ADER17

Analysis of Differential Expression with RNAseq

Instructors

Daniel Sobral

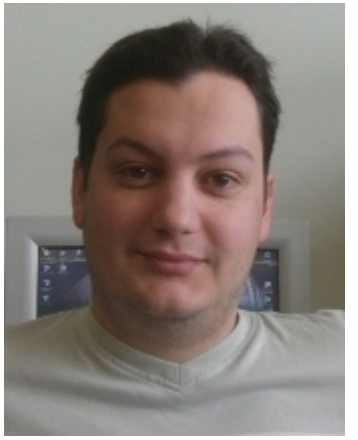
Head of the

Bioinformatics Unit at IGC

Mauro Truglio

Bioinformatics Specialist in the

Bioinformatics Unit at IGC



Daniel Faria

Postdoctoral fellow in the

Bioinformatics Unit at IGC

Learning Objectives

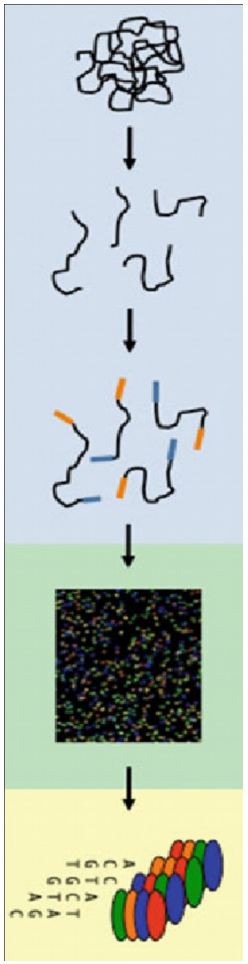
At the end of the course, we expect to create enhanced capabilities to:

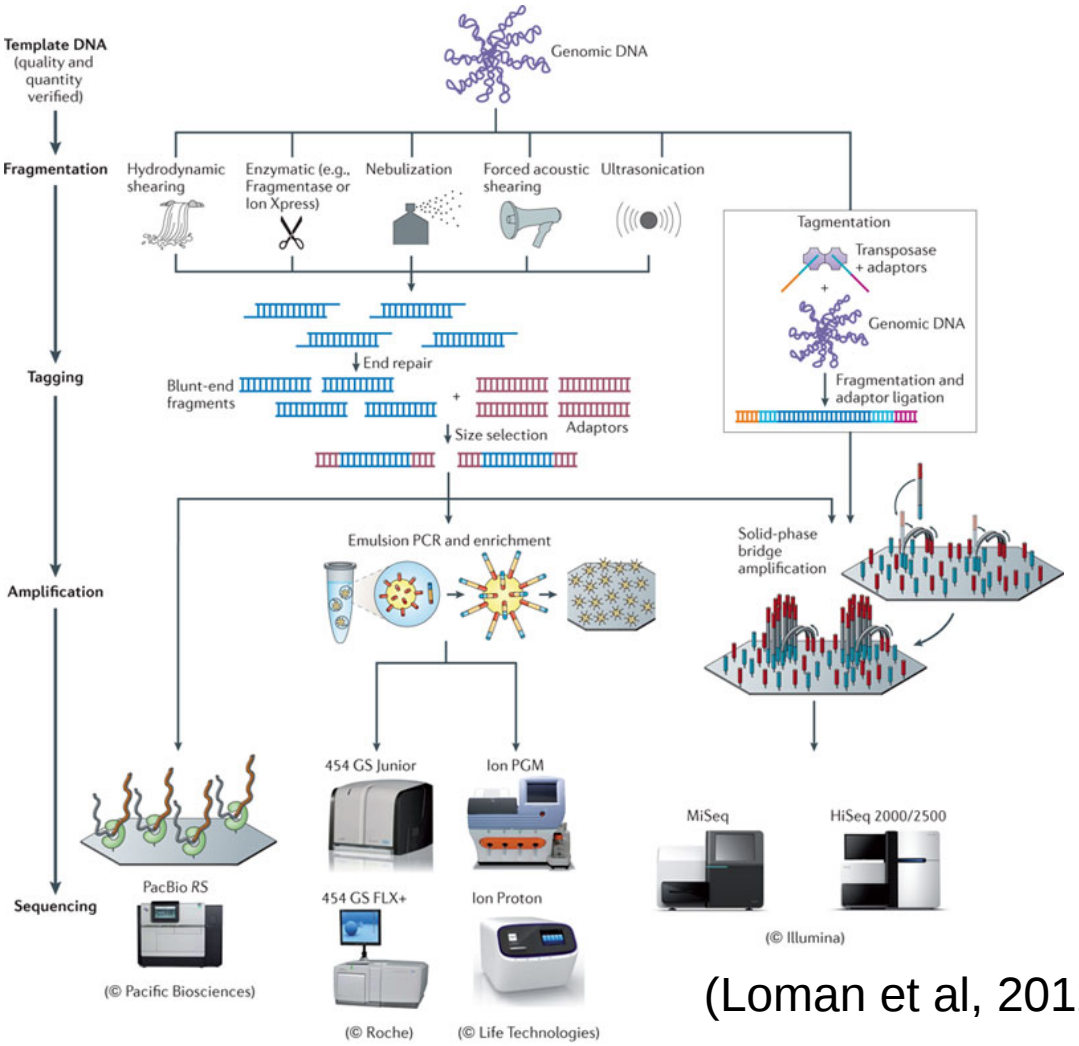
1. List characteristics of NGS technologies and choose adequate sequencing
2. Have broad overview of steps in analysis of RNA-Seq differential expression experiments
3. Assess the general quality of the raw data from the sequencing facility
4. Do simple processing operations in the raw data to improve its quality
5. Generate alignments against a reference genome
6. Assess the general quality of the alignments and detect possible problems
7. Generate tables of counts using the alignment and a reference gene annotation
8. Generate lists of differentially expressed genes, at least for a simple pairwise comparison
9. Perform simple functional enrichment analysis and understand the concepts behind them

Learning Outcome 1

Understand
NGS technologies
and plan your experiment

Overview of NGS technologies





(Loman et al, 2012)

Obtaining your “DNA”

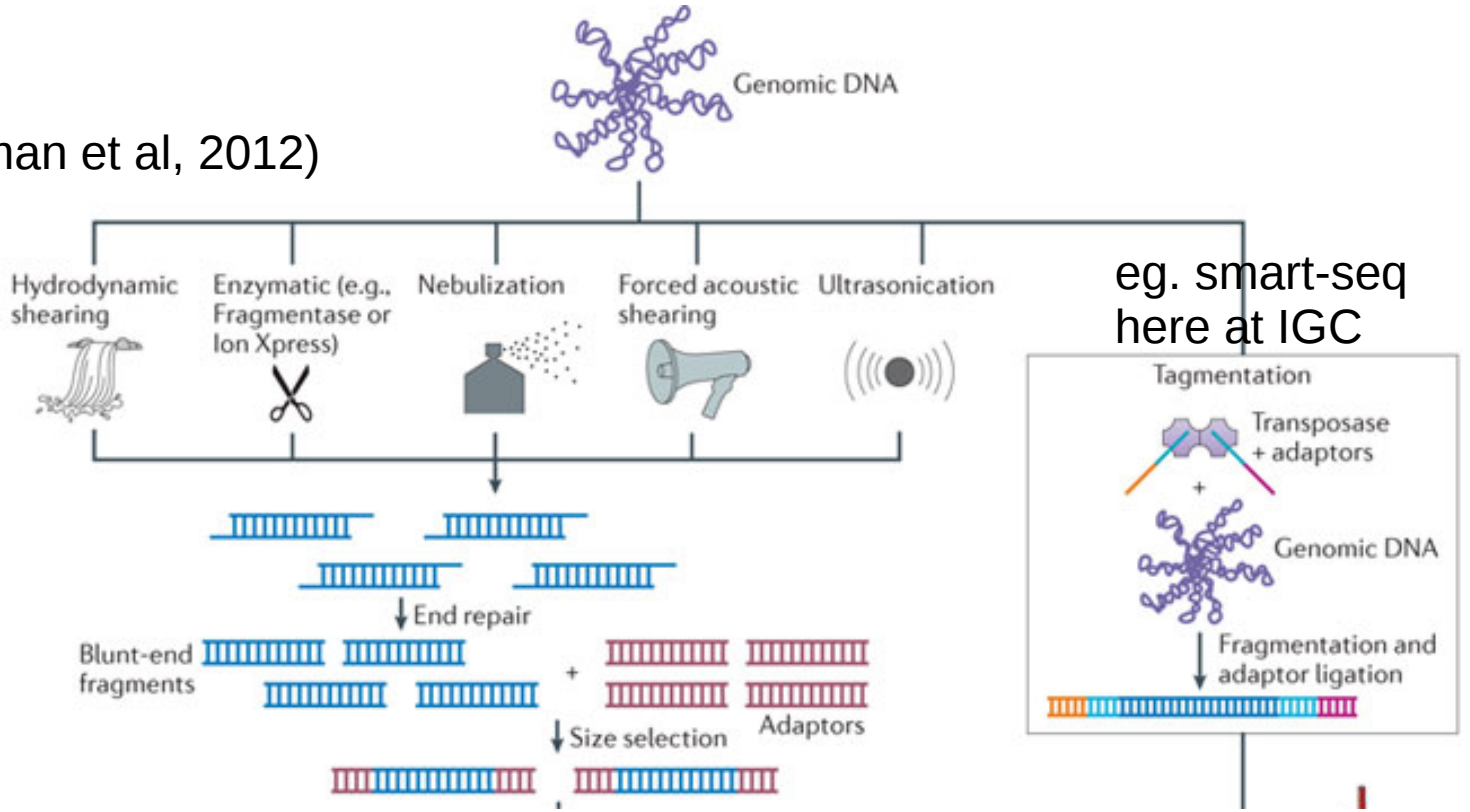
- PolyA+ RNA,
  - Less noisy, only mRNAs
    - ~single-cell becoming popular (eg. smart-seq here at IGC)
  - polysomal RNA
    - Enriches for mRNAs, but can bring other stuff
  - Stranded or not (eg. dUTP marking)

- Total RNA
  - Has lots of information (small RNA, etc..), but also “junk”
  - Random (not really!) priming
  - Nuclear ribosome-depleted RNA essential (otherwise all signal will come from this)

- Then, to be read, RNA needs to be converted (cDNA)
  - Often (almost always?) with amplification – possible PCR artifacts

Breaking and tagging the “DNA”

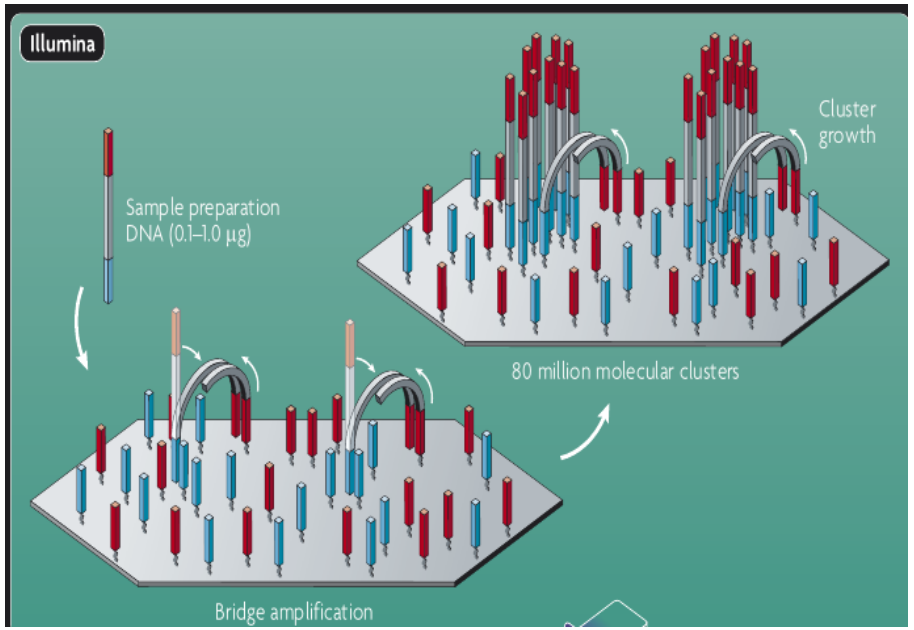
(Loman et al, 2012)



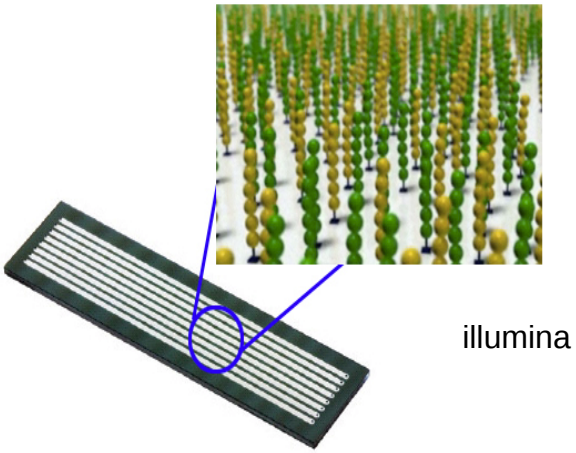
You may need to take this into consideration when interpreting your results

Sequencing with Illumina

The most commonly used technology at the moment



(Metzker, 2009)

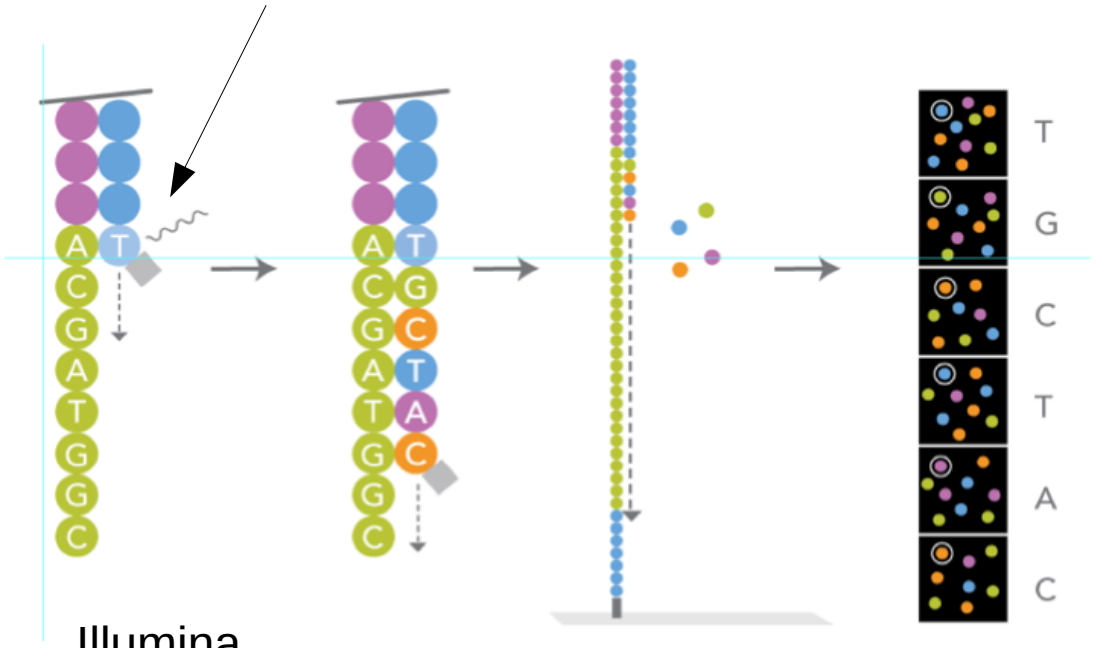


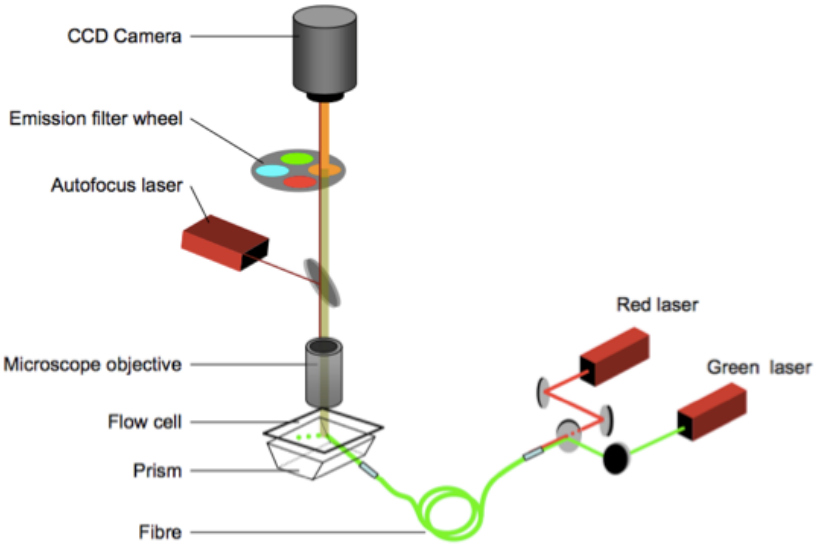
illumina

MILLIONS of sequences simultaneously!

Sequencing with Illumina

Terminator nucleotides like with Sanger

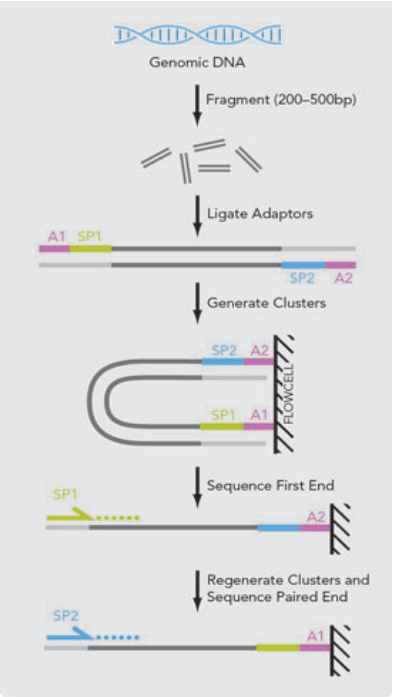


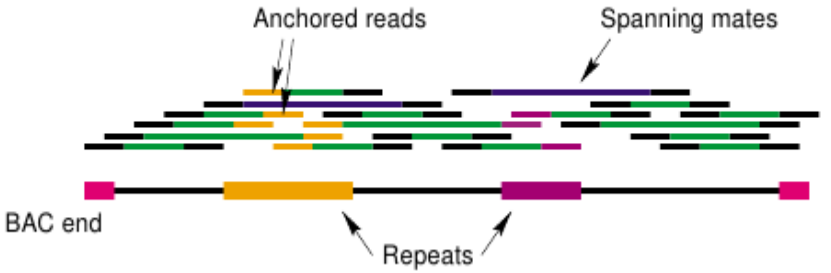


Illumina

Single vs Paired-end Reads

Paired-end reads are an alternative to longer reads





pairs act as long fragments

Learning Outcome 1

Parameters to consider for Library Preparation:

- mRNA; Total RNA (rRNA depleted?); (un)Stranded?
- How many replicates (do we need technical replicates?)

Parameters to consider for Sequencing Facility:

- Single-End or Paired-End
- Read Length: (50-150bp are most common)
- How much to sequence per sample (coverage)
  - Also related to number of replicates

RNA-seqlopedia

http://rnaseq.uoregon.edu

“Long reads, paired-end reads, and stranded library preparation methods are not as important for DGE especially if a reference genome is available. Instead DGE experiments need to focus time and expense on replicates in order to obtain accurate measures of variances

...

10M reads as bare minimum to get genes with above average expression levels”

ENCODE recommendations (2011)

- https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf
  - "Experiments whose purpose is to evaluate the similarity between the transcriptional profiles of two polyA+ samples may require only modest depths of sequencing (e.g. 30M pair-end reads of length > 30NT, of which 20-25M are mappable)"
  - "The ability to detect reliably low copy number transcripts depends upon depth of sequencing and on a sufficiently complex library. For experiments from a typical mammalian tissue or in which sensitivity of detection is important, a minimum depth of 100-200M, 2x76bp"
  - "Should have 2 or more biological replicates"
    - Correlation of Biological Replicates should be between 0.92 and 0.98 (Pearson R^2)

Coverage (ENCODE 2011)

How can I know how much sequence I need?

- To estimate the sequence coverage per mRNA of an average length (ignoring that there is actually a broad length distribution) present at 1 copy per cell based on an estimated input of the number of mRNAs the following calculation can be used: (Total sequence NT in the sequencing reaction / Estimate of the Number of Molecules of mRNA/cell) /(1,500NT/mRNA).
- Example: 10^10 nucleotides sequenced / 2X10^6 mRNAs/cell = 5X10^3NT sequence coverage per/mRNA.
- 5X 10^3 NT / 1.5 X^3 NT/mRNA ~3X sequence coverage of an RNA present at one copy per cell.

In practice, you never know, unless you do a pilot

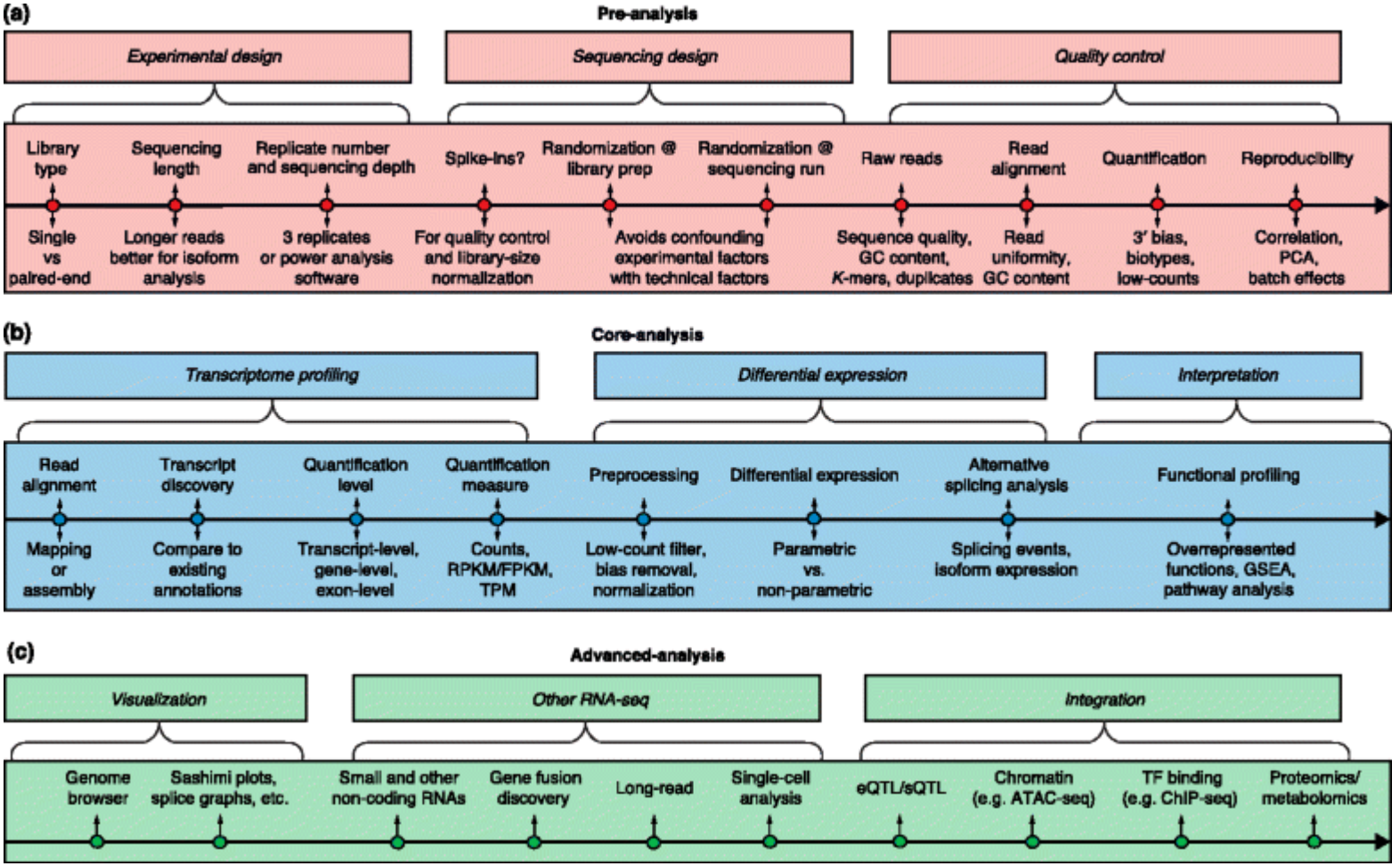
http://scotty.genetics.utah.edu/scotty.php

Learning Outcome 2

List steps in the analysis of RNA-Seq:

- QC of Raw Data; (LO 3)
- Preprocessing of Raw Data (if needed); (LO 4)
- Alignment of “clean” reads to reference genome (LO 5)
- QC of Aligments (LO 6)
- Generate table of counts of genes/transcripts (LO 7)
- Differential Analysis tests (LO 8)
- Post-analysis: Functional Enrichment (LO 9)

RNA-Seq workflow



(Conesa et al. 2016)

Learning Outcome 3

Assess general quality of raw data from sequencing facility

LO 3.1 - Interpret fastq files and their content

LO 3.2 - Use software like FastQC to produce QC reports

LO 3.3 - Read QC reports of raw data to assess the general quality of data and presence of sequence bias

Learning Outcome 4

Do simple processing operations in the raw data to improve its quality

LO 4.1 - Use tools such as seqtk and trimmomatic to remove low quality bases from your reads

LO 4.2 - Use tools such as cutadapt to remove adaptors and other artefactual sequences from reads

Learning Outcome 5

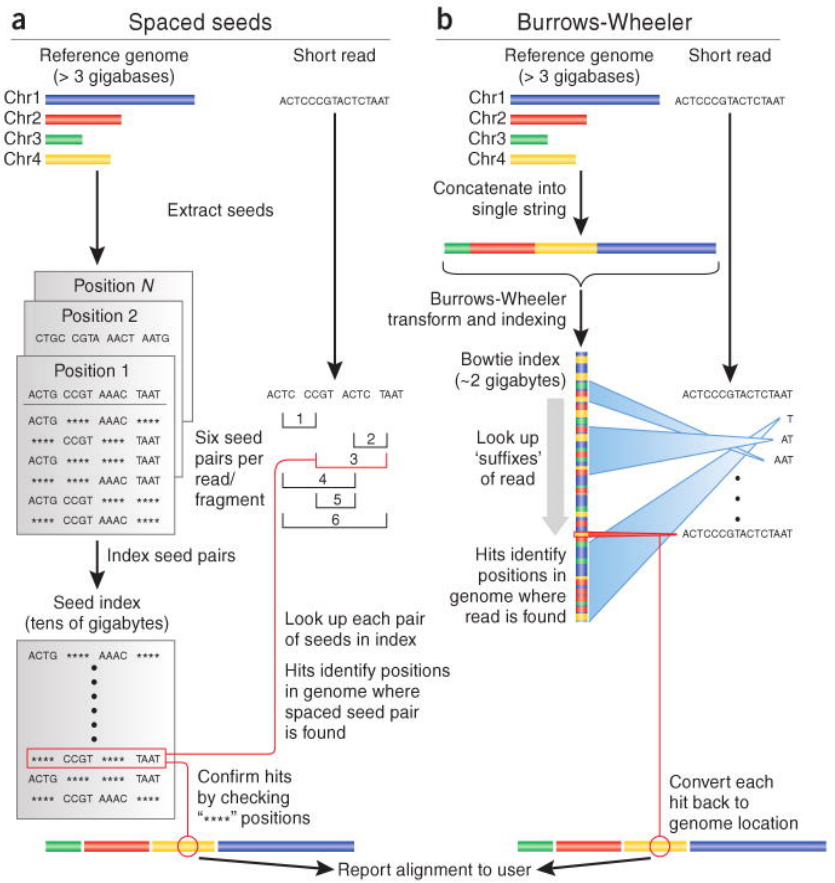
- Generate alignments of processed reads against a reference genome

  LO 5.1 - What is a reference genome, versioning and where to obtain genomes

  LO 5.2 - Alignment software: tophat2/hisat2; bwa; sailfish/salmon

  LO 5.3 - Run an alignment: the SAM/BAM alignment format

LO 5.2 - Alignment software: tophat2/hisat2; bwa; sailfish/salmon



(Trapnell & Salzberg, 2009)

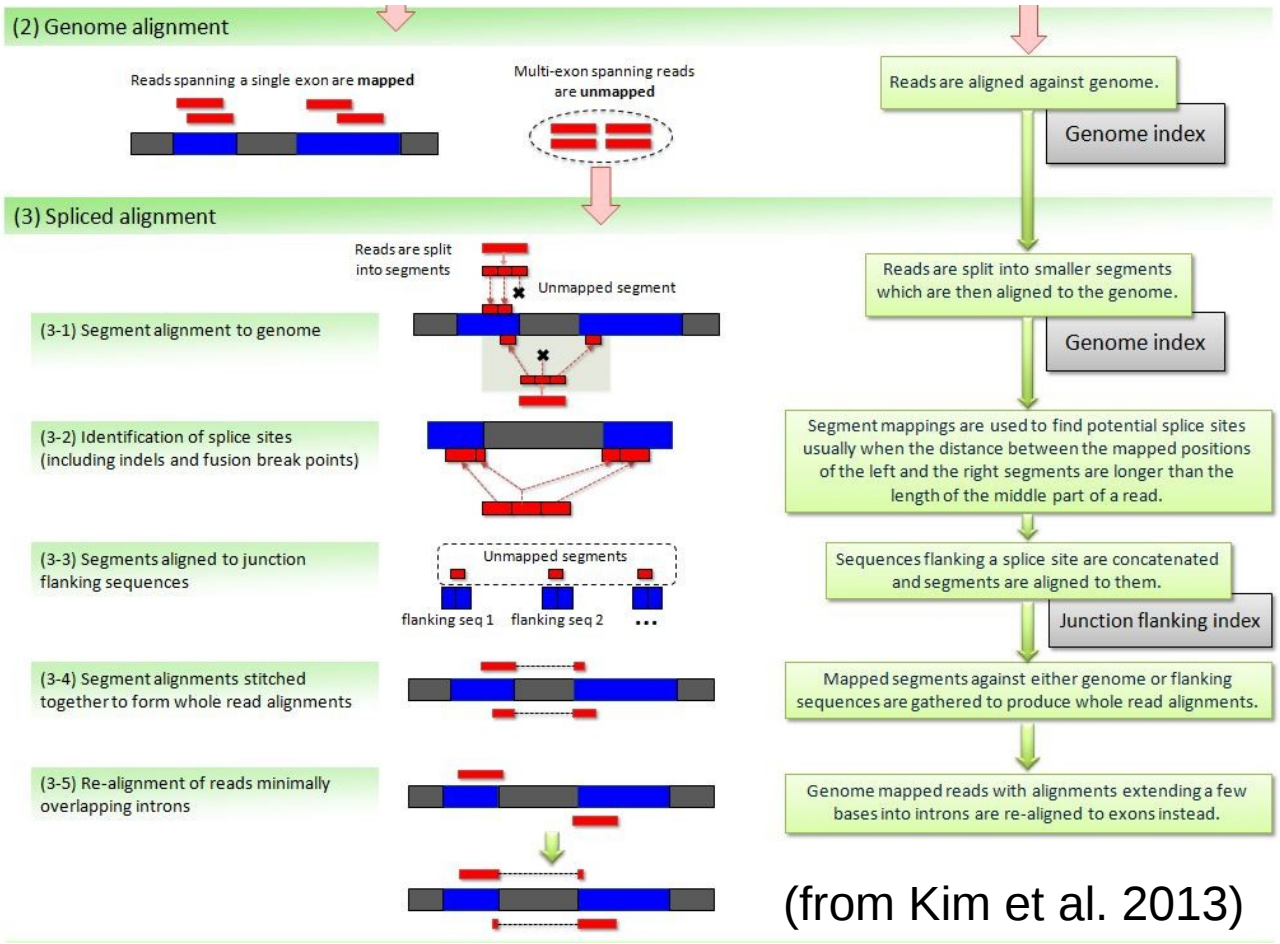
How to align millions of small reads to a large genome?

Running blast would be just too slooooow! (besides other issues)

With some assumptions we can speed up things

eg. Burrows-Wheeler aligners (Li & Durbin, 2009)

Spliced Alignment



(from Kim et al. 2013)

Learning Outcome 6

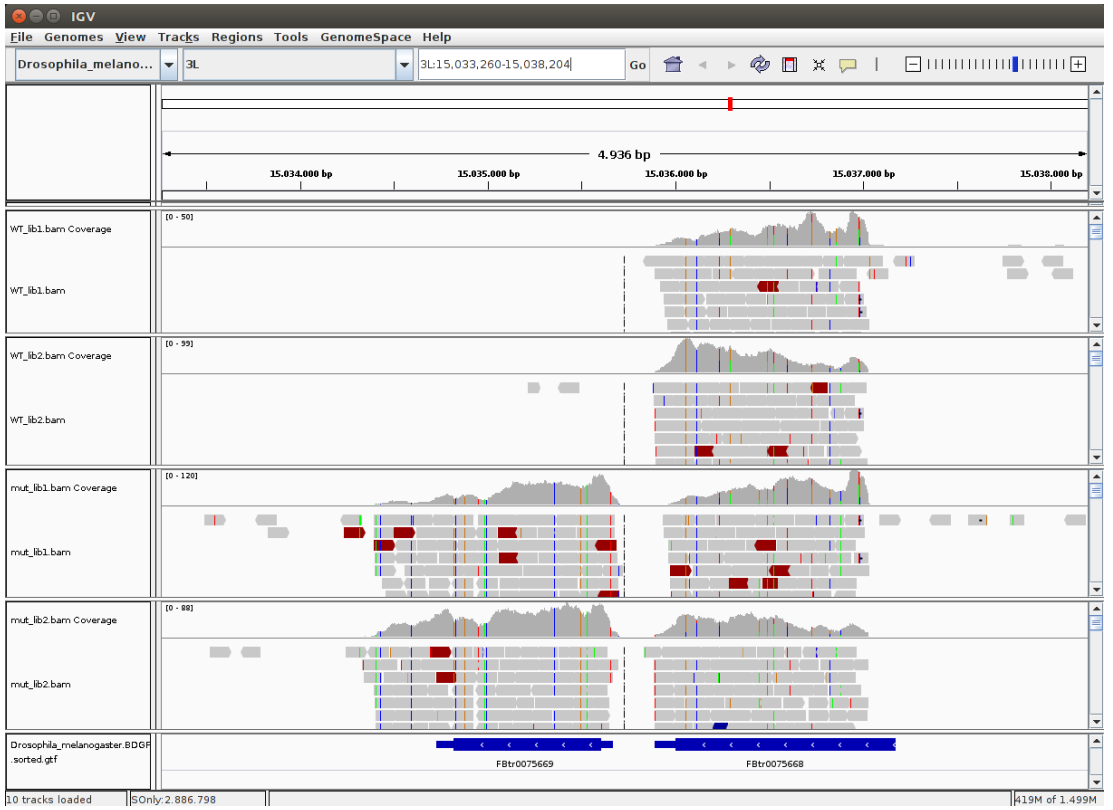
Assess the general quality of the alignments and detect possible problems

LO 6.1 - Visualizing alignments in IGV for single genes

LO 6.2 - Use tools such as RSeQC and Qualimap to assess quality of alignments

Visualizing results

IGV to view alignments in specific areas

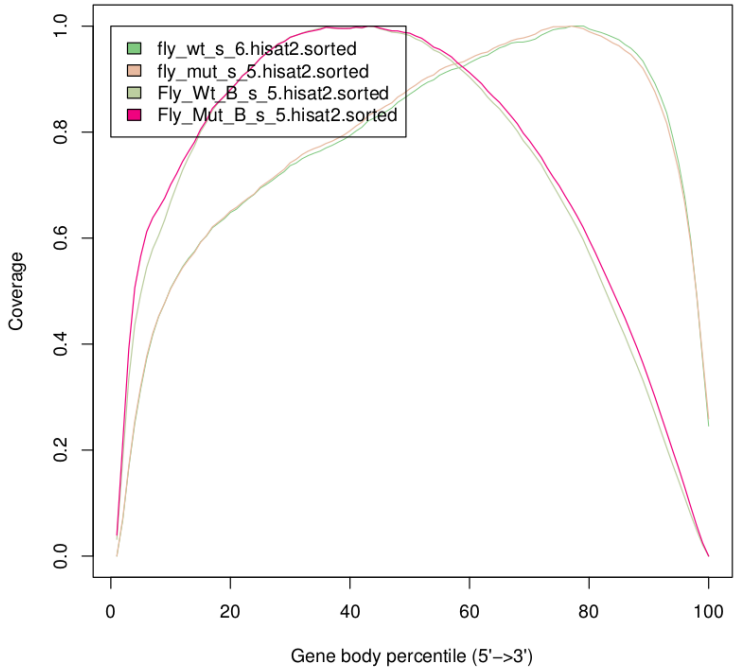


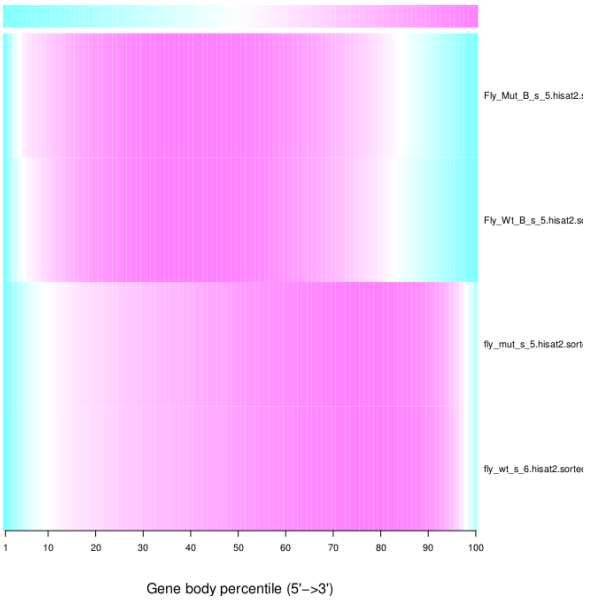
Assess quality of alignments

- Number of mapped reads
- Duplicates (how many do we expect?)
- Coverage of Reads along Gene
- Are reads mapping against known annotation?

Assess quality of alignments

- Coverage of Reads along Gene



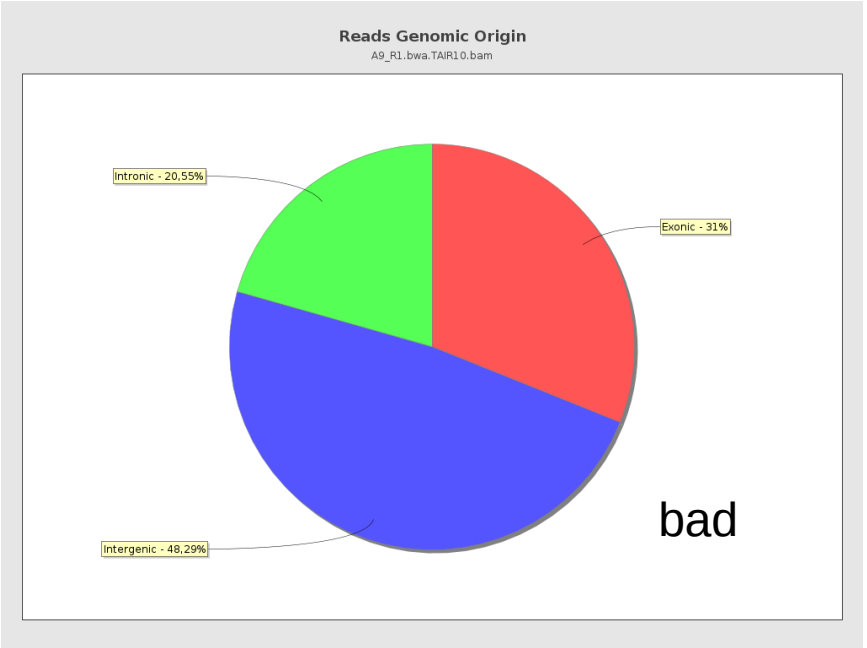


data from guilgur et al.

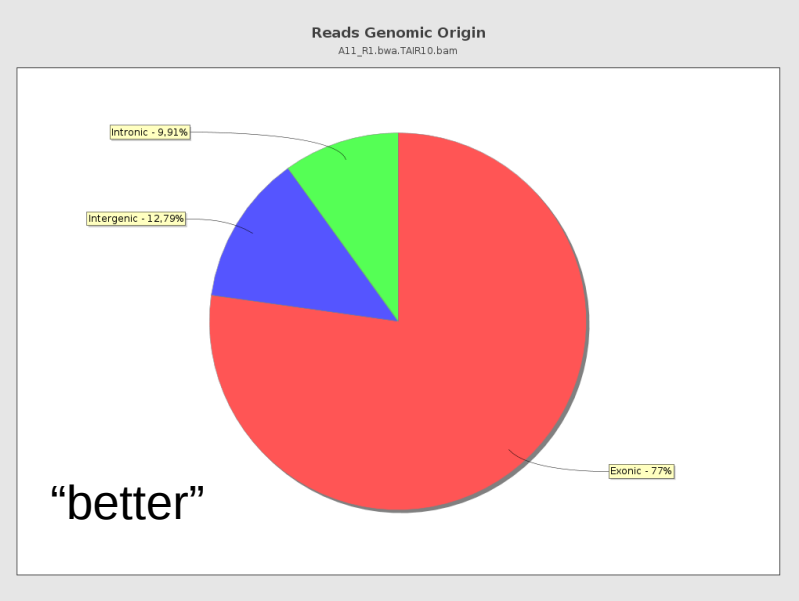
Assess quality of alignments

- Are reads mapping against known annotation?

<20% alignments



~80% alignments



data from jorg becker’s group (unpublished)

Learning Outcome 7

Generate tables of counts using the alignment and a reference gene annotation

- LO 7.1 - What is a reference gene annotation, versioning and where to obtain
- LO 7.2 - The process of generating gene counts from genome aligments
- LO 7.3 - Use tools such as htseq-counts and Qualimap to generate table of gene counts

Learning Outcome 8

Generate lists of differentially expressed genes, at least for a simple pairwise comparison

LO 8.1 - Using the R package edgeR to produce a pairwise differential expression analysis

LO 8.2 - Interpretation and visualization of results

LO 8.3 - Use more complex settings: Generalized Linear Models

Normalization (DESeq)

http://seqanswers.com/forums/showthread.php?t=586 (Simon Anders)

- “To estimate the library size, simply taking the total number of (mapped or unmapped) reads is, in our experience, not a good idea. Sometimes, a few very strongly expressed genes are differentially expressed, and as they make up a good part of the total counts, they skew this number. After you divide by total counts, these few strongly expressed genes become equal, and the whole rest looks differentially expressed.

- The following simple alternative works much better:

  - Construct a "reference sample" by taking, for each gene, the geometric mean of the counts in all samples.

  - To get the sequencing depth of a sample relative to the reference, calculate for each gene the quotient of the counts in your sample divided by the counts of the reference sample. Now you have, for each gene, an estimate of the depth ratio.

  - Simply take the median of all the quotients to get the relative depth of the library.

  This is what the 'estimateSizeFactors' function of our DESeq package does.“

  In edgeR you can do method="RLE" in estimate size factors (as alternative to TMM)

Normalization

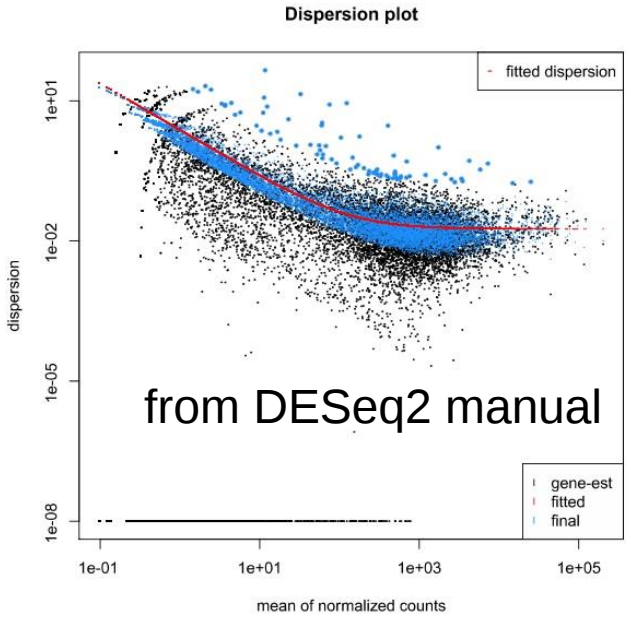
- The previous normalization is to compare the same gene between samples
- For comparing gene expression between genes in one sample, we need to normalize for gene length and other factors such as GC content
  - Eg. RPKM (single-end) or FPKM (paired-end)
    - Simply: CPM (counts per million) / Gene Length
  - These measures should NOT be used as input in the differential expression analysis

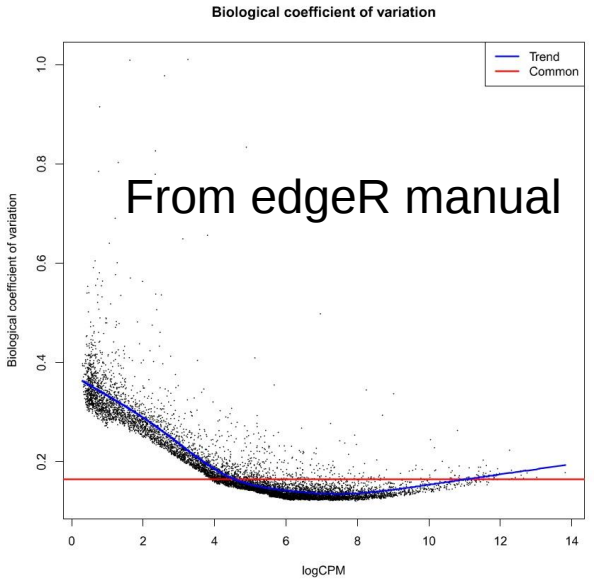
Estimating Variation

- Variation is how much (normalized) counts vary between the different samples
- This variation is clearly gene dependent
  - Highly expressed genes vary more in terms of absolute value, and low expressed genes vary more in terms of % of gene expression (fold change)
  - If you only look at fold change without taking variation into account, you're more likely to have low expressed genes as differentially expressed

Estimating Variation

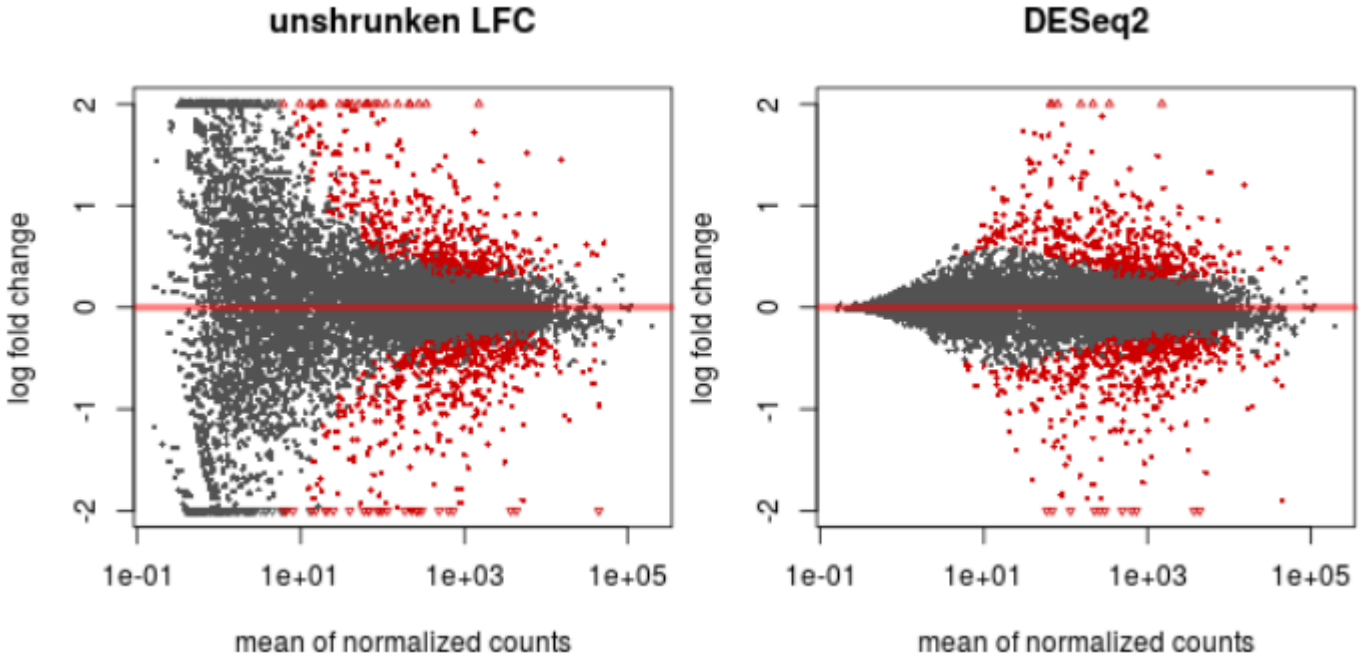
- We need to accurately estimate variation per gene
  - But we usually do not have enough replicates to do this
  - We can bin genes with similar expression and fit curve





Effect of considering variation

- “True” fold change is calculated based on curve



(from DESeq2 manual)

Multiple Testing and Filtering

- Each gene is tested individually, and since we test thousands of genes, some genes get good p-values just by chance
  - p-values need to be corrected for multiple testing. One way is to multiply with number of tests (Bonferroni), but this is too strict. A popular method is Benjamini-Hochberg (calculates a False Discovery Rate)

  (Daniel Faria will come back to this later)

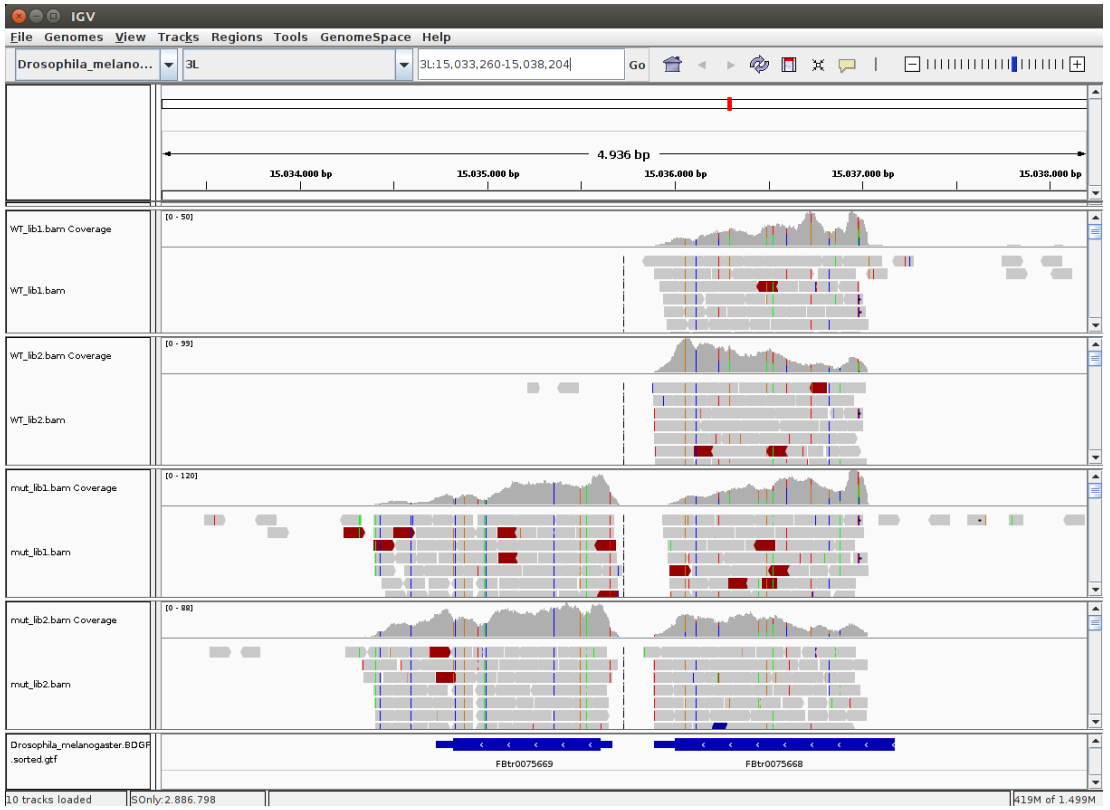
- To reduce number of genes, filter out genes not expressed or at very low level
  - Reduces problem of multiple testing
  - Should not be biased to any group of samples

(Lin et al., 2016) Drosophila

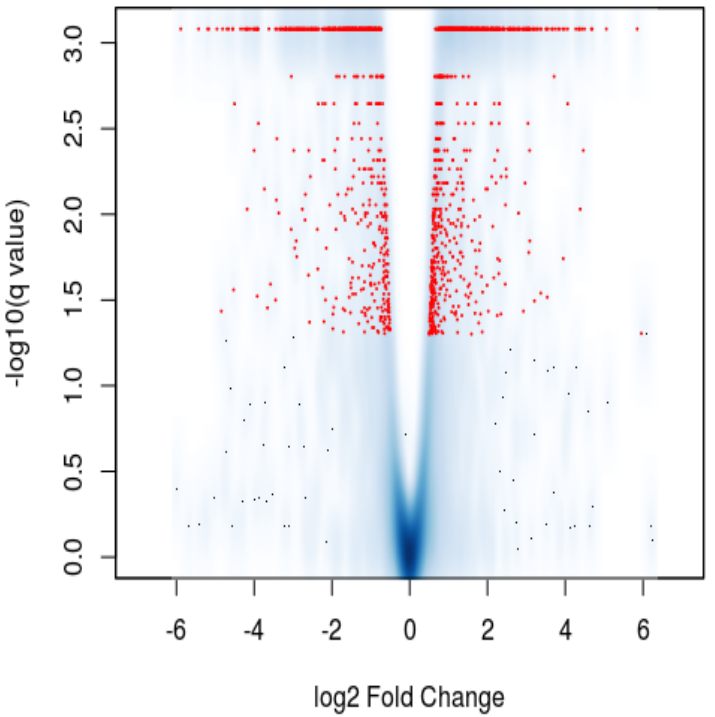
- The best analysis approach to our data was to normalize the read counts using the DESeq method and apply a generalized linear model assuming a negative binomial distribution using either edgeR or DESeq software. Genes having very low read counts were removed after normalizing the data and fitting it to the negative binomial distribution. We describe the results of this evaluation and include recommended analysis strategies for RNA-Seq read count data.
- In addition, at least three biological replicates per condition were required in order to have sufficient statistical power to detect expression differences among the three-way interaction of genotype, environment, and sex.

Visualizing results

IGV to view alignments in specific areas

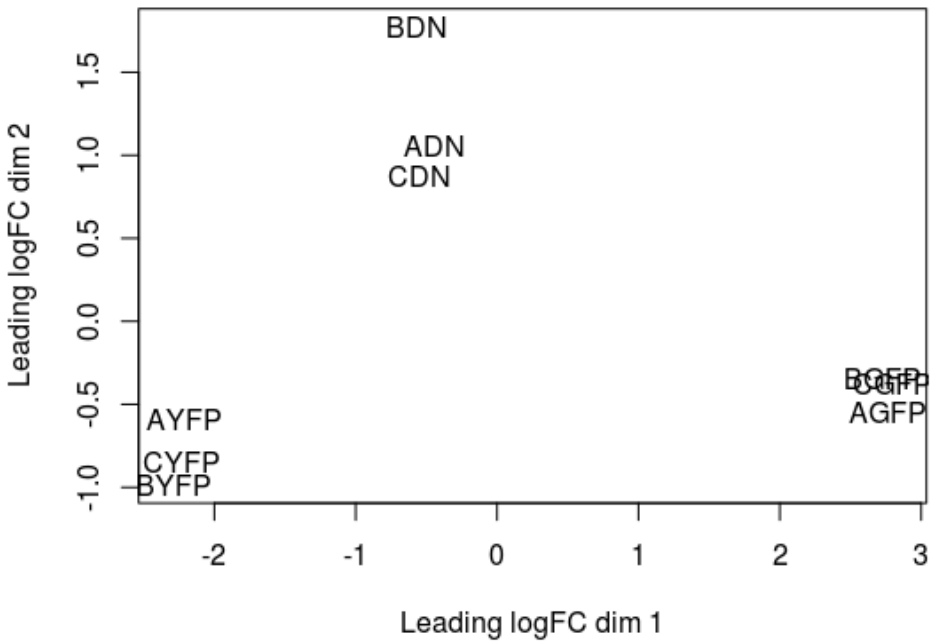


Volcano Plot

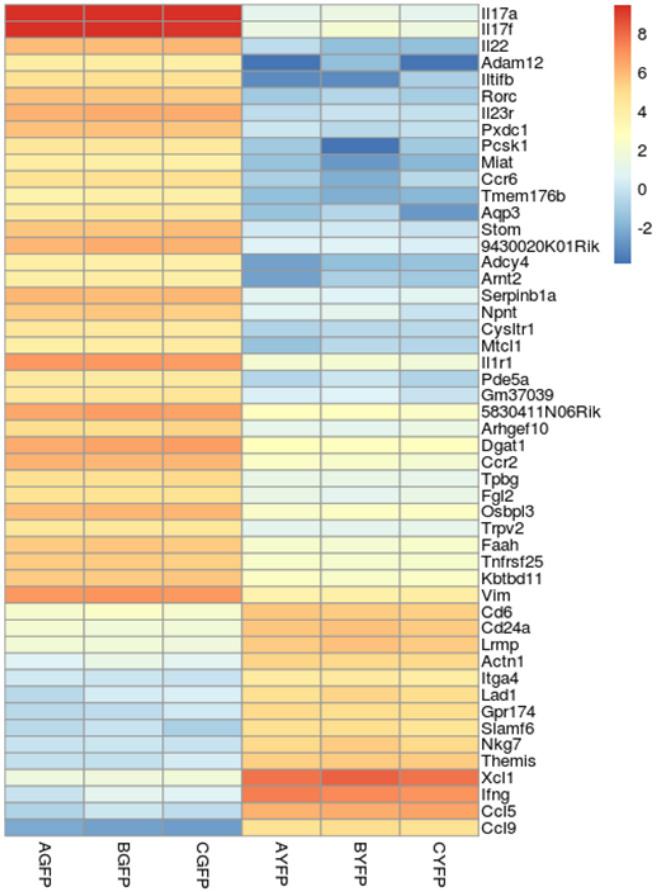


Visualizing results

PCA plots of samples



Expression of Selected Genes



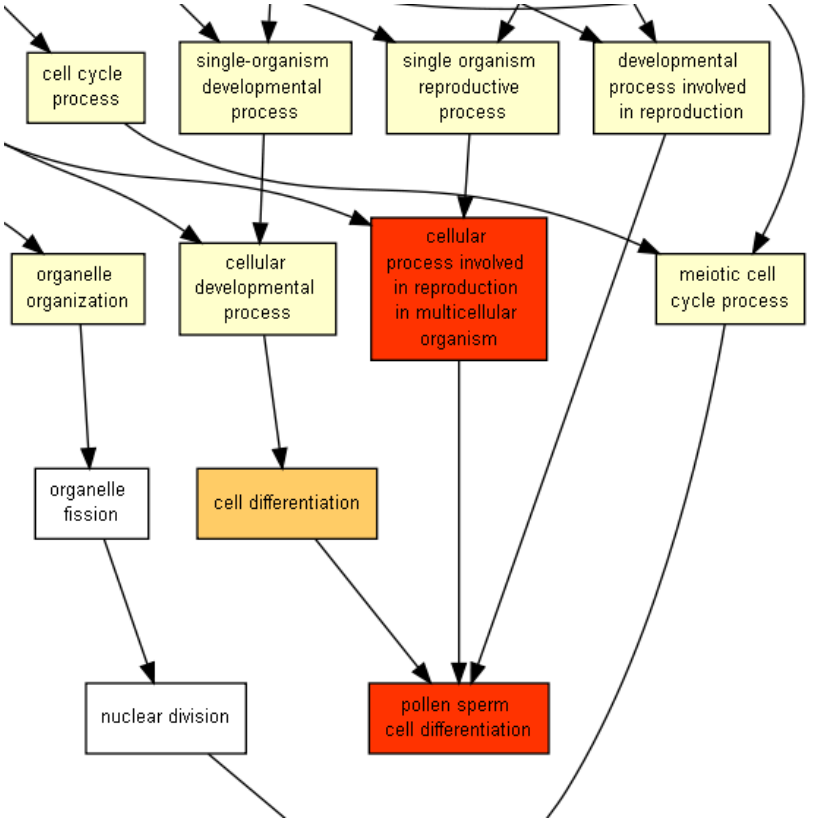
Learning Outcome 9

Perform simple functional enrichment analysis and understand the concepts behind them

- LO 9.1 - The statistics behind functional enrichment analysis
- LO 9.2 - Functional annotations: what are these and where to get them
- LO 9.3 - Using functional enrichment analysis with your lists of genes

Functional Enrichment Analysis

GO term	Description			
	pollen sperm cell differentiation	1.92E-12	6.75E-9	50.10 (18186,11,264,8)
	cellular process involved in reproduction in multicellular organism	1.65E-11	2.9E-8	29.45 (18186,14,397,9)
	DNA metabolic process	4.6E-10	5.4E-7	2.89 (18186,146,1854,43)
	negative regulation of biological process	1.17E-8	1.03E-5	2.71 (18186,190,1482,42)
	cellular response to DNA damage stimulus	2.15E-8	1.51E-5	3.29 (18186,99,1730,31)

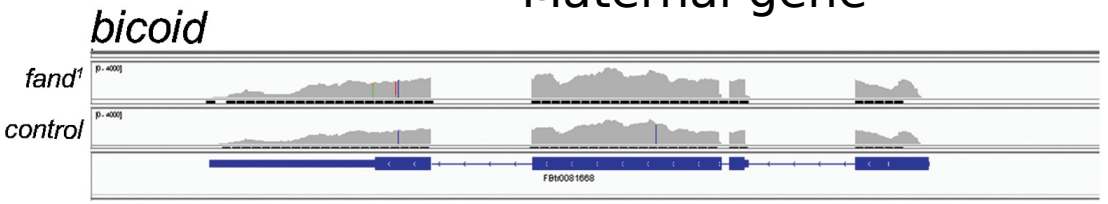


Single Cell data – Arabidopsis Thaliana Polen Sperm

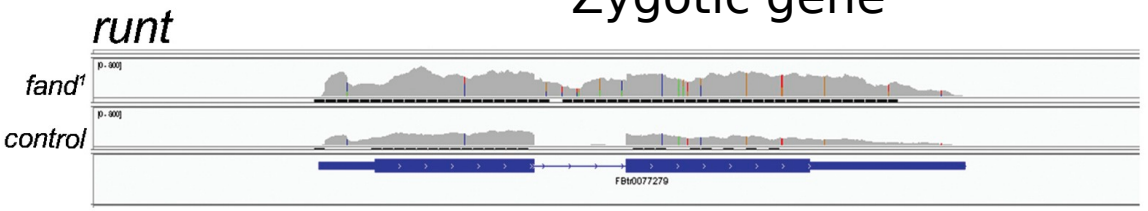
A few EXTRAS

Differential Analysis of Splicing

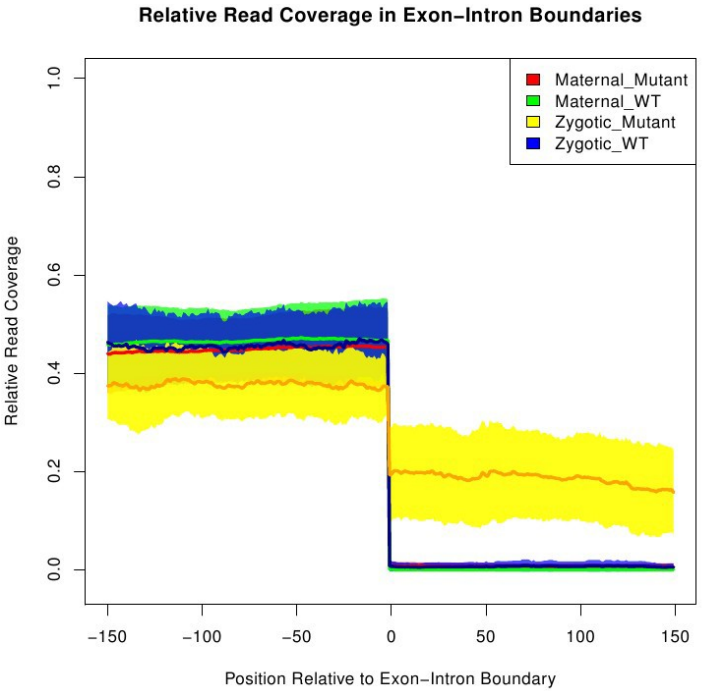
Maternal gene



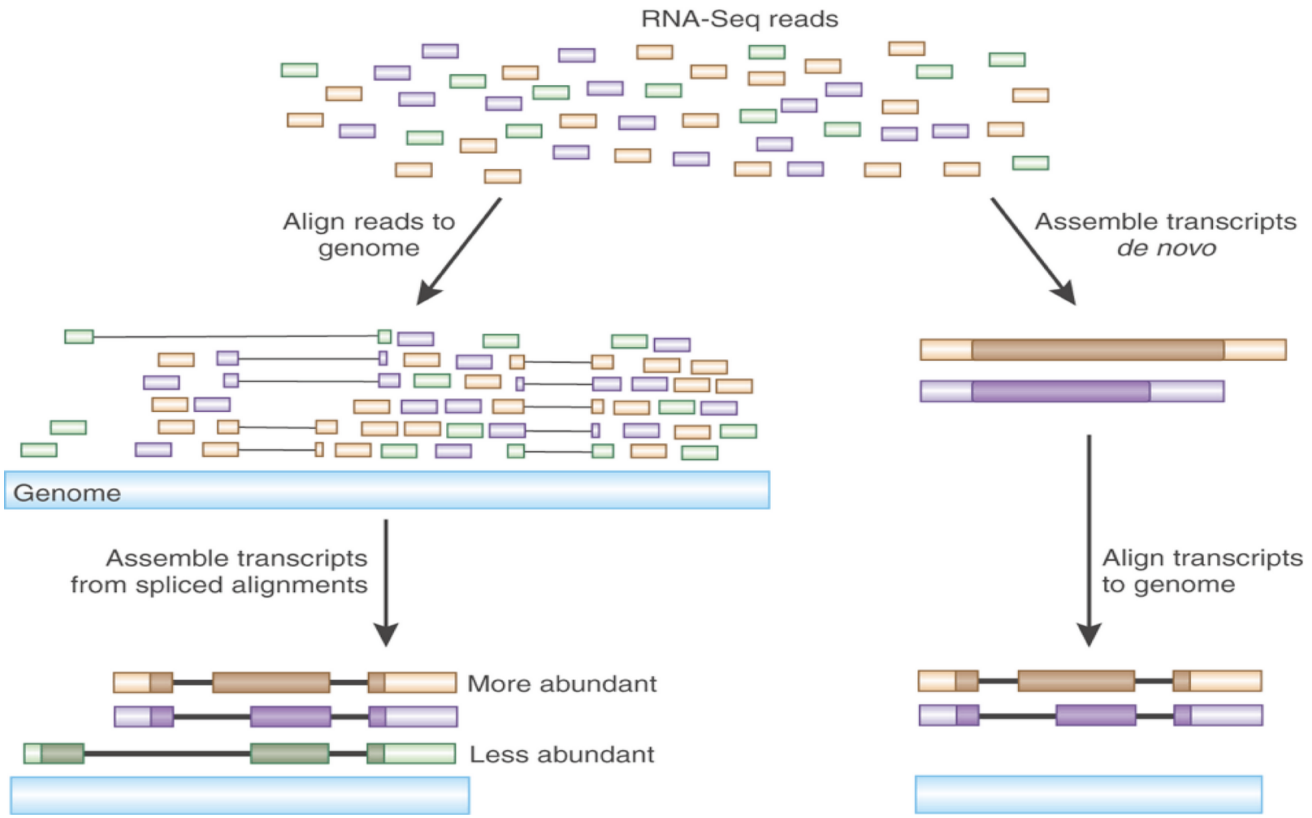
Zygotic gene



(Guilgur et al, 2014)



Denovo assembly of transcriptomes



Assembling transcripts denovo is even harder than the genome

Finding alternative Transcripts with RNA-Seq is a very active research area

MUCH easier if genome is already available

You still need the painstaking work done before NGS