GOOD PRACTICES FOR RNA-SEQ DATA ANALYSIS

PAOLO SONEGO

FEBRUARY 5TH, 2020

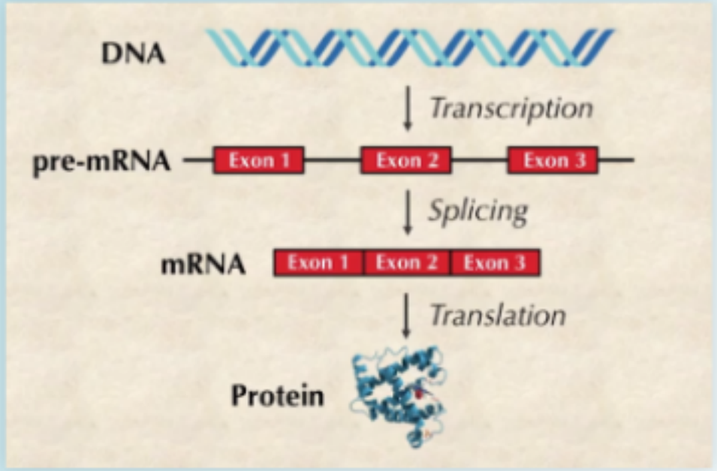
BASIC SETUP FOR REPRODUCIBILITY

- Docker
- Python/Jupyter Notebook or RMarkdown
- Github
- Organize your data, scripts, resources
- Keep track of versions for both annotation and software used in the analysis
- Literate programming paradigm for reproducible research by Donald E. Knuth

OVERVIEW

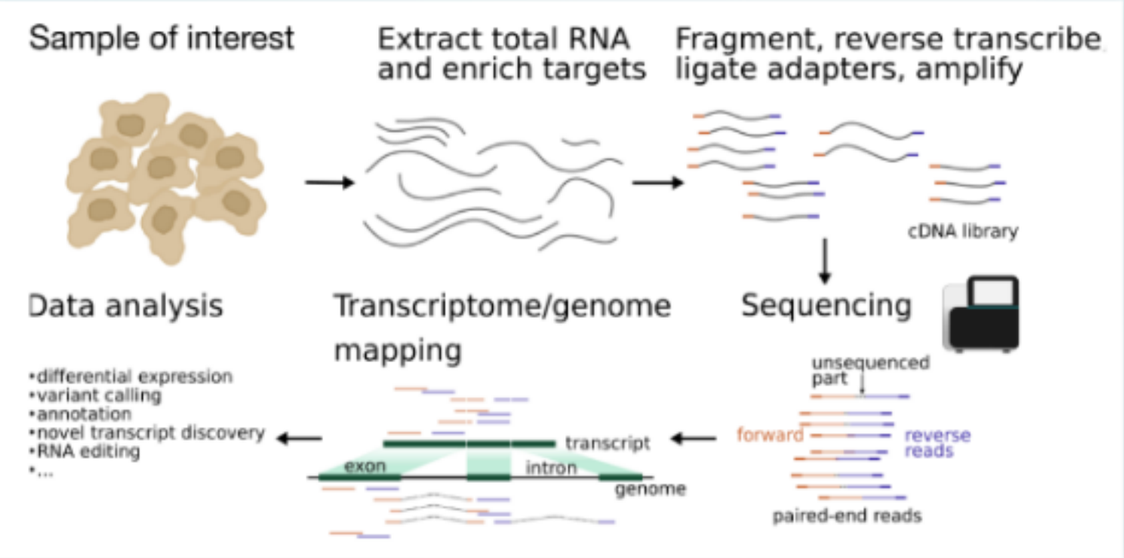
- General issues
  - What is RNA-seq and why do you want to perform a sequencing experiment?
  - Data mining for gene epression experiments
  - Issues in experimental design of a RNA-seq experiment
  - Advices and tips for gene expression analysis
- Steps for the analysis of gene expression data
  - From raw data (*FASTQ*) to counts
  - From a counts matrix to lists of differentially expressed genes
  - From a list of DE genes to biological inference

WHAT IS RNA-SEQ?



- RNA-Seq uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment - It allows to take a snapshot of a tissue/cell at a particular time and knowing which genes are expressed at that point
- It works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced
- it is used as a proxy for the changes in the encoded proteins
- Why are we not doing protein profiling?
- Why are we not doing microarray (MA) instead of RNA-seq?

RNA-SEQ LIBRARY CONSTRUCTION



© Koen Van den Berge et al. (2019)

- After total RNA extraction from a sample of interest, rRNA is depleted (either using poly(A)-selection or rRNA depletion) and the remaining RNA molecules are fragmented ideally achieving a uniform size distribution (~300-500 bp)
- Single-stranded target RNA are reverse transcribed to cDNA
- Double-stranded cDNA is synthesized and the adapters for sequencing are added to construct the final library which will be carried out on a Illumina Sequencer Flow Cell (Zeng and Mortazavi, 2012)*

[*] Zeng,W. and Mortazavi,A. (2012) *Technical considerations for functional sequencing assays.* Nat Immunol, 13, 802–807.

WHY DO WE NEED RNA-SEQ?

- **Gene expression profiling: which genes are active and how much they are transcribed**
- Alternative splicing
- Detect novel transcripts

GENE EXPRESSION DATA ANALYSES

- Class Discovery (Clustering):
  - Discover groups of genes with similar gene expression profiles
  - Discover groups of samples with similar gene expression profiles
- Class Prediction (Classification):
  - Using gene expression profiles train an algorithm on samples with known class membership (training set) in order to establish a prediction rule to classify new samples (test set).
- **Class Comparison (Differential Expression):**
  - **Identify over (up-regulated) and under (down-regulated) expressed genes in selected comparisons.**

While a good design does not guarantee a successful experiment, a suitably bad design guarantees a failed experiment

Kathleen Kerr

GENERAL ADVICES

- There is no way to get meaningful results from data of bad quality coming from poorly designed experiments!
- Take away message:
  - Clearly define the biological question of interest
  - Carefully design the experiment
  - Always check the quality of experiments

GARBAGE IN, GARBAGE OUT!

EXPERIMENTAL DESIGN FOR RNA-SEQ

- Technical or biological replicates?
- Higher sequencing depth (library size) or more replicates?
- How many samples I need (power calculation)?
- Single-end or paired-end?
- Stranded or unstranded?

TECHNICAL OR BIOLOGICAL REPLICATES?

- **Technical replicates:**
  - Different library preparations from the same RNA sample (ENCODE consortium)
- **Biological replicates:**
  - Samples representing population under analysis
  - RNA from an independent growth of cells/tissue (ENCODE)
- Which replicates I need?
  - **Biological replicates are mandatory for statistical inference!**
  - Technical replicates can be used to measure technical variation in the same biological sample
  - **Technical replicates can not be a substitute to biological replicates!**

A NOTE ON TECHNICAL REPLICATES

- Technical variation in count data, in the absence of any other substantial biases or technical effects, follows approximately a Poisson distribution, and the sum of independent Poisson-distributed variables is also Poisson distributed.
- This means *you can safely combine technical replicates by adding the counts with no loss of information, provided the assumptions hold.*

SEQUENCING DEPTH (LIBRARY SIZE)

- Sequencing depth or library size refers to the number of sequenced reads for a given sample. As the sample is sequenced to a deeper level, the reads are likely to cover a larger proportion of the genome/transcriptome, allowing more transcripts to be detected with more precise quantification.
- Optimal sequencing depth depends on the aims of the experiment and on the complexity of the target transcriptome.

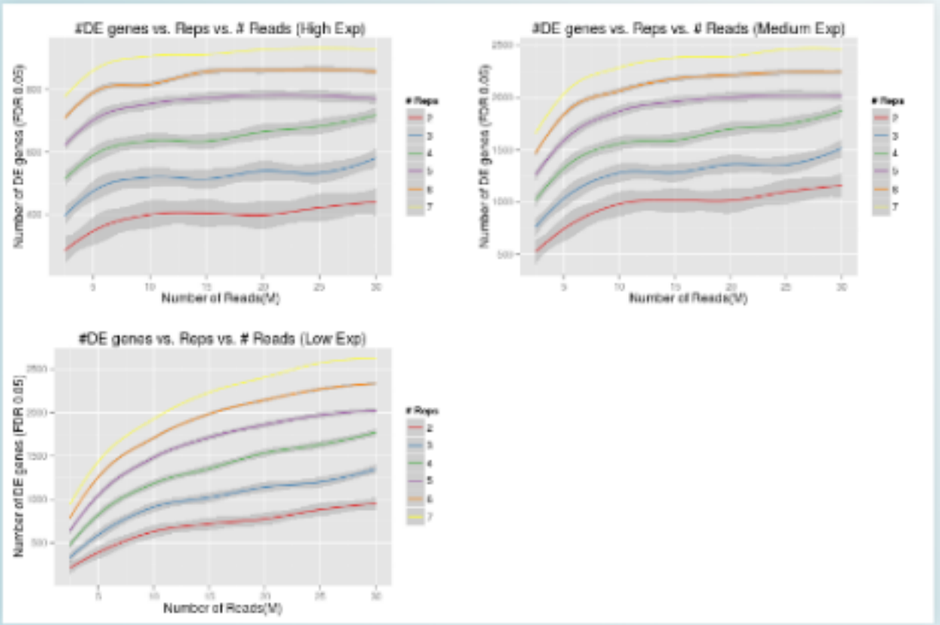
HOW MANY BIOLOGICAL REPLICATES I NEED ? (POWER AND SAMPLE SIZE ESTIMATION)

- Ideally best practice would suggest to apply methods of inferential statistics for selecting the number of samples that maximize the power of the analysis (reduce number of false negative without raise the number of false positive), e.g., Zhao, S., et al. (2018) *
- Practically:
  - Biological material at disposal
  - Funds for the experiment
- Same thing for the technology:
  - Trade-off between money and outcome
  - technology at disposal in the facility lab
- **Ultimately RNA-seq is a hypothesis generating tool: the fist driver of sample size is the budget!**

* Zhao,S., Li,C.-I., Guo,Y., Sheng,Q. and Shyr,Y. (2018) RnaSeqSampleSize: real data based sample size estimation for RNA sequencing. BMC Bioinformatics, 19, 191.

SEQUENCING DEPTH VS BIOLOGICAL REPLICATES

- More sequence or more replication?
- Liu, Y., et al. [1] showed that:
  - for high expressed genes: increasing sequencing depth has little effect on increasing number of DE genes, while biological replicates are clearly more beneficial.
  - for low expressed genes: both are beneficial.



[1]1. Liu,Y., Zhou,J. and White,K.P. (2014) RNA-seq differential expression studies: more sequence or more replication? Bioinformatics, 30, 301–304.

SINGLE END OR PAIRED END?

- PE improve mapping for repetitive regions in the genome
- PE improve accuracy for detection of differential expression for low-expressed genes
- If the species of interest lacks of a reference genome and you need to assembly de novo a transcriptome, PE is a far better choice

Take away message:

- SE is cheaper and sufficient for DE analysis allowing more biological replicates
- If you have specific needs:
  - identify novel transcripts
  - alternative splicing event

goes with PE and greater sequencing depth

STRANDED VS UNSTRANDED LIBRARY

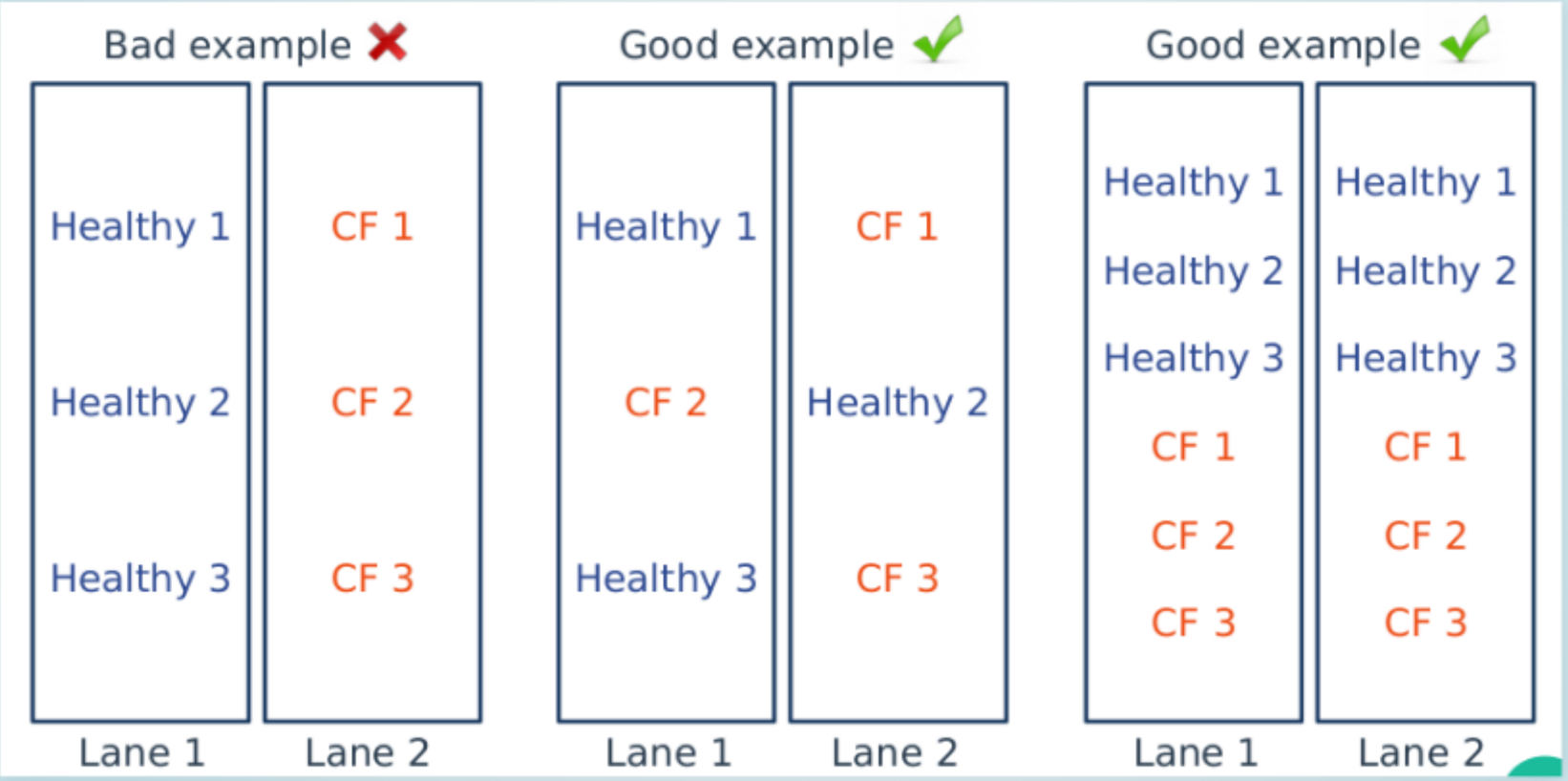
- RNA-seq libraries are generated by the synthesis of double stranded DNA followed by the addiction of sequencing adapters
- Unstranded libraries don’t keep track of the direction of the transcription
- For stranded libraries only one strand from the cDNA synthesis is sequenced retaining directionality allowing the transcripts to be mapped back to the reference genome in a trans-specific manner
- Useful for people studying quantifying sense and anti sense transcription as well as resolving overlapping transcripts in small transcriptomes (bacteria)

"A strand-specific protocol should be used in library preparation to generate the most reliable and accurate profile of expression. Ideally PE reads are also recommended particularly for transcriptome assembly. Whilst SE reads produce a DEG list with around 5% of false positives and false negatives, this method can substantially reduce sequencing cost and this saving could be used to increase the number of biological replicates thereby increasing the power of the experiment. As SE reads, when used in association with gene set enrichment, can generate accurate biological results, this may be a desirable trade-off." *

* Susan M. Corley el al. (2017) *Differentially expressed genes from RNA-Seq and functional enrichment results are affected by the choice of single-end versus paired-end reads and stranded versus non-stranded protocols.* BMC Genomics, 18.

SEQUENCING DESIGN

Avoid any confounding technical effect (day, Laboratory, lane, etc.) to the factor of interest.

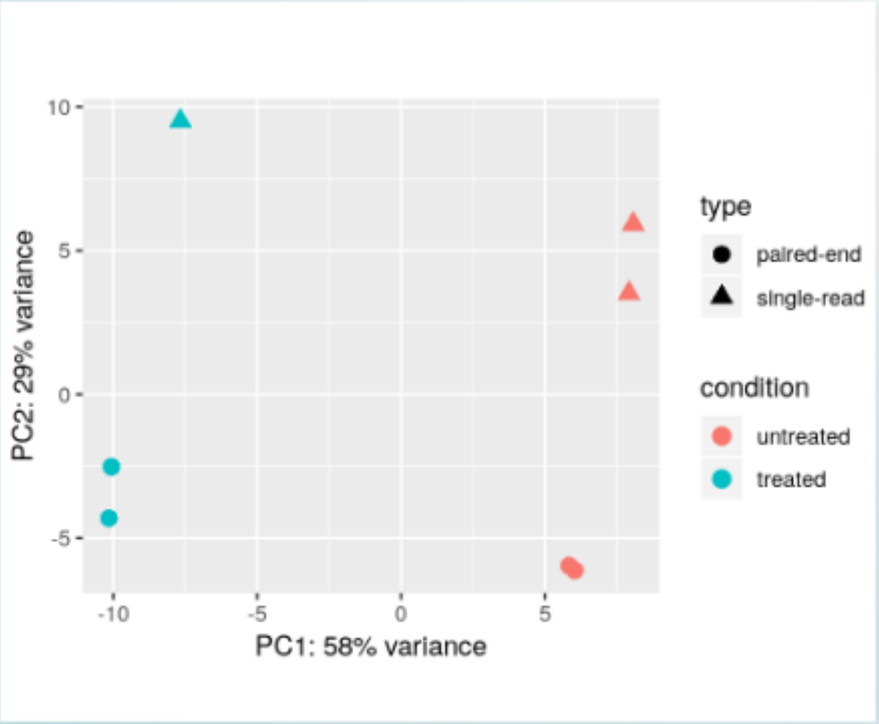


ADVICE

- The biological question must be well defined in order to build an experimental design which will be able to address it.
- Identify all the sources of variability:
  - Change of biological condition
  - Within replicates variability
  - Experimentalist, laboratory or day to day library preparation
  - Library effect (PE vs SE)
  - Sequencing machine, flowcell and lane
  - ...

BATCH EFFECT

- When you are not able to avoid the presence of non-biological factors in your experiment (bad design).
- It can be detected by EDA plots (e.g., PCA, MDS, Clustering).
- It can be incorporated in the statistical model in order to isolate only the effects of interest.



ANALYSIS OF GENE EXPRESSION DATA

BASIC ANALYSIS WORKFLOW

1. Retrieve data from public databases
2. Pre-process raw data (fastq)
3. Alignment or mapping of raw/pre-processed data
4. Summarise reads over genes to generate a count table
5. Determine differentially expressed genes
6. Gene enrichment analysis of lists of DE genes

RETRIEVE DATA FROM ON LINE DATABASES

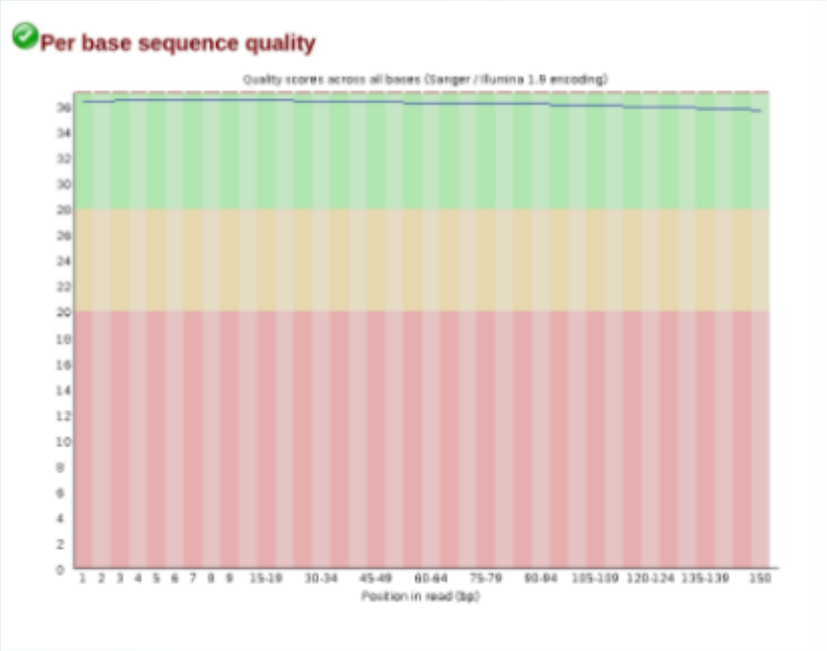
- NCBI Gene Expression Omnibus (*GEO*)
- EMBL-EBI *ArrayExpress* Archive of Functional Genomics Data
- NCBI Sequence Read Archive (*SRA*)
  - SRP - Study
  - SRX - Experiment
  - SRS - Sample
  - SRR - Run

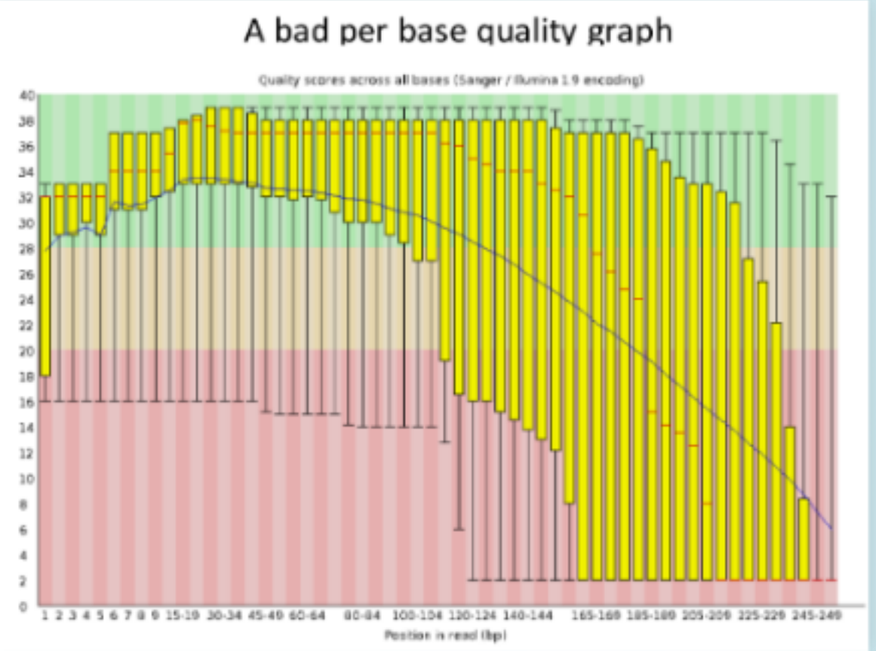
FASTQ

```
@SRR8979823.1 1/1
AACAAGATTCAAGGAACATTTCTCAAGAAGGATGATCCAAAGATAACTGCCTTGATGCAACAAGCAGAATTGCTCAGCTCGCTTGCACTTAAAGTTAATGCAGAGAACACAGACCAGAGCCTTGAAAATGCATGGAAGGTATTCCAAGAT
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@SRR8979823.2 2/1
CCAGCTGGGGTGAGAATGGCTACATCAGGATGGAGCGGAATATGGCTGACGCCTTCTCTGGAAAGTGTGGAATCGCAATGGAGGCCTCGTACCCCATCAAGAATGGAGAAAATCCTTCCAAAACTTACTTGTCTTTTGGAACTGCTGGAG
+
AAFFFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFJJFJFJJJJJJJJJJJJJJJJJJJJJJJJJJFJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
@SRR8979823.3 3/1
AAAGACCCTACCTGTTTTCCTGTCTTAGATGCGGCTGGTCTCCCTTCTGCCTTCATGGATGCTATAATGGATGGCATTCTCTGCTCTGCTGAAGCCATAGCATGCATACCTCAGTGCTTGGATGCCTTATGCCTGAATAACAATGGCCTT
+
AAAFFJJJJJJJJJJJ<JJJJJJAJJJJJJFJFJJJJJJJFJJJJJJJJJJJJFJJFJJF7AJJJ-FJJJFJJAFJJJJJJJJJJJJJFJJJFFAFFFAJFFJJJJJJJJJJJJFJJFFFFF<AJJJJFJFJJFJJJJ-FJFFJAFFJFJJJ
```

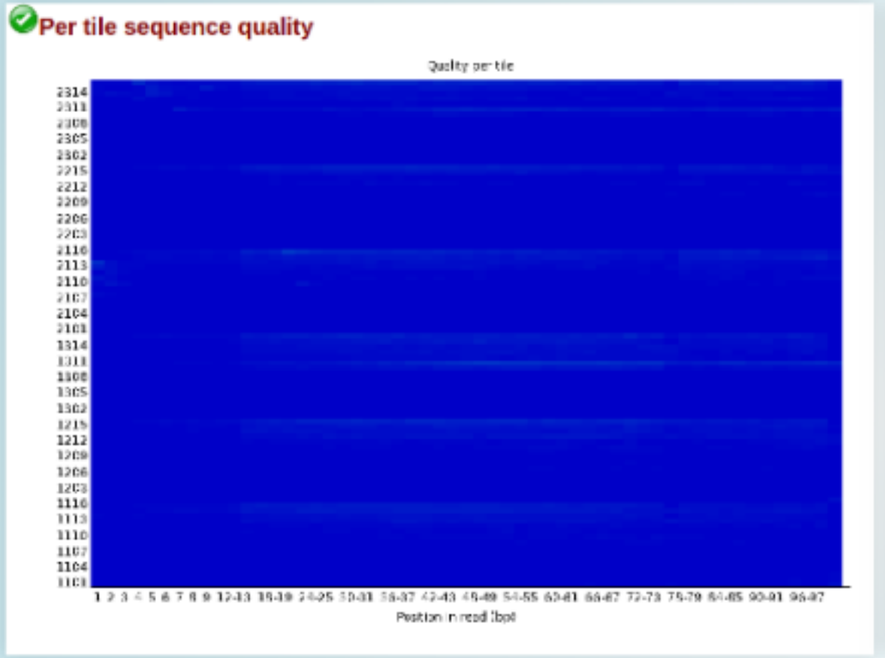
- The quality at each position is captured as *Phred quality score*, a.k.a. as *Q score*, which is an integer value representing the estimated probability of an error, i.e. that the base is incorrect. If p is the probability the base was called incorrectly, then $Q = -10 log_{10} p$ or $p = 10^{-Q/10}$.
- For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.
- You can use either this reference or the Wikipedia page to figure out wich encodings was used for your data. Recent experiments should present a Sanger format (*Phred+33*).

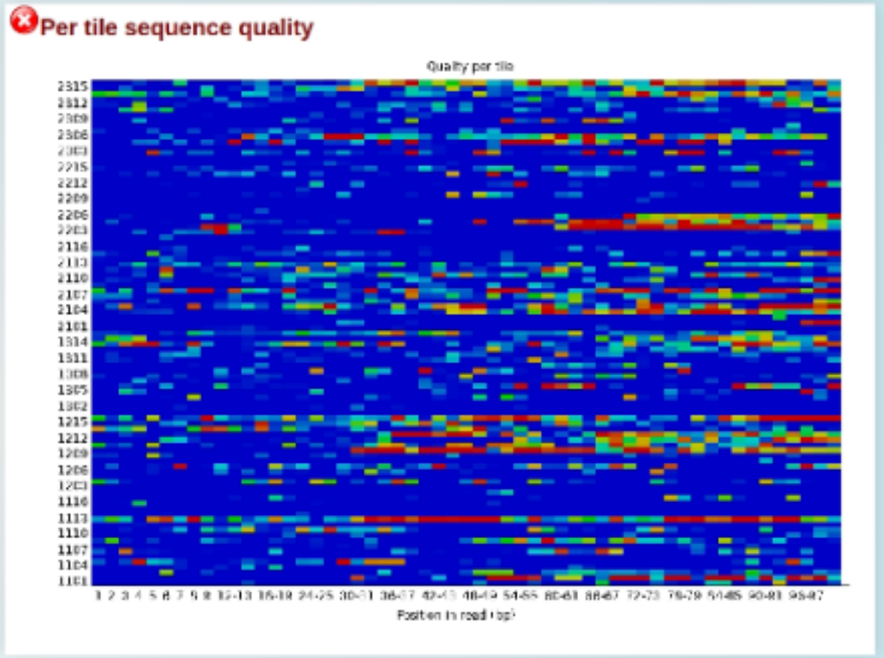
FASTQC - PER BASE SEQUENCE QUALITY





FASTQC - PER TILE SEQUENCE QUALITY





take a look at either this guide, this video or the official documentation for more details.

TRIMMING

- Adapter Trimming
  - Should increase mapping rates
  - Essential for smallRNA
  - May improves de novo assemblies
- Quality Trimming
  - Should increase mapping rates
  - loose information
  - In paired reads if a read is removed, its pair has to be removed as well
- Lots of different tools (cutadapt, trimmomatic, …) and tuning parameters.

TO TRIM OR NOT TO TRIM?

- Liao, Y. and Shi, W. [*] found that adapter sequences can be effectively removed by the read aligner and many low-sequencing-quality bases, which would be removed by read trimming tools, were rescued by the aligner.
- Accuracy of gene expression quantification from using untrimmed reads was found to be comparable to or slightly better than that from using trimmed reads. This study suggests that read trimming is a redundant process in the quantification of RNA-seq expression data.

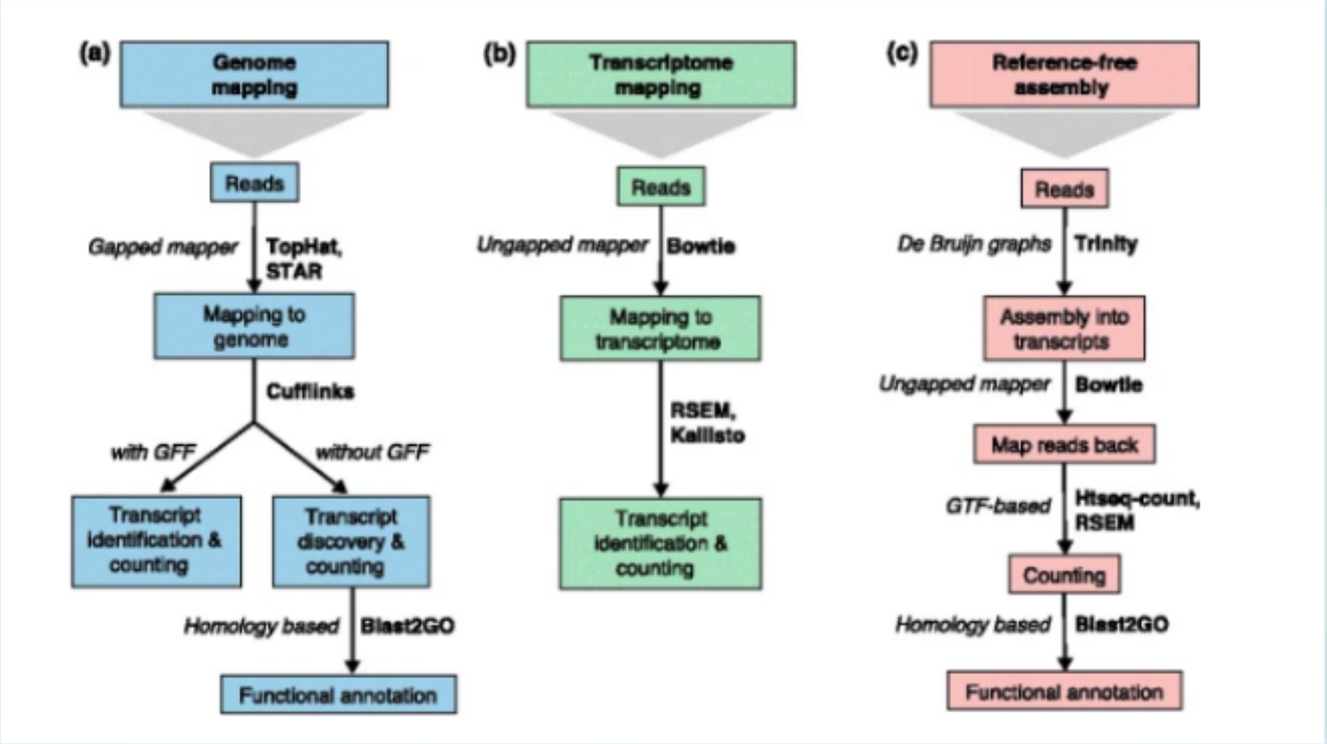
[*] Liao,Y. and Shi,W. (2019) Read trimming is not required for mapping and quantification of RNA-seq reads. bioRxiv, 10.1101/833962.

During the practical you are going to perform the analysis using either trimmed or untrimmed data.

ALIGNMENT AND MAPPING

- The number of reads that align/map to each gene provides a quantification of how many RNA transcripts of that gene were in the sample.
- The reads in the fastq files must first be aligned to a reference genome or transcriptome, or the abundances and estimated counts per transcript can be estimated without alignment. There are several choices for this step, see, for example Baruzzo et al. 2017:
  - alignement methods:
    - STAR
    - Subread/Rsubread
  - pseudo-alignement methods
    - Salmon
    - Kallisto

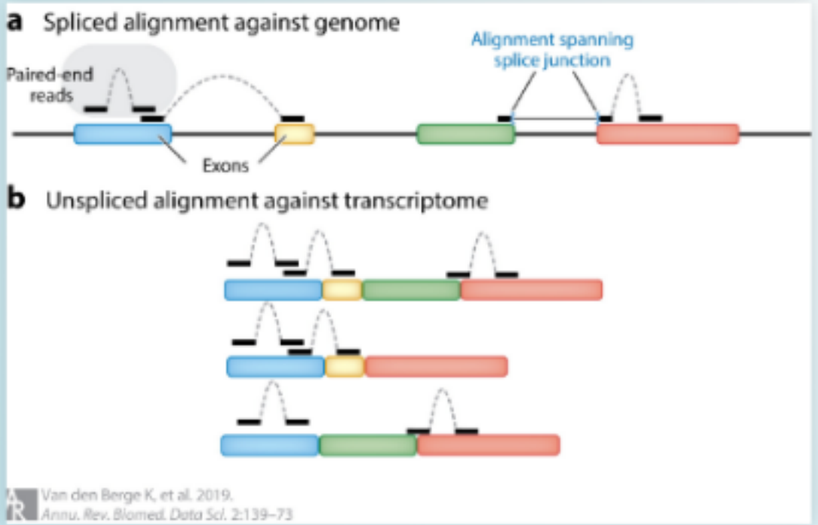
ALIGNMENT AND MAPPING



What is the difference between aligning and mapping?

ALIGNMENT

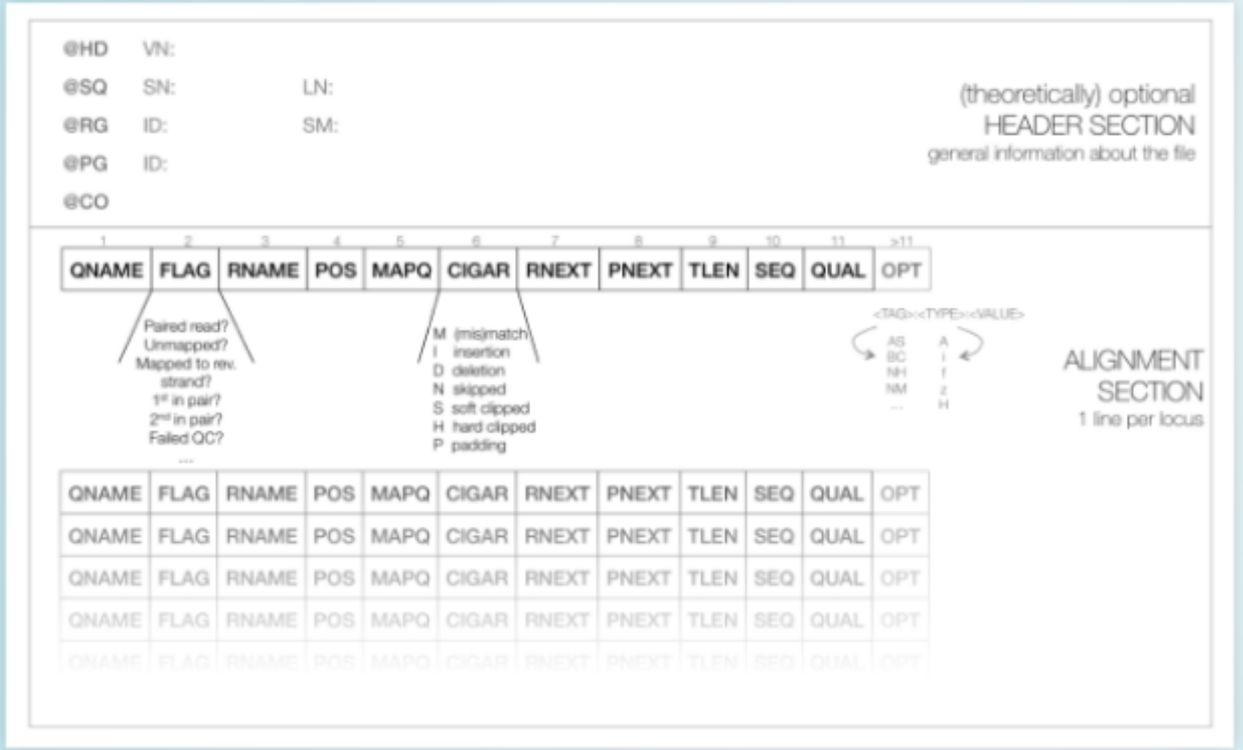
- Build an index -> convert genome FASTA file to a data structure which is faster to search
- Use splice-aware aligners for mapping RNA-Seq reads:
  - reads derived from mature mRNA, i.e., no introns in the sequence
  - a read spans two exons where the reference see an exon and then an intron
  - splice-aware aligners would not try to align reads to introns and will try to indentify downstream exons for the alignment



The output of alignement tools is a BAM file, which is a compressed ("B"inary) form of a Sequence Alignment/Map format (SAM) file.

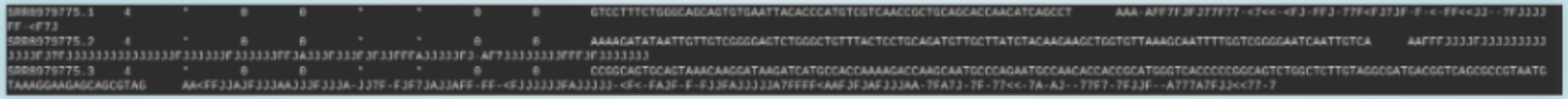
SAM/BAM

- The header section includes information about how the alignment was generated and stored.



SAM/BAM

The alignment section shows a sequence read for each line. For each read, there are 11 mandatory fields that always appear in the same order



Pos.	Field	Example entry	Description	NA value
1	QNAME	Read1	Query template (= read) name (PE: read pair name)	required
2	FLAG	83	Information about the read's mapping properties encoded as bit-wise flags (see next section and Table 6).	required
3	RNAME	chrI	Reference sequence name. This should match a `@SQ` line in the header.	*
4	POS	15364	1-based leftmost mapping position of the first matching base. Set as 0 for an unmapped read without coordinates.	0
5	MAPQ	30	Mapping quality of the alignment. Should be a Phred-scaled posterior probability that the position of the read is incorrect, but the value is completely dependent on the alignment program. Some tools set this to 0 if multiple alignments are found for one read.	0
6	CIGAR	51M	Detailed information about the alignment (see below).	*
7	RNEXT	=	PE reads: reference sequence name of the next read. Set to "=" if both mates are mapped to the same chromosome.	*
8	PNEXT	15535	PE reads: leftmost mapping position of the next read.	0
9	TLEN	232	PE reads: inferred template length (fragment size).	0
10	SEQ	CCA...GGC	The sequence of the aligned read on the forward strand (not including indels).	*
11	QUAL	BBH...1+B	Base quality (same as the quality string in the `FASTQ` format, but always in Sanger format [ASCII+33]).	*
12ff	OPT	NM:i:0	Optional fields (format: `<TAG>:<TYPE>:<VALUE>`; see below).	

For a complete explanation of the format see SAM/BAM and related specifications

`samtools` is a powerful suite of tools designed to interact with SAM and BAM files (Li et al., 2009).

COUNTING READS TO GENES

- Summarization is the process that assign mapped reads to genomic features such as genes, exons, etc.
- For gene-level differential expression the number of reads overlapping annotated exons of each gene can be used as a measure of the expression level of that gene.
- Methods that perform this tasks take as input a set of files that contain read mapping results (SAM/BAM) and an annotation file that includes genomic features (GFF/GTF) and return a read count for each gene in each sample, producing a matrix of read counts (integer).

	SRR5227652_trimmed.bam	SRR5227653_trimmed.bam	SRR5227654_trimmed.bam	SRR5227655_trimmed.bam	SRR5227656_trimmed.bam
VIT_01s0010g00020	193	65	69	19	9
VIT_01s0010g00060	1016	675	356	369	430
VIT_01s0010g00240	2626	2175	770	260	255
VIT_01s0010g00330	142	126	69	196	186
VIT_01s0010g00340	32	32	23	79	93
VIT_01s0010g00360	12	10	4	7	7

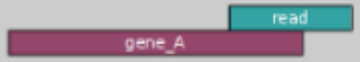
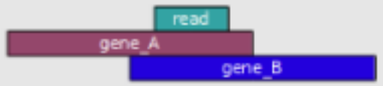
GFF/GTF

```
#!genome-build 12X
#!genome-version 12X
#!genome-date 2011-11
#!genome-build-accession GCA_000003745.2
14 iggp gene 3387 4364 . + . gene_id "VIT_14s0060g00010"; gene_source "iggp"; gene_biotype "protein_coding";
14 iggp transcript 3387 4364 . + . gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; gene_source "iggp"
14 iggp exon 3387 3863 . + . gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; exon_number "1";
14 iggp CDS 3690 3863 . + 0 gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; exon_number "1";
14 iggp start_codon 3690 3692 . + 0 gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; exon_number "1";
14 iggp exon 4269 4364 . + . gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; exon_number "2";
14 iggp CDS 4269 4361 . + 0 gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; exon_number "2";
14 iggp stop_codon 4362 4364 . + 0 gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; exon_number "2";
14 iggp five_prime_u• 3387 3689 . + . gene_id "VIT_14s0060g00010"; transcript_id "VIT_14s0060g00010.t01"; gene_source "iggp"
14 iggp gene 23717 26288 . - . gene_id "VIT_14s0060g00040"; gene_source "iggp"; gene_biotype "protein_coding";
14 iggp transcript 23717 26288 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; gene_source "iggp"
14 iggp exon 26129 26288 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "1";
14 iggp CDS 26129 26288 . - 0 gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "1";
14 iggp start_codon 26286 26288 . - 0 gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "1";
14 iggp exon 25956 26043 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "2";
14 iggp CDS 25956 26043 . - 2 gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "2";
14 iggp exon 25744 25852 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "3";
14 iggp CDS 25744 25852 . - 1 gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "3";
14 iggp exon 24810 25556 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "4";
14 iggp CDS 24810 25556 . - 0 gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "4";
14 iggp exon 23717 23978 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "5";
14 iggp CDS 23919 23978 . - 0 gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; exon_number "5";
14 iggp stop_codon 23916 23918 . - 0 gene_id "VIT_14s0060g00040"; tranecript_id "VIT_14s0060g00040.t01"; exon_number "5";
14 iggp three_prime_• 23717 23915 . - . gene_id "VIT_14s0060g00040"; transcript_id "VIT_14s0060g00040.t01"; gene_source "iggp"
```

Position index	Position name	Description
1	sequence	The name of the sequence where the feature is located.
2	source	Keyword identifying the source of the feature, like a program (e.g. Augustus or RepeatMasker) or an organization (like TAIR).
3	feature	The feature type name, like "gene" or "exon". In a well structured GFF file, all the children features always follow their parents in a single block (so all exons of a transcript are put after their parent "transcript" feature line and before any other parent transcript line). In GFF3, all features and their relationships should be compatible with the standards released by the Sequence Ontology Project.
4	start	Genomic start of the feature, with a **1-base offset**. This is in contrast with other 0-offset half-open sequence formats, like BED.
5	end	Genomic end of the feature, with a **1-base offset**. This is the same end coordinate as it is in 0-offset half-open sequence formats, like BED.[citation needed]
6	score	Numeric value that generally indicates the confidence of the source in the annotated feature. A value of "." (a dot) is used to define a null value.
7	strand	Single character that indicates the strand of the feature; it can assume the values of "+" (positive, or 5'->3'), "-" (negative, or 3'->5'), "." (undetermined).
8	phase	phase of CDS features; it can be either one of 0, 1, 2 (for CDS features) or "." (for everything else). See the section below for a detailed explanation.
9	attributes	All the other information pertaining to this feature. The format, structure and content of this field is the one which varies the most between the three competing file formats.

here you can find more info about the two formats.

HTSEQ COUNT

	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 (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

See Counting reads in features with htseq-count for more details.

TWO PATHS

- Transformations and Exploratory Data Analysis (EDA)
- Differential Expression Analysis

ESPLORATORY DATA ANALYSIS (EDA)

- Transformations
  - Within sample transformations
  - Transformations for EDA
- Visualizations
  - Scatterplot
  - PCA/MDS
  - Clustering
  - MA-plot (Mean-Difference MD-plot)

WITHIN SAMPLE TRANSFORMATIONS

This kind of transformations try to deal with bias present in comparing different genes within a sample, they take care of the facts that:
- Sequencing runs with more depth will have more reads mapping to each gene (“Million part”)
- Longer genes will have more reads mapping to them (“Kilobase part”)

- Counts per million (CPM): counts are divided by the library size (in millions)

$$CPM_i = \frac{\frac{X_i}{N}}{10^6} = \frac{X_i}{N}.10^6$$

- Reads/fragments per kilobase per million (RPKM/FPKM): counts are divided by the transcript length (kb) times the total number of millions of mapped reads

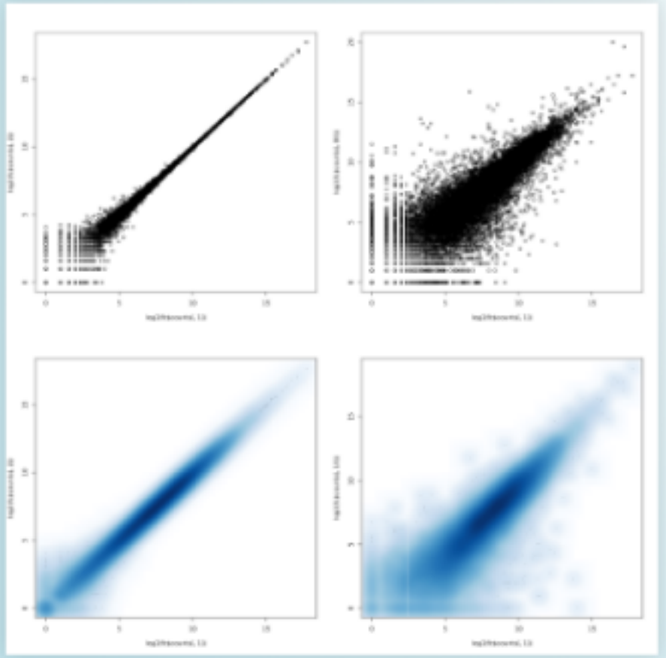
$$FPKM_i = \frac{X_i}{\left(\frac{\bar{l}_i}{10^3}\right)\left(\frac{N}{10^6}\right)} = \frac{X_i}{\tilde{l}_i N}.10^9$$

TRANSFORMATIONS FOR EDA

- Common methods for exploratory analysis of multidimensional data work best for data with the same range of variance at different ranges of the mean values
- When the expected amount of variance is approximately the same across different mean values, the data is said to be *homoskedastic*
- In RNA-seq raw counts the variance grows with the mean
- Transformations for count data that stabilize the variance across the mean are:
  - regularized-logarithm transformation (*rlog*)
  - variance stabilizing transformation (*vst*)
- *rlog/vst*-transformed data become approximately homoskedastic, and can be used directly for computing distances between samples, allowing Clustering and PCA
- Used just for EDA **NOT** for differential testing

EDA - SCATTERPLOTS

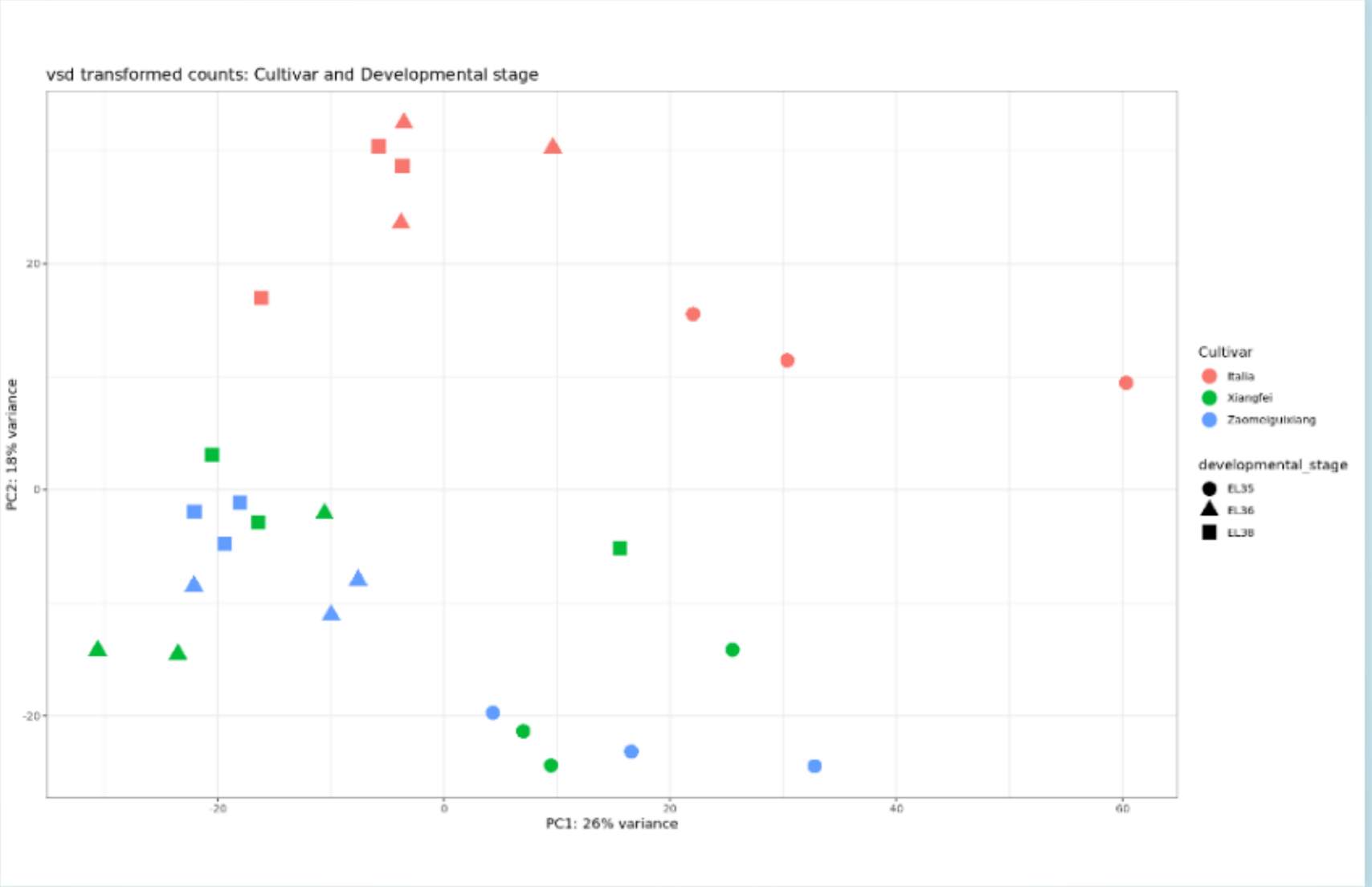
- The scatterplot is the most frequently used plot when you want to compare samples and understand the nature of relationship between two variables.
- Scatterplots are useful when you are looking for abnormal behavior: the samples can be too different between them but also they can be too similar!



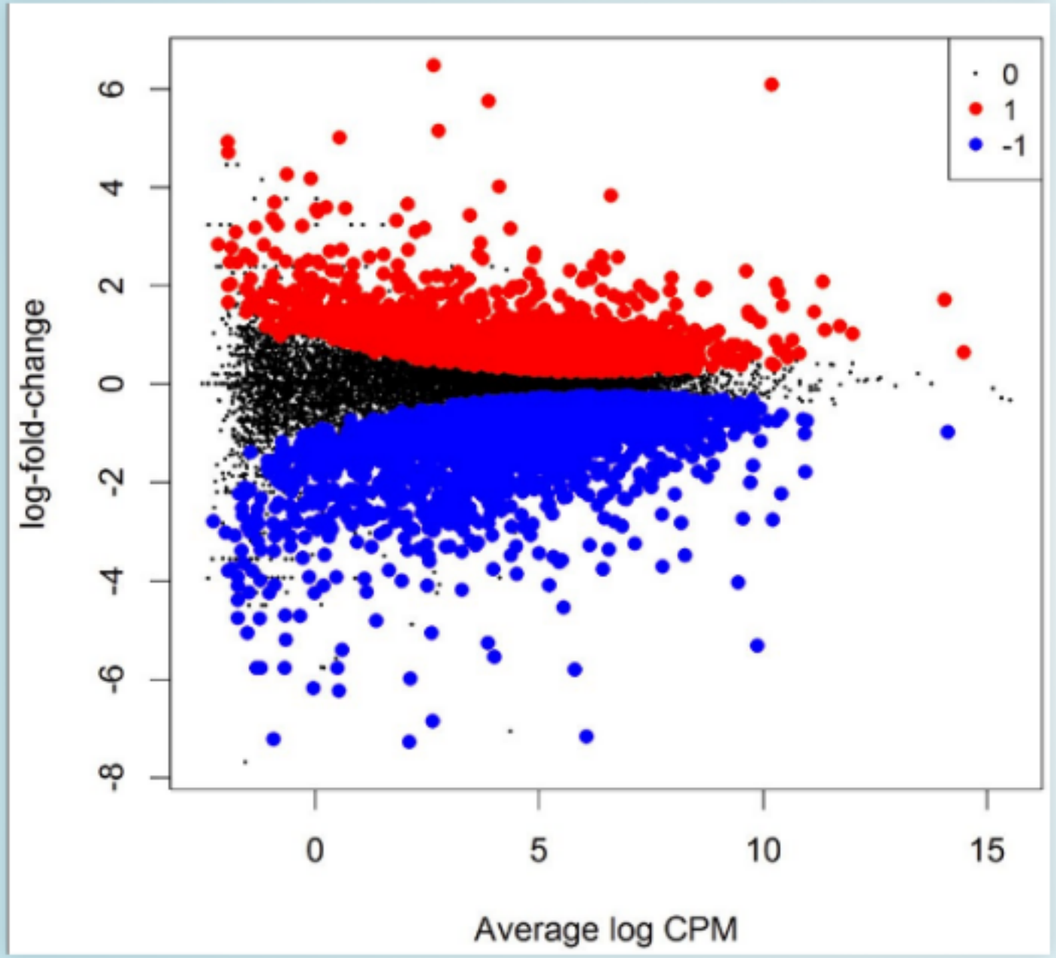
EDA - PCA

- Principal component analysis (PCA) allows us to summarize the information in a data set containing observations described by multiple inter-correlated quantitative variables
- PCA is used to extract the important information from a multivariate data table and to express this information as a set of few new variables called principal components
- These new variables correspond to a linear combination of the original features
- The number of principal components is less than or equal to the number of original variables
- The information in a given data set corresponds to the total variation it contains: the goal of PCA is to identify directions along which the variation in the data is maximal
- PCA reduces the dimensionality of a multivariate data to two or three principal components, that can be visualized graphically
- **Counts must be transformed (made homoskedastic) before building the PCA.**

EDA - PCA


vsd transformed counts: Cultivar and Developmental stage
PC2: 18% variance
PC1: 26% variance
20
0
-20
-20
0
20
40
60
Cultivar
Italia
Xiangfei
Zaomeiguiliang
developmental_stage
EL35
EL36
EL38

MEAN-DIFFERENCE PLOT



FILTERING

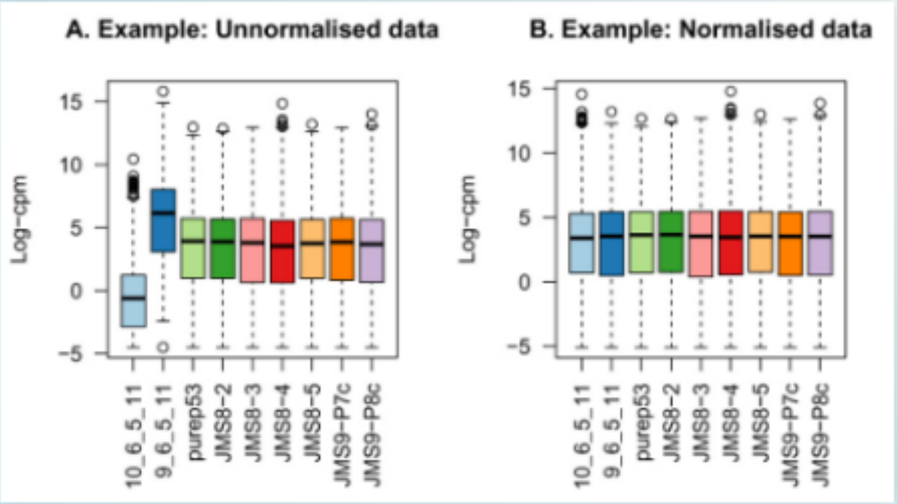
- Genes that have very low counts across all the libraries should be removed prior to downstream analysis
- From a *biological point of view*, a gene must be expressed at some minimal level before it is likely to be translated into a protein or to be considered biologically important
- From a *statistical point of view*, genes with consistently low counts are very unlikely be assessed as significantly DE because low counts do not provide enough statistical evidence for a reliable judgement to be made
- Filtering, as side effect, allows to better deal with the multiple testing issue related to the statistical tests used to determine differentially expressed genes

FILTERING

- Remove all genes having zero reads in all samples
- Filter genes that are less then n reads counts across a certain number of samples, e.g., require at least 1 CPM in at least 3 samples to keep (with 3 replicates for condition)
- A CPM value of 1 means that a gene is *expressed* if it has at least 20 counts in the sample with library size ~ 20 million
- Alternatively: filter based on minimum variance across all samples, so if a gene isn't changing (constant expression) its not interesting and no need to be tested

NORMALIZATION

- Identify and correct for systematic technical bias and make the counts comparable between samples
- Mandatory for any kind of *-omics* (data) analysis
- Normalization assumptions tailored to gene expression data (both MA and RNA-seq):
  1. The majority of the genes is not differentially expressed between contrasts
  2. As many down- as up-regulated genes



COMPOSITION BIAS

"Estimated normalization factors should ensure that a gene with the same expression level in two samples is not detected as differentially expressed. Imagine we have a sequencing experiment comparing two RNA populations, A and B. Suppose every gene that is expressed in B is expressed in A with the same number of transcripts. However, assume that sample A also contains a set of genes equal in number and expression that are not expressed in B. Thus, sample A has twice as many total expressed genes as sample B, that is, its RNA production is twice the size of sample B. Suppose that each sample is then sequenced to the same depth. Without any additional adjustment, a gene expressed in both samples will have, on average, half the number of reads from sample A, since the reads are spread over twice as many genes. Therefore, the correct normalization would adjust sample A by a factor of 2."

Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11, R25.

NORMALIZATION BY TRIMMED MEAN OF M-VALUES (TMM)

- TMM Normalization calculates a set of normalization factors, one for each sample, to eliminate composition biases between libraries
- The product of these factors and the library sizes defines the effective library size, which replaces the original library size in all downstream analyses
- The normalization factors of all the libraries multiply to unity.
- A normalization factor below one indicates that a small number of high count genes are monopolizing the sequencing, causing the counts for other genes to be lower than would be usual given the library size
- As a result, the effective library size will be scaled down for that sample

DIFFERENTIAL EXPRESSION

MEMENTO MORI

To consult the statistician bioinformatician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.

Sir Ronald Fisher (1890-1962)

STATISTICAL TESTS

- Make inference about the population beyond the data
- Check if statistics (e.g. averages) between groups of interests are reliably different from each other
- Measure the difference between the groups and compare it with the difference within the group
- Provide a confidence measure (p-value, confidence interval) for the evaluated statistics

STATISTICAL TESTS FOR DE

- For each gene, is the mean expression level under one condition significantly different from the mean expression level under a different condition?
- Use negative binomial distribution to **model the counts** directly:
  - `edgeR`
  - `DESeq2`
  - ...
- **Transform the counts** to be normally distributed using precision weights and then use normal-based methods
  - `limma voom`
- **Statistical tests rely on approximations: e.g., for gene expression profiling assume that the majority of the transcriptome is unchanged between the two conditions. If this assumption is not met by the data, results are not reliable!**

WHY LIMMA + VOOM?

- Same procedure for the analysis of both microarray and RNA-Seq data
- Designed and developed specifically for experiments with few biological replicates
- Accurate, e.g. see Soneson, C., Delorenzi (2013)
- Robust and reliable (It has been developed since 2003)
- Voom transformation was specifically designed for dealing with RNA-Seq data (since 2011)

DOWNSTREAM ANALYSIS WORKFLOW

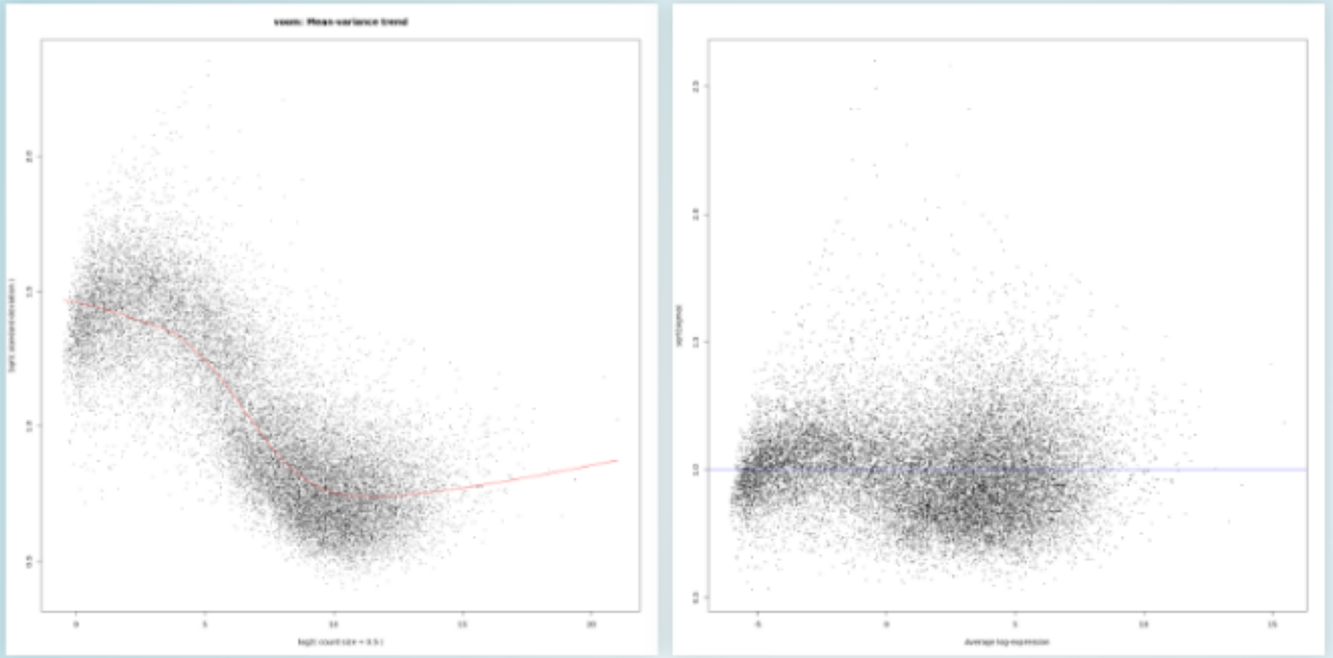
- Experimental design (design matrix and contrast matrix)
- Trimmed Mean of M-values (TMM) normalization
- Voom transformation
- Statistical test - linear modelling (limma)

VOOM TRANSFORMATION *

- Unlike methods that model counts using a negative binomial distribution, limma performs linear modelling on the $log_2(CPM)$ transformed values assumed to be normally distributed and the relationship between mean and variance is taken care using precision weights calculated by the voom function
- The read counts of gene g in sample s do not follow a normal distribution however the ($log_2(CPM)$) transformed response variable converges quickly to normality
- Voom models the log counts per million and fits a loess trend line (robust against highly variable genes) to the scatterplot of variance vs. mean to create weight that are then fed into a standard linear model analysis

* Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology, 15, R29

QC - VOOM PLOTS



- Left image: means (x-axis) and variances (y-axis) of each gene are plotted to show the dependence between the two before voom is applied to the data
- Right image: how the trend is removed after voom precision weights are applied to the data: plots log2 residual standard deviations against mean $log_2(CPM)$ values
- Moreover, the voom-plot provides a visual check on the level of filtering performed upstream. If filtering of lowly-expressed genes is insufficient, a drop in variance levels can be observed at the low end of the expression scale due to very small counts

STATISTICAL TESTING

- The Null hypothesis (H_0) is the mean (or an other statistics) of gene expression for a specific comparison does not change
- Alternative hypothesis (H_1) is that the gene expression can be higher or lower
- We repeat the test for all the genes in the specific samples
- Multiple comparisons problem (Look-elsewhere effect) arises


JELLY BEANS CAUSE ACNE.
SCIENTISTS! INVESTIGATE!
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE. (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
News
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!

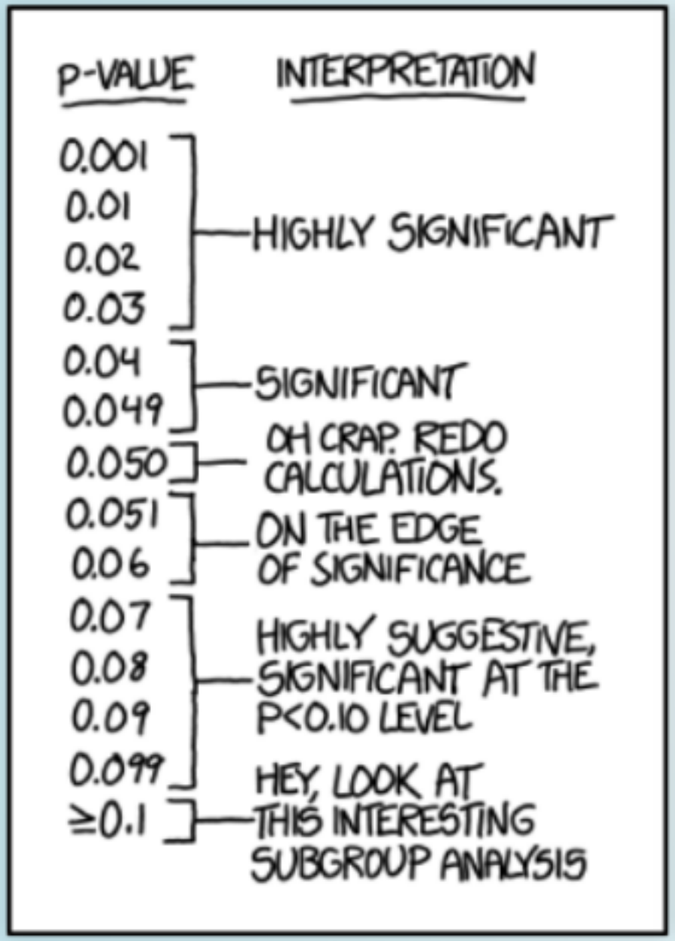
FAMILYWISE ERROR RATE (FWE) AND TYPE I ERROR

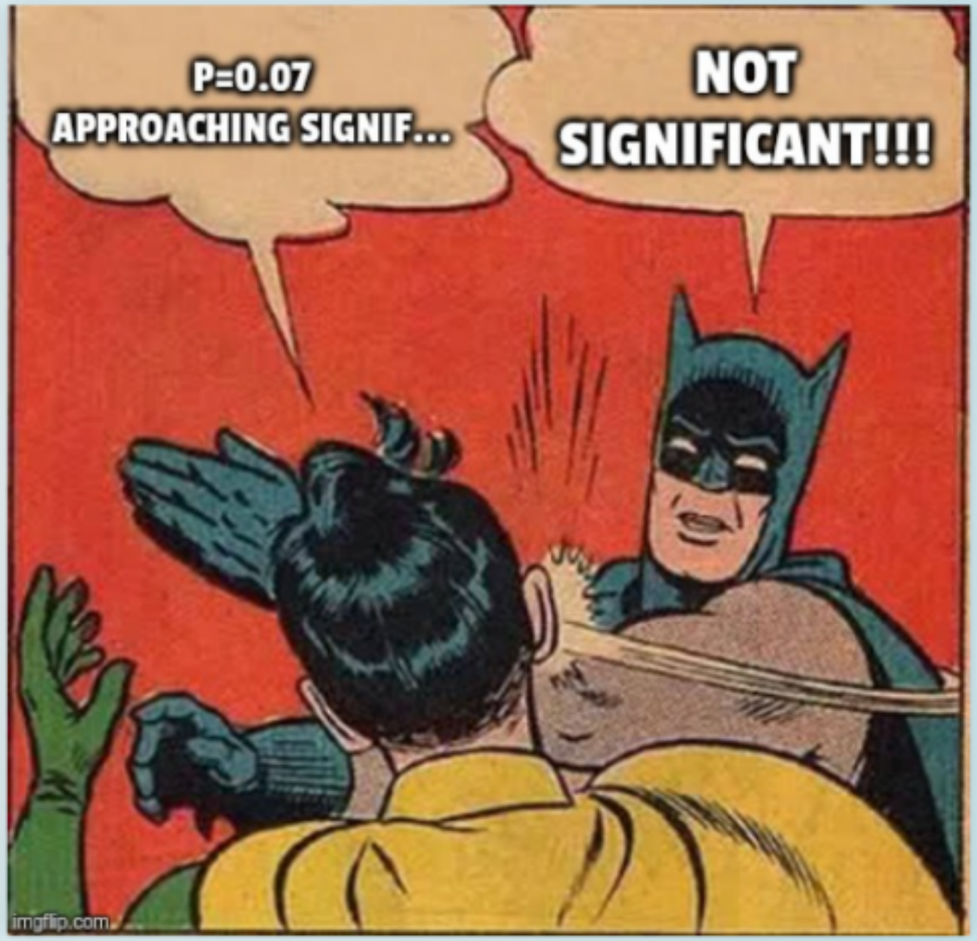
- **Type I error** = is the incorrect rejection of the H_0, when the null hypothesis is in fact *true* in the population
- The FWE rate is the probability of a coming to at least one false conclusion in a series of hypothesis tests, i.e., the probability of making at least one type I error
- The common practice is to test a null hypothesis with a maximum type I error of 5% ($p - value \leq 0.05$)
- If you apply multiple statistical test on the same data the formula for FWE is: $\alpha_{FW} \leq 1 - (1 - \alpha_{IT})^c$ with
  α_{IT} = alpha level for an individual test (e.g. .05)
  c = Number of comparisons
- For example with an α level of 5% and a series of ten tests, the FWER is: $FWE = \leq 1 - (1 - .05)^{10} = .401$ This means that the probability of a type I error is just over 40% (high considering only ten tests were performed)

MULTIPLE TESTING FOR GENE EXPRESSION

- If you test $30,000$ genes for differential gene expression, and you use a significance cut off of $p < 0.05$, then you should expect to call approximately 1500 (i.e., 5% of 30000) genes to show differential expression just by chance
- If your list of differentially expressed genes at $p < 0.05$ contains 1500 genes, then
  - either there are no genes differentially expressed between the two conditions or
  - your experiment does not have a sufficient number of replicates for recognize true DE genes (underpowered)
- Taking care of this issue by
  - Non-specific filtering
  - Multiple test correction (Bonferroni, B-H adjusted P-Value, etc.)

WHY A $p - value \leq 0.05$?





DESIGN AND CONTRAST MATRICES

- The *design matrix* links each group to the samples that belong to it.

	Xiangfei.EL35	Xiangfei.EL36	Xiangfei.EL38	Italia.EL35	Italia.EL36	Italia.EL38	Zaomeiguixiang.EL35	Zaomeiguixiang.EL36	Zaomeiguixiang.EL38
1	1	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0	0
7	0	0	1	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	0
9	0	0	1	0	0	0	0	0	0
10	0	0	0	1	0	0	0	0	0
11	0	0	0	1	0	0	0	0	0
12	0	0	0	1	0	0	0	0	0

- The *contrast matrix* specifies the comparisons of interest between groups (**group-means parametrization**)

	Xiangfei.EL35vsXiangfei.EL36	Xiangfei.EL35vsXiangfei.EL38	Xiangfei.EL36vsXiangfei.EL38	Italia.EL35vsItalia.EL36	Italia.EL35vsItalia.EL38
Xiangfei.EL35	1	1	0	0	0
Xiangfei.EL36	-1	0	1	0	0
Xiangfei.EL38	0	-1	-1	0	0
Italia.EL35	0	0	0	1	1
Italia.EL36	0	0	0	-1	0
Italia.EL38	0	0	0	0	-1
Zaomeiguixiang.EL35	0	0	0	0	0
Zaomeiguixiang.EL36	0	0	0	0	0
Zaomeiguixiang.EL38	0	0	0	0	0

LIMMA FITTING

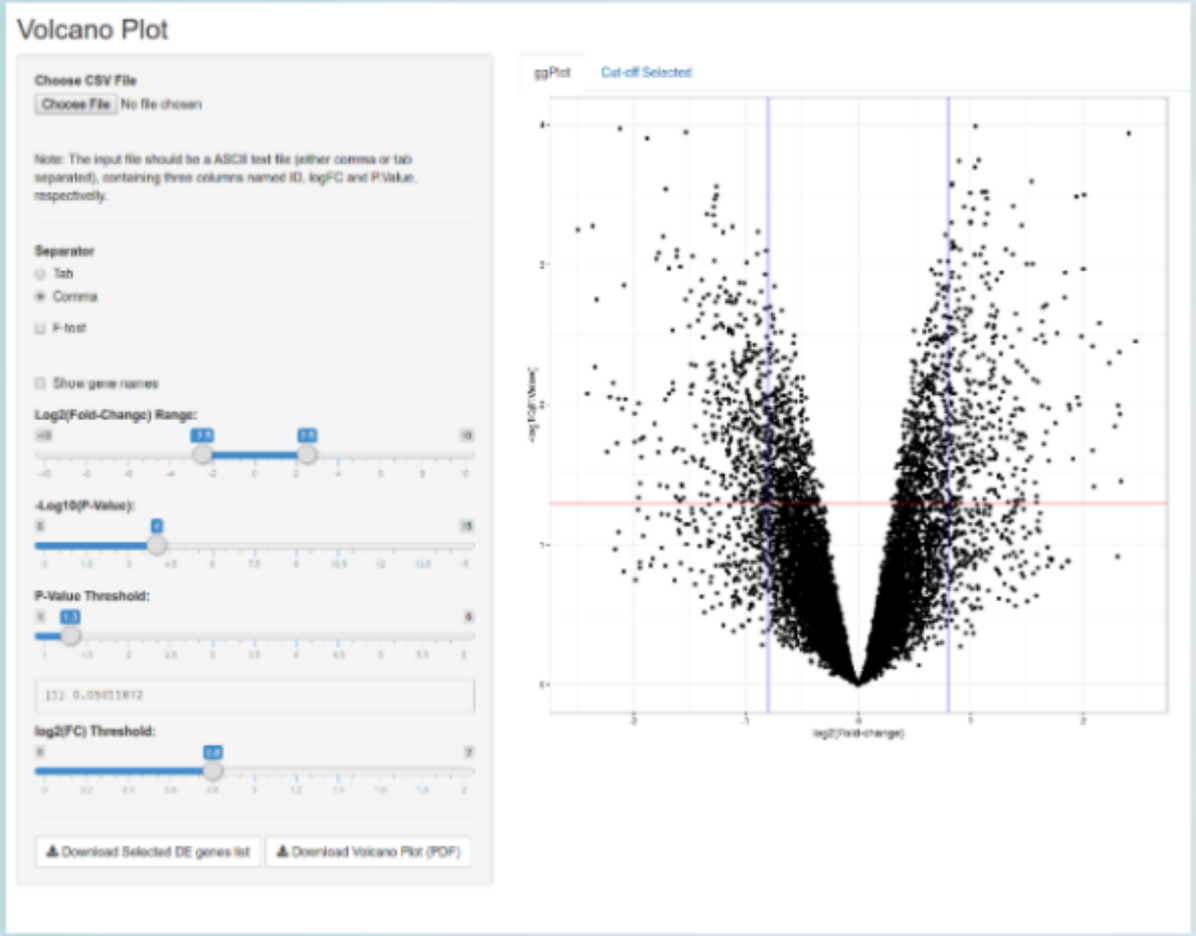
- Fitting a separate linear model to the expression values for each gene
- ebayes moderated t-test which borrow information across all genes to obtain precise estimate of gene-wise variability
- The method: voom transformation + linear modeling aims to remove the dependency of the variance from the mean expression level and apply a statistical test (moderated t-test) robust and accurate tailored to experiments with few biological replicates

LIMMA OUTPUT

	ID	logFC	AveExpr	t	P.Value	adj.P.Val	B
VIT_02s0025g04330	VIT_02s0025g04330	-5.967939	3.903416	-26.01021	1.311512e-11	2.419346e-07	16.32686
VIT_06s0004g05460	VIT_06s0004g05460	-3.438538	6.257815	-21.98922	8.794390e-11	3.955772e-07	15.26717
VIT_19s0093g00550	VIT_19s0093g00550	-6.354018	3.811761	-24.29892	2.839847e-11	2.619333e-07	15.26463
VIT_13s0019g02200	VIT_13s0019g02200	-2.411133	7.400294	-21.56152	1.097765e-10	3.955772e-07	15.09002
VIT_04s0044g00130	VIT_04s0044g00130	-3.106525	5.974479	-20.83956	1.611715e-10	3.955772e-07	14.71222
VIT_01s0026g00220	VIT_01s0026g00220	-2.121529	7.314954	-20.77744	1.666848e-10	3.955772e-07	14.68831
VIT_17s0000g00430	VIT_17s0000g00430	-3.373297	7.214778	-20.72442	1.715519e-10	3.955772e-07	14.64948
VIT_03s0038g01380	VIT_03s0038g01380	-3.059084	6.422147	-20.45446	1.988447e-10	3.992256e-07	14.51462
VIT_14s0171g00360	VIT_14s0171g00360	-4.849779	4.405474	-21.29683	1.261908e-10	3.955772e-07	14.45467
VIT_02s0236g00130	VIT_02s0236g00130	-1.799988	5.688626	-20.14180	2.364765e-10	3.992256e-07	14.34860

logFC	log2 fold change between compared groups
AveExpr	average log2 expression (across complete dataset)
t	moderated t-statistics
P.Value	raw p-value
adj.P.Val	adjusted p-value for multiple testing issue (FDR correction by default)
B	B-statistics, log-odds that the gene is differentially expressed

VISUALIZE DE ANALYSIS RESULTS



ShinyVolcanoPlot WebApp

CAVEAT

Results of any analysis must be evaluated in the context of the biological knowledge.
J. Quackenbush

- **Statistical significance does not imply biological significance!**
- If your results are suggesting expression changes drammatically different from everything you would have expected base on literature and previous knowledge, be very cautious!
- The **exploratory** nature of gene expression experiments requires external validation:
  - More experiments
  - Orthogonal validation with different methods (e.g. western blot, RT-qPCR - one gene approach)
  - Combine results from multiple studies for increase power: meta-analysis

BEYOND GENE LISTS

ENRICHMENT ANALYSIS

- In a differential expression analysis hundreds of genes are found differentially expressed between contasts. Manually inspecting such a large list of genes would be painful
- Functional Enrichment allows the inspection of functional terms that appear associated to the given set of differentially expressed genes more often than expected by chance
- The functional terms usually are associated to multiple genes. Thus, genes can be grouped into sets by shared functional terms.
- You have a list of genes DE expressed from an analysis you want to check if your genes are enriched for a particular set
- Count the number of genes in your DE list belonging to that set and compare them to the proportion of genes which belong to that set from the universe (e.g. your genome) using some statistical test
- It is important to have an agreed upon controlled vocabulary on the list of terms used to describe the functions of genes

ONTOLOGY

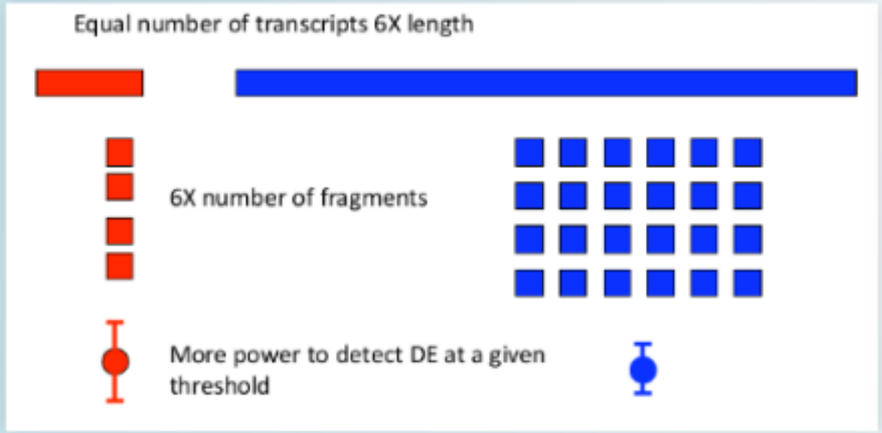
- An ontology is a specification of a conceptualization, a hierarchical mapping of concepts within a given frame of reference.
- An ontology is a restricted structured vocabulary of terms that represent domain knowledge.
- An ontology specifies a vocabulary that can be used to exchange queries and assertions in a consistent way.

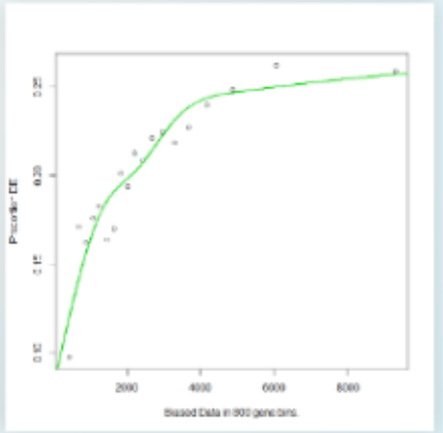
GENE ONTOLOGY

- It is a controlled vocabulary of terms and descriptions for basic biological functions related to genes and gene product (RNA, proteins).
- It was designed and created by observing that similar genes in different organisms conserved a similar function.
- The Gene Ontology (GO) consortium, which provides a central repository for all organisms, produces three independent ontologies for gene products:
  - *Molecular Function (MF)* describes the function carried out by the gene, such as binding or catalysis.
  - *Biological Process (BP)* a set of molecular functions, with a defined beginning and end, makes up a biological process. This describes biological phenomenon like DNA replication.
  - *Cellular Component (CC)* describes where in a cell a gene acts, what cellular unit the gene is part of.

LENGTH BIAS IN RNA-SEQ

- For genes of the same expression level longer transcripts will have more reads
- Therefore there is more information for longer transcripts than shorter ones
- Longer genes have higher power to detect DE at a given threshold





- The x-axis of the pwf plot (on the right) are binned lengths of genes and the y-axis is the ratio of differentially detected genes.
- We can see a clear bias in the detection of differential expression with longer genes.

GOSEQ

- GOseq [*] is a method to conduct Gene Ontology (GO) analysis tailored to RNA-seq data as it accounts for the gene length bias in detection of over-representation (Young et al. 2010)
- Three steps procedure:
  1. determine which genes are differentially expressed
  2. define a probability weighting function (pwf)
  3. generate many random samples to produce a null distribution in order to calculate significance of a category
     - Results:
       - Categories with short genes get a higher rank
       - Categories with long genes get a lower rank

[*] Young, M. D., Wakefield, M. J., Smyth, G. K., Oshlack, A. (2010) Gene ontology analysis for RNA-seq: accounting for selection bias Genome Biology Date

ON LINE RESOURCES

- Bioconductor
- F1000 Bioconductor Gateway
- Biostars
- Seqanswers

Respect Netiquette rules!

ACKNOWLEDGEMENTS

I'd like to say thank you to José Tomás Matus for having me here and of course to my colleague Marco Moretto to keep Colombos/Vespucci dream alive.

Thank you for your attention!