CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

UNIVERSITAS MASARYKIANA BRUNENSIS

DNA re-sequencing - Analysis

Vojtěch Bystrý

18. November 2019

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and Development for Innovation

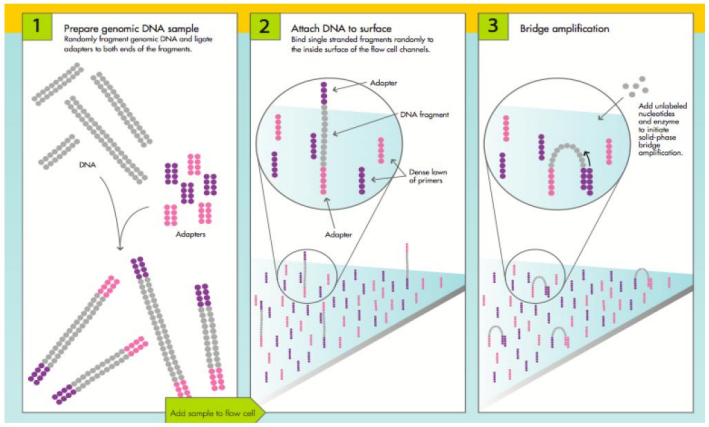
Goals of the presentation

- **Overview of NGS bioinformatics**
  - **NGS bioinformatics < Sequence analysis < Bioinformatics**
- **What to think about when you**
  - **plan experiment**
  - **discuss data analyses**
  - **check results**

- **Not to teach you how to do bioinformatics**

NGS Bioinformatics

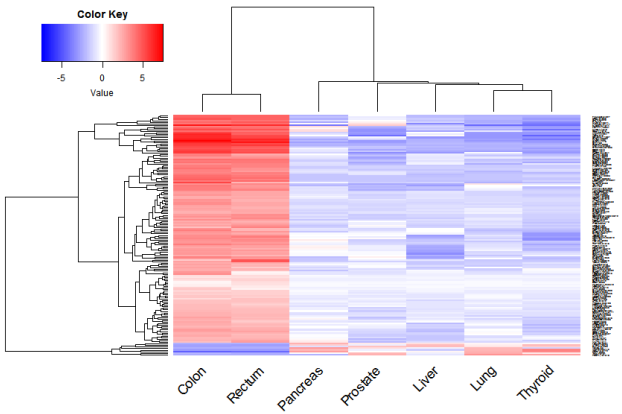
Illumina 1

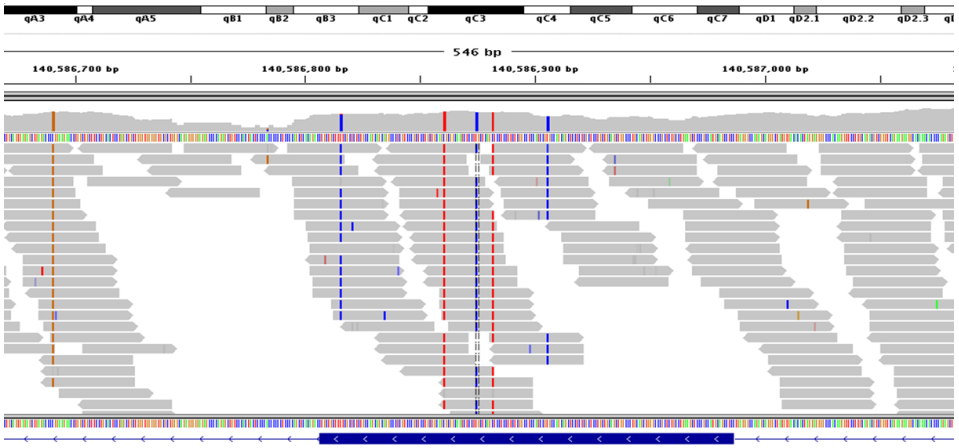


ECE/BioE 416
Lecture 24

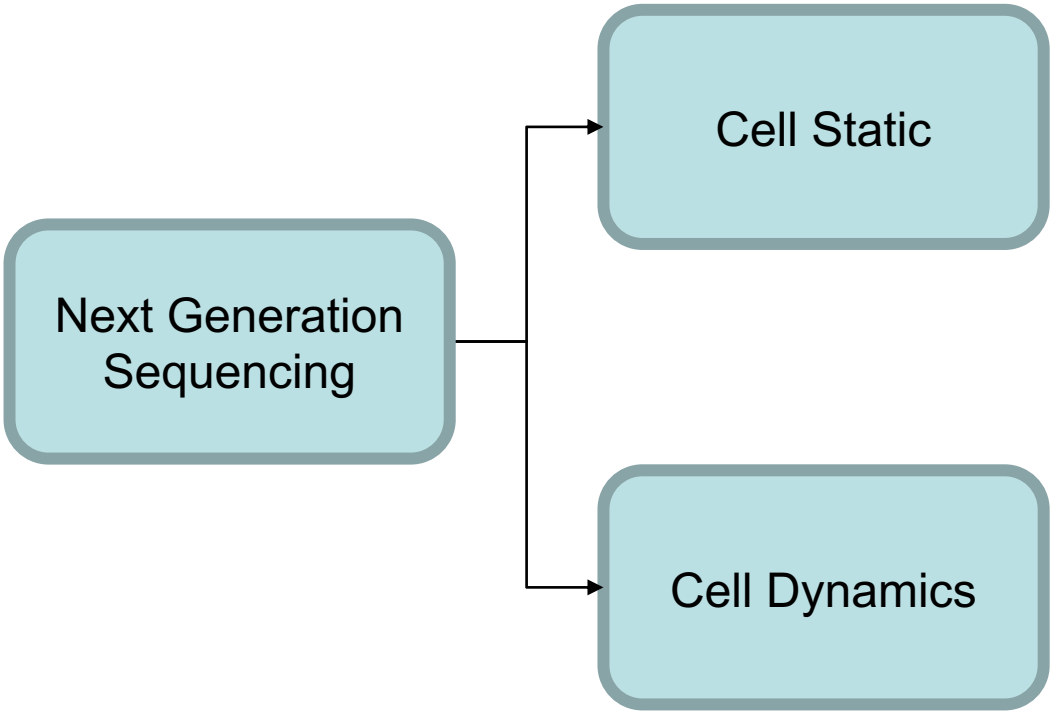
10

Your raw sequence data

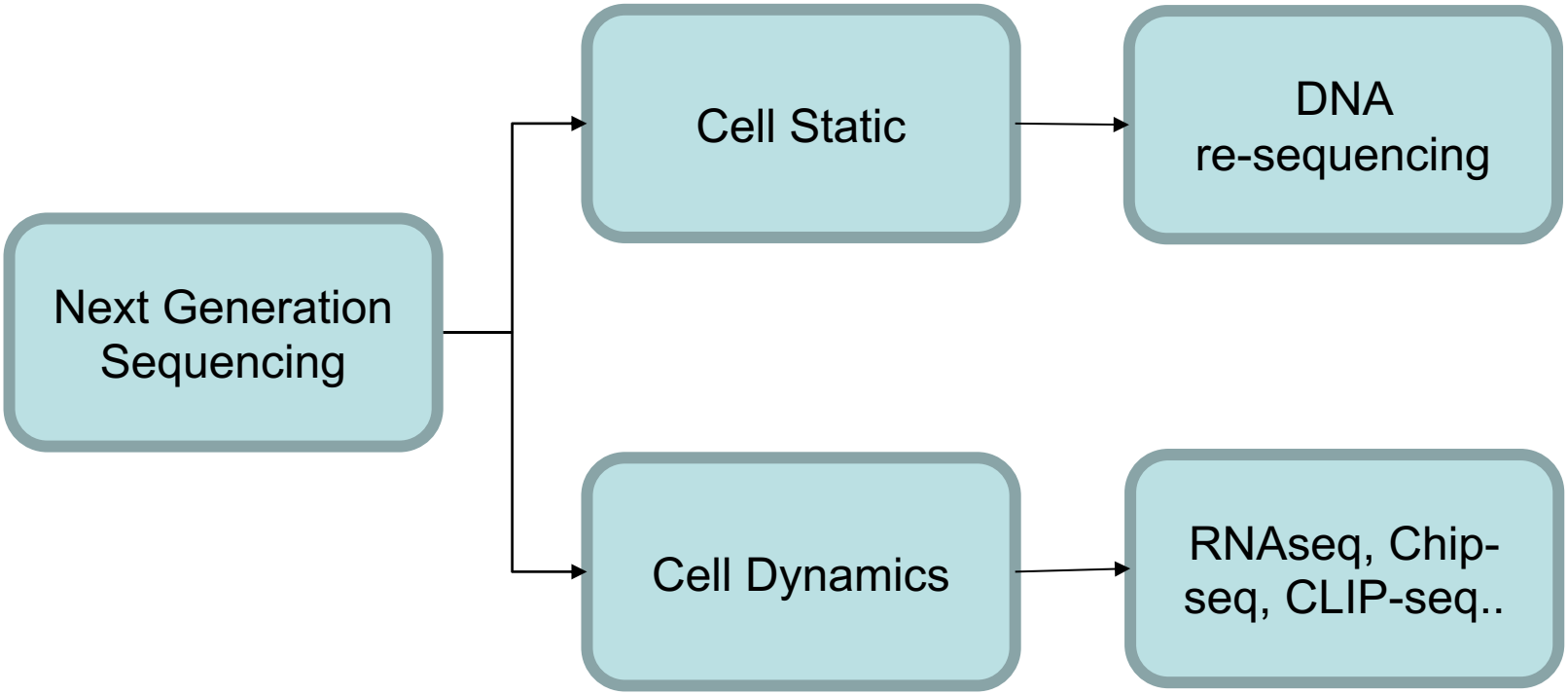




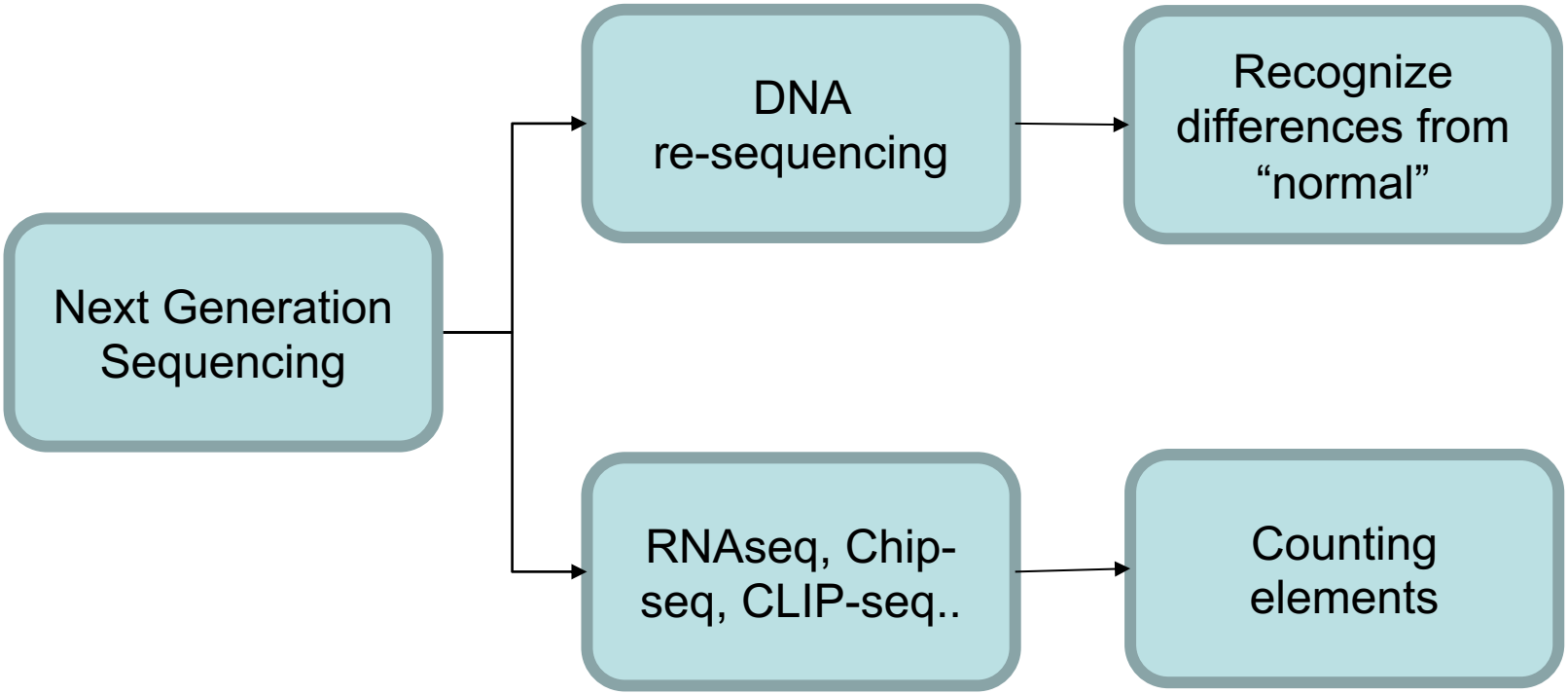
NGS experiments



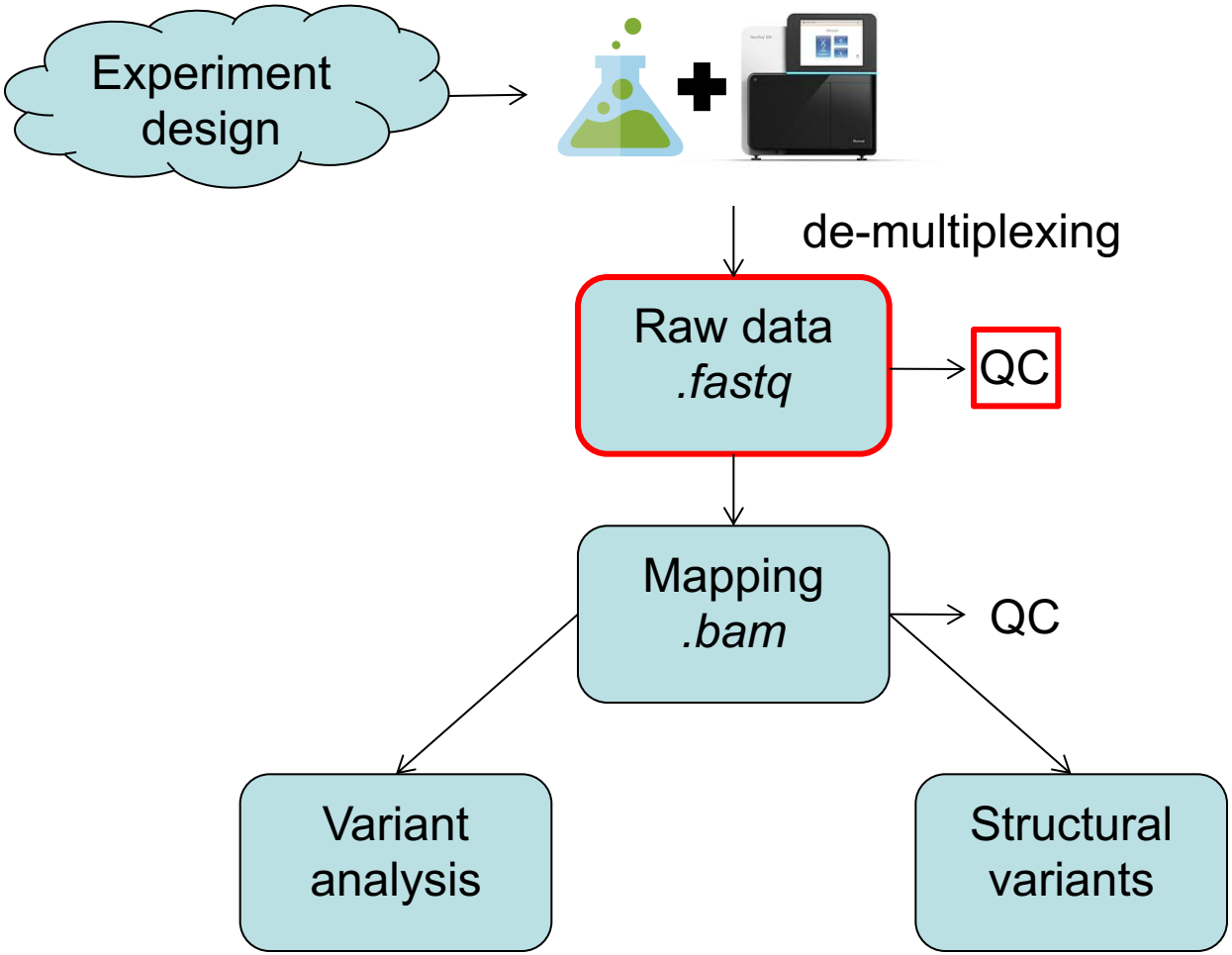
NGS experiments



NGS experiments



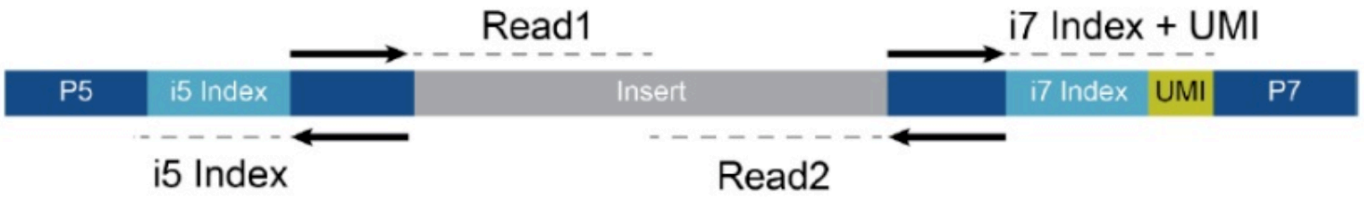
NGS data analysis



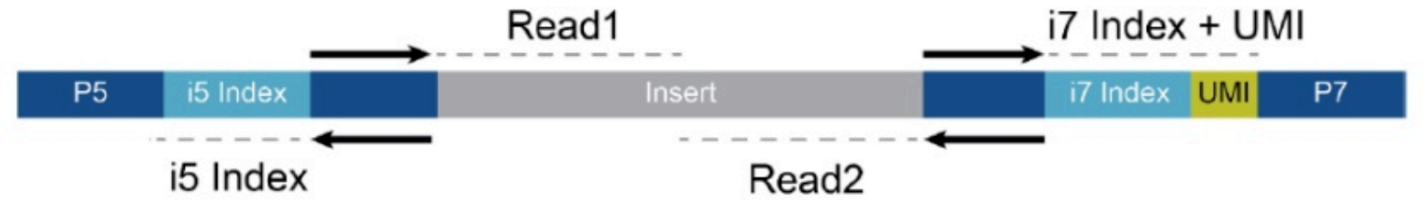
Data pre-processing

- Primer (adaptor) trimming
  - To cut adapter usually not necessary but good practice
  - Primer removal is necessary

- UMI extraction



UMI – unique molecular identifiers



- Each molecular fragment gets unique n-base sequence (n ~ 8-12)
- Usage:
  - Mark duplicates
  - Consensus sequence
    - sequencing (PCR) error removal

Raw data - QC

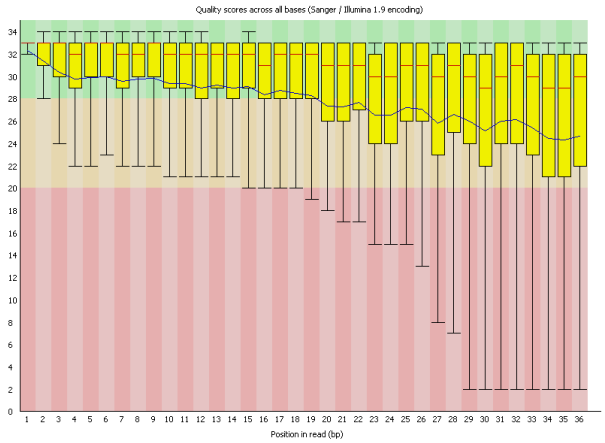
- **Fastq - q stands for quality – coded phred score**

```
CFFFFEFFGCEEGECFGGGGAFF87@E:++6C<++3:,8,33,,:,,,:,,:,,,
```

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

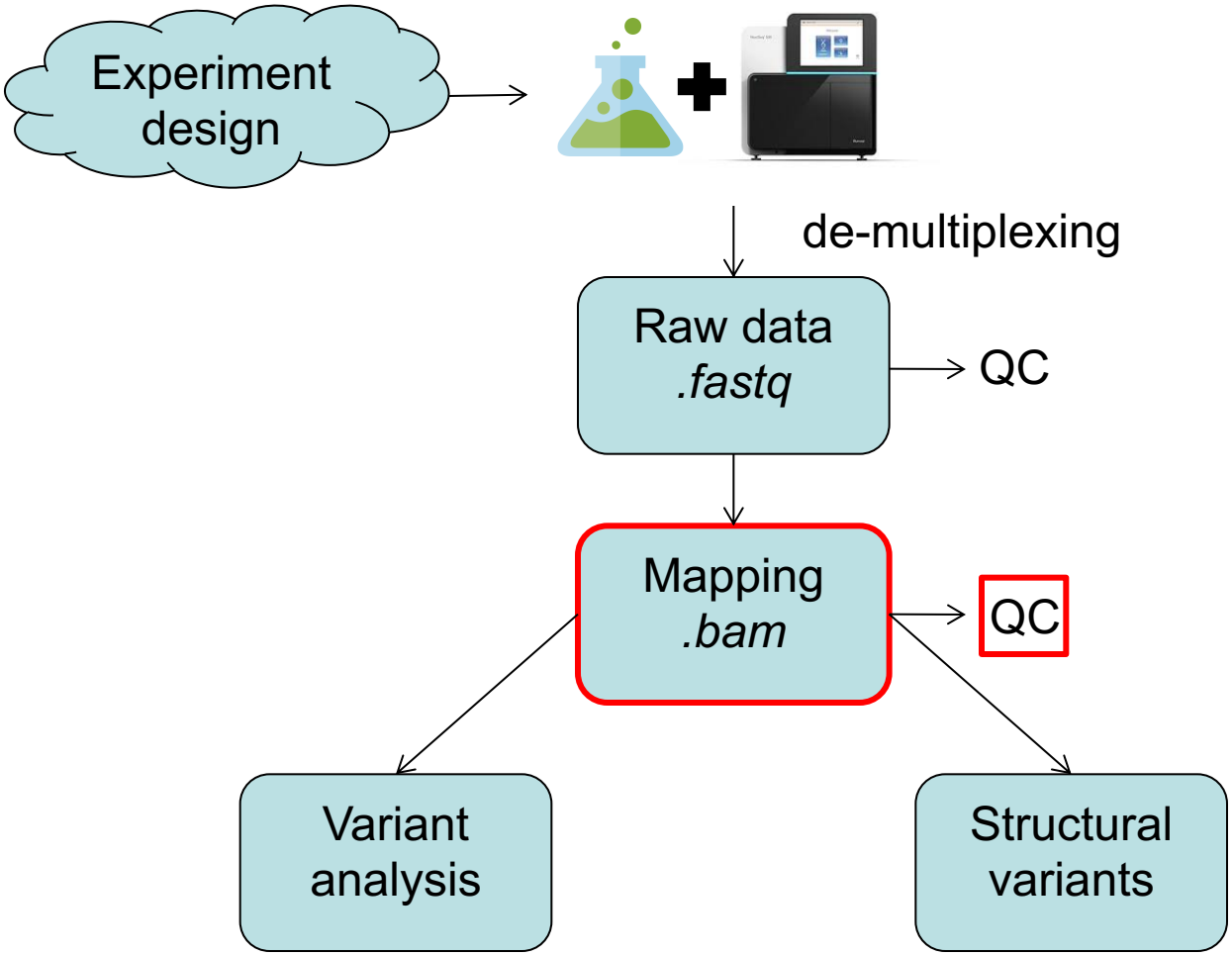
Quality	Error probability
5	31%
10	10%
20	1%
30	0.1%

- **Very good for early problem detection**
- **Reasonable for trimming and read filtering**
  - **RNA seq - above phred score 5**
- **Not good for individual variant analysis**



CEITEC

NGS data analysis



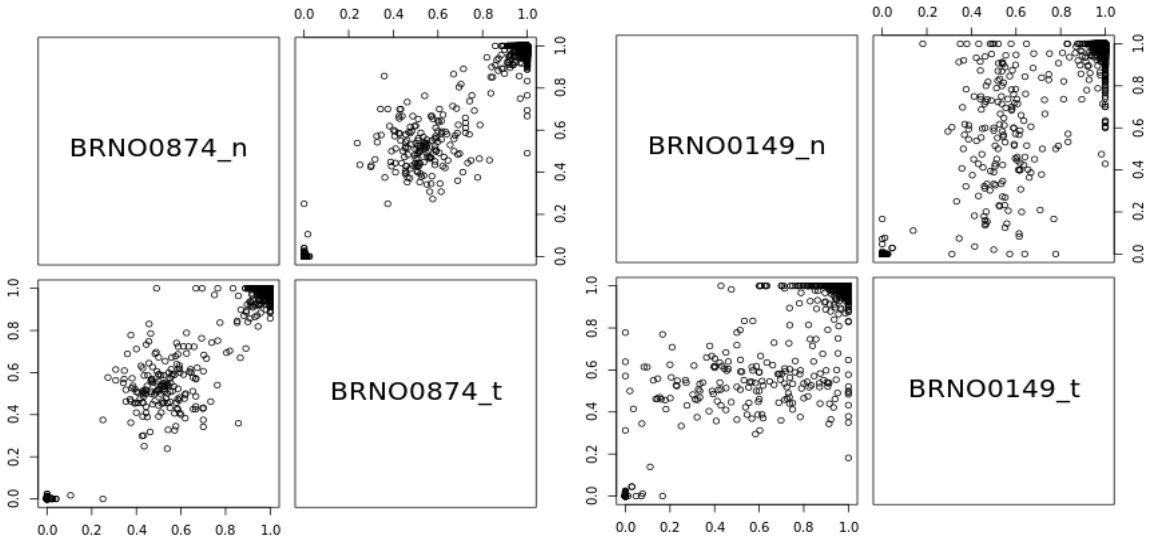
Alignment

- **Computationally most demanding**
- **More or less standardized**

- **Align to genome then select region of interest (ROI) <- .bed file**
  - **Don’t force alignment**
  - **Keep the information about wrongly aligned for QC**
  - **Exception targeted structural variant detection**

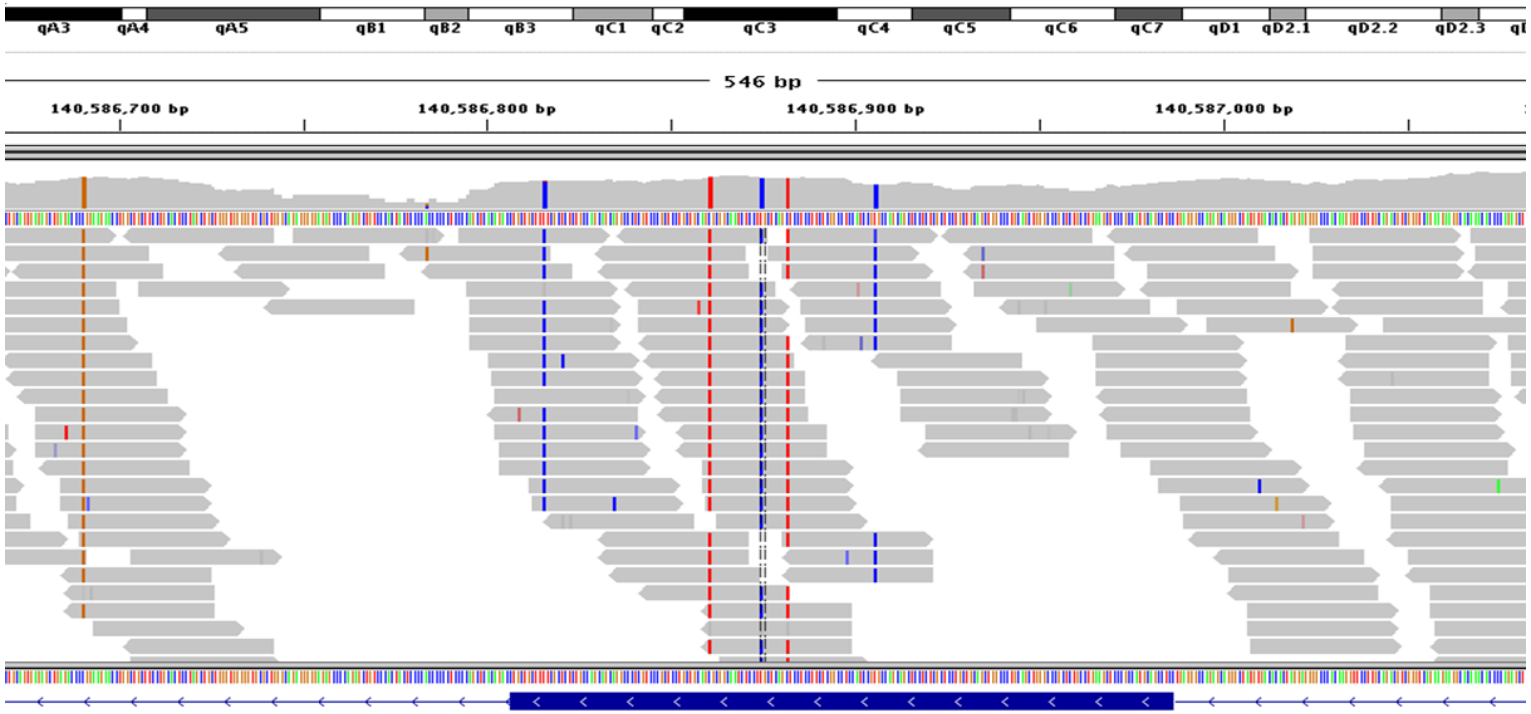
Alignment - QC

- **Mean coverage and variance**
- **Percentage of covered with at least**
  - **In WES we define good quality if at lest 90% of positions are covered at least 20x**
- **Insert size**
- **BAM cross-contamination**
- **Cross-sample snp allele frequency correlation**



CEITEC

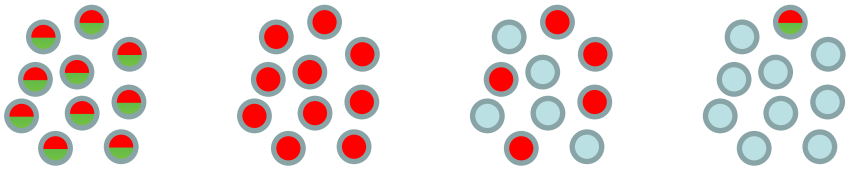
Variant Calling



CEITEC

Variant Calling

- Type of comparison
  - Germline
  - Somatic
    - Tumor - normal
    - Somatic variant calling without normal needs high coverage
  - Expected variant heterogeneity
  - Indirectly corelates to the necessary coverage



Variant Calling

- **Scope**

Scope	genes	~bp	~% of WG	~ Germ vars
WGS	~22000	3 200 mil	100%	700 000
WES	22000	30 mil	1%	60 000
PanCancer	1049	1.2 mil	0.04%	3000
CZECANCA	219	250 000	0.0083%	400
TP53	1	25772	0.000859%	30

Variant Calling - planning

- **Sample design**
  - **Germline**
  - **Somatic (Tumor - Normal \0**

- **Any relationship between samples for comparison improve specificity dramatically**
  - **Not sensitivity**
- **Somatic variant calling without normal needs high coverage**

- **RNA**
  - **Depends on gene expression levels**
  - **Variant might not be there! – gtex, previous runs QC**

Variant Calling

- **Specificity vs. Sensitivity**
- **Tools**
  - **varscan – no statistics = no assumptions**
  - **vardict**
  - **gatk haplotype caller**
  - **mutect – only snp**
  - **pindel – only indels**
  - **freebayes**
- **Callers combining – usual strategy**
- **Variant Annotation**
  - **Annovar – good database**
  - **snpEff**
  - **vep – variant effect predictor**

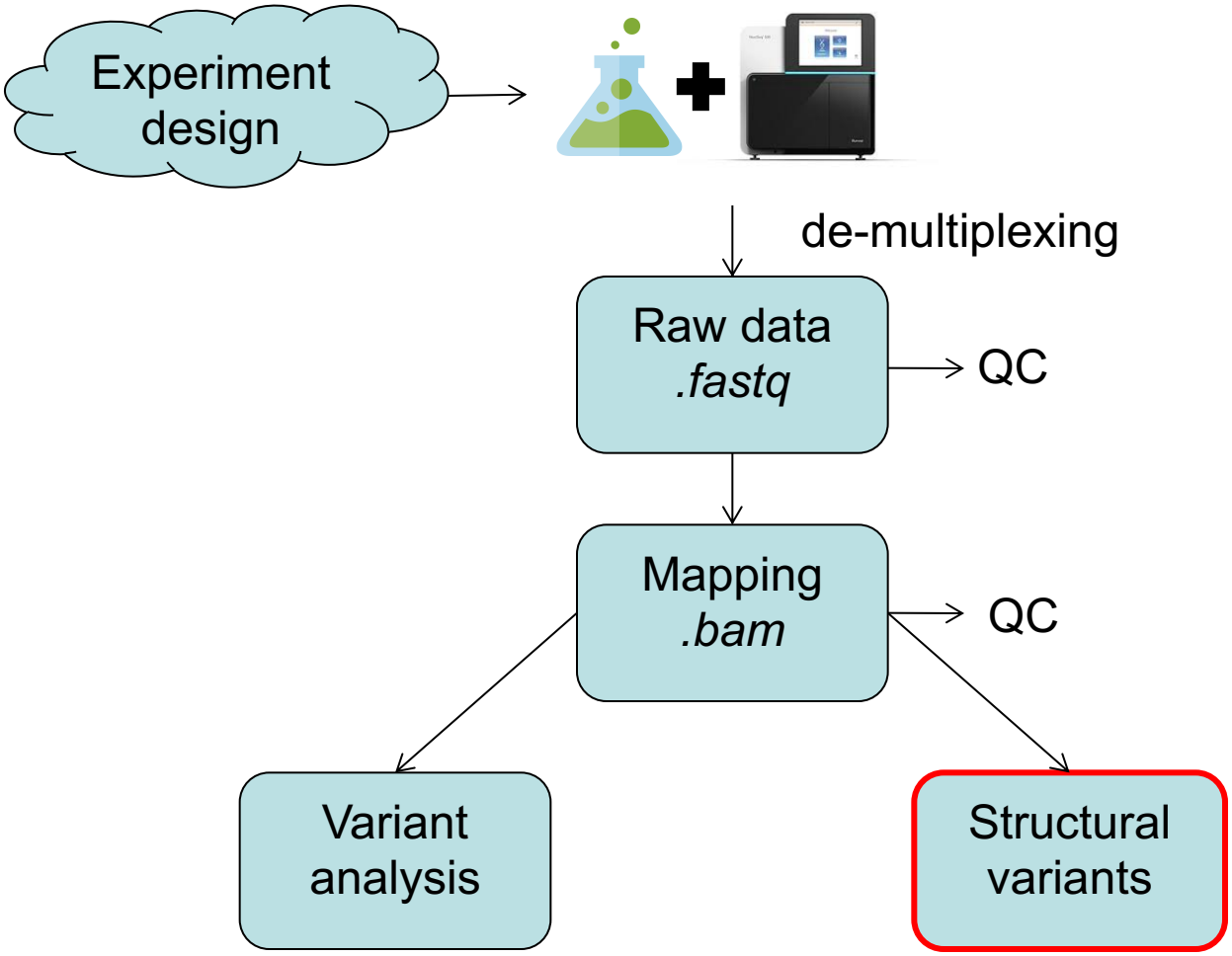
Variant Calling

- **Variant annotation can help variant calling significantly**
- **Variant occurrence in normal population**
  - **1000 genome project – above 5%**
- **Variant consequences cut off**

* SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codo	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequenc	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequenc	SO:0001822	Inframe deletion	MODERATE
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW

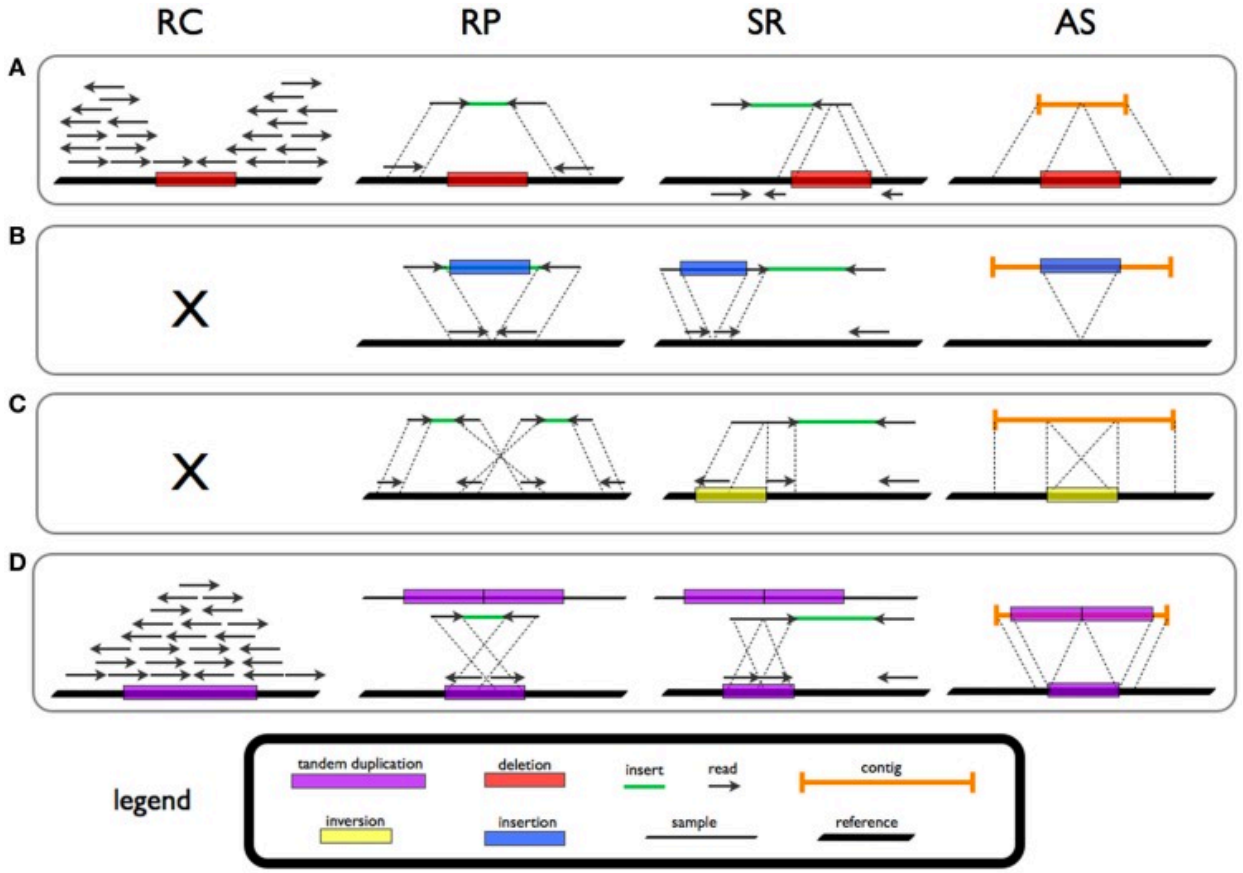
- **Database can help significantly – Sophia Genetics**

NGS data analysis



Structural variants

- **discordant read(-pairs) mapping**
- **copy number variants (CNV)**



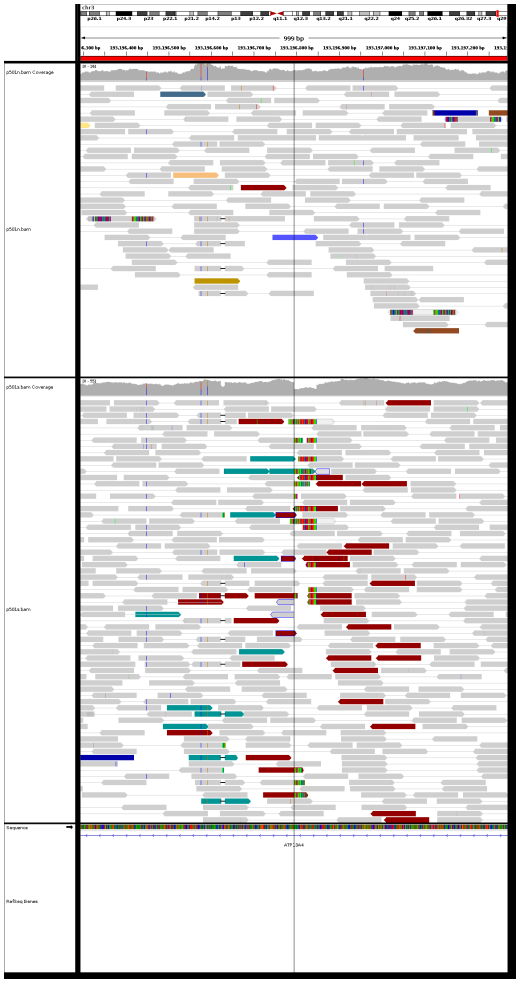
Structural variants

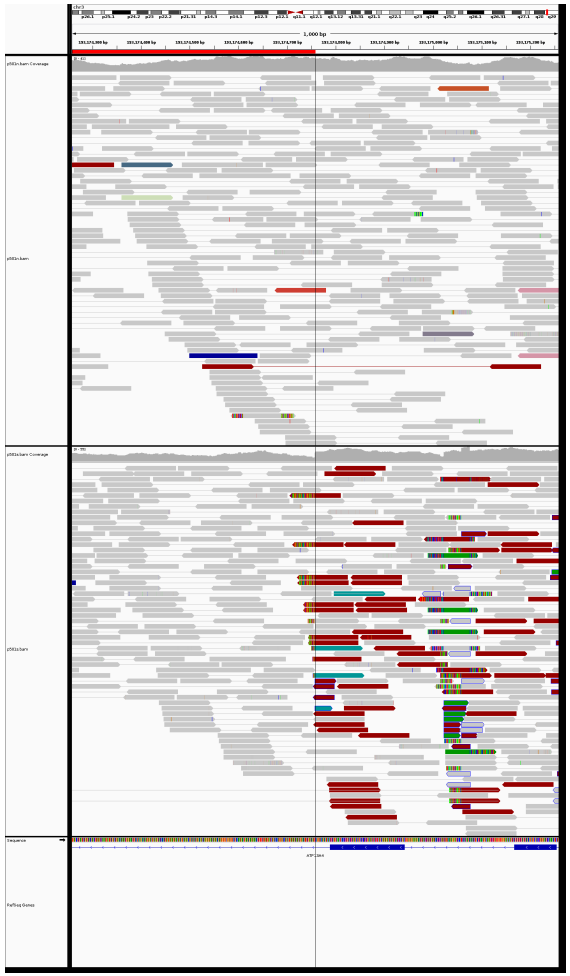
- CNV
- long variants in WGS – ControlFreec
- Smaller variants for WES / target panel
  - Somatic – tumor,normal
  - Germline - lot of references
    - XHMM

- Read-pairs very noisy expect a lot of FP
- BreakPoint
  - Target panel with short reads
- Delly
  - everything else

Structural variants

- **Manual check with IGV**





Thank you for your attention

CEITEC

Central European Institute of Technology
Masaryk University
Kamenice 753/5
625 00 Brno, Czech Republic

www.ceitec.muni.cz | info@ceitec.muni.cz

UNIVERSITAS MASARYKIANA BRUNENSIS

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and Development for Innovation