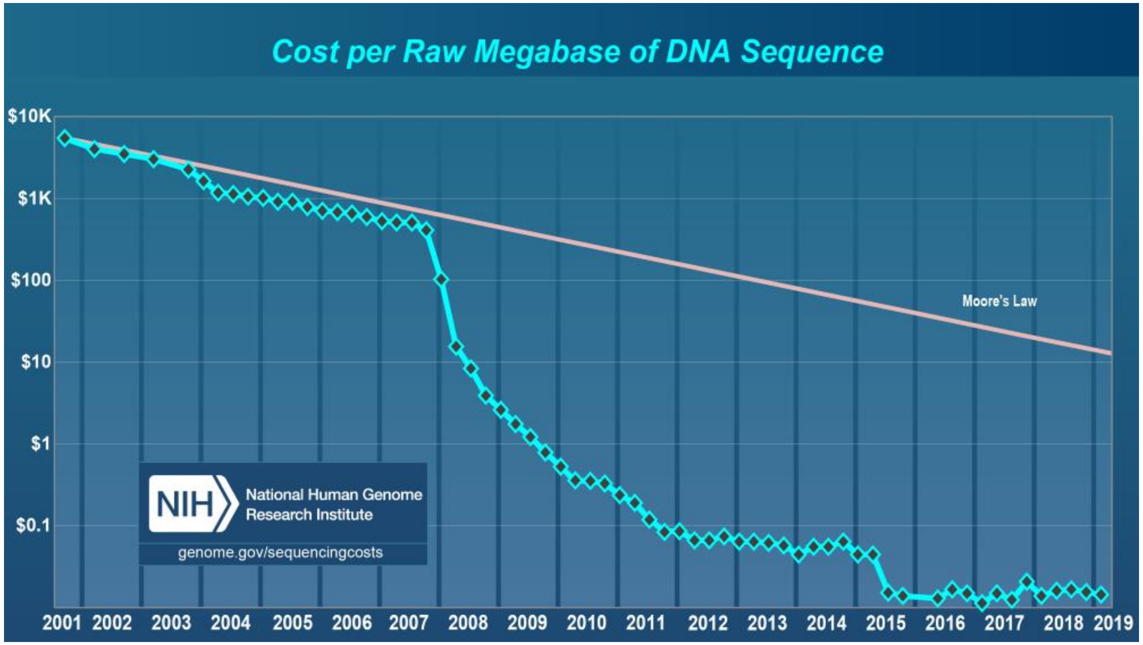
NGS data analysis and quality control

Dr. rer. nat. Marc Sturm

Institut für Medizinische Genetik und angewandte Genomik,
Tübingen

marc.sturm@med.uni-tuebingen.de

Sequencing vs. data analysis cost



Sequence data grows much faster than compute power!

The cost of sequencing will soon be dominated by the cost for analysis and storage of the data: servers, cooling, administration or CPU-hours for cloud-computing.

Thus, **faster algorithms** are needed to analyze the data. Also, a high **sensitivity and specificity** of variant calling is needed to avoid follow-up on false-positive variants.

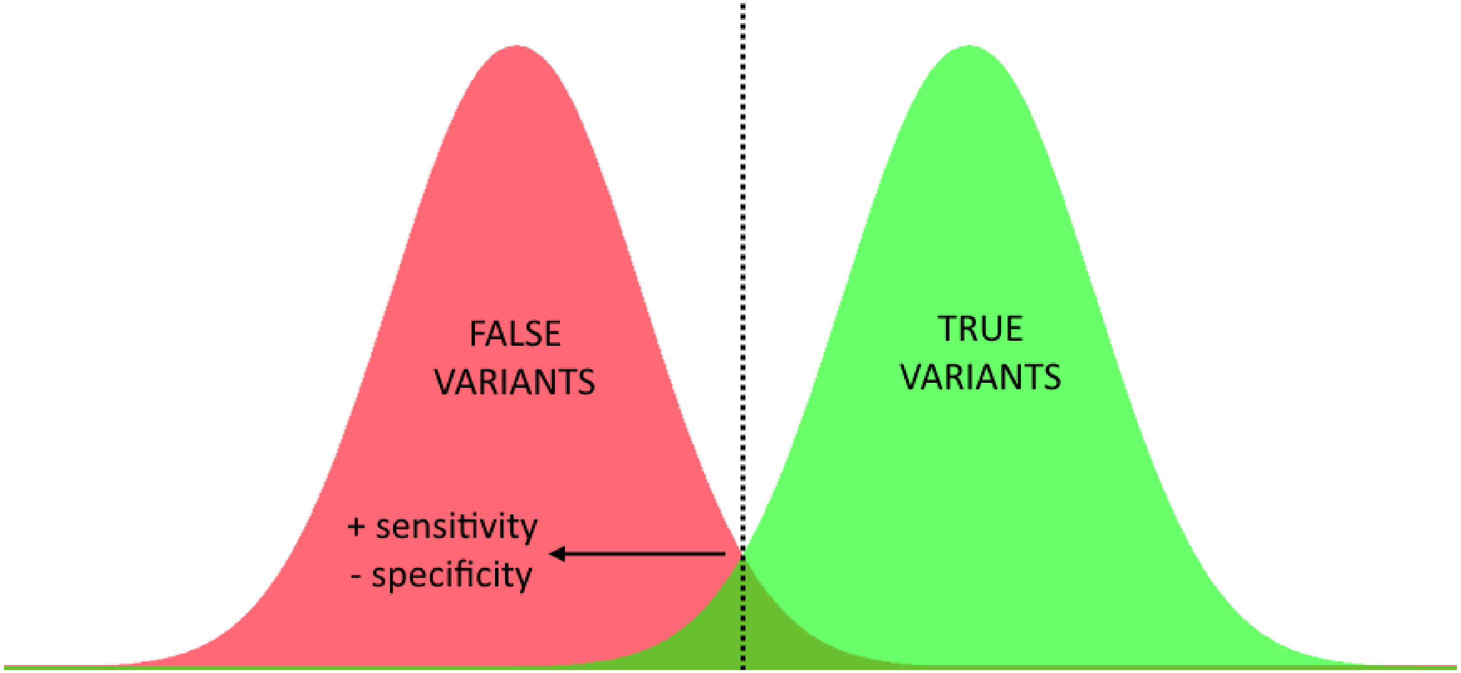
Motivation

Sensitivity and specificity of clinical tests are always negatively correlated.

In diagnostics, sensitivity is very important, which comes at the cost of lost specificity, i.e. many false-positive variants (artefacts).

Artefacts need to be recognized to avoid wrong diagnosis

> the bioinformatics pipeline and possible sources of errors have to be known!



Overview

Part 1: Basics

- NGS library preparation
- Illumina sequencing
- Raw data (FASTQ format)

Part 2: Analysis pipeline

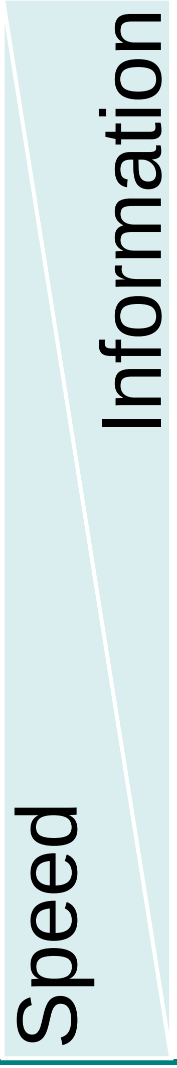
- Mapping
- Variant calling
- Variant annotation
- Variant filtering

Part 3: Quality control

- Run QC
- Sample identity
- Sample QC
- Variant QC

Genom vs. Exom vs. Panel

Comparison genome and targeted sequencing:

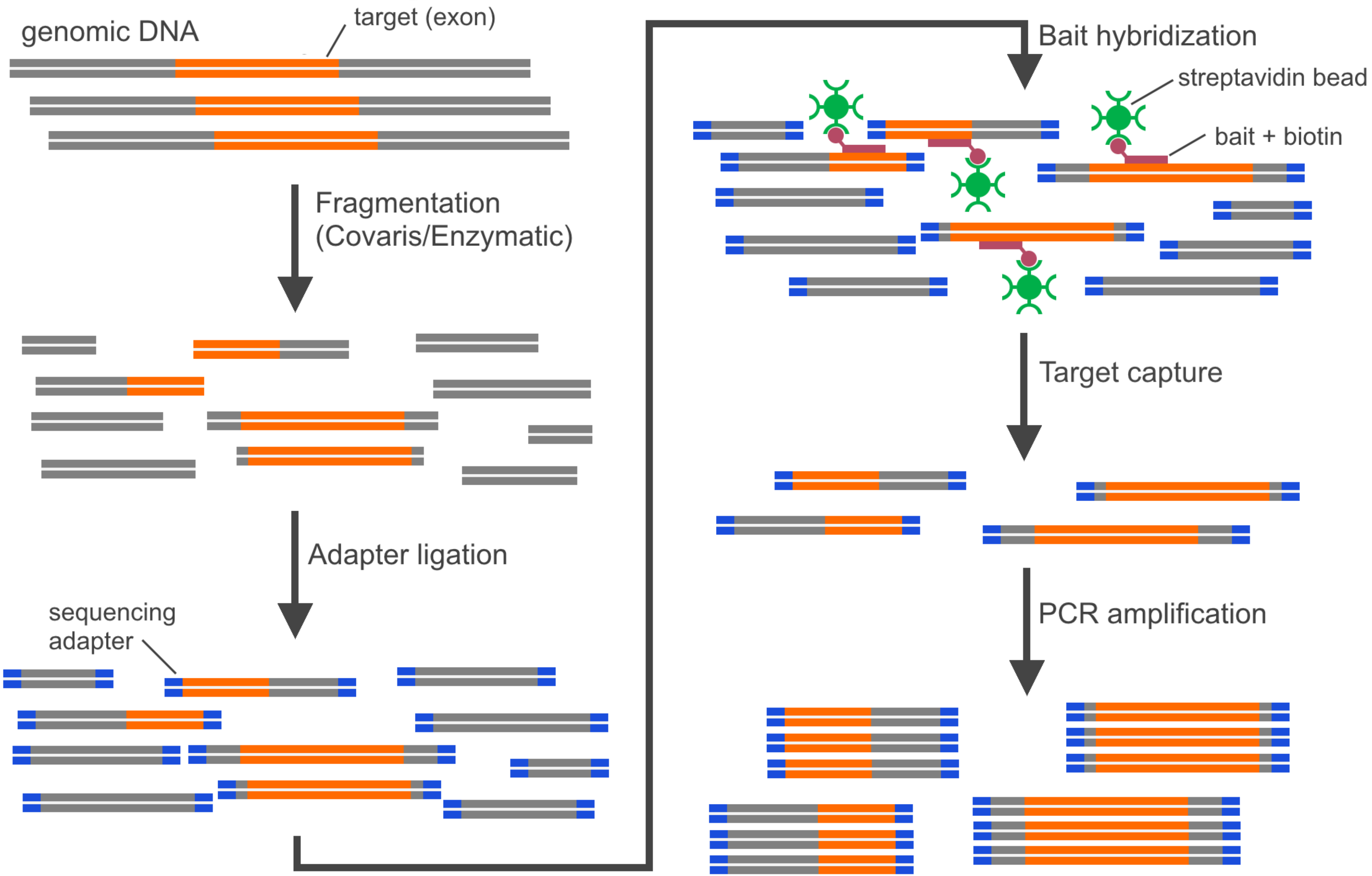


Information

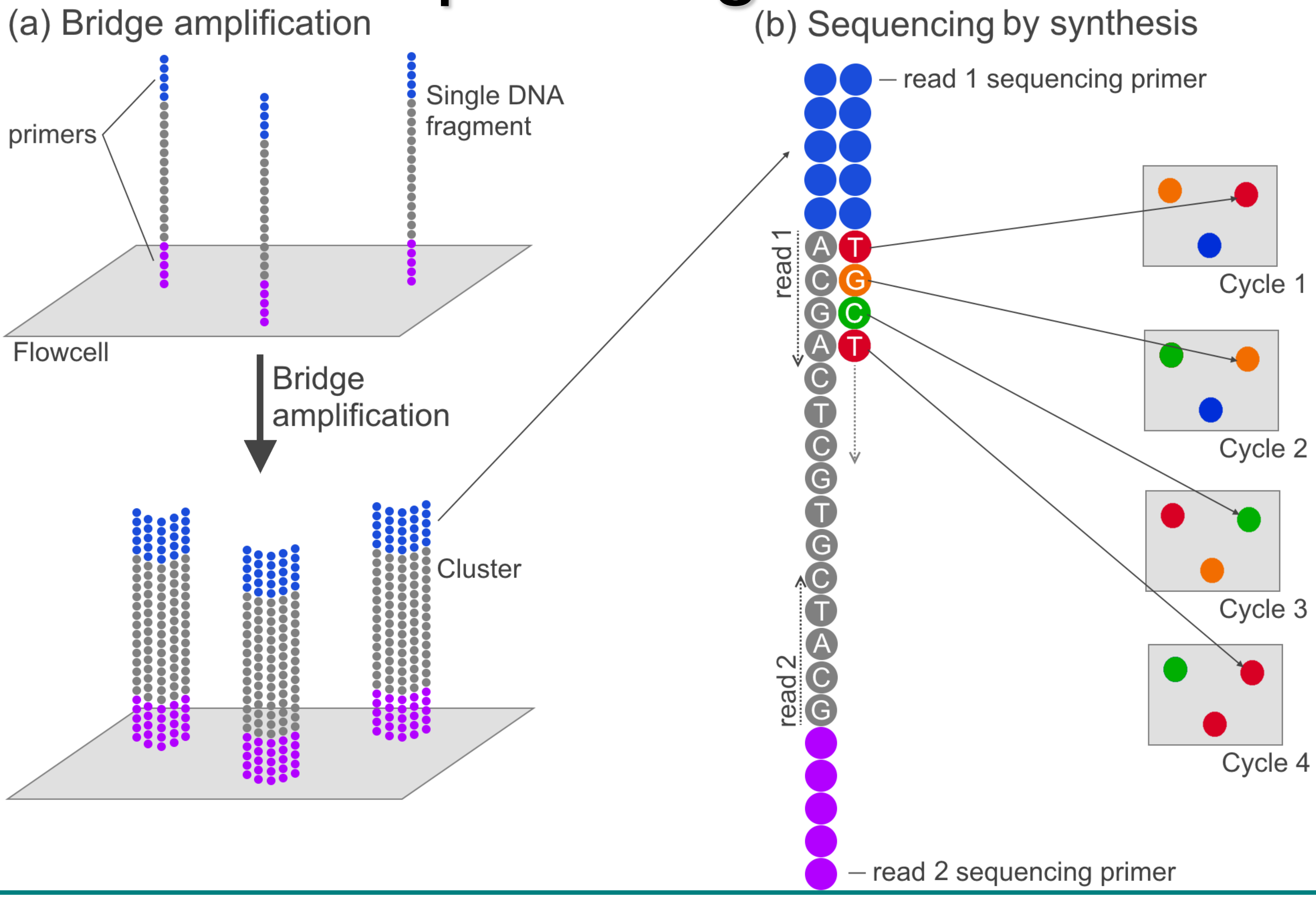
Speed

	Size	Pro	Contra
Genome	3.1 GB	- Structural variants - High resolution of CNVs	- Sequencing cost - Non-coding variants difficult to interpret
Exome	~47 MB	- Less than 1.5% of genome - All exons and splice regions	- No structural variants
Clinical Exome	~16 MB	- Less than 30% of exome	- Only variants in known disease genes, thus no disease gene discovery - Needs update when new disease genes are discovered
Panel	< 1 MB	- Typically less than 5% of clinical exome - Very fast and cost-efficient	- Needs update when new disease genes are discovered

NGS library preparation (exome)



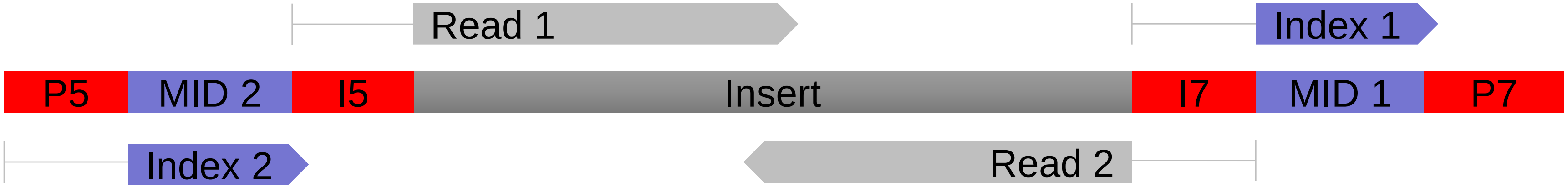
Illumina sequencing



Illumina sequencing - indices

In most sequencing runs **index reads** are generated in addition to the actual insert reads. Index reads are used to assign the randomly placed reads on a flowcell lane to the corresponding sample (**demultiplexing**).

The adapter layout for a typical Illumina library including primer binding sites is shown here:



The order in which the reads are sequenced is normally:

- Read1
- Index1 (optional - MID needed for de-multiplexing)
- Index2 (optional - second MID [index hopping] or UMI [tumor])
- Read2 (optional - helpful for InDels and SVs)

The raw data (FASTQ)

The FASTQ format is similar to the well-known FASTA format, but contains quality information (indicated by the Q at the end) in addition to the sequence.

Raw data:

```
AATTAAAGTCAGCTACAAATGACTTGCCAGTGTCTTCAA  ———————  Read 1
#++2+-*+++@@@@@177/5@@@@@@7@@@33/337877  ———————  Qualities read 1

AAGAAAGTAAAGAATATTCTTGGTAGCTAAGCATTATAT  ———————  Read 2
DH@IIGII<I@BGG;IIFBIGBD:@GEEDEE@D>E>GGG  ———————  Qualities read 2
```

Quality scores:

ASCII characters (‚!‘-‘J‘) encode for Phred scores (0-41), which represent the error probability of a base call ($P = 10^{\frac{-Q}{10}}$):

10 = 10%

20 = 1%

30 = 0.1%

40 = 0.01%

Understanding Q-scores is important!
They are also used for mapping quality and variant quality.

Overview

Part 1: Basics

- NGS library preparation
- NGS sequencing
- Raw data (FASTQ format)

Part 2: Analysis pipeline

- Mapping
- Variant calling
- Variant annotation
- Variant filtering

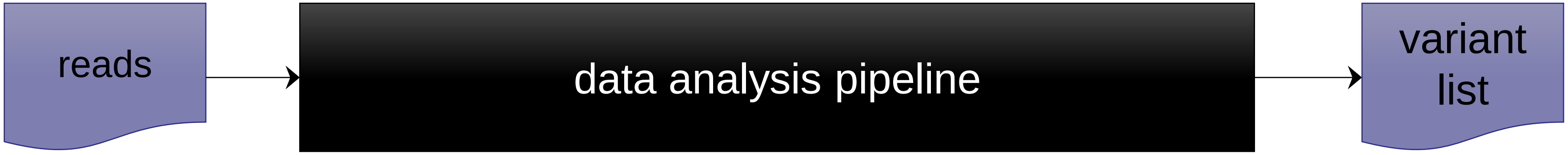
Part 3: Quality control

- Run QC
- Sample identity
- Sample QC
- Variant QC

Blackbox view of the data analysis

Input:

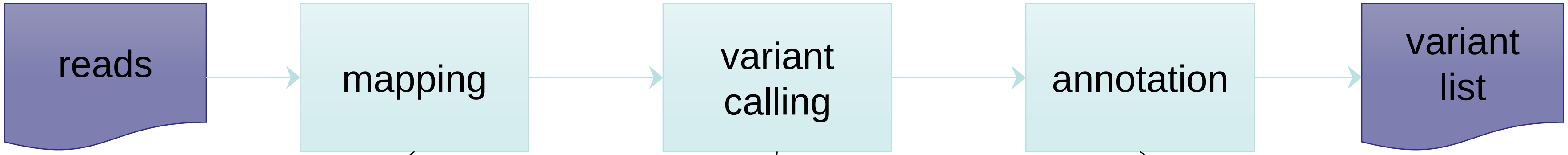
```
AATTAAAGTCAGCTACAAATGACTTGCCAGTGTCTTCAA  ———  Read 1
#++2+-*+++@@@@@177/5@@@@@@7@@@33/337877  ———  Qualities read 1

AAGAAAGTAAAGAATATTCTTGGTAGCTAAGCATTATAT  ———  Read 2
DH@IIGII<I@BGG;IIFBIGBD:@GEEDEE@D>E>GGG  ———  Qualities read 2
```



Output:

#chr	start	end	ref	obs	genotype	gene	dbSNP
chr1	871159	871159	G	A	het	SAMD11	
chr1	881627	881627	G	A	het	NOC2L	rs2272757
chr1	887801	887801	A	G	hom	NOC2L	rs3828047
chr1	888639	888639	T	C	hom	NOC2L	rs3748596
chr1	888659	888659	T	-	hom	NOC2L	rs3748597
chr1	894573	894573	G	A	hom	NOC2L	rs1330301
chr1	897325	897325	G	C	hom	KLHL17	rs4970441

Overview analysis pipeline



reads → mapping → variant calling → annotation → variant list

Assignment of reads to the corresponding position of the **reference genome.**

Determine deviations from the reference genome:

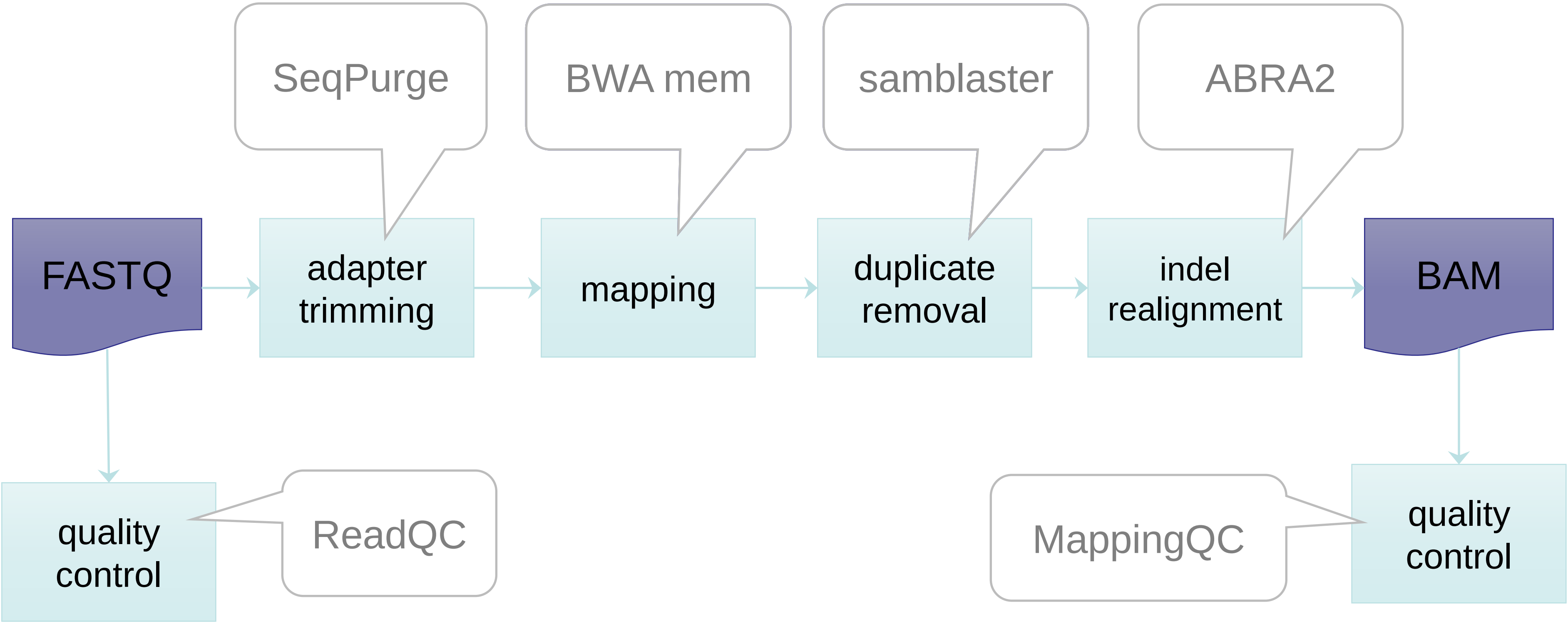
- Single nucleotide variants (SNVs)
- Insertions and deletions (indels)
- Copy-number alterations (CNAs)
- Structural variants (SVs)

Annotate variant list with information from various **databases**:

- cDNA/protein change
- allele frequencies
- conservedness
- pathogenicity predictions
- clinical databases

Mapping - details

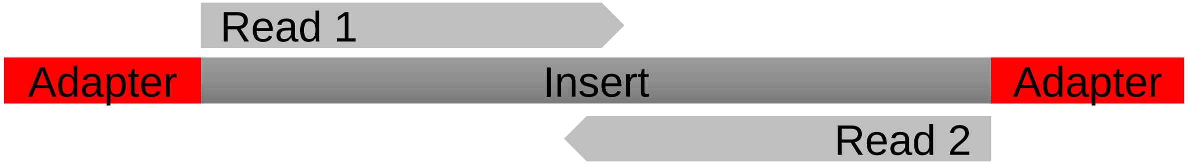
Mapping transforms the raw read data (**FASTQ format**) to reads mapped to the reference genome (**BAM format**). Besides the actual read mapping, several additional steps are involved:



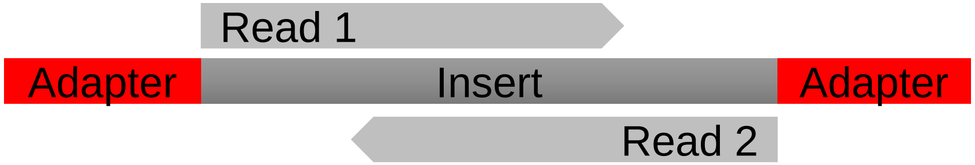
Mapping - adapter trimming

Adapter contamination occurs if the **insert length is smaller than the read length**, see case (c). Adapter sequences should be removed from the reads, because they can **lead to incorrect mapping** and thereby to **false-positive variants calls**.

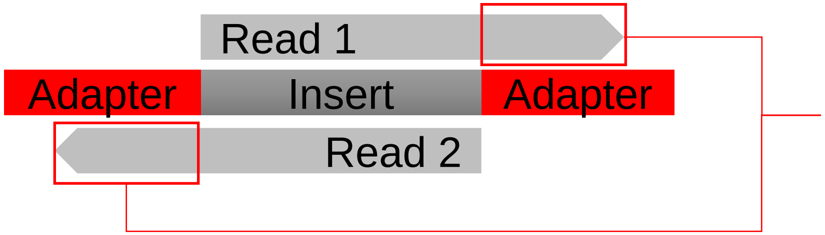
(a) Large insert: no overlap, no adapter contamination



(b) Medium insert: partial read overlap, no adapter contamination



(c) Small insert: complete read overlap, adapter contamination

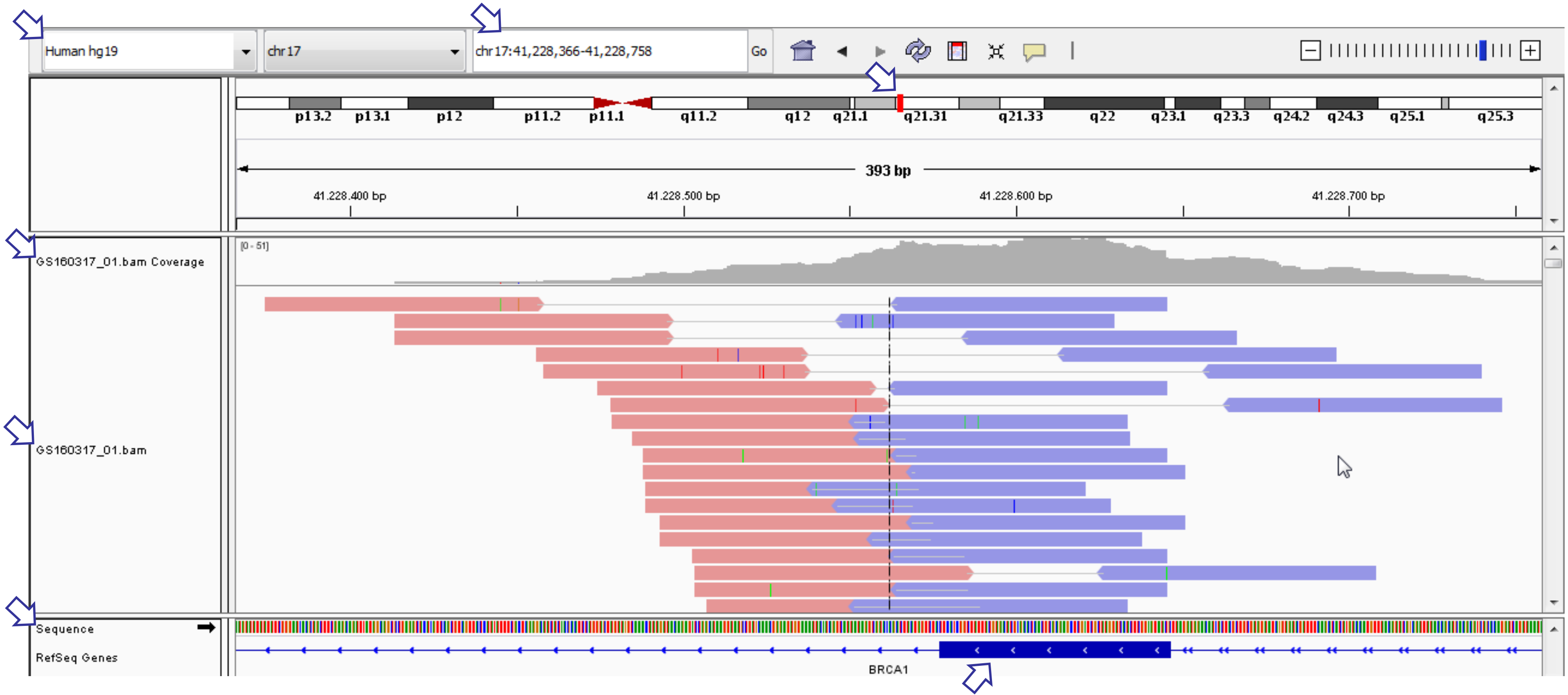


Adapter contamination that needs to be removed from the reads.

Mapping - mapping

During read mapping, each read is individually assigned to the corresponding position of the reference genome. The similarity/uniqueness of the fit are measured by the **mapping quality** (Q-score).

Sequence analysis based on a reference genome is called **re-sequencing**. This is computationally much easier than de-novo assembly of a genome without a reference genome.



Human hg19

chr17

chr17:41,228,366-41,228,758

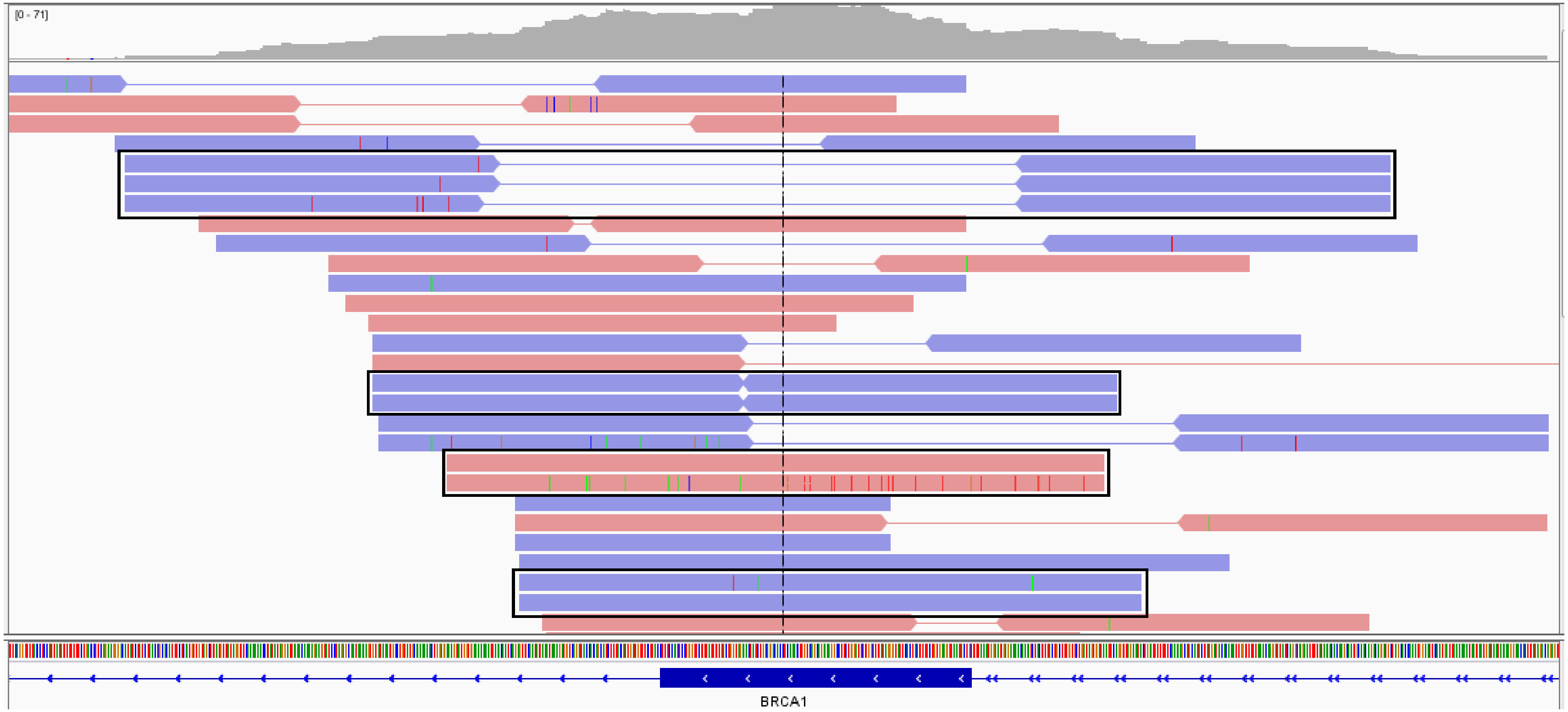
Mapping - duplicate removal

When **random DNA fragmentation** is performed during library preparation, duplicate reads (same start/end position) are likely PCR or are optical artefacts that represent only **one underlying DNA molecule.** Thus, duplicates are removed/marked.

Note:

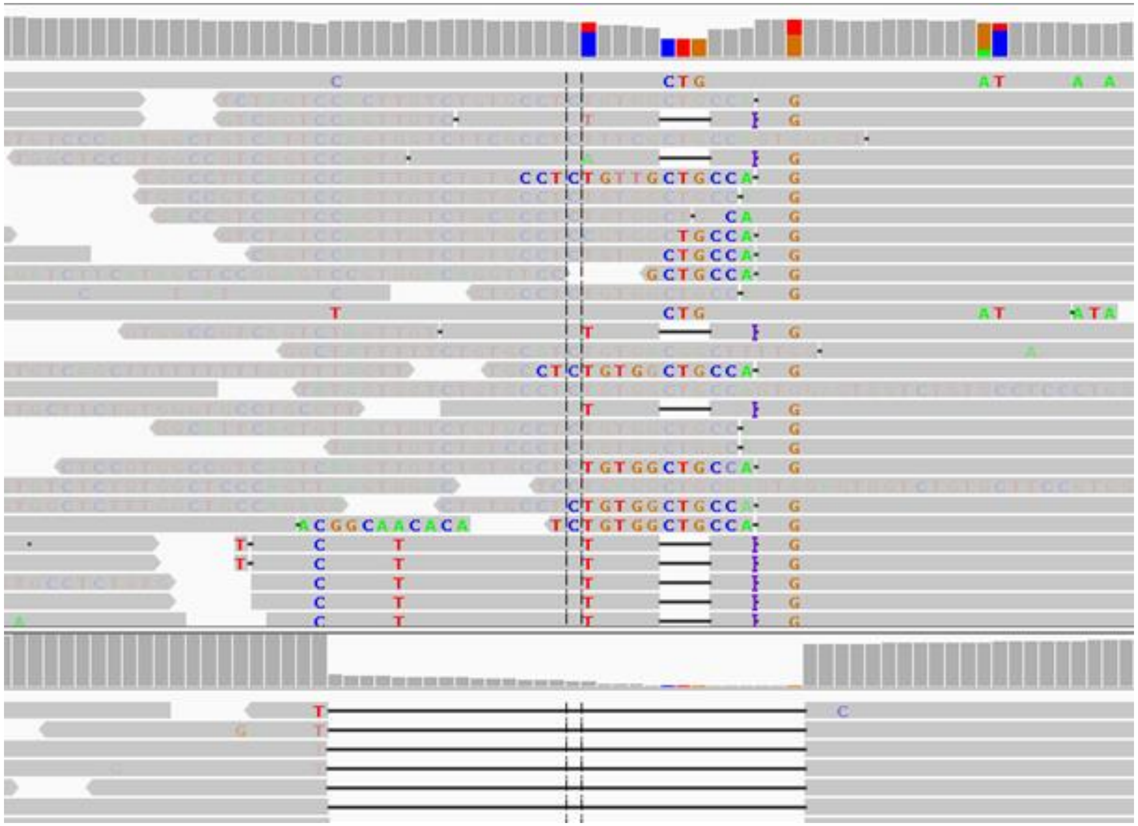
- Duplicate removal is not performed in amplicon-based panels.

- Enzymatic fragmentation is often not completely random.



Mapping - indel realignment

Larger insertions/deletions (indels) are difficult to map. During the initial mapping **each read is individually mapped** to the reference genome. During indel realignment regions with excessive mismatches, are re-aligned taking all reads at the locus into account. This 30bp deletion illustrates the problem:

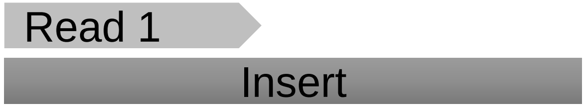


Screenshot from ABRA website:
https://github.com/mozack/abra

Single-end vs. paired-end

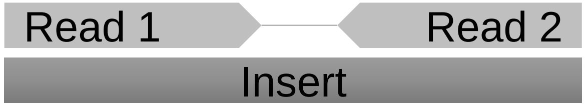
PRO single-end:

- No read overlap, which gives no additional information



PRO paired-end:

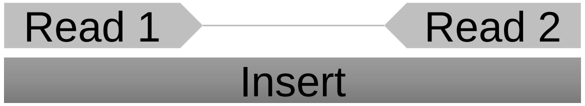
- More reads mappable, because one read can make the other uniquely mappable
- **Less duplicates**, because there are more possibilities when taking both ends into consideration for duplicate-removal
- Easier detection of structural variants (genome, unlikely in exome)



Longer vs. shorter reads

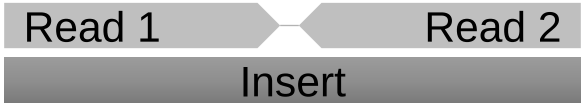
PRO shorter reads:

- Less overlapping reads (also depends on insert size)
- Less sequencing errors, which increase with read length



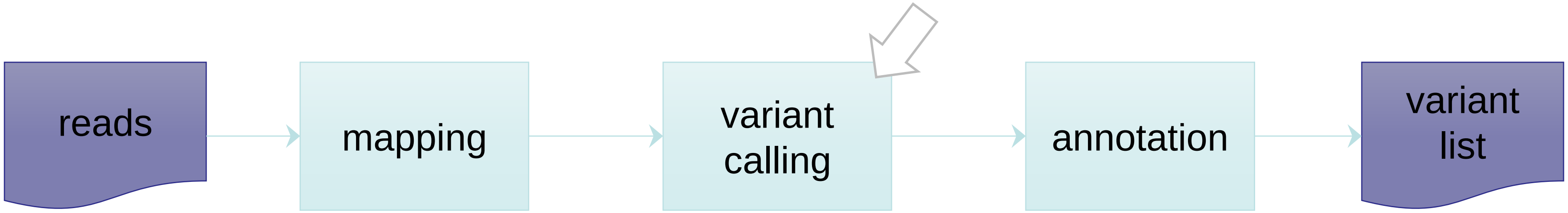
PRO longer reads:

- More uniquely mappable reads
- **Improved calling of indels in repeat regions**

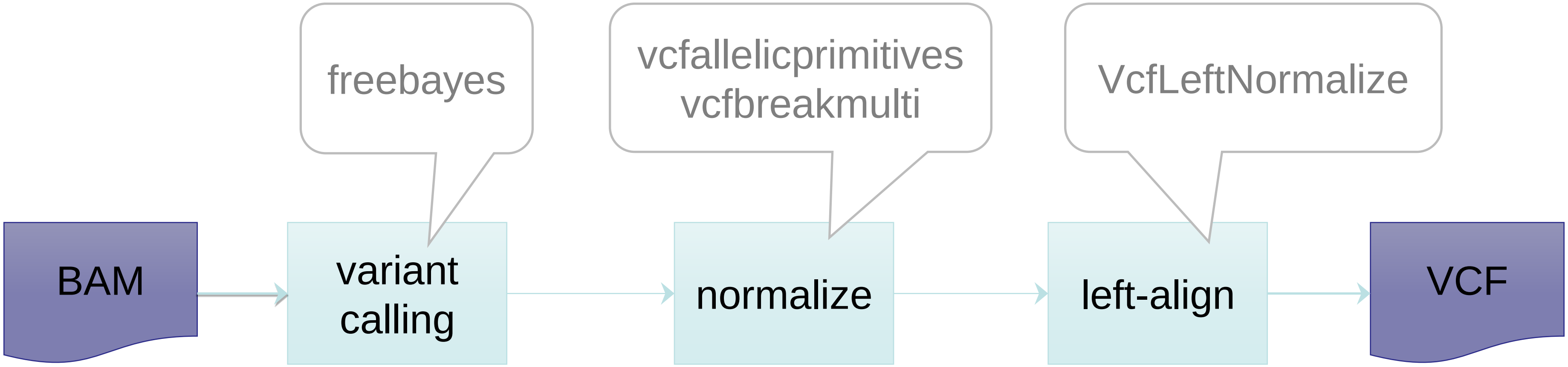


Note: For genome/exome re-sequencing, the consensus currently is 100bp paired-end sequencing.

Variant calling - details

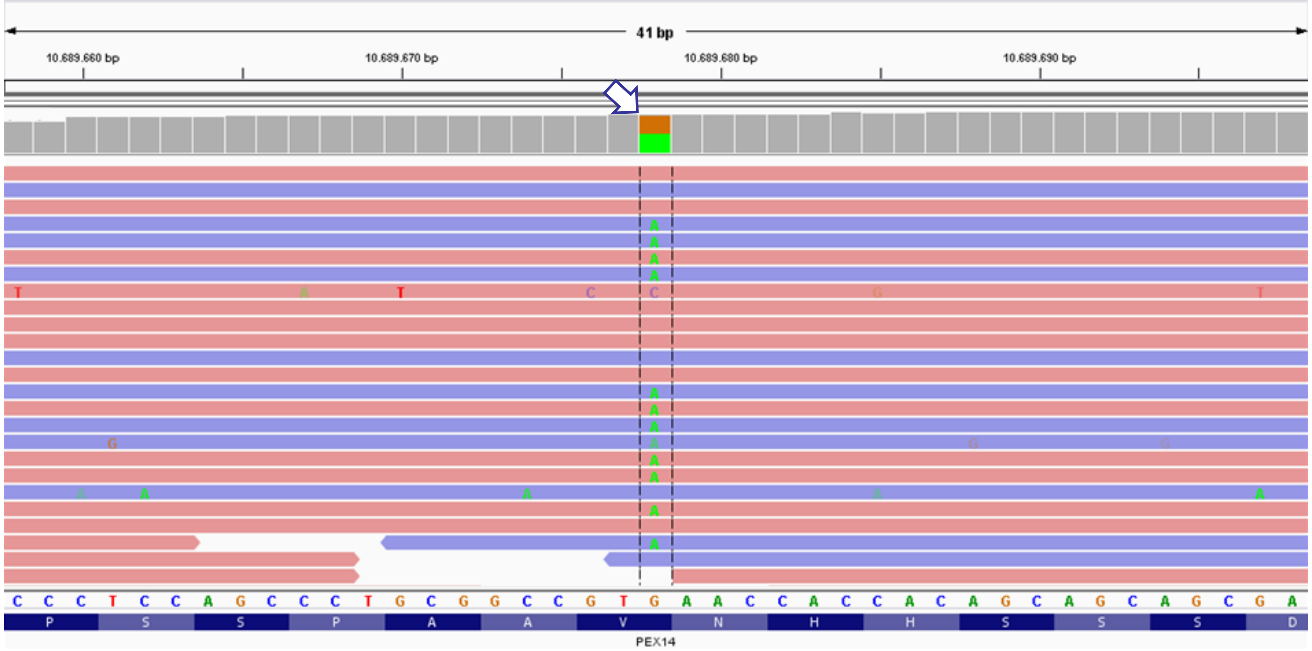


During variant calling, **deviations from the reference genome** are detected and stored in a VCF file. After the initial variant calling, several post-processing steps can be performed to normalize the variants:

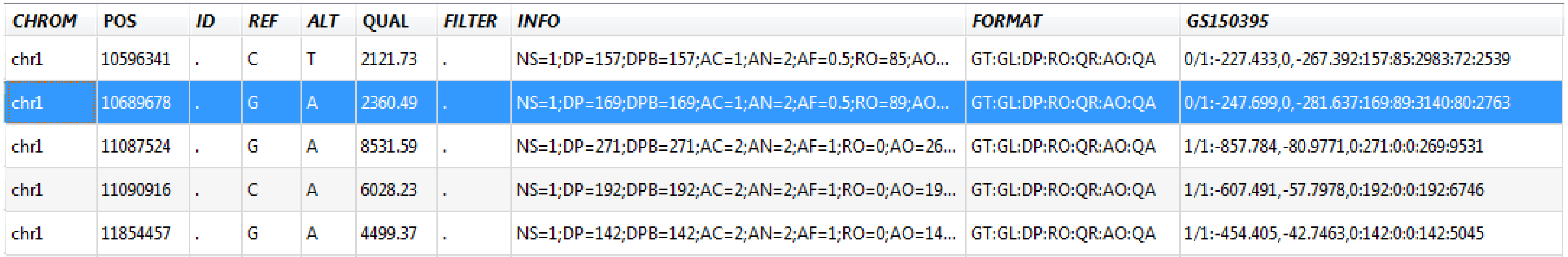


Variant calling

During variant calling, SNVs and small indels are called and stored in VCF format.


41 bp
10.689.660 bp
10.689.670 bp
10.689.680 bp
10.689.690 bp
C C C T C C A G C C C T G C G G C C G T G A A C C A C C A C A G C A G C A G C G A
P S S P A A V N H H S S S D
PEX14

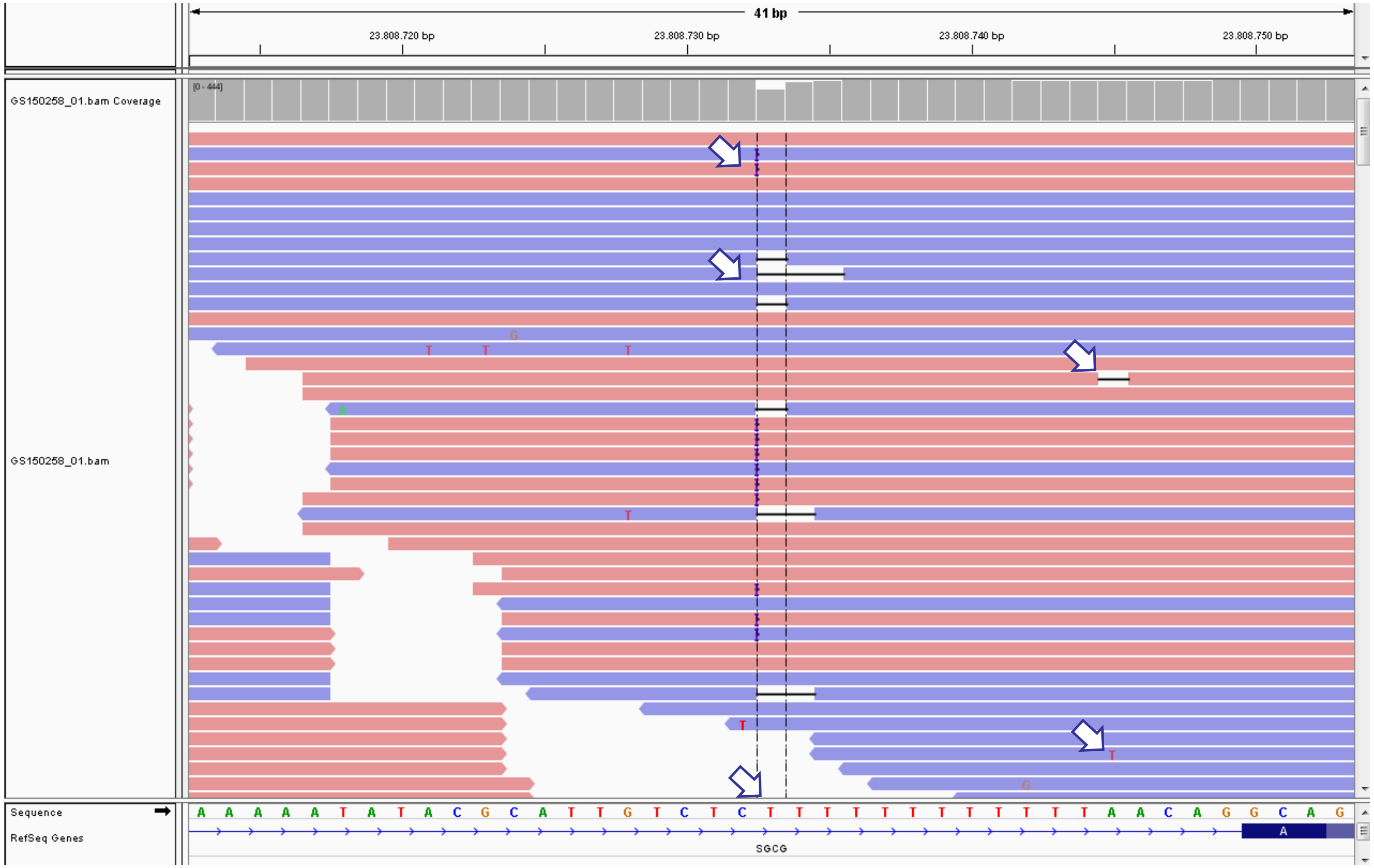
VCF format example (http://www.1000genomes.org/wiki/Analysis/vcf4.0)



CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GS150395
chr1	10596341	.	C	T	2121.73	.	NS=1;DP=157;DPB=157;AC=1;AN=2;AF=0.5;RO=85;AO...	GT:GL:DP:RO:QR:AO:QA	0/1:-227.433,0,-267.392:157:85:2983:72:2539
chr1	10689678	.	G	A	2360.49	.	NS=1;DP=169;DPB=169;AC=1;AN=2;AF=0.5;RO=89;AO...	GT:GL:DP:RO:QR:AO:QA	0/1:-247.699,0,-281.637:169:89:3140:80:2763
chr1	11087524	.	G	A	8531.59	.	NS=1;DP=271;DPB=271;AC=2;AN=2;AF=1;RO=0;AO=26...	GT:GL:DP:RO:QR:AO:QA	1/1:-857.784,-80.9771,0:271:0:0:269:9531
chr1	11090916	.	C	A	6028.23	.	NS=1;DP=192;DPB=192;AC=2;AN=2;AF=1;RO=0;AO=19...	GT:GL:DP:RO:QR:AO:QA	1/1:-607.491,-57.7978,0:192:0:0:192:6746
chr1	11854457	.	G	A	4499.37	.	NS=1;DP=142;DPB=142;AC=2;AN=2;AF=1;RO=0;AO=14...	GT:GL:DP:RO:QR:AO:QA	1/1:-454.405,-42.7463,0:142:0:0:142:5045

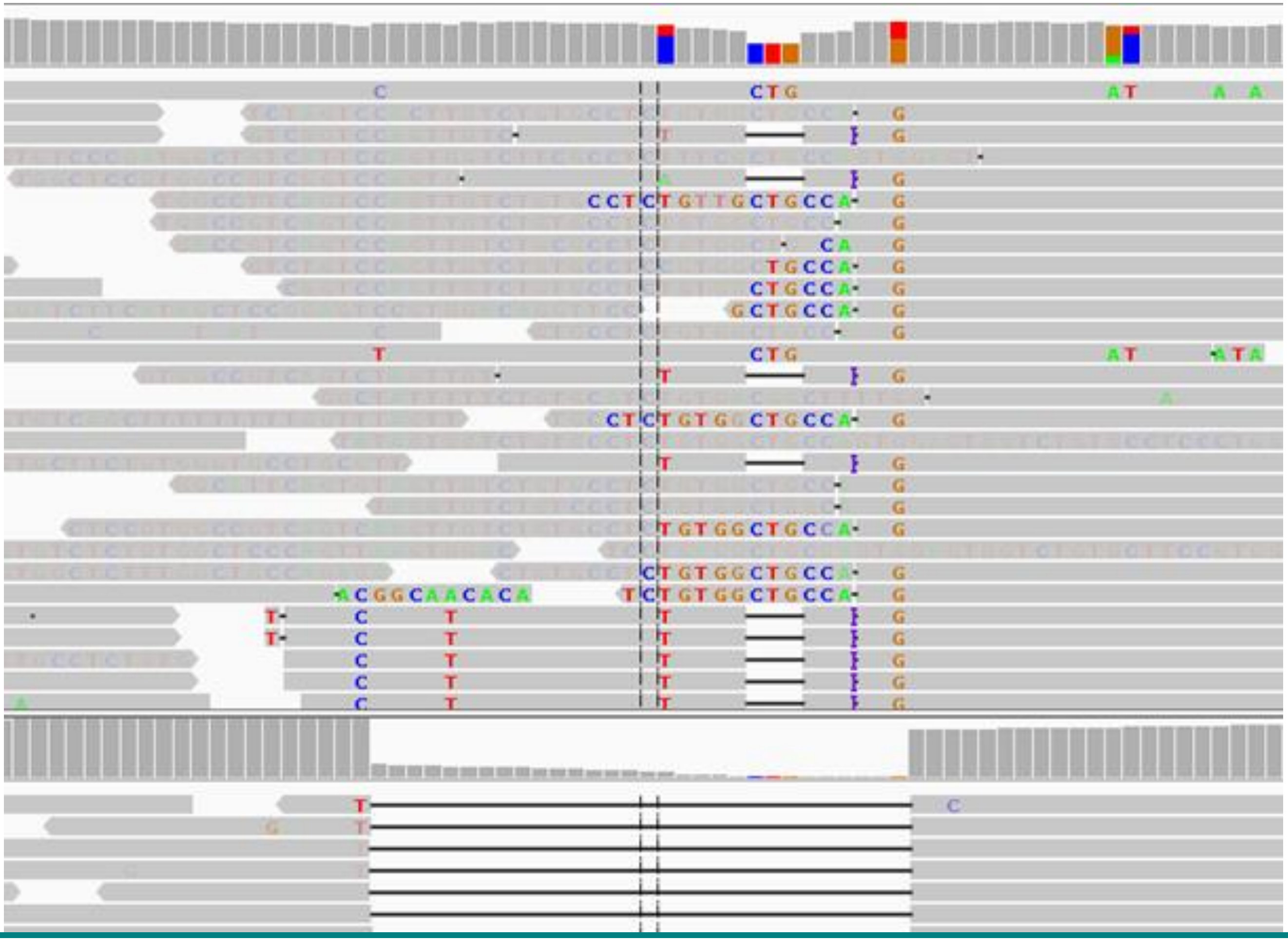
Variant calling - repeat problems

Single-base repeats lead to sequencing errors, which can cause false-positive variant calls.



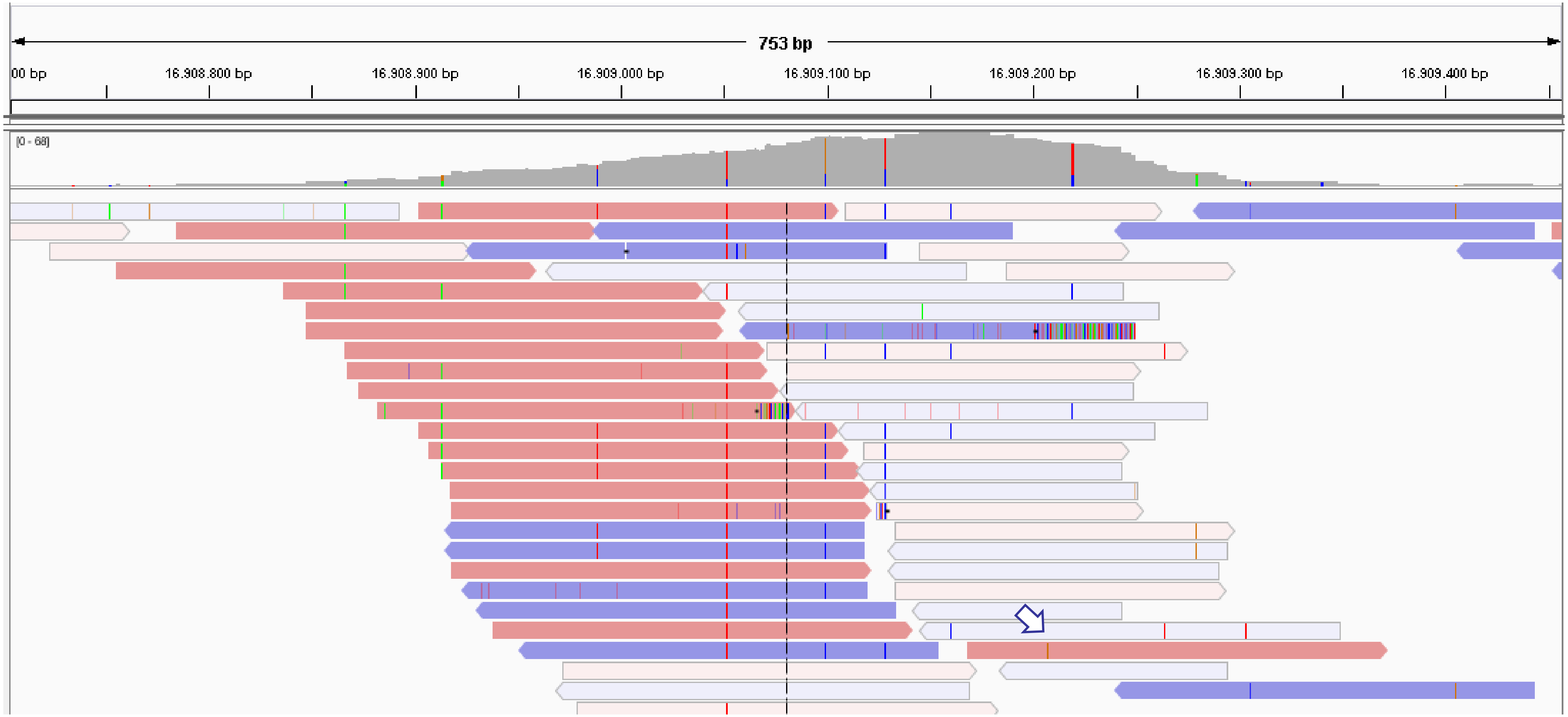
Variant calling - alignment problems

Incorrect alignments can lead to false-positive variant calls while missing the actual variants. Without indel realignment, this 30 bp deletion, would be called as several small SNVs and deletion.



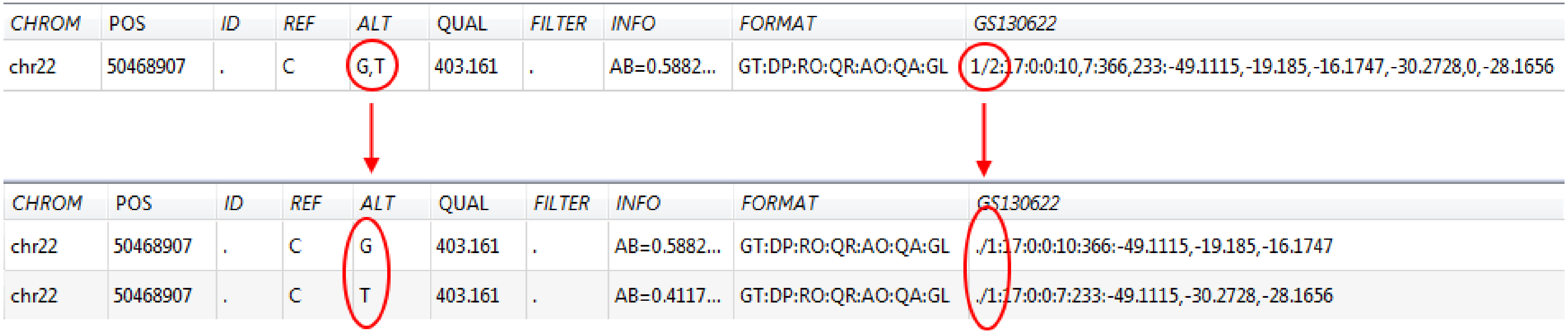
Variant calling - reference problems

Errors in the reference genome also lead to problems during variant calling. A common problem is that one of two homologous regions is missing in the reference genome. Then, reads from both regions are mapped to one locus, which typically results in several heterozygous SNVs with an allele frequency around 25%.



Variant calling - normalization

Several variants at the same genomic position are called as a **single multi-allelic variant** by some variant callers. To make variant lists from different variant callers comparable and to facilitate left-alignment and annotation of variant lists, these variants should be **split into several variants**:



CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GS130622
chr22	50468907	.	C	G,T	403.161	.	AB=0.5882...	GT:DP:RO:QR:AO:QA:GL	1/2:17:0:0:10,7:366,233:-49.1115,-19.185,-16.1747,-30.2728,0,-28.1656

CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	GS130622
chr22	50468907	.	C	G	403.161	.	AB=0.5882...	GT:DP:RO:QR:AO:QA:GL	./1:17:0:0:10:366:-49.1115,-19.185,-16.1747
chr22	50468907	.	C	T	403.161	.	AB=0.4117...	GT:DP:RO:QR:AO:QA:GL	./1:17:0:0:7:233:-49.1115,-30.2728,-28.1656

Format field descriptions (freebayes):
GT: Genotype (0/.=REF, 1=ALT1, 2=ALT2, …)
DP: Total read depth at the locus
RO: Reference allele observation count
QR: Reference allele quality sum in Phred
AO: Alternate allele observations
QA: Alternate allele quality sum in Phred
GL; Genotype Likelihood, log10-scaled likelihoods of the data given the called genotype for each possible genotype generated from the reference and alternate alleles given the sample ploidy

Variant calling - left-alignment

For indels, several valid alignments can be possible. Currently, the consensus is to shift them to the leftmost position, i.e. to the lowest genomic coordinate, to facilitate annotation of variants.

Example:

```
Ref (chrZ): ACATATATCGTGA
Read      : ACATATCGTGA
```

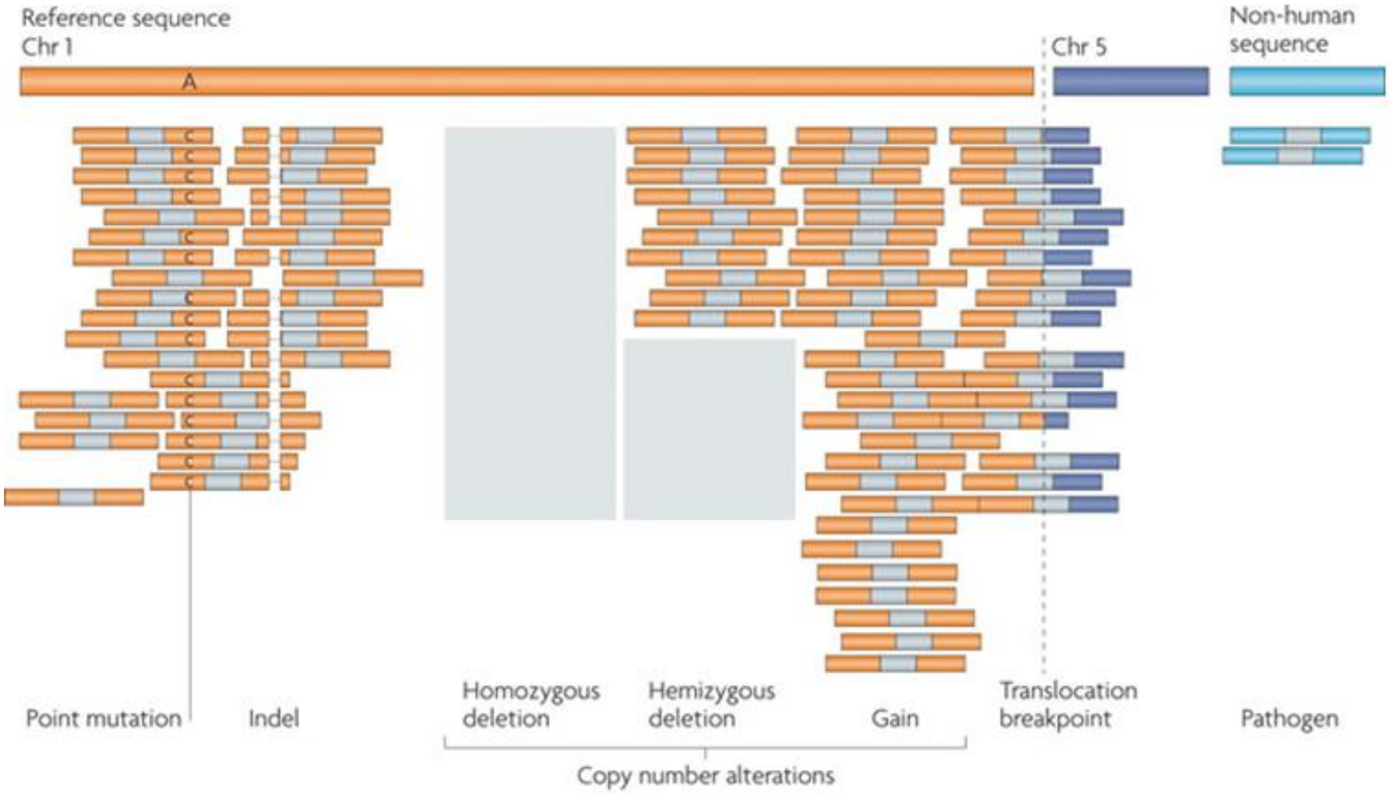
Possible valid alignments and variant calls:

```
Alignment1: ACATAT--CGTGA   chrZ:6 TAT>T
Alignment2: ACAT--ATCGTGA   chrZ:4 TAT>T
Alignment3: AC--ATATCGTGA   chrZ:2 CAT>C
```

If alignment 1 or 2 are called during variant calling, those alignments are converted to **alignment 3** during indel left-alignment.

Variants calling - SVs

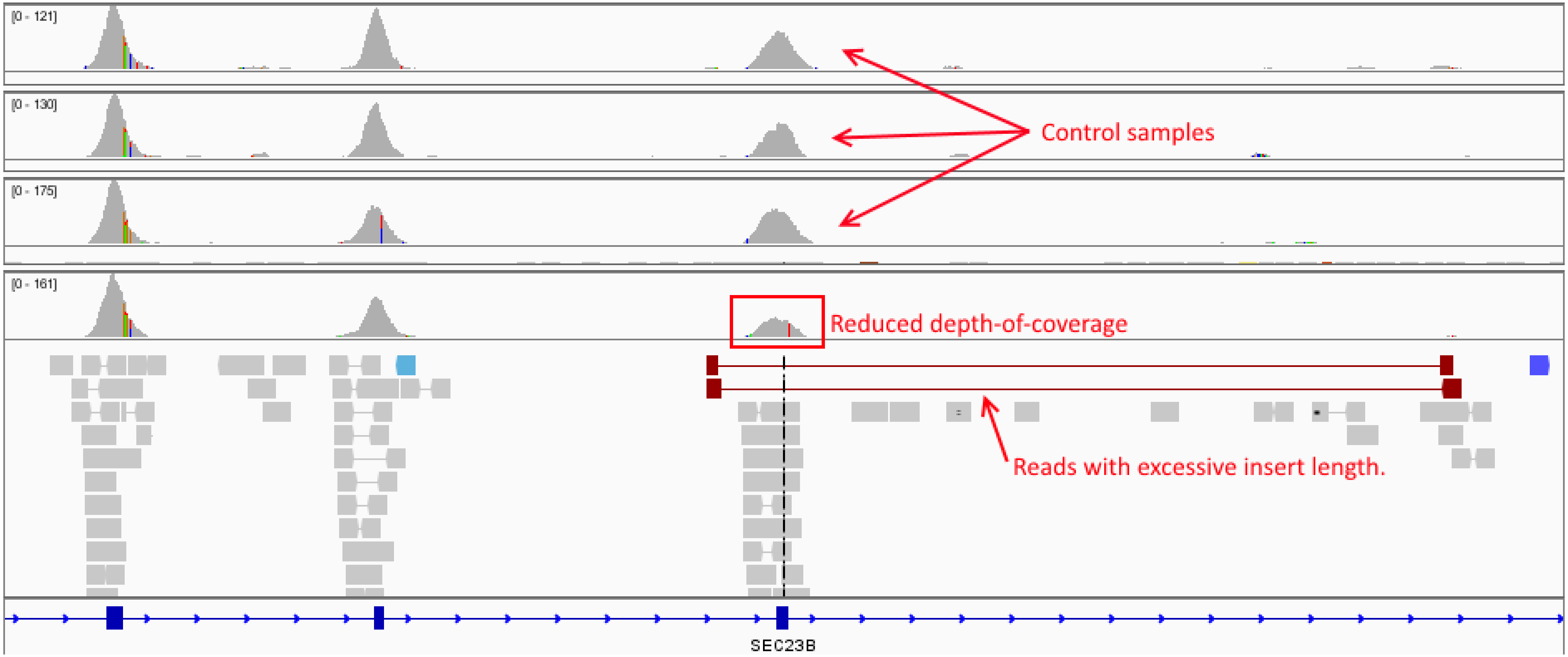
So far, we have only looked at small SNVs and InDel variants. However, for a complete genetic analysis, also large variants have to be taken into consideration. **Balaced structural variants** (inversions and translocations) can only be detected in **WGS** experiments (unless by chance a breakpoints lies inside the target region in exome/panel sequencing). **Copy-number variants** (deletions and gains) can also be detected in **exome/panel** based on depth-of-coverage analysis.



Source:
Nature Reviews Genetics
doi:10.1038/nrg2841

Structural variants - CNVs

The screenshot shows a heterzygous deletion in an exome, detected by a 50% reduction of the depth of coverage. In this case we can even determine the exact start/end of the deletion using the reads with excessive insert size (marked red).



Control samples

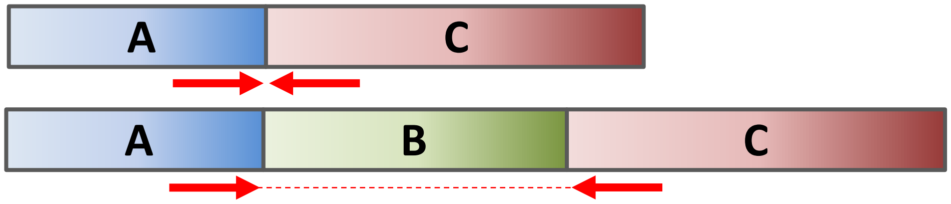
Reduced depth-of-coverage

Reads with excessive insert length.

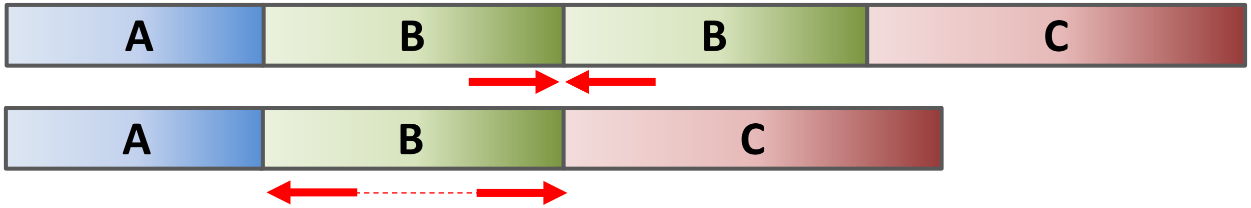
Note: depth-based CNV calling works reliably only if all samples are processed with the same procedure: dna extraction, library prep/enrichment, sequencing, mapping tools.

Structural variants – PE representation

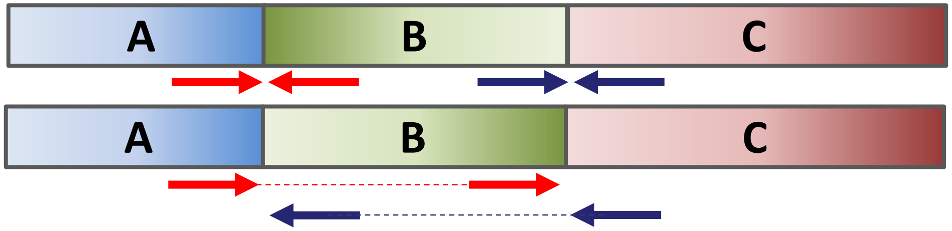
Deletion B



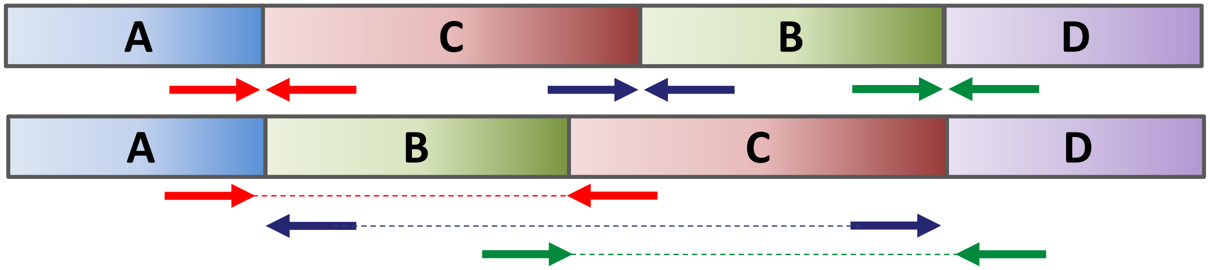
Duplication B (tandem)



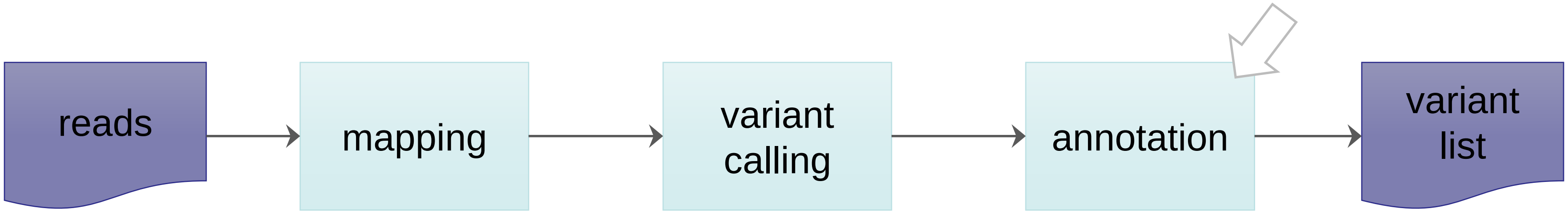
Inversion B

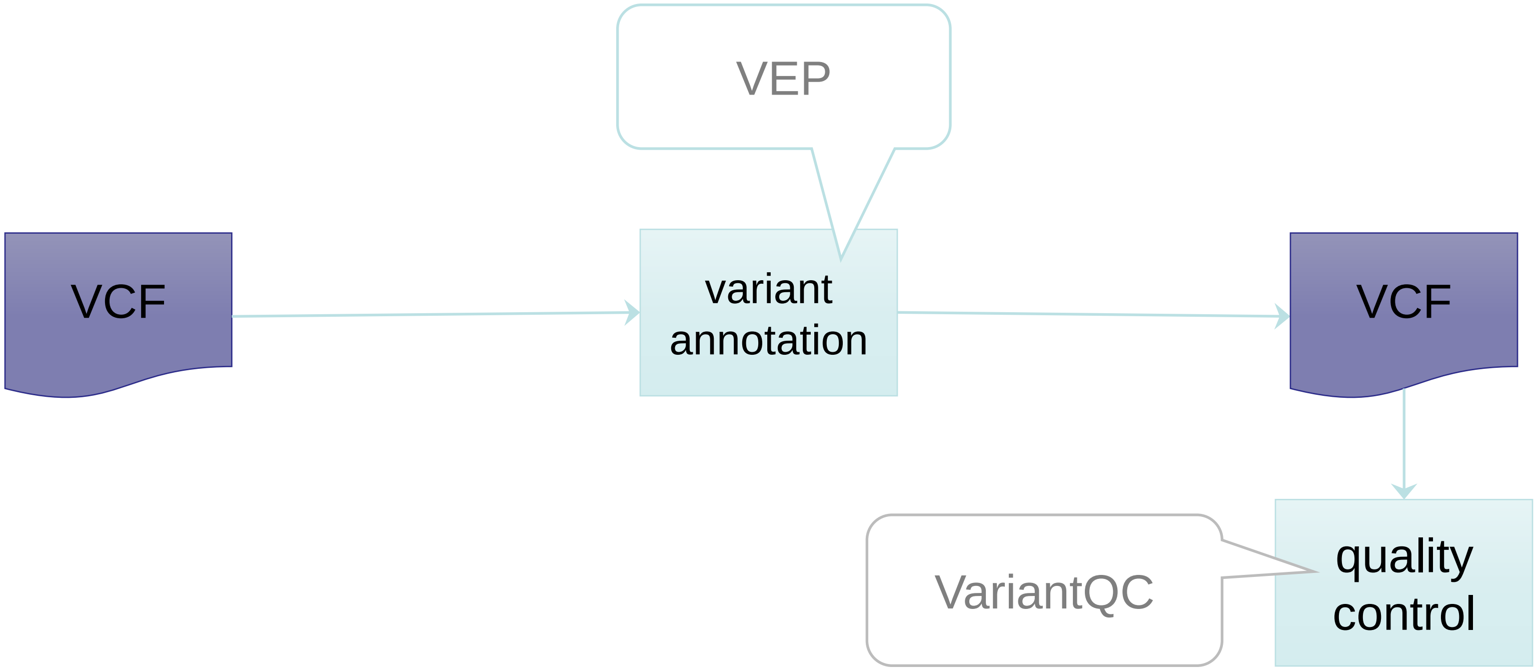


Translocation B



Annotation - details





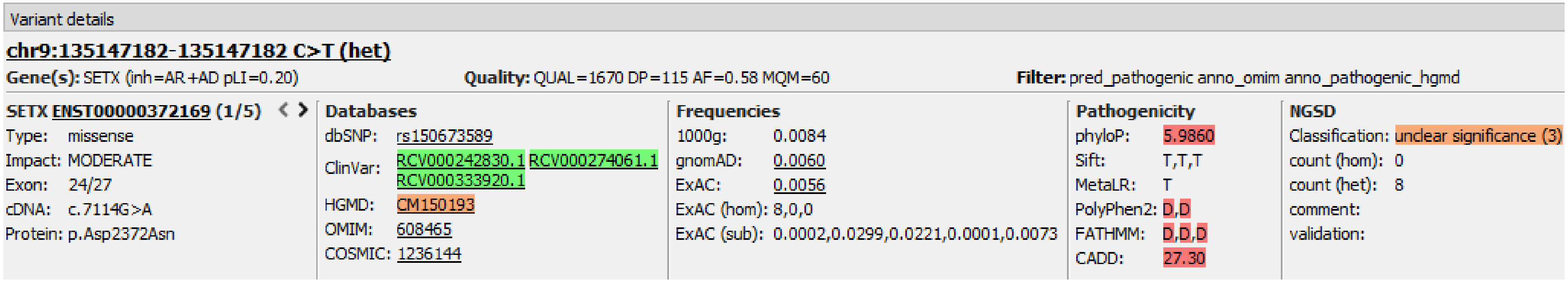
Annotation - example

During the annotation, variants are annotated with additional information from various databases and stored as VCF file again.

Commonly used annotation are:

- effect on cDNA/protein: Gene, Transcript, Type (missense, deletion, …),
- allele frequencies (dbSNP, 1000g, gnomAD)
- Conservedness (phyloP)
- pathogenicity predictions (Sift, PolyPhen2, MutationTaster2, CADD, FATHMM-MKL)
- clinical databases (ClinVar, OMIM, HGMD, COSMIC)

The screenshot below shows example variant details from our in-house variant analysis software GSvar:



Variant details

chr9:135147182-135147182 C>T (het)

Gene(s): SETX (inh=AR+AD pLI=0.20) **Quality:** QUAL=1670 DP=115 AF=0.58 MQM=60 **Filter:** pred_pathogenic anno_omim anno_pathogenic_hgmd

SETX ENST00000372169 (1/5) ‹ ›

Type:	missense
Impact:	MODERATE
Exon:	24/27
cDNA:	c.7114G>A
Protein:	p.Asp2372Asn

Databases

dbSNP:	rs150673589
ClinVar:	RCV000242830.1 RCV000274061.1 RCV000333920.1
HGMD:	CM150193
OMIM:	608465
COSMIC:	1236144

Frequencies

1000g:	0.0084
gnomAD:	0.0060
ExAC:	0.0056
ExAC (hom):	8,0,0
ExAC (sub):	0.0002,0.0299,0.0221,0.0001,0.0073

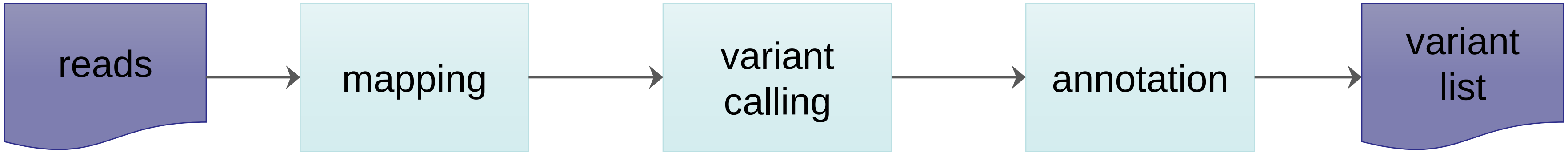
Pathogenicity

phyloP:	5.9860
Sift:	T,T,T
MetaLR:	T
PolyPhen2:	D,D
FATHMM:	D,D,D
CADD:	27.30

NGSD

Classification:	unclear significance (3)
count (hom):	0
count (het):	8
comment:	
validation:	

Statistics primary data analysis



Data reduction Exome (SureSelect Human All Exon v7):

Format	Size	Details
FASTQ	~4.7 GB	~87 million reads of 100bp (paired-end)
BAM	~4.0 GB	~94x depth on target region ~96% of target region at 20x (MAPQ=0 excluded) ~75% of reads on target
VCF	~9.0 MB	~62000 variants (92% SNVs, 8% indels) 2.6 transition/transversion ratio

Variant filtering - disease variants

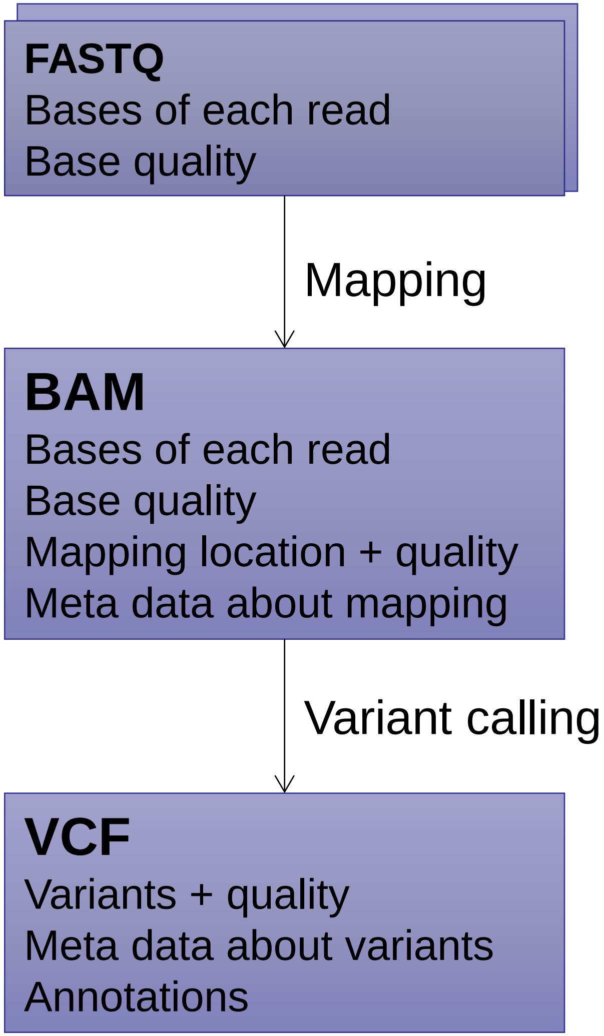
Finding putative disease-causing variants, requires several filtering steps based on various annotations.

Example variant filtering (SureSelect Human All Exon v7):

#variants	Filter
61893	None
1899	Allele frequency (<1% AF in 1000g/gnomAD)
505	Impact (Coding-change or consensus splice site)
238	IHDB (<20x with same genotype in-house DB) - removes pipeline-specific artefacts
221	Quality (depth>20, mapping-quality>50, variant-quality>30)
0-30	Mode of inheritance (dominant/recessive) Phenotype-specific target region (e.g. via HPO)

Note: Technical filters (AF, Impact, IHDB, Quality) cannot reduce the variant list of an exome/genome to a manageable size. Disease or inheritance information is needed!

Overview file formats



FASTQ
Bases of each read
Base quality

Mapping

BAM
Bases of each read
Base quality
Mapping location + quality
Meta data about mapping

Variant calling

VCF
Variants + quality
Meta data about variants
Annotations

FASTQ
Text format, normally zipped
https://en.wikipedia.org/wiki/FASTQ_format

Binary Alignment/Map
Compressed version of SAM text format
https://en.wikipedia.org/wiki/SAM_(file_format)

Variant Call Format
Text format, normally zipped
https://en.wikipedia.org/wiki/Variant_Call_Format

Overview

Part 1: Basics

- NGS library preparation
- Illumina sequencing
- Raw data (FASTQ format)

Part 2: Analysis pipeline

- Mapping
- Variant calling
- Variant annotation
- Variant filtering

Part 3: Quality control

- Run QC
- Sample identity
- Sample QC
- Variant QC

Sources of errors

1. Sample swaps
   - by the sender of the sample
   - during in-house sample processing

2. Problems during sequencing (sample)
   - bad DNA quality
   - bad sample prep kit quality
   - bad sequencing chemistry quality

3. Problems during data analysis (variant)
   - alignment problems around indels
   - errors in reference genome

For diagnostics, we need extensive quality control!

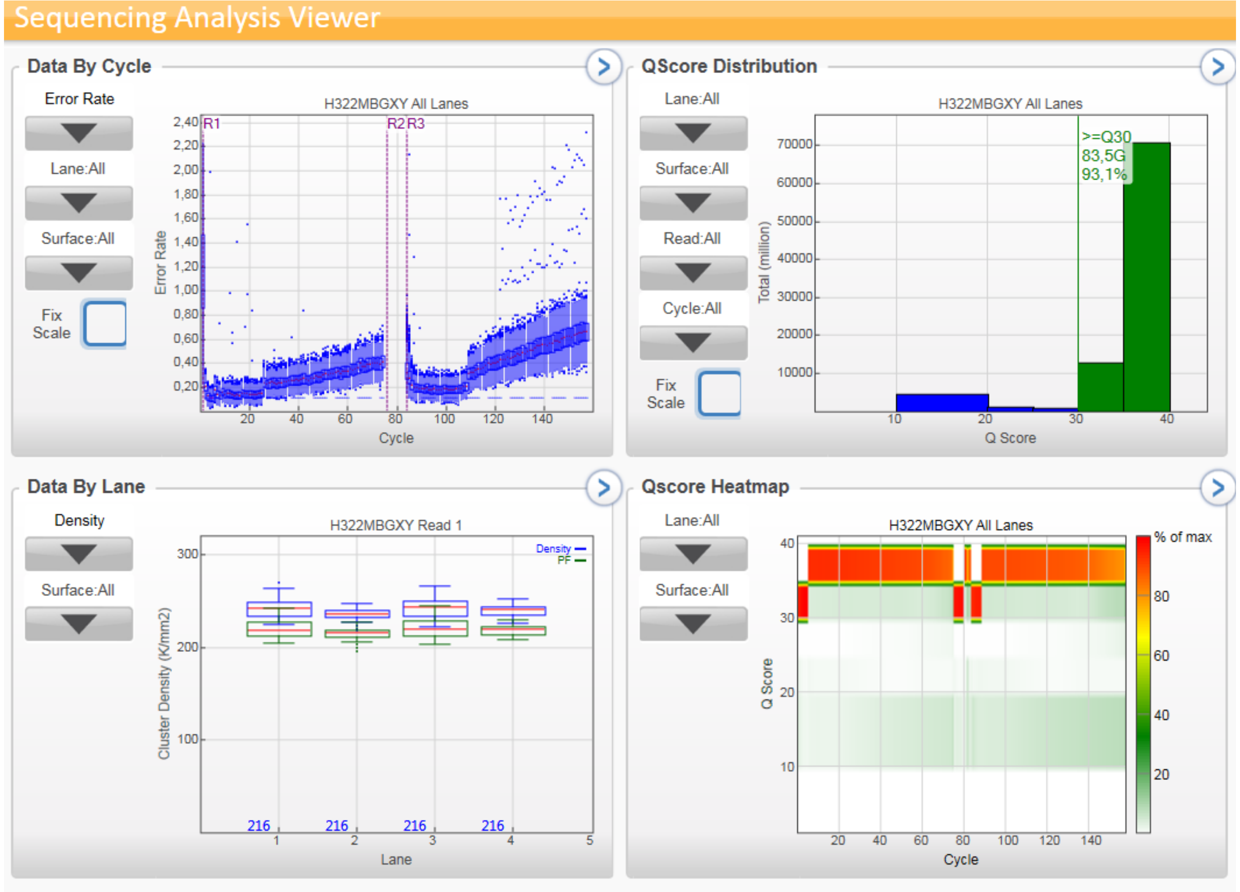
Run QC

⇨ Run QC
- Sample identity
- Sample QC
- Variant QC

Run QC (per lane)

The first QC step is the run QC using the Illumina Sequence Analysis Viewer:

- error rate (PhiX spike-in)
- Q-score distribution
- cluster density
- Q-score by cycle



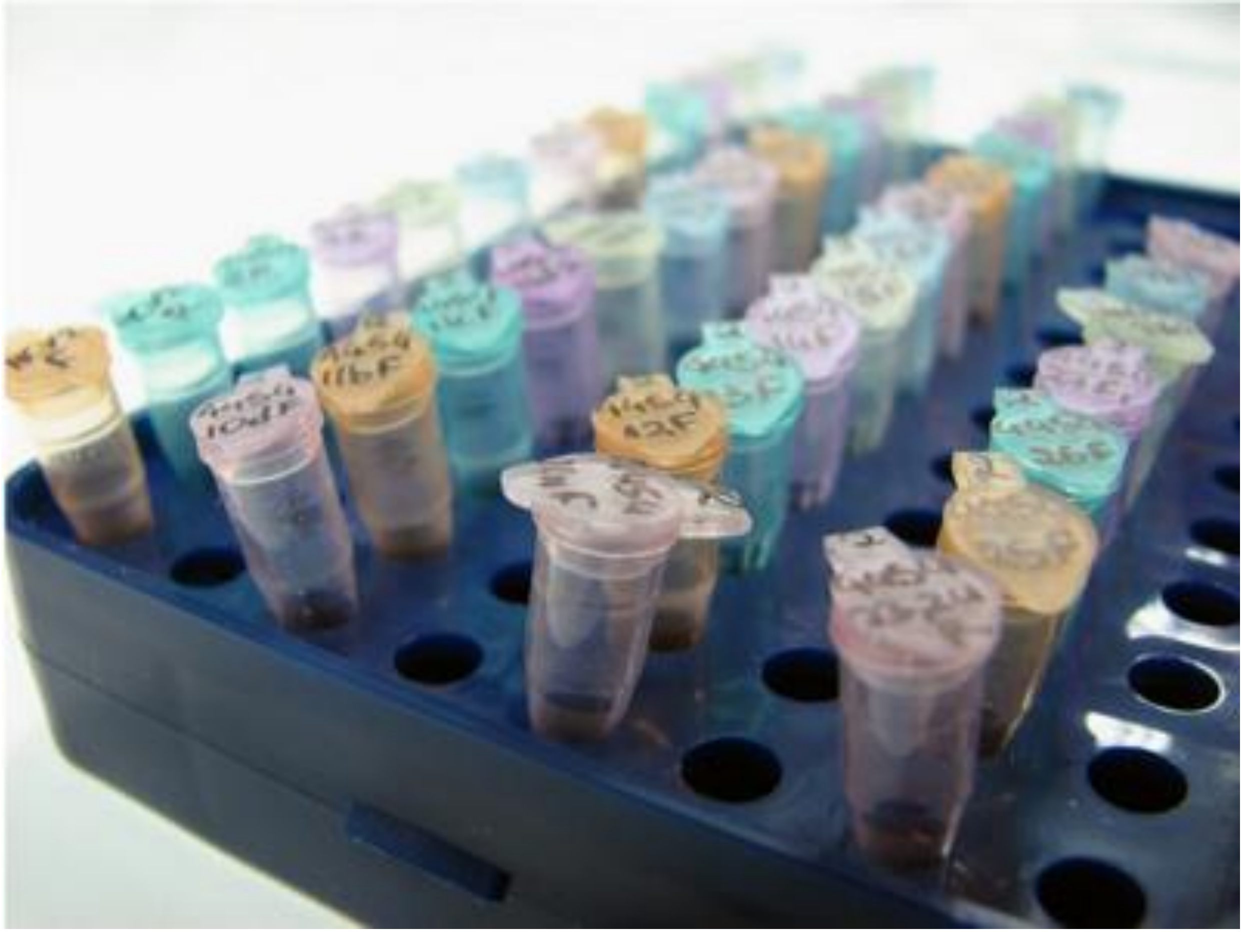
This QC step is performed by the wet-lab and gives an impression of the run quality on a per-lane basis.

Usually several samples are pooled on one lane, but no QC of individual samples is shown.

http://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html

Sample identity

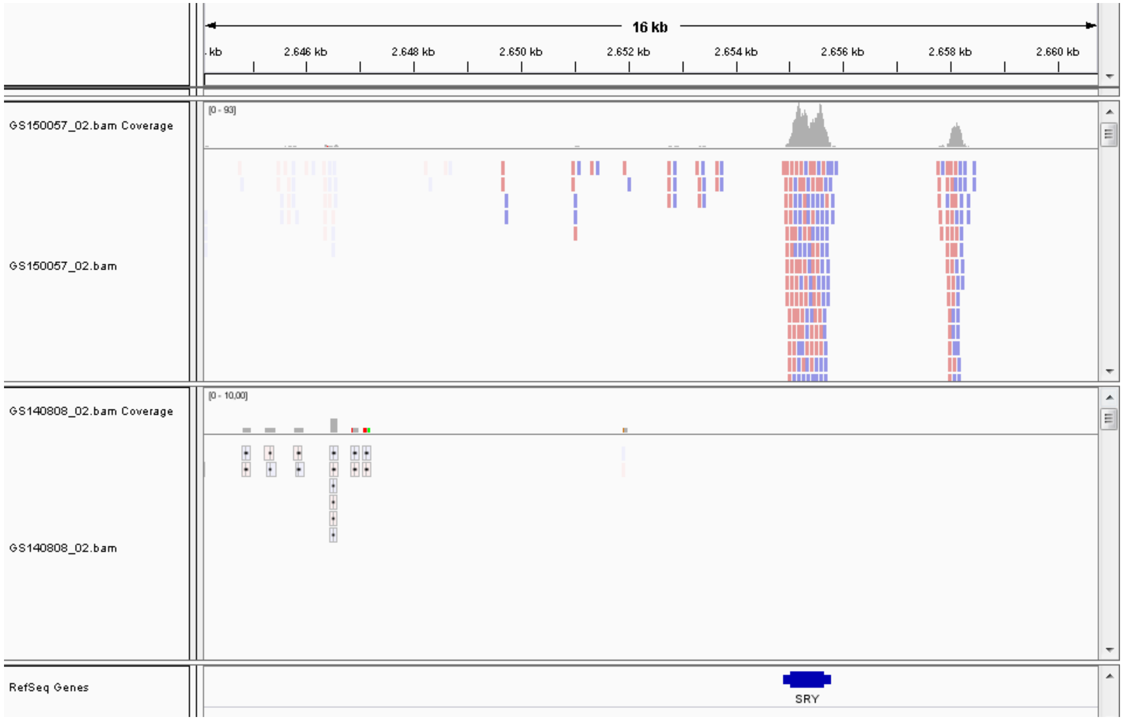
- Run QC
- ⇨ Sample identity
- Sample QC
- Variant QC



Sample identity - Gender

Checking gender can identify 50% of sample swaps (even sample swaps by sender). There are several possible methods:

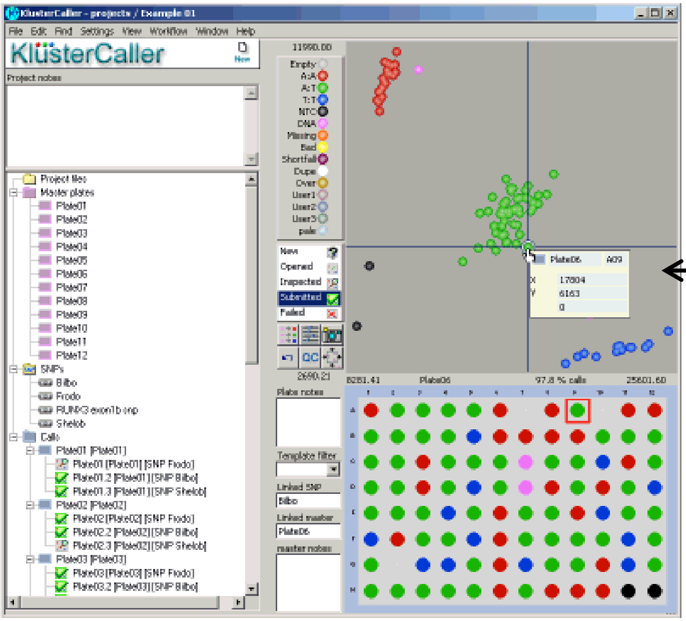
- coverage of the SRY gene (screenshot)
- ratio of reads mapped to chrX / chrY
- percentage of heterozygous SNPs on chrX

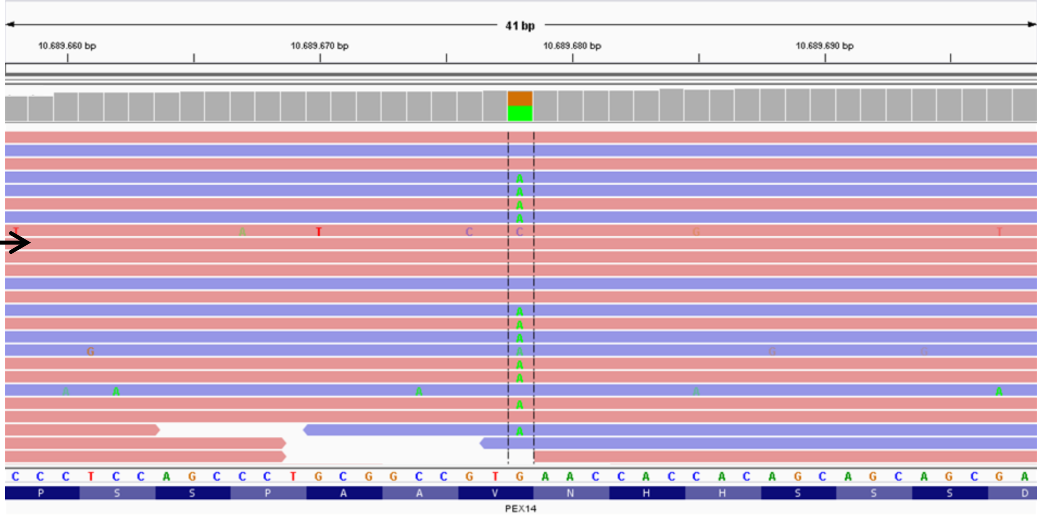


Sample identity - KASP

Sample identification based on SNP genotypes:

- Upon sample receipt, KASP assay is used to determine genotypes of 14 common SNPs
- After sequencing, KASP and NGS genotypes are compared

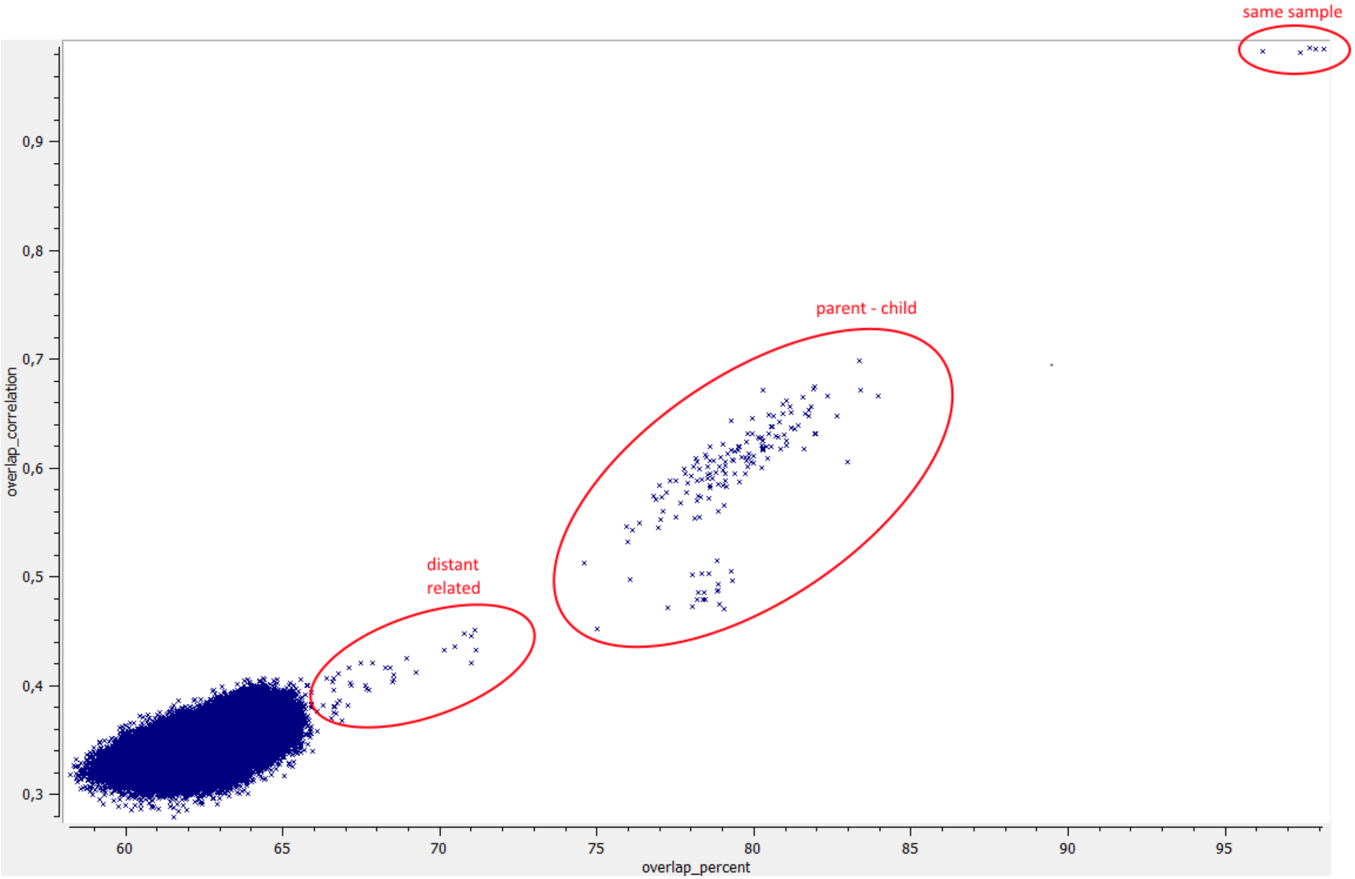



41 bp
10.689.660 bp
10.689.670 bp
10.689.680 bp
10.689.690 bp
C C C T C C A G C C C T G C G G C C G T G A A C C A C C A C A G C A G C A G C G A
P S S P A A V N H H S S S D
PEX14

Sample identity - Correlation

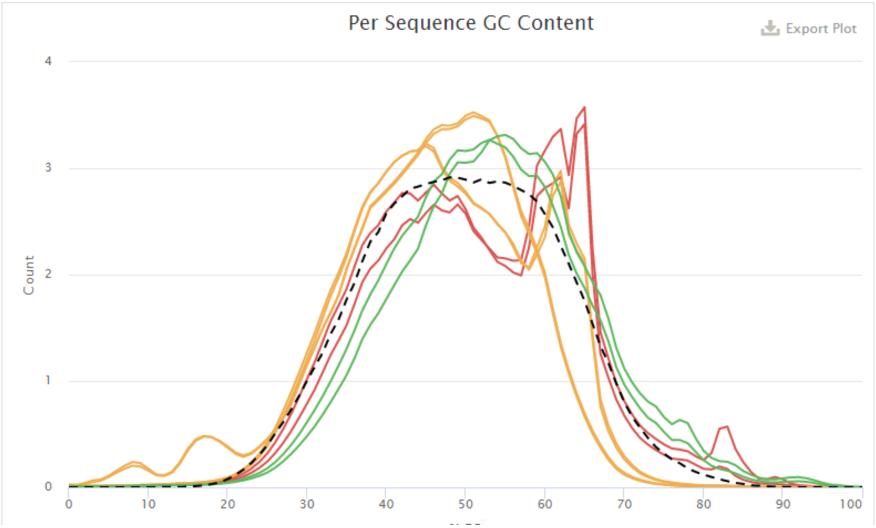
The overlap and genotype correlation of two variant lists can be used to check that similar samples show a high concordance.
Possible use-cases are for example:

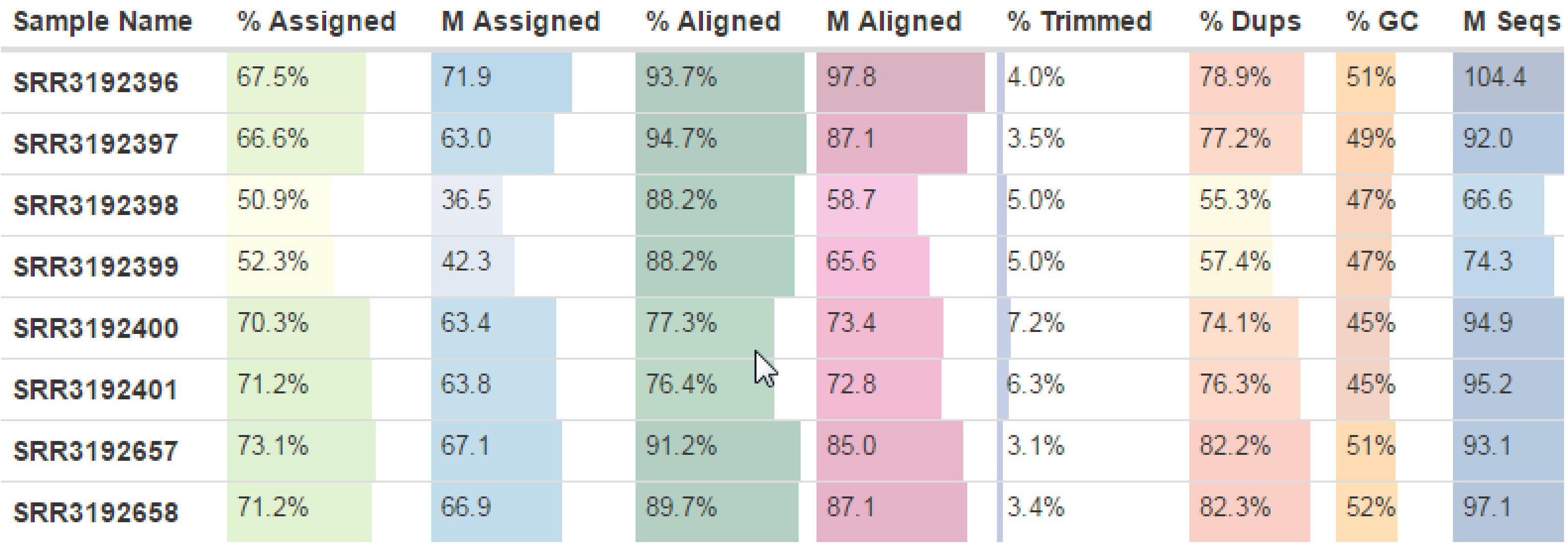
- tumor-normal pairs
- parent-child trios
- affected siblings



Sample QC

- Run QC
- Sample identity
- ⇨ Sample QC
- Variant QC

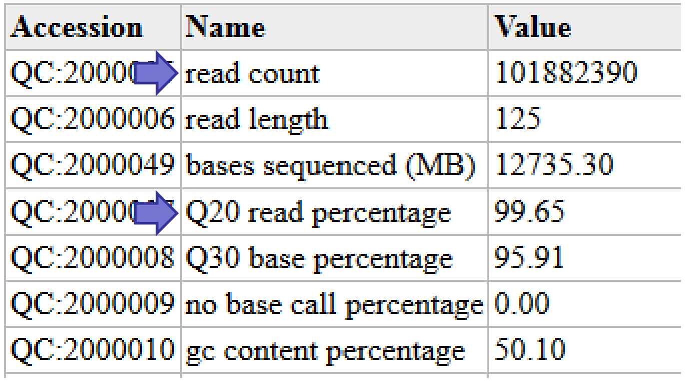




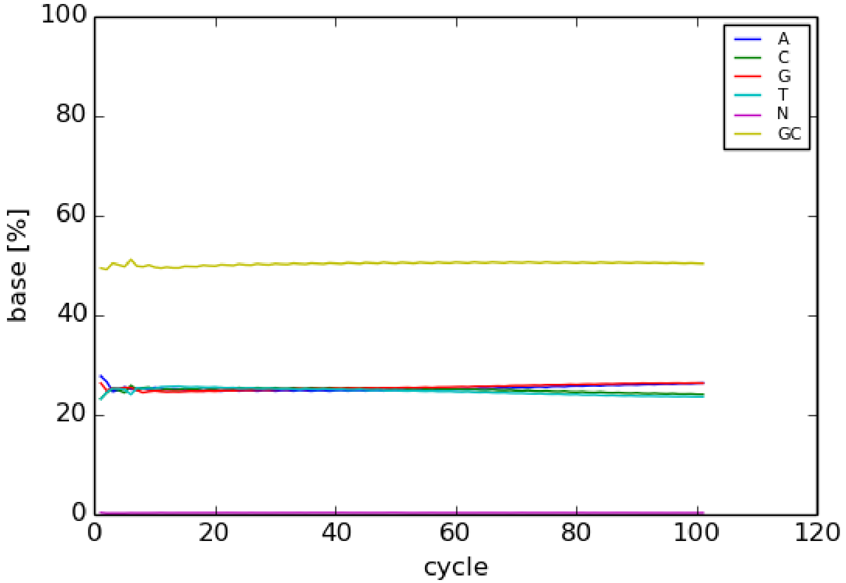
Sample Name	% Assigned	M Assigned	% Aligned	M Aligned	% Trimmed	% Dups	% GC	M Seqs
SRR3192396	67.5%	71.9	93.7%	97.8	4.0%	78.9%	51%	104.4
SRR3192397	66.6%	63.0	94.7%	87.1	3.5%	77.2%	49%	92.0
SRR3192398	50.9%	36.5	88.2%	58.7	5.0%	55.3%	47%	66.6
SRR3192399	52.3%	42.3	88.2%	65.6	5.0%	57.4%	47%	74.3
SRR3192400	70.3%	63.4	77.3%	73.4	7.2%	74.1%	45%	94.9
SRR3192401	71.2%	63.8	76.4%	72.8	6.3%	76.3%	45%	95.2
SRR3192657	73.1%	67.1	91.2%	85.0	3.1%	82.2%	51%	93.1
SRR3192658	71.2%	66.9	89.7%	87.1	3.4%	82.3%	52%	97.1

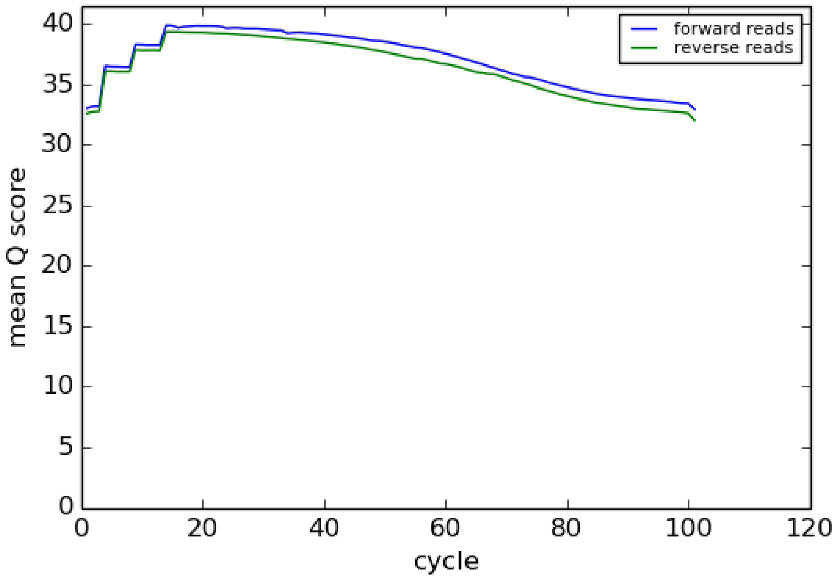
Sample QC - raw data

For each sample, several levels of QC can be performed. Several QC metrics can be calculated from the raw data (FASTQ):



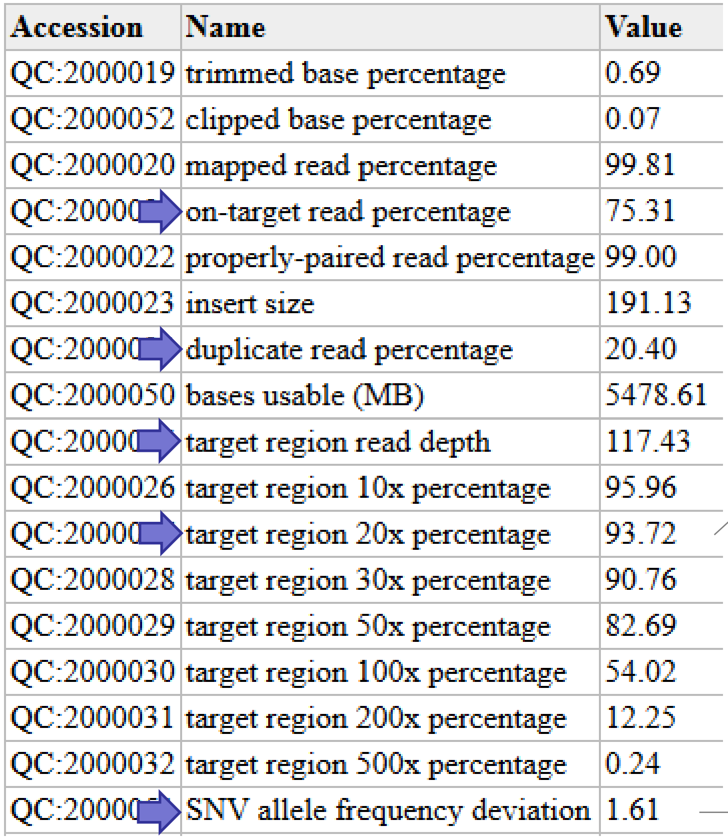
Accession	Name	Value
QC:20000➡	read count	101882390
QC:2000006	read length	125
QC:2000049	bases sequenced (MB)	12735.30
QC:20000➡	Q20 read percentage	99.65
QC:2000008	Q30 base percentage	95.91
QC:2000009	no base call percentage	0.00
QC:2000010	gc content percentage	50.10



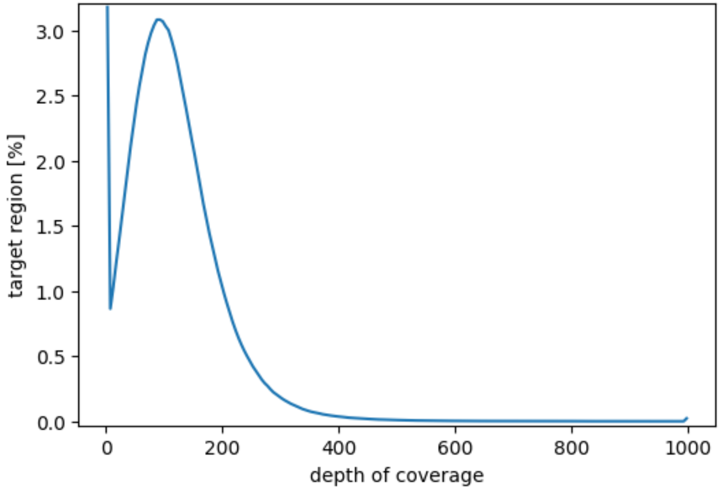


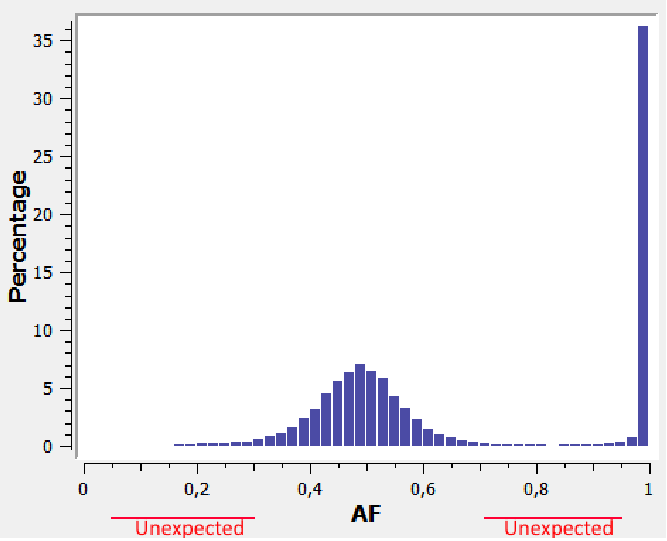
Sample QC - mapped reads

After mapping, the QC metrics can be calculated from the BAM file:



Accession	Name	Value
QC:2000019	trimmed base percentage	0.69
QC:2000052	clipped base percentage	0.07
QC:2000020	mapped read percentage	99.81
QC:20000	on-target read percentage	75.31
QC:2000022	properly-paired read percentage	99.00
QC:2000023	insert size	191.13
QC:20000	duplicate read percentage	20.40
QC:2000050	bases usable (MB)	5478.61
QC:20000	target region read depth	117.43
QC:2000026	target region 10x percentage	95.96
QC:20000	target region 20x percentage	93.72
QC:2000028	target region 30x percentage	90.76
QC:2000029	target region 50x percentage	82.69
QC:2000030	target region 100x percentage	54.02
QC:2000031	target region 200x percentage	12.25
QC:2000032	target region 500x percentage	0.24
QC:20000	SNV allele frequency deviation	1.61

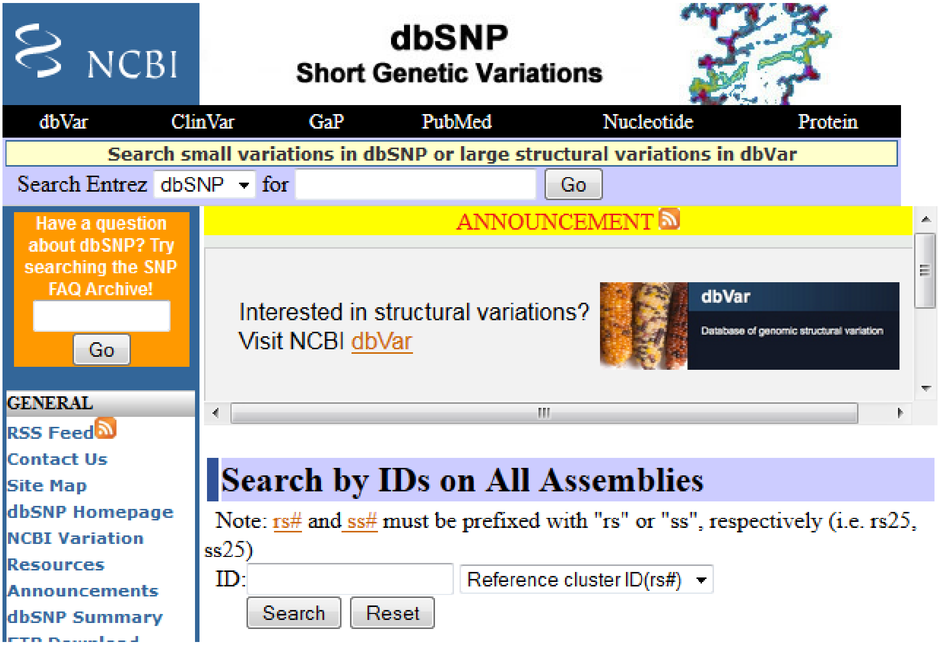




Sample QC - variants

Finally, several variant list quality metrics can be calculated from the VCF file:

Accession	Name	Value
QC:20000	variant count	39525
QC:20000	known variants percentage	99.25
QC:2000015	high-impact variants percentage	1.93
QC:2000016	homozygous variants percentage	38.16
QC:2000017	indel variants percentage	6.89
QC:20000	transition/transversion ratio	2.67



dbSNP
Short Genetic Variations

Search small variations in dbSNP or large structural variations in dbVar

Search Entrez dbSNP for

Have a question about dbSNP? Try searching the SNP FAQ Archive!

ANNOUNCEMENT

Interested in structural variations? Visit NCBI dbVar

dbVar
Database of genomic structural variation

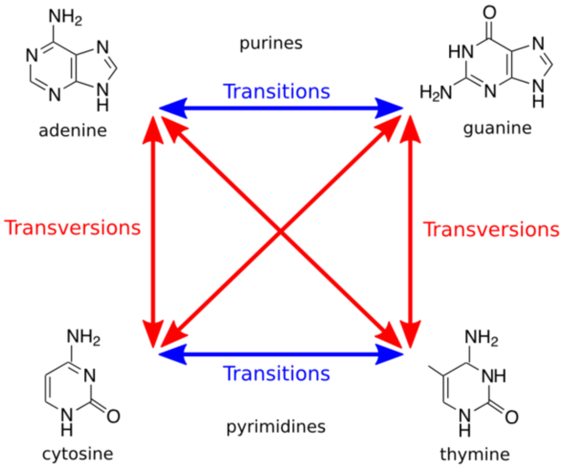
GENERAL

- RSS Feed
- Contact Us
- Site Map
- dbSNP Homepage
- NCBI Variation Resources
- Announcements
- dbSNP Summary

Search by IDs on All Assemblies

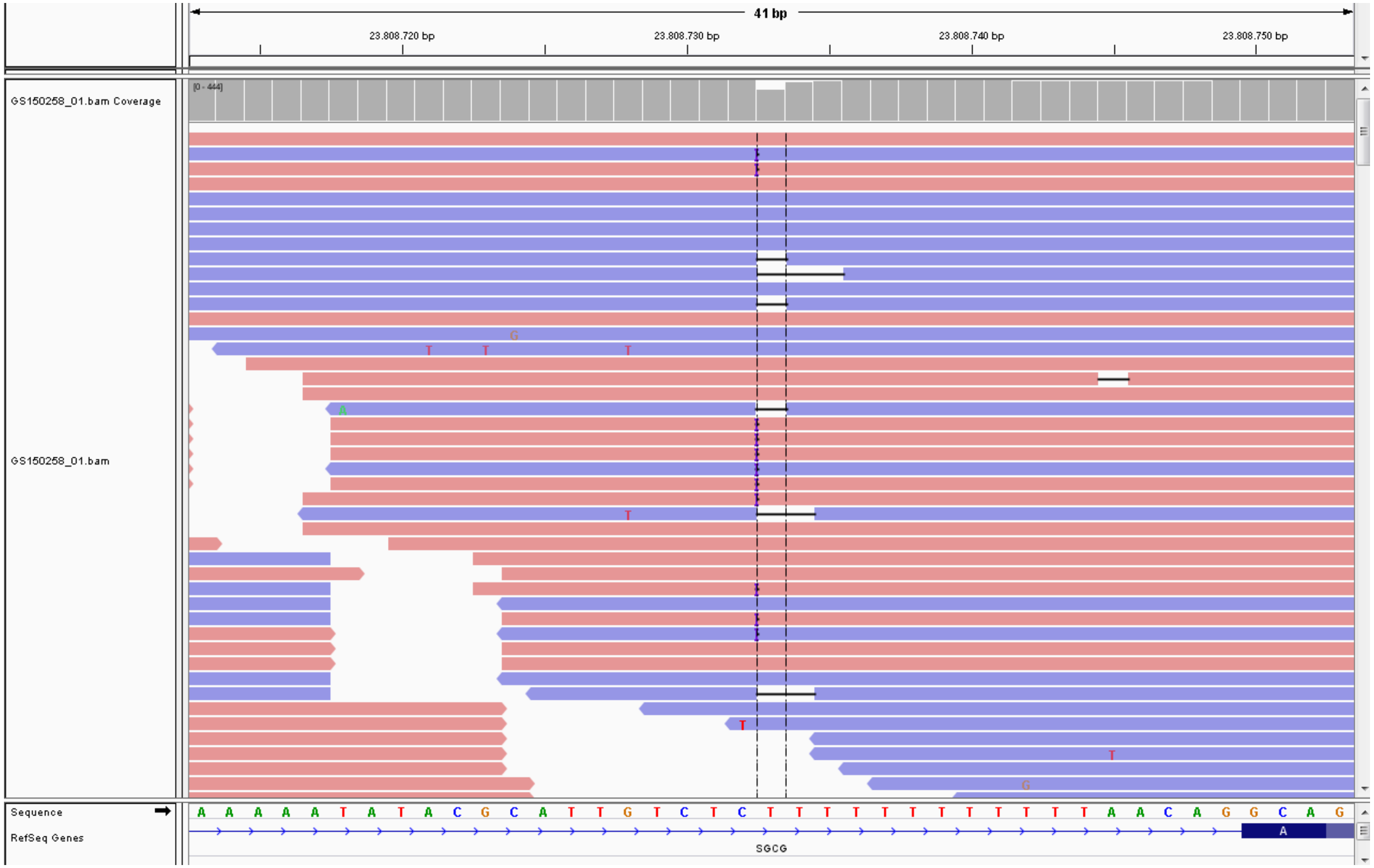
Note: rs# and ss# must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss25)

ID: Reference cluster ID(rs#)

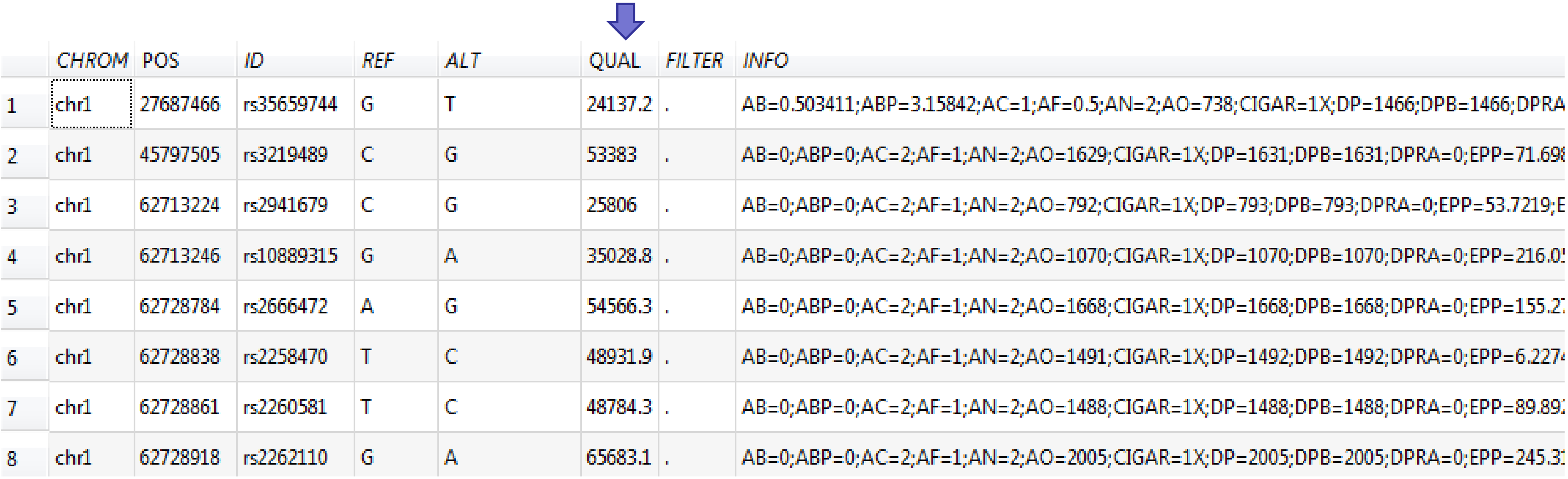


Variant QC

- Run QC
- Sample identity
- Sample QC
- Variant QC



Variant QC – variant calling



	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	chr1	27687466	rs35659744	G	T	24137.2	.	AB=0.503411;ABP=3.15842;AC=1;AF=0.5;AN=2;AO=738;CIGAR=1X;DP=1466;DPB=1466;DPRA
2	chr1	45797505	rs3219489	C	G	53383	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1629;CIGAR=1X;DP=1631;DPB=1631;DPRA=0;EPP=71.698
3	chr1	62713224	rs2941679	C	G	25806	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=792;CIGAR=1X;DP=793;DPB=793;DPRA=0;EPP=53.7219;E
4	chr1	62713246	rs10889315	G	A	35028.8	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1070;CIGAR=1X;DP=1070;DPB=1070;DPRA=0;EPP=216.05
5	chr1	62728784	rs2666472	A	G	54566.3	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1668;CIGAR=1X;DP=1668;DPB=1668;DPRA=0;EPP=155.27
6	chr1	62728838	rs2258470	T	C	48931.9	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1491;CIGAR=1X;DP=1492;DPB=1492;DPRA=0;EPP=6.2274
7	chr1	62728861	rs2260581	T	C	48784.3	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1488;CIGAR=1X;DP=1488;DPB=1488;DPRA=0;EPP=89.892
8	chr1	62728918	rs2262110	G	A	65683.1	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2005;CIGAR=1X;DP=2005;DPB=2005;DPRA=0;EPP=245.31

Variant quality score:

Phred score - error probability of a variant ($P = 10^{\frac{-Q}{10}}$):

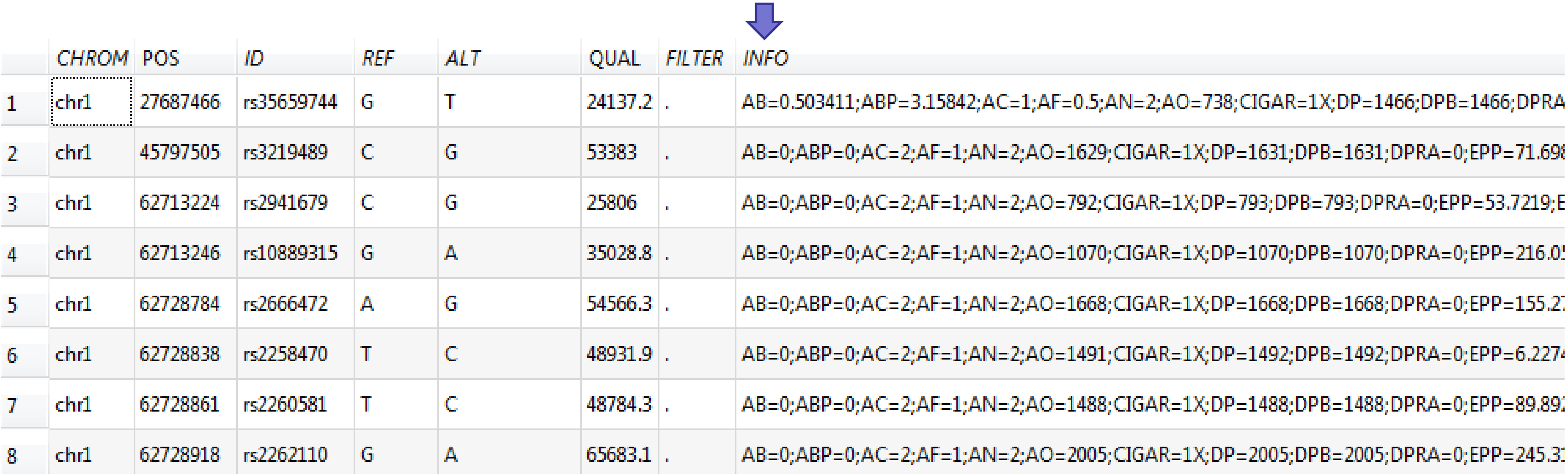
40 = 0.01%

30 = 0.1%

20 = 1%

10 = 10%

Variant QC – variant annotations



	CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	chr1	27687466	rs35659744	G	T	24137.2	.	AB=0.503411;ABP=3.15842;AC=1;AF=0.5;AN=2;AO=738;CIGAR=1X;DP=1466;DPB=1466;DPRA
2	chr1	45797505	rs3219489	C	G	53383	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1629;CIGAR=1X;DP=1631;DPB=1631;DPRA=0;EPP=71.698
3	chr1	62713224	rs2941679	C	G	25806	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=792;CIGAR=1X;DP=793;DPB=793;DPRA=0;EPP=53.7219;E
4	chr1	62713246	rs10889315	G	A	35028.8	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1070;CIGAR=1X;DP=1070;DPB=1070;DPRA=0;EPP=216.0
5	chr1	62728784	rs2666472	A	G	54566.3	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1668;CIGAR=1X;DP=1668;DPB=1668;DPRA=0;EPP=155.2
6	chr1	62728838	rs2258470	T	C	48931.9	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1491;CIGAR=1X;DP=1492;DPB=1492;DPRA=0;EPP=6.227
7	chr1	62728861	rs2260581	T	C	48784.3	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=1488;CIGAR=1X;DP=1488;DPB=1488;DPRA=0;EPP=89.89
8	chr1	62728918	rs2262110	G	A	65683.1	.	AB=0;ABP=0;AC=2;AF=1;AN=2;AO=2005;CIGAR=1X;DP=2005;DPB=2005;DPRA=0;EPP=245.3

Besides variant quality score, other relevant quality annotations are not standardized between variant callers.

freebayes:
DP – Total read depth at the locus
AF – Variant allele frequency
MQM – Mean mapping Q-score of observed alternate alleles

Questions

Part 1: Basics
- NGS library preparation
- Illumina sequencing
- Raw data (FASTQ format)

Part 2: Analysis pipeline
- Mapping
- Variant calling
- Variant annotation
- Variant filtering

Part 3: Quality control
- Run QC
- Sample identity
- Sample QC
- Variant QC

Used tools

This table summarizes the tools used for this presentation:

Step	Tool	URL
adapter trimming	SeqPurge	https://github.com/imgag/ngs-bits
mapping	BWA	http://bio-bwa.sourceforge.net/
duplicate removal	samblaster	https://github.com/GregoryFaust/samblaster
indel realignment	ABRA2	https://github.com/mozack/abra2
variant calling	freebayes	https://github.com/ekg/freebayes
variant normalization	vcfallelicprimitives vcfbreakmulti	https://github.com/vcflib/vcflib
indel left-alignment	VcfLeftNormalize	https://github.com/imgag/ngs-bits
Variant annotation	VEP	https://www.ensembl.org/info/docs/tools/vep/index.html
QC	ReadQC MappingQC VariantQC	https://github.com/imgag/ngs-bits

Our analysis pipeline **megSAP** can be found here: https://github.com/imgag/megSAP

Alternative tools

This table lists alternative widely-used tools:

Step	Tool	URL
adapter trimming	Skewer	https://sourceforge.net/projects/skewer/
mapping	Bowtie 2	https://sourceforge.net/projects/bowtie-bio/
duplicate removal	Picard MarkDuplicates	http://broadinstitute.github.io/picard/
indel realignment variant calling	GATK	https://www.broadinstitute.org/gatk/
annotation	Annovar	http://annovar.openbioinformatics.org/ ⚡ License
annotation	SnpEff	http://snpeff.sourceforge.net/
QC	FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/