Введение в анализ данных NGS

Анастасия Жарикова

10 ноября 2020

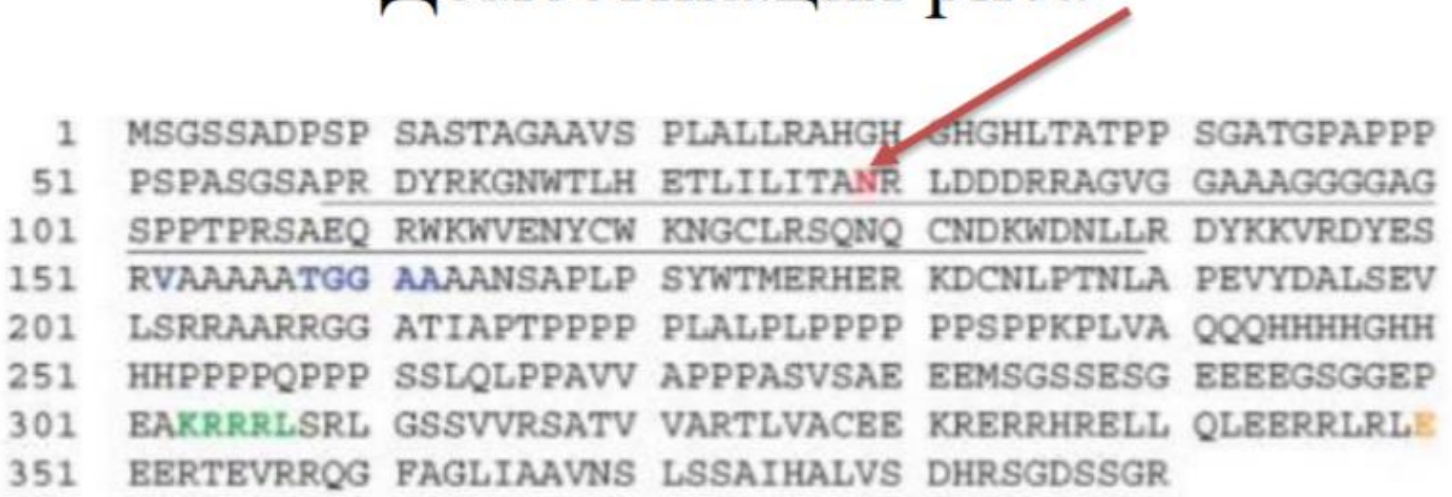
azharikova89@gmail.com

Для чего нужно секвенирование

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

Эволюция

Доместикация риса



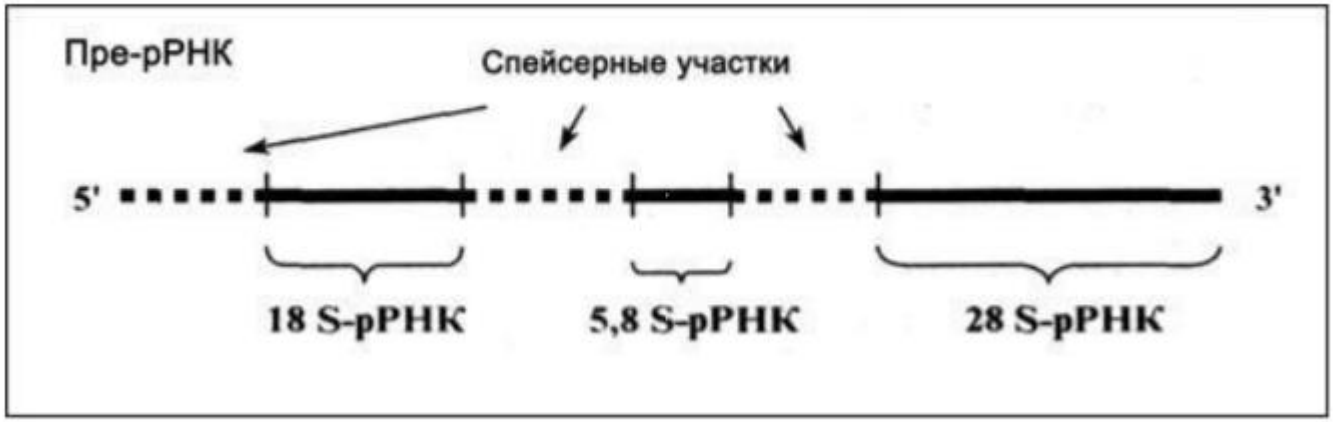
sh4

Li et al., Science, 2006

Дикий рис	Культурный рис
AAG	AAT
Лизин	Аспарагин

Филогения

Транскрибируемые спейсеры



Спейсерные последовательности наиболее вариабельные с точки зрения эволюционной консервативности. Секвенирование и анализ транскрибируемый спейсеров используется для изучения видового разнообразия и классификации близкородственных организмов

Популяционные и клинические исследования

- 1000 геномов
  - Частоты snp в популяциях
- GWAS
  - Поиск полиморфизмов, ассоциированных с болезнями:

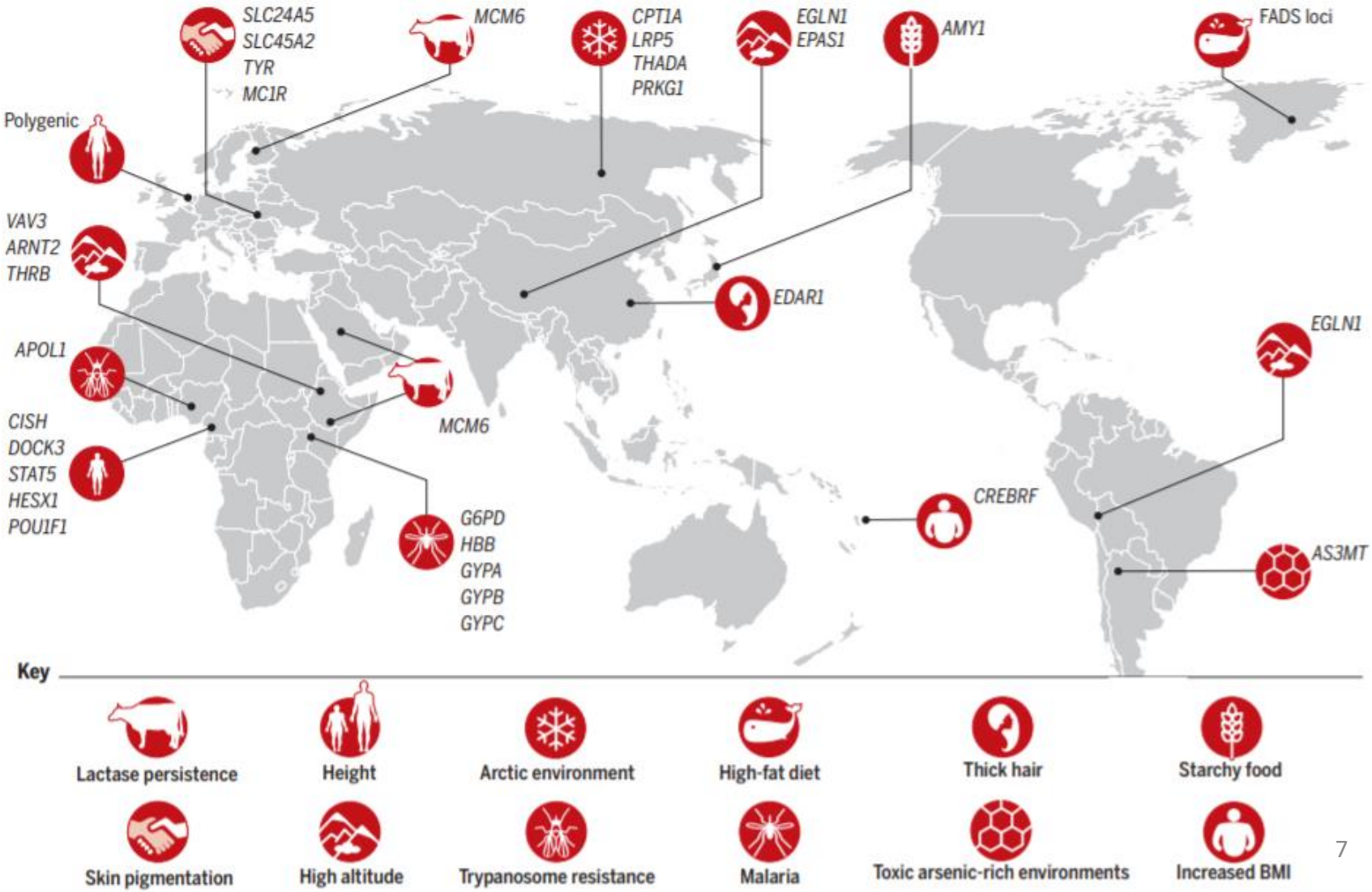
    - моногенные (муковисцидоз, ген CFTR)

    - полигенные (ишемическая болезнь сердца, шизофрения)
- Фармакогенетика
  - Индивидуальное лечение

    - Варфарин – предотвращает образование тромбов. Генетические факторы определяют до 53-54 % вариабельности дозы. Гены CYP2C9, CYP4F2, VKORC1.

Going global by adapting local: A review of recent human adaptation

Shaohua Fan,[1*] Matthew E. B. Hansen,[1*] Yancy Lo,[1,2*] Sarah A. Tishkoff[1,3†]



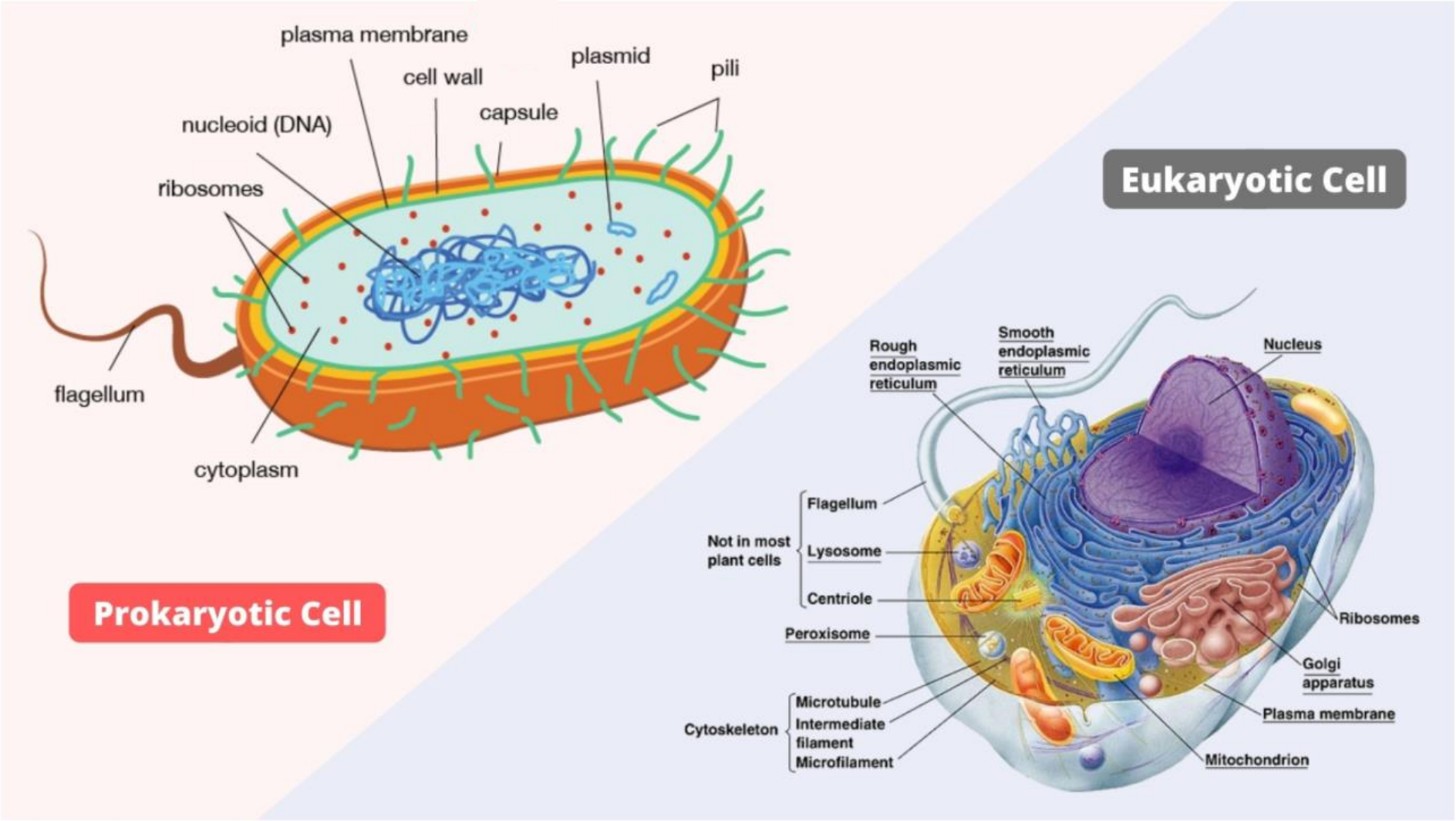
Что бывает

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

Что бывает

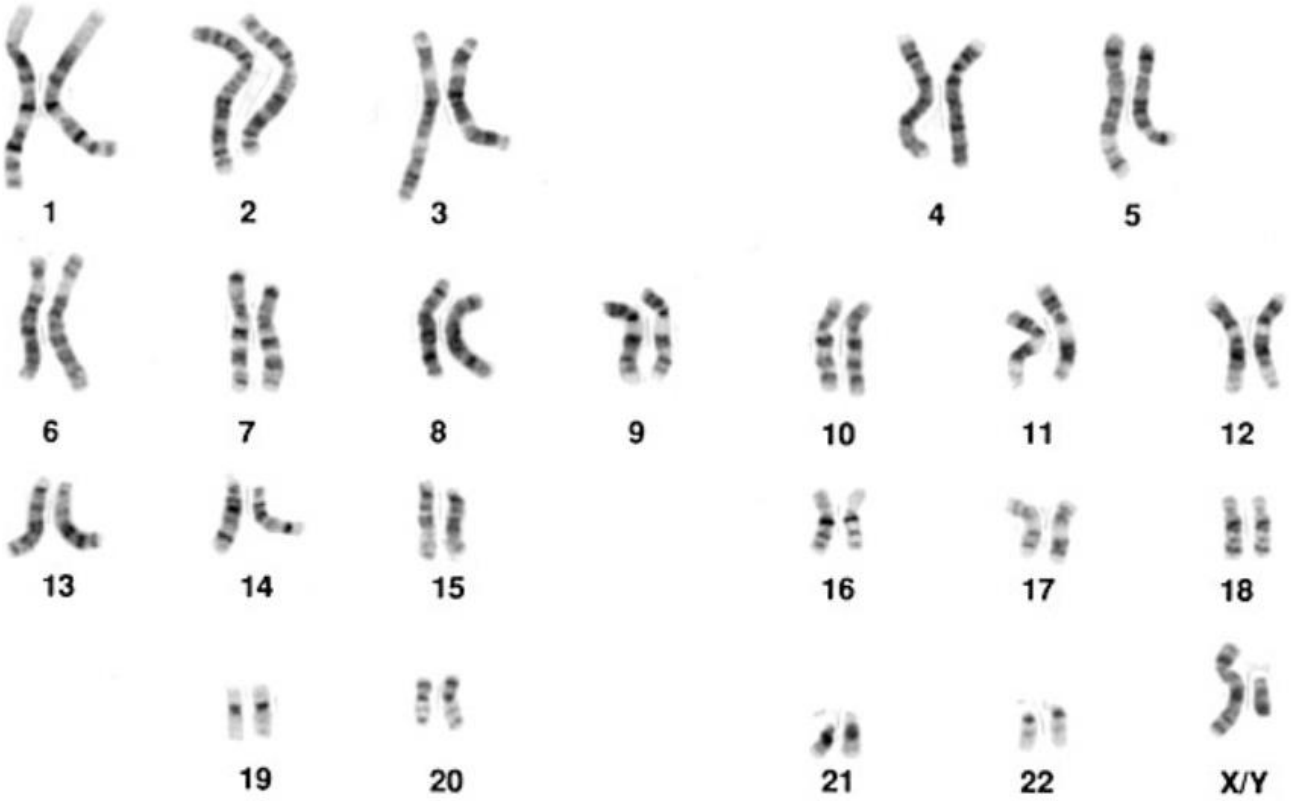
- **DNA-seq**
- RNA-seq
- Chip-seq
- HiC
- ATAC-seq
- DNase-seq
- eClip
- GRO-seq
- CRISPR-seq
- Ribo-seq
- …..

Где ДНК?



https://microbiologynote.com/difference-between-prokaryotic-and-eukaryotic-cells/

У человека 23 пары хромосом
Много или мало?



https://ru.wikipedia.org/wiki/Геном_человека

Число хромосом у разных видов

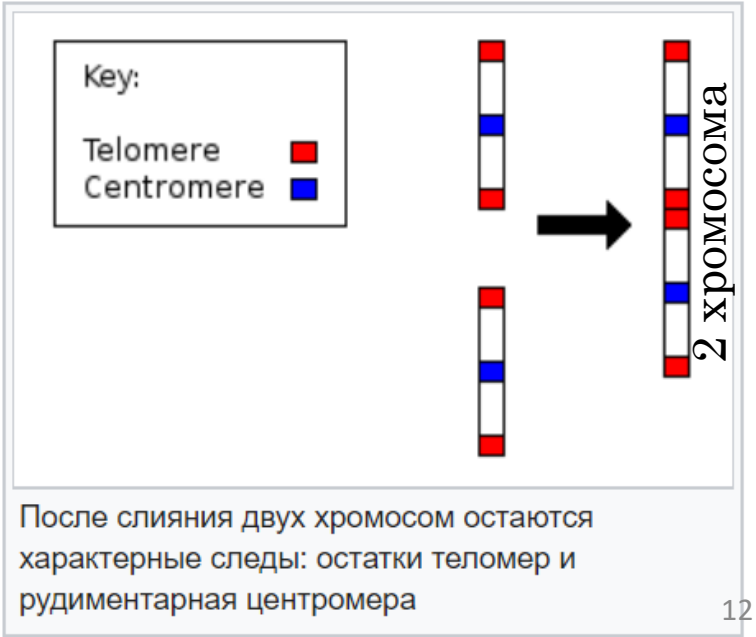
Гиббоны - 44

48

46

Макака - 42

Капуцин - 54



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

https://ru.wikipedia.org/wiki/2-%D1%8F_%D1%85%D1%80%D0%BE%D0%BC%D0%BE%D1%81%D0%BE%D0%BC%D0%B0_%D1%87%D0%B5%D0%BB%D0%BE%D0%B2%D0%B5%D0%BA%D0%B0

Число хромосом у разных видов

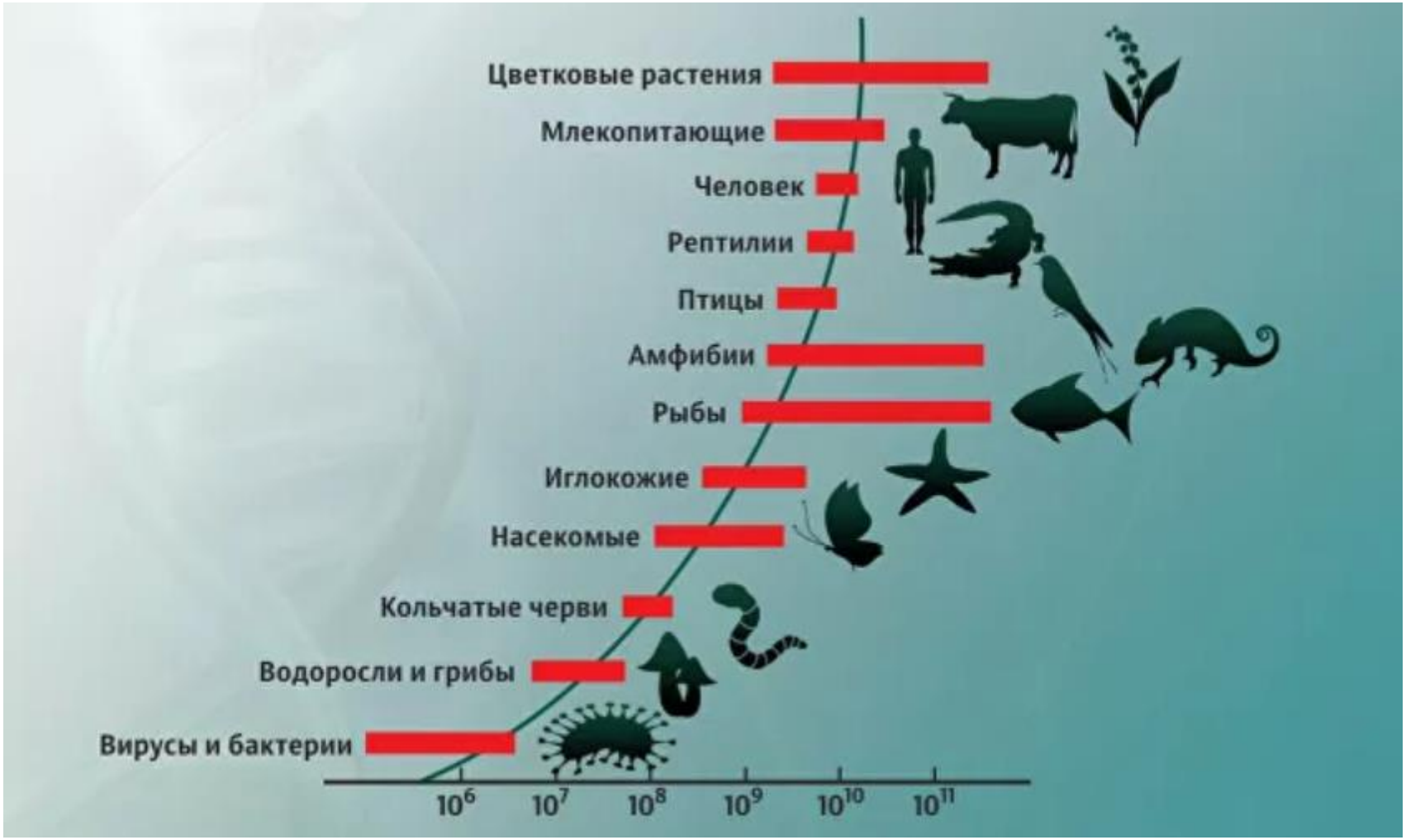
Муравей (Myrmecia pilosula) – 2
Плодовая мушка - 8
Арабидопсис – 10
Голубь – 16
Кошка – 38
Лиса - 34
Мышь - 40
Собака – 78
Утка – 80
Сазан - 104
Корова – 120
Рак (Cambarus clarkii) – 200
Хвощ – 216
Краб - 254
Бабочка – 380

https://pikabu.ru/story/chislo_khromosom_u_raznyikh_vidov_5141209

Число хромосом у разных видов

Подроды (в скобках) и виды муравьев	Число хромосом	
	2n (самцы)	2n (самки)
Род Formica		
F. (Formica) aquilonia - северный лесной муравей	26	52
F. (Serviformica) cinerea - серый лесной муравей	27	54
F. (Serviformica) cunicularia - прыткий муравей	27	54
F. (Coptoformica) exsecta - тонкоголовый муравей	26	52
F. (Serviformica) fusca - бурый лесной муравей	27	54
F. (Formica) lugubris - волосистый лесной муравей	26	52
F. (Serviformica) picea - черный блестящий муравей	26	52
F. (Formica) polyctena - малый лесной муравей	26	52
F. (Formica) pratensis - луговой муравей	26	52
F. (Formica) rufa - рыжий лесной муравей	26	52
F. (Serviformica) rufibarbis - краснощекий муравей	27	54
F. (Raptiformica) sanguinea - кровавый муравей-рабовладелец	26	52
F. (Formica) truncorum - красноголовый муравей	26	52
F. (Serviformica) uralensis - черноголовый муравей	26	52
Род Lasius		
L. (Lasius) alienus - бледноногий муравей	15	30
L. (Cautolasius) flavus - желтый земляной муравей	15	30
L. (Dendrolasius) fuliginosus - пахучий муравей-древоточец	14	28
L. (Lasius) niger - черный садовый муравей	15	30

Размер генома у разных видов



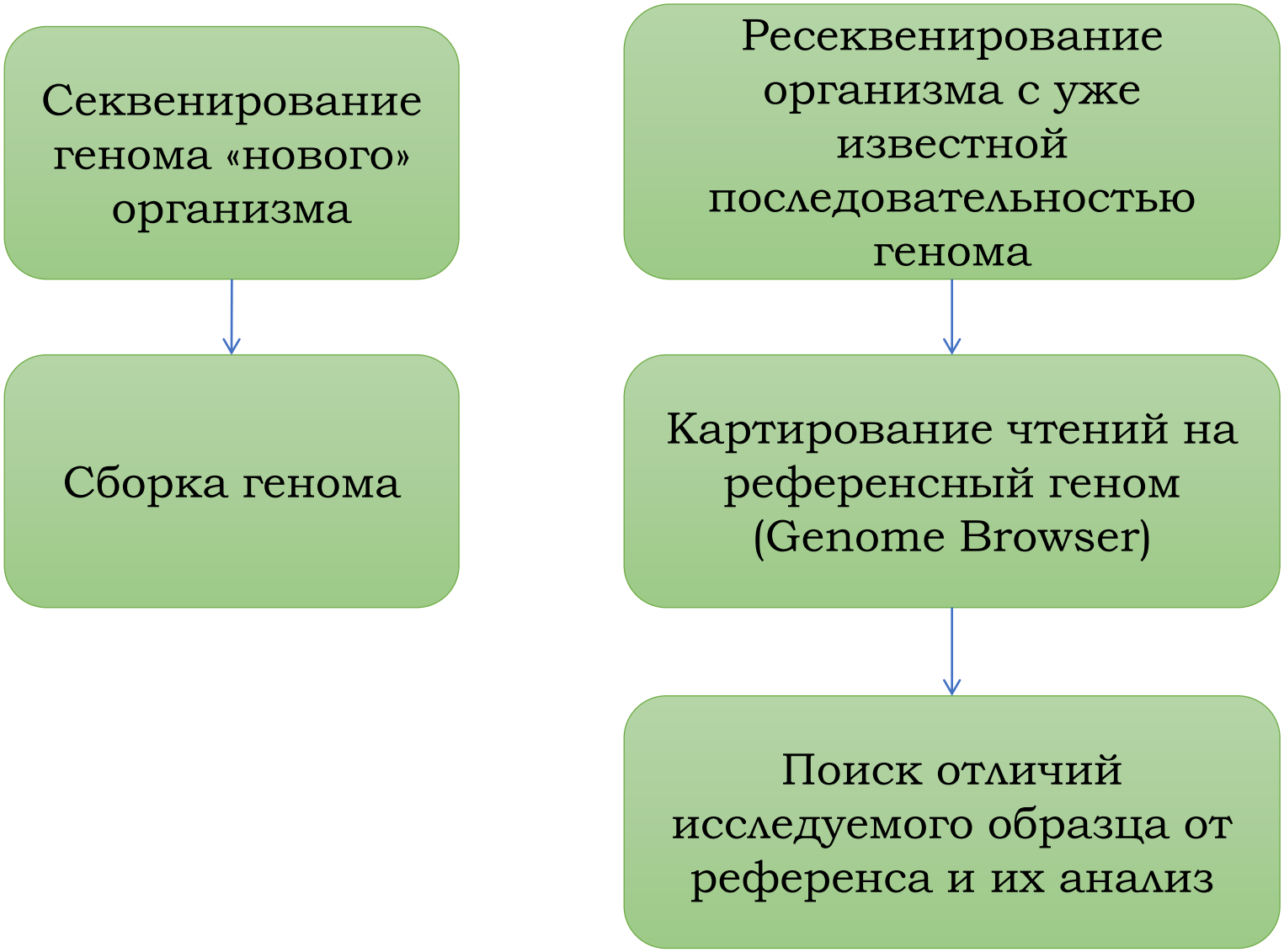
https://ria.ru/20170704/1497743424.html

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?

Секвенирование ДНК бывает …



Возможности ресеквенирования

Можно ресеквенировать:

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

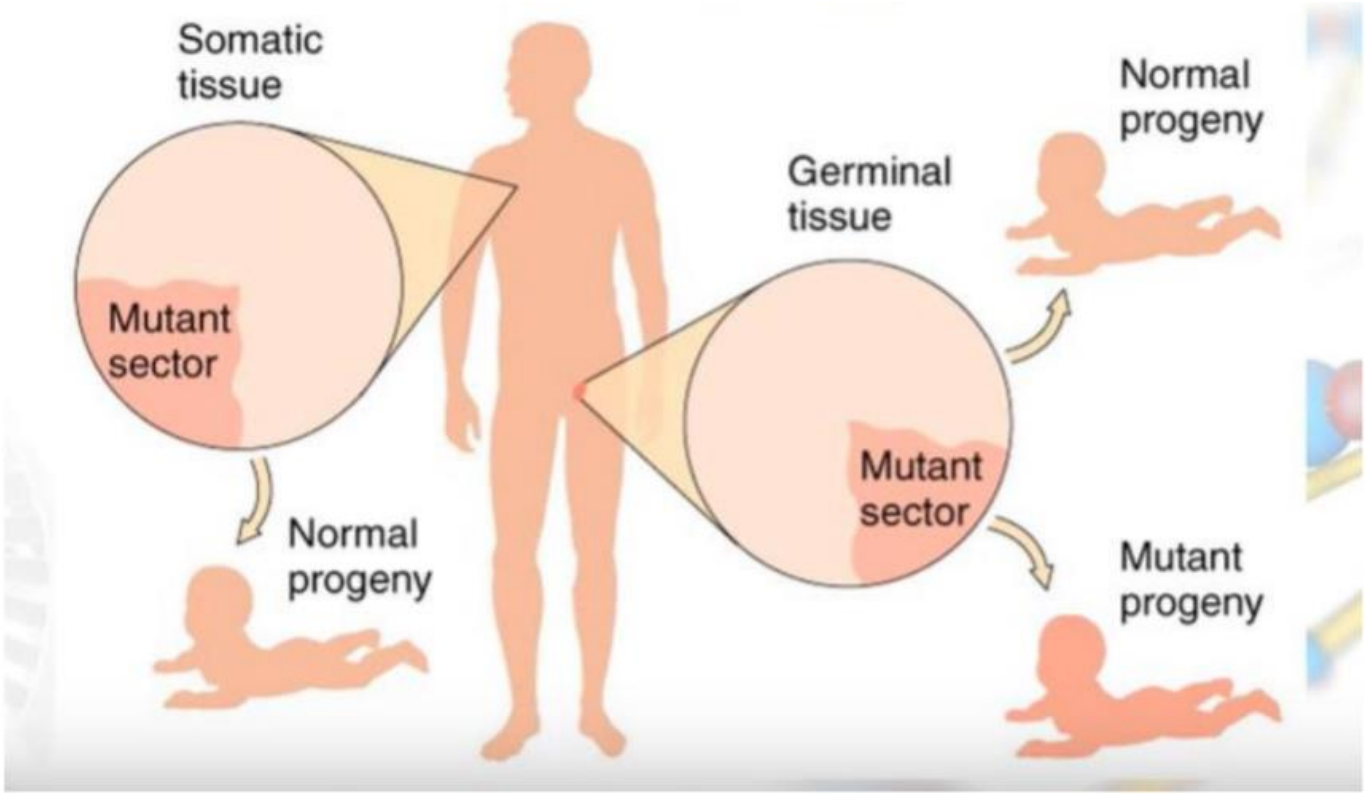
!!!Выбор в зависимости от бюджета и целей исследования!!!

Экзомное ресеквенирование

• **«Плюсы»**
- Небольшой объем кодирующих последовательностей – ниже цена
- Кодирующие последовательности лучше изучены
- Большое число болезнетворных мутаций находится в кодирующей последовательности (особенно менделевские заболевания)

• **«Минусы»**
- Нет информации о некодирующих участках
- Неравномерность покрытия экзонов

Соматические и герминальные мутации



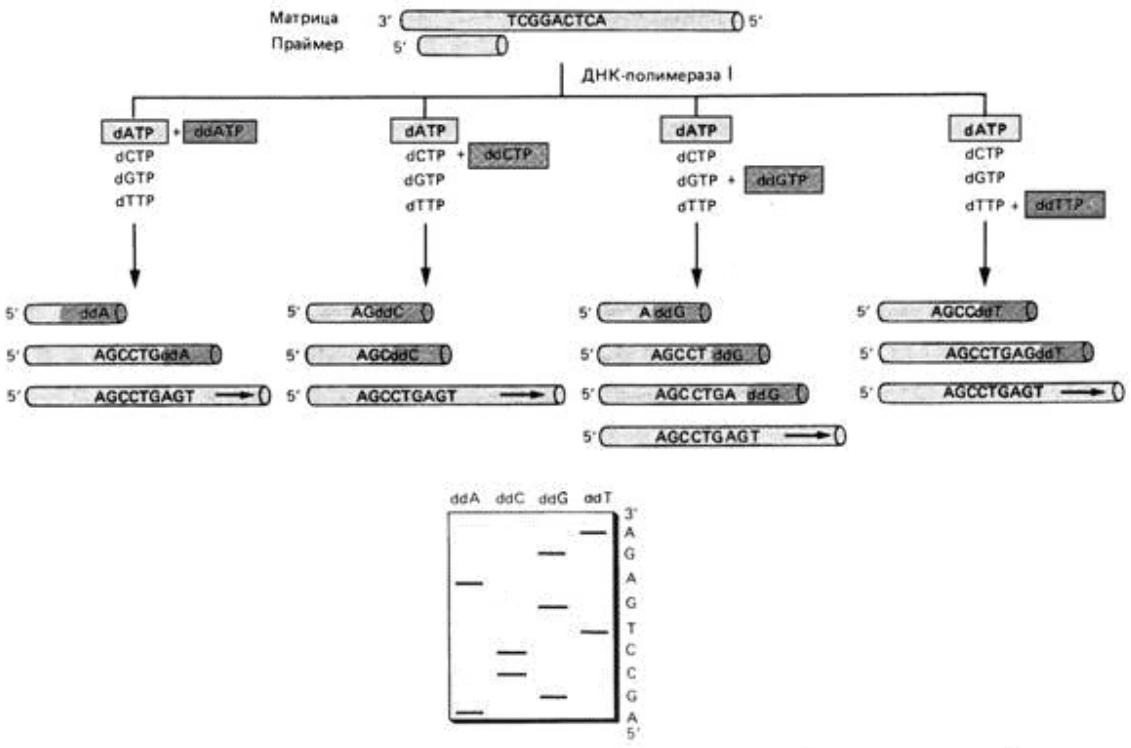
https://www.youtube.com/watch?v=3MiHSeH6yLg

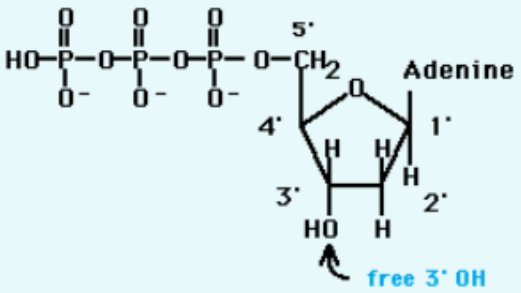
Какие бывают мутации

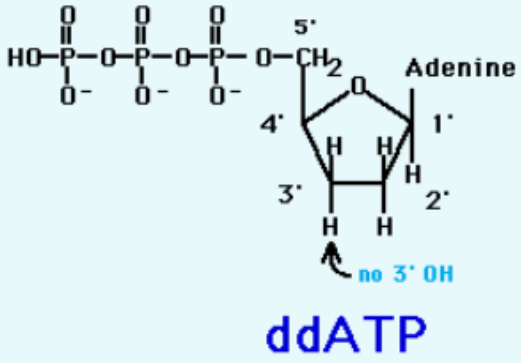
SNV – однонуклеотидные варианты, т.е. изменение одного нуклеотида
Короткие вставки и делеции (~ до 50 п.н.)
Структурные варианты: инверсии и транслокации; CNV
Анеуплоидии: нульсомии, моносомии, трисомии, полисомии
Полиплоидизация

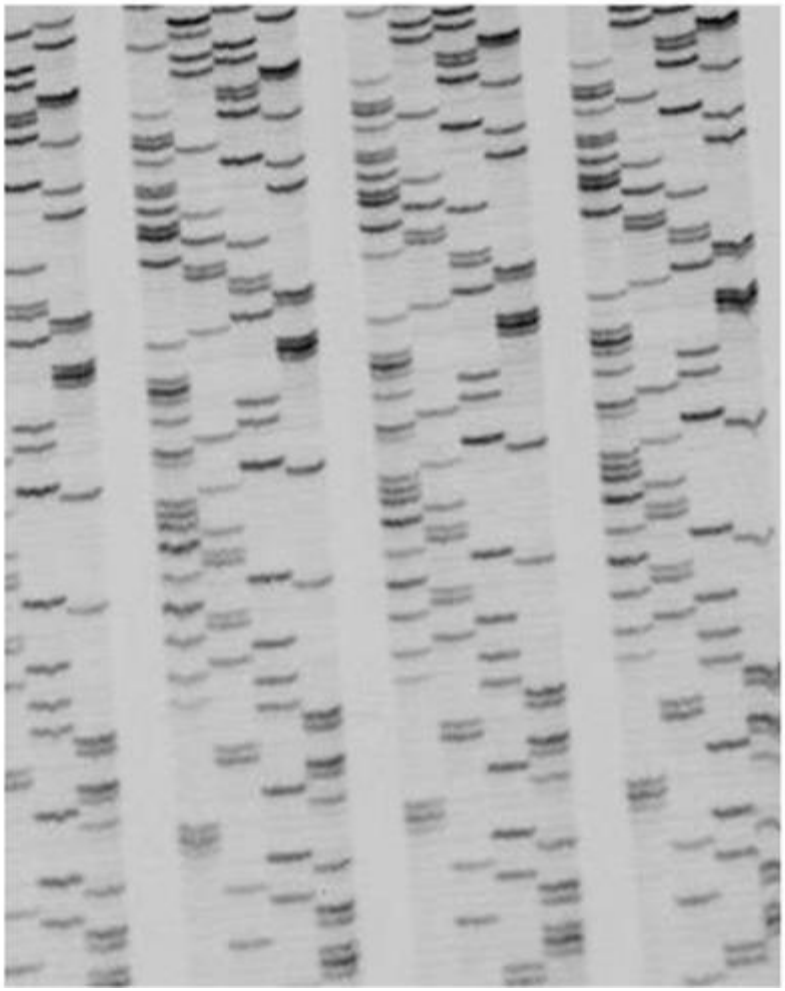
Секвенирование ДНК

Метод «терминаторов»



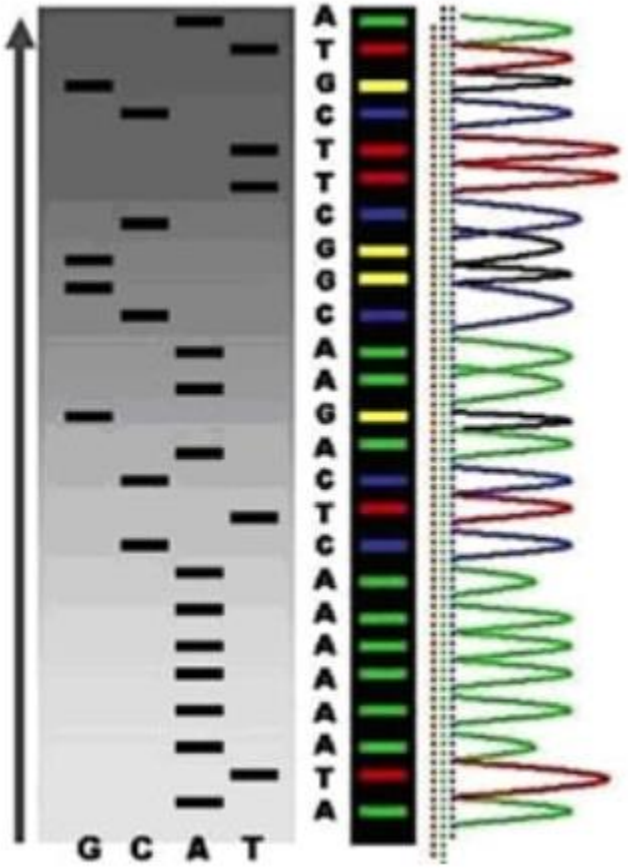






Секвенирование ДНК

Метод «терминаторов»



~ 1000 п.н.

«Золотой стандарт»

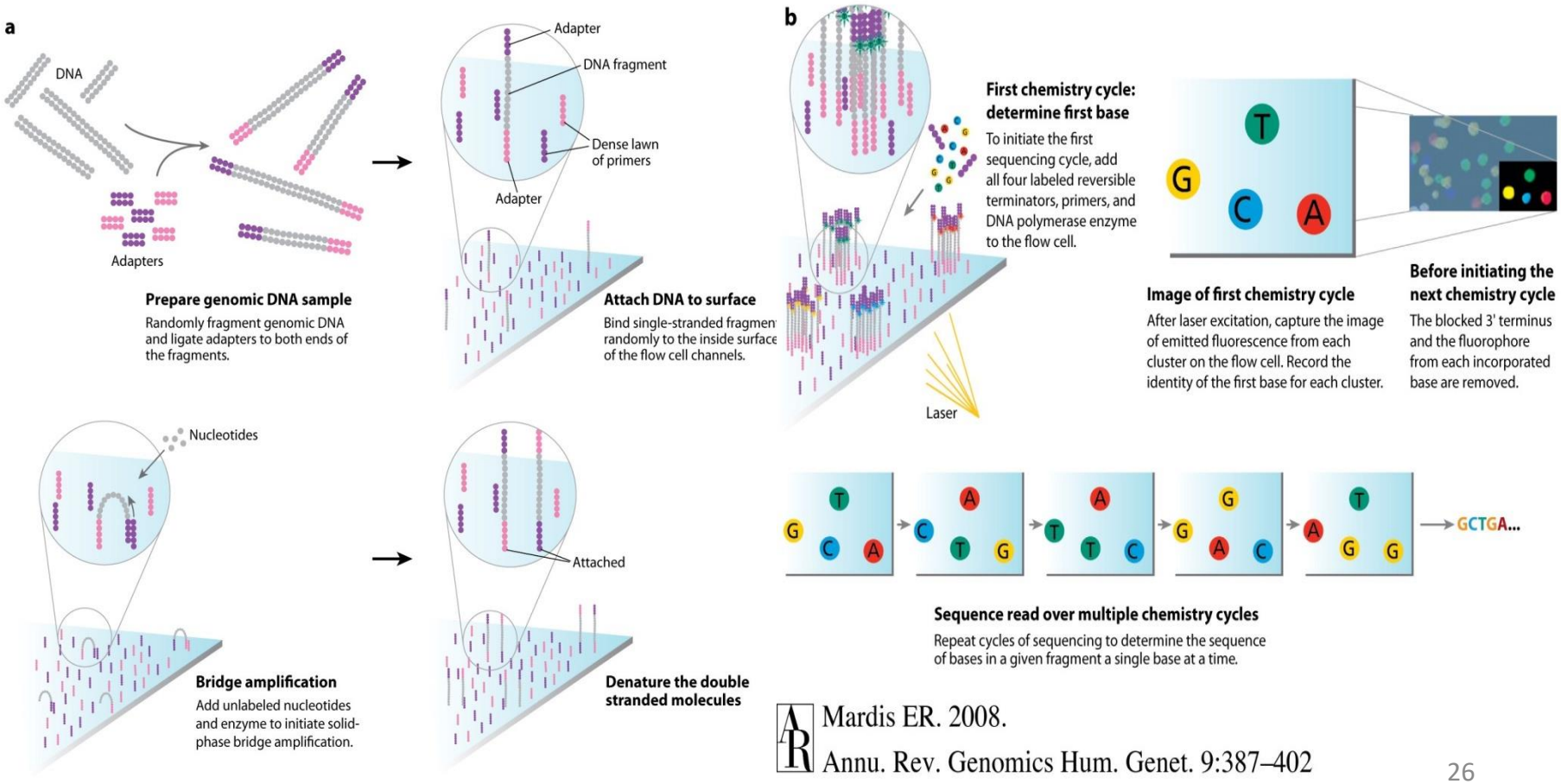
Секвенирование второго поколения Next-generation sequencing (NGS)

“+” : одновременно идет сиквенс большого количества разных фрагментов

“-” : прочтения длиной 75-150 нуклеотидов

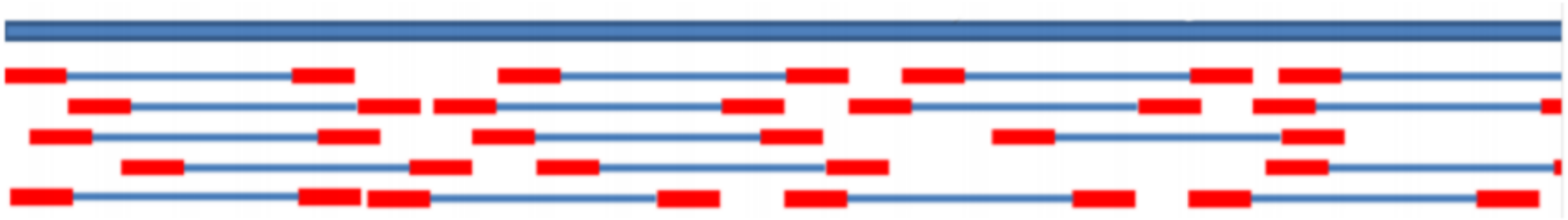
Секвенирование второго поколения Next-generation sequencing (NGS)

Illumina (но есть и другие приборы)



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

Парные и одноконцевые чтения



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

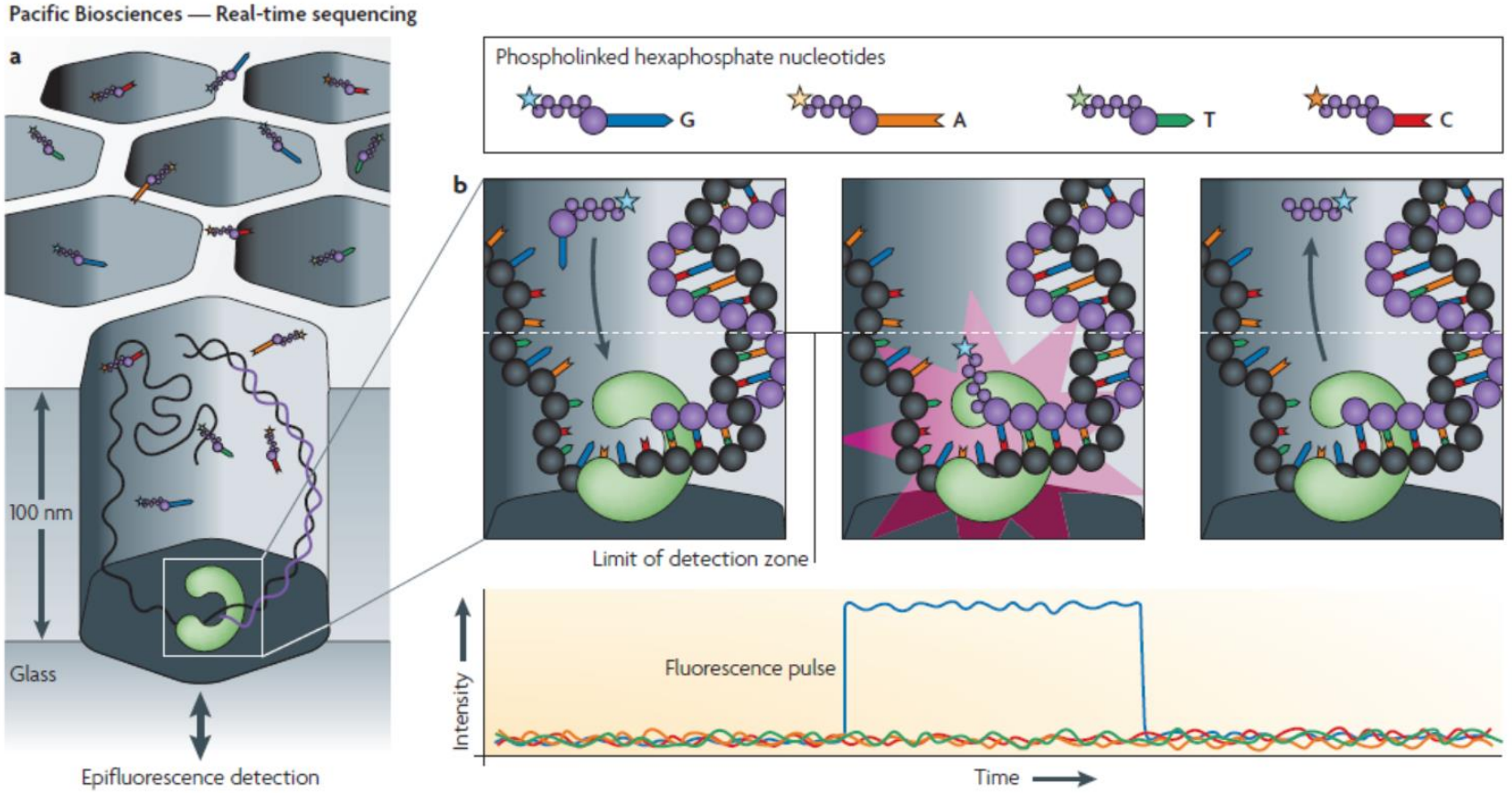
Для Illumina характерная длина чтения 100-200 п.н.

Что может пойти не так

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

Фрагмент ДНК слишком короткий, чтение захватывает последовательности адаптеров

Одномолекулярное секвенирование в реальном времени Pacific Biosciences



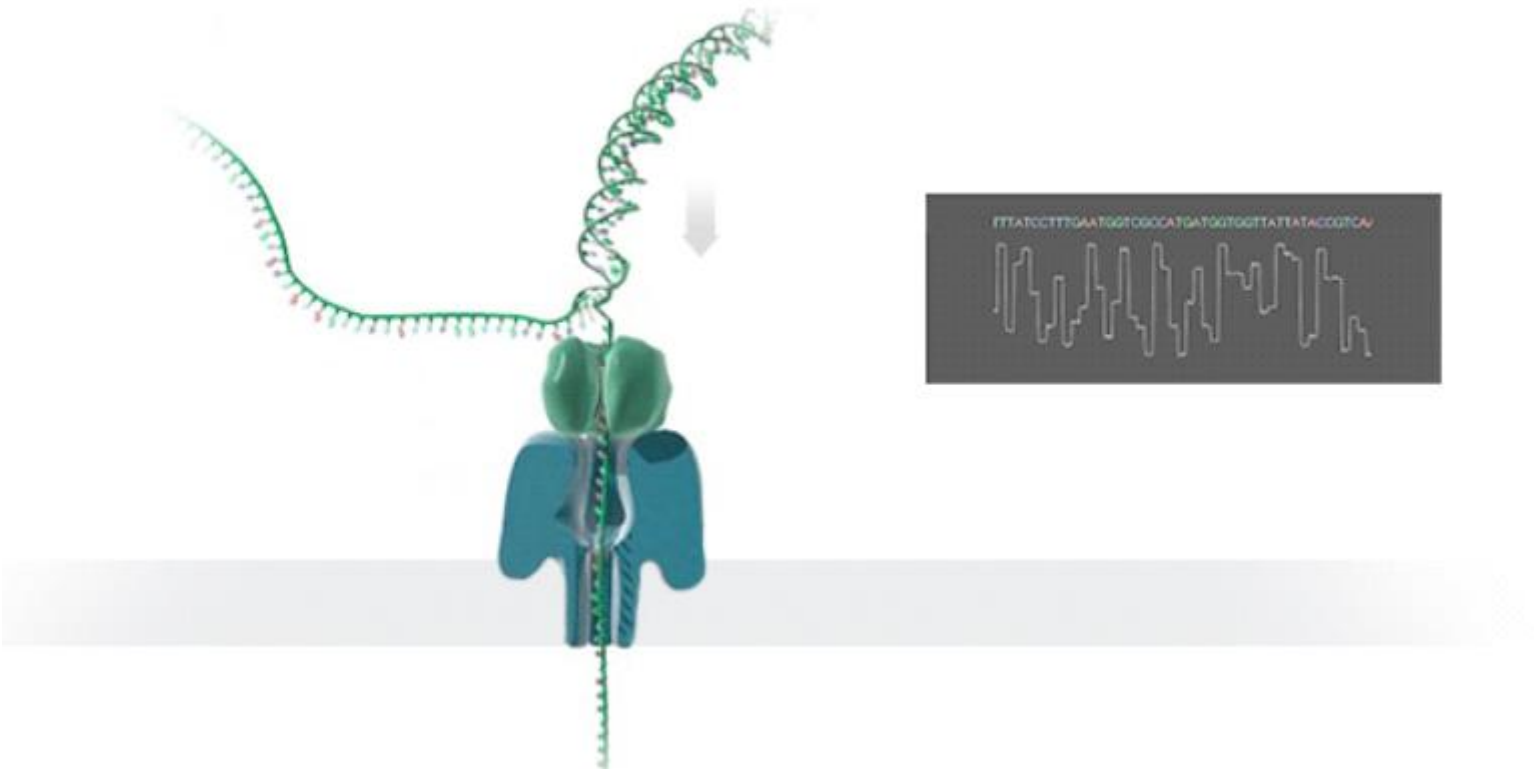
ru.wikipedia.org/wiki/Одномолекулярное_секвенирование_в_реальном_времени

Одномолекулярное секвенирование в реальном времени Pacific Biosciences

“+” : длина прочтений 20000-60000
без амплификации
быстро

“-” : большой процент ошибок
цена

Нанопоровое секвенирование Oxford Nanopore



https://www.skygen.com/katalog/oborudovanie/oxford_nanopore_technologies/nanoporovyy_sekvenator_minion/

Нанопоровое секвенирование Oxford Nanopore

“+” : длина прочтений 20000-60000
без амплификации
быстро
компактность и мобильность

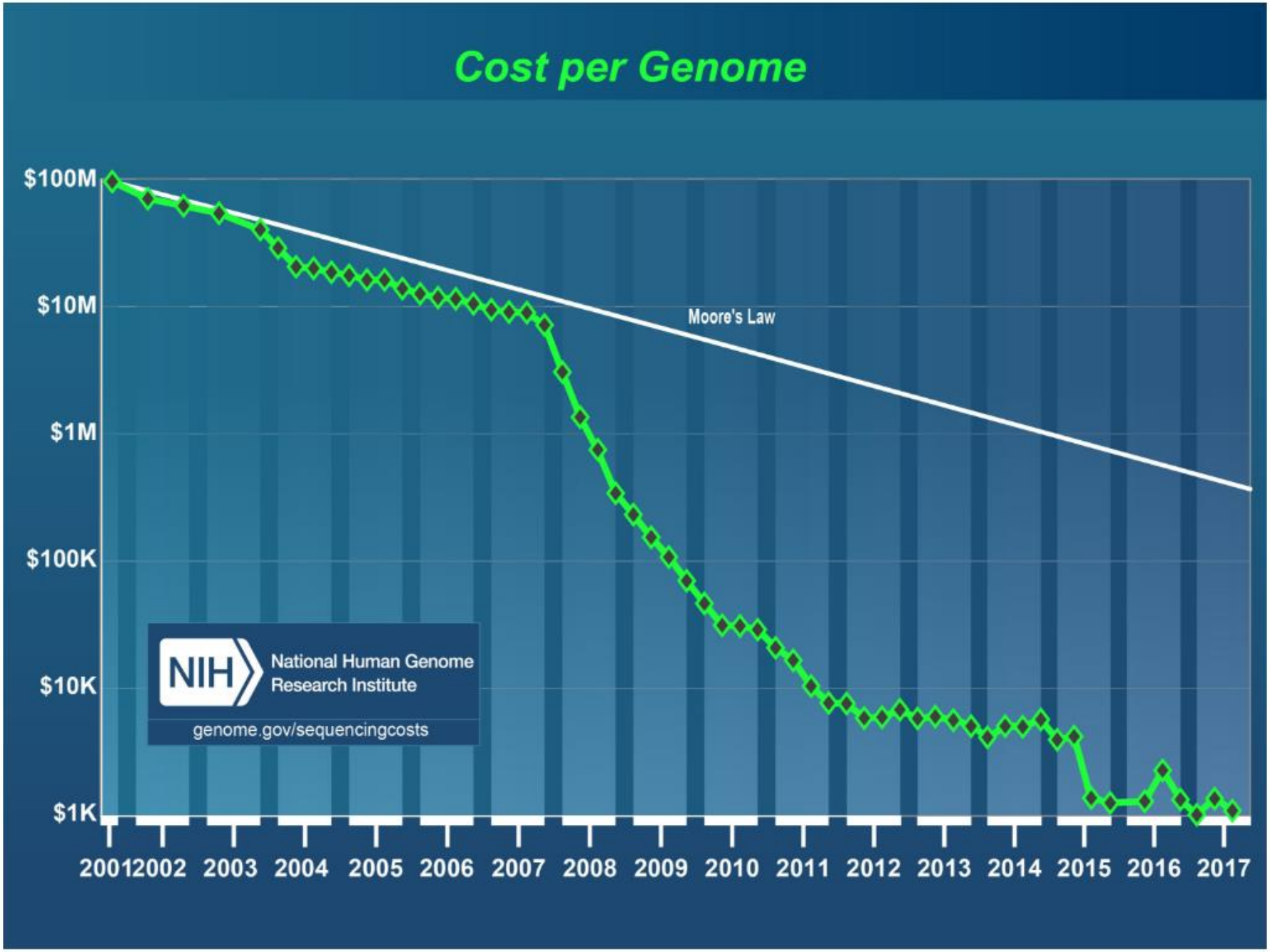
“-” : большой процент ошибок

Что же выбрать?

Все зависит от задачи!

Комбинировать платформы

Увеличивать покрытие



Источник: Source: DNA Sequencing Costs: Data

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

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

Важная информация!

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...)

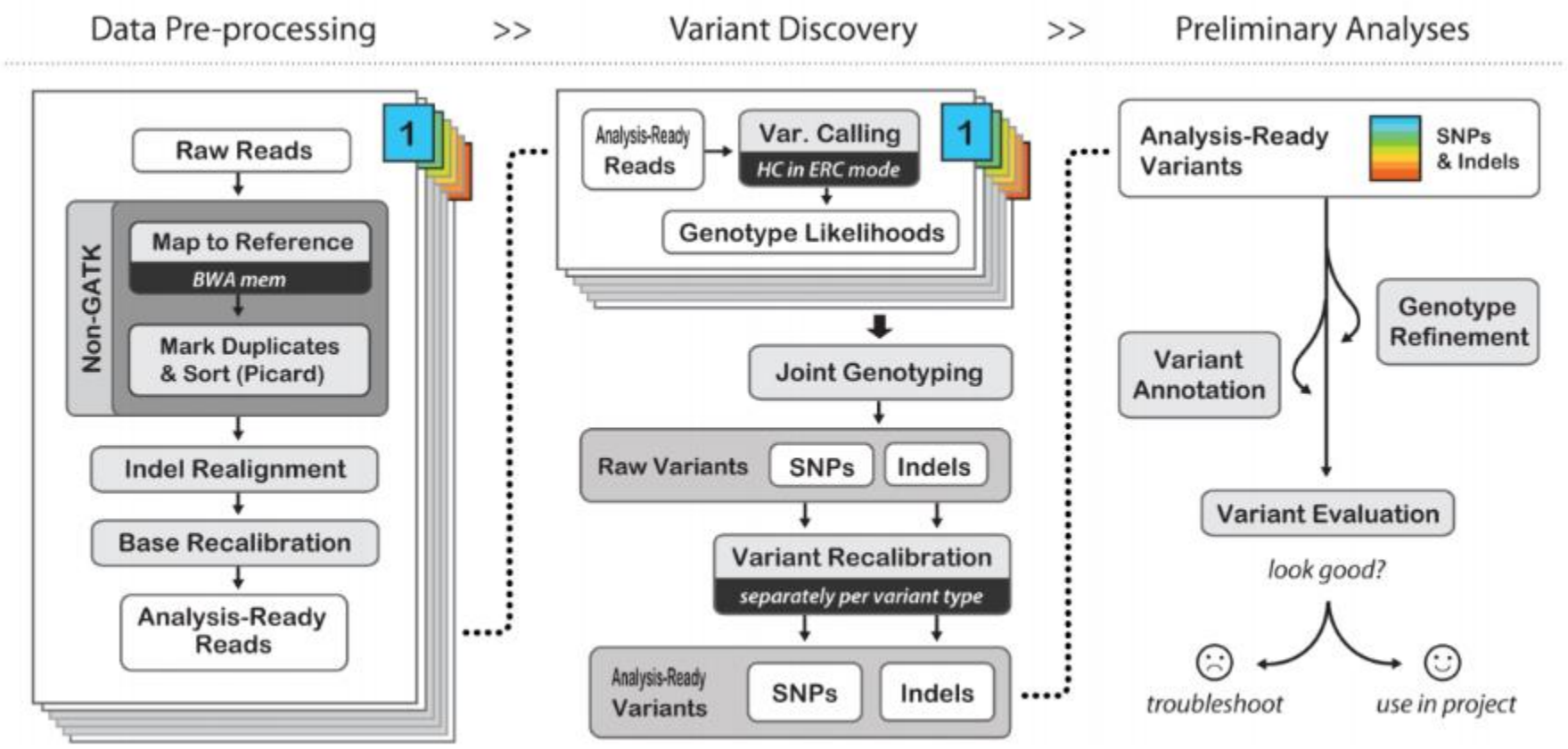
Runs: 1 run, 31.4M spots, 6.3G bases, 2.3Gb

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

ID: 11430914

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

Обработка данных



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)
```

Качество чтений

P – вероятность ошибки

Q – параметр качества. (Phred Quality Score)

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

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

Типичные значения Q от 1 до 40

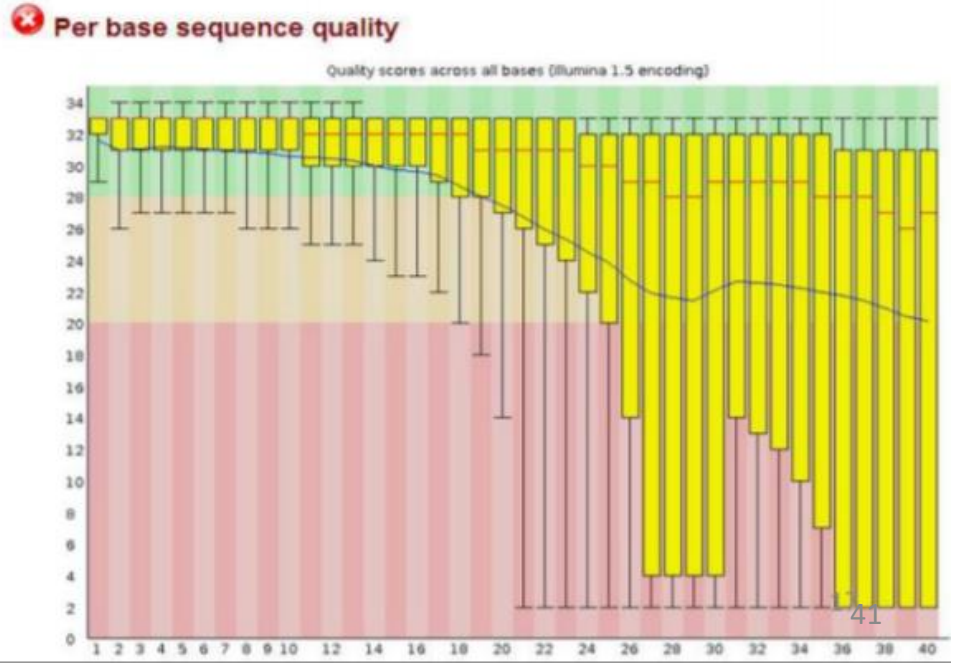
Q>20 – «хорошее качество»

Пересчет качества в вероятность ошибки

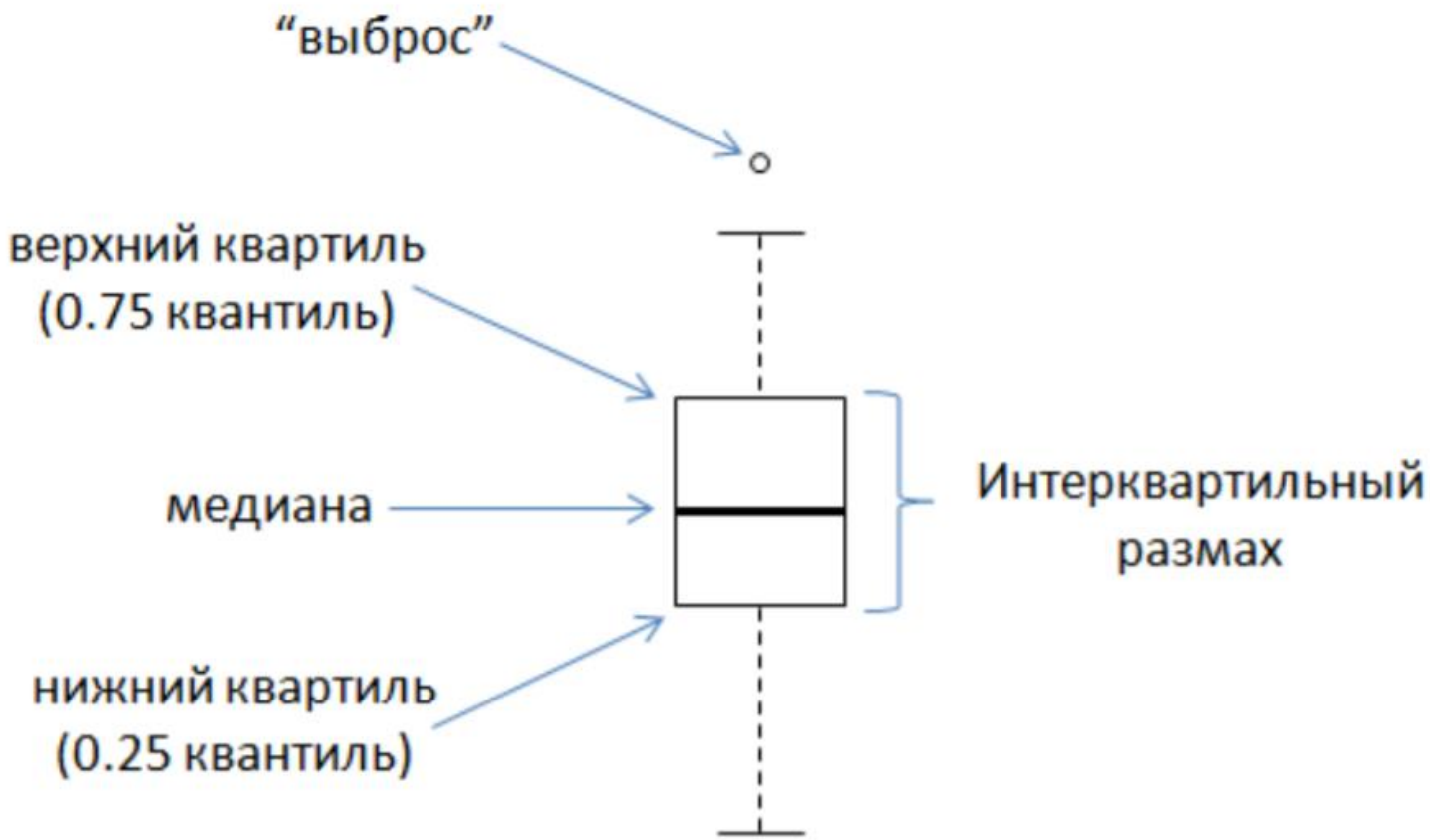
Phred Quality Score	Символ	Вероятность ошибки	Точность
10	+	1/10	90%
20	5	1/100	99%
30	?	1/1000	99,9%
40	I	1/10 000	99,99%
50	S	1/100 000	99,999%
60	]	1/1 000,000	99,9999%

fastQC


Per base sequence quality
Quality scores across all bases (Illumina 1.5 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12


Per base sequence quality
Quality scores across all bases (Illumina 1.5 encoding)
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40

«Ящик с усами» / диаграмма размахов / boxplot



fastQC

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

Программа FasqQC стоит на kodomo

Версию с графическим интерфейсом можно поставить на свой компьютер.

На сайте отличное руководство!