

From Experimental Design to Data Interpretation

AN INTRODUCTION TO BIOINFORMATICS FOR FUNCTIONAL
GENOMICS

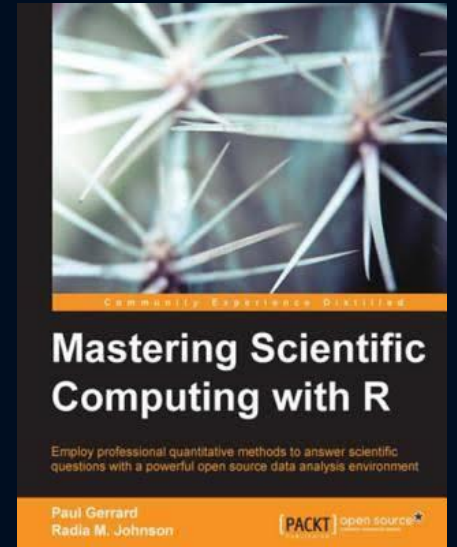
RADIA JOHNSON, PH.D.

About me

- **Research Associate Bioinformatician**

- ❖ Ph.D. in Immunology (University of Toronto)
- ❖ Post-Docs in Functional Genomics (University of Cambridge, UK & UdeM)
- ❖ Specialized in Immunology, Hematology & Cancer Biology
- ❖ Programming experience in R, Bash Shell Scripting, Perl, Python, Java Script
- ❖ Co-author of the book “**Mastering Scientific Computing with R**”

- Office: GCRC, room 536A
- Email: radia.johnson@mcgill.ca
- Webpage: <https://rjbioinformatics.com>



Outline

- Overview of bioinformatics & its role in functional genomics
- RNA-seq data challenges & recommendations
- Overview of the material covered for the upcoming workshops

What is Bioinformatics?

- Integrates biology and computer science
- Apply mathematical, statistical and computing methods to solve biological problems

Examples:

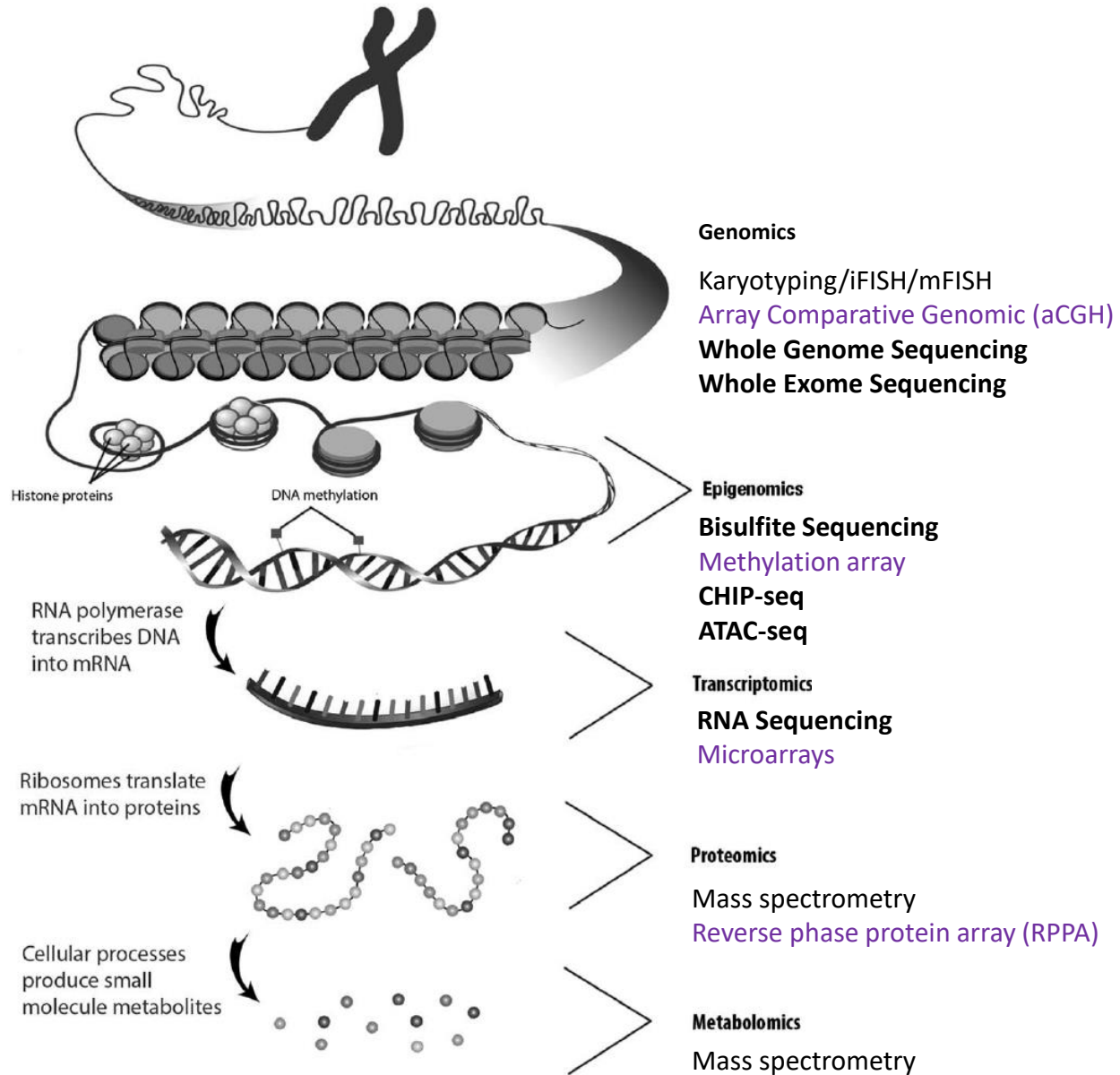
- ❑ *Sequence analysis e.g. mutations*
- ❑ *Protein structure analysis*
- ❑ *Quantify gene expression*
- ❑ *Mining Databases*

Functional genomics

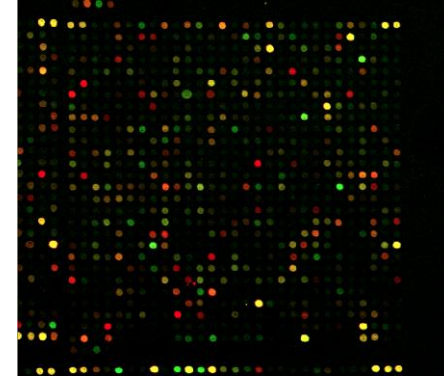
- “Functional genomics uses genomic data to **study gene and protein expression and function on a global scale** (genome-wide or system-wide), focusing on **gene transcription, translation** and **protein-protein interactions**, and often involving high-throughput methods.”

Nature.com

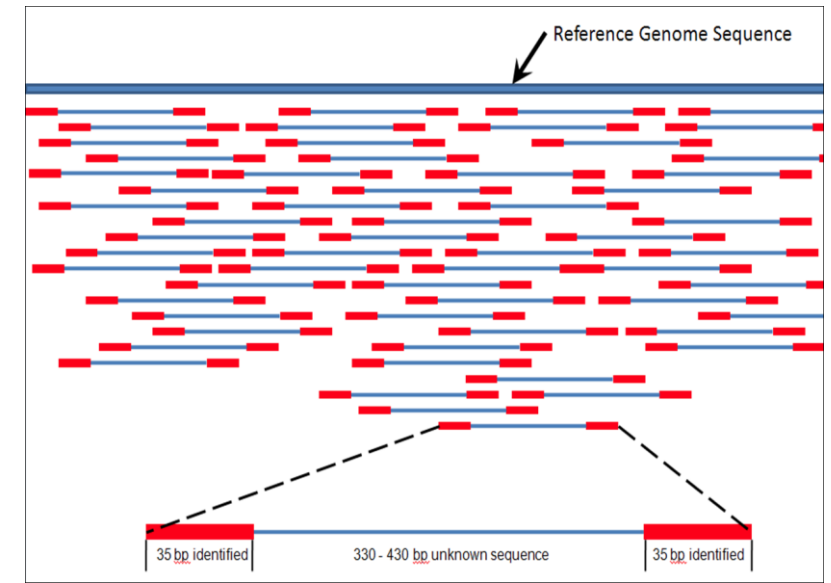
Methods used in functional genomics



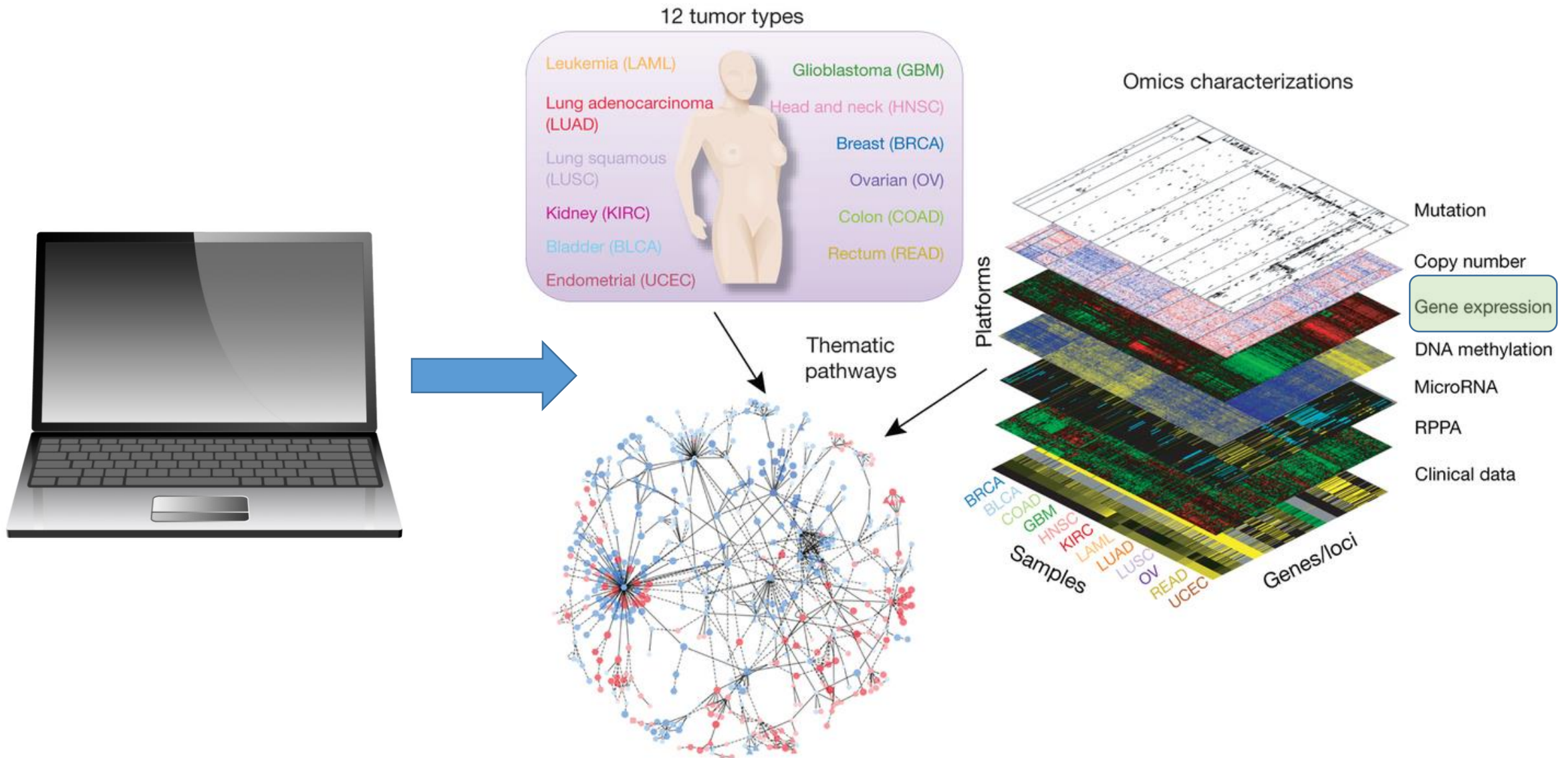
Array-based



Next Generation Sequencing

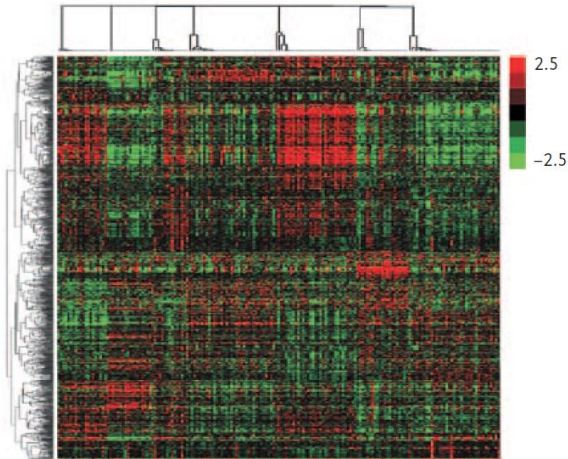


Bioinformatics plays an integral role in data analysis and integration

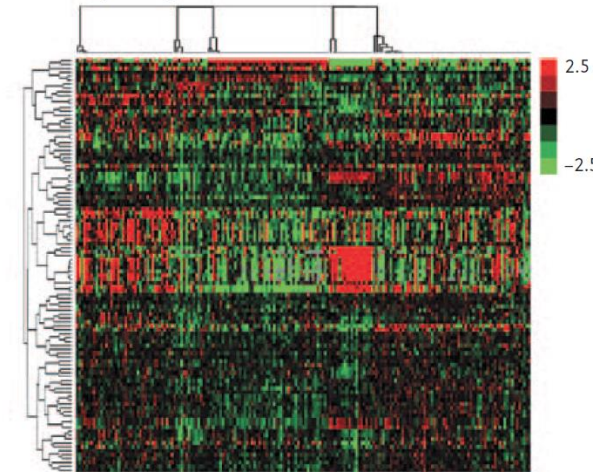


How can we visualize this data?

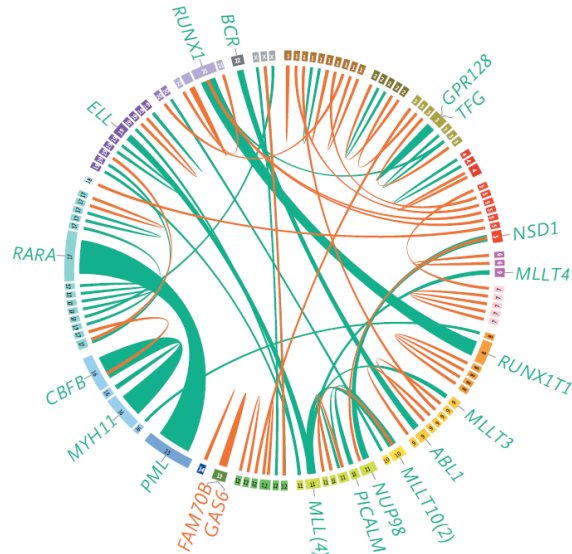
RNA Sequencing



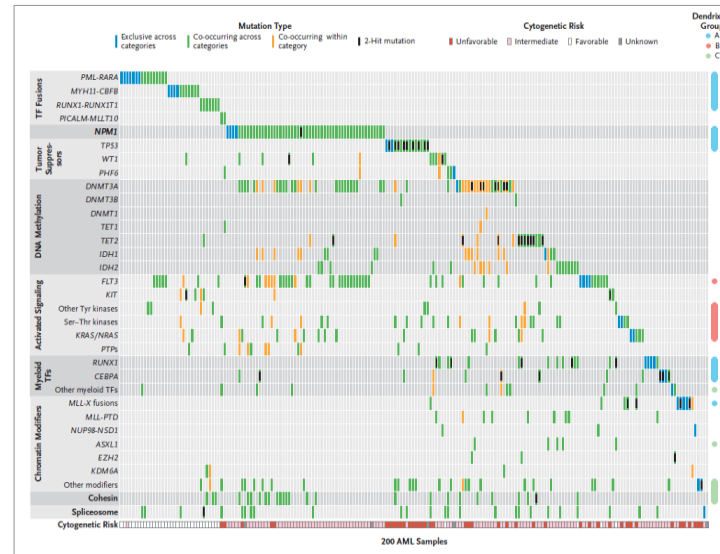
miRNA Sequencing



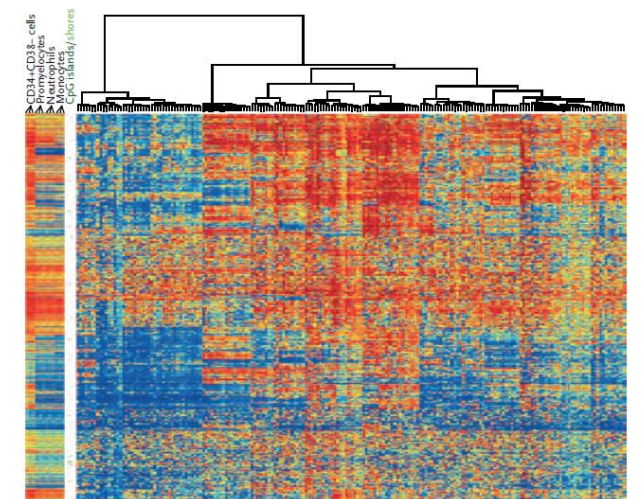
Gene fusions



Clinical Features and Mutations



Methylation CpG Sparse regions



Data from TCGA Research Network, NEJM 2013

How do we analyse genomics data?



www.r-project.org
www.bioconductor.org

- Open source
- Free
- Easy installation
- Helpful community
- Regularly maintained and updated
- Tons of documentation & tutorials
- Every package comes with example vignettes to walk you through standard tasks.

When a laptop or desktop is just not enough...

Guillimin



Briaree

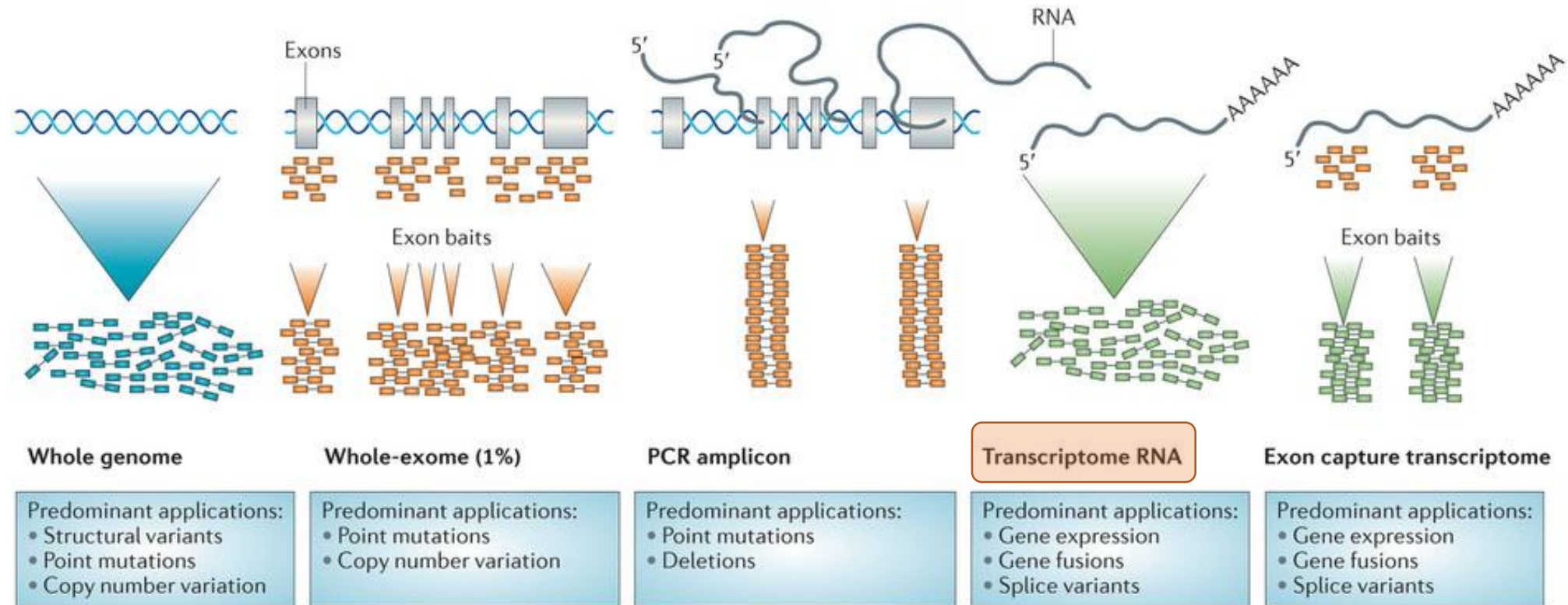


<http://www.calculquebec.ca/en/resources/compute-servers>

Working with NGS expression data

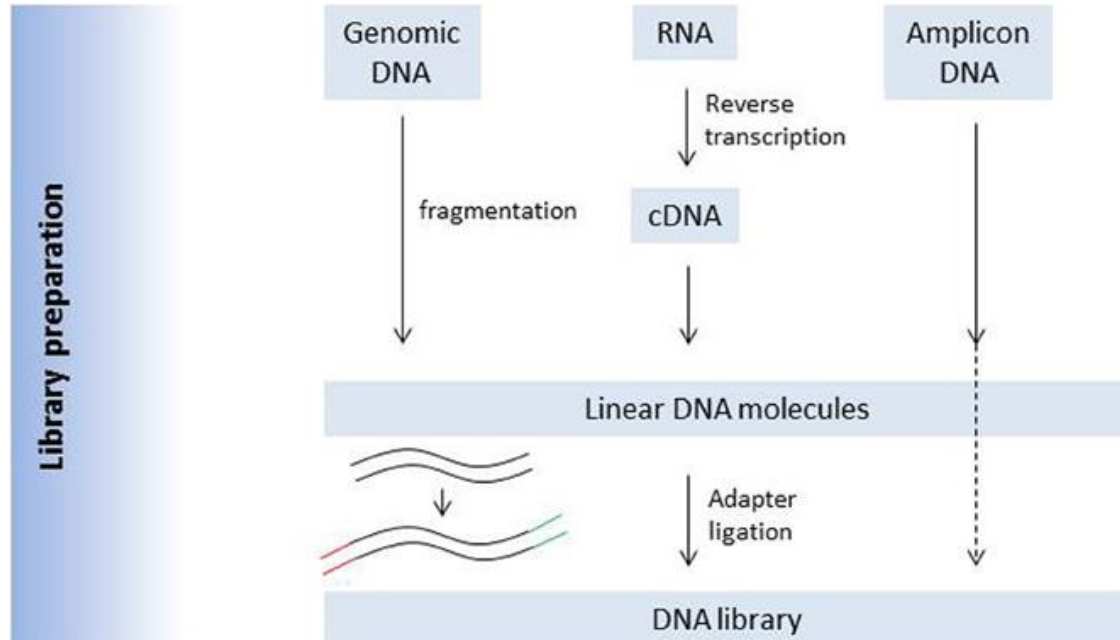
EXPERIMENTAL DESIGN & DATA ANALYSIS

Next Generation Sequencing(NGS)

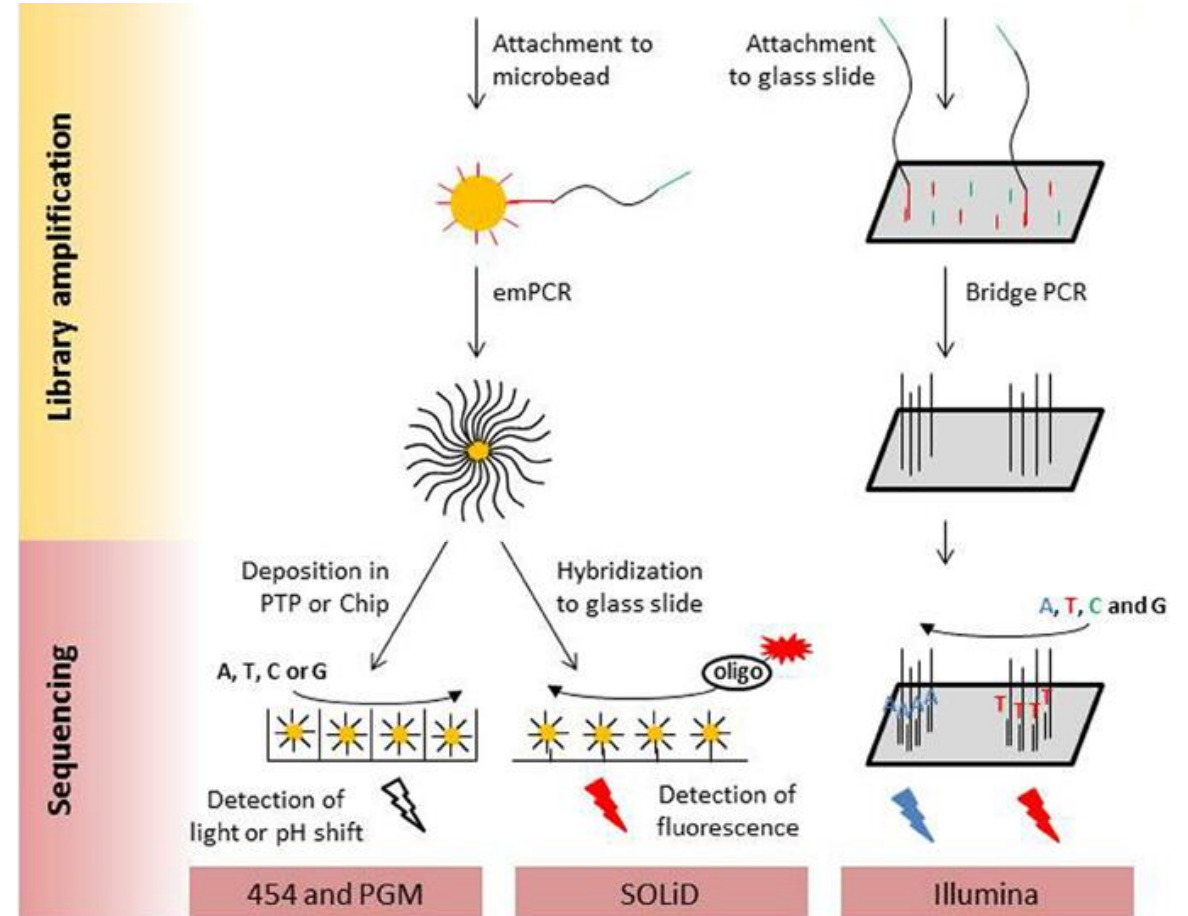


NGS Technology Workflow

1)



2)



End result of an NGS experiments visualized in IGV



<https://www.biostarhandbook.com/unit/rnaseq/stranded/img/rnaseq-dutp-paired.png>

Advantages of RNA-seq over microarrays

- Microarrays limited to the probes on the chip
- Low background noise
- Large dynamic range
- High technical reproducibility
- Identify novel transcripts and splicing
- Quantify rare transcripts

RNA-seq experiment

STEPS:

1. Experimental design
2. Quality Control
3. Read alignment
4. Assigning reads to genes or transcripts
5. Estimating gene or transcript abundance

Experimental Design Considerations

Minimize technical variation due to differences in

- Quality & quantity of RNA recovered during sample preparation
- **Library preparation** batch effects
 - largest but low compared to biological variation
- Flow cell and lane effects with Illumina technology

Recommendations:

- RIN > 9
- Randomize samples during preparation
- Dilute RNA to the same concentration
- Index & multiplex samples with all samples included on all lanes/flow cell or a blocking design => include some samples from each group on each lane

Importance of replicates over pooling samples

- Biological variance estimated from replicates
- Replicates add power to detect subtle changes statistically for events with LOW biological variance

For pooled samples:

- Genes with high variance in expression may appear differentially expressed (problem for lowly expressed genes)

PCR Duplicates

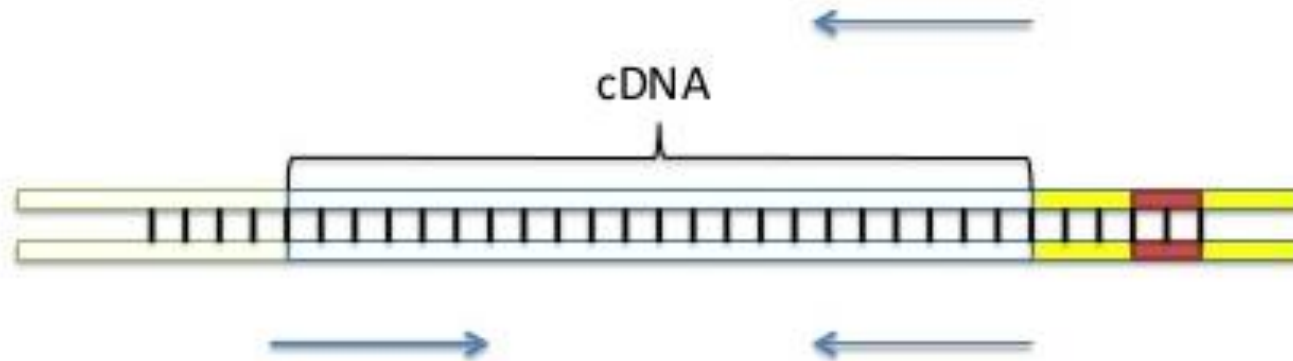
- PCR bias causes more amplification of some fragments and less of others
depends on sequencing depth, library complexity (i.e. number of transcripts expressed), expression levels => some fragments can completely overlap by chance
- PCR duplicates removed for mutation detection in WGS, Exome sequencing
- PCR duplicates should not be removed when estimating transcript abundance => underestimates abundant transcripts

WHY?

Cannot distinguish PCR duplicates from fragments that overlap in highly expressed genes

Single vs Paired End Sequencing

Single Read (SR) : only one end from each cDNA fragment is sequenced to generate one read per fragment *miRs



Paired End (PE) : the cDNA fragment is sequenced from both ends to generate two reads per fragment from two directions

**Better for splice variants, indels, inversions

**Needed for gene fusions, rearrangements,

With 50bp SE reads 95% of reads map uniquely

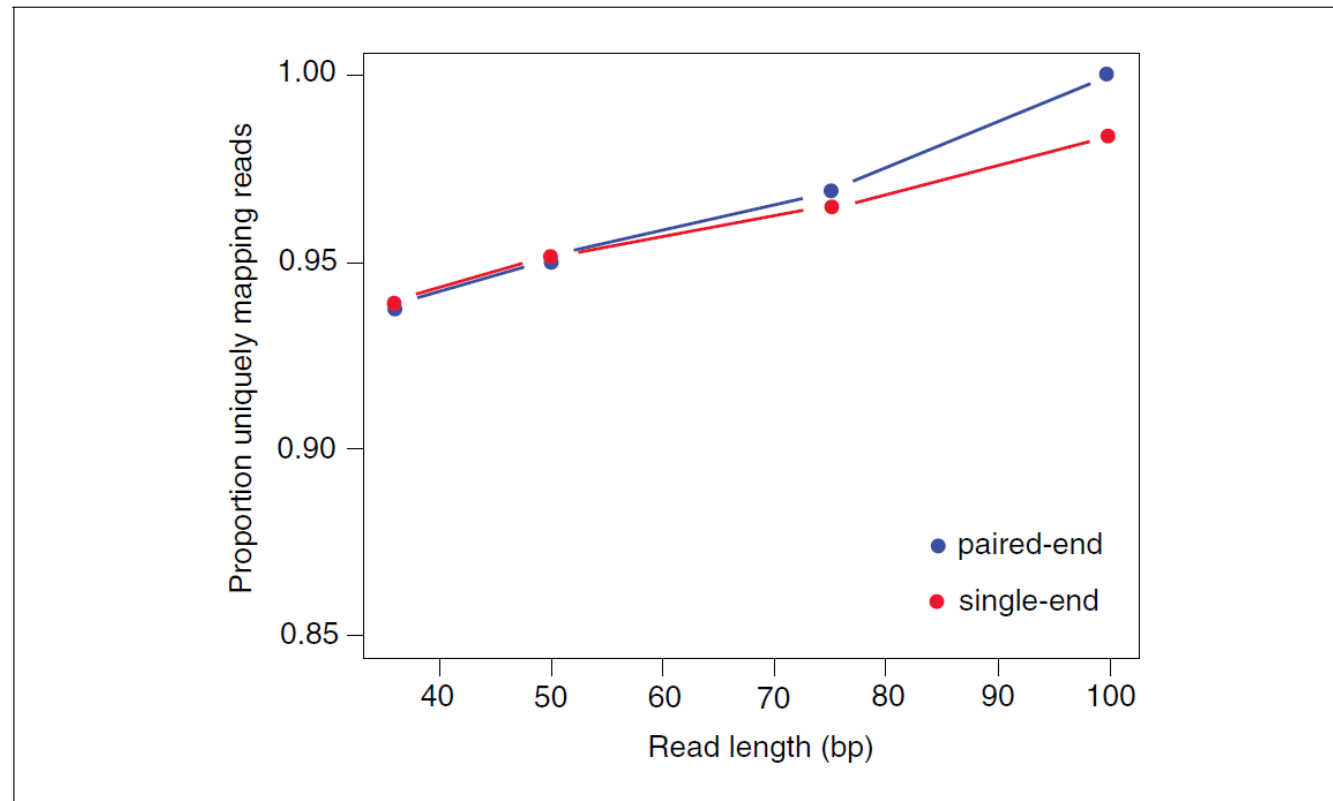
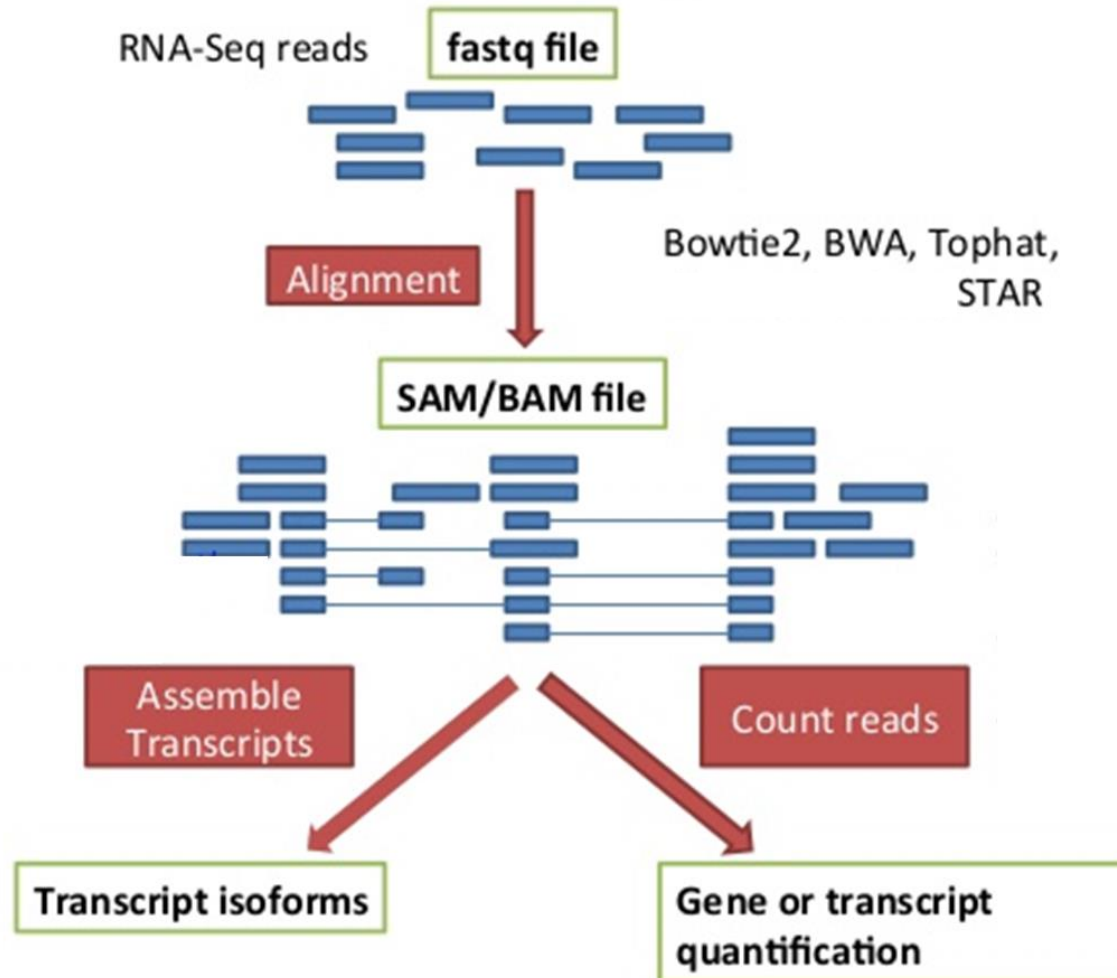
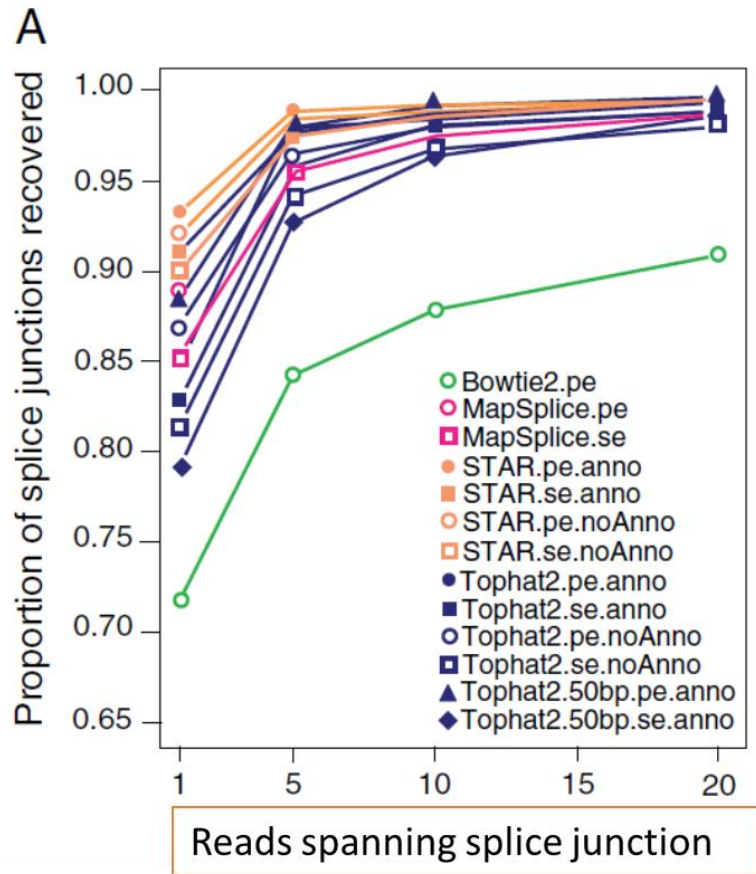


Figure 11.13.2 Proportion of reads mapping uniquely from PE and SE alignments relative to 100-bp PE results using Tophat2.

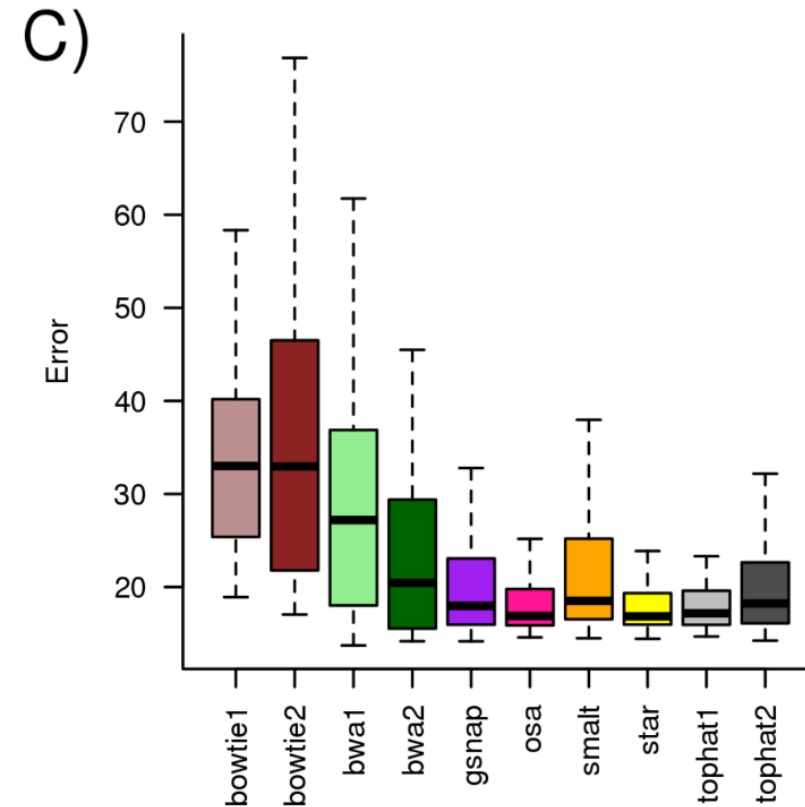
RNA-seq analysis workflow



Aligning reads to the genome

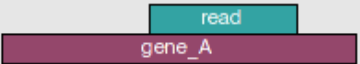
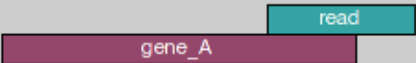


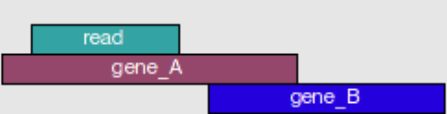



Williams et al., 2014 *Current Protocols in Human Genetics*

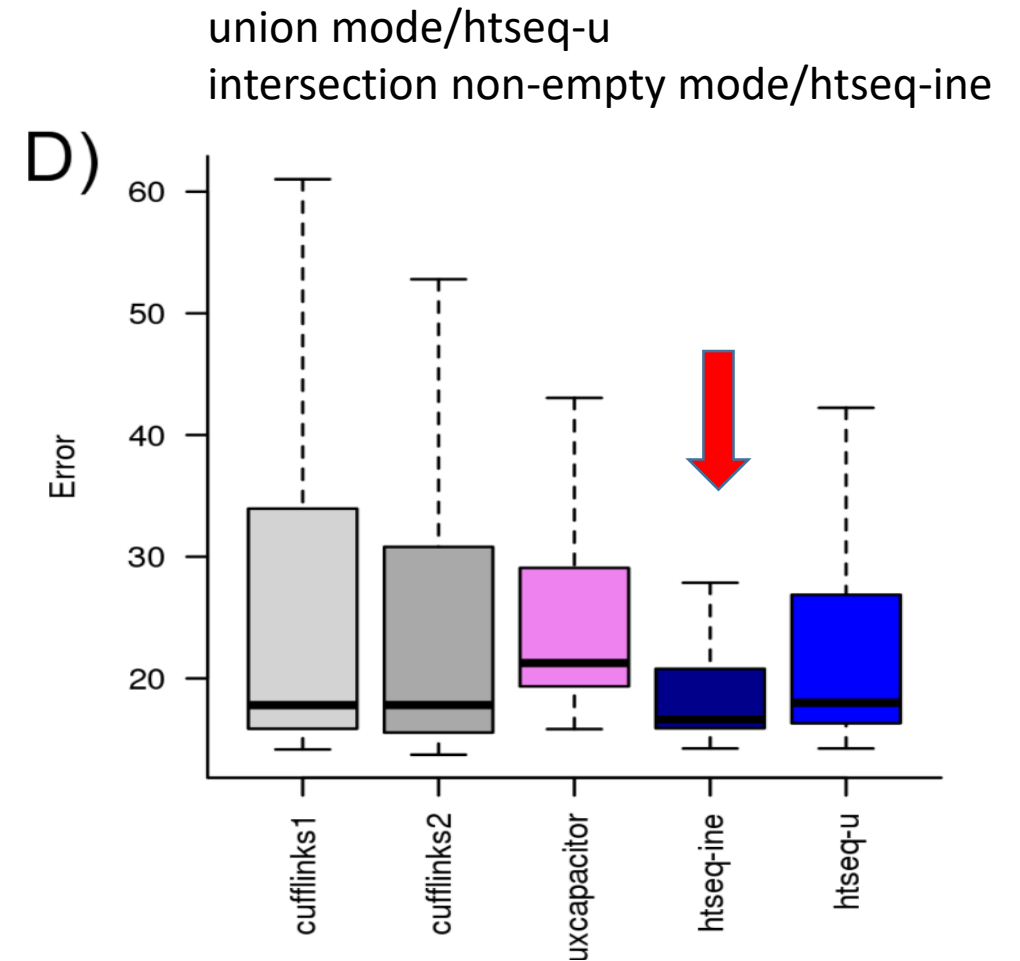


Fonseca NA, et al. 2014 *PLoS One*

Quantifying reads

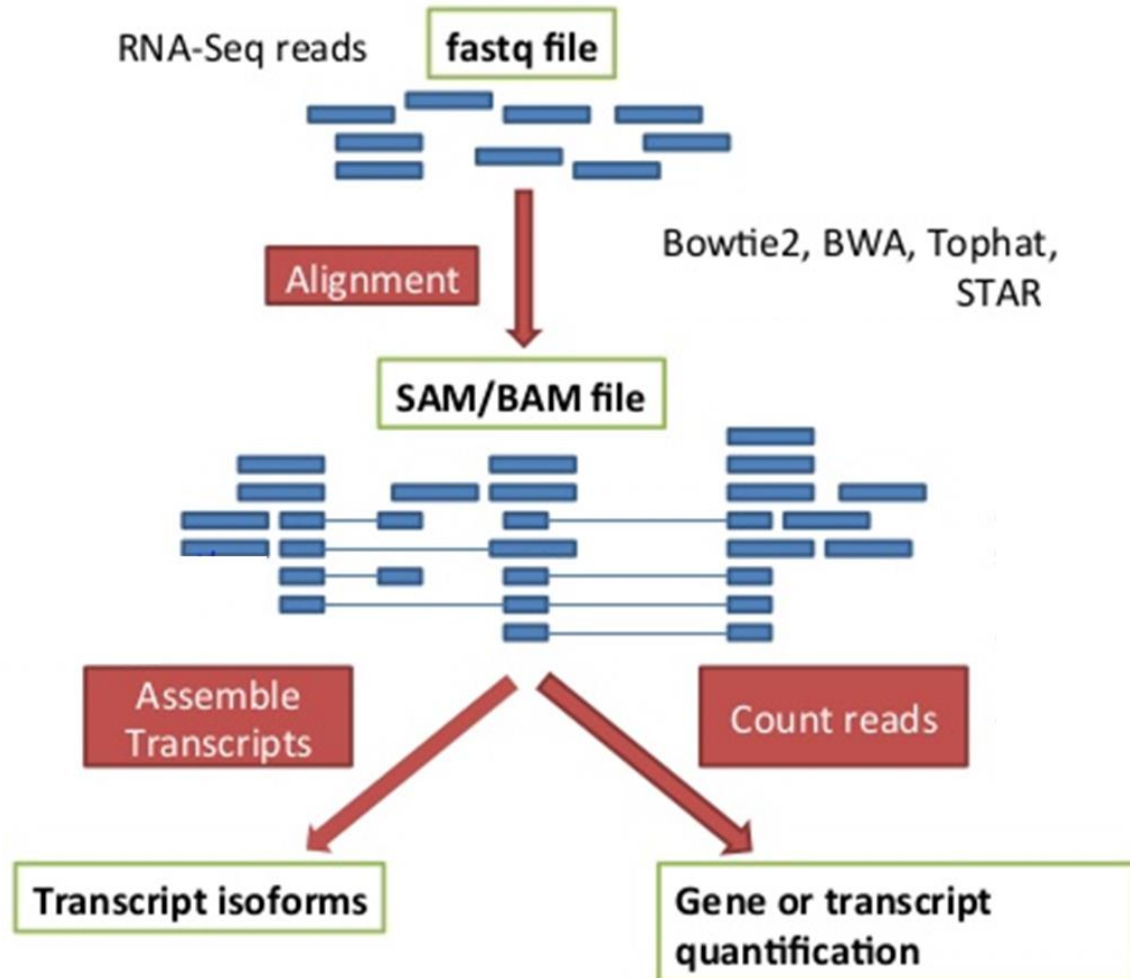
	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

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

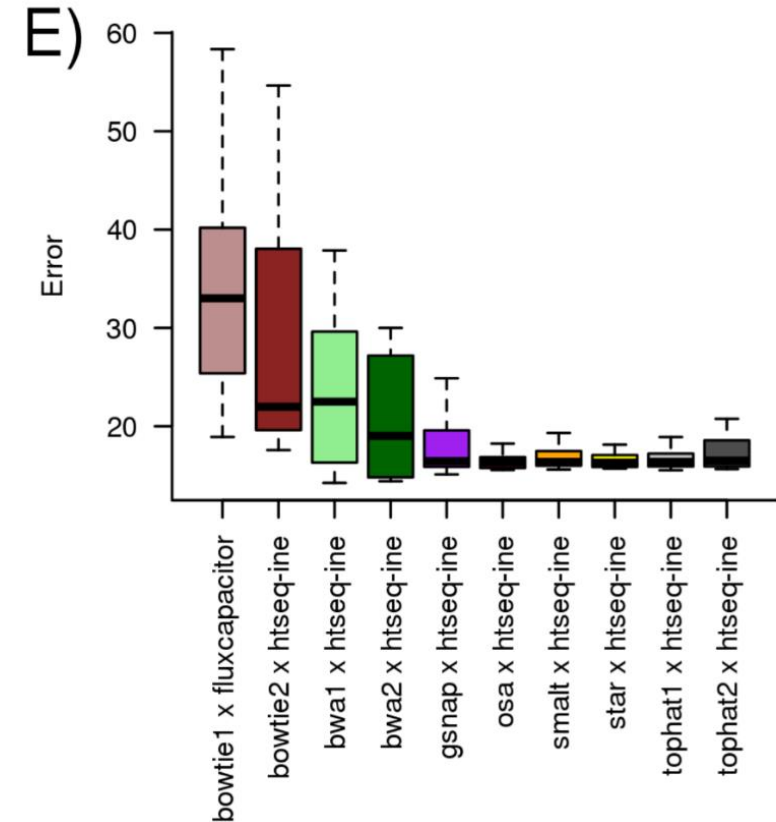


Fonseca NA, et al. 2014 *PLoS One*

Recommended workflow

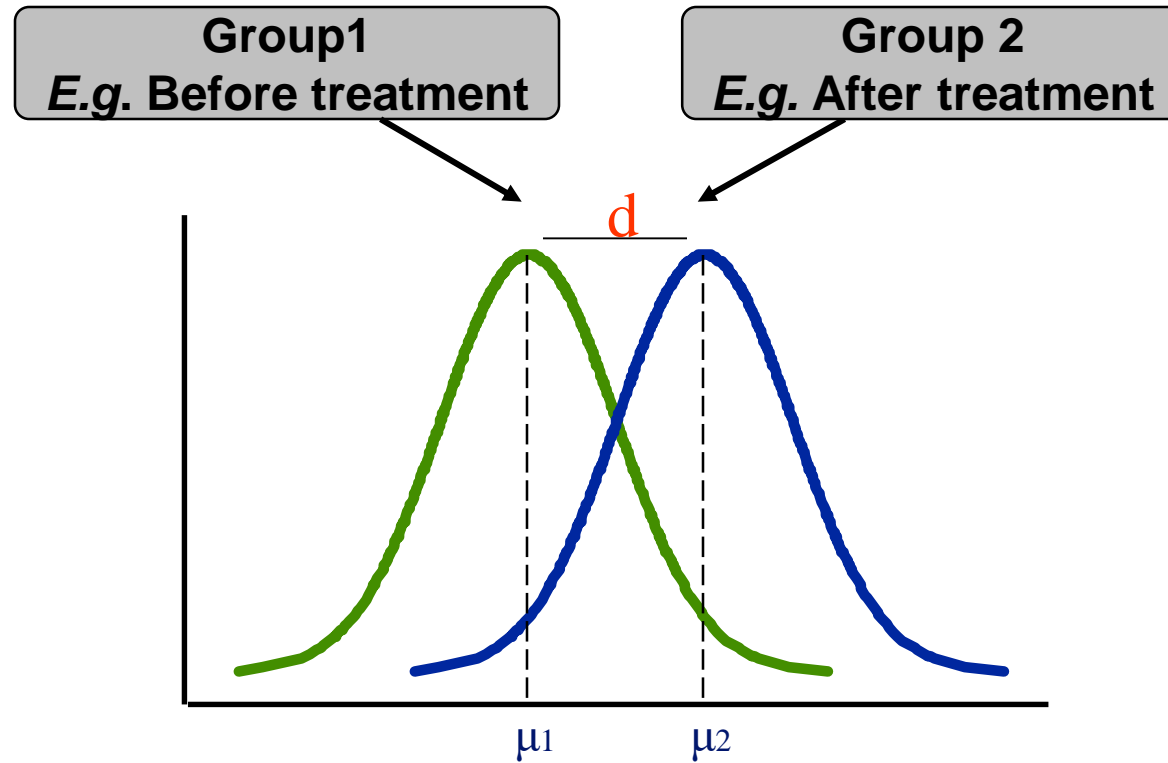


STAR with --quantmode or HTSeq



Fonseca NA, et al. 2014 *PLoS One*

Hypothesis Testing for Differentially Expressed Genes



Null hypothesis $H_0 : \mu_1 = \mu_2$

Alternative hypotheses $H_1 : \mu_1 \neq \mu_2$

Testing for differential expression (DE)

Factors affecting our ability to detect DE i.e. model biological variance:

- Sequencing depth
- Number of replicates

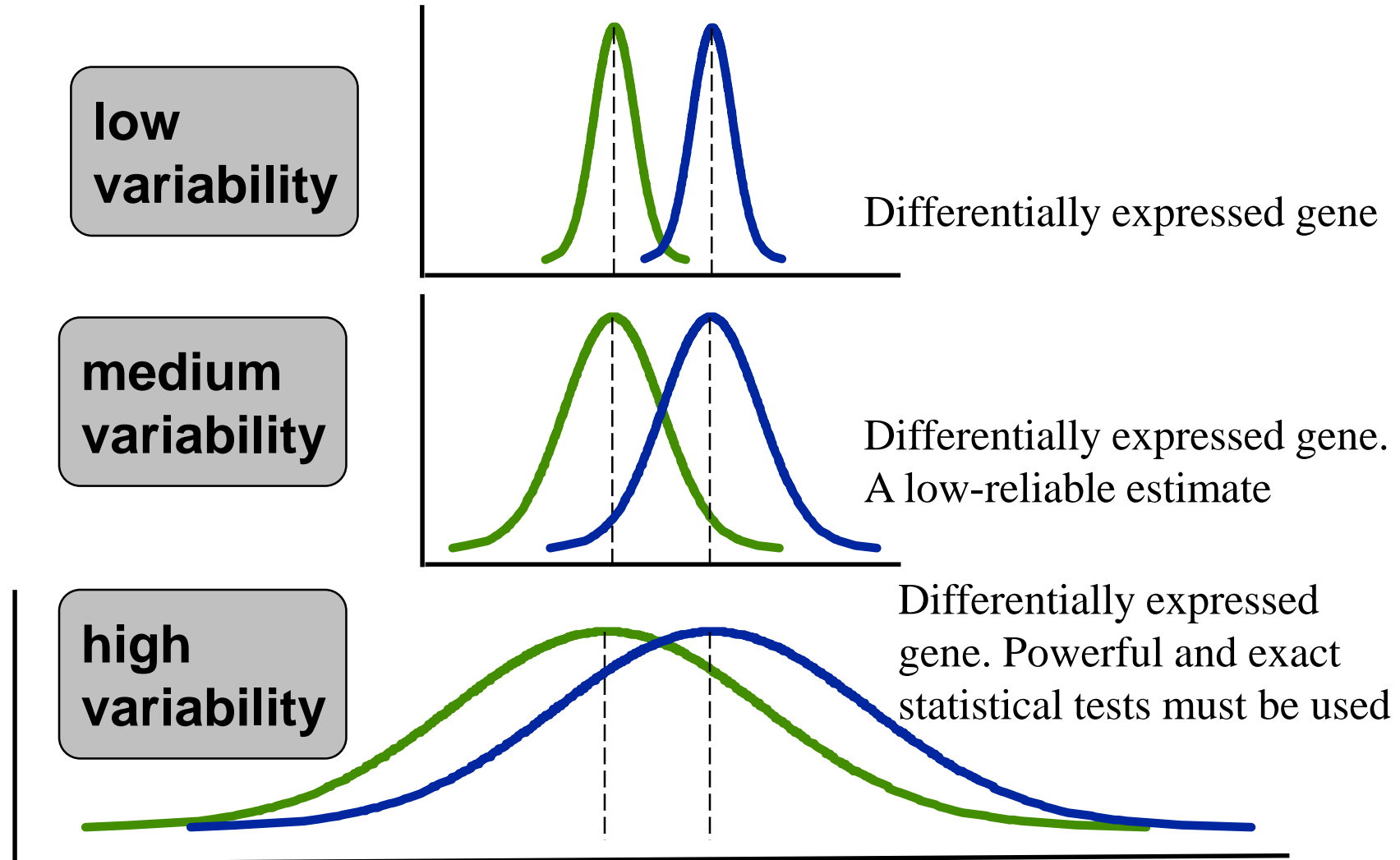
Other factors:

- Lowly expressed transcripts always difficult to measure accurately
- Samples used

Cell lines, inbred strains => minimal biological variability

Human samples collected at different time points or conditions => significant biological, environmental, and technical source of variation

Modeling Biological Variation



Other sources of variation

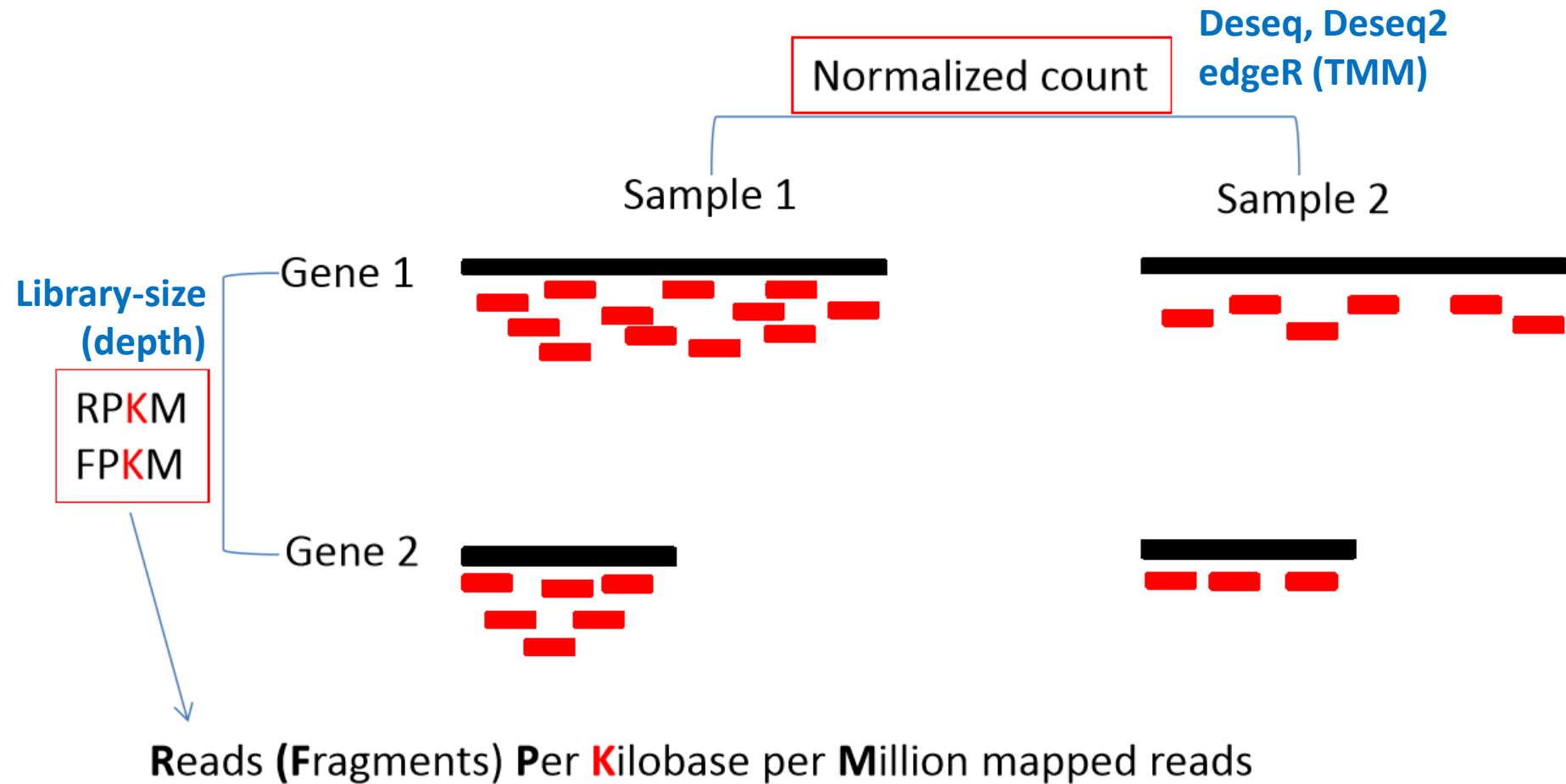
- Total RNA output per sample
- Variation in sequence composition (GC content)
- Fragment size selection
- Sequencing depth
- RNA composition

Normalisation seeks to correct for these biases



Only then can we reliably begin to draw inferences about differential expression

RNA-seq normalization



Library-size normalization assumes the underlying population of mRNA is similar between samples.

Normalization Methods

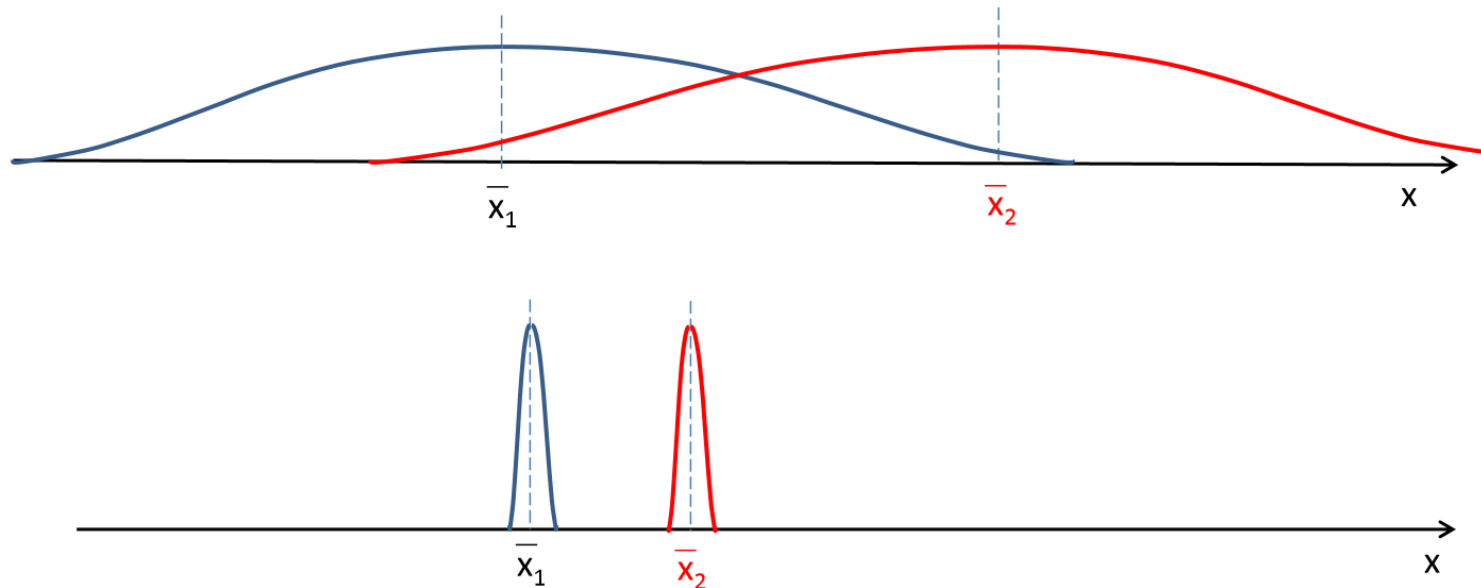
- **RPKM**: This approach quantifies gene expression from RNA-Seq data by normalizing for the total transcript length and the number of sequencing reads.

Highly expressed genes that are not highly expressed in all samples can bias library-size normalization.

- **DESeq/DESeq2**: DESeq is a differential gene expression analysis method based on a negative binomial distribution model. Uses a virtual reference sample to compare each sample to obtain a scaling factor.
- **TMM/edgeR**: The trimmed mean of M-values is a scaling normalization method. The scaling factor is calculated for library sizes that minimize the log-fold change between samples, and then rescaled gene counts are used for downstream analysis.

Biological replicates exhibit higher levels of variance than accounted for by the Poisson distribution

Dispersion affects ability to differentiate



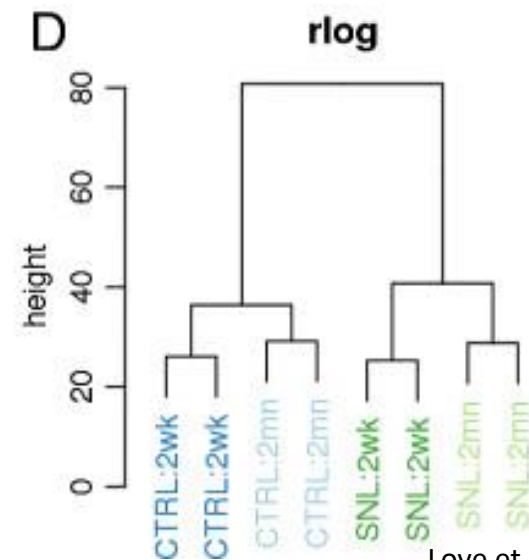
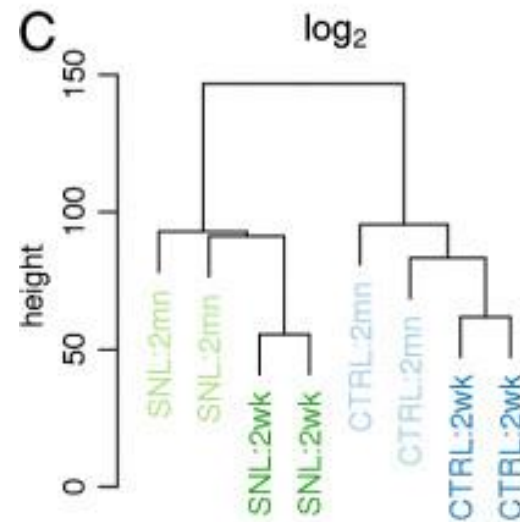
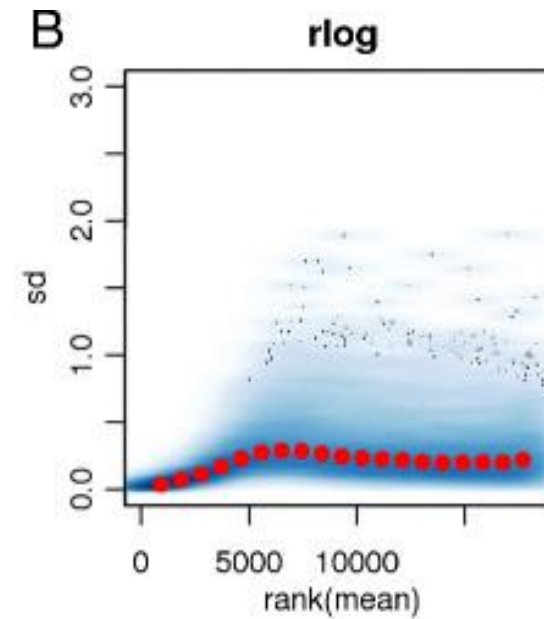
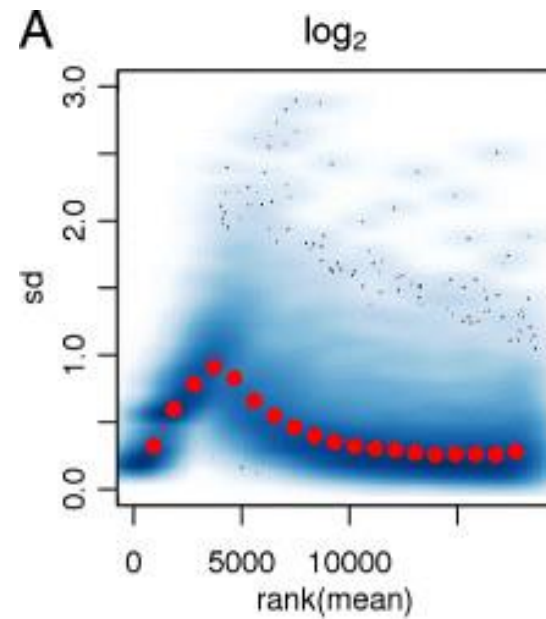
Poisson distributions assumes the mean and variance is equal => overdispersion

Negative binomial distribution => allows you to include a variance parameter estimated from data

DESeq/DESeq2
gene-specific dispersion parameter
for each gene with Wald test for p values

DESeq2 implements rlog transformation to stabilize variance for downstream analysis

rlog stabilizes the variance through the range of the mean of counts and helps to find meaningful patterns in the data



Summary

- Importance of bioinformatics for analysing and integrating functional genomics data
- Overview of NGS experimental design and considerations when analyzing the data

What will you learn in the next workshops?

1. Basics of working in a terminal
2. How to run programs in command line
3. How to write bash scripts to run programs in linux/cluster
4. From BAM to fastq.gz
5. Align fastq.gz files to the genome with STAR
6. Quantify genes with STAR --quantmode and HTSeq
7. Go from raw counts to normalized expression values with DESeq2

What will you learn in the next workshops?

- 8. Basics of R programming
- 9. Quality control steps i.e. sample distance clustering, PCA plots
- 10. Look at differentially expressed genes using models (paired & unpaired)
- 11. Basics of data visualization i.e. heatmaps, MA plots, boxplots, etc
- 12. Basic pathway analysis with the clusterprofiler package

Quick Links

[Access to Resources](#)[Calcul Québec Portal](#)[Events](#)[FAQs](#)[News](#)[Our Twitter feeds](#)[Server Status](#)[Wiki - Documentation](#)

Subscribe
to our
newsletter

Access to Resources

Any researcher eligible for funding from a National Research Council (NSERC, SSHRC, MRC) can become a member of Calcul Québec and Calcul Canada and use our compute resources.

Typically, regular professors are eligible while postdoctoral researchers or graduated students are not. Researchers with the status of associate professor can use compute resources if the project for which the resources are used is eligible for funding from CFI or other funding councils. Researchers from the CEGEP network who are eligible for funding from a National Research Council can also become members.

[Calcul Québec is also open to collaborations with companies from Québec and Canada.](#) Small and medium-sized businesses can take advantage of the expertise of our technical team to support them in their projects. For more information, contact Suzanne Talon, Calcul Québec Executive Director (514 343-6111, Ext 5502).

Accessing Resources as a Researcher

- The Principal Investigator (PI) must first register on the Compute Canada database (CCDB). Once the registration is confirmed, the PI will receive a CCRI (Compute Canada Role Identifier) code in the format "abc-123-01" and will become the sponsor of his future users group.
- The regular (sponsored) user then registers using the same form. This operation requires the sponsor's CCRI.
- Users (PI or sponsored) can then request an account with a regional partner or consortium using the CCDB website.

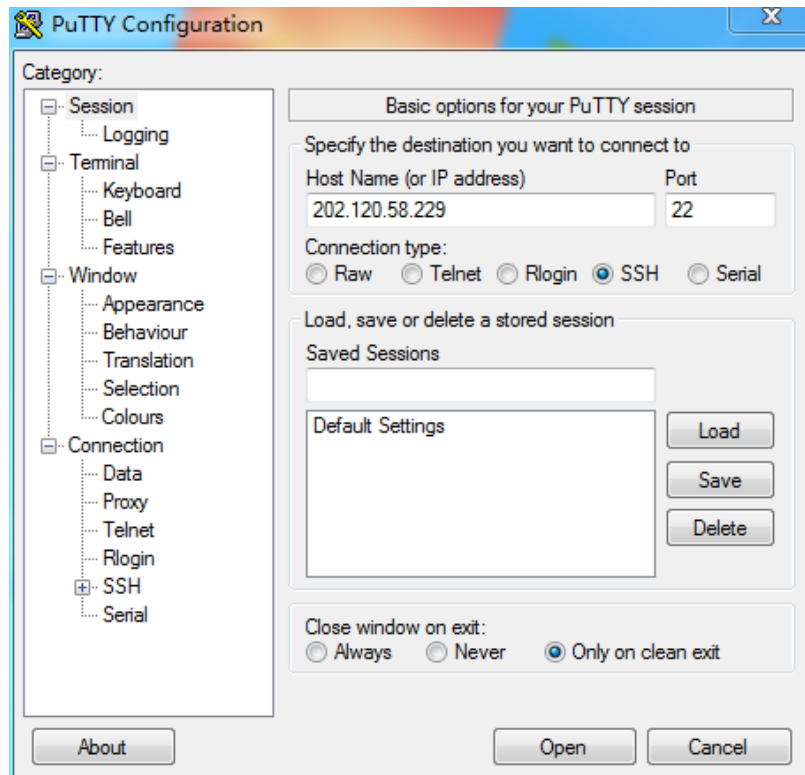
To complete these steps, go to the CCDB website and click "Register".

Accounts on Calcul Québec Servers


- Once your Calcul Québec account is created, you may request access to various servers using the Calcul Québec portal, under the "My Profile" tab.
- Note that there is a delay between the request for an account and its creation. Please wait one full business day before contacting us to indicate a problem.
- For detailed information on the Calcul Québec servers, see the "[Table summarizing properties of Calcul Québec servers](#)" on the Calcul Québec wiki.

How do we access these clusters?

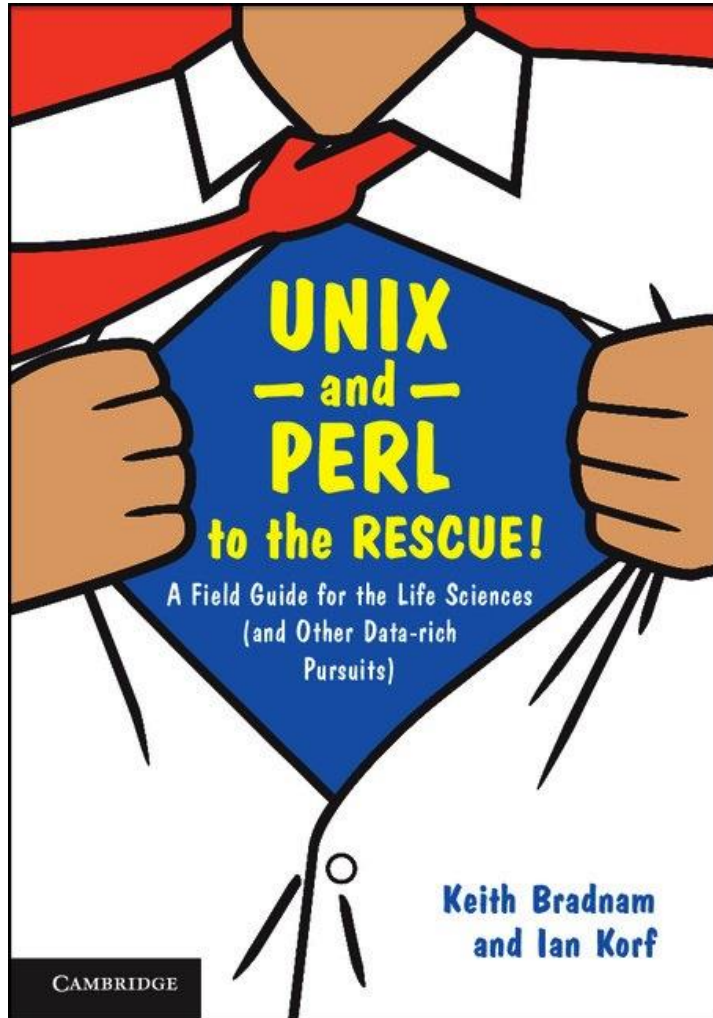
Windows



Mac Unix/Linux

```
web029235:~ radiamariejohnson$ !ssh  
ssh johnsonr@briaree.calculquebec.ca  
johnsonr@briaree.calculquebec.ca's password: 
```

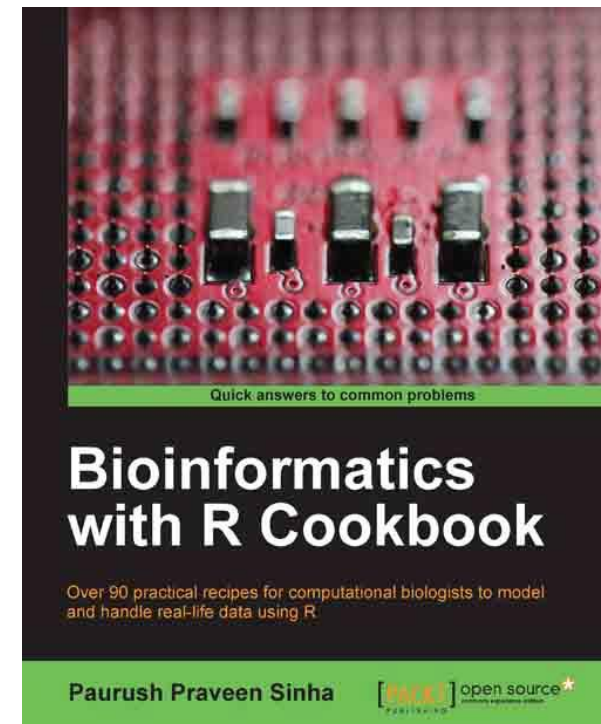
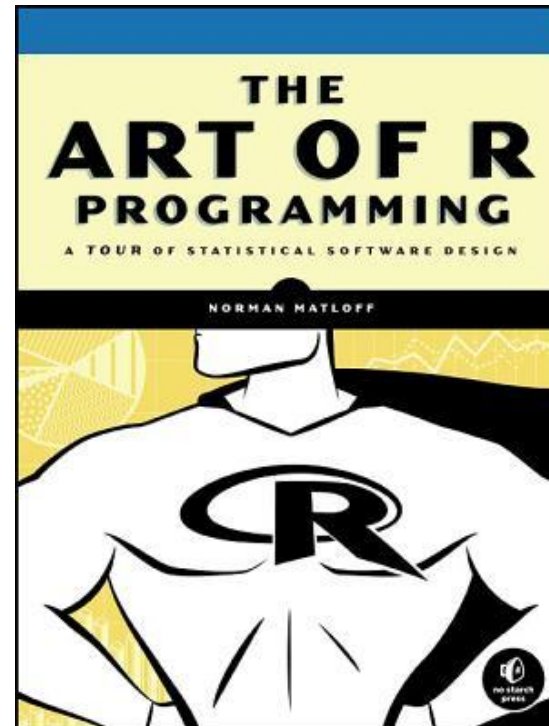
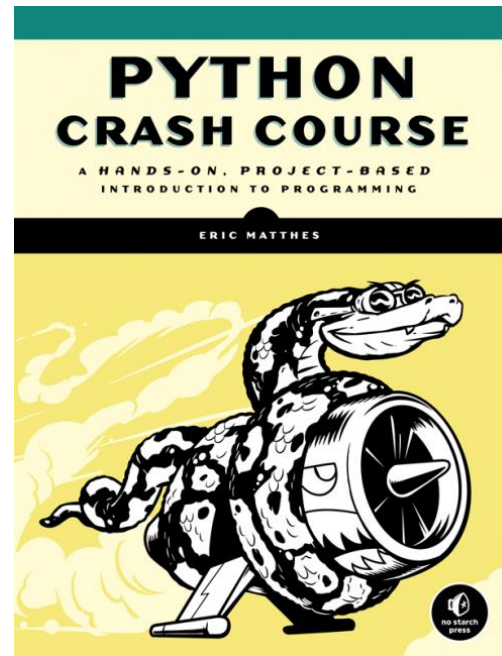
Working with a terminal & BASH shell scripting



```
jordan — bash — 96x26
Shard:~ jordan$ ls -l
total 0
drwx-----+ 10 jordan  staff   340 12 Jun 17:00 Desktop
drwx-----+ 13 jordan  staff   442 27 May 15:03 Documents
drwx-----+ 172 jordan  staff  5848 12 Jun 17:16 Downloads
drwx-----@ 27 jordan  staff   918 11 Jun 23:14 Dropbox
drwx-----@ 75 jordan  staff  2550 11 Jun 23:14 Library
drwx-----+  8 jordan  staff   272 17 Apr 17:20 Movies
drwx-----+  8 jordan  staff   272 12 Jun 10:56 Music
drwx-----+ 33 jordan  staff  1122  9 May 10:48 Pictures
drwxr-xr-x+  5 jordan  staff   170 23 Mar 12:17 Public
drwxr-xr-x   3 jordan  staff   102 11 Jun 17:03 Sites
Shard:~ jordan$
```


Resources

- <https://rjbioinformatics.com/gcrc-bioinformatics-workshops/>
- <http://onlinelibrary.wiley.com/doi/10.1002/0471142905.hg1113s83/pdf>





THANK YOU