ВЛИЯНИЕ СТЕПЕНИ ФИЛЬТРАЦИИ ДАННЫХ СЕКВЕНИРОВАНИЯ НА КАЧЕСТВО И ПОЛНОТУ DE NOVO СБОРКИ ТРАНСКРИПТОМА

Мегер Я.В.[1], Лантушенко А.О.[1], Водясова Е.А.[1, 2]

[1] ФГАОУ ВО «Севастопольский государственный университет»

ул. Университетская, 33, г. Севастополь, 299053, РФ; e-mail: meger_yakov@mail.ru

[2] ФИЦ «Институт биологии южных морей им. А.О. Ковалевского»

пр. Нахимова, 2, г. Севастополь, 299011, РФ

Поступила в редакцию: 01.08.2020

Аннотация. Для сборки de novo транскриптома существует множество сборщиков, которые имеют различающиеся алгоритмы. В тоже время этап фильтрации, являясь одним из ключевых, также имеет несколько подходов и алгоритмов. Однако, на сегодняшний день работ по изучению влияния степени фильтрации на сборку de novo транскриптома крайне мало. В данной работе были проанализированы транскриптомы, полученные с помощью двух наиболее распространенных программ (rnaSPADES и Trinity), а также применены различные подходы к этапу фильтрации прочтений. Были показаны ключевые различия для двух сборок и выявлены параметры, которые оказались чувствительными к степени фильтрации и длине входных прочтений. Также был предложен эффективный алгоритм фильтрации, который является двухэтапным и позволяет максимально сохранить объем входных данных при необходимом качестве всех прочтений после фильтрации и обрезки.

Ключевые слова: *rnaSPADES, Trinity, сборка de novo транскриптома, RNA-seq, фильтрация прочтений*

ВВЕДЕНИЕ

В настоящее время для более глубокого понимания физиологии процессов, протекающих в организмах при различных условиях, требуется изучение экспрессии генов или целых комплексов генов, как одного из аспектов реакции организма на раздражитель**.** С развитием и уменьшением стоимости технологий RNA-Seq увеличивается количество исследований, основанных на секвенировании полного транскриптома целого организма, отдельных тканей или клеток [1]. Параллельно развиваются и биоинформатические методы, направленные на улучшение сборок транскриптомов, анализа химмер и изоформ, аннотирование контигов и т.д. [2-4]. В случае работы с немодельным или недостаточно изученным объектом, возникает необходимость в сборке de novo без референсного генома. Поскольку большинство метрик оценки качества полученных сборок носят относительный (сравнительный) характер, оценка качества полученного транскриптома без качественного референса является сложной задачей. Эта ситуация осложняется тем, что выбор эффективного и оптимального алгоритма пред- и постобработки полученных прочтений до сих пор остается предметом дискуссий. Существует определенный программный конвейер для сборки de novo (рис. 1). Алгоритмы и их зависимость от входных данных разнятся у различных программ для сборки транскриптомов [5].

Первый этап de novo сборки транскриптомов обязательно включает в себя оценку качества полученных прочтений после секвенирования и их дальнейшую фильтрацию. Подходы на этом этапе отличаются: или стараются максимально очистить полученные транскриптомы от низкокачественных и коротких прочтений, или сохранить все отсеквенированные последовательности, или найти компромиссный вариант между достоверностью и полнотой. Данный этап является ключевым, так как качество собранного транскриптома напрямую зависит от входных данных (принцип “garbage in, garbage out”). Одним из факторов, который может

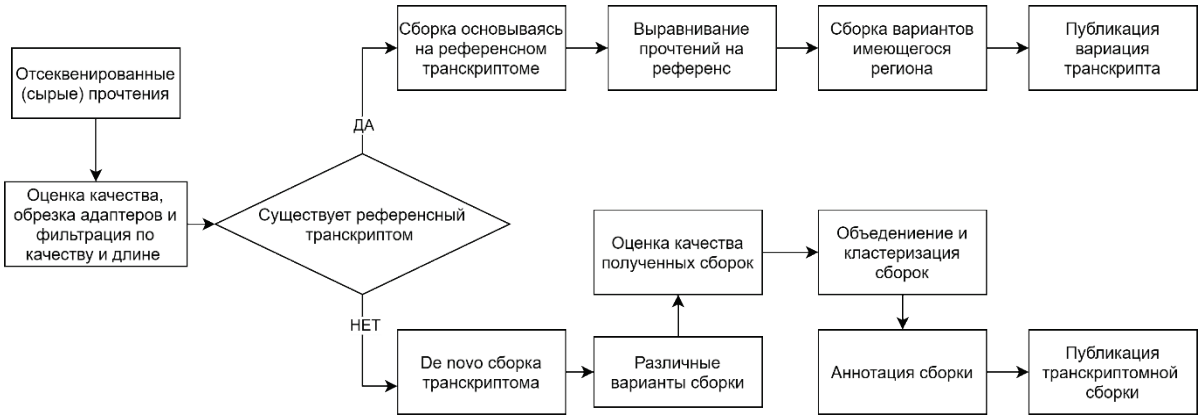

Рисунок 1. Общий алгоритм, применяемый при сборке транскриптома

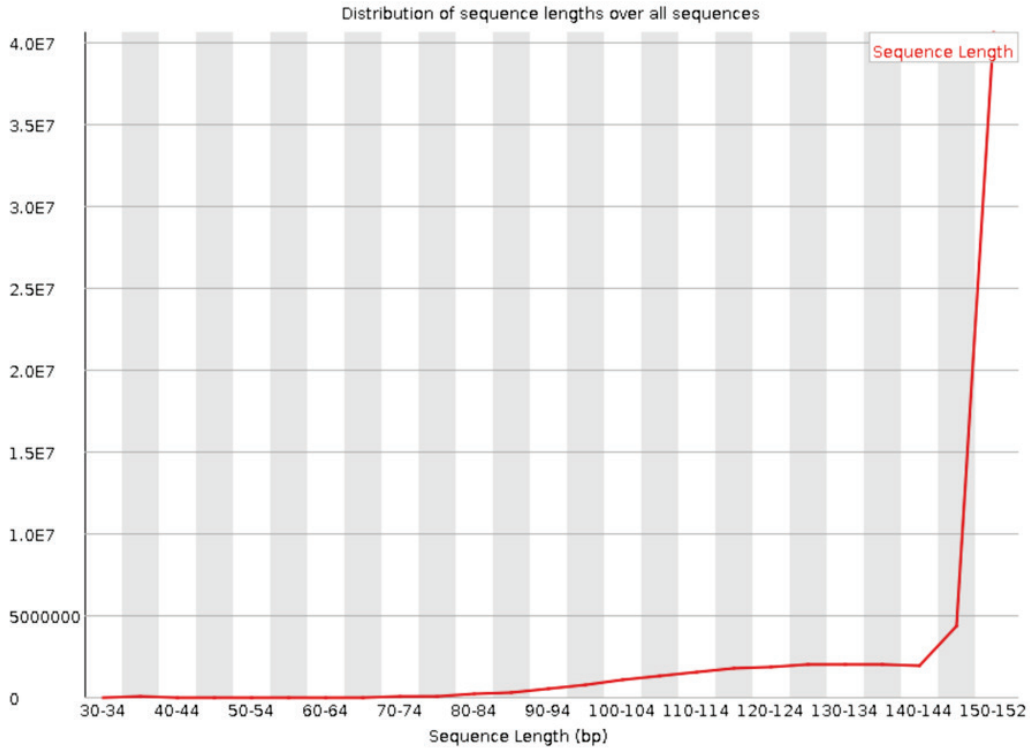

Рисунок 2. Распределение длин прочтений в неотфильтрованном виде

влиять на качество сборки de novo транскриптомов – это длина прочтений после фильтрации, так как длина прочтений является решающей для выбора длины k-меров[6]. Исследований в этом вопросе крайне мало [7].

Таким образом, целью данной работы было изучение влияние различной длины прочтений на качество и полноту сборок de novo транскриптомов с помощью различных сборщиков.

МАТЕРИАЛЫ И МЕТОДЫ

В данной работе сборка проводилась на одно-концевых прочтениях длиной 150 п.н., полученных с помощью NGS-секвенирования суммарной РНК из жаберных тканей *Mytilus galloprovincialis* на платформе Illumina Hi-Seq. Оценка качества прочтений проводилась с помощью программы fastqc [8]. У всех прочтений к концу падало качество и был высокий уровень дупликаций, что связано с наличием в пробах рРНК. В данном случае рРНК не убиралась из массива, и сборка проводилась для полного набора данных. Длина прочтений от 35 до 151, более 90% которых имеет длину более 149, в оставшихся подавляющее большинство в районе 90-140 п.н. (рис. 2).

Обработка прочтений проводилась при помощи программы fastp v.0.20.0 [9] различными методами:

1) Грубая – два этапа фильтрации, отбрасывающий прочтения короче 100 п.н. и обрезкой начала и окончания прочтения в местах падения качества.

2) Средняя – два этапа фильтрации, фильтрация отброшенных прочтений длиной менее 85 п.н., обрезка начала и окончания.

3) Мягкая - два этапа фильтрации, фильтрация отброшенных прочтений длиной менее 65 п.н., обрезка начала и окончания.

4) Минимальная – фильтрация в один этап, с отбрасыванием прочтений короче 50 п.н., обрезкой начала и окончания прочтений и прочтений с низким качеством (GC % <20)

Фильтрация, которая делалась в два этапа, осуществлялась по следующему алгоритму.

1 этап:

Обрезка начала и конца прочтений, фильтрация по качеству (phred >20)

> -q 20 -b 140 -f 10 -n 35 -y -l 80 --failed_out

Однако, таким образом очень сильно обрезалась большая часть прочтений. Поэтому потом отдельно анализировались прочтения, не прошедшие фильтрацию. Там были два кластера по длинам: в районе 70 и 140 нуклеотидов. Каждый кластер нес в себе разные ошибки, поэтому дальнейшая фильтрация была отдельно для каждого диапазона длин (2 этап):

> -Q -A -l 60 --length_limit 80 -b 70 -f 3 -t 2

> -Q -A -l 130 -3 -W 3 -M 30 -b 65 -t 5

После чего все три файла соединяли вместе и делали три варианта отфильтрованных прочтений по минимальной длине: 65, 85 и 100 нуклеотидов.

Было отсеквенировано 7 транскриптомов, все fastq файлы (в соответствии с минимальной длиной прочтений) были объединены в один и затем проводилась еще раз оценка качества с помощью fastqc.

Таблица 1. Объем входящих прочтений с разной степенью фильтрации

Длина прочтений	35-151 п.н. До фильтрации	50-140 п.н.	65-140 п.н.	85-140 п.н.	100-140 п.н.
Количество прочтений	63 592 364	60 894 123	61 727 260	59 690 644	56 364 226
Доля прочтений от исходного	100	95,8	97,1	93,9	88,6

De novo сборка проводилась тремя алгоритмами, реализованными в транскриптомных сборщиках Trinity v. 2.1.1. [10] и rnaSPADES v3.13.0 [11]. Качество сборок оценивалось программой QUAST v.4.6.3 [12], оценка процента картирования прочтений на сборку, выполнялась в программе Bowtie2 v. 2.3.4.3 [13], предсказание белок кодирующих областей проводилось на основе анализа открытых рамок считывания с помощью программы Transdecoder v. 5.5.0 входящий в пакет Trinity, поиск консервативных генов выполнялся на основе базы генов-ортологов mollusca_odb10 (https://busco.ezlab.org/list_of_lineages) в программе BUSCO v. 4.0.5 [14]. Была проведена кластеризация собранных контигов с использованием Usearch v.11.0.667 [15], процент идентичности выбран 0,95, центроиды выбирались по наибольшей длине.

РЕЗУЛЬТАТЫ

Суммарный объем прочтений составил 63,6 Gb. Длина прочтений до фильтрации составила от 35 до 151 п.н. (более 90% имело длину более 149 п.н.). При самой грубой чистке из массива данных исключается 11,4% прочтений (табл. 1).

Характеристика полученных сборок с помощью различных сборщиков достаточно сильно различается (табл. 2). Длина наибольшего контига в сборках с использованием Trinity составил 28175 п.н. для всех степеней фильтрации, с использованием rnaSPADES 17500±1000 п.н. Количество контигов в сборках с использованием Trinity в среднем составила ~250 тыс., с использование rnaSPADES – 170 тыс. контигов. GC состав всех сборок составил 34,8±0,2 %.

Помимо отличий между сборками на основе стандартных характеристик, были выявлены принципиальные отличия и по другим параметрам.

Процент картирования прочтений на сборки для всех вариантов составил 96±2 %, однако количество уникальных выравниваний различается: для сборок с использованием Trinity 16,6±1%, для сборок с использованием rnaSPADES 40±4%.

Поиск консервативных генов по базе mollusca_odb10 для сборок с использованием Trinity показал, что 4655±30 генов найдено из 5295 существующих в базе, с использованием rnaSPADES 4452±40 генов, что соизмеримо между собой (рисунок 4). Если рассматривать число уникальных генов, то опять наблюдается сильное отличие между сборщиками Trinity и rnaSPADES (~1850 и ~3440 соответственно).

Таблица 2. Характеристики полученных сборок QUAST

	Trinity v.2.1.1				
	35-151 п.н.	50-140 п.н.	65-140 п.н.	85-140 п.н.	100-140 п.н.
Количество контигов ≥ 0 п.н.	272 971	246 069	259 251	256 846	248 355
Количество контигов ≥ 500 п.н.	112 757	100 429	106 239	105 782	102 741
N50 (≥ 500 п.н.)	1 790	1 257	1 798	1 799	1 800
Длина наибольшего контига, п.н.	28 175	28 171	28 175	28 175	28 175
GC, %	34,69	34,89	34,74	34,75	34,72
	rnaSpades v.3.13.0				
Количество контигов ≥ 0 п.н.	174 049	173 267	173 747	173 517	172 288
Количество контигов ≥ 500 п.н.	75 371	66 762	69 790	69 697	69 138
N50 (≥ 500 п.н.)	1 442	1 436	1 439	1 438	1 438
Длина наибольшего контига, п.н.	17 202	17 193	18 672	16 815	16 815
GC, %	34,79	34,89	34,84	34,84	34,83

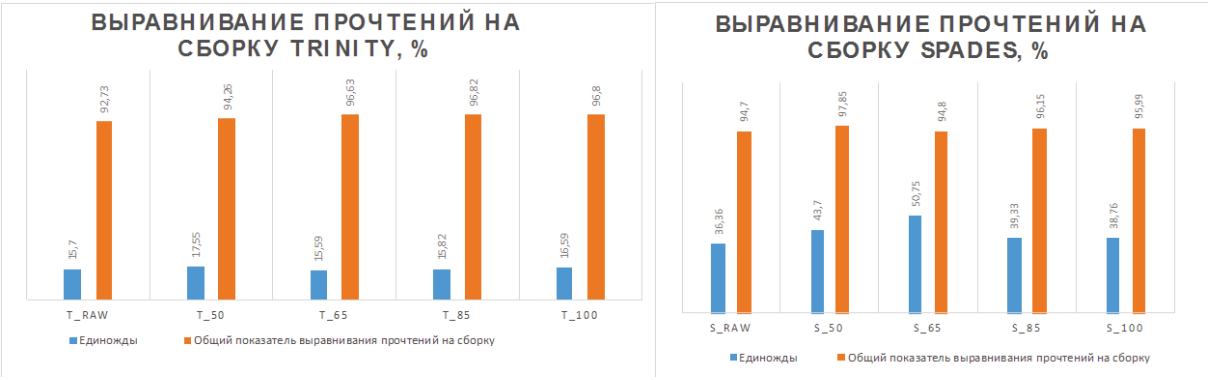

Рисунок 3. Выравнивание прочтений на полученную сборку, выполненное с помощью Bowtie2

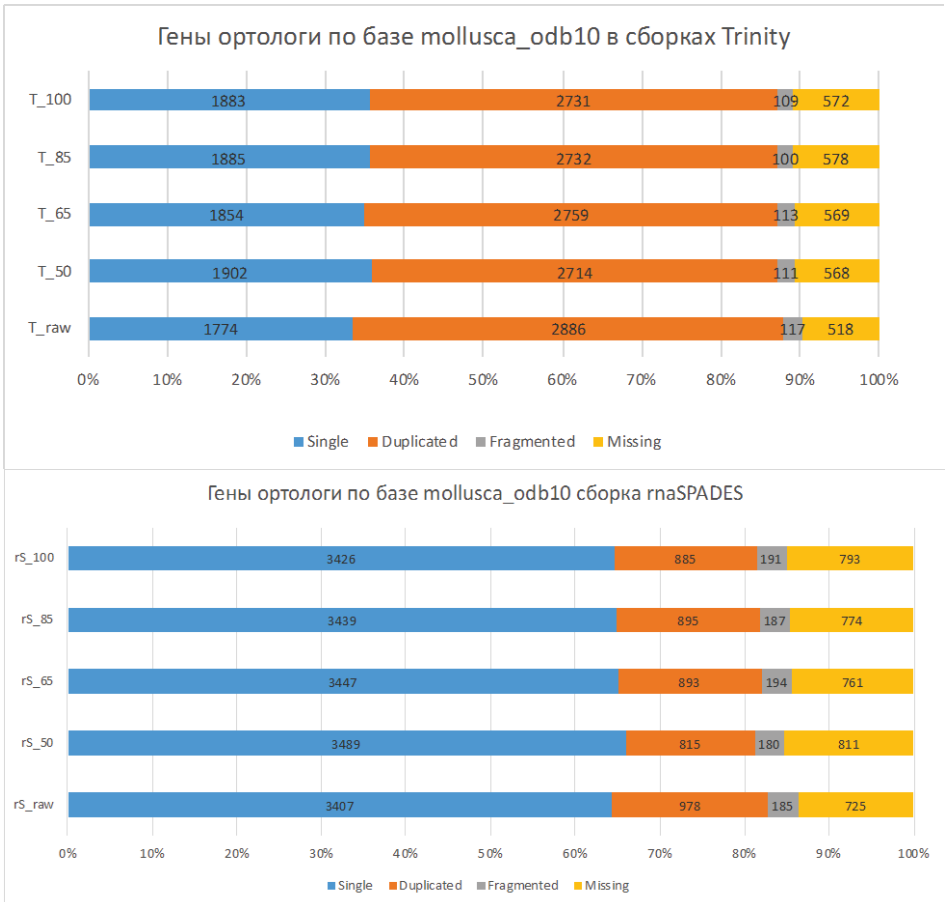

Рисунок 4. Результаты поиска генов-ортологов для всех вариантов сборок с помощью BUSCO

Поиск консервативных генов по базе mollusca_odb10 для сборок с использованием Trinity показал, что 4655±30 генов найдено из 5295 существующих в базе, с использованием rnaSPADES 4452±40 генов, что соизмеримо между собой (рисунок 4). Если рассматривать число уникальных генов, то опять наблюдается сильное отличие между сборщиками Trinity и rnaSPADES (~1850 и ~3440 соответственно).

Наибольшее отличие между сборщиками продемонстрировал анализ по поиску открытых рамок считывания (ОРС). Для сборок с использованием Trinity было найдено ~110 тыс. белок кодирующих транскриптов, с использованием rnaSPADES – 78 тыс (табл. 3). Были обнаружены контиги с 7 рамками считывания. По умолчанию минимальная длина ОРС была оставлена 100 а.к. (300 н.к.). По данным сборок Trinity число предсказанных белков гораздо больше, в тоже время распределение длин ОРС демонстрирует большую частоту коротких мРНК. Так число ОРС с длиной 300 н.к. у сборок Trinity в 2 раза больше, чем у сборок SPADES (рис. 5).

Таблица 3. Характеристики транскриптов с открытыми рамками считывания

	Trinity v.2.1.1				
	35-151 п.н.	50-140 п.н.	65-140 п.н.	85-140 п.н.	100-140 п.н.
Количество предсказанных белков	132 860	120 327	103 075	102 228	99 150
Частота в транскриптах, %	40,07	40,05	39,76	39,80	39,92
Кол-во контигов с ≥7 ОРС	25	24	0	0	0
Число контигов с 5'UTR и 3'UTR	38 079	34 362	32 273	32 784	31 860
	rnaSpades v. 3.13.0				
Количество предсказанных белков	86 710	78 481	81 460	81 449	80 636
Частота в транскриптах, %	56,02	51,53	46,84	46,94	46,80
Кол-во контигов с ≥7 ОРС	12	10	15	17	16
Число контигов с 5'UTR и 3'UTR	21 449	19 078	19 972	19 978	19 749

Число контигов после кластеризации для сборок с использованием Trinity составило 78,1±0,7% относительно начального количества, с использованием rnaSPADES 94,4±2,5%. После кластеризации, число оставленных контигов стало более одинаково по всем сборкам, диапазон составляет от 168330 до 197879 (в сборках диапазон от 172288 до 272971). Сильно различается кластеризация по сборщикам Trinity и SPADES. Распределение длин центроидов опять же смещено в сторону более коротких у Trinity, при этом максимальная длина также характерна для этого сборщика, что в результате обуславливает почти одинаковую среднюю длину. На рисунке 6 представлено распределение размеров кластеров (такое распределение размеров кластеров наблюдается для всех вариантов фильтрации). При сборке Trinity появляются очень большие кластеры размером до 140 транскриптов, однако число таких кластеров небольшое.

ОБСУЖДЕНИЕ

Отличие между различными сборщиками.

Проведенный анализ выявил определенную закономерность в сборках Trinity и rnaSPADES. Самый длинный контиг в сборке Trinity в 1.6 раза длиннее, чем при сборке SPADES. Такое же соотношение наблюдается для следующих параметров: число контигов, L50, количество предсказанных белков с помощью TransDecoder.

Число контигов, полученных в результате сборок Trinity и rnaSPADES, существенно отличается (на 35%). Тем не менее этот показатель следует рассматривать критично, так как такое число контигов может обуславливаться наличием большого числа химерных контигов и несуществующих изоформ, возникающих в результате сборки, а не свидетельствовать о качестве. Вероятно, сборка Trinity является более полной, но в тоже время более «грязной».

Это предположение подтверждается малым процентом выравнивания прочтений на сборку и огромным числом коротких контигов. Более того, процент транскриптов с открытыми рамками считывания также существенно отличается. Это свидетельствует, что большее число контигов, полученных при сборке Trinity, или слишком короткие (менее 300 п.н.), или являются химерами в результате чего для них не обнаружены ОРС длиной более 100 аминокислот.

Кроме того, большее число предсказанных белков (для сборки Trinity) имеет небольшую длину (рис. 5), но среднее значение длин практически одинаковы, что объясняется наличием более длинных контигов в сборках Trinity по сравнению со сборками SPADES (28175 п.н. и 17000 п.н. соответственно).

Тем не менее, с учетом анализа BUSCO и TransDecoder, сборку Trinity нельзя отбрасывать, так как это может повлечь за собой потерю информации о некоторых генах.

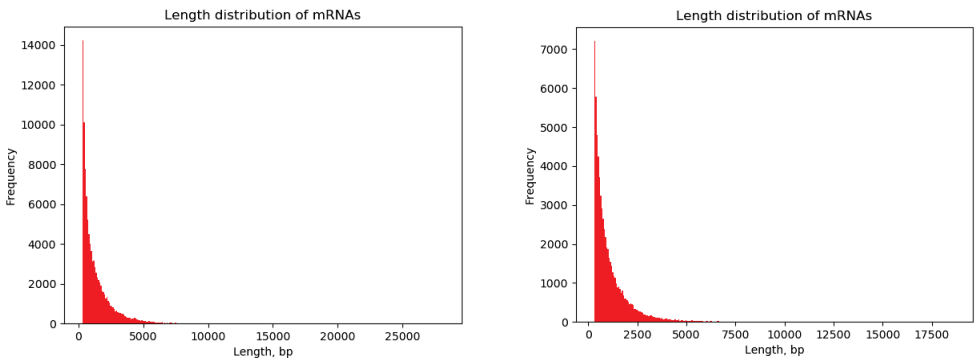

Рисунок 5. Распределение длин мРНК по результатам TransDecoder (Trinity справа, SPADES слева). Графики приведены для сборок с минимальной длиной прочтений в 100 п.н.

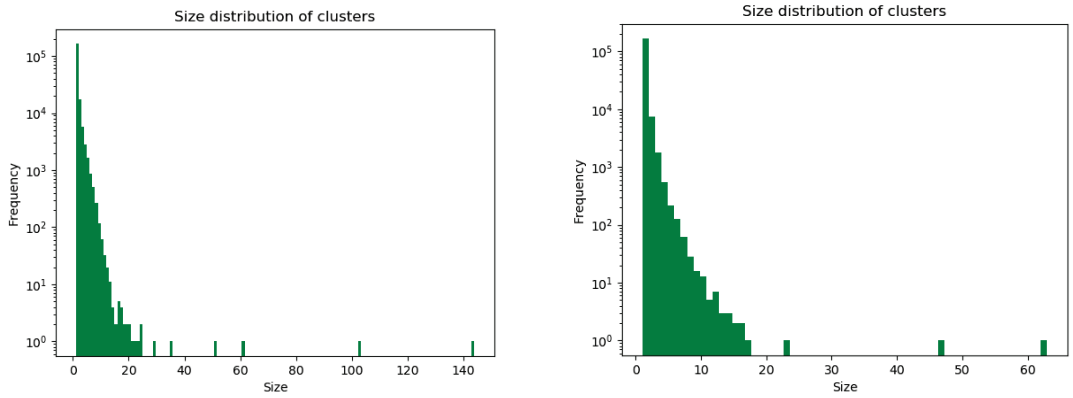

Рисунок 6. Распределение размеров кластеров для Trinity (слева) и SPADES (справа). Минимальная длина прочтений в fastq файле 100 п.н.. Шкала частоты является логарифмической

Отличие в степени фильтрации входных прочтений.

Анализ влияния фильтрации выявил некоторые особенности полученных сборок. Для обоих сборщиков было установлено, что фильтрация в два этапа эффективнее. После первого исключения из массива данных неудовлетворяющих по качеству прочтений и грубой обрезки по концам, повторно анализируя «плохие» прочтения мы часть из них возвращаем в массив данных для дальнейшей сборки. Если мы сравним число оставленных прочтений, то оказывается, что при более узком диапазоне длин (от 65 до 140 п.н.), число прочтений на 3 % больше, чем для длин от 50 до 140 п.н. (табл. 1). Кроме того, для сборки Trinity возрастает при двух этапной фильтрации число собранных контигов, и что самое главное для обоих сборщиков увеличивается число контигов с большей длиной (табл. 2). Последнее особенно заметно при сборке rnaSPADES. Вероятно, это связано, с возвращением на втором этапе фильтрации в массив прочтений, которые являются крайними участками транскриптов и поэтому могут быть меньшей длины, чем основная часть прочтений. При этом они могут нести важную информацию и при поиске ОРС оказаться белок-кодирующими последовательностями. Это предположение подтверждается результатами TransDecoder для rnaSPADES по увеличению числа предсказанных белков, при этом относительное число контигов с ОРС уменьшился (табл. 3). Следует отметить, что количество предсказанных белков у сборки Trinity падает с уменьшением длины. Это отличие между сборщиками является отражением реализованных в них различных алгоритмах и возможным наличием большего числа химер и несуществующих изоформ в сборке Trinity, о чем говорилось выше.

Оценка выравнивания прочтений на сборку выявила зависимость от фильтрации только для сборки rnaSPADES (рис. 3). Наибольший процент уникальных выравниваний выявлен опять же для двухэтапной чистки прочтений, но с максимально сохраненной длиной (диапазон 60-140 п.н.).

Поиск генов-ортологов не показал явной зависимости от фильтрации для обоих сборщиков.

ЗАКЛЮЧЕНИЕ

Таким образом, с одной стороны, de novo сборка транскриптомов, проведенная с использованием rnaSPADES больше подвержена влиянию степени фильтрации и отличается меньшей степенью полноты (но возможно большей чистотой сборок), по сравнению с de novo сборкой Trinity. С другой стороны, сборки, полученные с помощью Trinity, имеют большое число химмерных контигов и изоформ, что может приводить к трудностям на этапе оценки дифференциальной экспрессии транскриптов и некорректному анализу. Проведенный анализ показал, что для получения референсного транскриптома мы рекомендуем проводить объединение двух сборок (минимальная фильтрация в два этапа, сборщик rnaSPADES и максимальная фильтрация, сборщик Trinity), так как использование только одной программы для de novo сборки транскриптомов может приводить к искаженным или неполным данным. Кроме того, фильтрацию прочтений следует проводить дважды, максимально стараясь отсечь действительно низкокачественные прочтения, но максимально сохранив весь набор данных.

Работа выполнена в рамках государственной бюджетной темы (№ 0828-2018-0003), при поддержке Министерства образования и науки РФ (грант № 14.W03.31.0015) и внутреннего гранта СевГУ 2020 № 33/06-31.

Список литературы/ References:

1. Marinov G.K. On the design and prospects of direct RNA sequencing. *Briefings in functional genomics*, 2017, vol. 16, pp. 326-335.

2. Liu L., Song B., Ma J., Song Y., Zhang S.Y., Tang Y., Wu X., Wei Z., Chen K., Su J., Rong R., Lu Z., de Magalhães J.P., Rigden D.J., Zhang L., Zhang S.W., Huang Y., Lei X., Liu H., Meng J. Bioinformatics approaches for

deciphering the epitranscriptome: Recent progress and emerging topics. *Computational and structural biotechnology journal*, 2020, vol. 18, pp. 1587-1604.

3. Fu M., Su H., Su Z., Yin Z., Jin J., Wang L., Zhang Q., Xu X. Transcriptome analysis of Corynebacterium pseudotuberculosis-infected spleen of dairy goats. *Microbial pathogenesis*, 2020, vol. 34, pp. 104-120.

4. Seweryn M.T., Pietrzak M., Ma Q. Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics. *Computational and structural biotechnology journal*, 2020, vol. 18, pp. 1830-1837.

5. Tamames J., Cobo-Simón M., Puente-Sánchez F. Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes. *BMC genomics*, 2019, vol. 20, pp. 960.

6. Hölzer M., Manja M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience*, 2019, vol. 8, pp. 247-260.

7. Longone P. Percolation of aligned rigid rods on two-dimensional triangular lattices. *Physical review. E*, 2019, vol. 100, pp. 52-64.

8. Andrews S. FastQC: *A Quality Control Tool for High Throughput Sequence Data* [Online], 2010. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

9. Chen S. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, 2018, vol. 34, pp. 884-890.

10. Grabherr M.G., Haas B.J., Yassour M., Levin J.Z., Thompson D.A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B.W., Nusbaum C., Lindblad-Toh K., Friedman N., Regev A. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol*, 2011, vol. 29, pp. 644-702.

11. Bushmanova E., Antipov D., Lapidus A., Prjibelski A.D. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 2019, vol. 8, pp.103-147.

12. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 2013, vol. 29(8), pp. 1072-1075.

13. Langmead B., Wilks C., Antonescu V., Charles R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 2019, vol. 35, pp. 421-432.

14. Seppey M., Manni M., Zdobnov E.M. ***B***USCO: Assessing Genome Assembly and Annotation Completeness**.** *Methods in Molecular Biology*, 2019, vol. 6, pp.19-62.

15. Edgar R.C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 2010, vol. 26, pp. 2460-2461.

INFLUENCE OF THE DEGREE OF SEQUENCING DATA FILTERING ON THE QUALITY AND COMPLETENESS OF THE DE NOVO TRANSCRIPTOME ASSEMBLY

Meger Y.V.[1], Lantushenko A.O.[1], Vodiasova E.A.[2]

[1] Sevastopol State University

Universitetskaya str., 33, Sevastopol, 299053, Russia; e-mail: meger_yakov@mail.ru

[2] A.O. Kovalevsky Institute of Biology of the Southern Seas of RAS 1

Nachimov av., 2, Sevastopol, 299011, Russia

Abstract. There are many assemblers that have different algorithms to assemble a de novo transcriptome. At the same time, the filtering stage, being one of the key stages, also has several approaches and algorithms. However, to date, there is very little work on the influence of filtration degree on the de novo transcriptome Assembly. In this paper, we analyzed transcripts obtained using two of the most common programs (rnaSPADES and Trinity), and applied various approaches to the stage of filtering readings. Key differences were shown for the two assemblies and parameters were identified that were sensitive to the degree of filtering and the length of input reads. We also proposed an effective filtering algorithm that is two-stage and allows you to save the maximum amount of input data with the necessary quality of all readings after filtering and cropping.

Key words*: RNA-seq, rnaSPADES, Trinity, de novo transcriptome assembly, read filtering.*