

Секвенирование ДНК: методы первого, второго и третьего поколения

Секвенирование ДНК представляет собой совокупность методов, позволяющих определить последовательность нуклеотидов в молекуле ДНК. Если ПЦР отвечает прежде всего на задачу многократного копирования определенного участка, а электрофорез позволяет оценить размер полученных фрагментов, то секвенирование дает значительно более подробную информацию — оно позволяет установить порядок расположения аденина, тимина, гуанина и цитозина в исследуемой последовательности.

Определение последовательности ДНК стало одним из важнейших технологических достижений молекулярной биологии. Благодаря секвенированию стало возможным исследовать структуру генов и целых геномов, выявлять генетические варианты, изучать наследственные заболевания, анализировать микроорганизмы, исследовать эволюционные связи и получать огромные массивы геномных данных. Современная геномика практически невозможна без технологий секвенирования.

Развитие методов секвенирования условно разделяют на три поколения. К первому поколению относят классические методы определения последовательности отдельных молекул или сравнительно небольших фрагментов, прежде всего секвенирование по Сэнгеру. Второе поколение представлено технологиями массового параллельного секвенирования, которые обычно объединяют термином NGS — Next-Generation Sequencing. Третье поколение основано преимущественно на анализе отдельных молекул ДНК или РНК в реальном времени и позволяет получать значительно более длинные прочтения.

Исторически одним из ключевых этапов стало появление в 1970-х годах метода Фредерика Сэнгера. В 1977 году был опубликован метод определения последовательности ДНК с использованием принципа контролируемого прекращения синтеза цепи. Впоследствии метод был автоматизирован и на протяжении многих лет оставался основным инструментом секвенирования.

Принцип секвенирования по Сэнгеру основан на синтезе новой цепи ДНК. Для реакции необходимы исследуемая матрица, праймер, ДНК-полимераза и обычные дезоксинуклеотиды — dATP, dTTP, dGTP и dCTP. Однако дополнительно используются модифицированные нуклеотиды — дидезоксинуклеотиды, ddNTP.

Ключевая особенность ddNTP заключается в отсутствии химической группы, необходимой для дальнейшего удлинения цепи. Поэтому после включения такого нуклеотида синтез данной молекулы прекращается. Это явление называют терминацией цепи.

В современной автоматизированной версии четыре типа ddNTP имеют различные флуоресцентные метки. Во время реакции образуется смесь фрагментов различной длины, каждый из которых заканчивается определенным меченым нуклеотидом. Затем эти фрагменты разделяются, например методом капиллярного электрофореза. Более короткие фрагменты проходят систему раньше длинных, а детектор регистрирует флуоресцентную метку каждого фрагмента.

В результате компьютер формирует электрофорограмму, или хроматограмму, представленную последовательностью цветных пиков. Каждый пик соответствует

определенному нуклеотиду. Последовательность пиков позволяет восстановить последовательность исследуемого фрагмента ДНК.

Метод Сэнгера обладает высокой точностью и остается полезным для исследования отдельных генов и сравнительно небольших участков ДНК. Например, если после ПЦР необходимо подтвердить последовательность одного ампликона, секвенирование по Сэнгеру часто является рациональным вариантом.

Предположим, что при генотипировании исследуется определенный SNP. Сначала участок, содержащий этот вариант, амплифицируется с помощью ПЦР. Затем PCR-продукт очищается и секвенируется. На электроферограмме исследователь анализирует соответствующую позицию и определяет присутствующий нуклеотид или сочетание сигналов, соответствующее генотипу.

Таким образом, классическая последовательность анализа выглядит следующим образом:

ДНК → ПЦР целевого участка → получение ампликона → секвенирование по Сэнгеру → электроферограмма → анализ последовательности → определение генетического варианта.

Однако метод Сэнгера имеет ограниченную производительность. Он прекрасно подходит для отдельных фрагментов, но становится менее эффективным, если необходимо исследовать тысячи генов или весь геном. Именно необходимость значительно увеличить объем получаемых данных привела к развитию технологий секвенирования второго поколения.

Методы второго поколения получили общее название Next-Generation Sequencing, NGS. Их фундаментальное отличие от классического подхода заключается в массовом параллельном анализе огромного числа фрагментов ДНК.

Вместо последовательного чтения одного относительно небольшого участка NGS позволяет одновременно анализировать миллионы фрагментов. Это резко увеличило производительность секвенирования и снизило стоимость получения большого объема генетической информации.

Работа NGS начинается с подготовки исследуемого материала. ДНК выделяют из биологического образца, оценивают ее количество и качество, после чего формируют секвенирующую библиотеку. В данном контексте библиотека представляет собой подготовленную совокупность молекул ДНК, совместимых с конкретной технологической платформой.

При подготовке библиотеки к молекулам добавляются необходимые технологические последовательности, называемые адаптерами. Они позволяют фрагментам взаимодействовать с системой секвенирования и могут содержать дополнительные элементы, необходимые для идентификации образцов.

Особое значение имеют индексы, или баркоды. Благодаря им библиотеки нескольких образцов можно объединить и секвенировать одновременно. После получения данных компьютер определяет, к какому исходному образцу относится каждое прочтение. Такой подход называют мультиплексированием.

После подготовки библиотеки проводится собственно секвенирование. Конкретный химический принцип зависит от используемой платформы. Одной из наиболее распространенных концепций является sequencing by synthesis — определение последовательности в процессе синтеза новой цепи ДНК.

В результате получают огромное количество относительно коротких последовательностей, называемых reads — прочтениями. Далее начинается биоинформатическая часть исследования.

Прочтения проходят контроль качества, после чего могут быть сопоставлены с референсным геномом. Этот процесс называется alignment — выравнивание. Если подходящего референса нет или задача требует восстановления последовательности из самих данных, может применяться сборка — assembly.

После выравнивания можно искать отличия исследуемого образца от референсной последовательности. Этот этап называется variant calling. В результате выявляются SNP, небольшие вставки и делеции, а при соответствующем дизайне и анализе — другие типы геномных изменений.

Затем найденные варианты проходят аннотацию. На этом этапе определяется, где расположен вариант, затрагивает ли он ген, кодирующую область или другой функциональный элемент и какие известные сведения с ним связаны. После этого проводится биологическая или клиническая интерпретация в зависимости от задачи исследования.

Поэтому общий NGS-процесс можно представить следующим образом:

биологический образец → выделение ДНК → контроль качества → подготовка библиотеки → секвенирование → получение reads → контроль качества данных → выравнивание или сборка → поиск вариантов → аннотация → интерпретация.

NGS представляет собой не один конкретный анализ. На основе технологий второго поколения выполняются различные типы исследований. Одним из них является targeted sequencing — таргетное секвенирование, при котором исследуется заранее выбранная группа генов или участков генома.

Если исследователя интересует определенная группа генов, например связанных с конкретным биологическим процессом, таргетная панель позволяет сосредоточить секвенирование именно на этих участках. Такой подход уменьшает объем данных по сравнению с полногеномным секвенированием и упрощает последующую интерпретацию.

Другим вариантом является Whole Exome Sequencing, WES — секвенирование экзона. Экзом представляет собой совокупность преимущественно белок-кодирующих экзонных областей генома. Хотя он составляет относительно небольшую часть генома человека, в этих областях находится значительное количество вариантов, способных непосредственно изменять структуру белков.

Более масштабным подходом является Whole Genome Sequencing, WGS — полногеномное секвенирование. В этом случае анализируется практически весь доступный геномный материал, включая кодирующие и некодирующие области. WGS предоставляет наиболее широкий объем информации, но одновременно требует больших вычислительных ресурсов, хранения данных и более сложной интерпретации.

Кроме анализа ДНК, технологии массового секвенирования используются для исследования РНК. RNA-seq позволяет анализировать транскриптом — совокупность РНК, присутствующих в клетке или ткани в определенных условиях. Благодаря этому можно сравнивать экспрессию генов, исследовать различные транскрипты и изучать изменения клеточных программ.

NGS также является основой современной метагеномики. Вместо выделения и культивирования каждого микроорганизма отдельно исследователь может анализировать генетический материал целого микробного сообщества. Это особенно важно при изучении микробиоты, почвенных и водных экосистем и других сложных биологических объектов.

Несмотря на огромную производительность, секвенирование второго поколения имеет ограничения. Одно из основных связано с относительно небольшой длиной отдельных прочтений у распространенных short-read платформ. Если геном содержит длинные повторяющиеся последовательности, определить их точную организацию по коротким фрагментам бывает сложно.

Представим книгу, которую разрезали на множество небольших фрагментов по несколько слов, а затем попросили восстановить исходный текст. Если предложения уникальны, задача сравнительно проста. Но если в книге много одинаковых повторяющихся предложений, определить исходное положение каждого короткого фрагмента становится значительно сложнее. Аналогичная проблема возникает при сборке некоторых участков генома из коротких reads.

Именно потребность в чтении более длинных молекул стала одним из факторов развития секвенирования третьего поколения.

К технологиям третьего поколения обычно относят методы, способные анализировать отдельные молекулы нуклеиновых кислот без обязательной массивной клональной амплификации каждой мишени, причем многие такие платформы позволяют получать длинные и сверхдлинные прочтения.

Одним из важных направлений является Single Molecule Real-Time sequencing — SMRT sequencing. В этой технологии наблюдается синтез отдельной молекулы ДНК в реальном времени. Современные подходы позволяют получать длинные прочтения и при определенных режимах высокую консенсусную точность.

Другой принцип используется в нанопоровом секвенировании. В упрощенном виде молекула нуклеиновой кислоты проходит через нанометровую пору, а система регистрирует изменения электрического сигнала. Характер сигнала зависит от последовательности нуклеотидов, находящихся в области поры. Затем специальное программное обеспечение преобразует зарегистрированный сигнал в последовательность оснований — этот процесс называют basecalling.

Одним из главных преимуществ технологий третьего поколения являются long reads — длинные прочтения. Они могут охватывать участки генома, которые трудно корректно анализировать с помощью коротких прочтений.

Длинные reads особенно полезны для исследования повторяющихся областей, крупных структурных перестроек, сложных геномных локусов и построения более непрерывных сборок геномов. Они также позволяют лучше определять, какие варианты находятся на одной молекуле или хромосомной копии, что важно для фазирования генетических вариантов.

Технологии длинных прочтений применяются и для анализа транскриптов. Возможность прочитать значительную часть или целую молекулу транскрипта облегчает исследование различных изоформ РНК и альтернативного сплайсинга.

Отдельное преимущество некоторых платформ третьего поколения заключается в возможности получать дополнительную информацию о модификациях нуклеиновых кислот из особенностей измеряемого сигнала. Конкретные возможности и точность зависят от используемой технологии и алгоритмов анализа.

Если сравнить три поколения секвенирования, становится видно, что каждое решает свою группу задач. Секвенирование первого поколения характеризуется высокой точностью при анализе отдельных сравнительно небольших фрагментов и удобно для подтверждения конкретных последовательностей. Второе поколение обеспечивает огромную параллельность и позволяет исследовать тысячи генов, экзомы и целые геномы. Третье поколение делает акцент на анализе отдельных молекул и получении длинных прочтений, позволяющих исследовать сложную структуру генома.

Важно понимать, что понятия «первое», «второе» и «третье поколение» не означают, что новая технология полностью делает предыдущую ненужной. Метод выбирается в зависимости от биологического вопроса.

Если необходимо проверить последовательность одного PCR-продукта, применение классического секвенирования по Сэнгеру может быть вполне рациональным. Если необходимо одновременно исследовать большое количество генов или весь экзом, преимущество получает NGS. Если задача связана со сложными структурными вариантами, повторяющимися областями или необходимостью получения длинных последовательностей, технологии long-read sequencing могут предоставить дополнительную информацию.

Современные исследования нередко используют гибридный подход, при котором short-read и long-read данные дополняют друг друга. Короткие прочтения могут обеспечивать высокую глубину покрытия, а длинные помогают разрешать сложную структуру генома. Выбор стратегии определяется задачей, типом образца, необходимой точностью и доступными ресурсами. При анализе секвенирования необходимо учитывать понятие глубины покрытия, или sequencing depth. Если определенный участок генома прочитан многократно независимыми молекулами, исследователь получает больше данных для оценки последовательности в этой позиции. Другим понятием является coverage — покрытие. В зависимости от контекста оно может характеризовать, какая часть целевой области представлена в данных и насколько многократно она прочитана. Поэтому при оценке результатов недостаточно просто сказать, что образец был секвенирован; необходимо учитывать качество прочтений и характеристики покрытия. Огромное значение имеет биоинформатика. Современный секвенатор генерирует данные, но биологический смысл появляется только после их обработки и интерпретации. Сначала оценивается качество исходных reads. Затем они выравниваются относительно референсного генома или используются для сборки. После этого выявляются варианты, проводится их фильтрация и аннотация.

Упрощенную биоинформатическую цепочку можно представить следующим образом:

raw reads → quality control → filtering → alignment/assembly → variant calling → annotation → interpretation.

Количество данных, получаемых при современном секвенировании, может быть очень большим. Поэтому геномика объединяет молекулярную биологию, статистику, программирование и вычислительные технологии. Секвенирование применяется в фундаментальной генетике для изучения структуры и вариативности геномов. В медицинской генетике оно используется для поиска вариантов, потенциально связанных с наследственными заболеваниями. В онкологических исследованиях секвенирование позволяет исследовать соматические изменения опухолевых клеток. В микробиологии секвенирование используется для идентификации и характеристики микроорганизмов, исследования геномов бактерий и вирусов и анализа микробных сообществ. В эволюционной биологии сравнение последовательностей позволяет исследовать генетическое родство организмов.

В сельском хозяйстве секвенирование применяется для геномного анализа растений и животных, исследования генетического разнообразия и поиска маркеров,

связанных с хозяйственно значимыми признаками. В экологии оно используется при анализе eDNA и биоразнообразия.

Однако получение последовательности еще не означает автоматического получения биологического вывода. Обнаруженный вариант необходимо правильно интерпретировать. Один и тот же ген может содержать множество вариантов, часть из которых не оказывает заметного функционального эффекта. Поэтому особенно в биомедицинских исследованиях необходимо различать факт обнаружения генетического варианта и доказательство его функционального или клинического значения.

Качество исходной ДНК также остается критически важным. Для short-read NGS умеренная фрагментация в некоторых протоколах может быть допустима или даже являться частью подготовки библиотеки. Для long-read sequencing, напротив, сохранение высокомолекулярной ДНК часто имеет принципиальное значение. Если исходные молекулы были сильно разрушены при выделении, потенциальное преимущество длинных прочтений уменьшается.

Поэтому тема секвенирования непосредственно связана с ранее рассмотренными методами выделения и контроля качества ДНК. Правильная последовательность работы выглядит следующим образом:

качественный биологический образец → выделение ДНК → оценка концентрации, чистоты и целостности → выбор технологии секвенирования → подготовка материала → получение последовательностей → биоинформатический анализ → биологическая интерпретация.

Если обобщить развитие технологий, можно представить следующую логическую последовательность:

Первое поколение — чтение отдельных фрагментов.

Второе поколение — массовое параллельное чтение миллионов фрагментов.

Третье поколение — преимущественно чтение отдельных длинных молекул в реальном времени. Главный принцип выбора метода остается тем же, что и для других ДНК-технологий: сначала формулируется биологический вопрос, а затем выбирается технология, способная наиболее подходящим образом предоставить необходимые данные.

Таким образом, секвенирование ДНК прошло путь от определения последовательности отдельных фрагментов до массового анализа целых геномов и чтения длинных отдельных молекул. Метод Сэнгера заложил фундамент современной геномики. NGS сделал возможным высокопроизводительный анализ огромного количества последовательностей. Технологии третьего поколения расширили возможности исследования сложной организации геномов благодаря длинным прочтениям и анализу отдельных молекул.

Для студента важно понимать не только названия поколений, но прежде всего различие их принципов. Сэнгер основан на терминеции синтеза цепи; NGS второго поколения — на массовом параллельном секвенировании большого числа фрагментов; технологии третьего поколения — на анализе отдельных молекул с возможностью получения длинных прочтений. Эти различия определяют возможности, ограничения и области применения каждого подхода.