Введение в анализ данных NGS

Анастасия Жарикова

15 ноября 2022

Что такое секвенирование?

Что такое NGS?

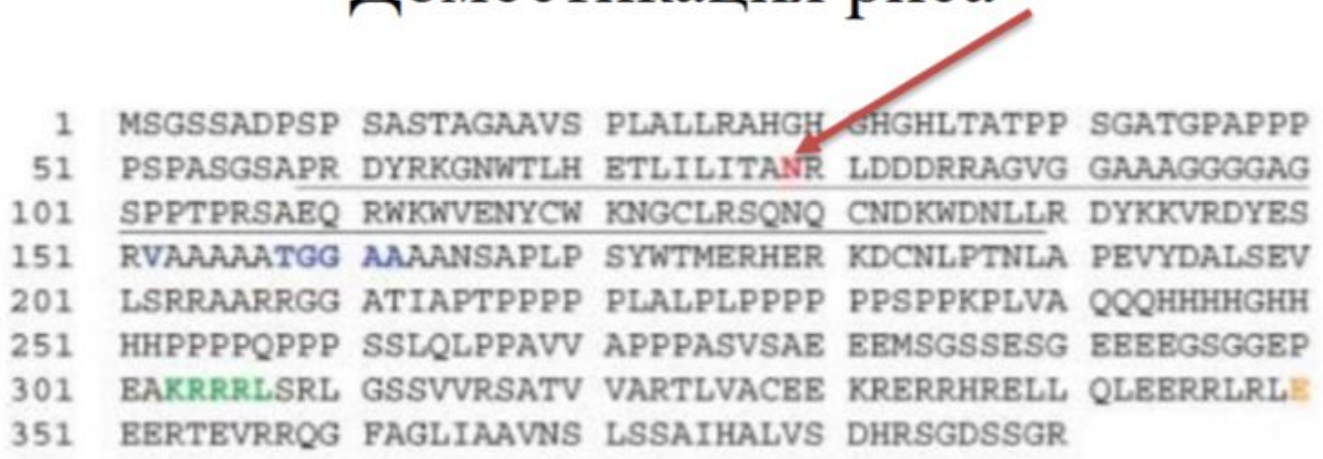
Что можно секвенировать?

Для чего нужно секвенирование?

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

Эволюция

Доместикация риса



sh4

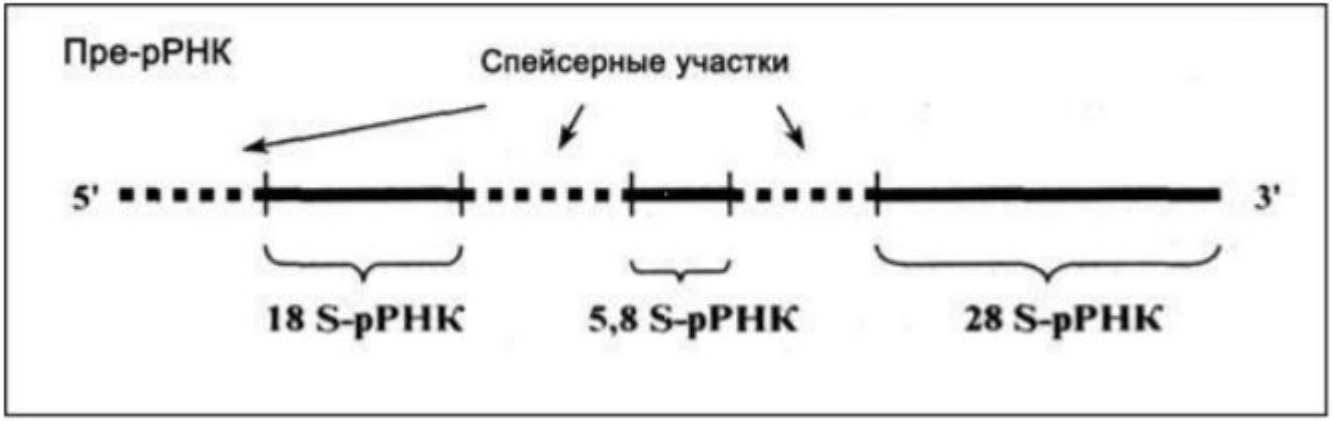
Li et al., Science, 2006

Дикий рис	Культурный рис
AAG	AAT
Лизин	Аспарагин

Филогения

Спейсерные последовательности наиболее вариабельны с точки зрения эволюционной консервативности.

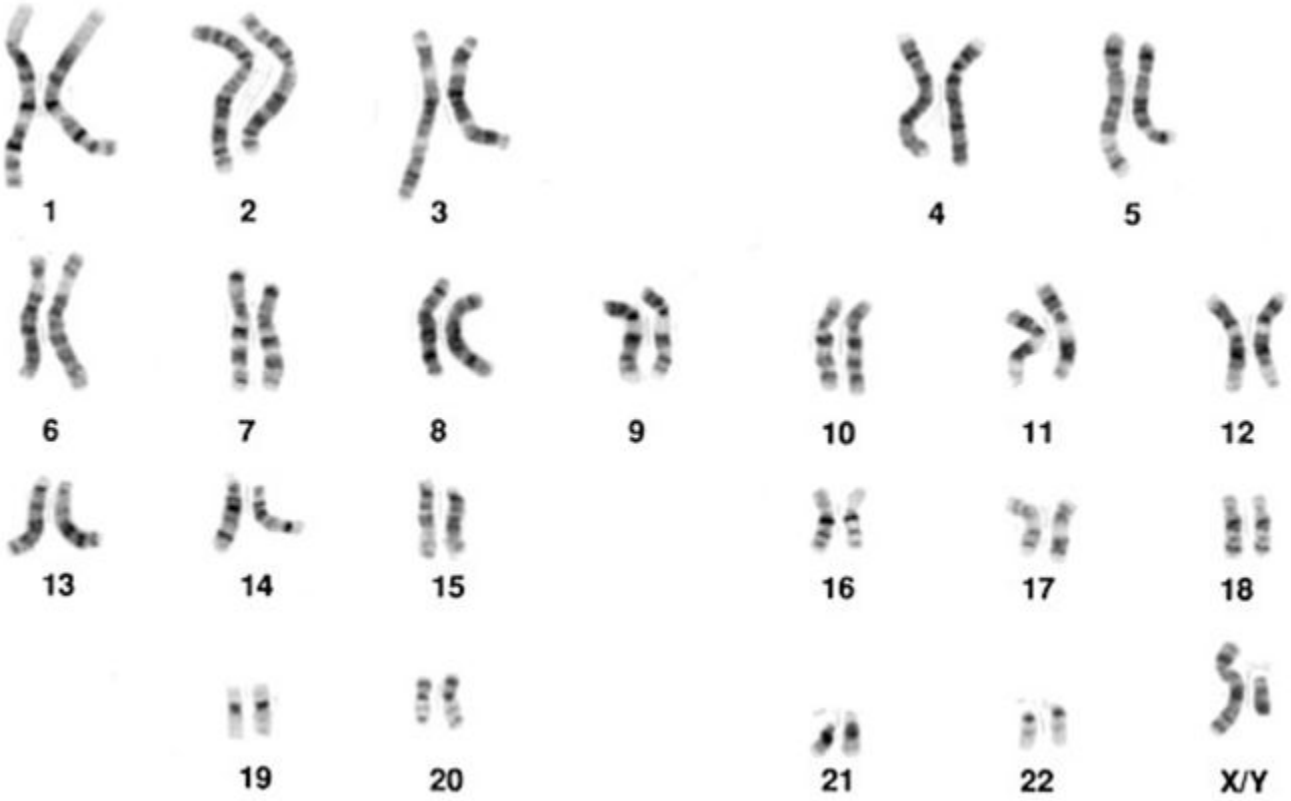
Секвенирование и анализ транскрибируемых спейсеров используется для изучения видового разнообразия и классификации близкородственных организмов.



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

У человека 23 пары хромосом. Много или мало?



Кариотип

Число хромосом у разных видов

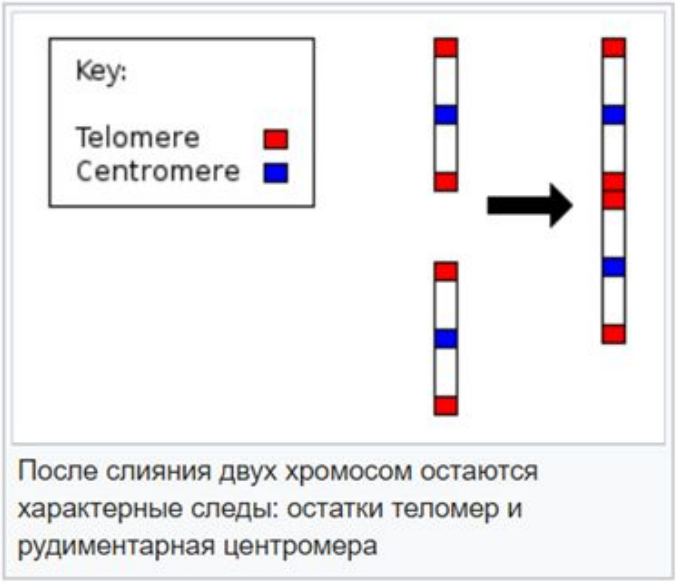
Гиббоны - 44

48

46

Макака - 42

Капуцин - 54



После слияния двух хромосом остаются характерные следы: остатки теломер и рудиментарная центромера

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

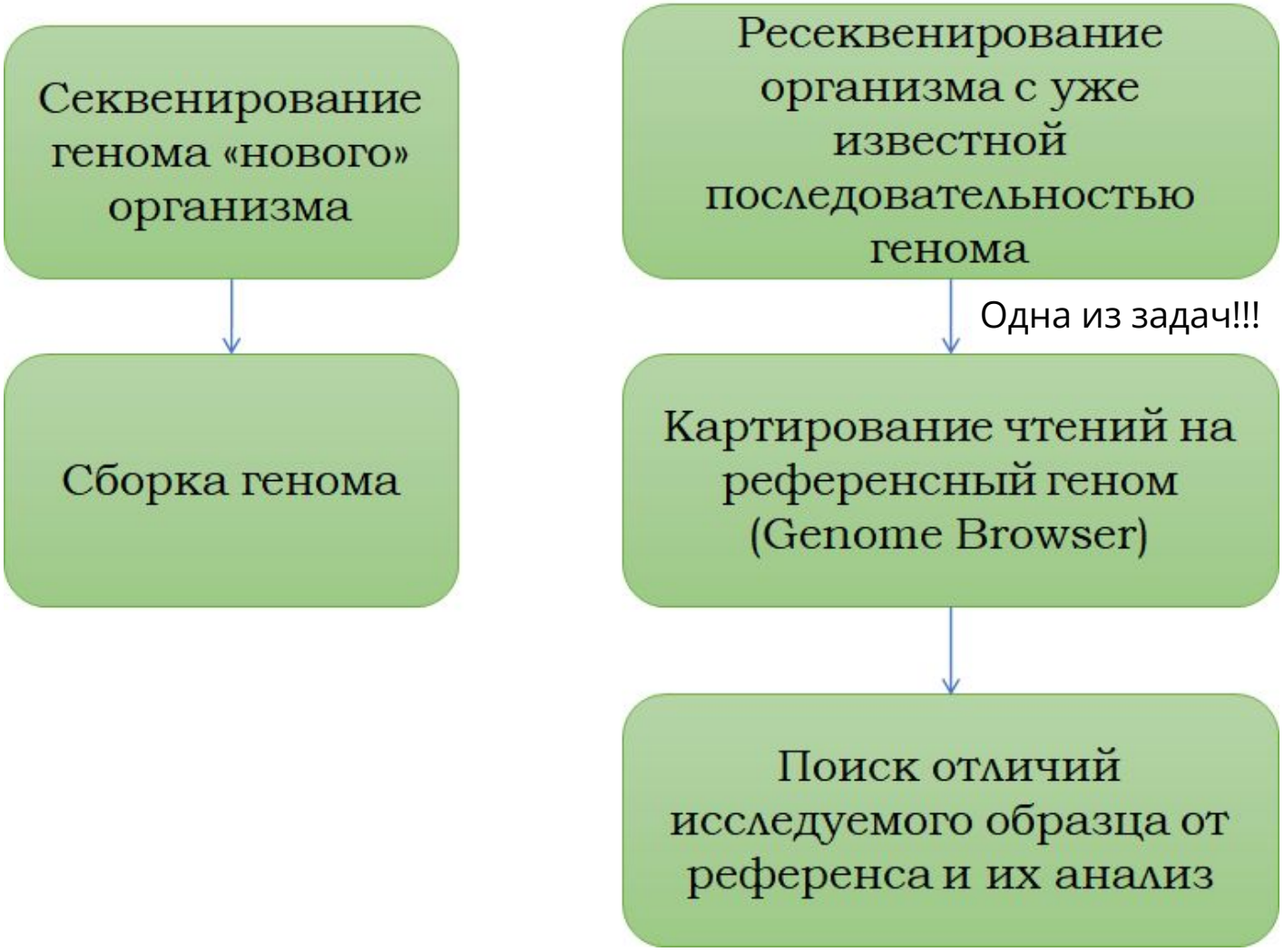
Картофель – 39 000
Человек ~ 20 000
Черви – 14 000
Мухи – 12 000
Грибы – 6 000
Бактерии – 2 000 – 4 000
Микоплазмы - 500
Вирус гриппа – 12

Какие еще гены бывают?

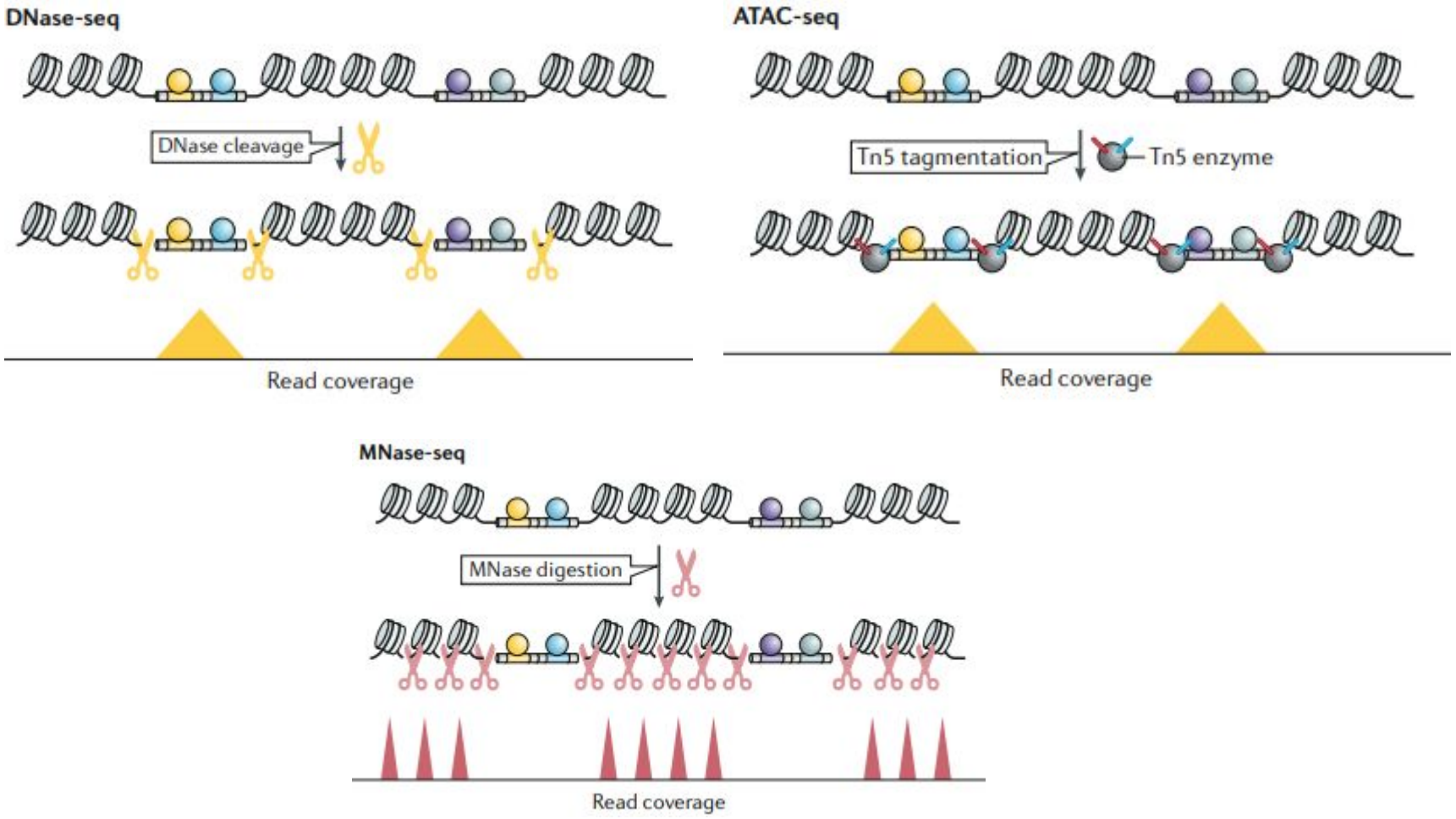
Книжка

http://book.bionumbers.org/

Секвенирование ДНК бывает

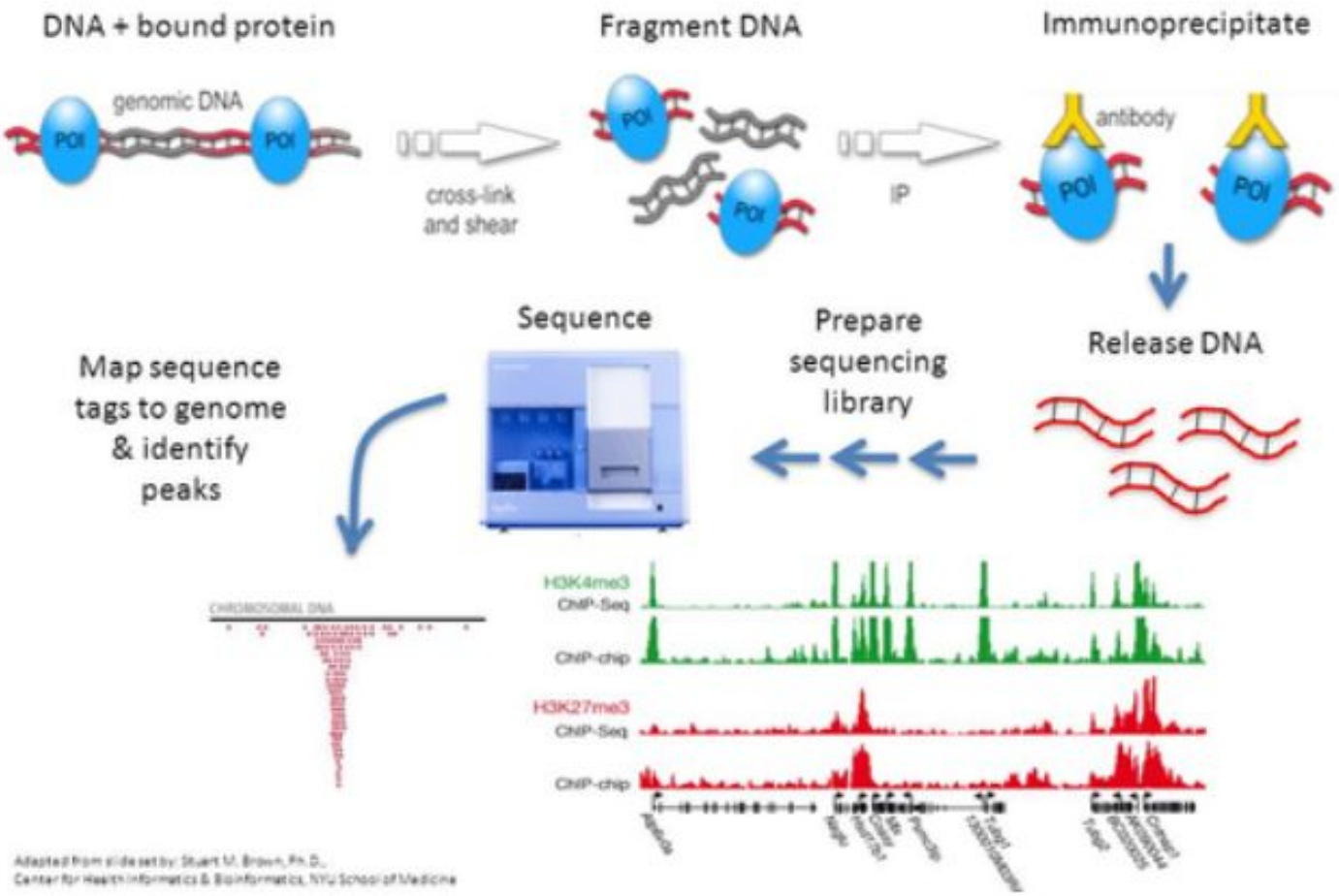


Доступность хроматина

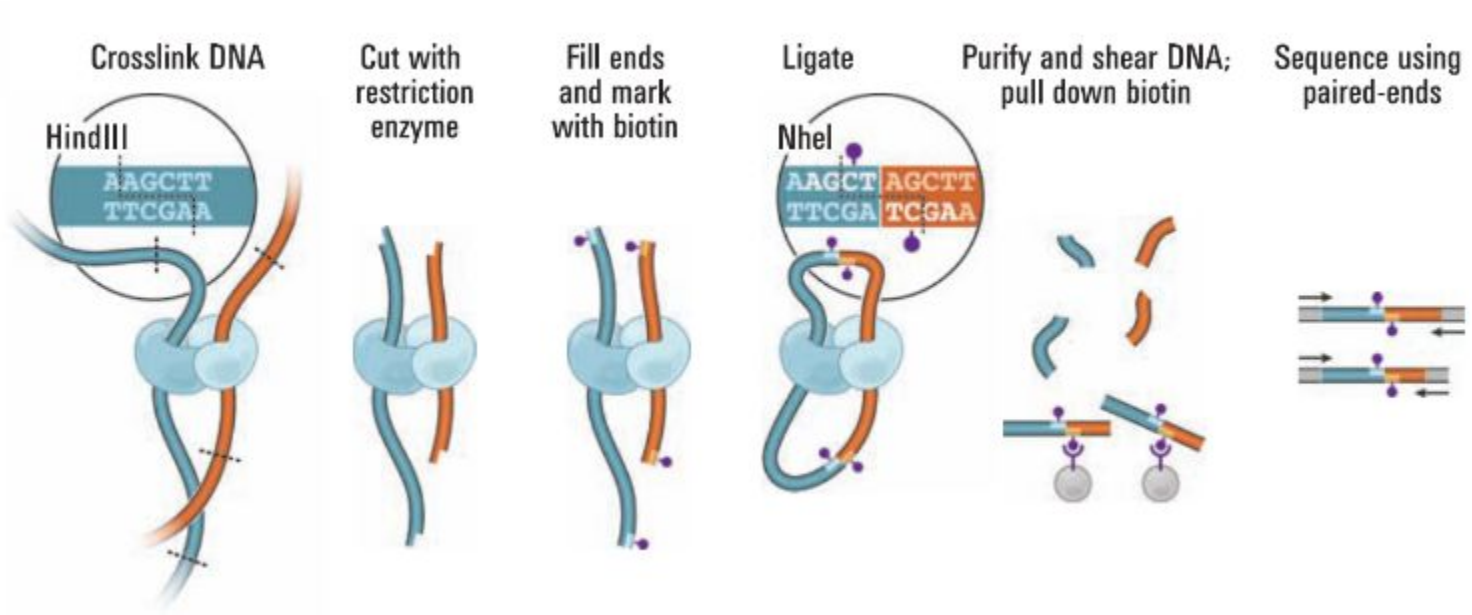


L. Minnoye, 2021

ChIP-seq - взаимодействие ДНК-белок



Hi-C - трехмерная структура хроматина



Lieberman-aiden et al., 2009

Что бывает

enseqlopedia

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

Зачем?

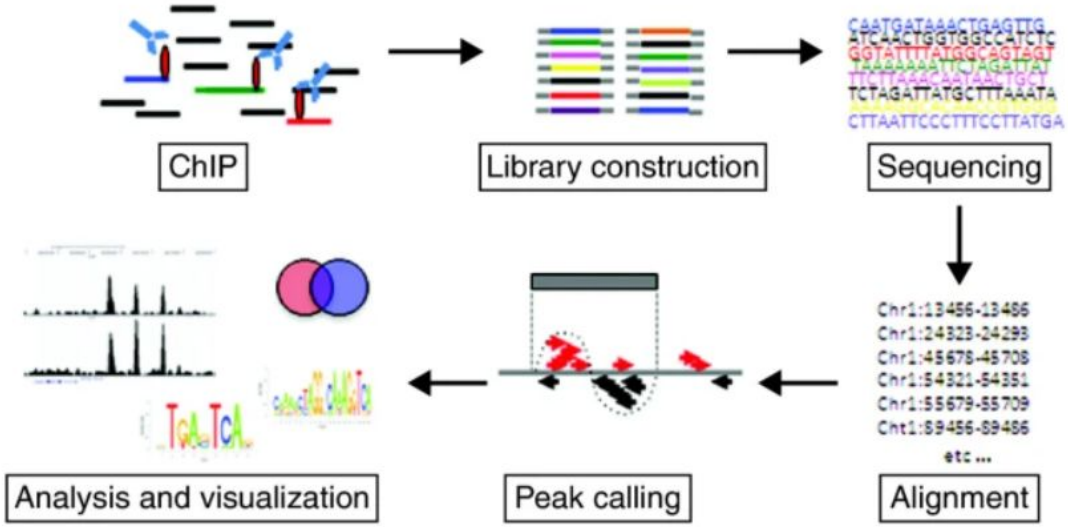
Зачем биоинформатику знать, что намешали в пробирке?

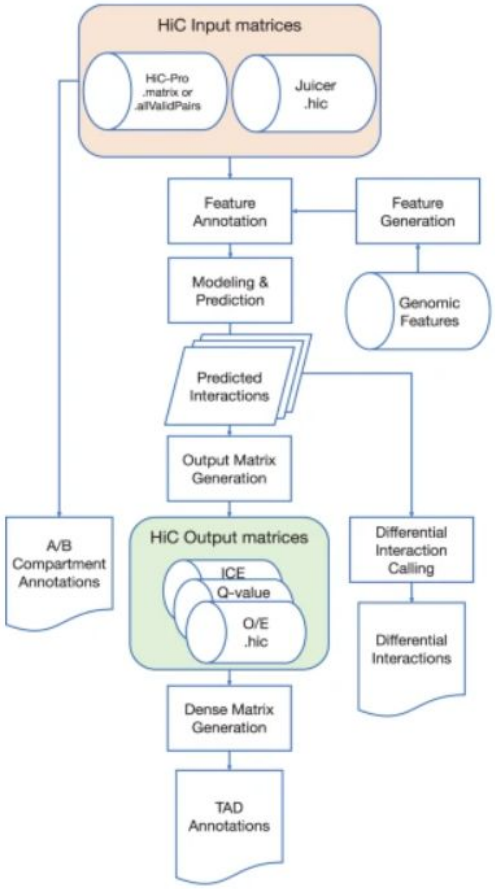
Протоколы

Экспериментальная часть

Биоинформатическая обработка результатов секвенирования

Для каждого протокола - свой анализ!





https://www.nature.com/articles/s41467-021-23749-x

Что бывает

enseqlopedia

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

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

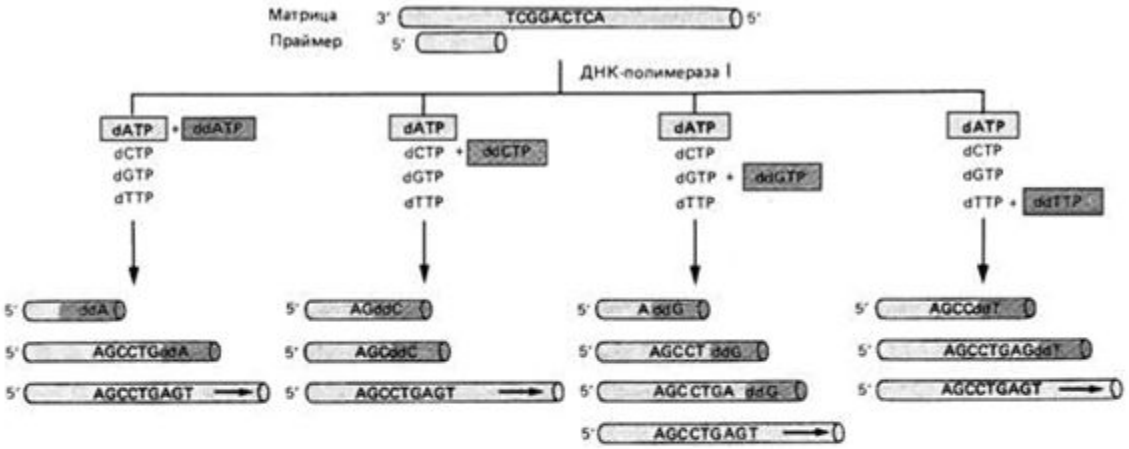
“Минусы”

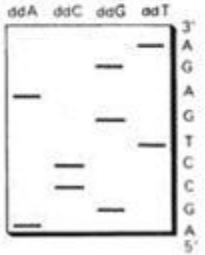
- Нет информации о некодирующих участках
- Неравномерность покрытия экзонов

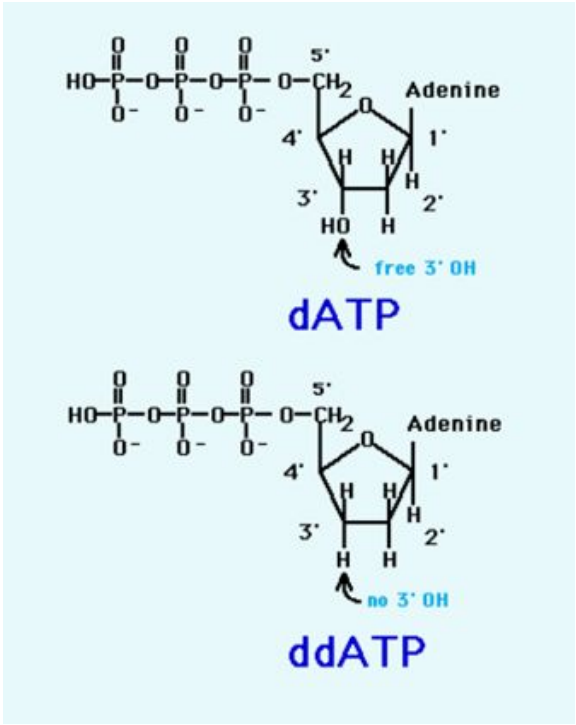
Какие бывают мутации

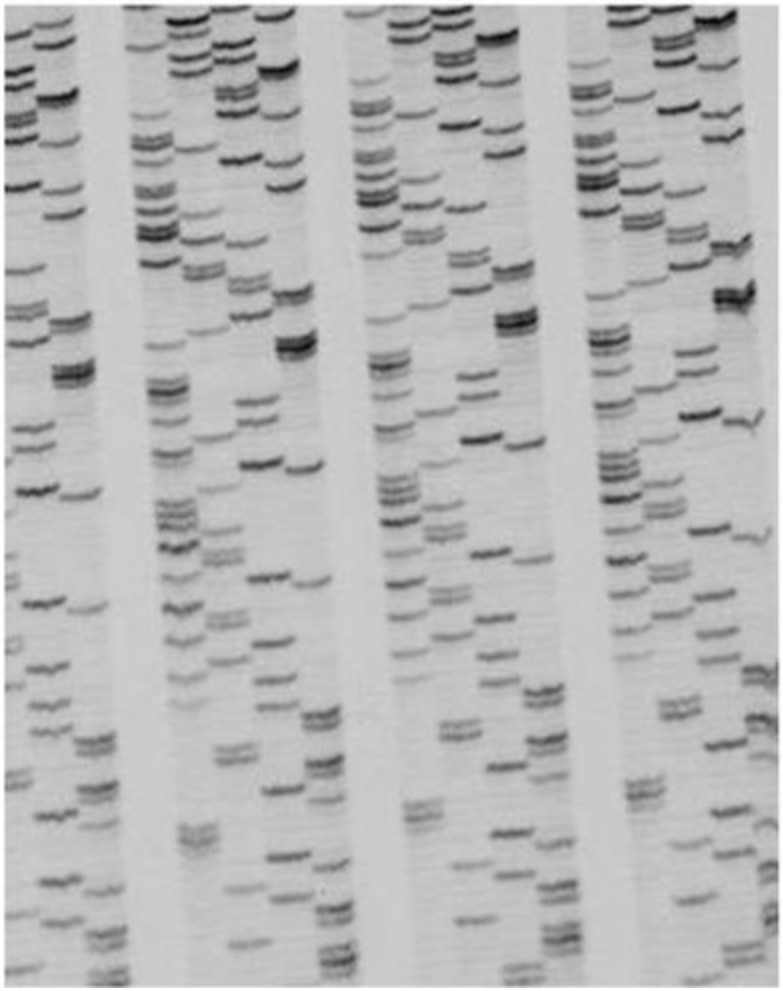
- SNV: однонуклеотидные варианты, т.е. изменение одного нуклеотида
- Короткие вставки и делеции (~ 50 п.н.)
- Структурные варианты: инверсии и транслокации; CNV
- Анеуплоидии: нульсомии, моносомии, трисомии, полисомии
- Полиплоидизация

Секвенирование ДНК - метод “терминаторов”





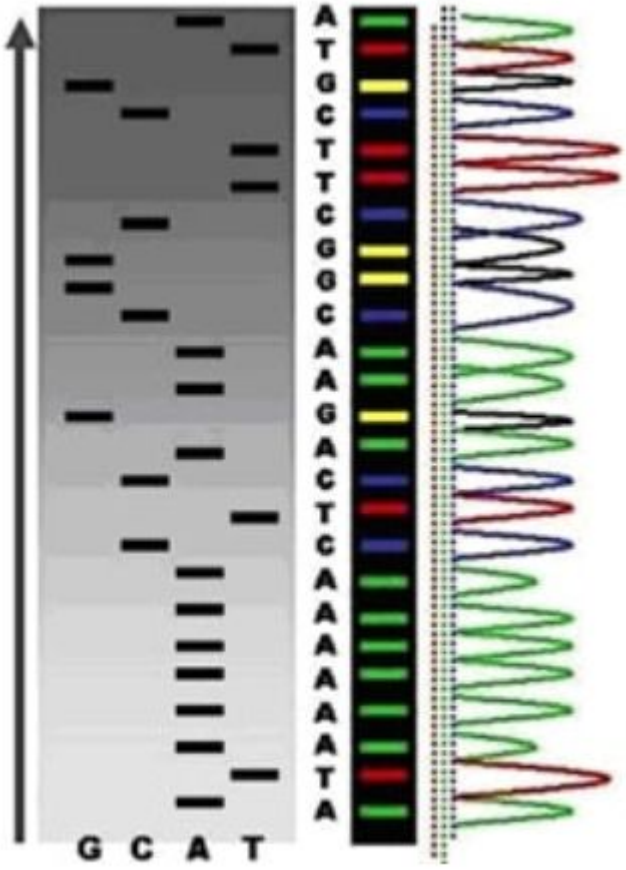




Секвенирование ДНК - метод “терминаторов”

~ 1000 п.н.

“Золотой стандарт”



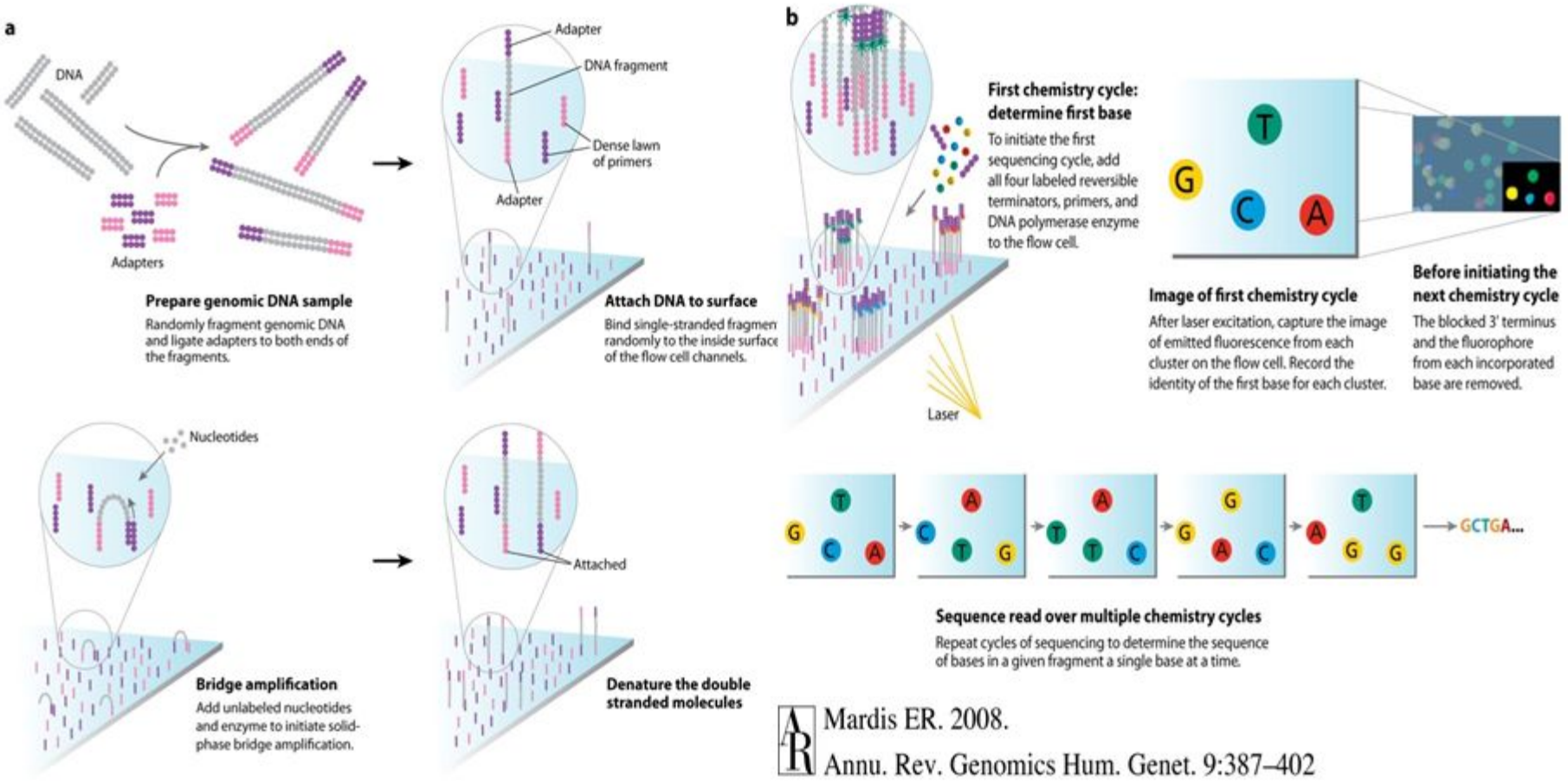
Next-generation sequencing (NGS)

“+” - 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

мультик

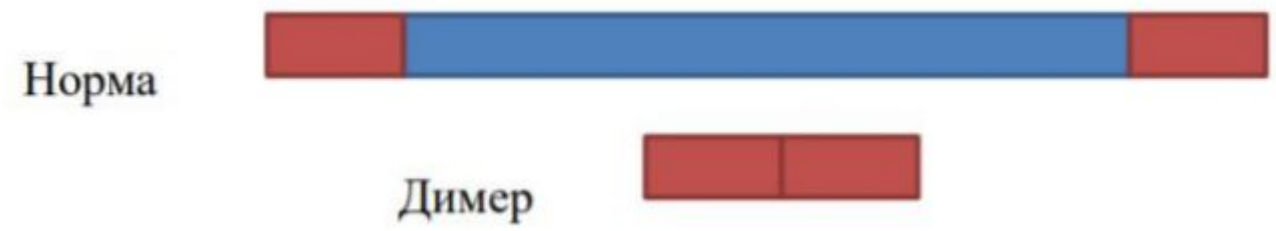
Парно-концевые и одно-концевые чтения

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

Для Illumina характерная длина чтения 75-150 нуклеотидов

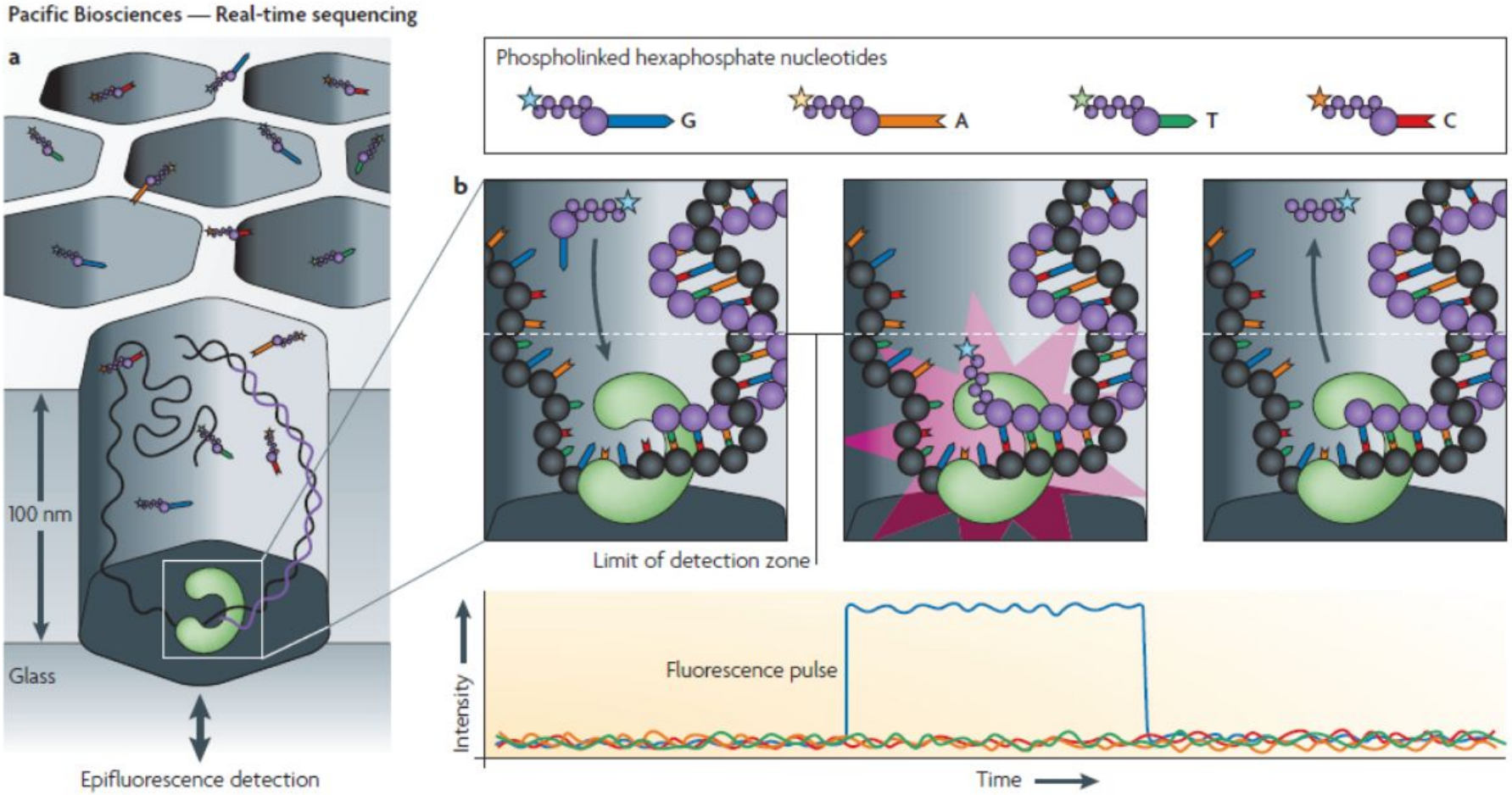
Что может пойти не так?

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



Фрагмент ДНК слишком короткий, чтение захватывает последовательность адаптера

Одномолекулярное секвенирование Pacific Biosciences



Pacific Biosciences

Pacific Biosciences

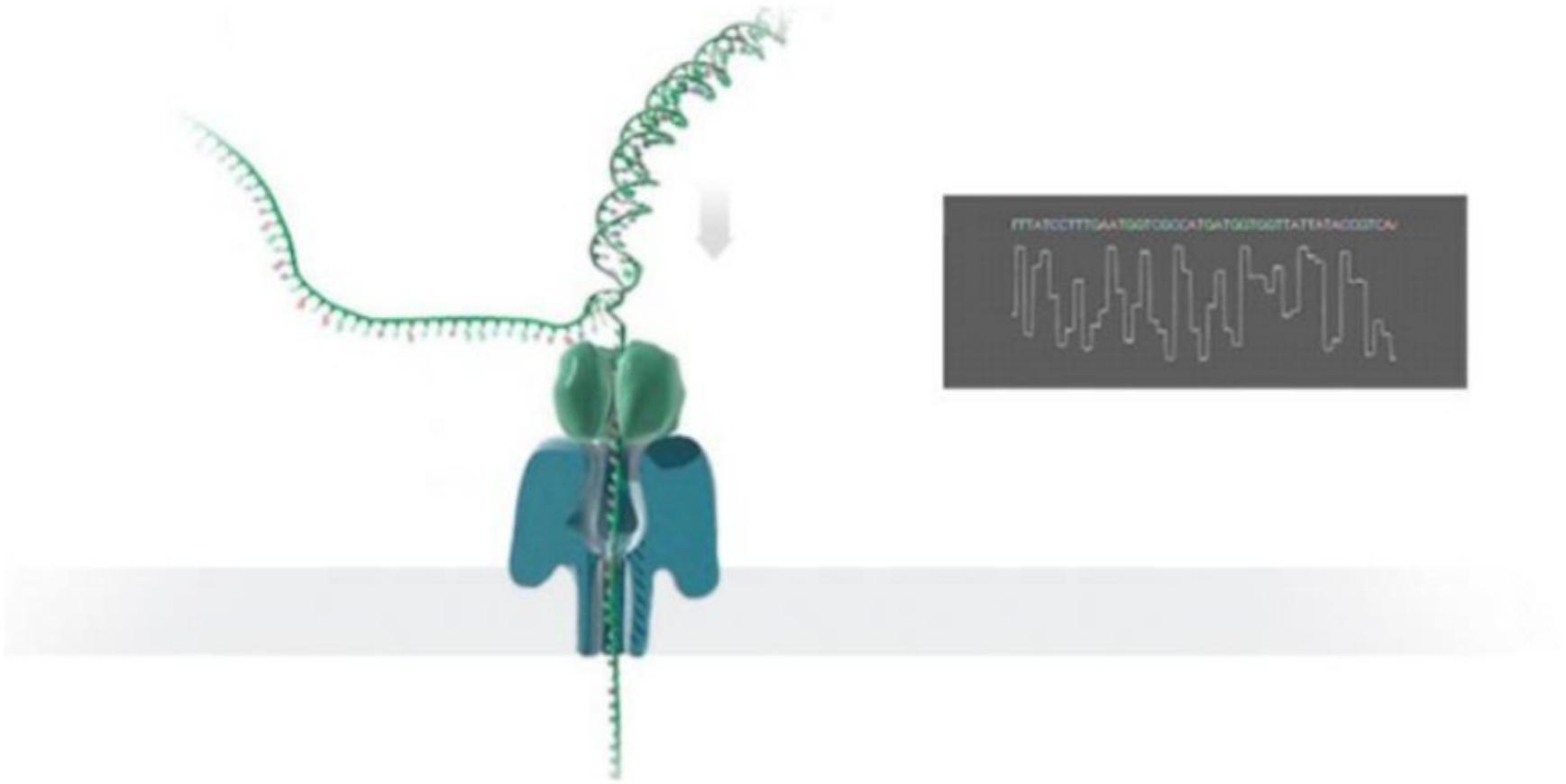
“плюсы”

- 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

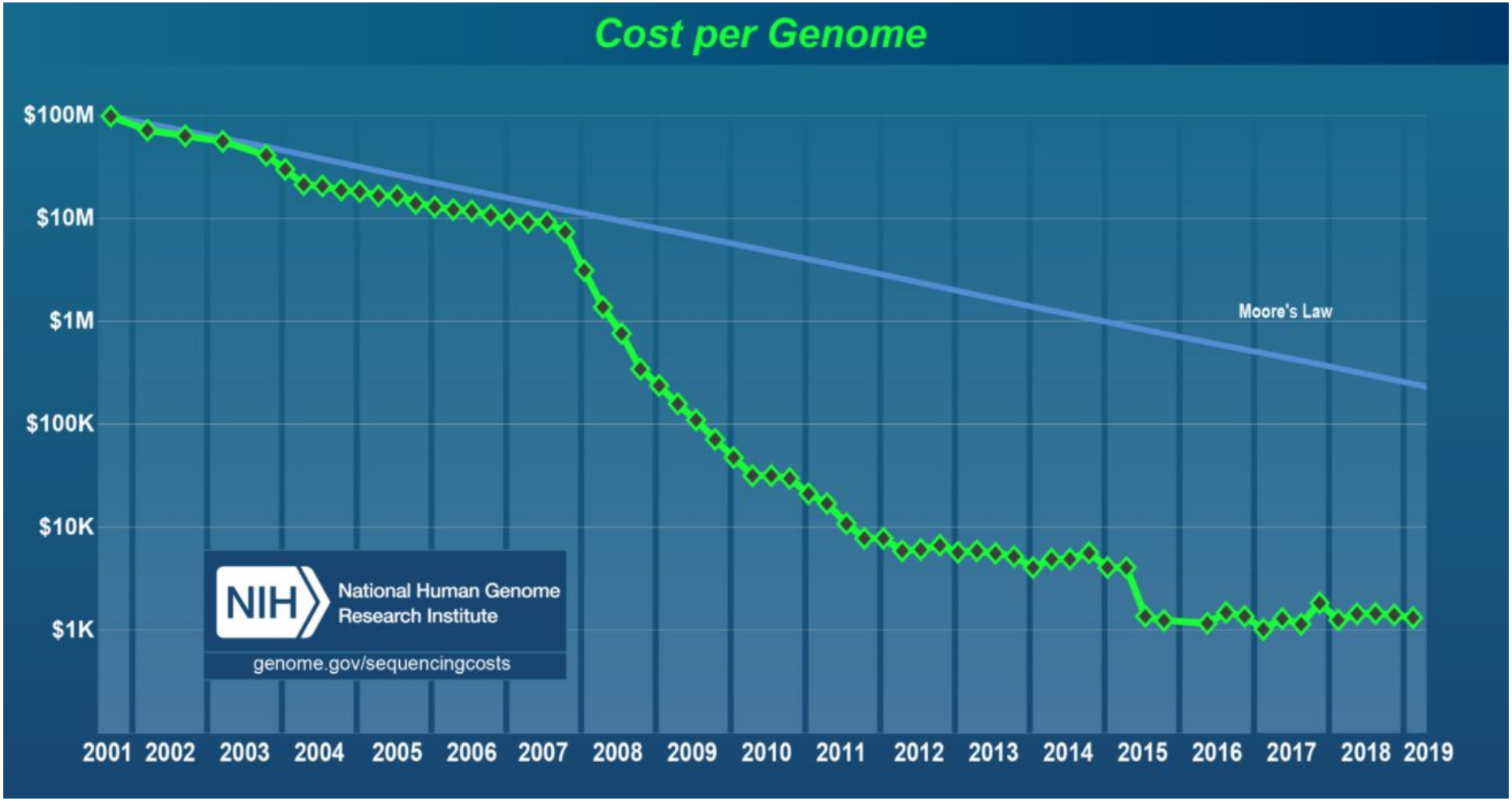
Увеличивать покрытие

Что почитать

skygen

Нанопоровое секвенирование

Обзор технологий секвенирования


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019

источник

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

ID: 11430914

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...)

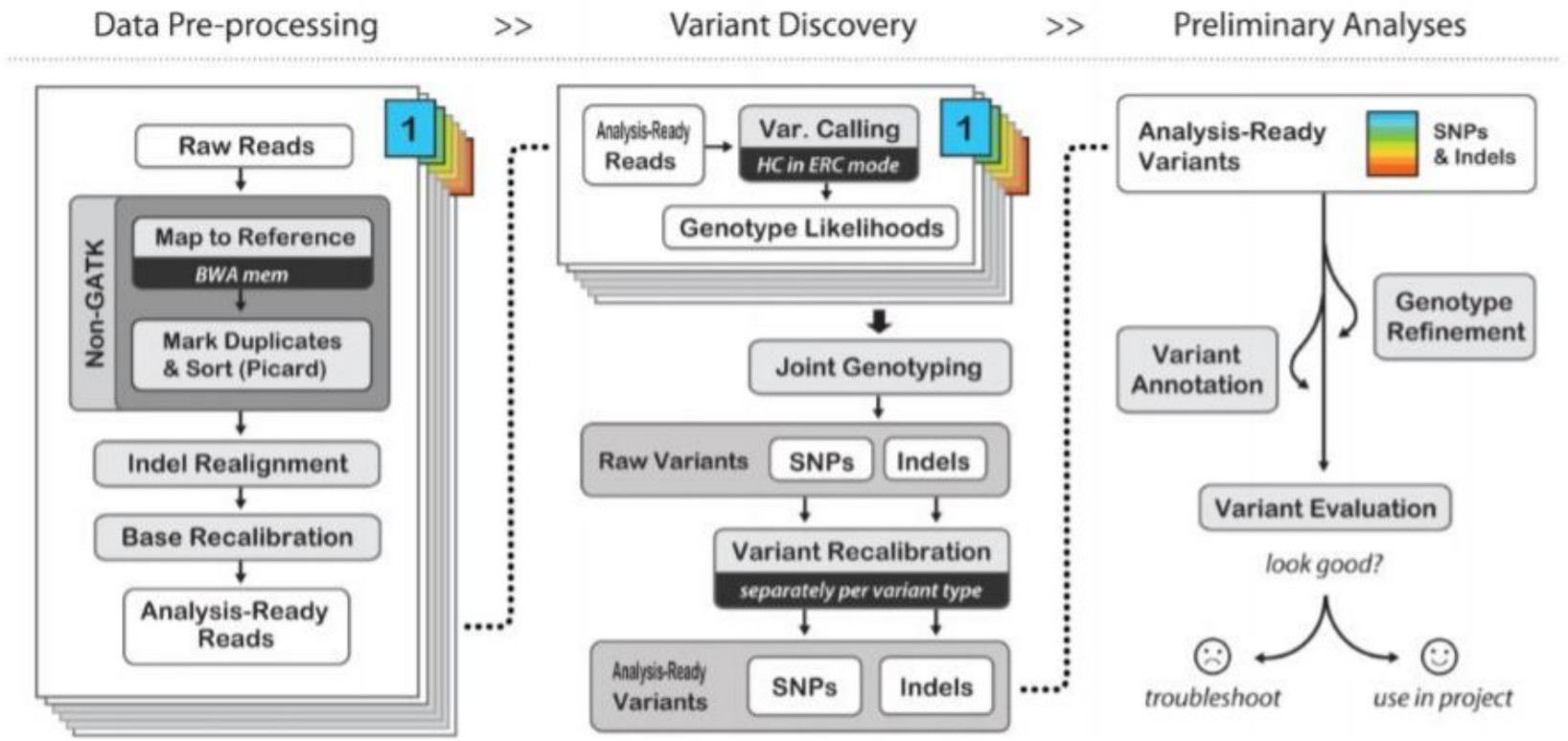
Runs: 1 run, 31.4M spots, 6.3G bases, 2.3Gb

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



При анализе данных NGS необходимо:

- Четко представлять общую задачу
- Знать биологический объект (организм, клеточная линия, ткань)
- Представлять особенности пробоподготовки и дизайн эксперимента
- Узнать на каком приборе было проведено секвенирование
- Выбрать версию референсного генома, если он используется, и оценить его качество
- При использовании дополнительных данных (например, разметка генов) зафиксировать версию файла и соотнести с версией выбранного генома
- Четко фиксировать все шаги программного конвейера, включая версии программ и пакетов, сохранять и комментировать код
- Вести лабораторный журнал (!)
- Бэкапы!
- Результаты лучше всего хранить в статьях =)

Формат fastq

```
@NB551509:7:HHJTJBGXC:1:11101:2231:1116 1:N:0:TGACCA
CATTACGGAATGTATCATCTTCTGAATGTGAACCACATCAGATGCAATACAGAGAAACACACACTCTCCAGGCAC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551509:7:HHJTJBGXC:1:11101:7127:1116 1:N:0:TGACCA
TTTTTTTCCCCCTCATTACTTTGCTTTTAGCTCACTCCTTGCAGGAATCTTTCCAGCTGCCTACCTAGCCCTTCC
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEEA
@NB551509:7:HHJTJBGXC:1:11101:2059:1116 1:N:0:TGACAA
CAAATATATTAGACCTTGTCCTGATTTGGAGTATGGCAAAAAATGTGCCATATCATATTCTTACCAAAACATTTG
+
AAAAAEEEEEEAEEEAE/EEAEEEEEEEEEAEEEEEAEEEEEEA/6A/AE/EEAEEE6EEEE/EEEEEE6E/EEE
@NB551509:7:HHJTJBGXC:1:11101:3510:1116 1:N:0:TGACCA
AATGGTTAGAGGTTCTAAATCTTGGGACACGCAGCAAGGAGAAGCAGATGCTTCTGGATTTATGGTATTATATA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEE
@NB551509:7:HHJTJBGXC:1:11101:8048:1117 1:N:0:TGACCA
CCCCCTTCTACAGCTTATAGAGTGTTGGATCCAGGACTGTCAGTCTCTGGAGATCCCAATCGATCCTTCCTTC
+
AAAAAEEEEEE/EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@NB551509:7:HHJTJBGXC:1:11101:5801:1117 1:N:0:TGACCA
CAAAACCTATAACATATTGTATACATATATATAATATATAAACACACATACACAATATAGACTTATCTTGCTCTT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
```

Fastq формат

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

Последовательность

Качество

```
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126
 0........................26...31.......40
                          -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
 0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

wiki

Качество чтений

P - вероятность ошибки

Q - параметр качества (Phred Quality Score)

Значения Q: 1-40

Q > 20 считается хорошим качеством

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

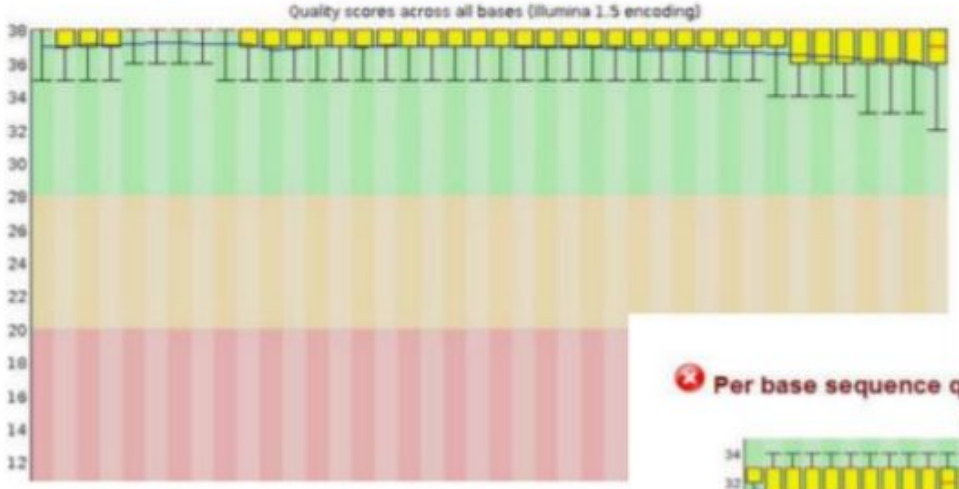
Вероятность ошибки	Q
0.001 (*точность 99,9%*)	30
0.01 (*точность 99%*)	20
0.1 (*точность 90%*)	10

Пересчет качества в вероятность ошибки

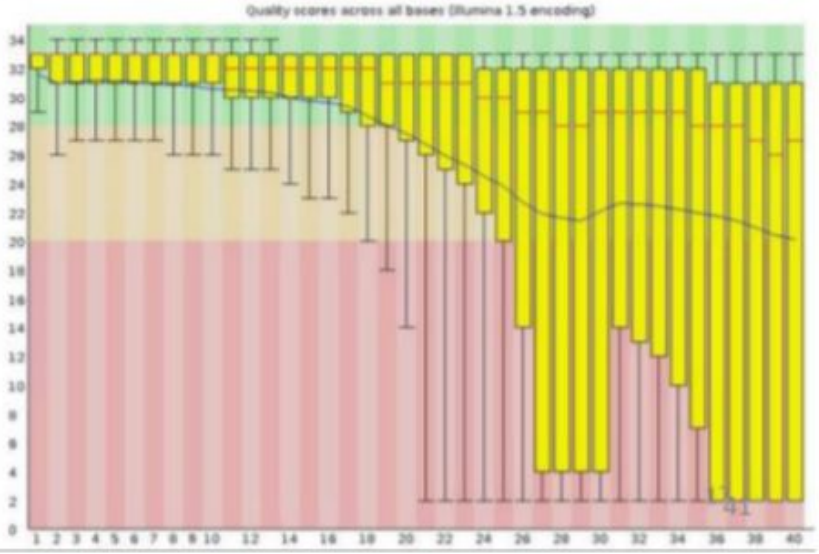
Phred Quality Score	Символ	Вероятность ошибки	Точность
10	+	1/10	90%
20	5	1/100	99%
30	?	1/1000	99,9%
40	I	1/10 000	99,99%
50	S	1/100 000	99,999%
60	]	1/1 000,000	99,9999%

fastQC

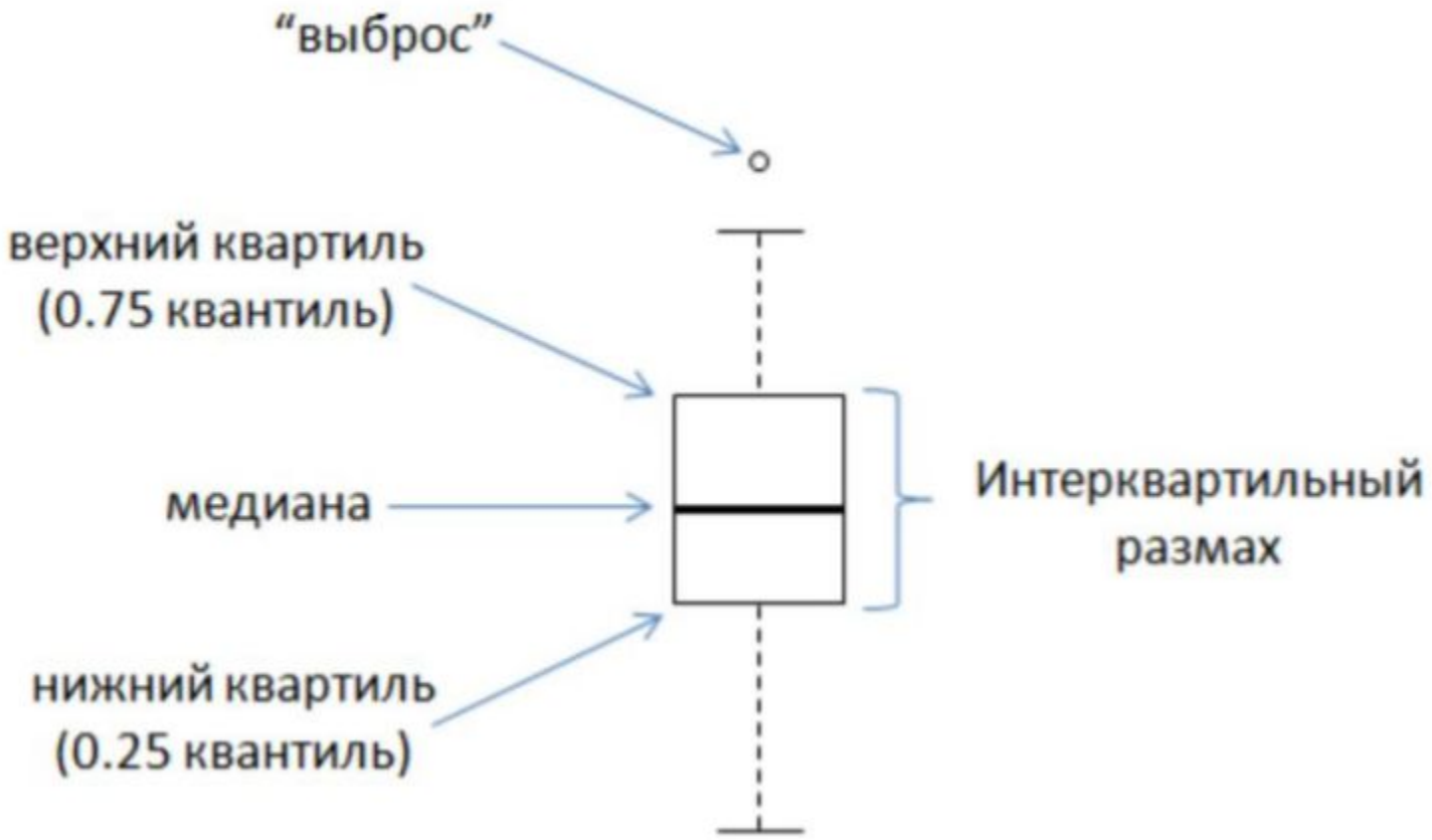
Per base sequence quality



Per base sequence quality



“Ящик с усами” / диаграмма размахов / boxplot



fastQC

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

Программа fastQC установлена на kodomo

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

Что делать?

Нужно удалить «плохие» фрагменты чтений:

- Адаптеры

- Нуклеотиды с неудовлетворительным качеством (< 20)

Trimmomatic

http://www.usadellab.org/cms/?page=trimmomatic

В результате получаем только те чтения, качество которых нас устраивает
С ними можно смело работать дальше!

Что делать дальше?

Дано:

- «очищенные» чтения хорошего качества (fastq)

- Последовательность референсного генома (fasta)

Задача:

Каждому чтению найти свое место на геноме - картирование

Что делать дальше?

Программы:

- bowtie

- bwa

- hisat2

Есть много других!

Шаг 0. Подготовка референса: индексирование

Для каждой программы свой индекс!

Шаг следующий – картирование чтений на референс

Получаем .sam или .bam