UiO : Department of Biosciences
University of Oslo

MBV4410/9410 Fall 2016

Dec. 6 - Analysing transcriptome data (using R) – part 2

Outline

Monday

Before lunch:

- Transcriptomics (lectures/practical)
  - Sequencing technologies
  - Transcriptome assembly
  - Gene expression

After lunch:

- Basic R/RStudio (lecture)
- Installing/setting up R/RStudio
- Basic R (practical)

Tuesday

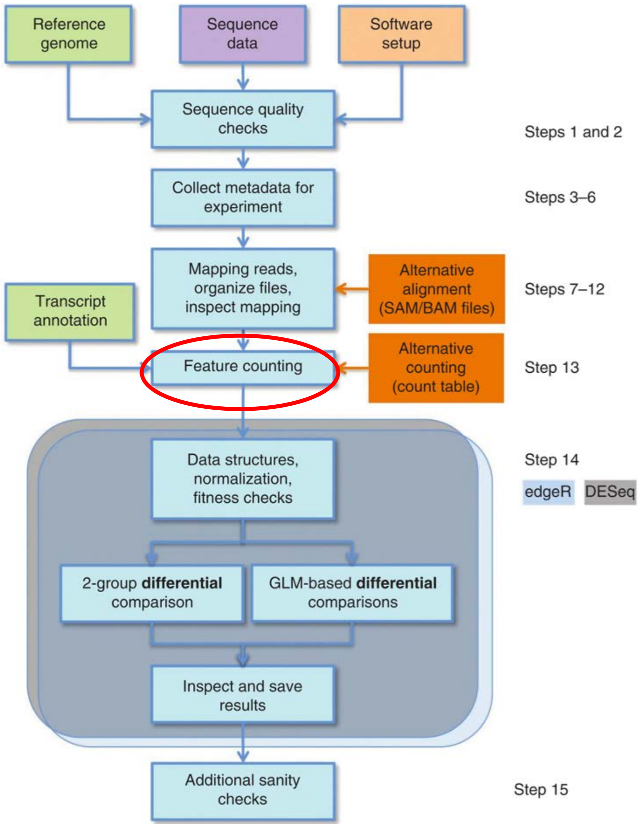
Continue the transcriptomics pipeline (lectures/practical)

- Count gene expression
- Experimental design
- Quality assessment
- Differential gene expression

After lunch:

- Bioconductor (lecture)
- Transcriptomics/DE-test (lecture/practical)

Summary

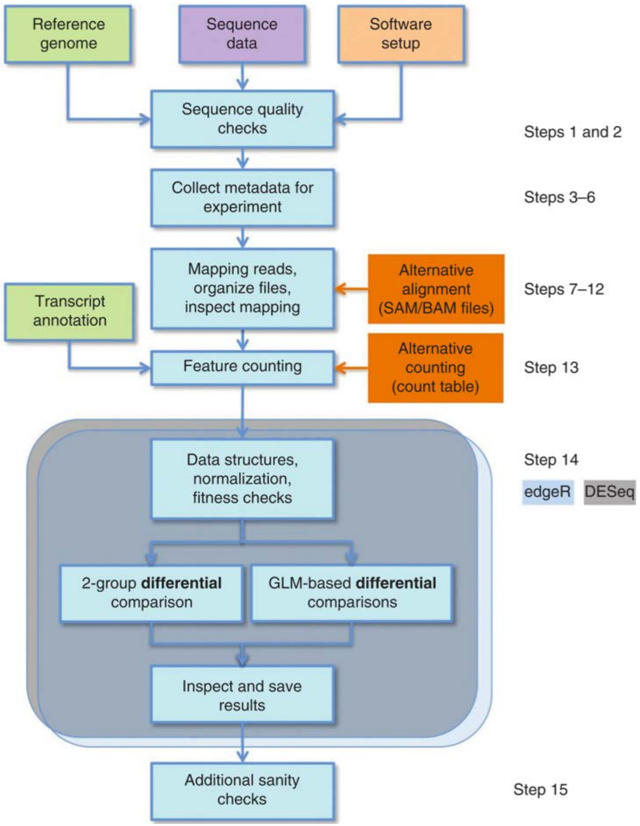


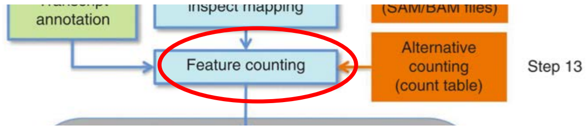
FastQC – *view* fastq files (fastq.gz / fq.gz)

trim-galore – *trims* the fastq files on quality and/or adapters

TopHat2 – *maps* the trimmed reads to the genome

Counting gene expression





Counting gene expression

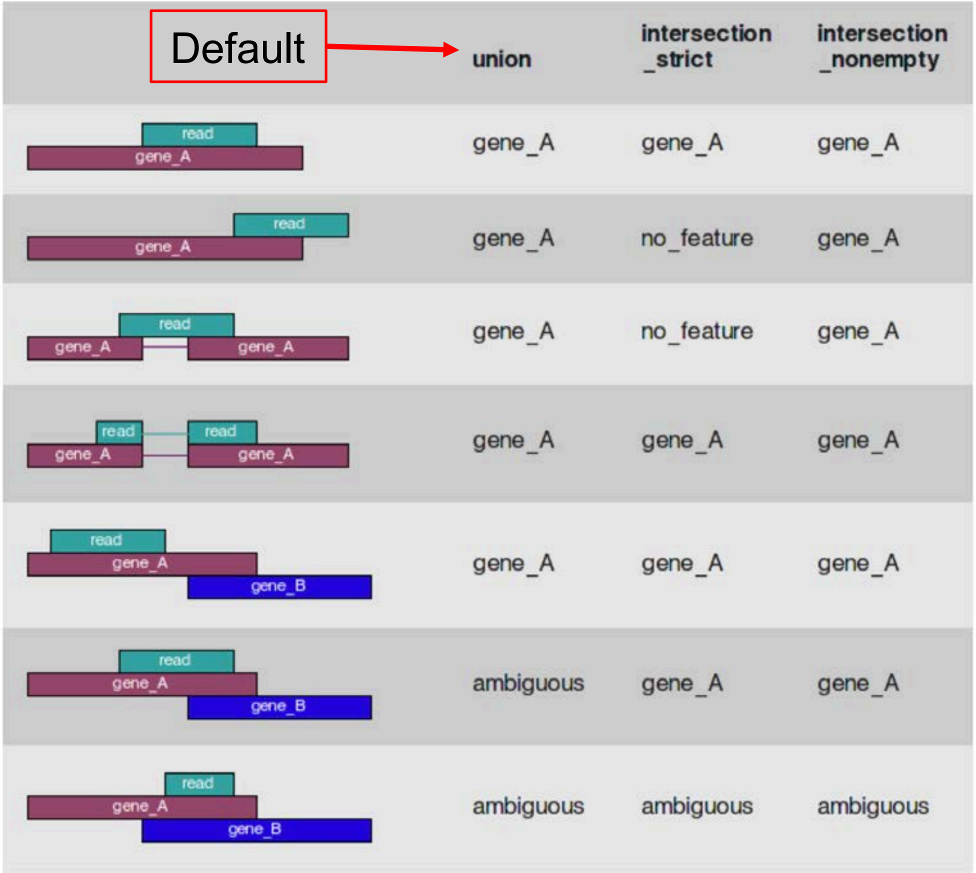
HTSeq-counts

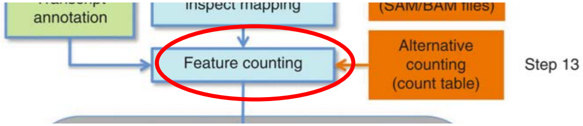
http://www-huber.embl.de/HTSeq/doc/count.html

HTSeq gives "raw counts"

Many programs to count/estimate expression:

- HTSeq (python) – gives raw counts
- Cufflinks (tuxedo pipeline) – fpkm values
- RSEM (de novo transcriptomes) – expected counts
- summarizeOverlaps (R) – similar to HTSeq
- …





Counting gene expression

Raw counts
The number of reads (pairs counting as one) mapping to a feature.
Not scaled by length (i.e. longer fragments = higher count) or sequencing depth (i.e. more sequences = higher count).

Counts per million (cpm)
Scaled by sequencing depth, not length.

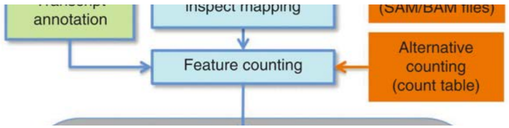
TPM
Transcripts per million. Scaled by sequencing depth and length

fpkm/rpkm
Reads/fragments per kilobase of exon per million reads mapped.
Similar to TPM. Scaled by sequencing depth and length

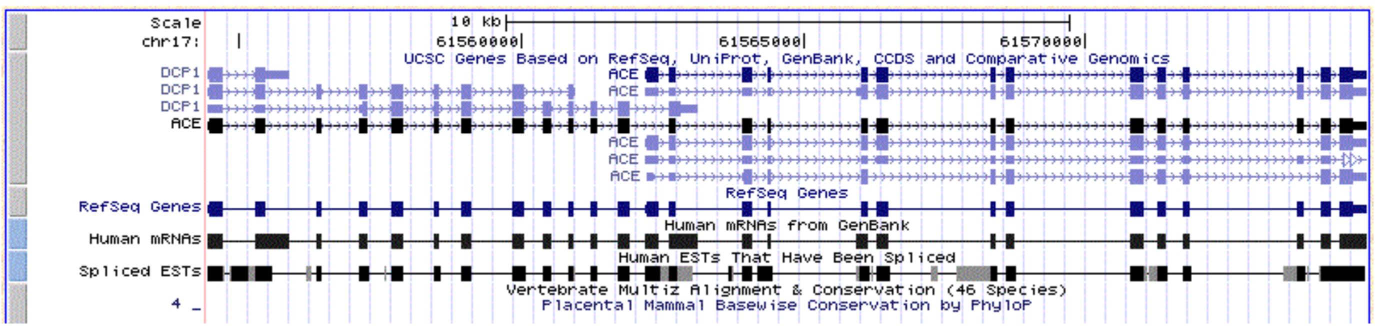
https://haroldpimentel.wordpress.com/2014/05/08/what-the-fpkm-a-review-rna-seq-expression-units/

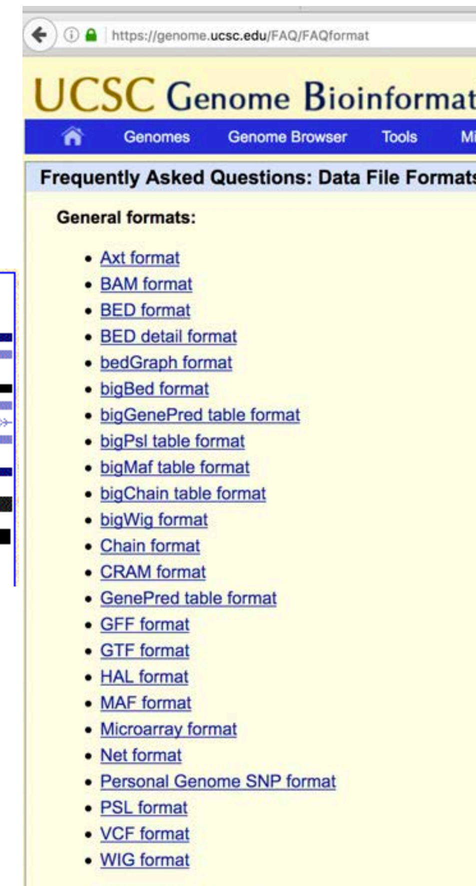
Annotation files

.gtf .gff .wig .gtf2 .gff3 .bed .vcf



- Like “Tracks” in a genome browser
- Specify coordinates in a genome
- A multitude of formats…

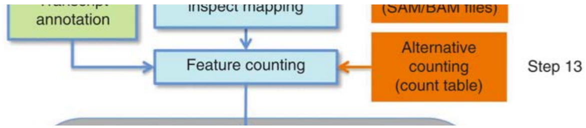




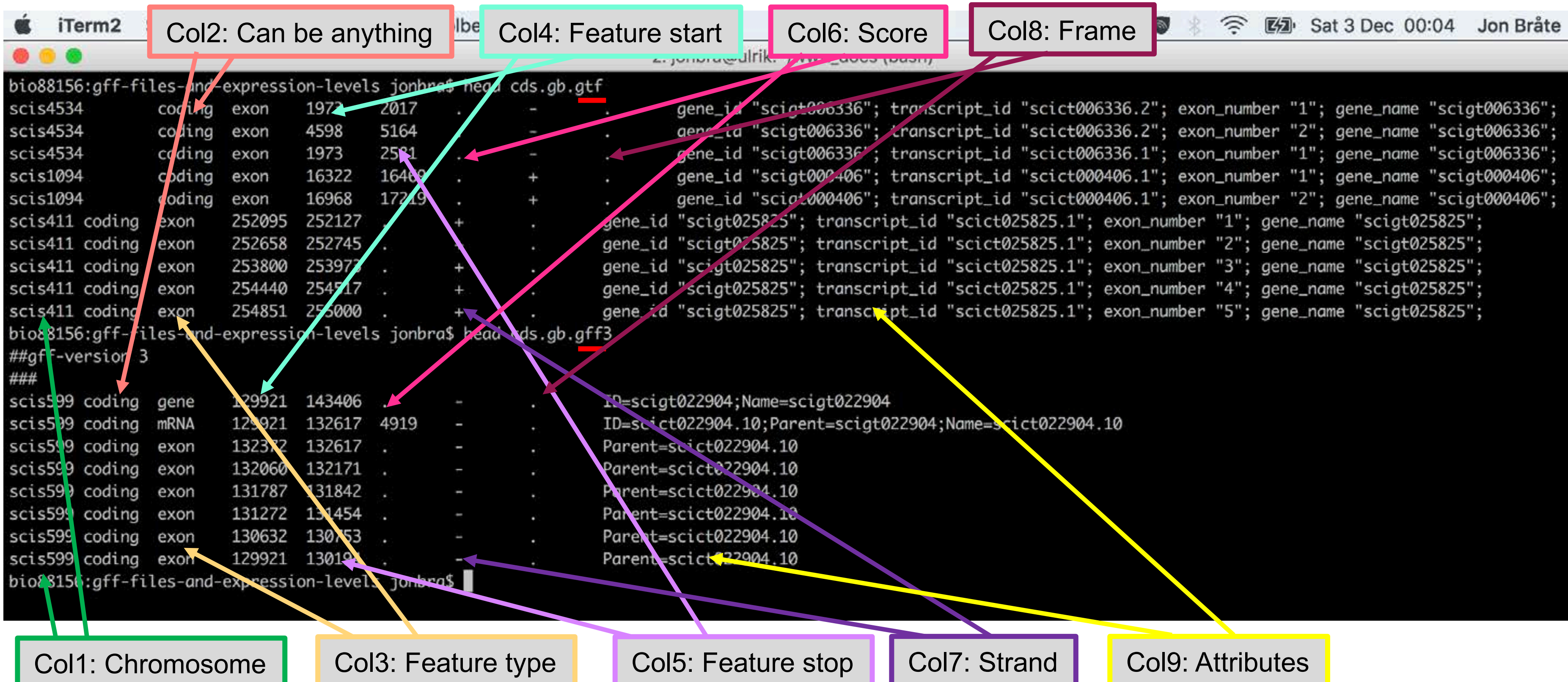
UiO : Department of Biosciences
University of Oslo

.gtf .gff .wig
.gtf2 .gff3
.bed .vcf

Annotation files

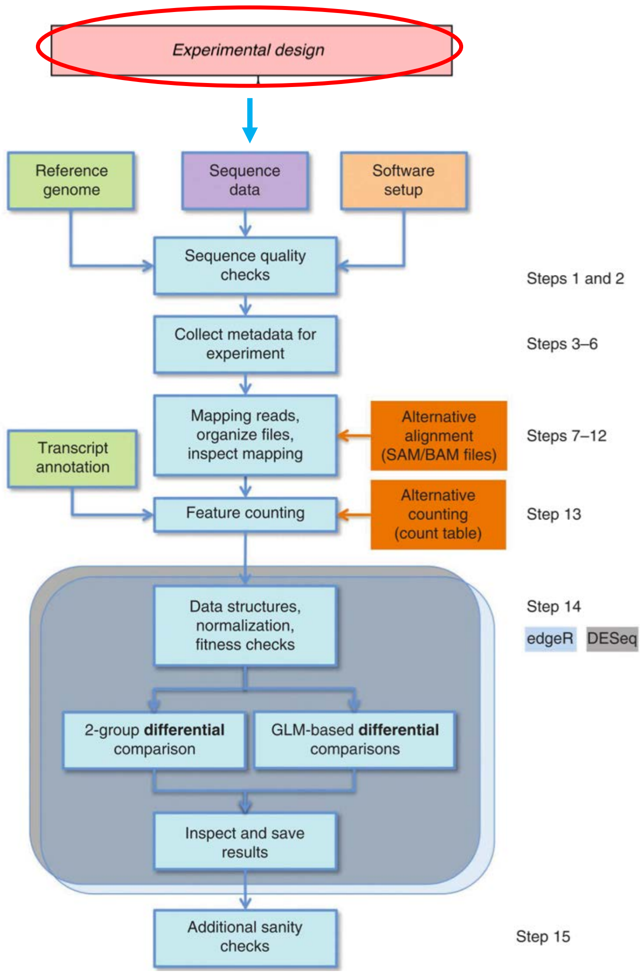


.gtf and .gff3 most common (perhaps…). 9 tab-separated columns

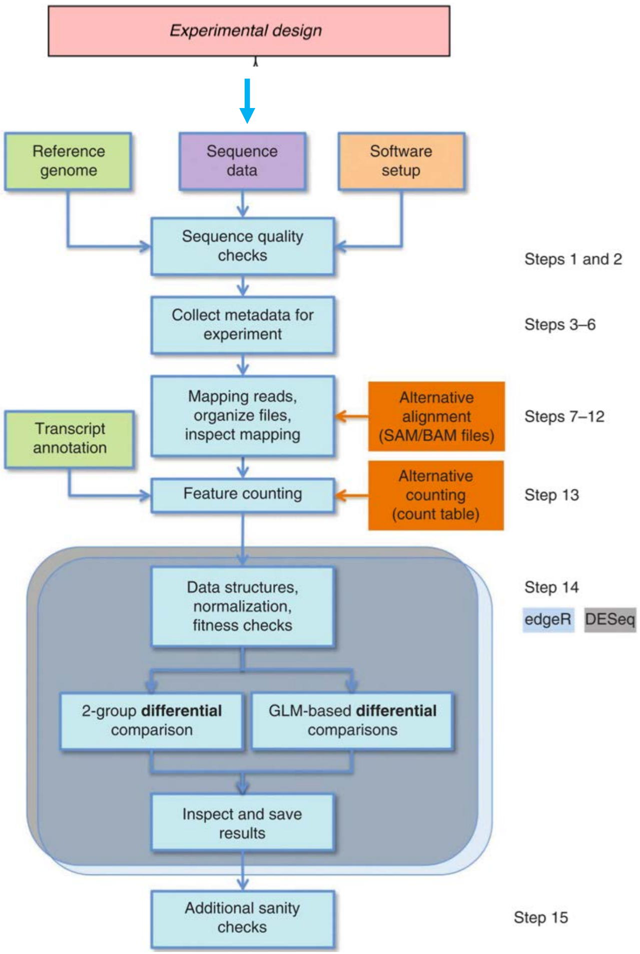


Exercise 3 – Counting gene expression

Experimental design



Experimental design



Experimental design

A crucial prerequisite for a successful RNA-seq study is that the data generated have the potential to answer the biological questions of interest. This is achieved by first defining a good experimental design, that is, by choosing the library type, sequencing depth and number of replicates appropriate for the biological system under study,

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

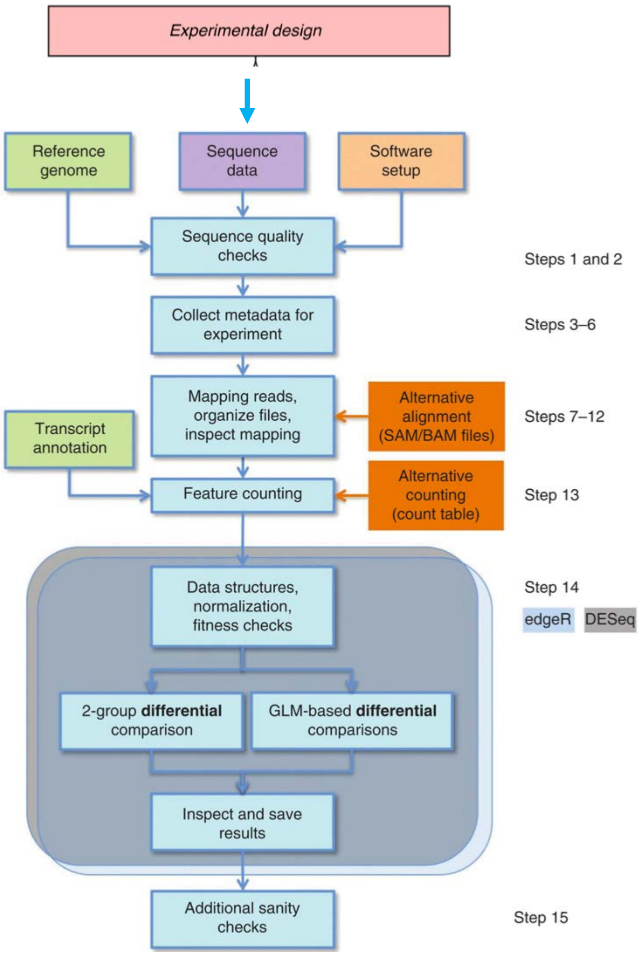
Genome Biology

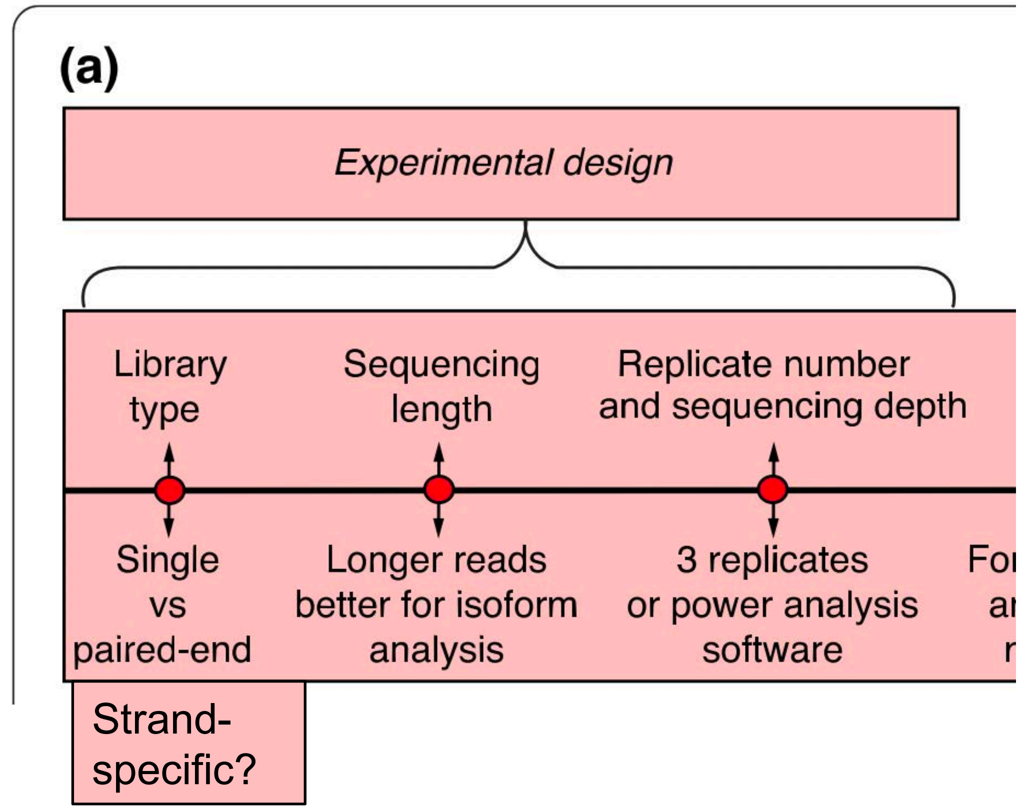
REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

Experimental design





Experimental design

Goal

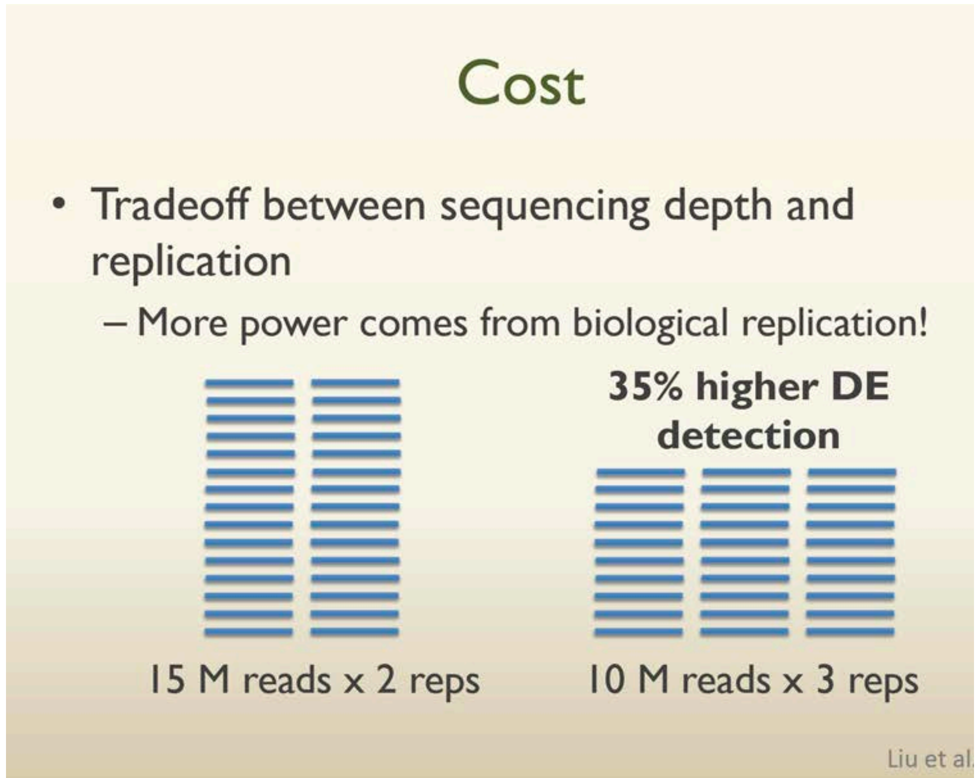
- To answer your research question, given logistical constraints.
- You can't do it all!

Experimental design - replicates

Differential expression analysis

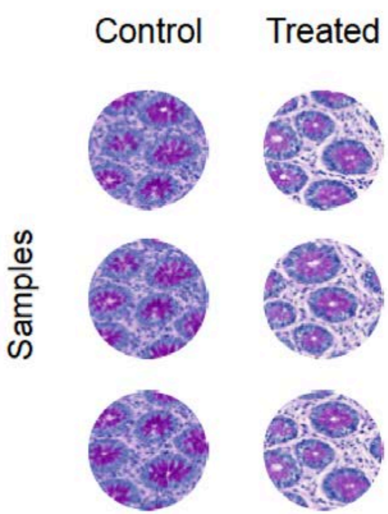
- Statistical power
  - The ability to distinguish differential expression due to treatment effect from background noise

Experimental design - replicates

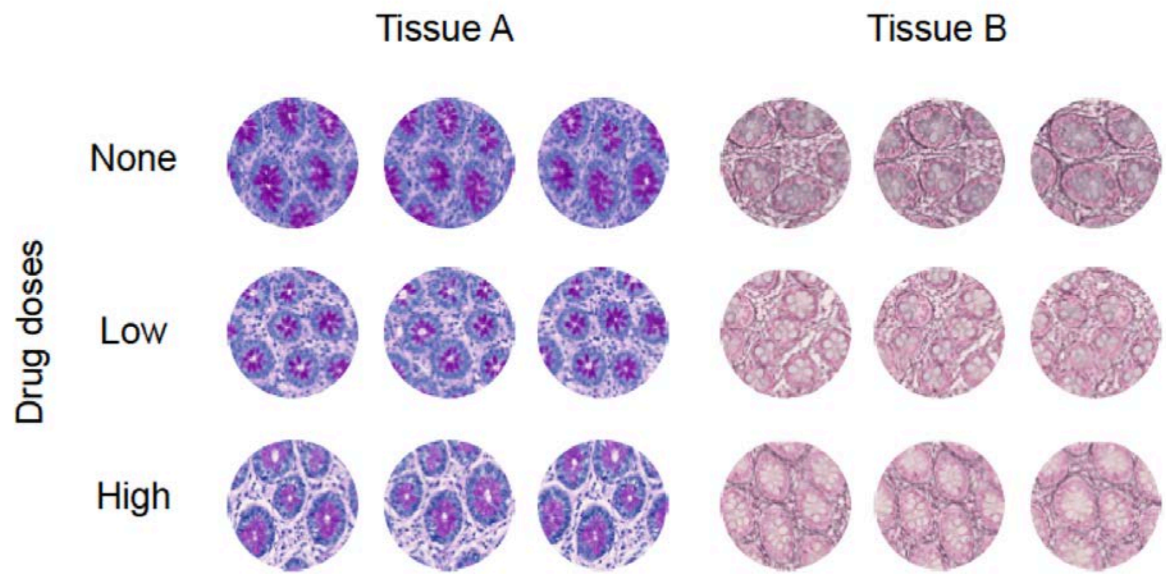


Experimental design - replicates

Quickly becomes many samples!



Simple design: control vs. treated



Complex design: Two factors, *tissue* in two levels (A and B) and *drug* in three levels.

Experimental design – systematic bias

- Ensure that you will not have any systematic biases:
  - Distribute the biological groups in a balanced way.
  - Divide into batches of the same sizes, limited by the capacity on each step.
  - Tip: in excel (or similar program) color code sample name according to biological group, and in next column color code by batch.
- Randomize and balance according to the biology your are interested in.

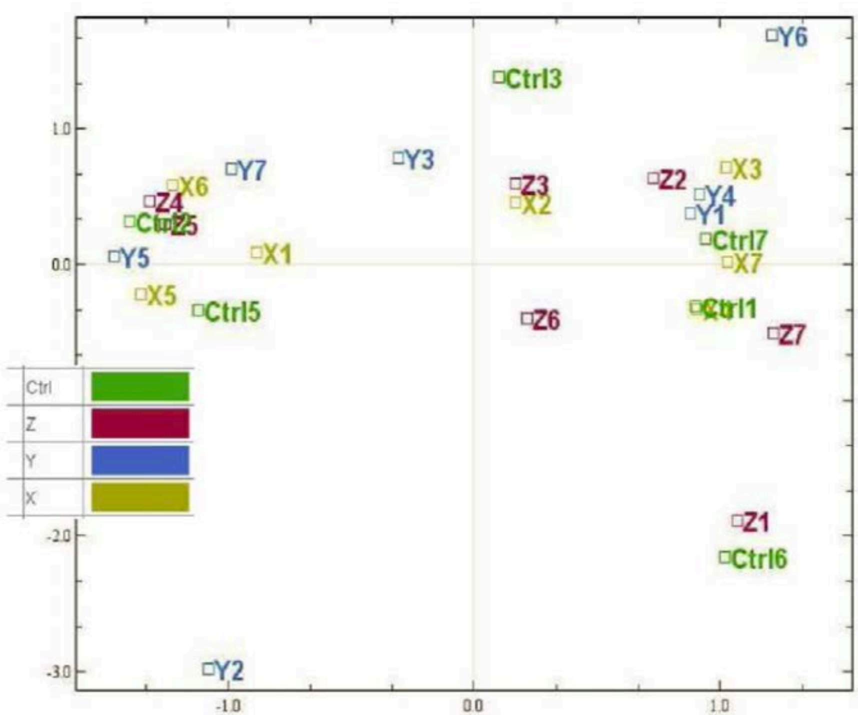
Experimental design: an example

Biology
A1
A2
A3
A4
A5
A6
B1
B2
B3
B4
B5
B6
C1
C2
C3
C4
C5
C6

Biology	Sample preparation order
A1	1
B4	2
C2	3
A3	4
B6	5
C4	6
A5	7
B2	8
C6	9
A2	10
B3	11
C1	12
A4	13
B5	14
C3	15
A6	16
B1	17
C5	18

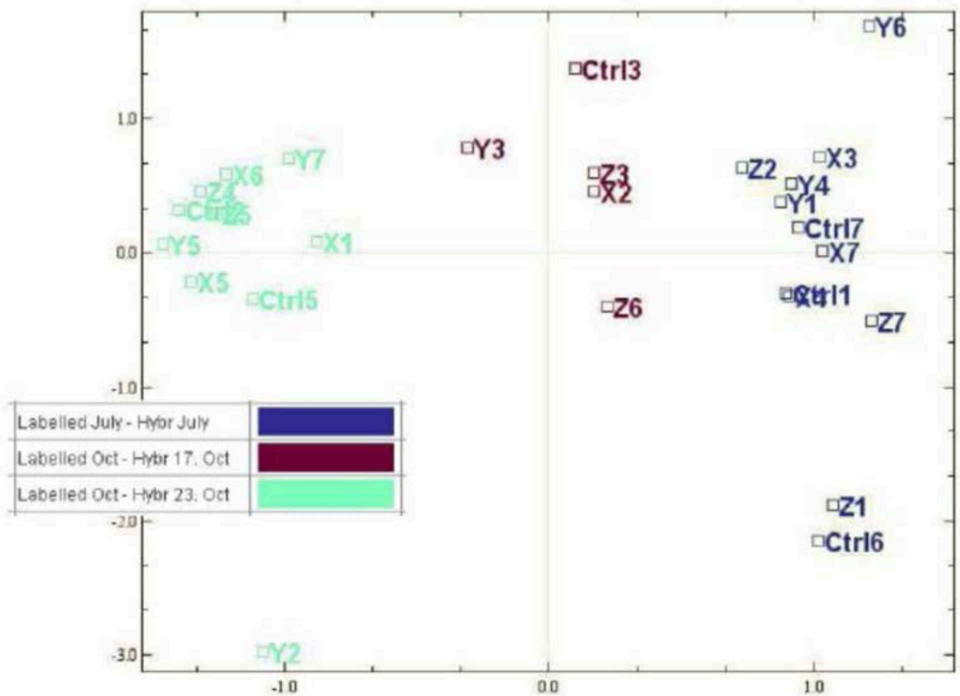
Biology	Sample preparation order	Extraction order
A2	10	1
B6	5	2
C1	12	3
A5	7	4
B5	14	5
C6	9	6
A6	16	7
B4	2	8
C5	18	9
A3	4	10
C3	15	11
B2	8	12
A4	13	13
C4	6	14
B1	17	15
A1	1	16
B3	11	17
C2	3	18

Experimental design: batch effect



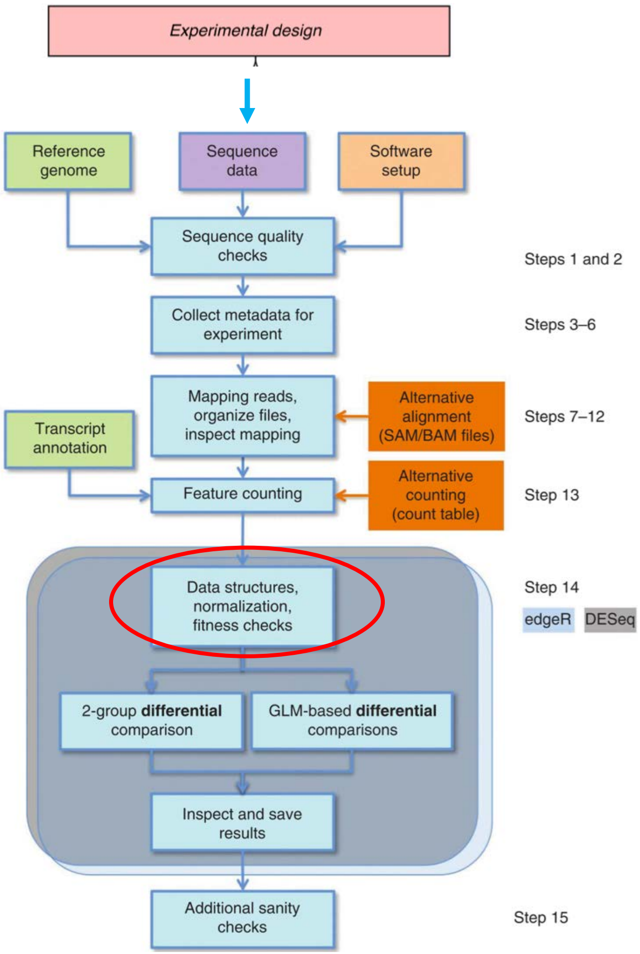
Samples color coded according to biology

Experimental design: batch effect



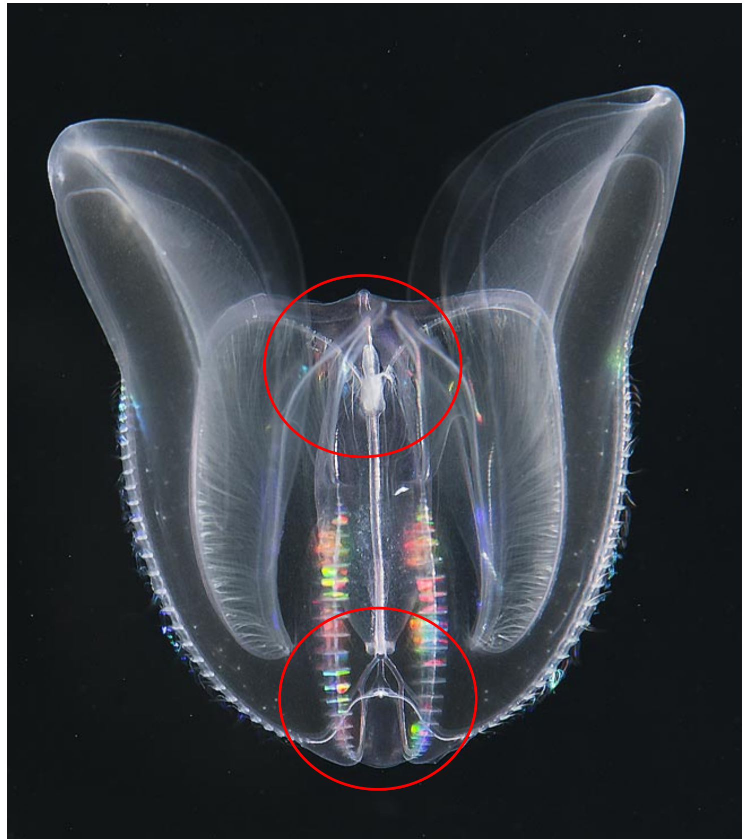
Samples color coded according to labeling date

Data exploration and quality assessment



Mnemiopsis leidyi

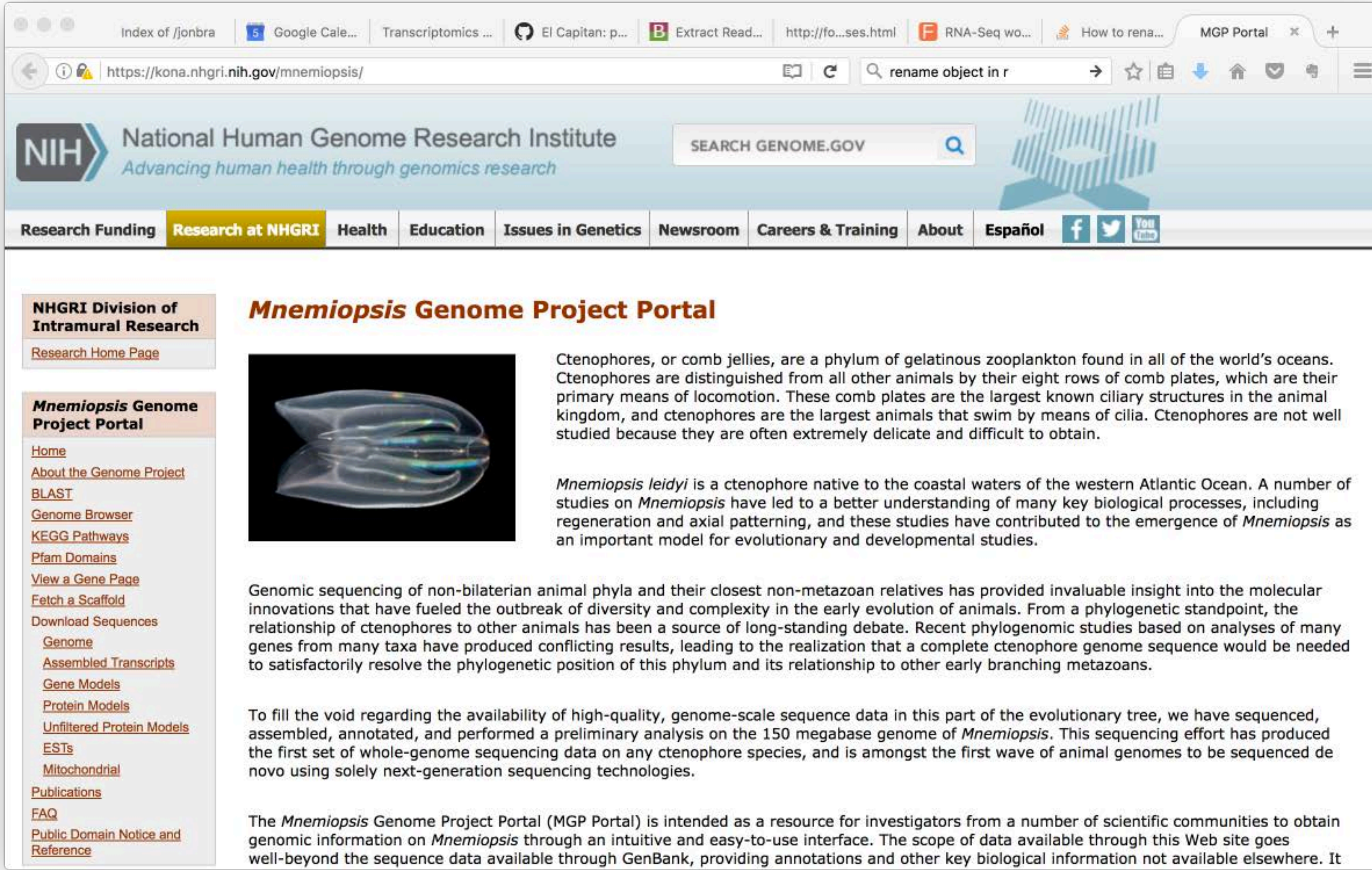
Goal: Find genes upregulated in the aboral organ



Oral organ X 4 replicates

Aboral organ X 4 replicates

UiO **Department of Biosciences**
University of Oslo


Index of /jonbra | Google Cale... | Transcriptomics ... | El Capitan: p... | Extract Read... | http://fo...ses.html | RNA-Seq wo... | How to rena... | MGP Portal

https://kona.nhgri.nih.gov/mnemiopsis/

rename object in r

NIH National Human Genome Research Institute
Advancing human health through genomics research

SEARCH GENOME.GOV

Research Funding | Research at NHGRI | Health | Education | Issues in Genetics | Newsroom | Careers & Training | About | Español

NHGRI Division of Intramural Research

Research Home Page

***Mnemiopsis* Genome Project Portal**

- Home
- About the Genome Project
- BLAST
- Genome Browser
- KEGG Pathways
- Pfam Domains
- View a Gene Page
- Fetch a Scaffold
- Download Sequences
  - Genome
  - Assembled Transcripts
  - Gene Models
  - Protein Models
  - Unfiltered Protein Models
  - ESTs
  - Mitochondrial
- Publications
- FAQ
- Public Domain Notice and Reference

Mnemiopsis Genome Project Portal

Ctenophores, or comb jellies, are a phylum of gelatinous zooplankton found in all of the world's oceans. Ctenophores are distinguished from all other animals by their eight rows of comb plates, which are their primary means of locomotion. These comb plates are the largest known ciliary structures in the animal kingdom, and ctenophores are the largest animals that swim by means of cilia. Ctenophores are not well studied because they are often extremely delicate and difficult to obtain.

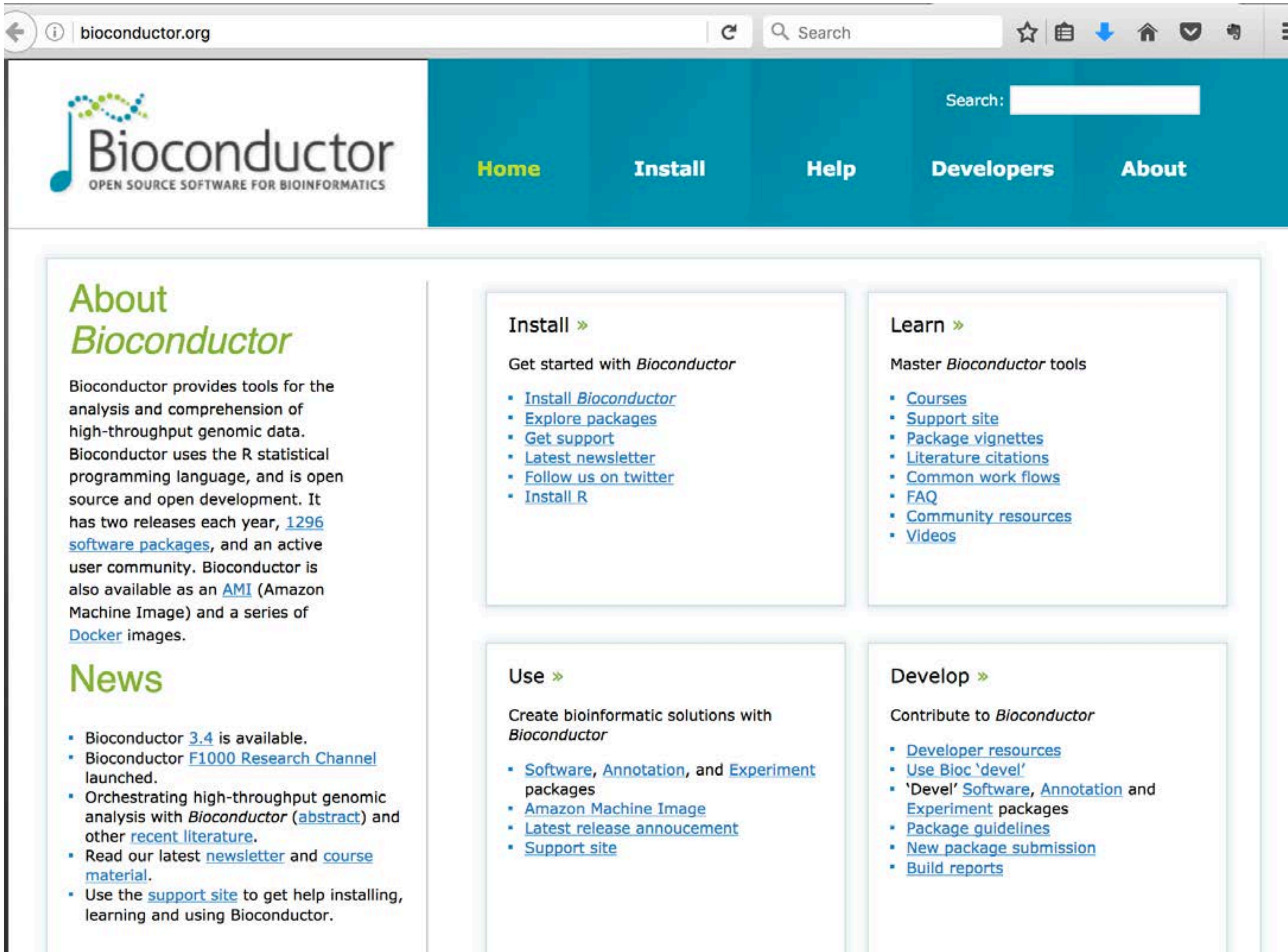
Mnemiopsis leidyi is a ctenophore native to the coastal waters of the western Atlantic Ocean. A number of studies on *Mnemiopsis* have led to a better understanding of many key biological processes, including regeneration and axial patterning, and these studies have contributed to the emergence of *Mnemiopsis* as an important model for evolutionary and developmental studies.

Genomic sequencing of non-bilaterian animal phyla and their closest non-metazoan relatives has provided invaluable insight into the molecular innovations that have fueled the outbreak of diversity and complexity in the early evolution of animals. From a phylogenetic standpoint, the relationship of ctenophores to other animals has been a source of long-standing debate. Recent phylogenomic studies based on analyses of many genes from many taxa have produced conflicting results, leading to the realization that a complete ctenophore genome sequence would be needed to satisfactorily resolve the phylogenetic position of this phylum and its relationship to other early branching metazoans.

To fill the void regarding the availability of high-quality, genome-scale sequence data in this part of the evolutionary tree, we have sequenced, assembled, annotated, and performed a preliminary analysis on the 150 megabase genome of *Mnemiopsis*. This sequencing effort has produced the first set of whole-genome sequencing data on any ctenophore species, and is amongst the first wave of animal genomes to be sequenced de novo using solely next-generation sequencing technologies.

The *Mnemiopsis* Genome Project Portal (MGP Portal) is intended as a resource for investigators from a number of scientific communities to obtain genomic information on *Mnemiopsis* through an intuitive and easy-to-use interface. The scope of data available through this Web site goes well-beyond the sequence data available through GenBank, providing annotations and other key biological information not available elsewhere. It

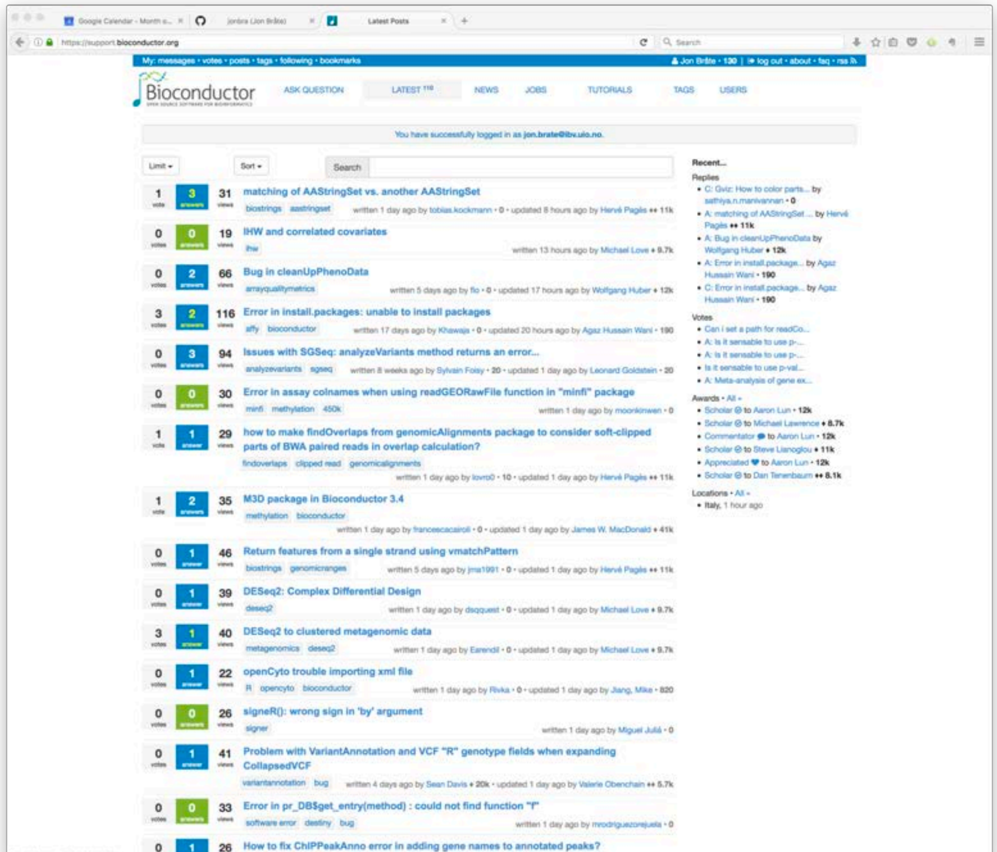
	aboral1	aboral2	aboral3	aboral4	oral1	oral2	oral3	oral4
ML000110a	69	175	141	139	108	146	133	63
ML000111a	0	0	0	0	0	1	0	0
ML000112a	1	10	8	3	2	13	6	1
ML000113a	383	546	402	471	290	190	282	317
ML000114a	188	214	257	230	289	215	162	128
ML000115a	493	455	540	501	413	403	419	452
ML000116a	404	462	464	362	516	336	285	336
ML000117a	266	361	301	273	396	277	239	277
ML000118a	177	158	162	153	164	131	107	136
ML000119a	382	339	362	295	254	310	259	308
ML00011a	37	26	33	29	24	46	34	26
ML000120a	227	225	250	141	333	241	130	169
ML000121a	385	294	398	213	385	351	188	270
ML000122a	352	336	336	283	442	300	245	276
ML000123a	1353	1232	1534	1162	1919	1272	976	1130
ML000124a	882	1694	1025	1001	979	834	655	849


bioconductor.org
Search
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS
Search:
Home
Install
Help
Developers
About
About Bioconductor
Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development. It has two releases each year, 1296 software packages, and an active user community. Bioconductor is also available as an AMI (Amazon Machine Image) and a series of Docker images.
News
Bioconductor 3.4 is available.
Bioconductor F1000 Research Channel launched.
Orchestrating high-throughput genomic analysis with Bioconductor (abstract) and other recent literature.
Read our latest newsletter and course material.
Use the support site to get help installing, learning and using Bioconductor.
Install »
Get started with Bioconductor
Install Bioconductor
Explore packages
Get support
Latest newsletter
Follow us on twitter
Install R
Learn »
Master Bioconductor tools
Courses
Support site
Package vignettes
Literature citations
Common work flows
FAQ
Community resources
Videos
Use »
Create bioinformatic solutions with Bioconductor
Software, Annotation, and Experiment packages
Amazon Machine Image
Latest release annoucement
Support site
Develop »
Contribute to Bioconductor
Developer resources
Use Bioc 'devel'
'Devel' Software, Annotation and Experiment packages
Package guidelines
New package submission
Build reports

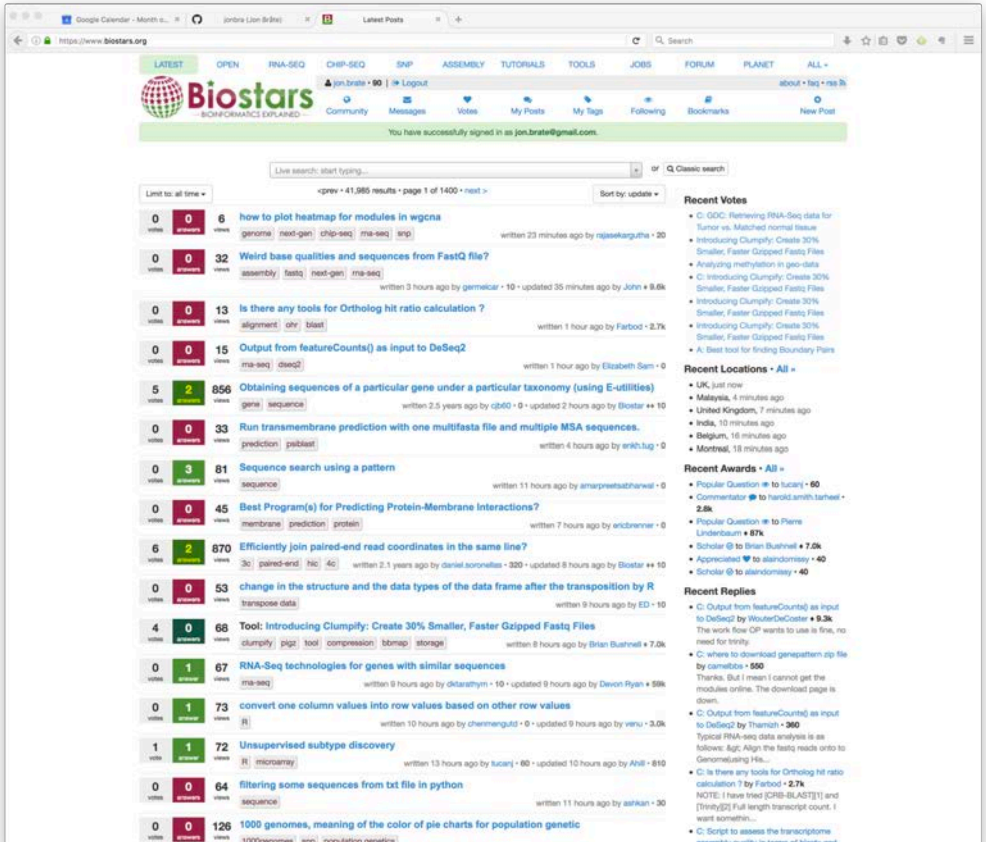
- **DESeq2 and edgeR** – two of the most common pacakges for RNA-seq analysis (differential expression).
- **DESeq2 and edgeR** – based on “raw counts” such as from HTSeq

- **Tuxedo pipeline** (TopHat+Cufflinks+Cuffdiff) also very common – fpkm-based. –
- Often people run all three procedures and compare

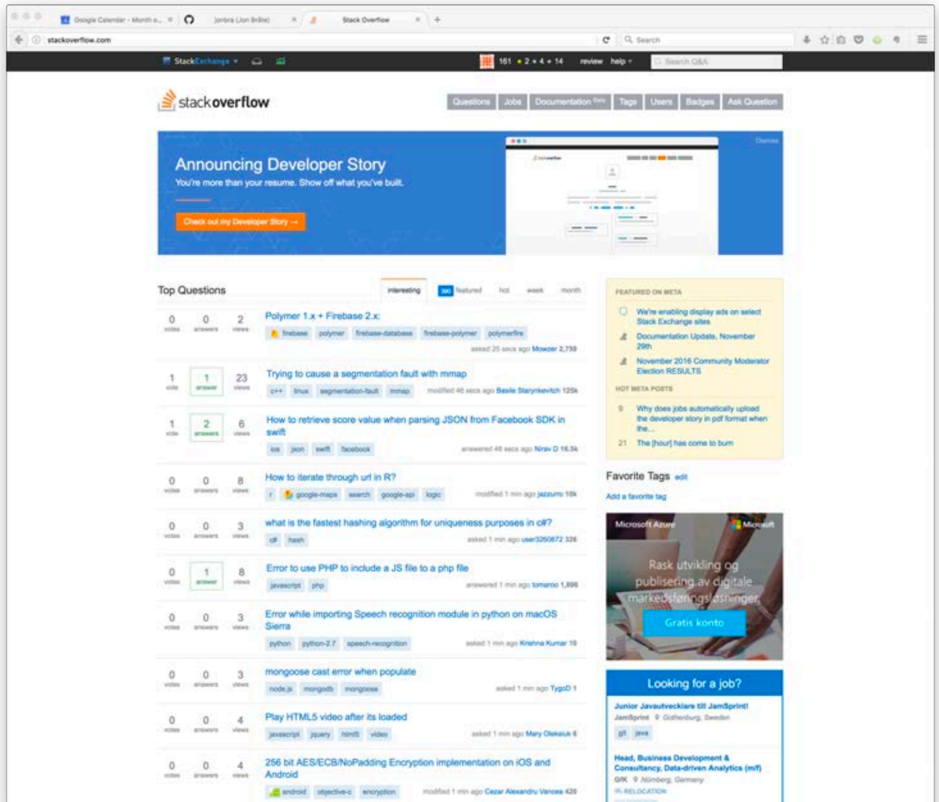
Useful sites – Bioconductor support



Useful sites – Biostars.org



Useful sites – Stackoverflow.com



Useful literature on RNA-seq analysis

PROTOCOL

Count-based differential expression analysis of RNA sequencing data using R and Bioconductor

Simon Anders[1], Davis J McCarthy[2,3], Yunshun Chen[4,5], Michal Okoniewski[6], Gordon K Smyth[4,7], Wolfgang Huber[1] & Mark D Robinson[8,9]

[1]Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, C
Centre for Human Genetics, University of Oxford, Oxford, UK. [4]Bioinformatic
Medical Biology, University of Melbourne, Melbourne, Victoria, Australia. [6]Fu
Statistics, University of Melbourne, Melbourne, Victoria, Australia. [8]Institute o
Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence sho

Published online 22 August 2013; doi:10.1038/nprot.2013.099

PROTOCOL

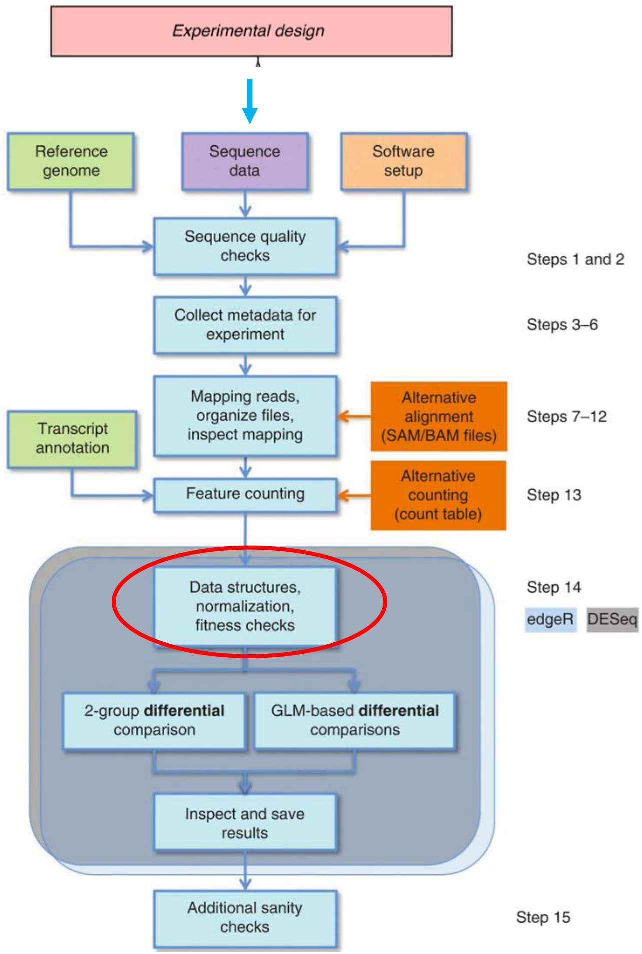
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell[1,2], Adam Roberts[3], Loyal Goff[1,2,4], Geo Pertea[5,6], Daehwan Kim[5,7], David R Kelley[1,2], Harold Pimentel[3], Steven L Salzberg[5,6], John L Rinn[1,2] & Lior Pachter[3,8,9]

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [2]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [3]Department of Computer Science, University of California, Berkeley, California, USA. [4]Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [5]Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [6]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [7]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. [8]Department of Mathematics, University of California, Berkeley, California, USA. [9]Department of Molecular and Cell Biology, University of California, Berkeley, California, USA. Correspondence should be addressed to C.T. (cole@broadinstitute.org).

Published online 1 March 2012; doi:10.1038/nprot.2012.016

Data exploration and quality assessment



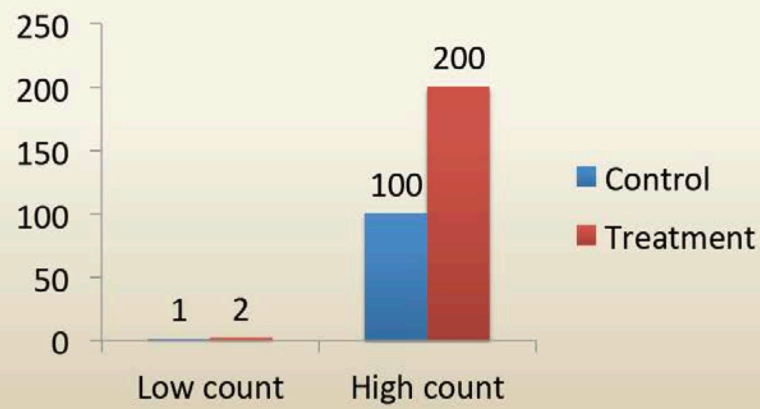
Transformation

- For visualization
- Homoskedastic data – the variance is the same across the means.
- For RNA-seq raw counts, however, the variance grows with the mean. => Higher counts, more variance.
- E.g. PCA plot dominated by highly expressed genes.
- log2-transform common – but now then, small numbers tend to dominate due to strong poisson noise

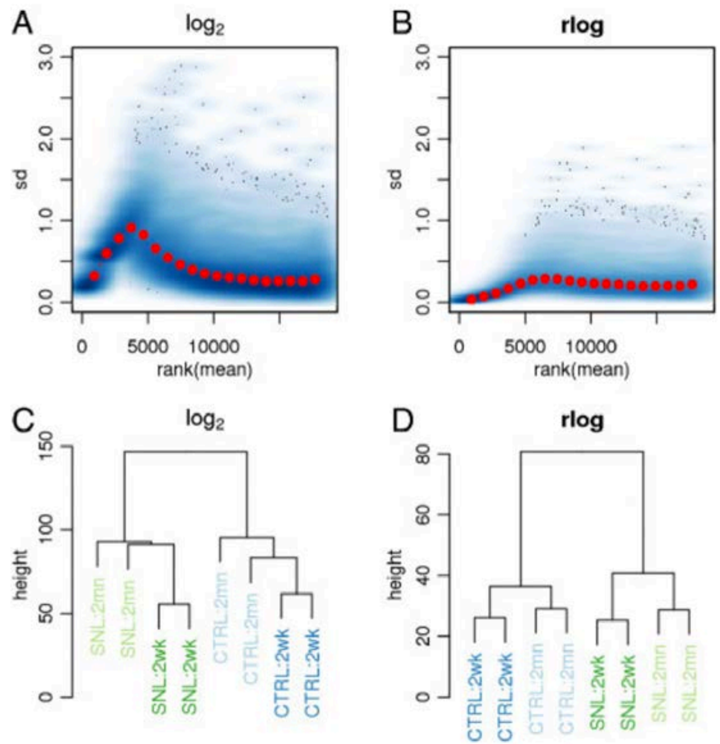
Anders and Huber, 2010
similar approach: Robinson and Oshlack 2010

Strong poisson noise for low count values

1) Poisson counting error
   - Uncertainty in count-based measurements
   - Disproportionately large for low-count data



DESeq2 – variance stabilizing transformation (rlog)

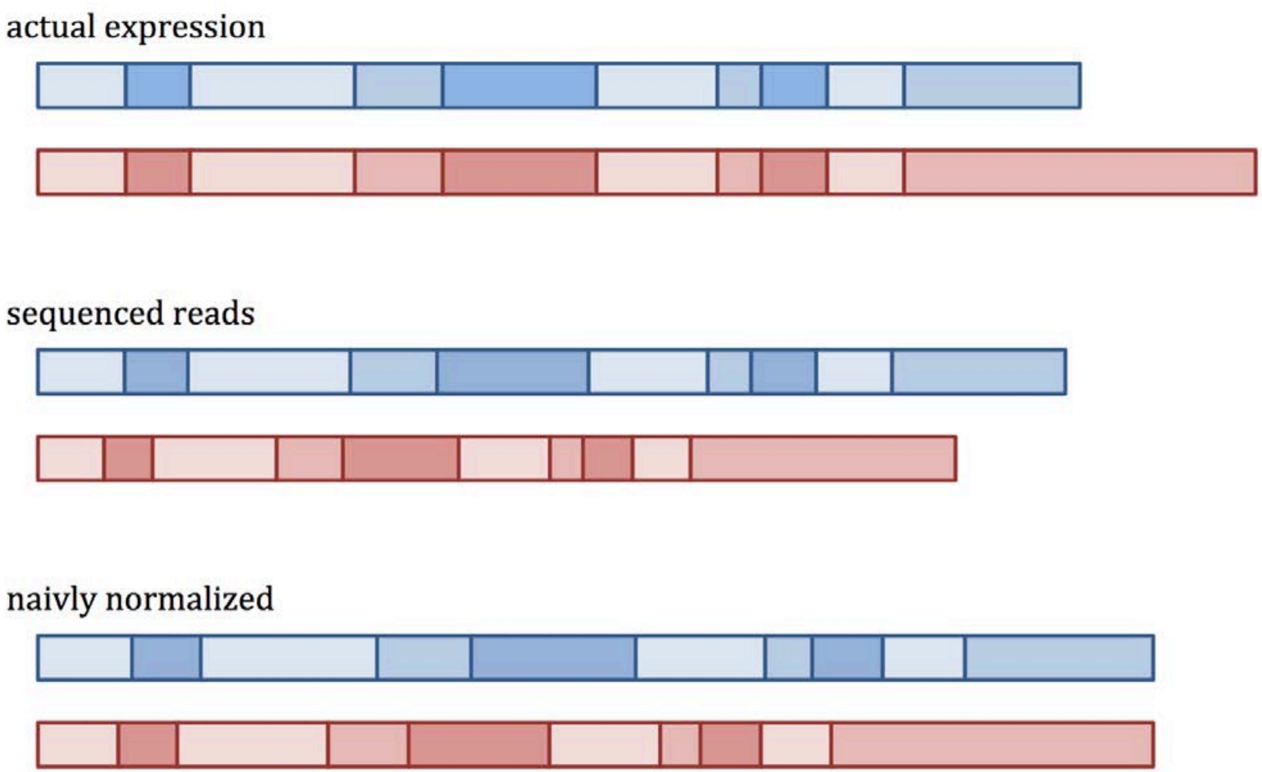


Love et al. 2014 Genome Biology

Normalization for library size

- If sample A has been sampled deeper than sample B, we expect counts to be higher.
- Naive approach: Divide by the total number of reads per sample
- Problem: Genes that are strongly and differentially expressed may distort the ratio of total reads.

Normalization for library size



Normalization for library size

- To compare more than two samples:
- Form a “virtual reference sample” by taking, for each gene, the geometric mean of counts over all samples

- **DESeq2:** Normalize each sample to this reference, to get one scaling factor (“size factor”) per sample.

Anders and Huber, 2010
similar approach: Robinson and Oshlack 2010

Differential expression analysis - distributions

Variation summary, intuitively

Total CV^2 = Technical CV^2 + Biological CV^2

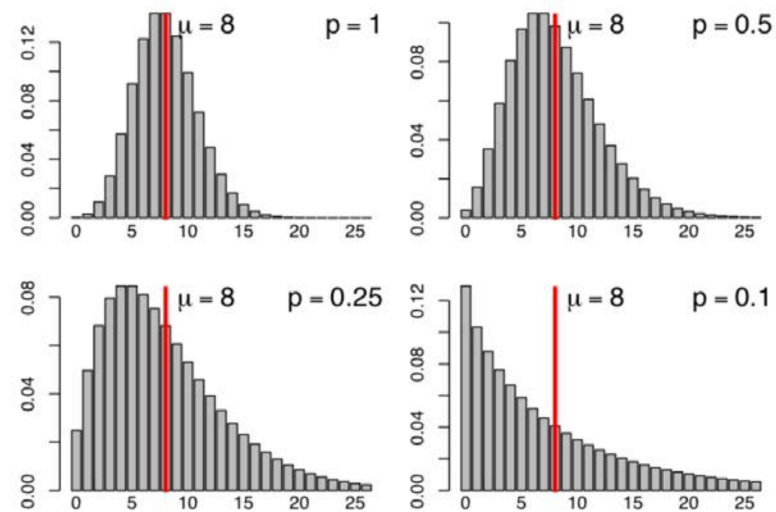
For **low counts**, the Poisson (technical) variation or the measurement error is dominant.

For **higher counts**, the Poisson variation gets smaller, and another source of variation becomes dominant, the **dispersion** or the **biological variation**. Biological variation does not get smaller with higher counts.

VIB

Differential expression analysis

- DESeq2 uses the negative binomial distribution.
- In pairwise DE tests performs a Wald test
- Many genes have zero counts
- Some genes have high counts



$$\Pr(Y = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k \quad \text{for } k = 0, 1, 2, \ldots$$

DE testing– adjusted p-values

Multiple hypothesis testing

- Thousands of genes = thousands of hypothesis tests (simultaneously)
- Increased chance of false positives! (Type I error)
  - e.g. you test for differential expression in 1000 genes <u>that are not differentially expressed</u>
  - You would expect $1000 \times 0.05 = 50$ of them to have a *P*-value < 0.05
- Individual *P*-values not useful
  - Need multiple testing statistic instead

DE testing– adjusted p-values

False discovery rate

(Benjamini & Hochberg 1995)

- The expected proportion of Type I errors among the rejected hypotheses
  - i.e. the proportion of false positives
- Tends to be conservative if many genes are DE
  - FDR = 0.05 common for exploratory/broad scope studies
  - FDR < 0.05 common for medical applications and hunts for candidate genes

Try Bioconductor (DESeq2 and edgeR) yourself

http://folk.uio.no/jonbra/R_DESeq2_exercises.html