Математическая биология и биоинформатика
2023. Т. 18. № 2. С. 418–433. doi: 10.17537/2023.18.418

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

Анализ эффективности микс-сборки метатранскриптомных наборов данных в исследовании вирусных сообществ

Букин Ю.С., Бондарюк А.Н., Бутина Т.В.*

Федеральное государственное бюджетное учреждение науки Лимнологический институт Сибирского отделения Российской академии наук, Иркутск, Россия

Аннотация. В данной работе проведен сравнительный анализ раздельной и комбинированной («микс») сборки метатранскриптомных данных для исследования вирусных сообществ в нескольких образцах на примере четырех метатранскриптомов эндемичных байкальских моллюсков *Benedictia baicalensis.* Анализ показал, что микс-сборка по сравнению с раздельной сборкой образцов увеличивает количество вирусных контигов (или скаффолдов) на образец, количество идентифицированных виротипов, среднюю длину скаффолдов на образец и долю собранных вирусных прочтений от общего количества прочтений в образцах. Микс-геномные *de novo* сборки с использованием скрытых марковских моделей для идентификации вирусов представляют данные в виде таблицы с количеством прочтений из разных образцов для каждого скаффолда (таблица представленности). Такая таблица позволяет сравнивать образцы по представленности всех вирусных скаффолдов, в том числе, не имеющих аналогов в известных базах данных, то есть для которых не удалось установить таксономическую принадлежность. Таким образом, микс-геномные сборки позволяют проводить сравнительный анализ с учетом скрытого разнообразия вирусов. В работе предложен конвейер по анализу данных метатранскриптомов с применением микс-геномной *de novo* сборки для исследования вирусов, которым могут воспользоваться другие исследователи.

Ключевые слова: *метагеномика, транскриптомика, вирусы, вирусные сообщества, метагеномная сборка, микс-сборка, метатранскриптомный анализ, вирусные скаффолды.*

ВВЕДЕНИЕ

Метагеномика совершила прорыв в медицинских и биологических исследованиях; весомый вклад и особую значимость она приобрела в исследовании генетического разнообразия вирусов в природе [1]. Вирусы, как известно, не имеют универсальных генов-маркеров; поэтому, высокопроизводительные технологии секвенирования пришлись как нельзя кстати для исследования вирусных сообществ в различных экосистемах, и раскрытия глобального генетического разнообразия вирусов. Метагеномика позволяет обнаруживать вирусы в различных образцах из окружающей среды, а также вирусы, ассоциированные с различными многоклеточными организмами (растениями, животными и человеком) [2–5].

Одно из направлений вирусной метагеномики это исследование разнообразия РНК-содержащих вирусов в окружающей среде или в составе многоклеточных организмов. Для таких исследований используются или выделенные из образцов вирусные частицы

*tvbutina@mail.ru

(обогащенный вирусный материал) или изолированная тотальная РНК (метатранскриптомный анализ) [6–8]. Метатранскриптомика позволяет идентифицировать в образцах не только генетический материал РНК-содержащих вирусов, но и экспрессирующиеся гены ДНК-содержащих вирусов. Исследование разнообразия вирусов в процессе анализа метатранскриптомов это многостадийный процесс, начиная от отбора проб и пробоподготовки до биоинформатического анализа геномных прочтений; и от каждой стадии зависит конечный результат.

Биоинформатический анализ метатранскриптомов для поиска генетического материала вирусов включает множество этапов (первичная обработка и тримминг по качеству первичных геномных прочтений, *de novo* сборка полногеномных данных, идентификация фрагментов вирусных геномов среди скаффолдов или контигов метагеномной сборки, таксономическая идентификация фрагментов вирусных геномов и др.); для выполнения каждого этапа существуют различные подходы, программы и ресурсы [9–12]. Развитие биоинформатики привело к широкому внедрению методов машинного обучения на основе скрытых марковских моделей (HMM, Hidden Markov Model) [13–15], которые позволяют среди скаффолдов или контигов геномной сборки с высокой эффективностью обнаруживать фрагменты вирусных геномов или полные вирусные геномы. Предварительное получение HMM-профилей для идентификации вирусов осуществляется на основе полногеномных баз данных. Во многих случаях вероятностные оценки такой идентификации показывают высокую достоверность, но близкородственные геномы для таких предсказанных вирусов в геномных базах данных не обнаруживаются. Таким образом, HMM-алгоритмы дают возможность исследовать скрытое разнообразие вирусов.

Часто в ходе исследования встает задача сравнения нескольких образцов, при этом для каждого набора данных производится геномная сборка, которая анализируется одной из программ, использующих HMM-алгоритм для поиска вирусов. Затем полученные контиги или скаффолды – фрагменты вирусных геномов, идентифицируются путем сопоставления с базами данных известных вирусных геномов. Дальнейшее сравнение нескольких образцов может быть выполнено на основе таксономического или функционального состава скаффолдов каждого набора данных. При таком сравнении из рассмотрения часто выпадают «гипотетические» вирусы, предсказанные HMM алгоритмом, но не идентифицируемые до какого-либо таксономического уровня, хотя в некоторых случаях такие «гипотетические» вирусы входят в доминирующий пул вирусного сообщества.

Существует другой подход, при котором после фильтрации по качеству наборов данных из нескольких образцов все прочтения объединяются в один массив и производится единая *de novo* сборка («cross-assembly»). В наших предыдущих метавиромных исследованиях мы использовали этот подход для сравнения вирусного разнообразия нескольких образцов [16–18]. В полученной геномной сборке (далее микс-сборке) скаффолды тестируются одним из алгоритмов на основе HMM метода для поиска потенциальных фрагментов вирусных геномов (фрагментов геномов РНК-содержащих вирусов и транскрибируемых генов ДНК-содержащих вирусов в случае метатранскриптомов). На следующем этапе исходные метатранскриптомные прочтения каждого образца по отдельности картируются на вирусные скаффолды, так определяется представленность потенциальных вирусов в каждой пробе по отдельности. Данный подход позволяет сравнить анализируемые образцы с учетом вирусных скаффолдов, для которых не установлена таксономическая принадлежность.

В данной работе мы поставили цель сравнить результаты исследования разнообразия вирусных сообществ в метатранскриптомных наборах данных (на примере полученных ранее наборов из нескольких образцов моллюсков, отобранных в различных географических районах озера Байкал), анализируемых с помощью

раздельной и микс *de novo* геномных сборок. Проведенное сравнение демонстрирует преимущество второго подхода и возможные ограничения при его применении.

МЕТОДЫ

Исходные данные метатранскриптомного анализа

В настоящем исследовании мы использовали наборы данных (табл. 1), полученные нами из образцов моллюсков *Benedictia baicalensis* – одного из наиболее многочисленных и распространенных эндемичных видов озера Байкал. Данные были загружены в базу NCBI в виде SRA архивов (BioProject PRJNA1029953). Всего в анализе участвовало 4 образца (каждый содержал пул из пяти особей) с трех станций отбора (табл. 1). Образцы секвенировали в двух повторностях с помощью секвенатора DNBSEQ-400 (MGI Tech Co., Ltd., China), в результате чего были получены парноконцевые прочтения длинной по 150 пар оснований. Повторности одной пробы объединяли в один массив данных и в итоге обрабатывали четыре пробы с кодировками RNA_57_58, RNA_59_60, RNA_61_62 и RNA_63_64. Первичный тримминг прочтений по качеству проводился в программе «Trimmomatic» [19] (опции для тримминга MAXINFO:40:0.05 AVGQUAL:15 MINLEN:90).

Таблица 1. Наборы метатранскриптомных данных, использованные для сравнительного анализа

Образцы	Станции отбора проб	Координаты	NCBI SRA BioSamples	Количество пар прочтений
RNA_57_58	залив Лиственничный	51°51’51.77”N 104°50’37.80”E	SAMN37882679	64 793 960
RNA_59_60	Ушканьи о-ва	53°51’05.76”N 108°42’28.46”E	SAMN37882680	65 597 110
RNA_61_62	пос. Большие Коты	51°54’08.66”N 105°06’13.04”E	SAMN37882681	67 042 790
RNA_63_64	залив Лиственничный	51°51’51.77”N 104°50’37.80”E	SAMN37882682	63 723 122

De novo сборка метатранскриптомных данных

Сборка метатранскриптомных данных проводилась с помощью программы SPAdes 3.13.1 [20] в режиме сборки метагеномных данных (metaSPAdes), при запуске программы задавались разные длины *k*-mer ($k \in \{21, 33, 55, 77\}$). Сборка данных проводилась для каждого образца по отдельности (RNA_57_58, RNA_59_60, RNA_61_62 и RNA_63_64) и для смешанного (микс) набора данных, включающего прочтения всех четырех образцов. Для дальнейшего анализа из сборок были отобраны только скаффолды длинной 500 и более пар оснований. Полная схема конвейера анализа данных приведена на рисунок 1.

Идентификация вирусных скаффолдов

Для идентификации вирусных скаффолдов и открытых рамок считывания вирусных белков использовалось приложение VirSorter2 [21] с использованием встроенных баз данных ДНК- и РНК-содержащих вирусов. Использование этих баз данных позволило идентифицировать скаффолды – фрагменты геномов РНК-содержащих вирусов и продукты транскрипции геномов ДНК-содержащих вирусов. Результат идентификации вирусных скаффолдов был дополнительно протестирован в программе CheckV [22]. Скаффолды, идентифицированные как вирусные в приложении VirSorter2 и

подтверждённые как вирусные в приложении CheckV, использовались для дальнейшего анализа.

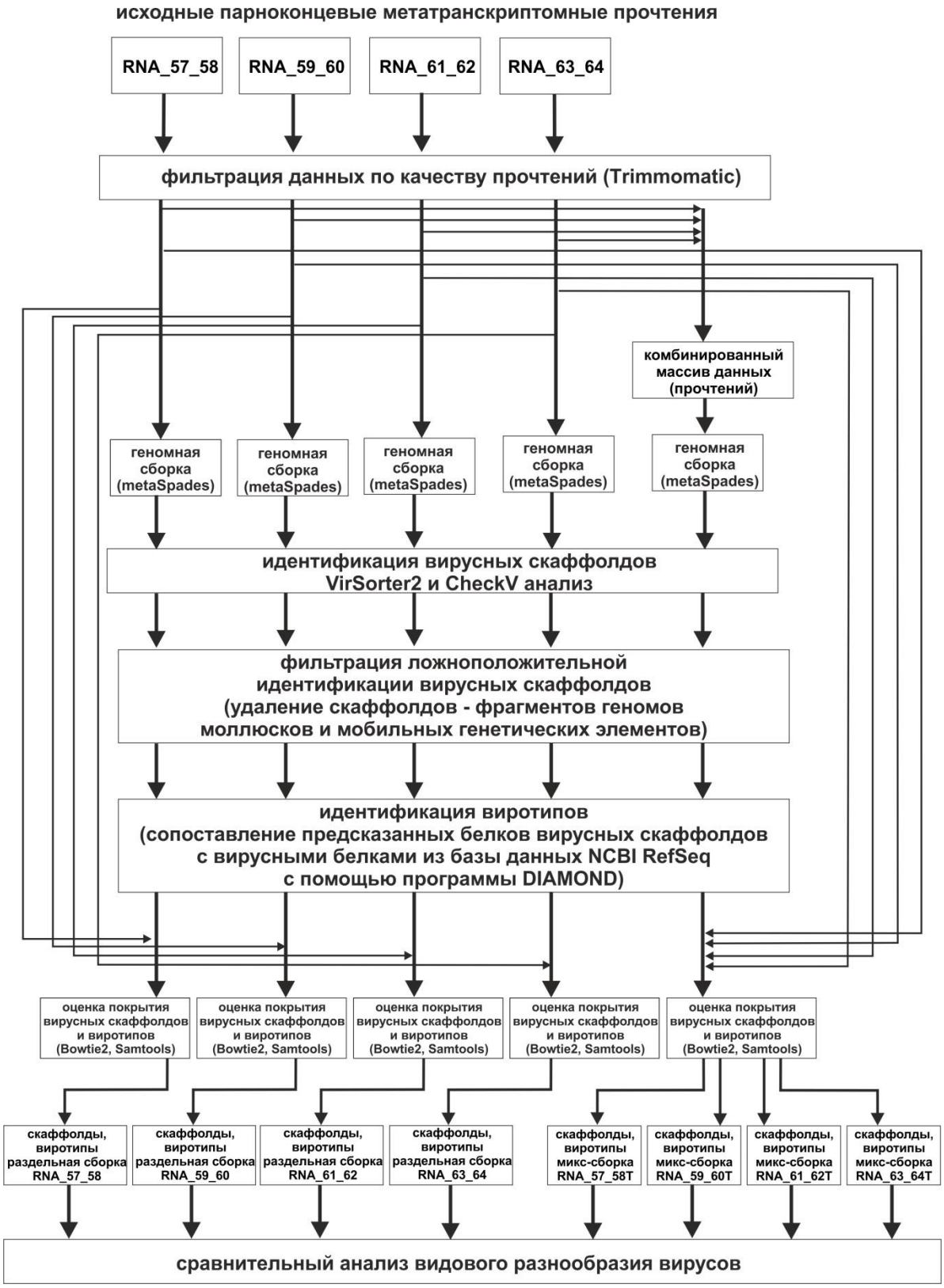

Рис. 1. Общая схема конвейера по обработке метатранскриптомных данных для идентификации ДНК- и РНК-содержащих вирусов в образцах. Стрелками показаны потоки исходных и промежуточных данных на каждой стадии анализа.

Фильтрация ложноположительных результатов

Часть скаффолдов, идентифицированных приложениями VirSorter2 и CheckV как вирусные, могли оказаться транскрибируемыми фрагментами геномов или мобильными генетическими элементами моллюсков. Поэтому полученные вирусные скаффолды были сопоставлены с полными геномами моллюсков из базы NCBI, собранными до хромосомного уровня точности (*Biomphalaria glabrata*, *Pomacea canaliculata*, *Gigantopelta aegis* и *Patella vulgata*), с помощью приложения BLASTn (опции для анализа: word size = 15, gapopen = 2, gapextend = 1, reward = 1, penalty = 1) и полными протеомами этих же моллюсков с помощью приложения DIAMOND [23] (опция для запуска more-sensitive). Если предполагаемый вирусный скаффолд выравнивался более чем на 30 % своей длины с фрагментом генома моллюсков (*e*-value ≤ 0.00001) или 300 п.н. скаффолда (*e*-value ≤ 0.00001) имело высокое сходство с каким-либо белком моллюсков, то такой скаффолд считался невирусным и удалялся из анализа.

В оставшихся вирусных скаффолдах был проведен поиск мобильных элементов с помощью алгоритма, основанного на HMM, и приложения DfamScan [24]. Для анализа использовались HMM профили различных видов беспозвоночных (*Anopheles coluzzii*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Halyomorpha halys*, *Heliconius erato demophoon*, *Heliconius melpomene*), доступные в базе данных Dfam (версия 3.7) [25]. Если в предполагаемом вирусном скаффолде обнаруживался мобильный элемент с *e*-value ≤ 0.00001, то такой скаффолд также считался невирусным и удалялся из анализа.

Таксономическая идентификация вирусных скаффолдов

Таксономическую идентификацию вирусных скаффолдов проводили путем сравнения транслированных открытых рамок считывания (ORFs) вирусных белков, предсказанных программой VirSorter2, с полной базой вирусных протеомов NCBI RefSeq [26]; данная процедура подробно описана в наших предыдущих работах [17, 18]. Сравнение проводилось путем выравнивания с помощью приложения DIAMOND [23] (опция для запуска more-sensitive). Вирусные ORFs считались идентифицированными при *e*-value ≤ 0.00001 и bit-score ≥ 50. Для каждой ORF в скаффолде выбиралось выравнивание с наибольшим bit-score из нескольких найденных вариантов. Если в одном скаффолде было несколько ORFs, соответствующих разным таксонам (NCBI RefSeq ID), то в качестве виротипа выбирался RefSeq ID, выравнивание которого с каким-либо ORF из скаффолда имело наибольший процент сходства. Если скаффолд содержал несколько (две и более) ORFs, имеющих сходство с разными белками одного вируса, то этот вирус выбирался в качестве виротипа этого скаффолда в независимости от значения bit-score и процента сходства.

Оценка представленности вирусных скаффолдов в образцах

Для анализа покрытия вирусных скаффолдов прочтениями использовалась комбинация приложений Bowtie2 [27] и SAMtools [28]. Для скаффолдов, полученных путем раздельных сборок проб, проводилось картирование первичных прочтений каждой пробы на свою отдельную сборку. Количество прочтений на каждый скаффолд в пробе пересчитывалось в относительное количество (долю в %) и нормировалось на среднее количество вирусных прочтений на пробу (относительное количество прочтений на скаффолд умножалось на среднее количество вирусных прочтений, картированных на вирусные скаффолды в исследуемых пробах).

На вирусные скаффолды, полученные в результате микс-сборки, первичные прочтения каждой пробы картировались по отдельности. Из микс-сборки было сформировано четыре отдельных набора вирусных скаффолдов. В каждый из наборов

вошли только те скаффолды, на которые картировались прочтения соответствующей пробы. В каждом из четырех наборов вирусных скаффолдов была проведена нормировка на относительное количество (долю в %) и среднее количество вирусных прочтений на пробу. Итоговые данные были представлены в виде таблиц, в которых строки соответствовали вирусным скаффолдам, а четыре столбца – показателям представленности (доли либо усреднённого количества прочтений на пробу) этих скаффолдов в исходных пробах.

На основе принадлежности скаффолдов к определенным виротипам были составлены таблицы представленности вирусных семейств (доля прочтений либо количество прочтений, нормированных на усреднённое количество прочтений на пробу) и виротипов в пробах. Доля прочтений и количество прочтений, нормированных на усреднённое количество прочтений на пробу, для скаффолдов одного семейства и виротипа суммировались.

Сравнительный анализ видового разнообразия вирусов в метатранскриптомных данных

Для всех проб на основе данных о количестве прочтений, нормированных на среднее количество вирусных прочтений, были рассчитаны индексы Шеннона H по формуле:

$$H = -\sum_{i=1}^{N} p_i \ln(p_i), \text{ где } p_i = \frac{n_i}{\sum_{j=1}^{N} n_j},$$

где N – общее количество вирусных скаффолдов в пробе, p_i – доля прочтений, приходящихся на i-тый вирусный скаффолд из общего количества вирусных прочтений в пробе, n_i – количество прочтений, приходящихся на i-тый вирусный скаффолд в пробе. Для расчётов индексов Шеннона использовался пакет vegan [29] языка программирования R. Данные о представленности вирусных семейств в раздельных и микс-сборках были агрегированы в единый массив данных и визуализированы в виде тепловой карты с помощью пакета gplots языка программирования R. Столбцы (пробы) в тепловой карте были скластеризованы по степени сходства состава семейств вирусов с помощью метода average linkage clustering на основе расстояний Брея – Кертиса. Для расчета расстояний Брея – Кертиса применялся пакет vegan [29] языка программирования R. Визуализация данных в виде столбчатых диаграмм, круговых диаграмм и гистограмм распределения проводилась стандартными средствами языка программирования R.

РЕЗУЛЬТАТЫ

Сравнение разнообразия вирусных сообществ в раздельной и микс-геномной сборках на уровне вирусных скаффолдов

После метагеномной сборки и идентификации вирусов приложениями VirSorter2 и CheckV [21, 22] в раздельных сборках метатранскриптомных данных число потенциальных вирусных скаффолдов лежало в диапазоне от 141 до 297, а в микс-сборке данных число потенциальных вирусных скаффолдов равнялось 343 (табл. 2). В раздельных сборках среди потенциальных вирусных скаффолдов количество ложноположительных идентификаций, совпадающих с фрагментами геномов и протеомов моллюсков, находилось в пределах от 86 до 156. В микс-сборке количество ложноположительных идентификаций равнялось 214. В раздельных сборках количество потенциальных вирусных скаффолдов, идентифицированных как мобильные элементы геномов, находилось в пределах от 15 до 31. В микс-сборке количество потенциальных вирусных скаффолдов, идентифицированных как

мобильные генетические элементы, равнялось 32. Из таблицы 2 видно, что количество ложноположительных идентификаций вирусов среди сборок метатранскриптомных данных не зависит от первоначального количества потенциальных вирусных скаффолдов. Доля ложноположительных идентификаций вирусных скаффолдов от первоначального количества потенциальных вирусных скаффолдов лежит в пределах от 54 % до 76 %. В микс-сборке доля ложноположительных идентификаций вирусных скаффолдов составила 63 %.

Таблица 2. Общие характеристики метатранскриптомных сборок при идентификации вирусов

Количество скаффолдов	**Идентификатор сборки**				
	RNA_57_58	**RNA_59_60**	**RNA_61_62**	**RNA_63_64**	**Микс-сборка**
Количество скаффолдов метагеномной сборки метатранскриптомных данных	100 644	160 521	131 610	92 229	218 982
Количество потенциальных вирусных скаффолдов после анализа с помощью VirSorter2 и CheckV	194	207	167	141	343
Количество потенциальных вирусных скаффолдов, идентифицированных как фрагменты геномов и протеомов моллюсков	102	156	117	86	214
Количество потенциальных вирусных скаффолдов, идентифицированных как мобильные элементы	15	31	20	23	32
Итоговое количество вирусных скаффолдов	89	48	50	50	126

Результаты сравнения разнообразия вирусных сообществ в метатранскриптомных образцах представлены на рисунке 2. При использовании микс-сборки, по результатам картирования исходных прочтений (рис. 2,a) количество вирусных скаффолдов во всех пробах возросло на 7–83 %. Индекс разнообразия Шеннона (рис. 2,b) для вирусных сообществ, подсчитанный на основе информации о представленности скаффолдов (см. раздел Методы), увеличился для трех проб (RNA_59_60, RNA_61_62 и RNA_63_64). Количество идентифицированных виротипов во всех четырех пробах (рис. 2,c) возросло на 10–93 %. Во всех пробах (рис. 2,d) также возросла общая доля прочтений (на 2–93 %), приходящаяся на вирусные последовательности. Анализ разнообразия вирусных сообществ на уровне скаффолдов показывает, что применение микс-сборки увеличивает (в некоторых образцах значительно) количество вирусных скаффолдов, виротипов и долю вирусных прочтений. Таким образом, микс-сборка метатранскриптомных данных позволяет раскрыть в виромном исследовании большее разнообразие вирусов в каждой из исследуемых проб. Это происходит благодаря вовлечению в анализ на уровне *de novo* сборки большего количества исходных прочтений.

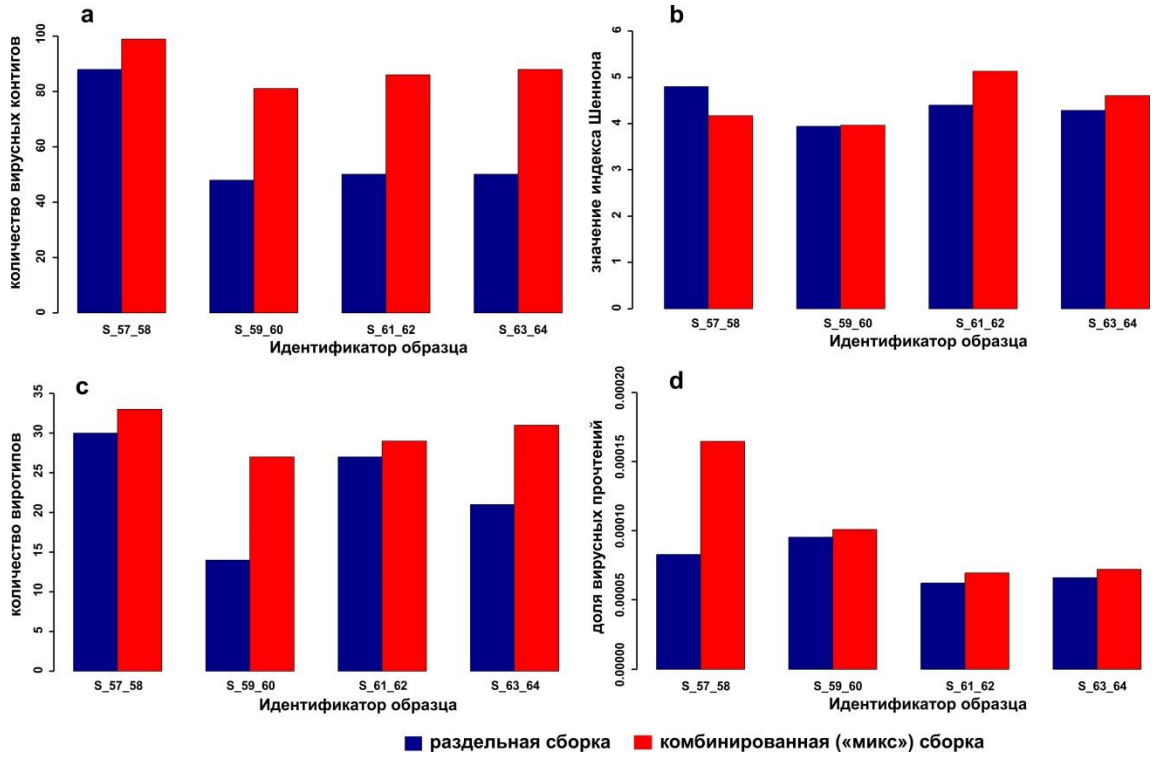

Рис. 2. Результаты сравнительного анализа разнообразия вирусных сообществ байкальских моллюсков *B. baicalensis*, полученные при использовании раздельной и микс-геномной сборки: **a**) столбчатая диаграмма сравнения количества вирусных скаффолдов в пробах; **b**) значения индексов разнообразия Шеннона, подсчитанные на основе информации о представленности скаффолдов (нормированном количестве прочтений); **c**) сравнение количества идентифицированных в пробах виротипов; **d**) доли вирусных прочтений в общем пуле метатранскриптомных прочтений при использовании разного типа сборок.

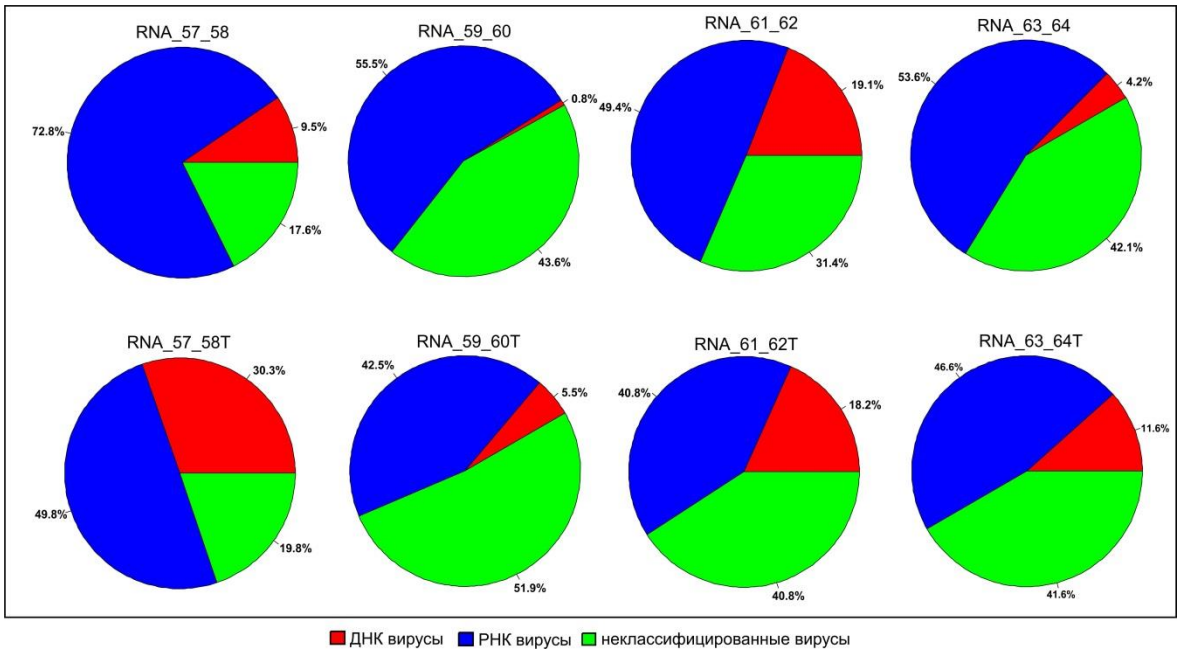

Рис. 3. Круговые диаграммы сравнения количества прочтений (в %), приходящихся на разные категории вирусов, по результатам таксономического анализа вирусных скаффолдов в метатранскриптомных сборках. RNA_57_58, RNA_59_60, RNA_61_62 и RNA_63_64 – раздельные геномные сборки. RNA_57_58T, RNA_59_60T, RNA_61_62T и RNA_63_64T – микс-геномная сборка прочтений из образцов *B. baicalensis*.

Распределение прочтений среди различных категорий вирусов

В число идентифицированных вирусных скаффолдов (рис. 3) вошли фрагменты геномов РНК-содержащих вирусов и активно транскрибируемые гены ДНК-содержащих вирусов. Некоторые скаффолды были распознаны методами машинного обучения, заложенными в приложениях VirSorter2 и CheckV, как вирусные, но с помощью алгоритма DIAMOND, основанного на попарном выравнивании последовательностей, в базе данных полных вирусных протеомов NCBI RefSeq не было идентифицировано для них каких-либо близкородственных вирусов (виротипов) с *e*-value ≤ 0.00001. В микс-сборке по сравнению с раздельными сборками (рис. 3) увеличилось количество прочтений, приходящихся на гены ДНК-содержащих вирусов, и количество не идентифицированных до виротипа вирусов.

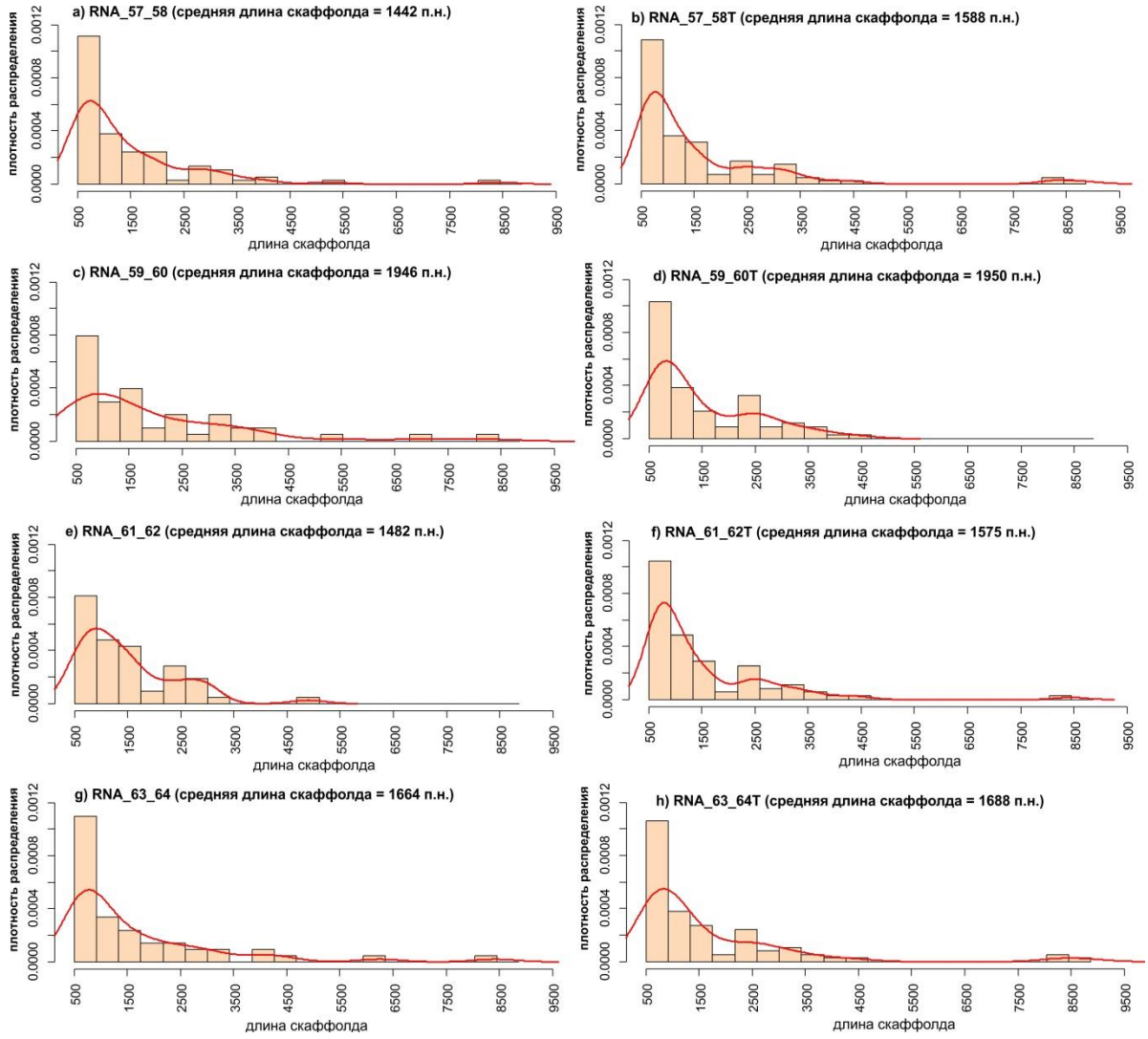

Рис. 4. Гистограммы распределения длин вирусных скаффолдов в сборках метатранскриптомных наборов данных из образцов байкальских моллюсков *B. baicalensis*. RNA_57_58, RNA_59_60, RNA_61_62 и RNA_63_64 – раздельные геномные сборки. RNA_57_58T, RNA_59_60T, RNA_61_62T и RNA_63_64T – микс-сборка.

Анализ распределения длин вирусных скаффолдов

Визуализация распределения длин вирусных скаффолдов (рис. 4) показывает, что использование микс-сборки приводит к увеличению (в среднем на 4 %) средних длин вирусных скаффолдов в анализируемых пробах по сравнению с раздельными сборками. Это связанно с тем, что вовлечение в процесс сборки большего количества прочтений

может приводить к увеличению покрытия и объединению в один скаффолд двух или более коротких скаффолдов – фрагментов одного генома РНК-содержащего вируса или фрагментов одного гена ДНК-содержащего вируса.

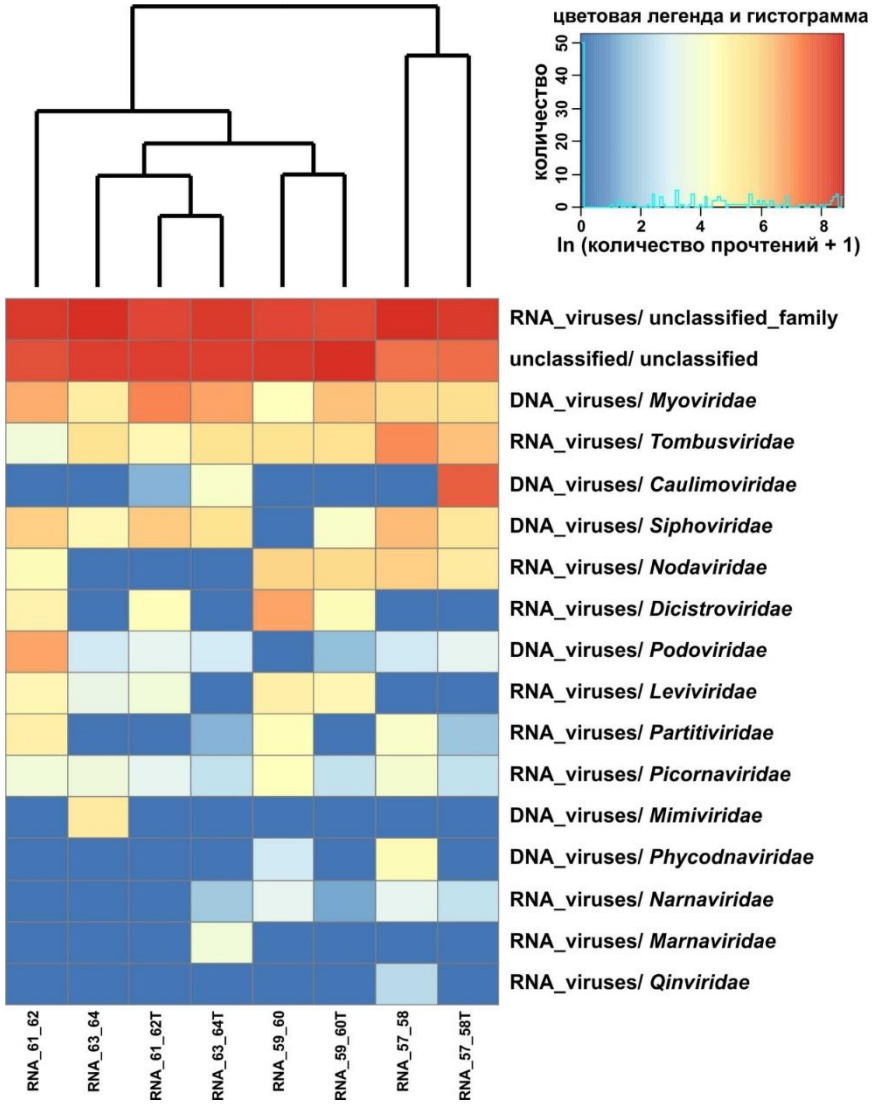

Рис. 5. Тепловая карта представленности (нормированного количества прочтений) вирусных семейств в раздельной (RNA_57_58, RNA_59_60, RNA_61_62 и RNA_63_64) и комбинированной сборке (RNA_57_58T, RNA_59_60T, RNA_61_62T и RNA_63_64T) метатранскриптомных наборов данных из образцов моллюсков *B. baicalensis*. Кластеризацию проводили методом average на основе расстояний Брея – Кертиса.

Распределение вирусов по семействам

Среди идентифицированных вирусных скаффолдов доминировали представители семейств РНК-содержащих вирусов (рис. 5). Большая часть вирусных прочтений приходилась на виротипы РНК-содержащих вирусов, не классифицированных до уровня семейства. Остальные РНК-содержащие вирусы принадлежали следующим семействам: *Tombusviridae*, *Nodaviridae*, *Dicistroviridae*, *Leviviridae*, *Partitiviridae*, *Picornaviridae*, *Narnaviridae*, *Marnaviridae* и *Qinviridae*. Все перечисленные семейства, кроме *Leviviridae*, включают вирусы, поражающие многоклеточные эукариотические организмы (растения, грибы, животные, в том числе позвоночные и беспозвоночные). Семейство *Leviviridae* включает прокариотические вирусы (РНК-содержащие фаги). Экспрессирующиеся в виде РНК гены ДНК-вирусов принадлежали следующим семействам: *Myoviridae*, *Siphoviridae*, *Podoviridae* (бактериофаги), *Caulimoviridae*

(вирусы растений), *Mimiviridae* (вирусы протист) и *Phycodnaviridae* (вирусы эукариотических микроводорослей). Спектр семейств вирусов, ассоциированных с моллюсками *B. baicalensis*, содержит вирусы, потенциальными хозяевами которых могут являться как сами моллюски, так и ассоциированная с ними микрофлора или объекты питания. Также существенное количество прочтений приходилось на скаффолды (рис. 5), не идентифицированные до виротипа.

На тепловой карте (рис. 5) пары проб RNA_59_60 и RNA_59_60T, RNA_57_58 и RNA_57_58T раздельной и микс-геномной сборок кластеризуются совместно, что указывает на близкий состав семейств вирусов. Пробы RNA_61_62 и RNA_61_62T, RNA_63_64 и RNA_63_64T при кластеризации не образуют парных кластеров. Проблема кластеризации сообществ по сходству состава вирусов в раздельных сборках осложняется невозможностью сопоставления скаффолдов разных сборок, для которых не определен виротип. Такие скаффолды входят в одну категорию – неидентифицированные вирусы, состав которых в разных пробах может сильно отличаться.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Наше исследование показало, что при изучении разнообразия вирусных сообществ и сравнении нескольких наборов метатранскриптомных данных использование микс-сборки имеет преимущество перед раздельными сборками. Основные преимущества при этом следующие: 1) увеличение количества вирусных контигов в каждом из образцов, 2) увеличение количества идентифицированных виротипов при таксономической идентификации вирусных скаффолдов в каждом из образцов, 3) увеличение средней длины вирусных скаффолдов в каждом из образцов и 4) увеличение доли вирусных прочтений в каждом из образцов. Все это свидетельствует о большей эффективности использования прочтений при анализе метатранскриптомов. Необходимо отметить, что во избежание получения химерных контигов на этапе микс *de novo* сборки метатранскриптомных данных с целью поиска вирусов необходимо применять специализированные метагеномные сборщики, такие как metaSPAdes [20], MEGAHIT [30], IDBA-UD [31] и другие [32].

Другим, более очевидным преимуществом микс-сборки является структура получаемых данных. При картировании исходных прочтений на объединенную сборку в итоговой таблице каждый из скаффолдов характеризуется количеством попадающих на него прочтений из каждой пробы по отдельности. Таким образом, по представленности скаффолдов в разных образцах можно напрямую сравнивать вирусные сообщества методами вычислительной экологии [33], не имея информации о таксономии этих скаффолдов. Микс-сборка позволяет в полной степени использовать методы машинного обучения, основанные на HMM алгоритмах [14, 21, 34], для идентификации вирусных скаффолдов. Программы, использующие HMM алгоритмы, во многих случаях с высокой долей вероятности дают информацию о принадлежности скаффолда к фрагменту генома без таксономической принадлежности этого вируса.

Для сравнения образцов по составу вирусов в случае раздельной сборки необходима дополнительная, сложная процедура группировки или кластеризации («binning») скаффолдов по сходству частот встречаемости нуклеотидов, частот встречаемости нуклеотидных *k*-mer и сходству генов, реализующих одинаковую функцию. Для прокариотических и эукариотических сообществ такая группировка (binning-процедура) облегчается наличием общих для всех клеточных организмов генов «домашнего хозяйства» («housekeeping genes») и сходством протяжённых участков генома у близкородственных видов; у вирусов же нет никаких общих для всех таксонов генов. Таксономическая идентификация вирусов также затруднена тем, что даже близкородственные виды (например, вирусы разных родов одного семейства)

значительно отличаются по нуклеотидным последовательностям и имеют лишь 30–35 % сходства при выравнивании белковых последовательностей (нижний предел чувствительности BLAST-анализа для поиска последовательностей подобных организмов) [35, 36]. Для сообществ прокариот разработан широкий спектр программных пакетов [37–39], которые применяются для binning-группировки скаффолдов из разных образцов; однако, для анализа вирусных последовательностей применение таких пакетов невозможно. Использованный и рекомендуемый в данной работе подход (микс-сборка) позволяет избежать применения сложной и неоднозначной процедуры сопоставления скаффолдов разных геномных сборок и непосредственно сравнивать вирусные скаффолды, в том числе фрагменты геномов вирусов, таксономия которых не установлена.

В данной работе произведено сравнение нескольких проб на основе результатов микс- и раздельных геномных сборок с использованием данных о выявленных виротипах – ближайших вирусным скаффолдам вирусов из базы данных NCBI RefSeq. Методами кластерного анализа устойчиво парной кластеризации одинаковых проб по данным разных типов сборок получено не было. Это связано с тем, что значительное количество скаффолдов, распознанных программой VirSorter2 (реализующей HMM-подход) как вирусные, не было идентифицировано по полной базе данных вирусных протеомов NCBI RefSeq до уровня виротипа (отнесены к неидентифицированным вирусам). В категории неидентифицированных вирусов могли присутствовать скаффолды разных видов, представленные в одних пробах и отсутствующие в других. Таким образом, результаты кластерного анализа были искажены. Использование исключительно данных микс-сборки позволяет избежать этих проблем и использовать для кластеризации непосредственно данные о показателях представленности скаффолдов совместно с данными таксономии (или без них).

Микс-сборка метагеномных и метатранскриптомных данных требует наличия большого количества вычислительных ресурсов. Исследователям необходим доступ к высокопроизводительным компьютерам с большим объемом оперативной памяти. В данном исследовании нами был задействован высокопроизводительный компьютер с двумя процессорами AMD EPYC 7513 32-Core с 1 Тб оперативной памяти. При запуске программы SPAdes максимальный объем используемой оперативной памяти в микс-сборке составил 380 Гб. В раздельных сборках каждый запуск задействовал от 70 до 90 Гб оперативной памяти. В данной работе анализ проводился для четырех проб метатранскриптомных данных. Вовлечение в анализ большего количества проб или исследование более сложноустроенных сообществ с большим спектром видов потребует наличие еще большего объема памяти высокопроизводительных компьютеров. Это может быть проблемой и ограничением в использовании данного подхода сборки для анализа метаданных – как для анализа вирусных сообществ, так и для исследования других групп организмов.

В области исследования сообществ РНК-содержащих вирусов методами метагеномики проводится большое количество работ. Изучается разнообразие вирусов в окружающей среде [40], а также вирусов, ассоциированных с различными организмами [41, 42], в том числе с человеком [43]. Большая часть таких исследований проведена с использованием раздельных сборок данных, где в конечном итоге рассматривались только те скаффолды и вирусные геномы, которые удалось идентифицировать до определённого виротипа (таксона). Информация о гипотетических вирусных геномах ускользает от внимания исследователей. В работе [44] авторы применили приложение CD-HIT-EST [45] для кластеризации и поиска сходства между скаффолдами раздельных геномных сборок при исследовании сообществ РНК-содержащих вирусов с помощью алгоритмов выравнивания нуклеотидных последовательностей. Подход с применением CD-HIT-EST в некоторой степени решает проблему сопоставления геномных сборок без таксономии, но может

столкнуть с проблемой, когда одна геномная сборка содержит скаффолд – фрагмент генома какого либо вируса, а другая сборка содержит скаффолд – другой фрагмент этого же вируса. В таком случае алгоритм выравнивания не сработает. Микс-сборка в этой ситуации из-за увеличения количества данных потенциально может привести к объединению таких разрозненных фрагментов в единый вирусный геном, присутствующий одновременно в разных пробах.

В нашей предыдущей работе мы использовали объединенную комбинированную геномную сборку для анализа вирусных сообществ РНК-содержащих вирусов байкальских губок [18] на основе метавиромных наборов данных (РНК выделяли из предварительно изолированного вирусного материала). Такой подход позволил сравнить сообщества РНК-содержащих вирусов, ассоциированных с различными особями губок, путем сопоставления скаффолдов, идентифицированных программой VirSorter2 (HMM алгоритмом) как вирусные. Микс-сборка позволила вовлечь в сравнительный анализ скаффолды, не идентифицированные до виротипа, находящиеся в доминирующем пуле вирусных последовательностей.

Предлагаемый в работе подход может быть использован и в других исследованиях вирусов на основе данных метагеномного и метатранскриптомного секвенирования ДНК и РНК-сообществ окружающей среды (из образцов воды, почвы и др.), или ассоциированных с различными организмами. Исследователи могут использовать схему конвейера анализа данных, предложенную в данной работе (рис. 1). Возможности проведения такого анализа будут возрастать с увеличением доступности высокопроизводительных компьютеров, суперкомпьютеров и кластеров с большим объемом памяти на одну вычислительную ячейку.

Нуклеотидные последовательности вирусных скаффолдов, аминокислотные последовательности белков в вирусных скаффолдах, результаты таксономического анализа вирусных скаффолдов и данные о представленности вирусных скаффолдов в образцах раздельных и микс-геномных сборок доступны по ссылке: https://disk.yandex.ru/d/WPmWDxJ70nqj5Q.

Работа выполнена при финансовой поддержке Российского научного фонда, проект № 22-24-01120.

СПИСОК ЛИТЕРАТУРЫ

1. Zhang L., Chen F.X., Zeng Z., Xu M., Sun F., Yang L., Bi X., Lin Y., Gao Y.J., Hao H.X. et al. Advances in Metagenomics and Its Application in Environmental Microorganisms. *Front. Microbiol.* 2021. V. 12. P. 1–15. doi: 10.3389/fmicb.2021.766364
2. Roux S., Matthijnssens J., Dutilh B.E. Metagenomics in Virology. In: *Encyclopedia of Virology*. Ed. Bamford D.H., Zuckerman M. Cambridge: Academic Press. 2020. P. 133–140. doi: 10.1016/B978-0-12-809633-8.20957-6
3. Sommers P., Chatterjee A., Varsani A., Trubl G. Integrating Viral Metagenomics into an Ecological Framework. *Annu. Rev. Virol*. 2021. V. 8. P. 133–158. doi: 10.1146/annurev-virology-010421-053015
4. Santiago-Rodriguez T.M., Hollister E.B. Potential Applications of Human Viral Metagenomics and Reference Materials: Considerations for Current and Future Viruses. *Appl. Environ. Microbiol.* 2020. V. 86. № 22. P. 1–12. doi: 10.1128/AEM.01794-20
5. Santiago-Rodriguez T.M., Hollister E.B. Unraveling the viral dark matter through viral metagenomics. *Front. Immunol.* 2022. V. 13. P. 1–13. doi: 10.3389/fimmu.2022.1005107
6. Leinonen R., Sugawara H., Shumway M. The sequence read archive. *Nucleic Acids Res.* 2011. V. 39. P. 2010–2012. doi: 10.1093/nar/gkq1019

7. Shi M., Lin X.D., Tian J.H., Chen L.J., Chen X., Li C.X., Qin X.C., Li J., Cao J.P., Eden J.S. et al. Redefining the invertebrate RNA virosphere. *Nature.* 2016. V. 540. № 7634. P. 539–543. doi: 10.1038/nature20167
8. Zhang Y.Y., Chen Y., Wei X., Cui J. Viromes in marine ecosystems reveal remarkable invertebrate RNA virus diversity. *Sci. China Life Sci.* 2022. V. 65. № 2. P. 426–437. doi: 10.1007/s11427-020-1936-2
9. Thomas T., Gilbert J., Meyer F. Metagenomics - a guide from sampling to data analysis. *Microb. Inform. Exp.* 2012. V. 2. № 1. P. 3. doi: 10.1186/2042-5783-2-3
10. Nooij S., Schmitz D., Vennema H., Kroneman A., Koopmans M.P.G. Overview of virus metagenomic classification methods and their biological applications. *Front. Microbiol.* 2018. V. 9. P. 749. doi: 10.3389/fmicb.2018.00749
11. Sutton T.D.S., Clooney A.G., Ryan F.J., Ross R.P., Hill C. Choice of assembly software has a critical impact on virome characterisation. *Microbiome.* 2019. V. 7. № 1. P. 1–15. doi: 10.1186/s40168-019-0626-5
12. Hiltemann S., Rasche H., Gladman S., Hotz H.R., Larivière D., Blankenberg D., Jagtap P.D., Wollmann T., Bretaudeau A., Goué N. et al. Galaxy Training: A powerful framework for teaching! PLoS Comput. Biol. 2023. V. 19. № 1. P. 1–18. doi: 10.1371/journal.pcbi.1010752
13. Skewes-Cox P., Sharpton T.J., Pollard K.S., DeRisi J.L. Profile hidden Markov models for the detection of viruses within metagenomic sequence data. *PLoS One.* 2014. V. 9. № 8. P. e105067. doi: 10.1371/journal.pone.0105067
14. Ren J., Song K., Deng C., Ahlgren N.A., Fuhrman J.A., Li Y., Xie X., Poplin R., Sun F. Identifying viruses from metagenomic data using deep learning. *Quant. Biol.* 2020. V. 8. № 1. P. 64–77. doi: 10.1007/s40484-019-0187-4
15. Reyes A.P., Alves J.M., Durham A.M., Gruber A. Use of profile hidden Markov models in viral discovery: current insights. *Adv. Genomics Genet.* 2017. V. 7. P. 29–45. doi: 10.2147/AGG.S136574
16. Butina T.V., Bukin Y.S., Petrushin I.S., Tupikin A.E., Kabilov M.R., Belikov S.I. Extended evaluation of viral diversity in Lake Baikal through metagenomics. *Microorganisms.* 2021. V. 9. № 4. P. 1–31. doi: 10.3390/microorganisms9040760
17. Butina T.V., Petrushin I.S., Khanaev I.V., Bukin Y.S. Metagenomic Assessment of DNA Viral Diversity in Freshwater Sponges, *Baikalospongia bacillifera*. *Microorganisms.* 2022. V. 10. № 2. P. 480. doi: 10.3390/microorganisms10020480
18. Butina T.V., Khanaev I.V., Petrushin I.S., Bondaryuk A.N., Maikova O.O., Bukin Y.S. The RNA Viruses in Samples of Endemic Lake Baikal Sponges. *Diversity.* 2023. V. 15. № 7. P. 1–20. doi: 10.3390/microorganisms10020480
19. Bolger A.M., Lohse M., Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics.* 2014. V. 30. № 15. P. 2114–2120. doi: 10.1093/bioinformatics/btu170
20. Nurk S., Meleshko D., Korobeynikov A., Pevzner P.A. MetaSPAdes: A new versatile metagenomic assembler. *Genome Res.* 2017. V. 27. № 5. P. 824–834. doi: 10.1101/gr.213959.116
21. Guo J., Bolduc B., Zayed A.A., Varsani A., Dominguez-Huerta G., Delmont T.O., Pratama A.A., Gazitúa M.C., Vik D., Sullivan M.B. et al. VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome.* 2021. V. 9. № 1. P. 1–13. doi: 10.1186/s40168-020-00990-y
22. Nayfach S., Camargo A.P., Schulz F., Eloe-Fadrosh E., Roux S., Kyrpides N.C. CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat. Biotechnol.* 2021. V. 39. № 5. P. 578–585. doi: 10.1038/s41587-020-00774-7
23. Buchfink B., Xie C., Huson D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* 2014. V. 12. № 1. P. 59–60. doi: 10.1038/nmeth.3176

24. Wheeler T.J., Clements J., Eddy S.R., Hubley R., Jones T.A., Jurka J., Smit A.F.A., Finn R.D. Dfam: A database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res.* 2013. V. 41. P. 70–82. doi: 10.1093/nar/gks1265
25. *Dfam release 3.7 (January 2023)*. URL: https://www.dfam.org/ (accessed 02.11.2023).
26. O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D. et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 2016. V. 44. P. D733–D745. doi: 10.1093/nar/gkv1189
27. Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012. V. 9. № 4. P. 357–359. doi: 10.1038/nmeth.1923
28. Danecek P., Bonfield J.K., Liddle J., Marshall J., Ohan V., Pollard M.O., Whitwham A., Keane T., McCarthy S.A., Davies R.M. Twelve years of SAMtools and BCFtools. *Gigascience.* 2021. V. 10. № 2. P. 1–4. doi: 10.1093/gigascience/giab008
29. *Oksanen J. Package 'vegan'*. URL: https://github.com/vegandevs/vegan (accessed 03.11.2023).
30. Li D., Liu C.M., Luo R., Sadakane K., Lam T.W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics.* 2015. V. 31. № 10. P. 1674–1676. doi: 10.1093/bioinformatics/btv033
31. Peng Y., Leung H.C.M., Yiu S.M., Chin F.Y.L. IDBA-UD: A *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* 2012. V. 28. № 11. P. 1420–1428. doi: 10.1093/bioinformatics/bts174
32. Yang C., Chowdhury D., Zhang Z., Cheung W.K., Lu A., Bian Z., Zhang L. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* 2021. V. 19. P. 6301–6314. doi: 10.1016/j.csbj.2021.11.028
33. Petrovskii S., Petrovskaya N. Computational ecology as an emerging science. *Interface Focus.* 2012. V. 2. № 2. P. 241–254. doi: 10.1098/rsfs.2011.0083
34. Kieft K., Zhou Z., Anantharaman K. VIBRANT: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome.* 2020. V. 8. P. 1–23. doi: 10.1186/s40168-020-00867-0
35. Moya A., Elena S.F., Bracho A., Miralles R., Barrio E. The evolution of RNA viruses: A population genetics view. *Proc. Natl. Acad. Sci. U.S.A.* 2000. V. 24. № 13. P. 6967–6973. doi: 10.1073/pnas.97.13.6967
36. Bondaryuk A.N., Kulakova N.V., Belykh O.I., Bukin Y.S. Dates and Rates of Tick-Borne Encephalitis Virus—The Slowest Changing Tick-Borne Flavivirus. *Int. J. Mol. Sci.* 2023. V. 24. № 3. P. 2921. doi: 10.3390/ijms24032921
37. Kang D.D., Froula J., Egan R., Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ.* 2015. V. 3. P. e1165. doi: 10.7717/peerj.1165
38. Wu Y.W., Simmons B.A., Singer S.W. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics.* 2016. V. 32. № 4. P. 605–607. doi: 10.1093/bioinformatics/btv638
39. Tamames J., Puente-Sánchez F. SqueezeMeta, a highly portable, fully automatic metagenomic analysis pipeline. *Front Microbiol.* 2019. V. 9. P. 3349. doi: 10.3389/fmicb.2018.03349
40. Rosario K., Breitbart M. Exploring the viral world through metagenomics. *Curr. Opin. Virol.* 2011. V. 1. № 4. P. 289–297. doi: 10.1016/j.coviro.2011.06.004
41. Gudenkauf B.M., Hewson I. Comparative metagenomics of viral assemblages inhabiting four phyla of marine invertebrates. *Front. Mar. Sci.* 2016. V. 3. P. 1–12. doi: 10.3389/fmars.2016.00023

42. Waldron F.M., Stone G.N., Obbard D.J. Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS Genet.* 2018. V. 14. № 7. P. e1007533. doi: 10.1371/journal.pgen.1007533
43. Bai G.H., Lin S.C., Hsu Y.H., Chen S.Y. The Human Virome: Viral Metagenomics, Relations with Human Diseases, and Therapeutic Applications. *Viruses.* 2022. V. 14. P. 278. doi: 10.3390/v14020278
44. Richard J.C., Blevins E., Dunn C.D., Leis E.M., Goldberg T.L. Viruses of Freshwater Mussels during Mass Mortality Events in Oregon and Washington, USA. *Viruses.* 2023. V. 15. № 8. P. 1–18. doi: 10.3390/v15081719
45. Li W., Godzik A. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics.* 2006. V. 22. № 13. P. 1658–1659. doi: 10.1093/bioinformatics/btl158

Рукопись поступила в редакцию 05.11.2023, переработанный вариант поступил 19.11.2023.
Дата опубликования 20.11.2023.

========== **BIOINFORMATICS** ==========

Performance Analysis of Cross-Assembly of Metatranscriptomic Datasets in Viral Community Studies

Bukin Yu.S., Bondaryuk A.N., Butina T.V.

Limnological Institute Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia

Abstract. We conducted a comparative analysis of individual and cross-assemblies of several metatranscriptomic data sets to study viral communities using several metatranscriptomes of endemic Baikal mollusks. We have shown that, compared to individual dataset assemblies, a hidden Markov model-based cross-assembly procedure increases the number of viral contigs (or scaffolds) per sample, the number of virotypes identified, and the average length of scaffolds per sample. The proportion of assembled viral reads from the total number of reads in samples is higher in cross-assembly. *De novo* cross-genomic assemblies combined with a virus identification algorithm using HMM present the data in a table with the number of reads from different samples for each scaffold. The table allows comparison of samples based on the representation of all viral scaffolds, including those not taxonomically identified, i.e. those that have no analogues in the NCBI RefSeq database. Thus, cross-genomic assemblies allow for comparative analyzes taking into account the latent diversity of viruses. We propose a pipeline for metatranscriptomic data analysis using *de novo* cross-genomic assembly to study viral diversity.

Key words: *metagenomics, transcriptomics, viruses, viral communities, metagenomic assembly, cross-assembly, metatranscriptomic analysis, viral scaffolds.*