УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА.
СЕРИЯ ЕСТЕСТВЕННЫЕ НАУКИ

2018, Т. 160, кн. 2
С. 214–226

ISSN 2542-064X (Print)
ISSN 2500-218X (Online)

УДК 575.113

СБОРКА ГЕНОМА АНГИДРОБИОТИЧЕСКОГО НАСЕКОМОГО *Polypedilum vanderplanki* С ИСПОЛЬЗОВАНИЕМ ДАННЫХ ILLUMINA И PACBIO

О.С. Козлова, З.И. Абрамова

Казанский (*Приволжский*) *федеральный университет, г. Казань, 420008, Россия*

Аннотация

В статье представлена новая версия сборки генома африканского ангидробиотического комара-звонца *Polypedilum vanderplanki* на основе глубокого секвенирования ДНК, выделенной из клеточной линии Pv11, полученной из эмбриональной массы насекомого. Исходными данными послужили парные прочтения с варьирующей длиной вставки, дополненные ультрадлинными прочтениями по технологии Pacific Biosciences. Показано, что результирующий набор скаффолдов обладает более высокими метриками целостности и полноты, кроме того, позволяет более корректно выполнять предсказание кодирующих последовательностей по сравнению с предыдущей версией сборки генома, что было продемонстрировано на примере белков теплового шока HSP20 и HSP70.

Ключевые слова*: Polypedilum vanderplanki*, ангидробиоз, секвенирование ДНК, сборка генома

Введение

Polypedilum vanderplanki (африканский комар-звонец семейства Chironomidae) – наиболее высокоорганизованный биологический вид, способный входить в состояние ангидробиоза и сохранять жизнеспособность в условиях экстремальной засухи [1]. В случае обезвоживания личинки данного вида теряют до 97% воды, что приводит к остановке всех метаболических и физиологических процессов, однако в течение часа после помещения их в воду личинки восстанавливают свою активность и способны к продолжению жизненного цикла [2]. Интересно, что подобную же удивительную толерантность личинки *P. vanderplanki* демонстрируют и по отношению к другим разновидностям экстремального абиотического стресса, например к воздействию повышенных концентраций солей [3] и критических для других организмов доз ионизирующей радиации [4]. Таким образом, ангидробиоз у данного вида можно рассматривать в качестве примера уникальной реакции на абиотический стресс у насекомых; его изучение с практической точки зрения обеспечит глубокое понимание молекулярных механизмов ангидробиоза *P. vanderplanki* и откроет широкие перспективы для разработки технологий безводного хранения клеток и тканей.

Первая версия сборки генома *P. vanderplanki* была опубликована в 2014 г. и имела сравнительно высокие статистические показатели целостности и полноты относительно референсного набора генов эукариот [5]. Тем не менее данная версия оказалась не вполне пригодной для дальнейшей исследовательской работы, в частности, в области сравнительной геномики по причине не всегда корректной структурной и функциональной аннотации белок-кодирующих областей. Среди вероятных причин ограниченных возможностей использования текущей сборки – выделение ДНК из целой личинки, что не гарантирует отсутствия бактериальных и иных контаминаций, применение неоптимальных методов аннотации, а также объективные неточности в самой сборке, повлекшие за собой некорректное моделирование структуры генов.

Исходя из вышесказанного было принято решение сконструировать альтернативную версию генома *P. vanderplanki* с использованием ДНК, выделенной из культуры клеток Pv11, полученной из эмбриональной массы насекомого и после преинкубации в трегалозе также способной входить в состояние ангидробиоза. Библиотеки ДНК были секвенированы на платформах Illumina (секвенаторы HiSeq и Miseq, включая парноконцевые прочтения с варьирующей длиной вставки), а с целью повышения целостности сборки набор коротких прочтений был дополнен данными секвенирования длинных прочтений по технологии Pacific Biosciences (PacBio) [6].

1. Методы

1.1. Исходные данные и предварительная обработка. В зависимости от технологий секвенирования и особенностей подготовки библиотек весь исходный набор данных секвенирования делился на три части: парноконцевое секвенирование с короткой длиной вставки (paired-end), парноконцевое секвенирование с длинной вставкой (mate-paired) и секвенирование длинных прочтений по технологии PacBio. Данные секвенирования библиотек с короткой вставкой включали в себя 62 млн пар чтений в режиме 101 + 101 (Hiseq2000), 75 млн пар чтений в режиме 101 + 101 (Hiseq1500) и 17 млн пар чтений в режиме 262 + 262 (MiSeq). Парноконцевое секвенирование с длинной вставкой (режим 101 + 101) было представлено библиотеками с размером вставки 5–6 Кб (21.6 млн пар), 6–7 Кб (24.8 млн пар) и 8–10 Кб (38.9 млн пар), полученными на секвенаторе Hiseq2000. Данные секвенирования, полученные по технологии PacBio, включали в себя 1.17 млн прочтений, или 74180 откорректированных с помощью выравнивания последовательностей в fasta-формате со средней длиной 15231 и стандартным отклонением 4941 нуклеотид.

Выделение ДНК осуществлялось в соответствии со стандартными протоколами подготовки образцов. Секвенирование по технологии Illumina проводилось на базе лаборатории эволюционной геномики (ФББ МГУ, г. Москва), лаборатории «Экстремальная биология» (КФУ, г. Казань) и в Национальном институте общей биологии (National Institute for Basic Biology, Япония), в котором было проведено и секвенирование по технологии PacBio. Данные секвенирования были любезно предоставлены сотрудниками указанных лабораторий.

Для предобработки чтений Illumina с короткой вставкой была использована программа Trimmomatic [7] в режимах фильтрации нуклеотидов с конца про-

чтения по качеству и удаления адаптеров. Предобработка чтений с длинной вставкой осуществлялась в программе NxTrim [8] с параметрами по умолчанию.

1.2. Анализ распределения кратности *k*-меров. Дополнительным источником информации о способности чтений привести к результативной сборке является анализ *k*-меров, имеющий важное прогностическое значение, поскольку позволяет оценить уровень гетерозиготности и более точно вычислить покрытие генома чтениями. Под *k*-мером следует понимать небольшую последовательность нуклеотидов длиной k ($k <$ длины чтения). Анализ распределения кратности *k*-меров проводился в программе Kmergenie [9] (с аппроксимацией согласно диплоидной модели) при $k = 21$.

1.3. Методы сборки генома. Известны три базовые концепции использования длинных чтений PacBio для *de-novo* сборки и улучшения существующих сборок геномов:

1) сборка генома *de-novo* исключительно с использованием чтений PacBio;

2) гибридная сборка *de-novo*, в рамках которой на разных этапах сборки короткие и длинные чтения используются совместно;

3) скаффолдинг (объединение последовательностей в более крупные) и закрытие гэпов (замена неизвестных нуклеотидов известными) в контигах, полученных на основе коротких чтений, с помощью чтений платформы PacBio.

Начальные версии геномных сборок были получены с использованием каждой из трёх вышеописанных концепций. Сборка генома с помощью только чтений PacBio (программа HGAP4 [10]) была предоставлена Национальным институтом общей биологии (National Institute for Basic Biology, Япония). В качестве программы, реализующей концепцию гибридной сборки, была выбрана программа DBG2OLC [11], название которой отражает два базовых подхода к геномной сборке: построение графов де Брюйна (de Bruijn Graph, DBG) и консенсус по перекрытию (Overlap Layout Consensus, OLC).

Для реализации третьей концепции сначала потребовалось получить сборку на основе всех имеющихся чтений Illumina в программе Platanus [12]. Скаффолдинг и закрытие гэпов в этой предварительной сборке осуществлялся в программном конвейере PBJelly [13]. Отметим, что именно этот подход использовался для создания предыдущей версии сборки генома *P. vanderplanki* [5], и, таким образом, сборку генома клеточной линии Pv11 на основе программы Platanus с дальнейшей обработкой в конвейере PBJelly, с точки зрения методологии, можно считать аналогом сборки генома личинки.

Предыдущая версия сборки генома *P. vanderplanki* была загружена из базы данных MidgeBase (bertone.nises-f.affrc.go.jp/midgebase).

1.4. Получение мета-сборки. Традиционно из нескольких вариантов сборки генома, полученных с использованием разных данных и программных продуктов, выбирается наиболее полная и целостная. Однако есть и альтернативный способ решения этой задачи, который заключается в создании новой, гибридной сборки, объединяющей в себе преимущества разных сборок-предшественников. Данный методологический подход успешно зарекомендовал себя в ряде сравнительных

тестов проекта Assemblathon, причём как в работе с симулированными данными секвенирования генома человека, так и на реальных чтениях ДНК рыбы псевдотрофеус-зебра, волнистого попугайчика и удава обыкновенного [14]. Именно такой подход, заключающийся в создании новой метасборки из трёх сборок, реализующих различные концепции использования чтений PacBio, был использован для получения оптимального набора скаффолдов.

С целью объединения трёх предварительных вариантов геномных сборок в одну был использован программный конвейер Metassembler [14], который картирует парноконцевые чтения с длинной вставкой на геномы и, таким образом, определяет соответствие друг другу скаффолдов из разных наборов. Особенностью конвейера Metassembler является итеративный подход: так, на каждой итерации происходит объединение двух сборок, а на следующем шаге новая сборка добавляется к результатам объединения предыдущих. Таким образом, последовательность, в которой сборки будут подаваться в работу конвейера, – важный детерминант его работы. В данном случае в качестве базовой сборки использовали наиболее полную, наименее дуплицированную, однако и наименее целостную сборку (Platanus + PBJelly), объединив её со сборкой с минимальным количеством скаффолдов (HGAP4), а на второй итерации добавили к этой гибридной сборке оставшуюся (DBG2OLC).

1.5. Методы скаффолдинга. Дополнительный скаффолдинг сборки Platanus + PBJelly, а также скаффолдинг мета-сборки были проведены с помощью программы SSPACE [15] и GapFiller [16], работа которых основана на данных парноконцевого секвенирования с длинной вставкой. Кроме того, для скаффолдинга можно использовать также данные РНК-секвенирования, которые в случае культуры клеток Pv11 представлены библиотеками РНК-чтений, отражающих разные стрессовые условия (в общей сложности 1.6 млрд одноконцевых чтений длиной 50 п.н., полученных на приборах Illumina). Все имеющиеся данные секвенирования РНК были собраны в *de-novo* транскриптом с помощью программы Trinity [17], что дало 41077 транскриптов, а затем выровнены на геномные последовательности программой blat [18], после чего был проведён скаффолдинг (программа lRNAScaffolder [19]).

1.6. Оценка качества сборки генома. Основные статистические показатели геномных сборок оценивались в программе Quast [20]. Полнота сборки, то есть её способность содержать в себе последовательности, кодирующие однокопийные белки, специфичные для определённой группы организмов, оценивалась в программе BUSCO [21], и в качестве меры полноты сборки использовались референсные белки двукрылых (2799 белков).

2. Результаты

2.1. Варианты геномных сборок и их характеристики. После незначительной предварительной обработки все короткие чтения Illumina были объединены в один файл для анализа кратности *k*-меров, который позволил оценить ожидаемый размер генома и более точно рассчитать его покрытие. Чтения PacBio не были вовлечены в этот процесс по причине, несопоставимой с короткими

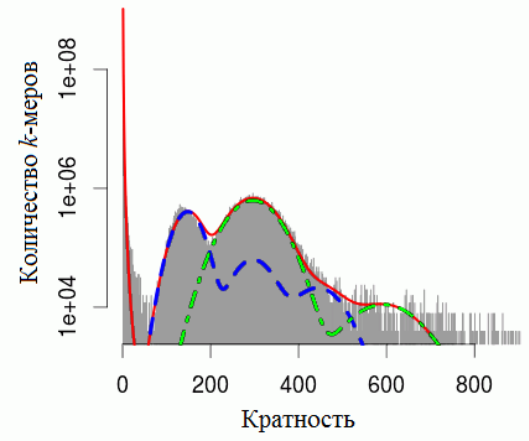

Рис. 1. Гистограмма распределения кратностей 21-меров на основе всех данных Illumina. Красная линия отражает аппроксимацию всей статистической модели (ошибочные *k*-меры и геномные *k*-меры), синяя и зелёная – аппроксимацию моделей для гетерозиготных и гомозиготных *k*-меров соответственно

чтениями длины, а также существенно меньшего объёма данных, учёт которого не оказал бы сильного влияния на результат. Результатом анализа *k*-меров является так называемая гистограмма распределения кратностей, представленная на рис. 1 и позволяющая определить, какое количество последовательностей заданной длины *k* имеется в данных с определённой кратностью, при этом считается, что в одном чтении не может содержаться двух одинаковых *k*-меров. Как следует из рис. 1, в среднем на каждый гомозиготный 21-мер приходится по 300 чтений, и, зная длину чтений (~ 100 п.н.), легко перейти от покрытия чтениями на *k*-мер к покрытию чтениями на нуклеотид [22], которое можно оценить как 375-кратное. Гистограмма распределения кратностей 21-меров демонстрирует чёткое разделение массы коротких последовательностей на потенциально ошибочные, редкие *k*-меры (*k*-меры, которым соответствуют кратности от 1 до начала первого подъёма гистограммы, что в данном случае составляет порядка 50) и действительно геномные *k*-меры, что позволяет ожидать получение качественной, релевантной геномной сборки. Оценка размера генома методом *k*-мер анализа составляет порядка 112 Мб.

Статистические метрики трёх вариантов геномных сборок, а также ранее опубликованной версии сборки генома *P. vanderplanki* представлены в табл. 1. В последнем столбце таблицы приведён процент ультраконсервативных одно-копийных белков двукрылых, классифицированных как полностью представленные в геноме (C), как дуплицированные (D) и как пропущенные (M).

Очевидно, что ни один из трёх вариантов сборок генома не может считаться однозначно предпочтительным. Сборка, полученная при помощи программы HGAP4, хоть и обладает уникальным для *de-novo* сборки показателем N50 (более 2 Мб), тем не менее имеет самый низкий уровень полноты (C = 82.7%), а также наивысший уровень дуплицированности, что хорошо согласуется с её слегка увеличенным размером (133 Мб). Напротив, версия сборки с наилучшими показателями полноты (Platanus + PBJelly) имеет сравнительно скромный N50

Табл. 1

Статистические метрики трёх первоначальных вариантов сборки генома и предыдущей сборки

Сборка	Всего скаффолдов	Скаффолдов, ≥ 10 Кб	Размер, Мб	N50, п.н.	BUSCO
HGAP4	401	324	133.19	2 276 022	C: 82.7%; D: 4%; M: 6.8
DBG2OLC	989	984	101.54	182 666	C: 88.4%; D: 1.3%; M: 6.8
Platanus + PBJelly	83195	1103	120.23	185 080	C: 96.9%; D: 0.8%; M: 1.4
Ранее опубликованная сборка	80283	619	117.14	264 320	C: 94.8%; D: 0.6%; M: 1.7

(185 Кб) и наибольшее количество скаффолдов. Именно поэтому вместо выбора из трёх вариантов наилучшего все они были объединены в одну мета-сборку. Однако ввиду расхождений между статистическими характеристиками сборок HGAP4 и Platanus + PBJelly на порядок, прежде чем приступать к их объединению, последняя была предварительно подвергнута дополнительным скаффолдингу и закрытию гэпов с использованием чтений с длинной вставкой, в результате чего N50 удалось увеличить до 627 Кб, незначительно снизив общее количество последовательностей (81 374) и оставив показатели полноты сборки неизменными. Одно из типичных следствий скаффолдинга – увеличение числа неизвестных нуклеотидов (*N*), которые образуются между составленными в скаффолд контигами и которые не удаётся разрешить с помощью закрытия гэпов; в данном случае, для сборки Platanus + PBJelly после скаффолдинга число *N* выросло с 5.88 до 2910.2 на 100 Кб.

Результатом объединения трёх первоначальных вариантов сборок в одну стал набор из 581 скаффолда, из которых 390 имели длину более 10 Кб. Метрика N50 по скаффолдам новой мета-сборки составила 658565 п.н., общий размер генома – 120.18 Мб, количество *N* (неизвестных нуклеотидов) на 100 Кб – 1584.39, а полнота согласно BUSCO – C: 96.8%, D: 0.9%, M: 1.6%. Таким образом, хотя с точки зрения метрики N50 новая мета-сборка не существенно улучшает результат той, которая использовалась в качестве базовой, количество последовательностей в ней существенно снижено, кроме того, отсутствуют дуплицированные участки генома, что позволяет сделать вывод об объединении всех преимуществ сборок, полученных с использованием разных подходов, в одной.

Статистика трёх этапов сборки после объединения (мета-сборка (M), мета-сборка после скаффолдинга геномными библиотеками с длинной вставкой (MS), мета-сборка после скаффолдинга геномными и транскриптомными библиотеками (MST)) приведена в табл. 2.

Таким образом, удалось сконструировать набор протяжённых скаффолдов, полнота которых согласно BUSCO составляет 96.8%. Длина собранного генома

Табл. 2

Статистические метрики трёх этапов сборки генома после объединения

Сборка	Всего скаффолдов	Скаффолдов, ≥ 10 Кб	N50, п.н.	Количество N на 100 Кб
M	581	390	658 565	1584.39
MS	451	281	968 491	1677.24
MST	440	272	1 001 272	1699.08

равна 120.38 Мб, при этом скаффолды длиной более 1000 п. н. составляют 120.00 Мб, а 118.5 Мб (98.4% от общей длины сборки) представлено скаффолдами длиной более 50 Кб. Полнота представленности консервативных белков двукрылых в новом варианте сборки генома несколько превышает результат предыдущей сборки, а число неизвестных нуклеотидов *N* на 100 Кб длины, напротив, существенно меньше, чем в предыдущей сборке (1699.08 против 4167.81). Особенно следует отметить, что в новом варианте сборки удалось почти вчетверо увеличить основной показатель целостности сборки – N50, который теперь превышает 1 Мб.

3.2. Уточнение структур генов, представляющих интерес. Преимущества нового варианта сборки заключаются не только в увеличенных значениях статистических показателей, но и в возможности получения более точной информации о структуре и свойствах кодирующих последовательностей. Даже полностью автоматическая структурная разметка генома на кодирующие участки, произведённая на основе 1.6 млрд чтений РНК культуры клеток, позволила сделать вывод, что модели генов, полученные на основе предыдущего варианта сборки генома, не всегда являются корректными по причине недостаточной протяжённости и целостности скаффолдов именно в местах расположения генов.

В подтверждение этому приведём два примера исправленной разметки генов, кодирующих белки теплового шока, поскольку эта функция является интересной с точки зрения вовлечённости в ответ на обезвоживание. Ранее было показано, что по крайней мере несколько белков теплового шока (БТШ или HSP, heat-shock proteins) вовлечены в процесс ангидробиоза [5, 23–24]. С точки зрения сравнительной геномики возникает необходимость в получении максимально точных последовательностей соответствующих им генов, так как это позволит с уверенностью искать ортологичные гены в других организмах и проверять различные гипотезы о специфичности или универсальности ответа на разный стресс в разных видах.

Ранее считалось, что в геноме *P. vanderplanki* содержится два паралогичных гена, кодирующих малые белки теплового шока HSP20 (Pv.12121 и Pv.02260) и расположенные в разных скаффолдах, при этом один из них (Pv.12121) имел характерный для этого семейства альфа-кристаллиновый домен. Примечательно, что гены, кодирующие данные белковые последовательности, согласно информации из базы данных MidgeBase, имели практически идентичный профиль экспрессии, кроме того, они обращали на себя внимание максимальным показателем экспрессии среди всех белков данного семейства (до 40 000 RPKM на 48-й час обезвоживания). Сходный профиль экспрессии, а также отсутствие альфа-кристаллинового домена у одного из белков позволили предположить, что в данном

случае паралогизация является мнимой, и вызвана она дефектом сборки. Действительно, создав выборку из ДНК-чтений, картирующихся на каждый из фрагментов, которые кодируют эти мнимые паралоги, а затем собрав только эти чтения *de-novo*, можно получить одну протяжённую кодирующую последовательность. Именно эта последовательность, образованная из двух предыдущих с частичным перекрытием, была найдена в новом варианте сборки, и, таким образом, было независимо подтверждено, что две последовательности, ранее считавшиеся двумя разными генами, наиболее вероятно, кодируют всего один белок. Следует заметить, что предсказанная молекулярная масса белка, кодируемого данной последовательностью, нехарактерно велика для малых БТШ (82 кДа), и сам белок не имеет ортологов в других организмах. Разумеется, природа появления этого интересного гена в геноме *P. vanderplanki* требует дальнейшего исследования, хотя в качестве гипотезы можно предположить, что такая последовательность появилась в результате слияния генов (gene fusion).

Второй сходный пример касается белка теплового шока с молекулярной массой 70 кДа (HSP70). Ранее считалось, что в геноме *P. vanderplanki* содержится три разных гена с двумя экзонами (Pv.12516, Pv.17811, Pv.15766), причём все они были расположены в начале трёх разных скаффолдов. Подобно ранее рассмотренному HSP20 гены демонстрировали одинаковый профиль экспрессии, а соответствующие им чтения собирались в один протяжённый фрагмент. Примечательно, что длины кодирующих последовательностей двух генов из трёх были подозрительно малы, как и предсказанные молекулярные массы (Pv.12516 – 23 кДа, Pv.17811 – 11 кДа, Pv.15766 – 10 кДа), и, хотя в описании базы данных они классифицировались как БТШ70, поскольку успешно выравнивались на соответствующие гены других хирономид, в реальности скорее представляли собой лишь части их кодирующих последовательностей. В настоящем варианте сборки белки, соответствующие этим генам, картируются на один и тот же белок с предсказанной молекулярной массой 69 кДа. Таким образом, можно предположить, что мы имеем дело не с тремя паралогичными генами, в каждом из которых по два экзона, а с одним, мультиэкзонным геном. Это подтверждается высокой гомологией результирующей последовательности с генами, кодирующими БТШ70 у других хирономид, в частности *Chironomus riparus* и *Diamesa cinerella*.

Заключение

Предлагаемый вариант сборки генома ангидробиотического насекомого *Polypedilum vanderplanki* представляет собой качественно новый набор протяжённых скаффолдов, который является гораздо более перспективным для проведения исследований в области геномики, по сравнению с предыдущим. Это подтверждается не только увеличенными показателями целостности и полноты сборки, но также более корректными предсказаниями последовательностей генов, кодирующих белки, представляющие особый интерес. Использование новой версии сборки открывает перспективы для проведения более точной структурной и функциональной аннотации кодирующих последовательностей, исследования пространственной организации хроматина методом HiC и реализации многих других молекулярно-биологических и биоинформатических методик.

Благодарности. Работа выполнена за счет средств субсидии, выделенной в рамках государственной поддержки Казанского (Приволжского) федерального университета в целях повышения его конкурентоспособности среди ведущих мировых научно-образовательных центров.

Литература

1. *Cornette R., Kikawada T.* The induction of anhydrobiosis in the sleeping chironomid: Current status of our knowledge // IUBMB Life. – 2011. – V. 63, No 6. – P. 419–429. – doi: 10.1002/iub.463.
2. *Nakahara Y., Watanabe M., Fujita A., Kanamori Y., Tanaka D., Iwata K., Furuki T., Sakurai M., Kikawada T., Okuda T.* Effects of dehydration rate on physiological responses and survival after rehydration in larvae of the anhydrobiotic chironomid // J. Insect. Physiol. - 2008. – V. 54, No 8. – P. 1220–1225. – doi: 10.1016/j.jinsphys.2008.05.007.
3. *Watanabe M., Kikawada T., Okuda T.* Increase of internal ion concentration triggers trehalose synthesis associated with cryptobiosis in larvae of *Polypedilum vanderplanki* // J. Exp. Biol. – 2003. – V. 206, Pt. 13. – P. 2281–2286. – doi: 10.1242/jeb.00418.
4. *Ryabova A., Mukae K., Cherkasov A., Cornette R., Shagimardanova E., Sakashita T., Okuda T., Kikawada T., Gusev O.* Genetic background of enhanced radioresistance in an anhydrobiotic insect: Transcriptional response to ionizing radiations and desiccation // Extremophiles. – 2017. – V. 21, No 1. – P. 109–120. – doi: 10.1007/s00792-016-0888-9.
5. *Gusev O., Suetsugu Y., Cornette R., Kawashima T., Logacheva M.D., Kondrashov A.S., Penin A.A., Hatanaka R., Kikuta S., Shimura S., Kanamori H, Katayose Y., Matsumoto T., Shagimardanova E., Alexeev D., Govorun V., Wisecaver J., Mikheyev A., Koyanagi R., Fujie M., Nishiyama T., Shigenobu S., Shibata T.F., Golygina V., Hasebe M., Okuda T., Satoh N., Kikawada T.* Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge // Nat. Commun. – 2014. – V. 5. – Art. 4784, P. 1–9. – doi: 10.1038/ncomms5784.
6. *Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B., Bibillo A., Bjornson K., Chaudhuri B., Christians F., Cicero R., Clark S., Dalal R., Dewinter A., Dixon J., Foquet M., Gaertner A., Hardenbol P., Heiner C., Hester K., Holden D., Kearns G., Kong X., Kuse R., Lacroix Y., Lin S., Lundquist P., Ma C., Marks P., Maxham M., Murphy D., Park I., Pham T., Phillips M., Roy J., Sebra R., Shen G., Sorenson J., Tomaney A., Travers K., Trulson M., Vieceli J., Wegener J., Wu D., Yang A., Zaccarin D., Zhao P., Zhong F., Korlach J., Turner S.* Real-time DNA sequencing from single polymerase molecules // Science. – 2009. – V. 323, No 5910. – P. 133–138. – doi: 10.1126/science.1162986.
7. *Bolger A.M., Lohse M., Usadel B.* Trimmomatic: A flexible trimmer for Illumina sequence data // Bioinformatics. – 2014. – V. 30, No 15. – P. 2114–2120. – doi: 10.1093/bioinformatics/btu170.
8. *O'Connell J., Schulz-Trieglaff O., Carlson E., Hims M.M., Gormley N.A., Cox A.J.* NxTrim: Optimized trimming of Illumina mate pair reads // Bioinformatics. – 2015. – V. 31, No 12. – P. 2035–2037. – doi: 10.1093/bioinformatics/btv057.
9. *Chikhi R., Medvedev P.* Informed and automated k-mer size selection for genome assembly // Bioinformatics. – 2014. – V. 30, No 1. – P. 31–37. – doi: 10.1093/bioinformatics/btt310.
10. *Chin C.S., Alexander D.H., Marks P., Klammer A.A., Drake J., Heiner C., Clum A., Copeland A., Huddleston J., Eichler E.E., Turner S.W., Korlach J.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data // Nat. Methods. – 2013. – V. 10, No 6. – P. 563–569. – doi: 10.1038/nmeth.2474.

11. *Ye Ch., Hill C.M., Wu Sh., Ruan J., Ma Zh.* (*Sam*). DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies // Sci. Rep. – 2016. – V. 6. – Art 31900, P. 1–9. – doi: 10.1038/srep31900.

12. *Kajitani R., Toshimoto K., Noguchi H., Toyoda A., Ogura Y., Okuno M., Yabana M., Harada M., Nagayasu E., Maruyama H., Kohara Y., Fujiyama A., Itoh T.* Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads // Genome Res. – 2014. – V. 24, No 8. – P. 1384–1395. – doi: 10.1101/gr.170720.113.

13. *English A.C., Richards S., Han Y., Wang M., Vee V., Qu J., Qin X., Muzny D.M., Reid J.G., Worley K.C., Gibbs R.A.* Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology // PLoS ONE. – 2012. – V. 7, No 11. – Art. e47768, P. 1–12. – doi: 10.1371/journal.pone.0047768.

14. *Wences A.H., Schatz M.C.* Metassembler: Merging and optimizing de novo genome assemblies // Genome Biol. – 2015. – V. 16. – Art. 207, P. 1–10. – doi: 10.1186/s13059-015-0764-4.

15. *Boetzer M., Henkel C.V., Jansen H.J., Butler D., Pirovano W.* Scaffolding pre-assembled contigs using SSPACE // Bioinformatics. – 2011. – V. 27, No 4. – P. 578–579. – doi: 10.1093/bioinformatics/btq683.

16. *Nadalin F., Vezzi F., Policriti A.* GapFiller: A de novo assembly approach to fill the gap within paired reads // BMC Bioinf. – 2012. – V. 13, Suppl. 14. – Art. S8, P. 1–16. – doi: 10.1186/1471-2105-13-S14-S8.

17. *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. – V. 29, No 7. – P. 44–52. – doi: 10.1038/nbt.1883.

18. *Kent W.J.* BLAT – the BLAST-like alignment tool // Genome Res. – 2002. – V. 12, No 4. – P. 656–664. – doi: 10.1101/gr.229202.

19. *Xue W, Li J.T., Zhu Y.P., Hou G.Y., Kong X.F., Kuang Y.Y., Sun X.W.* L_RNA_scaffolder: Scaffolding genomes with transcripts // BMC Genomics. – 2013. – V. 14. – Art. 604, P. 1–14. – doi: 10.1186/1471-2164-14-604.

20. *Gurevich A., Saveliev V., Vyahhi N., Tesler G.* QUAST: Quality assessment tool for genome assemblies // Bioinformatics. – 2013. – V. 29, No 8. – P. 1072–1075. – doi: 10.1093/bioinformatics/btt086.

21. *Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M.* BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs // Bioinformatics. – 2015. – V. 31, No 19. – P. 3210–3212. – doi: 10.1093/bioinformatics/btv351.

22. *Li X., Waterman M.S.* Estimating the repeat structure and length of DNA sequences using L-tuples // Genome Res. – 2003. – V. 13, No 8. – P. 1916–1922. – doi: 10.1101/gr.1251803.

23. *Mazin P.V., Shagimardanova E., Kozlova O., Cherkasov A., Sutormin R., Stepanova V.V., Stupnikov A., Logacheva M., Penin A., Sogame Y., Cornette R., Tokumoto S., Miyata Y., Kikawada T., Gelfand M.S., Gusev O.* Cooption of heat shock regulatory system for anhydrobiosis in the sleeping chironomid *Polypedilum vanderplanki* // Proc. Natl. Acad. Sci. U S A. – 2018. – V. 115, No 10. – P. E2477–E2486. – doi: 10.1073/pnas.1719493115.

24. *Kozlova O., Cherkasov A., Przhiboro A., Shagimardanova E.* Complexity of expression control of HSP70 genes in extremophilic midges // BioNanoScience. – 2016. – V. 6, No 4. – P. 388–391. – doi: 10.1007/s12668-016-0256-3.

Поступила в редакцию
06.03.18

Козлова Ольга Сергеевна, аспирант кафедры биохимии и биотехнологии

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия
E-mail: *olga-sphinx@yandex.ru*

Абрамова Зинаида Ивановна, доктор биологических наук, профессор кафедры биохимии и биотехнологии

Казанский (Приволжский) федеральный университет
ул. Кремлевская, д. 18, г. Казань, 420008, Россия
E-mail: *ziabramova@mail.ru*

ISSN 2542-064X (Print)
ISSN 2500-218X (Online)

UCHENYE ZAPISKI KAZANSKOGO UNIVERSITETA. SERIYA ESTESTVENNYE NAUKI

(Proceedings of Kazan University. Natural Sciences Series)

2018, vol. 160, no. 2, pp. 214–226

Assembly of Anhydrobiotic Midge *Polypedilum vanderplanki* Genome Using Illumina and PacBio Data

O.S. Kozlova[*], *Z.I. Abramova*[**]

Kazan Federal University, Kazan, 420008 Russia
E-mail: [*]*olga-sphinx@yandex.ru*, [**]*ziabramova@mail.ru*

Received March 6, 2018

Abstract

A completely new version of assembly of the African anhydrobiotic midge *Polypedilum vanderplanki* genome derived by deep DNA sequencing of the Pv11 cell line has been discussed. The input data include paired-end and mate-paired reads with various insert sizes supplemented with ultra-long reads of Pacific Biosciences platform sequencing. We have shown that the resulting set of scaffolds has higher continuity and completeness metrics and, besides, can provide more correct predictions of coding sequences as compared to the previous assembly version, which has been demonstrated based on heat-shock proteins HSP20 and HSP70.

Keywords: *Polypedilum vanderplanki*, anhydrobiosis, DNA sequencing, genome assembly

Acknowledgments. The work is performed according to the Russian Government Program of Competitive Growth of Kazan Federal University.

Figure Captions

Fig. 1. Factor distribution histogram of 21-mers based on all Illumina data. Red line – approximation of the entire statistical model (erroneous *k*-mers and genomic *k*-mers), blue and green lines – approximation of the models for heterozygous and homozygous *k*-mers, respectively.

References

1. Cornette R, Kikawada T. The induction of anhydrobiosis in the sleeping chironomid: Current status of our knowledge. *IUBMB Life*, 2011, vol. 63, no. 6, pp. 419–429. doi: 10.1002/iub.463.
2. Nakahara Y, Watanabe M, Fujita A, Kanamori Y, Tanaka D, Iwata K, Furuki T, Sakurai M, Kikawada T, Okuda T. Effects of dehydration rate on physiological responses and survival after rehydration in larvae of the anhydrobiotic chironomid. *J. Insect Physiol.*, 2008, vol. 54, no. 8, pp. 1220–1225. doi: 10.1016/j.jinsphys.2008.05.007.

3. Watanabe M., Kikawada T, Okuda T. Increase of internal ion concentration triggers trehalose synthesis associated with cryptobiosis in larvae of *Polypedilum vanderplanki. J. Exp. Biol.*, 2003, vol. 206, pt. 13, pp. 2281–2286. doi: 10.1242/jeb.00418.

4. Ryabova A., Mukae K., Cherkasov A., Cornette R., Shagimardanova E., Sakashita T., Okuda T., Kikawada T., Gusev O. Genetic background of enhanced radioresistance in an anhydrobiotic insect: Transcriptional response to ionizing radiations and desiccation. *Extremophiles*, 2017, vol. 21, no. 1, pp. 109–120. doi: 10.1007/s00792-016-0888-9.

5. Gusev O., Suetsugu Y., Cornette R., Kawashima T., Logacheva M.D., Kondrashov A.S., Penin A.A., Hatanaka R., Kikuta S., Shimura S., Kanamori H., Katayose Y., Matsumoto T., Shagimardanova E., Alexeev D., Govorun V., Wisecaver J., Mikheyev A., Koyanagi R., Fujie M., Nishiyama T., Shigenobu S., Shibata T.F., Golygina V., Hasebe M., Okuda T., Satoh N., Kikawada T. Comparative genome sequencing reveals genomic signature of extreme desiccation tolerance in the anhydrobiotic midge. *Nat. Commun.*, 2014, vol. 5, art. 4784, pp. 1–9. doi: 10.1038/ncomms5784.

6. Eid J., Fehr A., Gray J., Luong K., Lyle J., Otto G., Peluso P., Rank D., Baybayan P., Bettman B., Bibillo A., Bjornson K., Chaudhuri B., Christians F., Cicero R., Clark S., Dalal R., Dewinter A., Dixon J., Foquet M., Gaertner A., Hardenbol P., Heiner C., Hester K., Holden D., Kearns G., Kong X., Kuse R., Lacroix Y., Lin S., Lundquist P., Ma C., Marks P., Maxham M., Murphy D., Park I., Pham T., Phillips M., Roy J., Sebra R., Shen G., Sorenson J., Tomaney A., Travers K., Trulson M., Vieceli J., Wegener J., Wu D., Yang A., Zaccarin D., Zhao P., Zhong F., Korlach J., Turner S. Real-time DNA sequencing from single polymerase molecules. *Science*, 2009, vol. 323, no. 5910, pp. 133–138. doi: 10.1126/science.1162986.

7. Bolger A.M., Lohse M., Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 2014, vol. 30, no. 15, pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.

8. O'Connell J., Schulz-Trieglaff O., Carlson E., Hims M.M., Gormley N.A., Cox A.J. NxTrim: Optimized trimming of Illumina mate pair reads. *Bioinformatics*, 2015, vol. 31, no. 12, pp. 2035–2037. doi: 10.1093/bioinformatics/btv057.

9. Chikhi R., Medvedev P. Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, 2014, vol. 30, no. 1, pp. 31–37. doi: 10.1093/bioinformatics/btt310.

10. Chin C.S., Alexander D.H., Marks P., Klammer A.A., Drake J., Heiner C., Clum A., Copeland A., Huddleston J., Eichler E.E., Turner S.W., Korlach J. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods*, 2013, vol. 10, no. 6, pp. 563–569. doi: 10.1038/nmeth.2474.

11. Ye Ch., Hill C.M., Wu Sh., Ruan J., Ma Zh. (Sam). DBG2OLC: Efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.*, 2016, vol. 6, art. 31900, pp. 1–9. doi: 10.1038/srep31900.

12. Kajitani R., Toshimoto K., Noguchi H., Toyoda A., Ogura Y., Okuno M., Yabana M., Harada M., Nagayasu E., Maruyama H., Kohara Y., Fujiyama A., Itoh T. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.*, 2014, vol. 24, no. 8, pp. 1384–1395. doi: 10.1101/gr.170720.113.

13. English A.C., Richards S., Han Y., Wang M., Vee V., Qu J., Qin X., Muzny D.M., Reid J.G., Worley K.C., Gibbs R.A. Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE*, 2012, vol. 7, no. 11, art. e47768, pp. 1–12. doi: 10.1371/journal.pone.0047768.

14. Wences A.H., Schatz M.C. Metassembler: Merging and optimizing de novo genome assemblies. *Genome Biol.*, 2015, vol. 16, art. 207, pp. 1–10. doi: 10.1186/s13059-015-0764-4.

15. Boetzer M., Henkel C.V., Jansen H.J., Butler D., Pirovano W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, 2011, vol. 27, no. 4, pp. 578–579. doi: 10.1093/bioinformatics/btq683.

16. Nadalin F., Vezzi F., Policriti A. GapFiller: A de novo assembly approach to fill the gap within paired reads. *BMC Bioinf.*, 2012, vol. 13, suppl. 14, art. S8, pp. 1–16. doi: 10.1186/1471-2105-13-S14-S8.

17. 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, no. 7, pp. 44–52. doi: 10.1038/nbt.1883.

18. Kent W.J. BLAT – the BLAST-like alignment tool. *Genome Res.*, 2002, vol. 12, no. 4, pp. 656–664. doi: 10.1101/gr.229202.
19. Xue W., Li J.T., Zhu Y.P., Hou G.Y., Kong X.F., Kuang Y.Y., Sun X.W. L_RNA_scaffolder: Scaffolding genomes with transcripts. *BMC Genomics*, 2013, vol. 14, art. 604, pp. 1–14. doi: 10.1186/1471-2164-14-604.
20. Gurevich A., Saveliev V., Vyahhi N., Tesler G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 2013, vol. 29, no. 8, pp. 1072–1075. doi: 10.1093/bioinformatics/btt086.
21. Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 2015, vol. 31, no. 19, pp. 3210–3212. doi: 10.1093/bioinformatics/btv351.
22. Li X., Waterman M.S. Estimating the repeat structure and length of DNA sequences using L-tuples. *Genome Res.*, 2003, vol. 13, no. 8, pp. 1916–1922. doi: 10.1101/gr.1251803.
23. Mazin P.V., Shagimardanova E., Kozlova O., Cherkasov A., Sutormin R., Stepanova V.V., Stupnikov A., Logacheva M., Penin A., Sogame Y., Cornette R., Tokumoto S., Miyata Y., Kikawada T., Gelfand M.S., Gusev O. Cooption of heat shock regulatory system for anhydrobiosis in the sleeping chironomid *Polypedilum vanderplanki*. *Proc Natl. Acad. Sci. U S A.*, 2018, vol. 115, no. 10, pp. E2477–E2486. doi: 10.1073/pnas.1719493115.
24. Kozlova O., Cherkasov A., Przhiboro A., Shagimardanova E. Complexity of expression control of HSP70 genes in extremophilic midges. *BioNanoScience*, 2016, vol. 6, no. 4, pp. 388–391. doi: 10.1007/s12668-016-0256-3.

***Для цитирования*:** *Козлова О.С., Абрамова З.И.* Сборка генома ангидробиотического насекомого *Polypedilum vanderplanki* с использованием данных Illumina и Pacbio // Учен. зап. Казан. ун-та. Сер. Естеств. науки. – 2018. – Т. 160, кн. 2. – С. 214–226.

***For citation*:** Kozlova O.S., Abramova Z.I. Assembly of anhydrobiotic midge *Polypedilum vanderplanki* genome using Illumina and PacBio data. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Estestvennye Nauki*, 2018, vol. 160, no. 2, pp. 214–226. (In Russian)