



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

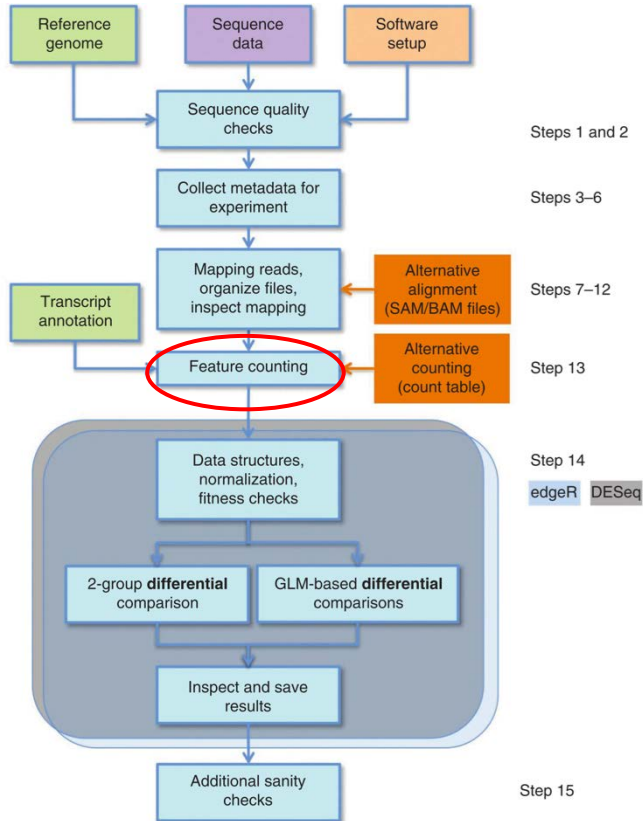
Continue the transcriptomics pipeline
(lectures/practical)

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

After lunch:

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

Summary

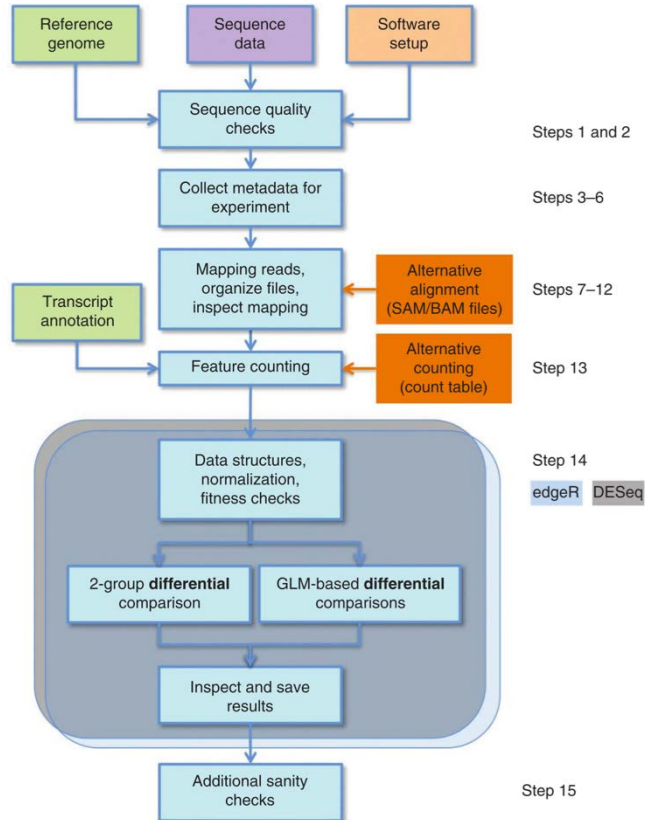


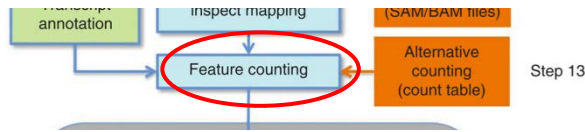
FastQC – *view* fastq files (fastq.gz / fq.gz)

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

TopHat2 – *maps* the trimmed reads to the genome

Counting gene expression





HTSeq-counts

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

HTSeq gives “raw counts”

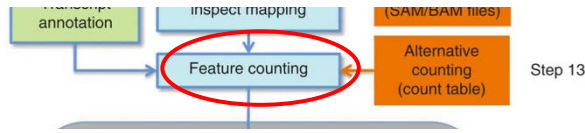
Many programs to count/estimate expression:

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

Counting gene expression

Default	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

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

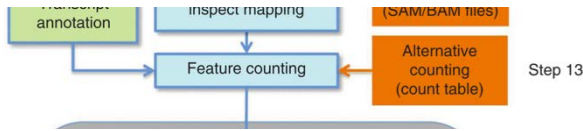
Transcripts per million. Scaled by sequencing depth and length

fpkm/rpkm

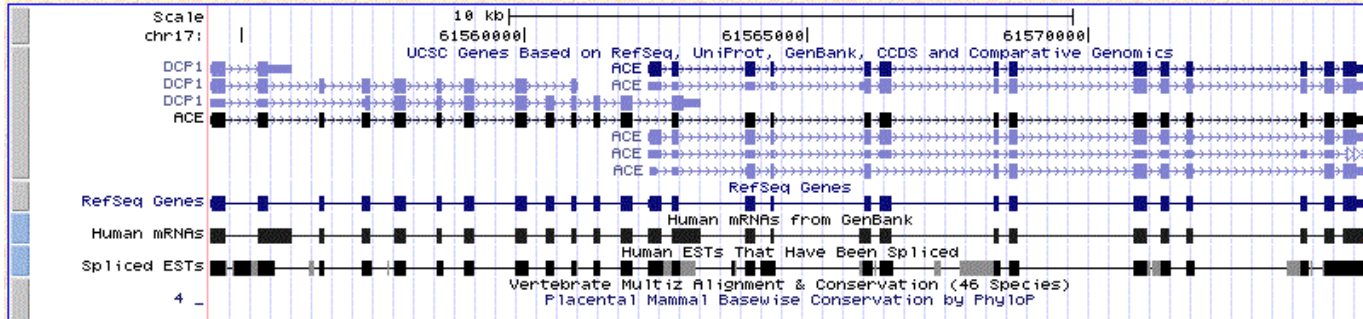
Reads/fragments per kilobase of exon per million reads mapped.

Similar to TPM. Scaled by sequencing depth and length

Annotation files



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



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

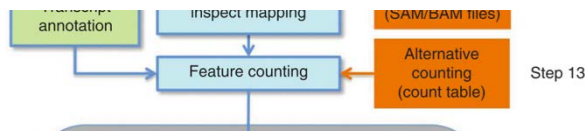
UCSC Genome Bioinformatics

Genomes Genome Browser Tools Mirrors

Frequently Asked Questions: Data File Formats

General formats:

- [Axt format](#)
- [BAM format](#)
- [BED format](#)
- [BED detail format](#)
- [bedGraph format](#)
- [bigBed format](#)
- [bigGenePred table format](#)
- [bigPsl table format](#)
- [bigMaf table format](#)
- [bigChain table format](#)
- [bigWig format](#)
- [Chain format](#)
- [CRAM format](#)
- [GenePred table format](#)
- [GFF format](#)
- [GTF format](#)
- [HAL format](#)
- [MAF format](#)
- [Microarray format](#)
- [Net format](#)
- [Personal Genome SNP format](#)
- [PSL format](#)
- [VCF format](#)
- [WIG format](#)



Annotation files

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

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

Col1: Chromosome

Col2: Can be anything

Col3: Feature type

Col4: Feature start

Col5: Feature stop

Col6: Score

Col7: Strand

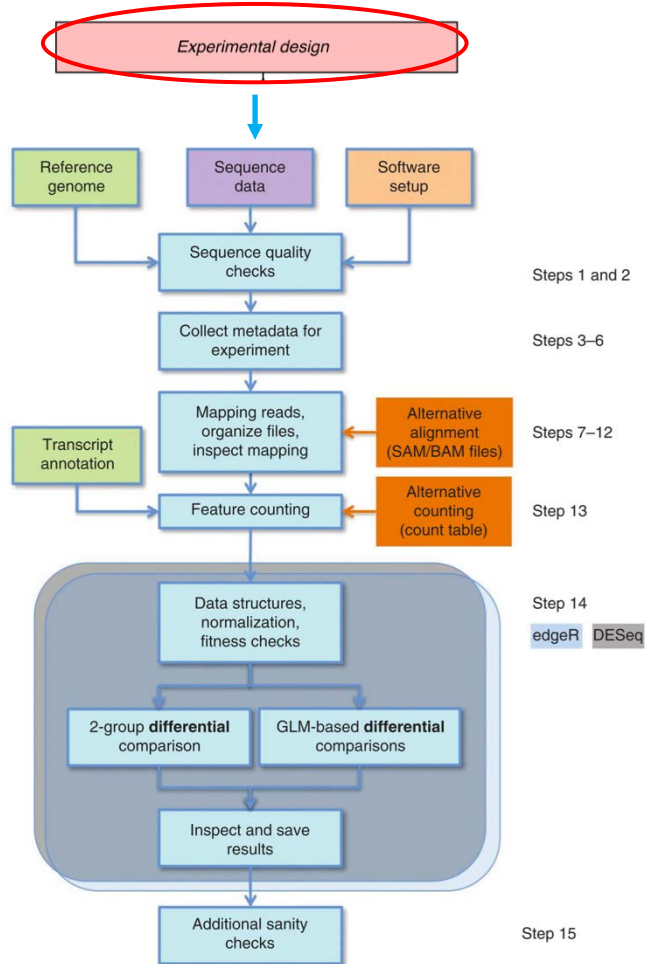
Col8: Frame

Col9: Attributes

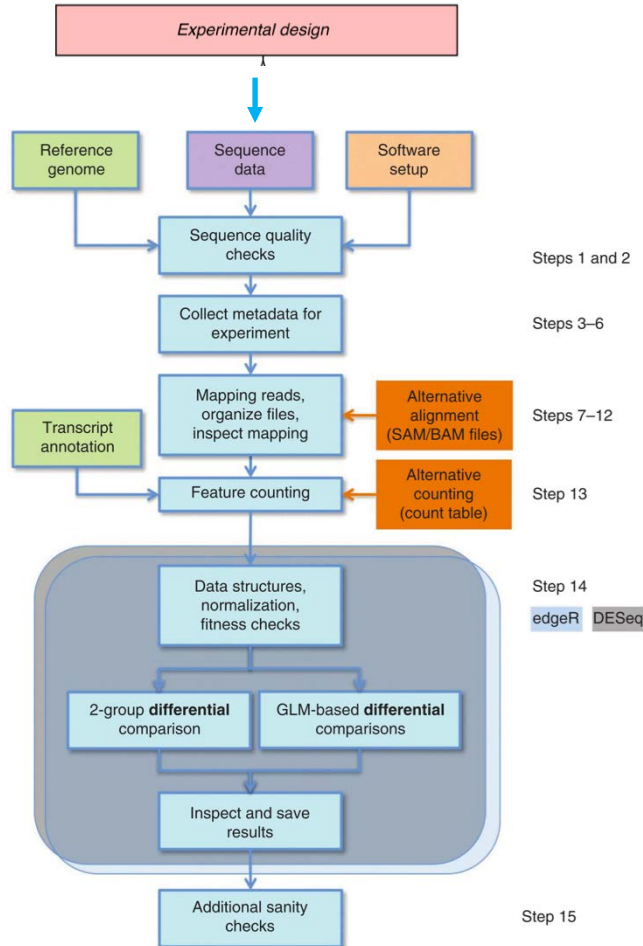
```
bio88156:gff-files-and-expression-levels jonbra$ head cds.gb.gtf
scis4534 coding exon 1972 2017 . - gene_id "scigt006336"; transcript_id "scict006336.2"; exon_number "1"; gene_name "scigt006336";
scis4534 coding exon 4598 5164 . - gene_id "scigt006336"; transcript_id "scict006336.2"; exon_number "2"; gene_name "scigt006336";
scis4534 coding exon 1973 2541 . - gene_id "scigt006336"; transcript_id "scict006336.1"; exon_number "1"; gene_name "scigt006336";
scis1094 coding exon 16322 16468 . + gene_id "scigt000406"; transcript_id "scict000406.1"; exon_number "1"; gene_name "scigt000406";
scis1094 coding exon 16968 17219 . + gene_id "scigt000406"; transcript_id "scict000406.1"; exon_number "2"; gene_name "scigt000406";
scis411 coding exon 252095 252127 . + gene_id "scigt025825"; transcript_id "scict025825.1"; exon_number "1"; gene_name "scigt025825";
scis411 coding exon 252658 252745 . - gene_id "scigt025825"; transcript_id "scict025825.1"; exon_number "2"; gene_name "scigt025825";
scis411 coding exon 253800 253977 . + gene_id "scigt025825"; transcript_id "scict025825.1"; exon_number "3"; gene_name "scigt025825";
scis411 coding exon 254440 254917 . + gene_id "scigt025825"; transcript_id "scict025825.1"; exon_number "4"; gene_name "scigt025825";
scis411 coding exon 254851 255000 . + gene_id "scigt025825"; transcript_id "scict025825.1"; exon_number "5"; gene_name "scigt025825";
bio88156:gff-files-and-expression-levels jonbra$ head cds.gb.gff3
##gff-version 3
###
scis599 coding gene 129921 143406 . - ID=scigt022904;Name=scigt022904
scis599 coding mRNA 129921 132617 4919 . - ID=scigt022904.10;Parent=scigt022904;Name=scigt022904.10
scis599 coding exon 132372 132617 . - Parent=scigt022904.10
scis599 coding exon 132066 132171 . - Parent=scigt022904.10
scis599 coding exon 131787 131842 . - Parent=scigt022904.10
scis599 coding exon 131272 131454 . - Parent=scigt022904.10
scis599 coding exon 130632 130753 . - Parent=scigt022904.10
scis599 coding exon 129921 130191 . - Parent=scigt022904.10
bio88156:gff-files-and-expression-levels jonbra$
```


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

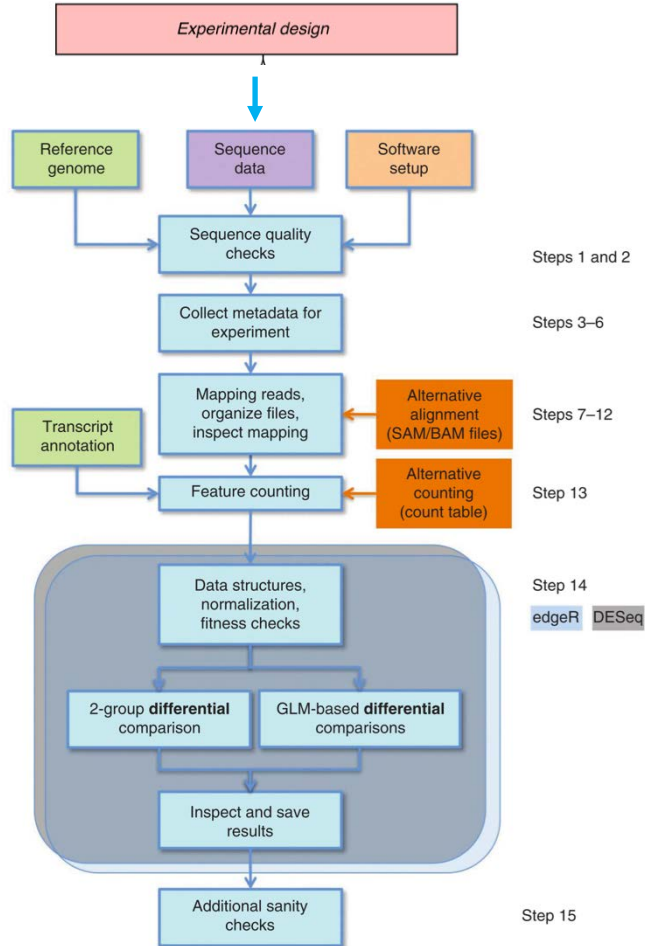
Open Access

A survey of best practices for RNA-seq data analysis

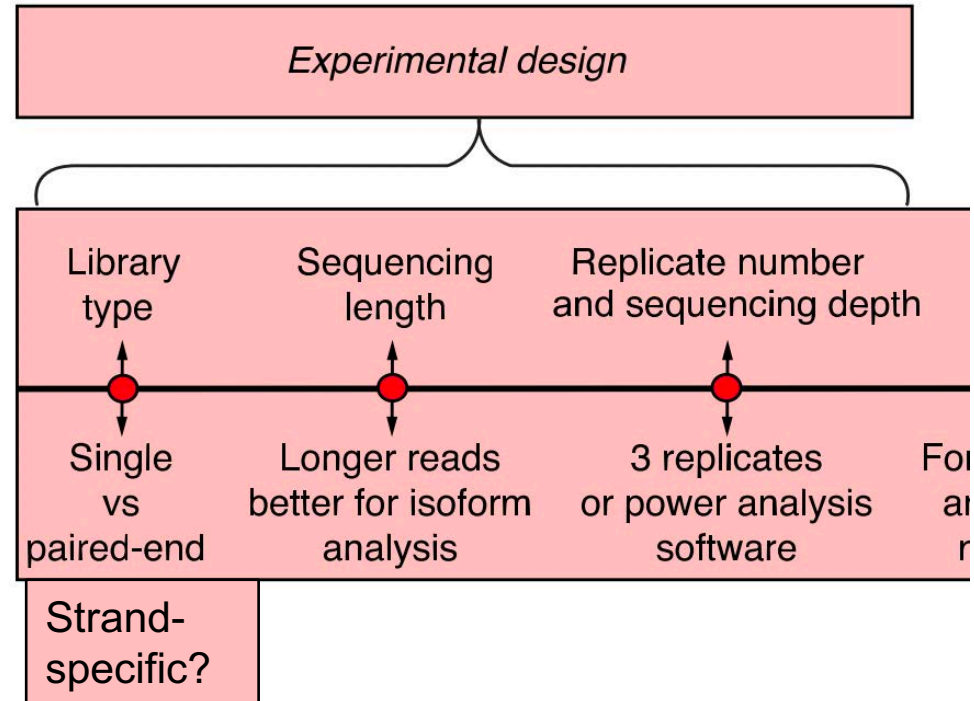


Ana Conesa^{1,2*}, Pedro Madrigal^{3,4*}, Sonia Tarazona^{2,5}, David Gomez-Cabrero^{6,7,8,9}, Alejandra Cervera¹⁰, Andrew McPherson¹¹, Michał Wojciech Szczęśniak¹², Daniel J. Gaffney³, Laura L. Elo¹³, Xuegong Zhang^{14,15} and Ali Mortazavi^{16,17*}

Experimental design



(a)



Experimental design

Goal

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

Experimental design - replicates

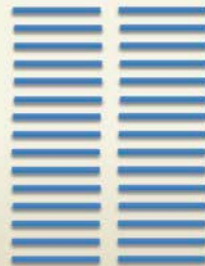
Differential expression analysis

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

Experimental design - replicates

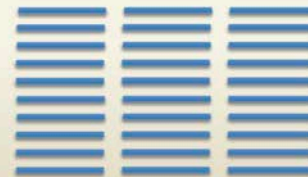
Cost

- Tradeoff between sequencing depth and replication
 - More power comes from biological replication!



15 M reads x 2 reps

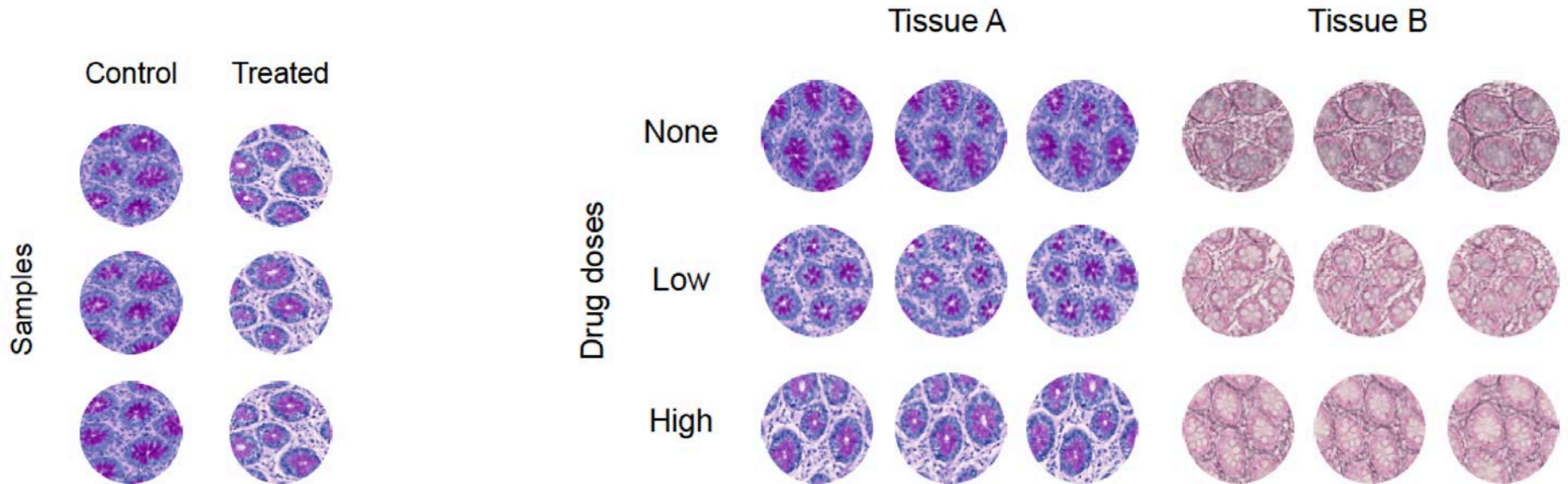
**35% higher DE
detection**



10 M reads x 3 reps

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.

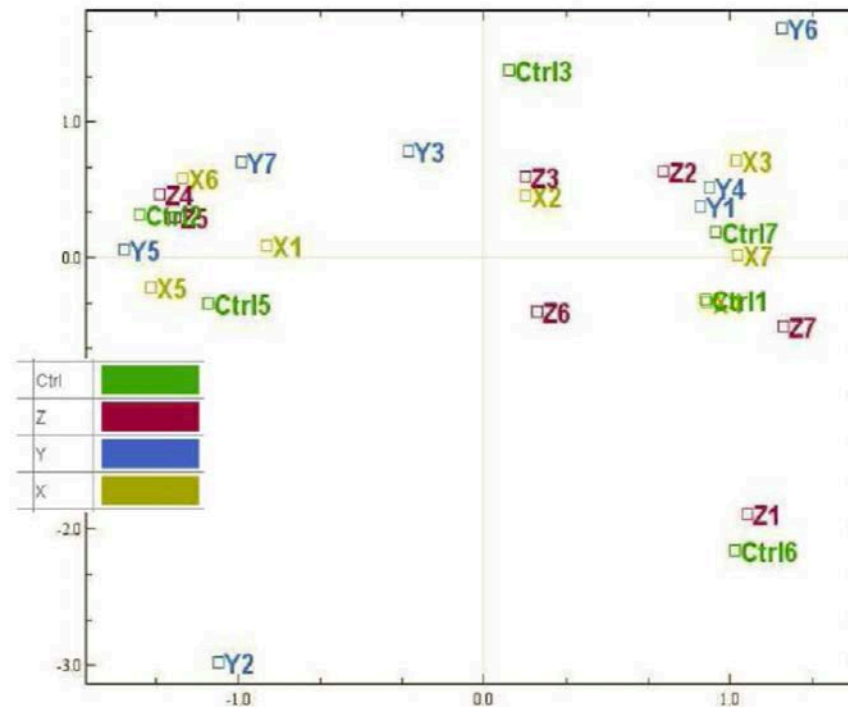
Experimental design: an example

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

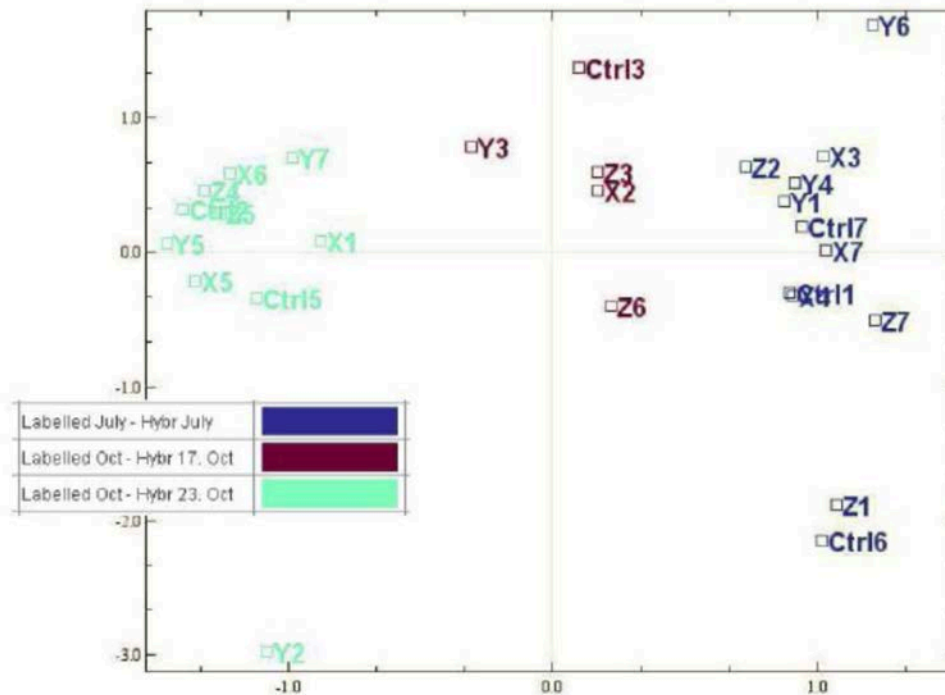
Biology	Sample preparation order	Extraction order
A2	10	1
B6	5	2
C1	12	3
A5	7	4
B5	14	5
C6	9	6
A6	16	7
B4	2	8
C5	18	9
A3	4	10
C3	15	11
B2	8	12
A4	13	13
C4	6	14
B1	17	15
A1	1	16
B3	11	17
C2	3	18

Experimental design: batch effect



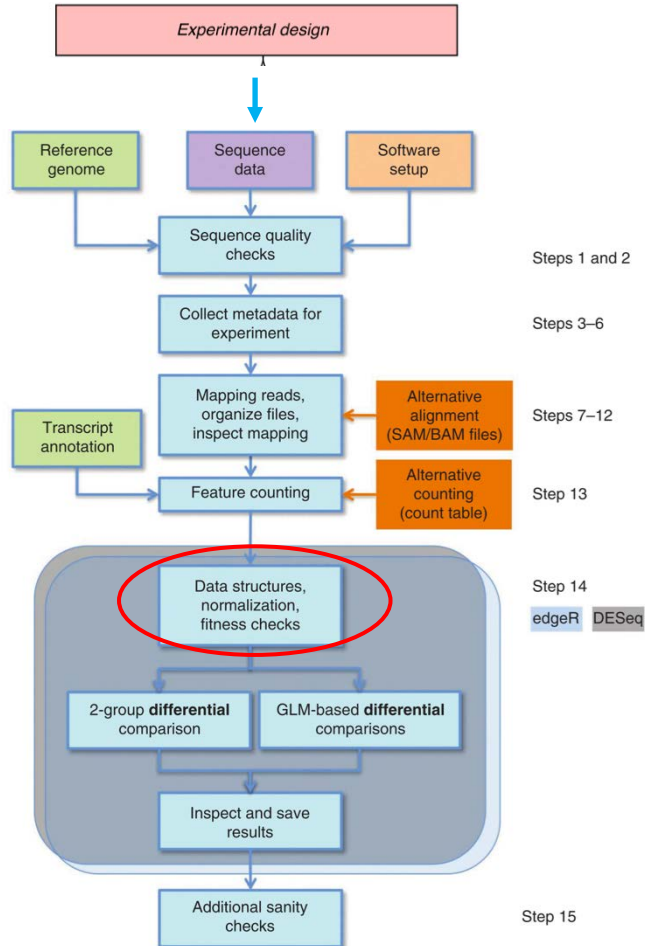
Samples color coded according to biology

Experimental design: batch effect



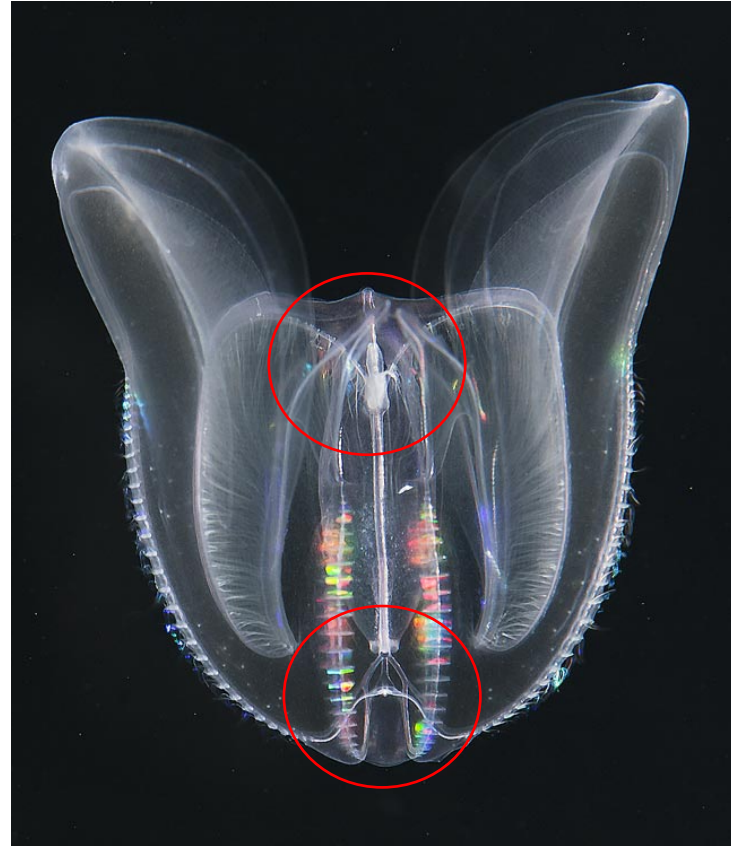
Samples color coded according to labeling date

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

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


National Human Genome Research Institute
Advancing human health through genomics research

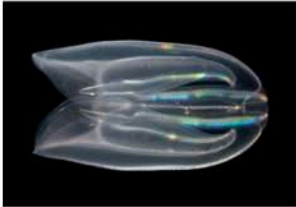
SEARCH GENOME.GOV

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.

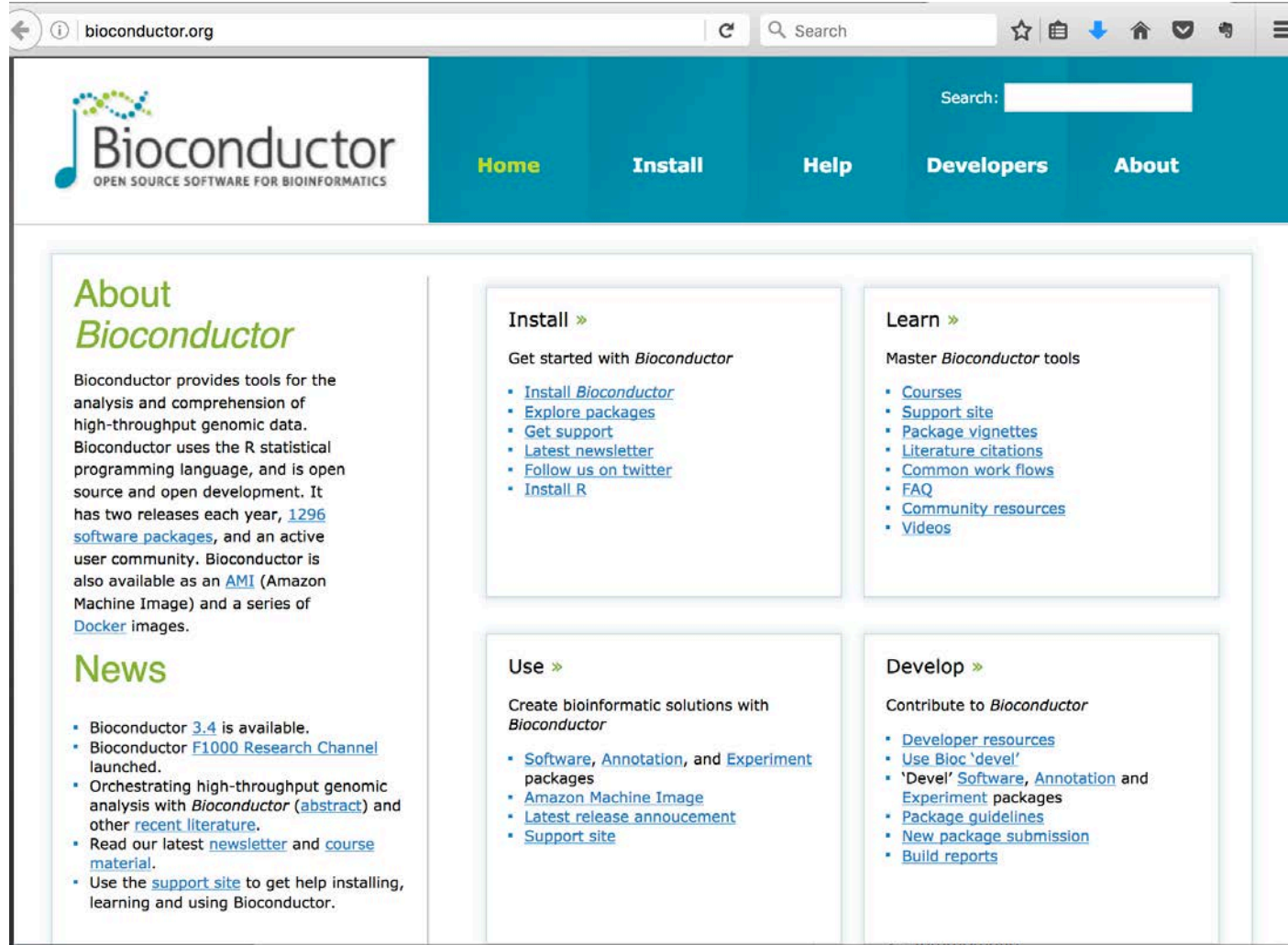
Mnemiopsis leidyi is a ctenophore native to the coastal waters of the western Atlantic Ocean. A number of studies on *Mnemiopsis* have led to a better understanding of many key biological processes, including regeneration and axial patterning, and these studies have contributed to the emergence of *Mnemiopsis*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

Genomic sequencing of non-bilateria 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



The screenshot shows the Bioconductor website in a web browser. The browser's address bar displays 'bioconductor.org'. The website's header features the Bioconductor logo on the left, which includes a stylized DNA helix and the text 'Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS'. To the right of the logo is a teal navigation bar with links for 'Home', 'Install', 'Help', 'Developers', and 'About'. A search bar is located in the top right corner of the header. The main content area is divided into several sections. On the left, there is an 'About Bioconductor' section with a green heading and a paragraph describing the project. Below this is a 'News' section with a green heading and a list of recent updates. To the right of the 'About' section are two columns of links. The top column is titled 'Install »' and lists links for installing Bioconductor, exploring packages, getting support, the latest newsletter, following on Twitter, and installing R. The bottom column is titled 'Learn »' and lists links for courses, a support site, package vignettes, literature citations, common work flows, FAQ, community resources, and videos. Below the 'Install' section are two more columns. The top column is titled 'Use »' and lists links for software, annotation, and experiment packages, Amazon Machine Image, latest release announcement, and support site. The bottom column is titled 'Develop »' and lists links for developer resources, using Bioc 'devel', 'Devel' software, annotation and experiment packages, package guidelines, new package submission, and build reports.

bioconductor.org

Search:

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

[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 announcement](#)
- [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 packages 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

The screenshot shows the Bioconductor support forum homepage. The browser address bar displays <https://support.bioconductor.org>. The page features a navigation bar with links for 'My messages', 'votes', 'posts', 'tags', 'following', and 'bookmarks'. A user profile for 'Jon Brille' is visible in the top right corner. The main content area is a list of forum posts, each with a title, a brief description, and a timestamp. The posts are sorted by 'Latest' and include a search bar. The right sidebar contains sections for 'Recent...' (a list of recent posts), 'Votes' (a list of votes), and 'Awards' (a list of awards). The footer indicates that 49 users are online in the last hour.

Google Calendar - Month <...> jandra Lin (Brille) Latest Posts

https://support.bioconductor.org

My messages | votes | posts | tags | following | bookmarks Jon Brille | 130 | log out | about | flag | msg to

Bioconductor
an open source project for bioinformatics

ASK QUESTION LATEST 118 NEWS JOBS TUTORIALS TAGS USERS

You have successfully logged in as jon.brille@biu.uio.no.

Limit Sort Search

1 3 31 matching of AASTringSet vs. another AASTringSet
biostrings astringset written 1 day ago by tobias.kockmann • 0 • updated 8 hours ago by Hervé Pagès • 11k

0 0 19 H/W and correlated covariates
rhr written 13 hours ago by Michael Love • 9.7k

0 2 66 Bug in cleanUpPhenoData
arrayqualitymetrics written 5 days ago by flo • 0 • updated 17 hours ago by Wolfgang Huber • 12k

3 2 116 Error in install.packages: unable to install packages
afly bioconductor written 17 days ago by Khawaja • 0 • updated 20 hours ago by Agnès Hussain Wani • 190

0 3 94 Issues with SGSeq: analyzeVariants method returns an error...
analyzevariants sgseq written 8 weeks ago by Sylvain Foley • 20 • updated 1 day ago by Leonard Goldstein • 20

0 0 30 Error in assay colnames when using readGEORawFile function in "minfi" package
minfi methylation 450k written 1 day ago by moorken • 0

1 1 29 how to make findOverlaps from genomicAlignments package to consider soft-clipped parts of BWA paired reads in overlap calculation?
findoverlaps clipped read genomicalignments written 1 day ago by jordi • 10 • updated 1 day ago by Hervé Pagès • 11k

1 2 35 M3D package in Bioconductor 3.4
methylation bioconductor written 1 day ago by francescaccioli • 0 • updated 1 day ago by James W. MacDonald • 41k

0 1 46 Return features from a single strand using vmatchPattern
biostrings genomicranges written 5 days ago by jma1991 • 0 • updated 1 day ago by Hervé Pagès • 11k

0 1 39 DESeq2: Complex Differential Design
deseq2 written 1 day ago by dqqquest • 0 • updated 1 day ago by Michael Love • 9.7k

3 1 40 DESeq2 to clustered metagenomic data
metagenomics deseq2 written 1 day ago by Elenndi • 0 • updated 1 day ago by Michael Love • 9.7k

0 1 22 openCyto trouble importing xml file
R opencyto bioconductor written 1 day ago by Rvika • 0 • updated 1 day ago by Jiang, Mike • 820

0 0 26 signR(): wrong sign in 'by' argument
signr written 1 day ago by Miguel Julá • 0

0 1 41 Problem with VariantAnnotation and VCF "R" genotype fields when expanding CollapsedVCF
variantannotation bug written 4 days ago by Sean Davis • 20k • updated 1 day ago by Valerie Owencham • 5.7k

0 0 33 Error in pr_DBSget_entry(method): could not find function "pr_DBSget_entry"
software error dplyr bug written 1 day ago by meodigazquez • 0

0 1 26 How to fix ChIPPeakAnno error in adding gene names to annotated peaks?
chippeakanno annotation written 1 day ago by meodigazquez • 0

Recent...
Replies
• C: Gviz: How to color parts... by sethys.n.mariavannan • 0
• A: matching of AASTringSet... by Hervé Pagès • 11k
• A: Bug in cleanUpPhenoData by Wolfgang Huber • 12k
• A: Error in install.packages... by Agnès Hussain Wani • 190
• C: Error in install.packages... by Agnès Hussain Wani • 190

Votes
• Can I set a path for readCo...
• A: Is it sensible to use p...
• A: Is it sensible to use p...
• Is it sensible to use p...
• A: Meta-analysis of gene ex...

Awards • All •
• Scholar @ to Aaron Lun • 12k
• Scholar @ to Michael Lawrence • 8.7k
• Commentator @ to Aaron Lun • 12k
• Scholar @ to Steve Liangou • 11k
• Appreciated @ to Aaron Lun • 12k
• Scholar @ to Dan Tenenbaum • 8.1k

Locations • All •
• Italy, 1 hour ago

Traffic: 49 users online in the last hour

Useful sites – Biostars.org

The screenshot shows the Biostars.org website interface. At the top, there's a navigation bar with links for LATEST, OPEN, RNA-SEQ, CHIP-SEQ, SNP, ASSEMBLY, TUTORIALS, TOOLS, JOBS, FORUM, PLANET, and ALL. Below this is a search bar and a login section. The main content area displays a list of questions and answers, each with a title, a brief description, and a link to the full post. The questions are sorted by relevance, and the list includes various topics related to bioinformatics, such as RNA-Seq, ChIP-Seq, and genome analysis. On the right side, there are sections for 'Recent Votes' and 'Recent Locations', which provide additional context and resources for the user.

Biostars
BIOINFORMATICS EXPLAINED

Latest Posts

Search

about • tags • new

You have successfully signed in as jon.brat@gmail.com.

Live search: start typing... or Q Classic search

Limit to: all time <prev • 41,985 results • page 1 of 1400 • next >

Sort by: update

0 0 6 how to plot heatmap for modules in wgcna
genome next-gen chip-seq rna-seq snp
written 23 minutes ago by rajeekargutha • 20

0 0 32 Weird base qualities and sequences from FastQ file?
assembly fastq next-gen rna-seq
written 3 hours ago by gemelcar • 19 • updated 35 minutes ago by John • 8.6k

0 0 13 Is there any tools for Ortholog hit ratio calculation ?
alignment phy blast
written 1 hour ago by Farbod • 2.7k

0 0 15 Output from featureCounts() as input to DeSeq2
rna-seq dseq2
written 1 hour ago by Elizabeth Sam • 0

5 2 856 Obtaining sequences of a particular gene under a particular taxonomy (using E-utilities)
gene sequence
written 2.5 years ago by q660 • 9 • updated 2 hours ago by BioStar • 10

0 0 33 Run transmembrane prediction with one multifasta file and multiple MSA sequences.
prediction postblast
written 4 hours ago by erik.tug • 0

0 3 81 Sequence search using a pattern
sequence
written 11 hours ago by amargnatsatharwal • 0

0 0 45 Best Program(s) for Predicting Protein-Membrane Interactions?
membrane prediction protein
written 7 hours ago by ericbrenner • 0

6 2 670 Efficiently join paired-end read coordinates in the same line?
3c paired-end hic 4c
written 2.1 years ago by daniel.xoromates • 320 • updated 8 hours ago by BioStar • 10

0 0 53 change in the structure and the data types of the data frame after the transposition by R
transpose data
written 9 hours ago by ED • 10

4 0 68 Tool: Introducing Clumpify: Create 30% Smaller, Faster Gzipped Fastq Files
clumpify gzip tool compression bitmap storage
written 8 hours ago by Brian Bushnell • 7.0k

0 1 67 RNA-Seq technologies for genes with similar sequences
rna-seq
written 9 hours ago by dltarshym • 10 • updated 9 hours ago by Devon Ryan • 59k

0 1 73 convert one column values into row values based on other row values
R
written 10 hours ago by chemengut • 0 • updated 9 hours ago by versu • 3.0k

1 1 72 Unsupervised subtype discovery
R microarray
written 13 hours ago by fucanj • 60 • updated 10 hours ago by Anil • 810

0 0 64 filtering some sequences from txt file in python
sequence
written 11 hours ago by askan • 30

0 0 126 1000 genomes, meaning of the color of pie charts for population genetic
1000genomes snp population genetics

Recent Votes

- C: GGC: Removing RNA-Seq data for Tumor vs. Matched normal tissue
- Introducing Clumpify: Create 30% Smaller, Faster Gzipped Fastq Files
- Analyzing methylation in geo-data
- C: Introducing Clumpify: Create 30% Smaller, Faster Gzipped Fastq Files
- Introducing Clumpify: Create 30% Smaller, Faster Gzipped Fastq Files
- Introducing Clumpify: Create 30% Smaller, Faster Gzipped Fastq Files
- A: Best tool for finding Boundary Plans

Recent Locations • All •

- UK, just now
- Malaysia, 4 minutes ago
- United Kingdom, 7 minutes ago
- India, 10 minutes ago
- Belgium, 16 minutes ago
- Montreal, 18 minutes ago

Recent Awards • All •

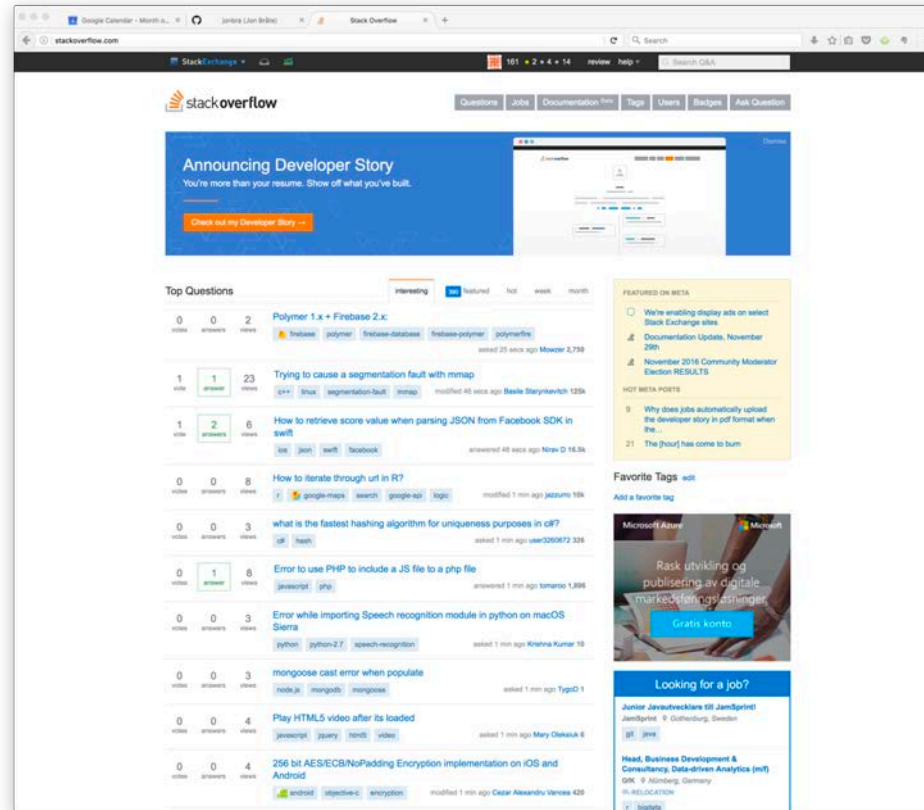
- Popular Question • to fucanj • 60
- Commentator • to hanikah.sahel • 2.8k
- Popular Question • to Pierre Underbaum • 87k
- Scholar • to Brian Bushnell • 7.0k
- Appreciated • to alandromsey • 40
- Scholar • to alandromsey • 40

Recent Replies

- C: Output from featureCounts() as input to DeSeq2 by WouterDeCoster • 9.3k
The work flow OP wants to use is fine, no need for trinity
- C: where to download genePattern.zip file by camellia • 500
Thanks. But I mean I cannot get the modules online. The download page is down.
- C: Output from featureCounts() as input to DeSeq2 by Tharsh • 360
Typical RNA-seq data analysis is as follows: Agt: Align the fastq reads onto the Genome using HIS...
- C: Is there any tools for Ortholog hit ratio calculation? by Farbod • 2.7k
NOTE: I have tried [CRIS-BLAST1] and [Trinity2] Full length transcript count. I want something...
- C: Script to assess the transcriptome assembly quality in terms of blast and

Traffic: 120k views visited in the last hour

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 Anders¹, Davis J McCarthy^{2,3}, Yunshun Chen^{4,5}, Michal Okoniewski⁶, Gordon K Smyth^{4,7},
Wolfgang Huber¹ & Mark D Robinson^{8,9}

¹Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany. ²Department of Statistics, University of Oxford, Oxford, UK. ³Department of Statistics, University of Oxford, Oxford, UK. ⁴Bioinformatics Centre for Human Genetics, University of Oxford, Oxford, UK. ⁵Bioinformatics Centre for Human Genetics, University of Oxford, Oxford, UK. ⁶Bioinformatics Centre for Human Genetics, University of Oxford, Oxford, UK. ⁷Statistics, University of Melbourne, Melbourne, Victoria, Australia. ⁸Statistics, University of Melbourne, Melbourne, Victoria, Australia. ⁹Institute of Bioinformatics, University of Zurich, Zurich, Switzerland. Correspondence should be addressed to Mark D Robinson (m.d.robinson@euzh.unizh.ch).

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

PROTOCOL

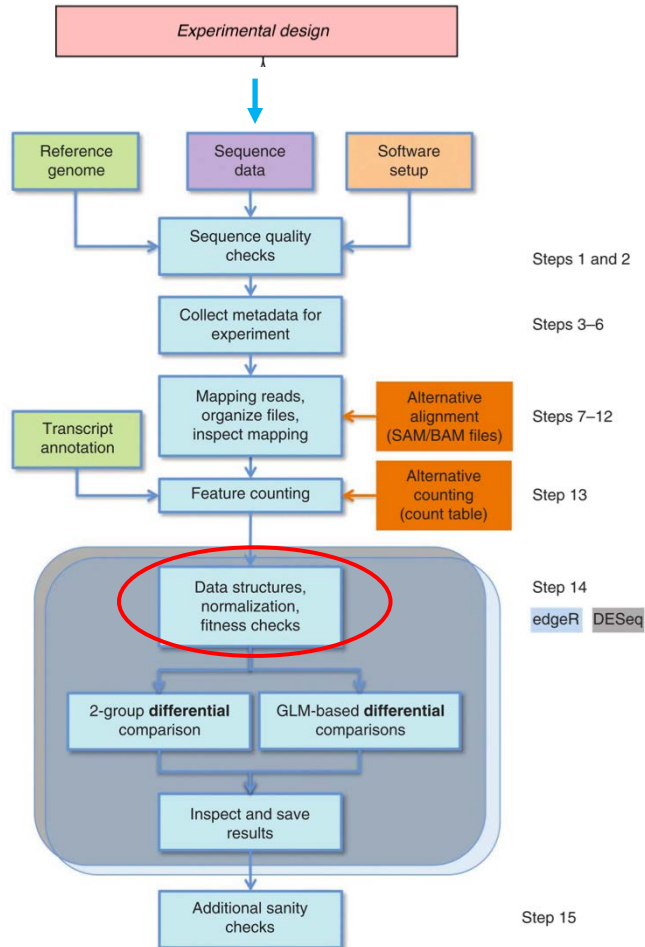
Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell^{1,2}, Adam Roberts³, Loyal Goff^{1,2,4}, Geo Pertea^{5,6}, Daehwan Kim^{5,7}, David R Kelley^{1,2}, Harold Pimentel³, Steven L Salzberg^{5,6}, John L Rinn^{1,2} & Lior Pachter^{3,8,9}

¹Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. ²Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. ³Department of Computer Science, University of California, Berkeley, California, USA. ⁴Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. ⁵Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. ⁶Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. ⁷Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. ⁸Department of Mathematics, University of California, Berkeley, California, USA. ⁹Department 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



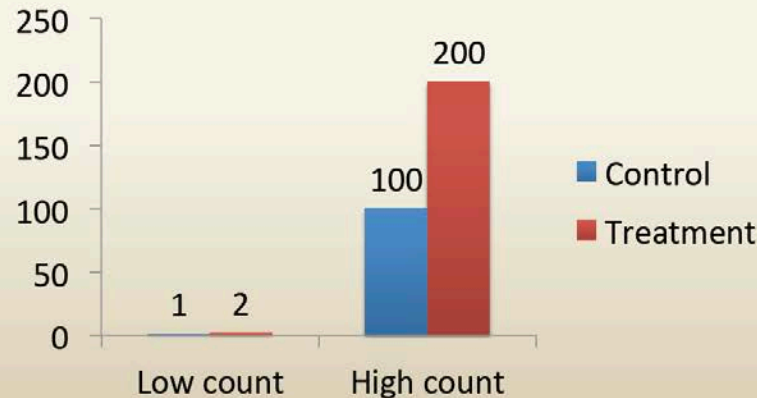
Transformation

- 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

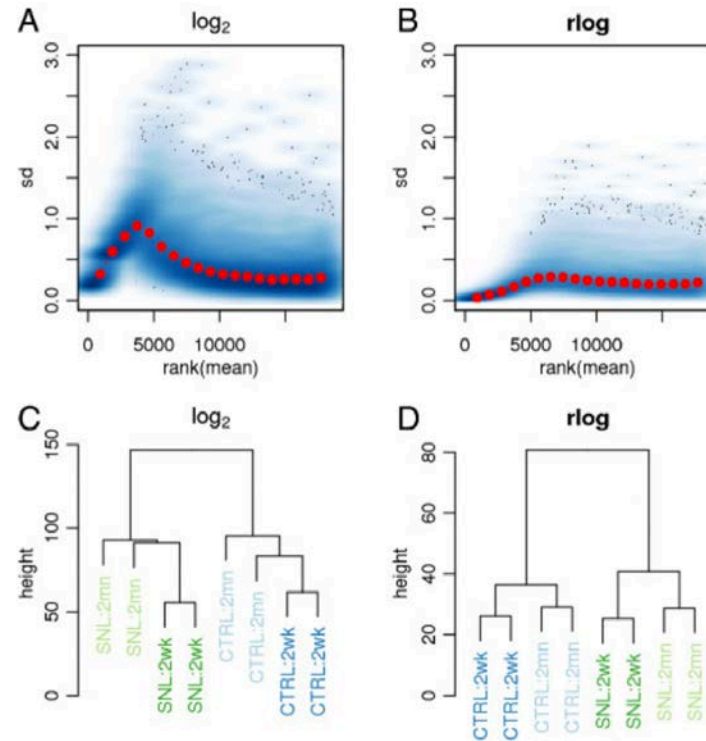
Strong poisson noise for low count values

1) Poisson counting error

- Uncertainty in count-based measurements
- Disproportionately large for low-count data



DESeq2 – variance stabilizing transformation (rlog)



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

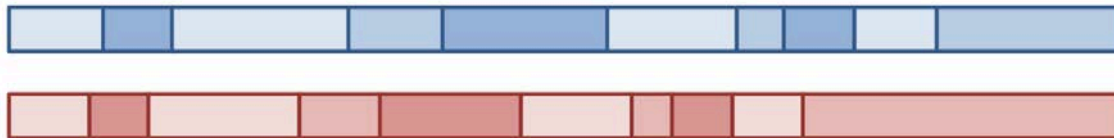
actual expression



sequenced reads



naively normalized



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.

Differential expression analysis - distributions

Variation summary, intuitively

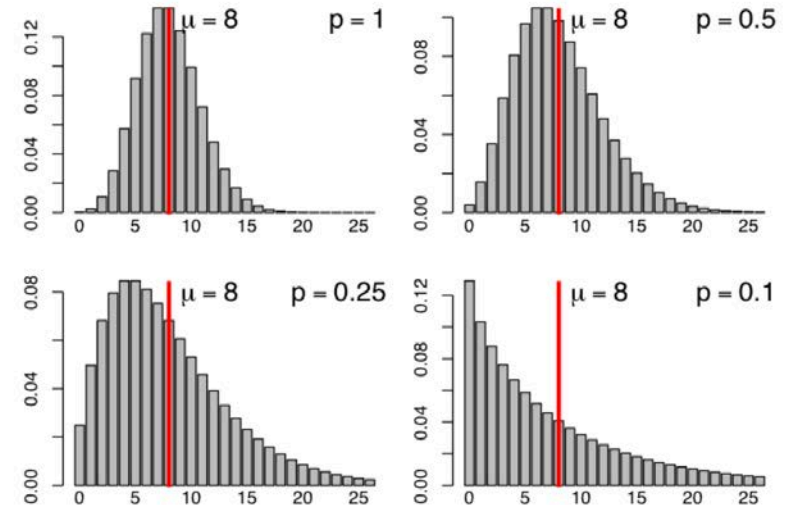
$$\text{Total CV}^2 = \text{Technical CV}^2 + \text{Biological CV}^2$$

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.

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



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

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 that are not differentially expressed
 - You would expect $1000 \times 0.05 = 50$ of them to have a $P\text{-value} < 0.05$
- 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
 - $\text{FDR} = 0.05$ common for exploratory/broad scope studies
 - $\text{FDR} < 0.05$ common for medical applications and hunts for candidate genes

Try Bioconductor (DESeq2 and edgeR) yourself

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