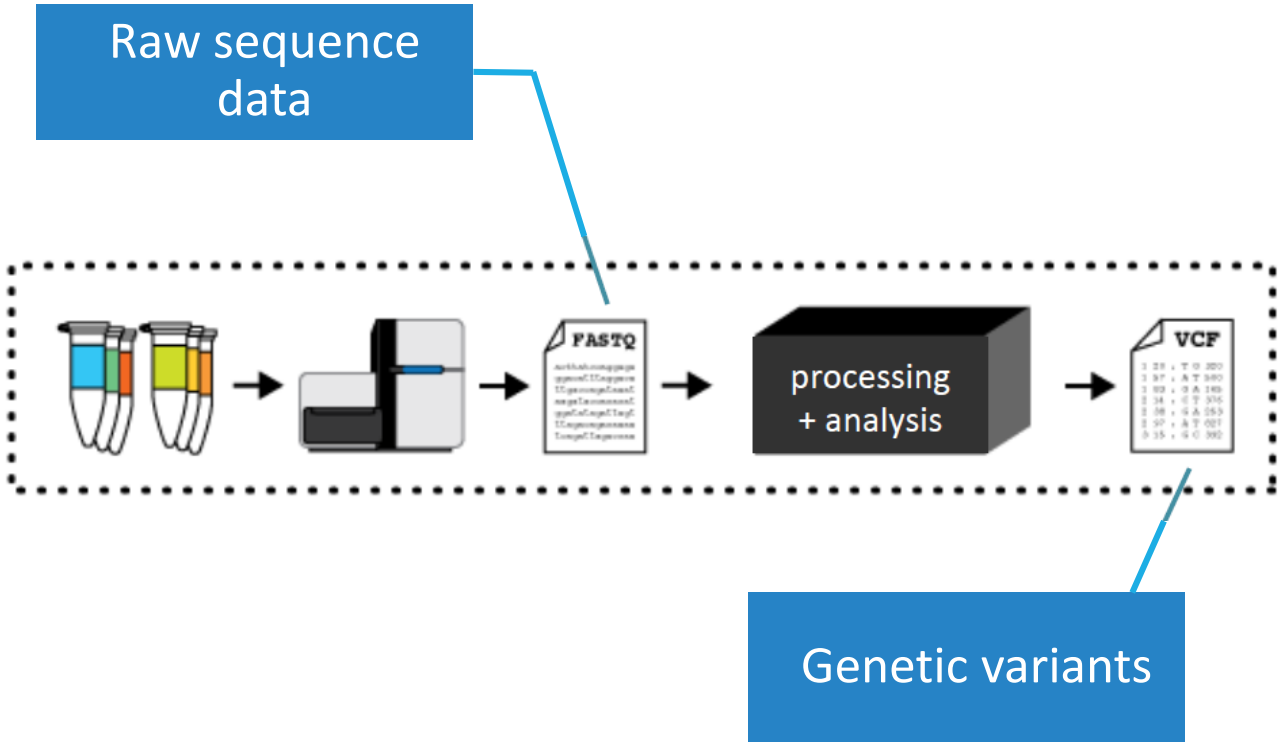
Rare variant analysis with Variant Analyzer

ANDRÁS GÉZSI (BME: MIT; SE: DGCI)

gezsi@mit.bme.hu

16/10/2018

So far: Primary and secondary analysis of raw sequence data

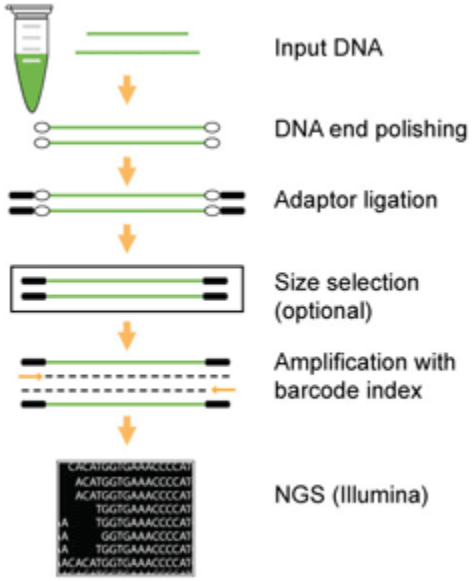


Secondary analysis

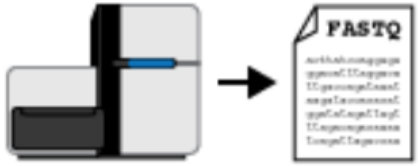
Volume of NGS raw data

	 HiSeq 2500	 HiSeq 3000	 HiSeq 4000	 MiSeq
Output Range	10 - 1000 Gb	125 - 750 Gb	125 - 1500 Gb	0.5 - 15 Gb
Maximum Read Length	2 x 150 bp	2 x 150 bp	2 x 150 bp	2 x 300 bp
Reads per Run	300 million - 4 billion	2.5 billion	2.5 - 5 billion	15 million
Run Time	7 hr - 6 days	<1 - 3.5 days	<1 - 3.5 days	4hr - 55 hr
Key Methods	Exome, transcriptome, & whole-genome sequencing. For Research Use Only. Not for use in diagnostic procedures.	Exome, transcriptome, & whole-genome sequencing. For Research Use Only. Not for use in diagnostic procedures.	Exome, transcriptome, & whole-genome sequencing. For Research Use Only. Not for use in diagnostic procedures.	Small genome, amplicon, & targeted gene panel sequencing. For Research Use Only. Not for use in diagnostic procedures.
Samples per Run*	1 - 8	6	6 - 12	1 - 96

Basics

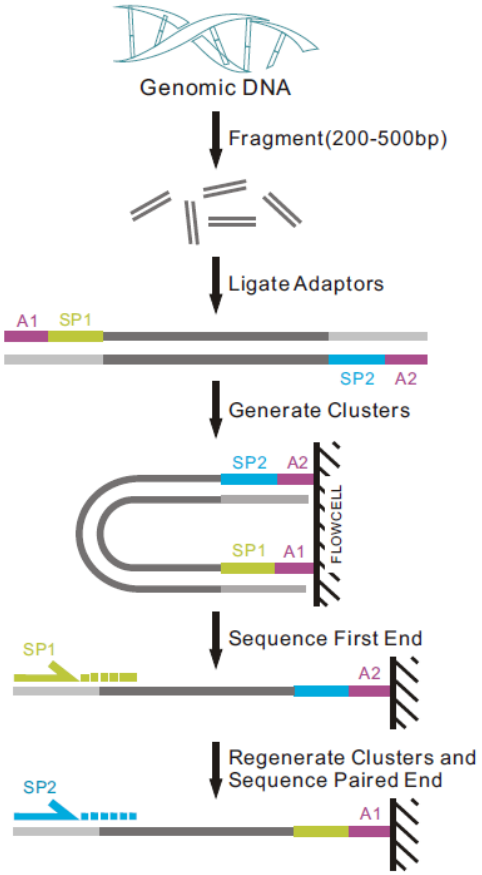


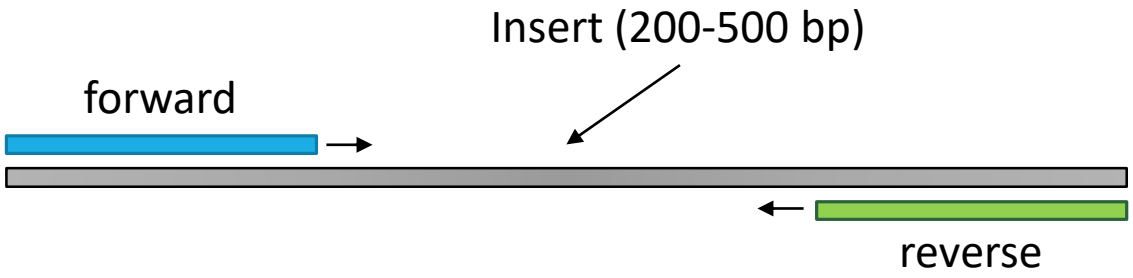
Library preparation



Raw sequences

Paired-end sequencing





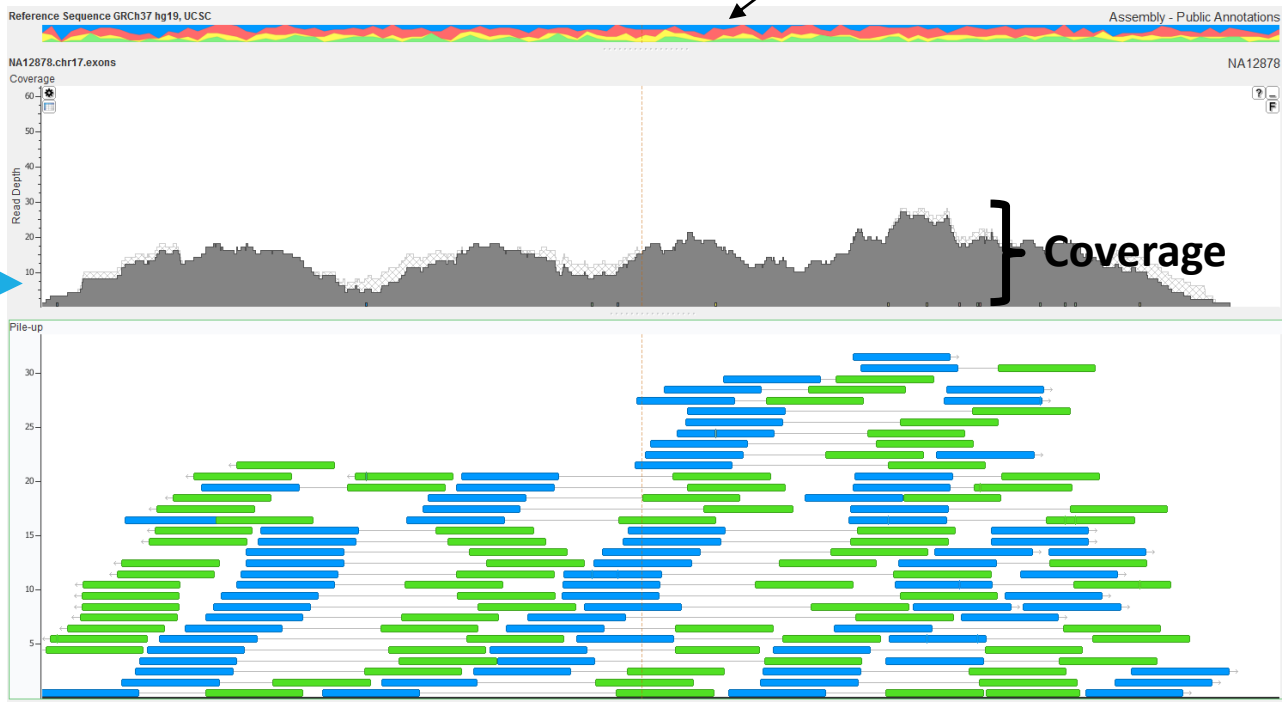
- More accurate mapping (even in the region of repetitive sequences)
- Structural variant detection

Mapping

Raw sequences

Mapping software
+ reference sequence

Reference sequence



Reference sequence

Where does the reference sequence come from?

- Human Genom Project
- Genome Reference Consortium, UCSC

- Human reference genome
  - 2003. 07: (NCBI34/hg16) (hg16)
  - 2004. 05: (NCBI35/hg17) (hg17)
  - 2006. 03: (NCBI36/hg18) (hg18)
  - 2009. 02: (GRCh37/hg19) (hg19)
  - 2013. 12: (GRCh38/hg38) (hg38)

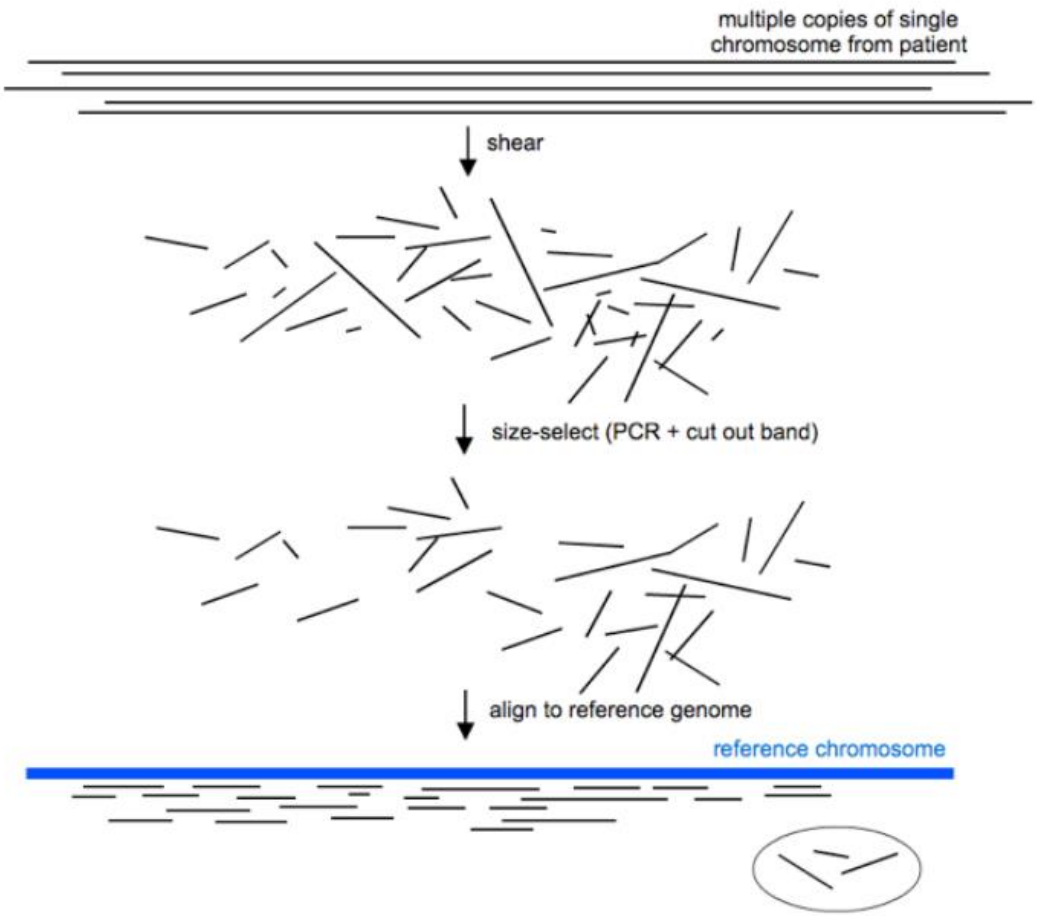
Actual version: GRCh38.p12

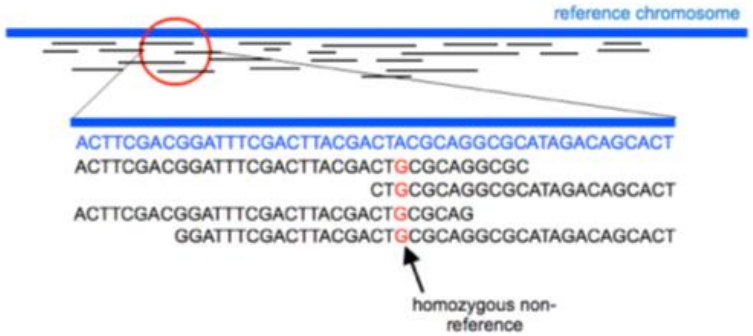
Versions:

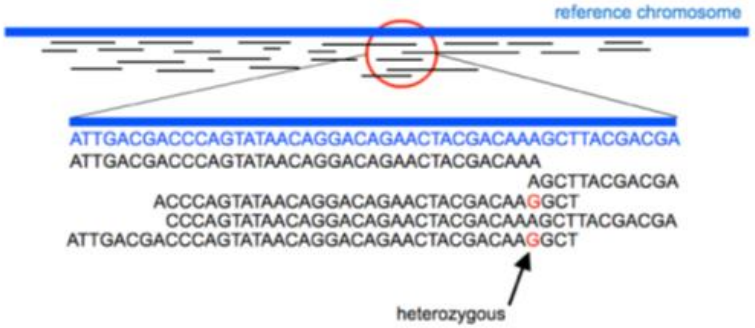
- Patches – every 4 months
- Main versions – by arrangement

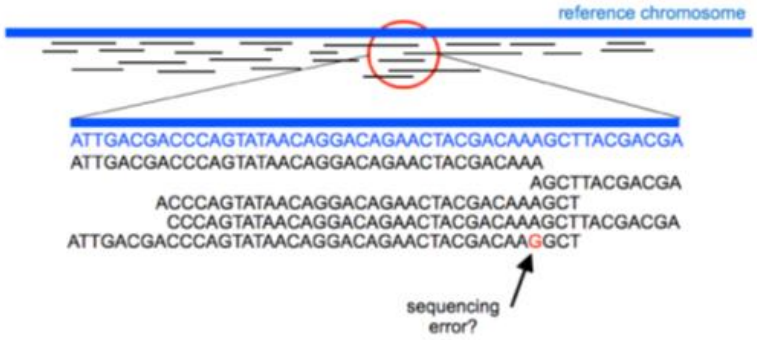
Genomic coordinates may vary between main versions!

Variant calling

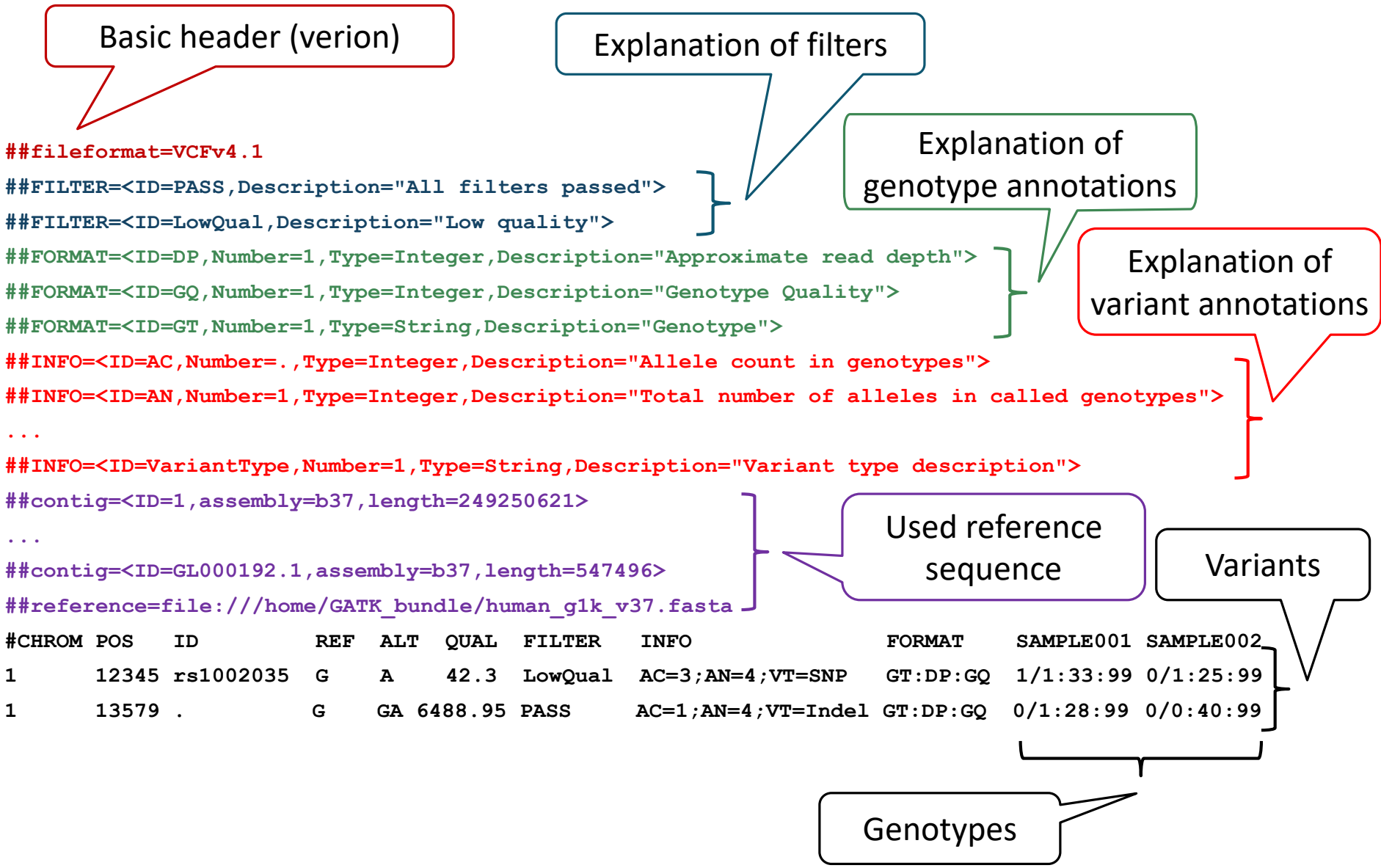








VCF file format



Phred score

- Transformation of probability of error

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

Phred score	Odds of error	Accuray of base calling
10	1 from 10	90%
20	1 from 100	99%
30	1 from 1000	99.9%
40	1 from 10,000	99.99%
50	1 from 100,000	99.999%
60	1 from 1,000,000	99.9999%

Types of variants

- Variants that affect a small number of nucleotides
  - Single nucleotide polymorphisms (SNP)
  - Short insertions and deletions (indel)
  - Multi nucleotide polymorphisms (MNP)

- Structural variants
  - Variations that change copy number (large insertions, deletions, copy number variants (CNV))
  - Variations that do not change copy number (inversions, translocations)

Quality annotations of variants

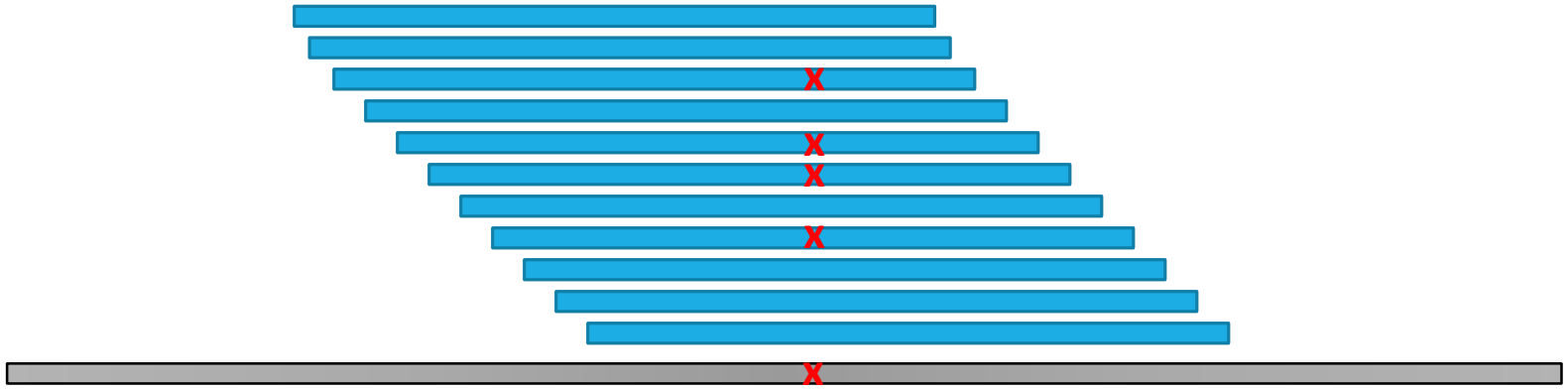
Variant callers report quality parameters for each called variant, these are called **annotation**s.

For example:

- Depth of coverage
- Allele balance
- Mapping quality
- Mapping quality bias
- Strand bias
- Bias in the position of the variant within the sequence

Annotation example: Allele balance

Allele balance = number of reads containing alternative allele / number of all reads



X: Mismatch from reference sequence

Homozygous wild-type => 0.0

Heterozygous => 0.5

Homozygous mutant => 1.0

Annotation example: Mapping quality bias

Mapping quality bias: Is there a difference between the mapping quality score of reads containing alternative allele and reads containing reference allele?

X: Mismatch from reference sequence

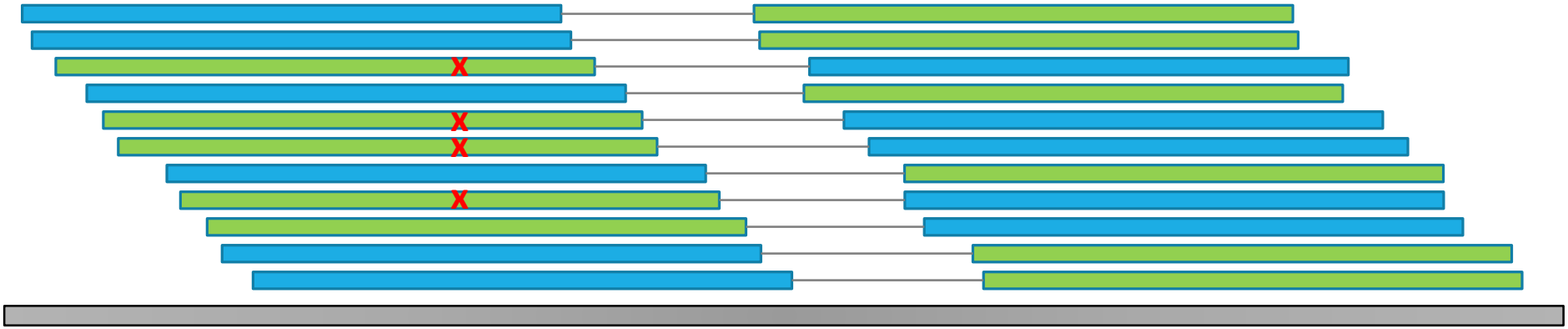
: Low mapping quality

: High mapping quality

In case of real variants:
No difference

Annotation example: Strand bias

Strand bias: Is there a difference between the strand (forward/reverse) of reads containing alternative allele and reads containing reference allele?



X: Mismatch from reference sequence

: Forward strand read

: Reverse strand read

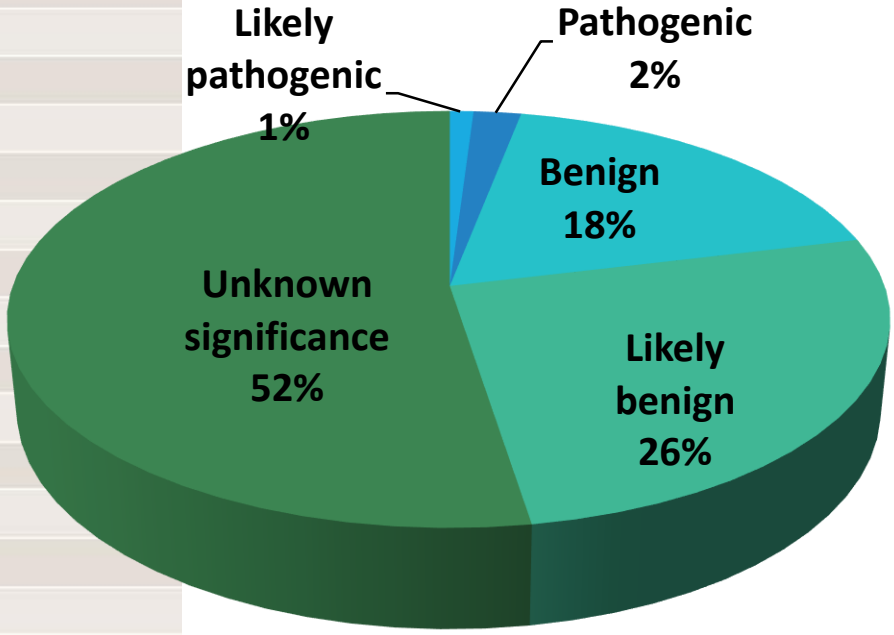
In case of real variants:
No difference

Annotation and analysis of variants

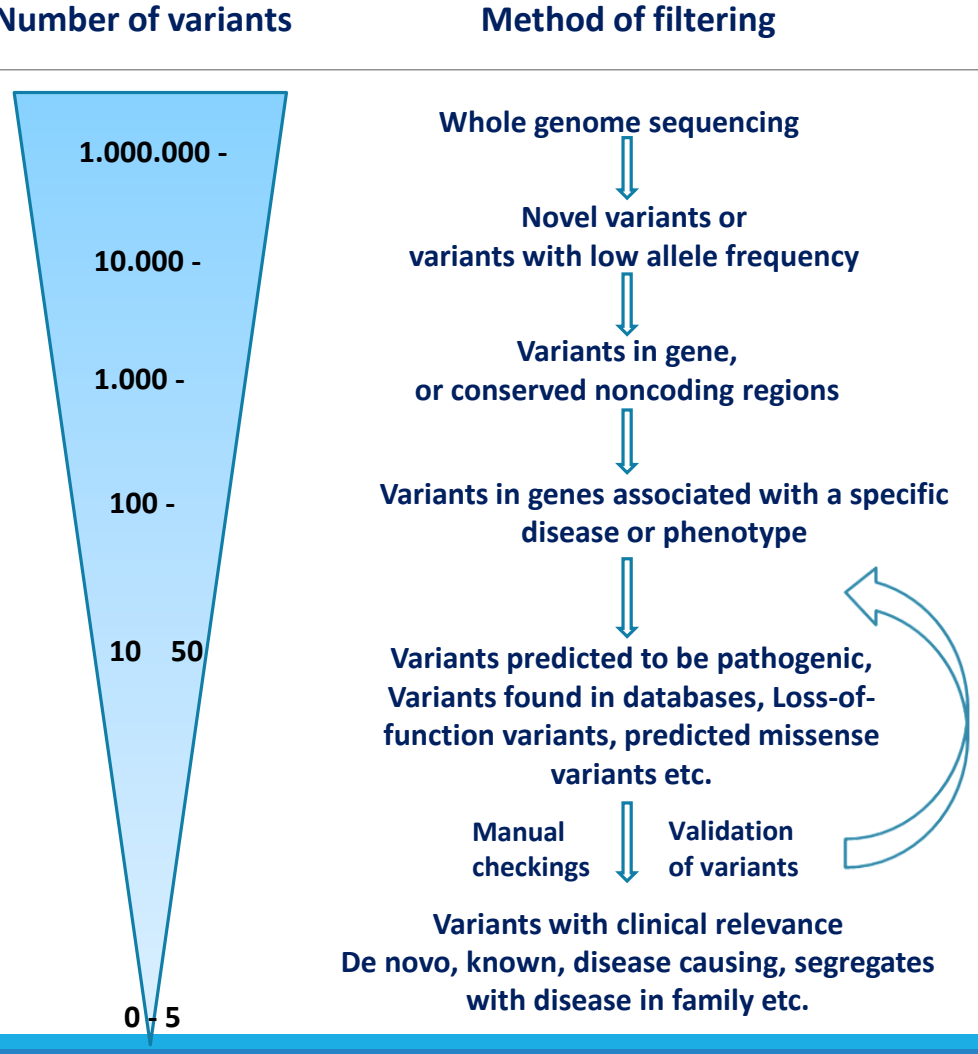
Whole exome sequencing

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

The table lists the mean number (± 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; realignment around insertion–deletions and merging with the Genome Analysis Toolkit (GATK)[91]; and removal of duplicates with PICARD. Variants were called using the following parameters: quality score > 50, allele balance ratio < 0.75; homopolymer run > 3; and quality by depth < 8. Variants were called from a RefSeq37.2 target (35,804,408 bp).

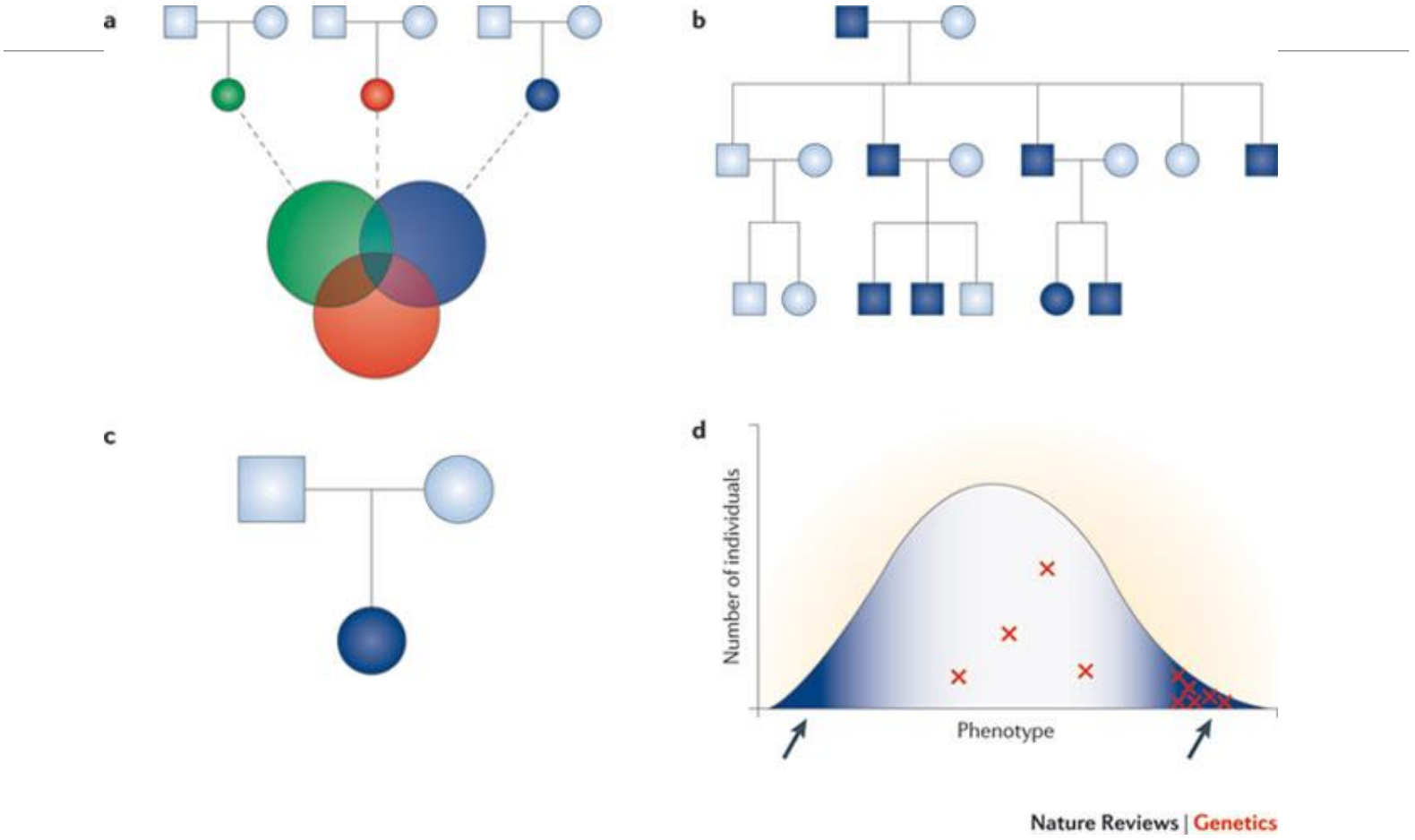


Identifying causal variants (by filtering or prioritization)



„Challenges in the Clinical Use of Genome Sequencing". Green RC, Rehm HL, Kohane IS.
Chapter in Genomic and Personalized Medicine, 2nd Edition, 2012, in press. Ginsburg and Willard (eds.)

Strategies for WES, WGS studies



Prediction softwares

Two main types of prediction softwares

Trained methods	Untrained methods
• Uses machine learning methods (random forest classifier, support vector machine, neural networks etc.) • Compare and learn the annotations/characteristics of known pathogenic / benign variants • Important to consider the nature of the training data E.g. Polyphen2, Mutation Taster	• Based on a priori models to distinguish pathogenic / benign variants • Not very specific, may be more generalizable E.g. SIFT, Mutation Assessor, FATHMM

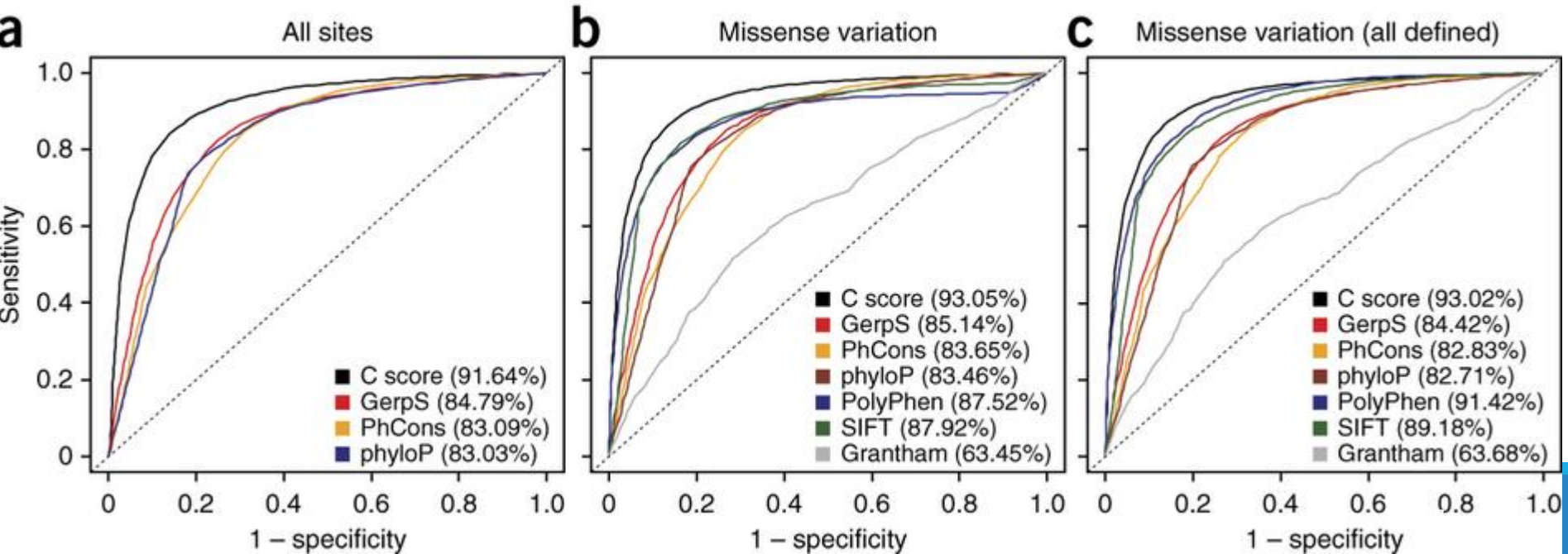
Method	Website	Features	Method description
SIFT	http://sift.bii.a-star.edu.sg/	Sequence based	Statistical method using PSSM with Dirichlet priors
PolyPhen	http://genetics.bwh.harvard.edu/pph/index.html	Sequence based, structure based, annotation	Rule-based model
SNAP	http://www.rostlab.org/services/SNAP/	Sequence based, annotation	Standard feed-forward neural networks with momentum term
MSRV	http://bioinfo.au.tsinghua.edu.cn/member/ruijiang/english/software.html	Sequence based	Multiple selection rule voting strategy using random forest
LRT	http://www.genetics.wustl.edu/jflab/lrt_query.html	Sequence based	Log ratio test
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/index.shtml	Sequence based, structure based	Naïve Bayes approach coupled with entropy-based discretization
MutationTaster	http://www.mutationtaster.org/	Sequence based, annotation	Naïve bayes model based on integrated data source
KGGSeq	http://statgenpro.psychiatry.hku.hk/limx/kggseq/	Sequence based, annotation	A three-level framework to combine a number of filtration and prioritization functions
SInBaD	http://tingchenlab.cmb.usc.edu/sinbad/	Sequence based	Separate mathematical models for promoters, exons, and introns, using logistic regression algorithm
GERP (score)	http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html	Sequence based	A "Rejected Substitutions" score computation to infer the constrained region
PhyloP (score)	http://hgdownload.cse.ucsc.edu/goldenPath/hg18/phyloP44way	Sequence based	An exact *P* value computation under a continuous Markov substitution model

Accuracy of predictions

Generally speaking the accuracy of the predictions are in the range of 65-80%

Many publications, with contradictory results

Methods and testing circumstances (test data) are not the same



Population databases

Exome Aggregation Consortium

http://exac.broadinstitute.org/

60,706 non-relative human,

- diseases-specific and
- populational studies

Potentially many samples with genetic disorders

Severe pediatric disorders were excluded, therefore it can be used for pediatric Mendelian diseases

Exome Aggregation Consortium (ExAC): aggregating and calling 92,000 exomes

Consortia	Samples
Type 2 diabetes case/control	16,167
Heart disease case/control	14,352
Schizophrenia/bipolar case/control	12,361
Inflammatory bowel disease case/control	1,933
The Cancer Genome Atlas (TCGA)	8,566
NHLBI-GO Exome Sequencing Project (ESP)	6,943
1000 Genomes Project	2,520
Sanger (schizophrenia/migraine)	1,348

Subset of **60,706 "reference" samples**:

- high-quality exomes
- unrelated individuals
- consent for public data sharing
- free of **known** severe pediatric disease

All data reprocessed with BWA/Picard

Joint calling across all samples with GATK 3 Haplotype Caller

Exome Variant Server

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

NHLBI GO Exome Sequencing Project (ESP): to identify rare variants in patients with heart, lung and hematological diseases

6,503 exomes, afro-american and european populations

Rare diseases, e.g. 418 exomes (ESP 6%) with cystic fibrosis

1000 Genomes Project

http://browser.1000genomes.org

2500 samples from 14 populations >79 million variants

No phenotypic information

Some populations are underrepresented, with few samples => if a variant is missing from the database that does not necessarily mean it is very rare

Primary goal: find the 95% of all variants with >1% MAF

Pros:

- Diverse populations
- Whole genome data (not just exomes)

dbSNP

http://www.ncbi.nlm.nih.gov/snp

Integrates many databases and self reports

Contains many pathogenic variations

Allele frequency information:

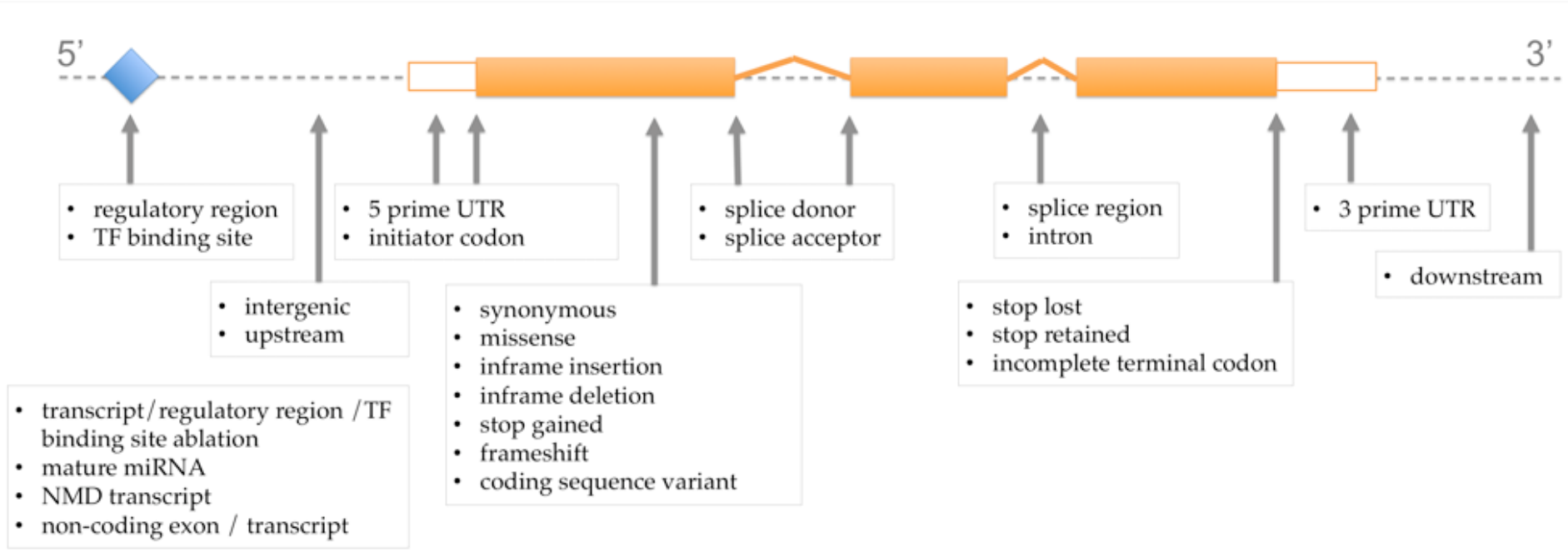
- 1000 Genome Project
- Many *not so reliable* data sources

Contains many false SNPs (15-17%)

- Bioinformatics (2004) 20 (7):1022-1032.)

Softwares for functional predictions

Effect of variants on genes



SnpEff, SnpSift

SnpEff

- Command line tool, Java – Platform independent
- Integrates with many NGS pipelines – Galaxy, GATK
- Input: VCF, Output: VCF
- Many organisms

SnpSift

- Many other annotations (e.g. prediction scores)
- Filtering options

Other softwares

VEP http://www.ensembl.org/info/docs/tools/vep/index.html

Web interface: easy to use, but limited

Input file formats: CSV, VCF, Pileup and HGVS

SeattleSeq http://snp.gs.washington.edu/SeattleSeqAnnotation141/

Web interface: easy to use, but not so many annotations

Input: VCF, or 1 variation

ANNOVAR http://annovar.openbioinformatics.org/

Standalone, perl scripts (+ web interface)

Input: VCF, text file

Many annotations

Need to download databases (~35 GB for human genome)

Filtering based on diseases, phenotypes

Disease databases

Can provide valuable information about a variant (whether it is associated with a given disease)

But (consider the following)

- Variant classification (clinical significance) may be subjective or wrong
- When was it updated?
- Is it curated? Where do the information come from?
- HGVS nomenclature?
- Transcript version is the same?
- An affected person may be present in many databases (not necessarily mean many observations)

dbSNP

„Clinical significance: Assertions of clinical significance for alleles of human sequence variations are reported as provided by the submitter and not interpreted by NCBI. Submissions based on processing data from OMIM® were assigned the value of 'probable-pathogenic', based on a personal communication from Ada Hamosh, director of OMIM. If there is a published authoritative guideline about the pathogenicity of any allele, that is included in the report."

Categories of clinical significance:

- unknown
- untested
- non-pathogenic
- probable-non-pathogenic
- probable-pathogenic
- pathogenic
- drug-response
- histocompatibility
- other

Online Mendelien Inheritance in Man (OMIM)

http://www.omim.org

Contains only selected variants, Manually curated, based on scientific literature

E.g.: first discovered mutation, common occurrence, variants that cause special phenotype, historical relevance, mutation with unusual mechanism, special inheritance, some polymorphisms

Number of Entries in OMIM (Updated October 14th, 2018) :

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	15,166	731	49	35	15,981
Gene and phenotype, combined +	47	0	0	2	49
Phenotype description, molecular basis known #	4,977	327	4	31	5,339
Phenotype description or locus, molecular basis unknown %	1,449	124	4	0	1,577
Other, mainly phenotypes with suspected mendelian basis	1,653	105	3	0	1,761
Totals	23,292	1,287	60	68	24,707

ClinVar

http://www.ncbi.nlm.nih.gov/clinvar

Category of analysis	Current total (Oct 15, 2018)
Records submitted	736432
Records with assertion criteria	612495
Records with an interpretation	718802
Total genes represented	30219
Unique variation records	467058
Unique variation records with interpretations	456787
Unique variation records with assertion criteria	391424
Unique variation records with practice guidelines (4 stars)	23
Unique variation records from expert panels (3 stars)	10438
Unique variation records with assertion criteria, multiple submitters, and no conflicts (2 stars)	63801
Unique variation records with assertion criteria (1 star)	297670
Unique variation records with assertion criteria and a conflict (1 star)	19492
Unique variation records with conflicting interpretations	19649
Genes with variants specific to one gene	6142
Genes with variants specific to one protein-coding gene	6030
Genes included in a variant spanning more than one gene	30182
Variants affecting overlapping genes	15821
Total submitters	1086

HGMD

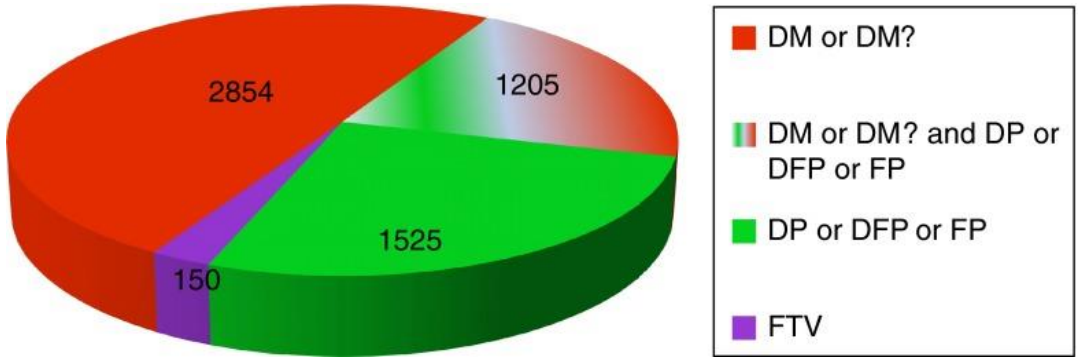
http://www.hgmd.cf.ac.uk/ac/index.php

>157,000 variants, >6600 genes

Public (4-5 years delayed) and commercial versions

No somatic or mithocondrial variants

Manually curated,
based on scientific literature



DM disease causing mutation
DP disease associated polymorphism
FP Functional polymoprhism
DFP Disease-associated polymorphisms with supporting functional evidence
FTV frameshift or truncating variants

Thank you for your attention!

What is uncommon in case of the following variant?

What is its cause?

17:41243800

How many variants have a minor allele frequency greater than 5% in dbSNP in at least 1 population but less than 5% in in this population (regarding all samples)?

How many variants are included in dbSNP?

Among these how many are there that are (nominally) significant according to the allelic statistic test?

How many missense variants have at least 5% minor allele frequencies in the 1000 Genome EUR population?

What can be the disease-causing variants?

How do you set up a filter cascade?