Ресеквенирование
Поиск полиморфизмов у человека

Анастасия Жарикова, 2019

Для чего нужно секвенирование?

Эволюция

Филогения

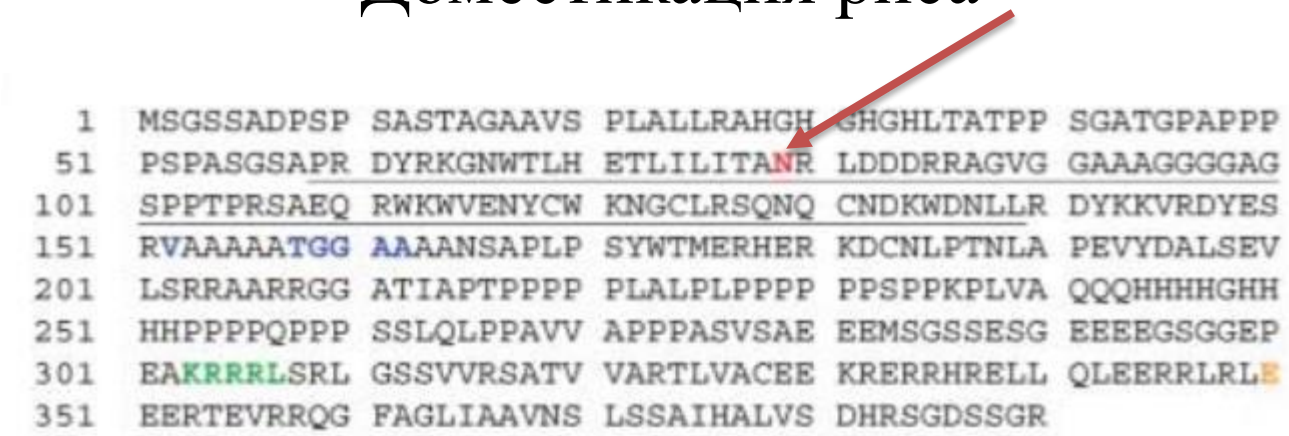
Клиника

Метагеномика

…

Эволюция

Доместикация риса



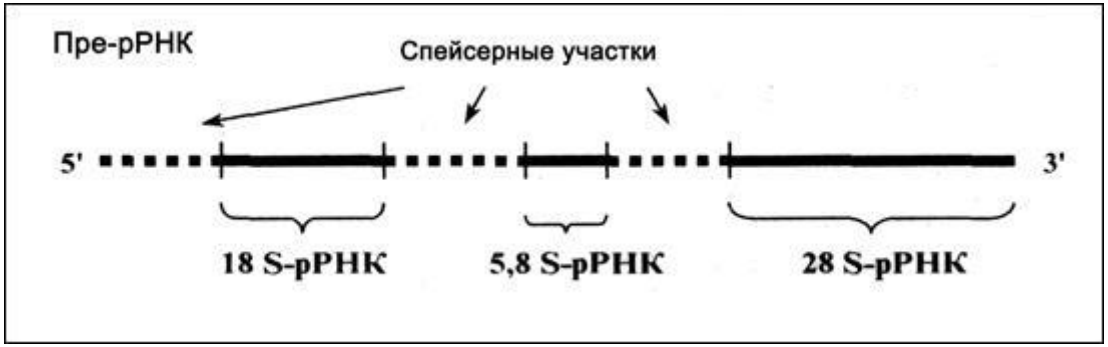
sh4

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 трансрибируемых спейсеров используется для изучения внутривидового разнообразия и классификации близкородственных организмов.

Популяционные и клинические исследования

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

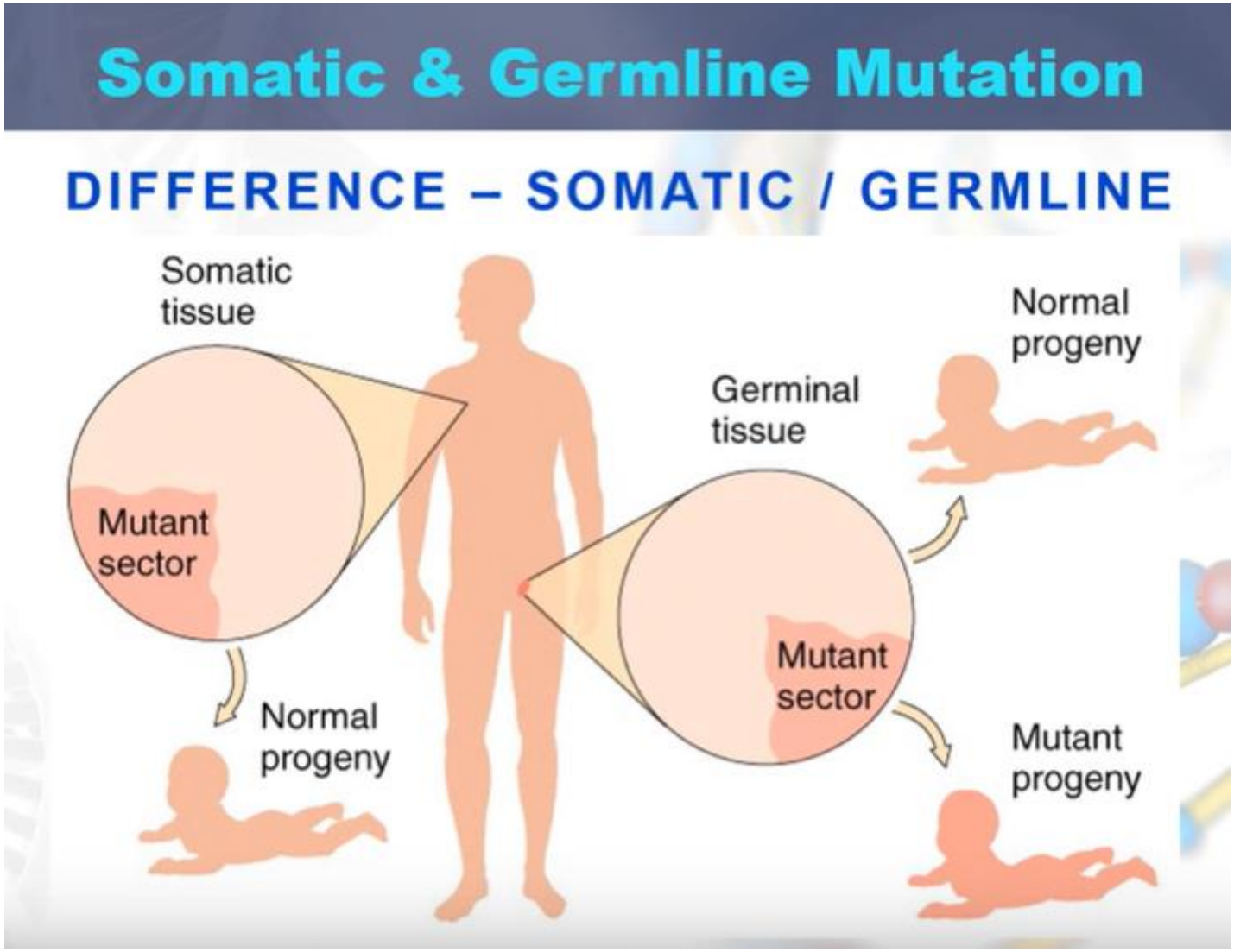
    - моногенные (муковисцидоз, ген CFTR)

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

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

Откуда берутся мутации?

Somatic & Germline Mutation

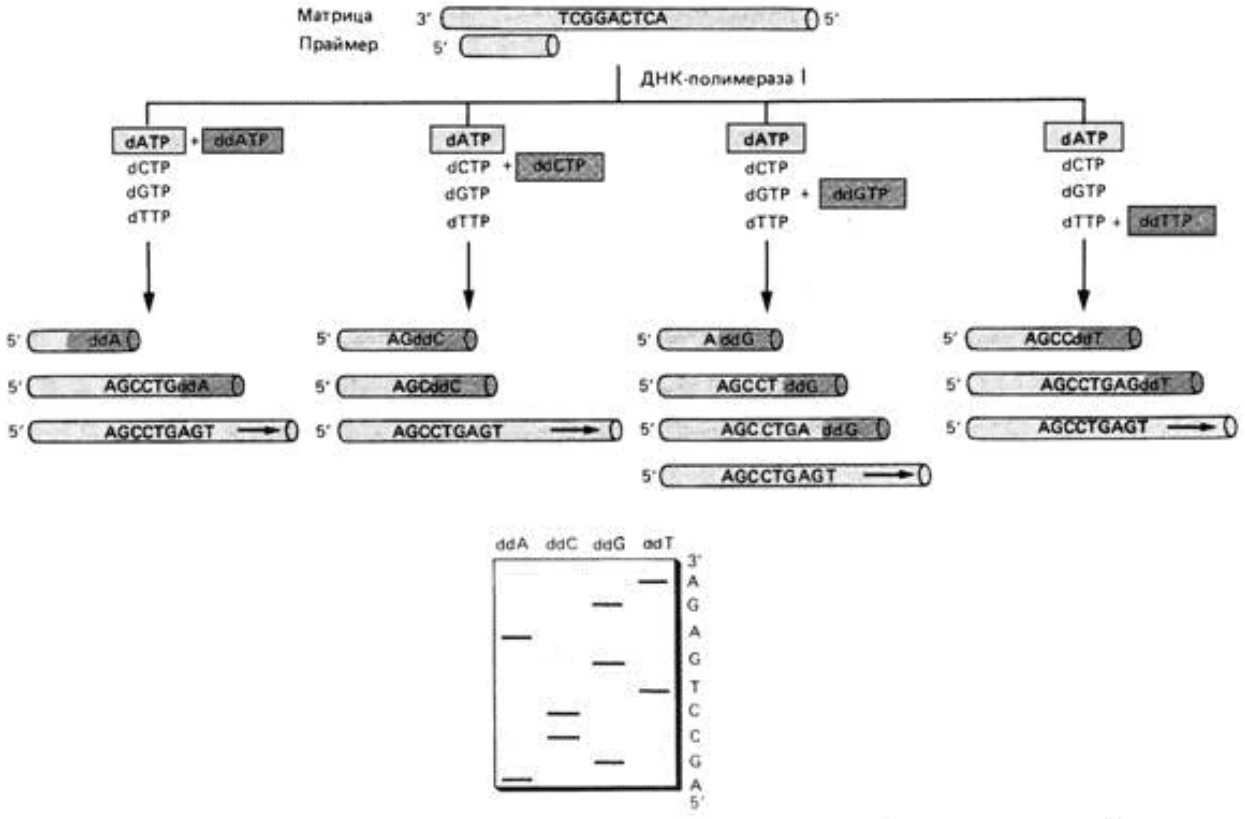


https://www.youtube.com/watch?v=3MiHSeH6yLg

Какие бывают мутации?

https://studopedia.info/5-45139.html

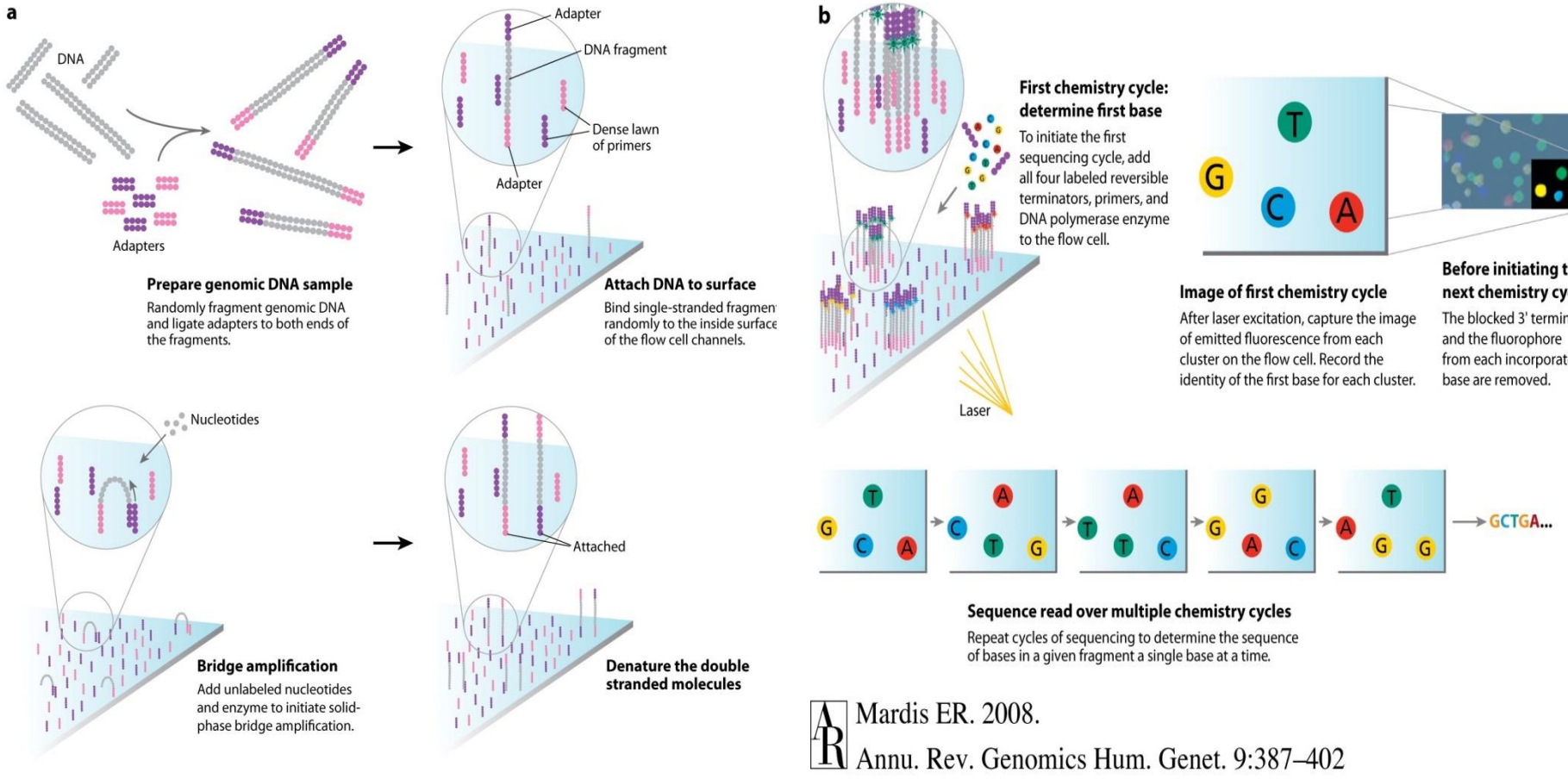
Секвенирование по Сэнгеру



Фредерик Сенгер – Нобелевская премия по химии 1980г.

Секвенирование второго поколения

Illumina

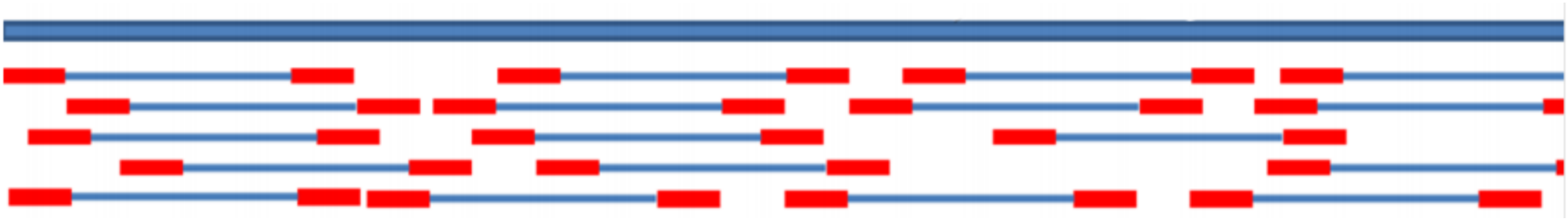


Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

http://postnauka.ru/longreads/468 - здесь хорошо и подробно рассказано о секвенировании (на русском)

http://www.youtube.com/watch?v=77r5p8IBwJk – наглядная демонстрация процесса секвенирования на Illumina

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

Норма

Димер

Фрагмент ДНК слишком короткий, чтение захватывает последовательности адаптеров

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 ридов

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»

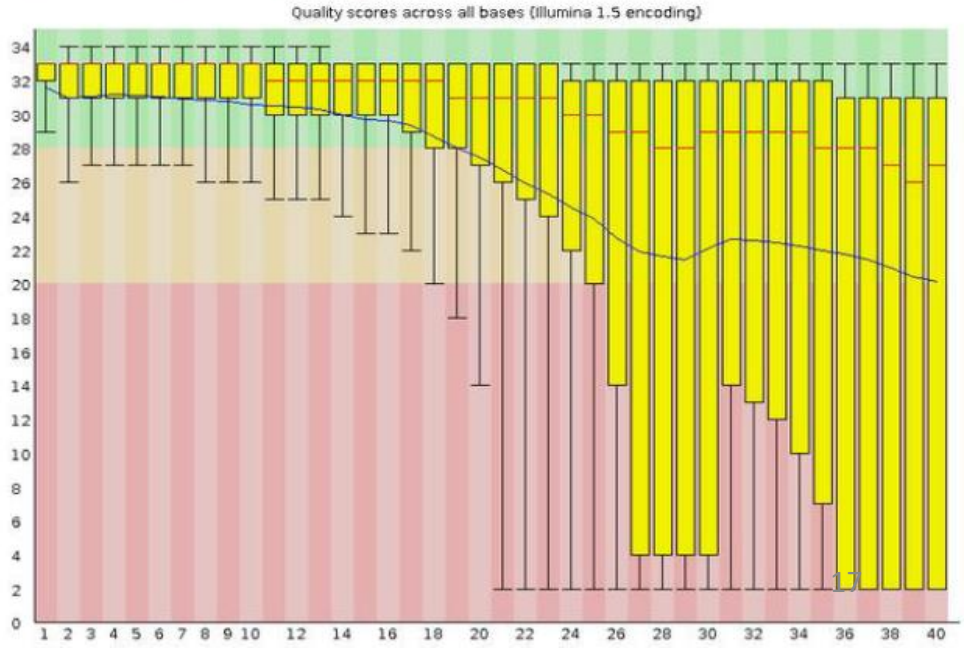
Пересчёт качества в вероятность ошибки

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%

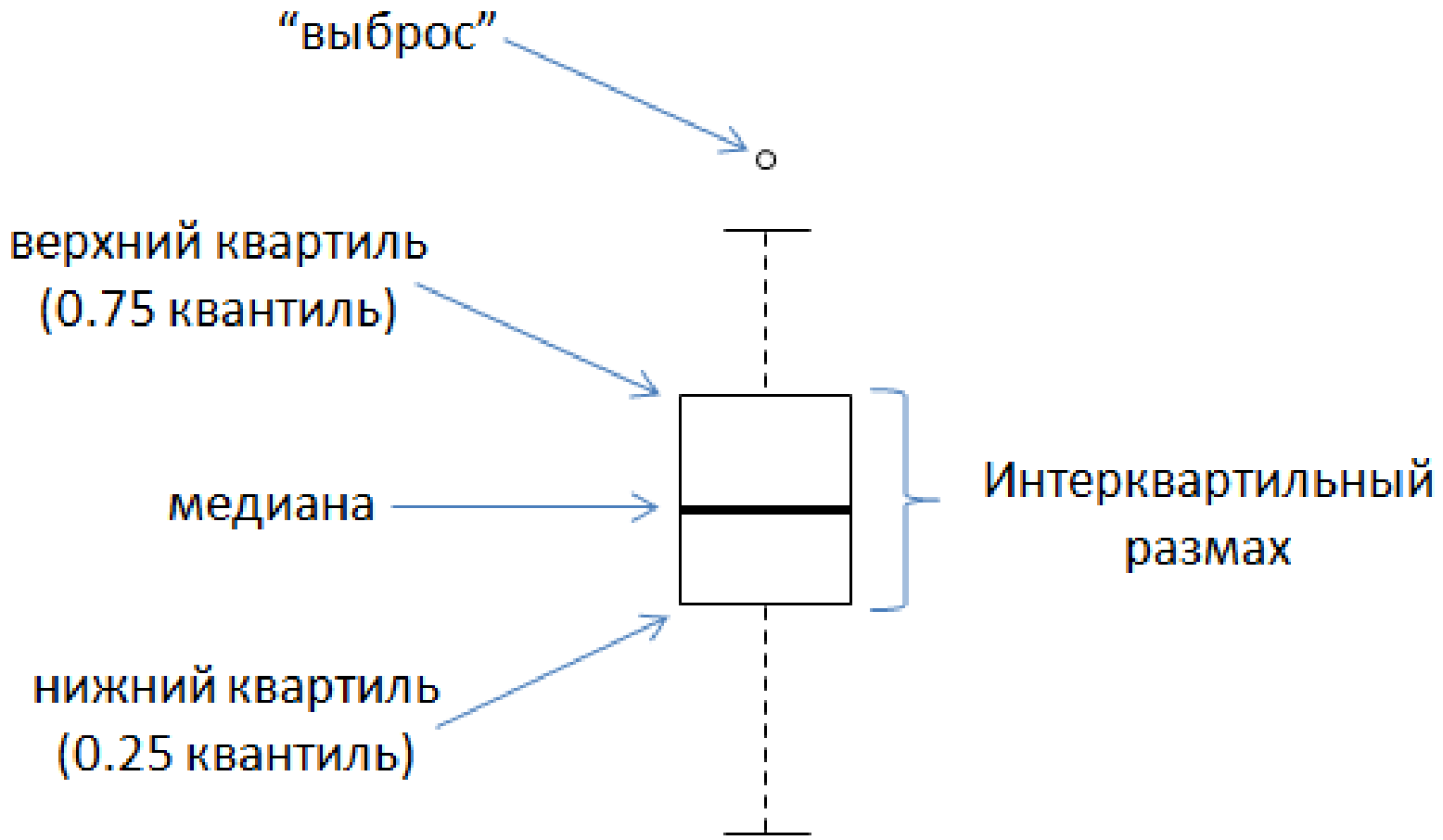
FastQC

Per base sequence quality

Per base sequence quality



Ящик с усами / диаграмма размахов / boxplot



FastQC

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!

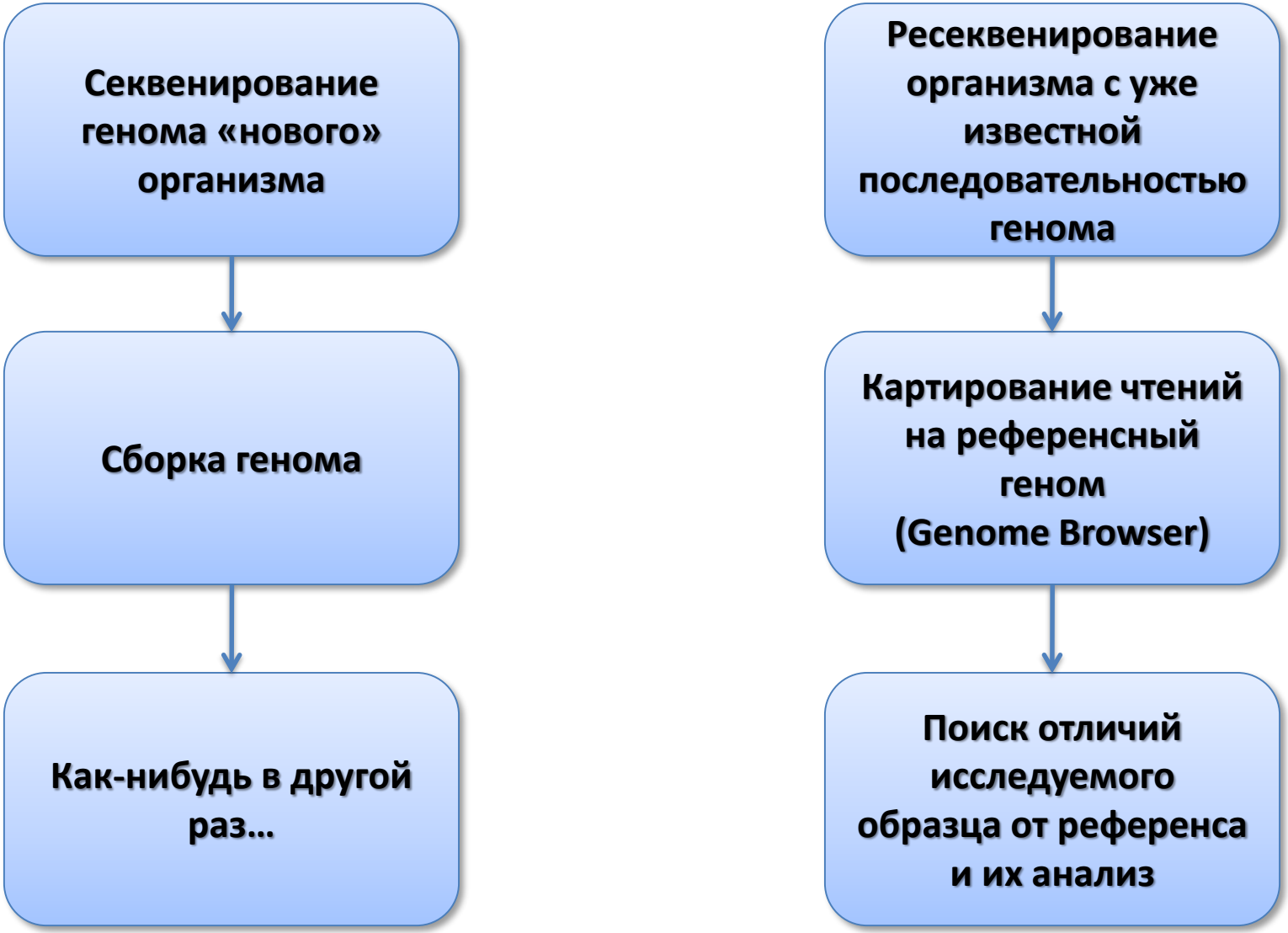
Что делать?

Нужно удалить «плохие» фрагменты чтений – тримминг:
- Удаление адаптеров
- Отсечение с конца чтения нуклеотидов с неудовлетворительным качеством (< 20)

Trimmomatic

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

Секвенирование бывает…



Возможности ресеквенирования

Можно ресеквенировать:

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

!!!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

Картирование

Задача: каждому чтению найти свое месте на референсном геноме.
Экзом – 1 человек - ~ 20-60 млн ридов

Последняя версия сборки генома человека – **hg38** (hg19 (GRCh37) – предыдущая)

Наиболее распространенные программы для картирования:
hisat2, bwa, bowtie1/2 (есть много других!)

1. Сначала индексируем референс. Для каждой программы свой индексный файл!

Для многих геномов индексные файлы можно скачать.

2. Процедура картирования → .sam
3. .sam → .bam (конвертирование)
   - .sam – текстовый файл
   - .bam – бинарный вариант .sam
4. Сортировка .bam файла
5. Индексирование отсортированного .bam файла

Hisat2

hisat2-build file.fasta file – аннотация референса

Некоторые параметры hisat2:
-x – путь к индексу
-U – путь к чтениям
--no-softclip – запрет подрезания чтений
--no-spliced-alignment – картирование без разрывов

SAM

Содержит заголовок и информацию о картировании чтений
http://samtools.github.io/hts-specs/SAMv1.pdf

```
SRR2776256.15395984     0       chr12   9822304 60      100M    *       0
0       AGATCACTCATAGAAACTGGAGGCAAAATGCATGACAGTAACAATGTGGAGAAAGACATTACACCATCTGAA
TTGCCTGCAAAGCCAGGTAAGAAGCTGG    ?@@DFFFDHHHHHJIJIHEGFAGHEG;FCFDFHI<GIJCFFDH?<<00
?98929/0.=B:8B78CC=CCEAAH=)=ECCB;7B;>@362@;@@C@CD359    AS:i:-4 XN:i:0  XM:i:1
XO:i:0  XG:i:0  NM:i:1  MD:Z:83C16       YT:Z:UU NH:i:1
SRR2776256.23192736     16      chr12   9822307 60      100M    *       0
0       TCACTCATAGAAACTGGAGGCAAAATGCATGACAGTAACAATGTGGAGAAAGACATTACACCATCTGAATTG
CCTGCAAACCCAGGTAAGAAGCTGGGCT    CCCC>;CEECEEEC@=DBC>ACHEHCD@=;G@GGGEHF=C<>IHFFGB
HGCDDGHGDFD?HGHEGGHFFGFA>GFH@HFADCHEHHBFHHHFFDDDD@@@    AS:i:0  XN:i:0  XM:i:0
XO:i:0  XG:i:0  NM:i:0  MD:Z:100         YT:Z:UU NH:i:1
```

SRR2776256.15395984 – ID чтения

chr12 9822304 - хромосома и координата, куда «легло» чтение

100M – CIGAR: сжато кодирует информацию о выравнивании чтения

NM:i – расстояние до генома

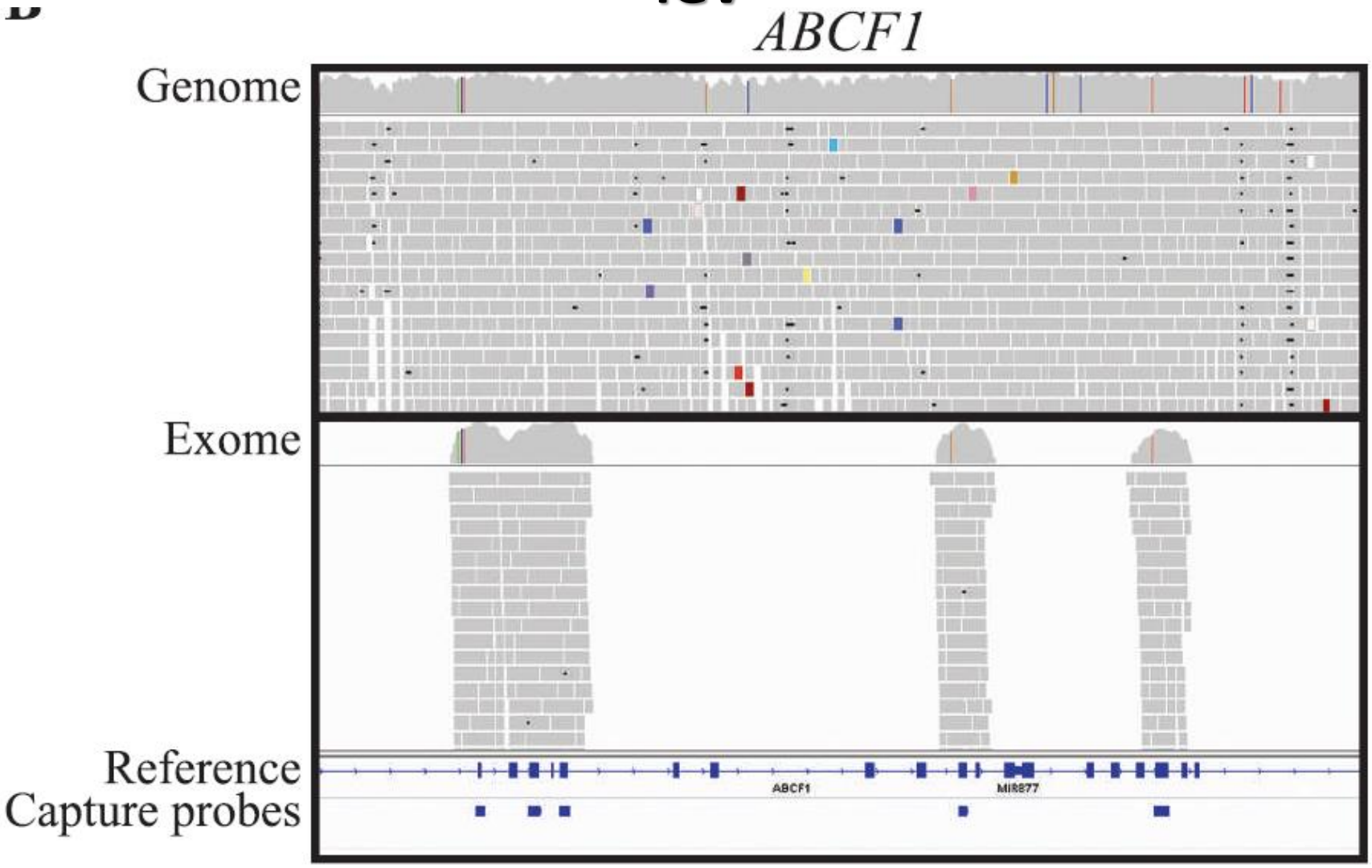
NH:I – количество картирований для данного чтения

Samtools – пакет для работы с BAM файлами
http://www.htslib.org/doc/samtools.html

Этот пакет поможет вам отсортировать и проиндексировать .bam файлы, узнать покрытие фрагмента генома и многое другое

Читайте руководство и подсказки к заданию!

IGV



Emily M. Coonrod, Jacob D. Durtschi, Rebecca L. Margraf, and Karl V. Voelkerding (*2013*) Developing Genome and Exome Sequencing for Candidate Gene Identification in Inherited Disorders: An Integrated Technical and Bioinformatics Approach. Archives of Pathology & Laboratory Medicine: March 2013, Vol. 137, No. 3, pp. 415-433.

Поиск SNP/INDELs

Читайте подсказки к заданию!

file.sorted.bam → file.bcf → file.vcf

file.sorted.bam – файл, полученный после картирования
file.vcf – файл содержит информацию об отличиях исследуемого образца от референса
file.bcf – бинарный file.vcf

.vcf файлы

Состоит из заголовка (строки, помеченные ##), в котором расшифровываются названия полей. Например, ##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth"> - т.е. поле DP показывает, сколько чтений пересекло конкретную позицию.

```
##FORMAT=<ID=SP,Number=1,Type=Integer,Description="Phred-scaled strand bias P-value">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
#CHROM  POS       ID  REF  ALT  QUAL  FILTER  INFO                                                  FORMAT      chr3.sorted.bam
chr3    41291081  .   G    A    140   .       DP=26;VDB=0.0371;AF1=1;AC1=2;DP4=0,0,7,17;MQ=60;FQ=-99  GT:PL:GQ   1/1:173,72,0:99
chr3    41310211  .   G    A    5.46  .       DP=1;AF1=1;AC1=2;DP4=0,0,0,1;MQ=60;FQ=-30               GT:PL:GQ   1/1:34,3,0:3
chr3    41310663  .   A    C    4.13  .       DP=2;AF1=0.5;AC1=1;DP4=0,1,0,1;MQ=60;FQ=3.81;PV4=1,1,1,1  GT:PL:GQ   0/1:32,0,31:3
```

Annovar

http://annovar.openbioinformatics.org/en/latest/

Программа, позволяющая аннотировать SNP на основании различных баз данных.
Мы возьмем некоторые (но есть очень много других!!!):

refgene
dbsnp
1000 genomes
GWAS
Clinvar

Описание каждой базы и запуск самой программы есть на сайте.

Клиническая интерпретация результатов

Это уже дело врачей-генетиков

СПАСИБО!