UCLA Institute for Quantitative and Computational Biology

March 2020

# QCB Workshop W5a: RNAseq-1 (day 1)

**Nicolas Rochette**

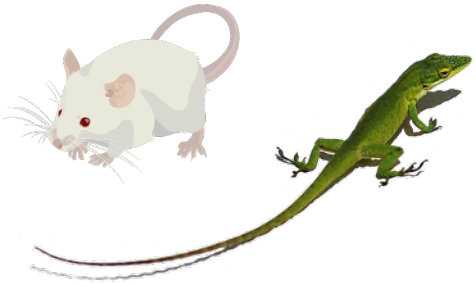
(EEB/ISG)

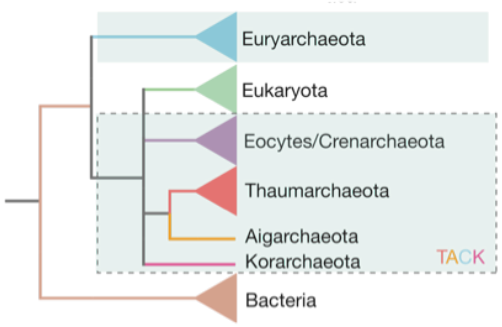
Karolina Kaczor-Urbanowicz

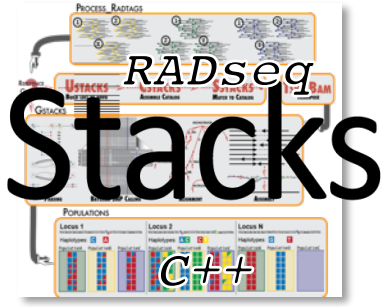
(Oral Biology & Medicine)

# My background

*Peromyscus maniculatus*






https://commons.wikimedia.org/wiki/File:201410_BALB.png
https://commons.wikimedia.org/wiki/File:Anolis_smaragdinus.jpg

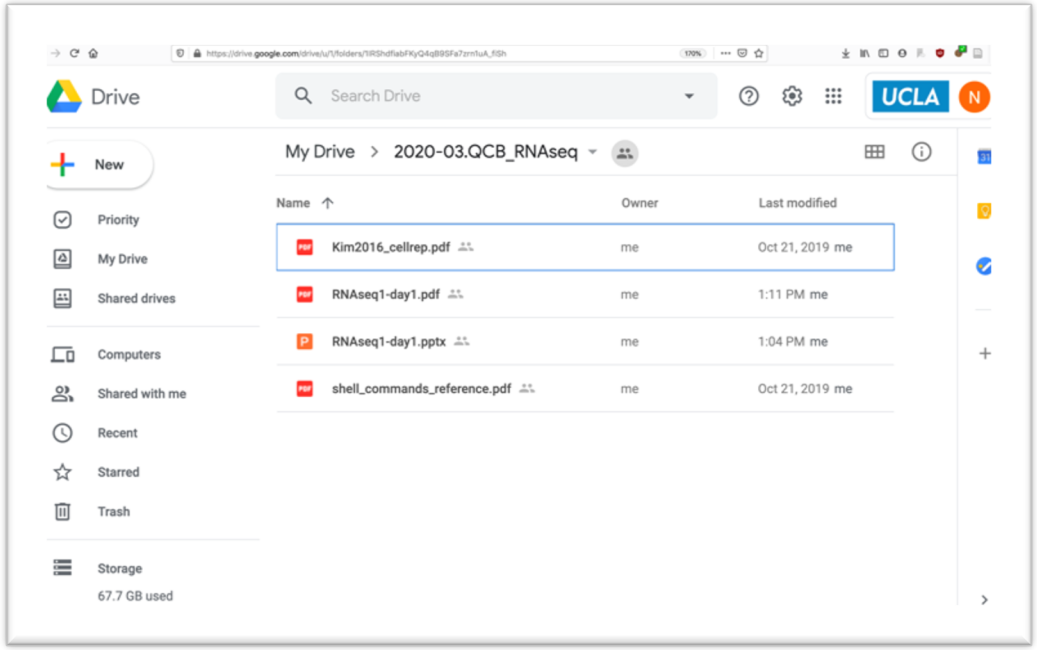
# Workshops' google drive folder

Slides & supporting documents:

https://drive.google.com/drive/folders/1IRShdfiabFKyQ4qB9SFa7zrn1uA_fiSh

https://tinyurl.com/**sshouvw**

Please bookmark it / write it down!



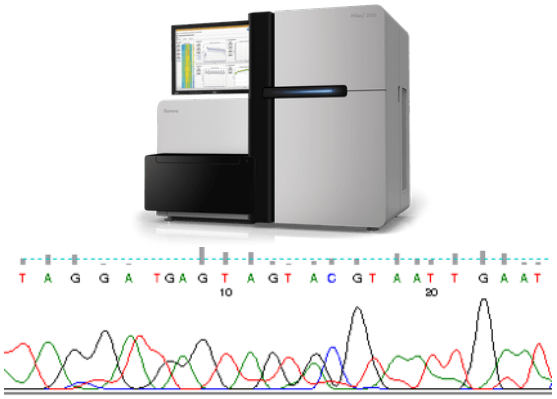
# Workshop Overview

# RNAseq workshops 1 & 2

- **RNAseq I: "from raw sequence data to gene counts"**
  1. Overview of RNAseq
     Workshop data: *Kim JW et al. (2016), Dev Cell 37(6)*
     Raw sequence reads: format, quality controls, filtering
  2. Alignments to an annotated genome: formats & software
  3. Counting gene/exon abundancies, visualizing alignments
- **RNAseq II: "from gene counts to biology"**
  - Normalization of expression counts
  - Visualization of gene expression levels (heatmaps, PCA)
  - Test for differentially expressed genes, functional enrichments

# RNAseq workshops 1 & 2

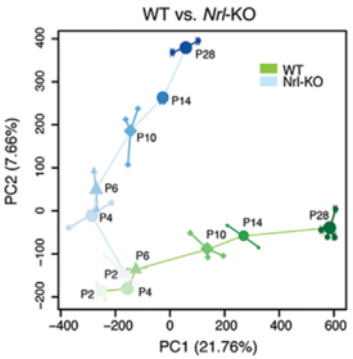
- RNAseq I: “from raw sequence data to gene counts”

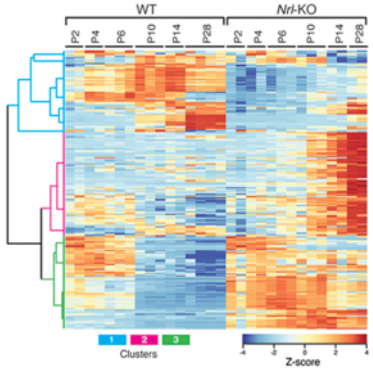


| | sample1 | sample2 | sample3 | … |
|---|---|---|---|---|
| gene1 | 999 | 701 | 616 | |
| gene2 | 532 | 520 | 41 | |
| gene3 | 14 | 36 | 305 | |
| … | | | | |

- RNAseq II: “from gene counts to biology”

| | sample1 | sample2 | sample3 | … |
|---|---|---|---|---|
| gene1 | 999 | 701 | 616 | |
| gene2 | 532 | 520 | 41 | |
| gene3 | 14 | 36 | 305 | |
| … | | | | |



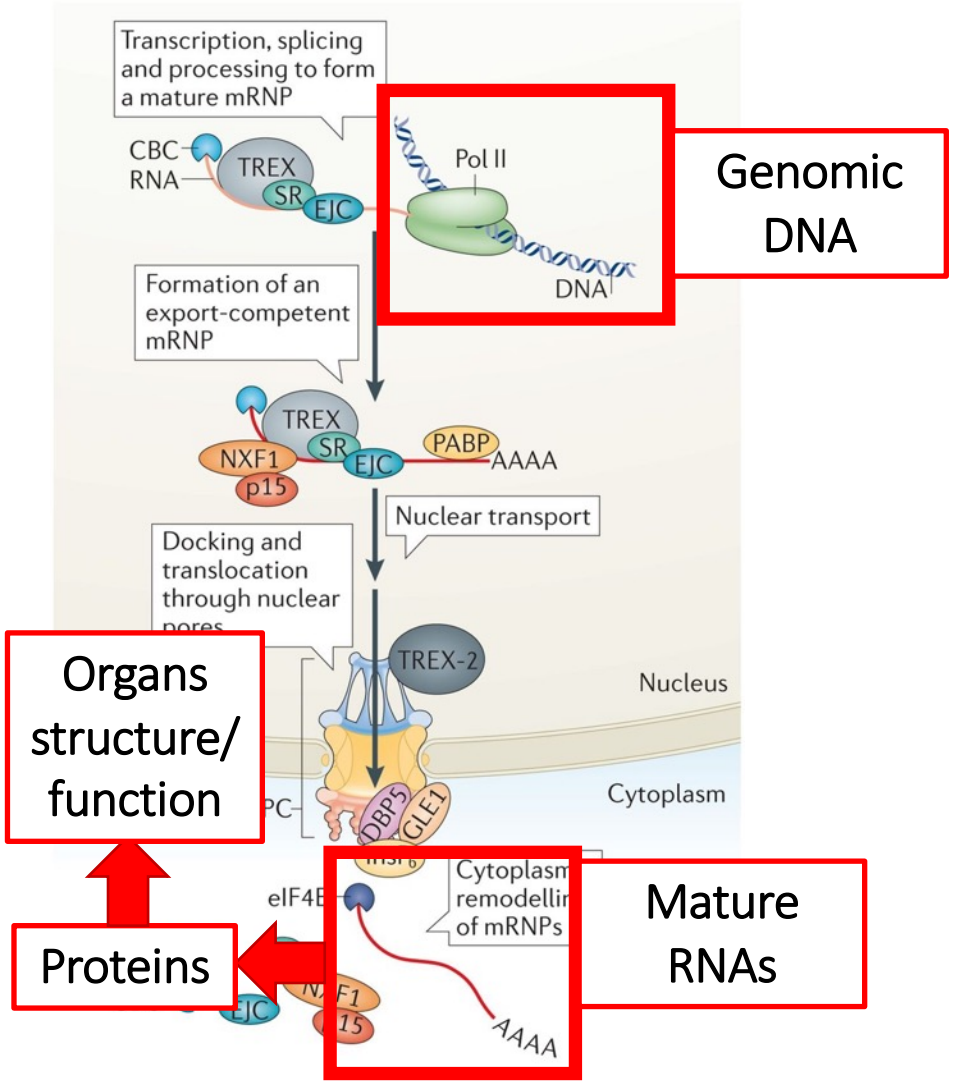



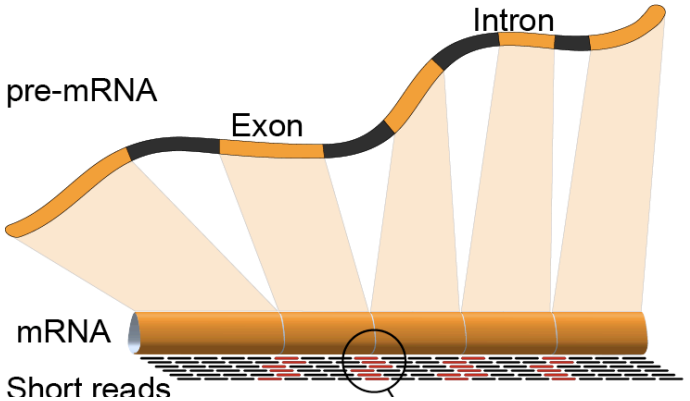
# Today

- Introduction to RNAseq analysis
- Workshop data & biological context
- Working on the Hoffman2 server
- Raw sequence data:
  - General properties
  - Format
  - Quality controls, filtering

# Introduction to RNAseq analysis

# *What is RNA-seq?*



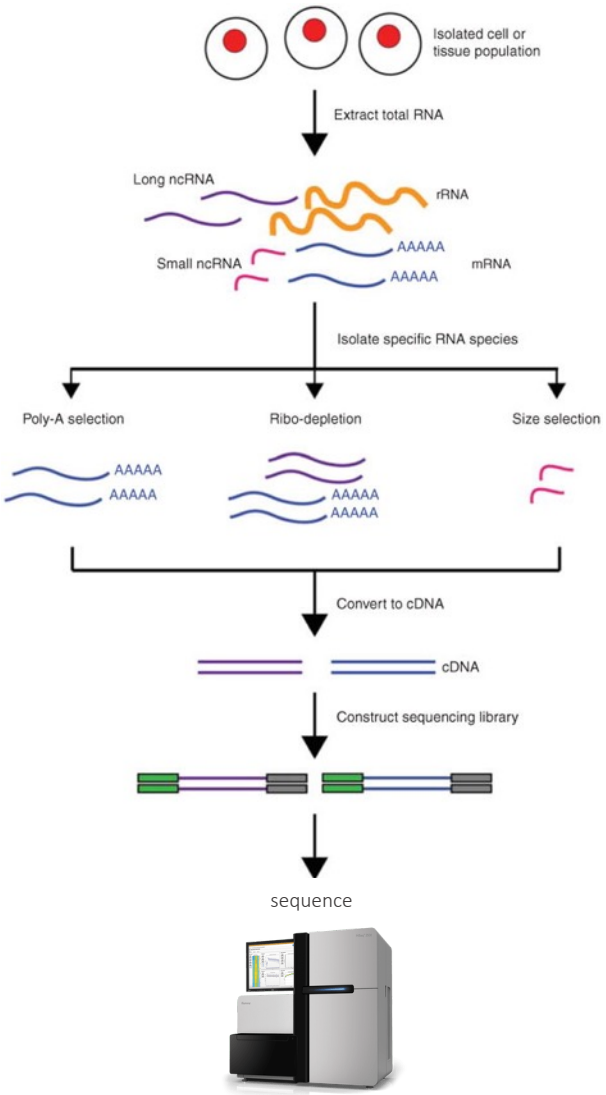
# RNA-seq: Purposes

Functional genomics: high-throughput data on molecular phenotypes, such as:

epigenetics, **transcriptomics (RNA-seq)**, proteomics, metabolomics…

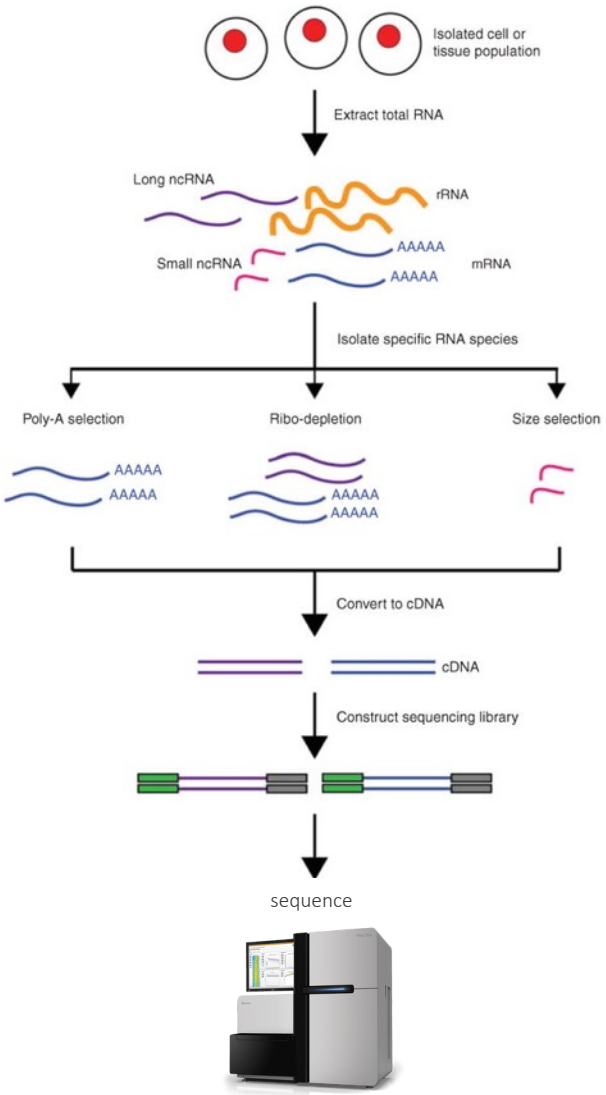
- Study of **genotype ↔ phenotype** mapping
- Discovery/**annotation of genes & transcripts**
- **Which genes are expressed in which tissues** & how much?
- Gene **regulatory networks** (co-regulated genes)
- Expression variation in **response to environmental stimuli**
- Medical **diagnostic**

# RNA-seq protocol




- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
  Gaida SM et al. (2015) Nat Comm 6:7045

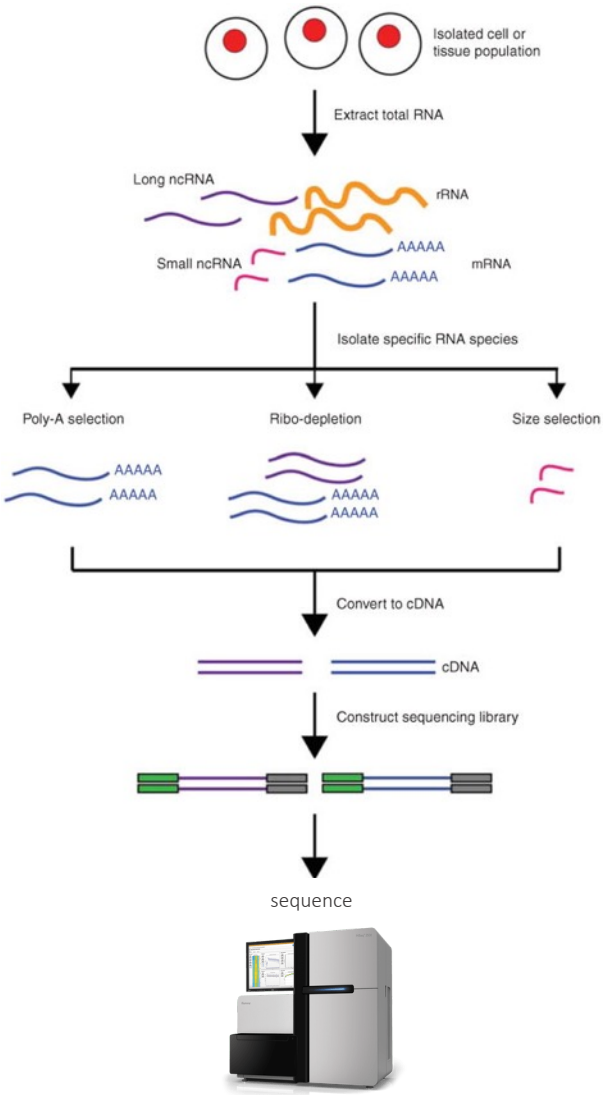
# RNA-seq protocol




- Whole-tissue, single-cell, cell type-specific, nuclei…

- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
- Gaida SM et al. (2015) Nat Comm 6:7045

# RNA-seq protocol



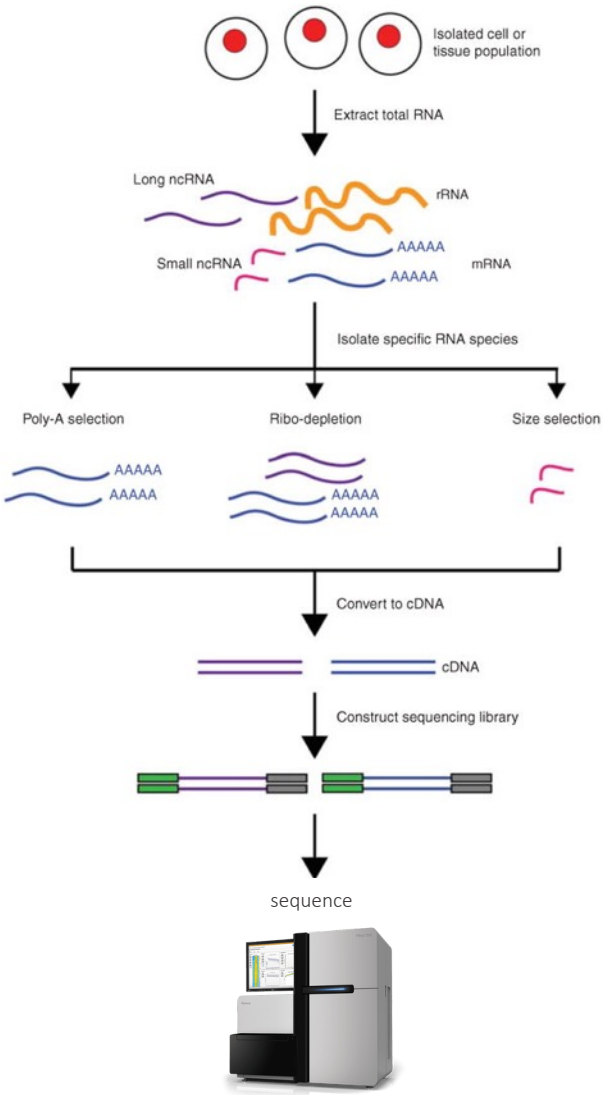

➢ RNA selection

RNAs come in many flavors:

- ribosome-related RNAs:
  *rRNA, tRNA, snoRNA* Up to 90% of RNAs…
- protein-coding RNAs:
  *mRNA (poly-adenylated), pre-mRNA*
- regulatory RNAs:
  *miRNA, siRNA, piRNA, lncRNA, tdRNA…*

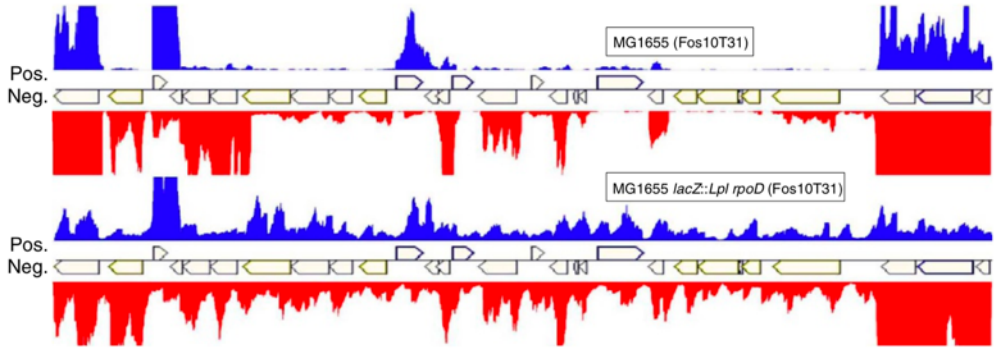
- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
  Gaida SM et al. (2015) Nat Comm 6:7045

# RNA-seq protocol



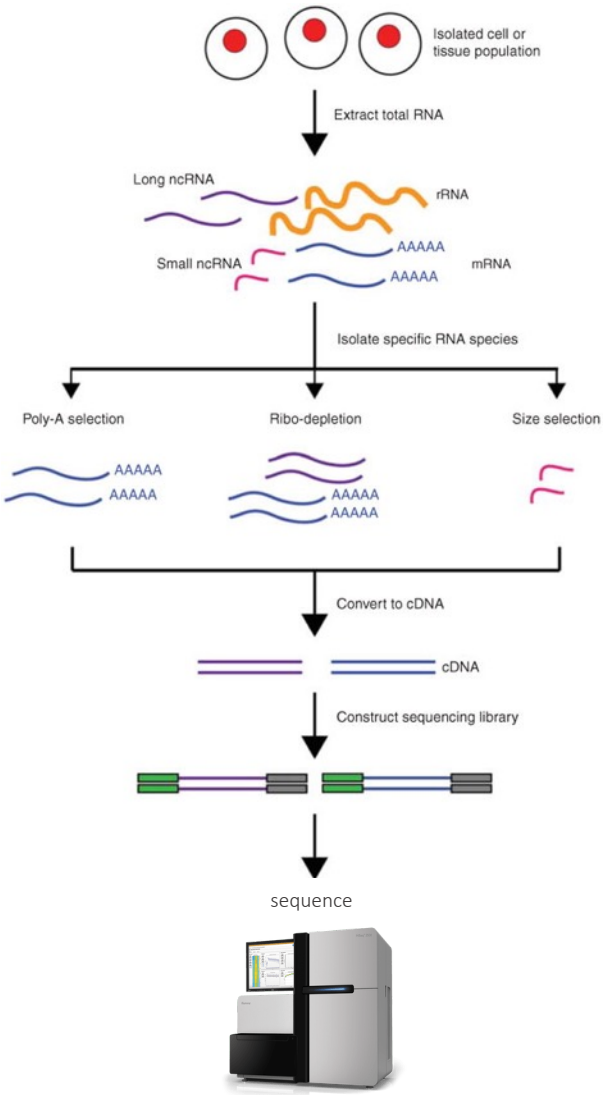

- Reverse transcription:

  **Stranded or unstranded**




- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
  Gaida SM et al. (2015) Nat Comm 6:7045

# RNA-seq protocol




- Sequencing technology:
  - Illumina: **short** reads; **single-end** or **paired-end**
  - Pacbio, Nanopore: **long** reads

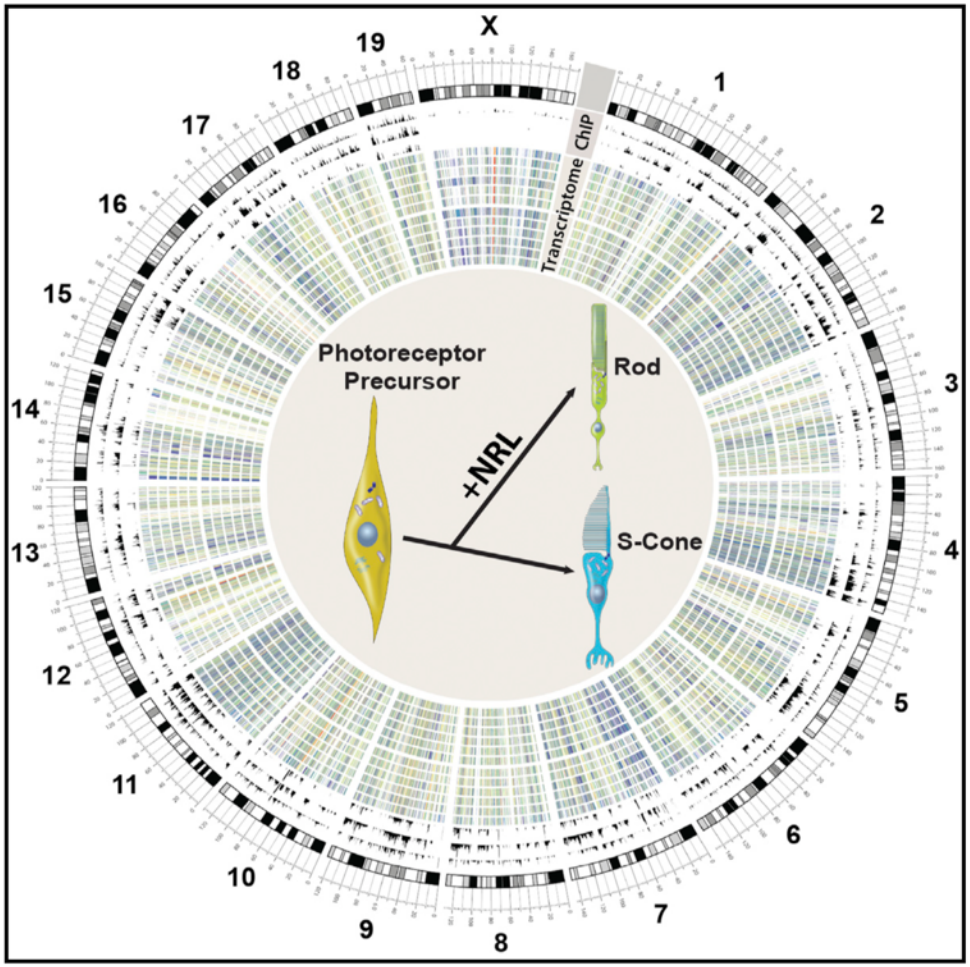
- Kukurba, Mongomery (2015) Cold Spring Harb. Protoc.
  Gaida SM et al. (2015) Nat Comm 6:7045

# The Data

*Kim JW et al. (2016), Dev Cell 37(6)*
*Kim JW et al. (2016) Cell Reports, 17(9)*

Kim JW et al. (2016)

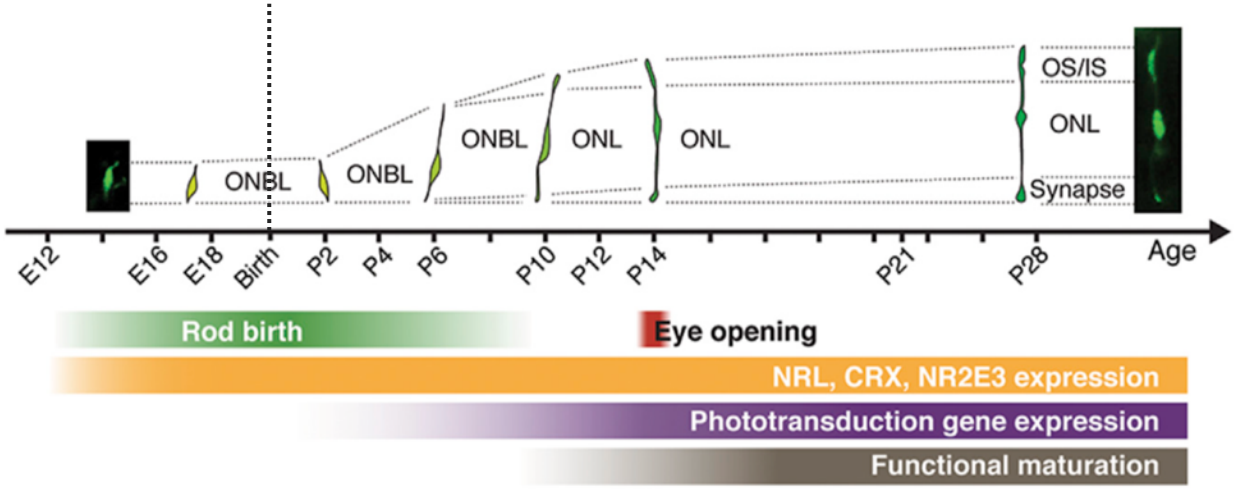
# NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors.

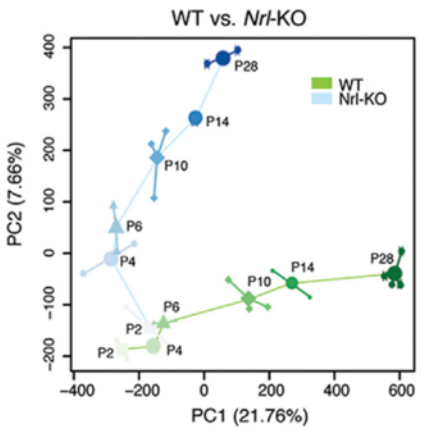


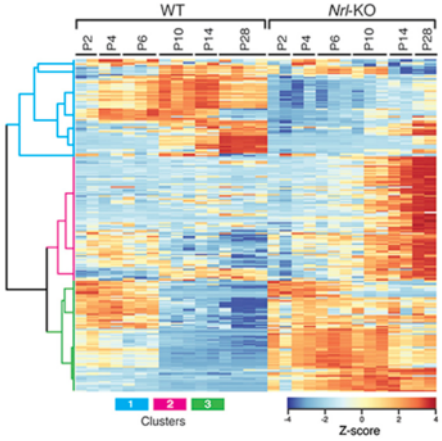

https://commons.wikimedia.org/wiki/File:Mouse-19-Dec-2004.jpg

Kim JW et al. (2016). **NRL-Regulated Transcriptome Dynamics of Developing Rod Photoreceptors.** *Cell Reports*, *17*(9).

- Analyzed the gene regulatory network activated by the NRL transcription factor, across developmental stages
- Gene expression with “stranded total RNAseq”

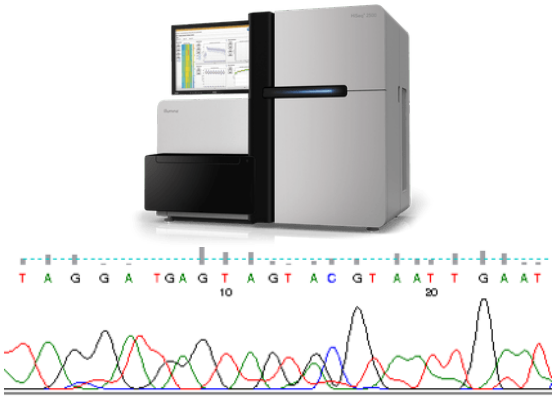


# RNAseq workshops 1 & 2

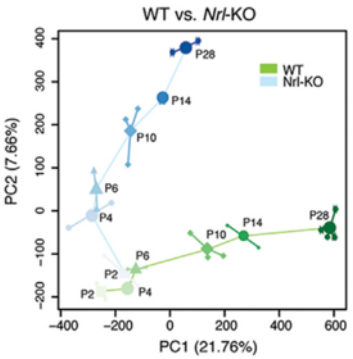
- RNAseq I: “from raw sequence data to gene counts”

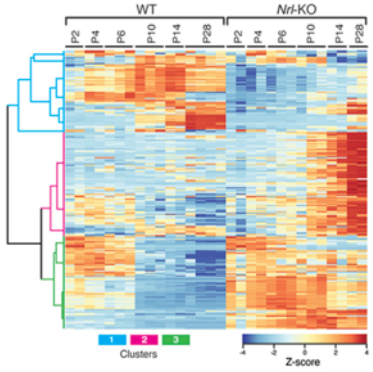


| | sample1 | sample2 | sample3 | … |
|---|---|---|---|---|
| gene1 | 999 | 701 | 616 | |
| gene2 | 532 | 520 | 41 | |
| gene3 | 14 | 36 | 305 | |
| … | | | | |

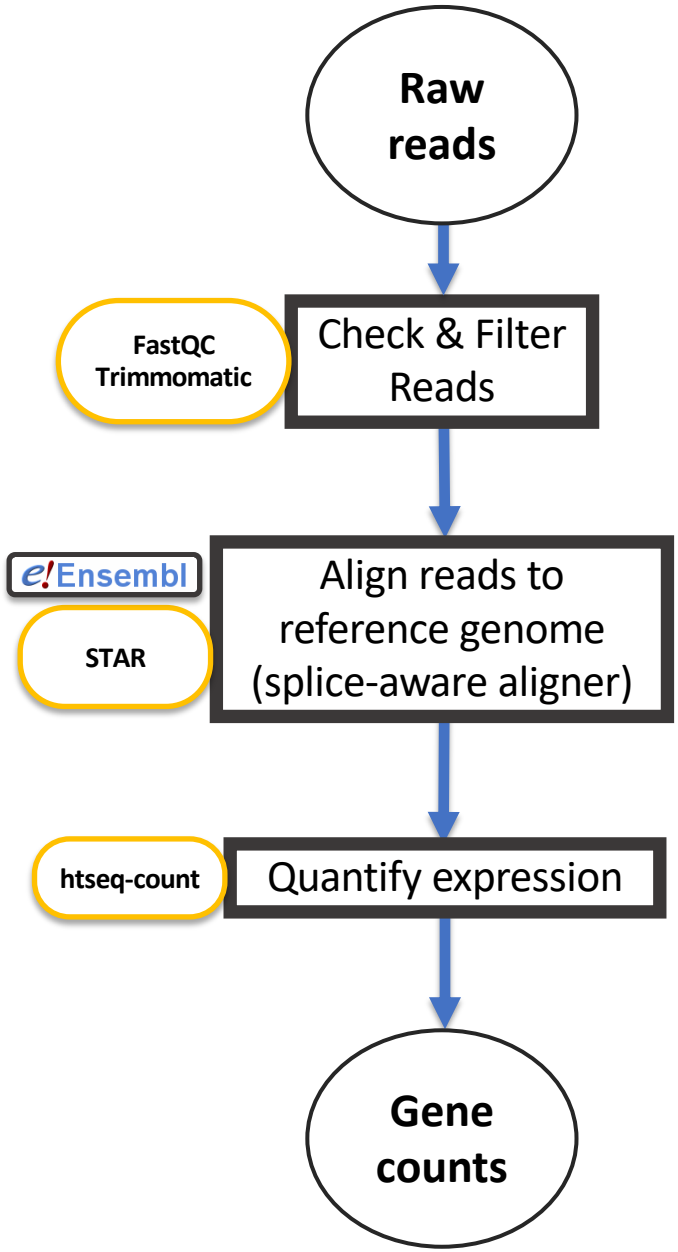
- RNAseq II: “from gene counts to biology”

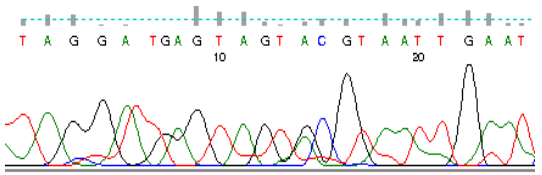
| | sample1 | sample2 | sample3 | … |
|---|---|---|---|---|
| gene1 | 999 | 701 | 616 | |
| gene2 | 532 | 520 | 41 | |
| gene3 | 14 | 36 | 305 | |
| … | | | | |

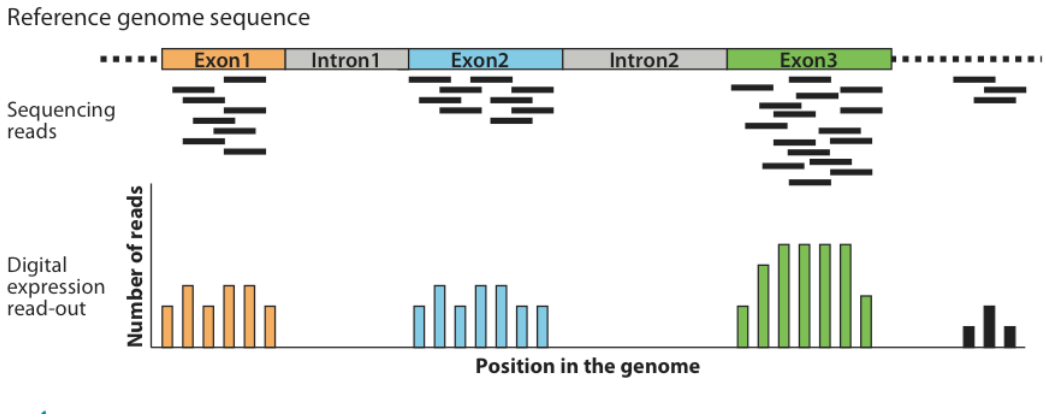


# RNA-seq analysis: Overview



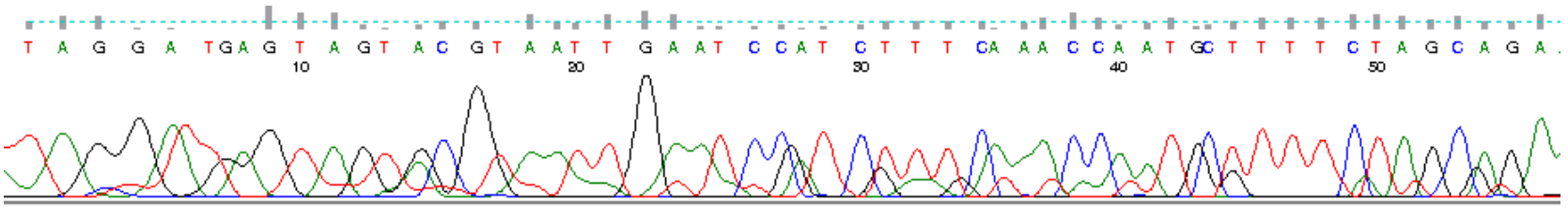




| | sample1 | sample2 | sample3 | ... |
|---|---|---|---|---|
| gene1 | 999 | 701 | 616 | |
| gene2 | 532 | 520 | 41 | |
| gene3 | 14 | 36 | 305 | |
| ... | | | | |

# The raw output of sequencing

# What is a read?




**Sanger**

One chromatogram = 1 'read':

1. Sequence
2. Per-basepair confidence

**Illumina**

- Similar in principle…
  …but higher-throughput & fully automated
- Typical RNAseq:
  **~10M reads of ~100bp each (=1Gbp's), per sample**
  ⇒ Larges files (even with compression)
  *e.g.* for Kim et al. 2016, the raw data alone is **53 GB**'s

# The FASTQ format

- "Plain text" format (could have a ".txt" extension)
- Usually compressed with GZIP (".fastq.gz")

```
@J00153:333:HF2VNBBXY:1:1101:1347:1086 1:N:0:NGCTCATT+NAGGACGT
ACCGGATTGCATCCCCAATCTGCTGTTTGTTTTCCACTGCGTACGTTCC
+
9<AFAJAFJJJJJJJJJ7AA-<FFF-<F-<FJJAJJJJJJJ-FFJFFJ<
@J00153:333:HF2VNBBXY:1:1101:2402:1086 1:N:0:NGCTCATT+NAGGACGT
CTGGTGGTGCCCTTCCGTCAATTCCTTTAAGTTTCAGCTTTGCAACCAT
+
BAAFAJJFJJJJJJFJJJJ<A<<AJJFFJJJJJFJJJ-F7JJJJFJJJJ
(etc.)
```

ID
Seq
+
Qual

Read 1

Read 2

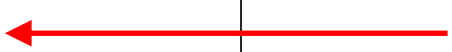
*etc.*

One read = 4 lines:

1. Read ID + properties (forward/reverse, barcodes)
2. Read sequence
3. (spacer line)
4. Base qualities according to the sequencer

# Base qualities: “PHRED” scale

```
@J00153:333:HF2VNBBXY:1:1101:1347:1086 1:N:0:NGCTCATT+NAGGACGT
ACCGGATTGCATCCCCAATCTGCTGTTTGTTTTCCACTGCGTACGTTCC
+
9<AFAJAFJJJJJJJJJ7AA-<FFF-<F-<FJJAJJJJJJJ-FFJFFJ<
```



Logarithmic scale to write **confidence probabilities** as integers:

$$Prob(WrongBase) = 10^{-Score*0.1}$$
$$Prob(CorrectBase) = 1 - Prob(WrongBase)$$

| Score | Error rate | Confidence |
|---|---|---|
| 10 | $10^{-1} = 0.1$ | 90% |
| 20 | $10^{-2} = 0.01$ | 99% |
| 30 | $10^{-3} = 0.001$ | 99.9% |
| 40 | $10^{-4} = 0.0001$ | 99.99% |

# ASCII encoding of PHRED scores

***ASCII codes** (1-byte characters)*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0 | <NUL> | 32 | <SPC> | 64 | @ | 96 | ` |
| 1 | <SOH> | 33 | ! | 65 | A | 97 | a |
| 2 | <STX> | 34 | " | 66 | B | 98 | b |
| 3 | <ETX> | 35 | # | 67 | C | 99 | c |
| 4 | <EOT> | 36 | $ | 68 | D | 100 | d |
| 5 | <ENQ> | 37 | % | 69 | E | 101 | e |
| 6 | <ACK> | 38 | & | 70 | F | 102 | f |
| 7 | <BEL> | 39 | ' | 71 | G | 103 | g |
| 8 | <BS> | 40 | ( | 72 | H | 104 | h |
| 9 | <TAB> | 41 | ) | 73 | I | 105 | i |
| 10 | <LF> | 42 | * | 74 | J | 106 | j |
| 11 | <VT> | 43 | + | 75 | K | 107 | k |
| 12 | <FF> | 44 | , | 76 | L | 108 | l |
| 13 | <CR> | 45 | - | 77 | M | 109 | m |
| 14 | <SO> | 46 | . | 78 | N | 110 | n |
| 15 | <SI> | 47 | / | 79 | O | 111 | o |
| 16 | <DLE> | 48 | 0 | 80 | P | 112 | p |
| 17 | <DC1> | 49 | 1 | 81 | Q | 113 | q |
| 18 | <DC2> | 50 | 2 | 82 | R | 114 | r |
| 19 | <DC3> | 51 | 3 | 83 | S | 115 | s |
| 20 | <DC4> | 52 | 4 | 84 | T | 116 | t |
| 21 | <NAK> | 53 | 5 | 85 | U | 117 | u |
| 22 | <SYN | 54 | 6 | 86 | V | 118 | v |
| 23 | <ETB> | 55 | 7 | 87 | W | 119 | w |
| 24 | <CAN> | 56 | 8 | 88 | X | 120 | x |
| 25 | <EM> | 57 | 9 | 89 | Y | 121 | y |
| 26 | <SUB> | 58 | : | 90 | Z | 122 | z |
| 27 | <ESC> | 59 | ; | 91 | [ | 123 | { |
| 28 | <FS> | 60 | < | 92 | \ | 124 | \| |
| 29 | <GS> | 61 | = | 93 | ] | 125 | } |
| 30 | <RS> | 62 | > | 94 | ^ | 126 | ~ |
| 31 | <US> | 63 | ? | 95 | _ | 127 | <DEL> |

- Codes are standard since the 60s
- Lets us write letters, numbers, punctuation in one byte (ie. 8 bits, ie. a number in 0-255)

Examples:

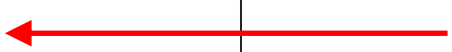
- 72 is ‘H’
- 101 is ‘e’
- The byte sequence 72, 101, 108, 108, 111 is “Hello”
- **Conversely** — ‘A’ is 65.

ASCII **codes 33-126 are proper text characters**. The **PHRED-33** encoding uses the 33-73 range to associate characters to scores 0 to 40 (confidences from 0 to 99.99%)

⇒ An ‘A’ quality (65) corresponds to: a PHRED score of 25, a confidence of 99.7%

# FASTQ base qualities: summing up

```
@J00153:333:HF2VNBBXY:1:1101:1347:1086 1:N:0:NGCTCATT+NAGGACGT
ACCGGATTGCATCCCCAATCTGCTGTTTGTTTTCCACTGCGTACGTTCC
+
9<AFAJAFJJJJJJJJJ7AA-<FFF-<F-<FJJAJJJJJJJ-FFJFFJ<
```



Qualities **9<AFJA…** correspond to:

- ASCII codes 57, 60, 65, 70, 74, 65
- Scores 24, 27, 32, 37, 41, 32 in PHRED-33
- Confidences of 99.60; 99.80; 99.94; 99.98; 99.99; 99.94 %

# Exercise 1

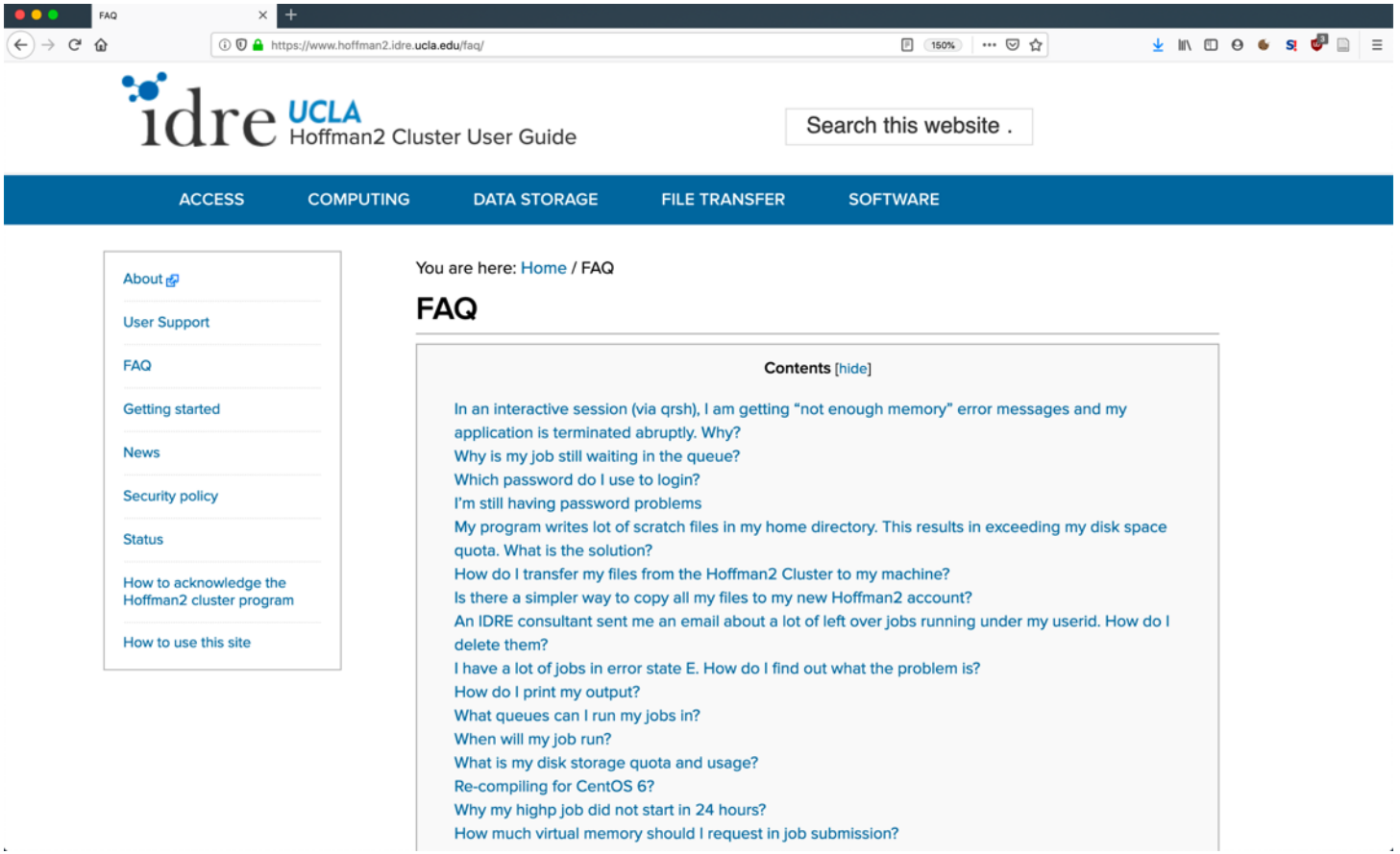
- The workshop data is in: **`/u/home/n/nrochett/QCB_RNAseq`**
- Copy this directory and its content to ***your*** home directory (note: this is only 5% of the raw data, for only two samples)
- **How many samples** are listed in `samples.txt` ?
- According to the `Kim2016.infoTable.tsv` table:
  - Has this RNAseq library been sequenced **single-end or paired-end**?
  - **How many samples** are there in the KO group ($NRL^{-/-}$)? In the control group?
- For file `P10KO_rep1.five_percent.fastq.gz` :
  - **How many lines** are there?
  - **How many reads** does this represent?
  - **What is the confidence** for the first base call of the first read?
  - **How long** are the reads?

# Working on the Hoffman2 cluster

# Hoffman2 Homepage & Resources

**https://www.hoffman2.idre.ucla.edu/**

Also, https://idre.ucla.edu/



# Connecting to the cluster

- Remote shell
  ```
  ssh USER@hoffman2.idre.ucla.edu
  ```
  - (through WSL/Putty/Mobaxterm on Windows)
  - Once connected, start an 'interactive job' requesting 4 hours and 4GB memory
    ```
    qrsh -l h_rt=4:00:00,h_data=4G
    ```

- Remote copy/File transfer

  Download:
  ```
  scp USER@hoffman2.idre.ucla.edu:/PATH/TO/FILE ~/Desktop/
  ```
  Upload:
  ```
  scp ~/Desktop/FILE USER@hoffman2.idre.ucla.edu:~/
  ```
  - Interactive clients exist:
    Cyberduck, FileZilla…

# Essential Hoffman2 commands

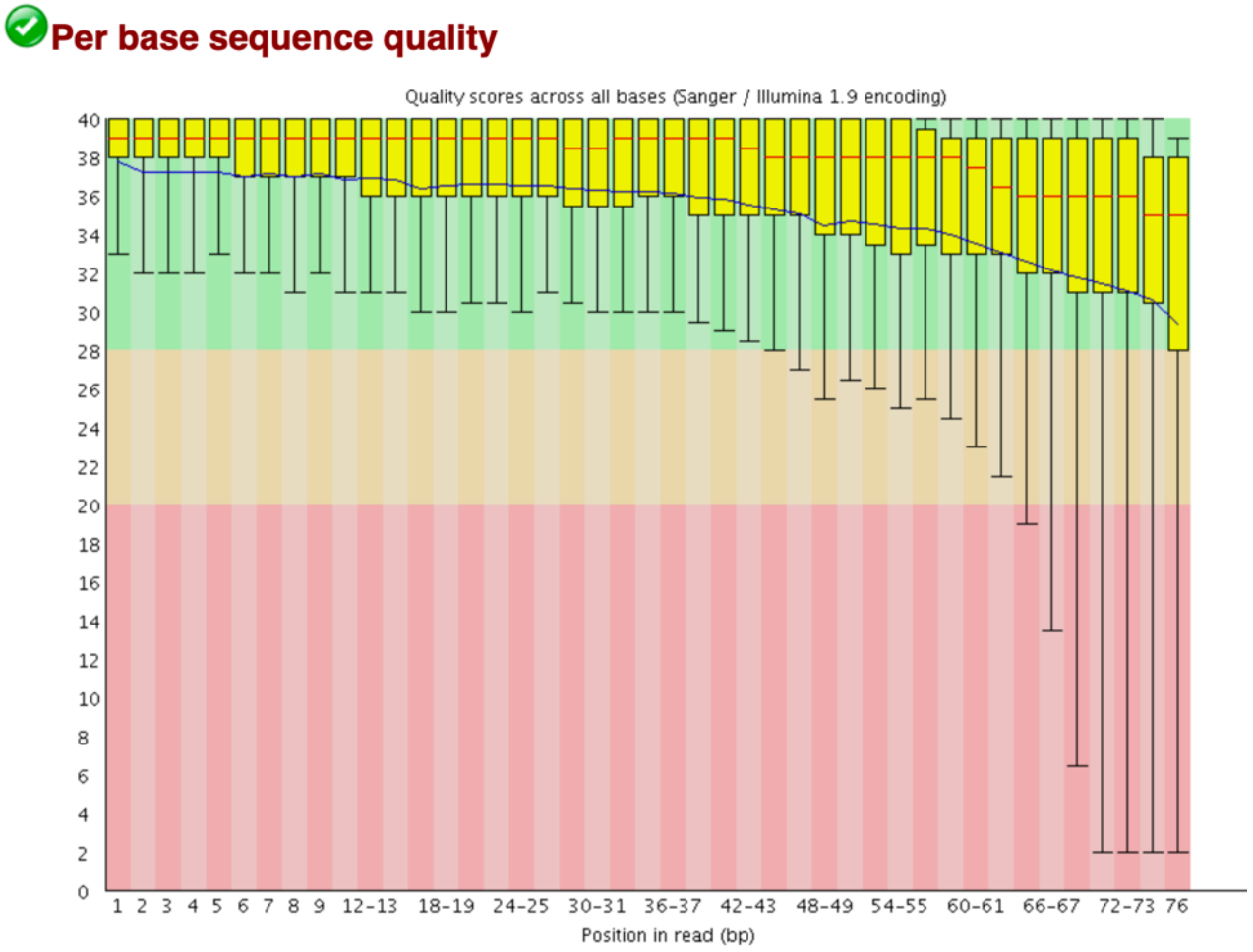
- Request an interactive job:
  **`qrsh`**
  *(Do not run computationally heavy commands on the login node!)*
- Submit a batch/script job:
  **`qsub –N JOBNAME script.sh`**
- Main resource options for jobs:
  Time: **`-l h_rt=4:00:00`**
  Memory / RAM: **`-l h_data=4G`**
- List current jobs: **`myjobs`** (or `qstat –u $USER`
  Main job states are **`r`** (for running) and **`qw`** (for queued-waiting)
- Modules:
  **`module load`** `NAME`
  `module avail`

# Hoffman2 disk spaces

- **Home directory** (20Gb max.):
  **`/u/home/?/USER`**
- **Scratch** (2Tb max., 'temporary'):
  `/u/flashscratch/?/USER`
  *N.B. Scratch data is considered abandoned after two weeks & it is not backuped*
- **Purchased space** (cf. IDRE website):
  `/u/project/LABNAME/USERNAME`

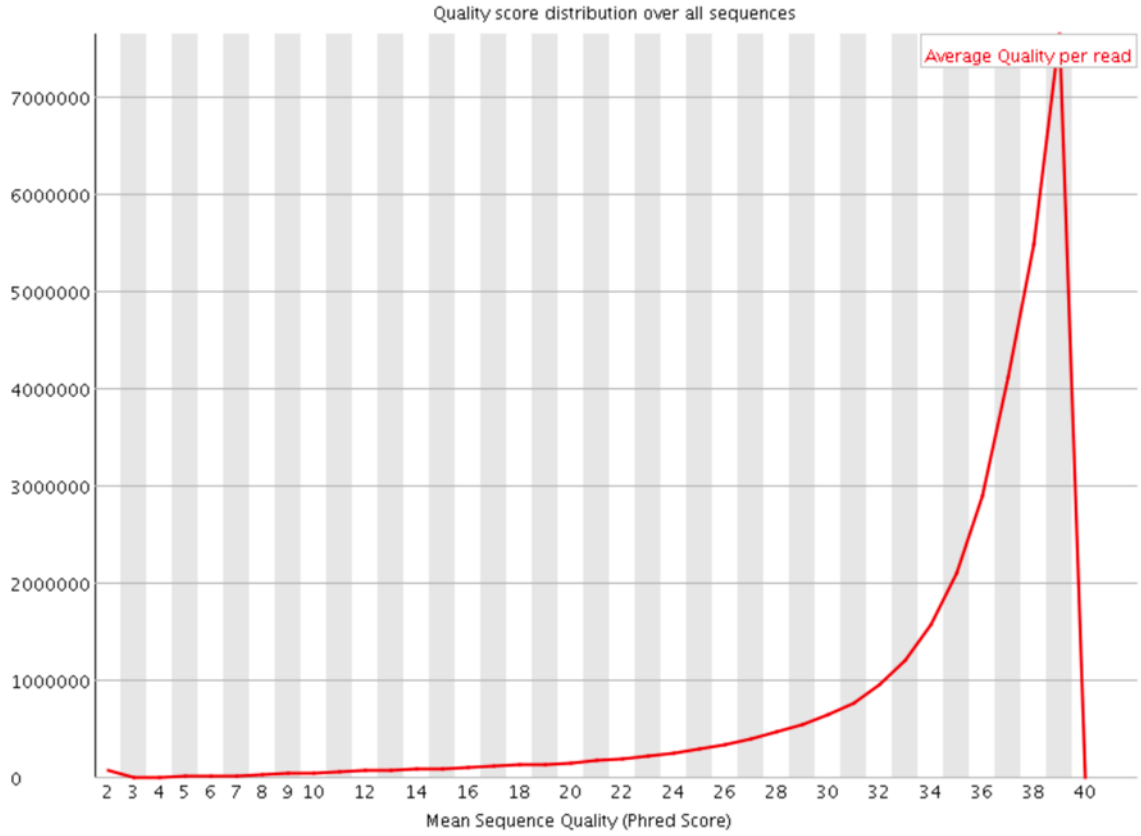
# Checking Read qualities with FastQC

# FastQC: base qualities

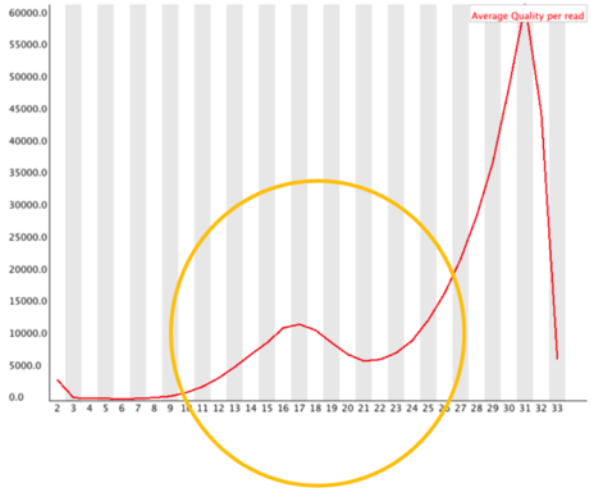


# FastQC: per-read mean base qualities

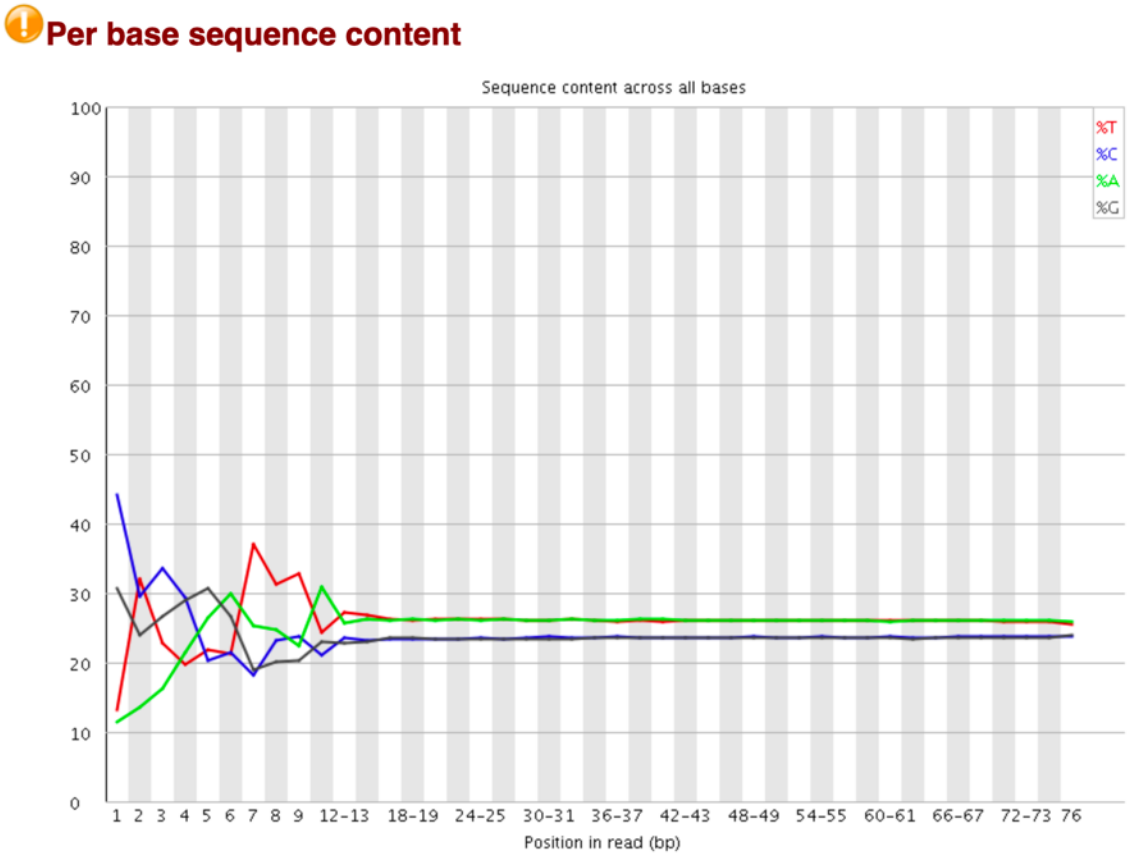
**Per sequence quality scores**



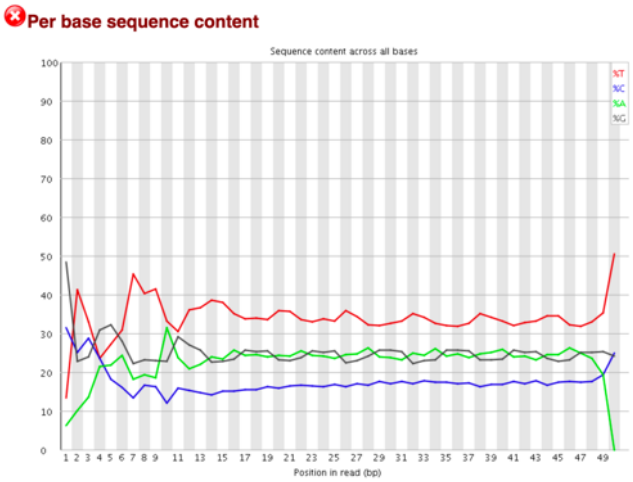

(bad)



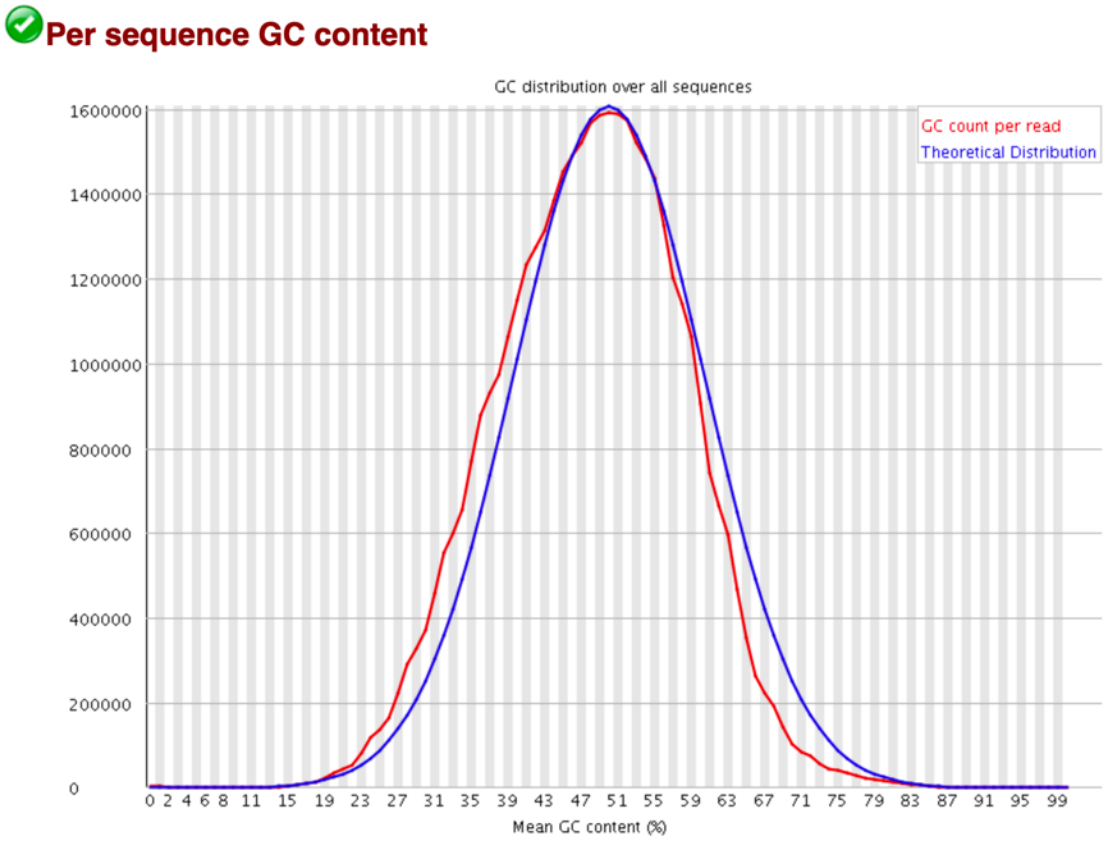
# FastQC: %ACGT over read length



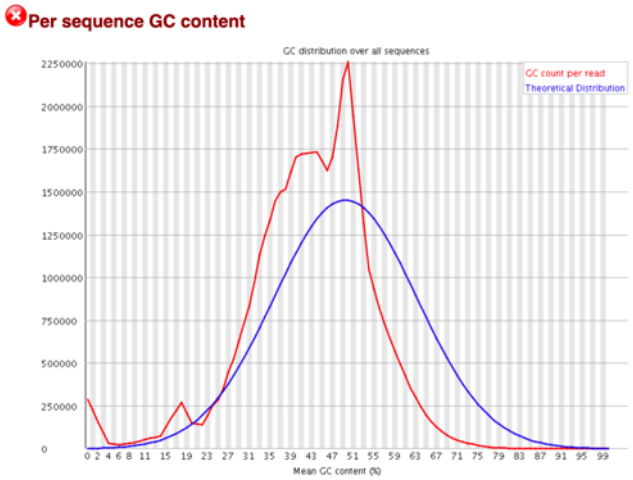

(bad)



# FastQC: per-read %GC

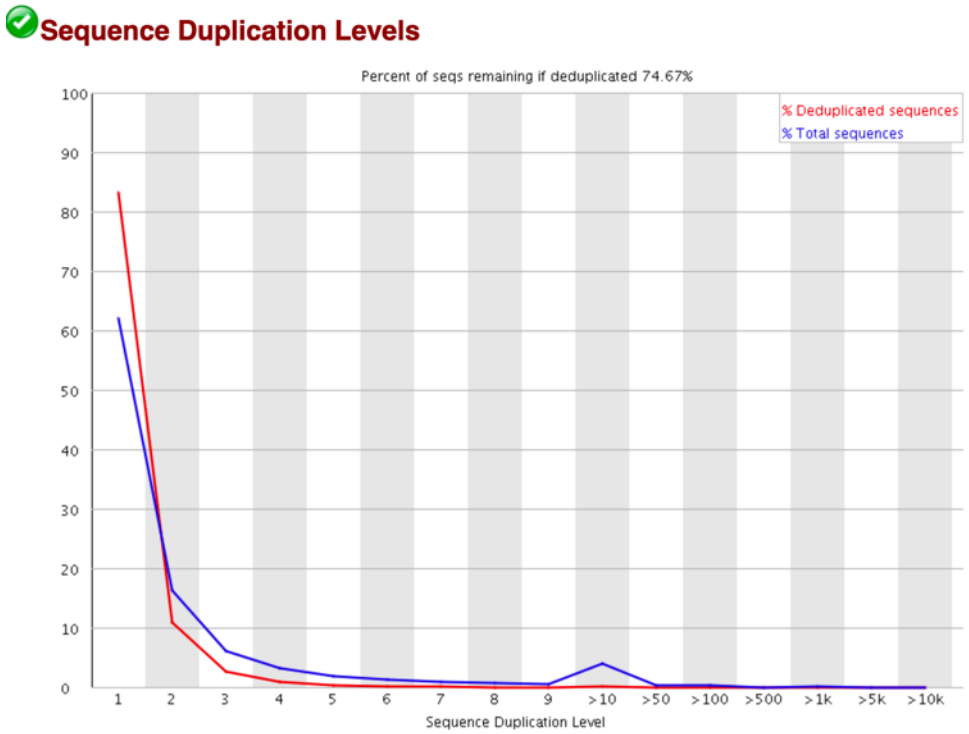



(bad)



# FastQC: over-represented sequences

- A little over-representation is normal in RNAseq (highly expressed genes)



# Exercise 2: let's run FastQC

- (Installation demonstration)
- Run FastQC on `P10KO_rep1.five_percent.fastq.gz`
  - You will need to specify the full path to the program!
  - For most programs the `-h` option asks for the 'help'
- Download the HTML results file (*e.g.* with SCP)
  & review it in your web browser

# FastQC: code

- Usage is given by:

```
./programs/fastqc/fastqc --help
```

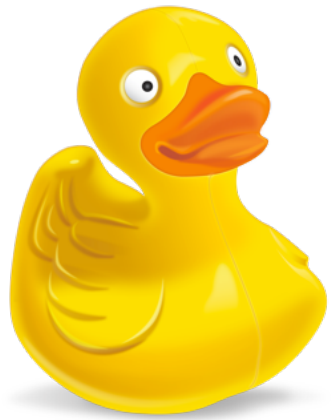
- We can run it with just:

```
./programs/fastqc/fastqc P10KO_rep1.five_percent.fastq.gz
```

- To download the HTML file (from *your* computer):

```
scp  USER@hoffman2.idre.ucla.edu:~/QCB_RNAseq/P10KO_rep1.five_percent.fastq.gz  ~/
```

*(Or with Cyberduck/etc.)*



# 6. Read trimming & filtering with ‘Trimmomatic’

# Exercise 3: let's run Trimmomatic

- Website & manual
- Run it for sample **P10KO_rep1** and/or sample **P10WT_rep1**

```
jar_file=~/QCB_RNAseq/programs/trimmomatic/trimmomatic-0.39.jar
java -jar $jar_file <ARGUMENTS...>
```

  - Ask for:
    - for minimum leading & trailing base qualities of 20 (`LEADING` and `TRAILING` parameters)
    - for clipping when the base quality over 5 bases drops under 20 (`SLIDINGWINDOW` parameter)
    - a minimum final length of 60 (`MINLEN` parameter)
    - *(for simplicity let's not use the ILLUMINACLIP parameter; otherwise we would have used the file TrueSeq2-SE to match the library preparation)*
- How many reads passed filters? What proportion of the raw reads have been discarded?

# Trimmomatic: code

- According to the online "quick start", the usage for single-end data is:

```
java -jar trimmomatic-0.35.jar SE -phred33 input.fq.gz
output.fq.gz ILLUMINACLIP:TruSeq3-SE:2:30:10 LEADING:3
TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

- We adapt it to (we also need to specify to specify a single thread/processor using `-threads 1`, as the default is to use several):

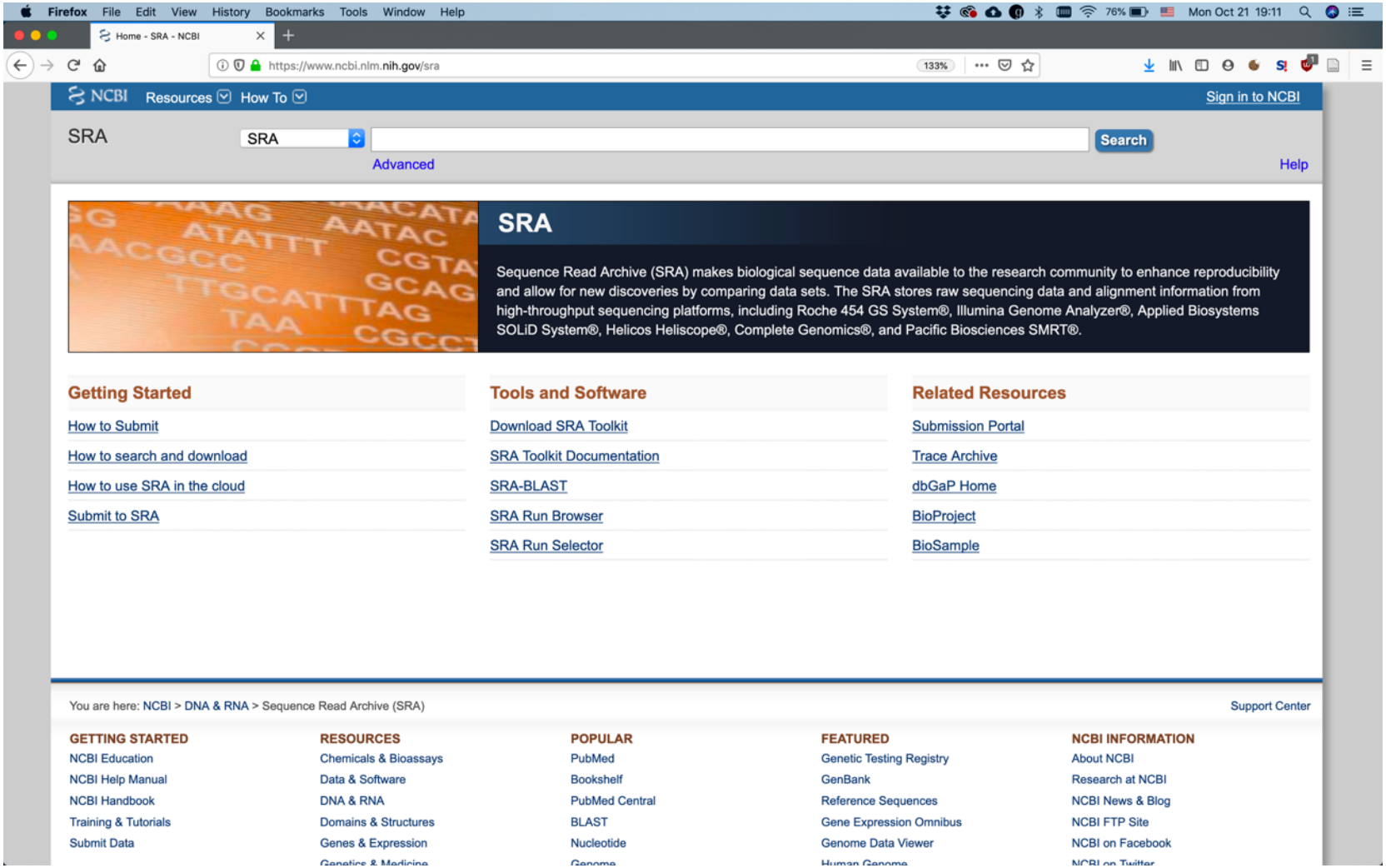
```
jar_file=~/QCB_RNAseq/programs/trimmomatic/trimmomatic-0.39.jar

java -jar $jar_file \
    SE -phred33 -thread 1 \
    P10KO_rep1.five_percent.fastq.gz \
    P10KO_rep1.five_percent.trimmo.fastq.gz \
    LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20 MINLEN:60
```

(***Note:*** *a backslash* `\` *at the end of a line means "continued line".)*

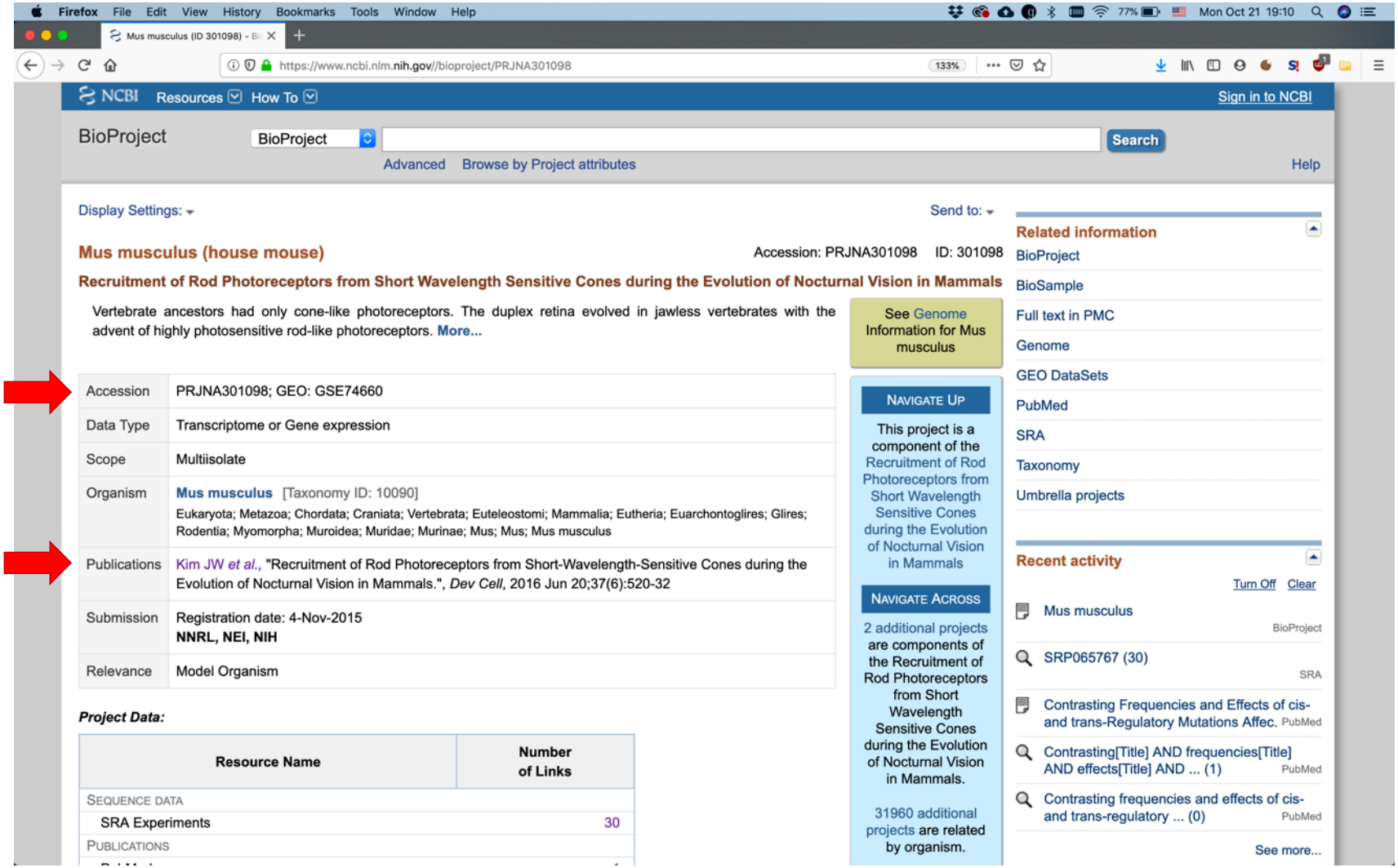
# the NCBI SRA (Sequence Read Archive)

https://www.ncbi.nlm.nih.gov/sra

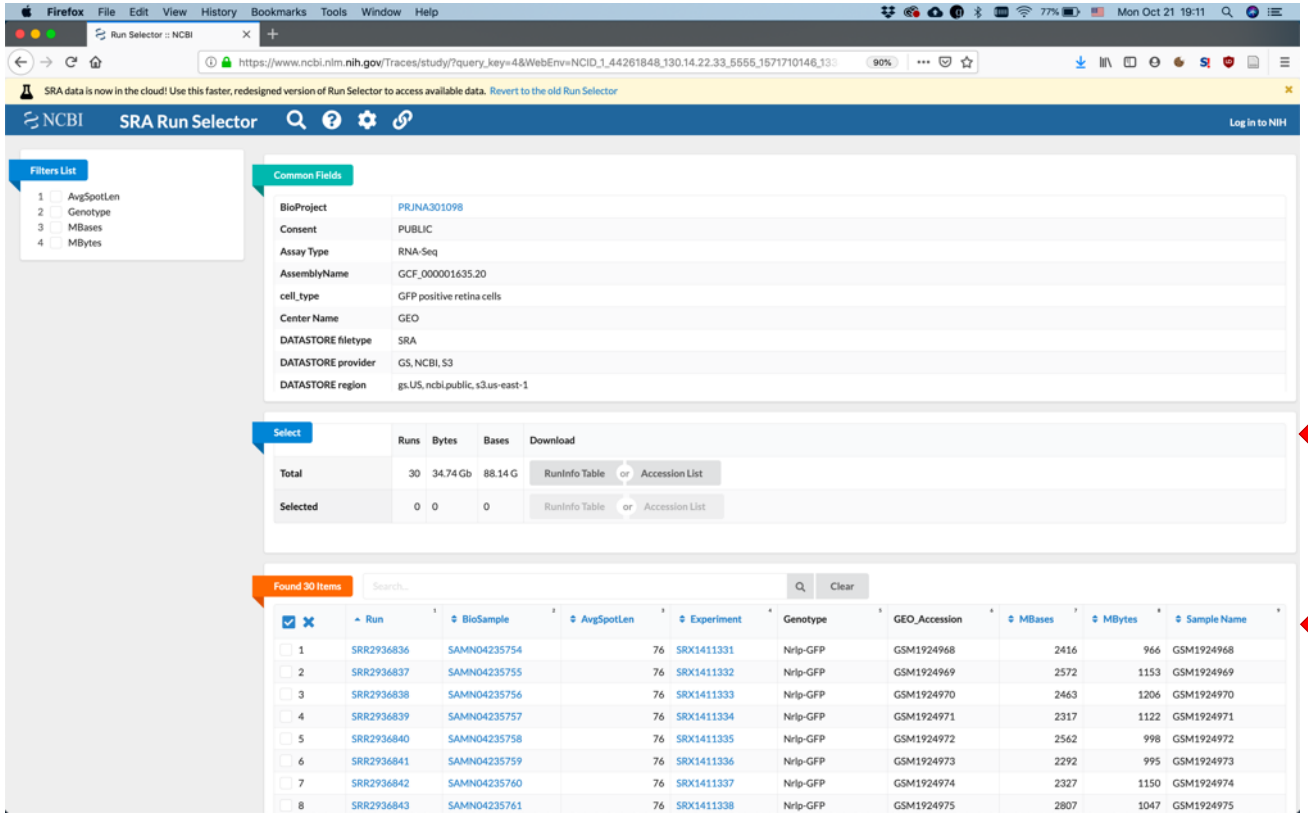


# NCBI SRA & BioProjects

eg. project **PRJNA301098**



# Downloading from the SRA Via “SRA run selector”



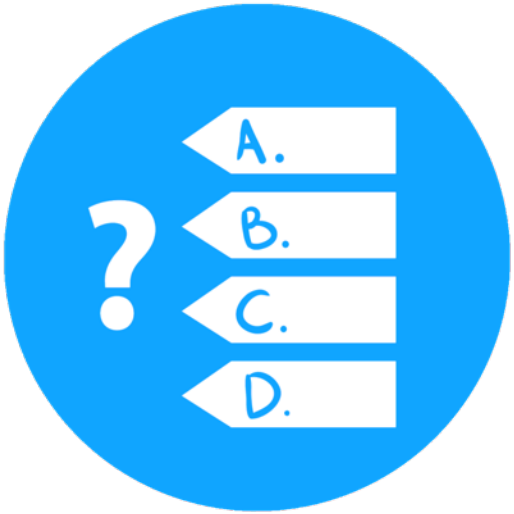

Information on samples & libraries

Sequence files
Each “Run” (SRR identifiers) is one sample

Reads for every sample were retrieved using the SRA toolkit, eg.:

```
fastq-dump --gzip $SRR_ID
```

# Quizzes




https://tinyurl.com/...

# *RNA-seq analysis: Overview*