
Введение в анализ данных NGS

— Анастасия Жарикова —
16 ноября 2021

Что такое секвенирование?

Что такое NGS?

Что можно секвенировать?

Для чего нужно секвенирование?

- Эволюция
- Филогения
- Клиника
- Метагеномика
- Анализ транскриптомов
- Single cell (различные приложения)
-

Эволюция

Доместикация риса

1 MSGSSADPSP SASTAGAAVS PLALLRAHGH GHGHLTATPP SGATGPAPPP
51 PSPASGSAPR DYRKGNWTLH ETLILITANR LDDDRRAGVG GAAAGGGGAG
101 SPPTPRSAEQ RWKWVENYCW KNGCLRSQNG CNDKWDNLLR DYKKVRDYES
151 RVAAAAATGG AAAANSAPLP SYWTMERHER KDCNLPTNLA PEVYDALSEV
201 LSRRAARRGG ATIAPT PPPP PLALPL PPPP PPSPPKPLVA QQQH HHHHGH
251 HH PPPP QPPP SSLQLPPAVV APPPASVSAE EEMSGSSESG EEEEGSGGEP
301 EAKRRRLSRL GSSVVR SATV VARTLVACEE KRERRHRELL QLEERRRLRL
351 EERTEVRRQG FAGLIAAVNS LSSAIHALVS DHRSGDSSGR

sh4

Li et al., Science, 2006

Дикий рис

AAG

Лизин

Культурный рис

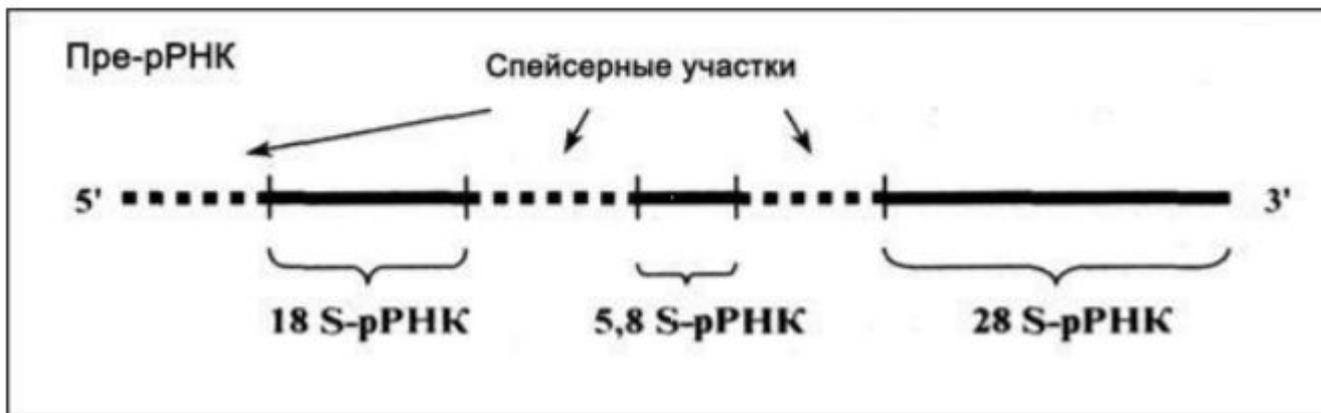
AAT

Аспарагин

Филогения

Спейсерные последовательности наиболее вариабельны с точки зрения эволюционной консервативности.

Секвенирование и анализ транскрибируемых спейсеров используется для изучения видового разнообразия и классификации близкородственных организмов.



Популяционные и клинические исследования

- 1000 genomes, gnomAD: частоты вариантов в популяциях
- GWAS: поиск полиморфизмов, ассоциированных с болезнями:
 - моногенные (муковисцидоз, ген CFTR)
 - полигенные (ишемическая болезнь сердца, шизофрения, ...)
- Фармакогенетика и индивидуальные особенности
 - варфарин
 - исследование генов из системы свертывания крови

Что бывает

enseqlopedia

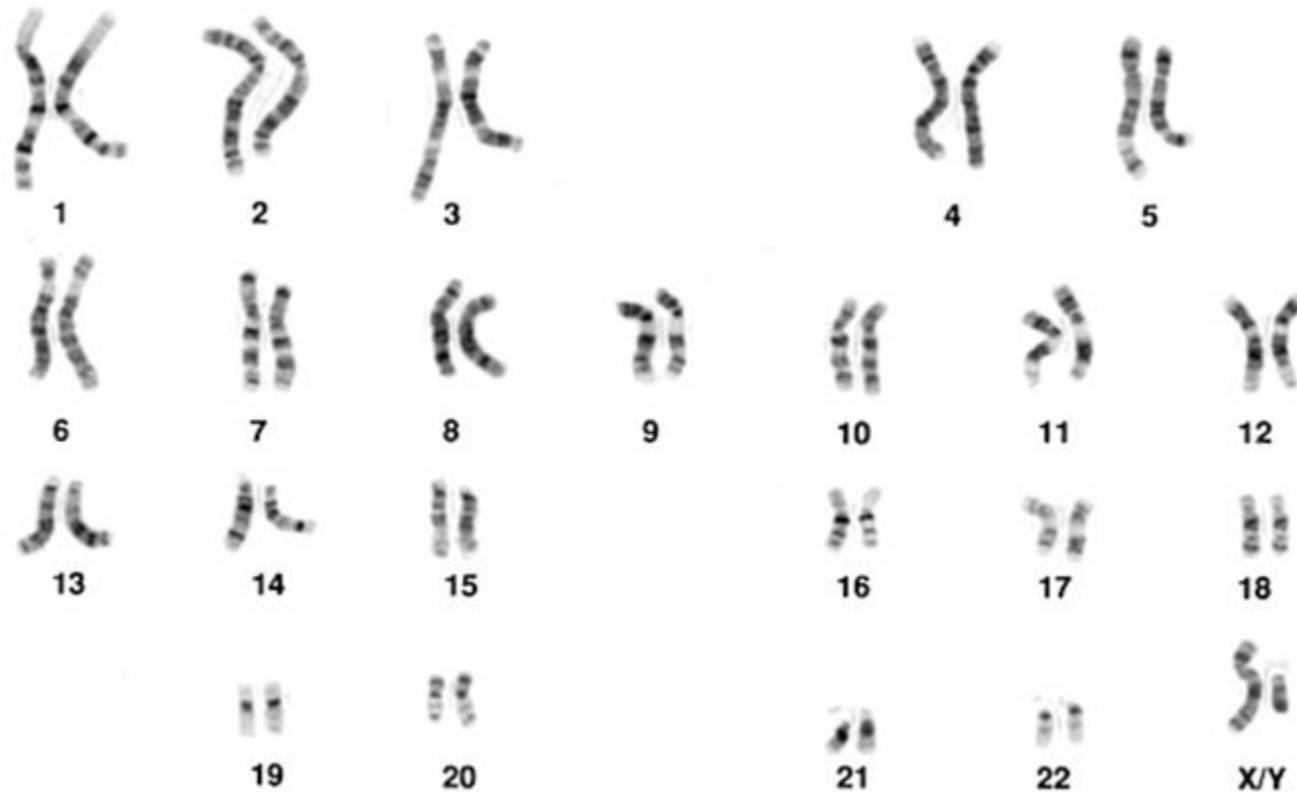
- DNA-seq
- RNA-seq
- HiC
- Chip-seq
- ATAC-seq
- DNase-seq
- GRO-seq
- Ribo-seq
- ...

Что бывает

enseqlopedia

- **DNA-seq**
- RNA-seq
- HiC
- Chip-seq
- ATAC-seq
- DNase-seq
- GRO-seq
- Ribo-seq
- ...

У человека 23 пары хромосом. Много или мало?



Число хромосом у разных видов



Гиббоны - 44



Макака - 42



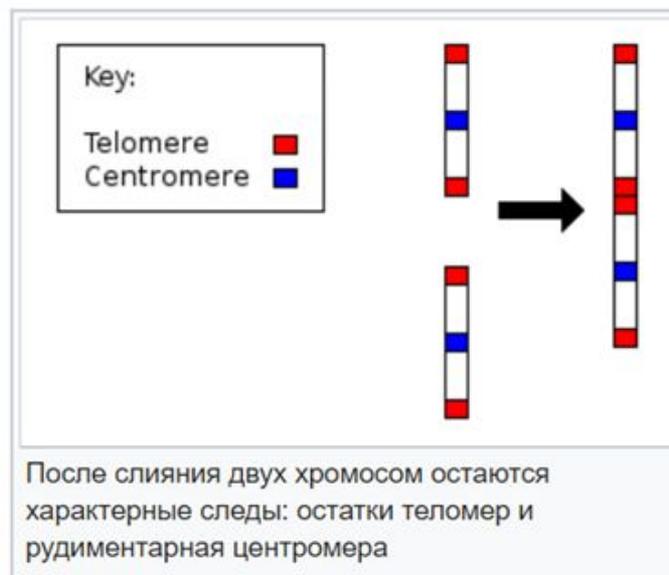
Капуцин - 54



48



46



Число хромосом у разных видов

Муравей (*Myrmecia pilosula*) – 2

Плодовая мушка - 8

Арабидопсис – 10

Голубь – 16

Кошка – 38

Лиса - 34

Мышь - 40

Собака – 78

Утка – 80

Сазан - 104

Корова – 120

Рак (*Cambarus clarkii*) – 200

Хвощ – 216

Краб - 254

Бабочка – 380



Размер генома у разных видов



Количество белок-кодирующих генов у разных видов

Картофель – 39 000

Человек ~ 20 000

Черви – 14 000

Мухи – 12 000

Грибы – 6 000

Бактерии – 2 000 – 4 000

Микоплазмы - 500

Вирус гриппа – 12

Какие еще гены бывают?

Книжка

<http://book.bionumbers.org/>

Секвенирование ДНК бывает



Возможности ресеквенирования

Можно ресеквенировать:

- полный геном
- экзом (кодирующую часть генома)
- отдельные таргетные гены или области

!!! Выбор зависит от бюджета и целей исследования!!!

Экзомное ресеквенирование

“Плюсы”

- Небольшой объем кодирующих данных - ниже цена
- Кодирующие последовательности лучше изучены
- Большое число болезнетворных мутаций находится в кодирующей последовательности

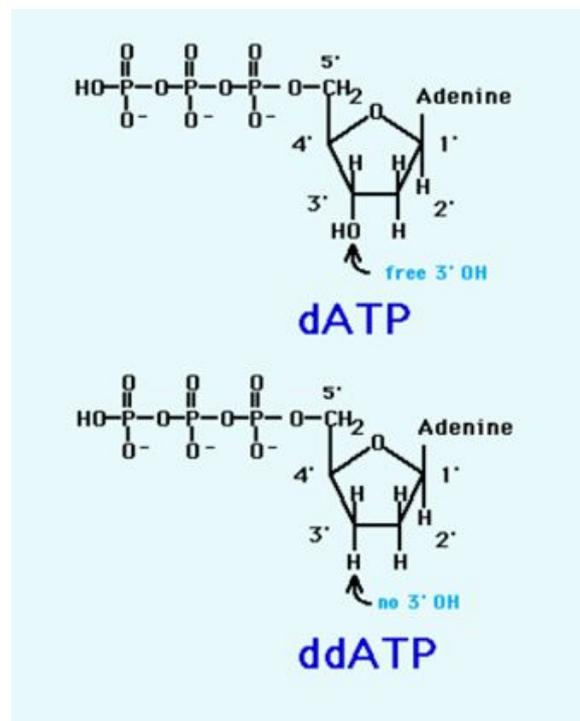
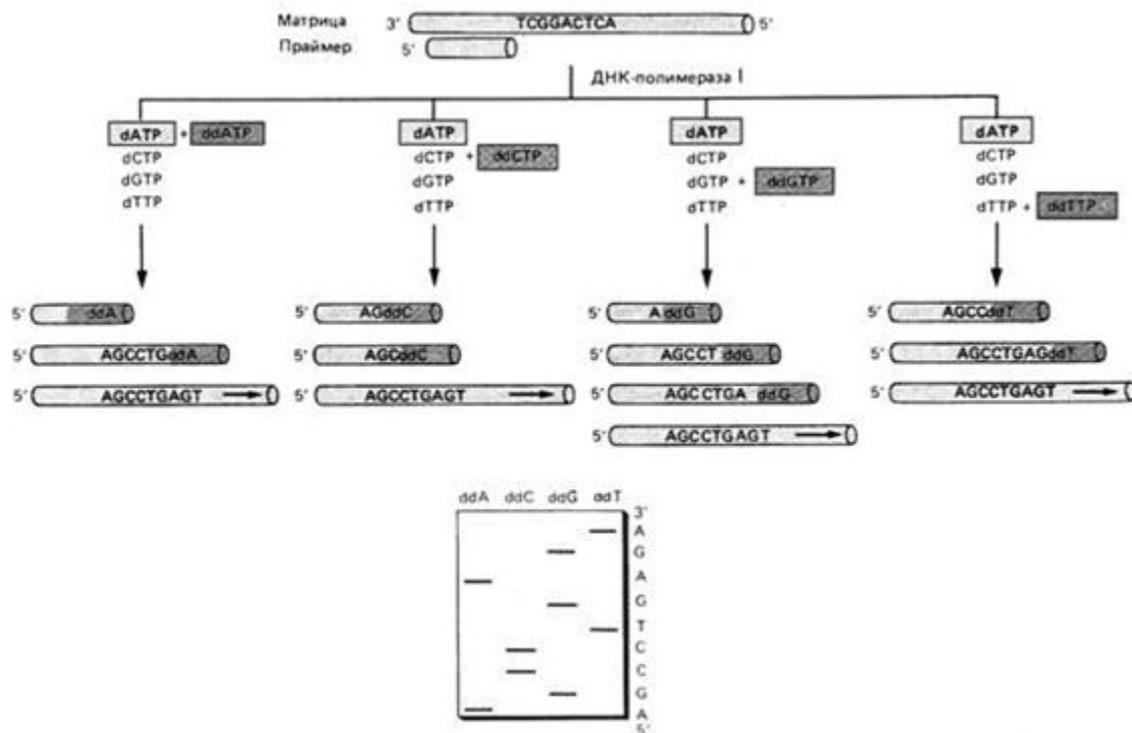
“Минусы”

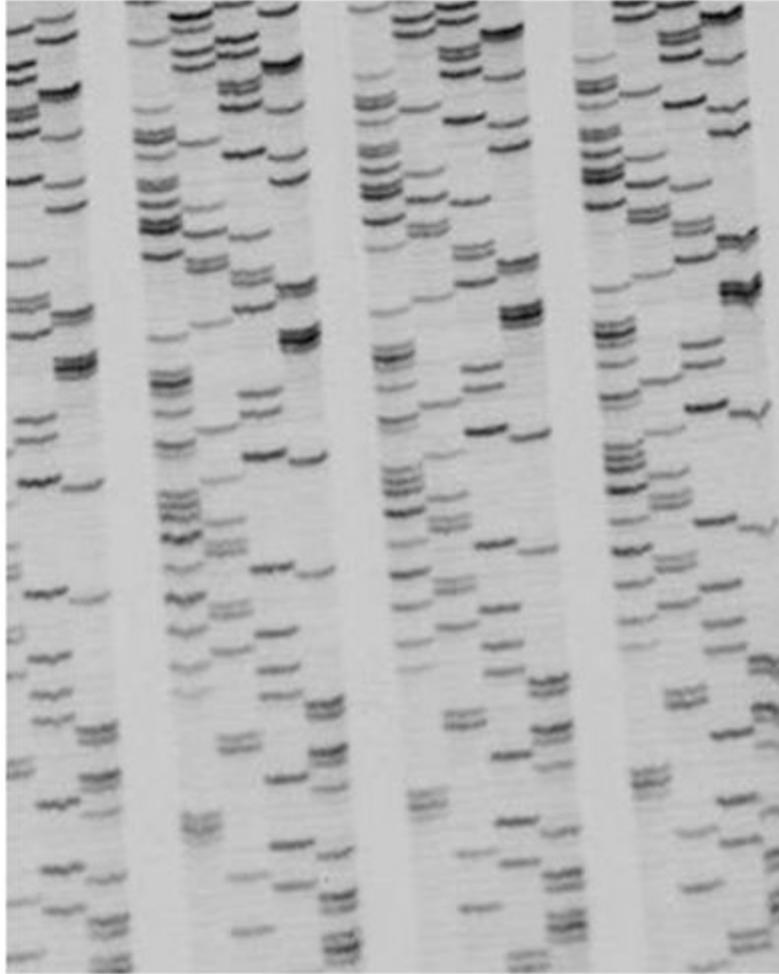
- Нет информации о некодирующих участках
- Неравномерность покрытия экзонов

Какие бывают мутации

- SNV: однонуклеотидные варианты, т.е. изменение одного нуклеотида
- Короткие вставки и делеции (~ 50 п.н.)
- Структурные варианты: инверсии и транслокации; CNV
- Анеуплоидии: нульсомии, моносомии, трисомии, полисомии
- Полиплоидизация

Секвенирование ДНК - метод "терминаторов"

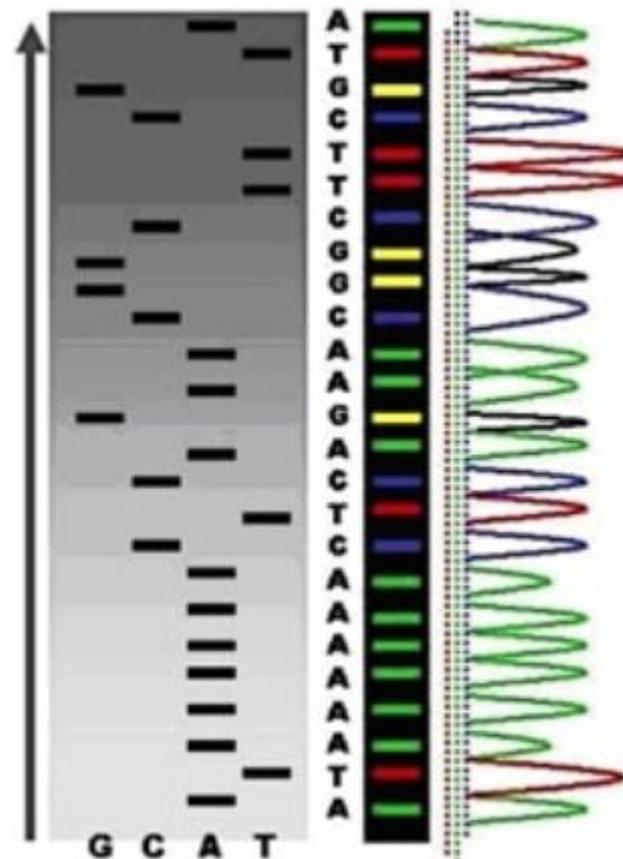




Секвенирование ДНК - метод “терминаторов”

~ 1000 п.н.

“Золотой стандарт”



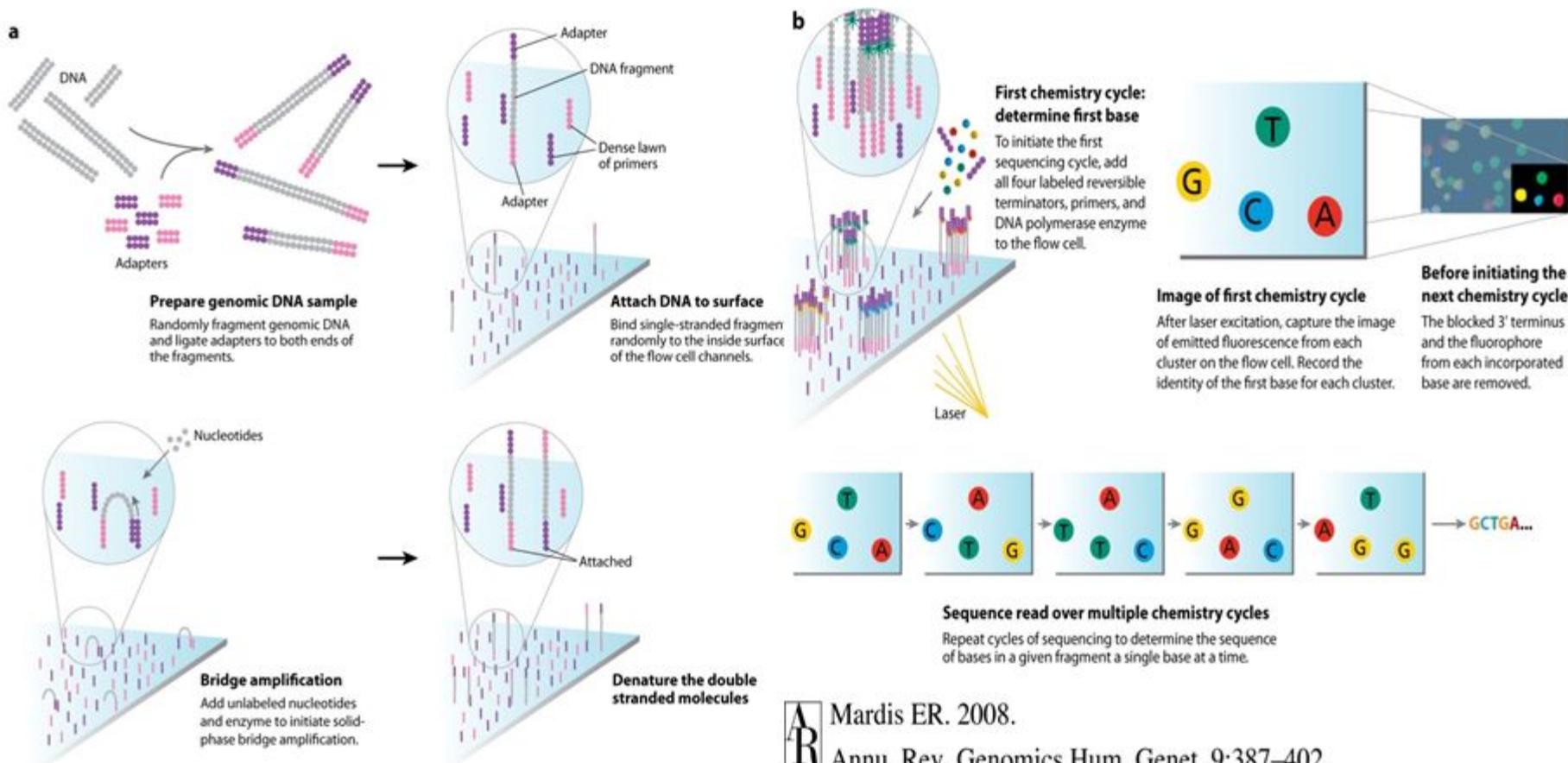
Next-generation sequencing (NGS)

“+” - одновременно идет сиквенс большого количества разных фрагментов

“-” - прочтения длиной 75 - 150 нукл

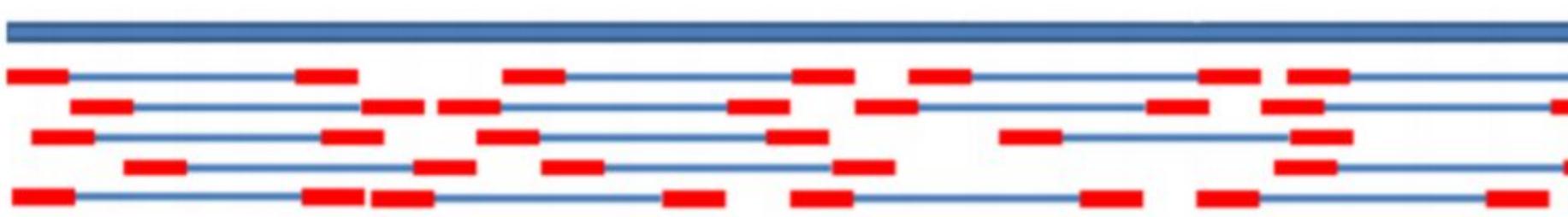
Next-generation sequencing (NGS) - Illumina

Есть и другие приборы!



Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387-402

Парно-концевые и одно-концевые чтения



ATGCAGA????????????????CACTTTA

Для Illumina характерная длина чтения 75-150 нуклеотидов

Что может пойти не так?

Димеры адаптеров: адаптеры соединяются друг с другом без фрагмента ДНК между ними

Норма



Димер

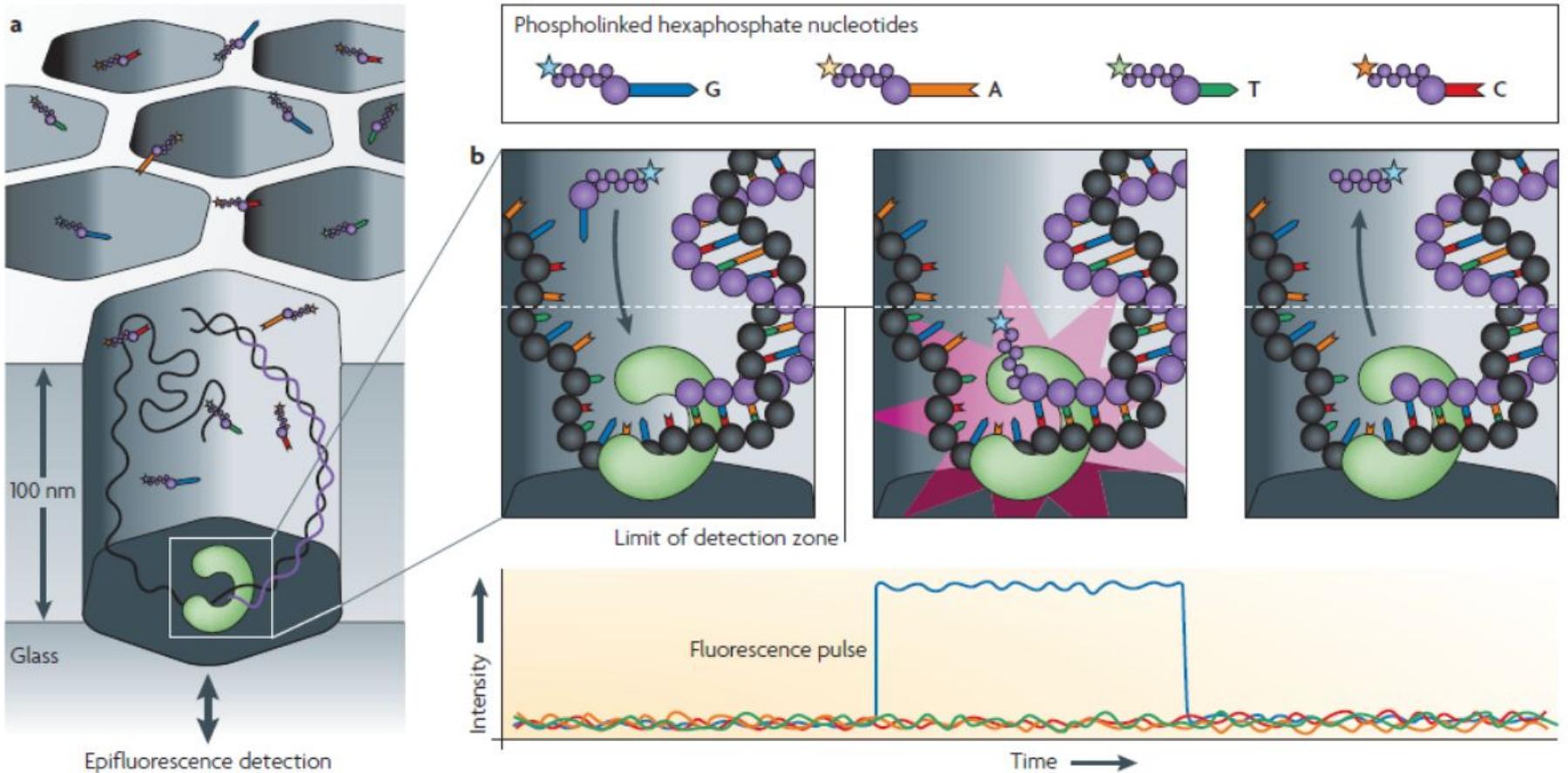


Фрагмент ДНК слишком короткий, чтение захватывает последовательность адаптера



Одномолекулярное секвенирование Pacific Biosciences

Pacific Biosciences — Real-time sequencing



Pacific Biosciences

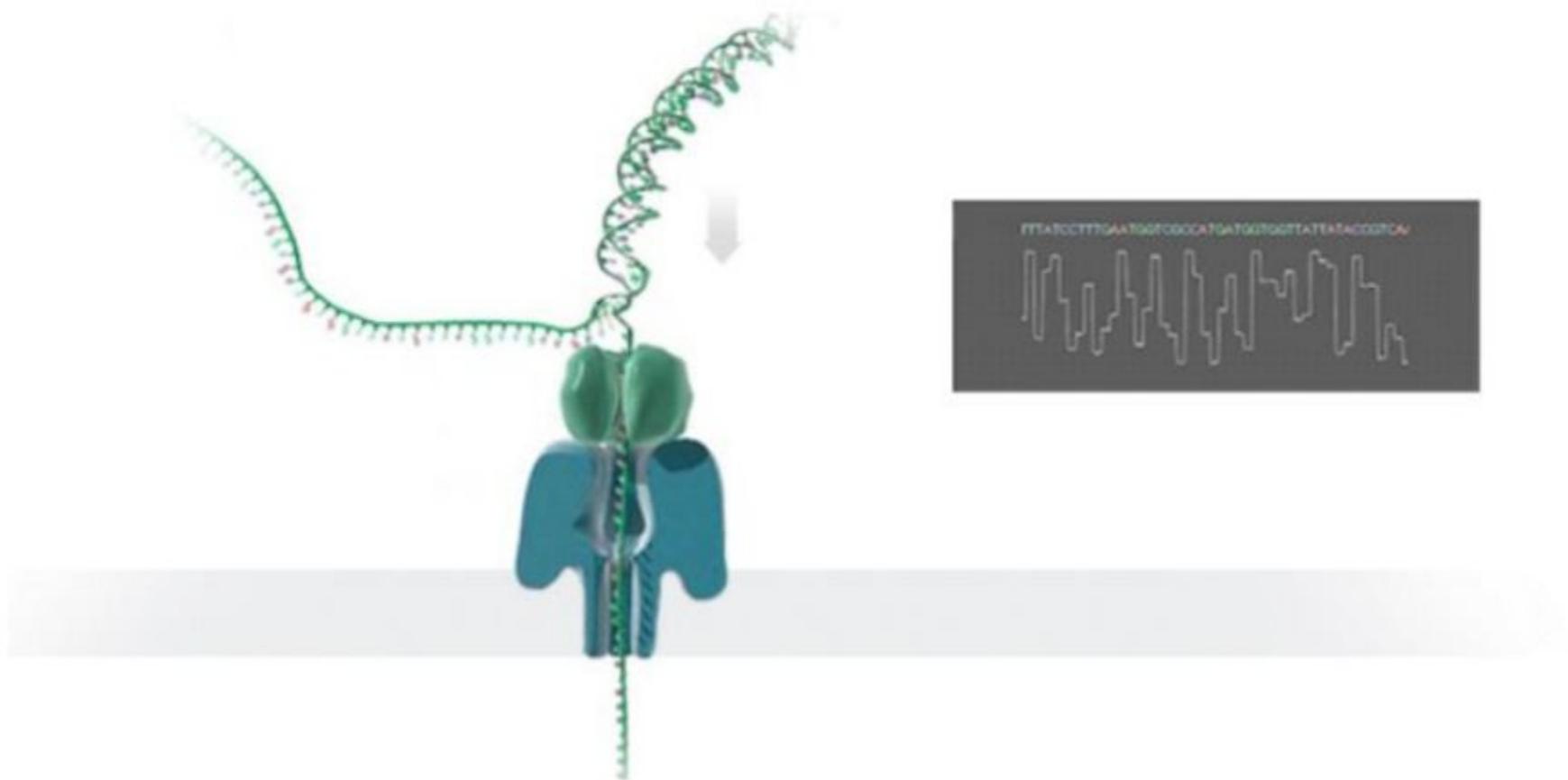
“плюсы”

- длина прочтений 20000-60000
- без амплификации
- быстро

“минусы”

- большой процент ошибок
- цена

Oxford Nanopore



МУЛЬТИК

Oxford Nanopore

“плюсы”

- длина прочтений 20000-60000
- без амплификации
- быстро
- компактность и мобильность

“минусы”

- большой процент ошибок



Что же выбрать?

Все зависит от задачи

Комбинировать платформы

Увеличивать покрытие

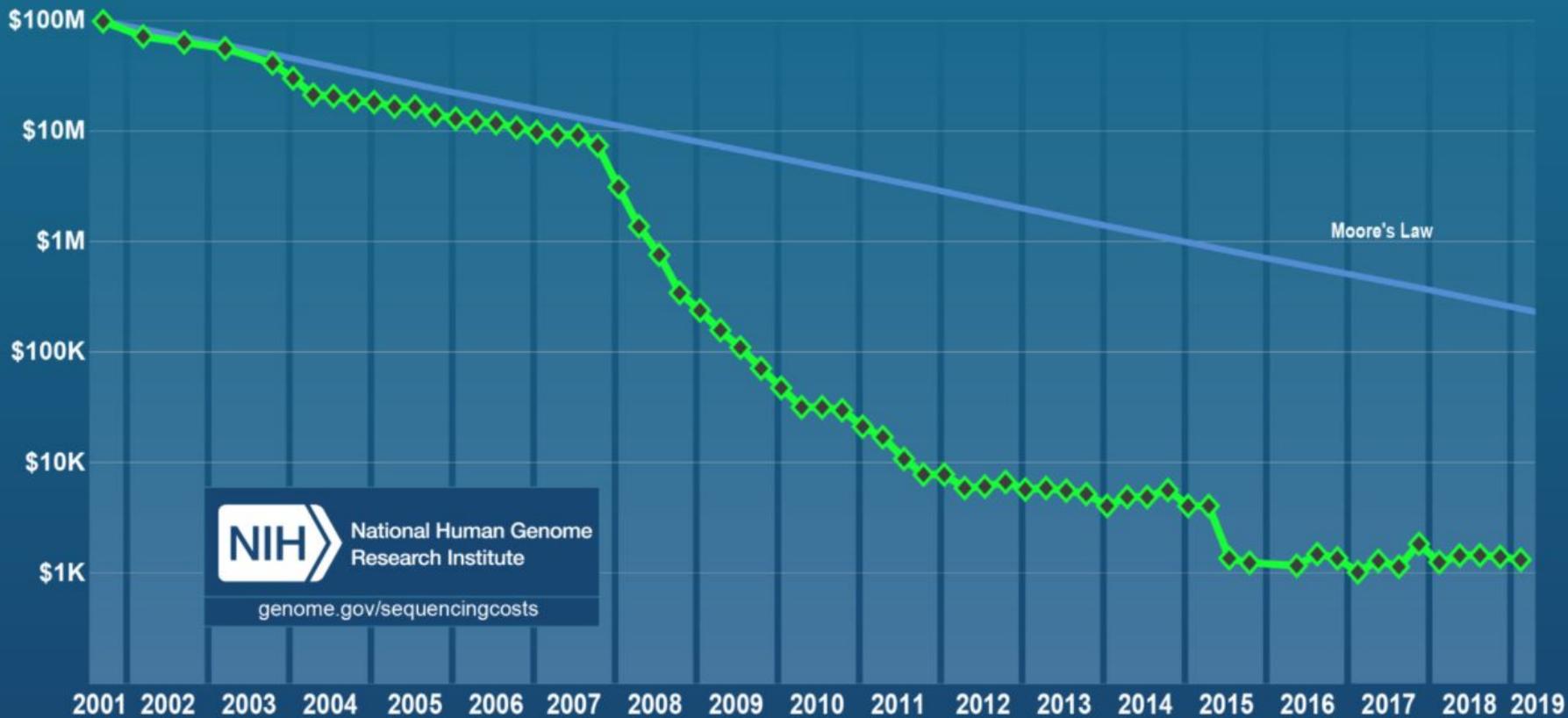
Что почитать

[skygen](#)

[Нанопоровое секвенирование](#)

[Обзор технологий секвенирования](#)

Cost per Genome



Откуда взять чтения?

<https://www.ncbi.nlm.nih.gov/sra>

SRX8794662: Whole exome seq of primary culture established from PDX tumor: Sample E9

1 ILLUMINA (Illumina HiSeq 4000) run: 31.4M spots, 6.3G bases, 2.3Gb downloads

Design: "Exom enrichment with Agilent SureSelect Human All Exon V6, based on UCSC hg19, GRCh37, February 2009"

Submitted by: NIH-phs002051

Study: DNA methylation in rhabdomyosarcoma PDX and PDX-derived primary cells

[PRJNA641459](#) • [SRP273116](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: Tumor DNA sample from N/A of a human participant in the dbGaP study "DNA Methylation in Rhabdomyosarcoma PDX and PDX-Derived Primary Cells"

[SAMN15468651](#) • [SRS7062698](#) • [All experiments](#) • [All runs](#)

Organism: [Homo sapiens](#)

Library:

Name: ON-2018/8626: E9

Instrument: Illumina HiSeq 4000

Strategy: WXS

Source: GENOMIC

Selection: PCR

Layout: PAIRED

The SRA run(s) below contain human sequence [\(more...\)](#)

Runs: 1 run, 31.4M spots, 6.3G bases, 2.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR12291396	31,383,123	6.3G	2.3Gb	2020-08-27

Как достать чтения из SRA?

Sra toolkit - <https://www.ncbi.nlm.nih.gov/sra/docs/sradownload/>

Обратите внимание, что при запуске sra toolkit важно сразу указывать одноконцевые чтения или парноконцевые!!!

Важная информация

SRX8794662: Whole exome seq of primary culture established from PDX tumor: Sample E9

1 ILLUMINA (Illumina HiSeq 4000) run: 31.4M spots, 6.3G bases, 2.3Gb downloads

Design: "Exom enrichment with Agilent SureSelect Human All Exon V6, based on UCSC hg19, GRCh37, February 2009"

Submitted by: NIH-phs002051

Study: DNA methylation in rhabdomyosarcoma PDX and PDX-derived primary cells

[PRJNA641459](#) • [SRP273116](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: Tumor DNA sample from N/A of a human participant in the dbGaP study "DNA Methylation in Rhabdomyosarcoma PDX and PDX-Derived Primary Cells"

[SAMN15468651](#) • [SRS7062698](#) • [All experiments](#) • [All runs](#)

Organism: [Homo sapiens](#)

Library:

Name: ON-2018/8626: E9

Instrument: Illumina HiSeq 4000

Strategy: WXS

Source: GENOMIC

Selection: PCR

Layout: PAIRED

The SRA run(s) below contain human sequence ([more...](#))

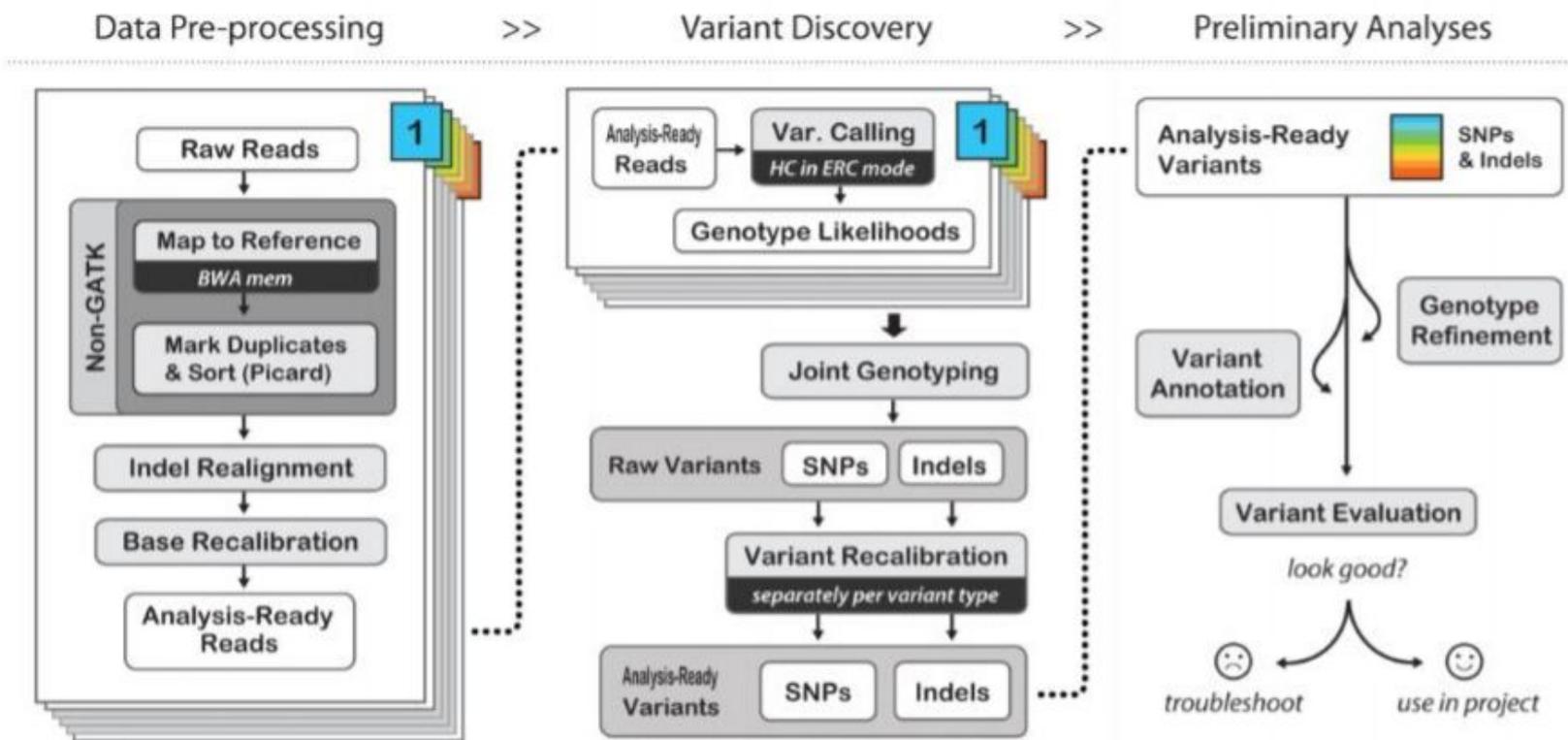
Runs: 1 run, 31.4M spots, 6.3G bases, 2.3Gb

Run	# of Spots	# of Bases	Size	Published
SRR12291396	31,383,123	6.3G	2.3Gb	2020-08-27

ID: 11430914

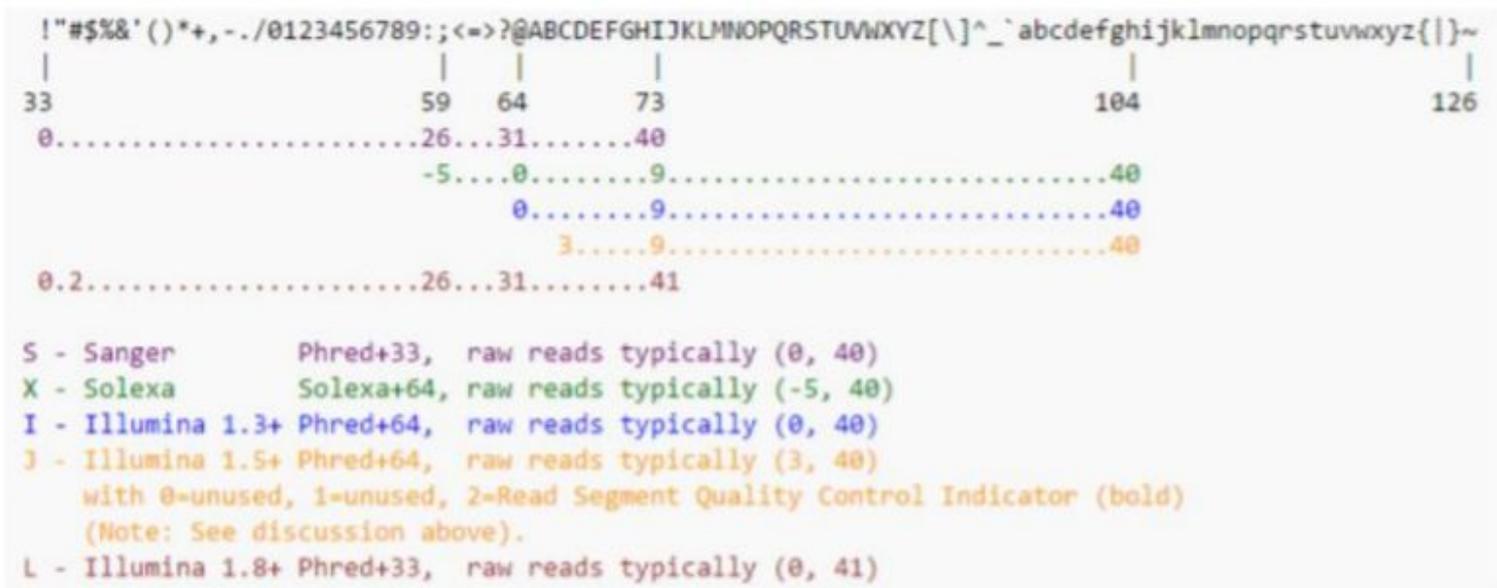
Обработка данных

Создаем программный конвейер



Fastq формат

```
                @HWI-ST992:147:D22HDACXX:3:1112:14175:15297 2:N:0:GGCTAC
Последовательность TAATGGCTTTTCCAAAACGCTCCACTCTTAAAGATGTGTATAAGAGACAGCAACAACAATTA
+
Качество           8??DDDBEDHHFHJJJJJJAFAFGIIIIIIGIGEEGIIIIHBFGGEEGCGIJIFFIDIIJJIIII
```



Качество чтений

P - вероятность ошибки

Q - параметр качества (Phred Quality Score)

Значения Q: 1-40

Q > 20 считается хорошим качеством

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

Вероятность ошибки	Q
0.001 (точность 99,9%)	30
0.01 (точность 99%)	20
0.1 (точность 90%)	10

Пересчет качества в вероятность ошибки

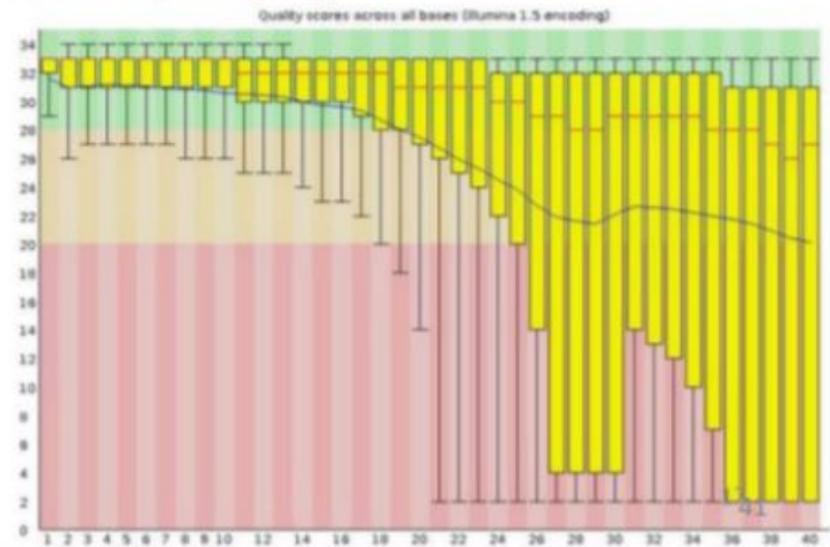
Phred Quality Score	Символ	Вероятность ошибки	Точность
10	+	1/10	90%
20	5	1/100	99%
30	?	1/1000	99,9%
40	!	1/10 000	99,99%
50	S	1/100 000	99,999%
60]	1/1 000,000	99,9999%

fastQC

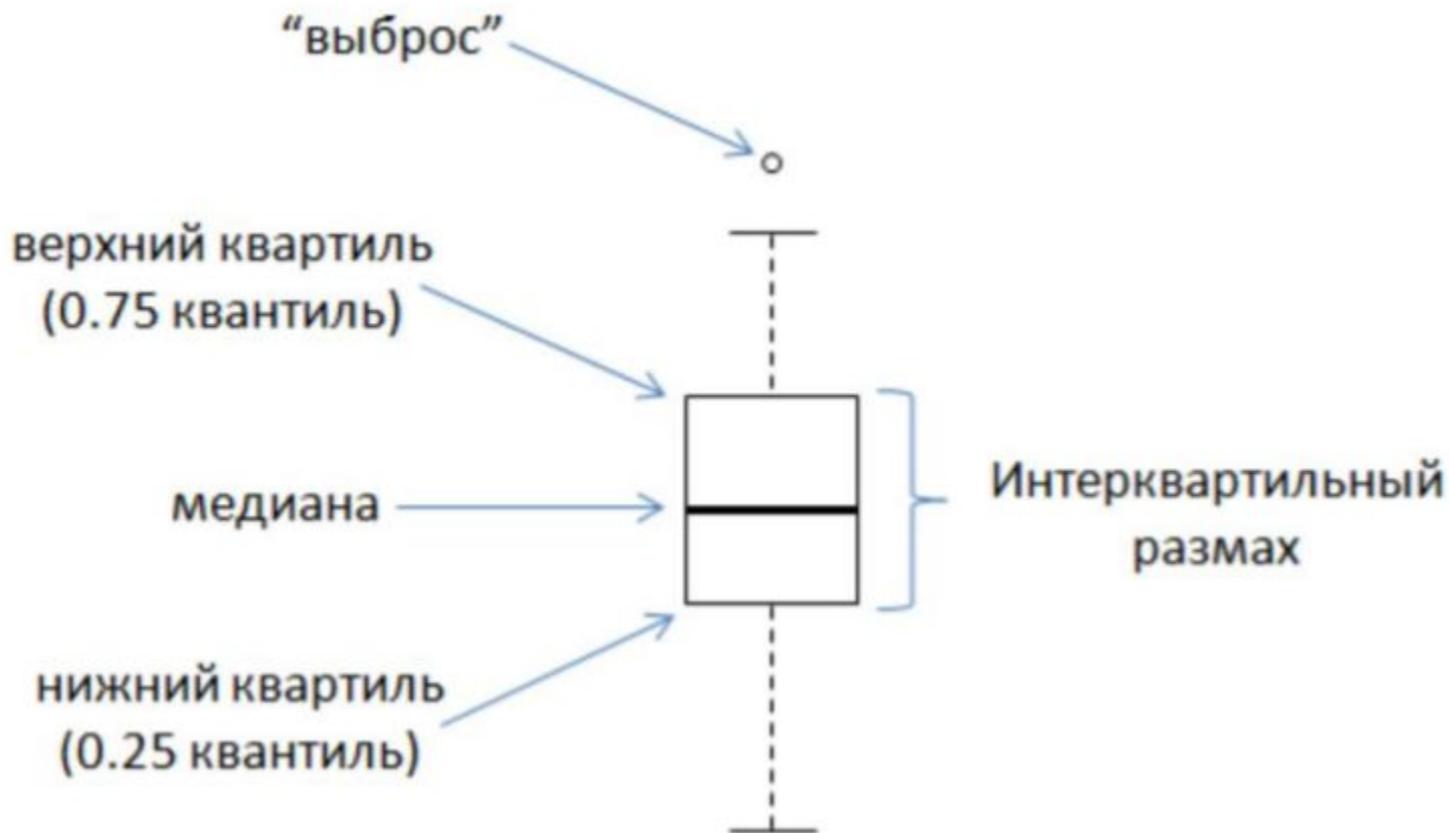
✔ Per base sequence quality



✘ Per base sequence quality



“Ящик с усами” / диаграмма размахов / boxplot



fastQC

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

Программа fastQC установлена на kodo mo

Версию с графическим интерфейсом можно поставить на свой компьютер

На сайте отличное руководство с примерами данных хорошего и плохого качества

<https://www.bioinformatics.babraham.ac.uk/training.html> - полезности