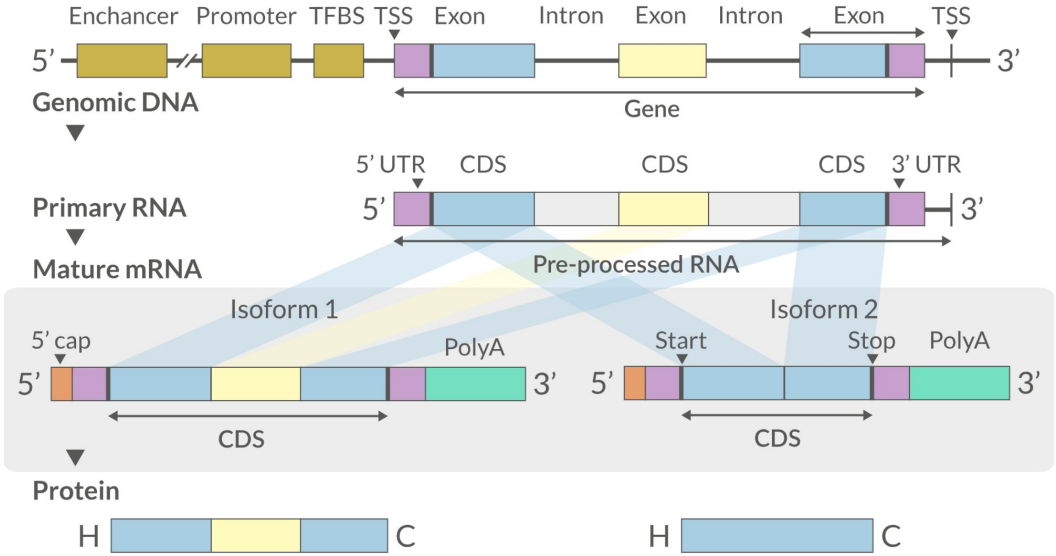
WM7

Computational and Systems Biology

RNA sequencing data analysis

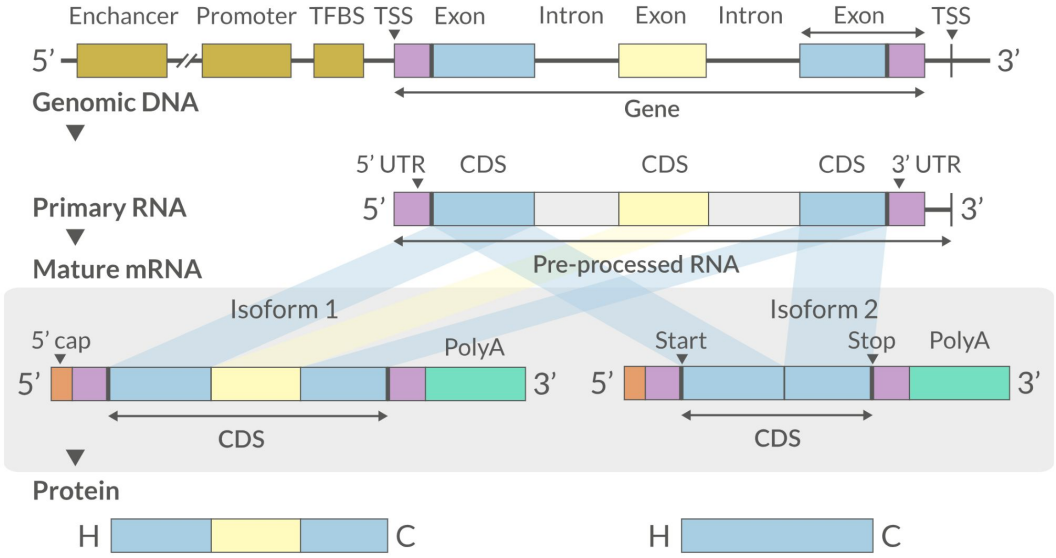
Dietmar Rieder

What is RNA?



- The transcriptome is spatially and temporally dynamic
- Data comes from functional units (coding regions)
- Only a tiny fraction of the genome

What is RNA?



RNA-sequencing can be used to measure RNA produced by cells

What is RNA-sequencing used for

RNA-seq can be used to carry out accurate analysis of:

- differential gene expression (DGEA)
- pathway analysis
- Whole gene coexpression network analysis (WGCNA)
- alternative splicing (AS)
- novel transcript reconstruction and annotation
- allele-specific expression of variants
- RNA editing and other modifications
- ...

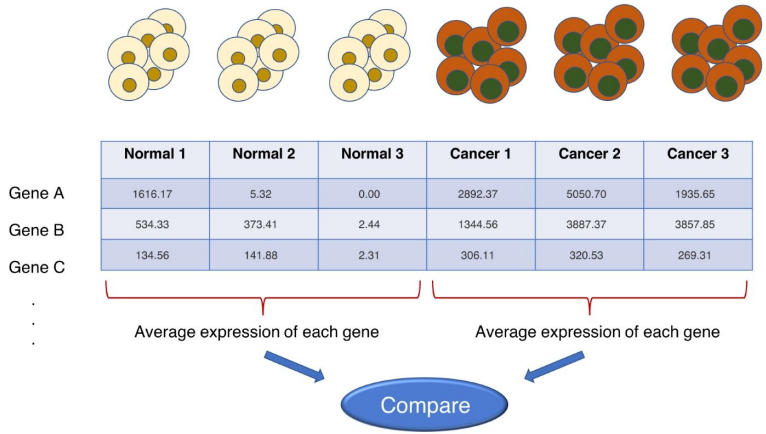
We will focus on DGEA

Differential gene expression analysis - DGEA

The identification of genes that are expressed in significantly different quantities in distinct groups of samples:

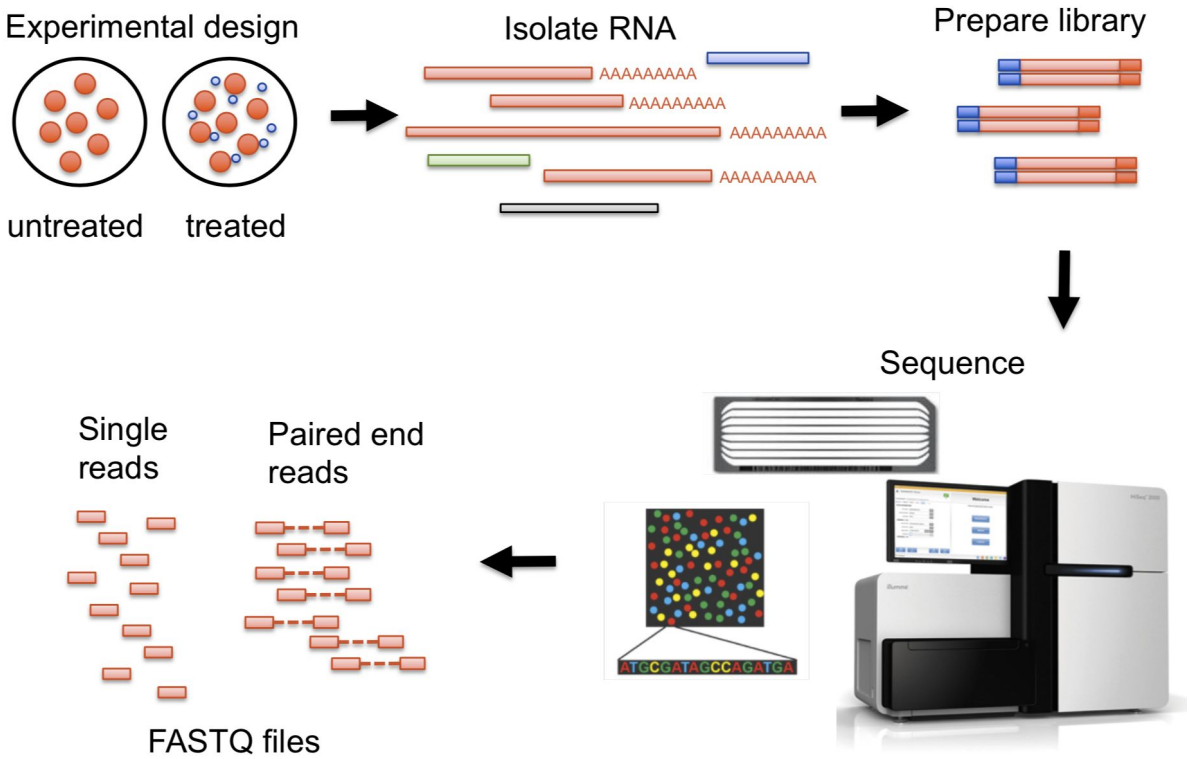
e.g.

- biological conditions (drug-treated vs. controls)
- diseased vs. healthy individuals
- different tissues
- different stages of development
- …

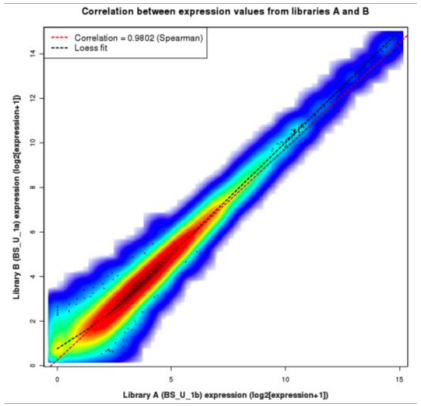


Typically univariate analysis (one gene at a time) – even though we know that genes are not independent.

RNA sequencing overview



RNA sequencing considerations



Biological replicates:

- Assess variability between individuals / “normal” biological variation
- Necessary for drawing conclusions about biology

Intuitively, the variation **between** the groups that you want to compare should be large compared to the variation **within** each group to be able to say that we have differential expression.

The more biological replicates, **the better** you can estimate the variation.

But **how many replicates** are needed?

Depends:

- Homogeneous **cell lines**, **inbred mice** etc: maybe **3** samples / group enough
- **Clinical** case-control studies on patients: can need a **dozen**, **hundreds** or **thousands**
- use a tool such as RNASeqPower to calculate formally

RNA sequencing considerations

Sequencing depth:

- Number of reads per library
- Need enough reads for detecting also weakly expressed genes
- Some analyses are impossible without sufficient sequencing depth
    - Alternative splicing
    - novel transcripts
    - allele-specific expression …

But if you are doing standard **differential gene expression** analysis the benefit from using **more replicates** is much higher than from increasing read depth.

You may be unable to do alternative splicing analysis and/or novel transcript annotation

... BUT if you weren't planning on validating them anyway, they may not add that much to your experiment ...

Replicates vs read depth

Example:

Effect size (fold change)	Replicates @ 20 x coverage		
	3	6	12
1.25 (+/- 25%)	0.26	0.47	0.76
1.75 (+/- 75%)	0.45	0.74	0.95
2 (+/- 100%)	0.62	0.89	0.99

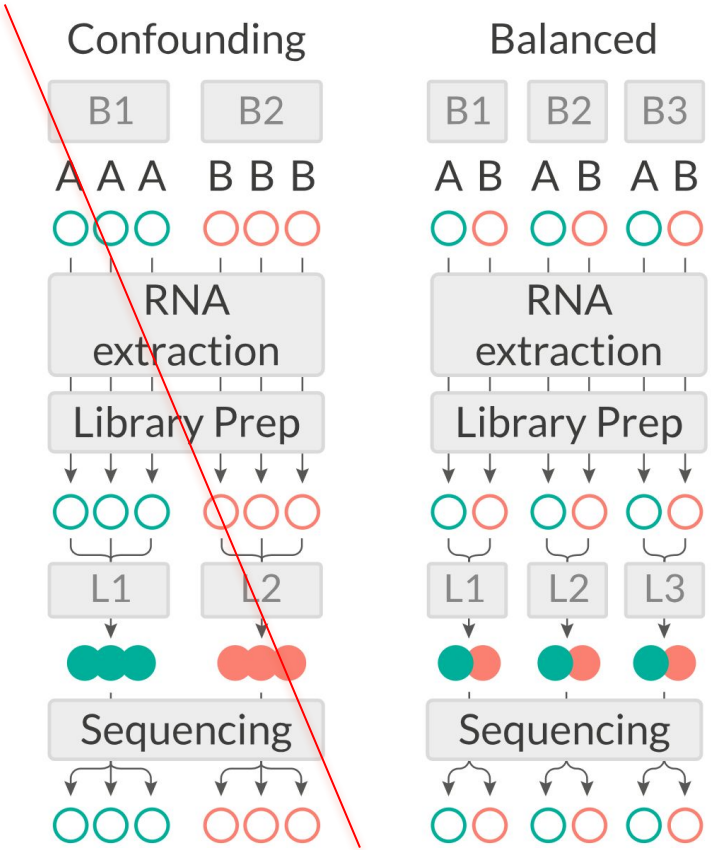
Effect size (fold change)	Depth @ 3 replicates		
	20	40	80
1.25 (+/- 25%)	0.26	0.31	0.34
1.75 (+/- 75%)	0.45	0.52	0.57
2 (+/- 100%)	0.62	0.70	0.75

coefficient of variation of counts: 0.3, FP rate: 0.05

Batch effects

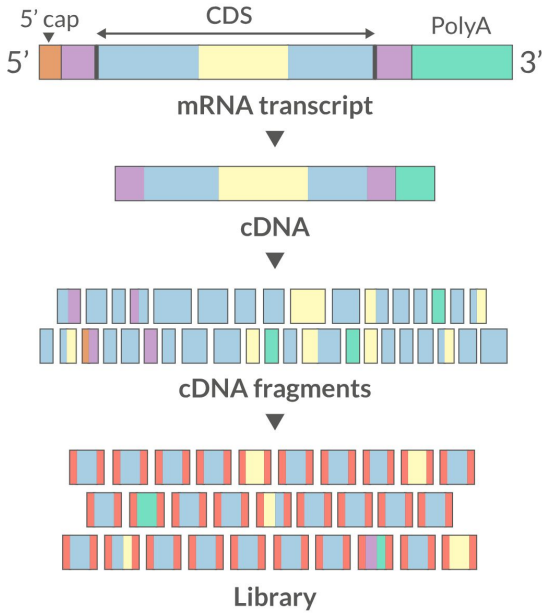
Differences in baseline between replicate samples prepared under different “conditions”

- Balanced design to avoid batch effects
- If you expect to have batch effects (samples prepared on different days, different chemistry….), ENSURE you have all conditions you expect to use in a differential analysis in EVERY batch
- Use batch as factor in your model for DGEA
- Or use tools to correct for batch effects, e.g. ComBat-seq



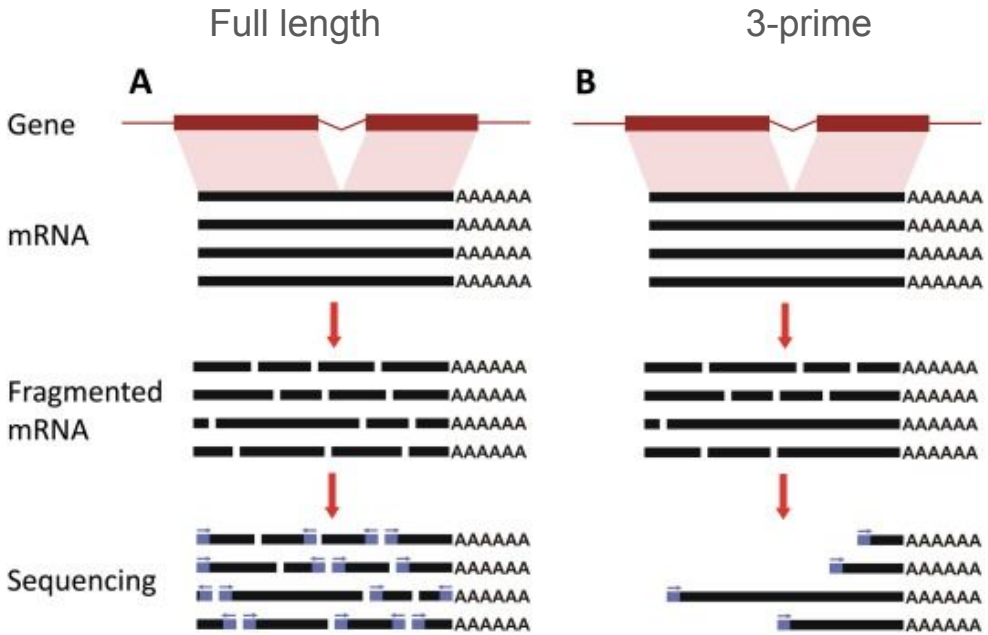
General considerations

1. Library type
    a. full length
    b. 3-prime
    c. paired-end
    d. single-end
2. (poly-A) selected vs total RNA
3. stranded vs unstranded



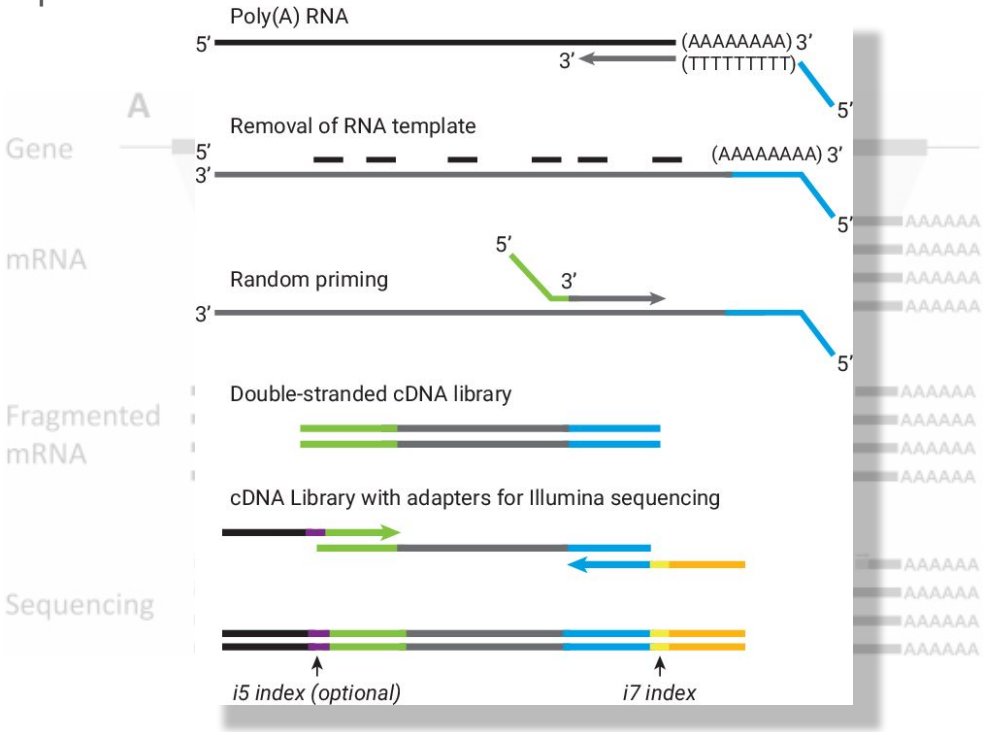
General considerations

Library type



General considerations

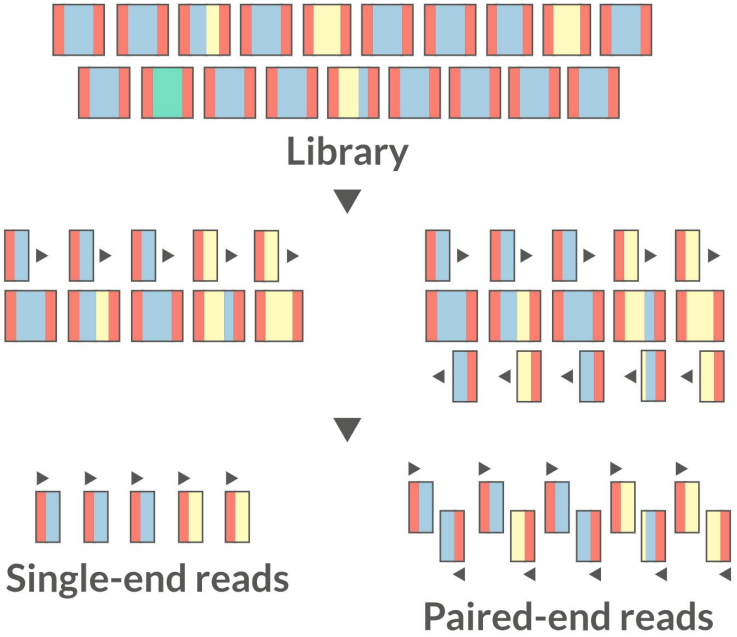
Lexogen QuantSeq 3 prime



General considerations

Library type

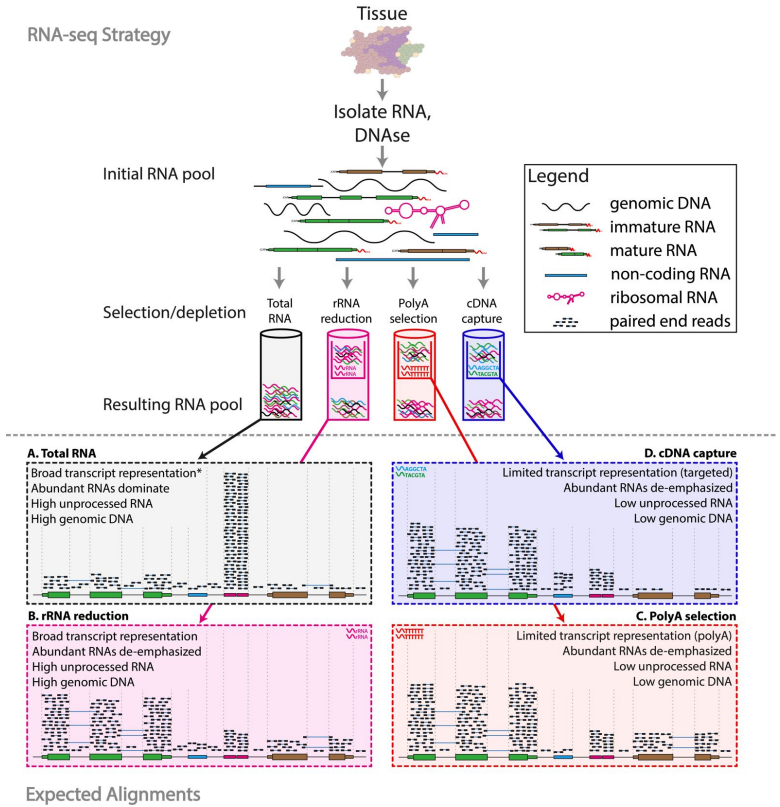
single-end vs paired-end



General considerations

Library type

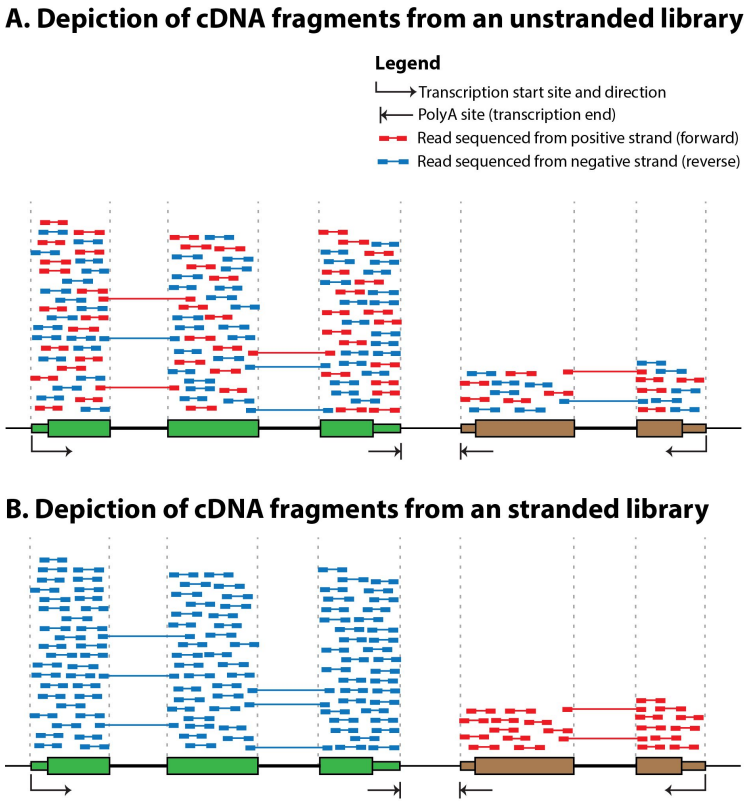
total vs selected



General considerations

Library type

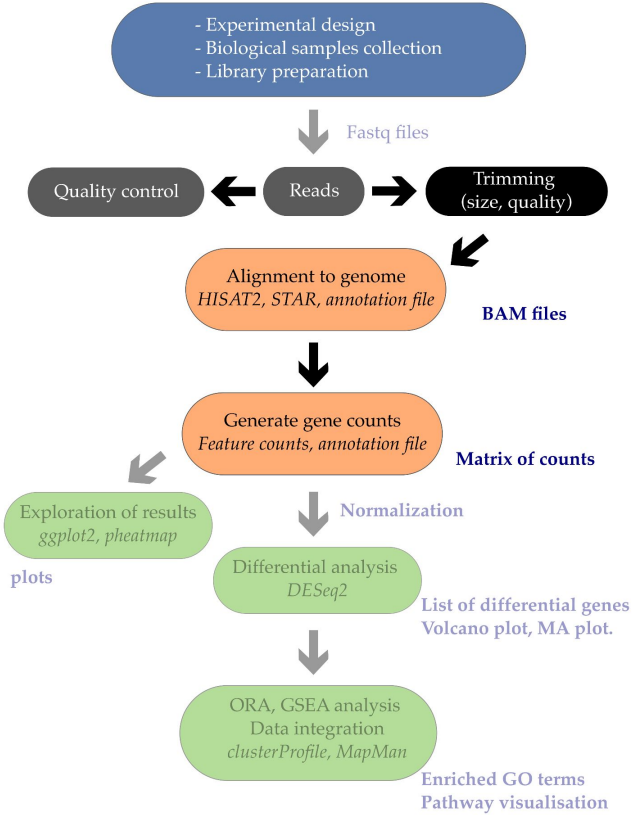
stranded vs unstranded



RNA-seq data analysis: workflow

Each type of RNA-seq analysis has distinct requirements and challenges but also a **common theme**:

1. Obtain **raw data** (convert format)
2. **Align** reads (e.g. STAR)
3. **Process alignment** with a tool specific to the goal e.g:
   a. ‘salmon’ or ‘feature-counts’ for expression analysis
   b. ‘StringTie2’ for transcript assembly
   c. ‘defuse’ or ‘arriba’ for fusion detection
   d. …
4. **Post process**:
   downstream software (R, Cytoscape, etc.)
   DGEA, ORA, Pathways….
5. **Summarize** and visualize
6. Create gene lists, prioritize **candidates** for validation, etc.

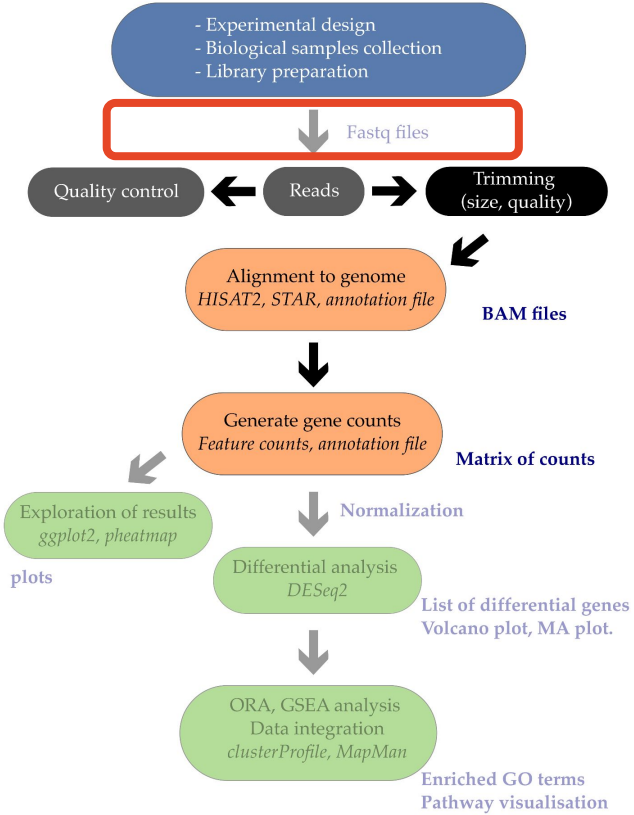


RNA-seq data analysis: FastQ files

- Simple **extension** of **FASTA** format
- Each **read** is represented as **block** of **4 lines**:
    - Sequence Name (read ID)
    - Sequence (actual read)
    - + (optional: Sequence name again)
    - Associated base call quality score

Example:

```
@HWI-ST1240:228:C42CRACXX:3:1101:1401:1993 1:N:0:CGTACTAG
NCCCTTAGAGCCAATCCTTATCCCGAAGTTACGGATCCGGCTTGCCGACTTCCCTTACCTACATTG
+
#0<BFFF<B<BBFFBBFFIIFFFIIFFBBFFIIIFFFFFBFIFFFFFIFFB<BBBBBBFFBB<BFB
```



FastQ files: Q-score

Phred system

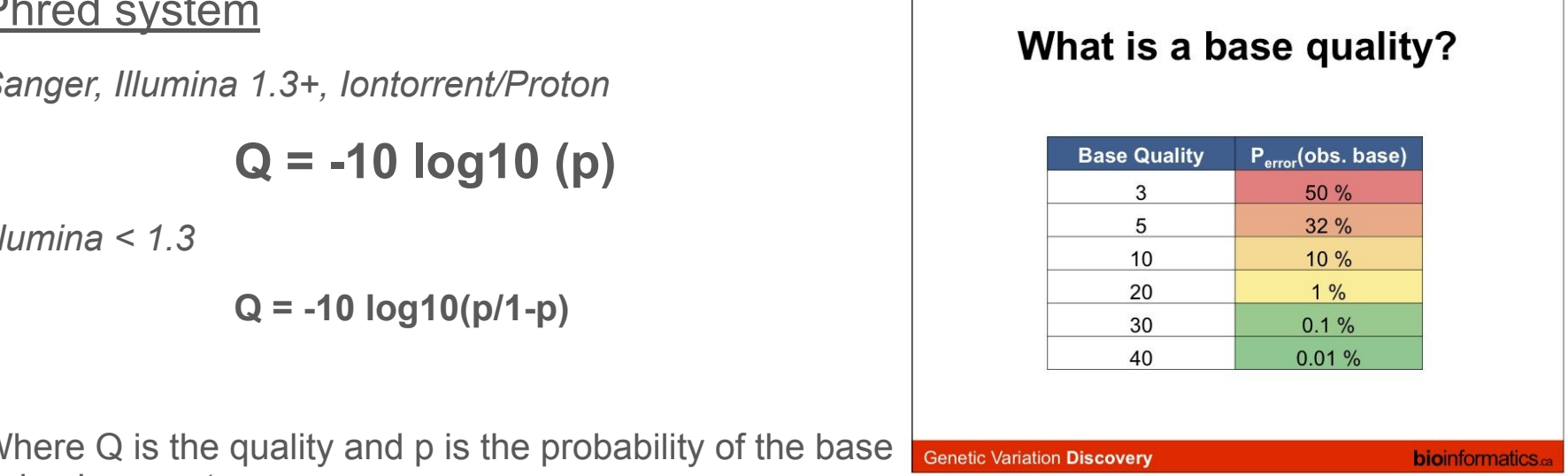
Sanger, Illumina 1.3+, Iontorrent/Proton

$$Q = -10 \log_{10} (p)$$

Illumina < 1.3

$$Q = -10 \log_{10}(p/1-p)$$

Where Q is the quality and p is the probability of the base being incorrect.

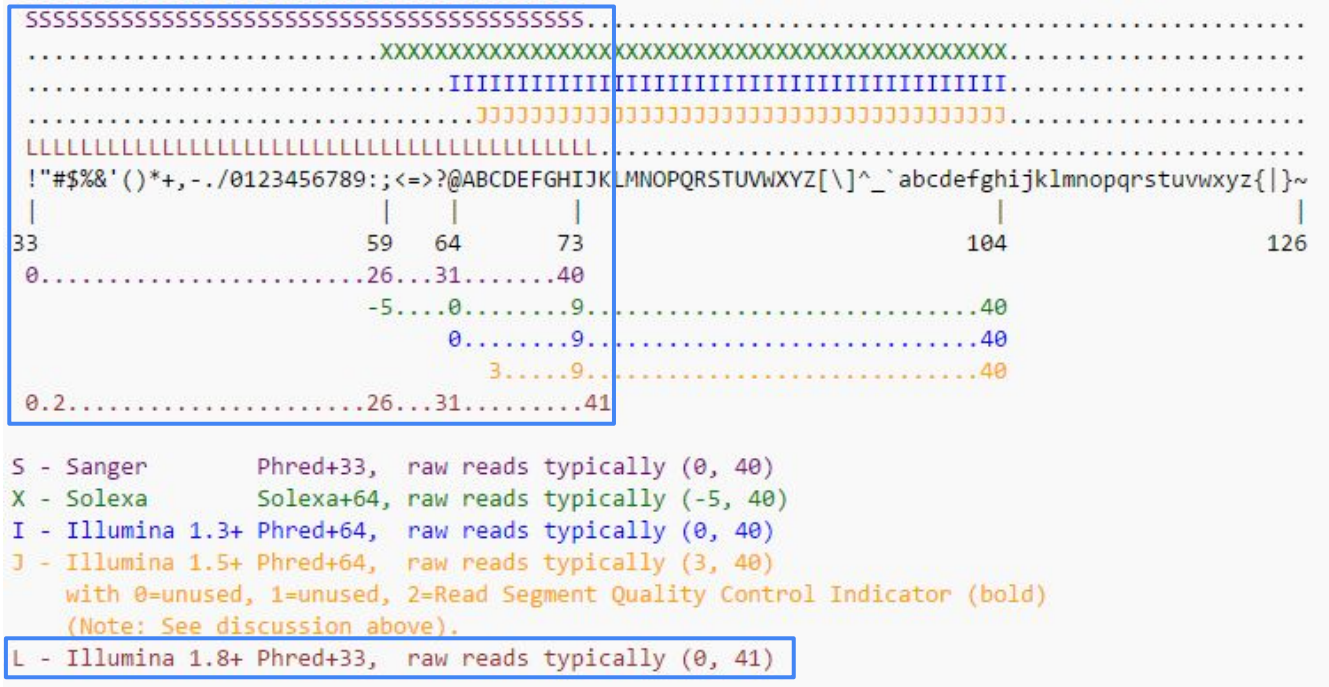



```
@HWI-ST1240:228:C42CRACXX:3:1101:1401:1993 1:N:0:CGTACTAG
NCCCTTAGAGCCAATCCTTATCCCGAAGTTACGGATCCGGCTTGCCGACTTCCCTTACCTACATTG
+
#0<BFFF<B<BBFFBBFFIIFFFIIFFBBFFIIIFFFFFBFIFFFFFIFFB<BBBBBBFFBB<BFB
```

FastQ files: Q-score

Phred system

Q = -10 log10 (p)



What is a base quality?

Base Quality	P_{error}(obs. base)
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

FastQ files: Q-score

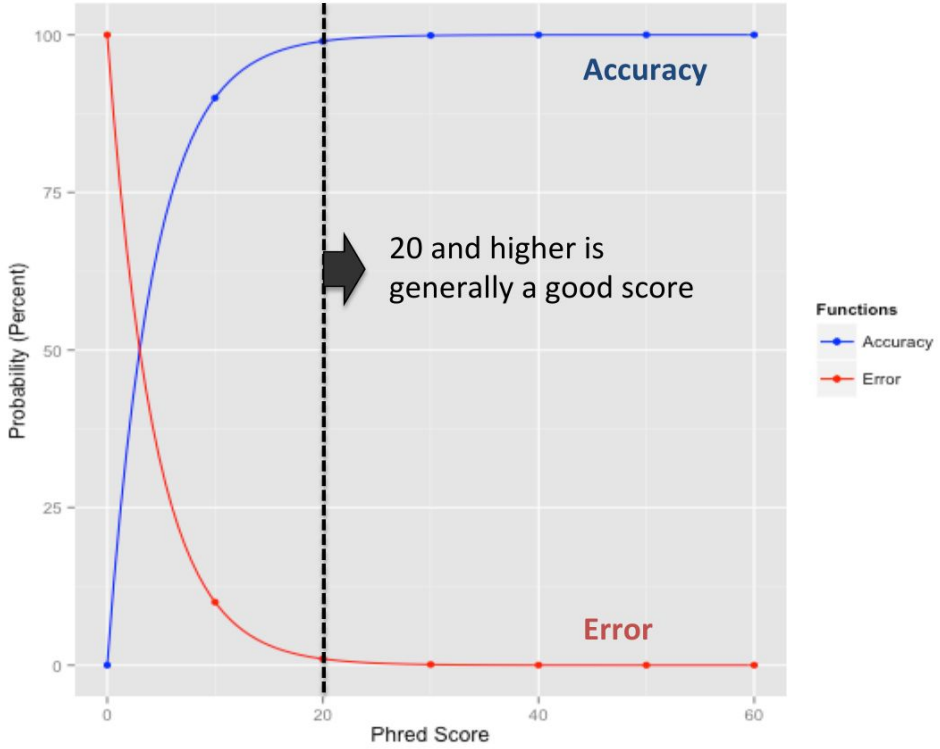
Phred system

Q = -10 log10 (p)

What is a base quality?

Base Quality	P_{error}(obs. base)
3	50 %
5	32 %
10	10 %
20	1 %
30	0.1 %
40	0.01 %

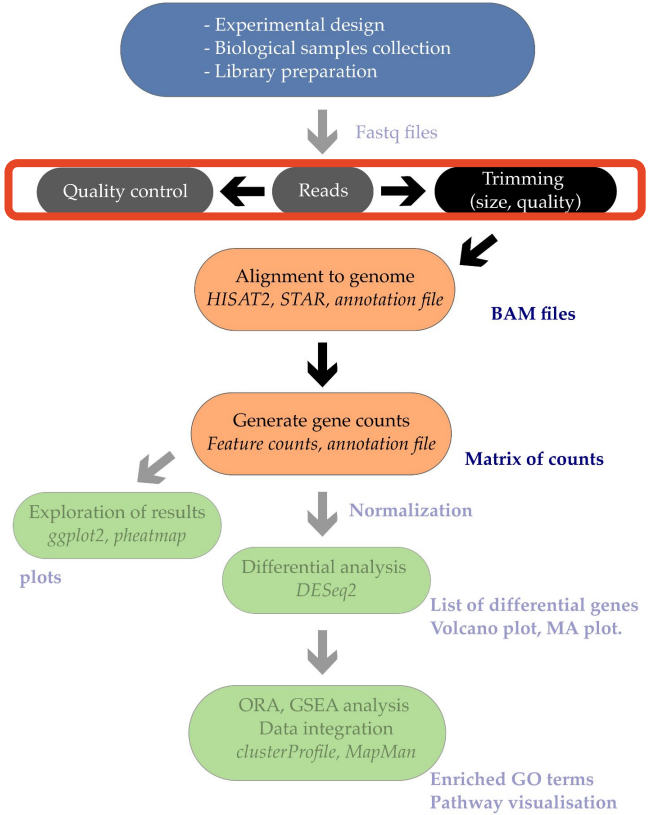
Genetic Variation **Discovery** bioinformatics.ca



RNA-seq data analysis: initial quality assessment

Common problems that can affect analysis

- Low confidence **base calls**
  - typically toward ends of reads
  - criteria vary by application
- Presence of **adapter** sequence in reads
  - poor fragment size selection
  - protocol execution or artifacts
- Over-abundant sequence **duplicates**
- Library **contamination**

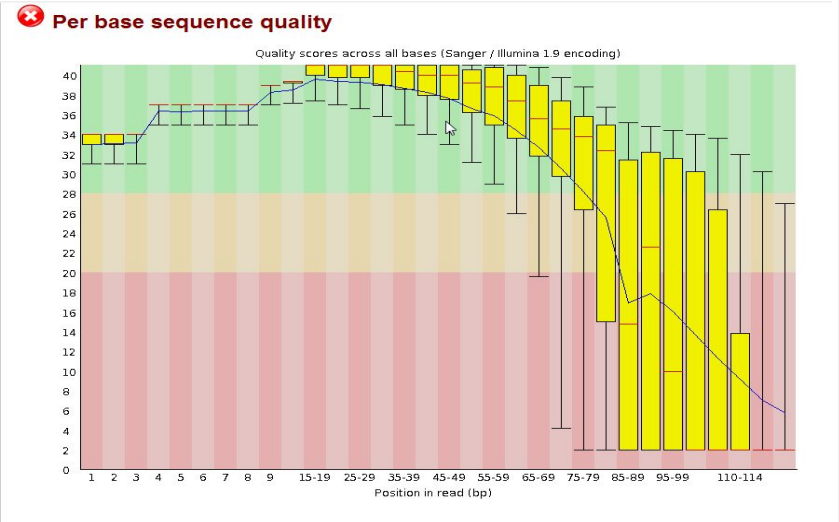


Quality assessment: FastQC

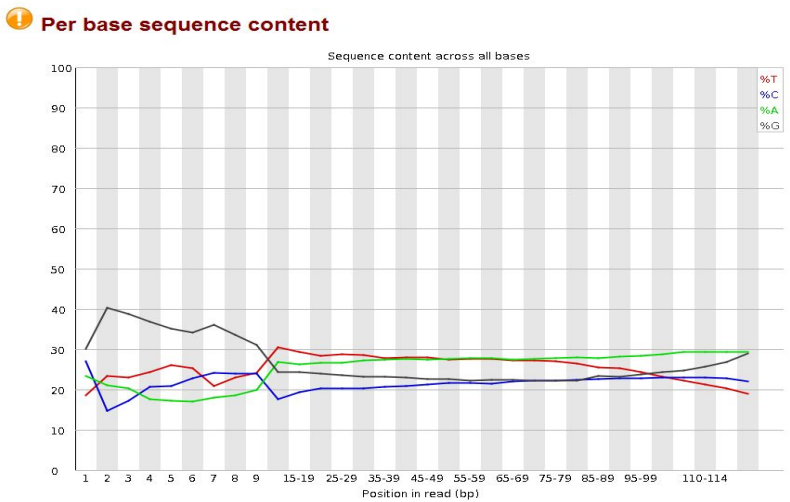
Free tool for assessing read quality

- Overall **base call quality**
- **Positional bias**
- Presence of **adapter** sequences in reads
- Over-abundant sequence **duplicates**
- Library **contamination**
- **…**

Quality assessment: FastQC

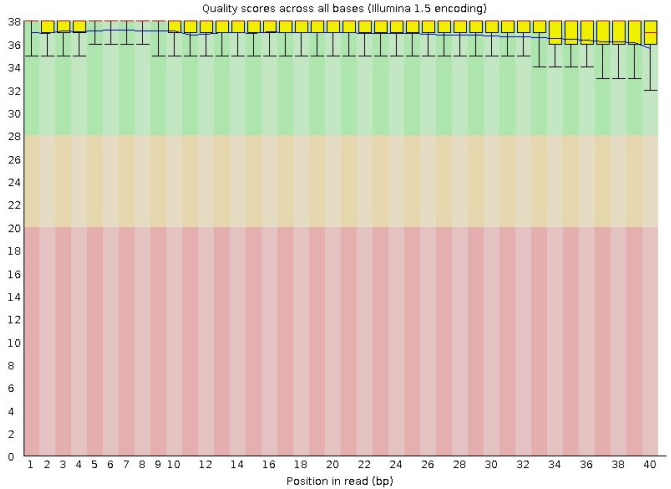


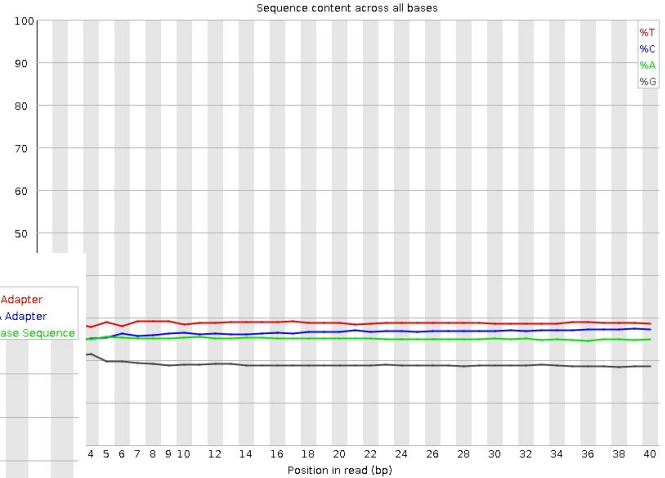
BAD base quality
(run or library issue)

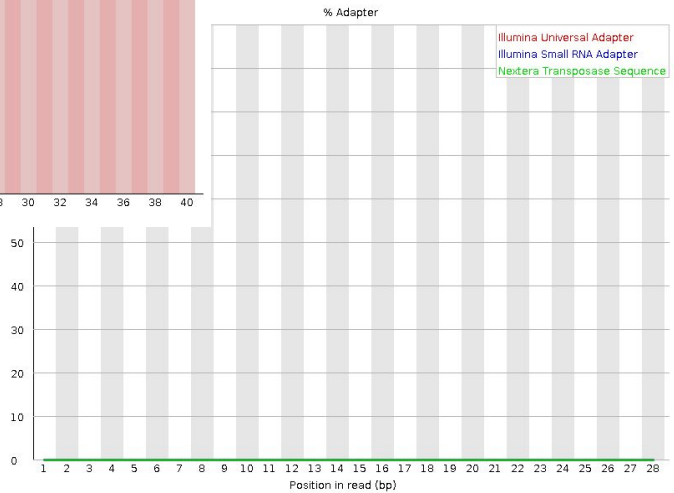


Leading bases!?
(library artifact)

Quality assessment: FastQC

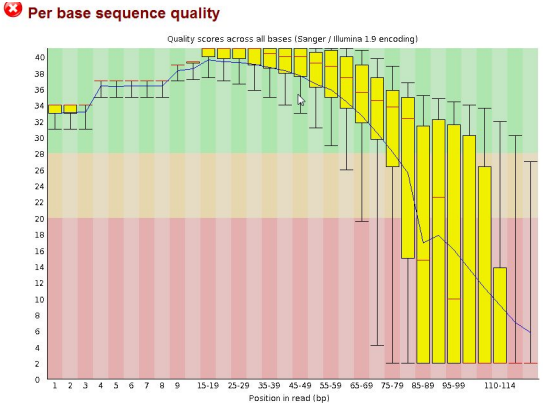




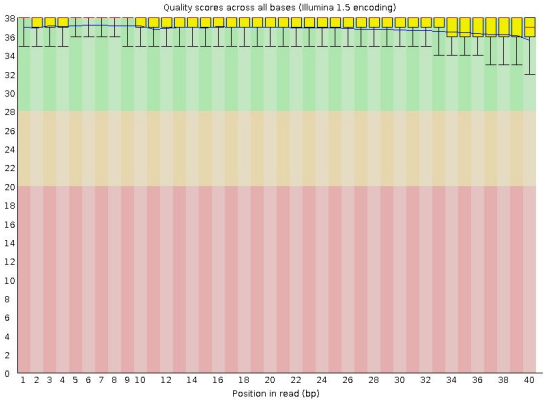


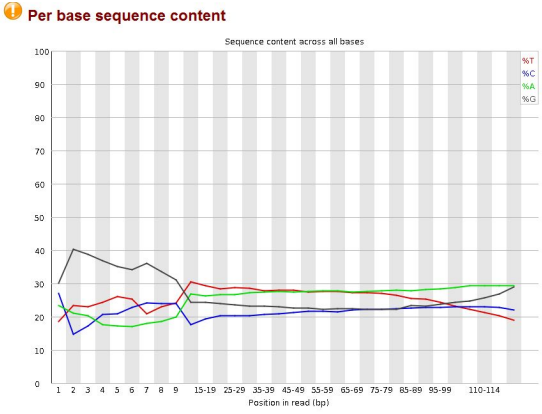
Good Illumina data

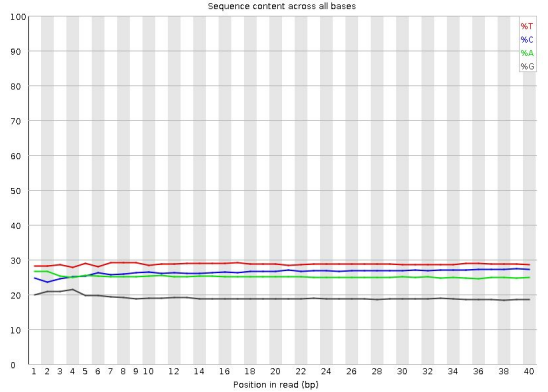
How to get from Bad to Good?



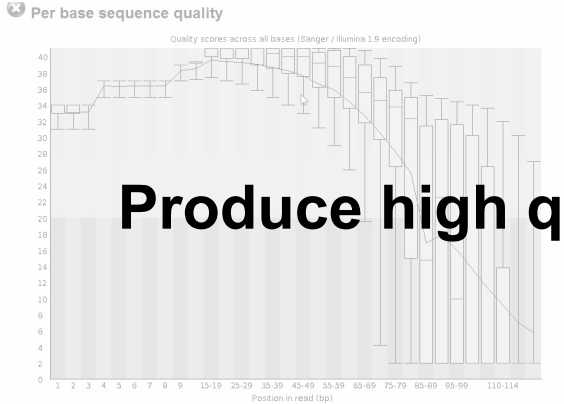
?





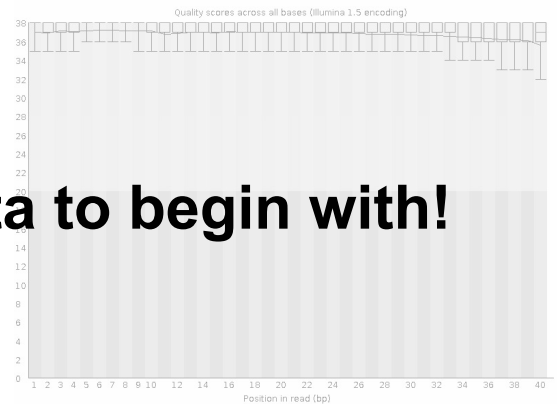


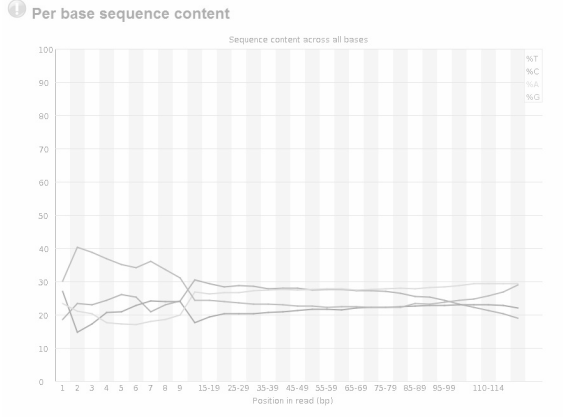
How to get from Bad to Good?

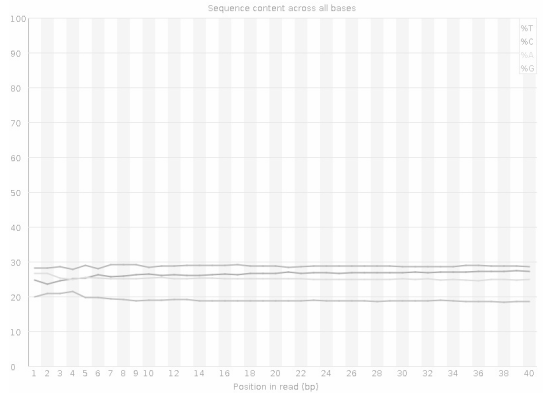


Produce high quality data to begin with!

?



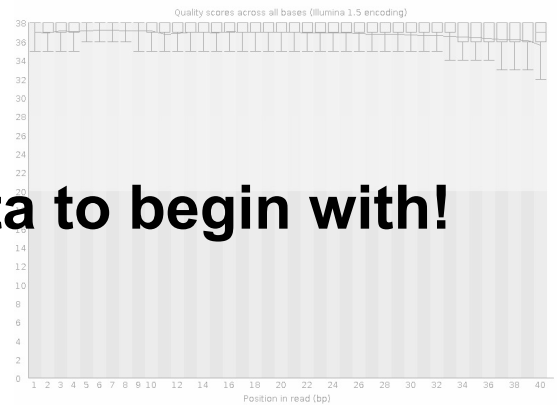


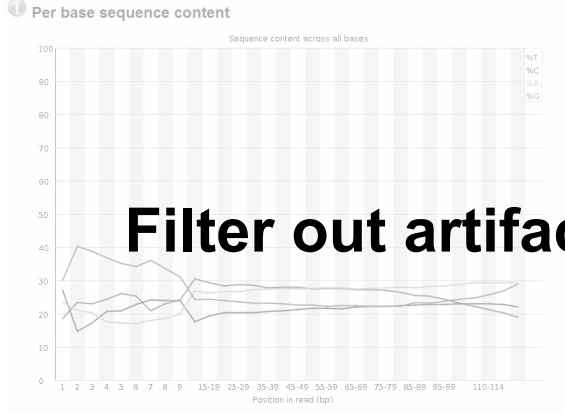


How to get from Bad to Good?

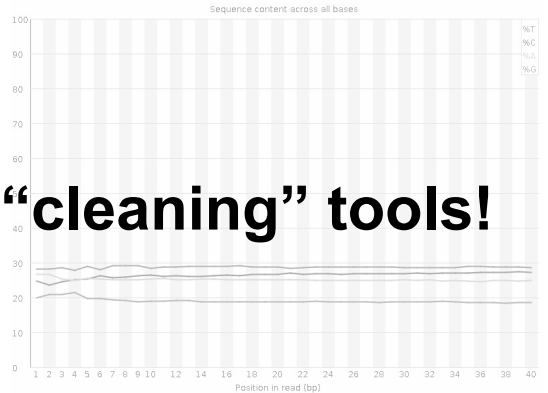
Produce high quality data to begin with!

?





Filter out artifacts using “cleaning” tools!



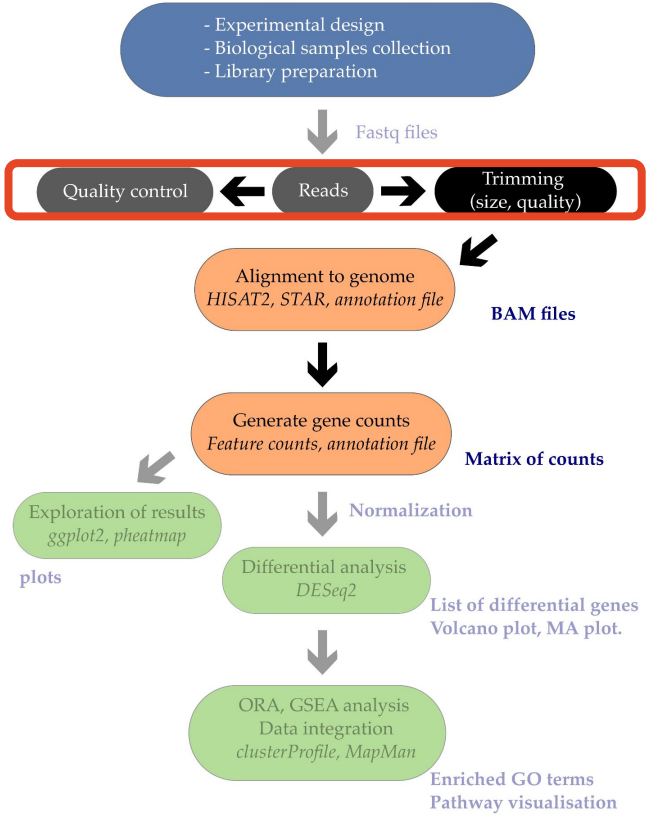
RNA-seq data analysis: read trimming / filtering

Trimming

- Trimming reads to remove adapter/readthrough or low quality bases
- Related options are hard clipping, filtering reads
- Sliding window trimming
- Filter by min/max read length
  - Remove reads less than ~18nt
- Demultiplexing/Splitting

Common tools

- Trimmomatic
- cutadapt
- fastp



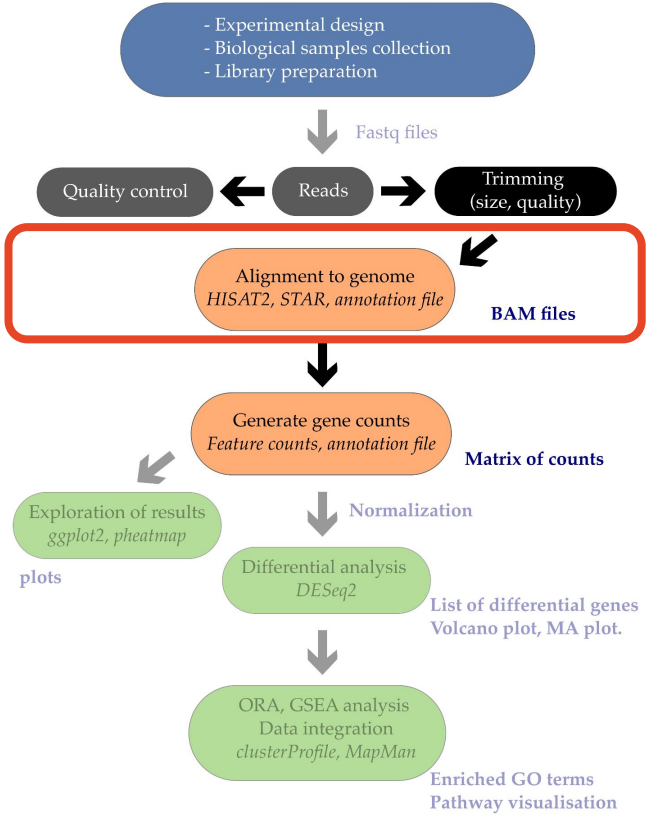
RNA-seq data analysis: read alignment

Mapping

- Aligning reads back to a **reference** sequence
- Mapping to **genome** or **transcriptome**
- **Splice-aware** alignment (genome)
- Pseudo-mapping
- Results stored in **SAM/BAM** file

Common tools

- STAR (genome)
- HISAT2 (genome)
- Bowtie (transcriptome)
- Salmon, Kallisto (pseudo)



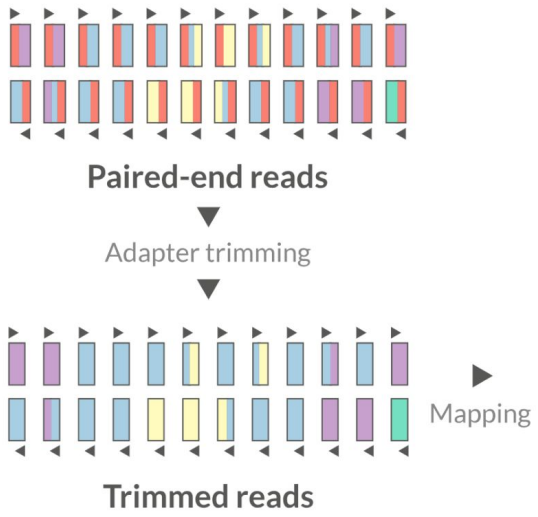
Genome vs transcriptom read alignment

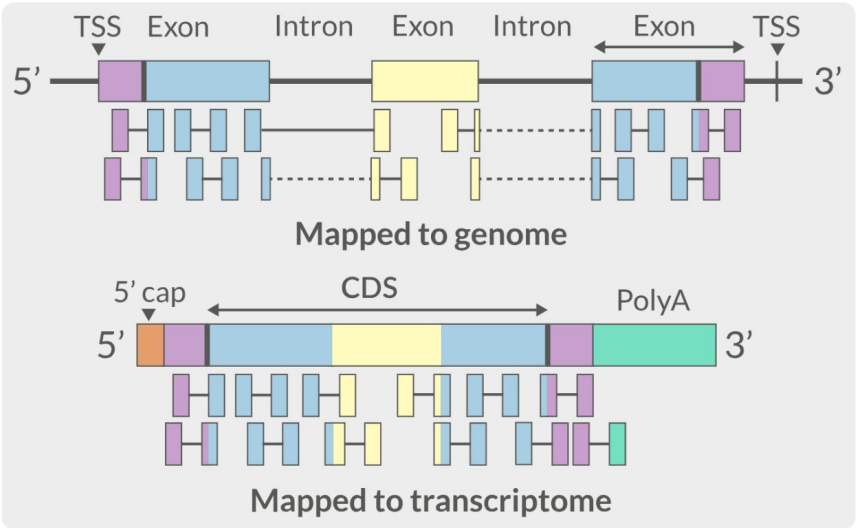
Genome

- Entire genome sequence as reference

Transcriptome

- Known transcript sequences as reference

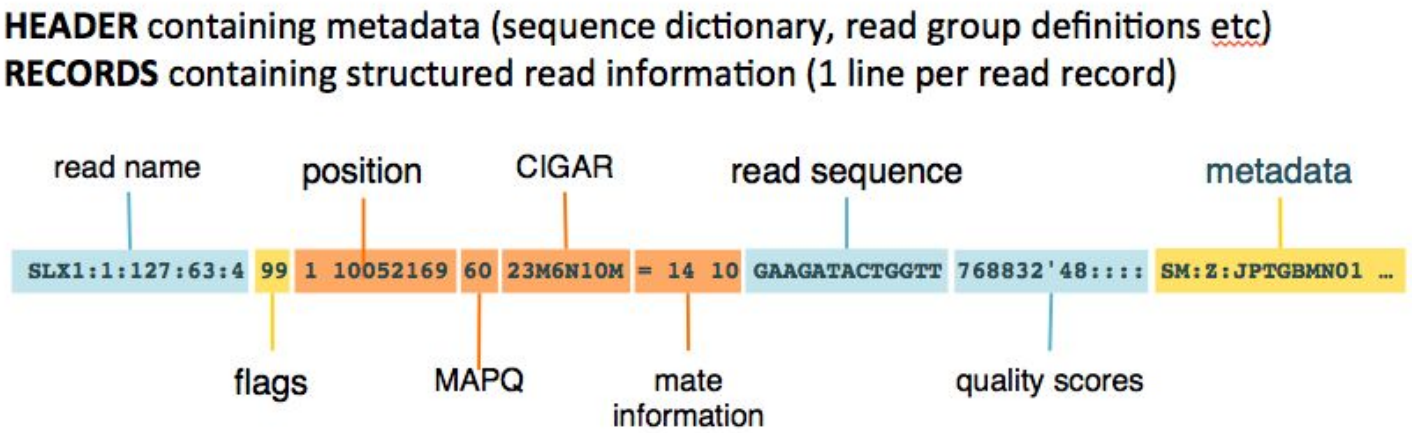




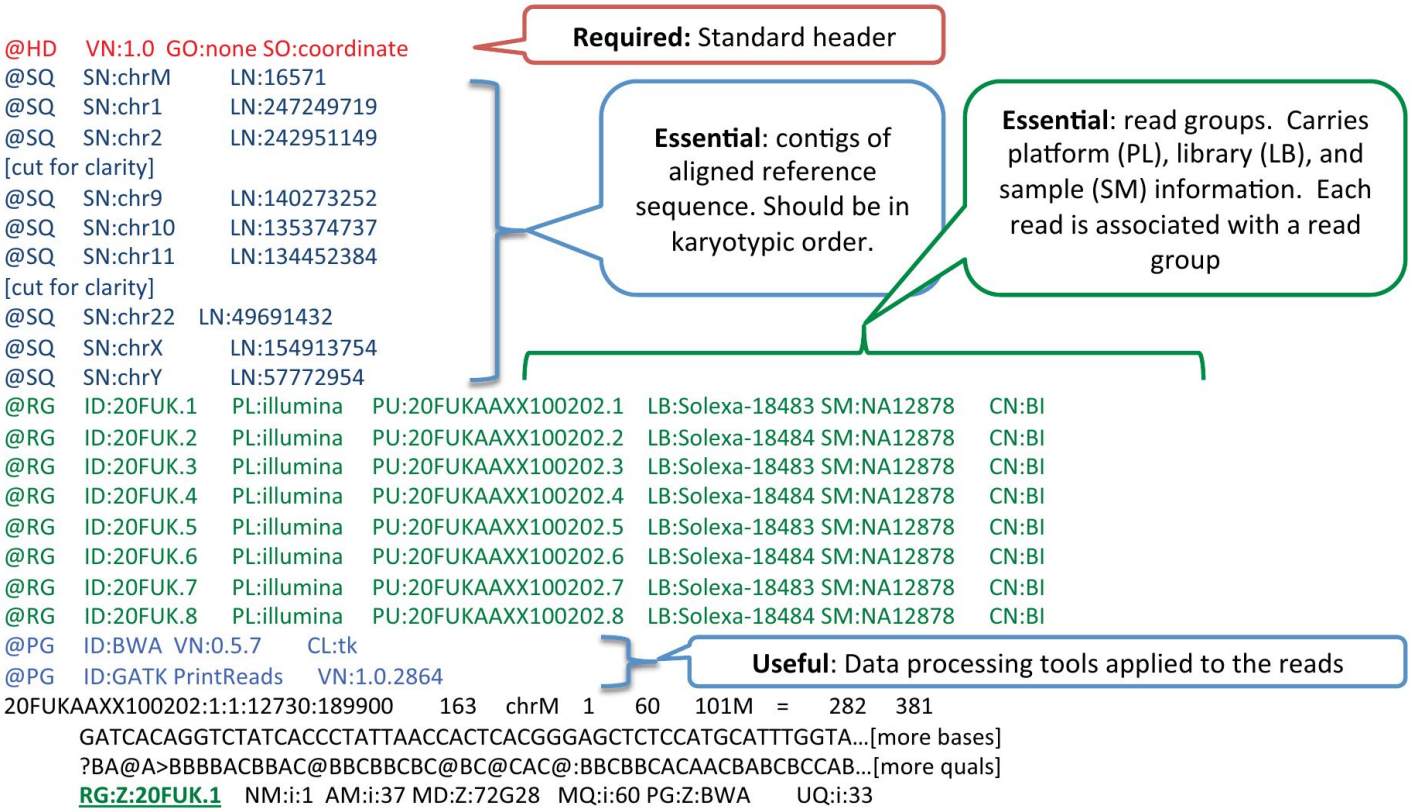
RNA-seq data analysis: SAM/BAM files

Sequence Alignment Map / Binary Alignment Map

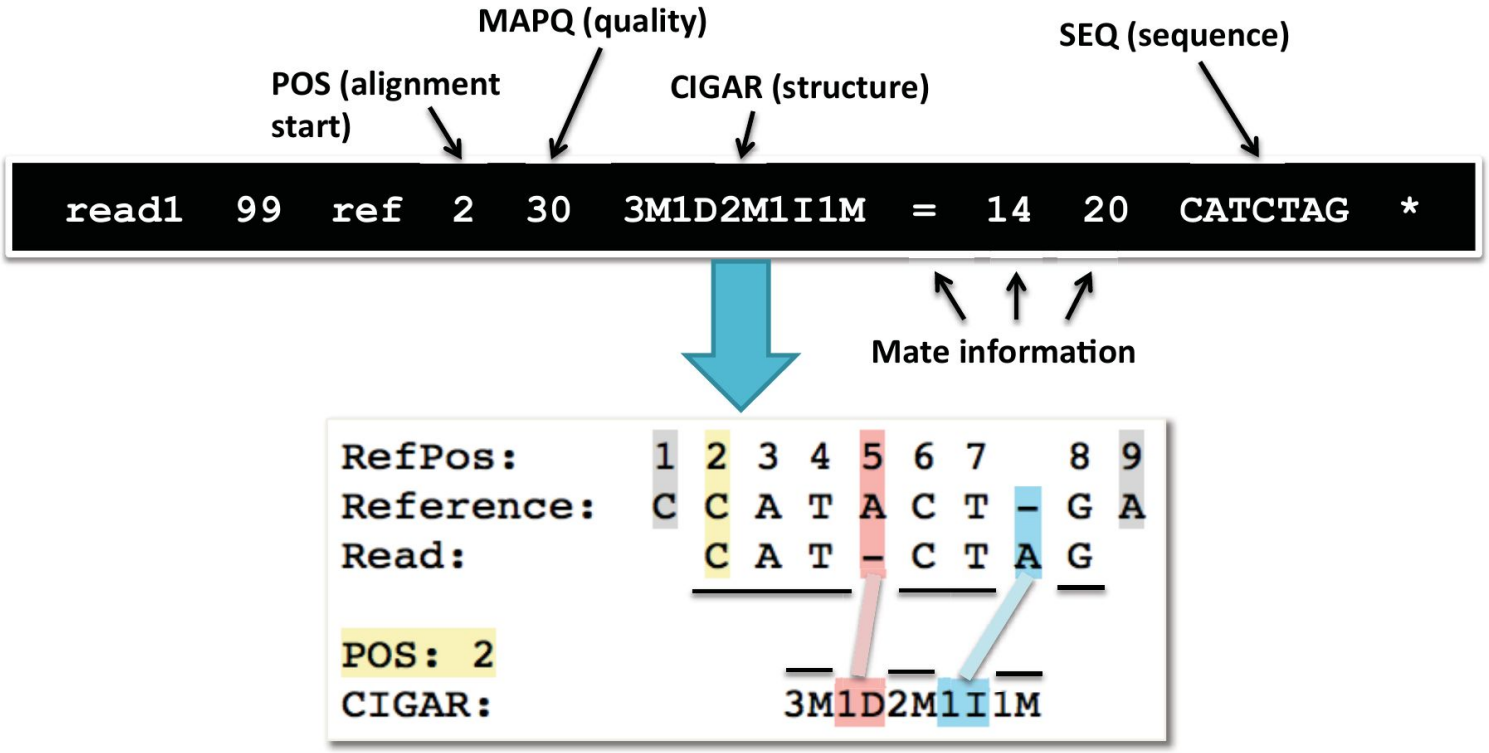
- Universal standard format (technology independent)
- Human-readable (SAM)
- Binary representation, compressed (BAM)
- SAM and BAM contain the exact same information



SAM/BAM files: Header



SAM/BAM files: Cigar string



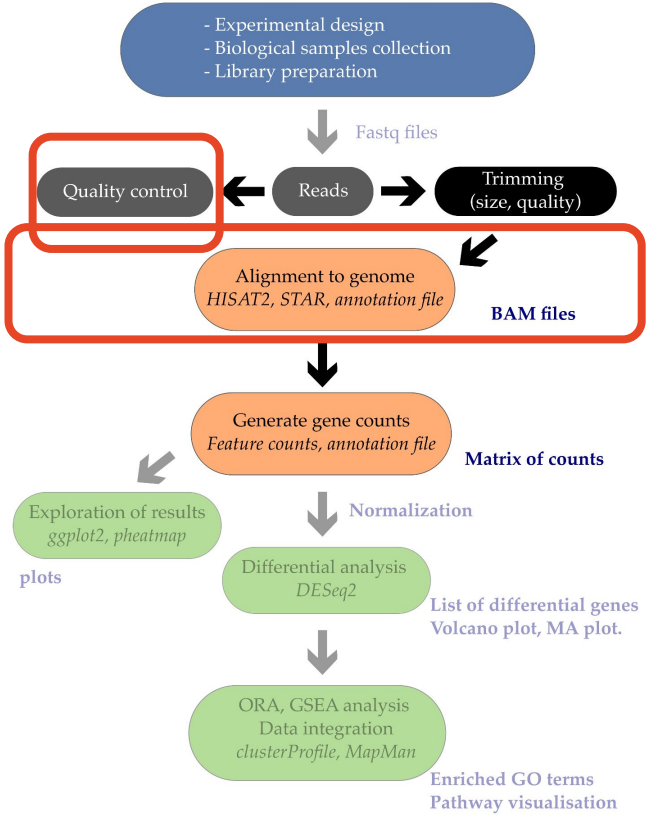
RNA-seq data analysis: alignment QC

Metrics

- Number of reads mapped/unmapped/paired etc.
- Uniquely mapped
- Insert size distribution
- Coverage
- Gene body coverage
- Biotype counts / Chromosome counts
- Counts by region: gene/intron/non-genic
- Sequencing saturation
- Strand specificity

Common tools

- samtools
- Picard
- QaliMap
- RSeQC
- STAR
- featureCounts
- MultiQC



Alignment QC: examples

General Statistics

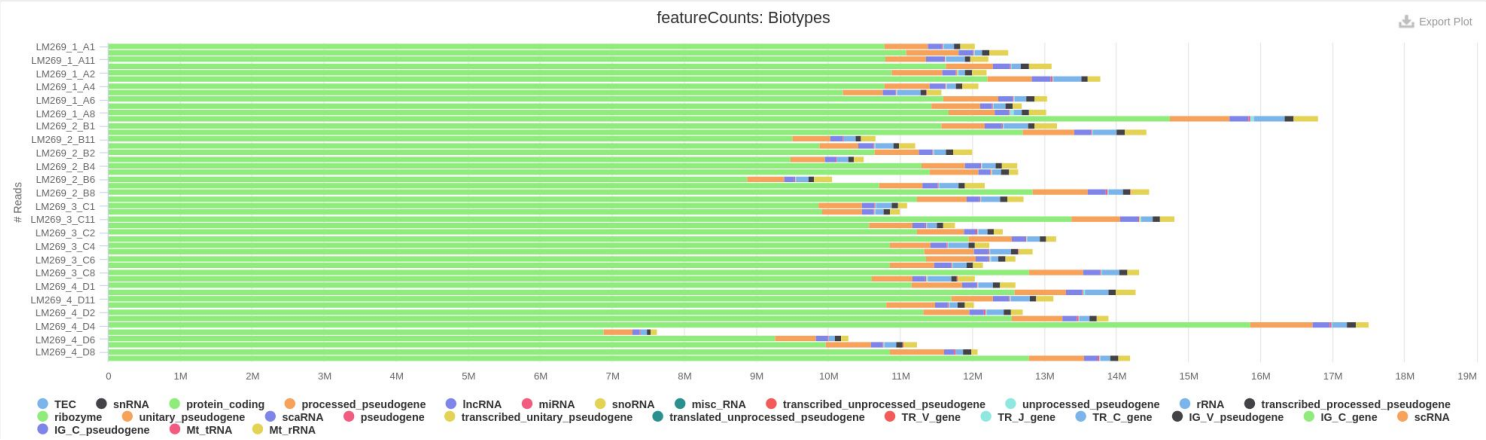
Copy table | Configure Columns | Plot | Showing 48/48 rows and 20/30 columns.

Sample Name	dupInt	% rRNA	% Dups	5'-3' bias	M Aligned	Error rate	M Non-Primary	M Reads Mapped	% Mapped	M Total seqs	M Reads Mapped	% Aligned	M Aligned	% Dups	% GC	M Seqs	% BP Trimmed
LM269_1_A1	0.16%	1.18%	81.1%	0.53	16.7	1.10%	4.9	16.7	100.0%	16.7	21.6	76.0%	14.3	51.8%	41%	18.8	14.1%
LM269_1_A10	0.24%	0.86%	84.3%	0.22	17.4	1.19%	5.0	17.4	100.0%	17.4	22.5	75.2%	14.9	58.2%	43%	19.8	16.8%
LM269_1_A11	0.21%	2.10%	82.3%	0.45	17.1	1.23%	5.2	17.1	100.0%	17.1	22.2	75.8%	14.6	55.6%	42%	19.2	14.2%

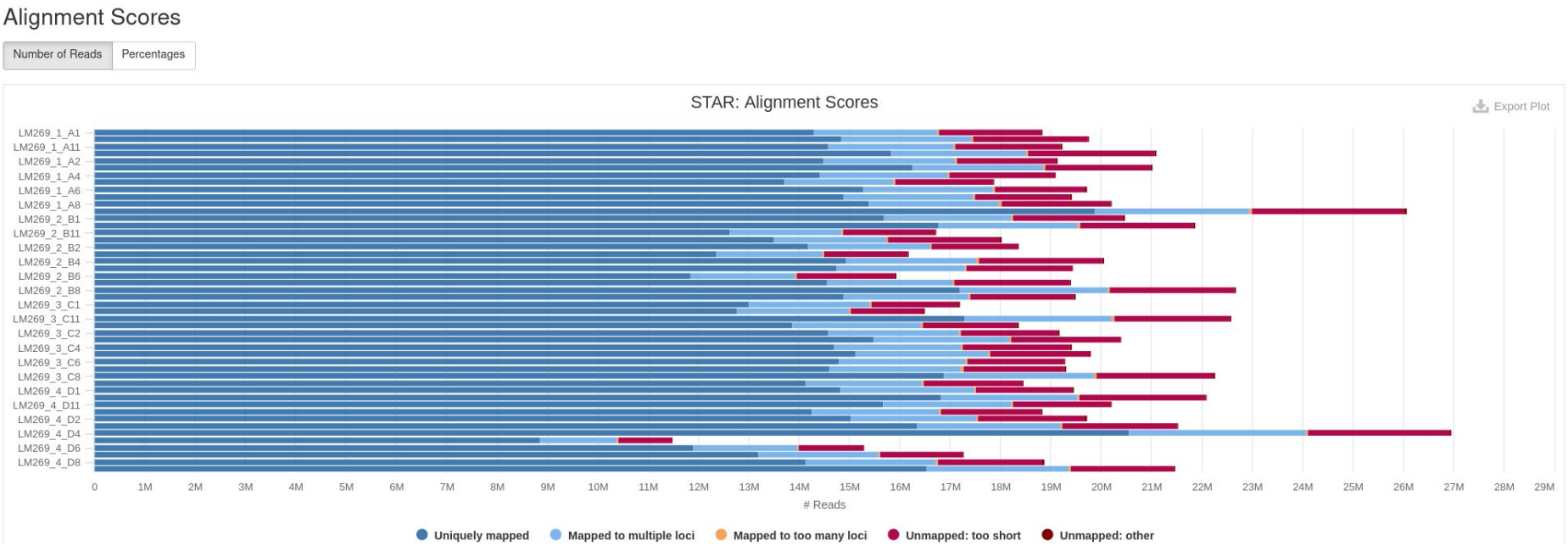
Biotype Counts

shows reads overlapping genomic features of different biotypes, counted by featureCounts.

Number of Reads | Percentages



Alignment QC: examples

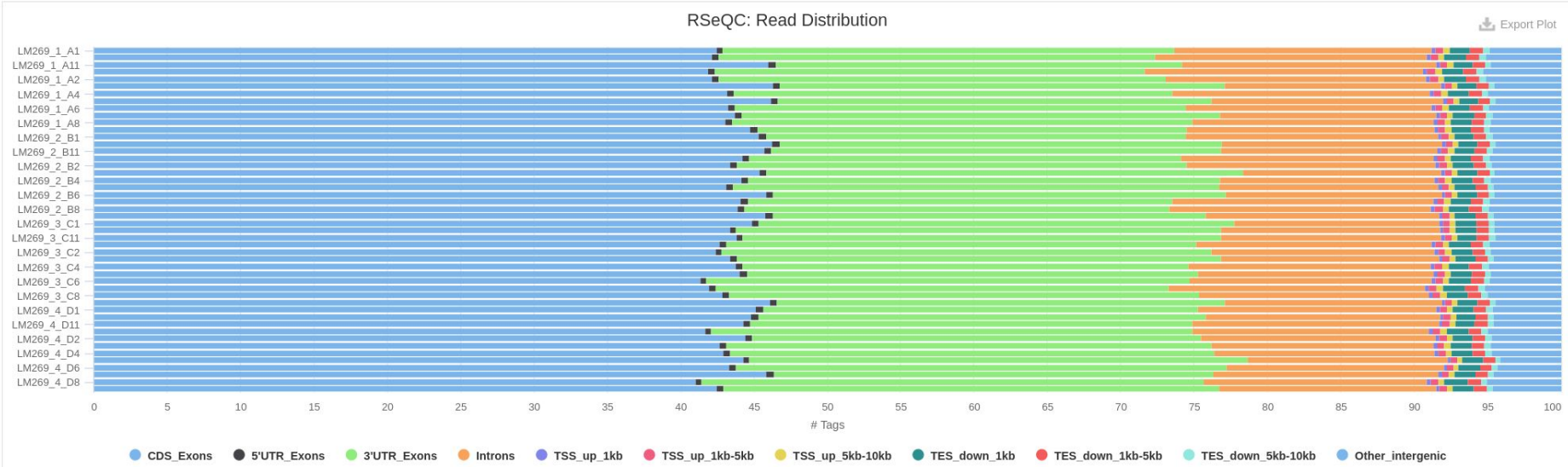


Alignment QC: examples

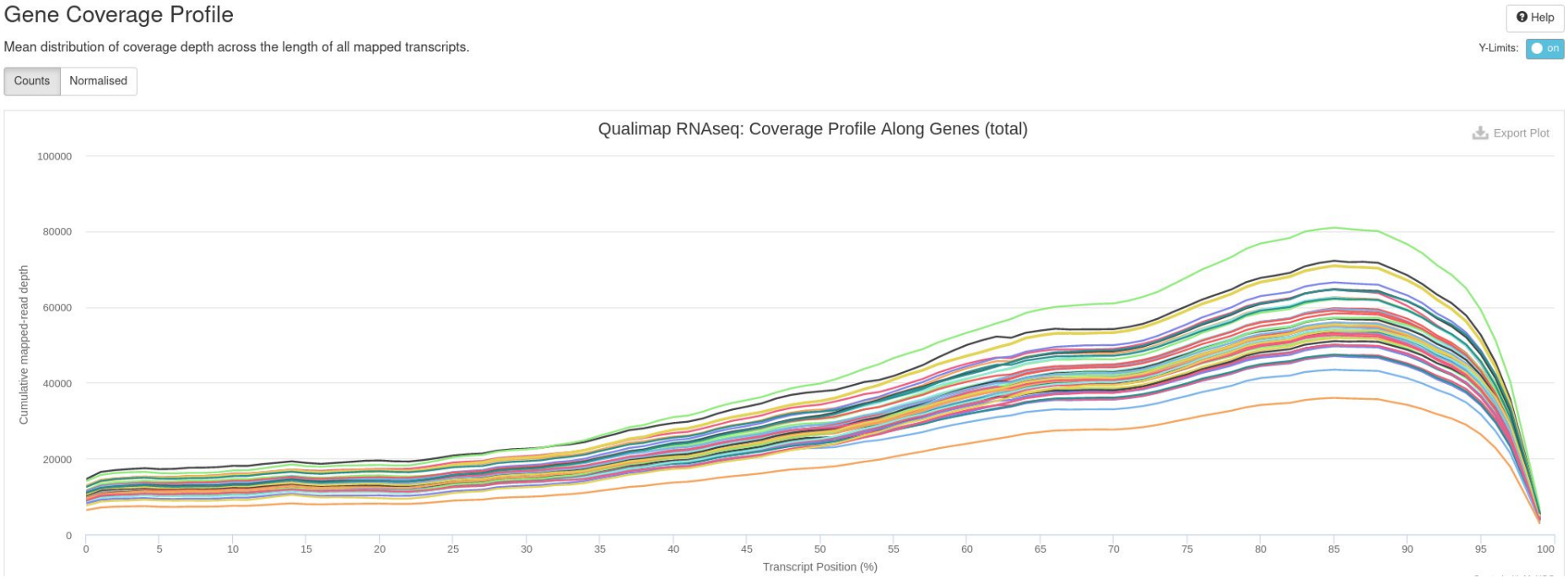
Read Distribution

Read Distribution calculates how mapped reads are distributed over genome features.

Number of Tags | Percentages



Alignment QC: examples



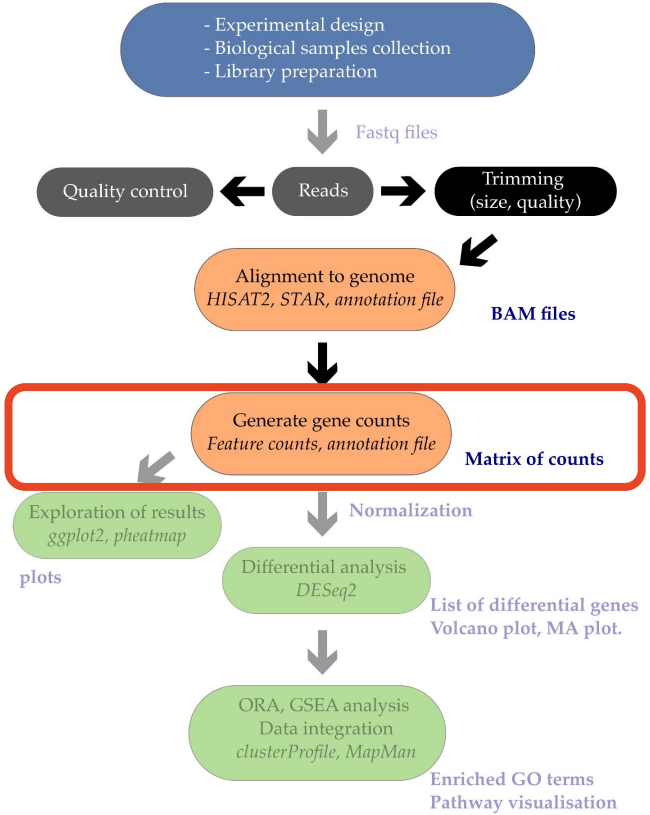
RNA-seq data analysis: quantification

Counts

- Read counts ~ gene expression
- Reads can be quantified on any feature
  - gene
  - transcript
  - exon etc.
- Intersection on gene models
- Gene/Transcript level

Common tools

- featureCounts
- htseq-count
- Salmon (pseudo mapping / pre-aligned reads)
- Kallisto (pseudo mapping)



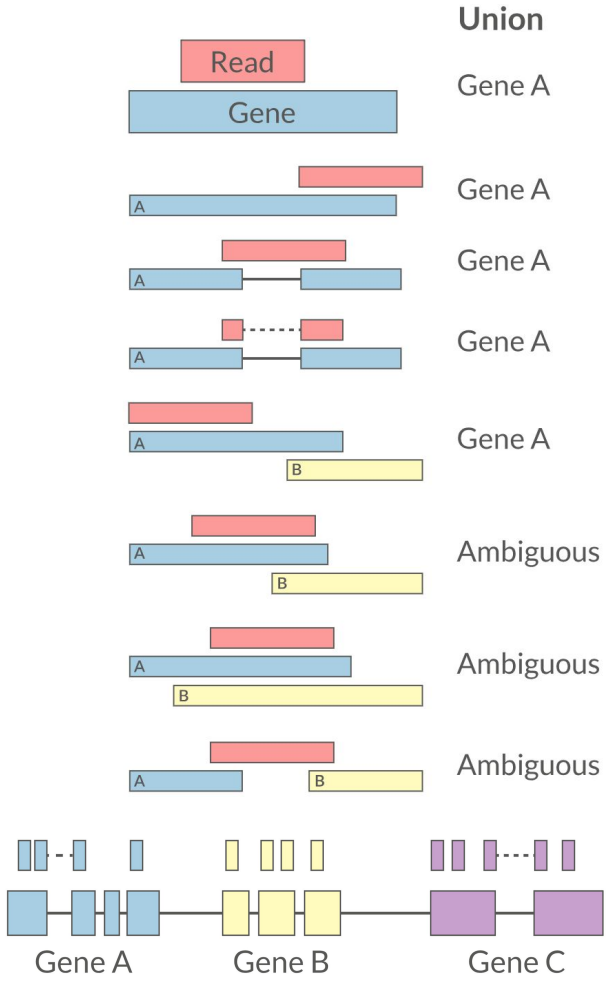
RNA-seq data analysis: quantification

Count rules

- Count each read at most once
- Discard a read if:
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene

Gene level counts

- "Collapse" transcript annotation to gene level
- Overlap read with collapsed gene

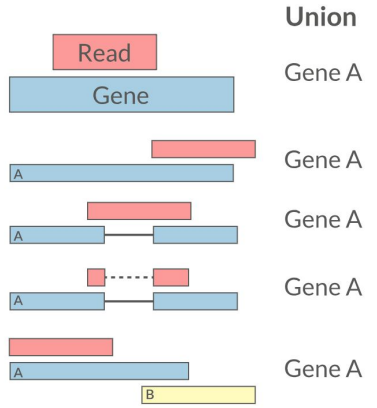


RNA-seq data analysis: quantification

Counting

- Provide no inference on isoforms
- Cannot accurately measure fold change

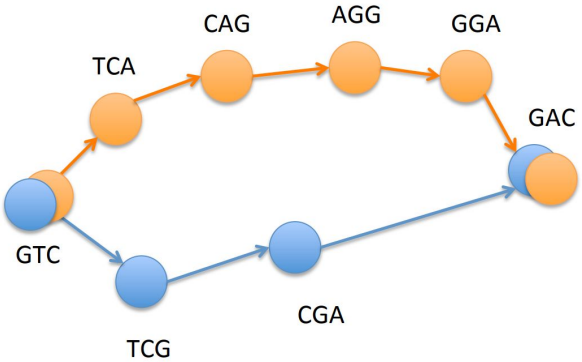
htseq-count, featureCounts



Probabilistic assignment

- Deconvolute ambiguous mappings
- Transcript-level
- cDNA reference

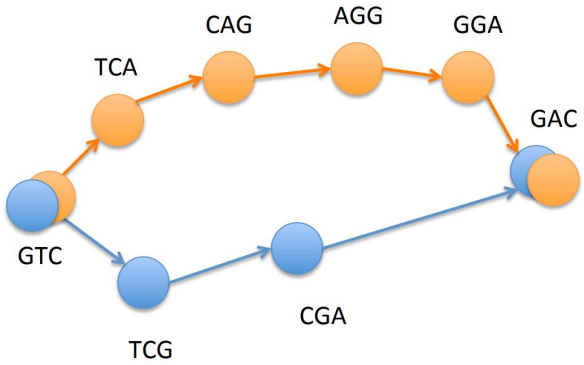
Salmon, Kallisto, RSEM



Pseudo-mapping + quantification

Indexing

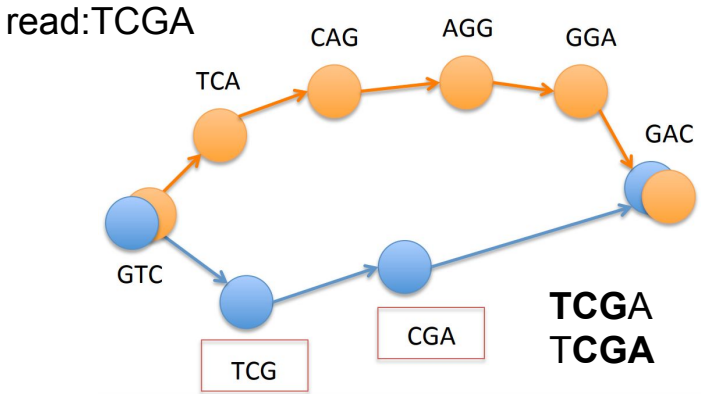
- Build an index by chopping ref transcriptomes into k-mers and putting them into graphs



Probabilistic assignment

- chop read(s) into k-mers
- find compatible transcript in graph

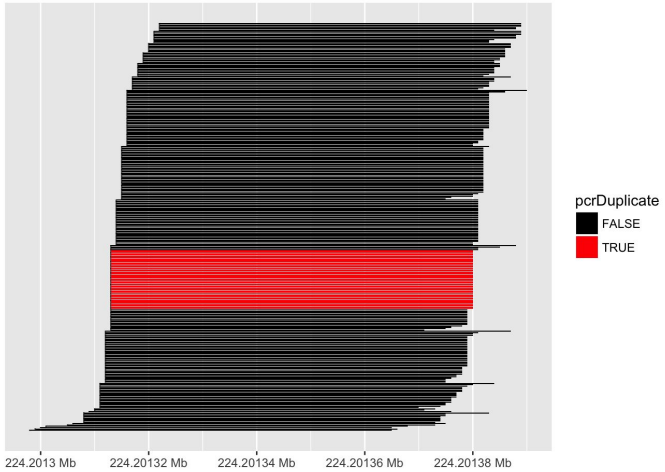
Salmon, Kallisto, RSEM

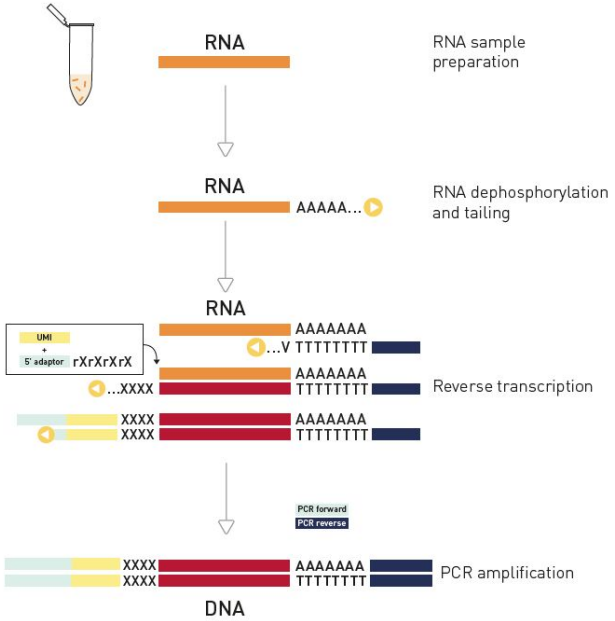


RNA-seq data analysis: quantification

PCR duplicates

- Computational deduplication (Don’t)
- Use PCR-free library-prep kits
- Use UMIs during library-prep





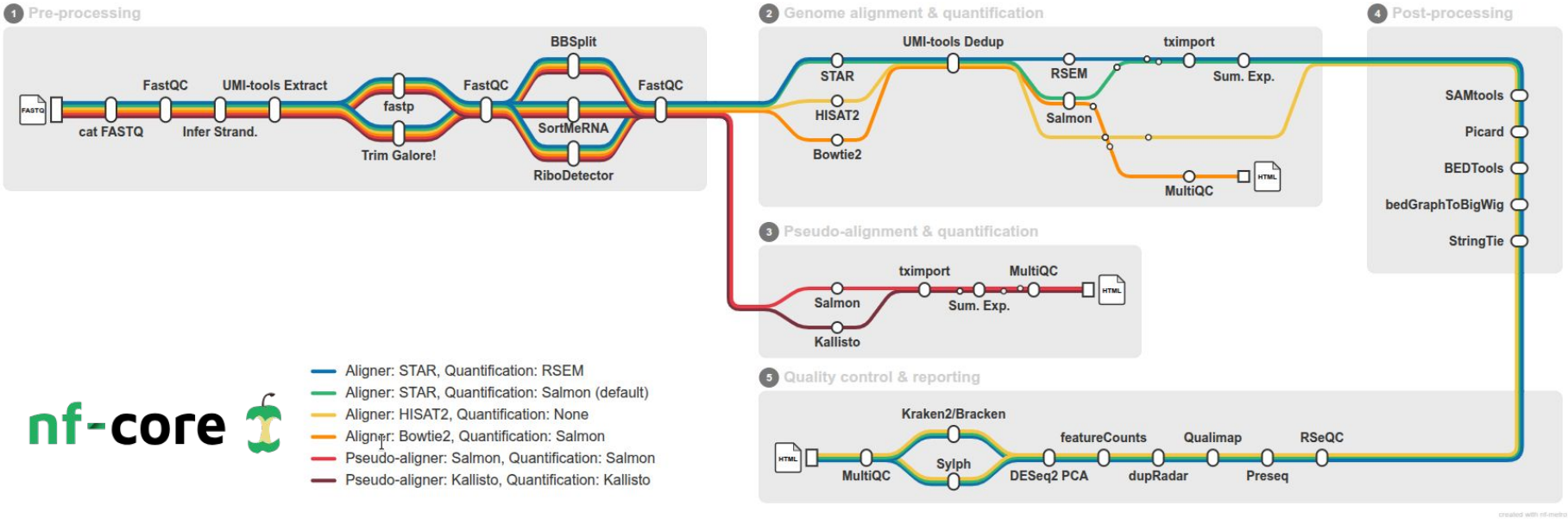
RNA-seq data analysis: quantification

Count table

- high-dimensional
- Genes x Samples
- raw counts not normalized

Gene	Sample1	Sample2	Sample 3
ENSG00000237613.2	10	12	9
ENSG00000268020.3	0	0	0
ENSG00000240361.2	2	7	7
ENSG00000186092.6	0	0	0
ENSG00000238009.6	0	0	0
ENSG00000239945.1	1092	987	432
ENSG00000233750.3	0	0	0
...	0	0	0
56000+ more rows ...			

RNA-seq analysis pipeline: nf-core/rnaseq



RNA-seq data analysis: Differential Gene Expression

For **each gene** we have a **measure of abundance**

# of reads mapping to each gene in each library

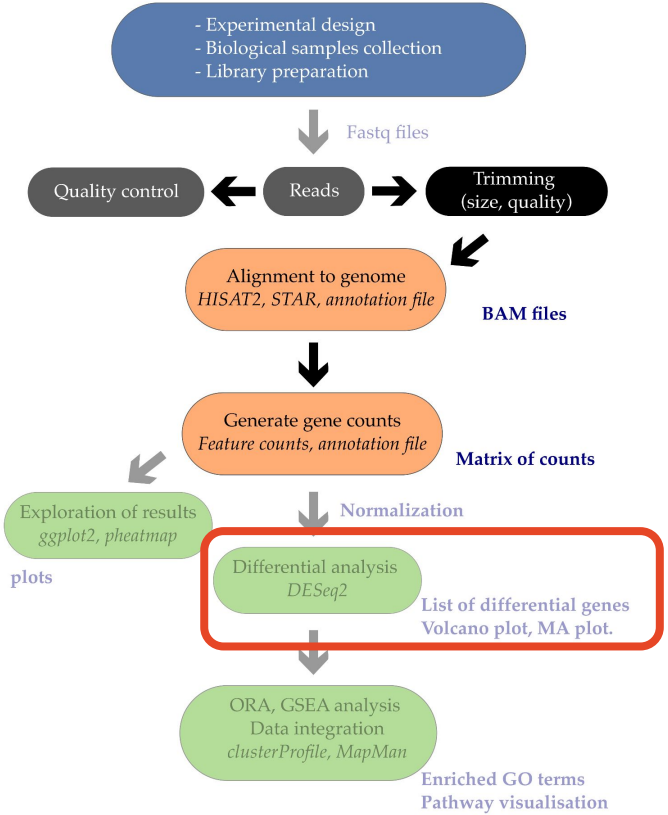
Question?

Is the number of reads mapping to the gene in a set of samples (condition A) different to the number of reads mapping to the gene in another set of samples (condition B)

read counts are not directly comparable

Need to account for:

- Library depth
- Data is not distributed normally!
- Lots of zero or near-zero counts (most genes are not expressed or expressed at low levels)



RNA-seq data analysis: Normalization

1. **Removal of systematic biases**:
   - Differences in library preparation
   - Sequencing depth
   - other technical variations

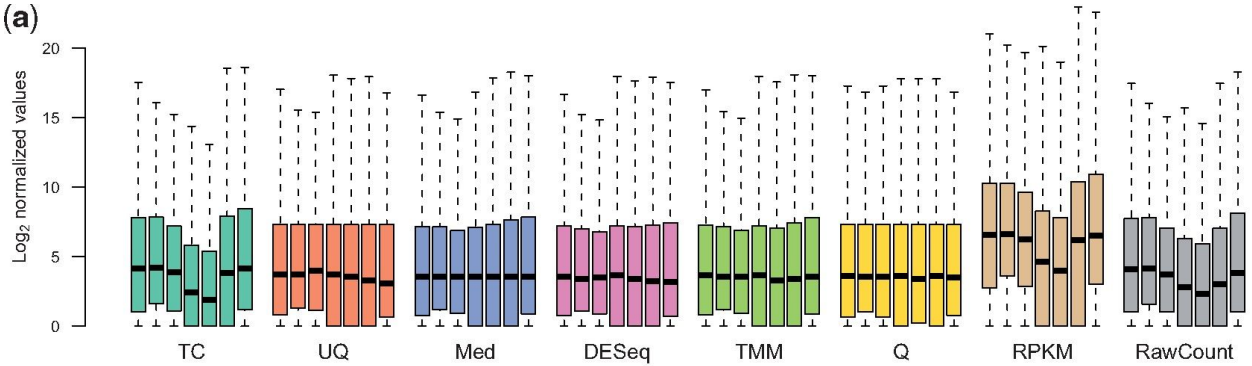
   Normalization aims to remove these biases to ensure that the gene expression measurements accurately reflect biological differences rather than technical artifacts.

2. **Comparison across samples**:
   - Comparison of gene expression levels across different samples
   - Identify true biological differences

3. **Data integration**:
   - Integration of RNA-seq data from different sources
   - Ensures that the gene expression measurements are on a comparable scale
   - Facilitates downstream analyses: clustering, DGEA, pathway enrichment analysis

RNA-seq data analysis: Normalization

Control for Sequencing depth & compositional bias



Total Count (TC), Upper Quartile (UQ), Median (Med), DESeq, Trimmed Mean of M values (TMM), Quantile (Q), Reads Per Kilobase per Million mapped reads (RPKM), Transcript Per Million (TPM)

Brief Bioinform, Volume 14, Issue 6, November 2013, Pages 671–683,https://doi.org/10.1093/bib/bbs046

RNA-seq data analysis: Normalization

- **Total Count (TC)**:

  $$\text{normalized count}_{(g)} = \text{count}_{(g)} / \text{TC}_{(s)} \times \text{mean}(\text{TC}_{(ds)})$$

- **Upper Quartile (UQ)**:

  $$\text{normalized count}_{(g)} = \text{count}_{(g)} / \text{UQ}_{(s)} \times \text{mean}(\text{UQ}_{(ds)})$$

- **Median (Med)**:

  $$\text{normalized count}_{(g)} = \text{count}_{(g)} / \text{MED}_{(s)} \times \text{median}(\text{MED}_{(ds)})$$

RNA-seq data analysis: Normalization

FPKM (RPKM):

Fragments (**R**eads) **P**er **K**ilobase of transcript per **M**illion mapped reads

The relative expression of a transcript is proportional to the number of cDNA fragments that originate from it.

However

- # fragments is **biased** towards **larger** genes → **P**er **K**ilobase of transcript
- # fragments is related to total **library depth** → per **M**illion mapped reads

RNA-seq data analysis: Normalization

FPKM (RPKM):

FPKM attempts to normalize for gene size and library depth

C: number of mapped fragments (reads) for a gene (transcript)
N: total number of mappable fragments in the library (sample)
L: gene (transcript) length in bases

$$\textbf{FPKM} = (C / (N \times L)) \times 1{,}000 \times 1{,}000{,}000$$

$$FPKM = (C / (N / 1{,}000{,}000)) / (L / 1000)$$

RNA-seq data analysis: Normalization

TPM - **T**ranscript **P**er kilobase **M**illion

TPM attempts to normalize for gene size and library depth

$$\mathrm{TPM} = \frac{(C_g * 1e^3 / L_g) * 1e^6}{\sum_{g=1}^{N}(C_g * 1e^3 / L_g)}$$

C_g: # mapped fragments for a gene (transcript)

N: # of genes in the library (sample)

L_g: gene (transcript) length in bases

1. Divide the read counts by the length of each gene in kilobases = RPK
2. Sum up all the RPK values in a sample and divide by 1,000,000 = “per million” scaling factor
3. Divide the RPK values by the “per million” scaling factor. This gives you TPM.

The **sum** of all **TPMs** in each sample is the **same**. Easier to **compare** across **samples**!

RNA-seq data analysis: Normalization

FPKM/RPKM

- **total number** of normalized **counts** is **different** in each sample
  - accounts for gene length and sequencing depth
  - **Can not compare** across samples

TPM

- **total number** of normalized **counts** is **same** in each sample
  - accounts for gene length and sequencing depth
  - **Can compare** across samples

gene	sampleA	sampleB
TP53	3.2	3.2
UTX	21.4	10.8
…	…	…
∑RPKM	1,000,000	1,600,000
3.2/1,000,000 > 3.2/1,600,000		

RNA-seq data analysis: Differential Gene Expression

DESeq2

- Comparing the counts between sample groups for the same gene
- Gene length does not need to be accounted
- **Sequencing depth** and **RNA composition** do need to be taken into account.

To **normalize** for sequencing depth and RNA composition, **DESeq2** uses the **median of ratios** method.

- **Assumption**: not ALL genes are differentially expressed
- therefore, the normalization factors
    - should account for **sequencing depth** and **RNA composition**
    - large **outlier** genes will **not represent** the median ratio values

This method is robust to imbalance in up-/down-regulation and large numbers of differentially expressed genes

RNA-seq data analysis: Differential Gene Expression

DESeq2 - normalization

1. For each gene calculate the **geometric mean** over all samples = "**pseudo-reference**"
2. Divide raw count for each gene in each sample by the corresponding "pseudo-reference" = **ratio**
3. Calculate the **normalization factor = median(ratio$_{sample}$)**
4. Divide each raw count by the normalization factor

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

gene	sampleA	sampleB	pseudo-reference (Ψref)	ratio A/Ψref	ratio B/Ψref	norm count A	norm count B
EF2A	1489	906	$(1489 \times 906)^{½} = 1161.5$	1489/1161.5 = 1.28	906/1161.5 = 0.78	1489/1.3 = 1145.39	906/0.77 = 1176.62
ABCD1	22	12	$(22 \times 13)^{½} = 16.9$	22/16.9 = 1.30	13/16.9 = 0.77	22/1.3 = 16.92	13/0.77 = 16.88
…	…	…	…	…	…	…	…
median(ratio)				**1.3**	**0.77**		

RNA-seq data analysis: Differential Gene Expression

DESeq2 - GLM (generalized linear model)

Assumption:

- Count values for a gene in sample *j* is generated by NB (negative binomial) distribution
  - with mean μ_{ij}
  - and dispersion α_i.
- dispersion α_i = variance deviation from mean

$$c_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

Null hypothesis:

- All samples have the same μ.

Alternative hypothesis:

- Mean (μ) is the same only within the same group

RNA-seq data analysis: Differential Gene Expression

DESeq2 - GLM

$$\log_2(q_{ij}) = \beta_0 + x_j\beta_\tau$$

with $x_j = 0$ if j is control, $x_j = 1$ if j is treatment sample

q_{ij} … Expected Relative Abundance: This is the biological "signal" of transcript *i* in sample *j*.
s_{ij} … Size Factor: Represents technical variation (sequencing depth) for sample j.
μ_{ij} … Expected Raw count for gene *i* in sample *j*.

$$c_{ij} \sim NB(\mu_{ij}, \alpha_i)$$

$$\mu_{ij} = s_{ij} \times q_{ij}$$

$$q_{ij} = 2^{\beta 0 + xj\beta\tau}$$

$$\mu_{ij} = s_{ij} \times 2^{\beta 0 + xj\beta\tau}$$

Calculate the coefficients β that fit best the observed data.

Is β_τ significantly different from null? Can we reject the null hypothesis?

Wald test is used in DESeq2 and reports a **p-value** that indicates if the observed difference between treatment and control (β_τ) is effect a true of the treatment (small p.)

RNA-seq data analysis: Differential Gene Expression

DESeq2 - multiple hypothesis testing problem

- p-value:
  - gene with a significance cut-off of $p < 0.05$, means a **5% chance** of being a **false positive** (FP)
- We test **20,000 genes** with a significance cutoff of 0.05
  - we can expect **1,000** genes that are **FP**
- If we found 3,000 genes differentially expressed ~ ⅓ of them are FP

- p-value is a result from a single test, the **more tests** we do the **more FP** we can expect

RNA-seq data analysis: Differential Gene Expression

DESeq2 - multiple hypothesis testing - p-value correction

There are a few common approaches for multiple test correction:

- Bonferroni
  - $\text{p-value}_{\text{adjusted}}$ = p-value * m, where m = total number of tests
  - very conservative, may lead to many false negatives
- FDR/Benjamini Hochberg
  - False Discovery Rate, controls the expected FDR below a specified level
    1. sort p-values in ascending order
    2. assign ranks to each p-value, starting from the lowest (1, 2, 3, …)
    3. define a FDR cutoff (e.g. 0.1)
    4. calculate adjusted p-value

$$\textbf{p-value}_{\textbf{adjusted}} = \textbf{(rank / \# tests) * FDR}$$