УДК 004.94:57.011

МЕТОДОЛОГИЧЕСКИЕ ОСОБЕННОСТИ КАРТИРОВАНИЯ РИДОВ И СБОРКИ ТРАНСКРИПТОМА ДЛЯ ТКАНЕЙ НЕРВНОЙ СИСТЕМЫ *RATTUS NORVEGICUS*

© 2014 г. **П.Н. Меньшанов[1, 2], Н.Н. Дыгало[1, 2]**

[1] Федеральное государственное бюджетное учреждение науки
Институт цитологии и генетики Сибирского отделения
Российской академии наук, Новосибирск, Россия,
e-mail: menshanovpn@icg.sbras.ru;
[2] Федеральное государственное автономное образовательное учреждение высшего образования
«Новосибирский национальный исследовательский государственный университет»,
Новосибирск, Россия

Поступила в редакцию 15 сентября 2014 г. Принята к публикации 31 октября 2014 г.

Идентификация и количественная оценка уровней экспрессии всех вариантов транскриптов в исследуемых биологических образцах являются главной задачей при анализе транскриптомов тканей. Особое значение данная задача приобретает при анализе транскриптомов тканей нервной системы, для которых характерны процессы регулируемого альтернативного сплайсинга. В данной статье рассматривается ряд проблем, ассоциированных с локализацией последовательностей транскриптома на референсном геноме и последующей сборкой транскриптомов для образцов, полученных из ткани головного мозга. Внимание уделено вопросу однозначности локализации ридов; вопросу о размере ридов, а также проблеме неполной аннотированности геномов как у всех немодельных видов, так и у модельных организмов. В работе показано, что при одинаковых условиях секвенирования, геномной локализации ридов и последующей их сборки удельное количество межэкзонных ридов в транскриптомах, полученных из образцов ткани головного мозга, достоверно меньше по сравнению с удельным количеством таких ридов в транскриптомах других тканей. Подобная недопредставленность межэкзонных последовательностей в мРНК из тканей мозга может свидетельствовать об экспрессии значительного количества транскриптов, несущих новые неаннотированные сайты сплайсинга. Проведенный анализ стоимости–эффективности свидетельствует о необходимости использования при изучении транскриптома мозга технологий секвенирования, дающих длины ридов не менее 75 пар нуклеотидов. Для геномной локализации ридов и последующей сборки транскриптомов тканей ЦНС целесообразно использовать аннотации, включающие не только уже известные мРНК, но и последовательности транскриптов, предсказанные методом *ab initio* – например, аннотации консорциума Ensemble.

Ключевые слова: NG-секвенирование, транскриптом, геномное картирование последовательностей, сборка последовательностей, аннотация генома.

ВВЕДЕНИЕ

Одной из основных задач, решаемых в ходе изучения транскриптома методами NGS-секвенирования, являются идентификация и количественная оценка уровней всех вариантов транскриптов, экспрессируемых в исследуемом биологическом образце (Martin, Wang, 2011). Особое значение данная задача приобретает при анализе транскриптомов тканей нервной системы, в которой активно протекают процессы регулируемого альтернативного сплайсинга (Ule *et al*., 2006). Альтернативный сплайсинг предопределяет количественные и качественные особенности транскриптома в ходе индивидуального развития головного мозга (Lin *et al*., 2010; Irimia *et al*., 2011; Tollervey *et al*., 2011). Преобладание тех или иных вариантов

транскриптов в ткани мозга также может стать отправной точкой развития ряда заболеваний ЦНС, в том числе нейродегенеративных и онкологических (Tollervey *et al*., 2011; Mills, Janitz, 2012; Oltean, Bates, 2014).

Современная биоинформатика имеет в своем арсенале целый набор методов для локализации последовательностей и сборки транскриптома, позволяющих в разумные сроки провести идентификацию и количественную оценку уровней мРНК, транскрибируемых в ткани (Treangen, Salzberg, 2011). Некоторые из этих методов, используемые в программах Tophat/Tophat2, STAR, MapSplice, RUM и Novoalign, были специально оптимизированы для детекции сплайс-вариантов и в настоящее время активно используются научным сообществом (Martin, Wang, 2011; Treangen, Salzberg, 2011; Salzberg *et al*., 2012). В данной работе рассмотрен ряд проблем, с которыми можно столкнуться при картировании ридов и сборке транскриптома из образцов нервной ткани.

ПРОБЛЕМА НЕПОЛНОЙ АННОТАЦИИ ГЕНОМОВ

Анализ статистики экзон-интронной организации генов у позвоночных свидетельствует о высокой консервативности ключевых параметров структурной организации генов среди животных этой группы (табл. 1) (Entrez Genome Database, 2014). Так, медианный размер экзона у позвоночных практически всегда находится в диапазоне 126–146 п.н, а среднее число экзонов на ген колеблется от 9 до 12 (табл. 1). Вместе с тем следует отметить, что у модельных видов позвоночных средняя длина экзонов составляет 340–360 п.н., тогда как для немодельных видов она чаще всего лежит в диапазоне 230–280 п.н. (табл. 1).

Подобное рассогласование является прямым следствием неполной аннотированности геномов у всех немодельных организмов, что может существенно осложнить процедуру локализации ридов транскриптома на референсном геноме и последующей сборки транскриптомов для таких видов.

Исходным материалом для сборки транскриптома является массив ридов – последовательностей, полученных в ходе секвенирования тотальной РНК (Martin, Wang, 2011). После отбраковки последовательностей низкого качества для ридов высокого качества устанавливается их геномная локализация и последовательности с установленной локализацией могут быть использованы для сборки транскриптома. Очевидно, что все последовательности с установленной локализацией могут быть подразделены на два класса: (1) риды с локализацией внутри экзона и (2) риды с межэкзонной локализацией.

Опираясь на статистику экзон-интронной организации, можно оценить ожидаемое удельное количество внутриэкзонных и межэкзонных ридов. Поскольку длина гена не оказывает существенного влияния на интенсивность транскрипции (Grishkevich, Yanai, 2014), то при заданных параметрах распределения длин экзонов на долю внутриэкзонных ридов при длине ридов 100 п.н. должно приходиться около 70–75 % от общей длины транскриптома, тогда как остальные 25–30 % должны составлять межэкзонные риды. Вместе с тем при одинаковых условиях секвенирования, геномной локализации ридов и последующей их сборки удельное количество межэкзонных ридов в транскриптомах, полученных из образцов ткани головного мозга, достоверно меньше по сравнению с удельным количеством таких ридов в транскриптомах других тканей у крысы ($F_{(10, 309)} = 268{,}4$, $p < 0{,}00001$ – SRA архивы взяты из публикации Yu с соавт. (2014), данные по удельной доле выявленных межэкзонных ридов взяты из аннотации Entrez Genome Database, 2014 – Rat 105) и мыши (рис. 1) ($F_{(1, 11)} = 24{,}12$, $p < 0{,}00046$; использованы SRA архивы SRS362207–SRS362218, данные по удельной доле выявленных межэкзонных ридов взяты из аннотации Entrez Genome Database, 2014 – Mouse 104). Подобная недопредставленность межэкзонных последовательностей в транскриптах из тканей мозга может свидетельствовать либо о преимущественной экспрессии транскриптов с более протяженными экзонами, что противоречит уже имеющимся фактам, либо об экспрессии значительного количества транскриптов, несущих новые неаннотированные сайты сплайсинга. Таким образом, проблема недоаннотированности геномов может существенно осложнить локализацию ридов транскриптома на референсном гено-

Таблица 1

Размеры экзонов и интронов у позвоночных

Биологический вид	Аннотация	Экзоны (длина п.н.)		Интроны (длина п.н.)		Число экзонов на транскрипт	
		Медианная длина	Средняя длина	Медианная длина	Средняя длина	Медиана	Среднее число
Homo sapiens GRCh38	106	141	345	1675	6957	8	11,4
Chlorocebus sabaeus	100	145	359	1765	7523	8	11,8
Tarsius syrichta	100	131	225	1470	4678	8	10,0
Mus musculus GRCm38.p2	104	146	362	1456	5955	8	10,9
Rattus norvegicus RNor6	105	144	352	1391	4990	7	10,1
Peromyscus maniculatus	100	132	236	1390	4766	8	10,8
Cricetulus griseus	101	131	239	1334	4324	7	9,9
Bos taurus Btau_4.6.1	103	135	268	1323	5331	5	8,6
Bubalus bubalis	100	135	291	1365	5452	8	10,7
Camelus ferus	100	128	194	1221	4005	9	11,0
Vicugna pacos	100	129	198	1235	4375	9	11,0
Physeter catodon	100	131	251	1335	4323	8	11,0
Lipotes vexillifer	100	130	209	1336	5140	9	10,9
Balaenoptera acutorostrata	100	131	259	1373	5625	9	11,9
Eptesicus fuscus	100	128	191	1265	4579	8	10,9
Equus caballus	101	133	275	1353	5015	8	10,3
Felis catus	101	131	256	1369	4931	8	10,4
Panthera tigris altaica	100	128	236	1424	4918	8	10,3
Oryctolagus cuniculus	101	135	278	1574	6391	8	10,5
Orycteropus afer afer	100	129	196	1550	7909	8	10,6
Pteropus alecto	100	131	227	1198	4120	9	11,0
Ursus maritimus	100	132	255	1360	4999	7	9,2
Tupaia chinensis	100	130	229	1665	6003	7	10
Ornithorhynchus anatinus	102	129	223	1351	4073	5	8,1
Leptonychotes weddellii	100	130	211	1230	4897	7	9,1
Erinaceus europaeus	100	127	183	1568	5432	8	10,6
Alligator mississippiensis	100	129	204	1334	4442	7	9,9
Anolis carolinensis	101	136	287	1609	5128	8	11,1
Chelonia mydas	100	127	183	1548	5020	8	10,6
Python bivittatus	100	131	250	1298	3004	8	10,2
Calypte anna	100	126	177	901	2825	8	11,0
Astyanax mexicanus	100	135	278	907	2957	8	10,33
Callorhinchus milii	100	131	242	965	3012	8	11,15
Cynoglossus semilaevis	100	133	223	268	1140	9	11,65
Lepisosteus oculatus	100	126	167	792	2453	8	10,73
Neolamprologus brichardi	100	134	264	377	1805	8	11,37
Poecilia formosa	100	135	258	443	1756	9	11,34

ме и последующую сборку транскриптомов даже для образцов, полученных из модельных организмов.

ПРОБЛЕМА ОДНОЗНАЧНОЙ ЛОКАЛИЗАЦИИ МЕЖЭКЗОННЫХ РИДОВ

Все современные технологии NGS-секвенирования (Illumina, SOLID, PacBio, Ion Proton) дают риды с длиной не менее 50 п.н. (Buermans, den Dunnen, 2014), что позволяет достаточно уверенно определить происхождение большинства внутриэкзонных последовательностей, даже если они приходятся на высокогомологичные участки генома (Treangen, Salzberg, 2011). Вместе с тем для успешной идентификации ридов с межэкзонной локализацией эти риды должны содержать не только стык сайта сплайсинга, но и прилегающие к данному стыку последовательности, протяженности которых должно хватать для однозначной локализации рида в транскриптоме. Однако какой должна быть длина прилегающего к стыку сайта сплайсинга участка для однозначной локализации последовательности?

В литературе до сих пор не существует какой-либо внятной, теоретически обоснованной позиции по вопросу выбора длины прилегающего к стыку сайта сплайсинга участка, а рекомендуемые значения данного параметра были подобраны эмпирически и лежат в диапазоне от 8 до 11 п.н. (van Bakel *et al*., 2010; Rogers *et al*., 2012; Parada *et al*., 2014). Для проверки пригодности предлагаемых в литературе значений для идентификации стыков сайтов сплайсинга в данной работе была проведена локализация непарных 100-нуклеотидных ридов, полученных из 6 отдельных образцов полиА-обогащенной тотальной РНК гиппокампа 3-дневных крысят линии Вистар (виварий ИЦиГ СО РАН) путем определения последовательностей нуклеотидов на секвенаторе Illumina 2000, относительно референсного генома (сборка rn5) и транскриптома (сборка Ensemble 76) (Hubbard *et al*., 2002) *Rattus norvegicus*. Локализацию ридов проводили с использованием программ пакета Tuxedo: Tophat 2.0.9 и Bowtie 2.1.0.0.

В ходе анализа было показано, что изменение минимальной длины прилегающего к стыку сайта сплайсинга участка, используемого для локализации ридов, практически не влияет на число последовательностей, которые могут быть идентифицированы (рис. 2, а). В то же время увеличение длины прилегающего участка приводило к приросту числа ридов с однозначной локализацией, при этом при значениях показателя 14–15 п.н. число однозначно локализованных последовательностей стабилизировалось (рис. 2, б) $F_{(9, 45)} = 427{,}3$, $p < 0{,}00001$).

Вместе с тем само по себе увеличение числа ридов с однозначной локализацией не является достаточным маркером качества геномной локализации таких ридов и полученных на основании такой локализации сборок транскриптома. По этой причине значения представленности транскриптов в исследованных образцах были сопоставлены в сборках транскриптома, полученных при задании различных длин участка, прилегающего к стыку сайта сплайсинга. Сборку проводили с использованием программ пакета Tuxedo: Cufflinks, Cuffmerge, Cuffcompare и CuffDiff версии 2.1.1. Было показано, что существенных достоверных различий по представленности транскриптов в сборках, полученных при длинах прилегающего к стыку сайта сплайсинга участка в 8, 16 и 24 п.н., выявлено не было, что свидетельствует об однозначной локализации большинства ридов при подобных значениях данного параметра. Таким образом,

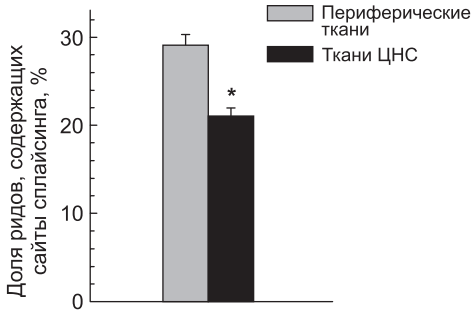

Рис. 1. Доля ридов, содержащих стыки сайта сплайсинга, в общем массиве ридов, полученных из образцов тканей ЦНС и периферических тканей мыши.

***** $p < 0{,}05$ по сравнению с другими периферическими тканями.

Fig. 1. The fraction of reads that contain splice sites in the total number of reads derived from the CNS and non-CNS tissue samples in mice.

***** $p < 0{,}05$ vs other non-CNS tissues.

при картировании непарных ридов транскриптома, полученных из образцов тканей ЦНС, целесообразно использовать значения 14–16 п.н.

ПРОБЛЕМА ВЫБОРА ДЛИНЫ И ТИПА РИДОВ

Поскольку однозначная локализация межэкзонных ридов возможна лишь при наличии прилегающей к стыку сайта сплайсинга последовательности определенной длины, не все межэкзонные риды, содержащие стык сайта сплайсинга, могут быть использованы для идентификации сплайс-вариантов. Вместе с тем вероятность попадания стыка сайта сплайсинга в зоны рида, не позволяющие однозначно локализовать последовательность, уменьшается по мере увеличения протяженности рида (рис. 3). Если сопоставить удельную цену секвенирования при различных длинах ридов на платформах Illumina и SOLID, при которой будет выявлено одинаковое количество стыков (табл. 2), то становится очевидно, что для повышения качества распознавания сплайсированных транскриптов необходимо применять технологии секвенирования, дающие более длинные последовательности, желательно не менее 75 п.н.

Альтернативным вариантом решения проблемы качества распознавания сплайсированных транскриптов является использование технологии парных ридов, которая позволяет верифицировать локализацию первого рида путем проверки местоположения соседней парной последовательности. Парные риды, использованные в данной работе, также были локализованы с помощью программ пакета Tuxedo – Tophat 2.0.9 и Bowtie 2.1.0.0. Процедура верификации подтвердила корректность местоположения для 90 % непарных ридов с ранее идентифицированной позицией. В то же время число ридов с множественной локализацией сократилось на 30 %, что свидетельствует о преимуществе использования данной технологии при секвенировании транскриптомов.

Как и в случае с непарными ридами, изменение минимальной длины прилегающего к стыку сайта сплайсинга участка, используемого для локализации последовательностей, практически не влияло на число пар, которые были идентифицированы (рис. 4, а). В то же время по сравнению с ранее проведенной локализацией непарных ридов число однозначно локализованных пар ожидаемо стабилизировалось при меньших значениях длины прилегающего к

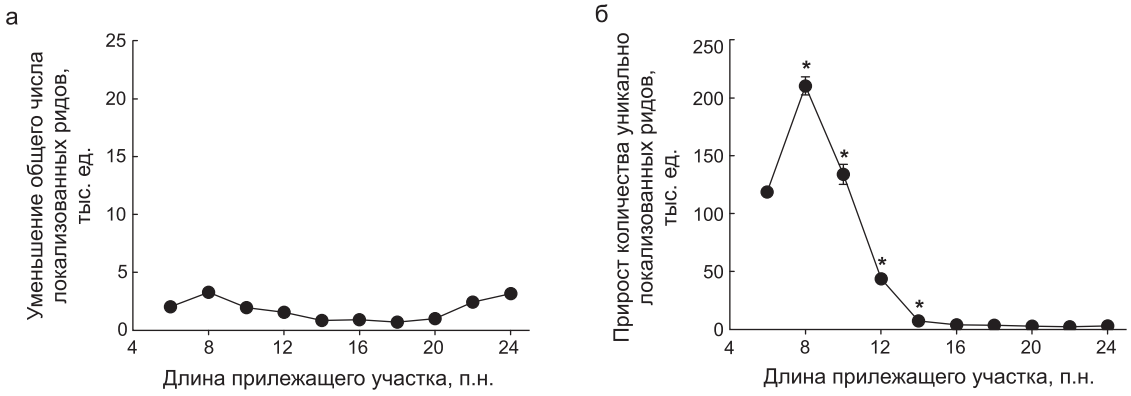

Рис. 2. Зависимость общего числа локализованных непарных ридов (а) и числа уникально локализованных непарных ридов (б) от длины прилежащего к стыку сайта сплайсинга участка, который был использован для локализации стыков сайтов сплайсинга.

* $p < 0{,}05$ по сравнению с приростом количества ридов, локализованных при меньшем значении длины прилежащего участка.

Fig. 2. Variation in the number of mapped single reads with the fixed minimal length of anchors on each side of the junction used to detect the splice site: (a) total number of mapped single reads, (b) uniquely mapped single reads.

* $p < 0{,}05$ vs. the increase in the number of mapped single reads, mapped with shorter anchors on each side of the junction used to detect the splice site.

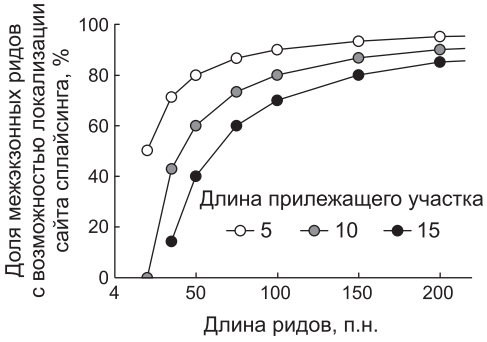

Рис. 3. Максимально возможная доля межэкзонных ридов в общей массе ридов определенной длины (ось X – п.н.), которая может быть локализована при заданной длине прилегающего к стыку сайта сплайсинга участка.

Fig. 3. The maximal portion of interexonic reads in the total number of reads of a fixed size (X axis: base pairs), that can be mapped under the condition of a preset minimal length of anchors on each side of the junction used to detect the splice site.

Таблица 2

Зависимость удельной цены при секвенировании 1 Гб для различных длин непарных ридов

Платформа	Длина рида, п.н.	Стоимость секвенирования 1 Гб, долларов США	Коэффициент пересчета стоимости (при длине прилегающего участка к стыку 15 п.н.)	Стоимость получения одинакового количества межэкзонных ридов, долларов США
SOLID	50	70	1,75	122,5
SOLID	75	55,6	1,17	64,9
Illumina	50	110	1,75	192,5
Illumina	75	90	1,17	105,0
Illumina	100	75	1,00	75,0

стыку сайта сплайсинга участка – 12–13 п.н. (рис. 4, б). Следовательно, при картировании парных ридов транскриптома, полученных из образцов тканей ЦНС, целесообразно использовать значения 12–13 п.н.

ПРОБЛЕМА ВЫБОРА АННОТАЦИИ РЕФЕРЕНСНОГО ГЕНОМА

Отдельно стоит отметить проблему выбора аннотации референсного генома для локализации ридов у организмов, для которых существует более чем одна референсная аннотация. Так, для крысы существует три общедоступных и постоянно обновляемых ресурса с аннотацией генома и транскриптома – Ensemble, RefSeq и Genscan. К сожалению, ни одна из предложенных аннотаций не позволяет идентифицировать все транскрипты, представленные в общем пуле мРНК. Вместе с тем из 10 000–12 000 транскриптов с представленностью не ниже 1 рида на 1 000 п.н., выявляемых при аннотировании ридов с использованием ресурсов Ensemble и Genscan, неидентифицированными остаются лишь порядка 300 и 500 транскриптов соответственно. В то же время применение аннотации RefGen, включающей лишь известные мРНК, неидентифицированными остаются порядка 800 транскриптов. Следовательно, для геномной локализации ридов и последующей сборки транскриптомов тканей ЦНС целесообразно использовать аннотации, включающие не только уже известные мРНК, но и последовательности транскриптов, предсказанные методом *ab initio*.

ЗАКЛЮЧЕНИЕ

Представленные данные подтверждают сложность процесса сборки транскриптомов для тканей нервной системы и необходимость дальнейшего совершенствования как методов локализации и сборки, так и существующих аннотаций. Показано, что для транскриптомов из нервной ткани характерна недопредставленность межэкзонных ридов по сравнению с удельной долей таких ридов в транскриптомах из других тканей. Проведенный анализ стоимости–эффективности также однозначно свидетельствует о необходимости использования при изучении

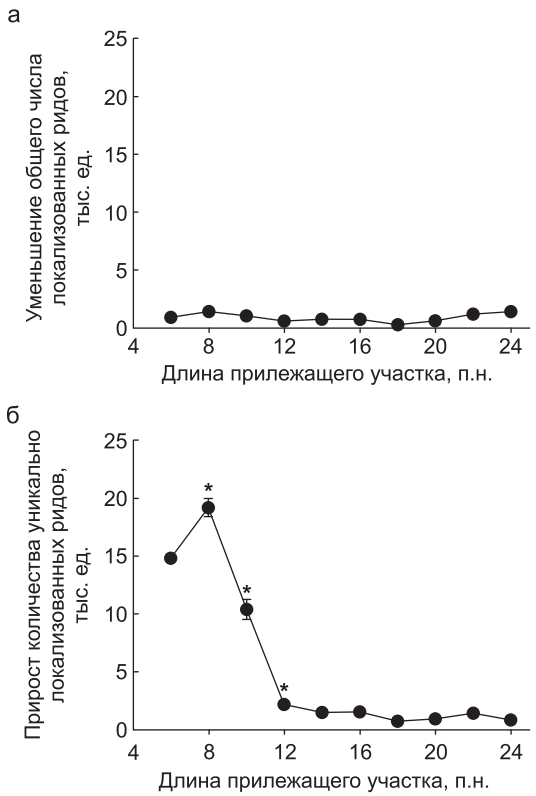

Рис. 4. Зависимость общего числа локализованных парных ридов (а) и числа уникально локализованных парных ридов (б) от длины прилежащего к стыку сайта сплайсинга участка, который был использован для локализации стыков сайтов сплайсинга.

***** $p < 0{,}05$ по сравнению с приростом количества ридов, локализованных при меньшем значении длины прилежащего участка.

Fig. 4. The numbers of mapped reads as functions of the fixed minimal length of anchors on each side of the junction used to detect the splice site: (a) total number of mapped single reads, (b) uniquely mapped single reads.

* $p < 0{,}05$ vs the increase in the number of reads mapped with smaller anchor lengths on each side of the junction used to detect the splice site.

транскриптома мозга технологий секвенирования, дающих длины ридов не менее 75 п.н. Более предпочтительным является использование методов секвенирования, дающих парные риды. Следует отметить, что для геномной локализации ридов и последующей сборки транскриптомов тканей ЦНС целесообразно использовать аннотации, включающие не только уже известные мРНК, но и последовательности транскриптов, предсказанные методом *ab initio*, – например аннотации консорциума Ensemble.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке гранта РНФ 14-15-00115.

ЛИТЕРАТУРА

Buermans H.P., den Dunnen J.T. Next generation sequencing technology: Advances and applications // Biochim. Biophys. Acta. 2014. V. 1842. No. 10. P. 1932–1941.

Entrez Genome Database [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); [Updated 01.09.2014]. Available from: http://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/

Grishkevich V., Yanai I. Gene length and expression level shape genomic novelties // Genome Res. 2014. V. 24. No. 9. P. 1497–1503.

Hubbard T., Barker D., Birney E., Cameron G., Chen Y., Clark L., Cox T., Cuff J., Curwen V., Down T. *et al*. The Ensembl genome database project // Nucl. Acids Res. 2002. V. 30. No. 1. P. 38–41.

Irimia M., Denuc A., Burguera D., Somorjai I., Martín-Durán J.M., Genikhovich G., Jimenez-Delgado S., Technau U., Roy S.W., Marfany G., Garcia-Fernàndez J. Stepwise assembly of the Nova-regulated alternative splicing network in the vertebrate brain // Proc. Natl Acad. Sci. USA. 2011. V. 108. No. 13. P. 5319–5324.

Lin L., Shen S., Jiang P., Sato S., Davidson B.L., Xing Y. Evolution of alternative splicing in primate brain transcriptomes // Hum. Mol. Genet. 2010. V. 19. No. 15. P. 2958–2973.

Martin J.A., Wang Z. Next-generation transcriptome assembly // Nat. Rev. Genet. 2011. V. 12. No. 10. P. 671–682.

Mills J.D., Janitz M. Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases // Neurobiol. Aging. 2012. V. 33. No. 5. P. 1012.e11-24.

Oltean S., Bates D.O. Hallmarks of alternative splicing in cancer // Oncogene. 2014 [Epub ahead of print] doi:10.1038/onc.2013.533.

Parada G.E., Munita R., Cerda C.A., Gysling K. A comprehen-sive survey of non-canonical splice sites in the human transcriptome // Nucl. Acids Res. 2014. [Epub ahead of print] doi:10.1093/nar/gku744.

Rogers M.F., Thomas J., Reddy A.S., Ben-Hur A. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data // Genome Biol. 2012. V. 13. No. 1. P. R4.

Salzberg S.L., Phillippy A.M., Zimin A., Puiu D., Magoc T., Koren S., Treangen T.J., Schatz M.C., Delcher A.L., Roberts M., Marçais G., Pop M., Yorke J.A. GAGE: A critical evaluation of genome assemblies and assembly algorithms // Genome Res. 2012. V. 22. No. 3. P. 557–567.

Tollervey J.R., Wang Z., Hortobágyi T., Witten J.T., Zarnack K., Kayikci M., Clark T.A., Schweitzer A.C., Rot G., Curk T., Zupan B., Rogelj B., Shaw C.E., Ule J. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain // Genome Res. 2011. V. 21. No. 10. P. 1572–1582.

Treangen T.J., Salzberg S.L. Repetitive DNA and next-generation sequencing: computational challenges and solutions // Nat. Rev. Genet. 2011. V. 13. No. 1. P. 36–46.

Ule J., Stefani G., Mele A., Ruggiu M., Wang X., Taneri B., Gaasterland T., Blencowe B.J., Darnell R.B. An RNA map predicting Nova-dependent splicing regulation // Nature. 2006. V. 444. No. 7119. P. 580–586.

van Bakel H., Nislow C., Blencowe B.J., Hughes T.R. Most «dark matter» transcripts are associated with known genes. // PLoS Biol. 2010. V. 8. No. 5. P. e1000371.

Yu Y., Fuscoe J.C., Zhao C., Guo C., Jia M., Qing T., Bannon D.I., Lancashire L., Bao W., Du T., Luo H., Su Z., Jones W.D., Moland C.L., Branham W.S., Qian F., Ning B., Li Y., Hong H., Guo L., Mei N., Shi T., Wang K.Y., Wolfinger R.D., Nikolsky Y., Walker S.J., Duerksen-Hughes P., Mason C.E., Tong W., Thierry-Mieg J., Thierry-Mieg D., Shi L., Wang C. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages // Nat. Commun. 2014. V. 5. P. 3230.

METHODOLOGICAL ASPECTS OF READ MAPPING AND ASSEMBLY OF TRANSCRIPTOMES DERIVED FROM BRAIN TISSUE SAMPLES OF *RATTUS NORVEGICUS*

P.N. Menshanov[1, 2], N.N. Dygalo[1, 2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia,
e-mail: MenshanovPN@icg.sbras.ru;
[2] Novosibirsk State University, Novosibirsk, Russia

Summary

Identification of all transcripts expressed in a biological sample and quantification of transcript levels are two major objectives in transcriptome research. They are particularly challenging in case of CNS transcriptomes, since alternative splicing is of wide occurrence in the brain. This paper recognizes and analyzes several problems associated with read mapping and subsequent assembly of transcriptomes derived from samples of CNS tissue, in particular: unambiguous read identification, read size, and incomplete reference annotations available for both model and non-model species. It is shown that the relative abundance of interexonic reads in transcriptomes derived from CNS tissues is lower than in those derived from non-CNS tissues provided that identical procedures of read sequencing, genomic mapping, and transciptome assembly are applied. The underrepresentation of interexonic reads in the transcriptomes derived from CNS vs. non-CNS tissues appears to be indicative of the existence of a large number of transcripts with novel sites of splicing not annotated yet. Cost-benefit analysis affirms the superiority of sequencing technologies that generate reads with lengths ≥ 75 bp. For genomic mapping of reads and subsequent transcriptome assembly, it is advisable to use genomic annotations that include both known and ab initio predicted transcripts, such as Ensemble transcriptome annotations.

Key words: next generation sequencing, transcriptome, read mapping, transcriptome assembly, genome annotation.