CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Moderní metody analýzy genomu
Bioinformatika II

Karol Pál

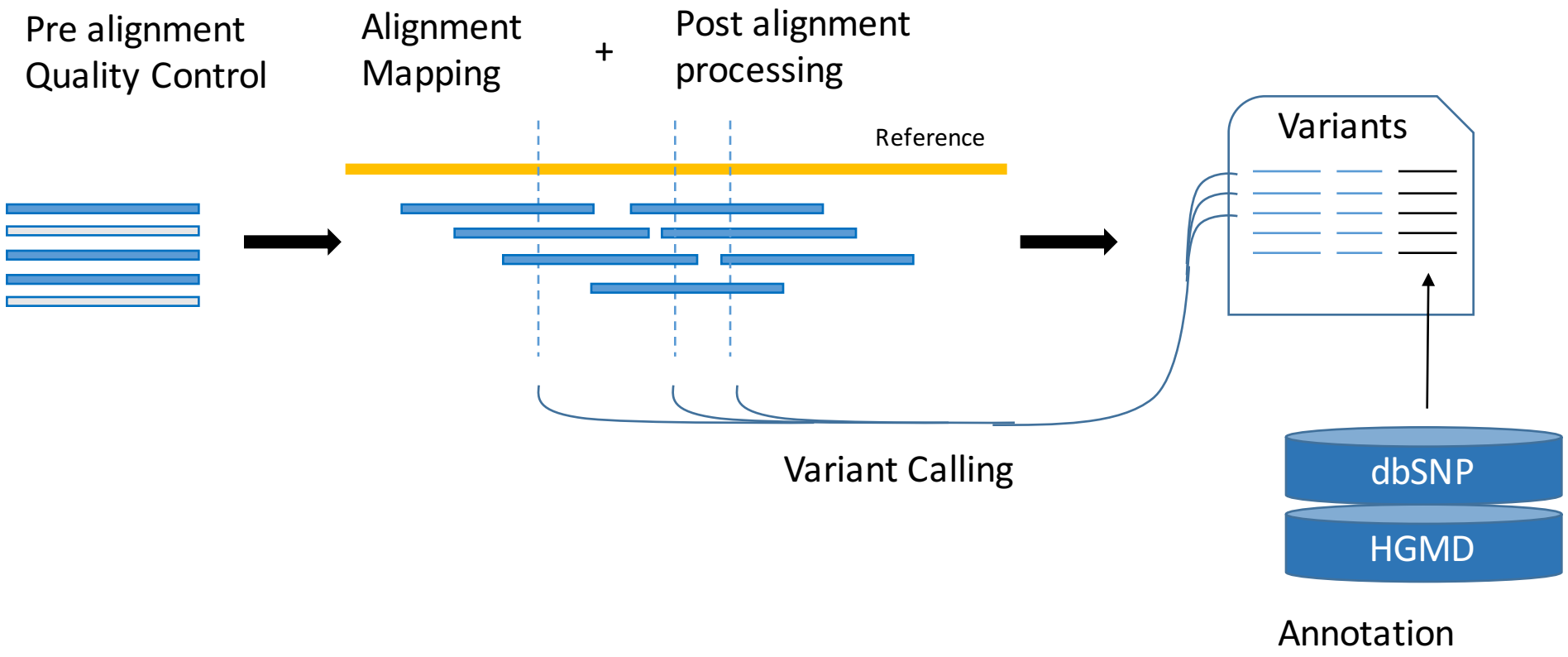
Brno, 25.11.2016

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

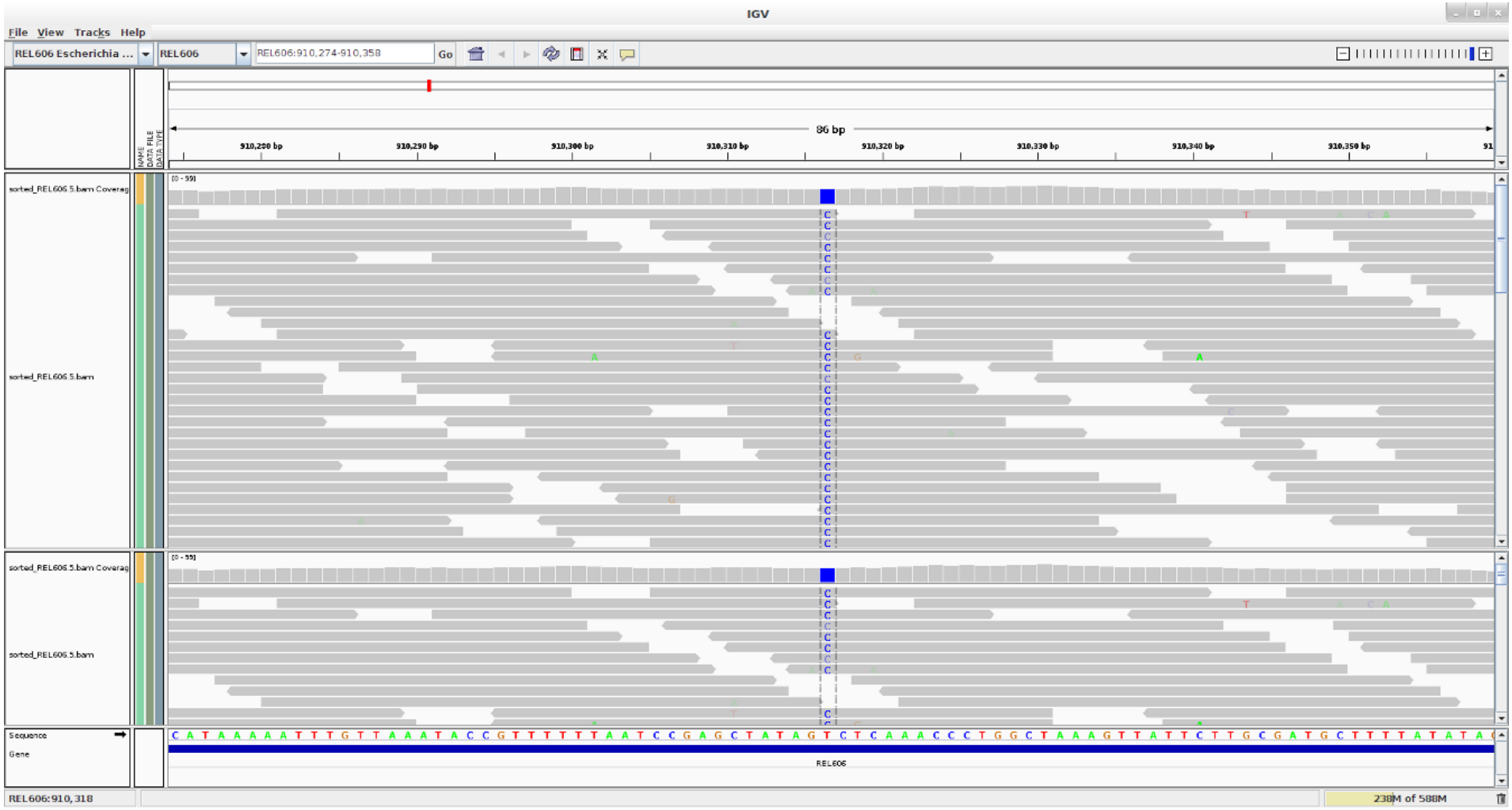
2007–13
OP Research and Development for Innovation

Recap from last lecture

- NGS data analysis
  - Commercial software vs “In house” pipeline
  - (Different license for different kind of experiment)



IGV (Showing a bam file)



Experiment design

- Modifications to the basic **pipeline** depending on **input material** and expected **output**

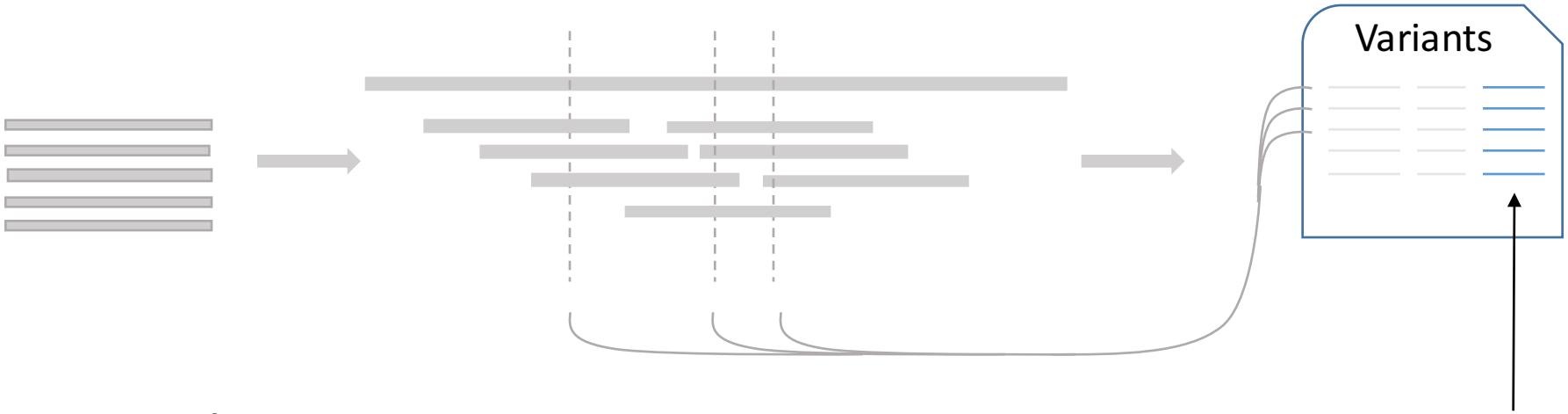
- Next-gen Sequencing
  - Whole Genome Sequencing (WGS)
  - Targeted Sequencing
    - Whole Exome Sequencing
    - Gene panels
    - PCR based
  - RNA Sequencing

- 3rd generation Sequencing
  - Various applications

Exome Sequencing (WES)

- Input material :
  - Coding regions + small RNA's
  - Represents 1% of DNA (human)
- Coverage ~ 80x
- Scenarios:
  - Single individual
  - Family with phenotype
  - Paired cancer + healthy tissue from one individual

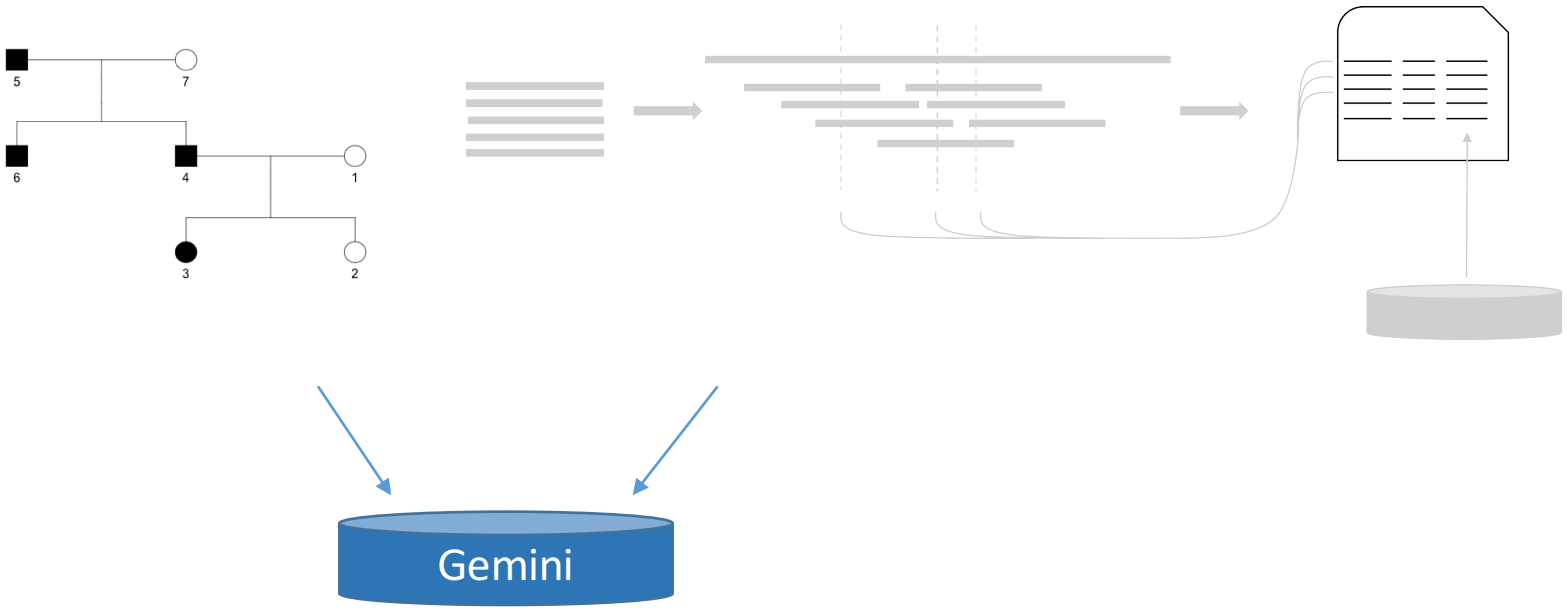
WES single individual



Expected output:

- (rare) germline variants
  - **ExAC**
  - **ESP6500**
  - **Kaviar**
  - ~~**dbSNP**~~
- Inherited disease
  - **HGMD**
  - **CLINVAR**

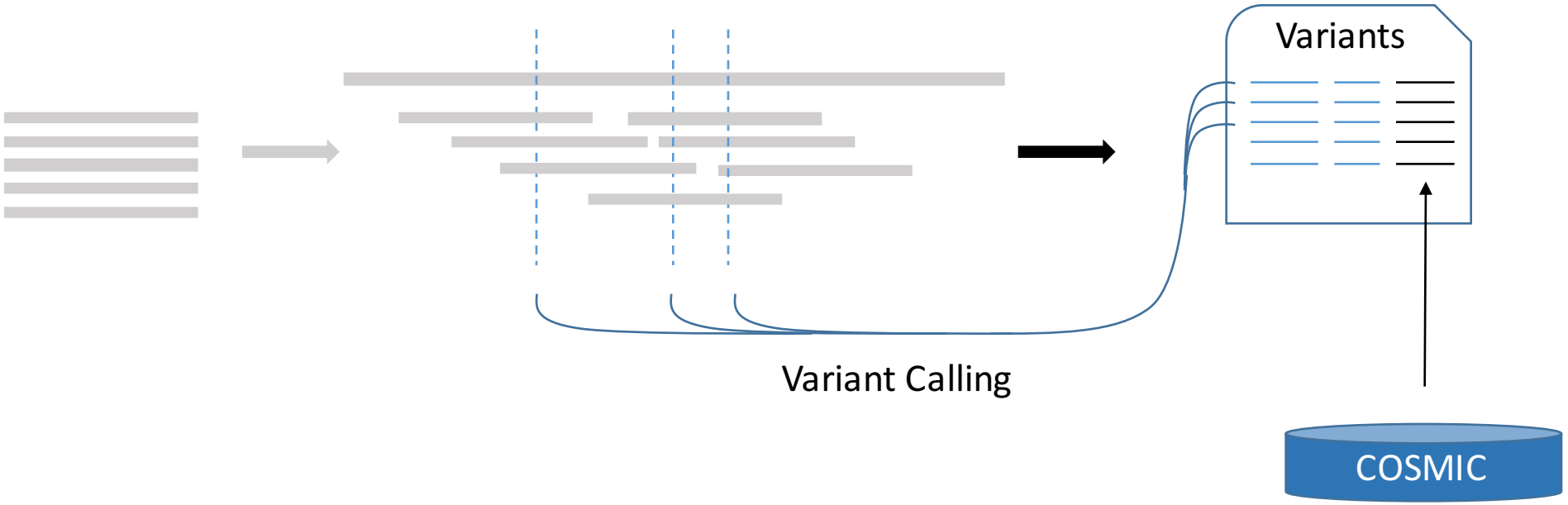
WES Family with phenotype



Expected output: find disease causing mutations
GEMINI = SQLite database

- SQL queries
- Preformatted queries for Autosomal Dominant and Autosomal Recessive phenotypes

WES paired samples (healthy + cancer)



Expected output: Somatic mutations
WES not ideal (low sensitivity)

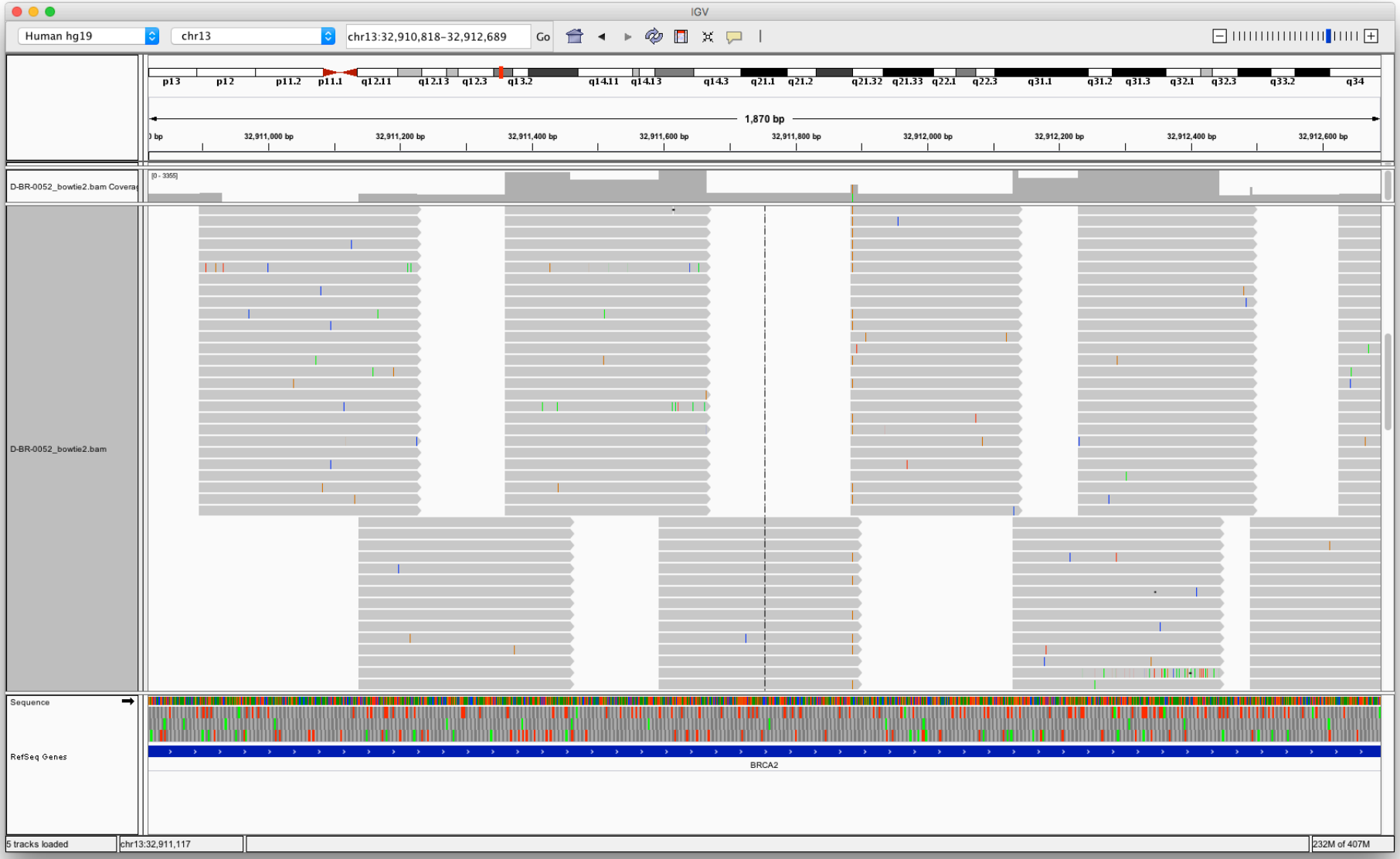
Mutect:

- Variant caller specialized for somatic point mutations in cancer genomes
- Takes two bam files as input

Targeted sequencing PCR + Panels

- Higher coverage (up to 10 000x)
- High sensitivity - somatic variants VAF up to 0.1%
- Methods used in diagnostics
  - CZECANCA (panel) targeting 219 cancer susceptibility genes
  - BRCA (PCR)

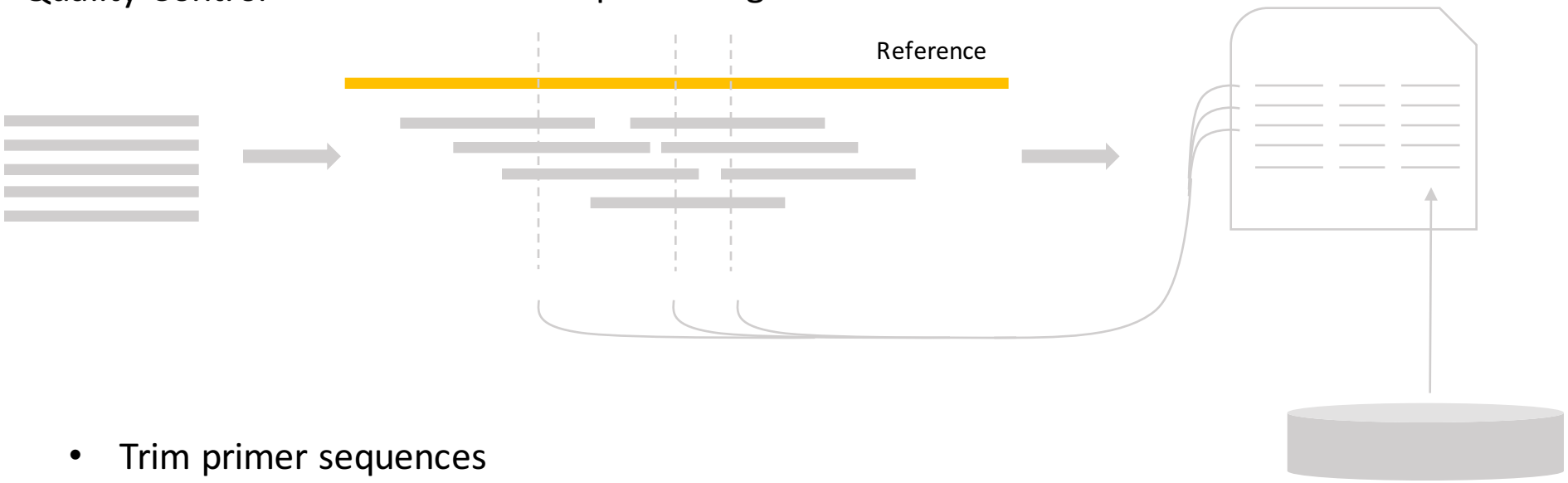
Targeted sequencing PCR based



Targeted Sequencing

Pre alignment
Quality Control

Post alignment
processing

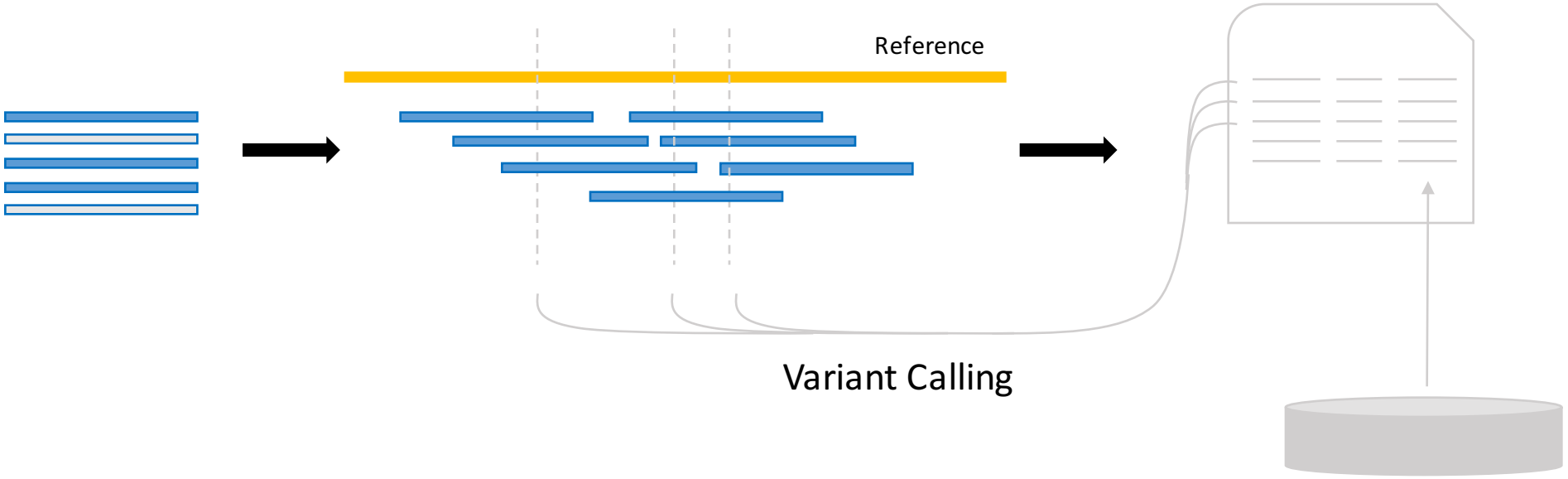


- Trim primer sequences
- Do not remove PCR duplicates (PCR based)

RNA Seq

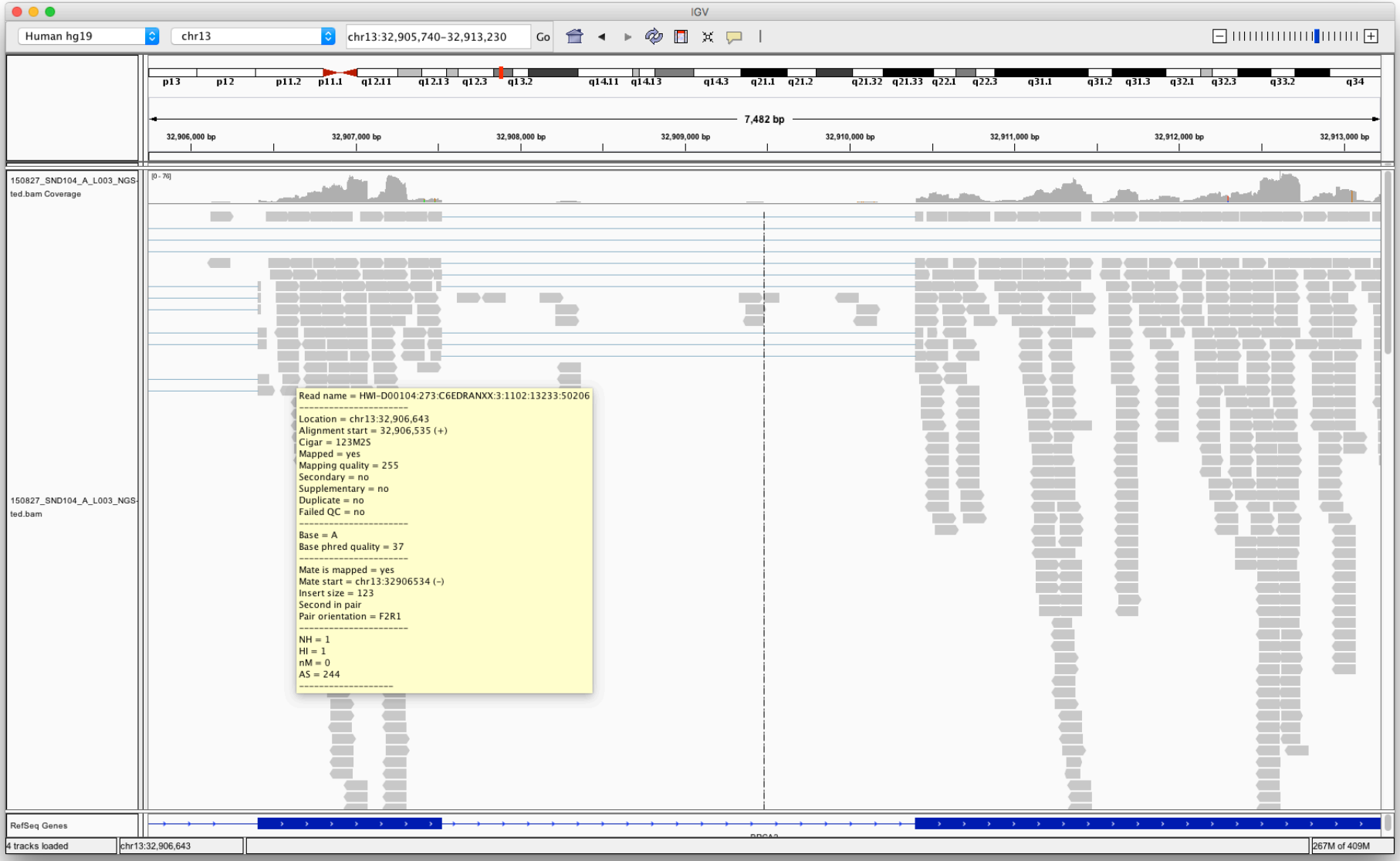
- Input material: cDNA
  - Poly A tailing
  - Specific capture
- Possible outputs:
  - Expression levels (RPKM,FPKM or tag counting)
  - Structural variants
  - (Variant calling)

RNA Seq gapped alignment



Gapped alignment vs. alignment to “transcriptome”

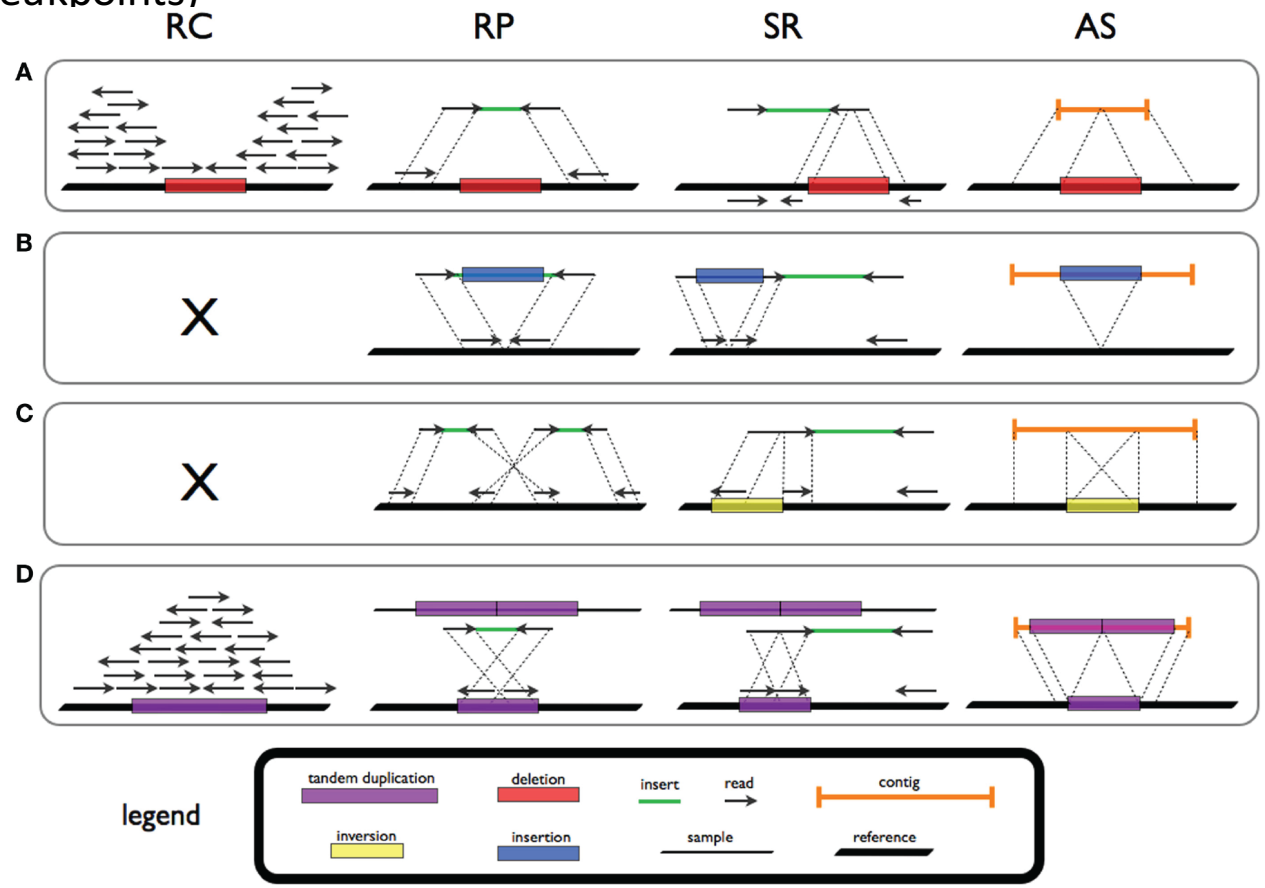
RNA Seq gapped alignment



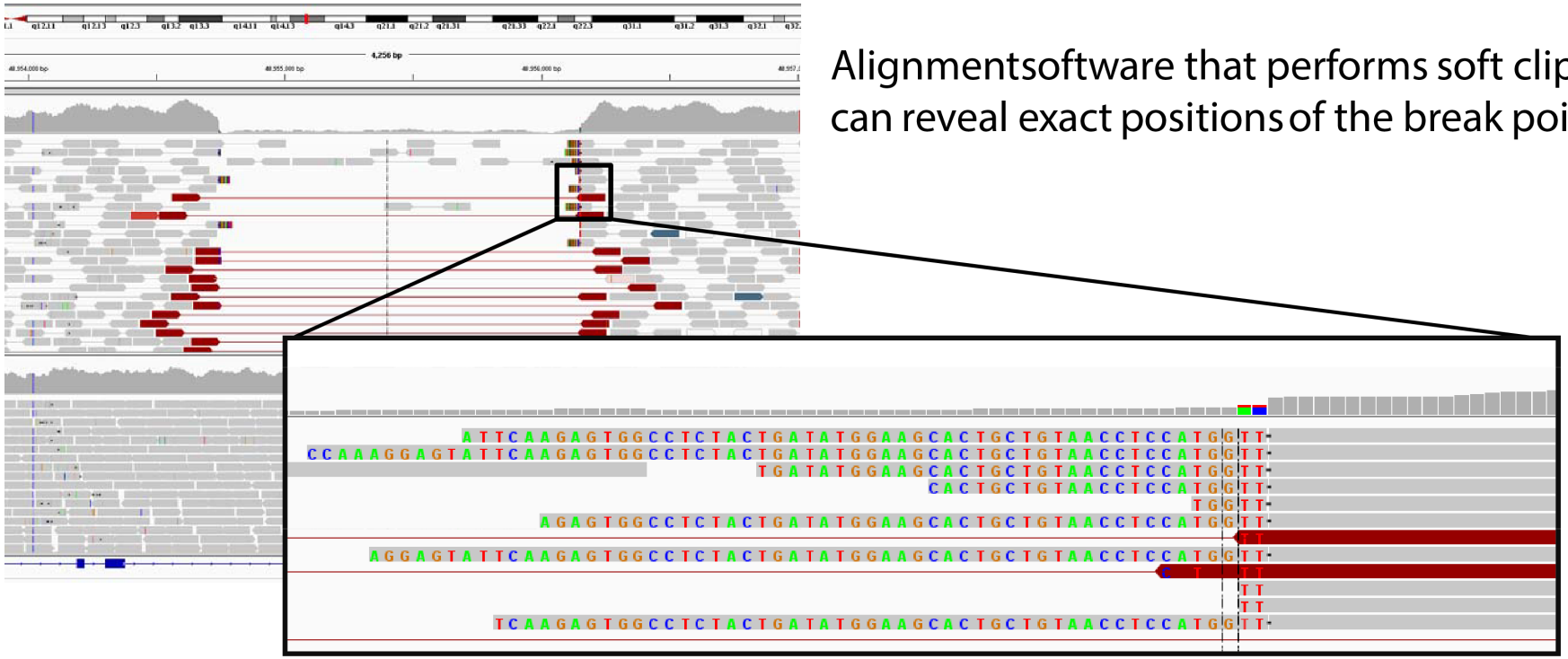
WGS structural variants

Methods used

- Read counts (Coverage) can be used for WES
- Read pair (span and orientation of reads)
- Split reads (breakpoints)
- Assembly

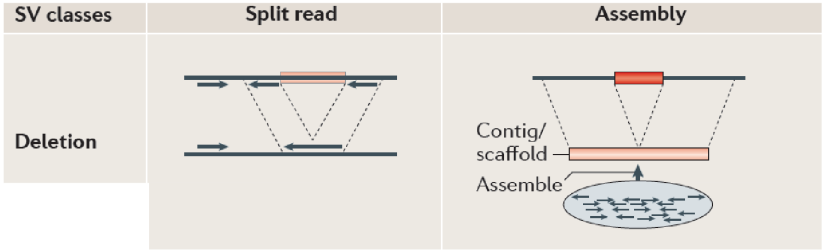


Clusters of soft clipping indicate rearrangement break points



Alignment software that performs soft clipping can reveal exact positions of the break points

Further realignment of the clipped sequences produces split reads

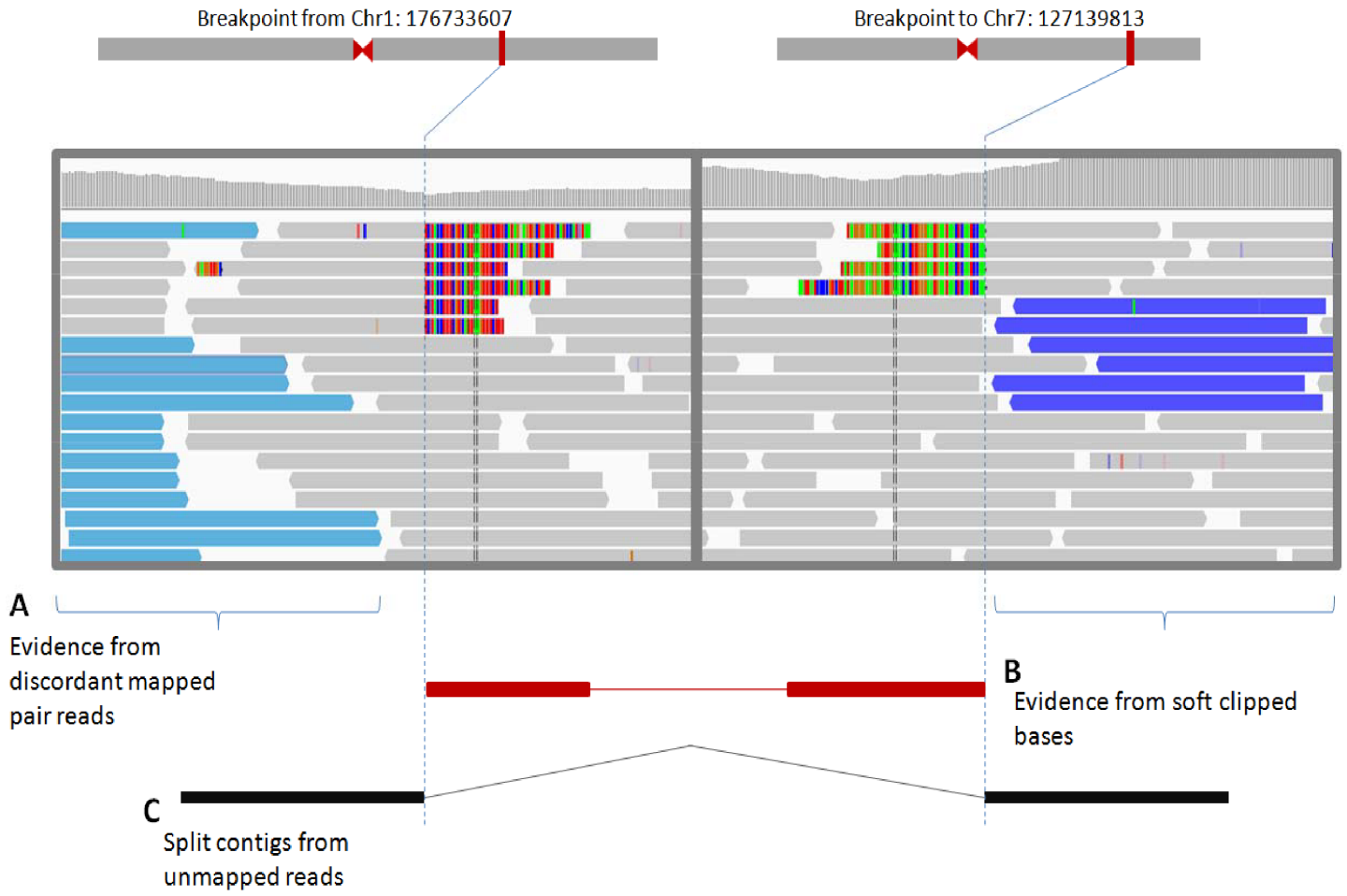


Reads with soft clipping and unmapped reads can be assembled into contigs that span break points

qSV : Detecting Somatic Structural Variants

qSV detects 3 types of supportingevidence

Resolves all lines of evidence to identify breakpoints to base pair resolution



Felicity Newell

Nanopore

- Long reads high error rate
- Scaffold for de novo assembly
- Transcripts
- Presence of pathogens

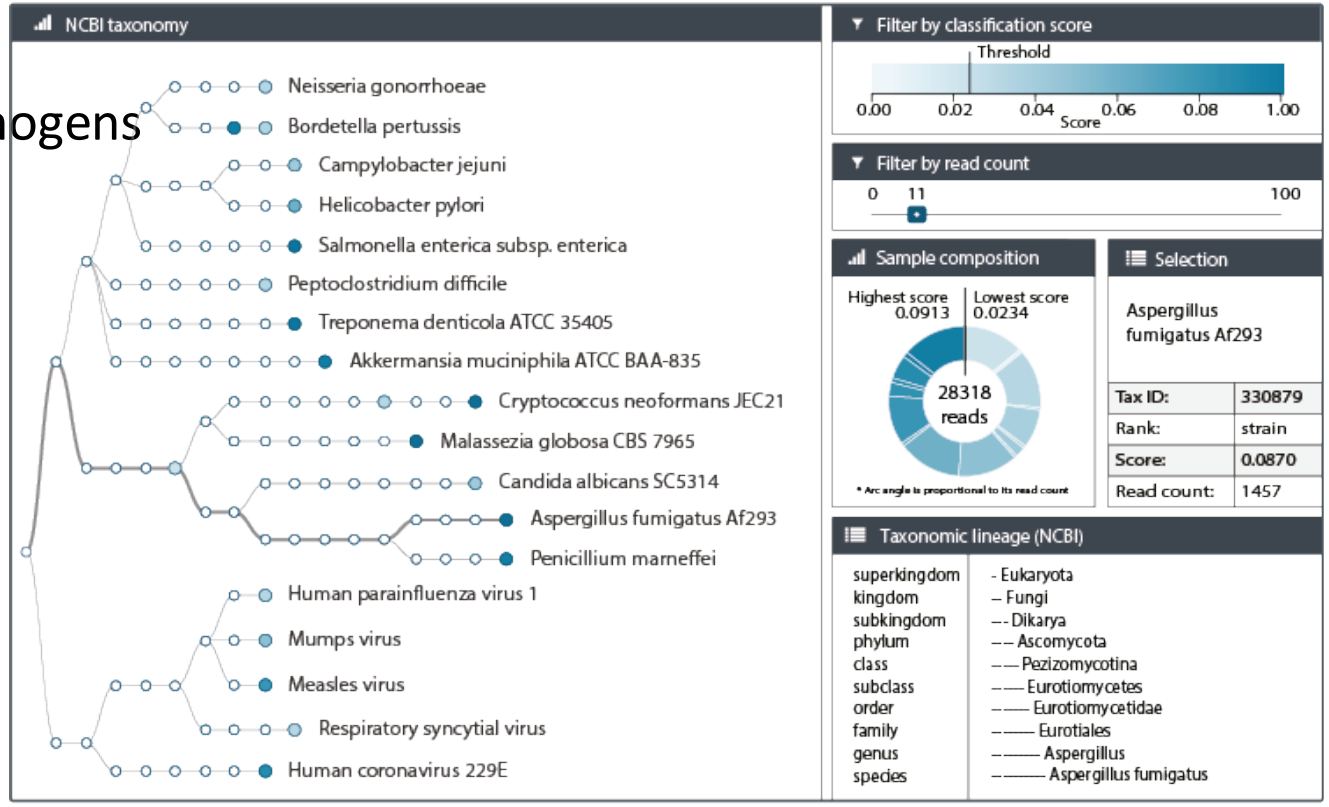


Fig. 1 Metrichor WIMP report, shown for a sample containing bacteria, viruses and fungi

Thank you for your attention

links

- http://exac.broadinstitute.org/
- http://evs.gs.washington.edu/EVS/
- http://db.systemsbiology.net/kaviar/
- https://gemini.readthedocs.io/en/latest/
- http://archive.broadinstitute.org/cancer/cga/mutect
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4479793/