

# NGS

ВЫСОКОПРОИЗВОДИТЕЛЬНОЕ  
СЕКВЕНИРОВАНИЕ



ИЗДАТЕЛЬСТВО

**БИНОМ**

# NGS

## ВЫСОКОПРОИЗВОДИТЕЛЬНОЕ СЕКВЕНИРОВАНИЕ

Под редакцией  
д-ра биол. наук *Д. В. Ребрикова*



Москва  
БИНОМ. Лаборатория знаний

УДК 577.21+616  
ББК 28.04  
N10

А в т о р ы:

Д. В. Ребриков, Д. О. Коростин, Е. С. Шубина,  
В. В. Ильинский

NGS: высокопроизводительное секвенирование /  
N10 Д. В. Ребриков [и др.] ; под общей редакцией Д. В. Реб-  
рикова. — М. : БИНОМ. Лаборатория знаний, 2014. —  
232 с. : ил.

ISBN 978-5-9963-1784-4

Рассмотрены различные варианты и особенности современных методов определения структуры нуклеиновых кислот (методов секвенирования второго и третьего поколений). Описаны принципы наиболее популярных технологий высокопроизводительного секвенирования (NGS). Дана классификация высокопроизводительных методов секвенирования по нескольким параметрам. Приведены основные элементы первичного анализа данных масштабного секвенирования. Отдельные главы посвящены применению NGS для решения различных биологических задач: секвенирования про- и эукариотических геномов и транскриптомов, метагеномного секвенирования, использования NGS в медицинской практике.

Для сотрудников генно-инженерных и медицинских диагностических лабораторий, а также для преподавателей и студентов, специализирующихся в области молекулярной биологии и биотехнологии.

УДК 577.21+616  
ББК 28.04

---

*Научное издание*

Ребриков Денис Владимирович  
Коростин Дмитрий Олегович  
Шубина Екатерина Сергеевна  
Ильинский Валерий Владимирович

**NGS: ВЫСОКОПРОИЗВОДИТЕЛЬНОЕ СЕКВЕНИРОВАНИЕ**

Главный художник *И. Е. Марев*  
Художественный редактор *Н. А. Новак*  
Технический редактор *Е. В. Денюкова*. Корректор *Е. Н. Клитина*  
Компьютерная верстка: *Л. В. Катуркина*

Подписано в печать 12.04.14. Формат 60×90/16.

Усл. печ. л. 14,5. Тираж 1200 экз. Заказ 2045

Издательство «БИНОМ. Лаборатория знаний»

125167, Москва, проезд Аэропорта, д. 3

Телефон: (499) 157-5272, e-mail: binom@Lbz.ru, <http://www.Lbz.ru>

---

ISBN 978-5-9963-1784-4

© БИНОМ. Лаборатория знаний, 2014

Отпечатано способом ролевой струйной печати  
в ОАО «Первая Образцовая типография» Филиал «Чеховский Печатный Двор»  
142300, Московская область, г. Чехов, ул. Полиграфистов, д. 1  
Сайт: [www.chpd.ru](http://www.chpd.ru), E-mail: [sales@chpd.ru](mailto:sales@chpd.ru), т/ф. 8(496)726-54-10

# ОГЛАВЛЕНИЕ

Предисловие Е. Д. Свердлова .....	7
Предисловие М. С. Гельфанда .....	8
Предисловие авторов .....	9
Перечень компаний, упомянутых в тексте .....	11
Список сокращений .....	12
Глава 1. Обзор методов определения последовательности нуклеиновых кислот .....	13
1.1. Методы, основанные на детекции сигнала от множества одинаковых молекул ДНК (методы с предварительной амплификацией фрагментов ДНК) ..	14
1.2. Методы, основанные на детекции сигнала от одной молекулы ДНК (секвенирование одиночных молекул ДНК) .....	34
1.3. Другие методы секвенирования .....	40
Список литературы .....	40
Глава 2. Технологии создания библиотек фрагментов ДНК для NGS .....	43
2.1. Очистка нуклеиновых кислот для NGS .....	45
2.2. Оценка концентрации нуклеиновых кислот и полногеномная амплификация (WGA) .....	46
2.3. Способы разрушения ДНК для приготовления библиотеки .....	47
2.4. Оценка длин фрагментов ДНК .....	51
2.5. Присоединение адаптеров .....	52
2.6. Предварительная амплификация библиотеки .....	53
2.7. Отбор фракции фрагментов нужной длины (size-select) ..	53
2.8. Мечение смешиваемых образцов специфичными адаптерами («штрих-кодирование») ....	56
2.9. Клональная амплификация фрагментов ДНК .....	57
2.10. Типы библиотек фрагментов ДНК для NGS .....	60
Список литературы .....	65

<b>Глава 3. Коммерческие технологии высокопроизводительного секвенирования</b> .....	<b>66</b>
3.1. Технология 454 Life Sciences компании Roche (эмульсионная ПЦР + пиросеквенирование) .....	66
3.2. Технология SOLiD компании Life Technologies Thermo Fisher Scientific (эмульсионная ПЦР + секвенирование лигированием).....	69
3.3. Illumina Genome Analyser компании Illumina (мостиковая ПЦР + секвенирование синтезом) .....	71
3.4. Платформы Ion PGM и Ion Proton компании Life Technologies Thermo Fisher Scientific (эмульсионная ПЦР + полупроводниковое секвенирование) .....	74
3.5. Платформа PacBio компании Pacific Biosciences (секвенирование синтезом одиночных молекул) .....	78
3.6. Платформа Heliscope компании Helicos Biosciences (секвенирование синтезом одиночных молекул) .....	80
Список литературы .....	84
<b>Глава 4. Общие принципы обработки данных NGS</b> .....	<b>85</b>
4.1. Оценка качества первичных данных .....	85
4.2. Сборка геномов <i>de novo</i> .....	89
4.3. Алгоритмы сборки .....	91
4.4. Аппаратные и биологические особенности данных NGS .....	94
4.5. Объединение контигов в скэффолды .....	97
4.6. Вариации в близкородственных геномах .....	100
4.7. Картирование прочтений при повторном секвенировании.....	101
4.8. Поиск однонуклеотидного полиморфизма (SNP) .....	104
4.9. Поиск структурных вариаций: протяженных вставок, делеций, инверсий и транслокаций .....	105
4.10. Аннотация обнаруженных вариаций с использованием баз данных.....	106
4.11. Предсказание функциональных и клинически значимых изменений белка на основе обнаруженных мутаций.....	107
Список литературы .....	108
<b>Глава 5. Оборудование и программные решения для обработки данных NGS</b> .....	<b>112</b>
5.1. Локальные центры обработки данных NGS: архитектура и программные решения.....	112
5.2. Программное обеспечение для локального центра обработки данных NGS .....	116

5.3. Сетевые сервисы и простые решения для обработки данных NGS .....	117
5.4. Специализированные проекты по обработке данных NGS.....	120
Список литературы .....	121
<b>Глава 6. Планирование эксперимента с использованием NGS .....</b>	<b>122</b>
6.1. Общие принципы планирования биологических экспериментов.....	122
6.2. Рандомизация в NGS.....	123
6.3. Повторности в NGS .....	124
6.4. Основные типы ошибок при секвенировании.....	125
6.5. Варианты применения NGS .....	126
Список литературы .....	127
<b>Глава 7. Секвенирование индивидуальных геномов и транскриптомов прокариот .....</b>	<b>128</b>
7.1. Роль NGS в микробиологии .....	128
7.2. История секвенирования бактериальных геномов.....	129
7.3. Определение полной последовательности бактериального генома <i>de novo</i> .....	130
7.4. Пример протокола секвенирования образца бактериальной ДНК.....	132
7.5. Анализ данных геномного секвенирования бактерий .....	141
7.6. Секвенирование транскриптома прокариот.....	142
Список литературы .....	145
<b>Глава 8. Исследование микробных сообществ методами NGS .....</b>	<b>146</b>
8.1. Очистка ДНК для метагеномных исследований.....	147
8.2. Анализ микробного сообщества секвенированием ампликонов .....	148
8.3. Метагеномное секвенирование .....	151
8.4. Биоинформатический анализ данных метагеномного секвенирования.....	152
8.5. Комбинированный алгоритм анализа таксономического состава сообщества .....	154
8.6. Сравнение метагеномов между собой.....	155
8.7. Метатранскриптом .....	155
Список литературы .....	157

<b>Глава 9. Секвенирование геномов эукариот.....</b>	<b>162</b>
9.1. Общие аспекты секвенирования сложных геномов....	162
9.2. Секвенирование эукариотических геномов <i>de novo</i> ....	164
9.3. Повторное секвенирование (ресеквенирование) .....	166
9.4. Фазирование при ресеквенировании диплоидных геномов .....	169
9.5. Секвенирование генома отдельной клетки .....	171
Список литературы .....	174
<b>Глава 10. Секвенирование транскриптомов эукариот.....</b>	<b>176</b>
10.1. Применение NGS для исследования РНК.....	176
10.2. Общие моменты очистки РНК и синтеза кДНК .....	178
10.3. Ферменты для обратной транскрипции.....	180
10.4. Подготовка библиотеки кДНК для NGS .....	182
Список литературы .....	188
<b>Глава 11. Повышение концентрации определенных последовательностей в библиотеке для NGS (таргетное секвенирование).....</b>	<b>191</b>
11.1. Параметры методов целевого обогащения.....	191
11.2. Обогащение библиотеки фрагментов ДНК только на основе ПЦР.....	192
11.3. Обогащение библиотеки фрагментов ДНК при помощи гибридизации с пробой.....	198
11.4. Обогащение при помощи гибридизации в растворе с отбором методом ПЦР (инвертированные молекулярные пробы) .....	201
11.5. Обогащение библиотеки белок-связывающими последовательностями хроматина (ChIP-Seq) .....	203
Список литературы .....	206
<b>Глава 12. Применение высокопроизводительного секвенирования в медицинской практике .....</b>	<b>207</b>
12.1. Генетическое тестирование с использованием NGS....	207
12.2. Исследование патогенов и микробиома человека.....	220
Список литературы .....	222
<b>Глава 13. Перспективы высокопроизводительного секвенирования .....</b>	<b>223</b>
Список литературы .....	227
<b>Предметный указатель .....</b>	<b>228</b>

# 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. Сverdlov,  
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 нарек 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  
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 artificial chromosome, искусственная бактериальная хромосома
cffDNA	cell-free fetal DNA, внеклеточная ДНК плода
ChIP	chromatin immunoprecipitation, иммунопреципитация хроматина
Indel	insertion/deletion, вставка/делеция
MALDI-TOF	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
SV	большие структурные вариации
A (A)	аденин
G (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
T (T)	тимин
т. п. н.	тысяча пар нуклеотидов
Ц (C)	цитозин

# 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 дегградации (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

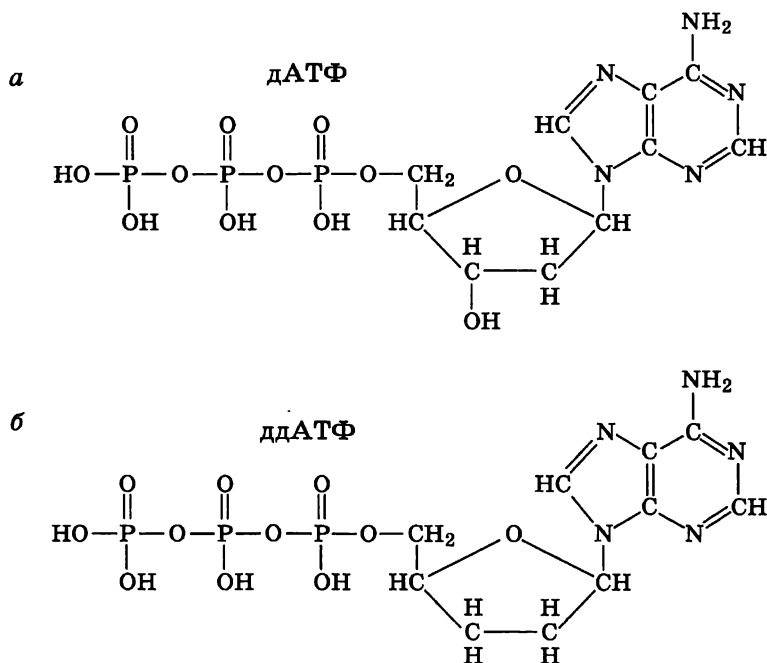


После обработки все четыре образца наносят параллельно в денатурирующий полиакриламидный гель и проводят электрофорез так, чтобы получить разделение фрагментов, отличающихся на один нуклеотид. Далее с помощью рентгеновской пленки получают изображение (электрофореграмму), по которому можно восстановить 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-полимеразы (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 дезоксинуклеозидтрифосфатов (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 реакцию (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-

ветствующих ддНТФ (А, Т, Г и Ц), меченых каждый своим флуоресцентным красителем (до применения флуорофоров в качестве метки долгое время использовали изотопы (обычно  $P^{32}$ ), а реакцию проводили в четырех отдельных пробирках (отдельно для каждого азотистого основания), как и в методе Максама–Гилберта).

В ходе реакции мечения (ферментативного синтеза ДНК) в каком-то положении случайным образом происходит включение в строящуюся цепь вместо дНТФ меченого ддНТФ, что ведет к остановке синтеза (так как отсутствие 3'-ОН-группы блокирует образование фосфодиэфирной связи со следующим нуклеотидом). Реакцию проводят циклически (аналогично ПЦР), многократно повторяя синтез ферментом новых одноцепочечных фрагментов. Так как ддНТФ составляют около 5% от дНТФ, а мечение включает 40–50 циклов реакции, в конце такой (линейной) амплификации получается набор одиночных цепей ДНК, отличающихся по длине и всегда заканчивающихся меченым нуклеотидом (рис. 1.3). После мечения проводят разделение полученных одноцепочечных фрагментов методом электрофореза в геле (обычно полиакриламидном).

Для удобства проведения электрофореза были созданы специальные приборы (автоматические секвенаторы), проводящие разделение флуоресцентно-меченых фрагментов ДНК в тонком капилляре, заполненном гелем (рис. 1.4). Детекция разделенных фрагментов происходит на дальнем конце капилляра за счет регистрации флуоресценции терминальных ддНТФ у проходящих через детектор молекул. В зависимости от типа терминального основания прибор регистрирует свой спектр флуоресценции (рис. 1.5). Анализ данных капиллярного секвенирования по сути сводится к «прочтению» последовательных пиков флуоресценции. В настоящее время с использованием современных автоматических секвенаторов длина одного прочтения (рида) по методу Сенгера составляет 800–1000 нуклеотидов.

Создание автоматических секвенаторов настолько упростило и удешевило процесс определения последовательности ДНК, что позволило к середине 1980-х годов говорить о возможности определить полную последовательность генома человека. Это вылилось в крупнейшее исследование под названием Human Genome Project (HGP), в которое были вовлече-

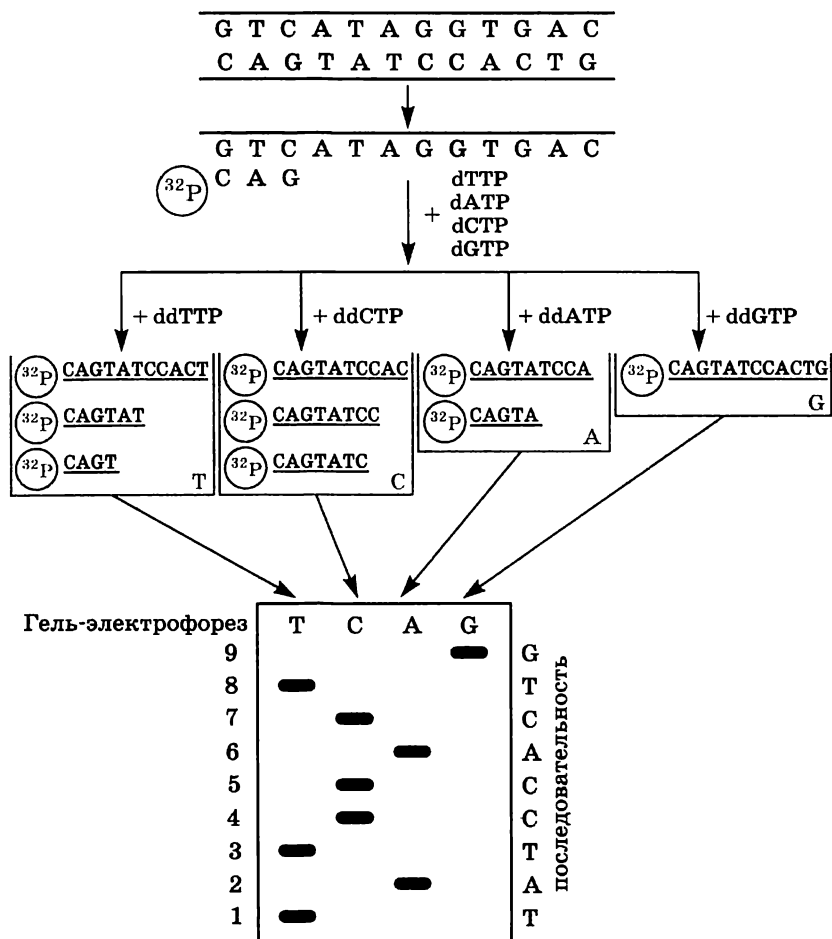
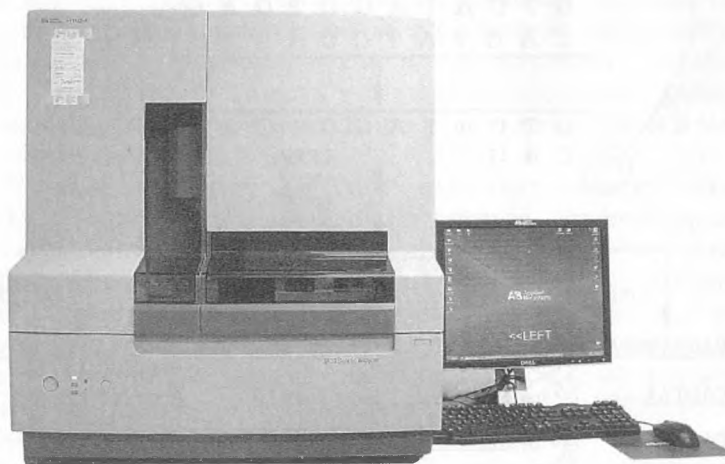


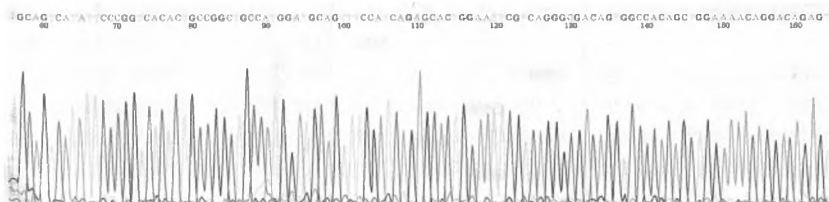
Рис. 1.3. Принцип метода Сенгера: удлинение цепи ДНК ферментом происходит до момента включения дидезоксинуклеотида.

Разделение полученных фрагментов методом электрофореза в геле позволяет определить последовательность нуклеотидов

ны научные центры по всему миру. Целью проекта, помимо определения последовательности нуклеотидов, была идентификация всех генов человека. Проект стартовал в 1990 году и финишировал в 2001 году публикацией в журнале *Nature* [5]. Однако более или менее полный анализ полученных данных был закончен только через два года.



**Рис. 1.4.** Современный капиллярный автоматический секвенатор для определения последовательности нуклеиновых кислот методом Сенгера (ABI 3130xl, Applied Biosystems)



**Рис. 1.5.** Хроматограмма, полученная в результате секвенирования по методу Сенгера на автоматическом секвенаторе

Подход к определению полной последовательности генома человека был следующим: весь геном разделили на фрагменты по 150 000 п. н., которые вставили в искусственные бактериальные хромосомы (ВАС), причем для каждого такого фрагмента было определено его расположение на хромосоме. Вставку каждой ВАС секвенировали так называемым методом дробовика (shotgun sequencing), для чего фрагмент генома из ВАС расщепляли на более короткие фрагменты (по 2000–3000 п. н.), каждый из которых субклонировали в бактериальный вектор и определяли последовательность фрагмента методом Сенгера. Ввиду случайности фрагментации получаемые последовательности частично перекрывались, что

позволяло при помощи специального программного обеспечения «собирать» из отдельных прочтений ВАС-вставки, а затем и целые хромосомы.

Стоит отметить, что в 1998 году, параллельно с мировым научным сообществом, секвенированием генома человека занялась компания Celera под управлением Крейга Вентера. Последовательность в 3 млрд п. н. ее сотрудниками была получена в течение 9 месяцев, т. е. в 20 раз быстрее, чем участниками консорциума HGP. Среди прочего это связано с тем, что Celera стартовала на автоматических секвенаторах последнего поколения, сильно выигрывающих по производительности. Кроме того, компания не применяла клонирование генома в ВАС, а сразу использовала метод дробовика для всего генома [6].

### **1.1.3. Гибридизация на твердой фазе (принцип комплементарности цепей ДНК)**

В конце 1980-х был предложен подход к определению последовательности ДНК, получивший название секвенирования путем гибридизации (sequencing by hybridization, SBH), или секвенирования на чипе [7, 8]. Метод предполагает гибридизацию меченой одноцепочечной ДНК, разрушенной до коротких фрагментов, с синтетическими олигонуклеотидами известной структуры и определенной длины, точно расположенными на подложке. При этом на подложке присутствуют все возможные варианты последовательности олигонуклеотида данной длины (например, все 65 536 вариантов олигонуклеотида длиной в 8 оснований). Условия гибридизации подбирают так, чтобы только полностью комплементарные фрагменты ДНК взаимодействовали с олигонуклеотидом на подложке. Таким образом, после удаления несвязавшихся молекул ДНК можно зарегистрировать сигнал в тех позициях чипа, где находится олигонуклеотид, последовательность которого есть в секвенируемом образце ДНК. Полученный гибридизационный паттерн можно использовать для восстановления исходной последовательности путем сборки перекрывающихся участков сработавших проб (рис. 1.6).

Из-за низкой дискриминирующей способности гибридизационного подхода (невозможно подобрать условия, при которых к олигонуклеотидам будут «прилипать» только полностью комплементарные фрагменты, всегда найдутся GC-богатые

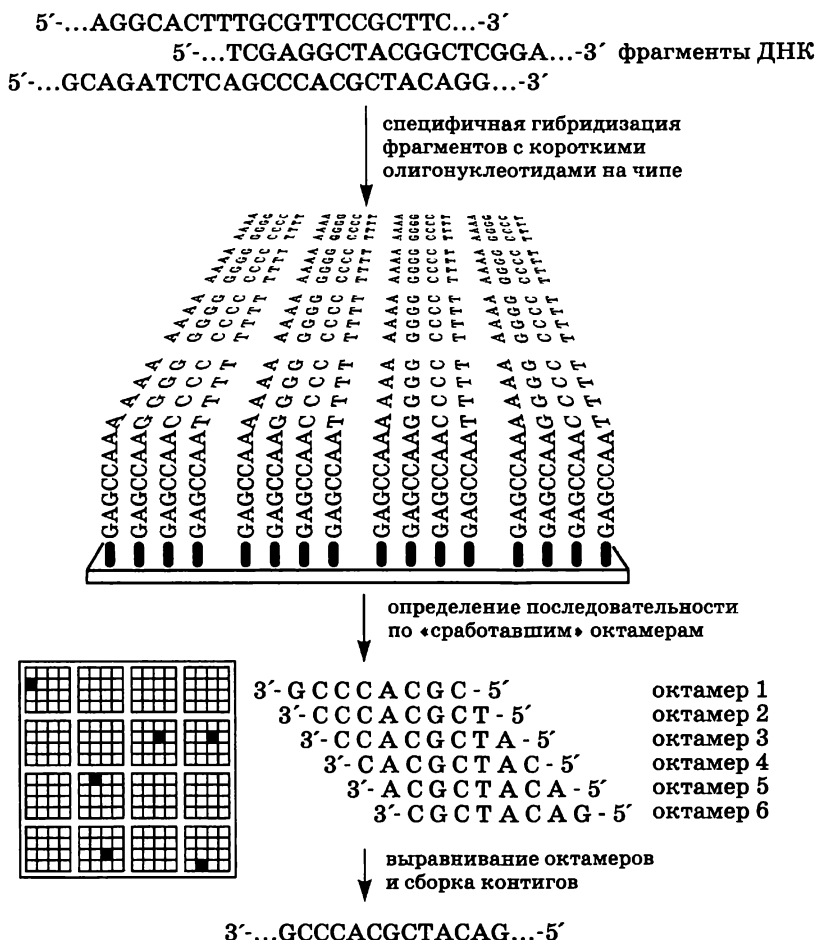


Рис. 1.6. Принцип работы метода секвенирования гибридизацией на твердой фазе (чипе)

участки, которые будут гибридизоваться и при наличии одного или даже нескольких неспаренных оснований) метод SBH пока не нашел практического применения в секвенировании ДНК. Однако алгоритмы, разработанные авторами SBH для сборки коротких прочтений в более длинные фрагменты, стали основой для последующих алгоритмов высокоскоростной сборки и выравнивания, в том числе используемых в современных технологиях NGS.

В настоящее время разработкой технологий NGS, основанных на гибридизационном подходе, занимаются компании Affymetrix и Perlegen (США).

#### 1.1.4. MALDI–TOF масс-спектрометрия (определение нуклеотида по массе и заряду)

Масс-спектрометрия позволяет идентифицировать компоненты гетерогенной смеси биомолекул по разнице их молекулярных масс. В варианте масс-спектрометрии MALDI–TOF (matrix-assisted laser desorption/ionization–time-of-flight) образец для анализа помещают на поглощающую УФ-излучение подложку и подвергают воздействию короткого лазерного импульса. Ионизированные молекулы летят в электрическом поле в направлении детектора, причем время достижения детектора зависит от соотношения масса/заряд для каждой молекулы.

В первые годы после разработки MALDI–TOF масс-спектрометрию применяли для анализа структуры пептидов, однако позже были разработаны подходы, позволяющие определять и последовательность нуклеотидов в нуклеиновых кислотах. В начале 2000-х годов стали появляться публикации, предлагающие варианты секвенирования ДНК на основе так называемой тандемной (MS/MS) MALDI–TOF масс-спектрометрии [9–12].

Ключевые элементы определения последовательности ДНК методом MALDI–TOF масс-спектрометрии следующие: гомогенный фрагмент ДНК (или РНК) высушивают на поверхности в составе 3-гидроксипиридиновой кислоты. ДНК обрабатывают коротким импульсом УФ-лазера, в результате чего ионы ДНК переходят в газовую фазу. Заряженные молекулы ДНК в газовой фазе под действием высокого напряжения ускоряются в электрическом поле и попадают на детектор. Затем проводят повторный раунд расщепления и определения масс более мелких фрагментов. Между раундами возможен дополнительный этап сепарации на основе газовой хроматографии. На основании полученных данных может быть вычислена масса анализируемой молекулы и расшифрована последовательность сравнительно короткого гомогенного фрагмента (рис. 1.7).

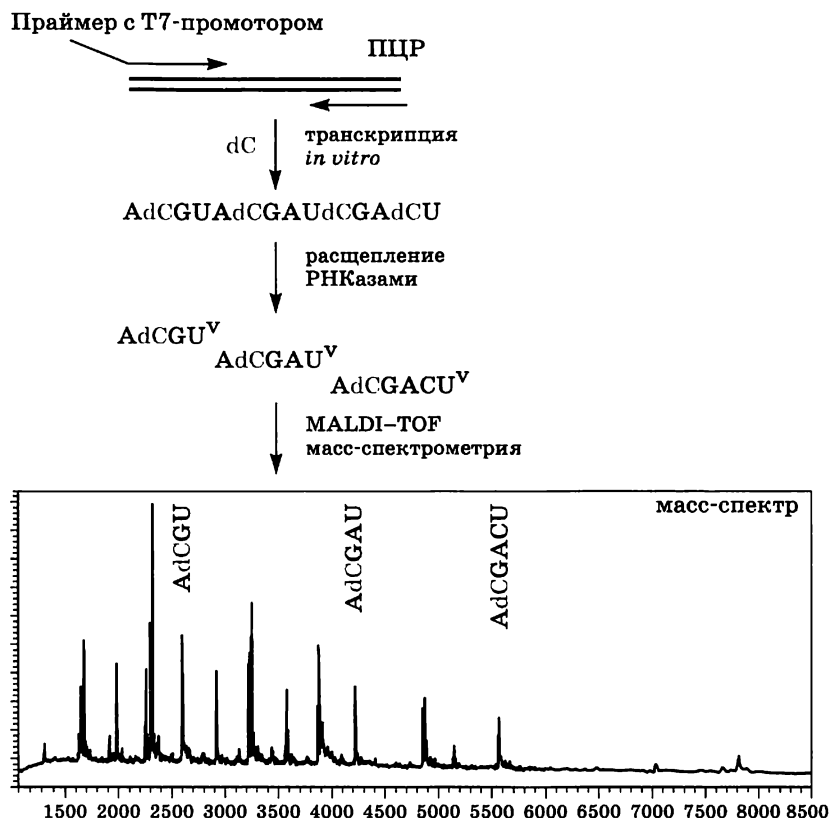


Рис. 1.7. Принцип метода секвенирования ДНК с помощью MALDI-TOF масс-спектрометрии

В настоящее время метод MALDI-TOF масс-спектрометрии не используется в коммерческих вариантах секвенаторов.

### 1.1.5. Секвенирование лигированием (принцип комплементарности цепей ДНК)

Первой технологией высокопроизводительного секвенирования можно считать разработанную в начале 1990-х годов в компании Lynx Therapeutics (США) технологию массового параллельного секвенирования при помощи уникальных меток (massively parallel signature sequencing – MPSS). Метод осно-

вывался на секвенировании лигированием с удалением присоединенного олигонуклеотида с помощью эндонуклеазы рестрикции. Вначале ПЦР-фрагменты с помощью адаптеров прикрепляют к микросферам (микрочастицам) так, что на каждой частице оказывается множество одинаковых фрагментов. Частицы располагают на подложке и выполняют несколько раундов определения последовательности путем лигирования флуоресцентно-меченых олигонуклеотидов с последующим удалением эндонуклеазой BbvI. Таким образом удавалось прочесть до 20 п. н. с каждой микросферы. Однако из-за сложности технологии она не нашла широкого применения и использовалась лишь внутри компании. В 2004 году Lynx Therapeutics объединилась с компанией Solexa (а позже их приобрела компания Illumina), что привело к развитию другого, более простого подхода: секвенирования синтезом (который Solexa купила у изобретателя технологии – компании Manteia Predictive Medicine, см. ниже). Тем не менее ключевые свойства подхода MPSS оказались типичны для многих разработанных позже технологий высокопроизводительного секвенирования, в том числе получения в результате секвенирования сотен тысяч или даже миллионов коротких прочтений ДНК [13, 14].

Современные методы секвенирования лигированием основаны на использовании коллекции коротких (как правило, от 8 до 10 оснований) флуоресцентно-меченых (с помощью четырех красителей) вырожденных олигонуклеотидов, так, что каждому флуорофору соответствует определенный нуклеотид (или два нуклеотида) в определенной позиции (рис. 1.8).

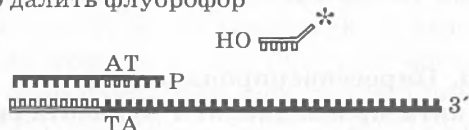
Сначала любым методом создают иммобилизованную на твердой фазе клональную библиотеку одноцепочечных фрагментов ДНК (например, методом эмульсионной ПЦР, разд. 2.9.2). Секвенирование начинают с отжига праймера, комплементарного адаптеру на одном из концов библиотеки ДНК. Затем к библиотеке добавляют флуоресцентно-меченые вырожденные олигонуклеотиды и проводят реакцию лигирования, что приводит к фиксации олигонуклеотида на фрагменте в случае его полного соответствия. Затем считывают флуоресценцию, определяя тем самым, какой нуклеотид (или пара нуклеотидов) находится в определенной позиции. Флуорофор удаляют и лигируют следующий олигонуклеотид (всего проводят 10–15 последовательных лигирований). Затем проводят «перезагрузку» путем отсоединения праймера с при-



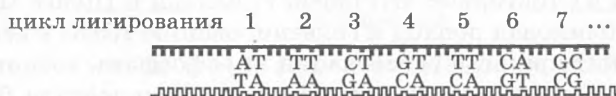
## 3. Блокировать несработавшие цепи



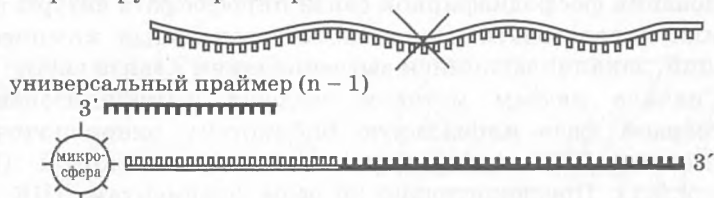
## 4. Удалить флуорофор



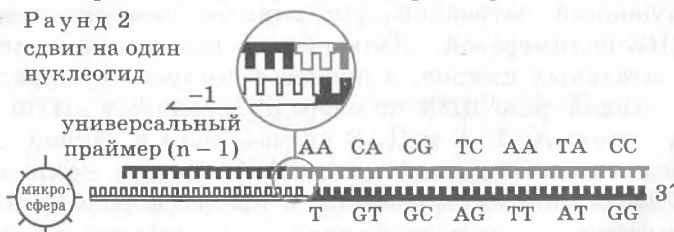
## 5. Повторить шаги 1–4



## 6. Перезагрузка праймера



## 7. Повторить шаги 1–5 с новым праймером



## 8. Повторить перезагрузку с праймерами (n - 2), (n - 3), ...

**Рис. 1.8 (окончание).** Секвенирование с использованием лигирования. Последовательность этапов на примере технологии SOLiD

лигированными мечеными олигонуклеотидами и повторяют цикл с другим праймером со сдвигом на одну букву.

Принцип секвенирования лигированием в настоящее время используют коммерческие технологии Polonator (Dover/Harvard) и SOLiD (Life Technologies Thermo Fisher Scientific) (см. разд. 3.2).

### **1.1.6. Пиросеквенирование (регистрация акта присоединения нуклеотида по образующемуся пирофосфату)**

В 1996 году специалистами Королевского технологического института в Стокгольме Мустафой Рональди и Полом Ниреном был опубликован подход к секвенированию ДНК, в основе которого лежит принцип регистрации пирофосфата, возникающего в результате присоединения очередного нуклеотида ДНК-полимеразой [15]. Для детекции выделяющегося в процессе образования фосфодиэфирной связи пирофосфата авторы предложили использовать каскад последовательных химических реакций, заканчивающийся высвечиванием кванта света.

Сначала любым методом создают иммобилизованную на твердой фазе клональную библиотеку одноцепочечных фрагментов ДНК (например, методом мостиковой ПЦР, разд. 2.9.1). Предварительно ко всем фрагментам ДНК присоединяют адаптер, на который будет гибридизоваться праймер, служащий затравкой для синтеза комплементарной цепи ДНК-полимеразой. Дальнейшая реакция состоит из последовательных циклов, в процессе которых к закрепленной на твердой фазе ДНК по очереди добавляют дНТФ всех четырех типов: А, Т, Г и Ц. В случае, если в данной ДНК-колонии на секвенируемой цепи ДНК имеется комплементарный добавленному нуклеотид, в процессе формирования ДНК-полимеразой фосфодиэфирной связи побочным продуктом реакции станет пирофосфат. Он активирует каскад химических реакций, в результате возникает световой сигнал, интенсивность которого прямо пропорциональна числу включенных в цепь нуклеотидов (если подряд идут несколько одинаковых нуклеотидов, сигнал будет ярче). Ферментативные реакции осуществляются АТФ-сульфурилазой, люциферазой и апиразой, также вместе с ними в ячейке присутствуют аденозинсульфофосфат (APS) и люциферин (стехиометрически выделившийся пирофосфат вместе с АСФ при помощи АТФ-

сульфуриказы образует АТФ, являющийся источником энергии для проведения люциферазой реакции окисления люциферина в оксильюциферин, в процессе которой и генерируется свет в видимом спектре в количестве, пропорциональном количеству включенных нуклеотидов). Световой сигнал обычно детектируется ПЗС-матрицей (аналогичной встроенным в обычный цифровой фотоаппарат) и анализируется при помощи программного обеспечения, преобразующего так называемую пирогамму в последовательность нуклеотидов (рис. 1.9). Не вовлеченные в синтез новой цепи нуклеотиды, а также

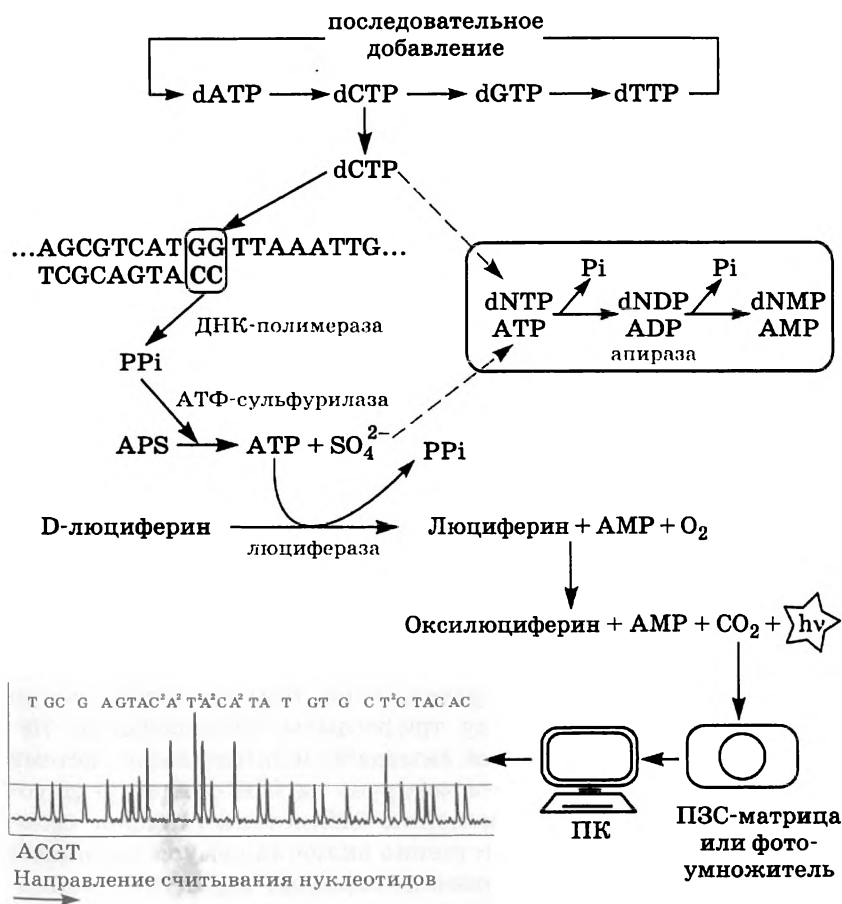


Рис. 1.9. Принцип метода пиросеквенирования

АТФ деградируются при помощи апиразы. После этого можно начинать следующий цикл, т. е. добавлять другой тип нуклеотида.

Позднее Маргулис с соавторами развили технологию, предложив сочетание пиросеквенирования с эмульсионной ПЦР [16]. На принципе пиросеквенирования основана коммерческая технология 454 Life Sciences компании Roche (см. разд. 3.1).

### **1.1.7. Обратимые терминирующие нуклеотиды (регистрация каждого присоединенного нуклеотида по отщепляемой метке)**

Концепция секвенирования синтезом была предложена Баласубрамьяном и Кленерманом, работавшими на химическом факультете Кембриджского университета [17]. Так же как и в пиросеквенировании, принцип заключается в регистрации факта присоединения очередного нуклеотида, но не по побочным продуктам реакции, а непосредственно по сигналу от присоединенного основания. При этом должны быть выполнены два требования: за один цикл реакции может быть добавлен только один нуклеотид (что легко обеспечить использованием 3'-блокированных трифосфатов с возможностью снятия блока) и метка должна быть отщепляемой.

Сначала любым методом создают иммобилизованную на твердой фазе клональную библиотеку одноцепочечных фрагментов ДНК (например, методом мостиковой ПЦР, разд. 2.9.1). Секвенирование начинают с отжига праймера, комплементарного адаптеру на одном из концов библиотеки ДНК. Затем к библиотеке добавляют четыре типа флуоресцентно-меченых обратимых терминирующих нуклеозидтрифосфатов (так называемых RT-оснований). ДНК-полимераза присоединяет подходящий нуклеотид к затравке, и на этом синтез временно останавливается (поскольку трифосфаты блокированы). Невключившиеся нуклеотиды смывают, и оптическая система считывает (например, фотографирует на ПЗС-матрицу) флуоресценцию каждой ДНК-колонии библиотеки (каждая колония флуоресцирует соответственно включившемуся на данном шаге нуклеотиду). После этого флуорофор, наряду с 3'-концевым блокатором, химически удаляют из синтезируемой цепи, что позволяет повторить цикл сначала (рис. 1.10).

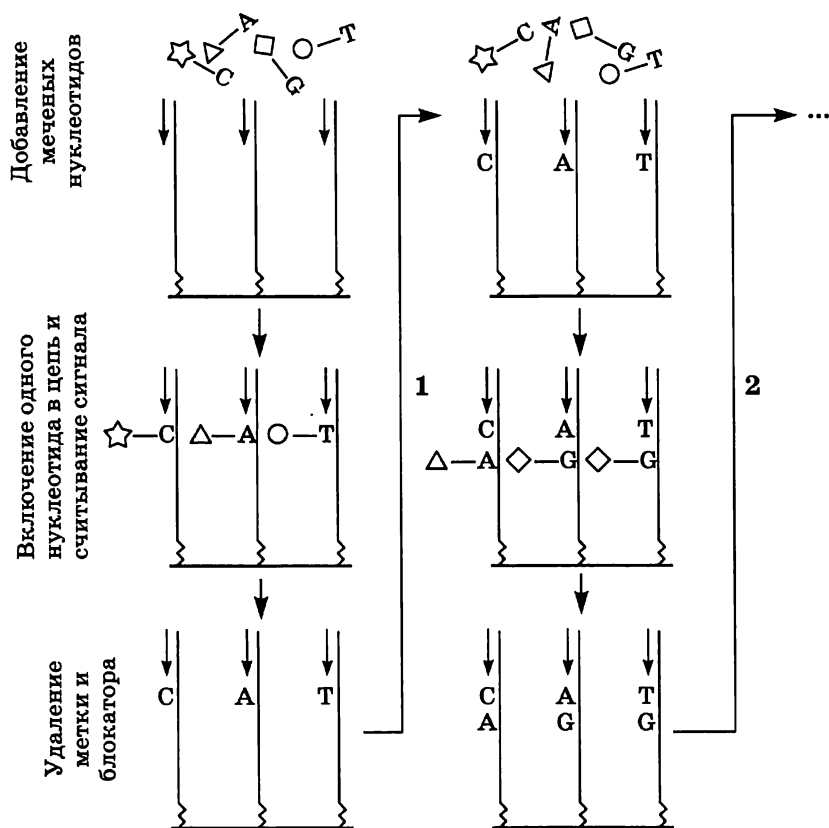


Рис. 1.10. Принцип секвенирования синтезом клональной библиотеки одноцепочечных фрагментов ДНК на твердой фазе

В отличие от пиросеквенирования в данном подходе флуоресцентный сигнал можно регистрировать в течение длительного времени после присоединения очередного нуклеотида, что позволяет «в спокойной обстановке» с использованием хорошей оптической системы считать огромный массив ДНК-колоний.

На принципе секвенирования синтезом основаны коммерческие технологии компаний Illumina и Pacific Bioscience (в варианте одиночных молекул, см. разд. 3.5).

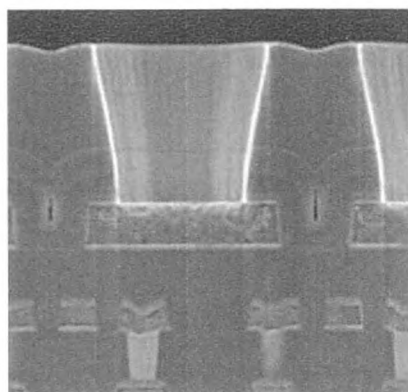
### 1.1.8. Полупроводниковое секвенирование (регистрация акта присоединения нуклеотида по образующимся ионам водорода)

Полупроводниковое секвенирование – это метод определения последовательности ДНК, основанный на детекции ионов водорода, выделяющихся в среду в ходе синтеза ДНК ферментом.

Процесс в целом очень похож на пиросеквенирование. Сначала одним из стандартных методов создают иммобилизованную на твердой фазе клональную библиотеку одноцепочечных фрагментов ДНК (например, методом эмульсионной ПЦР, разд. 2.9.2). Важно, чтобы метод создания библиотеки позволял некоторым образом отделить каждую колонию ДНК от других так, чтобы выравнивание рН (в случае его изменения в районе колонии) происходило не слишком быстро. При использовании эмульсионной ПЦР это обеспечивается закатыванием микросфер в соответствующие им по размерам микрореакторы, так, что в реакторе помещается только одна частица, а сообщаются реакторы только одной поверхностью на проточном чипе (рис. 1.11).



а



б

Рис. 1.11. Полупроводниковое секвенирование: схема детекции сигнала (а) и микрофотография поперечного среза чипа для секвенирования Ion Torrent (б)

Секвенирование начинают с отжига праймера, комплементарного адаптеру на одном из концов библиотеки ДНК. Затем к микрореакторам (с микросферами) по очереди добавляют обычные трифосфаты. Если добавленный нуклеотид оказывается комплементарен матрице, ДНК-полимераза встраивает его в синтезируемую цепь. Реакция образования фосфодиэфирной связи приводит к выделению пирофосфата и протона, вызывающего локальное изменение pH раствора в микрореакторе, которое детектируется подключенным к каждому микрореактору сенсором. Если нуклеотид не подходит, сигнал отсутствует. После каждого добавленного в реакцию нуклеотида прибор выполняет промывку системы буфером для очистки от остатков не включившихся дНТФ данного типа. Как и в случае пиросеквенирования, у полупроводникового секвенирования есть трудности с детекцией гомополимерных участков – в случае протяженного моонуклеотида (к примеру, TTTTTTTT), сигнал теряет дискретность, и определить, сколько именно нуклеотидов (5, 6 или 7) присутствует в последовательности, становится сложно.

Важным отличием полупроводникового секвенирования от трех других наиболее популярных методов (пиросеквенирования, секвенирования лигированием и секвенирования синтезом) является отсутствие оптического детектора сигнала, что значительно упрощает и удешевляет конструкцию прибора. Оптическая детекция, по сути представляющая собой встроенный в секвенатор микроскоп, имеет ограничения, связанные с полем зрения микроскопа и его разрешающей способностью: на площади чипа (слайда) возможно разместить ограниченное количество микроцентров секвенирования (колоний ДНК, микрореакторов, иммобилизованных на подложке микросфер и т. д.), достаточно удаленных друг от друга физически для достоверного отличия сигнала от соседних центров. Детекция на полупроводниковом чипе такого ограничения практически не имеет.

На принципе полупроводникового секвенирования основана коммерческая технология Ion Torrent от Life Technologies Thermo Fisher Scientific. Разработки в данном направлении ведутся также компанией Roche.

## 1.2. МЕТОДЫ, ОСНОВАННЫЕ НА ДЕТЕКЦИИ СИГНАЛА ОТ ОДНОЙ МОЛЕКУЛЫ ДНК (СЕКВЕНИРОВАНИЕ ОДИНОЧНЫХ МОЛЕКУЛ ДНК)

Описанные ниже методы отличаются отсутствием этапа клональной амплификации библиотеки (или отдельного фрагмента) ДНК. Это упрощает методику, но, главное, убирает один из этапов искажения исходного материала (ведь в перечисленных выше методах секвенируется не исходная ДНК, а примерно ее 20-я копия: попробуйте последовательно 20 раз скопировать эту страницу на Хегох, делая новую копию с предыдущей, – увидите влияние амплификации). Однако регистрация сигнала от единственной молекулы накладывает чрезвычайно высокие требования к соответствующим детекторам.

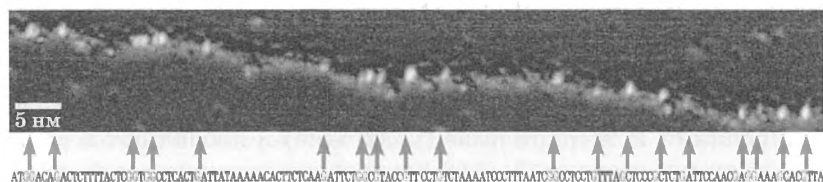
Отметим, что ряд авторов называют описанные ниже варианты NGS как технологии третьего поколения (в отличие от технологий NGS второго поколения, требующих клональной амплификации ДНК).

### 1.2.1. Секвенирование при помощи электронного микроскопа

Использование электронного микроскопа для определения последовательности нуклеиновых кислот было предложено Ричардом Фейнманом еще в конце 50-х годов XX века [18]. Электронная микроскопия включает в себя три технологических разновидности: сканирующая электронная микроскопия (СЭМ, SEM), просвечивающая электронная микроскопия (ПЭМ, TEM) и сканирующая просвечивающая электронная микроскопия (СПЭМ, STEM).

В 1960-х и 70-х годах методы просвечивающей электронной микроскопии активно разрабатывались, тогда же были предложены подходы для определения последовательности ДНК [19, 20]. В 1970 году Альберт Крю предложил метод визуализации в сканирующем электронном микроскопе (high-angle annular dark-field imaging, HAADF). Используя эту технику, можно обнаружить отдельные тяжелые атомы на тонких пленках аморфного углерода [21]. Чтобы визуализировать отдельные основания в ДНК, они должны быть помечены атомами тяжелых металлов. Однако метод не был доведен до практического использования из-за быстрого разрушения молекулы ДНК пучком электронов.

В 1990-х годах получила широкое распространение технология секвенирования с помощью метода сканирующей туннельной микроскопии, но воспроизводимость публикуемых результатов была слишком низкой, и в какой-то момент новые исследования к публикации принимать перестали. В 2009 году японские исследователи сообщили, что им удалось на отдельной одноцепочечной молекуле ДНК визуализировать гуанин по геометрическим характеристикам (рис. 1.12) [22]. В этом же направлении работу проводят и производители микроскопов.



**Рис. 1.12.** Визуализация гуаниновых нуклеотидов на цепи ДНК посредством возможностей сканирующей туннельной микроскопии

В 2010 году Криванек с коллегами предложили усовершенствование метода ПЭМ, что позволило видеть одиночные замены атомов в монослое нитрида бора [23].

Несмотря на появление множества различных технологий секвенирования, исследователи не оставляют попыток применения электронной микроскопии для секвенирования одиночных молекул ДНК. Теоретически электронная микроскопия может обеспечить чрезвычайно длинные прочтения, что очень важно для сборки эукариотических геномов, в основном состоящих из повторяющихся последовательностей. Например, в 2012 году коллективом ученых из Гарвардского университета, Университета штата Нью-Гемпшир и компании ZS Genetics была продемонстрирована возможность прочтения длинных последовательностей ДНК при помощи электронной микроскопии [24], однако подобные технологии секвенирования по-прежнему далеки от широкого коммерческого применения.

### 1.2.2. Использование обратимых терминирующих нуклеотидов на одиночных молекулах нуклеиновых кислот (регистрация каждого присоединенного нуклеотида по отщепляемой метке)

Этот подход аналогичен описанному выше варианту обратимых терминирующих нуклеотидов (см. разд. 1.1.7), но без этапа создания клональной библиотеки.

Один из вариантов метода предложен исследователями Корнелльского университета (США) в 2003 году [25]. Исследователи разработали технологию оптической детекции включения флуоресцентно-меченых нуклеотидов в строящуюся цепь отдельной молекулы ДНК. Используя принципы ближнего (эванесцентного) поля с апертурой порядка нескольких нанометров, можно регистрировать флуоресценцию единственного флуорофора. Для секвенирования молекулу ДНК-полимеразы прикрепляют к твердой фазе (подложке) и добавляют в реакцию матрицу и специальные флуоресцентно-меченые трифосфаты (каждый помечен своим цветом). В ходе синтеза ферментом комплементарной цепи ДНК при образовании фосодиэфирной связи происходит отщепление флуорофора от вновь присоединенного основания, а прибор фиксирует флуоресценцию и длительность испускания света (рис. 1.13). Преимуще-

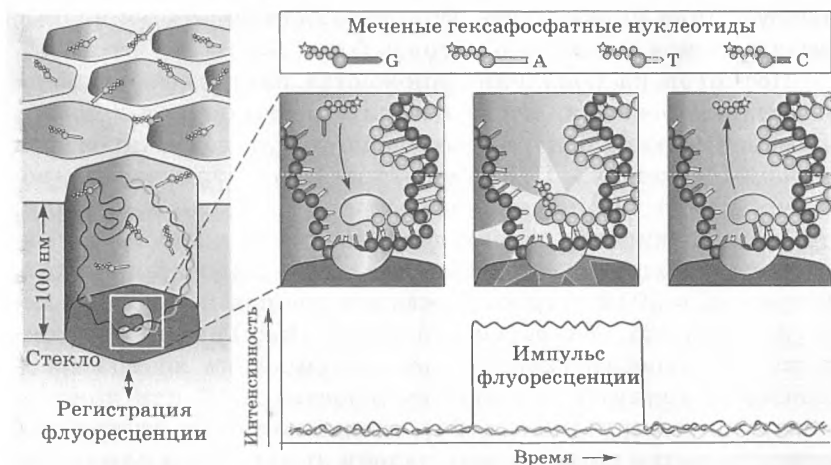


Рис. 1.13. Технология секвенирования синтезом одиночных молекул

ствами подхода являются очень длинные прочтения (десятки тысяч нуклеотидов) и теоретическая возможность определять модифицированные основания в ДНК.

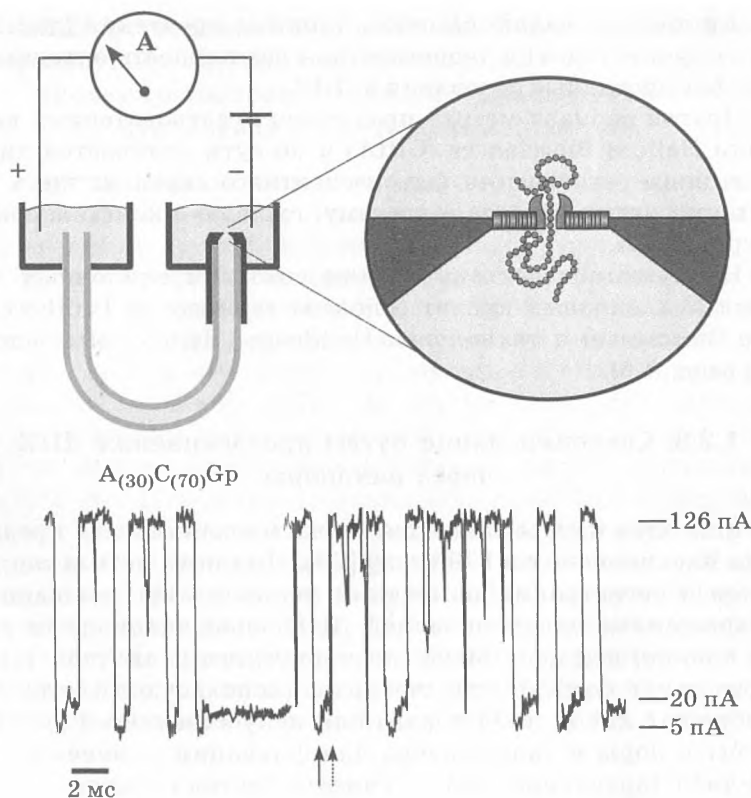
Другой вариант метода предложен исследователями компании Helicos Biosciences (США) и по сути отличается лишь принципом регистрации флуоресцентного сигнала: здесь используют четыре лазера и систему, схожую с конфокальным микроскопом.

На принципе секвенирования синтезом одиночных молекул нуклеиновых кислот основана технология PacBio (Pacific Bioscience) и технология HeliScope (Helicos Biosciences) (см. разд. 3.6).

### **1.2.3. Секвенирование путем протаскивания ДНК через нанопоры**

Еще одна интересная идея секвенирования была предложена Касьяновичем в 1996 году [26]. Принцип метода заключается в регистрации изменений ионного тока, вызванных прохождением одноцепочечной ДНК через нанопору (в тонкой пленке) под действием электрического поля (рис. 1.14). Поры могут быть биологическими (используют клеточную мембрану с какой-либо порой) или искусственными (это могут быть поры в виде сенсора для фиксации изменения какой-либо характеристики – туннельного тока, емкости, ионного тока, флуоресценции и т. д.). При переходе через пору каждый тип азотистых оснований по-своему «закупоривает» пору и влияет на ток. Предварительные результаты показали возможность различать длинные гомополимерные отрезки (например, 30 аденинов от следующих за ними 70 цитозинон). Однако чрезвычайно короткое время прохождения основания сквозь пору (1 мкс) и тепловые флуктуации пока не позволяют определять количество стоящих подряд однотипных нуклеотидов, и метод не получил практического применения.

Несколько компаний в мире разрабатывают свои технологии секвенирования одиночных молекул с помощью нанопор, но ни одного такого устройства пока не представлено. Дальше всех по этой технологии продвинулась компания Nanopore, заявлявшая выпуск прибора в 2013 году, но на начало 2014 года коммерческого варианта технологии на рын-



**Рис. 1.14.** Секвенирование путем протаскивания ДНК через нанопоры. Проходящая через пору молекула одноцепочечной ДНК или РНК меняет потенциал на мембране

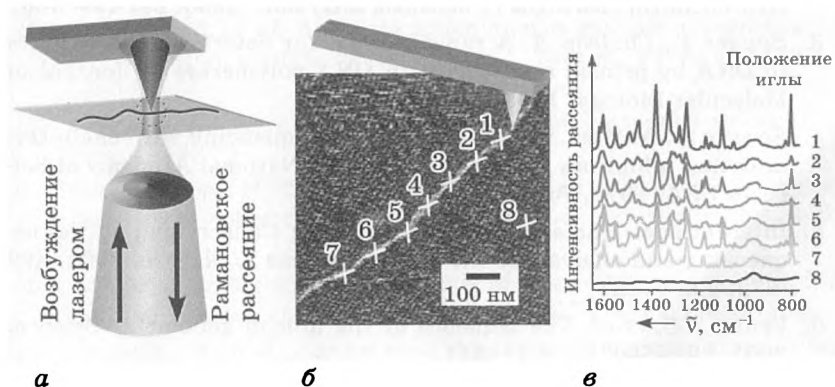
ке не появилось. Основная проблема – высокая чувствительность к факторам внешней среды, в результате чего происходит разрыв билипидного слоя, в котором находятся поры.

#### 1.2.4. Секвенирование методом спектроскопии комбинационного рассеяния

Перспективное направление для секвенирования ДНК и РНК – использование спектроскопии комбинационного рассеяния. Комбинационное рассеяние света (эффект Рамана) – неупругое рассеяние оптического излучения на молекулах вещества (твёрдого, жидкого или газообразного), сопровожда-

ющееся заметным изменением частоты излучения. В отличие от рэлеевского рассеяния в случае комбинационного рассеяния света в спектре рассеянного излучения появляются спектральные линии, которых нет в спектре первичного (возбуждающего) света. Число и расположение появившихся линий определяются молекулярным строением вещества. Спектроскопия комбинационного рассеяния света (или рамановская спектроскопия) – эффективный метод химического анализа, изучения состава и строения веществ.

В 2007 году немецкий исследователь Волкер Дэкерт показал, что с помощью методики TERS (tip-enhanced raman spectroscopy) можно, пройдя по цепочке РНК, определить по спектрам, какие именно нуклеотиды находятся в цепи (рис. 1.15). За счет усиления сигнала комбинационного рассеяния на таких металлах, как золото и серебро, получается спектр, который можно детектировать. По усиленному спектру удастся расшифровать последовательность нуклеотидов, причем со всеми модификациями, поскольку спектр содержит необходимую информацию. В 2012 году Дэкерт на одной из конференций показал работу, где он различил отдельные модификации на молекуле ДНК.



**Рис. 1.15.** Определение последовательности нуклеотидов РНК с использованием методики TERS: *а* – схема эксперимента (показана игла атомно-силового микроскопа над молекулой ДНК, усиливающая сигнал лазера); *б* – движение иглы по подложке с молекулой ДНК (точки считывания сигнала 1, 2, ..., 7); *в* – пример графика результатов измерения рамановского спектра РНК-гомополимера из цитозинов в семи точках, а также в точке 8 для определения уровня шума

В настоящее время работы в этом направлении ведут крупнейшие компании, прежде всего Intel, IBM и HP. Так, Intel пытается секвенировать ДНК с помощью разбиения ее на отдельные нуклеотиды с последующим получением сигнала с наночастиц серебра [27].

### 1.3. ДРУГИЕ МЕТОДЫ СЕКВЕНИРОВАНИЯ

Кроме перечисленных выше подходов есть множество гораздо менее технически и идейно проработанных технологий: метод колебаний [28], секвенирование при помощи вращающегося поля [29] и т. д. Вследствие недостаточной проработанности методов мы не сочли нужным раскрывать детали этих подходов в столь кратком обзоре технологий.

### СПИСОК ЛИТЕРАТУРЫ

1. *Maxam A., Gilbert W.* A new method for sequencing DNA // Proceedings of the National Academy of Sciences USA, 1977, 74 (2): 560–564.
2. *Maxam A., Gilbert W.* Sequencing end-labeled DNA with base-specific chemical cleavages // Methods Enzymol., 1980, 65: 499–560.
3. *Sanger F., Coulson A.* A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase // Journal of Molecular Biology, 1975, 94 (3): 441–448.
4. *Sanger F., Nicklen S., Coulson A.* DNA sequencing with chain-terminating inhibitors // Proceedings of the National Academy of Sciences USA, 1977, 74 (12): 5463–5467.
5. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome // Nature, 2001, 409 (6822): 860–921.
6. *Venter J.C. et al.* The sequence of the human genome // Science, 2001, 291 (5507): 1304–1351.
7. *Pevzner P.A., Borodovsky M.Y., Mironov A.A.* Linguistics of nucleotide sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words // Journal of Biomolecular Structure and Dynamics, 1998, 6 (5): 1013–1026
8. *Idury R.M., Waterman M.S.* A new algorithm for DNA sequence assembly // J Comput Biol., 1995, 2 (2): 291–306.

9. *Marvin L.F., Roberts M.A., Fay L.B.* Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry in clinical chemistry // *Clinica Chimica Acta*, 2003, 337 (1–2): 11–21.
10. *Edwards J., Ruparel H., Ju J.* Mass-spectrometry DNA sequencing // *Mutat Res.*, 2005, 573 (1–2): 3–12.
11. *Ragoussis J., Elvidge G., Kaur K., Colella S.* Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research // *PLoS Genet.*, 2006, 2 (7): e100.
12. *Mauger F. et al.* DNA sequencing by MALDI-TOF MS using alkali cleavage of RNA/DNA chimeras // *Nucleic Acids Res.*, 2007, 35 (8): e62.
13. *Zhou D. et al.* Massively parallel signature sequencing // *Methods Mol Biol.*, 2006, 331: 285–311.
14. *Reinartz J. et al.* Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms // *Brief Funct Genomic Proteomic*, 2002, 1 (1): 95–104.
15. *Ronaghi M. et al.* Real-time DNA sequencing using detection of pyrophosphate release // *Analytical Biochemistry*, 1996, 242 (1): 84–89.
16. *Margulies M. et al.* Genome Sequencing in Open Microfabricated High Density Picoliter Reactors // *Nature*, 2005, 437 (7057): 376–380.
17. *Bentley D.R. et al.* Accurate whole human genome sequencing using reversible terminator chemistry // *Nature*, 2008, 456 (7218): 53–59.
18. *Feynman R.* There's plenty of room at the bottom // *Caltech lecture*, 1959.
19. *Beer M., Zobel R.* Electron stains II: Electron microscopic studies on the visibility of stained DNA molecules // *J. Mol. Biol.*, 1961, 3 (6): 717–726.
20. *Cole M. et al.* Molecular microscopy of labeled polynucleotides: Stability of osmium atoms // *J. Mol. Biol.*, 1977, 117 (2): 387–400.
21. *Crewe A., Wall J., Langmore J.* Visibility of a single atom // *Science*, 1970, 168 (3937): 1338–1340.
22. *Tanaka H., Kawai T.* Partial sequencing of a single DNA molecule with a scanning tunnelling microscope // *Nat. Nanotechnol.*, 2009, 4 (8): 518–22.
23. *Krivanek O.L. et al.* Atom-by-atom structural and chemical analysis by annular dark-field electron microscopy // *Nature*, 2010, 464 (7288): 571–574.

24. *Bell D. et al.* DNA Base Identification by Electron Microscopy // Microscopy and Microanalysis, 2012, 18 (5): 1049–1053.
25. *Levene M.J. et al.* Zero-Mode Waveguides for Single-Molecule Analysis at high concentrations // Science, 2003, 299: 682–686.
26. *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 USA, 1996, 93 (24): 13770–13773.
27. *Bailo E., Deckert V.* Tip-enhanced Raman spectroscopy of single RNA strands: towards a novel direct-sequencing method // Angew Chem Int Ed Engl., 2008, 47(9): 1658–1661.
28. *Schönherr G., Noolandi J.* Fluctuating bond model of DNA gel electrophoresis // Electrophoresis, 1991, 12 (6): 432–435.
29. *Tsai Y.S., Chen C.M.* Driven polymer transport through a nanopore controlled by a rotating electric field: off-lattice computer simulations // J Chem Phys., 2007, 126: 144910.

# ТЕХНОЛОГИИ СОЗДАНИЯ БИБЛИОТЕК ФРАГМЕНТОВ ДНК ДЛЯ NGS

Определение последовательности нуклеиновых кислот (НК) методами высокопроизводительного секвенирования, по сути, представляет собой одновременное (параллельное) прочтение последовательности нескольких миллионов разных (относительно коротких) фрагментов исходной ДНК. Наиболее популярные на рынке технологии NGS не позволяют читать непосредственно геномную или кДНК и предполагают амплификацию исходных молекул (за исключением технологий PacBio и HeliScope) (см. разд. 3.5 и 3.6).

Как уже было сказано в главе 1, перед загрузкой в прибор исследуемую ДНК необходимо модифицировать путем создания коллекции случайных фрагментов нужной структуры (а зачастую и прикрепленных к твердофазному носителю – предметному стеклу или стеклянной микрочастице). Образец ДНК, определенным образом фрагментированный и амплифицированный, называют библиотекой случайных фрагментов ДНК для NGS. Если фрагменты при этом прикреплены к предметному стеклу или микрочастице, библиотеку называют иммобилизованной.

Напомним, что общая последовательность этапов высокопроизводительного секвенирования для наиболее популярных платформ следующая:

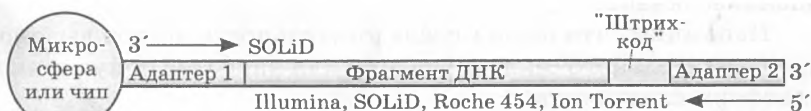
- 1) разрушение ДНК с получением фрагментов определенной длины;
- 2) присоединение синтетических олигонуклеотидных адаптеров по краям фрагментов;
- 3) наработка (амплификация) каждого фрагмента ДНК в отдельном микрореакторе с микрочастицей (эмульсионная ПЦР) и/или непосредственно на поверхности предметного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

Приготовление библиотеки фрагментов происходит на первых трех этапах и, по мнению многих пользователей, определяет 90% успеха сиквенсного проекта.

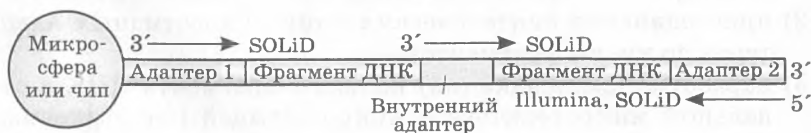
В настоящее время в NGS используют два основных типа библиотек фрагментов ДНК: обычная (paired-end library) и инвертированная (mate-pair library) (рис. 2.1). Во избежание путаницы отметим, что производители наборов реагентов для конструирования библиотек выделяют гораздо больше типов (названий) библиотек (например, деля их по виду стартового материала или числу и направлению прочтений), однако такое деление не несет под собой научного обоснования и является маркетинговым приемом. Вдобавок существует множество разночтений при переводе терминов. Например, термином «парно-концевая библиотека» иногда называют paired-end library, а иногда mate-pair library (на самом деле так можно назвать любую из библиотек в случае прочтения фрагментов двух сторон). Поэтому далее в отношении библиотек мы будем использовать только термины «обычная» и «инвертированная».

Поскольку процесс создания любой библиотеки фрагментов ДНК для NGS включает несколько стандартных этапов: фрагментирование ДНК, лигирование адаптеров, предварительная амплификация библиотеки, отбор фракции фрагментов определенной длины (size-select), нанесение на твердую фазу и т. п.,

### Обычная библиотека фрагментов



### Инвертированная библиотека фрагментов



**Рис. 2.1.** Схема обычной и инвертированной библиотек.

Чтение по технологии SOLiD может вестись в направлении 3'–5', поскольку метод не использует полимеразу. «Штрих-код» может присутствовать или отсутствовать в любой из библиотек рядом с любым из адаптеров (здесь показан только для обычной библиотеки)

рассмотрим прежде особенности каждого из этапов, а затем покажем их место в каждой из методик создания библиотек.

## 2.1. ОЧИСТКА НУКЛЕИНОВЫХ КИСЛОТ ДЛЯ NGS

В случае исследования РНК первым этапом после очистки от примесей является обратная транскрипция. Следовательно, метод очистки РНК должен эффективно удалять из биологического образца примеси, способные помешать работе ревертазы. 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, пожалуй, фенол-хлороформная экстракция [1].

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 с соавторами [2]. Этот метод основан на использовании для лизиса клеток сильного хаотропного агента – гуанидина тиоционата (GuSCN) – и последующей сорбции ДНК на твердом носителе (сейчас основанные на оксиде кремния сорбенты можно встретить под названиями «стеклянные бусы», «диатомовая земля», «стеклянное молоко» и т. д.). После промывки сорбента на нем остается чистая НК, легко снимающаяся 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 нуклеиновых кислот раствором с низкой ионной силой.

В целом требования к чистоте препарата НК при постановке NGS не слишком высоки и примерно соответствуют таковым для обычного клонирования в плазмидный вектор.

## 2.2. ОЦЕНКА КОНЦЕНТРАЦИИ НУКЛЕИНОВЫХ КИСЛОТ И ПОЛНОГЕНОМНАЯ АМПЛИФИКАЦИЯ (WGA)

Концентрацию полученных препаратов ДНК или РНК (перед обратной транскрипцией, разрушением или уже разрушенных образцов) можно измерить непосредственно по спектру поглощения нуклеиновых кислот (на спектрофотометре), по флуоресценции интеркалирующего красителя (после электрофореза или при помощи флуориметра) или методом количественной ПЦР.

Производители оборудования для NGS рекомендуют осуществлять измерение с помощью компактных флуориметров (типа Qubit от Life Technologies или Quantus от Promega) (рис. 2.2). Принцип их действия основан на измерении флуоресценции интеркалирующего в ДНК красителя типа SYBR. Точность измерения определяется 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, внутри которого флуориметр выдает достаточно точные результаты.

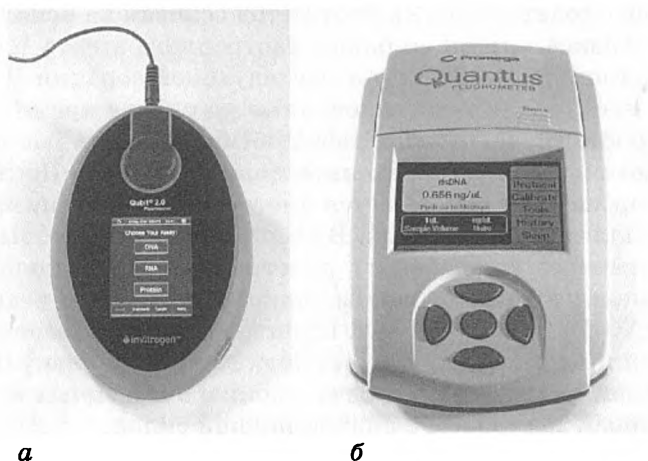


Рис. 2.2. Флуориметры Qubit (Life Technologies) (а) и Quantus (Promega) (б)

Метод количественной ПЦР 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 кДНК – можно оценивать количество одного или двух средне- или низкопредставленных транскриптов. Праймеры для таких ПЦР-систем желательно располагать близко, оставляя место только под олигонуклеотидную пробу (обычно типа TaqMan).

Отдельно следует упомянуть ситуацию, когда по каким-либо причинам невозможно получить достаточное стартовое количество ДНК или кДНК (мало биологического материала, исследование единичных клеток и т. п.). В этом случае можно перед началом создания библиотеки фрагментов амплифицировать НК одним из стандартных способов.

В случае малого количества геномной ДНК можно прибегнуть к методике полногеномной амплификации (whole genome amplification, WGA), основанной на изотермическом синтезе ДНК полимеразой *Phi29* [3]. Существуют достаточно хорошо работающие коммерческие наборы реагентов для WGA (например, REPLI-g Single Cell Kit от QIAGEN).

Если работа начинается с малого количества РНК, можно использовать амплификацию кДНК [4]. Для этих целей также подходит множество наборов реагентов, имеющихся на рынке (например, набор Mint компании «Евроген»).

В любом случае при использовании предварительной амплификации ДНК или кДНК в проектах NGS необходимо помнить об искажениях, которые неизбежно вносит этот прием. Для анализа результатов секвенирования единичных клеток разработаны специальные алгоритмы, позволяющие учесть часть искажений и получить лучшие результаты [5] (см. разд. 4.4.2).

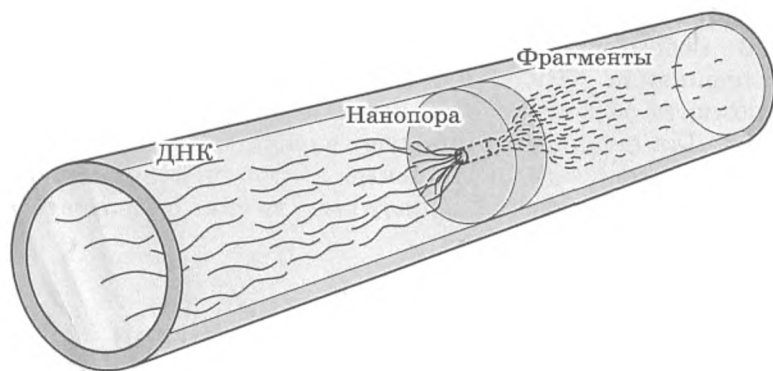
### 2.3. СПОСОБЫ РАЗРУШЕНИЯ ДНК ДЛЯ ПРИГОТОВЛЕНИЯ БИБЛИОТЕКИ

Исходным материалом для исследования могут служить любые нуклеиновые кислоты: ДНК всех типов, РНК всех типов, включая микроРНК, продукты ПЦР, отдельные области

генома (таргетное секвенирование или обогащение), комплексы нуклеиновых кислот с белками и т. д. Несмотря на различие в целях и задачах экспериментов, создание библиотеки в итоге сводится к разрушению ДНК до фрагментов определенного размера (или отбор определенной фракции из существующего спектра фрагментов), лигированию адаптеров и увеличению количества ДНК одним из методов наработки *in vitro* (амплификации). Если исходным материалом служит РНК, ее переводят в кДНК и продолжают подготовку библиотеки, как для генома.

Фрагментировать ДНК можно физическими или энзиматическими способами. В ряде случаев, например если стартовым материалом служит микроРНК или короткие продукты ПЦР, этап разрушения не требуется (однако может потребоваться этап отбора фракции фрагментов нужной длины).

Физически ДНК можно разрушить ультразвуком (часто процесс называют соникацией), пропусканием через микроотверстие (небулизацией) (рис. 2.3), гидродинамическим воздействием (гидроширингом) и др. К преимуществам физических методов перед энзиматическими можно отнести хорошую воспроизводимость результатов (стабильный диапазон длин получаемых фрагментов) даже для разного по качеству стартового материала и низкую зависимость от последовательности.



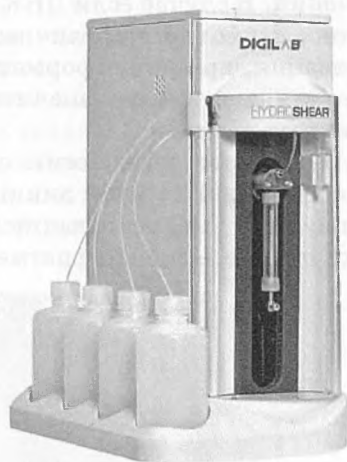
**Рис 2.3.** Принцип разрушения ДНК пропусканием через микроотверстие

Наиболее распространенными приборами для разрушения ДНК ультразвуком (в рамках технологии NGS) являются приборы фирмы Covaris (рис. 2.4). С их помощью можно проводить разрушение ДНК в диапазоне 200–1500 п. н. К недостаткам оборудования этой фирмы можно отнести высокую стоимость и сравнительно низкую производительность приборов (в случае с моделями М- и S-серий, рассчитанных на один образец).

Гидродинамические фрагментаторы, в частности Hydro-shear от Digilab (рис. 2.5), позволяют проводить эффективное разрушение в диапазоне 1–10 т. п. н. и хорошо подходят для конструирования инвертированных библиотек.



**Рис. 2.4.** Прибор для ультразвуковой фрагментации ДНК (Covaris)



**Рис. 2.5.** Прибор для гидродинамической фрагментации ДНК HydroShear (Digilab)

Энзиматические методы расщепления ДНК предполагают использование различных эндонуклеаз (наборы реагентов для энзиматического расщепления ДНК во множестве представлены на рынке как производителями платформ для секвенирования, так и биотехнологическими компаниями, предлагающими альтернативные, как правило, более дешевые решения для приготовления библиотек – например, DS Fragmentase от New England Biolabs). Чаще всего для расщепления ДНК использу-

ют смеси модифицированных нуклеаз (типа эндонуклеазы T7 и нуклеазы из *Vibrio vulnificus*), а также транспозаз.

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. Оптимизация протокола заключается в подборе периода обработки нуклеазами и температуры реакции: чем короче фрагменты нужно получить, тем дольше проводят реакцию (рис. 2.6).

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 «перерезана» или «недорезана», можно, несмотря на отличные аппаратные показатели секвенирования, при биоинформатическом анализе получить «бедные» данные, так как значительная часть ДНК оказалась 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.

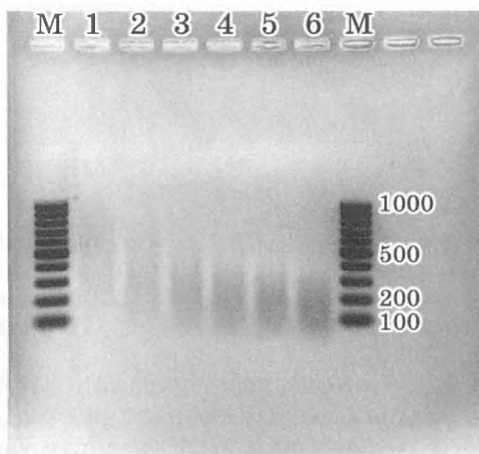


Рис. 2.6. Гель-электрофореграмма ферментативного фрагментирования геномной ДНК человека эндонуклеазой T7 с шагом инкубации в 5 минут (дорожка 1 – 10 мин, дорожка 2 – 15 мин, дорожка 3 – 20 мин, дорожка 4 – 25 мин, дорожка 5 – 30 мин, дорожка 6 – 35 мин, М – маркер длин фрагментов, п. н.)

К преимуществам ферментативного разрушения ДНК перед физическими методами можно отнести образование однотипных концов необходимой структуры («липких», или «тупых»), по которым и проводится присоединение сиквенсных адаптеров (тогда как физически разрушенная ДНК требует ферментативной «полировки» концов). Заметим, что 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 материала.

## 2.4. ОЦЕНКА ДЛИН ФРАГМЕНТОВ ДНК

Наиболее распространенным способом оценки длины фрагментированной ДНК является метод электрофореза в геле. Это может быть 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 (так называемая лаборатория-на-чипе) (рис. 2.6 и 2.7).

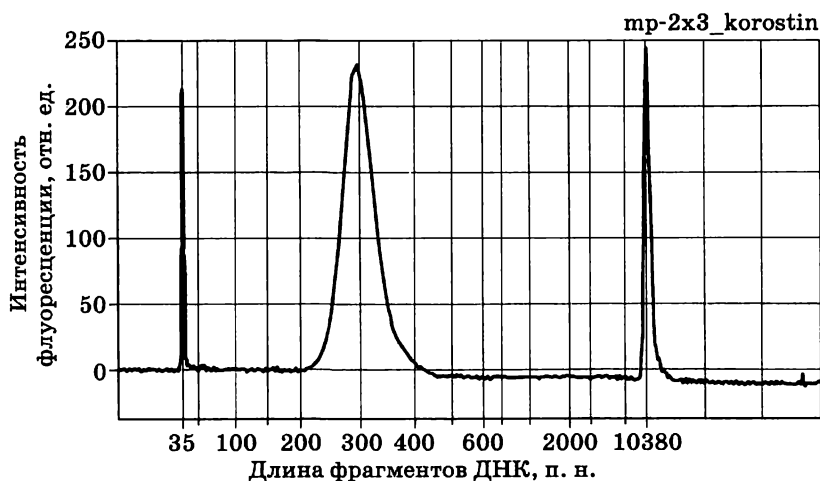


Рис. 2.7. Пример анализа фрагментированной ДНК на приборе Agilent 2100 Bioanalyzer (после отбора фракции). Видно, что средний размер фрагментов составляет около 300 п. н.

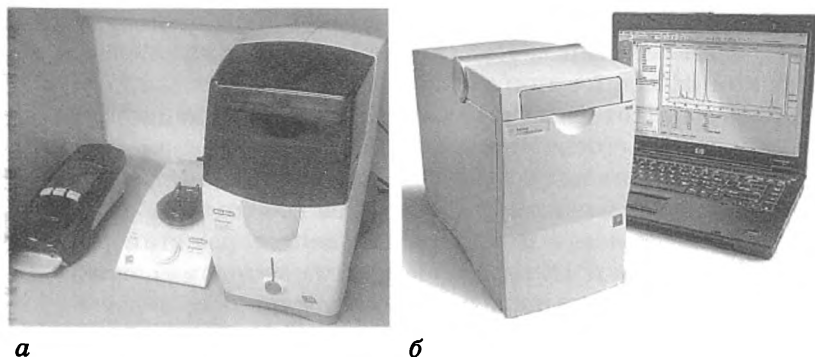


Рис. 2.8. Приборы для оценки длин фрагментированной ДНК:  
а – Bio-Rad Experion, б – Agilent 2100 Bioanalyzer

Среди встречающихся на отечественном рынке приборов для капиллярного гель-электрофореза можно отметить Bio-Rad Experion и Agilent 2100 Bioanalyzer (рис. 2.8). Эти аналоги проводят высокочувствительный электрофорез в капиллярах с регистрацией ДНК по флуоресцирующему интеркалятору. Преимуществом таких устройств перед обычным электрофорезом в агарозном геле является большая чувствительность (до 1 нг ДНК) и воспроизводимость результатов. Несмотря на довольно высокую стоимость приборов и расходных материалов для автоматического капиллярного гель-электрофореза, можно рекомендовать их использование с целью повышения стабильности работы NGS-лаборатории. Кроме оценки степени разрушения исходной ДНК это же оборудование можно использовать для выявления димеров праймеров и не сработавших адаптеров, а также определения размеров полученной библиотеки.

## 2.5. ПРИСОЕДИНЕНИЕ АДАПТЕРОВ

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.

Во всех коммерчески доступных технологиях NGS к фрагментам присоединяют два разных адаптера, причем в секве-

нирование пойдут только фрагменты ДНК, несущие по концам разные адаптеры (фрагменты с одним адаптером не будут амплифицированы из-за эффекта супрессии ПЦР) [6]. В большинстве случаев фрагменты должны иметь «тупые» концы.

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.

## 2.6. ПРЕДВАРИТЕЛЬНАЯ АМПЛИФИКАЦИЯ БИБЛИОТЕКИ

Данный этап требуется в том случае, если количество исходной ДНК гораздо ниже рекомендуемого.

Опасность любой амплификации ДНК *in vitro* (не важно, методом ПЦР или другими методами) заключается в неизбежном искажении исходной представленности фрагментов. В случае ПЦР искажение происходит вследствие разной эффективности амплификации для разных по длине и последовательности фрагментов ДНК.

Для снижения таких искажений количество циклов при амплификации библиотеки должно быть минимальным (но в любом случае не превышающим 13–15 циклов ПЦР). Большое количество циклов ПЦР для образцов геномной или кДНК ведет к снижению сложности смеси фрагментов ДНК (потере некоторых последовательностей из библиотеки) и значительному искажению результатов секвенирования.

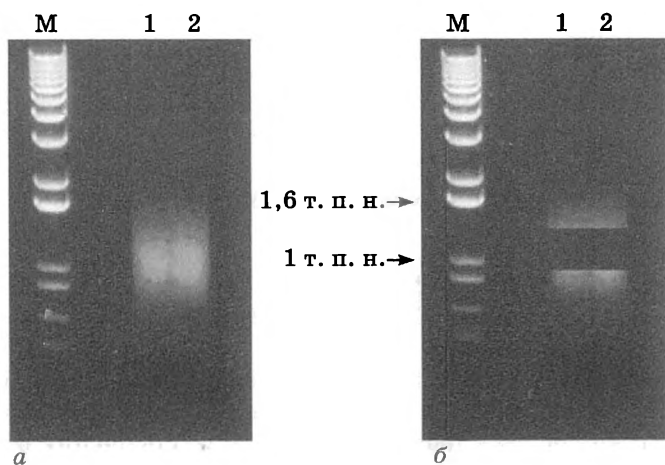
## 2.7. ОТБОР ФРАКЦИИ ФРАГМЕНТОВ НУЖНОЙ ДЛИНЫ (SIZE-SELECT)

«Сужение» профиля распределения длин фрагментов библиотеки существенно повышает эффективность ее секвенирования (так как более короткие фрагменты в сравнении с оптимальными дают слишком короткие прочтения, а более длинные искажают результаты измерений и подбор соотношений для оптимального сиквенса).

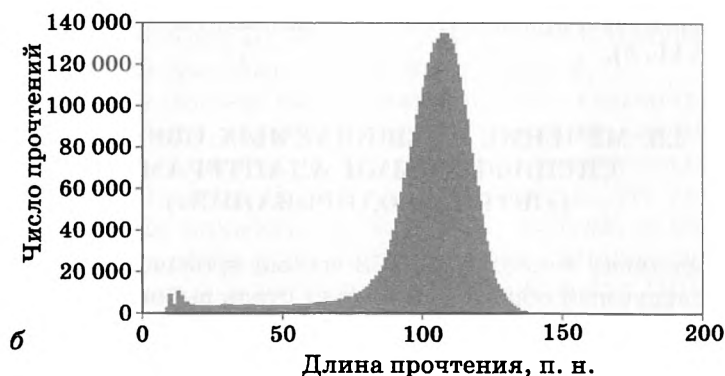
Для отбора нужной фракции библиотеки (size-select) можно использовать как обычный электрофорез в агарозном геле, так и аппаратные решения разных производителей (также на основе гель-электрофореза). Во всех случаях осуществляют разделение библиотеки ДНК в геле вместе с маркером длин (например, с шагом 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 с элюцией из него ДНК набором реагентов для экстракции из геля (например, производства «Евроген» или Qiagen) (рис. 2.9). К минусам 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.). Также практика показывает, что качество селектированной таким образом библиотеки оказывается ниже полученной с использованием автоматизированных подходов (см. ниже) (рис. 2.10).

Более предпочтительными для отбора определенной фракции фрагментов ДНК являются специализированные системы, работающие на картриджных гелях, такие как E-gel от Life Technologies Thermo Fisher Scientific или Pippin Prep от Sage



**Рис. 2.9.** Электрофореграмма до (а) и после (б) вырезания фрагментов в диапазоне от 800 до 1400 п. н.



**Рис. 2.10.** Распределение длин прочтений библиотеки фрагментов (на приборе Ion PGM с реагентами для прочтения 100 п. н.), прошедшей отбор фракции с помощью элюции из агарозного геля вручную (а) и прошедшей size-select с помощью прибора E-gel компании Life Technologies (б)

Science. E-gel представляет собой прибор в компактном корпусе, состоящий из источника тока, камеры для картриджа и транслюминатора. 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 (рис. 2.11, а). Всего процедура разделения и элюции занимает около полутора часов. Pippin Prep – полностью автоматизированный прибор,

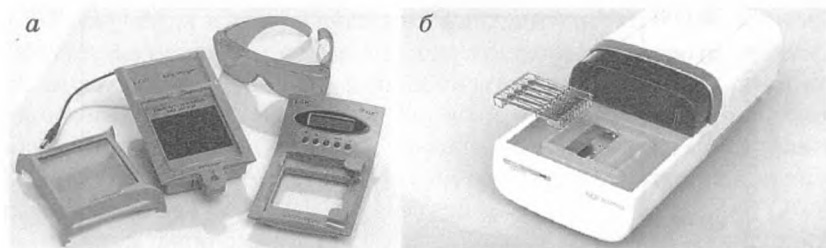


Рис. 2.11. Система для отбора фракций: *а* – E-gel (Life Technologies),  
*б* – Pippin Prep (Sage Science)

в котором диапазон нужной фракции задается исследователем на компьютере, правда, стоит эта система существенно дороже (рис. 2.11, б).

## 2.8. МЕЧЕНИЕ СМЕШИВАЕМЫХ ОБРАЗЦОВ СПЕЦИФИЧНЫМИ АДАПТЕРАМИ («ШТРИХ-КОДИРОВАНИЕ»)

Поскольку технологии NGS весьма производительны, часто исследуемый образец не требует столь высокого покрытия и объема данных на выходе, сколько дает NGS-платформа с одного запуска. 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 «штрих-кодирования» (или «бар-кодирования»), по аналогии с широко используемыми в маркировке товаров уникальными метками – штрих-кодами (barcodes).

Пример рационального использования «штрих-кодов» – секвенирование экзоса 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 экзоса экономически неоправданно.

Рациональнее пометить уникальным образом несколько разных образцов, смешать их эквимоллярно («пулировать») 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 (см. рис. 2.1). Их присоединяют к фрагментам ДНК на стадии конструирования библиотеки (при лигировании сиквенсных адаптеров к фрагментам ДНК). Таким образом «штрих-код» представляет собой участок адаптера для секвенирования, расположенный после участка отжига праймера, с которого начинается сиквенс. Как правило, для разных образцов используют один и тот же адаптер 1 (отжигающийся на микросферах для проведения эмульсионной ПЦР), и набор адаптеров 2, отличающихся штрих-кодирующими участками (см. рис. 2.1).

Важным условием проведения хорошего секвенирования для всех образцов, внесенных в одну реакцию, является их эквимоллярное смешивание, что обеспечит равномерную представленность прочтений для каждого образца. На это стоит обратить особое внимание: нужно очень аккуратно измерить концентрацию каждой библиотеки перед смешиванием (для этого можно использовать прибор Bioanalyzer или ПЦР «в реальном времени» – второй вариант предпочтителен вследствие большей точности).

## 2.9. КЛОНАЛЬНАЯ АМПЛИФИКАЦИЯ ФРАГМЕНТОВ ДНК

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.

Как уже было сказано, большинство технологий NGS недостаточно чувствительны, чтобы регистрировать сигнал от единичной молекулы ДНК, и требуют применения так называемого 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, приготовленная, например, в 96-, 384- или 1536-луночном планшете так, чтобы в среднем в каждую реакцию попала единственная стартовая молекула ДНК [6]. Таким образом, все наработанные в данной лунке фрагменты будут клонами единственной молекулы (очевидно, что в некоторые лунки попадет две или более стартовых молекул, а в некоторые – ни одной, такие реакции являются неизбежным «браком» методики).

На принципе амплификации фланкированных адаптерами отдельных фрагментов ДНК основано множество молекулярных технологий [4, 7].

### 2.9.1. Мостиковая ПЦР

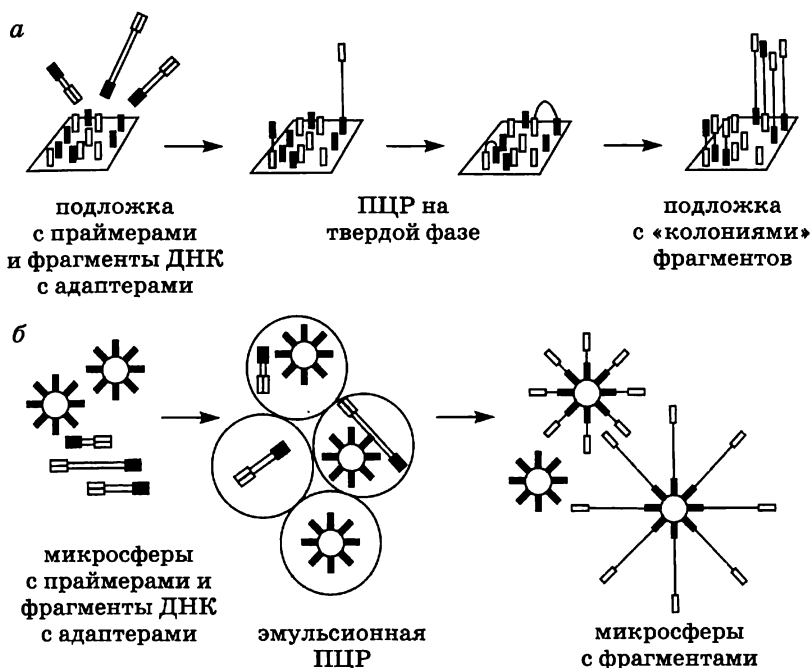
Однако для высокопроизводительного секвенирования 1536-луночного планшета (и даже сотни таких планшетов) недостаточно. Нужны миллионы отдельных реакций, в каждой из которых будет получен гомогенный продукт ПЦР.

Одним из подходов, дающих такой результат, является мостиковая ПЦР (bridge PCR) [8]. Принцип заключается в проведении ПЦР с праймерами, прикрепленными к твердой фазе (подложке). С пришитого к поверхности (наподобие предметного стекла) 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 (рис. 2.12, а). Видно, что при повторении циклов ПЦР из точки синтеза первого фрагмента колония фрагментов ДНК будет быстро «расти».

Создание клональной библиотеки фрагментов ДНК методом мостиковой ПЦР используется в технологии компании Illumina (см. разд. 3.3).

### 2.9.2. Эмульсионная ПЦР

Эмульсионная ПЦР (emulsion PCR, ePCR) представляет собой другой распространенный вариант клонирования фрагментов ДНК *in vitro*. Эта методика позволяет амплифицировать ДНК на покрытых праймерами микросферах так, что каждая микросфера оказывается «облеплена» только одним типом фрагментов [9, 10].



**Рис. 2.12.** Клональная амплификация фрагментов ДНК с применением: мостиковой ПЦР (а) и эмульсионной ПЦР (б)

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 «стартовый» фрагмент ДНК. В ходе ПЦР праймеры на микросфере затравляют синтез фрагментов (начиная с единственной мишени), а праймер в растворе достраивает вторую цепь (возможен вариант без свободного праймера, по аналогии с описанным выше вариантом мостиковой ПЦР). В итоге получают миллионы микросфер, каждая из которых несет миллионы идентичных фрагментов ДНК (рис. 2.12, б).

Для оценки качества библиотеки после эмульсионной ПЦР исходят из правила, что соотношение сработавших микросфер к несработавшим должно быть примерно 30/70.

Большой процент сработавших микросфер указывает на их высокую поликлональность. Оценить процент сработавших микросфер можно путем измерения флуоресценции несущих ДНК микросфер относительно несработавших (для этого применяют набор реагентов Ion Sphere Quality Control Kit от Life Technologies, см. разд. 7.4.6).

После проведения эмульсионной ПЦР проводят обогащение фракции сработавших микросфер, например, за счет гибридации с магнитными частицами, несущими стрептавидин (сработавшие микросферы в этом случае должны нести биотин на дистальном адаптере).

Многие пользователи считают эмульсионную ПЦР трудоемкой и весьма нестабильной. Действительно, в первых вариантах NGS 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 PGM и Ion Proton 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) (см. гл. 3).

## **2.10. ТИПЫ БИБЛИОТЕК ФРАГМЕНТОВ ДНК ДЛЯ NGS**

Специально, чтобы запутать пользователя и продать ему старый товар под новым соусом, компании постоянно придумывают новые названия давно известным вещам. Так возникли два термина, активно используемых в протоколах NGS, но не обозначающих ничего нового: «paired-end» и «mate-pair libraries». Переведем их так: обычная и инвертированная библиотеки фрагментов ДНК соответственно. Если используемая технология NGS 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 (если для обычной библиотеки это расстояние соответствует физической длине фрагмента и, как правило, не превышает 2–3 т. п. н., то для инвертированной библиотеки это могут быть и десятки тысяч пар нуклеотидов).

### **2.10.1. Обычная (paired-end) библиотека фрагментов ДНК для NGS**

Технология создания обычной, фланкированной разными адаптерами заданной структуры, библиотеки фрагментов ДНК известна давно. Такие библиотеки уже более 20 лет используют для амплификации кДНК, вычитающей гибридизации геномов, дифференциального дисплея и т. п. [4, 6, 7, 11]. Принцип создания обычной библиотеки с разными адаптерами по концам прост: надо расщепить ДНК на фрагменты требуемой длины и пришить по концам фрагментов разные по последовательности адаптеры (самый простой способ – использовать так называемые псевдо-двуцепочечные супрессионные адаптеры [6]).

Этапы создания обычной библиотеки фрагментов ДНК для NGS показаны на рис. 2.13 и включают:

- 1) фрагментирование ДНК;
- 2) лигирование сиквенсных адаптеров;
- 3) предварительную амплификацию библиотеки (требуется не всегда);
- 4) отбор фракции нужной длины (size-select);
- 5) клональную амплификацию селектированной библиотеки.

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, однако на рынке она



**Рис. 2.13.** Схема создания обычной библиотеки ДНК. Представлены два варианта – с использованием «штрих-кодирования» и без.  
Каждая линия обозначает двуцепочечную ДНК

пока не появилась. 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 перекрытия прочтений, – их общую длину (а при отсутствии перекрытия дает примерное расположение прочтений друг относительно друга). Заметим, что качество сигнала быстро падает от начала секвенирования к концу фрагмента, так что прочтение короткого фрагмента с двух сторон насквозь позволяет компенсировать низкое качество прочтения «на излете».

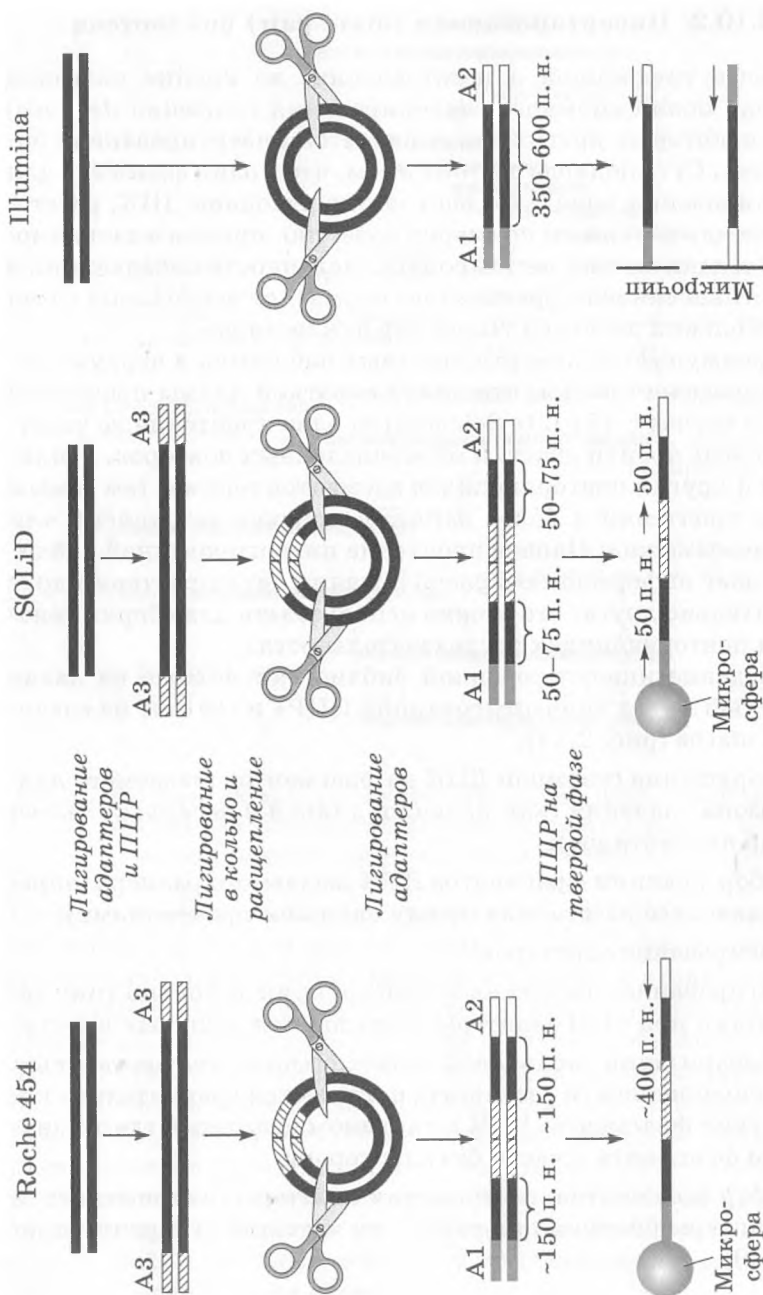
### 2.10.2. Инвертированная (mate-pair) библиотека

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.

Преимущество инвертированных библиотек в первую очередь определяется тем, что ввиду короткой длины прочтений (даже в случае с 454 Life Sciences) за одно прочтение не удастся насквозь пройти участки многочисленных повторов, дубликаций и других повторяющихся элементов генома, тем самым сборка прочтений в более длинные контиги затруднена или даже невозможна. Парное прочтение инвертированной библиотеки дает информацию о расположении двух прочтений друг относительно друга, что можно использовать для сборки геномов на повторяющихся последовательностях.

Создание инвертированной библиотеки похоже на давно известный метод «инвертированной ПЦР» и состоит из следующих шагов (рис. 2.14);

- 1) разрушение геномной ДНК до фрагментов желаемого диапазона значений (как правило, длиной в несколько тысяч пар нуклеотидов);
- 2) отбор фракции фрагментов ДНК желаемого размера (определяющего расстояние между парными прочтениями);
- 3) лигирование адаптеров;
- 4) лигирование полученных конструкций в кольцо (циклизация), при этом адаптеры оказываются «сшиты» вместе;
- 5) линейаризация кольцевой конструкции множественным расщеплением (в результате получают сравнительно короткие фрагменты ДНК с тандемом адаптеров где-то внутри фрагмента и части без адаптеров);
- 6) отбор фрагментов, содержащих адаптеры (например, если адаптеры биотинилированы – на частицы со стрептавидином);



**Рис. 2.14.** Подготовка инвертированной библиотеки для наиболее распространенных платформ NGS. Каждая линия обозначает одноцепочечную ДНК

- 7) лигирование к полученным коротким фрагментам адаптеров для секвенирования (таких же, как и в случае обычной библиотеки);
- 8) предварительная амплификация библиотеки (требуется не всегда);
- 9) отбор фракции нужной длины (size-select);
- 10) клональная амплификация селектированной библиотеки.

### СПИСОК ЛИТЕРАТУРЫ

1. *Chomczynski P., Sacchi N.* Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chlorophorm extraction // *Anal. Biochem.*, 1987, 162: 156–159.
2. *Boom R. et al.* Improved silica-guanidiniumthiocyanate DNA isolation procedure based on selective binding of bovine alphacasein to silica particles // *Journal of Clinical Microbiology*, 1999, 37: 615–619.
3. *Zhang L. et al.* Whole genome amplification from a single cell: implications for genetic analysis // *Proc Natl Acad Sci US A*, 1992, 89(13): 5847–51.
4. *Lukyanov K. et al.* Construction of cDNA libraries from small amounts of total RNA using the suppression PCR effect // *Biochem Biophys Res Commun*, 1997, 230 (2): 285–8.
5. *Bankevich A. et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // *J Comput Biol.*, 2012, 9 (5): 455–77.
6. *Lukyanov K.A. et al.* Molecule by molecule PCR amplification of complex DNA mixtures for direct sequencing: an approach to in vitro cloning // *Nucleic Acids Res.*, 1996, 24 (11): 2194–2195.
7. *Rebrikov D.V. et al.* Mirror orientation selection (MOS): a method for eliminating false positive clones from libraries generated by suppression subtractive hybridization // *Nucleic Acids Res.*, 2000, 28 (20): e90.
8. *Kawashima E.H., Farinelli L., Mayer P.* Patent: Method of nucleic acid amplification, retrieved 2012-12-22.
9. *Dressman D. et al.* 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.
10. *Williams R. et al.* Amplification of complex gene libraries by emulsion PCR // *Nature methods*, 2006, 3 (7): 545–550.
11. *Akowitz A., Manuelidis L.* A novel cDNA/PCR strategy for efficient cloning of small amounts of undefined RNA // *Gene*, 1989, 81 (2): 295–306.

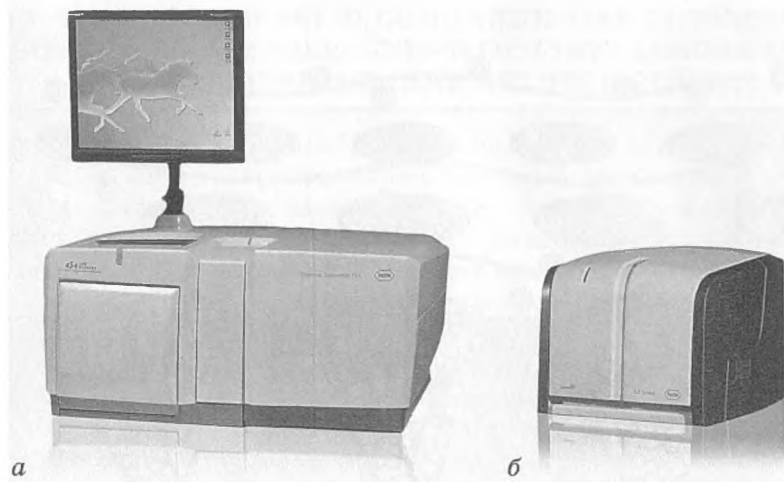
# **КОММЕРЧЕСКИЕ ТЕХНОЛОГИИ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ**

В данной главе описаны присутствующие в настоящее время на рынке секвенаторы второго и третьего поколений. Проведено сравнение технологий, их основных достоинств и недостатков.

## **3.1. ТЕХНОЛОГИЯ 454 LIFE SCIENCES КОМПАНИИ ROCHE (ЭМУЛЬСИОННАЯ ПЦР + ПИРОСЕКВЕНИРОВАНИЕ)**

Основанная на принципе пиросеквенирования платформа 454 Life Sciences [1, 2], пожалуй, является первой широкодоступной технологией секвенирования второго поколения, на базе которой в 2005 году был представлен первый коммерчески доступный автоматический секвенатор. Название технологии возникло из номера проекта в компании CuraGen Corporation и не несет какого-то скрытого смысла. Разработавшее технологию подразделение компании CuraGen Corporation в 2007 году было приобретено компанией Roche Diagnostics.

В настоящее время Roche Diagnostics предлагает две базирующиеся на технологии 454 Life Sciences модели секвенаторов: Genome Sequencer FLX+ и Genome Sequencer Junior (рис. 3.1). Первый прибор позволяет получать до 1 млн прочтений с длиной до 800–1000 п. н., второй – до 100 000 прочтений с длиной 600–800 п. н. В 2014 году производитель обещает достичь длины прочтения на обоих приборах в 1000 п. н. Длина прочтения и довольно высокое его качество – важнейшее преимущество 454 Life Sciences по сравнению с технологиями-конкурентами (не считая гораздо менее распространенных PacBio RS). Она сравнима по длине прочтения с методом Сенгера и позволяет довольно просто решать некоторые специфические задачи, например определение гаплотипов HLA при скрининговых исследованиях или



**Рис. 3.1.** Секвенаторы Genome Sequencer FLX+ (а) и Genome Sequencer Junior (Roche) (б)

метагеномные исследования микробиомов на основе анализа гена 16S рибосомальной РНК.

Подготовку образцов проводят стандартным способом: исследуемый образец ДНК расщепляют (например, пропуская через небольшое отверстие под высоким давлением газа (азота) – методика небулизации). Затем к полученным фрагментам (длиной около 1000 п. н.) присоединяют разные по последовательности адаптеры и проводят клональную наработку фрагментов ДНК (методом эмульсионной ПЦР, см. разд. 2.9.2). Полученные сферы с однотипными молекулами ДНК помещают в матрицу с микрореакторами так, что в каждую ячейку попадает только одна сфера (рис. 3.2, а; см. также рис. 1.9). Матрицу размещают в секвенаторе, где осуществляется последовательная подача всех четырех типов нуклеотидов и детекция уровня экстинкции света от каждой пиколитровой ячейки с расположенной в ней микросферой (рис. 3.2, б). Свет передается по световоду к ПЗС-матрице (по сути, как у цифрового фотоаппарата). Несмотря на то что время осуществления всей последовательности химических реакций измеряется долями секунды, общее время одного цикла (добавление нуклеотида, каскад химических реакций, детекция света, обработка апиразой для удаления остатков дНТФ) лимитируется скоростью доставки реактивов до каждой отдельной

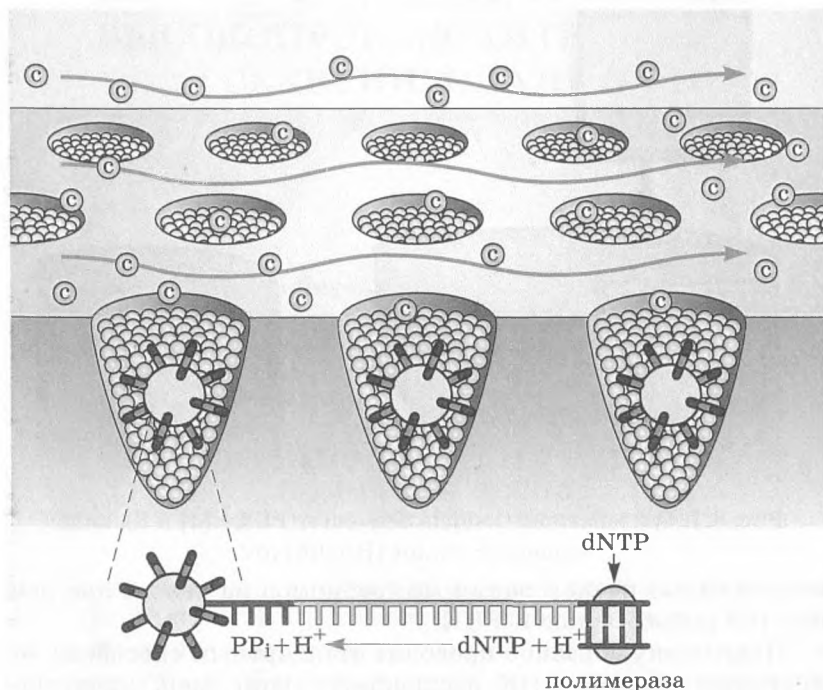


Рис. 3.2. Принцип работы технологии 454 Life Sciences

пиколитровой ячейки и составляет уже минуты, что с учетом числа циклов дает в итоге порядка 10 ч работы.

Как и технология Ion Torrent, 454 Life Sciences имеет проблемы с прочтением гомополимеров. Дела тут обстоят чуть лучше, но все же 8–10 повторяющихся нуклеотидов приводят к ошибке в подсчете точного числа оснований в повторе.

Несмотря на статус пионера и отставание от конкурентов по производительности и стоимости определения одного нуклеотида, технология 454 Life Sciences лучше других подходит в тех случаях, когда важна длина одного прочтения. По сути, это аналог капиллярных секвенаторов, но на порядки более производительный.

С 2016 года Roche снимает с производства секвенаторы на основе технологии 454 Life Sciences и делает ставку на технологии третьего поколения (в частности, на совместные разработки с Pacific Biosciences, см. разд. 3.5).

### 3.2. ТЕХНОЛОГИЯ SOLiD КОМПАНИИ LIFE TECHNOLOGIES THERMO FISHER SCIENTIFIC (ЭМУЛЬСИОННАЯ ПЦР + СЕКВЕНИРОВАНИЕ ЛИГИРОВАНИЕМ)

Технология SOLiD (sequencing by oligonucleotide ligation and detection – секвенирование с помощью лигирования и детекции олигонуклеотидов) основана на методе, разработанном исследователями Гарвардского университета, США [3]. В настоящее время она принадлежит компании Thermo Fisher Scientific. Сейчас компания предлагает платформу SOLiD в двух вариантах: SOLiD 5500 (на один чип) и SOLiD 5500xl (на два чипа) (рис. 3.3). Также существует версия SOLiD 5500 Wildfire, в которой вместо эмульсионной ПЦР для клональной амплификации применяется мостиковая ПЦР, однако эта платформа не получила широкого распространения.



Рис. 3.3. Секвенатор SOLiD 5500xl  
(Life Technologies Thermo Fisher Scientific)

В отличие от большинства других платформ технология SOLiD использует принцип секвенирования с помощью лигирования (без применения ДНК-полимеразы). Это позволяет «читать» одну и ту же цепь ДНК в любом направлении (тогда как методы на основе ДНК-полимераз могут «читать» только в направлении растущего 3'-конца затравки).

Подготовка библиотек аналогична таковой в технологии 454 Life Sciences. Методом эмульсионной ПЦР (см. разд. 2.9.2) создают иммобилизованную на микрочастицах библиотеку фрагментов ДНК (так, что каждая микрочастица несет множество одинаковых фрагментов ДНК). Микрочастицы впрыскивают в плоскую прозрачную камеру (наподобие расположенных на расстоянии нескольких микрометров двух предметных стекол) так, чтобы микрочастицы монослоем распределились между стеклами (такую камеру называют проточной – flow cell или flow chip – поскольку с одной стороны в нее можно впрыскивать реагент, с другой стороны он вытекает (рис. 3.4, а).

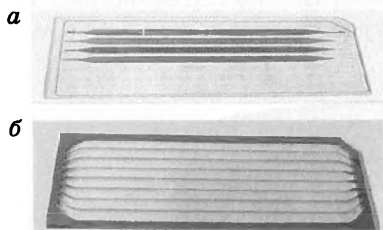


Рис. 3.4. Проточные чипы для аппаратов SOLiD 5500 (а) и Illumina (б). В чипе для SOLiD 5500 микросферами загружены четыре дорожки из шести (темные на фото). Размеры чипов – примерно 5 × 10 см

Метод секвенирования лигированием по технологии SOLiD основан на использовании флуоресцентно-меченых (с помощью четырех красителей) октамерных олигонуклеотидов (проб), у которых специфичны только два основания с 3'-конца, остальные позиции вырождены. Таким образом, всего существует 16 типов проб, по 4 каждого цвета (см. рис. 1.8).

Секвенирование начинается с отжига праймера, комплементарного адаптеру на одном из концов библиотеки ДНК. Затем в проточную ячейку добавляют все 16 типов проб, которые находят себе места посадки на фрагментах ДНК по принципу комплементарности. После отжига проб лигаза зашивает разрыв между пробой и праймером (в том случае, если проба отоглась вплотную к праймеру). Несвязавшиеся пробы смывают, а у связавшихся прибор детектирует флуоресценцию.

Затем три 5'-концевых нуклеотида пробы вместе с флуорофорами химически отсоединяются и удаляются из проточной ячейки, оставляя только часть с известным динуклеотидом и тремя случайными основаниями, а на 5'-конце появляется

место для присоединения новой пробы. По прошествии 10–15 последовательных лигирований производится так называемая перезагрузка праймера – новосинтезированная цепь отплавляется от матрицы, а на ней отжигается праймер, отличающийся по длине от первого на один нуклеотид. Таких перезагрузок праймеров проводят 4, что приводит к гарантированному прочтению каждого основания исходной матрицы два раза в разных пробах. Технология SOLiD является довольно производительной (до 300 Гбайт данных за один запуск автоматического секвенатора), но сильно уступает по длине прочтения Illumina (75 нуклеотидов у SOLiD против 250 у Illumina).

Из-за высокой стоимости секвенирования, сложности и длительности процедур в настоящее время технология SOLiD не привлекает особого интереса, и можно прогнозировать ее скорый уход с рынка.

### **3.3. ILLUMINA GENOME ANALYSER КОМПАНИИ ILLUMINA (МОСТИКОВАЯ ПЦР + СЕКВЕНИРОВАНИЕ СИНТЕЗОМ)**

Первоначально технология секвенирования синтезом была разработана Баласубраманияном и Кленерманом в Кембриджском университете в 1998 году [4]. В том же году они основали компанию Solexa с целью коммерциализации метода секвенирования. В 2004 году компания Solexa приобрела компанию Manteia Predictive Medicine, среди прочего для того, чтобы получить более мощную технологию параллельного секвенирования на основе ДНК-колоний (получаемых с использованием клональной амплификации ДНК на поверхности). В этом способе праймеры прикреплены к подложке, а отжигающиеся на них молекулы ДНК нарабатываются полимеразой. В результате мостиковой ПЦР образуются ДНК-колонии или колонии ДНК (см. разд. 2.9.1, рис. 2.12).

В 2006 году Solexa выпустила свой первый коммерческий Genome Analyzer. Машина давала 1 млрд п. н. за один запуск (~4 дня работы прибора) [5]. В 2007 году компания Illumina приобрела Solexa, и в 2011 году был выпущен обновленный вариант секвенатора – HiSeq 2500, дающий 200 Гбайт данных за один запуск (по 25 Гбайт в день). Прибор позволяет делать парные прочтения ( $2 \times 100$  п. н.) или чтение с одной стороны (рис. 3.5, а).

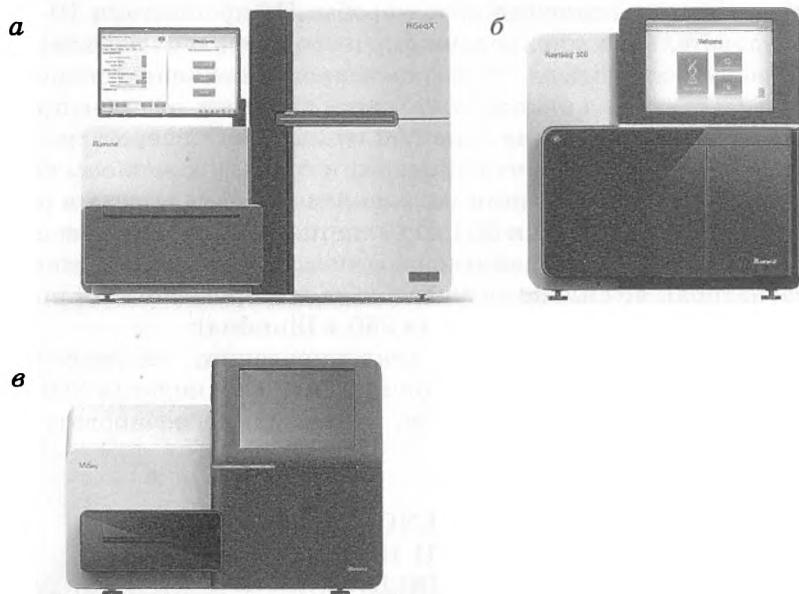
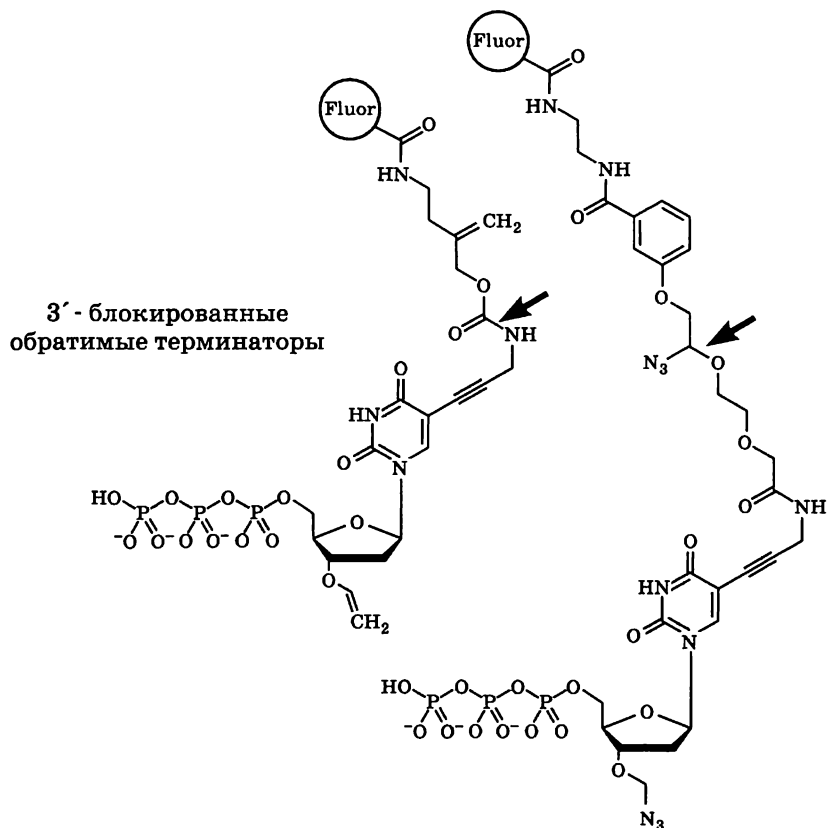


Рис. 3.5. Секвенаторы HiSeq X (а), NextSeq (б) и MiSeq (в) (Illumina)

В 2012 году Illumina выпустила на рынок секвенатор MiSeq с пониженной частотой ошибок, что позволяет выполнять более длинные прочтения (до 250 п. н.). Прибор позиционируется для небольших научно-исследовательских или клинических лабораторий, так как позволяет за относительно небольшие деньги (если иметь в виду стоимость запуска прибора, а не стоимость прочтения миллиона пар нуклеотидов) секвенировать экзом человека (рис. 3.5, в). В 2014 году вышел средний по производительности NextSeq 500. Результаты прочтения на секвенаторах Illumina обычно имеют средний процент ошибок  $<1\%$ .

Библиотеку фрагментов для всех трех аппаратов готовят на основе мостиковой ПЦР. Секвенирование на аппаратах от Illumina осуществляется с использованием четырех меченых разными флуорофорами нуклеотидов, 3'-конец которых заблокирован (так называемых обратимых терминаторов), что не позволяет включать ДНК-полимеразе в синтезируемую цепь более одного основания за цикл (рис. 3.6).



**Рис. 3.6.** Схема строения обратимых терминирующих нуклеотидов, используемых в секвенировании по технологии Illumina. Fluor – флуоресцирующая метка. Стрелками показано место отщепления группы с флуорофором

Каждый цикл состоит из следующих этапов (см. рис. 1.10):

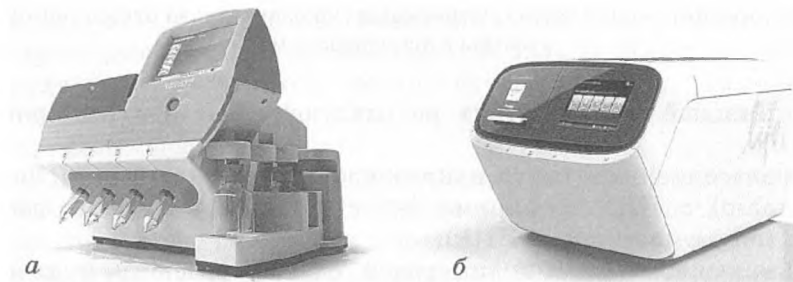
- 1) внесение всех типов нуклеотидов в проточный чип (flow chip) с ПЦР-колониями (кластерами) на твердой фазе (после мостиковой ПЦР);
- 2) включение ДНК-полимеразой одного меченого нуклеотида;
- 3) детекция флуоресценции от каждой ПЦР-колонии;
- 4) снятие 3'-блока вместе с флуорофором.

Последовательность операций для приборов HiSeq, Next-Seq и MiSeq одинакова.

Платформа Illumina выгодно отличается от конкурентов более простой методикой подготовки библиотек ДНК к секвенированию (так как лишена этапов, связанных с проведением эмульсионной ПЦР), а также возможностью получать прочтения до 250 п. н., причем с двух сторон.

### **3.4. ПЛАТФОРМЫ ION PGM И ION PROTON КОМПАНИИ LIFE TECHNOLOGIES THERMO FISHER SCIENTIFIC (ЭМУЛЬСИОННАЯ ПЦР + ПОЛУПРОВОДНИКОВОЕ СЕКВЕНИРОВАНИЕ)**

В начале 2010 года компанией Ion Torrent Systems Inc. была представлена технологическая платформа, разработанная по лицензии DNA Electronics Ltd. и основанная на принципе регистрации изменения проводимости среды за счет появляющихся в результате акта присоединения нуклеотида протонов. Первый полупроводниковый секвенатор, Ion PGM – Personal Genome Machine (рис. 3.7, а) – был представлен Ion Torrent Inc. в 2010 году (в настоящее время вся технология принадлежит компании Life Technologies, в феврале 2014 года приобретенной Thermo Fisher Scientific). В 2012 году на рынке появился усовершенствованный (более мощный) аппарат Ion Proton, работающий на том же принципе (рис. 3.7, б).



**Рис. 3.7. Секвенаторы Ion PGM (а) и Ion Proton (б)  
(Life Technologies Thermo Fisher Scientific)**

Основы технологии полупроводникового секвенирования описаны в разд. 1.1.8, поэтому здесь мы остановимся на практических особенностях реагентной и приборной базы данной технологической платформы.

Подготовка образцов для Ion PGM и Ion Proton является стандартной: исследуемый образец ДНК расщепляют любым способом до фрагментов длиной 100–500 п. н. в зависимости от используемого типа реагентов для секвенирования (см. ниже). Затем к полученным фрагментам присоединяют разные по последовательности адаптеры и проводят клональную наработку фрагментов ДНК (методом эмульсионной ПЦР): фрагменты амплифицируют на ISP-сферах (ion sphere particles), на которых закреплены комплементарные одному из адаптеров праймеры. Как было сказано выше, для эмульсионной ПЦР важнейшим условием является соблюдение соотношения концентрации сфер и фрагментов ДНК – в случае избытка фрагментов в один микрореактор эмульсионной ПЦР попадет более чем один фрагмент ДНК, тем самым сфера окажется поликлональной и даст множественный сигнал. При малом количестве фрагментов падает выход сработавших микросфер, «облепленных» фрагментами ДНК, и секвенирование не даст ожидаемого количества данных (см. разд. 2.9.2).

В настоящее время эмульсионную ПЦР для платформ Ion PGM Torrent и Ion Proton можно проводить в автоматическом режиме с помощью приборов серии OneTouch (рис. 3.8). Приборы OneTouch сконструированы таким образом, что их «заряжают» реактивами на несколько запусков одного типа: например, набор реагентов для эмульсионной ПЦР от Ion Proton рассчитан на проведение 8 реакций. Технологически возможен переход с одного типа реагентов на другой (например, если в лаборатории имеется еще и Ion PGM) и для этой процедуры разработан специальный протокол. Однако ее проведение сопряжено с некоторыми потерями масла и recovery-реагента, что при частых переходах приведет к тому, что пользователю не хватит набора на обещанные 8 реакций. Поэтому логичным видится планирование работ, при котором последовательно будет проводиться эмульсионная ПЦР одного типа (на одинаковых реагентах).

В 2013 году появилась полностью автоматизированная система пробоподготовки – Ion Chef, выполняющая эмульсионную ПЦР, обогащение и загрузку чипа в автоматическом



**Рис. 3.8.** Система OneTouch 2 (Life Technologies Thermo Fisher Scientific) для проведения эмульсионной ПЦР. Слева – модуль ES, служащий для обогащения фракции сработавших сфер, справа – OneTouch 2 для проведения эмульсионной ПЦР

режиме. Но покупка такой станции оправдана только в высокопроизводительных лабораториях, где работа секвенаторов осуществляется в непрерывном режиме.

Стоит отметить, что Life Technologies Thermo Fisher Scientific ведет разработку технологии амплификации библиотеки для платформ Ion PGM и Ion Proton без использования эмульсионной ПЦР, по-видимому, идеологически сходной с мостиковой ПЦР, используемой в технологии Illumina.

В настоящее время для Ion PGM поставляются реагенты, обеспечивающие длину прочтения в 200 или 400 нуклеотидов. Регулировать длину прочтения можно в ручном режиме, и практика показывает, что на тех же реагентах можно добиться прочтений в 250 и 440 п. н. с достаточно хорошим качеством чтения последних 50 нуклеотидов. Также для PGM предлагаются три варианта чипов: 314, 316 и 318, имеющие 1,2, 6,2 и 11,1 млн пикселей (ячеек, в которые загружаются обогащенные микросферы), соответственно, обеспечивающие приблизительно 3–5-кратное увеличение масштаба получаемых данных. К примеру, на чипе 318 с реагентами для про-

чтения 400 п. н. можно получить порядка 3 млрд п. н. необработанных данных.

Напомним, что чип для полупроводникового секвенирования представляет собой матрицу из микрореакторов, размер которых таков, что в один реактор может поместиться только одна ISP-сфера с прикрепленными к ней (одинаковыми) молекулами ДНК (см. рис. 1.11). Такая частица несет на себе несколько миллионов одинаковых (клонированных) ДНК единственной исходной молекулы. Каждый микрореактор расположен над ISFET-сенсором (ISFET – ion sensitive field-effect transistor, канальный транзистор измерения концентрации ионов), передающим сигнал от возникающих ионов водорода в программу, переводящую его в последовательность нуклеотидов.

Полупроводниковые микрочипы обладают гораздо большим разрешением в сравнении с оптическими системами детекции: самый «слабый» чип, 314, на первом полупроводниковом приборе Ion PGM имеет разрешение в 1,2 млн микрореакторов, а чип PI у следующей версии данной технологии – Ion Proton, незначительно отличающийся от чипа 314 размерами, несет уже около 150 млн микрореакторов. Производитель анонсирует выпуск реагентов для прочтений в 600 п. н., что существенно расширит спектр применения PGM.

Важным преимуществом полупроводникового секвенирования, помимо достаточно простой и дешевой реагентной составляющей, является скорость: один цикл реакции, включающий введение раствора нуклеотида, включение комплементарных нуклеотидов, измерение сигнала и очистку чипа от несвязавшихся молекул нуклеотида, занимает всего несколько секунд (а время прохождения секвенирования – всего несколько часов). К примеру, секвенирование 200 п. н. на чипе 318 осуществляется в течение 4 ч и на выходе дает более 1 млрд п. н. данных. Детекция сигнала происходит сразу в виде цифрового значения, т. е. не требуется анализировать изображение, полученное с микроскопа.

Для присутствующего на рынке с 2012 года более производительного варианта – Ion Proton – пока выпускаются только чипы PI в 150 мегапикселей, а реагенты доступны с длиной прочтения до 200 п. н. Таким образом, производительность Proton сейчас достигает 10–12 млрд п. н. за запуск. В 2014 году должны появиться чипы следующего масштаба, PII, которые, по-видимому, обеспечат 3–5-кратное увеличе-

ние производительности прибора. Пробоподготовка для Ion Proton такая же, как и на Ion PGM.

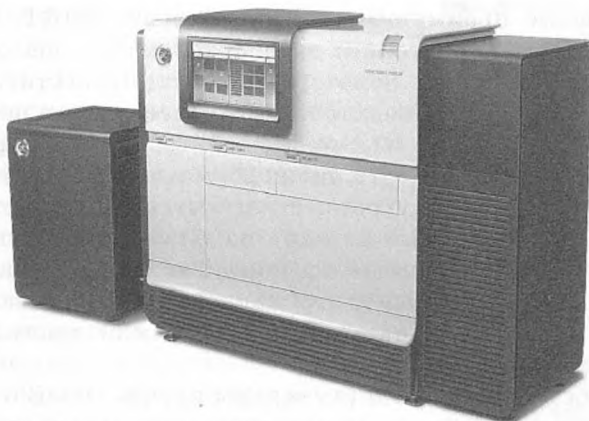
Проблемой полупроводникового секвенирования остается чтение гомополимеров: если во фрагменте встречается повтор из более чем пяти одинаковых оснований, выделяемое количество протонов не позволяет точно определить число нуклеотидов, и результаты чтения оснований после гомополимера отбраковываются («тримминг») программным обеспечением прибора. Кроме того технология Ion Torrent крайне чувствительна к качеству воды, из которой готовят рабочие буферы. mQ-Вода на открытом воздухе быстро теряет качество из-за растворения в ней диоксида углерода, этого буфера может быть вполне достаточно для снижения чувствительности прибора при секвенировании, поэтому в непосредственной близости от секвенатора должен располагаться источник mQ-воды.

Заметим, что система подачи реактивов в PGM и Proton пневматическая и работает на инертных газах (аргон или азот), поэтому, несмотря на его настольный формат, следует иметь в виду, что рядом с секвенатором должен располагаться баллон высокого давления, по правилам хранящийся в специальном металлическом шкафу. Также оператор PGM обязан иметь сертификаты, удостоверяющие его навыки работы с газобаллонным оборудованием.

### **3.5. ПЛАТФОРМА PacBio КОМПАНИИ PACIFIC BIOSCIENCES (СЕКВЕНИРОВАНИЕ СИНТЕЗОМ ОДИНОЧНЫХ МОЛЕКУЛ)**

Технологии секвенирования одиночных молекул только начинают выходить на рынок, но им можно прочесть большое будущее. Недаром некоторые авторы называют методы, основанные на прочтении единичных молекул нуклеиновых кислот, методами секвенирования третьего поколения (в сравнении с технологиями второго поколения, перечисленными выше).

Pacific Biosciences – один из поставщиков оборудования и реагентов для секвенирования отдельных молекул ДНК. Компания была создана в 2004 году на базе лабораторий Корнелльского университета США, а в 2011 году выпустила свой первый коммерчески доступный секвенатор PacBio RS. Весной 2013 года вышла вторая версия прибора – PacBio RSII (рис. 3.9). На начало 2014 года в мире установлено порядка



**Рис. 3.9.** Секвенатор одиночных молекул PacBio RSII  
(Pacific Biosciences)

80 таких устройств. В данный момент это нишевый прибор, используемый в качестве дополнительного устройства к приборам Illumina или Ion Torrent.

Технология секвенирования PacBio базируется на принципах ближнего (эванесцентного) поля – апертура имеет размер порядка нескольких нанометров. Фактически такое устройство представляет собой совокупность тысяч флуоресцентных микроскопов. Метод получил название SMRT (single molecule real time) – секвенирование одиночных молекул «в реальном времени», и относится к классу «секвенирования синтезом», т. е. с использованием ДНК-полимеразы.

В основе метода SMRT лежит разработанная основателями компании технология оптической детекции включения отдельных флуоресцентно-меченых нуклеотидов в строящуюся цепь молекулы ДНК. По сути, в чипе для секвенирования проделаны сверхмалые рабочие камеры (zero-mode waveguides – ZMW), на дне которых закреплена молекула ДНК-полимеразы  $\phi 29$ . При синтезе цепи ДНК происходит включение нуклеотидов по принципу комплементарности. Каждый нуклеотид мечен своим флуорофором, отсоединяющимся от основания при образовании фосфодиэфирной связи (в этот момент прибор фиксирует флуоресценцию отделенной единичной молекулы флуорофора, а также длительность испускания света).

Основным преимуществом технологии SMRT является огромная длина прочтений: в среднем сейчас она составляет 8500 п. н., причем некоторые прочтения достигают длины в 30 т. п. н. Производитель утверждает, что в перспективе по этой технологии он сможет добиться длины прочтения в 70 т. п. н. Очевидно, что биоинформатический анализ массивов данных такого уровня с легкостью позволяет прочитать многие проблемные для других платформ участки генома – к примеру, tandemные повторы. Также анализ кинетики включения нуклеотидов в строящуюся цепь, по заявлению разработчиков, позволяет проводить прямой анализ уровня метилирования геномов.

К недостаткам PacBio (во всяком случае, PacBio RS) стоит отнести низкую точность прочтения отдельных оснований. До недавнего времени для качественного секвенирования результаты PacBio приходилось в обязательном порядке объединять с результатами по тому же эксперименту, полученными на другой платформе. Возможно, новая версия прибора и реагентов для него решит этот вопрос, что станет понятно после появления публикаций от независимых коллективов. Также к минусам PacBio можно отнести громоздкость прибора – он весит более тонны.

Насколько известно авторам, в России пока нет секвенаторов от Pacific Biosciences.

### **3.6. ПЛАТФОРМА HeliScope КОМПАНИИ HELICOS BIOSCIENCES (СЕКВЕНИРОВАНИЕ СИНТЕЗОМ ОДИНОЧНЫХ МОЛЕКУЛ)**

Еще одна технология секвенирования отдельных молекул представлена компанией Helicos Biosciences. Компания была основана в 2003 году (в настоящее время является банкротом и не ведет деятельности). В 2008 году компания выпустила прибор HeliScope (рис. 3.10), который на то время обладал выдающимися характеристиками и являлся первым аппаратом, проводящим секвенирование одиночных молекул по технологии tSMS (true single molecule sequencing, истинное секвенирование отдельных молекул), т. е. без предварительной аппликации ДНК.

Технология HeliScope похожа на технологию Illumina. Определение последовательности нуклеиновой кислоты осно-

вано на регистрации флуоресцентного сигнала при присоединении нуклеотида. Единственное серьезное отличие от Illumina состоит в том, что на чипе Helicos не требуется амплификация. Отдельные фрагменты из библиотеки присоединяются к поверхности без мостиковой амплификации. Фрагменты ДНК распределяются таким образом, чтобы в один пиксель матрицы попадало не более одной молекулы, чтобы можно было зарегистрировать единичный сигнал.

Для секвенирования фрагментированную ДНК модифицируют с 3'-конца добавлением участка поли-А, чтобы матрица (денатурированная, в виде одноцепочечных фрагментов) могла отжечься на иммобилизованных на подложке праймерах (олиго-dT). Секвенирование осуществляется с помощью ДНК-полимеразы и четырех флуоресцентно-меченых нуклеотидов, которые добавляются в реакцию последовательно в течение одного цикла секвенирования. Детекция сигнала флуоресценции от отдельных нуклеотидов осуществляется в конце каждого цикла с помощью четырех лазеров и системы, сходной с конфокальным микроскопом. Так как молекулы ДНК неподвижны на чипе для секвенирования, в процессе секвенирования получается набор изображений, на которых имеются (в случае включения соответствующего нуклеотида) яркие точки – флуорофоры, излучающие свет под воздействием лазера. Сопоставляя сигнал на изображении с координатами точек, программное обеспечение Helicos определяет последовательность отсеквенированных нуклеотидов для каждого из фрагментов матрицы. Максимальная длина прочтения у Helicos составляет 55 нуклеотидов и не имеет блестящей точности, что связано в первую очередь с высоким уровнем шума.

На начало 2014 года в мире установлено лишь около 20 приборов. Насколько известно авторам, в России нет секвенаторов Heliscope.

В табл. 3.1 приведены показатели производительности, цена и ключевые параметры технологии для присутствующих на рынке платформ NGS.



Рис. 3.10. Секвенатор HeliScope (Helicos Biosciences)

Таблица 3.1

## Сравнение производительности и цены разных технологий

Технология	Прибор	Клональная ПЦР	Принцип действия	Чтение с 2 сторон	Длина прочтения, п. н.	Производительность, млн п. н. в день	Продолжительность одного запуска, дней	Цена за млн п. н., руб.	Цена оборудования, млн руб.
Roche 454 Life Sciences	454 Genome Sequencer FLX+	эмульсионная	полимераза (пиросеквенирование)	нет	800	750	1	600	10
	454 Genome Sequencer Junior	эмульсионная	полимераза (пиросеквенирование)	нет	500	40	0,4	1500	5
SOLiD	SOLiD 5500	эмульсионная	лигаза (меченые октамеры)	есть	75	20 000	7–14	5	20
Illumina	HiSeq 2500 (2 чипа)	мостиковая	полимераза (обратимые терминаторы)	есть	100	50 000	3–10	4	20
	MiSeq	мостиковая	полимераза (обратимые терминаторы)	есть	300	4 700	0,5	5	10

Окончание табл. 3.1

Технология	Прибор	Клональная ПЦР	Принцип действия	Чтение с 2 сторон	Длина прочтения, п. н.	Производительность, млн п. н. в день	Продолжительность одного запуска, дней	Цена за млн п. н., руб.	Цена оборудования, млн руб.
Ion Torrent	Ion PGM – 314 чип	эмульсионная	полимераза (полу-проводник)	нет	200	120	0,1	400	5
	Ion PGM – 318 чип	эмульсионная	полимераза (полу-проводник)	нет	200	1 600	0,1	40	5
	Ion Proton – P1 чип	эмульсионная	полимераза (полу-проводник)	нет	150	22 000	0,1	7	10
Pacific Biosciences	PacBio RSII	нет (одна молекула)	полимераза (обратимые терминаторы)	нет	30 000	5 000	0,1	20	40
Helicos Biosciences	Heliscope	нет (одна молекула)	полимераза (обратимые терминаторы)	нет	50	3 000	–	30	30
Dover Systems	Polonator	эмульсионная	лигаза (меченые нонамеры)	есть	13	30 000	–	10	500

## СПИСОК ЛИТЕРАТУРЫ

1. *Ronaghi M. et al.* A sequencing method based on real-time pyrophosphate // *Science*, 1998, 281 (5375): 363.
2. *Margulies M. et al.* Genome Sequencing in Open Microfabried High Density Picoliter Reactors // *Nature*, 2005, 437 (7057): 376–80.
3. *Shendure J. et al.* Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome // *Science*, 2005, 309 (5741): 1728–1732.
4. *Osborne M.A., Furey W.S., Klenerman D., Balasubramanian S.* Single-molecule analysis of DNA immobilized on microspheres // *Anal Chem.*, 2000, 72(15): 3678–81.
5. *Bentley D.R. et al.* Accurate whole human genome sequencing using reversible terminator chemistry // *Nature*, 2008, 456 (7218): 53–59.

# ОБЩИЕ ПРИНЦИПЫ ОБРАБОТКИ ДАННЫХ NGS

Предыдущие три главы описывали инструментальную и технологическую часть получения первичных (необработанных) данных о последовательностях нуклеиновых кислот. Но 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. Данная глава рассматривает особенности обработки первичных данных NGS и некоторые аспекты анализа последовательностей с биологической точки зрения.

## 4.1. ОЦЕНКА КАЧЕСТВА ПЕРВИЧНЫХ ДАННЫХ

Как уже было сказано в главе 3, приборы для определения последовательности ДНК (автоматические секвенаторы) в процессе работы считывают некоторый сигнал, соответствующий определенному нуклеотиду в последовательности ДНК: электрический (технология полупроводникового секвенирования) или оптический (большинство других технологий NGS). Исследователь же в большинстве случаев имеет дело с уже обработанными данными в виде последовательности букв А, Т, G, С и приписанных каждой букве значений, соответствующих достоверности полученного результата. Прибор сохраняет результаты в файл определенного формата. Наиболее распространенные форматы файлов приведены в табл. 4.1.

Таблица 4.1

**Наиболее распространенные форматы файлов  
для записи данных NGS**

<b>Формат файла</b>	<b>Адрес подробного описания формата данных</b>	<b>Использующие формат программы</b>	<b>Тип формата данных</b>
ALN	<a href="http://www.biostat.jhsph.edu/~hji/cisgenome/index_files/tutorial_seqpeak.htm">http://www.biostat.jhsph.edu/~hji/cisgenome/index_files/tutorial_seqpeak.htm</a>	CisGenome	TSV
BED	<a href="http://genome.ucsc.edu/FAQ/FAQformat.html#format1">http://genome.ucsc.edu/FAQ/FAQformat.html#format1</a>	UCSC BEDTools	TSV
FASTA	<a href="http://en.wikipedia.org/wiki/FASTA_format">http://en.wikipedia.org/wiki/FASTA_format</a>	FASTA	TSV
FASTQ	<a href="http://en.wikipedia.org/wiki/FASTQ_format">http://en.wikipedia.org/wiki/FASTQ_format</a>	SAMTools	TSV
CFF3	<a href="http://gmod.org/wiki/GFF3">http://gmod.org/wiki/GFF3</a>	Gmod	TSV
SAM/ BAM	<a href="http://samtools.sourceforge.net/SAM-13.pdf">http://samtools.sourceforge.net/SAM-13.pdf</a>	SAMTools	TSV/ Binary
USEQ	<a href="http://useq.sourceforge.net/useqArchiveFormat.html">http://useq.sourceforge.net/useqArchiveFormat.html</a>	IGB	Binary
WIG	<a href="http://genome.ucsc.edu/goldenPath/help/wiggle.html">http://genome.ucsc.edu/goldenPath/help/wiggle.html</a>	UCSC	TSV

Рассмотрим организацию файла с данными на примере наиболее распространенного формата секвенирования – FASTQ. Для каждого прочтения (последовательности) запись в формате FASTQ состоит из четырех строк:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!'*(((((***+)))%% %% ++)(%% %% %).1***+*'))**55CCF>>>>>CCCCCCC65
```

Первая строка начинается со знака @, в ней записывается имя последовательности (иногда информация о ее длине). Вторая строка – это сама последовательность. Третья строка начинается со знака +, изначально в ней повторялось название последовательности, но сейчас она чаще всего не содержит ничего, кроме «+». В четвертой строке записана информация о качестве прочтения каждого нуклеотида [1].

Обычно для записи информации о качестве прочтения используют значение Phred, определяемое по формуле:

$$Q = -10 \log_{10} P,$$

где  $Q$  – значение Phred,  $P$  – вероятность ошибки.

Таким образом,  $Q = 10$  означает, что вероятность ошибки в данной позиции составляет 10%,  $Q = 20$  соответствует 1% и т. д.

Изначально такая оценка качества данных секвенирования была введена для автоматизации обработки данных секвенирования по методу Сенгера (см. разд. 1.1.2). Она основывалась на форме пиков и расстоянии между ними. Расчет вероятности ошибки производили на основании данных секвенирования известной последовательности [2, 3]. Пример такой хроматограммы приведен на рис. 1.5. Для каждого типа приборов калибровку проводит производитель по собственной технологии. В случае NGS используют иные параметры сигнала, однако общий принцип калибровки остается тем же.

Значение  $Q$  – число, но для экономии места удобно записывать каждое значение  $Q$  в виде символа. Чтобы файл был читаем для человека, были выбраны печатаемые символы кодировки ASCII. Для данных сенгеровского секвенирования



сборки и поиска вариаций. При необходимости можно воспользоваться доступными в Интернете скриптами<sup>1</sup> и перевести данные о качестве прочтений из специальных форматов в формат FASTQ.

## 4.2. СБОРКА ГЕНОМОВ *DE NOVO*

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 (см. гл. 2) 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 [4].

При сборке генома исследователь обычно имеет дело с большим количеством коротких последовательностей (прочтений), случайным образом соответствующих исходной ДНК, причем для возможности сборки из коротких перекрывающихся фрагментов более протяженных регионов (в идеале – целых хромосом), суммарная длина этих фрагментов должна в несколько раз превышать длину исследуемого генома. Чем больше превышение (покрытие), тем более длинные последовательности можно составить. Необходимую суммарную длину коротких прочтений, достаточную для получения определенного процента собранного генома (так называемую степень покрытия секвенирования или покрытие генома), можно оценить с помощью уравнения Ландера–Ватермана (Lander-Waterman model), которое базируется на предположении, что фрагменты случайно распределены по геному (т. е. подчиняются распределению Пуассона).

В соответствии с уравнением Ландера–Ватермана вероятность того, что нуклеотид не будет определен, равна:

$$P(X = 0) = e^{-c},$$

---

<sup>1</sup> Например, <https://github.com/WimS83/XSQConverter> или <http://seqanswers.com/wiki/Sff2fastq>

где  $c = NL_R/L_G$  – степень покрытия, здесь  $N$  – число прочтений,  $L_R$  – длина прочтения,  $L_G$  – длина генома.

Таблица 4.2

**Доля несеквенированных нуклеотидов  
в зависимости от степени покрытия генома человека**

Глубина секвенирования, покрытий	Доля пропущенных нуклеотидов, %	Число пропущенных нуклеотидов	Глубина секвенирования, покрытий	Доля пропущенных нуклеотидов, %	Число пропущенных нуклеотидов
1	36,79	1 млрд	16	$1,13 \cdot 10^{-5}$	338
2	13,53	400 млн	17	$4,14 \cdot 10^{-6}$	124
3	4,98	150 млн	18	$1,52 \cdot 10^{-6}$	45,7
4	1,83	50 млн	19	$5,60 \cdot 10^{-7}$	16,8
5	0,67	20 млн	20	$2,06 \cdot 10^{-7}$	6,18
6	0,25	7 млн	21	$7,58 \cdot 10^{-8}$	2,27
7	0,09	3 млн	22	$2,79 \cdot 10^{-8}$	0,84
8	0,03	1 млн	23	$1,03 \cdot 10^{-8}$	0,31
9	0,01	400 тыс.	24	$3,78 \cdot 10^{-9}$	0,11
10	$4,54 \cdot 10^{-3}$	150 тыс.	25	$1,39 \cdot 10^{-9}$	0,04
11	$1,67 \cdot 10^{-3}$	50 тыс.	26	$5,11 \cdot 10^{-10}$	0,02
12	$6,14 \cdot 10^{-4}$	20 тыс.	27	$1,88 \cdot 10^{-10}$	0,01
13	$2,26 \cdot 10^{-4}$	7 тыс.	28	$6,91 \cdot 10^{-11}$	$2,07 \cdot 10^{-3}$
14	$8,32 \cdot 10^{-5}$	3 тыс.	29	$2,54 \cdot 10^{-11}$	$7,63 \cdot 10^{-4}$
15	$3,06 \cdot 10^{-5}$	1 тыс.	30	$9,36 \cdot 10^{-12}$	$2,81 \cdot 10^{-4}$

В табл. 4.2 приведена доля пропущенных нуклеотидов и их число для генома человека (3 млрд п. н.) в зависимости от степени покрытия секвенирования ( $c$ ). Видно, что при  $c = 5$  будет определено 99% генома, однако около 20 млн нуклеотидов будут пропущены, при  $c = 22$  теоретически будет определена почти вся последовательность генома.

### 4.3. АЛГОРИТМЫ СБОРКИ

Целью сборки является получение минимального числа контигов максимальной длины.

В настоящее время для сборки протяженных регионов из сравнительно коротких последовательностей нуклеотидов (прочтений) используют два основных алгоритма [5].

Первый из алгоритмов был предложен в 1980 году [6] и получил название OLC (overlap-layout-consensus – перекрытие–расположение–согласованность). Это интуитивно понятный алгоритм, в котором сначала проводят попарное сравнение и выравнивание всех прочтений, после чего строится граф, где узлами являются прочтения, и если перекрытие между прочтениями оказывается не меньше, чем заданная величина  $T$ , узлам приписывается связь. Число узлов графа равно числу прочтений, и оно линейно растет с увеличением глубины секвенирования, число связей при этом возрастает по логарифмическому закону.

При использовании алгоритма OLC можно оценить число контигов (непрерывных последовательностей, собранных из перекрывающихся между собой отдельных прочтений) по формуле:

$$n = N \exp \left( -c \frac{L_R - T + 1}{L_R} \right),$$

где  $c = NL_R/L_G$  – глубина секвенирования,  $N$  – число прочтений,  $L_R$  – длина прочтения,  $L_G$  – длина генома,  $T$  – минимальное перекрытие,  $-c \frac{L_R - T + 1}{L_R}$  – вероятность для прочтения оказаться крайним правым в контиге [7].

Из формулы видно, что число контигов зависит от количества и длины прочтений, длины генома и минимального перекрытия, которое считается приемлемым. Если зафиксировать длину необходимого перекрытия прочтений и построить

график зависимости числа контигов от глубины секвенирования для разной длины прочтения (рис. 4.1), можно заметить, что для получения с помощью фрагментов длиной 50 п. н. такого же числа контигов, что и для 10-кратного покрытия фрагментами по 500 п. н., необходимо 30-кратное покрытие.

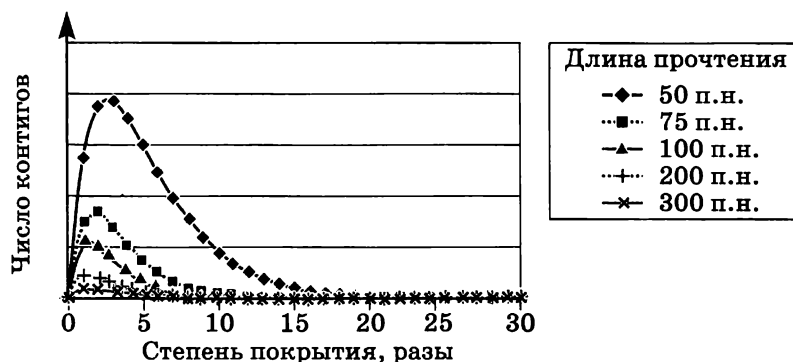


Рис 4.1. График зависимости числа контигов от глубины секвенирования

Этот алгоритм используется во многих широко распространенных программах, таких как: Arachne [8], Celera Assembler, CAP3 [9], Phrap [10], Phusion [11], Newbler и т. п.

Второй алгоритм основан на использовании графов де Брёйна (de-bruijn-graph) и на первый взгляд не так очевиден. Все прочтения разбивают на более короткие участки одинаковой длины  $k$ , которые являются узлами в графе, тогда как ребрами являются  $k + 1$ -меры. На рис. 4.2 представлен пример графа для  $k = 2$  (указаны последовательности, соответствующие ребрам графа).

В идеальном случае, когда каждый  $k$ -мер встречается в геноме только один раз, количество узлов графа будет примерно равно длине генома ( $L_G - k + 1$ ), число связей ( $L_G - k$ ) тоже почти равно длине генома. Однако на практике так не бывает, поскольку в геноме всегда присутствуют повторяющиеся последовательности, и кроме того – ошибки секвенирования, ведущие к появлению дополнительных связей.

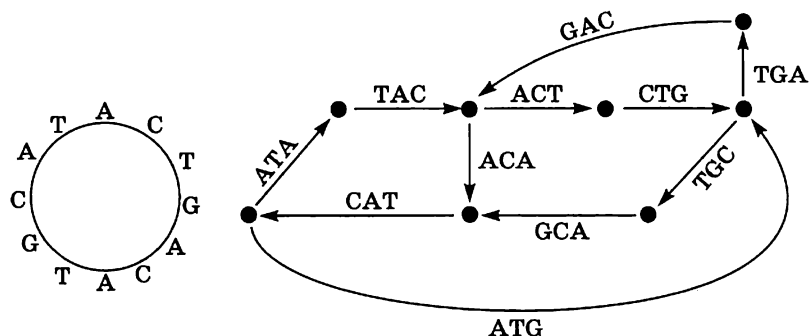


Рис. 4.2. Пример графа де Брёйна для  $k = 2$

Число контигов в данном алгоритме можно оценить по аналогии с числом контигов для модели Ландера–Ватермана. Считая, что минимальная величина перекрытия равна  $k$ , число контигов можно оценить по формуле:

$$n = N \exp(-d_k),$$

где  $d_k = N(L_R - K + 1)/L_G$ ,  $N$  – число прочтений,  $L_R$  – длина прочтения,  $L_G$  – длина генома,  $K$  – длина  $k$ -мера.

Этот алгоритм был предложен в 1995 году, а первый основанный на этом алгоритме программный продукт был представлен в 2001 году [12]. Вначале, при работе с данными секвенирования методом Сенгера, этот алгоритм был непопулярен, но с появлением технологий высокопроизводительного секвенирования он оказался крайне востребован и в настоящее время используется в программах Euler-USR [13], Velvet [14], ABySS [15], AllPath-LG [16] and SOAPdenovo [17]. Сначала алгоритм на основе графов де Брёйна плохо справлялся со сборкой больших геномов, и его применяли в основном для сборки геномов прокариот. Однако программное обеспечение и аппаратная база совершенствовались, и в настоящее время разработано несколько основанных на графах де Брёйна программных продуктов, позволяющих собирать большие геномы эукариот [18, 19].

Из-за меньших требований к объему оперативной памяти ЭВМ (в сравнении с программным обеспечением на основе алгоритма OLC) и отсутствия вычислительно затратного этапа

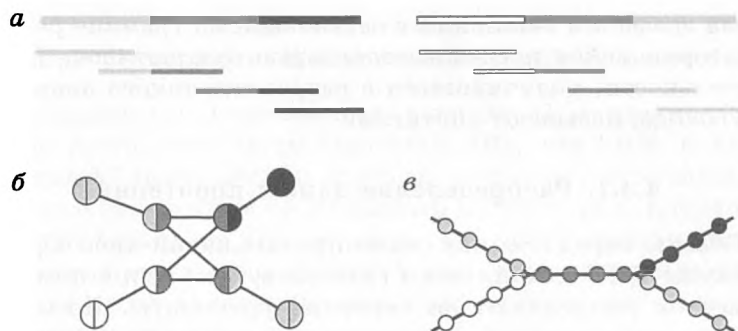
выравнивания программы, основанные на графах де Брёйна, часто оказываются даже более эффективны для сборки геномов из большого количества коротких прочтений.

#### 4.4. АППАРАТНЫЕ И БИОЛОГИЧЕСКИЕ ОСОБЕННОСТИ ДАННЫХ NGS

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. Выше уже было сказано, что, помимо 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.

Тем не менее даже после фильтрации часть ошибок секвенирования все равно остается. Для программ, работающих на базе алгоритма OLC, эта проблема решается сравнительно просто: можно «разрешить» алгоритму допускать 1–2 несовпадающих основания при выравнивании и определить наличие ошибок секвенирования на основании вероятностной модели, полагая, что правильные прочтения встречаются значительно чаще, чем неправильные. Для алгоритма де Брёйна из-за ошибок секвенирования возникают новые пути, усложняющие работу программы, но, поскольку ошибки секвенирования все-таки сравнительно редки, возможен их поиск и исправление на основании частоты встречаемости  $k$ -меров.

Кроме ошибок секвенирования существует еще одна распространенная проблема – повторяющиеся последовательности, присутствующие в любом геноме. Проблемой являются повторы, превышающие длину одного прочтения для используемой технологии NGS. На рис. 4.3 изображено представление повторов для разных алгоритмов сборки. Для алгоритма OLC прочтения, относящиеся к повторяющимся участкам генома, соответственно будут выравниваться с большим чис-



**Рис. 4.3.** Иллюстрация разницы между алгоритмами OLC и де Брёйна: *а* – два участка в геноме, имеющие одинаковую повторяющуюся последовательность нуклеотидов (средняя часть фрагментов); *б* – представление для OLC: из-за наличия повтора в графе появляются дополнительные узлы; *в* – представление для алгоритма де Брёйна. Узлы графа –  $k$ -меры.

Одинаковые  $k$ -меры сливаются в один узел.

Слева и справа от повтора возможны разные варианты сборки

лом прочтений, что требует существенного увеличения необходимого объема аппаратной памяти для хранения информации о связях. Поэтому для программного обеспечения на базе алгоритмов OLC повторяющиеся элементы обычно исключают из сборки. С увеличением длины прочтения и заданной величины перекрытия количество повторов, не поддающихся сборке, падает.

В случае использования графов де Брёйна повторяющиеся  $k$ -меры собираются («схлопываются») в один узел и вычислительная сложность не увеличивается. Однако при этом увеличение длины прочтения почти не дает дополнительных преимуществ, поскольку последовательности в любом случае разбиваются на  $k$ -меры, длина которых, как правило, не превышает 31 п. н. (иногда до 127 п. н.). С увеличением длины  $k$ -меров быстро возрастает потребность в вычислительных ресурсах, и при использовании более длинных  $k$ -меров становится сложнее обнаружить участки гетерозиготности и ошибки секвенирования, что снижает качество сборки.

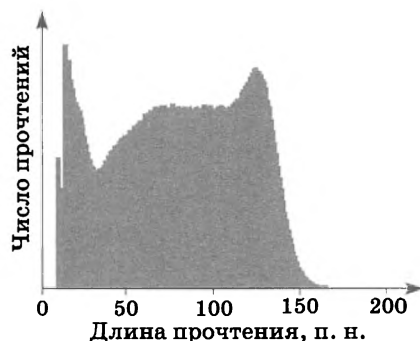
Таким образом, основная задача при использовании программного обеспечения для сборки длинных регионов из коротких прочтений заключается в поиске непрерывного пути

обхода графа без ветвлений с остановкой на границе региона с повторяющейся последовательностью нуклеотидов. Последовательности, получившиеся в результате такого непрерывного обхода, называют контигами.

#### 4.4.1. Распределение длины прочтений

Обычно перед тем, как секвенировать какой-либо образец, молекулы ДНК механически (ультразвуком) или с помощью ферментов расщепляют на короткие фрагменты. Поскольку разброс длин при этом получается достаточно большим, разделяя полученные фрагменты с помощью электрофореза, для дальнейшей работы, отбирают фракцию с оптимальным размером фрагментов (см. разд. 2.4). Насколько точно были отобраны фрагменты можно оценить по результатам секвенирования. На рис. 2.10, б приведен пример распределения длины прочтений для хорошо отобранного сегмента библиотеки. Видно, что большая часть прочтений приходится на достаточно узкий пик. При необходимости короткие прочтения можно убрать из дальнейшего анализа.

На рис. 4.4 представлен результат неудачного запуска прибора. Видно, что что-то пошло не так (либо выбор фрагментов для секвенирования был совсем плох, либо по каким-то причинам качество прочитанных последовательностей было низким и они были обрезаны программно на этапе обработки сигнала). Такой эксперимент необходимо повторить.



**Рис. 4.4.** Распределение длин прочтений в результате неудачного запуска прибора

#### 4.4.2. Учет искажений при секвенировании генома отдельных клеток

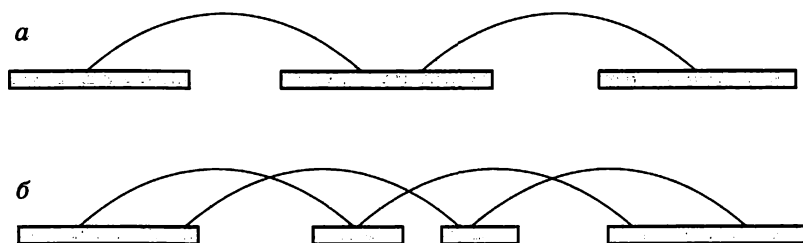
Большинство современных технологий секвенирования требует достаточно много стартовой ДНК или РНК. Если исследуемый биологический материал доступен и возобновляем, проблем с получением необходимого количества нуклеиновых кислот не возникает. Однако, если речь идет о секвенировании транскриптома отдельной клетки или генома некультивируемых или плохо культивируемых бактерий, стартовое количество нуклеиновых кислот может быть крайне мало. Для решения данной проблемы были разработаны методы амплификации ДНК и кДНК из малого количества стартового материала (см. гл. 2).

Обычно для некультивируемых бактерий секвенирование идет *de novo* и референсного генома не существует. Поэтому исследователи сталкиваются с проблемой сборки такого генома (хотя бы в достаточно протяженные контиги). Наличие этапа амплификации сильно усложняет эту задачу. Во-первых, покрытие амплифицированного генома оказывается очень неравномерным (вследствие искажений амплификации), во-вторых, при использовании инвертированных библиотек с двусторонним прочтением гораздо сложнее контролировать длину вставки. Кроме того, в процессе амплификации возникают ошибки и химерные прочтения (прочтения, состоящие из нескольких фрагментов, которые на самом деле находятся в разных местах в геноме). Программы для сборки, используемые для других задач, плохо справляются с такими данными. Специально для искаженных данных группой биоинформатиков из Санкт-Петербурга был разработан сборщик SPAdes, который учитывает особенности секвенирования отдельных клеток и справляется со сборкой значительно лучше [20, 21].

#### 4.5. ОБЪЕДИНЕНИЕ КОНТИГОВ В СКЭФФОЛДЫ

Поскольку длина одного прочтения (и, как следствие, максимальная длина региона с повторяющимся мотивом, который можно прочесть насквозь) ограничена технологически, для дальнейшего объединения контигов в скэффолды (системы правильно расположенных друг относительно друга, но разъединенных неизвестными регионами контигов) использу-

ют инвертированные библиотеки со вставками различных размеров (см. главу 2) и/или чтением фрагментов с двух сторон (парно-концевое прочтение библиотеки). Если из сцепленной пары прочтений одно прочтение выравнивается на один контиг, а прочтение с другой стороны фрагмента – на другой, то, зная примерное расстояние между прочтениями (по средней длине фрагментов приготовленной библиотеки), можно предположить, как расположены контиги друг относительно друга (рис. 4.5, *а*). Проблемы возникают с контигами, содержащими повторы или встречающимися в геноме более одного раза. Обычно такие контиги отличаются по величине покрытия и распознаются программами для сборки как повторяющиеся. Еще одна возможная сложность – чередование контигов (рис. 4.5, *б*). Такие контиги можно правильно расположить, используя для NGS библиотеку с меньшим размером фрагментов ДНК.



**Рис. 4.5.** Использование инвертированных библиотек для сборки контигов: *а* – вариант с уникальными последовательностями; *б* – вариант с повторяющимися последовательностями

Многие современные программы для сборки последовательностей, такие как Velvet [14], ABySS [15], SOAP [17], ALLPATHS-LG [16], позволяют учитывать информацию, полученную от парно-концевых прочтений и эффективно собирать контиги в скэффолды. При использовании программного обеспечения без этой функции можно воспользоваться отдельными программами для выстраивания контигов в скэффолды на основании данных парно-концевых прочтений, например: SSPACE [22], Bambus [23], GRASS [24] и др.

Помимо инвертированных библиотек, прочитанных с двух сторон, информацию о расположении контигов в геноме мож-

но 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.

Недавно был опубликован подход к сборке скэффолдов, основанный на данных о пространственном расположении участков ДНК в хроматине – метод Hi-C [25]. При использовании данного метода ДНК непосредственно в составе хроматина расщепляют с помощью эндонуклеаз рестрикции, дающих выступающий («липкий») 5'-конец. 5'-концы достраивают ДНК-полимеразой с использованием биотинилированного основания, после чего проводят лигирование в условиях, при которых вероятность реакции максимальна для близко расположенных концов расщепленной ДНК. В результате образец ДНК содержит продукты лигирования, помеченные биотином в месте стыка соседних фрагментов. Продукты лигирования отбирают на частицы со стрептавидином и секвенируют. Поскольку вероятность лигирования пропорциональна расстоянию между участками и значительно выше внутри хромосомы, чем между хромосомами, данный метод позволяет установить распределение контигов по хромосомам и их относительное расположение. Авторами показана принципиальная возможность использования этого метода на геномах человека, мыши и дрозофилы. Кроме того, было определено положение некоторых контигов в геноме человека, для которых оно до сих пор не было известно.

После сборки контигов в скэффолды между контигами остаются участки. Длина их приблизительно известна, а последовательность – нет. Наличие таких разрывов, как правило, обусловлено либо протяженной повторяющейся последовательностью (которую невозможно правильно собрать из коротких прочтений), либо недостаточным покрытием генома. В первом случае дополнительную информацию можно попробовать получить из прочтений, отфильтрованных перед сборкой из-за отнесения их к повторяющимся последовательностям. Во втором случае необходимо дополнительное секвенирование.

Несмотря на огромный поток экспериментов по секвенированию геномов и широкий спектр методических приемов, на сегодняшний день нет однозначного ответа на вопрос, ка-

кие комбинации технологий секвенирования и программ для сборки дают наилучшие результаты при секвенировании *de novo*. Нет и единых критериев оценки качества сборки.

Одним из способов оценки качества сборки является значение параметра N50, отвечающего такой длине контига, для которой более длинные контиги в сумме включают не менее 50% от общей длины всех контигов в проекте. Аналогичным образом определяется значение N90 (превышающие данный показатель контиги в сумме содержат 90% последовательностей). Однако информативность оценки качества сборки с помощью этой метрики вызывает сомнения.

Для выбора наилучшего программного обеспечения для сборки последовательностей некоторые авторы проводят их сопоставление с помощью различных тестовых баз данных [26, 27, 28]. Сравнения проводят разными способами, в том числе в виде соревнования (*assemblathon*), в котором разработчикам программ для сборки коротких прочтений предлагается посоревноваться в качестве получаемого результата [28].

#### 4.6. ВАРИАЦИИ В БЛИЗКОРОДСТВЕННЫХ ГЕНОМАХ

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. 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 методов высокопроизводительного секвенирования, как правило, в варианте повторного секвенирования генома.

#### 4.7. КАРТИРОВАНИЕ ПРОЧТЕНИЙ ПРИ ПОВТОРНОМ СЕКВЕНИРОВАНИИ

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 и т. п.) штаммов микроорганизма. Для эукариот – накопление данных о вариабельности генома внутри вида, ассоциативные исследования (в том числе поиск генетических детерминант определенных фенотипов).

Если перед исследователем стоит задача определения последовательности генома организма, данные для которого уже есть в базе данных (т. е. 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 (скажем, в 5 млн п. н.) для каждого фрагмента в ходе выравнивания (и если искать только места точного совпадения) 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 алгоритму позиционировать последовательность с учетом наличия единственного несовпадающего нуклеотида на 100 п. н., то мы получим  $3 \cdot 100 \cdot 5\,000\,000$  вариантов расположения фрагмента на референсном геноме, т. е. 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 общем виде задачу картирования прочтений на референсном геноме можно сформулировать следующим образом: дан набор полученных в эксперименте последовательностей  $Q$ , набор референсных последовательностей  $R$ , набор ограничений и пороговое значение дистанции  $k$ . Необходимо найти все подстроки  $m$  в  $R$ , удовлетворяющие всем ограничениям, в том числе и такому, что расстояние между  $m$  и любой последовательностью  $q$  из  $Q$  не превышает  $k$ , т. е.  $d(q, m) \leq k$ , где  $d(q, m)$  – функция расстояния. Ее обычно рассчитывают на основании количества замен, делеций или вставок. Также она может учитывать протяженность делеций или вставок и характерные ошибки использованной платформы секвенирования.

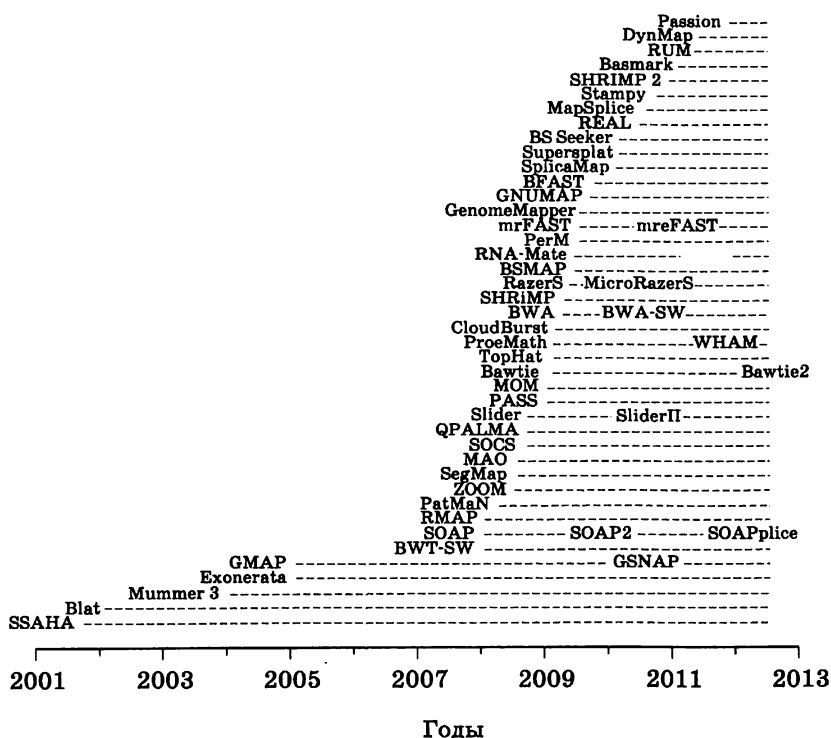
Основная цель картирования – найти правильное расположение в геноме каждой последовательности  $q$  из  $Q$  с учетом ошибок секвенирования и структурных вариантов. При этом важно отличать ошибки секвенирования от реальных различий в сравниваемых последовательностях.

Для картирования прочтений, полученных методом Сенгера, был разработан набор программ, успешно справляющихся с такими задачами: SSAHA [29], BLAST [30, 31], BLAT [32]. 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. Поэтому были разработаны новые алгоритмы, требующие меньше аппаратной мощности и времени на выполнение.

Число программ для картирования прочтений уже превысило полсотни и постоянно растет. На рис. 4.6 приведен список некоторых программ для картирования прочтений на референсный геном в порядке их появления. Видно, что после разработки методов высокопроизводительного секвенирования, количество программ для картирования начало быстро расти [33].

Большая часть алгоритмов выравнивания использует дополнительные структуры данных – индексы. Индексируются или референсный геном, или прочтения. В зависимости от используемой структуры данных можно выделить 3 группы алгоритмов:

- 1) основанные на хэш-таблицах: ZOOM [34], SHRiMP [35], RazerS [36];



**Рис. 4.6.** Список программ для картирования прочтений на референсный геном в порядке их появления. Видно, что после разработки методов секвенирования второго поколения количество программ для картирования начало быстро расти

- 2) основанные на суффиксных деревьях (suffix tree): MUMmer Bowtie [37], BWA [38], SOAP2 [39];
- 3) основанные на сортировке путем слияния (merge sort) Slider [40].

Наиболее популярными являются программы, использующие первые две группы алгоритмов.

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, однонуклеотидные замены и ухудшение качества данных к концу прочтений для Illumina) и особенности формата данных (цветовое кодирование для данных SOLiD).

При работе с данными секвенирования транскриптомов важным является алгоритм поиска мест сплайсинга. Он может производиться как на основании существующих баз данных (таких как X-mate [41] или же *de novo* на основе машинного обучения – например MapSplice [42] либо GMAP [43]). В первом случае, если в образце присутствуют ранее не описанные варианты сплайсинга, программа может их не найти [33].

#### 4.8. ПОИСК ОДНОНУКЛЕОТИДНОГО ПОЛИМОРФИЗМА (SNP)

Как было сказано выше, при картировании прочтений они не всегда точно совпадают с референсной последовательностью. Причиной таких несовпадений может быть реальное отличие референсной последовательности и исследуемой ДНК, ошибка секвенирования или неверное картирование прочтения.

Поскольку наиболее частым и интересным с биологической точки зрения различием в близкородственных геномах является однонуклеотидный полиморфизм (SNP), исследователи часто решают задачу его обнаружения в близкородственных геномах. Основной сложностью при поиске SNP является необходимость отличать реальные изменения последовательности от ошибок секвенирования и картирования. На первый взгляд ничего сложного тут нет: ошибки секвенирования происходят случайно, а следовательно, вероятность оказаться даже в двух независимых прочтениях в одном и том же месте очень мала (за исключением случая, когда ошибка возникает на этапе амплификации ДНК на первых циклах ПЦР). Вместе с тем, если речь идет о геномах диплоидных эукариот, для переменных позиций внутри полученных прочтений соотношение аллелей должно быть 50/50. Однако в случае секвенирования, например, ДНК опухолевой ткани в образце могут

встречаться клетки с соматическими мутациями, так что соотношение частот разных аллелей при этом может быть любым (и отфильтровывать такие вариации нельзя, поскольку именно они могут быть предметом поиска).

Для оценки того, является ли данный полиморфизм достоверным, учитывают, какая доля прочтений несет такой вариант и соответствует ли частота варианта ожиданиям исследователя.

Задача становится еще более сложной, если речь идет об исследованиях, при которых для экономии средств в одном образце смешивают ДНК нескольких организмов и полученную смесь секвенируют (например, подход часто используют при полногеномном поиске ассоциаций – genome-wide association study, GWAS). Поскольку значимыми в таких экспериментах часто оказываются именно редко встречающиеся варианты, необходимы подходы, отличающие их от ошибок секвенирования [44]. Для поиска таких вариантов используют методы, основанные на машинном обучении. Они позволяют оценивать достоверность обнаруженного полиморфизма на основании большого числа параметров. Для обучения используют два набора данных: соответствующие истинным полиморфизмам и ошибки секвенирования. На основании предварительного обучения определяют параметры, позволяющие достоверно отличить реальный полиморфизм от ложного, обусловленного технологическими ошибками ([45]).

Для поиска SNP обычно используют результаты выравнивания на референсный геном в виде файлов формата SAM. Для работы с этими файлами можно использовать программные пакеты SAMTools [46], GATK [47], Atlas [48] и ряд других. Наиболее популярным является GATK. При этом программы постоянно совершенствуются, появляются новые, проводятся исследования для выявления наилучших алгоритмов [49].

#### **4.9. ПОИСК СТРУКТУРНЫХ ВАРИАЦИЙ: ПРОТЯЖЕННЫХ ВСТАВОК, ДЕЛЕЦИЙ, ИНВЕРСИЙ И ТРАНСЛОКАЦИЙ**

Сразу отметим, что структурные изменения в геноме можно обнаружить не только с помощью NGS. Существуют альтернативные подходы: флуоресцентная гибридизация *in situ*

(FISH), сравнительная геномная гибридизация (CGH), супрессионная вычитающая гибридизация (SSH) и ряд других. FISH позволяет обнаружить изменения размером не менее 5–10 млн п. н., а CGH – изменения порядка 10–25 т. п. н. Несмотря на то что SSH позволяет найти изменения любого размера – места стыка для крупных вставок либо делеций или же обнаружить короткие делеции целиком, этот подход является довольно сложным в исполнении и неустойчив к минимальным отклонениям от протокола. В этой связи для поиска сравнительно небольших изменений в геноме исследователи все чаще прибегают к NGS.

Существует два основных способа поиска структурных изменений в геноме. Первый использует информацию о глубине и равномерности покрытия референсного генома прочтениями. В этом подходе подсчитывают число прочтений вдоль генома в «окне» определенной длины и устанавливают, являются ли различия с соседними участками статистически значимыми. Поскольку для многих технологий NGS частота прочтений зависит от GC-состава секвенируемого региона, перед анализом обычно проводят предварительную нормализацию данных, сглаживающую подобные отклонения, используя CNV-seq [50], SegSeq [51] и другие инструменты. Второй способ, помимо покрытия, учитывает информацию о выравнивании парно-концевых прочтений: если направление парного прочтения или расстояние до него отличается от ожидаемого, это может говорить о наличии структурных изменений. Инструменты, учитывающие информацию о парно-концевых прочтениях: CNVer [52], CNaseg [53], CNVnator [54] и др. В первом подходе возможно обнаружение только вставок и делеций, тогда как во втором – любые структурные вариации, включая инверсии и транслокации [55].

#### **4.10. АННОТАЦИЯ ОБНАРУЖЕННЫХ ВАРИАЦИЙ С ИСПОЛЬЗОВАНИЕМ БАЗ ДАННЫХ**

После того как вариации обнаружены, возникает вопрос об их роли в изучаемом биологическом процессе (заметим, что большинство вариаций является вариантами нормы). Первое, что необходимо сделать, – посмотреть, что же известно про эти изменения к текущему моменту. Существует множество баз данных, в которых собрана информация о значимости того

или иного изменения. Более всего таких баз существует для генома человека и среди них наиболее известной является dbSNP<sup>1</sup>. В ней собраны собственно SNP, короткие вставки и делеции. Большинство программ для поиска SNP сопоставляют полученные результаты с имеющимися в этой базе данными. В зависимости от того, какие именно данные интересуют исследователя, он может выбрать соответствующую базу данных. Так, информация о клинически значимых вариациях генома человека собрана в базе данных OMIM<sup>2</sup>, а информация о вариантах нормы аккумулирована в базе данных проекта HapMap<sup>3</sup>. Для описания встречающихся в норме крупных вставок и делеций была создана база данных DGV<sup>4</sup>. На том же ресурсе, что и dbSNP, сформирована dbVAR<sup>5</sup>, включающая более крупные изменения, чем SNP. Для описания крупных клинически значимых вставок и делеций создана база данных ISCA<sup>6</sup>, но она не является открытой. Информация во многих открытых базах данных дублируется. Как правило, между ними есть некоторая синхронизация, но она не всегда актуальна. Кроме того, существуют базы данных, посвященные отдельным заболеваниям человека и другим организмам. Многие базы данных предоставляют интерфейс программирования приложений API (application programming interface), что позволяет осуществлять поиск в них с помощью скриптов.

#### **4.11. ПРЕДСКАЗАНИЕ ФУНКЦИОНАЛЬНЫХ И КЛИНИЧЕСКИ ЗНАЧИМЫХ ИЗМЕНЕНИЙ БЕЛКА НА ОСНОВЕ ОБНАРУЖЕННЫХ МУТАЦИЙ**

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.

<sup>1</sup> <https://www.ncbi.nlm.nih.gov/SNP/>

<sup>2</sup> <http://www.omim.org/>

<sup>3</sup> <http://hapmap.ncbi.nlm.nih.gov/>

<sup>4</sup> <http://dgv.tcag.ca/dgv/app/home>

<sup>5</sup> <http://www.ncbi.nlm.nih.gov/dbvar/>

<sup>6</sup> <http://www.iscaconsortium.org/>

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 переменная 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. [56].

### СПИСОК ЛИТЕРАТУРЫ

1. *Cock P.J. et al.* The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants // *Nucleic acids research*, 2010, 38 (6): 1767–1771.
2. *Ewing B., Hillier L., Wendl M. C., Green P.* Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment // *Genome Research*, 1998, 8 (3): 175–185.
3. *Ewing B., Green P.* Base-calling of automated sequencer traces using phred. II. Error probabilities // *Genome research*, 1998, 8 (3): 186–194.
4. *Compeau P.E.C., Pevzner P.A.* How to apply de Bruijn graphs to genome assembly // *Nat Biotechnol*, 2011, Nov 8; 29(11): 987–991.
5. *Li Z. et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph // *Briefings in functional genomics*, 2012, 11 (1): 25–37.
6. *Staden R.* A new computer method for the storage and manipulation of DNA gel reading data // *Nucleic Acids Res.*, 1980, 8 (16): 3673–3694.
7. *Lander E.S., Waterman M.S.* Genomic mapping by fingerprinting random clones: a mathematical analysis // *Genomics*, 1988, 2 (3): 231–239.
8. *Batzoglou S. et al.* ARACHNE : A Whole-Genome Shotgun Assembler // *Genome Research*, 2002, 12: 177–189.
9. *Huang X.* CAP3: A DNA Sequence Assembly Program // *Genome Research*, 1999, 9 (9): 868–877.
10. *Gordon D., Green P.* Consed: a graphical editor for next-generation sequencing // *Bioinformatics*, 2013, Nov 15; 29 (22): 2936–2937.
11. *Mullikin J.C., Ning Z.* The Phusion Assembler // *Genome Research*, 2003, 13 (1): 81–90.

12. *Pevzner P.A., Tang H., Waterman M.S.* An Eulerian path approach to DNA fragment assembly // *Proceedings of the National Academy of Sciences of the United States of America*, 2001, 98 (17): 9748–9753.
13. *Chaisson M. J., Pevzner P.A.* Short read fragment assembly of bacterial genomes // *Genome research*, 2008, 18 (2): 324–330.
14. *Zerbino D.R., Birney E.* Velvet: algorithms for de novo short read assembly using de Bruijn graphs // *Genome Research*, 2008, 18 (5): 821–829.
15. *Simpson J.T. et al.* ABySS: A parallel assembler for short read sequence data // *Genome Research*, 2009, 19: 1117–1123.
16. *Gnerre S. et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data // *Proceedings of the National Academy of Sciences of the United States of America*, 2011, 108 (4): 1513–1518.
17. *Li R. et al.* De novo assembly of human genomes with massively parallel short read sequencing // *Genome research*, 2010, 20 (2): 265–272.
18. *Li R. et al.* The sequence and de novo assembly of the giant panda genome // *Nature*, 2010, 463 (7279): 311–317.
19. *Huang S. et al.* The genome of the cucumber, *Cucumis sativus* L. // *Nature genetics*, 2009, 41 (12): 1275–1281.
20. *Bankevich A. et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // *J Comput Biol.*, 2012, 19(5): 455–77.
21. *Nurk S. et al.* Assembling single-cell genomes and mini-metagenomes from chimeric MDA products // *J Comput Biol.*, 2013, 20(10): 714–37.
22. *Boetzer M. et al.* Scaffolding pre-assembled contigs using SSPACE // *Bioinformatics*, 2011, 27 (4): 578–9.
23. *Pop M., Kosack D.S., Salzberg S.L.* Hierarchical Scaffolding With Bambus // *Genome Research*, 14 (1): 149–159.
24. *Gritsenko A.A., Nijkamp J.F., Reinders M.J.T., Ridder D.De.* GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies // *Journal of Southern African Studies*, 2012, 1–9.
25. *Korbel J.O., Lee C.* Genome assembly and haplotyping with Hi-C // *Nature biotechnology*, 2013, 31 (12): 1099–1101.
26. *Salzberg S.L. et al.* GAGE : A critical evaluation of genome assemblies and assembly algorithms // *Genome Research*, 2012, 22 (3): 557–567.

27. *Earl D. et al.* Assemblathon 1: A competitive assessment of de novo short read assembly methods // *Genome Research*, 2011, 21: 2224–2241.
28. *Bradnam K.R. et al.* Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species // *GigaScience*, 2013, 2 (1): 10.
29. *Ning Z., Cox A.J., Mullikin J.C.* SSAHA: A Fast Search Method for Large DNA Databases // *Genome Research*, 11: 1725–1729.
30. *Altschul S., Gish W., Miller W.* Basic local alignment search tool // *J. Mol. Biol.*, 1990, 215: 403–410.
31. *Camacho C. et al.* BLAST+: architecture and applications // *BMC bioinformatics*, 2009, 10: 421.
32. *Kent W.J.* BLAT – The BLAST-Like Alignment Tool // *Genome Research*, 2002, 12: 656–664.
33. *Fonseca N.A., Rung J., Brazma A., Marioni J.C.* Tools for mapping high-throughput sequencing data // *Bioinformatics*, 2012, 28 (24): 3169–77.
34. *Lin H. et al.* ZOOM! Zillions of oligos mapped // *Bioinformatics*, 2009, 24 (21): 2431–2437.
35. *Rumble S.M. et al.* SHRiMP: accurate mapping of short color-space reads // *PLoS computational biology*, 2009, 5 (5): e1000386.
36. *Weese D., Holtgrewe M., Reinert K.* RazerS 3: faster, fully sensitive read mapping // *Bioinformatics*, 2012, 28 (20): 2592–2599.
37. *Langmead B., Salzberg S.L.* Fast gapped-read alignment with Bowtie 2 // *Nature methods*, 2012, 9 (4): 357–9.
38. *Li H., Durbin R.* Fast and accurate short read alignment with Burrows-Wheeler transform // *Bioinformatics*, 2009, 25 (14): 1754–1760.
39. *Li R. et al.* SOAP2: an improved ultrafast tool for short read alignment // *Bioinformatics*, 2009, 25 (15): 1966–1967.
40. *Malhis N., Butterfield Y.S.N., Ester M., Jones S.J.M.* Slider-maximum use of probability information for alignment of short sequence reads and SNP detection // *Bioinformatics*, 2009, 25 (1): 6–13.
41. *Wood D.L.A. et al.* X-MATE: a flexible system for mapping short read data // *Bioinformatics*, 27 (4): 580–581.
42. *Wang K. et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery // *Nucleic acids research*, 2010, 38 (18): e178.
43. *Wu T.D., Watanabe C.K.* GMAP: a genomic mapping and alignment program for mRNA and EST sequences // *Bioinformatics*, 2005, 21 (9): 1859–1875.

44. *Cirulli E.T., Goldstein D.B.* Uncovering the roles of rare variants in common disease through whole-genome sequencing // Nature reviews. Genetics, 2010, 11 (6): 415–25.
45. *Fang Y.-H., Chiu, Y.-F.* A novel support vector machine-based approach for rare variant detection // PLoS one, 8 (8): e71114.
46. *Li H. et al.* The Sequence Alignment/Map format and SAMtools // Bioinformatics, 2009, 25 (16): 2078–2079.
47. *McKenna A. et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data // Genome Research, 2010, 20: 1297–1303.
48. *Evani U. S. et al.* Atlas2 Cloud: a framework for personal genome analysis in the cloud // BMC genomics, 2012, 13 (6): S19.
49. *Liu X. et al.* Variant callers for next-generation sequencing data: a comparison study // PLoS one, 2013, 8 (9): e75619.
50. *Xie C., Tammi M.T.* CNV-seq, a new method to detect copy number variation using high-throughput sequencing // BMC bioinformatics, 10: 80.
51. *Chiang D.Y. et al.* (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing // Nat Methods, 2009, 6 (1): 99–103.
52. *Medvedev P. et al.* Detecting copy number variation with mated short reads // Genome Research, 20: 1613–1622.
53. *Ivakhno S. et al.* CNAsseg – a novel framework for identification of copy number changes in cancer from second-generation sequencing data // Bioinformatics, 2010, 26 (24): 3051–3058.
54. *Abyzov A., Urban A., Snyder M., Gerstein M.* CNVnator : An approach to discover genotype and characterize typical and atypical CNVs from family and population genome sequencing // Genome Research, 2011, 21 (6): 974–984.
55. *Duan J., Zhang J.-G., Deng H.-W., Wang Y.-P.* Comparative studies of copy number variation detection methods for next-generation sequencing technologies // PLoS one, 2013, 8 (3): e59128.
56. *McLaren W. et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor // Bioinformatics, 2010, 26 (16): 2069–2070.

# ОБОРУДОВАНИЕ И ПРОГРАММНЫЕ РЕШЕНИЯ ДЛЯ ОБРАБОТКИ ДАННЫХ NGS

Обработка данных высокопроизводительного секвенирования требует существенных аппаратных ресурсов. Обычный персональный компьютер (ПК) подойдет разве что для сборки небольших бактериальных геномов. Работа со сложными проектами (метагеномные исследования, транскриптомы, геномы эукариот) требует специального вычислительного оборудования. Данная глава посвящена описанию вариантов стандартных аппаратно-программных решений для получения необходимых (оптимальных) мощностей при работе с данными NGS.

Можно выделить два принципиально разных аппаратно-программных решения для обработки данных NGS: локальное и сетевое (облачное). Разница состоит в том, что локальный вычислительный центр (если это не обычный ПК), как правило, обеспечивает большую вычислительную мощность и требует при этом больших капиталовложений. Сетевые решения дешевы, но обычно менее производительны и надежны.

## 5.1. ЛОКАЛЬНЫЕ ЦЕНТРЫ ОБРАБОТКИ ДАННЫХ NGS: АРХИТЕКТУРА И ПРОГРАММНЫЕ РЕШЕНИЯ

Типовой набор задач, решаемых секвенсным центром, примерно следующий: прием данных с секвенаторов, их первичная обработка, сборка *de novo* геномов или транскриптомов, картирование, аннотирование, выравнивание разных типов. К ним также относятся различные задачи вычислительной филогенетики, сравнительной геномики, популяционной генетики и др.

Для разных научно-исследовательских и инженерных целей вычислительные центры строят с начала 50-х годов XX века. Стандартные вычислительные задачи, решаемые, скажем, в области физики высоких энергий, гидродинами-

ки, молекулярной динамики, требуют от центра высокой вычислительной мощности при сравнительно небольшом объеме хранения (обычно это вычислитель с большим числом ядер и множеством серверов, связанных очень быстрой сетью, у каждого сервера сравнительно небольшая оперативная память).

Однако для хранения данных NGS необходимы довольно большие хранилища информации (например, один запуск секвенатора SOLiD 5500xl дает около терабайта данных). Облачные решения здесь плохо подходят по причине сложности пересылки таких объемов данных по обычным каналам связи. Таким образом «классическая» схема организации вычислительного центра (рис. 5.1) для обработки данных высокопроизводительного секвенирования не подходит. В случае NGS стоит задача обработки большого объема данных при сравнительно простых действиях с этими данными.

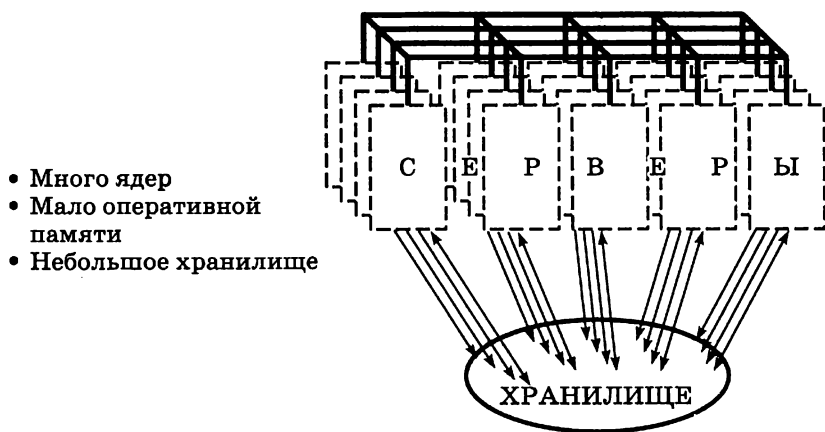
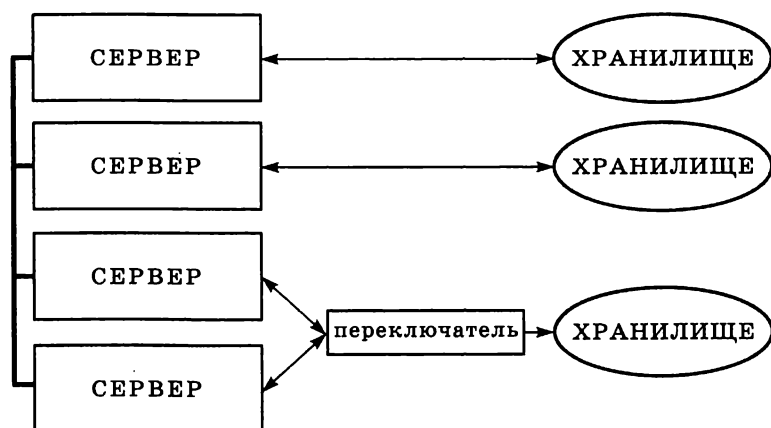


Рис. 5.1. «Классическая» схема организации вычислительного центра

Среди современных задач по структуре данных и операциями с ними высокопроизводительное секвенирование больше всего похоже на компьютерные системы компаний мобильной связи (для которых уже давно существуют типовые решения). Это некоторое количество вычислительных серверов и сопоставимое число хранилищ данных (рис. 5.2).



**Рис. 5.2.** Оптимальная организация вычислительного центра для обработки данных NGS

Для сборки эукариотических геномов нужен сервер с большой оперативной памятью – около терабайта. Собрать такой сервер можно далеко не на всякой платформе. Обычно для данной цели применяется платформа HiEnd самого высокого класса (наподобие устанавливаемых на атомных станциях), хотя для обработки сиквенсных данных столь высокий уровень надежности не требуется, нужен лишь большой объем оперативной памяти, а на надежности можно существенно сэкономить.

Производители оборудования для NGS (Illumina, Life Technologies) уже поставляют со своими приборами мини-кластеры, в которых можно проанализировать кое-что из NGS-данных (например, BioScope Cluster, LifeScope Cluster и LifeScope Workstation от Life Technologies Thermo Fisher Scientific).

На рис. 5.3 показан конкретный пример организации центра обработки данных NGS. Центр включает три системы хранения по 144 Тбайта и отдельные системы хранения (а не просто диски, вставленные в сервера). В этом классе решений системы хранения необходимо выносить отдельно по многим причинам. В представленном на рисунке варианте использован один узел с 512 Гбайтами памяти для сборки геномов и других мощных проектов и 30 вычислительных узлов с 48 Гбайтами памяти. Хранилище можно организовать по технологии Fiber

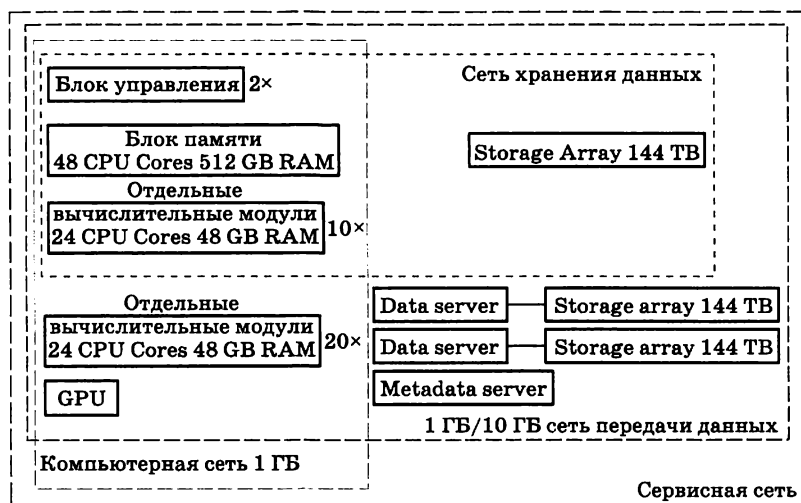


Рис. 5.3. Пример организации центра обработки данных NGS

Channel, при которой можно динамически подключать диски к каждому серверу. Есть часть для развертывания на ней распределенной файловой системы (например, Lustre, используемой в большинстве суперкомпьютеров и позволяющей равномерно распределить нагрузку на дисковые подсистемы).

При организации собственного центра обработки данных NGS необходимо уже на старте иметь хорошо проработанный технический проект, учитывающий наличие серверов с большой памятью, специализированных систем хранения (а не серверов с дисками), расчетных серверов, использование Fiber Channel для гибкого подключения дисков, специализированных систем вентиляции и электроснабжения.

Для установки программного обеспечения, модификации и администрирования системы желательно иметь своих ИТ-специалистов в команде. Чтобы получить максимальный эффект от работы столь «гетерогенного» коллектива, желательно расположить рабочие места ИТ-специалистов, биоинформатиков и биологов близко друг к другу для их тесного взаимодействия.

Отдельное внимание необходимо уделить обеспечению бесперебойного электроснабжения. Если, например, ваш институт относится ко второму классу энергопотребления,

диспетчер местной локальной сети может вас отключить от сети на пять минут без предупреждения, поэтому при организации вычислительного центра (как, впрочем, и при организации сиквенсного центра) необходимо закупить мощные источники бесперебойного питания.

Важна и система охлаждения помещений и оборудования. Аппаратура генерирует большое количество тепла, которое летом необходимо эффективно удалять из помещения (то же относится и к организации сиквенсного центра).

## 5.2. ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ ДЛЯ ЛОКАЛЬНОГО ЦЕНТРА ОБРАБОТКИ ДАННЫХ NGS

Чаще всего для работы с данными NGS используют операционную систему Linux (в частности, Scientific Linux) и очередь задач Torque, для мониторинга – алгоритмы Nagios и Puppet. Кроме того, существует множество биоинформатических пакетов для конкретных биологических задач (см. гл. 4).

Ниже перечислены некоторые популярные биоинформатические пакеты для решения биологических задач разного типа.

Abyss	Clustal	Pal2nal
AdapterRemoval	Cuda	Paml
Agalma	GATK	PHYLYP
Annovar	Geneid	Platanus
Bambus	HaploMerger	Python
Bamtools	Jellyfish	RAXML
Beagle-Lib	Libsequence	RepeatMasker
BioPerl	MCSanX	SHRiMP
Biopython	Megan	Soapdenovo
Blast	Mrbayes	STAR
Bowtie	MUMmer	Statistics-Descriptive
Cegma	OrthoMCL	Velvet

### 5.3. СЕТЕВЫЕ СЕРВИСЫ И ПРОСТЫЕ РЕШЕНИЯ ДЛЯ ОБРАБОТКИ ДАННЫХ NGS

Доступ к мощному центру обработки данных есть не всегда. Обычная лаборатория, как правило, не имеет собственного отдельного биоинформатического центра и программистов. Для биологического эксперимента исследователю необходимо выбрать технологию секвенирования и на стадии получения данных обратиться в какой-то геномный центр для их обработки.

Однако в случае не слишком сложных задач (если не требуется сборка эукариотического генома *de novo*) можно воспользоваться обычным ПК или сетевыми (облачными) решениями, предлагаемыми рядом компаний.

#### 5.3.1. Обычный персональный компьютер

Поскольку кластер может себе позволить далеко не каждая лаборатория, одним из простейших решений является запуск на персональном компьютере имеющейся в свободном доступе программы (такой как SOAP Stack, Bowtie или BLAST). Минусом такого подхода является низкая скорость вычислений. Даже если анализировать экзом человека, для обработки данных потребуется от 20 до 50 ч работы компьютера. К тому же исследователю потребуются навыки программиста, поскольку большинство свободных программ не обладают графическим интерфейсом, а чтобы правильно фрагментировать данные, нужно скомпилировать исходный код на C++ или Perl.

Другое решение – проприетарные (платные) программы типа CLC Bio. У них есть графический интерфейс (что упрощает использование), но они довольно дороги, а скорость их работы не намного выше. При этом интерфейс все-таки достаточно сложный и один программный продукт, как правило, все равно не решает все три стандартных задачи анализа данных NGS: выравнивание, поиск вариаций и их аннотацию. Приходится пользоваться разными продуктами.

#### 5.3.2. Многоядерный кластер

Как же быстро и малозатратно решить все три задачи анализа данных NGS? Можно купить одну высокопроизводительную машину, в которой 64 или 128 ядер и порядка 16 Гбайт памяти (такой объем памяти позволяет разместить всю ре-

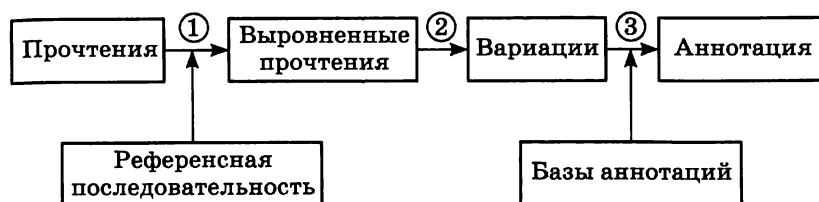
ференсную последовательность). Данное решение позволяет использовать большинство свободных программ, и оно достаточно дешево – подобный сервер стоит порядка 200 тыс. руб. Недостатками решения остаются сравнительно низкая скорость (лишь в 64 раза выше, чем на одном ПК) и немасштабируемость (в случае увеличения нагрузки). Для сравнения скажем, что максимальной нагрузкой на такой сервер можно считать несколько эукариотических геномов (в варианте повторного секвенирования) в сутки.

Как было сказано выше, производители оборудования для NGS (Illumina, Life Technologies) поставляют со своими секвенаторами такие мини-кластеры. Например, Life Technologies Thermo Fisher Scientific предлагает кластеры и рабочие станции LifeScore (кластер с 48-ядерным процессором, 96 Гбайт RAM и около 10 Тбайт обойдется примерно в 300 тыс. руб.).

### 5.3.3. Распределенные вычисления в облаке

Другим типовым решением многих современных вычислительных задач является использование так называемых распределенных вычислений. Такое решение базируется на использовании множества обычных ПК, как правило, с 2–16 ядрами и 4–16 Гбайтами оперативной памяти. На всех машинах параллельно запускаются вычисления, причем, если алгоритм удалось преобразовать для параллельного запуска (распараллелить), можно вводить в бой сколько угодно ПК. Этот подход дешев и легко масштабируем. Однако технически это довольно сложное решение: сложно разворачивать облачный кластер, сложно его настраивать, сложно распараллелить алгоритмы.

Несколько слов о том, что такое облачное хранение. Оно складывается из трех частей. Первое – это хранилище, которое должно быть надежным, достаточно дешевым и быстрым. Если хранилище медленное, а вы пробуете читать из него сотни машин, хранилище не справится. Второе – это система обработки данных. Две самые распространенные на сегодняшний день парадигмы – MapReduce и MPI (рис. 5.4). Третья – это сервис, позволяющий получать дополнительные машины. Желательна система распределенных вычислений, способная изменять задействованное число машин в зависимости от требуемых ресурсов.



- ① Вывравнивание прочтений — map-фаза
- ② Поиск вариаций — reduce-фаза
- ③ Аннотация вариаций — map-фаза

Рис. 5.4. Схема алгоритма, запускаемого на MapReduce

### Amazon Cloud Stack

Одно из наиболее популярных облачных решений – Amazon Cloud Stack. В качестве хранилища здесь выступает Amazon S3. Надежное решение, но скорость закачки данных невысока и процесс занимает довольно много времени. Для сокращения времени загрузки данных в облако можно закачивать данные, пока они обрабатываются на секвенаторе. Некоторые модели секвенаторов складывают данные на диск по мере их получения. В этом случае можно сразу же закачивать их в облако, тогда к моменту окончания работы секвенатора все данные будут в облаке. В качестве вычислительной системы в Amazon Cloud Stack используется Elastic MapReduce – некоторая надстройка, позволяющая автоматически изменять количество задействованных машин. Amazon EC2 – это решение по разворачиванию машин. Недостатком Amazon Cloud Stack является почасовая гранулярность оплаты. Существует открытое MapReduce-решение – Hadoop.

### Microsoft Azure

Другой распространенный вариант облачного решения – Microsoft Azure. Оно также может использовать Hadoop, обладает надежным хранилищем. Минусом, как и в предыдущем примере, является почасовая гранулярность оплаты сервиса.

#### 5.4. СПЕЦИАЛИЗИРОВАННЫЕ ПРОЕКТЫ ПО ОБРАБОТКЕ ДАННЫХ NGS

В Интернете можно обнаружить множество специализированных проектов для решения конкретной задачи по обработке данных NGS (см. разд. 5.2). Такие проекты позволяют картировать на референсные геномы, сравнивать геномы по базовым аннотациям, искать полиморфизм и т. п. [1–3]. Есть ряд узкоспециализированных решений, например поиск дифференциально представленных генов, аннотация внутриклеточных сигнальных каскадов и т. п. [4–6].

При решении подобных задач оптимальной является система с модульным принципом в подходе к организации и хранению информации. Проект предлагает лишь базовый уровень организации сервиса (аналог Nessus PHP Interface), в котором пользователь может запускать собственные приложения или легко модифицировать уже имеющиеся под конкретные проекты. У пользователя должна быть возможность выбора из широкого спектра программ картирования, аннотации, инструментов анализа данных.

Одной из первых подобных платформ стал Galaxy Project, запущенный в 2008 году группой ученых из Пенсильванского университета в США. Этот проект полностью отражает идеологию модульного построения систем. У него неплохие внутренние интерфейсы, и в скором времени с его помощью можно будет анализировать NGS-данные. Плюсы проекта – он бесплатный и в нем есть возможность масштабировать на облако.

Другой пример – платформа Postparser.net. С его помощью к настоящему моменту уже реализовано несколько NGS-проектов: поиск новых сайтов метилирования в геноме, новых профилей некодирующих РНК (например, микроРНК при раке), сравнение множества близкородственных геномов эукариот, поиск полиморфных вставок в геном, анализ регуляторных участков мобильных генетических элементов человека и т. д. Одним из преимуществ подобных платформ является приближение биолога к пониманию алгоритма анализа данных NGS. Зачастую биоинформатики приносят биологу готовые картинки для статей, и биолог понятия не имеет, как были сделаны вычисления. В сетевых сервисах биолог получает прямой доступ к данным, может их самостоятельно редактировать, комментировать, отбирать.

В качестве специализированных сетевых программных решений следует отметить такие, как CLC Bio<sup>1</sup> или Knome<sup>2</sup>.

### СПИСОК ЛИТЕРАТУРЫ

1. *Brouwer R.W., van den Hout M.C., Grosveld F.G., van Ijcken W.F.* NARWHAL, a primary analysis pipeline for NGS data // *Bioinformatics*, 2012, 28 (2): 284–285.
2. *Altmann A. et al.* A beginners guide to SNP calling from high-throughput DNA-sequencing data // *Hum Genet.*, 2012, 131 (10): 1541–1554.
3. *D'Antonio M. et al.* WEP: a high-performance analysis pipeline for whole-exome data // *BMC Bioinformatics*, 2013, 14 (7): S11.
4. *Marian A.J.* Molecular genetic studies of complex phenotypes // *Transl Res.*, 2012, 159 (2): 64–79.
5. *Coutant S. et al.* EVA: Exome Variation Analyzer, an efficient and versatile tool for filtering strategies in medical genomics // *BMC Bioinformatics*, 2012, 13 (14): S9.
6. *Pope B.J. et al.* FAVR (Filtering and Annotation of Variants that are Rare): methods to facilitate the analysis of rare germline genetic variants from massively parallel sequencing datasets // *BMC Bioinformatics*, 2013, 14: 65.

---

<sup>1</sup> <http://www.clcbio.com/>

<sup>2</sup> <http://www.knome.com/>

# ПЛАНИРОВАНИЕ ЭКСПЕРИМЕНТА С ИСПОЛЬЗОВАНИЕМ NGS

Любой эксперимент требует тщательной предварительной подготовки. В биологии планирование экспериментальной части исследования имеет большое значение по причине широкой вариабельности 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. 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 NGS – от сбора биологических образцов до обработки fastq-файлов. Планируя эксперимент, всегда помните, что говорит закон Мерфи об NGS: все, что может пойти не так в вашем сиквенсном проекте, уже пошло не так.

## 6.1. ОБЩИЕ ПРИНЦИПЫ ПЛАНИРОВАНИЯ БИОЛОГИЧЕСКИХ ЭКСПЕРИМЕНТОВ

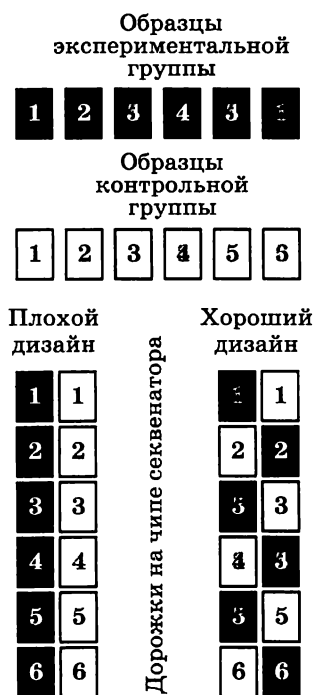
Планирование эксперимента – это процедура выбора количества и условий проведения опытов, необходимых и достаточных для решения поставленной задачи с требуемой точностью. Цель планирования эксперимента заключается в создании схемы, позволяющей получить наибольший объем информации при наименьших затратах. Среди классических характеристик экспериментального плана выделяют: сравнение (как правило, в научном эксперименте объект сравнивается либо с неким заранее заданным стандартом, либо с контрольным объектом), рандомизацию, повторяемость (или репликацию, не путать с удвоением ДНК), однородность и блокировку (страти-

фикацию). Рассмотрим чуть подробнее некоторые из перечисленных характеристик применительно к NGS.

## 6.2. РАНДОМИЗАЦИЯ В NGS

Рандомизация представляет собой процесс, используемый для группировки объектов таким образом, чтобы у каждого из них была равная вероятность попасть в контрольную или опытную группу. Другими словами, выбор участников исследования должен происходить случайно, чтобы исследование не было отклонено в сторону «предпочтительного» для исследователя результата. Рандомизация необходима для применимости статистических методов для анализа данных исследования, она помогает предотвратить смещения, обусловленные причинами, которые не были непосредственно учтены в плане эксперимента. Для этого, например, формирование экспериментальных групп лабораторных животных производится случайным образом.

Рассмотрим важность рандомизации на конкретном примере. Допустим, вы заинтересованы в поиске различий между двумя образцами ДНК. Для повышения надежности были собраны по шесть повторностей для каждого типа образцов (шесть контрольных и шесть экспериментальных). Предположим, вы секвенировали образцы на платформе Illumina HiSeq/GAIIx, используя две дорожки, причем каждая дорожка содержит шесть (мультиплексированных) повторностей. На рис. 6.1 показаны хороший и плохой варианты мультиплексирования. Неверным является подход с нанесением одинаковых образцов в одну дорожку секвенсного чипа. При планировании такого экспери-



**Рис. 6.1.** Сравнение хорошего и плохого экспериментальных дизайнов для параметра «рандомизация»

мента необходимо помнить пословицу «Не клади все яйца в одну корзину». Нанесение обоих типов образцов вперемешку позволяет исключить систематическую ошибку (в данном случае – эффект дорожки), а также получить хотя бы половину данных в случае неудачного прочтения одной из дорожек.

Конкретный пример: поиск дифференциально представленных транскриптов методом секвенирования транскриптомов контрольной и экспериментальной групп образцов. Если одна из дорожек по какой-либо причине даст меньше прочтений, при плохом дизайне можно принять аппаратно возникшую разницу за биологический сигнал. Гораздо надежнее в приведенном примере нанести в каждую из дорожек равное количество образцов из контрольной и экспериментальной групп. Выбор образцов из каждой группы для нанесения в одну дорожку лучше всего производить случайным образом [1, 2].

Другой пример возникновения систематической ошибки – влияние «штрих-кодирования» образцов (см. разд. 2.8). Дело в том, что присоединение адаптеров также влияет на качество секвенирования и разные по последовательности адаптеры влияют по-разному. При планировании эксперимента необходимо принимать в расчет влияние адаптеров и строить экспериментальную схему в направлении минимизации систематической ошибки.

### 6.3. ПОВТОРНОСТИ В NGS

Для выявления источника вариабельности необходимо провести несколько испытаний – повторов. Использование повторностей в NGS не очень распространено по двум причинам: из-за высокой стоимости исследования и того факта, что один из типов экспериментальных повторностей уже встроен в сам алгоритм секвенирования – это число прочтений каждого участка последовательности (так называемая степень покрытия). Однако другие типы повторностей, такие как биологические повторности, инструментальные повторности и повторности для разных технологий NGS могут быть крайне информативны и важны для корректной постановки эксперимента [3].

Биологические повторности (когда собирают по несколько однотипных биологических образцов) являются важной

составляющей любого биологического эксперимента, позволяя снизить экспериментальную ошибку (стохастические ошибки секвенирования).

Благодаря возможности кодирования образцов присоединением адаптеров заданной последовательности исследователь может довольно легко решить задачу инструментальных повторностей (секвенирование одного и того же образца несколько раз).

Использование разных NGS-платформ (особенно основанных на разных принципах секвенирования) также может существенно улучшить результаты. Например, комбинация длинных, но не очень точных прочтений на платформе PacBio и точных, но коротких прочтений на Illumina.

С повторностями связана чувствительность и специфичность эксперимента. В частности, повторности могут быть использованы для определения чувствительности и специфичности методов поиска полиморфизма последовательностей ДНК.

#### **6.4. ОСНОВНЫЕ ТИПЫ ОШИБОК ПРИ СЕКВЕНИРОВАНИИ**

Ниже приведен перечень типовых ошибок, допускаемых в ходе экспериментов с использованием с NGS. На этапе планирования эксперимента исследователю следует обратить особое внимание на перечисленные ниже источники ошибок во избежание их совершения.

Источники ошибок на этапе пробоподготовки:

- 1) ошибки лаборанта (например, неправильная маркировка образцов);
- 2) деградация ДНК и РНК в ходе хранения и транспортировки образцов;
- 3) загрязнение (контаминация) образца чужеродными (не относящимися к исследованию) нуклеиновыми кислотами;
- 4) недостаточное количество (или низкая концентрация) образца ДНК.

Источники ошибок на этапе подготовки библиотеки для NGS:

- 1) ошибки лаборанта – например, перекрестное загрязнение (кросс-контаминация) образцов;

- 2) искажения в ходе ПЦР (разная эффективность амплификации фрагментов, низкая начальная концентрация образца и большое число циклов ПЦР);
- 3) искажения в ходе обогащения поли-А-фракции РНК;
- 4) аппаратные ошибки (например, некорректная работа ПЦР-амплификатора);
- 5) образование химерных прочтений;
- 6) ошибки «штрих-кодирования» и других подходов с присоединением адаптеров (использование несовместимых «штрих-кодов»).

Источники ошибок на этапе собственно секвенирования:

- 1) ошибки лаборанта (например, перекрестные помехи в получении сигнала от разных микросфер в результате перегрузки чипа);
- 2) сбой фазы считывания (за счет неполного удлинения или добавления нескольких нуклеотидов вместо одного нуклеотида);
- 3) сложные для прочтения регионы (GC-богатые или гомополимерные участки);
- 4) аппаратные ошибки (например, выключение электропитания, выход из строя лазера, помп и приводов, жесткого диска, сбой программного обеспечения).

## 6.5. ВАРИАНТЫ ПРИМЕНЕНИЯ NGS

За последние несколько лет наблюдается шквал публикаций с применением высокопроизводительного секвенирования для различных целей. Наиболее популярные приложения NGS включают в себя: полногеномное секвенирование (*de novo* или повторное); исследование различных РНК (часто называемое RNA-Seq); крупномасштабный анализ метилирования ДНК; изучение ДНК-белковых взаимодействий (ChIP-seq) и ряд других (табл. 6.1). В ближайшие годы список приложений NGS, несомненно, будет расти благодаря непрерывному усовершенствованию существующих подходов и выходу на рынок принципиально новых технологий (в частности, технологий секвенирования одиночных молекул нуклеиновых кислот).

Таблица 6.1

## Варианты применения NGS

Вариант применения	Решаемая биологическая задача
Полногеномное секвенирование <i>de novo</i>	Реконструкция работы клетки и организма на молекулярном уровне, эволюционная геномика
Полногеномное повторное секвенирование	Поиск генетических вариаций
Метагеномное секвенирование	Исследование биоценоза, поиск новых видов живых систем
Секвенирование транскриптомов	Исследование генной экспрессии, аннотация генома
Секвенирование малых РНК	Исследование генной экспрессии
Таргетное секвенирование	Поиск генетических вариаций
Секвенирование обработанной бисульфитом ДНК	Исследование профиля метилирования
Секвенирование иммунопреципитированного хроматина	Полногеномное картирование ДНК-белковых взаимодействий
Секвенирование единичных клеток	Исследование генной экспрессии, секвенирование некультивируемых бактерий

## СПИСОК ЛИТЕРАТУРЫ

1. Auer P.L., Doerge R.W. Statistical design and analysis of RNA sequencing data // Genetics, 2010, 185 (2): 405–416.
2. Churchill G.A. Fundamentals of experimental design for cDNA microarrays // Nat Genet., 2002, 32: 490–495.
3. Robasky K., Lewis N.E., Church G.M. The role of replicates for error mitigation in next-generation sequencing // Nat Rev Genet., 2014, 15 (1): 56–62.

# **СЕКВЕНИРОВАНИЕ ИНДИВИДУАЛЬНЫХ ГЕНОМОВ И ТРАНСКРИПТОМОВ ПРОКАРИОТ**

В секвенировании прокариотических геномов в настоящее время можно выделить два основных направления: секвенирование отдельных микробов (в том числе некультивируемых) с целью восстановления метаболических путей и понимания эволюционных закономерностей существования данного вида и секвенирование микробных сообществ (с теми же целями, плюс исследование структуры сообщества). Различные аспекты секвенирования бактериальных сообществ описаны в главе 8, в этой главе рассмотрены некоторые особенности секвенирования генома отдельного вида бактерий.

## **7.1. РОЛЬ NGS В МИКРОБИОЛОГИИ**

Появление высокопроизводительного секвенирования дало значительный толчок развитию микробиологии и смежным областям. Дело в том, что наблюдающийся в течение последних трех лет быстрый рост количества практически полных последовательностей геномов бактерий в открытых базах данных позволил по-новому взглянуть на эволюцию прокариот, обнаружить принципиально новые типы генетических событий на уровне возникновения видов живых организмов, с тем чтобы эффективно восстанавливать метаболизм одноклеточных. Для некультивируемых микроорганизмов, например микроорганизмов, живущих в экстремальных условиях среды, геномные методы становятся важнейшим способом их изучения и отправной точкой для последующей разработки биохимических методов и методов геномной инженерии.

Другая область применения NGS в микробиологии – это медицина. Исследование генома отдельных микробов методами NGS позволяет быстро выявлять «островки патогенности» и локусы, определяющие устойчивость к антибиотикам, а изучение микробных сообществ человека – понять взаимосвязь микробиоценозов и состояния здоровья пациента (см. гл. 8).

Еще одно важное направление – биотехнологии. Применяя бактерии в различных промышленных процессах, на основе геномных данных можно идентифицировать новые ферменты и понять, как модификации генома связаны с функционированием микроорганизмов.

Наконец, NGS имеет значение для молекулярной биологии в целом в части понимания основных механизмов функционирования живой клетки.

На начало 2014 года определено около 1500 полных геномов микроорганизмов, а число частично собранных геномов (представленных в виде множества неупорядоченных фрагментов) составляет уже около десяти тысяч и быстро растет.

## 7.2. ИСТОРИЯ СЕКВЕНИРОВАНИЯ БАКТЕРИАЛЬНЫХ ГЕНОМОВ

Первые полные геномы бактерий, определяемые методом Сенгера, начали публиковать, начиная с конца 1990-х годов. Геномы кишечной палочки (*Escherichia coli*) и *Helicobacter pylori* были опубликованы в 1997 году [1, 2]. А далее последовал шквал полногеномных исследований прокариот.

Секвенирование бактериального генома методом Сенгера производили, создавая сначала библиотеку длинных фрагментов (30–40 т. п. н.) в векторах на основе геномов фагов (космидах или фагмидах) с получением картированной коллекции фрагментов, покрывающих весь геном бактерии. Затем каждый крупный фрагмент субклонировали в плазмидах фрагментами по 2–3 т. п. н. и секвенировали их насквозь (часто используя не только стандартные праймеры на вектор, но и множество геном-специфичных праймеров) [3, 4]. Такой проект обычно продолжался более года и стоил сотни тысяч (если не миллионы) долларов США.

Позднее, с удешевлением стоимости секвенирования (за счет выпуска более производительных секвенаторов, но работающих по-прежнему на принципе Сенгера), исследователи стали использовать более простой подход: метод дробовика, когда геном сразу клонируют в плазмидный вектор вставками по 2–3 т. п. н. и секвенируют с многократным покрытием генома по «сырым» данным (разд. 1.1.2).

В настоящее время методы NGS полностью вытеснили принцип Сенгера из области геномного секвенирования. Однако современный подход очень похож на метод дробовика: получают последовательность коротких фрагментов генома, добиваясь сборки 95–98% последовательности генома.

### 7.3. ОПРЕДЕЛЕНИЕ ПОЛНОЙ ПОСЛЕДОВАТЕЛЬНОСТИ БАКТЕРИАЛЬНОГО ГЕНОМА *DE NOVO*

Так же как и при использовании метода Сенгера, определение последовательности бактериального генома методами NGS начинают с секвенирования библиотеки случайных фрагментов ДНК.

Наилучшие результаты показывает алгоритм секвенирования, использующий оба типа библиотек: обычную и инвертированную (см. гл. 2). После секвенирования обычной библиотеки исследователь получает хорошее покрытие генома, но сборка дает много сравнительно коротких контигов. Для сборки контигов в более длинные регионы можно использовать повторное прочтение генома с использованием инвертированной библиотеки. При совмещении данных из двух библиотек можно добиться 99% сборки генома.

Наиболее продолжительным и трудозатратным этапом любого бактериального геномного проекта является его завершение с получением полной кольцевой молекулы ДНК. Сегодня, при использовании технологий NGS, 98% генома можно получить в течение недели, а получение полной кольцевой молекулы может занять несколько лет.

Для закрытия пробелов в последовательности и сборки контигов в полный геном существует несколько подходов. Один из наиболее надежных, но и трудоемких подходов – создание космидной библиотеки крупных фрагментов бактери-

ального генома (по 40 т. п. н.). Путем скрининга космидной библиотеки при помощи ПЦР (с праймерами на концевой регион контига) можно быстро и недорого определить, на какой именно фрагмент космидной библиотеки приходится пробел, после чего определить последовательность данного клона методом дробовика (по Сенгеру).

Другой подход: классическая «прогулка по геному» (genome walking). Для этого получают библиотеку фрагментов бактериального генома (расщепленного по сайту редкощепящей эндонуклеазы рестрикции) с прилигированными супрессионными адаптерами, после чего проводят ПЦР с праймером на концевой участок контига и дистальным праймером на адаптер [5].

Если в базах данных есть геном близкого вида, можно провести сравнение полученных последовательностей с родственным геномом и попытаться расположить контиги друг относительно друга. Для контигов, предположительно расположенных рядом (на расстоянии не более 5–7 т. п. н.), синтезируют праймеры на концевые участки и проводят «дальнобойную» ПЦР, оптимизированную для наработки длинных фрагментов (long-range PCR). Полученные фрагменты секвенируют пошагово методом Сенгера (всякий раз заказывая новый праймер на концевой участок известной последовательности).

Наиболее сложными при сборке являются регионы повторяющихся последовательностей, длиннее одного прочтения методом Сенгера (т. е. свыше 800–1000 п. н.). Такие регионы приходится собирать опосредованными методами, например ориентируясь на длину сквозного ПЦР-продукта, предполагая, что весь регион состоит из одинаковых повторов (но следует помнить, что это лишь допущение и истинная структура региона не известна).

Еще один вариант: если не удастся перекрыть весь участок повторов с помощью «дальнобойной» ПЦР, можно попробовать найти какие-то зацепки в дистальной части уже имеющейся последовательности. Это могут быть полиморфные основания в повторе, микроделеции, «неправильные» стыки повторов и т. п. Существуют примеры программного обеспечения, специально адаптированного для поиска таких мест в повторяющихся регионах<sup>1</sup>.

---

<sup>1</sup> Например, <http://metagenome.ru/ru/data-analysis>.

## 7.4. ПРИМЕР ПРОТОКОЛА СЕКВЕНИРОВАНИЯ ОБРАЗЦА БАКТЕРИАЛЬНОЙ ДНК

В данном разделе рассмотрен протокол приготовления и секвенирования библиотеки геномной ДНК бактерии при помощи технологии секвенатора Ion PGM (с длиной прочтения 200 п. н.) с практическими рекомендациями пользователю.

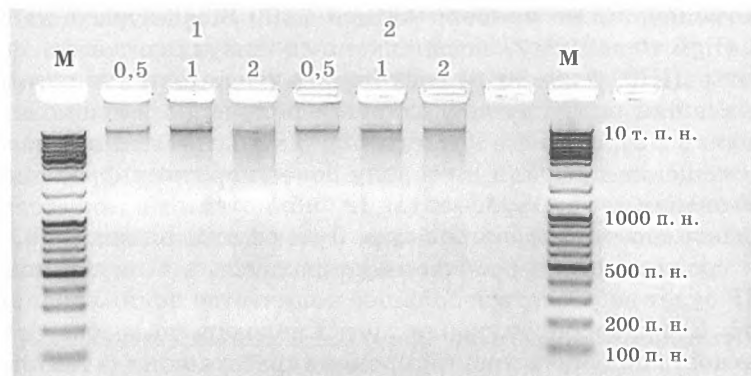
Протокол состоит из следующих этапов.

1. Оценка качества и количества исходной ДНК.
2. Ферментативное разрушение ДНК.
3. Лигирование адаптеров для секвенирования, амплификация библиотеки.
4. Отбор оптимальной фракции библиотеки (size-select).
5. Оценка качества и количества библиотеки ДНК.
6. Эмульсионная ПЦР и отбор сработавшей фракции микросфер.
7. Подготовка библиотеки к загрузке на чип.
8. Загрузка на чип.
9. Секвенирование.

### 7.4.1. Оценка качества и количества исходной ДНК

Важным этапом NGS является контроль качества и количества исходной ДНК: необходимо, чтобы она была как можно менее деградированной, а ее количество было в несколько раз больше требуемого по протоколу (это позволяет сохранить ДНК для повторения эксперимента или иных исследований того же самого генома позже).

Предварительно оценить качество и количество ДНК можно с помощью обычного гель-электрофореза: экспериментатор не должен наблюдать больших количеств РНК в примеси (практика показывает, что РНК в целом не оказывает существенного влияния на результат, но может повлиять на оценку количества ДНК некоторыми методами) и значительной смазанности в сторону легких фрагментов в результате деградации ДНК (рис. 7.1).



**Рис. 7.1.** Гель-электрофорез геномной ДНК (6 образцов) удовлетворительного для приготовления библиотек качества. Крайние дорожки – маркер длин фрагментов (100–10 000 п. н.)

Для более точной количественной оценки ДНК, помимо метода ПЦР «в реальном времени», можно использовать флуориметры типа Qubit от Life Technologies Thermo Fisher Scientific. Это достаточно точный прибор, позволяющий получить результат за пару минут.

#### **7.4.2. Ферментативное расщепление ДНК и очистка от фермента**

В качестве метода разрушения ДНК можно использовать ферментативный подход с использованием набора реагентов Ion Shear Plus Reagents или аналогов (например, компании New England Biolabs), так как он не требует еще одной операции («затупления» концов фрагментов), а также легко оптимизируется под задачи большинства лабораторий. Так как в описываемом примере планируется секвенировать библиотеку с длиной прочтений в 200 п. н., протокол рекомендует использовать время инкубации с нуклеазами около 15 мин.

Желательно перед работой с «боевой» ДНК провести пробное расщепление какой-нибудь геномной ДНК, имеющейся в достаточном количестве с разным временем инкубации – от 5 до 30 мин, для подбора оптимальных условий. Контролировать качество получившегося распределения фрагментов

рекомендуется на приборе Agilent 2100 Bioanalyzer с наборами High Sensitivity, позволяющими визуализировать фрагменты ДНК от 20–30 до нескольких тысяч пар нуклеотидов. Идеальным разрушением является получение «колокола» на графике с вершиной в области 220–280 п. н. Если наблюдается смещение профиля в сторону более коротких фрагментов, это говорит о гиперфрагментации библиотеки – в таком случае эффективность секвенирования будет ниже ожидаемой, так как средняя длина прочтения сократится, а в эмульсионной ПЦР будет наблюдаться большое количество поликлональных сфер. Если, наоборот, наблюдается недовар, то возрастает вероятность получить низкий процент сработавших («темплированных») микросфер (от англ. *template* – матрица, модель), и, тем самым, секвенирование даст мало первичных данных.

После расщепления необходимо провести очистку ДНК от фермента, буфера и низкомолекулярной фракции фрагментов. В настоящее время доступно множество решений, начиная от обычного переосаждения ДНК спиртом и заканчивая наборами реагентов с микроколонками с сорбционной очисткой. Популярными являются наборы реагентов AMPure XP beads от Beckman, представляющие собой микрогранулы из полистирола, покрытые магнитным и полимерным слоями, последний из которых содержит карбоксильные группы. ДНК может ковалентно связываться с карбоксильными остатками на поверхности частиц в присутствии полиэтиленгликоля и большого количества соли, а в буфере Трис–ЭДТА или дистилляте переходить в раствор. Преимуществом таких колонок является высокая емкость сорбента (1 мкл частиц может связать до 3 нг ДНК), а также низкие потери при очистке – все это позволяет работать с даже очень небольшими количествами ДНК без опасения потерять ценную материю при промывках.

Очистка с использованием сорбирующих частиц состоит из следующих этапов:

- 1) связывание ДНК на поверхности сорбирующих частиц в присутствии соли и ПЭГ;
- 2) отбор сорбирующих частиц из раствора при помощи магнитного штатива;
- 3) промывка сорбента этанолом;
- 4) элюция ДНК с поверхности частиц с помощью Трис–ЭДТА или mQ-воды.

Еще одним преимуществом очистки на сорбирующих частицах является возможность проводить селективный отбор фрагментов нужной длины, регулируя соотношение концентрации сорбента и ДНК – чем меньше сорбента, тем более длинные фрагменты ДНК будут осаждаться на частицах и проходить очистку.

Заметим, что для хорошей работы набора реагентов AMPure XP beads крайне важно перед использованием дождаться нагрева всех реагентов до комнатной температуры.

### 7.4.3. Лигирование адаптеров для секвенирования

После разрушения и очистки к фрагментам ДНК лигируют адаптеры, необходимые для проведения эмульсионной ПЦР и непосредственно секвенирования. В целом эта процедура не представляет особой сложности. Однако, если исходное количество ДНК было меньше рекомендуемых 100 нг (на порядок и более), необходимо подбирать оптимальные условия лигирования, разбавляя адаптеры. Практика показывает, что, если в лигирование пошло 3–5 нг ДНК (вместо 100 нг), следует разводить адаптеры в 4 раза.

В протоколе производитель обращает на это внимание, но все же отметим еще раз: при работе со «штрих-кодированными» адаптерами нужно быть крайне внимательным и пипетировать очень аккуратно, так как кросс-контаминация «штрих-кодов» приведет к невозможности правильной интерпретации результатов секвенирования.

Как и на предыдущем этапе, после ферментативной обработки проводят процедуру очистки с AMPure XP beads. Отметим, что на этапе подсушивания частиц с ДНК перед элюцией Трис-ЭДТА не следует пересушивать микрочастицы до состояния порошка, когда они свободно падают на дно пробирки, снятой с магнитного штатива, так как это ухудшает элюцию.

В случае малого количества исходной ДНК или больших потерь на предыдущих этапах можно провести амплификацию библиотеки с праймеров, комплементарных лигированным адаптерам. Отрицательной стороной амплифицированной библиотеки является искажение представленности фрагментов из-за разной эффективности ПЦР для разных молекул-мишеней. В результате количество полезной для анализа информа-

ции может значительно (в несколько раз) сократиться, если библиотеку «переамплифицировать». Поэтому для начала желательно провести тестовую амплификацию, взяв на ПЦР небольшую аликвоту и поставив несколько реакций с разным числом циклов – к примеру, от 6 до 12 с шагом в 2 цикла. По окончании тестовой ПЦР необходимо измерить количество и качество полученной ДНК на флуориметрах Qubit или Agilent 2100 Bioanalyzer. Для экспериментальной амплификации со всей библиотекой необходимо выбрать минимальное число циклов, дающее достаточное для анализа количество ДНК, а также возможность сохранить не менее половины объема библиотеки (если с первой частью в процессе эмульсионной ПЦР возникнут проблемы).

#### **7.4.4. Отбор оптимальной фракции библиотеки (size-select)**

Существенное улучшение качества секвенирования достигается за счет оптимизации размера библиотеки: лучше всего, если длина фрагментов будет укладываться в узкий рекомендуемый диапазон, в этом случае можно будет ожидать невысокого уровня поликлональности, длинных прочтений и в целом высокого качества данных. Процедура отбора фракции библиотеки подробно описана в разделе 2.7.

#### **7.4.5. Оценка качества и количества библиотеки фрагментов ДНК**

После size-select (а по большому счету и до него) желательно провести контроль качества библиотеки на Agilent 2100 Bioanalyzer – это позволит оценить, нужна ли фракция была отобрана, нет ли примесей праймеров или димеров праймеров, которые использовались на предыдущем этапе амплификации библиотеки фрагментов. Если библиотека удовлетворяет всем критериям (см. рис. 2.7), то необходимо оценить ее количество для определения фактора разведения библиотеки (TDF – template dilution factor).

Напомним, что эмульсионная ПЦР – это реакция амплификации ДНК, при которой реакция осуществляется в отдельных водных микрокаплях, взвешенных в гидрофобной среде (см. разд. 2.9.2). Задачей эмульсионной ПЦР в практике

NGS является формирование микрореакторов такого размера, чтобы в одну водную каплю попадали лишь одна молекула из библиотеки фрагментов ДНК и одна микросфера, – именно это позволяет получить микросферу, покрытую копиями (~5–10 млн штук) исходной молекулы, сигнал от которых будет достаточным для качественного секвенирования. Эмульсионная ПЦР крайне чувствительна к соотношению реагентов: если «переложить» ДНК, то в одном микрореакторе окажется более одной молекулы, – такая микросфера окажется поликлональной и будет исключена из результатов секвенирования. Если же «недоложить», на выходе окажется мало сработавших микросфер – загрузка чипа для секвенса будет невысокой. В итоге, отклонение от области оптимума (узкий диапазон концентраций, отличающихся от оптимального значения не более чем в 2–2,5 раза) приводит к существенному ухудшению результата секвенирования.

В принципе, количественно оценить библиотеку можно с помощью как Qubit, так и Agilent 2100 Bioanalyzer, однако Qubit будет измерять не только ДНК, фланкированную обоими типами адаптеров, но и фрагменты с одинаковыми адаптерами (A–A или B–B), а также любые другие примеси ДНК, присутствующие в образце. Bioanalyzer, в свою очередь, определяет концентрацию образца, сравнивая ее с концентрацией размерного стандарта, который в единственном повторе в объеме 1 мкл добавляется на чип для анализа пользователем; поэтому уровень искажений данной методики также довольно высок и выходит за пределы оптимальных значений.

Наиболее точным способом измерения концентрации библиотеки авторам видится использование количественной ПЦР. Как и на предыдущих этапах, у пользователя есть выбор: можно использовать фирменные наборы реагентов от Life Technologies Thermo Fisher Scientific (например, Ion Library PCR Quantitation Kit) или аналоги от независимых производителей. Как правило, и те, и другие работают хорошо. Измерение проводится относительно серии разведений стандартной контрольной библиотеки, концентрация которой известна заранее. В анализ попадают только молекулы, имеющие адаптеры A–B, так как именно на них садятся меченые праймеры. В этом типе измерения важно аккуратно приготовить разведения контрольной и экспериментальной

библиотек в независимых повторах и хорошо перемешать ПЦР-смесь с образцами ДНК.

#### 7.4.6. Эмульсионная ПЦР и обогащение микросфер

Следующим этапом после приготовления библиотеки является эмульсионная ПЦР. Действовать следует строго по протоколу, не забывая, что добавлять в реакцию необходимо разведенную согласно TDF библиотеку. Еще одним тонким местом является добавление микросфер – их очень важно тщательно перемешивать перед добавлением к водной фазе. Можно даже исключить центрифугирование в течение 2 с после встряхивания, чтобы не осадить часть сфер на дно микропробирки.

Эмульсионная ПЦР проходит в полностью автоматическом режиме и длится 4–6 ч в зависимости от используемого типа реагентов для Ion PGM. После окончания реакции необходимо аккуратно промыть микросферы согласно протоколу – это нужно сделать оперативно, сразу после остановки центрифуги OneTouch 2 в 10-минутном режиме Final Spin.

Отмытые микросферы необходимо обогатить, т. е. избавиться от несработавших сфер. Принцип обогащения основан на использовании магнитных микрочастиц, на внешней поверхности которых закреплены молекулы стрептавидина. Один из праймеров, использующихся для амплификации в эмульсионной ПЦР (отжигающийся на адаптере A) мечен биотином, поэтому сработавшие микросферы легко можно «вытащить» магнитом благодаря комплексу биотин–стрептавидин с магнитными частицами.

Процесс обогащения длится около 35 мин и проводится с помощью прибора OneTouch ES, по сути представляющего собой пипетирующую станцию, работающую со специальным 8-пробирочным стрипом и имеющую магнитные зоны в основании. Как правило, проблем с обогащением у пользователей не возникает, и хорошим результатом считается обогащение на уровне 95–100%.

Стоит отметить, что на начальных стадиях серий экспериментов, а также в случае смены типа реагентов (например, с 200-й серии на 400-ю и наоборот) существует способ качественной оценки эффективности эмульсионной ПЦР еще до проведения секвенирования. Для этого можно использовать

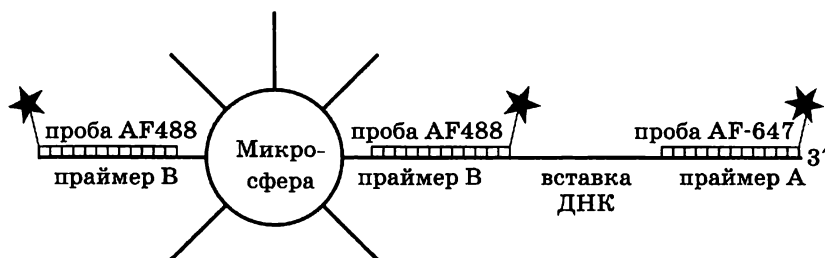


Рис. 7.2. Схема локализации меченых олигонуклеотидов из набора Ion Sphere Quality Control Kit

набор реагентов Ion Sphere Quality Control Kit. Принцип его действия основан на использовании двух олигонуклеотидов, меченых флуороформами AF488 и AF647 соответственно. Олигонуклеотид 488 комплементарен участкам В на микро-сферах и может отщепиться там в том случае, если В не занят амплифицированной на сфере молекулой библиотеки ДНК. Олигонуклеотид 647 комплементарен участку А, «обладателями» которого являются только сработавшие сферы (рис. 7.2). Таким образом, сравнивая интенсивность флуоресценции по флуорофорам AF488 и AF647, можно определить долю сработавших сфер. В качестве материала для анализа используют небольшую аликвоту необогащенных ISP-сфер (т. е. взятых сразу после проведения эмульсионной ПЦР). Измерение проводят на флуориметре Qubit 2, который используется для определения концентрации ДНК на предыдущих этапах. Оптимальной является доля сработавших микро-сфер 20–25%.

#### 7.4.7. Подготовка библиотеки к загрузке на чип

Обогащенные сработавшие микро-сферы с проверенным результатом эмульсионной ПЦР готовы к загрузке на чип. В настоящее время для платформы Ion PGM выпускают три типа чипов – 314, 316 и 318, оснащенных разным числом ячеек для секвенирования и, тем самым, позволяющих проводить секвенирование с разным объемом генерируемой информации. Практика показывает, что наиболее сложными в работе являются чипы 314 с наименьшей емкостью пространства

чипа (всего 7 мкл). В случае с 316 и 318 чипами емкость составляет уже 30 мкл.

Подготовка микросфер к загрузке на чип состоит из следующих этапов:

- 1) добавление к сработавшим микросферам контрольных микросфер;
- 2) отжиг праймера для секвенирования;
- 3) добавление к сферам полимеразы для секвенирования.

Здесь следует обратить особое внимание на качественное перемешивание всех компонентов: микросферы легко осаждаются на дно пробирки и могут остаться без «своих» праймера или полимеразы, что приведет к низкому качеству их секвенирования и удалению этих результатов из финальных данных секвенирования.

#### **7.4.8. Загрузка микросфер на чип**

В настоящее время доступны два способа загрузки микросфер на чип: с помощью пипетирования и методом центрифугирования (центрифуга поставляется в комплекте с Ion PGM, с адаптерами со смещенным центром тяжести). Советовать, какой из способов стоит предпочесть, авторы не будут, так как все зависит от личных предпочтений каждого исследователя. В принципе, и тот, и другой способы дают одинаково хорошую загрузку при достаточном опыте. Стоит лишь отметить, что крайне важно потренироваться («набить руку») в пипетирующем способе на б/у-чипе, так как поведение потока жидкости на каждом из трех типов чипов отличается.

#### **7.4.9. Запуск процесса секвенирования на приборе**

Процесс секвенирования на Ion PGM начинается с промывки прибора. Отклоняться от протокола здесь не стоит, так как хлоритные и водные промывки, которые нужно регулярно проводить, являются простой процедурой и не занимают много времени. При этом важно следить за качеством газа (азота или аргона) и уровнем давления в газовом баллоне, от которого работают системы подачи жидкости в Ion PGM – падение давления ниже критического приведет к прерыванию работы прибора.

Инициализация прибора, т. е. приведение всех реагентов в рабочее состояние, проводится перед каждым запуском (заметим, что многие комбинации чипов и реагентов позволяют проводить подряд два секвенирования на одной инициализации) согласно протоколу, что позволяет сэкономить реагенты. Следует обратить внимание на качество воды (производитель рекомендует Milli-Q Water Purification System компании Millipore Corporation), используемой для приготовления W2-буфера, – это очень важный параметр. Система Milli-Q должна находиться неподалеку от прибора, чтобы сократить время доставки емкости с W2 от Milli-Q-станции до Ion PGM Torrent. Дело в том, что растворяющийся в воде  $\text{CO}_2$  из воздуха приводит к ее «забуфериванию» и снижает качество сигнала при секвенировании. Напомним, что Ion PGM детектирует изменение pH в каждой отдельной ячейке чипа, которое происходит при включении нуклеотида в строящуюся цепь ДНК и выделении в среду  $\text{H}^+$  (см. разд. 4.1).

## 7.5. АНАЛИЗ ДАННЫХ ГЕНОМНОГО СЕКВЕНИРОВАНИЯ БАКТЕРИЙ

Анализ геномных данных, несомненно, сложнее процесса их получения. Вообще, в настоящее время мы наблюдаем в научной сфере интересный парадокс: после выхода на рынок технологий NGS скорость накопления данных о последовательностях геномов существенно превысила скорость из анализа. Купить оборудование для NGS и освоить коммерческую технологию любая лаборатория (имеющая не это деньги) может за очень короткий срок (за пару месяцев). На создание же эффективной научной группы из высококлассных профессионалов – биологов, биоинформатиков и программистов – в некоторых случаях могут уйти годы.

Анализ геномных данных можно разделить на две составляющих: аннотацию генома (т. е. идентификацию генов в последовательности и предсказание их функций) и биологический анализ (реконструкцию метаболизма, установление взаимосвязи фенотипа и генотипа и т. д.). Задача аннотирования прокариотических геномов на сегодняшний день имеет ряд неплохих решений (а продолжающиеся биоинформатические разработки в этой области сулят появление в скором времени

еще лучших алгоритмов). Существующие программные продукты позволяют в автоматическом режиме получить достаточно хорошо аннотированный геном<sup>1</sup>.

Отдельно следует указать на особенности полногеномного секвенирования *de novo* некультивируемых бактерий. Вследствие малого количества стартовой ДНК ее, как правило, подвергают амплификации (по технологии WGA). Наличие этапа амплификации сильно усложняет задачу последующего анализа данных, поскольку покрытие амплифицированного генома оказывается очень неравномерным (вследствие искажений амплификации), а при использовании инвертированных библиотек с двусторонним прочтением длину вставки гораздо сложнее контролировать. Кроме того, в процессе амплификации возникают ошибки и химерные прочтения. Существуют специальные программы для сборки таких (амплифицированных) геномов [6, 7].

Биологический анализ – это понимание того, как продукты генов работают в системе и как микроорганизм использует имеющийся набор генов в тех или иных условиях. Этот этап пока не подлежит автоматизации и зависит от квалификации и эрудиции исследователя.

## 7.6. СЕКВЕНИРОВАНИЕ ТРАНСКРИПТОМА ПРОКАРИОТ

Бактериальный транскриптом (по аналогии с транскриптомом эукариотическим) – это совокупность всех транскриптов данной бактериальной клетки. В среднем на рибосомальную РНК у бактерий приходится 80% от суммарной РНК (по массе), на транспортную РНК – 15%, на матричную и не кодирующие РНК – 4–5%.

Сложность бактериального транскриптома долгое время была недооценена. Лишь с созданием технологий высокопроизводительного секвенирования удалось показать, что у бактерий есть практически все те же механизмы регуляции экспрессии, что и у эукариот. Так, у бактерий были найдены *транс*-кодируемые малые РНК. Это участки, расположенные в пространстве между белок-кодирующими генами и способные взаимодействовать с другими РНК либо белками. Найде-

---

<sup>1</sup> <http://metagenome.ru/ru/data-analysis>

ны у бактерий и *цис*-кодируемые малые РНК. Они могут занимать 1–2 гена по длине или же иметь протяженность в несколько генов [8]. Найдены рибопереклюкатели – структуры, чаще всего сенсорные, расположенные в 5'-концевой нетранслируемой области РНК, способные изменять конфигурацию транскрибируемой молекулы и тем самым либо препятствовать, либо помогать экспрессии. У бактерий обнаружены безлидерные РНК (РНК без транслируемой области, в связи с чем меняется схема сборки рибосом на таких транскриптах).

У прокариот обнаружен широкий спектр эпигенетических модификаций, влияющих на спектр транскриптов в транскриптоме. Обнаружена компартиментализация транскриптов: было показано, что транскрипты белков, которые в дальнейшем будут взаимодействовать, обычно транскрибируются рядом. Так как у бактерий транскрипция и трансляция сопряжены, время сборки необходимых белковых комплексов у бактерий сокращено.

Описаны различные модификации бактериального сплайсинга, полиаденилирования, редактирования РНК и т. д. [9].

Упрощенная схема исследования бактериального транскриптома такова: выделение РНК из интересующей бактерии, синтез кДНК, создание библиотеки фрагментов и собственно секвенирование. На каждой стадии возможны дополнительные манипуляции. Например, можно провести обогащение РНК интересующей последовательностью. Можно отобрать фракцию только малых РНК либо рибосомальных РНК.

В конечном итоге после секвенирования исследователь получает набор файлов (например, в формате FASTQ, см. разд. 4.1). Базовая схема анализа данных секвенирования бактериального транскриптома стандартна: вначале проводят контроль качества полученных файлов (например, с помощью программы FASTQC). Затем проводят фильтрацию прочтений и удаление адаптерных последовательностей. После этого проводят картирование прочтений на референсный геном (при этом могут быть использованы программы Bowtie, BWA, SOAP и др.). Файлы, содержащие картированные последовательности, отбирают и проводят качественный либо количественный анализ данных (а чаще – оба вида анализа). В ходе качественного анализа транскриптома выполняют визуализацию – распределение прочтений по геному бактерии. При количественном анализе определяют функциональные

категории или метаболические пути, в которых могут быть представлены найденные гены.

В результате качественного анализа бактериальных транскриптов можно найти новые гены – как короткие (кодирующие белки), так и гены новых некодирующих РНК. Часто в ходе анализа уточняют границы генов, выявляют некодирующие и нетранслируемые участки. Так как у бактерий большинство генов собрано в опероны, можно выявлять опероны (как области покрытия генома прочтениями, включающие сразу несколько генов). Наконец, можно выявлять антисмысловую РНК – это покрытие прочтениями генома по цепи, противоположной смысловой цепи известного гена. Для многих бактерий антисмысловая транскрипция может составлять существенный процент транскриптома. В ряде случаев он может превышать объем транскрипции по смысловой цепи. За счет сопоставления количества прочтений в двух и более библиотеках можно проводить поиск дифференциально представленных транскриптов.

Важно отметить, что в ряде биологических задач исследование транскриптома (или протеома) может быть крайне информативным. Так, было показано, что бактерии *Helicobacter pylori* из разных участков одного и того же пораженного желудка не отличаются генетически, но сильно отличаются по уровню экспрессии генов (т. е. бактерии находятся в разных функциональных состояниях). При этом профиль экспрессии может значительно отличаться – до сотни генов с существенно разным уровнем экспрессии.

Ряд работ демонстрирует, что даже в культуре каждая бактерия уникальна по профилю экспрессии генов. Анализируя транскриптом в популяции бактерий, исследователь измеряет «среднюю температуру по больнице». Поэтому в перспективе исследование бактериальных транскриптомов методами NGS должно стремиться к возможности определения профиля представленности транскриптов одной клетки [10].

## СПИСОК ЛИТЕРАТУРЫ

1. *Blattner F.R. et al.* The complete genome sequence of *Escherichia coli* K-12 // *Science*, 1997, 277 (5331): 1453–1462.
2. *Tomb J.F. et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori* // *Nature*, 1997, 388 (6642): 539–547.
3. *Kogan Y. et al.* *Rhodobacter capsulatus* genome project. Half Way Through // *Microb. Compar. Genomics*, 1998, 3: 78–79.
4. *Haselkorn R. et al.* The *Rhodobacter capsulatus* genome // *Photosynthesis Research*, 2001, 70: 43–52.
5. *Siebert P.D. et al.* An improved PCR method for walking in uncloned genomic DNA // *Nucleic Acids Res.*, 1995, 23 (6): 1087–1088.
6. *Bankevich A. et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // *J Comput Biol.*, 2012, 19(5): 455–477.
7. *Nurk S. et al.* Assembling single-cell genomes and mini-metagenomes from chimeric MDA products // *J Comput Biol.*, 2013, 20 (10): 714–737.
8. *Baroni D., Arrigo P.* MicroRNA target and gene validation in viruses and bacteria // *Methods Mol Biol.*, 2014, 1107: 223–231.
9. *Bandyra K.J., Luisi B.F.* Licensing and due process in the turnover of bacterial RNA // *RNA Biol.*, 2013, 10 (4): 627–635.
10. *Waldbauer J.R., Rodrigue S., Coleman M.L., Chisholm S.W.* Transcriptome and proteome dynamics of a light-dark synchronized bacterial cell cycle // *PLoS One.*, 2012, 7 (8): e43432.

# ИССЛЕДОВАНИЕ МИКРОБНЫХ СООБЩЕСТВ МЕТОДАМИ NGS

К настоящему времени описано около 10 000 видов бактерий, однако некоторые исследователи полагают, что общее число видов прокариот на Земле достигает миллиона (причем большая часть из них не культивируется). Истинное разнообразие некультивируемых микроорганизмов стало понятно после появления молекулярно-генетических методов исследования бактериальных сообществ [1]. При исследовании микроорганизмов из различных сред, в том числе живущих в экстремальных условиях среды, геномные методы стали важнейшим способом их изучения и отправной точкой для последующей разработки биохимических методов и методов геной инженерии. В этом смысле секвенирование – и особенно высокопроизводительное секвенирование – стало новым этапом в развитии микробиологии как науки.

Для исследования сообществ организмов, обитающих в различных средах, в последние годы все чаще обращаются к «мета»-исследованиям, т. е. секвенированию образцов, получаемых напрямую из окружающей среды [2]. Это могут быть метабеномные, метатранскриптомные и даже метапротеомные данные. Такие исследования позволяют обнаружить ранее неизвестные некультивируемые микроорганизмы, описать их свойства и функции, обнаружить еще не изученные биомолекулы, а также получить более общую картину того, что происходит в той или иной среде.

Основными результатами секвенирования являются таксономический анализ сообщества либо реконструкция типа сообщества (например, по результатам анализа метабенома можно установить, что сообщество хемолитотрофно и доминирующим процессом в нем является сульфаторедукция).

Методы высокопроизводительного секвенирования хорошо подходят для исследования бактериальных сообществ, причем как для определения видового разнообразия и коли-

чественного состава сообщества (микробиоценоза), так и для секвенирования полных геномов присутствующих в нем (часто некультивируемых) организмов. А поскольку для прокариот характерен горизонтальный перенос генов, зачастую для понимания особенностей микробиоценоза важно не столько наличие гена (скажем, устойчивости к антибиотику) у конкретного вида, сколько его присутствие в сообществе в целом.

## 8.1. ОЧИСТКА ДНК ДЛЯ МЕТАГЕНОМНЫХ ИССЛЕДОВАНИЙ

Любое исследование с применением секвенирования начинается с экстракции ДНК и, в некоторых случаях, ее амплификации. При проведении метагеномных исследований в зависимости от задачи, стоящей перед исследователем, могут возникнуть следующие трудности на этапе очистки ДНК.

1. Для метагеномных исследований используют сильно неоднородный (гетерогенный) исходный биологический материал: почву, грунтовые воды, воздушные фильтры, мазки со слизистых человека, фекалии и т. п. Каждый из этих материалов содержит собственный набор примесей (ингибиторов), негативно влияющих на последующие этапы NGS, поэтому для каждого типа образцов разработаны специальные методы очистки ДНК и РНК [3–5]. Следует обратить внимание и на то обстоятельство, что для возможности сравнения полученных в разное время или в разных лабораториях результатов необходима стандартизация методов очистки нуклеиновых кислот.
2. При исследовании симбиотических микробных сообществ необходимо как-то избавляться от ДНК организма-хозяина, иначе она может составить значительную часть секвенированных последовательностей. Для этого применяют различные варианты фракционирования: фильтрацию, центрифугирование, проточную цитометрию и др. [2, 6]. Эти же подходы можно применять и в том случае, когда необходимо секвенировать лишь часть микробного сообщества, например только бактерий или только вирусы [7].
3. Если по каким-то причинам ДНК в образце мало, ее можно амплифицировать. Однако это может привести к искажению информации о представленности организмов в сообществе и числе прочтений для каждого отдельного ге-

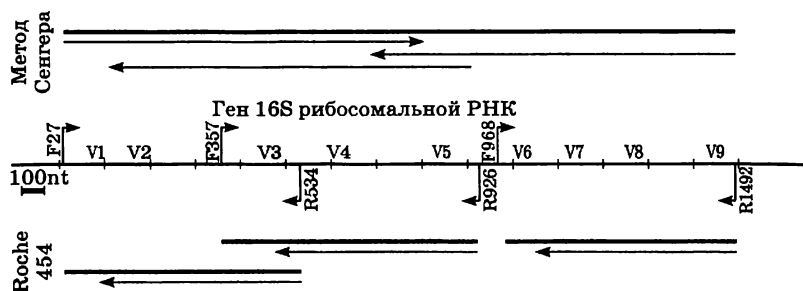
нома, а также образованию химерных прочтений [8]. При наличии альтернативных вариантов получения необходимого количества ДНК амплификации стоит избегать. Для сборки и анализа результатов секвенирования амплифицированных геномов существуют специальные программы, учитывающие искажения амплификации [9–10].

## 8.2. АНАЛИЗ МИКРОБНОГО СООБЩЕСТВА СЕКВЕНИРОВАНИЕМ АМПЛИКОНОВ

Для исследования видового состава прокариот в какой-либо среде обитания можно использовать метод, основанный на секвенировании высококонсервативных регионов генома. Чаще всего для этих целей выбирают ген 16S рибосомальной РНК (16S рРНК). У гена 16S рРНК (длиной около 1500 п. н.) есть константные и переменные участки (рис. 8.1). Константные участки позволяют создать «универсальные» праймеры для ПЦР, взаимодействующие с ДНК подавляющего большинства прокариот с примерно равной эффективностью. Методом ПЦР можно наработать расположенные между такими праймерами переменные участки гена 16S рРНК. Секвенирование одного или нескольких переменных регионов позволяет идентифицировать последовательности, принадлежащие разным видам.

До появления технологий высокопроизводительного секвенирования полученные продукты ПЦР клонировали в плазмидные векторы (в *Escherichia coli*), после чего каждый клон секвенировали с помощью метода Сенгера. Использование сенгеровского секвенирования позволяет получить последовательности высокого качества, почти полностью покрывающие ген 16S рРНК (что существенно для точной идентификации и различения близких видов) (см. рис. 8.1). Однако подход с клонированием достаточно дорог, а само клонирование вносит дополнительные искажения в конечный результат (см. разд. 10.3).

Появление технологий высокопроизводительного секвенирования с длиной прочтения в несколько сотен нуклеотидов позволило использовать NGS для анализа микробных сообществ по гену 16S рРНК без этапа клонирования, что позволяет ускорить и удешевить исследование. Первой для подобных исследований была применена технология 454 Life Sciences [11]. Показана эффективность применения при сек-



**Рис. 8.1.** Расположение консервативных регионов в гене 16S рРНК. Показано стандартное расположение праймеров и ампликонов для метагеномных исследований гена 16S рРНК методом Сенгера и технологией 454 Life Sciences. ОТ-ПЦП – обратная транскрипция–полимеразная цепная реакция

венировании гена 16S рРНК парно-концевых прочтений на платформах Illumina и Ion Torrent [12–13]. Использование NGS позволяет получать гораздо большее число прочтений, а следовательно, обнаруживать организмы с низкой представленностью в сообществе.

Недостатком методов NGS в данном применении является по-прежнему меньшая, чем в методе Сенгера, длина прочтения (секвенирование клонов методом Сенгера с двух сторон дает до 1400 п. н. единой последовательности). По короткому участку гена 16S рРНК (в 200–400 п. н.) зачастую можно определить бактерию лишь до рода. Кроме того, вероятность ошибок для технологий NGS гораздо выше, чем для сенгеровского секвенирования, таким образом, возникает проблема фильтрации ошибок секвенирования (см. разд. 4.4). Существуют работы, показывающие, что в ряде метагеномных проектов якобы большое видовое разнообразие сообщества в результате оказалось следствием ошибок секвенирования [14].

Крайне важным является подбор пары праймеров для амплификации регионов гена 16S рРНК. Даже в константных участках этого гена между последовательностями некоторых видов могут встречаться однонуклеотидные вариации, в результате чего эффективность амплификации ДНК разных видов может в значительной мере различаться (а для части видов может и вовсе не нарабатываться). Этой проблеме посвящен ряд исследований, предлагающих различные варианты пар

праймеров для получения ампликонов разной длины [15, 16]. В любом случае, при выборе пары праймеров для ПЦР стоит проверить, будут ли с ее помощью амплифицироваться последовательности гена 16S рРНК тех видов, которые ожидаются для данного микробиома, и какие виды при этом могут быть утеряны [11, 17].

Несомненным плюсом использования в качестве маркера гена 16S рРНК при анализе симбиотического микробиоценоза является исключение из анализируемого образца примеси ДНК организма-хозяина (за счет наличия этапа амплификации).

Помимо разной эффективности амплификации, при попытке количественно оценить состав микробиоты возникает еще одна сложность: копияность гена 16S рРНК может варьировать от 2 до 15 копий на гаплоидный геном [18]. Причем она может отличаться даже для микроорганизмов, относящихся к одному виду [19]. Оценка представленности различных видов с учетом числа копий гена 16S рРНК [20] или с помощью референсных генов, копияность которых постоянна [21], может улучшить точность получаемых результатов.

Для биоинформатической обработки данных высокопроизводительного секвенирования ампликонов используют два подхода [22].

1. Сравнение полученных последовательностей с различными базами данных. В частности, для гена 16S рРНК можно использовать базы данных: SILVA [23], RDB [24], NCBI<sup>1</sup>. Базы данных имеют встроенные инструменты поиска. Кроме того, некоторые базы предоставляют специальные инструменты для анализа данных высокопроизводительного секвенирования (например, SINA [25]). Минусом сравнения с базами данных являются проблемы при аннотации последовательностей, принадлежащих новым или плохо описанным видам некультивируемых бактерий.
2. Сравнение полученных последовательностей между собой, кластеризация и выявление «операционных таксономических единиц» (таксон, аналогичный виду и определяемый на основании степени сходства последовательности ДНК). Этот подход не опирается на уже известные последовательности и поэтому позволяет обнаруживать неизвестные виды микроорганизмов. Разработан ряд простых в

---

<sup>1</sup> <http://blast.ncbi.nlm.nih.gov/Blast.cgi>

применении инструментов, использующих этот принцип: QIIME [26], mothur [27], CLUSTOM [28].

Путем секвенирования ампликонов можно не только определить таксономический состав микробного сообщества, но и провести количественную оценку содержания того или иного микроорганизма в образце. В первом приближении можно принять долю прочтений ампликона (фрагмента гена 16S рРНК) с генома данного микроорганизма за долю микроорганизма в сообществе. Современные программы для анализа последовательностей позволяют строить удобные диаграммы процентного содержания бактерий в образце. Но, как уже было сказано, число генов рРНК у разных видов микроорганизмов варьирует, поэтому желательно делать на это поправку.

### 8.3. МЕТАГЕНОМНОЕ СЕКВЕНИРОВАНИЕ

Помимо использования для анализа микробного сообщества отдельных консервативных регионов генома (таких, как ген 16S рРНК или других филогенетических маркеров), можно напрямую выделить и секвенировать содержащуюся в образце ДНК. Важным преимуществом такого подхода является отсутствие этапа амплификации, вносящего значительные искажения в результаты.

Впервые термин «метагеном» был предложен Хандельсманом с соавт. в 1998 году [29]. Он обозначает суммарный набор генов, выявленных в исследуемом образце (как геном одного псевдоорганизма), используемый для воссоздания свойств исследуемого микробного сообщества.

Первый эксперимент по метагеномному секвенированию был проведен в 2004 году [30] и позволил обнаружить 148 ранее неизвестных видов бактерий и более 1,2 млн ранее неизвестных генов. Первые работы, так же как и в случае с 16S рРНК, проводили с помощью клонирования фрагментов ДНК в плазмидных векторах в клетках бактерии кишечной палочки и последующего секвенирования методом Сенгера. Длинные прочтения в значительной степени упрощали сборку и анализ данных, но из-за клонирования возникали искажения, поскольку некоторые последовательности плохо поддаются клонированию (например, если участок бактериальной ДНК несет ген, токсичный для *E. coli*, то такой фрагмент может

быть утерян в ходе клонирования). Эту проблему можно обойти, используя несколько разных организмов для клонирования [31].

Как уже было сказано, технологии высокопроизводительного секвенирования позволяют избежать клонирования и, соответственно, получить намного больше данных за меньшие деньги. Для метагеномного секвенирования используют практически все доступные на сегодняшний день технологии: 454 Life Sciences [32, 33], Illumina [34], Ion Torrent [35]. В некоторых случаях такое секвенирование позволяет получить полную последовательность генома для некультивируемых бактерий [36].

Существуют работы по сравнению технологий высокопроизводительного секвенирования применительно к метагеномам. Так, в работе Тяхта с соавт. было проведено метагеномное секвенирование одного и того же образца с помощью четырех технологий: 454 Life Sciences, SOLiD, Ion Torrent и Illumina. Авторы получили высокую сходимость результатов – больше 0,93 [37].

## **8.4. БИОИНФОРМАТИЧЕСКИЙ АНАЛИЗ ДАННЫХ МЕТАГЕНОМНОГО СЕКВЕНИРОВАНИЯ**

### **8.4.1. Сборка последовательностей**

Важным и одним из наиболее сложных этапов работы с метагеномными данными является их сборка и анализ. Использование высокопроизводительного секвенирования наряду с преимуществами (производительность и стоимость единицы информации) имеет по сравнению с сенгеровским секвенированием и существенный недостаток, а именно – короткие прочтения (для большинства платформ не превышающие 200 п. н.). При этом большая часть программ для сборки коротких прочтений в более длинные контиги рассчитана на данные, содержащие последовательности ДНК единственного организма. К настоящему времени на основе алгоритма де Брёйна разработан ряд программ, специально адаптированных для работы с метагеномными данными, например Meta-IDBA [38] и MetaVelvet [39]. Как и в случае других применений NGS, при метагеномных исследованиях более длинные прочтения упрощают дальнейший анализ, аннотацию генов, сравнение с уже известными геномами и поиск ранее неизвестных генов.

#### 8.4.2. Таксономический анализ метагеномных данных

Сборка прочтений в более длинные контиги необходима для дальнейшего таксономического анализа состава метагенома, называемого биннингом (от англ. binning – разбиение на интервалы), и для аннотации генов. В процессе биннинга устанавливают, к организму какой таксономической группы принадлежит та или иная последовательность. Как и для анализа данных секвенирования гена 16S рРНК, в метагеномных исследованиях существует два принципиально разных подхода к определению таксономической структуры микробного сообщества [2, 40].

1. Алгоритмы, основанные на сравнении полученных последовательностей с различными базами данных (в том числе с геномами референсных организмов). Обычно такой анализ проводят при помощи алгоритмов типа BLAST [41, 42]. Специализированные программы, разработанные на базе такого подхода: IMG/M [43], MG-RAST [44], MEGAN [45], CARMA [46], SOrt-ITEMS [47], MetaPhyler [48], RAiPhy [49]. Существуют специализированные инструменты для анализа данных секвенирования вирусных сообществ, например ProViDE [50]. Перечисленные программы позволяют проводить анализ сообщества по достаточно коротким прочтениям, например для MG-RAST хватает фрагментов длиной 75 п. н. Очевидно, что, как и в случае с геном 16S рРНК, использование баз данных плохо применимо для микробных сообществ, состоящих в основном из некультивируемых и плохо изученных видов (например, сообществ из почв и водоемов).
2. Алгоритмы, основанные на анализе последовательностей по определенным показателям, таким как степень покрытия, GC-состав, частота встречаемости кодонов и  $k$ -меров и др. К программам, использующим подобные алгоритмы анализа, относятся: TETRA [51], Phylopythia [52], S-GSOM [53], PCANIER [54] и TACOА [55]. Перечисленные инструменты позволяют классифицировать последовательности длиной от 800 п. н., однако алгоритмы, анализирующие нуклеотидный состав, лучше работают на более длинных фрагментах [52].

Помимо программ, использующих один из двух описанных выше алгоритмов, существуют гибридные варианты,

сочетающие преимущества обоих подходов, например SPHINX [54], PhymmBL, MetaCluster [56].

### 8.4.3. Аннотация метагеномных данных

Аннотация – поиск кодирующих участков генома и описание их функций. Если после сборки прочтений были получены достаточно большие контиги (длиной 30 т. п. н. и более), возможна аннотация традиционными способами, такими как RAST [57] или IMG [58]. Если же речь идет о коротких контигах или собственно прочтениях, то обычные инструменты могут оказаться неэффективными.

Как и в случае с аннотацией отдельных геномов, процесс поиска генов в метагеномных данных можно разделить на два этапа: поиск характерных участков генома и предсказание функций методами сравнительной геномики (путем сравнения с уже известными генами). Для таких задач существует ряд специализированных инструментов, учитывающих специфику метагеномных данных: FragGeneScan [59], MetaGeneMark [60], Metagene [61], Orphelia [62].

Несмотря на быстрое развитие вычислительных инструментов, в настоящее время поддаются аннотации лишь 20–50% последовательностей в метагеномных данных [2]. Алгоритмы для аннотации оставшейся части данных пока находятся в разработке.

## 8.5. КОМБИНИРОВАННЫЙ АЛГОРИТМ АНАЛИЗА ТАКСОНОМИЧЕСКОГО СОСТАВА СООБЩЕСТВА

Каждый из перечисленных выше способов анализа таксономического состава сообщества (по 16S рРНК и по метагеному) имеет свои достоинства и недостатки. Праймеры по-разному амплифицируют 16S рРНК разных микроорганизмов и могут существенно исказить реальное соотношение бактерий в образце. В то же время, метагеном как метод гораздо менее чувствителен к минорным представителям сообщества и контиги получаются только для доминирующих видов.

В этой связи оптимальным представляется комбинированный алгоритм, сочетающий в себе как чтение коротких переменных участков, так и метагеномное секвенирование.

Последовательность шагов в таком алгоритме может быть следующей:

- 1) секвенирование короткого вариабельного участка (чаще всего это фрагмент или несколько фрагментов 16S рРНК);
- 2) секвенирование метагенома;
- 3) анализ прочтений 16S рРНК в метагеноме и их классификация;
- 4) удаление из результатов всех не классифицируемых данных;
- 5) анализ сообщества по кратности прочтения контигов с 16S рРНК (чем больше прочтений 16S рРНК, тем больше данного микроба содержится в сообществе);
- 6) сопоставление результатов двух подходов.

В качестве дополнительных информативных шагов можно распределить контиги между микроорганизмами (например, по покрытию или по частотам встречаемости тетра-нуклеотидов).

## 8.6. СРАВНЕНИЕ МЕТАГЕНОМОВ МЕЖДУ СОБОЙ

Некоторые биологические задачи предполагают исследование схожих, но территориально разделенных сообществ микроорганизмов. Например, сравнение образцов из разных буровых скважин или микробиоты кишечника разных людей и т. п.

Кроме сопоставления видового состава сообщества и процентного соотношения микроорганизмов в ряде задач имеет смысл исследовать генетическое разнообразие отдельных видов бактерий в сообществе, пользуясь нуклеотидным полиморфизмом. Например, для микробиоценозов человека (кишечного или урогенитального) можно обнаружить, что в разных странах или при разных физиологических состояниях человека бактерии одного вида в сообществе могут быть очень далеки филогенетически [37].

## 8.7. МЕТАТРАНСКРИПТОМ

Биоинформатическое обнаружение гена в последовательности генома далеко не всегда означает, что он работоспособен. Для понимания функциональных возможностей сообщества в дополнение к метагеномным данным может быть полезно

определение метатранскриптома (секвенирование суммарных мРНК сообщества). Исследования метатранскриптомов осложнены нестабильностью молекул РНК и, как следствие, необходимостью быстрой и эффективной фиксации образца. Особенно это актуально при исследовании образцов почв (из-за наличия РНКаз и адсорбции молекул РНК на частицы почвы). Для экстракции РНК из столь сложных образцов разрабатываются специальные методы [63].

Кроме того, поскольку 95% от всех РНК в клетке составляют рибосомальная и транспортная РНК, «забивающие» библиотеку, от них тоже приходится избавляться [63]. Чтобы понять, какие гены экспрессируются и в какой степени, необходимо выравнивание полученных данных метатранскриптома на метагеном, аннотация и нормализация. Возможно и выравнивание на референсные геномы из баз данных, но более информативным будет выравнивание и аннотация относительно метагеномных данных для того же образца.

\* \* \*

В заключение главы отметим, что «мета»-исследования сегодня являются одной из быстро развивающихся областей биологии, в том числе благодаря появлению методов NGS. Метагеномика вместе с метатранскриптомикой и метапротеомикой позволяют понять состав и функционирование микробных сообществ (в том числе и симбиотических человеку), что, в свою очередь, может помочь предсказывать их поведение в зависимости от условий среды обитания<sup>1</sup>.

В 2009 году стартовал проект «Микробиом человека» [64], ставящий своей целью охарактеризовать обитающие на теле человека микробные сообщества для здоровых людей и пациентов различными заболеваниями. Первые результаты подобных исследований показывают, что некоторые изменения в составе микробиоценозов человека могут серьезно угрожать его здоровью [65, 66].

---

<sup>1</sup> Например, <http://metagenom.ru/>

## СПИСОК ЛИТЕРАТУРЫ

1. *Sleator R.D., Shortall C., Hill C.* Metagenomics // Letters in applied microbiology, 2008, 47 (5): 361–366.
2. *Thomas T., Gilbert J., Meyer F.* Metagenomics – a guide from sampling to data analysis // Microbial informatics and experimentation, 2012, 2 (1): 3.
3. *Simon C., Herath J., Rockstroh S., Daniel R.* Rapid identification of genes encoding DNA polymerases by function-based screening of metagenomic libraries derived from glacial ice // Applied and environmental microbiology, 2009, 75 (9): 2964–2968.
4. *Abulencia C.B. et al.* Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments // Environmental Whole-Genome Amplification To Access Microbial Populations in Contaminated Sediments // Appl Environ Microbiol., 2006, 72: 3291–3301.
5. *Hardeman F., Sjöling S.* Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment // FEMS microbiology ecology, 2007, 59 (2): 524–534.
6. *Burke C., Kjelleberg S., Thomas T.* Selective extraction of bacterial DNA from the surfaces of macroalgae // Applied and environmental microbiology, 2009, 75 (1): 252–256.
7. *Angly F.E. et al.* The marine viromes of four oceanic regions // PLoS biology, 2006, 4 (11): e368.
8. *Abbai N.S., Govender A., Shaik R., Pillay B.* Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater // Molecular biotechnology, 2012, 50 (1): 39–48.
9. *Bankevich A. et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // J Comput Biol., 2012, 19 (5): 455–477.
10. *Nurk S. et al.* Assembling single-cell genomes and mini-metagenomes from chimeric MDA products // J Comput Biol., 2013, 20 (10): 714–737.
11. *Conlan S., Kong H.H., Segre J.A.* Species-level analysis of DNA sequence data from the NIH Human Microbiome Project // PLoS one, 2012, 7 (10): e47075.
12. *Mizrahi-Man O., Davenport E.R., Gilad Y.* Taxonomic classification of bacterial 16S rRNA genes using short sequencing reads: evaluation of effective study designs // PLoS one, 2013, 8 (1): e53608.

13. *Jünemann S. et al.* Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing // PLoS one, 2012, 7 (8): e41606.
14. *Kunin V., Engelbrektson A., Ochman H., Hugenholtz P.* Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates // Environmental microbiology, 2010, 12 (1): 118–123.
15. *Nossa C.W.* Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome // World Journal of Gastroenterology, 2010, 16 (33): 4135.
16. *Klindworth A. et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies // Nucleic acids research, 2013, 41 (1): e1.
17. *Soergel D., Dey N., Knight R., Brenner S.E.* Selection of primers for optimal taxonomic classification of environmental 16S rRNA gene sequences // The ISME journal, 2012, 6 (7): 1440–1444.
18. *Crosby L.D., Criddle C.S.* Understanding bias in microbial community analysis techniques due to rrn operon copy number heterogeneity // BioTechniques, 2003, 34 (4): 790–794, 796, 798.
19. *Bodilis J., Nsigue-Meilo S., Besaury L., Quillet L.* Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in *Pseudomonas* // PLoS one, 2012, 7 (4): e35647.
20. *Kembel S.W., Wu M., Eisen J.A., Green J.L.* Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance // PLoS computational biology, 2012, 8 (10): e1002743.
21. *Case R.J. et al.* Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies // Applied and environmental microbiology, 2007, 73 (1): 278–288.
22. *Gevers D., Pop M., Schloss P.D., Huttenhower C.* Bioinformatics for the Human Microbiome Project // PLoS computational biology, 2012, 8 (11): e1002779.
23. *Quast C. et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools // Nucleic acids research, 2013, 41 (Database issue): D590–596.
24. *Cole J.R. et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis // Nucleic acids research, 2013, 42: 633–642.
25. *Pruesse E., Peplies J., Glöckner F.O.* SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes // Bioinformatics, 2012, 28 (14): 1823–1829.

26. *Caporaso J.G. et al.* Correspondence QIIME allows analysis of high-throughput community sequencing data Intensity normalization improves color calling in SOLiD sequencing // Nature Publishing Group, 2010, 7 (5): 335–336.
27. *Schloss P.D. et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities // Applied and environmental microbiology, 2009, 75 (23): 7537–7541.
28. *Hwang K. et al.* CLUSTOM: a novel method for clustering 16S rRNA next generation sequences by overlap minimization // PLoS one, 2013, 8 (5): e62623.
29. *Handelsman J. et al.* Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products // Chemistry & biology, 1998, 5 (10): R245–249.
30. *Venter J. C. et al.* Environmental genome shotgun sequencing of the Sargasso Sea // Science, 2004, 304 (5667): 66–74.
31. *Wooley J.C., Godzik A., Friedberg I.* A primer on metagenomics // PLoS computational biology, 2010, 6 (2): e1000667.
32. *Ghai R. et al.* Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing // The ISME journal, 2010, 4 (9): 1154–1166.
33. *Rademacher A. et al.* Characterization of microbial biofilms in a thermophilic biogas system by high-throughput metagenome sequencing // FEMS microbiology ecology, 2012, 79 (3): 785–799.
34. *Albertsen M. et al.* A metagenome of a full-scale microbial community carrying out enhanced biological phosphorus removal // The ISME journal, 2012, 6 (6): 1094–1106.
35. *Albertsen M. et al.* Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes // Nature biotechnology, 2013, 31 (6): 533–538.
36. *Whiteley A.S. et al.* Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform // Journal of microbiological methods, 2012, 91 (1): 80–88.
37. *Tyakht A.V. et al.* Human gut microbiota community structures in urban and rural populations in Russia // Nat Commun, 2013, 4: 2469.
38. *Peng Y., Leung H.C.M., Yiu S.M., Chin F.Y.L.* Meta-IDBA: A de Novo assembler for metagenomic data // Bioinformatics, 2011, 27 (13): i94–101.
39. *Namiki T., Hachiya T., Tanaka H., Sakakibara Y.* MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly

- from short sequence reads // *Nucleic acids research*, 2012, 40 (20): e155.
40. *Simon C., Daniel R.* Metagenomic analyses: past and future trends // *Applied and environmental microbiology*, 2011, 77 (4): 1153–1161.
  41. *Altschul S., Gish W., Miller W.* Basic local alignment search tool // *J. Mol. Biol.*, 1990, 215: 403–410.
  42. *Camacho C. et al.* BLAST+: architecture and applications // *BMC bioinformatics*, 2009, 10: 421.
  43. *Markowitz V.M. et al.* IMG/M 4 version of the integrated metagenome comparative analysis system // *Nucleic acids research*, 2013, 42 (5): 1–6.
  44. *Meyer F. et al.* The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes // *BMC bioinformatics*, 2008, 9: 386.
  45. *Huson D.H., Auch A.F., Qi J., Schuster S.C.* MEGAN analysis of metagenomic data // *Genome Research*, 2007, 17 (3): 377–386.
  46. *Gerlach W., Stoye J.* Taxonomic classification of metagenomic shotgun sequences with CARMA3 // *Nucleic acids research*, 2011, 39 (14): e91.
  47. *Monzoorul H.M., Ghosh T.S., Komanduri D., Mande S.S.* SOrT-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences // *Bioinformatics*, 2009, 25 (14): 1722–1730.
  48. *Liu B., Gibbons T., Ghodsi M., Treangen T., Pop M.* Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences // *BMC genomics*, 12 (2): S4.
  49. *Nalbantoglu O.U., Way S.F., Hinrichs S.H., Sayood K.* RAIPhy: phylogenetic classification of metagenomics samples using iterative refinement of relative abundance index profiles // *BMC bioinformatics*, 2011, 12 (1): 41.
  50. *Ghosh T.S., Mohammed M.H., Komanduri D., Mande S.S.* ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples // *Bioinformation*, 2011, 6 (2): 91–94.
  51. *Teeling H. et al.* TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences // *BMC bioinformatics*, 2004, 5: 163.
  52. *McHardy A.C. et al.* Accurate phylogenetic classification of variable-length DNA fragments // *Nature Methods*, 2007, 4 (1): 63–72.
  53. *Chan C.-K. K., Hsu A.L., Halgamuge S.K., Tang S.-L.* Binning sequences using very sparse labels within a metagenome // *BMC bioinformatics*, 2008, 9: 215.

54. *Zheng H., Wu H.* Short Prokaryotic DNA Fragment Binning Using a Hierarchical Classifier Based on Linear Discriminant Analysis and Principal Component Analysis // *Journal of Bioinformatics and Computational Biology*, 2010, 08 (6): 995–1011.
55. *Diaz N.N. et al.* TACO: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach // *BMC bioinformatics*, 2009, 10: 56.
56. *Wang Y., Leung H.C.M., Yiu S.M., Chin F.Y.L.* MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample // *Bioinformatics*, 2012, 28 (18): i356–i362.
57. *Aziz R.K. et al.* The RAST Server: rapid annotations using subsystems technology // *BMC genomics*, 2008, 9: 75.
58. *Markowitz V.M. et al.* IMG ER: a system for microbial genome annotation expert review and curation // *Bioinformatics*, 2009, 25 (17): 2271–2278.
59. *Rho M., Tang H., Ye Y.* FragGeneScan: predicting genes in short and error-prone reads // *Nucleic acids research*, 2010, 38 (20): e191.
60. *Zhu W., Lomsadze A., Borodovsky M.* *Ab initio* gene identification in metagenomic sequences // *Nucleic acids research*, 2010, 38 (12): e132.
61. *Noguchi H., Park J., Takagi T.* MetaGene: prokaryotic gene finding from environmental genome shotgun sequences // *Nucleic acids research*, 2006, 34 (19): 5623–5630.
62. *Hoff K.J., Lingner T., Meinicke P., Tech M.* Orphelia: predicting genes in metagenomic sequencing reads // *Nucleic acids research*, 2009, 37 (Web Server issue): W101–105.
63. *Carvalhais L.C., Dennis P.G., Tyson G.W., Schenk P.M.* Application of metatranscriptomics to soil environments // *Journal of microbiological methods*, 2012, 91 (2): 246–51.
64. *Peterson J. et al.* The NIH Human Microbiome Project // *Genome research*, 2009, 19 (12): 2317–2323.
65. *Wang J. et al.* Metagenomic sequencing reveals microbiota and its functional potential associated with periodontal disease // *Scientific reports*, 2013, 3: 1843.
66. *Cox M.J., Cookson W.O.C.M., Moffatt M.F.* Sequencing the human microbiome in health and disease // *Human molecular genetics*, 2013, 22 (R1): R88–94.

# СЕКВЕНИРОВАНИЕ ГЕНОМОВ ЭУКАРИОТ

Прокариотические и эукариотические геномы значительно отличаются друг от друга по своей структуре. У прокариот геном по большей части представлен уникальными кодирующими последовательностями ДНК, в то время как у эукариот лишь несколько процентов ДНК несут гены, а подавляющая часть генома состоит из повторяющихся элементов различной природы (микросателлитов, мобильных элементов, остатков вирусных геномов и т. п.). Особенно много повторяющихся последовательностей в геномах растений и млекопитающих. В большинстве случаев исследователя интересуют именно кодирующие участки эукариотического генома, в то время как повторы лишь мешают (повышают стоимость проекта). Для удешевления определения именно кодирующих регионов разработан ряд подходов, позволяющих перед секвенированием обогатить образец кодирующими последовательностями (см. гл. 11).

Сразу отметим, что, говоря об эукариотическом геноме, обычно подразумевают лишь ядерную ДНК. Генетический материал митохондрий и хлоропластов не принято считать частью генома организма, такие проекты обычно называют «митохондриальный геном» и «пластом» соответственно.

В данной главе рассматриваются особенности NGS-проектов геномов эукариот.

## 9.1. ОБЩИЕ АСПЕКТЫ СЕКВЕНИРОВАНИЯ СЛОЖНЫХ ГЕНОМОВ

Секвенирование генома человека и других организмов в 1990-х и 2000-х годах осуществлялось впервые (*de novo*), из-за чего ученые сталкивались со сложной задачей по сборке коротких прочтений в единую последовательность [1]. Осуществление такой сборки тем сложнее, чем больше размер

генома исследуемого организма и чем сложнее его организация. Размеры геномов эукариот варьируют от десятков миллионов пар нуклеотидов (например, 29 млн п. н. у паразитической микроспоридии *Encephalitozoon cuniculi*) до сотен миллиардов пар нуклеотидов (150 млрд п. н. у растения *Paris japonica*). Столь существенные различия в размерах геномов возникают главным образом по причине увеличения так называемых некодирующих последовательностей (в основном за счет амплификации разного рода мобильных генетических элементов и сателлитных повторов), что приводит к накоплению в геномах большого количества повторяющихся последовательностей. Таким образом, различия в размерах эукариотических геномов между видами зачастую определяются количеством повторов [2].

Трудность сборки генома из секвенированных фрагментов впервые была продемонстрирована в середине 1980-х годов [3, 4]. В зависимости от длины прочтений и длины повторов в исследуемой последовательности задача по сборке генома может быть тривиальной (если все повторы короче прочтений), вычислительно неразрешимой (когда требуется перебрать такое число способов перестановки прочтений, которое не обеспечивает даже использование суперкомпьютера) и неразрешимой в принципе (если информации, полученной из прочтений, недостаточно, чтобы выбрать правильный контиг из равнозначных вариантов) (рис. 9.1). Тривиальной бывает сборка только небольшого (вирусного) генома. В случае секвенирования геномов большего размера ассемблеры – программы для сборки геномов – не могут собрать геном полностью, давая на выходе несколько отдельных фрагментов или ошибочно собранный геном (см. разд. 4.2).



Рис. 9.1. Особенности сборки геномов эукариот *de novo*



Рис. 9.2. Использование парно-концевых прочтений для картирования в повторяющихся регионах

Одновременно с разработкой биоинформатических подходов развивались и методы секвенирования. Сборка становится более точной с увеличением длины одного прочтения и при добавлении информации о взаимном расположении нескольких прочтений друг относительно друга (рис. 9.2).

Кроме того, ввиду необходимости в значительном перекрытии прочтений при построении контигов возникает необходимость в многократном прочтении одного и того же участка генома. Однозначного мнения о достаточном покрытии при секвенировании генома *de novo* пока нет. Однако очевидно, что с увеличением покрытия (количества прочтений для одного и того же участка генома) увеличивается точность сборки, снижение числа контигов и повышение доли прочитанного генома. В публикациях последних лет в зависимости от исследуемого организма и секвенсной платформы авторы используют различные покрытия – от 6-кратного (для «чернового» прочтения) до более чем 100-кратного (для создания точной референсной последовательности генома нового организма) [5].

## 9.2. СЕКВЕНИРОВАНИЕ ЭУКАРИОТИЧЕСКИХ ГЕНОМОВ *DE NOVO*

Метод Сенгера, позволивший определить первые геномы эукариот, обладает достаточно большой длиной прочтения (около 800 п. н.) и низкой частотой ошибок, однако из-за вы-

сокой стоимости определения одного нуклеотида и трудоемкости в конце 2000-х годов метод перестали использовать для секвенирования геномов [6].

В настоящее время исследователю приходится выбирать между максимальной длиной прочтений и наличием информации об их взаимном расположении, поскольку технология, позволяющая совмещать оба параметра, пока не разработана.

Технология секвенирования 454 Life Sciences сегодня позволяет осуществлять прочтения длиной до 1200 п. н. (что соответствует лучшим показателям метода Сенгера) с возможностью получить до нескольких миллионов прочтений в течение суток. Этот метод секвенирования второго поколения появился одним из первых и активно применялся во второй половине 2000-х годов для определения геномов эукариот [7]. Однако с появлением более дешевых методов секвенирования и возможности получать информацию о взаимном расположении прочтений технология 454 Life Sciences от Roche теряет свои позиции.

Еще более длинные прочтения позволяют получить секвенаторы третьего поколения, такие как PacBio (прочтения длиной до 20 000 п. н.). Однако такие приборы только появились на рынке, и стоимость определения нуклеотида и новизна технологии пока сдерживают их распространение.

К наиболее популярным технологиям, дающим прочтение фрагмента с двух сторон, относятся технология SOLiD, которая позволяет прочесть несколько десятков нуклеотидов с двух концов для нескольких миллиардов фрагментов и технология Illumina, способная определить последовательности длиной в несколько сотен нуклеотидов для чуть большего числа фрагментов (см. гл. 3) [8]. Видно, что «двусторонние» технологии (в литературе их часто называют парно-концевыми) дают гораздо более короткие прочтения (в сравнении с 454 Life Sciences), зато дают информацию о расстоянии между прочтениями (и расстояние это исследователь может регулировать на уровне создания библиотеки фрагментов для NGS, см. гл. 2).

Выбор средней длины фрагментов при использовании «двусторонней» технологии NGS является важным нюансом секвенирования *de novo*. Слишком короткие фрагменты дают мало информации для ассемблера, а слишком длинные снижают качество и объем получаемых с одного запуска секвенатора данных. Большинство авторов при секвенировании больших

эукариотических геномов *de novo* для двустороннего прочтения рекомендуют использовать инвертированные библиотеки с длиной фрагментов 3–10 т. п. н.

Таким образом, при запуске проекта секвенирования эукариотического генома *de novo* можно рекомендовать исследователю следующий алгоритм: 1) использовать две разных NGS-платформы (с длинными односторонними прочтениями и короткими двусторонними), 2) для двусторонних прочтений сделать две (или несколько) библиотек фрагментов ДНК с разной длиной фрагментов (например, в 3 т. п. н. и 10 т. п. н.) и секвенировать их независимо, 3) после сборки контигов попытаться закрывать пробелы методом ПЦР, оптимизированной для наработки длинных фрагментов (long-range PCR), с секвенированием получаемых фрагментов методом Сенгера (рис. 9.3).

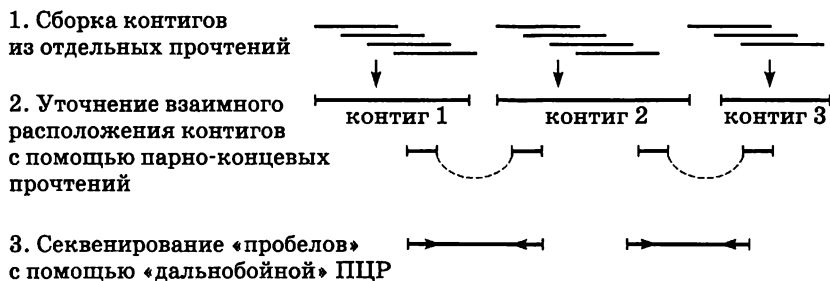


Рис. 9.3. *De novo*-секвенирование геномов эукариот

Заметим, что использование нескольких технологий секвенирования в одном проекте вдобавок позволяет повысить надежность результатов за счет учета типовых ошибок приборо.

### 9.3. ПОВТОРНОЕ СЕКВЕНИРОВАНИЕ (РЕСЕКВЕНИРОВАНИЕ)

Полное или частичное секвенирование генома организма, при условии что геном этого организма (или близкого родственника) уже есть в базах данных (так называемый

референсный геном), называют ресеквенированием. Проект по ресеквенированию может преследовать различные цели: изучение однонуклеотидного полиморфизма (SNP), небольших вставок или делеций (Indel), больших структурных вариаций (SVs), поиск новых участков (отсутствующих в референсном геноме), гаплотипирование и пр. [9].

Теоретически полиморфные позиции, делеции и вставки можно обнаружить даже в результате одно- или двукратного прочтения с последующей коррекцией ошибок секвенирования. Однако референсный геном также может содержать ошибки. Например, в настоящее время в референсном геноме человека содержится примерно 0,01% ошибок, поэтому однонуклеотидное несоответствие исследуемого генома референсному не всегда является следствием проявления SNP. Наличие повторов или SNP в последовательностях, фланкирующих делецию, также затрудняет ее обнаружение.

Сравнительно небольшие делеции (вставки) могут быть достоверно обнаружены при прочтении «насквозь» (охватывая смежную область, содержащую разрыв). Практика показывает, что большинство делеций (вставок) в уникальных последовательностях генома человека обнаруживаются уже при длине прочтения в 75–100 п. н.

Делеции (вставки), превышающие длину прочтения, можно искать двумя способами: 1) анализируя среднее покрытие (read depth), при котором частота прочтений определенного сегмента ДНК отражает его копийность; 2) картируя парно-концевые прочтения (в основном для инвертированных библиотек) и сравнивая расстояние между прочтениями в исследуемом и референсном геноме (этот метод подходит также для поиска инверсий).

Ни один из существующих методов не является оптимальным для поиска больших структурных вариаций. Многие из SVs находятся в повторяющихся последовательностях, что не позволяет точно расположить конкретное прочтение на геноме. Кроме того, SVs часто являются результатом совокупности событий, произошедших близко друг от друга, что не позволяет обнаружить их при использовании коротких прочтений NGS. Например, для обнаружения транспозонов требуются специальные алгоритмы. Тем не менее обилие повторяющихся последовательностей в геномах приводит к тому, что большинство SVs обнаружить не удастся [10].

Методы поиска новых последовательностей также недостаточно надежны, особенно при малой длине прочтения. Теоретически, использование метода парно-концевых прочтений в случае, когда один край фрагмента лежит в известной области, а второй – во вновь обнаруженной, может позволить обнаружить такие участки и затем, собрав их в контиги, встроить в референсный геном. Однако подобные алгоритмы плохо автоматизированы и требуют ручной доработки.

Требования к получаемым данным со стороны алгоритмической обработки при ресеквенировании менее строгие, чем при сборке *de novo*. Однако принципы выбора технологий аналогичны – отдается предпочтение парно-концевым прочтениям с длиной 50–250 п. н. (с небольшим размером фрагментов геномной библиотеки) Среднее покрытие, как правило, находится в диапазоне  $\times 12$ –50, что позволяет добиваться точности определения SNP на уровне 98–99%. Дальнейшее увеличение покрытия не ведет к заметному росту полноты секвенирования и точности определения SNP (рис. 9.4).

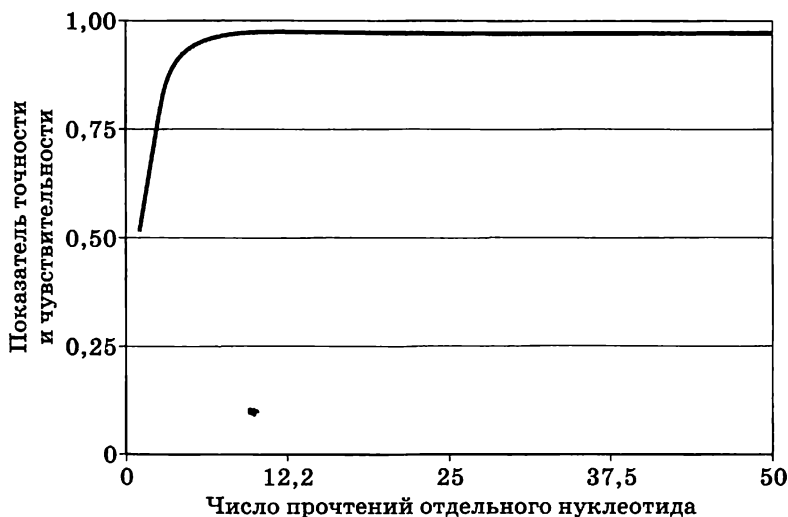


Рис. 9.4. Зависимость точности определения однонуклеотидных полиморфизмов от покрытия

#### 9.4. ФАЗИРОВАНИЕ ПРИ РЕСЕКВЕНИРОВАНИИ ДИПЛОИДНЫХ ГЕНОМОВ

Дополнительную сложность при анализе данных ресеквенирования эукариотического генома создает его диплоидность. Наличие двух наборов хромосом приводит к тому, что могут быть одновременно получены прочтения, как соответствующие референсному геному, так и отличающиеся от него. Более того, диплоидность требует определения, в какой из хромосом локализован данный вариант (аллель). Процесс определения принадлежности аллеля одной из хромосом называется фазированием (phasing).

Если ресеквенируют геном индивида, геномы обоих родителей которого были секвенированы ранее, фазирование становится сравнительно простой задачей, которую можно решить сравнением генотипов. Когда же, как чаще всего бывает на практике, геномы родителей недоступны, применяют статистические методы, предполагающие, что гаплотипы объединены в генотипы случайным образом. Основная идея такого анализа состоит в том, что, если проанализировать геномы индивидов, гомозиготных по большому количеству локусов, можно рассчитать частоты встречаемости гаплотипов. Статистические методы могут быть основаны на методе максимального правдоподобия, парсимонии, комбинаторике или использовании априорных распределений из теории коалесценции. Последний подход лег в основу работы программы PHASE [11]. Недавно разработан подход, предлагающий рассматривать фазирование как способ восстановления последовательности в случае потери части данных при анализе диплоидного генома [12].

Экспериментальное фазирование очень трудозатратно. Для точного фазирования при отсутствии информации о близких родственниках необходимо использовать специальные методы уже на этапе секвенирования. Некоторые методы позволяют полностью фазировать хромосомы, в то время как другие позволяют разделить лишь отдельные участки. Во втором случае требуется повторная сборка таких фрагментов *de novo*, чтобы получить гаплотип. В такой ситуации принято говорить о «проблеме реконструкции гаплотипа отдельного индивидуума».

Подходы к экспериментальному фазированию можно разделить на полногеномные (когда секвенирование проводят

таким образом, чтобы получить как можно больше информации о гаплотипе) и основанные на выделении информации о гаплотипе из прочтений после обычного секвенирования.

Референсный геном человека, установленный International Human Genome Sequencing Consortium (проект «Геном человека»), был получен с использованием библиотеки, содержащей длинные вставки с последующим секвенированием методом дробовика. Вставка в каждом клоне библиотеки представляла отдельный гаплотип, что позволяло получать гаплоидные последовательности. Собственно такой подход можно было бы использовать при секвенировании, чтобы получить максимум информации для последующего фазирования, однако это очень дорого и трудозатратно. При замене метода дробовика современным NGS стоимость полногеномного фазирования снизилась до 150 тыс. руб за образец. В результате были получены фрагменты гаплотипов, которые было необходимо объединять в полный геном. На этом этапе могут возникать ошибки из-за наличия гомозиготных участков, превышающих длину вставки в одном клоне библиотеки. Такой модифицированный подход не позволил полностью разделить все однонуклеотидные маркеры (SNP), но тем не менее от 94 до 99% маркеров было фазировано (рис. 9.5) [13].

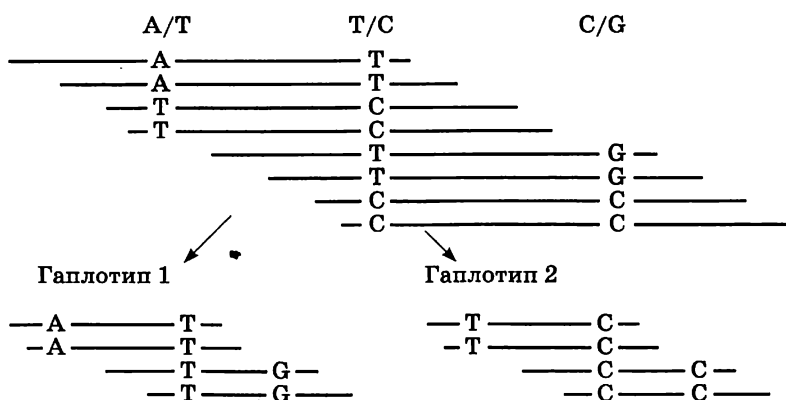


Рис. 9.5. Фазирование при секвенировании

Избежать ошибок фазирования можно, разделив хромосомы до начала секвенирования. Сейчас появляются методы, позволяющие при помощи микрофлюидики разделить метафазные хромосомы отдельной клетки, которые затем можно секвенировать [14].

Прямое секвенирование также дает некоторую информацию для фазирования. При секвенировании по Сенгеру, 454 Life Sciences и PacBio получаются довольно длинные (более 700 п. н.) прочтения, дающие достаточно информации для разделения гаплотипов. Другие технологии секвенирования (например, Illumina и Ion Torrent) не позволяют получить прочтений такой длины, однако при использовании парно-концевых прочтений с последующим применением специальных компьютерных программ (Haplotype Improver и «read-backed phasing», алгоритм программного пакета Genome Analysis Tool Kit) они позволяют достигать аналогичных показателей при более низкой стоимости эксперимента [15].

## 9.5. СЕКВЕНИРОВАНИЕ ГЕНОМА ОТДЕЛЬНОЙ КЛЕТКИ

С развитием технологий не только растет производительность секвенаторов, но и наблюдаются значительные улучшения в процессе пробоподготовки – требуется все меньше ДНК для проведения анализа. В последнее время стало возможным секвенирование генома отдельной клетки. Это открывает перед научным сообществом недоступные ранее возможности: секвенирование геномов некультивируемых микроорганизмов, анализ циркулирующих в кровотоке опухолевых клеток или клеток плода, генетическое исследование клеток на ранних этапах эмбрионального развития, изучение транскрипционного шума, гетерогенности опухолевых клеток, микроэволюции и т. д.

Секвенирование отдельных клеток применимо как к одноклеточным, так и многоклеточным организмам, в данном разделе будет уделено особое внимание использованию этого подхода для изучения геномов млекопитающих.

Любой проект по изучению клеточного генома начинается с изоляции индивидуальных клеток (поскольку ткани, как правило, состоят из десятков и сотен клеток различных

типов). Все способы получения отдельных клеток можно разделить на две большие группы: рандомизированные (случайные) и таргетные (направленные). Протопоподготовка, подразумевающая рандомизацию, лучше отражает состав исследуемой ткани, а таргетный подход позволяет выделить и изучить редкие типы клеток.

Получение клеток из плотных тканей включает два основных шага: взятие фрагмента ткани (биопсия) с последующим разделением его на отдельные клетки (обычно для этого используют ферменты) и помещение клеток в отдельные реакционные ячейки для последующего лизиса и дальнейшей работы.

Клетки можно разделить, используя микроманипуляторы (стеклянные капилляры) или же проведя серию разбавлений. Методы, основанные на микроманипуляции, дешевы, но низкопроизводительны, кроме того, при идентификации клеток под микроскопом крайне важен опыт исследователя. Использование клеточных сортеров (метод FACS, fluorescence-activated cell sorting – сортировка флуоресцентно-меченых клеток) – позволяет реализовать как рандомизированный, так и таргетный подходы, исключает человеческий фактор и значительно повышает выход сортированных клеток. Однако для разделения на сортере требуется значительно большее количество исходного материала, что особенно критично для небольших клеточных популяций.

Если два вышеописанных метода подразумевали обязательное использование клеточных суспензий, то лазерная захватывающая микродиссекция (laser-capture microdissection – LCM) позволяет изолировать клетки непосредственно из образцов тканей. Образец ткани, покрытый специальной пленкой, анализируют под микроскопом, затем в результате воздействия лазера выбранные клетки прилипают к пленке и извлекаются из образца. Сама процедура извлечения клеток накладывает ряд ограничений на использование данного метода. Он подразумевает высокую квалификацию исследователя, часть цитоплазмы может быть потеряна при извлечении или же, напротив, частично может захватываться цитоплазма соседних клеток, ядро также может быть частично повреждено.

С развитием микрофлюидики стали появляться устройства для выделения отдельных клеток и даже отдельных хро-

мосом. Вероятно, за такими устройствами будущее пробоподготовки для изучения геномов индивидуальных клеток.

После получения отдельной клетки необходимо амплифицировать ее геномную ДНК, так как единственной копии, содержащейся в клетке, недостаточно для проведения анализа. Впервые метод полногеномной амплификации (WGA) был предложен в 1992 году для случаев, когда количество ДНК в образце мало, но требуется провести множество анализов (например, при сравнительной геномной гибридизации). Методы WGA можно разделить на методы с использованием ПЦР и методы, основанные на амплификации со множественным замещением цепи (multiple displacement amplification – MDA). Существует два основных ПЦР-метода WGA: degenerate oligonucleotide PCR (DOP-PCR) и primer extension preamplification (PEP). Основное различие этих методов состоит в том, что в PEP используют случайные праймеры и низкую температуру отжига, в то время как в DOP-PCR – частично вырожденные праймеры (CGACTCGAGNNNNNNATGTGG) и повышенную температуру отжига. Использование *Taq*-полимеразы ограничивает длину фрагментов в 3 т. п. н. и приводит к возникновению ошибок. Кроме того, оба этих метода могут приводить к неполной или неравномерной амплификации генома вследствие предпочтительного отжига праймеров на некоторых участках генома.

Амплификация со множественным замещением цепи лишена недостатков ПЦР-подходов. В этом методе октамерные олигонуклеотиды связываются с денатурированной ДНК, после чего синтез осуществляется полимеразой *Phi29* при постоянной температуре. Двигаясь вдоль геномной ДНК, *Phi29*, достигнув следующего праймера, не диссоциирует, а продолжает синтез, замещая (отталкивая) ранее синтезированную копию ДНК. На такие «ДНК-ветки» также отжигаются праймеры, и они тоже начинают «ветвиться». Таким образом из копий ДНК образуются разветвленные структуры. В конце процедуры *S1* нуклеаза разрезает полученную ДНК-сеть. Процессивность и точность полимеразы *Phi29* позволяют получать фрагменты до 100 т. п. н.

После амплификации генетический материал можно подготовить к секвенированию, используя обычные коммерческие наборы реагентов.

\* \* \*

В заключение главы обобщим вышесказанное: секвенирование геномов эукариот является сложной задачей из-за размера и наличия повторяющихся элементов. При секвенировании геномов *de novo* и ресеквенировании применяются схожие подходы, отличающиеся длиной фрагментов библиотек и биоинформатическими алгоритмами. Выбор платформы для секвенирования, количества данных, способа пробоподготовки и обработки данных зависит от множества параметров и зачастую носит эмпирический характер. В то же время предпочтение можно отдать парно-концевым прочтениям максимальной длины при минимальной стоимости данных.

### СПИСОК ЛИТЕРАТУРЫ

1. *C. elegans Sequencing Consortium*. Genome sequence of the nematode *C. elegans*: a platform for investigating biology // *Science*, 1998, 282 (5396): 2012–2018.
2. *Thudi M. et al.* Current state-of-art of sequencing technologies for plant genomics research // *Briefings in functional genomics*, 2012, 11 (1): 3–11.
3. *Peltola H., Söderlund H., Ukkonen E.* SEQAID: a DNA sequence assembling program based on a mathematical model // *Nucleic Acids Res*, 1984, 12 (1 Pt 1): 307–321.
4. *Coulson A., Sulston J., Brenner S., Karn J.* Toward a physical map of the genome of the nematode *Caenorhabditis elegans* // *Proceedings of the National Academy of Sciences of the United States of America*, 1986, 83 (20): 7821–7825.
5. *Pareek C.S., Smoczynski R., Tretyn A.* Sequencing technologies and genome sequencing // *Journal of applied genetics*, 2011, 52 (4): 413–435.
6. *Shendure J., Mitra R.D., Varma C., Church G.M.* Advanced sequencing technologies: methods and goals // *Nature reviews. Genetics*, 2004, 5 (5): 335–344.
7. *Natali L. et al.* The repetitive component of the sunflower genome as shown by different procedures for assembling next generation sequencing reads // *BMC genomics*, 2013, 14 (1): 686.
8. *Nowrousian M.* Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems // *Eukaryotic cell*, 2010, 9 (9): 1300–1310.
9. *Stratton M.* Genome resequencing and genetic variation // *Nature biotechnology*, 2008, 26 (1): 65–66.

10. *Hall I.M., Quinlan A.R.* Detection and interpretation of genomic structural variation in mammals. *Methods Mol Biol.*, 2012, 838: 225–248.
11. *Stephens M., Donnelly P.* A comparison of Bayesian methods for haplotype reconstruction from population genotype data // *American Journal of Human Genetics*, 2003, 73: 1162–1169.
12. *Hosomichi K. et al.* Phase-defined complete sequencing of the HLA genes by next-generation sequencing // *BMC Genomics.*, 2013, 14: 355.
13. *Kitzman J.O. et al.* Haplotype-resolved genome sequencing of a Gujarati Indian individual // *Nat Biotechnol.*, 2011, 29 (1): 59–63.
14. *Fan H.C., Wang J., Potanina A., Quake S.R.* Whole-genome molecular haplotyping of single cells // *Nat Biotechnol.*, 2011, 29 (1): 51–57.
15. *Van der Auwera G.A. et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline // *Current Protocols in Bioinformatics*, 2013, 11.10.1–11.10.33.

## СЕКВЕНИРОВАНИЕ ТРАНСКРИПТОМОВ ЭУКАРИОТ

С появлением высокопроизводительного секвенирования исследование сложных смесей молекул РНК вышло на новый уровень. По ряду показателей NGS отличается от разработанных ранее методов – ПЦР и гибридизации на чипах. Можно сказать, что NGS объединило в себе плюсы обеих методик: широкий динамический диапазон измеряемых концентраций ДНК методом количественной ПЦР и огромный спектр анализируемых транскриптов, свойственный чипам.

Исследования бактериального транскриптома и метатранскриптома описаны в главах 7 и 8. Данная глава посвящена особенностям использования NGS применительно к анализу РНК эукариотических клеток. Сразу отметим, что рассматриваемые ниже подходы чаще всего предполагают ту или иную степень обогащения исследуемого образца молекулами определенного типа (полиА-фракцией РНК, poly A+ RNA, малыми РНК и т. п.), поэтому их можно также отнести к вариантам таргетного секвенирования (см. гл. 11).

### 10.1. ПРИМЕНЕНИЕ NGS ДЛЯ ИССЛЕДОВАНИЯ РНК

Методы высокопроизводительного секвенирования оказались крайне востребованы в области исследования рибонуклеиновых кислот. Такие задачи, как измерение уровня экспрессии всех генов (в ткани или отдельной клетке), поиск соматических мутаций, альтернативно-сплайсированных форм (изоформ транскриптов) и белок-некодирующих РНК, секвенирование различных типов малых РНК и т. д., оказались хорошо совместимы с NGS-подходами. Отметим, что в литературе высокопроизводительное секвенирование с целью изучения РНК получило некорректное название – РНК-секвенирование (RNA-seq) (в большинстве технологий NGS секвенируют не РНК, а кДНК) [1].

В табл. 10.1 показаны некоторые из задач транскриптомики и способы их решения методами высокопроизводительного секвенирования.

Таблица 10.1

## Примеры исследования РНК методами NGS

Цель исследования	Целевые молекулы для библиотеки	Используемый подход	Ссылка на работу
Уровень экспрессии генов	ПолиА-мРНК	Обогащение полиА-фракции. Часто достаточно коротких прочтений (50–100 п. н.) с 3'-конца	[2, 3]
Альтернативный сплайсинг	Экзон-интронные стыки	Необходимы длинные прочтения (более 300 п. н.) или парно-концевые прочтения по 100–150 п. н.	[4, 5]
Малые РНК (miRNA, snoRNA, piRNA, snRNA, tRNA и др.)	Короткая фракция кДНК	Использование 5'-концевого фосфата. Достаточно коротких прочтений (50–100 п. н.)	[6]
Антисмысловые и некодирующие РНК	Антисмысловые и некодирующие РНК	Необходимо создание 5'–3' ориентированной библиотеки кДНК	[7–9]
Транскриптом одной клетки	ПолиА-мРНК	Необходима предварительная амплификация библиотеки кДНК	[10–13]
Транслируемые РНК	РНК на рибосомах	Ферментативное разрушение РНК (кроме участков внутри рибосом)	[2, 14]
Двухцепочечные РНК и вторичные структуры на РНК	Двухцепочечные участки РНК	Обработка ацилирующими агентами	[15]

## 10.2. ОБЩИЕ МОМЕНТЫ ОЧИСТКИ РНК И СИНТЕЗА кДНК

Несмотря на появление технологий NGS, позволяющих анализировать непосредственно молекулы РНК (например, технологии PacBio, см. гл. 3), наиболее распространенные платформы NGS по-прежнему используют в качестве стартового материала копию кДНК с исходной молекулы РНК. В этой связи одним из первых этапов секвенирования РНК является перевод РНК в кДНК путем обратной транскрипции (реакции синтеза копии ДНК (комплементарной ДНК, кДНК) на матрице РНК с использованием специального фермента – обратной транскриптазы, или ревертазы).

Эффективность синтеза кДНК зависит от качества очистки РНК. Сегодня на рынке присутствует множество наборов реагентов для очистки РНК, но методика классической фенольной экстракции остается одним из наиболее стандартных подходов.

Для удаления геномной ДНК из библиотеки можно рекомендовать на финальных этапах очистки осаждение РНК хлоридом лития и обязательную последующую обработку препарата ДНКазой (с маркировкой «RNase free»).

Для синтеза кДНК на матрице РНК можно использовать несколько типов олигонуклеотидных затравок (праймеров), существенно отличающихся по сложности и количеству получаемой в результате реакции кДНК, а также по-разному влияющих на соотношение отдельных транскриптов после обратной транскрипции. Основные варианты стадии затравления следующие: 1) самозатравление (без использования специально добавляемых в реакцию олигонуклеотидов); 2) с использованием случайных затравок (random primers); 3) с использованием затравки олиго-dT; 4) с применением затравки, специфичной к отдельным транскриптам. Следует отметить, что температура гибридизации относительно коротких (обычно длиной 6–8 нуклеотидов) случайных затравок, а также олиго-dT праймера, значительно ниже оптимальной температуры для термостабильной реакции обратной транскрипции, поэтому эти типы затравок не могут быть использованы в работе с термостабильными ферментами для обратной транскрипции без предварительного этапа инкубации при более низкой температуре [16].

Вне зависимости от того, какой тип затравок будет использован в реакции обратной транскрипции, начало синтеза кДНК может происходить вследствие случайного эндогенного затравления обрывками РНК и ДНК, образовавшимися в ходе очистки нуклеиновых кислот и за счет формирования вторичных структур в РНК. В экспериментах по постановке обратной транскрипции с меченым  $P^{32}$ -стандартом при использовании обратных транскриптаз AMV, MMLV и мутантов MMLV было установлено, что по количеству полученной кДНК продукты реакций, выполненных с добавлением и без добавления затравок, практически идентичны. Высокая эффективность самозатравочного механизма обратной транскрипции типична для реакций, проводимых с использованием обратных транскриптаз AMV или MMLV при стандартных условиях. На это следует обратить особое внимание в тех случаях, когда необходимо получить первую цепь кДНК с определенной затравки и без примеси второй цепи (например, при получении ориентированных библиотек кДНК). Для этого можно рекомендовать добавление ревертазы при высокой температуре и использование актиномицина D в качестве агента, снижающего спонтанный синтез второй цепи [17].

Другим вариантом проведения обратной транскрипции является использование случайных затравок. В реакцию обратной транскрипции добавляют смесь всех возможных вариантов коротких дезоксирибонуклеотидов (чаще всего – гексамеров). Короткие затравки обеспечивают начало синтеза кДНК во множестве точек на всем протяжении молекулы РНК, производя, таким образом, несколько (укороченных) молекул кДНК с одной молекулы мишени. По некоторым данным, этот метод позволяет получить наибольшее количество кДНК и лучше всего применим к транскриптам с развитой вторичной структурой. Также использование этого подхода рекомендовано в случае использования деградированной (разрушенной на короткие фрагменты) РНК.

Синтез первой цепи кДНК (цепи, непосредственно синтезируемой на молекуле РНК) с использованием смеси вырожденных гексамеров можно проводить при комнатной температуре. Однако при использовании суммарной фракции РНК в качестве стартового материала большая часть полученной библиотеки кДНК будет соответствовать рРНК и секвенирование даст миллионы ненужных прочтений. Поэтому при ис-

пользовании случайной затравки необходимо предварительно избавиться от рРНК (см. разд. 10.4.2 и 10.4.3).

Синтез кДНК с использованием праймера олиго-dT более специфичен к мРНК (в сравнении со смесью случайных гексамеров), поскольку этот тип затравки плохо работает на неполиаденилированной рРНК. Для получения кДНК с использованием праймера олиго-dT крайне важным является отсутствие деградации и качество очистки РНК. Выбор этого типа затравления не рекомендован, если РНК фрагментирована.

### 10.3. ФЕРМЕНТЫ ДЛЯ ОБРАТНОЙ ТРАНСКРИПЦИИ

Несмотря на наличие широкого спектра обратных транскриптаз (табл. 10.2), разработчиками методики по-прежнему не решено несколько проблем, связанных с проведением обратной транскрипции (ОТ): все ревертазы (РНК-зависимые ДНК-полимеразы), в сравнении с ДНК-зависимыми ДНК-полимеразами, имеют низкую точность копирования матрицы (вследствие своей вирусной природы). Также «в противофазе» находятся эффективность фермента и его термостабильность. Низкая термостабильность ревертаз – серьезная проблема, поскольку для эффективной обратной транскрипции желательно денатурировать развитые вторичные структуры РНК, что возможно в случае проведения реакции при высокой температуре.

Таблица 10.2

Некоторые представленные на рынке ферменты  
для обратной транскрипции

Фермент	Производитель	Оптимум температуры, °C	Активность РНКазы Н	Особенности фермента или его использования
AMV-ОТ	много	37	высокая	—
MMLV-ОТ	много	45	средняя	—
<i>Tth</i> -полимераза	много	60–72	низкая	ОТ-ПЦР в одной пробирке

Продолжение таблицы на след. стр.

Продолжение табл. 10.2

Фермент	Производитель	Оптимум температуры, °C	Активность РНКазы Н	Особенности фермента или его использования
<i>Tfl</i> -полимераза	Promega	60–72	низкая	ОТ–ПЦР в одной пробирке
MMLV PM	Promega	42–55	низкая	точечная мутация в домене РНКазы Н MMLV–ОТ
Sensiscript	Qiagen	37	средняя	не является производным AMV или MMLV
Omniscript	Qiagen	37	средняя	не является производной AMV или MMLV
StrataScript	Stratagene	42–50	низкая	точечная мутация в домене РНКазы Н MMLV–ОТ
Expand RT	Roche	42	низкая	точечная мутация в домене РНКазы Н MMLV–ОТ
RevertAid	Fermentas	42–45	низкая	точечная мутация в домене РНКазы Н MMLV–ОТ
ImPromII	Promega	55	низкая	—
Powerscript	Clontech	42	низкая	—
Superscript II	Invitrogen	42–50	низкая	—
Thermoscript	Invitrogen	50–65	низкая	—

## 10.4. ПОДГОТОВКА БИБЛИОТЕКИ кДНК ДЛЯ NGS

Общие рекомендации подготовки библиотек фрагментов ДНК для высокопроизводительного секвенирования описаны в главе 2. Здесь рассмотрим лишь особенности подготовки библиотек применительно к исследованию различных типов РНК.

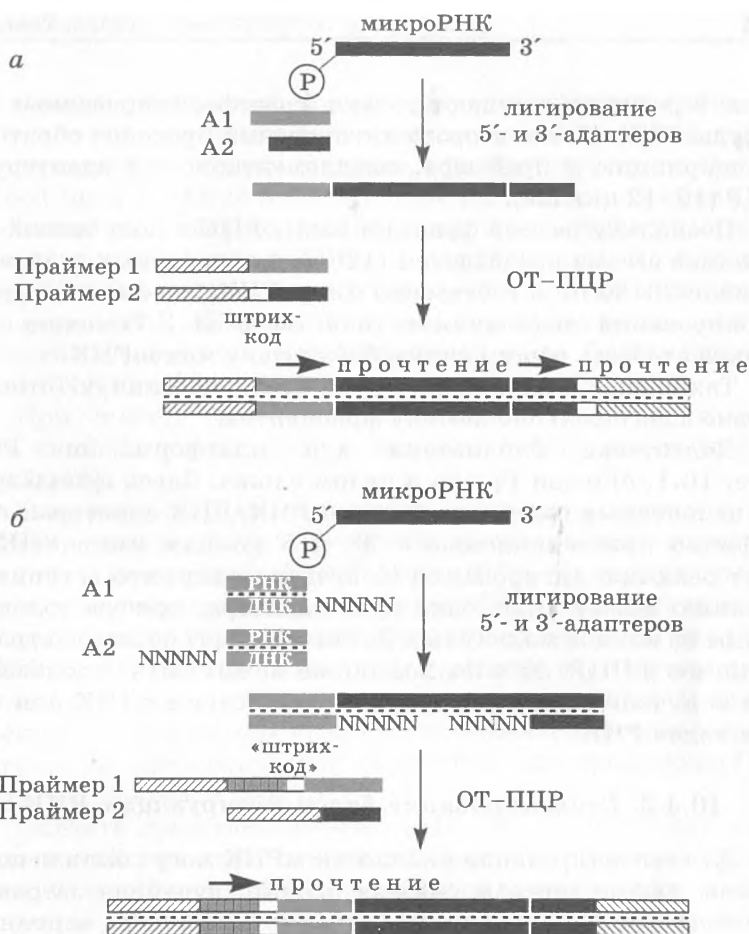
В зависимости от того, какие транскрипты предполагается изучать, необходимо создать соответствующую библиотеку кДНК. Для комплексных проектов необходимо, чтобы в библиотеку вошли все классы РНК, включая белок-кодирующие мРНК, некодирующие РНК, антисмысловые РНК и пр. Однако часто исследователь планирует изучить только белок-кодирующие мРНК или только малые РНК – в таком случае существуют подходы выборочного обогащения библиотеки определенными видами молекул РНК.

### 10.4.1. Секвенирование малых РНК

Секвенирование малых РНК (miRNA, snoRNA, piRNA, snRNA, tRNA и др.) чаще всего предполагает обогащение фракции целевых молекул со «штрих-кодированием» получаемой библиотеки (для возможности смешивания нескольких образцов). Обогащение может быть основано на различных свойствах малых РНК, в частности на размере (отбор фракции фрагментов определенной длины). Рассмотрим создание библиотеки на примере микроРНК (miRNA).

Протокол приготовления библиотеки для секвенирования микроРНК сравнительно прост и может быть выполнен в один шаг [18, 19] (рис. 10.1). Для этого используют тот факт, что микроРНК несут на 5'-конце фосфат, по которому может быть проведена реакция лигирования. В протоколе для Illumina (рис. 10.1, а), аденилированный ДНК-адаптер с заблокированным 3'-концом лигируют к образцу РНК с помощью модифицированной РНК-лигазы 2 фага T4.

Фермент модифицирован так, что в качестве субстрата использует только 3'-аденилированные адаптеры, что позволяет исключить лигирование фрагментов РНК между собой. Только 3'-аденилированные адаптеры могут быть пришиты к свободным 3'-концам молекул РНК (а поскольку адаптеры 3'-блокированы, они не могут образовывать димеры). Затем в реакцию добавляют 5'-адаптер, АТФ и РНК-лигазу 1. На этом



**Рис. 10.1.** Подготовка библиотеки для секвенирования микроРНК. *а* – Протокол для платформы Illumina, включающий лигирование 3'-аденилированного ДНК-адаптера к 3'-концу микроРНК в препарате суммарной РНК. Затем РНК-адаптер лигируют к 5'-концу микроРНК. Фланкированные адаптерами молекулы амплифицируют в ходе обратной транскрипции с последующей ПЦР. Первое прочтение дает последовательность микроРНК, второе – «штрих-код». *б* – Протокол для платформы Ion PGM включает лигирование с помощью РНК-лигазы адаптеров, представляющих собой РНК/ДНК гетеродуплексы.

Таким образом, адаптеры присоединяются ориентированно. В ходе последующей обратной транскрипции и ПЦР в библиотеку вводят «штрих-коды». «Штрих-код» и последовательность вставки секвенируют за одно прочтение

этапе в реакцию вступают только 5'-фосфорилированные молекулы РНК. После второго лигирования проводят обратную транскрипцию (с праймера, комплементарного 3'-адаптеру) и ПЦР (10–12 циклов).

Поскольку размер фракции микроРНК в полученной библиотеке весьма предсказуем (120 п. н. суммарных адаптеров и около 20–30 п. н. собственно микроРНК), с помощью фракционирования (вырезания из геля, см. разд. 2.7) можно прицельно отобрать обогащенную библиотеку микроРНК.

Такой способ дает в итоге 5'–3' ориентированную (относительно адаптеров) библиотеку фрагментов.

Подготовка библиотеки для платформ Ion PGM (рис. 10.1, б) и Ion Proton в целом схожа. Здесь используют двуцепочечные гетеродуплексные РНК/ДНК-адаптеры, специфично присоединяемые к 3'- и 5'-концам микроРНК за одну реакцию лигирования (благодаря тому, что вступить в реакцию может лишь одна цепь адаптера, причем только с одним из концов молекулы). Затем проводят обратную транскрипцию и ПЦР. Этот подход также может быть использован для получения ориентированной библиотеки кДНК для любых видов РНК.

#### 10.4.2. Секвенирование белок-кодирующих РНК

Для секвенирования библиотек мРНК могут быть использованы любые способы синтеза кДНК: случайная затравка, олиго-dT-праймер или присоединение различных вариантов адаптеров (с возможностью последующей амплификации библиотеки кДНК).

В случае использования случайной затравки перед реакцией обратной транскрипции необходимо удалить фракцию рРНК (например, используя наборы реагентов Ribo-Zero (Epicenter) или RiboMinus (Life Technologies Thermo Fisher Scientific), либо отобрать фракцию полиаденилированных мРНК на колонках с олиго-dT).

В некоторых случаях важно, чтобы получаемая библиотека фрагментов была ориентирована относительно 5'–3'-концов молекулы РНК (например, при изучении антисмысловых РНК или некодирующих РНК). Секвенирование такой (ориентированной) библиотеки кДНК называют ориентированным секвенированием [20].

Один из вариантов создания ориентированной библиотеки – это избирательное уничтожение одной из цепей кДНК путем введения дезоксиуридинтрифосфата на этапе синтеза второй цепи кДНК (с последующим расщеплением урацилсодержащей цепи ферментом [21], для чего разработаны такие наборы реагентов, как NEBNext Ultra Directional RNA Library Prep Kit (для платформы Illumina)) или с использованием в последующей ПЦР модифицированной полимеразы, «не узнающей» урацил в матричной цепи (Illumina TruSeq Stranded Total RNA Kit).

При синтезе первой цепи с целью снижения спонтанного синтеза второй цепи в реакцию добавляют актиномицин D [17].

Еще один подход использует случайную затравку или праймер олиго-dT с адаптерной последовательностью на 5'-конце для синтеза первой цепи кДНК. Благодаря способности ревертазы продолжать синтез кДНК по «подставленному» на 5'-конце РНК адаптеру (эффект смены цепи, template-switching effect или cap-switching effect), адаптер сразу добавляется и на другом конце первой цепи кДНК [22, 23]. Преимуществом метода является возможность использования полученной одноцепочечной кДНК сразу для проведения ПЦР (без синтеза второй цепи).

Снизить представленность рРНК в полученной библиотеке можно использованием супрессионных адаптеров определенной структуры. На рынке имеются коммерческие наборы реагентов для подавления амплификации рРНК в получаемой библиотеке кДНК (NuGEN Ovation RNA-seq).

Еще один подход использует для синтеза первой цепи 749 специально отобранных гексамеров (из 4096 возможных), которые плохо отжигаются на рРНК, при этом можно в пять раз снизить представленность рРНК в библиотеке [24].

### **10.4.3. Нормализация библиотеки кДНК перед секвенированием**

Состав эукариотического транскриптома хорошо укладывается в принцип Парето: 20% разных по последовательности транскриптов составляют 80% от общего количества мРНК. Разница в уровне представленности некоторых транскриптов составляет более  $10^5$  раз. Для идентификации редкопредстав-

ленных транскриптов необходимо секвенировать библиотеку фрагментов с большим покрытием, при этом последовательности средних и мажорных транскриптов будут секвенированы многие тысячи раз. Потеря фракции редкопредставленных транскриптов приводит к неполному анализу набора последовательностей, отличающих образцы кДНК.

Если исследователь стремится определить наибольшее число разных по последовательности транскриптов и при этом готов пренебречь информацией об уровне их представленности, можно существенно удешевить процедуру секвенирования и повысить вероятность обнаружения наибольшего разнообразия минорных транскриптов за счет предварительной нормализации кДНК. Для этого готовят нормализованные библиотеки, т. е. библиотеки, в которых разные виды кДНК имеют приблизительно равную представленность. За счет выравнивания концентрации молекул разных типов нормализация позволяет дополнительно снизить примесь рибосомальной и транспортной кДНК.

К настоящему моменту описано множество методик получения нормализованных библиотек кДНК [25–29]. Наиболее эффективный принцип создания нормализованных библиотек основан на том, что реассоциация денатурированной двуцепочечной ДНК является реакцией второго порядка, и, следовательно, мажорные последовательности реассоциируют значительно быстрее, чем минорные. Таким образом, если образец двуцепочечной кДНК денатурировать и снова ренатурировать, основная масса высокопредставленных молекул перейдет в двуцепочечную (ds) форму, и одноцепочечная (ss) фракция кДНК станет в значительной мере нормализованной [30]. Однако отбор ss-фракции представляет наибольшую проблему при создании нормализованных библиотек фрагментов. В разных методиках эта проблема решалась физическим разделением ss- и ds-фракций с помощью гидроксипатитных колонок [25, 26, 27], магнитных частиц [28] или использованием супрессионной ПЦР [31].

Одним из наиболее эффективных методов нормализации кДНК является подход на основе дуплекс-специфичной нуклеазы камчатского краба (DSN) [32, 33]. Синтез первой цепи и амплификацию полноразмерной кДНК осуществляют стандартными методами (например, с использованием затравки олиго-dT с «подставленным» к 5'-концу РНК адаптером). Затем получен-

ную библиотеку денатурируют и снова ренатурируют. В ходе ренатурации более представленные транскрипты образуют двуцепочечную форму быстрее. Со временем одноцепочечная фракция становится все более «ровной» по представленности молекул разного типа. Двуцепочечную фракцию разрушают при помощи дуплекс-специфичной нуклеазы камчатского краба (DSN-нормализация), а одноцепочечную снова амплифицируют в коде ПЦР.

DSN-нормализация успешно применяется в различных биологических моделях и на разных объектах [34].

#### **10.4.4. Приготовление библиотеки кДНК из малого количества РНК (транскриптом одной клетки)**

Среднестатистическая эукариотическая клетка содержит примерно 10 пг суммарной РНК, от которой белок-кодирующие РНК могут составлять лишь 1%. Следовательно, все подходы по изучению РНК из одной клетки требуют какой-либо амплификации кДНК для получения необходимого количества материи с целью приготовления библиотеки для секвенирования [11, 16].

Для получения библиотеки кДНК из малых количеств РНК (например, из РНК, полученной из единственной клетки) используют различные стратегии амплификации. На рынке представлено множество наборов реагентов, подходящих для этой цели (SMARTer Ultra Low RNA Kit от Clontech, набор реактивов Mint-2 от «Евроген»).

#### **10.4.5. Влияние амплификации на представленность транскрипта (искажение профиля транскрипции)**

Недостатком всех технологий амплификации сложной смеси нуклеиновых кислот является неизбежное искажение представленности молекул и потеря отдельных последовательностей [35]. Причина этого кроется в том, что, когда исходное количество молекул, добавляемых в реакционную пробирку (в ПЦР), крайне мало, случайные колебания числа стартовых молекул становятся заметны [36]. Вместе с тем чем ниже эффективность амплификации, тем выше вклад случайных колебаний начального числа матриц в результат. Наличие

существенной погрешности измерения низкопредставленных транскриптов методом секвенирования транскриптома означает, что достоверно можно зафиксировать лишь существенные изменения уровня экспрессии гена в результате экспериментального воздействия (минимум – в несколько раз для транскриптов уровня менее 1 копии на клетку). Ожидаемый эффект (например, изменение уровня представленности транскрипта) для конкретного дизайна эксперимента желательно оценить еще до начала эксперимента. Это позволит выбрать правильный размер образца, необходимый для достижения статистически и биологически значимого результата.

#### 10.4.6. Рибосомальный футпринтинг и изучение вторичных структур РНК

Данный подход позволяет выявить молекулы РНК, находящиеся в данный момент в процессе трансляции [2, 14]. Протокол включает обработку клеточного лизата РНКазами, которым оказывается недоступен защищенный рибосомой фрагмент мРНК примерно в 30 оснований. Затем рибосомы очищают (например, центрифугированием в градиенте плотности сахарозы). Фрагменты РНК отделяют от рибосом, проводят обратную транскрипцию и амплификацию библиотеки.

Еще одно направление исследования РНК методами высокопроизводительного секвенирования – изучение вторичных структур в РНК (SHAPE-Seq, selective 2-hydroxyl acylation analyzed by primer extension) [15]. Метод включает модификацию (ацилирование) неспаренных оснований в РНК с последующим синтезом кДНК. Сравнение результатов секвенирования модифицированного и необработанного препаратов РНК позволяет выявить места образования вторичных структур.

#### СПИСОК ЛИТЕРАТУРЫ

1. Wang, Z., Gerstein M., Snyder M. RNA-Seq: A revolutionary tool for transcriptomics // Nat. Rev. Genet., 2009, 10: 57–63.
2. Ingolia N.T. et al. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosomeprotected mRNA fragments // Nat. Protoc, 2012, 7: 1534–1550.
3. Mäder U. et al. Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods // Curr. Opin. Biotechnol, 2011, 22: 32–41.

4. *Eswaran J. et al.* RNA sequencing of cancer reveals novel splicing alterations // *Scientific reports*, 2013, 3: 1689.
5. *Zhao C. et al.* RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. *BMC // Genomics*, 2013, 14: 21.
6. *Röther S., Meister G.* Small RNAs derived from longer non-coding RNAs // *Biochimie*, 2011, 93: 1905–1915.
7. *Saxena A., Carninci P.* Long non-coding RNA modifies chromatin: epigenetic silencing by long non-coding RNAs. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 2011, 33: 830–839.
8. *Morris K.V., Vogt P.K.* Long antisense non-coding RNAs and their role in transcription and oncogenesis // *Cell Cycle*, 2010, 9: 2544–2547.
9. *Wang H. et al.* Genome-wide identification of long noncoding natural antisense transcripts and their responses to light in *Arabidopsis* // *Genome Res*, 2014 Mar; 24(3): 444–53.
10. *Hashimshony T., Wagner F., Sher N., Yanai I.* CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification // *Cell Rep*, 2012, 2: 666–673.
11. *Sasagawa Y. et al.* Quartz-Seq: A highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene expression heterogeneity // *Genome Biol.*, 2013, 14: R31.
12. *Zong C., S. Lu A.R. Chapman X., Xie. S.* Genome-wide detection of single-nucleotide and copy-number variations of a single human cell // *Science*, 2012, 338: 1622–1626.
13. *Wang J., Fan H.C., Behr B., Quake S.R.* Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm // *Cell*, 2012, 150: 402–412.
14. *Ingolia N.T. et al.* Genomewide annotation and quantitation of translation by ribosome profiling // *Curr Protoc Mol Biol*, 2013, Chapter 4: Unit 4, 18.
15. *Mortimer S.A. et al.* SHAPE-Seq: High- Throughput RNA Structure Analysis // *Curr Protoc Chem Biol.*, 2012, 4: 275–297.
16. *Lukyanov K. et al.* Construction of cDNA libraries from small amounts of total RNA using the suppression PCR effect // *Biochem Biophys Res Commun*, 1997, Jan 13; 230(2): 285–8.
17. *Perocchi F., Xu Z., Clauder-Munster S., Steinmetz L.M.* Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D // *Nucleic Acids Res.*, 2007, 35: e128.
18. *Vigneault F., Sismour A.M., Church G.M.* Efficient microRNA capture and bar-coding via enzymatic oligonucleotide adenylation // *Nat. Methods*, 2008, 5: 777–779.

19. *Buermans H.P. et al.* New methods for next generation sequencing based microRNA expression profiling // *BMC Genomics*, 2010, 11: 716.
20. *Levin J.Z. et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods // *Nat. Methods*, 2010, 7: 709–715
21. *Parkhomchuk D. et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA // *Nucleic Acids Res.*, 2009, 37: e123
22. *Matz M. et al.* Amplification of cDNA ends based on template-switching effect and step-out PCR // *Nucleic Acids Res.*, 1999, Mar 15; 27(6): 1558–60.
23. *Zhu Y.Y. et al.* Reverse transcriptase template switching: A SMART approach for fulllength cDNA library construction // *Biotechniques*, 2001, 30: 892–897.
24. *Armour C.D. et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis // *Nat. Methods*, 2009, 6: 647–649.
25. *Ko M.S.* An equalized cDNA library by the reassociation of short double-stranded cDNAs // *Nucl. Acids. Res.*, 1990, 18, 5705–5711.
26. *Patanjali S.R., Parimoo S., Weissman Sh.M.* Construction of a uniform-abundance (normalized) cDNA library // *Proc. Natl. Acad. Sci. USA*, 1991, 88, 1943–1947.
27. *Soares M.B. et al.* Construction and characterization of a normalized cDNA library // *Proc. Natl. Acad. Sci. USA*, 1994, 91, 9228–9232.
28. *Sasaki Y.F., Ayusava D., Oishi M.* Construction of a normalized cDNA library by introduction of a semi-solid mRNA-cDNA hybridization system // *Nucl. Acids Res.*, 1994, 22, 987–992.
29. *Coche Th., Dewar M.* Reducing bias in cDNA sequence representation by molecular selection // *Nucl. Acids Res.*, 1994, 22, 4545–4546.
30. *Galau G., Klein W., Britten R., Davidson E.* Significance of rare mRNA sequences in liver // *Arch. Biochem. Biophys.*, 1977, 179, 584–599.
31. *Rebrikov D.V., Desai S.M., Siebert P.D., Lukyanov S.A.* Suppression subtractive hybridization // *Methods Mol Biol.*, 2004, 258: 107–34.
32. *Shagin D.A. et al.* A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas // *Genome Res.*, 2002, 12 (12): 1935–1942.
33. *Zhulidov P.A. et al.* Simple cDNA normalization using kamchatka crab duplex-specific nuclease // *Nucleic Acids Res.*, 2004, 32 (3): e37.
34. *Bogdanova E.A., Shagin D.A., Lukyanov S.A.* Normalization of full-length enriched cDNA // *Mol Biosyst.*, 2008, 4 (3): 205–212.
35. *Bhargava V.H. et al.* Technical variations in low-input RNA-seq methodologies // *Sci Rep.*, 2014, 4: 3678.
36. *Peccoud J., Jacob C.* Theoretical uncertainty of measurements using quantitative polymerase chain reaction // *Biophys J.*, 1996, Jul; 71(1): 101–108.

# **ПОВЫШЕНИЕ КОНЦЕНТРАЦИИ ОПРЕДЕЛЕННЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ В БИБЛИОТЕКЕ ДЛЯ NGS (ТАРГЕТНОЕ СЕКВЕНИРОВАНИЕ)**

Полное секвенирование эукариотических геномов пока остается довольно дорогим исследованием. Кроме того, анализ полногеномных данных и их хранение требуют дополнительных материальных и временных ресурсов. В то же время для большинства биологических задач (включая медицинскую генетическую диагностику) повторное секвенирование полного генома не требуется. Гораздо эффективнее определить лишь участки ДНК или РНК, несущие полезную информацию, такие данные дешевле и их легче анализировать.

Как следствие, значительные усилия были приложены для разработки методов целевого обогащения определенных последовательностей в библиотеке для секвенирования, что подразумевает отбор интересующих фрагментов ДНК из общего пула перед секвенированием. Данная глава посвящена технологиям обогащения библиотек фрагментов ДНК определенными последовательностями, включая комбинацию иммунопреципитации хроматина и NGS (ChIP-seq).

## **11.1. ПАРАМЕТРЫ МЕТОДОВ ЦЕЛЕВОГО ОБОГАЩЕНИЯ**

При рассмотрении методов целевого обогащения можно выделить ряд параметров, характеризующих тот или иной подход:

- 1) эффективность обогащения – доля последовательностей, картируемых на целевых участках генома по отношению к общему объему получаемых данных;
- 2) однородность покрытия целевых регионов;
- 3) воспроизводимость результатов экспериментальных повторов;

- 4) стоимость;
- 5) простота методики;
- 6) количество ДНК, необходимое для проведения эксперимента, или количество образца ДНК на миллион пар нуклеотидов целевого региона.

Подходы, обеспечивающие высокую однородность и эффективность обогащения, позволяют получить оптимальное покрытие при меньшем объеме первичных данных, что удешевляет процесс секвенирования. Кроме того, для выбора подходящего метода целевого обогащения для конкретного проекта нужно учитывать размер интересующего региона, количество образцов и необходимость смешивания образцов (со «штрих-кодированием») для оптимального использования мощности секвенатора.

Ниже рассмотрены наиболее распространенные методы целевого обогащения.

## **11.2. ОБОГАЩЕНИЕ БИБЛИОТЕКИ ФРАГМЕНТОВ ДНК ТОЛЬКО НА ОСНОВЕ ПЦР**

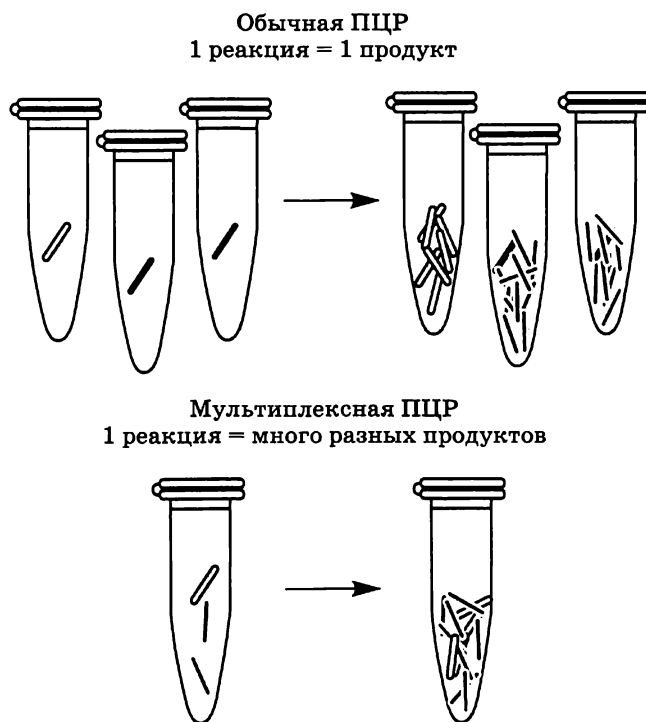
Так или иначе, но почти все методы обогащения библиотеки фрагментов ДНК целевыми последовательностями используют на определенном этапе метод полимеразной цепной реакции. Для некоторого структурирования подходов мы разделили их на две группы – в одной метод ПЦР является ключевым «сепаратором», тогда как в другой дискриминирующим фактором является иной принцип, а ПЦР используется лишь для увеличения концентрации библиотеки.

### **11.2.1. Классическая и мультиплексная ПЦР**

ПЦР остается самым распространенным видом подготовки образцов к секвенированию на протяжении 30 лет. Но если для метода Сенгера необходимо было поставить одну реакцию с получением единственного продукта, то технологии NGS требуют получения многих тысяч разных по последовательности фрагментов ДНК для одновременного их секвенирования. Следовательно, при использовании ПЦР для обогащения библиотек перед NGS требуется так называемое мультиплексирование реакций (т. е. ПЦР со множества пар праймеров одновременно). Кроме того, для оптимального использования

производительности большинства NGS-платформ требуется «штрих-кодирование» множества образцов с дальнейшим одновременным секвенированием в рамках одного запуска.

Мультиплексная ПЦР представляет собой крайне нестабильную систему. Без особых ухищрений можно проводить реакцию с использованием лишь небольшого числа пар праймеров (5–10 пар). Большее количество пар праймеров в одной реакции почти всегда дает неспецифичные продукты и димеры праймеров (вследствие неизбежного взаимодействия между праймерами), а амплификация некоторых целевых продуктов может не пройти вовсе. Для проведения мультиплексной ПЦР были разработаны специальные методы (например [1]), однако зачастую на практике эффективнее провести несколько одиночных ПЦР в отдельных пробирках (в формате 96-, 384- или 1536-луночных планшетов) (рис. 11.1).



**Рис. 11.1.** Обогащение целевыми продуктами с помощью классической ПЦР и мультиплексной ПЦР

После амплификации концентрации продуктов для разных образцов должны быть точно выровнены до объединения их в пул, чтобы избежать преимущественного секвенирования отдельных образцов. Существует несколько способов выравнивания, из которых оптимальным можно считать подход, сочетающий визуальную оценку длин и числа фрагментов на электрофореграмме и количественную оценку с помощью флуориметров (см. разд. 2.2).

Можно отметить, что применение ПЦР для обогащения библиотеки целевыми регионами перед высокопроизводительным секвенированием целесообразно при необходимости получить сравнительно короткие участки генома (ограниченные длиной нескольких ПЦР-реакций – 10–30 т. п. н.) для множества биологических образцов с высоким покрытием. Как правило, в такую схему укладываются экзоны нескольких интересующих исследователя генов.

В частности, ПЦР в качестве инструмента для обогащения таргетными фрагментами широко применяется в медицинской практике при секвенировании нескольких экзонов (от 1 до 50) сравнительно протяженных генов – длина которых превышает несколько тысяч нуклеотидных пар. Например, при выявлении мутаций в гене *LDLR* в ходе диагностики семейной гиперхолестеринемии используют 6 пар праймеров (6 независимых ПЦР-реакций) для амплификации кодирующих участков гена *LDLR* (длиной 46 477 п. н.) с последующим секвенированием ампликонов с помощью Ion Torrent [2].

Как было сказано в главе 8, таргетное обогащение с помощью ПЦР применяется при метагеномном секвенировании. В этом случае проводится амплификация одного или нескольких участков 16S рРНК длиной до 1000 п. н. Использование ПЦР позволяет детектировать образцы ДНК, встречающиеся в образце в низкой концентрации, а низкая стоимость реактивов и широкая распространенность необходимого оборудования (нужен лишь обычный ДНК-ампликатор) делает такой вариант наиболее рациональным [3].

### **Технология Ion AmpliSeq (Life Technologies Thermo Fisher Scientific)**

Ion AmpliSeq – коммерческая технология мультиплексной ПЦР, адаптированная для использования совместно с протоколами секвенирования Ion PGM и Ion Proton. Технология

позволяет в одной пробирке одновременно амплифицировать несколько тысяч разных фрагментов (с использованием соответствующего числа специфичных пар праймеров) суммарной длиной до 5 млн п. н. Для работы необходимо не менее 10 мг геномной ДНК, причем есть протоколы, по которым можно работать с образцами из парафиновых блоков. Процедура занимает около 3,5 ч. Смешиваемые образцы можно «штрих-кодировать».

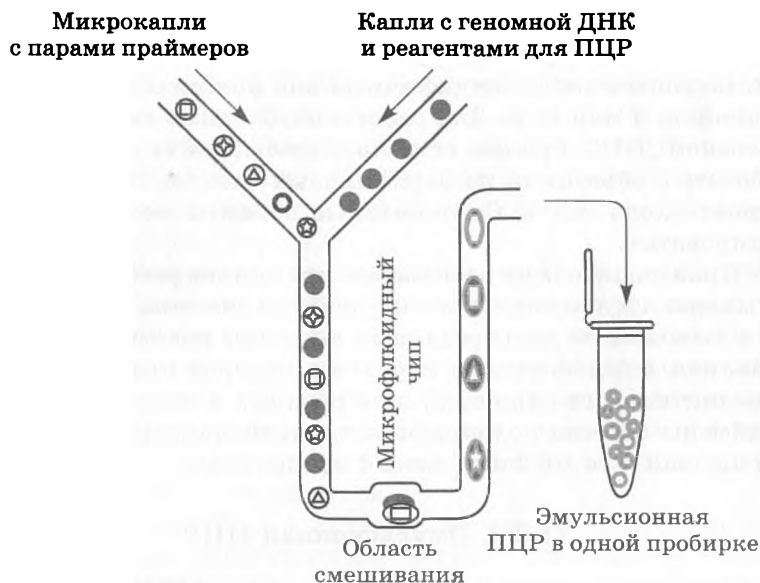
Производитель не раскрывает принципов работы системы, но можно предположить использование системы step-out PCR [4] с выходом на универсальный внешний праймер и ингибированием амплификации коротких димеров праймеров. Для дополнительного снижения димеризации в такой системе все праймеры обычно подбирают заканчивающимися (на 3'-конце) на одни и те же 3 или даже 4 нуклеотида.

### 11.2.2. Эмульсионная ПЦР

Описанная в разделе 2.9.2 эмульсионная ПЦР может быть использована и для получения обогащенной библиотеки. На этом принципе основано несколько коммерческих технологий.

#### Технология RainStorm (RainDance Technologies)

Примером целевого обогащения при помощи ПЦР со множеством праймеров может служить платформа RainStorm от RainDance Technologies. Эта платформа основана на использовании микрокапель в эмульсии. С применением специального оборудования на основе микрофлюидики (см. также разд. 11.2.3) в одной пробирке создается несколько миллионов капель, содержащих геномную ДНК и одну из нескольких тысяч пар праймеров (рис. 11.2). В каждой капле проходит независимая ПЦР, на эффективность которой не влияют другие праймеры и продукты. Совокупность таких капель (эмульсия) позволяет проводить реакцию с тысячами разных пар праймеров одновременно. По завершении ПЦР эмульсию разрушают и очищают продукты реакции. Смесь ампликонов затем можно использовать для создания библиотеки фрагментов и последующего высокопроизводительного секвенирования. Поскольку во время проведения описанной ПЦР каждая пара праймеров пространственно отделена от остальных, реакция лишена недостатков мультиплексной ПЦР [5].



**Рис. 11.2.** Принцип создания эмульсии, содержащей микрокапли с отдельными парами праймеров

Однако даже использование автоматизированных систем пробоподготовки не позволяет сделать ПЦР экономически оправданным подходом обогащения библиотеки многими миллионами пар нуклеотидов. Основные причины заключаются в высокой стоимости коллекции праймеров, а также из-за необходимости использовать значительное количество стартовой ДНК (что не всегда возможно). Как следствие, для больших целевых регионов (например, экзона человека) были разработаны другие методы таргетного обогащения.

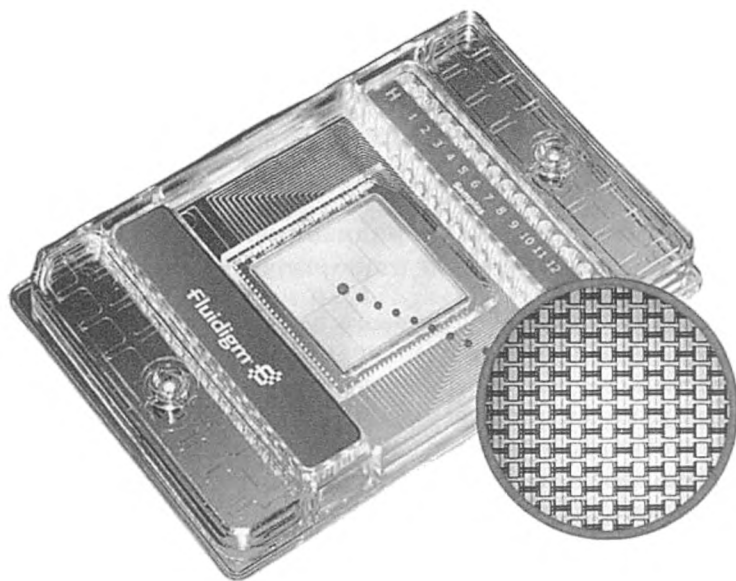
### 11.2.3. Подходы, основанные на микрофлюидике

Бурно развивающиеся в последнее время методы работы с микрообъемами (технологии микрофлюидики) и множество созданных аппаратных решений для микросмешивания позволяют упростить и удешевить процесс подготовки библиотек.

### Fluidigm Access Array (Fluidigm Corporation)

Технология Fluidigm Access Array позиционируется производителем для таргетного секвенирования. Принцип основан на постановке тысяч обычных ПЦР с нужными парами праймеров, но приготовление реакций существенно упрощено за счет автоматизации.

В микрочип Fluidigm с одной стороны загружают образцы, с другой – праймеры. На чип можно загрузить до 96 пар праймеров и до 96 образцов ДНК (есть форматы  $24 \times 192$ ,  $48 \times 48$  и др.). Система в автоматическом режиме смешивает тысячи комбинаций образцов и праймеров и проводит ПЦР в реакционных камерах объемом 35 нл (есть чипы с объемами меньше нанолитра, рис. 11.3). Таким образом с одного чипа можно получить до  $96 \times 96 = 9216$  разных ампликонов.



**Рис. 11.3.** Чип для микрофлюидного смешивания реагентов (Fluidigm Corporation)

Одна из проблем при подготовке обогащенных библиотек фрагментов для NGS – необходимость выравнивания концен-

трации целевых продуктов. Система Fluidigm позволяет провести такую нормировку автоматически по получаемым концентрациям ампликонов.

Fluidigm Access Array является открытой системой и производитель предлагает заранее адаптированные протоколы для Ion PGM, Ion Proton, Roche 454 Life Sciences и Illumina. Стоимость на образец при использовании таких чипов составляет порядка 500 руб.

### **11.3. ОБОГАЩЕНИЕ БИБЛИОТЕКИ ФРАГМЕНТОВ ДНК ПРИ ПОМОЩИ ГИБРИДИЗАЦИИ С ПРОБОЙ**

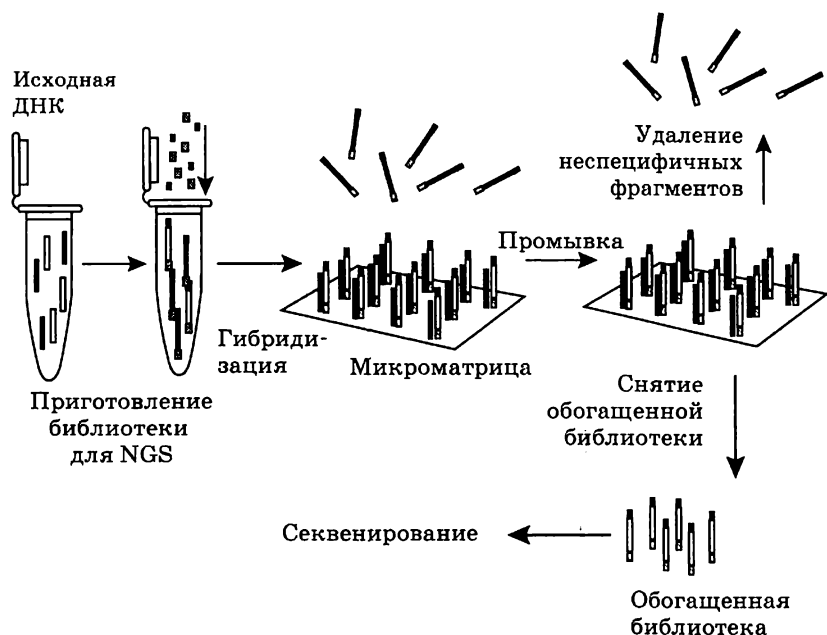
Этот принцип обогащения целевыми последовательностями давно и успешно используется в различных областях молекулярной биологии. Например, на нем основан отбор полиаденилированной (полиА) фракции РНК из суммарной РНК (за счет взаимодействия синтетического праймера олиго-dT с полиА). Ниже рассмотрены два подхода, применяемых сегодня в NGS: с заранее посаженными на твердую фазу пробами (разд. 11.3.1) или с отбором проб после гибридизации (разд. 11.3.2).

#### **11.3.1. Обогащение гибридизацией на твердой фазе (микрочипе или микросферах)**

Селекцию целевых фрагментов можно проводить путем их гибридизации с олигонуклеотидами на твердой фазе (чипе или микросферах). Для этого заранее подготовленную (амплифицированную) библиотеку фрагментов ДНК гибридизуют с иммобилизованными на чипе (или микрочастицах) олигонуклеотидными пробами, специфичными целевым регионам генома (или транскриптома). Непровзаимодействовавшие фрагменты ДНК удаляют в ходе промывки. Затем целевые фрагменты элюируют и снова амплифицируют методом ПЦР (рис. 11.4).

На этом принципе основана работа ряда коммерческих платформ. Так, компания NimbleGen адаптировала данный метод для NGS. Метод был изначально разработан для секвенатора 454 Life Sciences от Roche, но впоследствии адаптирован для Illumina и других платформ.

Преимуществом обогащения на твердой фазе в сравнении с различными вариантами мультиплексных ПЦР является



**Рис. 11.4.** Принцип обогащения целевыми фрагментами ДНК на твердой фазе (чипе)

более равномерное обогащение и практически отсутствующее ограничение на размер требуемого региона (на одном чипе или частицах можно разместить пробы на весь экзом человека).

В то же время, для работы с чипами требуется дополнительное оборудование, а максимальное число образцов, которое сотрудник может «обогащать» с использованием чипов, редко превышает 24 образца в сутки.

### **11.3.2. Обогащение при помощи гибридизации в растворе с последующим отбором на твердую фазу**

Обогащение библиотеки фрагментов ДНК за счет гибридизации с мечеными олигонуклеотидами в растворе принципиально не отличается от метода обогащения на твердой фазе за тем лишь исключением, что специфические пробы не прикреплены изначально к макрообъекту, а находятся в растворе (что повышает их доступность для фрагментов из библиотеки). После гибридизации с целевыми фрагментами необходи-

мо отобрать пробы из раствора (наиболее распространенный подход – аффинное взаимодействие стрептавидин–биотин).

Примером аппаратной платформы, основанной на этом принципе, является технология Ion TargetSeq (Life Technologies Thermo Fisher Scientific). Библиотеку фрагментов ДНК гибридизуют с коллекцией биотинилированных олигонуклеотидных проб. Пробы отбирают из раствора на стрептавидин, присоединенный к ферромагнитным микросферам. Затем отмывают частицы от «неприлипших» фрагментов и смывают обогащенную библиотеку (рис. 11.5). Методика занимает около 6 ч (в основном аппаратного времени приборов Ion Chef System или Ion OneTouch 2 System). Производитель позиционирует метод для платформ Ion PGM и Ion Proton.



**Рис. 11.5.** Принцип обогащения целевыми фрагментами ДНК гибридизацией в растворе

Аналогичным принципом работы обладают такие коммерческие аппаратные решения, как Sure Select (Agilent Technologies), TrueSeq Enrichment (Illumina) и ряд других.

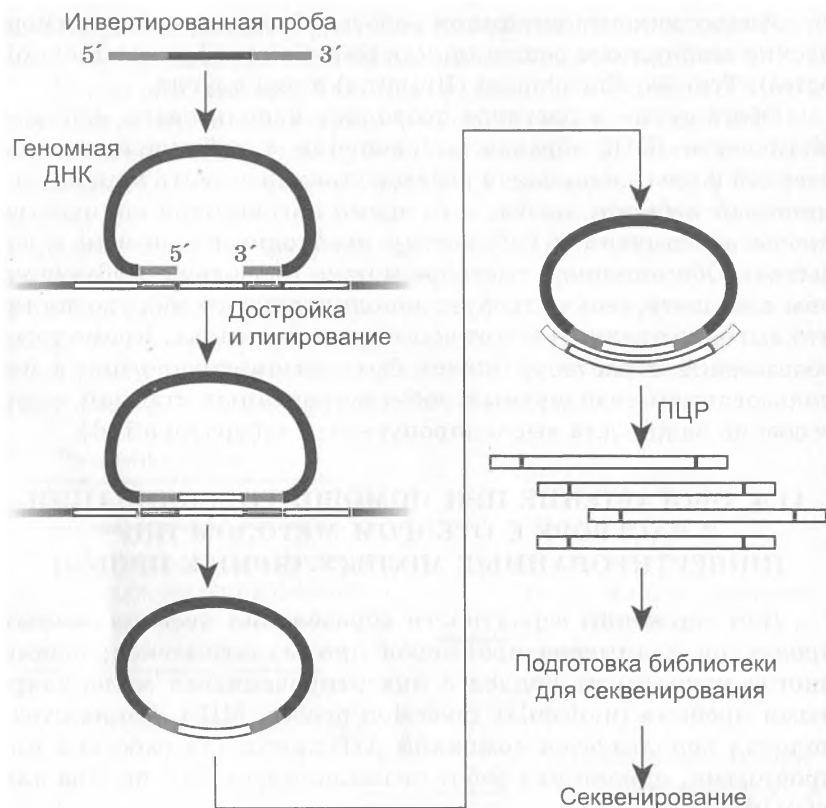
Обогащение в растворе позволяет использовать меньшее количество ДНК образца в сравнении с гибридизацией на твердой фазе (поскольку в растворе можно создать концентрационный избыток пробы, в то время как на чипе количество пробы ограничено и библиотеку необходимо наносить в избытке). Обогащение в растворе можно проводить в 96-луночной планшете, оно не требует дополнительного оборудования, что выгодно отличает его от обогащения на чипах. Кроме того, обогащение в растворе может быть автоматизировано с использованием стандартных роботизированных станций – это особенно важно для высокопропускных лабораторий [6].

#### **11.4. ОБОГАЩЕНИЕ ПРИ ПОМОЩИ ГИБРИДИЗАЦИИ В РАСТВОРЕ С ОТБОРОМ МЕТОДОМ ПЦР (ИНВЕРТИРОВАННЫЕ МОЛЕКУЛЯРНЫЕ ПРОБЫ)**

Для снижения вероятности образования неспецифичных продуктов и димеров праймеров при мультиплексировании иногда используют подход с инвертированными молекулярными пробами (molecular inversion probes, MIP). Данная технология используется компаний Affimetrix для работы с микрочипами, однако ряд работ позиционирует этот подход для NGS [6].

Принцип метода заключается в использовании длинного синтетического одноцепочечного олигонуклеотида, фланкированного последовательностями, комплементарными краям целевого региона и содержащего сайты посадки праймеров и регион со «штрих-кодом» (рис. 11.6).

В одной реакции можно смешивать сотни таких проб с разными флангами. Каждая проба взаимодействует с краями целевого участка последовательности, и ДНК-полимераза достраивает весь регион с 3'-конца пробы («упираясь» в конце пути в 5'-конец той же самой пробы). Затем лигаза ковалентно замыкает синтезированное таким образом кольцо. Незакольцованные фрагменты разрушают с помощью экзонуклеаз, а кольцевые молекулы амплифицируют с универсальными (одинаковыми) праймерами, отжигающимися на пробу. Подход позволяет обогащать до нескольких миллионов пар нуклеотидов.



**Рис. 11.6.** Принцип обогащения таргетными регионами с помощью инвертированных молекулярных проб

Можно выделить сразу несколько достоинств метода:

- 1) как и при обогащении с использованием обычной ПЦР, целевые участки нарабатывают непосредственно с геномной ДНК (без предварительного создания библиотеки), что позволяет снизить необходимое количество ДНК-матрицы до 200 нг;
- 2) возможность «мультиплексировать» сотни ампликонов;
- 3) в пробу можно заложить «штрих-код», что позволяет в дальнейшем объединять разные образцы в пул.

Однако МІР как метод целевого обогащения фрагментов не лишен и недостатков, основные из которых:

- 1) разная эффективность наработки (захвата) целевых участков;
- 2) дороговизна синтеза большой коллекции проб под заказ, что сводит использование метода к предлагаемым компаниями наборам проб под конкретные перечни генов.

### **11.5. ОБОГАЩЕНИЕ БИБЛИОТЕКИ БЕЛОК-СВЯЗЫВАЮЩИМИ ПОСЛЕДОВАТЕЛЬНОСТЯМИ ХРОМАТИНА (CHIP-SEQ)**

Секвенирование иммунопреципитированных элементов хроматина (chromatin immunoprecipitation, ChIP) представляет собой метод анализа ДНК-белковых взаимодействий. ChIP является мощным методом выборочного обогащения последовательностей ДНК, связанных с конкретным белком в живых клетках. До появления технологий NGS широкое использование этого метода было ограничено отсутствием достаточно надежных и недорогих методов определения всех обогащенных последовательностей ДНК. Сочетание иммунопреципитации хроматина с высокопроизводительным секвенированием (ChIP-seq) дало возможность массово определять сайты посадки на геном любого интересующего белка.

Чаще всего ChIP-seq используют для определения местоположения сайтов связывания (на геноме) факторов транскрипции и модифицированных гистонов. Существенным преимуществом метода ChIP является возможность фиксации ДНК-белковых взаимодействий непосредственно в живых клетках, позволяя получить «снимок» этих взаимодействий в экспериментально важный момент.

Алгоритм исследования включает два этапа: иммунопреципитацию и собственно секвенирование полученных фрагментов. На этапе ChIP ковалентно сшитые за счет воздействия формальдегидом или ультрафиолетовым излучением ДНК-белковые комплексы отбирают (обогащают) с использованием антитела против исследуемого белка (метод иммунопреципитации) [7]. Затем к отобраннным таким образом коротким фрагментам ДНК присоединяют адаптеры, амплифицируют и секвенируют при помощи любой платформы NGS [8, 9].

Первые исследования с использованием метода ChIP-seq были посвящены определению локализации транскрипционных факторов и модификации гистонов в Т-клетках человека [8, 9] и в клетках HeLa S3 [10]. ChIP-seq можно использовать и для исследования модификаций молекулы ДНК (в частности – сайтов метилирования).

Преимуществами ChIP-seq в сравнении с обычным вариантом ChIP или технологией ChIP-chip (когда полученные фрагменты гибридизуют с посаженными на микрочип олигонуклеотидами) являются более высокое разрешение метода, возможность идентифицировать сайты связывания в повторяющихся регионах и меньшая стоимость всего исследования [11–13].

\* \* \*

Сравнение методов целевого обогащения последовательностей показывает, что ПЦР на данный момент является безусловным лидером по специфичности и однородности обогащения. В то же время ПЦР не позволяет проводить обогащение протяженных регионов или большого числа участков.

У обогащения гибридизацией и MIP также есть преимущества. Эти методы позволяют работать с большими целевыми регионами. Так, к примеру, для покрытия экзома человека потребуется провести десятки тысяч индивидуальных ПЦР, каждая реакция потребует оптимизации, полученные продукты нужно будет нормализовать перед объединением в пул, а для того, чтобы начать работу потребуется более 100 мкг геномной ДНК. Подобный эксперимент с использованием обогащения гибридизацией займет один экспериментальный день для подготовки библиотеки, и до двух дней уйдет на гибридизацию и элюцию обогащенного материала, а для начала работы потребуется всего несколько нанограмм ДНК.

Сравнение распространенных методов таргетного обогащения по ключевым параметрам представлено в табл. 11.1, а общий принцип выбора технологии обогащения в зависимости от количества образцов и размера целевого региона – на рис. 11.7.

# Сравнение распространенных методов таргетного обогащения по ключевым параметрам

Метод Параметр	ПЦР	Инвертирован- ные пробы	Гибриди- зация на твердой фазе (чипе)	Гибриди- зация в растворе
Цена за единицу длины целевого региона	высокая	больше 100 образцов – низкая; меньше 10 образцов – высокая	средняя	низкая
Простота пробоподготовки	просто	сложно	сложно	средне
Количество исходной ДНК	от нескольких нанограмм до нескольких микрограмм	$\geq 200$ нг	10–15 мкг	$\geq 200$ нг
Воспроизводи- мость	до 100%	~ 92%	> 95%	> 96%
Специфичность обогащения	70–95%	> 98%	~ 70%	~ 80%
Однородность покрытия	высокая	низкая	средняя	средняя

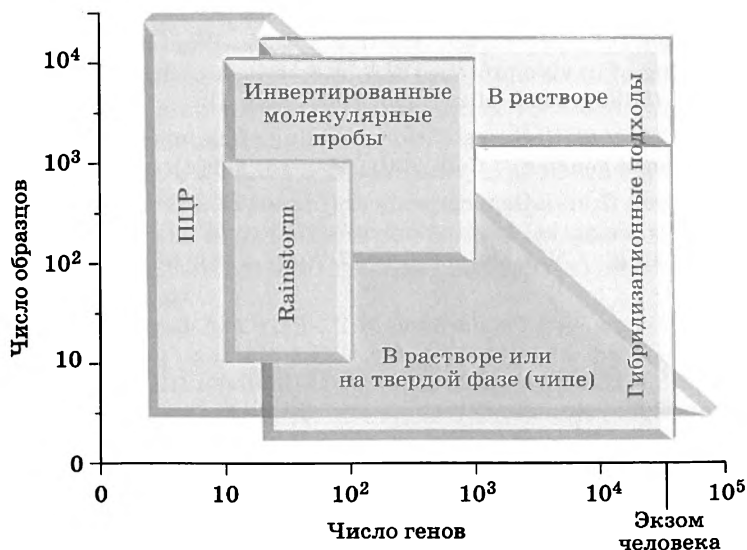


Рис. 11.7. Предпочтительные технологии в зависимости от количества образцов и размера таргетного региона для обогащения

## СПИСОК ЛИТЕРАТУРЫ

1. *Maricic T., Whitten M., Pääbo S.* Multiplexed DNA sequence capture of mitochondrial genomes using PCR products // *PLoS One*, 2010, 5 (11): e14004.
2. *Mejia-Leon M.E. et al.* Fecal microbiota imbalance in Mexican children with type 1 diabetes // *Sci Rep.*, 2014, 4: 3814.
3. *Faiz F., Allcock R.J., Hooper A.J., van Bockxmeer F.M.* Detection of variations and identifying genomic breakpoints for large deletions in the LDLR by Ion Torrent semiconductor sequencing // *Atherosclerosis*, 2013, 230 (2): 249–255.
4. *Matz M. et al.* Amplification of cDNA ends based on template-switching effect and step-out PCR // *Nucleic Acids Res.*, 1999, 27 (6): 1558–1560.
5. *Tewhey R. et al.* Microdroplet-based PCR enrichment for large-scale targeted sequencing // *Nature Biotech.*, 2009, 27: 1025–1031.
6. *Teer J.K. et al.* Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing // *Genome Res.*, 2010, 20 (10): 1420–1431.
7. *Solomon M.J., Larsen P.L., Varshavsky A.* Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene // *Cell*, 1988, Jun 17; 53(6): 937–47.
8. *Johnson D.S., Mortazavi A., Myers R.M., Wold B.* Genome-wide mapping of in vivo protein-DNA interactions // *Science*, 2007, Jun 8; 316 (5830): 1497–502. Epub, 2007, May 31.
9. *Barski A. et al.* High-resolution profiling of histone methylations in the human genome // *Cell*, 2007, May 18; 129(4): 823–37.
10. *Robertson G. et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing // *Nat Methods*, 2007 Aug; 4 (8): 651–7. Epub 2007 Jun 11.
11. *Kharchenko P.V., Tolstorukov M.Y., Park P.J.* Design and analysis of ChIP-seq experiments for DNA-binding proteins // *Nat Biotechnol*, 2008, Dec; 26 (12): 1351–1359. doi: 10.1038/nbt.1508. Epub, 2008, Nov 16.
12. *Bourque G. et al.* Evolution of the mammalian transcription factor binding repertoire via transposable elements // *Genome Res*, 2008 Nov; 18 (11): 1752–1762.
13. *Hoffman B.G., Jones S.J.* Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing // *J Endocrinol*, 2009, Apr; 201 (1): 1–13.

# **ПРИМЕНЕНИЕ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ В МЕДИЦИНСКОЙ ПРАКТИКЕ**

Переход любого нового метода из научно-исследовательской лаборатории в клиническую практику, как правило, занимает годы. Необходимо не только создать надежные решения, с которыми справится обычный лаборант, но и зарегистрировать их в соответствующих инстанциях в качестве медицинской технологии или медицинского изделия (а часто – и то, и другое). Нередко метод «добирается» до клиники лишь через 10–15 лет после изобретения (как это было, скажем, для полимеразной цепной реакции).

Однако благодаря своей революционности технологии NGS уже активно внедряются в практическую медицину. В ближайшее время можно ожидать появления широкого спектра разрешенных к использованию в медицинской практике методов диагностики, основанных на высокопроизводительном секвенировании.

В качестве наиболее очевидных направлений применения NGS в медицине являются различные варианты генетического тестирования и исследование микробиоценозов человека. Данная глава посвящена обзору вариантов использования NGS в медицине.

## **12.1. ГЕНЕТИЧЕСКОЕ ТЕСТИРОВАНИЕ С ИСПОЛЬЗОВАНИЕМ NGS**

Генетическое тестирование может потребоваться отдельному пациенту или семье в целом ряде случаев, но обычно показанием к исследованию служит наличие симптомов генетически обусловленного заболевания или же наличие родственников, страдавших таким заболеванием. Тестирование

может быть диагностическим, прогностическим или проводится для определения статуса носителя наследственного заболевания. По возрасту пациента и причине проведения анализа генетические исследования можно разделить на следующие группы:

- 1) преимплантационное (зная, что у будущих родителей высок риск передачи наследственного заболевания ребенку, врач (обычно в рамках экстракорпорального оплодотворения – ЭКО) может назначить преимплантационную генетическую диагностику (ПГД) эмбриона, чтобы избежать рождения больного ребенка);
- 2) пренатальное (пара, ожидающая рождения ребенка, узнает, что у плода высок риск развития определенного заболевания, и решает проверить плод на наличие данного генетического заболевания);
- 3) неонатальное (новорожденный ребенок может быть подвержен генетическому исследованию, чтобы поставить диагноз или составить прогноз);
- 4) детское, педиатрическое (ребенок с отставанием в развитии может быть подвергнут генетическому исследованию в целях постановки диагноза);
- 5) взрослое, терапевтическое (пациент, имеющий в истории семьи родственников, страдавших определенным заболеванием, желает выяснить, имеет ли он генетическую предрасположенность к развитию данного вида заболевания (чаще всего применяют в области онкологии)).

Использование генетической диагностики в клинической практике позволяет применять оптимальные схемы лечения и повышать процент выздоравливающих пациентов. В ряде случаев генетические исследования позволяют в значительной степени снизить смертность и заболеваемость вследствие своевременного врачебного вмешательства (например, в результате колоноскопии и удаления полипов во избежание развития наследственного неполипозного рака толстой и ободочной кишки или превентивной мастэктомии для снижения риска развития наследственного рака молочной железы). Даже в тех случаях, когда заболевание неизлечимо, молекулярно-генетическая диагностика позволяет принять своевременное решение при планировании семьи.

### 12.1.1. Диагностика наследственных заболеваний

В настоящее время в клинике широко применяют два типа генетических исследований: таргетная генетическая диагностика (проверяющая наличие конкретных генетических особенностей) и полногеномное исследование. Первый подход включает в себя генотипирование для определения состояния полиморфных позиций или соматических мутаций известной локализации, направленного анализа вставок, делеций или дупликаций (например, экспансии триплетных повторов), секвенирование отдельных генов или протяженного хромосомного региона [1]. Полногеномные исследования применяют при диагностике заболеваний с нечеткой клинической картиной, в случае вероятности развития заболевания из-за мутаций в нескольких генах или при необходимости проведения высокоточного кариотипирования [2].

Специалист в области клинической генетики для постановки диагноза сначала изучает фенотип пациента, а затем проводит проверку наличия предполагаемого генетического нарушения. Однако в ближайшем будущем можно прогнозировать рост направлений на полногеномное исследование высокого разрешения, что позволяет без жесткой привязки к фенотипу выявлять мутации, приводящие к редким или неочевидным фенотипам.

Чтобы подчеркнуть преимущества высокопроизводительного секвенирования перед традиционным секвенированием по Сенгеру, можно упомянуть два основных направления диагностического применения NGS: мультигенные панели для диагностики генетически гетерогенных заболеваний и выявление молекулярно-генетической природы заболеваний, имеющих только клиническое (феноменологическое) описание. В подобных случаях поочередная проверка наличия мутаций каждого вовлеченного гена с использованием секвенирования по Сенгеру нецелесообразна (в том числе по экономическим соображениям).

Для изучения молекулярных основ развития наследственных заболеваний высокопроизводительное секвенирование стало широко применяться с 2009 года, появилось множество публикаций об использовании экзомного секвенирования для диагностики редких наследственных заболеваний [3]. Такой

подход позволяет определить последовательность кодирующих участков (экзонов) всех генов человека (без определения ненужных в данном случае некодирующих регионов генома). Подход оказался особенно эффективен в случае редких или спорадических заболеваний и состояний, влияющих на фертильность, которые практически не диагностируются традиционными методами, такими как позиционное клонирование и анализ сцепления, из-за отсутствия достаточно большого числа родственников.

Хотя до широкого применения полногеномного и экзомного секвенирования в клинической практике еще не дошло, польза от внедрения этих технологий очевидна как при диагностике, так и при выборе схемы лечения пациента.

В качестве примера можно привести синдром Элерса–Данлоса. Дисплазия соединительной ткани – достаточно обширная группа заболеваний, характеризующихся множественными дефектами волокнистых структур и основного вещества соединительной ткани. Нарушения соединительной ткани приводят к расстройству гомеостаза на тканевом и организменном уровнях в виде различных морфофункциональных нарушений в висцеральных и локомоторных органах. Существует несколько подвидов соединительнотканной дисплазии. Дифференцированные дисплазии – основная группа, на которую направлено секвенирование экзома. Это синдромы Элерса–Данлоса, Марфана, Стиклера и ряд других.

В геноме человека обнаружено порядка 50 генов, кодирующих коллагены. В настоящее время известно 27 различных типов коллагеновых белков, нарушения в структуре которых приводят к различным дисплазиям соединительной ткани. Точную причину большого количества нарушений соединительной ткани установить практически невозможно ввиду того, что клиническая картина часто очень сходна, а генетическая природа – различна.

Так, существует 6 типов синдрома Элерса–Данлоса, которые клинически очень схожи. В работе Шаталова с соавт. [4] в результате секвенирования экзома у обследованного пациента была обнаружена мутация в гене *COL45A1* (замена в 1062-й позиции – аргинин на стоп-кодон, что, согласно данным предыдущих исследований, приводит к нарушению альфа-1 цепи коллагена V типа). На основании этих данных ребенку был поставлен диагноз синдрома Элерса–Данлоса

I типа и проведена генная терапия. В настоящее время пациент здоров.

Другой пример – атипичный гемолитико-уремический синдром, генетическое заболевание, которое по своей природе является прогрессирующим. Наблюдается крайне высокий риск внезапной смерти и необратимое индивидуальное повреждение жизненно важных органов, таких как почки, печень, сердце и головной мозг. Основными симптомами данного заболевания являются гемолитическая анемия с наличием фрагментированных эритроцитов (шизоцитов) и тромбоцитопения. Существует множество причин возникновения атипичного гемолитико-уремического синдрома. Вероятность его появления во всех случаях относительно невелика, что значительно затрудняет дифференциальную диагностику заболевания. Существующие диагностические методы позволяют исследовать причины возникновения заболевания, но по отдельности, что малоэффективно. Секвенирование экзона в данной ситуации позволяет исследовать все возможные причины одновременно. В работе Шаталова с соавторами [4] использовали секвенирование экзона пациента, результат которого позволил выявить мутацию в участке G11194D гена *CFH*, которая и являлась причиной атипичного гемолитико-уремического синдрома. На основании полученных данных пациенту была успешно проведена терапия.

### 12.1.2. Скрининг на статус носителя наследственных заболеваний

Генетический скрининг будущих родителей позволяет определить, являются ли они носителями нежелательных вариантов генов, что в свою очередь может повлиять на стратегию планирования семьи. Преимуществом такого подхода является свобода выбора пациента между решением иметь приемных или собственных детей, а в последнем случае выбирать между преимплантационной и пренатальной диагностикой (с возможностью прерывания беременности) или принятием того факта, что ребенок может иметь наследственные заболевания.

Такой скрининг рекомендуют как пациентам, входящим в группу риска по конкретному заболеванию, так и внешне здоровым лицам. На социальном уровне выделяют два (по при-

роде своей однотипных) фактора, влияющих на наследование аутосомно-рецессивных признаков генетического заболевания: близкородственные браки и одинаковая этническая принадлежность. Близкородственными принято называть браки между двоюродными братьями и сестрами или лицами, состоящими в еще большем родстве. По оценкам до 10% населения Земли состоят в близкородственном браке или являются детьми, рожденными в таком браке. Несмотря на некоторые социальные и экономические преимущества, близкородственных браках увеличивает вероятность передачи аутосомно-рецессивных заболеваний потомству, причем чем реже заболевание, тем сильнее влияет родство родителей на вероятность появления этого заболевания у потомства.

Некоторые аллели чаще встречаются у представителей определенных этнических групп из-за географических особенностей местности или вследствие эффекта основателя. Например, муковисцидоз больше распространен в Западной Европе, Средиземноморье и на Ближнем Востоке, в то время как серповидноклеточная анемия чаще встречается у лиц африканского происхождения, а талассемия – среди населения Средиземноморья, Ближнего Востока и в Южной Азии [5, 6].

Хотя на уровне государств пока запущено лишь несколько программ генетического скрининга родителей перед зачатием, те, что существуют, нацелены на выявление заболеваний, наиболее распространенных среди представителей данной этнической группы. Ряд стран Ближнего Востока и Средиземноморья реализуют программу по выявлению носителей талассемии. Например, в Иране, где генетический скрининг является частью обязательного добрачного исследования, рождаемость младенцев, наделенных этим заболеванием, снизилась на 70% [7]. В Канаде скрининг на статус носителя болезни Тея-Сакса у евреев-ашкенази позволил снизить рождаемость детей с данным заболеванием более чем на 90% [8].

Помимо государственных программ, существует возможность пройти генетическое исследование в индивидуальном порядке, как в отношении конкретного заболевания, так и целого ряда (панели) заболеваний. Например, британская организация Jewish Care предлагает панель ашкенази, которая позволяет выявить носительство синдрома Блума, кистозного фиброза, болезней Гоше и Тея-Сакса. Тесты на статус носи-

теля более 50 наиболее частых наследственных заболеваний предлагают многие компании, в том числе 23andMe, Counsyl, Pathway Genomics, Генотек. Согласно статистике, каждый второй человек является носителем одного или нескольких частых наследственных заболеваний. Однако пары, где оба родителя являются носителями одного негативного признака (а именно это является риском для потомства), встречаются только в 0,8–1,0% случаев.

Хотя секвенирование полного генома является всеобъемлющим тестом на носительство наследственных заболеваний, обработка данных (с игнорированием непатогенных или ошибочных полиморфных позиций) может оказаться довольно трудной задачей. Более того, в случае редких заболеваний ошибки секвенирования будут встречаться чаще, чем собственно полиморфизм. Тем не менее технологии NGS создают беспрецедентные возможности проведения полного скрининга на носительство наследственных заболеваний.

### 12.1.3. Пренатальный скрининг

Программа пренатальной диагностики в России сегодня включает проверку плода на наличие анеуплоидий: трисомии по 21-й хромосоме – синдром Дауна, по 13-й – синдром Патау, по 18-й – синдром Эдвардса (приказ Минздравсоцразвития РФ от 28.12.2000 и приказ департамента здравоохранения города Москвы от 04.04.2005 № 144). Действующий протокол включает несколько этапов, в том числе анализ крови матери на различные биохимические маркеры и ультразвуковое исследование плода. При наличии показаний к инвазивному исследованию генетический материал плода используют для проверки на анеуплоидии кариотипированием. Несмотря на значительные улучшения в методике проведения пренатальных исследований, определенный процент ответов является ложноположительным. Кроме того, использование инвазивных методов в 1–2% случаев может приводить к выкидышу [9]. Очевидно, что проведение неинвазивного исследования является более перспективной и предпочтительной методикой.

Множество работ по изучению циркулирующей в кровотоке матери внеклеточной ДНК плода (cell-free fetal DNA, cffDNA) показали возможность диагностики ряда заболева-

ний плода (синдрома Дауна и некоторых других хромосомных нарушений) по крови матери. ДНК плода попадает в плазму крови беременной женщины из клеток плаценты в результате апоптоза. Концентрация cffDNA значительно варьирует у пациентов, но в большинстве случаев она может быть обнаружена в плазме матери уже на пятой неделе беременности (составляя около 3–7% от всей свободной циркулирующей (внеклеточной) ДНК и достигая 25–30% к середине второго триместра) [10].

Существует несколько методов количественного измерения cffDNA с определением снижения или повышения доли зародышевой ДНК, относящейся к конкретной хромосоме [11]. Возможно, самый многообещающий метод – это секвенирование всей внеклеточной ДНК плазмы матери с использованием технологий NGS. По числу прочтений, относящихся к интересующей хромосоме относительно референсной хромосомы, можно определить случаи анеуплоидии. Поскольку ДНК плода представляет собой лишь небольшую фракцию внеклеточной ДНК плазмы матери, для детекции небольших отклонений в количестве ДНК различных хромосом зародыша необходимо большое покрытие (число прочтений каждого региона).

Анализ циркулирующей в плазме матери ДНК плода позволяет не только проводить исследование на анеуплоидии, но и определять пол будущего ребенка, выявлять небольшие хромосомные перестройки, определять «дозу гена» и т. п.

По предварительным данным, анализ cffDNA является довольно точным методом, однако в будущем необходимо для каждого варианта использования установить его чувствительность, специфичность, клиническую эффективность, надежность и другие характеристики. В 2012 году данный метод был впервые одобрен в США в качестве дополнения к действующим протоколам и назначается только по показаниям (высокий риск появления ребенка, страдающего наследственным заболеванием) либо по желанию пациента.

Применение NGS для секвенирования всей внеклеточной ДНК материнской плазмы (в сумме ДНК и плода, и матери) позволяет диагностировать практически любое наследственное заболевание. Несмотря на то что ДНК плода представляет собой лишь минорную часть исследуемого образца, полногеномная карта плода может быть составлена на основе сравне-

ния генотипов родителей. Отцовские аллели в геноме плода могут быть обнаружены напрямую, если они отличаются от материнских, в остальных случаях аллели плода могут быть детектированы по изменению «дозы гена» из-за присутствия ДНК плода. Составление генетической карты позволяет генотипировать плод по интересующему локусу.

На сегодняшний день была показана принципиальная возможность проведения пренатальной диагностики этим методом на ряде примеров в США. В России также существует коммерческое пренатальное экзомное секвенирование (Генотек). Однако техническая сложность процедур и высокая стоимость проведения такого исследования все еще ограничивают широкое внедрение этого метода в клиническую практику.

#### **12.1.4. Генетический скрининг новорожденных**

В ряде стран, в том числе и в России, действуют программы скрининга новорожденных. Как правило, для исследования берут образец крови из пятки ребенка и проводят как биохимический анализ, так и проверку носительства наследственных заболеваний (приказ Минздравсоцразвития РФ от 22.03.2006 № 185 «О массовом обследовании новорожденных детей на наследственные заболевания»).

Удешевление технологий NGS в ближайшем будущем, вероятно, приведет к определению полного генома каждого новорожденного ребенка, что заменит традиционный скрининг. Однако не только технологические и финансовые трудности стоят на пути применения такого подхода. Сейчас не вполне понятно, какие этические, социальные и правовые последствия может повлечь за собой массовое секвенирование геномов всех людей.

#### **12.1.5. Использование секвенирования в онкологии**

Известно, что опухоли возникают в результате накопления мутаций в соматических клетках, приводящих к нарушению клеточных механизмов регуляции. Вероятность появления критического набора мутаций увеличивается с возрастом (резкий рост количества онкологических заболеваний наблюдается в группе людей старше 50 лет). По разным оценкам, около 80–90% мутаций, приводящих к развитию опухолей,

имеют ненаследственный характер и не передаются потомству. Однако изучение генома клеток конкретной опухоли позволяет точно классифицировать опухоль и понять, как шло развитие именно этого вида рака.

Ранние представления о связи генома с развитием опухоли базировались на наблюдении хромосомных aberrаций в некоторых раковых клетках. Эти наблюдения также позволили впервые предположить, что опухолевые клетки являются потомками (клонами) здоровых клеток, имеющими нарушения в наследственной программе. Гипотеза нашла подтверждение в экспериментах по пересадке геномной ДНК раковых клеток человека в мышинные клеточные линии, что приводило к превращению фенотипически нормальных клеток *NIH3T3* в опухолевые. Так, в 1982 году был открыт первый человеческий онкоген – *HRAS* [12]. Впоследствии были открыты многие другие онкогены, а также гены, подавляющие развитие опухолей, – супрессоры (tumour suppressor, TS).

Изменения в геноме, связанные с развитием опухолей, затрагивают гены, контролирующие клеточный цикл и смерть (апоптоз) или участвующие в процессах репарации (восстановление и защита геномной ДНК). На сегодняшний день известно порядка 400 таких генов. Онкологические заболевания называют спорадическими, так как для их развития необходимо сочетание ряда факторов, как генетических (наследственная предрасположенность), так и особенностей образа жизни, питания и окружающей среды. В первом приближении гены, вовлеченные в развитие онкологических заболеваний, можно разделить на онкогены и гены-супрессоры (TS). Примерно 90% онкологических заболеваний связаны с работой доминантных онкогенов, т. е. достаточно мутации в одной копии гена (из двух), чтобы клетка стала раковой. Оставшаяся доля связана с рецессивными мутациями в TS-генах, т. е. мутированы должны быть обе копии гена (ген перестает функционировать). Поскольку мутация в одной копии гена-супрессора не приводит к развитию рака, такие мутации могут передаваться по наследству.

В норме онкогены контролируют клеточный цикл и апоптоз. Мутации нарушают их функцию, приводя к неконтролируемому делению клеток и отсутствию ответа на сигналы апоптоза, что и приводит к развитию опухоли. Выделяют пять классов онкогенов: факторы роста, рецепторы факторов

роста, трансдукторы, транскрипционные факторы и регуляторы апоптоза. Поскольку пути активации онкогенов разнообразны, только секвенирование полного генома соматической клетки дает возможность гарантированно обнаружить все возможные изменения.

Гены, предотвращающие развитие опухолей, отвечают за блокировку неконтролируемого клеточного роста и деления, запуская апоптоз, или защищают геном от повреждений. Мутации, вызывающие потерю функций этих генов, приводят к неконтролируемой пролиферации и, как следствие, развитию рака (например, гены *RB1*, *TP53* и *MLH1*). Гены-супрессоры также могут выключаться в результате эпигенетических изменений: метилирования, изменения структуры хроматина и регуляции экспрессии (*CDKN2A*, *RB1* и *MLH1*). Тотальное гиперметилование генома характерно для многих видов опухолевых клеток. Обычные методы секвенирования не позволяют детектировать эпигенетические изменения, для этого есть специальные протоколы NGS, хотя некоторые технологии могут обнаруживать метилирование напрямую [13].

В течение жизни человека его клетки накапливают ряд изменений в геноме, начиная от точечных мутаций, инсерций, делеций и инверсий на генном уровне и заканчивая более крупными геномными и хромосомными перестройками. Набор изменений в геноме раковой клетки по сравнению с геномом хозяина можно считать генетическим отпечатком пальца каждой конкретной опухоли. По оценкам, для злокачественной трансформации необходимо от 3 до 12 мутаций в зависимости от вида рака. Раннее выявление каждого изменения в геноме опухоли еще несколько лет назад требовало отдельного анализа, сейчас же NGS-технологии позволяют проанализировать все мутации генома опухолевой клетки в одном секвенировании экзона. Мутации в некоторых генах встречаются в клетках многих видов опухолей (*KRAS* и *TP53*), но есть и опухоль-специфичные (*RB1*). Для отдельных форм рака практически нет данных об онкогенах, однако даже в этих случаях возможно выявление индивидуальных биомаркеров для каждого пациента, что позволяет оценивать ответ опухоли на терапию.

С появлением нового взгляда на рак как на геномное заболевание генетическое тестирование все чаще применяют при

постановке диагноза, выборе схемы лечения и составления прогноза. Микрочипы и секвенирование по Сенгеру позволили лучше понять особенности «ракового» генома, технологии NGS открывают перед исследователями и врачами еще более широкие возможности.

Однако при работе с раковыми геномами возникает ряд специфических трудностей. Образцы для выделения ДНК при изучении генома опухоли сильно отличаются от тех, что обычно используют при изучении геномной ДНК. Биоптат, как правило, представляет собой небольшой фрагмент ткани, содержащий ограниченное количество ДНК. Качество ДНК снижается из-за того, что ткань фиксируют в формалине или заливают парафином. В то же время низкое качество ДНК может быть следствием некротизации опухоли. В образце часто присутствуют примеси здоровых клеток (фибробласты, лимфоциты). Тем не менее многократное покрытие генома при использовании методов NGS позволяет преодолеть эти трудности. NGS также можно использовать для изучения других образцов пациента (кровь, моча, фекалии).

Опухоли зачастую генетически гетерогенны. Более того, внутри одной опухоли могут находиться генетически различные клоны. Благодаря многократному покрытию генома, которое достигается при использовании NGS, даже редкие варианты изменений могут быть детектированы. Однако опухолевый геном нестабилен и продолжает меняться и после исследования генома биоптата. В случаях рецидивирующих опухолей, появления лекарственной резистентности или метастазов ресеквенирование опухолевого генома может помочь скорректировать схему лечения.

Анализ опухолевого генома имеет ряд особенностей:

- 1) необходимость одновременно анализировать геном пациента и опухоли и сравнивать их с референсным геномом;
- 2) необходимость выявлять многочисленные мутации и геномные перестройки. Отличия опухолевого генома от референсного могут быть так велики, что оптимальным решением может оказаться сборка генома *de novo*;
- 3) необходимость работать со смешанным образцом генома пациента и различных клонов опухоли сильно осложняет процесс определения соматических мутаций.

### 12.1.6. Перспективные направления генетических исследований в медицинской практике

Основной тенденцией в медицине сейчас является переход от лечения болезни к ее предупреждению, поэтому модели, предсказывающие риск развития заболеваний, вызывают все больший интерес. Большинство социально значимых заболеваний (диабет второго типа, инсульт, ишемическая болезнь сердца, рак молочной железы, прямой кишки и простаты) имеют сложную этиологию. Риск развития таких заболеваний у конкретного человека зависит от стиля жизни, семейной истории, физических нагрузок и наследственности. Существующие математические модели учитывают все эти факторы при оценке рисков. Открытие новых молекулярно-генетических маркеров позволит улучшить точность предсказаний, поэтому одно из важных потенциальных приложений полногеномного секвенирования – улучшение методов оценки риска развития комплексных хронических заболеваний.

Генетическая информация более или менее стабильна на протяжении жизни пациента (в отличие от биохимических маркеров), поэтому однократное составление полногеномного профиля пациента позволит использовать полученную информацию для оценки рисков в течение всей жизни. Кроме того, создание геномных профилей большой группы людей позволит выявить ранее неизвестные генетические маркеры. Создание полногеномного профиля пациентов позволит разделить их на группы риска, согласно генетической предрасположенности, и тем самым минимизировать количество людей, подвергающихся неприятному или дорогостоящему исследованию. Сейчас подобную доскрининговую классификацию пациентов применяют при назначении маммограммы, в качестве фактора риска используют возраст. Однако на сегодняшний день известен целый ряд локусов, вовлеченных в развитие рака молочной железы.

Гены могут влиять на ответ организма на то или иное лекарство – через изменение его абсорбции или метаболизма либо вследствие модификации терапевтической мишени. Достижения в области генетических исследований позволяют персонализировать медицину: геномные профили пациентов будут определять выбор препарата и схему лечения. Это, в свою очередь, повлечет за собой создание персонализиро-

ванных лекарств. Традиционно клинические испытания новых препаратов проводили на большой группе добровольцев, случайно отобранных из популяции. Усредненный эффект, которой оказывал препарат на эту группу, использовали в качестве отправной точки для применения в клинике. В рамках концепции фармакогенетики геномная информация каждого пациента используется для оценки персонального ответа на лекарство, выбора препарата и его дозы.

Исторически фармакогенетика ограничивалась изучением взаимосвязи узкого круга генетических маркеров с ответом на определенные препараты. Основной трудностью в фармакогенетических исследованиях всегда являлось создание корректной выборки, что довольно трудно, так как побочные эффекты чаще проявляются у пациентов с редким генотипом. После выхода препарата на рынок только добровольные сообщения врачей и пациентов могут служить источником подобной информации. До появления NGS в фармакогенетических исследованиях использовали подход «генов-кандидатов». Суть его состоит в том, что исходя из априорного предположения о терапевтической мишени и метаболизме препарата для исследования отбирали очень небольшое число генов. Такой подход дает хорошие результаты, но только в том случае, если начальные гипотезы были верны.

Современные технологии секвенирования позволяют отказать от старого подхода и перейти к поиску ассоциаций по всему геному (GWAS). Это не требует априорных знаний о генах, влияющих на эффект от исследуемого препарата. В то же время накопление данных о вредных мутациях и улучшающиеся алгоритмы предсказания третичной структуры белков делают возможным целенаправленное создание лекарств. Технологии высокопроизводительного секвенирования позволяют получать генетическую информацию о полных путях метаболизма конкретного препарата.

## 12.2. ИССЛЕДОВАНИЕ ПАТОГЕНОВ И МИКРОБИОМА ЧЕЛОВЕКА

Отдельным направлением применения технологий NGS являются инфекционные заболевания. Технологии NGS позволяют секвенировать микробный геном за несколько часов,

что может быть крайне полезно в случае эпидемии. NGS уже использовались для мониторинга распространения метициллин-устойчивого штамма *Staphylococcus aureus*, секвенирования генома *E. coli* после массовых случаев инфекции в Европе. Возможно, вскоре NGS станет золотым стандартом эпидемиологии.

Еще одним направлением применения NGS в медицине является исследование симбиотических микробных сообществ человека. С пониманием важности состояния микробиоценозов для здоровья пациента исследователи и врачи начали активно внедрять методы оценки их состояния. Сейчас для этого активно используют подходы на основе количественной ПЦР (например, набор реагентов «ФЕМОФЛОР», компания «ДНК-технология»). Однако удешевление методов NGS делает данный подход пригодным для экспресс-анализа сложных сообществ в диагностических целях.

\* \* \*

Таким образом, методы высокопроизводительного секвенирования уже сейчас находят активное применение в диагностике наследственных заболеваний, выявлении рисков мультифакторных заболеваний, определении статуса носителя рецессивных заболеваний и в пренатальной диагностике.

Отдельного внимания заслуживает использование секвенирование в онкологии, что позволяет выявлять мутации, ставшие причиной злокачественной трансформации, а в некоторых случаях – предсказывать эффективные пути терапии.

Ввиду быстрого развития технологий секвенирования далеко не все экспериментальные методики сейчас доступны в клинической практике, однако целая плеяда перспективных методов проходит клинические испытания во многих странах и появляется на локальных рынках.

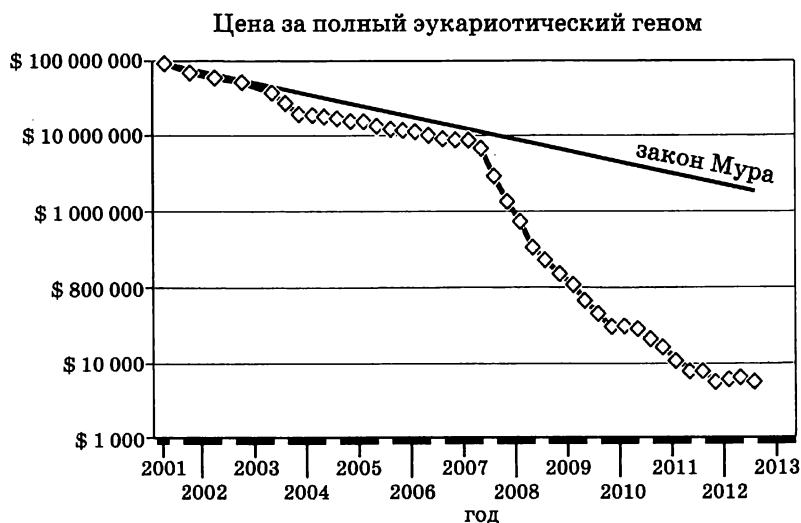
## СПИСОК ЛИТЕРАТУРЫ

1. *Mehdizadeh H.A. et al.* Analysis of CFTR Gene Mutations in Children with Cystic Fibrosis, First Report from North-East of Iran // Iran J Basic Med Sci., 2013, 16 (8): 917–921.
2. *Worthey E.A.* Analysis and annotation of whole-genome or whole-exome sequencing-derived variants for clinical diagnosis // Curr Protoc Hum Genet., 2013, 79: 9.24.1–9.24.24.
3. *Atwal P.S. et al.* Clinical whole-exome sequencing: are we there yet? // Genet Med., 2014, Feb 13. doi: 10.1038/gim.2014.10.2014.
4. *Шаталов П.А.* Московский НИИ Педиатрии и Детской хирургии МЗ РФ, компания Генотек: из доклада на конференции NGS-2013, ИБХ РАН, Москва, 2013.
5. *van der Ven K. et al.* Cystic fibrosis mutation screening in healthy men with reduced sperm quality // Hum Reprod., 1996, 11(3): 513–517.
6. *Bernadette M., Matthew D.* Global epidemiology of haemoglobin disorders and derived service indicators // Bull World Health Organ., 2008, 86 (6): 480–487.
7. *Abolghasemi H. et al.* Thalassemia in Iran: epidemiology, prevention, and management // J Pediatr Hematol Oncol., 2007, 29 (4): 233–238.
8. *Mitchell J.J., Capua A., Clow C., Scriver C.R.* Twenty-year outcome analysis of genetic screening programs for Tay-Sachs and beta-thalassemia disease carriers in high schools // Am J Hum Genet., 1996, 59 (4): 793–798.
9. *Corrado F. et al.* Pregnancy outcome following mid-trimester amniocentesis // J Obstet Gynaecol., 2012, 32 (2): 117–119.
10. *Zheng J. et al.* Effective noninvasive zygosity determination by maternal plasma target region sequencing // PLoS One, 2013, 8 (6): e65050.
11. *Liao G.J., Gronowski A.M., Zhao Z.* Non-invasive prenatal testing using cell-free fetal DNA in maternal circulation // Clin Chim Acta, 2014, Jan20; 428: 44–50.
12. *Taparowsky E. et al.* Activation of the T24 bladder carcinoma transforming gene is linked to a single amino acid change // Nature, 1982, 300 (5894): 762–765.
13. *Feng Z. et al.* Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic // PLoS Comput Biol., 2013, 9 (3): e1002935.

## ПЕРСПЕКТИВЫ ВЫСОКОПРОИЗВОДИТЕЛЬНОГО СЕКВЕНИРОВАНИЯ

За последние два десятилетия технологии секвенирования позволили на много порядков снизить стоимость и время получения данных о последовательностях ДНК и РНК. Существующие сегодня технологии позволяют одному исследователю за несколько дней решать задачи, которые два десятилетия назад потребовали бы усилий нескольких научных институтов и годы работы.

Стоимость секвенирования генома человека с появлением первых технологий NGS снижалась быстрее закона Мура на протяжении 2007–2011 года (рис. 13.1) [1]. Столь значительное падение в цене объясняется не только совершенствова-



**Рис. 13.1.** Динамика изменения стоимости секвенирования генома человека в 2001–2013 годах

нием конструкций приборов и реагентов, но и значительной конкуренцией среди производителей, что вынуждало их прибегать к демпингу. Однако к 2011–2012 году сложилась ситуация, в которой ключевые технологии секвенирования уже были выведены на рынок по «справедливой» цене, а новые технологии еще не получили достаточного развития, чтобы стать массовыми.

В настоящее время ближайшим ценовым рубежом для производителей секвенаторов является достижение психологически значимой отметки в \$1000 за полный эукариотический геном. В январе 2012 года Life Technologies презентовала секвенатор Ion Proton, якобы позволяющий к концу 2012 года секвенировать геном человека за \$1000. Однако до сих пор эта отметка по сути не достигнута, а цена зафиксировалась на уровне \$4000–5000 при выполнении проектов со значительным количеством образцов.

В начале 2014 года Illumina выступила с аналогичным обещанием в ходе презентации секвенатора HiSeq X. Однако, несмотря на очевидную общую тенденцию к снижению стоимости секвенирования (пусть и не экспоненциальную в последние годы), причин для скептицизма достаточно, ведь по сути компании развивают старые технологии, стараясь выжать из них максимум.

Можно утверждать, что в течение ближайших 3–5 лет на рынке произойдет замещение существующих технологий еще более производительными. Подтверждением этому является снятие с 2016 года с производства платформы 454 Life Sciences от Roche, первой массовой технологии NGS. Взамен Roche поглотила целый ряд молодых компаний, развивающих альтернативные подходы к секвенированию и активно сотрудничает с Pacific Biosciences. Сама Pacific Biosciences представила обновленные секвенаторы со значительно более высоким качеством прочтений, которые, однако, все еще достаточно дороги, чтобы стать массовыми.

Нет сомнений, что в ближайшее время отметка в \$1000 за геном человека будет взята. Однако открытым остается вопрос о сроках достижения отметок в \$100 и пока невероятные \$10 за полный геном человека с использованием любой из разрабатываемых сегодня платформ (см. гл. 1).

Для метода NGS, как и для большинства других сложных научных методов, прослеживается тенденция к отказу

от владения оборудованием в отдельных лабораториях в пользу заказа услуг в специализированных сервисных центрах. По разным оценкам, уже сегодня центры секвенирования по всему миру владеют более 50% секвенаторов и производят до 80% всех геномных данных<sup>1</sup>. Причина подобной тенденции заключается в довольно высокой стоимости NGS-секвенаторов и минимизации расходов при увеличении объема заказов (запускать секвенирование множества образцов на высокопроизводительном оборудовании в непрерывном режиме гораздо выгоднее поддержания возможности периодических запусков в собственной лаборатории). Важно также учитывать, что секвенаторы стремительно устаревают и требуют постоянных вложений в модификации и улучшения. Для небольших научных групп, не использующих секвенирование постоянно, гораздо выгоднее провести современные исследования в сервисном центре.

Накопление и обработка данных – вторая основная проблема развития NGS.

Согласно статистике GenBank, количество данных о последовательностях нуклеиновых кислот растет экспоненциально на протяжении последних 30 лет (рис. 13.2). К этому добавляются данные архива необработанных материалов (sequence read archive), который был создан и активно пополняется данными NGS в последние годы<sup>2</sup>.

Однако уже сейчас далеко не все данные успевают пройти научную обработку и осмысление с биологической точки зрения. Это вызвано не только отсутствием эффективных автоматизированных инструментов обработки и интерпретации, но и недостатком знаний исследователей в области биоинформатики (проще говоря – мощность думающих над смыслом ДНК-текста биологов стала ниже производительности капающих в пробирку лаборантов). Как следствие – научные лаборатории уже не успевают обрабатывать данные, генерируемые сервисными центрами. В попытке догнать эту гигантскую волну «сырых» данных крупные институты строят вычислительные центры для биоинформатических задач, а математики трудятся над оптимизацией алгоритмов. Наиболее вероятным сценарием развития биоинформатической обработки данных

<sup>1</sup> По данным Omicsmaps.com

<sup>2</sup> <http://www.ncbi.nlm.nih.gov/sra>

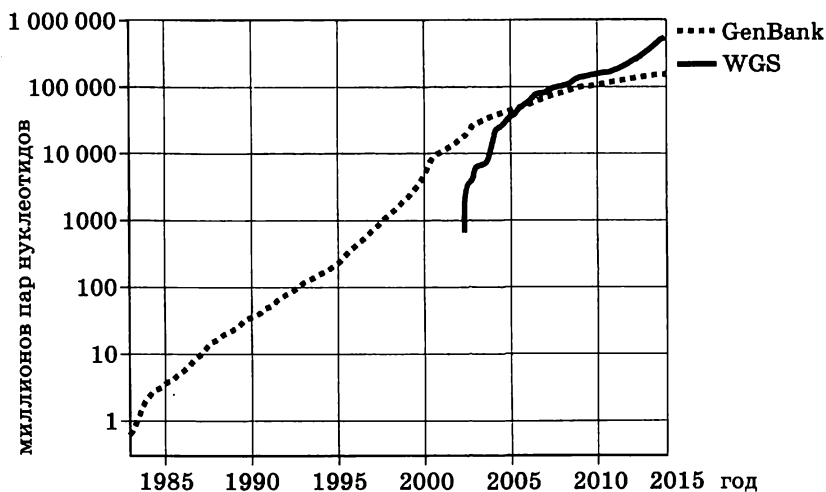


Рис. 13.2. Динамика накопления генетических данных в базах GenBank и WGS (1980–2014 годы)

в ближайшие годы видится разделение труда и переход на облачные технологии (см. гл. 5).

Прежде небольшие объемы данных обрабатывались исследователями самостоятельно, однако уже сейчас очевидно, что данные NGS невозможно обрабатывать без специальных навыков из совершенно разных областей знания (например, одновременно хорошего знания биохимии, генетики, программирования и системного администрирования). Штатный биоинформатик – практически обязательное условие нормального функционирования любой исследовательской группы в области генетики и молекулярной биологии. Что же касается вычислительных мощностей, то современные платформы (например, Amazon, см. разд. 5.3.3) могут предоставить в аренду на минуты или часы сотни тысяч процессоров, что является хорошей заменой владения и администрирования собственного кластера.

Технологии секвенирования пока не очень распространены в медицине, что объясняется закономерным отставанием экспериментальных методик от диагностических процедур. Но в 2012–2014 годах в США наблюдается появление первых рутинных диагностических технологий, основанных на

методах NGS (особенно важно первое полученное одобрение Управлением по контролю качества пищевых продуктов и лекарственных средств на применение Illumina MiSeqDx в медицинской практике). Пока применение ограничивается диагностикой наследственных заболеваний, однако ожидаемо, что со снижением стоимости технологии NGS могут найти применение и в диагностике инфекционных заболеваний, оценке состояния микробных биоценозов человека, ряде задач экологического мониторинга среды и пр. [2].

Технологии NGS третьего поколения бурно развиваются, стремительно снижая стоимость и повышая качество результатов. Широкие возможности современных технологий секвенирования второго поколения позволяют прогнозировать их быстрое и повсеместное распространение в медицине и сельском хозяйстве в ближайшие годы.

#### СПИСОК ЛИТЕРАТУРЫ

1. Barba M., Czosnek H., Hadidi A. Historical perspective, development and applications of next-generation sequencing in plant virology // *Viruses*, 2014, 6 (1): 106–136.
2. Sherry N.L. et al. Outbreak investigation using high-throughput genome sequencing within a diagnostic microbiology laboratory // *J Clin Microbiol.*, 2013, 51 (5): 1396–1401.

# ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

23andMe 213  
454 Life Sciences 30, 60, 66, 165, 171

## Адаптеры

присоединение 52  
супрессионные 131  
аденозинсульфофосфат 28  
алгоритм де Брёйна 152  
алгоритмы 22, 102  
для сборки коротких прочтений  
в более длинные фрагменты  
22, 91  
основанные на сортировке путем  
слияния 103  
основанные на суффиксных  
деревьях 103  
основанные на хэш-таблицах 102  
амплификация  
библиотеки 53  
полногеномная 47, 173  
анализ  
биоинформатический 85  
генома  
обнаружение новой вариации 107  
таксономический 146, 153  
анеуплоидия 100, 214  
аннотация 154  
апираза 28  
апоптоз 216  
ассемблеры 163  
АТФ-сульфуриказа 28  
ацилирование 188

## Базы данных 106

«бар-кодирование» 56  
библиотека  
гиперфрагментация 134  
иммобилизованная 43  
инвертированная 44, 49, 130, 167  
кДНК, ориентированная 179, 185  
клональная 25, 28, 30  
космидная 130  
обычная 44, 130  
парно-концевая 44  
случайных фрагментов ДНК 43  
фрагментов ДНК для NGS 60  
инвертированная 63  
обычная 61  
биннинг 153  
биотехнологии 129  
биотин 200  
болезнь 212  
Гоше 212  
Тей-Сакса 212

## Вариация

внеклеточная ДНК плода 213  
см. также cffDNA 213  
поиск, 105  
структурная 105, 167  
числа копий региона 100  
высокопроизводительное  
секвенирование 13

## Гаплотипирование 167

гаплотипы 169  
гели  
картриджные 54  
элюция 54  
гель-электрофореграмма 50  
ген 16S рРНК 148  
генетическая диагностика 208  
таргетная 209  
генетические исследования 219  
перспективные направления  
в медицинской практике 219  
генетический скрининг  
новорожденных 215  
генетическое тестирование 207  
геном  
бактериальный 130  
референсный 97, 101, 156  
генома  
сборка 163  
геномы  
прокариотические 162  
эукариотические 162  
Генотек 213, 215  
гибридизация на твердой фазе 21  
гибридизация с пробой 198  
гидразин 15  
гидролиз 15  
гидроширинг 48  
гиперметилирование генома 217  
гомополимерные участки  
последовательности 33, 37, 68  
граф де Брёйна 93

## Данные 146

метагеномные 146  
метапротеомные 146  
метатранскриптомные 146  
о последовательностях нуклеиновых  
кислот, первичные 85  
дезоксинуклеозидтрифосфаты, дНТФ  
17  
делеции 167  
дидезоксинуклеозидтрифосфаты,  
ддНТФ 17

дисплазия соединительной ткани 210  
дифференцированные дисплазии 210  
длина генома 90  
длина прочтения 90, 95  
распределение 96  
ДНК 14, 47  
клональная амплификация  
фрагментов 57  
метилирование 14  
однопочечная 35  
оценка длин фрагментов 51  
способы разрушения 47  
ДНК-колония 28, 30, 31, 71  
ДНК-технология 221  
дуплекс-специфичная нуклеаза 187

Евроген 47, 54

Загрузка микросфер на чип 140  
затравка олиго-dT 178

Иммунопреципитация хроматина  
191, 203  
инверсии 100, 105, 167  
инвертированные молекулярные  
пробы 201  
искусственные бактериальные  
хромосомы 20  
*см. также* ВАС 20  
исследование  
микробиома человека 220  
микробиоценозов человека 207  
патогенов 220  
исследования  
ассоциативные 101  
метагеномные 67, 153  
скрининговые 66

Кариотипирование 209  
кДНК 178  
кистозный фиброз 212  
клональная амплификация  
фрагментов ДНК 57  
комбинационное рассеяние света 38  
континги 91, 164, 168  
объединение в скэффолды 97

Лазерная захватывающая  
микродиссекция 172  
*см. также* LCM 172  
лигирование адаптеров  
для секвенирования 135  
люцифераза 28  
люциферин 28

Малые РНК 177, 182  
масс-спектрометрия 23  
метагеном 151  
метагеномное секвенирование 151  
метилирование геномов 80  
метод  
Hi-C 99  
визуализации в сканирующем  
электронном микроскопе 34  
дидеокситерминаторов 16  
дробовика 20  
инвертированной ПЦР 63  
колебаний 40  
Максама-Гилберта 14  
очистки нуклеиновых кислот 45  
Сенгера 16, 164  
терминаторов 16  
электрофореза в геле 51  
эмульсионной ПЦР 75  
методы расщепления ДНК  
физические 48  
энзиматические 49  
микробиоценоз 147  
микрореактор 32  
микроРНК 182  
микроскопия электронная 34  
микросфера 25, 32  
мостиковая ПЦР 58  
мРНК 180, 184, 188  
мутации 107, 209, 215  
рецессивные 216  
соматические 176, 209, 215

Небулизация 48, 67  
нормализация библиотеки DSN 187  
нуклеаза 50  
нуклеотиды 35  
визуализация 35

Обогащение 48  
в растворе 199  
гибридизацией 198  
микрофер 138  
на твердой фазе 198  
обработка данных NGS  
вычислительный центр 113, 114  
локальный центр 112  
программное обеспечение 116  
сетевые сервисы 117  
специализированные проекты 120  
обратимые терминирующие  
нуклеотиды 30  
однонуклеотидные вариации 100  
однонуклеотидный полиморфизм  
104, 167  
*см. также* SNP 104

- однородность покрытия 191
- онкоген 216, 217
  - доминантный 216
  - классы 216
- отбор 136
  - оптимальной фракции библиотеки 136
  - фракции фрагментов нужной длины 53
- отжиг 16
- оценка 136
  - качества и количества библиотеки фрагментов ДНК 136
  - качества и количества исходной ДНК 132
- очистка
  - ДНК от фермента 133, 147
  - нуклеиновых кислот 45
- ошибки секвенирования 94
- Пиперидин 15
- пиросеквенирование 28
- пирофосфат 28
- планирование биологических экспериментов 122
- повтор
  - триплетный 209
  - экспансия 209
- повторности 124
- подготовка библиотеки к загрузке на чип 139
- поиск ассоциаций по всему геному 220
- покрытие 89
  - среднее 167
- полиаденилированная (полиА) фракция РНК 198
- полиакриламидный гель 16
- полимераза
  - AMV 180
  - MMLV 180, 181
  - Phi29 47, 173
  - Taq 173
  - Tfi 181
  - Tth 180
  - ДНК 180
- полногеномное исследование 209
- полногеномный поиск ассоциаций 105
- полупроводниковое секвенирование 32
- праймер 25, 47, 59, 173, 178
  - перезагрузка 71
- программы для картирования прочтений 102
- «прогулка по геному» 131
- проект
  - «Геном человека» 170
  - «Микробиом человека» 156
- проточный чип 70, 73
- прочтения 89, 97
  - парно-концевые 164
  - химерные 97
- ПЦР 58, 192
  - «дальнобойная» 131, 166
  - классическая 192
  - мостиковая 58
  - мультиплексная 192
  - оптимизированная для наработки длинных фрагментов 131, 166
  - эмульсионная 58, 138, 195
- Рамановская спектроскопия 39
- рандомизация 123
- распределение Пуассона 89
- распределенные вычисления в облаке 118
  - Amazon Cloud Stack 119
  - Microsoft Azure 119
- реакции мечения 18
- репарация 216
- ресеквенирование диплоидных геномов 166
  - фазирование 169
- референсный геном 168
- рибосомальный футпринтинг 188
- Сборка
  - de novo* 89
  - генома 163
  - последовательностей 152
- секвенаторы 16, 33, 67, 69, 72, 74, 79, 81, 165
  - автоматические 16, 20
  - полупроводниковые 74
  - третьего поколения 165
- секвенирование 16, 48, 79, 87, 125, 128, 164, 184, 191, 215
  - алгоритм 130
  - второго поколения 13, 34, 66, 165
  - генома 162, 163
  - генома отдельной клетки 171
  - учет искажений 97
  - достаточное покрытие 164
  - иммунопреципитированных элементов хроматина 203
  - индивидуальных геномов и транскриптомов прокариот 128
  - использование в онкологии 215
  - лигированием 24
  - метагеномное 194

- методом дробовика 170  
методом Сенгера 87, 129  
методом спектроскопии  
    комбинационного рассеяния 38  
микробных сообществ 128  
на чипе 21  
одиночных молекул «в реальном  
    времени» 79  
одиночных молекул ДНК 34  
ориентированное 184  
основные типы ошибок 125  
отдельных микробов 128  
ошибки 104  
«плюс-минус» 16  
повторное 100  
    картирование прочтений 101  
пренатальное экзомное 215  
при помощи вращающегося поля 40  
при помощи электронного  
    микроскопа 34  
путем протаскивания ДНК через  
    нанопоры 37  
синтезом 25, 31, 79  
синтезом одиночных молекул 36  
стоимость 223  
таргетное 48, 191  
третьего поколения 13, 34, 78, 165  
серповидноклеточная анемия 212  
синдром 210  
    атипичный гемолитико-  
        уремический 211  
Блума 212  
Дауна 213  
Марфана 210  
Патау 213  
Стиклера 210  
Эдвардса 213  
Элерса–Данлоса 210  
скрининг  
    на статус носителя наследственных  
        заболеваний 211  
    пренатальный 213  
скэффолды 97  
случайная затравка 178  
соникация 48  
сорбенты силикатные 45, 135  
спектроскопия комбинационного  
    рассеяния света 39  
спектрофотометр 46  
сравнительная геномная  
    гибридизация 106  
стоимость секвенирования генома  
    человека NGS 223  
стрептавидин 200  
супрессионная вычитающая  
    гибридизация 106  
супрессионные адаптеры 185  
    псевдо-двущепочечные 61  
супрессор 216  
T7-промотор 24  
талассемия 212  
тестирование 207  
    взрослое 208  
    детское 208  
    диагностическое 208  
    для определения статуса  
        носителя наследственного  
        заболевания 208  
    неонатальное 208  
    преимплантационное 208  
    пренатальное 208  
    прогностическое 208  
транскрипт одной клетки 187  
транскриптом 198  
транскрипция *in vitro* 24  
транслокации 100  
транспозаза 50  
транспозоны 167  
Уравнение Ландера–Ватермана 89  
участки гетерозиготности 95  
Фазирование 169, 170  
фактор разведения  
    библиотеки 136  
фармакогенетика 220  
ферментативное расщепление  
    ДНК 133  
флуоресцентная гибридизация  
    *in situ* 105  
флуориметр 46  
флуорофор 30, 36  
Химерные прочтения 148  
хранение данных NGS 113  
хроматограмма 87  
«Штрих-код» 44, 56  
штрих-кодирование 56, 192  
Электронная микроскопия 34  
электрофореграмма 54  
электрофорез 16  
эндонуклеаза  
    T7 50  
    рестрикции 131  
эпигенетические изменения 217  
эффект  
    Рамана 38  
    супрессии ПЦР 53  
эффективность обогащения 191

- Affimetrix** 201  
**Agilent Technologies** 201  
**Amazon Cloud Stack** 119
- BAC** 20, 21
- CffDNA** 213, 214  
**CGH** 106  
**ChIP-chip** 204  
**ChIP-seq** 126, 191, 203, 204  
**Councyl** 213
- DOP-PCR** 173  
**DSN** 186
- FASTQ** 86, 87  
**FASTQC** 143  
**FISH** 106  
**Fluidigm Access Array** 197  
**Fluidigm Corporation** 197
- Genome Analysis Tool Kit** 171  
 genome walking 131  
**GWAS** 105, 220
- HAADF** 34  
**Haplotype Improver** 171  
**Human Genome Project** 18
- Illumina** 25, 31, 62, 165, 171, 201  
**Indel** 167  
**Ion AmpliSeq** 194  
**Ion Chef** 60, 200  
**Ion OneTouch** 75, 76, 138, 200  
**Ion PGM** 60, 74, 83, 132, 138, 139, 141  
**Ion Proton** 60, 83, 194, 200  
**Ion TargetSeq** 200  
**Ion Torrent** 61, 83, 149, 152, 171  
**ISFET-сенсор** 77  
**ISP-сфера** 75, 77
- LCM** 172  
**Life Technologies Thermo Fisher Scientific** 33, 54, 74, 118, 194, 200  
**long-range PCR** 131; 166
- MALDI-TOF масс-спектрометрия** 23  
**mate-pair library** 44, 63
- MDA** 173  
**MIP** 201, 204  
**mQ-вода** 78
- N50** 100  
**N90** 100  
**NGS** 13, 126, 128  
     варианты применения для различных целей 126  
     накопление и обработка данных 225  
     роль в микробиологии 128  
**NimbleGen** 198
- OLC** 91, 95
- PacBio** 43, 78, 79, 80, 125, 165, 171  
**paired-end library** 44  
**Pathway Genomics** 213  
**PEP** 173  
**PHASE** 169  
**Phred** 87, 88  
**Polonator** 28, 60, 83
- RainDance Technologies** 195  
**RainStorm** 195  
 «read-backed phasing» 171  
**read depth** 167
- S1 нуклеаза** 173  
**SBH** 21  
**SEM** 34  
**SHAPE-Seq** 188  
**shotgun sequencing** 20  
**size-select** 53, 136  
**SNP** 100, 104, 167  
**SOLiD** 60, 62, 69, 165  
**SSH** 106  
**STEM** 34  
**Sure Select** 201  
**SVs** 167
- TaqMan** 47  
**TDF** 136  
**TERS** 39  
**TrueSeq Enrichment** 201  
**TS** 216  
**tSMS** 80
- WGA** 47, 173

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 секвенсного центра.

ISBN 978-5-9963-1784-4



9 785996 317844