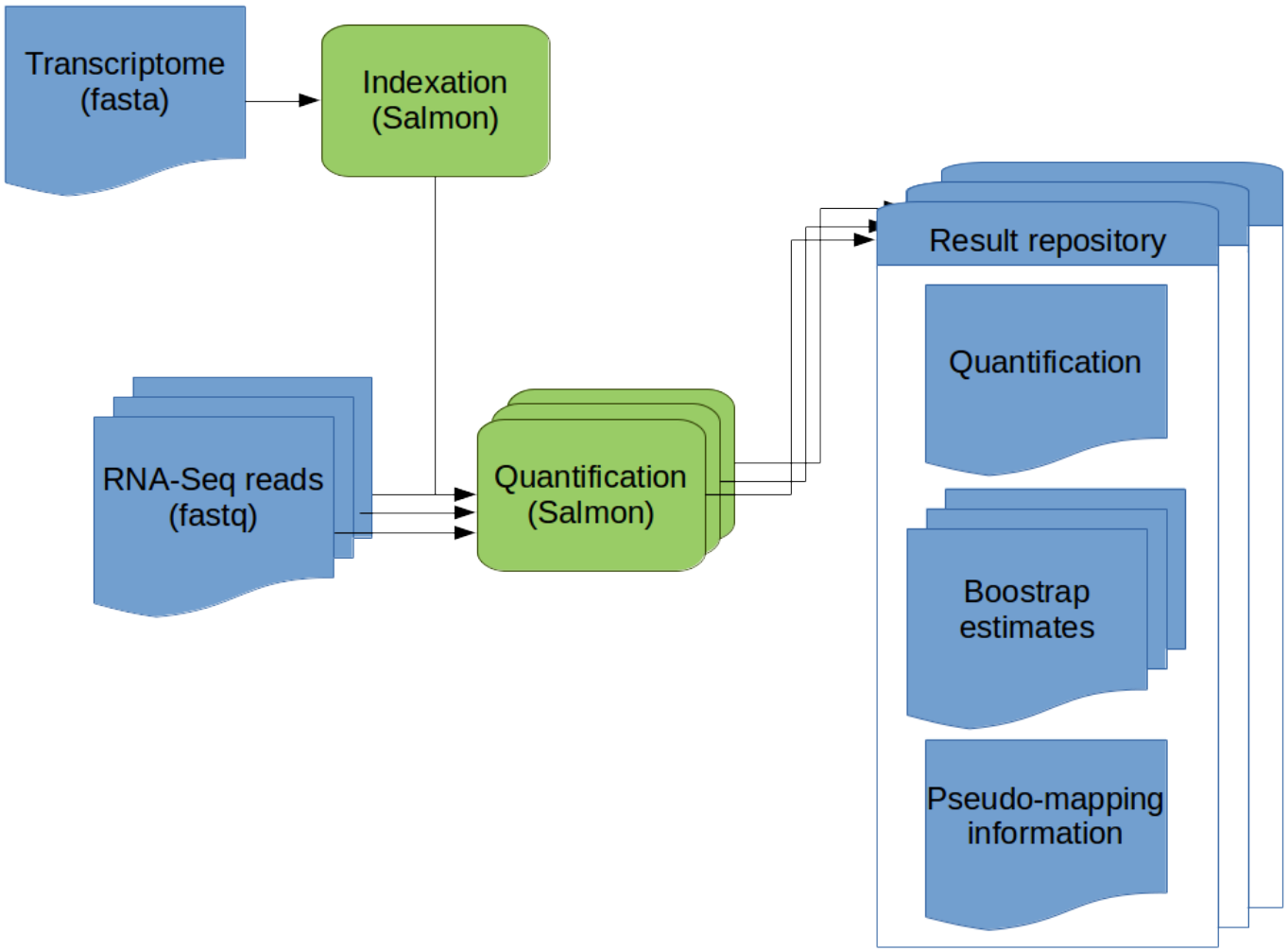
Differential Transcript Expression Analysis

Thibault Dayris, Claire Toffano-Nioche

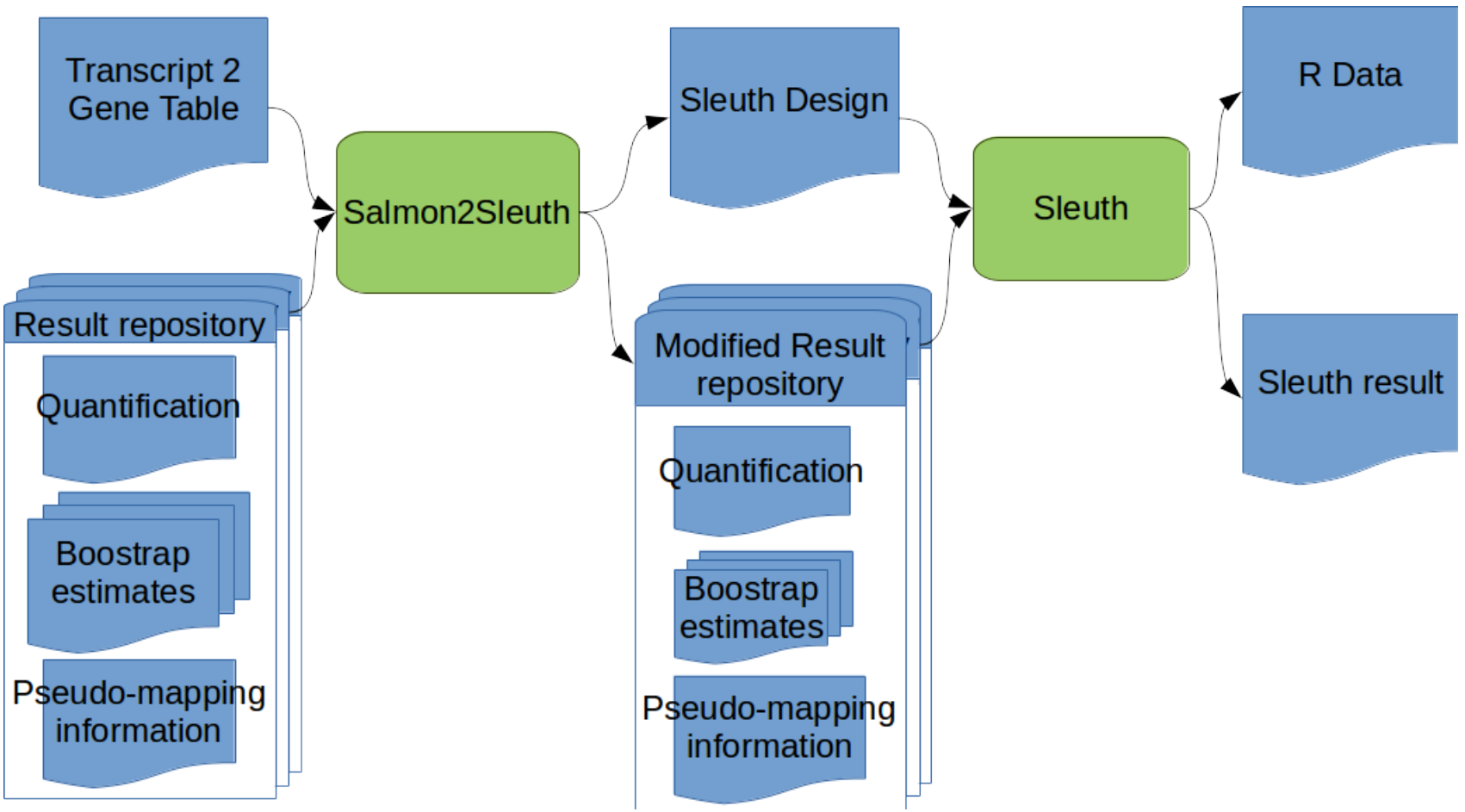
2017-11-14

Pipeline

Remember the Salmon output

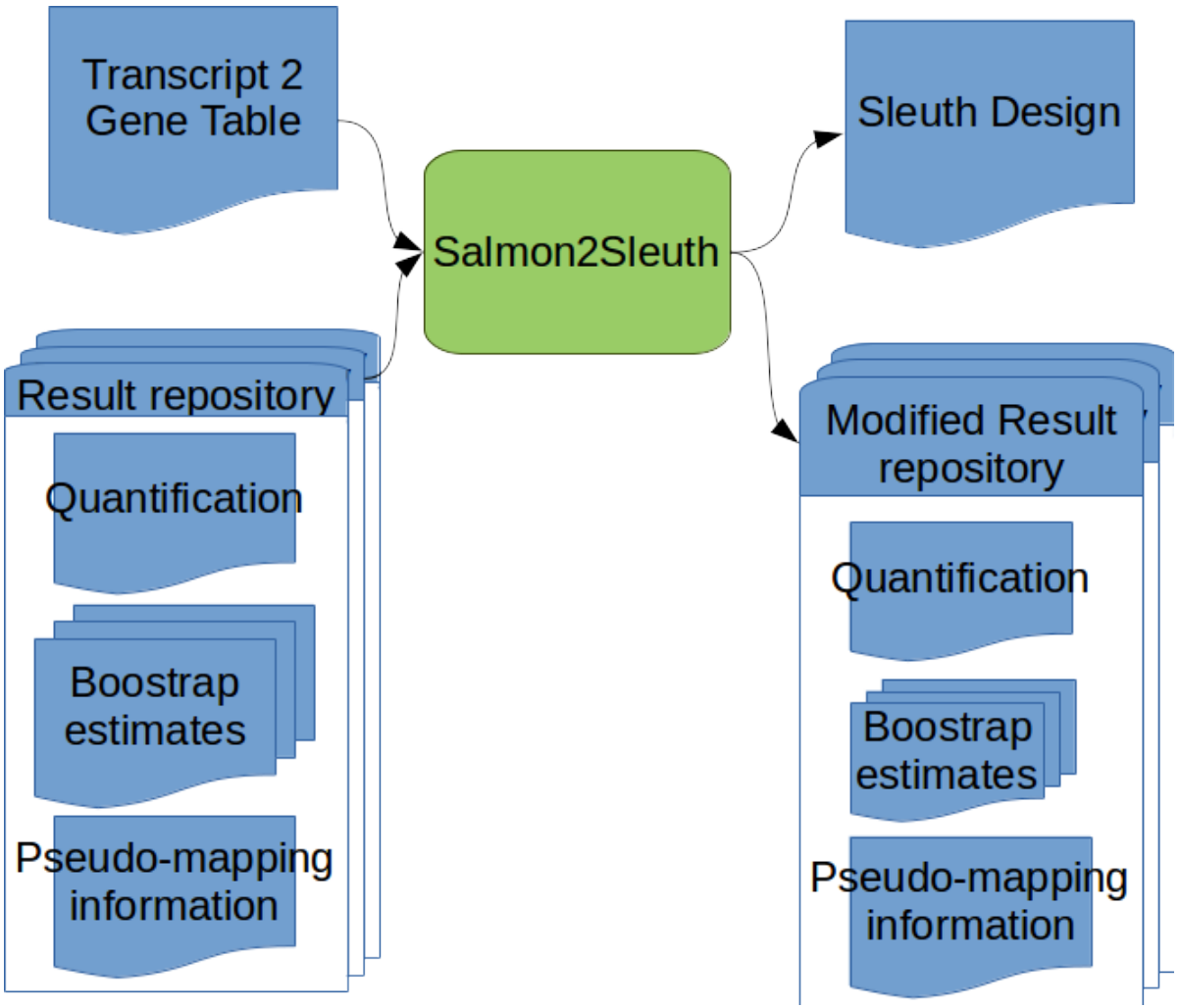


Whole Differential Transcript Expression Analysis



From Salmon to Sleuth

Salmon2Sleuth section



Remember the Salmon output

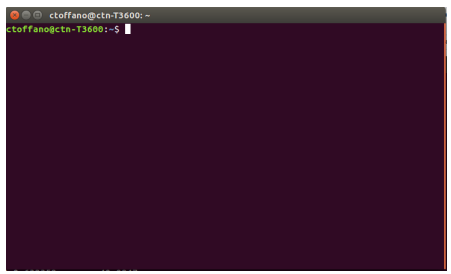
Salmon output is composed of multiples files and repositories.

However Sleuth has specific requirements:

- *quant.sf* must be in HDF5 format (Hierarchical Data Format)
- bootstraps must be gzipped and available in *aux_dir/bootstrap/*

Cluster acces

If you are not yet `log` onto the ABIMS cluster (Roscoff) and into a `qlogin` mode: open an "Xterm" window (traditionally a black window)




```
ssh my_user_name@bioinfo.sb-roscoff.fr # link cluster-PC
qlogin -pe thread 4 # login with parallel specification
cdprojet # http://abims.sb-roscoff.fr/resources/cluster
source "${CONDA3}"/activate eba2017_rnaseq_ref # use tools
RNASEQDIR="/projet/sbr/ggb/RNAseq/" # define shortcut
resize # avoid command line wrapping
```

Salmon2Sleuth.R

Salmon2Sleuth is a R script written in the purpose of this course to make the conversion of Salmon repositories easier.

The R package Wasabi[1] performs quant.sf format conversion within the R environment.

```
# Copy salmon directories in your working directory
cp -r "${RNASEQDIR}"/salmon_results/* .
# Copy Salmon2Sleuth.R script in your working directory
cp "${RNASEQDIR}"/Salmon2Sleuth.R .
# Run Salmon2Sleuth.R with R
Rscript Salmon2Sleuth.R
```

[1] https://github.com/COMBINE-lab/wasabi

Salmon2Sleuth.R run

Salmon2Sleuth.R is composed of two steps:

- Automatic format conversion with Wasabi
- Interactive creation of our design file

Salmon2Sleuth.R run

```
Library loaded
Wed Nov  8 23:15:00 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/KO1 to kallisto HDF5 format"
Wed Nov  8 23:15:02 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/KO2 to kallisto HDF5 format"
Wed Nov  8 23:15:04 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/KO3 to kallisto HDF5 format"
Wed Nov  8 23:15:06 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/WT1 to kallisto HDF5 format"
Wed Nov  8 23:15:08 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/WT2 to kallisto HDF5 format"
Wed Nov  8 23:15:10 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/WT3 to kallisto HDF5 format"
Salmon directories converted
Enter comma separated list of conditions according to your samples:
```

Salmon2Sleuth.R run

```
Library loaded
Wed Nov  8 23:15:00 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/KO1 to kallisto HDF5 format"
Wed Nov  8 23:15:02 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/KO2 to kallisto HDF5 format"
Wed Nov  8 23:15:04 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/KO3 to kallisto HDF5 format"
Wed Nov  8 23:15:06 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/WT1 to kallisto HDF5 format"
Wed Nov  8 23:15:08 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/WT2 to kallisto HDF5 format"
Wed Nov  8 23:15:10 2017
[1] "Successfully converted sailfish / salmon results in /projet/sbr/ggb/RNAseq/TP_isoforms/WT3 to kallisto HDF5 format"
Salmon directories converted
Enter comma separated list of conditions according to your samples:
KO,KO,KO,WT,WT,WT
```

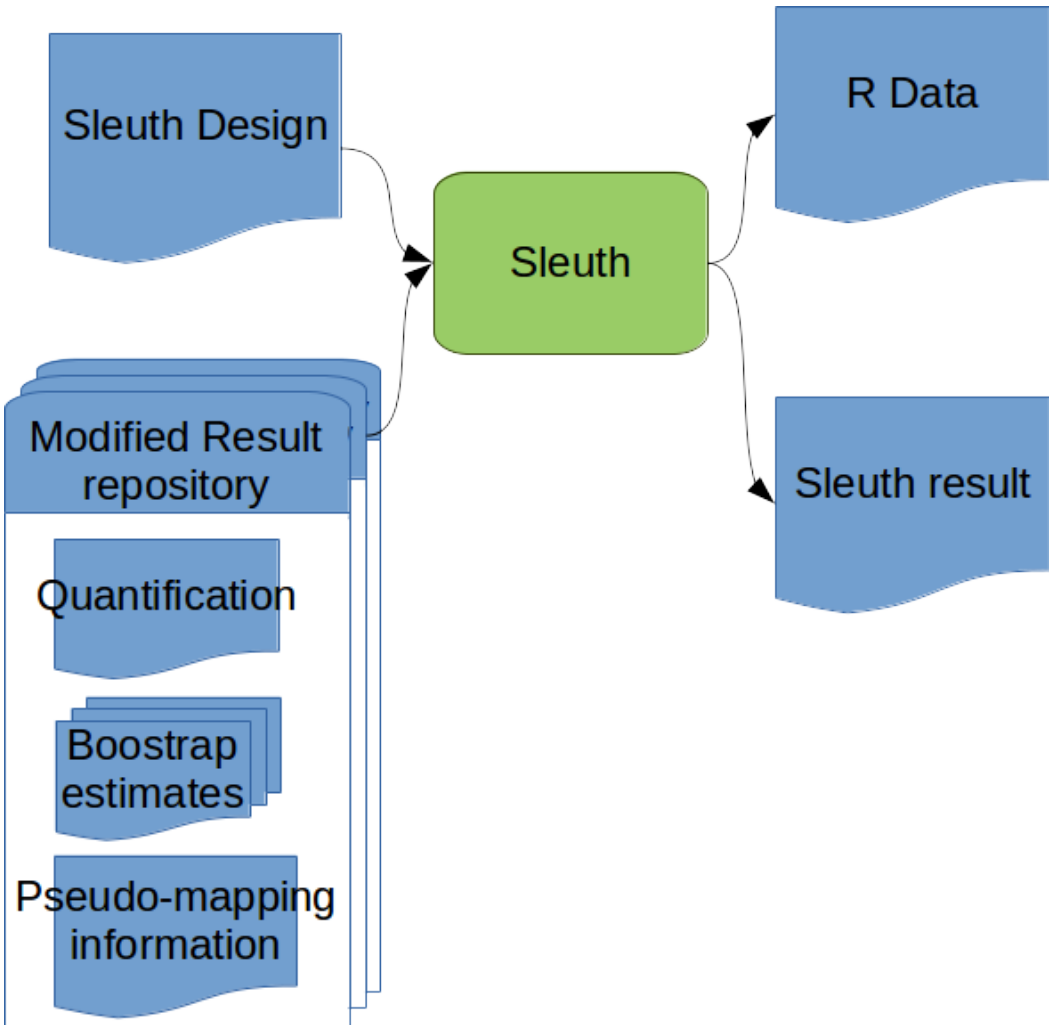
`Enter a comma separated list of conditions according to your samples:`
`KO,KO,KO,WT,WT,WT`

Salmon2Sleuth.R results

```
sample  path    test
KO1 /projet/sbr/ggb/RNAseq/TP_isoforms/KO1  KO
KO2 /projet/sbr/ggb/RNAseq/TP_isoforms/KO2  KO
KO3 /projet/sbr/ggb/RNAseq/TP_isoforms/KO3  KO
WT1 /projet/sbr/ggb/RNAseq/TP_isoforms/WT1  WT
WT2 /projet/sbr/ggb/RNAseq/TP_isoforms/WT2  WT
WT3 /projet/sbr/ggb/RNAseq/TP_isoforms/WT3  WT
```

Differential Transcript Expression

Sleuth section



Run Sleuth.R

```
# Copy Sleuth.R script in your working directory
cp "${RNASEQDIR}"/Sleuth.R .
# Run SLeuth.R script with R
Rscript Sleuth.R
```

Sleuth method

In RNA-Seq, variations between conditions can be identified in 2 groups:

- Biological and experimental variance, which is what we're looking for
- Inferential variance, which is artefactual
  - sequencing methods (reads number, sequences number, ...)
  - bioinformatical methods (pseudo-mapping, ...)

The Graal, is the accurate identification of the observed variance as biologically relevant or artefactual.

Bootstraping

Remember Salmon uses Bootstraps to estimate the inferential variance for each sample.

Based on this estimation, Sleuth produces both P-Value and FDR to identify the variance between samples as biological or artefactual.

Sleuth Output detailed

Sleuth Live

Sleuth provides a nice interface to navigate through results within R.

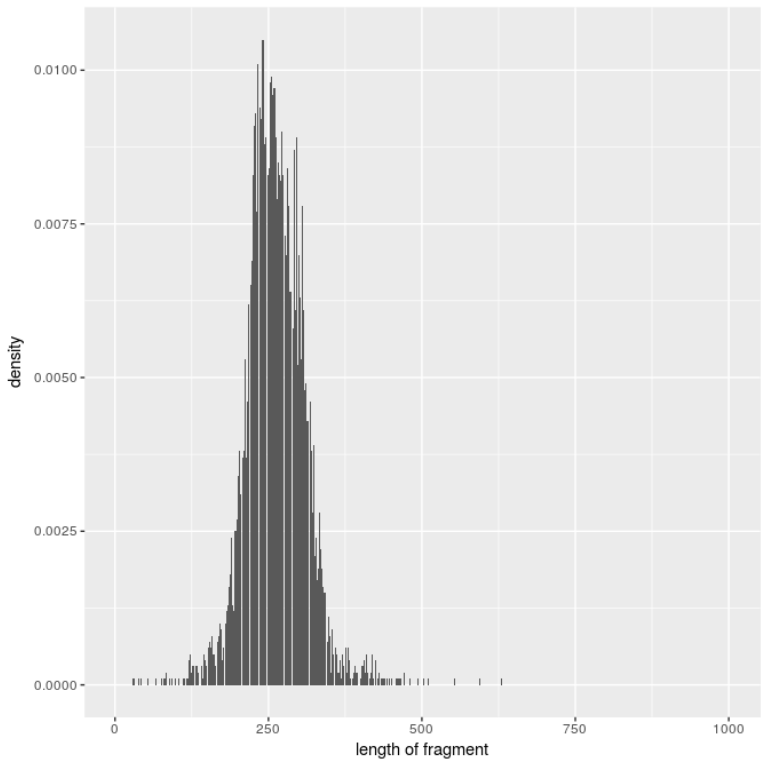
```
library("sleuth")
so <- readRDS("Sleuth_results.rds");
sleuth_live(so);
```

List all Sleuth functions

Most of R packages provide a “vignette”, use it!

```
help(package="sleuth");
```

Fragment distribution

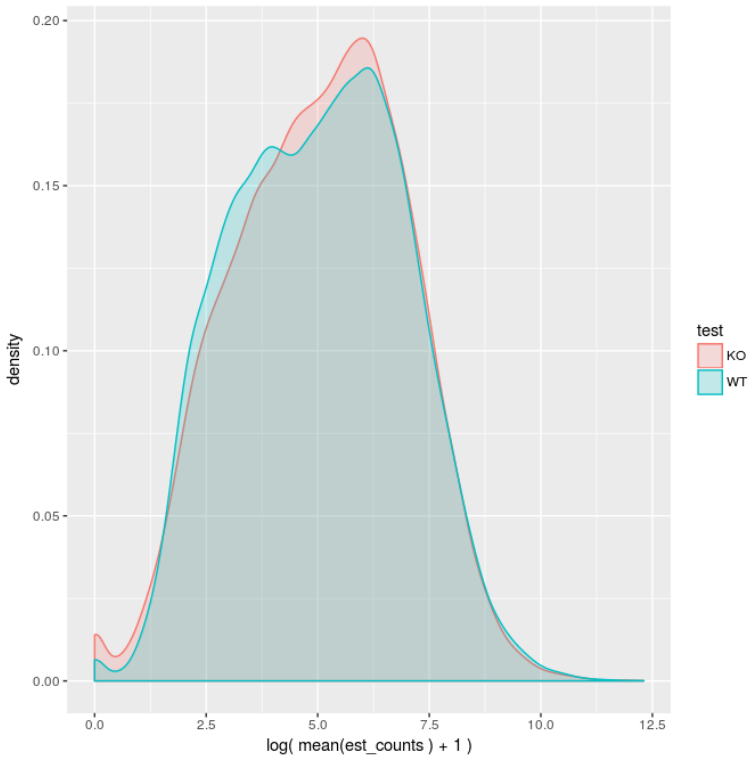


The fragment distribution graph illustrates the fragment length observation in the given sample.

- X axis: length of the fragments
- Y axis: frequency of observation

We expect a Gaussian distribution, with a maximum equal to the theoretical fragment length distribution given by the sequencing platform.

Group Density

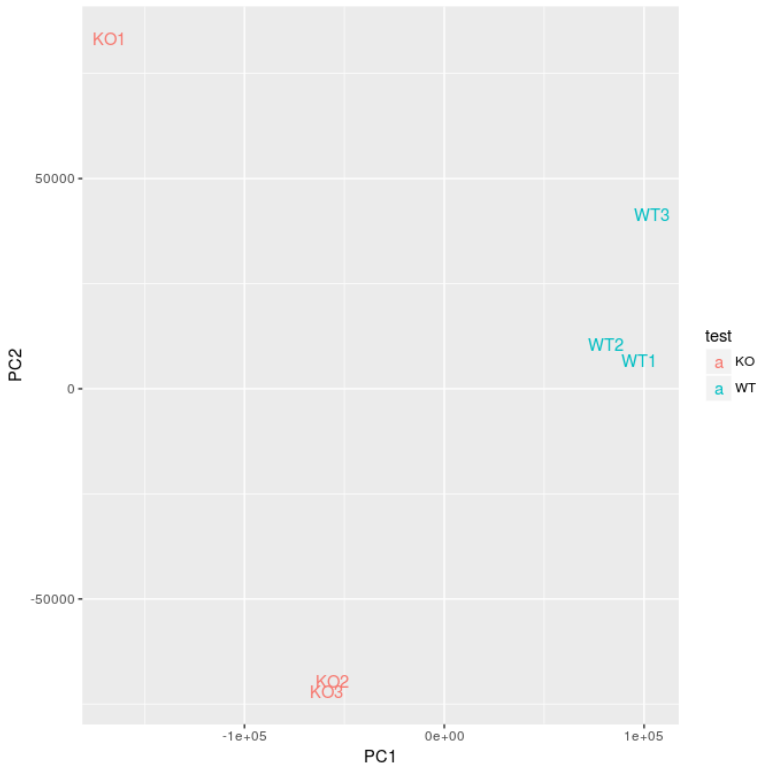


The group density illustrates the count density across conditions.

- X axis: mean counts, +1 to avoid negative values
- Y axis: number of observation
- Each color represent a condition

We expect each condition to behave quite similarly. The extreme values should be almost equal to 0. The highest value should match between conditions.

PCA

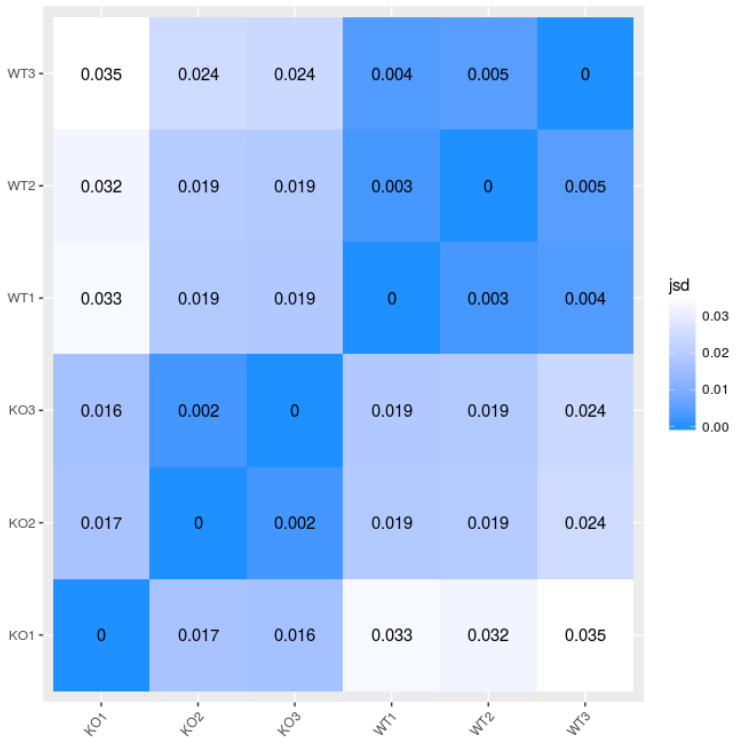


The PCA illustrates the divergence between samples.

- Each point is a sample
- Distance between samples is proportional to their divergence to each other

We expect samples across conditions to be well grouped, and widely separated from other conditions. Ultimately, a straight line should be drawable between conditions.

Heatmap

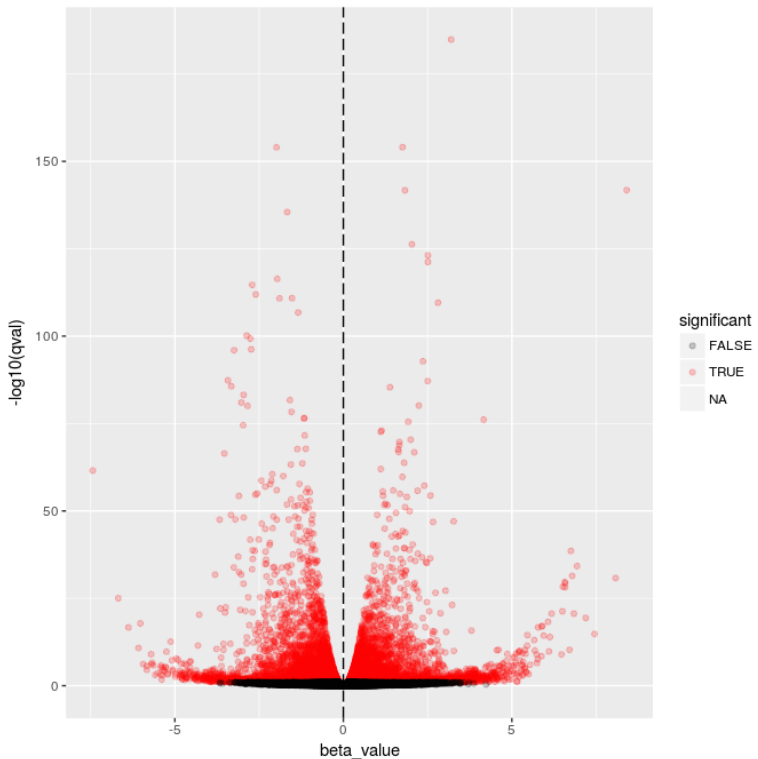


The Heatmap illustrates the divergence between samples.

- Each cell corresponds to a sample against another
- The colormap runs from blue (identical) to white (totally different)
- The numbers are jsd, which starts at 0 (identical) and raises with difference.

We expect each condition to be self-contained in a blue square, and separated from each other by white squares.

Volcano plot

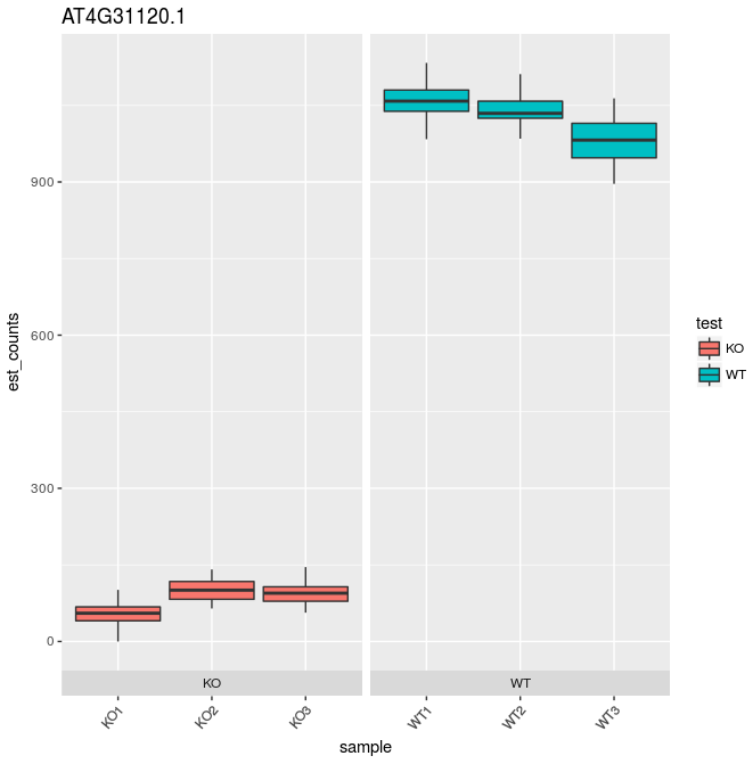


The Volcano plot illustrates the fold change and the confidence of differentially expressed events.

- X axis: beta_value, a log2(Fold_Change)
- Y axis: confidence in the DTE
- Red dots are transcripts differentially expressed
- Black dots are transcripts equally expressed

High confidence events are higher in the graph, most differentially expressed events are on the side of the graph.

Bootstrap summaries

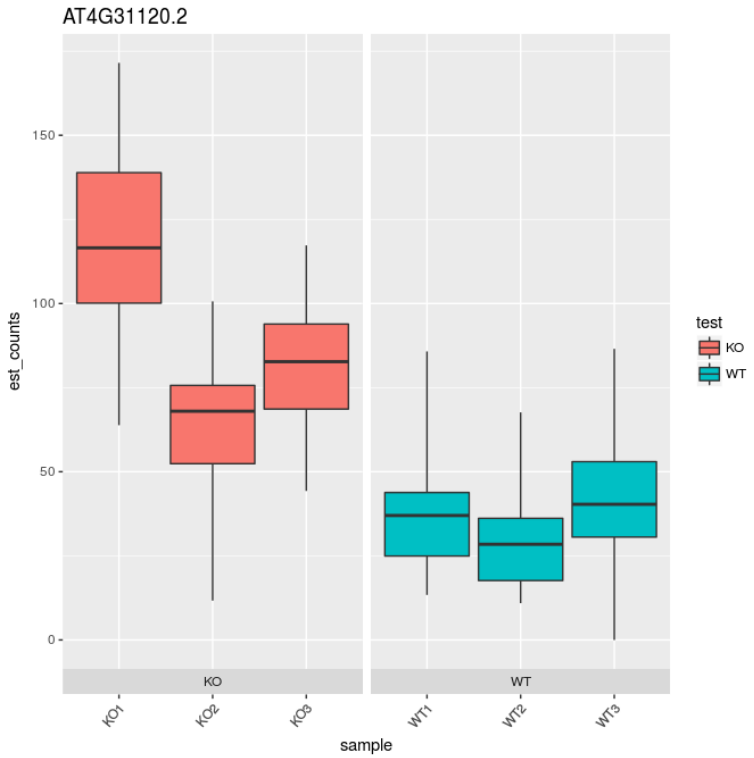


This graph illustrates the counts estimates across conditions

- X axis: Samples
- Y axis: Estimated counts

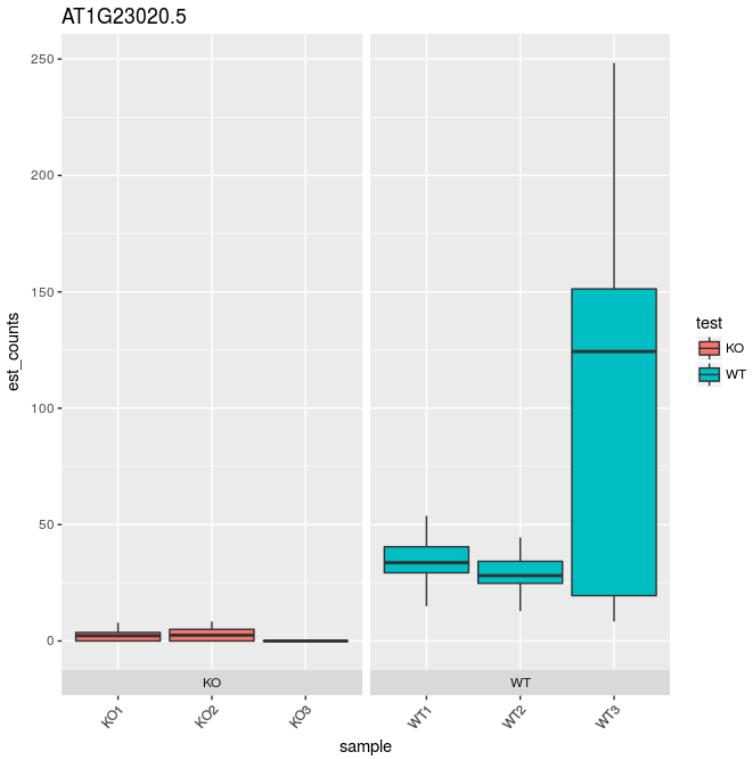
We expect the condition counts to be well grouped with each others, and well separated from other conditions

Bootstrap summaries



Here, the wide gaps in transcripts abundance within a condition leads to a low qvalue.

Bootstrap summaries



Low counts leads to troubleshot in EM algorithm and quantification.

Conclusion

Differential expression at isoform level

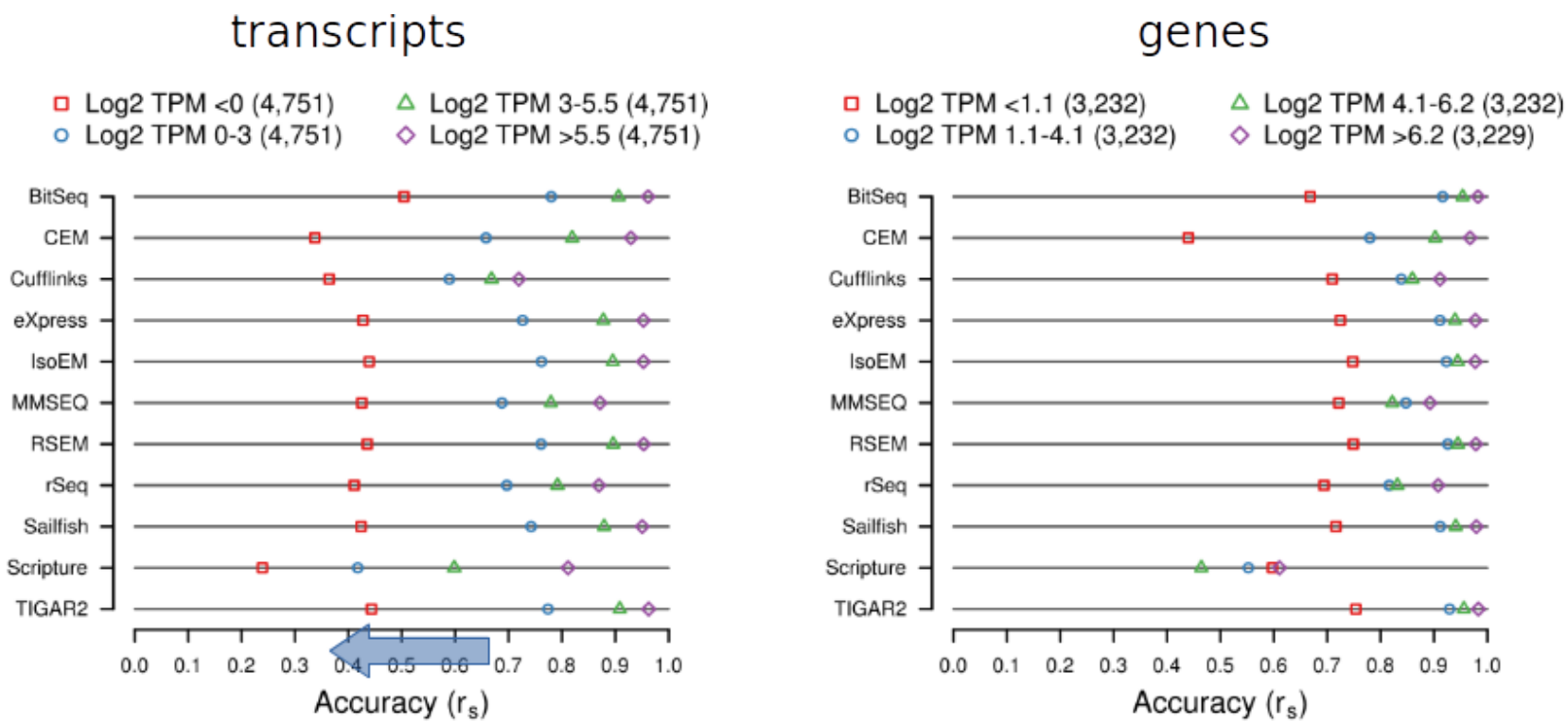
Not yet a well-established protocol because evaluate its accuracy is difficult:

- too few number of isoforms with experimental validation strategies (ex. qRT-PCR)
- synthetically generated datasets (simulation of reads) may not capture adequately the complexities of RNA-Seq experiments

Some keys:

- methods based on transcriptome are generally better than methods based on genome (eukaryotes)
- methods based on EM-algo are better than count-based methods
- the more abundant is the isoform, the more accurately it is inferred

Quantification gene/isoform level: deepness issue



Kanitz A, Gypas F, *et al. "Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data." Genome Biol. (2015)*

Isoforms discovery?

Quantification of isoform and discovery new isoform could be seen as 2 different tasks. Using the same RNAseq for these 2 tasks is common (begin by discover) but:

	quantification	discovery
sequencing lib. type	paired or single	paired/long reads
stranded sequencing lib.	better	compulsory
reads length	+	+++
mapping	transcriptome	genome/*de novo*
annotations	relies on	create new

Improve?

- make protocols like ribodepletion but for highly expressed housekeeping genes, to enrich with interesting transcripts
- isoform quantification:
  - don't forget the micro-arrays designed for (model organism)
  - gain statistical power with *spike-in* measurements
  - restrict the number of isoforms by gene (principal, following specific tissues, ...) to improve the number of reads/k-mers
- isoform discovery:
  - full-length cDNAs technology, long reads
  - complete isoform definitions by other NGS studies (and data integration with mixOmics):
    - ChIPSeq with a protein from the spliceosome as target
    - capturing the 5' or 3' end of RNAs

RNA-Seq: just a photo

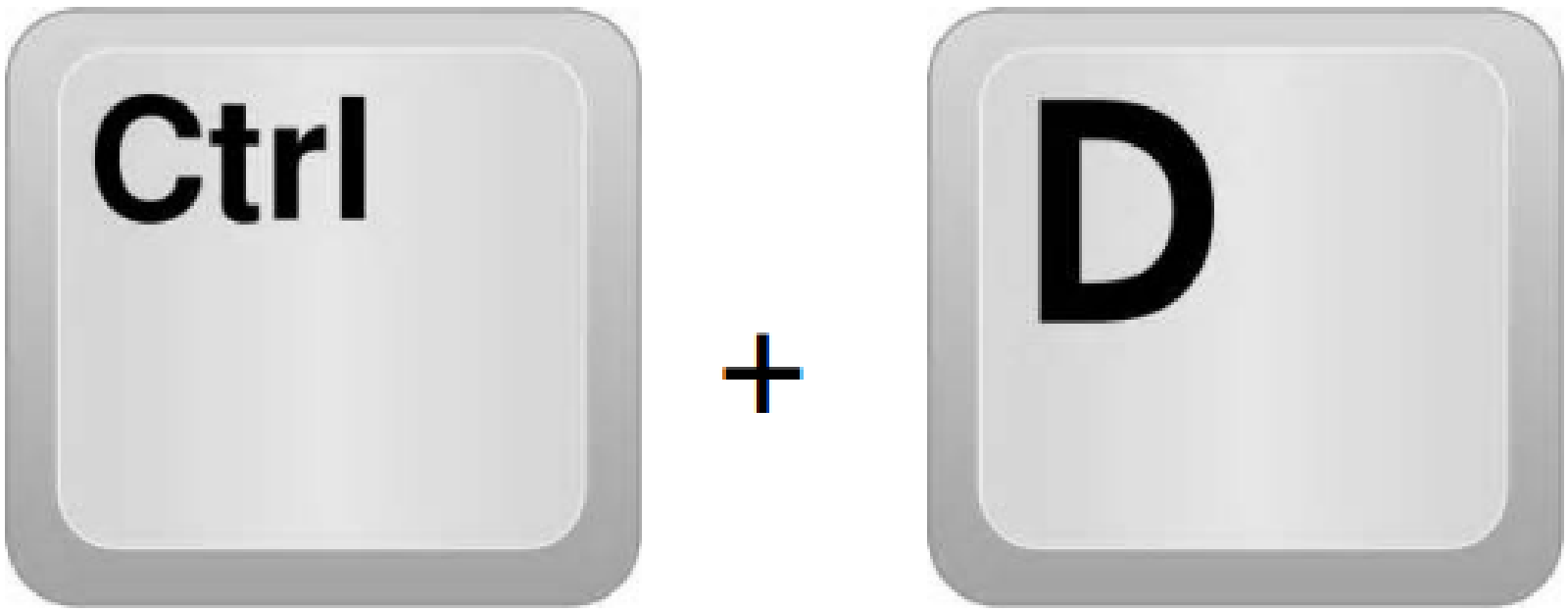
Adapt! biological query + organism + data + money

- sequencing protocols (single or paired-end, stranded or not), software, parameters

RNA-Seq is just an unique and sampled RNA capture in a given position, at a given time, of one biological experiment

. . . *ie.* a poor quality photo comparing to real life

closing session



Bonus

Run Sleuth.R as a SGE job

Save the following lines in a file, let's say sleuth_qsub_script.sh :

```
#!/bin/bash             ==> shell to use in local run
# -S /bin/bash          ==> shell to use in SGE run (SGE)
# -N Salmon2Sleuth      ==> set a name to the process (SGE)
# -pe parallel 4        ==> run 4 threads in parallel (SGE)
# -V                    ==> set Verbose mode (SGE)
# -cwd                  ==> set the working diretory (SGE)
bash Sleuth.R
```

Sleuth on SGE

You can now run *locally* with:

```
bash sleuth_qsub_script.sh
```

Or send it to the *SGE queue* with:

```
qsub sleuth_qsub_script.sh
```