Next Generation Sequencing

Overview - variant analysis and workflow QC

Christian Gilissen PhD

christian.gilissen@radboudumc.nl

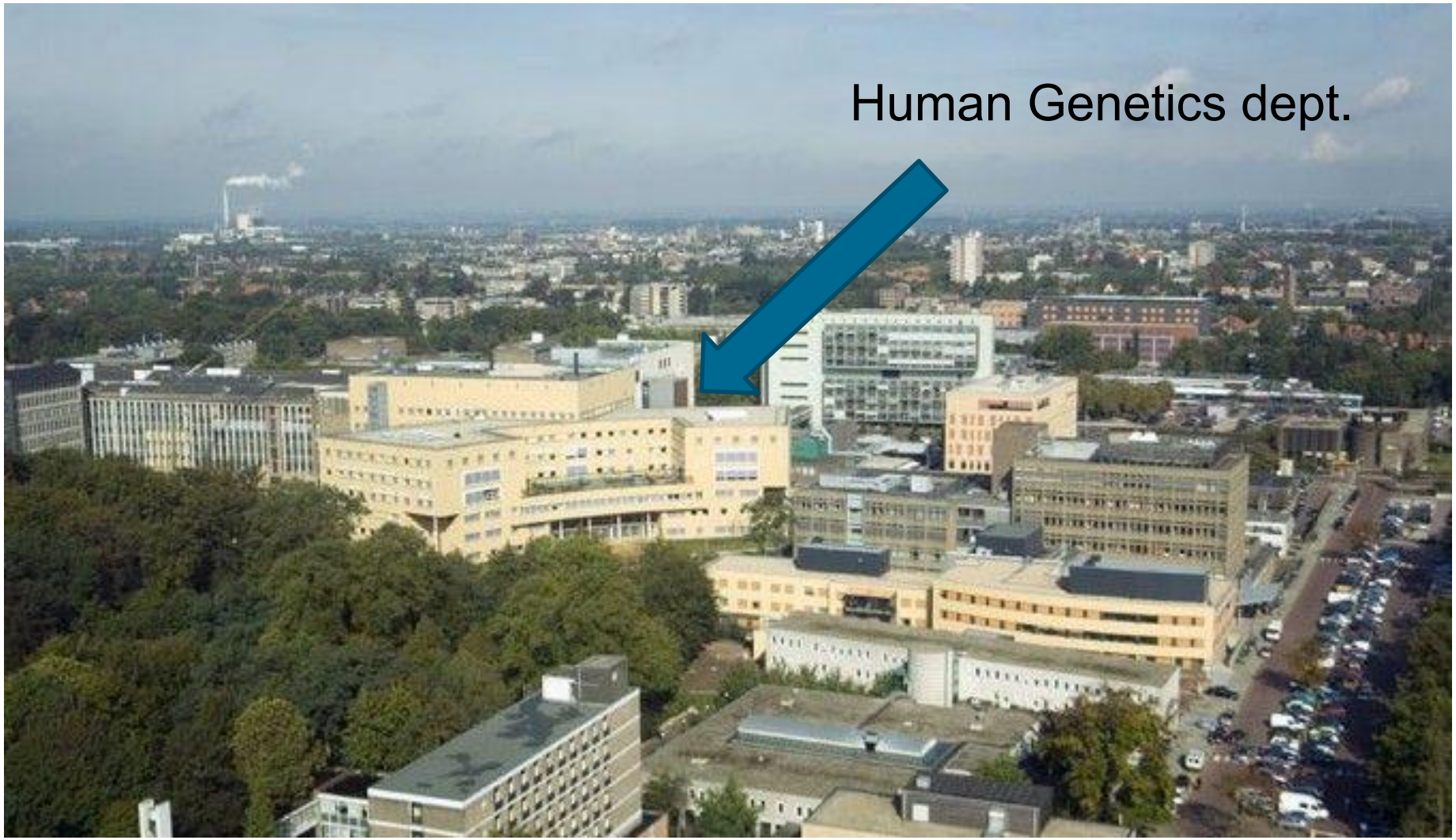
12-12-2022

Human Genetics Nijmegen

Radboudumc
university medical center

Nijmegen – the Netherlands

Human genetics Nijmegen



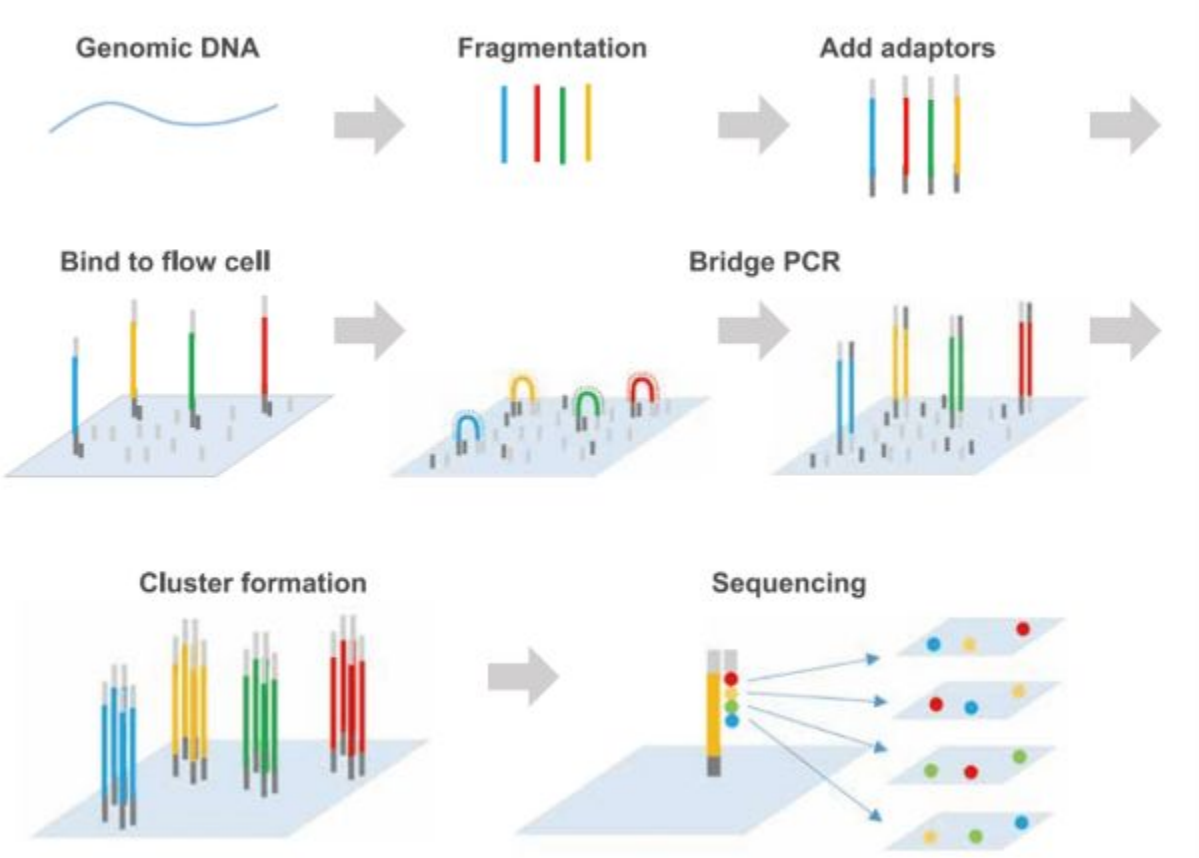
Contents

- Quick introduction to NGS analysis
- Files and file formats for NGS
- Quality control:
  - Raw data
  - Mapping
  - Coverage
  - Variant calling
  - Interpretation

Exome sequencing

Next Generation Sequencing

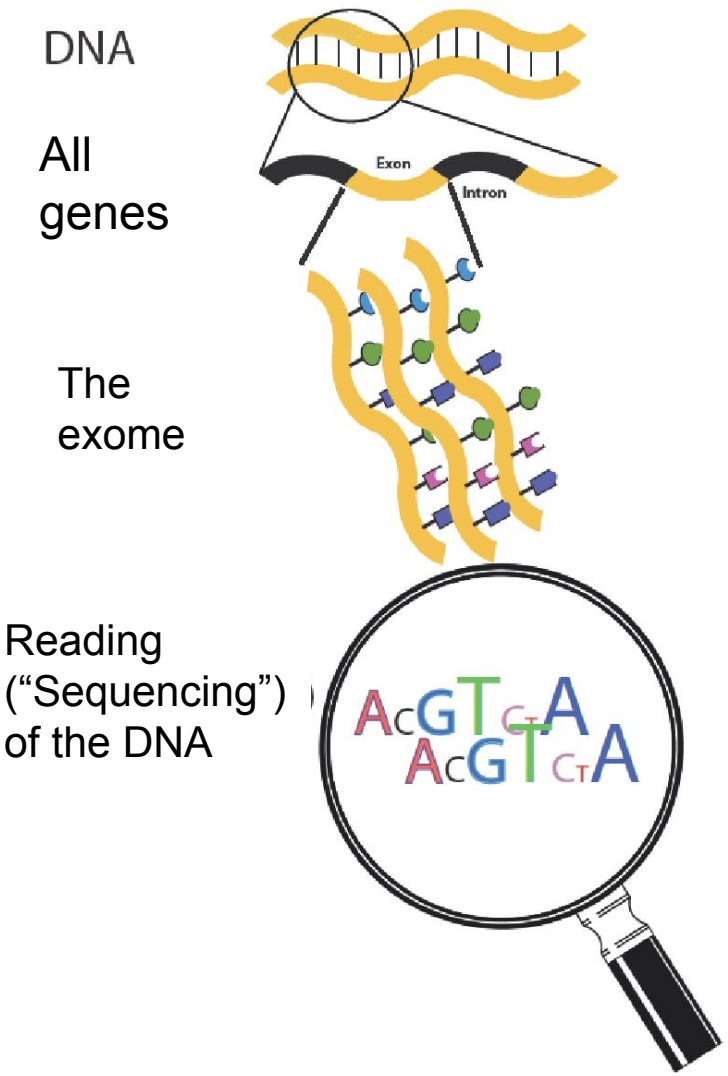
- Also called: massively parallel sequencing:
  - Reading millions of small DNA fragments in parallel



Sequencer

What is exome sequencing?

- Reading of the DNA of all coding sequences:
  - ~ 50 million base pairs
  - Out of ~3,000 million in the genome
- Enrichment methods capture the "interesting" part of the DNA
- Advantages: cheaper and faster!



How does it work?

Coding regions in **bold**

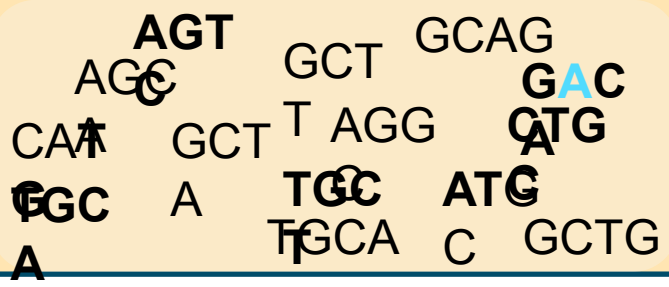
DNA

AGCA**TGCTGCAGTCAT**GCTT

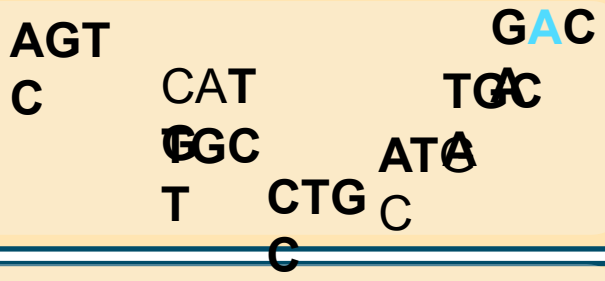
AGCA**TGCTGCAGACAT**GCTT

AGGCTA

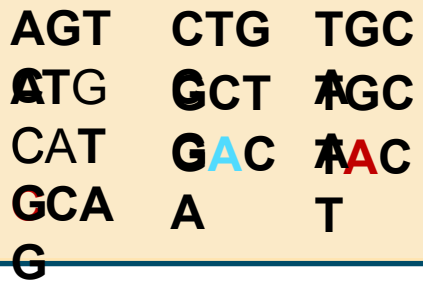
DNA Fragmentation



DNA Enrichment



Sequencing



Radboudumc

How does it work? - analyze

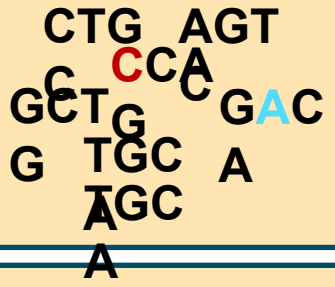
Sequence reads

FASTQ file

Reference

AGCATGCTGCAGTCATGCTTAGGCTA

Alignment / Mapping



Mapped reads

BAM file

Variant calling

Chr5, position 9, G → C
Reads = 5, Variants reads = 1, (20%), score = 10

Chr5, position 13, T → A
Reads = 2, Variants reads = 1, (50%), score = 50

VCF file

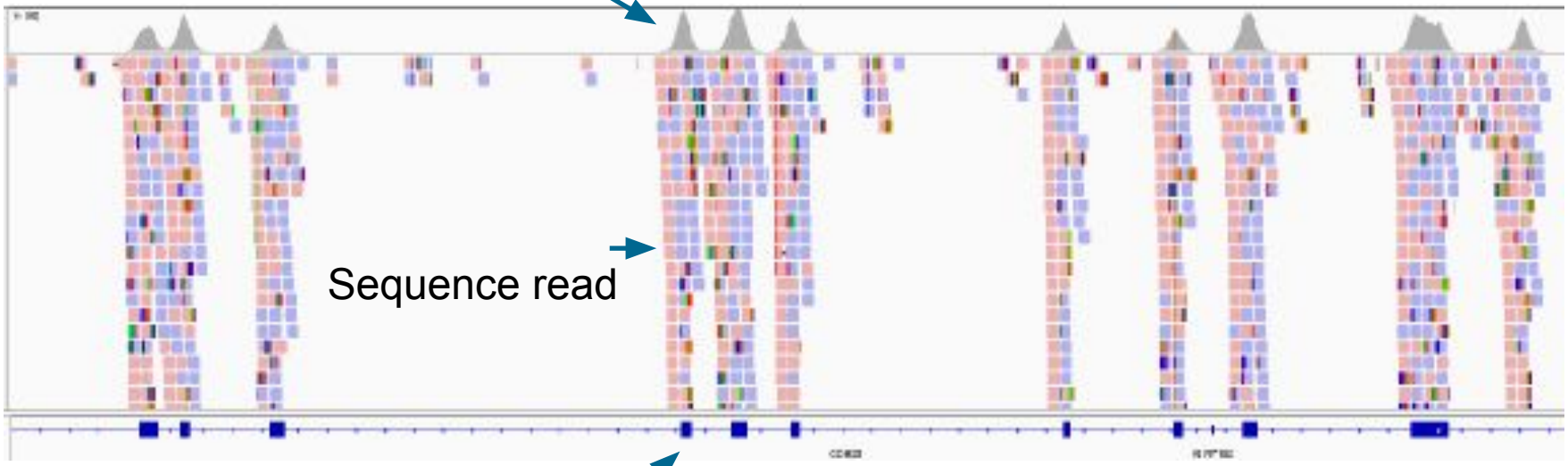
Variant Annotation

Annotated VCF file

Variant Interpretation

What do you get?

Total number of reads
(coverage)

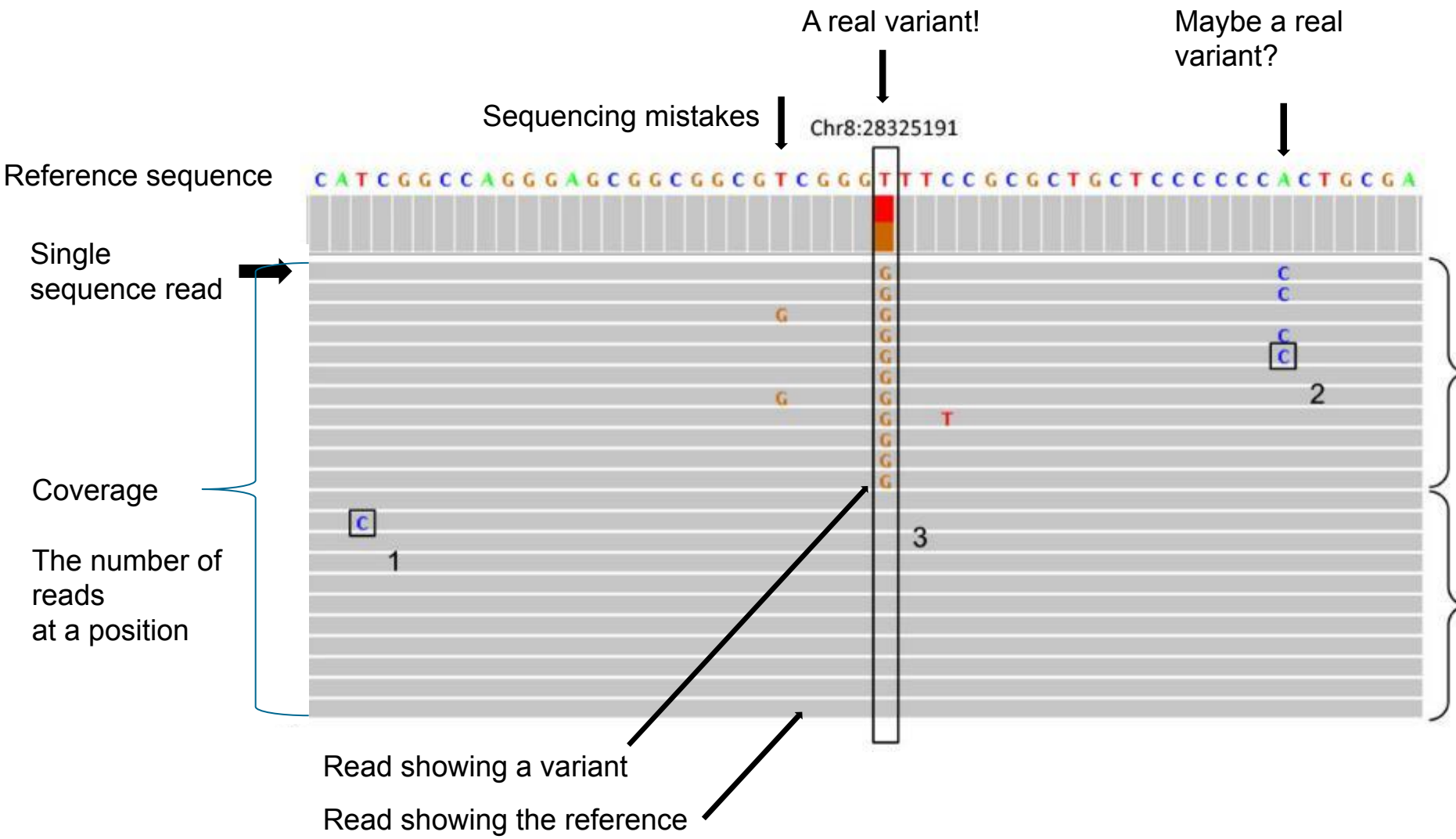


Sequence read

Exon of a gene

IGV screenshot: https://software.broadinstitute.org/software/igv/

Sequencing data and variants



IGV screenshot: https://software.broadinstitute.org/software/igv/

Summary

Sequencing

FASTQ

Alignment

BAM

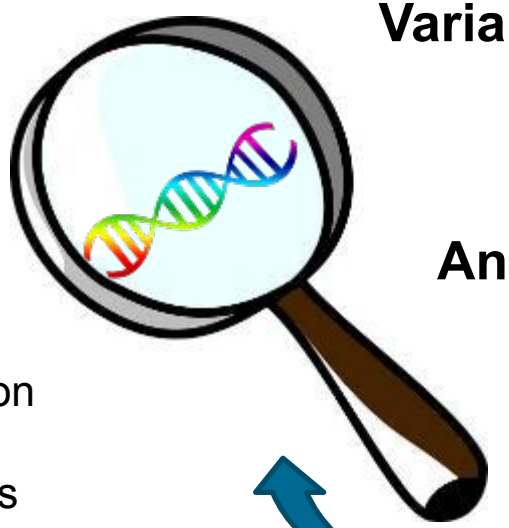
Variant calling

VCF

Annotation

Annotated VCF

Interpretation



- ✅ Effect on the protein
- ✅ Rare in the population
- ✅ Predicted deleterious

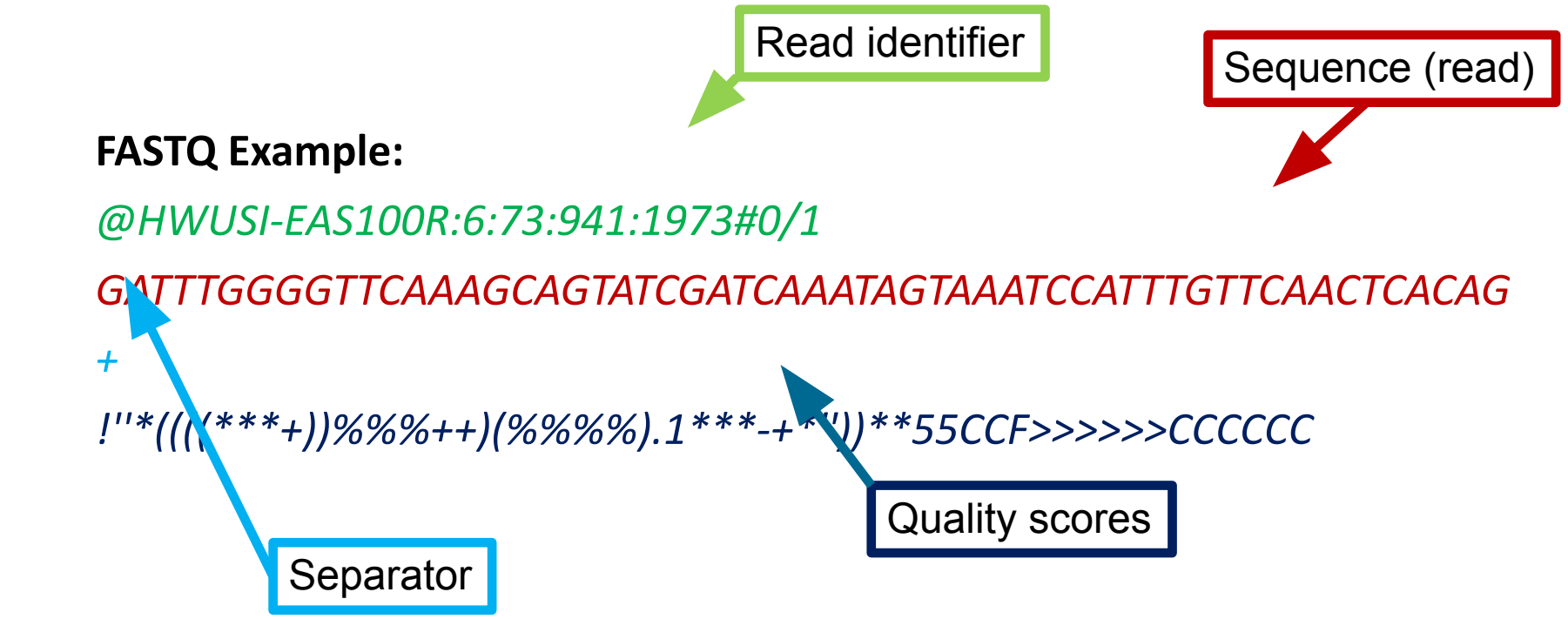
Data files

FASTQ

Purpose: store sequence information

Format: 1 header line followed by sequence, separator and quality scores

Size: ±15Gb per exome



FASTA / FASTQ - 2

Tools:

- *BWA alignment of reads to a reference*
- FASTQC, qualimap quality control
- FASTX suite read trimming, quality control

Remarks:

- Inefficient storage, typically files are gzipped (fq.gz or fastq.fz)
- Many, many lines (4 lines per read!)
- Come in pairs for paired-end sequencing
- RunX.1.fq, RunX.2.fq, sometimes also called forward/reverse

SAM / BAM / CRAM

Purpose: storing sequence alignments

Format: headers, and then one read alignment per line

Per line:

- the (genomic) position of the alignment,
- the way of the alignment (cigar string),
- the quality of the aligned read,
- the sequence of the aligned read

Size: bam = ±10-15Gb

Example

@HD VN:1.5 SO:coordinate

@SQ SN:ref LN:45

chr1 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *

Chr1 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *

Chr1 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;

SAM / BAM / CRAM - 2

Tools:

- *GATK, FreeBayes identification of variants*
- Samtools bam/sam conversion
- IGV integrated genomics viewer for viewing
- Picard tools bam2fastq, mark duplicates

Remarks

- Bam is the binary version of Sam, and much smaller
- Typically indexed with a bai file for speed-up
- The most important file for sequencing
- Usually, original FASTQ can fully be retrieved from bam
- Additional compression possible **CRAM** format now natively supported by samtools

VCF / GVCF

Purpose: storing variant calls

Format: header, followed by one line per position

Size: ± 100Mb

##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">

##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Sample1 Sample2 Sample3

chr2 4370 rs6057 G A 29 . NS=2;DP=13;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:52,51 1|0:48:8:51,51 1/1:43:5:.,.

Position

identifier

Reference / variant

Quality score

Info: read depth, genotype, quality etc.

VCF / GVCF - 2

Tools:

- VCFTools manipulation of VCF (splitting merging)
- Annovar Annotation of variants
- VEP online

Remarks:

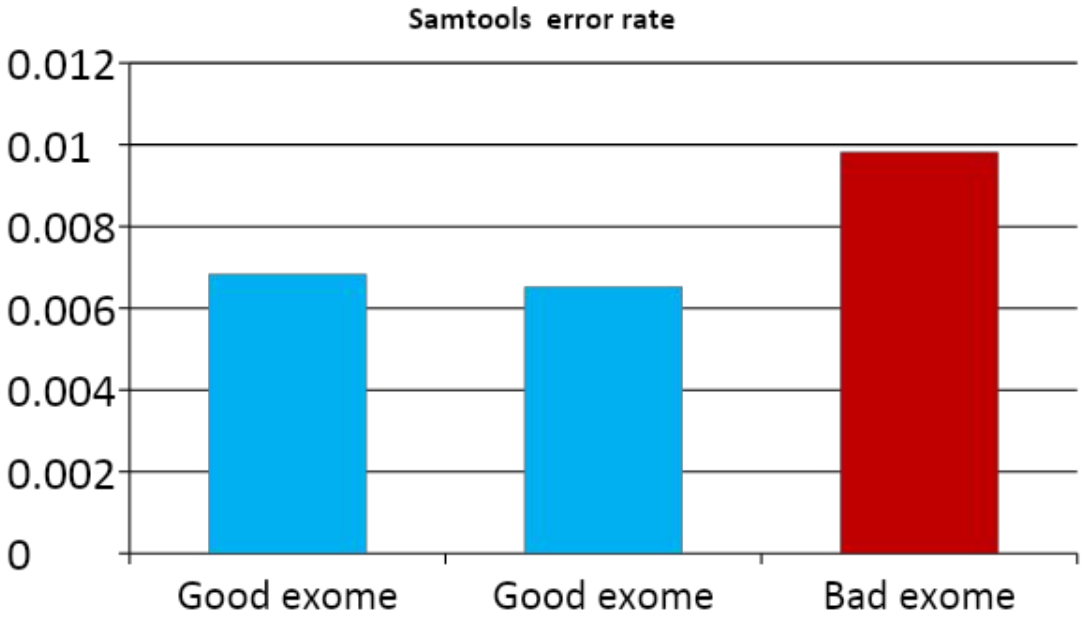
- Difficult to handle format, because multiple variants per line
- Big differences between VCF in the wild
- Can handle multiple samples in one file
- Difficult to view manually (excel)
- Can also be compressed / indexed

Raw sequencing data

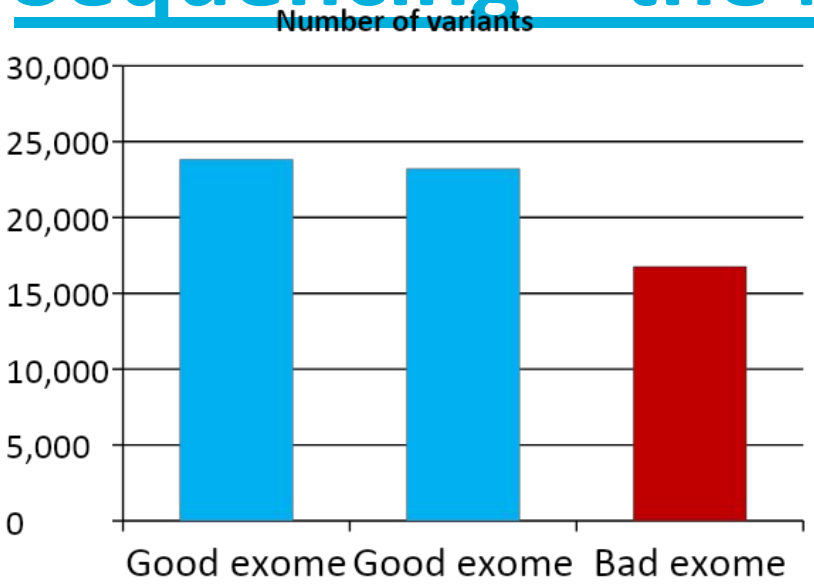
Sequencing – the good

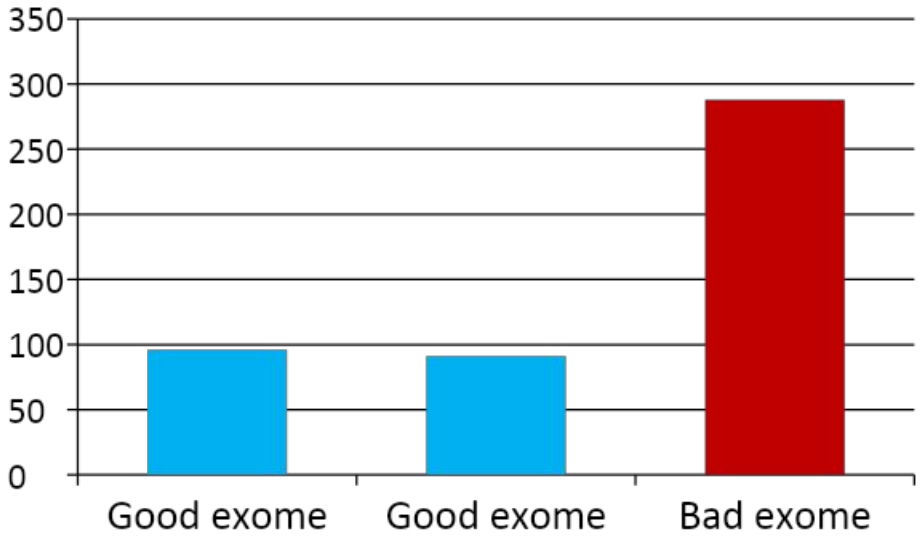
- Sequenced several samples in a single run.

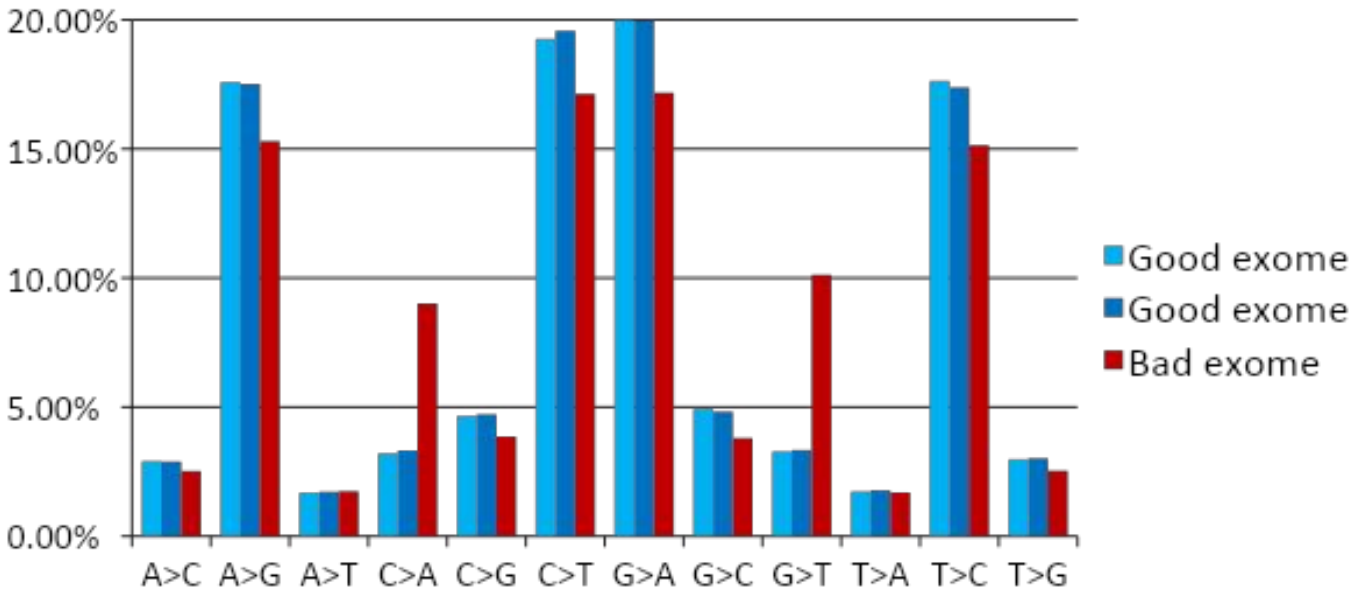
- **All quality metrics were good**.
  - Only the error rate was slightly higher for one sample.
  - And exome coverage was a lot lower for the same sample



Sequencing – the bad

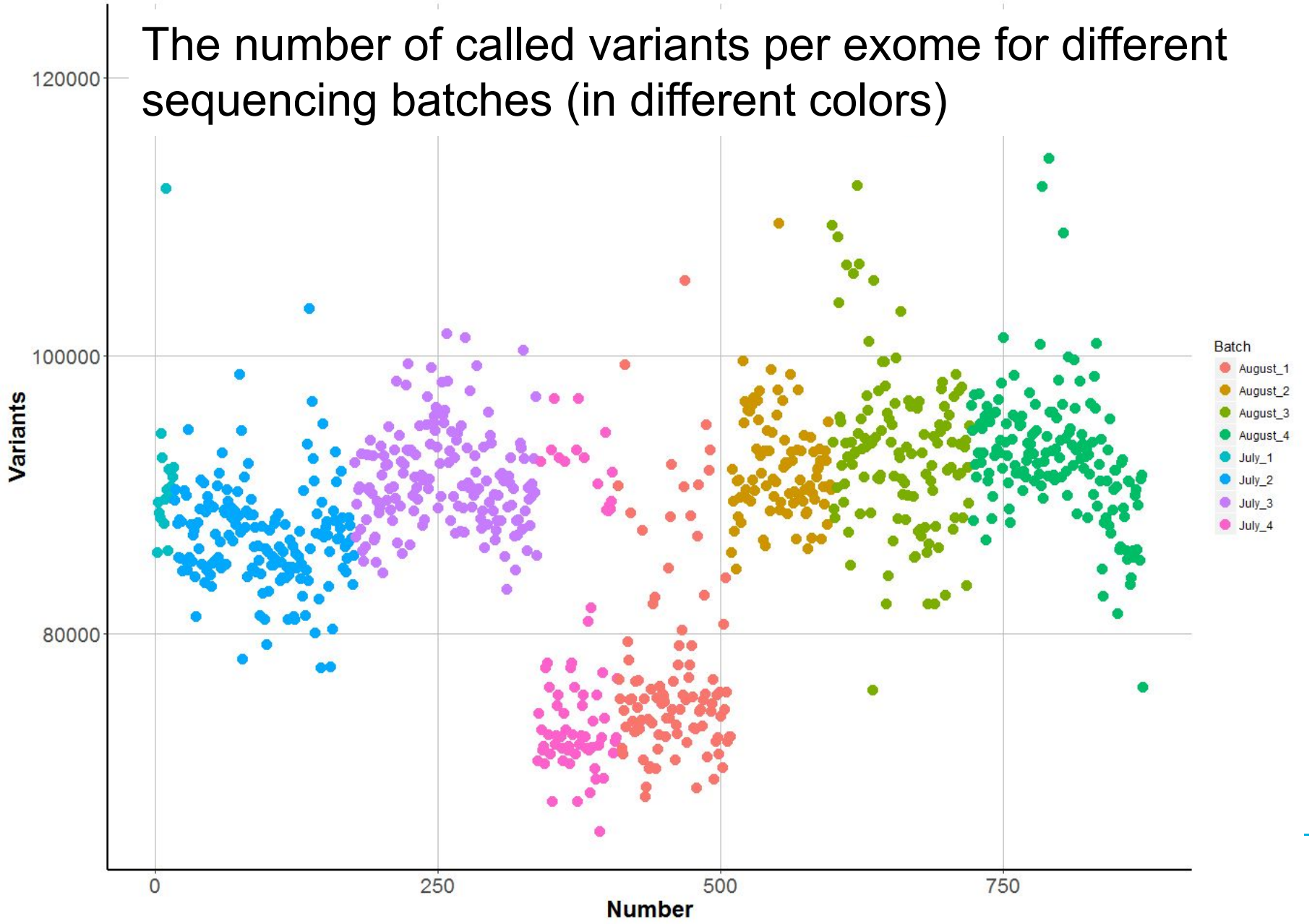






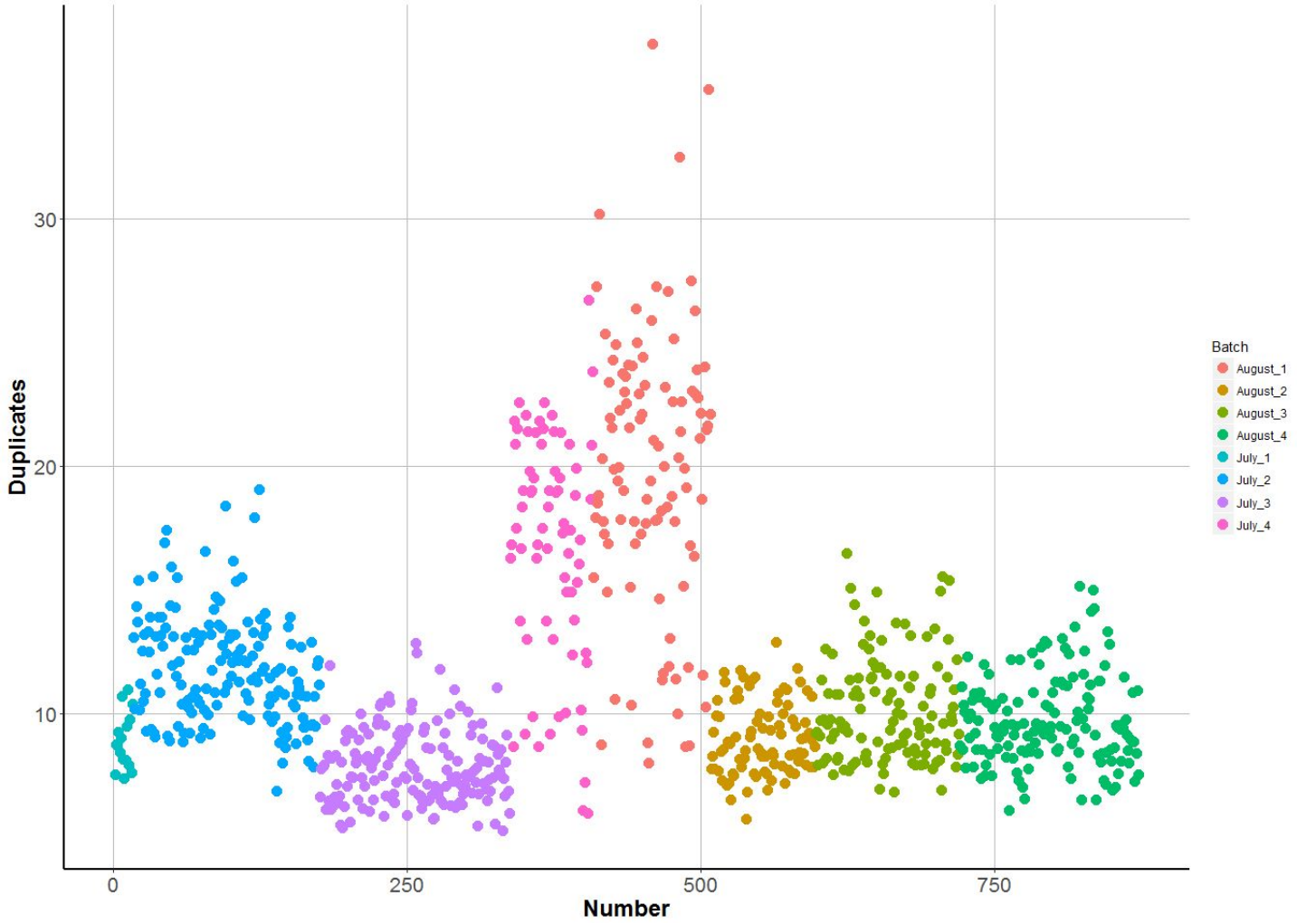
Calling variants in exomes

The number of called variants per exome for different sequencing batches (in different colors)



Exome coverage?

% of duplicate reads



Solution: we check coverage, Q40 and % duplicates of data

Take home message

Always check the raw quality of your sequence data:

- Error rate
- Q40
- % duplicates

Use programs like FastQC or qualimap

Mapping / alignment

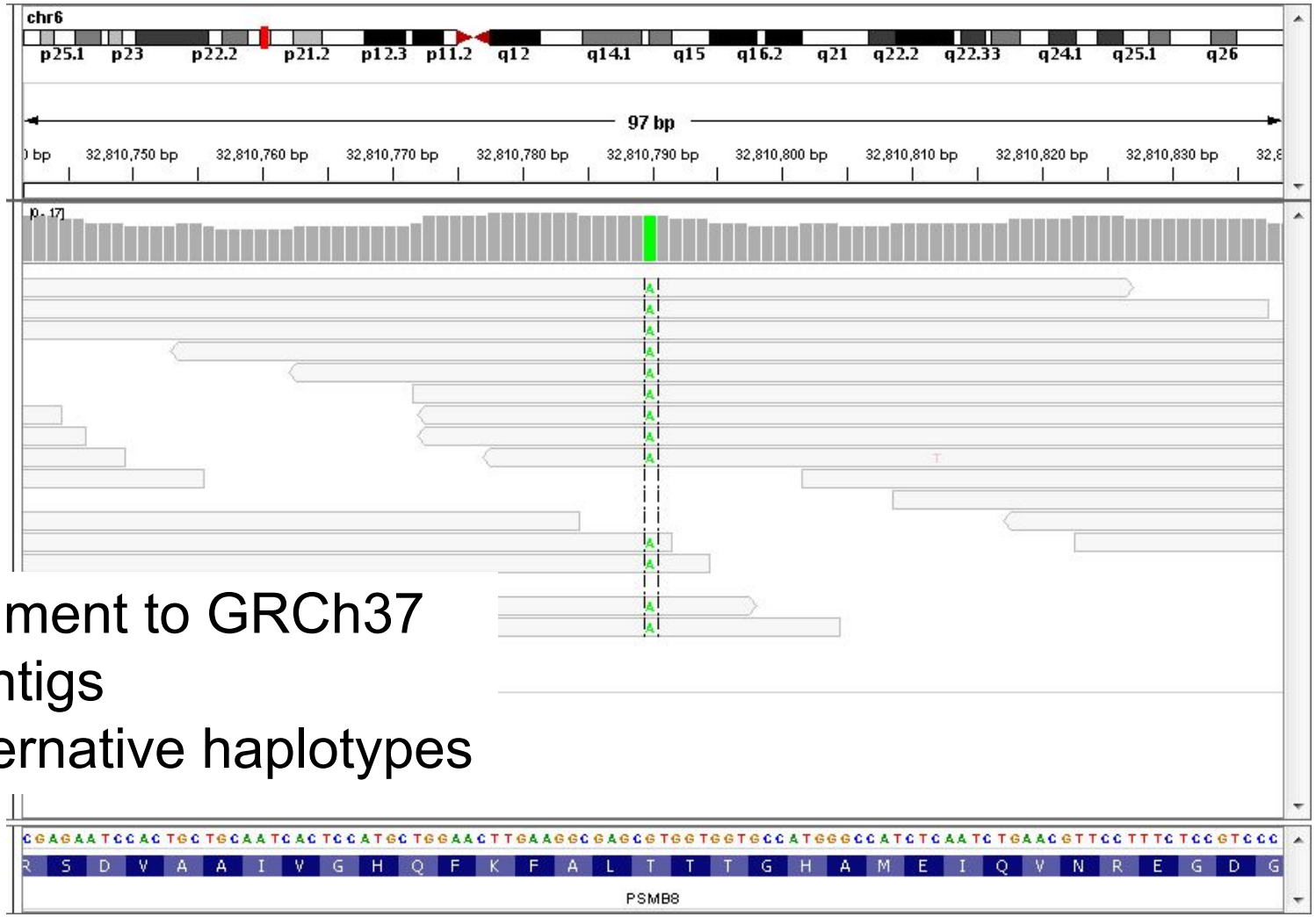
Easy example: Adaptor contamination

Analysis with wrong adaptor sequence specified

FASTQ → BAM → VCF → Annot

Radboudumc

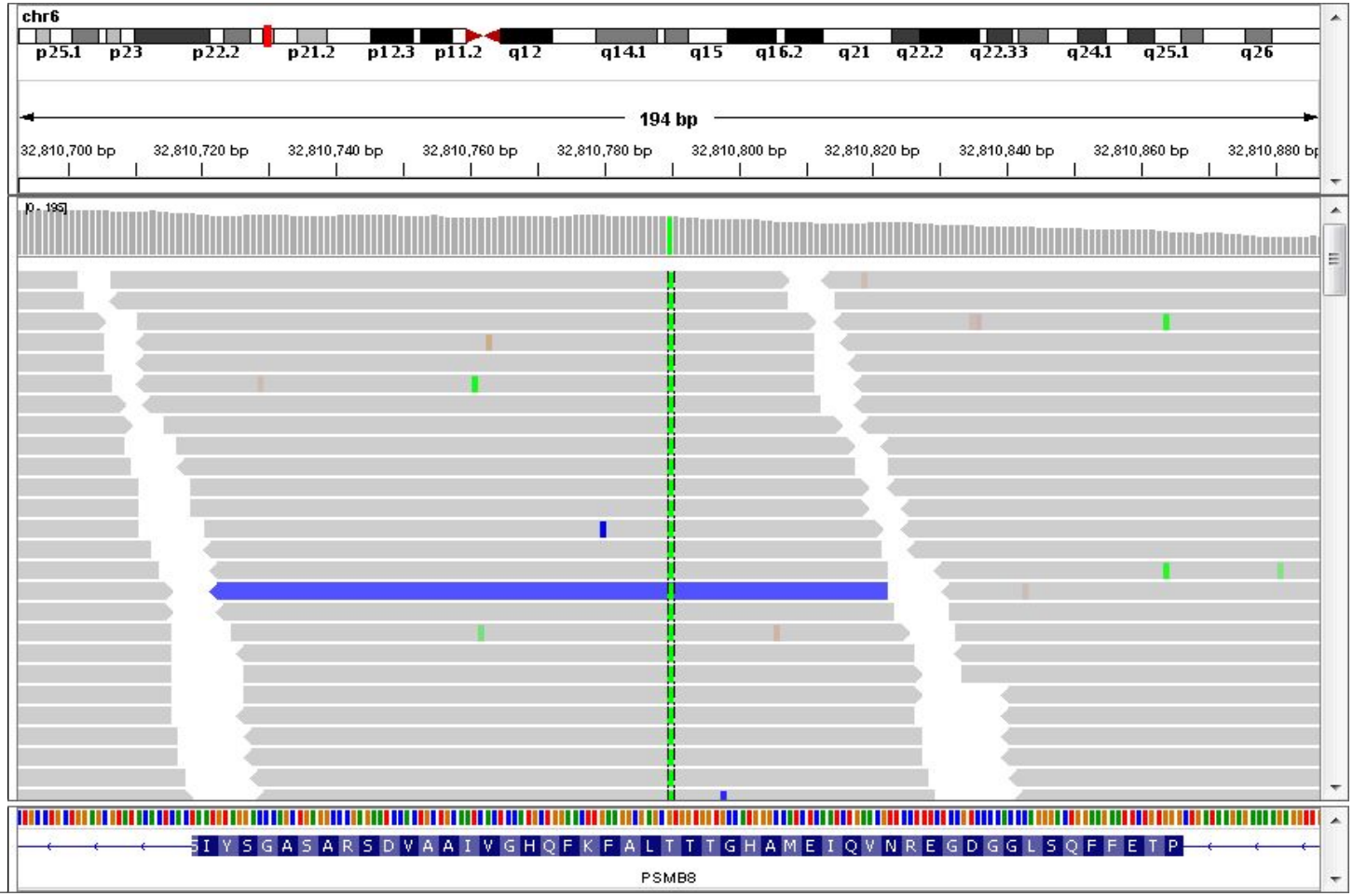
Missed variants because of alignment



Alignment to GRCh37
+ contigs
+ alternative haplotypes

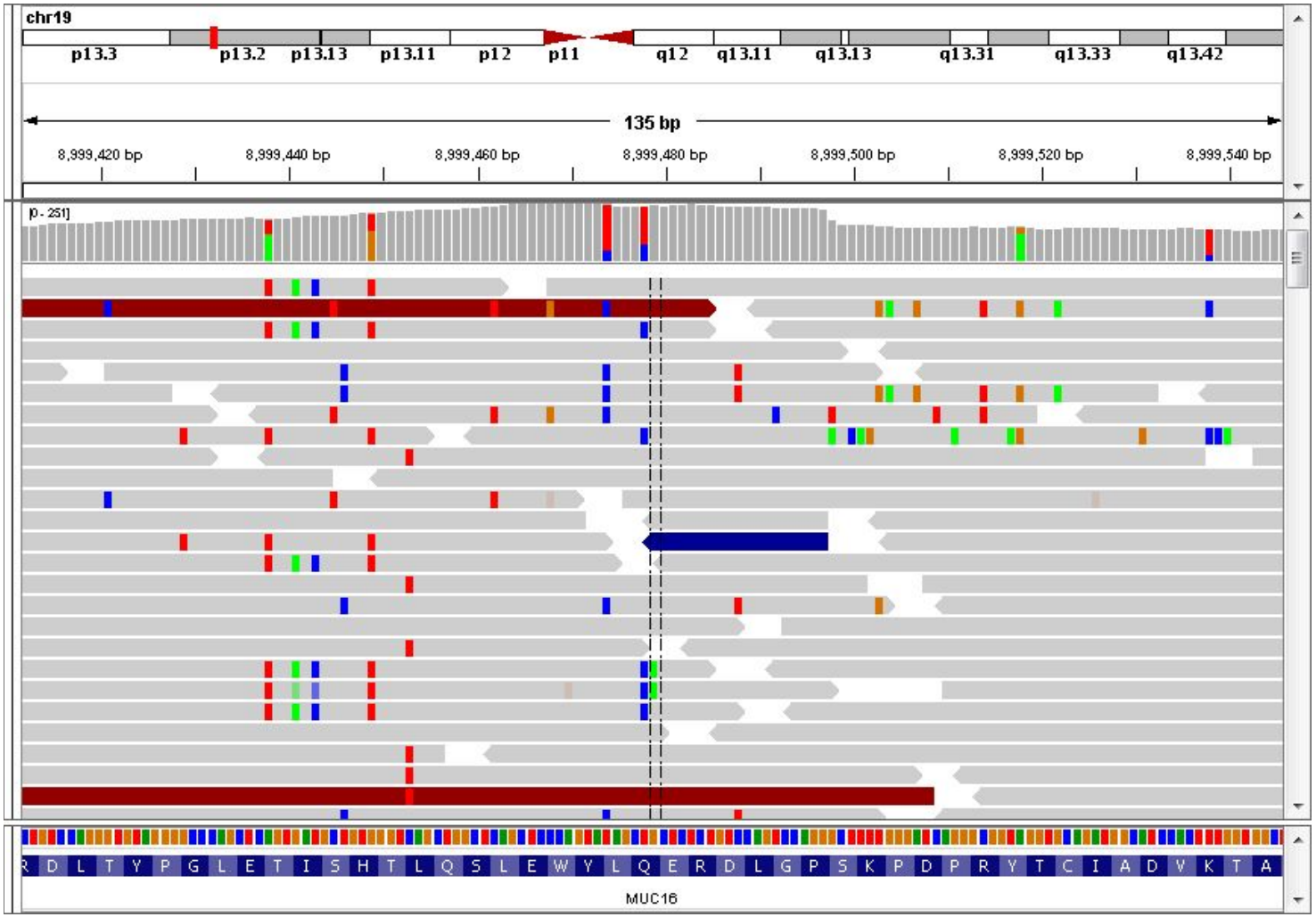
Sequencing → Alignment → Variant calling → Annotation

Realignment of the same data



Solution: We use no alternative haplotypes for the reference

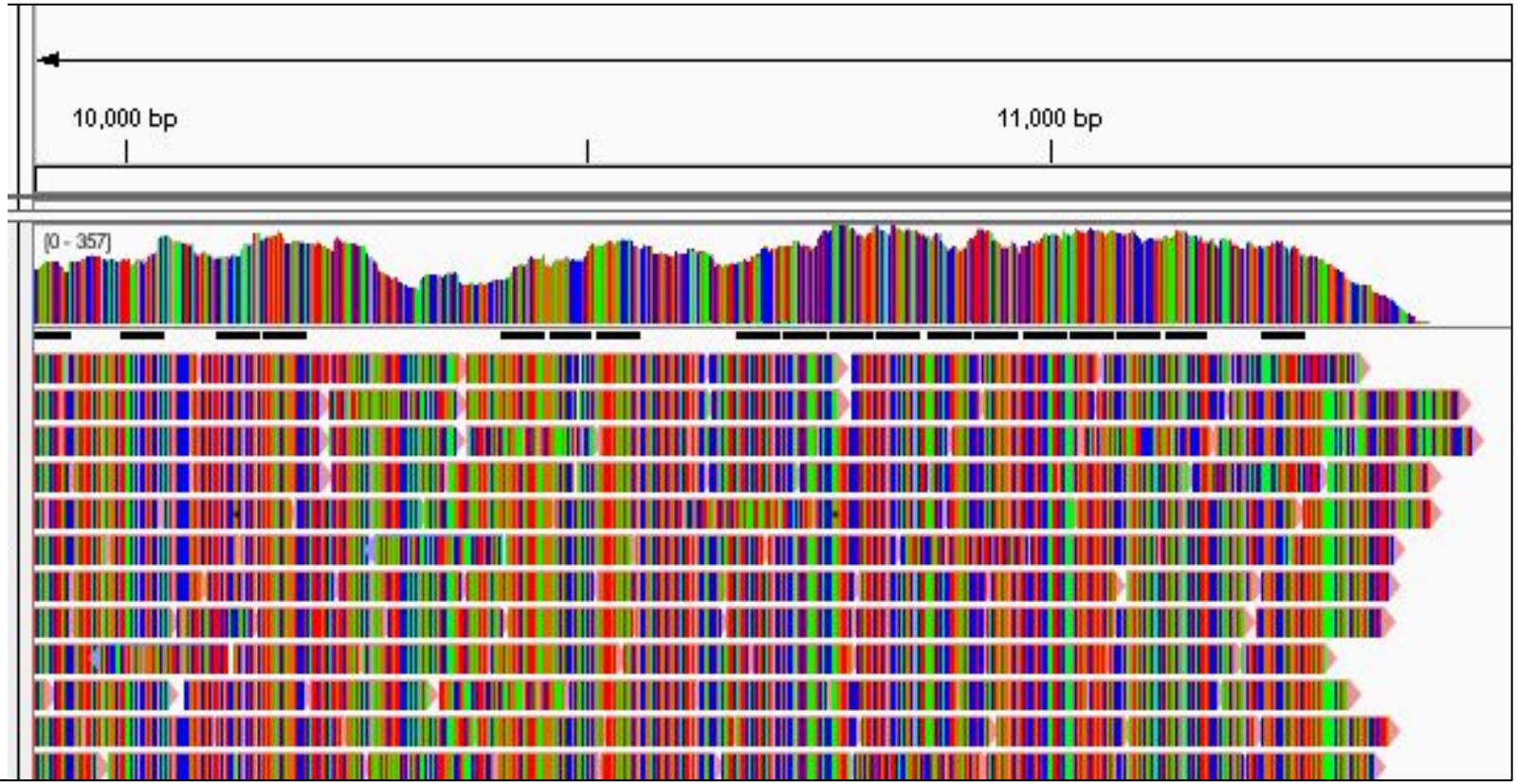
Badly aligned reads generate false positives



FASTQ → BAM → VCF → Annot

Radboudumc

Question: What has gone wrong here?



- Wrong reference genome used (e.g GRCh37 versus GRCh38)
- All bases show up as a variant

Frequently mutated genes

From Shyr. et al. BMC Medical Genomics (2014)

- Note that these genes are frequently mutated because of either:
  - Alignment issues (e.g. MUC16)
  - The sheer size of the gene. So these genes may still be relevant for your specific case (e.g. TTN)

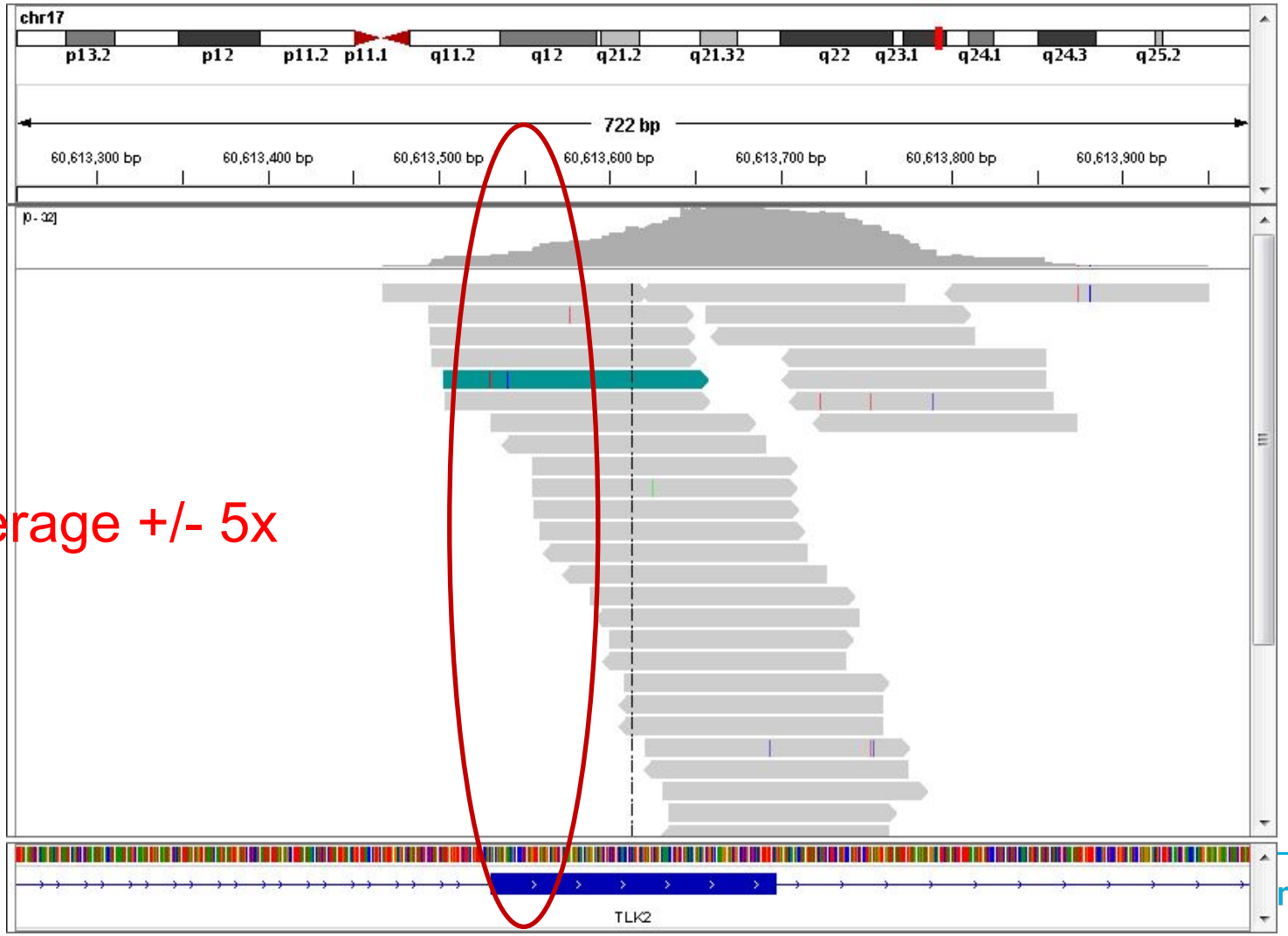
Take home message

- Mapping is not simple: try to do it right
- Be aware of mapping artifacts

Sequence coverage

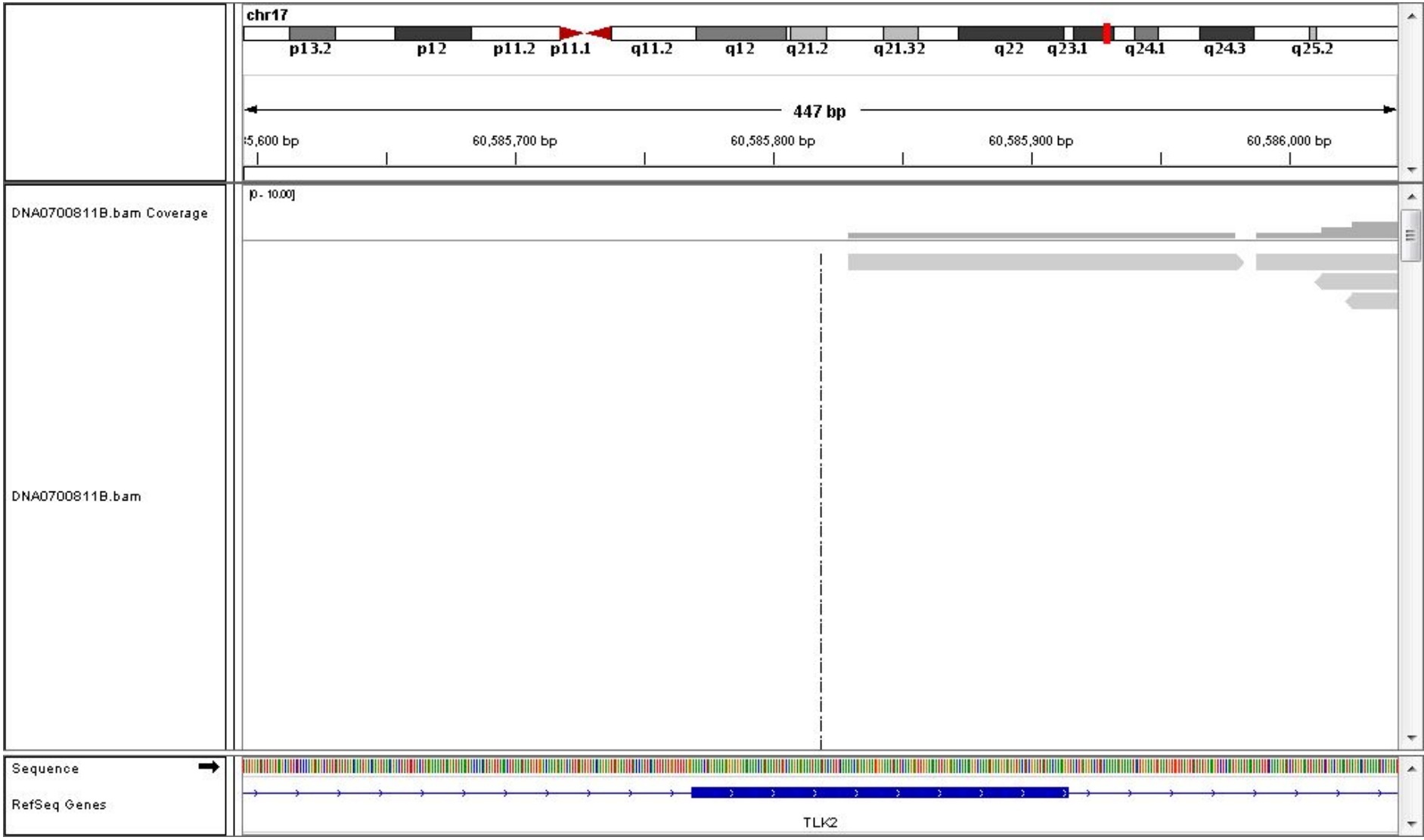
False negatives - coverage

Average coverage exome: 100x;
Average coverage exon: 23x



Coverage +/- 5x

TLK2 exon 3 – no target



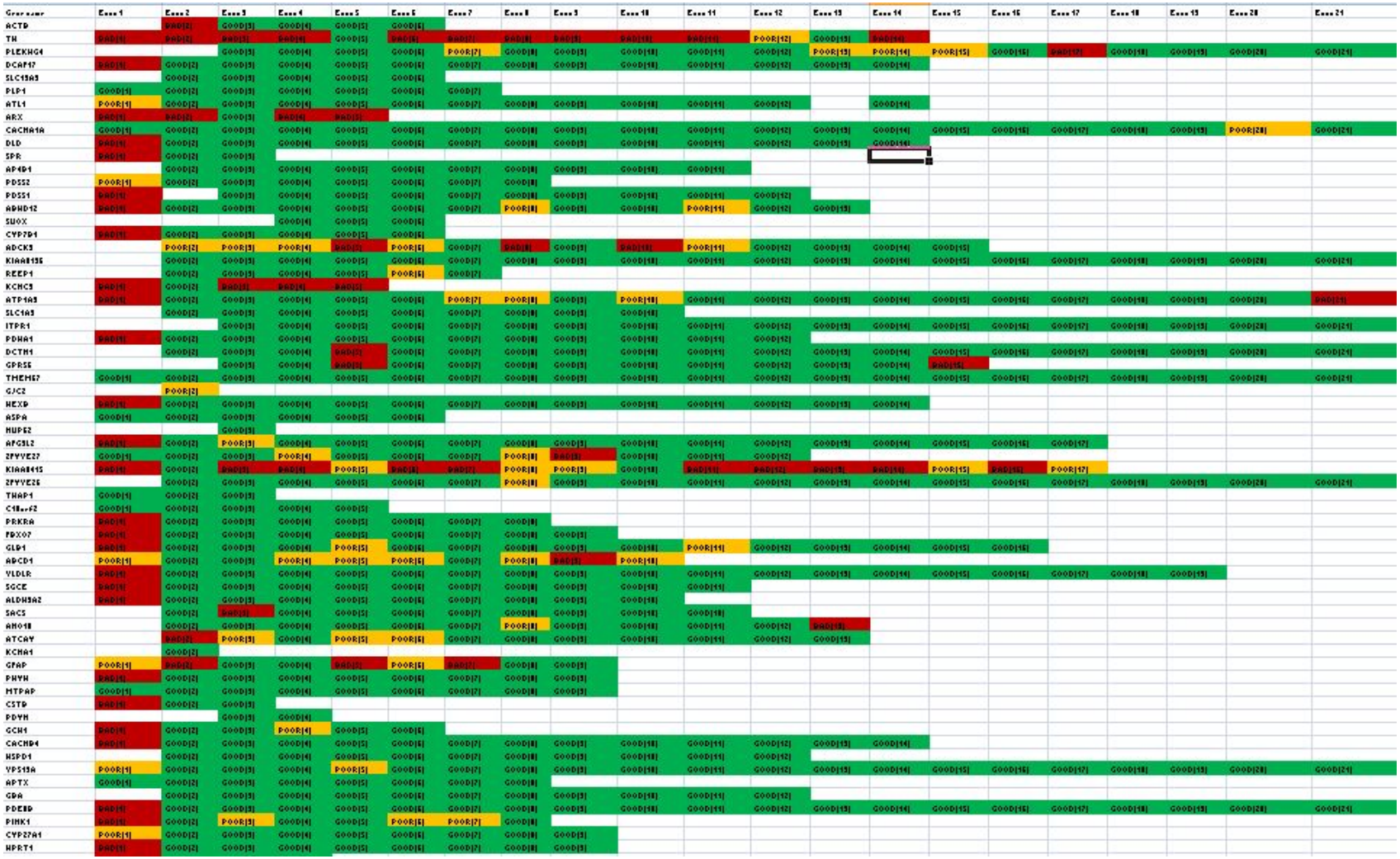
FASTQ → BAM → VCF → Annot

Radboudumc

Enrichment problems

Exons

Genes



Sequence coverage

Coverage: the number of times a single position on the DNA has been interrogated by a sequence reads

Why do we need coverage?

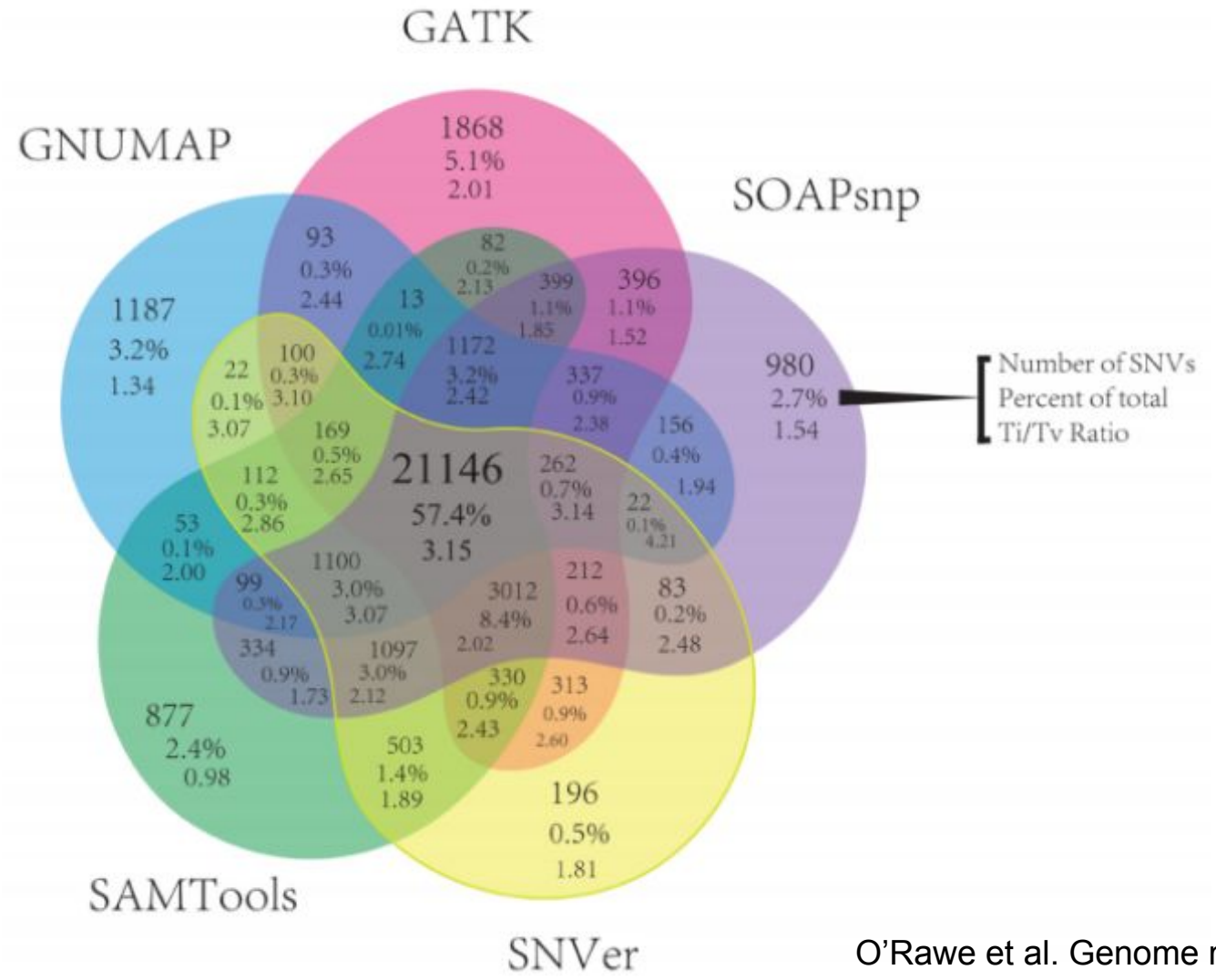
1. There are two alleles that we sequence at random: we need to be sure that we see each allele at least once.

2. We want to be able to distinguish variants from sequencing errors.

3. Some regions don't enrich very well, if we sequence more we will have a higher chance of sequencing these regions as well.

Variant calling

Variant calling... Easy?



O'Rawe et al. Genome med. 2013

FASTQ → BAM → VCF → Annot

Radboudumc

Real or not?

Variants

Are they true or not?

Yes = TRUE VARIANT
No = FALSE VARIANT

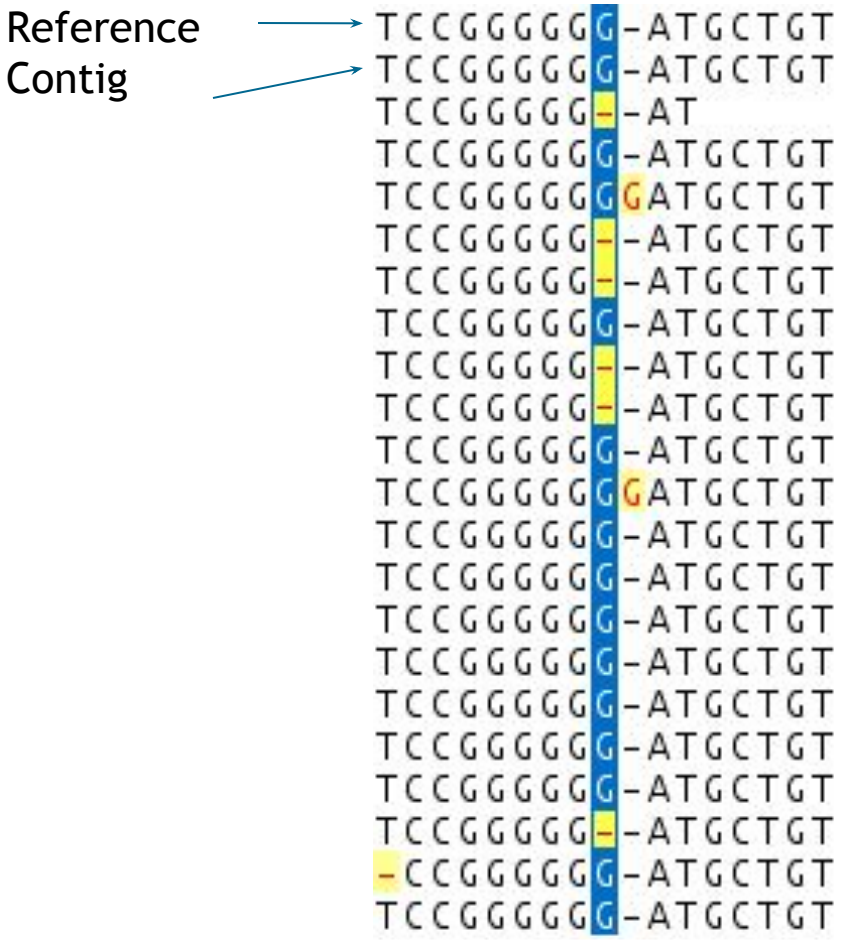
Variation detection (1/5)

Is this variant real?

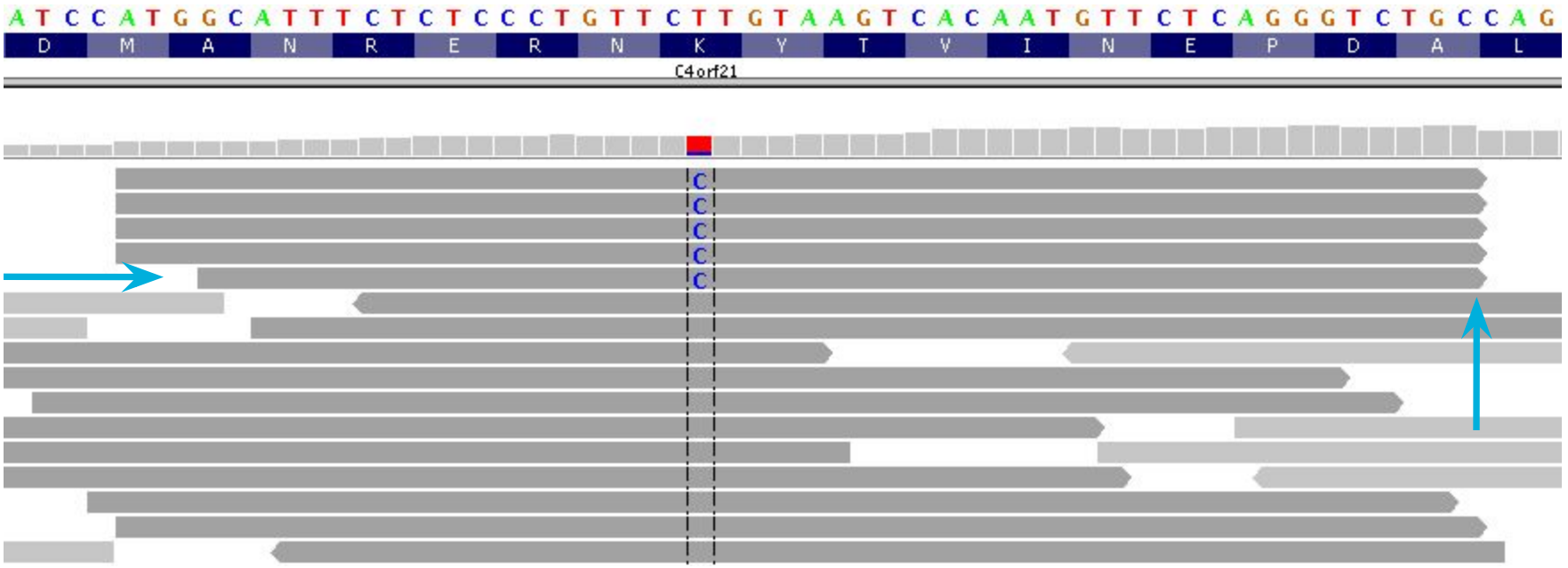
FASTQ → BAM → VCF → Annot

Radboudumc

Variation detection (2/5)



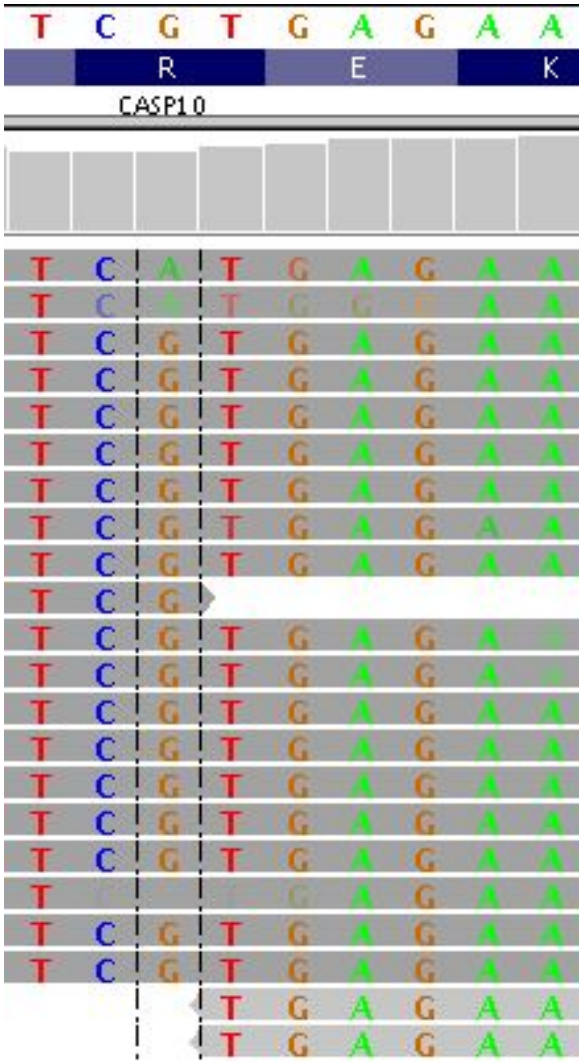
Variant detection (3/5)



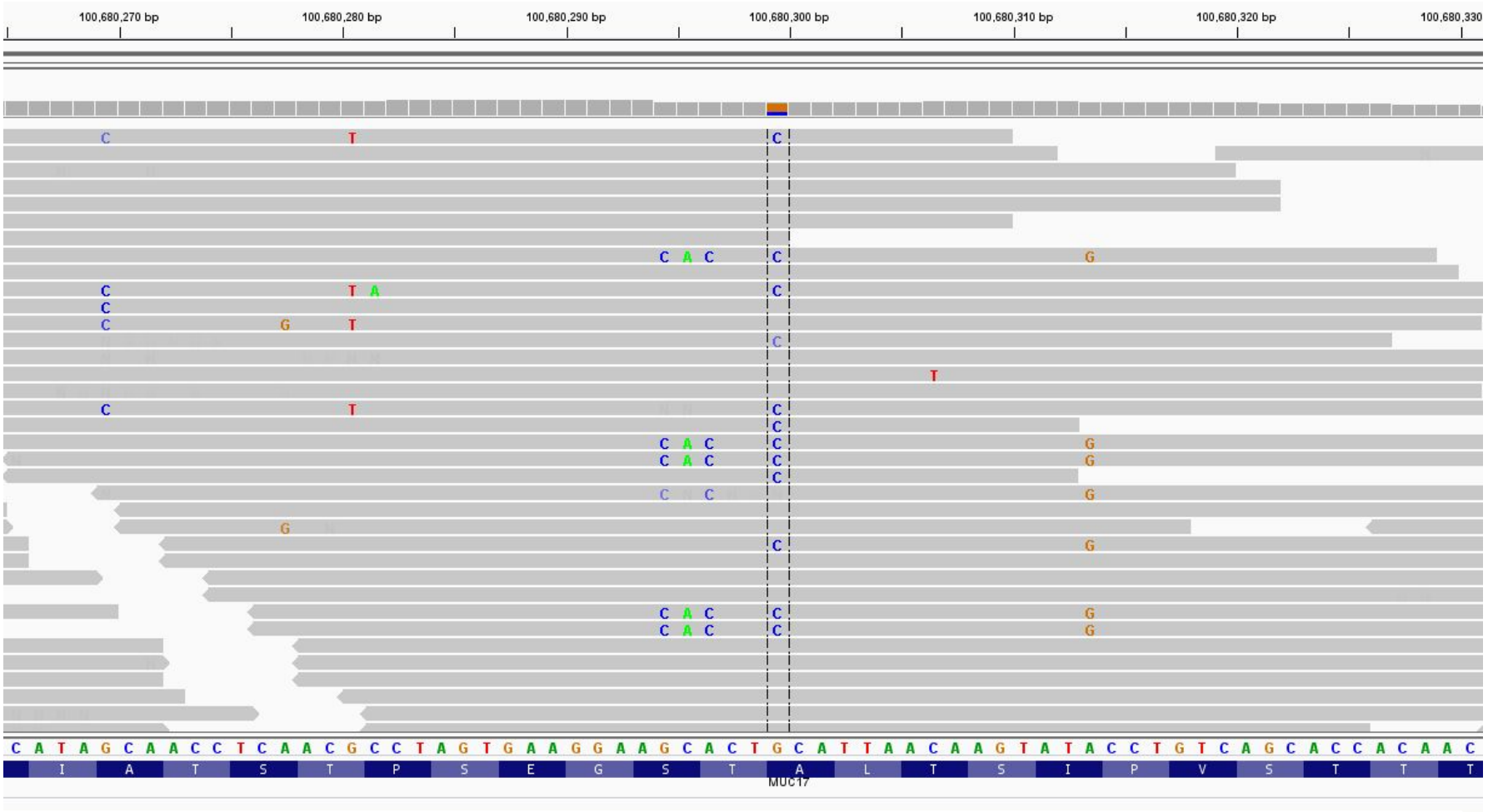
Real variant?

Variant detection (4/5)

Real?



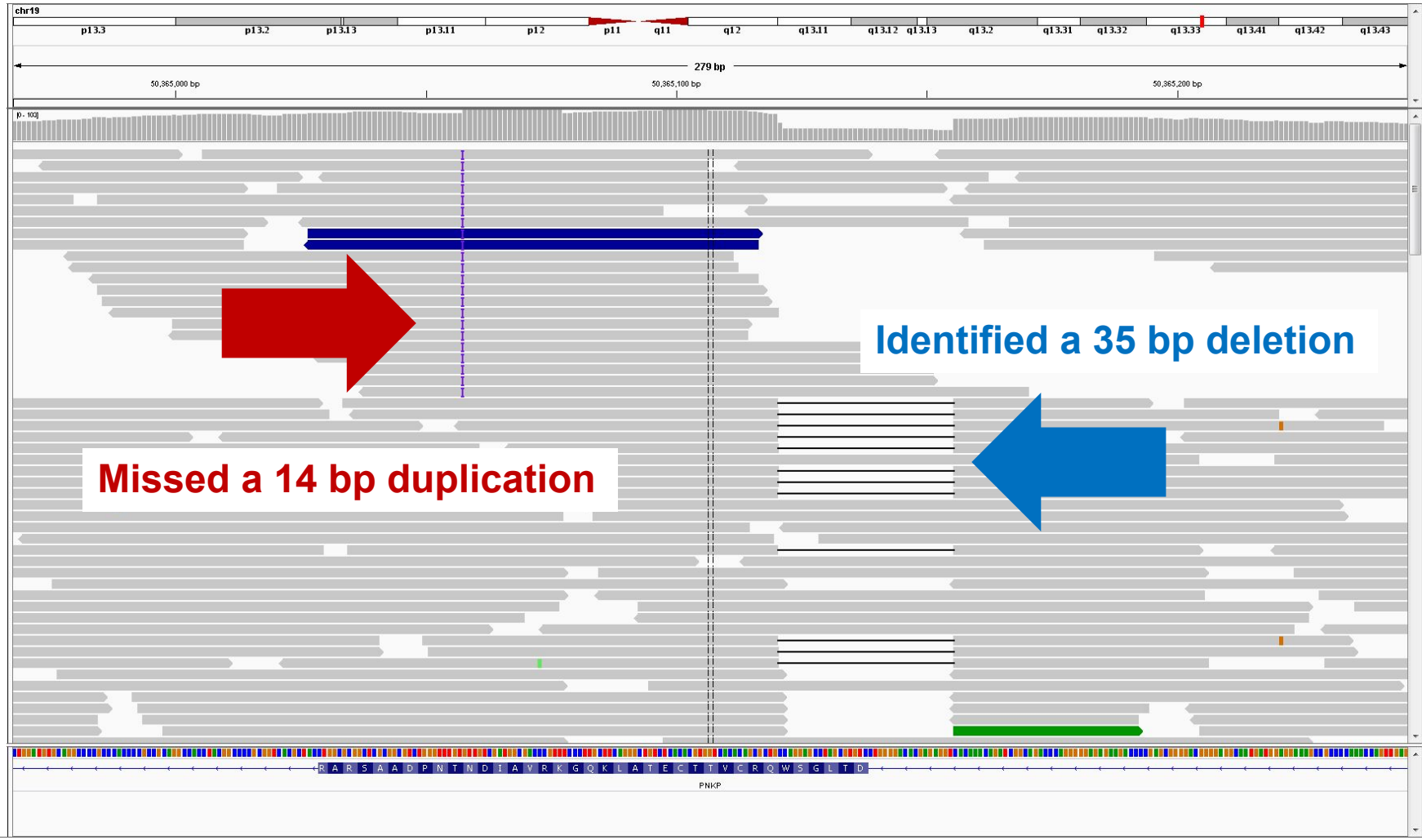
Variant detection (5/5)



FASTQ → BAM → VCF → Annot

Radboudumc

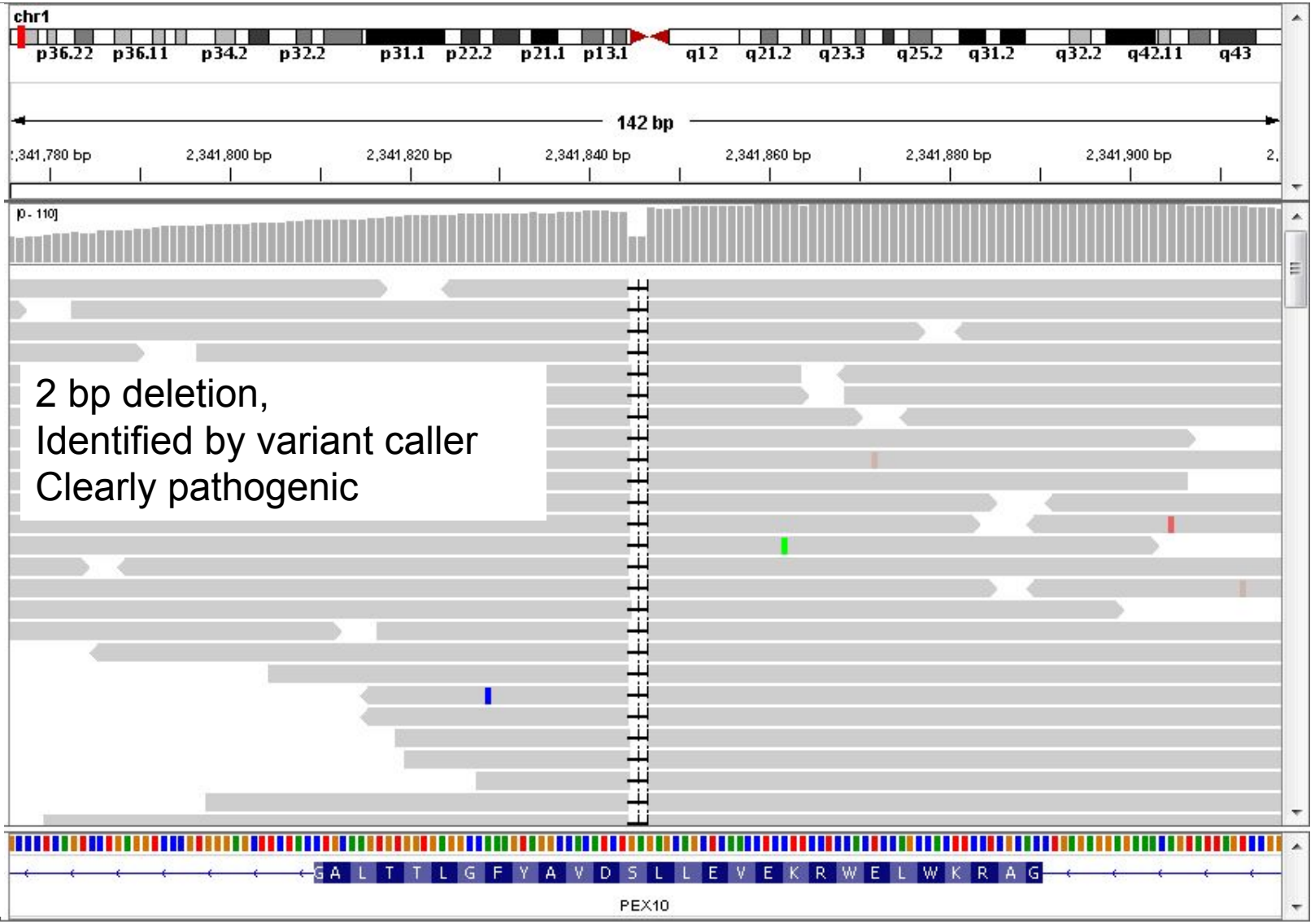
Complex variants



FASTQ → BAM → VCF → Annot

Radboudumc

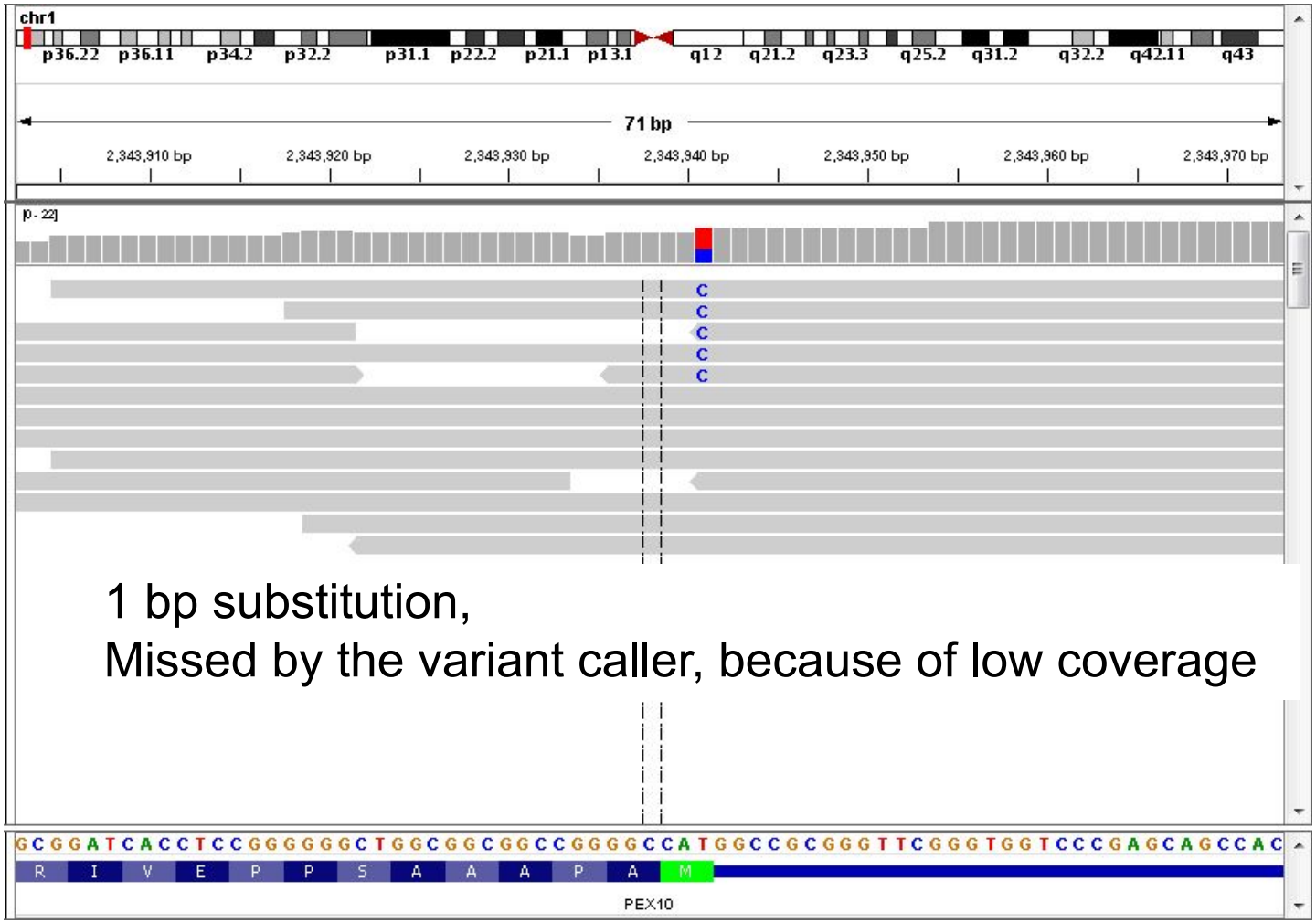
Mutations in recessive *PEX10* gene



FASTQ → BAM → VCF → Annot

Radboudumc

Mutations in recessive *PEX10* gene

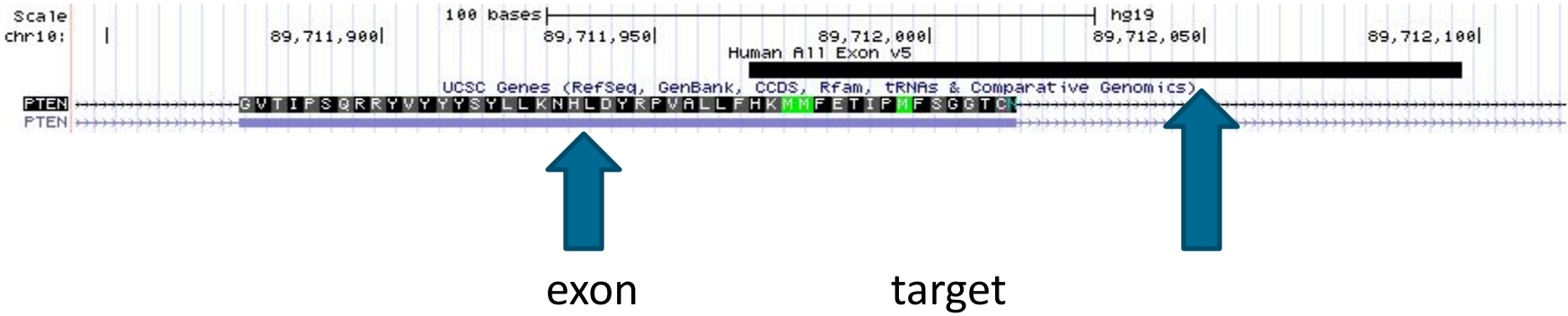


FASTQ → BAM → VCF → Annot

Radboudumc

Target regions

- Calling was done by BGI, using the exome kit targets, but:



- We missed 11% of coding sequence.

- **Solution**: We call all variants on targets +/- 200b

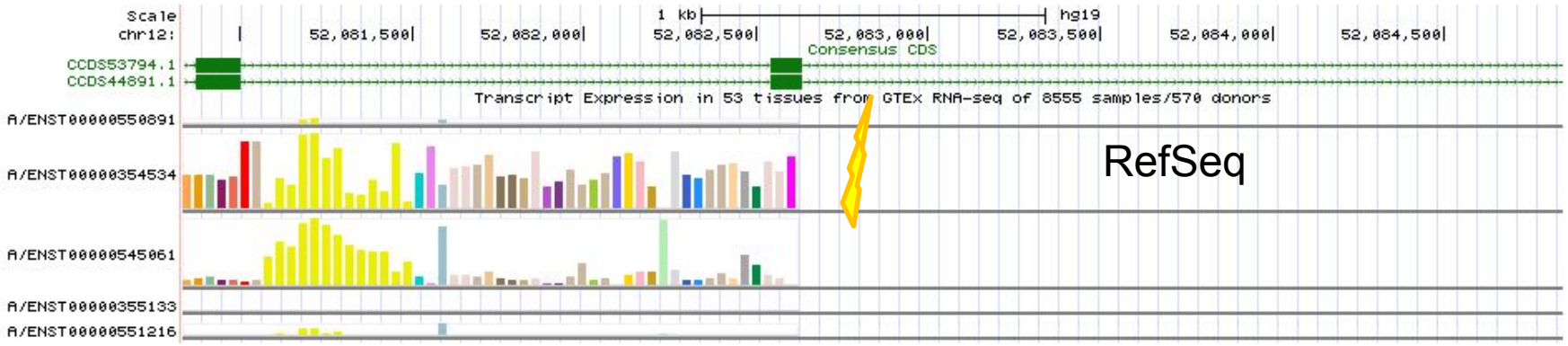
Take home message

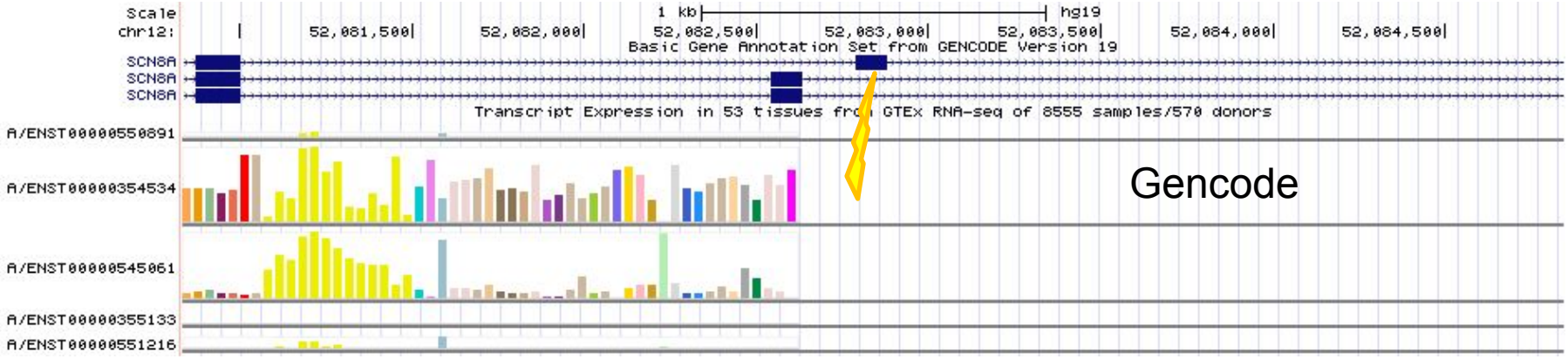
Variant calling is not perfect!

Multiple approaches and re-analysis may be required

Variant annotation

Gene annotation - *SCN8A*

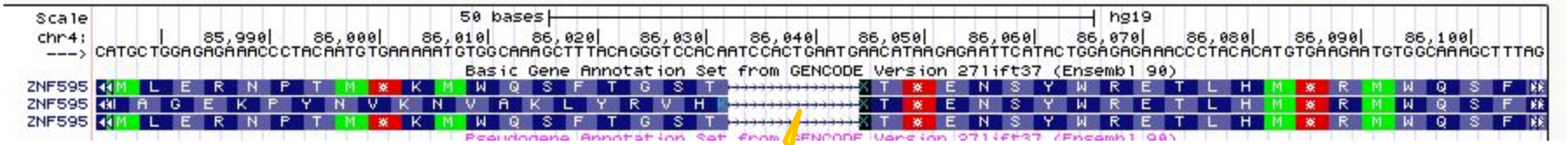




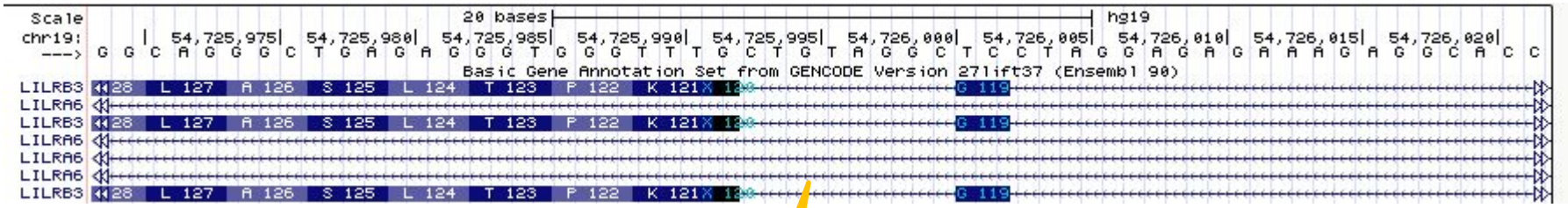
Solution: We moved from RefSeq to Gencode, but still we check the annotation for all genes.

Gene annotation issues

ZNF595



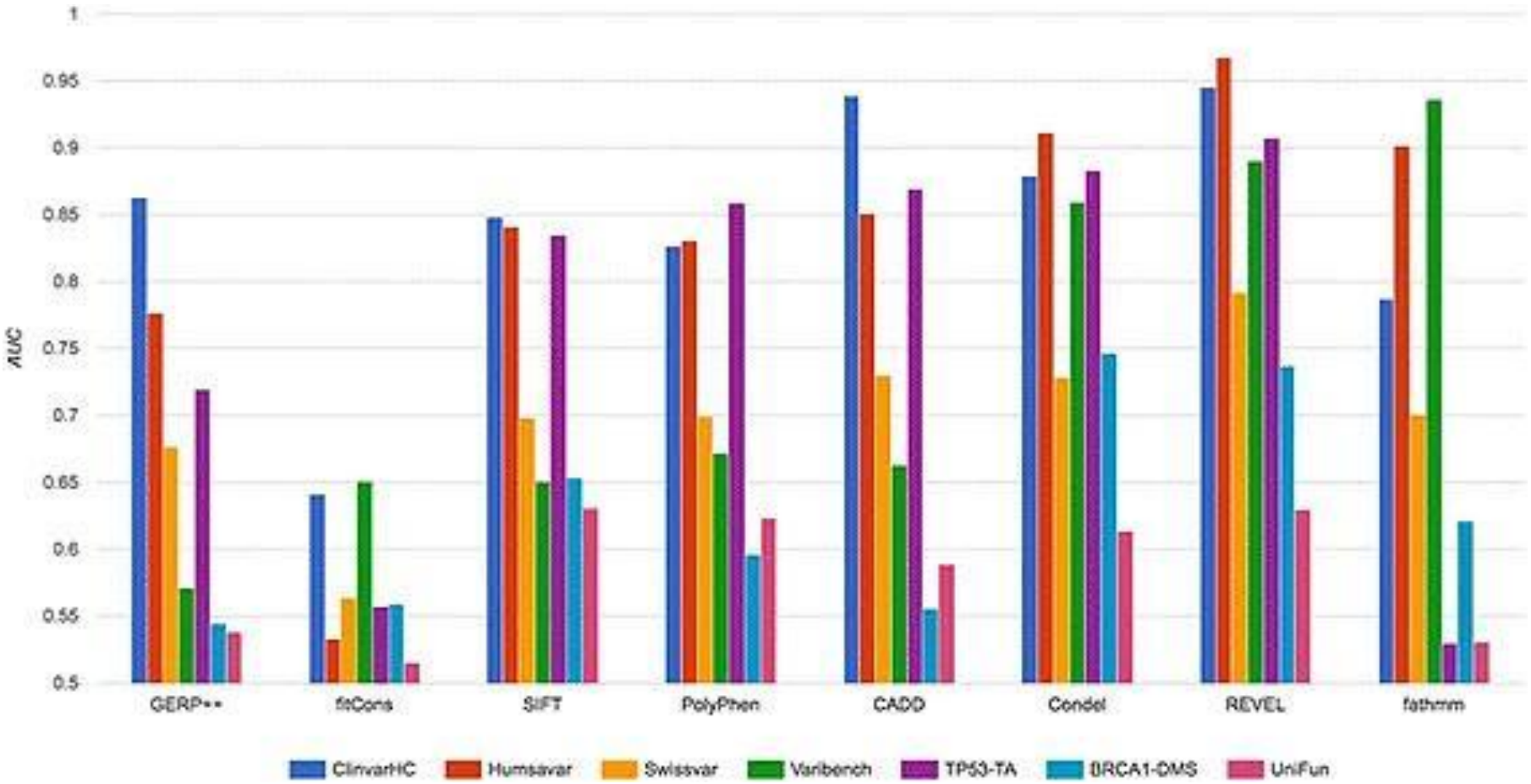
LILRB3



Coding or non-coding mutation?

Solution: we moved from in-house effect prediction to the Ensembl VEP

Variant prediction tools



Mahmood, K. *et al. Hum Genomics* **11**, 10 (2017).

Take home messages

- Always check the quality of your samples
  (Use standard tools like Qualimap)
- Be aware of mapping artifacts
- Always check coverage of samples and genes
  (80x exomes, 40x genome)
- Variant calling is not perfect
  (Perform re-analyses)
- Gene annotations are not perfect
  (Perform re-analyses)

Corominas *et al.* Clinical exome sequencing-Mistakes and caveats. Hum Mutat. 2022 Feb 22. PMID: 35191116.