RNA seq: differential expression analysis

For INF-BIO 4121/9121
Fall semester 2015

Monica Hongrø Solbakken
m.h.solbakken@ibv.uio.no

UiO : Centre for Ecological and Evolutionary Synthesis
University of Oslo

Outline II

- The RNAseq module day 11
  - Recap differential expression
- Today we will cover
  - Continue on differential expression analysis
    - Independent time series (edgeR)
    - Nested time series (DESeq)
  - GO:annotation + differential expression = GO enrichment analysis

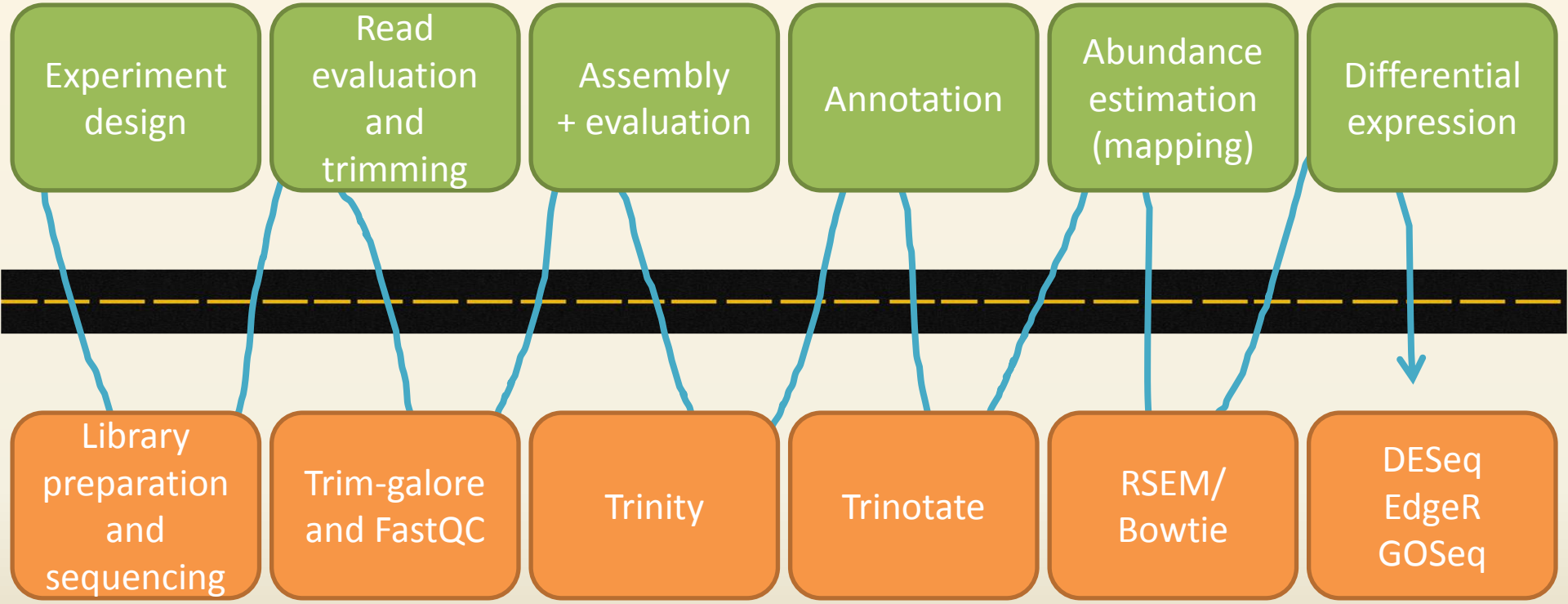
Theoretical aims

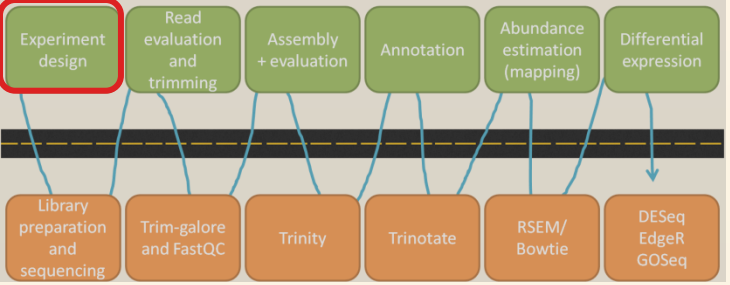
- Give the definition of a trancriptome
- To understand that RNAseq is only a proxy for biological function
- To understand how to choose a transcriptomics strategy – what do I have to consider?
- To understand the multifaceted nature of experimental design
- To understand the underlying biases and assumptions in RNAseq and RNAseq statistics
  - Not detailed assembly algorithms

Methodological ims

- Learn how to evaluate RNAseq data
- Learn how to evaluate a transcriptome assembly
- Learn about assembly annotation
- Learn how to do abundance estimation
- Learn how to do simple differential expression analyses
- Try out more complex differential expression analyses
- Experience the differences between two of the most popular DE analysis packages

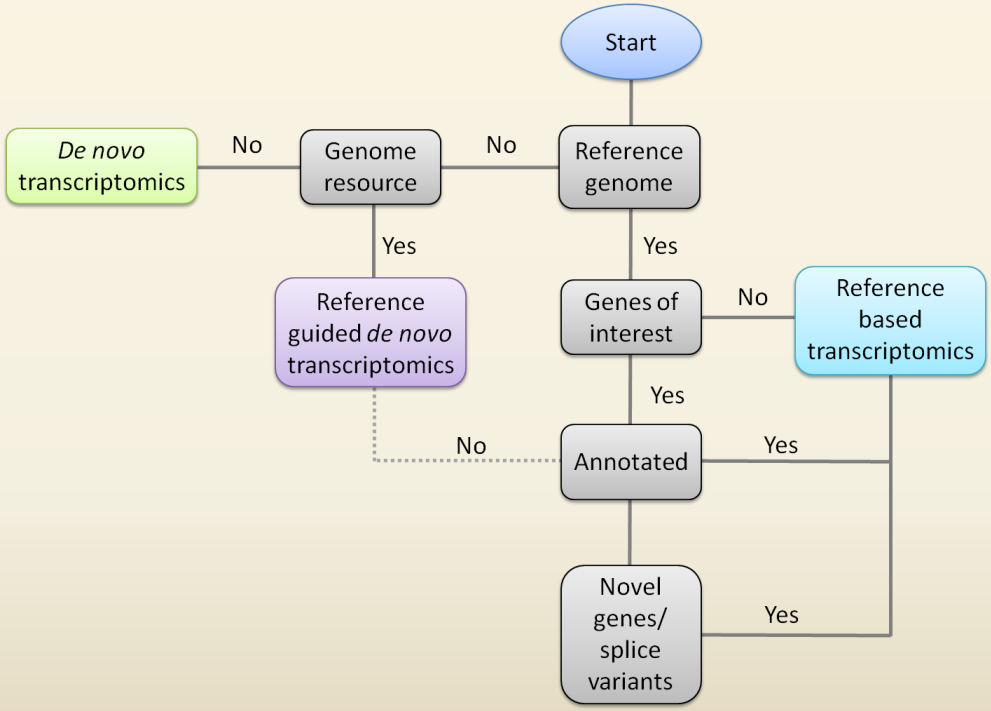
The flow of RNAseq

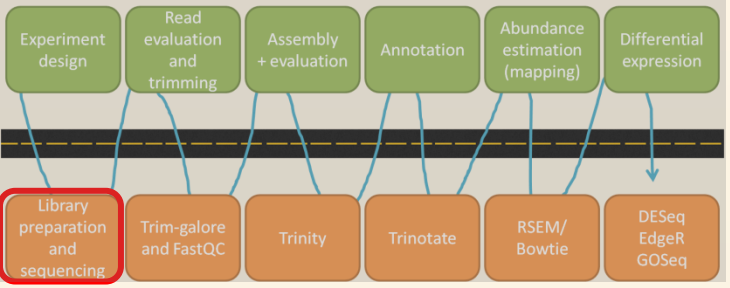




Recap – experimental design

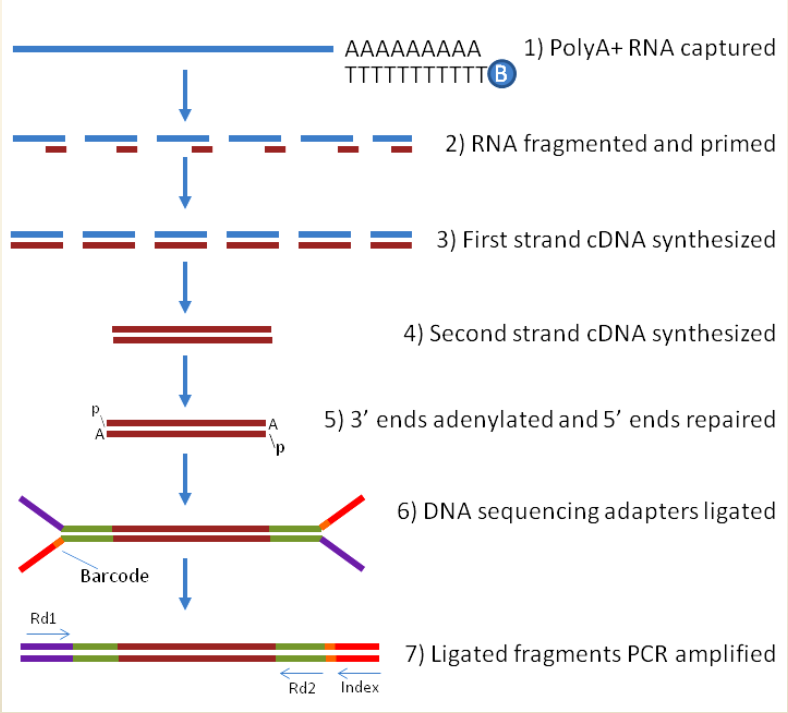
- What do you have, want and need?
- Model organism?
- Contrast or multi-factor?
- Replicates!
- Sequencing depth!



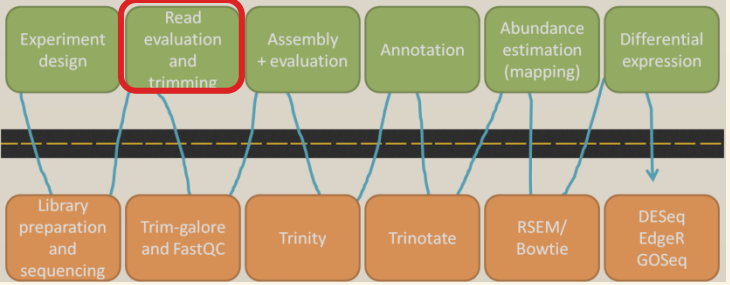


Recap – library prep and sequencing

- Long, medium or short read technology?
- Paired end or single read?
- TotalRNA or RNA subset?

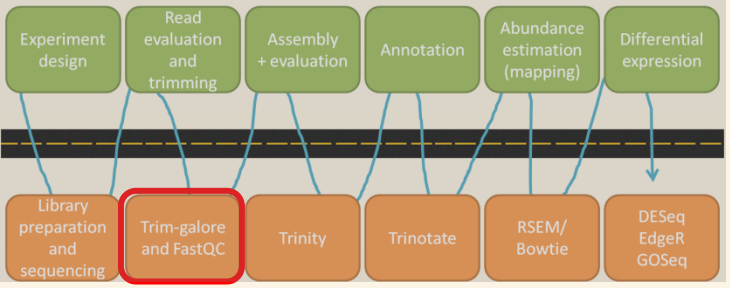


Illumina HiSeq 4000



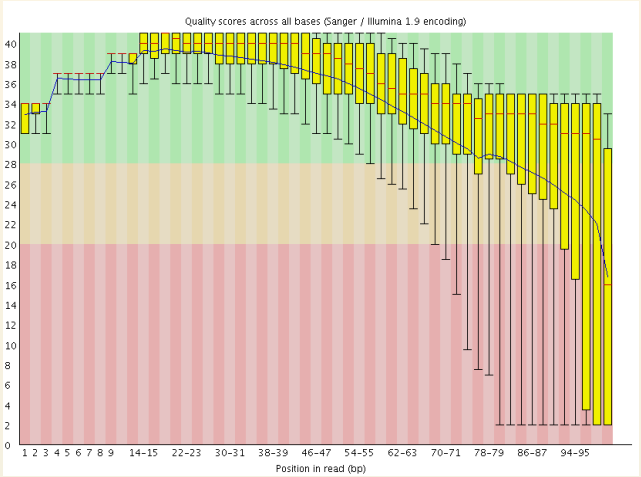
Recap – read evaluation and trimming

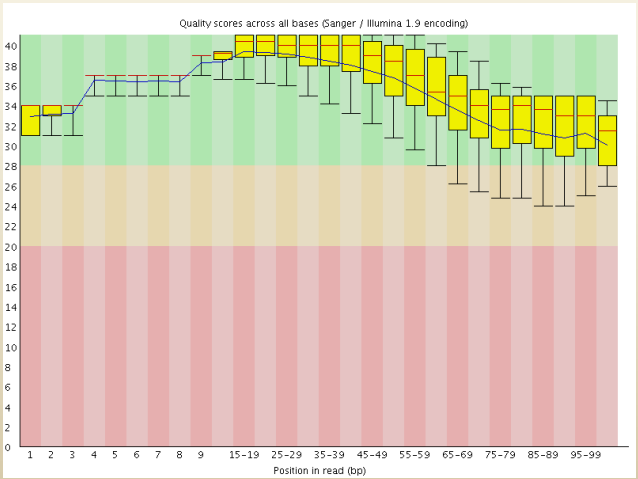
- Recommended to do adapter and quality trim
- A stringent and/or global trimming increases data loss
  - Assemblies benefit from a more stringet trim
  - Differential expression analyses suffers from stringet trims – loosing rare transcripts

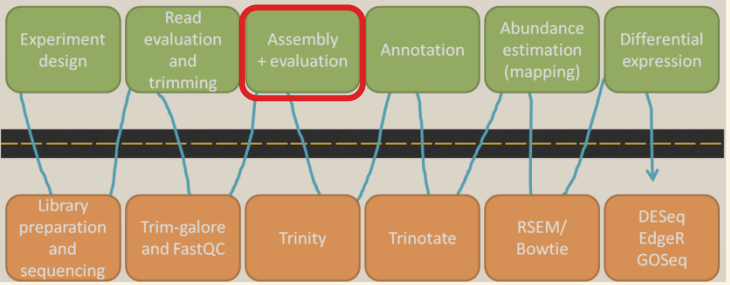


Recap – read trim-galore / fastqc

- Trim-galore trims both adapters and by quality
- In fastqc look for
  - Poor R2
  - Consecutive failed cycles
  - Trends in sequence content
  - Improvement compared to raw reads

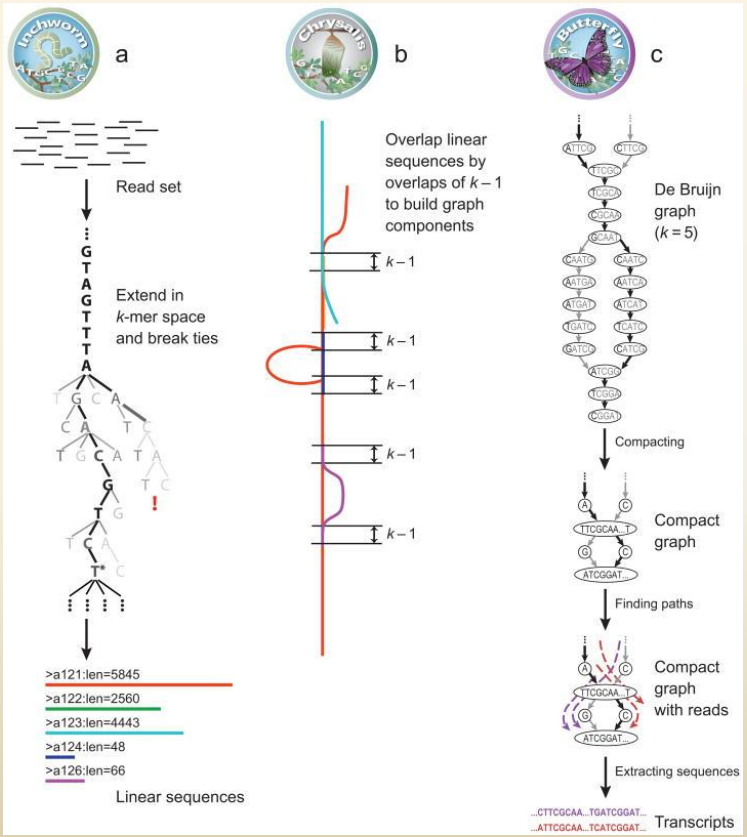


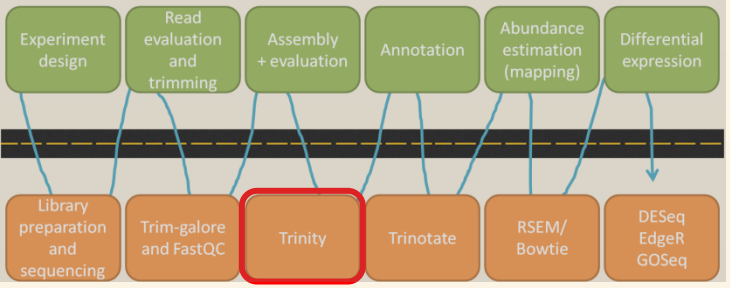




Recap – assembly

- *Ab initio*
  - Needs a **good** reference
  - Shorter time
  - Novel/rare transcripts resolved
- *De novo*
  - No reference needed
  - More coverage needed
  - Resolves complex splice variants
- Mixed approach
  - Polyploid organisms
  - Production of reference transcriptome

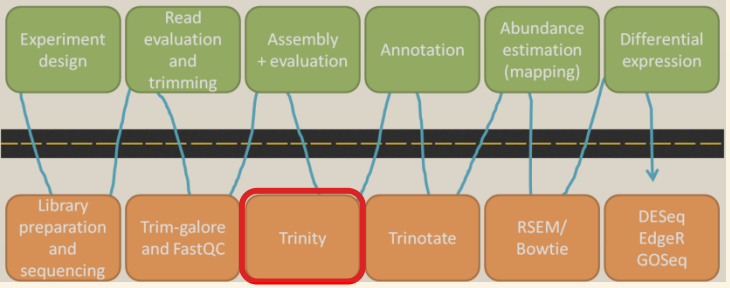




Recap – assembly evaluation I

- Evaluation by stats (TrinityStats.pl)
- Easy to focus on number of transcripts
- Contig N50 and median contig length are more important

	Complete	GG	Alternative
Total trinity 'genes'	320 520	342 099	380 658
Total trinity transcripts	468 626	454 484	569 062
Percent GC	47.31	47.64	47.41
Stats based on ALL transcript contigs			
Contig N10	3 657	5 607	4 648
Contig N20	2 645	3 962	3 330
Contig N30	2 042	2 986	2 524
Contig N40	1 597	2 276	1 930
Contig N50	1 235	1 716	1 463
Median contig length	459	505	472
Average contig	784.28	972.96	873.39
Total assembled bases	367 534 825	442 195 867	497 015 429

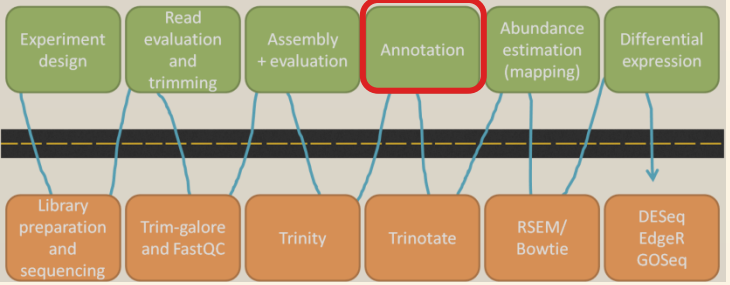


Recap – assembly evaluation II

- Abundance estimation (RSEM) with **all** samples
  - May be used for filtering (if making a reference)
  - Indicative of artefacts
- Full length estimation (BLAST)
  - Poorer resolution of length with complex eukaryotes

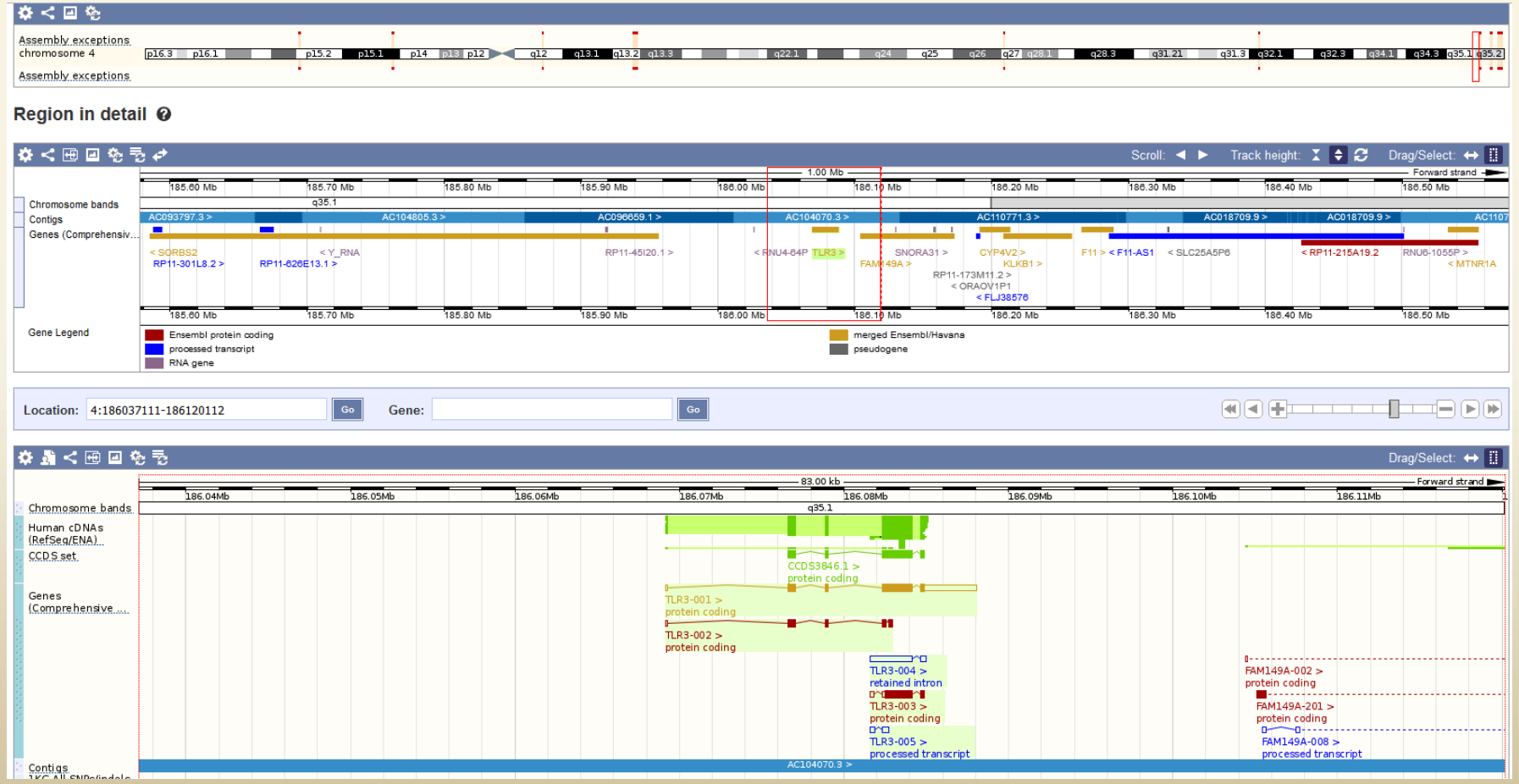
GG		DN
num features	neg_min fpkm	num features
-83	1	-65033
-69	2	-18941
	3	-16920
-66	4	-15806
-65	5	-11258
	6	-11201
…	…	…
73371	-6	16072
86545	-5	18267
104769	-4	21055
135418	-3	24869
199789	-2	30779
315811	-1	46015
342099	0	320520

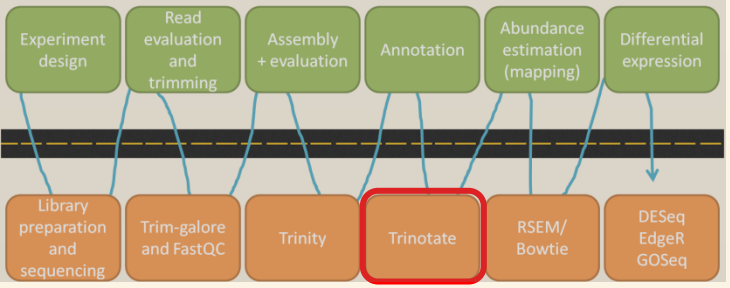
hit_pct cov bin	count_in bin	>bin below
100	5027	5027
90	2008	7035
80	1841	8876
70	1915	10791
60	2189	12980
50	2491	15471
40	2793	18264
30	3213	21477
20	2961	24438
10	906	25344



Recap – annotation

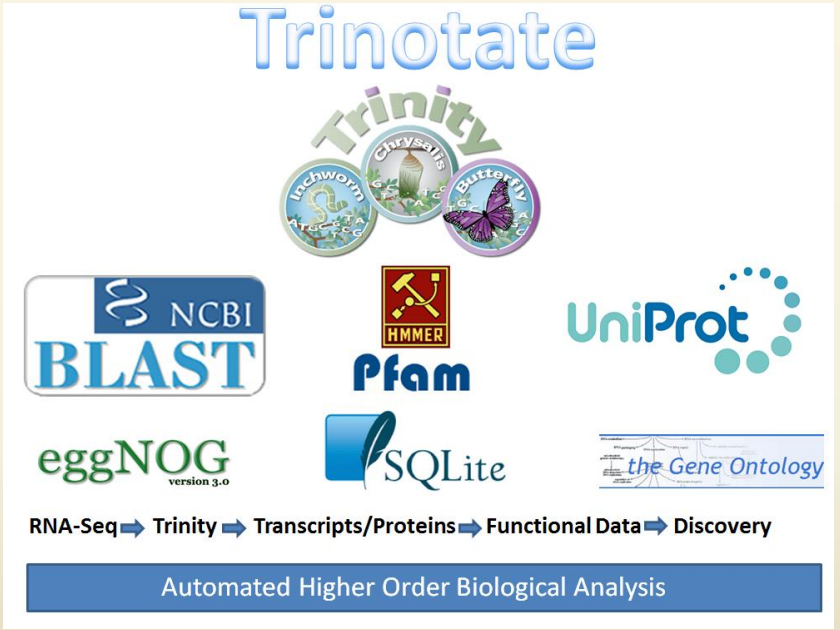
- Genome / transcriptome metadata

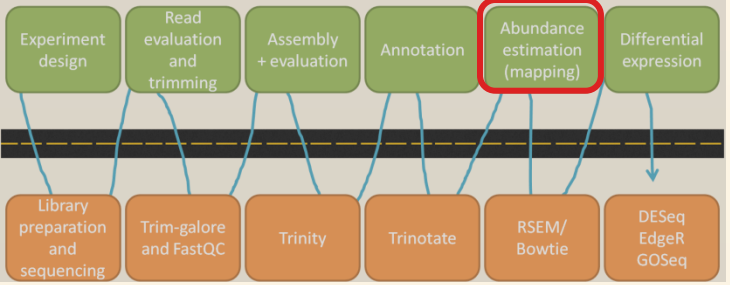




Recap – Trinotate

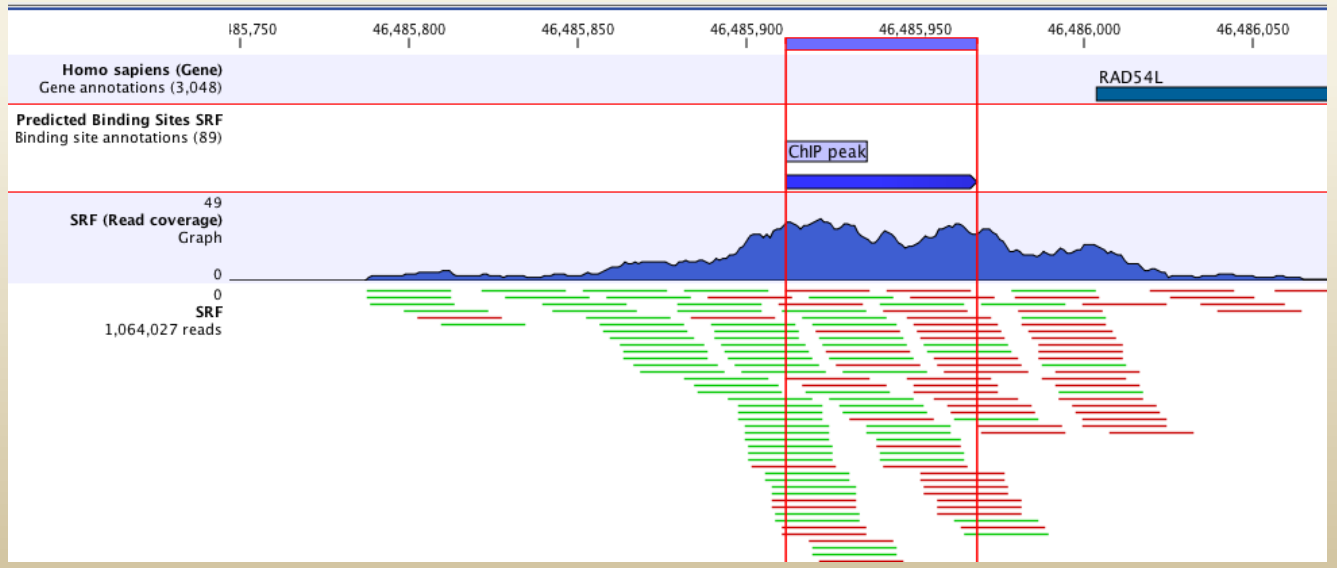
- One of the most extensive annotation pipelines
- BLAST, PFAM, transmembrane helices, sorting signals, GO categories
- For non-mammalian species: tends to miss the right family member

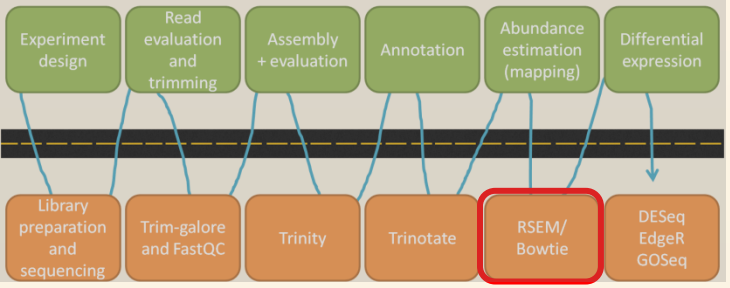




Recap – abundance estimation (mapping)

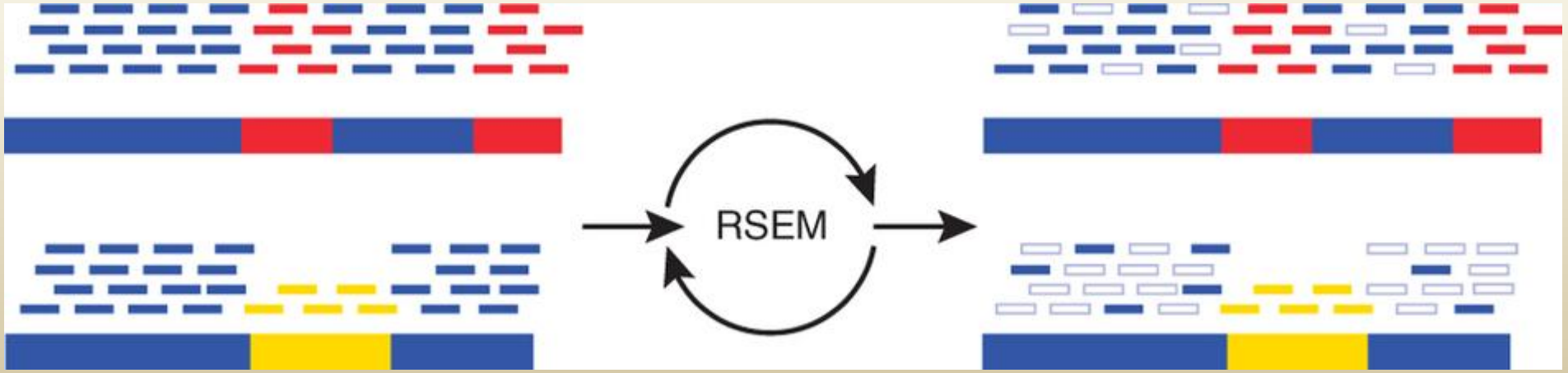
- Counting RNAseq reads that maps to each transcript for differential expression analysis
- Various RNA-mappers available
- They handle multi-mapping reads differently



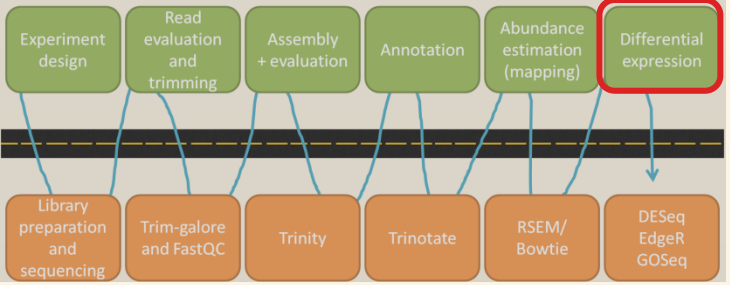


Recap – RSEM (bowtie)

- Applies a likelihood based algorithm for multimapping reads
- Optimized for RNAseq -> differential expression analyses
- Also used for abundance estimation when evaluating Trinity.fasta quality

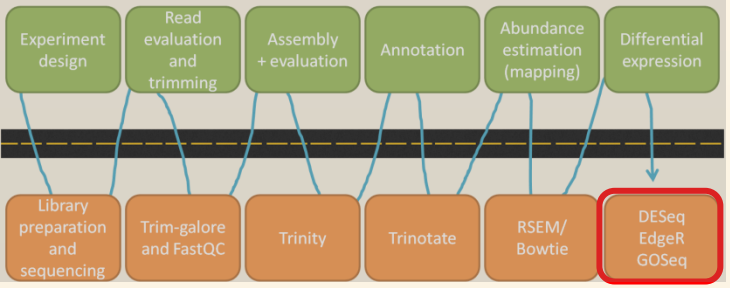


doi:10.1038/nprot.2013.084



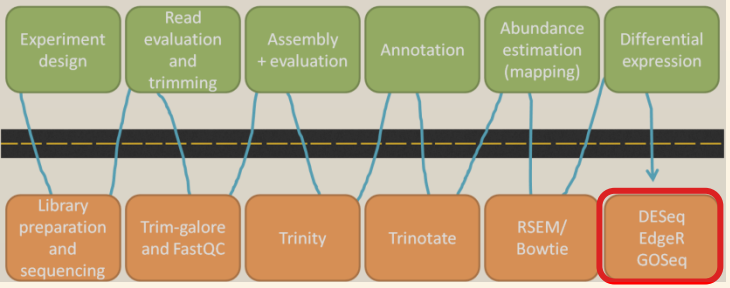
Recap – differential expression

- Differential expression analyses on RNAseq data has its challenges
  - Lack of replicates due to $
  - Sequencing biases and under/over sampling
  - Assumptions related to variance estimations and selection of variance distribution
  - Heavily dependent on "correct" experimental design
- Several approaches available
- Mostly performed with R



Recap – DESeq

- No replicate, partial replicate and replicate comparisons
- Limited multi-factorial abilities
- Raw counts needed
- Negative binomial distribution assumed
- Per-gene dispersion for large variance genes, fitted dispersion for little variance genes
- Conservative for lowly expressed genes
- Average type I correction

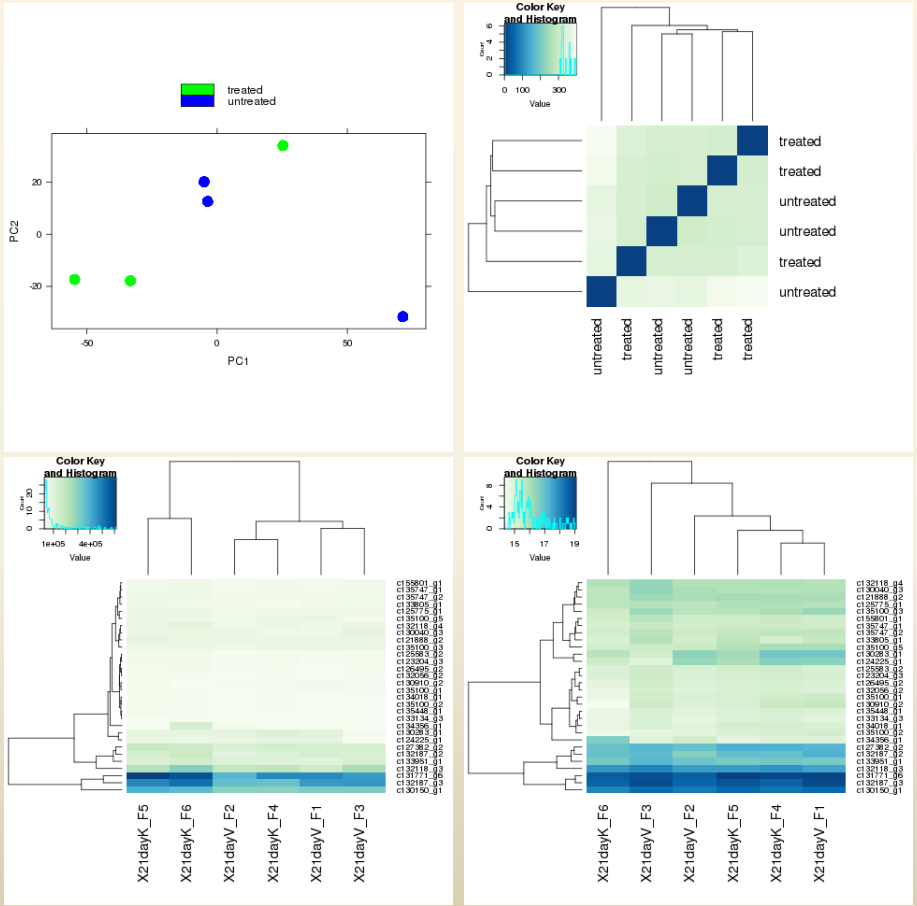


Recap – EdgeR

- Partial replicate and replicate comparisons
- Exact test or GLM
- Extended multi-factorial abilities (GLM)
- Negative binomial distribution assumed
- Dispersion is modeled using a maximum likelihood method
  - Gene wise dispersion used for highly variable genes and trended for little variable genes
- Non-conservative for lowly expressed genes
- "Skewed" type I error correction

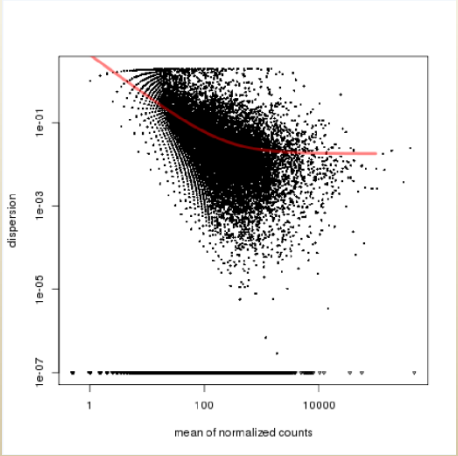
Recap – sample overview

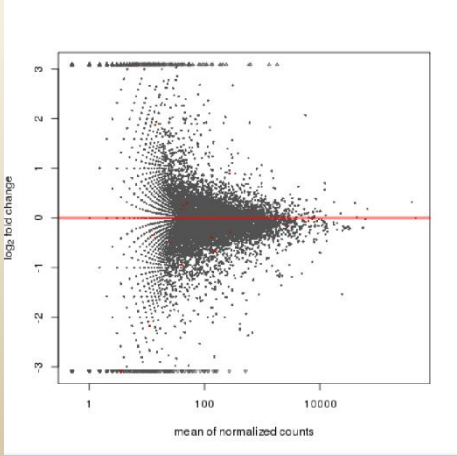
- PCA / MDS plots
  - How similar are the samples within each group?
  - Outlier removal?
- Heat maps and matrices
  - Complementary to PCA/MDS
  - Clusters samples according to expression profiles



Recap – no replicates

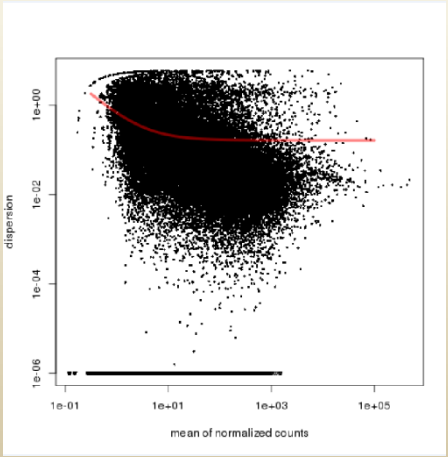
- No-replicate comparisons
  - Assuming that both samples behave similarly
  - Skewed dispersion estimates
  - False positives
  - Detects signals that conquer data noise

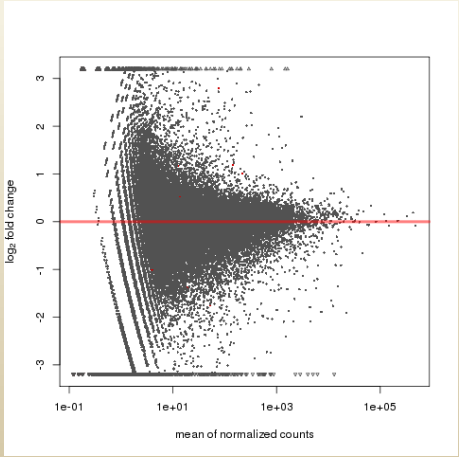




Recap - replicates

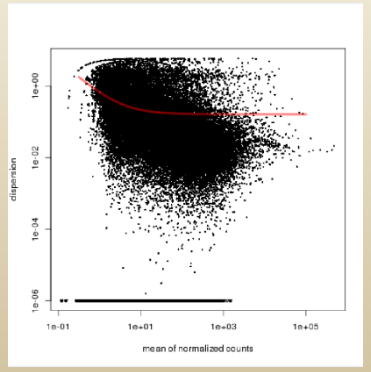
- Replicate comparisons
  - "Proper" estimation of dispersion
  - Dispersion estimates fit better
  - Less false positives
  - More replicates / more sequencing increases sensitivity for high/low expressed genes

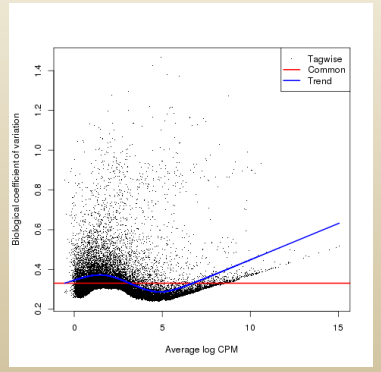


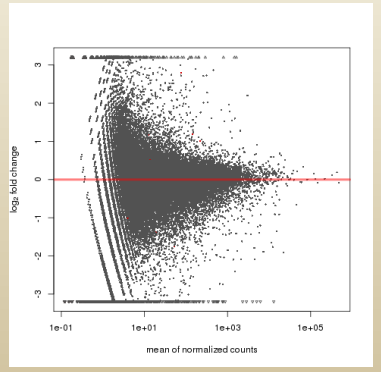


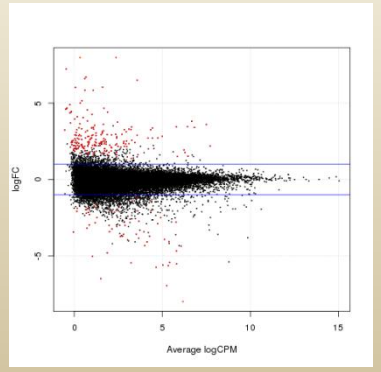
Recap – DESeq vs EdgeR

- DESeq vs EdgeR
  - Dispersion estimates are different
  - Handles lowly expressed genes differently
  - Have different false discovery rates
  - EdgeR became intermediate between no-replicate and replicate DESeq with respect to # of reported DE genes





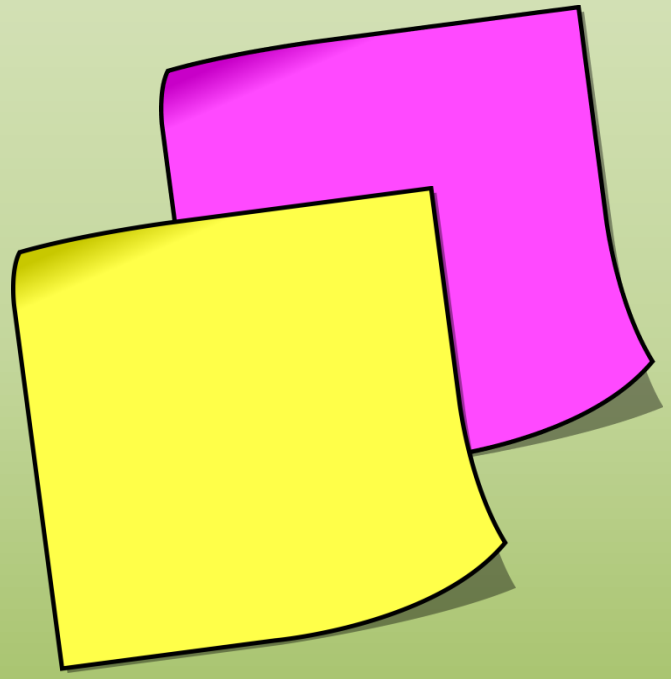




Questions?

The sticky notes!

- Put up YELLOW if command is running nicely
- Put up PINK if error or other issues



Support file

Use the support file (READ_ME - multi-factorial comparison in edgeR)

R

Use the support file (READ_ME - multi-factorial comparison in edgeR)

Before loading R do:

export PATH=/cluster/software/INF_BIOX121_H15/R/R-3.2.2/bin:$PATH

/cluster/software/INF_BIOX121_H15/R/R-3.2.2/bin/R

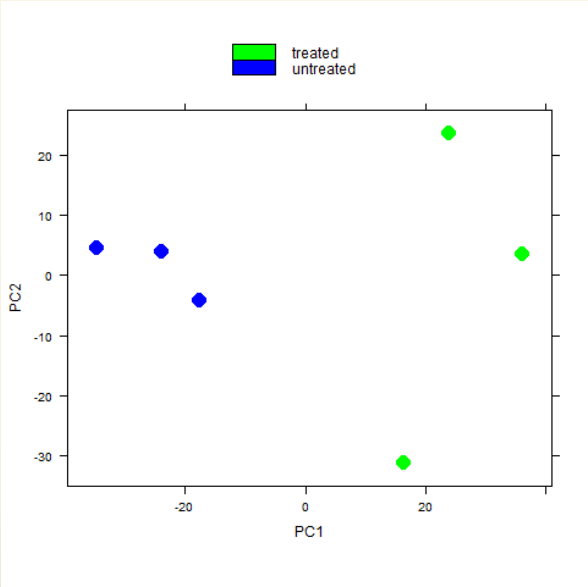
Yesterday : 21day DESeq

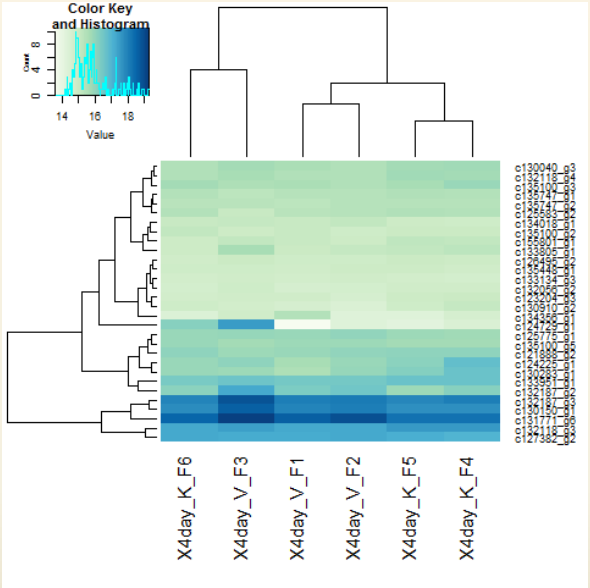
- Reported 133 DE genes
- Do the same analysis using the 4 day matrix

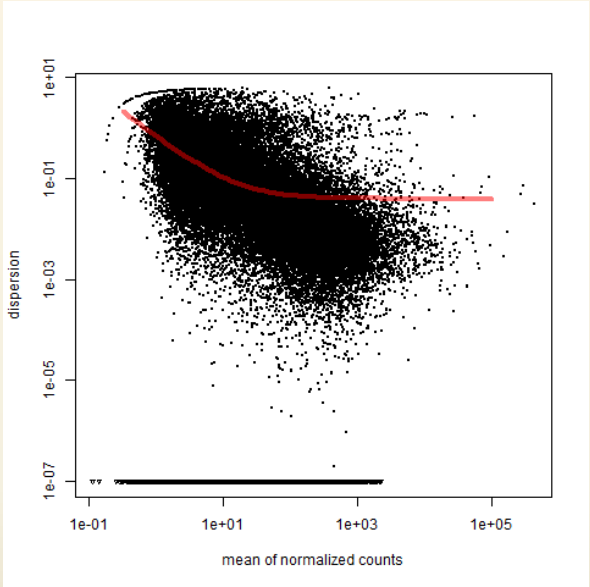
```
/data/RNAseq2/differential_expression/4day.counts.matrix
```

```
#############################################################
###### three replicate comparison DESeq 4 day #######
#############################################################
```

Today : 4day DESeq







- 197 DE genes

One at a time or all at once?

- Are the results comparable if time-points are treated together or independently?

Independent time series (edgeR)

- Two (assumed independent) time points are given in file 4day_21day.counts.matrix
- We will analyse these compared to their relative controls using a GLM approach

Time-series

All samples are from different fish

->

Independent time-series analysis

Control vs vaccinated 4 days after vaccination

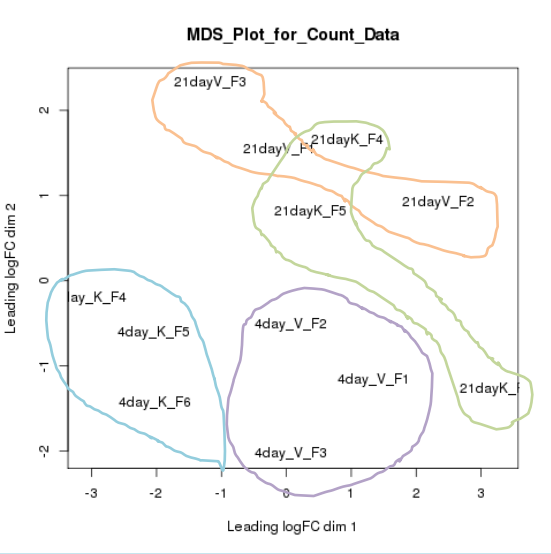
Control vs vaccinated 21 days after vaccination

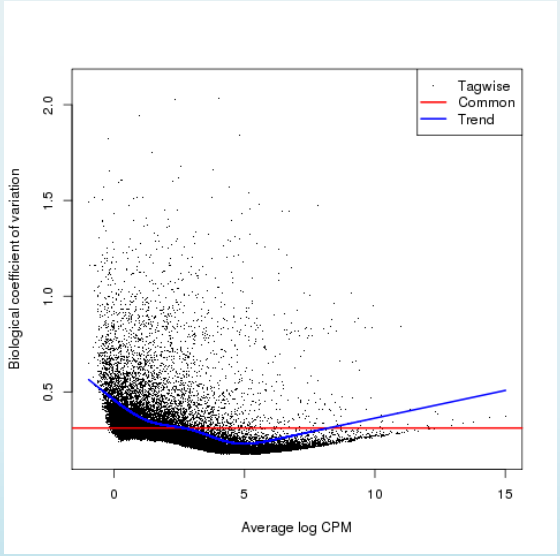
*Assuming that the expression-level of gene A at time A does not influence gene A at time B

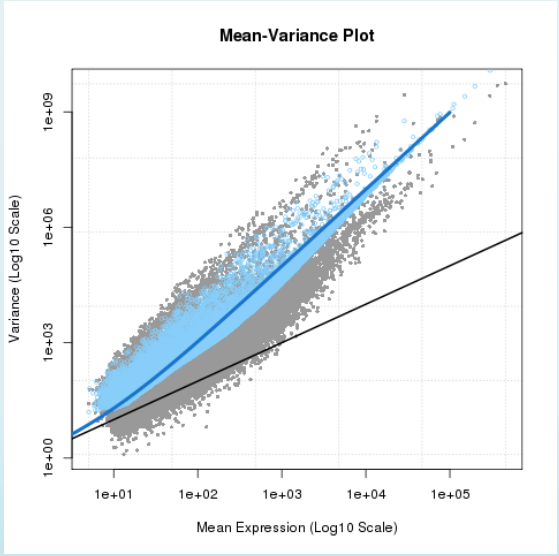
Assumption makes test 4 day and test 21 day comparable

Time-series

Preliminary investigations:





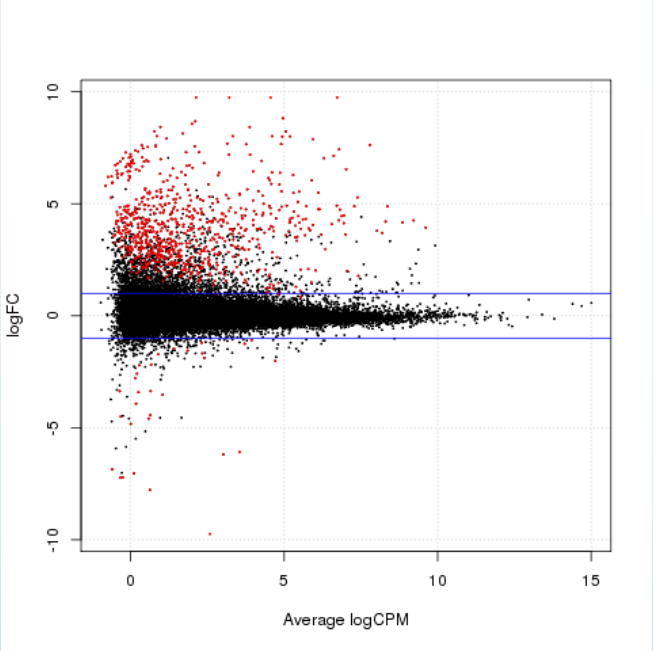


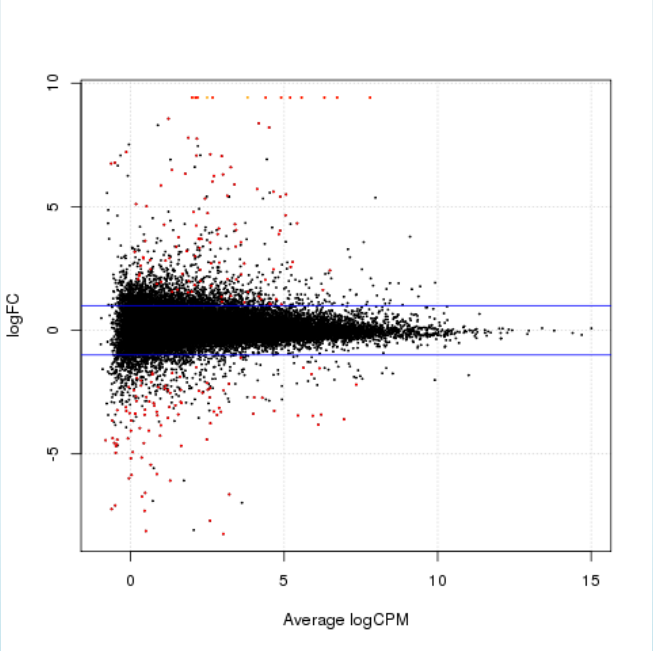
4 day samples are more homogenous across replicates

Trended dispersion fits the tagwise dispersions well

Time-series

Compare the outputs of the two analyses.





4 day contains more DE genes, mostly upregulated

775 vs 209 DE genes

One by one or all at once?

One by one DESeq:

4 day 197 DE genes

21 day 133 DE genes

All together edgeR:

4 day 775 DE genes

21 day 209 DE genes

One by one or all at once? II

Normalization!

Dispersion estimation – especially lowly expressed genes

Nested time series (DESeq)

- A more complex approach using a time 0
- GLM based
- Condition (treatment/control) vs time
- Various interactions between condition and time
- You still assume independent samples
- Still assuming that the expression level at time 1 does not affect expression and time 2

Nested time series (DESeq) II

- In READ_ME there are 3 different approaches
  - Fit 0 and fit 1
  - **Fit 2 and fit 3 – use this one**
  - Fit 4 and fit 5

Time-series

Compare the outputs of the two analyses. DE genes

Time-series

Compare the outputs of the two analyses. DE genes

GOSeq

- GO:terms – the collaborative effort to make consistent descriptions of gene products across databases
- Consists of up to three elements
  - Molecular function
  - Biological process
  - Cellular component

Example: Interleukin 1b

- Molecular functions:
  - Cytokine aktivity, interleukin-1 receptor binding, protein domain specifc binding
- Biological processes:
  - Activation of MAPK acitivty, aging, apoptopic process +++
- Cellular components:
  - Extracellular region, cytosol, vesicle +++

GOSeq II

- Uses the output of a DE analysis to look for enriched GO:terms
- Considers gene length biases using Wallenius distribution
  - The probability a gene will be DE due to length (more reads mapped)
- Each GO:category is tested for over/under representation among the DE genes

GOSeq III

- Model species is implemented however non-model species may be analyzed given you can provide:
  - GO categories
  - Gene lengths
  - Factor labels

GO:enrichment

- Copy the following files to ~ from `data/RNAseq2/differential_expression`
  - 21day_factor_labeling
  - lengths_effective_genes.txt
  - go_annotations.txt
- Run the GOseq code in the READ_ME file

GO:enrichment

Check the following GO:term

GO:0002921

Which gene IDs has this GO:term?

GO:enrichment

Which gene IDs has this GO:term?

c102983_g1
c112700_g1
c113312_g1
c113694_g1
c128632_g3
c133203_g4
c133343_g7
c134163_g2
c134694_g1
c200658_g1

GO:enrichment

What are these genes?

GO:enrichment

What are these genes?

C1S Complement C1s subcomponent

C4BPA C4b-binding protein alpha chain

CASP Calcium-dependent serine proteinase

FHOD3 FH1/FH2 domain-containing protein 3

HECAM Hepatocyte cell adhesion molecule

IC1 Plasma protease C1 inhibitor

MASP1 Mannan-binding lectin serine protease 1

PTN6 Tyrosine-protein phosphatase non-receptor type 6

RET1 Retinol-binding protein 1

GO:enrichment

Are the enriched/depleted GO:terms relevant?

GO:enrichment

Are the enriched/depleted GO:terms relevant?

This experiment did not give the expected immune response and the detected effects are more related to growth, external stimuli etc.