



CEITEC



Central European Institute of Technology
BRNO | CZECH REPUBLIC

Moderní metody pro analýzu genomu: Bioinformatika I

Vojtěch Bystrý

29. October 2018



EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE



OP Research and
Development for Innovation

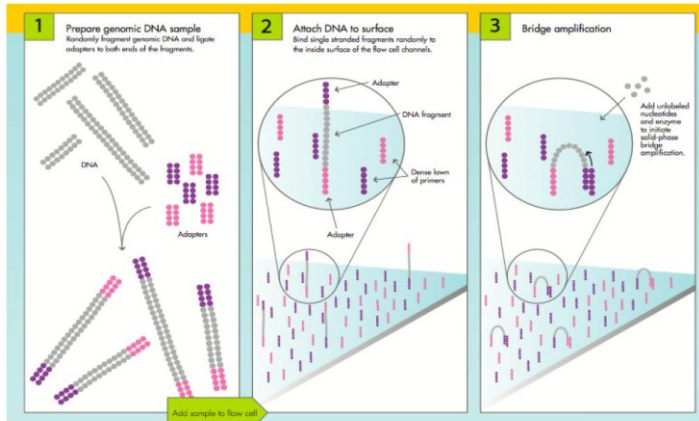


Goals of the presentation

- **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



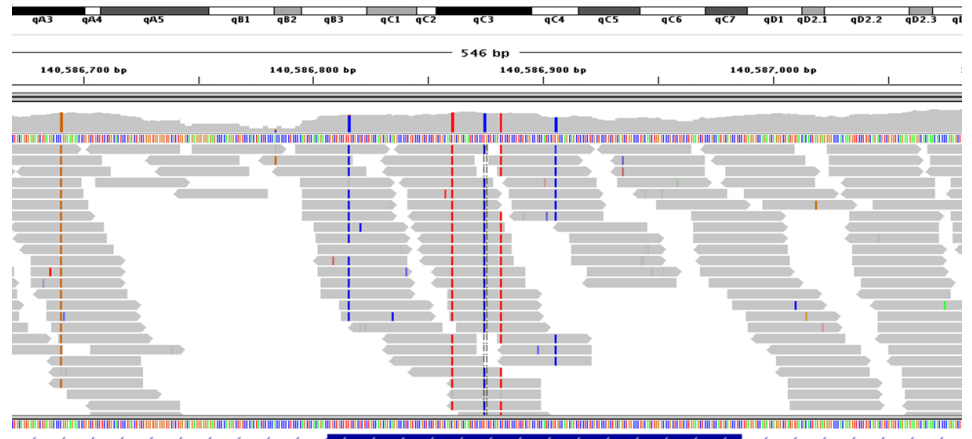
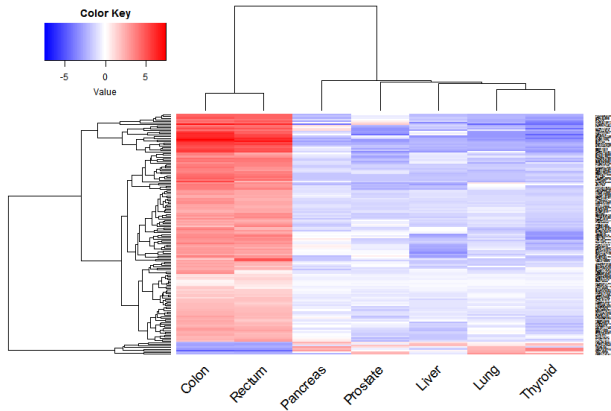
ECE/BioE 416
Lecture 24

10

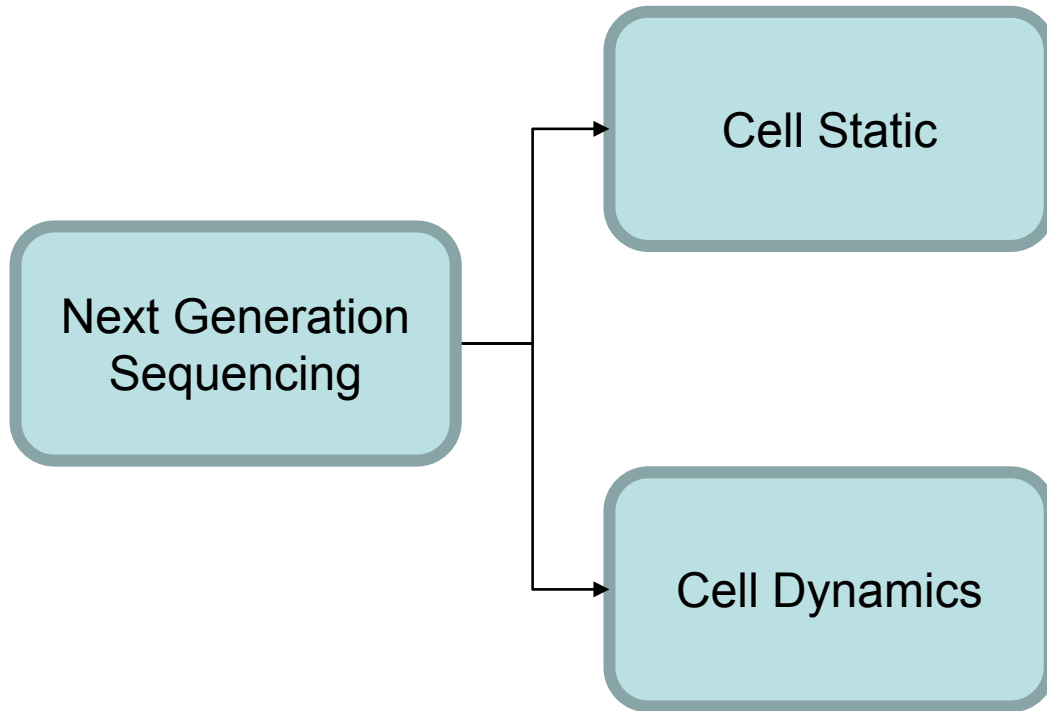
```

410188 >gl139879376[ref|NM_001191668.3] Rattus norvegicus
410189 (Dosey2). mRNA
410190 CCTTCCCECCGAGATGTAAAGTGGACCCAGAGGAGCGCTGTAA
410191 TATGGAGAGGCGCTTAAAGAGTATTCGAGTCTCCACGGAAATGGCCGAT
410192 CTCTGCAGAGC#ACCTGAAGTATTCCTGCTACAGGGGGCTCATCAT
410193 GCGCTGCCTAGTGTGTCACCTCANNAGCCCTAGAAACCTACGAGATC
410194 CAGGACCTTTGTTCTTACAGCTGTGGCTTGTTCCTCCCTCCGCGCTAT
410195 GGCTGTGTSAGAACTACTTCCCTCCGCTGCAAGAGCTCCTCTGGGAG
410196 GTTCTATGCTGCTGCTGGGGAGTGTCTCTGACAAAGCCCTTCCATCCG
410197 TCAGCAGGACCTCACCAGGCAAGGAGGAGAGTACATCTGGGTACGTA
410198 TCACACTGCTGATCCCAATGTTCTCTGTGCAGGAAATAACTGGAAATC
410199 GAGCCCGGATGAGAGACCACTCCCTACTGAGAGAGGACATTGTTCAT
410200 GAGGGAGATGCTCTCAACAGAGAGCTCAGCCCTGCTAGGTTCTC
410201 TCTGAAATTCAGACTTCTACGAGAGGAGGCTCATACTCTTGTATTC
410202 GGCTGAGTCTCCACCAAGAGTCTCTTANNAGCCGACTTAGAGAGAGGCT
410203 TAGTACTGCTGTCACCAAGCCGAAATGAGACTCAAGTGTGAGATC
410204 TCTTACTGCTGATGCTCTGGGCGCTGAGCTTAGGCTTAGTACTACTC
410205 AGAAGAGAGGACTCCTGAGATGTGAGAGCTGCTCACTTCTGCTAGT
410206 ACATGGAGAGGCTCTGAGAGGAGTCTGAGAGAGCTGAGAGAGAGAG
410207 ACCGCTGAGAGCTGACACCTCCCTGAGAGGCTCTGAGAGAGAGAG
410208 GTACCTGAGAGAGAGTCTGGGCTGCTGAGAGAGCTCTGAGAGAGAG
    
```

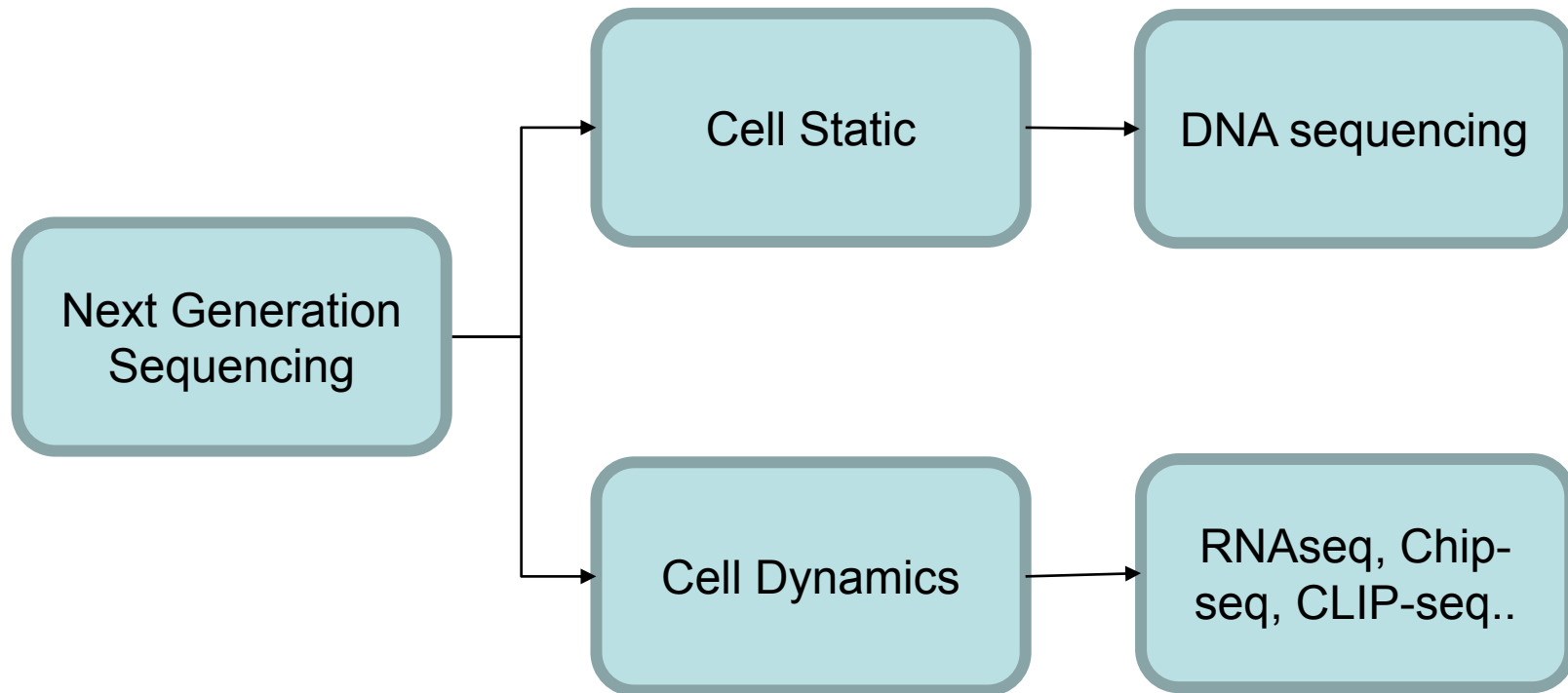
Your raw sequence data



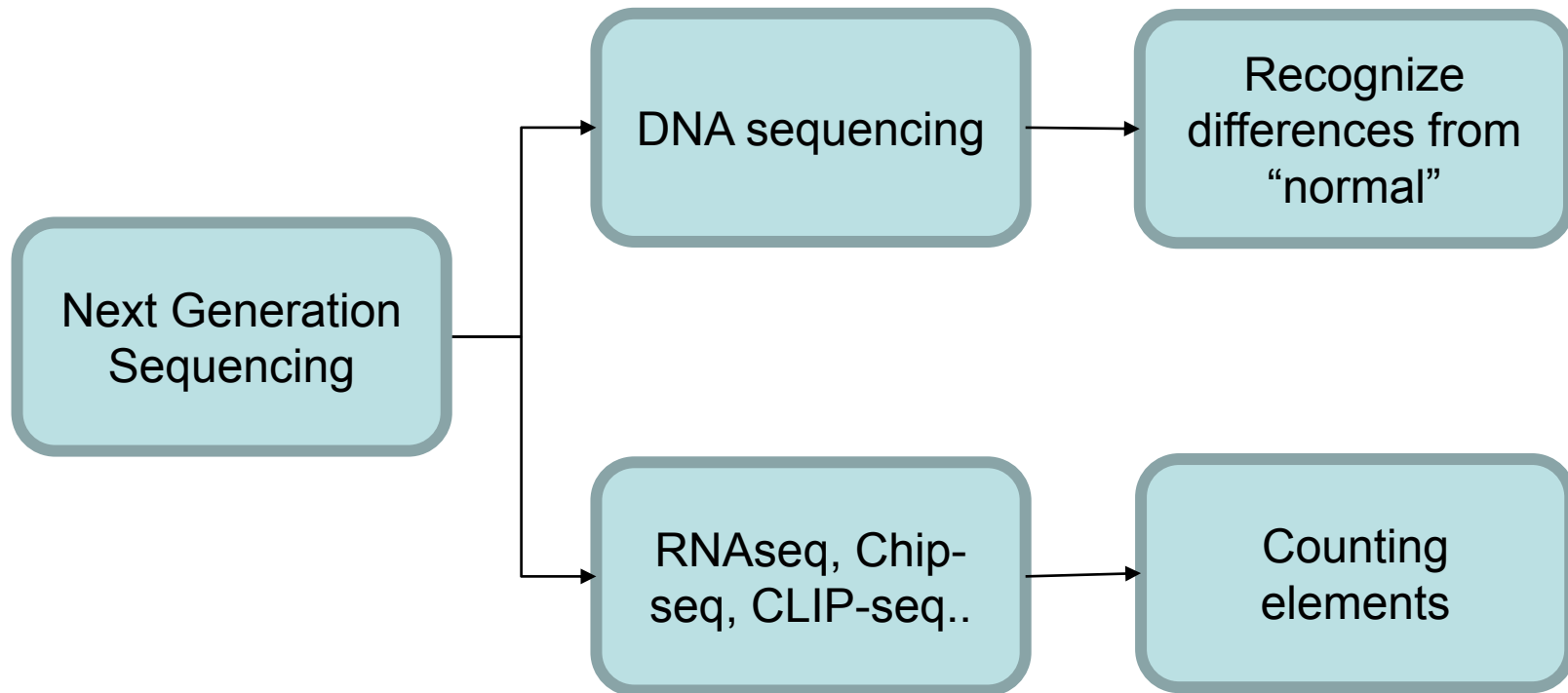
NGS experiments



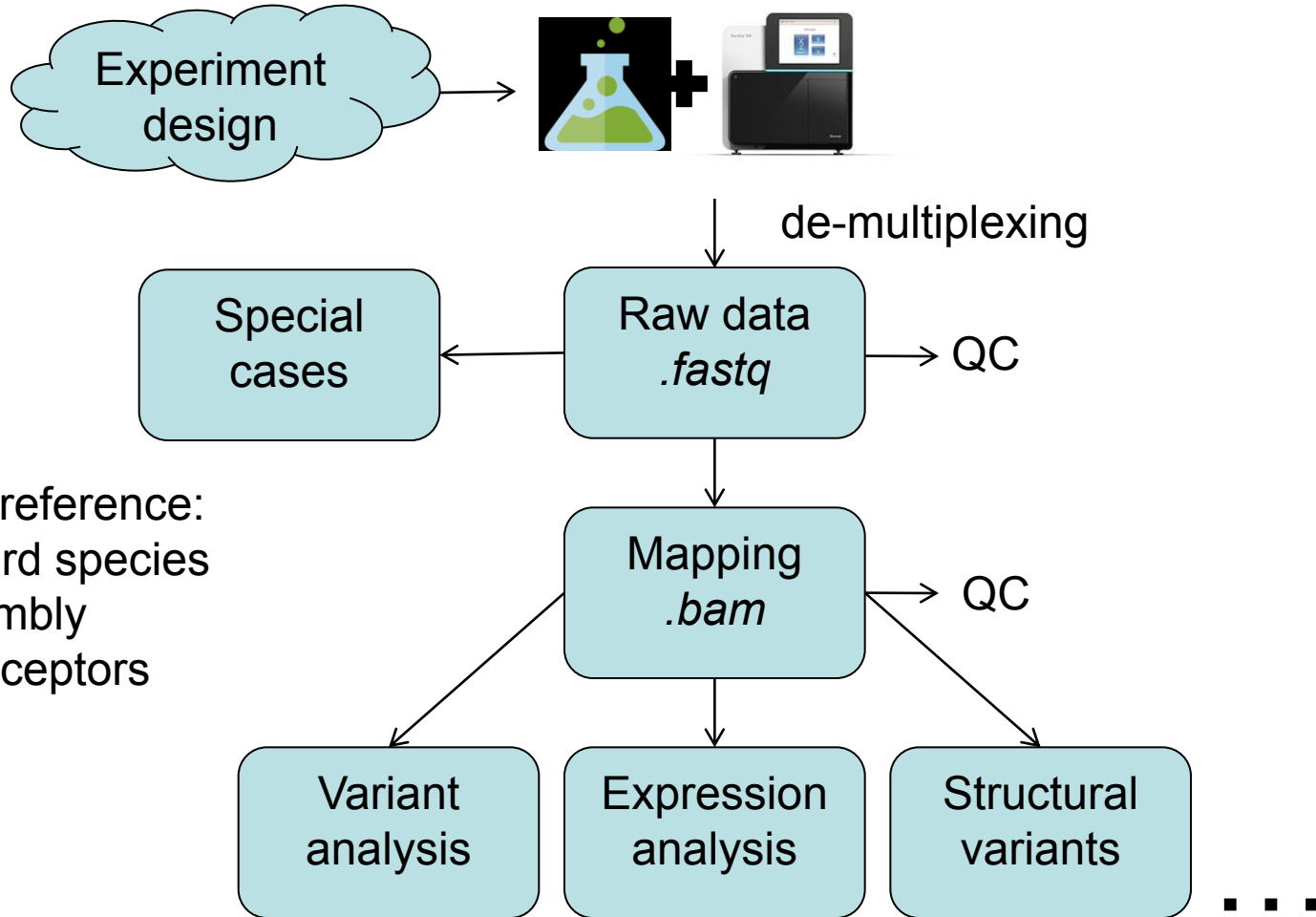
NGS experiments



NGS experiments

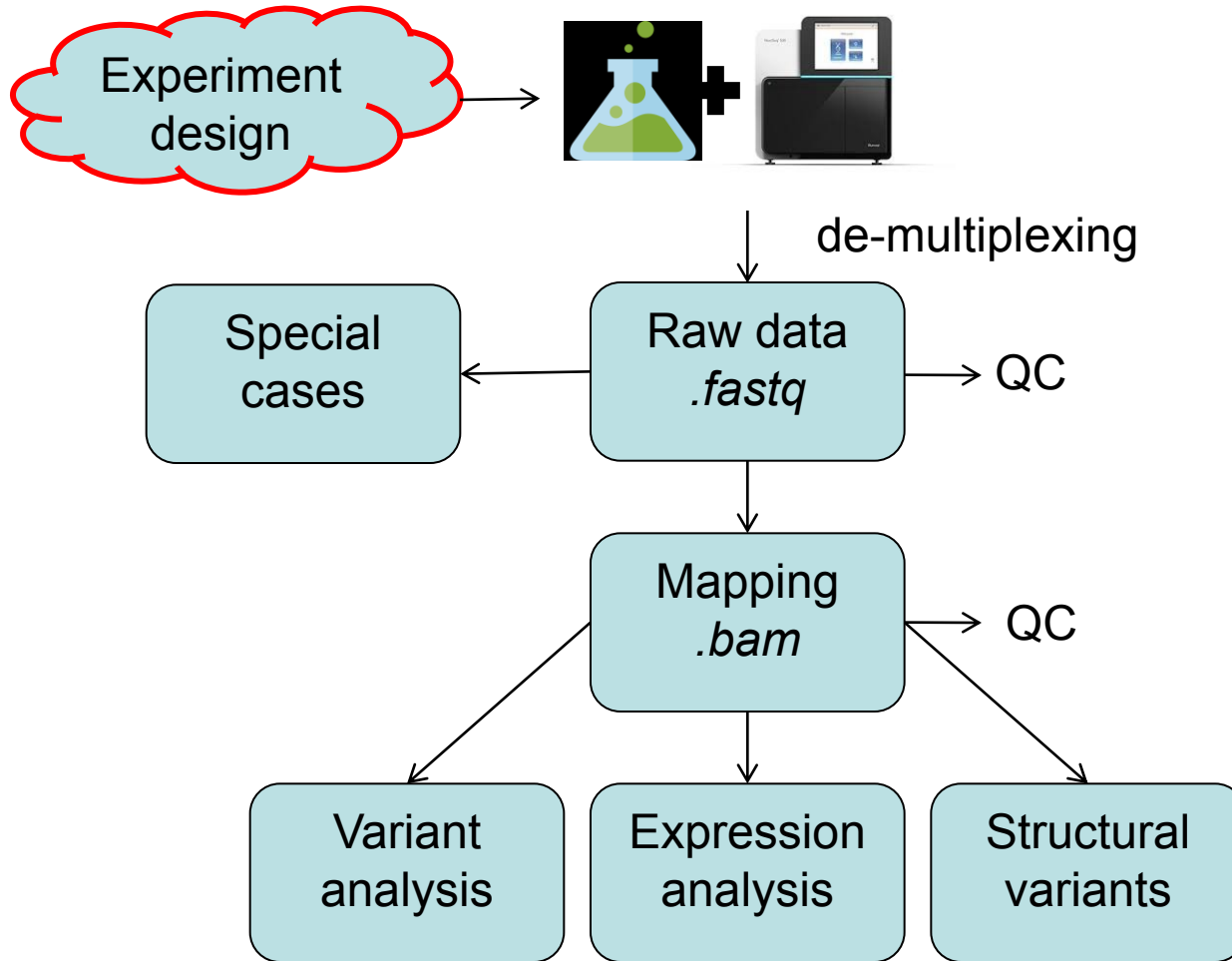


NGS data analysis workflow



Not known reference:
Non-standard species
Assembly
IG/TR receptors

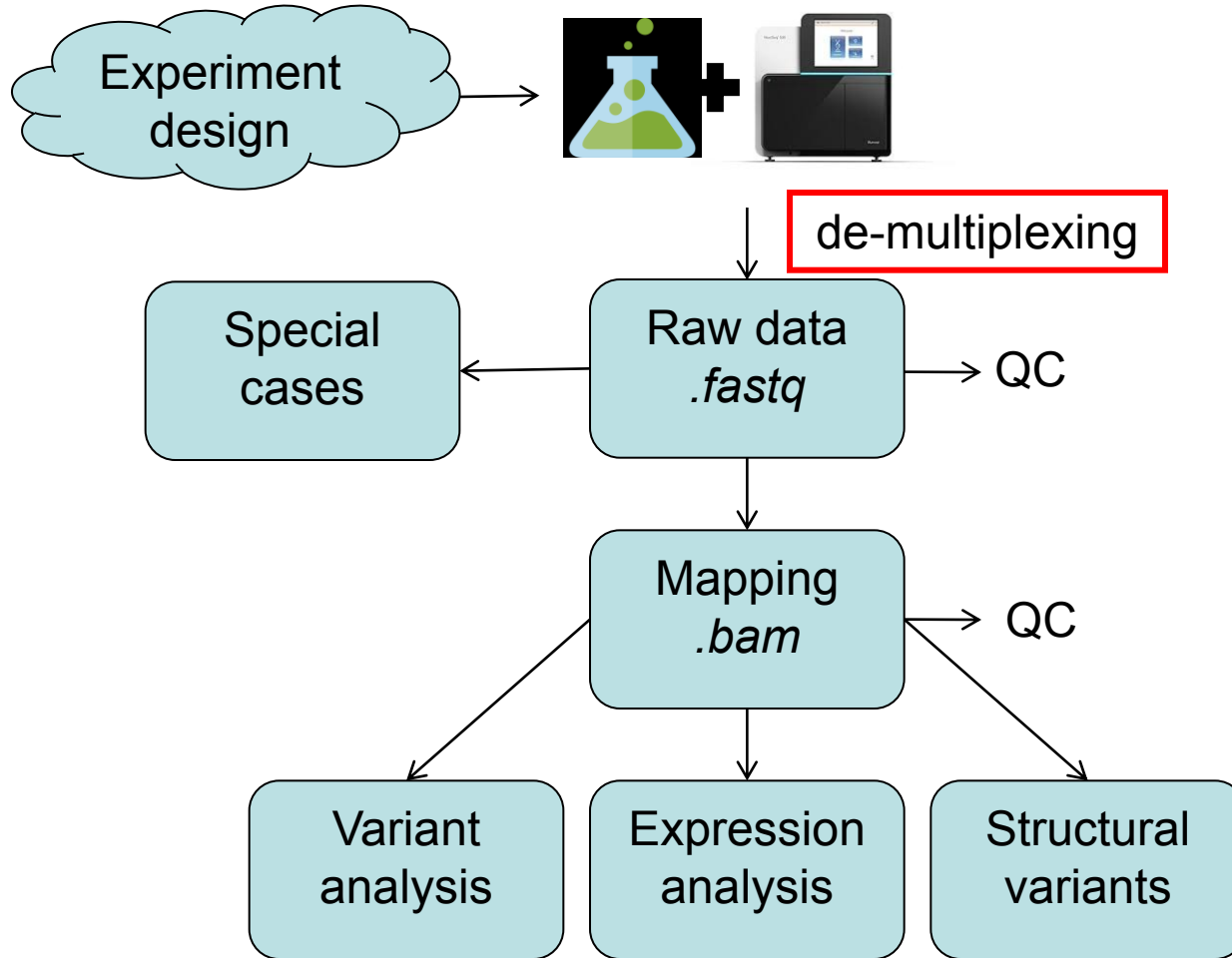
NGS data analysis workflow



Experimental design

- **Have a hypotheses!**
 - **Consult with sequencing expert and **bioinformatician****
1. **If experiment you have in mind can be done in a way you are planning to.**
 2. **If the results you want can be obtained from the planned sequencing. (desired outcome)**
 3. **If the bioinformatician knows how to perform specific types of analyses and how long it will probably take.**

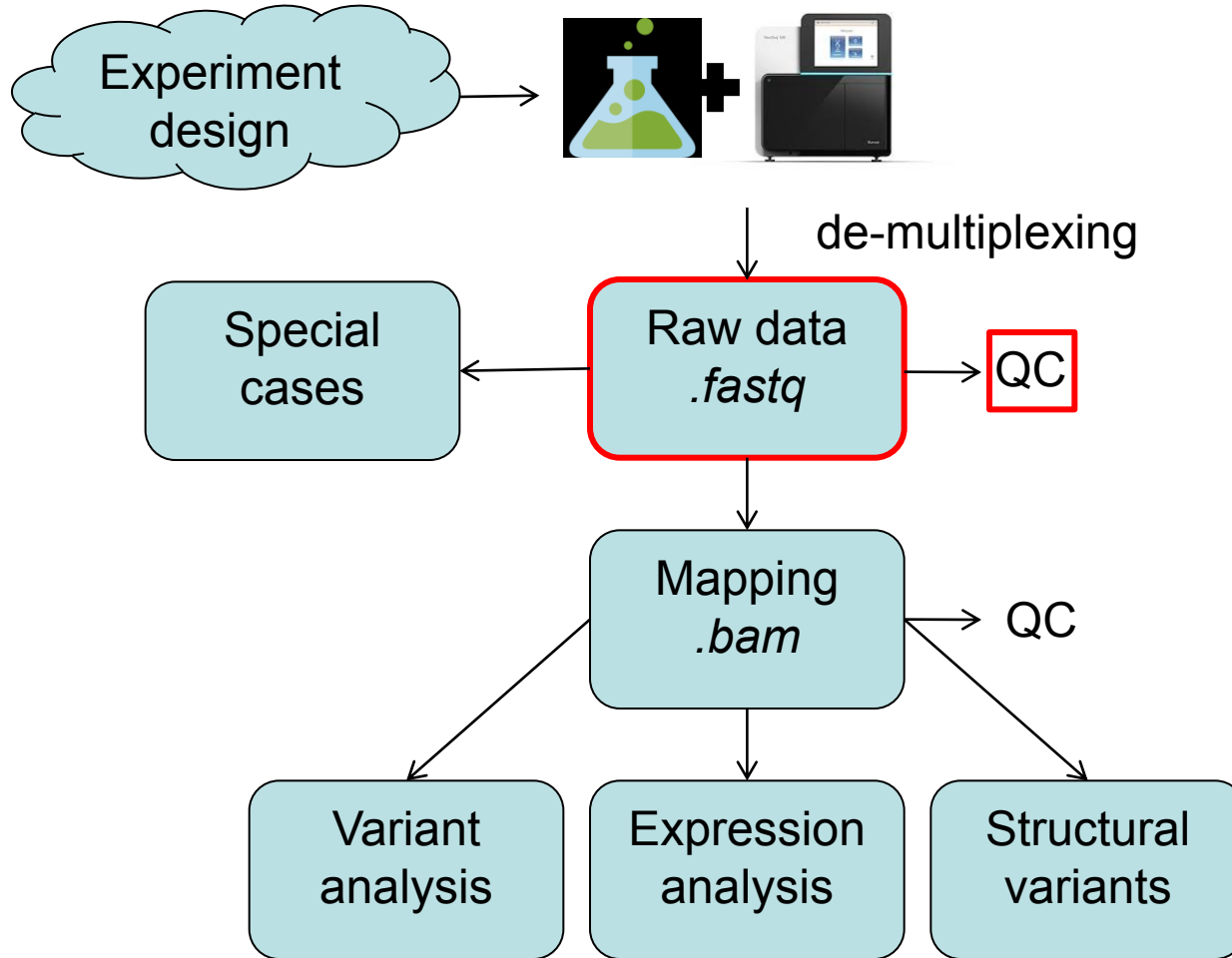
NGS data analysis



De-multiplexing

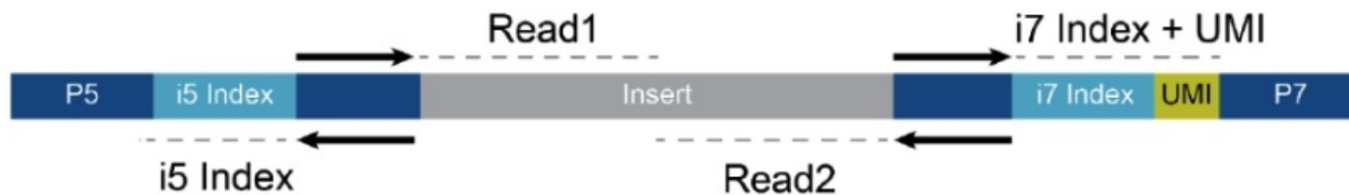
- **Not perfect**
 - In silico contamination – problem for MRD detection
- **Sample naming and organisation**
- **Naming**
 - Unique names
 - _ vs – vs .
 - Special characters: \$&|@+- ...
 - Really tricky: - vs –
- **Organization**
 - Should not be your worries
 - For any longer ‘operation’ comprehensive database is necessary
 - Currently working on it ourselves 😊
 - Please fill the forms carefully and as much as possible

NGS data analysis

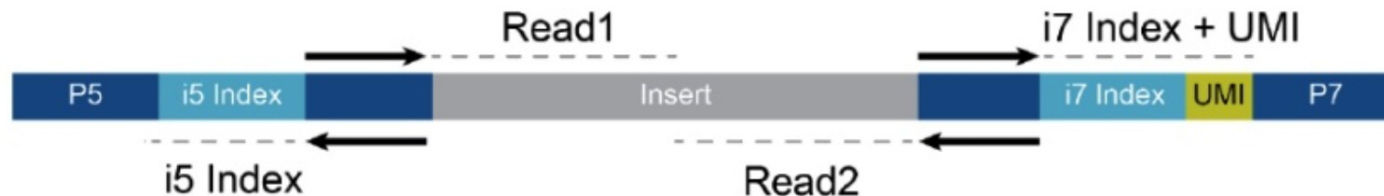


Data pre-processing

- 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 ($n \sim 8-12$)
- Usage:
 - Mark duplicates
 - Consensus sequence
 - sequencing (PCR) error removal

Raw data - QC

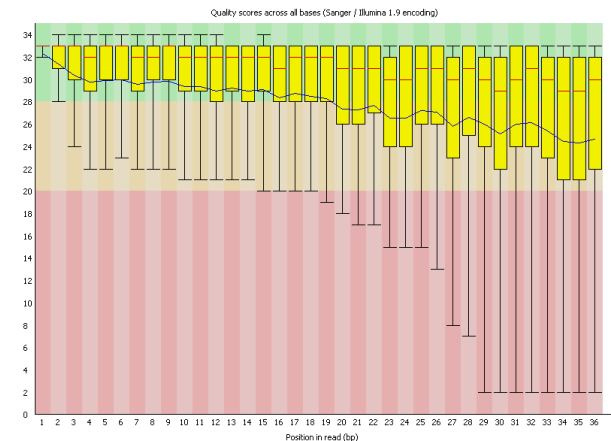
- **Fastq - q stands for quality – coded phred score**

CFFFFEFFF GCEEGECFGGGGAFF87@E:++6C<++3: , 8 , 33 , , : , , , : , , : , , ,

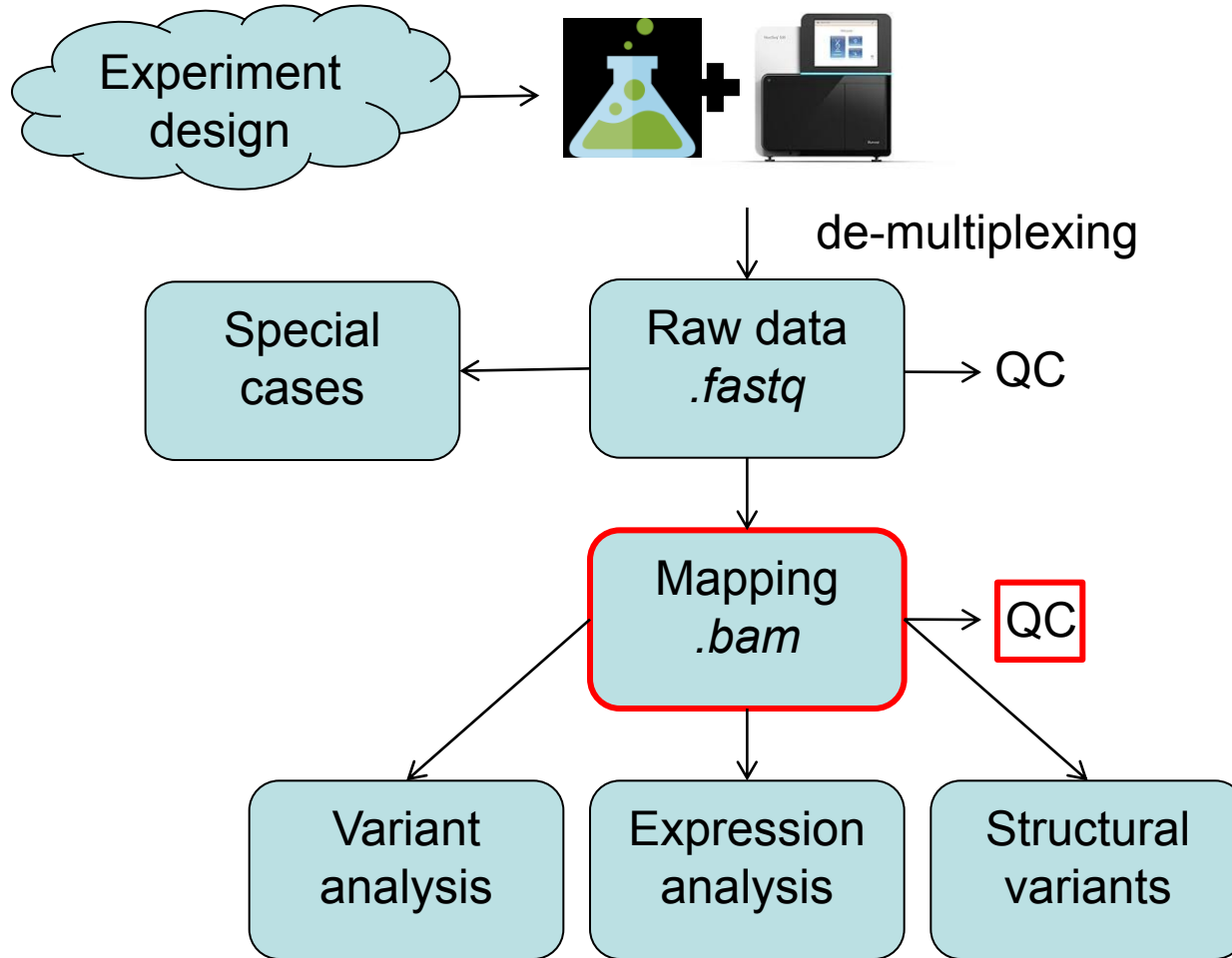
$$Q = -10 \log_{10} P$$

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**



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 SV detection
- **Our standard procedure:**
 - BWA - DNA
 - STAR - RNA
 - Chimera – sRNA

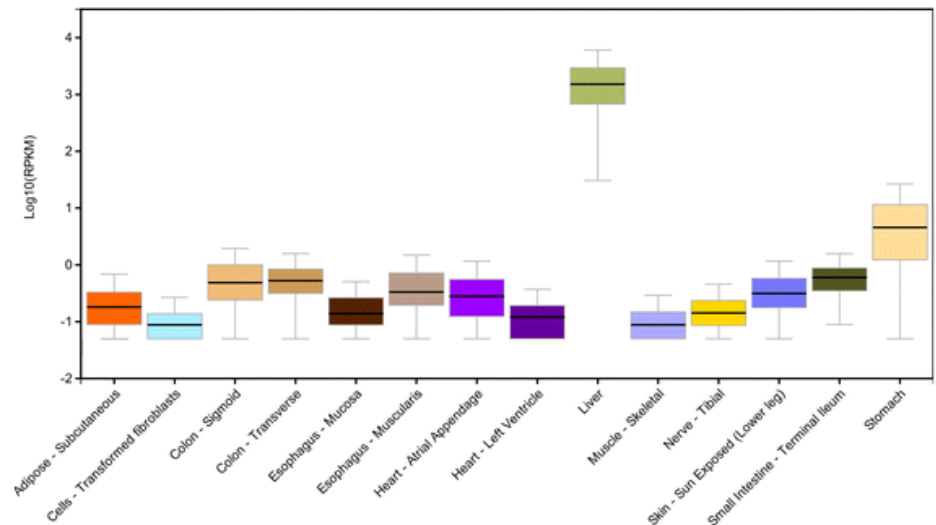
Alignment - QC

- **DNA**

- **Mean coverage and variance**
- **Percentage of covered with at least**
 - In WES we define good quality if at least 90% of positions are covered at least 20x
- **Per base coverage – in smaller experiments**

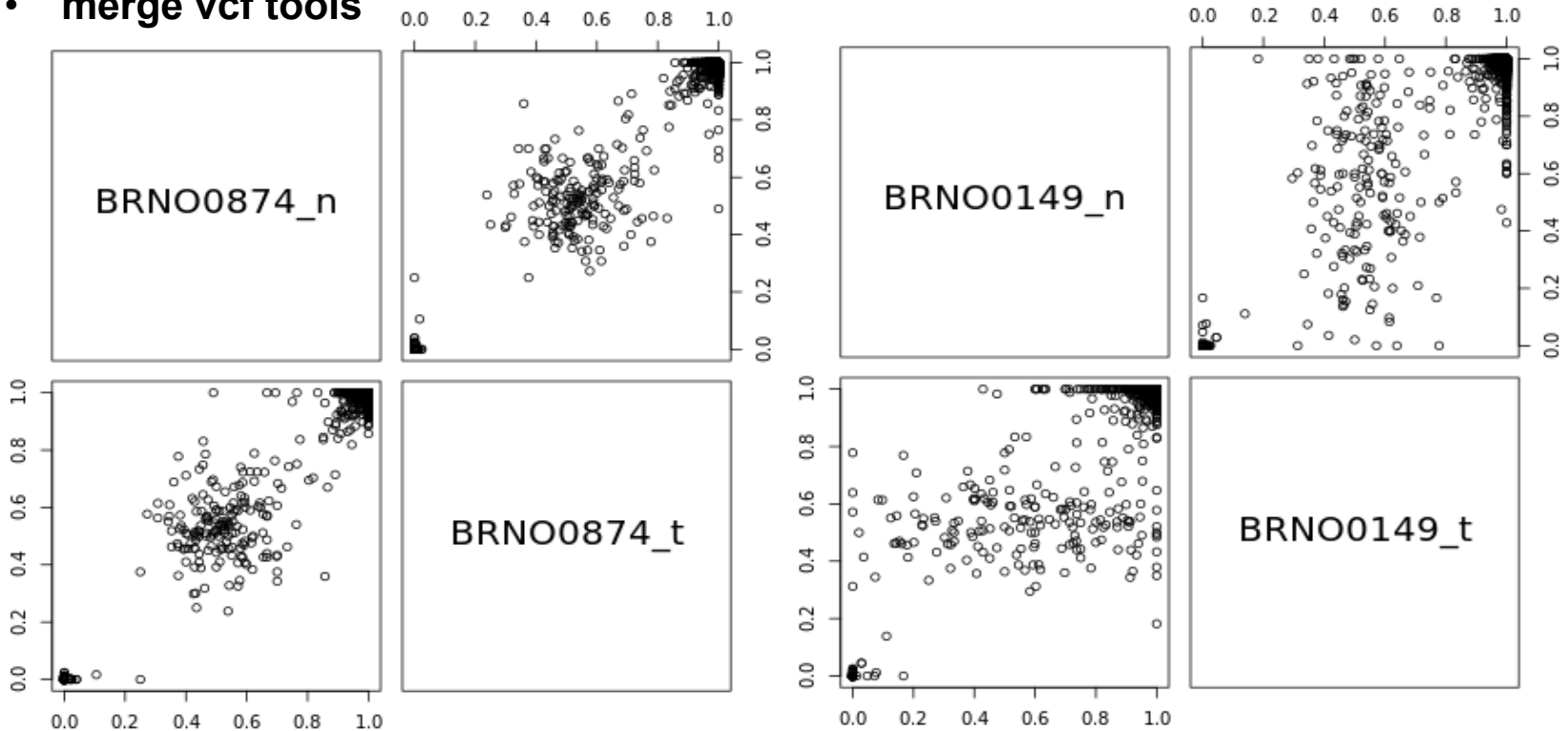
- **RNA**

- **Per gene coverage**
- **Variability of per gene mapping**
- **Gene counts distribution**
- **rRNA content estimate**
- **Tissue expression check - gtex**

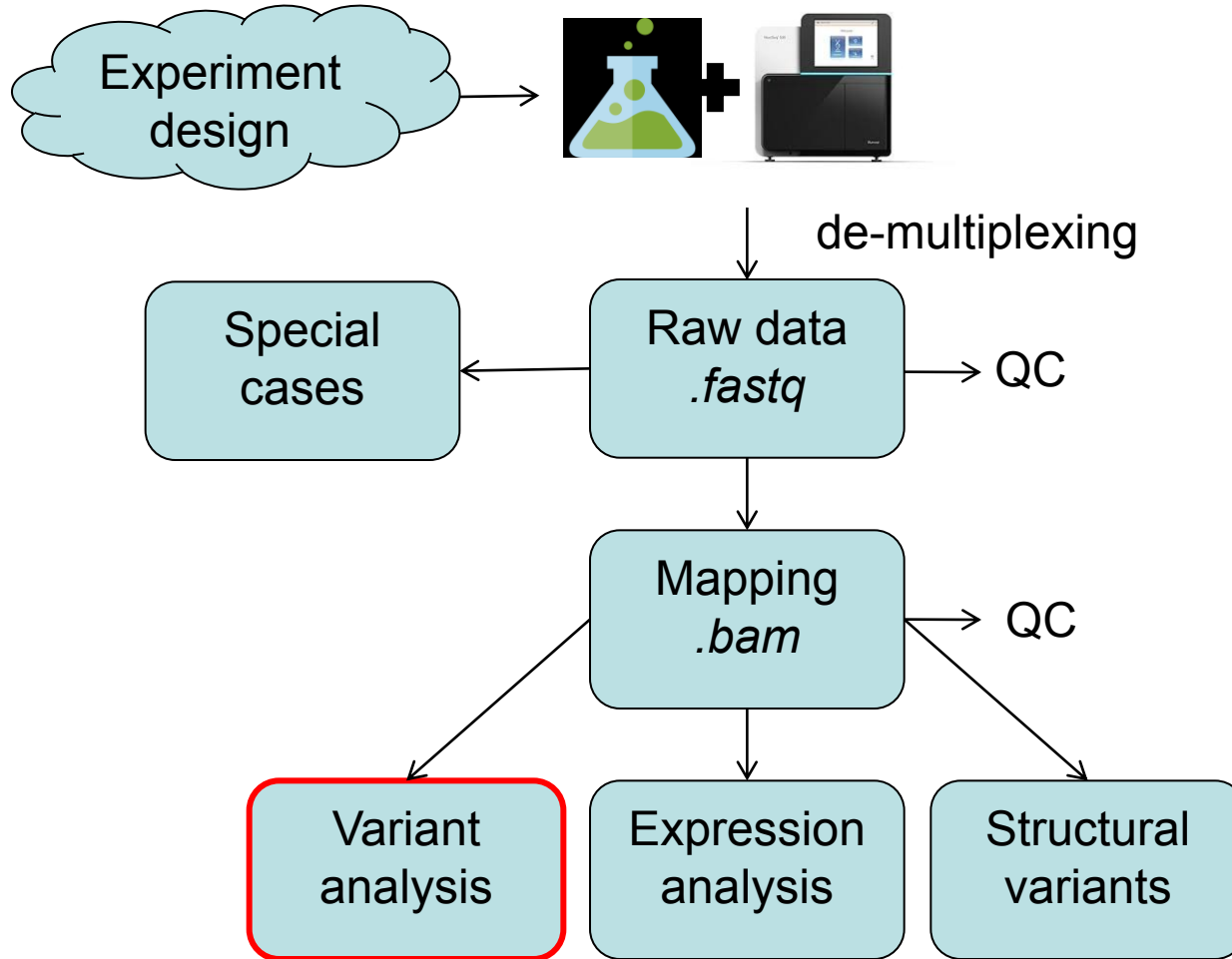


Alignment - QC

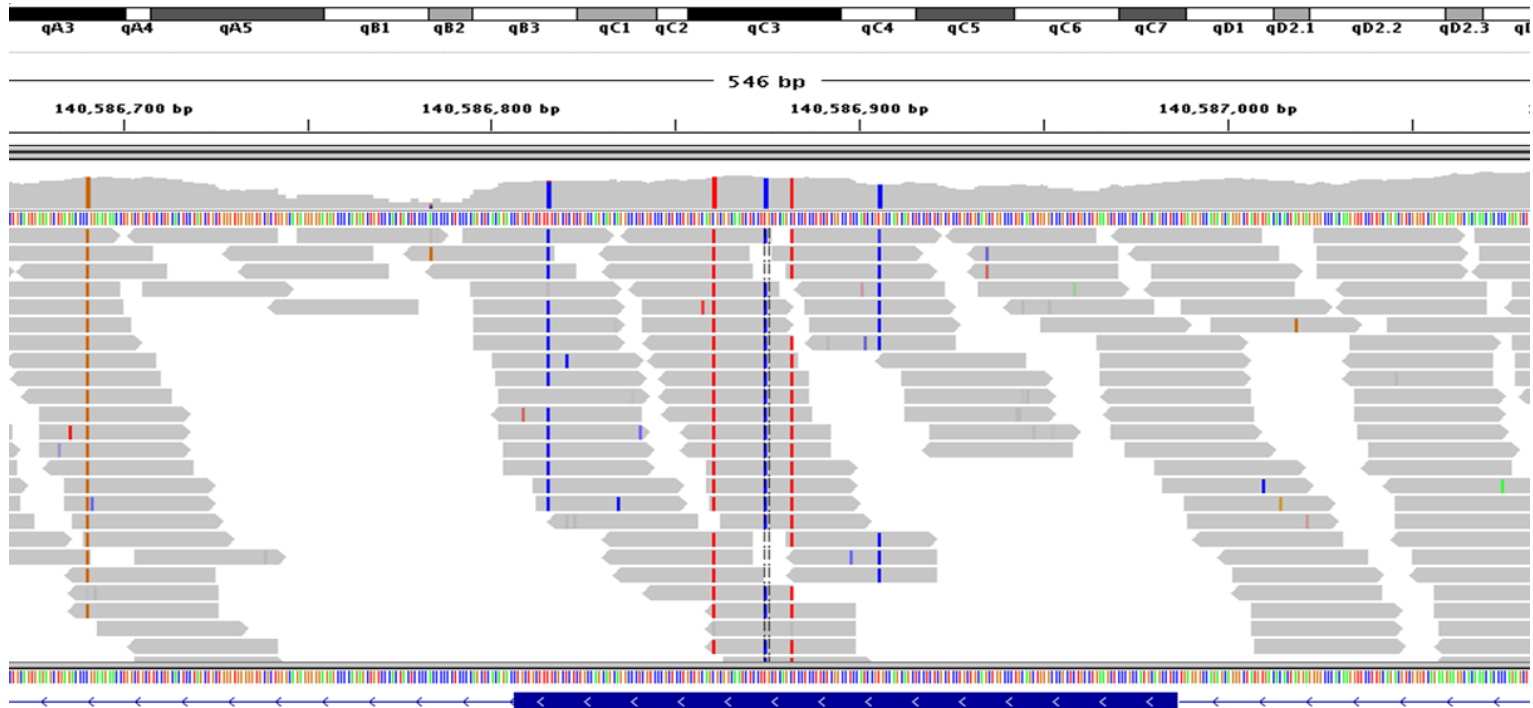
- **BAM cross-contamination**
 - verifyBamID
 - FREEMIX - below 0.03 = OK
- **Cross-sample snp allele frequency correlation**
 - **merge vcf tools**



NGS data analysis

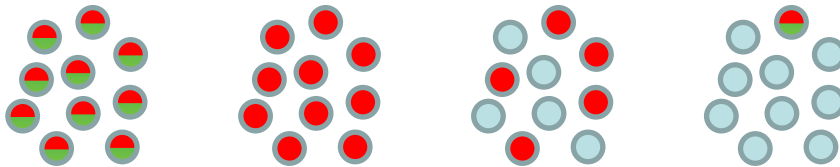


Variant Calling



Variant Calling

- Type of comparison
 - ermline – to reference genome
 - Somatic – to other sample(s)
- Expected variant heterogeneity
 - Indirectly correlates to the necessary coverage



Variant Calling - planning

- **Scope**

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

- **With fixed cost of bp-read it seems the price is linear**
- **The “price” of the analysis must be considered**
 - **Power of the results (sensitivity, specificity)**

Variant Calling - planning

- **Example**

Lets have an analysis with per base false positive error rate 0.0001.

Results in:

2 false variants in TP53 gene

3000 false variants in WES!

- **WES on a single healthy person with a question:
Are there any variants?**
- **Answer is YES**

Variant Calling - planning

- **Sample design**
 - Germline
 - Somatic
 - Tumor - Normal
 - Family
- **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**

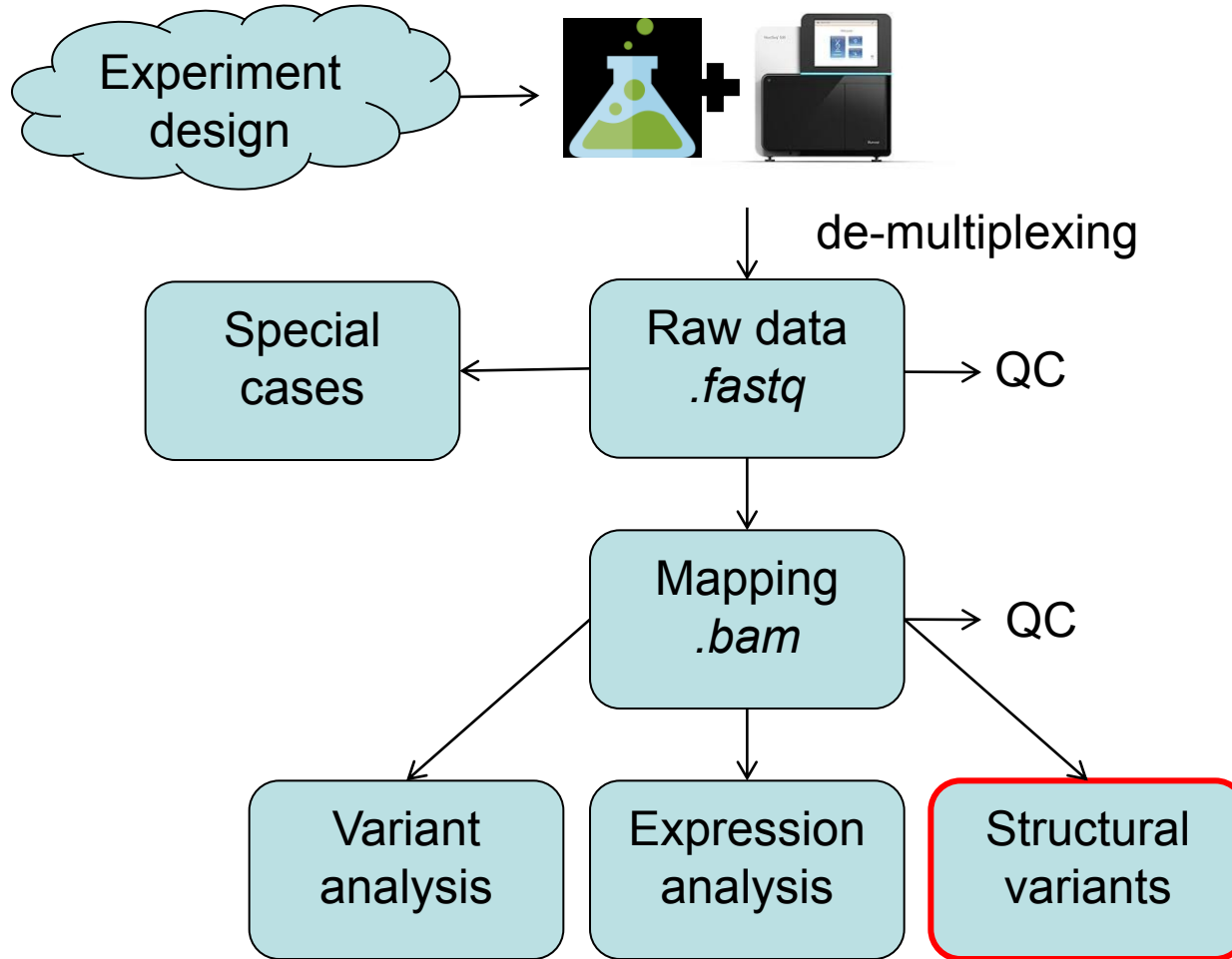
Variant Calling

- **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:0001567	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 codon	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

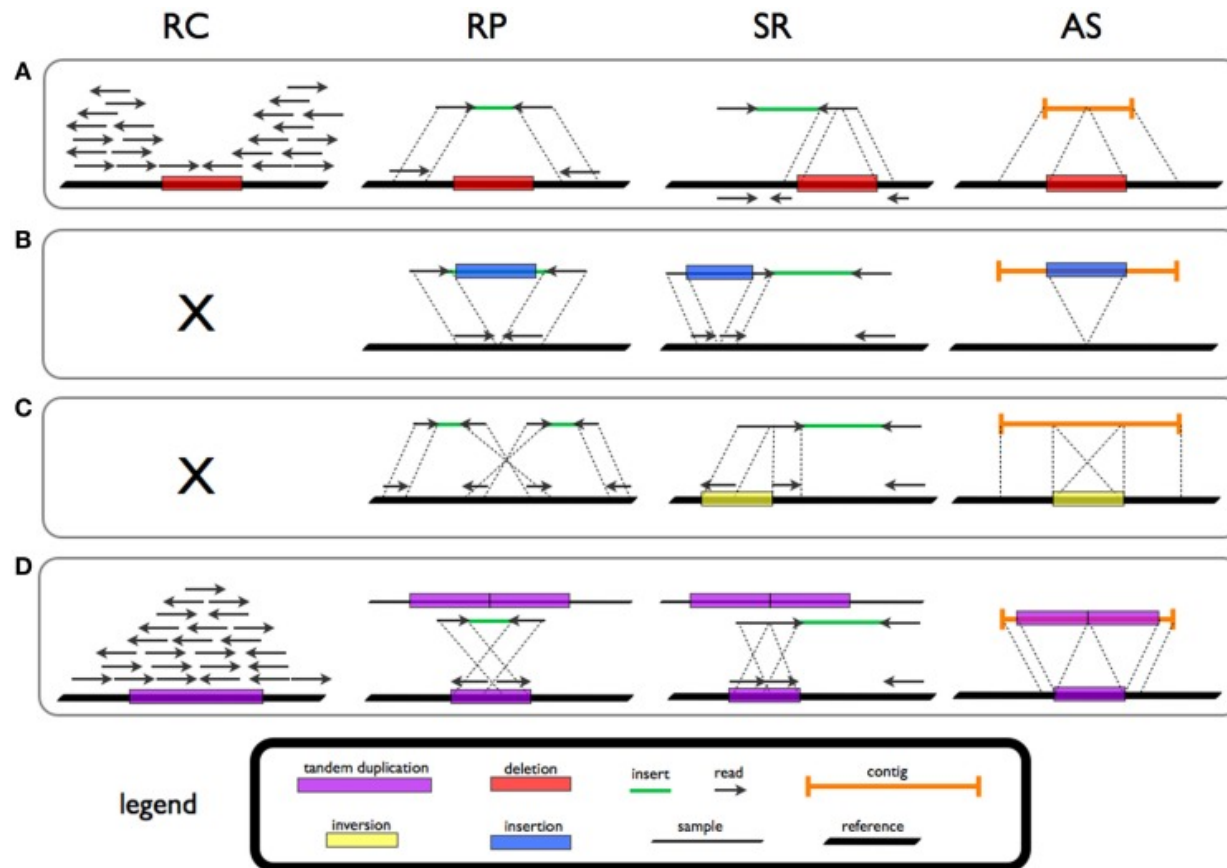
- **Database can help significantly – Sophia Genetics**

NGS data analysis



Structural variants

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

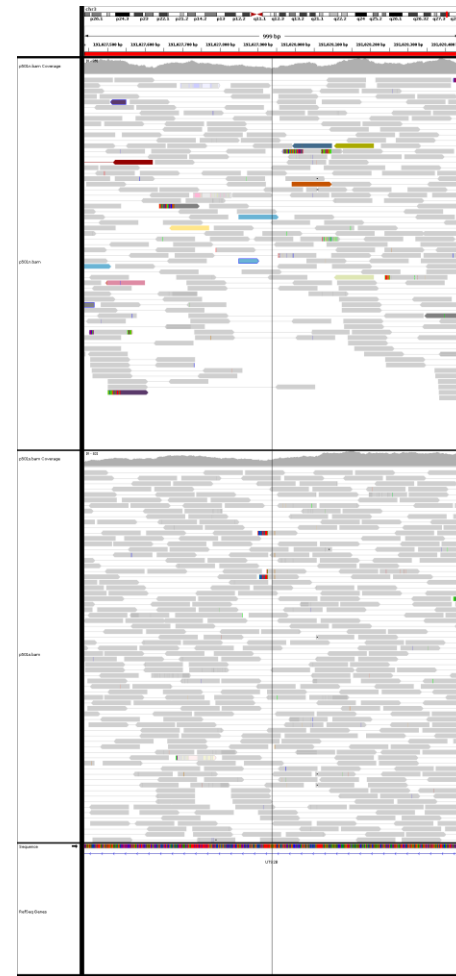
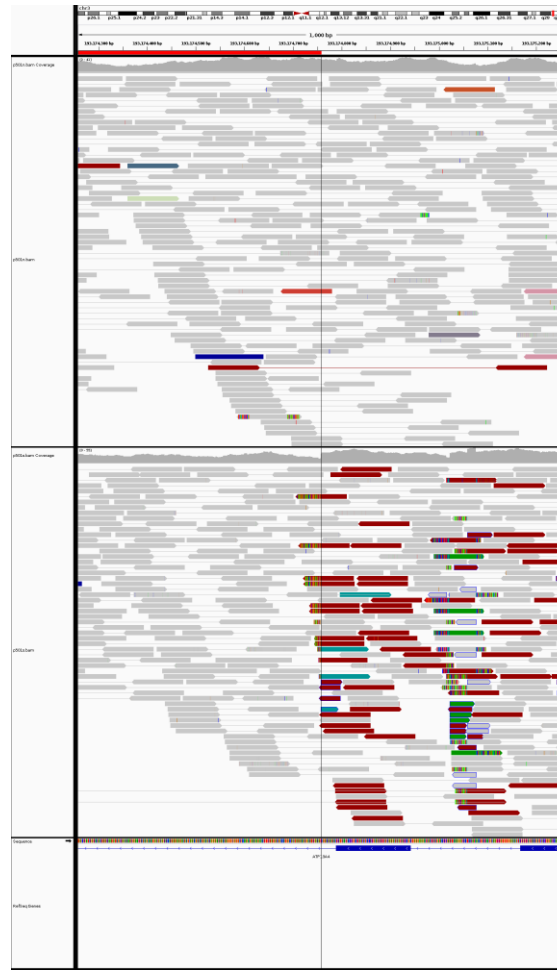
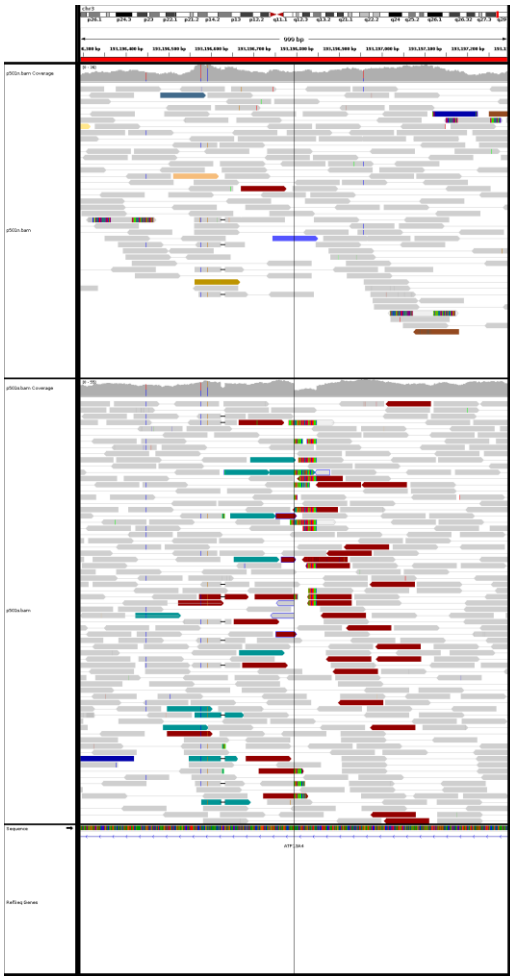


Structural variants

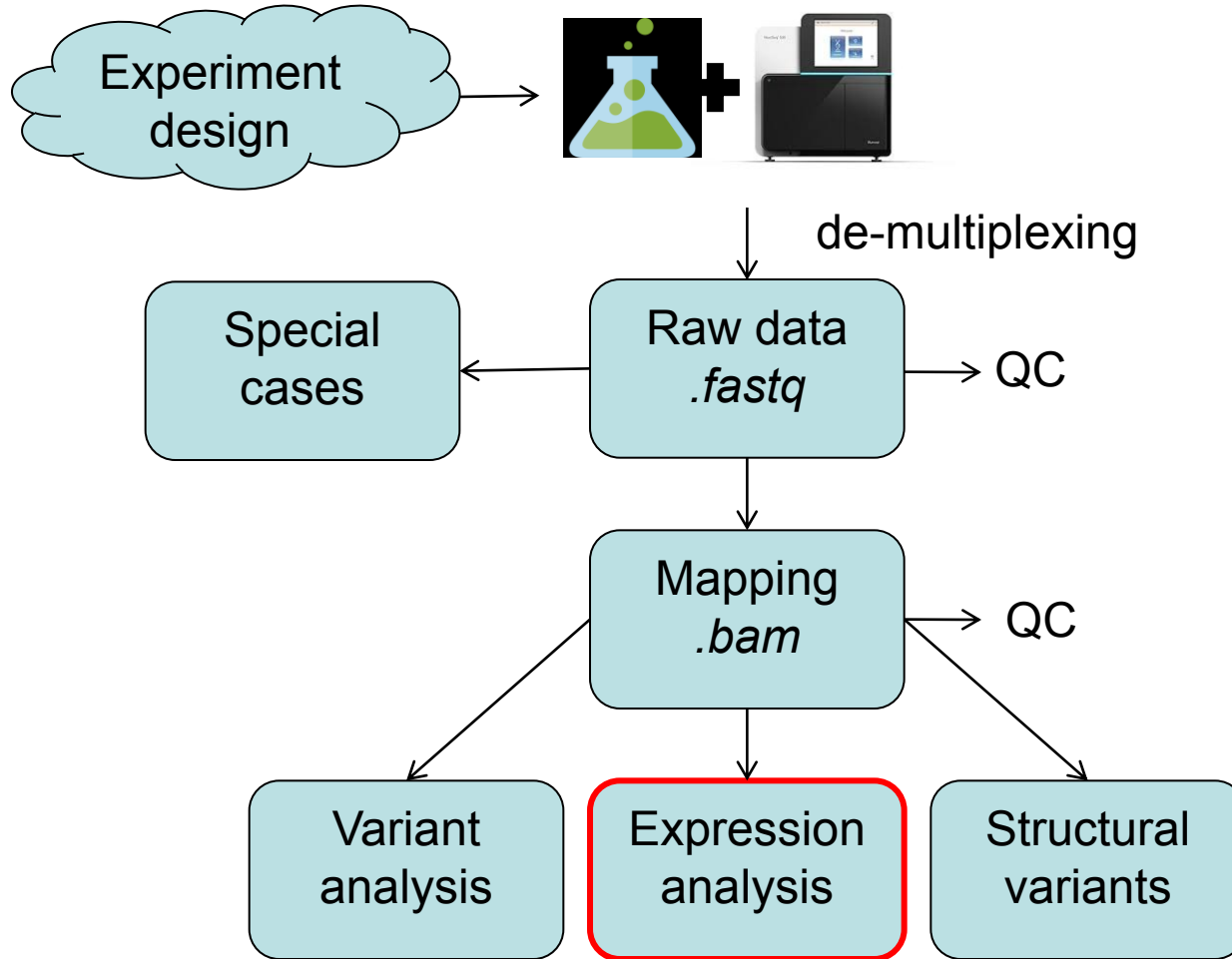
- **CNV**
- **long variants in WGS – ControlFreec**
- **Smaller variants for WES / target panel**
 - **Somatic – tumor,normal**
 - **Germline - lot of references**
 - **XHMM**
- **Read-pairs very noisy expect a lot of FP**
- **BreakPoint**
 - **Target panel with short reads**
- **Delly**
 - **everything else**

Structural variants

- Manual check with IGV (batchmode)



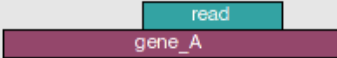
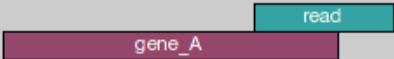


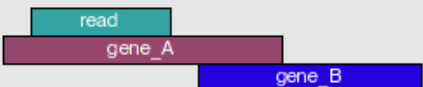
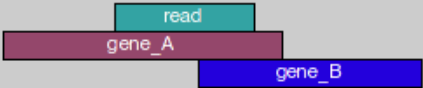
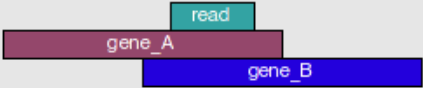
NGS data analysis



Expression analysis



Counting schemes

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

Expression analysis - planning

- **3 way balance**
 - **Read depth**
 - **Biological replicates**
 - **Fold change (number of genes) sensitivity**

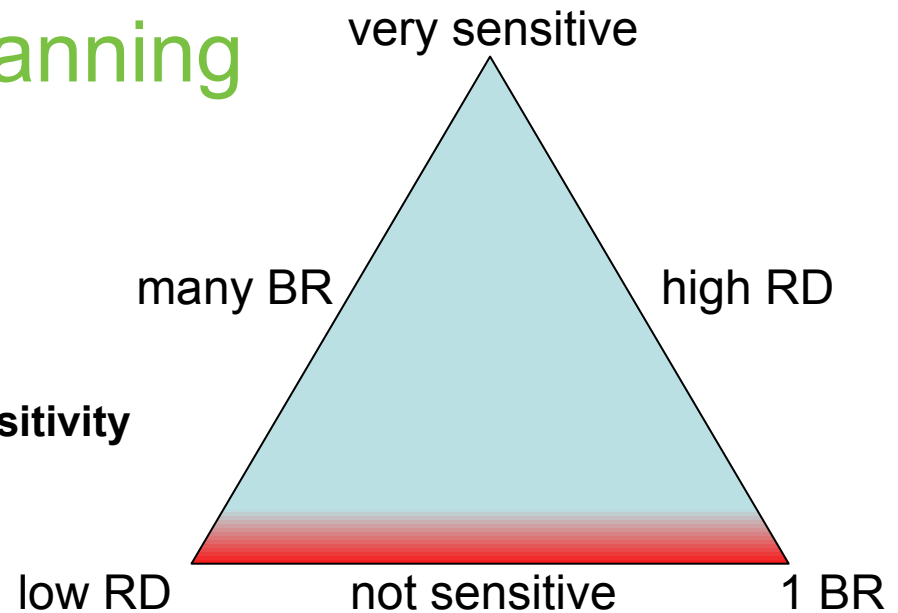


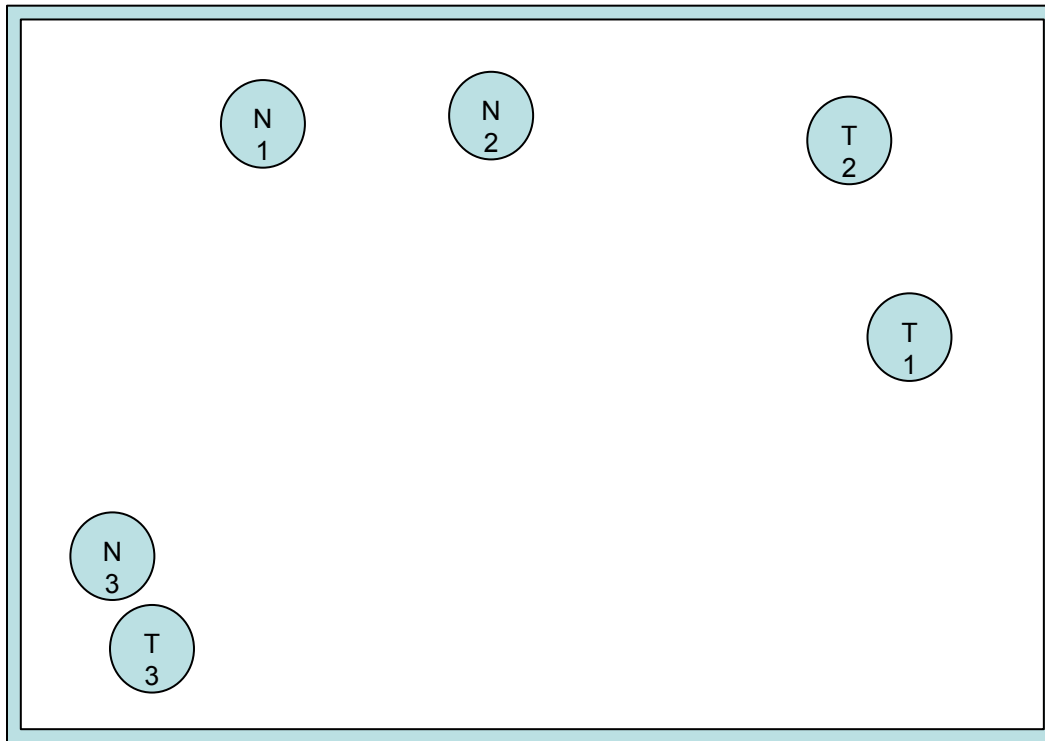
Table 1

Statistical power to detect differential expression varies with effect size, sequencing depth and number of replicates

	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Expression analysis - planning

- **Replicates**
- **Technical vs. biological**
 - Technical only for technique testing
- **Highly suggested minimum = 4 rep**



Expression analysis - planning

- **Depth**
- **Human ~ 22 000 genes = minimum 20 mil mapped reads**
- **Good 25 mil mapped reads**

- **Mapped reads!**
 - rRNA removal
 - Size selection for sRNA

- **Trade-off**
 - 4 replicates with 20 mil vs. 3 replicates with 30 mil
 - 9 replicates with 25 mil vs. 10 replicates with 20 mil

Expression analysis - planning

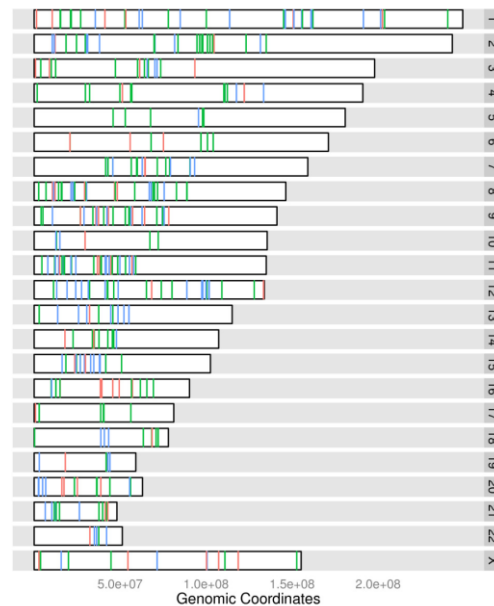
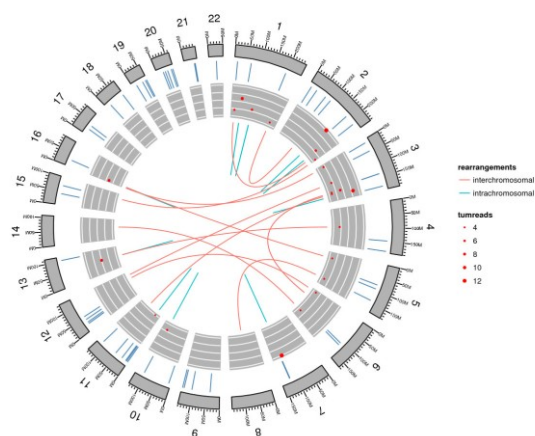
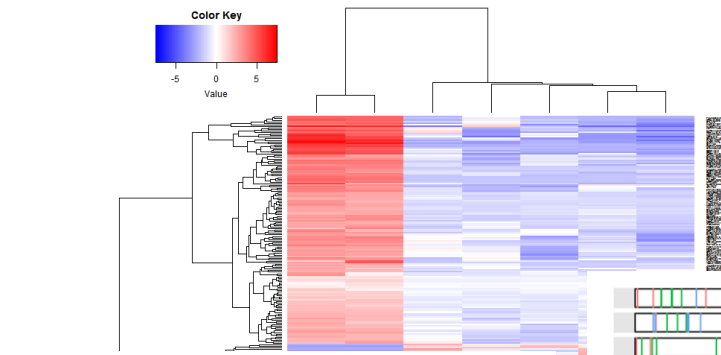
- **Depth**
- **Human ~ 22 000 genes = minimum 20 mil mapped reads**
- **Good 25 mil mapped reads**

- **Mapped reads!**
 - **rRNA removal – 90% rRNA**
 - **Size selection for sRNA**

- **Trade-off**
 - 4 replicates with 20 mil** vs. 3 replicates with 30 mil
 - 9 replicates with 25 mil vs. **10 replicates with 20 mil**
 - Reality: 3 replicates with 15 mil reads

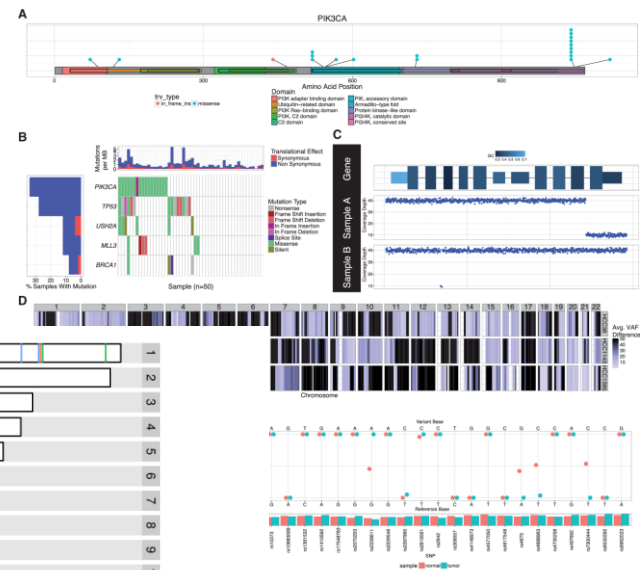
Visualization

- Bioinformatician provide all he think might be helpful
- Researcher requests exactly what he wants
- Good to find a balance



seqReg

- Exon
- Intron
- Other



Thank you for your attention



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

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



EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE



**OP Research and
Development for Innovation**

