CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

Moderní metody analýzy genomu

Mgr. Lenka Radová, Ph.D.

Brno, 4.12.2015

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND
INVESTING IN YOUR FUTURE

2007–13
OP Research and Development for Innovation

Pipeline/Workflow

Base calling

Reads pre-processing

Mapping on reference

Post-processing

Quality based variant detection

Local realignment

Model based variant detection

Variant annotation

Biological interpretation

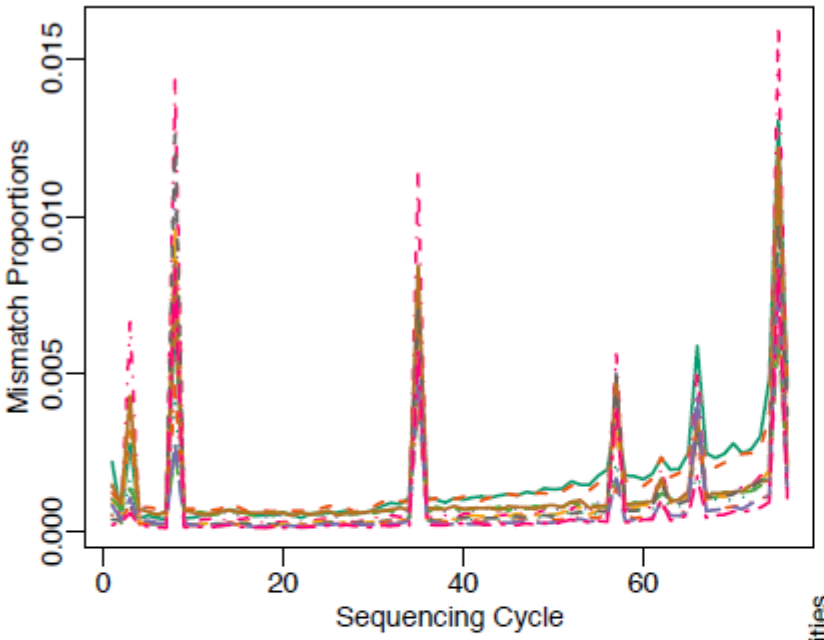
results

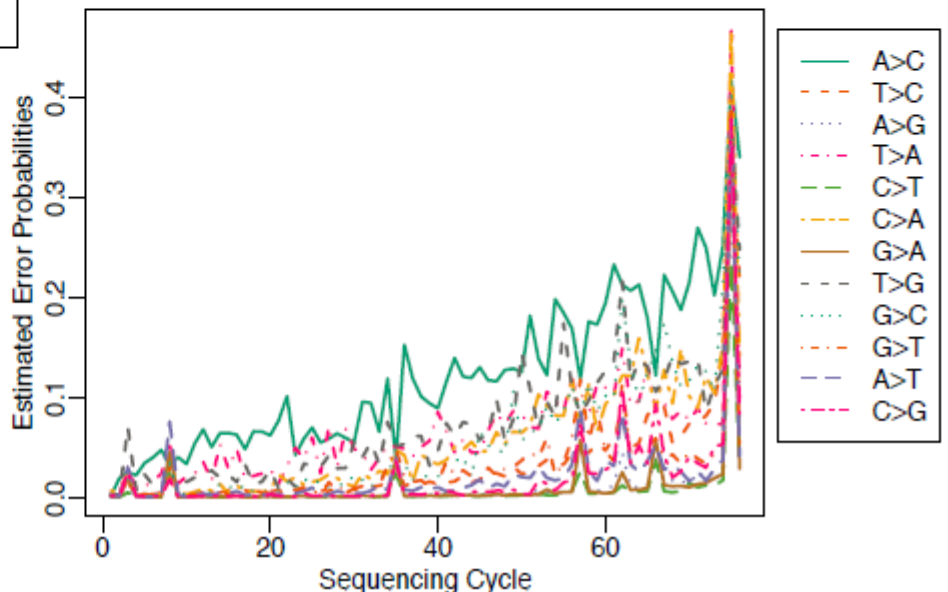
results

results

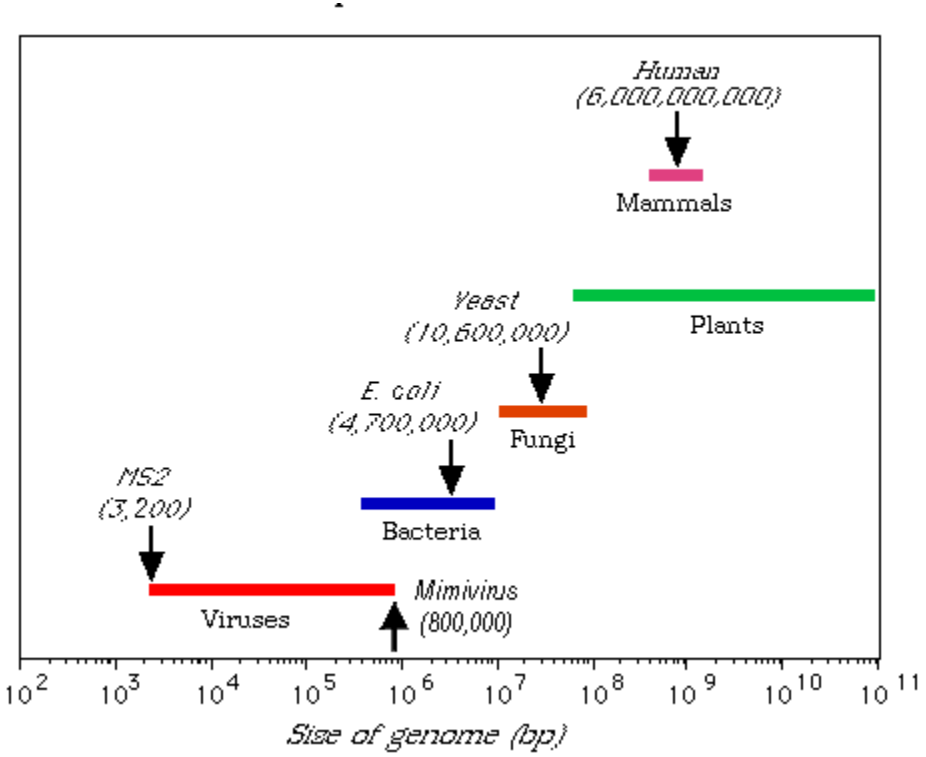
But…

many steps in experimental process may introduce errors and biases

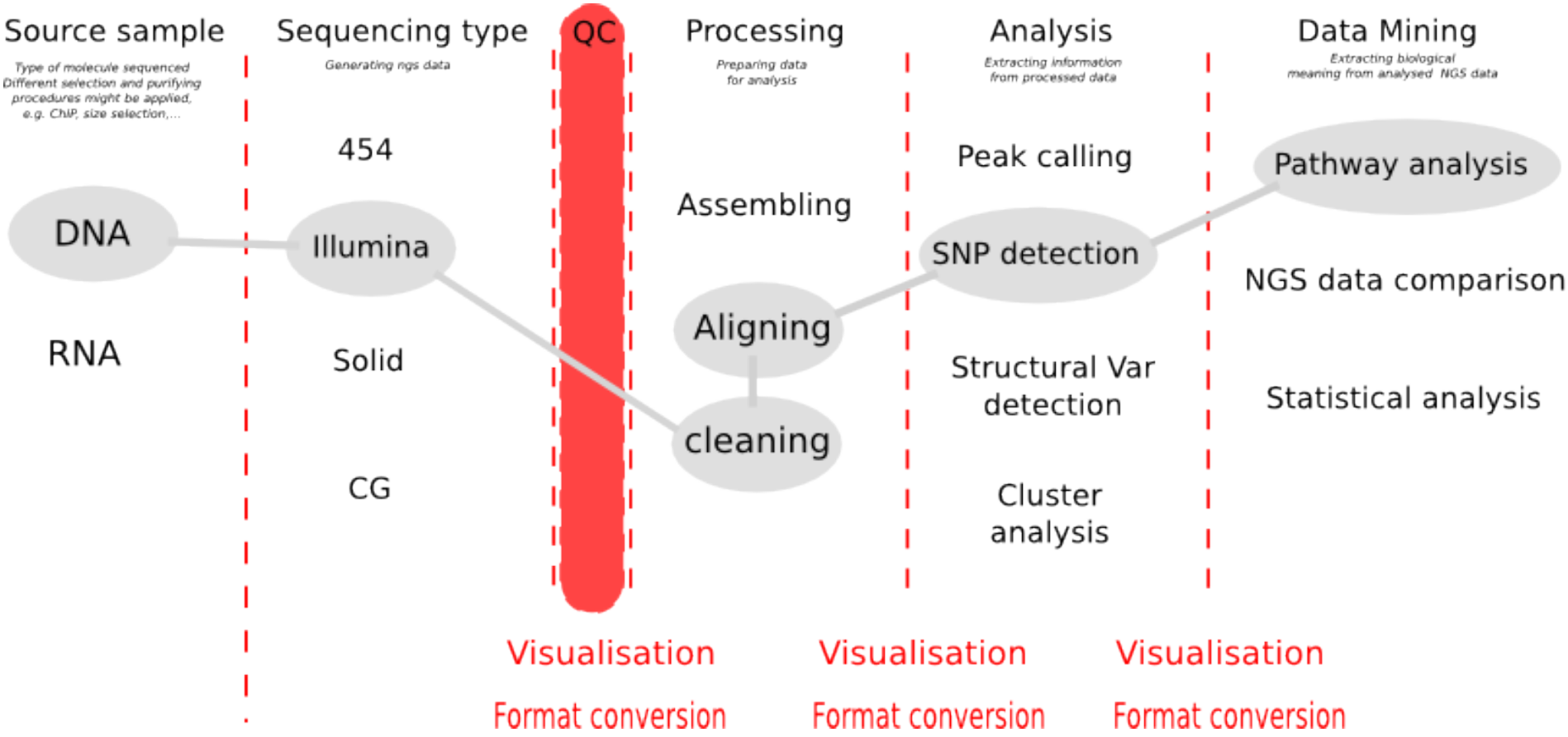




Scales of genome size



Quality control

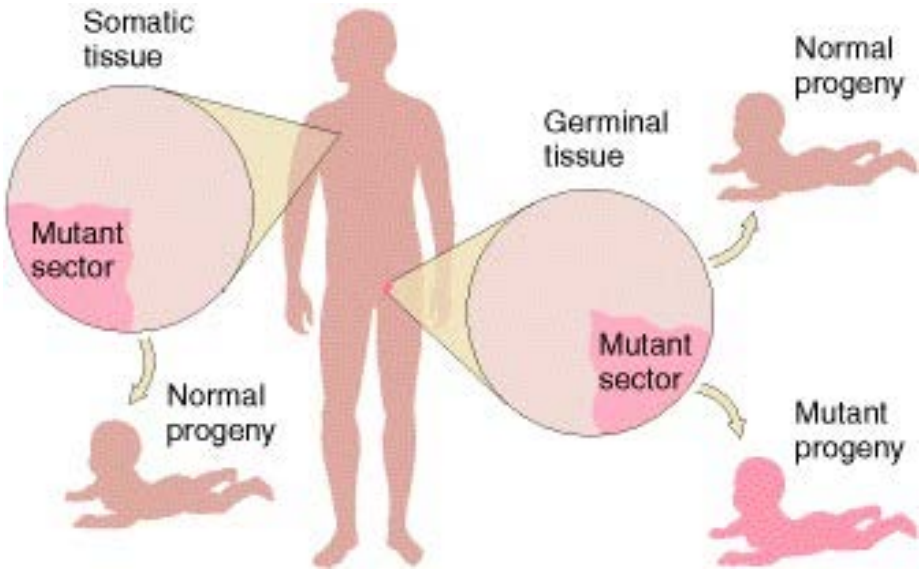


FASTQ format

- The first line starts with '@', followed by the label
- The third line starts with '+'. In some variants, the '+' line contains a second copy of the label
- The fourth line contains the Q scores represented as ASCII characters

Basic biological problems

- Identification of mutations
  - somatic
  - germinal

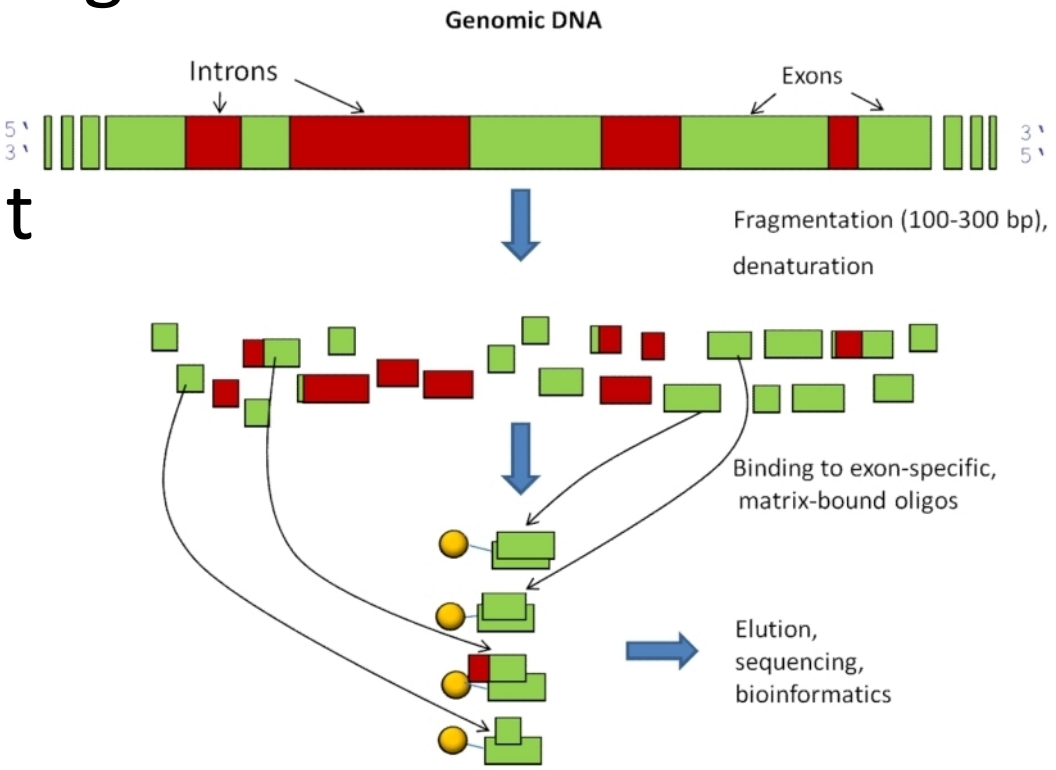


- Expression analyses - genes, miRNAs, etc.

An Introduction to Genetic Analysis. 7th edition. Griffiths AJF, Miller JH, Suzuki DT, et al. New York: W. H. Freeman; 2000

Mutation identification

- Whole exome or whole genome data, ultra-deep sequencing
- Output: VCF-format



Mutation identification

- Aim: identification of point mutations

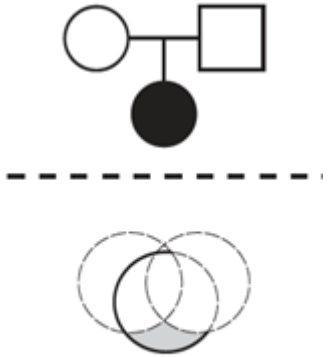
- Application: diagnostic of diseases

  - inherited (germinal, de-novo mutations)

    e.g. familiar hypercholesterolemia, hemophylia, cystic fibrosis…

  - gained (somatic mutations)

    e.g. cancer, leukemia, …



De novo based strategy

Germinal mutations

- Comparison with reference genome
- Expected allele frequency: 30-100%
- Softwares: GATK, VarScan, …
- Usage: e.g. prenatal diagnostic

Somatic mutations

- Comparison tumor-normal (matched, unmatched)
- Expected allele frequency:

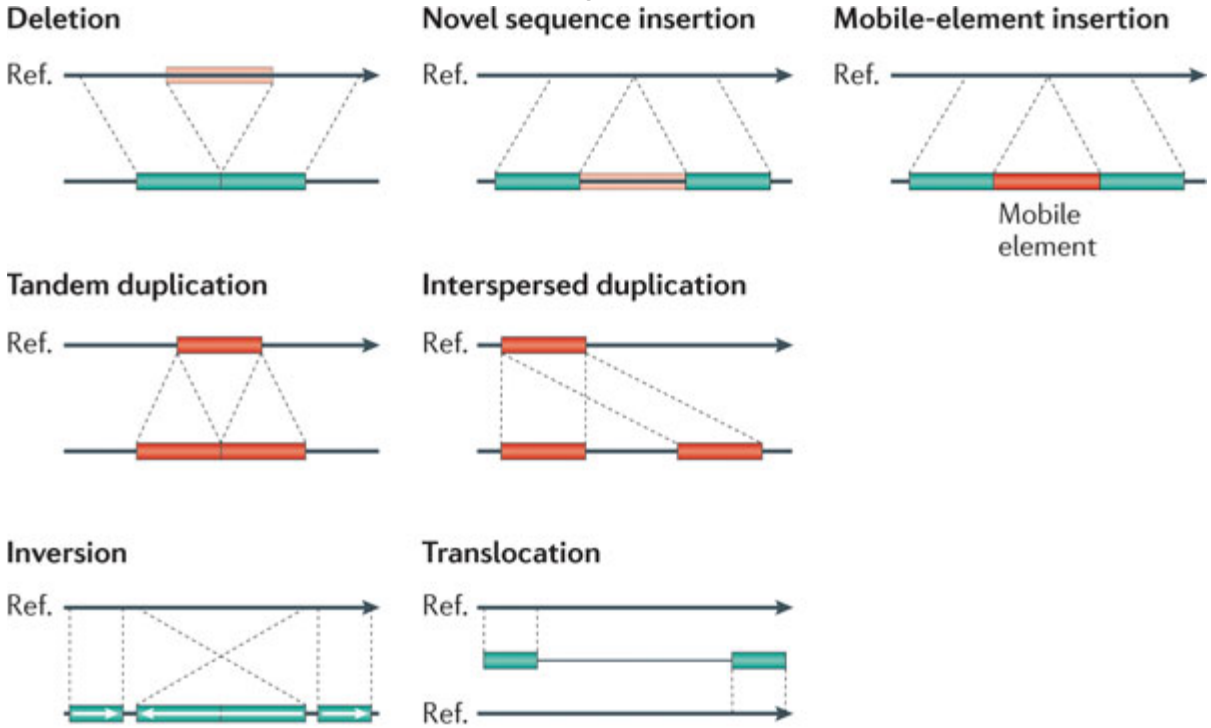
  >0,2%

- Softwares: MuTect, FreeBayes, deepSNV, …
- Usage: translational research, cancer diagnostic, personalized medicine,…

Advanced biological problems

- Structural variant discovery

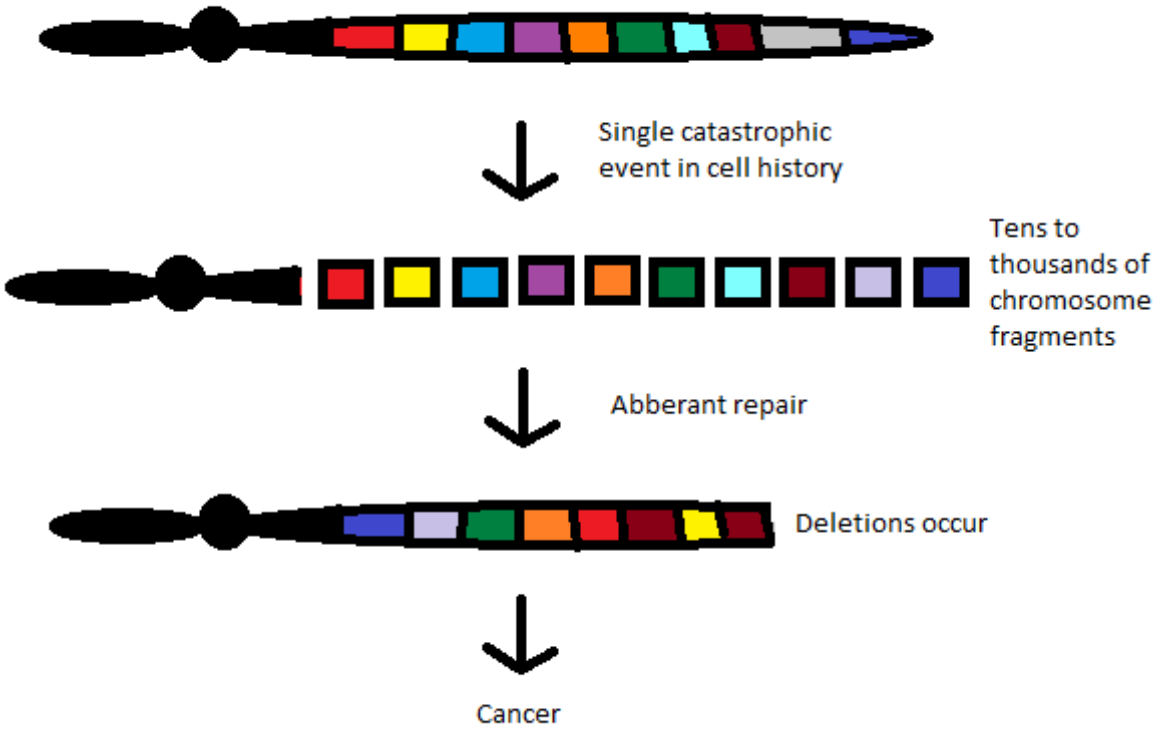
(deletions, duplications, CN variants, insertions, inversions, translocations)



Nature Reviews | Genetics

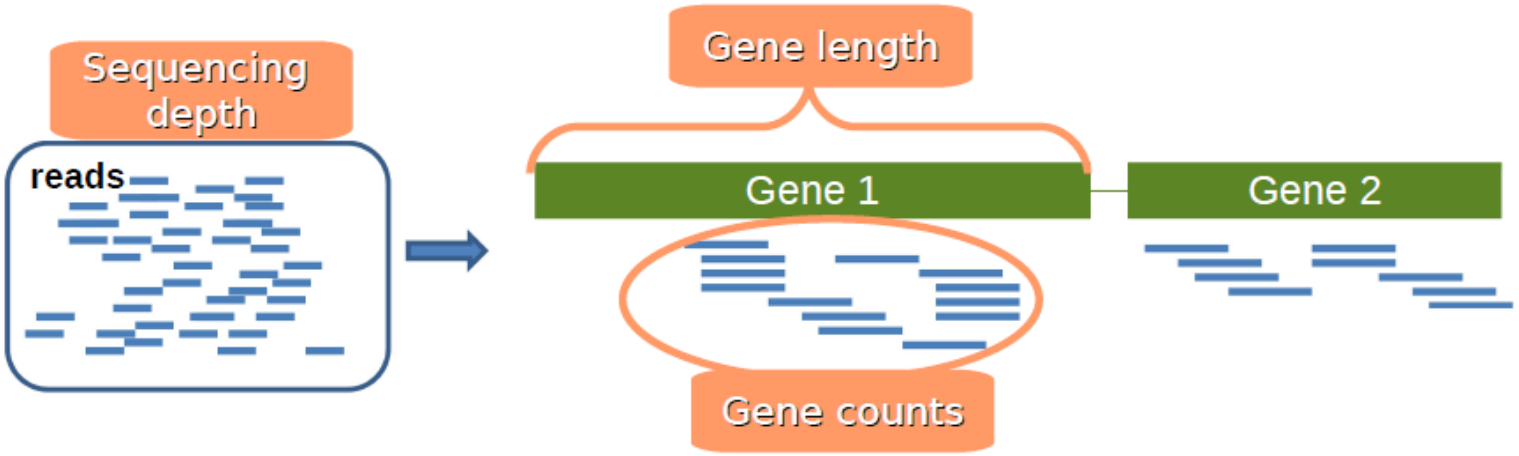
Advanced biological problems

- Chromotripsis = thousands of clustered chromosomal rearrangements occur in a single event in localised and confined genomic regions in one or a few chromosomes



Expression analyses – RNA-seq

- characterization of gene expression in cells via measurement of mRNA levels



- Output: expression level table

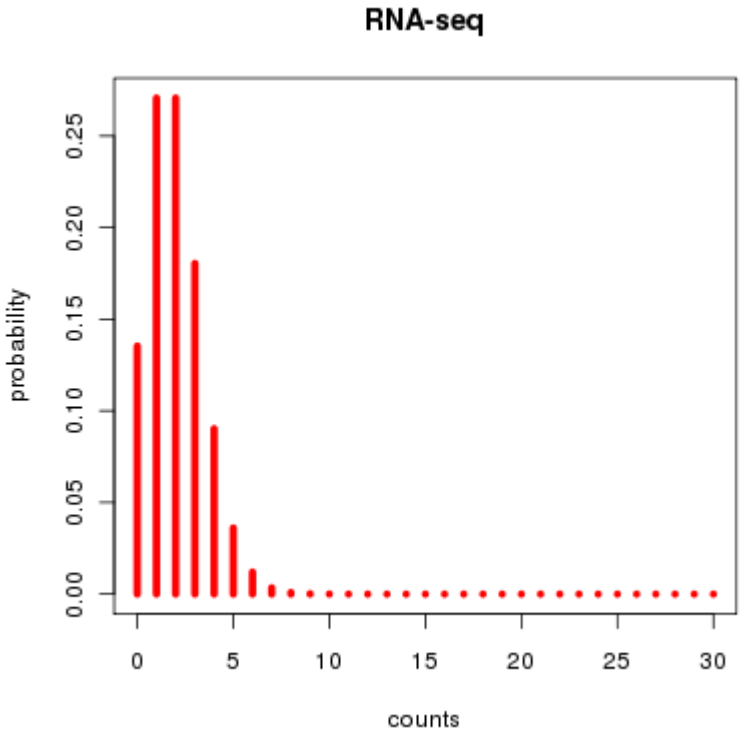
RNA-seq

- Aim: identification of genes differentially expressed in tissues with different conditions (tumor vs normal, treated vs untreated, different stages of illness, …)

- Application: translational research, diagnostic of diseases

Expression level in RNA-seq

= The number of reads (counts) mapping to the biological feature of interest (gene, transcript, exon, etc.) is considered to be linearly related to the abundance of the target feature

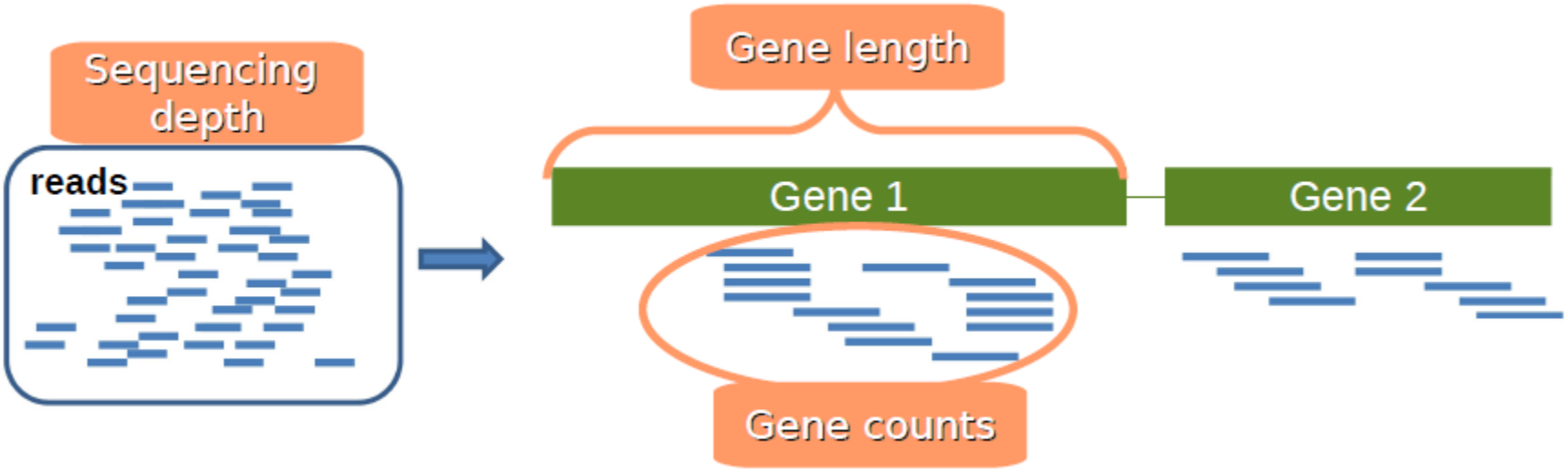


What is differential expression?

- A gene is declared **differentially expressed** if an observed difference or change in read counts between two experimental conditions is statistically significant, i.e. whether it is greater than what would be expected just due to natural random variation.
- Statistical tools are needed to make such a decision by studying counts probability distributions.

Definitions

- Sequencing depth: Total number of reads mapped to the genome. Library size.
- Gene length: Number of bases.
- Gene counts: Number of reads mapping to that gene (expression measurement)



Experimental design

- Pairwise comparisons: Only two experimental conditions or groups are compared.
- Multiple comparisons: More than 2 conditions or groups.

Replicates

- Biological replicates. To draw general conclusions: from samples to population.
- Technical replicates. Conclusions are only valid for compared samples.

RNA-seq biases

- Influence of sequencing depth: The higher sequencing depth, the higher counts.
- Dependence on gene length: Counts are proportional to the transcript length times the mRNA expression level.
- Differences on the counts distribution among samples.

Options

1. Normalization: Counts should be previously corrected in order to minimize these biases.

2. Statistical model should take them into account.

Normalization methods

- **RPKM** (Mortazavi et al., 2008) = Reads per kilo base per million: Counts are divided by the transcript length (kb) times the total number of millions of mapped reads

$$RPKM = \frac{\text{number of reads of the region}}{\frac{\text{total reads}}{1000000} \times \frac{\text{region length}}{1000}}$$

- **Upper-quartile** (Bullard et al., 2010): Counts are divided by upper-quartile of counts for transcripts with at least one read.

- **TMM** (Robinson and Oshlack, 2010): Trimmed Mean of M values.

- **Quantiles**, as in microarray normalization (Irizarry et al., 2003).

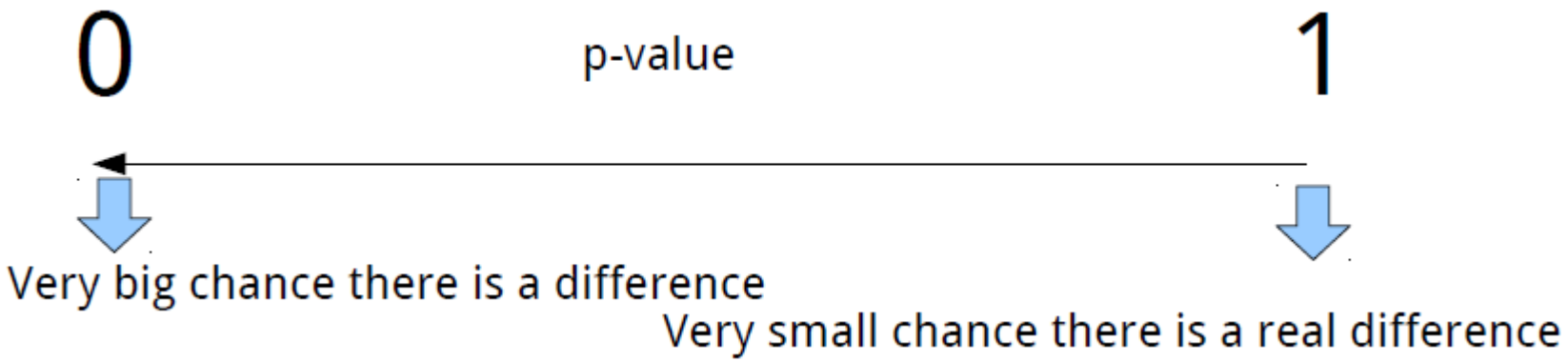
- **FPKM** (Trapnell et al., 2010): Instead of counts, Cufflinks software generates FPKM values (Fragments Per Kilobase of exon per Million fragments mapped) to estimate gene expression, which are analogous to RPKM.

Differential expression

- Parametric assumptions: Are they fulfilled?
- Need of replicates.
- Problems to detect differential expression in genes with low counts.

Goal

- Based on a **count table**, we want to detect differentially expressed genes between **conditions** of interest.
- We will assign to each gene a p-value (0-1), which shows us 'how surprised we should be' to see this difference, when we assume there is no difference.



Algorithms under active development

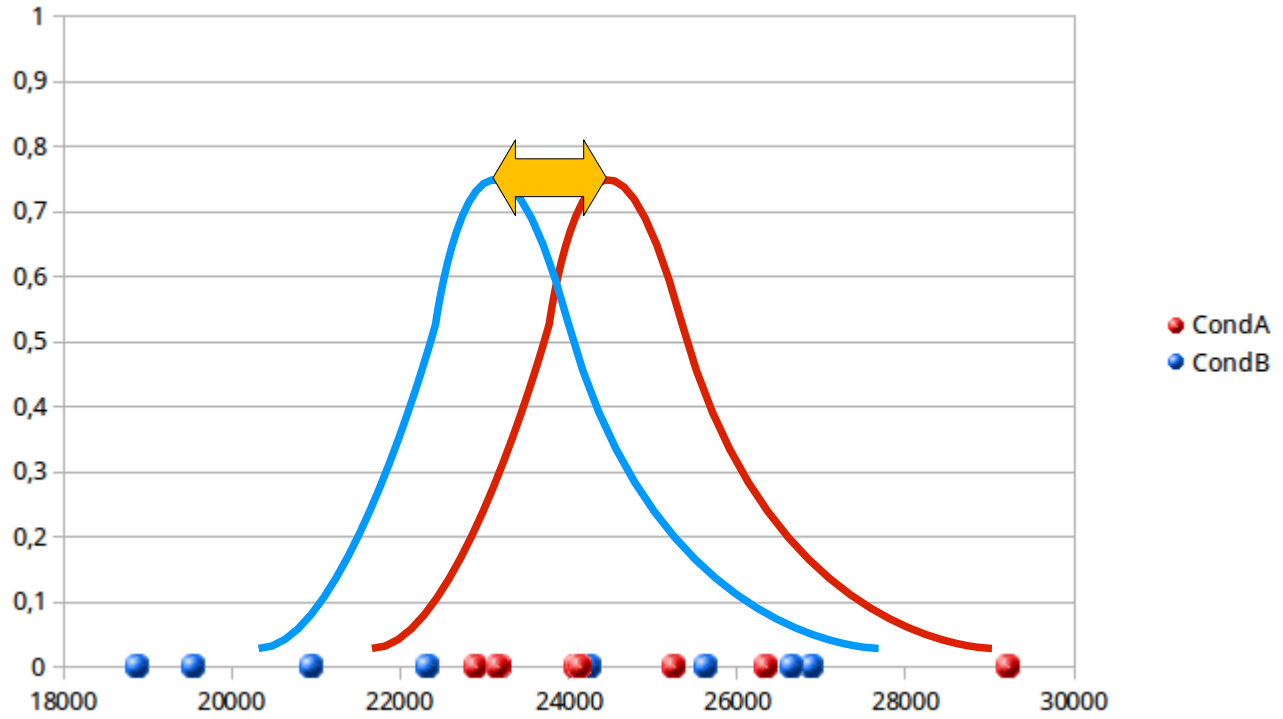
Detecting differential expression by count analysis

- edgeR - DE on the gene level from counts - TOP
- DEseq - DE on the gene level from counts - TOP
- tweeDEseq - DE on the gene level from counts
- NBPSeq - DE on the gene level from counts
- TSPM - DE on the gene level from counts
- SAMseq - non-parametric method on the gene level from counts - TOP if large number of replicates
- ShrinkSeq - DE on the gene level from counts
- BBSeq - DE on the gene level
- Bayseq - DE on the gene level from counts - TOP
- DEGseq - DE on the gene level
- sydSeq - improved DE on the gene level for low replicate studies
- DEXSeq - DE on the exon level
- NOIseq - Non-parametric method from counts
- CuffLinks cuffdiff2 - DE on the isoform level - TOP
- BitSeq - DE on the isoform level
- EBSeq - DE on the isoform level from counts
- Myrna - cloud computing for large RNA-seq datasets
- sSeq - optimized for small sample size experiments.
- MRFSeq - optimized for small read counts
- QuasiSeq - apply the QL, QLShrink and QLSpline methods to RNA-seq data for DE

http://wiki.bits.vib.be/index.php/RNAseq_toolbox#Detecting_differential_expression_by_count_analysis

Intuition

Difference is quantified and used for p-value computation

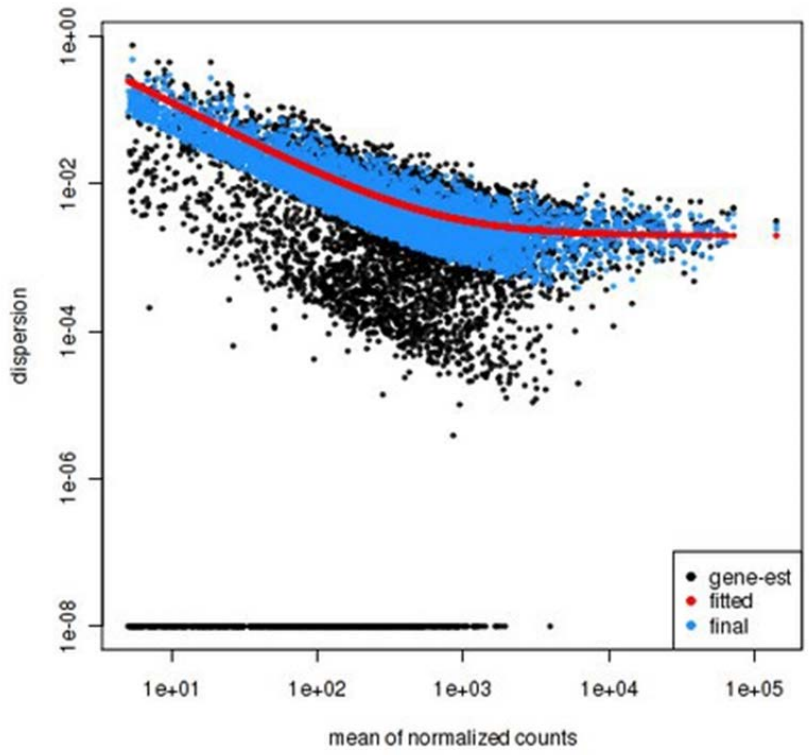


Dispersion estimation

- For every gene, a NB is fitted based on the counts. The most important factor in that model to be estimated is the dispersion.

- DESeq2 estimates dispersion by 3 steps:
  1. Estimates dispersion parameter for each gene
  2. Plots and fits a curve
  3. Adjusts the dispersion parameter towards the curve ('shrinking')

Dispersion estimation

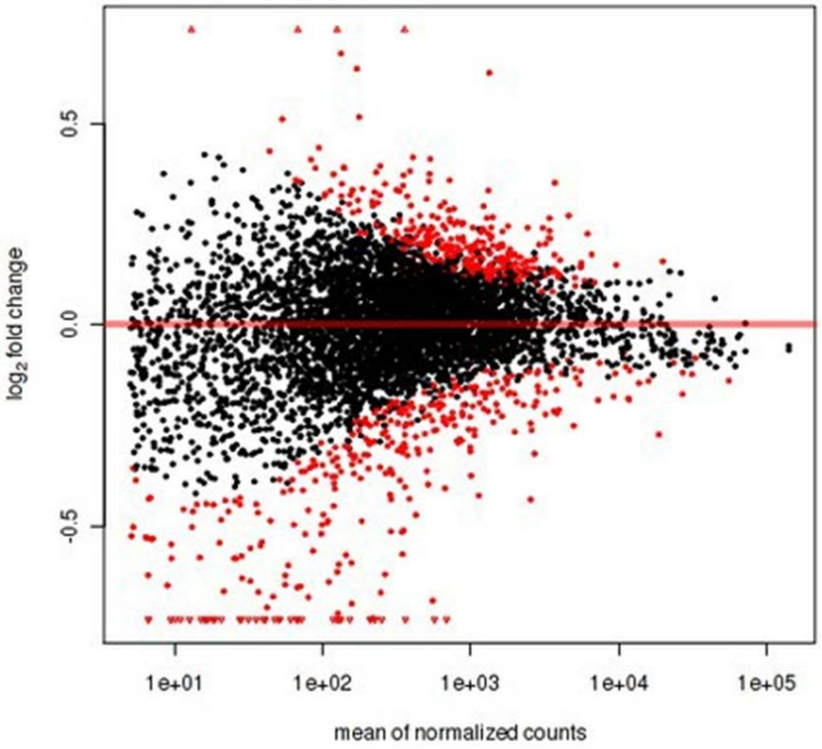


- Black dots = estimates from the data
- Red line = curve fitted
- Blue dots = final assigned dispersion parameter for that gene

Model is fitted

Test runs between 2 conditions

- for each gene 2 NB models (one for each condition) are made, and a Wald test decides whether the difference is significant (red in plot).



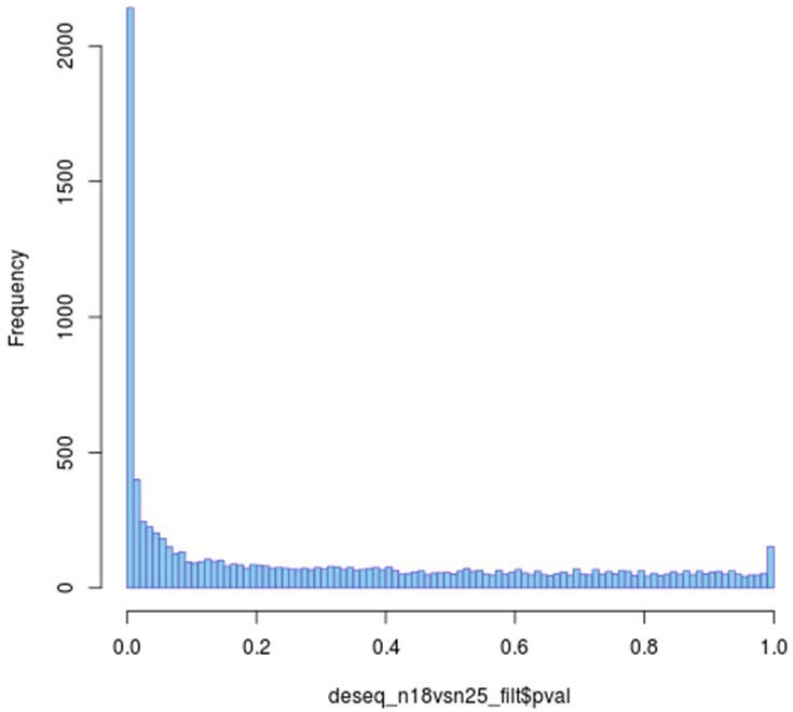
Test runs between 2 conditions

- for each gene 2 NB models (one for each condition) are made, and a Wald test decides whether the difference is significant (red in plot).

i.e. we are going to perform thousands of tests…

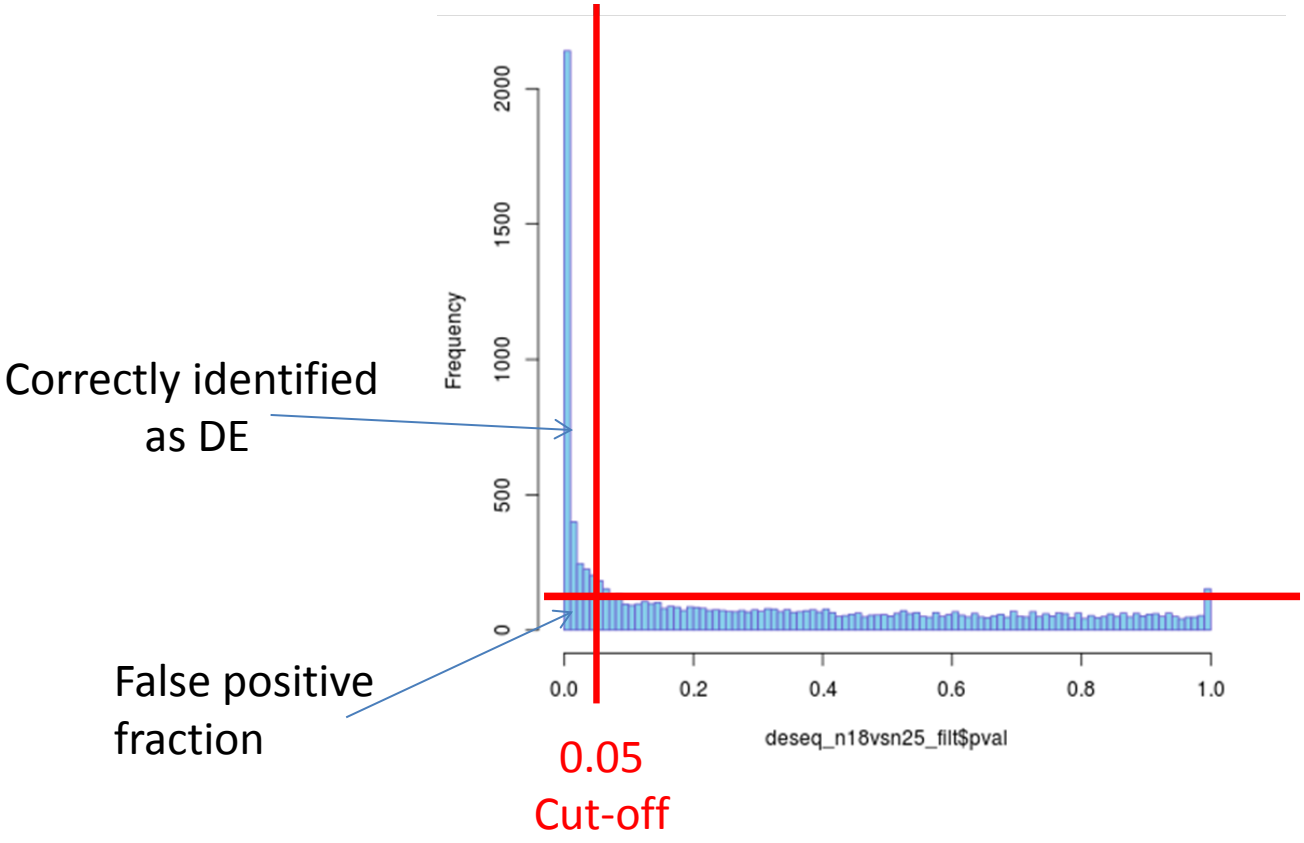
(if we set set a cut-off on the p-value of 0,05 and we have performed 20000 tests, 1000 genes will appear significant by chance)

Check the distribution of p-values



- If the histogram of the p-values does not match a profile as shown here, the test is not reliable. Perhaps the NB fitting step did not succeed, or confounding variables are present.

Improve test results



Improve test results

- Avoid testing = apply a filter before testing, an independent filtering

- Apply multiple testing correction

Multiple testing corrections

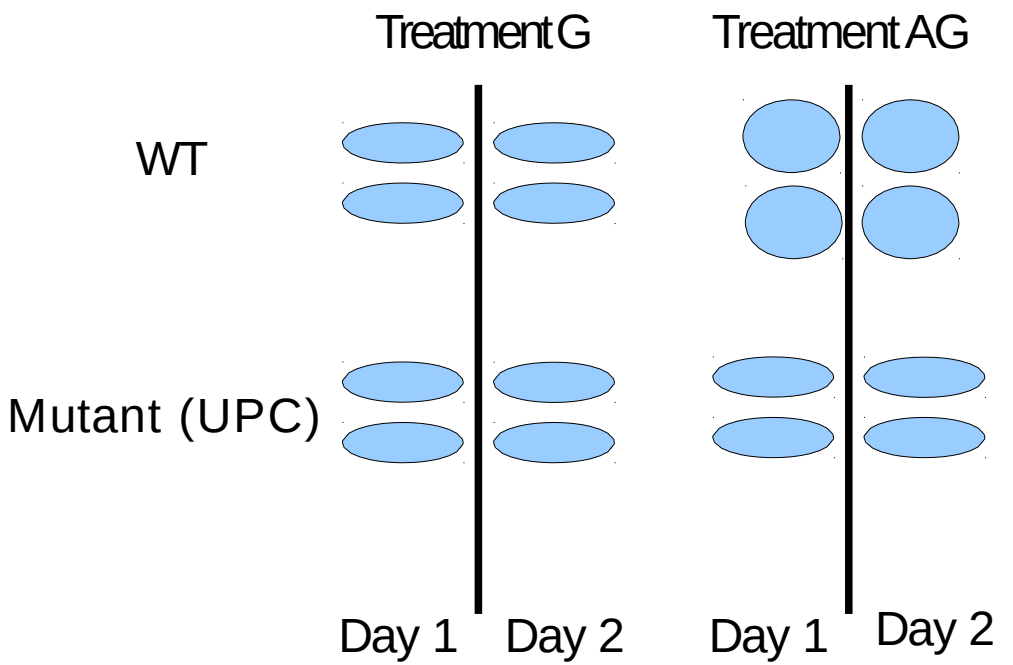
- Bonferroni or Benjamini-Hochberg correction, to control **false discovery rate** (FDR).

- FDR is the fraction of false positives in the genes that are classified as DE.

alpha	0.0001	0.001	0.01	0.025	0.05	0.1
Uncorrected	31	57	93	118	134	188
Bonferroni	0	6	13	21	24	31
FDR	0	19	44	63	73	91

- If we set a threshold α of 0,05, **20%** of the DE genes will be false positives.

Including different factors



sample	strain	treatment	day
sample1	WT	G	1
sample2	WT	AG	1
sample3	WT	G	1
sample4	WT	AG	1
sample5	UPC	G	1
sample6	UPC	AG	1
sample7	UPC	G	1
sample8	UPC	AG	1
sample9	WT	G	2
sample10	WT	AG	2
sample11	WT	G	2
sample12	WT	AG	2
sample13	UPC	G	2
sample14	UPC	AG	2
sample15	UPC	G	2
sample16	UPC	AG	2

Additional metadata
(batch factor)

Which genes are DE between UPC and WT?

Which genes are DE between G and AG?

Which genes are DE in WT between G and AG?

Statistical model

Gene = strain + treatment + day

- export results for unique comparisons

Goal

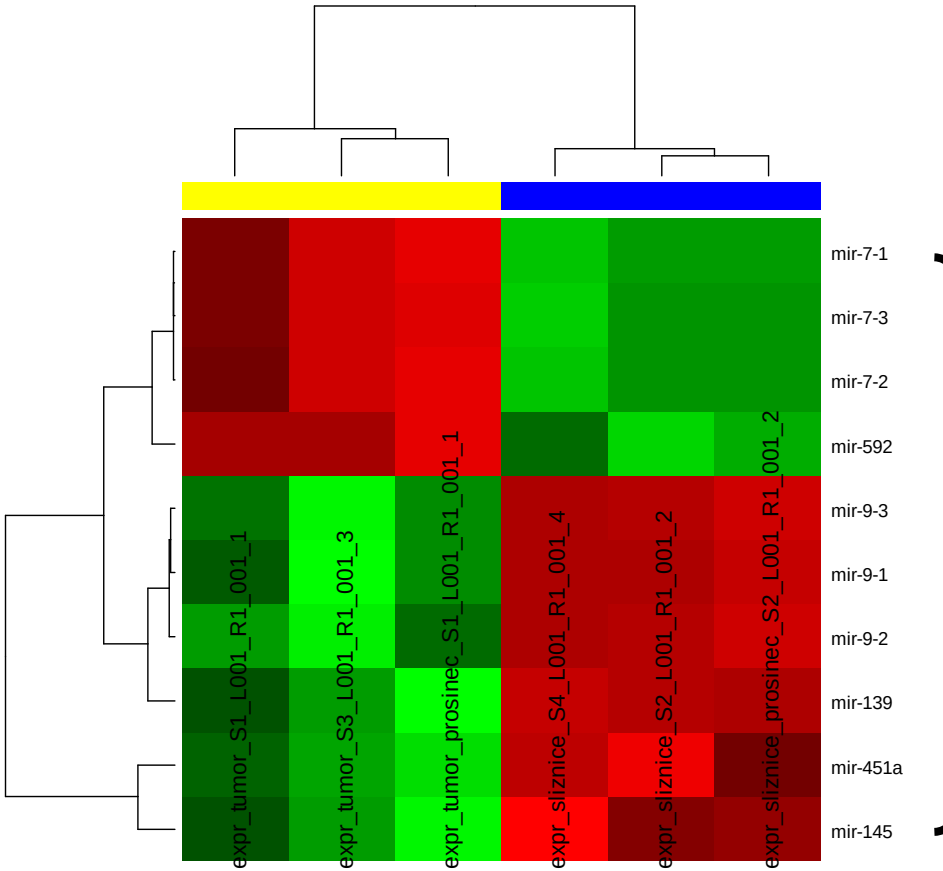
Galaxy / BITS — Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User

gene_id	sample1	sample2	sample3	sample4	sample5	sample6	sample7	sample8	sample9	sample10	sample11	sample12	sample13	sample1
CAF0006876	23171	22903	29227	24072	23151	26336	25252	24122	19527	26898	18880	24237	26640	2231
CAF0006885	647	698	854	765	797	816	868	767	532	761	563	654	748	72
CAF0006887	10	3	8	8	5	8	5	3	7	8	2	10	7	
CAF0006888	1	2	1	1	0	0	0	0	1	0	1	0	0	
CAF0006889	2	0	1	0	1	0	2	0	1	1	1	0	0	
CAF0006890	852	735	1032	810	1476	1437	1575	1358	644	859	549	747	1320	94
CAF0006891	475	465	624	505	538	624	654	562	431	586	410	550	639	47
CAF0006892	85	67	73	80	151	91	114	93	81	65	47	84	91	7

Galaxy / BITS — Analyze Data | Workflow | Shared Data | Visualization | Admin

	baseMean	log2FoldChange	lfcSE	pvalue	padj
CAF0006965	236.95771532567	0.319894269325064	0.0795476625084231	5.78492554744642e-05	0.00484865585947968
CAF0006989	152.753854809905	-0.47673982481625	0.120420053359006	7.52725227015407e-05	0.00561314522325369
CAF0007413	394.18013915485	0.545507459785333	0.103161564037881	1.23732350682432e-07	2.42600739993209e-05
CAL0000006	3840.73677986616	-0.675753238608597	0.0614877057756516	4.26668298965338e-28	6.06508986979228e-25
CAL0000023	97.9171191032388	0.42580183962291	0.109195747881053	9.64169841515241e-05	0.00668569477909227
CAL0000038	292.453306221006	-0.290563708698689	0.0702804475299353	3.55966374624607e-05	0.00343055051883985
CAL0000039	724.903093908146	-0.209063501932311	0.0523592353116698	6.52789812704274e-05	0.00515522621532848

Visualization of results - heatmap



Differentially expressed miRNAs with adjusted p<0,01

samples