

Секвенирование организмов с известным геномом ("пересеквенирование")

с особым вниманием к экзоному
секвенированию в медицине

Елена Набиева
enabieva@gmail.com

План лекции

- Пересеквенирование ДНК
 - Референсный геном
 - Пересеквенирование в медицине, экзомное секвенирование
- Биоинформатика
 - картирование и постобработка
 - поиск полиморфизмов
 - функциональная классификация полиморфизмов

"Пересеквенирование" (resequencing)

- Секвенирование организма с уже известной последовательностью генома
 - *S. cerevisiae*, *D. melanogaster*, *H. sapiens*,....
- Существует "референсный геном" (reference genome)
- Genome Reference Consortium (GRC)
 - референсный геном человека, мыши и (теперь) *Danio rerio*

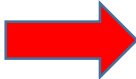
Референсный геном

- Гаплоидная последовательность генома организма
 - в нашем случае: *Homo Sapiens*
- Не (обязательно) является геномом конкретного индивида
 - человеческий: основан на геномах нескольких человек
- Не является "идеальным" геномом
 - может содержать редкие/неблагоприятные аллели
 - GRC старается заменять редкие аллели частыми гаплотипами
- Пробелы в знании
 - Неизвестные последовательности в хромосоме: "NNNNN"
 - Последовательности, не встроенные в хромосомы: chr*_random, chrUn*
- Вариация в популяции, варьибельные локусы (иммунитет, ...)
 - Решение: альтернативные гаплотипы в референсе

Референсный геном человека

- Текущая сборка генома человека: GRCh38/hg38
 - например, <http://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/>
- "Analysis set" – для использования с данными NGS
 - <ftp://hgdownload.cse.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/>
- GRCh38:
 - «альтернативные контиги», в т.ч. для HLA
 - Требуется аккуратности в биоинформатической обработке

Пересеквенирование в медицине

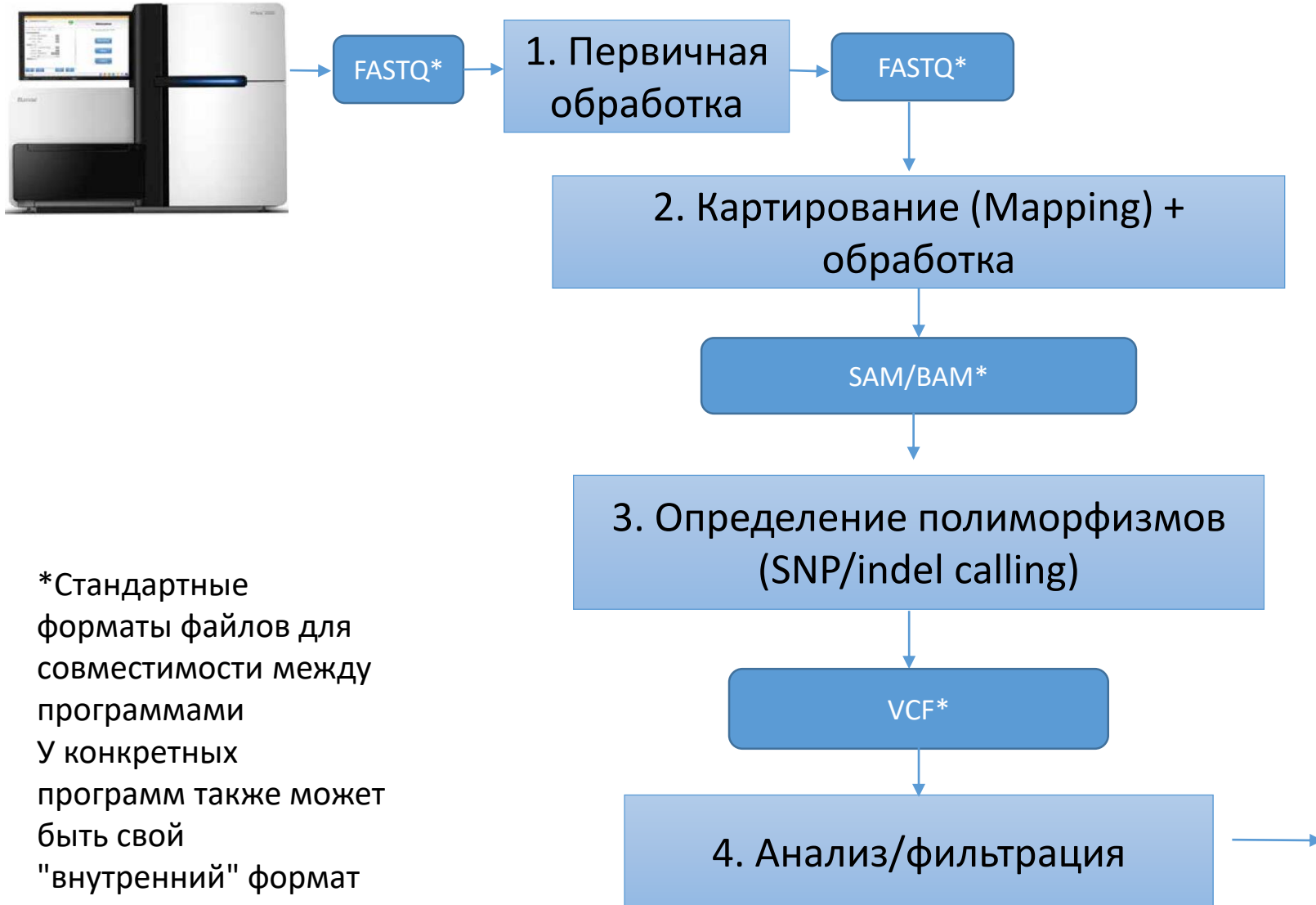
- Некоторые применения:
 - выявление причины менделевского заболевания
 - молекулярная диагностика пациентов с менделевскими заболеваниями
 - выявление предрасположенностей
 - выявление соматических (т.е. не врожденных) мутаций в раковых опухолях
 - ...
- В настоящее время распространено пересеквенирование отдельных участков
 -  • полного экзона (всех кодирующих последовательностей)
 - интересующих участков (например, набора генов)

Экзомное секвенирование

- Плюсы. Цена:информативность
 - Цена: экзм ~ 1.5% от генома человека
 - Информативность: содержат подавляющее большинство известных болезнетворных мутаций
 - Анализ: изучены лучше всего
- Минусы. Ограниченность информации и шум
 - Ограниченность участками, входящими в набор зондов*
 - Дополнительный экспериментальный шум
 - неравномерность покрытия экзонов
- В то же время:
 - Наборы часто включают в себя микроРНК, UTR и т.д.
 - На практике часть чтений ложится на участки, не являющиеся "мишенями" (off-target)

Биоинформатика

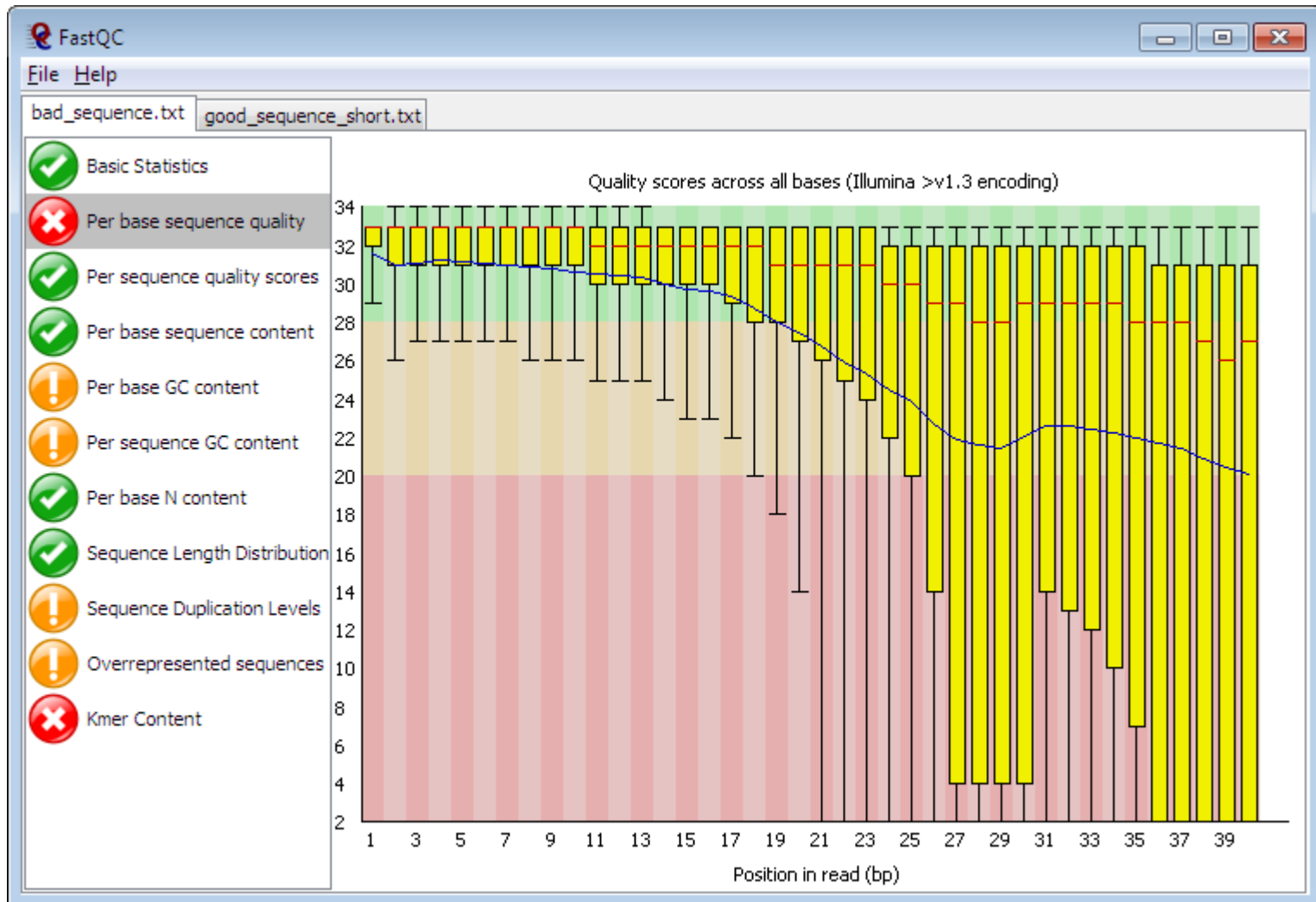
Биоинформатический процесс и стандартные форматы файлов (поиск полиморфизмов)



*Стандартные форматы файлов для совместимости между программами
У конкретных программ также может быть свой "внутренний" формат



Оценка качества чтений: FastQC

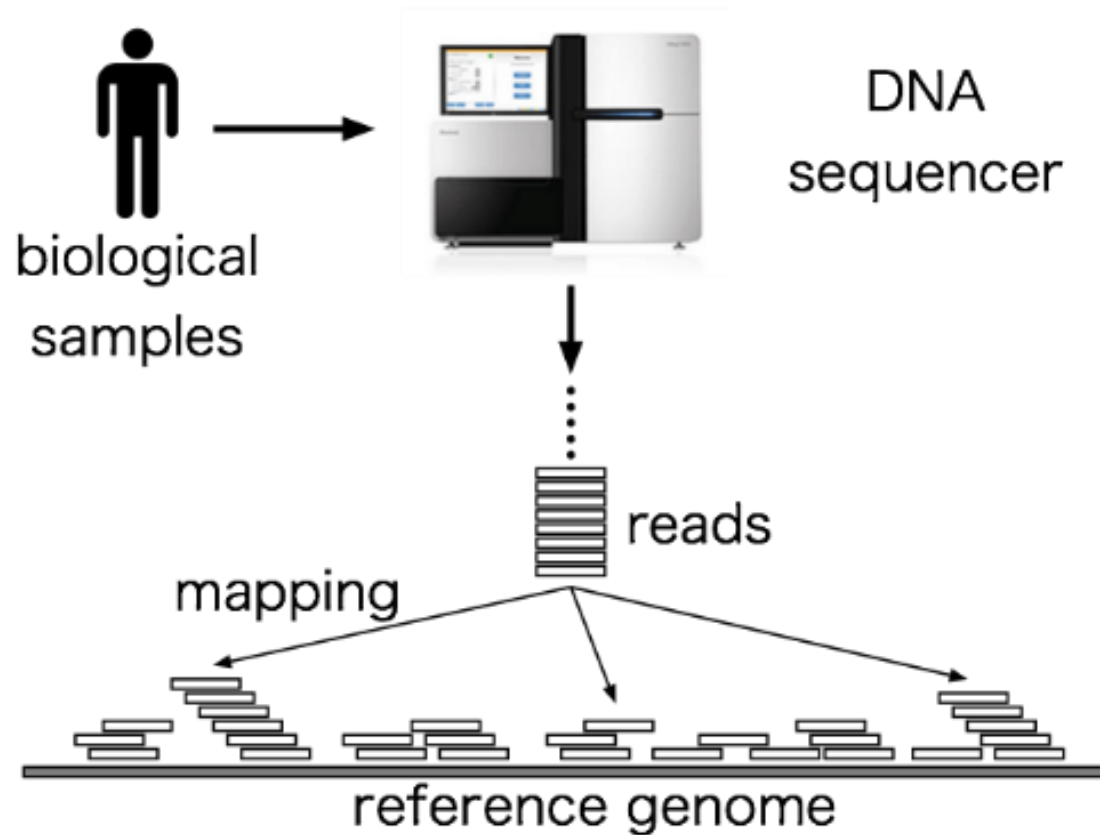


<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

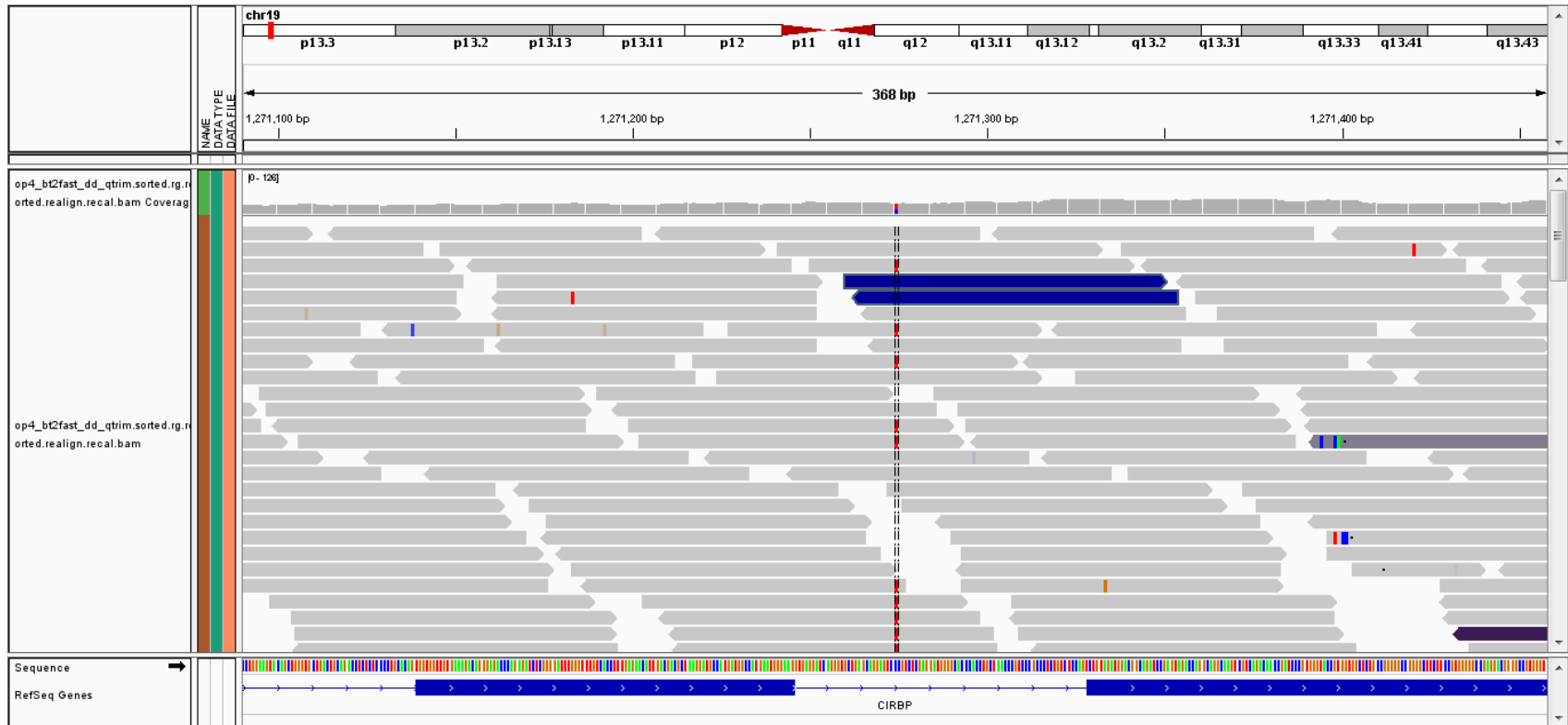
Предварительная обработка чтений (reads)

- Обрезание оставшихся артефактов секвенирования (адаптеров)
- Обрезание концов чтений с низким BAQ
 - Осторожно, если планируется дедупликация
- Программы:
 - cutadapt, trimmomatic

Read mapping ("картирование")



Read mapping ("картирование")



IGV

Mapping ("картирование", выравнивание)

- Дано: чтения (например, пары по 100 пн)
- Задача: найти им место в геноме
 - пары: предпочтительно, с соблюдением парности
- отличия от обычного выравнивания:
 - **много** коротких чтений на **один** длинный референс
 - чтения очень близки к референсу (обычно), часто парные
 - источники несовпадений:
 - технические ошибки
 - естественные отличия между людьми
 - несовпадения распределены неслучайно (см. лекцию про секвенирование)
- На выходе: файл в формате BAM

Алгоритмы картирования

- На основе хэширования
 - Хэширование помогает найти положение «зерна», потом выравнивание продляется от «зерна»
- На основе трансформации Барроуза-Уилера
 - Ср. суффиксные массивы

Mapping: программы

- Популярны: семейства **bowtie** и **bwa** (и др.)
 - Предварительный шаг: *индексация* референсного генома
 - преобразование Барроуза — Уилера (**B**urrows-**W**heeler Transform)
 - индекс создается один раз (для данного референса)
 - индекс -> быстрый поиск
 - Относительно мало памяти
 - порядка 3-4 GB для генома человека
 - Идея:
 - индекс для выравнивания "зёрен" (seeds)
 - расширяем выравнивание для всего чтения
 - ср. BLAST

Семейство bowtie

<http://bowtie-bio.sourceforge.net/>

- bowtie (1) (Langmead, Trapnell, Pop, Salzberg 2009)
 - для относительно коротких чтений (до 50 bp, но максимум в районе 1000 bp)
 - выравнивание без разрывов
- bowtie2 (Langmead, Salzberg 2012)
 - для более длинных чтений (50+, ограничения по длине нет)
 - разрешаются разрывы
 - большая гибкость при выравниваниях
- Каждая версия требует своего индекса генома!
 - bowtie-build/bowtie2-build (или скачать с сайта для стандартных геномов)
- bowtie положен в основу т.н. "tuxedo suite", с программами для анализа РНК-сек данных

Семейство bwa

<http://bio-bwa.sourceforge.net/>

- "bwa backtrack" (Li, Durbin 2009) – для чтений до 100 bp
 - два шага: bwa aln + bwa samse/sampe
bwa aln ref.fa short_read.fq > aln_sa.sai
bwa samse ref.fa aln_sa.sai short_read.fq > aln-se.sam (одноконечные чтения)
bwa sampe ref.fa aln_sa1.sai aln_sa2.sai read1.fq read2.fq > aln-pe.sam (парные чтения)
- bwa bwasm (Li, Durbin 2010) – для чтений длиной 70bp-1Mbp
 - создает проблемы для picard MarkDuplicates (по моему опыту)
 - о дедупликации позже
 - позволяет разрывы
- bwa mem (Li H. (2013) [arXiv:1303.3997v2] - для чтений длиной 70bp-1Mbp
 - "быстрее и лучше", чем bwa bwasm
 - флажок -M для последующей работы с picard MarkDuplicates
- Индексация: bwa index
 - -a is - для коротких геномов (вариант по умолчанию)
 - -a bwtsv – работает для длинных геномов (напр., человека)
 - Внимание: разные версии (релизы) bwa могут не работать с индексными файлами, созданными другими версиями!



SNAP

- Основан на **хэшировании**.
- *“3-20x faster and just as accurate as existing tools like BWA-mem, Bowtie2 and Novoalign”*
- Zaharia et al. arXiv:1111.5572v1, November 2011.

Некоторая специфика

- GATK требует ReadGroups:
- Добавлять при картировании:

```
bwa mem -R $rg_string ...
```

где \$rg_string =
@RG\tID:justchild_0.01_1367\tPL:ILLUMINA\tCN:BI\tSM:justchild\tPI:220\tLB:justchild_0.01_1367\tDS:justchild_0.01_1367\tPU:justchild_0.01_1367 ... (пример)
- Добавлять с помощью picard AddOrReplaceReadGroups

Альтернативные контиги (GRCh38)

- Как вариант, использовать скрипт: bwa-postalt.js:

```
bwa mem -M -R $rg_string $reference_fasta_file $fastq_file1 $fastq_file2 > ${bam_prefix}.sam
```

bwa.kit/k8 /uge_mnt/home/enabieva/bwa.kit/bwa-postalt.js -p \$id.alt \$reference_fasta_file.alt \$bam_prefix.sam
- См. обсуждение и разбор на <https://software.broadinstitute.org/gatk/blog?id=8180>
- <https://github.com/lh3/bwa/blob/master/README-alt.md>

Формат SAM/BAM (binary SAM)

<http://samtools.sourceforge.net/SAMv1.pdf>

- Заголовок (header) + информация о картировании чтений

[illegible]

Также: формат cram для *сжатых* выравниваний. См. пакет cramtools

SAM FLAG

Кодирует, насколько хорошо "легло" чтение и его пара

Каждый бит имеет свое значение

| Bit | Description | Запросы с помощью масок: |
|------|---|--------------------------|
| 1 | 0x1 template having multiple segments in sequencing | $> 99 \& 0x1$ |
| 2 | 0x2 each segment properly aligned according to the aligner | $> 99 \& 0x2$ |
| 4 | 0x4 segment unmapped | $> 99 \& 0x4$ |
| 8 | 0x8 next segment in the template unmapped | $> 99 \& 0x8$ |
| 16 | 0x10 SEQ being reverse complemented | $> 99 \& 0x10$ |
| 32 | 0x20 SEQ of the next segment in the template being reverse complemented | $> 99 \& 0x20$ |
| 64 | 0x40 the first segment in the template | $> 99 \& 0x40$ |
| 128 | 0x80 the last segment in the template | $> 99 \& 0x80$ |
| 256 | 0x100 secondary alignment | $> 99 \& 0x100$ |
| 512 | 0x200 not passing quality controls | $> 99 \& 0x200$ |
| 1024 | 0x400 PCR or optical duplicate | $> 99 \& 0x400$ |
| 2048 | 0x800 supplementary alignment | $> 99 \& 0x800$ |

Если "флаг" чтения = 99,
то "флаг" пары = 147

<https://samtools.github.io/hts-specs/SAMv1.pdf>

99 = 0b01100011

147 = 0b10010011

"Decoding SAM flags" онлайн

The screenshot shows a web browser window with the address bar displaying <https://broadinstitute.github.io/picard/explain-flags.html>. The page header for Picard includes a 'build passing' status and links for 'Latest Jar Release', 'Source Code ZIP File', 'Source Code TAR Ball', and 'View On GitHub'. The main section is titled 'Decoding SAM flags' and contains the following text: 'This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties. To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.'

Below the text is a form with the label 'SAM Flag:' followed by an input field and an 'Explain' button. A 'Switch to mate' button is also present, with a tooltip that reads 'Toggle first in pair / second in pair'.

The interface is divided into two columns: 'Find SAM flag by property:' and 'Summary:'. The 'Find SAM flag by property:' column contains a list of properties with checkboxes: 'read paired', 'read mapped in proper pair', 'read unmapped', 'mate unmapped', 'read reverse strand', 'mate reverse strand', 'first in pair', 'second in pair', 'not primary alignment', 'read fails platform/vendor quality checks', 'read is PCR or optical duplicate', and 'supplementary alignment'. The 'Summary:' column is currently empty.

<https://broadinstitute.github.io/picard/explain-flags.html>

Работа с BAM-файлами: пакет samtools

<http://samtools.sourceforge.net/samtools.shtml>

- samtools view: показать содержимое bam-файла
 - -h : показать включая заголовок
 - -f INT : показать только выравнивания с битами INT в поле FLAG (в 16-ричной системе)
 - -F INT: *не* показывать выравнивания с битами INT в поле FLAG
 - -q INT: показывать только выравнивания с MAPQ > INT
 - -bS: получить SAM-файл на вводе (-S), создать BAM-файл на выводе (-b)


фильтры!

MAPQ

- "MAPping Quality"
 - оценка картировщиком качества/уникальности картирования
 - значения зависят от картировщика
 - выглядит как BAQ (Phred score = $-10 \log_{10} P(\text{ошибка})$)
 - больше = лучше

Работа с BAM-файлами: пакет samtools

<http://samtools.sourceforge.net/samtools.shtml>

- samtools sort: сортировать bam-файл
 - samtools index: индексировать bam-файл
 - samtools tview: простой текстовой просмотрщик bam-файла
- 
- Они или их аналоги
нужны почти всегда

samtools flagstat – простая статистика на bam-файле

- Пример (парные чтения):

```
% samtools flagstat somefile.bam
```

```
63204618 + 0 in total (QC-passed reads + QC-failed reads)
```

```
0 + 0 duplicates
```

← Информативно только после дедупликатора!

```
62888312 + 0 mapped (99.50%:-nan%)
```

```
63204618 + 0 paired in sequencing
```

```
31602309 + 0 read1
```

```
31602309 + 0 read2
```

```
61305324 + 0 properly paired (97.00%:-nan%)
```

```
62751130 + 0 with itself and mate mapped
```

```
137182 + 0 singletons (0.22%:-nan%)
```

```
718912 + 0 with mate mapped to a different chr
```

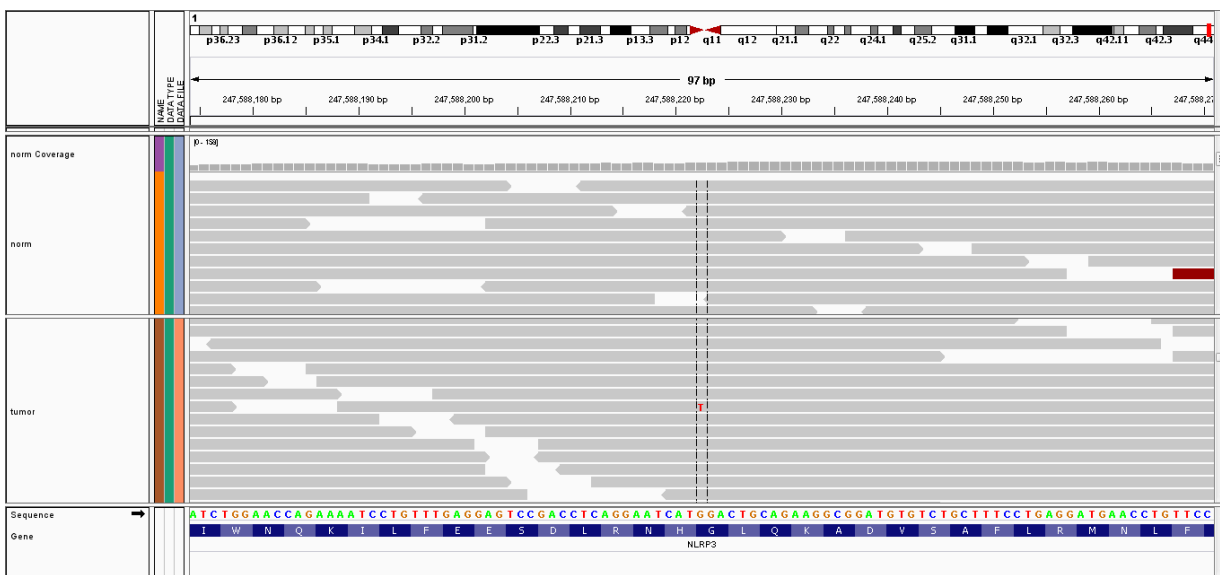
```
474976 + 0 with mate mapped to a different chr (mapQ>=5)
```

CRAM – формат файлов

- Более компактный формат, чем BAM
- Кодирование картирования относительно конкретного референса
 - Референс должен совпасть при прочитывании
 - Смотрит на MD5 sum
- Для большего сжатия: полная или частичная потеря информации о BAQ
- samtools, cramtools

Integrative Genomics Viewer

- <http://www.broadinstitute.org/igv/>
- Программа для визуального просмотра bam-файлов



- Важно проверять найденные варианты "глазами"

Требует индекс бам-файла

Альтернативы: Tablet...

Пост-обработка выравниваний

- Дедупликация чтений
 - убрать чтения, дублицированные в результате ПЦР
 - `picard markDuplicates`
 - `biobambam bammarkduplicates`
 - `samtools rmdup`

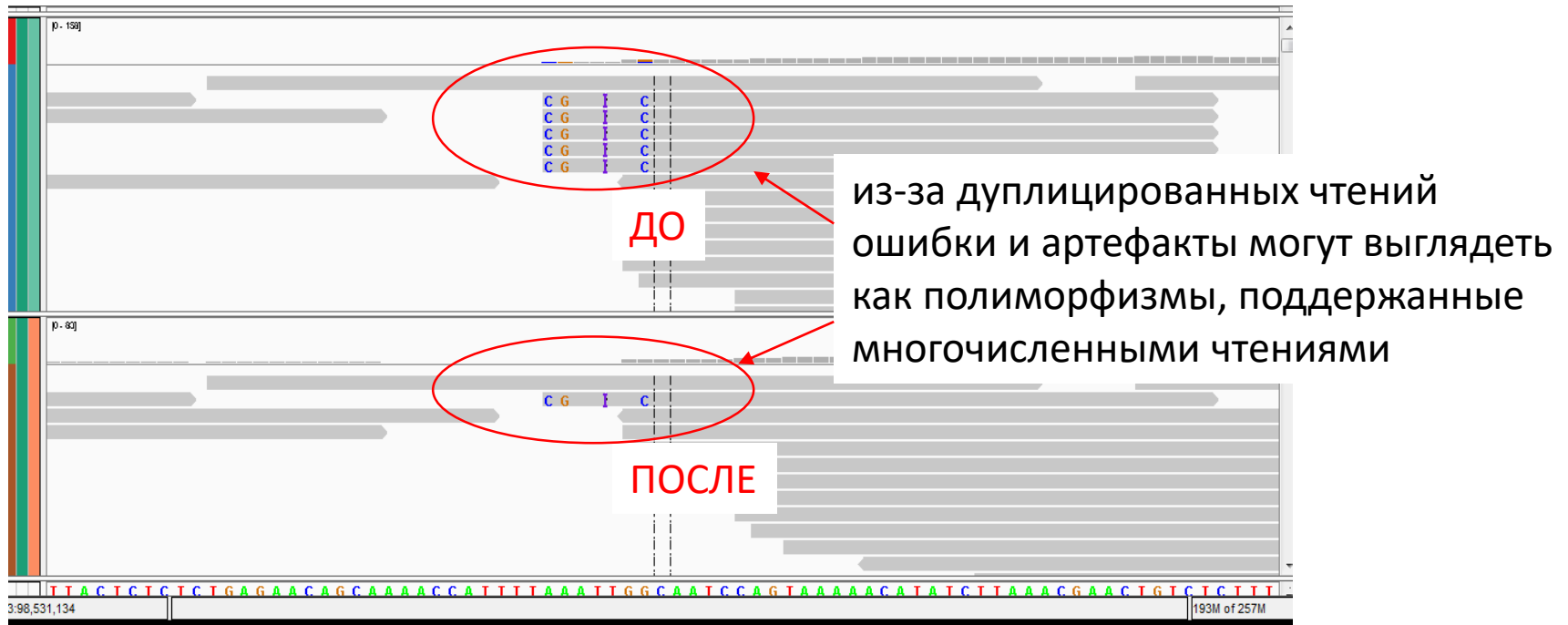
- Пост-обработка от Genome Analysis Toolkit (GATK):

- `indel realignment`
перевыравнивание вокруг инделов

← Не обязательно, если variant caller делает локальную пересборку

- `base recalibration`
 - рекалибрация качества оснований

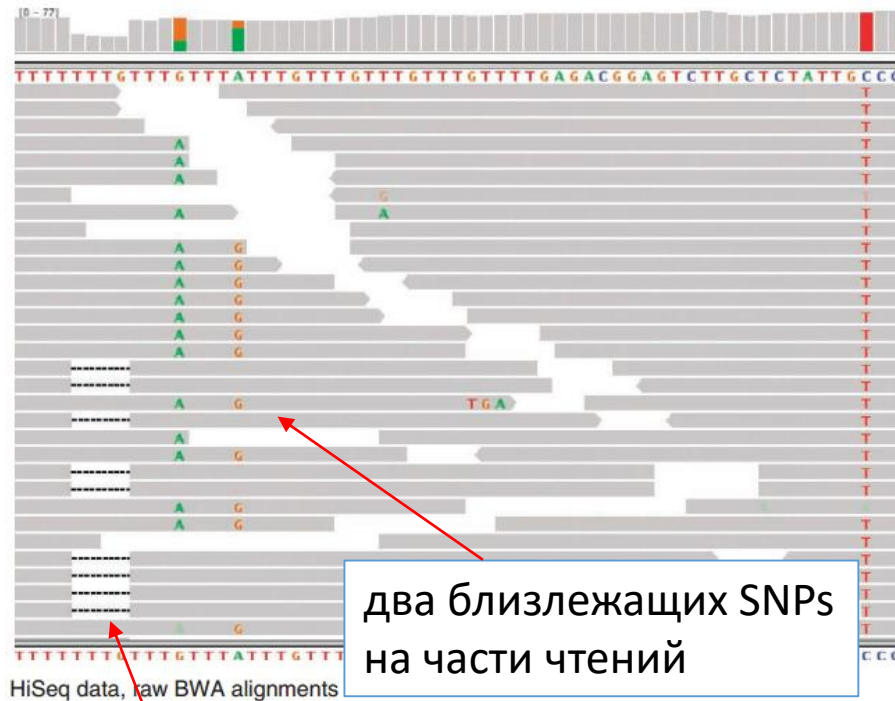
Иллюстрация: дедупликация



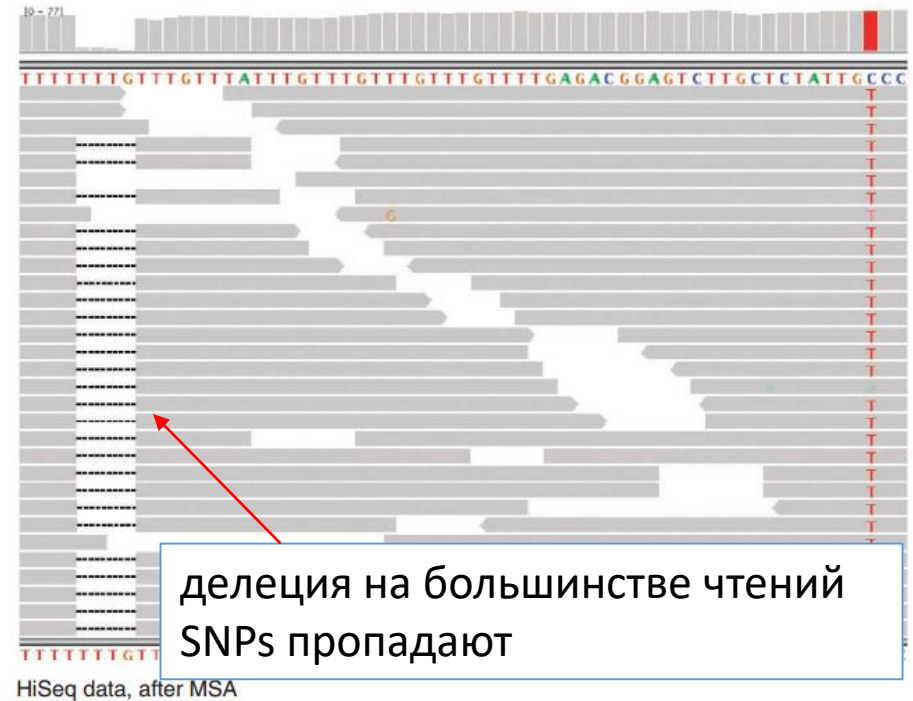
- picard MarkDuplicates: дублицированные пары чтений выровнены одинаково
- Внимание: при очень высоком покрытии (например, РНК-сек) могут возникать "ложные" дубликаты

Иллюстрация: Indel realignment

ДО



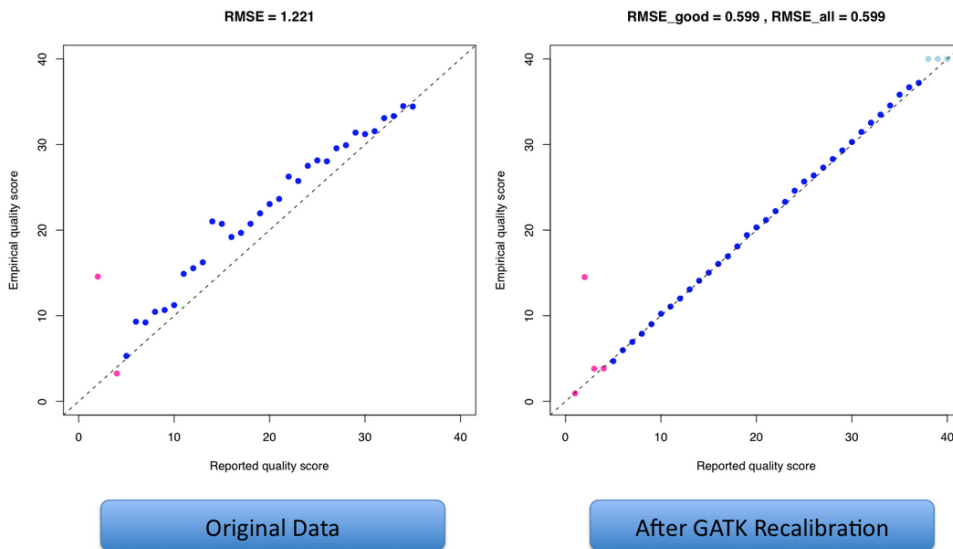
ПОСЛЕ



Также: Base quality score recalibration (GATK)

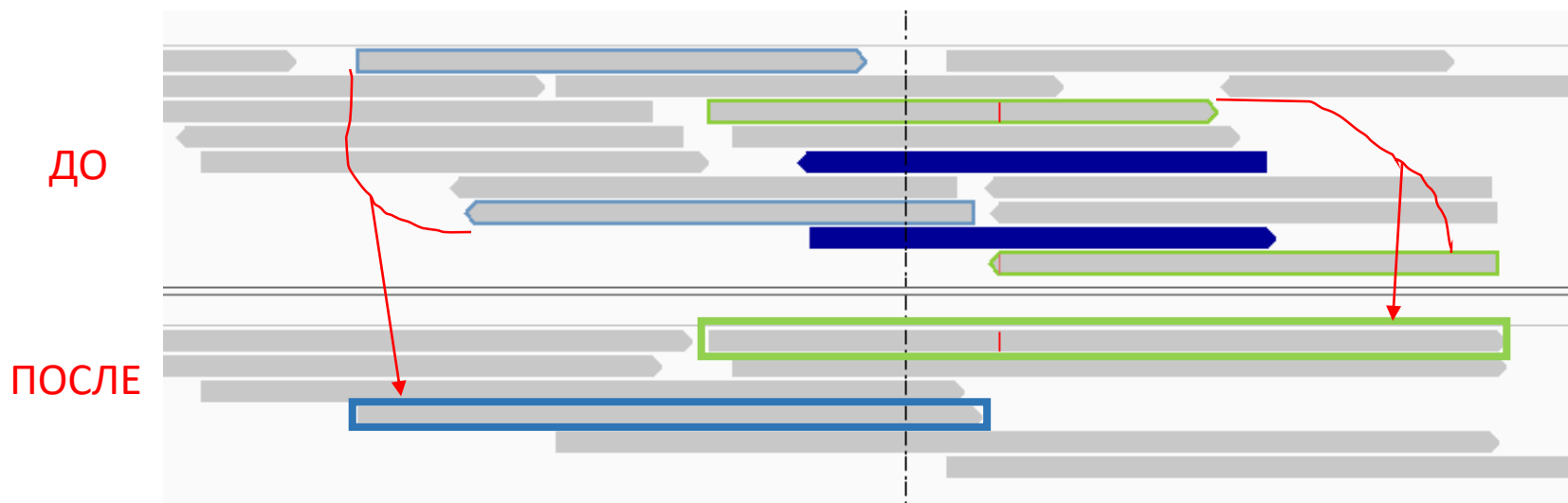
- Приводит предсказанные секвенатором оценки качества прочтения нуклеотида к эмпирически наблюдаемым (по несовпадениям с референсом)

Reported Quality vs. Empirical Quality



Другие аспекты: пересекающиеся чтения

- Если длина парных чтений близка к длине фрагмента, чтения пары будут существенно пересекаться
 - пример: длина фрагмента 300, длина чтения 250
- Имеет смысл объединять пересекающиеся чтения (до картирования)
 - особенно если SNP-caller не учитывает парную принадлежность чтений



Контроль качества - минимум

- % картированных чтений, % уникально картированных чтений
 - samtools flagstat
 - отчет картировщика (напр. bowtie2)
 - посчитать чтения с соотв. флажками (-q) используя samtools view ... | wc
 - на практике MAPQ может не совсем точно отражать "уникальность" выравнивания
- % чтений, закартированных на "мишени"
(для экзомного / направленного секвенирования)
 - см. следующий слайд
- среднее покрытие
 - просто: делить количество закартированных чтений на длину интервала
 - сложнее: см. следующий слайд

Пакет bedtools – работа с геномными интервалами

- <http://bedtools.readthedocs.org/en/latest/content/bedtools-suite.html>

- Пример интересующих интервалов: кодирующие участки

```
% intersectBed -wa -u -abam input.bam -b intervals.bed |  
samtools view -h - | samtools view -bS - >  
input.X.interval.bam
```

Создает пересечение input.bam с интервалами в intervals.bed

```
%coverageBed -abam input.bam -b intervals.bed -d >  
coveragedepth.bed
```

#Пишет глубину покрытия для каждой координаты в файле intervals.bed

(предупреждение: создает огромный файл!)

... и многое другое

Также: %samtools depth -b intervals.bed bam.bam

Примечание: .bed – простой формат файла для описания геномных интервалов

Для экзоминого секвенирования: получить от поставщика набора

Variant Calling

Программы

- Наиболее распространенные:
 - samtools
 - <https://www.htslib.org/workflow/>
 - Genome Analysis Toolkit (GATK)
 - Более-мене стандарт для человеческих данных
 - GATK3
 - GATK4 (значительные изменения)
 - <https://gatkforums.broadinstitute.org/gatk/discussion/11145/germline-short-variant-discovery-snps-indels>
 - ... многие другие

Выявление однонуклеотидных полиморфизмов и коротких инделов (SNP/indel calling)

индел = инсерция или делеция

Общая идея:

- определить сайты, где все или часть чтений отличаются от референса
- для каждого сайта оценить "достоверность" отличия
 - учитывая количество и качество "альтернативных" чтений

Пре-фильтрация:

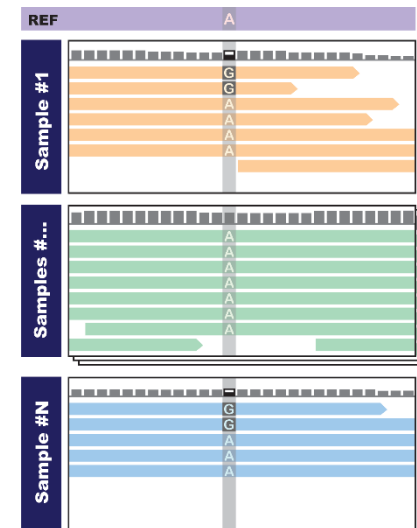
- часто используется предварительная фильтрация
 - нуклеотидов по BAQ (качество прочтения нуклеотида)
 - чтений по MAPQ (качество "картирования")
 - чтений по количеству несовпадений с референсом
 - обычно это устанавливается в настройках SNP-caller

Пост-фильтрация:

- имеет смысл не рассматривать SNPs/инделов в "подозрительных" сайтах

"Передовые" алгоритмы

- Локальная сборка гаплотипов
 - "Интересные" регионы собираются de novo
 - Нет необходимости в indel realignment (?)
 - Кто делает? (примеры)
 - GATK HaplotypeCaller
 - FreeBayes
- Если образцов много
 - Можно учитывать информацию со всех образцов
 - опасность "потерять" редкие варианты?
 - Кто умеет? (примеры)
 - bcftools
 - GATK



GATK

GATK для многих образцов

- Для каждого образца:
 1. HaplotypeCaller с флагом -ERC GVCF
 - Создает «промежуточный» файл формата .g.vcf
 - долго
 2. GenotypeGVCFs со всеми .g.vcf файлами
 - Создает файл формата .vcf
 - относительно быстро
- Зачем?
 - облегчает добавление новых образцов

Формат VCF

- Стандарт описания вариантов
- Довольно гибок; вывод разных программ обычно отличается (иногда достаточно сильно)
- CHROM, POS, REF, ALT сообщают основную информацию об аллеле
- GT (genotype) сообщает информацию о сайте:
 - 0/0 = гомозиготен по референсному аллелю
 - 0/1 = гетерозиготен
 - 1/1 = гомозиготен по альтернативному аллелю
 - (в мультиаллельных сайтах цифр может быть больше)
- Эта информация используется последующими программами для аннотирования вариантов
- Разнообразные программы могут (не) писать свою информацию о полиморфизме в поле INFO (а также QUAL, FILTER) – где-то в этих полях будет оценка программой качества найденного полиморфизма

Формат VCF: заголовок (header)

```
emacs@SAMSUNG2016
File Edit Options Buffers Tools Help
[Icons]
##fileformat=VCFv4.2
##ALT=<ID=NON_REF,Description="Represents any possible alternative allele at this location">
##FILTER=<ID=LowQual,Description="Low quality">
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alt alleles in the order listed">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (reads with MQ=255 or with bad mates are filtered)">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=MIN_DP,Number=1,Type=Integer,Description="Minimum DP observed within the GVCf block">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information, describing how the alternative alleles are phased in relation to one another">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID information, where each unique ID within a given sample (but not across samples) connects records within a phasing group">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification">
##FORMAT=<ID=RGQ,Number=1,Type=Integer,Description="Unconditional reference genotype confidence, encoded as a phred quality -10*log10 p(genotype call is wrong)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistics which comprise the Fisher's Exact Test to detect strand bias.">
##GATKCommandLine.GenotypeGVCFs=<ID=GenotypeGVCFs,Version=3.8-0-ge9d806836,Date="Tue Jul 10 10:25:13 UTC 2018",Epoch=1531218313317,CommandLineOptions="analysis_type=GenotypeGVCFs input_file=[] showFullBamList=false read_buffer_size=null read_filter=[] disable_read_filter=[] intervals=[/u/mnt/home/enabieva/hg19/gencode.v27lift37.annotation.exon.positions.coding.withflank50.bed] excludeIntervals=null interval_set_rule=UNION interval_merging=ALL interval_padding=0 reference_sequence=/u/mnt/home/enabieva/hg19/ucsc.hg19.fasta nonDeterministicRandomSeed=false disabledDithering=false maxRuntime=-1 maxRuntimeUnits=MINUTES downsampling_type=BY_SAMPLE downsample_to_fraction=null downsample_to_coverage=1000 baq=OFF baqGapOpenPenalty=40.0 refactor_NDN_cigar_string=false fix_misencoded_quality_scores=false allow_potentially_misencoded_quality_scores=false useOriginalQualities=false defaultBaseQualities=-1 performanceLog=null BQSR=null quantize_qual=0 static_quantized_qual=null round_down_quantized=false disable_indel_qual=false emit_original_qualities=false preserve_qscores_less_than=6 globalQScorePrior=-1.0 secondsBetweenProgressUpdates=10 validation_strictness=SI
```

Формат VCF

```
file edit options buffers tools help
chrM 14581 . G A 93546.90 . AC=6;AF=1.00;AN=6;DP=2607;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=54.84;QD=29.43;SOR=15.697 GT:AD:DP:GQ:PL 1/1:0,711:711:99:30828,2158,0 1/1:0,295:295:99:1
0111,886,0 1/1:0,1555:1555:99:52634,4665,0
chrM 14906 . A G 12451.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=1.25;ClippingRankSum=0.00;DP=836;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=59.90;MQRankSum=0.00;QD=30.09;ReadPosRankSum=0.476;SOR=1
0.175 GT:AD:DP:GQ:PL 1/1:1,323:324:99:12489,935,0 0/0:41,0:41:99:0,122,1600 0/0:463,0:463:99:0,120,1800
chrM 15627 . C T 2115.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=1.25;ClippingRankSum=0.00;DP=836;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=59.90;MQRankSum=0.00;QD=30.09;ReadPosRankSum=0.476;SOR=1
0.175 GT:AD:DP:GQ:PL 1/1:1,323:324:99:12489,935,0 0/0:41,0:41:99:0,122,1600 0/0:463,0:463:99:0,120,1800
chrM 15933 . C T 18632.90 . AC=6;AF=1.00;AN=6;DP=593;ExcessHet=3.0103;FS=0.000;MLEAC=6;MLEAF=1.00;MQ=60.00;QD=34.51;SOR=12.587 GT:AD:DP:GQ:PL 1/1:0,66:66:99:2188,198,0 1/1:0,19:19:57:711
,57,0 1/1:0,455:455:99:15760,1366,0
chr1 69511 . A G 295.13 . AC=2;AF=0.00;AN=4;DP=44;ExcessHet=3.0103;FS=0.000;MLEAC=2;MLEAF=0.500;MQ=27.58;QD=26.83;SOR=4.977 GT:AD:DP:GQ:PL 0/0:33,0:33:99:0,99,1131 1/1:0,11:11:33:329,33,0 ./
.:0,0:0:0:0,0
chr1 69569 . T C 259.89 . AC=1;AF=0.00;AN=4;BaseQRankSum=2.74;ClippingRankSum=0.00;DP=46;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.250;MQ=30.12;MQRankSum=-1.022e+00;QD=18.56;ReadPosRankSum=0.650;SOR=2.303
GT:AD:DP:GQ:PL 0/0:32,0:32:81:0,81,1215 0/1:3,11:14:57:289,0,57 ./.:0,0:0:0:0,0
chr1 135134 . G A 14.82 . AC=2;AF=0.500;AN=6;DP=81;ExcessHet=0.7918;FS=0.000;MLEAC=1;MLEAF=0.167;MQ=21.42;QD=7.41;SOR=2.303 GT:AD:DP:GQ:PL 0/0:33,0:33:99:0,99,1288 0/0:45,0:45:99:0,111,1665
1/1:0,2:2:6:49,6,0
chr1 135301 . G A 15.86 . AC=2;AF=0.500;AN=6;DP=49;ExcessHet=0.7918;FS=0.000;MLEAC=1;MLEAF=0.167;MQ=20.00;QD=7.93;SOR=2.303 GT:AD:DP:GQ:PL 1/1:0,2:2:6:49,6,0 0/0:45,0:45:99:0,111,1665
0/2,0:2:6:0,6,66
chr1 135374 . CGAG C 132.86 . AC=1;AF=NaN;AN=2;BaseQRankSum=0.608;ClippingRankSum=0.00;DP=38;ExcessHet=3.01;FS=0.000;MLEAC=1;MLEAF=0.500;MQ=33.16;MQRankSum=-4.320e-01;QD=3.91;ReadPosRankSum=-8.820e-01;SOR=0.0
--(Unix)--- justchild.frac0.4.seed1756_trio.vcf 1% L166 (Fundamental)
```

хромосома,
координата

референсный аллель, наблюдаемый
альтернативный аллель

INFO: эту информацию пишет
программа-SNP-caller
Ее значение объясняется в заголовке

Формат VCF: данные о варианте в каждом образце (GATK HaplotypeCaller)

В этом файле три образца

```
emacs@SAMSUNG2016
File Edit Options Buffers Tools Help
[Icons]
##contig=<ID=chrUn_gl000248,length=39786,assembly=hg19>
##contig=<ID=chrUn_gl000249,length=38502,assembly=hg19>
##reference=file:///uge_mnt/home/enabieva/hg19/ucsc.hg19.fasta
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT father justchild mother
chrM 3595 . C T 2652.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=2.69;ClippingRankSum=0.00;DP=
438;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=60.00;MQRankSum=0.00;QD=30.14;ReadPosRankSum=2.09;SOR=5.670 GT
:AD:DP:GQ:PL 1/1:1:2,113:99:3727,265,0 0/0:38,0:38:99:0,102,1363 0/0:267,0:267:99:0,120,1800
chrM 4105 . INFO: инфо о локусе (все образцы) AC=2;AF=0.500;AN=6;BaseQRankSum=1.41;ClippingRankSum=0.00;DP=
350;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=36.63;MQRankSum=-1.743e+00;QD=31.96;ReadPosRankSum=-5.400e-01;SO
R=4.977 GT:AD:DP:GQ:PL 1/1:1,23:24:39:805,39,0 0/0:18,0:18:48:0,48,720 0/0:308,0:308:99:0,120,1800
chrM 4768 . A G 1724.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=1.22;ClippingRankSum=0.00;DP=
254;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=37.99;MQRankSum=0.442;QD=31.35;ReadPosRankSum=0.076;SOR=5.784 GT
:AD:DP:GQ:PL 1/1:2,53:55:94:1762,94,0 0/0:17,0:17:51:0,51,547 0/0:182,0:182:99:0,120,1800
chrM 6262 . G A 23315.17 . AC=4;AF=0.500;AN=6;BaseQRankSum=0.300;ClippingRankSum
=0.00;DP=792;ExcessHet=0.7918;FS=0.000;MLEAC=4;MLEAF=0.667;MQ=58.86;MQRankSum=3.34;QD=33.74;ReadPosRankSum=1.13;SOR=8
.662 GT:AD:DP:GQ:PL 0/0:71,0:71:99:0,120,1800 1/1:3,35:38:23:1128,23,0 1/1:5,648:653:99:22225,1825,0
chrM 7176 . T C 3689.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=1.27;ClippingRankSum=0.00;DP=
418;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=42.23;MQRankSum=-1.860e+00;QD=32.65;ReadPosRankSum=1.75;SOR=7.24
0 GT:AD:DP:GQ:PL 1/1:2,111:113:99:3727,265,0 0/0:38,0:38:99:0,102,1363 0/0:267,0:267:99:0,120,1800
chrM 7257 . C T 2916.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=-3.862e+00;ClippingRankSum=0.
00;DP=407;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=31.86;MQRankSum=-3.833e+00;QD=28.59;ReadPosRankSum=-4.140e
-01;SOR=5.974 GT:AD:DP:GQ:PL 1/1:4,98:102:99:2954,143,0 0/0:38,0:38:99:0,102,1363 0/0:267,0:267:99:0,120,
1800
chrM 7275 . C T 3464.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=-1.808e+00;ClippingRankSum=0.
00;DP=424;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=31.86;MQRankSum=-3.833e+00;QD=28.59;ReadPosRankSum=-4.140e
-01;SOR=6.753 GT:AD:DP:GQ:PL 1/1:4,98:102:99:2954,143,0 0/0:38,0:38:99:0,102,1363 0/0:267,0:267:99:0,120,
1800
chrM 7301 . C T 4628.14 . AC=2;AF=0.500;AN=6;BaseQRankSum=0.00;ClippingRankSum=0.00;DP=
452;ExcessHet=0.7918;FS=0.000;MLEAC=2;MLEAF=0.333;MQ=34.50;MQRankSum=-1.925e+00;QD=31.70;ReadPosRankSum=-7.920e-01;S
OR=5.974 GT:AD:DP:GQ:PL 1/1:4,98:102:99:2954,143,0 0/0:38,0:38:99:0,102,1363 0/0:267,0:267:99:0,120,
1800
-(Unix)--- justchild.frac0.4.seed1756_trio.vcf 1% L127 (Fundamental)
```

INFO: инфо о локусе (все образцы)

Инфо о данном сайте в первом образце

Инфо о данном сайте во втором образце

Инфо о данном сайте в третьем образце

Заголовок/header

Заголовок/header

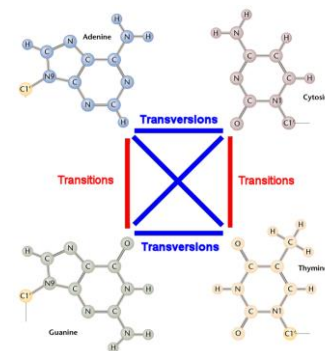
```
chr1      69569      .      T      C      259.89      .  
#6;ExcessHet=3.0103;FS=0.000;MLEAC=1;MLEAF=0.250;MQ=30.  
#      GT:AD:DP:GQ:PL      0/0:32,0:32:81:0,81,1215  
chr1      135134      .      G      A      14.82      .  
#EAF=0.167;MQ=21.42;QD=7.41;SOR=2.303      GT:AD:DP:GQ:PH  
#      1/1:0,2:2:6:49,6,0  
chr1      135301      .      G      A      15.86      .  
#EAF=0.167;MQ=20.00;QD=7.93;SOR=2.303      GT:AD:DP:GQ:PH  
#0:2,0:2:6:0,6,66  
chr1      135374      .      CGAG      C      13.86      .
```

GT = 0/0 : сайт гомозиготен по референсному аллелю
AD = 32, 0 : 32 чтения с реф. аллелем., 0 с альт. Аллелем
DP = 32 : общая глубина 32 чтения (с учетом фильтрации)
GQ = 81 : Genotype Quality
PL = 0, 81, 1215 – нормализованное PHRED-шкалированное
правдоподобие генотипов 0/0, 0/1, 1/1

Специфика GATK HaplotypeCaller

Контроль качества

- Количество отличий от референса
 - Оценка от 1000 Геномов (Nature 467, (28 October 2010)) – в кодирующих последовательностях
 - 10,000–11,000 несинонимичных
 - 10,000–12,000 синонимичных
 - 80–100 nonsense
 - см. след. слайд
- % новых полиморфизмов (сравнение с базой данных dbSNP)
 - Например, используя Annovar (см. позже)
 - Должно быть мало
 - см. след. слайд
- ts/tv (aka ti/tv: transition/transversion)
 - в геноме человека: ~2
 - в кодирующих последовательностях человека: ~3
 - Как считать: программа: SnpSift tstv
<http://snpeff.sourceforge.net/SnpSift.html>
(Также: GATK VariantEval
после использования samtools/bcftools: также `vcfutils.pl qstats`)



Среднее количество кодирующих вариантов в двух популяциях

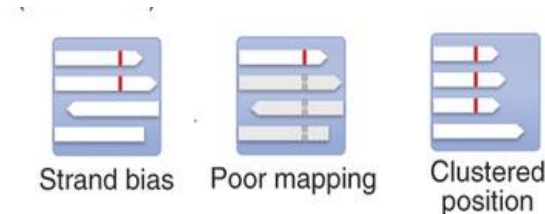
| Variant type | Mean number of variants (\pm sd) in African Americans | Mean number of variants (\pm sd) in European Americans |
|---------------------------|---|--|
| <i>Novel variants</i> | | |
| Missense | 303 (\pm 32) | 192 (\pm 21) |
| Nonsense | 5 (\pm 2) | 5 (\pm 2) |
| Synonymous | 209 (\pm 26) | 109 (\pm 16) |
| Splice | 2 (\pm 1) | 2 (\pm 1) |
| Total | 520 (\pm 53) | 307 (\pm 33) |
| <i>Non-novel variants</i> | | |
| Missense | 10,828 (\pm 342) | 9,319 (\pm 233) |
| Nonsense | 98 (\pm 8) | 89 (\pm 6) |
| Synonymous | 12,567 (\pm 416) | 10,536 (\pm 280) |
| Splice | 36 (\pm 4) | 32 (\pm 3) |
| Total | 23,529 (\pm 751) | 19,976 (\pm 505) |
| <i>Total variants</i> | | |
| Missense | 11,131 (\pm 364) | 9,511 (\pm 244) |
| Nonsense | 103 (\pm 8) | 93 (\pm 6) |
| Synonymous | 12,776 (\pm 434) | 10,645 (\pm 286) |
| Splice | 38 (\pm 5) | 34 (\pm 4) |
| Total | 24,049 (\pm 791) | 20,283 (\pm 523) |

The table lists the mean number (\pm standard deviation (sd)) of novel and non-novel coding single nucleotide variants from 100 sampled African Americans and 100 European Americans. Non-novel variants refer to those found in dbSNP131 or in 200 other control exomes. Capture was performed using the Nimblegen V2 target. The analysis pipeline consisted of: alignment using the Burrows–Wheeler alignment tool; recalibration;

Bamshad et al. Nat. Rev. Genet. 2011

Фильтрация вариантов

- Жесткая фильтрация
 - Пороги для разных значений
 - Качество варианта (согласно оценке от , количество и/или MAPQ поддерживающих чтений, нахождение на одной цепи, нахождение близко к концам чтения, и т.д.



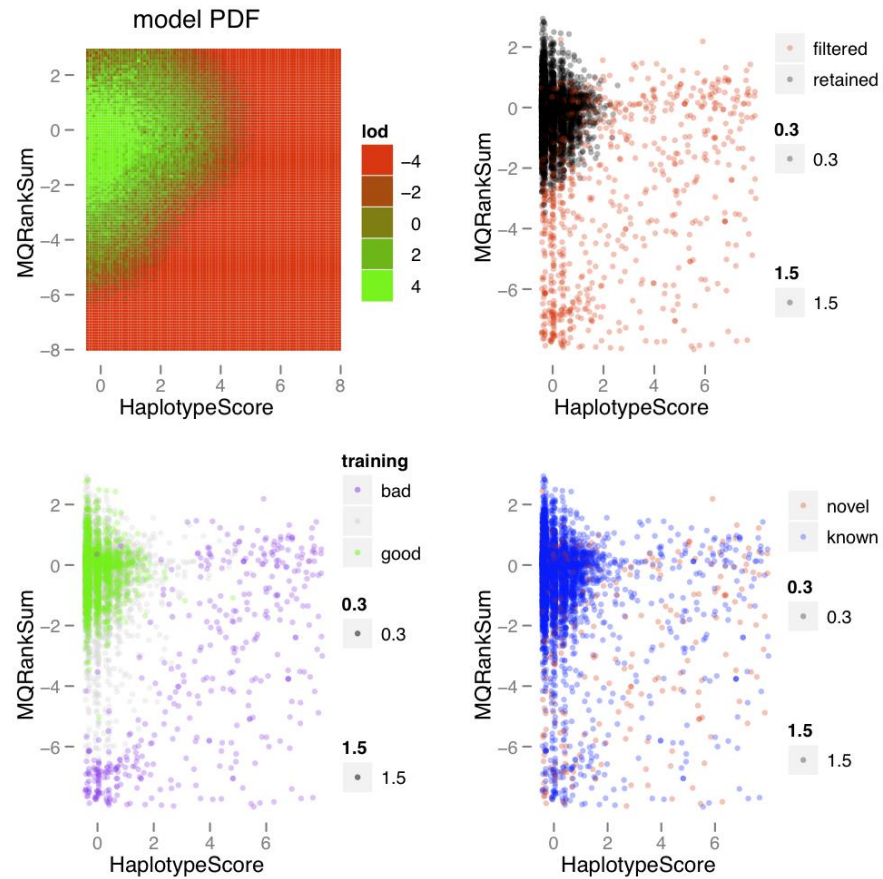
Иллюстрации: MuTest

- например:
<https://software.broadinstitute.org/gatk/documentation/article.php?id=2806>

Фильтрация вариантов

- «Умная» фильтрация от GATK
 - Много данных: GATK Variant Quality Score Recalibration
 - популяционные варианты, в которых мы уверены (1000G) => вероятностная модель зависимости от разных характеристик вместе
 - найденные варианты сравниваются с моделью
 - Нужно достаточно точек для построения модели: геном или ≥ 30 экзонов
 - <https://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr>

GATK VQSR



Фильтрация вариатов

- «Умная» фильтрация от GATK (GATK4)
 - Один образец: CNNScoreVariants
 - Использует сверточные нейронные сети!



Выявление значимых мутаций

Фильтрация вариантов

Информация о варианте:

- «семантика» варианта в последовательности
- популяционная информация
- предсказательные модели

Информация о гене:

- pLI/s-het

Биология в широком смысле:

базы данных аннотаций, модельные организмы,
функциональная информация...

Эффект на уровне генетического кода

- Довольно точно определяется с помощью аннотации генома
- Вредные мутации:
 - Новый стоп-кодон (nonsense) – вредно
 - Инсерция/делеция со смещением рамки
 - Мутация в сайте сплайсинга
 - и др.
- "Средне-вредные" мутации:
 - Несинонимичные замены (missense, меняют аминокислоту) – см. следующий шаг
- Вероятно, безвредные мутации:
 - синонимичные замены
 - некодирующие замены (?)
- Мутации в функциональных некодирующих областях:
 - можно охарактеризовать, имея знания соответствующей области.
- Программы: SnpEff <http://snpeff.sourceforge.net/SnpEff.html> и другие

Классификация полиморфизмов от SnpEff

| Impact | Effects |
|----------|---|
| High | SPLICE_SITE_ACCEPTOR SPLICE_SITE_DONOR START_LOST EXON_DELETED FRAME_SHIFT STOP_GAINED STOP_LOST RARE_AMINO_ACID |
| Moderate | NON_SYNONYMOUS_CODING CODON_CHANGE CODON_INSERTION CODON_CHANGE_PLUS_CODON_INSERTION CODON_DELETION CODON_CHANGE_PLUS_CODON_DELETION UTR_5_DELETED UTR_3_DELETED |

| Impact | Effects |
|----------|---|
| Low | SYNONYMOUS_START NON_SYNONYMOUS_START START_GAINED SYNONYMOUS_CODING SYNONYMOUS_STOP SPLICE_SITE_REGION |
| Modifier | UTR_5_PRIME UTR_3_PRIME REGULATION UPSTREAM DOWNSTREAM GENE TRANSCRIPT EXON INTRON_CONSERVED INTRON INTRAGENIC INTERGENIC INTERGENIC_CONSERVED NONE CHROMOSOME CUSTOM CDS |

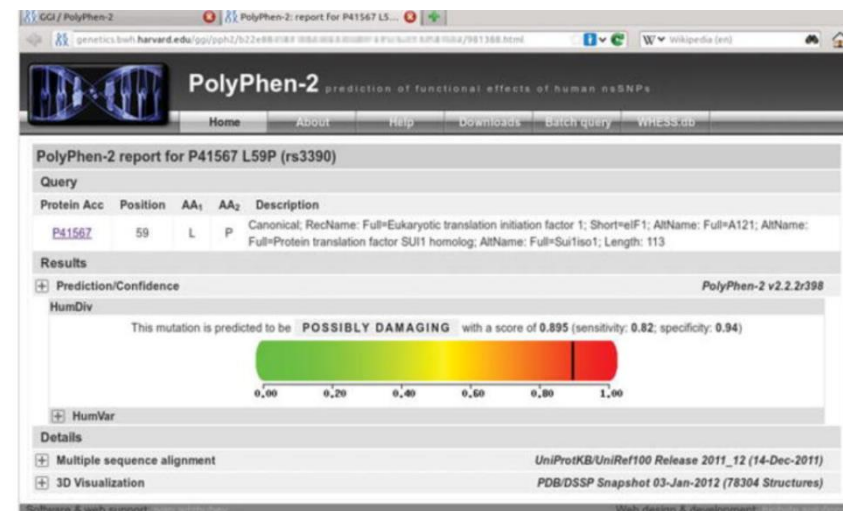
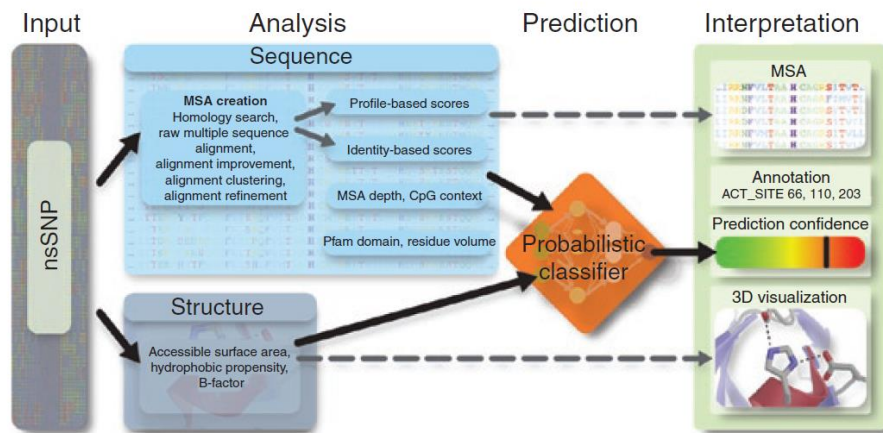
Предсказание вредности вариантов

- Информация:
 - Эволюционная консервативность
 - Эффект на структуру белка (для кодирующих вариантов)
 - Свойства последовательности
 -
 - Некоторые программы учитывают «важность» белка
- Пример: Polyphen2
- Мета-классификаторы (например, REVEL, CADD...)
- dbNSFP: агрегатор предсказаний большого числа классификаторов

Пример: PolyPhen2

Adzhubei et al. 2010

- Предсказание эффекта несинонимичных замен
- Наивный байесовский метод
- Признаки основаны на
 - Эволюционной консервативности
 - Белковой структуре



PolyPhen2: training sets

- HumDiv:
 - Damaging: 3,155 аллелей, вызывающих менделевские заболевания в человеке и влияющих на стабильность или функцию белка
 - Neutral: 6,321 различий между белками человека и близких гомологов в млекопитающих
- HumVar:
 - Damaging: 13,032 мутаций, вызывающих заболевания в человеке
 - Neutral: 8,946 несинонимичных замен без известного участия в заболеваниях

Популяционная информация

- Эволюция против вредных вариантов!
- Можно оценить предполагаемый порог частоты искомого варианта по информации о частоте фенотипа, его наследовании и генетической архитектуре
- gnomAD: 125,748 экзонов и 15,708 геномов

Важность гена

- pLI: “probability of loss-of-function intolerance”
 - \leq сравнение наблюдаемого и ожидаемого количества protein-truncating/missense вариантов

Article | Published: 03 August 2014

A framework for the interpretation of *de novo* mutation in human disease

Kaitlin E Samocha, Elise B Robinson, [...] Mark J Daly 

Nature Genetics **46**, 944–950(2014) | [Cite this article](#)

2852 Accesses | 410 Citations | 72 Altmetric | [Metrics](#)

Article | [Open Access](#) | Published: 17 August 2016

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek, Konrad J. Karczewski, [...] Exome Aggregation Consortium

Nature **536**, 285–291(2016) | [Cite this article](#)

42k Accesses | 3905 Citations | 939 Altmetric | [Metrics](#)

- S_{het} : популяционно-эволюционно-генетическая оценка отбора против protein-truncating вариантов в белке

Letter | Published: 03 April 2017

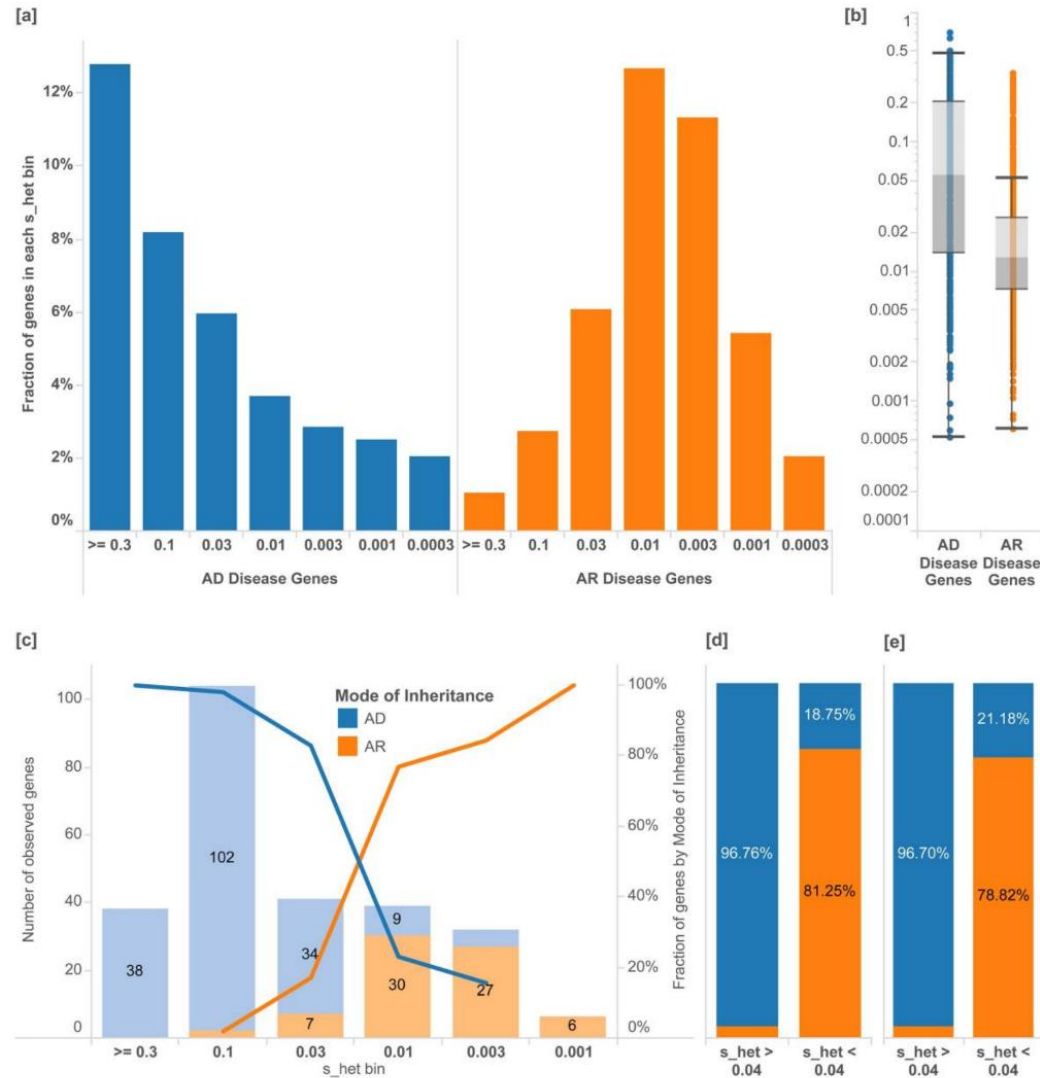
Estimating the selective effects of heterozygous protein-truncating variants from human exome data

Christopher A Cassa, Donate Weghorn, Daniel J Balick, Daniel M Jordan, David Nusinow, Kaitlin E Samocha, Anne O'Donnell-Luria, Daniel G MacArthur, Mark J Daly, David R Beier  & Shamil R Sunyaev 

Nature Genetics **49**, 806–810(2017) | [Cite this article](#)

1234 Accesses | 26 Citations | 56 Altmetric | [Metrics](#)

s_het



Cassa et al.

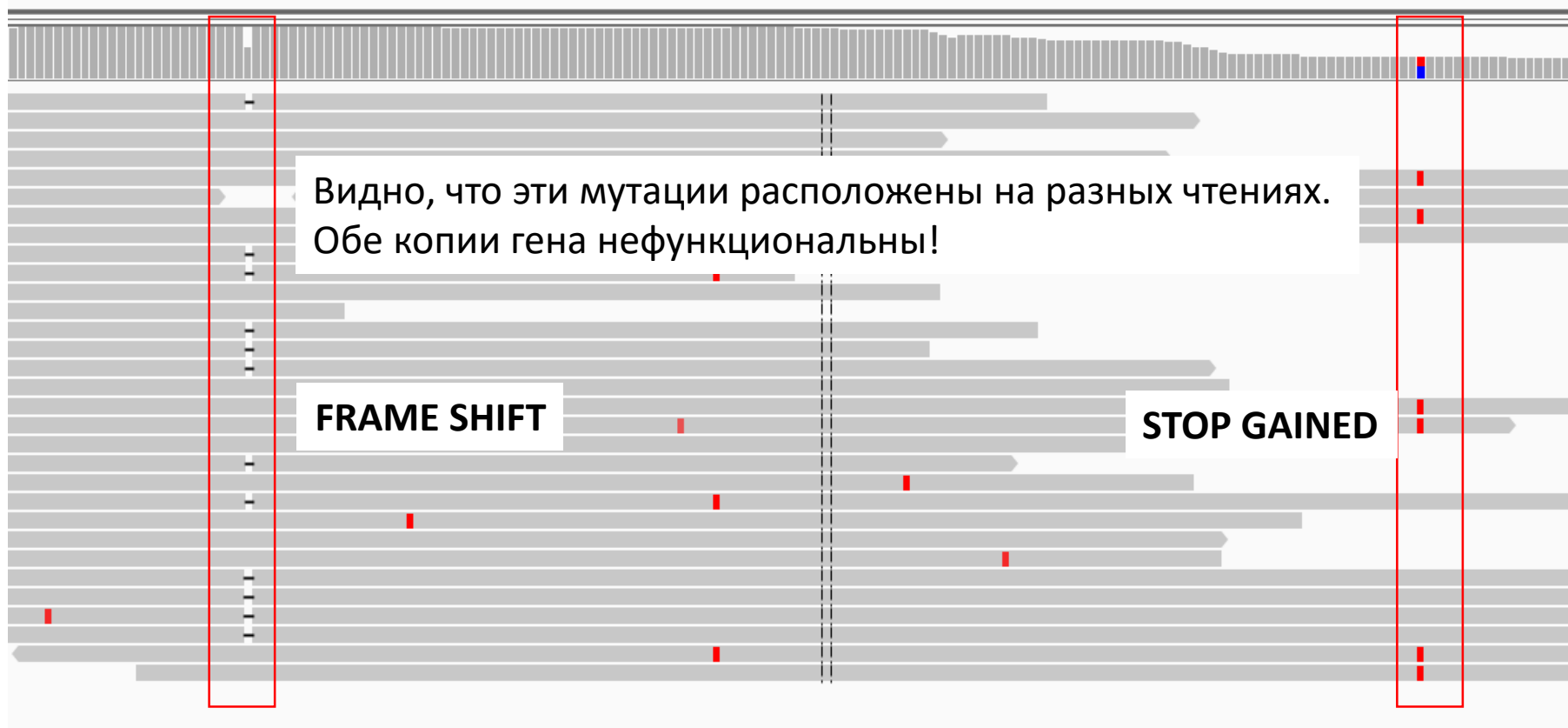
Базы данных биологических знаний

- ClinVar
 - Оценка вредности + свидетельств о ней
- OMIM
 - Генетика заболеваний
- Мышиные модели
 - Mammalian Phenotype Ontology
 - http://www.informatics.jax.org/vocab/mp_ontology

Инструменты для аннотации вариантов

- Annovar
 - GEMINI
 - GATK Funcotator
-
- Онлайн: Ensembl VEP

Одна история: причина наследственного заболевания сетчатки



ген, кодирующий фоторецепторный белок

Мысли в конце

- Стандартные форматы файлов позволяют строить *pipelines* для анализа
 - <https://bcbio-nextgen.readthedocs.io/>
 - GATK WDL
 - Snakemake
- Инструменты быстро развиваются
- Важна проверка качества на разных этапах
- NGS + биоинформатика => много информации
 - Как интерпретировать?
 - Развивать: технологии и алгоритмы
 - Интеграция данных