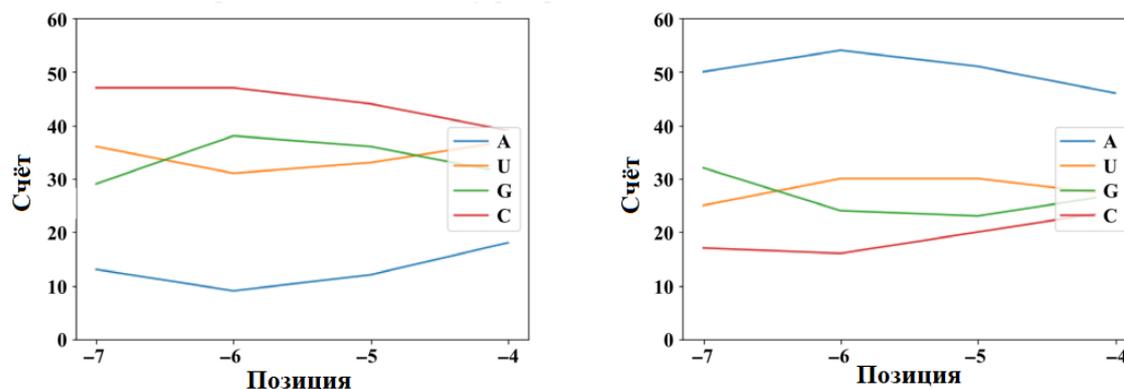


Рисунок 53. Средние значения интенсивностей флуоресценции белков CER и RFP и пропорция ридов, прочтений каждого варианта последовательности, в каждой из фракций после сортировки. Фракции F1-F6 соответствуют наборам клеток, отсортированных по увеличению эффективности трансляции белка CER от F1 до F6. Общие значения интенсивностей флуоресценции клеток в каждой фракции, измеренные флуориметром, показаны в виде зелёных (для CER) и красных (для RFP) столбцов в левой части панели. Для удобства изображения шкала значений интенсивности флуоресценции представлена в виде десятичного логарифма измеренных относительных единиц и начинается с 3 порядка. На правой панели показаны пропорции индивидуальных ридов, найденных в наборе данных, полученных после NGS, для всех вариантов последовательностей в каждой фракции, отсортированной по увеличению эффективности трансляции.

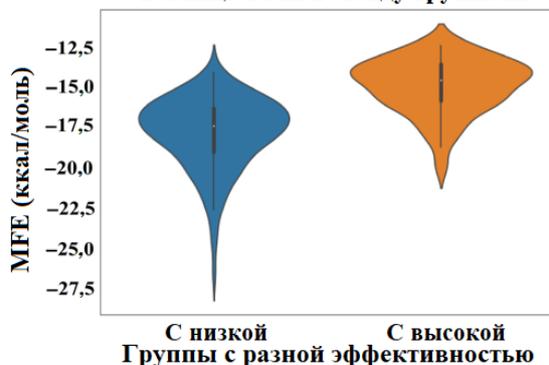
### 1.5. Оценка влияния особенностей спейсерного участка на эффективность трансляции

Сначала проводилась оценка такой особенности спейсерного участка, как нуклеотидный состав. Сравнение его для рандомизированного фрагмента в спейсерной области (Рис. 54А) в эффективно и плохо транслирующихся мРНК выявило достоверное различие в нуклеотидном составе для всех позиций этого участка. При этом значимость различий между двумя группами проверяли с помощью критерия хи-квадрат ( $\chi^2$ ), а оценка случайности наблюдаемой разницы, известная как р-уровень значимости или р-value, составила  $< 10^{-6}$ , что свидетельствует о том, что наблюдаемая разница в нуклеотидном составе между двумя группами не случайна. Во всех четырёх позициях рандомизированной области аденозин оказался наиболее предпочтительным для трансляции, так как встречался чаще всех нуклеотидов в эффективно транслирующихся мРНК и реже остальных – в плохо транслирующихся. В то время как для цитидинов наблюдалась обратная картина – они оказались неблагоприятными для трансляции, так как реже всего встречались в хорошо транслирующихся мРНК и чаще остальных – в плохо транслирующихся.

**А** Частоты нуклеотидов в группе с низкой эффективностью Частоты нуклеотидов в группе с высокой эффективностью



**Б** Разница в MFE между группами



**В** Разница в сходстве с SD между группами

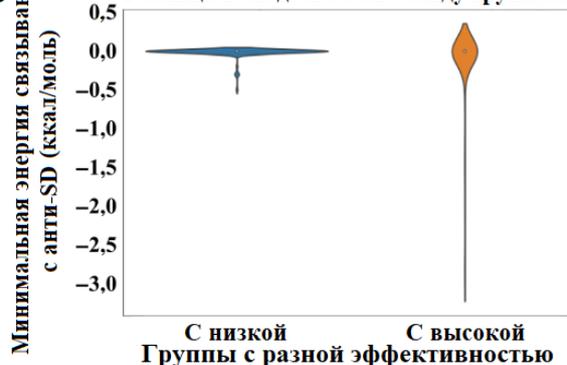


Рисунок 54. Особенности спейсерной области мРНК, кодирующих белок CER и обеспечивающих низкую и высокую эффективность трансляции. А. Частота нуклеотидов в определённых позициях рандомизированного участка мРНК, обеспечивающих низкую (левая панель) и высокую (правая панель) эффективность трансляции. Позиции нуклеотидов нумеруются относительно расстояния до старт-кодона. Б. Распределения минимальной свободной энергии (MFE) сворачивания вторичной структуры в группах с различной эффективностью трансляции. В. Распределения минимальной энергии гибридизации спейсерной области мРНК между константной последовательностью SD и старт-кодом AUG, содержащей весь рандомизированный участок и фланкирующие 1 нт перед ним и 4 нт после него, с 3'-концевой областью 16S рРНК в группах мРНК с различной эффективностью трансляции.

Полученные результаты согласуются с описанными в более ранней работе [209], в которой изучалось влияние спейсерного участка длиной 9 нт на эффективность трансляции. При этом в спейсерной области 6 нт в позициях от -6 до -1 относительно стартового кодона были рандомизированы, а кодоны в положениях +2 и +3 кодирующей области изменялись путём синонимических замен, не меняющих последовательность двух сенсорных белков, перед которыми встраивались разные 5'-UTR. Варьирование спейсерного участка перед старт-кодом позволило создать библиотеку размером 13914 и 25861 из 24576 и 49152 возможных вариантов соответственно для каждого белка. В результате разница между самым низким и самым высоким

уровнем синтезированного белка на выходе получилась  $10^3$ -кратной, тогда как в ходе нашего эксперимента при варьировании 4 нт в спейсерной области 5'-UTR мРНК она была  $10^2$ -кратной. В описанной выше работе также использовали метод Flow-seq и проводили схожий анализ полученных данных.

Аналогичным образом, в этих данных олигоаденозиновые фрагменты из 3 или 4 оснований в позициях от -6 до -3 могли быть обнаружены преимущественно в мРНК, показывающих высокую эффективность экспрессии. Среди последовательностей, заключённых между сильной SD (AAGAAGGA), расположенной на оптимальном расстоянии и оптимальной длины, и старт-кодоном (AUG) и обеспечивающих наибольший уровень экспрессии, можно выделить следующие: gacAAAACC, gacUGUAAG, gacUAGAGC, gacGUGGUU (где “gac” – константная, а “NNNNNN” – вариативная части спейсерной области).

Многие эффективно транслируемые мРНК в нашей библиотеке 4N содержат А-богатые последовательности наряду с U основанием в спейсерном участке. Например, мРНК со следующими вариантами спейсерных последовательностей: сAAAUcau, сAAUAcau, сAUAAcau, сAUAUcau, имеют среднее значение фракции более 5 (Приложение 2), демонстрируя одну из самых высоких эффективностей трансляции для представленной библиотеки наряду с вариантами сAAAАcau, сGAAАcau в окружении сильной последовательности SD (GGAG) и старт-кодона (AUG). В то время как варианты сCCGСcau, сCUCUcau, сCGCUcau, сCCGUcau вошли в число самых неэффективных.

Интересное наблюдение было замечено в обоих экспериментах (нашем и более раннем [209]): среди вариантов мРНК, обеспечивающих наибольшую эффективность трансляции, в 5'-UTR встречается последовательность СААААС.

Последующий анализ GC-состава рандомизированного участка 5'-UTR для мРНК, обладающих различными эффективностями трансляции (Рис. 55), также свидетельствует о положительной роли AU-состава исследуемого участка мРНК на выход белка. Данное наблюдение может отражать положительное влияние AU-богатых энхансеров [78], которые могут связывать рибосомный белок bS1 [14, 48, 78, 137], хотя обычно предполагается, что эти энхансеры предшествуют последовательности SD.

В результате анализа полученных с помощью NGS десятков тысяч различных вариантов последовательностей мРНК также было показано в более ранней работе, что уменьшение содержания GC-нуклеотидов в исследуемой спейсерной области приводит к увеличению выхода синтезируемого белка [209].

Альтернативным объяснением предпочтения в отношении аденозинов в спейсерной области эффективно транслирующихся мРНК является усиленное взаимодействие с рибосомой. В то время как спейсерный участок мРНК не участвует в образовании пар оснований с 16S рРНК,

нуклеотиды в положениях от -3 до -1 спейсерного участка длиной 4 нт образуют стекинг-взаимодействия ( $\pi$ - $\pi$ -взаимодействия между плоскими ароматическими молекулами/”stacking”) друг с другом и нуклеотидом G926 в 16S рРНК [107]. Нуклеотид A1503 в 16S рРНК контактирует с нуклеотидом мРНК в -4 положении (относительно первого нуклеотида в Р-сайте) в нескольких инициаторных комплексах [107, 245] и с нуклеотидом в -2 положении в интермедиате при транслокации [246]. При условии, что мало известно о структуре инициаторных комплексов, содержащих более длинные спейсерные области мРНК, такие как, например, используемые в нашем исследовании длиной 8 нт, можно предположить, что рибосома может формировать некоторые специфичные по последовательности взаимодействия (наиболее вероятны стекинг-взаимодействия) с этим участком мРНК.

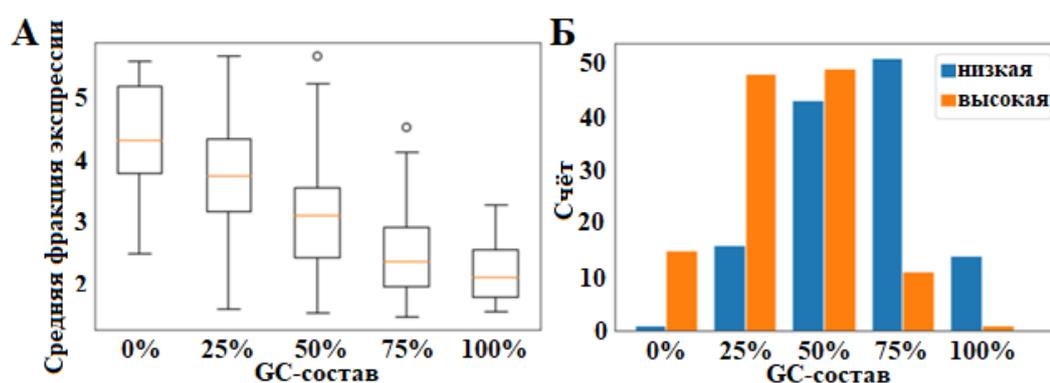


Рисунок 55. Анализ GC-состава рандомизированной последовательности из 4 нт в спейсерном участке в эффективно и плохо транслируемых мРНК, кодирующих белок CER. GC-состав оценивали и представляли на графике в зависимости от среднего значения фракции экспрессии (Приложение 2) для каждого варианта последовательности (А) или от количества вариантов в двух классах с низкой (синего цвета) и высокой (оранжевого цвета) эффективностью трансляции (Б). Чем больше среднее значение фракции, тем большую эффективность трансляции можно наблюдать для отдельного варианта последовательности.

Спейсерная область может быть частью вторичной структуры мРНК, возможно маскирующей другие элементы сайта инициации трансляции, такие как последовательность SD или старт-кодон. Действительно, образование вторичной структуры, второй особенностью спейсерного участка, мРНК приводит к ингибированию инициации трансляции [141, 156, 209]. С целью проанализировать влияние спейсерной последовательности на энергию сворачивания сайта инициации трансляции, а следовательно, и на эффективность трансляции, мы смоделировали вторичные структуры для всех последовательностей мРНК в нашем наборе данных, используя «окно», охватывающее всю 5'-UTR длиной 22 нт и первые 50 нт кодирующей области. Эти последовательности использовались для расчёта минимальной свободной энергии

(MFE), стандартной оценки предсказанной стабильности вторичной структуры. При этом более низкие значения MFE соответствуют более стабильным вторичным структурам (Рис. 54Б). Тест Колмогорова-Смирнова продемонстрировал, что разница в стабильности вторичной структуры между эффективно и плохо транслируемыми мРНК была достоверной на уровне  $10^{-15}$ . Для того, чтобы показать, что этот эффект не был тривиальным следствием различий в нуклеотидном составе, мы провели пермутационный анализ (см. описание в Методах) и продемонстрировали, что разница в MFE между фракциями была значительно больше, чем ожидалось, учитывая наблюдаемые частоты нуклеотидов в каждой позиции ( $p$ -value = 0,001, Рисунок в Методах).

Отсутствие или минимизация вторичной структуры мРНК в исследуемой спейсерной области приводят к увеличению выхода синтезируемого белка, как было показано в более ранней работе при анализе результатов NGS десятков тысяч различных вариантов мРНК [209].

Все варианты 5'-UTR в нашем наборе данных содержали последовательность SD длиной 4 нт, расположенную на расстоянии 8 нуклеотидов перед старт-кодом AUG. Далее проверили, может ли появление дополнительных участков комплементарности 3'-концевой области 16S рРНК, являющихся третьей особенностью, влиять на эффективность трансляции. С этой целью мы рассчитали свободную энергию гибридизации последовательности анти-SD (CACCUCCU) в 3'-концевой области 16S рРНК с участком 5'-UTR мРНК (CNNNNCAUA), содержащим фрагмент рандомизации (Рис. 54В). В целом два наблюдаемых распределения этих энергий существенно не отличались (для теста Колмогорова-Смирнова  $p$ -value > 0,05). Однако спейсерные области плохо транслируемых мРНК, кодирующих белок CER, не имели SD-подобных участков в дополнение к имеющейся SD, тогда как несколько спейсеров из эффективно транслируемого набора мРНК содержали последовательность, которая могла образовывать комплементарные взаимодействия с 3'-концом 16S рРНК. Примеры таких спейсеров включали следующие варианты фрагментов рандомизации: AGGA, AAGG, GAGG, GGGA, GGAG, GGGG. Данное наблюдение указывает на аддитивное влияние множественных SD-мотивов в 5'-UTR на эффективность трансляции. Но этот механизм может объяснить наблюдаемую высокую скорость трансляции только ограниченного набора эффективно транслируемых мРНК.

Несколько инструментов для предсказания эффективности трансляции опубликовано для моноцистронных [149-150, 215] и бицистронных [214, 247] конструкций. Сравнение измеренной эффективности трансляции мРНК в нашем наборе данных с предсказанной с помощью RBS Calculator [149-150] (Приложение 2) продемонстрировало относительно хорошую, хотя и не идеальную корреляцию (коэффициент корреляции Пирсона составляет:  $r = 0,62$ ).

Полученные данные позволяют предложить следующие рекомендации по выбору спейсерных участков между последовательностью SD и старт-кодом для бактериальных систем экспрессии. Для того, чтобы ускорить трансляцию, а следовательно, и увеличить количество

синтезируемого белка на выходе, кажется, разумно использовать олигоадениловые или другие А-богатые спейсерные последовательности, при этом избегая оснований цитидина. Однако нельзя исключать, что некоторые конкретные мРНК с А-богатыми спейсерными областями могут маскировать сайт инициации трансляции своей вторичной структурой в случае, если кодирующая область будет начинаться с U-богатой последовательности. Наши результаты могут помочь регулировать уровень экспрессии экзогенных генов для конкретной биотехнологической потребности. Для коэкспрессии генов, продукты которых должны быть синтезированы в определённом стехиометрическом соотношении, к примеру, белковые субчастицы входят в состав гетеромультимерного комплекса, в таком случае возможно отрегулировать уровни экспрессии этих генов с помощью правильного выбора спейсерных последовательностей между SD и старт-кодонами смоделированных мРНК.

## 2. Изучение влияния рандомизированных 5'-UTR на эффективность трансляции

При подготовке данного раздела диссертации использована следующая публикация, выполненная соискателем в соавторстве, в которой, согласно Положению о присуждении учёных степеней в МГУ, отражены основной результат, положения, выводы исследования: [19], общий объём публикации – 1 п.л. Автором лично проведены все экспериментальные процедуры, при этом FACS, NGS и биоинформатический анализ полученных данных проводились совместно с автором, при активном его участии.

### 2.1. Создание репортёрных конструкций с рандомизированной 5'-UTR перед стартовым кодоном гена *cer*

Для определения какая именно последовательность 5'-UTR мРНК является наиболее эффективной для посадки рибосомы и приводит к наибольшей экспрессии гена, был создан набор плазмид с рандомизированной вставкой длиной 20 и 30 нуклеотидов (20N и 30N соответственно) непосредственно перед старт-кодом гена *cer* по следующей схеме (Рис. 56) на основе существующих методик клонирования небольших рандомизированных участков [241, 248].

В качестве исходных фрагментов для последующей вставки в вектор были взяты синтетически синтезированные одноцепочечные олигонуклеотиды двух типов: прямой и обратный. Прямой представлен двумя вариантами последовательностей длиной 52 и 62 нт (оц олиги) с участками рандомизации в 20 и 30 нт соответственно, содержащими также стартовый кодон и начало кодирующей области гена *cer* длиной 19 нт наряду с сайтом узнавания эндонуклеазы рестрикции SacII:

5' – А С Т G **CCGCGG** **NNN...N** **ATGAAAGAGACGGACGAGAGCG** – 3',

которая вносит следующие разрывы:

SacII: 5' – CCGCGG GG – 3'  
 3' – GG CGCC – 5'.

Обратный в виде фрагментов длиной 26 нт, большей частью комплементарных к 3'-концу, включающему кодирующую область и старт-кодон, последовательностей прямого типа:

5' – GATC CGCTCTCGTCCGTCTCTTTCAAT – 3',

причём на его 5'-концах имеется участок, который будет соответствовать липким концам плазмиды, обработанной эндонуклеазой рестрикции BamHI. Дальнейшая подготовка проводилась путём отжига в течение 1' при 95°C с последующим охлаждением до комнатной температуры этих двух типов олигонуклеотидов друг на друга с образованием дуплексов.

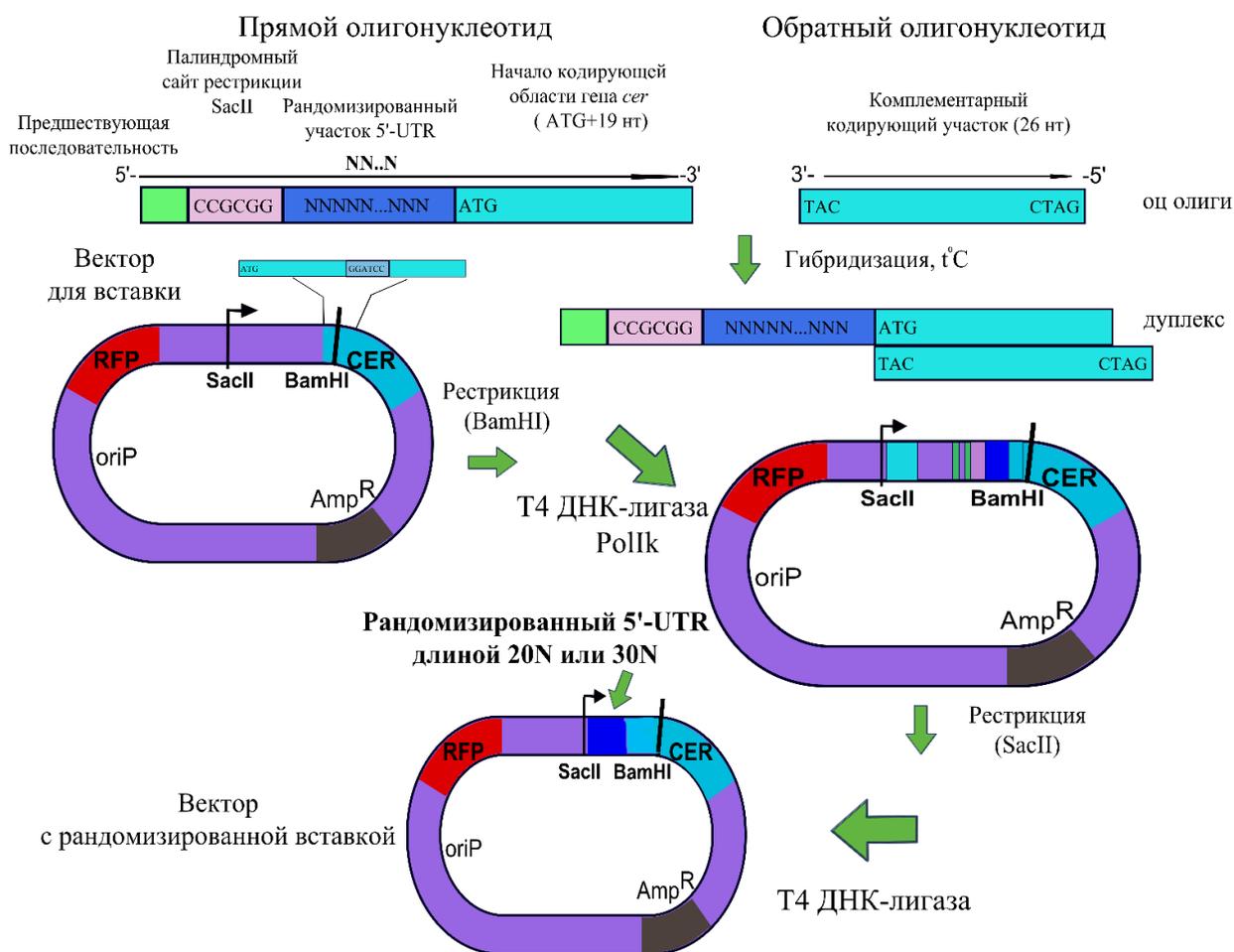


Рисунок 56. Схема создания плазмид с ДНК-библиотекой 5'-UTR. Оц олиги – одноцепочечные олигонуклеотиды: прямые длиной 52 или 62 нт, обратные длиной 26 нт. Дуплекс – отождённые друг на друга одноцепочечные олигонуклеотиды. Amp<sup>R</sup> – белок устойчивости к ампициллину. oriP – ориджин репликации плазмиды. Pollk – фрагмент Клёнова ДНК-полимеразы I *E. coli*. SacII, BamHI – эндонуклеазы рестрикции. Рисунок выполнен в программе Inkscape.

Следующим этапом была подготовка вектора для вставки. В данном случае в качестве вектора взяли вариант pRFP-CER BamHI, отличающийся от описанного ранее pRFP-CER [137, 156],

использованного для встраивания 5'-UTR с участком рандомизации в спейсерной области (Рис. 44). pRFPCER BamHI содержит вставку с сайтом узнавания эндонуклеазы рестрикции BamHI сразу за старт-кодоном, дублирующую начало кодирующей области гена *cer* (Рис. 57).

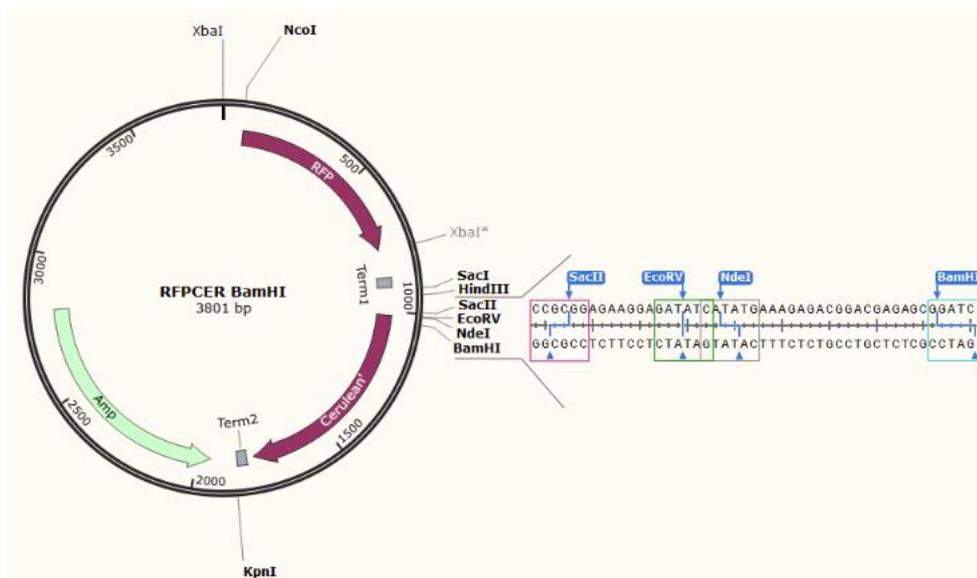


Рисунок 57. Схема плазмиды pRFPCER BamHI с выделенными использованными сайтами рестрикции эндонуклеаз SacII, EcoRV и BamHI. Участки узнавания выделены рамками, для EcoRV и NdeI они перекрываются. В векторе закодирован также белок устойчивости к ампициллину ( $Amp^R$ ), по которому в дальнейшем проводилась селекция. Схема создана в программе Clone Manager.

Встраивание полученных дуплексов в исходный вектор (Рис. 57) проводилось по сайту рестрикции BamHI, находящемуся на расстоянии 18 нт после стартового кодона в кодирующей области гена *cer*:



С образовавшимися таким образом липкими концами плазмиды с двух сторон при добавлении T4 ДНК-лигазы сшивались вышеуказанные дуплексы, имеющие комплементарные им выступающие 5'-концы из обратных олигонуклеотидов, о которых упоминалось ранее. Затем добавлялись фрагмент Клёнова ДНК-полимеразы I *E. coli* и смесь dNTP, и пришитые к плазмиде дуплексы достраивались до двухцепочечных. В результате получалась конструкция, содержащая три сайта рестрикции эндонуклеазы SacII. Именно эта рестриктаза использовалась на следующем этапе для разрезания с последующим лигированием T4 ДНК-лигазой. В итоге получался вектор теперь уже с одним участком рестрикции по SacII и встроенным рандомизированным участком длиной 20N или 30N непосредственно перед стартовым кодоном гена *cer* (Рис. 56).

В качестве проверки на наличие вставки выполнялась рестрикция по сайту EcoRV:

EcoRV: 5' – **GAT**ATC – 3'

3' – CTA**TAG** – 5',

который присутствует только в исходном векторе между участками узнавания эндонуклеаз SacII и NdeI (Рис. 57) и ещё в тех нескольких встроенных рандомизированных участках, которые могут встретиться, но на фоне большого размера всей библиотеки их количеством решили пренебречь.

Таким образом, получили плазмиды с различными вставками 5'-UTR, представленными всеми возможными вариантами последовательностей длиной 20N и 30N непосредственно перед стартовым кодоном гена *cer* (Рис. 56), при этом размер библиотеки ДНК в каждом случае можно оценить, как  $4^{20} \sim 10^{12}$  и  $4^{30} \sim 10^{18}$  различных вариантов соответственно. Наличие участков рандомизации уже в составе вектора pRFPCER BamHI проверялось секвенированием по Сэнгеру. С целью минимизировать влияние последовательности 5'-UTR на транскрипцию, наиболее чувствительную к идентичности первых нуклеотидов мРНК-транскрипта, сохранили на 5'-конце его последовательность GG, за которой сразу следовал участок рандомизации 20N или 30N, в итоге 5'-UTR получалась длиной 22 или 32 нт соответственно.

Данный метод имеет значительное преимущество перед ПЦР-амплификацией, поскольку использование метода ПЦР неминуемо бы привело к появлению мутаций за пределами участка рандомизации. И эти мутации, во-первых, никак нельзя было бы определить, так как планируется секвенировать лишь участок вставки. А во-вторых, они могут вносить совершенно неожиданный вклад как в абсолютные количества флуоресцентных белков, так и в их соотношение.

Следует отметить, что в данном эксперименте перед репортёрным сенсорным геном *cer* встраивали 5'-UTR длиной 22 нт или 32 нт, близкой к наиболее часто встречающимся у мРНК *E.coli* (представлены в диапазоне длин 20-40 нт) [110, 157]. Тогда как перед контрольным репортёрным геном *rfp* и перед геном *cer* в контрольной плазмиде находилась консервативная 5'-UTR длиной 17 нт, также близкой к самым часто встречающимся у мРНК *E.coli*, с оптимальной SD длиной 6 нт на расстоянии 9 нт от старт-кодона (при расчёте удаления центрального G в SD от IR) и оптимальным спейсерным участком длиной 7 нт (Рис. 57). В итоге, опираясь на литературные данные, создана библиотека конструкций с учётом оптимальных характеристик основных элементов 5'-UTR мРНК для экспрессии репортёрных генов в клетках *E. coli*.

## 2.2. Клонирование полученной библиотеки плазмидных ДНК

Учитывая очень большой размер библиотек с рандомизированными участками 20N и 30N, насчитывающих  $4^{20} \sim 10^{12}$  и  $4^{30} \sim 10^{18}$  различных вариантов соответственно, на следующем этапе трансформацию проводили с помощью электропорации электрокомпетентных клеток штамма JM109 *E.coli* библиотеками репортёрных конструкций. Затем к трансформированным клеткам

добавляли жидкую среду SOC с глюкозой и перемещали в термостат на 37°C для наработки белка устойчивости к ампициллину, по которому после высевания клеток на большую чашку Петри проводилась селекция колоний. В параллель с помощью титрования проводили подсчёт колоний.

Далее библиотеки конструкций выделяли из клеток и трансформировали ими с помощью электропорации клетки штамма BW25113 *E.coli* для последующей сортировки. С целью амплификации библиотеки плазмид трансформированные клетки растили в течение 16 ч при температуре 37°C в 100 мл жидкой среды LB с ампициллином. Оценка количества полученных клеток после электропорации с помощью титрования показала  $10^6$ - $10^7$  различных вариантов, лишь частично покрывающих размер возможных библиотек 20N и 30N. Достигнутый на этом этапе результат можно считать значимым, так как полученное методом Flow-seq разнообразие вариантов на 3-4 порядка превышает число 5'-UTR природных генов *E.coli* и этого вполне достаточно для поиска различных по эффективности сайтов посадки рибосомы в 5'-UTR мРНК.

Далее клеточную культуру центрифугировали после инкубирования при 37°C в течение 16 часов и промывали раствором PBS для подготовки клеток к дальнейшей сортировке. Параллельно контрольной плазмидой с идентичными 5'-UTR перед стартовыми кодонами генов обоих флуоресцентных белков и, как следствие, с приблизительно одинаковым соотношением CER/RFP трансформировали клетки JM109 и BW25113 *E.coli*, а затем готовили к сортировке.

### **2.3. Применение метода Flow-seq для разделения клеток *E. coli* с разными репортёрными конструкциями, содержащими рандомизированную 5'-UTR, и последующего анализа их последовательностей**

На следующем этапе применили метод Flow-seq для разделения клеток *E. coli*, которые трансформировали библиотекой плазмид с участками рандомизации 20N и 30N в 5'-UTR перед геном *cer*, и дальнейшего анализа соответствия определённого соотношения белков CER/RFP и задающих его переменных последовательностей 5'-UTR перед геном белка CER.

Клетки *E. coli*, трансформированные каждым видом библиотек, были разделены на две группы с целью получения биологических реплик для оценки воспроизводимости результатов сортировки. В ходе FACS для каждой группы оценили процент светящихся клеток: 96,3% и 96,9% в случае библиотек 20N и 30N соответственно, а затем их разделили на 8 фракций согласно соотношению белков CER/RFP, измеренному как отношение интенсивностей флуоресценции на длинах волн 405/530 нм для CER и 561/582 нм для RFP. Клетки, соответствующие определённой фракции, выделены наклонными рамками на реальной карте, выдаваемой FACS (Рис. 58).

Фракцию F5, в которую попали бы клетки с контрольной плазмидой или преимущественно с конструкцией 4/7 из контрольного набора, составляют теперь всего 2,6% и 8,3% клеток из библиотек 20N и 30N соответственно, в отличие от 91% клеток после первой сортировки с

контрольной плазмидой. Как и ожидалось, большинство вариантов последовательностей 5'-UTR с участками рандомизации 20N и 30N плохо функционировали в качестве сайтов инициации трансляции. При этом только 0,03% клеток, трансформированных библиотекой 20N, попадают во фракцию F8, соответствующую максимальной эффективности трансляции белка CER (Рис. 59А), тогда как 0,001% клеток в ней было при сортировке с контрольной плазмидой. Хотя клетки, трансформированные библиотекой 30N, также имеют очень небольшое количество эффективно транслируемых вариантов, их количество значительно выше (Рис. 59Б), чем у 20N. В этом случае во фракцию F8 с максимальной эффективностью трансляции попадает до 0,45% клеток.

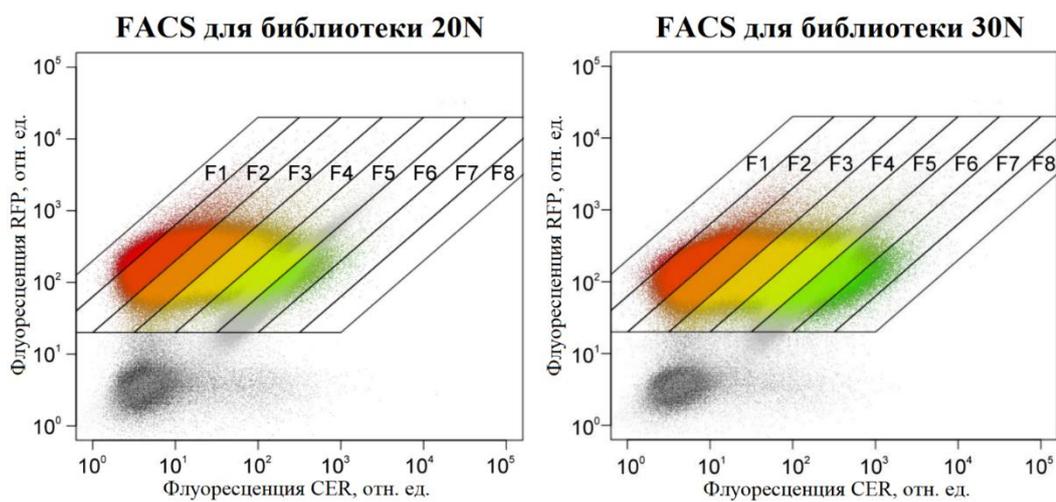


Рисунок 58. Распределение интенсивностей флуоресценции CER и RFP для клеток, которые трансформировали библиотеками репортёрных конструкций с участками рандомизации длиной 20 нт (20N, слева) и 30 нт (30N, справа) в 5'-UTR перед геном *cer*. Каждая точка на графике соответствует отдельной клетке. Интенсивности флуоресценции CER и RFP увеличиваются вдоль осей X и Y соответственно. Каждая фракция клеток F1-F8, отличающаяся по соотношению CER/RFP, ограничена наклонной рамкой. Наложение результатов сортировки для клеток, трансформированных конструкцией 4/7 из контрольного набора, показано серым цветом. Фракции F1-F8 отсортированы по увеличению эффективности трансляции.

Постепенное снижение пропорций клеток от слабо транслирующихся мРНК до небольшой доли высокоэффективных вариантов 5'-UTR в рандомизированных библиотеках свидетельствует о том, что движение рибосомы вдоль всей кодирующей области белка CER во время элонгации не ограничивает трансляцию в данной системе. В противном случае можно было бы наблюдать искусственное увеличение количества клеток во фракции, соответствующей минимальной эффективности трансляции, ограниченной элонгацией, с последующим резким уменьшением количества клеток, соответствующих более эффективной трансляции. Поскольку не наблюдали этот эффект экспериментально (Рис. 59), мы можем сделать вывод, что элонгация трансляции

кодирующей области белка CER происходит достаточно быстро, чтобы не ограничивать общий выход белка CER.

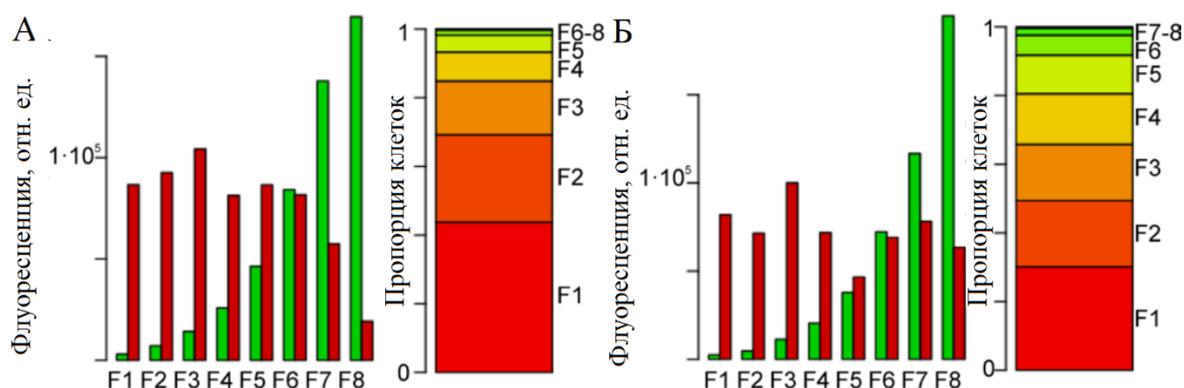


Рисунок 59. Средние пропорции клеток и значения интенсивностей флуоресценции CER и RFP в каждой фракции, полученной после сортировки. Фракции F1-F8 отсортированы по увеличению эффективности трансляции. Панели А и Б соответствуют клеткам, трансформированным библиотеками репортёрных конструкций с участками рандомизации 20N и 30N в 5'-UTR перед геном *cer*. Интенсивности общей флуоресценции клеток, измеренные спектрофлуориметром, в каждой фракции показаны зелёными для CER и красными для RFP столбцами слева на каждой панели. Распределение в каждой фракции пропорций клеток, трансформированных библиотеками 20N (А) и 30N (Б), показано справа на каждой панели.

Далее разделённые на фракции клетки каждой из двух групп собирали в отдельные пробирки и растили в течение 16 часов при интенсивном перемешивании при 37°C. После чего оценивали соотношение интенсивностей флуоресценции CER/RFP, измеряя с помощью спектрофлуориметра и нормируя на контрольное значение соотношения, соответствующего контрольной плазмиде. Количественный учёт отсортированных клеток проводился с помощью их титрования. Затем из клеток выделяли плазмидную ДНК и проводили ПЦР-амплификацию её фрагментов длиной 173 нт и 183 нт, содержащих 20N- и 30N-участки в 5'-UTR соответственно. Полученные ампликоны проверяли с помощью электрофореза в 2% агарозном геле. Далее для них проводили секвенирование следующего поколения, которое позволило установить соответствие между клетками, несущими определённые варианты вставок в плаزمиде, и их распределением по фракциям, разделённым по соотношению белков CER/RFP. В результате каждой отдельной конструкции присвоили определённое значение эффективности трансляции.

Последовательности, наблюдаемые в обеих репликах библиотек, использовались для оценки воспроизводимости сортировки. Для библиотек репортёрных конструкций, содержащих участок рандомизации 20N или 30N в 5'-UTR, мы наблюдали коэффициенты корреляции 0,95 и 0,98 соответственно. Поскольку воспроизводимость была достаточной, для дальнейшего анализа

мы объединили обе реплики, получив 11692 и 11889 уникальных 5'-UTR, содержащих фрагмент рандомизации 20N и 30N соответственно. Большинство вариантов было представлено от 10 до 1500 прочтениями (ридами), обычно обнаруживаемыми в одной или двух фракциях после сортировки. Таким образом, мы получили уникальную репрезентативную библиотеку 5'-UTR, которая примерно в 5 раз более разнообразна, чем коллекция 5'-UTR природных генов *E. coli* [249]. При этом репортёрные конструкции из обеих библиотек продемонстрировали  $10^4$ -кратную разницу между самым высоким и самым низким уровнем белка CER на выходе, причём для 20N библиотеки диапазон изменения эффективности трансляции уже, чем для 30N (Рис. 59). Данный диапазон значительно шире, чем наблюдаемый при рандомизации спейсерного участка ( $10^2$  раз), и обеспечивает значительный вклад последовательности 5'-UTR в общую скорость трансляции.

#### **2.4. Оценка влияния особенностей рандомизированных 5'-UTR на эффективность трансляции**

В первую очередь проводился анализ такой особенности рандомизированных 5'-UTR, различающихся по эффективности трансляции, как нуклеотидный состав. Он выявил почти равномерное распределение всех четырёх нуклеотидов в 5'-UTR с низкой эффективностью для обеих библиотек 20N и 30N (Рис. 60, панели F1). Более высокая эффективность трансляции привела к более смещённому нуклеотидному составу 5'-UTR (Рис. 60, панели F2-F8). Заметное обогащение основаниями А и G в позициях от -13 до -7 относительно стартового кодона AUG в высоко эффективных 5'-UTR (Рис. 60, панели F8) скорее всего отражает выборку из SD-подобных последовательностей (будет рассмотрено далее). Другие участки последовательности 5'-UTR эффективно транслируемых вариантов мРНК также демонстрируют отклонения в нуклеотидном составе. Значительное преобладание оснований А и, в меньшей степени, U было характерно для областей как до, так и после предполагаемых SD-мотивов в 5'-UTR из фракции с наибольшей экспрессией (Рис. 60, панели F8). Перепредставленность оснований А и U в последовательностях 5'-UTR мРНК с наибольшей эффективностью трансляции, скорее всего, отражает присутствие АU-богатых энхансеров трансляции [74], с которыми, как предполагалось ранее, связывается рибосомный белок bS1 [14, 74, 132-133]. Примечательно, что все участки высокоэффективных вариантов 5'-UTR имеют заметно сниженную долю оснований С (Рис. 60, панели F8). Данные наблюдения уже отмечены и ранее при оценке нуклеотидного состава в участке рандомизации 4N в спейсерной области 5'-UTR мРНК с высокой эффективностью трансляции.

Дальнейшее сравнение нуклеотидного состава эффективно транслируемых вариантов последовательностей 5'-UTR из библиотек, содержащих рандомизированные участки 20N и 30N, показало, что состав более коротких 5'-UTR в случае 20N имеет гораздо более сильные смещения по частоте встречаемости всех типов оснований, отражающие частое присутствие сильных

последовательностей SD, особенно в позициях от -13 до -7 относительно старт-кодона гена *ser* (Рис. 60, панели F8). В то же время в случае 30N 5'-концевой участок более длинных вариантов 5'-UTR в эффективно транслируемых мРНК, отсутствующий в коротких 5'-UTR из библиотеки 20N, имеет смещение по составу в сторону оснований AU по сравнению с нуклеотидами GC, что отражает важность этой области для эффективности трансляции.

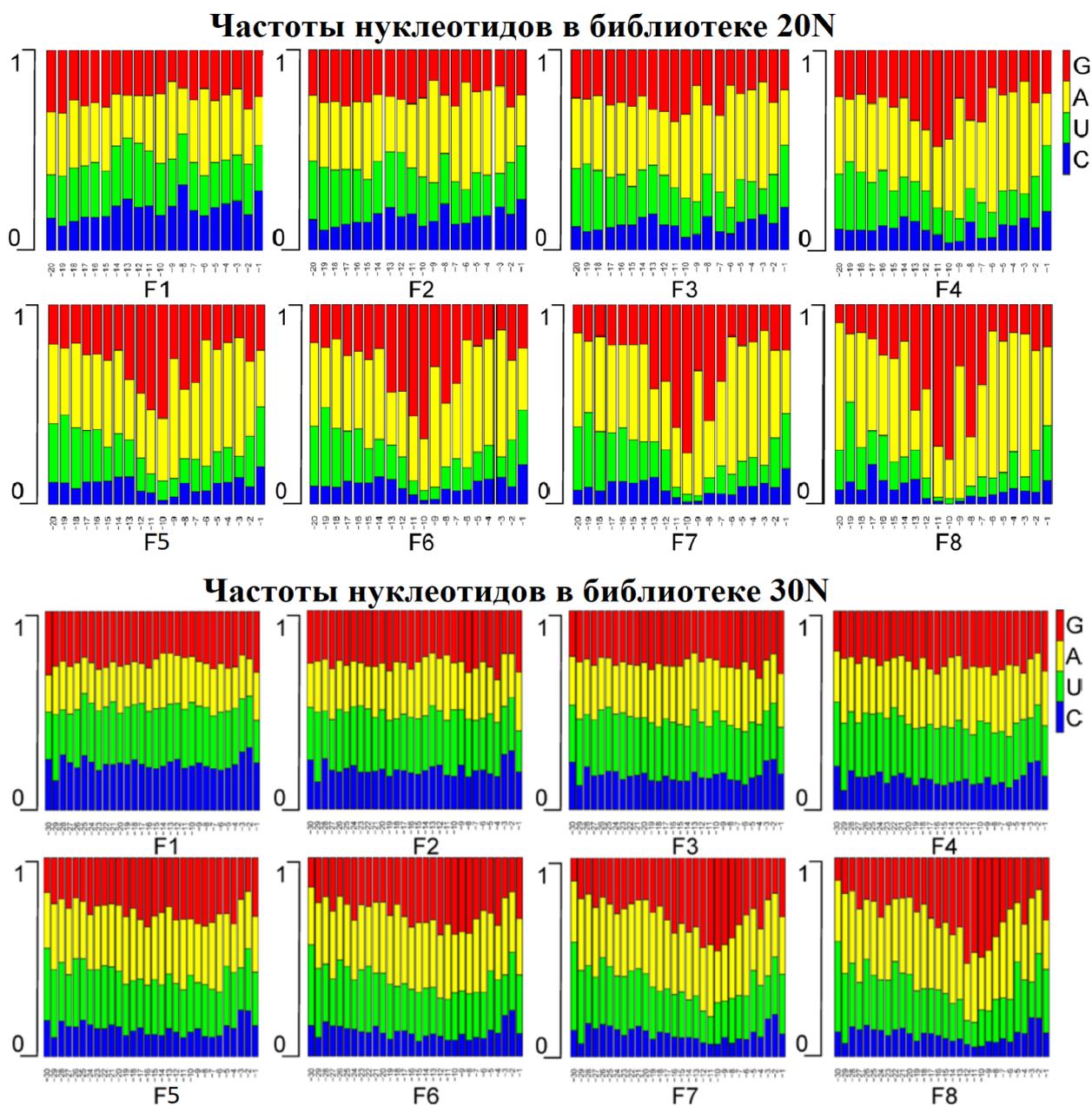


Рисунок 60. Влияние нуклеотидной последовательности рандомизированных фрагментов 5'-UTR длиной 20 нт (20N, сверху) и 30 нт (30N, снизу) на эффективность трансляции. Панели F1-F8 соответствуют фракциям, отсортированным по увеличению эффективности трансляции. Каждая полоска представляет собой долю нуклеотидов в каждой позиции 5'-UTR относительно старт-кодона AUG, как показано снизу под графиками. Нуклеотиды имеют цветовую кодировку, как показано в легенде справа от панелей F4: G – красный, A – жёлтый, U – зелёный, C – синий.

В предыдущей работе [137], сделанной в нашей лаборатории, показано, что удлинение короткой 5'-UTR (конструкция 4/7 из контрольного набора) с помощью AU-богатого энхансера с образованием 5'-UTR (конструкция 4/7AU из контрольного набора) длиной 29 нт привело к 5-кратному увеличению эффективности трансляции (Таблица 3). Природные транскрипты *E. coli* в среднем имеют 5'-UTR длиной 25-35 нт [110], что скорее больше согласуется с характеристиками 5'-UTR длиной 30 нт.

Подводя итог, можно сказать, что большинству коротких 5'-UTR необходимо иметь большую последовательность SD, чтобы обеспечить эффективную трансляцию, тогда как более длинные 5'-UTR могут достигать той же эффективности с SD меньшей длины и различными, например, AU-богатыми энхансерами трансляции, располагающимися в 5'-области перед SD.

На следующем этапе проводилась оценка влияния второй особенности 5'-UTR – вторичной структуры – на эффективность трансляции. Энергия укладки вторичной структуры для всей последовательности 5'-UTR длиной 22 нт (в случае библиотеки 20N) или 32 нт (в случае библиотеки 30N) и первых 50 нт кодирующей области гена *cer* рассчитывалась для всех вариантов 5'-UTR из полученных наборов данных для обеих библиотек. Затем строилось распределение энергии укладки для всех 5'-UTR, сгруппированных по эффективности экспрессии (Рис. 61). Анализ вторичной структуры для всех вариантов последовательностей 5'-UTR, обнаруженных с помощью метода Flow-seq, показал, что энергия сворачивания положительно коррелирует с эффективностью трансляции. Это означает, что для биосинтеза белка невыгодно формирование сильной вторичной структуры в 5'-UTR мРНК.

Далее провели оценку значений показателей спаривания оснований нуклеотидов (pfold score) в 5'- и 3'-направлениях и построили распределения данных показателей для всех вариантов последовательностей 5'-UTR из каждой фракции F1-F8 в обеих библиотеках 20N (Рис. 62) и 30N. При этом наблюдалось, что доля нуклеотидов с высокими показателями спаривания pfold score неуклонно снижается с увеличением эффективности трансляции. Более того, расположение спаренных участков 5'-UTR неслучайно на мРНК с разной эффективностью трансляции. 5'-UTR неэффективно транслируемых мРНК обогащены элементами вторичной структуры в сайте связывания рибосомы, в то время как элементы вторичной структуры 5'-UTR в эффективно транслируемых мРНК, если они и присутствуют, то, как правило, не перекрываются с сайтом связывания рибосомы. Паттернов вторичной структуры, последовательно увеличивающих эффективность трансляции, не наблюдалось.

Анализ распределения энергии укладки в пределах «скользящего окна» длиной 30 нт показал, что доля структурированных вариантов 5'-UTR максимальна для последовательностей, приводящих к низкой эффективности трансляции. Данные распределения построены для обеих библиотек 20N и 30N, но полученные серии графических изображений не представлены в работе.

## Анализ стабильности вторичной структуры 5'-UTR

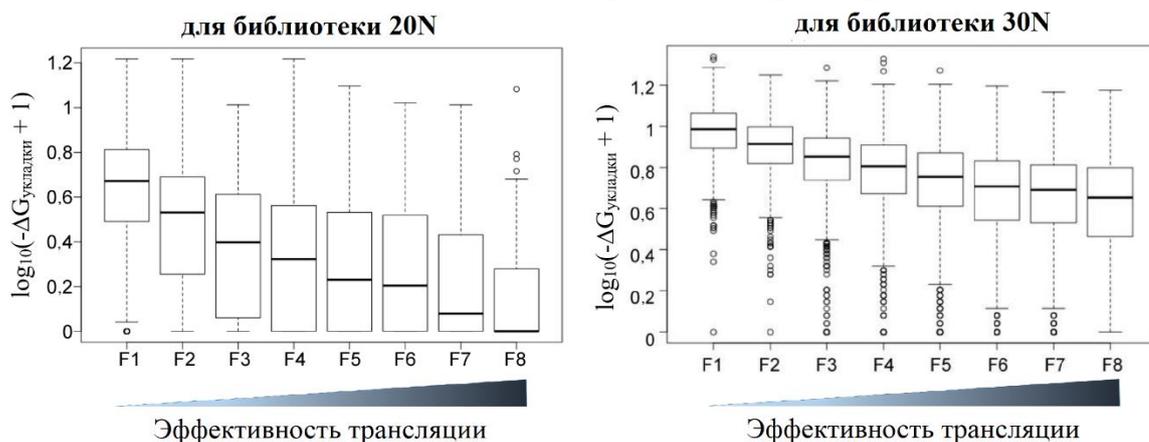


Рисунок 61. Влияние вторичной структуры на эффективность трансляции. Распределение энергии укладки 5'-UTR с 20N- (слева) и 30N- (справа) фрагментами рандомизации с различной эффективностью трансляции. Каждый прямоугольник соответствует фракции, отмеченной внизу графика, фракции отсортированы по возрастанию эффективности трансляции: от самой низкой F1 до самой высокой F8. Распределение множества значений оценки энергии укладки, рассчитанной как  $\log_{10}(-\Delta G_{\text{укладки}} + 1)$ , представлено в виде «ящика с усами», известного в статистике. 25-75 перцентиль множества значений изображен в виде «ящика», медианное значение (квартиль Q2) в нём показано горизонтальной линией, тогда как выпадающие значения за пределами «ящика», обозначающего межквартильный размах (MP, включающий диапазон значений Q1-Q3), то есть менее первого квартиля Q1 на 1,5 MP (-1,5 MP) и более третьего квартиля Q3 на 1,5 MP (+1,5 MP), изображены кружками.

Полученные результаты свидетельствуют о том, что стабильная вторичная структура приводит к ингибированию инициации трансляции, что согласуется с данными, приведёнными в более ранних работах [141, 156, 209] и описанными ранее при анализе вторичной структуры спейсерного участка 5'-UTR мРНК.

Дальнейший этап заключался в оценке таких важных особенностей 5'-UTR, как частота и длина последовательностей SD, встречающихся в 5'-UTR мРНК, отличающихся по уровню экспрессии. Последовательность SD, комплементарная 3'-концевой области 16S рРНК, является наиболее известным элементом 5'-UTR, повышающим эффективность трансляции. Для всех вариантов 5'-UTR, присутствующих в полученных наборах данных для двух библиотек 20N и 30N, мы рассчитали показатель сходства с канонической последовательностью SD (SD PSSM score), представленной на Рисунке 63В, для каждых 9 нт «скользящего окна» в 5'-UTR. Для каждого варианта 5'-UTR найдено «окно» с максимальным сходством с SD и построены распределения максимальных оценок для мРНК, различающихся по эффективности трансляции в обеих библиотеках (Рис. 63А-Б).

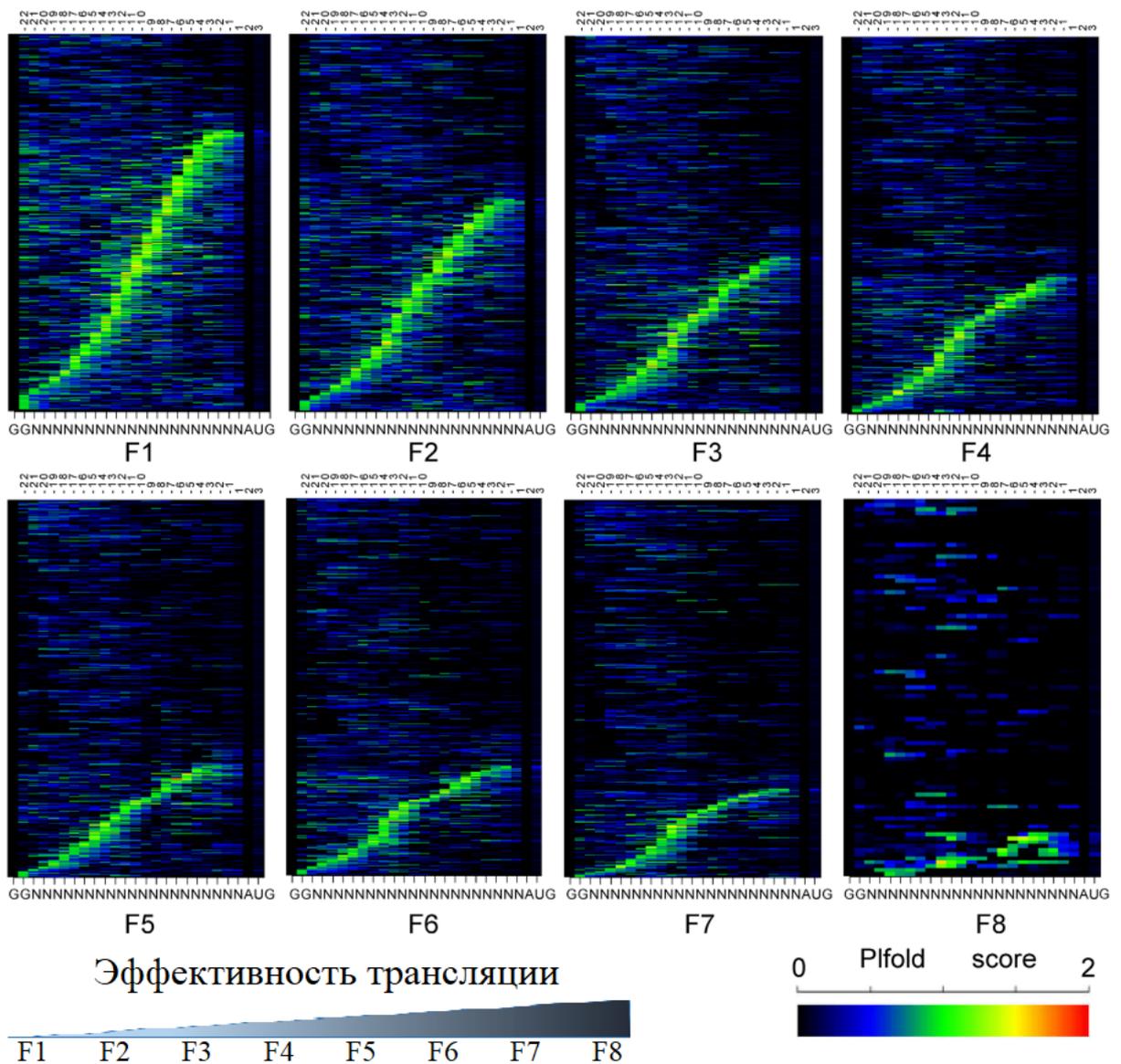


Рисунок 62. Распределение показателей спаривания оснований (plfold score) в 3'-направлении каждого взятого нуклеотида 5'-UTR из библиотеки с участком рандомизации 20N с различной эффективностью трансляции. Каждая панель соответствует фракции, отмеченной внизу. Фракции отсортированы по возрастанию эффективности трансляции: от самой низкой F1 до самой высокой F8. Позиции относительно старт-кодона (AUG) в последовательности мРНК показаны вверху каждой панели. Последовательность 5'-UTR с 20N-фрагментами рандомизации схематически изображена внизу панели. Цветные полосы соответствуют индивидуальным мРНК, которые сгруппированы по положению максимального значения показателя pifold score. Цвет полосок отражает значение показателя pifold score образования пары оснований в 3'-направлении, изменяясь от чёрного (когда спаривание маловероятно) до красного (спаривание весьма вероятно) согласно приведённой внизу справа шкале изменения значений pifold score и соответствующих им цветов. Аналогичные распределения в 5'-направлении, а также в обоих направлениях для библиотеки 30N построены, но не приведены в работе.

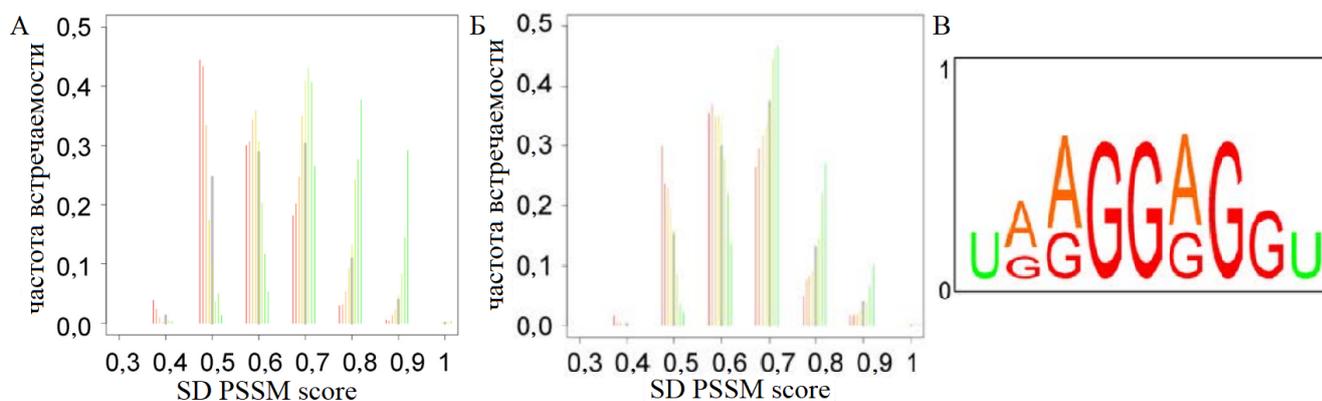


Рисунок 63. Частота встречаемости SD-подобных подпоследовательностей в 5'-UTR с участками рандомизации 20N и 30N, отличающихся по эффективности трансляции мРНК. А. Распределение максимальных оценок сходства с канонической SD (SD score) по позиционно-специфической матрице оценок (PSSM) по оси X и частоты встречаемости SD-подобных последовательностей по оси Y в 5'-UTR из библиотеки 20N с разной эффективностью трансляции. Цвета столбцов отражают эффективность трансляции от самой эффективной (зелёный цвет) до наименее эффективной (красный цвет). Высота каждого столбца соответствует частоте мотивов с определённой оценкой сходства с SD PSSM во фракции. Б. То же, что и в А, только для библиотеки 30N. В. Оценка канонической SD по матрице PSSM [250], используемой в данном исследовании. Частота встречаемости нуклеотидов представлена высотой буквы, обозначающей один из 4-х нуклеотидов.

В результате, как и ожидалось, наборы мРНК с более высокой эффективностью трансляции содержали последовательности, в среднем более близкие к последовательности SD.

Построение сходства с последовательностью SD для каждой позиции «скользящего окна» для всех последовательностей 5'-UTR из полученного набора данных (Рис. 64) выявило устойчивое увеличение частоты SD-подобных последовательностей при переходе от плохо транскрибируемых к эффективно транскрибируемым мРНК. Более того, в вариантах 5'-UTR мРНК, обеспечивающих низкий уровень экспрессии, SD-подобные последовательности в среднем могут чаще обнаруживаться за пределами их предпочтительного местоположения в природных мРНК, в то время как 5'-UTR мРНК, приводящие к высокому уровню синтезированного белка, обогащены благоприятно расположенными последовательностями SD. Если сравнивать панель F1, показывающую равномерное распределение нескольких SD-подобных последовательностей, с панелью F8, то можно заметить, что большинство SD-подобных последовательностей на последней расположено в оптимальных позициях в 5'-UTR относительно старт-кодона (Рис. 64).

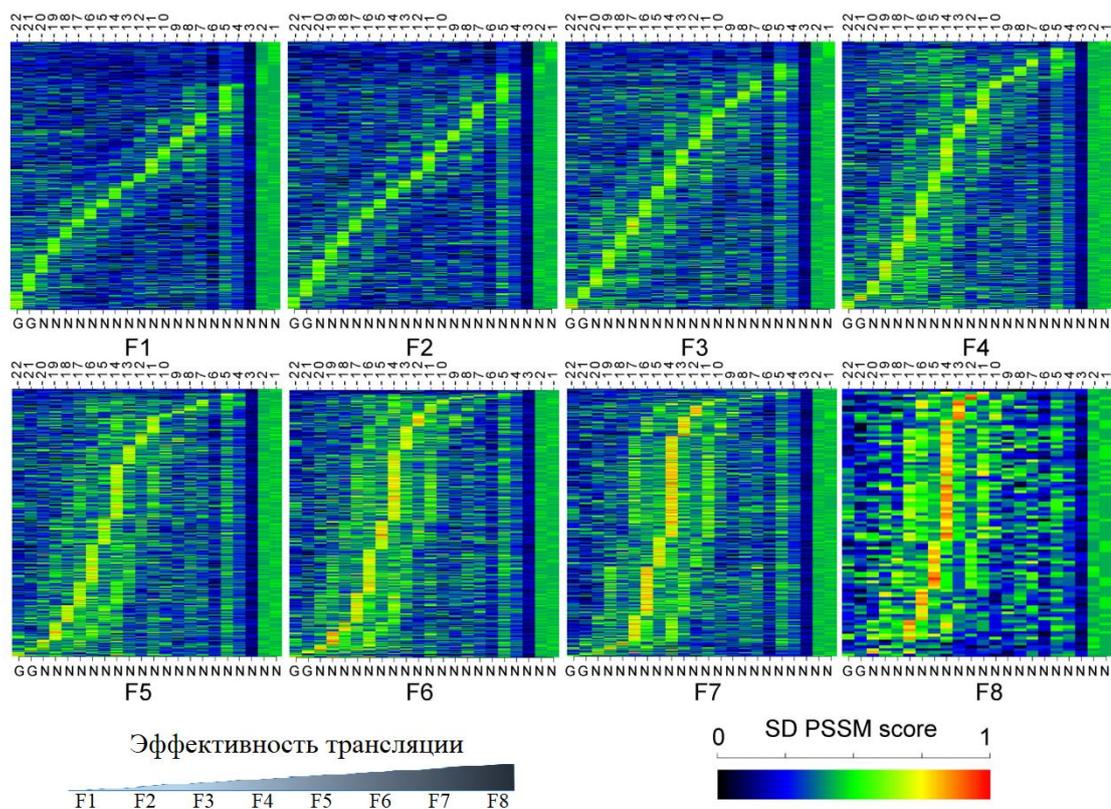


Рисунок 64. Распределение по позициям оценок SD-подобных подпоследовательностей по матрице PSSM (SD PSSM score) в 5'-UTR, отличающихся по эффективности трансляции мРНК, из библиотеки 20N. Каждая панель соответствует фракции, отмеченной внизу. Фракции отсортированы по возрастанию эффективности трансляции от самой низкой F1 до самой высокой F8. Позиции относительно последовательности мРНК показаныверху каждой панели. Последовательность 5'-UTR с 20N-фрагментами рандомизации схематически изображена внизу панели. Цветные полосы соответствуют индивидуальным мРНК, сгруппированным по положению максимального значения оценки SD PSSM score. Цвет полосок отражает сходство подпоследовательности, начинающейся с определённого нуклеотида до последовательности SD, изменяясь от чёрного (нет сходства) до красного (максимальное сходство) согласно приведённой внизу справа шкале изменения значений SD PSSM score и соответствующих им цветов. Аналогичное распределение построено для библиотеки 30N, но не представлено в работе.

Неожиданно, что среди эффективно транслируемых мРНК чаще встречаются 5'-UTR, содержащие несколько SD-подобных последовательностей. Это видно из общего увеличения мотивов, частично совпадающих с последовательностью SD, вдоль эффективно транслируемых 5'-UTR в дополнение к одной сильной последовательности SD. Однако точная количественная оценка этого явления в целом представляется затруднительной. Данный эффект можно частично объяснить тем, что последовательность SD содержит повтор мотива AGG, поэтому она может образовывать дуплексы с 3'-концевым участком 16S рРНК в трёх относительных ориентациях.

Эта особенность может объяснить очевидное присутствие двух дополнительных вторичных максимумов сходства с SD, окружающих основной максимум, который чётко виден на Рисунке 64. Однако, даже если вычесть этот эффект из наблюдаемых экспериментальных данных, всё равно будет видно, что множественные последовательности SD в среднем дают более сильное увеличение эффективности трансляции, чем одиночная SD. Схожее аддитивное влияние множественных SD-подобных последовательностей в 5'-UTR на эффективность трансляции наблюдалось и при рандомизации спейсерного участка 4N.

## 2.5. Нестандартные 5'-UTR с высокой эффективностью трансляции

В то время как все известные детерминанты в составе 5'-UTR: стабильность вторичной структуры, SD-подобные последовательности, AU-богатые энхансеры на 5'-конце, как показано, влияют на эффективность трансляции, но только их наличие не может объяснить всей наблюдаемой изменчивости в уровнях экспрессии по всему полученному набору данных. Для того, чтобы идентифицировать новые энхансеры трансляции помимо SD и AU-богатых, для подробного мутационного анализа выбрали три 5'-UTR с удивительно высокой эффективностью трансляции и минимальными SD-подобными последовательностями из библиотек 20N и 30N.

Среди отобранных вариантов 5'-UTR два были длиной 32 нт: A-богатая 5'-UTR и 5'-UTR с короткой SD, а третий – AG-богатая 5'-UTR – длиной 22 нт. С целью изучить их подробно, на основе вектора pRFPCER BamHI создали репортёрные конструкции, включающие один из данных вариантов 5'-UTR перед геном флуоресцентного белка CER, тогда как перед геном белка RFP сохранялась консервативная 5'-UTR длиной 17 нт. В качестве контроля для сравнения взяли тот же вектор, но с константными последовательностями 5'-UTR, одинаковыми перед обоими генами репортёрных белков CER и RFP и совпадающими с той, что присутствует перед геном *rfp* в описанных выше плаزمидях. Значения эффективности трансляции, измеренные *in vivo* как соотношение интенсивностей флуоресценции CER/RFP и предсказанные *in silico* с помощью RBS Calculator, оценивающего силу взаимодействия 30S комплекса с мРНК-транскриптами с различными вариантами 5'-UTR [149-150], нормировались на соответствующие значения для контроля (Таблицы 4-6).

Для контрольного и трёх отобранных вариантов 5'-UTR вместе с начальным участком кодирующей области гена *cer* построили вторичные структуры (Рис. 65). В них можно заметить присутствие шпилечных структур различной длины в случае трёх вариантов 5'-UTR, что, как известно из литературы, не способствует увеличению эффективности трансляции. Энергии этих структур сильно различаются: -1, -7,5, -0,2 ккал/моль соответственно для вариантов 5'-UTR: A-богатой, с короткой SD, AG-богатой. В то же время для контрольного варианта не наблюдается ни одной подобной структуры в 5'-UTR, однако, он уступает остальным по значениям

эффективности трансляции, измеренным *in vivo* (Таблицы 4-6). При оценке значений, предсказанных *in silico* с помощью RBS Calculator (Таблицы 4-6), можно отметить, что они ниже контрольного в случае вариантов А-богатой 5'-UTR и 5'-UTR с короткой SD, а для АG-богатой 5'-UTR данное значение лишь в 2 раза больше, чем у контроля, тогда как измеренные *in vivo* соотношения CER/RFP в 4,3-6 раз больше контрольного для всех трёх отобранных вариантов.

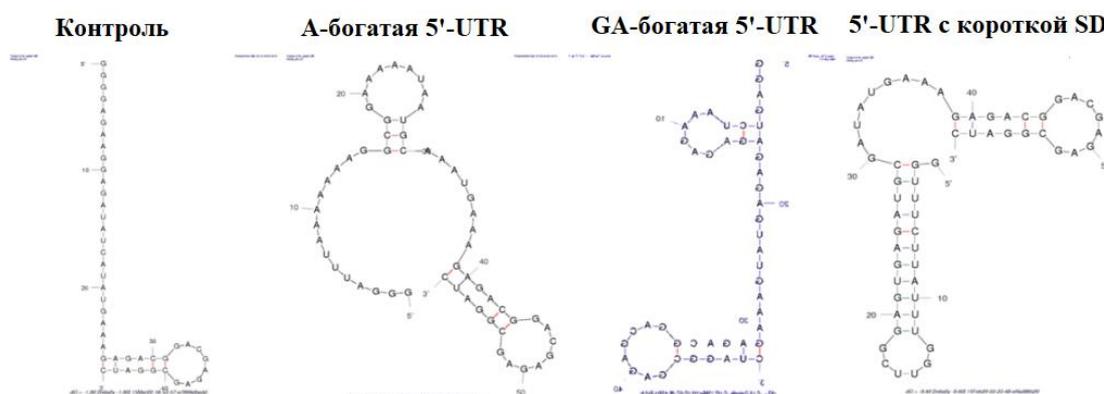


Рисунок 65. Вторичные структуры контрольного и отобранных вариантов 5'-UTR. Для укладки структур с помощью программы Mfold брали целиком 5'-UTR и начальный участок кодирующей области гена *cer* длиной 25 нт.

Далее рассмотрим результаты подробного мутационного анализа для всех отобранных вариантов 5'-UTR. Первый из них представлен А-богатой 5'-UTR, найденной среди эффективно транскрибируемых мРНК, принадлежащих к фракции F8 с наибольшим соотношением CER/RFP, а следовательно и наибольшей оцениваемой нами эффективностью трансляции *in vivo* (Таблица 4, А-богатая 5'-UTR). С другой стороны, данный вариант содержит последовательность SD длиной не более 4 нт. Измеренная эффективность трансляции в виде соотношения CER/RFP такой репортёрной конструкции оказалась в 6 раз выше, чем у контрольной (Таблица 4, Контроль). Полученные данные соответствуют ожиданиям, основанным на результатах эксперимента с применением Flow-seq: контрольная плазмида преимущественно определялась во фракции F5, а конструкция с вариативной 5'-UTR перед геном *cer* – в F8.

Последовательность А-богатой 5'-UTR имеет в составе 50% аденозинов, АU-богатый участок на 5'-конце и содержит две короткие последовательности SD (AAGG и GGA), которые расположены на большем, чем оптимальное, расстоянии от старт-кодона, и дополнительный инициаторный кодон AUG, находящийся в спейсерной области и в рамке считывания с геном *cer*. Мутация дополнительного старт-кодона снижает эффективность трансляции примерно в 2 раза (Таблица 4, А-богатая 5'-UTR без AUG), указывая на то, что предшествующий кодон AUG может использоваться для инициации трансляции в дополнение к стартовому кодону гена *cer*. Замена более короткой последовательности SD GGA на AAA (Таблица 4, А-богатая 5'-UTR -2G)

приводит к 1,3-кратному падению эффективности трансляции, свидетельствуя об умеренном вкладе этой добавочной SD в уровень экспрессии белка CER. В результате мутации обеих последовательностей SD и дополнительного кодона AUG наблюдается 16-кратное снижение эффективности биосинтеза белка CER (Таблица 4, А-богатая 5'-UTR -5G). Следует отметить, что последний мутант обладает не более, чем двумя нуклеотидами, комплементарными 3'-концевому участку 16S рРНК, но его эффективность трансляции всё же на порядок выше, чем у других репортёрных мРНК с SD длиной 2 нт (Таблица 4, конструкция 2/7) [137]. Таким образом, высокая эффективность трансляции данной мРНК с А-богатой 5'-UTR может быть объяснена аддитивным эффектом двух последовательностей SD, одна из которых вносит больший вклад, чем другая, дополнительного кодона AUG и ещё не известных элементов последовательности 5'-UTR, которые повышают уровень биосинтеза белка в дополнение к перечисленным выше.

Таблица 4. Эффективность трансляции мРНК с А-богатой 5'-UTR и её мутированных вариантов.

Название конструкции	Последовательность 5'-UTR	Относительная эффективность трансляции*	Предсказанная эффективность трансляции**
Контроль	GGAGAAGGAGAUUCAU	1	1
А-богатая 5'-UTR	GGGAUUUAAAAA <u>AAGGCGG</u> AAAA AUA <b>AUG</b> CAU	6,02	0,68
А-богатая 5'-UTR без AUG	GGGAUUUAAAAA <u>AAGGCGG</u> AAAA AUAAUACAУ	2,59	0,82
А-богатая 5'-UTR -2G	GGGAUUUAAAAA <u>AAGG</u> CAAAAA AUA <b>AUG</b> CAU	4,50	0,68
А-богатая 5'-UTR -5G	GGGAUUUAAAAAAAACAAAAAA AUAAUACAУ	0,38	1,57
2/7	CACACAACACCU <u>G</u> AUCAACU	0,03	0,14

\* Относительная эффективность трансляции оценивалась измеренным соотношением флуоресценции двух белков CER/RFP, одна мРНК содержала один из представленных вариантов 5'-UTR перед геном *cer*, а мРНК с закодированным белком RFP – контрольную 5'-UTR.

\*\* Предсказанная относительная эффективность трансляции с помощью RBS Calculator на основе известных особенностей, влияющих на трансляцию, для мРНК с представленными вариантами 5'-UTR перед геном *cer*, нормированная на предсказанное значение для контроля. Последовательность SD выделена подчёркиванием, дополнительный кодон AUG – красным, внесённые замены – жирным шрифтом.

Следует отметить, что наблюдаемая эффективность трансляции мРНК с А-богатой 5'-UTR и её мутантов не может быть объяснена и только особенностями, которые, как известно, важны для инициации трансляции и учтены при оценке эффективности на основе предсказаний с помощью RBS Calculator (Таблица 4, последний столбец). В то же время А-богатая 5'-UTR имеет заметно уменьшенное количество остатков цитозина (всего 2 нт), как и многие 5'-UTR с высокой эффективностью трансляции (Рис. 60).

Другая необычно эффективная мРНК включает последовательность SD длиной 4 нт (Таблица 5, 5'-UTR с короткой SD), расположенную на расстоянии 13 нт (если считать от центрального G в SD) от старт-кодона AUG. Аналогичная синтетическая мРНК (Таблица 5, конструкция 4/13) имеет на 2 порядка более слабый уровень трансляции. В 5'-UTR с короткой SD внесли несколько мутаций (Таблица 5) для определения областей, влияющих на биосинтез белка CER. С целью проверить возможность того, что необычно высокий уровень трансляция такой мРНК связан с наличием дополнительного кодона GUG в рамке, который может выступать как потенциальный инициаторный кодон, его преобразовали в GUC и GCG (Таблица 5, 5'-UTR с короткой SD -GUC и 5'-UTR с короткой SD -GCG соответственно). В результате этих замен наблюдалось отсутствие или незначительное снижение эффективности трансляции, что говорит скорее о незначительности этого кодона GUG в определении уровня экспрессии в данном случае.

Мутации последовательности SD (Таблица 5, 5'-UTR с короткой SD -SD) практически приводят к отмене трансляции данной мРНК. Однако изменения 5'-проксимальной области, оставляющие нетронутой SD, также снижают уровень биосинтеза белка в той же степени (Таблица 5, 5'-UTR с короткой SD -9UA). Даже замена трёх остатков уридина, предшествующих SD, значительно подавляет трансляцию (Таблица 5, 5'-UTR с короткой SD -3U). Мутации в спейсерном участке между последовательностью SD и старт-кодом лишь умеренно влияют на уровень синтезированного белка (Таблица 5, 5'-UTR с короткой SD -3'). Перемещение SD ближе к стартовому кодону (Таблица 5, 5'-UTR с короткой SD +SD), что оказалось очень полезным для трансляции синтетической мРНК с SD длиной 4 нт (Таблица 5, конструкция 4/7), уменьшает эффективность трансляции более, чем в 5 раз.

Неожиданно высокая эффективность трансляции данной 5'-UTR может быть объяснена наличием комбинации ранее не охарактеризованных энхансеров трансляции. Примечательно, что уровень биосинтеза белка CER очень чувствителен к мутациям в этой 5'-UTR. Не только наличие, но и расположение SD оказалось важным для её эффективности трансляции. Самым необычным открытием является то, что расположение SD, удалённое от старт-кодона и неблагоприятное для других 5'-UTR, в этом случае выгодно для трансляции. U-богатая последовательность, которая расположена перед отдалённой SD, как показано, является энхансером биосинтеза белка согласно эффективности трансляции мутантов 5'-UTR с короткой SD -3U и 5'-UTR с короткой SD -9UA.

Таблица 5. Эффективность трансляции мРНК с 5'-UTR с короткой SD и её мутантов.

Название конструкции	Последовательность 5'-UTR	Относительная эффективность трансляции*	Предсказанная эффективность трансляции**
Контроль	GGAGA <u>AAGGAG</u> AUAUCAU	1	1
5'-UTR с короткой SD	GGUUUCUUAUUUGGUUC <u>GGA</u> <u>GUG</u> GAGAUGCGAU	4,31	0,68
5'-UTR с короткой SD -SD	GGUUUCUUAUUUGGUU <u>CCCU</u> CUGAGAUGCGAU	0,076	0,13
5'-UTR с короткой SD -9UA	GG <u>CCCCCG</u> CCCGGUUC <u>GGA</u> <u>GUG</u> GAGAUGCGAU	0,073	1,18
5'-UTR с короткой SD -3U	GGUUUCUUA <u>CCCG</u> GUUC <u>GGA</u> <u>GUG</u> GAGAUGCGAU	0,51	0,65
5'-UTR с короткой SD -3'	GGUUUCUUAUUUGGUUC <u>GGA</u> <u>GUC</u> ACA <u>UCC</u> CAU	2,35	0,18
5'-UTR с короткой SD +SD	GGUUUCUUAUUUGGUU <u>CCAC</u> A <u>UGGAG</u> UCCCAU	0,41	0,92
5'-UTR с короткой SD -GUC	GGUUUCUUAUUUGGUUCGGA GUCAGAUGCGAU	3,21	0,11
5'-UTR с короткой SD -GCG	GGUUUCUUAUUUGGUUCGGA GCGAGAUGCGAU	4,32	0,56
4/13	CAC <u>CCGGAG</u> CAACAACAACU	0,06	0,04
4/7	CACACAAC <u>CCGGAG</u> CAACU	0,8	0,22

\* Относительная эффективность трансляции оценивалась измеренным соотношением флуоресценции двух белков CER/RFP, одна мРНК содержала один из представленных вариантов 5'-UTR перед геном *cer*, а мРНК с закодированным белком RFP – контрольную 5'-UTR.

\*\* Предсказанная относительная эффективность трансляции с помощью RBS Calculator на основе известных особенностей, влияющих на трансляцию, для мРНК с представленными вариантами 5'-UTR перед геном *cer*, нормированная на предсказанное значение для контроля.

Последовательность SD выделена подчёркиванием, дополнительный кодон AUG – красным, внесённые замены – жирным шрифтом.

Третий вариант 5'-UTR с неожиданно высокой эффективностью трансляции содержит подряд шесть AG-повторов (Таблица 6, AG-богатая 5'-UTR). Дополнительно 5'-UTR включает последовательность SD состава GGAG, расположенную на большом расстоянии, далёком от

оптимального: 20 нт от старт-кодона AUG, поэтому функциональность такой SD маловероятна. Клетки с данной репортёрной конструкцией обнаружены во фракции F7, второй по уровню эффективности трансляции. В соответствии с результатом Flow-seq при клонировании данной последовательности 5'-UTR перед геном *cer* в индивидуальную репортёрную конструкцию (Таблица 6, AG-богатая 5'-UTR) она транслировалась в 5 раз эффективнее, чем контрольная мРНК (Таблица 6, Контроль).

Таблица 6. Эффективность трансляции мРНК с AG-богатой 5'-UTR и её мутантов.

Название конструкции	Последовательность 5'-UTR	Относительная эффективность трансляции*	Предсказанная эффективность трансляции**
Контроль	GGAGAAGGAGAUUAUCAU	1	1
AG-богатая 5'-UTR	<u>GGAGUCU</u> AAAGAGAGAGAGAGU	5,09	2,05
AG-богатая 5'-UTR -AG <sub>1-3</sub>	<u>GGAGUCU</u> AAACACACAGAGAGU	0,15	0,13
AG-богатая 5'-UTR -AG <sub>4-6</sub>	<u>GGAGUCU</u> AAAGAGAGACACACU	1,40	0,58
AG-богатая 5'-UTR +SD1	<u>GGAGUCU</u> AAAGAG <u>GGAGA</u> AGU	1,52	4,18
AG-богатая 5'-UTR +SD2	<u>GGAGUCU</u> AAAG <u>GGAGG</u> AGAGAGU	1,56	12,37
AG-богатая 5'-UTR -G	<u>GGAGUCU</u> AAAGAGACAGAGAGU	1,14	0,31
4/16	CC <u>GGAGC</u> ACACACAACAACU	0,02	0,04

\* Относительная эффективность трансляции оценивалась измеренным соотношением флуоресценции двух белков CER/RFP, одна мРНК содержала один из представленных вариантов 5'-UTR перед геном *cer*, а мРНК с закодированным белком RFP – контрольную 5'-UTR.

\*\* Предсказанная относительная эффективность трансляции с помощью RBS Calculator на основе известных особенностей, влияющих на трансляцию, для мРНК с представленными вариантами 5'-UTR перед геном *cer*, нормированная на предсказанное значение для контроля. Последовательность SD выделена подчёркиванием, внесённые замены – жирным шрифтом.

Мутации, изменяющие первые (Таблица 6, AG-богатая 5'-UTR -AG<sub>1-3</sub>) и последние (Таблица 6, AG-богатая 5'-UTR -AG<sub>4-6</sub>) три AG-повтора, приводили к 34 и 3,6-кратному снижению эффективности трансляции соответственно. Возможно, что последовательность GAG,

обнаруженная в AG-повторах, выступает в качестве последовательности SD. С целью создать более сильные SD состава AGGAG и GGAGG мутировали соответственно проксимальную (Таблица 6, AG-богатая 5'-UTR +SD1) и дистальную (Таблица 6, AG-богатая 5'-UTR +SD2) части последовательности 5'-UTR с AG-повторами. Удивительно, но введение SD в любое из этих двух положений снижает эффективность трансляции примерно в 3,3 раза (Таблица 6). Более того, даже удаление одного G в середине области с AG-повторами (Таблица 6, AG-богатая 5'-UTR -G) уменьшает уровень биосинтеза белка в 4,5 раза. В результате получается, что участок с AG-повторами служит уникальным энхансерным элементом трансляции. Даже после удаления наиболее функционально значимой его части (Таблица 6, AG-богатая 5'-UTR -AG<sub>1-3</sub>) остаточный уровень синтезированного белка значительно выше, чем у синтетической репортёрной мРНК, содержащей сопоставимую последовательность SD (Таблица 6, конструкция 4/16).

С целью исключить возможность того, что необычно высокая эффективность трансляции является артефактным результатом взаимодействия между 5'-UTR и кодирующей областью гена *cer*, то есть путём формирования вторичной структуры большой протяжённости, кодирующие области генов *cer* и *rfp* поменяли. Таким образом, созданная репортёрная плазмида pRFPCER была использована для вставки последовательностей необычно эффективных 5'-UTR: А-богатой, с короткой SD и AG-богатой, и нескольких их мутантных вариантов теперь перед репортёрным геном *rfp*, а перед геном *cer* сохранялась константная контрольная 5'-UTR (Таблица 7). В результате обнаружили, что эффективности трансляции имеют одинаковые тенденции как для репортёрного гена *cer*, так и для гена *rfp*. «Переключение» кодирующих областей генов *cer* и *rfp* даже усиливает наблюдаемые различия между контролем, высоко эффективными 5'-UTR и их мутантными формами.

Далее провели поиск по вариантам последовательностей 5'-UTR природных генов *E. coli*, чтобы посмотреть, есть ли возможность идентифицировать похожие AG-богатые энхансеры трансляции, которые присутствуют на месте последовательности SD. Среди генов, содержащих AG-богатые вариации 5'-UTR вместо SD, нашли гены *ybaB* (кодирующий консервативный ДНК-связывающий белок, описанный сайт связывания рибосомы представлен последовательностью GAGAGAGAAACCUAUG, старт-кодон выделен красным), *lon* (кодирующий протеазу, RBS приведён в виде последовательности GAGAGAGCUCUAUG), *rplY* (кодирующий рибосомный белок bL25, RBS – AGAGAGAAAGAAUG) и *lrp* (кодирующий транскрипционный регулятор, RBS – AGAGAGACAAUAAUAUG), тогда как другие гены имели более короткие AG-фрагменты. Полные последовательности 5'-UTR генов *ybaB*, *lon*, *rplY*, *lrp* встроили перед геном репортёрного белка CER в вектор pRFPCER (Таблица 8). В случае, если для гена аннотированы несколько вариантов 5'-UTR, использовали самую короткую аннотированную форму.

Таблица 7. Эффективность трансляции мРНК с отобранными 5'-UTR и их мутированными вариантами в составе репортёра с «переключенными» кодирующими областями генов *cer* и *rfp*.

Название конструкции	Последовательность 5'-UTR	Относительная эффективность трансляции* (RFP/CER)	Относительная эффективность трансляции** (CER/RFP)
Контроль	GGAGAAGGAGAUUAUCAU	1	1
А-богатая 5'-UTR	GGGAUUUAAAAA <u>AAGGC</u> GGAAA AAUA <u>AUG</u> CAU	15,6	6,02
А-богатая 5'-UTR -5G	GGGAUUUAAAAAAAACAAAAA AAUAAUACAU	0,27	0,38
5'-UTR с короткой SD	GGUUUCUUAUUUGGUUC <u>GGAGU</u> <u>G</u> AGAUGCGAU	14	4,31
5'-UTR с короткой SD -9UA	GGCCCCCGCCCGGUUC <u>GGAGU</u> <u>G</u> AGAUGCGAU	0,002	0,073
AG-богатая 5'-UTR	<u>GGAGUC</u> UAAAGAGAGAGAGAGU	6,1	5,09
AG-богатая 5'-UTR -AG <sub>1-3</sub>	<u>GGAGUC</u> UAAACACACAGAGAGU	0,02	0,15

\* Относительная эффективность трансляции репортёрной конструкции с «переключенными» кодирующими областями генов *cer* и *rfp* оценивалась измеренным соотношением флуоресценции двух белков RFP/CER, одна мРНК содержала один из представленных вариантов 5'-UTR перед геном *rfp*, а мРНК с закодированным белком CER – контрольную 5'-UTR.

\*\* Относительная эффективность трансляции исходной конструкции оценивалась измеренным соотношением флуоресценции двух белков CER/RFP, одна мРНК содержала один из вариантов 5'-UTR перед геном *cer*, а мРНК с закодированным белком RFP – контрольную 5'-UTR.

Последовательность SD выделена подчёркиванием, дополнительный кодон AUG – красным, внесённые замены – жирным шрифтом.

С целью оценить значимость AG-фрагментов для эффективности трансляции AG-повторы заменялись AC-повторами и отдельно в параллель мутировали AG-фрагменты на сильные последовательности SD (Таблица 8). Следует отметить, что наблюдаемая эффективность трансляции мРНК с природными AG-богатыми 5'-UTR и их мутированными формами в целом согласуется с данными, полученными на основе предсказаний с помощью RBS Calculator (Таблица 8, последний столбец). Несмотря на то, что для всех протестированных вариантов природных AG-богатых 5'-UTR обнаружили, что внесённая последовательность SD лучше

способствует трансляции, чем встречающиеся в природе AG-повторы (Таблица 8, сравниваем конструкции lon и lon+SD, lrp и lrp+SD, ybaB и ybaB+SD, rplY и rplY+SD), мутации, приводящие к удалению таких AG-повторов во всех проверенных случаях приводили к снижению на порядок уровня биосинтеза белка (Таблица 8, сравниваем конструкции lon и lon-AG, lrp и lrp-AG, ybaB и ybaB-AG, rplY и rplY-AG). Как видно, эффективность трансляции мРНК с 5'-UTR, содержащей AG-повторы, по сравнению с 5'-UTR, включающей последовательность SD, зависит от контекста. При этом неизменно варианты с AG-повторами лучше стимулировали биосинтез белка, чем аналогичные последовательности 5'-UTR без данного элемента. Таким образом, можно предложить рассматривать их как энхансеры трансляции, зависящие от контекста.

Таблица 8. Эффективность трансляции мРНК с природными AG-богатыми 5'-UTR и их мутантов.

Название конструкции	Последовательность 5'-UTR	Относительная эффективность трансляции*	Предсказанная эффективность трансляции**
Контроль	GGAGAAGGAGAUUCAU	1	1
lon	UAAAACUAAGAGAGAGCUCU	0,25	0,36
lon-AG	UAAAACUAACACACAGCUCU	0,03	0,06
lon+SD	UAAAACUA <b>AAGGAGG</b> AGCUCU	1,84	3,14
lrp	AAUACAGAGAGACAAUAAU	1,04	0,67
lrp-AG	AAUACACACACACAAUAAU	0,13	0,12
lrp+SD	AAUAC <b>AGGAGG</b> ACAAUAAU	5,66	1,91
ybaB	CGUGAUUGAGAGAGAAACCU	0,8	2,86
ybaB-AG	CGUGAUUCACACAGAAACCU	0,04	0,32
ybaB+SD	CGUGAUU <b>GGAGG</b> AGAAACCU	4,84	9,31
rplY	UAAAACUAAGAGAGAGCUCU	0,19	0,36
rplY-AG	UAAAACUAACACACAGCUCU	0,03	0,06
rplY+SD	UAAAACUA <b>AAGGAGG</b> AGCUCU	1,81	3,14

\* Относительная эффективность трансляции оценивалась измеренным соотношением флуоресценции двух белков CER/RFP, одна мРНК содержала один из представленных вариантов 5'-UTR перед геном *cer*, а мРНК с закодированным белком RFP – контрольную 5'-UTR.

\*\* Предсказанная относительная эффективность трансляции с помощью RBS Calculator на основе известных особенностей, влияющих на трансляцию, для мРНК с представленными вариантами 5'-UTR перед геном *cer*, нормированная на предсказанное значение для контроля. Последовательность SD выделена подчёркиванием, внесённые замены – жирным шрифтом.

Суммируя полученные результаты при анализе выбранных последовательностей 5'-UTR и их мутированных вариантов, следует отметить несколько моментов. Во-первых, не все наблюдаемые эффекты можно объяснить, исходя только из стабильности вторичных структур вблизи старт-кодона или наличия, либо отсутствия последовательности SD оптимальной длины и на оптимальном расстоянии от инициаторного кодона. Во-вторых, отсутствие вторичной структуры в 5'-UTR не всегда обеспечивает высокую эффективность трансляции, если рассматривать различные варианты 5'-UTR относительно контрольной конструкции, у которой как раз этой структуры не наблюдается в 5'-UTR, тогда как для остальных вариантов отмечается её наличие. Возможно, что рибосоме иногда легче узнать небольшую шпилечную структуру на небольшом расстоянии в 5'-UTR от старт-кодона, связаться с ней и, не тратя энергию на её разворачивание, приступить к элонгации. В-третьих, такие особенности 5'-UTR, как AU-богатый участок (как и преимущественно U-богатая последовательность) на 5'-конце и нуклеотидный состав среднего участка, оказывают значительное влияние на эффективность трансляции. Это и не удивительно, так как с первой из них взаимодействует белок bS1 30S субчастицы рибосомы, который участвует в инициации трансляции неканонических мРНК, а вторая служит основным местом посадки рибосомы, после связывания с которым, она приступает как можно скорее к элонгации. Участок с AG-повторами в сайте посадки рибосомы может служить уникальным энхансерным элементом трансляции. Также можно отметить возможный аддитивный эффект нескольких последовательностей SD, дополнительного кодона AUG, приводящий к увеличению уровня биосинтеза белка. И наконец, в ходе мутирования наблюдается регулярное снижение эффективности трансляции при внесении в состав 5'-UTR большего количества цитидинов.

## **2.6. Природные 5'-UTR *E.coli* перед геном *cer* с высокой эффективностью трансляции**

Для изучения влияния известных особенностей мРНК, таких как переменные стартовые кодоны, последовательности SD с различными длиной и расположением от старт-кодона, элементы вторичной структуры и AU-богатые последовательности, на эффективность трансляции использовали двойную репортёрную конструкцию, основанную на генах двух флуоресцентных белков CER и RFP (Рис. 44) [156]. Ранее мы заметили, что большинство природных мРНК *E. coli* имеет относительно короткие последовательности SD, чаще всего длиной 4 нт, в 5'-UTR. Синтетические репортёрные мРНК, содержащие 5'-UTR с последовательностью SD со схожими длиной и расположением относительно стартового кодона, транслировались на среднем уровне в описанной системе [137]. Кроме того, идентификация AG-богатого энхансера трансляции в наборе синтетических и природных 5'-UTR послужила стимулом, чтобы проанализировать уровень эффективности трансляции большей выборки последовательностей природных 5'-UTR.

Для этого клонировали одни из самых коротких по данным базы RegulonDB 5'-UTR следующих 48 генов перед геном репортёрного белка CER: *tabA*, *ycdY*, *slyA*, *gmhB*, *yciN*, *ivy*, *yihD*, *pepD*, *yciI*, *cysK*, *cpxR*, *mipA*, *yrbL*, *gpmA*, *kdgR*, *adk*, *yefM*, *slyD*, *mprA*, *yobF*, *tig*, *glnS*, *mgrB*, *rpoZ*, *deoB*, *eno*, *rbsD*, *ubiD*, *gpsA*, *lrhA*, *lpxD*, *gyrA*, *malP*, *nfuA*, *ilvL*, *hisL*, *gapA*, *yccA*, *gmk*, *ribF*, *orn*, *fadR*, *ygiW*, *ymcE*, *yqiC*, *yoaB*, *tdk*, *dsrB* (Приложение 3). Данный набор охватывает диапазон длин последовательностей SD от 2 до 7 нт и расстояние от центрального гуанозина, например, в последовательности aagGagg он выделен как G, до старт-кодона от 1 до 17 нт. Эффективности трансляции измеряли как соотношение интенсивностей флуоресценции белков CER/RFP (Приложение 3). В то время как общий диапазон уровней биосинтеза белка сопоставим для природных и синтетических 5'-UTR, для синтетических вариантов с очень высоким уровнем экспрессии необходимо наличие длинных последовательностей SD от 6 до 8 нт, при этом некоторые природные 5'-UTR содержали последовательность SD длиной 3-5 нт, демонстрируя высокую эффективность трансляции. Среди выделяющихся примеров можно привести 5'-UTR гена *tdk*, включающую SD длиной 2 нт и состава GG на расстоянии 1 нт от старт-кодона, при этом, как и ожидалось, наблюдается для неё наименьший уровень биосинтеза белка из всех, что получили для 5'-UTR перечисленных генов (Рис. 66).

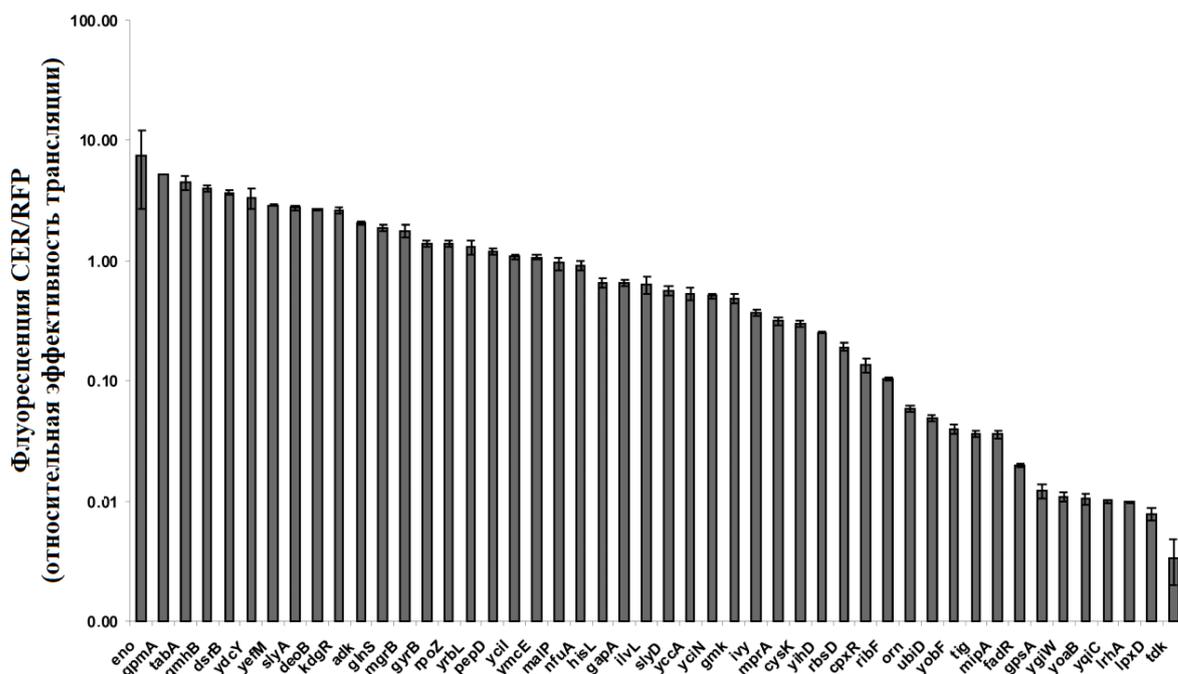


Рисунок 66. Зависимость эффективности трансляции от предшествующих гену белка CER последовательностей природных 5'-UTR, рассчитанной как соотношение интенсивностей флуоресценции белков CER/RFP, нормированное на таковое для контрольных мРНК.

Максировка сайта инициации трансляции за счёт образования вторичной структуры может резко снизить эффективность трансляции. Синтетические репортёрные мРНК созданы с таким

условием, чтобы свести к минимуму вторичную структуру сайтов связывания рибосом (Приложении 3), в то время как природные мРНК могут иметь элементы вторичной структуры, ингибирующие трансляцию. Следовательно, тогда как более низкие скорости биосинтеза белка природных 5'-UTR могут быть объяснены маскировкой сайтов инициации трансляции с помощью вторичной структуры, это не раскрывает причины повышенных уровней биосинтеза белка природных мРНК по сравнению с синтетическими последовательностями. Самой неожиданной среди протестированных природных вариантов оказалась высокоэффективная трансляция, индуцированная 5'-UTR гена *eno* (Рис. 66). Данная последовательность содержит SD длиной 4 нт, расположенную на расстоянии 9 нт перед старт-кодоном. Синтетические репортёрные мРНК, включающие элементы SD со сходными характеристиками, демонстрируют на порядок более низкие эффективности трансляции [137].

Наличие не более трёх AG-повторов отмечено в нескольких вариантах природных 5'-UTR: три повтора в 5'-UTR гена *ussA*, два – ещё в четырех последовательностях генов *yefM*, *hisL*, *yihD*, *ubiD*. Однако только 5'-UTR гена *yefM* продемонстрировал транслируемость заметно выше, чем у контроля, тогда как для всех остальных наблюдался уровень экспрессии ниже, чем у контроля. При этом для 5'-UTR генов *ussA* и *hisL* эффективность трансляции близка к контролю, а для вариантов 5'-UTR генов *yihD* и *ubiD* – она сильно ниже его (Рис. 66). Следует отметить, что значения эффективности для контроля соответствуют фракции F5 с умеренным уровнем экспрессии относительно всего диапазона  $10^4$  раз, полученного для рандомизированных библиотек 20N и 30N, но в то же время 5'-UTR контрольного репортёра содержит сильную SD оптимальной длины 6 нт на оптимальном расстоянии 9 нт от центрального G до старт-кодона.

Полученные результаты для природных 5'-UTR 48 генов, клонированных в двойную репортёрную конструкцию, ограничены разнообразием выборки и не дают возможности выявить более строгие закономерности.

Несмотря на десятилетия интенсивных исследований, до сих пор не хватает полного понимания правил, определяющих эффективность трансляции. На данном этапе мы провели анализ эффективности трансляции 48 природных 5'-UTR *E. coli* и свыше 20000 5'-UTR из рандомизированных библиотек последовательностей. Все известные особенности мРНК, влияющие на уровень биосинтеза белка, такие как последовательность SD с оптимальными длиной и расположением, отсутствие вторичной структуры и AU-богатые энхансеры трансляции, были чётко выявлены и в нашем исследовании.

В тоже время найдено много примеров 5'-UTR, как синтетических, так и природных, чью эффективность невозможно легко объяснить известными особенностями, и в целом точное предсказание эффективности трансляции произвольной 5'-UTR кажется проблематичным. Мы продемонстрировали благоприятное влияние множественных последовательностей SD и старт-

кодонов на трансляцию и нашли необычный сайт посадки рибосомы, состоящий из AG-повторов. Такие повторы также найдены в нескольких природных RBS *E. coli*, указывая на их возможную роль в качестве природных энхансеров трансляции, замещающих последовательность SD. В качестве отличительной особенности эффективных вариантов 5'-UTR отметили уменьшение доли остатков цитозина. Результаты данного исследования могут быть использованы для создания векторов для производства белков в биотехнологии и помочь разработать основы для тонкой настройки систем регуляции генов.

### **3. Изучение влияния природных 5'-UTR на эффективность трансляции**

При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором в соавторстве, в которых, согласно Положению о присуждении учёных степеней в МГУ, отражены основные результаты, положения и выводы исследования: [17], общий объём публикации – 0,5 п.л. Причём основополагающий вклад в ней принадлежит соискателю.

#### **3.1. Отбор природных 5'-UTR и создание репортёрных конструкций с ними перед стартовым кодоном гена *cer***

Исследование, выполненное для 5'-UTR 48 природных генов мы решили расширить на большую выборку и адаптировать к использованию Flow-seq. Выбор последовательностей природных 5'-UTR проводился в два этапа. Сначала решили ограничить библиотеку только 5'-UTR первых генов оперонов, принадлежащих штамму *E. coli* BW25113, чтобы избежать влияния возможных дополнительных систем регуляции. Далее исходный такой набор из 1451 варианта 5'-UTR сузили, отсекая последовательности с длиной более 60 нт из-за ограничений по синтезу олигонуклеотидов. Таким образом, библиотека, включающая 713 природных 5'-UTR длиной от 2 до 60 нт была создана, отличаясь от предыдущих рандомизированных библиотек, включающих переменные 5'-UTR в основном неизменной длины ввиду упрощения их синтеза. Полученная библиотека представлена для дальнейшей ПЦР-амплификации с использованием адапторных олигонуклеотидов, содержащих сайты эндонуклеаз рестрикции, необходимые для последующего клонирования в вектор. Выбранный диапазон включает варианты 5'-UTR с длинами 20-40 нт, как упоминалось ранее, являющимися наиболее часто встречающимися среди мРНК *E. coli* [110, 157].

Для анализа влияния на эффективность трансляции последовательностей природных 5'-UTR разной длины создана библиотека репортёрных конструкций, содержащих перед стартовым кодоном гена *cer* 713 отобранных 5'-UTR длиной от 2 до 60 нт, по следующей схеме (Рис. 67).

В качестве исходных фрагментов для последующей вставки в вектор взяли синтетически синтезированные одноцепочечные олигонуклеотиды переменной длины 40-98 нт (оц олиги) с промотором T7, различными последовательностями природных 5'-UTR длиной 2-60 нт, стар-

кодонам природных генов и первыми 15 нт белок-кодирующей области гена *cer*:

5' – **ТААТАСГАСТСАСТАТАГГГ** NN... **NDTKAAAGAGGACGAGAGC** – 3'.

После ПЦР-амплификации этих фрагментов с адапторными олигонуклеотидами, первый из которых содержит промотор T7, а второй комплементарен первым 5 кодонам гена *cer*:

5' – GTGCTC **AAGCTTTAATACGACTCACTATAGGG** – 3' и

5' – СТГГСС **GGATCCC**CGTCGCGGGTGCTCTCGTCCTCTTT – 3'.

Наряду с этим оба включают ещё и один из палиндромных сайтов узнавания эндонуклеаз рестрикции HindIII и BamHI соответственно, необходимых для последующего клонирования в вектор. Данные эндонуклеазы вносят следующие разрывы:

HindIII: 5' – **AAGCTT** – 3'

BamHI: 5' – **GGATCC** – 3'

3' – **TTCGA** – 5'

3' – **CCTAG** – 5'.

Дальнейшая подготовка проводилась путём разрезания указанными выше эндонуклеазами амплифицированных двухцепочечных олигонуклеотидов (дц олигов, 76-134 нт). В результате получены целевые дуплексы, которые после очистки использовались для встраивания в вектор. Проверка эффективности работы с олигонуклеотидами на каждой стадии проводилась путём электрофореза нуклеиновых кислот в 15% ПААГ в денатурирующих условиях.

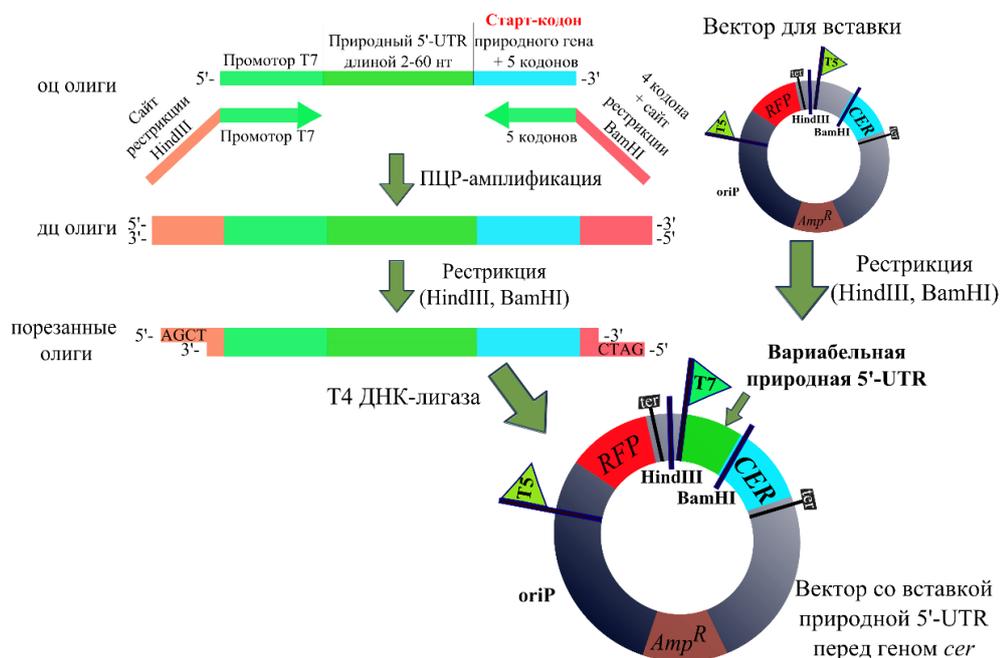


Рисунок 67. Схема создания библиотеки плазмид со вставкой природных 5'-UTR перед геном *cer*. Оц олиги – одноцепочечные олигонуклеотиды длиной 40-98 нт. Дц олиги – двухцепочечные олигонуклеотиды разной длины 76-134 нт после ПЦР-амплификации. Порезанные олиги – дц олиги после рестрикции эндонуклеазами HindIII и BamHI, используемые как целевой дуплекс для встраивания в плазмиду, порезанную то тем же сайтам. Amp<sup>R</sup> – белок устойчивости к ампициллину. oriP – ориджин репликации плазмиды. Рисунок выполнен в программе Inkscape.

Как и прежде, за основу вектора для последующего клонирования использовали плазмиду pRFPCER BamHI (Рис. 57) [137, 156], в которую внесли несколько модификаций. Полученная репортёрная конструкция содержала гены двух флуоресцентных белков, где первый (*rfp*) находился под контролем конститутивного промотора T5, а второй (*cer*) – под контролем промотора T7, причём оба гена имели свои терминаторы транскрипции (Рис. 68А). 5'-UTR гена *rfp* по-прежнему были идентичны во всех плазмидах в библиотеке, позволяя нам использовать его как внутренний контроль, тогда как 5'-UTR гена *cer*, выступающего в качестве сенсора, представлены вариативным набором природных 5'-UTR. При этом на 5'-конец последних были добавлены два основания гуанозина, чтобы обеспечить эффективную и единообразную транскрипцию этих последовательностей в составе плазмиды. Природные вставки 5'-UTR длиной не более 60 нт клонировали перед стартовым кодоном гена *cer* по сайтам рестрикции эндонуклеаз HindIII и BamHI, присутствующим также в векторе вместе с геном *amp<sup>R</sup>* белка, обеспечивающего устойчивость к ампициллину, по которому в дальнейшем проводился отбор (Рис. 68А).

С образовавшимися в результате рестрикции липкими концами плазмиды с двух концов при добавлении T4 ДНК-лигазы сшивались вышеуказанные целевые дуплексы, имеющие комплементарные им выступающие 5'-концы. Для удаления конструкций без вставок проводили рестрикцию по сайту узнавания эндонуклеазой SacII, присутствующему в исходной 5'-UTR.

В итоге полученная библиотека репортёрных плазмид содержала различные природные 5'-UTR непосредственно перед старт-кодоном гена *cer* (Рис. 67). Далее её трансформировали с помощью теплового шока в ультракомпетентные клетки штамма JM109 *E.coli*, который не содержит нуклеаз, ввиду чего подходит для наработки генно-инженерных конструкций. Затем к трансформированным клеткам добавляли жидкую среду SOC с глюкозой и перемещали в термостат на 37°C для наработки белка устойчивости к ампициллину, по которому после высевания клеток на большую чашку Петри проводилась селекция колоний. Выросшие на чашке колонии сканировали с помощью системы визуализации флуоресценции ChemiDoc (Bio-Rad) для оценки числа колоний с репортёрными конструкциями. С целью определения наличия вставки скололи 8 колоний с чашки и высеяли в жидкую среду LB с ампициллином. Затем из полученных ночных культур экстрагировали плазмиды и проверяли их секвенированием по Сэнгеру. Остальные колонии смывали с чашки с помощью LB и выделяли из клеток библиотеки конструкций с помощью набора для выделения плазмидной ДНК.

Далее выделенными библиотеками конструкций наряду с индивидуальными плазмидами трансформировали клетки штамма BL21 (DE3) *E.coli* для последующей сортировки (Рис. 68Б). В данном штамме присутствует T7 РНК-полимераза, чей ген находится под контролем lac-промотора, которая сможет узнать T7 промотор в нашем наборе плазмид. Для амплификации библиотеки конструкций трансформированные клетки растили в течение 16 ч при температуре

37°C на чашке Петри с агаризованной средой LB с ампициллином. Затем чашку с клетками также сканировали для оценки эффективности трансформации, а после смывали колонии 2 мл среды LB. Далее клеточную культуру центрифугировали и промывали раствором PBS для подготовки клеток к дальнейшей сортировке. Параллельно контрольными плазмидами pRFPCER BamHI и pRFPCER T7 с идентичными 5'-UTR перед старт-кодонами обоих флуоресцентных белков трансформировали клетки штаммов JM109 и BL21 (DE3) *E.coli*, а затем готовили к сортировке. При этом в первой конструкции оба гена находятся под контролем промоторов T5, тогда как во второй ген *rfp* – под контролем промотора T5, а ген *cer* – под контролем промотора T7.

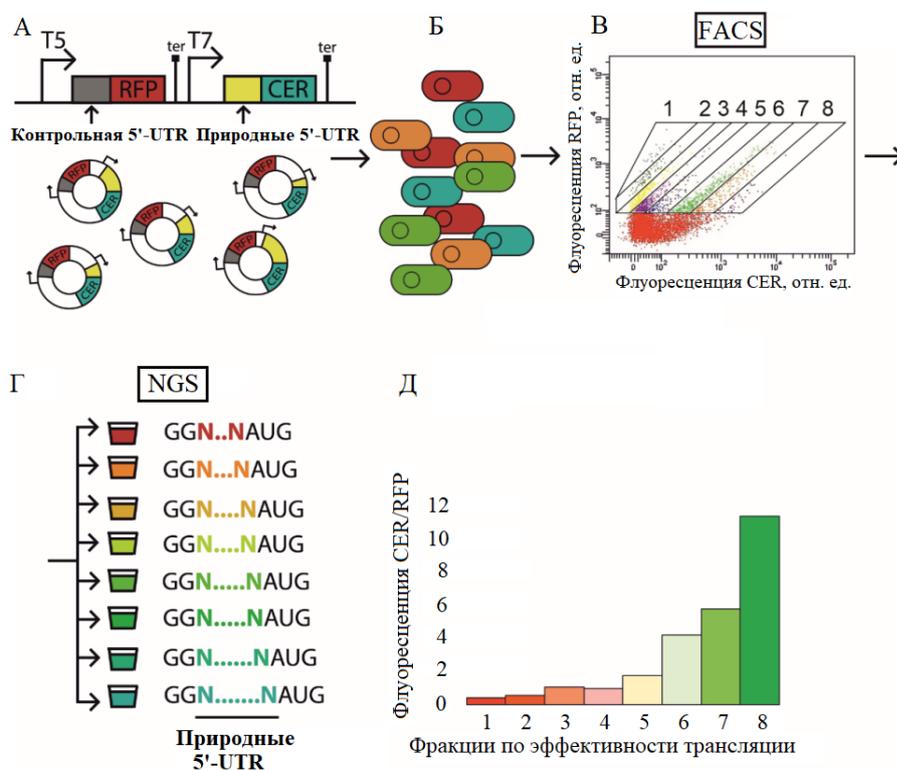


Рисунок 68. Схема эксперимента с Flow-seq. А. Клонирование ДНК-фрагментов, содержащих природные 5'-UTR в репортёрный вектор pRFPCER перед геном *cer*, при этом ген *rfp* сохраняет константную 5'-UTR. Промоторы, терминаторы, 5'-UTR, кодирующие области флуоресцентных белков CER и RFP отмечены. Б. Трансформация клеток штамма BL21 (DE3) *E. coli* библиотекой плазмид. В. Сортировка клеток по соотношению интенсивностей флуоресценции белков CER/RFP с помощью FACS – проточной цитофлуориметрии. Каждая фракция клеток F1-F8, отличающаяся по соотношению CER/RFP, ограничена наклонной рамкой. Г. Сбор фракций клеток с конкретным соотношением CER/RFP. Секвенирование нового поколения (NGS) ДНК-фрагментов с природными 5'-UTR, полученных после выделения плазмид из клеток каждой фракции и ПЦР-амплификации. Д. Соотношение интенсивностей флуоресценции CER/RFP, нормированное на таковое для контрольной плазмиды pRFPCER T7, в клетках из фракций F1-F8 (1-8), измеренное спектрофлуориметром. Рисунок создан в программе CorelDRAW.

### 3.2. Применение метода Flow-seq для разделения клеток *E. coli* с разными репортёрными конструкциями, содержащими природные 5'-UTR, и последующего анализа их последовательностей

На следующем этапе применили метод Flow-seq для разделения клеток *E. coli* (Рис. 68В), которые трансформировали библиотекой плазмид с природными 5'-UTR перед геном *cer*, и для дальнейшего анализа того, какая именно последовательность из всего полученного набора переменных 5'-UTR задаёт то или иное соотношение белков CER/RFP (Рис. 68Г-Д).

Сначала с помощью проточной цитофлуориметрии оценили контроли, в качестве которых выступили исходный штамм JM109 *E. coli* без плазмид, плазмиды pRFP-CER BamHI и pRFP-CER T7, упоминавшиеся ранее и трансформированные в клетки штаммов JM109 и BL21 (DE3) *E. coli*. В результате, как и ожидалось, для первого контроля – исходного штамма без плазмид – не наблюдалось ни сигнала от RFP, ни сигнала от CER. Для плазмиды pRFP-CER BamHI возможно детектирование сигналов от обоих флуоресцентных белков в обоих штаммах JM109 и BL21 (DE3) ввиду их способности узнавать промотор T5. При этом для плазмиды pRFP-CER T7 наблюдается флуоресценция только белка RFP, а CER – нет в штамме JM109, в то же время в штамме BL21 (DE3) видны оба сигнала от белков RFP и CER, так как он может узнавать и T5, и T7 промоторы.

Далее для клеток *E. coli*, содержащих библиотеку конструкций, проводилась сортировка в зависимости от соотношения интенсивностей флуоресценции белков CER/RFP на 8 фракций F1-F8 так, что F8 включала клоны с наибольшей интенсивностью флуоресценции CER. В общем проанализировано  $10^4$  клеток, что превышает разнообразие библиотеки 5'-UTR более, чем в 10 раз. В ходе сортировки доля светящихся клеток составила 91,5%. Только 0,3% клеток, трансформированных библиотекой природных 5'-UTR, попадало во фракцию F8, соответствующую максимальной эффективности трансляции белка CER (Рис. 68В), что на порядок больше, чем в случае рандомизированной библиотеки 20N, и близко к проценту клеток во фракции F8 в случае библиотеки 30N (0,45%).

На следующем этапе собранные клетки каждой фракции растили в течение 16 часов при интенсивном перемешивании при 37°C. Далее для оценки эффективности разделения клетки каждой фракции использовали для измерения соотношения флуоресценции белков CER/RFP. При этом, как и ожидалось, оно увеличивалось от фракции F1 к F8 (Рис. 68Д), тогда как общая разница между крайними фракциями составила приблизительно 30 раз, что значительно меньше диапазона, полученного для рандомизированных библиотек. Затем из клеток выделяли плазмидную ДНК и проводили ПЦР-амплификацию её фрагментов различной длины 94-149 нт, содержащих библиотеку природных 5'-UTR наряду с последовательностями промотора T7 и начала кодирующей области гена белка CER (первых 19 кодонов). Полученные ампликоны

проверяли с помощью электрофореза в 2% агарозном геле. Затем для них проводили высокопроизводительное секвенирование (NGS), которое позволило установить соответствие между клетками, несущими определённые варианты вставок в плазидах, и их распределением по фракциям, разделённым по соотношению белков CER/RFP. Каждый вариант 5'-UTR относили к той фракции, где найдено максимальное число ридов, соответствующих данному варианту. В результате полученное число последовательностей природных 5'-UTR, идентифицированное во всех 8 фракциях и представленное к последующему анализу, составило 648 из 713 вариантов, использованных для клонирования.

### **3.3. Оценка влияния вторичной структуры и состава природных 5'-UTR на эффективность трансляции**

Эффективность трансляции отдельной мРНК является скорее предметом оптимизации, нежели чем максимизации, ввиду различий в необходимых количествах каждого белка в клетке. Для набора полученных природных 5'-UTR мы провели анализ особенностей, способствующих биосинтезу белка. Прежде всего, мы оценили количество вариантов 5'-UTR, соответствующих каждой фракции, отражающей эффективность трансляции при использовании метода Flow-seq (Рис. 69А). Среди всей полученной библиотеки репортёрных конструкций большинство 5'-UTR обеспечивало умеренную экспрессию.

Далее провели оценку влияния вторичной структуры, способной маскировать сайты посадки рибосомы, на уровень биосинтеза белка [137]. Как известно, энергия сворачивания участка мРНК, взаимодействующего с рибосомой в ходе инициации трансляции, является важной особенностью, управляющей эффективностью трансляции. С этой целью провели анализ энергии укладки участка мРНК, включающего природную 5'-UTR и начало кодирующей области гена *cer* (16 нт) и вероятно контактирующего с рибосомой во время инициации (Рис. 69Б-В). Как и ожидалось, в среднем менее стабильные вторичные структуры наблюдались в мРНК, транслирующихся более эффективно, особенно в области, окружающей старт-кодон (Рис. 69В). Однако диапазон различий по этому признаку между фракциями небольшой. Данное наблюдение согласуется с идеей, что почти все природные 5'-UTR эволюционировали таким образом, чтобы свести к минимуму наличие вторичных структур. Минимальная медианная энергия укладки, наблюдаемая во фракции F1, составляет -6 ккал/моль; но это значение всё же несопоставимо со вторичными структурами, способными резко ингибировать трансляцию [137, 145].

Самым известным элементом 5'-UTR, определяющим эффективность трансляции, выступает последовательность Шайна-Дальгарно [13]. На следующем этапе проанализировали варианты последовательностей 5'-UTR, входящие в состав полученной библиотеки репортёрных конструкций, с целью поиска областей, которые с наибольшей вероятностью будут образовывать

контакты с 3'-концевым участком 16S рРНК (Рис. 69Г-Д). В соответствии с ожиданиями, для тех 5'-UTR, которые демонстрируют более сильную связь SD с 16S рРНК (Рис. 69Г-Д), наблюдалась в среднем более эффективная трансляция. При этом никакой разницы между фракциями F1-F8 в расположении последовательности SD относительно старт-кодона не обнаружено (Рис. 69Е).

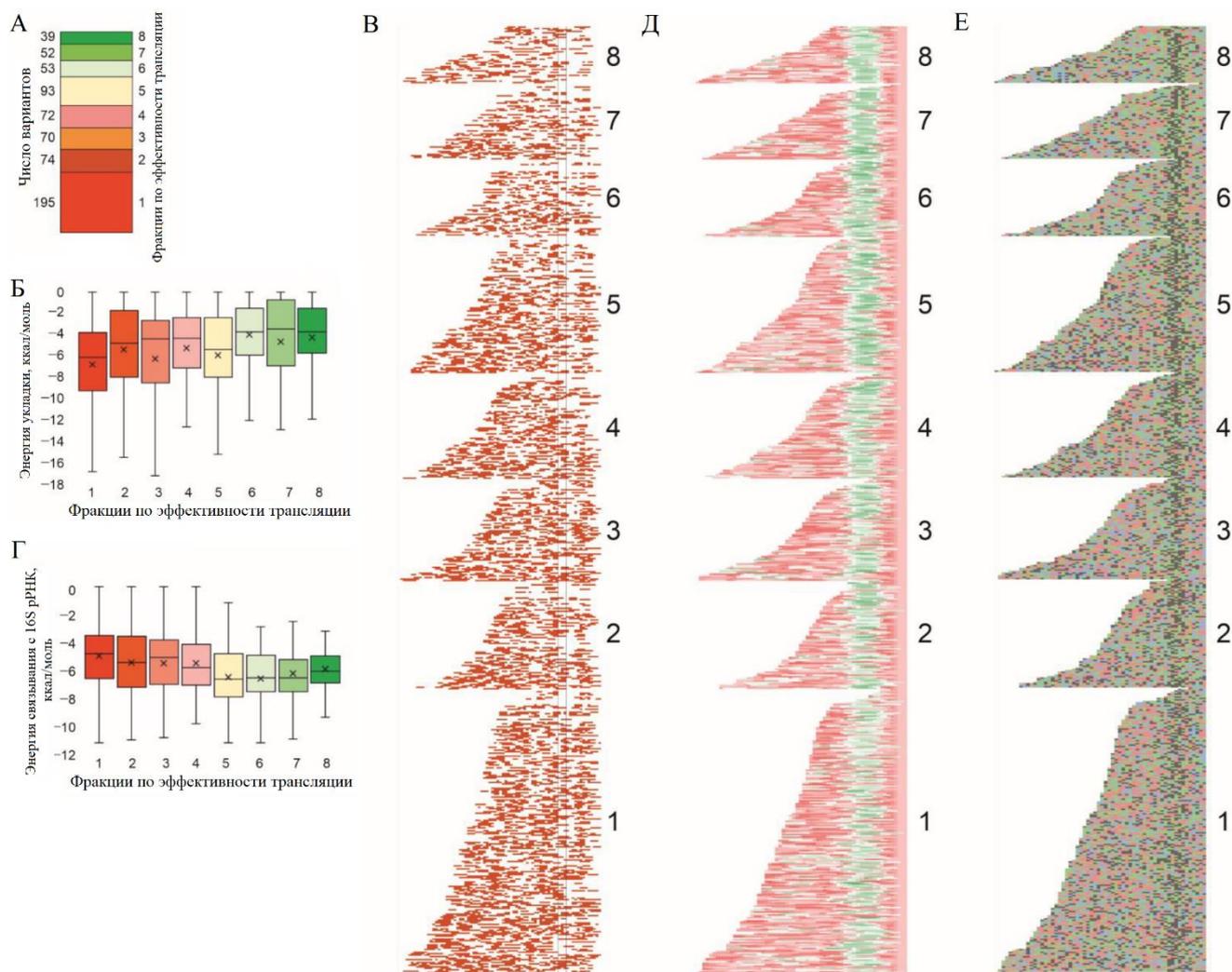


Рисунок 69. Оценка стабильности вторичной структуры и возможности образовывать взаимодействия с анти-SD 16S рРНК последовательностей 5'-UTR с разной эффективностью трансляции. Номера 1-8 соответствуют фракциям F1-F8, отсортированным по увеличению эффективности трансляции. А. Количество вариантов 5'-UTR (показано слева от диаграммы), попавших в каждую фракцию (показана справа). Б. Распределение во всех фракциях энергии укладки участка мРНК, включающего 5'-UTR и первые 16 нт кодирующей области гена *cer*. В. Распределение нуклеотидов, участвующих в образовании вторичных структур. Отдельная строка соответствует определённому участку мРНК от 5'-конца до положения +16 относительно начала открытой рамки считывания гена *cer*. Расположение старт-кодона выделено чёрными линиями. Красные прямоугольники соответствуют спаренным нуклеотидам, белые – одноцепочечным нуклеотидам. Все мРНК выровнены по их 3'-концу. Последовательности отсортированы по

эффективности трансляции (номер фракции указан справа) и по длине (внутри каждой фракции). Г. Распределение энергии взаимодействия 5'-UTR с участком анти-SD 16S рРНК во фракциях. Д. Распределение последовательностей SD вдоль проверенных 5'-UTR. Каждая отдельная строка соответствует отдельному варианту 5'-UTR мРНК. Цвет отражает энергию взаимодействия в пределах «скользящего окна» длиной 8 нт мРНК с 3'-концевой областью 16S рРНК. Зелёный соответствует эффективному взаимодействию, а красный – отсутствию контакта. Принцип сортировки мРНК такой же, как и для панели В. Е. Нуклеотидный состав последовательностей проанализированных 5'-UTR, отсортированных по эффективности трансляции. Каждая строка соответствует определённому варианту 5'-UTR мРНК. Синие прямоугольники соответствуют нуклеотиду С, красные – U, зелёные – А, серые – G. Принцип сортировки мРНК такой же, как и для панели В. Для панелей Б и Г диапазоны квартилей показаны в виде «ящика» (от 25% до 75%) и тонких линий, известных как «усы» (от 0% до 100%). Горизонтальная линия соответствует медиане, а крестик – среднему значению.

Кроме того, мы проверили наличие AU-богатых энхансеров трансляции [14] в участке мРНК, предшествующем последовательности SD, оценив частоту встречаемости каждого типа нуклеотидов в определённой позиции в 5'-UTR во всех восьми фракциях. В результате преобладание аденозинов и уридинов в потенциальном сайте энхансера трансляции отмечалось для всех наборов вариантов 5'-UTR, сгруппированных по эффективности трансляции. При этом наблюдалась тенденция к преобладанию А-богатых последовательностей для 5'-UTR с более высокой эффективностью трансляции, в то время как U-богатых – для вариантов с более низкой эффективностью. Данное наблюдение можно объяснить тем, что в эксперименте использовались природные 5'-UTR, которые универсально обладают оптимальными необходимыми элементами для связывания рибосом, а эффективность трансляции регулируется другими способами, такими как, например, корректировкой частоты кодонов. В то же время мы не смогли увидеть значимой разницы в содержании нуклеотидов AC, как ранее предполагалось на основе анализа библиотек рандомизированных 5'-UTR [212].

Дополнительно мы провели оценку того, как длины 5'-UTR распределены между фракциями, разделёнными по эффективности трансляции (Рис. 70). Оказалось, что распределение длин последовательностей 5'-UTR примерно одинаково для мРНК, обеспечивающих разные уровни биосинтеза белка.

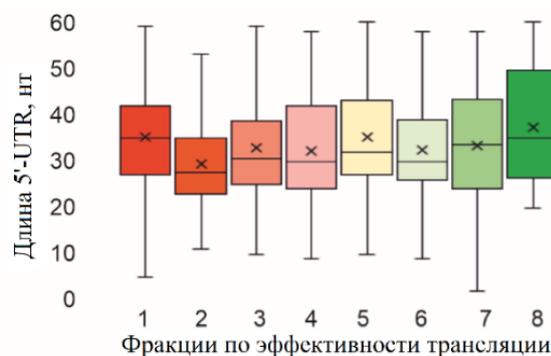


Рисунок 70. Влияние длины 5'-UTR на эффективность трансляции. Номера 1-8 соответствуют фракциям F1-F8, отсортированным по увеличению уровня синтезированного белка. Показано распределение длин 5'-UTR во фракциях. Диапазоны квартилей показаны в виде «ящика» (от 25% до 75%) и тонких линий, известных как «усы» (от 0% до 100%). Горизонтальная линия соответствует медиане, а крестик – среднему значению.

В полученной библиотеке природных 5'-UTR длиной 2-60 нт представлено 18 вариантов последовательностей из проверенных ранее 48 5'-UTR природных генов, при этом около половины найденных вариантов сохранили схожую тенденцию в распределении эффективности трансляции относительно контроля. Для остальных расхождения в уровне экспрессии скорее можно объяснить отличиями в использованных системах: в промоторах (ранее был T5 вместо T7), старт-кодонах (ранее везде кодирующая область начиналась с AUG, тогда как сейчас встречались природные GUG и AUG кодоны) и клеточных линиях (ранее была JM109, теперь BL21 (DE3) *E. coli*).

### 3.4. Сравнение результатов, полученных с помощью методов Flow-seq и Ribo-seq

На сегодняшний момент среди активно используемых подходов для изучения трансляции можно выделить метод Ribo-seq – рибосомный профайлинг, основанный на секвенировании нового поколения фрагментов мРНК, защищённых рибосомой во время трансляции. Применяя его, можно получить информацию об уровне трансляции мРНК каждого гена в клетке, отражённую в виде количества ридов отдельных мРНК, связанных с рибосомами. Таким образом, данная информация позволяет оценить эффективность трансляции каждой мРНК.

На следующем этапе провели сравнение эффективности трансляции, наблюдаемой с помощью Flow-seq для полученной нами библиотеки природных 5'-UTR перед репортёрным геном *cer* в мРНК и оцененной с помощью Ribo-seq [181, 251-253] для мРНК соответствующих генов, чьи 5'-UTR были в нашем наборе. В результате мы наблюдали, что наши данные в целом согласуются с полученными *in vivo* методом рибосомного профайлинга в 3 разных исследованиях для штамма BW25113, выращенного без обработки антибиотиками в разных средах (Рис. 71-72).

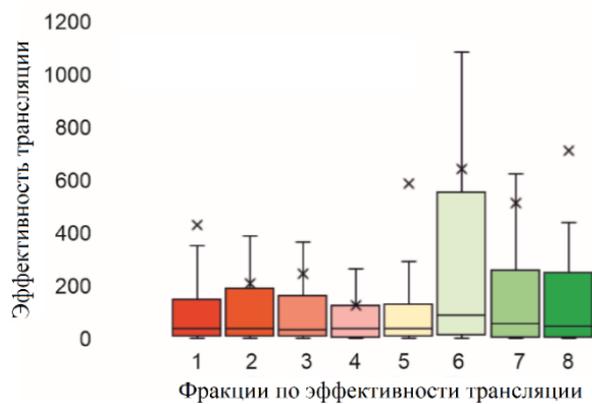


Рисунок 71. Сравнение эффективностей трансляции, наблюдаемых с помощью методов Flow-seq и Ribo-seq [253]. Номера 1-8 соответствуют фракциям F1-F8, отсортированным по увеличению эффективности трансляции. По оси Y представлена эффективность трансляции, определённая с помощью рибосомного профайлинга в виде количества ридов для конкретной мРНК. Диапазоны квартилей показаны в виде «ящика» (от 25% до 75%) и тонких линий, известных как «усы» (от 0% до 100%). Горизонтальная линия соответствует медиане, а крестик – среднему значению.

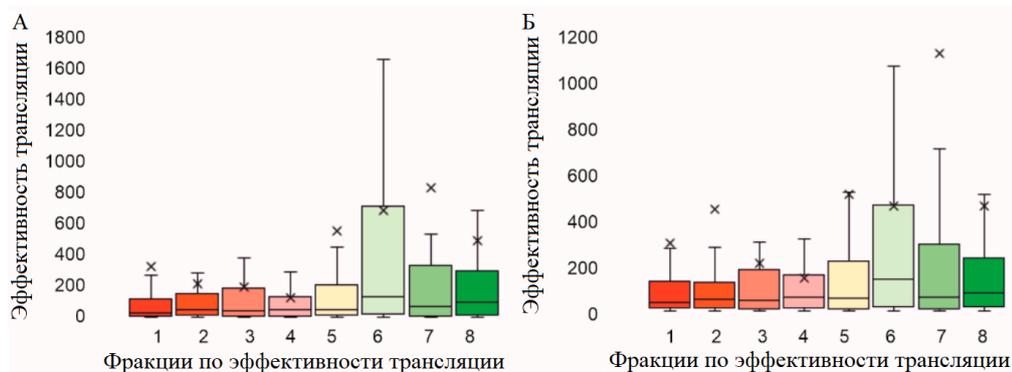


Рисунок 72. Сравнение эффективностей трансляции, наблюдаемых с помощью методов Flow-seq и Ribo-seq и представленных на панелях А и Б из исследований [251] и [252] соответственно. Номера 1-8 соответствуют фракциям F1-F8, отсортированным по увеличению эффективности трансляции. По оси Y представлена эффективность трансляции, определённая с помощью рибосомного профилирования в виде количества ридов для конкретной мРНК. Диапазоны квартилей показаны в виде «ящика» (от 25% до 75%) и тонких линий, известных как «усы» (от 0% до 100%). Горизонтальная линия соответствует медиане, а крестик – среднему значению.

Полученный результат ещё раз подтверждает успешность использования такого метода, как Flow-seq, для создания и анализа библиотек репортёрных конструкций с 5'-UTR различной длины перед геном сенсорного белка CER, а также для оценки влияния 5'-UTR на эффективность трансляции в системе *E. coli*.

#### 4. Особенности применения метода Flow-seq для оценки эффективности трансляции

При подготовке данного раздела диссертации использованы следующие публикации, выполненные автором в соавторстве, в которых, согласно Положению о присуждении учёных степеней в МГУ, отражены основные результаты, положения и выводы исследования: [16], общий объём публикации – 1 п.л. Причём вклад автора в ней был определяющим: анализ литературных данных, особенностей и применения метода Flow-seq в изучении бактериальной трансляции выполнен полностью автором лично.

Метод Flow-seq основан на комбинировании гибких способов геномной биоинженерии, клеточной сортировки с применением проточной цитометрии с детектированием флуоресценции и последующего высокопроизводительного секвенирования, что позволяет всесторонне оценить взаимосвязи генотипа и наблюдаемого фенотипа. Данный метод нашёл активное применение при изучении влияния на эффективность трансляции нуклеотидных последовательностей различных участков или целых 5'-UTR (Приложение 1). А именно, подобное исследование проводили для различных комбинаций промоторов и сайтов посадки рибосомы [205, 214], для RBS, где изменяли standby-сайт (4 нт) и спейсерный участок (6 нт), а последовательность SD была фиксированной длины и на определённом расстоянии от старт-кодона [212], или же для рандомизированного участка в районе расположения SD длиной 6 нт [215] в 5'-UTR одной мРНК или 9 нт в трёх разных мРНК [213], для вариативных standby-сайтов различной длины [151] или standby-сайта и спейсерного участка [210], для рандомизированного спейсерного участка и первых двух кодонов, следующих за инициаторным в CDS [209], для набора промоторов, RBS и первых 13 аминокислот в CDS [216] или же отдельно для начала кодирующей области сенсорного гена длиной 18 нт [219] или 30 нт [145]. Ряд данных работ дополнили и описанные выше исследования, в которых изучалось влияние рандомизированного участка в спейсерной области 5'-UTR, полной рандомизированной 5'-UTR вариативной длины и природных 5'-UTR различной длины перед геном сенсорного флуоресцентного белка на эффективность трансляции.

Анализ десятков тысяч различных вариантов последовательностей позволил получить диапазон изменения уровня синтезируемого репортёрного белка в пределах четырёх, а порой и пяти порядков. При этом если заменить один репортёрный белок на другой, то зачастую общая тенденция сохранялась для распределения последовательностей мРНК, обеспечивающих ту или иную эффективность трансляции, что свидетельствует о значимой роли именно вариативных 5'-UTR в определении наблюдаемых изменений в экспрессии сенсорного гена.

Наблюдаемая в представленных выше работах низкая стабильность вторичной структуры прослеживалась в вариантах мРНК, обеспечивающих высокую эффективность трансляции, и в большинстве более ранних исследований с применением метода Flow-seq [145, 151, 209-210, 212,

216, 219].

Присутствие последовательности SD в рандомизированных 5'-UTR, обеспечивающих высокий уровень трансляции, можно отметить как в данных научной литературы [212-213, 215], так и в нашей работе.

В результатах, полученных с помощью метода Flow-seq при изучении влияния различных участков 5'-UTR мРНК на эффективность трансляции в описанных выше и опубликованных ранее работах, нередко отмечается преобладание частоты определённых нуклеотидов в тех вариантах последовательностей, которые обеспечивают высокий уровень биосинтеза белка. В частности, в эффективно транслирующихся мРНК наблюдается низкое содержание GC-нуклеотидов в спейсерном участке [209], что также замечено и в нашем исследовании при рандомизации 4 нт в том же участке. Увеличение доли А и G и уменьшение доли С и U нуклеотидов детектировалось в 5'-UTR таких мРНК при рандомизации участка в районе SD длиной 9 нт [213]. Уменьшение доли цитозинов отмечено в нашей работе как при рандомизации 4 нт в спейсерном участке 5'-UTR, так и при рандомизации всей 5'-UTR длиной 20 нт или 30 нт. Наряду с этим в 5'-UTR встречались отдельные наблюдения, указывающие на присутствие множественных последовательностей SD, AU-богатых энхансеров, AG-повторов, которые обеспечивали высокую эффективность трансляции в наших исследованиях.

При дизайне направленных изменений с использованием генов флуоресцентных белков в репортёрных конструкциях возможно быстрое и эффективное определение вклада в уровень биосинтеза белка отдельных вариантов регуляторных последовательностей. Метод Flow-seq имеет ряд особенностей, которые необходимо учитывать при планировании многоступенчатого и комплексного эксперимента. Есть несколько ограничений, обусловленных, в первую очередь, сложностями, которые возникают на разных этапах: при клонировании ДНК-библиотеки, при проведении клеточной сортировки на основе разных соотношений интенсивностей флуоресценции репортёрных белков, при глубоком секвенировании и последующем анализе полученных результатов NGS. В то же время из-за возможности перекрытия спектров флуоресценции используемых белков или других реагентов наряду с последующими ошибками при регистрации сигналов следует ограничиваться только двумя детектируемыми объектами (флуоресцентными белками или реагентами). Несмотря на ограничения, метод Flow-seq нашёл широкое применение в разных областях научных исследований и продолжает быть актуальным уже много лет.

Далее рассмотрим несколько отмеченных рекомендаций по использованию различных последовательностей 5'-UTR для решения возможных биотехнологических задач.

Влияние определённой SD, как известно, сильно зависит от окружающих её последовательностей [149]. Поэтому следует с осторожностью повторно применять измеренные

уровни экспрессии в синтетической биологии или биоинженерной работе с разными белками и метаболическими путями ввиду сильной зависимости RBS от окружающей вторичной структуры мРНК. При возможности модификации SD и при минимальном воздействии на стабильность такой структуры в исследуемом участке мРНК может сохраниться относительное распределение уровня экспрессии белка с определённой SD в 5'-UTR [211]. Этот принцип учитывается при работе с алгоритмом в базе EMOPEC, позволяя с минимальными изменениями в SD оценить широкий спектр изменения уровней экспрессии генов, делая в свою очередь возможным параллельное и эффективное редактирование генома для настройки этих уровней.

Для бактериальных систем экспрессии при выборе спейсерных участков между последовательностью SD и старт-кодоном кажется разумным использование олигоадениловых или других А-богатых спейсерных последовательностей при возможном сведении к минимуму количества оснований цитидина, если необходимо увеличить эффективность трансляции белка. Но в то же время следует учитывать возможную маскировку сайта инициации за счёт образования вторичной структуры некоторыми отдельными мРНК, содержащими А-богатые спейсерные участки и U-богатую последовательность в начале кодирующей области. Правильный подбор спейсерной области в модельных мРНК может также представлять интерес при необходимости регулирования уровней экспрессии экзогенных генов белков, входящих, например, в гетеромультимерный комплекс в заданном стехиометрическом соотношении.

Относительно всей последовательности 5'-UTR можно сделать вывод о комплексном влиянии различных элементов, входящих в неё, на эффективность трансляции. Помимо потенциального использования отмеченных энхансеров, следует оценивать стабильность вторичной структуры 5'-UTR и начала кодирующей области белка, а также учитывать нуклеотидный состав в разных участках 5'-UTR. Также представляется разумным учитывать длину 5'-нетранслируемой области мРНК и избегать очень длинных последовательностей, при этом можно успешно использовать 5'-UTR длиной 20-40 нт. Данный диапазон показан как наиболее часто встречающийся среди природных 5'-UTR мРНК, эволюционно стремящихся к оптимизации процесса биосинтеза белка.

## ЗАКЛЮЧЕНИЕ

Одним из сложноустроенных процессов, лежащих в основе жизнедеятельности всех организмов, включая бактерий, является биосинтез белка, поэтому изучение его регуляции всегда представляет интерес для учёных в этой области. Несмотря на десятилетия исследований регуляции трансляции, до сих пор остаются не до конца изученные детали этого процесса. Наше понимание того, как и какие особенности мРНК влияют на эффективность трансляции у бактерий, формировалось на протяжении долгого пути, начатого с открытия последовательности Шайна-Дальгарно в 5'-нетранслируемой области для небольшого набора мРНК в 1970-х годах. Позже оказалось, что есть и другие элементы мРНК, важные в определении уровня экспрессии белка. Однако не все наблюдаемые эффекты в биосинтезе белка можно объяснить наличием или отсутствием известных на данный момент элементов мРНК.

Целью данной научно-исследовательской работы является изучение особенностей 5'-нетранслируемой области бактериальных мРНК, влияющих на эффективность трансляции, с помощью библиотек репортёрных конструкций для поиска новых эффективных сайтов посадки рибосомы в них и определения основных элементов таких сайтов, определяющих уровень экспрессии белка.

Для лучшего исследования такого сложного и важного процесса, как биосинтез белка, необходимо использовать несколько разных подходов, которые помогут взглянуть на него с многих сторон. С этой целью мы провели модификацию комплексного метода Flow-seq, чтобы оценить, насколько определённая последовательность 5'-UTR способна изменить эффективность трансляции, и благодаря каким её особенностям это может происходить. Далее мы адаптировали этот метод для одновременного определения уровня биосинтеза белка для целых библиотек различных вариантов сайтов посадки рибосомы, многократно превышающих их природное разнообразие. При этом мы создали наборы из более 24,5 тысяч репортёрных конструкций с различными синтетическими и природными 5'-UTR перед геном сенсорного флуоресцентного белка CER.

Для полученных библиотек конструкций провели анализ нуклеотидного состава и стабильности вторичной структуры мРНК. В результате показали наличие известных элементов, таких как последовательность SD, AU-богатый энхансер на 5'-конце мРНК, и присутствие элементов вторичных структур, которые легко расплавить, либо их полное отсутствие в 5'-UTR мРНК среди последовательностей, обеспечивающих высокий уровень экспрессии белка. Также представлены и ранее не описанные особенности 5'-UTR мРНК, способствующие увеличению эффективности трансляции, а именно низкую долю остатков цитозина, множественные SD, AG-повторы.

Для набора природных 5'-UTR продемонстрировали низкую вариабельность в уровне синтезируемого белка в результате эволюционного отбора в пользу более эффективных последовательностей. В то же время впервые проведённый анализ результатов, полученных для 648 природных 5'-UTR перед геном *cer* в составе репортёрных конструкций с помощью метода Flow-seq и для природных мРНК с тем же набором 5'-UTR генов *E. coli* с помощью метода Ribo-seq, выявил хорошее соответствие между ними.

Наряду с экспериментальными данными в данной работе представлен обзор, воедино собравший разнообразие подходов, используемых при изучении биосинтеза белка с индивидуальных мРНК или их набора, и результаты применения метода Flow-seq для оценки эффективности трансляции бактериальных мРНК. Также приведены некоторые рекомендации по выбору различных последовательностей 5'-UTR, которые могут быть полезны при решении биотехнологических задач.

Полученные результаты представляют большой интерес с фундаментальной точки зрения, поскольку проливают новый свет на то, как последовательность 5'-нетранслируемой области бактериальных мРНК влияет на эффективность такой сложной и важной составляющей жизнедеятельности клетки, как трансляция. В то же время они позволяют учесть больше тонкостей процесса биосинтеза белка на практике.

## ВЫВОДЫ

1. Метод Flow-seq успешно адаптирован для высокопроизводительного определения эффективности трансляции синтетических и природных 5'-нетранслируемых областей, а следовательно и участков посадки рибосомы, многократно превышающих их природное разнообразие у *E. coli*.
2. Показано, что высокая эффективность трансляции некоторых мРНК не может быть объяснена исключительно наиболее изученными особенностями, такими как взаимодействие рибосомы с последовательностью Шайна-Дальгарно, наличие AU-богатого энхансера на 5'-конце мРНК и отсутствие ингибирующего влияния вторичной структуры мРНК.
3. Продемонстрировано, что эффективность трансляции уменьшается с увеличением доли остатков цитозина в последовательности 5'-нетранслируемой области.
4. Отмечено, что дополнительный старт-кодон и множественные последовательности Шайна-Дальгарно имеют аддитивное влияние на эффективность трансляции.
5. Обнаружено, что AG-повторы в 5'-нетранслируемой области представляют собой энхансер трансляции.
6. Выявлено, что мРНК с природными 5'-нетранслируемыми областями менее вариабельны по эффективности трансляции, чем рандомизированные последовательности. Замечено, что в более длинных, 30-звенных 5'-нетранслируемых областях чаще встречаются благоприятные для трансляции варианты из случайных последовательностей, чем в 20-звенных.

# МАТЕРИАЛЫ И МЕТОДЫ

## 1. Материалы

### 1.1. Используемые приборы

1. Настольная центрифуга: Eppendorf Centrifuge 5415R;
2. Центрифуга: Eppendorf Centrifuge 5804 R;
3. Термостат для проведения ферментативных реакций: ДНК-Технология «Гном»;
4. Настольный орбитальный шейкер-инкубатор: Innova 40 фирмы New Brunswick
5. Система визуализации и анализа с камерой и УФ-лампой для гелей и Вестерн-блотов: ChemiDoc фирмы Bio-Rad для детектирования хемилюминесценции, мультиплексной флуоресценции, для работы по технологии без окраски;
6. Термостат для выращивания культур клеток;
7. Вакуумный концентратор: SpeedVac;
8. Измеритель оптической плотности: NanoDrop;
9. Электропоратор с высоковольтным мини-электродом и специальная кювета фирмы Bio-Rad;
10. Спектрофлуориметр (флуориметр): Victor X5 фирмы PerkinElmer;
11. Клеточный сортер: BD FACSAria™ III;
12. ПЦР-амплификатор: Mastercycler Gradient фирмы Eppendorf;
13. ПЦР-амплификатор: T100 Thermal Cycler фирмы Bio-Rad;
14. Сканер-денситометр для визуализации результатов гель-электрофореза: Typhoon FLA 9500;
15. Секвенатор: Ion Torrent [162] PGM (Life Technologies)
16. Секвенатор: MGIseq-2000 (MGI Tech)

### 1.2. Реактивы и биопрепараты

В ходе работы использовали следующие реактивы и биопрепараты: ампициллин (Sigma); дезоксирибонуклеозидтрифосфаты (Fermentas); N,N,N',N'-тетраметилэтилендиамин (ТЕМЕД) (Serva); акриламид, N,N'-метиленабисакриламид, мочеви́на (Gibco BRL); персульфат аммония, ацетат натрия, хлорид натрия, хлорид магния, додецилсульфат натрия, глицерин (Merck); соляная кислота (Химмед); этанол (Ферейн); хлорид калия, Tris, H<sub>3</sub>BO<sub>3</sub>, ксиленианол, ЭДТА, бакто-агар, агароза, глюкоза, бромфеноловый синий, бромид этидия (Helicon); дрожжевой экстракт, бакто-триптон (Amresco); набор для выделения плазмидной ДНК, набор для выделения ДНК из агарозного геля, набор для быстрого лигирования ДНК Rapid DNA Ligation Kit (Fermentas); маркеры со смесью ДНК-фрагментов определённой длины 1kb Plus, 100bp Plus и Ultra Low Range DNA Ladder GeneRuler™ (Invitrogen, Thermo Fisher); набор для выделения плазмидной ДНК QIAprep Spin Miniprep Kit, набор для выделения ПЦР-продуктов из смеси QIAquick PCR

Purification Kit (Qiagen); смесь дезоксинуклеотидов (dNTP) по 2 мкМ каждого (NEB); фосфатаза CIAP (Thermo Fisher).

В работе использовали ферменты фирмы MBI Fermentas (Литва):

1. Эндонуклеазы рестрикции:

SacII, NdeI, EcoRV, BamHI, Esp3I (BsmBI), HindIII и соответствующие им буферы.

2. Полимеразы:

Фрагмент Клёнова ДНК-полимеразы I *E. coli*, Taq ДНК-полимераза, 2x PCR Master Mix с Taq ДНК-полимеразой.

3. Лигазы:

T4 ДНК-лигаза.

Данные ферменты использовали согласно методике производителя Fermentas.

В работе использовали ферменты фирмы NEB (США):

1. Эндонуклеазы рестрикции:

HindIII, BamHI.

2. Полимеразы:

2x PCR Master Mix с Taq ДНК-полимеразой, High-Fidelity 2x Master Mix с Q5 ДНК-полимеразой высокой точности.

Данные ферменты использовали согласно методике производителя NEB.

### 1.3. Состав буферов и сред

В работе использовали следующие среды, буферы и растворы:

**LB:** 0,5 % дрожжевой экстракт; 1 % бакто-триптон; 1% NaCl.

**SOB:** 0,5% дрожжевой экстракт; 2% бакто-триптон; 0,5 мМ MgCl<sub>2</sub>; 2,5 мМ KCl; 8,55 мМ NaCl.

**SOC:** 0,5% дрожжевой экстракт; 2% бакто-триптон; 0,5 мМ MgCl<sub>2</sub>; 2,5 мМ KCl; 8,55 мМ NaCl; 20 мМ глюкоза.

**Агаризованные среды:** состав как у жидких с добавлением 1,5% бакто-агара.

**1x TBE:** 100 мМ Tris-HCl; 100 мМ H<sub>3</sub>BO<sub>3</sub>; 2 мМ ЭДТА (pH 8,3).

**1x PBS:** 140 мМ NaCl; 10 мМ Na<sub>2</sub>HPO<sub>4</sub>; 2,7 мМ KCl; 1,8 мМ KH<sub>2</sub>PO<sub>4</sub> (pH 7,3).

**Агарозные гели:** 1-2% агароза в буфере 1x TBE; 0,05 мкг/мл бромида этидия.

**6x буфер для нанесения ДНК на гель:** 0,1% бромфеноловый синий; 0,1% ксиленианол; 0,5% додецилсульфат натрия, 0,1М ЭДТА (pH 8,0), 50% глицерин.

**Полиакриламидные гели для разделения ДНК:** 15% полиакриламид в смеси с 1/20 бисакриламида; 1x TBE; 420 г/л мочевины; 0,1% персульфат аммония; 0,1% ТЕМЕД.

#### 1.4. Штаммы, плазмиды и олигонуклеотиды

Использованные штаммы	
Название штамма	Описание
JM109 <i>E. coli</i>	<i>endA1, recA1, gyrA96, thi, hsdR17</i> (rk <sup>-</sup> , mk <sup>+</sup> ), <i>relA1, supE44</i> , Δ( <i>lac-proAB</i> ), [F' <i>traD36, proAB, laqIqZΔM15</i> ]
BW25113 <i>E. coli</i>	<i>lacIq rrnBT14 ΔlacZWJ16 hsdR514 ΔaraBADAH33 ΔrhaBADLD78</i> [254]
BL21 (DE3) <i>E. coli</i>	<i>E. coli B F<sup>-</sup> dcm ompT hsdS</i> (r <sub>B</sub> - m <sub>B</sub> -) <i>gal λ</i> (DE3)
Использованные плазмиды	
Название плазмиды	Содержащиеся гены
pRFPCER	<i>rfp, cer, amp<sup>R</sup></i>
pRFPCER BamHI	<i>rfp, cer, amp<sup>R</sup></i>
pRFPCER T7	<i>rfp, cer, amp<sup>R</sup></i>
Использованные олигонуклеотиды	
Название	Последовательность
4N spacer 5'-UTR	5' – ACTG <b>CCGCGG</b> CACACACCGGAGC <b>NNNNCATATC</b> – 3'
T5 CER F	5' – CCATCTCATCCCTGCGTGTCTCATTGCTTTCAGGAAAATTTTCTG – 3'
CER R 4N	5' – CCACTACGCCTCCGCTTTCCTC <b>NNNN</b> TCACCAGGCCGCTCTCGTCC – 3'
20N 5'-UTR F	5' – ACTG <b>CCGCGG</b> <b>NNNNNNNNNNNNNNNNNNNN</b> <b>ATGAAAGAGACGGACGAGAGCG</b> – 3'
30N 5'-UTR F	5' – ACTG <b>CCGCGG</b> <b>NNNNNNNNNNNNNNNNNNNNNNNNNNNNNN</b> <b>ATGAAAGAGACGGACGAGAGCG</b> – 3'
CER BamHI R	5' – GATC <b>CGCTCTCGTCCGCTCTTTCAT</b> – 3'
BamHI F	5' – TGAAAGAGACGGACGAGAGCGGATCCC – 3'
BamHI R	5' – TTCAGGGATCCGCTCTCGTCCGCTCT – 3'
P1	5' – CCACTACGCCTCCGCTTTCCTCTCTATGGGCAGTCCGGTATTTGCTTTCAGGAAAATTTTCTG – 3'
A29-A60	5' – CCATCTCATCCCTGCGTGTCTCCGACTCAG <b>NNNNNNNN</b> CGATTCCACGAGC <b>CGCTCTCGTCC</b> – 3'
A-rich F	5' – GGGATTTAAAAAAGCGGAAAAATAATGCA – 3'
A-rich R	5' – TATGCATTATTTTCCGCCCTTTTTTAAATCCCGC – 3'
A-rich -5G F	5' – GGGATTTAAAAAACAACAAAAATAATACA – 3'
A-rich -5G R	5' – TATGTATTATTTTGTTTTTTTTAAATCCCGC – 3'
A-rich noAUG F	5' – GGGATTTAAAAAAGCGGAAAAATAATACA – 3'
A-rich noAUG R	5' – TATGTATTATTTTCCGCCCTTTTTTAAATCCCGC – 3'
A-rich -2G F	5' – GGGATTTAAAAAAGGCAAAAAATAATGCA – 3'
A-rich -2G R	5' – TATGCATTATTTTTCCTTTTTTAAATCCCGC – 3'
AG-rich F	5' – GGAGTCTAAAGAGAGAGAGAG – 3'
AG-rich R	5' – TACTCTCTCTCTTTAGACTCCGC – 3'
AG-rich –AG <sub>1-3</sub> F	5' – GGAGTCTAAACACACAGAGAG – 3'
AG-rich –AG <sub>1-3</sub> R	5' – TACTCTCTGTGTTTAGACTCCGC – 3'
AG-rich –AG <sub>4-6</sub> F	5' – GGAGTCTAAAGAGAGACACAC – 3'
AG-rich –AG <sub>4-6</sub> R	5' – TAGTGTCTCTCTTTAGACTCCGC – 3'
AG-rich +SD1 F	5' – GGAGTCTAAAGAGAGGAGAAG – 3'

AG-rich +SD1 R	5' - TACTTCTCCTCTCTTTAGACTCCGC - 3'
AG-rich +SD2 F	5' - GGAGTCTAAAGGAGGAGAGAG - 3'
AG-rich +SD2 R	5' - TACTTCTCCTCCTTTAGACTCCGC - 3'
AG-rich -G F	5' - GGAGTCTAAAGAGACAGAGAG - 3'
AG-rich -G R	5' - TACTTCTGTCTCTTTAGACTCCGC - 3'
ShortSD F	5' - GGTTTCTTATTTGGTTCCGGAGTGAGATGCGA - 3'
ShortSD R	5' - TATCGCATCTCACTCCGAACCAAATAAGAAACCGC - 3'
ShortSD -SD F	5' - GGTTTCTTATTTGGTTCCCTCTGAGATGCGA - 3'
ShortSD -SD R	5' - TATCGCATCTCAGAGGGAACCAAATAAGAAACCGC - 3'
ShortSD -9UA F	5' - GGCCCCCGCCCGTTCCGGAGTGAGATGCGA - 3'
ShortSD -9UA R	5' - TATCGCATCTCACTCCGAACCGGGCGGGGGCCGC - 3'
ShortSD -3' F	5' - GGTTTCTTATTTGGTTCCGGAGTCACATCCCA - 3'
ShortSD -3' R	5' - TATGGGATGTGACTCCGAACCAAATAAGAAACCGC - 3'
ShortSD +SD F	5' - GGTTTCTTATTTGGTTCCACATGGAGTCCCA - 3'
ShortSD +SD R	5' - TATGGGACTCCATGTGGAACCAAATAAGAAACCGC - 3'
ShortSD -3U F	5' - GGTTTCTTACCCGGTTCCGGAGTGAGATGCGA - 3'
ShortSD -3U R	5' - TATCGCATCTCACTCCGAACCGGTAAGAAACCGC - 3'
ShortSD-GTC_F	5' - GGTTTCTTATTTGGTTCCGGAGTCAGATGCGA - 3'
ShortSD-GTC_R	5' - TATCGCATCTGACTCCGAACCAAATAAGAAACCGC - 3'
ShortSD-GCG_F	5' - GGTTTCTTATTTGGTTCCGGAGCGAGATGCGA - 3'
ShortSD-GCG_R	5' - TATCGCATCTCGCTCCGAACCAAATAAGAAACCGC - 3'
RFP_R	5' - GAATCTATTATACAGAAAAATTTTCCTGAAAGCAAATAATTTTATGATTCTAGAGCG - 3'
A-rich_RFP_F	5' - GGGATTTAAAAAAGGCGGAAAAATAATGCATATGGGCGAGCTGATCAAGGAG - 3'
ShortSD_RFP_F	5' - GGTTTCTTATTTGGTTCCGGAGTGAGATGCGATATGGGCGAGCTGATCAAGGAG - 3'
AG-rich_RFP_F	5' - GGAGTCTAAAGAGAGAGAGATGAGATGAGATGCGATATGGGCGAGCTGATCAAGGAG - 3'
A-rich_RFP-G_F	5' - GGGATTTAAAAAACAACAAAAATAATACATATGGGCGAGCTGATCAAGGAG - 3'
ShortSD_RFP9UA_F	5' - GGCCCCCGCCCGTTCCGGAGTGAGATGCGATATGGGCGAGCTGATCAAGGAG - 3'
AG-rich_RFP-AG1-3_F	5' - GGAGTCTAAACACACAGAGATGAGATGAGATGCGATATGGGCGAGCTGATCAAGGAG - 3'
ybaB_F	5' - CGTGATTGAGAGAGAAAACC - 3'
ybaB_R	5' - TAGGTTTCTCTCTCAATCACGGC - 3'
ybaB-AG_F	5' - CGTGATTCACACAGAAAACC - 3'
ybaB-AG_R	5' - TAGGTTTCTGTGTGAATCACGGC - 3'
ybaB+SD_F	5' - CGTGATTGGAGGAGAAAACC - 3'
ybaB+SD_R	5' - TAGGTTTCTCTCAATCACGGC - 3'
lon_F	5' - TTAAACTAAGAGAGAGCTC - 3'
lon_R	5' - TAGAGCTCTCTTAGTTTAAGC - 3'
lon-AG_F	5' - TTAAACTAACACACAGCTC - 3'
lon-AG_R	5' - TAGAGCTGTGTGTAGTTTAAGC - 3'
lon+SD_F	5' - TTAAACTAAGGAGGAGCTC - 3'
lon+SD_R	5' - TAGAGCTCCTCCTTAGTTTAAGC - 3'
rplY_F	5' - TTAAACTAAGAGAGAGCTC - 3'
rplY_R	5' - TAGAGCTCTCTTAGTTTAAGC - 3'
rplY-AG_F	5' - TTAAACTAACACACAGCTC - 3'
rplY-AG_R	5' - TAGAGCTGTGTGTAGTTTAAGC - 3'
rplY+SD_F	5' - TTAAACTAAGGAGGAGCTC - 3'
rplY+SD_R	5' - TAGAGCTCCTCCTTAGTTTAAGC - 3'
lrp_F	5' - AATACAGAGAGACAATAAT - 3'
lrp_R	5' - TAATTATTGTCTCTGTATTGC - 3'
lrp-AG_F	5' - AATACACACACAATAAT - 3'
lrp-AG_R	5' - TAATTATTGTGTGTGTATTGC - 3'

lrp+SD_F	5' - AATACAGGAGGACAATAAT - 3'
lrp+SD_R	5' - TAATTATTGTCCTCTGTATTGC - 3'
tdk-F	5' - GCCTGTGGCT - 3'
tdk-R	5' - TAAGCCACAGGCGC - 3'
ymcE-F	5' - GCAAATACGCACTGAC - 3'
ymcE-R	5' - TAGTCAGTGCATTTGCGC - 3'
yoaB-F	5' - ACAAAT - 3'
yoaB-R	5' - TAATTTGTGC - 3'
yqiC-F	5' - GAAACTAC - 3'
yqiC-R	5' - TAGTAGTTTCGC - 3'
hisL-F	5' - CATCAGTTGAATAAACATTACAGAGACTTT - 3'
hisL-R	5' - TAAAAGTCTCTGTGAATGTTTATTCAACTGATGGC - 3'
ilvL-F	5' - ATTCCTCGAACAAGATGCAAGAAAAGACAA - 3'
ilvL-R	5' - TATGTCTTTTCTTGCATCTGTTCGAAGGAATGC - 3'
ribF-F	5' - GATTTTCACTGTTTGTAGCCAGA - 3'
ribF-R	5' - TATCTGGCTCAAAACAGTGAAAATCGC - 3'
gmk-F	5' - ATGTAGGCTTTATTTTCGCTAATCACATACGAAAGATACT - 3'
gmk-R	5' - TAAGTATCTTTCGTATGTGATTAGCGAAAATAAAGCCTACATGC - 3'
yccA-F	5' - ACATCTGTCATAGAGAGTGACTCA - 3'
yccA-R	5' - TATGAGTCACTCTCTATGACAGATGTGC - 3'
fadR-F	5' - GCTGTGTTATGGAAATCTCAC - 3'
fadR-R	5' - TAGTGAGATTTCCATAACACAGCGC - 3'
ygiW-F	5' - GAGTAATAAAC - 3'
ygiW-R	5' - TAGTTTATTACTCGC - 3'
gapA-F	5' - AACCTTTTATTCCTAACAATAGCTGGTGAATA - 3'
gapA-R	5' - TATATCCACCAGCTATTTGTTAGTGAATAAAAGGTTGC - 3'
orn-F	5' - GTGAATGGGTGGAAAATAG - 3'
orn-R	5' - TACTATTTCCACCCATTCACGC - 3'
gyrB-F	5' - AACCCAAGTATAAATGAGCGAGAAAACGTTG - 3'
gyrB-R	5' - TACAACGTTTCTCGCTCATTATACTTGGGTTGC - 3'
nfuA-F	5' - TAAAAATAGTCAACTATTAGGCCATTAC - 3'
nfuA-R	5' - TAGTAATGGCCTAATAGTTGACTATTTTAGC - 3'
rbsD-F	5' - AGCGAAACGTTTCGCTGATGGAGAAAAA - 3'
rbsD-R	5' - TATTTTTCTCCATCAGCGAAAACGTTTCGCTGC - 3'
malP-F	5' - ACGTTGTGTGAAAATCTAAGAAAAGTGGAACCTCCT - 3'
malP-R	5' - TAAGGAGTCCACTTTTCTTAGATTTCAACACAACGTGC - 3'
lrhA-F	5' - GTAAGTGATAATAT - 3'
lrhA-R	5' - TAATATTACTACTACGC - 3'
gpsA-F	5' - AAGAACATCAGGATGCCTG - 3'
gpsA-R	5' - TACAGGCATCCTGATGTTCTTGC - 3'
ubiD-F	5' - GCAGATTAAGAGGCTAC - 3'
ubiD-R	5' - TAGTAGCCTCTTAATCTGCGC - 3'
deoB-F	5' - ATCAITTAATTTGAAGCACTGAGTACGGAGAACA - 3'
deoB-R	5' - TATGTTCTCCGTACTCAGTGCTTCAAATTTAAATGATGC - 3'
lpxD-F	5' - TTAAATAAGT - 3'
lpxD-R	5' - TAACTTATTTAAGC - 3'
eno-F	5' - TTAAGTAGTACTTGAGGAAAACCTA - 3'
eno-R	5' - TATAGGTTTCTCAAGTCACTAGTTAAGC - 3'
glnS-F	5' - GTTATACGTGTTTACGCTTTGAGGAATCCAC - 3'
glnS-R	5' - TAGTGGATTCTCAAAGCGTAAACAACGTATAACGC - 3'

rpoZ-F	5' – CATTTCCTCACCTGTGGAGCTTTTAAAG – 3'
rpoZ-R	5' – TACTTAAAAAGCTCCACAGGTGAAGAAATGGC – 3'
yobF-F	5' – GGTTAAACGAGGTACAGTTCTGTTT – 3'
yobF-R	5' – TAAAACAGAACTGTACCTCGTTTAAACCGC – 3'
yrbL-F	5' – AGATCGTCAGAAAAATTTCCAGGAGATGG – 3'
yrbL-R	5' – TACCATCTCCTGGAATATTTTCTGACGATCTGC – 3'
yefM-F	5' – TACAAAAGAGGAGATTGAC – 3'
yefM-R	5' – TAGTCAATCTCCTCTTTGTAGC – 3'
mgrB-F	5' – TAAGGTAGGTGAAACGGAGATTGGA – 3'
mgrB-R	5' – TATCCAATCTCCGTTTCACCTACCTTAGC – 3'
kdgR-F	5' – TAAAACACTCGCAATGAGGTGATAA – 3'
kdgR-R	5' – TATTATCACCTCATGCGAGTGTTTTAGC – 3'
slyD-F	5' – CCACTATCTTCCCATGCTCAGGAGATATC – 3'
slyA-R	5' – TATTCATCTCCTTATAATTAGCTTGGC – 3'
tabA-F	5' – ACGTTGTTTAATTAAGGAGTGGAAAA – 3'
tabA-R	5' – TATTTTCCACTCCTTAATTAACAACGTGC – 3'
yciN-F	5' – CCTTGTTAAGGAGTAAATA – 3'
yciN-R	5' – TATATTTACTCCTTAAACAAGGGC – 3'
ydcY-F	5' – ATGAATCGATAAATTCACAAGTGGCTAAGGAGAAAG – 3'
ydcY-R	5' – TACTTTCTCCTTAGCCACTTGTGAAATTTATCGATTATGC – 3'
yihD-F	5' – CAAGAGTAAGGAGCATCAC – 3'
yihD-R	5' – TAGTGATGCTCCTTACTCTTGGC – 3'
T7-nat5'-UTR-start5	5' – <b>TAATACGACTCACTATAGGGNN...DTKAAAGAGGACGAGAGC</b> – 3'
HindIII-T7 F	5' – GTGCTC <b>AAGCTTTAATACGACTCACTATAGGG</b> – 3'
5-4 codons-BamHI R	5' – CTGGCC <b>GGATCCCCCGTCGGGGTCTCTCGTCTCTTT</b> – 3'
HindIII-T7 Final PCR F	5' – GTGCTCAAGCTTTAATACGACTCACTATAGGG – 3'
CER Final PCR R	5' – GCTCTCGTCCGTCTCTTC – 3'

## 2. Методы

### 2.1. Манипуляции с ДНК: клонирование

#### 2.1.1. Определение концентрации ДНК в растворе

Для определения концентрации ДНК в растворе брали 5-10 мкл образца, содержащего ДНК (олигонуклеотиды, ПЦР-фрагменты, плазмидную ДНК) и находящегося в воде либо в буферном растворе для элюции EB (Qiagen), и доводили его до 100 мкл 0,1 М Tris-HCl (pH 8,0), затем измеряли при 260 нм поглощение полученного раствора. Далее делали пересчёт измеренного значения  $A_{260}$  в концентрацию ДНК в мкг/мкл (при необходимости и в пмоль/мкл), учитывая разведение.

#### 2.1.2. Разделение в агарозном геле ДНК-фрагментов

Для разделения в агарозном геле ДНК-фрагментов готовили 1% агарозный гель, добавляя 0,8-1,0 г агарозы к 80-100 мл раствора 1x TBE, затем нагревая до полного растворения агарозы в микроволновой печи, при этом избегая кипения. Далее после охлаждения до ~ 40-50°C к полученному раствору добавляли 5 мкл водного раствора бромида этидия в концентрации 1

мг/мл, перемешивали, заливали в специальную форму и вставляли гребёнки с нужным количеством ячеек. Когда гель застыл, его вынимали из формы, удаляли гребёнки и переносили в специальную «ванночку» с буфером 0,5-1x TBE. В ячейки наносили образцы для разделения и проводили электрофорез при значениях силы тока ~ 25-50 мА. Результаты разделения ДНК-фрагментов анализировали с использованием системы ChemiDoc с камерой и УФ-лампой.

### **2.1.3. Выделение ДНК-фрагментов из агарозного геля**

Для выделения ДНК-фрагментов из агарозного геля использовали набор реагентов для выделения ДНК-фрагментов из агарозного геля фирмы Fermentas по протоколу работы с фирменными реагентами. Фрагмент агарозного геля с целевой последовательностью ДНК массой ~ 100-150 мг переносили в чистую 1,5 мл пробирку, к нему добавляли 100-150 мкл буферного раствора для связывания BS (Fermentas) и плавил при 55°C, пока гель не растворится полностью. Далее полученный раствор переносили на колонку для выделения ДНК (Fermentas) и центрифугировали (ц/ф) на скорости 13200 оборотов в минуту (об./мин) в течение 30-60" (сек) при комнатной температуре (t°C). Затем делали промывку этой колонки 2 раза, нанося на неё 500 мкл буферного раствора для промывки WB (Fermentas). Элюцию ДНК с колонки проводили с помощью 30-50 мкл буферного раствора для элюции EB (Fermentas). Для определения концентрации ДНК по поглощению при 260 нм использовали прибор NanoDrop или методику определения концентрации ДНК в растворе.

### **2.1.4. Приготовление векторов и вставок**

Приготовление векторов и вставок проводили путём рестрикции эндонуклеазами [255]. Реакционную смесь готовили из расчёта на 20 мкл со следующим составом: 1-4 мкг ДНК (плазмидной или ПЦР-продукта), 2 мкл 10x буферного раствора, подходящего для эндонуклеазы рестрикции, 0,5-1 мкл (10 ед.) эндонуклеазы рестрикции, доводили очищенной водой (mQ или без нуклеаз (Nuclease-free water)) до 20 мкл.

Полученную реакционную смесь инкубировали при 37°C в течение 2-4 ч. Далее к смеси добавляли 4 мкл 6x буфера для нанесения ДНК на гель, переносили на агарозный гель и проводили электрофоретическое разделение ДНК-фрагментов. ДНК-фрагмент с нужной длиной вырезали из геля и экстрагировали с использованием набора реагентов для выделения ДНК-фрагментов из агарозного геля фирмы Fermentas.

### **2.1.5. Очистка фенолом и хлороформом и осаждение этанолом**

Смесь со вставками либо с векторами доводили до удобного объема (например, 100 мкл) и добавляли к ней фенол в соотношении 1:1, затем центрифугировали 5' (мин) на 13200 об./мин

при комнатной  $t^{\circ}\text{C}$  и отбирали верхний слой в новую пробирку. Далее добавляли хлороформ в соотношении 1:1, ц/ф 5' на 13200 об./мин, снова отбирали верхний слой и добавляли ацетат натрия (1/10 от объема) и 96% этанол из расчёта 3 объёма спирта в случае вставок или 2,5 объёма спирта в случае плазмид и 1 объём образца с ацетатом натрия. Полученную смесь затем перемешивали, сбрасывали капли и инкубировали при  $-80^{\circ}\text{C}$  в течение 1 ч либо при  $-20^{\circ}\text{C}$  ночь. Далее ц/ф 15' при  $4^{\circ}\text{C}$  на 13200 об./мин, удаляли супернатант и добавляли 70% этанол в соотношении 1:1 (объёму супернатанта). После ц/ф 5' при  $4^{\circ}\text{C}$  на 13200 об./мин полностью отбирали супернатант, не задевая осадок, и высушивали его по тягой. Далее растворяли осадок в 20-30 мкл очищенной воды (mQ/без нуклеаз (Nuclease-free water)) и измеряли концентрацию.

### **2.1.6. Очистка ПЦР-продуктов из реакционной смеси**

Для очистки ПЦР-продуктов из реакционной смеси использовали специальный набор реактивов QIAquick PCR Purification Kit (Qiagen), следуя методике от производителя. Этот набор также подходит и для очистки реакционной смеси после рестрикции эндонуклеазами. Результат очистки ДНК-образцов проверяли с помощью гель-электрофореза.

### **2.1.7. Лигирование**

Лигирование проводили согласно методике [255]. Реакционную смесь готовили из расчёта на 20 мкл со следующим составом: 30-100 нг вектора, 10-50 нг вставки в 3х-10х мольном избытке по отношению к количеству вектора, 2 мкл 10х буферного раствора для Т4 ДНК-лигазы, 1 мкл (1 ед./мкл) Т4 ДНК-лигазы, доводили очищенной водой (mQ/без нуклеаз (Nuclease-free water)) до 20 мкл. Лигазную смесь перемешивали и инкубировали 4-16 ч при  $16-20^{\circ}\text{C}$ . Для трансформации клеток штаммов JM109, BW25113, BL21 (DE3) *E. coli* брали 2-6 мкл полученной смеси.

В качестве альтернативной методики также использовали набор реактивов для быстрого лигирования Rapid DNA Ligation Kit. С его помощью готовили реакционную смесь из расчёта на 20 мкл со следующим составом: 10-100 нг вектора, 10-50 нг вставки в 3х-10х мольном избытке по отношению к количеству вектора, 4 мкл 5х буферного раствора для Т4 ДНК-лигазы, 1 мкл (5 ед./мкл) Т4 ДНК-лигазы, доводили очищенной водой (mQ или без нуклеаз (Nuclease-free water)) до 20 мкл. Лигазную смесь перемешивали и инкубировали 15' при  $22^{\circ}\text{C}$ . Для трансформации клеток штаммов JM109, BW25113, BL21 (DE3) *E. coli* брали 2-6 мкл полученной смеси.

### **2.1.8. Выделение плазмидной ДНК**

Для выделения плазмидной ДНК использовали набор реагентов для выделения плазмидной ДНК фирмы Fermentas по протоколу работы с фирменными реагентами. Для получения ночной культуры сколотую колонию клеток штаммов JM109, BW25113, BL21 (DE3)

*E. coli* с определённой плазмидной ДНК помещали в 2-5 мл среды LB с ампициллином в концентрации 100 мкг/мл, по которому далее проводилась селекция. После инкубации при перемешивании при 37°C в течение 16 ч клетки осаждали путём ц/ф 5' на 5000-7000 об./мин. Осаждённые клетки ресуспендировали в 250 мкл раствора для ресуспендирования RS (фирмы Fermentas), затем добавляли 250 мкл раствора для лизиса LS (Fermentas) и перемешивали, пока раствор не становился вязким. Далее к полученной смеси добавляли 350 мкл раствора для нейтрализации NS (Fermentas) и перемешивали, пока не появились белые хлопья. После осаждения суспензии путём ц/ф при 13200 об./мин в течение 10' при комнатной t°C супернатант отбирали и переносили его на колонку для выделения плазмидной ДНК (Fermentas), далее снова ц/ф и делали промывку этой колонки 2 раза, на каждом шаге добавляя 500 мкл буферного раствора для промывания WB (Fermentas). Элюцию плазмидной ДНК с колонки проводили с помощью 50 мкл буфера для элюирования EB (Fermentas) или очищенной водой (mQ/без нуклеаз (Nuclease-free water)). Определяли концентрацию выделенной плазмидной ДНК по поглощению при 260 нм с использованием прибора NanoDrop или методики определения концентрации ДНК в растворе. В качестве альтернативного набора реактивов для выделения плазмидной ДНК использовали QIAprep Spin Miniprep Kit (Qiagen) согласно инструкции производителя.

### **2.1.9. Полимеразная цепная реакция (ПЦР)**

Для проведения полимеразной цепной реакции (ПЦР) [256] смесь компонентов готовили из расчёта на 50 мкл в пробирке объёмом 0,5-0,75 мл со следующим составом: 5-50 нг матричной ДНК (плазмидной/ПЦР-продукта), 5 мкл 10x буферного раствора для Taq-полимеразы, по 5-25 пмоль каждого из двух праймеров, по 0,2 мМ каждого типа dNTP, 2,5 мМ MgCl<sub>2</sub>, 1 мкл (1 ед./мкл) Taq-полимеразы, доводили очищенной водой (mQ/без нуклеаз (Nuclease-free water)) до 50 мкл.

Полученную смесь тщательно перемешивали и помещали в специальный прибор для проведения ПЦР-амплификации Mastercycler Gradient фирмы Eppendorf или T100 Thermal Cycler фирмы Bio-Rad. Реакцию ПЦР проводили при следующих параметрах: предварительный прогрев 2' при 94°C, 25-35 циклов амплификации: денатурация ДНК 45" при 94°C, отжиг праймеров 45" при 54-64°C (в зависимости от температуры отжига, рассчитанной по последовательности для каждого из пары праймеров), элонгация при 72°C 20"-4' (из расчёта процессивности полимеразы при синтезе в среднем 10<sup>3</sup> нт за 1'), после циклов амплификации проводили ещё достройку продуктов при 72°C 1-5'.

В качестве альтернативной методики также использовали 2x реакционную смесь с Taq ДНК-полимеразой или Q5 ДНК-полимеразой высокой точности фирмы NEB. При этом для проведения ПЦР смесь компонентов готовили из расчёта на 20 мкл в пробирке объёмом 0,5-0,75 мл со следующим составом: 10 мкл 2x реакционной смеси с Taq ДНК-полимеразой или Q5 ДНК-

полимеразой, 5-50 нг матричной ДНК (плазамидной/ ПЦР-продукта), по 5-25 пмоль каждого из двух праймеров для ПЦР, доводили очищенной водой (mQ/без нуклеаз (Nuclease-free water)) до 20 мкл. В данном случае параметры для проведения ПЦР были следующие: предварительный прогрев при 98°C 1', 25 циклов амплификации: денатурация ДНК при 98°C 30", отжиг праймеров при 54-64°C (в зависимости от температуры отжига, рассчитанной по последовательности для каждого из пары праймеров) 30", элонгация 72°C 20"-4' (из расчёта процессивности полимеразы при синтезе в среднем 10<sup>3</sup> нт за 1'), далее проводили ещё достройку продуктов при 72°C 1-5'.

Анализ ПЦР-продуктов проводили с помощью электрофореза в 10-15% ПААГ или 1-2% агарозном геле. При этом отбирали по 5 мкл смеси после ПЦР-амплификации и добавляли 1 мкл 6x буфера для нанесения ДНК на гель, и наносили на гель для разделения ДНК-фрагментов.

### **2.1.10. Создание репортёрной конструкции pRFPCER BamHI**

Для получения вектора pRFPCER BamHI для последующей вставки рандомизированных или природных 5'-UTR перед геном *cer* взяли успешно использованную ранее конструкцию pRFPCER [137, 156] и провели её рестрикцию по сайту эндонуклеазы Esp3I (BsmBI). В результате полученный линейный вектор напрямую лигировали с помощью T4 ДНК-лигазы с заранее отождёнными друг на друга комплементарными олигонуклеотидами BamHI F и BamHI R, содержащими дублирующие первые 6 кодонов кодирующей области гена *cer*, следующими за старт-кодомом ATG (с +2 по +7) и сайт рестрикции BamHI. Таким образом получался вектор с началом CDS гена *cer* следующего вида: старт-кодон ATG, далее 6 дублирующих кодонов и сайт рестрикции BamHI, за которым находился кодон CTG (вместо стартового ATG) и те же 6 кодонов, что и после инициаторного, за которыми идёт обычная кодирующая область гена *cer*, идентичная той, что в конструкции pRFPCER, следующая с +7 кодона.

### **2.1.11. Создание отдельных конструкций с нестандартными и их мутированными вариантами или природными 5'-UTR перед генами *cer* или *rfp***

Для получения отдельных конструкций с нестандартными и их мутированными вариантами или природными 5'-UTR перед генами *cer* или *rfp* использовали плазмиду pRFPCER, разрезанную с помощью эндонуклеаз рестрикции NdeI и SacII. Полученный линейный вектор напрямую лигировали с комплементарными олигонуклеотидами, заранее отождёнными друг на друга и содержащими необходимый вариант 5'-UTR.

### **2.1.12. Секвенирование ДНК**

#### **2.1.12.1. Секвенирование плазмидной ДНК или ПЦР-продуктов**

Для проверочного секвенирования по Сэнгеру плазмидной ДНК или ПЦР-продуктов в

пробирку объёмом 0,5-0,75 мл добавляли 300 нг ДНК-матрицы для анализа и 5 пмоль праймера для секвенирования нужного фрагмента матрицы, выпаривали досуха при 80°C, затем оправляли на анализ в ЦКП «Геном».

#### **2.1.12.2. Высокопроизводительное секвенирование (NGS)**

Секвенирование библиотеки ампликонов в случае рандомизации 4N в спейсерном участке 5'-UTR и 20N, 30N 5'-UTR проводилось с помощью Ion Torrent [158] PGM (Life Technologies) с использованием специального набора реактивов Ion PGM™ Template OT2 200 Kit для эмульсионной ПЦР-амплификации и чипов Ion Chips 314 или 318 наряду с набором реактивов для секвенирования Ion PGM™ Sequencing 200 Kit v2, следуя инструкциям производителя.

Секвенирование библиотеки ампликонов в случае природных 5'-UTR проводилось с помощью MGISEQ-2000 (MGI Tech) после подготовки с использованием специального набора реактивов MGIEasy Universal DNA Library Prep (MGI Tech, Китай) наряду с набором реактивов для секвенирования DNBSEQ-G400RS High-throughput Sequencing Set (FCL PE100, MGI Tech, Китай), следуя инструкциям производителя.

### **2.2. Манипуляции с клетками**

#### **2.2.1. Приготовление компетентных клеток *Escherichia coli***

Компетентные клетки готовили согласно методике [257]. Для получения культуры клеток штаммов JM109, BW25113, BL21 (DE3) *Escherichia coli* в логарифмической фазе роста свежую сколотую колонию клеток помещали в 50 мл среды SOB и растили в термостате на 18°C с перемешиванием (125 об./мин) до достижения оптической плотности  $A_{600} \sim 0,5-0,6$ . Затем клеточную культуру перемещали в лёд на 10', далее осаждали с помощью ц/ф в специальных пробирках в роторе JA 14 в течение 10' на скорости 5000 об./мин при 4°C. После удаления супернатанта проводили промывку: осадок клеток ресуспендировали в 20 мл буферного раствора ТВ, охлажденного до 4°C, клеточную суспензию инкубировали при 0°C в течение 10' при помешивании, далее ц/ф в роторе JA 14 в течение 10' на скорости 5000 об./мин при 4°C. После удаления супернатанта стадию промывки повторяли. Далее клетки без супернатанта ресуспендировали в 4 мл буферного раствора ТВ, затем добавляли 280 мкл ДМСО (до концентрации 7%) и переносили в лёд на 10'. Полученную клеточную суспензию разделяли на аликвоты по 200 мкл в 1,5 мл стерильных пробирках, сразу замораживали в жидком азоте и помещали на температуру -80°C для длительного хранения.

При приготовлении электрокомпетентных клеток необходимо было максимально эффективнее избавиться от солей. Для этого также растили простые химически компетентные клетки в среде до стадии экспоненциального роста, переносили в специальные пробирки и

промывали 10% раствором глицерина, заранее охлаждённым, а после ц/ф.

### 2.2.2. Трансформация и электропорация компетентных клеток *Escherichia coli*

Для трансформации компетентные клетки размораживали постепенно во льду, добавляли 2-6 мкл лигазной смеси либо 1-5 мкл (~ 100 нг) плазмидной ДНК, растворённой в буферном растворе для элюции EB (Fermentas/Qiagen) или очищенной воде (mQ/без нуклеаз (Nuclease-free water)). После инкубирования в течение 30' при 0°C смесь клеток с ДНК нагревали при 42°C в течение 40-50" в термостате «Гном», затем охлаждали до 0°C и добавляли 800 мкл стерильной среды SOC. Далее трансформационную смесь помещали в термостат на 37°C на 1 час. Для высевания на чашку Петри с твёрдой средой LB с добавлением 100 мкг/мл ампициллина брали полный объём этой смеси (в случае трансформации лигазной смесью) или аликвоту в 100–200 мкл (в случае трансформации плазмидной ДНК) и инкубировали 16 ч при 37°C в термостате.

Электропорацию проводили в соответствии с детальным описанием процедуры трансформации, предоставленным фирмой Bio-Rad, с использованием высоковольтного мини-электрода и специальной кюветы для образца этой же фирмы.

### 2.2.3. Сортировка клеток

Клетки, трансформированные библиотекой плазмид или отдельными конструкциями, выращивали в течение 16 ч при температуре 37°C в жидкой среде LB с добавлением 100 мкг/мл ампициллина при перемешивании или на твёрдой среде LB с добавлением 100 мкг/мл ампициллина в чашках Петри. Затем клетки скручивали в жидкой среде или смывали буферным раствором PBS с твёрдой среды. Далее промывали раствором PBS дважды и разводили в нём же клетки до  $A_{600} \sim 0.004$ . Подготовленные клетки сортировали с помощью Becton Dickinson FACSAria III, при этом одновременно проводился мониторинг интенсивностей флуоресценции белков RFP и CER на длинах волн в нанометрах (нм) 561/582 нм и 405/530 нм соответственно.

Далее на полученном с помощью флуоресцентной проточной цитометрии (FACS) графическом изображении клеток, где по осям располагались значения (в логарифмической шкале) интенсивностей флуоресценции белков CER (ось X) и RFP (ось Y), выделялись наклонными рамками фракции клеток приблизительно с одинаковым соотношением CER/RFP. Клеточные фракции по отдельности собирались в новые пробирки в соответствии с указанными наклонными рамками на графике. Число клеток, собранных в каждой фракции, пропорционально общему вкладу клеток с определённым соотношением интенсивностей флуоресценции CER/RFP, избегая искусственного обогащения клетками с редкими показателями соотношения CER/RFP. Общее число отсортированных клеток в несколько раз покрывало разнообразие анализируемого набора конструкций.

Отсортированные фракции клеток растили при перемешивании в течение 16 ч при температуре 37°C в жидкой среде LB с добавлением 100 мкг/мл ампициллина. Затем из клеток фракций выделяли библиотеки плазмид и использовали для ПЦР-амплификации интересующего участка с помощью праймеров: T5 CER F и CER R 4N, причём последний содержит 6 вариантов баркодных последовательностей длиной 4 нт (в случае библиотеки 4N), P1 и одного из набора A29-A60, содержащего 32 варианта баркодных последовательностей длиной 9 нт (в случае библиотек 20N и 30N), HindIII-T7 Final PCR F и CER Final PCR R (в случае библиотек природных 5'-UTR). Полученные ПЦР-продукты (ампликоны) после проверки с помощью электрофореза далее подвергались высокопроизводительному секвенированию (NGS).

#### **2.2.4. Измерение флуоресценции с помощью спектрофлуориметра**

Из ночной культуры клеток, выращенной в течение 16 ч при температуре 37°C в жидкой среде LB с добавлением 100 мкг/мл ампициллина, отбирали 200 мкл и переносили в 96-луночный планшет. После ц/ф на 4000 об./мин в течение 7-10' и удаления супернатанта осаждённые клетки промывали следующим образом: ресуспендировали в 200 мкл 0,9% NaCl и ц/ф в тех же условиях, затем супернатант удаляли. Стадию промывки повторяли 2 раза. Далее после полного удаления супернатанта клетки ресуспендировали в 200 мкл свежего раствора 0,9% NaCl и переносили в 96-луночный планшет с чёрными ячейками для измерения флуоресценции с помощью планшетного ридера Victor X5 (PerkinElmer) на длинах волн 531/595 нм и 430/486 нм соответственно для RFP и CER. После вычитания фонового сигнала для раствора 0,9% NaCl без клеток (в среднем 3000 для RFP и 1000 для CER) из измеренных значений вычисляли соотношение интенсивностей флуоресценций CER/RFP и нормировали его на соответствующее значение для контрольной плазмиды с идентичными 5'-UTR перед генами белков RFP и CER.

### **2.3. Биоинформатический анализ результатов NGS**

Общая обработка данных с отсеквенированными последовательностями проводилась с использованием программ, написанных на языке Python [207] и адаптированных для работы с большими массивами полученных данных. В то время как статистический анализ и построение графиков выполнялись в программах, написанных на языке R [208].

#### **2.3.1. Анализ результатов для контрольного набора конструкций**

Для извлечения последовательностей 5'-UTR из сырых данных с ридами, полученными с помощью NGS, в формате FASTQ все прочтения длиной от 50 до 150 нт использовали для поиска участков, отличающихся не более, чем двумя несовпадающими нуклеотидами (известные как мисматчи/"mismatches"), при этом включая вставки/делеции (инделлы/"indels"), в следующих

константных последовательностях `CTCGTCCGTCTCTTTCAT` и `CCGCGGCT`, окружающих различные варианты 5'-UTR из контрольного набора конструкций, остальные риды удалялись. Во всех отсортированных фракциях искали каждый вариант последовательностей, подсчёт правильных ридов проводился с сохранением вариантов с точным совпадением с известными последовательностями. Собранные данные обобщали в виде таблицы, в которой подсчитывались случаи появления каждого варианта последовательности в каждой из отсортированных фракций, затем по этим данным строилась диаграмма. Тогда как один и тот же вариант последовательности распределён по нескольким фракциям, в эксперименте каждому варианту присваивали средневзвешенный номер его фракции, исходя из распределения Гаусса.

### **2.3.2. Анализ результатов для 4N библиотеки**

#### **2.3.2.1. Обработка полученных ридов и их распределение по фракциям**

Для извлечения последовательностей 5'-UTR из сырых данных с ридами, полученными с помощью NGS, в формате FASTQ все прочтения длиной от 50 до 100 нт использовали для поиска участков, отличающихся не более, чем двумя несовпадающими нуклеотидами, включая вставки/делеции, в следующих константных последовательностях `GGCACACACCGGAGC` и `CATATGAAAGAGACGGACGAGAGCGGCCTGGTGA`, фланкирующих участок рандомизации 4N, остальные риды удалялись. Во всех отсортированных фракциях искали каждый вариант последовательности. Собранные данные обобщали в виде таблицы, в которой подсчитывались случаи появления каждого варианта последовательности в каждой из отсортированных фракций (Приложение 2). Тогда как один и тот же вариант последовательности распределён по нескольким фракциям, в эксперименте каждому варианту присваивали средневзвешенный номер его фракции, исходя из распределения Гаусса.

Из-за сильно отличающегося количества вариантов во фракциях и относительно узкого общего диапазона эффективности трансляции белка CER в эксперименте (Рис. 53) мы разделили все варианты 5'-UTR на две группы примерно одинакового размера. В группу (класс) с низкой эффективностью трансляции определили 125 из 249 вариантов 5'-UTR мРНК, у которых средневзвешенные значения фракций лежали в диапазоне от 1,47 до 3,07, тогда как в группу с высокой эффективностью вошли 124 из 249 вариантов 5'-UTR мРНК, средние значения фракций которых были в пределах от 3,08 до 5. Дальнейший анализ проводили для этих 2-х групп.

#### **2.3.2.2. Оценка нуклеотидного состава последовательностей 5'-UTR**

Для каждой отдельной позиции в 4 нт рандомизированном участке подсчитаны частоты встречаемости каждого из четырёх типов нуклеотидов (A, U, G, C) отдельно для двух групп с эффективно и слабо транслирующимися мРНК. По данным частотам построены графики для

каждого типа нуклеотидов во всех четырёх позициях для обеих групп.

### 2.3.2.3. Оценка стабильности вторичной структуры мРНК

Минимальную энергию укладки – MFE – рассчитывали для участка мРНК, охватывающего всю 5'-UTR длиной 22 нт и начало кодирующей области гена *cer* длиной 50 нт, с использованием программы RNAfold версии 2.1.7 из пакета Vienna RNA с параметрами по умолчанию [152, 258]. Для того, чтобы оценить статистическую значимость различий в MFE между эффективно и плохо транслируемыми мРНК при условии сохранения частот нуклеотидов в каждой позиции, использовали моделирование Монте-Карло. В ходе него сгенерировали  $10^3$  пар наборов перемешанных вариантов последовательностей путём случайной перестановки нуклеотидов в каждой позиции отдельно для обеих групп. Расчёт MFE выполнен для каждого случайно перемешанного варианта, распределение MFE построено для каждого набора, а затем рассчитана статистика Колмогорова-Смирнова для пар этих смоделированных распределений. В итоге получили распределение, представленное на Рисунке. Статистика Колмогорова-Смирнова для реальной пары групп последовательностей рассчитана и использована для оценки случайности наблюдаемой разницы в виде p-value. Полученное значение p-value составляет 0,001, свидетельствующее о том, что наблюдаемая разница в MFE между двумя группами не случайна.

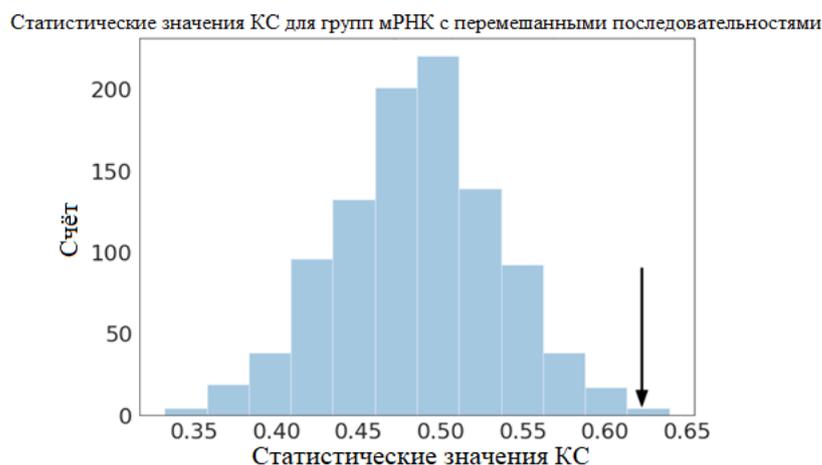


Рисунок. Гистограмма статистических значений Колмогорова-Смирнова (КС) для групп мРНК с перемешанными последовательностями. Последовательности в двух группах перемешаны 1000 раз, в результате чего получили наборы перемешанных последовательностей с одинаковым количеством случайных вариантов и одинаковыми частотами нуклеотидов в каждой позиции. Для каждой пары наборов распределения энергии вторичной структуры сравнивались, используя статистику Колмогорова-Смирнова. Статистическое значение КС указано чёрной стрелкой для реальных последовательностей в эксперименте. Рассчитанное значение p-value составляет 0,001.

#### **2.3.2.4. Оценка сходства последовательности 5'-UTR с SD**

Для оценки способности отдельного спейсерного участка к гибридизации с 3'-концевой областью 16S рРНК рассчитывали свободную энергию гибридизации фрагментов 5'-UTR, содержащих рандомизированный участок, CNNNNCAUA, с последовательностью анти-SD, CACCUCU, используя программу RNAfold из пакета Vienna RNA [152, 258]. Для обеих групп мРНК построены распределения энергий взаимодействия с последовательностью анти-SD.

#### **2.3.3. Анализ результатов для 20N и 30N библиотек**

##### **2.3.3.1. Обработка полученных ридов и их распределение по фракциям**

Для извлечения последовательностей 5'-UTR из сырых данных по ридам, полученным с помощью NGS, в формате FASTQ все прочтения длиной от 50 до 150 нт использовали для поиска участков, отличающихся не более, чем двумя несовпадающими основаниями, включая вставки/делеции, в следующих константных последовательностях CTCGTCCGTCTCTTTCAT и CCGCGGCT, окружающих участки рандомизации 20N и 30N, остальные риды удалялись. Во всех отсортированных фракциях искали каждый вариант последовательности, отбор ридов по длине проводился только с возможным отклонением в 1 нт от целевых длин 20 и 30 нт соответственно. Собранные данные были обобщены в виде таблицы, в которой подсчитывались случаи появления каждого варианта последовательности в каждой из отсортированных фракций.

Для исправления ошибок мы создали собственный рабочий процесс исправления ошибок, основанный на поиске похожих вариантов в разреженном наборе непохожих вариантов последовательностей. А именно, для всех вариантов рассчитывали расстояние Левенштейна (LD), значение которого сообщает, насколько различаются две взятые последовательности, и обозначает количество изменений, которое нужно внести в одну последовательность, чтобы получить другую, при этом допускались нуклеотидные замены и вставки/делеции. Следовательно, чем меньше LD, тем более схожи последовательности. Далее варианты последовательностей, отличающиеся по LD не более, чем на 3, объединялись. В ходе процедуры слияния варианты, встречающиеся более 10 раз, предпочитали считать правильными за счёт тех, что встречались реже. Точно так же варианты ожидаемой длины считались правильными за счёт тех, что имели одиночную вставку/делецию. Коррекция ошибок применялась только для библиотек, содержащих участки рандомизации 20N и 30N.

Тогда как один и тот же вариант последовательности распределён по нескольким фракциям, в эксперименте каждому варианту присваивали средневзвешенный номер его фракции, исходя из распределения Гаусса, которое в большинстве случаев было узким.

##### **2.3.3.2. Оценка нуклеотидного состава последовательностей 5'-UTR**

Для каждой отдельной позиции в 20 нт или 30 нт рандомизированном участке посчитали частоты встречаемости каждого из четырёх типов нуклеотидов отдельно для всех мРНК из каждой фракции. По данным частотам построили графики для каждого типа нуклетидов во всех 20-ти или 30-ти позициях для каждой фракции.

### 2.3.3.3. Оценка стабильности вторичной структуры мРНК

Для анализа стабильности вторичной структуры последовательностей мРНК использовали FASTA-формат вариантов последовательностей 5'-UTR с константными GG перед участками рандомизации 20N или 30N и первыми 50 нт кодирующей области гена *cer*. Для расчёта минимальной энергии укладки ( $\Delta G$ ) участка, охватывающего всю 5'-UTR длиной 22 нт (в случае библиотеки 20N) или 32 нт (в случае библиотеки 30N) и начало кодирующей области гена *cer* длиной 50 нт, применили программу RNAfold из пакета Vienna RNA версии 2.2.4 [258]. Для расчёта показателей спаривания оснований (plfold score) использовали программу RNAplfold из того же пакета [258] со значением параметра “-W” 30 нт. Для каждого нуклеотида 5'-UTR рассчитывали значения показателей образования пары между ним и последующим основанием в 5'- и 3'-направлениях для «окна» длиной 41 нт (в случае библиотеки 20N) или 51 нт (в случае библиотеки 30N). Для расчёта энергии укладки в пределах «скользящего окна» каждая ( $i, \dots, i+29$ ) подпоследовательность сворачивалась с помощью программы RNAfold для каждой начальной позиции  $i$  в диапазоне значений (1, ..., 20) для библиотеки 20N и (1, ..., 30) для библиотеки 30N.

### 2.3.3.4. Анализ SD-подобных последовательностей в 5'-UTR

Для оценки способности каждой 5'-UTR из обеих библиотек 20N и 30N к гибридизации с 3'-концевой областью 16S рРНК (SD score) использовали «скользящее окно» в пределах ( $i, \dots, i+8$ ) и весовые коэффициенты позиционно-специфической матрицы оценок (PSSM), взятые из базы данных Transterm Database [250], ID T0030, значение оценки (score) лежало в диапазоне от 0 до 1 для  $i$  из набора (1, ..., 22) в случае библиотеки 20N и (1, ..., 32) в случае библиотеки 30N. Расположение последовательности SD в рандомизированных 5'-UTR определяли как положение подпоследовательности с максимальной оценкой SD score.

Для того, чтобы проанализировать влияние более, чем одной последовательности SD на эффективность трансляции по сравнению с одной SD, мы рассчитали показатель SD score для реальных 5'-UTR и виртуальных последовательностей 5'-UTR, состоящих из цепочки поли(C) с одиночной SD, обнаруженной как последовательность с максимальной оценкой SD score в реальных 5'-UTR, для сопоставления. После вычитания последних из первых получали оценку SD score для 5'-UTR после виртуального исключения SD с наилучшей оценкой.

## **2.3.4. Анализ результатов для библиотеки природных 5'-UTR**

### **2.3.4.1. Обработка полученных ридов и их распределение по фракциям**

Для извлечения последовательностей 5'-UTR из сырых данных по ридам, полученным с помощью NGS, в формате FASTQ все прочтения длиной от 50 до 150 нт использовали для поиска участков, отличающихся не более, чем двумя несовпадающими основаниями, включая вставки/делеции, в окружающих 5'-UTR константных последовательностях, остальные риды удалялись. При этом отсечение ридов проводили с помощью программы cutadapt версии 4.0 [259], а их объединение с usearch версии 11.0.667 [260]. Объединённые прочтения картировали на референсные конструкции с использованием usearch. В результате сформировали таблицу, отражающую представленность каждого варианта последовательности 5'-UTR во всех 8 проанализированных фракциях.

Распределение ридов между фракциями соответствовало распределению Гаусса с ярко выраженным пиком. Таким образом, каждому варианту присваивали номер фракции, в которой количество прочтений конкретной последовательности было максимальным.

Фракции и соответствующие им варианты разделили на две группы по соотношению интенсивностей флуоресценции CER/RFP. В «неэффективную» группу по фракциям и последовательностям 5'-UTR вошли фракции F1-F4, для которых соотношение CER/RFP, а значит, и эффективность трансляции не превышали таковых для контрольной плазмиды. Фракции F5-F8, напротив, определялись как «эффективная» группа, так как измеренное соотношение CER/RFP в них было выше, чем у контрольной плазмиды.

### **2.3.4.2. Оценка стабильности вторичной структуры мРНК и анализ SD-подобных последовательностей в 5'-UTR**

Свободная энергия сворачивания вторичной структуры РНК  $\Delta G$  при 37°C для библиотеки 5'-UTR с добавлением первых 16 нт ORF гена *cer* (AUGAAAGAGGACGAGA) рассчитывалась с помощью RNAfold [261]. Для оценки эффективности взаимодействия всех вариантов последовательностей 5'-UTR с анти-SD (ACCUCUU) применили программу RNAduplex версии 2.5.1 из пакета Vienna RNA [141].

## БЛАГОДАРНОСТИ

Автор выражает глубокую благодарность своему научному руководителю д.х.н., профессору РАН, член-корреспонденту РАН Сергиеву Петру Владимировичу, позволившему работать под его началом и научившему решать сложные научные задачи, стремиться разобраться в деталях исследуемых процессов, ответственно и с разных сторон подходить к решению поставленных задач, не бояться трудностей и ошибок, а наоборот, выяснять причины их возникновения, искать способы их преодолеть и не повторять. Благодаря комфортной психологической атмосфере, создаваемой моим наставником, даже в сложных ситуациях и непростых условиях всегда было желание спешить в лабораторию, трудиться на всеобщее благо, порой не замечая времени, дней недели, усталости. Его заинтересованность в моих результатах воодушевляла на активную и плодотворную работу над экспериментами, а регулярная помощь и поддержка на разных этапах, постоянный разбор сложностей и ошибок позволили проделать длинный путь по осуществлению всех поставленных задач и создать основу для представленной диссертации.

Особую благодарность заслуживает д.х.н., с.н.с. Остерман Илья Андреевич. Когда я впервые пришла в эту лабораторию, будучи студенткой третьего курса, Илья Андреевич с вниманием, заботой и терпением обучал меня правильной постановке экспериментов, анализу полученных результатов, поиску решений при возникновении трудностей и дальнейших новых направлений работы, он постоянно делился своими навыками и опытом, что легло в основу при моём становлении как самостоятельного учёного на протяжении уже одиннадцати лет.

Глубокую признательность заслуживает заведующая лабораторией структуры и функций рибонуклеиновых кислот, где начиналось выполнение данной работы, д.х.н., академик РАН Донцова Ольга Анатольевна за ценные советы, постоянный интерес к работе и жизни сотрудников, регулярную помощь и поддержку в решении важных вопросов, возникающих в научных исследованиях и образовательном процессе. В её лаборатории всегда царит тёплая и душевная рабочая атмосфера, всегда много работающих студентов, аспирантов, сотрудников, и все готовы с радостью помочь, обучить, поделиться опытом, откликнуться в трудную минуту. И это вдохновляет на активную работу, стимулирует желание творить, помогать в ответ и быть частью такого замечательного коллектива.

Автор также выражает благодарность всем соавторам статей, которые легли в основу написания представленной диссертационной работы. Отдельной благодарности заслуживает весь коллектив лаборатории, студент Слесарчук А. Н. за выполнение отдельных экспериментов.

Автор признателен Зацепину Т. С. за синтез олигонуклеотидов, научным коллективам из ФНКЦ ФХМ им. Ю. М. Лопухина и ИХБФМ СО РАН, ЦКП «Геномика» в Новосибирске за

секвенирование библиотек образцов и коллегам из ИППИ им. А. А. Харкевича РАН за активное участие в обработке больших массивов данных, за применение актуальных биоинформатических и статистических подходов, и обучение им.

Автор бесконечно благодарен всей своей большой семье, особенно мужу и детям, родителям, сестре за моральную поддержку, помощь во всех начинаниях, постоянную веру, что всё получится, понимание и любовь, позволившие мне погружаться в мир науки и достигать всё новых успехов в ней.

## СПИСОК ЛИТЕРАТУРЫ

1. Brenner S., Jacob F., Meselson M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis // *Nature*, 1961. Т. 190. С. 576–581.
2. Gros F. и др. Unstable ribonucleic acid revealed by pulse labelling of *Escherichia coli* // *Nature*, 1961. Т. 190. С. 581–585.
3. Bradshaw R.A., Stahl P.D. Messenger RNA (mRNA): The Link between DNA and Protein // *Encyclopedia of Cell Biology*, 2016. С. 341-345.
4. Jacques N., Dreyfus M. Translation initiation in *Escherichia coli*: old and new questions // *Mol. Microbiol.*, 1990. Т. 4. № 7. С. 1063-7.
5. Steitz J.A. Polypeptide chain initiation: nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA // *Nature*, 1969. Т. 224. № 5223. С. 957-64.
6. Laursen B.S. и др. Initiation of protein synthesis in bacteria // *Microbiol. Mol. Biol. Rev.*, 2005. Т. 69. С. 101–123.
7. Geissmann T., Marzi S., Romby P. The role of mRNA structure in translational control in bacteria // *RNA Biology*, 2009. Т. 6. № 2. С. 153-160.
8. Jagodnik J., Chiaruttini C., Guillier M. Stem-Loop Structures within mRNA Coding Sequences Activate Translation Initiation and Mediate Control by Small Regulatory RNAs // *Mol. Cell*, 2017. Т. 68. № 1. С. 158-170.e3.
9. Shine J., Dalgarno L. Determinant of cistron specificity in bacterial ribosomes // *Nature*, 1975. Т. 254. С. 34–38.
10. May E.E., Vouk M.A., Bitzer D.L. Classification of *Escherichia coli* K-12 Ribosome Binding Sites // *IEEE Eng. Med. Biol. Mag.*, 2006. Т. 25. № 1. С. 90-7.
11. Munita J.M., Arias C.A. Mechanisms of Antibiotic Resistance // *Microbiol. Spectr.*, 2016. Т. 4. № 2. С. 1-37.
12. Reeve B. и др. Predicting translation initiation rates for designing synthetic biology // *Front. Bioeng. Biotechnol.*, 2014. Т. 2. № 1. С. 1-6.
13. Shine J., Dalgarno L. The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites // *Proc. Natl. Acad. Sci. USA*, 1974. Т. 71. № 4. С. 1342-6.
14. Boni I.V. и др. Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1 // *Nucleic Acids Res.*, 1991. Т. 19. С. 155–162.
15. Lauber M.A., Rappsilber J., Reilly J.P. Dynamics of Ribosomal Protein S1 on a Bacterial Ribosome with Cross-Linking and Mass Spectrometry // *Mol. Cell. Proteomics*, 2012. Т. 11. № 12. С. 1965-76.

16. Komarova E.S. и др. Flow-Seq Method: Features and Application in Bacterial Translation Studies // *Acta Naturae*, 2022. Т. 14. № 4. С. 20-37.
17. Komarova E.S. и др. Flow-Seq Evaluation of Translation Driven by a Set of Natural *Escherichia coli* 5'-UTR of Variable Length // *International Journal of Molecular Sciences*, 2022. Т. 23. № 20. С. 12293.
18. Komarova E.S. и др. Influence of the spacer region between the Shine-Dalgarno box and the start codon for fine-tuning of the translation efficiency in *Escherichia coli* // *Microbial Biotechnology*, 2020. Т. 13. № 4. С. 1254-1261.
19. Evfratov S.A. и др. Application of sorting and next generation sequencing to study 5'-UTR influence on translation efficiency in *Escherichia coli* // *Nucleic Acids Research*, 2017. Т. 45. № 6. С. 3487-3502.
20. Korostelev A. и др. Interactions and dynamics of the Shine-Dalgarno helix in the 70S ribosome // *Proc. Natl. Acad. Sci.*, 2007. Т. 104. № 43. С. 16840-3.
21. Jacobs G.H. и др. Transterm: a database of messenger RNA components and signals // *Nucleic Acids Research*, 2000. Т. 28. № 1. С. 293-295.
22. Saier M.H. Jr. Understanding the genetic code // *J. Bacteriol.*, 2019. Т. 201. С. e00091-19.
23. Frank J. The ribosome - a macromolecular machine par excellence // *Chem. Biol.*, 2000. Т. 7. № 6. С. R133-R141.
24. Fu Y. и др. Most RNAs regulating ribosomal protein biosynthesis in *Escherichia coli* are narrowly distributed to Gammaproteobacteria // *Nucleic Acids Research*, 2013. Т. 41. № 6. С. 3491-3503.
25. Dunkle J.A. и др. Structures of the bacterial ribosome in classical and hybrid states of tRNA binding // *Science*, 2011. Т. 332. № 6032. С. 981-984.
26. de Crécy-Lagard V., Jaroch M. Functions of Bacterial tRNA Modifications: From Ubiquity to Diversity // *Trends Microbiol.*, 2021. Т. 29. № 1. С. 41-53.
27. Yashiro Y. и др. Mechanism of aminoacyl-tRNA acetylation by an aminoacyl-tRNA acetyltransferase AtaT from enterohemorrhagic *E. coli* // *Nat. Commun.*, 2020. Т. 11. № 1. С. 5438.
28. Shepherd J., Ibba M. Bacterial transfer RNAs // *FEMS Microbiol. Rev.*, 2015. Т. 39 № 3. С. 280-300.
29. Tharp J.M., Ehnbohm A., Liu W.R. tRNA<sup>Pyl</sup>: Structure, function, and applications // *RNA Biol.*, 2018. Т. 15. № 4-5. С. 441-452.
30. Steiner R.E., Ibba M. Regulation of tRNA-dependent translational quality control // *IUBMB Life*, 2019. Т. 71. № 8. С. 1150-1157.
31. Opron K., Burton Z.F. Ribosome Structure, Function, and Early Evolution // *Int. J. Mol. Sci.*, 2019. Т. 20. № 1. С. 40.
32. Kim Y., Opron K., Burton Z.F. A tRNA- and Anticodon-Centric View of the Evolution of Aminoacyl-tRNA Synthetases, tRNAomes, and the Genetic Code // *Life*, 2019. Т. 9. № 2. С. 37.

33. Koh C.S., Sarin L.P. Transfer RNA modification and infection – Implications for pathogenicity and host responses // *BBA – Gene Regulatory Mechanisms*, 2018. Т. 1861. № 4. С. 419–432.
34. Hong S. и др. Mechanism of tRNA-mediated +1 ribosomal frameshifting // *PNAS*, 2018. Т. 115. № 44. С. 11226-11231.
35. Lei L., Burton Z.F. Evolution of the genetic code // *Transcription*, 2021. Т. 12. № 1. С. 28-53.
36. Berg M.D., Brandl C.J. Transfer RNAs: diversity in form and function // *RNA Biol.*, 2021. Т. 18. № 3. С. 316-339.
37. Milyn P., Rodnina M.V. Kinetic control of translation initiation in bacteria // *Critical Reviews in Biochemistry and Molecular Biology*, 2012. Т. 47. № 4. С. 334–348.
38. Wilson D.N. Ribosome-targeting antibiotics and mechanisms of bacterial resistance // *Nature Reviews Microbiology*, 2014. Т. 12. С. 35-48.
39. Polikanov Y.S. и др. The Mechanisms of Action of Ribosome-Targeting Peptide Antibiotics // *Front. Mol. Biosci*, 2018. Т. 5. С. 48.
40. Polacek N., and Mankin A.S. The ribosomal peptidyl transferase center: structure, function, evolution, inhibition // *Crit. Rev. Biochem. Mol. Biol.*, 2005. Т. 40. С. 285–311.
41. Moore P., Steitz T. The structural basis of large ribosomal subunit function // *Annu. Rev. Biochem.*, 2003. Т. 72. С. 813–850.
42. Ramakrishnan V. Ribosome structure and the mechanism of translation // *Cell*, 2002. Т. 108. С. 557–572.
43. Dallas A., Noller H.F. Interaction of translation initiation factor 3 with the 30S ribosomal subunit // *Mol Cell*, 2001. Т. 8. С. 855 –864.
44. Wegmann U., Horn N., Carding S.R. Defining the bacteroides ribosomal binding site // *Appl. Environ. Microbiol.*, 2013. Т. 79. № 6. С. 1980-9.
45. Lin Y.H. и др. Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes // *Gene*, 2008. Т. 416. № 1-2. С. 44-7.
46. Nakagawa S. и др. Dynamic evolution of translation initiation mechanisms in prokaryotes // *Proc. Natl. Acad. Sci. USA*, 2010. Т. 107. № 14. С. 6382-7.
47. Ma J., Campbell A., Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures // *J. Bacteriol.*, 2002. Т. 184. № 20. С. 5733-45.
48. Duval M. и др. Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation // *PLoS Biol.*, 2013. Т. 11. № 12. С. e1001731.
49. Accetto T., Avguštin G. Inability of *Prevotella bryantii* to form a functional Shine-Dalgarno interaction reflects unique evolution of ribosome binding sites in Bacteroidetes // *PLoS One*, 2011. Т. 6. № 8. С. e22914.

50. Gardner P.P., Eldai H. Annotating RNA motifs in sequences and alignments // *Nucleic Acids Res.*, 2015. Т. 43. № 2. С. 691-8.
51. Chang B., Halgamuge S., Tang S.-L. Analysis of SD sequences in completed microbial genomes: non-SD-led genes are as common as SD-led genes // *Gene*, 2006. Т. 373. С. 90–99.
52. Scharff L.B. и др. Local Absence of Secondary Structure Permits Translation of mRNAs that Lack Ribosome-Binding Sites // *PLoS Genet.*, 2011. Т. 7. № 6. С. e1002155.
53. Zheng X. и др. Leaderless genes in bacteria: clue to the evolution of translation initiation mechanisms in prokaryotes // *BMC Genomics*, 2011. Т. 12. С. 361.
54. Arenz S., Wilson D.N. Bacterial Protein Synthesis as a Target for Antibiotic Inhibition // *Cold Spring Harb. Perspect. Med.*, 2016. Т. 6. № 9. С. a025361.
55. Simonetti A. и др. A structural view of translation initiation in bacteria // *Cell. Mol. Life Sci.*, 2009. Т. 66. № 3. С. 423-36.
56. Yusupov M.M. и др. Crystal structure of the ribosome at 5.5 Å resolution // *Science*, 2001. Т. 292. № 5518. С. 883-96.
57. Karimi R. и др. Novel roles for classical factors at the interface between translation termination and initiation // *Mol. Cell*, 1999. Т. 3. № 5. С. 601-9.
58. Julián P. и др. The Cryo-EM structure of a complete 30S translation initiation complex from *Escherichia coli* // *PLoS Biol.*, 2011. Т. 9. № 7. С. e1001095.
59. Marzi S. и др. Structured mRNAs regulate translation initiation by binding to the platform of the ribosome // *Cell*, 2007. Т. 130. № 6. С. 1019-31.
60. Marzi S. и др. RNA switches regulate initiation of translation in bacteria // *Biol. Chem.*, 2008. Т. 389. № 5. С. 585-98.
61. Carter A.P. и др. Crystal structure of an initiation factor bound to the 30S ribosomal subunit // *Science*, 2001. Т. 291. № 5503. С. 498-501.
62. Simonetti A. и др. Involvement of protein IF2 N domain in ribosomal subunit joining revealed from architecture and function of the full-length initiation factor // *Proc. Natl. Acad. Sci. USA*, 2013. Т. 110. № 39. С. 15656-61.
63. Boelens R., Gualerzi C.O. Structure and function of bacterial initiation factors // *Curr. Protein Pept. Sci.*, 2002. Т. 3. № 1. С. 107-19.
64. Rodnina M.V. Translation in Prokaryotes // *Cold Spring Harb. Perspect. Biol.*, 2018. Т. 10. № 9. С. a032664.
65. Kaminishi T. и др. A snapshot of the 30S ribosomal subunit capturing mRNA via the Shine-Dalgarno interaction // *Structure*, 2007. Т. 15. № 3. С. 289-97.
66. Yusupova G. и др. Structural basis for messenger RNA movement on the ribosome // *Nature*, 2006. Т. 444. № 7117. С. 391-4.

67. Hui A., de Boer H.A. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli* // *Proc. Natl. Acad. Sci. USA*, 1987. T. 84. № 14. С. 4762-6.
68. Jacob W.F., Santer M., Dahlberg A.E. A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins // *Proc. Natl. Acad. Sci. USA*, 1987. T. 84. № 14. С. 4757-61.
69. Calogero R.A. и др. Selection of the mRNA translation initiation region by *Escherichia coli* ribosomes // *Proc. Natl. Acad. Sci. USA*, 1988. T. 85. № 17. С. 6427-31.
70. Melançon P. и др. The anti-Shine-Dalgarno region in *Escherichia coli* 16S ribosomal RNA is not essential for the correct selection of translational starts // *Biochemistry*, 1990. T. 29. № 13. С. 3402-7.
71. Lee K., Holland-Staley C.A., Cunningham P.R. Genetic analysis of the Shine-Dalgarno interaction: selection of alternative functional mRNA-rRNA combinations // *RNA*, 1996. T. 2. № 12. С. 1270-85.
72. Skorski P. и др. The highly efficient translation initiation region from the *Escherichia coli* rpsA gene lacks a shine-dalgarno element // *J. Bacteriol.*, 2006. T. 188. № 17. С. 6277-85.
73. Studer S.M., Joseph S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex // *Mol. Cell.*, 2006. T. 22. № 1. С. 105-15.
74. Yusupova G.Z. и др. The path of messenger RNA through the ribosome // *Cell*, 2001. T. 106. № 2. С. 233-41.
75. Selmer M. и др. Structure of the 70S ribosome complexed with mRNA and tRNA // *Science*, 2006. T. 313. № 5795. С. 1935-42.
76. Ringquist S. и др. Nature of the ribosomal mRNA track: analysis of ribosome-binding sites containing different sequences and secondary structures // *Biochemistry*, 1993. T. 32. № 38. С. 10254-62.
77. de Smit M.H., van Duin J. Translational standby sites: how ribosomes may deal with the rapid folding kinetics of mRNA // *J. Mol. Biol.*, 2003. T. 331. № 4. С. 737-43.
78. Komarova A.V. и др. AU-rich sequences within 5' untranslated leaders enhance translation and stabilize mRNA in *Escherichia coli* // *J. Bacteriol.*, 2005. T. 187. № 4. С. 1344-9.
79. Schlax P.J., Worhunsky D.J. Translational repression mechanisms in prokaryotes // *Mol. Microbiol.*, 2003. T. 48. № 5. С. 1157-69.
80. Ehresmann C. и др. Molecular mimicry in translational regulation: the case of ribosomal protein S15 // *RNA Biol.*, 2004. T. 1. № 1. С. 66-73.
81. Lovmar M., Ehrenberg M. Rate, accuracy and cost of ribosomes in bacterial cells // *Biochimie*, 2006. T. 88. № 8. С. 951-61.

82. La Teana A., Gualerzi C.O., Brimacombe R. From stand-by to decoding site. Adjustment of the mRNA on the 30S ribosomal subunit under the influence of the initiation factors // *RNA*, 1995. T. 1. № 8. С. 772-82.
83. Steitz J.A., Jakes K. How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in *Escherichia coli* // *Proc. Natl. Acad. Sci. USA*, 1975. T. 72. № 12. С. 4734-8.
84. Sergiev P.V. и др. The path of mRNA through the bacterial ribosome: a site-directed crosslinking study using new photoreactive derivatives of guanosine and uridine // *RNA*, 1997. T. 3. № 5. С. 464-75.
85. Takyar S., Hickerson R.P., Noller H.F. mRNA helicase activity of the ribosome // *Cell*, 2005. T. 120. № 1. С. 49-58.
86. Götz F., Dabbs E.R., Gualerzi C.O. *Escherichia coli* 30S mutants lacking protein S20 are defective in translation initiation // *Biochim. Biophys. Acta*, 1990. T. 1050. № 1-3. С. 93-7.
87. Rydén-Aulin M. и др. Ribosome activity and modification of 16S RNA are influenced by deletion of ribosomal protein S20 // *Mol. Microbiol.*, 1993. T. 7. № 6. С. 983-92.
88. Nikolay R. и др. *Ribosomal Proteins: Role in Ribosomal Functions* // In eLS, John Wiley & Sons, Ltd (Ed.), 2015.
89. Ciccarelli F.D. и др. Toward automatic reconstruction of a highly resolved tree of life // *Science*, 2006. T. 311. № 5765. С. 1283-7.
90. Letunic I., Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation // *Bioinformatics*, 2007. T. 23. № 1. С. 127-8.
91. Letunic I., Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments // *Nucleic Acids Res.*, 2019. T. 47. № W1. С. W256-W259.
92. Letunic I., Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation // *Nucleic Acids Res.*, 2021. T. 49. № W1. С. W293-W296.
93. Ban N. и др. A new system for naming ribosomal proteins // *Curr. Opin. Struct. Biol.*, 2014. T. 24. С. 165-9.
94. Cannone J.J. и др. The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs // *BMC Bioinformatics*, 2002. T. 3. С. 2.
95. Mueller F., Brimacombe R. A new model for the three-dimensional folding of *Escherichia coli* 16 S ribosomal RNA. I. Fitting the RNA to a 3D electron microscopic map at 20 Å // *J. Mol. Biol.*, 1997. T. 271. № 4. С. 524-44.
96. Schluenzen F. и др. Structure of functionally activated small ribosomal subunit at 3.3 angstroms resolution // *Cell*, 2000. T. 102. № 5. С. 615-23.

97. Grigoriadou C. и др. A quantitative kinetic scheme for 70 S translation initiation complex formation // *J. Mol. Biol.*, 2007. Т. 373. № 3. С. 562-72.
98. Milon P. и др. Kinetic checkpoint at a late step in translation initiation // *Mol. Cell*, 2008. Т. 30. № 6. С. 712-20.
99. Sohmen D. и др. Enhanced SnapShot: Antibiotic inhibition of protein synthesis II // *Cell*, 2009. Т. 139. № 1. С. 212-212.e1.
100. Ogle J.M., Ramakrishnan V. Structural insights into translational fidelity // *Annu. Rev. Biochem.*, 2005. Т. 74. С. 129-77.
101. Wen J.D., Kuo S.T., Chou H.D. The diversity of Shine-Dalgarno sequences sheds light on the evolution of translation initiation // *RNA Biol.*, 2021. Т. 18. № 11. С. 1489-1500.
102. Tolstrup N. и др. Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus* // *Extremophiles*, 2000. Т. 4. №3. С. 175-9.
103. Weiner J. 3rd, Herrmann R., Browning G.F. Transcription in *Mycoplasma pneumoniae* // *Nucleic Acids Res.*, 2000. Т. 28. № 22. С. 4488-96.
104. Nakagawa S., Niimura Y., Gojobori T. Comparative genomic analysis of translation initiation mechanisms for genes lacking the Shine-Dalgarno sequence in prokaryotes // *Nucleic Acids Res.*, 2017. Т. 45. № 7. С. 3922-3931.
105. Saito K., Green R., Buskirk A.R. Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing // *Elife*, 2020. Т. 9. С. e55002.
106. Tzareva N.V., Makhno V.I., Boni I.V. Ribosome-messenger recognition in the absence of the Shine-Dalgarno interactions // *FEBS Lett.*, 1994. Т. 337. № 2. С. 189-94.
107. Hussain T. и др. Large-Scale Movements of IF3 and tRNA during Bacterial Translation Initiation // *Cell*, 2016. Т. 167. № 1. С. 133-144.e13.
108. Jäger D. и др. Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis* // *BMC Genomics*, 2014. Т. 15. № 1. С. 684.
109. Schmitt E. и др. Recent Advances in Archaeal Translation Initiation // *Front. Microbiol.*, 2020. Т. 11. С. 584152.
110. Kim D. и др. Comparative analysis of regulatory elements between *Escherichia coli* and *Klebsiella pneumoniae* by genome-wide transcription start site profiling // *PLoS Genet.*, 2012. Т. 8. № 8. С. e1002867.
111. Babski J. и др. Genome-wide identification of transcriptional start sites in the haloarchaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq) // *BMC Genomics*, 2016. Т. 17. № 1. С. 629.
112. Lomsadze A. и др. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes // *Genome Res.*, 2018. Т. 28. № 7. С. 1079-1089.

113. Thomason M.K. и др. Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in *Escherichia coli* // *J. Bacteriol.*, 2015. Т. 197. № 1. С. 18-28.
114. Tierrafría V.H. и др. RegulonDB 11.0: Comprehensive high-throughput datasets on transcriptional regulation in *Escherichia coli* K-12 // *Microb. Genom.*, 2022. Т. 8. № 5. С. mgen000833.
115. Beck H.J., Moll I. Leaderless mRNAs in the Spotlight: Ancient but Not Outdated // *Microbiol. Spectr.*, 2018. Т. 6 № 4. С. 6.4.02.
116. Moll I. и др. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control // *Mol. Microbiol.*, 2002. Т. 43 № 1. С. 239-46.
117. Van Etten W.J., Janssen G.R. An AUG initiation codon, not codon-anticodon complementarity, is required for the translation of unleadered mRNA in *Escherichia coli* // *Mol. Microbiol.*, 1998. Т. 27. № 5. С. 987-1001.
118. Kaberdina A.C. и др. An unexpected type of ribosomes induced by kasugamycin: a look into ancestral times of protein synthesis // *Mol. Cell*, 2009. Т. 33. № 2. С. 227-36.
119. Giliberti J. и др. A 5'-terminal phosphate is required for stable ternary complex formation and translation of leaderless mRNA in *Escherichia coli* // *RNA*, 2012. Т. 18. № 3. С. 508-18.
120. Brock J.E. и др. Ribosomes bind leaderless mRNA in *Escherichia coli* through recognition of their 5'-terminal AUG // *RNA*, 2008. Т. 14. № 10. С. 2159-69.
121. Jones R.L. 3rd, Jaskula J.C., Janssen G.R. In vivo translational start site selection on leaderless mRNA transcribed from the *Streptomyces fradiae* aph gene // *J. Bacteriol.*, 1992. Т. 174. № 14. С. 4753-60.
122. Udagawa T., Shimizu Y., Ueda T. Evidence for the translation initiation of leaderless mRNAs by the intact 70 S ribosome without its dissociation into subunits in eubacteria // *J. Biol. Chem.*, 2004. Т. 279. № 10. С. 8539-46.
123. Balakin A.G. и др. Unusual ribosome binding properties of mRNA encoding bacteriophage lambda repressor // *Nucleic Acids Res.*, 1992. Т. 20. № 3. С. 563-71.
124. O'Donnell S.M., Janssen G.R. Leaderless mRNAs bind 70S ribosomes more strongly than 30S ribosomal subunits in *Escherichia coli* // *J. Bacteriol.*, 2002. Т. 184. № 23. С. 6730-3.
125. Grill S. и др. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation // *EMBO J.*, 2000. Т. 19. № 15. С. 4101-10.
126. Yamamoto H. и др. 70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria // *Proc. Natl. Acad. Sci.*, 2016. Т. 113. № 9. С. E1180-9.
127. Goyal A., Belardinelli R., Rodnina M.V. Non-canonical Binding Site for Bacterial Initiation Factor 3 on the Large Ribosomal Subunit // *Cell. Rep.*, 2017. Т. 20. № 13. С. 3113-3122.

128. Krishnan K.M., Van Etten W.J. 3rd, Janssen G.R. Proximity of the start codon to a leaderless mRNA's 5' terminus is a strong positive determinant of ribosome binding and expression in *Escherichia coli* // *J. Bacteriol.*, 2010. Т. 192. № 24. С. 6482-5.
129. Vesper O. и др. Selective translation of leaderless mRNAs by specialized ribosomes generated by MazF in *Escherichia coli* // *Cell*, 2011. Т. 147. № 1. С. 147-57.
130. Grill S. и др. Temperature-dependent translation of leaderless and canonical mRNAs in *Escherichia coli* // *FEMS Microbiol. Lett.*, 2002. Т. 211. № 2. С. 161-7.
131. Brenneis M., Soppa J. Regulation of translation in haloarchaea: 5'- and 3'-UTRs are essential and have to functionally interact in vivo // *PLoS One*, 2009. Т. 4. № 2. С. e4484.
132. Kozak M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes // *Gene*, 2005. Т. 361. С. 13-37.
133. Shultzaberger R.K. и др. Anatomy of *Escherichia coli* ribosome binding sites // *J. Mol. Biol.*, 2001. Т. 313. № 1. С. 215-28.
134. Rudd K.E. EcoGene: a genome sequence database for *Escherichia coli* K-12 // *Nucleic Acids Res.*, 2000. Т. 28. № 1. С. 60-4.
135. Nakagawa S. и др. Dynamic evolution of translation initiation mechanisms in prokaryotes // *Proc. Natl. Acad. Sci. USA*, 2010. Т. 107. № 14. С. 6382-7.
136. Vimberg V. и др. Translation initiation region sequence preferences in *Escherichia coli* // *BMC Mol. Biol.*, 2007. Т. 8. С. 100.
137. Osterman I.A. и др. Comparison of mRNA features affecting translation initiation and reinitiation // *Nucleic Acids Res.*, 2013. Т. 41. № 1. С. 474-86.
138. Chen H. и др. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of *Escherichia coli* mRNAs // *Nucleic Acids Res.*, 1994. Т. 22. № 23. С. 4953-7.
139. Gu W., Zhou T., Wilke C.O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes // *PLoS Comput. Biol.*, 2010. Т. 6. № 2. С. e1000664.
140. Gingold H., Pilpel Y. Determinants of translation efficiency and accuracy // *Mol. Syst. Biol.*, 2011. Т. 7. С. 481.
141. de Smit M.H., van Duin J. Secondary structure of the ribosome binding site determines translational efficiency: a quantitative analysis // *Proc. Natl. Acad. Sci. USA*, 1990. Т. 87. № 19. С. 7668-72.
142. Studer S.M., Joseph S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex // *Mol. Cell*, 2006. Т. 22. № 1. С. 105-15.
143. Stenström C.M., Isaksson L.A. Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3' side // *Gene*, 2002. Т. 288. № 1-2. С. 1-8.

144. Gonzalez de Valdivia E.I., Isaksson L.A. A codon window in mRNA downstream of the initiation codon where NGG codons give strongly reduced gene expression in *Escherichia coli* // *Nucleic Acids Res.*, 2004. Т. 32. № 17. С. 5198-205.
145. Osterman I.A. и др. Translation at first sight: the influence of leading codons // *Nucleic Acids Res.*, 2020. Т. 48. № 12. С. 6931-6942.
146. Park Y.S. и др. Design of 5'-untranslated region variants for tunable expression in *Escherichia coli* // *Biochem. Biophys. Res. Commun.*, 2007. Т. 356. № 1. С. 136-41.
147. Barendt P.A. и др. Broad-specificity mRNA-rRNA complementarity in efficient protein translation // *PLoS Genet.*, 2012. Т. 8. № 3. С. e1002598.
148. Barendt P.A. и др. Evidence for context-dependent complementarity of non-Shine-Dalgarno ribosome binding sites to *Escherichia coli* rRNA // *ACS Chem. Biol.*, 2013. Т. 8. № 5. С. 958-66.
149. Salis H.M., Mirsky E.A., Voigt C.A. Automated design of synthetic ribosome binding sites to control protein expression // *Nat. Biotechnol.*, 2009. Т. 27. № 10. С. 946-50.
150. Salis H.M. The ribosome binding site calculator // *Meth. Enzym.*, 2011. Т. 498. С. 19-42.
151. Borujeni A.E., Channarasappa A.S., Salis H.M. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites // *Nucleic Acids Res.*, 2014. Т. 42. № 4. С. 2646-59.
152. Hofacker I.L. Vienna RNA secondary structure server // *Nucleic Acids Res.*, 2003. Т. 31. № 13. С. 3429-31.
153. Farasat I. и др. Efficient search, mapping, and optimization of multi-protein genetic systems in diverse bacteria // *Mol. Syst. Biol.*, 2014. Т. 10. № 6. С. 731.
154. Nakeff A. и др. Application of flow cytometry and cell sorting to megakaryocytopoiesis // *Blood*, 1979. Т. 53. № 4. С. 732-45.
155. Solieri L., Dakal T.C., Giudici P. Next-generation sequencing and its potential impact on food microbial genomics // *Ann. Microbiol.*, 2013. Т. 63. С. 21-37.
156. Osterman I.A. и др. Attenuation-based dual-fluorescent-protein reporter for screening translation inhibitors // *Antimicrob. Agents Chemother.*, 2012. Т. 56. № 4. С. 1774-83.
157. Lesnik E.A. и др. Identification of conserved regulatory RNA structures in prokaryotic metabolic pathway genes // *Biosystems*, 2005. Т. 80. № 2. С. 145-54.
158. Gould P.S., Bird H., Easton A.J. Translation toeprinting assays using fluorescently labeled primers and capillary electrophoresis // *Biotechniques*, 2005. Т. 38. № 3. С. 397-400.
159. Shirokikh N.E. и др. Quantitative analysis of ribosome-mRNA complexes at different translation stages // *Nucleic Acids Res.*, 2010. Т. 38. № 3. С. e15.
160. Osterman I.A. и др. Tetracenomycin X inhibits translation by binding within the ribosomal exit tunnel // *Nat. Chem. Biol.*, 2020. Т. 16. № 10. С. 1071-1077.

161. Machida R.J., Lin Y.Y. Four methods of preparing mRNA 5' end libraries using the Illumina sequencing platform // *PLoS One*, 2014. Т. 9. № 7. С. e101812.
162. Rothberg J.M. и др. An integrated semiconductor device enabling non-optical genome sequencing // *Nature*, 2011. Т. 475. № 7356. С. 348-52.
163. Liu L. и др. Comparison of next-generation sequencing systems // *J. Biomed. Biotechnol.*, 2012. Т. 2012. С. 251364.
164. Ingolia N.T. и др. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling // *Science*, 2009. Т. 324. № 5924. С. 218-23.
165. Andreev D.E. и др. Translation of 5' leaders is pervasive in genes resistant to eIF2 repression // *Elife*, 2015. Т. 4. С. e03971.
166. Andreev D.E. и др. Oxygen and glucose deprivation induces widespread alterations in mRNA translation within 20 minutes // *Genome Biol.*, 2015. Т. 16. № 1. С. 90.
167. Meydan S. и др. Retapamulin-Assisted Ribosome Profiling Reveals the Alternative Bacterial Proteome // *Mol. Cell*, 2019. Т. 74. № 3. С. 481-493.e6.
168. Brar G.A., Weissman J.S. Ribosome profiling reveals the what, when, where and how of protein synthesis // *Nat. Rev. Mol. Cell. Biol.*, 2015. Т. 16. № 11. С. 651-64.
169. Reid D.W., Shenolikar S., Nicchitta C.V. Simple and inexpensive ribosome profiling analysis of mRNA translation // *Methods*, 2015. Т. 91. С. 69-74.
170. Ingolia N.T., Hussmann J.A., Weissman J.S. Ribosome Profiling: Global Views of Translation // *Cold Spring Harb Perspect. Biol.*, 2019. Т. 11. № 5. С. a032698.
171. Weaver J. и др. Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes // *mBio.*, 2019. Т. 10. № 2. С. e02819-18.
172. Meydan S. и др. Identification of Translation Start Sites in Bacterial Genomes // *Methods Mol. Biol.*, 2021. Т. 2252. С. 27-55.
173. Vazquez-Laslop N. и др. Identifying Small Open Reading Frames in Prokaryotes with Ribosome Profiling // *J. Bacteriol.*, 2022. Т. 204. № 1. С. e0029421.
174. O'Connor P.B., Andreev D.E., Baranov P.V. Comparative survey of the relative impact of mRNA features on local ribosome profiling read density // *Nat. Commun.*, 2016. Т. 7. С. 12915.
175. Andreev D.E. и др. Insights into the mechanisms of eukaryotic translation gained with ribosome profiling // *Nucleic Acids Res.*, 2017. Т. 45. № 2. С. 513-526.
176. Glaub A. и др. Recommendations for bacterial ribosome profiling experiments based on bioinformatic evaluation of published data // *J. Biol. Chem.*, 2020. Т. 295. № 27. С. 8999-9011.
177. Gerashchenko M.V., Gladyshev V.N. Ribonuclease selection for ribosome profiling // *Nucleic Acids Res.*, 2017. Т. 45. № 2. С. e6.

178. Marks J. и др. Context-specific inhibition of translation by ribosomal antibiotics targeting the peptidyl transferase center // *Proc. Natl. Acad. Sci.*, 2016. Т. 113. № 43. С. 12150-12155.
179. Vázquez-Laslop N., Mankin A.S. Context-Specific Action of Ribosomal Antibiotics // *Annu. Rev. Microbiol.*, 2018. Т. 72. С. 185-207.
180. Svetlov M.S. и др. Context-specific action of macrolide antibiotics on the eukaryotic ribosome // *Nat. Commun.*, 2021. Т. 12. № 1. С. 2803.
181. Li G.W., Oh E., Weissman J.S. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria // *Nature*, 2012. Т. 484. № 7395. С. 538-41.
182. Mohammad F. и др. Clarifying the Translational Pausing Landscape in Bacteria by Ribosome Profiling // *Cell. Rep.*, 2016. Т. 14. № 4. С. 686-694.
183. Dittrich W.M., Göhde W.H. Flow-through chamber for photometers to measure and count particles in a dispersion medium // *United States Patent*, 1973. № 3.761.187.
184. Fulwyler M.J. Electronic separation of biological cells by volume // *Science*, 1965. Т. 150. № 3698. С. 910–1.
185. Fulwyler M.J. Particle Separator, US3380584A. 1965.
186. Moldavan A. Photo-Electric Technique for the Counting of microscopical cells // *Science*, 1934. Т. 80. № 2069. С. 188-189.
187. Coons A.H., Kaplan M.H. Localization of antigen in tissue cells // *J. Exp. Med.*, 1950. Т. 91. № 1. С. 1-13.
188. Crosland-Taylor P.J. A Device for Counting Small Particles suspended in a Fluid through a Tube // *Nature*, 1953. Т. 171. № 4340. С. 37-38.
189. Sweet R.G. High Frequency Recording with Electrostatically Deflected Ink Jets // *Review of Scientific Instruments*, 1965. Т. 36. № 2. С. 131-6.
190. Picot J. и др. Flow cytometry: retrospective, fundamentals and recent instrumentation // *Cytotechnology*, 2012. Т. 64. № 2. С. 109–30.
191. Shapiro H.M. Practical flow cytometry (4th ed.) // New York, Wiley-Liss, 2003. ISBN 978-0-471-41125-3.
192. Goddard G., Kaduchak G. Ultrasonic particle concentration in a line-driven cylindrical tube // *J. Acoust. Soc. Am.*, 2005. Т. 117. С. 3440–3447.
193. Goddard G. и др. Ultrasonic particle-concentration for sheathless focusing of particles for analysis in a flow cytometer // *Cytometry A.*, 2006. Т. 69. С. 66–74.
194. McKinnon K.M. Flow Cytometry: An Overview // *Curr. Protoc. Immunol.*, 2019. Т. 120. С. 5.1.1–5.1.11.
195. Kamensky L.A., Melamed M.R., Derman H. Spectrophotometer: new instrument for ultrarapid cell analysis // *Science*, 1965. Т. 150. № 3696. С. 630-1.

196. Bonner W.A. и др. Fluorescence activated cell sorting // *Rev. Sci. Instrum.*, 1972. Т. 43. № 3. С. 404-9.
197. Robinson J.P. Flow cytometry past and future // *Biotechniques*, 2022. Т. 72. № 4. С. 159-169.
198. Petersen T.W., Ger van den Engh. Stability of the breakoff point in a high-speed cell sorter // *Cytometry A.*, 2003. Т. 56. № 2. С. 63-70.
199. Cossarizza A. и др. Guidelines for the use of flow cytometry and cell sorting in immunological studies (2nd edition) // *Eur. J. Immunol.*, 2019. Т. 49. С. 1457–1973.
200. Nakeff A. и др. Application of flow cytometry and cell sorting to megakaryocytopoiesis // *Blood*, 1979. Т. 53. № 4. С. 732-45.
201. Claflin A.J. и др. Inhibition of growth and guanylate cyclase activity of an undifferentiated prostate adenocarcinoma by an extract of the balsam pear (*Momordica charantia abbreviata*) // *Proc. Natl. Acad. Sci. USA*, 1978. Т. 75. С. 989–993.
202. Traganos F. и др. Simultaneous Staining of Ribonucleic and Deoxyribonucleic Acids in Unfixed Cells Using Acridine Orange in a Flow Cytofluorometric System // *J. Histochem. Cytochem.*, 1977. Т. 25. С. 46–56.
203. Luo Y. и др. Side population cells from human melanoma tumors reveal diverse mechanisms for chemoresistance // *J. Investig. Dermatol.*, 2012. Т. 132. С. 2440–2450.
204. Piccoli S. и др. White Paper on Recent Issues in Bioanalysis: FDA Immunogenicity Guidance, Gene Therapy, Critical Reagents, Biomarkers and Flow Cytometry Validation (Part 3—Recommendations on 2019 FDA Immunogenicity Guidance, Gene Therapy Bioanalytical Challenges, Strategies for Critical Reagent Management, Biomarker Assay Validation, Flow Cytometry Validation & CLSI H62) // *Bioanalysis*, 2019. Т. 11. С. 2207–2244.
205. Kosuri S. и др. Composability of regulatory sequences controlling transcription and translation in *Escherichia coli* // *Proc. Natl. Acad. Sci. USA*, 2013. Т. 110. № 34. С. 14024-9.
206. Mutalik V.K. и др. Quantitative estimation of activity and quality for collections of functional genetic elements // *Nat. Methods*, 2013. Т. 10. № 4. С. 347-53.
207. Sanner M.F. Python: a programming language for software integration and development // *J. Mol. Graph. Model.*, 1999. Т. 17. № 1. С. 57-61.
208. Dessau R.B., Phipper C.B. ["R"--project for statistical computing] // *Ugeskr. Laeger*, 2008. Т. 170. № 5. С. 328-30.
209. Mirzadeh K. и др. Enhanced Protein Production in *Escherichia coli* by Optimization of Cloning Scars at the Vector-Coding Sequence Junction // *ACS Synth. Biol.*, 2015. Т. 4. № 9. С. 959-65.
210. Sauer C. и др. Exploring the Nonconserved Sequence Space of Synthetic Expression Modules in *Bacillus subtilis* // *ACS Synth. Biol.*, 2018. Т. 7. № 7. С. 1773-1784.
211. Klausen M.S., Sommer M.O.A. Parts Characterization for Tunable Protein Expression // *Methods Mol. Biol.*, 2018. Т. 1671. С. 3-14.

212. Duan Y. и др. Deciphering the Rules of Ribosome Binding Site Differentiation in Context Dependence // *ACS Synth. Biol.*, 2022. Т. 11. № 8. С. 2726-2740.
213. Kuo S.T. и др. Global fitness landscapes of the Shine-Dalgarno sequence // *Genome Res.*, 2020. Т. 30. № 5. С. 711-723.
214. Mutalik V.K. и др. Precise and reliable gene expression via standard transcription and translation initiation elements // *Nat. Methods*, 2013. Т. 10. № 4. С. 354-60.
215. Bonde M.T. и др. Predictable tuning of protein expression in bacteria // *Nat. Methods*, 2016. Т. 13. № 3. С. 233-6.
216. Goodman D.B., Church G.M., Kosuri S. Causes and effects of N-terminal codon bias in bacterial genes // *Science*, 2013. Т. 342. № 6157. С. 475-9.
217. Tuller T. и др. An evolutionarily conserved mechanism for controlling the efficiency of protein translation // *Cell*, 2010. Т. 141. № 2. С. 344-54.
218. Allert M., Cox J.C., Hellinga H.W. Multifactorial determinants of protein expression in prokaryotic open reading frames // *J. Mol. Biol.*, 2010. Т. 402. № 5. С. 905-18.
219. Pechmann S., Frydman J. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding // *Nat. Struct. Mol. Biol.*, 2013. Т. 20. № 2. С. 237-43.
220. Bentele K. и др. Efficient translation initiation dictates codon usage at gene start // *Mol. Syst. Biol.*, 2013. Т. 9. С. 675.
221. dos Reis M., Savva R., Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection // *Nucleic Acids Res.*, 2004. Т. 32. № 17. С. 5036-44.
222. Shah P. и др. Rate-limiting steps in yeast protein translation // *Cell*, 2013. Т. 153. № 7. С. 1589-601.
223. Kudla G. и др. Coding-sequence determinants of gene expression in *Escherichia coli* // *Science*, 2009. Т. 324. № 5924. С. 255-8.
224. Welch M. и др. Design parameters to control synthetic gene expression in *Escherichia coli* // *PLoS One*, 2009. Т. 4. № 9. С. e7002.
225. Zhou M. и др. Non-optimal codon usage affects expression, structure and function of clock protein FRQ // *Nature*, 2013. Т. 495. № 7439. С. 111-5.
226. Navon S., Pilpel Y. The role of codon selection in regulation of translation efficiency deduced from synthetic libraries // *Genome Biol.*, 2011. Т. 12. № 2. С. R12.
227. Tuller T. и др. Translation efficiency is determined by both codon bias and folding energy // *Proc. Natl. Acad. Sci. USA*, 2010. Т. 107. № 8. С. 3645-50.
228. Subramaniam A.R., Pan T., Cluzel P. Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria // *Proc. Natl. Acad. Sci. USA*, 2013. Т. 110. № 6. С. 2419-24.

229. Zadeh J.N. и др. NUPACK: Analysis and design of nucleic acid systems // *J. Comput. Chem.*, 2011. Т. 32. № 1. С. 170-3.
230. Lehning C.E. и др. Assessing glycolytic flux alterations resulting from genetic perturbations in *E. coli* using a biosensor // *Metab. Eng.*, 2017. Т. 42. С. 194-202.
231. Zhai W. и др. Sequence and thermodynamic characteristics of terminators revealed by FlowSeq and the discrimination of terminators strength // *Synth. Syst. Biotechnol.*, 2022. Т. 7. № 4. С. 1046-1055.
232. Glanville D.G. и др. A High-Throughput Method for Identifying Novel Genes That Influence Metabolic Pathways Reveals New Iron and Heme Regulation in *Pseudomonas aeruginosa* // *mSystems*, 2021. Т. 6. № 1. С. e00933-20.
233. Cheung R. и др. A Multiplexed Assay for Exon Recognition Reveals that an Unappreciated Fraction of Rare Genetic Variants Cause Large-Effect Splicing Disruptions // *Mol. Cell*, 2019. Т. 73. № 1. С. 183-194.e8.
234. Andrianantoandro E. и др. Synthetic biology: new engineering rules for an emerging discipline // *Mol. Syst. Biol.*, 2006. Т. 2. С. 2006.0028.
235. Chappell J., Jensen K., Freemont P.S. Validation of an entirely in vitro approach for rapid prototyping of DNA regulatory elements for synthetic biology // *Nucleic Acids Res.*, 2013. Т. 41. № 5. С. 3471-81.
236. Zhang L. и др. Development and comparison of cell-free protein synthesis systems derived from typical bacterial chassis // *Bioresour. Bioprocess*, 2021. Т. 8. № 1. С. 58.
237. Seo S.W. и др. Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency // *Metab. Eng.*, 2013. Т. 15. С. 67-74.
238. Seo S.W. и др. Predictive combinatorial design of mRNA translation initiation regions for systematic optimization of gene expression levels // *Sci. Rep.*, 2014. Т. 4. С. 4515.
239. Zhu H., Wang Q. Prediction of Translation Initiation Site in Bacterial and Archaeal Genomes. *Current Bioinformatics* // *Curr. Bioinformat.*, 2014. Т. 9. С. 155–165
240. Laserson U., Gan H.H., Schlick T. Predicting candidate genomic sequences that correspond to synthetic functional RNA motifs // *Nucleic Acids Res.*, 2005. Т. 33. № 18. С. 6057-69.
241. Oliphant A.R., Nussbaum A.L., Struhl K. Cloning of random-sequence oligodeoxynucleotides // *Gene*, 1986. Т. 44. № 2-3. С. 177-83.
242. Holmqvist E., Reimegård J., Wagner E.G. Massive functional mapping of a 5'-UTR by saturation mutagenesis, phenotypic sorting and deep sequencing // *Nucleic Acids Res.*, 2013. Т. 41. № 12. С. e122.
243. Noderer W.L. и др. Quantitative analysis of mammalian translation initiation sites by FACS-seq // *Mol. Syst. Biol.*, 2014. Т. 10. № 8. С. 748.

244. Dvir S. и др. Deciphering the rules by which 5'-UTR sequences affect protein expression in yeast // *Proc. Natl. Acad. Sci. USA*, 2013. Т. 110. № 30. С. E2792-801.
245. Jenner L.B. и др. Structural aspects of messenger RNA reading frame maintenance by the ribosome // *Nat. Struct. Mol. Biol.*, 2010. Т. 17. № 5. С. 555-60.
246. Zhou J. и др. Crystal structures of EF-G-ribosome complexes trapped in intermediate states of translocation // *Science*, 2013. Т. 340. № 6140. С. 1236086.
247. Nieuwkoop T., Claassens N.J., van der Oost J. Improved protein production and codon optimization analyses in *Escherichia coli* by bicistronic design // *Microb. Biotechnol.*, 2019. Т. 12. № 1. С. 173-179.
248. Worthington M.T., Pelo J., Luo R.Q. Cloning of random oligonucleotides to create single-insert plasmid libraries // *Anal. Biochem.*, 2001. Т. 294. № 2. С. 169-75.
249. Blattner F.R. и др. The complete genome sequence of *Escherichia coli* K-12 // *Science*, 1997. Т. 277. № 5331. С. 1453-62.
250. Jacobs G.H. и др. Transterm: a database to aid the analysis of regulatory sequences in mRNAs // *Nucleic Acids Res.*, 2009. Т. 37(Database issue). С. D72-6.
251. Elgamal S. и др. EF-P dependent pauses integrate proximal and distal signals during translation // *PLoS Genet.*, 2014. Т. 10. № 8. С. e1004553.
252. Kannan K. и др. The general mode of translation inhibition by macrolide antibiotics // *Proc. Natl. Acad. Sci. USA*, 2014. Т. 111. № 45. С. 15958-63.
253. Li G.W. и др. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources // *Cell*, 2014. Т. 157. № 3. С. 624-35.
254. Datsenko K.A. and B.L. Wanner. One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products // *Proc. Natl. Acad. Sci. USA*, 2000. Т. 97. № 12. С. 6640-5.
255. Sambrook J., Russell D.W. *Molecular Cloning: A Laboratory Manual* // Cold Spring Harbor Laboratory Press, 2001. Т. 87.
256. Mullis K.B. and Faloona F.A. Specific synthesis of DNA in vitro via a polymerasecatalyzed chain reaction // *Methods Enzymol.*, 1987. Т. 155. С. 335-50.
257. Inoue H., Nojima H., Okayama H. High efficiency transformation of *Escherichia coli* with plasmids // *Gene*, 1990. Т. 96. № 1. С. 23-8.
258. Lorenz R. и др. ViennaRNA Package 2.0 // *Algorithms Mol. Biol.*, 2011. Т. 6. С. 26.
259. Martin M. Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads // *EMBnet J.*, 2011. Т. 17. С. 10-12.
260. Edgar R.C., Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads // *Bioinformatics*, 2015. Т. 31. № 21. С. 3476-82.
261. Zuker M., Stiegler P. Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information // *Nucleic Acids Res.*, 1981. Т. 9. С. 133-148.

# ПРИЛОЖЕНИЕ 1

Таблица. Применение метода Flow-seq для анализа эффективности трансляции.

Изучаемые элементы мРНК	Количество вариантов в полученных библиотеках в итоге после Flow-seq	Типы вариантов	Результаты	Ссылка
<b>Промоторы и сайты посадки рибосомы (RBS) в 5'-UTR</b>	11894 (94%) из 12653 возможных с комбинациями 114 промоторов и 111 RBS (одна комбинация привела к несовместимому сайту рестрикции)	Взятые из доступных баз данных и сгенерированные с помощью RBS Calculator	Диапазон изменения экспрессии $10^4$ раз. Выбор промотора оказывает наибольшее влияние на уровень РНК и в меньшей степени на уровень белка, так как на эффективность его трансляции влияет и выбор сайта посадки рибосомы, и, возможно, другие элементы. 55% из нескольких сотен проверенных отдельных колоний безошибочно определены при Flow-seq анализе	[205]
<b>Промоторы и сайты посадки рибосомы (RBS) в 5'-UTR</b>	~ 500 комбинаций из 14 промоторов и 22 RBS для двух детектируемых флуоресцентных белков и более 1200 комбинаций из рандомизированной библиотеки	Конкретные варианты и с участками рандомизации в изучаемых элементах	Динамический диапазон экспрессии $10^3$ раз. Получаемые комбинации приводят к экспрессии произвольного гена в пределах 2-х кратного изменения уровня с надёжностью 93%	[214]
<b>6 нт в спейсерном участке за SD в 5'-UTR и перед старт-кодоном и первые 6 нт, следующих за ним, (кодоны в +2 и +3 положениях кодирующей области (CDS)) ...SD-GAC-6N-AUG-6N<sup>syn</sup>- ...</b>	13914 (56%) для одного белка и 25861 (53%) для другого из 24576 и 49152 возможных соответственно	Рандомизированные спейсерные участки и кодоны в +2 и +3 положениях с синонимическими заменами, не изменяющими кодирующую последовательность двух сенсорных белков	Диапазон изменения экспрессии $10^3$ раз. Низкое содержание GC-нуклеотидов и уменьшение стабильности вторичной структуры изучаемых элементов важны для высокого уровня экспрессии, не ограничивающегося данными детерминантами. Распределение уровней флуоресценции белка, измеренных для нескольких десятков отдельных колоний с помощью планшетного ридера, совпадает с результатами Flow-seq	[209]
<b>4 нт в спейсерном участке за SD в 5'-UTR и перед старт-кодоном ...SD-C-4N-CAU-AUG-...</b>	249 (97%) из 256 возможных	Рандомизированные	Диапазон изменения экспрессии $10^2$ раз. Преобладание доли аденозинов и уменьшение доли цитидинов наблюдалось в эффективно транскрибирующихся вариантах. Низкое содержание GC-нуклеотидов и уменьшение стабильности вторичной структуры изучаемых элементов важны для высокого уровня экспрессии. SD-подобные последовательности также встречаются только в вариантах с высокой экспрессией	[18]
<b>SD длиной 6 нт в 5'-UTR</b>	4066 (99%) из 4096 возможных	Рандомизированные	Измеренные уровни белков (флуоресцентного и 5 природных) для 91% вариантов последовательностей были в пределах 2-х кратного изменения предсказанного уровня экспрессии в ЕМОРЕС, учитывающего контекстное окружение SD, сводящего к минимуму изменения вторичной структуры	[215]
<b>SD длиной 9 нт в 5'-UTR 3-х разных мРНК</b>	244401-257565 (99,641-99,995%) из 262144 возможных для 3-х разных мРНК	Рандомизированные	Диапазон изменения экспрессии $10^3$ раз. Увеличение доли нуклеотидов А и G и уменьшение доли нуклеотидов С и U наблюдалось в эффективно транскрибирующихся вариантах. Сила образования дуплекса SD с анти-SD значительно влияла на эффективность трансляции в отличие от расположения образования дуплекса и энергии укладки вторичной структуры мРНК. Сила взаимодействия SD с анти-SD была выше в вариантах мРНК, обеспечивающих высокую экспрессию	[213]
<b>Standby-сайты различной длины (20-164 нт) перед SD, дистальные в 5'-UTR</b>	136 5'-UTR с разной длиной и вторичной структурой, формой и числом модулей	Смоделированные варианты	Диапазон изменения эффективности трансляции $10^2$ раз. Скорость инициации трансляции мРНК контролируется площадью поверхности одноцепочечных участков, частичным разворачиванием структуры РНК для минимизации штрафа по свободной энергии связывания рибосомы, отсутствует кооперативное связывания, и, возможно, скольжение рибосомы в рассматриваемом участке. Разработана и проверена экспериментально биофизическая модель для предсказания скорости инициации трансляции. Рибосома может легко связываться с отдалёнными от старт-кодона модулями standby-сайтов и обеспечивать высокую эффективность трансляции.	[151]
<b>Сайт посадки рибосомы (RBS) в 5'-UTR с фиксированной последовательностью SD</b>	Более 20000 (10%) из ~ 200000 возможных вариантов для двух флуоресцентных белков	Рандомизированные и частично конкретные позиции с неполными вариациями	Диапазон изменения эффективности трансляции $10^4$ раз. На эффективность трансляции значительно влияет консервативность SD, тогда как AC-богатый спейсерный участок слабо зависит от контекста. Для высокой	[212]

(5 нт) и изменяющимися standby-сайтом (4 нт) и спейсерным участком (6 нт) RRRV-AGGAG-R-6N-AUG (R: A/G, V: A/G/C, N:A/U/C/G)			экспрессии наблюдалась низкая стабильность вторичной структуры изучаемого участка. Замена репортёрного белка на другой зачастую не влияла на общую тенденцию распределения последовательностей, задающих тот или иной уровень биосинтеза белка	
Почти полная последовательность 5'-UTR длиной 22 или 32 нт GG-20N/30N-AUG...	11692 (10 <sup>-6%</sup> из возможных вариантов), 11889 (10 <sup>-12%</sup> ) для 20N и 30N соответственно, 48 вариантов природных с вариациями	Рандомизированные, природные, конкретные	Диапазон изменения эффективности трансляции 10 <sup>4</sup> раз. Наблюдались низкая стабильность вторичной структуры и консервативность последовательности SD у вариантов с высокой экспрессией. Наличие AU-богатых энхансеров на 5'-конце в standby-сайте, низкая доля цитидиновых оснований, множественные SD, AG-повторы в 5'-UTR мРНК в отдельных случаях обеспечивают высокую эффективность трансляции	[19]
Последовательности 5'-UTR длиной 2-60 нт первых генов оперонов <i>E. coli</i> с GG на самом 5'-конце, сохранившихся при транскрипции GG-natural 5'-UTR	648 (91%) из 713 возможных с длиной 2-60 нт, (45%) из всех 1451 природных 5'-UTR первых генов оперонов	Природные	Диапазон изменения эффективности трансляции 30 раз. Наблюдалось влияние вторичной структуры РНК и последовательности SD на эффективность трансляции, но с меньшей вариабельностью по сравнению с рандомизированными библиотеками. При этом наблюдались низкая стабильность вторичной структуры и консервативность последовательности SD для вариантов с высокой экспрессией. Результаты оценки эффективности трансляции в случае с теми или иными 5'-UTR коррелировали с результатами, полученными с помощью Ribo-seq	[17]
Участки в промоторной области, standby-сайт длиной 10/20/30 нт, спейсерный участок длиной 8 нт 10N/20N/30N-SD-8N	~ 12000 (очень маленький процент из возможных)	Рандомизированные	Диапазон изменения эффективности трансляции 10 <sup>5</sup> раз. При высоком уровне экспрессии наблюдалась низкая стабильность вторичной структуры изучаемого участка	[210]
Промоторы, сайты посадки рибосомы (RBSs), первые 13 аминокислот кодирующей области белка	14234 комбинации из 2-х промоторов, 4-х сайтов посадки рибосомы (RBS) и последовательностей N-концевых пептидов, соответствующих первым 13-ти аминокислотам 137-ми природных генов <i>E. coli</i>	Природные	Диапазон изменения эффективности трансляции более 10 <sup>2</sup> раз. Использование редких кодонов на N-конце может увеличивать экспрессию в 14 раз независимо от RBS, обеспечивающих ту или иную эффективность трансляции. Уменьшение стабильности вторичной структуры, а не сама редкость кодонов отвечает за увеличение экспрессии	[216]
Первые 6 кодонов, следующих за старт-кодоном, в кодирующей области	10	Природные	Уменьшение стабильности вторичной структуры, а не сама редкость кодонов отвечает за увеличение эффективности трансляции. Редкие кодоны часто A/T-богаты в третьем положении, что чаще коррелирует с увеличением экспрессии, чем синонимичные кодоны, оканчивающиеся на G/C	[219]
Первые 10 кодонов, следующих за старт-кодоном, в кодирующей области	Более 30000	Рандомизированные	Уменьшение стабильности вторичной структуры, а не сама редкость кодонов отвечает за увеличение эффективности трансляции. Сильное влияние на экспрессию оказывают кодоны ближе к инициаторному. Дополнительные старт-кодоны в рамках считывания способствуют уровню трансляции. Наличие в N-концевой области белка аминокислот, на синтез которых клетка тратит много ресурсов, отрицательно сказывалось на эффективности биосинтеза таких белков	[145]

## ПРИЛОЖЕНИЕ 2

Таблица 1. Свойства вариантов 5'-UTR с 4N-участком рандомизации в спейсерной области библиотеки мРНК.

№	Вариабельный спейсерный участок	Число ридов во фракциях с различной эффективностью трансляции						Предсказанная эффективность трансляции*	Группа эффективности трансляции**	MFE*** (ккал/моль)	Энергия связывания 16S рРНК**** (ккал/моль)	Средняя фракция*****
		1	2	3	4	5	6					
1	AAAA	19	12	9	114	1781	3293	54.24	высокая	-12.3	0	5.58
2	AAAC	12	13	7	72	717	442	54.24	высокая	-12.3	0	5.21
3	AAAG	15	3	6	69	92	103	13.26	высокая	-13.8	0	4.84
4	AAAU	4	7	5	59	858	859	19.79	высокая	-14.3	0	5.42
5	AACA	8	9	40	151	180	61	37.33	высокая	-13.3	0	4.49
6	AACC	2	18	12	81	112	26	38.87	высокая	-13.2	0	4.44
7	AACG	4	3	22	27	16	3	12.68	высокая	-15.7	0	3.76
8	AACU	5	1	106	83	16	2	4.48	высокая	-18	0	3.52
9	AAGA	6	0	5	57	303	30	54.24	высокая	-12.3	0	4.85
10	AAGC	10	0	36	21	1	6	27.12	высокая	-14.6	0	3.28
11	AAGG	0	0	0	1	1	2	21.75	высокая	-13.8	-0.3	5.25
12	AAGU	6	4	29	147	23	24	17.85	высокая	-14.3	0	4.07
13	AAUA	5	4	9	62	585	214	54.24	высокая	-12.3	0	5.12
14	AAUC	3	20	45	104	94	79	20.8	высокая	-13.5	0	4.46
15	AAUG	4	17	21	50	35	0	21.75	высокая	-14.2	0	3.75
16	AAUU	12	15	14	78	302	24	17.29	высокая	-14.5	0	4.61
17	ACAA	38	30	56	465	665	61	54.24	высокая	-12.3	0	4.42
18	ACAC	12	14	52	182	40	11	54.24	высокая	-13.5	0	3.83
19	ACAG	13	10	93	53	21	23	13.26	высокая	-14.2	0	3.60
20	ACAU	18	67	92	216	275	21	19.79	высокая	-14.3	0	4.05
21	ACCA	8	62	117	109	39	10	39.05	высокая	-13.2	0	3.40
22	ACCC	13	27	98	59	18	12	40.66	высокая	-13.1	0	3.34
23	ACCG	23	80	39	18	9	2	12.68	низкая	-15.7	0	2.51
24	ACCU	22	53	199	32	26	15	4.69	высокая	-17.9	0	3.09
25	ACGA	1	27	143	124	39	10	24.23	высокая	-14.3	0	3.59
26	ACGC	104	99	25	32	7	10	5.37	низкая	-17.6	0	2.17
27	ACGG	8	27	47	28	6	5	21.75	высокая	-15.2	0	3.10
28	ACGU	69	136	94	18	22	16	6.75	низкая	-16.6	0	2.54
29	ACUA	0	6	29	213	55	14	32.62	высокая	-13.2	0	4.13
30	ACUC	20	87	93	17	8	4	11.58	низкая	-16.2	0	2.64
31	ACUG	1	27	122	39	5	1	21.75	высокая	-14.2	0	3.12
32	ACUU	9	9	136	68	27	2	25.93	высокая	-13.9	0	3.40
33	AGAA	10	6	9	118	254	57	54.24	высокая	-12.3	0	4.70
34	AGAC	20	4	11	80	19	10	50.7	высокая	-12.5	0	3.72
35	AGAG	2	1	5	26	38	0	13.26	высокая	-13.8	0	4.35
36	AGAU	9	6	8	224	135	17	19.79	высокая	-14.3	0	4.31
37	AGCA	28	90	105	70	49	16	13.26	высокая	-15.5	0	3.20
38	AGCC	44	38	19	8	3	1	13.81	низкая	-15.4	0	2.04
39	AGCG	35	57	72	45	15	7	9.68	низкая	-16.3	0	2.87
40	AGCU	110	112	91	36	7	3	1.59	низкая	-20.2	0	2.24
41	AGGA	9	20	35	66	7	10	79.55	высокая	-12.7	-1.5	3.49
42	AGGC	68	61	36	6	1	3	29.69	низкая	-14.6	-0.3	1.97
43	AGGG	1	8	7	5	2	1	35.71	высокая	-13.8	-0.6	3.08
44	AGGU	10	71	89	17	5	0	42.56	низкая	-16.1	-0.4	2.67
45	AGUA	11	28	103	57	20	13	29.67	высокая	-13.6	0	3.37
46	AGUC	39	30	16	13	5	1	4.46	низкая	-15.8	0	2.21
47	AGUG	11	68	11	6	1	0	21.75	низкая	-14.2	0	2.15
48	AGUU	17	60	82	27	9	16	4.69	низкая	-17.8	0	3.00
49	AUAA	26	5	23	35	1009	2209	54.24	высокая	-12.3	0	5.61
50	AUAC	9	7	11	33	383	387	54.24	высокая	-12.7	0	5.33
51	AUAG	0	20	8	42	22	18	13.26	высокая	-14.8	0	4.09
52	AUAU	17	0	2	50	488	1133	19.79	высокая	-13.3	0	5.60
53	AUCA	11	18	34	255	162	49	23.8	высокая	-13.3	0	4.30
54	AUCC	21	30	30	137	27	4	22.75	высокая	-13	0	3.53
55	AUCG	1	6	31	41	6	13	12.68	высокая	-15.7	0	3.86
56	AUCU	30	71	23	11	8	20	2.62	низкая	-17.2	0	2.73
57	AUGA	0	2	10	84	174	37	54.24	высокая	-12.3	0	4.76
58	AUGC	9	20	15	7	9	12	14.44	высокая	-15.4	0	3.32
59	AUGU	22	6	43	135	57	11	14.51	высокая	-14.3	0	3.85
60	AUUA	3	7	11	91	333	78	53.51	высокая	-12.3	0	4.87
61	AUUC	10	24	23	118	58	16	29.54	высокая	-13.9	0	3.96
62	AUUG	8	4	21	101	27	10	21.75	высокая	-14.2	0	3.96
63	AUUU	6	21	63	181	31	17	8.45	высокая	-14.6	0	3.82
64	CAAA	36	217	277	190	1202	1440	12.51	высокая	-15.8	0	4.97
65	CAAC	26	107	166	142	160	35	12.51	высокая	-15.8	0	3.64
66	CAAG	26	57	104	83	62	14	8.45	высокая	-16.6	0	3.40
67	CAAU	61	175	243	145	330	128	5.87	высокая	-17.4	0	3.82
68	CACA	99	205	294	245	174	24	9.68	высокая	-16.3	0	3.25
69	CACC	86	139	180	75	21	11	10.08	низкая	-16.2	0	2.69
70	CACG	68	212	120	70	19	6	7.06	низкая	-17.1	0	2.55
71	CACU	243	115	217	99	22	4	1.16	низкая	-21	0	2.36
72	CAGA	27	76	64	177	113	29	12.51	высокая	-15.8	0	3.74
73	CAGC	78	225	53	31	7	2	5.37	низкая	-17.6	0	2.17
74	CAGG	43	77	69	41	9	1	8.45	низкая	-16.6	0	2.58
75	CAGU	49	169	109	54	8	11	6.06	низкая	-17.9	0	2.59
76	CAUA	37	71	90	124	526	865	12.51	высокая	-15.8	0	5.12
77	CAUC	76	166	121	133	91	16	6.17	низкая	-16.6	0	3.07
78	CAUG	19	47	52	70	58	13	8.45	высокая	-16.6	0	3.54
79	CAUU	24	211	216	176	105	36	4.48	высокая	-18.6	0	3.31
80	CAAA	76	342	369	98	43	19	14.77	низкая	-15.4	0	2.73
81	CCAC	131	236	221	34	22	7	14.51	низкая	-15.4	0	2.39
82	CCAG	43	198	177	43	9	9	8.45	низкая	-16.6	0	2.59
83	CCAU	84	204	207	54	31	5	7.39	низкая	-16.9	0	2.59
84	CCCA	96	204	165	41	23	4	14.51	низкая	-15.4	0	2.44
85	CCCC	166	295	129	13	33	9	13.87	низкая	-15.4	0	2.19
86	CCCG	170	213	60	40	12	12	5.9	низкая	-17.4	0	2.11

87	CCCU	246	212	82	21	8	1	1.6	низкая	-20.1	0	1.84
88	CCGA	150	376	195	71	10	22	10.49	низкая	-16.2	0	2.37
89	CCGC	602	210	80	25	8	6	4.88	низкая	-17.8	0	1.54
90	CCGG	128	89	29	7	1	3	8.45	низкая	-16.6	0	1.73
91	CCGU	488	154	89	28	15	15	2.46	низкая	-19.4	0	1.70
92	CCUA	37	148	304	80	17	4	5.15	низкая	-17.1	0	2.84
93	CCUC	286	186	100	40	29	5	1.6	низкая	-20.1	0	2.00
94	CCUG	129	223	170	29	11	14	7.73	низкая	-17.6	0	2.33
95	CCUU	199	265	122	45	8	18	4.3	низкая	-17.9	0	2.17
96	CGAA	48	200	92	102	517	803	3.69	высокая	-18.4	0	4.84
97	CGAC	119	165	128	94	137	19	3.27	низкая	-18.7	0	3.03
98	CGAG	8	65	54	47	43	94	1.99	высокая	-19.8	0	4.07
99	CGAU	280	286	125	111	207	141	1.53	высокая	-19.7	0	3.09
100	CGCA	169	159	228	145	70	30	0.85	низкая	-21.7	0	2.85
101	CGCC	300	188	58	34	13	3	0.89	низкая	-21.6	0	1.79
102	CGCG	111	85	35	22	2	3	0.62	низкая	-22.5	0	1.95
103	CGCU	474	273	150	79	25	4	0.1	низкая	-26.4	0	1.93
104	CGGA	212	176	64	107	91	20	4.28	низкая	-18.9	-0.5	2.63
105	CGGC	337	122	92	22	7	2	4.28	низкая	-19	0	1.70
106	CGGG	27	18	36	24	2	4	3.27	низкая	-18.9	0	2.71
107	CGGU	292	263	92	35	8	7	0.77	низкая	-22.3	0	1.89
108	CGUA	146	84	130	175	188	41	1.91	высокая	-19.8	0	3.39
109	CGUC	278	184	91	45	14	8	1.45	низкая	-20	0	1.96
110	CGUG	205	130	55	15	15	3	1.46	низкая	-20.2	0	1.85
111	CGUU	229	158	132	93	35	18	0.3	низкая	-24	0	2.40
112	CUAA	22	53	457	553	290	60	13.08	высокая	-15.7	0	3.85
113	CUAC	30	96	301	137	39	7	13.08	высокая	-15.7	0	3.13
114	CUAG	30	144	222	111	11	9	8.45	низкая	-16.6	0	2.92
115	CUAU	44	95	337	244	81	25	8.05	высокая	-17.1	0	3.36
116	CUCA	224	516	413	131	56	9	11.58	низкая	-15.9	0	2.49
117	CUCC	268	257	135	72	19	0	8.38	низкая	-16.6	0	2.09
118	CUCG	226	255	86	9	20	8	4.63	низкая	-17.2	0	1.95
119	CUCU	600	340	164	29	12	8	1.39	низкая	-20.6	0	1.73
120	CUGA	28	192	263	130	39	18	13.08	низкая	-15.7	0	3.02
121	CUGC	299	273	115	36	9	11	5.37	низкая	-17.6	0	1.94
122	CUGG	86	197	94	17	16	16	8.45	низкая	-16.6	0	2.36
123	CUGU	221	425	221	57	32	32	4.84	низкая	-18.2	0	2.34
124	CUUA	32	91	412	351	132	27	13.08	высокая	-15.7	0	3.52
125	CUUC	154	243	232	36	23	4	5.39	низкая	-17.4	0	2.34
126	CUUG	55	195	201	90	17	2	8.45	низкая	-16.6	0	2.69
127	CUUU	391	346	351	80	38	21	4.48	низкая	-18.1	0	2.26
128	GAAA	2	3	3	41	857	2148	54.73	высокая	-13.3	0	5.68
129	GAAC	9	21	15	73	188	25	50.7	высокая	-13.3	0	4.47
130	GAAG	0	0	5	10	25	11	21.75	высокая	-13.8	0	4.82
131	GAAU	1	4	7	38	515	673	27.24	высокая	-14.5	0	5.49
132	GACA	13	14	24	55	412	63	54.73	высокая	-13.3	0	4.77
133	GACC	13	19	29	137	48	9	50.7	высокая	-13.3	0	3.84
134	GACG	1	1	35	65	15	0	12.68	высокая	-15.7	0	3.79
135	GACU	12	20	125	30	25	7	9.21	высокая	-17.7	0	3.26
136	GAGA	3	3	1	14	122	417	54.73	высокая	-13.3	-0.2	5.68
137	GAGC	18	23	57	27	6	4	18.09	низкая	-14.6	0	2.94
138	GAGG	0	0	0	9	4	2	35.69	высокая	-13.8	-3.1	4.53
139	GAGU	2	6	47	68	11	8	12.45	высокая	-14.6	0	3.73
140	GAUA	17	9	0	20	270	410	54.73	высокая	-13.3	0	5.41
141	GAUC	28	40	47	43	25	9	21.56	высокая	-14	0	3.12
142	GAUG	1	0	5	25	17	11	24.9	высокая	-14.2	0	4.53
143	GAUU	7	7	8	51	152	31	23.7	высокая	-15	0	4.67
144	GCAA	19	58	112	244	261	208	20.24	высокая	-16	0	4.43
145	GCAC	39	85	173	92	33	7	19.88	низкая	-16	0	3.04
146	GCAG	33	68	79	41	15	6	20.24	низкая	-16	0	2.81
147	GCAU	60	131	202	247	136	65	10.12	высокая	-17.2	0	3.55
148	GCCA	89	178	167	35	7	7	19.88	низкая	-15.7	0	2.41
149	GCCC	95	170	95	22	11	19	19.01	низкая	-15.9	0	2.37
150	GCCG	329	146	54	60	4	3	2.74	низкая	-19.1	0	1.78
151	GCCU	194	182	94	34	13	8	2.19	низкая	-20.7	0	2.07
152	GCGA	23	40	61	73	103	77	14.38	высокая	-16.8	0	4.12
153	GCGC	92	114	94	36	15	5	5.37	низкая	-17.6	0	2.39
154	GCGG	20	22	27	17	9	0	13.87	низкая	-16.8	0	2.72
155	GCGU	89	148	125	62	18	18	3.69	низкая	-17.4	0	2.62
156	GCUA	13	67	257	187	51	14	7.06	высокая	-17.7	0	3.40
157	GCUC	292	212	93	32	19	1	2.19	низкая	-20.7	0	1.89
158	GCUG	82	208	113	53	7	1	7.39	низкая	-18.2	0	2.35
159	GCUU	96	187	212	68	17	5	5.9	низкая	-18.4	0	2.55
160	GGAA	2	5	9	86	281	54	25.58	высокая	-14.1	-0.8	4.83
161	GGAC	21	26	47	121	22	5	22.65	высокая	-14.4	-0.5	3.46
162	GGAG	5	6	0	6	7	0	13.81	высокая	-15.5	-3.1	3.17
163	GGAU	6	8	21	86	149	18	10.59	высокая	-15.4	-0.6	4.45
164	GGCA	10	46	82	27	4	6	5.9	низкая	-17.4	0	2.93
165	GGCC	43	60	30	7	1	1	6.14	низкая	-17.3	0	2.06
166	GGCG	1	32	22	15	2	1	4.3	низкая	-18.2	0	2.84
167	GGCU	116	70	32	5	7	4	0.71	низкая	-22.1	0	1.84
168	GGGA	5	9	32	46	7	12	48.71	высокая	-14.6	-1.5	3.69
169	GGGG	1	7	30	24	1	0	37.18	высокая	-14.6	-2.2	3.27
170	GGGU	0	30	40	7	10	0	18.93	низкая	-18	-0.4	2.97
171	GGUA	5	7	35	15	12	4	13.2	высокая	-15.5	-0.1	3.44
172	GGUC	56	67	45	17	4	5	4.88	низкая	-17.1	0	2.28
173	GGUG	54	29	7	3	2	0	10.12	низкая	-15.9	-0.1	1.63
174	GGUU	11	55	105	38	10	10	2.09	низкая	-19.7	0	3.05
175	GUAA	14	7	16	129	372	49	92.99	высокая	-14.1	0	4.68
176	GUAC	10	15	43	95	11	6	48.2	высокая	-14.1	0	3.56
177	GUAU	13	36	80	146	147	15	40.63	высокая	-15.6	0	3.97
178	GUCA	36	70	153	79	37	9	32.44	высокая	-14.5	0	3.10
179	GUCC	153	108	19	19	9	2	11.02	низкая	-16.6	0	1.80
180	GUCG	33	39	13	0	3	1	18.07	низкая	-16.4	0	1.92
181	GUCU	267	106	39	17	8	5	7.02	низкая	-19.9	0	1.66
182	GUGA	42	43	50	11	4	3	57.71	низкая	-14.7	0	2.35
183	GUGC	139	51	10	6	2	0	28.22	низкая	-18.4	0	1.47
184	GUGU	73	24	4	6	2	0	15.51	низкая	-16.1	0	1.53
185	GUUA	0	11	65	147	43	26	40.63	высокая	-15.3	0	4.03
186	GUUC	31	78	71	14	3	4	12.61	низкая	-18.3	0	2.46
187	GUUG	4	1	28	1	8	0	44.45	высокая	-15.8	0	3.19
188	GUUU	25	83	74	22	16	5	22.63	низкая	-17.3	0	2.72
189	UAAA	25	24	40	465	552	92	32.32	высокая	-13.4	0	4.48
190	UAAAC	10	8	54	193	106	44	24.67	высокая	-13.4	0	4.23
191	UAAAG	6	5	21	140	34	12	18.17	высокая	-14.9	0	4.04

192	UAAU	6	9	116	377	129	53	12.06	высокая	-15.8	0	4.12
193	UACA	18	31	124	217	33	12	24.9	высокая	-13.5	0	3.58
194	UACC	3	22	78	92	10	14	24.67	высокая	-13.4	0	3.58
195	UACG	8	16	68	52	6	6	12.68	высокая	-15.7	0	3.32
196	UACU	16	48	105	16	11	1	2.99	низкая	-18.2	0	2.80
197	UAGA	14	8	23	94	27	5	32.32	высокая	-13.4	0	3.74
198	UAGC	5	7	42	9	5	4	11.48	высокая	-15.4	0	3.19
199	UAGU	5	21	152	70	7	12	5.79	высокая	-16.3	0	3.33
200	UAUA	17	4	23	262	120	17	32.32	высокая	-13.4	0	4.16
201	UAUC	20	17	101	124	15	19	9.59	высокая	-14.7	0	3.52
202	UAUG	3	19	49	82	17	6	18.17	высокая	-14.9	0	3.62
203	UAUU	11	12	153	154	24	15	8.05	высокая	-16	0	3.58
204	UCAA	116	352	499	170	66	25	10.83	низкая	-15.4	0	2.83
205	UCAC	165	292	111	38	21	3	7.73	низкая	-18.9	0	2.15
206	UCAG	61	108	71	20	9	12	10.83	низкая	-15.4	0	2.44
207	UCAU	221	383	212	64	24	22	5.39	низкая	-16.7	0	2.30
208	UCCA	120	132	64	22	29	0	12.12	низкая	-15	0	2.20
209	UCCC	144	123	50	42	6	1	8.38	низкая	-15.9	0	2.03
210	UCCG	67	55	25	3	5	1	4.63	низкая	-17.2	0	1.89
211	UCCU	228	106	54	18	4	0	3.92	низкая	-18.9	0	1.69
212	UCGA	95	148	58	17	13	4	7.39	низкая	-16.2	0	2.16
213	UCGU	221	74	48	13	10	0	3.11	низкая	-18.5	0	1.68
214	UCGU	380	104	47	24	9	3	2.35	низкая	-17.9	0	1.57
215	UCUA	260	206	93	38	12	11	3.6	низкая	-17.2	0	1.98
216	UCUC	418	177	53	25	20	5	1.34	низкая	-19.8	0	1.66
217	UCUG	187	135	39	6	14	5	5.39	низкая	-17.6	0	1.81
218	UCUU	322	178	44	4	11	1	3	низкая	-18	0	1.58
219	UGAA	24	22	178	163	37	23	9.94	высокая	-16.2	0	3.53
220	UGAC	6	39	62	37	21	3	8.8	высокая	-16.5	0	3.22
221	UGAG	2	6	43	61	23	2	5.37	высокая	-17.6	0	3.75
222	UGAU	32	51	192	167	15	3	4.11	высокая	-17.5	0	3.20
223	UGCA	52	170	140	30	6	23	2.29	низкая	-19.5	0	2.61
224	UGCC	162	95	65	29	8	1	2.39	низкая	-19.4	0	1.97
225	UGCG	28	63	2	5	15	0	1.67	низкая	-20.3	0	2.26
226	UGCU	138	107	49	6	4	0	0.28	низкая	-24.2	0	1.79
227	UGGA	18	94	146	33	26	26	11.53	высокая	-16.7	-1.2	3.10
228	UGGC	73	138	51	15	3	1	11.53	низкая	-16.8	0	2.07
229	UGGG	36	42	22	8	12	5	8.8	низкая	-16.7	-0.3	2.46
230	UGGU	49	158	61	14	19	0	4.1	низкая	-20.1	-0.2	2.32
231	UGUA	25	68	254	29	22	13	5.13	низкая	-17.6	0	2.99
232	UGUC	115	98	37	16	6	2	3.73	низкая	-18.3	0	1.93
233	UGUU	106	217	44	41	21	10	0.81	низкая	-21.8	0	2.28
234	UUAА	32	35	118	584	344	66	27.24	высокая	-13.3	0	4.16
235	UUAС	5	31	182	101	40	6	18.84	высокая	-16.7	0	3.43
236	UUAG	6	18	106	30	13	5	18.17	высокая	-14.9	0	3.23
237	UUАU	12	24	182	219	69	30	17.37	высокая	-15.2	0	3.74
238	UUAС	34	76	217	51	36	14	27.24	низкая	-14	0	3.05
239	UUCС	24	48	78	41	2	2	18.09	низкая	-14.9	0	2.77
240	UUCG	10	112	99	11	5	6	12.68	низкая	-15.7	0	2.62
241	UUCU	92	241	87	12	12	0	3.29	низкая	-18.5	0	2.12
242	UUGA	4	19	179	111	23	4	16.61	высокая	-14	0	3.42
243	UUGC	76	81	38	6	14	5	7	низкая	-15.7	0	2.16
244	UUGG	21	38	76	20	1	0	16.61	низкая	-14.9	0	2.63
245	UUGU	58	275	83	49	8	1	6.06	низкая	-16.3	0	2.32
246	UUUA	13	69	328	111	22	7	11.58	высокая	-14.6	0	3.15
247	UUUC	243	116	49	28	14	4	4.3	низкая	-17.2	0	1.82
248	UUUG	32	97	116	32	17	4	17.37	низкая	-15.2	0	2.72
249	UUUU	79	275	163	37	22	18	9.68	низкая	-16	0	2.50

\* Эффективность трансляции, предсказанная с помощью RBS Calculator [150].

\*\* Результат классификации всех мРНК на две группы: с плохо (низкая эффективность трансляции) и эффективно (высокая эффективность трансляции) транслируемыми мРНК.

\*\*\* MFE, минимальная энергия укладки.

\*\*\*\* «Энергия связывания 16S рРНК» - свободная энергия гибридизации спейсерного участка, содержащего 4 нт рандомизированный фрагмент мРНК и фланкирующие его 1 нт перед и 4 нт после, между константной последовательностью SD и AUG старт-кодоном с последовательностью анти-SD CACCUCU 3'-концевого участка 16S рРНК.

\*\*\*\*\* Средневзвешенный номер фракции, исходя из распределения Гаусса для отдельного варианта последовательности.

Таблица 2. Варианты последовательностей рандомизированного фрагмента, теоретически возможные, но не встретившиеся среди вариантов, полученных с помощью NGS.

№	Вариабельный спейсерный участок
1	AUGG
2	GGGC
3	GUAG
4	GUGG
5	UAGG
6	UCGG
7	UGUG

## ПРИЛОЖЕНИЕ 3

Таблица 1. Свойства вариантов природных 5'-UTR *E. coli* перед геном *cer*.

Название гена	Последовательность 5'-UTR	SD	Длина SD	Расстояние от SD до старт-кодона AUG (при расчёте от центрального G)	CER/RFP	Стандартное отклонение CER/RFP
<i>eno</i>	UUAACUAGUGACUUGAGGAAAACCUAA	AGGA	4	9	7,3548	4,726
<i>gpmA</i>	UUAUUUUCUUAUUAAAAGAUUUUGAGGAGUUAAGUUAU	AGGAG	5	12	5,1973	0,036
<i>tabA</i>	ACGUUGUUUUUUUUAAAGGAGUGGAAAAG	UAAGGAG	7	11	4,4163	0,626
<i>gmhB</i>	ACAUGCCUUUUUAAAGGAGCUUUAAAAG	UAAGGAG	7	13	3,9606	0,279
<i>dsrB</i>	AAGCCACAGGAGGAAAACGA	AGGAGG	6	9	3,6506	0,149
<i>ycsY</i>	AUGAAUCGUAUUUUUUCACAAGUGGCUAAGGAGAAAAGU	UAAGGAG	7	9	3,3045	0,633
<i>yefM</i>	UACAAAAGAGGAGAUUGACA	AGGAG	5	10	2,8703	0,044
<i>slyA</i>	CAAGCUAAUUUAAAGGAGAUUAAAAGU	UAAGGAG	7	11	2,7237	0,113
<i>deoB</i>	AUCAUUUUUUUUAAGGACUGAGUACGGAGAACAUC	GGAG	4	6	2,6506	0,042
<i>kdgR</i>	UAAAACACUCGCAUUGAGGUGUAAA	GAGGU	5	7	2,5857	0,163
<i>adk</i>	UCAAAAAUUCGACACAUUUUUAAGGGGAUUUCGCA	UAAGG	5	13	2,0337	0,066
<i>glnS</i>	GUUUUACGUUUUUUACGCUUUUGAGGAAUCCACG	AGGA	4	8	1,8669	0,094
<i>mgrB</i>	UAAGGUAAGGUAACGAGGAGUUGGAG	UAAGG	5	21	1,7638	0,201
<i>gyrB</i>	AACCCAAAGUUAUUUUAAGGAGGAGAAACGUUGA	GAG	3	14	1,3805	0,087
<i>rpoZ</i>	CAUUUCUUUACCCUGGAGGCUUUUUAAAGU	UAAG	4	11	1,3720	0,084
<i>yrbL</i>	AGAUUCGUCAGAAAUAUUUCCAGGAGAUUGGC	AGGAG	5	7	1,3004	0,178
<i>pepD</i>	GAUCGACAAGGAGACUUAACG	AAGGAG	6	11	1,1881	0,089
<i>yciI</i>	ACUUUCCCCGAACAAGGAGUUGUGCCCCG	AAGGAG	6	12	1,0751	0,053
<i>ymcE</i>	GCAAAUACGACUGACC	GA	2	2	1,0552	0,051
<i>malP</i>	ACGUUGUGUUGAAAAUCUUAAGAAAGUGGAACUCCUA	UAAG	4	8	0,9487	0,114
<i>nfuA</i>	UAAAAUAGUCAACUUAUAGGCCAUUACU	AGG	3	7	0,9097	0,081
<i>hisL</i>	CAUCAGUUAAUAAACAUCACAGAGACUUUU	GAG	3	6	0,6485	0,061
<i>gapA</i>	AACUUUUUUUACUUAACAUAUAGCUGGUGGAAUUAU	GGA	3	6	0,6481	0,036
<i>ihvL</i>	AUUCCUUCGAACAAGUAGCAAGAAAAGACAAA	AAG	3	20	0,6288	0,111
<i>styD</i>	CCACUUAUUUCCCCAUGCUCAGGAGAUUAUCA	AGGAG	5	8	0,5630	0,054
<i>ycsA</i>	ACAUUCUGUUAUAGAGAGUAGACUCAA	GAG	3	9	0,5329	0,068
<i>yciN</i>	CCUUUUUUUUAAGGAGUUAUUAA	UAAGGAG	7	11	0,5035	0,028
<i>gmK</i>	AUGUAGGCUUUUUUUCGCUUAUCACAUACGAAAGAUACUC	AAG	3	34	0,4843	0,046
<i>ivy</i>	AGGAGGUUAUAAC	AGGAGGU	7	9	0,3645	0,021
<i>mprA</i>	GAAAUUUAUUGAGGUCUACCCAA	GAGGU	5	14	0,3123	0,027
<i>cysK</i>	GUUAUCCAAUUUCAUACAGUUAAGGACAGGCC	UAAGGA	6	8	0,2989	0,017
<i>yihD</i>	CAAGAGUUAAGGAGCAUCU	UAAGGAG	7	11	0,2504	0,002
<i>rbsD</i>	AGCGAAACGUUUUCGUGAUGGAGAAAAA	GGAG	4	7	0,1903	0,014
<i>cpvR</i>	GGAGGUUUUUAAACAA	GGAGGU	6	11	0,1358	0,018
<i>ribF</i>	GAUUUUUACUGUUUUGAGCCAGAC	GAG	3	6	0,1042	0,004
<i>orn</i>	GUGAAUUGGUGGAAAAUAGC	GGA	3	10	0,0582	0,004
<i>ubiD</i>	GCAGAUUAAGGAGGCUACA	GAGG	4	5	0,0486	0,003
<i>yobF</i>	GGUUAACAGGAGUACAGUUCUGUUUA	GAGGU	5	14	0,0396	0,004
<i>tig</i>	GAGGUAAACAG	GAGGU	5	6	0,0359	0,002
<i>mipA</i>	CCCGCUAAACAUUUUUUUAAGGAAUGAUUUUUU	UAAGGA	6	15	0,0356	0,002
<i>fadR</i>	GCUGUGUUUUGGAAAUUCACU	GGA	3	10	0,0199	0,001
<i>gpsA</i>	AAGAACAUCAGGAGUCCUGA	AGGA	4	9	0,0121	0,001
<i>ygiW</i>	GAGUAAUAAACA	UAA	3	8	0,0109	0,001
<i>yoaB</i>	ACAAAAG	AA	2	7	0,0104	0,001
<i>yqiC</i>	GAAACUACA	UA	2	6	0,0101	0,000
<i>lrhA</i>	GUAAAGUAUAUAUA	UAAG	4	12	0,0098	0,000
<i>lpxD</i>	UUAAAUAAGUA	UAAG	4	6	0,0079	0,001
<i>tdk</i>	GCCUGGGCUG	GG	2	1	0,0034	0,001

Таблица 2. Свойства вариантов синтетических 5'-UTR *E. coli* перед геном *cer*.

Название конструкции	Последовательность 5'-UTR	SD	Длина SD	Расстояние от SD до старт-кодона AUG (при расчёте от центрального G)	CER/RFP	Стандартное отклонение CER/RFP
"8/7"	CACACAACAAGGAGGUACUUAUG	AAGGAGGU	8	7	17	1
"8/10"	CACACAAGGAGGUACAACUUAUG	AAGGAGGU	8	10	10	2
"8/13"	CACAAGGAGGUACAACAACUUAUG	AAGGAGGU	8	13	0,8	0,1
"8/16"	AAGGAGGUCACACAACAACUUAUG	AAGGAGGU	8	16	0,5	0,02
"6/7"	CACACAACACAGGAGGAACUUAUG	AGGAGG	6	7	2	0,2
"6/10"	CACACACAGGAGGAACAACUUAUG	AGGAGG	6	10	13	2
"6/13"	CACCAGGAGGAACAACAACUUAUG	AGGAGG	6	13	0,4	0,04
"6/16"	CAGGAGGACACACAACAACUUAUG	AGGAGG	6	16	0,7	0,08
"4/7"	CACACAACACCCGAGCAACUUAUG	GGAG	4	7	0,8	0,04
"4/10"	CACACACCCGAGCAACAACUUAUG	GGAG	4	10	0,6	0,02
"4/13"	CACCCGAGCAACAACAACUUAUG	GGAG	4	13	0,06	0,03
"4/16"	CCGAGCACAACAACAACUUAUG	GGAG	4	16	0,02	0,01
"2/7"	CACACAACACCGUACAACAACUUAUG	GA	2	7	0,03	0,005
"2/10"	CACACACCGUACAACAACAACUUAUG	GA	2	10	0,04	0,01
"2/13"	CACCCUGAUCACAACAACAACUUAUG	GA	2	13	0,008	0,006
"2/16"	CCUGAUCACAACAACAACAACUUAUG	GA	2	16	0,02	0,006
0	CCUCCUACAACAACAACAACUUAUG		0	0	0,001	0,001
4/7 A/U	UAUUUUAAUUAAUAAACACCCGAGCAACAACUUAUG	GGAG	4	7	3,7	1,2