# ОГЛАВЛЕНИЕ

# ПРЕДИСЛОВИЕ Е. Д. СВЕРДЛОВА

Нельзя не заметить, что методы молекулярной биологии со временем становятся все сложнее и дороже, и, вместе с тем, теряют в разнообразии. Многие задачи на этапе планирования эксперимента сводятся к типовым методическим шагам, выполняемым по принципу «заказа услуг на стороне». Уход сложных и дорогостоящих методов в сервисные центры, наряду с очевидными преимуществами такой специализации, имеет и ряд негативных последствий. Поверхностное знание методов исследования, провал между биологическим материалом и данными в компьютере приводят к непониманию границы возможностей используемых методик.

Стремительное развитие методов секвенирования в последние годы привело к ощущению, что с их помощью можно решить любые задачи генетики. Тем не менее высокопроизводительное секвенирование, как и любой другой метод, имеет ряд ограничений. Так, вопреки распространенному мнению, NGS не является панацеей при исследовании мультифакторных заболеваний и лишь помогает чуть быстрее выполнить определенные методические шаги.

Данная книга пытается сдержать растущий провал между объектом и анализируемыми данными, подробно и с практическими рекомендациями, перечисляя все этапы технологии NGS.

*Е. Д. Свердлов,*
*академик РАН и РАСХН, советник РАН*

# ПРЕДИСЛОВИЕ М. С. ГЕЛЬФАНДА

Как говорил один известный политический деятель, это «очень своевременная книга». Современные секвенаторы, наконец, начали появляться в российских лабораториях. Некоторые исследователи заранее знали, что они собираются изучать при помощи этих приборов, и готовились к их появлению, многие – лишь заполучив чудесную машинку, задумались, к чему же ее применить, третьи – только рассматривают возможность закупки в приложении к своим текущим задачам. Книга будет полезна исследователям из всех трех категорий. Первым – как набор ссылок на современные методы анализа данных и подготовки материалов, третьим – как пособие по выбору адекватной платформы и сборник практических советов, вторым же (если им вообще что-то может помочь) – как намек на то, что интересное можно было бы сделать.

Книга хорошо сбалансирована. В ней есть история методов секвенирования, биофизические и биохимические основы современных технологий, сравнение возможностей и недостатков платформ, описаны основы пробоподготовки, сведения о методах биоинформатического анализа получаемых данных, типичные задачи, решаемые при помощи таких приборов. Описания лаконичные, но точные, а обильные ссылки дадут возможность заинтересованному читателю глубже изучить конкретные проблемы. Книга глубоко погружена в современный российский контекст, и в ней имеются советы, которые не встретишь в стандартном обзоре: от необходимости проверить надежность энергоснабжения научного учреждения до правильной планировки серий экспериментов с целью экономии расходных материалов и организации совместной работы экспериментаторов и биоинформатиков и т. п.

Думаю, что эта книга необходима в каждой молекулярно-биологической лаборатории. В духе авторов добавлю, что желательно в нескольких экземплярах, один из которых будет храниться в сейфе завлаба и выдаваться под расписку.

*Михаил Гельфанд, д-р биол. наук,*
*профессор, член Academia Europaea*

# ПРЕДИСЛОВИЕ АВТОРОВ

*Учителю*
*Льву Абрамовичу Остерману*
*посвящается*

Развитие науки базируется на методах исследования. Создание новых технологий всегда приводило к прорыву в определенной области знаний. Причем зачастую развитие какого-то методического направления неожиданно дает эффект в иной (даже не смежной) научной области. Бурное развитие цитологии в какой-то момент стало следствием прогресса в области изготовления стеклянных линз. Появление методов высокопроизводительного секвенирования (next generation sequencing, NGS) стало возможно благодаря развитию компьютерной индустрии, технологий изготовления микропроцессоров и цифровых носителей информации. Оказалось, что эти же элементы могут быть использованы в совершенно ином назначении: для работы с биологическими макромолекулами. Так синтез микроэлектроники и биохимии дал новый метод исследования живых систем – секвенирование второго поколения. Вместе с тем вовремя подоспели адекватные вычислительные мощности для обработки получаемых данных.

Следует отметить, что кроме использования наработок из микроэлектроники технологии NGS включают в себя и ряд предшествующих молекулярно-биологических методик, в частности полимеразную цепную реакцию (ПЦР) и гибридизацию на микрочипах.

Изобретение и внедрение в практику технологий высокопроизводительного секвенирования вывело на новый уровень такие направления науки, как генетика, молекулярная биология, дало стимулы для становления персонализированной медицины. Сегодня область высокопроизводительного секвенирования объединяет широкую гамму различных технологий, базирующихся на разных принципах и разработанных более или менее независимо. В этой книге коллектив авторов, имеющих собственный опыт работы с технологиями NGS, излагает принципы основных современных методов высокопроизводи-

тельного секвенирования. Рассмотрены особенности наиболее популярных технологий секвенирования, формат типовых задач для NGS, варианты обработки биоинформатических данных, стандартные ошибки каждого из этапов исследования. Авторы постарались структурировать и систематизировать существующие подходы, сделав акцент на общих принципах и существенных отличиях.

Несколько слов о терминах. В 2011 году, в предисловии к переводу девятого издания книги Б. Льюина «Гены» редактор отечественного издания сказал по этому поводу, что «...по прошествии 3–5 лет в зоне.ру килобазы вытеснят т. п. н., а последовательность нуклеотидов окончательно превратится в сиквенс». Пожалуй, наступил тот момент, когда написать книгу об NGS, не используя термин «секвенирование», сложнее, чем согласиться с его появлением в русском языке. Однако не все позиции сданы авторами без боя, «т. п. н-ы» пока остались, а при выборе между «штрих-кодом» и «бар-кодом» предпочтение отдано первому варианту.

Отдельно следует сказать о термине «NGS». В лабораторной практике строжайше запрещено указывать на емкости с реагентом «new», ввиду неинформативности данного обозначения. Авторы сочли некорректным использование термина «секвенирование следующего поколения» – буквального перевода с английского (next generation sequencing) – для обозначения современных технологий секвенирования, прямо указывая на поколение методов или обозначая их как высокопроизводительные (что, безусловно, тоже относительно).

Авторы выражают благодарность коллегам, помогавшим на разных этапах подготовки рукописи: Дмитрию Алексееву (ФГБУН НИИ ФХМ, Москва), Андрею Гаража (ООО «Первый онкологический научно-консультационный центр», Москва), Игнатию Клесниченко (ООО «Бином», Москва), Николаю Равину (Центр «Биоинженерия» РАН, Москва), Владиславу Трошину (ООО «Троицкий инженерный центр», Троицк), Сергею Науменко (МГУ имени М.В. Ломоносова, Москва), Петру Шаталову (ООО «Генотек», Москва).

# ПЕРЕЧЕНЬ КОМПАНИЙ, УПОМЯНУТЫХ В ТЕКСТЕ

23andMe
454 Life Sciences
Affymetrix
Agilent Technologies
Amazon
Ambion
Applied Biosystems
Bio-Rad
Celera
Councyl
Covaris
CuraGen Corporation
Digilab
DNA Electronics Ltd.
Dover
Fluidigm Corporation
Helicos Biosciences
Illumina
JewishCare
Life Technologies Thermo Fisher Scientific
Lynx Therapeutics
Nanopore
New England Biolabs
NimbleGen
Pacific Biosciences
Pathway Genomics
Perlegen
Promega
Qiagen
RainDance Technologies
Roche
Sage Science
Solexa
ZS Genetics
Генотек
ДНК-Технология
Евроген

# СПИСОК СОКРАЩЕНИЙ

| | |
|---|---|
| 16S рРНК | РНК малой субъединицы бактериальной рибосомы |
| CGH | comparative genomic hybridization, сравнительная геномная гибридизация |
| BAC | bacterial arifichial chromosome, искусственная бактериальная хромосома |
| cffDNA | cell-free fetal DNA, внеклеточная ДНК плода |
| ChIP | chromatin immunoprecipitation, иммунопреципитация хроматина |
| Indel | insertion/deletion, вставка/делеция |
| MALDITOF | matrix-assisted laser desorption/ionization time-of-flight, матрично-активированная лазерная десорбция/ионизация с регистрацией времени пролета частиц |
| MIP | molecular inversion probe, инвертированная молекулярная проба |
| NGS | next-generation sequencing, секвенирование следующего поколения |
| OLC | overlap-layout-consensus, перекрытие–расположение–согласованность |
| P32 | изотоп фосфора-32 |
| SBH | sequencing by hybridisation, секвенирование путем гибридизации |
| SNP | single nucleotide polymorphism, однонуклеотидный полиморфизм |
| SVs | structured variations, большие структурные вариации |
| А (A) | аденин |
| Г (G) | гуанин |
| ддНТФ | дидезоксирибонуклеозидтрифосфат |
| ДНК | дезоксирибонуклеиновая кислота |
| дНТФ | дезоксирибонуклеозидтрифосфат |
| кДНК | комплементарная ДНК |
| НК | нуклеиновая кислота |
| п. н. | пара нуклеотидов |
| ПЗС-матрица | считывающая матрица прибора с зарядовой связью |
| ПЦР | полимеразная цепная реакция |
| РНК | рибонуклеиновая кислота |
| Т (T) | тимин |
| т. п. н. | тысяча пар нуклеотидов |
| Ц (C) | цитозин |

ГЛАВА 1

# ОБЗОР МЕТОДОВ ОПРЕДЕЛЕНИЯ ПОСЛЕДОВАТЕЛЬНОСТИ НУКЛЕИНОВЫХ КИСЛОТ

В данной главе приводится краткий обзор различных подходов к определению последовательности нуклеиновых кислот. К настоящему моменту можно выделить три поколения технологий секвенирования. К первому поколению относят изобретенные в середине 70-х годов XX века методы химической деградации (метод Максама–Гилберта) и остановки полимеразы на дидезоксинуклеотидах (метод Сенгера). Вторым поколением принято считать коммерческие технологии высокопроизводительного секвенирования, разработанные в середине 1990-х, хоть и основанные на разных принципах, но всегда требующие получения сигнала от множества одинаковых молекул ДНК. В настоящее время на рынок выходят технологии, способные регистрировать сигнал от единственной исследуемой молекулы нуклеиновой кислоты. В некоторых публикациях такие подходы стали называть секвенированием третьего поколения. Далее мы будем использовать лишь термины «NGS» и «высокопроизводительное секвенирование» (как равнозначные), объединяя под ними технологии второго и третьего поколений.

Сразу отметим, что методы определения последовательности РНК пока недостаточно эффективны (технологии секвенирования одиночных молекул, позволяющие работать непосредственно с РНК, только начали появляться, см. разд. 3.5 и 3.6). В то же время превращение РНК в ДНК путем обратной транскрипции настолько стандартно, что в настоящее время для определения последовательности РНК исследователи почти всегда используют секвенирование кДНК.

Мы постарались дать максимально широкий спектр подходов к определению последовательности ДНК, несмотря на то, что лишь некоторые из них к настоящему моменту нашли применение в высокопроизводительном секвенировании (коммерческие технологии NGS более подробно описаны в главе 3).

## 1.1. МЕТОДЫ, ОСНОВАННЫЕ НА ДЕТЕКЦИИ СИГНАЛА ОТ МНОЖЕСТВА ОДИНАКОВЫХ МОЛЕКУЛ ДНК (МЕТОДЫ С ПРЕДВАРИТЕЛЬНОЙ АМПЛИФИКАЦИЕЙ ФРАГМЕНТОВ ДНК)

Большинство современных методов молекулярной биологии предполагает использование множества идентичных макромолекул для получения детектируемого сигнала. К ним относятся различные виды хроматографии, рентгеноструктурный анализ, масс-спектрометрия и т. д. Секвенирование ДНК также требует усиления сигнала за счет использования в анализе множества одинаковых молекул ДНК. Ниже рассмотрены подходы с предварительной амплификацией ДНК (путем обычного или *in vitro* клонирования) для получения миллионов идентичных фрагментов, забираемых в дальнейший анализ.

### 1.1.1. Метод Максама–Гилберта (химическая деградация)

В середине 70-х годов XX века исследователями Гарвардского университета (США) Алланом Максамом и Уолтером Гилбертом был разработан метод определения последовательности нуклеотидов, основанный на нуклеотид-специфичной химической деградации при обработке ДНК различными химическими агентами [1]. На первом этапе образец ДНК, обычно представляющий собой сравнительно короткий (100–1000 п. н.) гомогенный фрагмент (полученный, например, вырезанием «полосы» из геля после электрофоретического разделения расщепленной эндонуклеазами плазмиды), с одного из концов метят радиоактивной меткой. Затем образец разделяют на четыре части, после чего каждую из частей обрабатывают своим реагентом, приводящим к гидролизу ДНК по конкретным основаниям (или сочетаниям оснований). Параметры каждой реакции подбирают таким образом, чтобы гидролиз проходил не полностью, а лишь по некоторым позициям в каждой молекуле ДНК (в среднем желательно получить одну модификацию на отдельную молекулу). В результате получают набор «расщепленных» фрагментов ДНК, соответствующих по длине местам нахождения нуклеотидов данного типа (рис. 1.1). Например, реакция определения положения гуанина выглядит так: при помощи диметилсульфата проводят метилирование ДНК, в результате которого гуанин метилируется по положению 3, а аденин – по

положению 7. Дальнейшая обработка соляной кислотой при 0° С приводит к выпадению из цепи метиладенина (апуринизации по остаткам аденина). Такую ДНК с «пустыми» остатками дезоксирибозы в позициях, где был аденин, можно гидролизовать при нагревании в щелочи. Гидролиз в случае с метилгуанином осуществляют при помощи пиперидина. Модификации по пиримидиновым (Ц и Т) основаниям проводят с гидразином. Если реакцию проводить в присутствии NaCl, модификация затронет только Ц. Гидролиз обработанной гидразином ДНК проводят пиперидином [2].

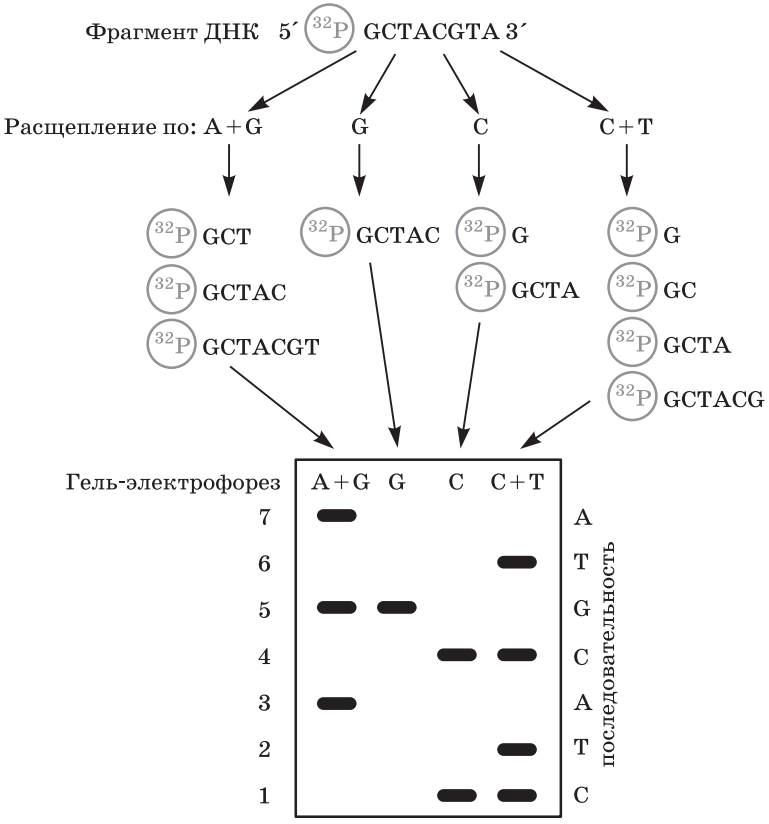


**Рис. 1.1.** Принцип метода Максама–Гилберта. Расщепление одинаковых, помеченных с одного из концов, фрагментов ДНК по разным позициям дает фрагменты разной длины, затем фрагменты могут быть разделены при помощи электрофореза в геле

После обработки все четыре образца наносят параллельно в денатурирующий полиакриламидный гель и проводят электрофорез так, чтобы получить разделение фрагментов, отличающихся на один нуклеотид. Далее с помощью рентгеновской пленки получают изображение (электрофореграмму), по которому можно восстановить последовательность нуклеотидов исследуемого фрагмента ДНК, отсчитывая, в какой из четырех дорожек оказался фрагмент, следующий за самым легким продуктом, наиболее удаленным от лунок в геле. Таким образом удается определить до 200 нуклеотидов за одно прочтение.

В настоящее время метод почти не используют ввиду сложности подготовки образцов ДНК и работы с вредными химическими веществами. Даже несмотря на появление в начале 1990-х годов автоматических секвенаторов, основанных на технологии Максама–Гилберта и существенно упростивших пробоподготовку, этот подход в итоге проиграл методу Сенгера (метод терминаторов, см. разд. 1.1.2). Преимуществами метода Максама–Гилберта (в сравнении с методом Сенгера) являются полная его независимость от вторичных структур и отсутствие необходимости знания участка последовательности интересующей ДНК (для отжига необходимой ферменту ДНК-полимеразе затравки), что позволяет избежать стадии клонирования. До последнего времени метод Максама–Гилберта использовали в случаях, когда фермент ДНК-полимераза (используемый в методе Сенгера) не мог пройти через вторичную структуру, например псевдоузел.

### 1.1.2. Метод Сенгера (остановка синтеза ДНК ферментом на дидезоксинуклеотидах)

В 1975 году, двумя годами ранее описанного выше метода Максама–Гилберта, Фредериком Сенгером и Аланом Кулзоном из лаборатории молекулярной биологии в Кембридже (Великобритания) был предложен метод определения последовательности ДНК, основанный на использовании ДНК-полимеразы и радиоактивно меченых нуклеотидов, названный авторами «плюс-минус секвенирование» [3]. Через два года Сенгер усовершенствовал технологию, создав метод дидезокситерминаторов (впоследствии получивший название «метод Сенгера») [4], а спустя всего три года, в 1980 году, Фредерик Сенгер за эту работу был удостоен Нобелевской премии по

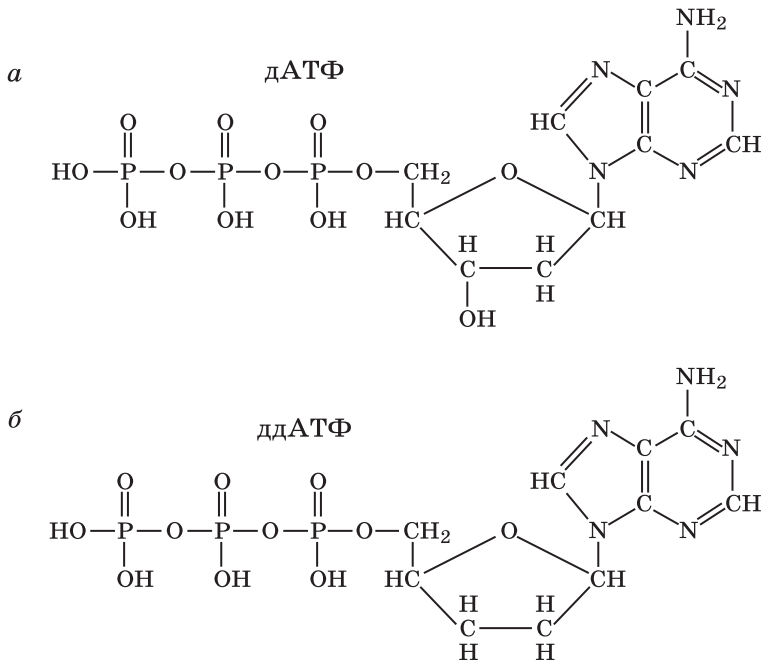


**Рис. 1.2.** Структурные формулы нуклеотидов, используемых для обычного синтеза (*а*) и остановки синтеза ДНК (*б*)

химии (которую он разделил с Уолтером Гилбертом, награжденным за метод химической деградации).

Основной идеей метода является использование модифицированных «нуклеотидов» – дидезоксинуклеозидтрифосфатов (ддНТФ) (рис. 1.2). В отличие от обычного субстрата ДНК-полимеразы дезоксинукдеозидтрифосфатов (дНТФ), они не несут ОН-группу в 3′-положении дезоксирибозы и вследствие этого не способны к присоединению полимеразой следующего нуклеотида. Участок ДНК, последовательность которого необходимо определить, добавляется в реакцию, технически похожую на обычную полимеразную цепную рекацию (ПЦР): в пробирке имеются термостабильная ДНК-полимераза, дНТФ всех четырех типов, а также олигонуклеотид, выступающий в качестве затравки для синтеза новой цепи. Помимо этих компонентов, в концентрации, примерно в 20 раз меньшей чем дНТФ, присутствуют четыре соот-

Коллектив авторов — сотрудники научного подразделения ЗАО «НПФ ДНК-Технология», ведущего в нашей стране разработчика оборудования и реагентов для медицинской ДНК-диагностики.

В книге в полной мере освещаются особенности методов определения структуры нуклеиновых кислот, дается точная, написанная доступным языком картина процессов, идущих в реакционной пробирке.
В первую очередь, книга предназначена сотрудникам научно-исследовательских лабораторий и студентам, стремящимся разобраться в фундаментальных принципах и особенностях высокопроизводительного секвенирования. Вместе с тем, каждая глава содержит много конкретных рекомендаций, делая книгу незаменимым практическим руководством для работников сиквенсного центра.