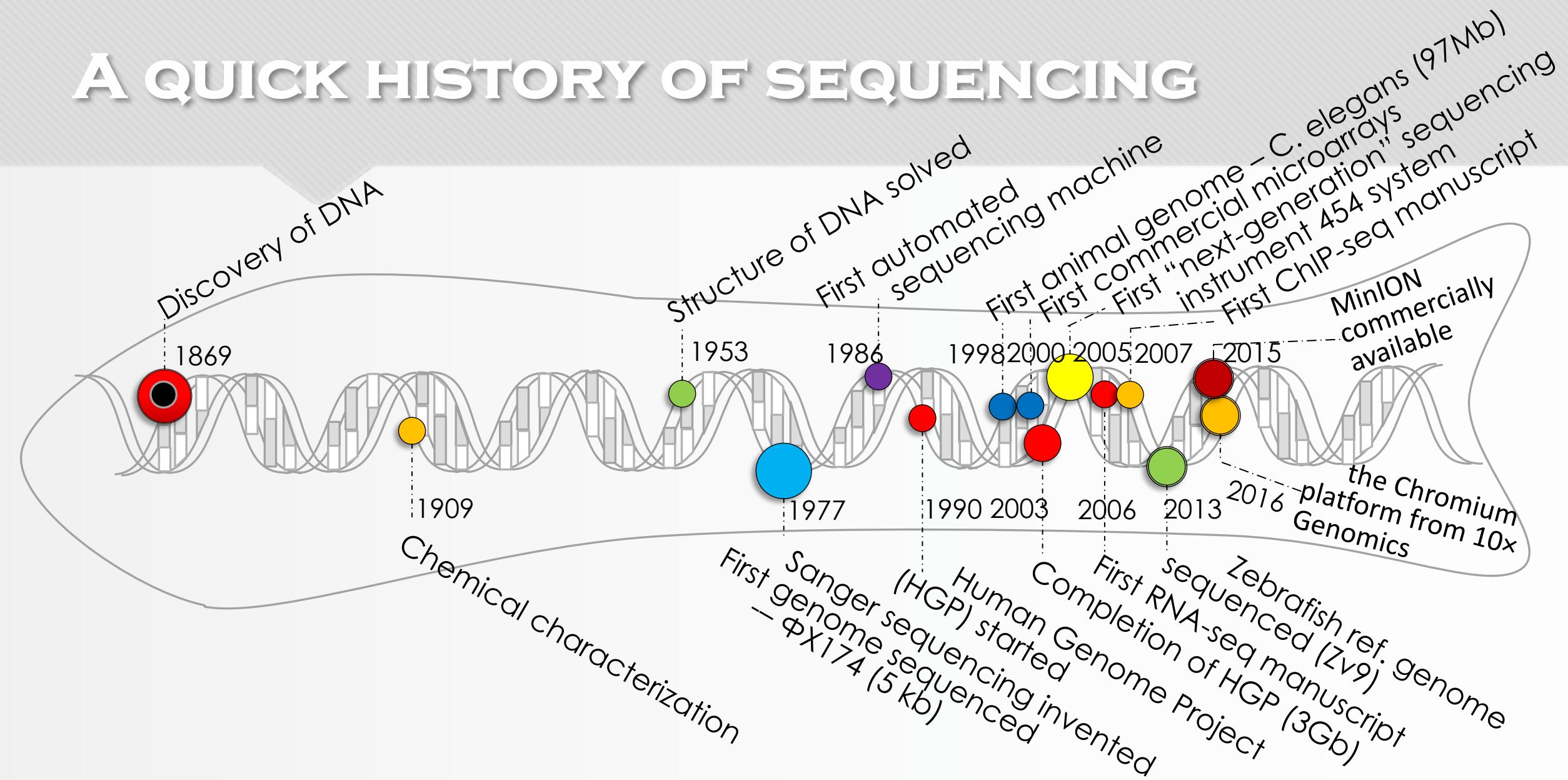


# **RNA-SEQ AND CHIP-SEQ ANALYSIS PRIMER**

Jianhong Ou @ DUKE, 2022

# A QUICK HISTORY OF SEQUENCING



# 22 Years of the Human Genome



**June  
2000: International  
Human Genome  
Sequencing  
Consortium  
Announces "Working  
Draft" of Human  
Genome**

# **3 MAIN STEPS FOR RNA-SEQ AND CHIP-SEQ**

01

## **Experimental design**

Library type / strategy  
Sequencing depth  
Number of replicates  
Batch effect

**Consult a data analyst or bioinformatician to make sure robust experimental design**

02

## **Bench work**

RNA/DNA extraction  
Library preparation  
Sequencing & imaging

**Double check**  
1. Library prep batch  
2. Layout of library on lane/ flow cell

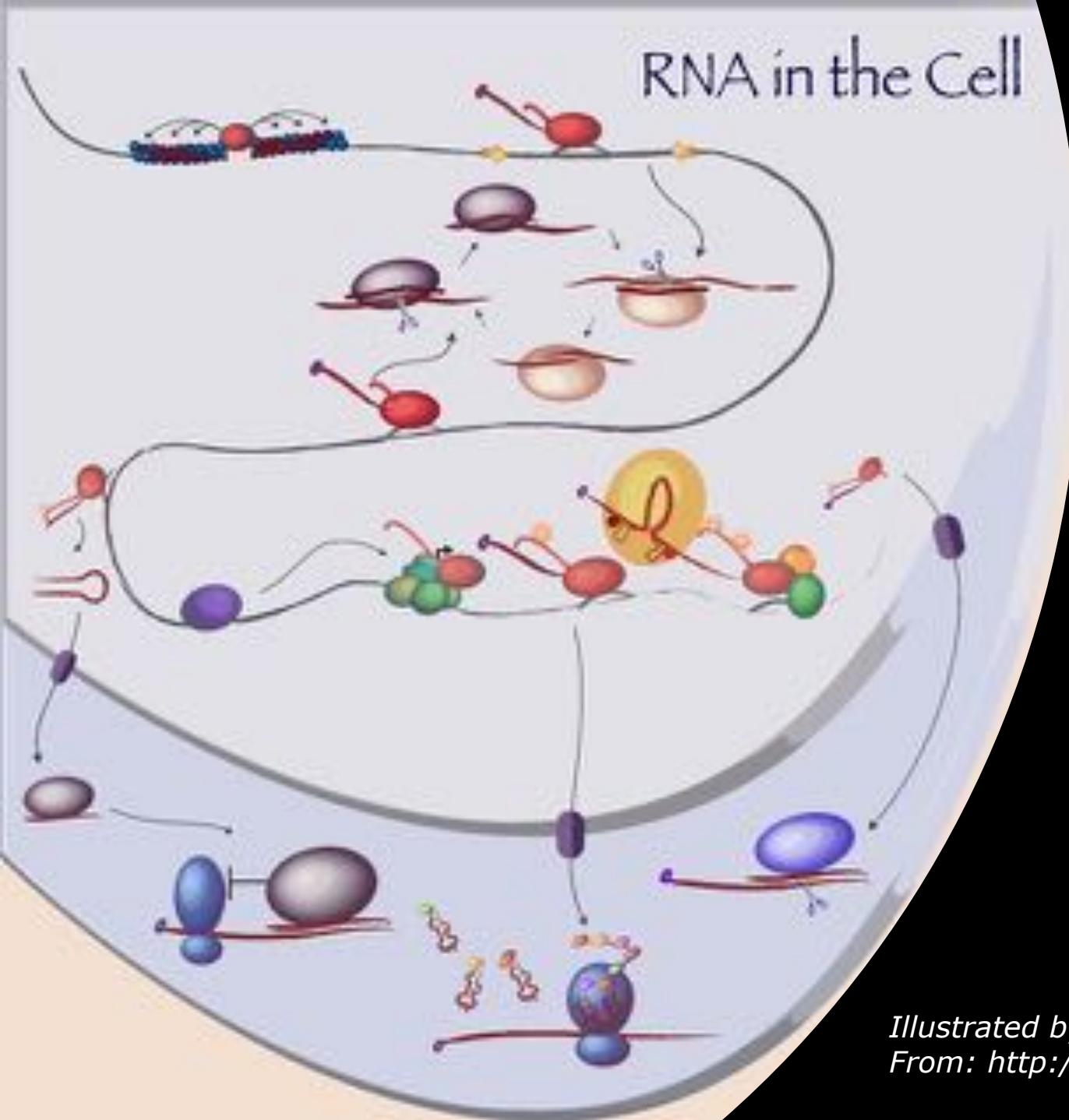
03

## **Data analysis**

Quality control  
Alignment & Quantification  
Differential Expression test  
Functional profiling

**Discuss and optimize the analysis to meet your experiment design**  
1. Blind analysis to avoid the bias  
2. Discuss in time to optimize the analysis

# RNA-SEQ



## THE TRANSCRIPTOME IS VERY COMPLEX

*Illustrated by Mary Lindstrom; designed by M.L. and Amy White  
From: <http://web.mit.edu/sharplab/researchsummary.html>*

## coding RNA

### mRNA

Most are polyadenylated

## Non-coding RNA

### Small RNAs

- siRNA
- miRNA
- snoRNA
- snRNA

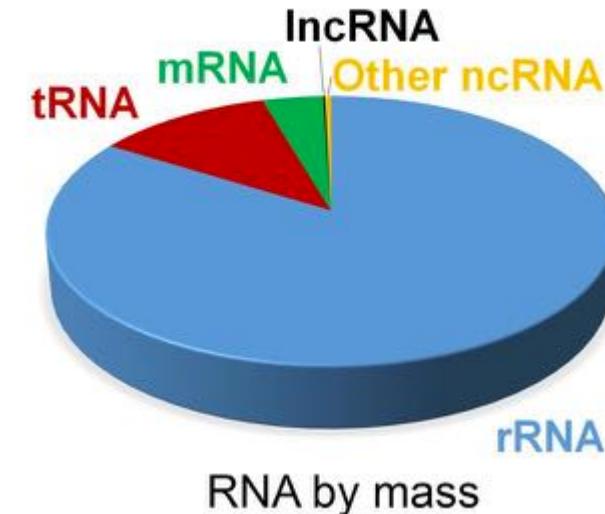
### Transcriptional RNAs

- rRNA
- tRNA

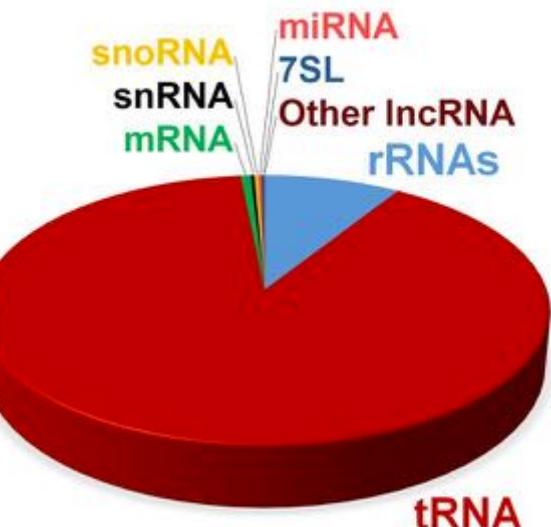
### lncRNAs

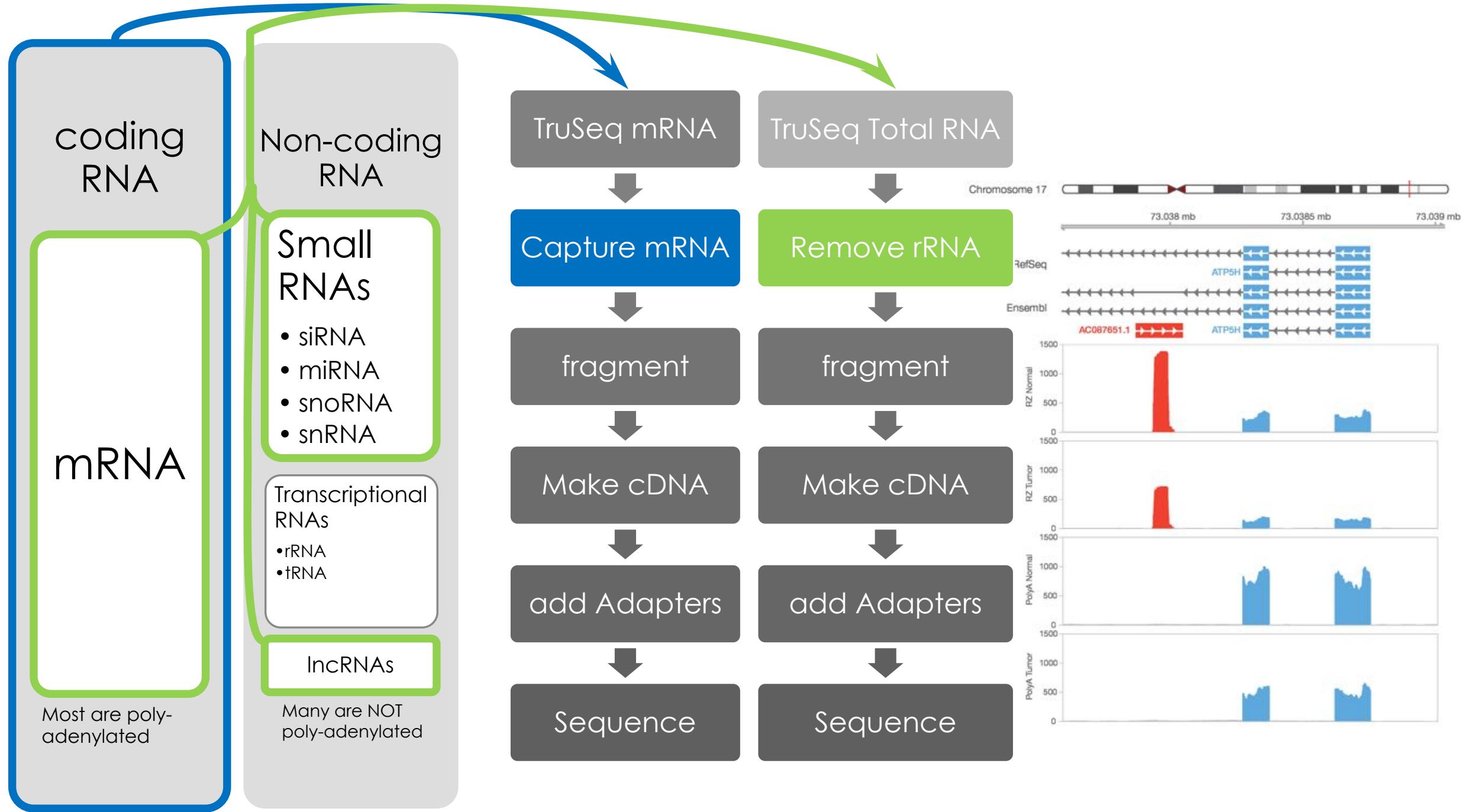
Many are NOT polyadenylated

A

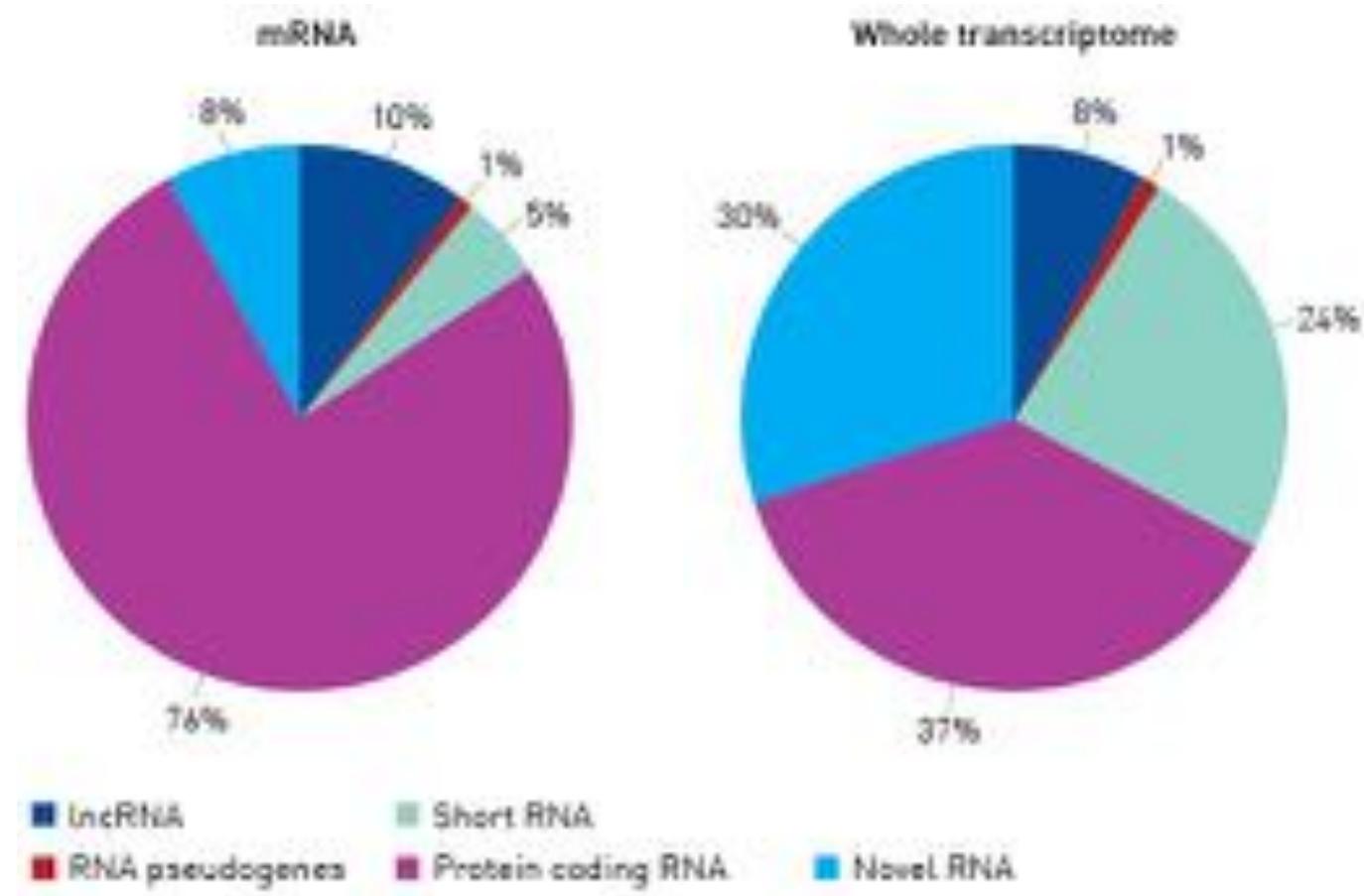


B

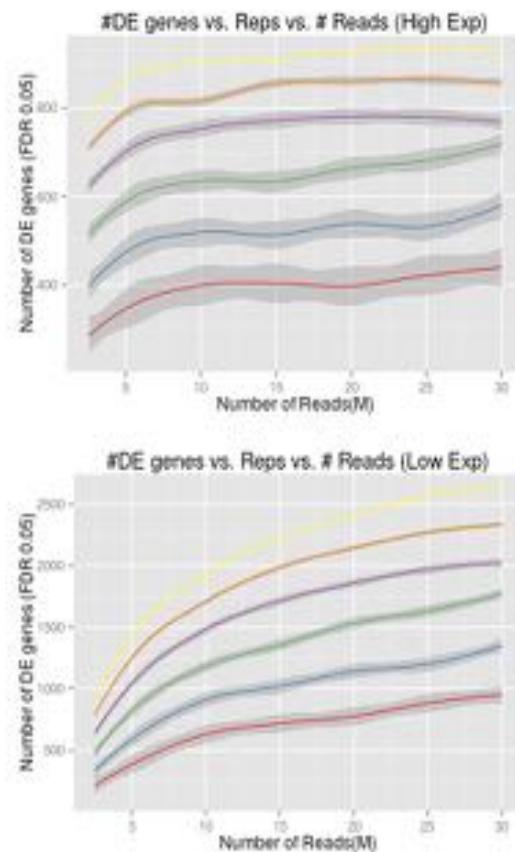




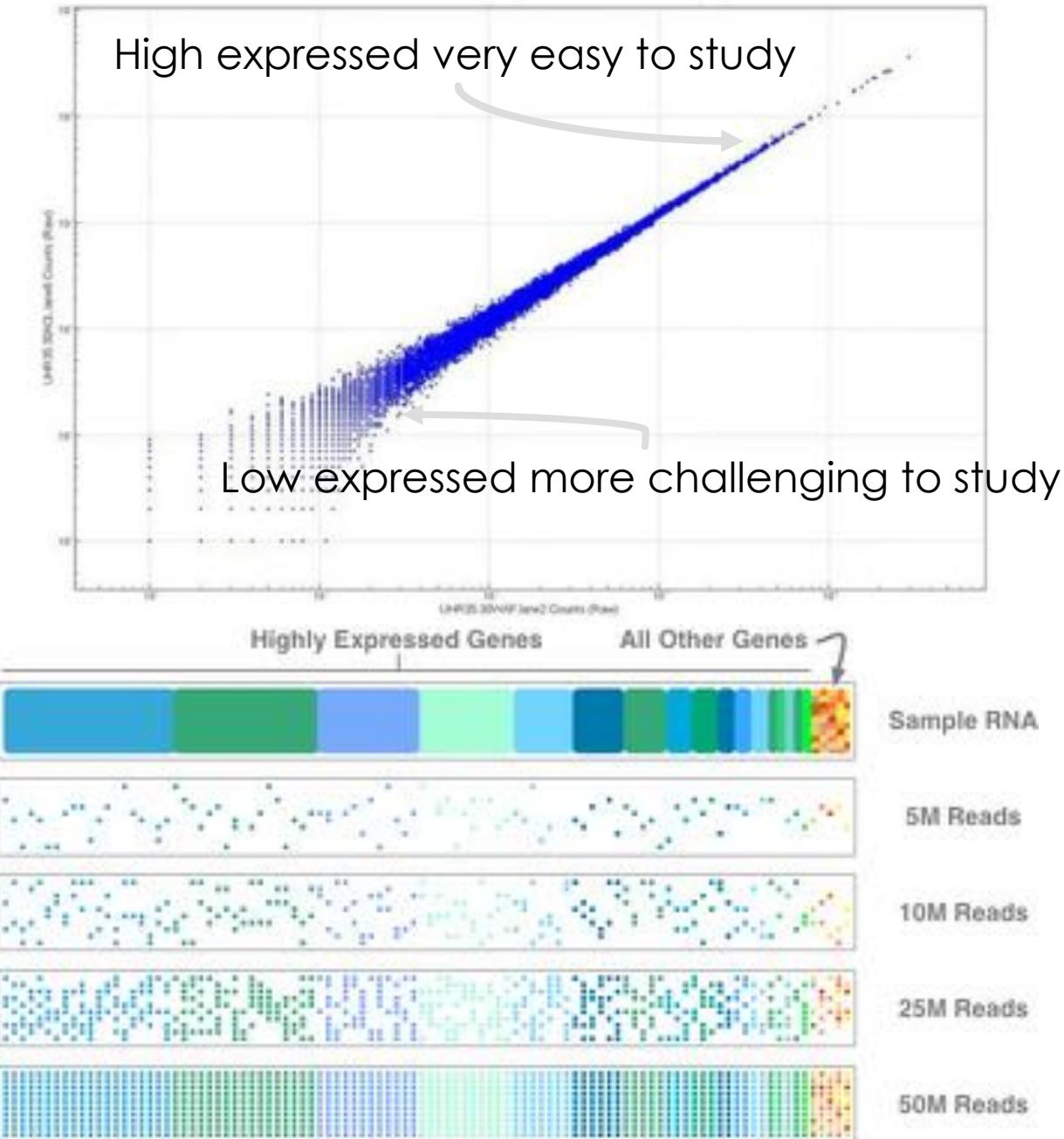
# Distribution of RNA species in NGS samples after poly-A enrichment (mRNA) and rRNA depletion (whole transcriptome) sequencing



# How many reads?

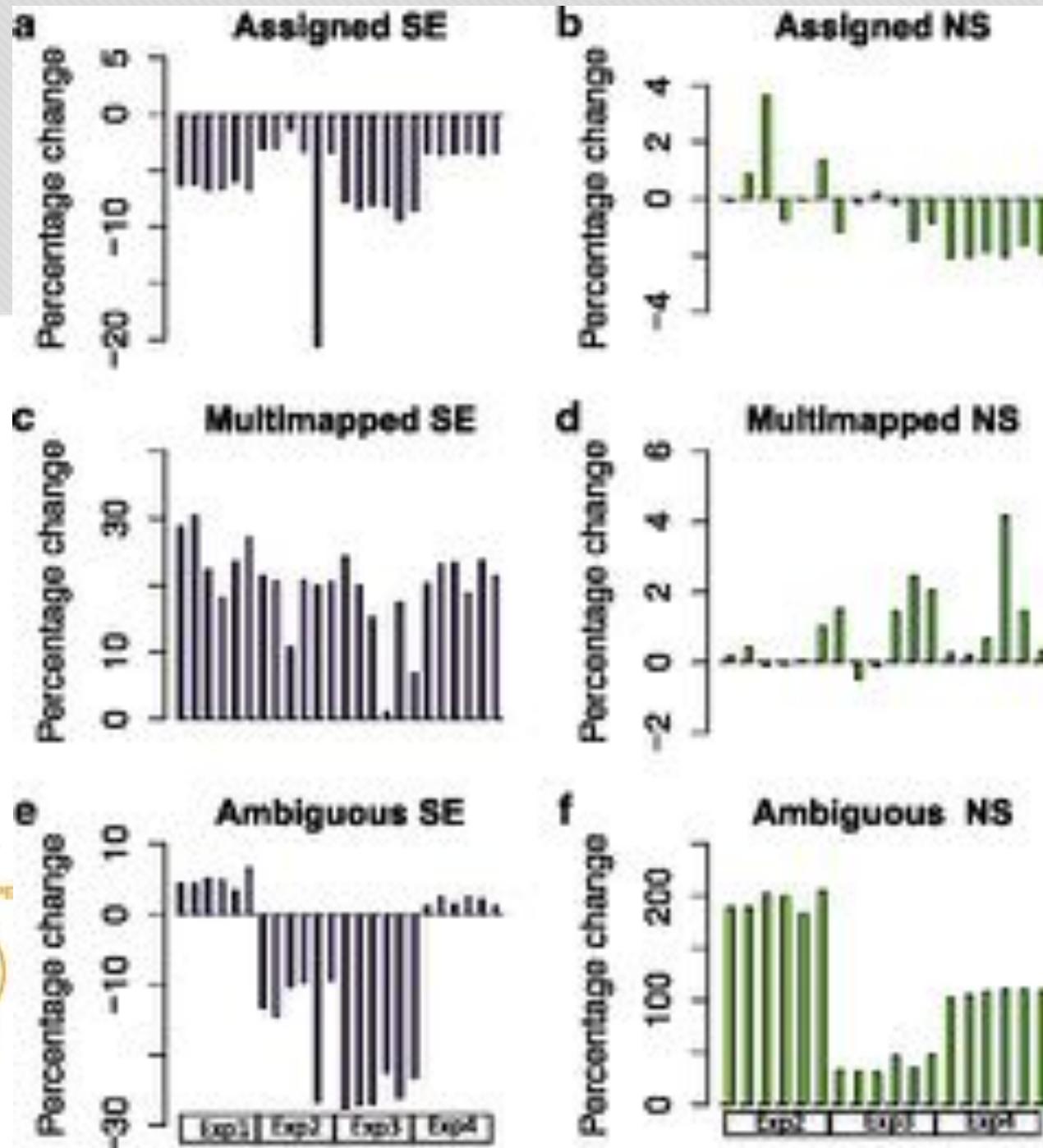
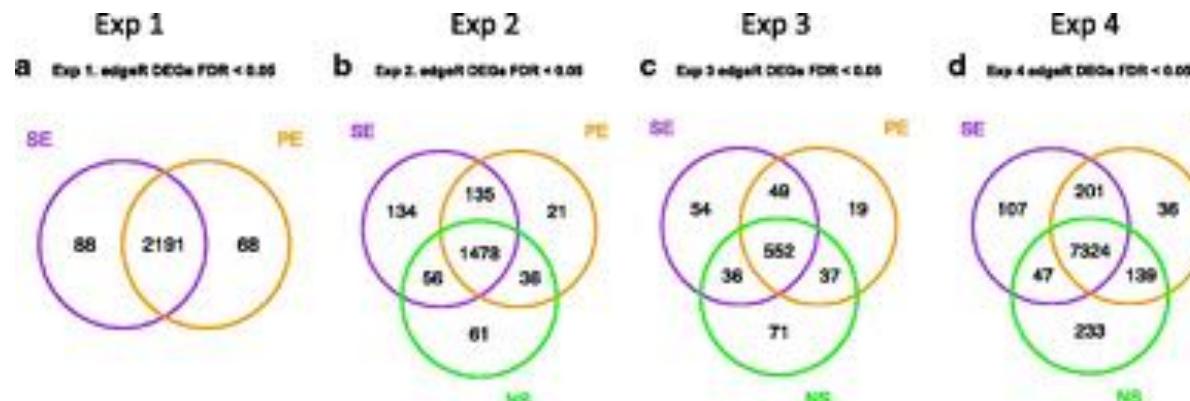


In most cases, increasing replicates is more beneficial than increasing sequencing depth.



# Single End vs Paired End

- Paired end (PE) sequencing improves mapping
  - to repeat sequences in the genome
  - across exon-exon junctions
- Single-end (SE) sequencing is cheaper than PE.



**Table 1:** Recommended sequencing depths for typical RNA-seq experiments for different genome sizes (Genohub, 2015). DGE = differential gene expression, SR = single read, PE = paired-end.

	Small (bacteria)	Intermediate (fruit fly, worm)	Large (mouse, hu- man)
No. of reads for DGE ( $\times 10^6$ )	5 SR	10 SR	20–50 SR
No. of reads for <i>de novo</i> transcriptome assembly ( $\times 10^6$ )	30–65 PE	70–130 PE	100–200 PE
Read length (bp)	50	50–100	>100

**For a basic RNA-seq differential expression experiment, 10M to 20M single end reads per sample is usually enough**

# VARIANCE AND BLOCKING

Good experimental designs mitigate experimental error and the impact of factors not under study.

Krzywinski et.al., 2014 doi: 10.1038/nmeth.3005

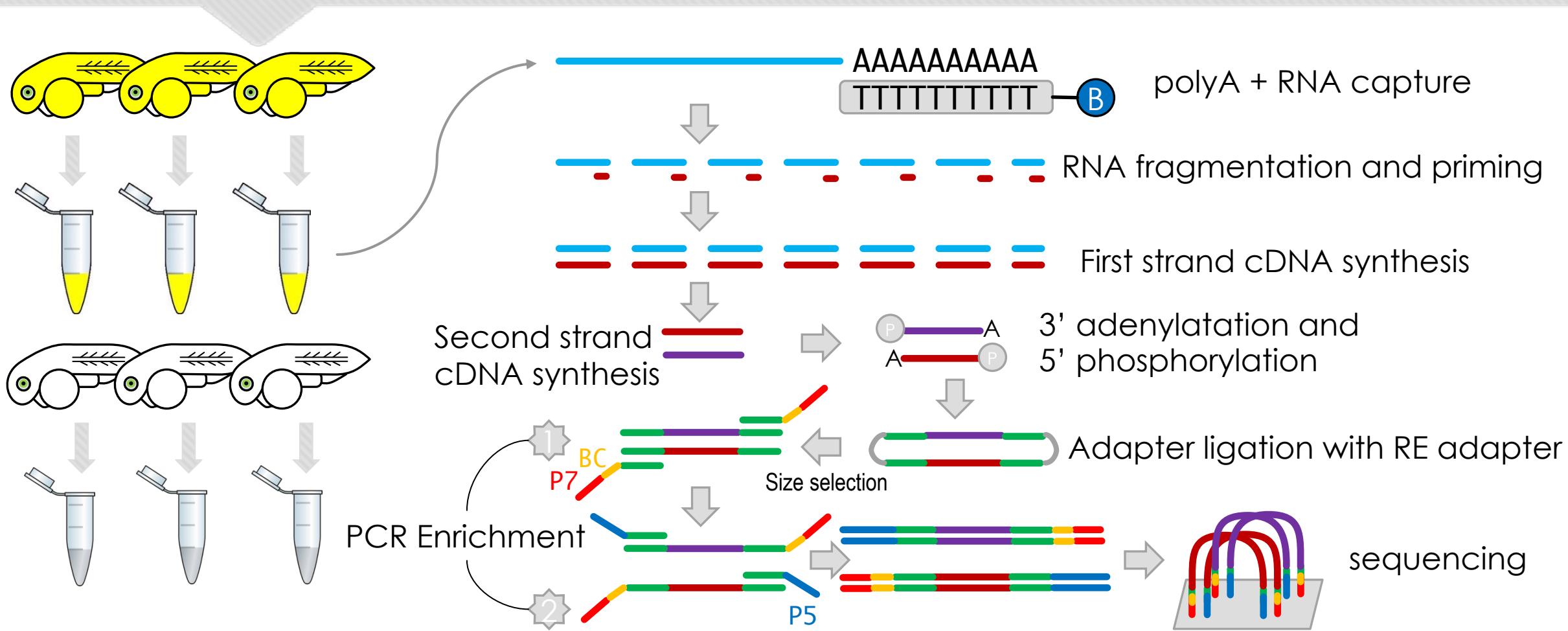
## DO

- Choosing organisms of **similar genetic background** (littermates)
- Choosing organisms of the **same sex** if possible
- Using a **constant sample collection time and sampling sites on tissues**.
- Having the **same laboratory technician** perform RNA prep and library prep, same **lots of reagents, limit processing time range**
- If variation between samples can not be removed, use **balanced blocking design** and **modeling the block effect** during analysis
- Randomizing samples to prevent a **confounding batch** effect if all samples can't be processed at one time.

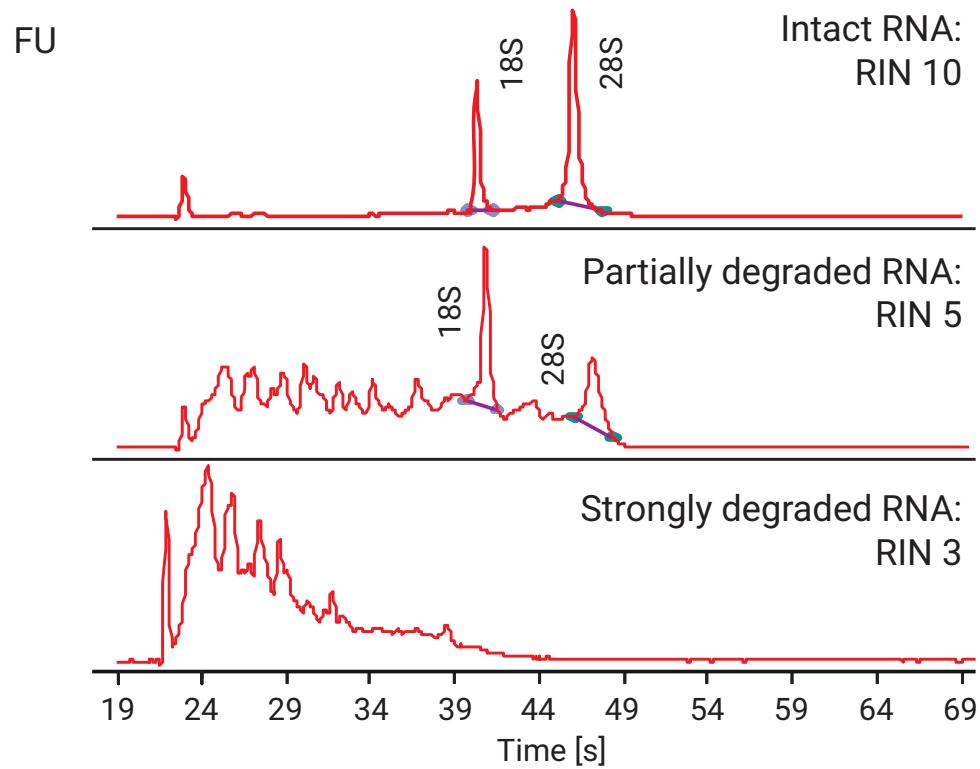
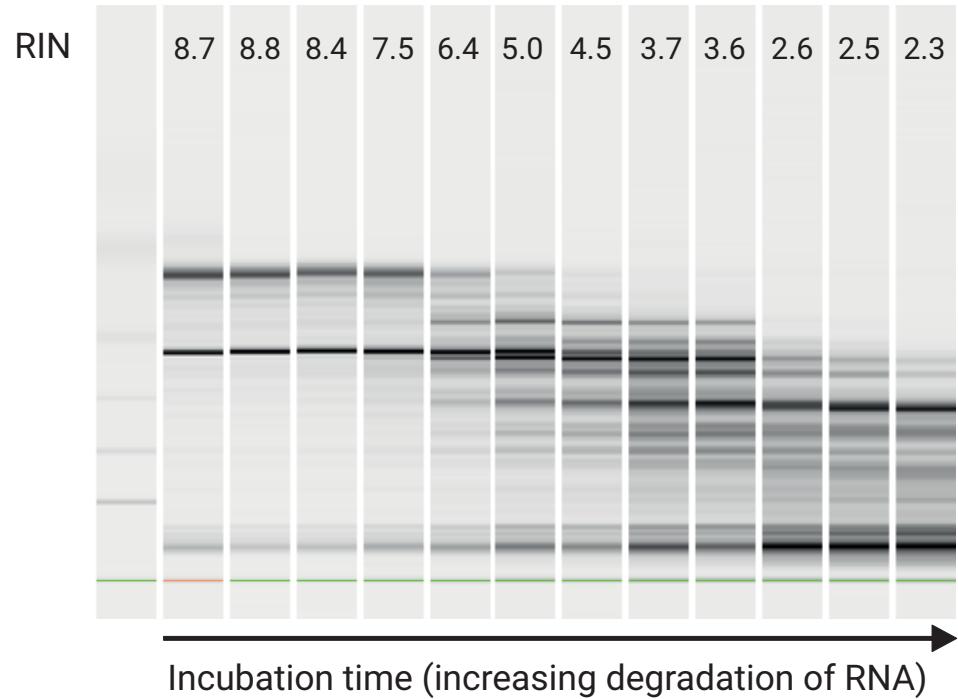
## DON'T

- Do all controls in one day and all exp another day.
- Use different reagents or kits within one experiment batch

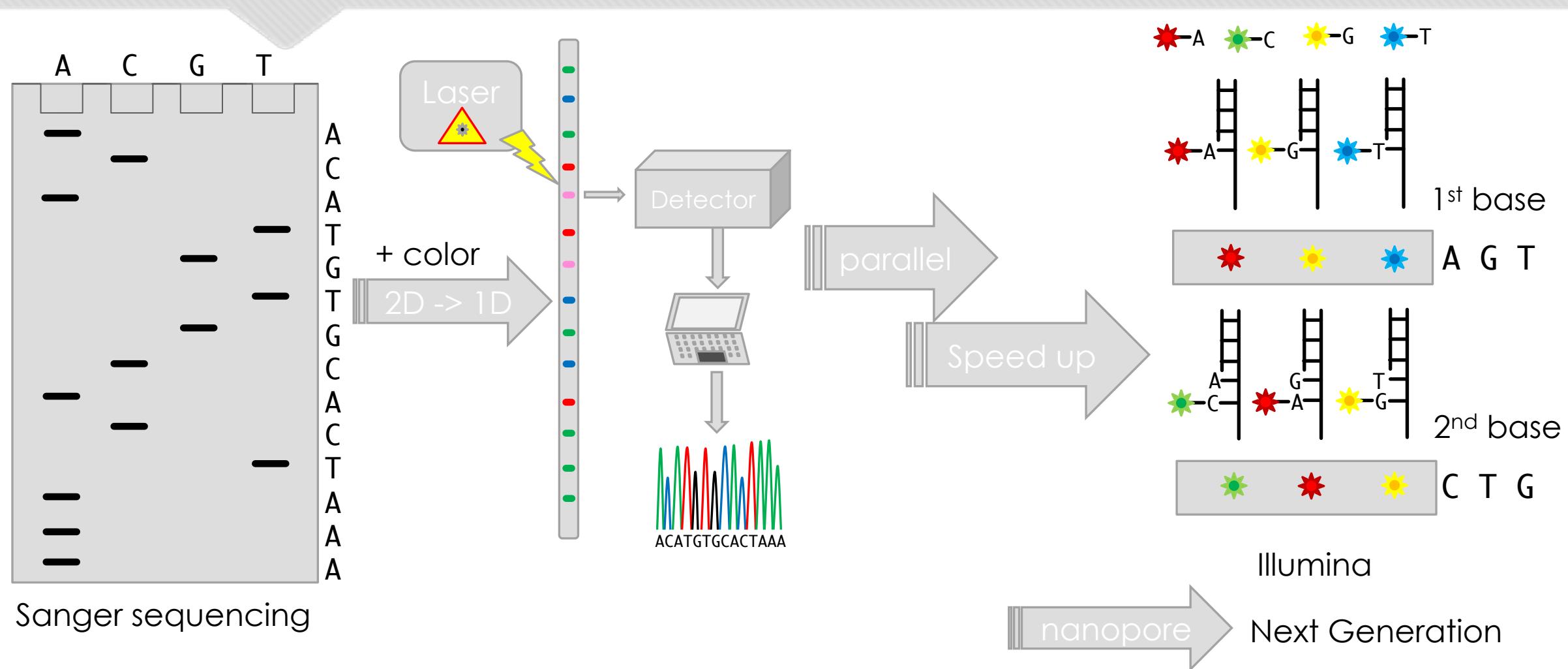
# Illumina TruSeq RNA-seq protocol



# RNA QUANTITY AND QUALITY (RIN)



# Next Generation Sequencing (NGS)



# RNA-seq data analysis pipeline

Reads

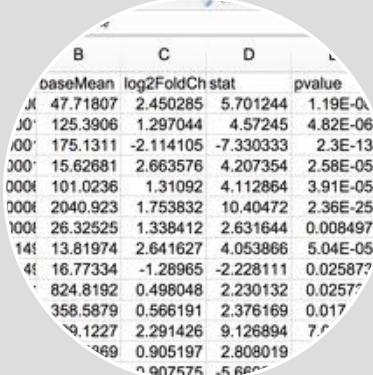


Alignments

Read counts

Gene	Read Count
a4a	11
oca5	1065
abcb3l1	1000
abcb4	3
abcb6a	144
abcb8	748
bcb9	217
cc2	237
1	8443

Differential Expression



# PREPROCESS



# Download files from seqServer

**From:** GCB Genome Sequencing Shared Resource  
**Sent:** Monday, May 28, 2018 3:00 PM  
**To:** Jianhong Ou, Ph.D. <[jianhong.ou@duke.edu](mailto:jianhong.ou@duke.edu)>;  
**Cc:** GCB Genome Sequencing Shared Resource <[sequencing@duke.edu](mailto:sequencing@duke.edu)>  
**Subject:** Your HiSeq Order #xxxx

xxx,

Your run performed within specifications.

See your report at: [http://seqweb.gcb.duke.edu/18/05/g2u4ouwzjboh0w\\_4789\\_180525B1.html](http://seqweb.gcb.duke.edu/18/05/g2u4ouwzjboh0w_4789_180525B1.html)

See instruction below to retrieve your data.

You can download your data using unix command line (sftp) or a sftp client such as FileZilla or Cyberduck on an Apple computer.

The example of unix ftp command for downloading a directory recursively: get -r directory

\*\*\*\*\*

Instruction for data download

\*\*\*\*\*

sftp server name: dnaseq2.igsp.duke.edu

Username: =====

Password: =====

Your data is under path/to/your/data

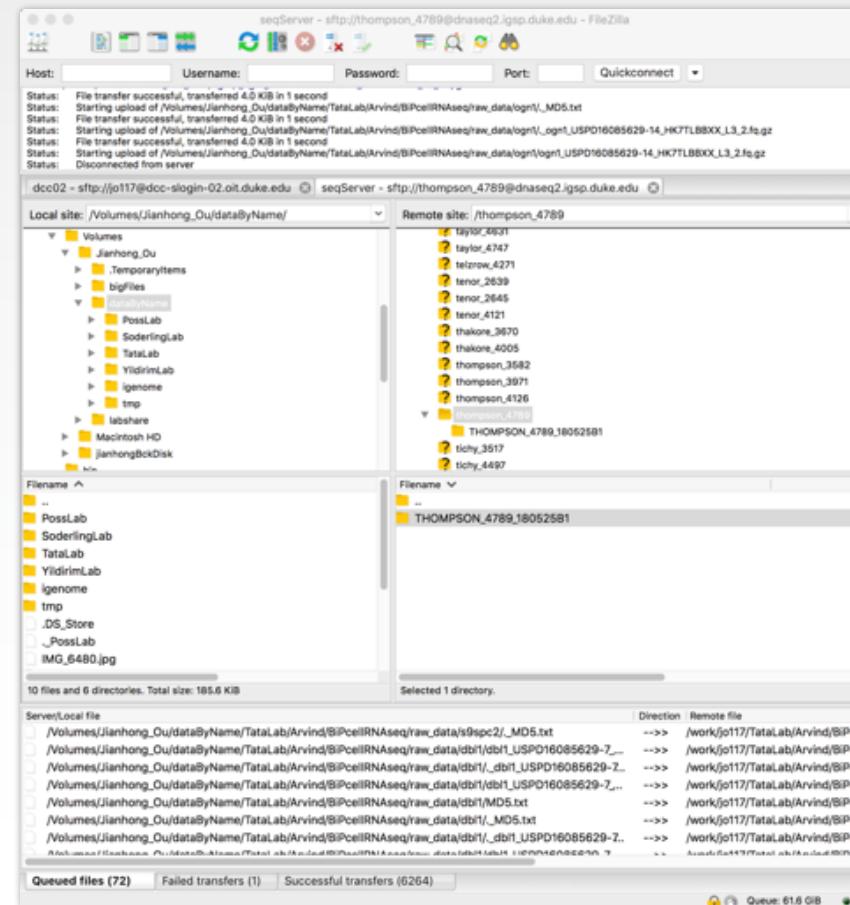
Reminder, your data will be available to download until 2018-06-27

Please, let us know if you have any questions or concerns.

Best regards and thank you for your business.

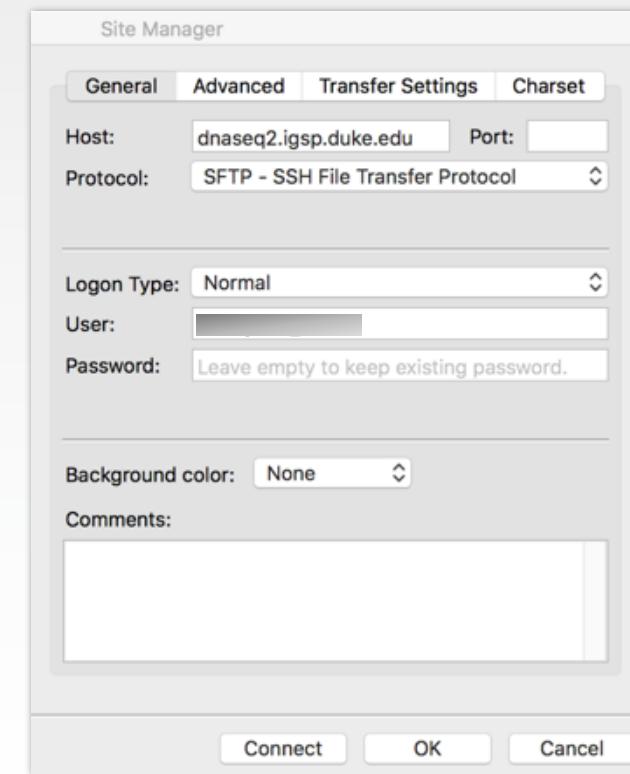
Nicolas

GCB Sequencing and Genomic Technology Shared Resource



# Install FileZilla and download a file

- Goto <https://software.duke.edu/>
- NetID Login
- Search for software: FileZilla
- Click FileZilla
- Click download here
- Download FileZilla Client
- Install FileZilla Client



# File formats: FASTA and FASTQ

## ○ FASTA file:

- Description line: each sequence started with a ">" symbol, and with or without sequence name
- Following the initial line is the actual sequences

## ○ FASTQ file:

- Line1: begins with a '@' character and is followed by a sequence identifier and an optional description
- Line2: is the raw sequence letters
- Line3: begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again
- Line4: encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence

```
>seq1
ACGTACGTACGT
ACGTACGTACGT
>seq2
TGCATGCATGCA

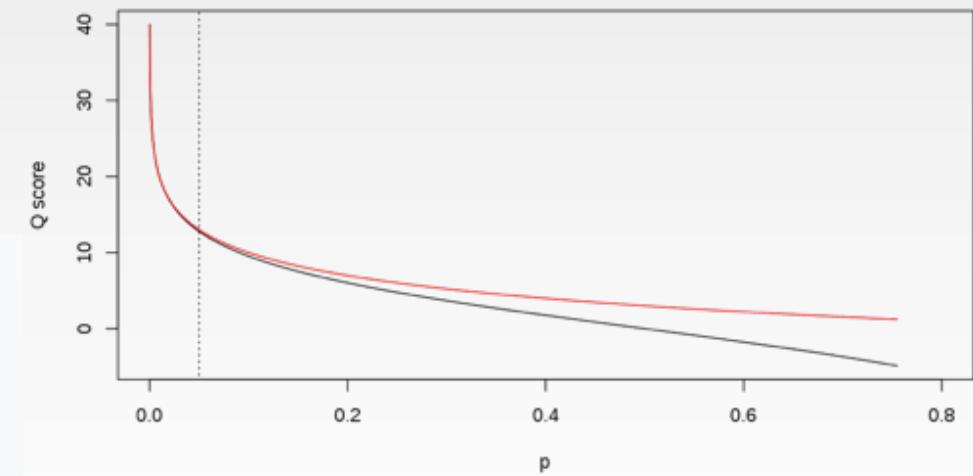
@seq1
ACGTACGTACGTACGTACGT
+
ABCDEFGHI!#%1234567890+-
@seq2
TGCATGCATGCA
+
ABCDEFGHI!#%
```

# FASTQ: quality score

# Quality

$$Q_{sanger} = -10 \log_{10} p \quad ; \quad Q_{solexa} = -10 \log_{10} \frac{p}{1-p}$$

# Encoding



Relationship between  $Q$  and  $p$  using the Sanger (red) and Solexa (black) equations (described above). The vertical dotted line indicates  $p = 0.05$ , or equivalently,  $Q \approx 13$ .

# Submitting high-throughput sequence data to Gene Expression Omnibus (GEO)

This screenshot shows the 'Raw Files' section of the GEO Metadata Template. It includes a table with columns for 'file name', 'file type', and 'file checksum'. The data is as follows:

file name	file type	file checksum
Run123abc.csfasta	solid_native_csfasta	6cc6ee3cce10b970e
Run123abc_QV.qual	solid_native_qual	88ceb0e0d056dda92
2011_01_gfh_qseq.txt	Illumina_native_qseq	95cf1d1fa509d871b2
DS18389-7_1.fastq	fastq	95cf1d1fa509d871b2
DS18389-7_2.fastq	fastq	0wd6ee3cce10b970e

This screenshot shows the 'Instrument Model' and 'Read Length' section of the GEO Metadata Template. It includes a table with columns for 'instrument model', 'read length', and 'single or paired-end'. The data is as follows:

instrument model	read length	single or paired-end
AB SOLiD System 3.0	50	single
AB SOLiD System 3.0	50	single
Illumina HiSeq 2000	72	single
Illumina HiSeq 2000	50	paired-end
Illumina HiSeq 2000	50	paired-end

# File formats: md5 and compressed files

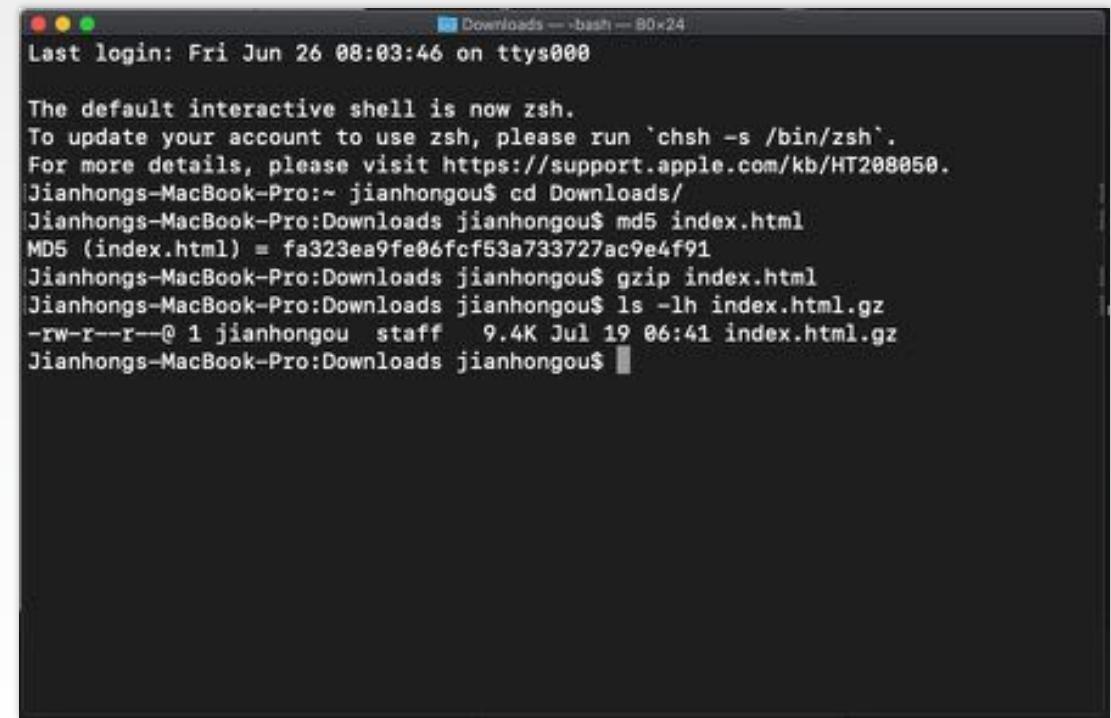
## ○ **md5**

- **Unix:** md5sum <file>
- **OS X:** md5 <file>
- **Windows:** Application required. Many are available for free download.  
Eg: FCIV -md5 <file>; certUtil -hashfile <file> MD5

- **Data File Compression:** Individual files can be compressed to speed transfer, but this is not required. Acceptable compression formats are gzip and bzip2 (i.e. files ending with a .gz or .bz2 extension). Never compress binary files (e.g., BAM, bigWig, bigBed), and DO NOT upload ZIP archives (files with a .zip extension).

# Exercises

- Get checksum for a file
- Compress and decompress a file



A screenshot of a Mac OS X terminal window titled "Downloads -- bash -- 80x24". The window shows a login message and then a series of shell commands. The commands include changing the default shell to zsh, navigating to the Downloads directory, calculating the MD5 checksum of "index.html", compressing it with gzip, and listing files with ls.

```
Last login: Fri Jun 26 08:03:46 on ttys000
The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
Jianhongs-MacBook-Pro:~ jianhongou$ cd Downloads/
Jianhongs-MacBook-Pro:Downloads jianhongou$ md5 index.html
MD5 (index.html) = fa323ea9fe06fcf53a733727ac9e4f91
Jianhongs-MacBook-Pro:Downloads jianhongou$ gzip index.html
Jianhongs-MacBook-Pro:Downloads jianhongou$ ls -lh index.html.gz
-rw-r--r--@ 1 jianhongou staff 9.4K Jul 19 06:41 index.html.gz
Jianhongs-MacBook-Pro:Downloads jianhongou$ █
```

# Backup your data

- Keep two copies in physically different storage spaces with their checksum.
  - Keep it in the cloud: Duke's Box (50G), OneDrive (1TB)
  - External Hard Drive
  - Burn it to a Blu-ray Disc
  - Put it on a USB Flash Drive
  - Save it to Network Attached Storage (NAS) Device

# FASTQC

Function

A quality control tool for high throughput sequence data.

Language

Java

Requirements

A suitable Java Runtime Environment

Code Maturity

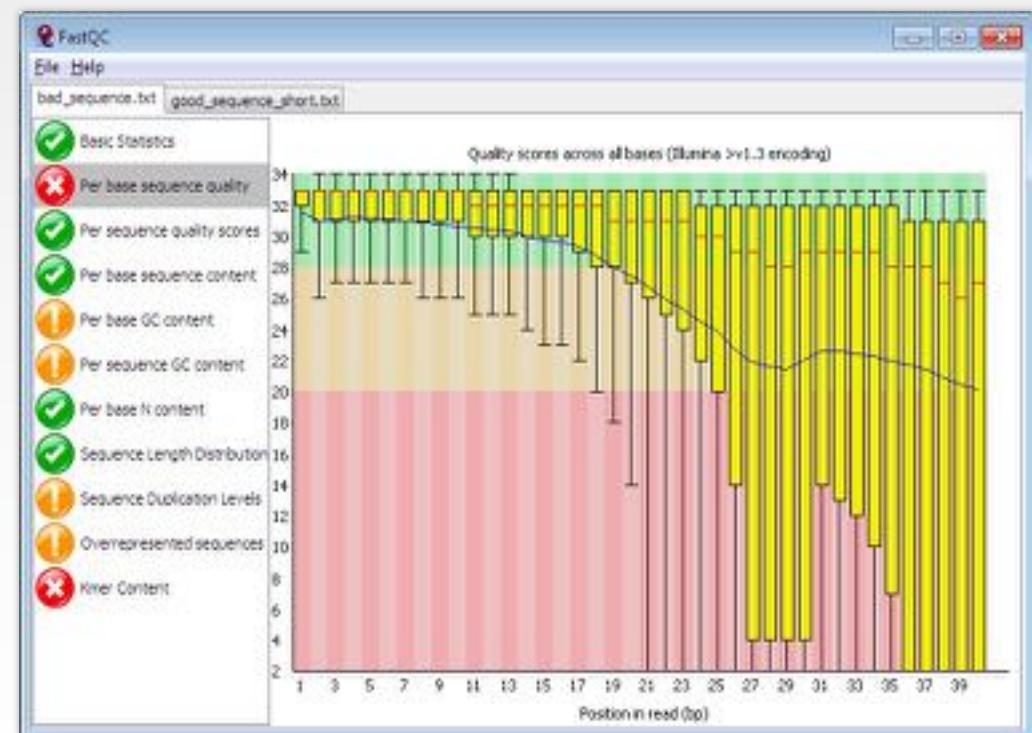
Stable.

Code Released

Yes, under GPL v3 or later.

Initial Contact

Simon Andrews

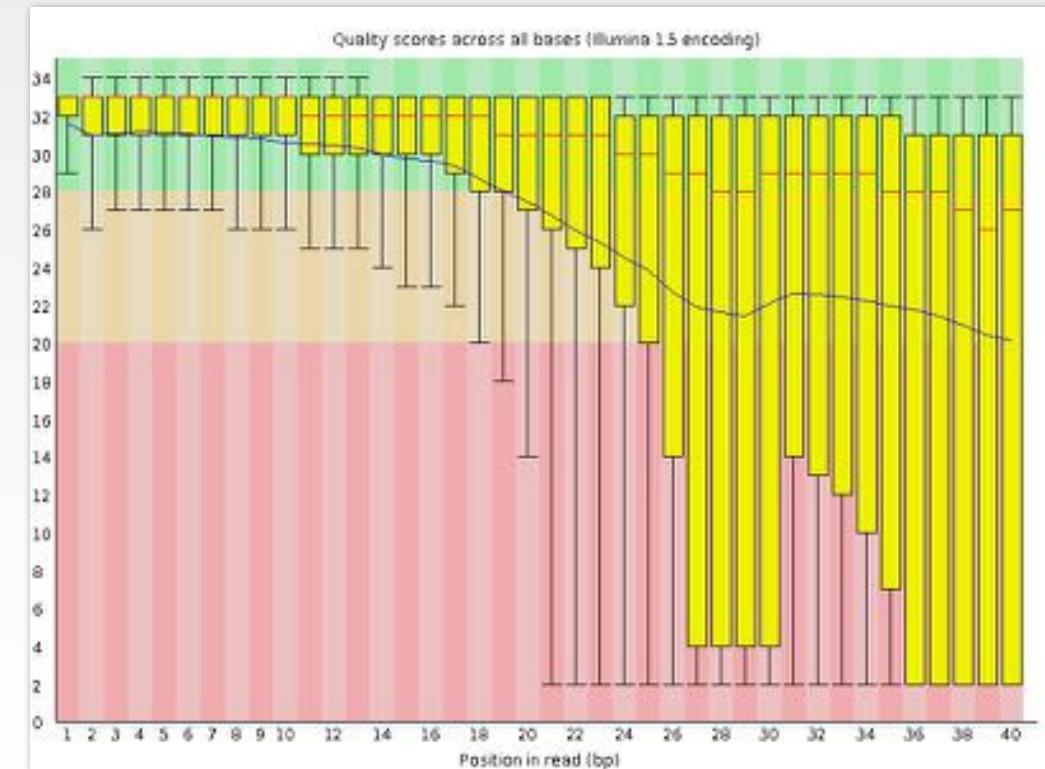
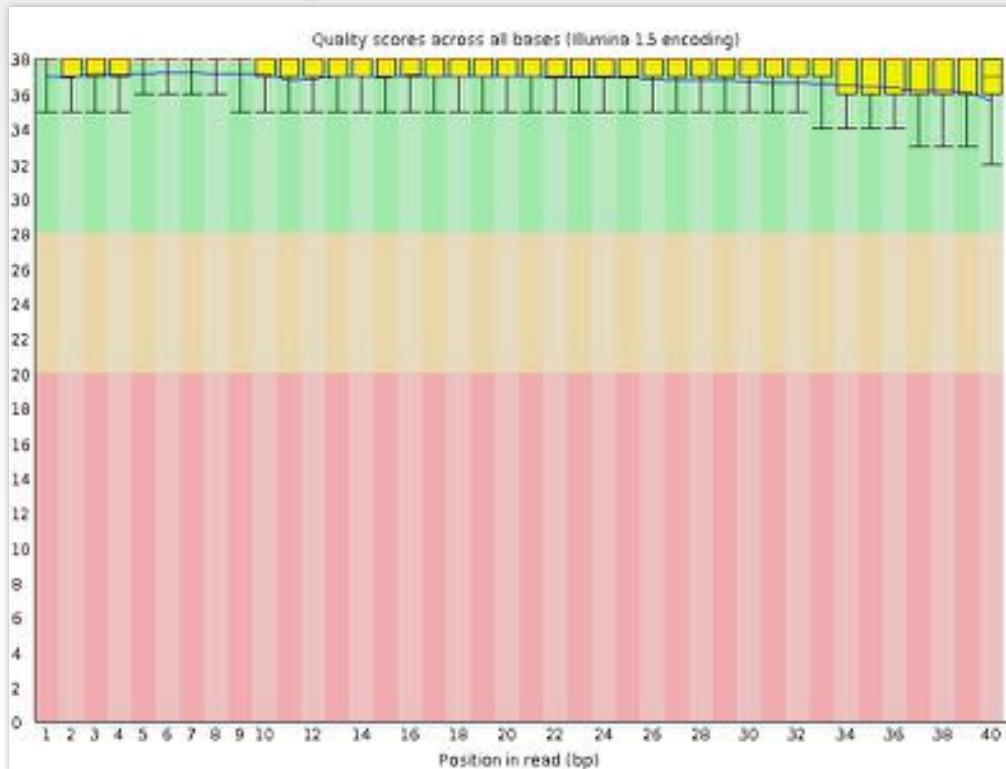


# FASTQC: Basic Statistics

Measure	Value	Description
Filename	good_sequence_short.txt	The original filename of the file which was analysed
File type	Conventional base calls	Says whether the file appeared to contain actual base calls or colorspace data which had to be converted to base calls
Encoding	Illumina 1.5	Says which ASCII encoding of quality values was found in this file
Total Sequences	250000	A count of the total number of sequences processed
Sequences flagged as poor quality	0	A count of the number of sequences flagged as poor quality
Sequence length	40	Provides the length of the shortest and longest sequence in the set
%GC	45	The overall %GC of all bases in all sequences

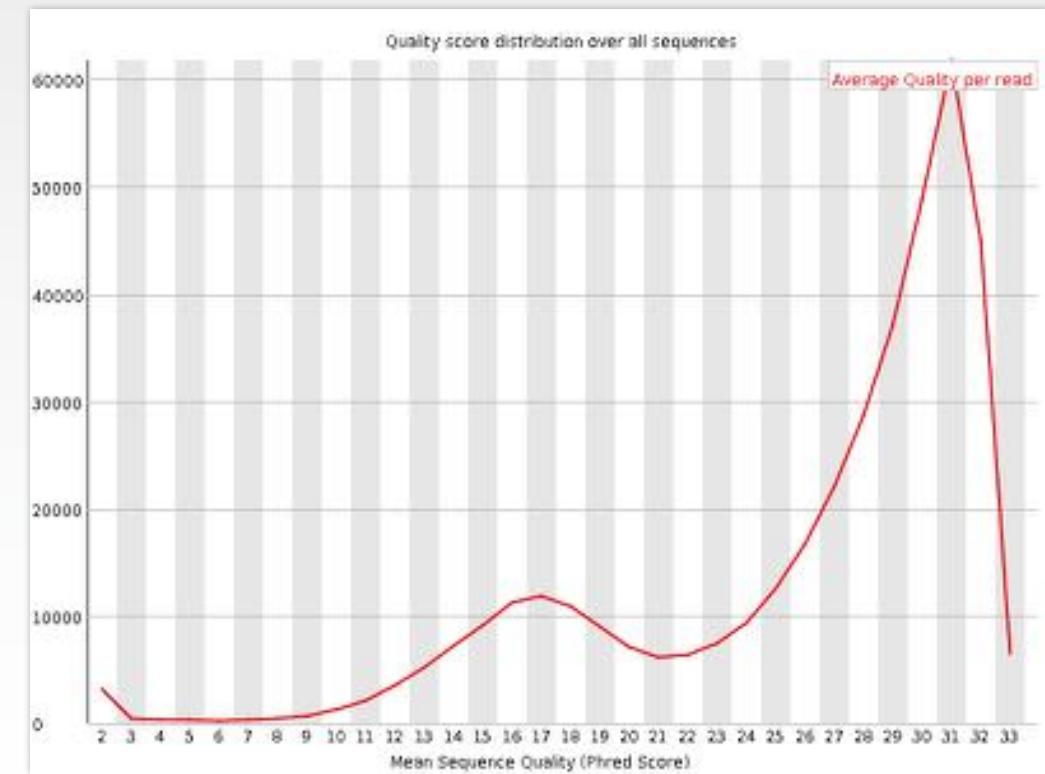
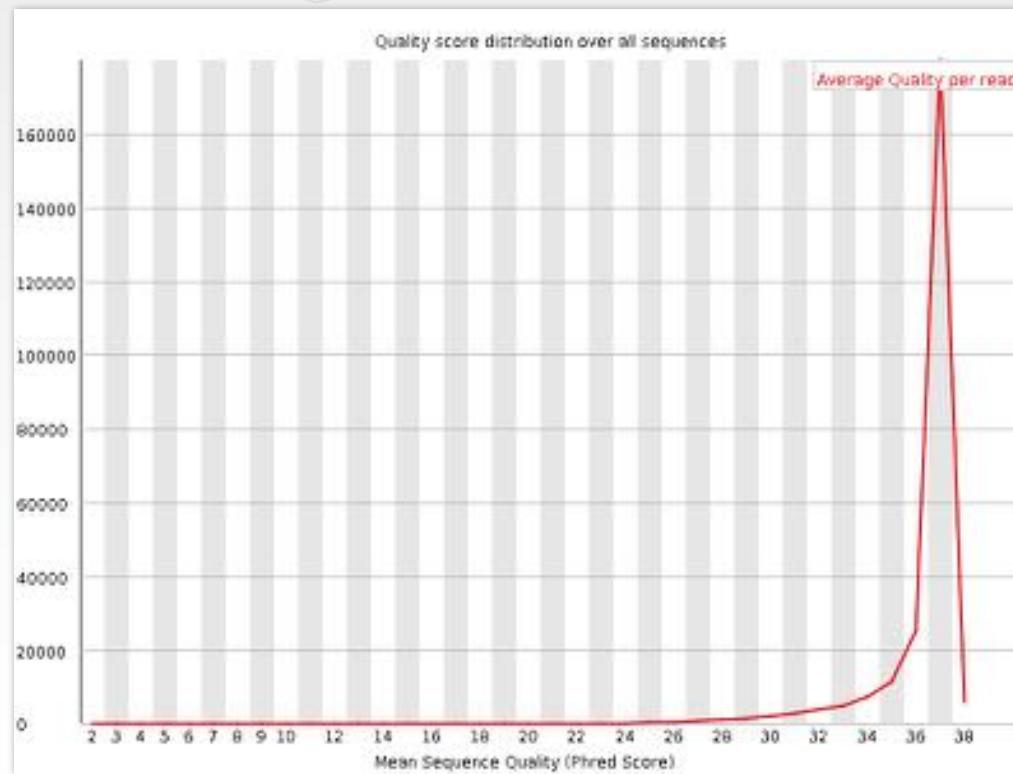
# FASTQC: Per base sequence quality

$$Q_{sanger} = -10 \log_{10} p \quad ; \quad Q_{solexa} = -10 \log_{10} \frac{p}{1-p}$$



**Warning:** A warning will be issued if the lower quartile for any base is less than 10, or if the median for any base is less than 25.  
**Failure:** This module will raise a failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20.

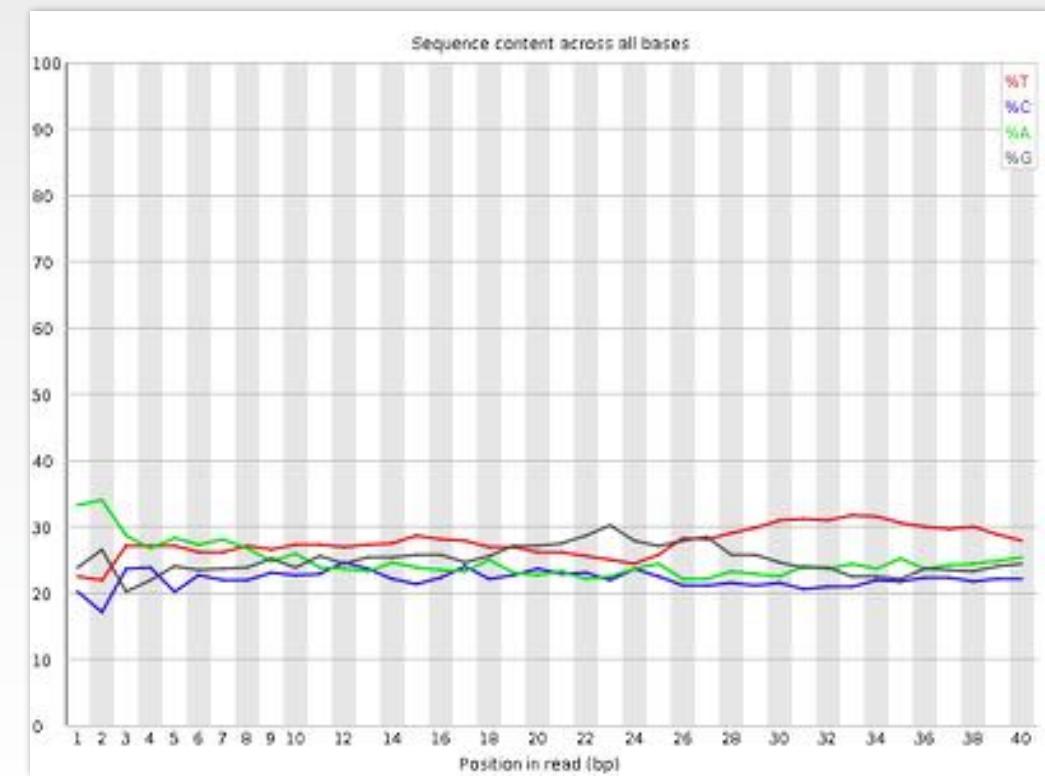
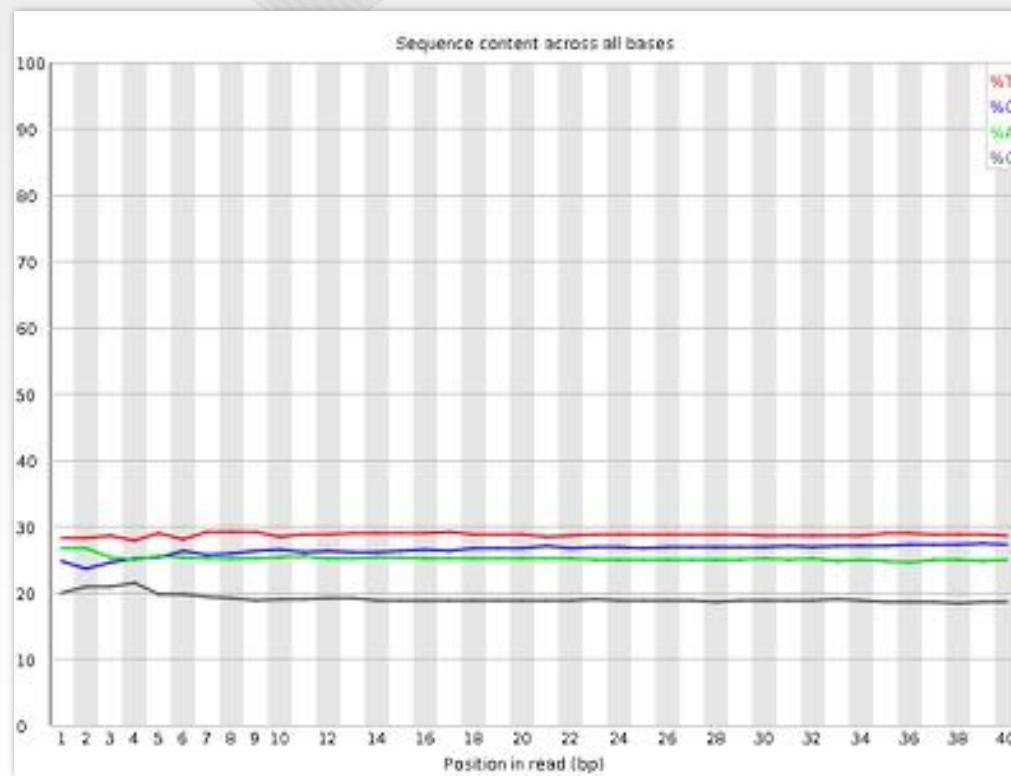
# FASTQC: Per sequence quality scores



**Warning:** A warning is raised if the most frequently observed mean quality is below 27 - this equates to a 0.2% error rate.

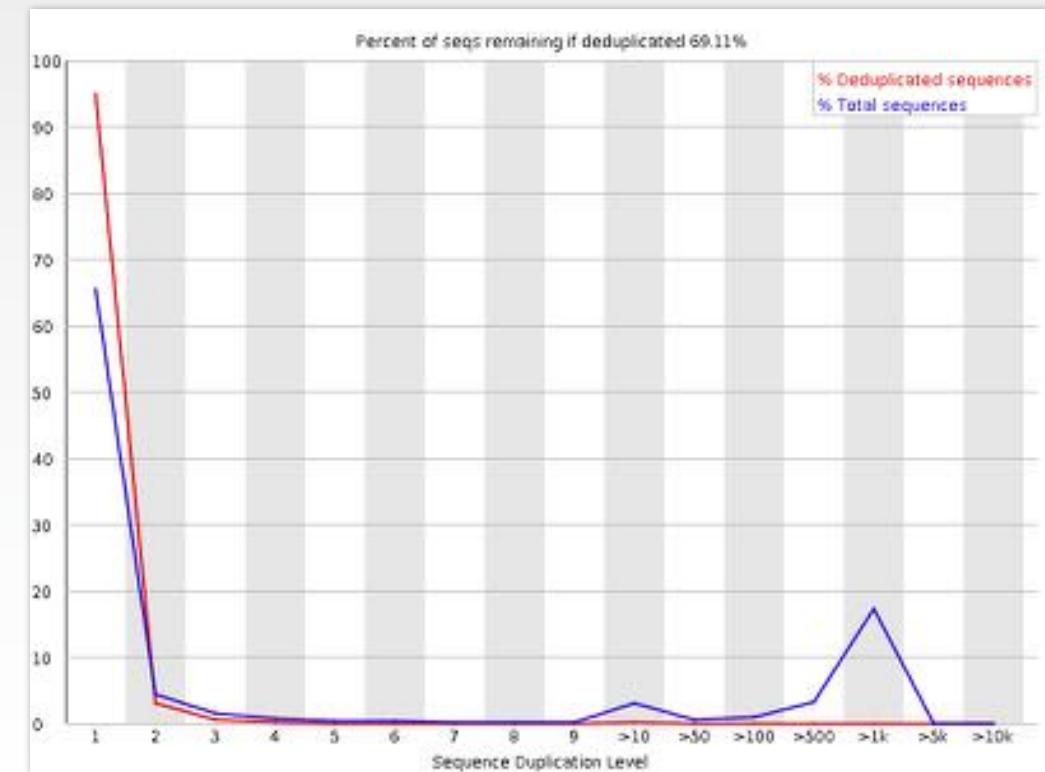
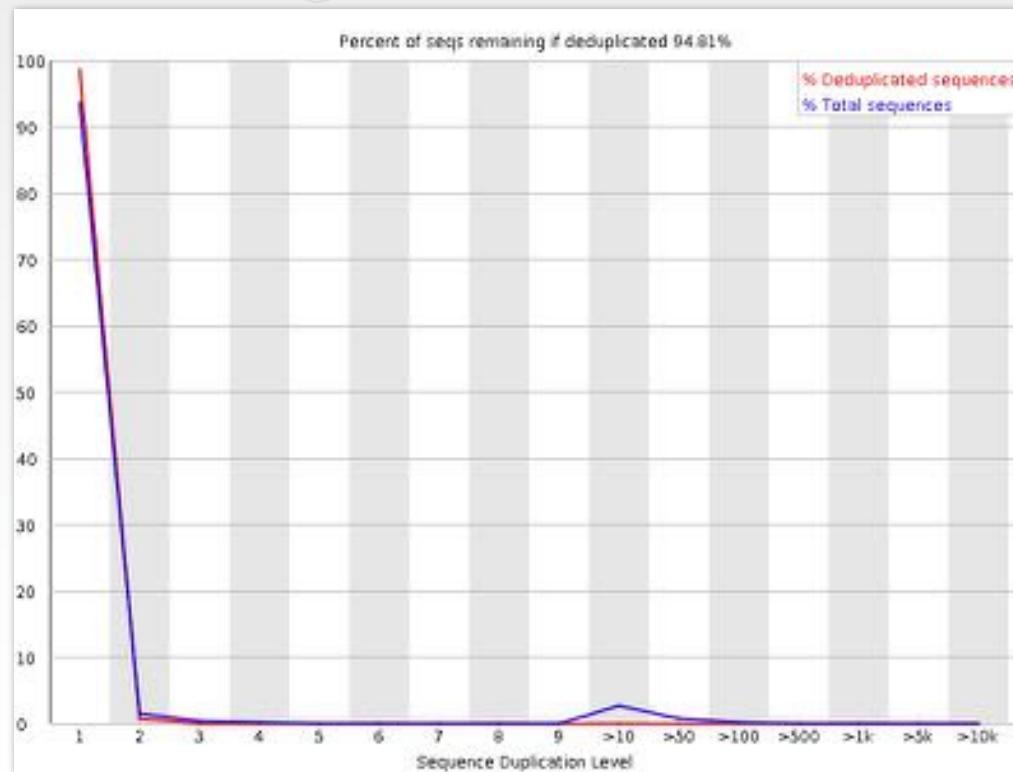
**Failure:** An error is raised if the most frequently observed mean quality is below 20 - this equates to a 1% error rate.

# FASTQC: Per base sequence content



**Warning:** This module issues a warning if the difference between A and T, or G and C is greater than 10% in any position.  
**Failure:** This module will fail if the difference between A and T, or G and C is greater than 20% in any position.

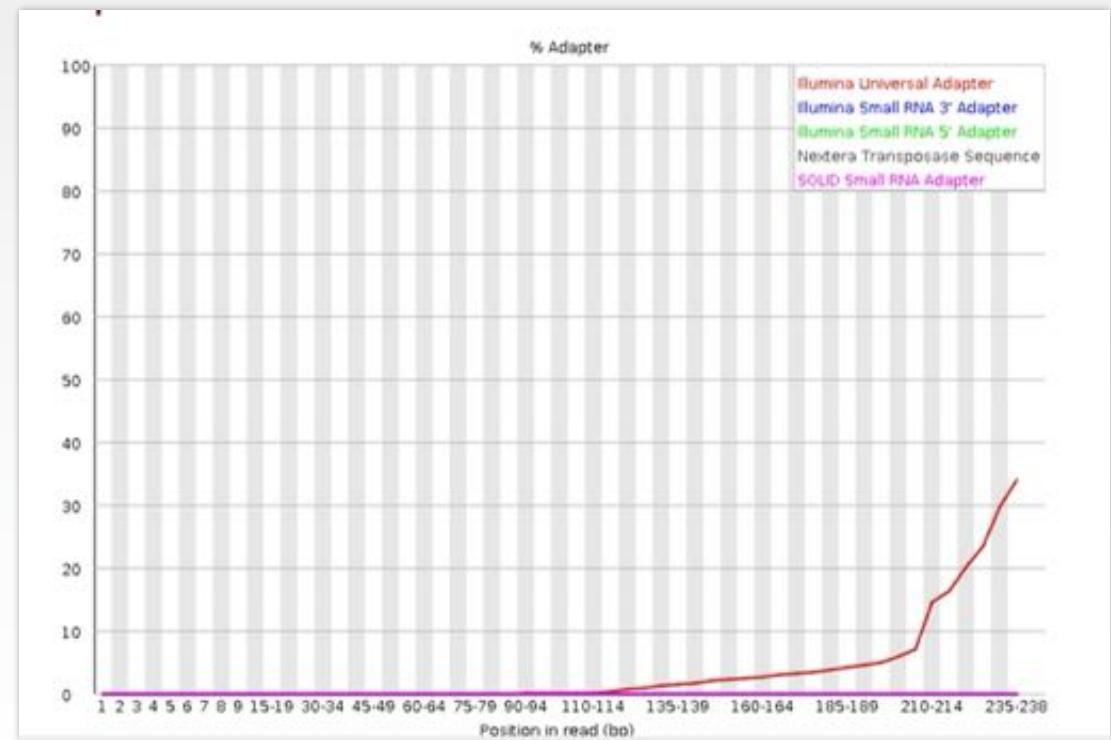
# FASTQC: Sequence Duplication Levels



**Warning:** This module will issue a warning if non-unique sequences make up more than 20% of the total.  
**Failure:** This module will issue an error if non-unique sequences make up more than 50% of the total.

# Trimming low-quality bases and adapters

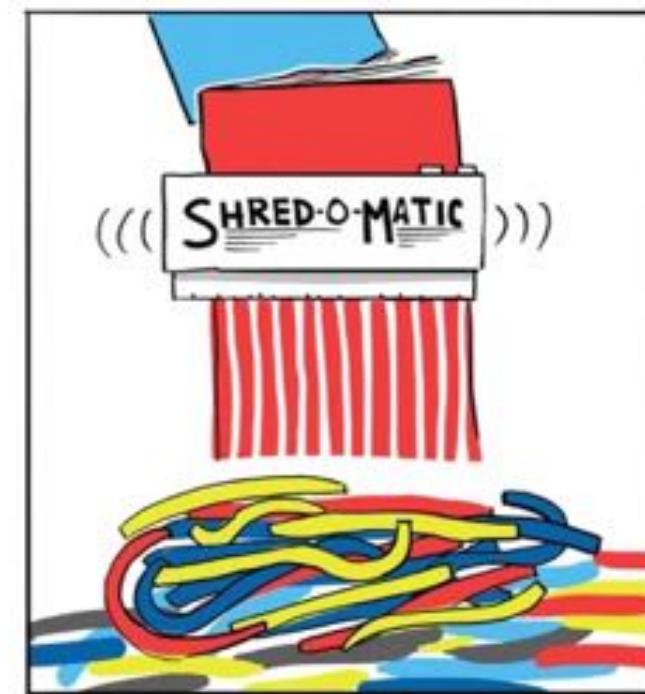
- Cutadapt, Trim Galore!
- Trimmomatic
  - Using relaxed trimming thresholds
  - Improve mappability
  - Reduce mapping artifacts



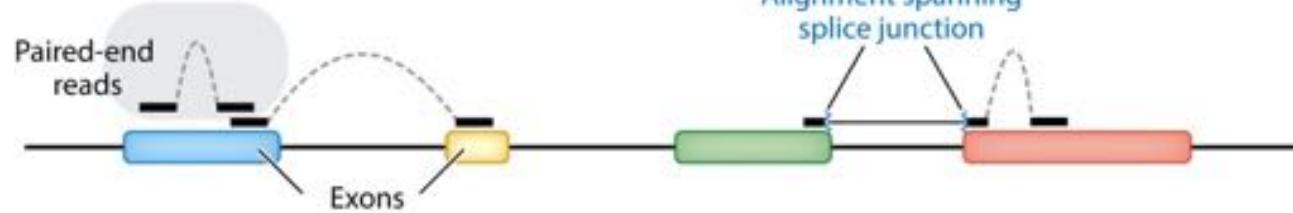
A large red 3D puzzle piece stands out from a background of white puzzle pieces. The red piece is positioned centrally and is slightly tilted, symbolizing a missing or unique element in a process.

# **ALIGNMENT AND QUANTIFICATION**

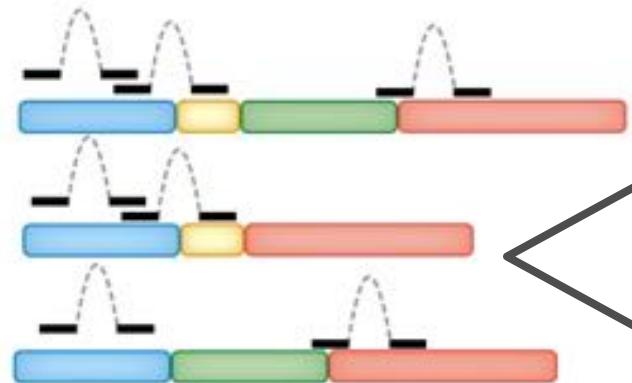
Genome reconstruction—akin to reassembling magazine articles after they have been through a paper shredder.



**a** Spliced alignment against genome using splice-aware aligner: STAR, HISAT2, GMAP, Tophat2



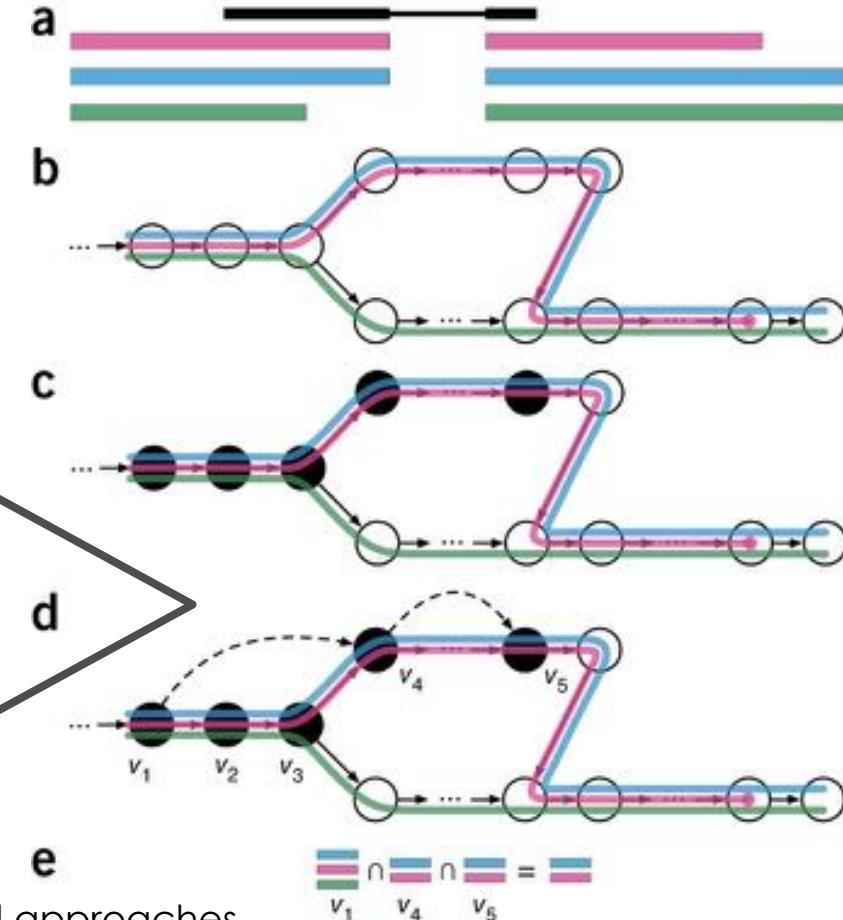
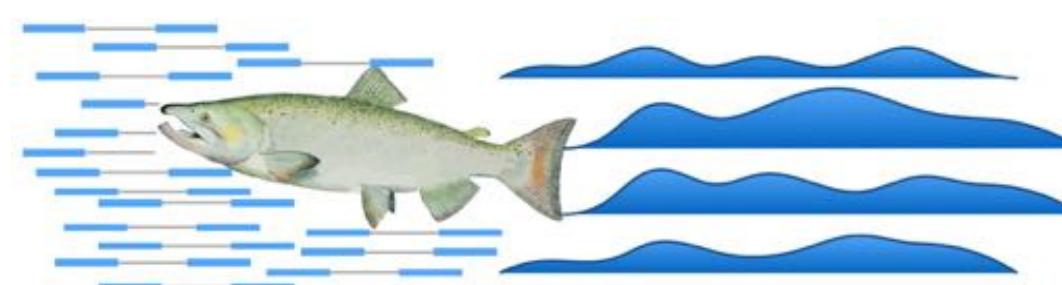
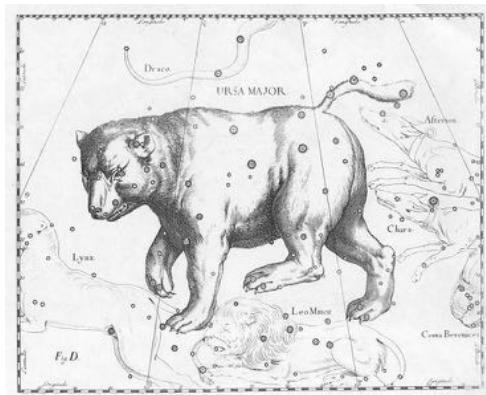
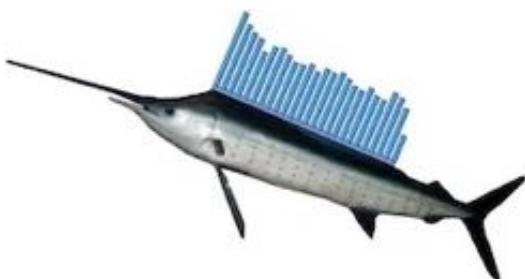
**b** Unspliced alignment against transcriptome using ungapped aligner: Bowtie, bwa



well  
annotated  
reference  
genome

Van den Berge K, et al. 2019.  
*Annu. Rev. Biomed. Data Sci.* 2:139–73

**Alignment free or Pseudo-alignment against transcriptome:** **Salmon, Sailfish, Kallisto**



Bray et al., 2016. doi: 10.1038/nbt.3519

# tximport + DESeq2/edgeR/limma-voom

- STAR (Spliced Transcripts Alignment to a Reference)
- RSEM(RNA-Seq by Expectation-Maximization)
- StringTie(downstream followed by Ballgown)
- =====
- Sailfish (Patro, Mount, and Kingsford 2014)
- Kallisto (Bray et al. 2016)
- Salmon (Patro et al. 2017)

**Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms**

**Fleximer: Accurate uantification of RNA-Seq via Variable-Length k-mers**

**Near-optimal probabilistic RNA-seq quantification**

**Salmon provides fast and bias-aware quantification of transcript expression**



“Lightweight”  
approaches

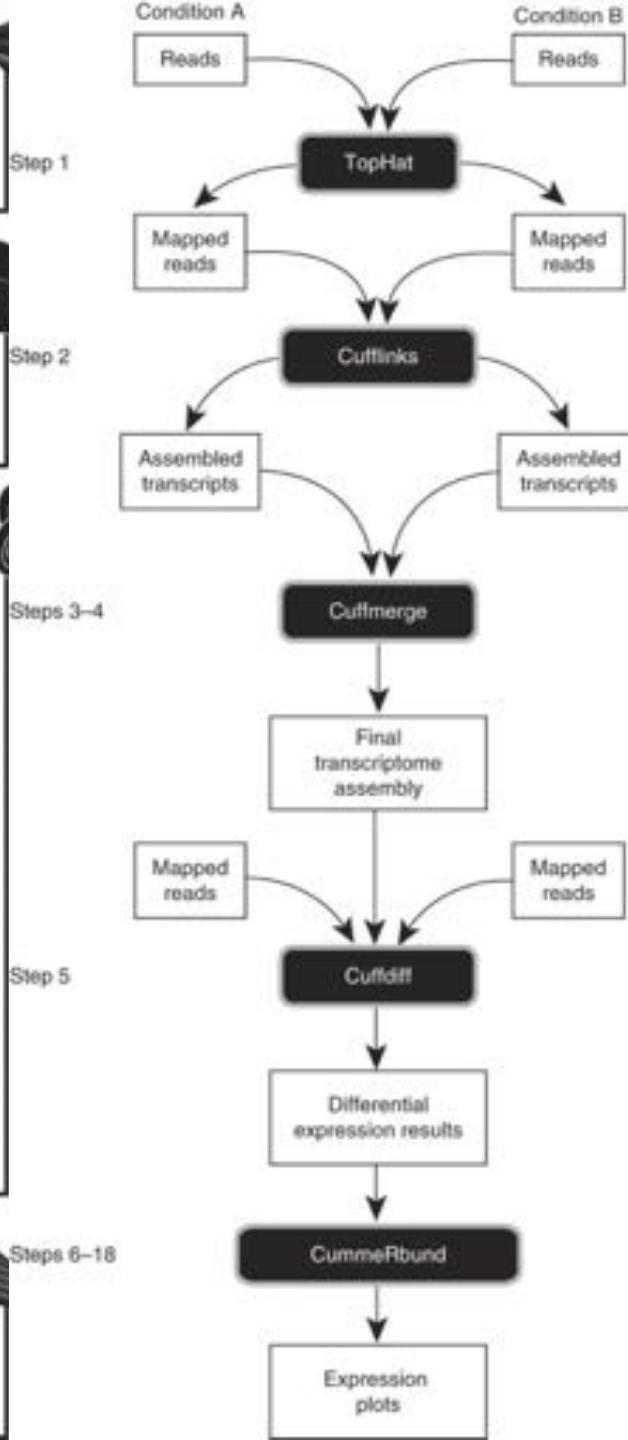
**Bowtie**  
Extremely fast, general purpose short read aligner

**TopHat**  
Aligns RNA-Seq reads to the genome using Bowtie  
Discovers splice sites

**Cufflinks package**

- Cufflinks Assembles transcripts
- Cuffcompare Compares transcript assemblies to annotation
- Cuffmerge Merges two or more transcript assemblies
- Cuffdiff Finds differentially expressed genes and transcripts  
Detects differential splicing and promoter use

**CummeRbund**  
Plots abundance and differential expression results from Cuffdiff



## PROTOCOL

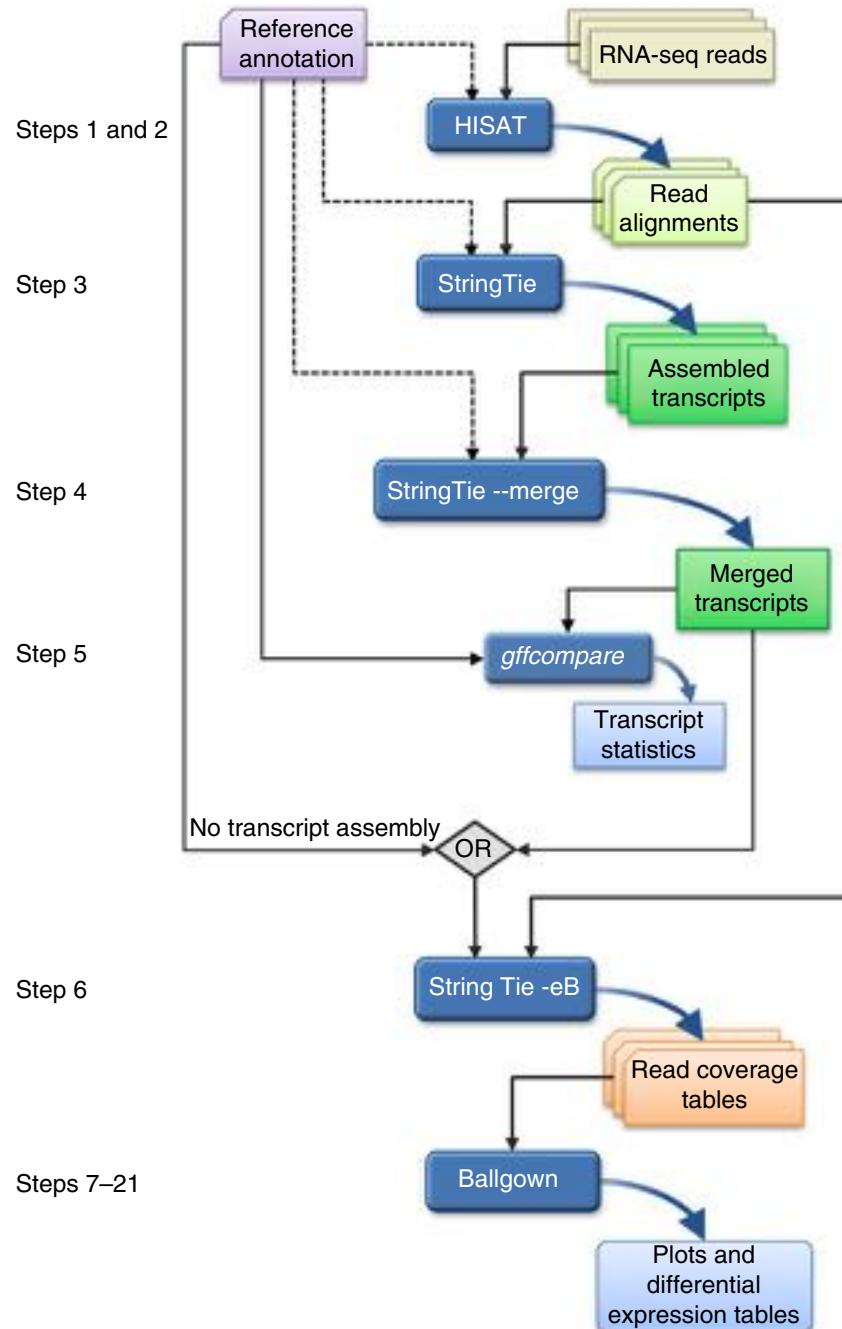
# Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

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

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. <sup>2</sup>Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. <sup>3</sup>Department of Computer Science, University of California, Berkeley, California, USA. <sup>4</sup>Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. <sup>5</sup>Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. <sup>6</sup>Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. <sup>7</sup>Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. <sup>8</sup>Department of Mathematics, University of California, Berkeley, California, USA. <sup>9</sup>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; corrected after print 7 August 2014; doi:10.1038/nprot.2012.016

- ❖ Reads are mapped directly to reference genome allowing for splice junctions across exons
- ❖ Methods require large amount of memory and CPU time
- ❖ Can only handle simple experiment design



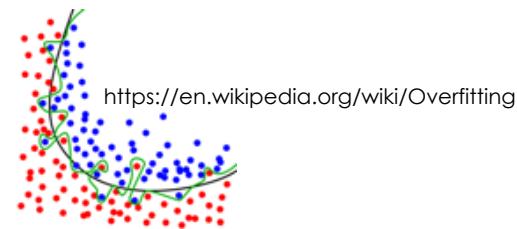
# PROTOCOL

# Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown

Mihaela Perteal<sup>1,2</sup>, Daehwan Kim<sup>1</sup>, Geo M Perteal<sup>1</sup>, Jeffrey T Leek<sup>3</sup> & Steven L Salzberg<sup>1-4</sup>

<sup>1</sup>Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. <sup>2</sup>Department of Computer Science, Whiting School of Engineering, Johns Hopkins University, Baltimore, Maryland, USA. <sup>3</sup>Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. <sup>4</sup>Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland, USA. Correspondence should be addressed to S.L.S. ([salzberg@jhu.edu](mailto:salzberg@jhu.edu)). 

Published online 11 August 2016; doi:10.1038/nprot.2016.095

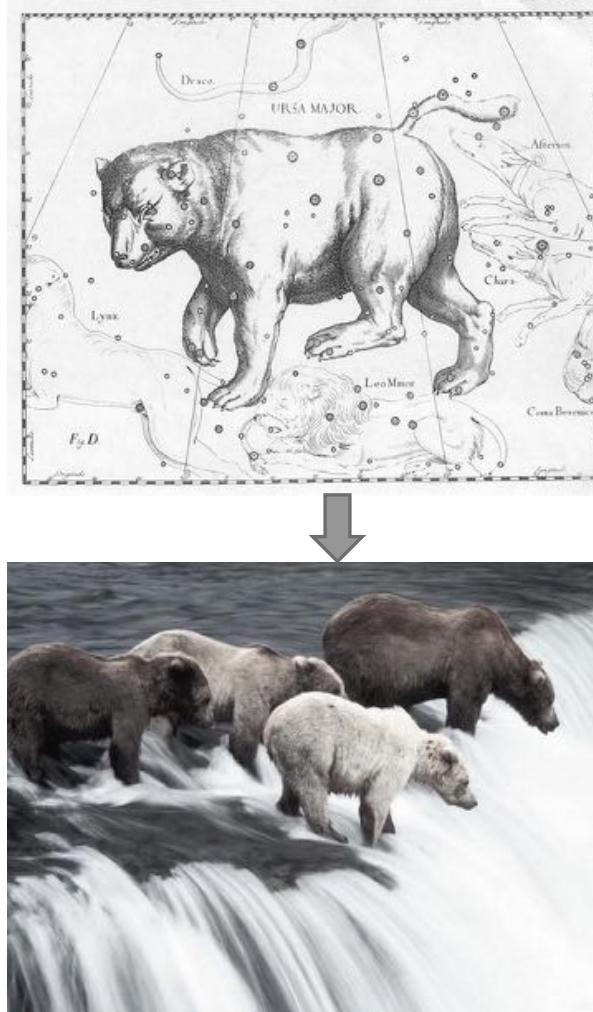


Note that Ballgown's statistical test is a standard linear model-based comparison. **For small sample sizes ( $n < 4$  per group), it is often better to perform regularization.** This can be done using the limma<sup>33</sup> package in Bioconductor. Other regularized methods such as DESeq<sup>23</sup> and edgeR<sup>20</sup> can be applied to gene or exon counts, but they are not appropriate for direct application to FPKM abundance estimates. The statistical test uses a cumulative upper quartile normalization<sup>34</sup>.

# KALLISTO, SLEUTH PIPELINE

- ❖ considerably faster than traditional alignment + counting
- ❖ Transcripts level estimates rather than gene level. Transcripts level estimates can be aggregated to gene level, but abundance estimates for lowly expressed transcripts are highly variable and should be interpreted with caution.
- ❖ No precise alignments ( no alignment files for visualization in genome browser )

<https://scilifelab.github.io/courses/rnaseq/labs/kallisto>



## Near-optimal probabilistic RNA-seq quantification

Nicolas L Bray<sup>1</sup>, Harold Pimentel<sup>2</sup>, Pál Melsted<sup>3</sup>  
& Lior Pachter<sup>2,4,5</sup>

We present kallisto, an RNA-seq quantification program that is two orders of magnitude faster than previous approaches and achieves similar accuracy. Kallisto pseudoaligns reads to a reference, producing a list of transcripts that are compatible with each read while avoiding alignment of individual bases. We use kallisto to analyze 30 million unaligned paired-end RNA-seq reads in <10 min on a standard laptop computer. This removes a major computational bottleneck in RNA-seq analysis.

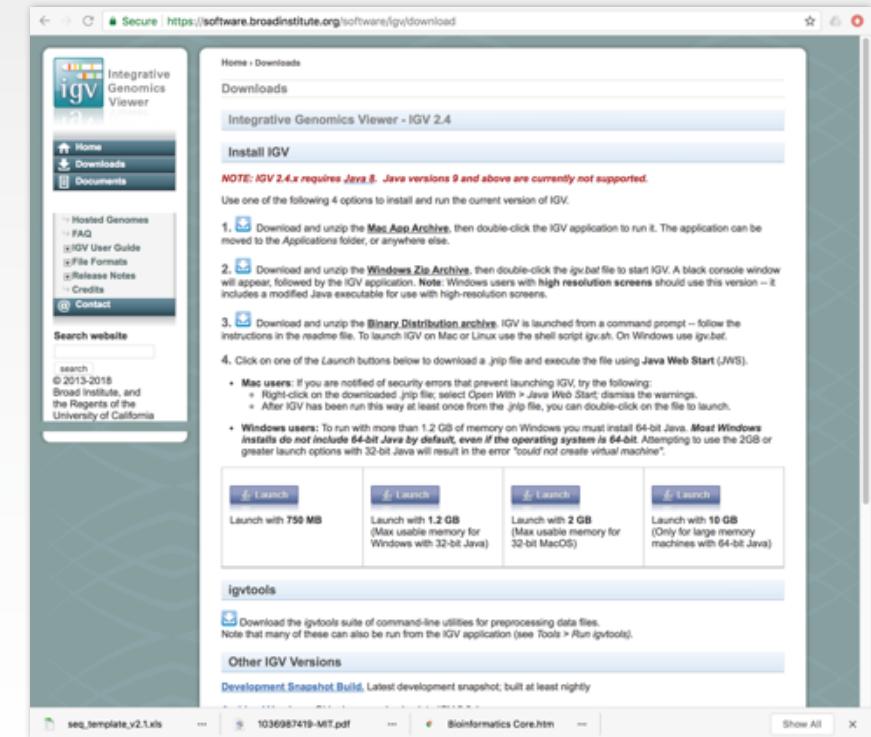
## Differential analysis of RNA-seq incorporating quantification uncertainty

Harold Pimentel<sup>1</sup>, Nicolas L Bray<sup>2</sup>, Suzette Puente<sup>3</sup>,  
Pál Melsted<sup>4</sup> & Lior Pachter<sup>5</sup>

We describe sleuth (<http://pachterlab.github.io/sleuth>), a method for the differential analysis of gene expression data that utilizes bootstrapping in conjunction with response error linear modeling to decouple biological variance from inferential variance. sleuth is implemented in an interactive shiny app that utilizes kallisto quantifications and bootstraps for fast and accurate analysis of data from RNA-seq experiments.

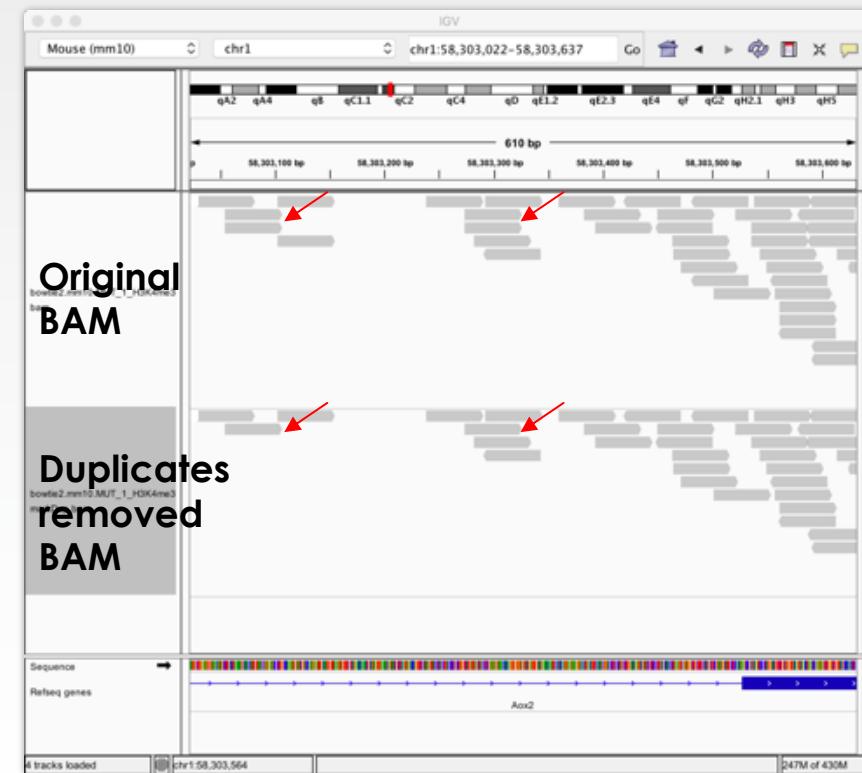
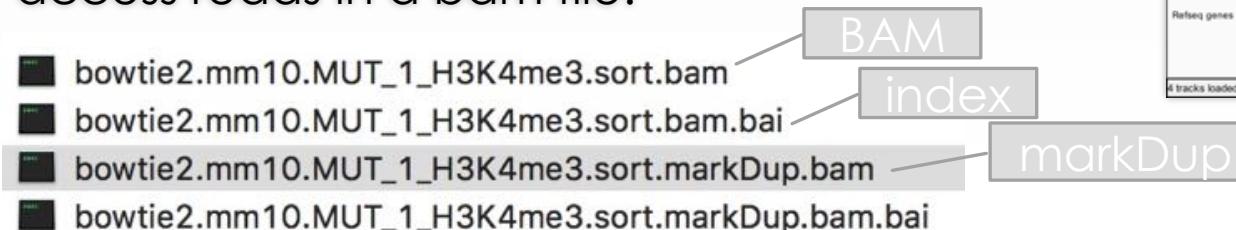
# IGV: Integrative Genomics Viewer

- ➊ Goto  
<https://software.broadinstitute.org/software/igv/>
- ➋ Click Downloads icon
- ➌ Download IGV



# File formats: sam/bam

- Sam/bam: alignment file contain mapping information.
- markDup: output of MarkDuplicates function of picard tools. Remove the duplicated reads for DNA-seq.
- Sort: sort the bam files by name or coordinates.
- Index: Bam index file for more efficiently access reads in a bam file.



# Exercises



Visualizing data in  
Integrative Genomics  
Viewer



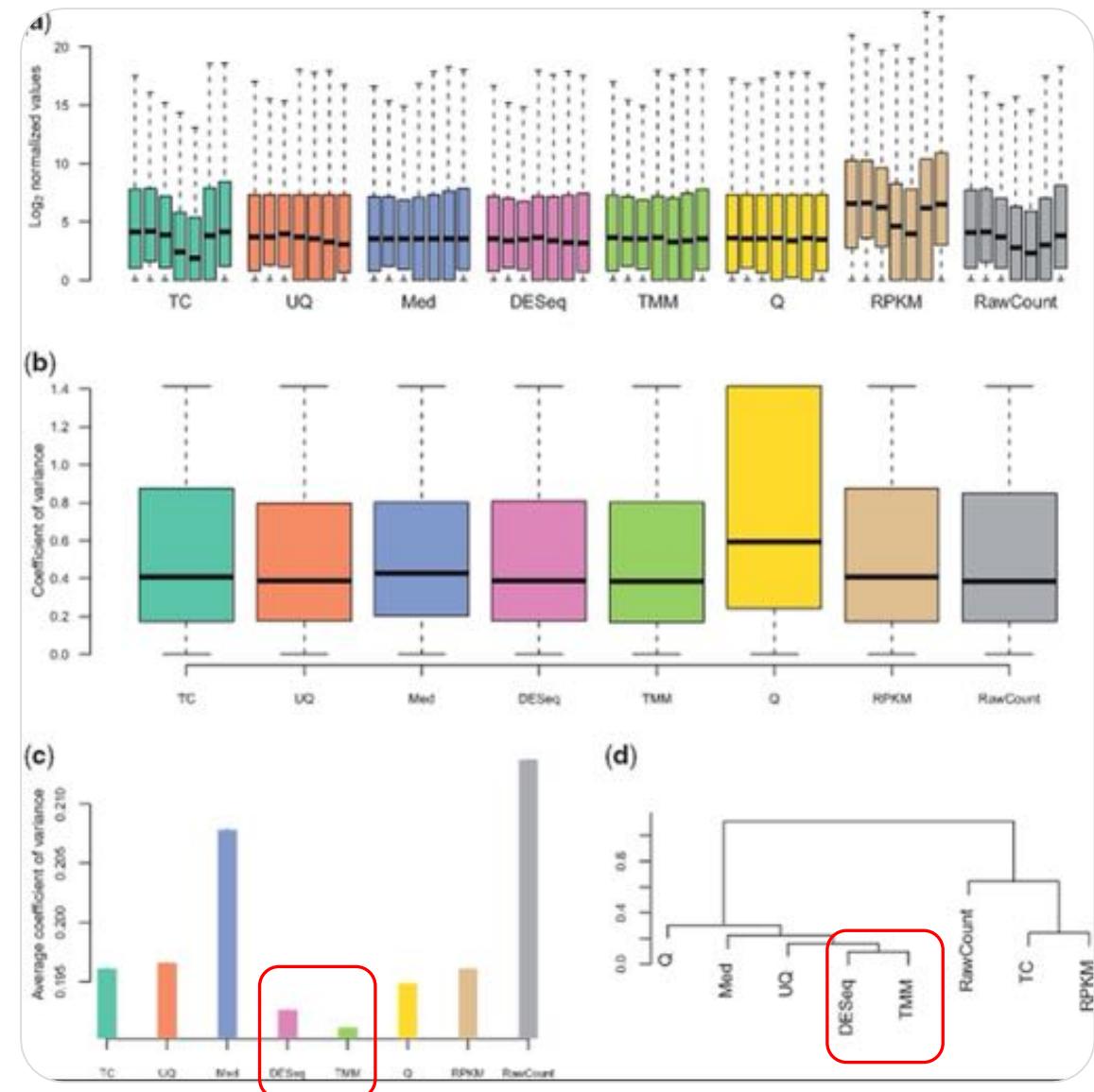
Visualizing data in UCSC  
genome browser

# COUNT TABLE

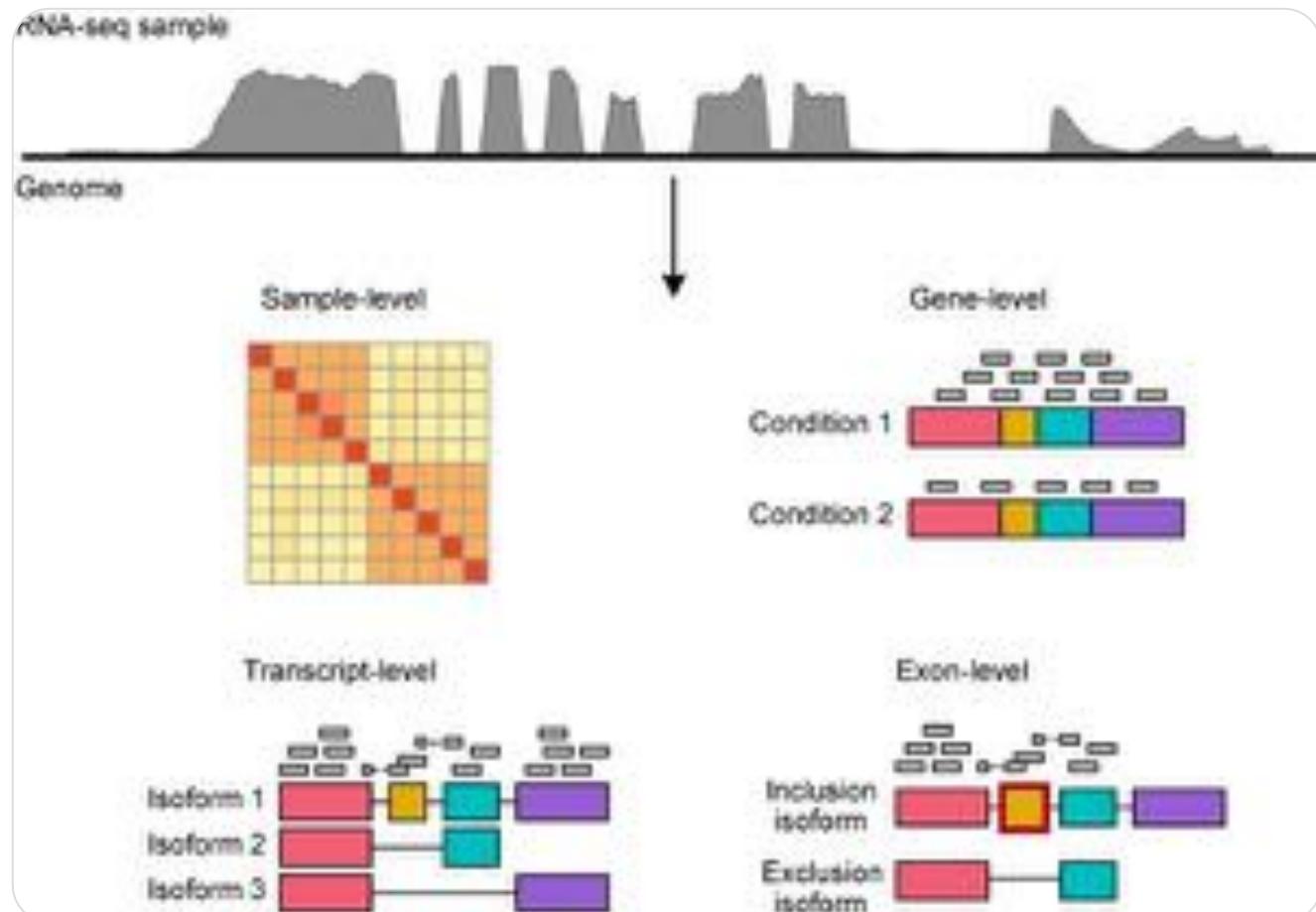
Gene Symbol	WT_rep1	WT_rep2	WT_rep3	KD_rep1	KD_rep2	KD_rep3
<b>ackr4b</b>	30	28	42	5	66	7
<b>aclya</b>	24	48	56	77	89	108
<b>acot14</b>	0	0	0	0	0	0
<b>acsbg1</b>	5	9	7	123	98	70
<b>actr8</b>	3536	3639	4586	225	545	898

# Normalization: most genes are not differential expressed

- TMM (Trimmed weighted mean)
- DESeq (median of log expression ratio)

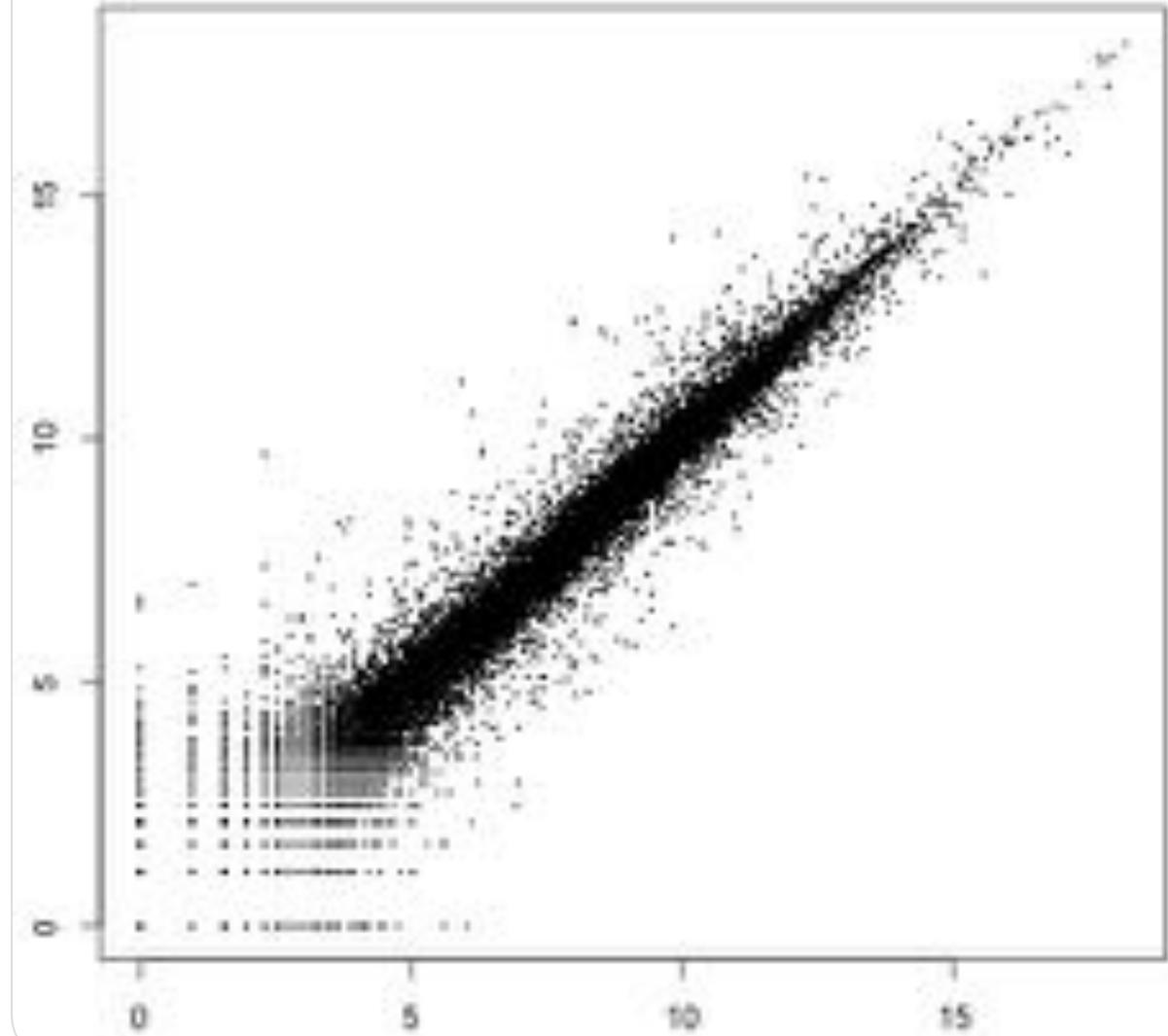
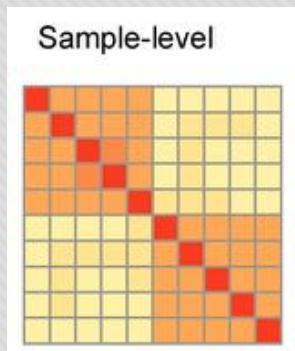


**RNA-seq analyses at four different levels: sample-level, gene-level, transcript-level, and exon-level.**

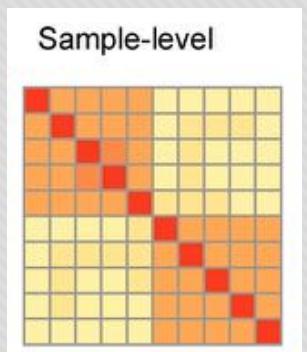


Li et.al., 2018. doi: 10.1007/s40484-018-0144-7

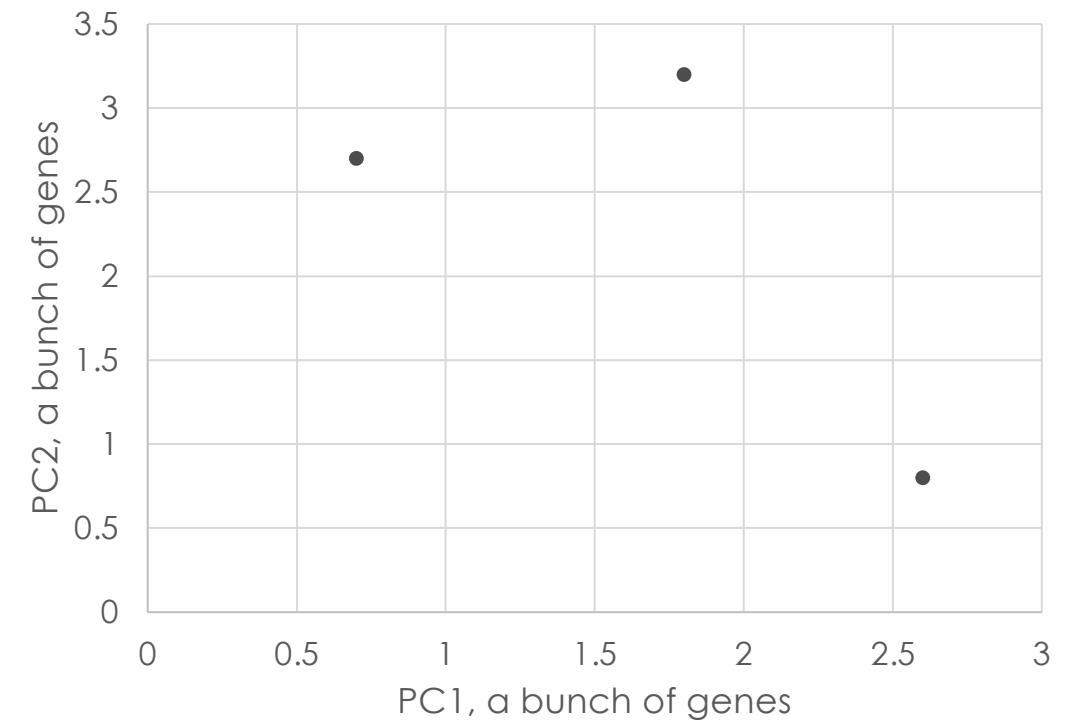
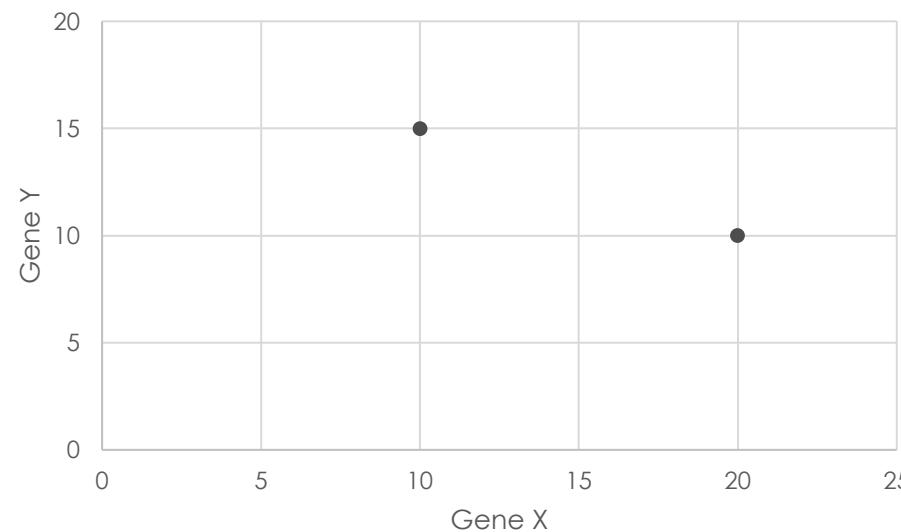
# Scatter plot for replicates



# Principal component analysis (PCA)

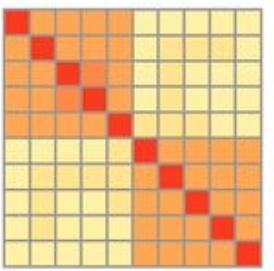


Gene	Sample1	Sample2
X	10	20
Y	15	10

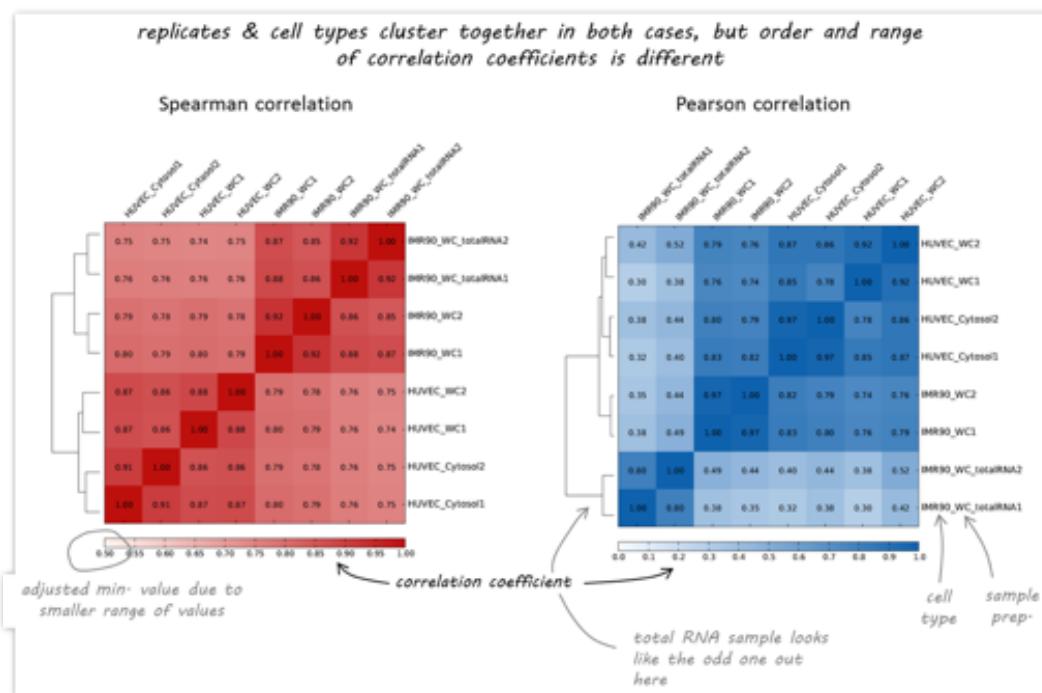


PCA reduces the complexity of expression data to show relationships on two axes

Sample-level

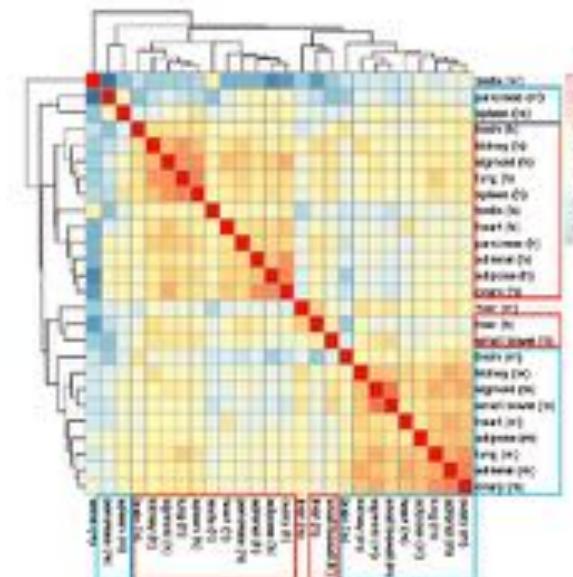


# Sample-sample distance plot

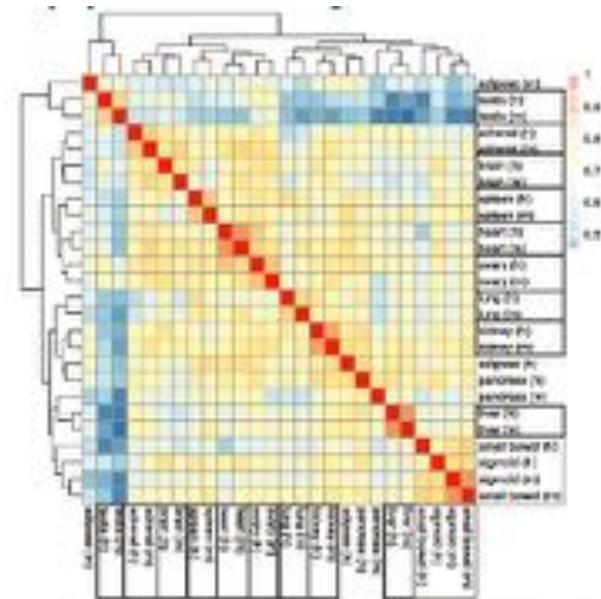


<https://deeptools.readthedocs.io/en/develop/content/tools/plotCorrelation.html>

Sample-level analysis is affected by batch effect



"Overall, our results indicate that there is considerable RNA expression diversity between humans and mice, well beyond what was described previously, likely reflecting the fundamental physiological differences between these two organisms."



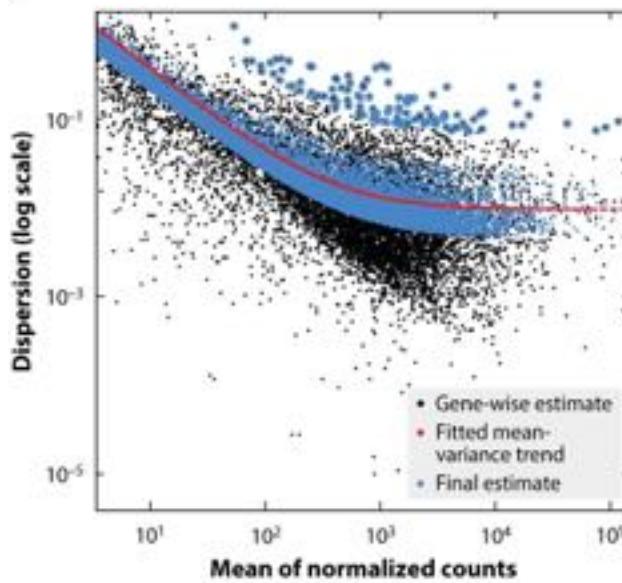
"Once we accounted for the batch effect (...), the comparative gene expression data no longer clustered by species, and instead, we observed a clear tendency for clustering by tissue."

# Summary of suitable analysis approaches for the three types of comparative analyses

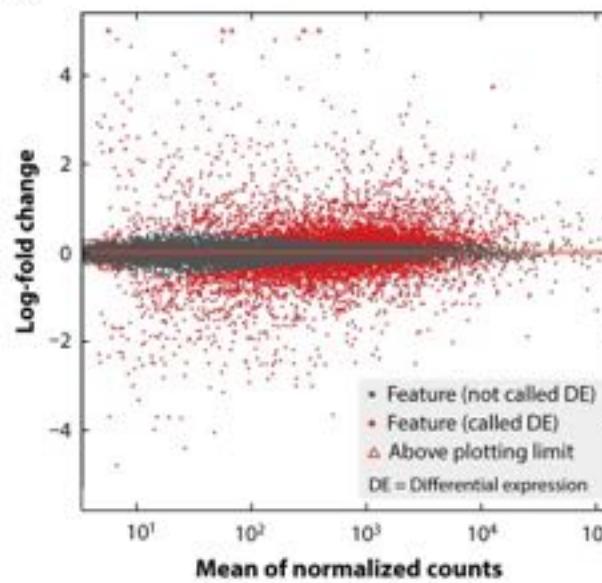
Task	Input data	Software (examples)
differential gene expression (DGE)	Gene counts	DESeq2, edgeR, voom/limma
differential transcript usage (DTU)	Transcript counts	DESeq2, edgeR, sleuth, voom/limma
differential exon usage (DEU)	Exon counts	DEXSeq, voom/limma, MISO, IsoformSwitchAnalyzeR

# Differential Expression

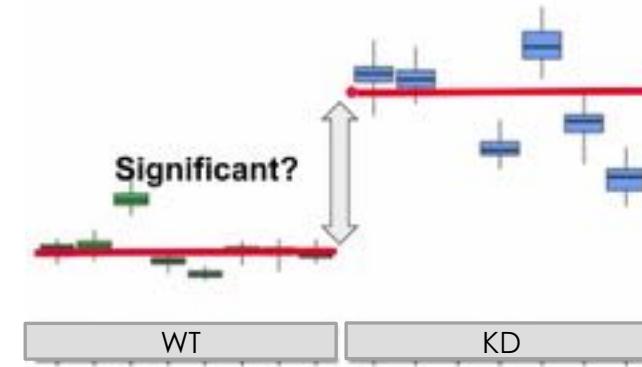
a



b



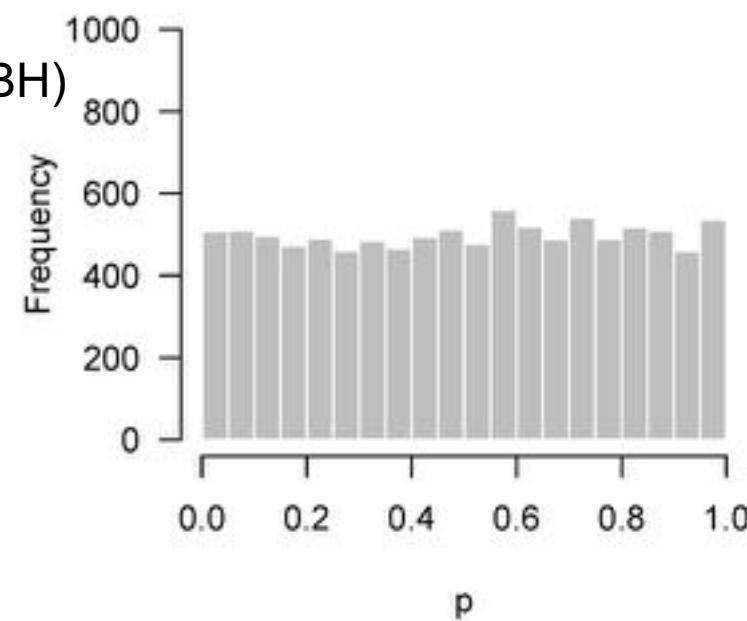
Testing:  
the variation **between** groups >  
the variation **within** groups



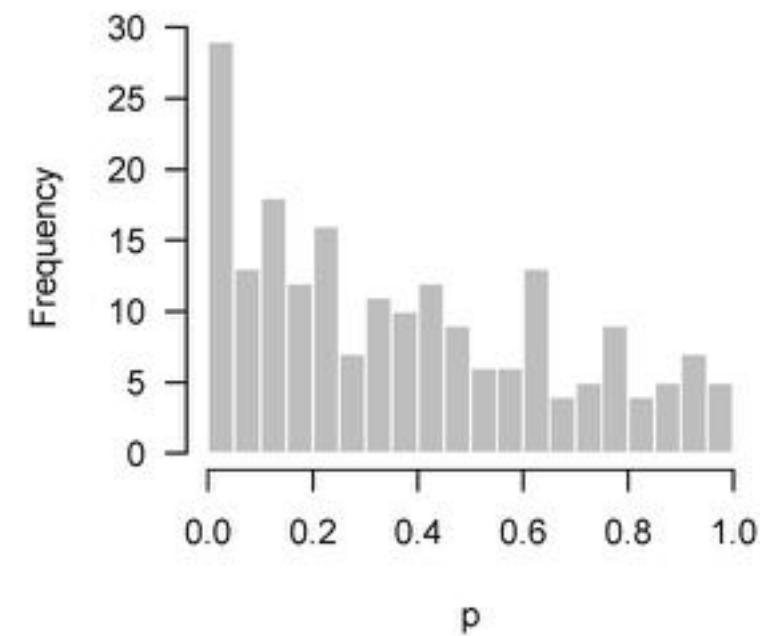
# Multiple comparisons

Benjamini-Hochberg Correction (BH)  
false-discovery rate (FDR)

From a uniform random number generator



From a sufficiently powered experiment



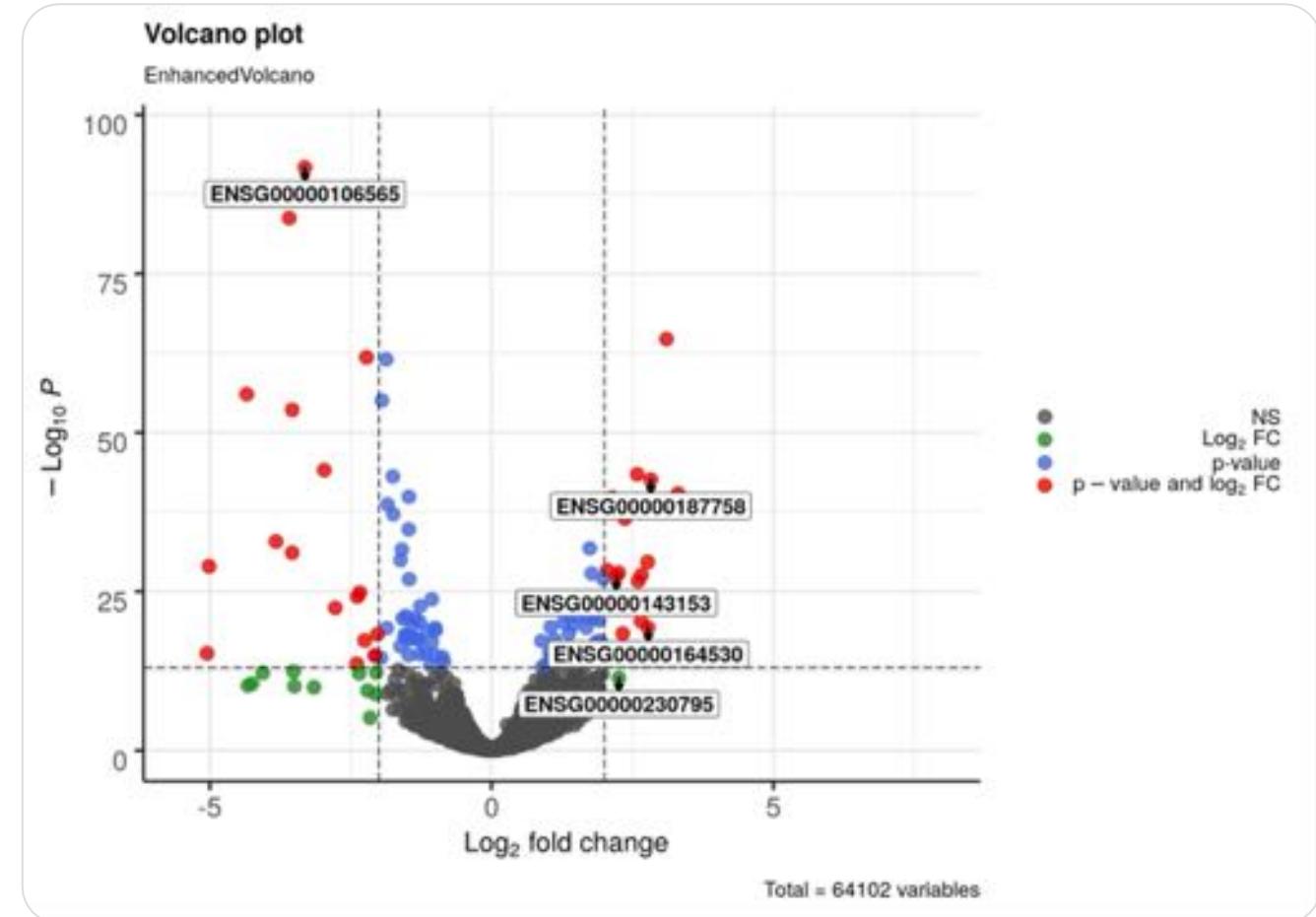
# Results

baseMean	log2FoldChange	IfcSE	stat	pvalue	padj	gene_name	ensembl_id
569.7493	-2.51313	0.352548	-7.12847	1.01E-12	6.32E-12	aacs	ENSDARG00000012468
228.7958	-2.64457	0.241543	-10.9487	6.74E-28	1.41E-26	aamdc	ENSDARG00000103957
3299.677	-1.25607	0.237625	-5.28595	1.25E-07	4.81E-07	aamp	ENSDARG00000045019
151.558	1.228695	0.449503	2.733449	0.006267	0.012898	aarsd1	ENSDARG00000015747
423.9228	-1.38447	0.159835	-8.66182	4.64E-18	4.53E-17	aass	ENSDARG00000051816
2594.372	1.402565	0.149554	9.37834	6.7E-21	8.28E-20	abca1b	ENSDARG00000079009
1215.935	-1.7795	0.223888	-7.94816	1.89E-15	1.52E-14	abcb3l1	ENSDARG00000036787
208.7542	-2.28293	0.19053	-11.982	4.42E-33	1.35E-31	abcb6a	ENSDARG00000063297
206.3376	-1.2775	0.346485	-3.68701	0.000227	0.000587	abcb9	ENSDARG00000056200
223.5531	2.194671	0.430981	5.092271	3.54E-07	1.29E-06	abcc2	ENSDARG00000014031

# Visualization of DEGs



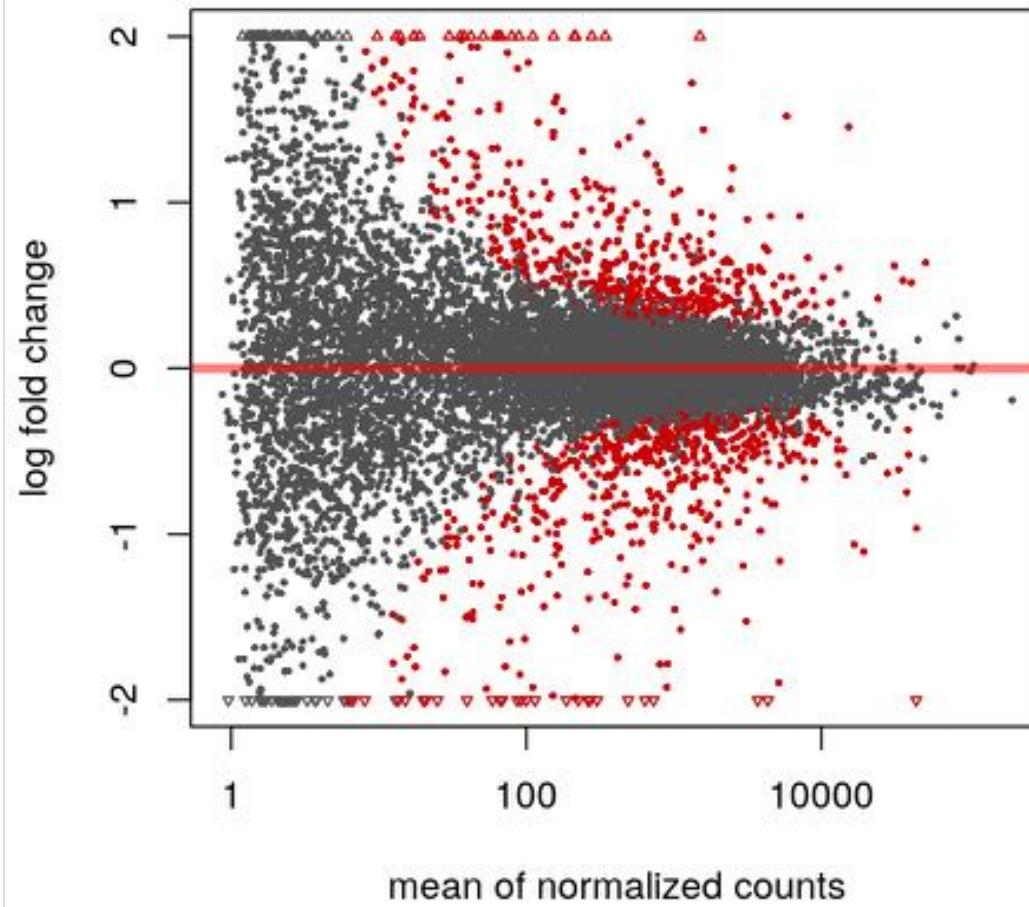
# Volcano plot



<https://bioconductor.org/packages/release/bioc/vignettes/EnhancedVolcano/inst/doc/EnhancedVolcano.html>

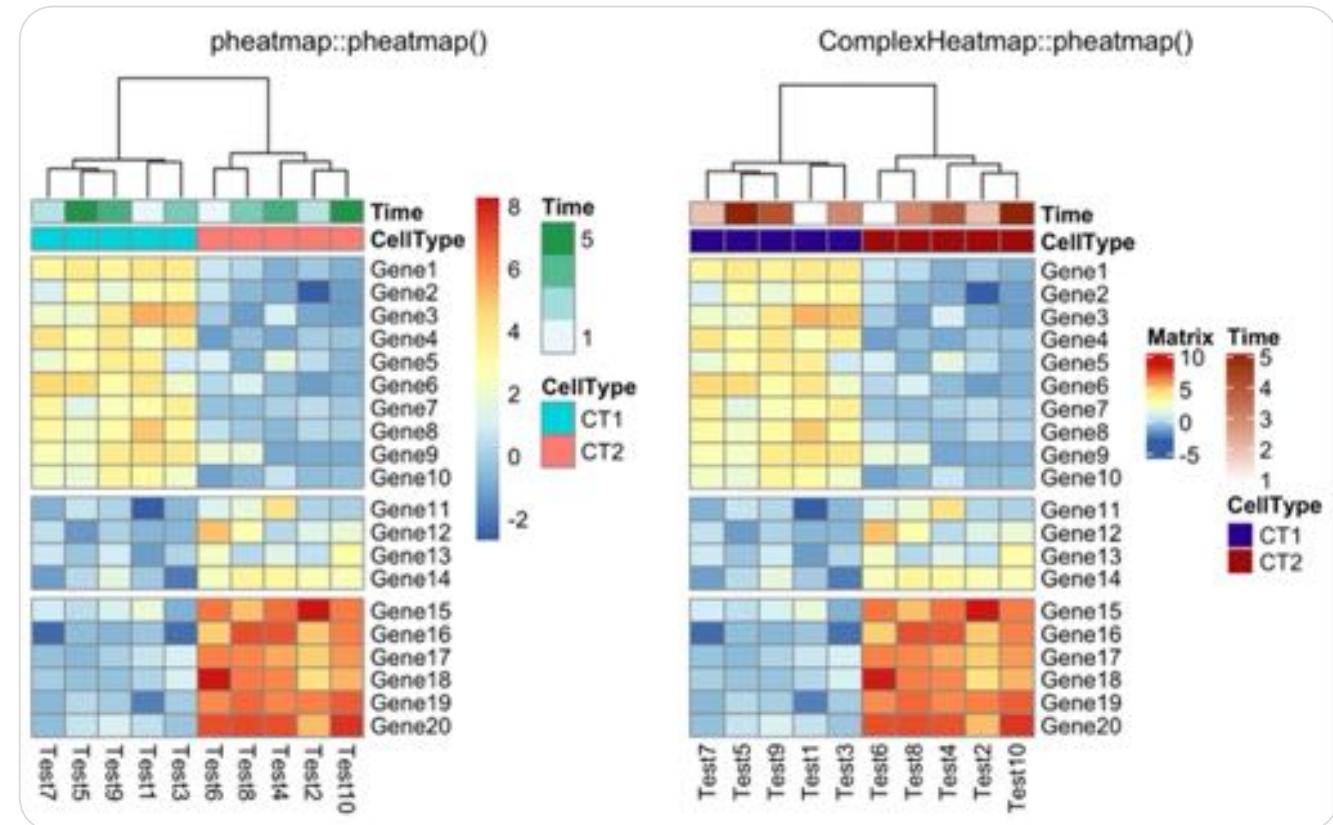
# MA plot

M (log ratio) and A (mean average) scales



# Heatmap

- Heatmap is a graphical reorientation of 3D data in 2D figure. It uses a system of color-coding to represent different values.
- Packages: pheatmap, ggplot2, ComplexHeatmap ...



<https://CRAN.R-project.org/package=pheatmap>

<https://bioconductor.org/packages/ComplexHeatmap/>

<https://jokergoo.github.io/2020/05/06/translate-from-pheatmap-to-complexheatmap/>

# Functional annotation of DE genes

Gene ontology  
(GO) enrichment  
analysis

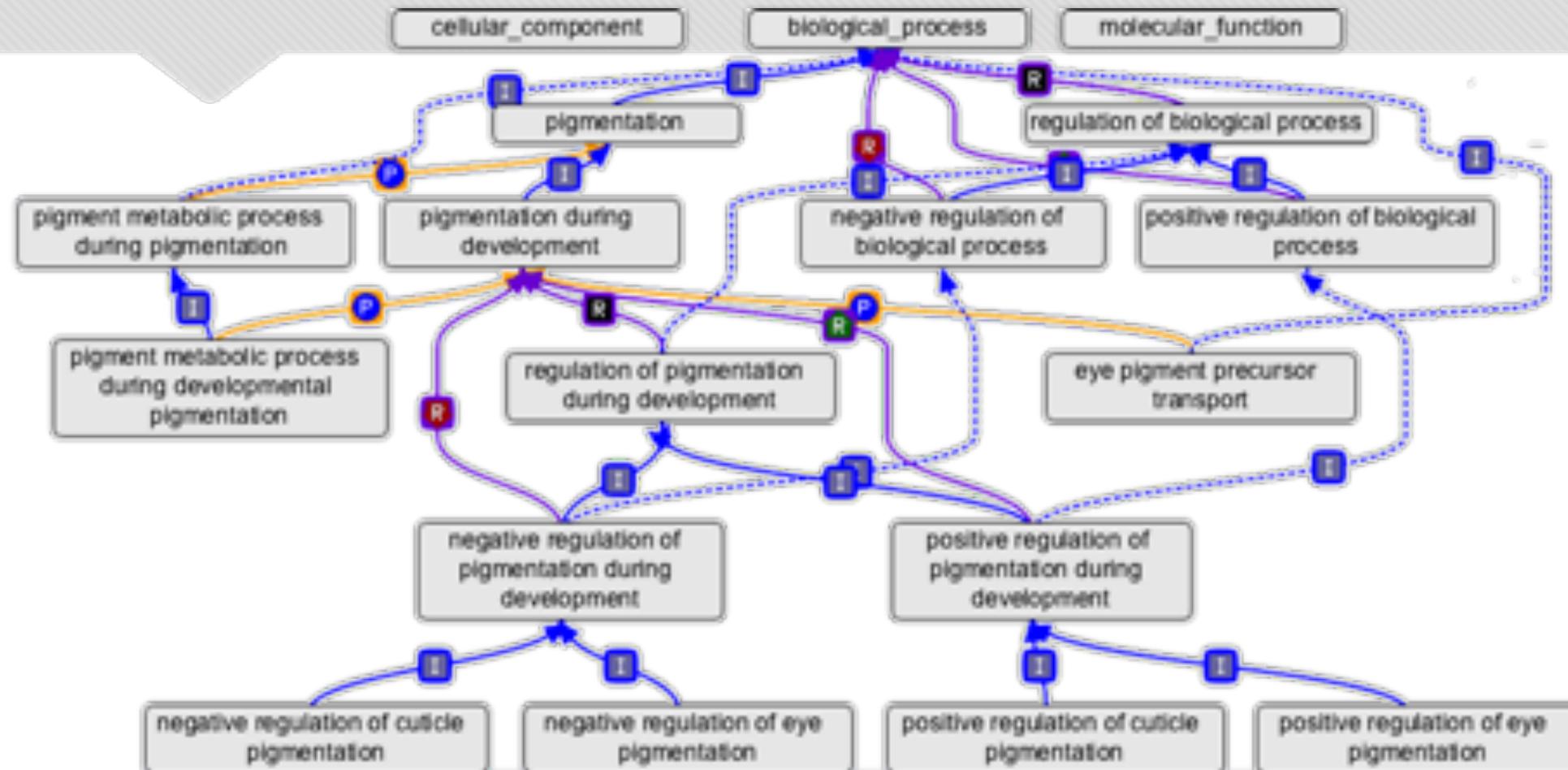
Kyoto Encyclopedia  
of Genes and  
Genomes (KEGG)  
pathway  
enrichment analysis

Gene set  
enrichment analysis  
(GSEA)

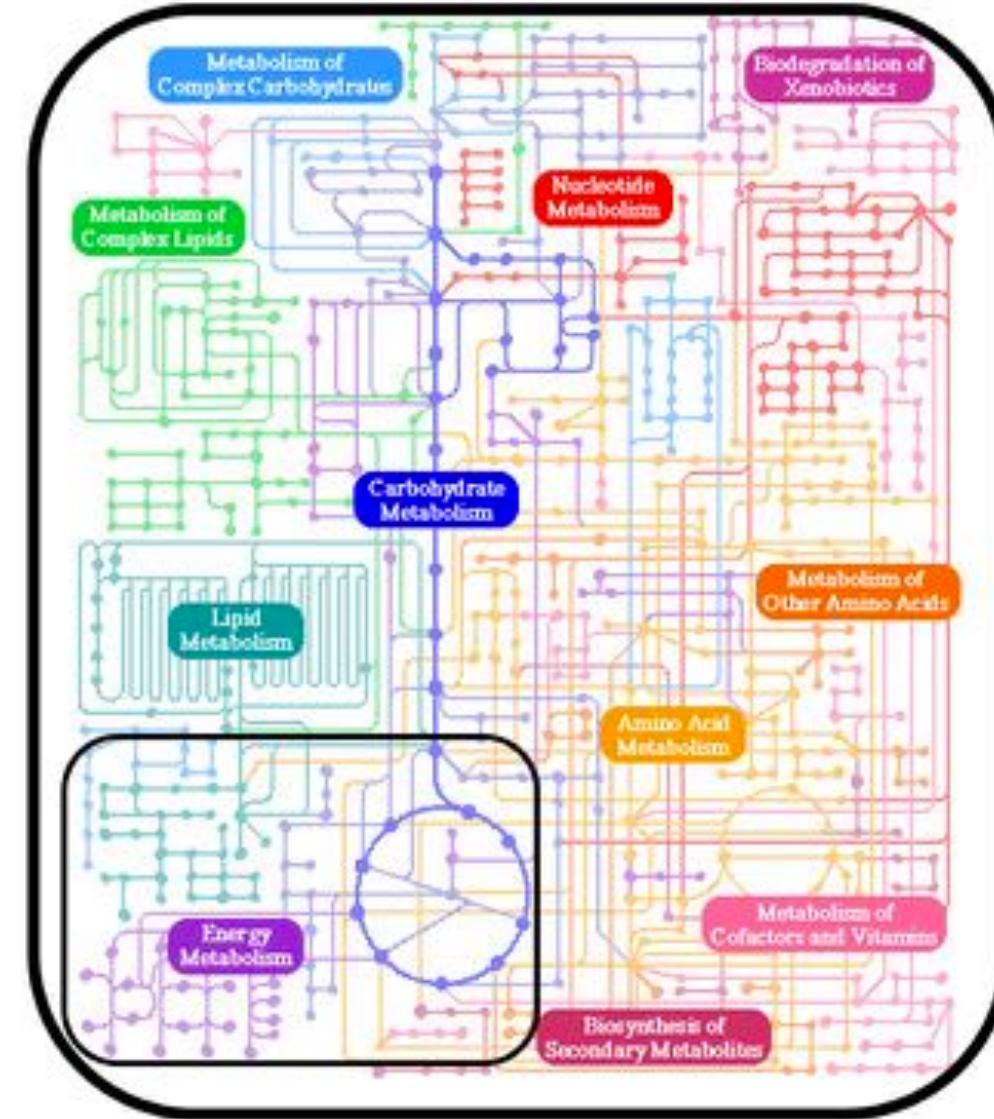
QIAGEN Ingenuity  
Pathway Analysis  
(IPA)

A gene set over-represented comparing to random sampling

# Gene ontology

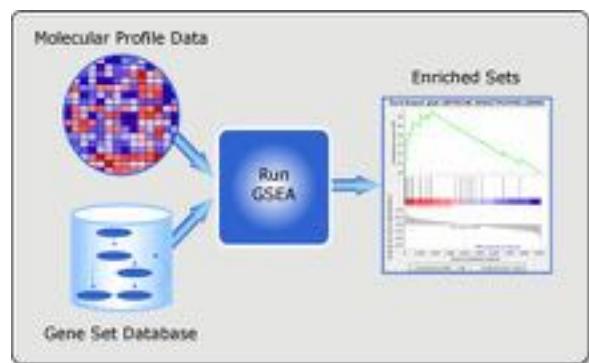


# KEGG pathway enrichment analysis



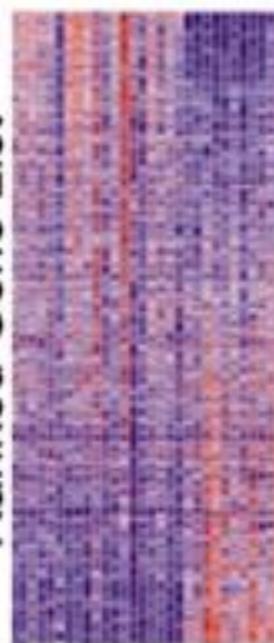
<https://www.genome.jp/kegg/pathway.html>

# GSEA



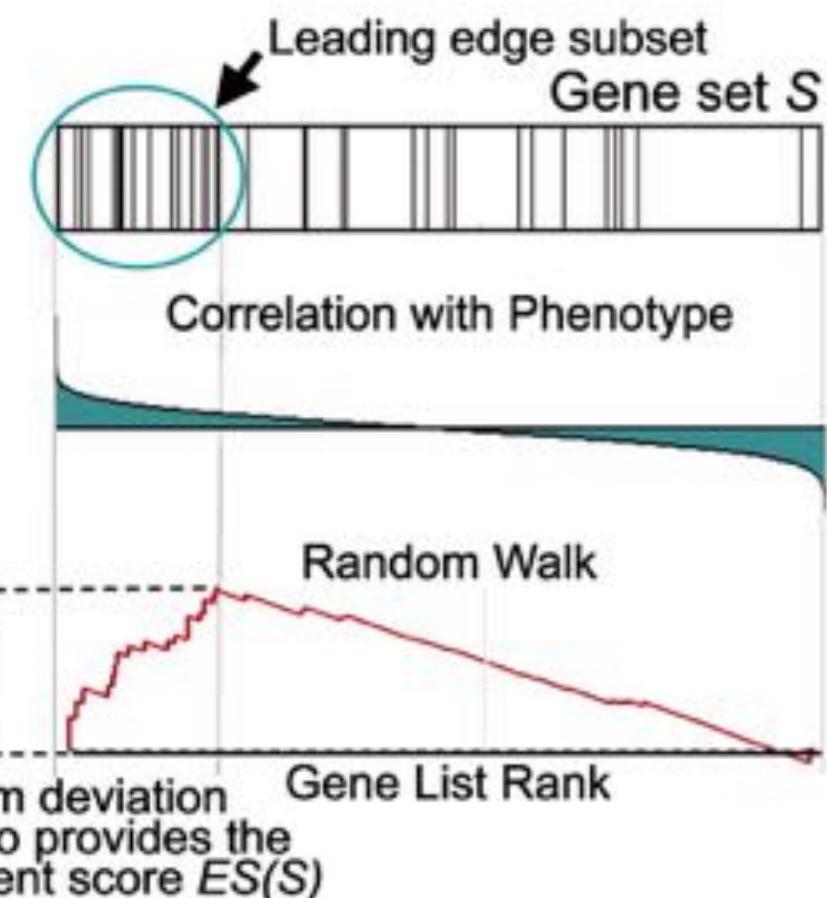
A Phenotype Classes  
A B

Ranked Gene List

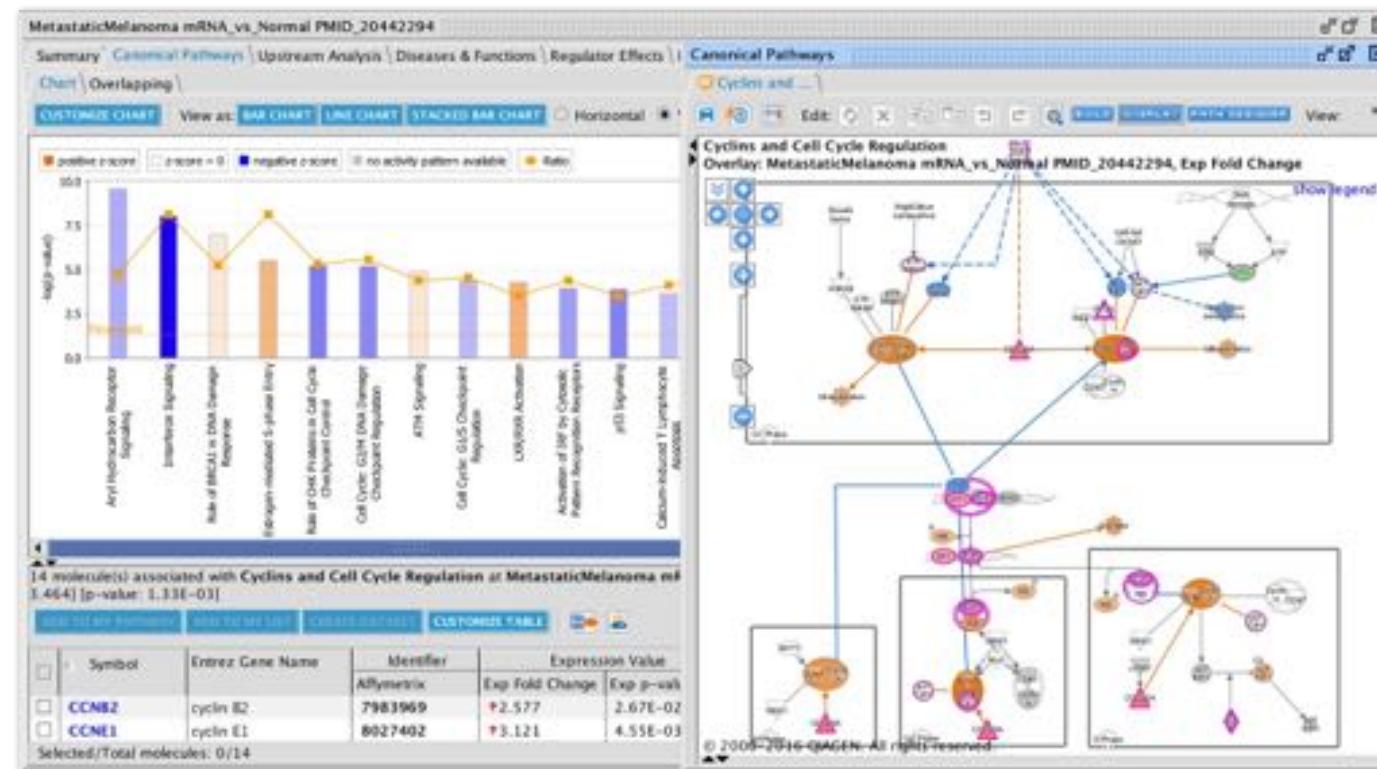


B

Gene set S

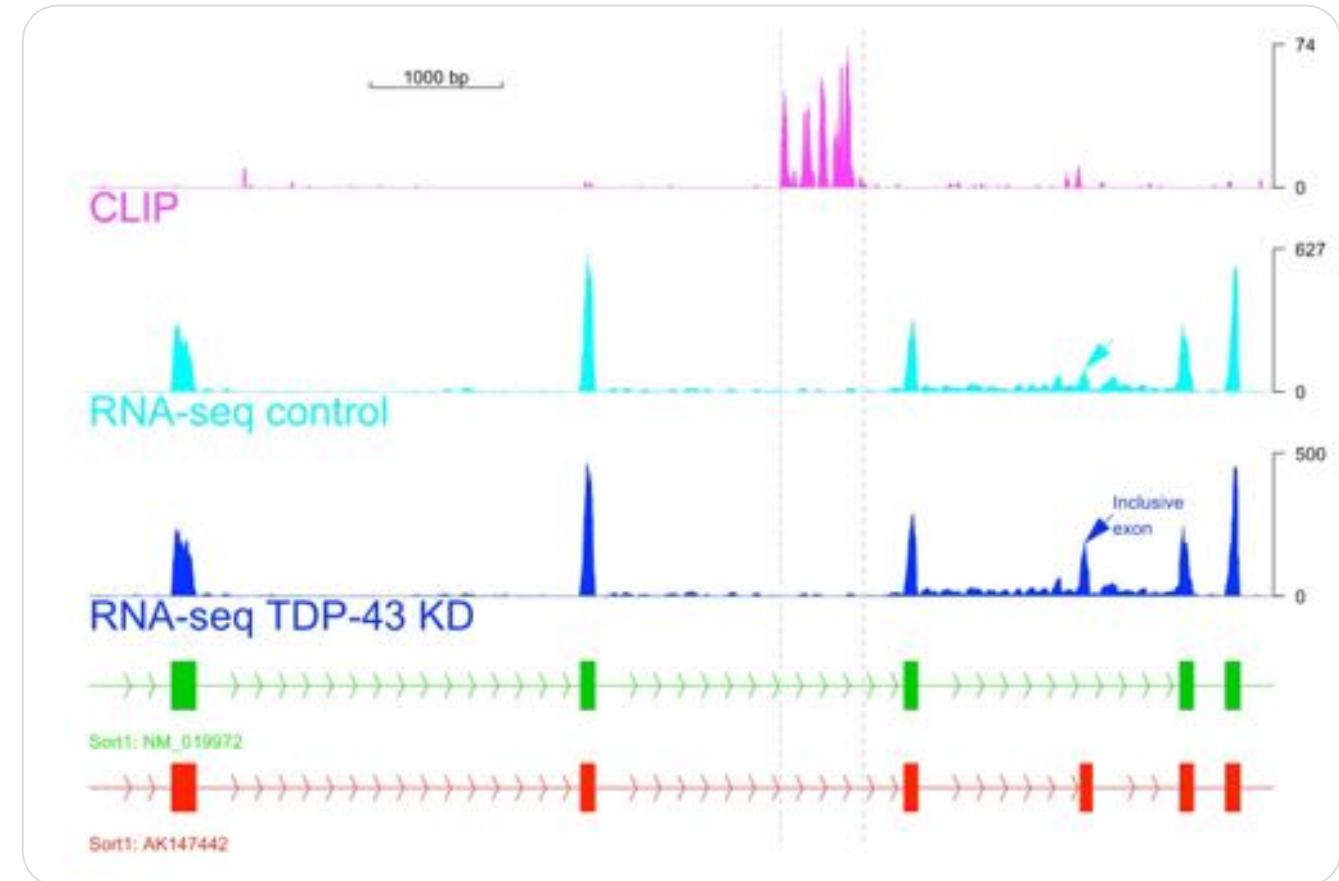


# IPA



<https://digitalinsights.qiagen.com/products/features/>

# View tracks by trackViewer



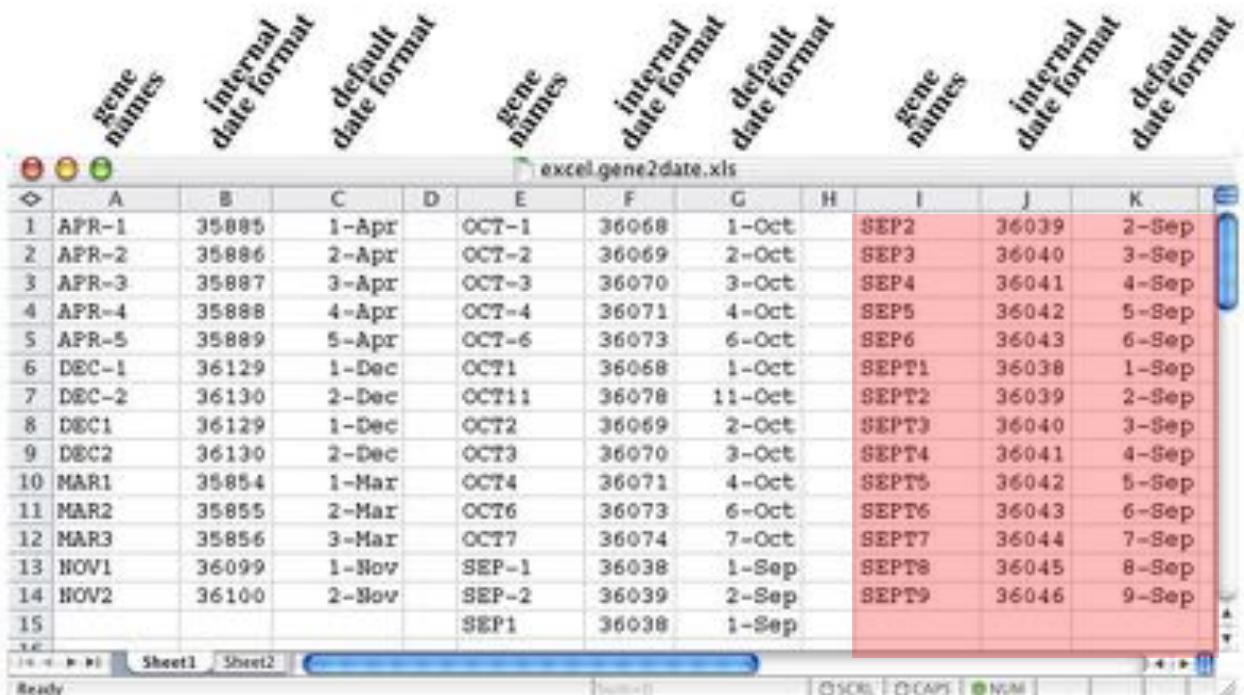
Ou et.al., 2019. doi: 10.1038/s41592-019-0430-y

# SAVE RESULTS

The results can be saved in XLS file format using [WriteXLS](#) package to avoid the gene name errors that can be inadvertently introduced when opened by Excel.

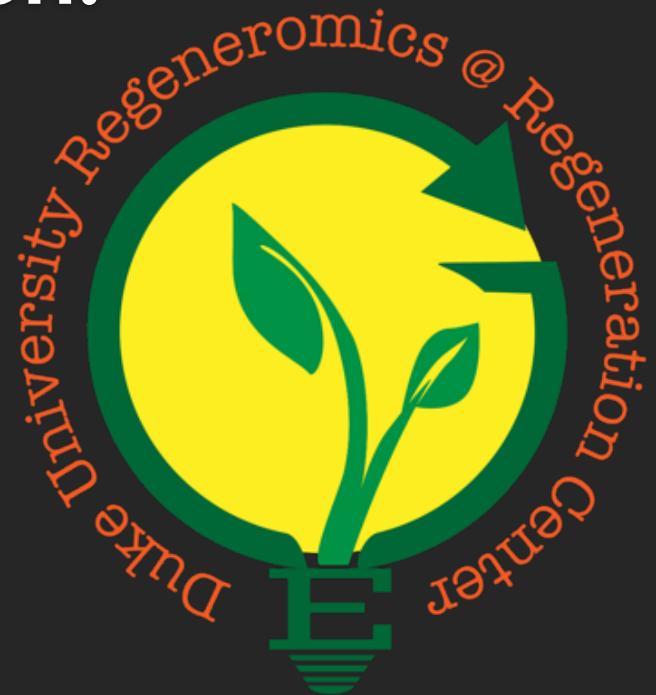


```
library(WriteXLS)  
WriteXLS(a_data_frame, "output.xls")
```



	A	B	C	D	E	F	G	H	I	J	K
1	APR-1	35885	1-Apr	OCT-1	36068	1-Oct	SEP2	36039	2-Sep		
2	APR-2	35886	2-Apr	OCT-2	36069	2-Oct	SEP3	36040	3-Sep		
3	APR-3	35887	3-Apr	OCT-3	36070	3-Oct	SEP4	36041	4-Sep		
4	APR-4	35888	4-Apr	OCT-4	36071	4-Oct	SEP5	36042	5-Sep		
5	APR-5	35889	5-Apr	OCT-6	36073	6-Oct	SEP6	36043	6-Sep		
6	DEC-1	36129	1-Dec	OCT1	36068	1-Oct	SEPT1	36038	1-Sep		
7	DEC-2	36130	2-Dec	OCT11	36078	11-Oct	SEPT2	36039	2-Sep		
8	DEC1	36129	1-Dec	OCT2	36069	2-Oct	SEPT3	36040	3-Sep		
9	DEC2	36130	2-Dec	OCT3	36070	3-Oct	SEPT4	36041	4-Sep		
10	MAR1	35854	1-Mar	OCT4	36071	4-Oct	SEPT5	36042	5-Sep		
11	MAR2	35855	2-Mar	OCT6	36073	6-Oct	SEPT6	36043	6-Sep		
12	MAR3	35856	3-Mar	OCT7	36074	7-Oct	SEPT7	36044	7-Sep		
13	HOV1	36099	1-Nov	SEP-1	36038	1-Sep	SEPT8	36045	8-Sep		
14	HOV2	36100	2-Nov	SEP-2	36039	2-Sep	SEPT9	36046	9-Sep		
15				SEP1	36038	1-Sep					

# Regeneromics Shared Resource can help your research!



We do: experimental design, bioinformatics analysis, manuscript preparation, grant applications



Jianhong Ou, Ph.D.

Email: regeneromics@duke.edu

<https://sites.duke.edu/dukeregenerationcenter/jobsrni/>

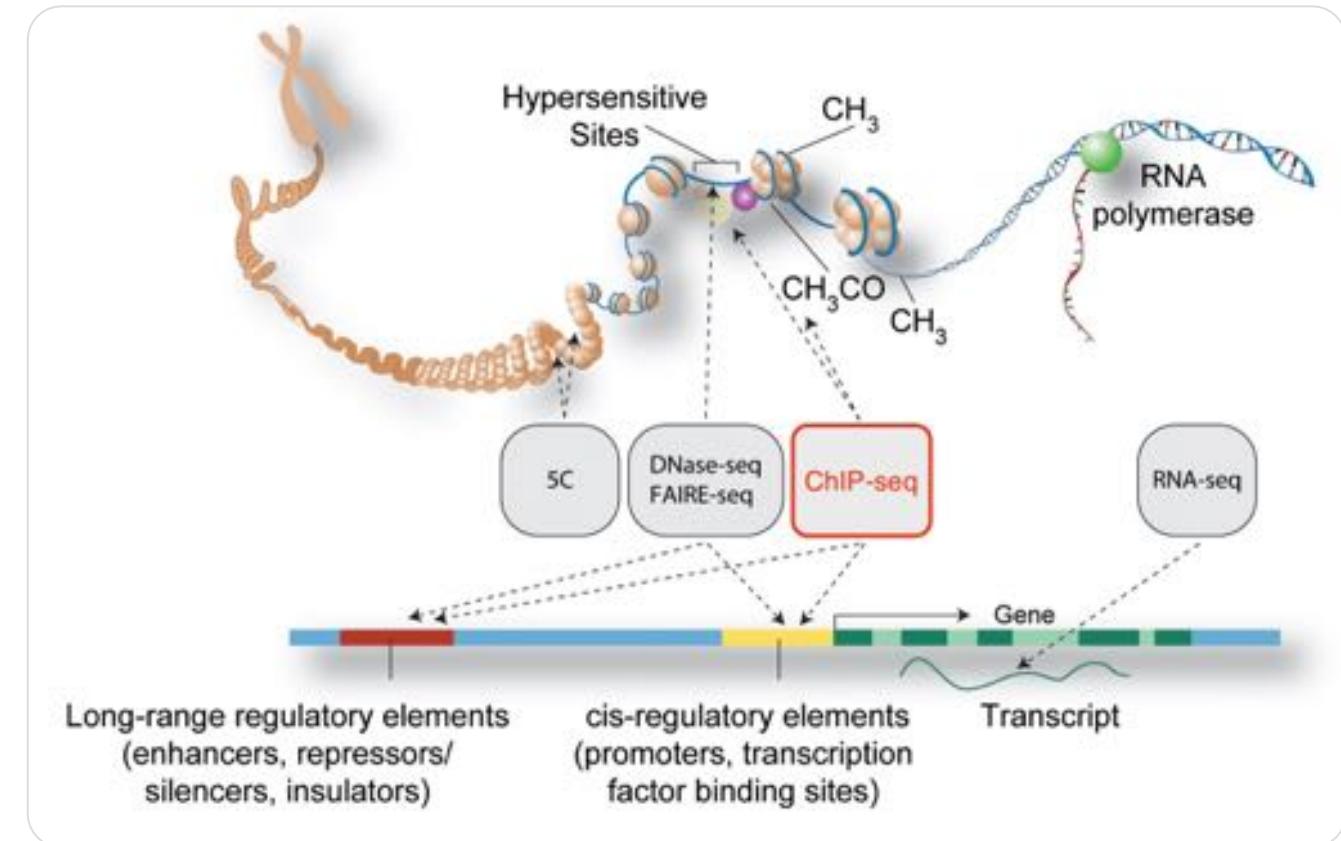
# CHIP-SEQ

<https://jianhong.github.io/genomictools/index.html>

# ChIP-seq technology

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is a technique for genome-wide profiling of DNA-binding proteins, histone modifications or nucleosomes. – Peter J. Park

Park, 2009. doi:10.1038/nrg2641



The ENCODE Project Consortium 2011. doi: 10.1371/journal.pbio.1001046

# Uses

- Determine the DNA targets for transcription factors (TFs) in genome-wide
- Capture the histone modifications across the entire genome
- Define the TF binding motifs
- Reveal gene regulatory networks in combination with RNA sequencing and methylation analysis

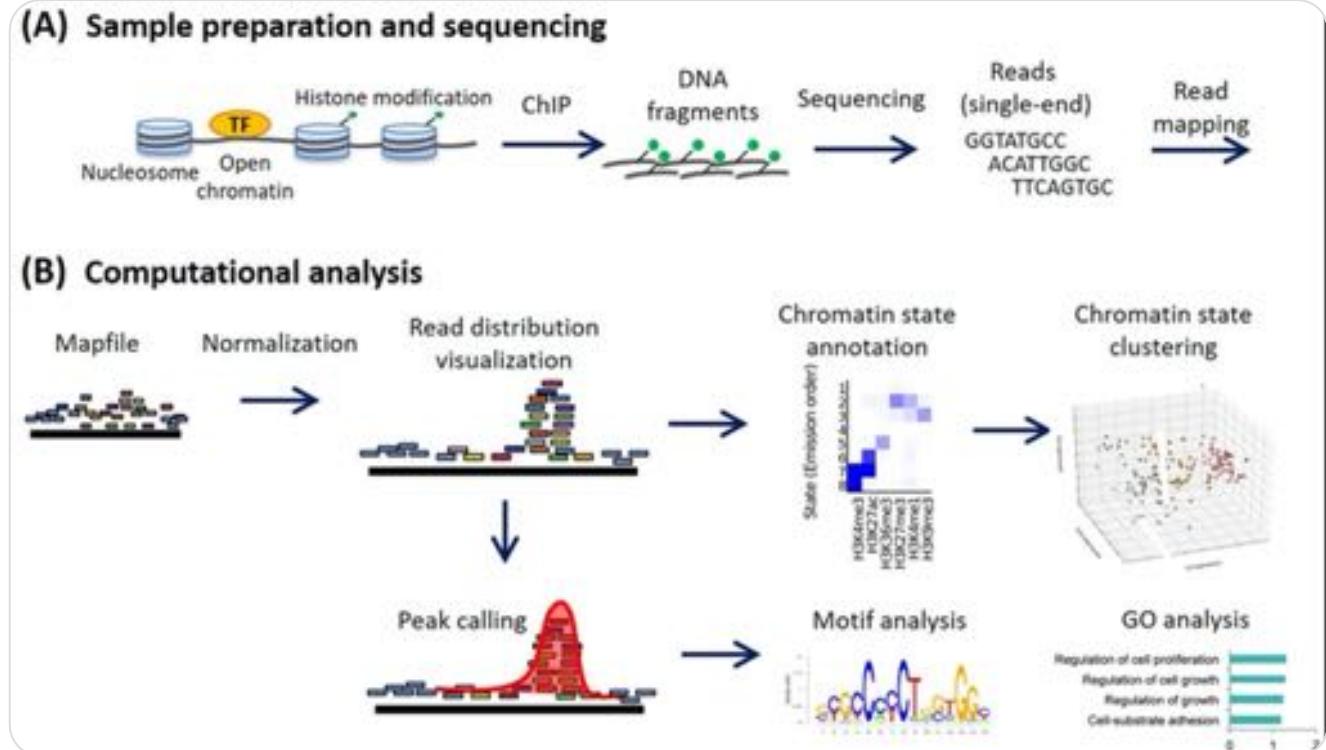
<https://www.illumina.com/techniques/sequencing/dna-sequencing/chip-seq.html>

# ENCODE and Epigenomics Roadmap

The collage consists of three main panels:

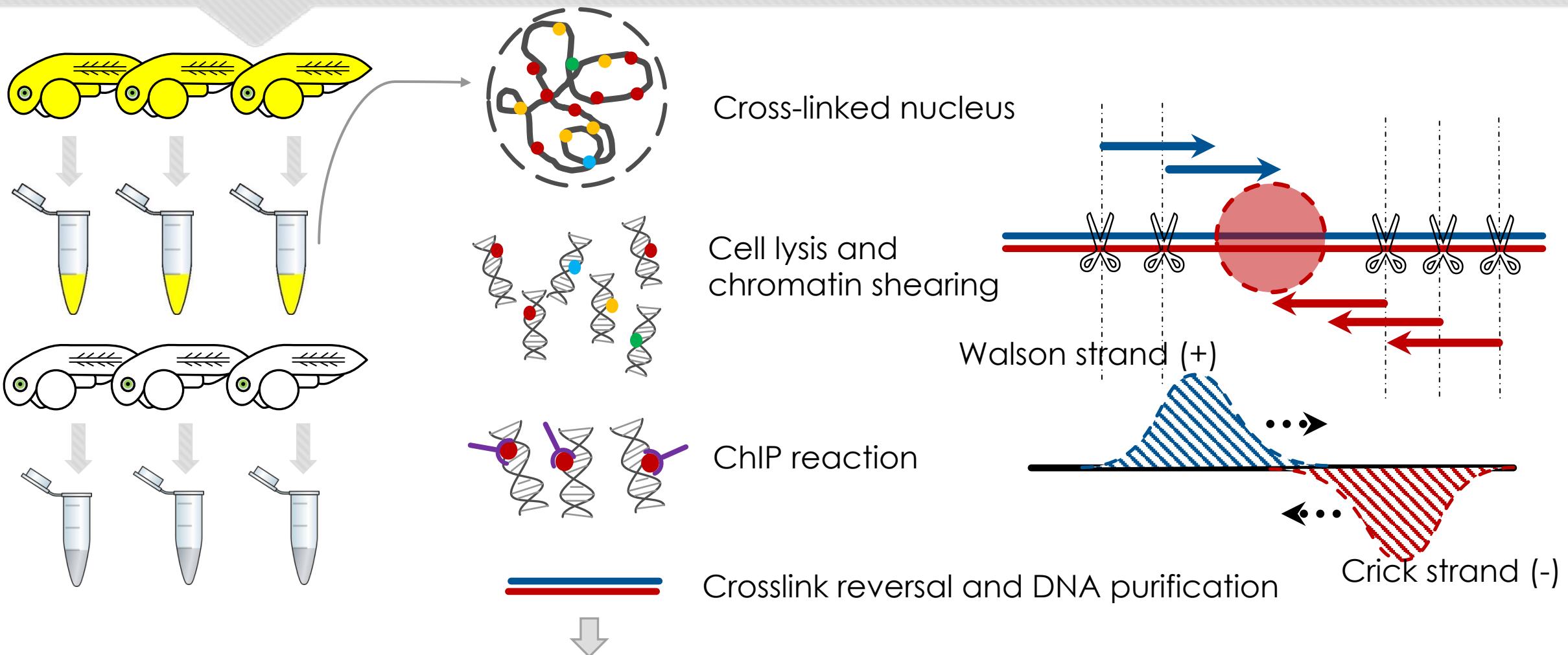
- Left Panel (nature magazine cover):** The front cover of the journal 'nature' from September 2012. It features a large circular graphic titled 'ENCODE' in the center, with the text 'GUIDEBOOK TO THE HUMAN GENOME' and 'The ENCODE project in print and online'. Below the main title, there are several news stories: 'LAST RAYS OF THE SUN' (Planetary Science), 'HARNESSING FOSSIL POWER' (Palaeontology), and 'RETHINK ON RISK DATA' (Toxicology). The ENCODE logo is prominently displayed in the center of the circle.
- Middle Panel (ENCODE and NIH Roadmap websites):** A composite image showing two screenshots of scientific databases.
  - ENCODE Portal:** A screenshot of the ENCODE Encyclopedia of DNA Elements portal. It shows a diagram of epigenetic marks (H3K4me1, H3K27me3, H3K36me3, H3K9me3, H3K27ac, H3K4me3, H3K9me2, H3K36me2) and their relationship to transcription factors, chromatin, and genes. Below the diagram are search and browse functions for the ENCODE project, including links for 'About ENCODE Project', 'Getting Started', 'Experiments', 'Search ENCODE portal', and 'Search for candidate Cis-Regulatory Elements'.
  - NIH Roadmap Epigenomics Mapping Consortium:** A screenshot of the NIH Roadmap Epigenomics Mapping Consortium website. It features a diagram of a cell nucleus with chromatin, RNA, and genes. Below the diagram is a section titled 'NIH Roadmap Epigenomics Mapping Consortium' with a brief description of the project's goals and methods.
- Right Panel (nature magazine cover):** The back cover of the journal 'nature' from February 2013. It features a colorful illustration of a futuristic city with flying cars, a hot air balloon, and a bridge. The title 'nature' is at the top, followed by 'THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE'. The main headline is 'EPIGENOME ROADMAP' with the subtitle 'Functional regulatory elements in genomes from human tissues'. Other news stories include 'SPECTRUM OF CONFUSION' (Gender), 'CREATIVE ACCOUNTING' (Finance), and 'LITHIUM IN THE STARS' (Cosmology).

# ChIP-seq analysis workflow



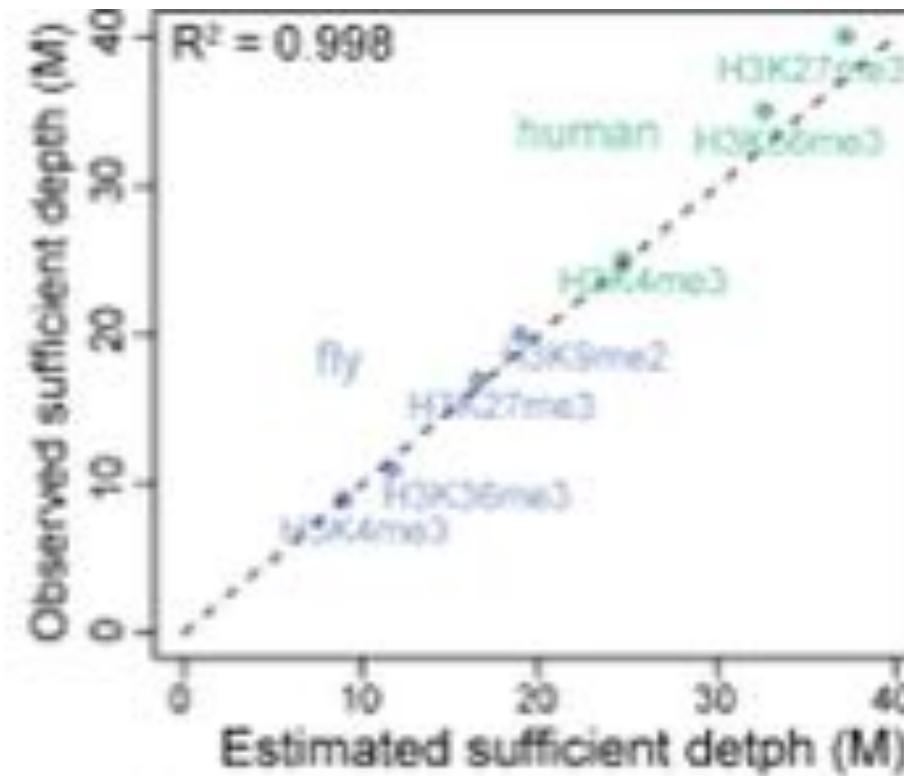
Nakato et.al., 2020. doi: 10.1016/j.ymeth.2020.03.005

# ChIP-seq protocol



# Sequencing Depth

- Sharp peaks (TFs)
  - > 10M reads human/mouse genomes
- Broad Peaks (Histones)
  - > 20M reads human/mouse genomes



# Single end or paired end

- Paired end (PE) sequencing
  - improved efficiency of alignment to repetitive regions
  - Greater ability to detect fragment sizes
- Single-end (SE) sequencing is cheaper than PE.

# Control sample

## Input DNA

- Controls for CNVs, sequencing biases, fragmentation and shearing biases

## IgG

- As with input but also controls for non-specific binding

# Replicates

- Biological replicate experiments are necessary.
- Use different antibodies are recommended.

# Short reads alignment for ChIP-seq



Quality Assessment of sequencing:  
FastQC

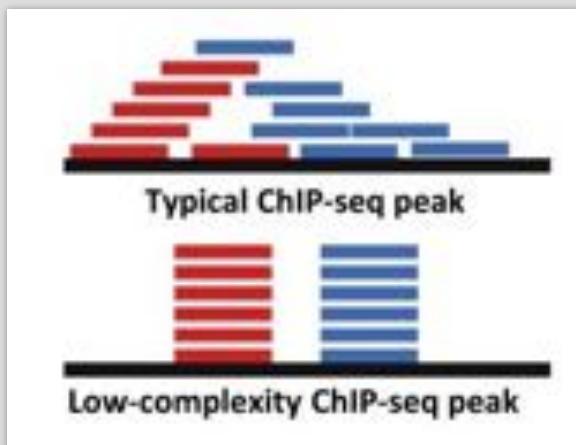


Mapping reads to reference genome:  
bowtie/2, bwa, Rsubread, STAR



Remove potential PCR duplicates:  
Picard software suite::MarkDuplicates,  
Samtools::rmdup

# Library complexity (Duplicates)



Landt et.al., 2012. doi: 10.1101/gr.136184.111



Duplication rates are a useful QC metric

(Duplicate reads/Total Mapped Reads) \*100  
Expected to be low



Non-Redundant Fraction (NRF)

ENCODE guidelines: NRF  
>= 0.8 for 10M reads

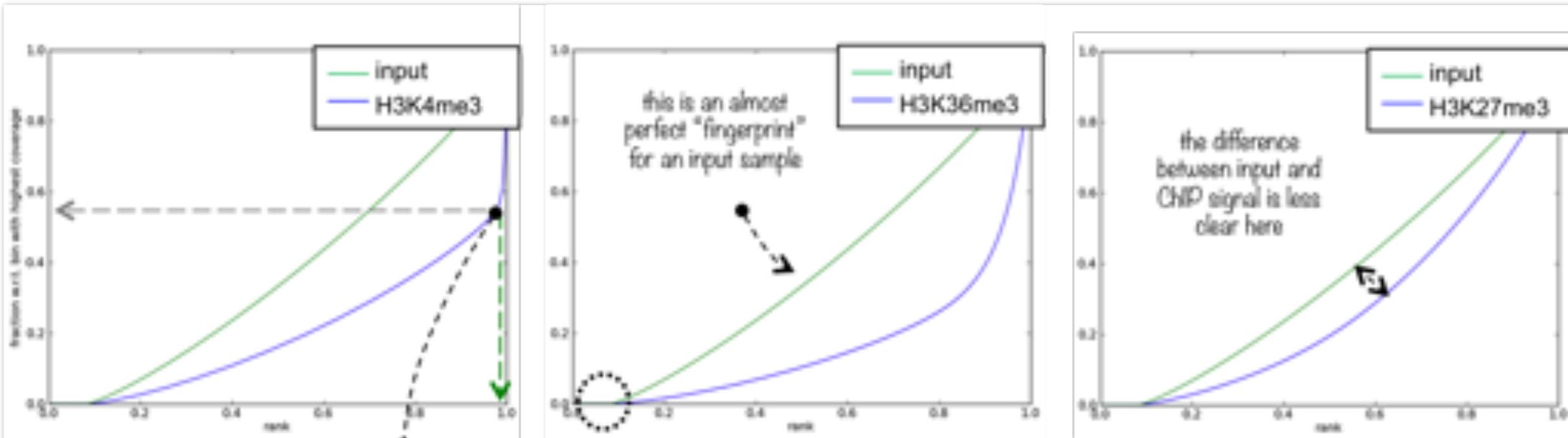


PCR Bottleneck Coefficient (PBC)

A measure of library complexity. High-quality ChIP-seq data set should have PBC values >= 0.9

<https://genome.ucsc.edu/ENCODE/qualityMetrics.html>

# CUMULATIVE CURVE



when counting the reads contained in 97% of all genomic bins, only ca. 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

→ this indicates very localized, very strong enrichments!  
(as every biologist hopes for in a ChIP for H3K4me3)

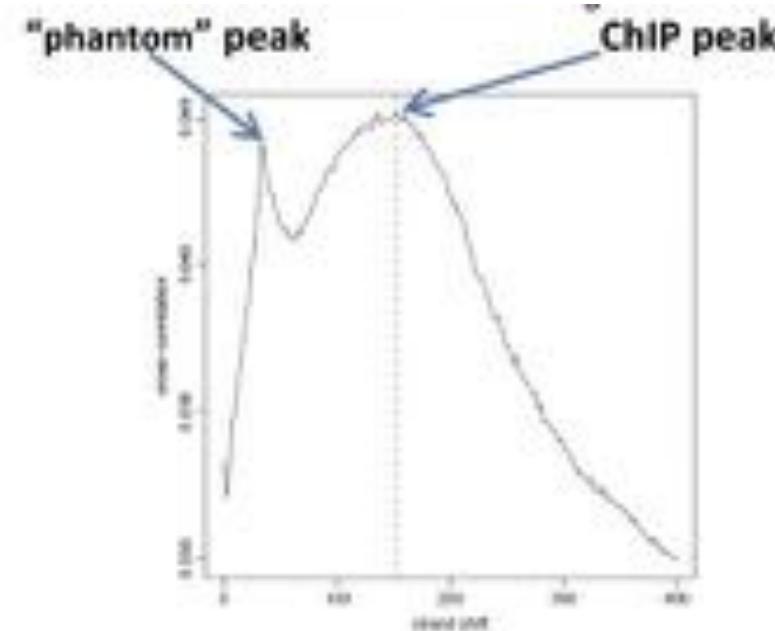
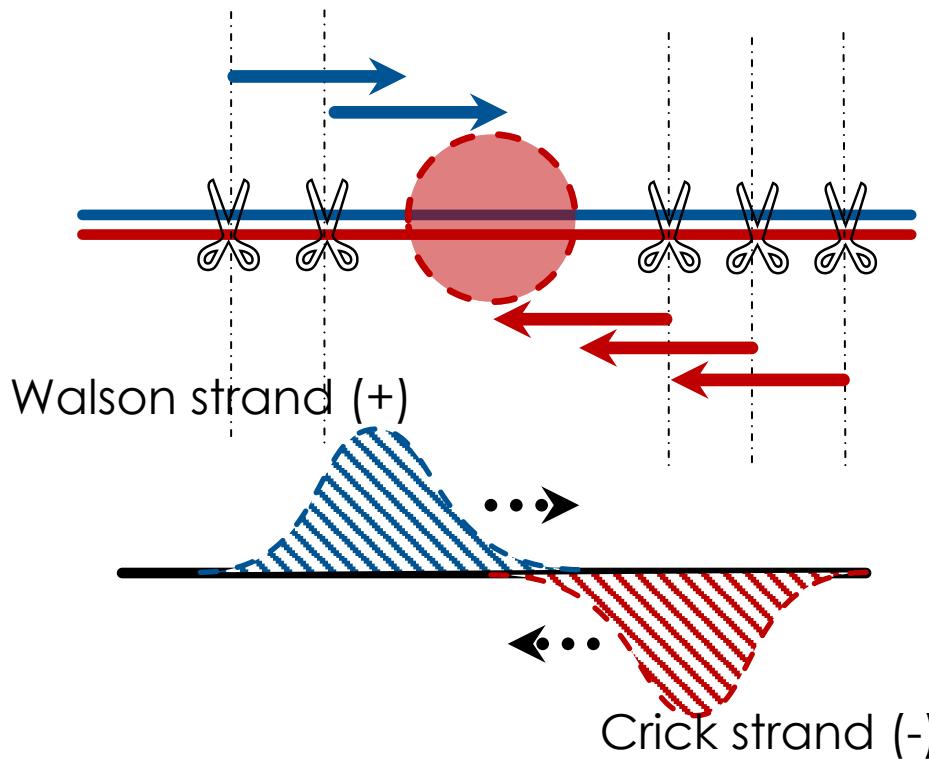
pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any read)

H3K27me3 is a mark that yields broad domains instead of narrow peaks



it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

# Cross-correlation



Strand cross-correlation is computed as the Pearson correlation between the positive and the negative strand profiles at different strand shift distances

# Quality Control

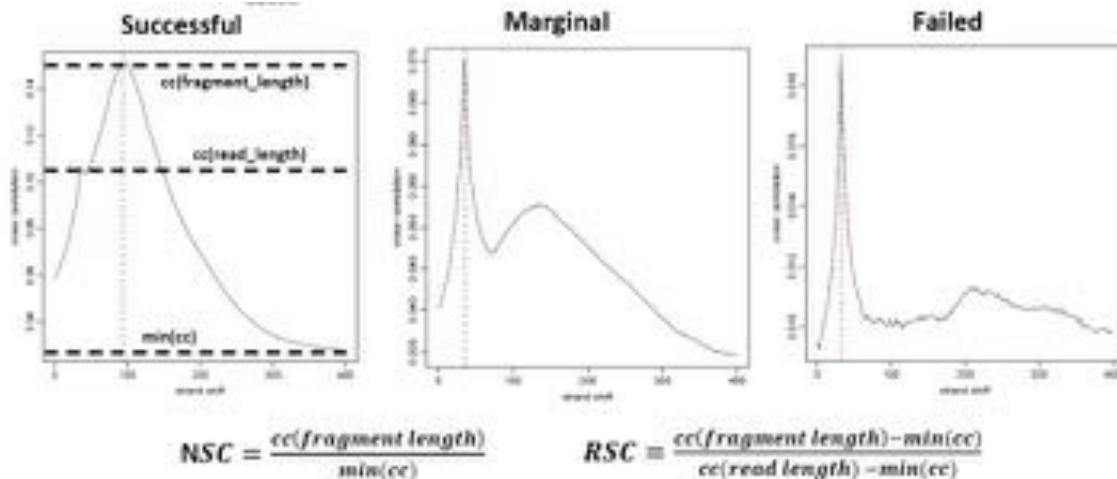


Normalized Strand Cross-correlation coefficient (NSC)  
High-quality ChIP-seq data sets should have NSC values  $\geq 1.05$ .



Relative Strand Cross-correlation coefficient (RSC)  
High-quality ChIP-seq data sets should have RSC values  $\geq 0.8$

<https://genome.ucsc.edu/ENCODE/qualityMetrics.html>



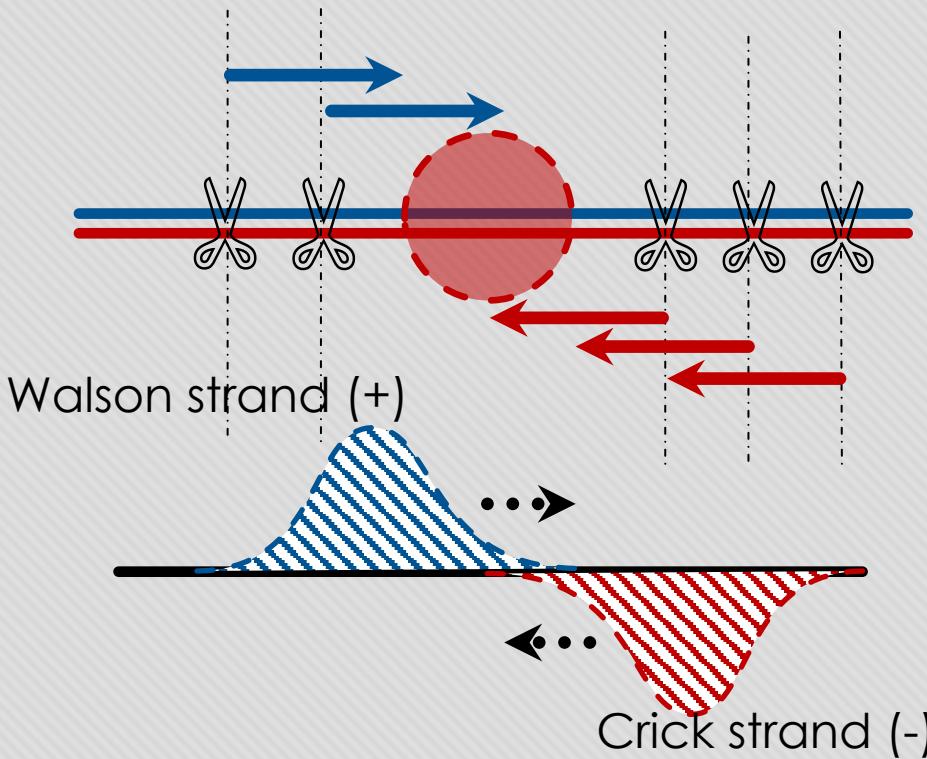
Landt et.al., 2012. doi: 10.1101/gr.136184.111

Calculate NSC and RSC by phantompeaktools

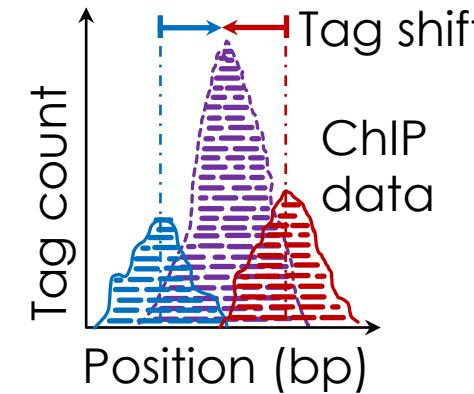
Rscript run\_spp.R -c=<tagAlign/BAMfile> -savp -out=<outFile>

<https://github.com/kundajelab/phantompeakqualtools>

# Call peaks

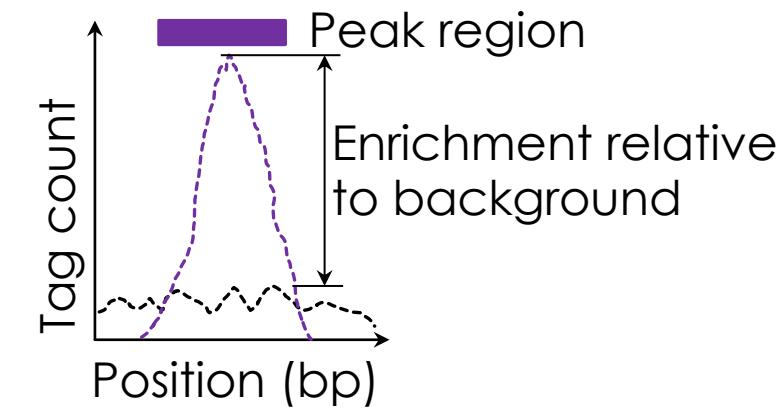
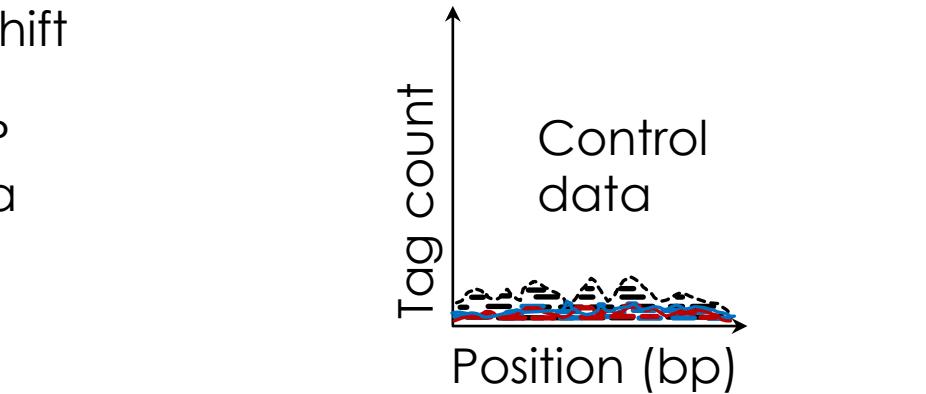


Generate signal profile  
along each chromosome



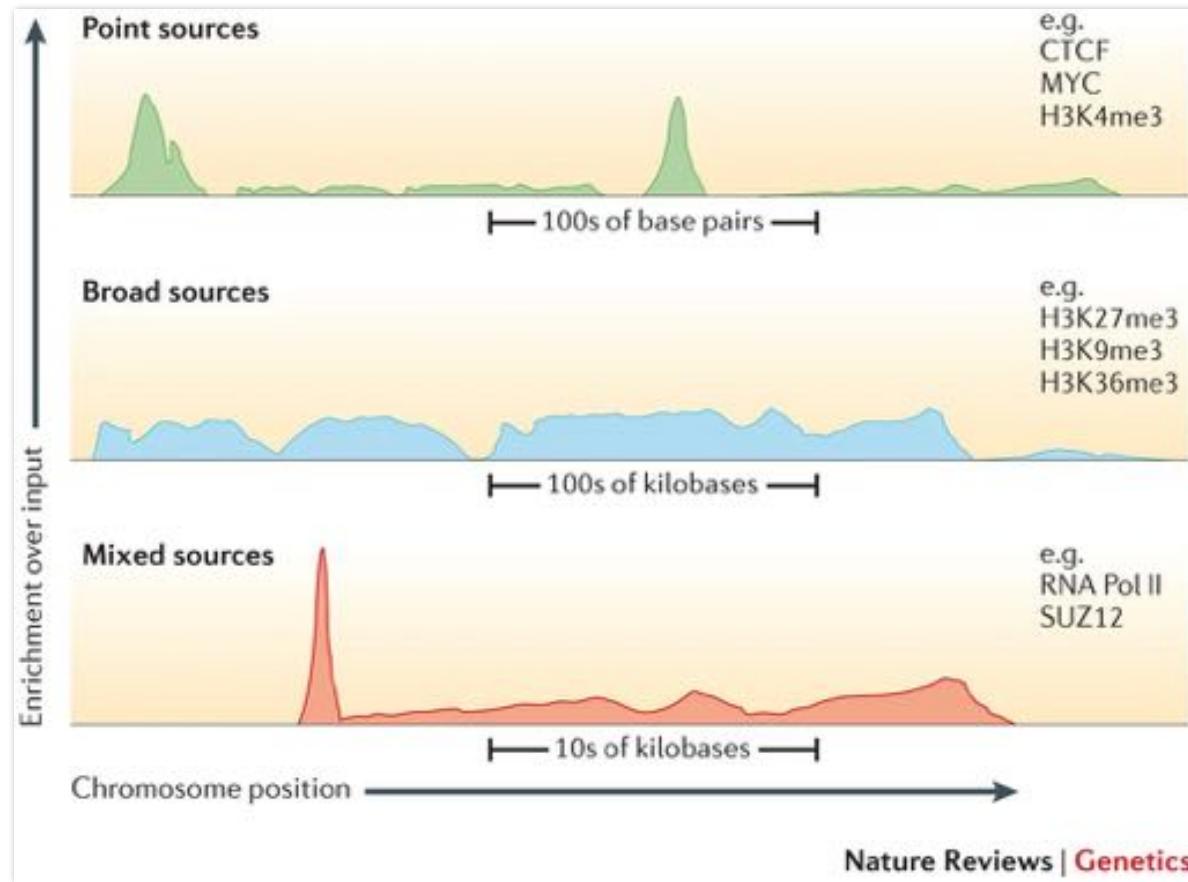
Identify  
peaks in  
ChIP  
signal

Define background  
(model or data)

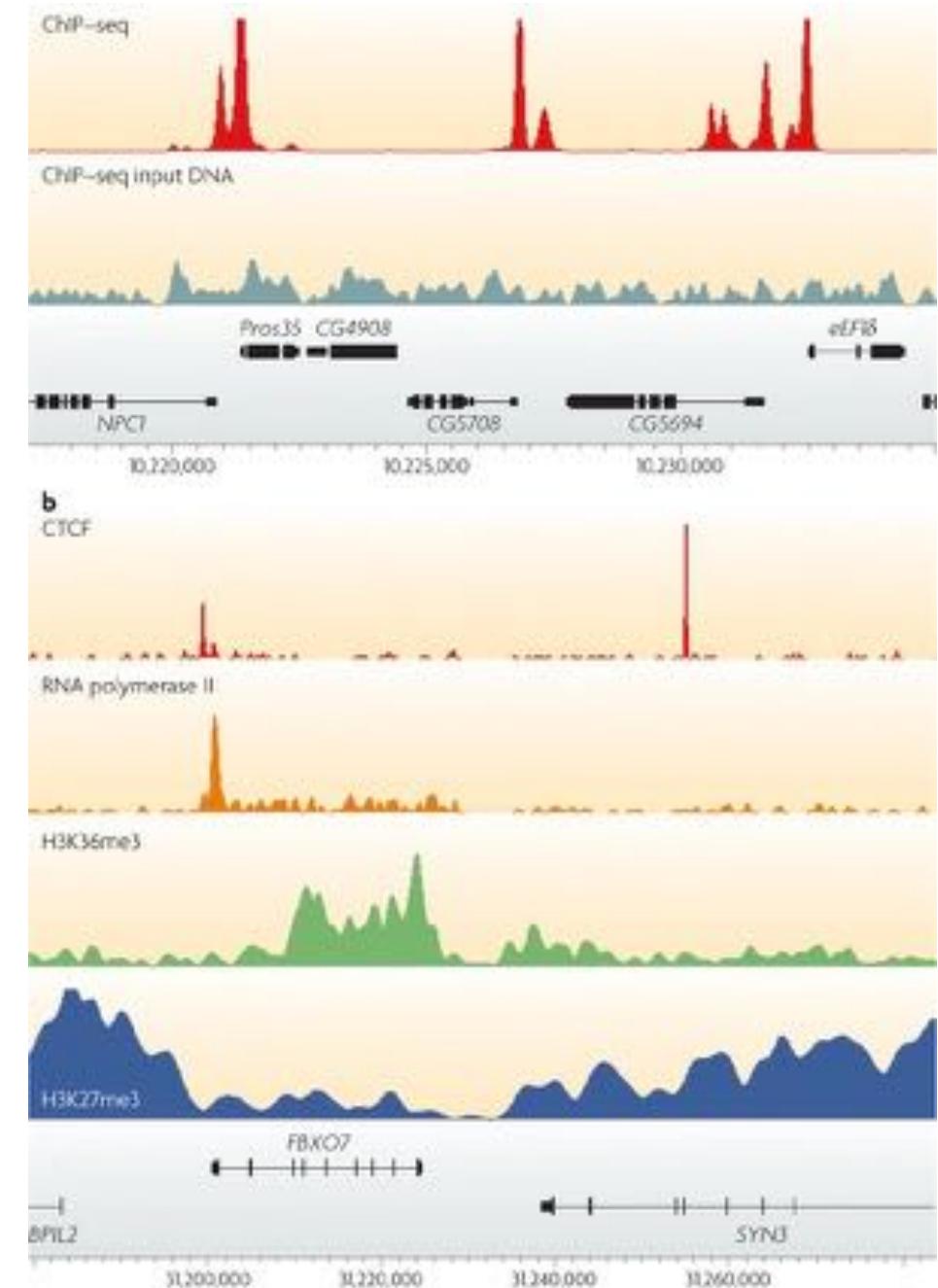


(Modified from [Nature Methods doi:10.1038/nmeth.1371](https://doi.org/10.1038/nmeth.1371),  
S.Pepke)

# THREE DIFFERENT TYPES OF PEAKS



Sims et.al., 2014. doi: 10.1038/nrg3642



Park, 2009. doi:10.1038/nrg2641,  
P Park

Nature Reviews | Genetics

# Histone modifications



- H3K27ac <--> active enhancers and promoters
- H3K4me3 <--> active promoter
- H3K4me1 <--> active enhancers
- H3K27me3 <--> silenced genes

Broad peaks	Narrow peaks
H3K27me3	H3ac
H3K36me3	H3K27ac
H3K4me1	H3K4me2
H3K79me2	H3K4me3
H3K79me3	H3K9ac
H3K9me1	
H3K9me2	
H4K20me1	

Zanconato et.al., 2015. doi:10.1038/ncb3216

<https://www.cellsignal.com/contents/resources-reference-tables/histone-modification-table/science-tables-histone>

# Peak caller

Transcription factor peaks:

- MACS
- Homer

Histone marks:

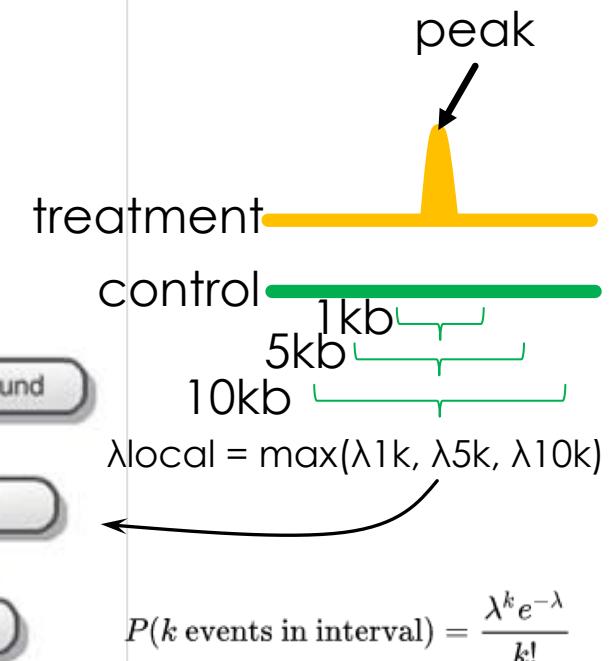
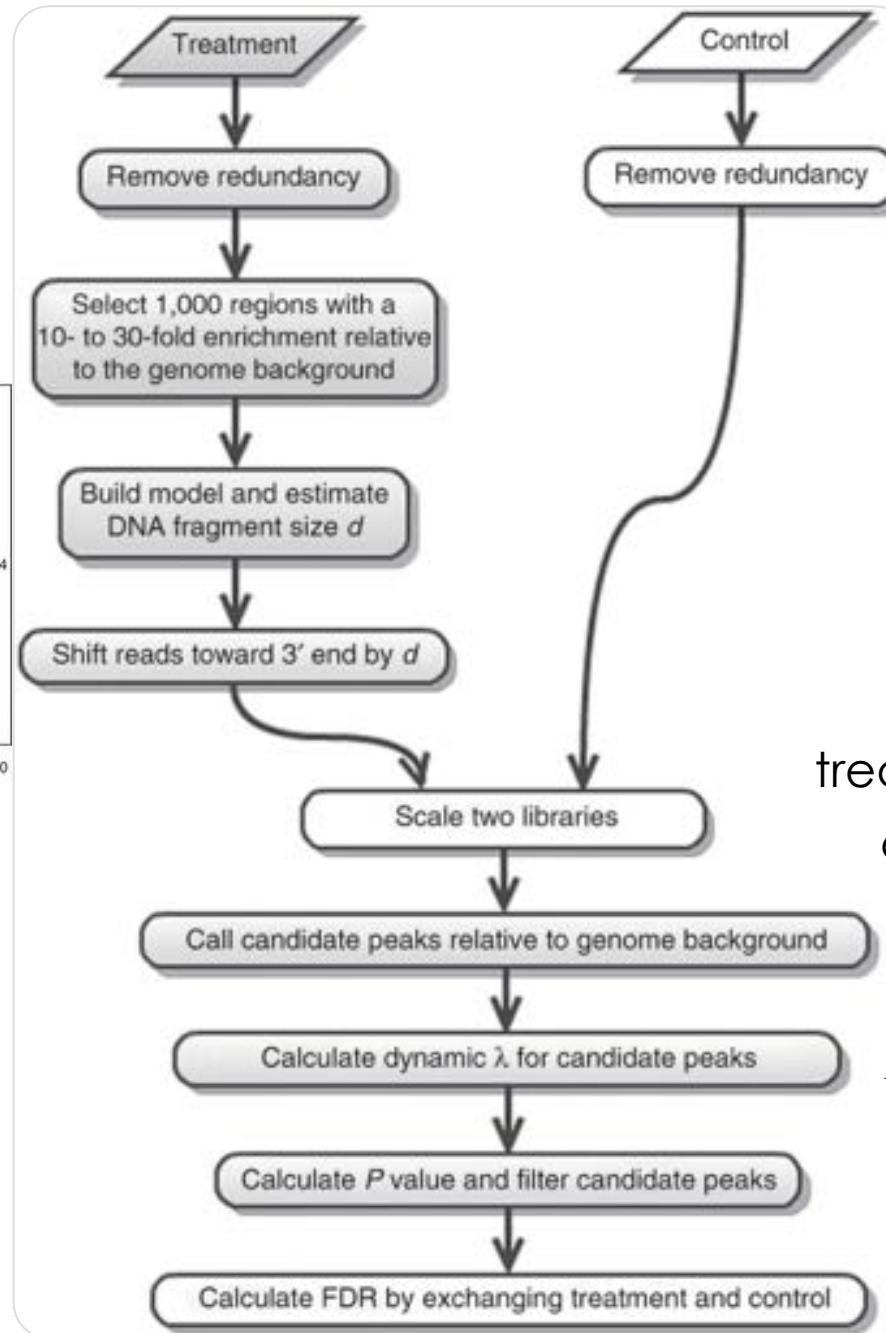
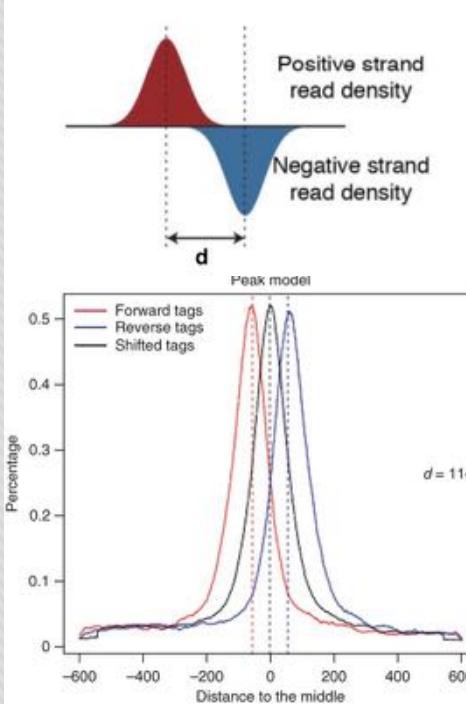
- SICER2
- SPP+MACS
- Homer

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FEE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FEE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X			X	X		X		X		conditional binomial model
Minimal ChIPSeq Peak Finder	16	2.0.1			X			X			X			
E-RANGE	27	3.1			X			X			X	X		chromosome scale Poisson dist.
MACS	13	1.3.5		X				X				X		local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromosome scale Poisson dist.
HPeak	29	1.1		X				X				X		Hidden Markov Model
Sole-Search	23	1	X	X				X		X		X		One sample t-test
PeakSeq	21	1.01			X			X				X		conditional binomial model
SiSSRS	32	1.4		X			X				X			
spp package (wid & mtc)	31	1.7		X			X		X	X'	X			

X\* = Windows-only GUI or cross-platform command line interface  
X\*\* = optional if sufficient data is available to split control data  
X' = method excludes putative duplicated regions, no treatment of deletions

Generating density profiles      Peak assignment      Adjustments w. control data      Significance relative to control data

# MACS/2 (Model-based Analysis of ChIP-seq)



$$P(k \text{ events in interval}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

# Running MACS2

Sharp peaks:

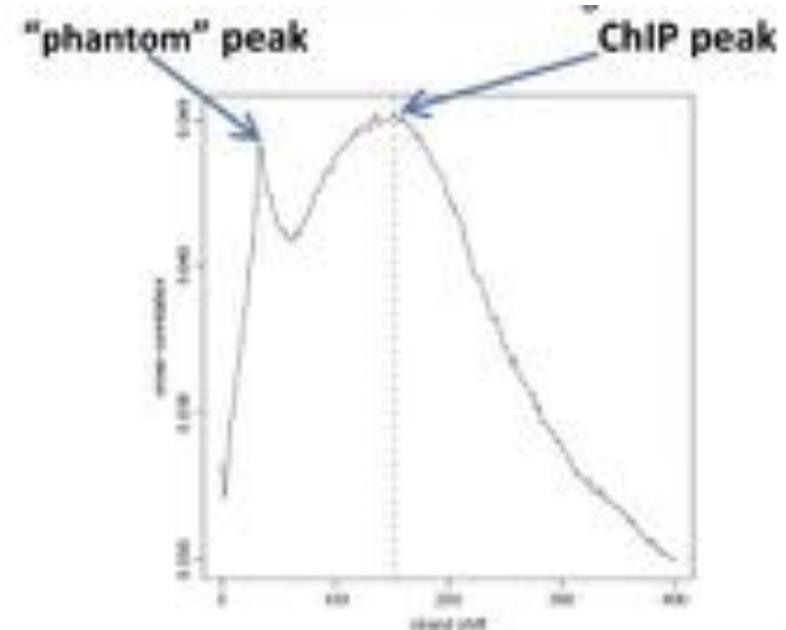
```
macs2 callpeak -t treatment.bam -c input.bam \
-f BAM -g hs -n sampleName --outdir outputFolder \
-q 0.01
```

Broad peaks:

```
macs2 callpeak -t treatment.bam -c input.bam \
-f BAM -g hs -n sampleName --outdir outputFolder \
--broad --broad-cutoff 0.1
```

# Running MACS2 with user-defined model

```
macs2 callpeak -t treatment.bam -c input.bam \
-f BAM -g hs -n sampleName \
--outdir outputFolder \
--nomodel --shift 80 --extsize 160
```



Calculate cross-correlation by phantompeaktools

Rscript run\_spp.R -c=<tagAlign/BAMfile> -savp -out=<outFile>

<https://github.com/kundajelab/phantompeakqualtools>

# Output of MACS2

- NAME\_peaks.xls
- NAME\_peaks.narrowPeak
- NAME\_summits.bed
- chromosome name
- start position of peak
- end position of peak
- length of peak region
- absolute peak summit position
- pileup height at peak summit
- $-\log_{10}(\text{pvalue})$  for the peak summit (e.g. pvalue = $1e-10$ , then this value should be 10)
- fold enrichment for this peak summit against random Poisson distribution with local lambda,
- $-\log_{10}(\text{qvalue})$  at peak summit

# File formats: bed, bedGraph, wig, bigwig, narrowPeak, broadPeak, ...

- Help documentation:  
<https://genome.ucsc.edu/FAQ/FAQformat.html>
- BED format:
  - chrom chromStat chromEnd name score strand ...
  - Sample file:

```
chr22 1000 5000 cloneA 960 +
chr22 2000 6000 cloneB 900 -
```
- WIG (Wiggle Track) format: (no strand information)
  - **variableStep format**

```
variableStep chrom=chrN
[span=windowSize]
chromStartA minValueA
chromStartB minValueB
... etc ... etc ...
```
  - **fixedStep format**

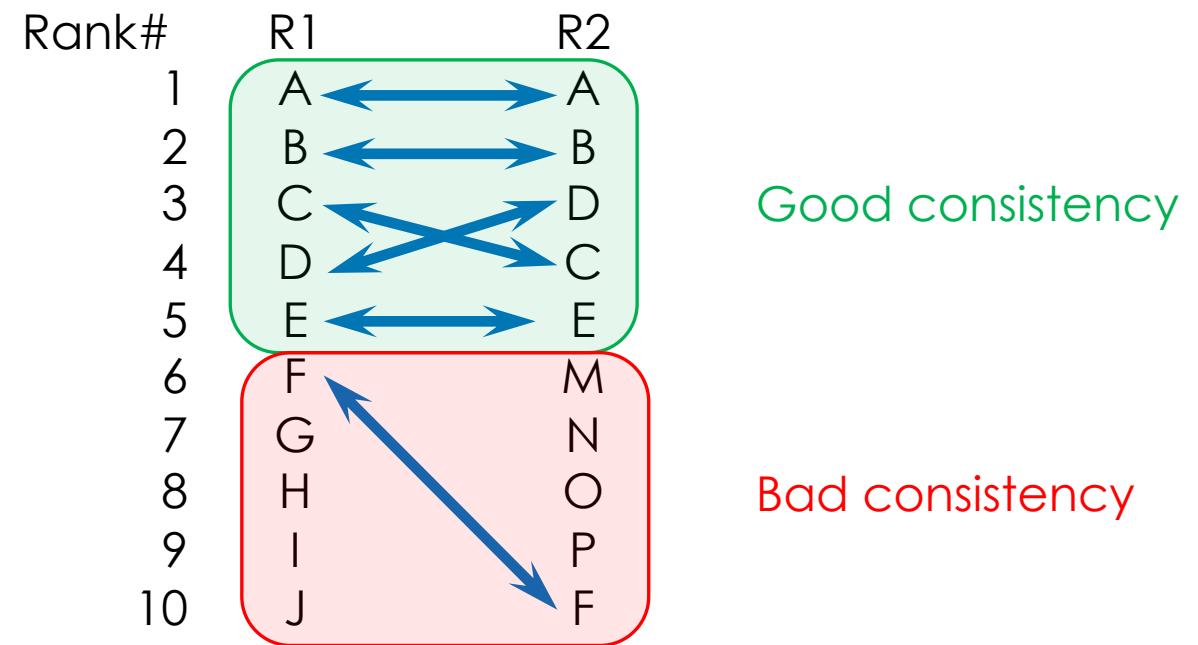
```
fixedStep chrom=chrN
start=position step=stepInterval
[span=windowSize]
dataValue1
dataValue2
... etc ...
```

# Blacklist/Hotspots

- Blacklisted regions are genomic regions with anomalous, unstructured, high signal or read counts in NGS experiments, independent of cell type or experiment.
- Where to download: <https://sites.google.com/site/anshulkundaje/projects/blacklists>

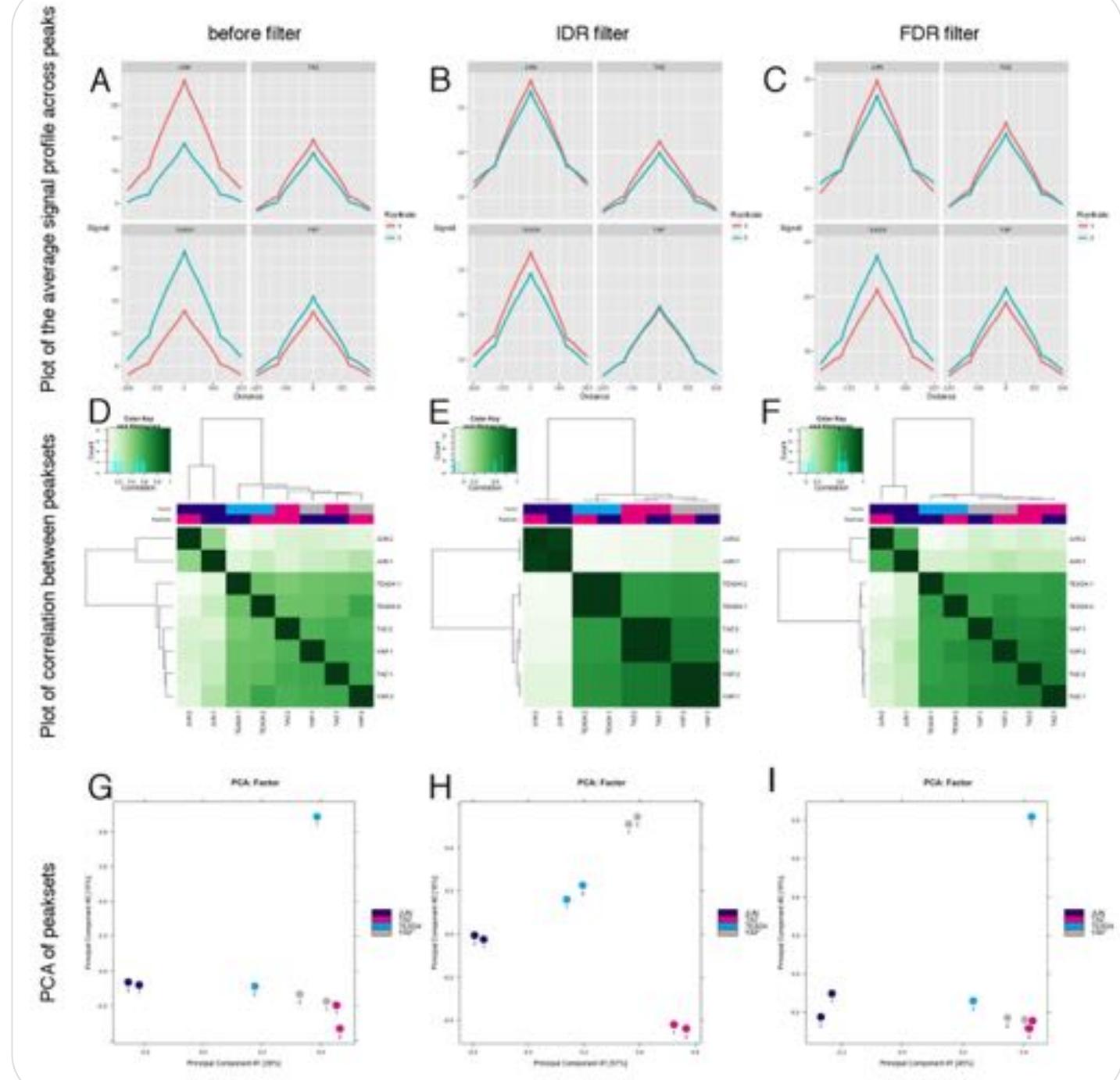
# Irreproducibility Discovery Rate (IDR) analysis

The IDR framework to filter the low reproducible peaks. The IDR framework is a unified approach to measure the reproducibility of findings identified from replicate experiments and provide highly stable thresholds based on reproducibility. ([Landt et al., 2012](#))



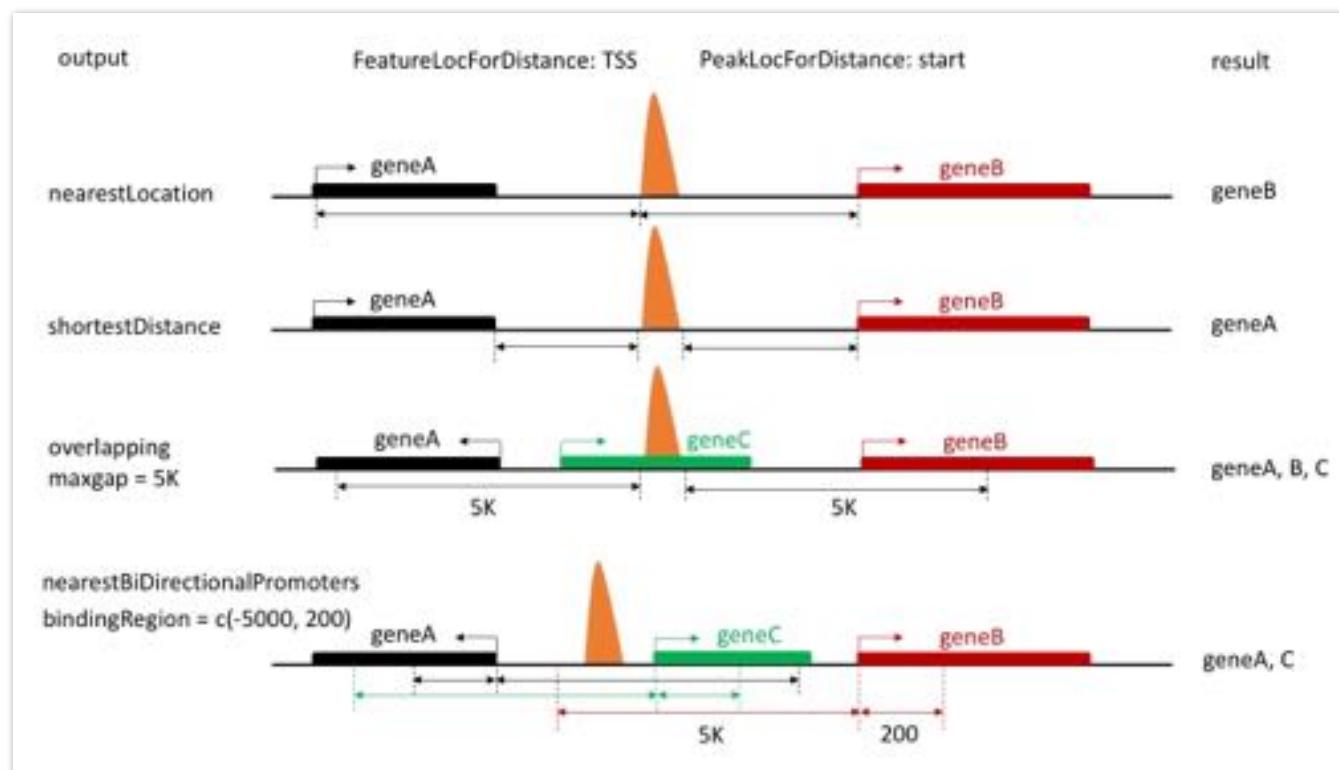
<https://arxiv.org/abs/1110.4705>  
<https://github.com/nboley/idr>

# The effect of IDR

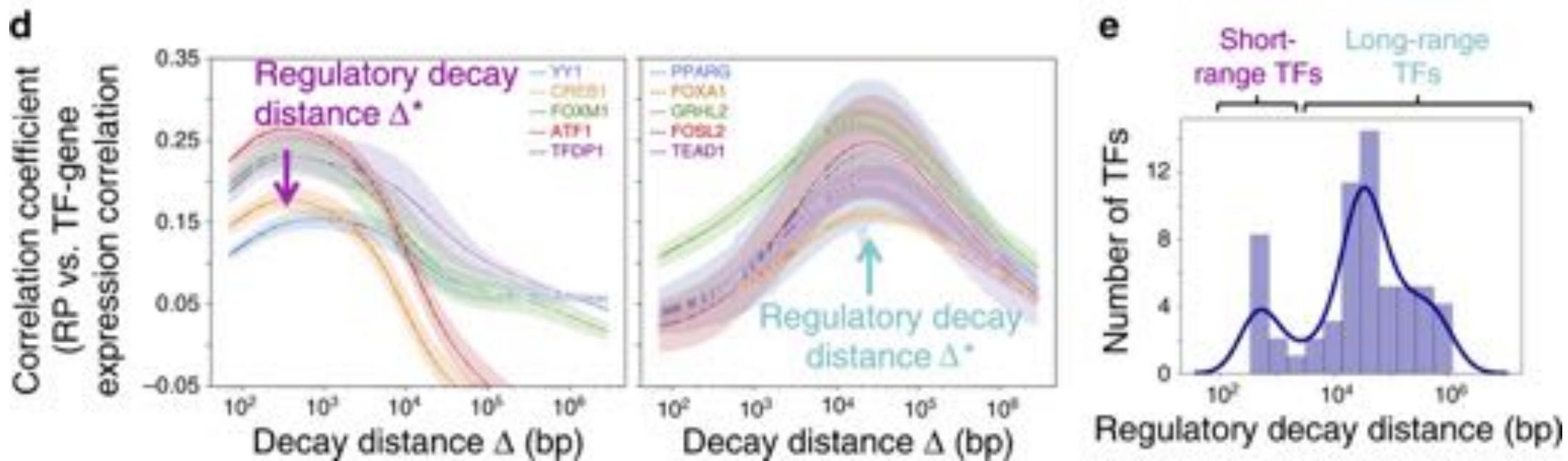


# Annotation by ChIPpeakAnno

The *annotatePeakInBatch* and *annoPeaks* functions in [ChIPpeakAnno](#) package can be used to annotate the peaks. Depend on the experiment, we can annotate the peaks in multiple ways.

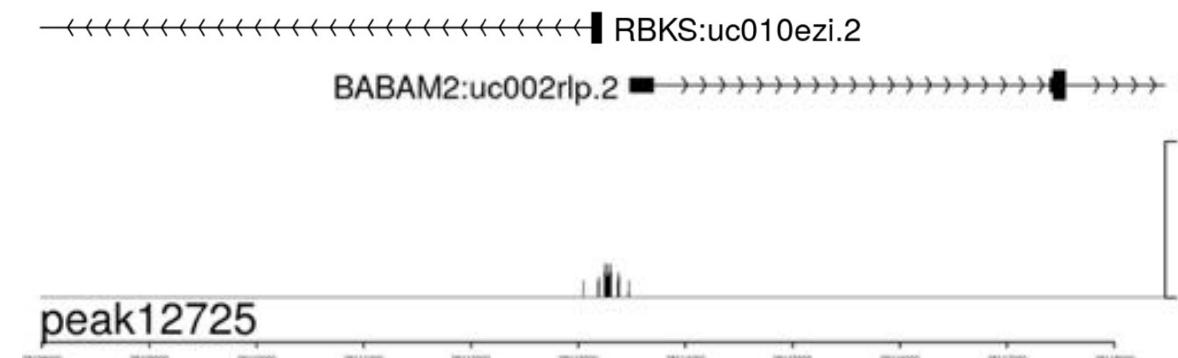


# Two distinct TF classes: short-range and long-range



# Bidirectional promoters

- Bidirectional promoters are the DNA regions located between the 5' ends of two adjacent genes coded on opposite strands. The two adjacent genes are transcribed to the opposite directions, and often co-regulated by this shared promoter region(Robertson et al., 2007).



# Functional annotation of DE genes

Gene ontology  
(GO) enrichment  
analysis

Kyoto Encyclopedia  
of Genes and  
Genomes (KEGG)  
pathway  
enrichment analysis

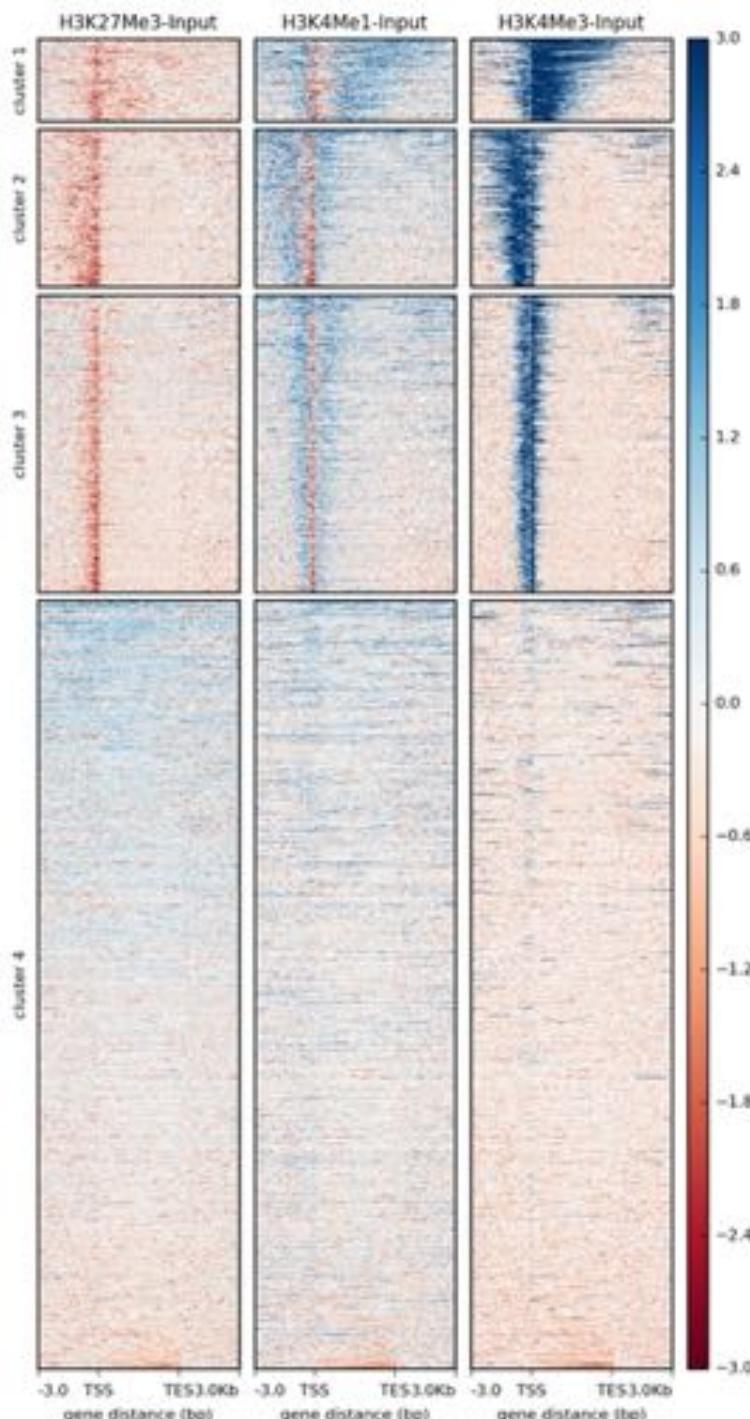
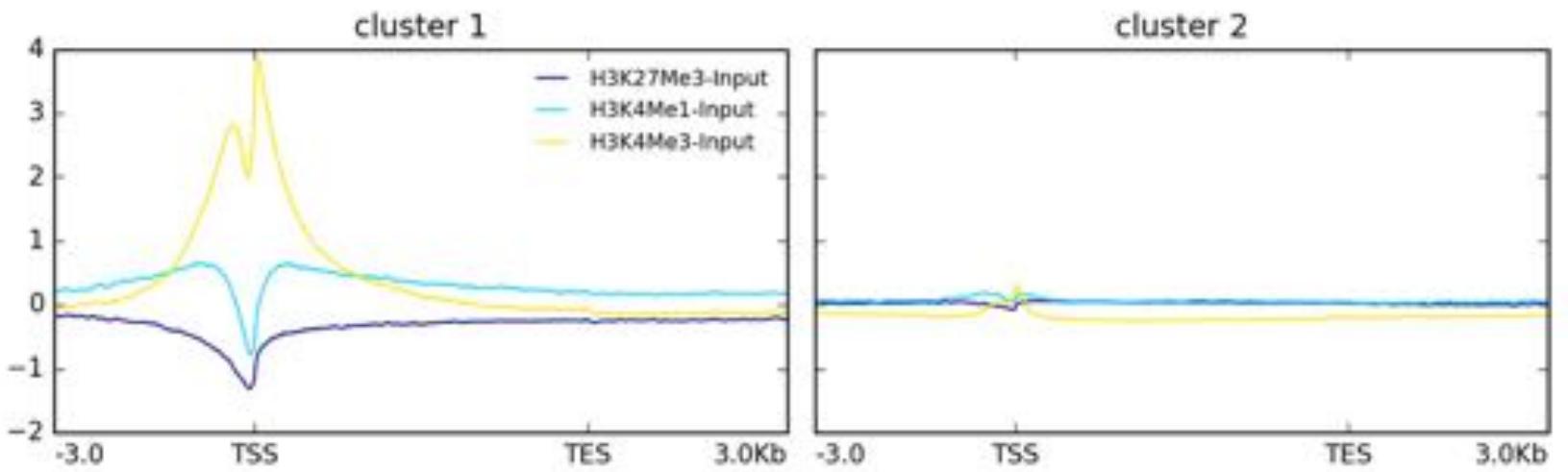
Gene set  
enrichment analysis  
(GSEA)

QIAGEN Ingenuity  
Pathway Analysis  
(IPA)

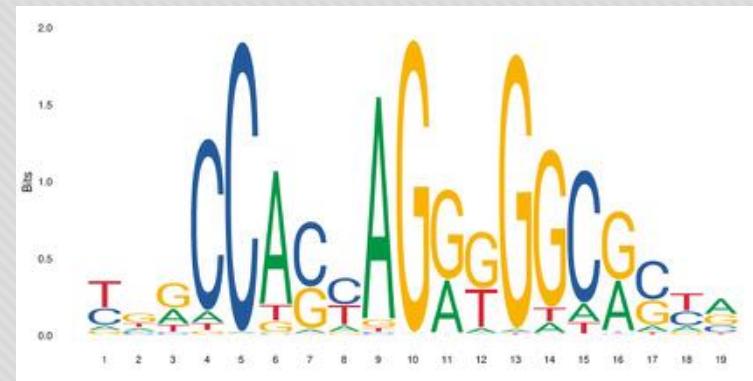
A gene set over-represented comparing to random sampling

# Metagene analysis

- A metagene analysis is an average of quantitative data over more than one genomic regions. The interested genomic regions include promoters, transcription factor binding sites, enhancers, etc.



# Motif search



- Sequence motif represents conserved characteristics such as DNA binding sites, where transcription factors bind, and catalytic sites in enzymes.
- A sequence logo is a graphical representation of sequence conservation in multiple sequences.
- To search an enriched motifs, multiple softwares could be used, such as [the MEME Suite](#)(Bailey et al., 1994), [Consensus](#)(Hertz et al., 1990), [rGADEM](#), [Homer](#), and et. al.

# Build Regulation Network: Integrating TF binding with transcriptomic data

GeneNetworkBuilder (GNB) can be used to facilitate identification of the complex regulatory network of TFs and how indirect targets are inter-connected.

Issue: hairball – too much information

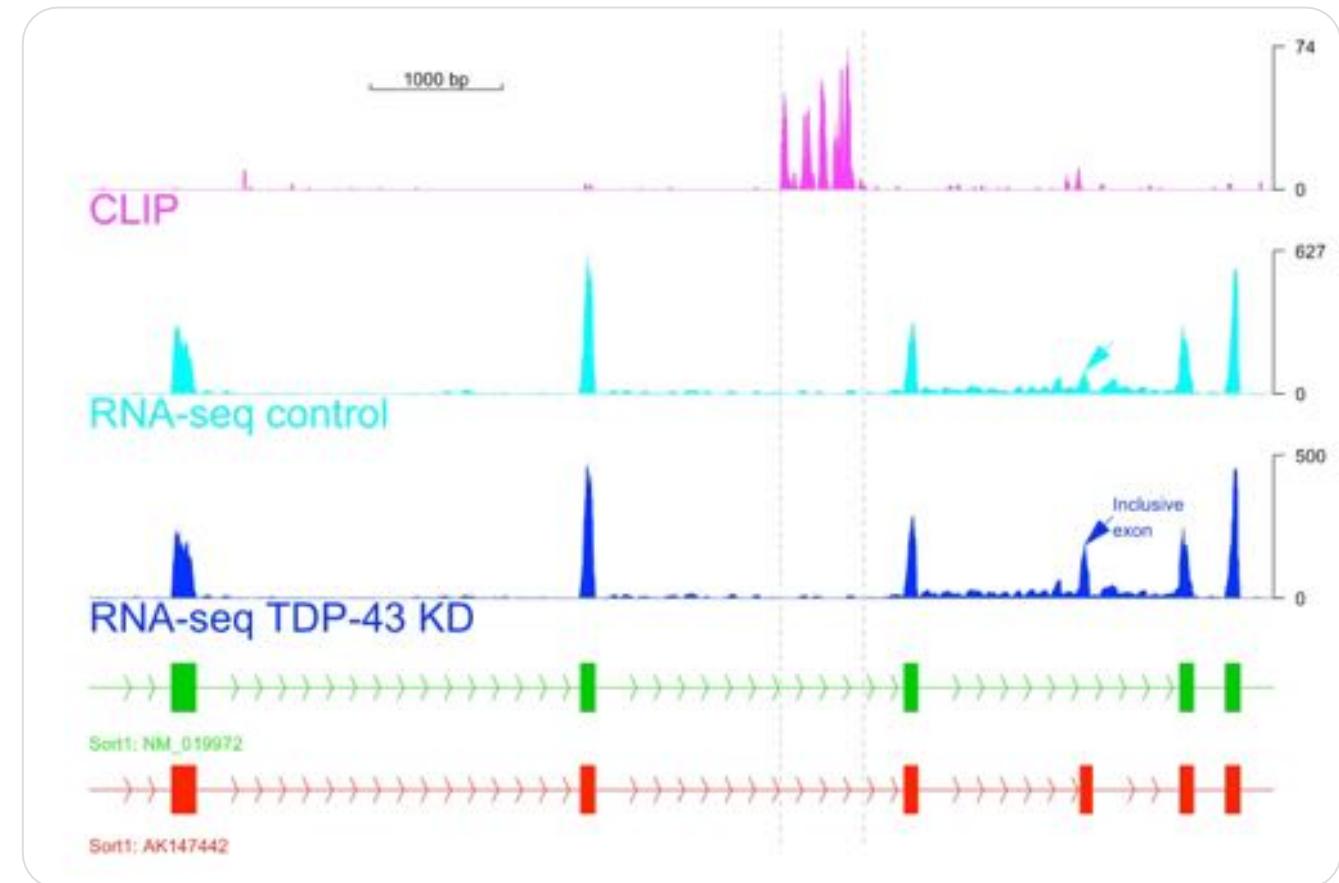
filter data by enrichment analysis: ClueGO

filter the interaction database by evidence.

Focus on a subset network

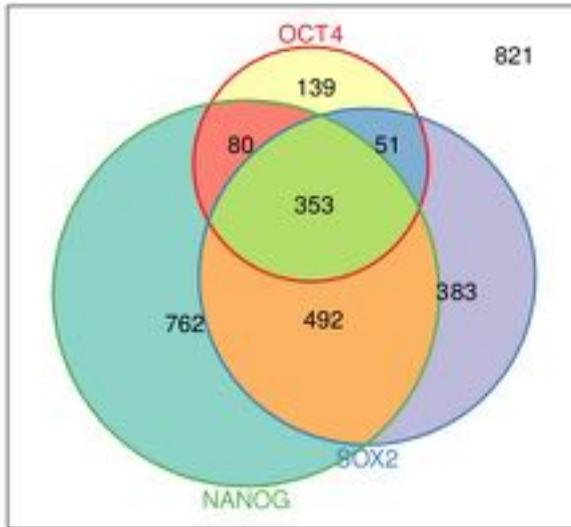
<https://bioconductor.org/packages/GeneNetworkBuilder/>

# View tracks by trackViewer



# Association among different sets of peaks

## Venn diagram and upset plot



<https://github.com/js229/Vennerable>

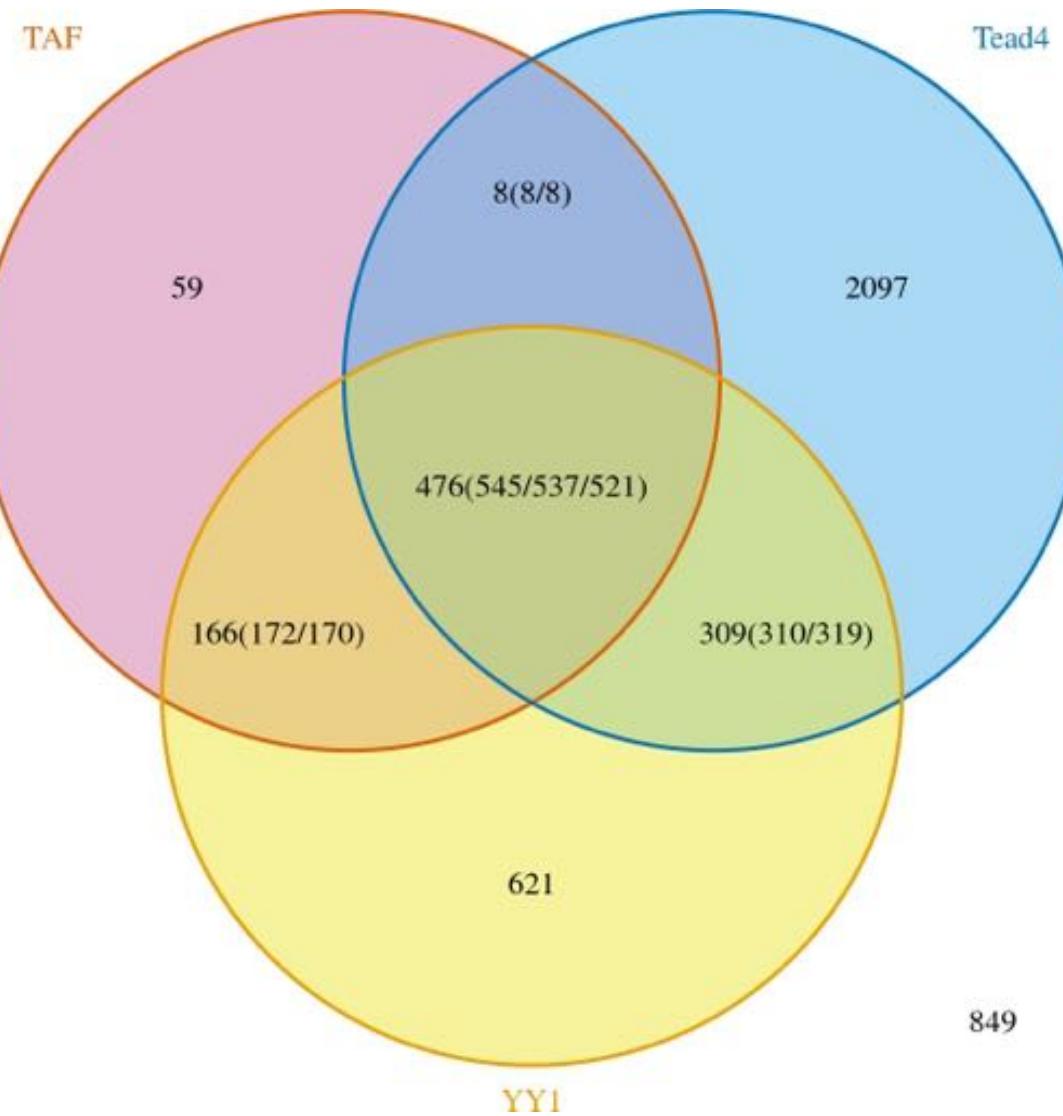


<http://vcg.github.io/upset/>

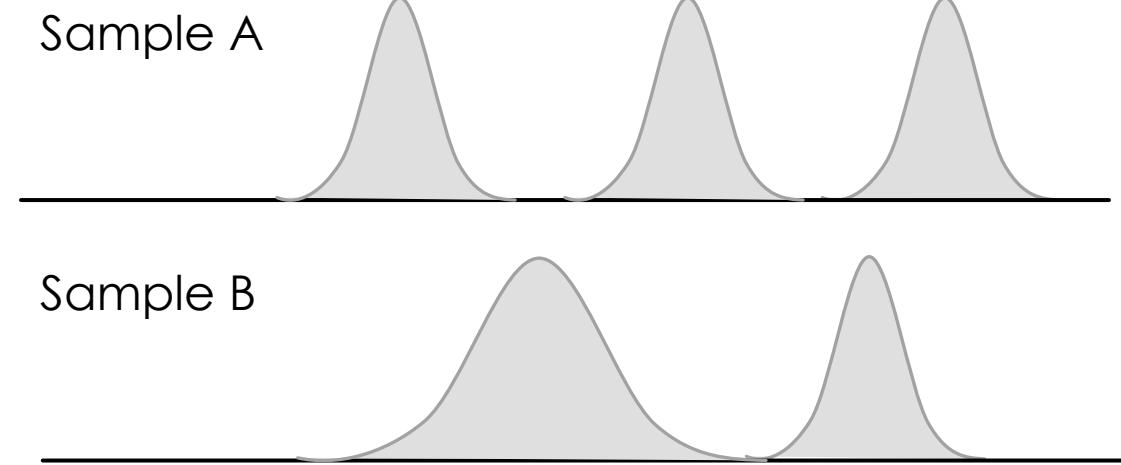
<http://caleydo.org/tools/upset/>



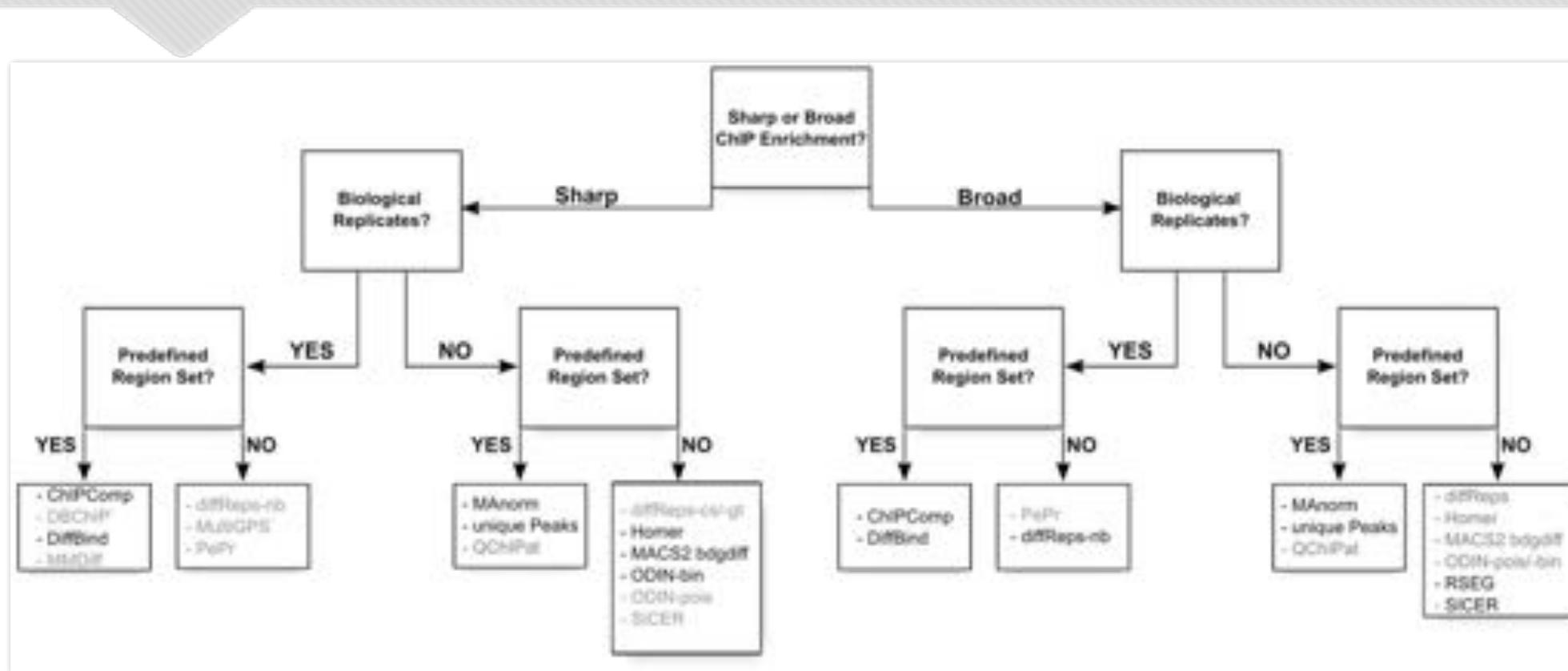
## FIND OVERLAPS FOR REPLICATES, *findOverlapsOfPeaks*



849



# Differential binding sites analysis



# Differential binding sites analysis



DiffBind vs.  
csaw

Peak  
based vs.  
window  
based



Using control or not

# Tools for ChIP-seq analysis

- Mapping: [bowtie1/2](#), [Rsubread](#), [bwa](#)
- Peak calling: [CCAT](#), [SICER](#), [MACS](#), [ZINBA](#), [BayesPeak](#), [chipseq](#), [ChIPseqR](#), [CSAR](#), [csaw](#), [GenoGAM](#), [iSeq](#), [PICS](#)
- Quality assessment: [ChIPQC](#)
- Analysis of differential binding: [DiffBind](#), [csaw](#), [ChIPDiff](#), [DBChIP](#), [MMDiff](#)
- Peak annotation and pathway analysis: [ChIPpeakAnno](#), [Homer](#), [PAVIS](#), [GREAT](#), [ChIPseeker](#), [clusterProfiler](#), [GOstats](#)
- Gene Network Building: [\*\*GeneNetworkBuilder\*\*](#), [ChIPXpress](#)
- Visualization: [trackViewer](#), [IGV](#), [UCSC genome browser](#), [rtracklayer](#)
- Motif enrichment analysis: [The MEME Suite](#), [homer](#)

# Regeneromics Shared Resource can help your research!



We do: experimental design, bioinformatics analysis, manuscript preparation, grant applications



Jianhong Ou, Ph.D.

Email: regeneromics@duke.edu

<https://sites.duke.edu/dukeregenerationcenter/jobsrni/>

# HANDS-ON

# Questions will be answered

- Prepare your system
- Basics of shell script
- Introduction of R programming language

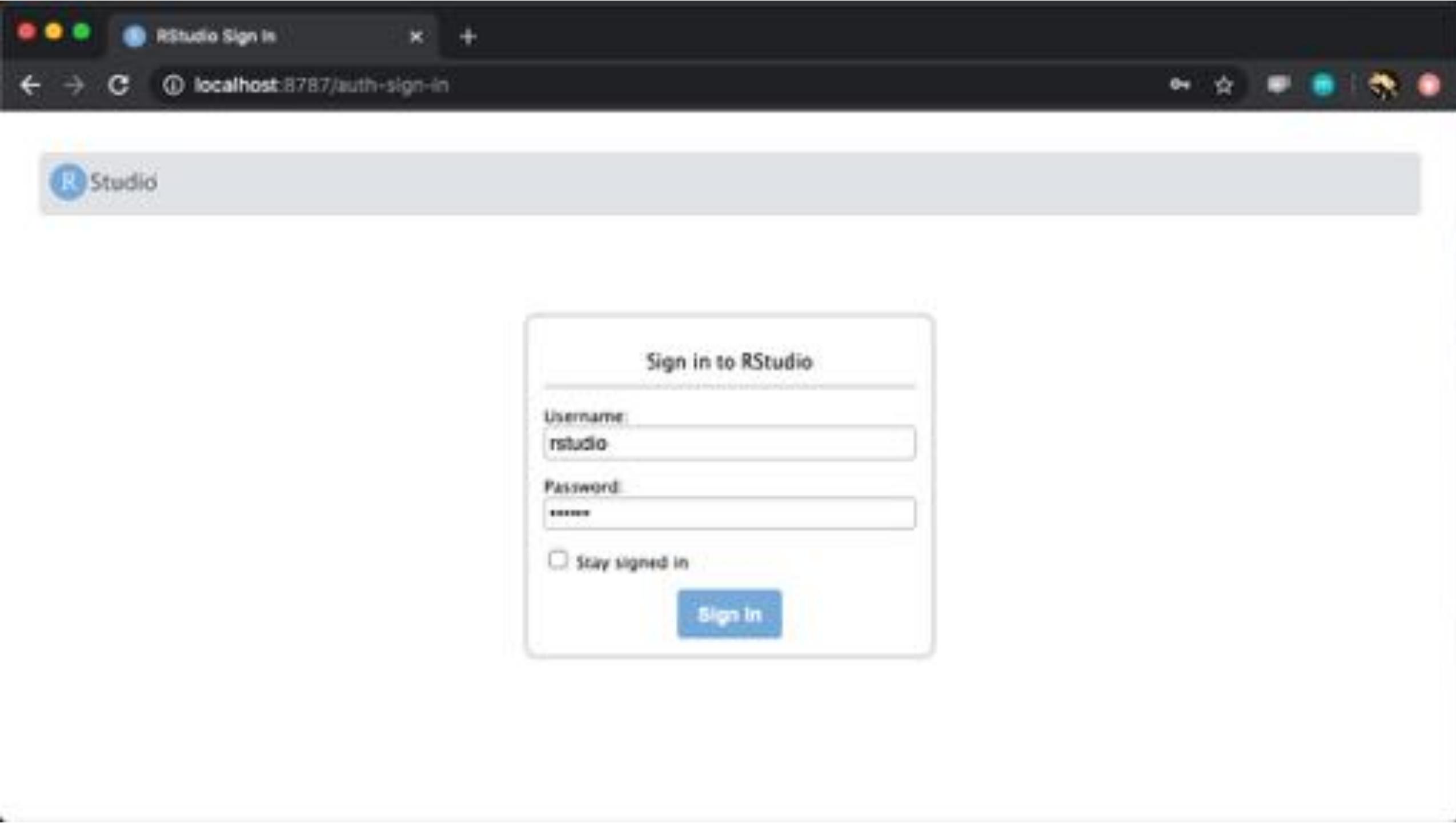
# Set up your computing environment (m1)

- BEST Way: Install from source

# Set up your computing environment (m2)

- Install Docker (<https://www.docker.com/get-started>) (30min)
- Check docker installed: docker --version
- docker pull jianhong/genomictools:latest
- cd ~
- mkdir tmp4genomictools
- docker run -e PASSWORD=123456 -p 8787:8787 -v \${PWD}/tmp4genomictools:/home/rstudio jianhong/genomictools:latest
  - localhost:8787 by username: rstudio password:123456

<https://github.com/jianhong/genomictools>



Studio

## Sign in to RStudio

Username:

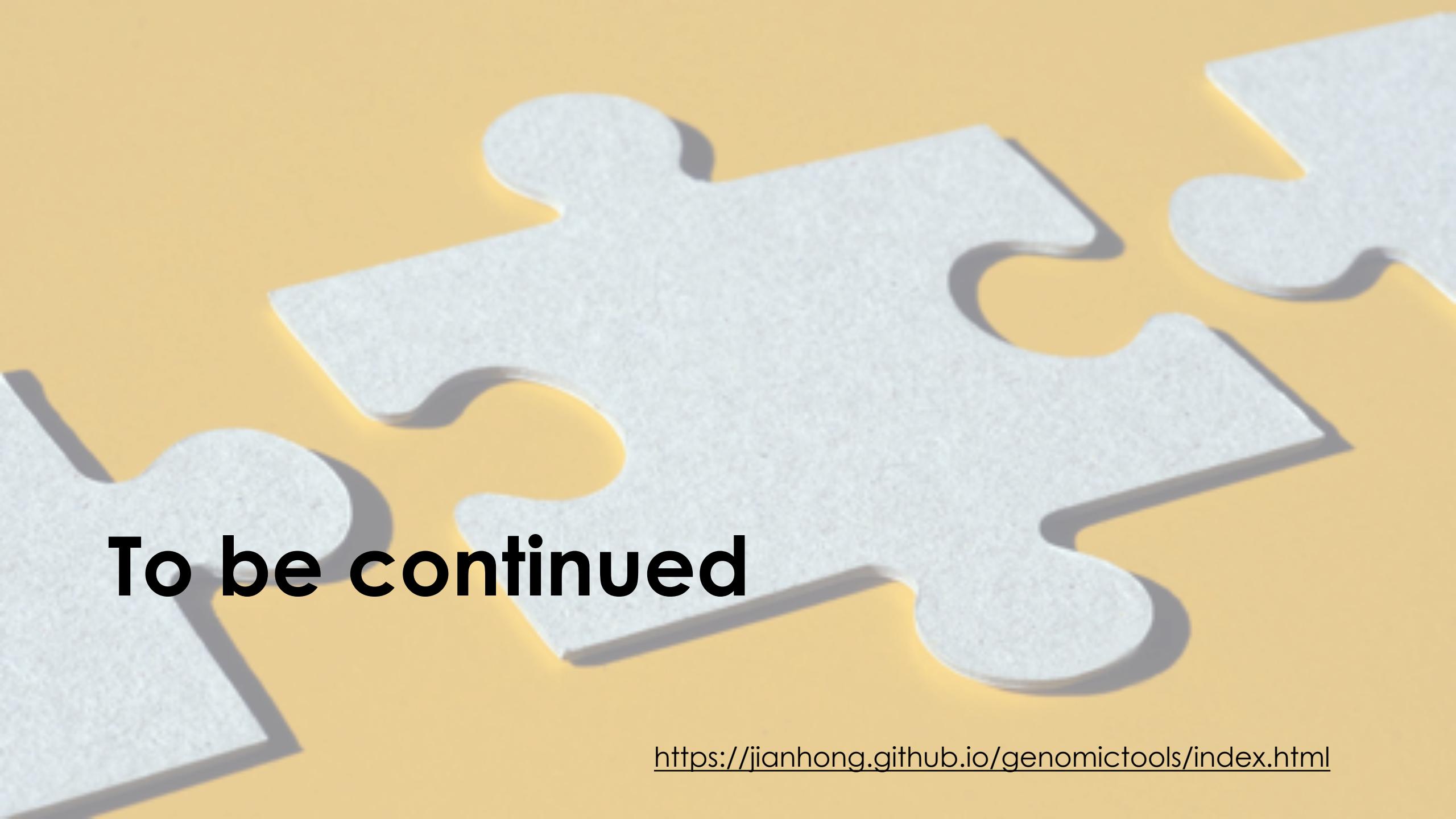
rstudio

Password:

\*\*\*\*\*

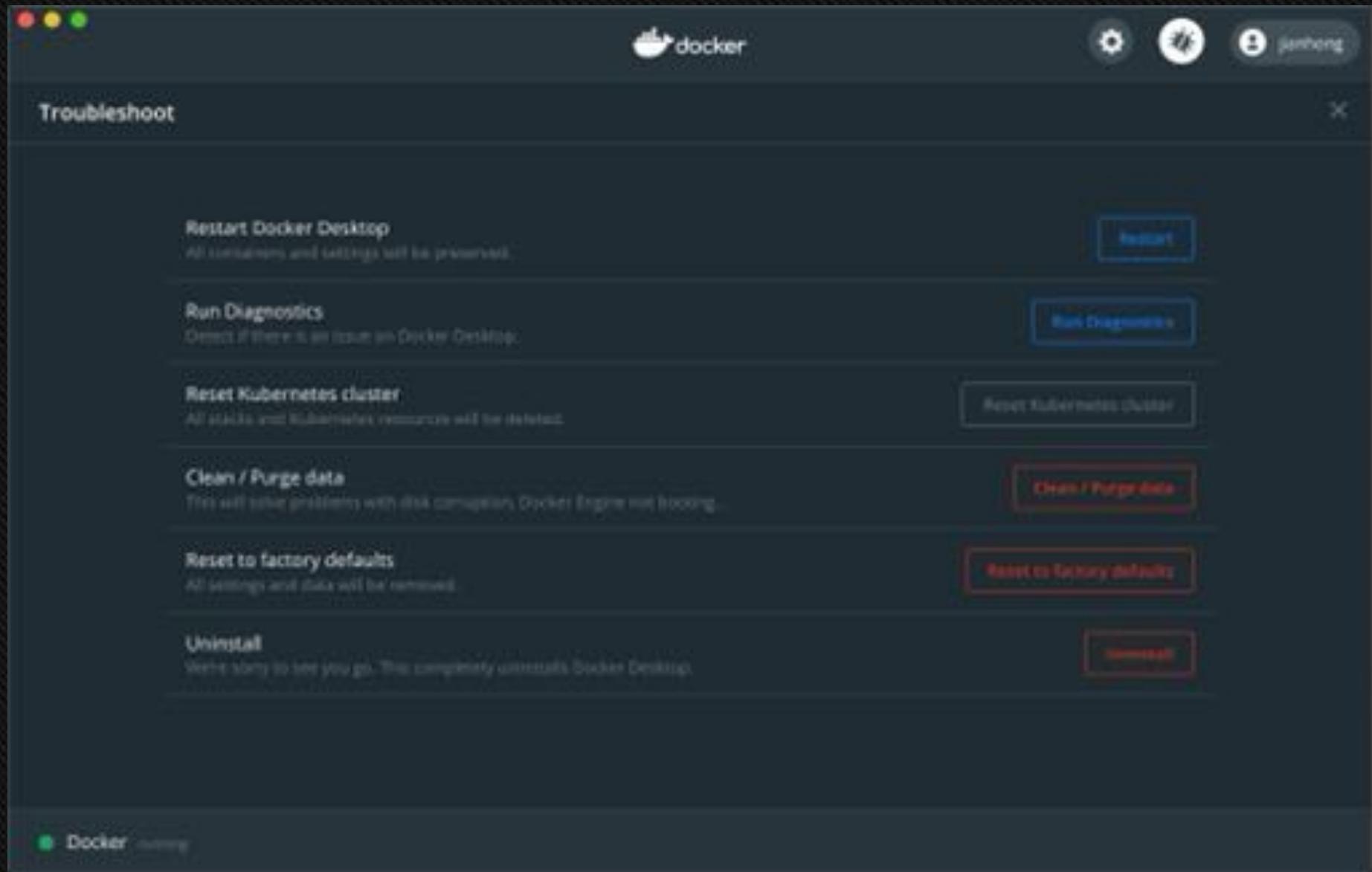
 Stay signed in

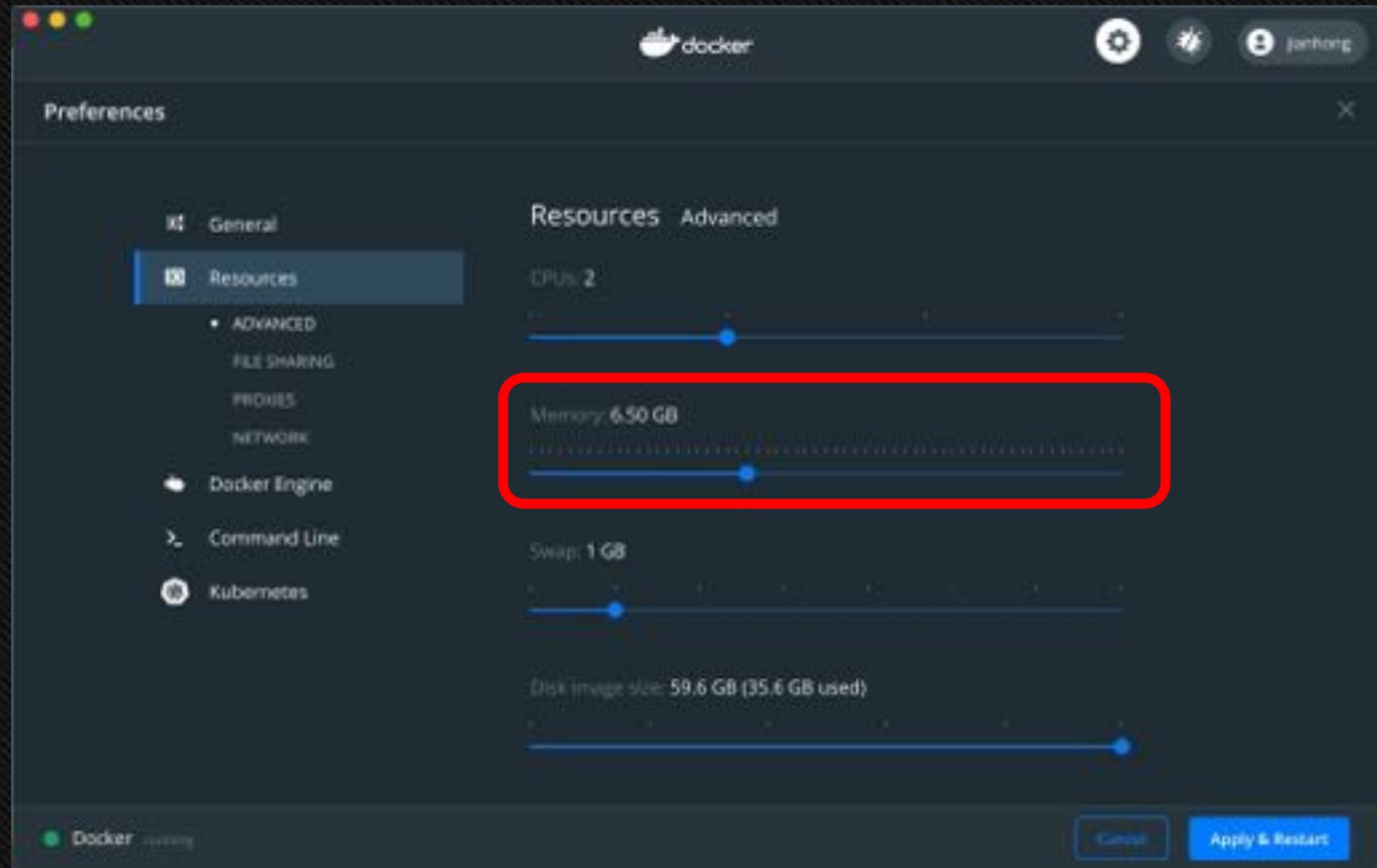
Sign In



**To be continued**

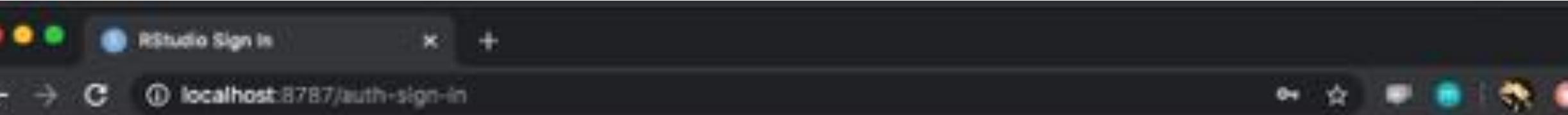
<https://jianhong.github.io/genomictools/index.html>





# Start your docker

```
cd ~  
mkdir tmp4genomictools  
docker run -e PASSWORD=123456 -p 8787:8787 \  
-v ${PWD}/tmp4genomictools:/home/rstudio \  
jianhong/genomictools :latest
```



# Basics of shell scripts

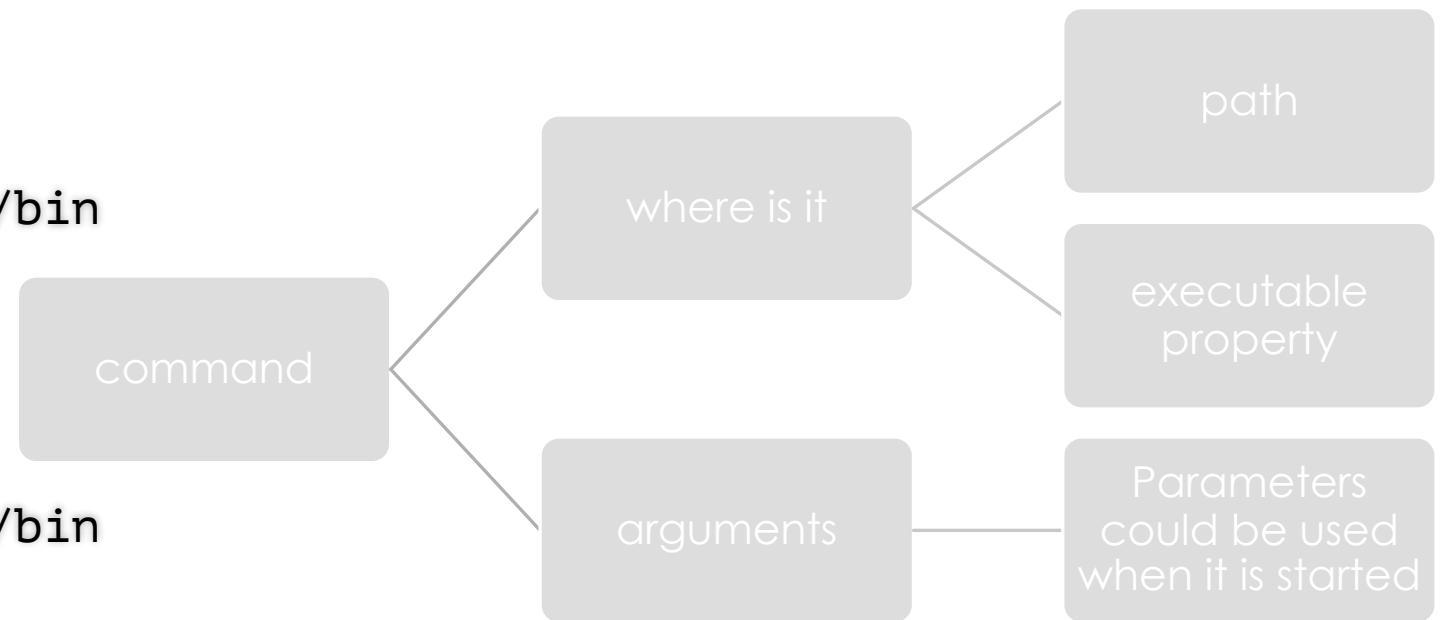
- Unix-like operating systems: “behave” like the original Unix operating system and comply(at least partially) with POSIX (portable operating system interface) standards.
  - Examples: Linux, OS X, Free BSD.
- A terminal: is a software that emulates a teletype writer terminal used in the early days of Unix.
- A shell: is a software that runs inside a terminal and interprets and executes user commands. One of the most popular shells being used today is called BASH (Bourne Again Shell)

# Hello Bash

```
command          Options or arguments
riboimager:bin jianhongou$ echo Hello Bash
Hello Bash
prompt          arguments
riboimager:bin jianhongou$ ssh user@dcc-slogin.oit.duke.edu
#####
command
#####
# You are about to access a Duke University computer network that is intended #
# for authorized users only. You should have no expectation of privacy in      #
# your use of this network. Use of this network constitutes consent to          #
# monitoring, retrieval, and disclosure of any information stored within the    #
# network for any purpose including criminal prosecution.                      #
#####
user@dcc-slogin.oit.duke.edu's password: user@dcc-slogin-03 ~ $ whoami
user
command
```

# command line

```
## to see working directory  
$ pwd  
/Users/jianhongou/miniconda3/bin  
$ which pwd  
/bin/pwd  
$ /bin/pwd  
/Users/jianhongou/miniconda3/bin  
$ ls -lh bowtie2  
-rwxrwxr-x 2 jianhongou staff 18K Jan 22 2017 bowtie2
```



# Getting help

```
## try to get documentation of bowtie
$ ./bowtie2 -h
## get help of `ls`
$ man ls
```

# Download transcript fasta

		Show 10 entries	Show/hide columns		Filter		
★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	
Y	Human <i>Homo sapiens</i>	<a href="#">FASTA</a>					
Y	Mouse <i>Mus musculus</i>	<a href="#">FASTA</a>					
Y	Zebrafish <i>Danio rerio</i>	<a href="#">FASTA</a>					

<http://useast.ensembl.org/info/data/ftp/index.html>

```
curl -0 ftp://ftp.ensembl.org/pub/release-100/fasta/danio_rerio/cdna/Danio_rerio.GRCz11.cdna.all.fa.gz
```

# Step1: build transcriptome index

## kallisto

```
kallisto index -i danRer.GRCz11_transcrits.idx Danio_rerio.GRCz11.cdna.all.fa.gz
```

## Salmon

```
salmon index -i danRer.GRCz11_transcrits.salmon.idx -t Danio_rerio.GRCz11.cdna.all.fa.gz
```

# Step 2: quantify

## kallisto

```
kallisto quant -i danRer.GRCz11_transcrits.idx \
-o kallisto_quant/$cond.rep$rep \
-b 30 -t 2 fastq/$cond.rep$rep.fastq.gz \
--single -l 200 -s 50
```

## Salmon

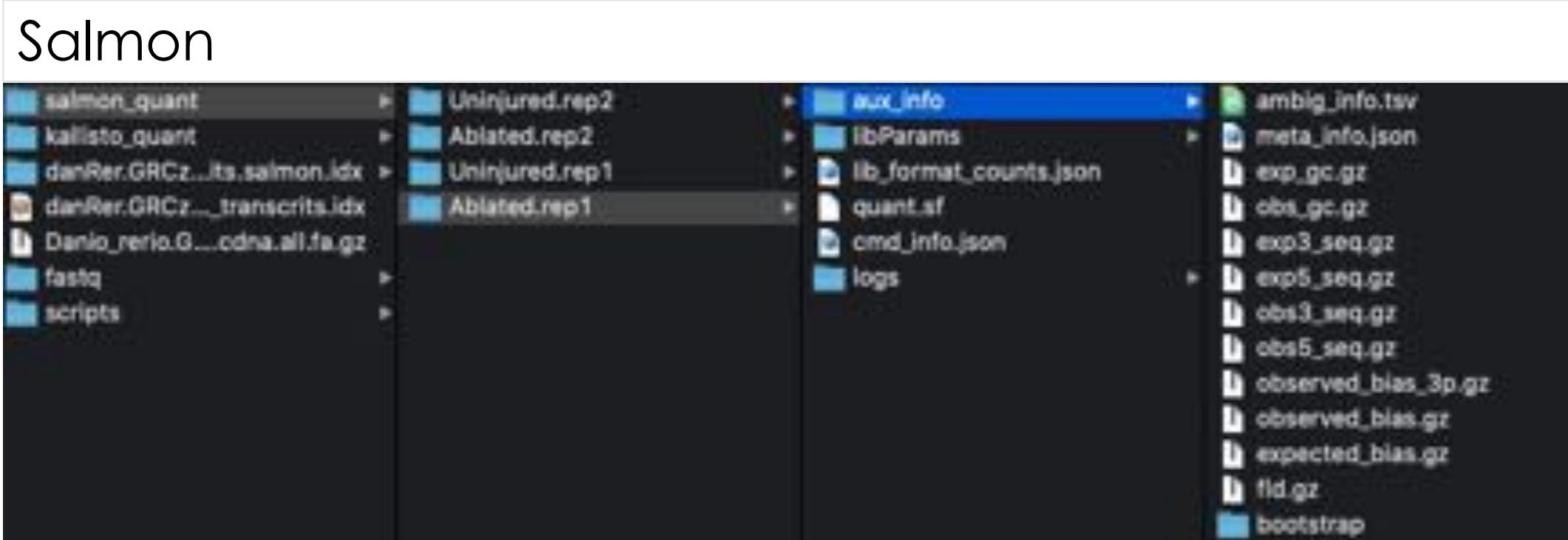
```
salmon quant -i danRer.GRCz11_transcrits.salmon.idx -l A \
-r fastq/$cond.rep$rep.fastq.gz \
--validateMappings -p 2 \
-o salmon_quant/$cond.rep$rep \
--numBootstraps 30 --seqBias --gcBias
```

# output

## kallisto



## Salmon



# output

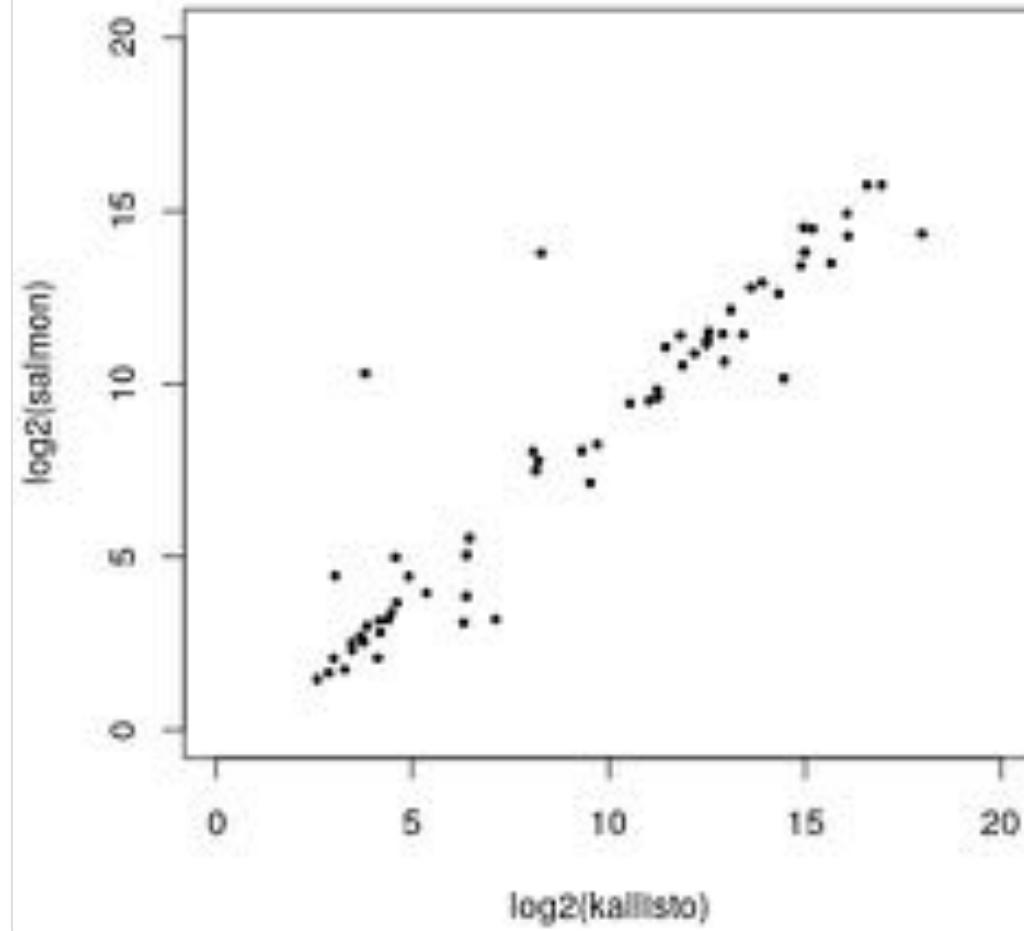
## kallisto

	target_id	length	eff_length	est_counts	tpm
256	ENSDART00000146315.3	2622	2422.990	260.627000	5793.73000
257	ENSDART0000021121.8	1181	981.994	200.373000	10990.60000
279	ENSDART00000140081.2	6779	6579.990	1.760960	14.41500
281	ENSDART00000138182.3	6809	6609.990	0.239042	1.94789
330	ENSDART00000169072.2	8455	8255.990	1178.350000	7687.70000
331	ENSDART00000147947.2	8535	8335.990	719.649000	4650.01000

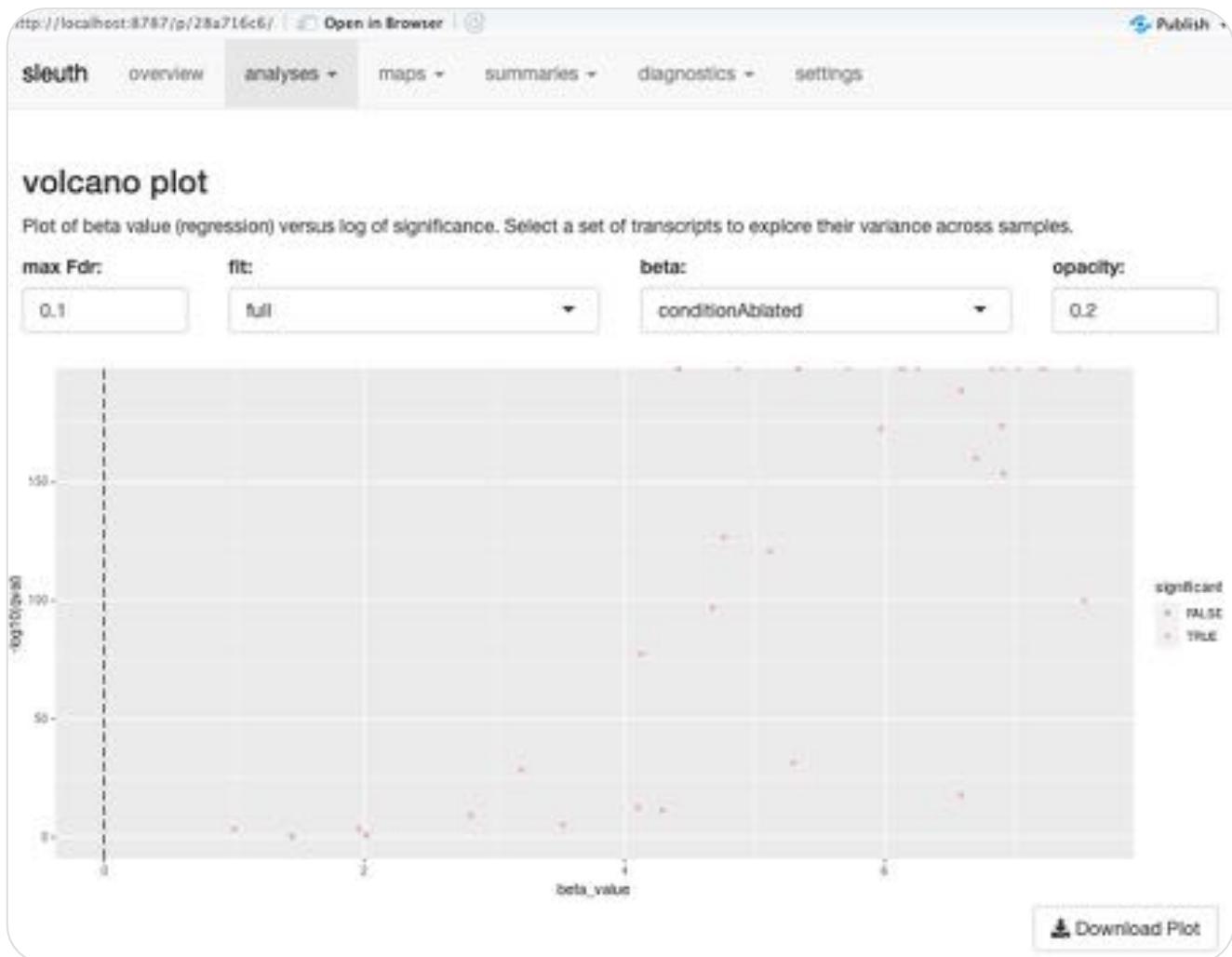
## Salmon

	Name	Length	EffectiveLength	TPM	NumReads
256	ENSDART00000146315.3	2622	2838.809	2434.698251	292.406
257	ENSDART0000021121.8	1181	1526.830	2749.368836	177.594
279	ENSDART00000140081.2	6779	5857.133	8.071238	2.000
330	ENSDART00000169072.2	8455	9663.830	2781.720176	1137.280
331	ENSDART00000147947.2	8535	9769.232	1884.152025	778.720
339	ENSDART0000060594.6	2004	1621.171	14.579657	1.000

Kallisto and  
Salmon  
gene  
counts  
overall  
similar



# Sleuth\_live



# Introduction of R programming language

```
    .operation == "MIRROR_X":  
    mirror_mod.mirror_object =  
    mirror_mod.use_x = True  
    mirror_mod.use_y = False  
    mirror_mod.use_z = False  
    .operation == "MIRROR_Y":  
    mirror_mod.use_x = False  
    mirror_mod.use_y = True  
    mirror_mod.use_z = False  
    .operation == "MIRROR_Z":  
    mirror_mod.use_x = False  
    mirror_mod.use_y = False  
    mirror_mod.use_z = True
```

```
#Selection at the end -add to  
    r_ob.select= 1  
    tier_ob.select=1  
    context.scene.objects.active =  
    ("Select "+ str(modifier))  
    mirror_mod = r_ob.data  
    mirror_mod.select = selected  
    data.objects[one_name].select =  
    selected
```

```
print("please select exactly one object. CLASSES -----")
```

```
types.Operator):  
    " X mirror to the selected object.mirror_mirror_X"  
    " mirror_X"
```

# Install R/Bioconductor/Rstudio

- <https://cloud.r-project.org/>



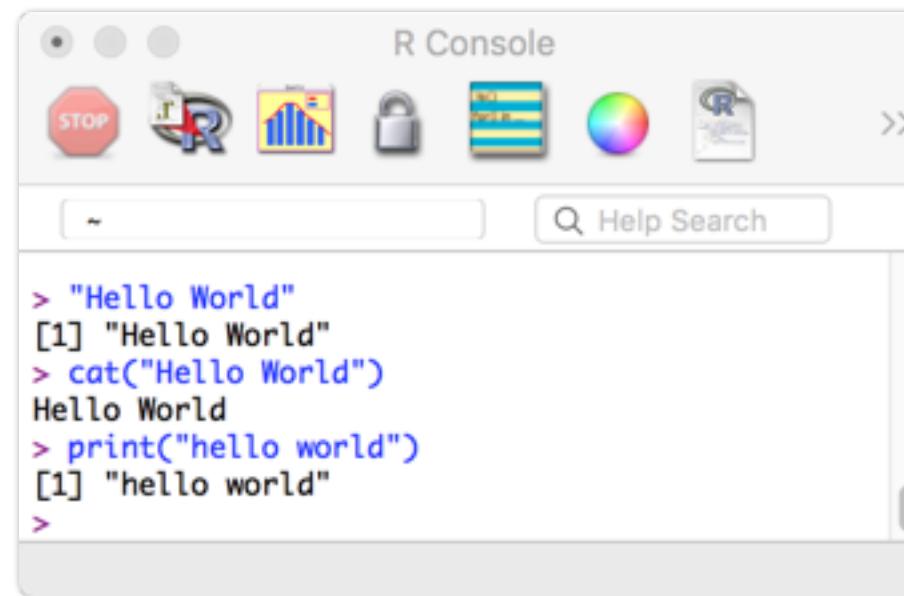
- <https://www.rstudio.com/products/rstudio/download/>



- <https://Bioconductor.org/>



# Exercise: Hello world



The image shows a screenshot of the R Console window. The title bar reads "R Console". Below the title bar is a toolbar with several icons: a red octagon labeled "STOP", a small R logo, a bar chart, a lock, a horizontal bar with vertical stripes, a color palette, and another R logo. To the right of the toolbar is a "Help Search" button. The main area of the window displays the following R session history:

```
> "Hello World"
[1] "Hello World"
> cat("Hello World")
Hello World
> print("hello world")
[1] "hello world"
>
```

# Where to start?

- From vignettes of a package  
<https://bioconductor.org/packages/4.0/bioc/vignettes/trackViewer/inst/doc/trackViewer.html#lolliplot>
- From help documents
- From a book

# Grammar of R

```
hairLength <- 25
waistHeight <- 110
height <- 168
marryMe <- FALSE
wait <- 100 * 365
for(day in 1:wait){
  if(hairLength >= height - waistHeight){
    marryMe <- TRUE
    break
  }else{
    if(runif(1) < 0.5){
      next
    }
    hairLength <- hairLength + .1
  }
}
day
```

- 待我长发及腰, When my hair is longer enough to make hair bun
- 少年娶我可好? Will you give me your promise ring?
- 待你青丝绾正, When you get your driver license,
- 铺十里红妆可愿? Will you roll out the red carpet for me?
- 却怕长发及腰, I am afraid before my hair grew till my waist,
- 少年倾心他人。 You will falling in love with another woman.
- 待你青丝绾正, I am afraid when you get your driver license,
- 笑看君怀她笑颜。 I see another woman burying herself in your arms.
- ——何晓道 《十里红妆女儿梦》

# Data structure of R

Dimension	Homogeneous	Heterogeneous
1d	Atomic vector	List
2d	Matrix	Data frame
Nd	Array	

Atomic vector: character, integer, logical, numeric, factor

Matrix, Array: atomic vector with dimension. Matrix is an array.

List: a chain for different kinds of data structures, including list.

Data frame: 2D version of List; All lists have same length of elements.

# Exercises

- What is the class of `c('A', TRUE, 1L, 1, factor(letters[1:3]))`
- What is the class of `c(TRUE, 1L, 1, factor(letters[1:3]))`
- What is the class of `c(TRUE, 1L, factor(letters[1:3]))`
- What is the class of `c(TRUE, factor(letters[1:3]))`
- ```
m <- array(1:18, dim=c(3, 3, 2))
i <- 1; j <- 3; k <- 2;
m[i, j, k]
m[(k-1)*(nrow(m)*ncol(m)) + (j-1)*nrow(m) + i]
```
- What is the difference between `as.numeric(as.character(factor(c(3, 5, 1))))` and `as.numeric(factor(c(3, 5, 1)))`
- What is the difference between `as.list(m)` and `as.list(as.data.frame(m))`

# How to get help

- `help(functionName)` or `?functionName` or `??functionName` : `help(help); help("for")`
- `browseVignettes("grid")`
- Google
- R mailing list: <https://www.r-project.org/mail.html#instructions>

# How to install a package

- Install R packages: `install.packages("BiocManager", repos="https://cloud.r-project.org")`
- Install Bioconductor packages:  
`BiocManager::install(c("org.Dr.eg.db", "TxDb.Drerio.UCSC.danRer10.refGene"))`

# Merge tables

- Question: how could we merge two excel sheet by gene symbols?

- Code:

```
m <- data.frame(symbol=letters[1:5], foldChange=runif(5))
n <- data.frame(gene=sample(letters[1:10], size=6), foldChange=runif(n=6))
z <- merge(m, n, by.x="symbol", by.y="gene", suffixes=c(".m", ".n"), all.x=TRUE, all.y=TRUE)
z
```

# Set operations

- Question: how to find the shared/unique genes from 2 lists of genes?

- Code:

```
A <- sample(letters[1:10], size=8)
B <- sample(letters[1:10], size=5)
intersect(A, B) ## shared genes
setdiff(A, B) ## only in A
setdiff(B, A) ## only in B
```

# How to read data from a file

- Tab-delimited file (tsv): `read.delim("file.txt")`
- Comma-delimited file (csv): `read.csv("file.csv")`
- Excel sheet (xls, xlsx): `library(gdata); read.xls("file.xlsx", sheet=1)`

# How to save a table

- Save to a csv file: `write.csv(data, "file.csv")`
- Save to a xls file: `library(WriteXLS); WriteXLS(dataframe, "file.xls")`

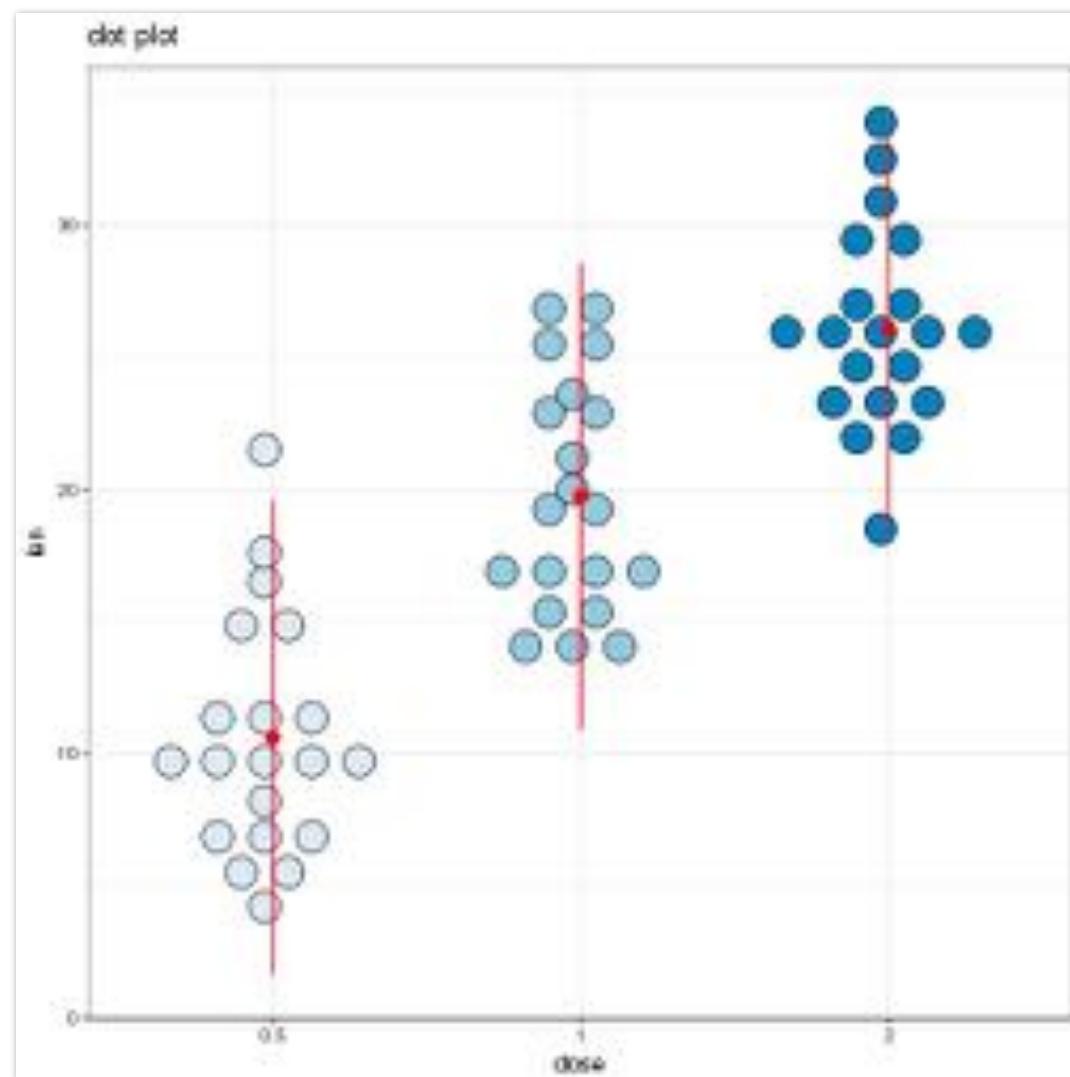
# Basics of R plot

- Dots plot: `plot(x=mtcars$wt, y=mtcars$mpg)`
- Bar plot: `counts <- table(mtcars$vs, mtcars$gear); barplot(counts)`
- Histogram: `hist(mtcars$vs)`
- Pie: `x<-table(mtcars$gear); pie(x)`
- Boxplot: `boxplot(mpg~cyl,data=mtcars)`

# ggplot2

Based on the grammar of graphics

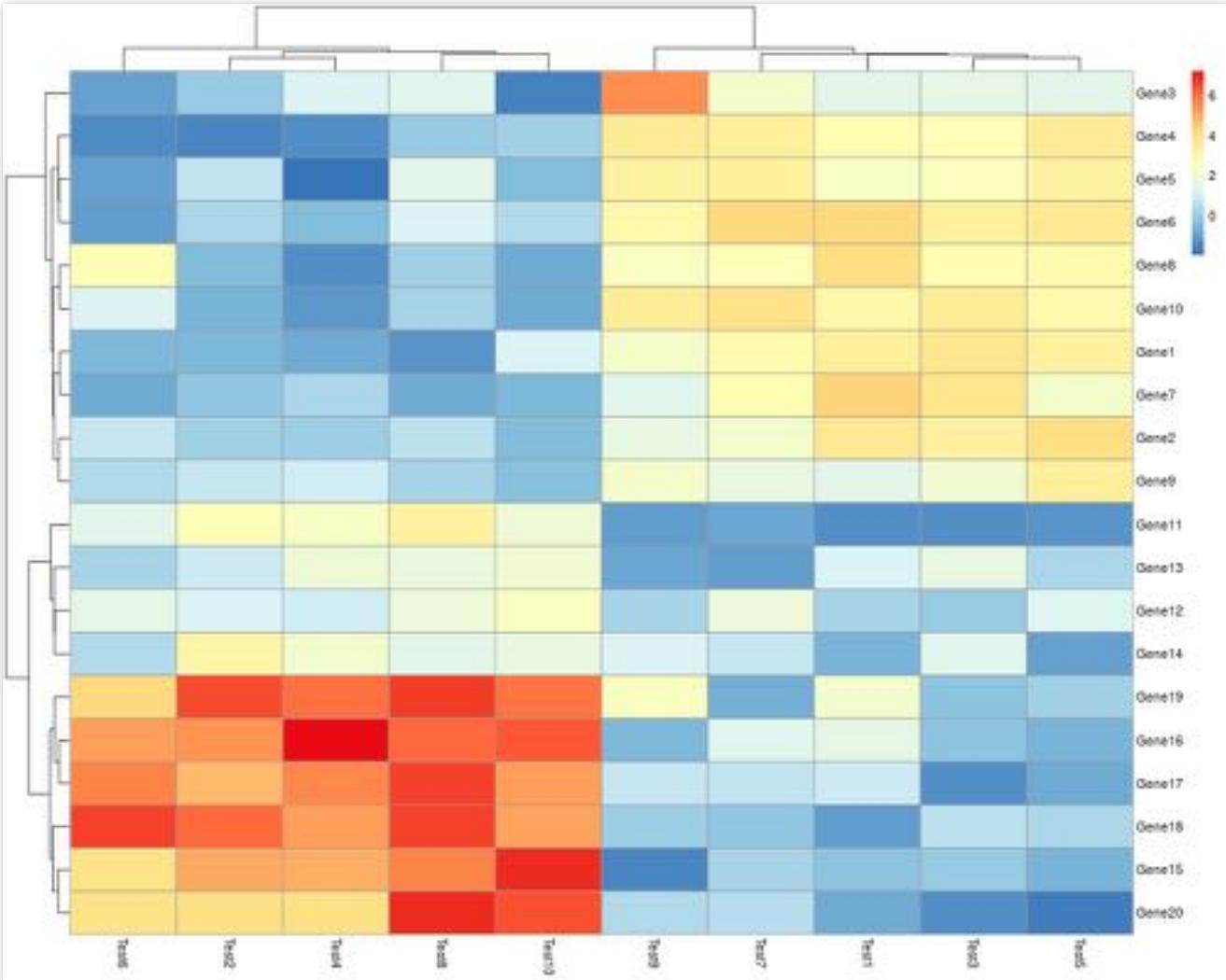
```
library(ggplot2)  
data("ToothGrowth")  
ToothGrowth$dose <- factor(ToothGrowth$dose)  
  
ggplot(ToothGrowth, aes(x=dose, y=len, fill=dose)) +  
  geom_dotplot(binaxis='y', stackdir='center',  
  stackratio=1.5, dotsize=1.2) +  
  stat_summary(fun.data=mean_sdl, mult=1,  
  geom="pointrange", color="red") +  
  scale_fill_brewer(palette="Blues", guide=FALSE) +  
  theme_bw() + labs(title="dot plot")
```



# pheatmap

## Pretty heatmap

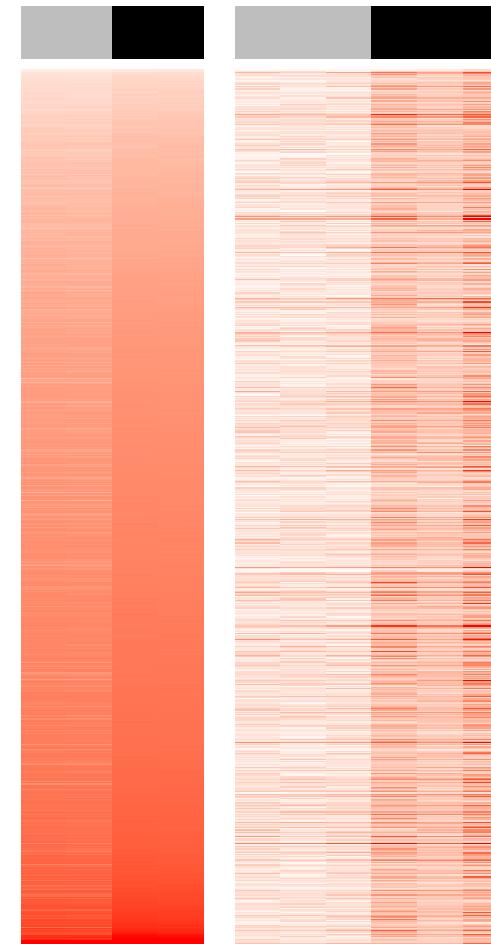
```
library(BiocManager)
install("pheatmap")
library(pheatmap)
test <- matrix(rnorm(200), 20, 10)
test[1:10, seq(1, 10, 2)] <- test[1:10, seq(1, 10, 2)] + 3
test[11:20, seq(2, 10, 2)] <- test[11:20, seq(2, 10, 2)] + 2
test[15:20, seq(2, 10, 2)] <- test[15:20, seq(2, 10, 2)] + 4
colnames(test) <- paste("Test", 1:10, sep = "")
rownames(test) <- paste("Gene", 1:20, sep = "")
pheatmap(test)
```



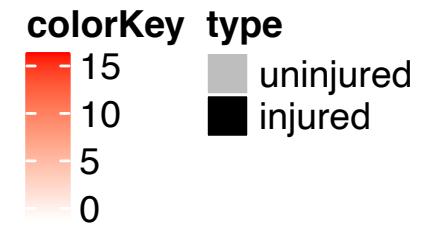
## ComplexHeatmap

Plot multiple heatmaps in one canvas  
rowAnnotation and colAnnotation

RNAseq ATACseq



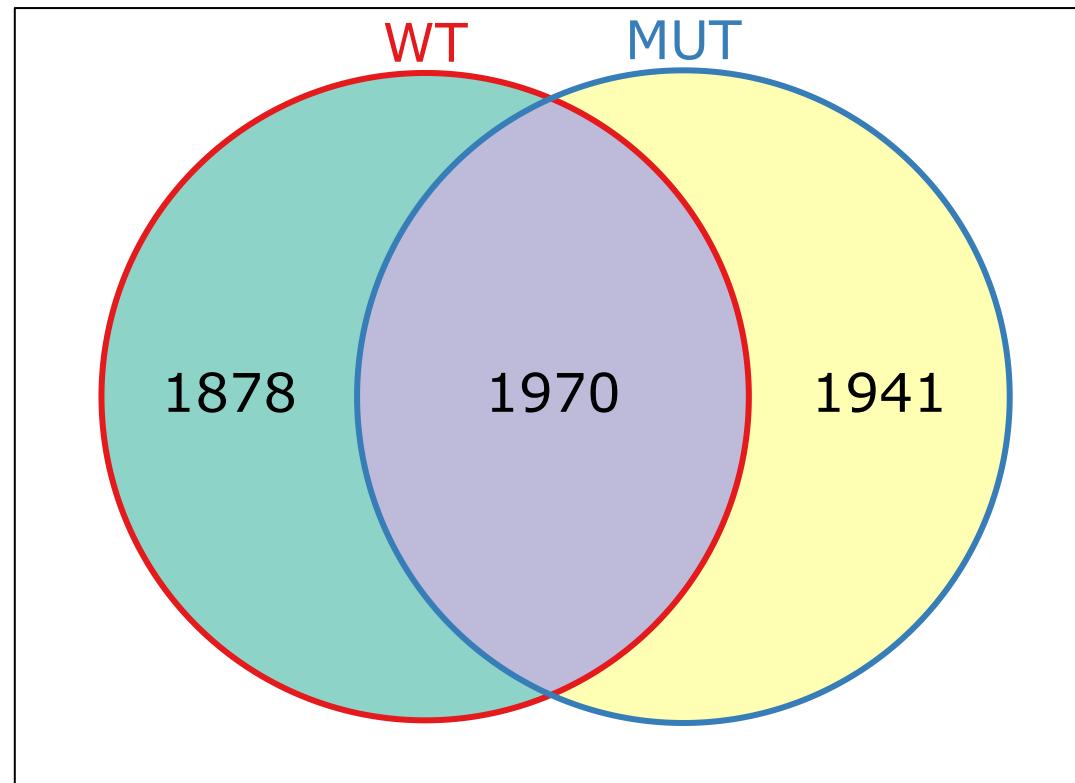
— fgf20a



— marcks1a  
— cx43  
— mdka

## Venn Diagram: Venerable

displays Venn and Euler diagrams



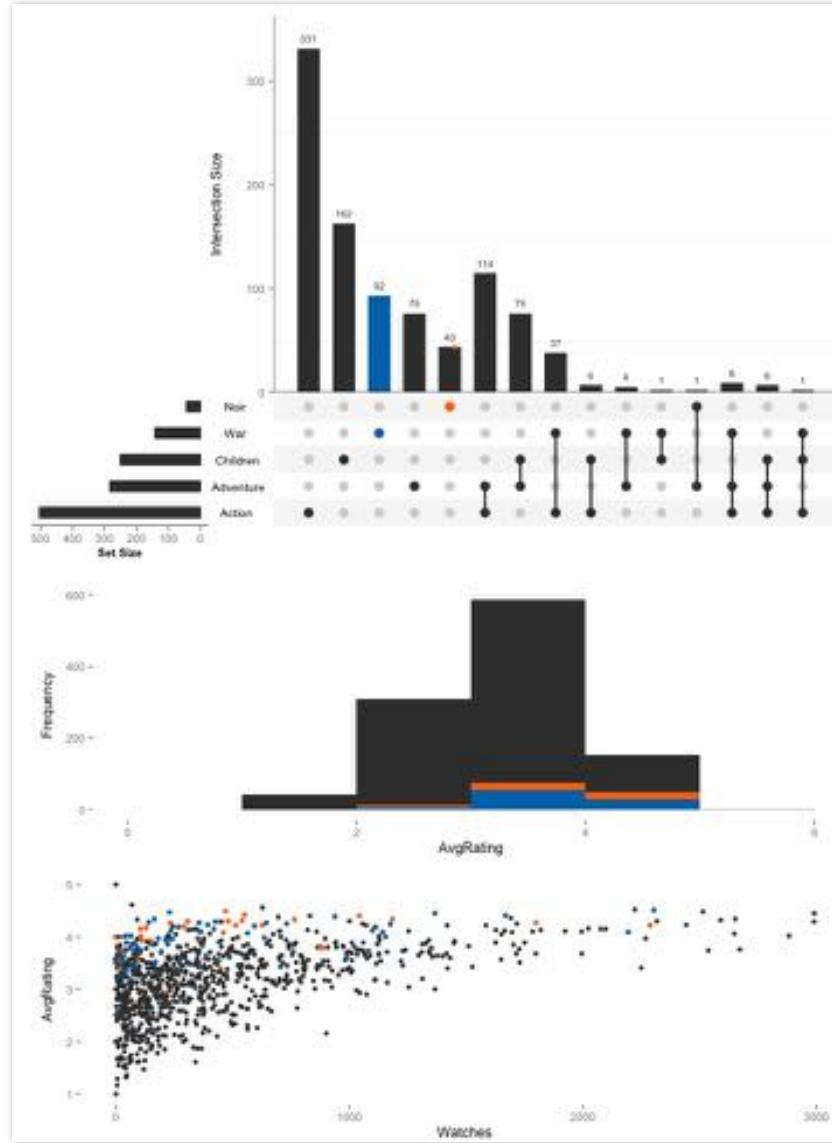
# UpSetR

visualizes set intersections in a matrix layout

```
install("UpSetR")
library(UpSetR)

movies <- read.csv(
  system.file("extdata", "movies.csv",
  package = "UpSetR"), header=TRUE,
  sep=";" )

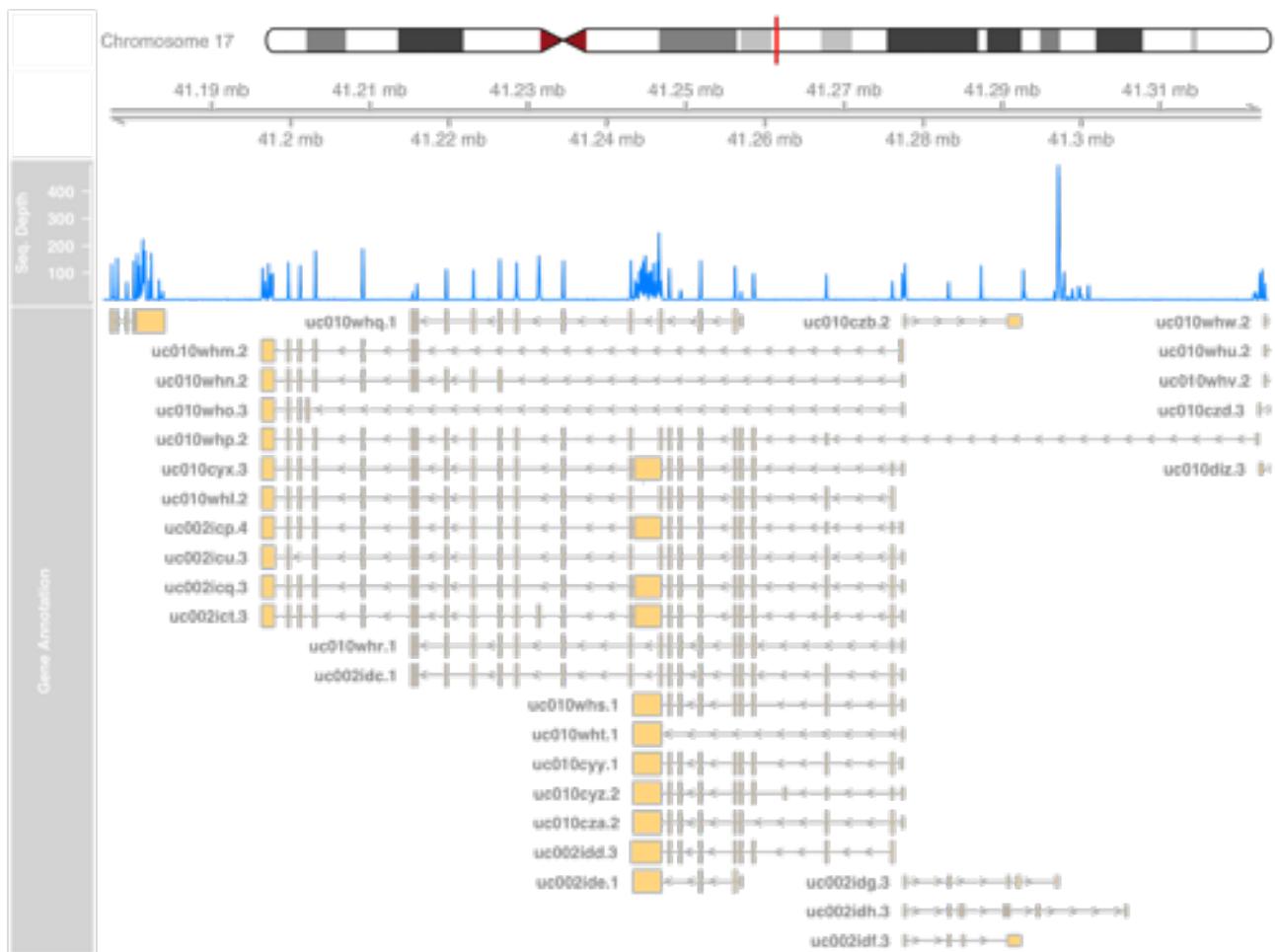
upset(movies, nsets = 7, nintersects =
  30, mb.ratio = c(0.5, 0.5), order.by =
  c("freq", "degree"), decreasing =
  c(TRUE, FALSE))
```



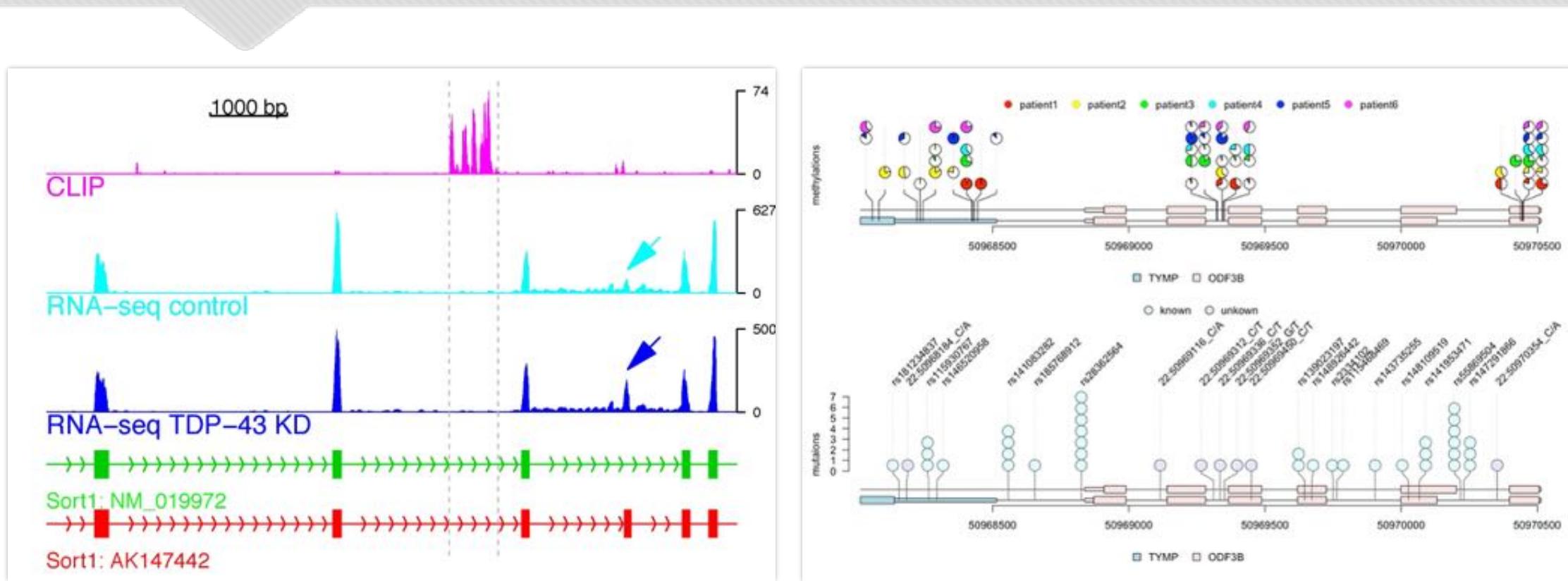
# Gviz

Plotting data and annotation information along genomic coordinates

```
library(Gviz)
library(GenomicRanges)
data(cpgIslands)
chr <- as.character(unique(seqnames(cpgIslands)))
gen <- genome(cpgIslands)
atrack <- AnnotationTrack(cpgIslands,
                           name="CpG")
gtrack <- GenomeAxisTrack()
data(geneModels)
grtrack <- GeneRegionTrack(geneModels,
                           genome=gen, chromosome=chr, name="Gene
                           Model")
plotTracks(list(gtrack, atrack, grtrack))
```

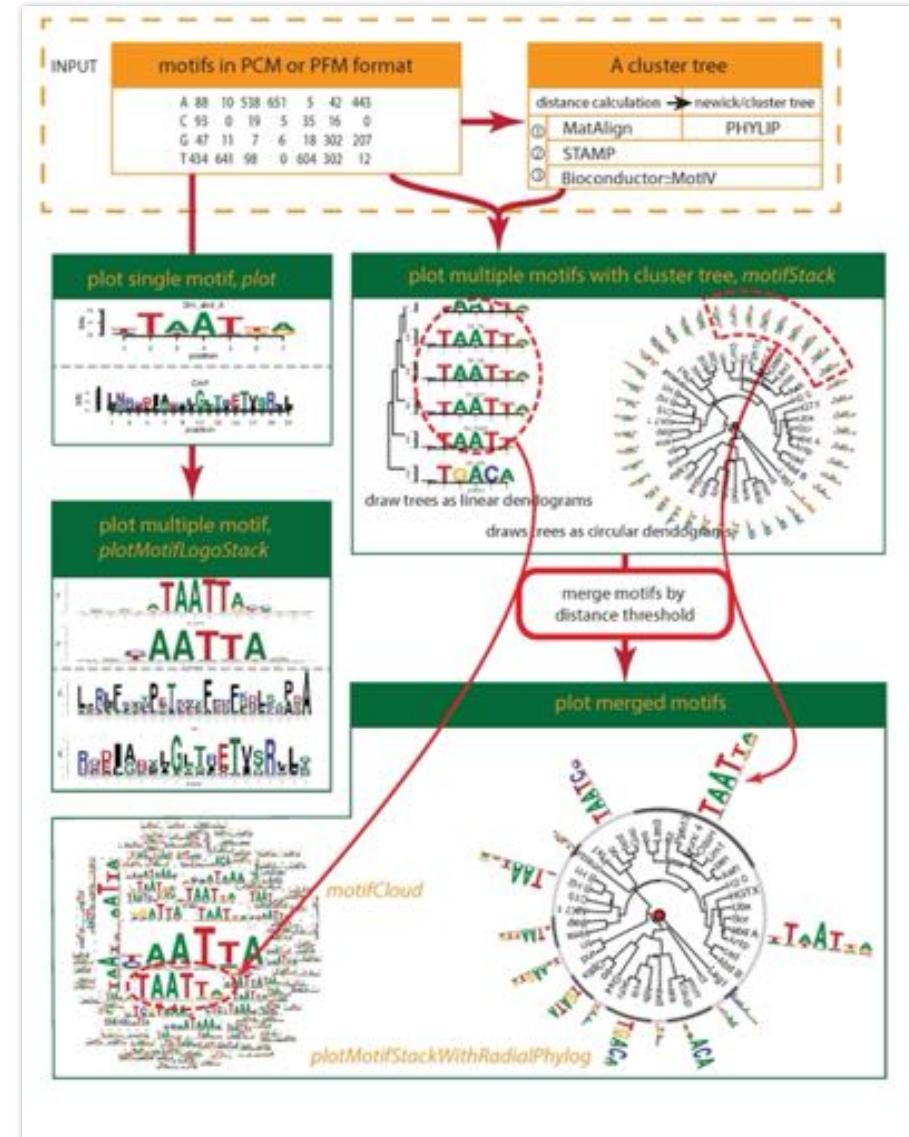


# trackViewer



# motifStack

Plot stacked logos for single or multiple DNA, RNA and amino acid sequence

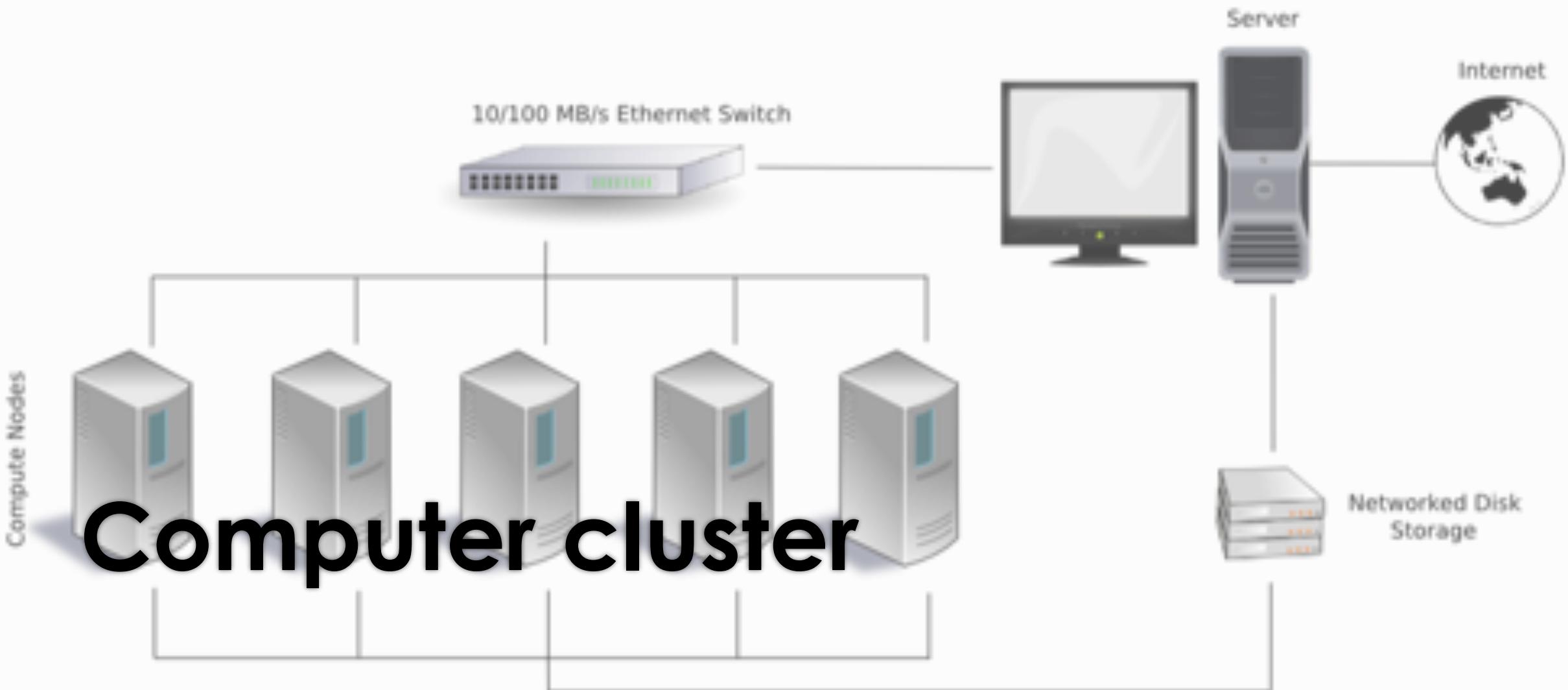


# Basics of Bioconductor objects

- GenomicRanges::GRanges:  
gr0 <- GRanges(Rle(c("chr2", "chr2", "chr1", "chr3"), c(1, 3, 2, 4)), IRanges(1:10, width=10:1))  
gr0
- SummarizedExperiment::RangedSummarizedExperiment:  
nrows <- 200; ncols <- 6  
counts <- matrix(runif(nrows \* ncols, 1, 1e4), nrows)  
rowRanges <- GRanges(rep(c("chr1", "chr2"), c(50, 150)),  
IRanges(floor(runif(200, 1e5, 1e6)), width=100),  
strand=sample(c("+", "-"), 200, TRUE),  
feature\_id=sprintf("ID%03d", 1:200))  
colData <- DataFrame(Treatment=rep(c("ChIP", "Input"), 3),  
row.names=LETTERS[1:6])  
rse <- SummarizedExperiment(assays=SimpleList(counts=counts),  
rowRanges=rowRanges, colData=colData)  
rse

# Basics of annotation

- annotationDBi: include ChipDb, OrgDb, GOdb, InparanoidDb and ReactomeDb
- BSgenome: The BSgenome class is a container for storing the full genome sequences of a given organism.
- GenomicFeatures: The GenomicFeatures package retrieves and manages transcript-related features from the UCSC Genome Bioinformatics and BioMart data resources.



# Cluster software

- Oracle Grid Engine, previously known as Sun Grid Engine (SGE):`qsub`
- IBM Platform Load Sharing Facility (LSF):`bsub`
- Slurm workload manager:`sbatch`
- More: [https://en.wikipedia.org/wiki/Comparison\\_of\\_cluster\\_software](https://en.wikipedia.org/wiki/Comparison_of_cluster_software)

# The Duke compute cluster

- Website: <https://rc.duke.edu/dcc/>
- Next class: **8/4 OR 9/1**, <https://rc.duke.edu/dcc-training/>
- DCC use slurm workload manager.

# Accessing the DCC

- Login: ssh NetID@dcc-slogin.oit.duke.edu
- Copy files: scp, rsync
  - scp -r sourceFolder netid@dcc-slogin.oit.duke.edu:/work/netid/targetFolder
  - rsync -av sourceFolder netid@dcc-slogin.oit.duke.edu:/work/netid/targetFolder

# DCC file systems

- /dscrhome: 250 GB group quota
- /work: temporary storage
- /datacommons: archival storage; available for \$80/TB/year

# module

- module avail
- module load
  - which STAR
  - module load STAR/2.5.3a
  - which STAR
  - echo \$PATH

# Submit a job

- **srun:**
  - Interactive mode: srun --pty bash -i
  - Run a script: srun --mem 1G /bin/hostname
- **sbatch:**
  - `#!/bin/bash`
  - `#SBATCH -J mapping #jobname`
  - `#SBATCH --array=1-2 # job array, can be 1,3,5,7 or 1-7:2 '%' will limit the number of job`
  - `#SBATCH -o pindel.out.%A_%a.txt # %A == $SLURM_ARRAY_JOB_ID #log folder must be there`
  - `#SBATCH -e pindel.err.%A_%a.txt # %a == $SLURM_ARRAY_TASK_ID`
  - `#SBATCH --mem-per-cpu=5G #5G`
  - `#SBATCH -c 1 # cpus-per-task`
  - `#SBATCH --partition=common`
  - `i=$((($SLURM_ARRAY_TASK_ID-1))`
  - `echo $i > job.id.txt`

# DCC partitions

- **common**, for jobs that will run on the DCC core nodes (up to 64 GB RAM).
- **common-large**, for jobs that will run on the DCC core nodes (64-240 GB RAM).
- **gpu-common**, for jobs that will run on DCC GPU nodes.
- **Group partitions** (partition name varies), for jobs that will run on lab-owned nodes

# Slurm commands

- `sbatch` – Submit a batch job
- `squeue` – show list of jobs
- `scancel` – delete one or more jobs



## Additional information

# Basic tools

- [Bedtools](#)
- [Samtools](#)
- [Picard](#)
- [Ucsc genomic tools](#)

# Samtools

- SAM, BAM file convert:
  - samtools view -h file.bam > file.sam
  - samtools view -b -S file.sam > file.bam
- Sorting BAM
  - samtools sort file.bam -o outputbam
- Create BAM index
  - samtools index file.bam

# Online resources

- [cBio Cancer Genomics Portal](#) (cBioPortal) & package [cgdsr](#)
- [The Cancer Genome Atlas \(TCGA\)](#) & package [TCGAbiolinks](#), [RTCGAToolbox](#)
- [Gene Expression Omnibus \(GEO\)](#) & package [GEOquery](#), [SRAdb](#)
- [Encyclopedia of DNA Elements \(ENCODE\)](#) & package [ENCODEExplorer](#)
- [recount2](#) & package [recount](#)

# Use cgdsr package download data

```
● BiocManager::install("cgdsr")
● library(cgdsr)
● # Create CGDS object
● mycgds <- CGDS("https://www.cbioportal.org/")
● # Test the connection
● test(mycgds)
● # Get list of cancer studies at server
● getStudies(mycgds)[, 1:2]
● # Get available case lists (collection of samples) for a
  given cancer study
● (mycancerstudy <- getStudies(mycgds)[2,1])
● getCaseLists(mycgds, mycancerstudy)
● mycaselist <- getCaseLists(mycgds, mycancerstudy)[1, 1]
● # Get available genetic profiles
● getGeneticProfiles(mycgds, mycancerstudy)
● (mygeneticprofile <-
  getGeneticProfiles(mycgds, mycancerstudy)[2,1])
● # Get data slices for a specified list of genes, genetic
  profile and case list
● getDataSlices(mycgds, c('BRCA1', 'BRCA2'), mygeneticprofile,
  mycaselist)
● # Get clinical data for the case list
● myclinicaldata <- getClinicalData(mycgds, mycaselist)
● dim(myclinicaldata)
● head(myclinicaldata)
```

# Download expression data from recount2

```
## in shell apt-get update && apt-get install libxml2-dev  
BiocManager::install("recount")  
library('recount')  
## Find a project of interest  
project_info <- abstract_search('GSE32465')  
## Explore info  
project_info  
## Download the gene-level RangedSummarizedExperiment data  
download_study(project_info$project)  
## Load the data  
load(file.path(project_info$project, 'rse_gene.Rdata'))  
## Delete it if you don't need it anymore  
unlink(project_info$project, recursive = TRUE)  
rse_gene  
## the sample phenotype data provided by the recount project  
head(colData(rse_gene))  
## the sample feature data provided by the recount project  
rowData(rse_gene)
```

- ## Scale counts by taking into account the total coverage per sample
- rse <- scale\_counts(rse\_gene)
- (title <- colData(rse)\$title)
- (group <- ifelse(grepl("uninduced", title), "uninduced", "induced"))
- (target <- sub("^.\*targeting (.\*) gene.\*\$", "\\\1", title))
- ## Add sample information for DE analysis
- colData(rse)\$group <- factor(group)
- colData(rse)\$gene\_target <- factor(target)
- library(DESeq2)
- ## Specify design and switch to DESeq2 format
- dds <- DESeqDataSet(rse, ~ gene\_target + group)
- ## Perform DE analysis
- dds <- DESeq(dds, test = 'LRT', reduced = ~ gene\_target, fitType = 'local')
- res <- results(dds)
- ## Explore results
- head(res)
- library(WriteXLS)
- WriteXLS(as.data.frame(res), "DE.results.GSE32465.xls")

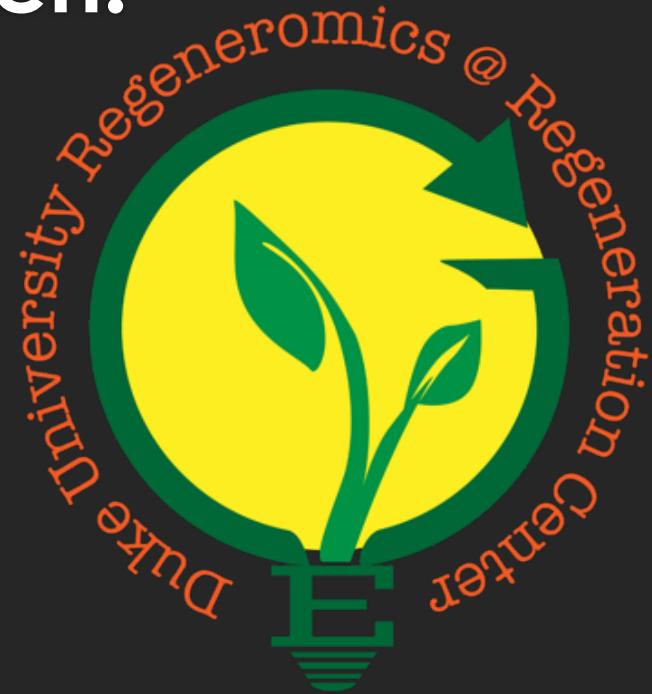
# DESeq2

```
library(DESeq2)
load(url("http://qiubio.com/bioconductor/RNA-seq/ds1.Rdata"))
ls()
head(counts[, 1:7], 3)
grp <- as.factor(substr(colnames(counts), 1, 2)) ##assign group
dds <- DESeqDataSetFromMatrix(counts,
                                colData=data.frame(grp),
                                design=formula(~1+grp))
design(dds)
dds <- DESeq(dds)
res <- results(dds)
res
```

```
mcols(res, use.names=TRUE)
plotMA(res)
plotDispEsts(dds)
## save results
library(WriteXLS)
WriteXLS(as.data.frame(res), "results.xls")
## try comparison of other groups
results(dds, contrast = c("grp", "PE", "DE"))
## question, how to compare group PE vs FE
```

# Regeneromics Shared Resource can help your research!



We do: experimental design, bioinformatics analysis, manuscript preparation, grant applications



Jianhong Ou, Ph.D.

Email: regeneromics@duke.edu

<https://sites.duke.edu/dukeregenerationcenter/jobsrni/>