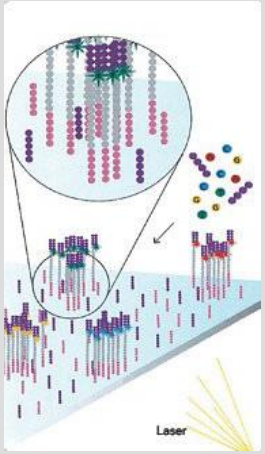
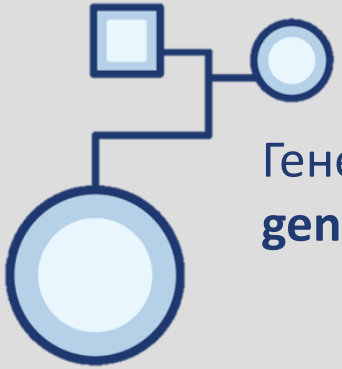
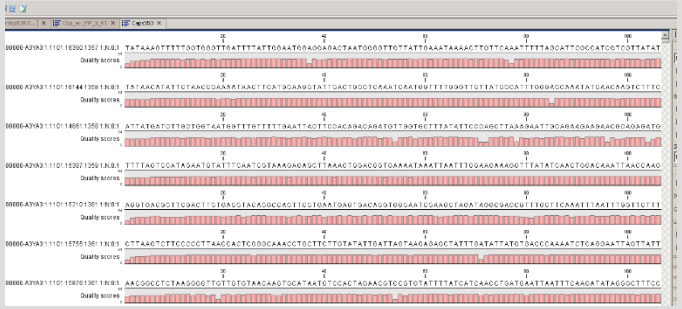
Секвенирование нового поколения:



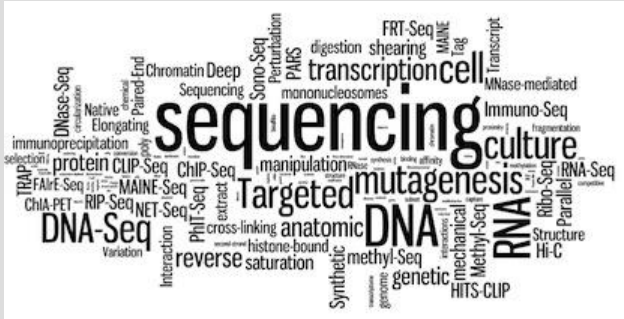
принципы



Генетические школы
genschool.ru



возможности



перспективы

Мария Логачева,
Сколковский институт науки и технологии, Genomics core facility

Лекция прочитана в рамках проекта Genschool.ru

Технологии секвенирования

- 1-е поколение

Sanger sequencing

N G S

- 2-е поколение

Roche

illumina

ion torrent

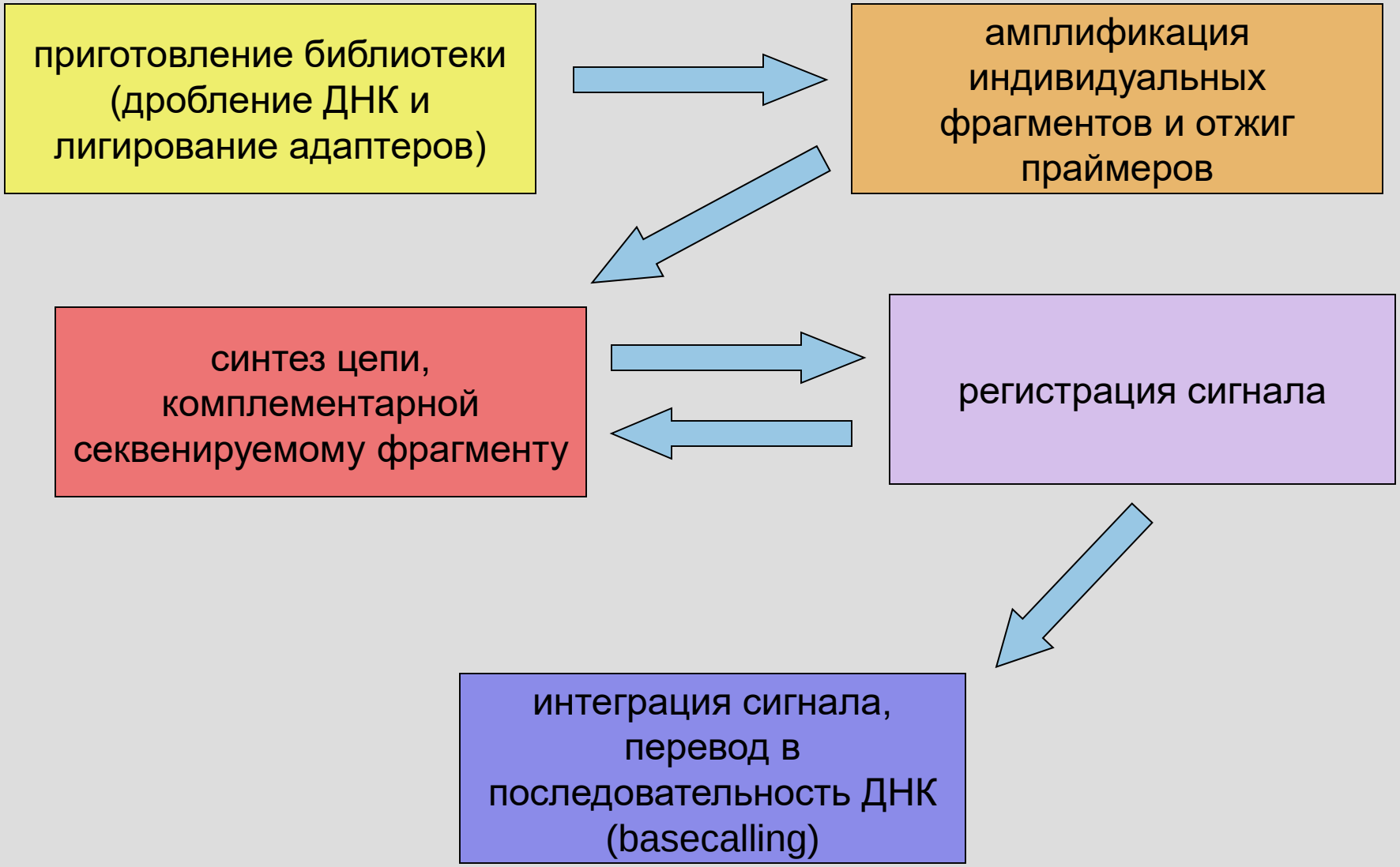
applied biosystems

- 3-е поколение

Helicos

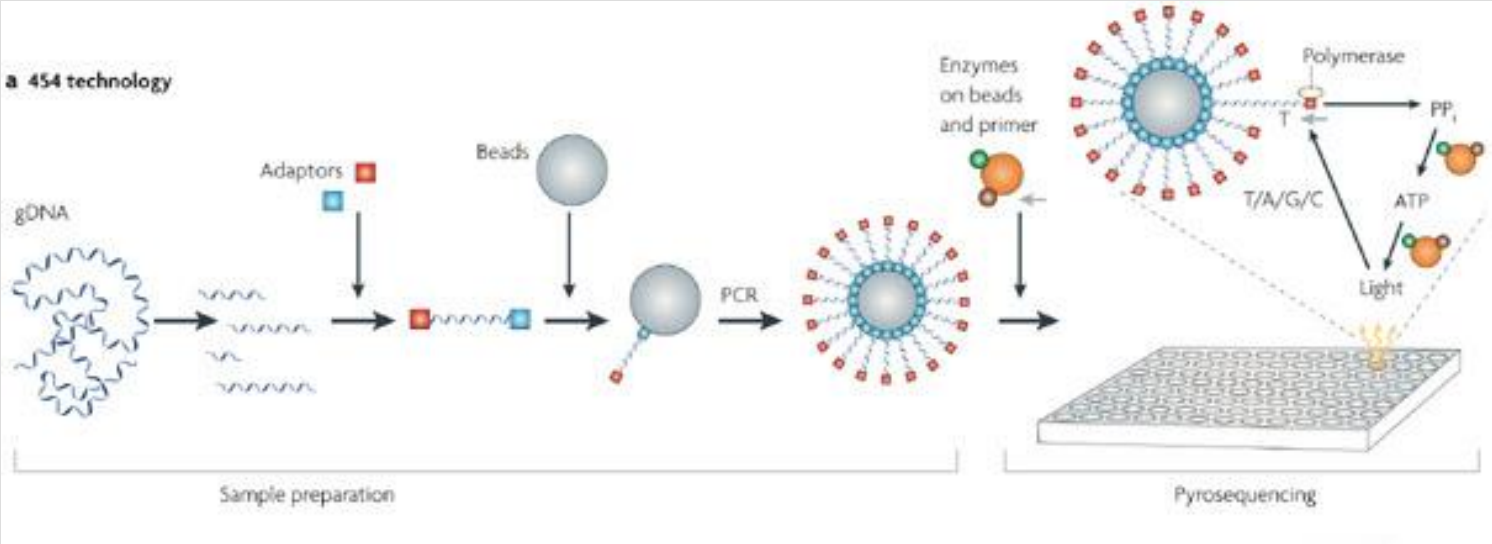
PACIFIC BIOSCIENCES

Общая схема работы NGS: от исходной ДНК до «букв» на экране



Технологии секвенирования 2 поколения: 454

Самая первая из технологий NGS (Margulies et al. Nature 2005; 441.7089)



Преимущества:

- большая длина чтения (сравнимо с секвенированием по Сэнгеру)
- короткое время работы
- наименее чувствительна к GC-составу

Недостатки

- неточность прочтения гомополимерных участков
- высокая цена в расчете на нуклеотид
- в 2016 году прекратилась поддержка

Illumina – секвенирование путем синтеза с обратимым терминированием

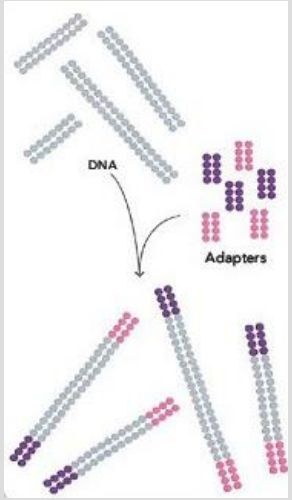
Самая распространенная из NGS-технологий (~ 85% данных)

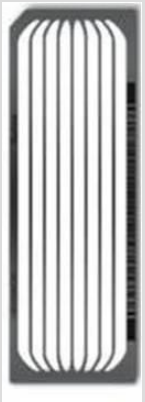
Преимущества:
- высокая точность (частота ошибок 0.1-0.5%)
- универсальность
- доступность ПО для обработки и анализа результатов
- наименьшая цена получаемых данных (в расчете на нуклеотид)

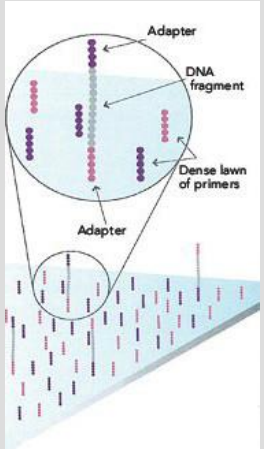
Недостатки
- высокая цена реагентов
- проблемы с секвенированием матриц с низкой сложностью
- большая длительность прогона (не для всех приборов)
- ошибки в GC-богатых участках
- артефакты неправильного соотнесения чтений

Секвенирование Illumina - принцип метода

1. Подготовка библиотеки

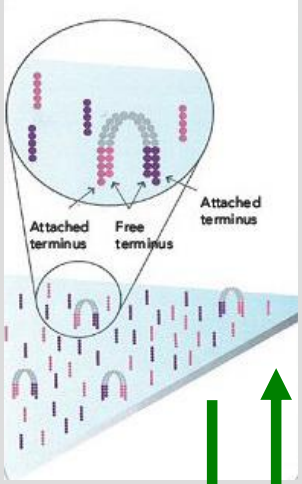


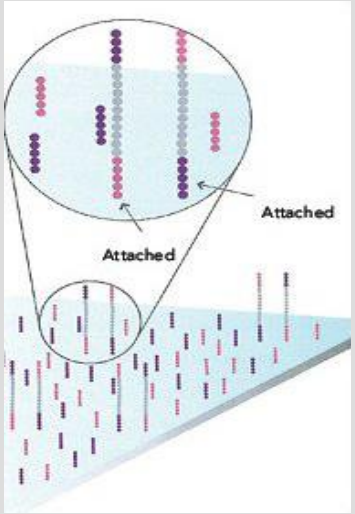




2. Библиотеку пропускают через каналы ячейки, покрытые праймерами, комплементарными концам адаптеров

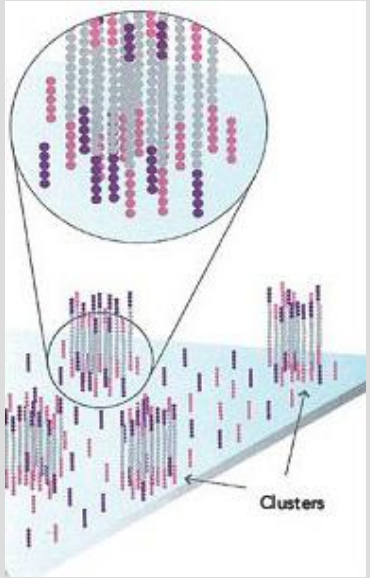
3. Через ячейку пропускают реагенты для достраивания второй цепи ДНК





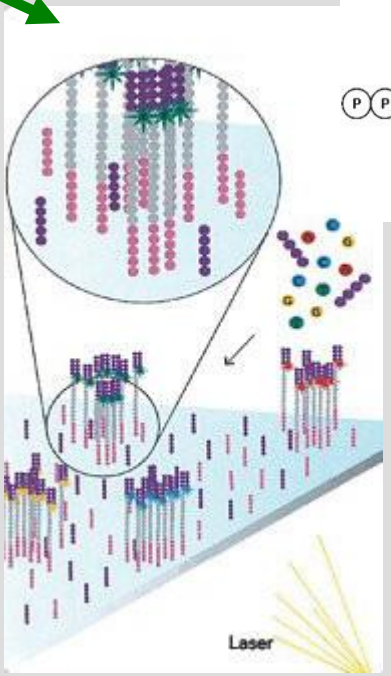
4. Двухцепочечные фрагменты денатурируют

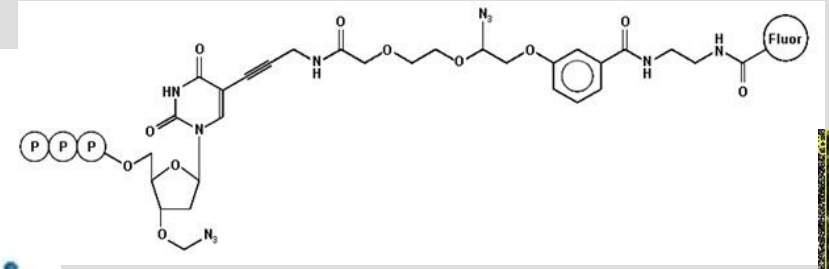
Стадии 3-4 повторяются 30-35 раз

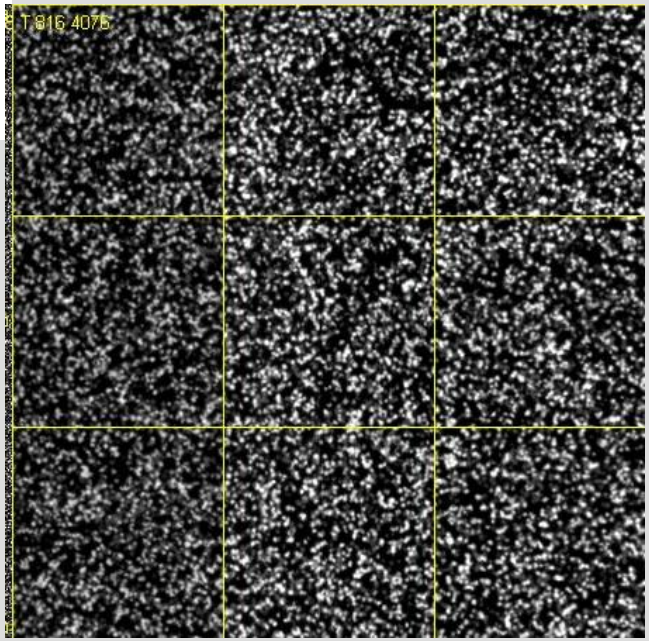


6. Каждый фрагмент оказывается окружен группой идентичных молекул («кластеры»).

Секвенирование Illumina - принцип метода







7. Синтез

8. Сканирование

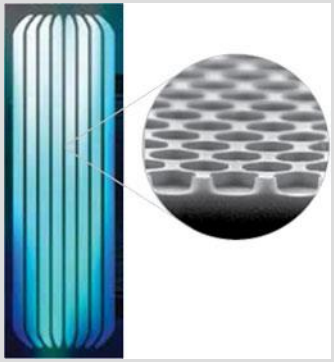
10. Повторение 7-9 нужное число раз (50-300). Число циклов соответствует длине чтения

9. Деблокирование и отмывка

Illumina: дальнейшее развитие

Упорядоченные (patterned) ячейки

Области, пригодные для формирования кластеров, занимают не всю ячейку, а расположены упорядоченно.



Преимущества: меньшая чувствительность к плотности кластеров, отсутствие артефактов неправильного соотнесения чтений
Недостатки: более строгие требования к длине библиотек, артефакты генерации кластеров

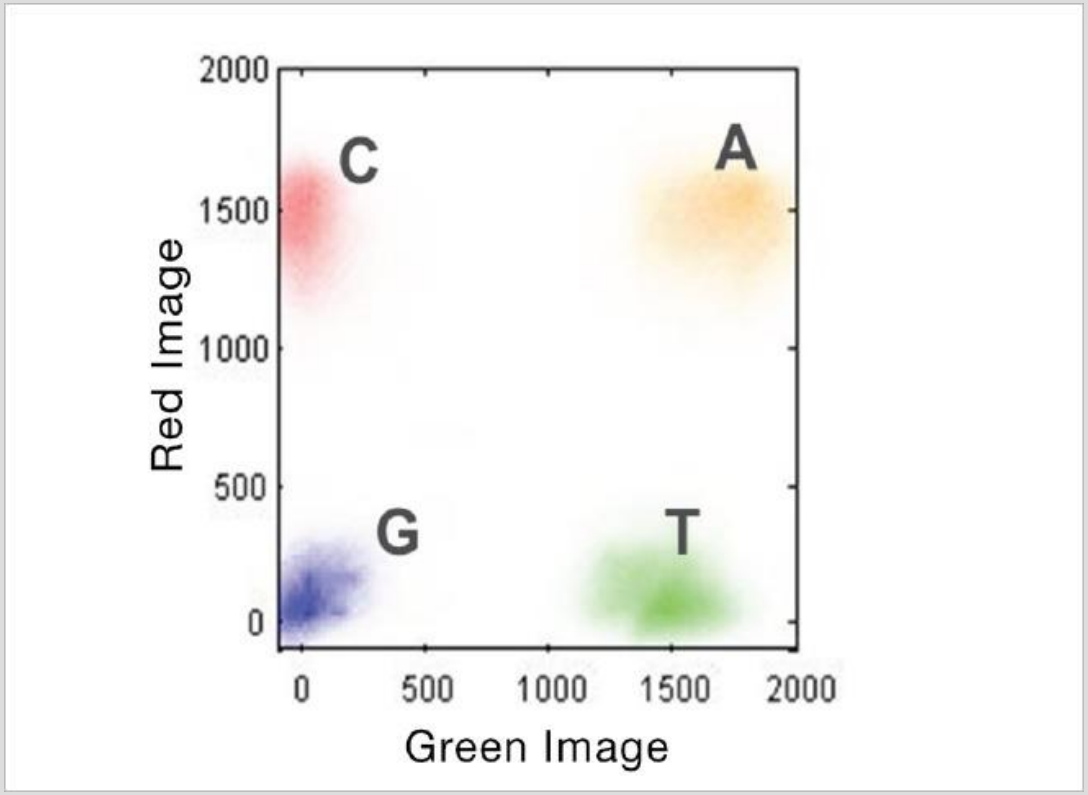
Исключающая амплификация (exclusion amplification)

Обеспечивает формирование моноклональных кластеров за счет быстрой изотермической амплификации.

Преимущества: меньшая чувствительность к плотности кластеров, большая доля моноклональных кластеров
Недостатки: более строгие требования к длине библиотек, артефакты генерации кластеров

Illumina: дальнейшее развитие

Двухцветное сканирование



Используются два типа флуорофоров.
C и T имеют метки каждого из типов, A – обе метки, G – без метки

Преимущества: скорость
Недостатки: большая требовательность к плотности кластеров, меньшая точность

Приборы:

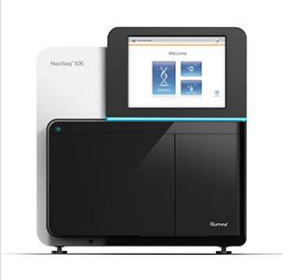
- MiniSeq – двухцветный аналог MiSeq
- NextSeq – средняя производительность, от 120 до 400 M чтений, длина от 75 до 150+150 пн
- NovaSeq – двухцветный аналог HiSeq

Illumina: линейка приборов

HiSeqX10

HiSeq2000/2500
до 1.6 млрд чтений,
длина чтения до 250+250
HiSeq4000
до 2.5 млрд чтений
длина чтения до 150+150

NovaSeq – 2-цветный, до 3 млрд чтений
длина чтения до 150+150



NextSeq – 2-цветный
120-400 млн чтений
длина чтения до 150+150

MiSeq 4-цветный
MiniSeq – 2-цветный
15-25 млн чтений
длина чтения до 300+300 (MiSeq)
или 150+150 (MiniSeq)

Полупроводниковое секвенирование

Самая новая из технологий секвенирования 2 поколения
Принцип сходен с 454, но регистрируется не свет, а pH

Преимущества:
- относительно низкая цена за запуск
- быстрота
- относительно большая длина чтения – 200-600 пн

Недостатки
- невысокая точность прочтения гомополимерных участков
- низкая производительность

- ресеквенирование небольших наборов генов (панелей) для диагностики и скрининга
- приложения, требующие высокой скорости (пренатальная диагностика, контроль распространения патогенов)
- секвенирование де ново небольших геномов
- «полевые» исследования

Полупроводниковое секвенирование: приборы

Ion PGM

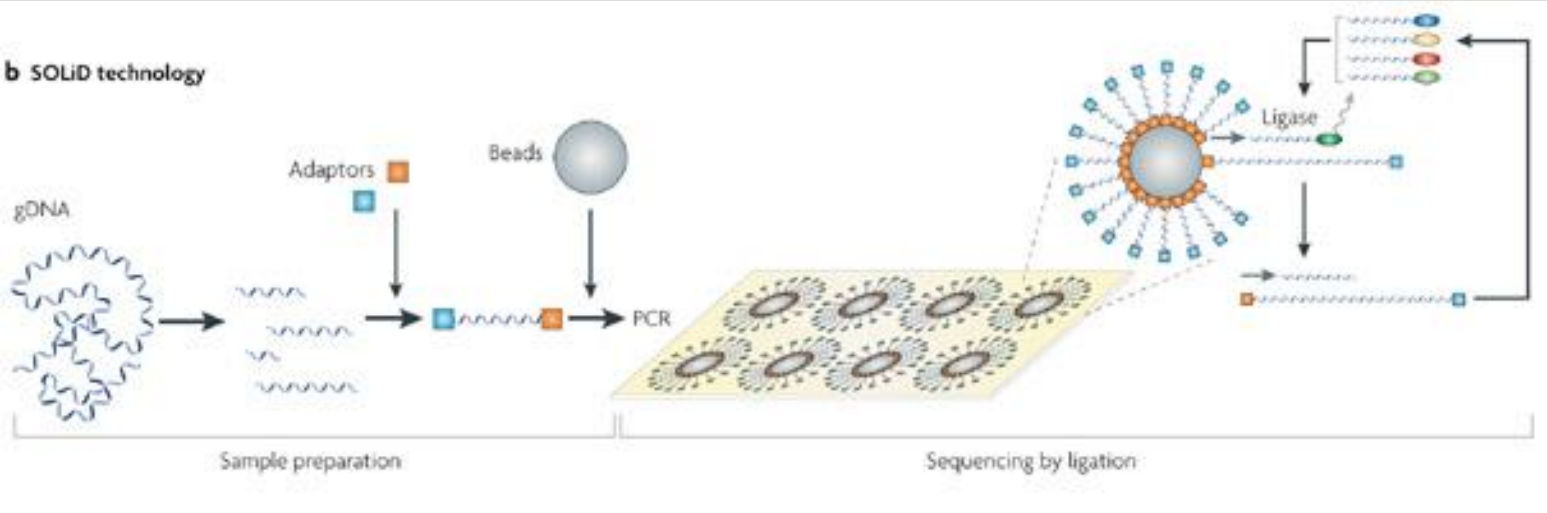
- 4-5 миллионов чтений
- время работы 4-7 часов
- длина чтения 200-400 пн

Ion Proton

- 60-80 миллионов чтений
- время работы 2-4 часа
- длина чтения до 200 пн

Ion S5 – сочетает режимы работы PGM и Proton. Длина чтения до 600 пн.

Секвенирование путем лигирования (SOLiD)



Преимущества:

- высокая точность
- возможность использовать часть дорожек на ячейке

Недостатки

- очень короткие чтения
- длительность работы
- относительно малая доступность свободного ПО

Основные ограничения секвенирования 2 поколения

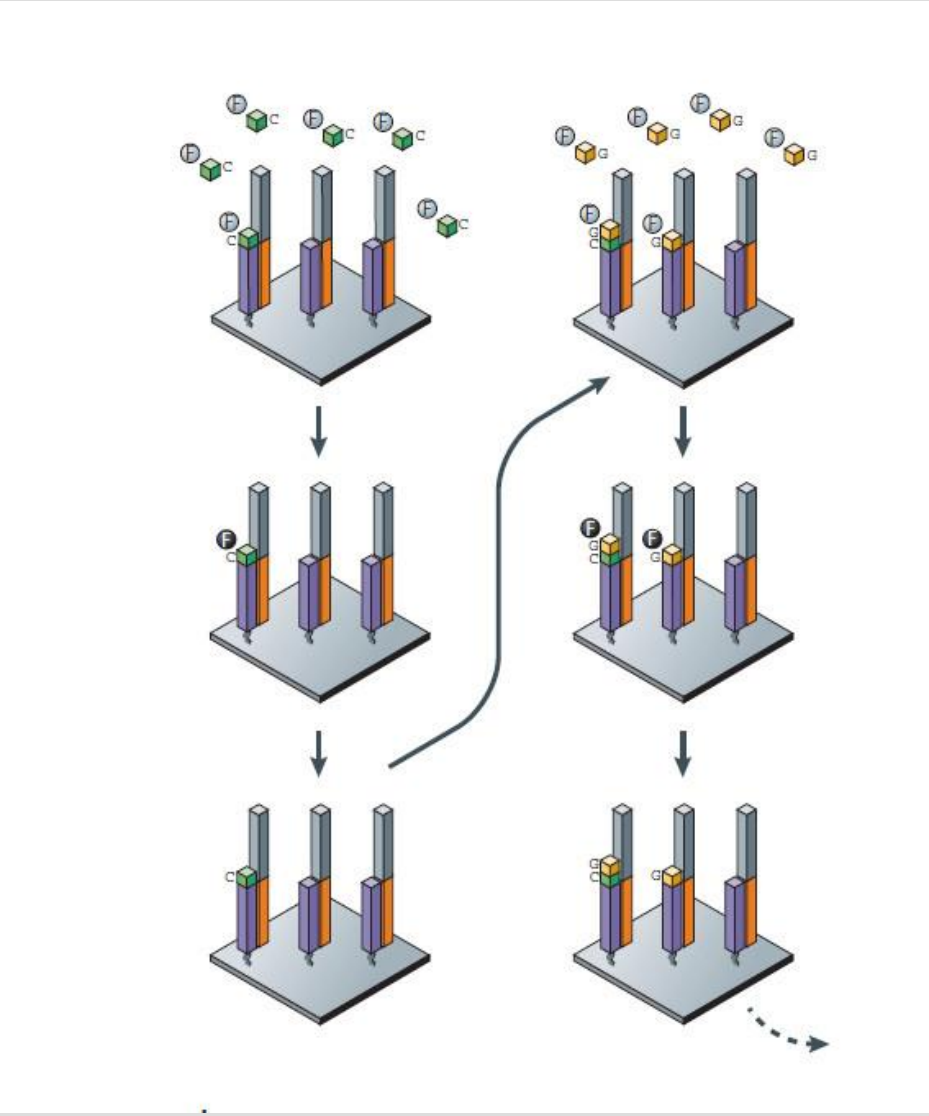
- **Дороговизна приборов и реагентов**
- **Малая длина прочтения (менее 1000 bp)**
- **ПЦР – источник ошибок и артефактов**
- **Невозможность детекции модифицированных нуклеотидов**

Секвенирование единичных молекул

Helicos

- **простая пробоподготовка** (ДНК фрагментируется, аденилируется и закрепляется на ячейке с олиго-dT адаптерами)
- **принцип секвенирования сходен с Illumina** (флуоресцентно меченые обратимо терминированные нуклеотиды)
- **небольшая длина чтения** (до 50 пн)
- **большое количество ошибок** (3-5%)
- **~ 35 Гб за один запуск**
- **время работы 8 дней**

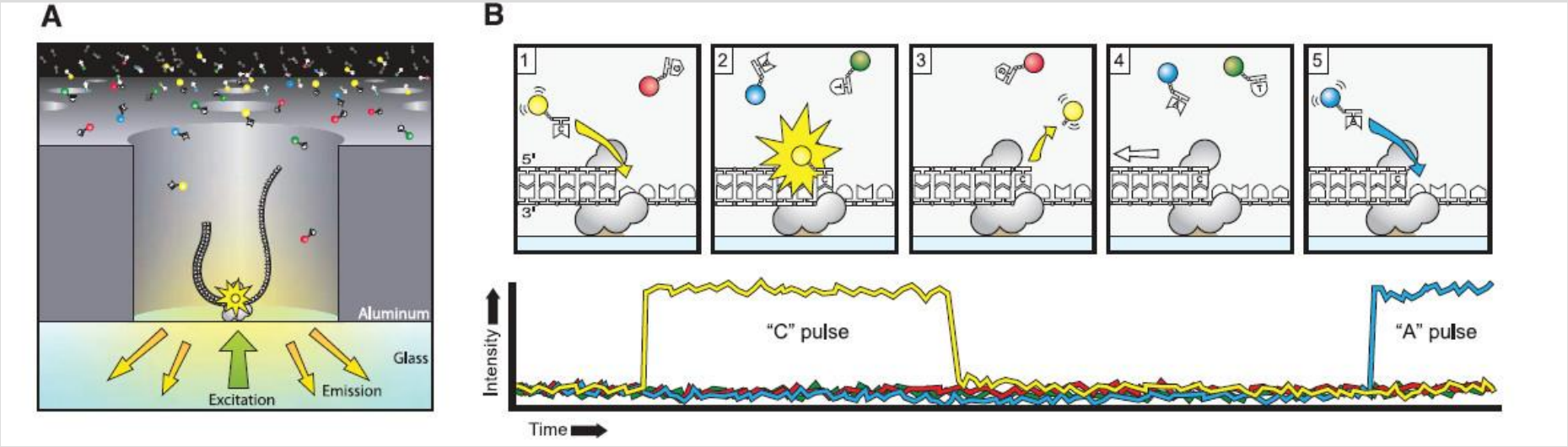
Область применения – высокоточный анализ экспрессии, оценка копийности участков генома



Секвенирование единичных молекул

SMRT sequencing (Pacific Biosciences)

Используется полимераза, иммобилизованная в 100-нм лунках, и флуоресцентно меченые dNTP. Возможны очень длинные чтения (> 10 000), но высокая частота ошибок (до 10%). Возможно прямое секвенирование метилированной ДНК, ведется работа над разработкой прямого секвенирования РНК.



PacBio: примеры применения

Method Секвенирование сложных участков генома

Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene

Erick W. Loomis,[1,5] John S. Eid,[2,5] Paul Peluso,[2] Jun Yin,[1] Luke Hickey,[2] David Rank,[2] Sarah McCalmon,[2] Randi J. Hagerman,[3,4] Flora Tassone,[1,4] and Paul J. Hagerman[1,4,

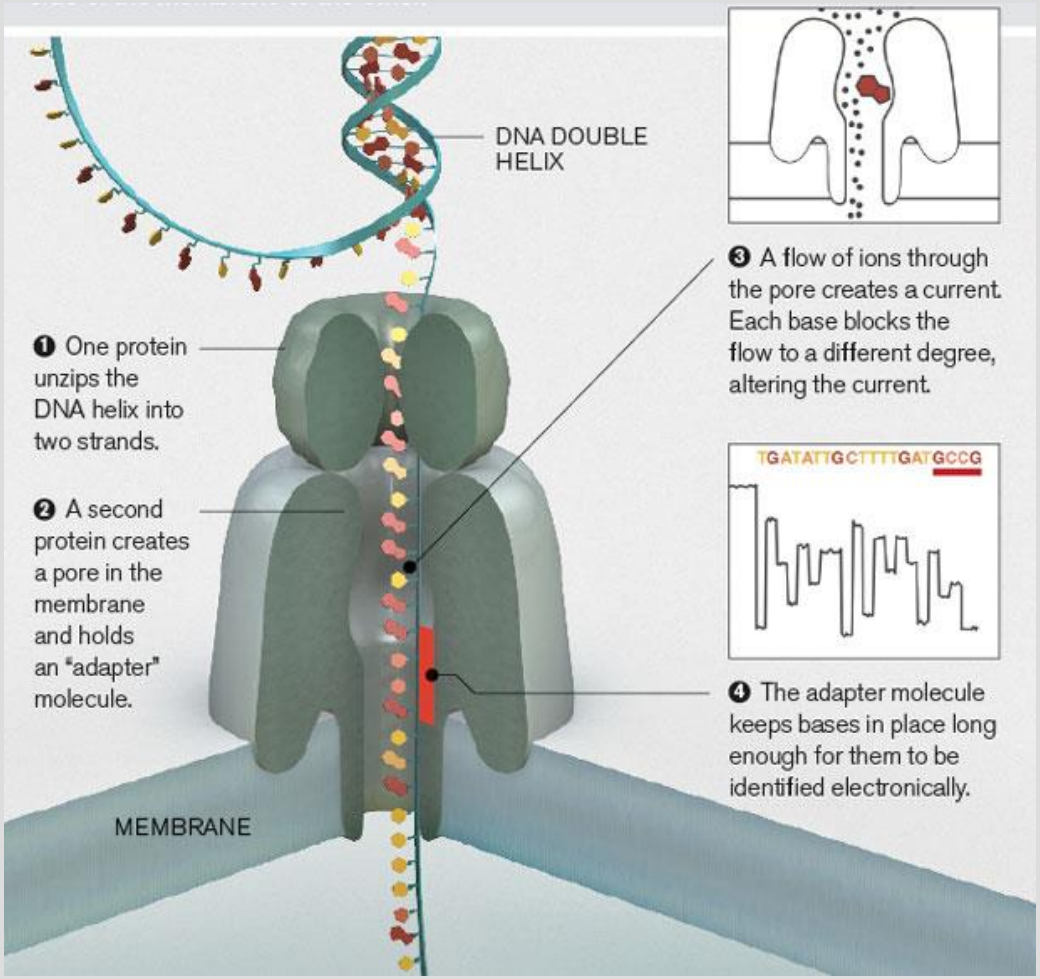
[1]Department of Biochemistry and Molecular Medicine, University of California, Davis, School of Medicine, Davis, California 956 USA; [2]Pacific Biosciences, Inc., Menlo Park, California 94025, USA; [3]Department of Pediatrics, University of California, Davis, Sc of Medicine, Sacramento, California 95817, USA; [4]MIND Institute, University of California Davis Medical Center, Sacramento, Calif 95817, USA

The human fragile X mental retardation 1 (*FMR1*) gene contains a $(CGG)_n$ trinucleotide repeat in its 5′ untranslated region (5′UTR). Expansions of this repeat result in a number of clinical disorders with distinct molecular pathologies, including fragile X syndrome (FXS; full mutation range, greater than 200 CGG repeats) and fragile X–associated tremor/ataxia syndrome (FXTAS; premutation range, 55–200 repeats). Study of these diseases has been limited by an inability to sequence expanded CGG repeats, particularly in the full mutation range, with existing DNA sequencing technologies. Single-molecule, real-time (SMRT) sequencing provides an approach to sequencing that is fundamentally different from other "next-generation" sequencing platforms, and is well suited for long, repetitive DNA sequences. We report the first sequence data for expanded CGG-repeat *FMR1* alleles in the full mutation range that reveal the confounding effects of CGG-repeat tracts on both cloning and PCR. A unique feature of SMRT sequencing is its ability to yield real-time information on the rates of nucleoside addition by the tethered DNA polymerase; for the CGG-repeat alleles, we find a strand-specific effect of CGG-repeat DNA on the interpulse distance. This kinetic signature reveals a novel aspect of the repeat element; namely, that the particular G bias within the CGG/CCG-repeat element influences polymerase activity in a manner that extends beyond simple nearest-neighbor effects. These observations provide a baseline for future kinetic

Секвенирование единичных молекул.

Нанопоровое секвенирование (Oxford Nanopore)

Принцип основан на использовании мембран с белковыми нанопорами, через которые протягивается молекула ДНК. Секвенатор размером с USB-диск. Высокая скорость, очень высокая частота ошибок, низкая производительность (пока!)



Method

Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome

Sara Goodwin,[1] James Gurtowski,[1] Scott Ethe-Sayers, Panchajanya Deshpande, Michael C. Schatz, and W. Richard McCombie

Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

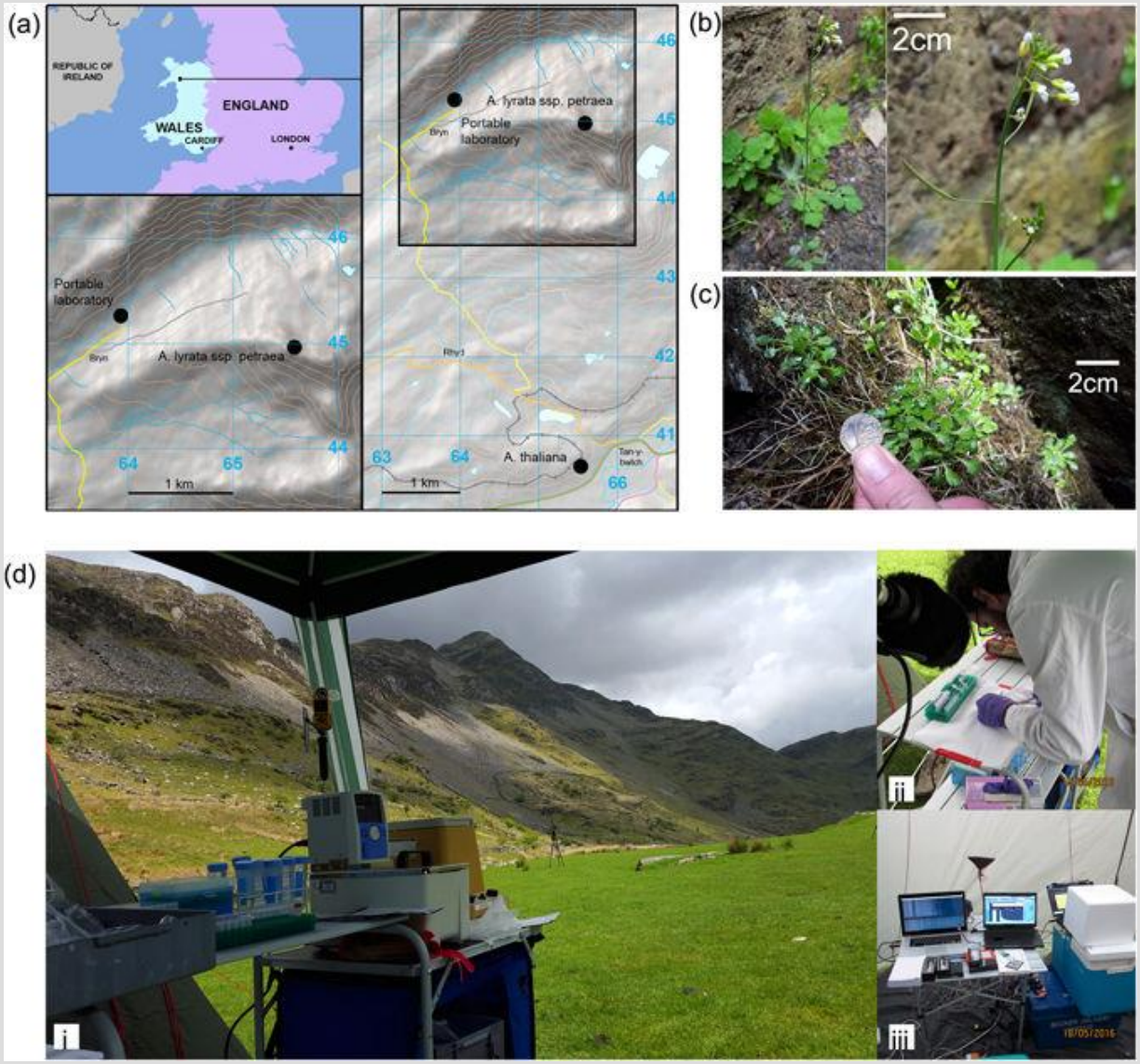
Monitoring the progress of DNA molecules through a membrane pore has been postulated as a method for sequencing DNA for several decades. Recently, a nanopore-based sequencing instrument, the Oxford Nanopore MinION, has become available, and we used this for sequencing the *Saccharomyces cerevisiae* genome. To make use of these data, we developed a novel open-source hybrid error correction algorithm Nanocorr specifically for Oxford Nanopore reads, because existing packages were incapable of assembling the long read lengths (5–50 kbp) at such high error rates (between ~5% and 40% error). With this new method, we were able to perform a hybrid error correction of the nanopore reads using complementary MiSeq data and produce a de novo assembly that is highly contiguous and accurate: The contig N50 length is more than ten times greater than an Illumina-only assembly (678 kb versus 59.9 kbp) and has >99.88% consensus identity when compared to the reference. Furthermore, the assembly with the long nanopore reads presents a much more complete representation of the features of the genome and correctly assembles gene cassettes, rRNAs, transposable elements, and other genomic features that were almost entirely absent in the Illumina-only assembly.

[Supplemental material is available for this article.]

Most DNA sequencing methods are based on either chemical cleavage of DNA molecules (Maxam and Gilbert 1977) or synthesis of new DNA strands (Sanger et al. 1977), which are used in the majority of today's sequencing routines. In the more common synthesis-based methods, base analogs of one form or another are

pin adapter is ligated to one end of a double-stranded DNA sample, while a "motor" protein is bound to the other to unwind the DNA and control the rate of nucleotides passing through the pore (Clarke et al. 2009). Under ideal conditions the leading template strand passes through the pore, followed by the hairpin adapter

Секвенируем, не отходя от места сбора?



Arabidopsis thaliana RTnS reads could be aligned to approx. 50% of the reference genome (53Mbp) with an average error rate of 20.9%.

The *A. lyrata* RTnS data were more problematic with significantly poorer mapping to the two *A. lyrata* assemblies

Parker et al. (2017). Field-based species identification of closely-related plants using real-time nanopore sequencing. *Scientific reports*, *7*(1), 8345.

Геном человека, сборка де ново

MENU nature biotechnology

Article | OPEN

Nanopore sequencing and assembly of a human genome with ultra-long reads

Miten Jain, Sergey Koren [...] Matthew Loose

Nature Biotechnology **36**, 338–345 (2018)
doi:10.1038/nbt.4060
Download Citation

DNA sequencing | Genomics | Next-generation sequencing

Received: 20 April 2017
Accepted: 11 December 2017
Published: 29 January 2018

Abstract

We report the sequencing and assembly of a reference genome for the human GM12878 Utah/Ceph cell line using the MinION (Oxford Nanopore Technologies) nanopore sequencer. 91.2 Gb of sequence data, representing ~30× theoretical coverage, were produced. Reference-

Недостатки

- **невысокая точность**

10-12% ошибок

- **невысокая производительность**
- **отсутствие стандартизированных протоколов**
- **низкая доступность некоторым группам пользователей, риск прекращения поддержки**

Maria Logacheva	Laboratory of evolutionary genomics Moscow State University 119192 Vorobyovy Gory MSU 1, Building 73, Moscow, Russia	Rejected. No shipment for any item permitted without a license due to Military links

In your email you had mentioned that "I note that these labs are not subdivisions of the Defense or Energy State Units". However this is the 2nd customer of yours that UK government has identified as having military links.

Области применения NGS

или

What can next generation sequencing do for you?

- **секвенирование геномов и транскриптомов de novo**

отправная точка большинства молекулярно-биологических и генетических исследований на немодельных объектах, поиск крупных геномных перестроек

- **полногеномное ресеквенирование**

поиск мутаций, ассоциированных с болезнями, картирование генов

- **направленное ресеквенирование**

биомедицина: скрининг мутаций с известной ролью в развитии болезней и поиск новых мутаций

- **анализ транскриптома (RNA-seq)**

сравнение уровней экспрессии, поиск новых генов и изоформ, аннотация de novo секвенированных геномов

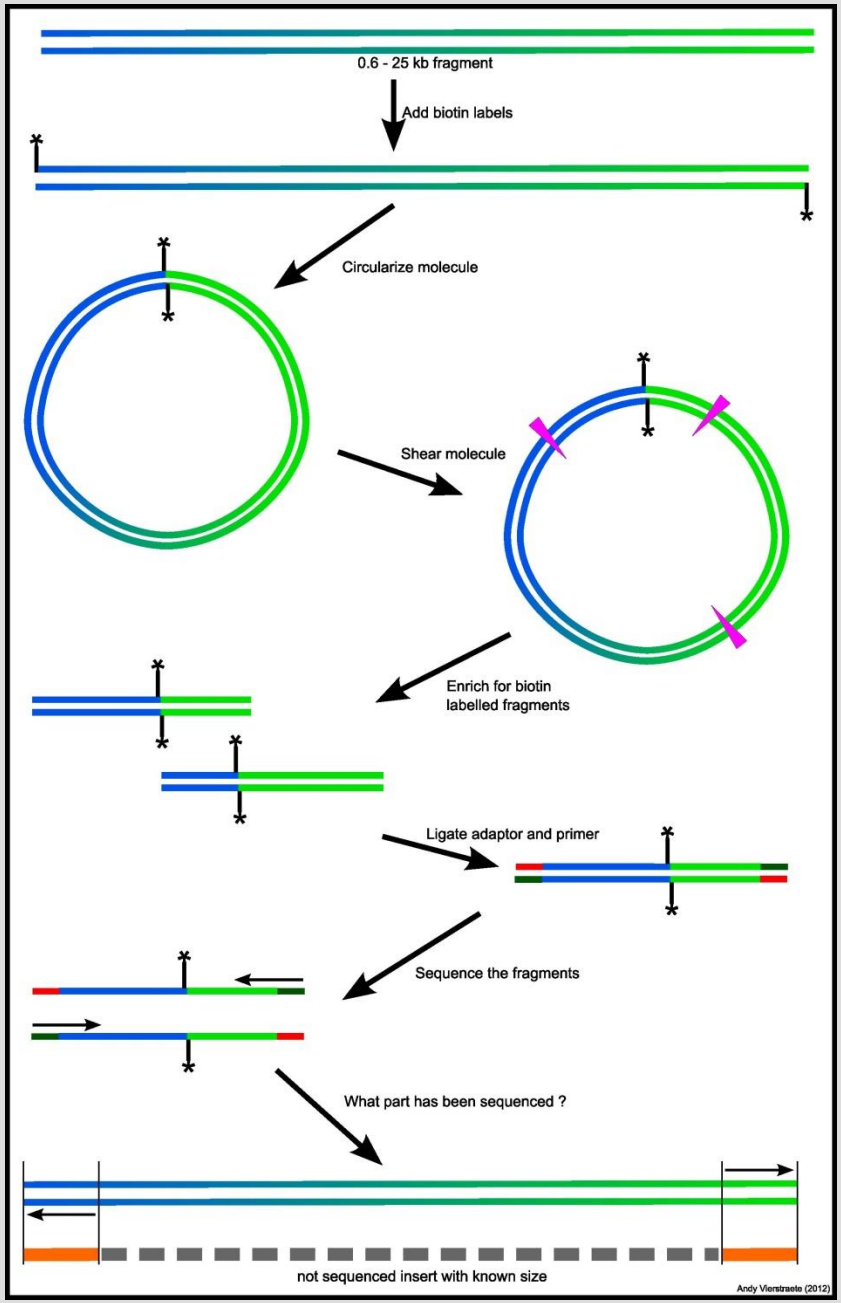
- **Эпигенетика**

ДНК-белковые и ДНК-ДНКовые взаимодействия: поиск сайтов связывания транскрипционных факторов, изучение пространственной организации хроматина, метилирование

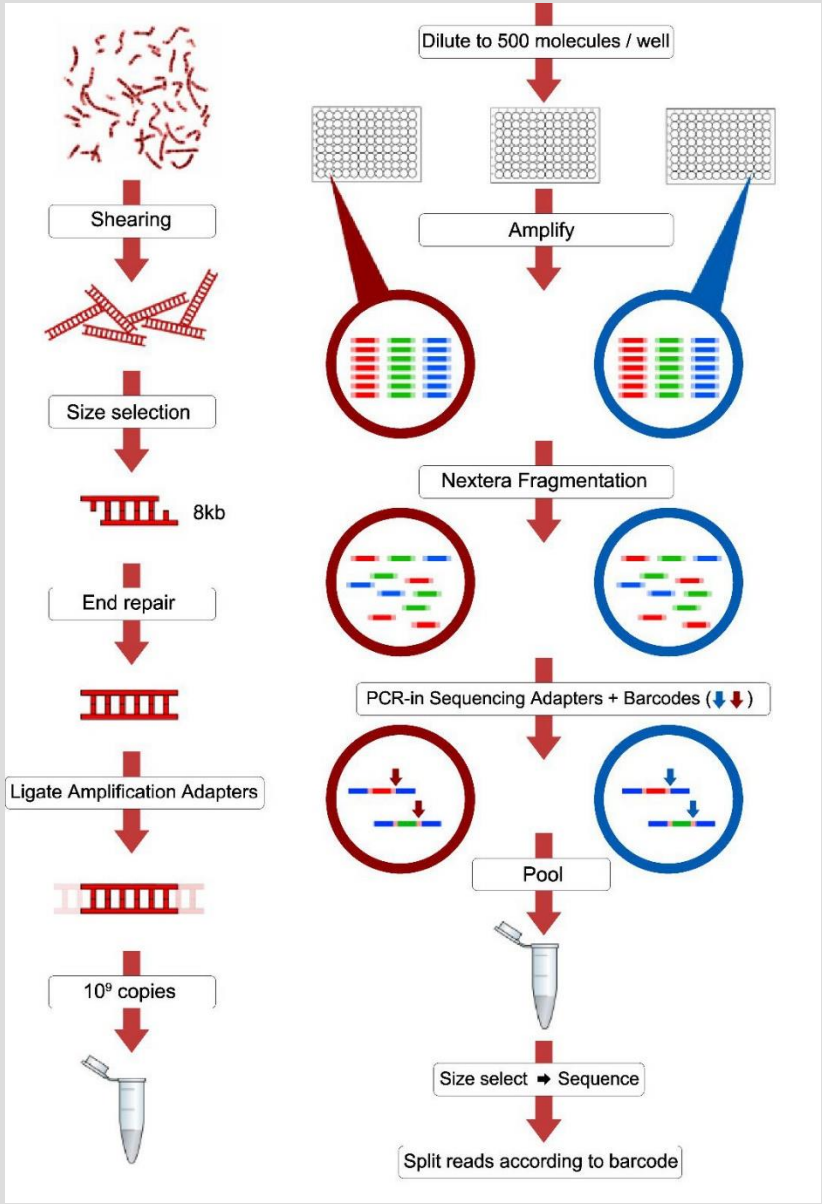
- **метагеномика**

анализ разнообразия микробных сообществ

Библиотеки с длинной вставкой (mate pair)



Синтетические длинные чтения



Voskoboynik A et al. eLife Sciences 2013;2:e00569

Принцип метода – разбиение ДНК на длинные фрагменты, разведение до нескольких сотен молекул и приготовление многих сотен независимых «мини-библиотек», представляющих небольшую часть генома. Секвенирование и сборка каждой такой мини-библиотеки проводится независимо.

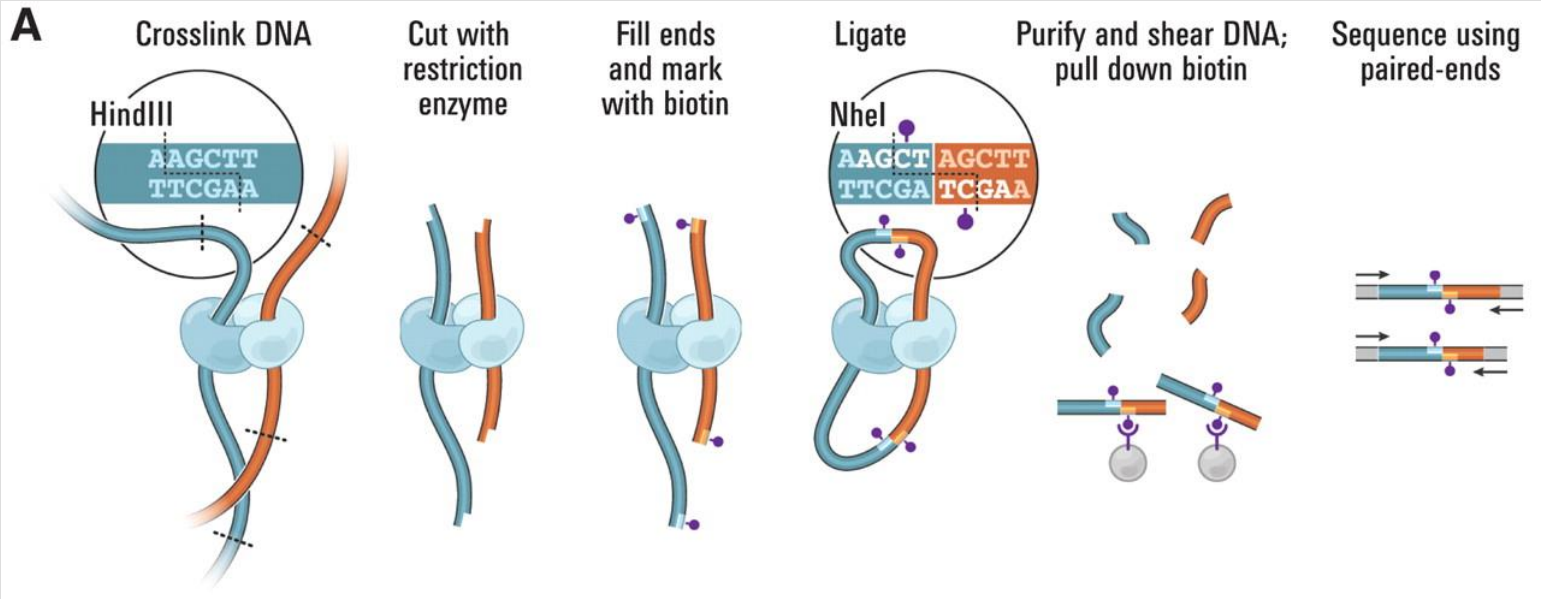
Применение:

- фазирование гаплотипов
- сборка геномов с большим количеством повторов
- поиск структурных перестроек

Реализация:

- Moleculo SLR (Illumina)
- 10x Genomics

Фиксация конформации хромосом (Hi-C)



Erez Lieberman-Aiden et al. Science 2009;326:289-293

Число контактов внутри одной хромосомы существенно выше, чем межхромосомных.

Дальнейшее развитие: cHiCago – используется сборка хроматина in vitro для полного исключения межхромосомных контактов (Putnam et al. Genome Res. 2016 Mar;26(3):342-50)

Таргетное ресеквенирование

Преимущества:

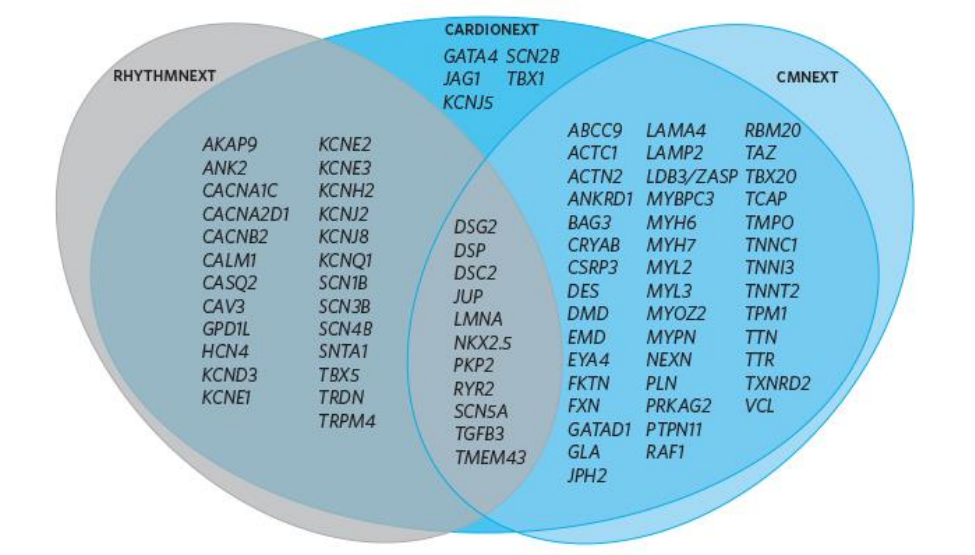
- легкость клинической интерпретации
- простота обработки и хранения данных
- быстрота получения данных
- дешевле полногеномного секвенирования
- минимальный риск нежелательных случайных находок

Недостатки:

- трудоемкий протокол
- требуется бОльшая глубина секвенирования
- невозможность детекции неизвестных вариантов
- затруднена детекция вариаций копийности
- затруднено сравнение результатов, полученных разными протоколами

Экзом:
~ 40-60 Мб (только кодирующие последовательности или кодирующие + UTR)

Панели генов
~ 0.5 – 5 Мб (от нескольких десятков до нескольких сотен генов)

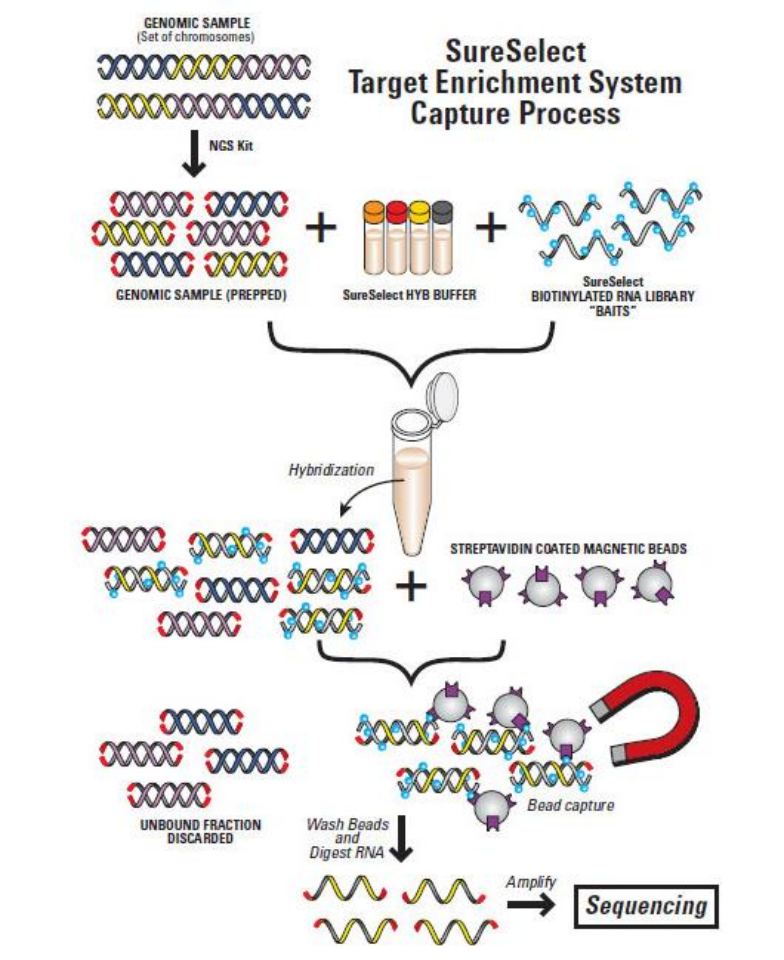


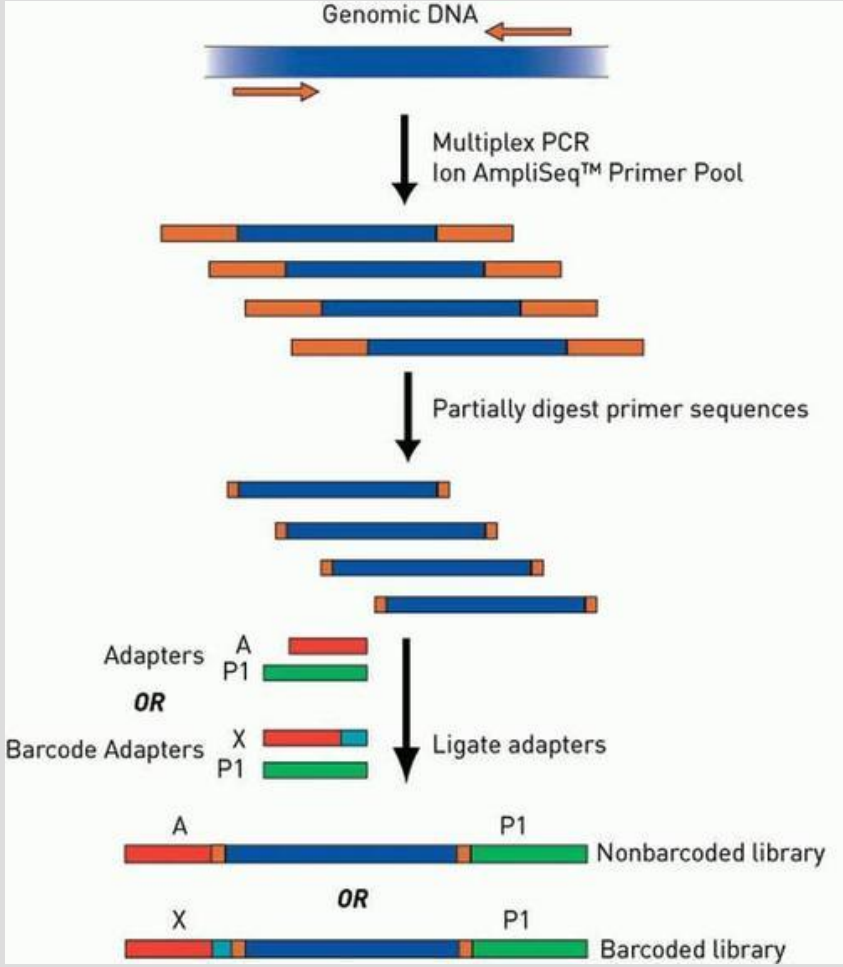
Cancer Gene Panel Includes:

ABL1	ERBB4	JAK2	PIK3CA
AKT1	FBXW7	JAK3	PTEN
ALK	FGFR1	KDR	PTPN11
APC	FGFR2	KIT	RB1
ATM	FGFR3	KRAS	RET
BRAF	FLT3	MET	SMAD4
CDH1	GNA11	MLH1	SMARCB1
CDKN2A	GNAQ	MPL	SMO
CSF1R	GNAS	NOTCH1	SRC
CTNNB1	HNF1A	NPM1	STK11
EGFR	HRAS	NRAS	TP53

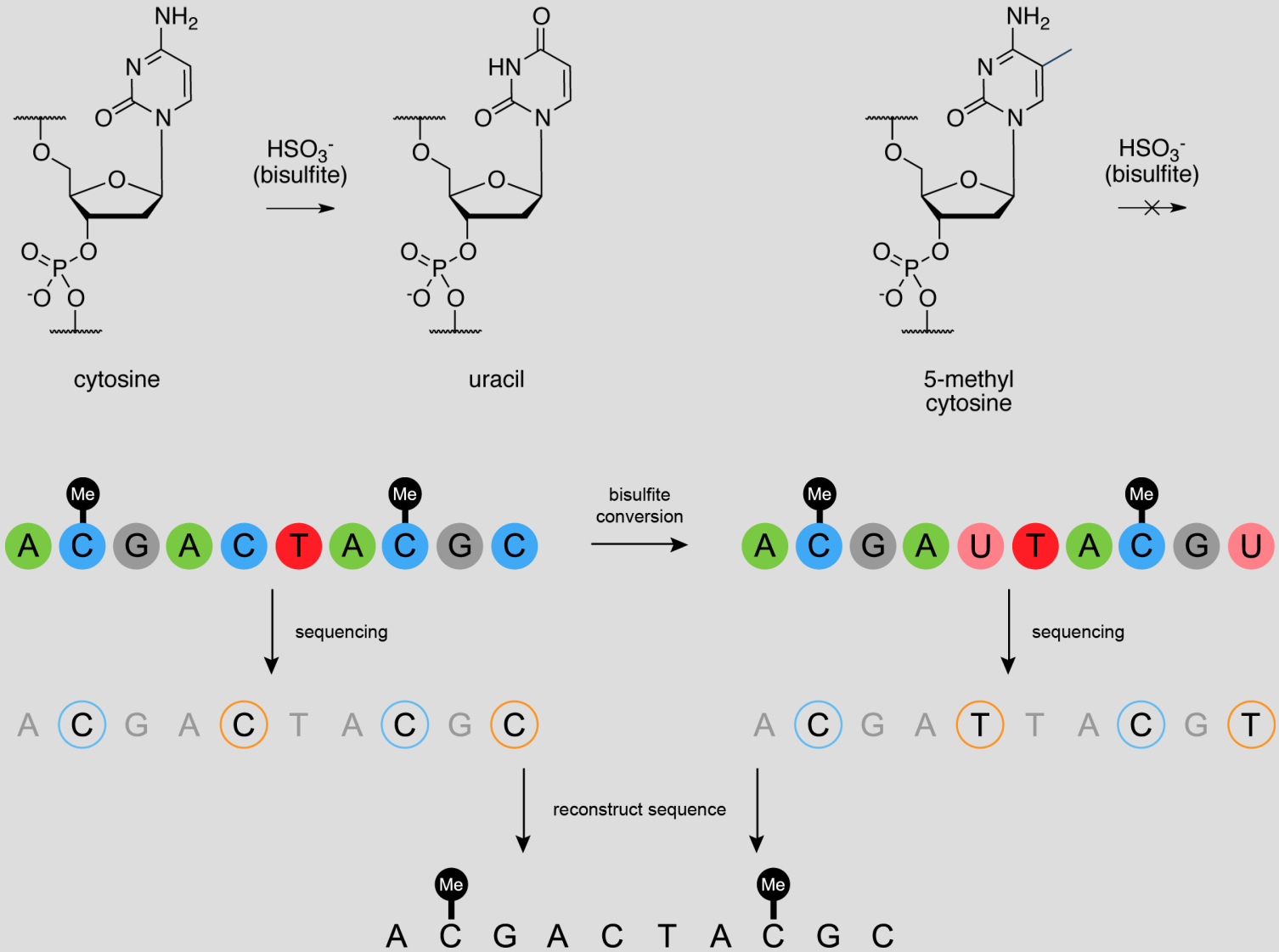
Таргетное ресеквенирование

Два основных метода отбора целевых участков генома для секвенирования:
обогащение и **амплификация**
Эффективность отбора далека от 100%!





Анализ метилирования – бисульфитная конверсия



RNA-seq

тотальная РНК

↓

отбор целевой фракции

↓

фрагментация (фрагменты длиной 100-300)

↓

синтез 1 цепи кДНК

↓

синтез 2 цепи кДНК

Отбор полиА-мРНК
+ быстрота, не требует обработки ДНКазой, легко масштабируется
- только для эукариот, критична целостность РНК

Деплеция рРНК
+ годится для прокариот, максимальная представленность регуляторных РНК, можно работать с деградированной РНК
- плохо подходит для немодельных объектов, требует обработки ДНКазой, большая глубина секвенирования

Случайные праймеры, для более равномерной представленности по длине. Термостабильные ревертазы

На этой стадии возможно приготовление strand-specific библиотек

NGS, области применения

платформа/задача	Illumina	Полупроводниковое секвенирование	Solid	SMRT (PacBio)	Helicos	Oxford Nanopore
Секвенирование de novo	++	++	-	+++	-	+++*
Полногеномное ресеквенирование	+++	++	++	+*	-	+*
Таргетное секвенирование	+++	+++	++	+*	-	+*
анализ экспрессии	+++	++	++	-	++	-
эпигенетика	+++	++	++	-	-	-
метагеномика	+++	+	-	-	+	+

* - при необходимости детекции перестроек, фазирования гаплотипов и т.д.

Спасибо за внимание!