Технологии высокопараллельного секвенирования в медико-генетических исследованиях*

Тощаков С.В., Доминова И.Н., Патрушев М.В.

ФГАОУ ВПО «Балтийский федеральный университет им. И. Канта»,
236041, г.Калининград, ул. А.Невского, д.14, факс +7(4012) 466313, e-mail: post@kantiana.ru

За последние 5 лет технологии секвенирования следующего поколения сыграли огромную роль в исследованиях генетических аспектов патогенеза различных заболеваний, и можно утверждать, что они стали мощнейшим толчком к появлению новой науки — медицинской геномики. Уже сейчас можно наблюдать многочисленные попытки внедрения высокопроизводительного секвенирования в клиническую диагностику, которые, однако, бессмысленны без глубокого понимания развития данных методик, их подводных камней и специфики. Задачей обзорной статьи являются детальное рассмотрение ныне существующих и активно разрабатываемых технологий высокопроизводительного секвенирования, а также обсуждение наиболее ярких примеров применения геномных технологий в медицине и потенциал их внедрения в рутинную практику медико-диагностических лабораторий.

Ключевые слова: геномика, секвенирование, секвенирование следующего поколения, нанопоры, биомаркёры мультифакториальных заболеваний

Введение

Анализируя современные тенденции развития медицины, нельзя не обратить внимание на то, что активное использование геномных данных является лейтмотивом большинства разрабатываемых диагностических и терапевтических подходов. Представляя собой некий опорный элемент, геномика определяет развитие сопутствующих системных дисциплин, которые в своей совокупности в иностранной литературе именуются как «-omics» (-омики). Геномика, как и другие «-омики», представляет собой самостоятельную научную дисциплину, основными характеристиками которой являются высокая технологичность и большой объём получаемых данных.

Рассматривая ретроспективу развития геномики, можно заметить, что её появление обусловлено двумя факторами, основным из которых является создание высоких технологий расшифровки геномов. Второй фактор — выраженная необходимость в получении максимально возможной информации о геноме каждого конкретного объекта исследования — приобрёл своё значение после завершения программы «Геном человека» и лёг в основу популярной в настоящее время концепции *персонализированной медицины*. Под данным термином принято понимать систему диагностических и терапевтических методов и подходов, учитывающих индивидуальные особенности организма, прежде всего, его генетическую конституцию.

Таким образом, возросшие технологические возможности и порождённый программой «Геном человека» спрос на качественно новый тип данных инициировали «геномную гонку», участниками которой стали крупнейшие биотехнологические компании, а главной целью — не разработка качественно новых исследовательских инструментов, а уменьшение стоимости процесса. В настоящее время мы наблюдаем активную фазу этого технологического соревнования, однако уже можем наблюдать некоторые промежуточные результаты: стоимость определения первичной последовательности ДНК снизилась почти на 5 порядков, оборудование, необходимое для расшифровки полного генома человека можно разместить в средней по размеру лаборатории. При этом нельзя не отметить те трудности, с которыми столкнулись исследователи при массовом внедрении технологий секвенирования следующего поколения:

- отсутствие универсальных алгоритмов аннотации геномов;
- слишком большое количество данных, депонированных в разнообразных геномных базах;
- необходимость тщательной верификации геномных данных.

Указанные проблемы послужили мощнейшим толчком к привлечению сложных вычислительных алгоритмов, что, в свою очередь, стало главным фактором развития биоинформатики, которая сегодня имеет определяющее значение в дальнейшей судьбе высокопроизводительных областей биологической науки.

Роль, которую играют технологии секвенирования следующего поколения в фундаментальных исследованиях, трудно переоценить. Несмотря на существующие проблемы, результаты, получаемые с их использованием, формируют новые представления о структуре и функционировании клеток и органов. Тенденцией же последнего времени стало постепенное внедрение высокопроизводительного секвенирования в различные прикладные дисциплины, в том числе в медицину.

* Данная работа выполнена при поддержке Министерства образования и науки Российской Федерации, соглашение 14.A18.21.0109.

Таким образом, представляя собой некий опорный элемент, геномика определяет развитие сопутствующих дисциплин (транскриптомика, эпигеномика, фармакогеномика), потенциал которых предопределён для использования в медицинской практике.

В данной обзорной статье мы детально рассматриваем ныне существующие и активно разрабатываемые технологии высокопроизводительного секвенирования, а также обсуждаем наиболее яркие примеры применения геномных технологий в медицине и потенциал их внедрения в рутинную практику медико-диагностических лабораторий.

Программа «Геном человека»

В феврале 2001 г. журнал Nature анонсировал завершение проекта «Геном человека», запущенного в 1990 г. [11]. Несмотря на то, что завершение этой работы обозначило начало новой, постгеномной эры исследований наследственного аппарата человека, полученные данные не дали возможность совершить стремительный рывок в медицине и фармакологии. Напротив, окончание этого проекта создало новый запрос в развитии технологий, позволяющих быстро и относительно недорого расшифровывать геномы отдельных индивидуумов и устанавливать причину фенотипических различий, определяя таким образом генетические детерминанты, вызывающие развитие различных патологий.

В основе технологической платформы, использованной для первичной расшифровки генома человека, лежал метод прямого ферментативного секвенирования ДНК с использованием искусственных терминаторов элонгации цепи, предложенный Ф. Сэнгером в 1977 г. [26]. Исходно данный подход использовал радиоактивно меченные терминаторы, требовал разделения реакции на 4 пробирки и использования сложных в приготовлении высокоразрешающих полиакриламидных гелей, что делало процесс достаточно трудоёмким. Впоследствии, с появлением флуоресцентно меченных терминаторов элонгации, стало возможным проводить разделение фрагментов по длинам в одной дорожке геля или капилляре (рис. 1А) [19]. Такое усовершенствование методики дало возможность автоматизировать систему и привело к появлению первых коммерческих капиллярных секвенаторов. К моменту завершения программы «Геном человека» наиболее высокопроизводительные приборы такого типа (ABI PRISM 3700 DNA Analyzer system, Applied Biosystems, США) имели 96 капилляров и могли прочитывать в среднем около 700 п.н. Полный цикл работы прибора с полной загрузкой — около 3 ч. Таким образом, получается, что геном человека размером 3 млрд п.н. можно расшифровать посредством одного капиллярного секвенатора приблизительно за 15 лет непрерывной работы. Кроме того, необходимо учитывать, что для достоверного секвенирования необходимо прочитать каждое основание по нескольку раз (при точности капиллярного секвенатора — как минимум 5), а также то, что не все прочтения могут быть достоверно отнесены к определённому участку генома. Это и определило масштабность проекта «Геном человека», над которым работали десятки исследовательских групп по всему миру и который обошёлся приблизительно в 3 млрд долл. США.

Полученные результаты данные дали учёным неоценимую возможность исследовать структуру генома человека, оценить количество генов и кодируемых ими белков, определить соотношение между генами и межгенными участками. Однако необходимо отметить, что первый расшифрованный геном человека представлял по своей сути «смесь» геномов разных индивидуумов [31]. Это определило предназначение полученных данных именно как референтной последовательности, т.е. мощного инструмента для проведения медико-генетических исследований, но никак не универсальной «энциклопедии жизни», способной дать ответ на все вопросы о генетических основах патогенеза. Стало понятно, что для определения строгого соответствия между генотипом и фенотипом необходимо иметь полную последовательность генома каждого конкретного организма. Первая успешная попытка секвенирования персонального генома методом автоматизированного секвенирования по Сэнгеру завершилась в 2007 г., т.е. через 6 лет после завершения проекта «Геном человека» [12]. В результате усовершенствования технологии и значительного удешевления реактивов этот проект обошёлся исследователям в 1000 раз дешевле, чем проект «Геном человека», тем не менее, и стоимость, и продолжительность этого проекта поставила под сомнение возможность реализации подобных инициатив с использованием классической технологии секвенирования.

Появление секвенаторов «второго поколения»

Впервые принцип высокопараллельного секвенирования был реализован в середине 90-х годов С. Бреннером [1]. Несмотря на то, что предложенная методика была очень трудоемка и, в результате этого, так и не была коммерциализирована, её принцип лёг в основу большинства ныне существующих методов секвенирования второго поколения. В данной группе методов секвенирования фрагментированная геномная ДНК наносится на поверхность проточной ячейки в виде молекулярных колоний (так называемых ПЦР-колоний или микросфер, несущих десятки тысяч копий одной молекулы ДНК). После этого происходит определение их последовательности путём оптической детекции сигнала с каждой из частиц (колоний), расположенных строго в определённом месте проточной ячейки. Процесс секвенирования происходит путём повторяющихся циклов подачи реагентов (например, флуоресцентно меченных нуклеотидов), отмывки их избытка и последующего сканирования поверхности слайда (рис. 1Б).

В 2005 г. Д. Черчу удалось значительно упростить методику пробоподготовки, применив эмульсионную ПЦР, позволяющую физически разделить амплифицируемые молекулы ДНК в пространстве. Вследствие этого из процесса подготовки библиотек был исключён этап клони-

рования фрагментированной геномной ДНК в *E.coli* со специфическими для каждой молекулы олигонуклеотидными адаптерами, что значительно упростило работу [28]. Впоследствии данный подход, сочетающий в себе эмульсионную ПЦР и процесс секвенирования путём лигирования флуоресцентно меченных олигонуклеотидов, лёг в основу системы полногеномного секвенирования SOLiD (Life Technologies, США) [30].

Одновременно с этим идеи, заложенные Бреннером, были реализованы в системе высокопроизводительного пиросеквенирования 454, увидевшей свет в конце 2005 г. [15]. В данном подходе используется метод регистрации включения нуклеотида в синтезирующуюся цепь ДНК путём люминометрической детекции высвобождающегося при этом пирофосфата. С помощью этой системы в достаточно короткие сроки (2 мес.) был полностью секвенирован второй индивидуальный геном человека и стоимость этого проекта составила около 1 млн долл. Результаты этого проекта во многом определили конец эпохи методов масштабного капиллярного секвенирования, которые стали постепенно сдавать позиции и уступать место высокопараллельным технологиям.

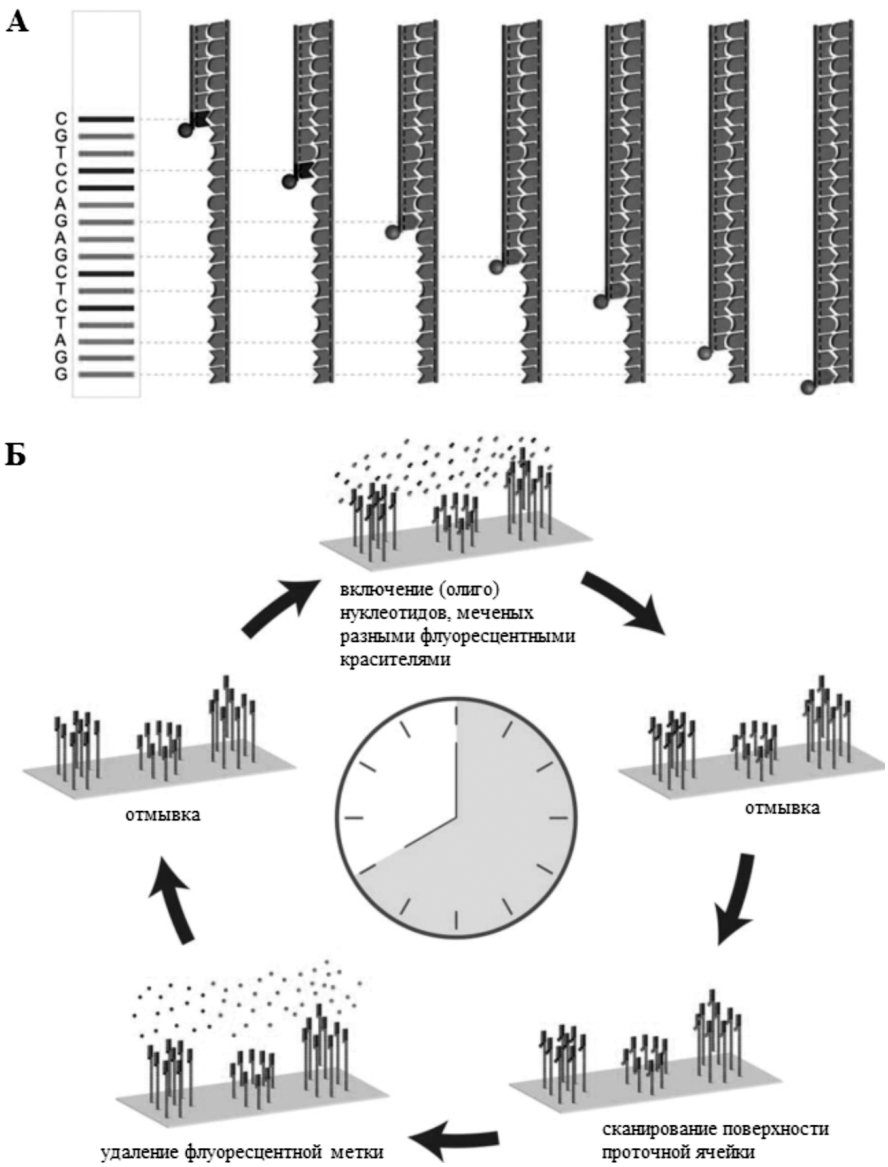

Рис. 1. Схема работы систем секвенирования первого и второго поколений:
А — современная схема секвенирования по Сэнгеру с использованием четырёх различных флуоресцентных меток и терминаторов полимеризации ДНК. После этапа полимеризации реакционная смесь разделяется по размеру с использованием методов капиллярного электрофореза;
Б — общая схема секвенирования, характерная для систем второго поколения. После включения нуклеотидов (система Illumina) или олигонуклеотидов (система SOLiD) в растущую цепь ДНК происходит отмывка избыточных реагентов и сканирование поверхности проточной ячейки. Затем метка удаляется и цикл повторяется заново[32]

Параллельно с технологиями прочтения нуклеотидной последовательности совершенствовались и подходы к клональной амплификации. Так, в системе компании Illumina вместо эмульсионной ПЦР был использован так называемый bridge PCR, или твердофазный ПЦР, идущий путём образования молекулярных мостиков между закреплёнными на поверхности проточной ячейки олигонуклеотидами. Такой подход позволил частично обойти ограничения, накладываемые эмульсионной ПЦР (длина фрагмента ДНК) и дал возможность более эффективно картировать геномные перестройки методами секвенирования второго поколения [2].

Однако, несмотря на достигнутые успехи в развитии технологии, применение таких подходов в реальной клинико-диагностической практике ставилось под сомнение в результате времени прогона прибора — время одного запуска полногеномного секвенатора может составлять до двух недель. Наконец, огромный объём данных, получаемый с одного запуска полногеномного секвенатора (до 700 млрд п.н.), затрудняет как хранение, так и дальнейший анализ таких данных. Поэтому в целях увеличения скорости секвенирования и уменьшения общей стоимости приборов основными компаниями-производителями секвенаторов второго поколения были созданы «младшие братья» существующих систем — более компактные, быстрые и недорогие приборы. Так, система Miseq (Illumina, США) обладает производительностью в 40 раз меньшей, нежели полная версия прибора — один запуск даёт возможность расшифровать последовательность общей длиной 15 млрд п.н., тогда как система HiSeq, основанная на том же принципе секвенирования, позволяет прочитать около 600 млрд нуклеотидов. Аналогичными системами обзавелись и другие биотехнологические компании, и сейчас активно идёт работа по получению одобрения FDA использования таких приборов в диагностических целях. Однако, с точки зрения авторов данного обзора, это может быть не совсем обоснованно с точки зрения технологических и функциональных особенностей таких систем. На этапах подготовки геномных библиотек и клональной амплификации исходный генетический материал подвергается многократным циклам ПЦР, что вносит свой негативный вклад в суммарную точность таких систем. Кроме того, поскольку эффективность включения нуклеотидов в цепь ДНК всегда меньше 100%, популяция растущих

Таблица

Основные существующие и разрабатываемые платформы высокопараллельного секвенирования

Наименование секвенатора	Основа процесса секвенирования и пробоподготовки	Длина прочтения, п.н.	Максимальный выход, млн п.н.	Время анализа	Уровень ошибок
Roche/454 GS FLX Titanium	Пиросеквенирование, эПЦР	500—1000	700	23 ч	0,4—1,5%
Roche/454 GS Junior	Пиросеквенирование, эПЦР	400	35	10 ч	0,4—1,5%
Ion Torrent PGM	Секвенирование синтезом, эПЦР	200—400	От 1000	4 ч	1—3%
Illumina HiSeq 2000	Секвенирование синтезом, обратимый терминатор, твердофазная ПЦР	До 2 100	До 600 000	До 11 дней	0,5—2%
Illumina MiSeq	Секвенирование синтезом, твердофазная ПЦР	До 2 250	До 15 000	До 27 ч	0,5—2%
Life Technologies 5500xl	Секвенирование лигированием меченых олигонуклеотидов, эПЦР	2 60	До 300 000	До 7 дней	0,20%
Ion Torrent Proton	Секвенирование синтезом, эПЦР	200	10 000	4 ч	—
HeliScope	Одномолекулярное секвенирование путём синтеза цепи	До 55	35 000	До 8 дней	3—5%
Pacific BioSciences	Секвенирование одиночных молекул в реальном времени	До 6000	90	1 ч	13—15%
Oxford Nanopore Technologies	Биологические и твердотельные нанопоры, электрическое считывание	—	—	—	—
Noblegen	Твердотельные нанопоры, оптическое считывание	—	—	—	—
Genia	Биологические нанопоры, электрическое считывание	—	—	—	—
IBM	Твердотельные нанопоры, электрическое считывание	—	—	—	—

Примечание. Прочерк (—) — информация недоступна

цепей ДНК с каждым шагом становится всё более десинхронизированной [32]. Это явление называется *дефазированием* и вызывает значительное уменьшение качества сигнала по мере продвижения к 3'-концу молекулы ДНК и, таким образом, является одним из основных факторов, снижающих максимальную длину прочтения. Наконец, нельзя не отметить, что небольшая длина прочтений высокопроизводительных систем по сравнению с сэнгеровскими (капиллярными) секвенаторами создаёт определённые затруднения в картировании данных прочтений на референтный геном [27]. Подытоживая вышесказанное, можно утверждать, что технологии секвенирования первого и второго поколений, несомненно, произвели революцию в геномных технологиях, однако тот факт, что для решения клинико-диагностических задач одновременно необходимы точность, скорость работы и относительно небольшая стоимость приборов, породждает значительный запрос на создание принципиально новых технологических подходов.

Переход к новым системам секвенирования

Логично рассудить, что следующее поколение систем секвенирования должно называться *секвенаторами третьего поколения*. Однако на данный момент в научном сообществе единое мнение по этому поводу отсутствует. В первую очередь, это связано с вопросом, что же должен собой представлять секвенатор третьего поколения, который сможет найти рутинное применение в клинических исследованиях? Очевидно, что такой прибор должен сочетать в себе отказ от оптических систем детекции вследствие очевидной дороговизны высокоразрешающих оптических систем и флуоресцентных меток. Кроме того, для повышения точности прибора очевидным шагом является отказ от многочисленных раундов амплификации в ходе пробоподготовки, значительно снижающих фактическую точность прибора. Таким образом, прибор должен обеспечивать электрическую детекцию сигнала от одиночных молекул ДНК. Однако, несмотря на достаточно чёткое определение, значительное количество существующих и разрабатываемых в настоящий момент систем высокопроизводительного секвенирования не вписывается в настолько ясно обозначенные границы.

Одной из технологий, находящихся на стыке второго и третьего поколений секвенаторов, служит технология IonTorrent (Life Technologies, США), в которой используется система детекции сигнала на полупроводниковых чипах. В основе работы лежит простой принцип: при включении нуклеотида в растущую цепь ДНК происходит выделение пирофосфата (детекция которого используется в системе 454) и иона водорода, используемого в качестве аналита в данной системе. Основное её достоинство — использование в качестве детекторов полупроводниковых чипов, содержащих от одного до нескольких десятков миллионов лунок (в зависимости от ёмкости чипа), в каждой из которых находится анализируемая микрочастица с клонально амплифицированными молекулами ДНК. Детекция иона водорода происходит путём прямого измерения изменения pH среды. Это позволяет производить детекцию сигнала без использования оптики, что значительно снижает стоимость как прибора, так и процесса секвенирования. При этом общий принцип пробоподготовки также включает в себя этап клональной амплификации при помощи эмульсионной ПЦР, что не позволяет отнести данный прибор к системам секвенирования третьего поколения [25].

Секвенирование одиночных молекул нуклеиновых кислот

Первым прибором, позволяющим проводить секвенирование одиночных молекул ДНК, стал прибор Helicos Genetic Analysis Platform (Helicos, США). Работа данной системы основывалась на визуализации роста цепи ДНК, т.е. по своей сути мало отличалась от систем секвенирования второго поколения. Однако высокое разрешение оптики и эффективность флуоресцентно меченных аналогов нуклеотидов позволяла визуализировать одиночные молекулы ДНК матрицы, не прибегая к клональной амплификации геномной библиотеки. При этом детекция одиночных молекул с использованием стандартной технологии происходила неточно. Уровень ошибки одиночного прочтения составлял около 0,5%, а максимальная длина составляла 32 нуклеотида. Помимо этого, стоимость прибора и реагентов была столь высока, что вкупе с остальными указанными параметрами не позволила этой технологии завоевать прочное место в научных лабораториях [24].

Секвенирование одиночных молекул в реальном времени

Более серьёзные позиции сейчас занимает система PacBio RS (Pacific Biosciences, США), позволяющая осуществлять секвенирование одиночных молекул ДНК в реальном времени. Данная технология позволяет в реальном времени следить за кинетикой включения нуклеотидов в растущую цепь ДНК посредством специальных устройств — волноводов с нулевой модой (zero-mode waveguide, ВНМ). В полупроводниковом чипе, содержащем около 75 тыс. таких волноводов, ДНК-полимераза иммобилизуется на дне ячеек. В ходе секвенирования матрица ДНК связывается с полимеразой, после чего подаются нуклеотиды и начинается регистрация сигнала в реальном времени при помощи системы оптики. Используемые аналоги нуклеотидов содержат флуоресцентную метку, присоединённую к 5'-фосфатным группам. При включении нуклеотидов флуоресцентная метка высвобождается и диффундирует в зону волновода, регистрируемую системой оптики. Поскольку включение нуклеотида происходит за милли-, а диффузия — за микросекунды, инкорпорация нуклеотида даёт сигнал, достаточный для того, чтобы отличить его от шума волновода, созданного диффун-

дирующими молекулами (рис. 2А). Помимо возможности секвенировать единичные молекулы важным достоинством данной системы является длина прочтения. В настоящий момент средняя длина прочтения составляет 3000 п.н., а время прогона при этом — от 30 до 60 мин. Кроме того, PacBio RS — единственная доступная технология, позволяющая детектировать метилированный цитозин с разрешением в 1 нуклеотид, не применяя при этом бисульфитной конверсии геномной ДНК [3]. Ограничения на применение данной системы в клинической практике определяются в первую очередь вероятностью ошибки одиночного прочтения, которая составляет 15—20% и является самой высокой для коммерчески доступных приборов, а также значительной стоимостью и громоздкостью аппарата, который весит около 1,5 т.

Системы секвенирования на основе нанопор

Под секвенированием с использованием нанопор понимается определение последовательности одиночных молекул нуклеиновых кислот, проходящих через поры размером несколько нанометров под воздействием электрического поля. Принцип основан на том, что при прохождении нуклеиновых кислот через тонкие плёнки с нанопорами за процессом прохождения можно наблюдать, измеряя при этом ионный ток через нанопору [8].

В настоящее время активно разрабатывается множество различных систем, основанных на этом принципе. К основным достоинствам такого подхода относятся отсутствие какой-либо пробоподготовки перед запуском прибора, длина прочтения, достигающая нескольких миллионов пар нуклеотидов и, наконец, то, что последо-

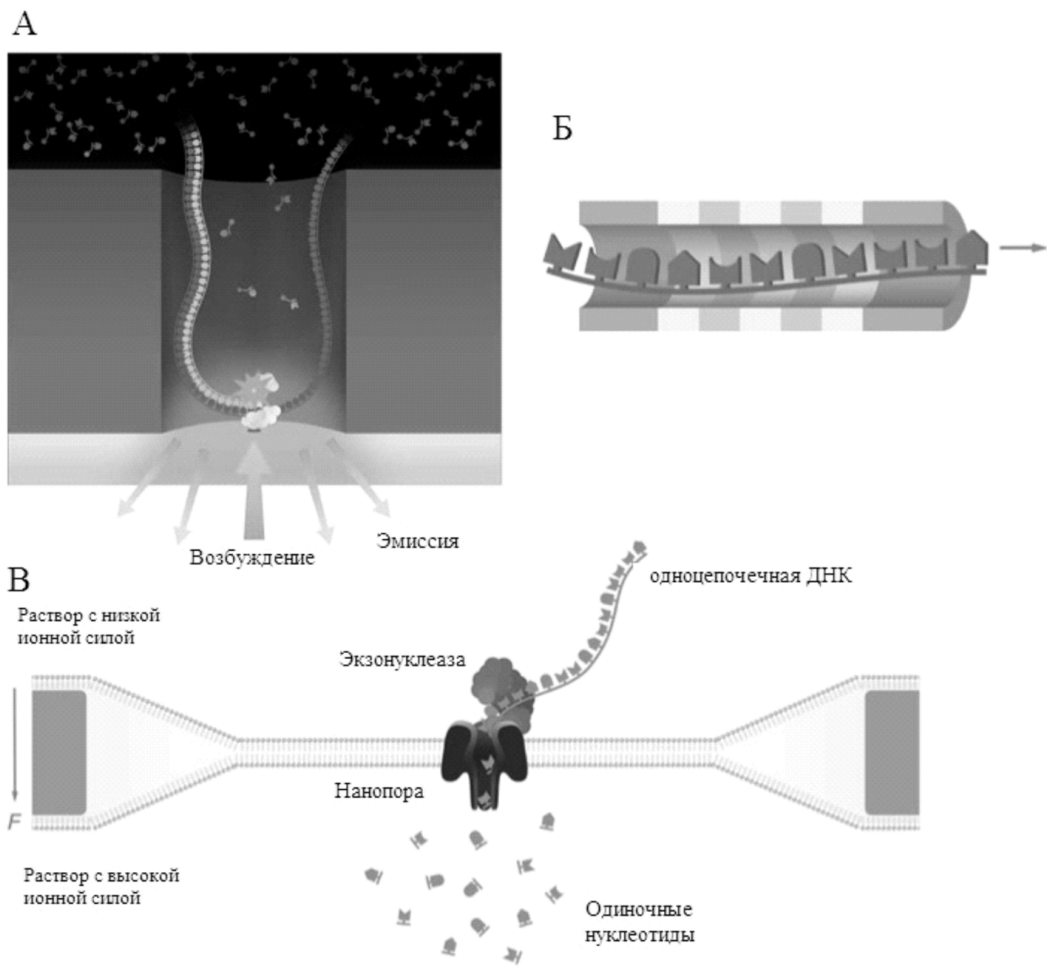

Рис. 2. Общие принципы работы некоторых систем секвенирования третьего поколения. Системы секвенирования третьего поколения определяются прямым определением последовательности одиночных молекул нуклеиновых кислот:
А — технология компании Pacific Biosciences. ДНК-полимераза иммобилизована в волноводе с нулевой фазой и включения нуклеотида регистрируется путем измерения интенсивности флуоресценции нуклеотидов, меченных по гамма-фосфату;
Б — ДНК-транзистор компании IBM, прочитывающий индивидуальные основания молекулы одноцепочечной ДНК в процессе их прохождение через твердофазную нанопору, чередующую слои металла (светло-серый) и диэлектрика (тёмно-серый);
В — технология Oxford Nanopore использует измерение параметров транслокации через пору одиночных нуклеотидов, «отрезанных» от молекулы ДНК экзонуклеазой [23]

вательность ДНК или РНК может считываться напрямую, без синтеза комплементарной цепи ДНК. Сейчас достаточно сложно оценить возможный потенциал такой методологии, однако нельзя исключать, что он лежит далеко за пределами секвенирования ДНК/РНК, а может быть применён для детекции малых молекул и биосенсорики [34]. Сейчас основные усилия по разработке технологии нанопор сконцентрированы на подборе оптимальных характеристик пор, которые будут позволять осуществлять контролируемую транслокацию и детекцию нуклеиновых кислот. Нанопоры конструируются на основе биологических, твердофазных материалов или же комбинации обоих [5]. Одна из наиболее перспективных технологий использует в качестве нанопоры а-гемолизин — трансмембранный белок золотистого стафилококка, формирующий гептамерные поровые структуры в липидном бислое [7]. Стабильность этого белка и относительно простой способ его производства, а также то, что структура этого белка известна, позволяет модифицировать его стандартными методами генной инженерии и таким образом достигать оптимальных свойств нанопор. Эта технология была приобретена компанией Oxford Nanopores (Великобритания) и в настоящий момент уже стала первой системой секвенирования на основе нанопор, представленной на современном рынке (рис. 2В). В последние годы, однако, внимание разработчиков всё больше сдвигается в сторону твердофазных нанопор на основе графена или нитрида кремния, в первую очередь, из-за экономической эффективности их производства [4]. Несомненно, детекция сигнала на основе изменения электрических параметров нанопор, из соображений простоты и стоимости, — оптимальная методика, однако системы оптической детекции также разрабатываются. Подход, разрабатываемый компанией Noblegen, требует предварительной конверсии нуклеиновых кислот в флуоресцентно меченную закодированную форму, которая может быть прочитана в виде последовательности флуоресцентных сигналов, проходящих через пору [16]. В компании IBM сейчас разрабатывается так называемый ДНК-транзистор, представляющий собой систему твердофазных нанопор, в которой транслокация молекулы ДНК через мембрану и определение последовательности осуществляются путём измерения модуляций напряжения (рис. 2Б) [13]. При этом необходимо отметить, что, несмотря на кажущуюся выигрышность многих подходов, оптимальная комбинация типа нанопор и метода детекции может быть специфической для каждого конкретного экспериментального приложения.

В настоящее время рядом лабораторий и компаний также ведутся разработки методов прямой визуализации молекул ДНК путём сканирующей туннельной или трансмиссионной электронной микроскопии. По заявлениям разработчиков, длина прочтения будет составлять несколько миллионов пар нуклеотидов, однако достоверных результатов пока не получено, о чём говорит полное отсутствие публикаций.

Применение систем секвенирования второго поколения в клинической практике

Основные приложения

Несмотря на бурное развитие таких методик, как РНК-секвенирование, секвенирование иммунопреципитированного хроматина и бисульфитное секвенирование, основным применением высокопроизводительных систем остаётся секвенирование геномной ДНК. Говоря о различных подходах к данному вопросу, стоит ввести несколько терминов, обозначающих ключевые параметры медико-генетического эксперимента по секвенированию:

а) *глубина покрытия* — количество единичных прочтений, приходящихся теоретически на каждый нуклеотид генома, рассчитываемое по формуле:

$$\text{Глубина покрытия} \quad \frac{\text{Суммарный выход прибора}}{\text{Размер генома}};$$

б) *точность единичного прочтения* — вероятность ошибки системы в каждом прочитанном нуклеотиде без учёта многократного покрытия;

в) *мультиплексирование образцов* — метод подготовки библиотек геномной ДНК, позволяющий проводить анализ геномного материала нескольких пациентов в одном запуске прибора, не проводя при этом физического разделения проточной ячейки. Это осуществляется при помощи ДНК-штрих-кодов — коротких, заранее известных последовательностей длиной 8—10 п.н., присоединяемых к геномной ДНК на этапе лигирования адаптеров. В ходе секвенирования система прочитывает эти последовательности и проводит сопоставление *прочтение—образец* в автоматическом режиме.

Медико-генетические эксперименты по секвенированию геномной ДНК условно можно разделить на две группы — полногеномное и таргетное ресеквенирование. Полногеномное ресеквенирование, как правило, применяется при полногеномных исследованиях ассоциации (GWAS) или выполняется в рамках крупных проектов по сбору данных (например, проект 1000 геномов). Такой подход подразумевает детальное описание всех мутаций и хромосомных перестроек в геноме. Ввиду того, что около 50% генома человека составляют разнообразные повторы [10], картирование коротких прочтений может быть неэффективным и требовать огромного покрытия (до 1000х). С целью решения этой проблемы используются методы секвенирования парных фрагментов. В таком подходе геномная ДНК расщепляется на крупные фрагменты заранее известной длины (размер может составлять от 1000 до 10 000 п.н.). После этого при помощи различных манипуляций из таких фрагментов готовится геномная библиотека, устроенная таким образом, что последующему секвенированию подвергаются два фрагмента, находящихся на определённом расстоянии один от другого. В таком случае даже если один из фрагментов при картировании попадёт на тот или иной повтор (что подразумевает «неоднозначное» картирование на не-

сколько участков генома), он будет иметь «якорь» в виде другого, однозначно картированного фрагмента. Поэтому он будет картироваться не на весь геном, а только на небольшой участок размером около 1000 п.н. Такой подход позволяет произвести однозначное картирование обоих прочтений и получить информацию о крупных геномных перестройках, ведь если фактически оба прочтения будут картированы на расстоянии, меньшем исходного размера фрагмента, это означает, что в геноме имеет место делеция, если на большем — инсерция. Однако, несмотря на такие ухищрения, эксперименты по полногеномному секвенированию требуют значительных экспериментальных мощностей, что связано с тем, что общепринятая минимальная глубина покрытия, достаточная для достоверной трактовки данных, составляет 30х. Это значит, что для расшифровки полного генома человека в ходе эксперимента нужно прочитать минимум 100 млрд п.н. Это определяет высокие материальные затраты и в настоящее время приводит к тому, что относительное число полногеномных проектов неуклонно снижается.

Более актуальным сейчас становится так называемое таргетное ресеквенирование — расшифровка последовательности определённых, интересующих исследователя элементов генома. Наиболее ярким примером такого подхода служит секвенирование экзома — всех кодирующих частей генов человека, которые суммарно составляют около 50 млн п.н. Вследствие того, что для определения последовательности экзома нужен гораздо меньший выход по сравнению с секвенированием полного генома, применение такой методики даже с большим покрытием делает возможным использование мультиплексирования образцов и таким образом получение необходимого результата с высокой достоверностью и экономической эффективностью.

Первым шагом в проведении таргетного ресеквенирования является обогащение образца геномной ДНК интересующими исследователя последовательностями. Для этого существуют различные подходы, которые делятся на две основные группы: обогащение путём гибридизации и обогащение посредством амплификации геномных мишеней. Выбор между ними определяется суммарной длиной таргетной последовательности — при длине, меньшей 2 млн п.н., как правило, пользуются методами, основанными на ПЦР-амплификации, при большей длине используется гибридизация. В настоящее время имеет место огромное разнообразие методов обогащения путём ПЦР: мультиплексная ПЦР [18], long-range ПЦР [20], автоматизированная эмульсионная ПЦР [33] и др. Обогащение путём гибридизации в настоящее время подразумевает фрагментацию геномной ДНК и последующую инкубацию полученных фрагментов с биотинилированными РНК-затравками. После гибридизации таргетные (находящиеся в смеси в виде РНК-ДНК дуплексов) элементы генома отделяются от остальной смеси путём магнитной иммуносепарации [26].

Следующим этапом являются стандартная подготовка геномной библиотеки с мультиплекс-адаптерами и последующее секвенирование. Выбор системы секвенирования в данном случае определяется, в первую очередь, экономической эффективностью, поскольку ввиду небольшой суммарной длины исследуемых элементов генома покрытие достаточно велико (100—200х), и поэтому такой параметр, как точность единичного прочтения, решающего значения не имеет.

Исследование моногенных заболеваний

Определение гена, ответственного за патогенез того или иного заболевания, осуществляется путём использования технологий позиционного клонирования. Эта достаточно сложная и длительная процедура начинается с установления многодетных семей с явными признаками наследования заболевания и определения его типа. Если удаётся найти семьи, отвечающие требованиям, то происходит картирование гена, для чего используется анализ сцепления, который позволяет выявить ту область, в которой расположен искомый ген. Технологии высокопроизводительного секвенирования позволяют определить последовательность всей группы сцепления и выявить ген-кандидат. Естественно, в результате расшифровки последовательности протяжённой группы сцепления обнаруживается большое количество мутаций, ассоциация которых с исследованным заболеванием определяется статистически.

Применение высокопараллельного секвенирования в выявлении функциональных мутаций, обусловливающих моногенные заболевания, во многих случаях избыточно. Это обусловлено тем, что главной проблемой в процессах определения генетических этиологических факторов моногенных заболеваний является подбор семей, отвечающих определённым критериям. И при наличии нескольких больших семей с определённым заболеванием и возможностью более узкой локализации области сцепления высокопроизводительное секвенирование может стать излишним. В то же время, некоторые работы доказывают возможность применения описываемых технологий для выявления генетических факторов патогенеза моногенных и полигенных болезней. Так, в 2010 г. было опубликовано исследование, в котором авторы установили генетическую природу невропатии Шарко—Мари—Тут. С использованием технологий параллельного секвенирования было показано, что в исследуемой семье причиной заболевания были мутации в гене *SH3TC2* в гетерозиготном состоянии. В данной работе нулевая гипотеза фактически отсутствовала, поэтому авторы прибегли к полногеномному секвенированию с использованием библиотек парных фрагментов. Как и следовало ожидать, было идентифицировано большое количество генетических вариаций: 234 вариации количества копий, около полумиллиона однонуклеотидных полиморфизмов. Из всего множества полиморфизмов были отобраны несинонимичные замены (около 149), а

из них были выделены варианты, связанные с нейродегенеративными заболеваниями. Их оказалось 54. Из этих 54 мутаций о двух имелась информация, что они связаны с развитием невропатии Шарко—Мари—Тут. Детальный сегрегационный анализ семьи позволил выявить единственную мутацию в гене *SH3TC2*, отвечающую за развитие этого заболевания. Данные, полученные при секвенировании ДНК, были подтверждены на животных моделях с нокаутированным геном *SH3TC2*, обладавших фенотипом, схожим с невропатией Шарко—Мари—Тут у человека [14]. Данная работа показательна потому, что залогом получения достоверного результата стала не только информация о генах, связанных с развитием заболевания, но и наличие обширной родословной, позволившей провести детальный сегрегационный анализ мутаций-кандидатов. Кроме того, нельзя не отметить, что в данном исследовании проведённый авторами полногеномный анализ оказался избыточным и для получения результата достаточно было бы провести секвенирование полного экзома.

Известно, что носителем патогенетической информации может быть не только геномная, но и митохондриальная ДНК, мутации в генах которой обусловливают несколько десятков как спорадических, так и наследственных патологий, передающихся по материнской линии. С технологической точки зрения, митохондриальная ДНК благодаря своей мультикопийности и небольшому размеру является очень удобным объектом для высокопроизводительного секвенирования. Это позволяет получать статистически достоверную информацию о её последовательности даже на «младших» платформах для параллельного секвенирования (например, Ion Torrent PGM или Illumina MiSeq.).

Исследование мультифакториальных заболеваний

Особую надежду на масштабные технологии секвенирования ДНК возлагают в связи с онкологическими и сердечно-сосудистыми болезнями, поскольку их патогенез ассоциирован с множественными и зачастую неописанными генетическими детерминантами. Результаты первого масштабного секвенирования опухоли были опубликованы в 2008 г. Авторы провели анализ 23 219 транскриптов, представлявших 20 661 белоккодирующий ген из 22 образцов глиобластомы, что привело к выявлению повторяющейся мутации изоцитратдегидрогеназы 1 (IDH1) у 12% пациентов преимущественно молодого возраста [21]. В 2010 г. были опубликованы две примечательные работы, в которых методы параллельного секвенирования использовались для анализа геномов мелкоклеточного рака лёгкого [23] и меланомы [22]. Было показано, что последовательность ДНК клеток меланомы содержит около 30 тыс. нуклеотидных замен, не представленных в лимфобластоидной линии того же пациента. Также наблюдалось значительное увеличение количества других изменений генома клеток — инсерций, делеций, перестановок, изменений числа копий генов.

Таким образом, несмотря на относительно недавний старт применения технологий параллельного секвенирования в исследованиях причин и механизмов онкогенеза, сегодня уже получен ряд результатов, способствующих более глубокому пониманию этого процесса. В настоящий момент Сэнгеровский институт фонда Wellcome Trust (Великобритания) проводит беспрецедентное по своим масштабам исследование геномов различных видов рака, результаты которого будут находиться в свободном доступе.

Следует отметить, что онкологические заболевания, в большинстве случаев являются результатом клональной пролиферации, что делает этот класс патологий весьма удобным для исследований методами высокопроизводительного секвенирования, но в то же время накладывает определённые ограничения на выбор системы секвенирования. Прежде всего, природа онкологических образцов подразумевает высокий уровень генетической гетерогенности материала. Это выводит на первый план такие параметры системы, как точность одиночного прочтения и общее количество прочтений, генерируемых за запуск системы. Таким образом, детальные исследования природы опухолей подразумевают использование «больших» систем секвенирования (SOLiD 5500xl, Illumina HiSeq), что определяет тот факт, что приоритет в исследовании рака отдаётся в первую очередь крупнейшим геномным центрам.

Другой областью онкогенетических исследований является поиск генов предрасположенности к раковым заболеваниям. Такие работы имеют более направленный характер и, как правило, характеризуются наличием известных генов-кандидатов. В настоящее время наиболее эффективным методом является обогащение генетического материала пациента с использованием высокомультиплексированной ПЦР участков генома, прилежащих к описанным полиморфизмам [9]. Это позволяет с минимальными временными затратами осуществлять скрининг пациентов по большому количеству локусов.

Другой группой нозологий, важность исследования которой трудно переоценить, являются сердечно-сосудистые болезни, занимающие, наряду с раком, ведущие позиции в списках основных причин смерти. Технологии параллельного секвенирования применяются для исследования генетических причин сердечно-сосудистых заболеваний не менее широко, однако, в целом, на сегодняшний день говорить о более глубоком понимании механистических причин возникновения сердечно-сосудистых патологий пока рано. Главным образом, это связано с большей комплексностью системных нарушений, индуцирующих патогенез сердечно-сосудистых заболеваний, участием в этом процессе большого количества различных типов клеток и тканей. Несмотря на это, попытки понять генетическую природу патологий данного типа предпринимаются различными исследовательскими группами.

В 2011 г. группой Медера (Meder) была предложена недорогая (в пересчёте на одну мутацию) стратегия диагностики кардиомиопатий, которая включала в себя специфическое обогащение геномного материала последо-

вательностями экзонов генов, ассоциированных с различными видами кардиомиопатий (суммарно 1092 экзона). За один запуск прибора было проанализировано 5 пациентов с дилатационной и 5 пациентов с гипертрофической кардиомиопатией, что позволило выявить 57 548 вариантов последовательностей ДНК, 459 из которых были несинонимичными заменами. Кроме того, в кодирующих частях исследуемых генов было выявлено 809 малых инсерций/делеций. Путём дальнейшего сопоставления полученных результатов с базой данных известных мутаций, вызывающих развитие кардиомиопатий, а также предиктивного анализа in silico функциональности несинонимичных замен, мутации, предположительно вызывающие развитие заболевания, были выявлены у 6 из 10 пациентов. Очевидно, что данное исследование не столько дало какую-либо информацию о механизмах патогенеза, сколько подтвердило факт комплексности сердечно-сосудистых заболеваний. Таким образом, полногеномные исследования мультифакториальных заболеваний в настоящий момент являются, по сути, масштабным методом идентификации генетических маркёров, но никак не универсальным приёмом изучения механизмов патогенеза. Так, например, исследование редких полиморфных вариантов на большой выборке пациентов с аневризмой аорты позволило выявить 700 почти уникальных полиморфизмов, статистически достоверно ассоциируемых с данным заболеванием. При этом, с точки зрения авторов, такое обилие различных вариантов затрудняет определение истинных механизмов и может приводить к большому количеству ложноположительных результатов [6]. В то же время, учитывая тот факт, что каждый геном содержит в среднем 150 тыс. индивидуальных полиморфизмов, не аннотированных в базе данных SNP, количество полиморфных вариантов, ассоциированных с аневризмой аорты, не кажется таким уж большим и при функциональной аннотации каждого из полиморфизмов эти данные могут быть использованы для верификации рисков.

Сегодня, после первого этапа массового применения технологий высокопроизводительного секвенирования, очевидно, что его применение в клинической практике не так очевидно, как это казалось ранее. Этому есть несколько причин.

1. Предиктивная значимость отдельных обнаруживаемых полиморфизмов, хотя и демонстрирует достоверные значения, всё ещё имеет значительный межпопуляционный разброс. Это не позволяет использовать их в качестве однозначных генетических детерминант, как используются, например, мутации в генах *BRCA1* и *BRCA2* для верификации рисков рака груди и яичников;

2. Большинство полиморфизмов, определяющих развитие мультифакториальных заболеваний, являются патогенетическими только при взаимодействии с другими факторами, природа которых часто неизвестна;

3. Отсутствие функциональной аннотации большинства полиморфизмов, выявляемых методами высокопроизводительного секвенирования, не позволяет делать однозначных выводов об их ассоциации с клиническим фенотипом даже при удовлетворительных статистических показателях;

4. Существуют и системные проблемы, не позволяющие напрямую транслировать получаемые геномные данные в повседневную клиническую практику. Главная из них — это отсутствие специалистов, одинаково знакомых как с геномикой, так и с клиникой;

5. Необходимо создавать глобальные базы данных, в которых будет депонироваться информация об ассоциативных связях геномных данных и фенотипов пациентов наряду с данными о влиянии той или иной схемы лечения (подобные базы существуют, однако их функциональность не позволяет оперировать необходимыми массивами информации);

6. Необходимо проводить подготовку специалистов, основной компетенцией которых будет трансляция геномики в рутинную клиническую практику.

Таким образом, применение методов полногеномного секвенирования в исследованиях генетических факторов развития мультифакториальных патологий, по сути, представляет собой способ массивного сбора данных, ощутимые клинические результаты которых будут видны лишь в весьма отдалённой перспективе.

Заключение

Сейчас ни для кого не составляет сомнений тот факт, что технологии секвенирования второго поколения совершили революцию в современных медико-генетических исследованиях. Большинство существующих платформ позволяет определить полную последовательность генома человека за несколько недель и с разумными по сравнению с программой «Геном человека» материальными затратами. Однако очевидно, что огромный прогресс, сделанный за последние 5—7 лет, является только началом технологической революции в современной геномике. С развитием систем секвенирования третьего поколения, позволяющих получать достоверный результат, применяя минимальное количество манипуляций с генетическим материалом, высокопараллельное секвенирование станет таким же рутинным методом диагностики, каким сейчас является полимеразная цепная реакция.

Сейчас же, несмотря на существование многочисленных подходов, позволяющих сузить область поиска значимых мутаций (таргетное ресеквенирование), основной проблемой остаются анализ огромных массивов данных и определение причинно-следственной взаимосвязи между мутацией и развитием заболевания. Это говорит о том, что полногеномное секвенирование само по себе является лишь мощным инструментом, который может быть успешно применён лишь при тщательном планировании и предварительном моделировании медико-генетического эксперимента.

Список литературы

1. Brenner S. et al. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays // Nature biotechnology. — 2000. — 18. — P. 630—634.
2. Chen W. et al. Mapping translocation breakpoints by next-generation sequencing // Genome research. — 2008. — 18. — P. 1143—1149.
3. Flusberg B.A. et al. Direct detection of DNA methylation during single-molecule, real-time sequencing // Nature methods. — 2010. — 7. — P. 461—465.
4. Garaj S. et al. Graphene as a subnanometre trans-electrode membrane // Nature. — 2010. — 467. — P. 190—193.
5. Hall A.R. et al. Hybrid pore formation by directed insertion of α-haemolysin into solid-state nanopores // Nature nanotechnology. — 2010. — 5. — P. 874—877.
6. Harakalova M. et al. Genomic DNA pooling strategy for next-generation sequencing-based rare variant discovery in abdominal aortic aneurysm regions of interest-challenges and limitations // Journal of cardiovascular translational research. — 2011. — 4. — P. 271—280.
7. Howorka S., Cheley S., Bayley H. Sequence-specific detection of individual DNA strands using engineered nanopores // Nature biotechnology. — 2001. — 19. — P. 636—639.
8. Kasianowicz J.J., Brandin E., Branton D., Deamer D.W. Characterization of individual polynucleotide molecules using a membrane channel // Proceedings of the National Academy of Sciences of the United States of America. — 1996. — 93. — P. 13770—13773.
9. Ku C.S. et al. Technological advances in DNA sequence enrichment and sequencing for germline genetic diagnosis // Expert Rev. Mol. Diagn. — 2012. — 12(2). — P. 159—173.
10. Lander E.S. et al. Initial sequencing and analysis of the human genome // Nature. — 2001. — 409(6822). — P. 860—921.
11. Lander E.S. Initial impact of the sequencing of the human genome // Nature. — 2011. — 470, — P. 187—197.
12. Levy S. et al. The diploid genome sequence of an individual human // PLoS biology. — 2007. — 5. — e254.
13. Luan B. et al. Base-by-base ratcheting of single stranded DNA through a solid-state nanopore // Physical review letters. — 2010. — 104, 238103.
14. Lupski J.R. et al. Whole-genome sequencing in a patient with Charcot—Marie—Tooth neuropathy // The New England journal of medicine. — 2010. — 362. — P. 1181—1191.
15. Margulies M. et al. Genome sequencing in microfabricated high-density picolitre reactors // Nature. — 2005. — 437. — P. 376—380.
16. McNally B. et al. Optical recognition of converted DNA nucleotides for single-molecule DNA sequencing using nanopore arrays // Nano letters. — 2010. — 10. — P. 2237—2244.
17. Meder B. et al. Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies // Circulation. Cardiovascular genetics. — 2011. — 4. — P. 110—122.
18. Michils G. et al. Molecular analysis of the breast cancer genes *BRCA1* and *BRCA2* using amplicon-based massive parallel pyrosequencing // J. Mol. Diagn. — 2012. — 14(6). — P. 623—630.
19. Nunnally B.K., He H., Li L.C., Tucker S.A., McGown L.B. Characterization of visible dyes for four-decay fluorescence detection in DNA sequencing // Analytical chemistry. — 1997. — 69. — P. 2392—2397.
20. Ozcelik H. Long-range PCR and next-generation sequencing of *BRCA1* and *BRCA2* in breast cancer // J. Mol. Diagn. — 2012. — 14(5). — P. 467—475.
21. Parsons D.W. et al. An integrated genomic analysis of human glioblastoma multiforme // Science (New York, N.Y.). — 2008. — 321. — P. 1807—1812.

Next generation sequencing technologies in medical genetic studies

Toshchakov S.V., Dominova I.N., Patrushev M.V.

Immanuel Kant Baltic Federal University,
236041, Russia, Kaliningrad, A.Nevskogo str., e-mail: post@kantiana.ru

In last 5 years next generation sequencing technologies made a huge impact on research of genetic aspects of pathogenesis and initiated the evolvement of the science of medical genomics. At the moment we can observe multiple attempts of implementation of high-throughput sequencing in clinical diagnostics. Nevertheless those attempts might be meaningless without deep knowledge of such techniques, especially of its development, specificity and possible pitfalls. In this review we take a detailed overlook of existing and developing technologies and discuss most spectacular examples of its applications in medical genomics and its potential for implementation in routine diagnostic laboratory practice.

Key words: genomics, sequencing, next generation sequencing technologies, nanopores, biomarkers of multifactorial diseases