



<https://doi.org/10.34883/PI.2024.13.2.011>
УДК 616-074:577.212.3



Костюк С.А.

Белорусский государственный медицинский университет, Минск, Беларусь

Секвенирование – технологическая часть получения первичных данных о последовательностях нуклеиновых кислот: особенности, которые следует учитывать при его осуществлении во избежание погрешностей в процедуре исследования

Конфликт интересов: не заявлен.

Подана: 06.02.2024

Принята: 19.04.2024

Контакты: s.kostiuk@mail.ru

Резюме

Изобретение и внедрение в практику технологий высокопроизводительного секвенирования вывели на новый уровень такие направления науки, как генетика, молекулярная биология, и дали стимулы для становления персонализированной медицины. В настоящее время область сиквенс-анализа объединяет широкую гамму различных технологий, базирующихся на разных принципах и разработанных независимо. Правильное планирование научного эксперимента в медицине крайне важно. Перед началом работы исследователь обязан перечислить для себя все элементы исследования, способные исказить результаты, приложив максимальные усилия к уменьшению или устранению таких искажений. Ошибки и искажения поджидают исследователя на каждом из этапов секвенирования – от сбора биологических образцов до обработки файлов. В данной статье описаны технические особенности получения первичных данных о последовательностях нуклеиновых кислот.

Ключевые слова: секвенирование, ДНК, РНК, молекулярная биология, ПЦР, первичные данные

Svetlana A. Kostiuk
Belarusian State Medical University, Minsk, Belarus

Sequencing is a Technological Part in Obtaining Primary Data on Nucleic Acid Sequences: Features to be Taken into Account when Performing it to Avoid Errors in the Test Procedure

Conflict of interest: nothing to declare.

Submitted: 06.02.2024
Accepted: 19.04.2024
Contacts: s.kostiuk@mail.ru

Abstract

Inventing and implementing high-throughput sequencing technologies brought to a new level such areas of science as genetics and molecular biology, and provided incentives for the development of personalized medicine. Today, the field of sequencing analysis includes a wide range of different technologies based on different principles and developed independently. Proper planning of a scientific experiment in medicine is extremely important. Before beginning a work, the researcher must list for himself or herself all elements of the study that could distort the results, making every effort to reduce or eliminate such distortions. Errors and distortions await the researcher at every stage of sequencing, from biological sampling to file processing. This article describes technical features of obtaining primary data on nucleic acid sequences.

Keywords: sequencing, DNA, RNA, molecular biology, PCR, primary data

Большинство современных методов молекулярной биологии предполагает использование множества идентичных макромолекул для получения детектируемого сигнала [1]. К ним относятся различные виды хроматографии, рентгеноструктурный анализ, масс-спектрометрия. Секвенирование ДНК также требует усиления сигнала за счет использования в анализе множества одинаковых молекул ДНК. Определение последовательности нуклеиновых кислот (НК) методами высокопроизводительного секвенирования представляет собой одновременное параллельное прочтение последовательности нескольких миллионов разных (относительно коротких) фрагментов исходной ДНК. Наиболее популярные на рынке технологии секвенирования не позволяют оценивать нуклеотидную последовательность непосредственно геномной или комплементарной ДНК и предполагают амплификацию исходных молекул [2, 3].

Перед загрузкой в секвенатор исследуемую ДНК необходимо модифицировать путем создания коллекции случайных фрагментов нужной структуры. Образец ДНК, определенным образом фрагментированный и амплифицированный, называют библиотекой случайных фрагментов ДНК для секвенирования. Если фрагменты при этом прикреплены к предметному стеклу или микрочастице, библиотеку называют иммобилизованной.



Общая последовательность этапов высокопроизводительного секвенирования для наиболее популярных платформ следующая [2, 3]:

- 1) разрушение ДНК с получением фрагментов определенной длины;
- 2) присоединение синтетических олигонуклеотидных адаптеров по краям фрагментов;
- 3) наработка (амплификация) каждого фрагмента ДНК в отдельном микрореакторе с микрочастицей (эмульсионная ПЦР) и/или непосредственно на поверхности предметного стекла (мостиковая ПЦР);
- 4) определение нуклеотидной последовательности фрагментов ДНК;
- 5) биоинформатический анализ данных – коротких прочтений.

Приготовление библиотеки фрагментов происходит на первых трех этапах и определяет 90% успеха сиквенсного проекта. В настоящее время в высокопроизводительном секвенировании используют два основных типа библиотек фрагментов ДНК: обычную и инвертированную. Во избежание путаницы отметим, что производители наборов реагентов для конструирования библиотек выделяют гораздо больше типов библиотек, например, разделяя их по виду стартового материала или числу и направлению прочтений, однако такое деление не несет под собой научного обоснования и является маркетинговым приемом.

Исходным материалом для исследования могут служить любые НК: ДНК всех типов, РНК всех типов, включая микроРНК, продукты полимеразной цепной реакции (ПЦР), отдельные области генома (таргетное секвенирование), комплексы нуклеиновых кислот с белками и т. д. Несмотря на различие в целях и задачах экспериментов, создание библиотеки для высокопроизводительного секвенирования в итоге сводится к разрушению ДНК до фрагментов определенного размера или отбору определенной фракции из существующего спектра фрагментов, лигированию адаптеров и увеличению количества ДНК одним из методов наработки *in vitro* (амплификации), отбору фракции фрагментов определенной длины, нанесению на твердую фазу и т. п. [1–3].

В случае исследования РНК первым этапом после очистки от примесей является обратная транскрипция, то есть создание комплементарной ДНК, и далее продолжают подготовку библиотеки, как для ДНК. Следовательно, метод очистки РНК должен эффективно удалять из биологического образца примеси, способные помешать работе ревертазы [4, 5]. Если же предполагается секвенирование генома, чувствительным к примесям этапом может стать ферментативное разрушение ДНК или этап лигирования адаптеров, если разрушение ДНК выполнялось физическими методами.

Все современные методы очистки нуклеиновых кислот можно глобально разделить на две группы: методы с поэтапным удалением примесей из водного раствора НК и методы, основанные на сорбции НК на твердой фазе. Наиболее известной методикой первого типа является фенол-хлороформная экстракция [1, 6]. Вторая группа методик основана на использовании силикатных сорбентов, эффективно связывающих НК в растворе с высокой ионной силой. Одной из первых появившихся методик такого типа является метод выделения ДНК, предложенный Boom R. с соавторами [7]. Эта технология основана на использовании для лизиса клеток сильного хаотропного агента гуанидина тиоционата (GuSCN) и последующей сорбции ДНК на твердом носителе. После промывки сорбента на нем фиксируется чистый образец НК, легко снимающийся дистиллированной водой. В последнее время чрезвычайно

распространены одноразовые пластиковые микроколоники с упакованным в них сорбентом, например, наборы реагентов фирм QIAGEN, Ambion и др. Промывка колонок растворами с высокой ионной силой удаляет белки и низкомолекулярные соединения, после чего проводят элюцию очищенных НК раствором с низкой ионной силой.

Концентрацию полученных препаратов ДНК или РНК перед обратной транскрипцией можно измерить непосредственно по спектру поглощения нуклеиновых кислот на спектрофотометре, по флуоресценции интеркалирующего красителя после электрофореза или при помощи флуориметра, а также методом количественной ПЦР в режиме реального времени [1, 4, 5].

Производители оборудования для секвенирования рекомендуют осуществлять измерение с помощью компактных флуориметров типа Qubit от Life Technologies или Quantus от Promega. Принцип их действия основан на измерении флуоресценции интеркалирующего в ДНК красителя типа SYBR. Точность измерения обеспечивается калибровкой прибора с помощью двух стандартов перед каждой серией измерений. Преимущество флуоресцентного определения концентрации перед спектрофотометрическим с помощью довольно распространенных приборов типа Nanodrop заключается в гораздо более широком диапазоне концентраций и более точных результатах.

Метод количественной ПЦР в режиме реального времени как способ определения концентрации ДНК несколько сложнее и дольше в исполнении, но дает гораздо более точные и воспроизводимые результаты в огромном линейном диапазоне значений. Если стартовым материалом является геномная ДНК, можно использовать праймеры на уникальный регион, представленный в геноме один раз, если это комплементарная ДНК, то можно оценивать количество одного или двух средне- или низкопредставленных транскриптов. Праймеры желательно располагать близко, оставляя место для олигонуклеотидной пробы типа TaqMan [1, 4, 5].

Отдельно следует упомянуть ситуацию, когда по каким-либо причинам невозможно получить достаточное стартовое количество ДНК или комплементарной ДНК, например, в случае если недостаточно биологического материала и проводится исследование единичных клеток. В этой ситуации можно перед началом создания библиотеки фрагментов амплифицировать НК одним из стандартных способов. При наличии малого количества геномной ДНК возможно прибегнуть к методике полногеномной амплификации, основанной на изотермическом синтезе ДНК [8]. Если работа начинается с малого количества РНК, можно использовать амплификацию комплементарной ДНК [9]. Важно при использовании предварительной амплификации ДНК или комплементарной ДНК в проектах секвенирования помнить об искажениях, которые неизбежно вносит этот прием.

Фрагментировать ДНК можно физическими или ферментатическими способами. В ряде случаев, например, если стартовым материалом служат микроРНК или короткие продукты ПЦР, этап разрушения не требуется, однако может потребоваться этап отбора фракции фрагментов нужной длины. Физически ДНК можно фрагментировать ультразвуком, процесс называют соникацией, пропусканием через микротверстие – небулизацией, гидродинамическим воздействием – гидроширингом. К преимуществам физических методов перед ферментатическими можно отнести хорошую воспроизводимость результатов, поскольку они создают стабильный диапазон длин получаемых фрагментов даже для разного по качеству стартового материала,



а также низкую зависимость от последовательности. Энзиматические методы расщепления ДНК предполагают использование различных эндонуклеаз, при этом наборы реагентов для энзиматического расщепления представлены на рынке как производителями платформ для секвенирования, так и биотехнологическими компаниями, предлагающими альтернативные, более дешевые решения для приготовления библиотек. Чаще всего для расщепления ДНК используют смеси модифицированных нуклеаз – типа эндонуклеазы T7 и нуклеазы из *Vibrio vulnificus*, а также транспозаз [2, 3].

Постановка ферментативной реакции не требует больших трудозатрат, протокол легко можно оптимизировать под особенности каждого типа разрушаемой ДНК и ожидаемых длин фрагментов в любом диапазоне значений. Данный подход заключается в подборе периода обработки нуклеазами и температуры реакции: чем короче фрагменты нужно получить, тем дольше проводят реакцию. Следует отметить, что правильно расщепленная ДНК представляет собой набор фрагментов, длина которых распределена по закону Гаусса с максимумом в требуемом диапазоне значений. В случае, если ДНК «перефрагментирована» или «недофрагментирована», можно, несмотря на отличные аппаратные показатели секвенирования, при биоинформатическом анализе получить неинформативные данные, так как значительная часть ДНК окажется вне оптимума размеров.

Некоторые технологии создания библиотек используют отбор фракций нужной длины сразу после расщепления ДНК: чаще – для инвертированных библиотек; другие берут в лигирование весь спектр фрагментов. К преимуществам ферментативного разрушения ДНК перед физическими методами можно отнести образование однотипных концов необходимой структуры, по которым и проводится присоединение сиквенсных адаптеров, тогда как физически разрушенная ДНК требует ферментативной обработки концов. После каждой ферментативной реакции необходимо проводить очистку от фермента и иных компонентов реакции спиртовым осаждением ДНК или сорбентными методами, что всегда сопряжено с потерями исследуемого биологического материала [2, 3].

Наиболее распространенным способом оценки длины фрагментированной ДНК является метод электрофореза в геле. Для этого может использоваться как обычный электрофорез в агарозном геле с окрашиванием бромистым этидием, так и готовые аппаратные решения, например, с использованием электрофореза в микрокапиллярах на небольшой пластиковой подложке. После получения фрагментов ДНК желаемого размера путем подбора условий разрушения ДНК или непосредственно отбором оптимальной фракции фрагментов выполняют лигирование адаптеров, с которых будет осуществляться амплификация библиотеки и ее секвенирование [1].

Во всех коммерчески доступных технологиях высокопроизводительного секвенирования к фрагментам присоединяют два разных адаптера, причем в секвенировании будут использованы только фрагменты ДНК, несущие по концам разные адаптеры, поскольку фрагменты с одним адаптером не будут амплифицированы из-за эффекта супрессии ПЦР [10, 11].

На этапе лигирования важно выдержать соотношение между концентрацией адаптеров и фрагментированной ДНК для того, чтобы лигирование прошло эффективно. Косвенным признаком эффективности присоединения адаптеров является количество циклов ПЦР с праймерами, отжигающимися на адаптерах, необходимое для получения детектируемого продукта реакции [2, 3].

Этап предварительной амплификации библиотеки требуется в том случае, если количество исходной ДНК гораздо ниже рекомендуемого. Сложность любой амплификации ДНК *in vitro* заключается в неизбежном искажении исходной представленности фрагментов. В случае ПЦР искажение происходит вследствие разной эффективности амплификации для разных по длине и последовательности фрагментов ДНК, поэтому количество циклов при амплификации библиотеки должно быть минимальным, не превышающим 13–15 циклов [4, 5]. Большее количество циклов ПЦР для образцов геномной или комплементарной ДНК ведет к снижению сложности смеси фрагментов ДНК и потере некоторых последовательностей из библиотеки, что значительно искажает результаты секвенирования.

На следующем технологическом этапе производят отбор фракций нужной длины фрагментов библиотеки, что существенно повышает эффективность ее секвенирования, поскольку более короткие фрагменты в сравнении с оптимальными дают слишком короткие прочтения, а более длинные фрагменты искажают результаты измерений и подбор соотношений для оптимального сиквенс-анализа [2, 3]. Для отбора нужной фракции библиотеки исследователь может использовать как обычный электрофорез в агарозном геле, так и аппаратные решения разных производителей, в том числе и на основе гель-электрофореза.

Во всех случаях осуществляют разделение библиотеки ДНК в геле вместе с маркером длин, например, с шагом в 50 п. н. на одной из дорожек, а затем вырезают из геля фракцию нужной длины. Самым дешевым способом является обычный агарозный гель-электрофорез с вырезанием участка геля при помощи скальпеля и элюцией из него ДНК набором реагентов для экстракции из геля. К недостаткам данного метода можно отнести низкую устойчивость к ошибкам лаборанта и высокую вероятность кросс-контаминации, связанную с многократным использованием камеры для электрофореза, заливочного столика, гребенок и т. д. [1–3].

Более предпочтительными для отбора определенной фракции фрагментов ДНК являются специализированные системы, работающие на картриджных гелях. Система позволяет визуализировать гель-электрофорез непосредственно во время его прохождения. Элюция нужной фракции осуществляется из специальных лунок, расположенных в конце геля, которые исследователь заполняет дистиллятом или трис-ЭДТА-буфером [2, 3].

Поскольку технологии секвенирования весьма производительны, часто исследуемый образец не требует столь высокого объема данных на выходе, поэтому для экономии средств можно смешать несколько исследуемых образцов и секвенировать их вместе, в ходе одной процедуры. Если ДНК предварительно пометить образец-специфичными адаптерами, секвенируемыми вместе с уникальной частью фрагментов ДНК, биоинформатический анализ полученных прочтений позволяет разделить данные от разных биологических образцов. Такой подход получил название «штрихкодирования» (или «бар-кодирования»), по аналогии с широко используемыми в маркировке товаров уникальными метками штрихкодами (barcodes).

Пример рационального использования «штрихкодов» – секвенирование экзона человека, который составляет около 1% от гаплоидного генома размером 3 млрд п. н., на платформе SOLiD 5500, использующей одну проточную камеру с шестью дорожками, с каждой из них получается около 24 млрд п. н. Наносить на одну дорожку с такой производительностью один образец экзона экономически неоправданно.



Рациональнее пометить уникальным образом несколько разных образцов, смешать их эквимольно и провести одновременное секвенирование, а результаты разделить уже биоинформатически на основании наличия в каждом прочтении специфичной метки. Метки представляют собой короткие около 10 п. н. последовательности, которые также прочитываются в процессе секвенирования. Их присоединяют к фрагментам ДНК на стадии конструирования библиотеки при лигировании сиквенсных адаптеров к фрагментам ДНК. Таким образом, «штрихкод» представляет собой участок адаптера для секвенирования, расположенный после участка отжига праймера, с которого начинается сиквенс-анализ. Как правило, для разных образцов используют один и тот же адаптер 1, отжигающийся на микросферах для проведения эмульсионной ПЦР, и набор адаптеров 2, отличающихся штрихкодированными участками.

Важным условием проведения эффективного секвенирования для всех образцов, внесенных в одну реакцию, является их эквимольное смешивание, что обеспечит равномерную представленность прочтений для каждого образца [2, 3]. На это стоит обратить особое внимание: нужно очень аккуратно измерить концентрацию каждой библиотеки перед смешиванием, для этого можно использовать метод ПЦР в реальном времени, который обладает более высокой точностью.

Ключевым элементом большинства технологий высокопроизводительного секвенирования является одновременная наработка множества разных по последовательности фрагментов ДНК так, чтобы получаемые ампликоны были локализованы вместе на одной микросфере или в одной точке на поверхности. Существует два метода такого ПЦР-клонирования: мостиковая ПЦР и эмульсионная ПЦР.

Большинство технологий секвенирования следующего поколения недостаточно чувствительны, чтобы регистрировать сигнал от единичной молекулы ДНК, и требуют применения клонирования *in vitro*: способа ферментативной наработки отдельных молекул ДНК с получением изолированных друг от друга пулов идентичных фрагментов. Простейшим способом клонирования фрагментов ДНК *in vitro* может быть обычная ПЦР [11]. Таким образом, все наработанные в данной пробирке фрагменты будут клонами единственной молекулы.

На принципе амплификации фланкированных адаптерами отдельных фрагментов ДНК основано множество молекулярных технологий [9, 12], например, мостиковая ПЦР (bridge PCR) [13]. Принцип этой технологии заключается в проведении ПЦР с праймерами, прикрепленными к твердой фазе или подложке. С пришитого к поверхности праймера синтезируется фрагмент ДНК, после этапа денатурации фрагмент снова взаимодействует с праймером на поверхности в пределах доступности по длине фрагмента, образуя дугу (мостик) между двумя точками на подложке, при повторении циклов ПЦР из точки синтеза первого фрагмента колония фрагментов ДНК будет быстро увеличиваться. Создание клональной библиотеки фрагментов ДНК методом мостиковой ПЦР используется в технологии компании Illumina.

Эмульсионная ПЦР (emulsion PCR, ePCR) представляет собой другой распространенный вариант клонирования фрагментов ДНК *in vitro*. Эта методика позволяет амплифицировать ДНК на покрытых праймерами микросферах таким образом, что каждая микросфера оказывается окружена только одним типом фрагментов. Для этого создают библиотеку фрагментов ДНК с адаптерами по концам, а затем смешивают полученные фрагменты с покрытыми праймерами микросферами, свободным

праймером и другими необходимыми для ПЦР компонентами в условиях мелкодисперсной водно-масляной эмульсии так, чтобы в каждую микрокаплю воды попали в среднем одна микросфера и один исходный матричный фрагмент ДНК. В ходе ПЦР праймеры на микросфере инициируют синтез фрагментов, начиная с единственной мишени, а с праймера в растворе достраивается вторая цепь. В итоге реакции получают миллионы микросфер, каждая из которых несет миллионы идентичных фрагментов ДНК [12, 13].

Для оценки качества библиотеки после эмульсионной ПЦР исходят из правила, что соотношение сработавших микросфер и несработавших должно быть 30%:70%. Большой процент сработавших микросфер указывает на их высокую поликлональность. Оценить процент сработавших микросфер возможно путем измерения флуоресценции несущих ДНК микросфер относительно несработавших. После эмульсионной ПЦР проводят обогащение фракции сработавших микросфер, например, за счет гибридизации с магнитными частицами, несущими стрептавидин. При этом сработавшие микросферы в этом случае должны нести биотин на дистальном адаптере.

Многие пользователи считают эмульсионную ПЦР трудоемкой и весьма нестабильной. Действительно, в первых вариантах секвенирования эмульсию приходилось готовить вручную, и малейшие отклонения в технологии приводили к разрушению всей эмульсии. В настоящее время разработаны автоматизированные подходы к проведению эмульсионной ПЦР, например, с помощью приборов серии OneTouch. В 2013 году для платформ Ion Torrent появился полностью автоматический аппарат для пробоподготовки Ion Chef, выполняющий эмульсионную ПЦР, обогащение и загрузку чипа в автоматическом режиме. На методе эмульсионной ПЦР базируются технологии секвенирования 454 Life Sciences (Roche), Ion PGM и Ion Proton (Life Technologies), Polonator (Dover/Harvard), SOLiD (Life Technologies).

Если используемая технология высокопроизводительного секвенирования предполагает прочтение каждого фрагмента библиотеки с двух сторон, существенным отличием этих библиотек является расстояние по исходной ДНК между получаемыми в результате секвенирования одного фрагмента последовательностями. Для обычной библиотеки это расстояние соответствует физической длине фрагмента и, как правило, не превышает 2–3 т. п. н., для инвертированной библиотеки это могут быть и десятки тысяч пар нуклеотидов. В процессе секвенирования обычная библиотека может быть прочитана как в одном направлении, так и в обоих. Платформы на технологиях Ion Torrent и 454 Life Sciences в настоящее время позволяют проводить только однонаправленное секвенирование фрагментов с дистального по отношению к микросфере адаптера. Для Ion Torrent показана возможность по окончании сиквенса в одном направлении провести денатурацию библиотеки ДНК прямо в чипе и «смыть» синтезированную фракцию ДНК, а затем провести секвенирование тех же фрагментов, расположенных в тех же микрореакторах, в противоположном направлении.

Технологии Illumina и SOLiD позволяют осуществлять секвенирование фрагмента с двух сторон. После прочтения с одного из адаптеров проводят чтение в обратном направлении с праймера, комплементарного адаптеру на противоположном конце. Чтение с двух сторон хорошо тем, что это позволяет значительно повысить объем данных с одной библиотеки и, в случае перекрывания прочтений, их общую длину, при отсутствии перекрывания дает примерное расположение прочтений друг относительно друга. При этом качество сигнала быстро падает от начала секвенирования



к концу фрагмента, так что прочтение короткого фрагмента с двух сторон насквозь позволяет компенсировать низкое качество прочтения [14].

Более трудоемкой в изготовлении, но крайне полезной в случае полногеномного секвенирования, особенно *de novo*, и для некоторых других задач является инвертированная библиотека. Суть подхода состоит в том, что в один фрагмент для секвенирования попадают два участка исходной ДНК, расстояние между которыми примерно известно, причем в зависимости от задачи можно регулировать удаленность попадающих в совместный сиквенс фрагментов генома от нескольких сотен до нескольких десятков тысяч пар нуклеотидов [15]. Преимущество инвертированных библиотек в первую очередь определяется тем, что ввиду короткой длины прочтений за одно прочтение не удастся насквозь пройти участки многочисленных повторов, дубликаций и других повторяющихся элементов генома, тем самым сборка прочтений в более длинные затруднена или даже невозможна. Парное прочтение инвертированной библиотеки дает информацию о расположении двух прочтений относительно друг друга, что можно использовать для сборки геномов на повторяющихся последовательностях.

Помимо инвертированных библиотек, прочитанных с двух сторон, информацию о расположении контигов в геноме можно получить и на основании сравнения с уже собранными геномами генетически близких видов, однако нужно помнить, что такие допущения необходимо подтверждать экспериментально из-за часто встречающихся даже для близких таксонов крупных перестроек в геномах.

Безусловно, получением первичных данных секвенирование не заканчивается, а скорее только начинается. Автоматический секвенатор дает миллионы коротких прочтений, которые необходимо собрать в более длинные последовательности и провести их биоинформатический анализ. К сожалению, в настоящее время ни одна из технологий секвенирования не позволяет определять последовательность генома непрерывно на протяжении многих миллионов пар геномной ДНК. Все технологии предполагают фрагментирование ДНК и дают в результате сравнительно короткие, но перекрывающиеся участки последовательности длиной 50–1000 п. н., поэтому в случае определения генома какого-либо организма впервые возникает необходимость его сборки, определения полной последовательности по огромному набору фрагментов.

К сожалению, на практике исследователь всегда имеет дело с данными, в которых присутствуют ошибки секвенирования, существенно усложняющие и без того непростую задачу по сборке *de novo*. Для частичного устранения ошибок секвенирования полученные данные предварительно обрабатывают. Кроме последовательности нуклеотидов, приборы для секвенирования выдают данные о качестве прочтения каждой буквы, отбрасывая данные, имеющие низкое качество, можно значительно сократить количество ошибок. Большинство программ для сборки геномов позволяют проводить такую фильтрацию, для тех же, где она не предусмотрена, желательно провести ее самостоятельно перед сборкой.

Существует еще одна распространенная проблема – повторяющиеся последовательности, присутствующие в любом геноме: это повторы, превышающие длину одного прочтения для использованной технологии секвенирования, что требует существенного увеличения необходимого объема аппаратной памяти для хранения информации о связях.

Несмотря на огромный поток экспериментов по секвенированию геномов и широкий спектр методических приемов, на сегодняшний день нет однозначного ответа на вопрос, какие комбинации технологий секвенирования и программ для сборки дают наилучшие результаты при секвенировании *de novo*. Нет и единых критериев оценки качества сборки.

Любой научный эксперимент требует тщательной предварительной подготовки. В медицинской науке планирование экспериментальной части исследования имеет большое значение по причине широкой вариабельности свойств, характерной для биологических объектов. Эта особенность также является причиной трудностей при интерпретации результатов, которые могут значительно различаться от опыта к опыту [16].

Известно, что геномы двух организмов одного и того же вида, как правило, не являются полностью идентичными. Наиболее частыми различиями в близкородственных геномах являются однонуклеотидные вариации отличия ДНК по нуклеотиду в определенной позиции, дающие так называемый внутривидовой однонуклеотидный полиморфизм (single nucleotide polymorphism, SNP), являющийся основной причиной существования аллелей [17]. Наряду с SNP часто встречаются вставки или делеции одного нуклеотида. Более протяженные вставки или делеции обычно называют вариацией числа копий региона, они могут быть разной протяженности: от нескольких нуклеотидов до целой хромосомы. Отклонение числа хромосом от нормального называют анеупloidией. Кроме того, возможны инверсии изменения структуры хромосомы, вызванные разворотом одного из ее участков на 180°, и транслокации – перенос участка хромосомы в новое место. Все эти изменения можно обнаружить с помощью высокопроизводительного секвенирования, как правило, в варианте повторного секвенирования генома.

Ряд биологических задач предполагает повторное определение последовательности ДНК вида организмов, геном которого уже есть в базе данных. Для прокариот это может быть секвенирование отличающихся определенными свойствами – патогенностью, продуктивностью – штаммов микроорганизма. Для эукариот – накопление данных о вариабельности генома внутри вида, ассоциативные исследования, в том числе поиск генетических детерминант определенных фенотипов [17].

Если перед исследователем стоит задача определения последовательности генома организма, данные для которого уже есть в базе данных, то есть стоит задача повторного секвенирования, ресеквенирования, ее решение существенно упрощается. В этом случае проводят сравнение полученных данных с референсным геномом, и от качества такого сравнения зависят результаты дальнейшего анализа данных.

Для сравнения полученных данных геномного секвенирования с уже имеющимися полными геномами совсем не обязательно собирать геном заново, достаточно выровнять (картировать) прочтения на уже имеющуюся референсную последовательность. Несмотря на то, что задача выравнивания на референсный геном намного проще, чем сборка *de novo*, и здесь существует ряд трудностей. Во-первых, даже для сравнительно небольшого генома бактерии – 5 млн п. н. – для каждого фрагмента в ходе выравнивания необходимо проверить 5 млн возможных вариантов расположения этого участка на геноме. Но в референсном геноме могут быть и небольшие отличия – однонуклеотидные замены, делеции или вставки. Если разрешить алго-



ритму позиционировать последовательность с учетом наличия единственного несовпадающего нуклеотида на 100 п. н., то мы получим $3 \times 100 \times 5\,000\,000$ вариантов расположения фрагмента на референсном геноме, то есть для каждого прочтения нужно выполнить уже 1,5 млрд сравнений. Если предположить наличие делеции или вставки, количество сравнений возрастет еще больше. Для современных персональных ЭВМ решение по выравниванию бактериального генома на референсный геном занимает около двух часов. В целом задачу картирования прочтений на референсном геноме можно сформулировать следующим образом: даны набор полученных в эксперименте последовательностей Q , набор референсных последовательностей R , набор ограничений и пороговое значение дистанции k . Необходимо найти все подстроки t в R , удовлетворяющие всем ограничениям, в том числе и такому, что расстояние между t и любой последовательностью q из Q не превышает A' , то есть $d(q, t) < k$, где $d(q, t)$ – функция расстояния. Ее обычно рассчитывают на основании количества замен, делений или вставок. Также она может учитывать протяженность делений или вставок и характерные ошибки использованной платформы секвенирования. Основная цель картирования – найти правильное расположение в геноме каждой последовательности q из Q с учетом ошибок секвенирования и структурных вариантов. При этом важно отличать ошибки секвенирования от реальных различий в сравниваемых последовательностях.

Однако технологии высокопроизводительного секвенирования дают на порядки больший объем данных, и методы, успешно применявшиеся для сенгеровского секвенирования, оказываются недостаточно эффективны для работы с большими массивами данных [2, 3]. Поэтому были разработаны новые алгоритмы, требующие меньше аппаратной мощности и времени на выполнение. Число программ для картирования прочтений уже превысило полсотни и постоянно растет. Выбирая подходящую программу для картирования прочтений, нужно учитывать, для каких данных и каких задач создавалась та или иная программа (ДНК, РНК, микроРНК, изучение паттерна метилирования ДНК посредством обработки ее бисульфитом или паттерна расположения ДНК-связывающих белков после иммунопреципитации хроматина и т. д.). Кроме того, некоторые программы разработаны специально для работы с данными конкретной технологической платформы для секвенирования и учитывают характерные особенности или ошибки именно этой платформы, например, прочтение гомополимерных участков для Ion Torrent и 454 Life Sciences, одонуклеотидные замены и ухудшение качества данных к концу прочтений для Illumina и особенности формата данных, цветовое кодирование для данных SOLiD.

Если в ходе анализа генома обнаружена новая вариация, возникает вопрос, какое биологическое значение она может иметь. Конечно, без экспериментальных данных биоинформатические подходы «сказать» что-то наверняка не могут, но могут дать направление для дальнейшего поиска.

Для этих целей разработаны инструменты, позволяющие на основании информации о месте расположения вариации сделать предположение о ее возможной биологической роли. Такие программные продукты позволяют определить, находится ли вариабельная позиция в кодирующей последовательности регуляторной области, в области некодирующей РНК и т. д. Если вариация обнаружена в кодирующей последовательности, то является ли она значимой, приводит ли к замене аминокислоты, влияет ли на рамку считывания и т. д.

Таким образом, любой лабораторный эксперимент в науке требует тщательной предварительной подготовки. В медицинской биологии планирование экспериментальной части исследования имеет большое значение по причине широкой вариативности свойств, характерной для биологических объектов. Эта особенность является основной причиной трудностей при интерпретации результатов, которые могут значительно различаться от опыта к опыту. Правильное планирование эксперимента в биологии крайне важно. Поэтому перед началом работы исследователь обязан перечислить для себя все элементы исследования, способные исказить результаты, приложив максимальные усилия к уменьшению или устранению таких искажений, лежащих в основе возникновения ошибок в процедуре исследования.

■ ЛИТЕРАТУРА/REFERENCES

1. Kostiuik S. *Molecular biological methods in medicine*. Minsk: Belarusian Medical Academy of Postgraduate Education. 2013; 327 p. (in Russian)
2. Kostiuik S. Sequencing – methods for deciphering the nucleotide sequence of DNA. Message 1. *Meditsinskie novosti*. 2017;6:18–21. (in Russian)
3. Kostiuik S. Sequencing – methods for deciphering the nucleotide sequence of DNA. Message 2. *Meditsinskie novosti*. 2017;7:11–14. (in Russian)
4. Kostiuik S., Poluyan O. Molecular biological factors influencing the diagnostic reliability of laboratory test results based on PCR analysis technologies. Message I. *Laboratory diagnostics. Eastern Europe*. 2020;3(9):314–323. (in Russian)
5. Kostiuik S., Poluyan O. Molecular biological factors influencing the diagnostic reliability of laboratory test results based on PCR analysis technologies. Message II. *Laboratory diagnostics. Eastern Europe*. 2020;4(9):366–374. (in Russian)
6. Chomczynski P., Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chlorophorm extraction. *Analytical Biochemistry*. 1987;162:156–159.
7. Boom R. Improved silica-guanidiniumthiocyanate DNA isolation procedure based on selective binding of bovine alphacasein to silica particles. *Journal of Clinical Microbiology*. 1999;37:615–619.
8. Zhang L. Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Science USA*. 1992;89(13):5847–5851.
9. Lukyanov K. Construction of cDNA libraries from small amounts of total RNA using the suppression PCR effect. *Biochemical and Biophysical Research Communications*. 1997;230(2):285–288.
10. Bankevich A. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*. 2012;9(5):455–477.
11. Lukyanov K.A. Molecule by molecule PCR amplification of complex DNA mixtures for direct sequencing: an approach to in vitro cloning. *Nucleic Acids Research*. 1996;24(11):2194–2195.
12. Rebrikov D.V. Mirror orientation selection (MOS): a method for eliminating false positive clones from libraries generated by suppression subtractive hybridization. *Nucleic Acids Research*. 2000;28(20):90.
13. Dressman D. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences USA*. 2003;100:8817–8822.
14. Williams R. Amplification of complex gene libraries by emulsion PCR. *Nature methods*. 2006;3(7):545–550.
15. Akoiuitz A., Manuelidis L. A novel cDNA/PCR strategy for efficient cloning of small amounts of undefined RNA. *Gene*. 1989;81(2):295–330.
16. Kostiuik S. Laboratory errors and risk management in the molecular genetic laboratory. *Medical news*. 2021;1:6–10. (in Russian)
17. Kostiuik S. Predisposition genes for multifactorial diseases and genetic polymorphism: what is it? *Medical news*. 2020;7:45–49. (in Russian)