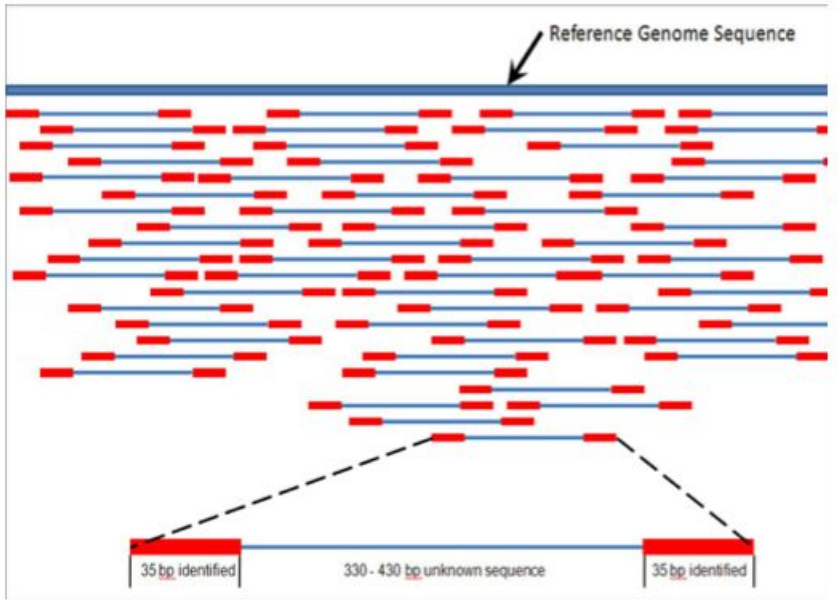
NGS data analysis: from FASTQ to VCF

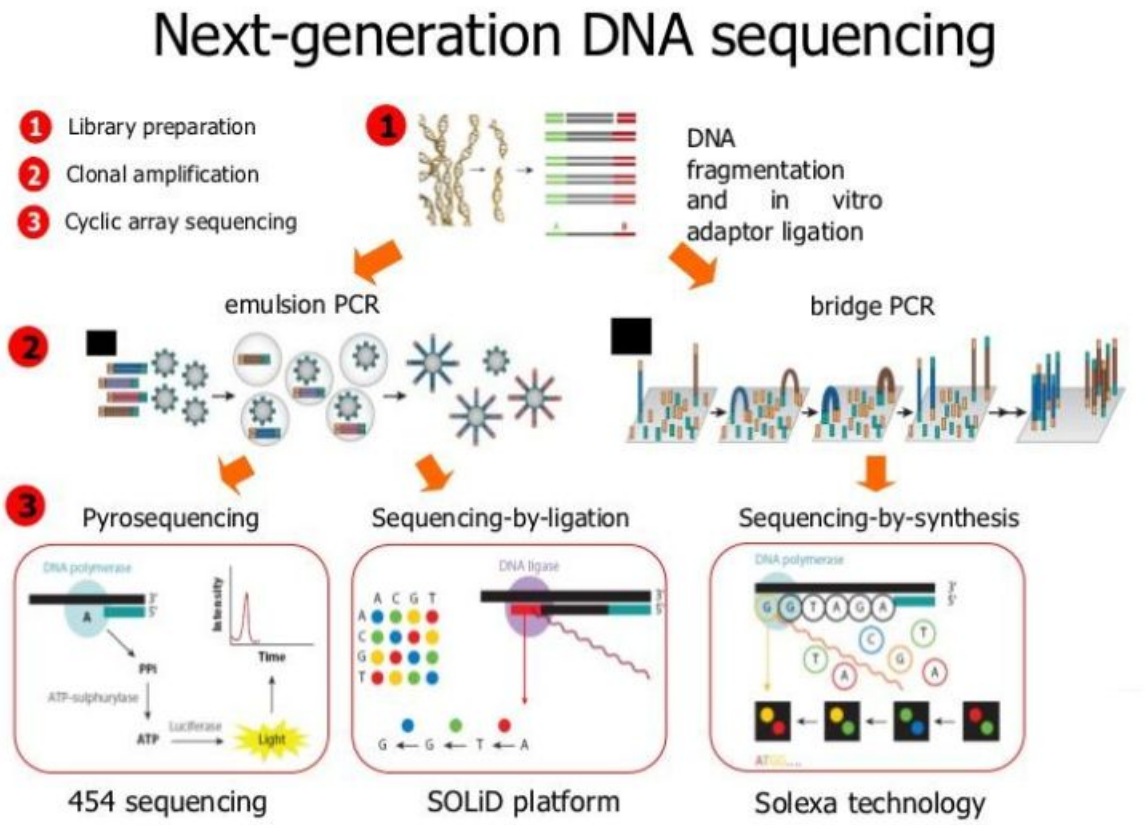
Биоинформатика и наука - NGS

- Genome assembly - сборка геномов немодельных организмов
- Resequencing - таргетное секвенирование, экзомы, геномы (как правило, пациентов)
- RNA-seq - характеризация транскриптома (зависит от клеточного типа)
- Transcriptome assembly - если геном слишком неудобный (важные растения, рыбы)
- Chip-seq - кто где сидит на ДНК (тоже зависит от клеточного типа)



Биоинформатика и индустрия

- Практически работа биоинформатика почти всегда сводится к NGS
- Работа с NGS почти всегда сводится к quality control (QC)
- Ну, и еще стандартные пайплайны



Подготовка

Скачивание генома
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/
Индексация
bwa index

Прочтения **.fastq** FastQC

bwa mem

Выравнивание **.sam** **.bam** **.bai**

NGSrich или qualimap или BamQC

samtools mpileup или
GATK UnifiedGenotyper

Поиск вариантов **.vcf**

Annovar или
SNPEff или VEP

Аннотация **.txt**

Аннотированный VCF файл

4. Фильтрация данных для выявления потенциального патогенного варианта

- Отталкиваясь от клинической информации, в том числе
  - по типу наследования:
    - AD – поиск гетерозигот.
    - AR – поиск гомозигот.
    - XL – поиск генов на X хромосоме.
    - …
  - Анализ функций генов
    - Проверка известного гена
    - Анализ группы генов, описанных для этого заболевания (OMIM)
- По частотам
  - Выбор геномных вариантов с низкими частотами, учитывая:
    - Локальные популяционные частоты
    - Случаи, когда патогенный (редкий) аллель может быть в референсе
- По скорам патогенности, включающим:
  - Консервативность позиции
  - Эффект-предикторы
  - Тип мутации (LoF, синонимичная, несинонимичная, сплайсинг, ..)

Кандидатные варианты

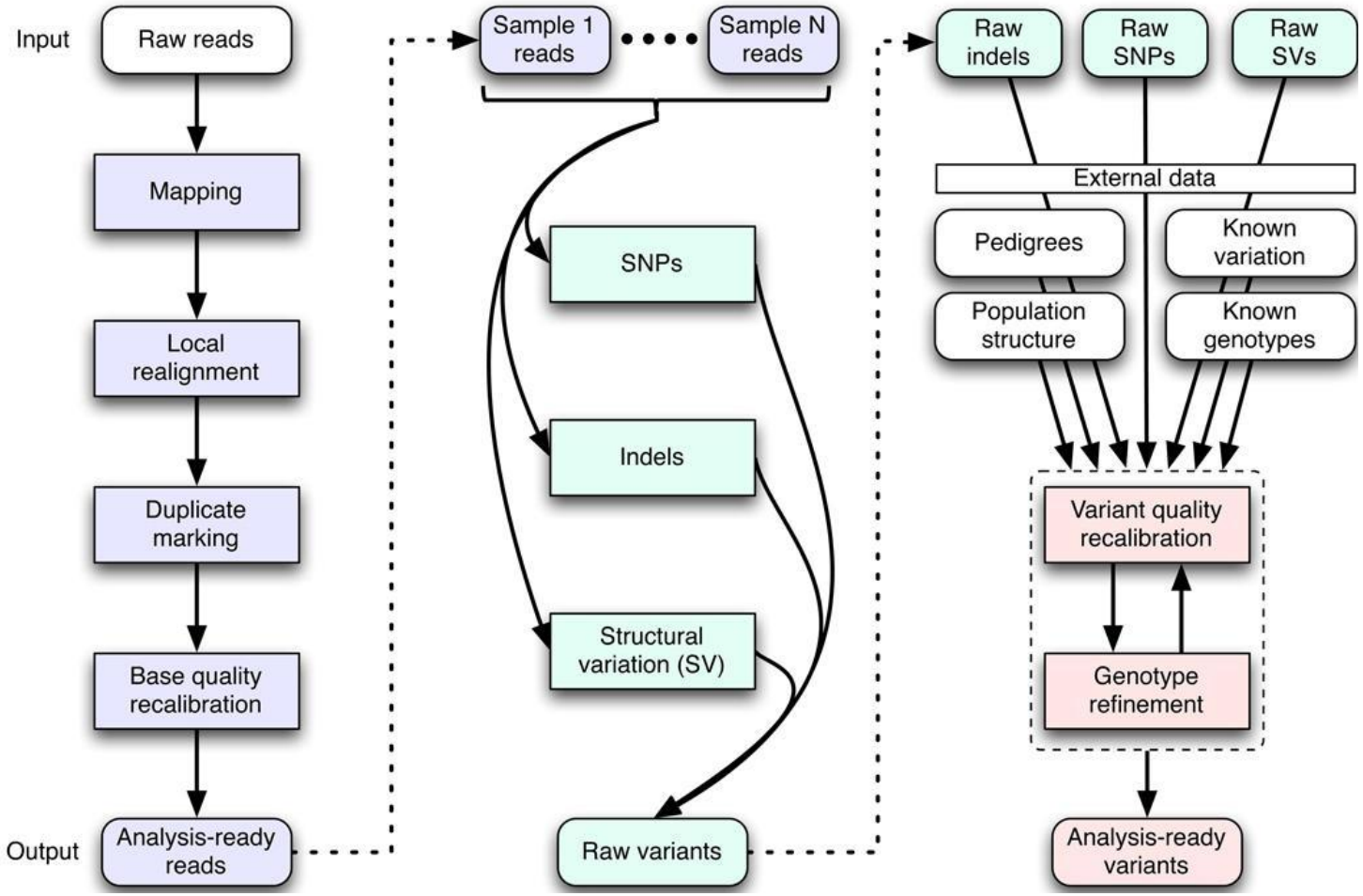
5. Детальный анализ кандидатных вариантов

- IGV browser
  - Strand biases
  - Гомополимеры
  - Local coverage
  - Unambiguous mapping
- BLAST/BLAT
  - Анализ гомологичных участков
  - Уникальность, в том числе псевдогены, паралоги, протяженные дупликации
  - Повторы
- Сбор информации о гене
  - Экспрессия в различных тканях (GTE browser)
  - Литературный анализ
  - Информация по модельным объектам

Отобранные варианты

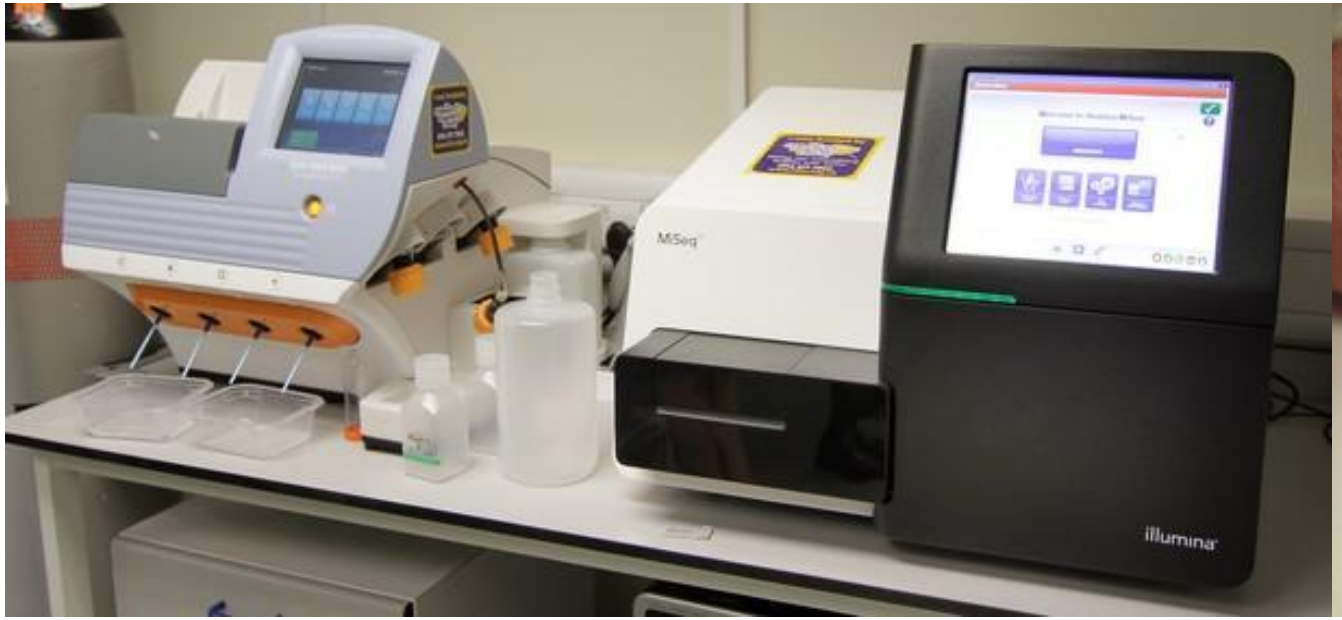
6. Формирование заключения (с резюме).

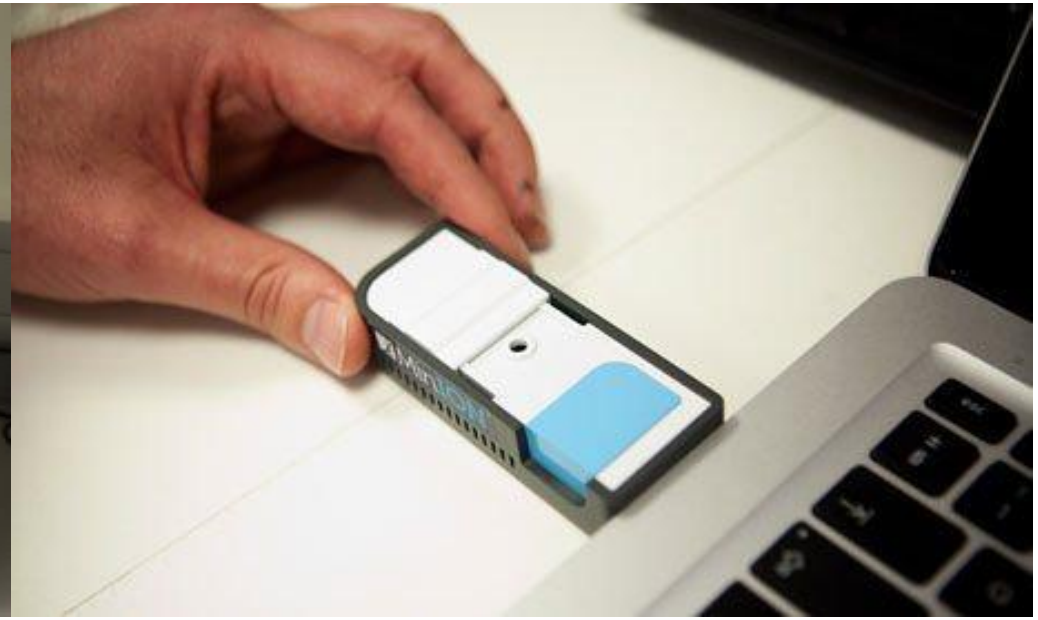
Стандарт от создателей GATK



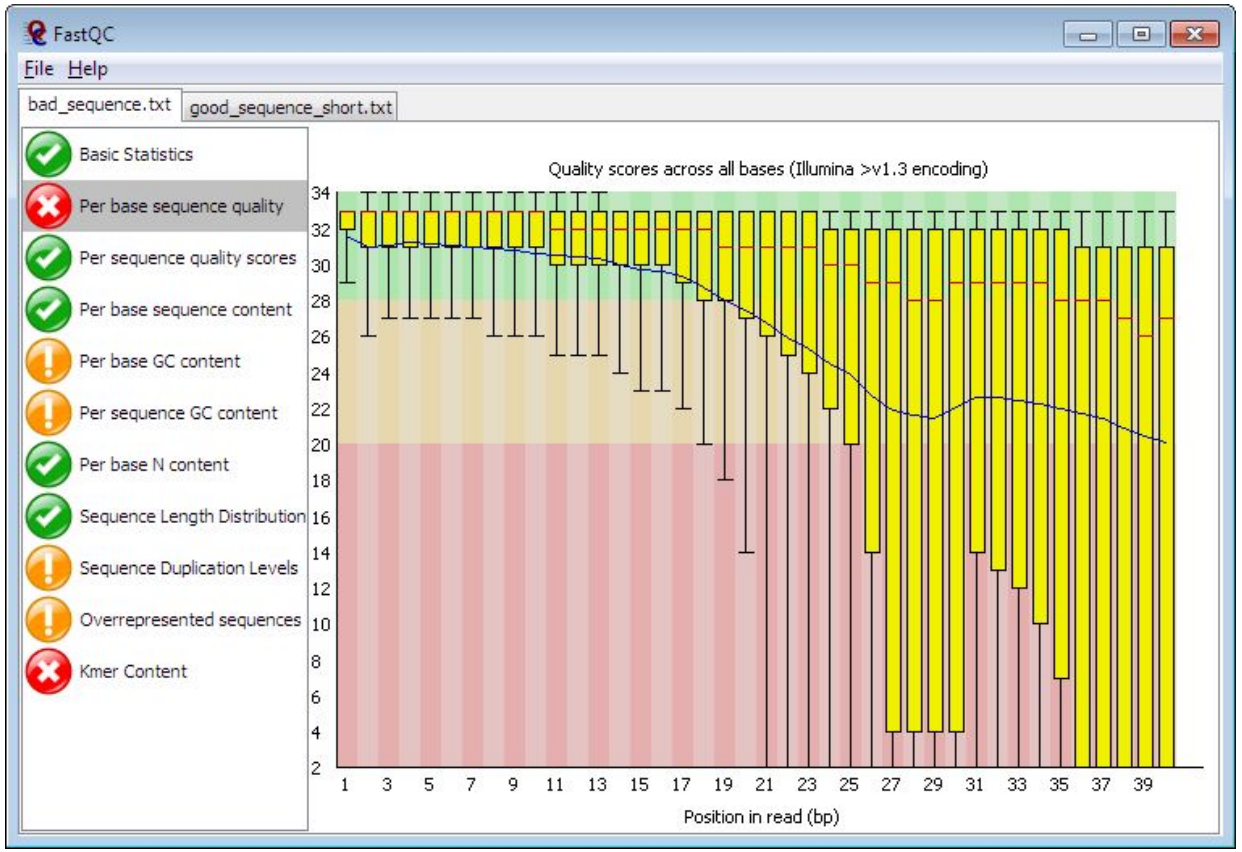
DePristo, Mark A., et al. "A framework for variation discovery and genotyping using next-generation DNA sequencing data." *Nature genetics* 43.5 (2011): 491-498.

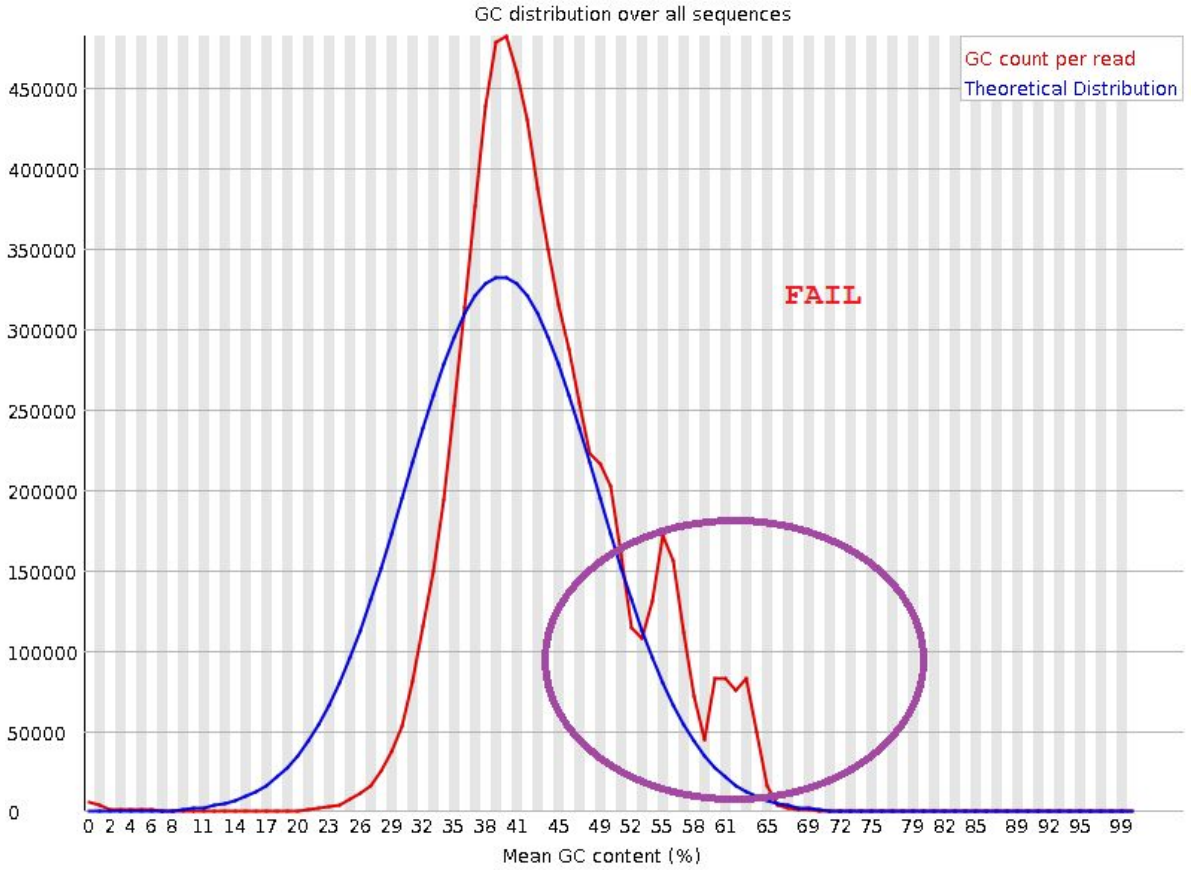
1. Получение данных (Fastq)





2. Контроль качества - FastQC (http://www.bioinformatics. babraham.ac.uk/projects/fastqc/)





2. Контроль качества - Что такое Q-score?

$$Q = -10 \log_{10} P \quad \text{Или} \quad P = 10^{-Q/10}$$

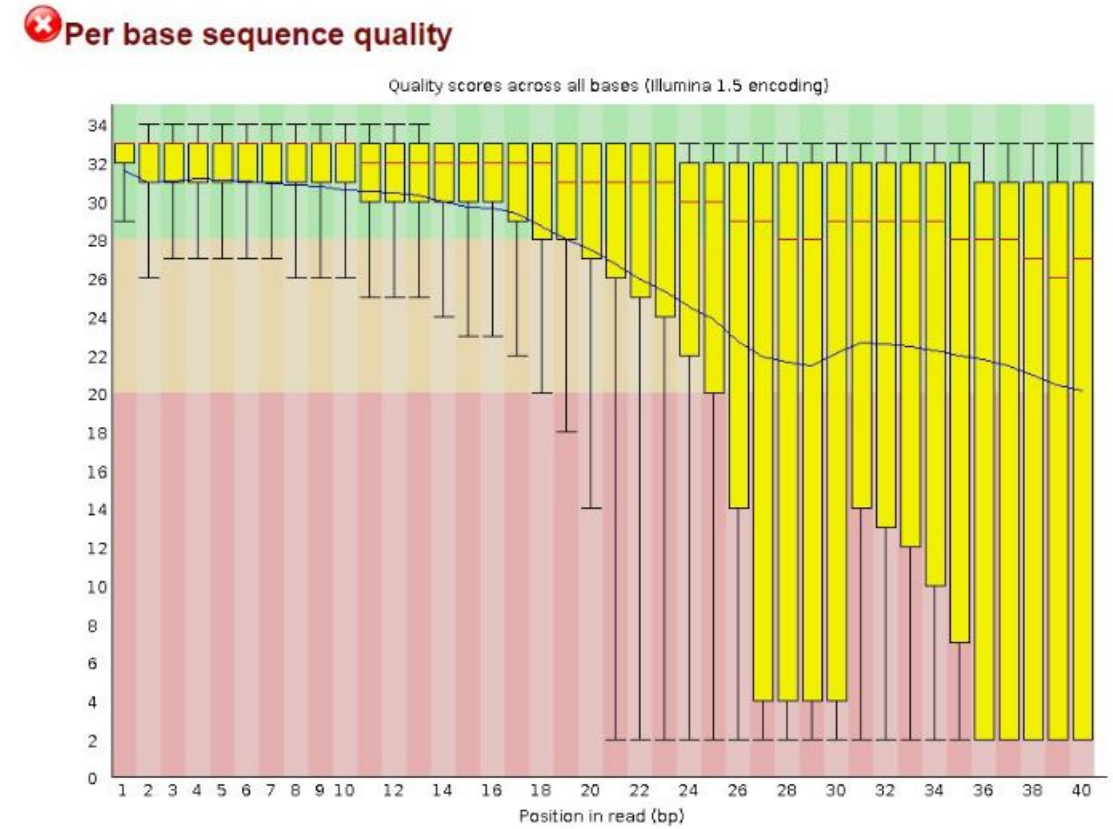
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Примеры жизненных неудач

- QCFail: https://sequencing.qcfail.com/

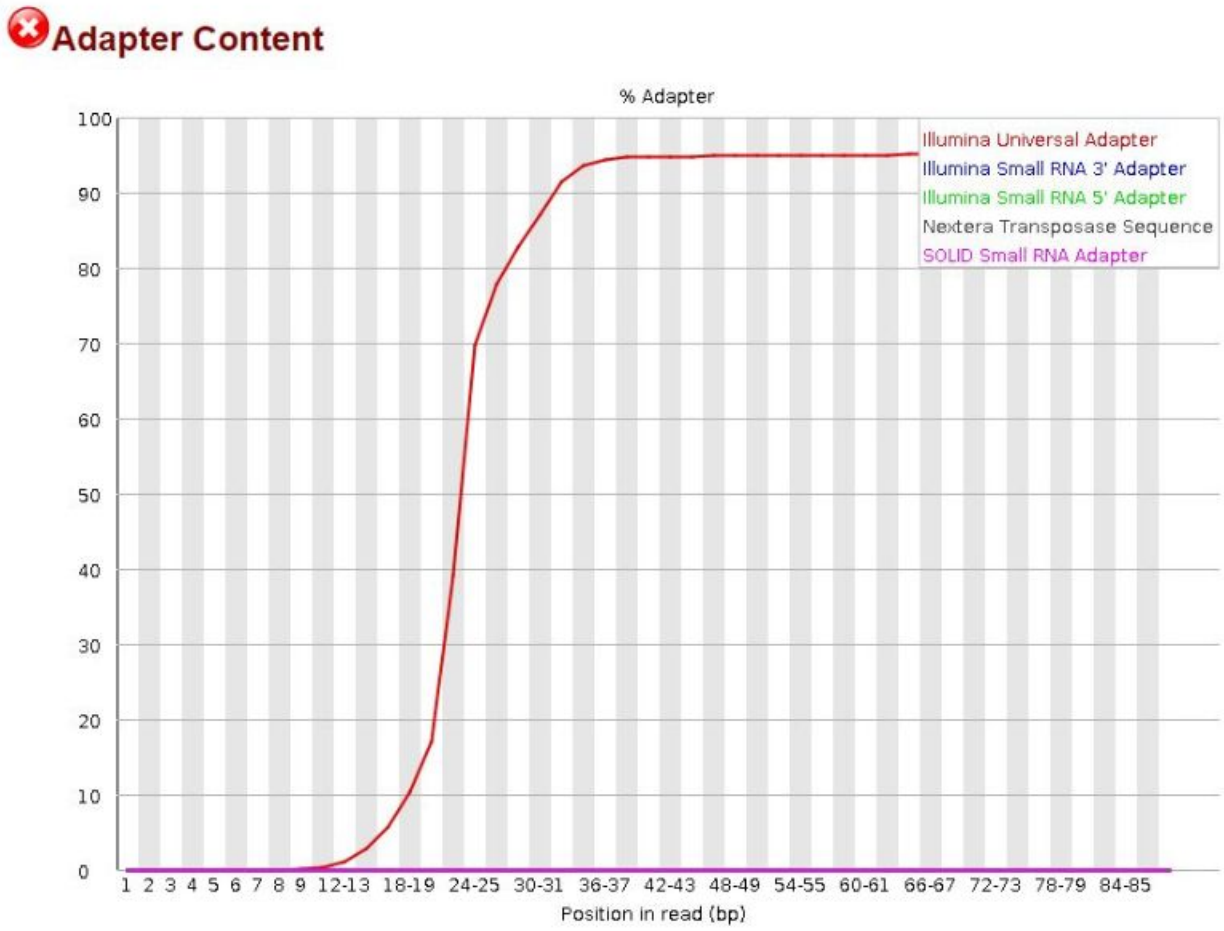
Пример 1

- Плохое секвенирование - Illumina



Пример 2

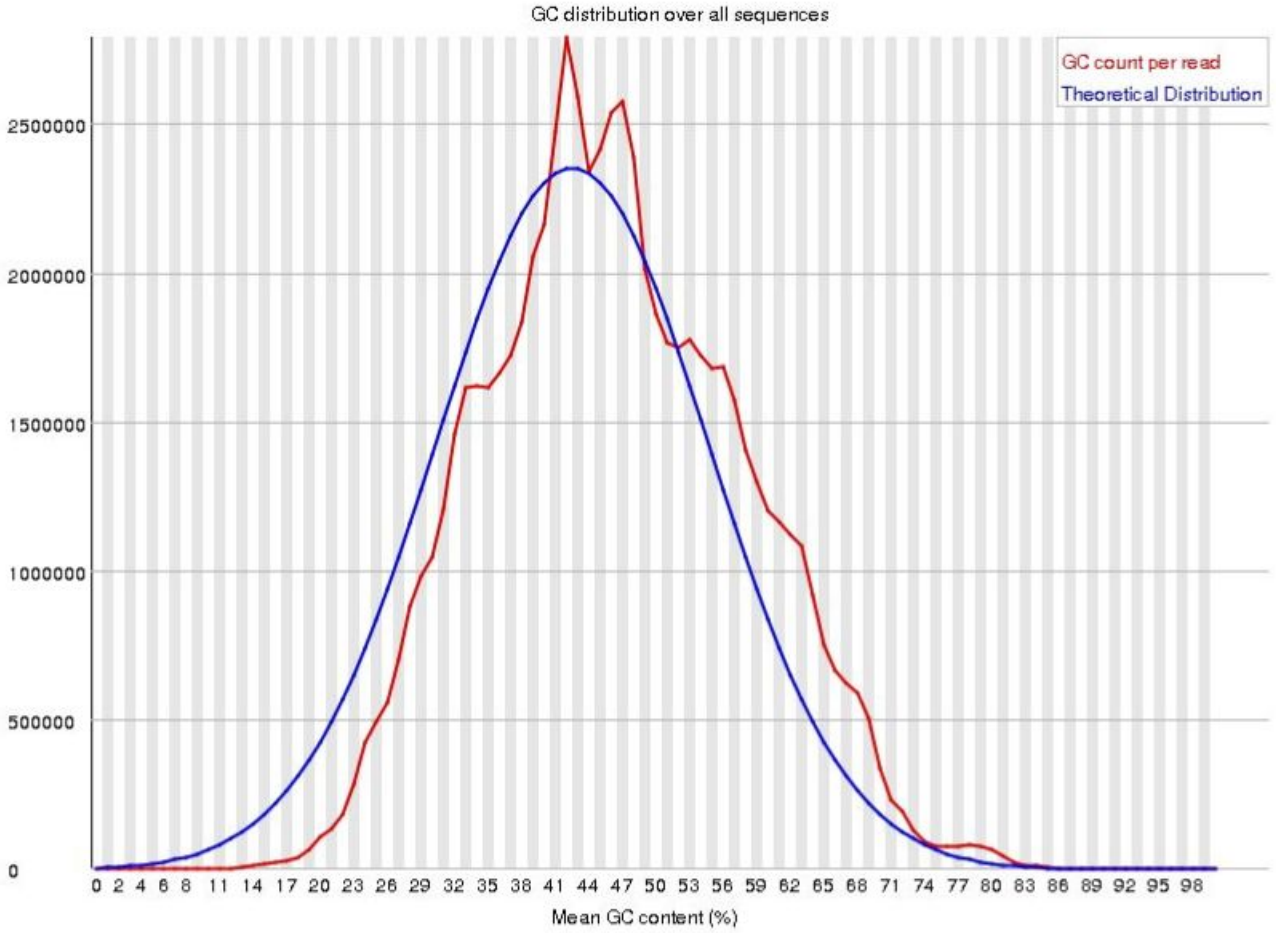
- Слишком короткие фрагменты
- Секвенирование доходит до конца и идет в адаптер



Пример 3

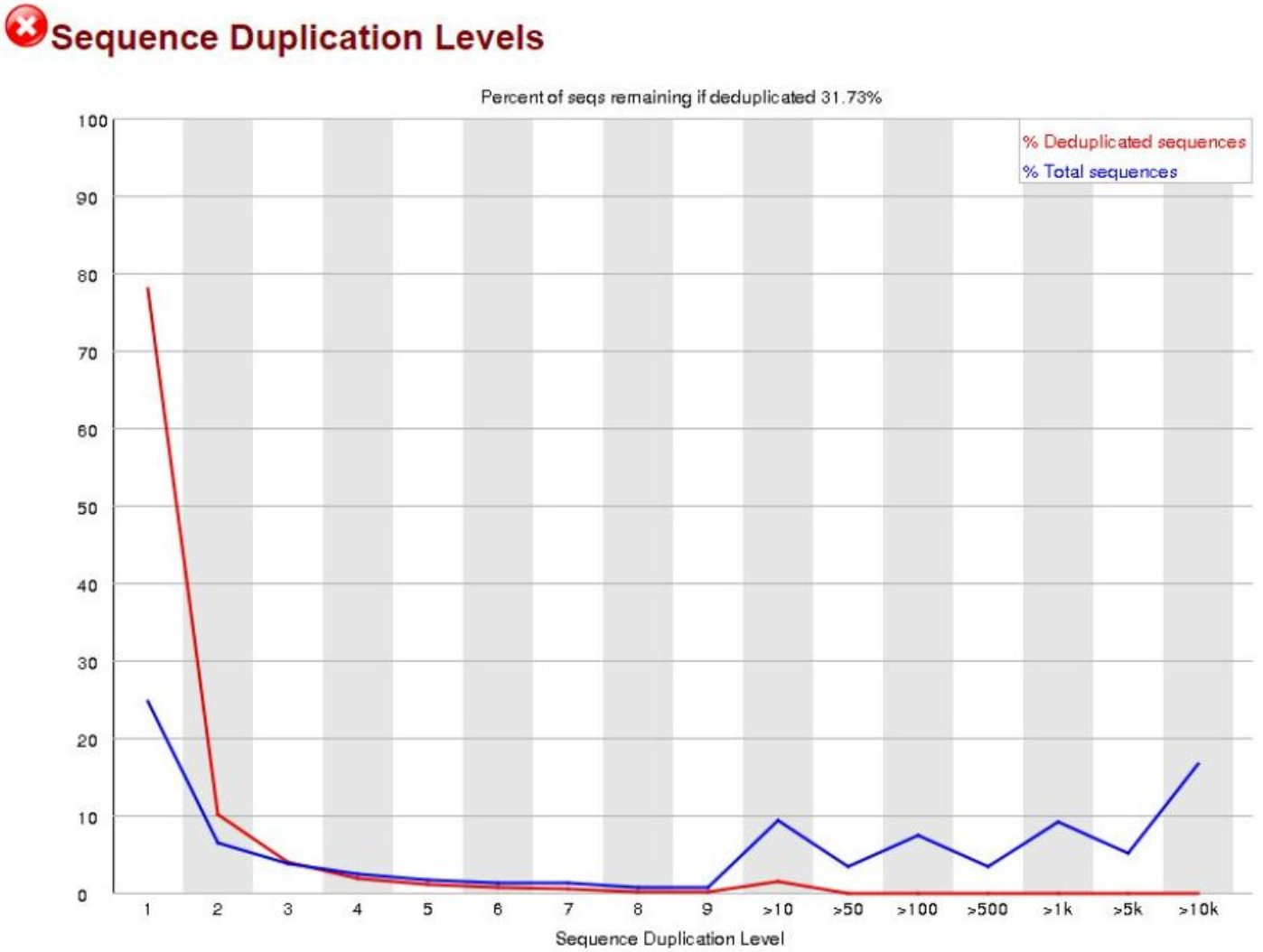
- Загрязнение бактериальным геномом

Per sequence GC content



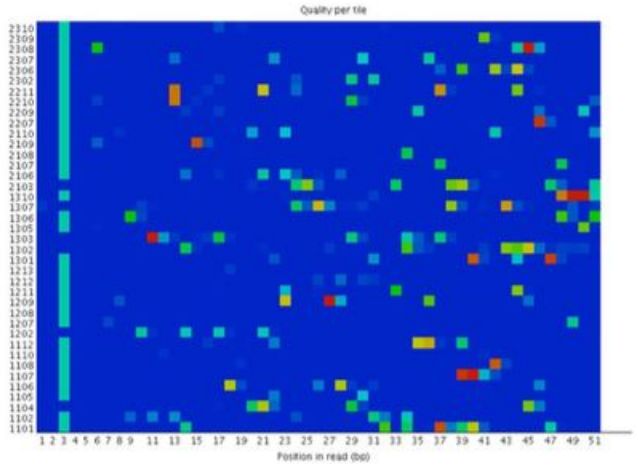
Пример 3

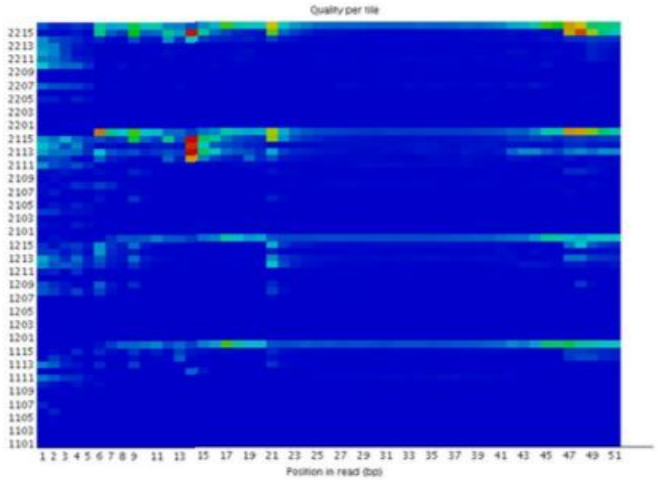
- Много повторяющихся последователь-ностей

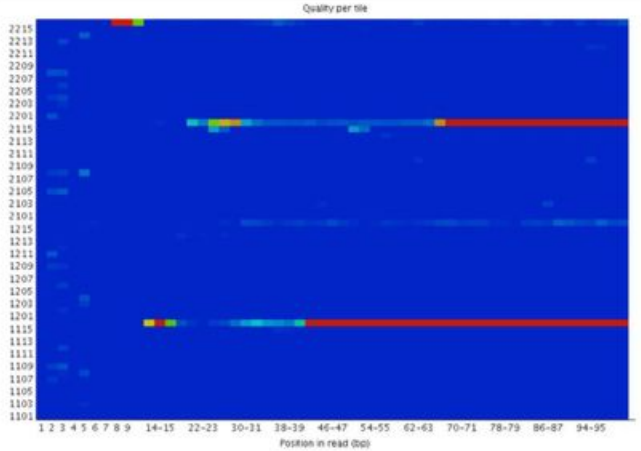


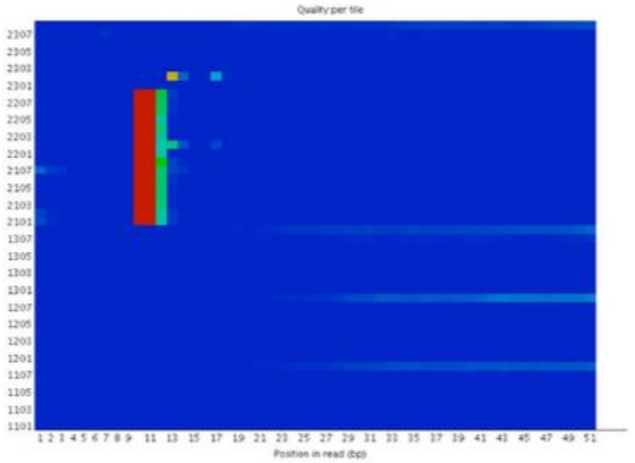
Пример 4

- Перегрузка
- Повторы
- Грязь на читающем элементе
- Пузырек





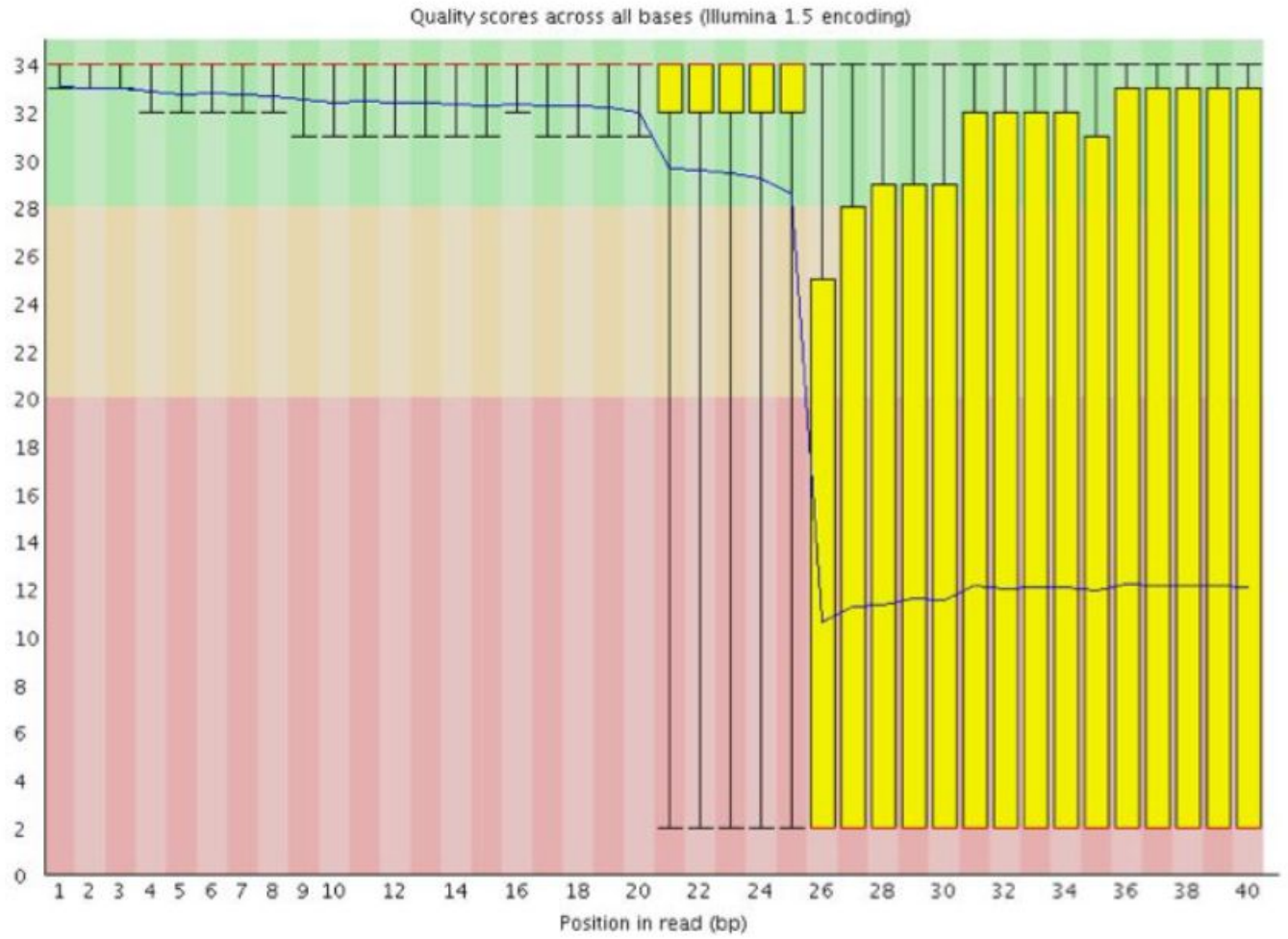




Пример 5

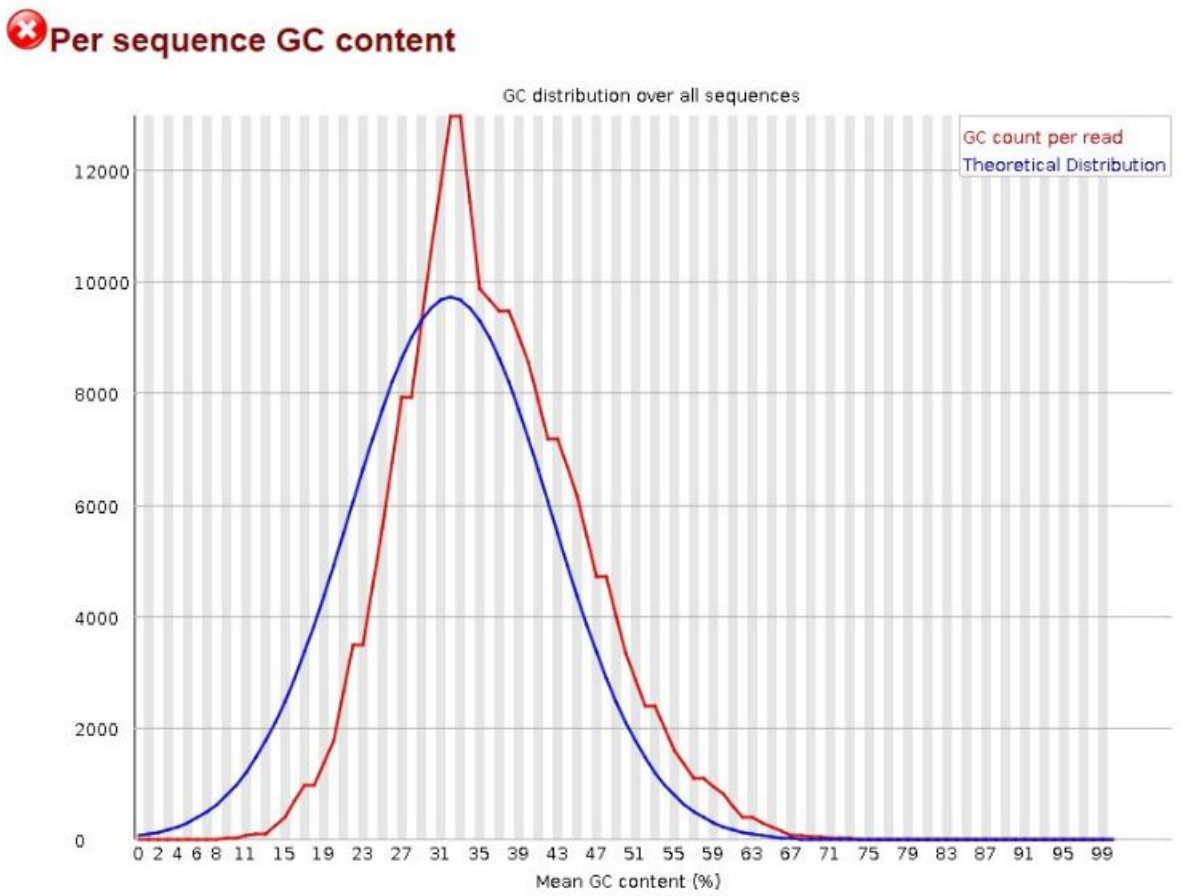
- NextSeq -

Ранние версии химии



Кастомизация FastQC

- Можно добавить свои последовательности адаптеров https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/
- Можно менять параметры графиков, биннинг и тд

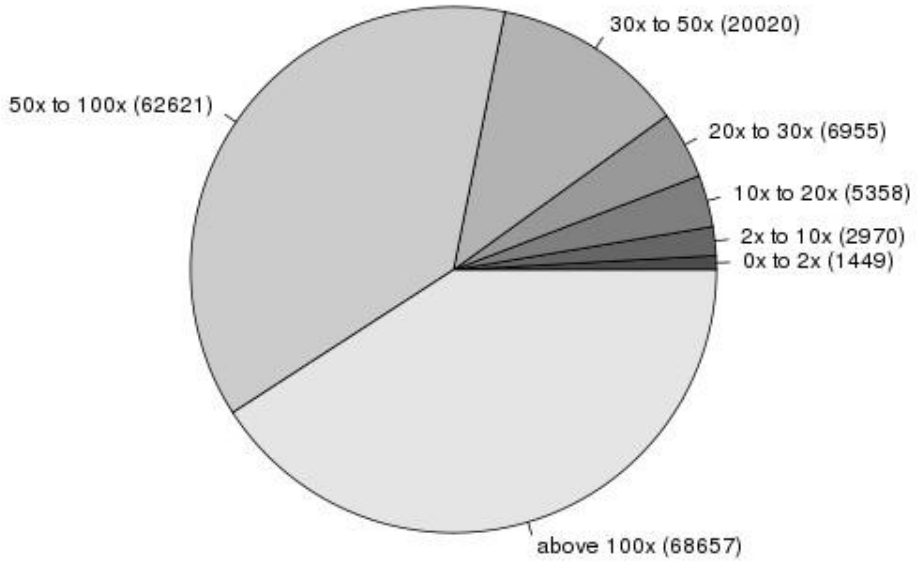


NGSrich

https://sourceforge.net/projects/ngsrich/

Summary Statistics

# Reads	34200378
# On Target ± 100 bp	32254979
Target Size (bp)	57911140
# Target Regions	168030
Coverage Mean	98.91
Coverage Std Dev	59.19
Covered 1x	98.99%
Covered 5x	97.58%
Covered 10x	96.03%
Covered 20x	92.02%
Covered 30x	86.92%
TPKM	16.29



Тримминг - удаление ошибок секвенирования

Как можно удалить ошибки секвенирования?

Тримминг (англ. trim – приводить в порядок) – основная операция для того, чтобы ваши риды стали лучше.

Две части тримминга:

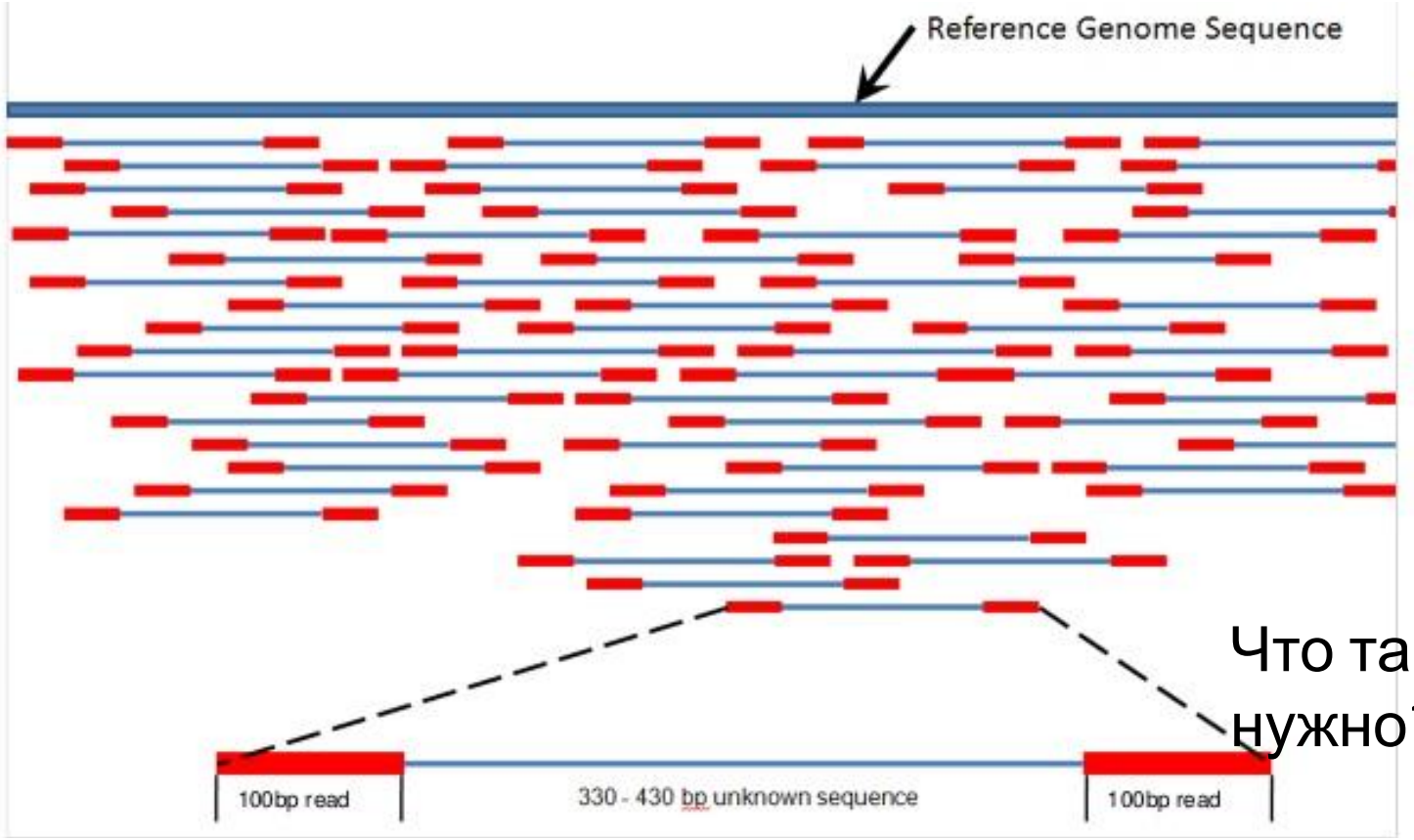
Удаление адаптеров в ридах

Отсечение с конца ридов нуклеотидов с качеством ниже данного (например, <20)

Триминг – **необходимая** процедура. Она всегда облегчит вам последующий анализ. А часто анализ ридов без тримминга и вовсе невозможен

Для тримминга рекомендуется программа Trimmomatic

3. Выравнивание ридов на геном – Как это выглядит



Нужно:

1. Риды – FastQ
2. Референс / Индекс

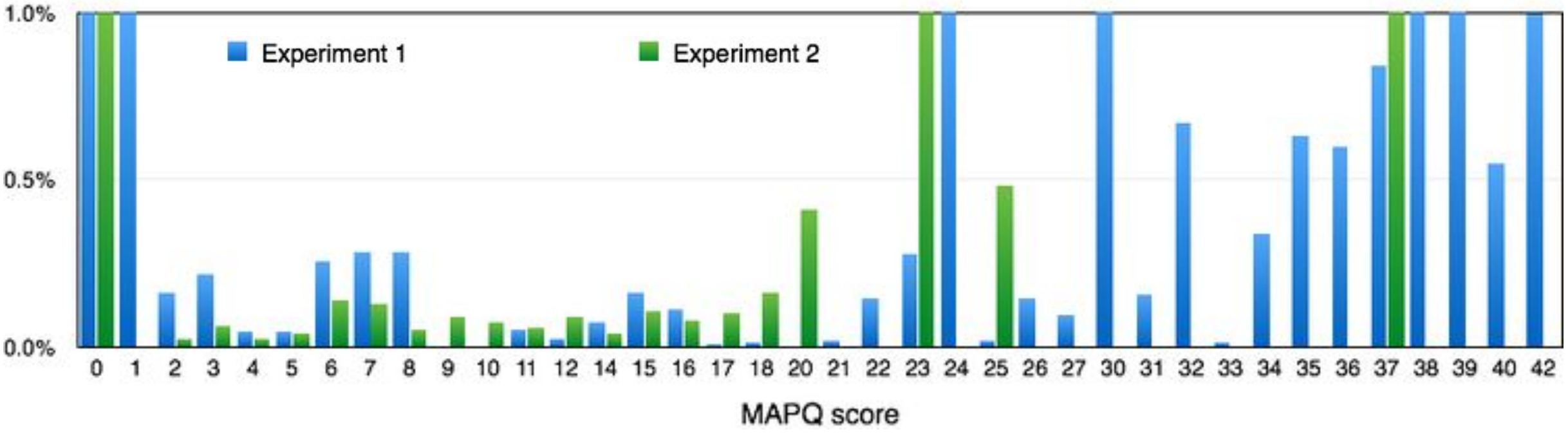
Что такое покрытие и сколько его нужно?

Выравнивание

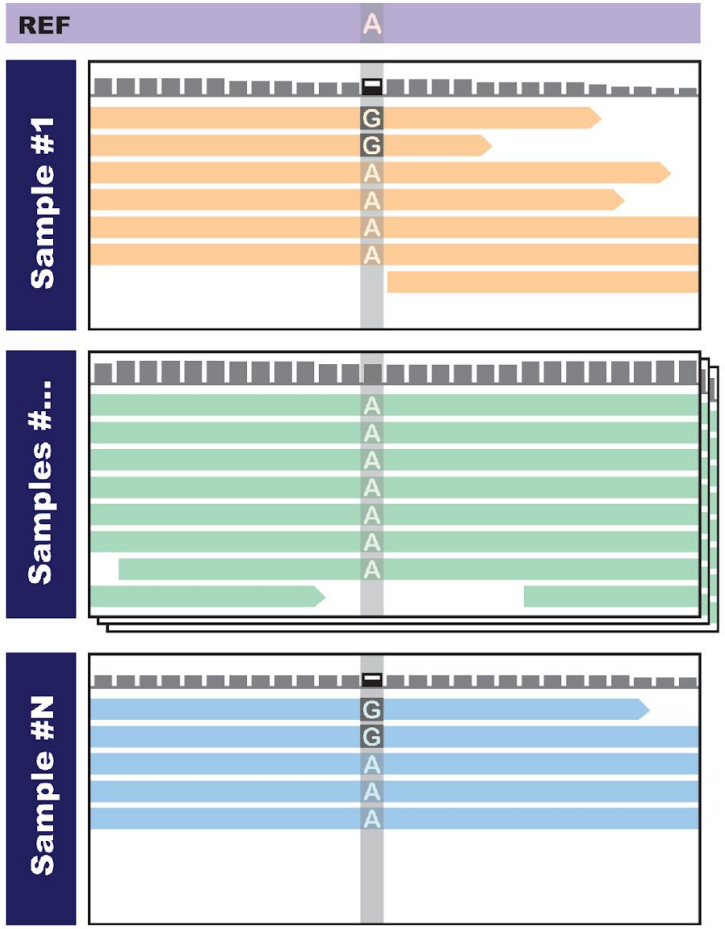
- Выравнивать можно на геном и на транскриптом
- Пользуйтесь современными, хорошо задокументированными и поддерживаемыми, open source программами
- Специализированные программы - каждая хороша для своего
  - Bwa - используется для выравнивания в WES/WGS, т.к. дает самые точные оценки качества выравнивания
  - STAR - оптимизирован для RNA-seq, очень быстрый, но надо 32 Гб RAM для мыши/человека
  - TMAP - для торрента. Но можно и STAR!
  - Bowtie - умеет colorspace
  - Bowtie2 - быстрый, универсальный, и неприхотливый

4. Проверка качества выравнивания - MapQ и его распределение

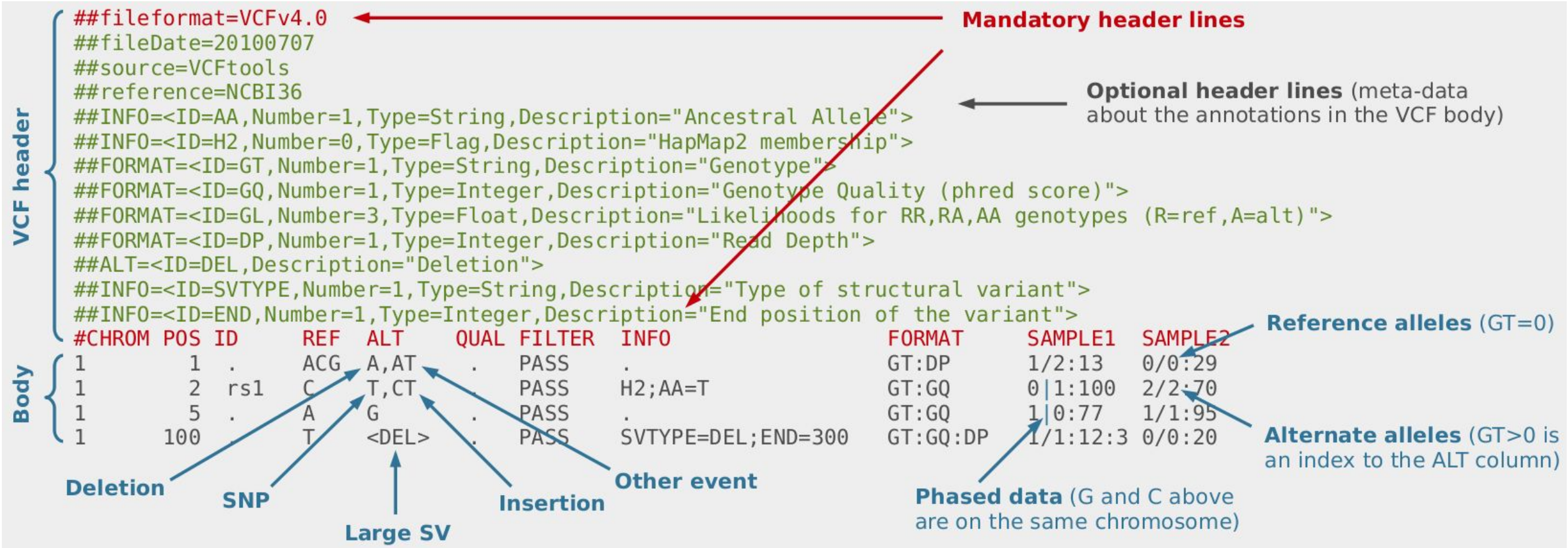
Score – Precision
20 – 0.99
30 – 0.999



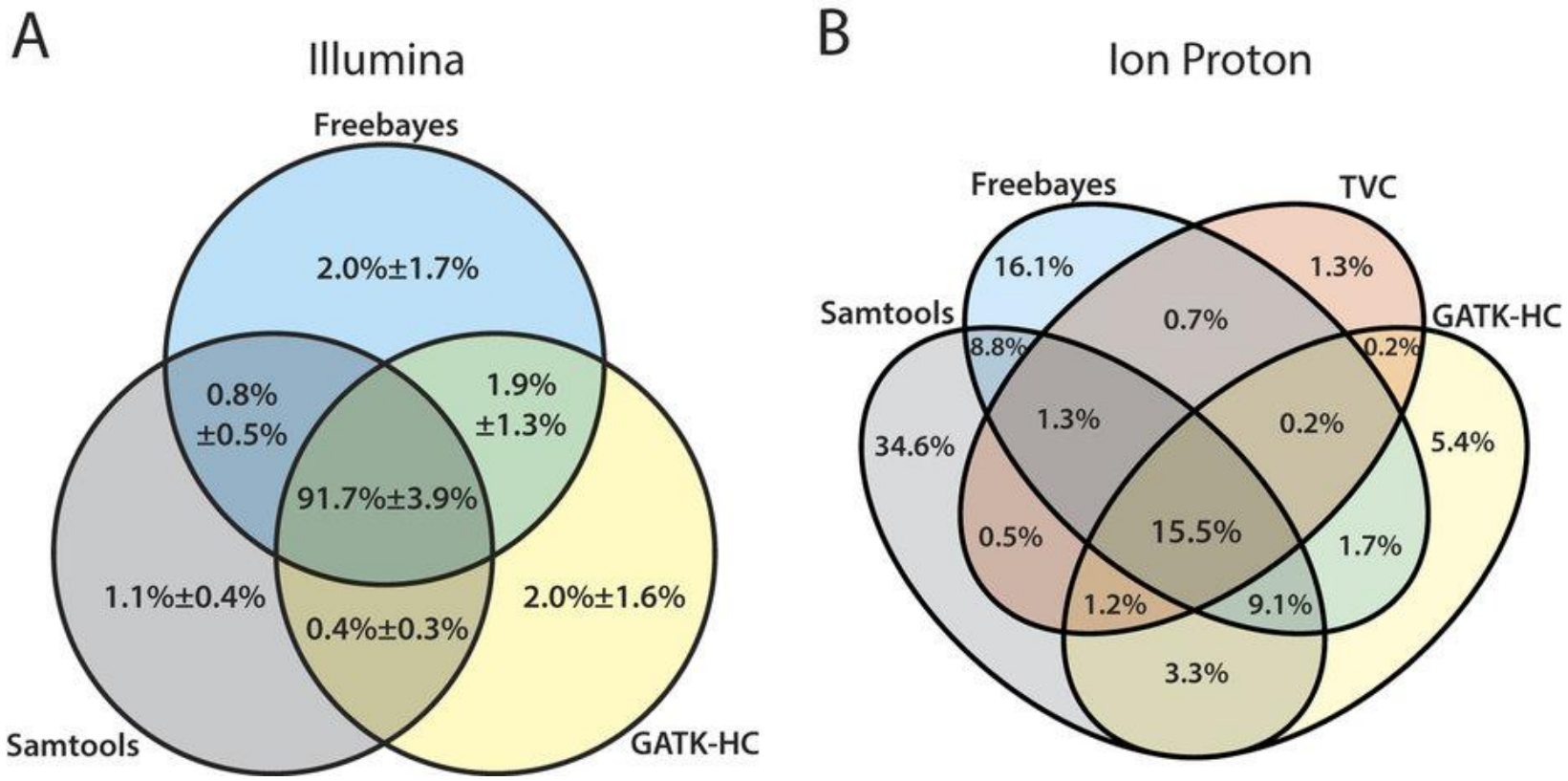
5. Поиск вариантов - Как это работает



VCF – формат данных



6. Контроль качества вариантов Сравнение платформ и методов



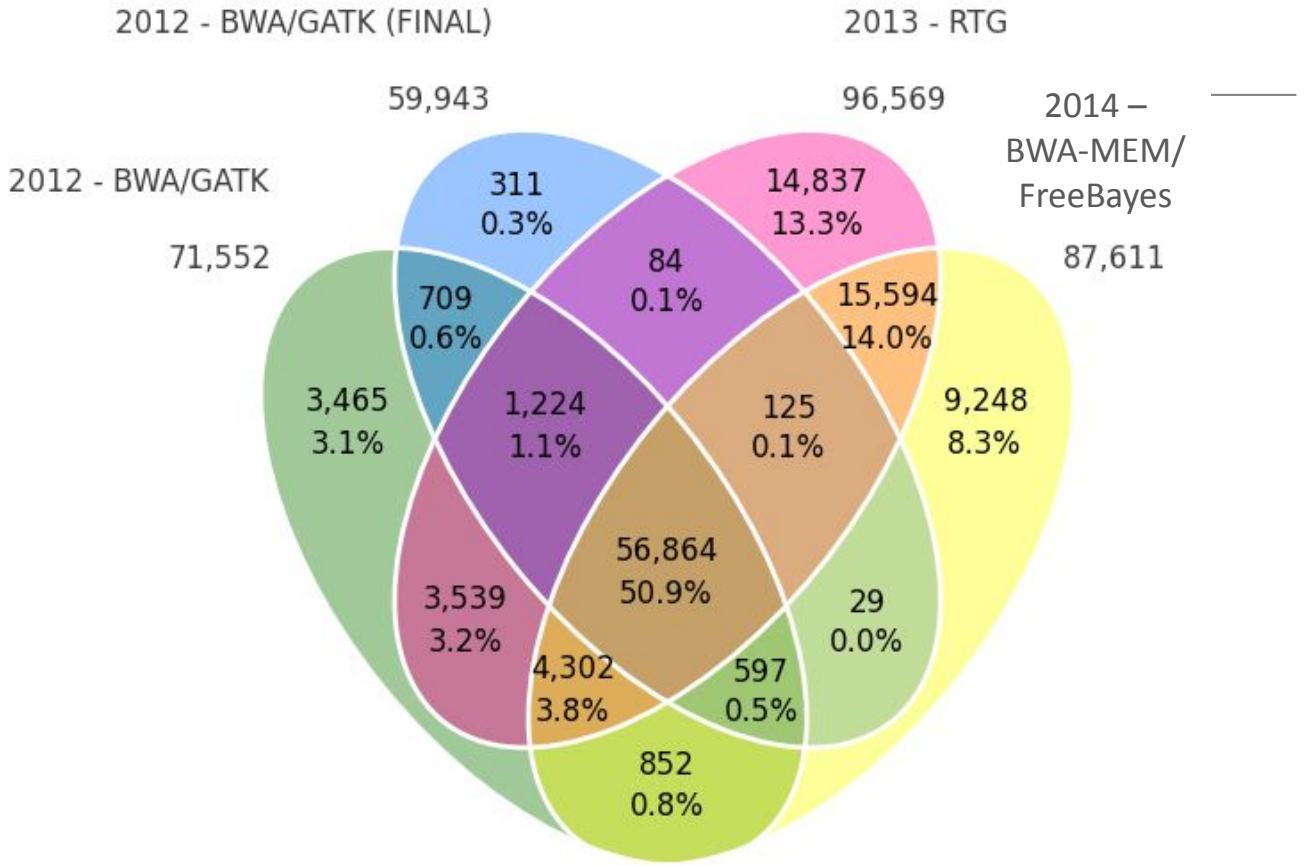
Hwang, Sohyun, et al. "Systematic comparison of variant calling pipelines using gold standard personal exome variants." *Scientific reports* 5 (2015).

Alignment and Variant Calling Broken Down

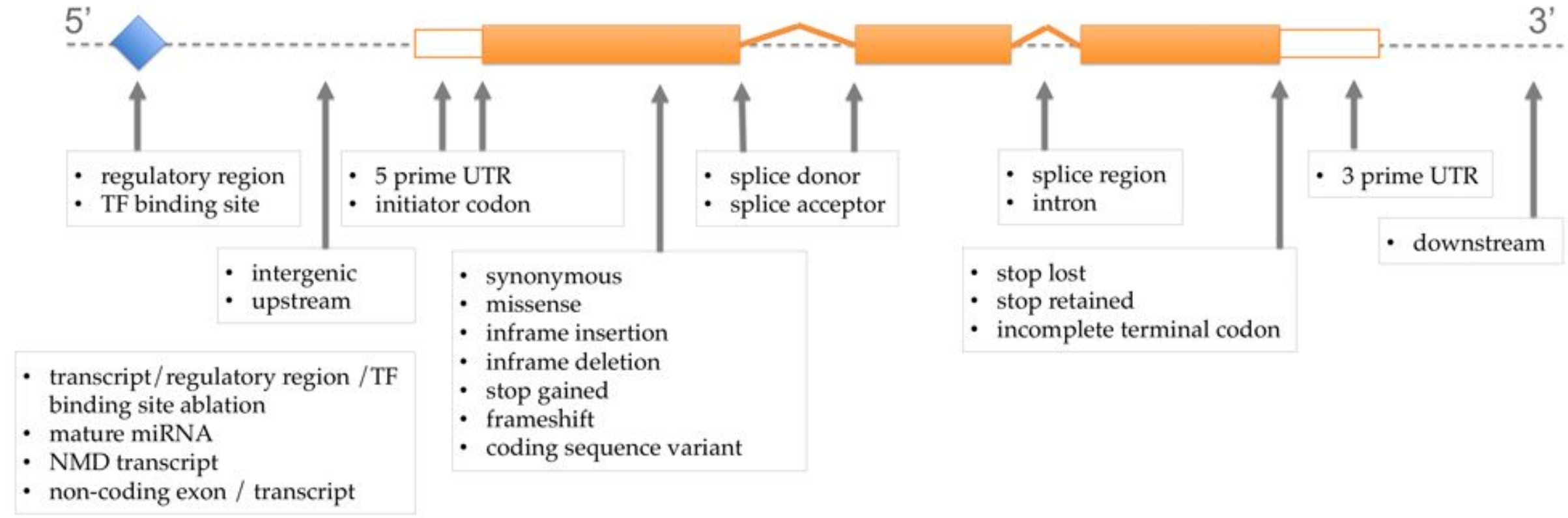
2012 2 VCFs from 23andMe
- BWA 0.6.1
- GATK (early & late 2012)

2013 Real Time Genomics
- v3.1.2 2013-05-02
- Called on Trio

2014 Rerun
- BWA 0.7.6 (2014-01-31)
- FreeBayes



7. Аннотация вариантов – Предсказание эффекта



7. Аннотация вариантов – SnpEff, SIFT, PolyPhen, VEP

Variant Effect Predictor Results:

Uploaded Variation	Location	Allele	Gene	Transcript	Consequence	Position in cDNA	Position in CDS	Position in protein	Amino acid change	Codon change	Co-located Variation	Extra
7_117171039_G/A	7:117171039	A	ENSG00000001626	ENST00000446805	DOWNSTREAM	-	-	-	-	-	rs1800077	-
7_117171039_G/A	7:117171039	A	ENSG00000001626	ENST00000003084	SYNONYMOUS_CODING	492	360	120	A	gcG/gcA	rs1800077	-
7_117171039_G/A	7:117171039	A	ENSG00000001626	ENST00000454343	SYNONYMOUS_CODING	492	360	120	A	gcG/gcA	rs1800077	-
7_117171039_G/A	7:117171039	A	ENSG00000001626	ENST00000426809	SYNONYMOUS_CODING	360	360	120	A	gcG/gcA	rs1800077	-
7_117171039_G/A	7:117171039	A	ENSG00000256365	ENST00000545164	INTRONIC	-	-	-	-	-	rs1800077	-
7_117171039_G/A	7:117171039	A	ENSG00000256792	ENST00000544858	UPSTREAM	-	-	-	-	-	rs1800077	-
7_117171092_T/C	7:117171092	C	ENSG00000001626	ENST00000446805	DOWNSTREAM	-	-	-	-	-	rs1800078	-
7_117171092_T/C	7:117171092	C	ENSG00000001626	ENST00000003084	NON_SYNONYMOUS_CODING	545	413	138	L/P	cTa/cCa	rs1800078	-
7_117171092_T/C	7:117171092	C	ENSG00000001626	ENST00000454343	NON_SYNONYMOUS_CODING	545	413	138	L/P	cTa/cCa	rs1800078	-
7_117171092_T/C	7:117171092	C	ENSG00000001626	ENST00000426809	NON_SYNONYMOUS_CODING	413	413	138	L/P	cTa/cCa	rs1800078	-
7_117171092_T/C	7:117171092	C	ENSG00000256365	ENST00000545164	INTRONIC	-	-	-	-	-	rs1800078	-
7_117171092_T/C	7:117171092	C	ENSG00000256792	ENST00000544858	UPSTREAM	-	-	-	-	-	rs1800078	-
7_117171122_T/C	7:117171122	C	ENSG00000256792	ENST00000544858	UPSTREAM	-	-	-	-	-	rs35516286	-
7_117171122_T/C	7:117171122	C	ENSG00000001626	ENST00000426809	NON_SYNONYMOUS_CODING	443	443	148	I/T	aTt/aCt	rs35516286	-
7_117171122_T/C	7:117171122	C	ENSG00000001626	ENST00000446805	DOWNSTREAM	-	-	-	-	-	rs35516286	-
7_117171122_T/C	7:117171122	C	ENSG00000001626	ENST00000454343	NON_SYNONYMOUS_CODING	575	443	148	I/T	aTt/aCt	rs35516286	-
7_117171122_T/C	7:117171122	C	ENSG00000001626	ENST00000003084	NON_SYNONYMOUS_CODING	575	443	148	I/T	aTt/aCt	rs35516286	-
7_117171122_T/C	7:117171122	C	ENSG00000256365	ENST00000545164	INTRONIC	-	-	-	-	-	rs35516286	-

VCF-Annotate

Пример запуска: vcf-annotate -f +/d=8/Q=10/q=10/-a >

Ключ	Описание [стандартное значение]
1, StrandBias FLOAT	Min P-value for strand bias (INFO/PV4) [0.0001]
2, BaseQualBias FLOAT	Min P-value for baseQ bias (INFO/PV4) [0]
3, MapQualBias FLOAT	Min P-value for mapQ bias (INFO/PV4) [0]
4, EndDistBias FLOAT	Min P-value for end distance bias (INFO/PV4) [0.0001]
a, MinAB INT	Minimum number of alternate bases (INFO/DP4) [2]
c, SnpCluster INT1,INT2	Filters clusters of 'INT1' or more SNPs within a run of 'INT2' bases []
D, MaxDP INT	Maximum read depth (INFO/DP or INFO/DP4) [10000000]
d, MinDP INT	Minimum read depth (INFO/DP or INFO/DP4) [2]
H, HWE FLOAT	Minimum P-value for HWE (plus F<0) (INFO/HWE and INFO/G3) [0.0001]
q, MinMQ INT	Minimum RMS mapping quality for SNPs (INFO/MQ) [10]
Q, Qual INT	Minimum value of the QUAL field [10]
r, RefN	Reference base is N []
v, VDB FLOAT	Minimum Variant Distance Bias (INFO/VDB) [0.015]
W, GapWin INT	Window size for filtering adjacent gaps [3]
w, SnpGap INT	SNP within INT bp around a gap to be filtered [10]

Аннотация вариантов

Способ приписать каждому варианту аннотацию

http://wannovar.wglab.org/

Chr
Start
End
Ref
Alt
Func
Gene
GeneDetail
ExonicFunc
AAChange
1000G ALL
1000G AFR
1000G AMR
1000G EAS
1000G EUR
1000G SAS
ExAC Freq
ExAC AFR
ExAC AMR
ExAC EAS
ExAC FIN
ExAC NFE
ExAC OTH
ExAC SAS
ESP6500si ALL
ESP6500si AA
ESP6500si EA
CG46
ClinVar SIG
ClinVar DIS
ClinVar STATUS
ClinVar ID ClinVar DB ClinVar DBID
GWAS DIS GWAS OR GWAS BETA
GWAS PUBMED
GWAS SNP
GWAS P
SIFT score
SIFT pred
Polyphen2 HDIV score
Polyphen2 HDIV pred
Polyphen2 HVAR score
Polyphen2 HVAR pred
LRT score
LRT pred MutationTaster score MutationTaster pred
MutationAssessor score
MutationAssessor pred
FATHMM score FATHMM pred RadialSVM score
RadialSVM pred
LR score
LR pred
VEST3 score
CADD raw
CADD phred
GERP++ RS
phyloP46way placental
phyloP100way vertebrate
SiPhy 29way logOdds

Table 1. Properties of genomic variation and phenotype databases

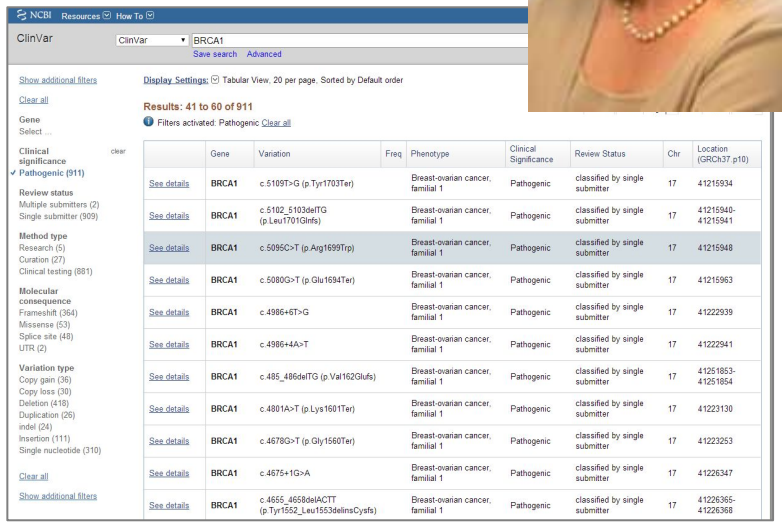
	1000 Genomes	NHLBI Exome Variant Server	dbSNP	Human Gene Mutation Database	Locus-specific databases	OMIM	GeneReviews	ClinVar
Focus	Genome/exome variation in diverse populations, germline only	Exome variation in well-phenotyped populations, germline only	Repository for all molecular variation, both germline and somatic	Detailed information on variants responsible for inherited disease, germline only	Gene-specific variants, some with expert curation, both germline and somatic	Literature review for genes and genetic phenotypes, germline and somatic variants	Expert clinical review based on the literature for genes and the phenotypes associated with germline and somatic variants	Clinical significance of variants across all genes, both germline and somatic
Variant source	Variants from sequence data in individuals from 26 populations	Variants from sequence data in phenotyped individuals, many with rare disorders	Submitted by research/ clinical groups	Variants mined from the literature, does not include unpublished variants	Submitted by research/clinical groups, database specific	Selected variants mined from the literature	Variants selected by authors based on their phenotypic relevance	Submitted by research/ clinical groups or extracted from public databases or expert consensus reports
Phenotype	None provided	Focused phenotype information available through dbGAP	May provide clinical significance of variant	Phenotypic information limited to associated disease	May provide detailed phenotype per submission	Thorough review of the phenotype	Thorough review of the phenotype	Limited phenotypic information
Clinical resource	None	None	None	None	None	Clinical synopsis/ literature review of clinical details	Includes clinical practice guidelines	Can include variant-specific practice guidelines
Prediction of Causation	None	None	May provide submitter prediction	Yes Interpretation of the literature with prediction of causation	Yes Two part prediction, submitter/curator	Yes Associated phenotype with interpretation of the literature	Yes Interpretation of the literature with prediction of causation	Yes Submitter prediction/ expert predictions
Model-based information	No	Yes PolyPhen2 Conservation	No	Yes Sift, MutPred Conservation	No	No	Not standard, may be included	Not standard, may be included
Accessibility	Public	Public	Public	Academic and non-profit limited access/fee-based full access	Public	Public	Public	Public
Curator	1000 Genomes	University of Washington	NCBI Limited curation	HGMD Subscribers can submit feedback	Various experts	Johns Hopkins	University of Washington-based editors	NCBI Individuals can review variants and submit a reviewed record
References	None	None	If provided by submitter, may be mined from PubMed	First report of all mutations, additional reports may be included	Variant-specific references when available	Gene- and variant-specific references	Gene- and variant-specific references	Gene-specific references, variant data linked to submitter, may or may not have reference

ClinVar

ClinVar is designed to provide a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence.

Submitters:

- OMIM: Johns Hopkins
- Samuels
- Lab for Molecular Medicine
- Invitae
- Emory Genetics Lab



Star rating system

- 0-4 stars – level of review

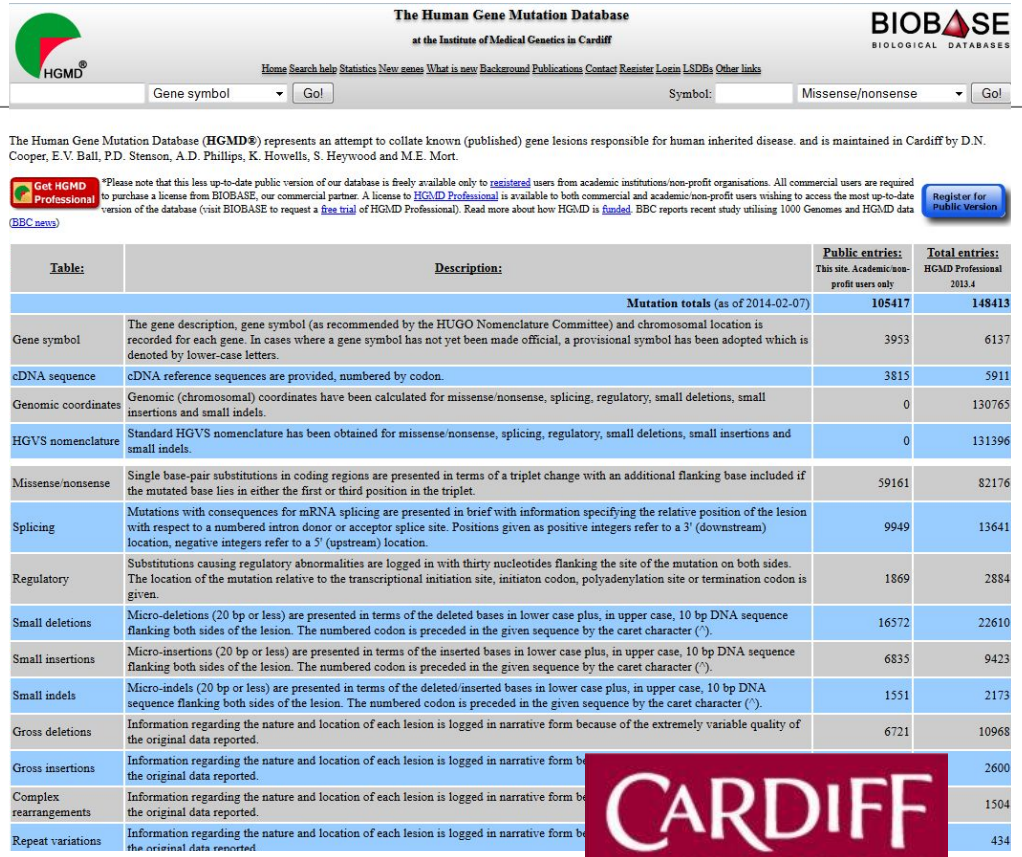
HGMD

Data mines academic papers for reported functional variants

Also takes submissions, corrections reviewed by team

First available in 1996

- Originally 10k variants
- 105k in Public (2014)
- 148k in “Pro” (2014)



The Human Gene Mutation Database

at the Institute of Medical Genetics in Cardiff

BIOBASE BIOLOGICAL DATABASES

Home Search help Statistics New genes What is new Background Publications Contact Register Login LSDBs Other links

Gene symbol Go! Symbol: Missense/nonsense Go!

The Human Gene Mutation Database (**HGMD®**) represents an attempt to collate known (published) gene lesions responsible for human inherited disease. and is maintained in Cardiff by D.N. Cooper, E.V. Ball, P.D. Stenson, A.D. Phillips, K. Howells, S. Heywood and M.E. Mort.

Get HGMD Professional *Please note that this less up-to-date public version of our database is freely available only to registered users from academic institutions/non-profit organisations. All commercial users are required to purchase a license from BIOBASE, our commercial partner. A license to HGMD Professional is available to both commercial and academic/non-profit users wishing to access the most up-to-date version of the database (visit BIOBASE to request a free trial of HGMD Professional). Read more about how HGMD is funded. BBC reports recent study utilising 1000 Genomes and HGMD data (BBC news)

Register for Public Version

Table:	Description:	Public entries: This site. Academic/non-profit users only	Total entries: HGMD Professional 2013.4
	Mutation totals (as of 2014-02-07)	**105417**	**148413**
Gene symbol	The gene description, gene symbol (as recommended by the HUGO Nomenclature Committee) and chromosomal location is recorded for each gene. In cases where a gene symbol has not yet been made official, a provisional symbol has been adopted which is denoted by lower-case letters.	3953	6137
cDNA sequence	cDNA reference sequences are provided, numbered by codon.	3815	5911
Genomic coordinates	Genomic (chromosomal) coordinates have been calculated for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.	0	130765
HGVS nomenclature	Standard HGVS nomenclature has been obtained for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.	0	131396
Missense/nonsense	Single base-pair substitutions in coding regions are presented in terms of a triplet change with an additional flanking base included if the mutated base lies in either the first or third position in the triplet.	59161	82176
Splicing	Mutations with consequences for mRNA splicing are presented in brief with information specifying the relative position of the lesion with respect to a numbered intron donor or acceptor splice site. Positions given as positive integers refer to a 3' (downstream) location, negative integers refer to a 5' (upstream) location.	9949	13641
Regulatory	Substitutions causing regulatory abnormalities are logged in with thirty nucleotides flanking the site of the mutation on both sides. The location of the mutation relative to the transcriptional initiation site, initiaton codon, polyadenylation site or termination codon is given.	1869	2884
Small deletions	Micro-deletions (20 bp or less) are presented in terms of the deleted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	16572	22610
Small insertions	Micro-insertions (20 bp or less) are presented in terms of the inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	6835	9423
Small indels	Micro-indels (20 bp or less) are presented in terms of the deleted/inserted bases in lower case plus, in upper case, 10 bp DNA sequence flanking both sides of the lesion. The numbered codon is preceded in the given sequence by the caret character (^).	1551	2173
Gross deletions	Information regarding the nature and location of each lesion is logged in narrative form because of the extremely variable quality of the original data reported.	6721	10968
Gross insertions	Information regarding the nature and location of each lesion is logged in narrative form b		2600
Complex rearrangements	Information regarding the nature and location of each lesion is logged in narrative form b the original data reported.		1504
Repeat variations	Information regarding the nature and location of each lesion is logged in narrative form b		434

CARDIFF UNIVERSITY

PRIFYSGOL CAERDYDD

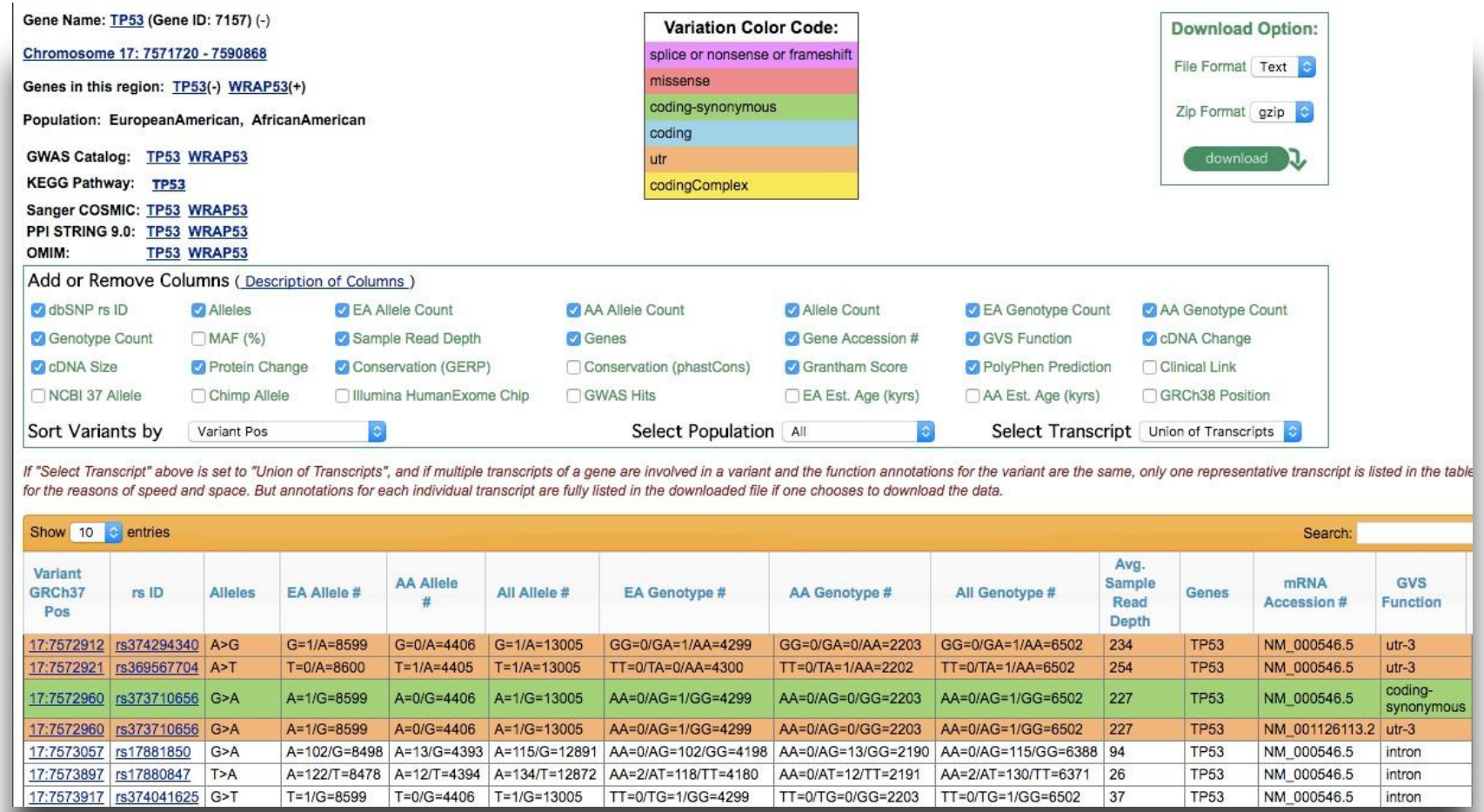
7. Аннотация вариантов – База 1000 человеческих геномов

7. Аннотация вариантов – База 1000 человеческих геномов

1000 Genomes Release	Variants	Individuals	Populations
Phase 3	84.4 million	2504	26
Phase 1	37.9 million	1092	14
Pilot	14.8 million	179	4

NHLBI Exome Sequencing Project (ESP)

http://evs.gs.washington.edu/EVS/



Gene Name: TP53 (Gene ID: 7157) (-)

Chromosome 17: 7571720 - 7590868

Genes in this region: TP53(-) WRAP53(+)

Population: EuropeanAmerican, AfricanAmerican

GWAS Catalog: TP53 WRAP53

KEGG Pathway: TP53

Sanger COSMIC: TP53 WRAP53

PPI STRING 9.0: TP53 WRAP53

OMIM: TP53 WRAP53

Variation Color Code:

- splice or nonsense or frameshift
- missense
- coding-synonymous
- coding
- utr
- codingComplex

Download Option:

File Format: Text

Zip Format: gzip

download

Add or Remove Columns (Description of Columns)

dbSNP rs ID	Alleles	EA Allele Count	AA Allele Count	Allele Count	EA Genotype Count	AA Genotype Count
Genotype Count	MAF (%)	Sample Read Depth	Genes	Gene Accession #	GVS Function	cDNA Change
cDNA Size	Protein Change	Conservation (GERP)	Conservation (phastCons)	Grantham Score	PolyPhen Prediction	Clinical Link
NCBI 37 Allele	Chimp Allele	Illumina HumanExome Chip	GWAS Hits	EA Est. Age (kyrs)	AA Est. Age (kyrs)	GRCh38 Position

Sort Variants by: Variant Pos Select Population: All Select Transcript: Union of Transcripts

If "Select Transcript" above is set to "Union of Transcripts", and if multiple transcripts of a gene are involved in a variant and the function annotations for the variant are the same, only one representative transcript is listed in the table for the reasons of speed and space. But annotations for each individual transcript are fully listed in the downloaded file if one chooses to download the data.

Show 10 entries Search:

Variant GRCh37 Pos	rs ID	Alleles	EA Allele #	AA Allele #	All Allele #	EA Genotype #	AA Genotype #	All Genotype #	Avg. Sample Read Depth	Genes	mRNA Accession #	GVS Function
17:7572912	rs374294340	A>G	G=1/A=8599	G=0/A=4406	G=1/A=13005	GG=0/GA=1/AA=4299	GG=0/GA=0/AA=2203	GG=0/GA=1/AA=6502	234	TP53	NM_000546.5	utr-3
17:7572921	rs369567704	A>T	T=0/A=8600	T=1/A=4405	T=1/A=13005	TT=0/TA=0/AA=4300	TT=0/TA=1/AA=2202	TT=0/TA=1/AA=6502	254	TP53	NM_000546.5	utr-3
17:7572960	rs373710656	G>A	A=1/G=8599	A=0/G=4406	A=1/G=13005	AA=0/AG=1/GG=4299	AA=0/AG=0/GG=2203	AA=0/AG=1/GG=6502	227	TP53	NM_000546.5	coding-synonymous
17:7572960	rs373710656	G>A	A=1/G=8599	A=0/G=4406	A=1/G=13005	AA=0/AG=1/GG=4299	AA=0/AG=0/GG=2203	AA=0/AG=1/GG=6502	227	TP53	NM_001126113.2	utr-3
17:7573057	rs17881850	G>A	A=102/G=8498	A=13/G=4393	A=115/G=12891	AA=0/AG=102/GG=4198	AA=0/AG=13/GG=2190	AA=0/AG=115/GG=6388	94	TP53	NM_000546.5	intron
17:7573897	rs17880847	T>A	A=122/T=8478	A=12/T=4394	A=134/T=12872	AA=2/AT=118/TT=4180	AA=0/AT=12/TT=2191	AA=2/AT=130/TT=6371	26	TP53	NM_000546.5	intron
17:7573917	rs374041625	G>T	T=1/G=8599	T=0/G=4406	T=1/G=13005	TT=0/TG=1/GG=4299	TT=0/TG=0/GG=2203	TT=0/TG=1/GG=6502	37	TP53	NM_000546.5	intron

7. Аннотация вариантов – The Exome Aggregation Consortium (ExAC)

Population	Male Samples	Female Samples	Total
African/African American (AFR)	1,888	3,315	5,203
Latino (AMR)	2,254	3,535	5,789
East Asian (EAS)	2,016	2,311	4,327
Finnish (FIN)	2,084	1,223	3,307
Non-Finnish European (NFE)	18,740	14,630	33,370
South Asian (SAS)	6,387	1,869	8,256
Other (OTH)	275	179	454
Total	33,644	27,062	**60,706**

7. Аннотация вариантов – The Exome Aggregation Consortium (ExAC)

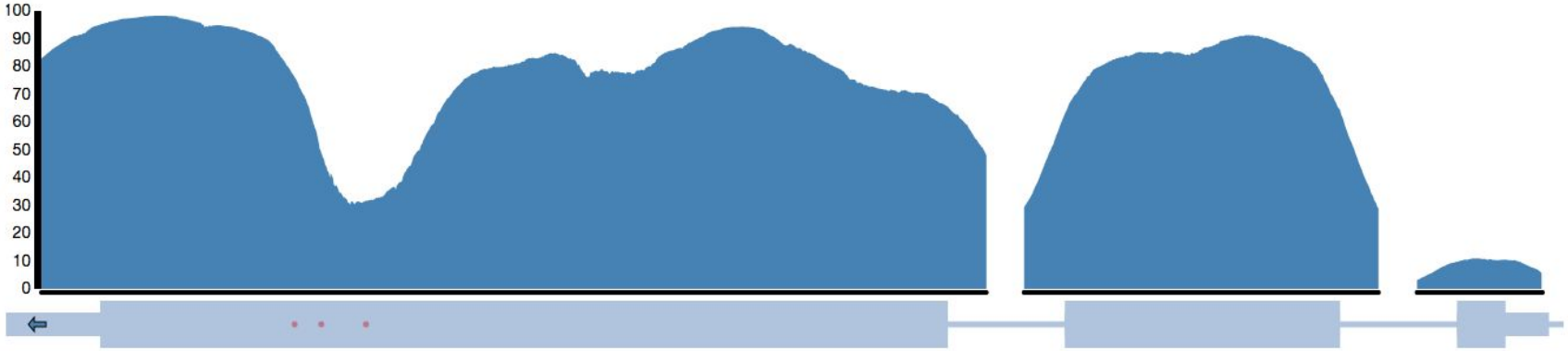
Variant: 20:126155 GCAAA / G

Filter Status	PASS
dbSNP	rs11467497
Allele Frequency	0.1523
Allele Count	18482 / 121328
UCSC	20-126155-GCAAA-G
ClinVar	Click to search for variant in Clinvar

Population Frequencies

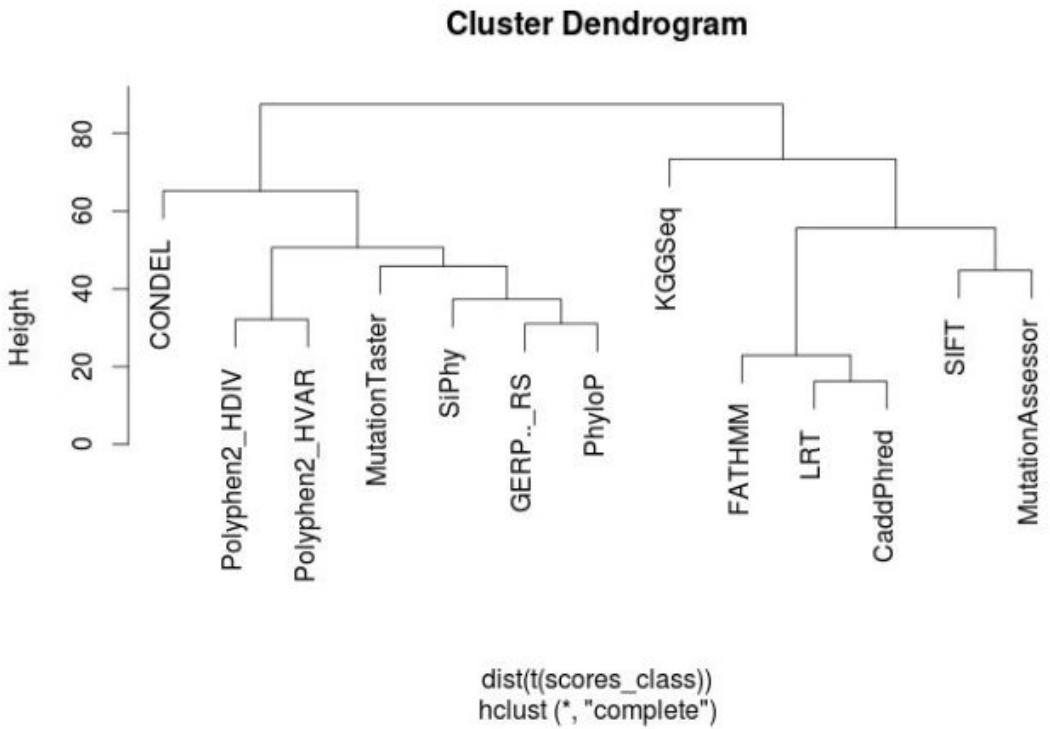
Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	2808	10402	373	0.2699
South Asian	2799	16508	263	0.1696
Other	150	908	17	0.1652
European (Non-Finnish)	10251	66706	781	0.1537
European (Finnish)	733	6610	40	0.1109
Latino	1196	11542	77	0.1036
East Asian	545	8652	21	0.06299
Total	**18482**	**121328**	**1572**	**0.1523**

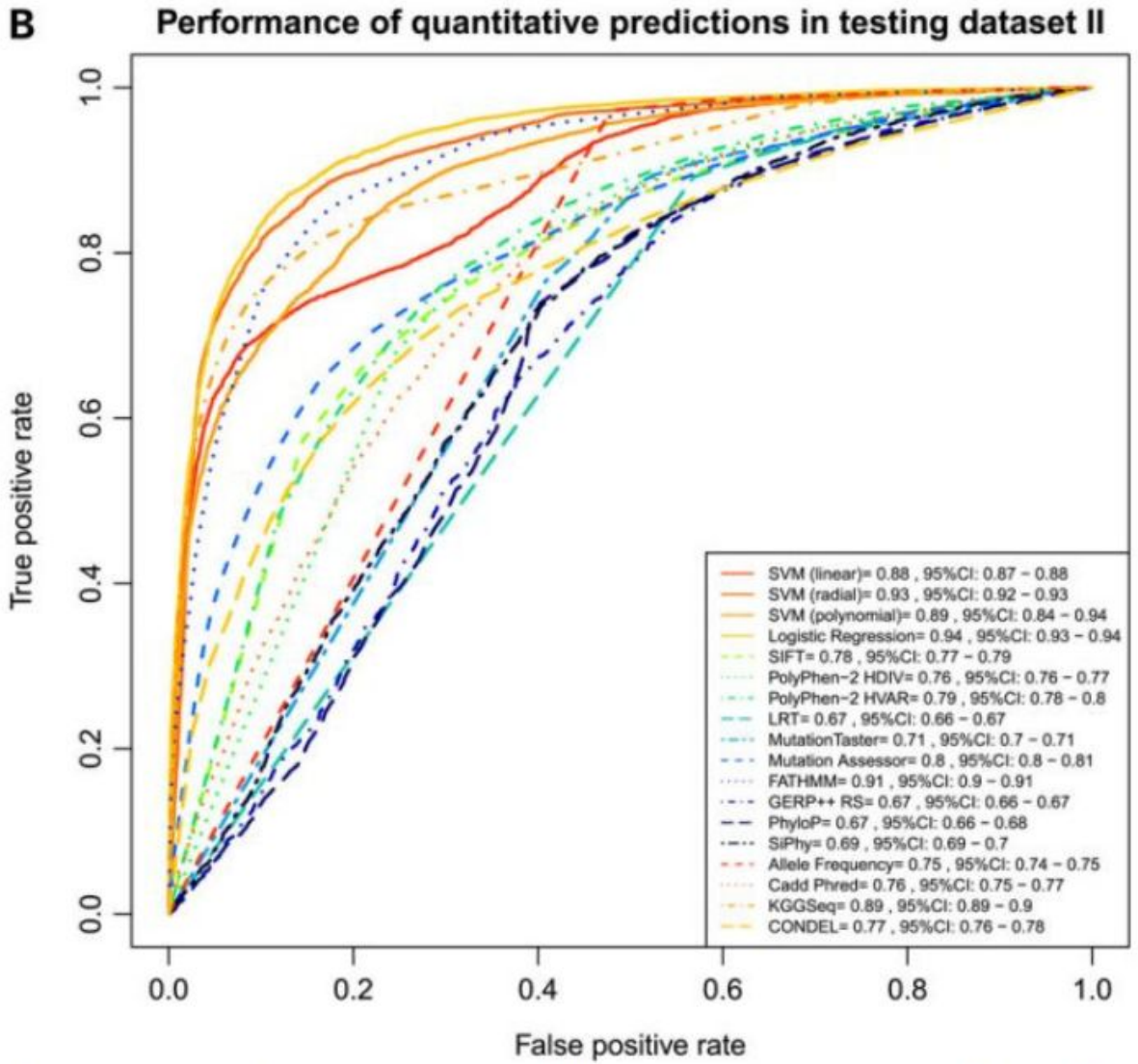
7. Аннотация вариантов – The Exome Aggregation Consortium (ExAC)



Variant	Chrom	Position	Consequence	Filter	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Number of Hemizygotes	Allele Frequency
X:153296070 A / AG	X	153296070	p.Glu416Ter	PASS	frameshift		1	84912	0	0	0.00001178
X:153296104 TCAGG / T	X	153296104	p.Pro403SerfsTer17	PASS	frameshift		1	82363	0	1	0.00001214
X:153296161 G / T	X	153296161	p.Ser385Ter	PASS	stop gained		1	81695	0	0	0.00001224
X:153296689 G / A (rs61749714)	X	153296689	p.Arg161Ter†	PASS	stop gained		48	87676	0	20	0.0005475

InSilico предсказание патогенности





Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies

Disease gene related

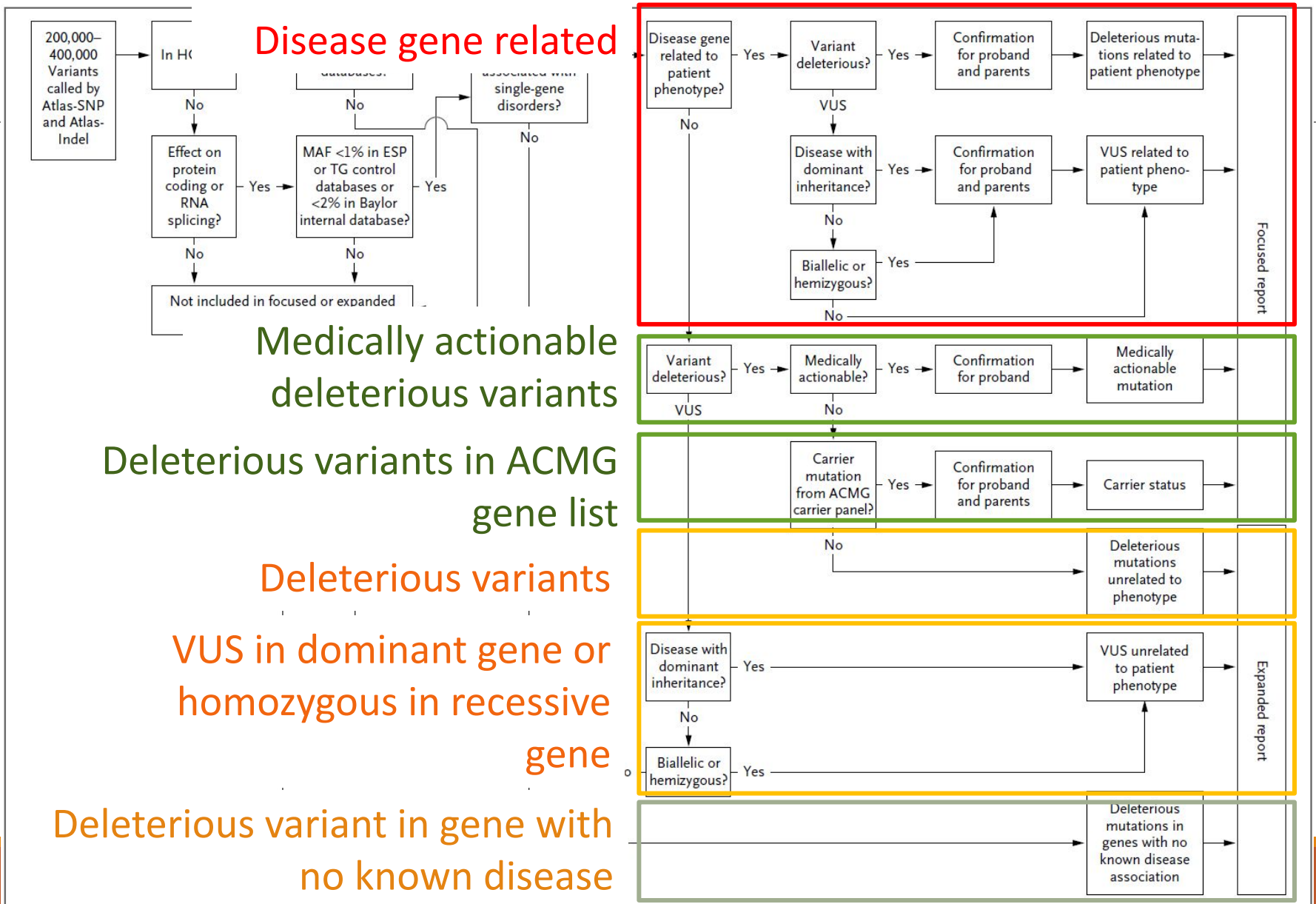
Medically actionable deleterious variants

Deleterious variants in ACMG gene list

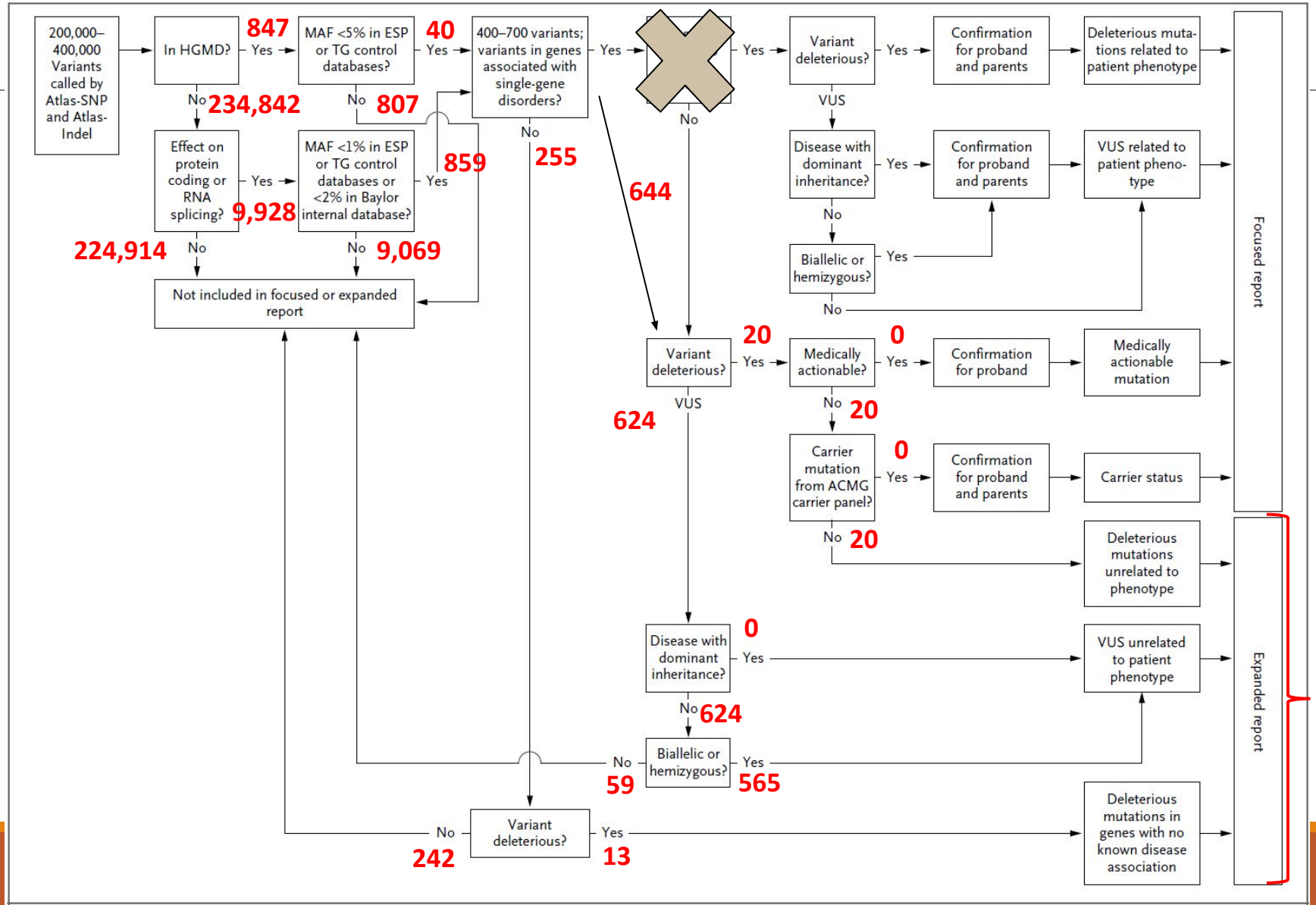
Deleterious variants

VUS in dominant gene or homozygous in recessive gene

Deleterious variant in gene with no known disease



Start: 235,689



Резюме по анализу экзомов

Детальное изучение клиники и семейной истории для формирования клинико-генетической гипотезы

2 основных подхода: анализ по списку генов и поиск *ab initio*

Привлечение информации о консервативности и популяционных данных, агрегированных в базах данных

В идеале: ведение собственной базы экзомных данных для учета локальных частот SNVs

Выбор кандидатных SNVs всегда должен осуществляться на основе данных по экспрессии гена, функции и локализации белка, через призму его возможной этиопатогенетической роли в заболевании.

ОЦЕНКА КЛИНИЧЕСКОЙ ЗНАЧИМОСТИ ВЫЯВЛЕННЫХ ИЗМЕНЕНИЙ

- Тип мутации. Делеции, инсерции, дупликации, мутации сплайсинга, сдвиги рамки считывания, нонсенс-мутации обычно патогенны. Миссенс-мутации требуют более тонкой оценки.

- Частота встречаемость в группах больных и здоровых (должна быть ниже популяционной частоты заболевания!)

- Сегрегация с заболеванием в семье (должна быть у всех больных). Критерий работает только в семейных случаях.

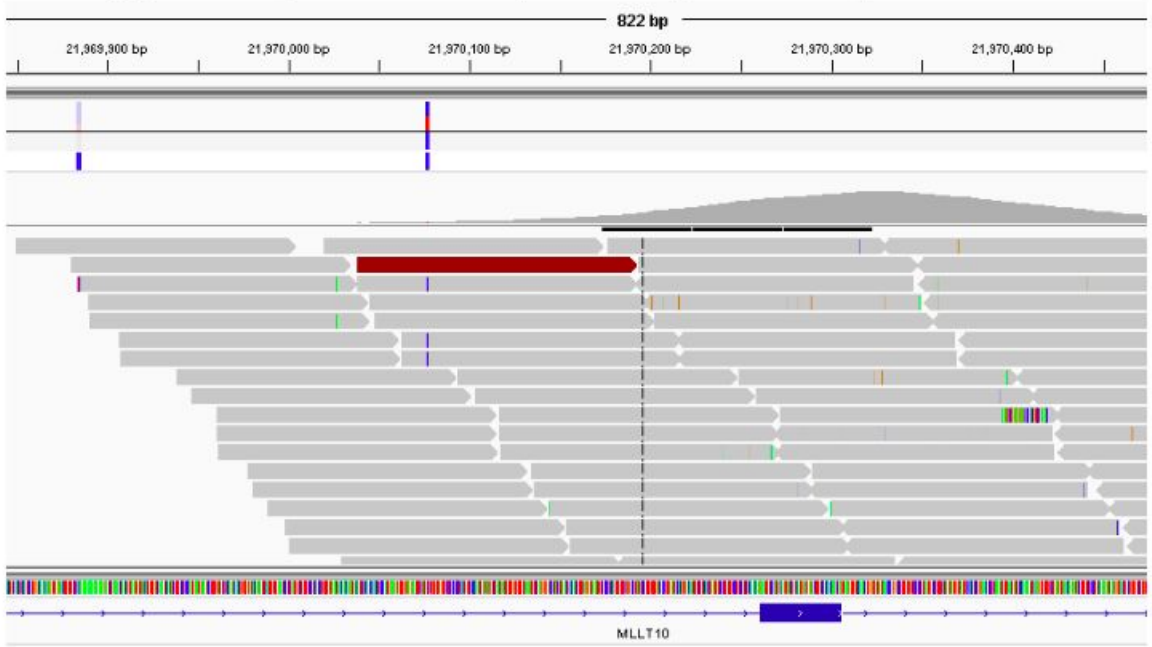
- Биоинформатические ресурсы (базы данных мутаций, публикации, Polyphen, SIFT).

- Данные экспериментальных исследований (клеточные линии, модельные животные).

Integrative Genomics Viewer

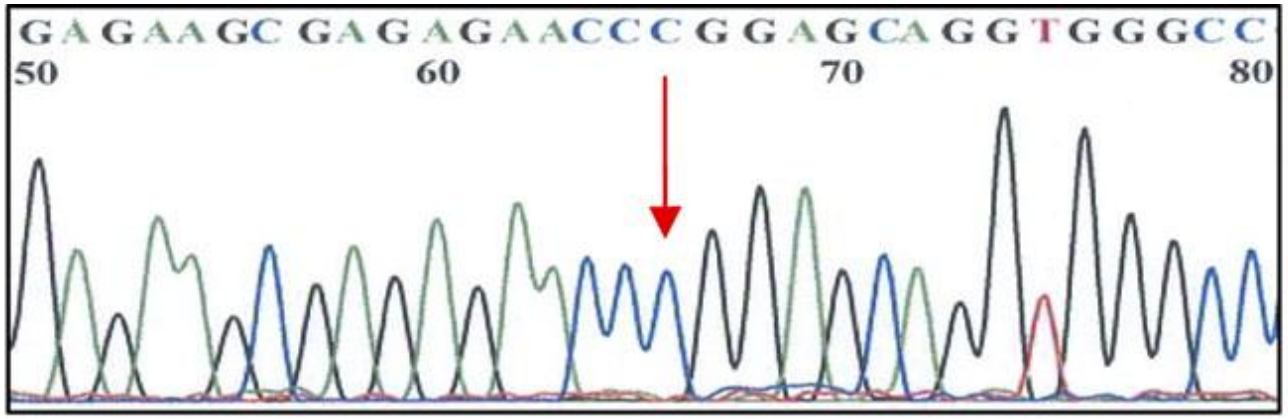
http://www.broadinstitute.org/igv/

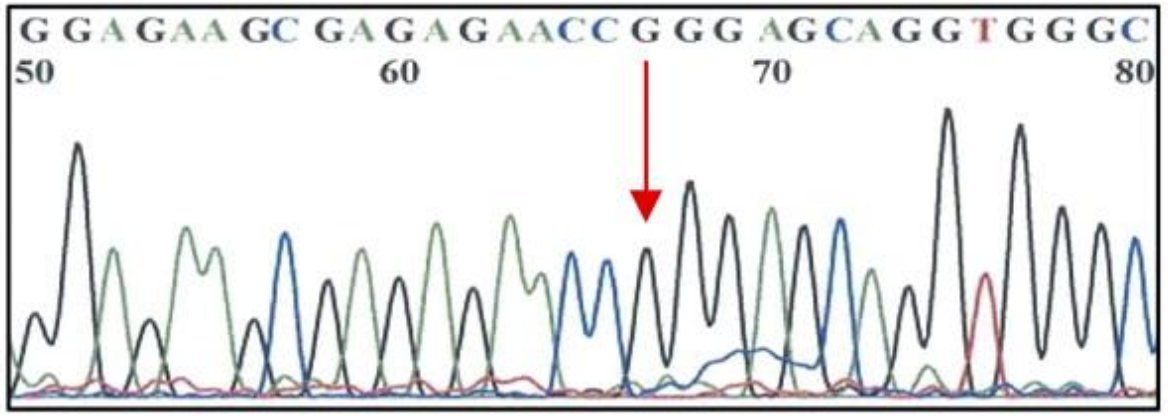
Программа для визуального просмотра bam-файлов



Важно проверять найденные варианты "глазами"

Верификация результата





ГДЕ МОЖНО «ПОТЕРЯТЬ» МУТАЦИИ

1. Качество секвенирования. Мутация МОЖЕТ находится в непокрытых или «скомпрометированных» по покрытию областях – потеря качества на уровне секвенирования конкретной пробы
2. Allelic drop-out. Потеря качества на уровне метода (методическая ошибка).
3. Ошибки коллинга. Частые ошибки прочтения не исключают наличия мутаций в природе.
4. Ошибки интерпретации
5. Ошибки в диагнозе
6. Фенокопии