ГЕНЕТИКА, 2017, том 53, № 6, с. 641–650

ОБЗОРНЫЕ И ТЕОРЕТИЧЕСКИЕ СТАТЬИ

УДК 577.212

ПОЛНОГЕНОМНОЕ СЕКВЕНИРОВАНИЕ ГЕНОМОВ ЭУКАРИОТ: ОТ СЕКВЕНИРОВАНИЯ ФРАГМЕНТОВ ДНК К СБОРКЕ ГЕНОМА

© 2017 г. К. С. Задесенец[1, *], Н. И. Ершов[1], Н. Б. Рубцов[1, 2]

[1]*Федеральный исследовательский центр Институт цитологии и генетики Сибирского отделения Российской академии наук, Новосибирск 630090*

[2]*Новосибирский государственный университет, Новосибирск 630090*

**e-mail: kira_z@bionet.nsc.ru*

Поступила в редакцию 25.07.2016 г.

Стремительное развитие технологий секвенирования второго и даже третьего поколений сделало полногеномное секвенирование рутинной процедурой. Однако способы сборки полученных последовательностей и ее результаты требуют отдельного рассмотрения. Современные ассемблеры основаны на эвристических алгоритмах, что приводит к фрагментированной сборке генома, состоящей из скэффолдов и контигов различной длины, порядок локализации которых в хромосоме, а также их принадлежность к конкретной хромосоме зачастую остаются неизвестными. В связи с этим полученная сборка генома может рассматриваться только как его черновой вариант. Принципиальное улучшение качества и повышение надежности сборки драфта может быть достигнуто раздельным секвенированием элементов генома, отличающихся по своему размеру: хромосомы, хромосомные районы, клонированные в различных векторах фрагменты ДНК, а также использованием референсного генома, оптического картирования, Hi-C-технологии. Такой подход кроме упрощения сборки драфта генома позволит более точно выявлять численные и структурные вариации и аномалии геномов изучаемых видов. В данном обзоре обсуждаются основные технологии секвенирования и сборки генома *de novo*, а также различные подходы для улучшения качества уже существующих драфтов генома.

Ключевые слова: прочтение, контиг, скэффолд, граф де Брейна, картирование хромосом, методы, ДНК.

DOI: 10.7868/S0016675817050137

СЕКВЕНИРОВАНИЕ ДНК И ЕГО ЗНАЧЕНИЕ ДЛЯ СОВРЕМЕННОЙ БИОЛОГИИ

В настоящее время секвенирование ДНК стало не только ключевым методом исследований во многих областях современной биологии, но и предопределило открытие и дальнейшее развитие новых направлений. История секвенирования уходит в 50-е годы прошлого столетия, когда были разработаны методы, позволяющие определять последовательность аминокислот в полипептидной цепи, а расшифровка генетического кода позволила частично определять последовательность нуклеотидов транскрибируемой нуклеиновой кислоты. В конце 60-х был разработан метод секвенирования РНК [1], что позволило У. Фирсу с коллегами сначала секвенировать ген белка оболочки бактериофага MS2 [2], а затем и всю его ДНК [3]. Примерно в это же время были разработаны методы прямого секвенирования ДНК: “плюс–минус” метод [4], метод “терминации цепи” [5] и “метод химической деградации” [6]. В течение следующих десятилетий “секвенирование по Сэнгеру” было полностью автоматизировано: на замену электрофорезу в геле и радиоактивно меченым нуклеотидам пришли капиллярный электрофорез и нуклеотиды, конъюгированные с флуорохромами [7]. Также удалось увеличить длину прочтений фрагмента ДНК в одной реакции до 500–1000 пн.

Ограничение размера секвенируемого фрагмента ДНК было преодолено секвенированием перекрывающихся фрагментов. Логичным развитием этого подхода стал “метод дробовика” (shotgun-sequencing), основанный на случайной физической или химической фрагментации ДНК-матрицы, клонировании полученных фрагментов (~2–3 тпн) и их последующем секвенировании. Вследствие случайной фрагментации полученные фрагменты перекрывали друг друга так, что при многократном покрытии анализируемого протяженного фрагмента ДНК возникала возможность его сборки. Этот подход был успешно использован более 20 лет назад при секвенировании первого бактериального генома: геном *Haemophilus influenza* был собран из ~24 × 10^3 прочтений фрагментов ДНК длиной ~460 пн [8]. При анализе геномов прокариот, содержащих небольшое количество повторов, полная последо-

вательность генома может быть собрана в результате анализа секвенированных фрагментов при относительно небольшом его покрытии (7–10×).

Вероятность ошибки при "секвенировании по Сэнгеру" варьирует от 10^{-5} до 10^{-4} при размере прочтений около 1000 пн. Несмотря на то, что секвенирование по Сэнгеру до сих пор считается "золотым стандартом" качества и широко используется в разнообразных исследованиях, оно имеет ряд существенных недостатков: высокую удельную стоимость и низкую производительность [9]. Технологии секвенирования нового или второго поколения (next-generation sequencing, **NGS**) не имеют этих недостатков, однако уступают по точности и длине прочтений. В настоящее время наиболее широко распространены высокопроизводительные платформы GL FLX Titanium/GS Junior (Roche), SoLiD/Ion Torrent PGM (Applied BioSystems), Genome Analyzer/HiSeq 2000/MiSeq (Illumina) [10, 11]. Благодаря их высокой производительности стала возможным реализация крупномасштабных программ и проектов, направленных на секвенирование большого числа видов про- и эукариот: Genome 10K (для позвоночных), i5K (для насекомых), 959 Nematode Genomes (для круглых червей), 1 KP (для растений), 3M (миллион научно- и экономически-значимых видов + + миллион геномов человека + миллион метагеномов и микробиомов) [10, 12–14]. Однако вопрос, насколько успешно идет реализация этих проектов, что представляют собой получаемые сборки геномов, требует отдельного рассмотрения.

DRAFT GENOME SEQUENCE – ЧЕРНОВАЯ ГЕНОМНАЯ ПОСЛЕДОВАТЕЛЬНОСТЬ

Прорыв в секвенировании во многом связан с масштабным проектом по секвенированию генома человека [15, 16], который выполнялся 20 группами исследователей в течение 15 лет. При секвенировании с использованием как классического метода Сэнгера, так и метода дробовика было получено и использовано для первичной сборки черновой геномной последовательности около 30×10^6 прочтений длиной до 800 пн [15, 16]. При сборке и ее верификации были использованы физические карты высокого разрешения, созданные с помощью различных подходов: гибридизация соматических клеток, панели радиационных гибридов, библиотеки бактериальных искусственных хромосом (BACs), гибридизация нуклеиновых кислот *in situ*, секвенирование клонированных последовательностей [15, 17–20].

Первый драфт генома человека (draft genome sequence) был анонсирован сотрудниками частной корпорации "Celera Genomics" в июне 2000 г., но детали работы были опубликованы только в феврале 2001 г. [16]. Практически одновременно были опубликованы результаты работы международного консорциума (International Human Genome Sequencing Consortium) [15]. В феврале 2001 г. в процессе подготовки совместных публикаций были выпущены пресс-релизы, в которых было заявлено, что проект был завершен обеими группами. Однако представленные драфты покрывали только 83% генома (~90% эухроматиновых районов хромосом). Причем порядок и ориентация значительной части полученных контигов, разделенных 150 тыс. брешей, оставались неустановленными. Последующие годы были ознаменованы регулярным выходом новых, улучшенных драфтов генома человека. Сегодня последний вариант (GRCh38.p7) представляет собой достаточно хорошую версию некого абстрактного генома, с небольшим числом брешей, с "белыми пятнами" в С-позитивных районах хромосом и без учета огромного генетического разнообразия, характерного для человека.

Практически все работы по секвенированию геномов эукариот в настоящее время находятся на стадии сборки черновой геномной последовательности. Согласно определению Е.Д. Свердлова, "черновая геномная последовательность представляет собой незавершенную сборку последовательности генома, в которой отсутствует ряд сегментов, не установлены окончательно порядок и ориентация участков последовательности, и существуют ошибки в последовательностях нуклеотидов" [21]. Из секвенированных геномов модельных видов эукариот действительно полная завершенная сборка последовательностей генома сделана, пожалуй, только для *C. elegans* и *S. cerevisiae* [22, 23]. Тем не менее, учитывая огромный полиморфизм по однонуклеотидным заменам (SNPs, single nucleotide polymorphism), вариации по числу копий участков ДНК (CNVs, copy number variations), корректно говорить об окончательной сборке эухроматиновой части генома конкретного индивида и описании генетического полиморфизма в определенных группах. Но даже такой подход кажется излишне оптимистичным, так как проблема сборки последовательностей нуклеотидов индивидуальных хромосом, составляющих пару гомологов, остается практически нерешенной. Секвенирование гаплоидных геномов [24] едва ли можно считать решением проблемы.

СОВРЕМЕННЫЕ ПОДХОДЫ И МЕТОДЫ ПОЛНОГЕНОМНОГО СЕКВЕНИРОВАНИЯ И СБОРКИ ЧЕРНОВЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

Сегодня многократное покрытие генома прочтениями длиной до нескольких сот нуклеотидов выполняется многочисленными специализированными центрами. Несмотря на стремительный прогресс компьютерной техники и программного

обеспечения, сборка драфта остается большой проблемой. Основной причиной этого являются повторенные последовательности. Они осложняют сборку генома даже у ряда прокариот [25]. Тем не менее высокая производительность NGS стимулировала разработку методов построения драфта генома путем прямой сборки полученных прочтений. Обычно исходным материалом для сборки генома оказывается пул из 10^6–10^{12} прочтений, длиной от 85 пн (SOLiD 5500 xl) до 700 пн (454 GS FLX Titanium system) [10, 11]. (Технология PacBio или Moleculo будет рассмотрена ниже.) Для прохождения повтора размер анализируемого фрагмента ДНК должен превышать размер повтора. Частично это может быть достигнуто с использованием метода парных прочтений (PE, paired-end) и методом спаренных концов (MP, mate-pair), обеспечивающих секвенирование концов фрагментов ДНК размером 200–500 пн и 2–20 тпн соответственно. Кроме относительной локализации пары прочтений эти методы определяют их взаимную ориентацию, позволяют выявлять ошибки секвенирования, структурные вариации генома, связывать отдаленные фрагменты сборки, увеличивая размер контигов [26]. Однако если проблема прохождения таких повторов, как SINEs и LINEs, была в принципе решена, то проблема с кластерами повторенных последовательностей и обогащенные повторами районы остались практически неразрешенными (например, С-позитивные районы хромосом).

Для сборки драфтов были созданы специализированные программы (ассемблеры), основанные на эвристических или приближенных алгоритмах и дающие не единую консенсусную последовательность, а набор контигов и скэффолдов. Можно выделить несколько типов алгоритмов ассемблеров: 1) “жадный” алгоритм (greedy-extension), используемый в ассемблерах SSAKE, SHARCGS; 2) граф перекрытий (OLC, Overlap-Layout-Consensus), используемый преимущественно для сборки длинных прочтений ассемблерами Celera, Arachne, Newbler; 3) граф де Брейна (DBG, de Bruijn graph), используемый для сборки коротких прочтений ассемблерами SOAPdenovo, Velvet, IDBA, SPAdes [27–31]. Иерархичные или гибридные ассемблеры Atlas и MaSuRCA [25, 28] объединяют алгоритмы (2) и (3), обеспечивая более длинные контиги с меньшим числом ошибок.

Ранее для исправления ошибок, возникших при секвенировании, использовался алгоритм Смита–Ватермана, допускающий небольшие различия в перекрытиях прочтений [32]. Он оказался наиболее эффективным при сборке простых геномов, но не работал при наличии в геноме повторов, длина которых превышает размер прочтения. Ассемблеры, использующие “жадный” алгоритм, объединяют все копии такого повтора в один фрагмент. Также они дают сбой при наличии повтора на конце фрагмента, давая “химерные” контиги. Уже в начале 90-х годов при сборке бактериальных геномов возникла необходимость в более сложных и совершенных методах сборки драфтов.

Как правило, для сборок контигов из коротких прочтений (<50 пн) используются алгоритмы на основе поиска Эйлерова пути в графе де Брейна, построенного по k-мерам этих прочтений [33]. Для этого прочтения сначала разбиваются на короткие фрагменты из k нуклеотидов (или k-меры), перекрывающиеся на $(k - 1)$ нуклеотид, с последующим конструированием графа. Стоит отметить, что значение k подбирается индивидуально для каждого случая, исходя как из длины прочтений (k не должен превышать прочтение минимальной длины, например, $k = 21$ для прочтений длиной 25 пн) и глубины секвенирования, так и особенностей анализируемого генома. На сегодняшний день реализован подход для подбора этого параметра с помощью программного алгоритма [34]. Ассемблеры IDBA и SPAdes используют не одно фиксированное значение k, а позволяют работать с несколькими значениями k, приводя, таким образом, к увеличению длины контигов, а главное – к разрешению проблемы гэпов и районов, обогащенных повторами. Например, ассемблер SPAdes позволяет эффективно решить проблему неоднородной представленности фрагментов ДНК после проведения полногеномной амплификации (MDA, multiple displacement amplification) из одной бактериальной клетки [31]. Стоит отметить, что упомянутые ассемблеры могут быть эффективно использованы лишь для работы с небольшими бактериальными геномами.

Для более длинных прочтений эффективнее ассемблер, основанный на графе перекрытий [35]. Хотя в теории размер графа де Брейна зависит только от размера генома и не должен зависеть от количества прочтений, из-за ошибок секвенирования в графах создаются “узлы”, а увеличение количества прочтений неизбежно увеличивает и размер графа де Брейна. На первом этапе сборки генома человека *de novo* из коротких прочтений (до конструирования графа де Брейна) при коррекции ошибок количество k-меров ($k = 25$) уменьшилось с 14.6×10^{12} до 5×10^{12} [36]. Коррекция ошибок заключалась в замене k-меров, встречаемых менее трех раз, k-мерами с большей частотой [32].

Стоит отметить, что изменение параметров настройки ассемблеров может приводить к разным результатам при использовании одних и тех же первичных данных, и использование одного и того же ассемблера с идентичными настройками может давать разные результаты при обработке данных, полученных в независимых экспериментах [37]. На низкое качество драфтов, полученных только на базе данных NGS, указывает результат, выполненной *de novo* с помощью ассемблера

SOAP сборки генома человека. В сборке отсутствовали около 2.3 тыс. кодирующих экзонов, 99% сегментных дупликаций (~5% генома человека), протяженные повторенные последовательности. В итоге полученный *de novo* драфт генома оказался на 16.2% меньше современного [38]. Выбор ассемблера часто обуславливается поставленными задачами и особенностями данных NGS: важны уровень покрытия генома, доля в геноме уникальных и повторенных последовательностей, частота ошибок секвенирования, необходимость получения максимально длинных скэффолдов, достижение высокого разрешения описания гаплотипа и т.д. Некоторые ассемблеры требуют определенный дизайн библиотек [39]. Сравнение результатов сборки драфта с использованием разных методов и алгоритмов сборки указывает на важность привлечения дополнительных независимых от NGS данных.

Из-за ограниченных возможностей сборки прочтений секвенирование генома *de novo* большинства видов обычно завершается созданием драфта генома, состоящего из скэффолдов и контигов различной длины. Их порядок, локализация в хромосоме и даже принадлежность к конкретной хромосоме часто остаются неизвестными. Сборка генома, основанная на эвристических алгоритмах, может рассматриваться только как драфт генома [28].

Увеличение размера прочтений может существенно улучшить качество сборки драфта. Последние поколения технологий секвенирования характеризуются увеличением их размера [25, 35, 40]. Платформы секвенирования третьего [41–43] и четвертого поколения [40, 44] имеют перед своими предшественниками значительное преимущество.

Технология определения последовательности нуклеотидов одиночной молекулы ДНК в режиме реального времени (SMRT, single molecule real-time sequencing) позволяет получать прочтения длиной в среднем ~10 тпн, с максимальной длиной до 60 тпн, но частота ошибок составляет около 15% [45]. В современных ассемблерах не предусмотрена корректировка таких ошибок [46]. Было предложено корректировать ошибки такого секвенирования покрытием длинных прочтений короткими прочтениями высокого качества, полученными из того же образца [47], или за счет увеличения уровня покрытия генома [45]. Оба предложенных подхода позволяют исправить до 99% ошибок в прочтениях, давая точность консенсусной последовательности до 99.999% при проведении более чем 50-кратном покрытии. К сожалению, такое покрытие проблематично при секвенировании больших геномов. Сегодня SMRT позволяет провести качественную сборку некоторых бактериальных геномов уровня “один контиг – одна хромосома” [25, 45, 47]. При секвенировании более крупных геномов предпочтительнее использовать комбинированный подход, использующий разные стратегии секвенирования (например, NGS и SMRT), и создание разных библиотек [48]. Кроме того, SMRT позволяет картировать паттерн метилирования даже в высокоповторенных районах генома за счет анализа изменения кинетики полимеразной реакции, обусловленного модификацией нуклеотидов [41, 49].

Дополнительные возможности при сборке драфта дает использование референсного генома; например, с его помощью были получены большинство драфтов индивидуальных генов человека. При сборке драфта генома нового вида использование в качестве референсного генома драфта генома близкородственного вида позволяет решить часть проблем и значительно уменьшить его фрагментированность. Так, алгоритм RACA (reference-assisted chromosome assembly) позволяет собирать *de novo* скэффолды, отображающие организацию генома на уровне хромосом [50]. Успешность его применения зависит от качества драфтов референсного генома и геномов видов “внешней” группы. Например, при реконструкции хромосом тибетской антилопы *Pantholops hodgsonii* в качестве референсного генома использовался геном *Bos taurus*, а в качестве генома “внешней” группы – геном человека [50]. При наличии хорошо собранного референсного генома удается получить протяженные скэффолды даже из очень коротких прочтений [28].

Эффективность использования драфта генома в качестве референсного зависит от его качества и сходства с геномом изучаемого вида. Так как эволюционно значимые перестройки генома часто происходят в районах, обогащенных повторами, т.е. наиболее проблематичных при сборке драфтов, то риск собрать драфт генома, более похожий на геном референсного вида, чем на геном изучаемого, достаточно велик. Причем часто драфты геномов получают из результатов одного или нескольких индивидов, что не дает представления о внутривидовом полиморфизме [51]. Например, секвенирование генома большой панды – это секвенирование генома самки из китайского центра Чэнду [52], а секвенирование генома собаки выполнялось из ДНК представителя одной породы [53]. При использовании результатов полногеномного секвенирования в дальнейших исследованиях необходимо четко осознавать, что представляют собой драфты геномов секвенированных *de novo* видов, особенно собранных из прочтений NGS с использованием референсного генома одного из индивидов “близкородственного вида”. В них полностью отсутствует информация не только о полиморфных маркерах, таких как SNPs и CNVs, но и о целом ряде структурных перестроек, отличающих секвенируемый геном от референсного

[32]. Отметим также, что значительно осложняют сборку драфтов геномов следы относительно недавней полной или частичной дупликации генома, имевшей место в эволюции изучаемого вида [51].

ВЕРИФИКАЦИЯ И УЛУЧШЕНИЕ КАЧЕСТВА СБОРКИ ГЕНОМА

Финишированная сборка генома представлена менее чем для 35% из около 1800 видов, вовлеченных в полногеномное секвенирование, и в среднем она составляет не более 80% генома [25, 38]. Значения N50/NG50 и размер контигов или скэффолдов могут значительно колебаться [27, 28, 53], что зависит как применяемых методов секвенирования, ассемблеров, так и особенностей генома. Значения основных метрик сборки генома следует интерпретировать с осторожностью. Большое значение N50 может не свидетельствовать о высокой надежности сборки, и для верификации сборки следует привлекать альтернативные методы анализа генома [28].

Надежность сборки генома во многом зависит от выбранной стратегии секвенирования, наличия референсного генома, выбранного ассемблера и самого генома: его размера, насыщенности повторами, сегментными дупликациями, следов частичной или полной дупликации генома при эволюции вида. Например, сборка на основании данных SMRT, схожих по размеру, но содержащих разное количество повторов, геномов *Arabidopsis thaliana* (120 Мпн) и *Drosophila melanogaster* (130 Мпн) дала разные результаты. У дрозофилы повторы длиннее и их больше, поэтому при средней длине прочтения менее 3600 пн сборка генома арабидопсиса содержала меньшее количество пробелов и разрывов [46].

Одним из способов улучшения сборки и ее верификации является оптическое картирование, представляющее собой метод создания рестрикционной карты высокого разрешения. Длинные молекулы ДНК прикрепляются к стеклу и слегка натягиваются за счет электростатических взаимодействий. После обработки рестриктазами натяжение нитей ДНК ослабляется. В результате свободные концы разрывов скручиваются, и между ними образуются гэпы размером около 1–2 мкм, что позволяет определять расстояние между сайтами рестрикции [54]. Метод ранее успешно использовался для сравнения структуры бактериальных геномов, их сборки и корректировки. Оптическое картирование может быть использовано и для улучшения качества сборки за счет разрешения длинных повторов, заполнения пробелов в сборке и для сборки генома *de novo* [55]. Первым геномом позвоночного, сборка которого была существенно улучшена с использованием этого подхода, является геном мыши [56]. Геном гриба рейши *Ganoderma lucidum* и геном домашней козы *Capra hircus* оказались первыми геномами, собранными *de novo* с использованием оптического картирования [57, 58]. Построение оптической карты не требует наличия референсной последовательности, позволяет детектировать делеции и инсерции, давая информацию об их размере. К недостаткам данного способа картирования генома относится низкое разрешение и проблематичность использования при работе с большими геномами. Работа с большими геномами предъявляет очень высокие требования к программному и техническому обеспечению. Однако в сочетании с NGS технологиями оптическое картирование дает ценный вклад в верификацию и улучшение качества сборки генома [55, 59].

Альтернативным подходом для изучения структурных вариаций генома и его сборки *de novo* является другой вариант оптического/геномного картирования генома – BioNano IrysSystem (BioNano Genomics), основанный на микрофлюидной технологии. Он заключается в введении меченых нуклеотидов в ники высокомолекулярной ДНК (10^5–10^6 пн), ее окрашивании и последующем расправлении фрагментов в наноканалах диаметром 45 нм (скрученная нить ДНК расправляется при проходе через градиент микро- и наноструктур) [60]. Данная технология позволяет получать оптические карты со средним размером фрагмента 225 тпн [61]. Для улучшения качества сборки производят выравнивание оптических контигов на референсный геном. При сборке генома *de novo* оптические “прочтения” сначала разбивают на *k*-меры и с помощью графов де Брейна собирают консенсусную последовательность генома. К недостаткам данного метода можно отнести меньшую длину контигов из-за наличия фрагильных сайтов в местах близкой локализации ников на противоположных нитях ДНК [62].

Проблемой является также определение хромосомной локализации контигов. Даже для протяженных контигов и скэффолдов ее можно определить, если известна локализация какой-либо уникальной последовательности, входящей в их состав, либо созданием ДНК-зонда и проведением гибридизации *in situ* [63]. Для соотнесения контигов и скэффолдов с генетической картой необходимо проведение скрещиваний и анализа потомства, несущего маркеры как интересующих контигов и скэффолдов, так и групп сцепления. К сожалению, в большинстве случаев такие исследования не проводятся, а “привязка” контигов и скэффолдов к конкретным хромосомам осуществляется при наличии референсного генома, оставляя открытым вопрос о степени его реорганизации относительно генома изучаемого вида. Создание карт сцепления, картирование с помощью панели радиационных гибридов и гибридизации соответствующих ДНК-зондов *in situ* является скорее исключением из правил, как и создание

гибридных карт высокого разрешения, построенных с использованием всего комплекса существующих методов [64, 65].

Рассматривая методы, позволяющие верифицировать и улучшить качество геномных драфтов, необходимо упомянуть методы Hi-C (high-resolution genome conformation capture), базирующиеся на анализе пространственной конформации всего генома [66, 67]. В их основе лежит постулат о том, что участки ДНК из одного района хромосомы в интерфазном ядре контактируют друг с другом чаще, чем ДНК разных хромосом или удаленные участки ДНК одной хромосомы. Использование базы данных, полученных с использованием Hi-C-технологий, позволяет не только провести проверку сборки прочтений, полученных методами NGS, но также определить порядок и ориентацию фрагментов генома относительно центромер, выявить и определить состав маркерных хромосом, хромосомные транслокации [66]. Для обработки данных Hi-C было разработано программное обеспечение LACHESIS (ligating adjacent chromatin enables scaffolding *in situ*), с помощью которого сначала производят кластеризацию контигов или скэффолдов в хромосомные группы, затем определяют порядок внутри каждой из них и ориентацию контигов или скэффолдов.

Несмотря на большие возможности новых методов секвенирования и сборки драфтов, разработка и внедрение в практику новых высокотехнологичных методов секвенирования, способных обойти существующие ограничения, а также пригодных и для верификации уже существующих сборок геномов, остается актуальной задачей.

ПЕРСПЕКТИВЫ УПРОЩЕНИЯ СБОРКИ ДРАФТА ГЕНОМА И ПОВЫШЕНИЯ ЕГО КАЧЕСТВА

Рассмотренное выше сравнение драфта генома человека, полученного *de novo* на базе данных NGS с использованием современного ассемблера SOAP, с современным драфтом [38] показало насущную необходимость привлечения дополнительных методов для повышения качества и верификации драфтов геномов новых видов. Основная идея заключается в разбиении генома на элементы, отличающиеся по размеру: от отдельных хромосом и их фрагментов до клонированных фрагментов ДНК. На этом пути создание ДНК-библиотек индивидуальных хромосом и их районов является важным шагом. К сожалению, использование хромосомного сортинга ограничено потребностью в культурах активно пролиферирующих клеток, получение которых зачастую практически невыполнимая задача. Это справедливо и для соматических гибридов, содержащих целые хромосомы изучаемого вида.

Существенно проще и технологичнее получение и секвенирование микродиссекционных ДНК-библиотек [68]. Однако такие ДНК-библиотеки содержат не только часть ДНК изолированной хромосомы или ее района, а также часто содержат постороннюю ДНК из биологического материала, использованного для производства компонентов необходимых реакционных смесей. Наш опыт получения и секвенирования таких микродиссекционных ДНК-библиотек, полученных, как правило, из десяти копий метафазных хромосом, показал, что они покрывают от 10 до 50% анализируемого района генома и могут содержать до 50% контаминирующих фрагментов ДНК.

К сожалению, абсолютная элиминация следовых количеств ДНК “хозяина” фермента практически невозможна [69], а высокая эффективность сиквенс-независимой амплификации ДНК, используемой для получения микродиссекционных ДНК-библиотек, приводит к наработке этой ДНК. В микродиссекционной ДНК-библиотеке могут присутствовать также побочные продукты ПЦР, возникшие вследствие отжига друг на друга праймеров.

Отдельно взятые результаты секвенирования микродиссекционных ДНК-библиотек не могут быть использованы для построения драфта секвенируемого *de novo* генома, но в комбинации с данными NGS по секвенированию всего генома они могут выявить большинство контигов, содержащих уникальные последовательности диссектированной хромосомы. Наличие в микродиссекционной ДНК-библиотеке фрагментов, гомологичных уникальным последовательностям, содержащимся в контигах, будет указывать на их принадлежность данной хромосоме или ее району. Этот подход является перспективным, так как не требует наличия клеточных культур и позволяет получать ДНК-библиотеки из относительно небольших хромосомных районов. Возможно, он также позволит частично решить проблему сегментных дупликаций и детектировать следы частичной или полной дупликации генома, имевшей место в эволюции анализируемого вида.

Несомненным преимуществом микродиссекционных ДНК-библиотек является достаточно высокая технологичность их получения и легкая проверка их качества проведением гибридизации *in situ*. Критичным для реализации этого подхода является получение цитологических препаратов митотических или мейотических хромосом и их идентификация. Наш опыт показал возможность использования описанного подхода для высокоразрешающего анализа хромосомных перестроек у человека и привязки контигов к конкретным хромосомам секвенируемых *de novo* видов. Секвенирование микродиссекционной ДНК-библиотеки малой сверхчисленной хромосомы человека

позволило, как минимум на порядок, повысить точность в определении точки разрыва, имевшего место при ее возникновении. Результаты секвенирования микродиссекционной ДНК-библиотеки первой хромосомы *Macrostomum lignano* позволило приступить к поиску контигов, входящих в состав этой хромосомы.

Дополнительные возможности улучшения качества сборки драфта генома возникают в случае выявления у представителей секвенируемого вида численных и структурных хромосомных аномалий. Для их использования критичным является надежность их описания. Качественный драфт генома человека и количественная оценка результатов секвенирования NGS позволяют выявлять анеуплоидии хромосом даже в случае, если они имеют место лишь в небольшом проценте клеток. Более того, такие нарушения генома плода могут быть обнаружены в результате секвенирования внеклеточной ДНК, выделенной из периферической крови матери [70]. При выявлении и детальном описании в кариотипе особи секвенируемого вида хромосомных аномалий, приводящих к нарушению баланса генов, решение обратной задачи позволит определить хромосомную или субхромосомную локализацию последовательностей, доля которых оказалась измененной по сравнению с той, которая была определена у особей с нормальным кариотипом. Такое исследование требует секвенирования с высоким уровнем покрытия (70×). Кроме того, необходимо учитывать возможность присутствия в сравниваемых геномах вариаций числа копий определенных последовательностей как варианта нормального полиморфизма. Основным препятствием в использовании такого подхода являются сложности поиска особей-носителей хромосомных аберраций, а также их детальное описание. Однако у некоторых видов частота хромосомных аномалий у особей лабораторных линий удивительно высока. Примером может служить модельный объект для изучения процессов старения и регенерации свободноживущий плоский червь *M. lignano* [71]. В ходе кариотипирования имеющихся линий и культур *M. lignano* было установлено, что некоторые лабораторные линии практически содержат особи с трисомиями и тетрасомиями по первой хромосоме [72]. На первый взгляд, использование анеуплоидных особей в исследованиях по секвенированию геномов эукариот кажется экзотическим вариантом, не имеющим большой перспективы. Однако у видов, в эволюции которых относительно недавно прошла полная или частичная дупликация генома, можно ожидать достаточно высокую частоту встречаемости анеуплоидных форм. Возможно, геном *M. lignano* представляет собой результат подобной недавно произошедшей дупликации генома. Использование такого подхода при секвенировании генома у этих видов позволит получить дополнительную информацию, которая не только упростит черновую сборку генома, но и позволит получить более надежный результат даже при оценке его размера.

ЗАКЛЮЧЕНИЕ

Развитие современных NGS технологий стимулировало многочисленные исследования по изучению геномов видов, относящихся к самым разным таксонам. Сравнительный анализ геномов несомненно расширит существующие представления о принципах их организации и механизмах эволюции. Современные возможности NGS технологий и биоинформатического анализа полученных данных позволяют проводить практически на конвейере полногеномное секвенирование эукариотических организмов. Однако необходимо признать, что информативность и надежность полученных драфтов генома существенно различаются, варьируя от практически бессмысленного массива коротких контигов до драфта, достаточно полно описывающего организацию генома. Эти различия зависят как от особенностей изучаемого генома, так и от наличия дополнительной информации, полученной другими методами.

Преимущественное использование NGS технологий на первых этапах таких исследований обусловило получение драфтов, представляющих собой набор относительно коротких контигов и скэффолдов, разделенных большим числом гэпов. На этом этапе из анализа практически полностью выпадали районы, обогащенные повторами, сегментные дупликации. В значительной степени были потеряны участки геномов, представляющие следы частичной или полной дупликации генома. Возможности сравнительного анализа геномов на этой стадии исследований были ограничены изучением особенностей организации небольших участков генома, содержащих преимущественно уникальные последовательности ДНК.

Более качественное и детальное секвенирование геномов за счет увеличения размера прочтений, применения оптического картирования, фрагментов ДНК, клонированных в YACs и BACs, Hi-C-технологий, широкого использования референсных геномов принципиально изменило ситуацию: увеличился размер контигов и скэффолдов, уменьшилось число гэпов. Это позволило приступить к решению новых задач, таких как выявление и изучение организации горячих точек эволюционно значимых хромосомных перестроек. Тем не менее приходится признать, что и сегодня драфты геномов изучаемых видов в значительной степени зависят от использованного комплекса методов и выбранного референсного генома. Причем важным является не только качество драфта референсного генома, но и степень

его дивергенции относительно секвенируемого *de novo* генома вида. Очевидно, что привлечение в исследования по полногеномному секвенированию новых методов и подходов, облегчающих черновую сборку и ее верификацию, является актуальной задачей. Такие подходы могут включать секвенирование индивидуальных хромосом и их районов. Кроме упрощения черновой сборки и ее верификации, они вместе с гибридизацией *in situ* клонированных и секвенированных фрагментов ДНК позволят определять хромосомную локализацию собранных контигов и скэффолдов.

Необходимо отметить огромное значение для формирования правильных представлений о принципах структурно-функциональной организации генома эукариот изучения внутривидового геномного разнообразия. Представления о его уровне дают исследования по секвенированию индивидуальных геномов человека. В настоящее время описаны миллионы SNPs, тысячи примеров CNVs, огромное число хромосомных перестроек. В ряде случаев выявлена их роль в формировании различных патологий или предрасположенности к некоторым заболеваниям, однако в большинстве случаев они, вероятно, могут быть отнесены к вариантам полиморфизма, значение которого остается невыясненным. Уже в настоящее время ведутся многочисленные исследования, посвященные выявлению ассоциаций между различными вариантами SNPs и CNVs, с одной стороны, и формированием различных признаков в фенотипе человека – с другой. В эти исследования вовлечены сотни тысяч геномов человека, миллионы SNPs, но предсказание многих признаков фенотипа по результатам секвенирования генома человека до сих пор уступают прогнозам, основанным на показателях этих признаков у его родителей. Возможно, что значение комбинаторики SNPs намного перекрывает влияние отдельных вариантов, и даже современные возможности биоинформатического анализа оказываются далеки от того, что позволило бы просчитать значение различных комбинаций.

Работа поддержана Российским фондом фундаментальных исследований (РФФИ) (грант № 16-34-60027 мол_а_дк).

СПИСОК ЛИТЕРАТУРЫ

1. *Sanger F., Brownlee G.G., Barrell B.G.* A two-dimensional fractionation procedure for radioactive nucleotides // J. Mol. Biol. 1965. V. 13. № 2. P. 373–398.
2. *Jou W.M., Haegeman G., Ysebaert M., Fiers W.* Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein // Nature. 1972. V. 237. № 5350. P. 82–88. doi 10.1038/237082a0
3. *Fiers W., Contreras R., Duerinck F. et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene // Nature. 1976. V. 260. № 5551. P. 500–507. doi 10.1038/260500a0
4. *Sanger F., Coulson A.R.* A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase // J. Mol. Biol. 1975. V. 94. P. 444–448. doi 10.1016/0022-2836(75)90213-2
5. *Sanger F., Nicklen S., Coulson A.R.* DNA sequencing with chain-terminating inhibitors // Proc. Natl Acad. Sci. USA. 1977. V. 74. № 12. P. 5463–5467.
6. *Maxam A.M., Gilbert W.* A new method of sequencing DNA // Proc. Natl Acad. Sci. USA. 1977. V. 74. № 2. P. 560–564.
7. *Smith L.M., Sanders J.Z., Kaiser R.J. et al.* Fluorescence detection in automated DNA sequence analysis // Nature. 1986. V. 321. № 6071. P. 674–679. doi 10.1038/321674a0
8. *Fleischmann R.D., Adams M.D., White O. et al.* Whole-genome random sequencing and assembly of *Haemophilus influenzae* // Science. 1995. V. 269. № 5223. P. 496–512.
9. *Kircher M., Kelso J.* High-throughput DNA sequencing – concepts and limitations. // Bioessays. 2010. V. 32. № 6. P. 524–536. doi 10.1002/bies.200900181
10. *Liu L., Li Y., Li S. et al.* Comparison of next-generation sequencing systems. // J. Biomed. Biotechnol. 2012. V. 2012. Article ID 251364. doi 10.1155/2012/251364
11. *Heather J.M., Chain B.* The sequence of sequencers: the history of sequencing DNA // Genomics. 2016. V. 107. № 1. P. 1–8. doi 10.1016/j.ygeno.2015.11.003
12. *Haussler D., O'Brien S.J., Ryder O.A. et al.* Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species // J. Hered. 2009. V. 100. № 6. P. 659–674. doi 10.1093/jhered/esp086
13. *Kumar S., Schiffer P.H., Blaxter M.* 959 Nematode Genomes: a semantic wiki for coordinating sequencing projects // Nucl. Acids Res. 2012. V. 40. D1295–D1300. doi 10.1093/nar/gkr826
14. 5K Consortium. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment // J. Hered. 2013. V. 104. № 5. P. 595–600. doi 10.1093/jhered/est050
15. *Lander E.S., Linton L.M., Birren B. et al.* Initial sequencing and analysis of the human genome // Nature. 2001. V. 409. P. 860–921. doi 10.1038/35057062
16. *Venter J.C., Adams M.D., Myers E.W. et al.* The sequence of the human genome // Science. 2001. V. 291. № 5507. P. 1304–1351. doi 10.1126/science.1058040
17. *Ruddle F.H., Creagan R.P.* Parasexual approaches to the genetics of man // Ann. Rev. Genet. 1975. V. 9. № 1. P. 407–486. doi 10.1146/annurev.ge.09.120175.002203
18. *Fan Y., Davis L.M., Shows T.B.* Mapping small DNA sequences by fluorescence *in situ* hybridization directly on banded metaphase chromosomes // Proc. Natl Acad. Sci. USA. 1990. V. 87. № 16. P. 6223–6227. doi 10.1073/pnas.87.16.6223
19. *Gyapay G., Schmitt K., Fizames C. et al.* A radiation hybrid map of the human genome // Human Mol. Genet. 1996. V. 5. № 3. P. 339–346.
20. *Stewart E.A., McKusick K.B., Aggarwal A. et al.* An STS-based radiation hybrid map of the human genome // Genome Res. 1997. V. 7. № 5. P. 422–433.
21. *Свердлов Е.Д.* Взгляд на жизнь через окно генома. М.: Наука, 2009. Т. 1. 592 с.

22. *Engel S.R., Dietrich F.S., Fisk D.G. et al.* The reference genome sequence of *Saccharomyces cerevisiae*: then and now // G3 (Bethesda). 2014. V. 4. № 3. P. 389–398. doi 10.1534/g3.113.008995
23. *Hillier L.W., Coulson A., Murray J.I. et al.* Genomics in *C. elegans*: so many genes, such a little worm // Genome Res. 2005. V. 15. № 12. P. 1651–1660. doi 10.1101/gr.3729105
24. *Steinberg K.M., Schneider V.A., Graves-Lindsay T.A. et al.* Single haplotype assembly of the human genome from a hydatidiform mole // Genome Res. 2014. V. 24. № 12. P. 2066–2076. doi 10.1101/gr.180893.114
25. *Koren S., Phillipy A.M.* One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly // Curr. Opin. Microbiol. 2015. V. 23. P. 110–120. doi 10.1016/j.mib.2014.11.014
26. *Nagarajan N., Pop M.* Sequence assembly demystified // Nature Rev.: Genetics. 2013. V. 14. № 13. P. 157–167. doi 10.1038/nrg3367
27. *Li Z., Chen Y., Mu D. et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph // Brief. Funct. Genomics. 2012. V. 11. № 1. P. 25–37. doi 10.1093/bfgp/elr035
28. *Ekblom R., Wolf J.B.W.* A field guide to whole-genome sequencing, assembly and annotation // Evol. Appl. 2014. V. 7. № 9. P. 1026–1042. doi 10.1111/eva.12178
29. *Deng X., Naccache S.N., Ng T. et al.* An ensemble strategy that significantly improves *de novo* assembly of microbial genomes from metagenomic next-generation sequencing data // Nucl. Acids Res. 2015. V. 43. № 7. doi 10.1093/nar/gkv002
30. *Peng Yu., Leung H.C.M., Yiu S.M., Chin F.Y.L.* IDBA – a practical iterative de Bruijn graph *de novo* assembler // Lect. Notes Comput. Sci. 2010. V. 6044. P. 426–440.
31. *Bankevich A., Nurk S., Antipov D. et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // J. Comp. Biol. 2012. V. 19. № 5. P. 455–477. doi 10.1089/cmb.2012.0021
32. *Schatz M.C., Delcher A.L., Salzberg S.L.* Assembly of large genomes using second-generation sequencing // Genome Res. 2010. V. 20. № 9. P. 1165–1173. doi 10.1101/gr.101360.109
33. *Pevzner P.A., Tang H., Waterman M.S.* An Eulerian path approach to DNA fragment assembly // Proc. Natl Acad. Sci. USA. 2001. V. 98. № 17. P. 9748–9753. doi 10.1073/pnas.171285098
34. *Chikhi R., Medvedev P.* Informed and automated *k*-mer size selection for genome assembly // Bioinformatics. 2014. V. 30. № 1. P. 31–37. doi 10.1093/bioinformatics/btt310
35. *Pendleton M., Sebra R., Chun Pang A.W. et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies // Nature Methods. 2015. V. 12. № 1. P. 780–786. doi 10.1038/nmeth.3454
36. *Li R., Zhu H., Ruan J. et al. De novo* assembly of human genomes with massively parallel short read sequencing // Genome Res. 2009. V. 20. № 2. P. 265–272. doi 10.1101/gr.097261.109
37. *Bradnam K.R., Fass J.N., Alexandrov A. et al.* Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species // GigaScience. 2013. V. 2. № 1.10. doi 10.1186/2047-217X-2-10
38. *Alkan C., Sajjadian S., Eichler E.E.* Limitations of next-generation genome sequence assembly // Nature Methods. 2011. V. 8. № 1. P. 61–65. doi 10.1038/nmeth.1527
39. *Love R.R., Weisenfeld N.I., Jaffe D.B. et al.* Evaluation of DISCOVAR *de novo* using a mosquito sample for cost-effective short-read genome assembly // BMC Genomics. 2016. V. 17. P. 187. doi 10.1186/s12864-016-2531-7
40. *Feng Y., Zhang Y., Ying C. et al.* Nanopore-based fourth-generation DNA sequencing technology // Genomics Proteomics Bioinformatics. 2015. V. 13. № 1. P. 4–16. doi 10.1016/j.gpb.2015.01.009
41. *Roberts R.J., Carneiro M.O., Schatz M.C.* The advantages of SMRT sequencing // Genome Biol. 2013. V. 14. № 7. 405. doi 10.1186/gb-2013-14-7-405
42. *Voskoboynik A., Neff N.F., Sahoo D. et al.* The genome sequence of the colonial chordate, *Botryllus schlosseri* // eLife. 2013. V. 2. e00569. http://dx.doi.org/. doi 10.7554/eLife.00569
43. *McCoy R.C., Taylor R.W., Blauwkamp T.A. et al.* Illumina TruSeq synthetic long-reads empower *de novo* assembly and resolve complex, highly repetitive elements // PLoS One. 2014. V. 9. № 9. e106689. http://dx.doi.org/. doi 10.1371/journal.pone.0106689
44. *Eisenstein M.* Oxford Nanopore announcement sets sequencing sector abuzz // Nature Biotechnol. 2012. V. 30. № 4. P. 295–296. doi 10.1038/nbt0412-295
45. *Chin C.-S., Alexander D.H., Marks P. et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data // Nature Methods. 2013. V. 10. № 6. P. 563–569. doi 10.1038/nmeth.2474
46. *Lee H., Gurtowski J., Yoo S. et al.* Error correction and assembly complexity of single molecule sequencing reads // bioRxiv. 2014. doi 10.1101/006395
47. *Koren S., Schatz M.C., Walenz B.P. et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads // Nat. Biotechnol. 2012. V. 30. № 7. P. 693–700. doi 10.1038/nbt.2280
48. *Faino L., Thomma B.P.H.J.* Get your high-quality low-cost genome sequence // Trends in Plant Sci. 2014. V. 19. № 5. P. 288–291. doi 10.1016/j.tplants.2014.02.003
49. *Flusberg B.A., Webster D.R., Lee J.H. et al.* Direct detection of DNA methylation during single-molecule, real-time sequencing // Nature. Methods. 2010. V. 7. № 6. P. 461–465. doi 10.1038/nmeth.1459
50. *Kim J., Larkin D.M., Cai Q. et al.* Reference-assisted chromosome assembly // Proc. Natl Acad. Sci. USA. 2013. V. 110. № 5. P. 1785–1790. doi 10.1073/pnas.1220349110
51. *Ellegren H.* Genome sequencing and population genomics in non-model organisms // Trends Ecol. Evol. 2014. V. 29. № 1. P. 51–63. doi 10.1016/j.tree.2013.09.008
52. *Li R., Fan W., Tian G. et al.* The sequence and *de novo* assembly of the giant panda genome // Nature. 2010. V. 463. № 7279. P. 311–317. doi 10.1038/nature08696
53. *Lindblad-Toh K., Wade C.M., Mikkelsen T.S. et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog // Nature. 2005. V. 438. № 7069. P. 803–819. doi 10.1038/nature04338
54. *Zhou S., Schwartz D.C.* The optical mapping of microbial genomes // ASM News. 2004. V. 70. № 7. P. 323–330.
55. *Howe K., Wood J.M.D.* Using optical mapping data for the improvement of vertebrate genome assemblies //

GigaScience. 2015. V. 4. № 10. doi 10.1186/s13742-015-0052-y

56. *Church D.M., Goodstadt L., Hillier L.W. et al.* Lineage-specific biology revealed by a finished genome assembly of the mouse // PLoS Biol. 2009. V. 7. № 5. e1000112. http://dx.doi.org/. doi 10.1371/journal.pbio.1000112
57. *Chen S., Xu J., Liu C. et al.* Genome sequence of the model medicinal mushroom *Ganoderma lucidum* // Nat. Commun. 2012. V. 3: 913. doi 10.1038/ncomms1923
58. *Dong Y., Xie M., Jiang Y. et al.* Sequencing and automated whole-genome optical mapping of the genome of a domestic goat (*Capra hircus*) // Nat. Biotechnol. 2013. V. 31. № 2. P. 135–141. doi 10.1038/nbt.2478
59. *Levy-Sakin M., Ebenstein Yu.* Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy // Curr. Opin. Biotech. 2013. V. 24. № 4. P. 690–698. doi 10.1016/j.copbio.2013.01.009
60. *Lam E.T., Hastie A., Lin C. et al.* Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly // Nat. Biotechnol. 2012. V. 30. № 8. P. 771–776. doi 10.1038/nbt.2303
61. *Shelton J.M., Coleman M.C., Herndon N. et al.* Tools and pipelines for BioNano data: molecule assembly pipeline and FASTA super scaffolding tool // BMC Genomics. 2015. V. 16: 734. doi 10.1186/s12864-015-1911-8
62. *Staňková H., Hastie A.R., Chan S. et al.* BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes // Plant Biotechnol. 2016. V. 14. № 7. P. 1523–1531. doi 10.1111/pbi.12513
63. *Богомолов А.Г., Карамышева Т.В., Рубцов Н.Б.* Флуоресцентная гибридизация *in situ* ДНК-проб, полученных из индивидуальных хромосом и хромосомных районов // Мол. биология. 2014. Т. 48. № 6. С. 881–890.
64. *Olson M., Hood L., Cantor C., Botstein D.* A common language for physical mapping of the human genome // Science. 1989. V. 245. № 4925. P. 1434–1435.
65. *Hudson T.J., Stein L.D., Gerety S.S. et al.* An STS-based map of the human genome // Science. 1995. V. 270. № 5244. P. 1945–1954.
66. *Burton J.N., Adey A., Patwardhan R.P. et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions // Nat. Biotechnol. 2013. V. 31. № 12. P. 1119–1125. doi 10.1038/nbt.2727
67. *Ay F., Noble W.S.* Analysis methods for studying the 3D architecture of the genome // Genome Biol. 2015. V. 16: 183. doi 10.1186/s13059-015-0745-7
68. *Seifertova E., Zimmerman L.B., Gilchrist M.J. et al.* Efficient high-throughput sequencing of a laser microdissected chromosome arm // BMC Genomics. 2013. V. 14: 357. doi 10.1186/1471-2164-14-357
69. *Rand K.H., Houck H. Taq* polymerase contains bacterial DNA of unknown origin // Mol. Cell Probes. 1990. V. 4. № 6. P. 445–450.
70. *Karlsson K., Sahlin E., Iwarsson E. et al.* Amplification-free sequencing of cell-free DNA for prenatal non-invasive diagnosis of chromosomal aberrations // Genomics. 2015. V. 105. № 3. P. 150–158. doi 10.1016/j.ygeno.2014.12.005
71. *Egger B., Ladurner P., Nimeth K. et al.* The regeneration capacity of the flatworm *Macrostomum lignano* – on repeated regeneration, rejuvenation, and the minimal size needed for regeneration // Dev. Genes Evol. 2006. V. 216. № 10. P. 565–577. doi 10.1007/s00427-006-0069-4
72. *Zadesenets K.S., Vizoso D.B., Schlatter A. et al.* Evidence for karyotype polymorphism in the free-living flatworm, *Macrostomum lignano*, a model organism for evolutionary and developmental biology // PLoS One. 2016. V. 11. № 10. e0164915. doi 10.1371/journal.pone.0164915

Whole-Genome Sequencing of Eukaryotes: from Sequencing of DNA Fragments to a Genome Assembly

K. S. Zadesenets[a], *, N. I. Ershov[a], and N. B. Rubtsov[a, b]

[a] *Institute of Cytology and Genetics Siberian Branch, Russian Academy of Sciences, Novosibirsk, 630090 Russia*

[b] *Novosibirsk State University, Novosibirsk, 630090 Russia*

**e-mail: kira_z@bionet.nsc.ru*

Rapid advances in sequencing technologies of second- and even third-generation made the whole genome sequencing a routine procedure. However, the methods for assembling of the obtained sequences and its results require special consideration. Modern assemblers are based on heuristic algorithms, which lead to fragmented genome assembly composed of scaffolds and contigs of different lengths, the order of which along the chromosome and belonging to a particular chromosome often remain unknown. In this regard, the resulting genome sequence can only be considered as a draft assembly. The principal improvement in the quality and reliability of a draft assembly can be achieved by targeted sequencing of the genome elements of different size, e.g., chromosomes, chromosomal regions, and DNA fragments cloned in different vectors, as well as using reference genome, optical mapping, and Hi-C technology. This approach, in addition to simplifying the assembly of the genome draft, will more accurately identify numerical and structural chromosomal variations and abnormalities of the genomes of the studied species. In this review, we discuss the key technologies for the genome sequencing and the *de novo* assembly, as well as different approaches to improve the quality of existing drafts of genome sequences. English translation of the paper published in Russian Journal of Genetics, 2017, Vol. 53, No. 6, is available ONLINE by subscription from: http://www.springer.com/, http://link.springer.com/journal/11177

Keywords: read, contig, scaffold, de Bruijn graph, chromosome mapping, methods, DNA.