Математическая биология и биоинформатика
2023. Т. 18. № 2. С. 602–620. doi: 10.17537/2023.18.602

========== **БИОИНФОРМАТИКА** ==========

Структура повторов в геномах сальмонелл

Мирошниченко Л.А.[*1], Арефьева Н.А.[2,3], Джиоев Ю.П.[2], Гусев В.Д.[1], Борисенко А.Ю.[2], Эрдынеев С.В.[2], Букин Ю.С.[4,5]**

[1]*Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия*
[2]*Иркутский государственный медицинский университет, Иркутск, Россия*
[3]*Научный центр проблем здоровья семьи и репродукции человека, Иркутск, Россия*
[4]*Лимнологический институт СО РАН, Иркутск, Россия*
[5]*Иркутский государственный университет, Иркутск, Россия*

Аннотация. Сальмоне́ллы (*Salmonella*) – род грамотрицательных, факультативных и неспороносных анаэробных бактерий. Род включает два вида: *S. bongori* и *S. enterica*. Все патогенные в отношении человека и животных сальмонеллы относятся к виду *S. enterica* spp. Он включает в свой состав семь подвидов куда входят более 2500 серотипов. По глобальной статистике они ежегодно вызывают около 2.8 миллиарда случаев диарейных заболеваний, а смертность достигает более 300000 случаев. Особо опасными стали штаммы *S. enterica* spp, приобретшие множественную лекарственную устойчивость к антибиотикам. На фоне этой глобальной проблемы актуальность приобретает детальное исследование полных геномов различных представителей рода *Salmonella*. Важную роль в регуляции основных генетических процессов в ходе жизненного цикла и эволюции играют повторы. В работе рассматриваются разнообразные проявления повторности в геномах *S. enterica*. Учет общих фрагментов разных геномов служит основой для формирования матрицы попарной относительной сложности, использующейся при построении филогенетического дерева. Длинные повторы внутри отдельных геномов, как правило, соответствуют крупным индивидуальным дупликациям. Основное внимание уделено локальным структурным закономерностям, большая часть которых представлена тандемными повторами. Значительный интерес представляют многозначные повторы, такие как тандемные повторы, образующие палиндромы, повторы с регулярными заменами или сложной структурой мономера.

Ключевые слова: *бактериальный геном, сальмонелла, сложностные разложения, повторы, локальные структурные закономерности, тандемные повторы.*

ВВЕДЕНИЕ

Сальмонеллы представляют собой факультативно-анаэробные грамотрицательные жгутиковые бактерии рода *Salmonella* из семейства *Enterobacteriaceae*. Среди них вид *Salmonella enterica* spp. основной возбудитель сальмонеллезных заболеваний человека и животных, таких как: брюшной тиф, паратиф. В случаях, когда *S. enterica* spp. в желудочно-кишечном тракте становится слишком много, они могут проникнуть в кровь и распространиться по всему организму, поселиться в других органах и тканях, вызывая их воспаление. При этих случаях могут также проявляться такие патологические состояния как: менингиты, эндокардиты, остеомиелиты, васкулиты. Вид *S. enterica* spp. включает в свой состав семь подвидов, каждый из которых имеет

[*]luba@math.nsc.ru
[**] arefieva.n4@gmail.com

множество серогрупп. Из них подвид *S. enterica enterica* включает пять серогрупп: A, B, C, D, E., но в эпидемиологическом отношении наиболее значимы для человека лишь несколько из них [1–3]. Так, 90 % случаев сальмонеллёзов приходится на серотипы: *typhimurium* (серогруппа B), *enteritidis* (D), *infantis* (C), *newport* (C), *agona (*B), *derby* (B) и *london* (E). Причем, количество заболевших сальмонеллезами, в том числе в развитых странах, в последние годы возрастает. Это связано с появлением штаммов сальмонелл *S. typhimurium* и *S. enteritidis*, устойчивых к современным антибиотикам и распространением этих штаммов по всему миру. Внутрибольничный (нозокомиальный) сальмонеллёз также является одной из серьёзных проблем современного здравоохранения. В 80 % случаев возбудителем нозокомиального сальмонеллёза является серотип *typhimurium.* В большинстве случаев сальмонеллез передается через зараженную пищу, молочные и мясные продукты, яйца птиц. Заражение может происходить и контактно-бытовым путем через поврежденную кожу или слизистые оболочки [4–6]. На сегодняшний день идентифицировано около 2600 серотипов *S. enterica* spp. и большинство из них могут вызывать перекрестное заражение между животными и людьми. Так, на сальмонеллезы приходятся 93.8 миллиона случаев гастроэнтерита и при этом ежегодно во всем мире регистрируется около 155 000 смертей [7–9].

Также важной глобальной проблемой стала множественная лекарственная устойчивость (МЛУ) патогенных бактерий, в том числе и многих серотипов *S. enterica* spp. Эта проблема усугубляется с каждым годом, представляет собой угрозу благополучию людей во всем мире и одну из наиболее серьезных проблем общественного здравоохранения [10–12]. В «золотую эру антибиотиков» они помогли спасти бессчетное количество жизней, однако продлилась эта эра недолго. В настоящее время МЛУ патогенных бактерии к антибиотикам угрожает самым основам современной медицины и в ближайшие десятилетия она может стать более частой причиной смерти, чем при онкологии. По научным прогнозам, уже к 2050 году число смертей, связанных с МЛУ бактерии может достичь 10 миллионов человек в год [13]. Результаты многих исследований свидетельствуют в пользу существования прямой зависимости между увеличением потребления антибиотиков и распространением устойчивости бактерий к их действию, а в результате глобализации социальных связей и экономики и антибиотикорезистентность стремительно распространяется по всему миру [14]. Особенно острой эта проблема стала в период пандемии COVID-19 в стационарах, когда лечение с разной и смешанной этиологией послужило триггером ускорения образования антибиотикорезистентных госпитальных штаммов [15].

Между тем поиск новых антибиотиков сильно замедлился. Несмотря на острую потребность в противомикробных средствах, в настоящее время разрабатывается очень мало новых соединений, большинство из которых к тому же принадлежат к классам уже используемых антибиотиков. Так, за последние 15 лет лишь один новый класс антибиотиков против грамположительных бактерий был введен в клиническую практику, а последний класс антибиотиков широкого спектра действия был введен в клинику в 60-х гг. прошлого века [16, 17].

Таким образом, человечество стоит на грани глобального кризиса и может быть отброшено назад в доантибиотическую эпоху. Однако, на фоне этих проблем распространения патогенов с МЛУ вновь актуальной становится фаготерапия против многих патогенных микроорганизмов. Препараты бактериофагов – лучшая альтернатива антибиотикам по целому ряду причин: 1) фаги уничтожают только бактерии определенного вида; 2) прием фагов не вызывает аллергии, не снижает функцию иммунной системы организма; 3) препараты фагов – более быстрое и высокоэффективное средство лечения патогенов, чем антибиотики; 4) производство препаратов фагов – экологически абсолютно чистый процесс; 5) препараты фагов можно принимать вместе с другими препаратами; 6) препараты фагов можно

эффективно применять во всех возрастных группах и пациентах группы высокого риска [18, 19]. Также, современные геномные и биоинформационные технологии позволяют целенаправленно моделировать процесс отбора высокоспецифичных и вирулентных фагов против патогенных микроорганизмов на основе их геномных структур и механизмов антагонистического взаимодействия посредством систем CRISPR/Cas бактерии и анти-CRISPR фагов. CRISPR/Cas системы бактерии – это древнейшая система иммунной защиты бактерии от патогенных для них бактериофагов и плазмид. Посредством этой системы бактерии распознают и эффективно расщепляют ДНК фагов, используя ферментную систему Cas (CRISPR-associated). Ферменты системы Cas узнают необходимую последовательность на ДНК фагов с помощью комплементарной им РНК, играющей роль гида, и разрезают ее в нужном месте. Таким образом, с помощью системы CRISPR/Cas можно осуществлять все виды манипуляций с геномами бактерий и фагов [20, 21].

Появляется все больше работ, в которых проводятся исследования по скринингу фаговых видов с таргетными характеристиками против конкретных патогенов [22–24]. Однако пока внедрение фаготерапии в практику имеет ограничения в связи с недостаточной информацией о геномных механизмах антагонистических взаимоотношении между бактериями и фагами. В этих взаимоотношениях значимую роль могут играть как системные модификации в CRISPR-Cas у бактерии и anti-CRISPR-Cas у фагов, так и структурные модификации в их генах.

На фоне глобальной проблемы множественной лекарственной устойчивости патогенных бактерий к антибиотикам особенную актуальность приобретает детальное исследование геномов различных представителей рода *Salmonella*. Особый интерес представляют любые проявления повторности. Достаточно длинные повторяющиеся фрагменты присутствуют в геномах самых разных организмов. Они играют важную роль в регуляции основных генетических процессов в ходе жизненного цикла и эволюции. Повторы весьма многообразны. Так, например, выделяют повторы прямые и симметричные (инвертированные); точные и несовершенные (с заменами, вставками, делециями); разнесенные по тексту, регулярно расположенные и тандемно повторяющиеся; а также повторы, допускающие несколько трактовок (например, тандемные повторы, образующие палиндромы).

В работе рассматриваются разнообразные проявления повторности в геномах *S. enterica*. Учет общих фрагменты разных геномов служит основой для формирования матрицы попарной относительной сложности, использующейся при построении филогенетического дерева. Длинные повторы внутри отдельных геномов, как правило, соответствуют крупным индивидуальным дупликациям. Основное внимание уделено локальным структурным закономерностям, большая часть которых представлена тандемными повторами.

Значительный интерес к тандемным повторам обусловлен тем, что многие из них используются в качестве молекулярных маркеров при проведении внутри- и межвидовой классификации организмов. Так, в [25] полные спектры тандемных повторов применяются для дифференциации близкородственных бактерий *Yersinia pseudotuberculosis* и *Yersinia pestis* – возбудителей псевдотуберкулеза и чумы, соответственно, в [26] тандемные повторы рассматриваются как РНК-маркеры для генотипирования вируса клещевого энцефалита. В [27] VNTR – тандемные повторы различной кратности используются для типирования различных штаммов сальмонеллы серовара *typhimurium*, а в [28] для типирования *Salmonella enterica* subsp. *enterica*.

ДНК-ОРИЕНТИРОВАННАЯ МЕРА СЛОЖНОСТИ

Интуитивные представления о сложности последовательностей обычно связаны со степенью их «случайности». «Случайные» последовательности представляются

наиболее сложными. Колмогоров был одним из первых, кто предложил измерять сложность объекта числом и указал способ такого измерения: сложность есть длина наиболее короткого описания объекта, по которому этот объект может быть однозначно восстановлен [29]. Однако колмогоровское определение сложности неконструктивно в том смысле, что программы, гарантированно осуществляющей поиск кратчайшего (из всех возможных) описаний, позволяющих восстановить последовательность, не существует. Поэтому колмогоровская сложность обычно рассматривается как гипотетическая нижняя граница длины описания объекта, а на практике используются алгоритмические приближения к вычислению этой длины, что приводит к большому разнообразию определений сложности [30].

Существует тесная взаимосвязь между способами оценивания сложности и алгоритмами сжатия текстов. Однако для большинства приложений, важным элементом является не столько достигаемая степень компрессии, сколько выявление структурных закономерностей, обусловивших сжатие текста (в частности, ДНК-последовательности), и их интерпретируемость. Значительный интерес представляют локальные структурные особенности, выявляемые в режиме скользящего окна. Зоны пониженной сложности характеризуются высоким содержанием различных повторов: как тандемных, так и разнесенных; как прямых, так и комплементарных инвертированных. Именно повторы обеспечивают регуляцию разнообразных генетических процессов в ходе жизненного цикла и являются основой наследственной изменчивости.

Среди всего многообразия мер сложности [30] для анализа ДНК-последовательностей мы традиционно используем ДНК-ориентированный вариант меры сложности [31] на основе определения Лемпеля и Зива [32].

Пусть $\Sigma = \{a, c, g, t\}$ – алфавит; S – конечная последовательность, составленная из элементов Σ; $N = |S|$ – длина S; $S[i] = S_i$ – элемент, стоящий в i-й позиции S, $S[i : j] = S_i \ldots S_j$ – фрагмент S, включающий элементы с i-го по j-й ($1 \le i < j \le N$); $S^R[i : j] = S^R_j \ldots S^R_i$ – инвертированный фрагмент, т.е. прочитанный в обратном направлении с учетом комплементарной подстановки $a \leftrightarrow t$, $g \leftrightarrow c$.

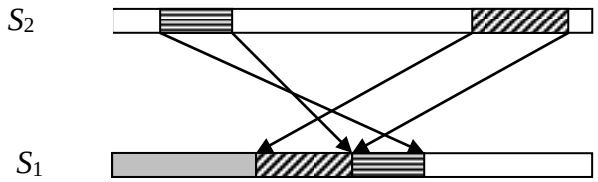


Рис. 1. Схема покрытия последовательности S_1 фрагментами из S_2. Серым цветом (без штриховки) выделена покрытая за $(k - 1)$ шагов часть S_1. Примыкающие к ней два фрагмента добавляются на k-м и $(k + 1)$-м шагах. Первый иллюстрирует случай прямого копирования (диагональная штриховка), второй – инвертированного (горизонтальная штриховка).

В случае двух последовательностей S_1 и S_2 одна из них, например, S_1, представляется в виде конкатенации фрагментов из S_2. Покрытие S_1 фрагментами из S_2 строится от начала к концу (слева–направо). На каждом шаге копируется максимальный фрагмент (прямой или инвертированный) из S_2, который совпадает с префиксом еще непокрытого участка S_1 (см. рис. 1). Если нечего копировать (символ из S_1 не встречается в S_2), используется операция «генерации символа». Получаемую таким образом последовательность разнотипных межтекстовых повторов $H(S_1 / S_2)$ назовем сложностным разложением S_1 по S_2, а число компонентов в нем $c(S_1 / S_2)$ – относительной сложностью. Эта характеристика вычислима за линейное время. На основе $H(S_1 / S_2)$ формируется множество совершенных межтекстовых повторов, длина которых превышает заданный порог.

Для сравнения группы последовательностей $P = \{S_1, S_2 \ldots S_p\}$ каждая из последовательностей сравнивается с каждой, т. е. формируется матрица, каждый элемент которой $c_{ij} = c(S_i / S_j)$, $1 \le i, j \le p$ есть сложность S_i относительно S_j. Эта матрица может быть использована для построения дерева филогенетических связей.

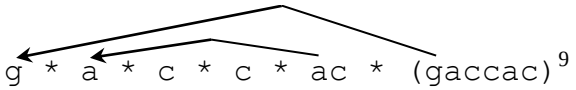
Сложность одной последовательности S измеряется числом шагов порождающего ее процесса. Допустимыми операциями при этом являются: копирование максимально длинного «готового фрагмента» из уже синтезированной части текста и генерация символа (используется только для синтеза элементов алфавита). Наряду с прямым копированием используется инвертированное копирование. При этом на каждом шаге процесса синтезирования S из двух операций копирования выбирается та, что соответствует более длинному повтору. При необходимости число операций может быть расширено добавлением симметричного или прямого комплементарного копирования.

Схема порождения последовательности S, называемая в дальнейшем сложностным разложением, может быть представлена в виде конкатенации фрагментов

$$H(S) = S[1 : i_1] * S[i_1 + 1 : i_2] * \ldots * S[i_{k-1} + 1 : i_k] * \ldots * S[i_{m-1} + 1 : N].$$

Фрагменту $S[i_{k-1} + 1 : i_k]$, синтезируемому на k-м шаге, соответствует прототип в предыстории (префиксе $S[1: i_k - 1]$), в виде прямого или инвертированного повтора. Т. е. существует позиция $j_k < i_k$), такая что либо $S[i_{k-1} + 1 : i_k] = S[j_{k-1} + 1 : j_k]$ ($j_k < i_k$), либо $S[i_{k-1} + 1 : i_k] = S^R[j_{k-1} + 1 : j_k]$. При этом компонент разложения $S[i_{k-1} + 1 : i_k]$ нерасширяем вправо, т.е. фрагмент $S[i_{k-1} + 1 : i_k + 1]$ не встречается в $S[1: i_k]$ ни в прямом ни в инвертированном виде. $C = m$ – число шагов процесса будем называть сложностью последовательности S.

Замечательным свойством сложностных разложений является возможность копирования с использованием элементов, синтезированных на данном шаге. Это свойство в сочетании с режимом скользящего окна лежит в основе алгоритма выявления всех точных тандемных повторов (периодичностей) [33]. Например, сложностное разложение для $S = (\text{gaccac})^{10}$ выглядит следующим образом:



Сложностные разложения полных геномов чаще всего используются для выделения наиболее крупных областей дупликации. Локальные структурные закономерности, подчас оказываются замаскированными в разложении полного генома. Инструментом для их обнаружения служит построение сложностных разложений в режиме скользящего окна фиксированной длины W. Несовершенные и регулярно расположенные повторы выявляются как на основе анализа позиционной информации о компонентах сложностных разложений и их прототипов, так и с помощью анализа позиционной информации, содержащейся в листьях вспомогательных деревьев без построения самих разложений.

Объектом исследования послужили полногеномные последовательности 29 штаммов вида *Salmonella enterica* spp., два генома вида *Salmonella bongori* из банка данных NCBI. *S. enterica* spp. представлен девятью типами сероваров, а *S. bongori* – двумя. Для контраста в подборку геномов включен геном *Atlantibacter subterranean*, ранее относящийся к роду *Salmonella*. Размеры геномов варьируют от 4 453 147 до 6 125 327 нуклеотидов.

Таблица 1. Геномы сальмонелл, использованные в исследовании, и их характеристики

№	Genbank ID	Геномы	Сокр. назв.	Размер генома	Источник	Регион выделения
1	CP100494	*Atlantibacter subterranea* strain LH84	Atlantibacter	4712132	курица	Швейцария
2	LR134156	*S. enterica* subsp. *arizonae* strain NCTC10047	S-arizonae1	4930903	неизвестен	Англия
3	CP082954	*S. enterica* subsp. *arizonae* strain S499	S-arizonae2	4916406	человек	Китай
4	CP030026	*S. enterica* subsp. *diarizonae* ser. 59	S-diarizonae1	6125327	неизвестен	Канада
5	CP029989	*S. enterica* subsp. *diarizonae* ser. 48	S-diarizonae2	5361355	неизвестен	Канада
6	CP009049	*S. enterica* subsp. *enterica* ser. *paratyphi A* strain 50973	Se-paratyphiA1	4599018	человек	Китай
7	CP009559	*S. enterica* subsp. *enterica* ser. *paratyphi A* strain CMCC 50503	Se-paratyphiA2	4612373	питание	США
8	CP074223	*S. enterica* subsp. *enterica* ser. *paratyphi B* str. CFSAN000546	Se-paratyphiB1	4453147	питание	США
9	NC010102	*S. enterica* subsp. *enterica* ser. *paratyphi B* str. SPB7	Se-paratyphiB2	4858887	дикое животное	Малайзия
10	CP040562	*S. enterica* subsp. *enterica* ser. *typhimurium* strain SAP17-7399	Se-typhimurium1	4902004	человек	США
11	NC003197	*S. enterica* subsp. *enterica* ser. *typhimurium* str. LT2	Se-typhimurium2	4857450	дикое животное	Малайзия
12	CP052777	*S. enterica* subsp. *enterica* ser. *infantis* strain CVM N18S1246	Se-infantis1	4662734	курица	США
13	CM001274	*S. enterica* subsp. *enterica* ser. *infantis* str. SARB27	Se-infantis2	4878146	неизвестен	Сенегал
14	CP012598	*S. enterica* subsp. *enterica* ser. *newport* strain 0211-109	Se-newport1	4962039	корова	США
15	CP025243	*S. enterica* subsp. *enterica* ser. *newport* str. CDC 2012K-0663	Se-newport2	4928771	человек	США
16	CP018657	*S. enterica* subsp. *enterica* ser. *enteritidis* strain 92-0392	Se-enteritidis1	4934513	человек	США
17	CP050721	*S. enterica* subsp. *enterica* ser. *enteritidis* strain SE81	Se-enteritidis2	4738616	кровь человека	Китай
18	NC016832	*S. enterica* subsp. *enterica* ser. *typhi* str. P-stx-12	Se-typhi1	4768352	больной человек	Малайзия
19	CP012091	*S. enterica* subsp. *enterica* ser. *typhi* strain PM016/13	Se-typhi2	4793553	вода	Малайзия
20	CP013226	*S. enterica* subsp. *enterica* ser. *anatum* strain GT-38	Se-anatum1	4830602	индейка	США
21	CP014659	*S. enterica* subsp. *enterica* ser. *anatum* str. USDA-ARS-USMARC-1765	Se-anatum 2	4945392	человек	США
22	CP074204	*S. enterica* subsp. *enterica* ser. *london* str. CFSAN001081	Se-london1	4631614	молоко	США
23	CP060134	*S. enterica* subsp. *enterica* ser. *london* strain HA1-SP5	Se-london2	4747268	свинья	Китай
24	CP051368	*S. enterica* subsp. *houtenae* ser. 43	S-houtenae1	4677885	бык	Китай
25	CP075174	*S. enterica* subsp. *houtenae* ser. 45	S-houtenae2	4651005	геккон	США
26	UGYB01000001	*S. enterica* subsp. *indica*	S-indica	4749918	коллекция	Англия
27	CP034697	*S. enterica* subsp. *salamae* ser. 42	S-salamae1	4861251	мясо	Руанда
28	CP079836	*S. enterica* subsp. *salamae* strain LHICA_SA1	S-salamae2	5045976	бройлер	Испания
29	CP074220	*S. enterica* subsp. VII str. CFSAN000554	S-VII-1	4676509	человек	США
30	CP053582	*S. enterica* subsp. VII serovar 1,40	S-VII-2	4517871	коллекция	Тонго
31	NC021870	*S. bongori* N268-08	S-bongori1	4683551	клиника	Швейцария
32	CP022120	*S. bongori* ser. 66:z41	S-bongori2	4468959	коллекция	Канада

Полный перечень исследованных геномов и принятые в дальнейшем изложении их сокращенные наименования приведены в таблице 1.

СРАВНЕНИЕ ПОЛНЫХ ГЕНОМОВ САЛЬМОНЕЛЛЫ

Попарное сравнение последовательностей проводится с помощью построения сложностных разложений $H(S_i / S_j)$ одной последовательности относительно другой. В первую очередь нас интересуют компоненты разложений, длина которых превышает некоторый порог (R = 500). Они соответствуют либо прямым повторам, либо инвертированным.

Максимальный совершенный прямой повтор длиной L = 130343 выявлен при сравнении двух разных цепей *Salmonella enterica* серовара *london*. Кроме этого повтора в этих же геномах встречаются точные совпадающие фрагменты длиной 124267, 121758, 115299 и 108362 и другие. Суммарная длина выделенных компонентов сложностных разложений (повторов, длина мономера которых превышает 500 нуклеотидов) сопоставима с длиной геномов.

Максимальный совершенный инвертированный повтор имеет длину L = 169106 нуклеотида. Эта повторяющаяся область выявляется при сравнении разных цепей серовара *paratyphi* A (Se-paratyphiA1 и Se-paratyphiA2). В обеих последовательностях область начинается геном fdnG (*formate dehydrogenase-N subunit alpha*), заканчивается геном cls (*cardiolipin synthase*) и содержит не одну сотню генов. Всего обнаружен 251 точный повтор между этими геномами с длиной мономеров 500 нуклеотидов и выше, подавляющее большинство из них представляют собой комплементарные инвертированные повторы, в шести случаях длина мономера превышает 100000. Длинные повторы практически полностью покрывают геномы.

В таблице 2 перечислены длины максимальных точных (прямых или инвертированных) повторов между разными геномами. Включены только значения, превышающие 10000.

Таблица 2. Длина максимальных точно совпадающих фрагментов разных геномов

№	Длина максимального повтора	Геном 1	Геном 2
1	169106	Se-paratyphiA1	Se-paratyphiA2
2	130343	Se-london1	Se-london2
3	115686	Se-newport1	Se-newport2
4	86856	Se-anatum1	Se-anatum2
5	83802	Se-typhimurium1	Se-typhimurium2
6	53395	Se-typhi1	Se-typhi2
7	52971	Se-typhimurium2	Se-enteritidis1
8	47891	Se-typhimurium1	Se-enteritidis1
9	27493	Se-paratyphiB1	Se-paratyphiB2
10	17402	Se-typhimurium2	Se-newport2
11	17161	Se-typhimurium1	Se-newport2
12	16721	Se-newport1 Se-newport2	Se-london1 Se-london1
13	15602	Se-newport1 Se-newport2	Se-london2 Se-london2
14	14103	Se-typhimurium2	Se-newport1
15	13383	S-VII-1	S-VII-1
16	11539	Se-paratyphiB2 Se-paratyphiB2	Se-typhimurium2 Se-enteritidis1
17	10635	Se-typhimurium1	Se-newport1

Обратим внимание, что самые длинные повторы, как правило, обнаруживаются между разными геномами внутри одинаковых сероваров. Исключение составляет геном

Se-enteritidis1, содержащий длинные фрагменты, точно совпадающие с фрагментами последовательностей серовара *Typhimurium*. Длина повторов между геномами разных подвидов сальмонеллы, как правило, превышает тысячу нуклеотидов. При сравнении всех геномов с *Atlantibacter subterranean* (представителем другого рода) выделяется только один фрагмент rRNA -16S рибосомной РНК длиной чуть более 500 нуклеотидных пар (н.п.), являющийся точной копией аналогичных фрагментов из других геномов.

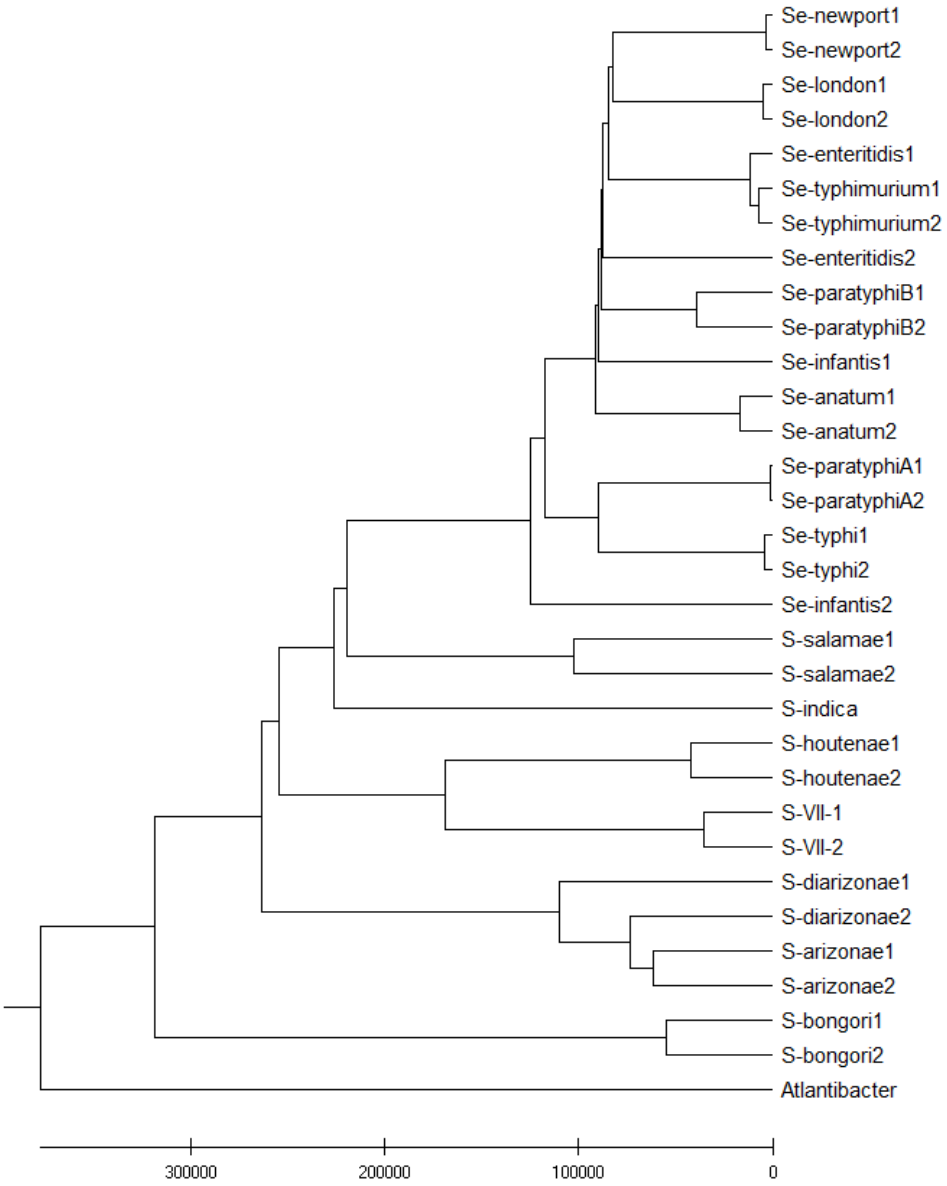


Рис. 2. Дерево филогенетических связей для 32 полных геномов рода *Salmonella*, построенное UPGMA методом из пакета программ MEGA на основе матриц попарной относительной сложности.

По значениям относительной сложности $c(S_i / S_j)$, $1 \le i, j \le 32$, полученным при сравнении каждой пары из 32 полных геномов всей подборки сформирована матрица попарной относительной сложности. Для построения дерева филогенетических связей (рис. 2) на основе полученной матрицы использован UPGMA метод [34] из пакета программ MEGA [35]. Деревья, построенные другими методами (например, NJ [36]) практически совпадают с представленным деревом.

Филогенетическое дерево, построенное по «относительной сложности», в целом отражает согласованность с общепринятой классификацией. Так, последовательности, относящиеся к подвиду *S. enterica* subsp. *enterica* образуют отдельный кластер на дереве, другой кластер образован последовательностями подвида *arizonae* и *diarizonae*.

Однако есть и некоторые исключения. Так, уже упомянутая последовательность Se-enteritidis1 оказывается ближе к последовательностям серовара *typhimurium*, чем к последовательности своего серовара Se-enteritidis2, а последовательности серовара *infantis* не демонстрируют высокой близости между собой.

СОВЕРШЕННЫЕ ПОВТОРЫ В ГЕНОМАХ САЛЬМОНЕЛЛЫ

Для выявления наиболее крупных повторяющихся фрагментов (прямых или инвертированных) в каждом из геномов используется построение сложностных разложений полных геномов на основе ДНК-ориентированной меры сложности. Дальнейшему исследованию подвергаются компоненты разложений, длина которых превышает некоторый порог.

Таблица 3. Длинные повторы в каждом из исследованных геномов

№	Геномы	Число компонентов длиной $L \geq 200$	Суммарная длина выделенных компонент	Максимальная длина повтора
1	Atlantibacter	36	38331	5007
2	S-arizonae1	93	48921	2873
3	S-arizonae2	89	81385	19224
4	S-diarizonae1	393	191922	3251
5	S-diarizonae2	58	93493	12809
6	Se-paratyphiA1	56	64347	4917
7	Se-paratyphiA2	53	84375	6572
8	Se-paratyphiB1	39	31716	4907
9	Se-paratyphiB2	53	48578	6289
10	Se-typhimurium1	63	53032	5477
11	Se-typhimurium2	57	52443	5205
12	Se-infantis1	59	41265	2046
13	Se-infantis2	48	29939	3107
14	Se-newport1	75	173721	28538
15	Se-newport2	67	54391	6590
16	Se-enteritidis1	70	54631	5015
17	Se-enteritidis2	21	41409	6121
18	Se-typhi1	50	67175	7895
19	Se-typhi2	49	67118	7886
20	Se-anatum1	20	226734	74175
21	Se-anatum 2	77	50853	5085
22	Se-london1	58	42011	4350
23	Se-london2	58	46387	5279
24	S-houtenae1	81	60051	5821
25	S-houtenae2	66	69341	5887
26	S-indica	82	61824	3230
27	S-salamae1	61	76020	5656
28	S-salamae2	88	68325	11268
29	S-VII-1	66	71784	5391
30	S-VII-2	65	55242	5668
31	S-bongori1	84	57398	5017
32	S-bongori2	59	48310	5500

В таблице 3 для каждого генома из таблицы 1 указано число длинных повторов (компонент сложностного разложения, длина которых превышает 200 н.п.), их суммарная длина, а также длина максимального повтора.

Максимальный совершенный внутритекстовый комплементарный инвертированный повтор обнаружен в геноме Se-anatum1 в позициях 2140190 и 1729912. Он имеет длину 74175 нуклеотида и содержит около сотни генов. На границах области лежат гены *bifunctional glycosyl transferase/transpeptidase* и *enterobactin non-ribosomal peptide synthetase EntF*. Этот же геном содержит еще две огромные области дупликации (уже в прямом направлении). Одна из них имеет длину 83398 с одной заменой (позиции 4276032 и 283029). К повтору длиной 22933 н.п. в позициях 4361979 и 2463654 примыкает повтор длиной 7082 (позиции 4384924 и 2486596). Хотя в этом геноме выделено всего 20 повторов, их суммарная длина составляет 2×226734 н.п. По-видимому, дупликации в этом геноме произошли сравнительно недавно.

Лидером по числу точных повторов длиной свыше 200 н.п. является геном S-diarizonae1. В нем выявлено 393 повтора, суммарная длина которых составляет 2×191922, но максимальная длина повторяющегося мономера составляет только 3251 н.п. Эти повторы образуют более длинные повторяющиеся области, но уже несовершенные. Например, интервалы [435117, 440311] и [1242757, 1247953] длиной 5197 представляют собой несовершенный повтор, содержащий 9 совершенных ядер, суммарной длиной 4837. Множество точечных мутаций внутри повторяющихся областей, по-видимому, свидетельствует о том, что крупные дупликации произошли довольно давно.

Геном Se-infantis2 характеризуется наименьшей суммарной длиной компонентов, превышающих 200 н.п. (2×29939) при относительно невысоком их числе (48). Максимальный точный повтор имеет длину 3107, но содержит при этом несколько небольших генов: *heme lyase CcmF/NrfE family subunit, cytochrome c maturation protein* CcmE, ccmB, CcmD и *heme ABC transporter permease*.

Наиболее типичными внутритекстовыми повторами, выявляющимися практически во всех геномах являются повторы рибосомных и транспортных генов, например фрагментов (16S ribosomal RNA + , tRNA-Glu + 23S ribosomal RNA).

СОВЕРШЕННЫЕ ТАНДЕМНЫЕ ПОВТОРЫ

В таблице 3 указано число тандемных повторов, длина которых не менее 25 нуклеотидов, а длина мономера не превышает 22 н.п. Сразу можно обратить внимание на то, что в геномах сальмонеллы выявляется слишком мало тандемных повторов. Этот эффект особенно заметен при сравнении с числом тандемных повторов бактерии *Y. pseudotuberculosis* [25]. В то же время некоторые геномы сальмонеллы содержат тандемные повторы с довольно длинными мономерами.

Моносерии обнаружены только в двух геномах: a^{48} у S-houtenae2; g^{60} и g^{28} у S-salamae1. Нет ни одного тандемного повтора с длиной мономеров 2 и 23. Единственный тандемный повтор с длиной мономера 25 выявлен у S-bongori1. При длине мономера, превышающей 25, тандемные повторы обнаруживаются в отдельных геномах. Исключение составляют только тандемы с длинами мономеров 33, 36, 39 и 45, обнаруженные в 13, 12, 17 и 12 из 32 геномов, соответственно.

Наибольшее число (13) тандемных повторов с длиной мономера 15 выявлено в Se-infantis1. Девять из них являются фрагментами одного длинного несовершенного повтора (см. следующий раздел). Также в S-salamae2 девять из одиннадцати совершенных тандемных повторов с длиной мономера 15 образуют длинный несовершенный повтор, включающий кроме того точные повторы с мономерами длиной 30, 75 и даже 120. Одиннадцать тандемных повторов с длиной мономера 9 в S-diarizonae1 при этом отличны друг от друга.

Максимальный совершенный тандемный повтор имеет длину $13000 = 6401 \times 2.03$. Он выявлен в геноме S-diarizonae2. Мономер состоит из 8 генов и межгенных участков, среди которых *helix-turn-helix transcriptional regulator, phage/plasmid replication protein, DNA-binding protein, Eex N family lipoprotein type IV secretion system protein, conjugal transfer protein* и еще пара гипотетических протеинов.

Таблица 4. Зависимость числа тандемных повторов от длины мономера

№	Геномы	Длина мономера																		
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	Atlantibacter				3	1		1		1	7		2				7				
2	S-arizonae1	1	1		2	1		4			2			3			2				
3	S-arizonae2	1	1		3	2		6		1	3						3				
4	S-diarizonae1	1	1	1	1	1	2	11	1		6	1			1					1	1
5	S-diarizonae2	1	1	1	1		1	5			5						3		1		
6	Se-paratyphiA1				3			5	2		1						2				
7	Se-paratyphiA2				3			6	2		1						1				
8	Se-paratyphiB1				2		1	1	3		1		2			1			1	1	
9	Se-paratyphiB2				1			1	2		2					1			1	1	1
10	Se-typhimurium1				2			2			4		2	2							
11	Se-typhimurium2				3			3			4		2	2							
12	Se-infantis1			1	2						1		1	13		1					
13	Se-infantis2	1	1	1	1		1				3	1		1							
14	Se-newport1	1	1		2			1			4					1					
15	Se-newport2	1	1		2			1			3					1					
16	Se-enteritidis1				2	1		3			4		2	4					1	1	
17	Se-enteritidis2				1	2	2	1			1		2	1	1	1	1				
18	Se-typhi1	1	1		4	1	1		1		3	3	2	2					1		1
19	Se-typhi2	1	1		4	1	1	1	1		3	2	2	2					1		
20	Se-anatum1					2					1			4	1	1					
21	Se-anatum 2				1	2		1						2	1	1		1			
22	Se-london1				1		1	1			2			1		1				1	
23	Se-london2				1		1				2			1		1				1	
24	S-houtenae1				1	1	3	2		1	1			2						3	
25	S-houtenae2	1	2			1	1	3	1	3	2	1		2			2			3	1
26	S-indica	1	1		3	2	2			2	6			4			1				
27	S-salamae1		2		3	1		2	1		3			1							
28	S-salamae2			2	2	1	2		1		3			11			1			1	
29	S-VII-1				1	2		2	1	2	4			2	1		1			1	
30	S-VII-2					1	2	3		2	1				1		2			1	
31	S-bongori1				2	1	4	2	1		4			4							
32	S-bongori2				1			1			6	1		1							

Геном Se-paratyphiA1 лидирует по общему числу точных тандемных повторов (42), длина которых превышает порог 25, а также по числу тандемных повторов (27) с длиной мономеров, превышающей 40 н.п. Он содержит девять тандемных повторов с мономерами длиной от 100 до 300 н.п., а также тандемные повторы с мономерами длиной 409, 640, 1743, 1916 и 4914. Геном этого же серотипа Se-paratyphiA2 содержит длинные тандемные повторы с мономерами длиной 693, 999, 1775, 2885, а также восемь тандемов с длиной мономеров от 106 до 242. Несмотря на высокое сходство этих геномов (см. выше), общими тандемными повторами являются только повторы с длинами мономеров 118 и 119, соответствующие повторам тРНК. Остальные дупликации на основе длинных мономеров, по-видимому, являются индивидуальными для отдельных геномов. Так, тандемному повтору с длиной мономера 1743 на цепи Se-paratyphiA1 соответствует дупликация области, содержащей два гена: *ribosome*

biogenesis GTPase Der и *zinc ribbon domain-containing protein*. В геноме Se-paratyphiA2 ген Der представлен одним экземпляром.

То, что тандемные дупликации длинных мономеров в основном характерны для отдельных геномов, подтверждается и тем фактом, что у 12 геномов из 32 отсутствуют точные тандемные повторы с мономерами длиной выше 125.

Среди тандемных повторов с высокой кратностью можно отметить лишь несколько. В геноме S-arizonae1 наибольшую кратность имеет повтор (cctgtt)^{16}cctgt в позиции 4274437, которому соответствует фрагмент белка *relaxase* (PV)[17]. Повтор (gcaagg)^{18} соответствует повтору (KG)[18] в геноме Se-typhi1, (gtactg)[20] согласован с (VL)[20] в этом же геноме. Повтор (atcatcgcc)[27] в позиции 351759 генома S-houtenae2 соответствует повтору (DDG)[27] у гена *autotransporter adhesin BigA*.

В некодирующих участках разных геномов выявлены, в основном, тандемные повторы с мономерами 5, 7 и 8. Так, повтор (taaattac)[19] обнаружен в позиции 1751490 генома S-diarizonae1, (ttgcc)[27] в позиции 4467134 генома S-diarizonae2, (cccag)[22] в геноме Se-infantis1, а (actcaga)[24] в Se-enteritidis1.

Однако не все повторы, длина мотива которых не кратна трём, лежат в некодирующих частях. Так фрагмент (ctgcgatg)[17] в позиции 1766341 S-houtenae1, соответствует шестикратному повтору из 8 аминокислот: (AASQHRSI)[6].

МНОГОЗНАЧНЫЕ ТАНДЕМНЫЕ ПОВТОРЫ

Среди тандемных повторов встречаются совершенные повторы, которые можно трактовать и как несовершенные, но с меньшей длиной мономера. Простейшими среди таких повторов являются повторы с регулярными заменами, т.е. повторы типа MPTR – Multi-Periodic Tandem Repeat [37]. Так,

cagta-cagca-
caata-cagca-
cagta-cagca-
caata-cagca

можно рассматривать как восьмикратный повтор мономера cagta с шестью заменами, четырехкратный повтор мономера (cagtacagca) с двумя заменами g на a в третьей позиции и совершенный повтор мономера (cagta cagca caata cagca).

Наряду с повторами с регулярными заменами, сложную структуру могут иметь и некоторые другие типы повторов. К повторам с многозначной трактовкой можно отнести, например, компаунды, т.е. тандемы из тандемных повторов. Если их мономеры имеют высокую степень сходства, невозможно установить грань между понятием компаунда и несовершенного повтора. В этом же разделе рассмотрим несовершенные тандемные повторы высокой кратности, имеющие в своем составе совершенные копии.

В геномах сальмонеллы обнаруживается намало тандемных повторов с регулярными заменами. Так, в S-arizonae2 точный тандемный повтор $194 = 78 \times 2.49$ (здесь 78 – длина мономера, 2.49 – кратность, 194 – длина всего фрагмента) можно трактовать также как несовершенный (с одной регулярной заменой) пятикратный повтор мономера длины 39. Повтор с тем же мономером (только в комплементарном виде) обнаружен и в геноме S-diarizonae2.

Фрагмент [4512995–4513484] внутри гипотетического протеина DOE63_23970 генома S-diarizonae1 содержит пять точных повторов с мономером 27, а также тандемные повторы с мономерами 108 и 81. Этот фрагмент можно трактовать как несовершенный восемнадцатикратный тандемный повтор с мономером 27, которому соответствует несовершенный повтор мотива TGDCSAASN (TGYRSAASN) на аминокислотном уровне. В этом же геноме точный тандем с мономером 36 соответствует неточному пятикратному тандему из 18 нуклеотидов с регулярными

заменами. На аминокислотном уровне этот повтор расширяем до 10 несовершенных копий мотива PDLRPA в составе *autotransporter outer membrane beta-barrel domain-containing protein*.

В геноме Se-infantis1 выявлено 13 точных тандемных повторов разной кратности с мономером `caggttggcatcgcg` длины 15, а также точные тандемы с мономерами 45 и 90, представляющие собой несовершенные повторы того же мономера длины 15. Совокупность этих точных повторов можно трактовать как несовершенный повтор указанного мономера кратности 25, которому на аминокислотном уровне соответствует (DANLR)[25] в области *pentapeptide repeat-containing protein*.

Точные тандемные повторы мономера длиной 33 н.п. разной кратности встречаются в 13 геномах из 32, в 19 геномах выявлены неточные тандемы с тем мономером кратности 3 и выше. Кроме того, повторы с мономерами длиной 66 и 132 представляют собой неточные повторы этого же мономера (33) в области *autotransporter domain-containing protein* (SGDDDVTPPDD). В результате, например, в Se-paratyphiA2, образуется 26-кратный несовершенный повтор мотива `tcgggcggggttacatcgtcatcgccgctatca` длины 33.

Более сложной структурой обладает точный тандемный повтор 480 = 108×4.44 в области *DNA translocase Fts*K генома Se-paratyphiA2, мономер которого длины 108 состоит из трех частей, две из которых совпадают с точностью до одной замены, а третья отличается от первых двух и точечными заменами и делецией 9 нуклеотидов:

```
caacagccggtagcgccgcagccgcagtatcagcagccg
caacagccggtagcaccgcagccgcagtatcagcagccg
cagcaaccgacagcgccacagccgcagtat
```

В некодирующих частях S-diarizonae2 выявлены два тандемных повтора с мономером длиной 51 н.п. Эти повторы комплементарны друг другу, а повтор с мономером 52 (105 = 52×2.02) примыкает ко второму повтору, отличается от него одной вставкой, т.е. образует компаунд (или несовершенный тандемный повтор). Повторы с теми же мономерами длины 51 или 52 наблюдаются еще в 6 геномах, но компаунд образуется только в описанном случае.

Точный тандемный повтор в позиции 4951496 генома S-diarizonae2 с мономером длиной 927 п.н.: 1910 = 927×2.06 и примыкающий к нему тандемный повтор с мономером 309 (674 = 309×2.18, поз. 4952114) образуют восьмикратный несовершенный тандемный повтор с мономером 309 (поз. 4950937), лежащий в области *Ig-like domain-containing protein*. Область [4950321..4974440] длиной 24120 н.п. (8039 аа) содержит большое число тандемно повторяющихся *Bacterial Ig-like domain*, длиной 309 нуклеотидов. Некоторые фрагменты этой области содержат довольно протяженные совершенные повторы с длиной мономеров 309 и 927, которые можно расширить несовершенными копиями этих же мономеров. Аналогичные повторы обнаруживается в геномах S-arizonae1 и S-diarizonae1. В S-arizonae1 несовершенные тандемные повторы мономера длины 309 образуют совершенный тандемный повтор с длиной мономера 618, а в S-diarizonae1 точный трехкратный тандемный повтор мономера длины 309 неточно расширен до кратности 4. Нуклеотидные последовательности мономеров длины 309, образующих тандемы в геномах S-arizonae1 и S-diarizonae2 совпадают, а в S-diarizonae1 комплементарна им. В S-arizonae2, а также некоторых других геномах (например, у представителей серовара *Typhimurium*) иммуноглобулинподобный домен проявляет себя несколькими несовершенными (сильно размытыми) повторами кратности порядка 600 и даже 1200. В ряде геномов следы таких повторов не выявлены.

Довольно много примеров, когда точный тандемный повтор составляет часть несовершенного повтора более высокой кратности. Так, точный тандемный повтор 1551 = 686×2.26 в позиции 3434309 S-VII-2 расширяем до пяти несовершенных копий (поз. 3433020), соответствующих пятикратному повторению гена *DUF2931 family*

protein длиной 681. А точный тандемный повтор 513 = 180×2.85 в геноме S-arizonae2 составляет часть пятикратного несовершенного повтора (поз. 3277251), три копии которого имеют длину 181. Этот повтор содержит нескольких копий *RtT sRNA*. Аналогичные повторы малых РНК наблюдаются и в других геномах. Длина мономера может при этом варьироваться.

ПАЛИНДРОМЫ И РЕГУЛЯРНЫЕ ПОВТОРЫ

Все многообразие палиндромов сводится к палиндромам четной длины, нечетной и комплементарным палиндромам (они, естественно, имеют четную длину). Наибольший интерес представляют палиндромы, длина которых превышает заданный порог. При пороге $R = 20$ в каждом геноме выявляется около десятка палиндромов четной длины, примерно столько же нечетной длины и примерно в три раза больше комплементарных палиндромов.

Некоторые из палиндромов довольно консервативны и присутствуют (иногда в комплементарном виде) у представителей различных подвидов и серотипов *S. enterica*. Это, например, палиндромы `tcgtgctcaga-agactcgtgct`, `atgcggctggc-g-cggtcggcgta` и комплементарный палиндром `cacaaaaaaaccgctcaa-ttgagcggtttttttgtg`.

Другие, например, `cattttcagtg-gtgacttttac`, `gcgcaactgt-t-tgtcaacgcg` и `aaaaaaagcgggtt-aacccgctttttt` представлены у подвидов *arizonae* и *diarizonae*.

По три копии комплементарного палиндрома `gtacgtaaaaa-tttttacgtac` выявлено в некодирующих частях геномов S-diarizonae1 и S-diarizonae2 с примерно одинаковыми расстояниями между копиями. В S-diarizonae1 это расстояние составляет около 9 тысяч, а у S-diarizonae2– 6400.

Самый длинный комплементарный палиндром (2×27) выявлен также в S-diarizonae1 `aaagtatctgctggataacgtttttag-ctaaaaacgttatccagcagatacttt` в позиции 2761837.

Некоторые палиндромы формируются на основе тандемных повторов с более коротким мономером. Например, на тандемный повтор `t-tgatgtagt-tgatgtagt-tgatgta` с симметричным мономером в Se-typhimurium2 накладывается палиндром `ttgatgtagt-tgatgtagtt` в позиции 2332009; длиной 2×10, а также палиндром нечетной длины `atgtagttgat-g-tagttgatgta` (поз. 2332012).

Аналогичный пример выявлен в области гена, содержащего анкириновые повторы генома S-salamae2. Здесь тандемный повтор в позиции 2043270 (`attgttac`)^{12}`attgtt` с почти симметричным мономером формирует множество палиндромов нечетной длины, с центрами симметрии g или с типа:

`ttgttacattgttacattgttacatt-g-ttacattgttacattgttacattgtt`

и

`attgttacattgttacattgtta-c-attgttacattgttacattgtta`

Максимальный из таких палиндромов имеет длину 101 = 2×50 + 1.

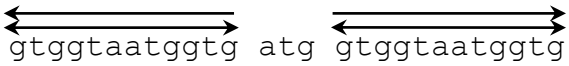
Структуры такого типа мы называем локальными фракталами [33]. Они характеризуются проявлениями самоподобия, основанного на свойстве симметрии или комплементарной симметрии. Элемент самоподобия проявляется в том, что повторение обычного палиндрома или комплементарного приводит к усилению конструкции, т.е. образованию нового палиндрома (соответственно, комплементарного палиндрома) вдвое большей длины. При кратности повторений выше двух возникают множественные структуры.

Однако в геномах сальмонелл, кроме приведенных примеров, проявления такой фрактальности крайне слабые. В ряде геномов вся фрактальность сводится к

фрагментам типа $(cgc)^5$ и $(gcg)^5$ и $(aaagaaa)^2$, а общее число выделяемых структур в несколько раз меньше, чем например, у бактерии *Y. pseudotuberculosis*. Это позволяет даже высказать гипотезу о наличии некоторого запрета на формирование таких структур.

В то же время, если рассматривать повторы палиндромов (комплементарных палиндромов), допуская некоторую вставку между ними, размер которой сравним с длиной мономера, выявляется ряд любопытных структур. Во всех геномах, за исключением *Atlantibacter subterranea* встречается фрагмент:

А характерным для всех геномов вида *Salmonella enterica* (отсутствующим у *S. bongori*) является также фрагмент



В некоторых геномах эта структура дополняется третьим вхождением симметричного мотива.

Разновидностей регулярно расположенных повторов, по-видимому, не так много в геномах сальмонеллы. Они представлены, например, повторами малых или транспортных РНК. Однако наиболее ярко проявляют себя CRISPR-локусы, представляющие собой точные повторы длиной 29 нуклеотидов с расстоянием между ними в 32 нуклеотида с различной кратностью повторений:

```
4228754   cggtttatccccgctggcgcggggaacacctcttccacacaacgttgtccatgtaaggcaa
4228815   cggtttatccccgctggcgcggggaacactcaatatgatccggcgtcagacgcggcgctga
4228876   cggtttatccccgctggcgcggggaacacgtccgtctgcgtaccctgattaacggctatta
4228937   cggtttatccccgctggcgcggggaacactcactgcctgaaagcgtgatgccgtgtttcac
4228998   cggtttatccccgctggcgcggggaacacgtttgtgacgtcgtcttttacgccaatgtgtt
4229059   cggtttatccccgctggcgcggggaacactggtccggtagttgtgttgggtcgccgtcgag
4229120   cggtttatccccgctggcgcggggaacacccgtgaatctcctccgccgcttggcgttgttc
...
```

Большинство представителей подвида *S. enterica enterica* содержат по два локуса CRISPR1 и CRISPR2 с расстоянием между ними примерно 16 т.п.н. Выделяется геном S-salamae1, содержащий три локуса CRISPR кратности 34, 28 и 19, соответственно. Примерно половина геномов разных подвидов содержит по одному локусу CRISPR, а в ряде геномов, например, у Se-paratyphiB и S-houtenae2 CRISPR-локусы не выявлены.

Наибольшая кратность повторений наблюдается у представителей сероваров *typhimurium* [38] и *infantis*. Однако лидером по кратности повторений в нашей подборке геномов является S-diarizonae2 (66 у CRISPR1 и 40 у CRISPR2). В то же время S-diarizonae1 содержит только один локус CRISPR кратности 17. В [38] обсуждается вопрос о положительной корреляции между CRISPR и патогенностью бактерий.

ЗАКЛЮЧЕНИЕ

Сравнение геномов сальмонеллы, относящихся к разным видам, подвидам и серогруппам на основе понятия относительной сложности, т.е. по числу компонентов в попарных разложениях последовательностей друг по другу, в целом отражает согласованность с общепринятой классификацией. Так, геномы, относящиеся к подвиду *Salmonella enterica* subsp. *enterica* образуют отдельный кластер на дереве филогенетических связей (рис. 2), другой кластер состоит из геномов подвидов *arizonae* и *diarizonae*. Однако есть и некоторые исключения. Так, последовательности серовара *infantis* не демонстрируют высокой близости между собой, а одна из последовательностей серовара *enteritidis* оказывается ближе к последовательностям серовара *typhimurium*, чем к последовательностям своего серовара.

Внутри каждого из проанализированных 32 геномов обнаружены довольно крупные области дупликаций, охватывающие большое количество (иногда сотни) генов. В некоторых геномах эти области консервативные, в других разрушены точечными заменами. Часть дупликаций носят тандемный характер.

Локальные структурные закономерности представлены, в основном, тандемными и регулярными повторами. Бактериальные геномы в целом бедны проявлениями повторности, в них отсутствуют микро- и минисателлиты, присущие эукариотами. Однако количество тандемных повторов в геномах сальмонеллы в несколько раз меньше, чем, например, в геномах бактерии *Y. pseudotuberculosis*. Если в каждом геноме *Y. pseudotuberculosis* выделяется порядка 300 тандемных повторов, длина которых 25 и выше, то геномы сальмонеллы содержат от 11 до 44 аналогичных структур. Несопоставимы также и длины мономеров, составляющих тандемные повторы у представителей бактериальных геномов псевдотуберкулеза и сальмонеллы, но уже в пользу сальмонеллы. Максимальная длина тандемно повторяющегося мономера в геномах бактерии *Yersinia pseudotuberculosis*, рассмотренных в [25] составляет 188 нуклеотидов. Самый длинный тандемный повтор в геноме S-diarizonae2 содержит несколько генов и межгенных участков и имеет длину $13000 = 6401 \times 2.03$. И этот случай не единичный, в рассматриваемой подборке геномов выявлено более 30 тандемных повторов, длина мономера которых превышает 188 нуклеотидов.

Выделяется крайне мало так называемых фракталоподобных структур, представляющих собой тандемные повторы палиндромов или комплементарных палиндромов, что позволяет даже высказать гипотезу о наличии некоторого запрета на формирование таких структур. Зато ярко проявляют себя регулярно расположенные повторы. Это и повторы малых РНК, но в большей степени CRISPR-локусы.

Многие повторы характеризуются возможностью их многозначного толкования. Это совершенные тандемные повторы, которые можно трактовать и как несовершенные, но с меньшей длиной мономера; компаунды (тандемы из тандемных повторов), которые также можно отнести и к несовершенным повторам; несовершенные повторы, некоторые копии которых совершенны; уже упомянутые фрактальные структуры, т.е. тандемные повторы, образующие длинные палиндромы или комплементарные палиндромы.

Работа выполнена при поддержке гранта Российского научного фонда (проект № 23-25-00520).

СПИСОК ЛИТЕРАТУРЫ

1. Popoff M.Y., Bockemüh J., Brenner F.W. Supplement 1998 (no. 42) to the Kauffmann-White scheme. *Res. Microbiol*. 2000. V. 151. № 1. P. 63–65. doi: 10.1016/s0923-2508(00)00126-1
2. Brenner F.W., Villar R.G., Angulo F.J., Tauxe R., Swaminathan B. *Salmonella* nomenclature. *Clin. Microbiol*. 2000. V. 38. № 7. P. 2465–2467. doi: 10.1128/JCM.38.7.2465-2467.2000
3. Knodler L.A., Elfenbein J.R. Salmonella enterica. *Trends Microbiol*. 2019. V. 27. № 11. P. 964–965. doi: 10.1016/j.tim.2019.05.002
4. Braden Ch.R. *Salmonella enterica* serotype Enteritidis and eggs: a national epidemic in the United States. *Clin. Infect. Dis*. 2006. V. 43. № 4. P. 512–517. doi: 10.1086/505973
5. Alali W.Q., Thakur S., Berghaus R.D., Martin M.P., Gebreyes W.A. Prevalence and distribution of Salmonella in organic and conventional broiler poultry farms. *Foodborne Pathog. Dis*. 2010. V. 7. № 11. P. 1363–1371. doi: 10.1089/fpd.2010.0566
6. Johnson R., Mylona E., Frankel G. Typhoidal Salmonella: Distinctive virulence factors and pathogenesis. *Cell Microbiol*. 2018. V. 20. № 9. P. e12939.

7. Crump J.A., Luby S.P., Mintz E.D. The global burden of typhoid fever. *Bull. World Health Organ*. 2004. V. 82. № 5. P. 346–353.
8. Buckle G.C., Walker C.L., Black R.E. Typhoid fever and paratyphoid fever: Systematic review to estimate global morbidity and mortality for 2010. *J. Glob. Health*. 2012. V. 2. № 1. P. 010401. doi: 10.7189/jogh.02.010401
9. Chen J., Long J.E., Vannice K., Shewchuk T., Kumar S., Duncan S.A., Zaidi A.K.M. Taking on Typhoid: Eliminating Typhoid Fever as a Global Health Problem. *Open Forum Infect Dis*. 2023. V. 10. № 1. P. S74–S81. doi: 10.1093/ofid/ofad055
10. Aljeldah M.M. Antimicrobial Resistance and Its Spread Is a Global Threat. *Antibiotics*. 2022. V. 11. № 8. P. 1082. doi: 10.3390/antibiotics11081082
11. Chang Y.-J., Chen C.-L., Yang H.-P., Chiu C.-H. Prevalence, Serotypes, and Antimicrobial Resistance Patterns of Non–Typhoid Salmonella in Food in Northern Taiwan. *Pathogens*. 2022. V. 11. № 6. P. 705. doi: 10.3390/pathogens11060705
12. Salam M.A., Al-Amin M.Y., Salam M.T., Pawar J.S., Akhter N., Rabaan A.A., Alqumber M.A.A. Antimicrobial Resistance: A Growing Serious Threat for Global Public Health. *Healthcare*. 2023 V. 11. № 13. P. 1946. doi: 10.3390/healthcare11131946
13. O'Neill J. *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations / the Review on Antimicrobial Resistance chaired by Jim O'Neill*. 2014. URL: https://wellcomecollection.org/works/rdpck35v (accessed 21.12.2023).
14. Wagenlehner F.M.E., Dittmar F. Global Burden of Bacterial Antimicrobial Resistance in 2019: A Systematic Analysis. *Eur. Urol*. 2022. V. 82. № 6. P. 658. doi: 10.1016/j.eururo.2022.08.023
15. Adebisi Y.A., Alaran A.J., Okereke M., Oke G.I., Amos O.A., Olaoye O.C., Oladunjoye I., Olanrewaju A.Y., Ukor N.A., Lucero-Prisno D.E. COVID-19 and Antimicrobial Resistance: A Review. *Infect. Dis*. (*Auckl*). 2021. V. 14. doi: 10.1177/11786337211033870
16. Lewis K. The science of antibiotic discovery. *Cell*. 2020. V. 181. № 1. P. 29–45. doi: 10.1016/j.cell.2020.02.056
17. Dheman N., Mahoney N., Cox E.M., Farley J.J., Amini T., Lanthier M.L. An Analysis of Antibacterial Drug Development Trends in the United States, 1980–2019. *Clin. Infect. Dis*. 2021. V. 73. № 11. P. e4444–e4450. doi: 10.1093/cid/ciaa859
18. Luong T., Salabarria A.C., Roach D.R. Phage Therapy in the Resistance Era: Where Do We Stand and Where Are We Going? *Clin. Ther*. 2020. V. 42. № 9. P. 1659–1680. doi: 10.1016/j.clinthera.2020.07.014
19. Hatfull G.F., Dedrick R.M., Schooley R.T. Phage Therapy for Antibiotic–Resistant Bacterial Infections. *Annu. Rev. Med*. 2022. V. 73. P. 197–211. doi: 10.1146/annurev-med-080219-122208
20. Biswas A., Gagnon J.N., Brouns S.J.J., Fineran P.C., Brown C.M. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol*. 2013. V. 10. № 5. P. 817–827. doi: 10.4161/rna.24046
21. Barrangou R., Marraffini L.A. CRISPR-Cas systems: Prokaryotes upgrade to adaptive immunity. *Mol. Cell*. 2014. V. 54. № 2. P. 234–244. doi: 10.1016/j.molcel.2014.03.011
22. Назаров П.А. Альтернативы антибиотикам: литические ферменты бактериофагов и фаговая терапия. *Вестник Российского государственного медицинского университета*. 2018. № 1. С. 5–15. doi: 10.24075/vrgmu.2018.002
23. Royer S., Morais A.P., da Fonseca Batistão D.W. Phage therapy as strategy to face post-antibiotic era: a guide to beginners and experts. *Arch. Microbiol*. 2021. V. 203. № 4. P. 1271–1279. doi: 10.1007/s00203-020-02167-5
24. Nale J.Y., Ahmed B., Haigh R., Shan J., Phothaworn P., Thiennimitr P., Garcia A., AbuOun M., Anjum M.F., Korbsrisate S., Galyov E.E., Malik D.J., Clokie M.R.J. Activity of a Bacteriophage Cocktail to Control Salmonella Growth Ex Vivo in Avian,

Porcine, and Human Epithelial Cell Cultures. *Phage* (*New Rochelle*). 2023. V. 4. № 1. P. 11–25. doi: 10.1089/phage.2023.0001

25. Miroshnichenko L.A., Gusev V.D. Complete spectra of periodicities in the problems of differentiation of closely related bacterial genomes. *J. Phys.: Conf. Ser*. 2021. V. 1715. P. 012026. doi: 10.1088/1742-6596/1715/1/012026
26. Гусев В.Д., Мирошниченко Л.А., Титкова Т.Н., Джиоев Ю.П., Козлова И.В., Парамонов А.П. Структурированные РНК-маркеры для генотипирования вируса клещевого энцефалита. *Математическая биология и биоинформатика*. 2018. Т. 13. № 1. С.13–37. doi: 10.17537/2018.13.13
27. Dimovski K., Cao H., Wijburg O.L., Strugnell R.A., Mantena R.K., Whipp M., Hogg G,. Holt K.E. Analysis of Salmonella enterica serovar Typhimurium variable-number tandem-repeat data for public health investigation based on measured mutation rates and whole-genome sequence comparisons. *J. Bacteriol*. 2014. V. 196. P. 3036–3044. doi: 10.1128/JB.01820-14.
28. Kjeldsen M.K., Torpdahl M., Pedersen K., Nielsen E.M. Development and comparison of a generic multiple-locus variable-number tandem repeat analysis with pulsed-field gel electrophoresis for typing of Salmonella enterica subsp. enterica. *J. Appl. Microbiol*. 2015. V. 119. P. 1707–1717. doi: 10.1111/jam.12965.
29. Колмогоров А.Н. Три подхода к определению понятия "количество информации". *Проблемы передачи информации*. 1965. Т. 1. № 1. С. 3–11.
30. Гусев В.Д., Мирошниченко Л.А. Сложность ДНК-последовательностей. Различные подходы и определения. *Математическая биология и биоинформатика*. 2020. Т. 15. № 2. С. 313–337. doi: 10.17537/2020.15.313
31. Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences. *Bioinformatics*. 1999. V. 15. № 12. P. 994–999.
32. Lempel A., Ziv J. On the complexity of finite sequences. *IEEE Trans. Inform. Theory*. 1976. V. IT-22. № 1. P. 75–81.
33. Гусев В.Д., Мирошниченко Л.А., Чужанова Н.А. Выявление фракталоподобных структур в ДНК-последовательностях. *Information Science & Computing. Classification, Forecasting, Data Mining*. 2009. № 8. P. 117–123. (International Book Series).
34. Sneath P.H.A., Sokal R.R. *Numerical Taxonomy*. San Francisco: Freeman, 1973.
35. Kumar S., Stecher G., Li M., Knyaz C., Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. *Molecular Biology and Evolution.* 2018. V. 35. P. 1547–1549.
36. Saitou N., Nei M. The neighbour-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 1987. V. 4. P. 406–425.
37. Hauth AM, Joseph DA. Beyond tandem repeats: complex pattern structures and distant regions of similarity. *Bioinformatics*. 2002. V. 18. № 1. P. S31–37. doi: 10.1093/bioinformatics/18.suppl_1.s31
38. Kushwaha S.K., Bhavesh N.L.S., Abdella B., Lahiri C., Marathe S.A. The phylogenomics of CRISPR-Cas system and revelation of its features in *Salmonella*. *Sci. Rep*. 2020. V. 10. Article No. 21156. doi: 10.1038/s41598-020-77890-6

Рукопись поступила в редакцию 13.12.2023, переработанный вариант поступил 19.12.2023.
Дата опубликования 30.12.2023.

Repeat Structure in *Salmonella* genomes

Miroshnichenko L.A.[1], Arefieva N.A.[2,3], Dzhioev Yu.P.[2], Gusev V.D.[1], Borisenko A.Yu.[2], Erdyneev S.V.[2], Bukin Yu.S.[4,5]

[1]*Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia*
[2]*Irkutsk State Medical University, Irkutsk, Russia*
[3]*Research Center for Family Health and Human Reproduction, Irkutsk, Russia*
[4]*Limnological Institute SB RAS, Irkutsk, Russia*
[5]*Irkutsk State University, Irkutsk, Russia*

Abstract. *Salmonella* is a genus of gram-negative, facultative and non-spore-forming anaerobic bacteria. The genus includes two species: *S. bongori* and *S. enterica*. All *Salmonella*, which are pathogenic to humans and animals, belong to the species *S. enterica*. It includes seven subspecies, which includes more than 2500 serotypes. According to global statistics, they cause about 2.8 billion cases of diarrheal diseases annually, and the mortality rate reaches more than 300000 cases. *S. enterica* spp strains that have acquired multidrug resistance to antibiotics have become especially dangerous. Against the backdrop of this global problem, a detailed study of the complete genomes of various representatives of the genus *Salmonella* becomes relevant. Repeats play an important role in the regulation of basic genetic processes during the life cycle and evolution. The work examines various manifestations of repetition in the genomes of *S. enterica*. Taking into account common fragments of different genomes serves as the basis for the formation of a matrix of pairwise relative complexity, which is used in constructing a phylogenetic tree. Long repeats within individual genomes typically correspond to large individual duplications. The main attention is paid to local structural regularities, most of which are represented by tandem repeats. Of significant interest are multivalued repeats, such as tandem repeats forming a palindrome, repeats with regular substitutions or complex monomer structure.

Key words: *bacterial genome, salmonella, complex decompositions, repeats, local structural regularities, tandem repeats.*