Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Московский государственный университет имени М.В.Ломоносова»

На правах рукописи

Ершова Анна Степановна

Анализ систем рестрикции-модификации в полногеномном контексте

03.01.09 – Математическая биология, биоинформатика

ДИССЕРТАЦИЯ
на соискание ученой степени
кандидата биологических наук

Научный руководитель:
кандидат физико-математических наук
Алексеевский А.В.

Москва, 2016

Содержание

Введение....4
Список принятых сокращений....9
Глава 1. Обзор литературы....10
1.1 Классификация и номенклатура систем рестрикции-модификации....10
1.1.1 Тип I....12
1.1.2 Тип II....20
1.1.3 Тип III....34
1.1.4 Метил-зависимые системы (Тип IV и IIM)....36
1.2 Организация генов систем рестрикции-модификации в геноме и их мобильность. .40
1.3 Одиночные гены систем рестрикции-модификации....42
1.4 Функции систем рестрикции-модификации в клетке....43
1.4.1 Защита от бактериофагов....43
1.4.2 Влияние метилирования генома на регуляцию экспрессии генов....45
1.5 Системы рестрикции-модификации как эгоистичный элемент генома....48
1.6 Влияние систем рестрикции-модификации на эволюцию геномов прокариот....50
1.6.1 Изменение олигонуклеотидного состава генома....50
1.6.2 Влияние на перестройки генома....53
1.6.3 Влияние на горизонтальный перенос генов и поддержание гетерогенности популяции....54
1.6.4 Взаимодействие между различными системами рестрикции-модификации....56
1.7 Методы сравнительной геномики....57
1.7.1 Сходство генов: гомологи, ортологи и паралоги....57
1.7.2 Аннотация систем рестрикции-модификации в БД REBASE....59
1.7.3 Оценка частот олигонуклеотидов в геномах....60
1.8 Заключение....62
Глава 2. Материалы и методы....64
2.1 Последовательности геномов и системы рестрикции-модификации....64
2.2 Анализ состава систем рестрикции-модификации....66
2.3 Поиск генов ДНК-метилтрансфераз....66
2.4 Поиск ортологичных белков....67
2.5 Поиск ортологичных систем рестрикции-модификации....68
2.6 Анализ геномного контекста для генов рассредоточенных систем рестрикции-модификации....68
2.7 Оценка недопредставленности сайтов в геноме....69
2.8 Сравнение распределений величины *Kr* и границы недо- и перепредставленности..70
2.9 Идентификация генов эндонуклеаз рестрикции, предположительно недавно полученных путем горизонтального переноса генов....71
2.10 Определение семейств белков....71
2.11 Определение семейств систем рестрикции-модификации....72
2.12 Построение модели влияния систем рестрикции-модификации на недопредставленность сайта в геноме....72
Глава 3. Результаты и обсуждение....74
3.1 Организация генов систем рестрикции-модификации....74

3.1.1 Идентификация одиночных эндонуклеаз рестрикции в полных геномах бактерий и архей и их классификация....74
3.1.2 Сравнение идентифицированных одиночных эндонуклеаз рестрикции с метил-зависимыми эндонуклеазами рестрикции....77
3.1.3 Группы ортологичных систем рестрикции-модификации....82
3.1.4 Рассредоточенные системы типа I....84
3.1.5 Рассредоточенные системы типа II....86
3.1.6 Геномный контекст генов рассредоточенных систем рестрикции-модификации…90
3.1.7 Одиночные эндонуклеазы рестрикции, для которых не были найдены парные ДНК-метилтрансферазы....90
3.1.8 Заключение по разделу....94
3.2 Недопредставленность сайтов систем рестрикции-модификации в геномах прокариот....95
3.2.1 Избегание сайтов систем рестрикции-модификации различных типов....96
3.2.2 Избегание палиндромных и непалиндромных сайтов....102
3.2.3 Перепредставленные сайты систем P-M....106
3.2.4 Влияние продолжительности жизни систем рестрикции-модификации в геноме на недопредставленность палиндромных сайтов....106
3.2.5 Следы потерянных систем рестрикции-модификации....111
3.2.6 Выделяющиеся сайты....114
3.2.7 Изучение недопредставленности сайта GATC....115
3.2.8 Заключение по разделу....127
Выводы....128
Список публикаций по теме диссертации....130
Список литературы....133

Введение

Актуальность темы исследования. Системы рестрикции-модификации (Р-М) широко распространены среди прокариот. Большинство прокариот содержит от одной до четырех систем Р-М. Классические системы Р-М включают ферменты с двумя типами активности: ДНК-метилтрансферазы способны метилировать определенные последовательности ДНК (сайты узнавания), эндонуклеазы рестрикции расщепляют ДНК, если соответствующий сайт неметилирован. Также к системам Р-М относят эндонуклеазы рестрикции, расщепляющие метилированную ДНК. Благодаря колокализации генов, системы Р-М способны перемещаться между геномами за счет горизонтального переноса генов и могут рассматриваться как своеобразные формы жизни, подобно вирусам или транспозонам [1].

Системы Р-М были открыты благодаря их способности защищать бактерий от бактериофагов, однако последующие исследования показали, что они влияют на различные эволюции и экологии бактерий: от олигонуклеотидного состава генома [2–4] до регуляции экспрессии генов [5] и патогенности [6]. Таким образом, изучение эволюции систем Р-М и их влияния на эволюцию прокариот необходимо для более глубокого понимания механизмов изменения патогенности бактерий и является актуальной научной задачей.

Эволюция систем Р-М и влияние систем Р-М на эволюцию кодирующих их геномов прокариот обсуждается в литературе с момента открытия систем Р-М. Однако большинство выводов основано на изучении небольшого числа геномов прокариот и систем Р-М. Например, избегание сайтов узнавания систем Р-М было показано только для палиндромных последовательностей длины 4-6 п.н. [2–4,7] для нескольких десятков бактериальных геномов.

В последние годы, благодаря развитию технологий секвенирования геномов, количество известных геномов выросло в сотни раз, также увеличилось

количество известных систем Р-М и их сайтов узнавания. В связи с этим появилась возможность изучения эволюции систем Р-М и их влияния на геномы прокариотических хозяев на основе тысяч доступных в настоящее время геномов прокариот.

Цели и задачи исследования. Цель данной работы заключалась в исследовании влияния систем Р-М на геномы прокариот методами биоинформатики и сравнительной геномики.

Для достижения данной цели были поставлены следующие задачи:

1. Поиск рассредоточенных систем Р-М в доступных геномах прокариот.
2. Исследование влияния систем Р-М на встречаемость их сайтов в геномах прокариот, кодирующих данные системы.
3. Исследование влияния свойств системы Р-М на недопредставленность сайтов систем Р-М в геномах прокариот, включая тип системы Р-М, особенности ее сайта узнавания (длина, вырожденность и (а)симметрия) и продолжительность жизни данной системы Р-М в геноме.

Научная новизна.

В результате анализа систем Р-М в полных геномах прокариот, впервые проведен систематический поиск рассредоточенных систем Р-М методами биоинформатики.

Полученные результаты по оценке влияния систем Р-М на недопредставленность своих сайтов в геномах прокариот являются новыми. На основе анализа сотен сайтов узнавания систем Р-М в тысячах геномов прокариот показано, что только системы Р-М типа II (исключая системы типа IIM и IIG) вызывают недопредставленность своих сайтов узнавания в соответствующих геномах прокариот. При этом впервые показано, что палиндромные и асимметричные сайты узнавания избегаются в равной степени.

Впервые показано, что уменьшение числа сайтов систем Р-М в геноме может свидетельствовать о прошлой активности потерянных в настоящее время систем Р-М. Такой вывод сделан на основании того, что в геномах бактерий недопредставлены не только сайты систем Р-М, закодированные в них, но также сайты систем Р-М, закодированные в геномах близких родственников.

На примере последовательности GATC впервые показано, что снижение числа сайтов систем Р-М может быть связано с наличием в популяции бактерий взаимоисключающих систем Р-М.

Теоретическая и практическая значимость. В работе исследованы системы Р-М в более, чем двух тысячах геномах прокариот. На основании анализа полученных данных в геномах были найдены рассредоточенные системы Р-М, гены которых не колокализованы, а разнесены на значительные расстояния в геноме. Существование таких систем дополняет существующие представления об эволюции систем Р-М.

Предложенная методика поиска рассредоточенных систем Р-М может быть использована при аннотации и реаннотации геномов прокариот.

Анализ встречаемости сайтов узнавания систем Р-М в геномах прокариот выявил, что только системы Р-М типа II, ЭР и МТаза которых действуют независимо, вызывают избегание своих сайтов в геномах прокариот, кодирующих эти системы. Выявленный сдвиг во времени между появлением и исчезновением генов систем Р-М в геномах и недопредставленностью их сайтов узнавания позволяет анализировать потерянные системы Р-М, что может быть интересно с точки зрения изучения эволюции прокариот и систем Р-М.

В работе найдена связь между недопредставленностью сайта узнавания GATC и наличием взаимоисключающих систем Р-М в популяции бактерий, способных к обмену ДНК. Данные результаты являются одним из немногих известных примеров взаимодействия между различными системами Р-М, и влиянием этого

взаимодействия на эволюцию геномов прокариот.

Основные результаты и положения, выносимые на защиту:

1. Гены белков, входящих в одну систему Р-М, могут быть не колокализованы, и находиться на большом (больше 4 т.п.н.) расстоянии друг от друга. Такие системы Р-М предложено называть рассредоточенными.
2. Предложен метод систематического поиска рассредоточенных систем Р-М в геномах прокариот, который заключается в поиске систем Р-М, содержащих белки, гомологичные одиночным эндонуклеазам рестрикции (и ДНК-метилтрансферазам).
3. Системы Р-М типа II, состоящие из независимо действующих эндонуклеазы рестрикции и ДНК метилтрансферазы, вызывают недопредставленность своих сайтов в кодирующих их геномах независимо от свойств сайта. Сайты систем Р-М типов I, III, IV, IIC/G, как правило, не избегаются в соответствующих геномах.
4. Продолжительность жизни систем Р-М в геномах прокариот влияет на недопредставленность соответствующих сайтов узнавания в данных геномах.
5. В геномах прокариот обнаруживается недопредставленность сайтов потерянных систем Р-М.
6. Избегание сайта узнавания системы Р-М может быть адаптацией к горизонтальному переносу генов между бактериями, имеющими взаимоисключающие системы, способные расщеплять один и тот же метилированный или неметилированный сайт.

Степень достоверности и апробация результатов. Материалы диссертации опубликованы в 4 статьях в рецензируемых научных журналах и в 14 тезисах сборников трудов конференций. Результаты работы были представлены на

международных конференциях Moscow Conference on Computational Molecular Biology (MCCMB) (Москва, Россия) в 2007, 2009, 2011, 2013, 2015 гг, Molecular Genetics of Bacteria and Phages Meeting, 2013 (Мэдисон, США), Симпозиумах DFG/RFBR 2007-2015 (Россия, Литва, Германия), BGRS'14 (Новосибирск, Россия), ECCB'14 (Страсбург, Франция). Список публикаций по теме диссертации приведен в конце работы.

Личный вклад автора. Результаты, изложенные в диссертации, получены лично автором. Постановка изложенных в диссертации цели и задач была сделана научным руководителем к. ф.-м.н. Алексеевским А.В. Диссертант участвовал в подготовке всех публикаций по теме диссертации. В совместных публикациях к. ф.-м.н. Алексеевскому А.В., к. ф.-м.н. Спирину С.А., д.б.н. Карягиной А.С. принадлежат постановки задач и указания основных направлений исследований, подготовка публикаций к печати. Русиновым И.С. в работах (Rusinov, 2015; Ershova, 2016) выполнена программная реализация формулы Карлина для подсчета ожидаемого числа сайтов в геноме, расчет представленности сайтов рестрикции в геномах. Васильевым М.О. выполнена программная реализация алгоритма поиска рассредоточенных систем P-M в работе (Ershova, 2012) и программная реализация анализа совместного влияния систем P-M, узнающих последовательность GATC, на недопредставленность этой последовательности в геноме методом линейной регрессии в работе (Ershova, 2016). Все остальные результаты получены диссертантом. В частности, в работе (Ershova, 2012) диссертантом найдены гены одиночных эндонуклеаз рестрикции в геномах прокариот, предложен алгоритм поиска рассредоточенных систем в геноме, выполнен поиск рассредоточенных систем в доступных геномах прокариот. В работе (Rusinov, 2015) диссертант выполнил анализ представленности сайтов рестрикции в геномах, поиск недавно приобретенных систем P-M. Обзор литературы о роли систем P-M в эволюции и экологии прокариот (Ershova, 2015) написан на основе обзора литературы для данной диссертации. В работе (Ershova, 2016) диссертант классифицировал GATC – специфичные эндонуклеазы

рестрикции и ДНК-метилтрансферазы, выявил и проанализировал феномен недопредставленнности последовательности GATC в присутствии взаимоисключающих систем Р-М в разных штаммах одного вида бактерий.

Структура и объем диссертации. Диссертация состоит из введения, 3 глав, выводов и библиографии. Общий объем диссертации, включая 15 рисунков и 9 таблиц, составляет 152 страницы, в том числе библиография включает 264 наименования на 24 страницах.

Список принятых сокращений

Система Р-М – система рестрикции модификации; ЭР – эндонуклеаза рестрикции; МТаза – ДНК-метилтрансфераза; АТФ – аденозинтрифосфат; ГТФ – гуанозинтрифосфат; SAM – S-аденозилметионин; БД – база данных; *Kr* – отношение наблюдаемого числа сайтов узнавания систем Р-М к ожидаемому, рассчитанное по формуле, предложенной в работе Karlin и Cardon [8].

Глава 1. Обзор литературы.

Обзор литературы описывает имеющие данные о системах Р-М, их влиянии на эволюцию геномов прокариот, а также методы и подходы биоинформатики, использующиеся в сравнительной геномике.

1.1 Классификация и номенклатура систем рестрикции-модификации

Классические системы Р-М включают ферменты с двумя видами активности. Эндонуклеаза рестрикции (ЭР) способна узнавать определенную последовательность ДНК (сайт узнавания) и расщеплять ДНК, если эта последовательность неметилирована. ДНК-метилрансфераза (МТаза) способна взаимодействовать с той же самой последовательностью ДНК и метилировать ее, защищая таким образом от гидролиза ЭР [9].

Несмотря на сходство функции, системы Р-М очень разнообразны по структурно-функциональной организации. В настоящее время принята классификация систем Р-М, предложенная в работе [9], в которой выделяется четыре типа систем Р-М I, II, III и IV, различающихся по составу МТазы и ЭР, строению сайта узнавания, требованиями кофакторов, и расстоянием от сайта узнавания до позиции, в которой происходит гидролиз ДНК (см. таблицу 1.1).

В работе [9] также предложена система наименования генов и белков систем Р-М, которая в настоящее время является общепринятой. Первая буква в названии системы Р-М соответствует роду, следующие две виду бактерии, из которой данная система была изолирована; дополнительные буквы и арабские цифры идентифицируют штамм или серотип.

Различные системы Р-М из одного и того же организма обозначаются римскими цифрами в порядке нахождения той или иной системы. Так, EcoRI была первой системой Р-М, изолированной из штамма *Escherichia coli* RY13.

Таблица 1.1

Общая характеристика типов систем Р-М. Взята из работы [10]

Характеристика	Тип I	Тип II	Тип III	Тип IV
Пример	EcoKI	EcoRI	EcoP1I	EcoMcrBC
Гены	hsdR, hsdM, hsdS	R, M	res, mod	mcrB, mcrC
Состав ЭР / МТазы	$R_2M_2S_1/M_2S_1$	R_2/M_1	mod_2res_1 с обеими активностями	McrBC
Сайт узнавания	ассиметричный, из двух частей AAC(N6)GTGC	часто палиндром, GAATTC	асимметричный AGACC	метилированный, из двух частей RmC(N30–4000)RmC
Расстояние от сайта узнавания до сайта гидролиза	большое, до 1000 п.н. ДНК режется в случайном месте	определенное место внутри или рядом с сайтом узнавания	25-27 п.н.	внутри сайта узнавания
Кофакторы	АТФ, SAM, Mg^{2+}	SAM, Mg^{2+}	АТФ, SAM, Mg^{2+}	ГТФ, Mg^{2+}

Для идентификации белка системы Р-М используют префикс с точкой, например R.EcoRI для ЭР системы EcoRI, M. EcoRI для МТазы и т.д. При этом R в названии эндонуклеазы рестрикции может быть опущена.

Если ЭР и МТазная субъединицы объединены в одну полипептидную цепь, как например в системе Eco57I, такой фермент обозначается RM.Eco57I.

Если с одной системой Р-М ассоциировано несколько генов ЭР или МТазы, кодирующих отдельные ферменты, их обозначают арабской цифрой после префикса, например, система HphI содержит МТазы M1.HphI и M2.HphI.

Если фермент включает несколько субъединиц, то их обозначают буквенными

суффиксами A, B, C. Например, ЭР BbvCI является гетеродимером, включающим субъединицы R.BbvCIA (или просто BbvCIA) и R.BbvCIB (или просто BbvCIB).

Это общие принципы наименования систем Р-М и входящих в них белков. Некоторые другие особенности номенклатуры, характериные для различных типов систем Р-М будут приведены при описании этих типов ниже. Особенности взаимодействия систем Р-М с их сайтами узнавания и регуляция активности систем Р-М в геноме могут сильно влиять на токсичность систем Р-М для бактерий, а следовательно, и на избегание соответствующих сайтов в геномах прокариот, поэтому эти вопросы подробно рассматриваются при описании систем Р-М различных типов.

1.1.1 Тип I

1.1.1.1 Гены системы рестрикции-модификации типа I

Система Р-М типа I кодируется тремя генами, которые обозначаются hsd (host specificity determinant): *hsdR* кодирует субъединицу рестрикции (R), *hsdM* – метилтрансферазную субъединицу, и *hsdS* (S – specificity) – субъединицу, узнающую сайт на ДНК с помощью ДНК-узнающего домена (Target Recognition Domain (TRD)) (см. рисунок 1.1 А).

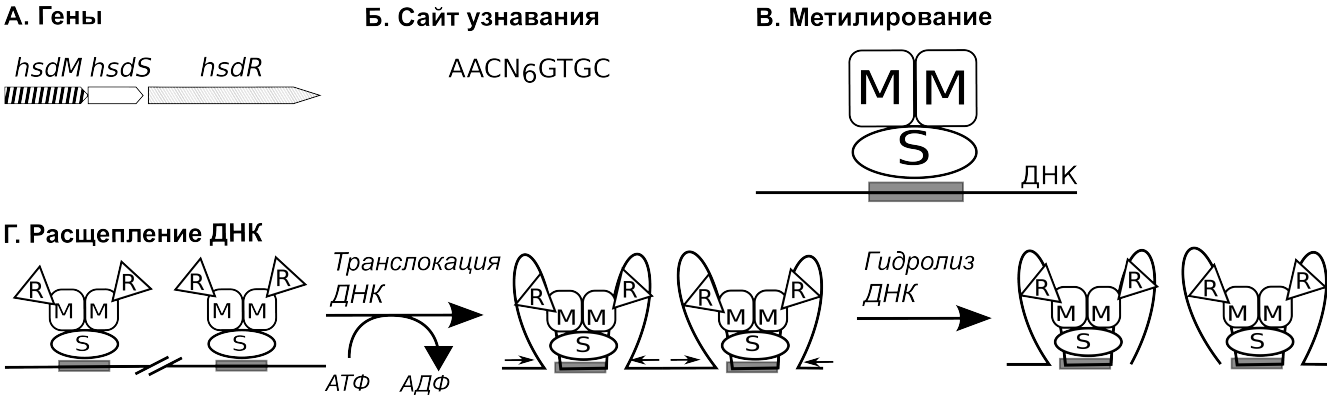


Рисунок 1.1 Характеристика типичной системы Р-М типа I. А. Схема организации генов. *hsdM* – ген ДНК-метилтрансферазы; *hsdR* – эндонуклеазы рестрикции; *hsdS* – ген белка, узнающего сайт ДНК в системах Р-М типа I. Б. Пример сайта узнавания. В. Состав комплекса для метилирования ДНК. M – метилтрансфераза, S – ДНК узнающий белок. Г. Состав комплекса, расщепляющего ДНК. R в треугольнике – эндонуклеаза рестрикции; M и S – как на рис. 1В. Черными стрелками показано направление транслокации ДНК.

Экспрессия этих генов регулируется двумя промоторами, один из которых

расположен перед геном *hsdR*, второй – перед генами *hsdM* и *hsdS* [11–13]. Биоинформатический анализ выявляет присутствие генов системы типа I примерно в половине всех известных геномов бактерий и архей. Около 40% геномов не содержат генов систем R-M типа I, и около 10% геномов содержат только один или два гена из трех [14]. Среди геномов, в которых присутствует система R-M типа I, большинство имеет один набор генов этой системы. Однако, встречаются геномы, содержащие две и более системы, в некоторых геномах их число доходит до восьми. В геномах *Staphylococcus aureus* на один ген *hsdR* приходится два гена *hsdM* и *hsdS*. При этом гены *hsdM* и *hsdS* колокализованы, а ген *hsdR* расположен на расстоянии, значительно большем, чем обычно характерно для генов систем R-M [15]. Для геномов некоторых видов, например *Mycoplasma*, характерно наличие одного или двух генов *hsdM* и *hsdR*, и нескольких генов S-субъединиц: в геномах *M.pneumoniae* было найдено 10 генов S-субъединиц, в *M. suis* – 13, и в *M. haemofelis* – 22 гена. При этом эти гены попеременно находятся то в активном, то в молчащем состоянии, что обеспечивает постоянное изменение специфичности и лучшую защиту от бактериофагов [16,17].

1.1.1.2 Классификация

В настоящее время тип I делят на пять семейств от A до E на основании сходства аминокислотной последовательности [18–20]. В один подтип объединяют системы R-M с очень сходными последовательностями всех субъединиц, различия есть только в областях, соответствующих ДНК-узнающим доменам (TRD) S-субъединиц [18].

1.1.1.3 Узнавание ДНК

Ферменты системы типа I узнают последовательность ДНК, состоящую из двух частей, находящихся на расстоянии друг от друга, например, AACNNNNNNGTGC. Такое строение сайта связано со структурой S-субъединицы, которая содержит два отдельных ДНК-узнающих домена (TRD),

каждый из которых взаимодействует с одной частью сайта узнавания (см. рисунок 1.2). Каждый TRD состоит из вариабельного домена, взаимодействующего с ДНК и консервативного альфа-спирального “димеризующего” домена. В последовательности белка TRD расположены как прямой повтор, но в пространственной структуре белка они инвертированы из-за антипараллельного взаимодействия димеризующих доменов.

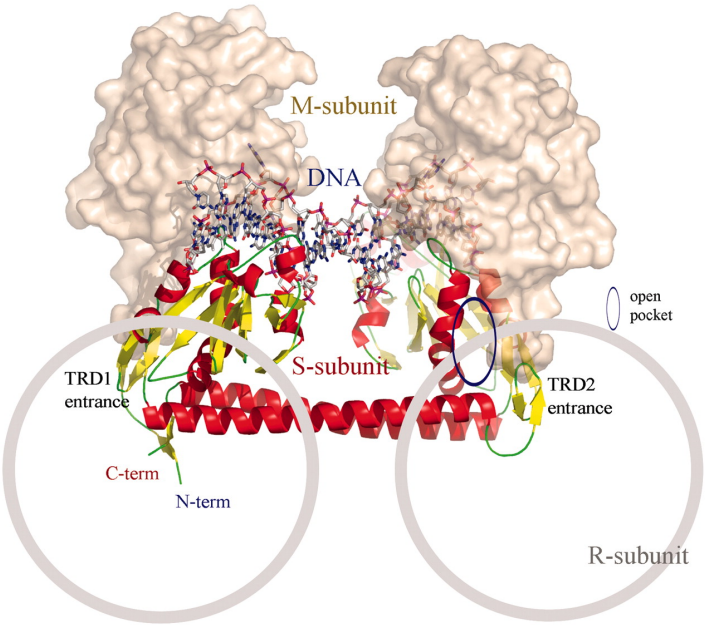


Рисунок 1.2 Модель сборки белковых субъединиц системы P-M типа I. S-субъединица показана в ленточном представлении, ДНК – в шаростержневом, M-субъединица TaqI показана в виде поверхности белка. Взаимное расположение M- и S- субъединиц определено методами молекулярного докинга, учитывая функциональные особенности белков. Возможные области связывания R-субъединиц показаны серыми кругами на основе особенностей структуры и доступных экспериментальных данных. Рисунок взят из работы Kim и соавт.[21].

Структура S-субъединицы способствует быстрому изменению специфичности систем P-M типа I за счет рекомбинации между разными S-субъединицами [22–24]. Последовательности, кодирующие TRD, могут по-разному комбинироваться при хромосомных перестройках [17], так что число возможных специфичностей для системы типа I больше, чем S-белков, закодированных в бактерии. Например, рекомбинации между ДНК-узнающими доменами (TRD) 22 S-субъединиц *M. haemophylis* могут приводить к возникновению более 500 вариантов специфичностей [14]. Системы типа I *Streptococcus pneumoniae* меняют свою

специфичность с помощью специфических рекомбиназ, которые способствуют рекомбинации между участками генома, кодирующими ДНК-узнающими доменами S-субъединиц, обеспечивая этим высокое разнообразие систем даже внутри одной популяции (см. рисунок 1.3) [25,26].

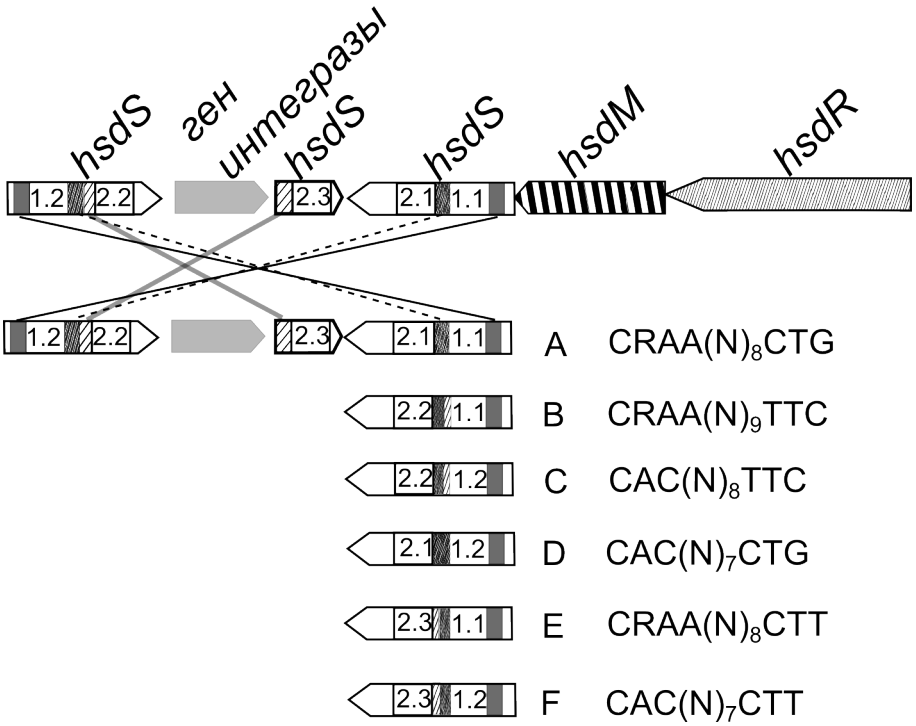


Рисунок 1.3 Организация генов системы R-M типа I *Streptococcus pneumoniae* TIGR4, специфичность которой меняется при внутригеномных рекомбинациях. Направление генов показано стрелками. В генах S-субъединицы прямоугольниками с цифрами отмечены ДНК-узнающие домены (TRD). Различные последовательности S-белка, получившиеся в результате рекомбинации, обозначены буквами, указаны сайты узнавания полученных белков [26]. Домены с различной специфичностью обозначены разными цифрами, вертикальными прямоугольниками обозначены инвертированные повторы, обеспечивающие внутрихромосомную рекомбинацию. Возможные способы рекомбинации между различными повторами показаны прямыми линиями. Рисунок взят из работы (Ершова и соавт., 2015) и изменен.

Сходные механизмы изменения специфичности системы типа I путем рекомбинации S-белка были показаны для *Bacteroides fragilis* и *Mycoplasma pulmonis* [17,27]. Кроме последовательности, может меняться расстояние между двумя частями сайта узнавания за счет изменения в числе повторов в димеризующем домене [14,28].

1.1.1.4 Расщепление ДНК

ЭР типа I представляет собой пентамерный белок, состоящий из двух R, двух M и одной S субъединицы. Этот комплекс требует АТФ, Mg^{2+} и SAM в качестве кофакторов, и способен как метилировать, так и гидролизовать ДНК. R субъединица необходима для эндонуклеазной активности. Она состоит из N-концевого эндонуклеазного домена и так называемого “моторного” домена [14].

В $R_2M_2S_1$ комплексе R-субъединица взаимодействует с M-субъединицей. Если ни одна из частей сайта узнавания не метилирована, комплекс связывает сайт узнавания, R-субъединица перемещает ДНК, пока не столкнется с препятствием, например, суперспирализованной ДНК или другим белком (см. рисунок 1.1 Г). В этот момент ЭР расщепляет обе цепи ДНК [29]. Расщепление ДНК может происходить с любой стороны от сайта узнавания [30].

Различные субстраты расщепляются с различной эффективностью. Так, линейная ДНК с единственным сайтом узнавания расщепляется только при большом избытке ЭР, который вызывает образование неспецифических комплексов с ДНК [31]. Линейная ДНК с двумя сайтами узнавания расщепляется в одном месте, примерно посередине между ними [29]. Кольцевая ДНК с одним сайтом узнавания, например, плазмидная, расщепляется в одном случайном месте. В результате образуется линейная ДНК размера, равного размеру плазмиды.

1.1.1.5 Метилирование ДНК

Метилирование сайта узнавания способны осуществлять как комплексы $R_2M_2S_1$, так и M_2S_1, используя кофактор SAM [20,32]. Обычно при этом в каждой части сайта узнавания один аденин превращается в N6-метиладенин (N6mA). Как и другие МТазы, МТаза типа I выворачивает основание из двойной спирали ДНК и модифицирует его [33]. Некоторые МТазы, например, M.EcoKI и другие МТазы Типа IA имеют наибольшее сродство к полуметилированному субстрату, а МТаза EcoAI Типа IB предпочитает неметилированный субстрат [34].

Большее сродство к полуметилированным сайтам приводит к тому, что такие

МТазы наиболее активно метилируют ДНК бактерии-хозяина после репликации, и очень медленно модифицируют чужеродную ДНК. Это свойство позволяет таким системам Р-М быть более эффективными против чужеродной ДНК [14].

1.1.1.6 Контроль активности систем рестрикции-модификации типа I

Эндонуклеазная и метилтрансферазная активность систем Р-М должна быть сбалансирована для обеспечения защиты клетки от внедрения чужеродной ДНК и, одновременно, от случайного расщепления ее собственной ДНК. Вероятность такого расщепления особенно велика, когда новая система Р-М проникает в бактериальную клетку, и вся ее ДНК полностью неметилирована по сайту узнавания данной системы. Такая полностью неметилированная ДНК является идеальным субстратом для ЭР, и медленно метилируется МТазой, которая, как правило, имеет большее сродство к полуметилированной ДНК [14].

Гидролиз собственной ДНК бактерии после интродукции системы Р-М предотвращается за счет временного промежутка между началом метилирования и началом гидролиза ДНК. Для систем типа I показано, что после интродукции в клетку новой системы Р-М метилтрансферазная активность детектируется практически сразу, а эндонуклеазная активность обнаруживается в клетках спустя достаточно продолжительное время. Например, после внедрения в клетку системы EcoKI метилтрансферазная активность обнаруживается практически сразу, а эндонуклеазная спустя примерно 15 поколений, достигая максимума через 30 поколений [12].

Существуют различные механизмы для контроля активности ЭР и МТазы, которые могут различаться у разных семейств [14]. Поддержание баланса между образованием МТазы и ЭР происходит за счет **различной активности промоторов** перед генами *hsdR* и *hsdM*, *hsdS*, а также перекрывания между генами *hsdM* и *hsdS* на одну пару оснований. В результате в клетке на одну единицу R-субъединицы синтезируется восемь частей М-субъединицы и четыре части S-субъединицы [35].

Также баланс метилтрансферазной и эндонуклеазной активности поддерживается **на посттрансляционном уровне** за счет различий в стабильности комплексов ЭР и МТазы [36] или протеолиза ЭР [37]. Например, для закодированной на плазмиде системы EcoR124I показано, что ее эндонуклеазная активность регулируется за счет различия в стабильности эндонуклеазного и метилтрансферазного комплекса [38]. Попадание плазмиды с этой системой в клетку не является летальным для клетки, эндонуклеазная активность обнаруживается через шесть поколений после начала коньюгации [13]. Исследования *in vitro* сборки белков этой системы [38] показали, что при смешивании белков S, M, и R образуются комплексы M_2S_1, $R_1M_2S_1$ и $R_2M_2S_1$. При этом только комплекс $R_2M_2S_1$ способен расщеплять ДНК. Присутствие ДНК оказывает влияние на этот процесс: в присутствии ДНК этот комлекс формировался быстрее, чем в ее отсутствие. Однако этот комплекс нестабилен, и легко распадается на $R_1M_2S_1$ и R-субъединицу с Kd $\sim 2.4\times10^{-7}$ M. Дальнейшей диссоциации комплекса не происходит, и комплекс $R_1M_2S_1$ является очень стабильным. Комплекс M_2S_1 также является очень стабильным. Сходные результаты были получены в других экспериментах, в т.ч. с использованием природных промоторов этих систем. Janscak и соавторы [38] делают вывод, что после попадания в клетку генов системы EcoR124I формируются стабильные комплекс МТазы M_2S_1 и комплекс $R_1M_2S_1$, который не способен расщеплять ДНК. Это позволяет МТазе модифицировать хозяйскую ДНК. Формирование нестабильного комплекса ЭР происходит, когда накапливается избыток R субъединицы по отношению к очень стабильным комплексам M_2S_1 и $R_1M_2S_1$, и вероятно, этого времени хватает, чтобы ДНК бактерии оказалась полностью метилированной. Эндонуклеазная активность системы EcoKI также регулируется за счет различной стабильности комплексов ЭР, МТазы и промежуточных вариантов. В случае EcoKI, в отличие от EcoR124I, комплекс $R_2M_2S_1$ является очень стабильным, а комплекс M_2S_1 легко диссоциирует на M_1S_1 и свободную M-субъединицу. Комплекс M_1S_1 связывает R-субъединицу в очень стабильный, но

неактивный комплекс $R_1M_1S_1$. Т.о., при низких концентрациях R-, M- и S-субъединиц, которые наблюдаются в клетке после интродукции системы P-M в бактериальную клетку, будут возникать неактивные комплексы $R_1M_1S_1$ и M_1S_1, а также МТаза M_2S_1, которая сможет модифицировать ДНК бактериальной клетки до того, как концентрация M- и S- субъединиц вырастет, и станет более вероятным формирование активной ЭР $R_2M_2S_1$ [39].

Активность некоторых систем типа I подавляется **за счет протеолиза ЭР** [37]. Это явление называется ослабление рестрикции (“restriction alleviation”). При этой форме регуляции активности ЭР протеаза ClpXP расщепляет R-субъединицу [37,40]. ClpXP действует на системы типа IA и IB. Система IC EcoR124I не чувствительна к действию ClpXP [13,41]. Более подробно механизм этого явления еще не исследован, в частности, непонятно, как новая система отличается от старой. В работе [42] показано, что R-субъединица системы EcoKI может быть фосфорилирована по остатку треонина. Это может играть роль в локализации и протеолизе ЭР [42]. Подавление активности ЭР в результате протеолиза также было показано после событий, повреждающих ДНК (например, облучения УФ, добавления 2-аминопурина или налидиксовой кислоты), когда необходимо предотвратить расщепление хромосомной ДНК в процессе репарации [40]. Подавление активности системы P-M проявляется в том, что после добавления 2-аминопурина к культуре бактерий, эффективность образования фаговых бляшек в ней увеличивалась на четыре порядка. Механизм ослабления рестрикции используется при гомологической рекомбинации для предотвращения рестрикции немодифицированных сайтов в бактериальной хромосоме [43].

Регуляции баланса метилирования и расщепления ДНК также может происходить **за счет различий в локализации ЭР и МТазы**. В работе [44] показано, что ЭР EcoKI может быть ассоциирована с внутренней мембраной клетки так, что МТазная часть пентамерного комплекса находится на внутренней поверхности мембраны и способна модифицировать хромосомную ДНК, а R-субъединица (или

ее часть) экспонирована на внешней поверхности цитоплазматической мембраны.

Некоторые механизмы антирестрикции бактериофагов используют смещение баланса метилирования и расщепления ДНК для преодоления рестрикционного барьера. Так, некоторые бактериофаги (например, лямбда) способны синтезировать Ral (restriction alleviation) белки, которые увеличивают сродство МТаз типа I (например, M.EcoKI) к неметилированной ДНК [45].

1.1.2 Тип II

Системы Р-М типа II состоят из МТазы и ЭР, которые узнают короткий (4-8 п.н.) сайт узнавания и расщепляют ДНК внутри или непосредственно рядом с ним. Особенности структурно-функциональной организации систем Р-М типа II показаны на рисунке 1.4.

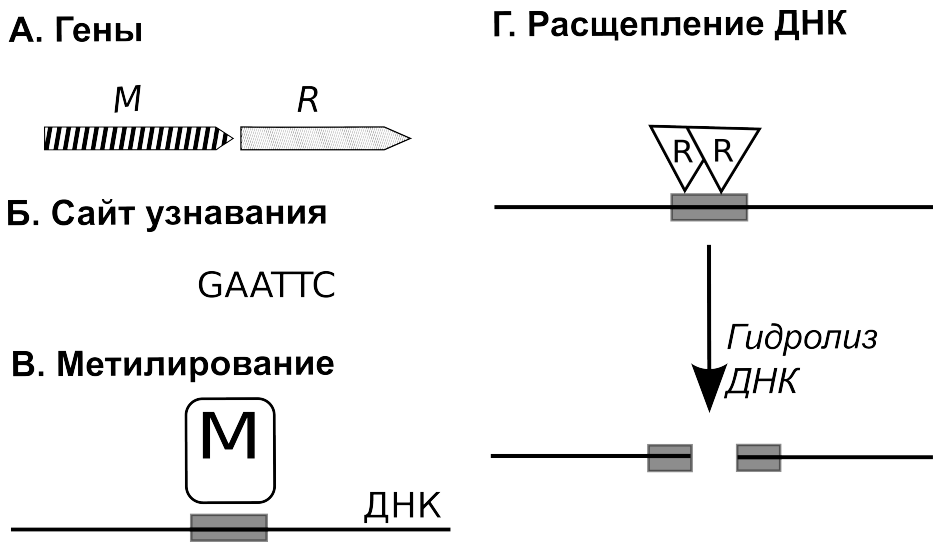


Рисунок 1.4 Характеристика типичной системы Р-М типа II. А. Схема организации генов. М – ген ДНК-метилтрансферазы; R – эндонуклеазы рестрикции. Б. Пример сайта узнавания. В. Метилирование ДНК. Г. Состав комплекса, расщепляющего ДНК. R в треугольнике – эндонуклеаза рестрикции

Системы Р-М типа II широко распространены среди прокариот [46]. Кроме прокариот системы Р-М, содержащие ЭР и МТазу [47,48], или одиночные МТазы [49] были найдены у группы вирусов эукариотических водорослей *Chlorella*. Системы Р-М типа II являются наиболее хорошо изученной группой систем Р-М, поскольку ЭР типа II являются ключевыми ферментами в генной инженерии.

1.1.2.1 Гены систем рестрикции-модификации типа II

ЭР и МТаза кодируются отдельными генами, кроме белков подтипа IIG (подтипы систем типа II описаны ниже). Эти гены колокализованы в геноме, и могут располагаться последовательно, в некоторых случаях ген ЭР закодирован первым (AccI), в других первым закодирован ген МТазы (BsuBI), также гены систем Р-М типа II могут быть ориентированы в противоположных направлениях [50]. В некоторых случаях эти два гена могут дополняться генами других белков: C (control), V (vsr reparation) и т. д. В системах типа IIG один ген кодирует белок с ЭР и МТазной активностью одновременно [51].

1.1.2.2 Классификация эндонуклеаз рестрикции типа II

ЭР типа II очень гетерогенны, объединение их в одну группу основано на их способности гидролизовать ДНК в определенной позиции внутри или рядом с сайтом узнавания [51]. В связи с разнообразием ЭР, их делят на 11 подтипов по сходству биохимических свойств [9], а не по сходству последовательностей (см. таблицу 1. 2). Подтипы не являются взаимоисключающими и один фермент может относиться к нескольким подтипам [9].

Таблица 1.2

Подтипы ЭР типа II. Таблица взята из работы [9]

Подтип	Объединяющий признак	Пример ЭР
A	Асимметричный сайт узнавания	FokI
B	Расщепляет ДНК с обеих сторон сайта узнавания	RM.BcgI
C	Эндонуклеазный и МТазный домены в одном полипептиде	GsuI
E	ЭР нуждается в двух сайтах узнавания: один является субстратом, второй - эффектором.	EcoRII
F	ЭР узнает два сайта и разрезает ДНК между ними	SfiI
G	Требует AdoMet в качестве кофактора	GsuI
H	Структура генов сходна с типом I	RM.BcgI

M	Расщепляют метилированные сайты	DpnI
P	Палиндромный сайт узнавания и расщепления	EcoRI
S	Расщепляет ДНК рядом с сайтом узнавания	FokI
T	ЭР представляет собой гетеродимер	Bpu10I

Тип IIA (assymmetric)

ЭР этого типа узнают асимметричный сайт и расщепляют ДНК внутри или на определенном расстоянии от него. Многие ЭР имеют одну или две парных МТазы, которые модифицируют только одну цепь ДНК. Некоторые ЭР слиты с МТазным доменом в один белок, при этом в некоторых случаях такой слитный белок сопровождается отдельной МТазой [51]. Обычно такие ЭР состоят из ДНК-узнающего и нуклеазного доменов. Примером такой системы является FokI. Эта система состоит из двух белков, МТазы и ЭР, узнающих асимметричный сайт GGATG/CATCC. M.FokI содержит два домена – N-концевой домен очень специфично взаимодействует с последовательностью GGATG, C-концевой домен гораздо менее специфически узнает последовательность CATCC. Оба домена метилируют адениновое основание в N6 положении [52]. ЭР FokI в растворе находится в виде мономера, но способна расщеплять ДНК только в димеризованном состоянии. Этот белок содержит два домена, разделенных гибким линкером. Один из доменов способен специфически узнавать сайт GGATG/CATCC и связывать его. Связывание изменяет конформацию белка, и каталитический домен связывается с обратной цепью ДНК на расстоянии от сайта узнавания. После димеризации с каталитическим доменом другой субъединицы ЭР, димер расщепляет ДНК в 9 п.н. от сайта узнавания по прямой цепи и 13 п.н. по обратной. Второй мономер белка, необходимый для димеризации может связываться из раствора, но ассоциация в этом случае достаточно слабая. Если оба мономера взаимодействуют ДНК, образуется более стабильный комплекс [53].

Таким образом, типичное взаимодействие ЭР с асимметричным сайтом узнавания устроено следующим образом. Сначала одна субъединица связывается

с асимметричным сайтом узнавания, затем белок на короткое время димеризуется, и расщепление ДНК осуществляется в димерном состоянии нуклеазного домена [54]. Такая кратковременная димеризация называется транзиентной (“transient dimerisation”).

Тип IIB (both)

ЭР данного типа гидролизуют ДНК с обеих сторон от сайта узнавания, при этом образуется небольшой фрагмент (около 30 п.н.), содержащий сайт узнавания. Ферменты типа IIB представляют собой белки, обладающие эндонуклеазной и метилтрансферазной активностями. Ферменты типа IIB включают метилтрансферазный и эндонуклеазный домены, сходные с последовательностями M- и R-субъединиц типа I, объединенных в одну полипептидную цепь. Узнавание сайта ферментами типа IIB осуществляется с помощью S-субъединицы, сходной с S-субъединицей типа I [55]. Некоторые ферменты типа IIB, например, BcgI, CspCI и BsaXI, имеют отдельную S-субъединицу, как системы типа I [56]. В других белках, таких как AloI, PpiI [57], S – субъединица входит в состав одной полипептидной цепи с метилтрансферазным и нуклеазным доменами. Сайт узнавания состоит из двух частей, разделенных короткой неспецифической последовательностью [58]. Как и в системах R-M типа I, ДНК-узнающий домен S-субъединицы может быстро менять специфичность путем рекомбинации [57].

МТазный домен метилирует адениновое основание в первой части сайта узнавания на прямой цепи ДНК, и адениновое основание второй части сайта на обратной цепи. Для метилирования, и, в некоторых случаях, для расщепления ДНК требуется SAM [59].

При взаимодействии с ДНК ферменты типа IIB могут не делать ничего, если сайт полностью метилирован, если сайт неметилирован или полуметилирован, эти ферменты способны расщеплять или метилировать ДНК. Если сайт метилирован наполовину, то он будет метилирован. Механизм выбора между расщеплением или метилированием в случае неметилированного сайта до сих пор остается

неясным. В отличие от ЭР типа IIP, которые расщепляют неметилированный сайт сразу, гидролизующая активность ЭР типа IIB зависит от того, сколько сайтов на ДНК являются неметилированными. Большинство ферментов типа IIB не активны как ЭР, когда связываются с одним сайтом узнавания [58,59]. Если таких сайтов много, то связанные с ДНК белки олигомеризуются, и становятся способны расщеплять ДНК [60], в противном случае нуклеазная активность подавляется, и сайт будет метилирован [61].

BcgI является белком, имеющим ЭР и МТазную активность в одном полипептиде. Активной формой этого белка является гетерогексамер, димер из тримеров 2RM+1S [55]. Эндонуклеазную активность этот гексамер осуществляет, связывая два сайта, и внося двуцепоченые разрывы в ДНК в четырех местах одновременно [61,62]. Данный процесс требует участия четырех дополнительных белков, которые могут быть находящимися по соседству гексамерами или отдельными субъединицами BcgI. Метилирование происходит при связывании гексамера с одним сайтом, при этом метилирование полуметилированного сайта происходит в 100 раз более эффективно, чем неметилированного [61].

Тип IIC (combined)

К этому типу отнесены белки, сочетающие эндонуклеазную и метилтрансферазную активность. Таким образом, этот тип полностью включает все белки, относящиеся к вышеописанному типу IIB, и ряд других ферментов.

Некоторые ферменты этого типа функционируют без парной МТазы, например, MmeI [63], другие, например Eco57I [64] имеют парную МТазу, а BpuSI [65] даже две МТазы. Большинство ферментов типа IIC связывают сайт узнавания как мономеры, поэтому для расщепления обеих цепей ДНК происходит “транзиентная димеризация” между каталитическими доменами различных молекул, подобно тому, как это было описано в случае системы FokI [54].

Эффективность гидролиза ДНК ферментами типа IIC значительно возрастает,

если субстрат содержит много сайтов узнавания. Это объясняется тем, что на такой ДНК возрастает локальная концентрация фермента и облегчается транзиентная димеризация [51]. ЭР типа IIC расщепляют ДНК на некотором расстоянии от сайта узнавания, которое может колебаться на 1-2 п.н. в зависимости от топологии ДНК, ионных условий и других физических причин [51].

Тип IIE (effector)

ЭР типа IIE содержат нуклеазный, а также специальный эффекторный домен, который связывается с еще одной копией сайта узнавания, изменяет конформацию каталитического домена и стимулирует гидролиз ДНК [66]. В отсутствие второго сайта, эффекторный домен ингибирует эндонуклеазную активность фермента. Поэтому для расщепления ДНК такие ЭР нуждаются в наличии двух сайтов узнавания. Примером таких ЭР является EcoRII. Интересно, что удаление эффекторного домена приводит к значительному увеличению активности EcoRII, которая перестает нуждаться в наличии двух сайтов для гидролиза ДНК [67].

Биологическая функция такой активации не известна, возможно, наличие такой регуляции защищает собственную ДНК клетки от расщепления случайно немодифицированного сайта [51].

Тип IIF (four DNA strand)

ЭР этого типа связывают два сайта узнавания, как и ЭР типа IIE, но, в отличие от них, ЭР типа IIF расщепляют одновременно оба сайта (четыре цепочки ДНК). Однако, расщепление возможно только если связаны оба сайта [68–70]. С ДНК эти ферменты связываются как гомотетрамеры. Сайты узнавания могут быть как симметричные, так и асимметричные [51].

Родственные ЭР могут относиться к типу IIE или типу IIF. Так, EcoRII (CCWGG), относится к типу IIE, а ЭР Ecl18kI (CCNGG) и SsoII (CCNGG) - к типу IIF. И EcoRII, и Ecl18kI взаимодействуют с ДНК как гомодимеры, и используют сходные

механизмы расщепления ДНК [51]. Структурно эти два фермента также сходны, за исключением того, что на N-конце EcoRII находится эффекторный домен, которого нет у Ecl18kI. Тем не менее, Ecl18kI более активна, если субстрат содержит несколько сайтов узнавания. Показано, что Ecl18kI осуществляет расщепление ДНК в виде временного тетрамерного комплекса [71].

Тип IIG (gamma-methyltransferase domain)

Функционирование ЭР типа IIG зависит от присутствия в клетке SAM. Ферменты состоят из нуклеазного и метилтрансферазного доменов. Эти ферменты взаимодействуют с сайтом узнавания с помощью S-субъединицы, которая может как входить в состав полипептидной цепочки, так и представлять собой отдельную субъединицу [51]. Большинство ЭР типа IIC и IIB также принадлежит к типу IIG.

SAM является донором метильной группы и необходим для метилирования ДНК. Зависимость эндонуклеазной активности от SAM защищает ДНК клетки от гидролиза в случае дефицита SAM, в результате которого хозяйская ДНК может быть недометилирована [51].

Необычными представителями этой группы являются MmeI и сходные с ним белки [63,72]. MmeI взаимодействует с асимметричным сайтом TCCRAC и модифицирует только адениновое основание на прямой цепи ДНК [72]. Таким образом, для защиты хромосомной ДНК используется модификация только одной цепи. В случае, если адениновое основание в этом сайте неметилировано, MmeI вносит двуцепочечный разрыв на расстоянии 18 п.н. от сайта. Для эффективного расщепления ДНК MmeI нуждается в двух сайтах узнавания, причем, в отличие от систем типа III, эти сайты могут быть в любой ориентации, а не только “голова-к-голове” [72]. Поэтому после репликации хозяйская полуметилированная ДНК может быть уязвима к расщеплению MmeI. Тем не менее, MmeI клонируется и экспрессируется в клетках *E.coli* [72]. Также найдено большое число ферментов, сходных с MmeI по последовательности и способу взаимодействия ДНК [63].

Каким образом происходит защита хозяйской полуметилированной ДНК от расщепления остается неясным. Возможно, она достигается за счет баланса метилтрансферазной и эндонуклеазной активности [72].

Тип IIH (hybrid)

Системы этого типа представляют собой переходную форму между системами типа I и II. Примером такой системы является AhdI [73].

Тип IIM (methylated)

Ферменты данного типа узнают метилированный сайт узнавания [51]. Более подробно они обсуждаются в разделе 1.1.4 Метил-зависимые ЭР.

Тип IIP (palindrome)

Тип IIP является наиболее широко распространенным и разнообразным из систем типа II. Ферменты этого типа узнают симметричную последовательность ДНК (палиндром), и симметрично расщепляют ДНК внутри этой последовательности, как EcoRI или, реже, на ее границе, как EcoRII. Система типа IIP, как правило, состоит из ЭР и парной МТазы с такой же специфичностью, в редких случаях система может включать две МТазы [9,51].

ЭР типа IIP может действовать как мономер, как, например, ЭР MvaI [74] или может быть гомодимером, как в случае с EcoRII, или гомотетрамером, как Cfr42I. В отличие от большинства других ЭР, в ЭР типа IIP аминокислотные остатки, ответственные за взаимодействие с сайтом узнавания и за гидролиз ДНК объединены в один домен [75].

Симметичный сайт узнавания обычно имеет длину 4-8 п.н., он может быть как вырожденным, например, GANTC, или невырожденный, например, AAGCTT. При расщеплении ДНК могут образовываться как “тупые”, так и “липкие” концы [51].

В настоящее время известны сотни палиндромных сайтов узнавания, и для каждого сайта несколько ЭР, узнающих этот сайт. Иногда эти ЭР представляют

собой родственные белки, гены которых распространяются с помощью горизонтального переноса в геномах различных бактерий и архей. При этом даже среди родственных ЭР могут наблюдаться значительные различия в биохимических свойствах [76]. Часто ЭР, узнающие похожие сайты, демонстрируют сходство аминокислотных последовательностей. Например, PstI (CTGCA|G) и SbfI (CCTGCA|GG), или BssHII (G|CGCGC) и AscI (GG|CGCGCC), вероятно, имеют общее происхождение [51].

ЭР с похожими сайтами узнавания не сходны по аминокислотной последовательности. Это может означать как исчезновение следов общего происхождения со временем, так и независимое возникновение таких ЭР [51].

Тип IIS (shifted)

ЭР типа IIS расщепляют ДНК в фиксированной позиции, сдвинутой (shifted), относительно сайта узнавания на один или два витка спирали ДНК [77]. Формально все ЭР групп IIB, IIC и IIG также являются ЭР группы IIS, однако они значительно отличаются от других ферментов типа IIS, классическим представителем этого подтипа считается описанная выше ЭР FokI [51]. В некоторых случаях ЭР типа IIS соответствуют две отдельные МТазы, каждая из которых модифицирует адениновое или цитозиновое основание на одной цепи ДНК. За узнавание и гидролиз ДНК отвечают различные домены ЭР [78].

Сайт узнавания обычно ассимметричный, гидролиз обеих цепей ДНК предполагает наличие транзиентной димеризации для гидролиза ДНК [51].

Тип IIT (two different subunits)

К этому типу первоначально были отнесены ЭР, представляющие собой гетеродимер из двух различных субъединиц. Сейчас к этой группе причисляют белки, имеющие два различных каталитических центра. Некоторые из них действительно являются гетеродимерами, например, BbvCI; Bpu10I; BtsI, BsrDI и BspD6I [79], другие являются одноцепочечными белками с двумя различными

каталитическими центрами, например, Mva1269I [80].

Ферменты взаимодействуют с асимметричным сайтом узнавания и расщепляют ДНК внутри него или поблизости. Как правило, ЭР типа IIT сопутствуют две отдельные МТазы, по одной для модификации каждой цепочки асимметричного сайта узнавания. Эти МТазы могут представлять собой отдельные белки или быть слитыми в одну полипептидную цепь [51].

1.1.2.3 Узнавание ДНК

Механизмы узнавания ДНК ЭР и парной МТазой различаются [81]. Так, например, взаимодействие с тиминовыми основаниями в сайте GAATTC существенно для M.EcoRI и менее важно для R. EcoRI [82]. Различия в механизмах узнавания ДНК связаны с отсутствием сходства последовательностей ЭР и МТазы и формы взаимодействия с сайтом узнавания – МТаза обычно представляет собой мономер, а ЭР – димер [51].

ЭР и МТаза специфично взаимодействуют с основаниями по малой и большой бороздке ДНК, зависимая от последовательности конформация остова также играет роль в специфичном взаимодействии. Специфическое взаимодействие достигается за счет формирования водородных связей между донорами и акцепторами электронов, а также гидрофобого взаимодействия с метильной группой основания тимина. Метилирование нуклеотидных оснований нарушает взаимодействие ЭР с ДНК и защищает ДНК от расщепления [51].

Изменение специфичности ЭР и МТаз типа II в большинстве случаев возможно за счет точечных мутаций, и, в случае систем Р-М типа IIG, за счет замены ДНК-узнающего домена целиком [57]. Главная сложность в эволюции специфичности систем Р-М типа II состоит в необходимости однонаправленного изменения специфичности одновременно ЭР и МТазы.

1.1.2.4 Метилирование ДНК

По типу метилирования МТазы можно разделить на белки, способные

переносить метильную группу с S-аденозилметионина (SAM) на адениновое основание в 6 позиции (N6mA), цитозиновое основание в 4 (N4mC) или пятой (5mC) позиции [83]. Среди МТаз типа II встречаются все три группы МТаз. Большинство изученных МТаз систем типа II действуют как мономеры, однако есть и примеры ферментов, активных в димерной форме [84] Как правило, наиболее предпочтительным субстратом для МТаз является полуметилированная ДНК [85].

Большинство сайтов узнавания систем Р-М типа II являются палиндромными, т.е. имеют одинаковую последовательность по обеим цепочкам ДНК, например, GATC. Некоторые МТазы, например, BamHI, модифицируют палиндромный сайт узнавания по обеим цепочкам одновременно, другие, например, EcoRI, модифицируют каждую цепь своего палиндромного сайта связывания независимо [84]. При этом, поскольку палиндромные сайты симметричны, один и тот же фермент может модифицировать свой сайт по обеим цепям.

Метилирование асимметричных сайтов узнавания, например, GAATC, устроено более сложно, поскольку МТаза на разных цепочках должна провзаимодействовать с различающимися последовательностями. В обзоре [85] выделено несколько способов метилирования асимметричных сайтов и защиты ДНК хозяйской клетки от расщепления после репликации: (i) каждая цепь асимметричного сайта модифицируется своей МТазой. Так, сайтом узнавания системы MboII является сайт 5′-GAAGA-3′/3′-CTTCT-5′. Система содержит две МТазы, одна из которых M1.MboII, модифицирует адениновое основание сайта 5′-GAAGA^{m6}-3′, другая, M2. MboII, модифицирует цитозиновое основание сайта 3′-CTTC^{m4}T-5′. Также известны случаи, когда две различные МТазы объединяются в один слитный полипептид. Так, например, МТаза M. FokI метилирует адениновые основания по обеим цепям асимметричного сайта 5′-GGATG/CATCC-3′, т.к. содержит два МТазных домена. N-концевой домен метилирует верхнюю цепь, а С-концевой – нижнюю; (ii) для расщепления ДНК нужно две неметилированные

копии сайта. Примером такой системы является система типа IIG MmeI.

1.1.2.5 Расщепление ДНК

Как правило, ЭР действуют в виде димеров, они способны узнавать сайт узнавания и гидролизовать ДНК внутри или в непосредственной близости от сайта узнавания, если он неметилирован. Известны примеры, когда ЭР взаимодействуют с ДНК как мономеры или тетрамеры [51]. Например, ЭР типа IIP MvaI взаимодействует с последовательностью CCWGG как мономер [74] и расщепляет две цепи ДНК, внося, последовательно, два одноцепочечных разрыва [86]. В отличие от нее, ЭР Eco29kI узнает и связывает свой палиндромный сайт узнавания CCGCGG как мономер, а затем димеризуется для расщепления ДНК [87].

В качестве кофактора эти белки, как правило, нуждаются в Mg2+, но многие могут использовать Mn2+ и другие двухвалентные ионы [88]. Различные подтипы ЭР характеризуются различными особенностями взаимодействия с ДНК. Эти особенности обсуждались при описании подтипов.

Эффективность расщепления ДНК для, по крайней мере, некоторых, ЭР зависит от последовательностей, окружающих сайт узнавания [89–91]. Такой эффект связан с тем, что последовательности, фланкирующие сайт узнавания, модулируют термодинамические и кинетические параметры взаимодействия между ДНК и ЭР. Так, в работе [92] показано, что EcoRI, которая узнает сайт GAATTC, расщепляет последовательность TGAATTCA в 200 раз менее эффективно, чем природную последовательность AGAATTCC бактериофага SV40. Также геномный контекст сайта узнавания влияет на конформацию ДНК [93], котрая, в свою очередь, влияет на эффективность ЭР.

Интересно, что ЭР, узнающие одну и ту же последовательность могут разительно отличаться по механизму расщепления ДНК. В работе [76] описывается пять механизмов расщепления ДНК для семи ЭР, узнающих один и тот же сайт GGCGCC. NarI связывает два сайта и вносит один одноцепочечный разрыв за

одно связывание с ДНК. KasI связывает один сайт, и вносит один одноцепочечный разрыв. Mly113I требует два сайта для эффективного расщепления, и вносит двуцепочечный разрыв в оба сайта. SfoI, EgeI и EheI связываются с одним сайтом и вносят двуцепочечный разрыв, действуя в соответствии с представлениями о том, как должны действовать типичные ЭР типа II. Наконец, BbeI связывается с двумя сайтами, которые должны располагаться на небольшом расстоянии друг от друга, и расщепляет оба сайта.

1.1.2.6 Контроль активности

Известно несколько уровней контроля баланса метилирования и гидролиза ДНК в системах типа II.

Метилирование промотора. Промоторные области некоторых систем R-M содержат соответствующий сайт узнавания. Для некоторых систем показана зависимость уровня экспрессии гена МТазы от статуса метилирования их сайта в промоторной области. Примером систем с такой регуляцией является CfrBI. Эта система закодирована двумя генами ЭР и МТазы, расположенными дивергентно. В перекрывающейся промоторной области находится один сайт узнавания системы CfrBI. Если этот сайт неметилирован, экспрессия гена МТазы выше, чем экспрессия гена соответствующей ЭР. Метилирование сайта узнавания приводит к снижению экспрессии гена M.CfrBI и увеличению экспрессии гена CfrBI [94,95]. Зависимость экспрессии генов от метилирования промотора была описана для систем LlaDII [96], FokI [97] и некоторых других.

Транскрипционные факторы. Некоторые системы, кроме ЭР и МТазы, содержат дополнительный С-белок (C-control), который является транскрипционным фактором, способным связываться с промоторной областью, регулируя, таким образом, экспрессию генов соответствующей системы. Примером системы R-M с такой регуляцией является PvuII [98]. Гены ЭР и МТазы этой системы ориентированы дивергентно. Перед геном ЭР расположен ориентированный в ту же сторону ген С-белка. Перекрывающаяся промотрная

область оперона *pvuIICR* и *pvuIIM* содержит два сайта связывания С-белка. При этом сродство С-белка к дальнему от гена *pvuIIC* сайту значительно выше, чем к ближнему. Исходный уровень экспрессии гена *pvuIIM* выше, чем *pvuIIC* и *pvuIIR*. Появление С-белка и его связывание с дальним от гена *pvuIIC* сайтом приводит к увеличению экспресси генов оперона pvuIICR за счет связывания С-белка с σ-субъединицей РНК-полимеразы. Дальнейшее накопление С-белка приводит к тому, что он связывается со вторым своим сайтом в промоторной области, и ингибирует экспрессию *pvuIICR* оперона за счет того, что механически мешает продвижению РНК-полимеразы.

В некоторых случаях регуляторный ДНК связывающий домен входит в состав МТазы, как, например, в системе SsoII [99], EcoRII [100], MspI [101]. Регуляция экспрессии генов в системе SsoII происходит следующим образом. Гены этой системы расположено дивергентно, промоторные области этих генов перекрываются. Экспрессия гена *ssoIIM* выше, чем гена *ssoIIR*. Синтезированный белок M.SsoII связывается с промоторной областью, что приводит к снижению уровня экспрессии гена *ssoIIM* и повышению уровня экспрессии гена *ssoIIR* [99].

Регуляция с помощью антисмысловой РНК показана для систем EcoRI и Eco29I. В этих системах гены ЭР и МТазы организованы в один оперон, где ген ЭР предшествует гену МТазы. Этот оперон содержит несколько промоторов, поэтому с него синтезируется несколько мРНК. Перед опероном находится промоторная область, которая регулирует синтез двухцистронной мРНК, содержащей последовательности ЭР и МТазы. Ген МТазы может также экспрессироваться отдельно, с промоторов, находящихся внутри гена ЭР [102].

Для оперона системы EcoRI экспериментально было определено шесть промоторов (включая две тандемные пары, которые могут рассматриватьсыя как два сложных промотора), обладающих различной силой PREV0 > PREV1,2 >=PM1,2 > PR [103,104]. Каждый из смысловых промоторов (PR, PM1,2) имеет соответствующий антисмысловой промотор (PREV1,2 и PREV0 соответственно),

локализованный внутри гена ЭР. С антисмысловых промоторов транскрибируется антисмысловая РНК [103]. Экспериментально показано, что транскрипция с промотора PREV0 уменьшает транскрипцию с промотора PM1,2 и наоборот. Транскрипция с промотора PREV0 также уменьшает транскрипцию с промотора PR [104]. Вероятный механизм этого ингибирования экспрессии состоит в РНК-интерференции антисмысловых РНК с мРНК, транскрибированных с генов системы Р-М.

Для системы Eco29kI также описана подобная схема регуляции экспрессии гена ЭР. Антисмысловая РНК, транскрибирующаяся с обратного промотора, расположенного в гене ЭР, комплементарно связывается с мРНК, которая кодирует ЭР, и приводит к ее деградации [105].

Регуляция на уровне сборки комплекса. МТаза, как правило, действует в виде мономера, а ЭР в качестве димера. Кроме того, большинство ЭР, узнающих асимметричный сайт, нуждаются в нескольких сайтах для гидролиза ДНК [61]. Некоторые из ЭР, узнающих палиндромы (например, EcoRII) также нуждаются в наличии двух сайтов для расщепления ДНК [67]. Такой механизм предотвращает расщепление случайно незаметилированного сайта в ДНК бактерии [51].

1.1.3 Тип III

1.1.3.1 Гены систем рестрикции-модификации типа III

Системы Р-М типа III состоят из двух рядом расположенных генов *mod* и *res*, кодирующих белки, способные узнавать последовательность ДНК и модифицировать ее (Mod) или гидролизовать (Res) [20,106,107]. Общая характеристика систем Р-М типа III показана на рисунке 1.5. В настоящее время экспериментально охарактеризованы 10 систем типа III. Гены *res* и *mod* субъединиц находятся под контролем независимых промоторов [108]. Примерами систем типа III являются системы EcoPI и HinfIII. Системы Р-М типа III найдены в большинстве секвенированных бактерий. Последовательности этих белков в различных микроорганизмах отличаются высоким сходством [109].

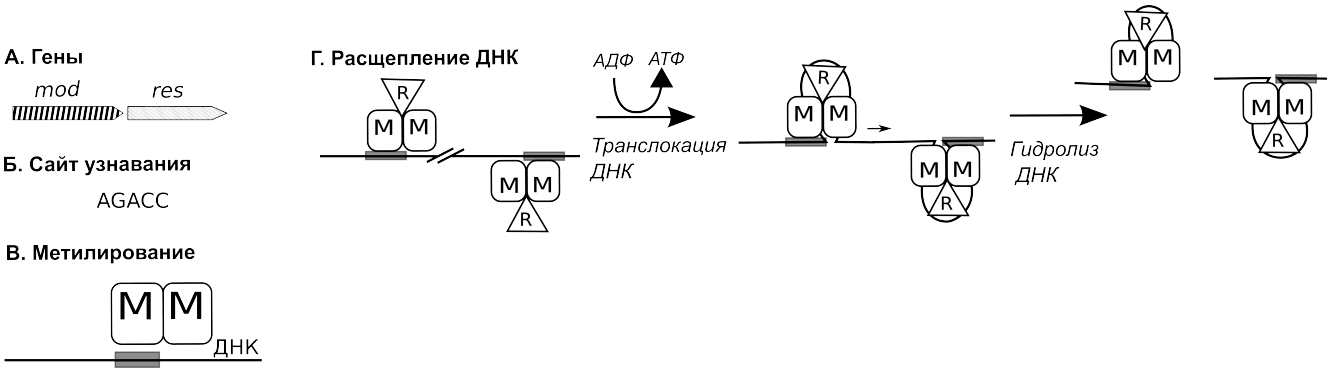


Рисунок 1.5 Характеристика типичной системы Р-М типа III. А. Схема организации генов. *mod* – ген ДНК-метилтрансферазы; *res* – эндонуклеазы рестрикции. Б. Пример сайта узнавания. В. Состав комплекса для метилирования ДНК. Г. Состав комплекса, расщепляющего ДНК. Черными стрелками показано направление транслокации ДНК.

1.1.3.2 Узнавание ДНК

Взаимодействие с ДНК осуществляется Mod субъединицей и в МТазном, и в ЭР комплексе. Системы Р-М типа III асимметричный сайт узнавания длиной 5-6 п.н [109].

1.1.3.3 Расщепление ДНК

Для гидролиза ДНК требуется комплекс Res_1Mod_2 [110,111] и наличие двух противоположно ориентированных копий сайта узнавания и АТФ в качестве кофактора [107]. При этом один комплекс Res_1Mod_2 связывает только один сайт [110]. Вопрос о том, как происходит коммуникация между ЭР, связанными с различными сайтами остается открытым: в некоторых работах были найдены петли ДНК, свидетельствующие о ее транслокации [112]. Для более детального понимания механизмов расщепления ДНК ЭР этого типа необходимы дальнейшие исследования. Гидролиз ДНК осуществляется при АТФ-зависимой транслокации ДНК на расстоянии примерно 25-27 п.н. от одного из сайтов узнавания. Образующиеся фрагменты имеют липкие концы, длина одноцепочечной части 2-3 н.о. [113,114]. Mod субъединица не только осуществляет специфическое взаимодействие с ДНК, но и способствует экспонированию Res субъединицы в сайте, который должен быть гидролизован [113,114]. Для гидролиза ДНК ферментами типа III необходимо наличие двух сайтов узнавания в противоположной ориентации [113,114].

1.1.3.4 Метилирование ДНК

Для метилирования ДНК достаточно только Mod субъединицы, которая во всех известных случаях модифицирует аденин до N6mA внутри сайта узнавания. Активная МТаза типа III представляет собой гомодимер Mod_2 [115–117]. Поскольку сайт системы типа III асимметричен, то он модифицируется только по одной цепи. Неполное метилирование не приводит к повреждению хозяйской хромосомы после репликации, поскольку для гидролиза ДНК ЭР типа III требует двух неметилированных сайтов в противоположной ориентации “голова-к-голове”, а после репликации все неметилированные сайты бактериальной ДНК находятся на одной цепи, и, следовательно, в одинаковой ориентации [109].

1.1.3.5 Контроль активности

В работе Arber и соавт. [118] было показано, что метилтрансферазная активность системы EcoP1I наблюдается почти сразу после попадания этой системы в клетку, а активность ЭР появляется спустя примерно три часа после этого. Различие во времени обнаружения эндонуклеазной и метилтрансферазной активностей позволяет предположить существование механизма регуляции активности ЭР для защиты ДНК хозяина. В обзорной статье [109] обсуждается, что, по-видимому, существует регуляция экспрессии ЭР на нескольких уровнях. На уровне транскрипции регуляция экспрессии гена *res* происходит за счет синтеза антисмысловой РНК с промотора в середине *res* гена [108]. В работе [119] показано, что экспрессия Mod белка положительно регулирует количество Res белка. Также в этой работе было показано, что взаимодействие с субъединицей Mod необходимо для правильной укладки субъединицы Res и предотвращения ее протеолиза.

1.1.4 Метил-зависимые системы (Тип IV и IIM)

В отличие от остальных систем Р-М, ферменты типа IV гидролизуют модифицированную ДНК, и, соответственно, содержат только ЭР. Белки McrA и McrB (modified cytosine restriction) способны узнавать модифицированные

основания цитозина, ферменты Mrr (modified DNA rejection and restriction) распознают как метилцитозин, так и метиладенин. ЭР типа IV отличаются низкой специфичностью, что позволяет им защищать клетку от широкого спектра чужеродной ДНК с различными паттернами метилирования [120]. Слабая специфичность систем типа IV объясняется тем, что бактериальная клетка не содержит гидроксиметилированной ДНК, в то время как геном бактериофага может быть полностью гидрокисметилирован [120]. В этих условиях отбор на специфичность ЭР слабый.

Выделяют две группы метил-зависимых ЭР: Тип IV и Тип IIM. Это разделение связано с процессом открытия этих ферментов [120], не найдено каких-либо фундаментальных свойств, разделяющих эти две группы ферментов. Поэтому некоторые авторы предлагают объединить ферменты типа IIM и типа IV в одну группу [120].

1.1.4.1 Гены метил-зависимых систем рестрикции-модификации

Ферменты этого типа кодируются одним или двумя генами. Большинство из них предсказано методами биоинформатики и не изучено экспериментально [120].

Метил-зависимые ЭР очень различны между собой по последовательности и особенностям взаимодействия с ДНК. Вероятно, такое разнообразие связано с тем, что эти белки несколько раз независимо возникали в процессе коэволюции бактерий и бактериофагов [120].

Интересны некоторые примеры, иллюстрирующие плавный переход от обычных ЭР к метил-зависимым. ЭР MspJI была отнесена к ЭР типа II, поскольку ее ген находился рядом с геном цитозиновой МТазы. Экспериментальная проверка показала, однако, что МТаза неактивна, а MspJI является метил-зависимой ЭР [121]. С другой стороны, белок, предсказанный как McrBC, при экспериментальной проверке оказался ЭР типа II LlaI [122].

ЭР типа II BamHI гидролизует сайт GGATCC, если предпоследнее цитозиновое

основание не модифицировано. При этом эта ЭР предпочитает гидролизовать последовательности, содержащие N6mA, и можно получить мутантные белки, которые способны взаимодействовать только с сайтами, содержащими N6mA [123]. Такая ЭР может быть отнесена и к типу II, и к типу IV. Подобные ферменты показывают возможные пути происхождения систем типа IV [120].

1.1.4.2 Классификация ферментов

Экспериментально изученные ферменты типов IV и IIM немногочисленны и очень разнообразны по последовательностям и биохимическим свойствам. Даже два фермента “McrA” EcoKMcrA и ScoA3McrA сходны по последовательности только нуклеазного С-концевого домена [120].

1.1.4.3 Узнавание ДНК

Модификации ДНК, которые защищают ДНК от гидролиза ЭР типов I-III, включают метилированный цитозин и аденин (5mC, N4mC, N6mA), а также гидроксиметилированный цитозин (5hmC) и глюкозилированный гидроксиметилцитозин (5ghmC). У прокариот нет известных ферментов, которые бы модифицировали цитозин в 5hmC, 5ghmC в сайт-специфичной манере. Таким образом обычно неспецифично модифицирована ДНК бактериофагов.

Для каждого типа модификации ДНК существуют ферменты, способные гидролизовать такую модифицированную ДНК. Например, EcoKMcrA [124], SauUSI [125], относящиеся к типу IV способны гидролизовать ДНК, содержащую 5-метилцитозин или 5-гидроксиметилцитозин.

Как правило, метил-зависимые ЭР осуществляют узнавание ДНК за счет отдельного домена или отдельной субъединицы белка, расщепление ДНК осуществляется другим доменом или субъединицей. Так, например, ЭР EcoKMcrA содержит два домена, N-концевой ДНК-связывающий домен (DBD), узнающий метилированное или гидроксиметилированное основание цитозина в последоваательности YCGR (Y=C или T; R=A или G) и С-концевой нуклеазный

домен HNH [126]. EcoKMcrBC содержит две субъединицы, McrB узнает ДНК, McrC содержит нуклеазный домен и способна расщеплять ДНК [127].

ЭР DpnI типа IIM узнает и гидролизует четырехбуквенный сайт Gm6ATC , когда обе цепи ДНК метилированы [128,129]. Высокая специфичность отличает ее от других метил-зависимых ЭР, которые низкоспецифичны. DpnI содержит ДНК-связывающий и нуклеазный домены. Узнавание ДНК осуществляется мономером, который связывается с ДНК по большой бороздке и узнает модифицированное адениновое основание на обеих цепях [130].

1.1.4.4 Расщепление ДНК

Как уже было отмечено, большинство метил-зависимых ЭР имеют низкую специфичность [121,131]. В некоторых случаях для гидролиза ДНК требуется два сайта. Механизмы взаимодействия с ДНК и расщепления очень разнообразны.

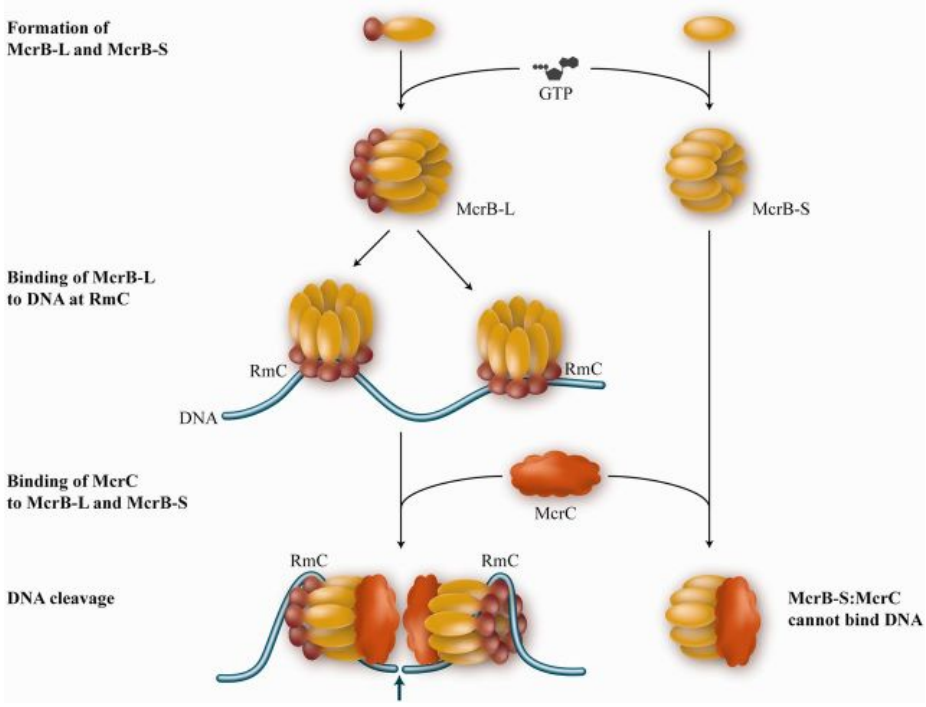


Рисунок 1.6 Модель комплекса McrBC. С гена mcrB экспрессируется два белка: полноразмерный McrB-L и более короткий белок McrB-S, у которого отсутствует N-концевой домен. McrB-S, как и McrB-L, способен связываться с McrC, но, в отличие от него, не способен взаимодействовать с ДНК. Гомодимеры McrB связывают ГТФ и формируют гептамерное кольцо с центральным каналом. Мультимеры связываются с метилированной ДНК за счет N-концевых доменов белка McrB-L.Для расщепления ДНК необходимо формирование комплекса с McrC, который обладает нуклеазной активностью. Рисунок взят из обзора Loenen и Raleigh [120].

Так, например, ЭР EcoKMcrBC узнает сайт Rm5C, где R=A или G, гидролиз ДНК осуществляется на расстоянии 30-35 п.н. от сайта, для гидролиза ДНК нужно, чтобы два таких сайта находились друг от друга на расстоянии 30-3000 п.н. [120]. McrBC взаимодействует с ДНК в виде мультимерного комплекса (см. рисунок 1.6), который осуществляет транслокацию ДНК, расходуя ГТФ. Расщепление ДНК происходит либо при взаимодействии двух таких комплексов либо при возникновении препятствия для транслокации. Гидролиз ДНК также требует ГТФ [19].

В отличие от McrBC, MspJI узнает сайт mCGNR (R=A или G), модифицированный только по одной цепи и расщепляет ДНК на расстоянии 12 п.н. по одной цепи и 16-17 по другой [131]. Активность этой ЭР возрастает при наличии второго сайта в цис- или транс- положении. MspJI взаимодействует с ДНК как тетрамер [132].

1.1.4.5 Контроль активности систем типа IV

Многие модификации ДНК, к которым специфичны эти ЭР, отсутствуют у прокариот (например, гидроксиметилирование). Поскольку некоторые метил-зависимые ЭР чувствительны к метилированной ДНК, они могут быть несовместимы с некоторыми МТазами [133]. Интересно, что несмотря на отсутствие токсичности для ДНК бактерии, существуют механизмы, предотвращающие случайный гидролиз ДНК. Например, многие ЭР типа IV для гидролиза требуют наличия двух модифицированных сайтов в цис- или транс-положении [120]. Механизм действия ЭР McrBC [19] также похож на регуляцию активности на уровне сборки комплекса, как это описано, например, для систем типа I.

1.2 Организация генов систем рестрикции-модификации в геноме и их мобильность

Подавляющее большинство систем Р-М кодируется генами, расположенными неподалеку друг от друга [1,50,134]. Колокализация способствует тому, что

системы Р-М могут распространяться между прокариотами путем горизонтального переноса на мобильных генетических элементах, таких как плазмиды [135,136], профаги [136,137], инсерционные последовательности (IS-элементы) и транспозоны [46,138], интегративные коньюгативные элементы (ICE) [46,139] и интегроны [46,140,141]. Таким образом, мобильные элементы позволяют системам Р-М распространяться между прокариотами, а системы Р-М стабилизируют мобильные элементы в геноме [140,142,143].

Такой симбиоз подтверждается данными работы Koonin и соавт. [144], где для 1055 прокариотических геномов показано, что гены систем Р-М и других систем токсин-антитоксин (но не CRISPR-Cas систем) часто встречаются вместе с типичными компонентами мобильных элементов (гены вирусов и транспозонов). В работе [46] анализ 2261 прокариотических геномов также выявил корреляцию между присутствием генов систем Р-М и мобильных генетических элементов, интегронов, CRISPR-Cas систем, а также со способностью бактерий к природной трансформации.

Надо отметить, что стабилизации мобильных элементов в геноме способствуют только системы типа II благодаря токсичности ЭР без присутствия МТазы. Остальные типы систем Р-М такими свойствами не обладают и не способны стабилизировать мобильные элементы в геноме, тем не менее, гены систем других типов тоже часто колокализованы с мобильными элементами [46].

Наличие различных иммунных систем в одной бактериальной клетке увеличивает ее устойчивость к атакам фагов. Экспериментально показано, что система CRISPR-Cas и система Р-М могут действовать в бактериальной клетке одновременно и независимо, при этом метилирование фага не оказывает влияние на действие системы CRISPR-Cas [145]. Однако в работе Koonin с соавт. [144] показано, что гены защитных систем не только одновременно присутствуют в геноме, но и колокализованы (“защитные острова”). Остается неясным, отражает ли колокализация генов различных защитных систем функциональную связь

между ними, или она связана с их приобретением с помощью горизонтального переноса и фиксацией в геноме за счет токсичности систем токсин-антитоксин в случае потери соответствующих генов [144].

1.3 Одиночные гены систем рестрикции-модификации

Помимо МТаз, входящих в состав систем Р-М, в геномах прокариот часто встречаются одиночные МТазы. В работе [146] было обнаружено, что 79% генов МТаз, найденных в геномах бактерий не ассоциированы с другими генами систем Р-М, т.е. являются одиночными. Одиночные МТазы могут быть гомологичны МТазам из полных систем Р-М. Это позволяет предположить, что такие одиночные МТазы могут быть результатом деградации полных систем Р-М [46,146].

Потеря токсичного компонента системы токсин-антитоксин должна приводить к полной элиминации бесполезного антитоксина из генома бактерии. Поэтому большое количество одиночных генов МТаз позволяет предположить, что они выполняют какие-то полезные функции, кроме того, что являются компонентами систем Р-М.

Одиночные МТазы могут защищать клетку при внедрении других систем Р-М с близкой специфичностью [147]. Процесс внедрения новой системы Р-М в бактериальную клетку представляет собой большую опасность для клетки, поскольку вся хозяйская ДНК должна быть метилирована прежде, чем активируется ЭР. В работе Takahashi и соавт. [147] показано, что одиночная МТаза Dcm, метилирующая последовательность CCWGG (W=A или T) может защитить хозяйскую ДНК от действия ЭР EcoRII.

Одиночные гены систем Р-М могут быть результатом того, что колокализованные гены в геноме оказались разнесены на большие, чем обычно расстояния внутри генома. Так, в системе Р-М типа I из *Staphylococcus aureus* колокализованные гены М- и S-субъединиц находятся на значительном расстоянии от гена R-субъединицы,

при этом система сохраняет функциональную активность [15,148]. В геномах *Lactococcus lactis* и *Mycoplasma pneumonia* [149,150] ген S-субъединицы локализован на плазмиде, а колокализованные гены R- и M- субъединиц на хромосоме.

Гены одиночных МТаз, найденные в бактериальных геномах, могут быть локализованы в профагах [46]. Известно, что в геномах фагов также содержатся одиночные гены мульти- и моноспецифичных МТаз для защиты фагов от бактериальных ЭР и регуляции своего собственного жизненного цикла [151].

Некоторые одиночные МТазы, выполняют важные функции, не связанные с действием систем Р-М. Например, Dam и CcrM вовлечены в процессы клеточной регуляции, репликации и репарации ДНК [5,152–154]. Одиночные МТазы, которые имеют регуляторный эффект у бактерий, обычно консервативны внутри таксономической группы, в отличие от модифицирующих ферментов, входящих в состав систем Р-М, которые распространены случайно [146].

1.4 Функции систем рестрикции-модификации в клетке

1.4.1 Защита от бактериофагов

Системы Р-М были впервые обнаружены и охарактеризованы по их способности защищать бактериальную клетку от внедрения чужеродной ДНК, в том числе, фаговой ДНК [155–157]. Защитная функция систем Р-М основана на их способности взаимодействовать с сайтами узнавания и различать их метилированное и неметилированное состояние. При наличии системы Р-М в бактериальной клетке собственная ДНК оказывается метилирована и защищена от гидролиза, а проникшая в клетку чужеродная ДНК оказывается неметилированной и подвергается гидролизу ЭР. Из-за этой способности различать свою и чужую ДНК систему Р-М часто рассматривают как иммунную систему прокариот [158,159]. Различные исследования демонстрируют от 10 до 10^8 –кратный уровень защиты хозяйских клеток от фагов для различных систем Р-

M [10]. Также защитную роль систем P-M подтверждает факт, что многие фаги избегают действия систем P-M, в т.ч. с помощью различных модификаций своего генома (метилирование, гликозилирование и другие модификации нуклеотидов) [157]. Для защиты бактерий от таких фагов в процессе коэволюции бактерий и бактериофагов появились системы P-M, которые наоборот, способны расщеплять только метилированную ДНК [120].

Однако, защитная функция систем P-M достаточно ограничена, и рано или поздно бактериофаги преодолевают защитный барьер систем P-M. После одного цикла размножения вируса в бактериальной клетке, содержащей систему P-M, ДНК потомков бактериофага модифицируется соответствующей МТазой, и все потомки этого фага становятся нечувствительными к данной системе P-M. В работе [160] обсуждается, что основная функция систем P-M может заключаться в том, что они позволяют снизить численность популяции бактерий, необходимую для внедерения бактерий в новые места обитания, наполненные фагами.

Кроме того, даже внутри одной популяции бактерии могут быть гетерогенны по набору систем P-M [161,25,162]. Такие популяции могут быть более устойчивыми к атакам бактериофагов, поскольку даже преодолев защиту систем P-M одной части популяции, бактериофаг остается восприимчив к системам P-M другой части популяции (см. рисунок 1.7) [162].

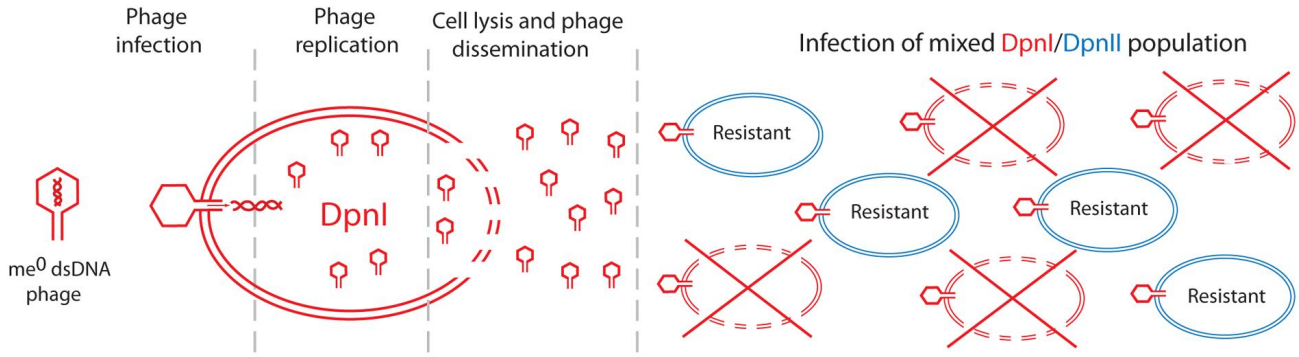


Рисунок 1.7 Преимущество наличия различных систем в смешанной популяции *Streptococcus pneumoniae* при атаке бактериофагов. Система типа IIM DpnI расщепляет метилированную последовательность GATC и система P-M DpnII метилирует последовательность GATC и расщепляет неметилированную последовательность GATC. Неметилированный двуцепочечный ДНК бактериофаг (me^0 dsDNA phage, показан красным шестиугольником) поражает смешанную

популяцию бактерий, часть из которых несет систему P-M DpnI (показаны красными овалами), а часть – систему DpnII (показаны синими овалами). При этом фаг поражает бактерии, несущие систему DpnI, а клетки, несущие систему DpnII, выживают благодаря тому, что расщепляют неметилированную ДНК фага. В случае заражения смешанной популяции фагом с метилированной ДНК будет наблюдаться обратная картина – клетки несущие систему DpnII окажутся уязвимыми к инфекции, а клетки, несущие систему DpnI, выживут благодаря ее способности расщеплять метилированную ДНК (не показано на рисунке). Рисунок взят из работы Johnston и соавт. [163].

Дальнейшие исследования систем P-M показали, что их биологическая роль не ограничивается защитой бактерий от бактериофагов. В частности, некоторые патогенные бактерии могут секретировать ЭР наружу. В результате системы P-M могут непосредственно влиять на взаимодействие бактерий с эукариотическими хозяевами, разрушая ДНК эукариотической клетки. Например, для *Neisseria gonorrhoeae* показано, что в процессе проникновения бактерии в эукариотическую клетку происходит повышение экспрессии генов ЭР и выпуск ЭР в эукариотическую клетку, где они проникают в ядро и повреждают хромосомную ДНК [164]. Для систем P-M вирусов хлорелл *Paramecium bursaria chlorella virus* (PBCV-1) также показано, что ЭР вируса способствует деградации ядерной ДНК хозяина [165].

1.4.2 Влияние метилирования генома на регуляцию экспрессии генов

Метилирование генома как одиночными, так и входящими в состав систем P-M МТазами может оказывать влияние на экспрессию генов, и, в результате, на регуляцию клеточного цикла, вирулентность и другие фенотипические проявления бактерий [5,166].

Хорошо изучены регуляторные функции описанных выше одиночных МТаз типа II Dam и CcrM, которые метилируют последовательности GATC и GANTC соответственно [167,168]. Роль систем P-M в регуляции вирулентности и экспрессии генов в настоящее время интенсивно изучается благодаря появлению доступных технологий секвенирования метилома, которые позволяют изучить, какие МТазы являются активными в геноме. Эти данные существенно изменяют

представления о функционировании систем P-M в геноме, сложившиеся при исследовании P-M систем, клонированных в экспрессионных мультикопийных векторах [169].

Так, в работе [170] для *Helicobacter pylori* показано, что нокаут одного из генов S-субъединицы системы P-M типа I (HPP12_0797) ведет к потере метилирования сайта GAA(N_8)TAG, что, в свою очередь, приводит к изменениям в транскриптоме за счет изменений в уровне экспрессии генов, которые содержали этот сайт.

Метилирование ДНК различными системами P-M способно влиять на патогенность бактерий. Так, для *Streptococcus pneumoniae* еще в 1933 г. Webster and Clow показали переключение между фенотипическими формами (“фазами”), которые вызывают бессимптомное носительство или инвазивное заболевание (например, пневмонию). В 2014 г. Manso и соавторы [26] показали, что механизм такого переключения основан на рекомбинации между геном S-субъединицы системы P-M типа I SpnD39III и двумя фрагментами генов S-субъединицы. Перестройки приводят к появлению одной из шести альтернативных специфичностей (маркированных от A до F). В результате геномы с различными вариантами SpnD39III различаются по паттернам метилирования, профилям экспрессии генов и вирулентности. В частности, оказалось, что B-аллель S-субъединицы ассоциирован с пониженным уровнем экспрессии оперона, кодирующего гены белков капсулы. Уменьшение экспрессии этих генов ведет к формированию прозрачных колоний. Такие штаммы не вызывали симптомов болезни у мышей, и были способны колонизировать носоглотку. Штаммы, несущие A-аллель S-субъединицы формировали капсулу и образовывали непрозрачные колонии. Эти штаммы вызывали инвазивную инфекцию, и не были способны колонизировать носоглотку. Широкое распространение системы SpnD39III среди пневмококков свидетельствует о важности этого механизма регуляции экспрессии генов.

В работе [171] для штамма *Campylobacter jejuni* SA (IA3902), вызывающего

аборты у овец, и геномов фенотипически отличающихся бактерий, обитающих в кишечном тракте, *Campylobacter jejuni* NCTC 11168 и 81-176 было показано, что в этих сходных по набору МТаз геномах уровень метилирования и расположение гипо- и гиперметилированных областей значительно различались. Различия затрагивали гены вирулентности, что может объяснять различную патогенность этих штаммов.

МТазы могут содержать гомополимерные последовательности, что приводит к фазовой вариации. Фазовая вариация заключается в том, что экспрессия белков изменяется в результате рекомбинации, а не в результате точечных мутаций. В работе [6] показано, что некоторые гены *H. pylori* содержат сдвиги рамки считывания, часто вызванные гомополимерными нуклеотидными повторами. Такие повторы могут меняться в длине из-за ошибочного спаривания оснований и это способствует фазовой вариации. Среди генов, способных к фазовой вариации, авторы выделяют несколько генов МТаз, которые из-за сдвигов рамки являются неактивными. Однако после коррекции сдвига рамки некоторые из ферментов, закодированные этими генами (4 из 7) оказались активными.

Также впервые были найдены системы P-M типа IIG, способные к изменению специфичности из-за влияния таких гомополимерных повторов. Каждая из двух гомологичных систем типа IIG JHP1272 (J99-R3) и HP1353-HP1354 (26695) содержала два гомополимерных повтора, которые приводили к преждевременному обрыву цепи. Коррекция обоих повторов приводила к появлению активного белка, узнающего последовательность GGWCN (JHP1272) и CRTCN (HP1353-HP1354). Коррекция только первого повтора приводила к изменению специфичности белка на GGWTA и CRTTA соответственно. Таким образом, один из повторов полностью выключал экспрессию гена, в то время как второй вызывал изменение специфичности. Наблюдались мутантные изогенные штаммы с различной активностью этих МТаз, что подтверждает возможность фазовой вариации этих генов в природных популяциях.

В работе [172]проводили исследование влияния фазовой вариации в генах МТаз *Helicobacter pylori* на способность этой бактерии к колонизации хозяина. В геноме штамма *H. pylori* OND79 было найдено пять генов, кодирующих МТазы или другие белки систем P-M, содержащих гомополимерные повторы, и, следовательно, способных к фазовой вариации. Из-за ошибок полимеразы на гомополимерном участке, экспрессия таких генов не всегда приводит к синтезу одного и того же функционального белка. Для исследования их роли в колонизации хозяина были использованы мутанты двух типов: мутант с делецией МТазы и мутант в состоянии ‘ON’, где гомодимерный повтор был заменен на неповторяющуюся синонимичную последовательность так, чтобы продуцировался функциональный полноразмерный белок. Для обоих типов мутантов была исследована их способность к колонизации на мышиной модели. Полученные данные позволили разделить эти гены на три категории: 1) не оказывающие эффекта на способность к колонизации 2) те, где экспрессия полноразмерного белка является вредной для колонизации 3) те, для которых как отсутствие, так и экспрессия полноразмерного белка являются вредными. Т.о., авторы показали, что фазово-вариабельные МТазы являются необходимыми для успешной колонизации *H. pylori*, что показывает важность метилирования генома и эпигенетического разнообразия для колонизации и патогенеза. Существование генов третьей категории показывает, что для колонизации хозяина может быть необходима именно дифференциальное метилирование геномов в популяции, которое способна обеспечить фазовая вариация.

Таким образом, в результате действия систем P-M происходят изменения паттернов метилирования генома, что приводит к изменению уровня экспрессии генов и влияет на различные аспекты жизни бактерии.

1.5 Системы рестрикции-модификации как эгоистичный элемент генома

Гены ЭР и МТазы, как правило, расположены неподалеку друг от друга, и

способны распространяться между геномами путем горизонтального переноса [1,46,136]. В работе I. Kobayashi [1] системы Р-М рассматриваются как эгоистичные элементы генома. В рамках этой концепции основной функцией систем Р-М является распространение своих собственных генов. Эта идея подтверждается наличием такого явления как гибель клеток, потерявших плазмиду, несущую гены соответствующей системы Р-М (“postsegregational killing”) [1,173]. Возможное объяснение этого феномена состоит в том, что ЭР более стабильна, чем МТаза, поэтому клетки, потерявшие плазмиду, кодирующую гены системы Р-М, быстро лишаются МТазы, а затем их ДНК подвергается гидролизу оставшейся ЭР, что приводит к гибели клеток.

Другое объяснение гибели клеток после потери генов системы Р-М состоит в том, что даже если время жизни ЭР и МТазы одинаково, то с каждым делением концентрация ЭР и МТазы в клетке будет снижаться, а риск повреждения бактериальной ДНК будет расти, т.к. чтобы избежать действия ЭР должны быть метилированы все сайты, а для ЭР достаточно одного незаметилированного сайта [174].

Системы Р-М типа II имеют сходство с системами токсин-антитоксин. Во-первых, они содержат стабильный токсин и лабильный антитоксин; во-вторых, они ассоциированы с мобильными элементами; в-третьих, могут демонстрировать эгоистичное поведение [136,175].

Модель системы Р-М как эгоистичного элемента генома позволяет объяснить, почему бактериальные клетки не теряют плазмиду, кодирующую систему Р-М. Эта модель хорошо объясняет данные, полученные для систем Р-М типа II: Bsp6I [135], EcoRI [176], EcoRII и SsoII [177], EcoRV [178], PaeR7I [143], PvuII [173]. Однако, такая модель взаимодействия системы Р-М с геномом бактерии-хозяиа не применима к системам типов I и III [179]. Это связано с тем, что в этих системах ЭР представляет собой комплекс, включающий МТазную субъединицу. В результате соотношение ЭР и МТазы в клетке не изменяется при потере локуса,

содержащего гены системы Р-М, и потеря таких систем не может привести к гибели клетки.

Способность систем Р-М к распространению с помощью горизонтального переноса позволяет рассматривать их как независимую систему генов, способную возникать и теряться в прокариотическом геноме. О высокой скорости процессов возникновения/утраты генов систем Р-М свидетельствует то, что различные штаммы одного и того же вида могут различаться по набору систем Р-М [25,46,161]. Таким образом, системы Р-М могут быть рассмотрены как внешний фактор, влияющий на эволюцию геномов прокариот.

1.6 Влияние систем рестрикции-модификации на эволюцию геномов прокариот

Способность Р-М системы к гидролизу ДНК может приводить к изменению частот сайтов систем Р-М в геномах [2–4]; возрастанию частоты внутрихромосомных перестроек [180], которые лежат в основе внутривидовой эволюции; ограничению горизонтального переноса между штаммами с различным набором систем Р-М [181–183]; стабилизации мобильных элементов в геноме [183].

1.6.1 Изменение олигонуклеотидного состава генома

Впервые снижение числа сайтов систем Р-М было экспериментально обнаружено в геномах бактериофагов. При выращивании в бактериях, содержащих систему Р-М EcoB, в геноме фага fd наблюдалась потеря обоих сайтов данной системы [184,185].

Статистический анализ частот сайтов узнавания систем Р-М в геномах бактериофагов показал, что снижение количества сайтов узнавания систем Р-М в геноме является распространенной стратегией антирестрикции для двуцепочечных ДНК-содержащих фагов [10]. Например, двуцепочечный ДНК Фаг Ф1 (phi) *Bacillus subtilis* несет много меньшее число сайтов определенных ЭР, чем

можно было бы ожидать статистически. Так, последовательность GGCC, узнаваемая ЭР BsuRI, которая должна была бы встретиться в геноме фага 400 раз по статистической оценке, не встретилась в нем ни разу [185]. В работе [186] была показана недопредставленность сайтов узнавания систем P-M в геномах колифагов на основе сравнения наблюдаемых частот всех известных сайтов систем P-M с ожидаемыми, рассчитанными на основании частот тринуклеотидов. При этом все фаги *E.coli*, кроме фага лямбда и G4, показали снижение числа сайтов систем P-M типа II энтеробактерий, а фаги *Bacillus* – снижение числа сайтов систем P-M *Bacillus*. В отличие от двуцепочечных ДНК-содержащих фагов, РНК-содержащий фаг MS2, эукариотические фаги, такие как вирус Эпштейна-Барр, аденовирус, вирус папилломы человека, SV40 и три митохондриальных генома не избегают каких-либо сайтов систем P-M [186].

Такое избегание сайтов систем P-M хозяев вместе с отсутствием избегания у вирусов, не подверженных действию систем P-M, позволяет объяснить недопредставленность сайтов систем P-M именно действием отбора против этих сайтов, как сайтов систем P-M [186,187]. При этом сайт может и не избегаться в геноме бактериофага, даже если бактерия-хозяин кодирует соответствующую систему P-M [4].

С появлением последовательностей геномов прокариот было обнаружено, что в геномах бактерий и архей сайты систем P-M также встречаются значительно реже, чем статистически ожидается [2–4,7,8,188,189]. Избегание сайтов систем P-M коррелирует с присутствием в геноме бактерий и архей соответствующих систем P-M [3,4,188]. Недопредставленность сайтов узнавания систем P-M в геномах прокариот связывают с возможностью спонтанного гидролиза хозйской ДНК клетки в случае ошибок метилирования ДНК [3,7]. Предположение о токсичности систем P-M для кодирующих их бактерий подтвердилось экспериментально [[190]].

Как показано в работе [3], сайты систем P-M, относящихся к данному виду

бактерий, находятся среди наиболее недопредставленных палиндромов в соответствующих геномах, и, более того, сайты систем Р-М, закодированных в данном геноме являются более избегаемыми, чем остальные палиндромы [4].

Избегание палиндромов не коррелирует с таксономическим сходством и наблюдается не для всех видов бактерий [4]. Возможно, это вызвано тем, что мутации, снижающие число сайтов, могут приводить к нарушениям каких-то клеточных процессов, и, тем самым, снижать жизнеспособность бактерий [183].

Данные о недопредставленности сайтов систем Р-М в геномах прокариот были позже подтверждены в работе [188], где были идентифицированы наиболее пере- и недопредставленные слова в ДНК четырех бактерий *Escherichia coli*, *Bacillus subtilis*, *Clostridium perfringens* и *Pseudomonas aeruginosa*. Авторы показали, что палиндромы являются значительно более недопредставленными, чем сходные с ними непалиндромные слова. Для трех из четырех видов бактерий авторы нашли слабую корреляцию между недопредставленностью палиндромов и частотой использования кодонов. Также Fuglsang и соавторы делают вывод о влиянии систем Р-М на недопредставленность своих сайтов в геноме, т.к. сайты систем Р-М, как правило, сильно недопредставлены по сравнению с другими палиндромами.

Интересно, что степень недопредставленности палиндромов и сайтов систем Р-М у бактериофагов значительно меньше, чем у бактерий [4]. Эти результаты свидетельствуют о том, что давление отбора на сайты систем Р-М и палиндромы больше для геномов бактерий, чем для фаговых геномов. Поэтому авторы предположили, что это свидетельствует о том, что системы Р-М скорее паразитируют на бактериях, чем защищают их от фагов.

Seshasayee и соавторами [146] была исследована недопредставленность сайтов узнавания МТаз в примерно 1000 геномов прокариот, и показано, что палиндромные сайты узнавания МТаз избегаются. При этом сайты недавно приобретенных систем Р-М избегаются реже. По мнению авторов, это может быть

связано с тем, что в таком случае отбор не имел достаточно времени/поколений для изменения олигонуклеотидного состава генома. Наиболее сильное избегание сайтов наблюдается для МТаз, входящих в состав систем Р-М, которые не являются продуктом недавнего горизонтального переноса. Однако для результатов этих авторов характерна большая статистическая погрешность, поэтому их выводы нуждаются в дополнительной проверке.

Полученные разными авторами с использованием различных методов данные свидетельствуют о связи недопредставленности коротких палиндромных последовательностей в геномах прокариот с присутствием систем Р-М.

Большинство работ по исследованию недопредставленности сайтов узнавания систем Р-М в геномах прокариот были выполнены на ограниченной выборке геномов бактерий и/ или бактериофагов, что позволило обнаружить недопредставленность сайтов узнавания систем Р-М, но не оценить количественно распространенность этого явления и влияющit на него факторы.

Практически все работы, за исключением [186] исследовали избегание сайтов только систем Р-М типа II, сайты которых представляют собой короткие (четырех-шестибуквенные), часто палиндромные последовательности. В работе [186] были исследованы сайты систем Р-М типа II и других типов (I и III), и показано, что только сайты систем Р-М типа II были недопредставлены в геномах бактериофагов.

Таким образом, имеющиеся данные свидетельствуют о влиянии систем Р-М на изменение олигонуклеотидного состава прокариотического генома.

1.6.2 Влияние на перестройки генома

Благодаря важности систем Р-М для бактерии, а также способности некоторых систем Р-М к расщеплению генома хозяина при утрате соответствующей системы [173], области, содержащие гены системы Р-М, реже подвергаются гомологической рекомбинации, элиминирующей гены системы Р-М, что приводит

к большей стабильности областей ДНК, соседствующих с генами систем Р-М [180,191]. С другой стороны, системы Р-М могут вызывать рекомбинации и перестройки генома за счет внесения двухцепочечных разрывов в ДНК [180,192].

ЭР гидролизует чужеродную ДНК на фрагменты, которые затем расщепляются экзонуклеазами или могут быть субстратом для белков, обеспечивающих рекомбинацию. Некоторые исследования показали, что фрагменты чужеродной ДНК, образующиеся при действии ЭР, могут стимулировать гомологическую рекомбинацию с хозяйским геномом [43,193]. Такая стимуляция гомологичной рекомбинации может служить двум целям: а) восстановлению хозяйской ДНК после случайного гидролиза и б) росту генетического разнообразия [194]. Кроме гомологической рекомбинации, фрагменты чужеродной ДНК могут быть встроены в хозяйский геном и при помощи негомологичной рекомбинации [195].

1.6.3 Влияние на горизонтальный перенос генов и поддержание гетерогенности популяции

Как обсуждалось выше, внутри одной популяции бактерии могут быть гетерогенны по набору систем Р-М [27,8,197]. В работе [46] был построен пан-геном для 43 видов бактерий. Среди систем Р-М только 4% были найдены в составе консервативной части пан-генома. Большинство систем Р-М любого типа относятся к дополнительным генам, которые встречаются не во всех геномах данного вида. При этом около 80% семейств генов представлены менее, чем в 1/3 штаммов.

Прокариоты способны приобретать [136,176] и терять системы Р-М [136,146] как на уровне генов системы целиком, так и на уровне изменения сайта узнавания системы Р-М, который является их важнейшей характеристикой. Эти изменения в наборе систем Р-М приводят к изменению паттернов метилирования генома бактерии, что в свою очередь, влияет на ее фенотипические свойства и позволяет лучше адаптироваться к меняющимся условиям окружающей среды [196].

Различия в наборе систем Р-М затрудняют обмен ДНК между популяциями бактерий [161,197]. Это создает предпосылки для накопления и других различий и способствует эволюции вида. Горизонтальный перенос генов у прокариот может осуществляться тремя путями: перенос генов в составе бактериофагов (трансдукция), поглощение клеткой ДНК из внешней среды (трансформация) и перенос генов с помощью плазмид или мобильных генетических элементов (коньюгация) [198]. При природной трансформации и коньюгации в цитоплазму клетки-реципиента переносится только одноцепочечная ДНК [199]. Хотя ЭР обычно не способны расщеплять одноцепочечную ДНК, они могут расщеплять участок генома, получившийся после встраивания перенесенного фрагмента и достройки комплементарной цепи [199].

Для *Neisseria meningitidis* показано, что популяция структурирована в филогенетические клады, ассоциированные с набором систем Р-М [161]. При этом наблюдается примерно равное число случаев переноса ДНК внутри и между кладами. Однако длина фрагментов ДНК, переносящихся между кладами значительно меньше, чем длина фрагментов, переносящихся между геномами внутри клады (в среднем 0,68 т.п.н. в сравнении с 3,68 т.п.н.). Это хорошо объясняется присутствием кладо-специфичных наборов систем Р-М.

Roer и соавторы показали, что эффективность переноса плазмиды, содержащей два сайта узнавания системы типа I EcoKI, в штамм *E.coli*, несущий эту систему, составляет 15% от эффективности переноса плазмиды в штамм, где система EcoKI была инактивирована [200].

Таким образом, системы Р-М препятствуют горизонтальному переносу генов, но не блокируют его полностью. Ограничение горизонтального переноса генов способствует сохранению генетического разнообразия внутри популяции.

1.6.4 Взаимодействие между различными системами рестрикции-модификации

В литературе довольно мало данных о взаимном влиянии систем Р-М. Известно, что присутствие одних систем Р-М в геноме может препятствовать появлению других систем в данном геноме [133]. Например, ЭР типа IV McrBC расщепляет метилированную последовательность $R_{me}C(N)40\text{-}2000R_{me}C$ (где R – A или G, N – A, T, G или C, meC – метилированное цитозиновое основание). В работе [201] показано, что наличие этой системы в геноме *E. coli* препятствует трансформации плазмиды, которая кодирует систему Р-М типа II PvuII. По-видимому, это происходит из-за того, что ЭР McrBC расщепляет геномную ДНК после метилирования МТазой M.PvuII ее сайта узнавания CAGCTG.

Различные системы Р-М могут взаимодействовать друг с другом, если их сайты узнавания перекрываются. Например, ЭР типа III для расщепления ДНК нуждаются в двух противоположно ориентированных неметилированных сайтах узнавания [107]. В работе [202] для двух систем Р-М типа III EcoPI (AGACC) и EcoP15I (CAGCAG) показано, что, действительно, каждая из этих ЭР не способна расщеплять ДНК, которая содержит по одному сайту каждой системы, которые находятся в противоположной ориентации друг к другу. При этом присутствие обоих ферментов приводит к гидролизу такого субстрата.

Взаимодействие между системами Р-М может происходить за счет изменений в экспрессии других генов. В работе [203] показано, что в штаммах *dam*- бактерии *E. coli*, не кодирующих МТазу Dam, наблюдается 100-кратное уменьшение как эндонуклеазной, так и метилтрансферазной активности системы Р-М типа I EcoK. Активность других систем типов I и III (EcoB, EcoD, и EcoP1) также снижалась. В отличие от них, активность системы типа II EcoRI не менялась. Авторы интерпретируют такое снижение активности систем Р-М в *dam*- штаммах, как последствие индукции в таких штаммах каких-то белков, которые ослабляют действие этих систем Р-М.

Таким образом, взаимное влияние систем P-M друг на друга может изменять их влияние на бактерии, однако известно всего несколько примеров такого взаимодействия. Возможно, накопление данных секвенирования геномов и анализа метиломов позволит лучше изучить этот вопрос.

1.7 Методы сравнительной геномики

Большинство систем P-M в полных геномах прокариот были предсказаны методами сравнительной геномики [204]. Также эти методы используются при сравнении как различных систем P-M, так и геномов бактерий и позволяют получить информацию о эволюции систем P-M [136,144] и коэволюции систем P-M и прокариот [4,146].

В этом разделе описаны концепции и подходы, применяющиеся в сравнительной геномике для изучения эволюции генов, которые были использованы в данной работе.

1.7.1 Сходство генов: гомологи, ортологи и паралоги

Для большинства генов можно найти сходные с ними гены в других организмах [205]. Такие сходные гены, имеющее общее происхождение, называют гомологами [206]. Гомологи подразделяются на несколько групп, в зависимости от типа эволюционных связей между ними.

Ортологами называются гены, которые произошли от одного предкового гена последнего общего предка сравниваемых видов.

Паралоги - гены, которые образуются при дупликации одного гена внутри генома.

Ксенологи - гомологичные гены, которые не являются ортологами, т.к. по крайней мере один из них появился в геноме путем горизонтального переноса [207].

Как только сравнение генов из различных геномов стало практически важной

задачей, возникли вопросы о том как определить “одинаковые” гены в геномах разных видов, т.е. как найти ортологичные гены.

Классическая схема идентификации ортологов включает филогенетический анализ [208], при этом топология филогенетического дерева, построенного на основании последовательностей данного гена сравнивается с филогенетическим деревом, построенным для видов, содержащих данный ген, и дерево для гена приводится к дереву видов в соответствии с принципом парсимонии. Принцип парсимонии состоит в описании эволюции данного гена минимальным числом элементарных эволюционных событий дупликации и потери гена. Предполагается, что полученное в результате дерево будет содержать ортологичные гены. Однако на практике этот подход сталкивается с некоторыми препятствиями. Одним из таких препятствий при анализе геномов прокариот является существование горизонтального переноса, который приводит к тому, что один и тот же вид дерева для двух разных генов может отражать совершенно различную эволюционную историю [206].

Важным предположением, которое лежит в основе большинства попыток филогенетической классификации ортологов и паралогов является гипотеза о том, что последовательности ортологичных генов (белков) более сходны друг с другом, чем с какими-либо другими генами (белками) из сравниваемых геномов. Т.о. они формируют “наилучшие совпадения при двунаправленном сравнении” (“symmetrical best hits”, “best bidirectional hits”) [206]. Обратное также верно — лучшие находки с большой вероятностью сформированы ортологами. Это позволяет использовать детекцию “наилучших совпадений при двунаправленном сравнении” как простой и надежный метод поиска возможных ортологов [209]. Данный метод лучше всего подходит для поиска ортологов в родственных геномах, но также хорошо работает для поиска ортологов конкретного гена в геномах эволюционно далеких организмов [206].

Поиск гомологов данного белка в различных геномах может быть осуществлен

как на основе полной последовательности данного белка, поданной на вход программе BLAST [210], так и с помощью профиля, построенного на основе скрытых Марковских моделей, позволяющих искать гомологов, учитывая наиболее консервативные части последовательности, если их возможно выделить [211].

1.7.2 Аннотация систем рестрикции-модификации в БД REBASE

Благодаря большому объему опубликованных и неопубликованных экспериментальных данных БД REBASE(REstriction dataBASE) [212] является важным ресурсом для аннотации и изучения систем Р-М в прокариотических геномах. На март 2015 г. REBASE содержала информацию о системах Р-М, закодированных в 8116 геномах бактерий и 226 геномах архей.

Поиск потенциальных систем Р-М во вновь секвенированных геномах микроорганизмов в REBASE осуществляется следующим образом [213]. В каждой последовательности ищутся гены, похожие на уже представленные в REBASE гены систем Р-М. Главным индикатором системы Р-М в геноме являются гены, содержащие консервативные мотивы, характерные для ДНК-метилтрансфераз, располагающиеся в определенном порядке на определенном расстоянии друг от друга [214]. Если новая последовательность имеет достаточно высокое сходство с уже охарактеризованными последовательностями генов систем Р-М, то становится возможным предсказать ее специфичность [213]. Предсказания REBASE подтверждаются при независимой проверке биоинформатическими [46] и экспериментальными методами [170,215].

Большинство систем Р-М, аннотированных сейчас в REBASE являются предсказанными по сходству последовательности, и их активность не изучена экспериментально [212]. Как показали работы [170,215] наличие генов системы Р-М в геноме не означает их функциональной активности в данном геноме. Распространение технологий секвенирования, позволяющих определить метилированные основания (Single molecule real time sequencing (SMRT))

[169] позволяет получить более достоверные данные об активности систем P-M в геноме.

1.7.3 Оценка частот олигонуклеотидов в геномах

Геномы различных бактерий значительно различаются по GC-составу [216], в то же время, гены внутри генома сходны по олигонуклеотидному составу [217]. При этом олигонуклеотидный состав геномов близких видов более сходен, чем более далеких [218]. В результате гены, перенесенные в данный геном из другого генома с помощью горизонтального переноса будут иметь олигонуклеотидный состав донорского генома и отличаться по своему олигонуклеотидному составу от генома-реципиента. Такие недавно приобретенные гены, находясь в геноме-реципиенте будут подвержены тому же мутационному процессу, что и остальной геном, а их олигонуклеотидный состав будет приближаться к олигонуклеотидному составу остального генома. Существование этого процесса “исправления” олигонуклеотидного состава (“amelioration”) было показано для большой группы генов кишечных бактерий [218]. Описанный процесс может быть использован для оценки количества времени, необходимого для того, чтобы интродуцированный фрагмент ДНК приблизился по составу к остальному геному.

Таким образом, существенно отличающиеся по олигонуклеотидному составу области прокариотического генома, вероятно, были недавно приобретены от неродственного организма [218]. Поэтому различия в олигонуклеотидном составе используют для поиска недавно приобретенных в процессе эволюции фрагментов генома.

Кроме того, сравнение наблюдаемой частоты встречаемости коротких олигонуклеотидных последовательностей в геноме со статистически ожидаемой позволяет оценить давление отбора на эту последовательность.

Существует несколько методов для оценки ожидаемой частоты коротких олигонуклеотидов в геноме. Наиболее простые основаны на учете частот моно-

или динуклеотидов в геноме. Более сложные методы позволяют учесть частоты более длинных “подслов” данного олигонуклеотидного “слова”. Поскольку именно эти методы будут использованы для оценки влияния систем Р-М на избегание их сайтов в геномах прокариот, рассмотрим их более подробно.

1.7.3.1 Марковская модель максимального порядка

Эта модель была описана в работе Schbath и соавторов [219].

Обозначим длину слова через *m*. Обозначим наблюдаемое число слова $W=w_1...w_m$ как *N(W)*. В соответствии с Марковской моделью максимального порядка *m*–2, ожидаемое число слов *W* в последовательности *S* будет:

$$K(w_1..w_m)=\frac{N(w_1..w_{m-1})\times N(w_2..w_m)}{N(w_2..w_{m-1})} \quad (1)$$

Контраст для слова *W* измеряет нормализованную разницу между наблюдаемым и ожидаемым числом слов *W*, и вычисляется как:

$$C(W)=\frac{T(W)}{\sigma} \quad (2)$$

где σ - стандартное отклоненение разности;

$$T(W)=\frac{N(W)-K(W)}{\sqrt{L}} \quad (3)$$

где *L* – длина последовательности *S*.

Для случая максимального порядка получим:

$$\sigma^2=\frac{K(W)}{L}\times\left(1-\frac{N(w_1..w_{m-1})}{N(w_2..w_{m-1})}\right)\times\left(1-\frac{N(w_2..w_m)}{N(w_2..w_{m-1})}\right) \quad (4)$$

1.7.3.2 Метод Карлина

В работе [8] предложен способ вычисления ожидаемого числа слова, учитывая все его подслова, включая разрывные.

Для тетрануклеотидов мера относительной частоты (τ) вычислялась как:

$$\tau(XYZW)=\frac{f(XYZW)\times f(XY)\times f(XNZ)\times f(XNNW)\times f(YZ)\times f(YNW)\times f(ZW)}{f(XYZ)\times f(XYNW)\times f(XNZW)\times f(YZW)\times f(X)\times f(Y)\times f(Z)\times f(W)}$$

(5)

где N – любой нуклеотид, f(XYZW) – средняя по обеим цепям частота тетрануклеотида XYZW и.т.д.

В работе [8] относительная распространенность характеризуется как экстремальная, для недопредставленности – если $\tau \leq 0.78$, перепредставленности – если $\tau \geq 1.23$

В работе [220] было проведено сравнение точности методов, предложенных S. Karlin и S. Schbath для четырех-, пяти- и шестибуквенных олигонуклеотидов в геноме *Escherichia coli*. Оказалось, что ожидаемые значения, предсказанные методом S.Karlin [8], лучше коррелировали с наблюдаемыми, чем ожидаемые значения, предсказанные с использованием Марковской модели максимального порядка [219] для всех исследованных длин олигонуклеотидов. Вероятно, большая точность метода, преложенного S. Karlin связана с учетом разрывных подслов [220].

1.8 Заключение

Литературные данные о структурно-функциональных особенностях различных систем Р-М и влиянии систем Р-М на эволюцию и экологию прокариот позволяют сделать следующие выводы.

- Системы Р-М различных типов значительно различаются по функциональной организации и регуляции метилирования и гидролиза ДНК, и таким образом, по скорости эволюции и своему влиянию на различные аспекты жизни прокариот.
- Гены систем Р-М в геноме, как правило, колокализованы.
- Палиндромные сайты узнавания систем Р-М типа II недопредставлены в геномах прокариот, что объясняется отрицательным отбором на эти сайты

из-за возможности случайного гидролиза хозяйской ДНК соответствующей эндонуклеазой рестрикции. Не все сайты систем Р-М типа II являются недопредставленными.

- Недопредставленность сайтов систем других типов в геномах прокариот не изучена.

Проведенный анализ литературы показывает, что исследование влияния систем Р-М различных типов на недопредставленность своих сайтов в геномах прокариот является актуальной биоинформатической задачей.

Глава 2. Материалы и методы.

2.1 Последовательности геномов и системы рестрикции-модификации

Данная работа сделана на трех различных списках полных геномов прокариот, поскольку за время выполнения работы количество известных полных геномов прокариот значительно выросло [212].

Полные геномы прокариот и их аннотации были взяты из БД NCBI, National Center for Biotechnology Information.

Информация о генах систем P-M, закодированных в полных геномах, и, в том числе, о сайтах узнавания соответствующих белков, была получена из БД REBASE [212].

Под геномом понимается набор всех последовательностей хромосом и плазмид данного организма. Были использованы только полные последовательности геномов.

Список прокариотических геномов 1 включает последовательности 1040 полных прокариотических геномов, доступных на февраль 2010 [221]. Закодированные в них системы P-M и одиночные ЭР и МТазы были получены из БД REBASE [222]. Эти геномы были использованы для поиска рассредоточенных систем P-M в разделе 3.1. Список геномов приведен в статье (Ershova, 2012).

Список прокариотических геномов 2 включает 1980 геномов бактерий и 134 генома археи (которые принадлежат 1213 видам 628 родов) с аннотированными в них системами P-M из БД REBASE [223]. Последовательности геномов (хромосом и плазмид) были взяты из БД NCBI [224]. Список проанализированных последовательностей приведен в работе (Rusinov, 2015). В списке отмечено присутствует ли в геноме хотя бы одна система P-M. Согласно данным БД REBASE, 1859 геномов бактерий и 133 генома архей кодируют хотя бы одну

систему P-M и 121 геном бактерии и один геном археи не кодируют известных систем P-M. Этот список геномов был использован для анализа недопредставленности сайтов систем P-M в геномах в разделе 3.2.

Для изучения влияния токсичности систем P-M на недопредставленность их сайтов в геномах прокариот была исследована недопредставленность сайтов узнавания эндонуклеаз рестрикции систем P-M в геномах, содержащих гены соответствующих систем по данным REBASE. Такие пары сайт-геном были названы назвали актуальными. Набор проанализированных 3449 бактериальных и 116 архейных ***актуальных пар*** геном-сайт приведен в работе (Rusinov, 2015).

Поскольку не все системы P-M, предсказанные в геноме, показывают функциональную активность [215], и предсказанная специфичность системы P-M может отличаться от реальной, была проанализирована недопредставленность сайтов узнавания систем P-M, чья активность была экспериментально показана. Для этой цели были отобраны системы P-M, которые входят в список REBASE Gold Standard [212], и определена недопредставленность их сайтов в геномах, кодирующих соответствующие системы, а также была оценена недопредставленность в геномах, сайтов систем P-M, которые были определены непосредственно методом Pacific Bio [6,169,170,225]. Такой набор пар сайт-геном был назван набором ***экспериментально подтвержденных пар***. Соответствующие пары приведены в работе (Rusinov, 2015).

Для сравнения влияния систем P-M на недопредставленность своих сайтов в геноме с влиянием других свойств последовательностей, соответствующих сайтам узнавания, в каждом прокариотическом геноме была оценена недопредставленность всех известных сайтов узнавания систем P-M. Этот набор пар сайт-геном был назван ***прокариотическим контролем***. Хотя среди пар прокариотического контроля содержится некоторое количество актуальных пар, их фракция довольно мала, и может быть оценена как примерно 1% пар, и не может оказывать заметного влияния на результат.

В качестве отрицательного контроля была оценена недопредставленность сайтов узнавания систем Р-М в геномах эукариотических вирусов, которые не встречаются с действием систем Р-М в течение своей жизни. Такой набор пар сайт-геном был назван ***вирусным контролем.*** Геномы эукариотических вирусов были взяты из БД NCBI [224], **список геномов эукариотических вирусов** приведен в работе (Rusinov, 2015). Известно несколько эукариотических вирусов (вирусы *Chlorella*, *Marseilleviridae* и *Phaeocystis globosa*), которые кодируют системы Р-М или одиночные МТазы. В них также была проанализирована недопредставленность соответствующих сайтов.

Список прокариотических геномов 3 был использован для исследования недопредставленности последовательности GATC, описанного в разделе 3.2.8 Главы 3. н включает последовательности 2316 геномов, кодирующих белки системы Р-М или одиночные ЭР или МТазы, узнающие последовательность GATC. Список последовательностей геномов и соответствующих GATC-специфичных белков приведен в работе (Ershova, 2016).

2.2 Анализ состава систем рестрикции-модификации

Гены систем Р-М, аннотированные как псевдогены (“pseudo”) в записи NCBI, были отмечены как “поврежденные”.

Системы Р-М, включающие гены ЭР и МТазы, ни один из которых не был аннотирован как “поврежденный”, были обозначены как “полные” системы Р-М.

Все остальные системы Р-М были отмечены как “неполные”.

2.3 Поиск генов ДНК-метилтрансфераз

Для всех неполных системы Р-М, содержащих ЭР, ген которой не был отмечен как “поврежденный” был предпринят поиск открытых рамок считывания, сходных с последовательностями МТаз. Поиск открытых рамок считывания, находящихся на расстоянии до 4 т.п.н от начала и конца гена ЭР (старт- и стоп-кодона,

соответственно), осуществляли программой getorf пакета EMBOSS. Для поиска последовательностей, сходных с последовательностями МТазных доменов использовали программу tblastn [16] с порогом E-value < 0.01.

В результате этой процедуры не было найдено ни одного нового гена предполагаемой МТазы по сравнению с уже аннотированными REBASE.

Если рядом с геном ЭР (на расстоянии 4 т.п.н.) не было найдено неповрежденного гена соответствующей МТазы, то такая ЭР была отмечена как “одиночная”.

Гены одиночных ЭР были дополнительно проанализированы на сходство с другими ЭР. Если при сравнении последовательности одиночной ЭР с последовательностями ее гомологов (% идентичности >20%) из полных систем Р-М, длина аминокислотной последовательности одиночной ЭР была более, чем на 20% короче любого гомолога, такая ЭР была маркирована как “возможный фрагмент”. Все проанализированные одиночные ЭР представлены в работе (Ershova, 2012).

2.4 Поиск ортологичных белков

Для поиска ортологов был использован метод “наилучших совпадений при двунаправленном сравнении” (“best bidirectional hits”) [209], описанный в разделе 1.7.1. Хотя этот метод не требует введения порогов на сходство последовательностей, они были использованы, чтобы найти ближайших ортологов. Две ЭР из различных геномов считались ортологичными, если их аминокислотные последовательности имели больше 40% идентичности при длине выравнивания больше 80% длины более длинной последовательности. Для МТаз ортологичными считались белки, имеющие больше 50% идентичности при длине выравнивания больше, чем 80% длины более длинной последовательности. Согласно этим критериям изученные белки, как правило, имели не более одного ортолога в геноме.

2.5 Поиск ортологичных систем рестрикции-модификации

Две пары ЭР и МТаз, закодированные в двух различных геномах, названы ортологичными, если их ЭР и МТазы являются ортологичными (см. рисунок 2.1). Пары ЭР-МТаза, которые не были ранее аннотированы в REBASE как система Р-М, но ортологичные таким парам в других геномах, были рассмотрены как потенциальные новые системы Р-М.

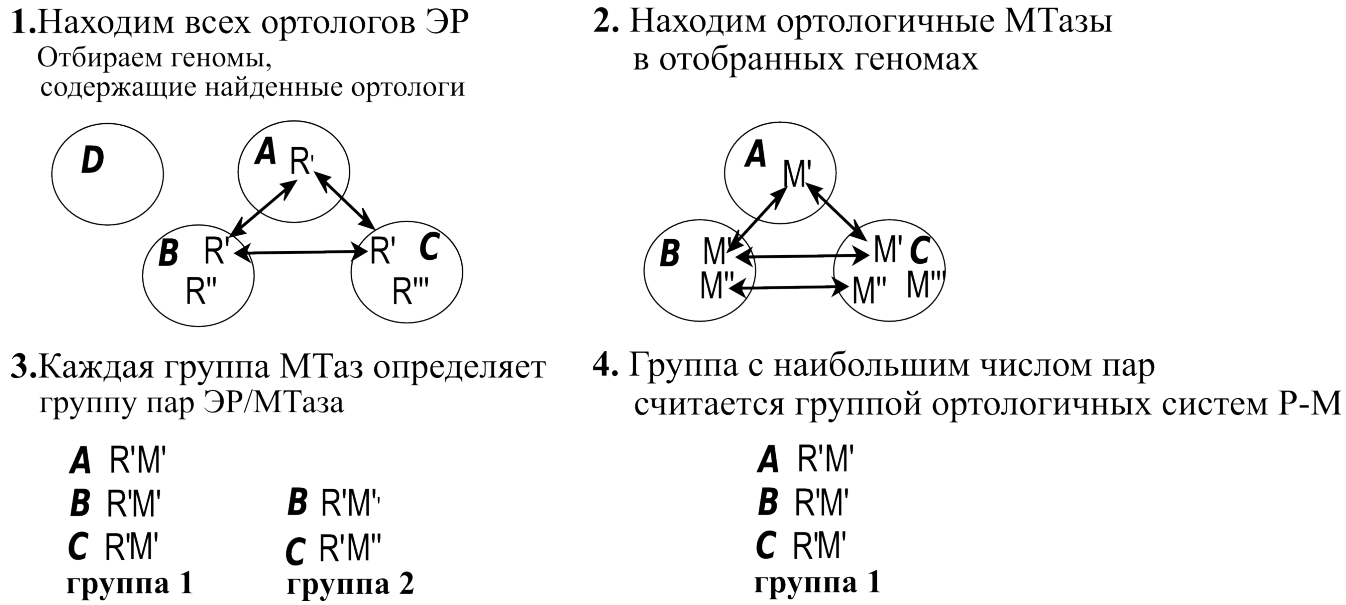


Рисунок 2.1 Алгоритм поиска ортологичных систем Р-М. Геномы обозначены большими кругами, стрелки соединяют ортологичные гены.

В случае, если гены ЭР и предположительно парной МТазы были найдены на большом расстоянии друг от друга в геноме (> 4 т.п.н.), такая система Р-М была названа рассредоточенной.

Список найденных групп ортологичных систем, включающих рассредоточенные системы, и выравнивания соответствующих белков приведены в работе (Ershova, 2012).

2.6 Анализ геномного контекста для генов рассредоточенных систем рестрикции-модификации

Для выявления причин образования рассредоточенных систем была проанализирована область размером 20 т.п.н. в обе стороны от генов ЭР и МТазы

для выявления мобильных генетических элементов [226]: IS (insertion sequence) элементов, генов транспозаз, интеграз и других белков бактериофагов. Поиск соответствующих элементов осуществлялся путем поиска соответствующих слов в геномной аннотации. Если мобильные элементы были найдены, осуществлялся поиск сходных элементов в окрестностях ортологичных систем P-M из других геномов (BLAST, E value < 0.01).

2.7 Оценка недопредставленности сайтов в геноме

Для характеристики давления отбора на сайт в геноме для данного сайта в данном геноме было оценено отношение наблюдаемого числа встреч такого сайта в геноме к статистически ожидаемому. Если эта величина близка к 1, давления на сайт нет, если эта величина значительно меньше 1, на сайт действует отрицательный отбор, если эта величина значительно больше 1, на сайт действует положительный отбор.

Для оценки ожидаемого числа сайтов был использован метод, предложенный в работе S. Karlin [8]. Этот метод описан в разделе 1.7.3.

Отношение наблюдаемого числа сайтов к ожидаемому в данной работе обозначено *Kr*. Для расчета *Kr* для сайта любой длины (site) формула 5 была применена следующим образом:

$$Kr = \frac{f_{obs}(site) \times \prod f_{obs}(evenN)}{\prod f_{obs}(oddN)} \qquad (6)$$

где $f_{obs}(site)$ - наблюдаемое в геноме число встреч данного сайта;

$f_{obs}(evenN)$ - наблюдаемое в геноме число подслов данного сайта с четным числом букв *N*, где *N* – любой из четырех нуклеотидов;

$f_{obs}(oddN)$ наблюдаемое в геноме число подслов данного сайта с нечетным числом букв *N*, где *N* – любой из четырех нуклеотидов.

2.8 Сравнение распределений величины *Kr* и границы недо- и перепредставленности

Для сравнения распределений *Kr* в двух наборах данных был использован критерий однородности Колмогорова-Смирнова.

В работе [8] последовательность считалась недопредставленной, если $Kr \leq 0.78$, и перепредставленной, если $Kr \geq 1.23$. В данной работе в большинстве случаев были использованы такие же границы недо- и перепредставленности. Использование других границ обсуждается в соответствующих местах текста.

В контрольном наборе, содержащем эукариотические вирусы наблюдается 1.7% недопредставленных пар сайт-геном и 1.9% перепредставленных пар сайт геном. Это показывает, что примерно в 2% случаев сайты могут избегаться по причинам, не связанным с действием систем Р-М. Это достаточно грубая оценка, поскольку эукариотические вирусы и прокариоты существенно различаются по размеру геномов, что может влиять на *Kr*, а также по образу жизни, что может влиять на причины отбора на различные короткие последовательности.

Для определения достоверности различий между двумя фракциями недопредставленных сайтов был использован метод $\chi 2$ или точный критерий Фишера.

В данной работе рассматриваются только те пары сайт-геном, где ожидаемое число сайтов в геноме было больше 15. Если ожидаемое число сайтов слишком маленькое, отношение наблюдаемого числа к ожидаемому может быть недостоверным. Граница в 15 сайтов была получена следующим образом. Ожидаемые значения были разбиты по карманам 6,5-7,5, 7,5-8,5 и т.д. Для каждого кармана было посчитано стандартное отклонение *Kr* путем сравнения *Kr* для непалиндромных сайтов со значением *Kr* для комплементарных сайтов. От кармана 14,5-15,5 и выше стандартное отклонение *Kr* не превышало 0,10 и уменьшалось с ростом ожидаемого числа сайтов. Поэтому, если ожидаемое число

сайтов 15 и более, вероятность того, что *Kr* будет меньше 0,78 по случайным причинам составляет 0,012 и быстро уменьшается с ростом числа ожидаемых сайтов.

2.9 Идентификация генов эндонуклеаз рестрикции, предположительно недавно полученных путем горизонтального переноса генов

Фрагменты, генома, предположительно полученные недавно путем горизонтального переноса были предсказаны с помощью программы Alien_Hunter с параметрами по умолчанию [227]. Для каждого фрагмента ДНК предположительно перенесенного путем горизонтального переноса эта программа возвращает его координаты, величину, характеризующую фоновый олигонуклеотидный состав данного генома (threshold) и величину, характеризующую олигонуклеотидный состав данного фрагмента (score). Чем сильнее олигонуклеотидный состав фрагмента отличается от значения, характерного для всего генома, тем более атипичным является данный фрагмент, и тем вероятнее, что он был недавно получен от отличающегося по олигонуклеотидному составу организма.

Полученные координаты предположительно горизонтально перенесенных фрагментов были сравнены с координатами генов систем Р-М. Для каждого фрагмента, содержащего гены систем Р-М, было посчитано отношение score/threshold, отражающее его сходство по олигонуклеотидному составу с остальным геномом.

2.10 Определение семейств белков

ЭР и МТазы, узнающие последовательность GATC были охарактеризованы по составу доменов БД Pfam [228]. Если последовательность белка содержала только один домен, то такой белок был отнесен к семейству, одноименному с соответствующим семейством доменов Pfam.

Если белок содержал несколько доменов, то применялась следующая процедура:

а), если белок содержал домены, принадлежащие к одному семейству Pfam, то этот белок был отнесен к семейству белков с одним доменом этого семейства;

б) если белок содержал домены нескольких семейств, и некоторые из них были аннотированы в БД Pfam как эндонуклеазные, в то время как другие из них были аннотированы, как МТазный домен, тогда белок был отнесен к семейству с объединенным названием, включающим названия обоих семейств Pfam (например, fused_D12/DpnII);

в) если белок содержал несколько различных доменов, и только один из этих доменов аннотирован Pfam как домен ЭР или МТазы, тогда такой белок был отнесен к семейству, одноименному с соответствующим эндонуклеазным или метилтрансферазным доменом.

2.11 Определение семейств систем рестрикции-модификации

Системы Р-М были отнесены к одному семейству если их эндонуклеазы рестрикции были отнесены к одному семейству белков и ДНК метилтрансферазы были отнесены к одному семейству белков. Детекция парных метилтрансфераз для одиночных ЭР осуществлялась методом, описанным в разделе 2.6.

2.12 Построение модели влияния систем рестрикции-модификации на недопредставленность сайта в геноме

Недопредставленность сайта узнавания системы Р-М в геноме (*Kr*) может быть описана как линейная функция влияния всех систем R-M и одиночных белков с соответствующей специфичностью, которые кодируются в геноме. Построенная модель влияния систем Р-М на недопредставленность сайта GATC учитывает информацию о видах, которые включают некоторые штаммы, кодирующие комплементарные системы Р-М, как независимую переменную p. Эта переменная была равна 1 для каждого генома таких видов (см. таблицу 3.6) и 0 для геномов

всех остальных видов. Чтобы найти лучшую (в терминах L^2-нормы) линейную функцию, был использован метод линейной регрессии с равными весами для всех геномов и без каких-либо дополнительных ограничений.

Глава 3. Результаты и обсуждение

3.1 Организация генов систем рестрикции-модификации

3.1.1 Идентификация одиночных эндонуклеаз рестрикции в полных геномах бактерий и архей и их классификация

Как описано выше, обязательными компонентами большинства систем Р-М являются ЭР и МТаза, которые закодированы отдельными генами. В случае систем типа IIC/G, один ген может содержать метилтрансферазный и эндонуклеазные домены. Кроме того, известны метил-зависимые ЭР, которые относят к типу IV и IIM, которые расщепляют метилированную ДНК, и, соответственно, не нуждаются в парной МТазе. Все эти случаи были отнесены к полным системам Р-М (см. таблицу 3.1). В некоторых случаях в геномах были аннотированы только отдельные белки, относящиеся к системам Р-М типа I-III. Такие системы в таблице 3.1 отнесены к неполным системам Р-М.

Анализ полных геномов и закодированных в них систем Р-М, представленных в БД REBASE, показал, что большинство прокариот содержат от 1 до 4 полных и столько же неполных систем Р-М, однако есть отдельные организмы (например, *Helicobacter pylori*), содержащие больше 20 систем в одном геноме. Также следует отметить, что кроме систем всего полных систем было найдено примерно столько же, сколько неполных (3039 и 3543 соответственно).

Рассмотрим более подробно, что представляют собой неполные системы Р-М. Как видно из таблицы 3.1, среди неполных систем РМ можно выделить несколько категорий. Наибольший интерес представляют собой найденные одиночные эндонуклеазы рестрикции.

Таблица 3.1

Аннотированные системы Р-М в полных геномах прокариот из набора 1

Категория систем Р-М	Хромосома	Плазмида	Всего
Полные системы Р-М	2762	277	3039
Неполные системы Р-М:	3232	298	3543
только МТазы:	2525	222	2747
сиротские, в т.ч. Dam и Dcm	420	11	431
только ЭР	254	18	272
другие белки систем Р-М	450	68	518
Всего	5994	575	6582

В 1040 геномах набора 1 (см. “Материалы и Методы”) было найдено 272 гена одиночных ЭР, среди которых есть представители всех типов систем Р-М (см. таблицу 3.2). Известно, что только ЭР типа II могут быть активны без соответствующей МТазы. Поэтому одиночные ЭР типа I и III для бактериальной клетки являются скорее бесполезными, но и безвредными. В отличие от них, присутствие одиночных ЭР типа II токсично для бактерии [173,215]. Присутствие одиночных генов предполагаемых ЭР может быть связано с различными причинами:

а) продукты этих генов, несмотря на сходство с ЭР, не проявляют эндонуклеазной активности;

б) нарушена структура гена, что приводит к отсутствию его экспрессии;

в) в геноме присутствует МТаза, которая метилирует сайт узнавания одиночной ЭР;

г) ЭР не проявляет активности в отсутствие соответствующей МТазы. Например, Gingeras и соавторы показали, что делеция гена МТазы системы Р-М PaeR7I *Pseudomonas aeruginosa* не вела к гибели клеток, содержащих оставшуюся одиночную ЭР [229,230]. Эта ЭР также не препятствовала инвазии бактериофагов,

что свидетельствует об инактивации ЭР. Трансформация клеток плазмидой, содержащей ген соответствующей МТазы восстанавливала активность ЭР и способность системы Р-М ограничивать инфекцию бактериофагов [229]. Секвенирование показало, что ген ЭР не содержал мутаций. Это свидетельствует о том, что для эндонуклеазной активности важны обе части данной системы Р-М.

д) геном бактерии может быть защищен от расщепления. В работе Vasu и соавторов [231] предполагается, что защита хромосомной ДНК структурными белками [232] может быть одним из возможных объяснений существования ЭР со случайной активностью. Кроме того, бактериальный геном может не содержать ни одного сайта узнавания данной системы Р-М. Например, ЭР PacI из *Pseudomonas alcaligenes* является одиночной ЭР без соответствующей МТазы. Shen и соавторы [233] объяснили существование бактерии с такой ЭР отсутствием сайтов узнавания этой ЭР в геноме бактерии.

е) бактерия, содержащая функциональную одиночную ЭР, может выжить благодаря высокой активности систем репарации [234,235] или присутствию ферментов, специфически гидролизующих ЭР [41].

В данной работе для 272 одиночных генов ЭР был произведен поиск парных МТаз методами сравнительной геномики. Как видно из таблицы 3.2, среди 272 найденных ЭР было найдено 109 генов, которые, вероятно, не экспрессируют активные ЭР. 10 из них не изучены экспериментально и не имеют ортологов среди ЭР, входящих в состав систем Р-М. Можно предположить, что они являются ошибкой аннотации. 99 генов содержат сдвиг рамки считывания или преждевременный стоп-кодон. Вероятно, эти гены являются фрагментом разрушенных систем Р-М.

Для того, чтобы обнаружить такие случаи, можно было бы сравнить сайты узнавания одиночных ЭР с сайтами узнавания МТаз, закодированных в тех же геномах. К сожалению, сайты узнавания и одиночных ЭР, и МТаз в большинстве случаев неизвестны. Другим способом обнаружить такие МТазы является поиск

корреляции между присутствием сходных ЭР и МТаз в одном геноме. Для этого был предпринят поиск ЭР, ортологичных одиночным, и поиск ортологичных МТаз в соответствующих геномах.

Таблица 3.2

Одиночные ЭР в геномах прокариот

Тип ЭР	Фрагменты	Гены парных МТаз не колокализованы	Гены парных МТаз колокализованы и повреждены	Не найдено парных МТаз	Всего
I	82	38 (3)*	29 (17)	10 (7)	159
II	23	19 (6)	11 (10)	21 (10)	74
III	4	0	34 (13)	1 (1)	39
Всего	109	57 (9)	74 (39)	32 (18)	272

*Числа в скобках указывают на число групп ортологичных ЭР.

Возможным объяснением существования в геномах оставшихся 173 генов, в случае, если они активны, может быть то, что их сайт узнавания перекрывается с какой-либо МТазой в геноме, которая метилирует его, и, тем самым, защищает от расщепления. Все МТазы, закодированные в соответствующих геномах можно считать кандидатами в парные МТазы к этим ЭР.

Оставшиеся 163 одиночные ЭР вошли в 58 ортологичных групп, каждая из которых включала две и более ЭР. При этом 26 групп включали две и больше одиночных ЭР. Для этих 163 одиночных ЭР был предпринят поиск парных МТаз. В результате этого анализа одиночные ЭР были разделены на три класса (см. таблицу 3.2): (i) ЭР, которые входят в состав рассредоточенных потенциальных систем Р-М; (ii) ЭР, гены которых локализованы рядом (на расстоянии не больше 4 т.п.н.) с поврежденными генами МТаз, содержащих сдвиги рамки или преждевременный стоп-кодон; (iii) ЭР, для которых не было найдено парных МТаз.

3.1.2 Сравнение идентифицированных одиночных эндонуклеаз рестрикции с метил-зависимыми эндонуклеазами рестрикции

Метил-зависимые ЭР типа IIM и IV часто имеют сходство по последовательности

с ЭР типа II [120], и, при этом, не нуждаются в парной МТазе. Если одиночные ЭР имеют высокое сходство по последовательности с метил-зависимыми ЭР, это может объяснить отсутствие рядом с ними генов парных МТаз. Для проверки этого предположения было предпринято сравнение аминокислотных последовательностей одиночных ЭР со всеми последовательностями известных метил-зависимых ЭР, доступными в БД REBASE.

Поиск гомологов одиночных ЭР среди известных ЭР типа IIM не дал ни одной находки.

Среди аминокислотных последовательностей метил-зависимых ЭР типа IV были найдены последовательности, гомологичные последовательностям некоторых одиночных ЭР из нашего списка (идентичность больше 40% при длине выравнивания больше 60% длины каждой последовательности). Последовательности одиночных ЭР и сходные с ними последовательности ЭР типа IV сформировали две несходные между собой группы: кластер 1 и кластер 2 (см. рисунки 3.1, 3.2).

В кластер 1 входят ЭР, сходные с субъединицей McrB ЭР типа IV. Это сходство связано с тем, что ЭР содержат ГТФ-связывающий домен (AAA_5 по БД Pfam), гомологичный соответствующему домену субъединицы McrB. Геномный контекст всех ЭР из кластера 1 включает дополнительную ЭР (см. рисунок 3.1.). Такие пары из двух колокализованных ЭР, одна из которых похожа по последовательности на метил-зависимые ЭР типа IV, описаны в литературе [O'Driscoll, 2006; O'Sullivan, 1995; Ohshima, 2002]. В этих работах было экспериментально показано, что продукты этих двух колокализованных генов формируют гетеродимер, который проявляет активность ЭР типа II.

Учитывая эти данные, можно предположить, что все ЭР, входящие в кластер 1, включая те, которые аннотированы REBASE как тип IV, в действительности являются компонентами ЭР системы типа II (см. рисунок 3.1).

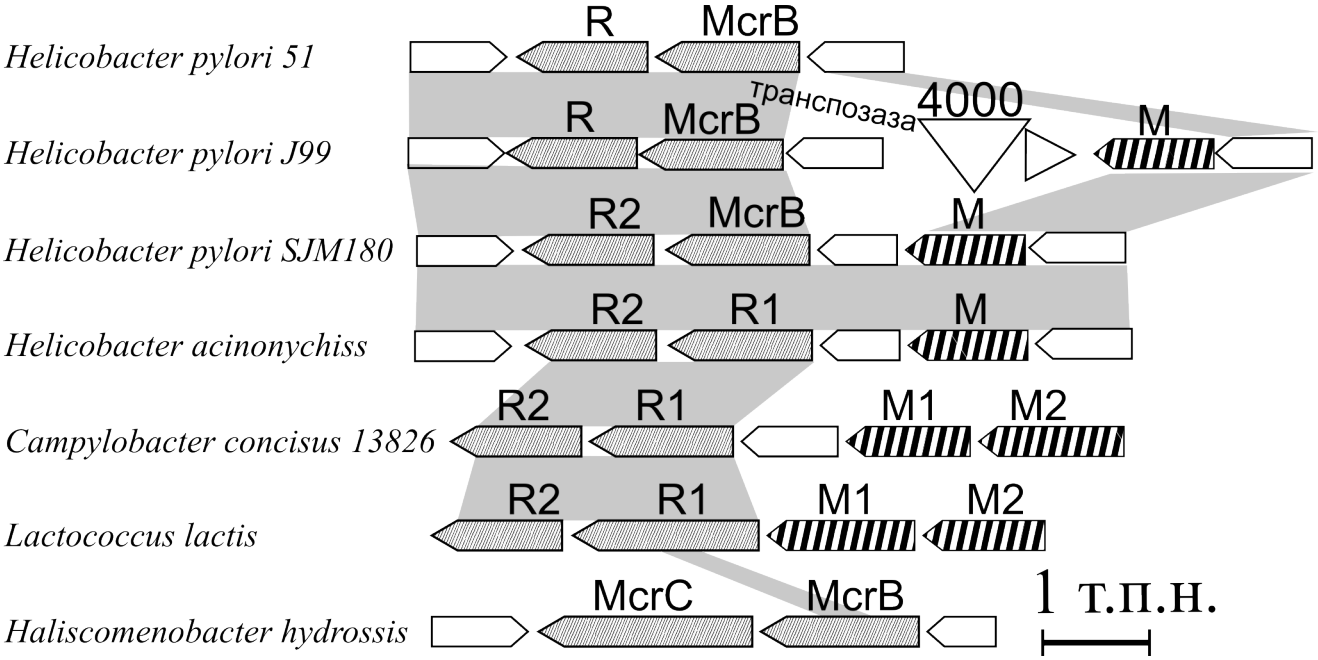
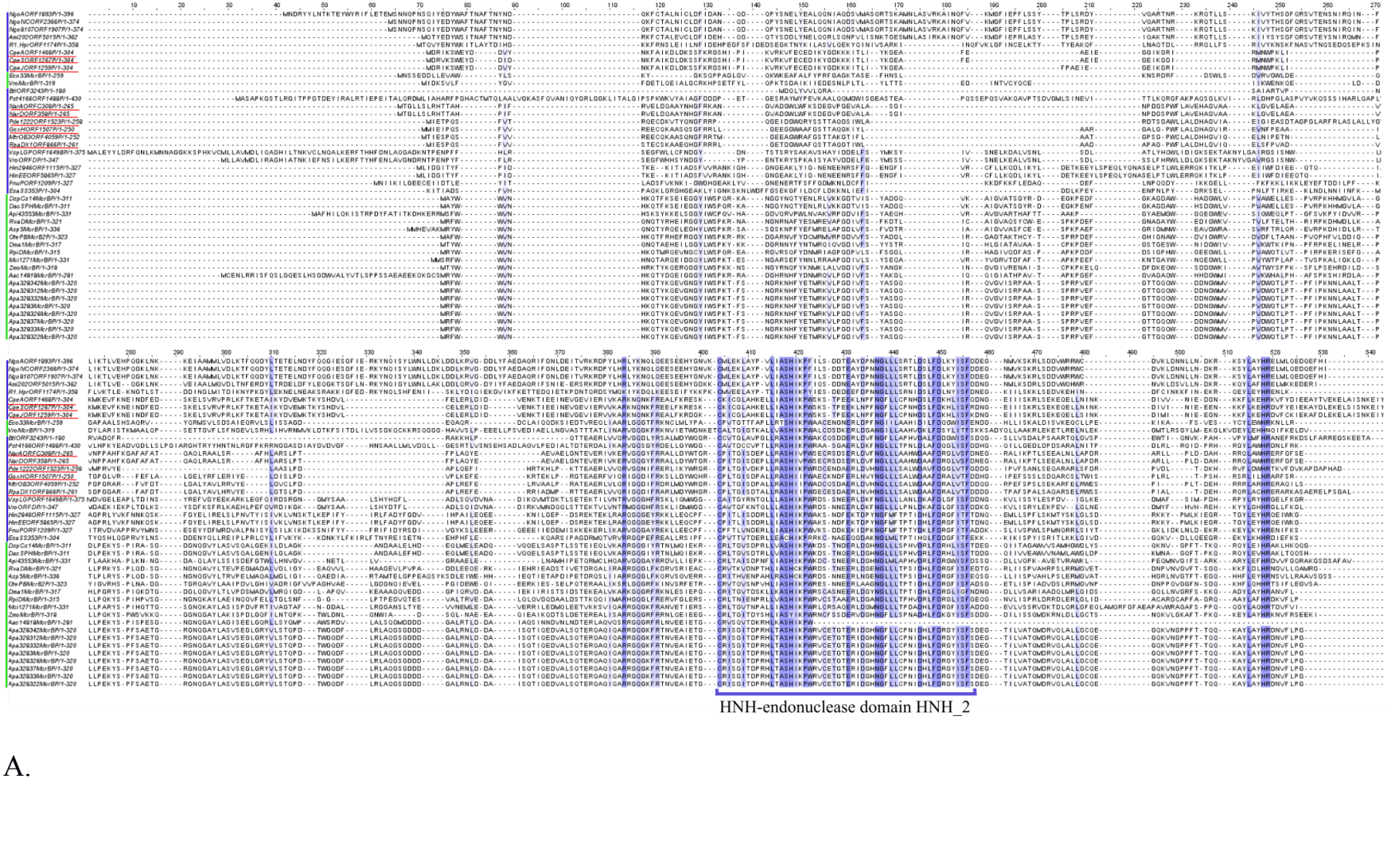
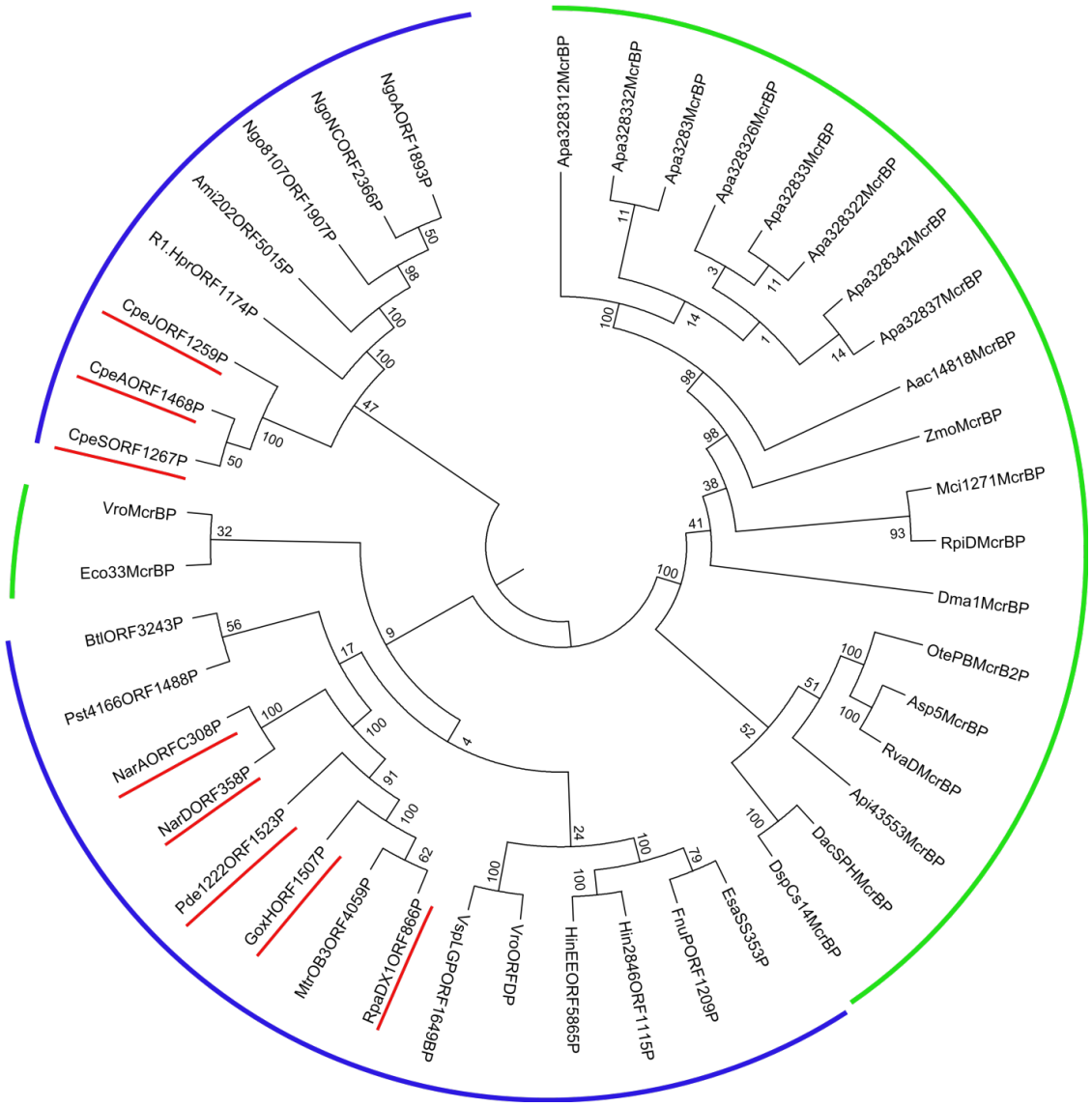


Рисунок 3.1 Пример организации генов ЭР, сходных по аминокислотной последовательности (>40% идентичности на >60% длины) с метил-зависимыми ЭР типа IV. Сходство этих ЭР типа II с ЭР типа IV обсуждается также в работе [236]. ЭР типа IV из *Haliscomenobacter hydrossis* не входят в кластер 1, и показаны для сравнения. Ортологичные гены соединены серыми прямоугольниками, гены, кодирующие МТазы, обозначены стрелками с частыми черно-белыми полосами, и буквой “M”, гены, кодирующие ЭР, показаны серыми стрелками и обозначены буквой “R”.



A.



Б.

Рисунок 3.2 В кластер 2 (>40% идентичности на более чем >60% длины) входят как ЭР типа II, так и метил-зависимые ЭР типа IV. При этом наряду с одиночными ЭР, в кластер 2 входят и ЭР из систем P-M, что позволяет предположить, что одиночные ЭР корректно отнесены к типу II, а не типу IV. ЭР, аннотированные в БД REBASE как ЭР типа II отмечены голубыми линиями, ЭР, аннотированые как ЭР типа IV маркированы зелеными линиями, одиночные ЭР выделены красным подчеркиванием. А. Выравнивание аминокислотных последовательностей ЭР кластера 2. Все последовательности содержат домен, аннотированный в БД Pfam как HNH_2. Цветом выделены позиции, содержащие аминокислотные остатки, консервативные среди большинства белков. Б. Филогенетическое дерево, построенное методом Neighbor-joining tree в программе MEGA, полученное методом бутстрэп-анализа, из множественного выравнивания представленного выше. Величины поддержки бутстрэп, показанные в узлах дерева, позволяют сделать вывод о сходстве ЭР. Видно, что ЭР типа IV (все, кроме двух) формируют ветку, достоверно отличающуюся от ЭР типа II. Одиночные ЭР формируют сходную группу, которая при этом ближе к ЭР типа II, чем типа IV.

Выравнивание аминокислотных последовательностей и филогенетический анализ, поддержанный бутсрэп-анализом, выявили, что одиночные ЭР, входящие в кластер 2, ближе к ЭР типа II, а не ЭР типа IV (см. рисунок 3.2). Таким образом, результаты анализа сходства последовательностей не позволяют отнести какие-либо из одиночных ЭР к метил-зависимым. Однако, нельзя исключить, что какие-то из одиночных ЭР в действительности являются метил-зависимыми ЭР, не имеющими сходства по последовательности с уже известными метил-зависимыми ЭР. Одиночные ЭР, входящие в кластер 1, вероятно, не являются токсичными в отсутствие второй субъединицы, позволяющей сформировать активный гетеродимер.

3.1.3 Группы ортологичных систем рестрикции-модификации

Для 163 одиночных ЭР был предпринят поиск парной МТазы (см. Материалы и Методы). Для 57 из них возможные парные МТазы были обнаружены. Полученные 57 пар ЭР-МТаза являются потенциальными рассредоточенными системами Р-М. Эти рассредоточенные системы входят в 11 групп ортологичных систем, относящихся к типам I и II (см. таблицу 3.3). Рассредоточенные системы типа III не были идентифицированы. С двумя исключениями, группы ортологичных систем Р-М, которые включают предсказанные рассредоточенные системы Р-М, состоят из систем Р-М, закодированных в неродственных таксонах: различные виды, классы, порядки или даже царства. По-видимому, это связано с распространением генов систем Р-М путем горизонтального переноса.

Все рассредоточенные системы Р-М предсказаны на основе сходства аминокислотной последовательности с другими белками, аннотированными в REBASE, как компоненты систем Р-М и имеют различную степень достоверности.

Наиболее вероятными системами Р-М представляются рассредоточенными системы Р-М, которые ортологичны аннотированым в REBASE системам Р-М, гены которых колокализованы.

Таблица 3.3

Список ортологичных групп систем Р-М, которые включают рассредоточенные системы Р-М.

Номер группы	Тип	Таксон	Общее число систем Р-М	Число систем Р-М, аннотированных в REBASE	Число рассредоточенных систем Р-М	Число одиночных ЭР
1	II	*Bordetella*: 3 вида	3	2	1	1
2	II	Proteobacteria: 3 класса	5	2	3	3
3	II	α-Proteobacteria: 4 порядка	5	0	5	5
4	II	*Clostridium perfringens*: 3 штамма	3	0	3	3
5	II	*Fibrobacter succinogenes*: 2 штамма	2	0	2	2
6	II	*Bacteroides*: 2 вида	2	0	2	2
7	II	*Bacteroides*: 3 вида	3	0	3	3
8	I	Archaea, Bacteria	79	29	50*	27**
9	I	Archaea, Bacteria	101	100	4	4
10	I	Archaea, Bacteria	24	23	1	1
11	I	Archaea, Bacteria	128	122	6	6

* Активность рассредоточенных систем Р-М из трех штаммов *Staphylococcus aureus* была подтверждена экспериментально [Waldron, 2006], см. дальнейшие пояснения в тексте.
** Из этих 27 одиночных ЭР, 25 закодированы в геномах *S. aureus*, 23 из которых содержат один ген R-субъединицы и две кассеты генов M- и S-субъединиц. Поэтому в этом случае число рассредоточенных систем получается больше числа одиночных генов R-субъединиц.
Полный список предсказанных рассредоточенных систем и их ортологов представлен в работе (Ershova, 2012).

Сохраняют ли такие рассредоточенные системы функциональную активность – вопрос, требующий дополнительных экспериментальных исследований.

В пользу сохранения активности этих систем говорит их высокое сходство по аминокислотной последовательности с системами Р-М, гены которых колокализованы (см. таблицу 3.3). Функционально неактивные гены быстро накапливают мутации и элиминируются из генома [237]. В то же время, даже при высоком сходстве по последовательности с функциональным белком, данный белок может оказать не функциональным. Например, в работе Zheng и соавторов

[215] показано, что белок HindVP (*Haemophilus influenzae*) имеет достаточно высокое сходство по аминокислотной последовательности (40% идентичности) с эндонуклеазами рестрикции, чья активность проверена экспериментально (в том числе, HgiDI, BsaHI). Однако этот белок не показал эндонуклеазной активности.

Кроме того, в процессе рекомбинаций, приведших к образованию рассредоточенных систем, может произойти потеря регуляторных элементов, что также может вести к отсутствию экспрессии соответствующих генов.

Таким образом, найденные пары являются кандидатами в рассредоточенные системы P-M, но их активность, также как активность любых предсказанных по сходству последовательностей систем P-M, нуждается в экспериментальной проверке.

3.1.4 Рассредоточенные системы типа I

Как видно из таблицы 3.3, группы ортологичных систем P-M типа I более многочисленны и всегда содержат аннотированные в REBASE системы P-M, гены которых колокализованы, чем группы ортологичных систем типа II. Это объясняется большей консервативностью R- и M-субъединиц типа I по сравнению с МТазами и, особенно, ЭР типа II [46]. В большинстве рассредоточенных систем P-M типа I гены R-субъединицы отделены от колокализованных генов M- и S-субъединицы.

Для 38 одиночных ЭР типа I были найдены возможные парные МТазы. Для трех из них (входящих в группу 8) экспериментально была показана функциональная активность [15].

На рисунке 3.3 показаны примеры организации генов в предсказанных рассредоточенных системах P-M типа I. Нумерация групп соответствует нумерации в Табл. 3.3.

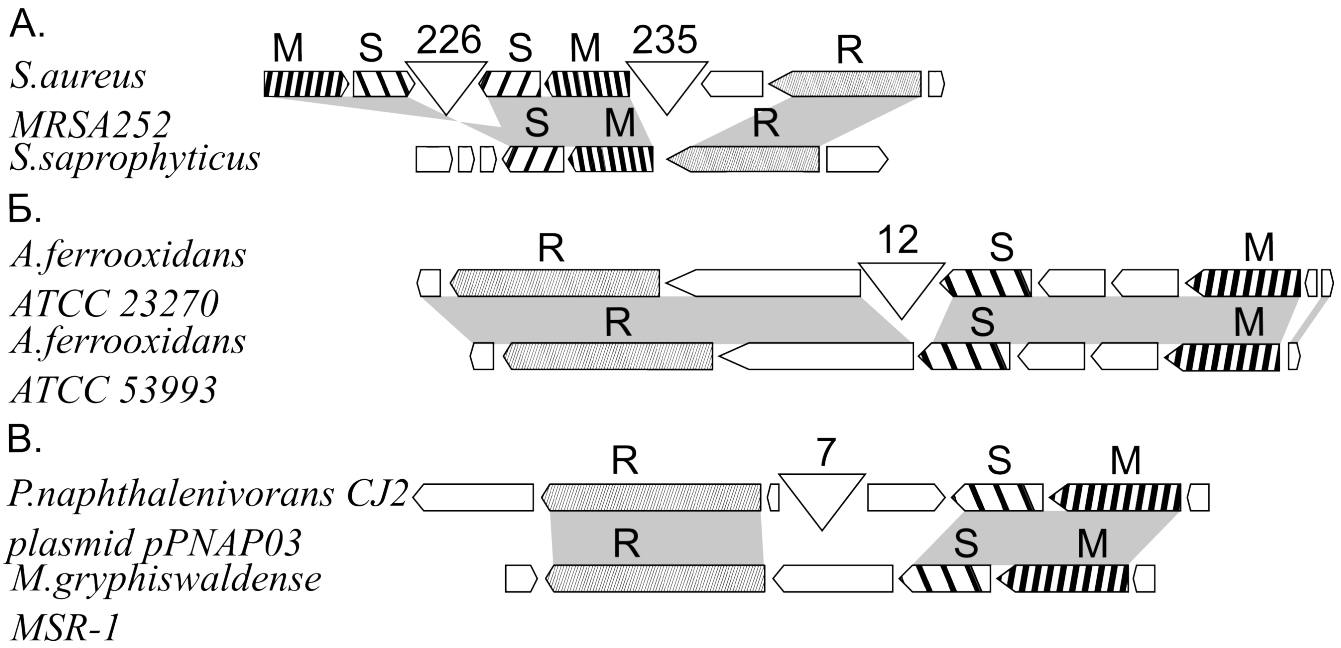


Рисунок 3.3 Примеры организации генов рассредоточенных систем типа I. Нумерация групп совпадает с таблицей 3.3. А. Представители группы 8 из *Staphylococcus aureus* и *Staphylococcus saprophyticus*. Б. Представители группы 10 из двух штаммов *Acidithiobacillus ferrooxidans*. В. Представители группы 9 из *Polaromonas naphthalenivorans* CJ2 и *Magnetospirillum gryphiswaldense* MSR-1. Ортологичные гены соединены серыми прямоугольниками, гены, кодирующие метилтрансферазные субъединицы, обозначены стрелками с частыми черно-белыми полосами, и буквой “M”, гены, кодирующие эндонуклеазные субъединицы, показаны серыми стрелками и обозначены буквой “R”, гены S-субъединиц обозначены стрелками с редкими черными полосами и буквой “S”. Остальные открытые рамки обозначены пустыми стрелками. Увеличенные расстояния между генами обозначены треугольниками, над которыми указано расстояние в т.п.н.

Рассмотрим более подробно группу 8, включающую системы P-M типа I, которые были изучены экспериментально.

Эта группа включает 79 систем P-M типа I из 25 штаммов *Staphylococcus aureus*, одного штамма *Anabaena variabilis* ATCC 27893, и одного штамма *Oscillatoria* sp. На рисунке 3.3 А показан пример организации генов рассредоточенной системы из *S. aureus*. Ген hsdR локализован на расстоянии более, чем 200 т.п.н. от локуса, содержащего гены hsdM и hsdS. На расстоянии около 200 т.п.н. от первого локуса расположен еще один локус, содержащий гены hsdM и hsdS. M-субъединицы в этих кассетах сходны по аминокислотной последовательности (85% идентичности). S-субъединицы различаются по последовательности в области ДНК-узнающих доменов (TRD). Поскольку S-субъединица в системах P-M типа I отвечает за специфическое взаимодействие с ДНК, такая организация генов

приводит к появлению двух систем P-M типа I с различной специфичностью.

Такая организация систем P-M типа I наблюдалась в 23 штаммах *S. aureus*. В работе [15]. для трех штаммов *S. aureus* (*S. aureus* 8325-4, 8325-4, 879R4RF, COL) было показано, что эти системы активны и предотвращают обмен ДНК между *S. aureus* и *E. coli*. В геномах *S. aureus* subsp. aureus ST398, *S. aureus* subsp. *aureus* JKD 6008, *Anabaena variabilis* ATCC 27893 и *Oscillatoria* sp. найдены одиночные гены hsdR и одна кассета, содержащая гены hsdM и hsdS.

Три представителя вида *Staphylococcus* и представители 26 других видов бактерий и архей содержат системы типа I, гены которых колокализованы, и ортологичны соответствующим генам рассредоточенных систем, описанных выше. На рисунке 3.3 A показана организация генов одной из рассредоточенных и ортологичной ей системы P-M, гены которой колокализованы. Во всех 26 штаммах *S. aureus* две пары генов hsdM и hsdS локализованы на двух геномных островах, содержащих многочисленные повторы [148,238]. В окрестности 20 т.п.н. от генов одиночных ЭР не было обнаружено каких-либо мобильных элементов: повторов, транспозонов, генов фаговых белков или генов, относящихся к рекомбинации.

Группа 10 ортологичных систем P-M включает рассредоточенную систему из *Acidithiobacillus ferrooxidans* ATCC 23270 (см. рисунок 3.3) и 23 обычные системы P-M, гены которых колокализованы из различных видов бактерий и архей.

Группа 9 включает четыре рассредоточенные системы из *Nitrosococcus oceani* ATCC 19707, *Nitrosococcus watsoni* C-113, *Polaromonas naphthalenivorans* CJ2, *Pseudomonas stutzeri* A1501 и 100 систем P-M, гены которых колокализованы, из различных бактерий и архей. Пример систем этой группы показан на рисунке 3.4 B.

3.1.5 Рассредоточенные системы типа II

Возможные парные МТазы были идентифицированы для 19 одиночных ЭР типа

II. Только две группы ортологичных систем Р-М типа II (группы 1 и 2 в таблице 3.3) включают как рассредоточенные системы Р-М, так и системы Р-М, гены которых колокализованы.

Группа 1 включает рассредоточенную систему Р-М типа II из *Bordetella pertussis*, которая состоит из ЭР (BpeTORF204P) и МТазы (M.BpeTORF740P), гены которых находятся на расстоянии около 160 т.п.н. Две ортологичные системы Р-М BbrRORF307P и BpaSORF304P, гены которых колокализованы, были найдены в геномах *Bordetella bronchiseptica* и *Bordetella parapertussis* соответственно. Организация генов систем группы 1 показана на рисунке 3.4А. Сходство аминокислотных последовательностей белков этих систем высокая, и составляет для ЭР >98% идентичности и для МТаз >99% идентичности. Такое высокое сходство позволяет предполагать происхождение всех трех систем Р-М от одного предка. Сходство геномного контекста для генов ЭР и МТазы рассредоточенной системы Р-М, найденной в геноме *B. pertussis* и колокализованных генов систем BbrRORF307P и BpaSORF304P позволяет предполагать сохранность регуляции транскрипции, и как следствие, экспрессию генов этой рассредоточенной системы.

Большое расстояние и изменение взаимной ориентации генов рассредоточенной системы *B. pertussis* могут быть объяснены значительными геномными перестройками в геноме *B. pertussis*, вызванными экспансией инсерционной последовательности семейства IS481 [239]. Инсерционные элементы IS 481 были найдены рядом с генами ЭР и МТазы. Ориентация этих повторов свидетельствует о том, что организация генов этой рассредоточенной системы Р-М связана действительно с внутрихромосомными перестройками, а не с ошибками сборки генома. То, что гены ЭР и МТазы локализованы в центре различных контигов (BX640411.1 и BX640413.1, соответственно) также свидетельствует о большом расстоянии между этими генами.

Группа 2 (см. рисунок 3.4 Б) включает пять ортологичных систем, из которых

гены системы RvaDORF1484P из *Rhodomicrobium vannielii* колокализованы, гены четырех других систем рассредоточены. Во всех случаях неподалеку от генов ЭР или МТазы находится открытая рамка считывания, аннотированная как никирующая эндонуклеаза (V-белок). Возможно, этот белок функционально важен для данных систем P-M.

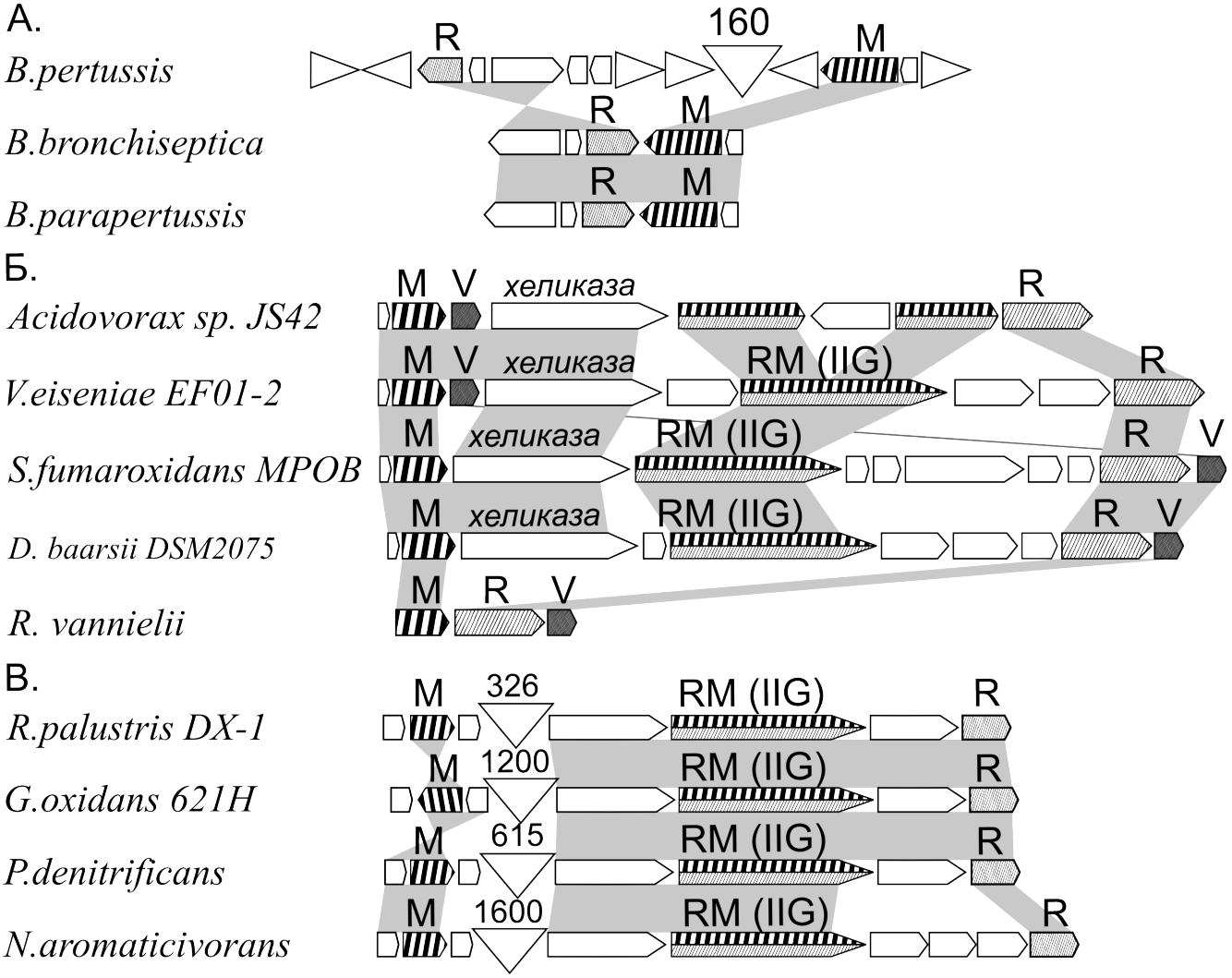


Рисунок 3.4 Организация генов в трех группах ортологичных систем P-M, которые включают в себя рассредоточенные системы. Обозначения такие же, как на рисунке 3.3. Кроме того, вытянутыми треугольнками обозначены инсерционные элементы. Номера групп соответствуют номерам из таблицы 3.3. А. Организация генов систем группы 1 из *Bordetella pertussis, Bordetella bronchiseptica*, и *Bordetella parapertussis*. Б. Представители ортологичных систем типа II из группы 2. В. Представители ортологичных систем типа II из группы 3.

В данном случае гены ЭР и МТазы находятся на расстоянии 7-14 т.п.н. друг от друга, что достаточно близко, хоть и дальше, чем обычно располагаются гены ЭР и МТазы в системах рестрикции-модификации. Такое положение сохраняет

возможность горизонтального переноса генов этой системы.

Организация генов систем P-M, входящих в группу 3 показана на рисунке 3.4 B. По аннотации REABSE, в геномах *Rhodopseudomonas palustris* DX-1, *Gluconobacter oxydans* 621H, *Paracoccus denitrificans* PD1222, и *Novosphingobium aromaticivorans* DSM 12444 предсказаны одиночные ЭР с неизвестной специфичностью. Эти белки ортологичны. Также в этих геномах содержатся ортологичные МТазы и белки типа IIG, обладающие как эндонуклеазной, так и метилтрансферазной активностью. Ортологичные МТазы сходны (>50% сходства на 90% длины) с орфанными МТазами, узнающими последовательность GANTC, в частности, M. CcrMI, которая имеет самостоятельное значение, например, в регуляции клеточного цикла [240]. Гены этих МТаз закодированы на расстоянии от 330 до 1600 т.п.н. от генов одиночных ЭР (см. рисунок 3.4.B).

Гены, кодирующие белки типа IIG, расположены неподалеку от генов предполагаемых одиночных ЭР. Системы типа IIG довольно часто [51] колокализованы с геном дополнительной одиночной МТазой с той же специфичностью, но в литературе не было найдено ни одного примера систем P-M, включающих слитный белок типа IIG и одиночную ЭР.

По-видимому, данные белки были аннотированы как ЭР из-за наличия HNH-эндонуклеазного домена в их последовательности (домен семейства HNH_2 согласно БД Pfam). Однако соответствующий домен характерен не только для ЭР типа II, но также для метил-зависимых ЭР типа IV, а также для других белков, например, колицинов, хоминг эндонуклеаз и т.д [241].

Без экспериментальной проверки полученные данные не позволяют сделать вывод о том, что данные белки являются ЭР типа II. В случае, если они действительно могут проявлять соответствующую активность, найденные одиночные МТазы или белки типа IIG могут защищать хозяйскую ДНК от действия соответствующих ЭР.

3.1.6 Геномный контекст генов рассредоточенных систем рестрикции-модификации

В окрестности 20 т.п.н. одного или нескольких генов, входящих в рассредоточенную систему P-M часто (в 38 из 57 случаев) были найдены различные мобильные генетические элементы. Это позволяет предположить, что гены этих систем были колокализованы и сейчас оказались на большом расстоянии друг от друга из-за геномных перестроек, вызванных мобильными элементами.

3.1.7 Одиночные эндонуклеазы рестрикции, для которых не были найдены парные ДНК-метилтрансферазы

Как видно из таблицы 3.2, для 106 одиночных ЭР не было найдено возможных парных МТаз. Для 74 из них рядом с генами ЭР найдены гены МТаз, содержащие сдвиги рамки или преждевременный стоп-кодон. Для всех этих 74 пар в других геномах были найдены ортологичные системы P-M, содержащие полноразмерные гены ЭР и МТазы (см. рисунок 3.5).

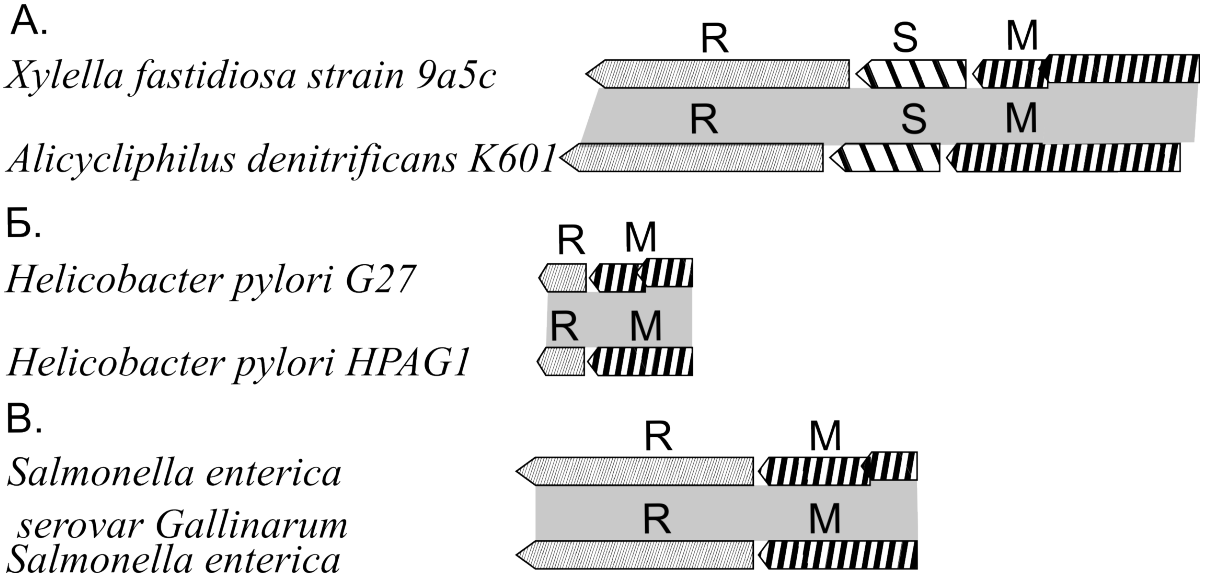


Рисунок 3.5 Примеры систем P-M с поврежденными генами МТаз. Обозначения те же, что на рисунках 3.3, 3.4. Сдвиг рамки показан сдвигом в области стрелки, обозначающей ген.

Некоторые ошибки в генах МТаз (сдвиг рамки или преждевременный стоп-кодон) могут быть ошибками секвенирования, и в действительности такие гены могут быть функциональны. О такой возможности свидетельствуют литературные

данные. Например, ген МТазы MJ1209 из системы MjaVIP *Methanocaldococcus jannaschii* (RefSeq ID NC_000909) был аннотирован как псевдоген с двумя сдвигами рамки считывания. Однако было экспериментально показано, что сдвиги рамки считывания являются ошибками секвенирования, этот ген экспрессируется, и его продукт является функционально активным [215]. В том же геноме в гене МТазы M. MjaIV также был найден сдвиг рамки считывания. Поскольку ЭР R.MjaIV является активной, Zheng и соавторы полагают, что сдвиг рамки считывания в этом гене МТазы также является ошибкой секвенирования [215]. Возможно, подобные ошибки достаточно распространены. Так, в работе Yu и соавторов [242] были проверены 138 генов *Brucella abortus* S19, содержащих сдвиги рамки считывания или преждевременные стоп-кодоны. Повторное секвенирование соответствующих фрагментов ДНК показало отсутствие повреждения в 109 из 138 исследованных генов. Сдвиги рамки считывания или преждевременные стоп-кодоны были подтверждены только в 29 генах.

Из таблицы 3.2 видно, что для 32 генов предполагаемых одиночных ЭР, не было найдено никаких кандидатов в парные МТазы на основании сходства одиночных ЭР с ЭР из других систем P-M. Возможно, что парная МТаза для данной ЭР закодирована в данном геноме, но не обнаружена в данной работе, т.к. не имеет гомологии с известными МТазами. Так, например, M.NruI не имеет значимого сходства по последовательности со всеми ранее найденными МТазами [243].

Системы P-M, показанные на рисунке 3.1, являются также хорошим примером случаев, когда при ортологичных ЭР соответствующие парные МТазы не обладают сходством аминокислотной последовательности. Системы P-M, включающие сходные ЭР и сильно различающиеся негомологичные МТазы описаны в работе [236]. Такие парные МТазы не могут быть найдены методом, использованным в данной работе. Можно предположить, что если одиночные ЭР действительно проявляют эндонуклеазную активность, их специфичность перекрывается с какими-то МТазами в геноме. Поэтому для этих 32

предполагаемых одиночных ЭР, как и для 10 ЭР, не имеющих близких ортологов, все МТазы того же типа рассматриваются как потенциальные. Интересно, что для некоторых предполагаемых одиночных ЭР (3 типа II и 4 типа I) не было найдено ни одной МТазы того же типа.

Среди этих 32 одиночных ЭР, 21 относится к типу II. Все эти ЭР имеют ортологов, входящих в аннотированные в REBASE системы P-M. Присутствие таких генов в геноме требует объяснений, поскольку ЭР типа II токсичны для клетки без соответствующей МТазы. Анализ аннотированных в REBASE систем P-M, ЭР которых ортологичны данным предполагаемым одиночным ЭР показал, что во всех случаях ген МТазы локализован непосредственно перед геном ЭР. Возможно, что при потере гена МТазы, приведшего к образованию одиночного гена ЭР, произошла потеря регуляторных элементов, что привело к отсутствию экспрессии гена такой одиночной ЭР. Известно, что в некоторых случаях такие гены транскрибируются [237], но нет данных об их трансляции.

В геномах *H. pylori* было найдено девять одиночных ЭР, которые относятся к четырем ортологичным группам. Пять из них (ортологичная группа 18) колокализованы с ранее описанными McrB-подобными ЭР (см. рисунок 3.1).

Шесть одиночных ЭР из группы 37 (*Bacteroides*) ортологичны пяти ЭР, для которых была предсказана парная МТаза (группы 6 и 7). Все 11 одиночных ЭР сходных друг с другом и аннотированы как HpaII-подобные ЭР. Такое высокое сходство может свидетельствовать о том, что эти белки функциональны. Согласно REBASE одна из них, BthVORF1149P, не проявляет эндонуклеазной активности. Поэтому можно предположить, что все эти 11 белков не являются ЭР, а выполняют какую-то иную функцию.

Наиболее важным результатом данной работы является обнаружение рассредоточенных систем P-M, гены которых находятся на значительном расстоянии друг от друга. Можно предположить два сценария формирования таких систем.

Во-первых, рассредоточенные системы Р-М могут возникать после геномных перестроек, которые часто происходят благодаря внедрению в геном транспозонов или профагов [244]. Таким способом могли быть сформированы, например, рассредоточенные системы в геномах *B. pertussis* и *S. aureus*, поскольку гены соответствующих рассредоточенных систем окружены мобильными элементами, и среди ортологичных систем Р-М найдены системы, гены которых колокализованы.

Во-вторых, рассредоточенная система Р-М может возникнуть при горизонтальном переносе участка, содержащего ген ЭР в геном, кодирующий МТазу с такой же или более широкой специфичностью [1,147]. Подобный механизм мог привести к возникновению систем Р-М из группы 3 (рисунок 3.4 В).

Можно предположить, что благодаря частым внутрихромосомным перестройкам, разобщение генов систем Р-М происходит достаточно часто. Вероятно, в большинстве случаев это сопровождается повреждением гена. Действительно, в геномах прокариот были найдены гены ЭР с преждевременным стоп-кодоном (см. таблицу 3.2). Однако, в некоторых случаях, такая рассредоточенная система может сохранять свою активность, как например, описанные системы Р-М *S. aureus*.

Гены рассредоточенных систем Р-М теряют способность к одновременному горизонтальному переносу, т.к. вероятность одновременного переноса для двух генов из различных частей хромосомы значительно ниже, чем для колокализованных генов. Поэтому, если рассматривать систему Р-М как эгоистичный элемент генома, рассредоточенные системы Р-М являются эволюционным тупиком. Это может объяснить редкость таких форм систем Р-М. Так, в проанализированных 1040 геномах было обнаружено 57 предположительных рассредоточенных систем Р-М, что значительно меньше 3000 аннотированных систем Р-М.

Ортологичные рассредоточенные системы были найдены в различных таксонах (см. таблицу 3.2). Это показывает, что рассредоточенные системы из одной и той

же колокализованной системы Р-М могли возникать несколько раз независимо. Это соображение хорошо согласуется с тем, что некоторые гены ортологичных рассредоточенных систем Р-М окружены неродственными мобильными элементами.

Возможно, наличие рассредоточенных систем Р-М несет какие-то дополнительные преимущества для бактерии. Например, одиночные ЭР типа II могут быть токсичны для другой бактерии при горизонтальном переносе в клетки других бактерий. Одиночная ЭР LhopHLHKP, закодированная на плазмиде, кажется перспективным кандидатом для такого применения.

Интересно, что гены всех предсказанных в данной работе парных МТаз также являются одиночными, хотя наша процедура поиска парных МТаз этого не требовала (см. Материалы и Методы). Это позволяет предположить, что, по крайней мере, часть одиночных МТаз может быть парными к неизвестным одиночным ЭР, что объясняет большое число одиночных МТаз в геномах прокариот, см. таблицу 3.1, а также работы Seshasayee с соавт. [146].

3.1.8 Заключение по разделу

Проведенный анализ позволил выявить новую форму существования систем Р-М - рассредоточенные системы, гены ЭР и МТазы которых не колокализованы, а разнесены на большие расстояния (до нескольких тысяч п.н.) в геноме. В этой форме системы Р-М могут сохранять функциональную активность, о чем свидетельствуют экспериментальные данные, полученные для рассредоточенных систем типа I *S. aureus*. По-видимому, рассредоточенные системы представляют собой один из этапов эволюции систем Р-М. Предложенный метод аннотации рассредоточенных систем может быть использован при аннотации систем Р-М во вновь секвенированных геномах.

3.2 Недопредставленность сайтов систем рестрикции-модификации в геномах прокариот

Как обсуждается в разделе 1.6.1 обзора литературы, различными авторами было показано, что короткие палиндромные последовательности статистически недопредставлены в геномах бактерий. Эту недопредставленность объясняют токсичностью ЭР для хозяйской клетки. Однако небольшое количество доступных на тот момент геномов и аннотированных в них систем Р-М с известным сайтом узнавания, не позволило авторам сделать выводы о влиянии конкретных систем Р-М, закодированных в данном геноме на недопредставленность ее сайта узнавания в данном геноме.

Методы детекции недопредставленности сайта могут существенно повлиять на результаты анализа [220]. Все методы основаны на сравнении числа наблюдаемых сайтов в геноме с числом сайтов, которые статистически ожидаются исходя из свойств последовательности данного генома. Для оценки ожидаемого числа сайтов применяются различные методы, описанные в обзоре литературы (см. Раздел 1.6). Как показано в работе Elhai [220], метод, предложенный в работе Karlin и Cardon [8] наилучшим образом подходит для анализа недопредставленности сайтов систем Р-М. Одной из причин, вероятно, является то, что этот метод позволяет учитывать вырожденные последовательности, а сайты систем Р-М часто являются вырожденными.

В данной работе в качестве меры недопредставленности или перепредставленности сайта использовано отношение (*Kr*) наблюдаемого числа сайтов к ожидаемому, которое оценивается согласно модели, предложенной в работе Karlin и Cardon [8]. Считается, что сайт узнавания системы Р-М избегается в геноме (и находится под действием отрицательного отбора), если он является недопредставленным, т.е. величина *Kr* для данной пары (сайт, геном) меньше некоторого порога.

Ограничением такого подхода является то, что он не позволяет обнаруживать избегание сайтов недавно приобретенных систем Р-М или систем, с изменившимся сайтом узнавания, поскольку для проявления действия отбора нужно некоторое время (поколения бактерий).

3.2.1 Избегание сайтов систем рестрикции-модификации различных типов.

Системы Р-М разных типов значительно различаются по структурно-функциональной организации белковых комплексов и структуре сайтов узнавания, поэтому сайты систем разных типов были проанализированы отдельно.

Сайты систем типа II были дополнительно разделены на сайты систем типа IIC/G, типа IIM и все остальные, которые были названы “ортодоксальными”. Такое деление связано с тем, что системы IIM расщепляют метилированные сайты узнавания, и, таким образом, функционально сходны скорее с системами типа IV [120], чем II, а системы типа IIC/G объединяют эндонуклеазную и метилтрансферазную функции в одном комплексе (часто и в одном полипептиде) [51] и, таким образом, функционально ближе к системам типа I и III, чем к типу II, классические представители которого включают два фермента (ЭР и МТазу), которые взаимодействуют с сайтом узнавания независимо.

В соответствии с работой [8], в данной работе сайт узнавания системы Р-М считается недопредставленным в геноме, если отношение наблюдаемого числа сайтов к ожидаемому (*Kr*) меньше или равно 0,78. В таблице 3.4 показано число случаев, когда сайт был недопредставлен в геноме для сайтов разных типов систем Р-М.

Недопредставленность сайтов в наборах актуальных пар была обнаружена только для сайтов ортодоксальных систем Р-М типа II (850 из 1774 случаев), а также для ЭР типа IIM, узнающих метилированный сайт GATC (в 38 из 42 случаев). В последнем случае недопредставленность сайта GATC трудно объяснить

токсичностью ЭР, поскольку метил-зависимые ЭР не токсичны для хозяйского генома [120].

Таблица 3.4

Процент недопредставленных сайтов систем различных типов в геномах прокариот и эукариотических вирусов (в скобках приведены абсолютные числа)

Серой заливкой выделены наборы пар, в которых отмечалась заметная недопредставленность сайтов в геномах

Тип системы P-M	Актуальные пары	Актуальные пары, включающие только сайты экспериментально подтвержденных систем P-M	Прокариотический контроль	Вирусный контроль
Тип I	0.0% (0/100)	0.0% (0/14)	0.1% (238/357501)	0.1% (21/18859)
Тип III	0.0% (0/76)	0.0% (0/7)	0.3% (213/82065)	0.2% (57/31571)
Тип IIC/G	0.0% (0/107)	0.0% (0/47)	0.1% (171/218322)	0.2% (66/27699)
Тип II ортодоксальные	47.9% (850/1774)	45.3% (58/128)	3.9% (21380/542911)	1.7% (2720/158921)
Тип IIM	70.4% (38/54[1])	14.3% (1/7)	0.6% (125/21128)	0.3% (79/29070)
Тип IV	0.0% (0/13)	0.0% (0/3)	1.0% (64/6342)	0.2% (25/10116)

[1]Все 38 недопредставленных сайтов являются сайтами GATC

Большинство эукариотических вирусов не взаимодействует с прокариотическими системами Р-М, и в геномах этих вирусов не наблюдается недопредставленность сайтов систем Р-М. Интересно, что в геномах некоторых эукариотических вирусов (вирусы *Chlorella*, *Marseilleviridae* и *Phaeocystis globosa*) были обнаружены гены ортодоксальных систем Р-М типа II. В этих геномах наблюдалась недопредставленность соответствующих сайтов в 8 из 17 случаев.

Разница в распределении *Kr* в наборах актуальных пар, содержащих сайты ортодоксальных систем типа II и пар, содержащих сайты остальных проанализированных типов систем Р-М (I, III, IIG) показана на рисунке 3.6. Видно, что распределения *Kr* сходны как в прокариотическом контрольном наборе, так и в наборе эукариотических вирусов. Как и ожидается, большинство значений близки к 1. При этом распределение *Kr* актуальных пар, содержащих сайты ортодоксальных систем типа II, значительно сдвинуто в область недопредставленности и значительно отличается от распределения *Kr* в контрольных наборах (см. рисунок 3.6).

Недопредставленность сайтов была обнаружена в 47,9% пар, содержащих сайты ортодоксальных систем типа II. В прокариотическом контрольном наборе недопредставленность обнаруживалась только в 3,9% пар, содержащих соответствующие сайты. Возможно, частично эта недопредставленность может быть объяснена следами потерянных систем Р-М (см. далее). Среди геномов, которые не кодируют известных систем Р-М, недопредставленность наблюдалась в 0,7% пар, содержащих сайты ортодоксальных систем типа II. Распределение *Kr* в этом наборе близко к распределению *Kr* для прокариотического контрольного набора.

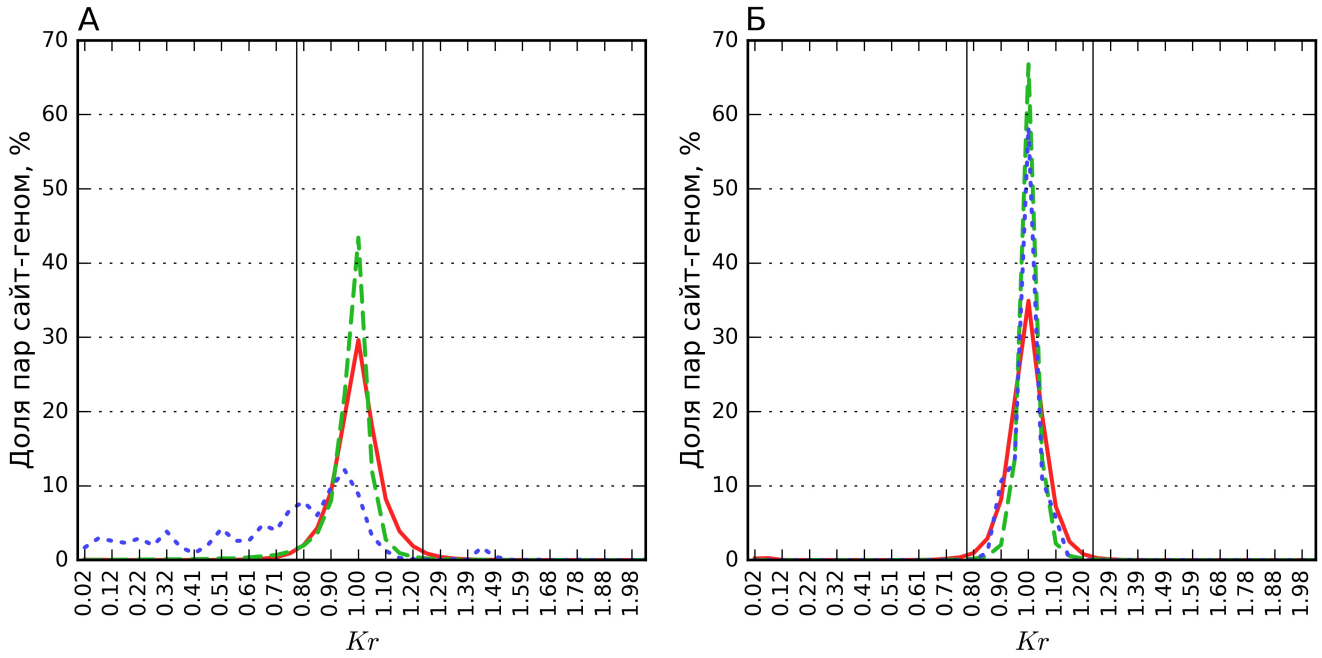


Рисунок 3.6 Гистограммы распределения *Kr* для различных наборов пар сайт-геном. Отрезок [0:2) разбит на 41 карман, процент пар с *Kr,* попадающими в этот карман, показан над его серединой. Границы недопредставленности и перепредставленности показаны вертикальными линиями. Распределение *Kr* для набора актуальных пар (содержащих только такие пары сайт-геном, что системы с таким сайтом закодированы в данном геноме) показано синим пунктиром, зеленой прерывистой линией показано распределение *Kr* для прокариотического контроля (набора пар, включающих все возможные пары сайт-геном, где сайт относится к той же группе, что и сайты актуального набора, но не требуется, чтобы система с таким сайтом была закодирована в данном геноме). Распределение *Kr* в геномах эукариотических вирусов (вирусный контроль) показано сплошной красной линией. А. Распределение *Kr* для наборов, включающих только сайты ортодоксальных систем Р-М типа II. Б. Распределение *Kr* для наборов, включающих только сайты систем Р-М типов I, IIC/G и III вместе.

В отличие от распределения *Kr* для пар, содержащих актуальные сайты ортодоксальных систем Р-М типа II, распределение *Kr* для актуальных пар, содержащих сайты систем типов I, IIC/G и III практически не отличается от распределения *Kr* в соответствующих контрольных наборах (см. рисунок 3.5Б). Некоторая разница между графиками наблюдается в области $0.8 < Kr < 0.95$, и связана, главным образом, только с сайтом CAGAG ЭР типа III из 20 штаммов *Salmonella enterica*.

Среди актуальных пар (сайт-геном), использованных в данной работе, встречается большое количество предсказанных сайтов. Ошибочные предсказания могут оказать влияние на наши оценки недопредставленности сайтов систем Р-М в геномах прокариот. Для оценки влияния ошибочных предсказаний было

исследовано избегание сайтов среди экспериментально подтвержденных систем Р-М. Среди актуальных пар, содержащих только сайты экспериментально подтвержденных ортодоксальных систем Р-М типа II недопредставленность наблюдалась в 45,3% случаев. Эта величина близка к 47,9% случаев недопредставленности, которая наблюдается среди всех актуальных пар, содержащих сайты ортодоксальных систем Р-М типа II. Полученные результаты свидетельствуют о том, что, хотя среди проанализированных актуальных пар могут быть пары, содержащие ошибочно предсказанные сайты, влияние таких ошибочных предсказаний невелико и не позволяет объяснить отсутствие избегания в примерно половине случаев.

Избегание сайтов ортодоксальных систем Р-М типа II было найдено примерно в половине всех исследованных случаев. При этом 59 сайтов систем Р-М избегаются в большинстве геномов, где закодированы соответствующие системы. Избегания сайтов других систем Р-М, в т.ч. типов I, III (кроме сайта CAGAG), IIC/G, IV, IIM (кроме сайта GATC) найдено не было.

Найденное в данной работе избегание сайтов ортодоксальных систем Р-М соответствует результатам, полученным в предшествующих работах [2–4,7,8,188]. Этот эффект может быть связан с возможным расщеплением ДНК бактерии ЭР при неполном метилировании соответствующих сайтов узнавания. Однако, в отличие от избегания сайтов узнавания систем Р-М в геномах бактериофагов [186], которые могут полностью элиминировать соответствующие сайты в своих геномах, в геномах бактерий даже сильная недопредставленность сайта (например, Kr = 0.12) может означать, что в геноме присутствуют сотни соответствующих сайтов. Кажется маловероятным, что такое снижение числа сайтов может защитить бактериальную ДНК от расщепления, поскольку для расщепления ДНК ЭР нужно 1-2 неметилированных сайта узнавания [51]. По-видимому, в случае избегания сайта у бактерий речь идет скорее о снижении вероятности взаимодействия между ЭР и ее сайтом узнавания в ДНК хозяина,

наряду с другими способами регуляции активности ЭР типа II (см. Главу 1, раздел 1.1.2.6).

Кроме ЭР, системы P-M содержат и МТазу. Метилирование ДНК может влиять на экспрессию генов [5,6]. Поэтому, возможно, что отрицательный отбор действует на сайты, метилирование которых приводит к изменению экспрессии генов, вредному для бактерий.

В данной работе показано отсутствие недопредставленности сайтов систем P-M типов I, III, и II C/G, несмотря на то, что все они также способны расщеплять неметилированную ДНК и могут влиять на экспресию генов.

Возможно, отсутствие недопредставленности сайтов этих систем может быть объяснено тем, что эти системы менее токсичны или время их жизни в соответствующем геноме меньше. Возможно также, что оба этих фактора действуют одновременно.

Во-первых, ЭР систем P-M типов I, III, и II C/G представляют собой комплекс, включающий МТазную субъединицу. Поэтому в случае потери гена МТазы или потери белком способности взаимодействовать с ДНК, ЭР также потеряет активность. В случае ортодоксальных систем типа II, в которых ЭР и МТаза действуют независимо, при потере гена МТазы произойдет расщепление хозяйской ДНК [173]. Таким образом, можно ожидать, что ортодоксальные системы типа II более токсичны для хозяина, и именно на их сайты в геноме действует отрицательный отбор.

Во-вторых, в системах P-M типов I, III, и II C/G узнавание ДНК осуществляется с помощью отдельного ДНК-узнающего домена (или отдельного S-белка), общего для ЭР и МТазы. Мутации в этом ДНК-узнающем модуле приводят к изменениям в последовательностис сайта узнавания соответствующей системы [170,245]. Механизмы для частого и даже программируемого (ведущего к фазовой вариации) изменения специфичности показаны для систем типов I и III [25,26,170]. Быстрая

эволюция специфичности белков типов IIC/G и IV обсуждается в работах [46,120]. В отличие от всех этих систем, ортодоксальные системы типа II, ЭР и МТаза которых узнают ДНК независимо, для изменения специфичности требуют одновременно двух независимых мутаций в двух разных генах (ЭР и МТаза, соответственно). Можно предположить, что изменение специфичности ортодоксальных систем типа II происходит значительно реже, чем изменение специфичности систем других типов. В результате сайты ортодоксальных систем типа II находятся под действием отрицательного отбора дольше, чем сайты остальных систем, что объясняет их заметную элиминацию в геноме.

3.2.2 Избегание палиндромных и непалиндромных сайтов

Ранее было показано, что короткие палиндромные последовательности наиболее недопредставлены в геномах прокариот [2–4,188]. Этот эффект объясняли влиянием систем Р-М, так как их сайты часто являются палиндромами. Однако многие сайты систем Р-М не являются палиндромами [51].

Сравнение распределений *Kr* в актуальных и контрольных наборах, содержащих палиндромные и непалиндромные сайты ортодоксальных систем типа II показано на рисунке 3.7. Как видно из рисунка 3.7, присутствие систем Р-М в геноме влияет на избегание обоих типов сайтов: распределение *Kr* для наборов пар сайт-геном, где сайт является сайтом системы Р-М, закодированной в данном геноме (актуальные пары) сдвинуто в область недопредставленности как для палиндромных, так и для непалиндромных сайтов (рисунок 3.7А и Б, соответственно).

Оба распределения отличаются от соответствующих распределений *Kr* в прокариотическом контрольном наборе пар сайт-геном, где сайтом является соответственно палиндромный или непалиндромный сайт узнавания системы Р-М вне зависимости от того, закодирована она в геноме или нет. Распределения *Kr* для контрольных множеств у палиндромных и непалиндромных сайтов значительно различаются (см. рисунок 3.8). При используемом в данной работе пороге

недопредставленности *Kr*=0,78, недопредставленность наблюдается в 5,5% (из 373604 случаев) пар прокариотического контроля, содержащих палиндромные сайты и только в 0,4% (из 169307 случаев) пар прокариотического контроля, содержащего непалиндромные сайты. Это различие в проценте недопредставленных пар является статистически значимым (p<<0.001, по критерию χ2).

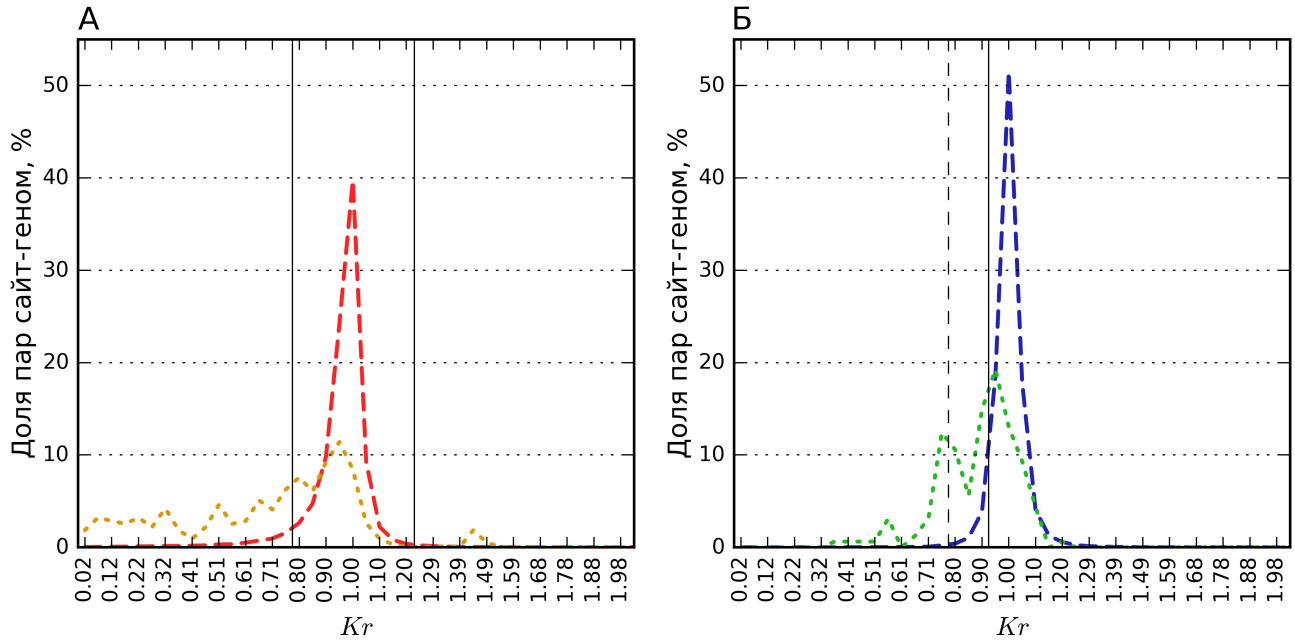


Рисунок 3.7 Распределение *Kr* для палиндромных (А) и непалиндромных (Б) сайтов ортодоксальных систем Р-М. Гистограммы распределения *Kr* для наборов актуальных пар показаны пунктирными оранжевой (А) и зеленой (Б) линиями, для контрольных наборов пар – прерывистой красной (А) и синей (Б) линиями, Границы недопредставленности и перепредставленности показаны вертикальными линиями на рисунке 3.6 А.. На рисунке 3.6 Б показаны две границы недопредставленности, см. пояснения в тексте. Граница перепредставленности не определялась.

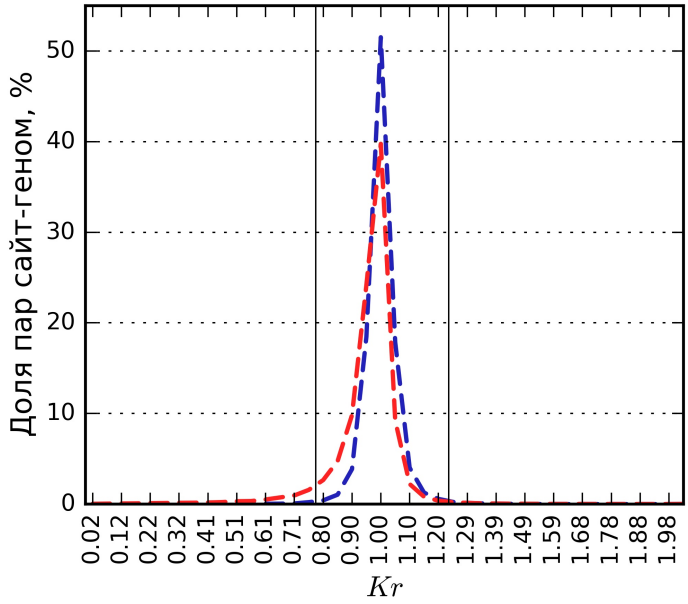


Рисунок 3.8 Распределение *Kr* для палиндромных (красная прерывистая линия) и непалиндромных (синяя прерывистая линия) сайтов ортодоксальных систем Р-М типа II для прокариотических геномов.

Поэтому при оценке влияния систем Р-М узнающих палиндромные и непалиндромные сайты, нужно принять во внимание разницу в распределении *Kr* в контрольных множествах (см. рисунок 3.8). Из-за этого различия один и тот же порог не может быть использован для оценки недопредставленности для обоих типов сайтов. При пороге *Kr*=0.78 недопредставленность наблюдается в 5,5% случаев в контрольном прокариотическом наборе, содержащем палиндромные сайты. Такой же процент недопредставленных сайтов в контрольном прокариотическом наборе, содержащем непалиндромные сайты наблюдается при пороге *Kr*=0,926. При этих порогах в наборе актуальных пар, содержащих палиндромные сайты, недопредставленность наблюдается в 50,5% случаев (порог недопредставленности *Kr*≤0,78) и в актуальных парах, содержащих непалиндромные сайты, недопредставленность наблюдается в 52,5% случаев (порог недопредставленности *Kr*≤0,926).

Для сравнения влияния систем P-M на палиндромные и непалиндромные сайты независимо от выбора порога был построен график зависимости процента недопредставленных пар в актуальном наборе от процента недопредставленных пар в контрольном наборе при различных порогах недопредставленности (см. рисунок 3.9).

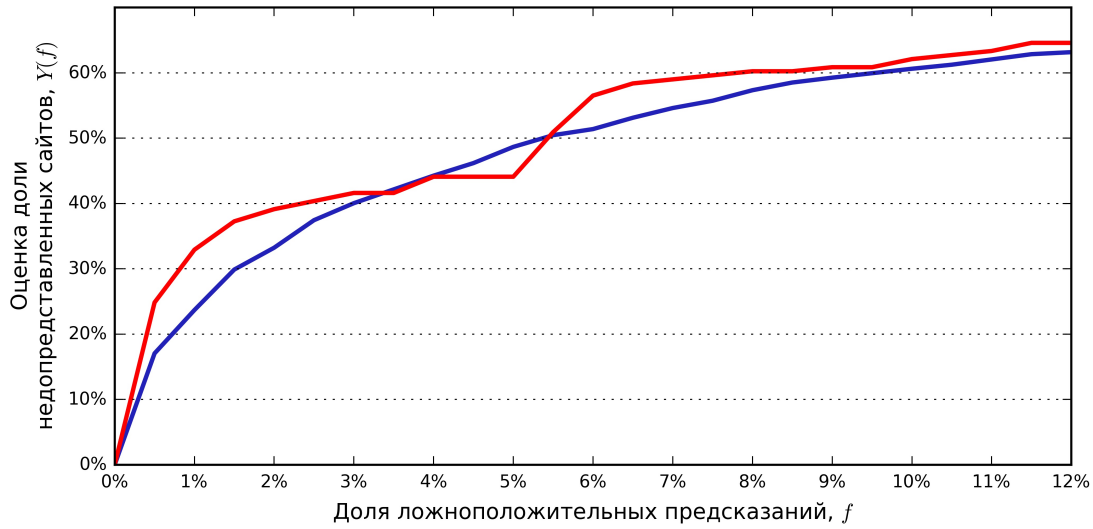


Рисунок 3.9 Графики избегания сайтов из-за присутствия палиндромных (синяя линия) и непалиндромных (красная линия) сайтов. По оси X показана доля *f* недопредставленных сайтов в прокариотическом контрольном наборе (доля ложноположительных результатов). Разница между процентом недопредставленных сайтов в соответствующих наборах актуальных пар и процентом ложноположительных результатов показана на оси Y. В каждой точке графиков граница недопредставленности зафиксирована в величине, дающей f% недопредставленных сайтов в контрольном наборе.

В этом случае процент случаев недопредставленности в контрольном наборе рассматривается как ложноположительный результат для оценки влияния систем P-M на недопредставленность своих сайтов в геноме.

Из рисунка 3.9 видно, что графики для палиндромных и непалиндромных сайтов практически совпадают. Отклонения графика для непалиндромных сайтов могут быть связаны с небольшим количеством актуальных пар, содержащих непалиндромные сайты (161 в сравнении с 1613 парами, содержащими палиндромные сайты).

Анализ влияния таких свойств сайтов как длина и вырожденность на их недопредставленность в геноме не выявил значимой связи этих факторов с

недопредставленностью сайтов.

Полученные данные позволяют сделать вывод, что системы Р-М в равной мере влияют на недопредставленность как палиндромных, так и непалиндромных сайтов. В то же самое время, палиндромные сайты, как правило, избегаются сильнее, чем непалиндромные. Это может быть связано как с особенностями действия систем Р-М, узнающих палиндромные и непалиндромные сайты [51], так и с другими причинами избегания палиндромов, не связанные с действием систем Р-М [2].

3.2.3 Перепредставленные сайты систем Р-М

Было найдено 47 случаев перепредставленности сайтов ортодоксальных систем типа II в актуальном наборе пар (сайт, геном). (см. пик на рисунке 3.5А для области *Kr*~1,4–1,5). В 41 случае речь идет о сайте CCGG и геноме одного из представителей рода *Helicobacter* (38 штаммов *H. pylori*, один штамм *H. acinonychis*, и два штамма *H. cetorum*). Найти объяснение этому феномену пока не удалось.

3.2.4 Влияние продолжительности жизни систем рестрикции-модификации в геноме на недопредставленность палиндромных сайтов

Снижение частоты сайта в геноме является длительным процессом, требующим много времени/поколений бактерий [218]. Поэтому недавно приобретенные системы Р-М могут не иметь достаточно времени для того, чтобы оказать заметное влияние на число своих сайтов в соответствующем геноме. Это предположение было высказано в работе [4] и подтверждено в работе Seshasayee с соавторами [146].

Как обсуждалось выше, недопредставленность характерна, главным образом, только для сайтов ортодоксальных систем типа II. Однако при этом недопредставленность наблюдалась только в половине всех проанализированных

случаев, когда геном кодирует соответствующую ортодоксальную систему типа II. Как показано в данной работе, это не может быть объяснено ошибками предсказания сайтов узнавания.

Отсутствие недопредставленности для актуальных пар может быть связано с тем, что соответствующие системы являются недавно приобретенными в данном геноме. Для того, чтобы избежать влияния различий в распределении *Kr* палиндромных и непалиндромных сайтов, для анализа были использованы только палиндромные сайты длины 4-6.

Поскольку прямо измерить продолжительность жизни систем P-M в геноме невозможно, было использовано несколько различных подходов для выделения групп пар сайт-геном, обогащенных сайтами недавно приобретенных или долгоживущих систем P-M.

В соответствии с гипотезой о влиянии времени жизни системы P-M на недопредставленность ее сайта в геноме, ожидается, что фракция недопредставленных пар будет больше в группе, обогащенной долгоживущими системами P-M.

Во первых, системы P-M были разделены на системы, закодированные на плазмидах или на хромосомах. Поскольку плазмиды являются мобильными элементами генома [246], можно ожидать, что среди систем P-M, закодированных на плазмидах, будет больше недавно приобретенных, чем среди систем P-M, закодированных на хромосомах.

Во-вторых, системы P-M были разделены на частые, которые встречаются во многих геномах одного вида, и редкие, которые редко встречаются в геномах одного вида. Можно предположить, что более широко распространенные системы часто являются долгоживущими.

В-третьих, для определения времени жизни систем P-M в геноме была использована разница в олигонуклеотидном составе фрагментов генома,

кодирующих гены систем Р-М и остального генома. Гены систем Р-М, закодированные в фрагменте, сильно отличающемся по олигонуклеотидному составу от остального генома, вероятно, являются недавно приобретенными путем горизонтального переноса.

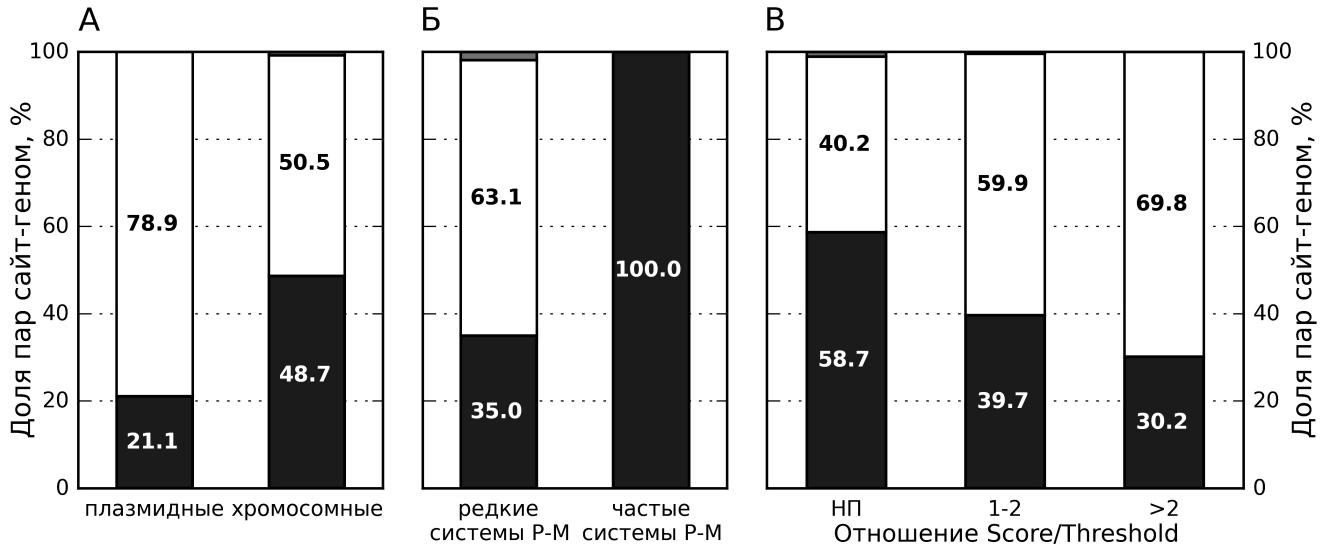


Рисунок 3.10 Доля недопредставленных сайтов среди сайтов предположительно недавно приобретенных и старых систем Р-М в актуальных наборах пар. Показаны фракции недопредставленных (черным), нормально представленных (белым) и перепредставленных (серым). А. Сайты систем Р-М, закодированных на хромосомах и плазмидах; Б. Сайты редких и частых систем Р-М; В. Сайты систем Р-М, закодированных на участках генома с различной степенью чужеродности в геноме по данным программы Alien_hunter (среди пар, которые не относятся ни к редким, ни к частым системам), чем больше отношение Score/Threshold, тем больше данный фрагмент генома отличается по олигонуклеотидному составу, и тем более вероятно, что он недавно попал в геном. НП обозначены пары, включающие сайты систем Р-М, гены которых не закодированы на фрагментах, предсказанных программой Alien_hunter как предположительно недавно приобретенных геномом. Геномы *Helicobacter pylori* не показаны на данных гистограммах (см. текст).

Увеличение фракции недопредставленных пар наблюдалось в группах, предположительно обогащенных долгоживущими системами Р-М, полученных всеми тремя способами (см. рисунок 3.10).

3.2.4.1 Сравнение недопредставленности сайта в геномах, если соответствующая система рестрикции-модификации закодирована на плазмиде или на хромосоме.

Сравнение актуальных наборов пар сайт геном, где соответствующие системы Р-М были закодированы на хромосоме или на плазмиде показало (см. рисунок 3.9А), что сайты систем Р-М, закодированных на плазмиде избегаются в соответствующем геноме в 21,1% случаев (общее число пар в соответствующем

наборе – 38), в то время как сайты систем, закодированных на хромосоме избегаются в 48,7% случаев (общее число пар в соответствующем наборе – 1137). Различие между наборами достоверны по критерию $\chi 2$, $p < 0.001$.

Полученный результат согласуется с предположением, что среди систем Р-М, закодированных на плазмидах, в среднем чаще встречаются недавно приобретенные, чем среди систем Р-М, закоированных на хромосомах.

3.2.4.2 Сравнение недопредставленности сайтов редких и широко распространенных систем рестрикции-модификации

Различные штаммы бактерий одного вида разлчиаются по набору систем Р-М [46]. Некоторые системы Р-М закодированы в большинстве штаммов, другие только в некоторых. В данной работе система обозначена как редкая, если системы с той же специфичностью встречаются менее, чем в 25% штаммов данного вида. Если системы Р-М с данной специфичностью обнаруживаются более, чем в 75% штаммов данного вида, такие системы считались частыми для данного вида. Рассматривались только виды с пятью и более представителями, и сайты только ортодоксальных систем Р-М типа II, закодированных на хромосомах. Представители вида *Helicobacter pylori* были проанализированы отдельно.

Сравнение недопредставленности сайтов редких и частых систем Р-М в актуальных наборах пар сайт – геном показало, что сайты редких систем недопредставлены в 35,0% случаев (общее число проанализированных пар – 103), в то время как сайты частых систем недопредставлены в 100% случаев (общее число проанализированных пар – 45) (см. рисунок 3.10Б). Различие между наборами достоверны по критерию $\chi 2$, $p<0.001$. Эти результаты могут быть объяснены тем, что среди редких систем Р-М чаще встречаются недавно приобретенные, чем среди частых систем Р-М.

H. pylori отличаются от общего правила. В наборе актуальных пар редкие системы Р-М избегаются в 86,5% случаев (общее число пар – 37), сайты частых систем избегаются в 56,6% случаев (общее число пар – 235). Более того, в 16,2%

пар, содержащих сайты частых систем Р-М наблюдается перепредставленность соответствующих сайтов в геноме. Возможно, это связано с тем, что *H. pylori* содержат аномально большое число систем Р-М в геноме [247].

3.2.4.3 Недопредставленность сайтов предположительно недавно перенесенных систем рестрикции-модификации

Предположительно недавно перенесенные с помощью горизонтального переноса генов фрагменты генома были предсказаны с помощью программы Alien_hunter [227]. Сайты систем, закодированных на предположительно недавно перенесенных фрагментах генома недопредставлены в соответствующих геномах реже, чем сайты систем Р-М, которые закодированы на участках генома, не отличающихся по своему олигонуклеотидному составу от остального генома (см. рисунок 3.10В). Эти результаты также подтверждают предположение о влиянии времени жизни систем Р-М в геноме на недопредставленность сайтов этих систем.

Поведение *Helicobacter pylori* отличается обратной зависимостью времени жизни системы и недопредставленности сайта и в данном случае. Сайты систем Р-М, закодированные на предположительно недавно перенесенных участках генома из бегаются в 81,0% случаев (общее число пар – 137), в то время как сайты систем, закодированных на фрагментах, не отличающихся по олигонуклеотидному составу от остального генома недопредставлены в 47,6% случаев (общее число проанализированных пар – 290) и перепредставлены в 12,4% случаев.

Полученные данные свидетельствуют о том, что в наборах данных, обогащенных сайтами недавно приобретенных ортодоксальных систем Р-М типа II, недопредставленность наблюдается значительно реже. Это позволяет предположить, что отсутствие избегания сайтов систем Р-М, закодированных в данном геноме, связано с тем, что данные системы являются недавно приобретенными в геноме.

В группе, содержащей сайты широко распространенных систем, недопредставленность наблюдается в 100% случаев (при этом надо отметить, что

было найдено всего 45 таких случаев). Такое избегание может быть связано с тем, что данные системы Р-М присутствовали у общего предка этих штаммов, и продолжительность их жизни в геномах достаточно длительная. Поэтому этот факт также может свидетельствовать о влиянии времени жизни системы на избегание ее сайтов в соответствующем геноме.

3.2.5 Следы потерянных систем рестрикции-модификации

В геномах прокариот часто недопредставлены сайты систем Р-М, которые закодированы не в данных геномах, а в геномах близких родственников (см. рисунок 3.11). Сайты систем Р-М, которые не закодированы в данном геноме, но закодированы в других геномах того же вида недопредставлены в 43,3% случаев (общее число проанализированных пар — 1930). Сайты систем, закодированные в других геномах того же рода, но не вида недопредставлены в 18,4 % случаев (общее число проанализированных пар — 5162), что больше, чем избегание палиндромов длиной 4 — 6 п.н. а контрольном прокариотическом наборе, которое составляет 6% (общее число проанализированных пар — 339646). Другими словами, близко родственные геномы имеют сходный профиль недопредставленных сайтов, безотносительно систем Р-М, закодированных в каждом конкретном геноме. Эти результаты подтверждают предположение, высказанное в работах [3,4], что недопредставленность сайтов может быть следом потерянной системы Р-М.

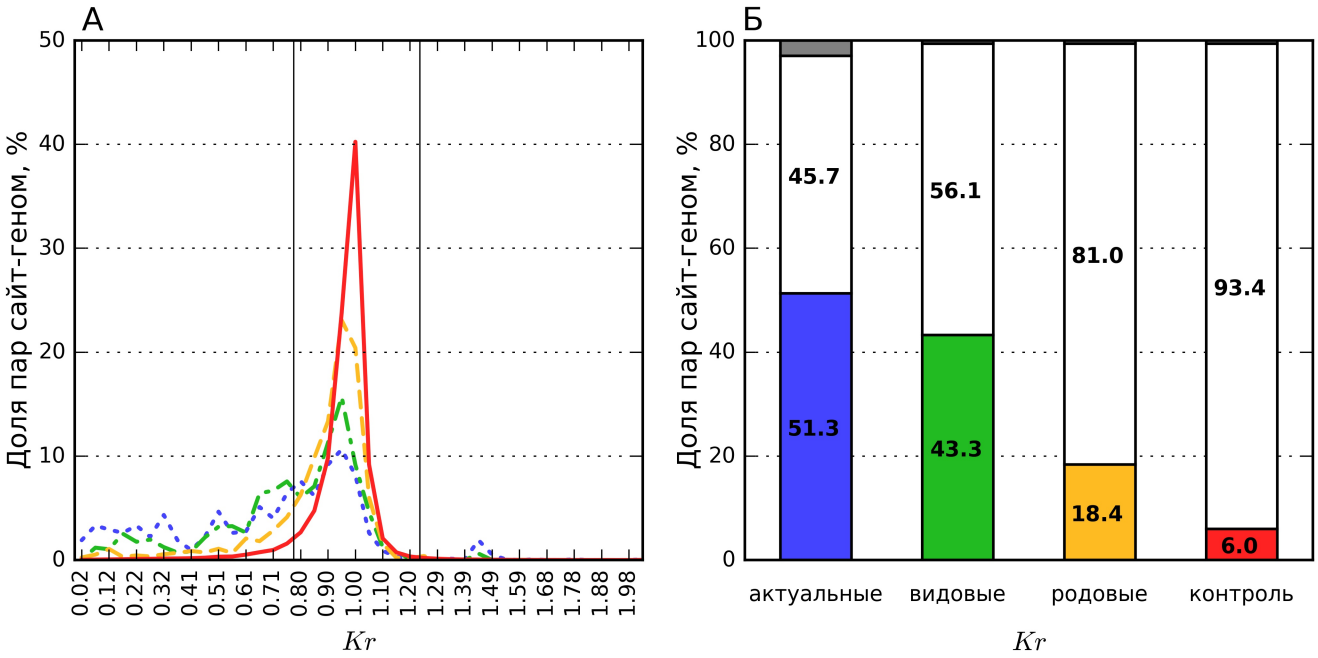


Рисунок 3.11 Избегание в геномах прокариот сайтов систем Р-М, закодированных в близкородственных геномах. Рассмотрены только палиндромные сайты ортодоксальных систем Р-М, длиной 4-6 п.н. А. Распределение *Kr* для актуальных пар сайт-геном показано синим пунктиром, сплошной красной линией показано распределение *Kr* прокариотического контрольного набора. Зеленая прерывистая линия обозначает распределение *Kr* для сайтов систем Р-М, не закодированных в данных геномах, но закодированных в геномах других штаммов того же вида, оранжевая прерывистая линия показывает распределение *Kr* для сайтов систем Р-М в случае, если эти системы Р-М закодированы только в других геномах того же рода, но не вида. Б. Процент недопредставленных сайтов для случаев, описанных на рисунке 3.10А. Цветовые обозначения соответствуют обозначениям на рисунке 3.10А.

Основываясь на данных о влиянии времени жизни систем Р-М в геноме на недопредставленность их сайтов узнавания, а также о недопредставленности сайтов, которые можно интерпретировать как сайты узнавания недавно потерянных систем Р-М, можно предположить следующую схему связи между наличием в геноме системы Р-М и избеганием соответствующего сайта узнавания (см. рисунок 3.12).

Приобретая новую систему Р-М, бактерия получает эффективную защиту от бактериофагов. Благодаря этому преимуществу система Р-М распространяется в бактериальной популяции.

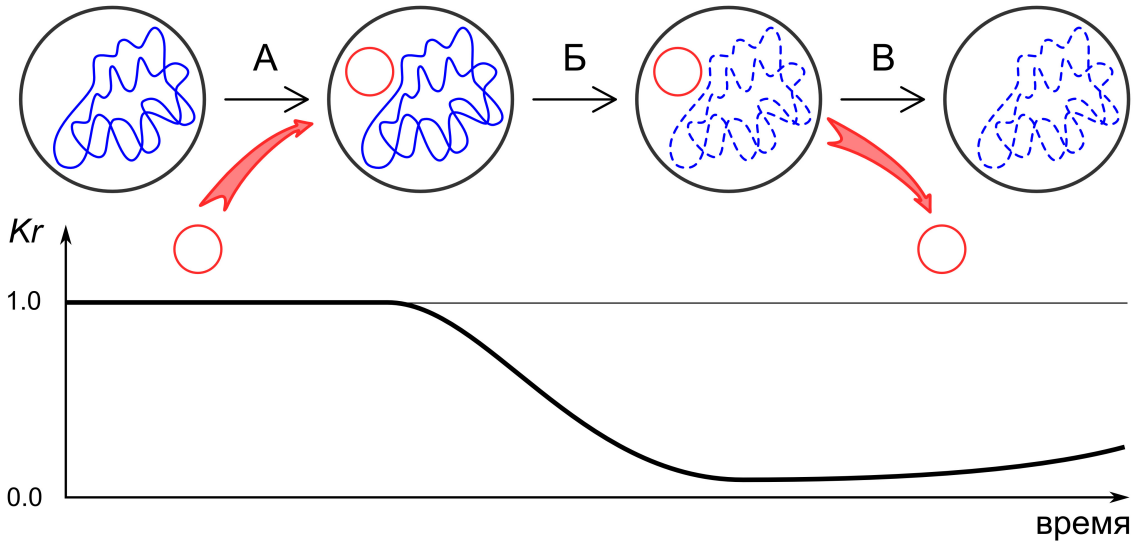


Рисунок 3.12 Схема связи между временем жизни системы Р-М в геноме и недопредставленностью ее сайта.

“А” - обозначает момент приобретения новой системы Р-М, “Б” - токсичность системы Р-М вызывает недопредставленность ее сайта узнавания в геноме, “В” - после потери системы Р-М недопредставленность ее сайта сохраняется. Красный кружок обозначает гены системы Р-М, синяя линияя – геномную ДНК бактерии, прерывистая синяя линия обозначает геномную ДНК с недопредставленными сайтами узнавания системы Р-М.

В то же время, из за токсичности ЭР происходит снижение числа сайтов в геноме бактерии. Тот же процесс элиминации сайтов, возможно, происходит и в геномах бактериофагов, возможно, параллельно с развитием других механизмов антирестрикции. В результате коэволюции бактерий, системы Р-М и бактериофага происходит их взаимная адаптация, в результате чего система Р-М перестает служить эффективной защитой бактерии от фага [10]. Бесполезные гены достаточно быстро элиминируются в бактериальном геноме [237], поэтому гены неэффективных систем Р-М теряются из генома.

Снижение числа сайтов систем Р-М в геноме под давлением систем Р-М происходит медленно и требует большого числа поколений бактерий. После потери генов систем Р-М, негативное давление на ее сайт узнавания исчезает, однако восстановление числа сайтов, вероятно, происходит медленно. Это время может быть сравнимо с изменением олигонуклеотидного состава генов, полученных путем горизонтального переноса генов. Lawrence and Ochman оценивают его как величину порядка сотен миллионов лет [218]. Таким образом,

недопредставленность сайта узнавания в геноме сохраняется даже после потери соответствующей системы P-M.

Количество сайтов, которые являются следами потерянных систем P-M, можно приблизительно оценить следующим образом. В геномах прокариот, которые не кодируют известные системы P-M избегается 0,7 % пар сайт геном (общее число проанализированных пар — 29318), где сайт является сайтом узнавания ортодоксальной системы типа II. Вероятно, эти случаи недопредставленности могут быть связаны с действием других факторов, не связанных с системами P-M. В прокариотическом контрольном множестве недопредставленность наблюдается примерно в 6% таких пар. Таким образом, подавляющее большинство недопредставленных сайтов может быть объяснено следами присутствия систем P-M в геноме.

3.2.6 Выделяющиеся сайты

Было найдено несколько сайтов, чье поведение отличается от остальных сайтов.

Сайт CTAG недопредставлен в 55,4% всех проанализированных геномов, в то время как соответствующая система P-M была найдена только в 0,9% геномов бактерий. Недопредставленность CTAG была также описана в предшествующих работах [2–4]. Учитывая широкое распространение недопредставленности этого сайта, и небольшое количество узнающих его систем P-M, можно предполагать, что недопредставленность последовательности CTAG может быть связана с какими-то другими факторами. Так, в работе [7] недопредставленность последовательности CTAG объясняется возможными структурными особенностями ДНК, действием системы репарации VSP, которая способствует переходу CTAG в CCAG из-за перехода A в G после дезаминирования [7,248,249]. Кроме того, последовательность CTAG часто является сайтом для встройки инсерционных последовательностей [250], и, таким образом, может подвергаться отрицательному отбору для предотвращения экспансии этих элементов в геном.

Сайт CCGG перепредставлен в геномах *Helicobacter pylori* (см. раздел 3.2.3).

Сайт GATC является единственным избегаемым сайтом метил-зависимых ЭР. Поскольку этот сайт также является не только сайтом узнавания систем Р-М, но является сайтом узнавания одиночных МТаз, регулирующих экспрессию генов бактерий, а также сайтом узнавания системы репарации MutHSL, его избегание в геномах прокариот было изучено более детально.

3.2.7 Изучение недопредставленности сайта GATC

Палиндромная последовательность 5'-GATC-3' является сайтом узнавания многих систем Р-М, а также одиночных МТаз. GATC является сайтом узнавания классических систем Р-М, которые включают ДНК-метилтрансферазы, способные метилировать последовательность GATC в N6 или N4 положении аденина или в С5 положении цитозина, и эндонуклеазы рестрикции, способные расщеплять неметилированный сайт. Кроме того, GATC является сайтом узнавания метил-зависимых ЭР типа IIM, которые способны расщеплять последовательность GATC, метилированную в N6 положении аденинового основания [139,251]. МТазы, модифицирующие последовательность GATC в N6 положении аденина и ЭР типа IIM, расщепляющие такие модифицированные последовательности, являются взаимоисключающими, и не могут быть активны в одном организме [128,252]. Представители одного и того же вида могут содержать различные системы Р-М или одиночные МТазы, узнающие сайт GATC. Например, в различных геномах *Streptococcus pneumoniae* найдено три различные системы Р-М, узнающие последовательность GATC: DpnI является ЭР типа IIM, которая расщепляет сайт GATC, метилированный в N6 положении аденина. Система Р-М DpnII состоит из ЭР, способной расщеплять неметилированную последовательность GATC, и двух МТаз DpnM и DpnA, способных метилировать последовательность GATC в N6 положении аденина, в двуцепочечной и одноцепочечной ДНК соответственно. МТаза системы Р-М DpnIII способна метилировать двуцепочечную ДНК в С5 положении цитозина, в то время как

соответствующая ЭР расщепляет неметилированную в этом положении последовательность GATC. Наличие различных систем, специфичных к одной и той же последовательности GATC в одной популяции *S. pneumoniae* способствует лучшей защите от бактериофагов [163] и поддержанию генетического разнообразия за счет ограничения горизонтального переноса генов внутри популяции (см. рисунок 1.7) [25,253].

Кроме МТаз из систем Р-М, известно большое число одиночных МТаз, узнающих последовательность GATC. Метилирование GATC одиночной МТазой Dam играет важную роль в регуляции клеточного цикла и экспрессии генов *E.coli* и некоторых других бактерий [5,254].

Метилирование последовательности GATC также связано с репарацией ошибочно вставленных нуклеотидов системой MutHSL [254,255].

Таким образом, последовательность GATC имеет отношение к множеству различных аспектов жизни прокариот, что может влиять на ее встречаемость в геноме. Разнообразие функций этой последовательности делает ее хорошей моделью для изучения причин избегания палиндромов в геномах прокариот.

Встречаемость сайта GATC в геномах прокариот не была изучена систематически, и не была связана с действием конкретных систем Р-М [2–4]. В работе [146] показано, что, как правило, сайты одиночных МТаз избегаются в геноме реже, чем сайты систем Р-М. Для генома *E.coli*, кодирующего МТазу Dam, показано, что сайты GATC представлены в геноме в том же количестве, которое статистически ожидалось, но расположены кластерами в определенных участках генома [256]. В работе [257] с действием GATC-специфичной МТазы связывается перепредставленность последовательности GCGATCGC в геномах цианобактерий.

В данном разделе предпринят анализ влияния всех известных одиночных МТаз, систем Р-М и ЭР типа IIM на встречаемость последовательности GATC в геномах прокариот, а также других факторов (таких как время жизни системы Р-М в

геноме). На основании полученных данных была построена модель, описывающая влияние различных факторов на встречаемость последовательности GATC в геномах прокариот.

3.2.7.1 Недопредставленность последовательности GATC в геномах прокариот.

Различия в распределении *Kr* для сайта GATC в геномах бактерий, несущих систему P-M, узнающую последовательность GATC, одиночную МТазу или не кодирующих ЭР или МТаз, узнающих GATC, показывают, что наибольшее влияние на недопредставленность GATC в геномах бактерий оказывают соответствующие системы P-M (см. рисунок 3.13).

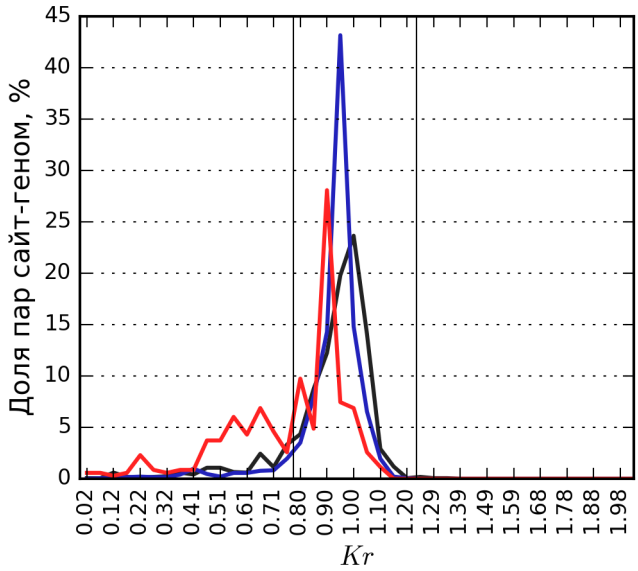


Рисунок 3.13 Распределение *Kr* для последовательности GATC в геномах, которые кодируют GATC-специфичные системы P-M (красный), одиночные МТазы (синий) или не кодируют системы P-M или одиночные МТазы, узнающие данную последовательность (черный).

Последовательность GATC недопредставлена в 164 из 391 геномов, кодирующих соответствующую систему P-M типа II, в 58 из 66 геномов, кодирующих ЭР типа IIM, и в 200 из 3124 геномов, кодирующих одиночную МТазу.

3.2.7.2 Классификация аминокислотных последовательностей по доменному составу

Последовательности белков ЭР и МТаз, которые узнавали сайт GATC были охарактеризованы по составу доменов БД Pfam. Белки, содержащие домены одного семейства по Pfam были объединены в один класс (см. таблицу 3.5).

Таблица 3.5

Семейства систем P-M и одиночных МТаз и ЭР, узнающих сайт GATC.

Семейство белков	#*	Семейство доменов Pfam для МТаз	Тип метилирования**	Семейство доменов Pfam для ЭР	Пример***
Системы P-M типа II					
Meth&MutH	104(92)	DNA_methylase	5mC	MutH	Sau3AI
D12&DpnII	35(29)	MethyltransfD12	N6mA	DpnII	MjaIII
fused_D12/DpnII	9(9)	MethyltransfD12	N6mA	DpnII	MhyORF641P
D12&B743	67(61)	MethyltransfD12	N6mA	Pfam-B_743	Cau700975II
D12&N6_N4&DpnII	88(77)	MethyltransfD12& N6_N4_Mtase	N6mA& N6mA или N4mC	DpnII	DpnII
D12&N6_N4&B743	8(8)	MethyltransfD12; N6_N4_Mtase	N6mA& N6mA или N4mC	Pfam-B_743	BovEORF2878P
N6_N4&DpnII	62(62)	N6_N4_Mtase	N6mA или N4mC	DpnII	HpyAIII
N6_N4&B743	9(9)	N6_N4_Mtase	N6mA или N4mC	Pfam-B_743	Hci611ORF1397P
A70&BglII	13(13)	MT-A70	N4mC	Endonuc-BglII	OgrORF13976P
Эндонуклеаза рестрикции типа IIM					
DpnI	73(64)			DpnI	DpnI
Одиночные ДНК-метилтрансферазы типа II					
D12	1974(1875)	MethyltransfD12	N6mA		M.YpsDam
Dam	629 (550)	Dam	N6mA		M.EcoGVII
A70	312 (290)	MT-A70	N4mC		M.AvaV
Hypoth	249 (245)	Cons_hypoth95	N6mA		M.CklADamP
N6_N4	87(71)	N6_N4_Mtase	N6mA		M.Hpy99VI
Meth	85(65)	DNA_methylase	5mC		M.FpsJIV
B15504	25(16)	Pfam-B_15504	N6mA		M.VchK139I
D12&N6_N4	21(20)	MethyltransfD12 & N6_N4_Mtase	N6mA& N6mA или N4mC		Nme1568OF1755P

* Указано число всех проанализированных белковых последовательностей, в скобках приведено число последовательностей, для которых доступны полные геномы.
** N6mA - метилирование аденинового основания в позиции 6, N4mC - метилирование цитозинового основания в позиции 4, 5mC - метилирование вцитозинового основания в пятой позиции.
***Приведены экспериментально подтвержденные представители класса, если они известны
“&” обозначает белки, гены которых колокализованы, “/” обозначает домены одного полипептида.

3.2.7.3 Недопредставленность GATC в геномах, кодирующих различные семейства белков

Различные классы систем P-M и одиночных МТаз по-разному влияют на недопредставленность последовательности GATC (см. рисунок 3.14). Для анализа были использованы только гены, которые не закодированы на “недавно приобретенных” по данным программы Alien_Hunter участках генома [227] .

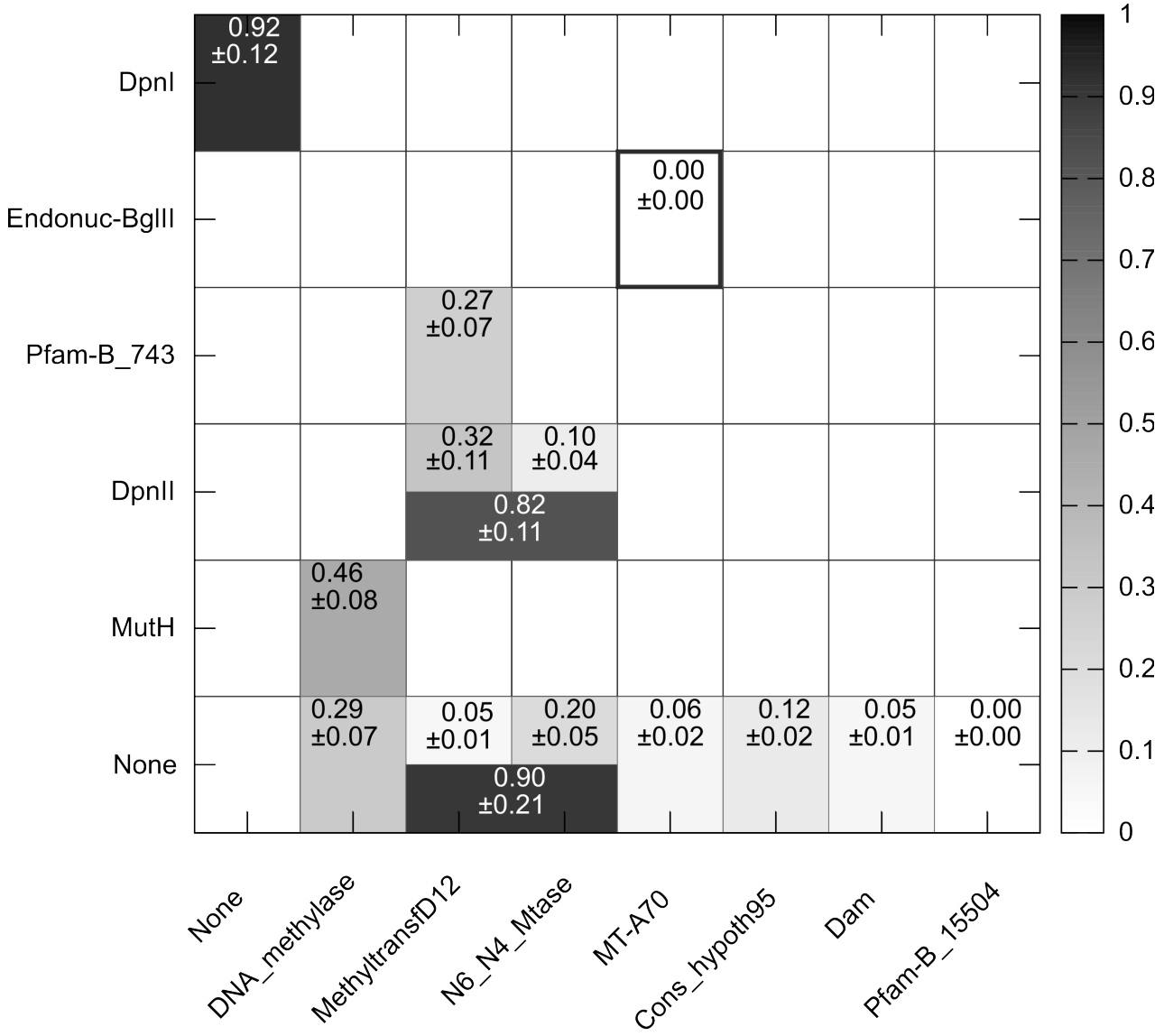


Рисунок 3.14 Недопредставленность GATC в геномах, кодирующих системы P-M и одиночные белки различных семейств из таблицы 3..5. По оси X представлены Pfam семейства МТаз, по оси Y – семейства ЭР. "None" означает, что в геноме не закодировано ЭР или МТазы, узнающей GATC. В каждой клетке таблицы указана доля геномов, кодирующих белки соответствующего семейства, в которых последовательность GATC недопредставлена и оценка погрешности из-за различного числа геномов в каждой группе. В двух случаях (семейства D12&N6_N4&DpnII и D12&N6_N4), геномы кодируют две МТазы, гены которых колокализованы. Такой ситуации соответствуют две объединенные клетки. Цвет клеток показывает фракцию геномов, избегающих последовательность GATC (см. цветовую шкалу справа). Белый цвет клеток, означает, что ни один из геномов, кодирующих белок данного семейства не избегает последовательность GATC, черный цвет – все геномы избегают последовательность GATC.

Из рисунка 3.14 можно сделать три важных наблюдения:

(1) Геномы, кодирующие только одиночные МТазы, избегают последовательность GATC реже, чем геномы, кодирующие системы Р-М. Исключение — геномы, кодирующие колокализованные одиночные МТазы семейства D12&N6_N4.

(2) Геномы, кодирующие одиночные МТазы, метилирующие цитозиновое основание в пятом положении (5mC, содержат домен Pfam DNA_methylase) избегают последовательность GATC в 30% случаев, в то время как геномы, кодирующие одиночные МТазы, метилирующие адениновое основание, избегают GATC в 20% случаев или реже (кроме МТаз семейства D12&N6_N4).

(3) Недопредставленность последовательности GATC наблюдается в 82% геномов, которые кодируют систему Р-М типа II семейства D12&N6_N4&DpnII, 92% геномов, которые кодируют ЭР типа IIM (DpnI), и 90% геномов, которые кодируют одиночные МТазы семейства D12&N6_N4. Избегание последовательности GATC в геномах, кодирующих нетоксичные одиночные МТазы и метил-зависимые ЭР, нуждается в объяснении.

3.2.7.4 Регуляторные одиночные ДНК метилтрансферазы не вызывают недопредставленность GATC

Гены одиночных МТаз встречаются в геномах прокариот примерно в 10 раз чаще, чем гены систем Р-М. Одиночные МТазы могут быть как фрагментами разрушенной системы Р-М, так и так называемыми “сиротскими” или “орфанными” (“orphan”) ДНК-метилтрансферазами, которые выполняют специфические функции, не связанные с действием систем Р-М. Например, такие МТазы осуществляют контроль клеточного цикла у некоторых бактерий. Одиночные МТазы, узнающие последовательность GATC, не вызывают избегания этой последовательности в соответствующих геномах. Единственным исключением являются МТазы семейства D12&N6_N4 (см. рисунок 3.14).

Чтобы оценить влияние регуляторных орфанных МТаз на недопредставленность

GATC в соответствующих геномах, была проанализирована представленность последовательности GATC в геномах бактерий порядков *Enterobacteriales, Vibrionales, Aeromonadales, Pasteurellales* и *Alteromonadales,* которые кодируют МТазу Dam и ее ортологов [258]. Для этих бактерий филогенетическое дерево, построенное на основе сходства последовательностей Dam-подобных МТаз соответствует дереву, построенному на основе сходства 16S РНК [258]. Геномы бактерий этих порядков кодируют только одиночные МТазы семейств D12 и/или Dam и не кодируют GATC-специфичных систем P-M. Последовательность GATC была недопредставлена только в 1.7% (22 of 1238) геномов, относящихся к шести (из 224 проанализированных) видов (*Alteromonas macleodii, Mangrovibacter sp.* MFB070*, Mannheimia haemolytica, Mannheimia varigena, Shewanella amazonensis, Yersinia enterocolitica*). Полученные результаты показывают, что Dam-подобные МТазы не влияют на недопредставленность последовательности GATC в соответствующих геномах. Эти результаты совпадают с результатами, полученными ранее в других работах [256,259].

3.2.7.5 Избегание GATC в геномах, кодирующих ДНК-метилтрансферазы, метилирующие цитозин в 5 положении может быть связано с мутациями метилированного цитозина в тимин

В геномах, кодирующих GATC-специфичные МТазы, метилирующие эту последовательность в шестой позиции аденинового основания (N6mA), последовательность GATC недопредставлена реже, чем в геномах, которые кодируют МТазы, метилирующие цитозиновое основание в пятой позиции (5mC) (см. рисунок 3.15A). Сдвиг распределения *Kr* для сайта GATT в область перепредставленности в геномах, кодирующих МТазы типа 5mC, позволяет, по крайней мере частично, объяснить избегание GATC в таких геномах увеличением вероятности мутаций метилированного цитозина в тимин.

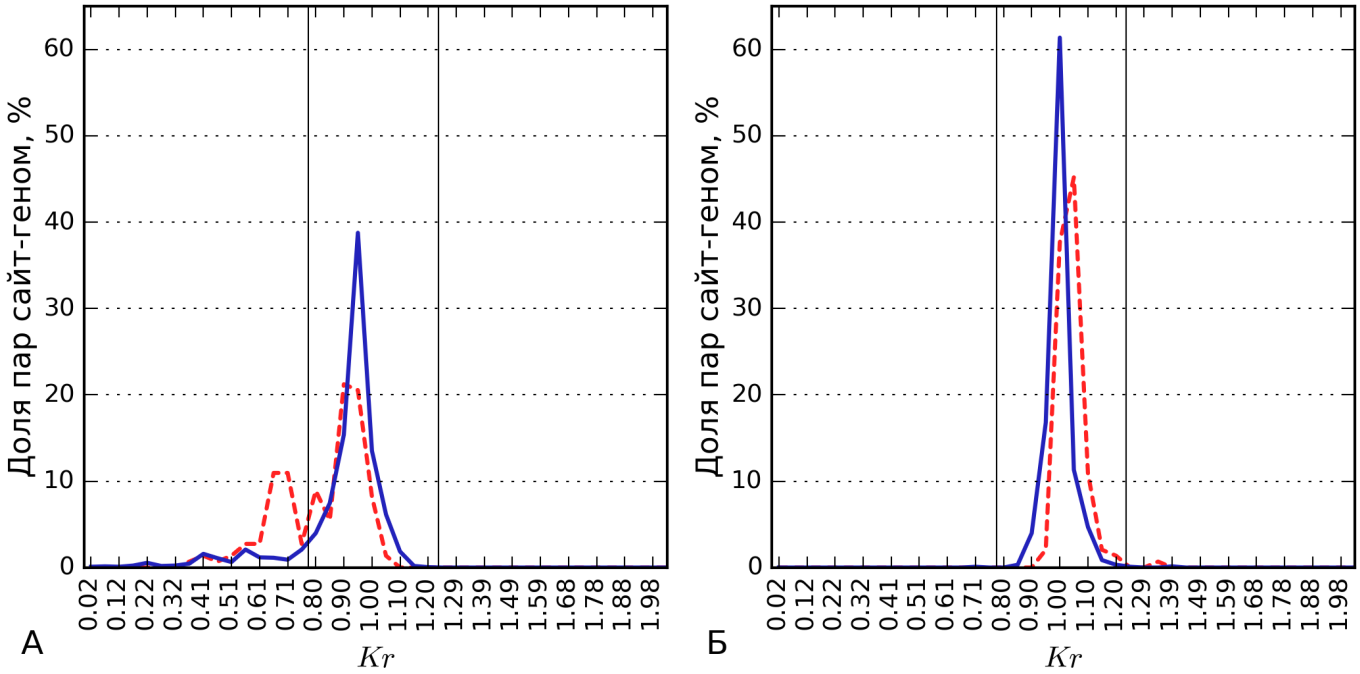


Рисунок 3.15 Распределение Kr для геномов, кодирующих N6mA МТазы (2214 генома) (сплошная синяя линяя) and 5mC МТазы (146 геномов) (красная пунктирная линия). Рассмотрены как одиночные МТазы, так и МТазы, входящие в состав систем Р-М. Границы недо-и перепредставленности показаны сплошными вертикальными линиями. А. Для сайта GATC. Распределения различаются значимо по тесту Колмогорова-Смирнова; В. Для сайта GATT. Распределения различаются значимо по тесту Смирнова.

3.2.7.6 GATC избегается в геномах видов, в которых представлены взаимоисключающие системы рестрикции-модификации

Недопредставленность GATC наблюдается во всех случаях (61 геном 4 неродственных видов *Eubacterium rectale, Moraxella catarrhalis, Neisseria meningitidis, Streptococcus pneumoniae*), когда в различных представителях одного вида встречаются МТазы, метилирующие адениновое основание в N6 позиции и ЭР типа IIM, которые расщепляют метилированную по N6 позиции аденинового основания последовательность GATC (см. таблицу 3.6).

Хотя условия отбора этого не требовали, видно, что в таблице 3.6 представлены те же семейства белков, присутствие которых в геномах связано с наибольшей недопредставленностью GATC (см. рисунок 3.14). Системы Р-М семейств D12&N6_N4&DpnII и N6_N4&DpnII, а также одиночные МТазы семейства D12&N6_N4 способны метилировать адениновое основание последовательности GATC.

Таблица 3.6

Kr и число геномов, кодирующих белки каждого семейства в видах, представители которых кодируют взаимно исключающие системы.

Тип	Вид	$Kr_{ср}$±σ	Семейства систем P-M и одиночных белков				
			Meth&MutH	D12&N6_N4&DpnII	N6_N4&DpnII	D12&N6_N4	DpnI
Firmicutes	*Eubacterium rectale*	0.59±0.01		1			1
Proteobacteria	*Moraxella catarrhalis*	0.55±0.01		1*			2
Proteobacteria	*Neisseria meningitidis*	0.43±0.01				14	14
Firmicutes	*Streptococcus pneumoniae*	0.57±0.01	1	10	2		17

* Этот геном кодирует полноразмерный ген ЭР типа IIM и два колокализованных фрагмента МТаз семейств D12 and N6_N4 Mtases, сходных с МТазами системы MboI (>95% identity).

В свою очередь, метил-зависимые ЭР семейства DpnI расщепляют сайты GATC, в которых метилирован аденин. Поэтому эти системы не могут быть активны в одном геноме [133,201,252]. При этом в геноме, кодирующем N6mA МТазу последовательности GATC метилированы, а в геноме, кодирующем ЭР семейства DpnI – нет.

Рассмотрим возможные причины недопредставленности последовательности GATC у бактерий, кодирующих взаимоисключающие системы.

Системы P-M семейства D12&N6_N4&DpnII могут быть токсичны для бактерии, подобно другим системам P-M типа II [1,2,190]. Таким образом, в случае *S. pneumonia*, *E. rectale* и *M. catarrhalis*, недопредставленность последовательности GATC могла бы быть объяснена токсичностью системы P-M или следами присутствия таких систем. Однако при этом ни одна система P-M с ЭР того же семейства (DpnII) не вызывает недопредставленность GATC в 100% случаев (см.

рисунок 3.13). Кроме того, токсичность систем Р-М не может объяснить наблюдаемую недопредставленность последовательности GATC в геномах *N. meningitides.* Таким образом, недопредставленность последовательности GATC в этих случаях трудно объяснить одной только токсичностью ЭР систем Р-М.

Интересно, что во всех случаях взаимоисключающие системы закодированы в гомологичных областях генома соответствующих штаммов *S. pneumoniae* [260], *N. meningitidis* [261]. Анализ геномного контекста показал, что это также верно для *E. rectale* и *M. catarrhalis*.

S. pneumoniae и *N. meningitidis* известны как бактерии, способные к обмену геномной ДНК и живущие в генетически неоднородных популяциях [181,163,262,263]. При этом наличие взаимоисключающих систем не является барьером для обмена ДНК, т.к. он происходит в форме одноцепочечной ДНК [25,163,260]. Однако, бактерии с взаимоисключающими системами часто принадлежат к различным кладам [25,161], что показывает ограничение обмена ДНК между ними. Учитывая то, что гены взаимоисключающих систем закодированы в гомологичных областях, такое ограничение может быть связано с возможностью внедрения генов системы Р-М в геном с противоположным статусом метилирования. Такое событие будет вредно для генома-реципиента в любом статусе метилирования [133,201,252,264]. Поэтому можно предположить, что недопредставленность GATC в данном случае является адаптацией бактерий к циркуляции в популяции взаимоисключающих систем. Этот способ защиты может дополняться другими способами защиты геномной ДНК от повреждения новой системой Р-М, например, регуляцией экспрессии ЭР [265].

3.2.7.7 Построение математической модели, описывающей влияние систем рестрикции-модификации на недопредставленность GATC в геноме

Геном может содержать гены более чем одной системы Р-М или одиночной МТазы. Для оценки вклада различных систем Р-М в недопредставленность GATC в геноме была построена математическая модель. Влияние всех систем,

закодированных в данном геноме суммировалось. Вклад каждой системы оценивался с учетом ее времени жизни в соответствующем геноме. Системы были поделены на старые и новые по данным Alien_Hunter следующим образом: если отношение score/threshold больше или равно 2, система считалась “новой”, все остальные системы считались “старыми”.

В основе модели лежит предположение, что отклонение *Kr* от 1 является линейной комбинацией следующего вида:

$$1 - Kr = b_1 x_{1,old} + c_1 x_{1,new} + \ldots + b_n x_{n,old} + c_n x_{n,new} + a\,p \qquad (7)$$

где – $x_{1,old} \ldots x_{n,old}, x_{1,new} \ldots x_{n,new}$ – количества систем, относящихся к семействам из таблицы 3.5, представленных в данном геноме. Если какой-то $x_{n,new/old}$ оказывался нулевым у всех рассматриваемых геномов, соответствующее слогаемое в линейной комбинации опускалось.

$b_1 \ldots b_n$ – коэффициент, описывающий влияние этих систем, если они старые в данном геноме;

$c_1 - c_n$ – коэффициент, описывающий влияние этих систем, если они новые в данном геноме;

p-признак отношения к видам, представители которых несут классические системы Р-М и МТазы или ЭР типа IIM;

a – коэффициент, оценивающий вклад этого фактора.

Подбор коэффициентов осуществлялся методом наименьших квадратов (линейной регрессии) на основе данных о *Kr* для сайта GATC 2315 проанализированных геномов прокариот и закодированных в них генах систем Р-М, одиночных МТаз и ЭР типа IIM.

Для оценки качества модели использовалось число ложноположительных и ложноотрицательных предсказаний недопредставленности.

Данная модель позволила описать 51,3% случаев недопредставленности (173 из

337 случаев действительной недопредставленности GATC в геномах), и ложно предсказала недопредставленность в 72 из 236 найденных случаев.

Коэффициенты, полученные в модели отражают закономерности найденные ранее: влияние систем P-M на недопредставленность значительно больше, чем влияние одиночных МТаз, недавно приобретенные системы P-M меньше влияют на недопредставленность. Также модель позволяет оценить вклад каждой системы P-M и одиночной МТазы в недопредставленность, если геном кодирует несколько различных генов. В таблице 3.7 представлены полученные коэффициенты, большие, чем 0,1.

Таблица 3.7

Коэффициенты модели

Название семейства	Коэффициент для “предположительно старых систем”	Величина	Коэффициент для предположительно молодых систем”	Величина
Meth&MutH	b1	0.183	c1	0.105
B200	b13	0.159	c13	nd*
D12&DpnII	b2	0.264	c2	0.084
N6_N4	b20	0.160	c20	0.232
Meth	b21	0.104	c21	0.007
D12&N6_N4	b22	0.429	c22	0.060
DpnI	b24	0.336	c24	0.201
D12&B743	b4	0.120	c4	0.261
D12&N6_N4&DpnII	b5	0.385	c5	0.362
D12&N6_N4&B743	b6	0.264	c6	0.050
N6_N4&DpnII	b7	0.132	c7	0.285
N6_N4&B743	b8	0.590	c8	0.488
p	a	0.138		

*nd означает, что недостаточно данных для расчета коэффициента

При этом модель предсказала только около половины случаев недопредставленности. Остальные случаи, по крайней мере частично, могут быть объяснены следами потерянных систем. Действительно, 146 геномов, недопредставленность в которых не описана данной моделью, относятся к 99 видам, 29 из которых представлены более, чем одним геномом. В 11 из этих 29 видов присутствуют геномы, которые кодируют систему P-M, и сайт GATC в

которых недопредставлен (и эта недопредставленность описывается предложенной моделью). Таким образом, по крайней мере для 16 геномов из 11 видов недопредставленность GATC может быть объяснена как след потерянной системы P-M. Для объяснения недопредставленности остальных 130 геномов в настоящее время недостаточно информации. Можно предположить, что недопредставленность GATC в них связана со следами потерянных систем P-M, наличием неизвестных систем P-M, горизонтальным переносом с неродственными видами, которые содержат комплементарные системы P-M или какими-то факторами, которые связаны с особенностями жизни этих бактерий или особенностями данных систем P-M.

Например, известно, что ЭР различаются по своей эффективности. Так, ЭР BbrIII примерно в 200 раз более эффективна, чем BbrI [266]. Можно предположить, что сайт более эффективной системы может избегаться сильнее. Действительно, в работе [190] показано, что системы P-M EcoRI и EcoRV *E.coli* различаются по эффективности защиты бактрии от бактериофагов. При этом более эффективная система EcoRI является более токсичной для генома хозяина, и ее сайты недопредставлены в геноме *E.coli*.

Кроме того, токсичность данной ЭР для данного генома может зависеть не только от числа, но и от распределения сайтов в геноме. Например, ЭР Ecl18kI, EcoRII и некоторые для гидролиза ДНК нуждаются в нескльких сайтах, расположенных неподалеку друг от друга [67,71]. Для некоторых ЭР эффективность гидролиза сайта зависит от его геномного контекста [89–91].

3.2.8 Заключение по разделу

В данной работе был предпринят анализ доступных в настоящее время тысяч полных геномов прокариот и закодированных в них систем P-M. Показано, что в геномах прокариот избегались только сайты ортодоксальных систем P-M типа II. При этом недопредставленность сайтов систем, закодированных в данном геноме, наблюдалась только в 47,9% всех изученных случаев. Чтобы объяснить отсутствие

недопредставленности остальных сайтов была исследована связь недопредставленности со свойствами сайтов: длиной, палиндромностью и вырожденностью.

Анализ недопредставленности актуальных палиндромных и непалиндромных сайтов ортодоксальных систем Р-М типа II в соответствующих геномах показал, что процент недопредставленных сайтов в обоих наборах сходен, что свидетельствует о том, что системы Р-М в равной степени влияют на недопредставленность как палиндромных, так и непалиндромных сайтов.

В данной работе не было выявлено влияния длины и вырожденности сайтов систем Р-М на их недопредставленность в геномах прокариот.

Показано, что время жизни соответствующих систем Р-М в геномах влияет на недопредставленность их сайтов. В наборах пар сайт-геном, предположительно обогащенных недавно приобретенными системами Р-М, недопредставленность сайтов наблюдалась достоверно реже. Это может объясняться тем, что для того, чтобы отбор на количество сайтов в геноме стал заметен, необходимо время. В таком случае, актуальные сайты систем Р-М, которые не являются недопредставленными в данном геноме, вероятно, являются сайтами недавно приобретенных систем Р-М.

На примере недопредставленности последовательности GATC показано, что избегание сайта может быть адаптацией к горизонтальному переносу генов комплементарных систем между бактериями, имеющими различный статус метилирования.

Выводы

1. Для 57 из 272 одиночных эндонуклеаз рестрикции с помощью методов сравнительной геномики были обнаружены гены парных ДНК метилтрансфераз. Таким образом, показано существование систем Р-М, гены которых не колокализованы, а находятся на значительном расстоянии друг от друга.

Предложен метод систематического предсказания таких систем на основе поиска ортологов.

2. Показано, что сайты ортодоксальных систем R-M типа II, закодированых в геномах прокариот, недопредставлены в данных геномах в 47,9% случаев, в то время как среди всех сайтов ортодоксальных систем R-M в геномах прокариот (контрольная группа) недопредставленность наблюдается только в 3,9% случаев.

3. Найдена связь недопредставленности сайтов систем R-M в геноме прокариот с продолжительностью существования генов этих систем в соответствующем геноме, оцененной тремя различными способами.

4. Сайты систем R-M, которые не закодированы в данном геноме, но закодированы в других геномах того же вида недопредставлены в 43,3% случаев, поэтому недопредставленные сайты могут быть следами потерянных систем R-M в данных геномах.

5. Недопредставленность сайта 5'-GATC-3' наблюдается в 100% случаев, если среди представителей вида встречаются как штаммы, кодирующие системы R-M, расщепляющие сайт 5'-GATC-3' в метилированном состоянии, так и штаммы, кодирующие системы R-M, которые расщепляют этот сайт в неметилированном состоянии.

Список публикаций по теме диссертации

Статьи в рецензируемых журналах:

1. Ershova A., Rusinov I., Vasiliev M., Spirin S., Karyagina A. Restriction-modification systems interplay causes avoidance of GATC site in prokaryotic genomes // Journal of Bioinformatics and Computational Biology.— 2016.— V. 14 (2).— P. 1641003.

2. Rusinov I., Ershova A., Karyagina A., Spirin S., Alexeevski A. Lifespan of restriction-modification systems critically affects avoidance of their recognition sites in host genomes // BMC Genomics.— 2015.— 16(1).— P. 1084.

3. Ершова А.С., Русинов И.С., Спирин С.А., Карягина А.С., Алексеевский А.В. Роль систем рестрикции–модификации в эволюции и экологии прокариот // Биохимия.— 2016.— 81(1).— С. 18—34.

4. Ershova, A.S., Karyagina, A.S., Vasiliev, M.O., Lyashchuk, A.M., Lunin, V.G., Spirin, S. A, Alexeevski, A.V. Solitary restriction endonucleases in prokaryotic genomes // Nucleic acids research.— 2012.— 40 (20)— P. 10107—10115.

Тезисы конференций:

1. Anna Ershova, Ivan Rusinov, Anna Karyagina, Sergei Spirin, Andrei Alexeevski GATC avoidance in bacteria with DpnI/DpnII complementary R-M systems // MCCMB 2015: Proceedings. July 16–19.— Moscow, Russia.— 2015.

2. Rusinov I.S., Ershova A.S., Karyagina A.S., Spirin S.A., Alexeevski A.V., 2014. Comparison of two methods for detection of exceptional words in genomes // The Ninth International Conference on Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS-2014).— Novosibirsk, Russia.— 2014.

3. I.S. Rusinov, A.S. Ershova, A.S. Karyagina, S.A. Spirin, A.V. Alexeevski. Avoidance of restriction sites in genomes of prokaryotes and their viruses // Workshop "Imaging (biology) down to the single molecule level" of the International Research Training Group Gießen/Marburg-Moscow (DFG/RFBR-funded) from 17th to 20th

of September 2014 in Schloss Rauischholzhausen.— Rauischholzhausen, Germany. — 2014.

4. Rusinov I.S., Ershova A.S., Karyagina A.S., Spirin S.A., Alexeevski A.V. Restriction sites avoidance is trace of lost restriction modification systems // The Ninth International Conference on Bioinformatics of Genome Regulation and Structure\Systems Biology (BGRS-2014).— Novosibirsk, Russia.— 2014.

5. Rusinov I.S., Ershova A.S., Karyagina A.S., Spirin S.A., Alexeevski A.V. Underrepresented Words in Prokaryotic Genome Shed Light on Lifespan of Restriction-Modification Systems in Genome // 13th European Conference on Computational Biology (ECCB'14).— Strasbourg, France.— 2014.

6. A.S. Ershova, A.N. Lyaschuk, V.G. Lunin, A.S. Karyagina, S.A. Spirin, A.V. Alexeevski Analysis of predicted separated restriction-modification systems in several bacterial genomes // Workshop “Global and mechanistic approaches in nucleic acid biology” of the International Research Training Group Gießen/Marburg Moscow (DFG/RFBR-funded), 17-20 Sept. 2013.— Sankt-Petersburg, Russia.— 2013.

7. A.S. Ershova, A.S. Karyagina, S.A. Spirin, A.V. Alexeevsky, A.N. Lyaschuk, V.G. Lunin Prediction and analysis of separated restriction-modification systems in prokaryotic genomes// MCCMB 2013: Proceedings, July 25–28.— Moscow, Russia. — 2013.

8. Anna Ershova, Ivan Rusinov, Anna Karyagina, Sergey Spirin, Andrey Alexeevski Occurrence of Restriction-Modification systems recognition sites in genomes of bacteria, archaea and their viruses // MCCMB 2013: Proceedings, July 25–28.— Moscow, Russia.— 2013.

9. I.S. Rusinov, A.S. Ershova, A.S. Karyagina, S.A. Spirin, A.V. Alexeevski Half of actual restriction-modification sites are underrepresented in bacterial and phage genomes//Program &abstracts The 2013 Molecular Genetics of Bacteria and Phages

Meeting, 6-10 August.— University of Wisconsin-Madison Madison, Wisconsin, USA.— 2013.— P. 254.

10. I.S. Rusinov, A.S. Ershova, A.S. Karyagina, S.A. Spirin, A.V. Alexeevski Occurrence of restriction sites in genomes of prokaryotes and their viruses // Offspring-Meeting of the International Research Training Group Gießen/Marburg-Moscow (DFG/RFBR-funded) from 26th to 29th of June 2013.— Moscow, Russia. — 2013.— P. 37.

11. A.S. Ershova, A.N. Lyaschuk, V.G. Lunin, A.S. Karyagina, S.A. Spirin, A.V. Alexeevski Separated restriction-modification systems in prokaryotic genomes // Offspring-Meeting of the International Research Training Group Gießen/Marburg-Moscow (DFG/RFBR-funded) from 26th to 29th of June 2013.— Moscow, Russia. — 2013.— P. 40.

12. A.Ershova, A.Karyagina, S. Spirin, A. Alexeevski. Diversity of the restriction-modification systems in full prokaryotic genomes.// Proceedings of the international Moscow conference on computational molecular biology, July 21-24.— Moscow, Russia.— 2011.— P. 102.

13. M.S.Krivozubov, A.S.Ershova, A.S.Karyagina, S.A. Spirin, A.V. Alexeevski. Occurrence of recognition sites of restriction-modification systems in bacteriophage genomes. // Proceedings of the Five International Conference on Bioinformatics of Genome Regulation and Structure.— Novosibirsk, Russia.— 2008.

14. A.Ershova, A.Karyagina, S. Spirin, A. Alexeevski. Restriction sites avoidance in bacteriophage genomes as a strategy against restriction-modification systems: a whole genome analysis. // Proceedings of the international Moscow conference on computational molecular biology, July 27-31.— Moscow, Russia.— 2007.— P. 83-84.

Список литературы

1. Kobayashi I. Behavior of restriction – modification systems as selfish mobile elements and their impact on genome evolution // Nucleic Acids Res. — 2001. — Vol. 29, № 18. — P. 3742–3756.
2. Karlin S., Mrazek J., Campbell A.M. Compositional Biases of Bacterial Genomes and Evolutionary Implications // J. Bacteriol. — 1997. — Vol. 179, № 12. — P. 3899–3913.
3. Gelfand M.S., Koonin E. V. Avoidance of palindromic words in bacterial and archaeal genomes : a close connection with restriction enzymes // Nucleic Acids Res. — 1997. — Vol. 25, № 12. — P. 2430–2439.
4. Rocha E.P.C., Danchin A., Viari A. Evolutionary Role of Restriction / Modification Systems as Revealed by Comparative Genome Analysis // Genome Res. — 2001. — Vol. 11. — P. 946–958.
5. Marinus M.G., Løbner-Olesen A., Lobner-Olesen A. DNA Methylation // Ecosal Plus. — 2014. — Vol. 2014, № 6. — P. 997–1003.
6. Krebes J., Morgan R.D., Bunk B., et al. The complex methylome of the human gastric pathogen Helicobacter pylori // Nucleic Acids Res. — 2014. — Vol. 42, № 4. — P. 2415–2432.
7. Burge C., Campbell A.M., Karlin S. Over- and under-representation of short oligonucleotides in DNA sequences // Proc. Natl. Acad. Sci. U. S. A. — 1992. — Vol. 89, № 4. — P. 1358–1362.
8. Karlin S., Cardon L.R. Computational DNA sequence analysis // Annu. Rev. Microbiol. — 1994. — Vol. 48. — P. 619–654.
9. Roberts R.J., Belfort M., Bestor T., et al. A nomenclature for restriction enzymes , DNA methyltransferases , homing endonucleases and their genes // Nucleic Acids Res. — 2003. — Vol. 31, № 7. — P. 1805–1812.
10. Tock M.R., Dryden D.T.F. The biology of restriction and anti-restriction // Curr. Opin. Microbiol. — 2005. — Vol. 8, № 4. — P. 466–472.
11. Loenen W.A., Daniel A.S., Braymer H.D., et al. Organization and sequence of the hsd genes of Escherichia coli K-12 // J. Mol. Biol. — 1987. — Vol. 198, № 2. — P. 159–170.
12. Prakash-Cheng A., Ryu J. Delayed expression of in vivo restriction activity following conjugal transfer of Escherichia coli hsdK (restriction-modification) genes // J. Bacteriol. — 1993. — Vol. 175, № 15. — P. 4905–4906.
13. Kulik E.M., Bickle T.A. Regulation of the activity of the type IC EcoR124I restriction enzyme // J. Mol. Biol. — 1996. — Vol. 264, № 5. — P. 891–906.

14. Loenen W.A.M., Dryden D.T.F., Raleigh E.A., et al. Type I restriction enzymes and their relatives // Nucleic Acids Res. — 2014. — Vol. 42, № 1. — P. 20–44.

15. Waldron D.E., Lindsay J. a. Sau1: a novel lineage-specific type I restriction-modification system that blocks horizontal gene transfer into Staphylococcus aureus and between S. aureus isolates of different lineages // J. Bacteriol. — 2006. — Vol. 188, № 15. — P. 5578–5585.

16. Sitaraman R., Dybvig K. The hsd loci of Mycoplasma pulmonis: organization, rearrangements and expression of genes // Mol. Microbiol. — 1997. — Vol. 26, № 1. — P. 109–120.

17. Dybvig K., Sitaraman R., French C.T. A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements // Proc. Natl. Acad. Sci. U. S. A. — 1998. — Vol. 95, № 23. — P. 13923–13928.

18. Murray N.E. Type I Restriction Systems : Sophisticated Molecular Machines (a Legacy of Bertani and Weigle) // Microbiol. Mol. Biol. Rev. — 2000. — Vol. 64, № 2. — P. 412–434.

19. Bourniquel A. a., Bickle T. a. Complex restriction enzymes: NTP-driven molecular motors // Biochimie. — 2002. — Vol. 84, № 11. — P. 1047–1059.

20. Dryden D.T., Murray N.E., Rao D.N. Nucleoside triphosphate-dependent restriction enzymes // Nucleic Acids Res. — 2001. — Vol. 29, № 18. — P. 3728–3741.

21. Kim J.-S., DeGiovanni A., Jancarik J., et al. Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme and its functional implications. // Proc. Natl. Acad. Sci. U. S. A. — 2005. — Vol. 102, № 9. — P. 3248–3253.

22. Gann A.A., Campbell A.J., Collins J.F., et al. Reassortment of DNA recognition domains and the evolution of new specificities // Mol. Microbiol. — 1987. — Vol. 1, № 1. — P. 13–22.

23. Gubler M., Braguglia D., Meyer J., et al. Recombination of constant and variable modules alters DNA sequence recognition by type IC restriction-modification enzymes // EMBO J. — 1992. — Vol. 11, № 1. — P. 233–240.

24. Thorpe P.H., Ternent D., Murray N.E. The specificity of sty SKI, a type I restriction enzyme, implies a structure with rotational symmetry // Nucleic Acids Res. — 1997. — Vol. 25, № 9. — P. 1694–1700.

25. Croucher N.J., Coupland P.G., Stevenson A.E., et al. Diversification of bacterial genome content through distinct mechanisms over different timescales // Nat. Commun. Nature Publishing Group, — 2014. — Vol. 5. — P. 1–12.

26. Manso A.S., Chai M.H., Atack J.M., et al. A random six-phase switch regulates

pneumococcal virulence via global epigenetic changes // Nat. Commun. — 2014. — Vol. 5. — P. 5055.

27. Cerdeño-Tárraga A.M., Patrick S., Crossman L.C., et al. Extensive DNA inversions in the B. fragilis genome control variable gene expression // Science. — 2005. — Vol. 307, № 5714. — P. 1463–1465.

28. Price C., Lingner J., Bickle T.A., et al. Basis for changes in DNA recognition by the EcoR124 and EcoR124/3 type I DNA restriction and modification enzymes // J. Mol. Biol. — 1989. — Vol. 205, № 1. — P. 115–125.

29. Dreier J., MacWilliams M.P., Bickle T.A. DNA cleavage by the type IC restriction-modification enzyme EcoR124II // J. Mol. Biol. — 1996. — Vol. 264, № 4. — P. 722–733.

30. Yuan R., Hamilton D.L., Burckhardt J. DNA translocation by the restriction enzyme from E. coli K // Cell. — 1980. — Vol. 20, № 1. — P. 237–244.

31. Szczelkun M.D., Janscak P., Firman K., et al. Selection of non-specific DNA cleavage sites by the type IC restriction endonuclease EcoR124I // J. Mol. Biol. — 1997. — Vol. 271, № 1. — P. 112–123.

32. Kennaway C.K., Obarska-Kosinska A., White J.H., et al. The structure of M.EcoKI Type I DNA methyltransferase with a DNA mimic antirestriction protein // Nucleic Acids Res. — 2009. — Vol. 37, № 3. — P. 762–770.

33. Su T.-J., Tock M.R., Egelhaaf S.U., et al. DNA bending by M.EcoKI methyltransferase is coupled to nucleotide flipping // Nucleic Acids Res. — 2005. — Vol. 33, № 10. — P. 3235–3244.

34. Suri B., Shepherd J.C., Bickle T.A. The EcoA restriction and modification system of Escherichia coli 15T-: enzyme structure and DNA recognition sequence // EMBO J. — 1984. — Vol. 3, № 3. — P. 575–579.

35. Weiserova M., Janscak P., Benada O., et al. Cloning, production and characterisation of wild type and mutant forms of the R.EcoK endonucleases // Nucleic Acids Res. — 1993. — Vol. 21, № 3. — P. 373–379.

36. Firman K., Dutta K., Weiserova M., et al. The role of subunit assembly in the functional control of Type I Restriction-Modification enzymes // Mol. Biol. Today. — 2000. — Vol. 1, № 2. — P. 35–41.

37. Makovets S., Doronina V.A., Murray N.E. Regulation of endonuclease activity by proteolysis prevents breakage of unmodified bacterial chromosomes by type I restriction enzymes // Proc. Natl. Acad. Sci. U. S. A. — 1999. — Vol. 96, № 17. — P. 9757–9762.

38. Janscak P., Dryden D.T.F., Firman K. Analysis of the subunit assembly of the type IC restriction – modification enzyme EcoR124I // Nucleic Acids Res. — 1998. — Vol. 26, № 19. — P. 25–31.

39. Dryden D.T., Cooper L.P., Thorpe P.H., et al. The in vitro assembly of the EcoKI type I DNA restriction/modification enzyme and its in vivo implications // Biochemistry. — 1997. — Vol. 36, № 5. — P. 1065–1076.

40. Doronina V., Murray N.E. The proteolytic control of restriction activity in Escherichia coli K-12 // Mol. Microbiol. — 2001. — Vol. 39, № 2. — P. 416–428.

41. Makovets S., Powell L.M., Titheradge A.J.B., et al. Is modification sufficient to protect a bacterial chromosome from a resident restriction endonuclease? // Mol. Microbiol. — 2004. — Vol. 51, № 1. — P. 135–147.

42. Cajthamlová K., Sisáková E., Weiser J., et al. Phosphorylation of Type IA restriction-modification complex enzyme EcoKI on the HsdR subunit // FEMS Microbiol. Lett. — 2007. — Vol. 270, № 1. — P. 171–177.

43. McKane M., Milkman R. Transduction, restriction and recombination patterns in Escherichia coli // Genetics. — 1995. — Vol. 139, № 1. — P. 35–43.

44. Holubová I., Vejsadová S., Weiserová M., et al. Localization of the type I restriction-modification enzyme EcoKI in the bacterial cell // Biochem. Biophys. Res. Commun. — 2000. — Vol. 270, № 1. — P. 46–51.

45. Zabeau M., Friedman S., Van Montagu M., et al. The ral gene of phage lambda. I. Identification of a non-essential gene that modulates restriction and modification in E. coli // Mol. Gen. Genet. — 1980. — Vol. 179, № 1. — P. 63–73.

46. Oliveira P.H., Touchon M., Rocha E.P.C. The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts // Nucleic Acids Res. — 2014. — № 21. — P. 1–14.

47. Van Etten J.L., Xia Y.N., Burbank D.E., et al. Chlorella viruses code for restriction and modification enzymes // Gene. — 1988. — Vol. 74, № 1. — P. 113–115.

48. Zhang Y., Nelson M., Nietfeldt J.W., et al. Characterization of Chlorella virus PBCV-1 CviAII restriction and modification system // Nucleic Acids Res. — 1992. — Vol. 20, № 20. — P. 5351–5356.

49. Nelson M., Burbank D.E., Van Etten J.L. Chlorella viruses encode multiple DNA methyltransferases // Biol. Chem. — 1998. — Vol. 379, № 4-5. — P. 423–428.

50. Wilson G.G. Organization of restriction-modification systems // Nucleic Acids Res. — 1991. — Vol. 19, № 10. — P. 2539–2566.

51. Pingoud A., Wilson G.G., Wende W. Type II restriction endonucleases-a historical perspective and more // Nucleic Acids Res. — 2014. — Vol. 42, № 12. — P. 7489–7527.

52. Friedrich T., Fatemi M., Gowhar H., et al. Specificity of DNA binding and methylation by the M.FokI DNA methyltransferase // Biochim. Biophys. Acta. — 2000. — Vol. 1480, № 1-2. — P. 145–159.

53. Pernstich C., Halford S.E. Illuminating the reaction pathway of the FokI restriction endonuclease by fluorescence resonance energy transfer // Nucleic Acids Res. — 2012. — Vol. 40, № 3. — P. 1203–1213.

54. Bitinaite J., Wah D.A., Aggarwal A.K., et al. FokI dimerization is required for DNA cleavage // Proc. Natl. Acad. Sci. U. S. A. — 1998. — Vol. 95, № 18. — P. 10570–10575.

55. Kong H. Analyzing the functional organization of a novel restriction modification system, the BcgI system // J. Mol. Biol. — 1998. — Vol. 279, № 4. — P. 823–832.

56. Kong H., Roemer S.E., Waite-Rees P.A., et al. Characterization of BcgI, a new kind of restriction-modification system // J. Biol. Chem. — 1994. — Vol. 269, № 1. — P. 683–690.

57. Jurenaite-Urbanaviciene S., Serksnaite J., Kriukiene E., et al. Generation of DNA cleavage specificities of type II restriction endonucleases by reassortment of target recognition domains // Proc. Natl. Acad. Sci. U. S. A. — 2007. — Vol. 104, № 25. — P. 10358–10363.

58. Marshall J.J.T., Halford S.E. The type IIB restriction endonucleases // Biochem. Soc. Trans. — 2010. — Vol. 38, № 2. — P. 410–416.

59. Marshall J.J.T., Gowers D.M., Halford S.E. Restriction endonucleases that bridge and excise two recognition sites from DNA // J. Mol. Biol. — 2007. — Vol. 367, № 2. — P. 419–431.

60. Smith R.M., Jacklin A.J., Marshall J.J.T., et al. Organization of the BcgI restriction-modification protein for the transfer of one methyl group to DNA // Nucleic Acids Res. — 2013. — Vol. 41, № 1. — P. 405–417.

61. Smith R.M., Jacklin A.J., Marshall J.J.T., et al. Organization of the BcgI restriction – modification protein for the transfer of one methyl group to DNA. — 2013. — Vol. 41, № 1. — P. 405–417.

62. Marshall J.J.T., Smith R.M., Ganguly S., et al. Concerted action at eight phosphodiester bonds by the BcgI restriction endonuclease // Nucleic Acids Res. — 2011. — Vol. 39, № 17. — P. 7630–7640.

63. Morgan R.D., Dwinell E.A., Bhatia T.K., et al. The MmeI family: type II restriction-modification enzymes that employ single-strand modification for host protection // Nucleic Acids Res. — 2009. — Vol. 37, № 15. — P. 5208–5221.

64. Janulaitis A., Petrusyte M., Maneliene Z., et al. Purification and properties of the Eco57I restriction endonuclease and methylase--prototypes of a new class (type IV) // Nucleic Acids Res. — 1992. — Vol. 20, № 22. — P. 6043–6049.

65. Sarrade-Loucheur A., Xu S., Chan S. The Role of the Methyltransferase Domain of Bifunctional Restriction Enzyme RM.BpuSI in Cleavage Activity // PLoS One. — 2013. — Vol. 8, № 11. — P. e80967.

66. Mucke M., Kruger D.H., Reuter M. Diversity of Type II restriction endonucleases that require two DNA recognition sites // Nucleic Acids Res. — 2003. — Vol. 31, № 21. — P. 6079–6084.

67. Szczepek M., Mackeldanz P., Möncke-Buchner E., et al. Molecular analysis of restriction endonuclease EcoRII from Escherichia coli reveals precise regulation of its enzymatic activity by autoinhibition // Mol. Microbiol. — 2009. — Vol. 72, № 4. — P. 1011–1021.

68. Wentzell L.M., Nobbs T.J., Halford S.E. The SfiI restriction endonuclease makes a four-strand DNA break at two copies of its recognition sequence // J. Mol. Biol. — 1995. — Vol. 248, № 3. — P. 581–595.

69. Siksnys V., Skirgaila R., Sasnauskas G., et al. The Cfr10I restriction enzyme is functional as a tetramer // J. Mol. Biol. — 1999. — Vol. 291, № 5. — P. 1105–1118.

70. Khan F., Furuta Y., Kawai M., et al. A putative mobile genetic element carrying a novel type IIF restriction-modification system (PluTI) // Nucleic Acids Res. — 2010. — Vol. 38, № 9. — P. 3019–3030.

71. Zaremba M., Owsicka A., Tamulaitis G., et al. DNA synapsis through transient tetramerization triggers cleavage by Ecl18kI restriction enzyme // Nucleic Acids Res. — 2010. — Vol. 38, № 20. — P. 7142–7154.

72. Morgan R.D., Bhatia T.K., Lovasco L., et al. MmeI: a minimal Type II restriction-modification system that only modifies one DNA strand for host protection // Nucleic Acids Res. — 2008. — Vol. 36, № 20. — P. 6558–6570.

73. Marks P., McGeehan J., Wilson G., et al. Purification and characterisation of a novel DNA methyltransferase, M.AhdI // Nucleic Acids Res. — 2003. — Vol. 31, № 11. — P. 2803–2810.

74. Kaus-Drobek M., Czapinska H., Sokołowska M., et al. Restriction endonuclease MvaI is a monomer that recognizes its target sequence asymmetrically // Nucleic Acids Res. — 2007. — Vol. 35, № 6. — P. 2035–2046.

75. Niv M.Y., Ripoll D.R., Vila J.A., et al. Topology of Type II REases revisited; structural classes and the common conserved core // Nucleic Acids Res. — 2007. — Vol. 35, № 7. — P. 2227–2237.

76. Gowers D.M., Bellamy S.R.W., Halford S.E. One recognition sequence, seven restriction enzymes, five reaction mechanisms // Nucleic Acids Res. — 2004. — Vol. 32, № 11. — P. 3469–3479.

77. Szybalski W., Kim S.C., Hasan N., et al. Class-IIS restriction enzymes--a review // Gene. — 1991. — Vol. 100. — P. 13–26.

78. Li L., Wu L.P., Chandrasegaran S. Functional domains in Fok I restriction endonuclease // Proc. Natl. Acad. Sci. U. S. A. — 1992. — Vol. 89, № 10. — P.

4275–4279.

79. Xu S.-Y., Zhu Z., Zhang P., et al. Discovery of natural nicking endonucleases Nb.BsrDI and Nb.BtsI and engineering of top-strand nicking variants from BsrDI and BtsI // Nucleic Acids Res. — 2007. — Vol. 35, № 14. — P. 4608–4618.

80. Halford S.E., Johnson N.P., Grinsted J. The reactions of the EcoRI and other restriction endonucleases // Biochem. J. — 1979. — Vol. 179, № 2. — P. 353–365.

81. Modrich P., Rubin R.A. Role of the 2-amino group of deoxyguanosine in sequence recognition by EcoRI restriction and modification enzymes // J. Biol. Chem. — 1977. — Vol. 252, № 20. — P. 7273–7278.

82. Berkner K.L., Folk W.R. EcoRI cleavage and methylation of DNAs containing modified pyrimidines in the recogintion sequence // J. Biol. Chem. — 1977. — Vol. 252, № 10. — P. 3185–3193.

83. Bheemanaik S., Reddy Y.V.R., Rao D.N. Structure, function and mechanism of exocyclic DNA methyltransferases // Biochem. J. — 2006. — Vol. 399, № 2. — P. 177–190.

84. Malygin E.G., Evdokimov A.A., Hattman S. Dimeric / oligomeric DNA methyltransferases : an unfinished story // Biol. Chem. — 2009. — Vol. 390. — P. 835–844.

85. Madhusoodanan U.K., Rao D.N. Diversity of DNA methyltransferases that recognize asymmetric target sequences // Crit. Rev. Biochem. Mol. Biol. — 2010. — Vol. 45, № 2. — P. 125–145.

86. Pein C.D., Cech D., Gromova E.S., et al. Interaction of the MvaI restriction enzyme with synthetic DNA fragments // Nucleic Acids Symp. Ser. — 1987. — № 18. — P. 225–228.

87. Ibryashkina E.M., Sasnauskas G., Solonin A.S., et al. Oligomeric structure diversity within the GIY-YIG nuclease family // J. Mol. Biol. — 2009. — Vol. 387, № 1. — P. 10–16.

88. Gasiunas G., Sasnauskas G., Tamulaitis G., et al. Tetrameric restriction enzymes: expansion to the GIY-YIG nuclease family // Nucleic Acids Res. — 2008. — Vol. 36, № 3. — P. 938–949.

89. Alves J., Pingoud A., Haupt W., et al. The influence of sequences adjacent to the recognition site on the cleavage of oligodeoxynucleotides by the EcoRI endonuclease // Eur. J. Biochem. — 1984. — Vol. 140, № 1. — P. 83–92.

90. Van Cleve M.D., Gumport R.I. Influence of enzyme-substrate contacts located outside the EcoRI recognition site on cleavage of duplex oligodeoxyribonucleotide substrates by EcoRI endonuclease // Biochemistry. — 1992. — Vol. 31, № 2. — P. 334–339.

91. Taylor J.D., Halford S.E. The activity of the EcoRV restriction endonuclease is influenced by flanking DNA sequences both inside and outside the DNA-protein complex // Biochemistry. — 1992. — Vol. 31, № 1. — P. 90–97.

92. Greene P.H., Poonian M.S., Nussbaum A.L., et al. Restriction and modification of a self-complementary octanucleotide containing the EcoRI substrate // J. Mol. Biol. — 1975. — Vol. 99, № 2. — P. 237–261.

93. Dickerson R.E., Drew H.R. Structure of a B-DNA dodecamer. II. Influence of base sequence on helix structure // J. Mol. Biol. — 1981. — Vol. 149, № 4. — P. 761–786.

94. Beletskaya I. V, Zakharova M. V, Shlyapnikov M.G., et al. DNA methylation at the CfrBI site is involved in expression control in the CfrBI restriction-modification system // Nucleic Acids Res. — 2000. — Vol. 28, № 19. — P. 3817–3822.

95. Zakharova M., Minakhin L., Solonin A., et al. Regulation of RNA polymerase promoter selectivity by covalent modification of DNA // J. Mol. Biol. — 2004. — Vol. 335, № 1. — P. 103–111.

96. Christensen L.L., Josephsen J. The methyltransferase from the LlaDII restriction-modification system influences the level of expression of its own gene // J. Bacteriol. — 2004. — Vol. 186, № 2. — P. 287–295.

97. Kita K., Kotani H., Sugisaki H., et al. The FokI restriction-modification system. I. Organization and nucleotide sequences of the restriction and modification genes // J. Biol. Chem. — 1989. — Vol. 264, № 10. — P. 5751–5756.

98. Mruk I., Rajesh P., Blumenthal R.M. Regulatory circuit based on autogenous activation-repression: roles of C-boxes and spacer sequences in control of the PvuII restriction-modification system // Nucleic Acids Res. — 2007. — Vol. 35, № 20. — P. 6935–6952.

99. Karyagina A., Shilov I., Tashlitskii V., et al. Specific binding of sso II DNA methyltransferase to its promoter region provides the regulation of sso II restriction-modification gene expression // Nucleic Acids Res. — 1997. — Vol. 25, № 11. — P. 2114–2120.

100. Som S., Friedman S. Regulation of EcoRII methyltransferase: effect of mutations on gene expression and in vitro binding to the promoter region // Nucleic Acids Res. — 1994. — Vol. 22, № 24. — P. 5347–5353.

101. Som S., Friedman S. Characterization of the intergenic region which regulates the MspI restriction-modification system // J. Bacteriol. — 1997. — Vol. 179, № 3. — P. 964–967.

102. Mruk I., Kobayashi I. To be or not to be: Regulation of restriction-modification systems and other toxin-antitoxin systems // Nucleic Acids Res. — 2014. — Vol.

42, № 1. — P. 70–86.

103. Liu Y., Kobayashi I. Negative regulation of the EcoRI restriction enzyme gene is associated with intragenic reverse promoters // J. Bacteriol. — 2007. — Vol. 189, № 19. — P. 6928–6935.

104. Mruk I., Liu Y., Ge L., et al. Antisense RNA associated with biological regulation of a restriction-modification system // Nucleic Acids Res. — 2011. — Vol. 39, № 13. — P. 5622–5632.

105. Nagornykh M., Zakharova M., Protsenko A., et al. Regulation of gene expression in restriction-modification system Eco29kI // Nucleic Acids Res. — 2011. — Vol. 39, № 11. — P. 4653–4663.

106. Mücke M., Reich S., Möncke-Buchner E., et al. DNA cleavage by type III restriction-modification enzyme EcoP15I is independent of spacer distance between two head to head oriented recognition sites // J. Mol. Biol. — 2001. — Vol. 312, № 4. — P. 687–698.

107. Janscak P., Sandmeier U., Szczelkun M.D., et al. Subunit assembly and mode of DNA cleavage of the type III restriction endonucleases EcoP1I and EcoP15I // J. Mol. Biol. — 2001. — Vol. 306, № 3. — P. 417–431.

108. Sharrocks A.D., Hornby D.P. Transcriptional analysis of the restriction and modification genes of bacteriophage P1 // Mol. Microbiol. — 1991. — Vol. 5, № 3. — P. 685–694.

109. Rao D.N., Dryden D.T.F., Bheemanaik S. Type III restriction-modification enzymes : a historical perspective // Nucleic Acids Res. — 2014. — Vol. 42, № 1. — P. 45–55.

110. Butterer A., Pernstich C., Smith R.M., et al. Type III restriction endonucleases are heterotrimeric: comprising one helicase-nuclease subunit and a dimeric methyltransferase that binds only one specific DNA // Nucleic Acids Res. — 2014. — Vol. 42, № 8. — P. 5139–5150.

111. Wyszomirski K.H., Curth U., Alves J., et al. Type III restriction endonuclease EcoP15I is a heterotrimeric complex containing one Res subunit with several DNA-binding regions and ATPase activity // Nucleic Acids Res. — 2012. — Vol. 40, № 8. — P. 3610–3622.

112. Reich S., Gössl I., Reuter M., et al. Scanning force microscopy of DNA translocation by the Type III restriction enzyme EcoP15I // J. Mol. Biol. — 2004. — Vol. 341, № 2. — P. 337–343.

113. Hadi S.M., Bächi B., Shepherd J.C., et al. DNA recognition and cleavage by the EcoP15 restriction endonuclease // J. Mol. Biol. — 1979. — Vol. 134, № 3. — P. 655–666.

114. Bächi B., Reiser J., Pirrotta V. Methylation and cleavage sequences of the EcoP1

restriction-modification enzyme // J. Mol. Biol. — 1979. — Vol. 128, № 2. — P. 143–163.

115. Brockes J.P. The deoxyribonucleic acid-modification enzyme of bacteriophage P1. Subunit structure // Biochem. J. — 1973. — Vol. 133, № 4. — P. 629–633.
116. Saha S., Rao D.N. ATP hydrolysis is required for DNA cleavage by EcoPI restriction enzyme // J. Mol. Biol. — 1995. — Vol. 247, № 4. — P. 559–567.
117. Ahmad I., Krishnamurthy V., Rao D.N. DNA recognition by the EcoP15I and EcoPI modification methyltransferases // Gene. — 1995. — Vol. 157, № 1-2. — P. 143–147.
118. Arber, W., Yuan, R. and Bickle T.. Strain-specific modification and restriction of DNA in bacteria // FEBS Proc.Symp. — 1975. — Vol. 9. — P. 3–22.
119. Redaschi N., Bickle T.A. Posttranscriptional regulation of EcoP1I and EcoP15I restriction activity // J. Mol. Biol. — 1996. — Vol. 257, № 4. — P. 790–803.
120. Loenen W.A.M., Raleigh E.A. The other face of restriction: modification-dependent enzymes // Nucleic Acids Res. — 2014. — Vol. 42, № 1. — P. 56–69.
121. Zheng Y., Cohen-karni D., Xu D., et al. A unique family of Mrr-like modification-dependent restriction endonucleases // Nucleic Acids Res. — 2010. — Vol. 38, № 16. — P. 5527–5534.
122. O'Sullivan D.J., Klaenhammer T.R. Control of expression of LlaI restriction in Lactococcus lactis // Mol. Microbiol. — 1998. — Vol. 27, № 5. — P. 1009–1020.
123. Whitaker R.D., Dorner L.F., Schildkraut I. A mutant of BamHI restriction endonuclease which requires N6-methyladenine for cleavage // J. Mol. Biol. — 1999. — Vol. 285, № 4. — P. 1525–1536.
124. Raleigh E.A., Wilson G. Escherichia coli K-12 restricts DNA containing 5-methylcytosine // Proc. Natl. Acad. Sci. U. S. A. — 1986. — Vol. 83, № 23. — P. 9070–9074.
125. Xu S.-Y., Corvaglia A.R., Chan S.-H., et al. A type IV modification-dependent restriction enzyme SauUSI from Staphylococcus aureus subsp. aureus USA300 // Nucleic Acids Res. — 2011. — Vol. 39, № 13. — P. 5597–5610.
126. Mulligan E.A., Hatchwell E., McCorkle S.R., et al. Differential binding of Escherichia coli McrA protein to DNA sequences that contain the dinucleotide m5CpG // Nucleic Acids Res. — 2010. — Vol. 38, № 6. — P. 1997–2005.
127. Krüger T., Wild C., Noyer-Weidner M. McrB: a prokaryotic protein specifically recognizing DNA containing modified cytosine residues // EMBO J. — 1995. — Vol. 14, № 11. — P. 2661–2669.
128. Lacks S.A. Purification and properties of the complementary endonucleases DpnI and DpnII // Methods Enzymol. — 1980. — Vol. 65, № 1. — P. 138–146.

129. Wei H., Therrien C., Blanchard A., et al. The Fidelity Index provides a systematic quantitation of star activity of DNA restriction endonucleases // Nucleic Acids Res. — 2008. — Vol. 36, № 9. — P. e50–e50.

130. Siwek W., Czapinska H., Bochtler M., et al. Crystal structure and mechanism of action of the N6-methyladenine-dependent type IIM restriction endonuclease R.DpnI // Nucleic Acids Res. — 2012. — Vol. 40, № 15. — P. 7563–7572.

131. Cohen-Karni D., Xu D., Apone L., et al. The MspJI family of modification-dependent restriction endonucleases for epigenetic studies // Proc. Natl. Acad. Sci. U. S. A. — 2011. — Vol. 108, № 27. — P. 11040–11045.

132. Horton J.R., Wang H., Mabuchi M.Y., et al. Modification-dependent restriction endonuclease, MspJI, flips 5-methylcytosine out of the DNA helix // Nucleic Acids Res. — 2014. — Vol. 42, № 19. — P. 12092–12101.

133. Ishikawa K., Fukuda E., Kobayashi I. Conflicts targeting epigenetic systems and their resolution by cell death: novel concepts for methyl-specific and other restriction systems // DNA Res. — 2010. — Vol. 17, № 6. — P. 325–342.

134. Williams R.J. Restriction endonucleases: classification, properties, and applications // Mol. Biotechnol. — 2003. — Vol. 23, № 3. — P. 225–243.

135. Kulakauskas S., Lubys A., Ehrlich S.D. DNA restriction-modification systems mediate plasmid maintenance // J. Bacteriol. — 1995. — Vol. 177, № 12. — P. 3451–3454.

136. Furuta Y., Abe K., Kobayashi I. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements // Nucleic Acids Res. — 2010. — Vol. 38, № 7. — P. 2428–2443.

137. Kita K., Kawakami H., Tanaka H. Evidence for horizontal transfer of the EcoT38I restriction-modification gene to chromosomal DNA by the P2 phage and diversity of defective P2 prophages in Escherichia coli TH38 strains // J. Bacteriol. — 2003. — Vol. 185, № 7. — P. 2296–2305.

138. Takahashi N., Ohashi S., Sadykov M.R., et al. IS-linked movement of a restriction-modification system // PLoS One. — 2011. — Vol. 6, № 1. — P. e16554.

139. Burrus V., Bontemps C., Decaris B., et al. Characterization of a novel type II restriction-modification system, Sth368I, encoded by the integrative element ICESt1 of Streptococcus thermophilus CNRZ368 // Appl. Environ. Microbiol. — 2001. — Vol. 67, № 4. — P. 1522–1528.

140. Kobayashi I., Nobusato A., Kobayashi-Takahashi N., et al. Shaping the genome--restriction-modification systems as mobile genetic elements // Curr. Opin. Genet. Dev. — 1999. — Vol. 9, № 6. — P. 649–656.

141. Rowe-Magnus D.A., Guerout A.M., Ploncard P., et al. The evolutionary history of

chromosomal super-integrons provides an ancestry for multiresistant integrons // Proc. Natl. Acad. Sci. U. S. A. — 2001. — Vol. 98, № 2. — P. 652–657.

142. Jeltsch A., Pingoud A. Horizontal Gene Transfer Contributes to the Wide Distribution and Evolution of Type II Restriction-Modification Systems // New York. — 1996. — P. 91–96.

143. Naito T., Kusano K., Kobayashi I. Selfish behavior of restriction-modification systems // Science. — 1995. — Vol. 267, № 5199. — P. 897–899.

144. Makarova K.S., Wolf Y.I., Snir S., et al. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems // J. Bacteriol. — 2011. — Vol. 193, № 21. — P. 6039–6056.

145. Dupuis M.-È., Villion M., Magadán A.H., et al. CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance // Nat. Commun. — 2013. — Vol. 4. — P. 2087.

146. Seshasayee N.S.A., Singh P., Krishna S. Context-dependent conservation of DNA methyltransferases in bacteria // Nucleic Acids Res. — 2012. — Vol. 40, № 15. — P. 7066–7073.

147. Takahashi N., Naito Y., Handa N., et al. A DNA Methyltransferase Can Protect the Genome from Postdisturbance Attack by a Restriction-Modification Gene Complex // Society. — 2002. — Vol. 184, № 22. — P. 6100–6108.

148. Kuroda M., Ohta T., Uchiyama I., et al. Whole genome sequencing of meticillin-resistant Staphylococcus aureus // Lancet. — 2001. — Vol. 357, № 9264. — P. 1225–1240.

149. Schouler C., Gautier M., Ehrlich S.D., et al. Combinational variation of restriction modification specificities in Lactococcus lactis // Mol. Microbiol. — 1998. — Vol. 28, № 1. — P. 169–178.

150. Dandekar T., Huynen M., Regula J.T., et al. Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames // Nucleic Acids Res. — 2000. — Vol. 28, № 17. — P. 3278–3288.

151. Murphy J., Mahony J., Ainsworth S., et al. Bacteriophage orphan DNA methyltransferases: insights from their bacterial origin, function, and occurrence // Appl. Environ. Microbiol. — 2013. — Vol. 79, № 24. — P. 7547–7555.

152. Schlagman S.L., Hattman S., Marinus M.G. Direct role of the Escherichia coli Dam DNA methyltransferase in methylation-directed mismatch repair // J. Bacteriol. — 1986. — Vol. 165, № 3. — P. 896–900.

153. Boye E., Løbner-Olesen A. The role of dam methyltransferase in the control of DNA replication in E. coli // Cell. — 1990. — Vol. 62, № 5. — P. 981–989.

154. Palmer B.R., Marinus M.G. The dam and dcm strains of Escherichia coli--a

review // Gene. — 1994. — Vol. 143, № 1. — P. 1–12.

155. W. Arber D.D. Host specificity of DNA produced by Escherichia coli. I. Host controlled modification of bacteriophage lambda // J Mol Biol. — 1962. — Vol. 5. — P. 18–36.

156. Arber W., Dussoix D. Host specificity of DNA produced by Escherichia coli. I. Host-controlled modification of bacteriophage lambda // J. Mol. Biol. — 1962. — Vol. 5. — P. 18–36.

157. Bickle T. a, Krüger D.H. Biology of DNA restriction // Microbiol. Rev. — 1993. — Vol. 57, № 2. — P. 434–450.

158. Bickle T.A. Restricting restriction // Mol. Microbiol. — 2004. — Vol. 51, № 1. — P. 3–5.

159. Loenen W.A.M. Tracking EcoKI and DNA fifty years on: a golden story full of surprises // Nucleic Acids Res. — 2003. — Vol. 31, № 24. — P. 7059–7069.

160. Korona R., Korona B., Levin B.R. Sensitivity of naturally occurring coliphages to type I and type II restriction and modification // J. Gen. Microbiol. — 1993. — Vol. 139 Pt 6. — P. 1283–1290.

161. Budroni S., Siena E., Hotopp J.C.D., et al. Neisseria meningitidis is structured in clades associated with restriction modification systems that modulate homologous recombination // Proc. Natl. Acad. Sci. U. S. A. — 2011. — Vol. 108, № 11. — P. 4494–4499.

162. Frank S.A. Polymorphism of bacterial restriction-modification systems: the advantage of diversity // Evolution (N. Y). — 1994. — Vol. 48, № 5. — P. 1470–1477.

163. Johnston C., Polard P., Claverys J.-P. The DpnI/DpnII pneumococcal system, defense against foreign attack without compromising genetic exchange // Mob. Genet. Elements. — 2013. — Vol. 3, № 4. — P. e25582.

164. Weyler L., Engelbrecht M., Mata Forsberg M., et al. Restriction endonucleases from invasive Neisseria gonorrhoeae cause double-strand breaks and distort mitosis in epithelial cells during infection // PLoS One. — 2014. — Vol. 9, № 12. — P. e114208.

165. Van Etten J.L. Unusual life style of giant chlorella viruses // Annu. Rev. Genet. — 2003. — Vol. 37. — P. 153–195.

166. Srikhanta Y.N., Fox K.L., Jennings M.P. The phasevarion : phase variation of type III DNA methyltransferases controls coordinated switching in multiple genes // Nat. Rev. Microbiol. Nature Publishing Group, — 2010. — Vol. 8, № 3. — P. 196–206.

167. Low D.A., Casadesu J. Review Clocks and switches : bacterial gene regulation by

DNA adenine methylation // Curr. Opin. Microbiol. — 2008. — Vol. 2.

168. Collier J. Epigenetic regulation of the bacterial cell cycle // Curr. Opin. Microbiol. — 2009. — Vol. 12, № 6. — P. 722–729.

169. Murray I.A., Clark T.A., Morgan R.D., et al. The methylomes of six bacteria // Nucleic Acids Res. — 2012. — Vol. 40, № 22. — P. 11450–11462.

170. Furuta Y., Namba-fukuyo H., Shibata T.F., et al. Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity // PLoS Genet. — 2014. — Vol. 10, № 4. — P. e1004272.

171. Mou K.T., Muppirala U.K., Severin A.J., et al. A comparative analysis of methylome profiles of Campylobacter jejuni sheep abortion isolate and gastroenteric strains using PacBio data // Front. Microbiol. — 2014. — Vol. 5. — P. 782.

172. Gauntlett J.C., Nilsson H., Fulurija A., et al. Phase-variable restriction / modification systems are required for Helicobacter pylori colonization // Gut Pathog. — 2014. — Vol. 6. — P. 35.

173. Handa N., Kobayashi I. Post-segregational killing by restriction modification gene complexes: observations of individual cell deaths // Biochimie. — 1999. — Vol. 81, № 8-9. — P. 931–938.

174. Ichige A., Kobayashi I. Stability of EcoRI restriction-modification enzymes in vivo differentiates the EcoRI restriction-modification system from other postsegregational cell killing systems. // J. Bacteriol. — 2005. — Vol. 187, № 19. — P. 6612–6621.

175. Mochizuki A., Yahara K., Kobayashi I., et al. Genetic addiction: selfish gene's strategy for symbiosis in the genome // Genetics. — 2006. — Vol. 172, № 2. — P. 1309–1323.

176. Kusano K., Naito T., Handa N., et al. Restriction-modification systems as genomic parasites in competition for specific sequences // Proc. Natl. Acad. Sci. U. S. A. — 1995. — Vol. 92, № 24. — P. 11095–11099.

177. Chinen A., Naito Y., Handa N., et al. Evolution of Sequence Recognition by Restriction-Modification Enzymes: Selective Pressure for Specificity Decrease // Mol. Biol. Evol. — 2000. — Vol. 17, № 11. — P. 1610–1619.

178. Nakayama Y., Kobayashi I. Restriction-modification gene complexes as selfish gene entities: roles of a regulatory system in their establishment, maintenance, and apoptotic mutual exclusion // Proc. Natl. Acad. Sci. U. S. A. — 1998. — Vol. 95, № 11. — P. 6442–6447.

179. O'Neill M., Chen A., Murray N.E. The restriction-modification genes of Escherichia coli K-12 may not be selfish: they do not resist loss and are readily replaced by alleles conferring different specificities // Proc. Natl. Acad. Sci. U. S.

A. — 1997. — Vol. 94, № 26. — P. 14596–14601.

180. Handa N., Nakayama Y., Sadykov M. Experimental genome evolution : large-scale genome rearrangements associated with resistance to replacement of a chromosomal restriction - modification gene complex // Mol. Microbiol. — 2001. — Vol. 40. — P. 932–940.

181. Hamilton H.L., Dillard J.P. Natural transformation of Neisseria gonorrhoeae: from DNA donation to homologous recombination // Mol. Microbiol. — 2006. — Vol. 59, № 2. — P. 376–385.

182. Humbert O., Salama N.R. The Helicobacter pylori HpyAXII restriction-modification system limits exogenous DNA uptake by targeting GTAC sites but shows asymmetric conservation of the DNA methyltransferase and restriction endonuclease components. // Nucleic Acids Res. — 2008. — Vol. 36, № 21. — P. 6893–6906.

183. Vasu K., Nagaraja V. Diverse functions of restriction-modification systems in addition to cellular defense // Microbiol. Mol. Biol. Rev. — 2013. — Vol. 77, № 1. — P. 53–72.

184. Arber W., Kühnlein U. Mutational loss of the B-specific restriction in bacteriophage fd. // Pathol. Microbiol. (Basel). — 1967. — Vol. 30, № 6. — P. 946–952.

185. Krüger D.H., Bickle T.A. Bacteriophage survival: multiple mechanisms for avoiding the deoxyribonucleic acid restriction systems of their hosts // Microbiol. Rev. — 1983. — Vol. 47, № 3. — P. 345–360.

186. Sharp P.M. Molecular Evolution of Bacteriophages : Evidence of Selection against the Recognition Sites of Host Restriction Enzymes // Mol. Biol. Evol. — 1986. — Vol. 3, № 1. — P. 75–83.

187. Blaisdell B.E., Campbell A.M., Karlin S. Similarities and dissimilarities of phage genomes // Proc. Natl. Acad. Sci. U. S. A. — 1996. — Vol. 93, № 12. — P. 5854–5859.

188. Fuglsang A. Distribution of potential type II restriction sites (palindromes) in prokaryotes // Biochem. Biophys. Res. Commun. — 2003. — Vol. 310, № 2. — P. 280–285.

189. Lamprea-Burgunder E., Ludin P., Mäser P. Species-specific typing of DNA based on palindrome frequency patterns // DNA Res. — 2011. — Vol. 18, № 2. — P. 117–124.

190. Pleška M., Qian L., Okura R., et al. Bacterial Autoimmunity Due to a Restriction-Modification System // Curr. Biol. — 2016. — Vol. 26, № 3. — P. 404–409.

191. Sadykov M., Asami Y., Niki H., et al. Multiplication of a restriction – modification gene complex // Mol. Microbiol. — 2003. — Vol. 48, № 2. — P. 417–427.

192. Asakura Y., Kojima H., Kobayashi I. Evolutionary genome engineering using a restriction-modification system // Nucleic Acids Res. — 2011. — Vol. 39, № 20. — P. 9034–9046.

193. Chang S., Cohen S.N. In vivo site-specific genetic recombination promoted by the EcoRI restriction endonuclease // Proc. Natl. Acad. Sci. U. S. A. — 1977. — Vol. 74, № 11. — P. 4811–4815.

194. Price C., Bickle T.A. A possible role for DNA restriction in bacterial evolution // Microbiol. Sci. — 1986. — Vol. 3, № 10. — P. 296–299.

195. Kusano K., Sakagami K., Yokochi T., et al. A new type of illegitimate recombination is dependent on restriction and homologous interaction // J. Bacteriol. — 1997. — Vol. 179, № 17. — P. 5380–5390.

196. Casadesús J., Low D.A. Programmed heterogeneity: epigenetic mechanisms in bacteria // J. Biol. Chem. — 2013. — Vol. 288, № 20. — P. 13929–13935.

197. Oliveira P.H., Touchon M., Rocha E.P.C. Regulation of genetic flux between bacteria by restriction-modification systems. // Proc. Natl. Acad. Sci. U. S. A. — 2016. — Vol. 113, № 20. — P. 5658–5663.

198. Heuer H., Smalla K. Horizontal gene transfer between bacteria // Environ. Biosaf. Res. — 2007. — Vol. 6. — P. 3–13.

199. Johnston C., Martin B., Fichant G., et al. Bacterial transformation: distribution, shared mechanisms and divergent control // Nat. Rev. Microbiol. — 2014. — Vol. 12, № 3. — P. 181–196.

200. Roer L., Aarestrup F.M., Hasman H. The EcoKI type I restriction-modification system in Escherichia coli affects but is not an absolute barrier for conjugation // J. Bacteriol. — 2015. — Vol. 197, № 2. — P. 337–342.

201. Fukuda E., Kaminska K.H., Bujnicki J.M., et al. Cell death upon epigenetic genome methylation: a novel function of methyl-specific deoxyribonucleases // Genome Biol. — 2008. — Vol. 9, № 11. — P. R163.

202. Kunz A., Mackeldanz P., Mücke M., et al. Mutual activation of two restriction endonucleases: interaction of EcoP1 and EcoP15 // Biol. Chem. — 1998. — Vol. 379, № 4-5. — P. 617–620.

203. Белогуров А.А., Ефимова Е.П., Дельвер Е.П. З.Г.Б. Ослабление рестрикции типа I у Escherichia coli: действие мутации dam // Молекулярная генетика, микробиология, вирусология. — 1987. — Vol. 9. — P. 10–16.

204. Roberts R.J., Vincze T., Posfai J., et al. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes // Nucleic Acids Res. — 2010. — Vol. 38, № Database issue. — P. D234–D236.

205. Wolf Y.I., Novichkov P.S., Karev G.P., et al. The universal distribution of

evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages // Proc. Natl. Acad. Sci. U. S. A. — 2009. — Vol. 106, № 18. — P. 7273–7280.

206. Koonin E. V. Orthologs, Paralogs, and Evolutionary Genomics 1 // Annu. Rev. Genet. — 2005. — Vol. 39. — P. 309–338.
207. Wolf Y.I., Rogozin I.B., Kondrashov a S., et al. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context // Genome Res. — 2001. — Vol. 11, № 3. — P. 356–372.
208. Mirkin B., Muchnik I., Smith T.F. A biologically consistent model for comparing molecular phylogenies // J. Comput. Biol. — 1995. — Vol. 2, № 4. — P. 493–507.
209. Tatusov R.L., Koonin E. V, Lipman D.J. A genomic perspective on protein families // Science. — 1997. — Vol. 278, № 5338. — P. 631–637.
210. Altschul S.F., Madden T.L., Schäffer A.A., et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. — 1997. — Vol. 25, № 17. — P. 3389–3402.
211. Eddy S.R. Accelerated Profile HMM Searches // PLoS Comput. Biol. — 2011. — Vol. 7, № 10. — P. e1002195.
212. Roberts R.J., Vincze T., Posfai J., et al. REBASE--a database for DNA restriction and modification: enzymes, genes and genomes // Nucleic Acids Res. — 2015. — Vol. 43, № Database issue. — P. D298–D299.
213. Roberts R.J., Vincze T., Posfai J., et al. REBASE--enzymes and genes for DNA restriction and modification // Nucleic Acids Res. — 2007. — Vol. 35, № Database issue. — P. D269–D270.
214. Malone T., Blumenthal R.M., Cheng X., et al. Structure-guided Analysis Reveals Nine Sequence Motifs Conserved among DNA Amino-methyl- transferases , and Suggests a Catalytic Mechanism for these Enzymes // J. Mol. Biol. — 1995. — Vol. 253. — P. 618–632.
215. Zheng Y., Posfai J., Morgan R.D., et al. Using shotgun sequence data to find active restriction enzyme genes // Nucleic Acids Res. — 2009. — Vol. 37, № 1. — P. e1.
216. Sueoka N. Directional mutation pressure, mutator mutations, and dynamics of molecular evolution // J. Mol. Evol. — 1993. — Vol. 37, № 2. — P. 137–153.
217. Fleischmann R.D., Adams M.D., White O., et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd // Science. — 1995. — Vol. 269, № 5223. — P. 496–512.
218. Lawrence J.G., Ochman H. Amelioration of Bacterial Genomes : Rates of Change and Exchange // J. Mol. Evol. — 1997. — Vol. 44. — P. 383–397.
219. Schbath S., Prum B., de Turckheim E. Exceptional motifs in different Markov

chain models for a statistical analysis of DNA sequences // J. Comput. Biol. — 1995. — Vol. 2, № 3. — P. 417–437.

220. Elhai J. Determination of Bias in the Relative Abundance of Oligonucleotides in DNA Sequences // J. Comput. Biol. — 2001. — Vol. 8, № 2. — P. 151–175.
221. NCBI, National Center for Biotechnology Information [Electronic resource]. URL: ftp://ftp.ncbi.nih.gov/genomes/ (accessed: 15.02.2010).
222. REBASE [Electronic resource]. URL: http://tools.neb.com/genomes/ (accessed: 02.02.2010).
223. REBASE [Electronic resource]. URL: http://tools.neb.com/genomes/ (accessed: 15.04.2015).
224. NCBI, National Center for Biotechnology Information [Electronic resource]. URL: ftp://ftp.ncbi.nih.gov/genbank/ (accessed: 03.03.2015).
225. Lluch-Senar M., Luong K., Lloréns-Rico V., et al. Comprehensive methylome characterization of Mycoplasma genitalium and Mycoplasma pneumoniae at single-base resolution // PLoS Genet. — 2013. — Vol. 9, № 1. — P. e1003191.
226. Frost L.S., Leplae R., Summers A.O., et al. Mobile genetic elements: the agents of open source evolution // Nat. Rev. Microbiol. — 2005. — Vol. 3, № 9. — P. 722–732.
227. Vernikos G.S., Parkhill J. Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the Salmonella pathogenicity islands // Bioinformatics. — 2006. — Vol. 22, № 18. — P. 2196–2203.
228. Finn R.D., Bateman A., Clements J., et al. Pfam: the protein families database // Nucleic Acids Res. — 2014. — Vol. 42, № Database issue. — P. D222–D230.
229. Gingeras T.R., Brooks J.E. Cloned restriction/modification system from Pseudomonas aeruginosa // Proc. Natl. Acad. Sci. U. S. A. — 1983. — Vol. 80, № 2. — P. 402–406.
230. Theriault G., Roy P.H., Howard K.A., et al. Nucleotide sequence of the PaeR7 restriction/modification system and partial characterization of its protein products // Nucleic Acids Res. — 1985. — Vol. 13, № 23. — P. 8441–8461.
231. Vasu K., Nagamalleswari E., Nagaraja V. Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria // Proc. Natl. Acad. Sci. U. S. A. — 2012. — Vol. 109, № 20. — P. E1287–E1293.
232. Dorman C.J. Nucleoid-associated proteins and bacterial physiology // Adv. Appl. Microbiol. — 2009. — Vol. 67. — P. 47–64.
233. Shen B.W., Heiter D.F., Chan S.-H., et al. Unusual target site disruption by the rare-cutting HNH restriction endonuclease PacI // Structure. — 2010. — Vol. 18, № 6. — P. 734–743.

234. Taylor J.D., Goodall A.J., Vermote C.L., et al. Fidelity of DNA recognition by the EcoRV restriction/modification system in vivo // Biochemistry. — 1990. — Vol. 29, № 48. — P. 10727–10733.

235. Hiom K.J., Sedgwick S.G. Alleviation of EcoK DNA restriction in Escherichia coli and involvement of umuDC activity // Mol. Gen. Genet. — 1992. — Vol. 231, № 2. — P. 265–275.

236. O'Driscoll J., Heiter D.F., Wilson G.G., et al. A genetic dissection of the LlaJI restriction cassette reveals insights on a novel bacteriophage resistance system // BMC Microbiol. — 2006. — Vol. 6. — P. 40.

237. Kuo C.-H., Ochman H. The extinction dynamics of bacterial pseudogenes // PLoS Genet. — 2010. — Vol. 6, № 8.

238. Tsuru T., Kawai M., Mizutani-Ui Y., et al. Evolution of paralogous genes: Reconstruction of genome rearrangements through comparison of multiple genomes within Staphylococcus aureus // Mol. Biol. Evol. — 2006. — Vol. 23, № 6. — P. 1269–1285.

239. Parkhill J., Sebaihia M., Preston A., et al. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica // Nat. Genet. — 2003. — Vol. 35, № 1. — P. 32–40.

240. Wright R., Stephens C., Shapiro L. The CcrM DNA methyltransferase is widespread in the alpha subdivision of proteobacteria, and its essential functions are conserved in Rhizobium meliloti and Caulobacter crescentus // J. Bacteriol. — 1997. — Vol. 179, № 18. — P. 5869–5877.

241. Stoddard B.L. Homing endonuclease structure and function // Q. Rev. — 2005.

242. Yu G.X., Snyder E.E., Boyle S.M., et al. A versatile computational pipeline for bacterial genome annotation improvement and comparative analysis, with Brucella as a use case // Nucleic Acids Res. — 2007. — Vol. 35, № 12. — P. 3953–3962.

243. Zhu Z., Pedamallu C.S., Fomenkov A., et al. Cloning of NruI and Sbo13I restriction and modification sstems in E. coli and amino acid sequence comparison of M.NruI and M.Sbo13I with other amino-methyltransferases // BMC Res. Notes. — 2010. — Vol. 3. — P. 139.

244. Hughes D. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes // Genome Biol. — 2000. — Vol. 1, № 6. — P. REVIEWS0006.

245. Furuta Y., Kawai M., Uchiyama I., et al. Domain movement within a gene: a novel evolutionary mechanism for protein diversification // PLoS One. — 2011. — Vol. 6, № 4. — P. e18819.

246. Smillie C., Garcillan-Barcia M.P., Francia M. V., et al. Mobility of Plasmids //

Microbiol. Mol. Biol. Rev. — 2010. — Vol. 74, № 3. — P. 434–452.

247. Lin L., Posfai J., Roberts R.J., et al. Comparative genomics of the restriction-modification systems in Helicobacter pylori // Proc. Natl. Acad. Sci. U. S. A. — 2001. — Vol. 98, № 5. — P. 2740–2745.

248. Bhagwat A.S., Mcclelland M. DNA mismatch correction by Very Short Patch repair may have altered the abundance of oligonucleotides in the E . coli genome // Nucleic Acids Res. — 1992. — Vol. 20, № 7. — P. 1663–1668.

249. Merkl R., Kröger M., Rice P., et al. Statistical evaluation and biological interpretation of non-random abundance in the E. coli K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair // Nucleic Acids Res. — 1992. — Vol. 20, № 7. — P. 1657–1662.

250. Mahillon J., Chandler M. Insertion sequences // Microbiol. Mol. Biol. Rev. — 1998. — Vol. 62, № 3. — P. 725–774.

251. Hermann A., Jeltsch A. Methylation sensitivity of restriction enzymes interacting with GATC sites // Biotechniques. — 2003. — Vol. 34, № 5. — P. 924–926, 928, 930.

252. Lacks S., Springhorn S.S. Transfer of recombinant plasmids containing the gene for DpnII DNA methylase into strains of Streptococcus pneumoniae that produce DpnI or DpnII restriction endonucleases // J. Bacteriol. — 1984. — Vol. 158, № 3. — P. 905–909.

253. Eutsey R.A., Powell E., Dordel J., et al. Genetic Stabilization of the Drug-Resistant PMEN1 Pneumococcus Lineage by Its Distinctive DpnIII Restriction-Modification System // MBio. — 2015. — Vol. 6, № 3. — P. e00173–15.

254. Barras F., Marinus M.G. The great GATC: DNA methylation in E. coli // Trends Genet. — 1989. — Vol. 5, № 5. — P. 139–143.

255. Marinus M.G., Casadesus J. Roles of DNA adenine methylation in host-pathogen interactions: mismatch repair, transcriptional regulation, and more // FEMS Microbiol. Rev. — 2009. — Vol. 33, № 3. — P. 488–503.

256. Hénaut A., Rouxel T., Gleizes A., et al. Uneven distribution of GATC motifs in the Escherichia coli chromosome, its plasmids and its phages // J. Mol. Biol. — 1996. — Vol. 257. — P. 574–585.

257. Elhai J. Highly Iterated Palindromic Sequences (HIPs) and Their Relationship to DNA Methyltransferases // Life. — 2015. — Vol. 5, № 1. — P. 921–948.

258. Løbner-Olesen A., Skovgaard O., Marinus M.G. Dam methylation: coordinating cellular processes // Curr. Opin. Microbiol. — 2005. — Vol. 8, № 2. — P. 154–160.

259. Barras F., Marinus M.G. Arrangement of Dam methylation sites (GATC) in the

Escherichia coli chromosome // Nucleic Acids Res. — 1988. — Vol. 16, № 20. — P. 9821–9838.

260. Lacks S.A., Mannarelli B.M., Springhorn S.S., et al. Genetic basis of the complementary DpnI and DpnII restriction systems of S. pneumoniae: an intercellular cassette mechanism // Cell. — 1986. — Vol. 46, № 7. — P. 993–1000.

261. Cantalupo G., Bucci C., Salvatore P., et al. Evolution and function of the neisserial dam -replacing gene // FEBS Lett. — 2001. — Vol. 495. — P. 178–183.

262. Rotman E., Seifert H.S. The Genetics of Neisseria Species // Annu. Rev. Genet. — 2014. — Vol. 48, № 1. — P. 405–431.

263. Treangen T.J., Ambur O.H., Tonjum T., et al. The impact of the neisserial DNA uptake sequences on genome evolution and stability // Genome Biol. — 2008. — Vol. 9, № 3. — P. R60.

264. Tesfazgi Mebrhatu M., Wywial E., Ghosh A., et al. Evidence for an evolutionary antagonism between Mrr and Type III modification systems // Nucleic Acids Res. — 2011. — Vol. 39, № 14. — P. 5991–6001.

265. Morozova N., Sabantsev A., Bogdanova E., et al. Temporal dynamics of methyltransferase and restriction endonuclease accumulation in individual cells after introducing a restriction-modification system // Nucleic Acids Res. — 2016. — Vol. 44, № 2. — P. 790–800.

266. O'Connell Motherway M., O'Driscoll J., Fitzgerald G.F., et al. Overcoming the restriction barrier to plasmid transformation and targeted mutagenesis in Bifidobacterium breve UCC2003 // Microb. Biotechnol. — 2009. — Vol. 2, № 3. — P. 321–332.