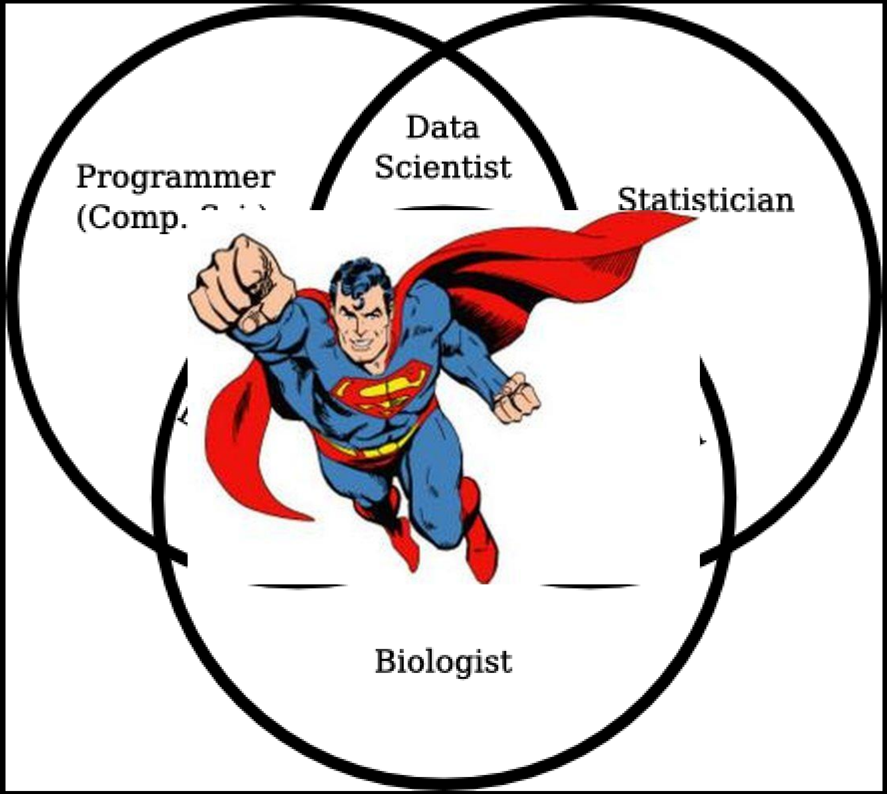
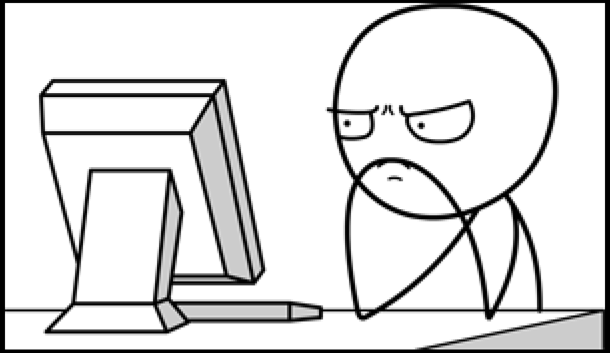
MSIB32500 Advanced Bioinformatics Spring 2017

RNAseq Data Analysis and Clinical Applications, Part I

Riyue Bao, Ph.D.
Research Assistant Professor (Bioinformatics)
Center for Research Informatics & Department of Pediatrics
The University of Chicago


Data
Scientist
Programmer
(Comp. Sci.)
Statistician
Biologist



Outline

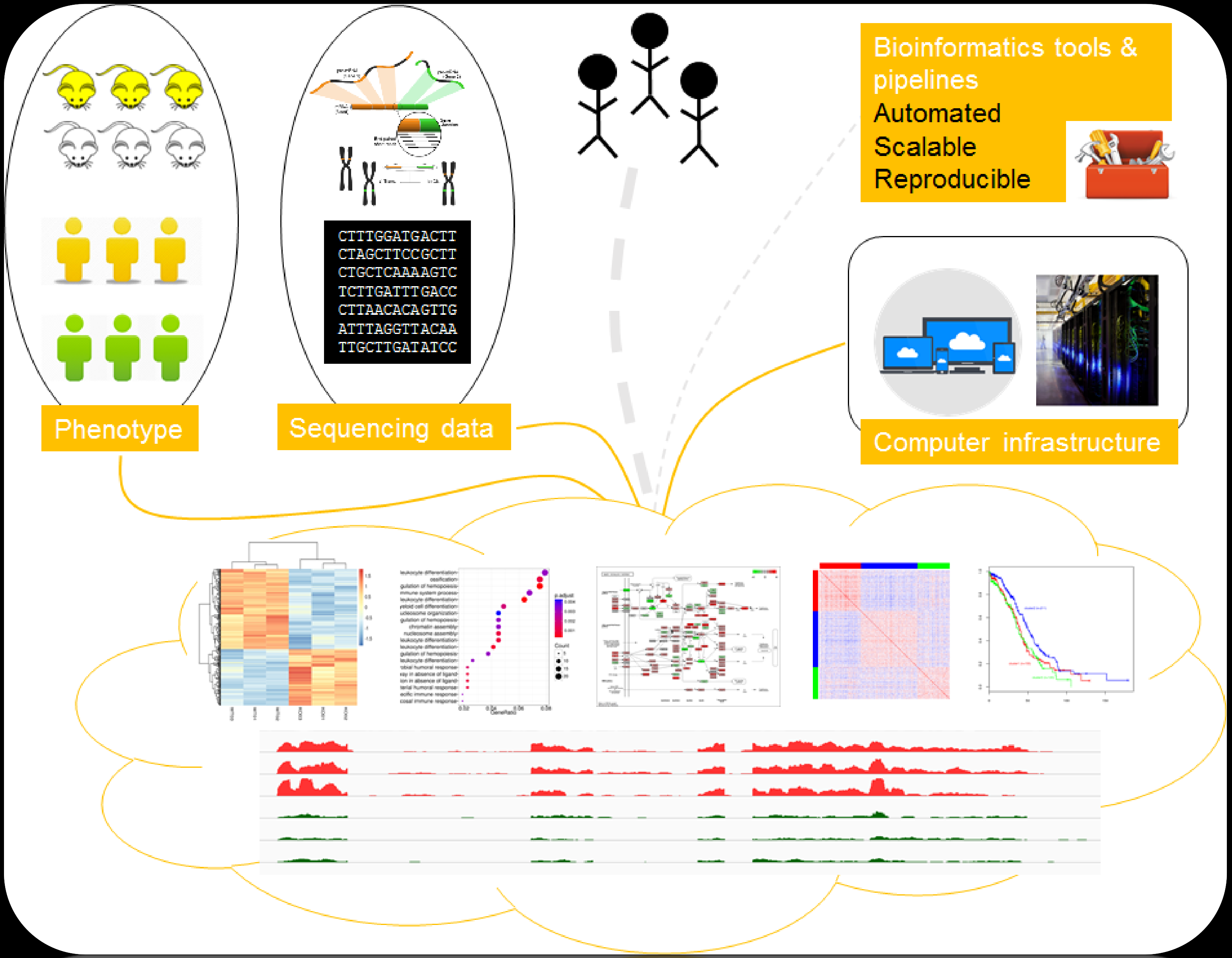
- Part I (05/27/2017)
  - Introduction to RNAseq technology and clinical applications
  - Hands on: From raw data to gene expression quantification
- Part II (06/03/2017)
  - Differential gene expression analysis and data visualization
  - Hands on: Identification of genes and pathways significantly changed under condition
  - ***Homework assignment***
- Part III (06/08/2017)
  - How to associate gene expression data with clinical outcome
  - Hands on: Use gene expression data to discover tumor subtypes and survival analysis

Class materials

- GitHub
  - https://github.com/MScBiomedicalInformatics/MSIB32500
  - This lecture note contains the same contents as the notebook. In addition, the notebook also contains hands-on materials
  - **lecture8.pdf**
  - **Handson8.Rmd**
- Tarbell high-performance computing (HPC) clusters (hands on practice)

Objective

- Learn the good-practice RNAseq analysis pipeline
- Learn commonly used bioinformatics tools
- Practice the automated, scalable pipeline
- Explore the quality metrics and input/output of the RNAseq pipeline
- Visualize result files and quality plots



Phenotype

```
CTTTGGATGACTT
CTAGCTTCCGCTT
CTGCTCAAAAGTC
TCTTGATTTGACC
CTTAACACAGTTG
ATTTAGGTTACAA
TTGCTTGATATCC
```

Sequencing data

Bioinformatics tools & pipelines

Automated

Scalable

Reproducible

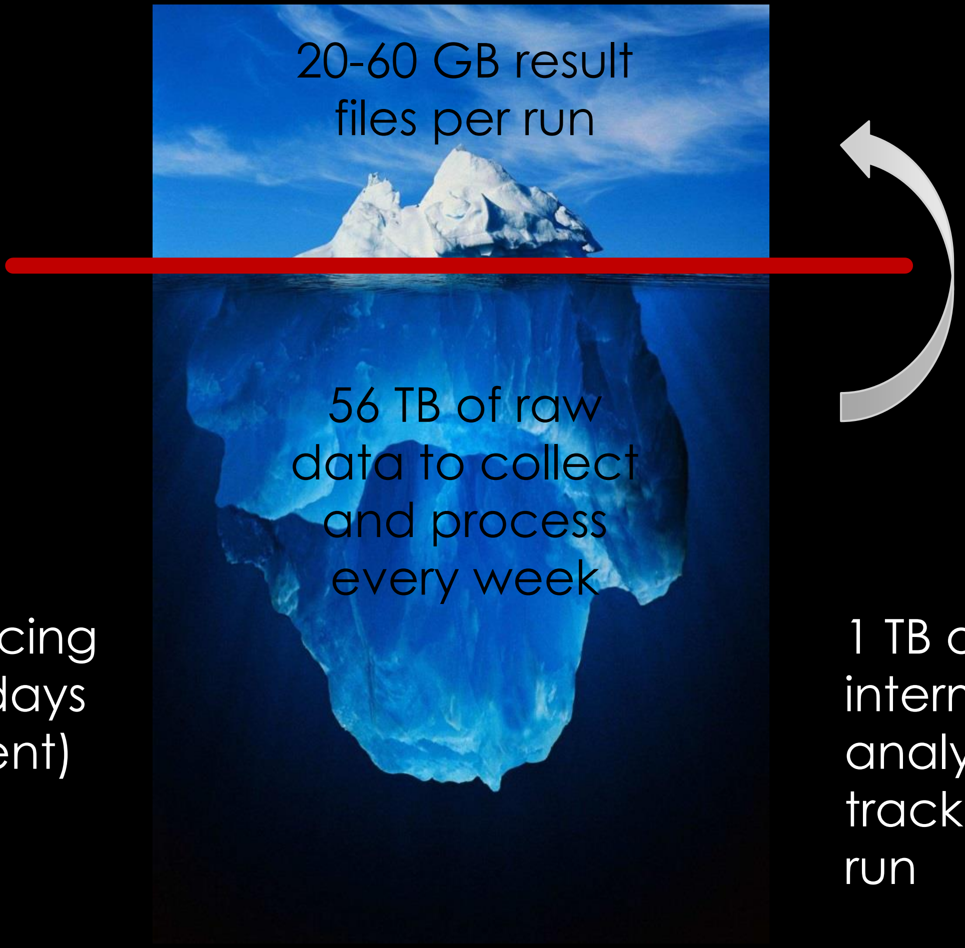
Computer infrastructure

Biological and clinical questions

- I am interested in studying transcriptional landscape shift before and after drug treatment in cell lines
- I want to identify which pathways are affected after knocking down my favorite gene in mice
- I have expression data of clinical isolates collected at various time points, when patient's response changed. Why?
- I have a cohort of patients and want to discover which gene signature predicts patient's response to treatment
- I want to detect gene fusions, expressed mutations, and disrupted isoforms in tumors that may be related to disease

... and more!

The Sequencing Iceberg



20-60 GB result files per run

Data storage
Data transfer
<u>Data analysis</u>

56 TB of raw data to collect and process every week

One sequencing run every 3 days (per instrument)

1 TB of images, intermediate analysis files and tracking data per run

28 instruments

Adopted from Thomas Keane, Wellcome Trust Sanger Institute

Current Sequencing Technologies

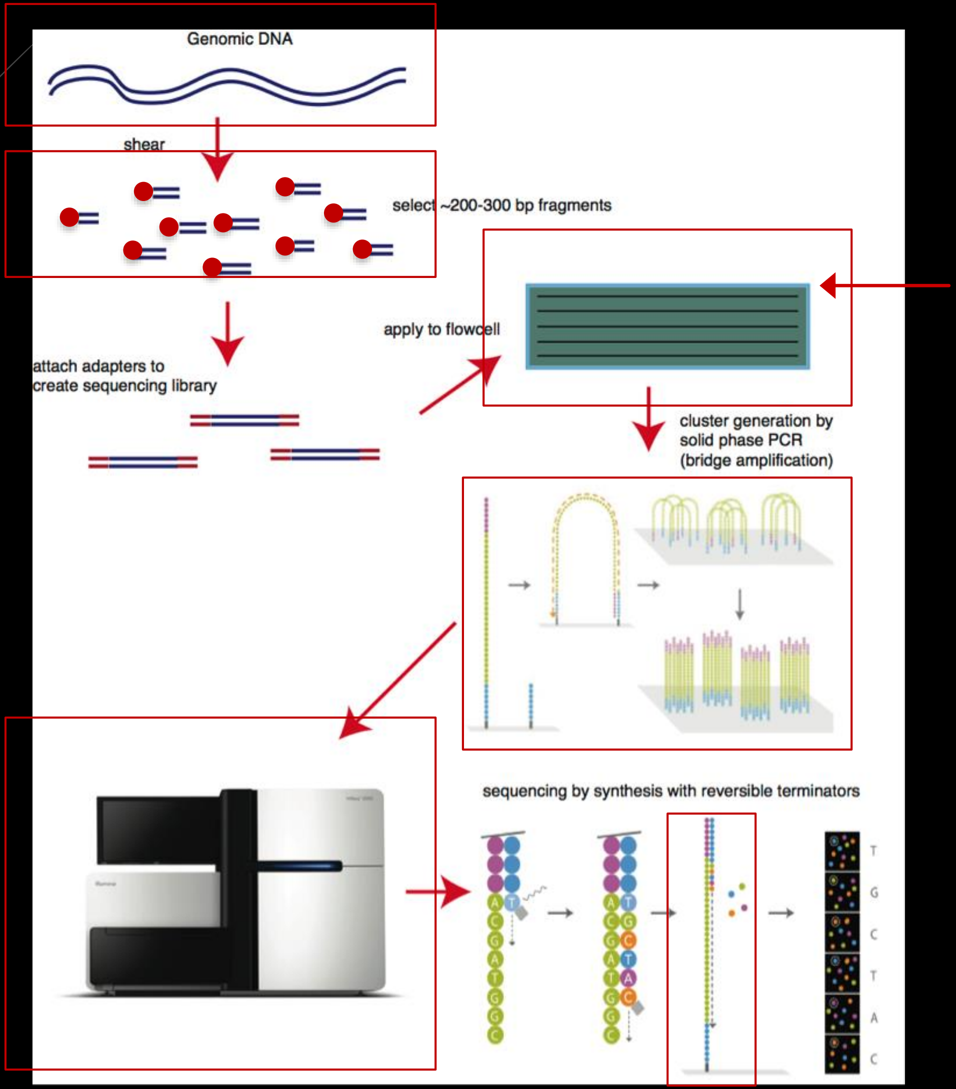
Illumina
HiSeq 2000/2500/X *TEN*

Illumina
MiSeq

Ion Torrent (Life Technologies)
Ion Proton

PacBio
RS

Illumina


Genomic DNA
shear
select ~200-300 bp fragments
attach adapters to create sequencing library
apply to flowcell
cluster generation by solid phase PCR (bridge amplification)
sequencing by synthesis with reversible terminators
T
G
C
T
A
C

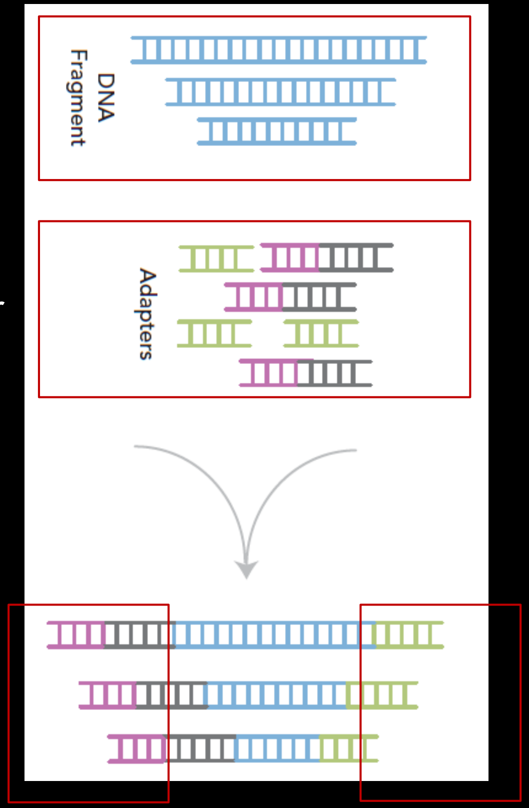
- Sample
- Library
- DNA fragment
- Barcode
- Run
- Flow cell
- Lane
- Cluster
- Read
- Adapter/Primer

Though we are talking about Illumina here, many of those terms can be applied to other sequencing technologies.

Illumina

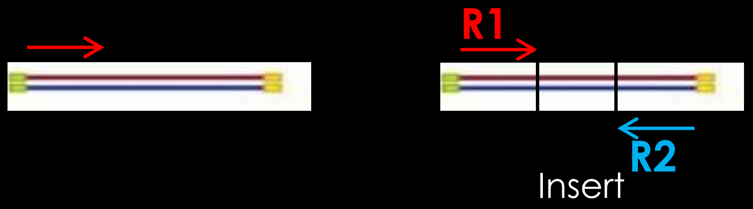
DNA fragment

Adapter/Primer



Single-End

Paired-End



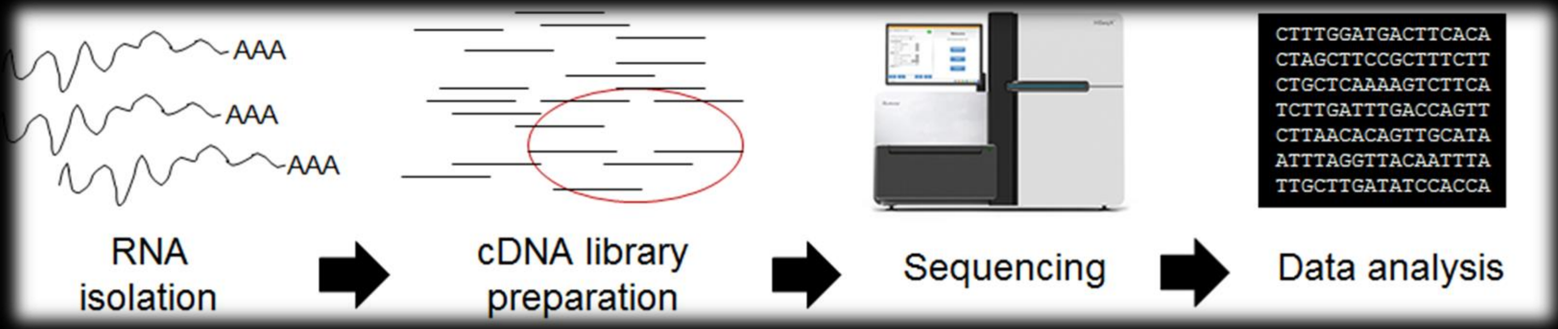
- Sample
- Library
- DNA fragment
- Barcode
- Run
- Flow cell
- Lane
- Cluster
- Read
- Adapter/Primer

Fragment size = R1+R2+insert size

What is RNAseq?

High-throughput sequencing of RNA: Profile, identify or assemble transcripts

- Detect gene expression changes between conditions
- Identify novel splice sites / exons, mutations, fusion genes, etc.
- Broad detection range, high sensitivity, low requirement of RNA amount
- Available for all species (reference genome is optional): reference genome-guided alignment or *de novo* assembly



RNA isolation → cDNA library preparation → Sequencing → Data analysis

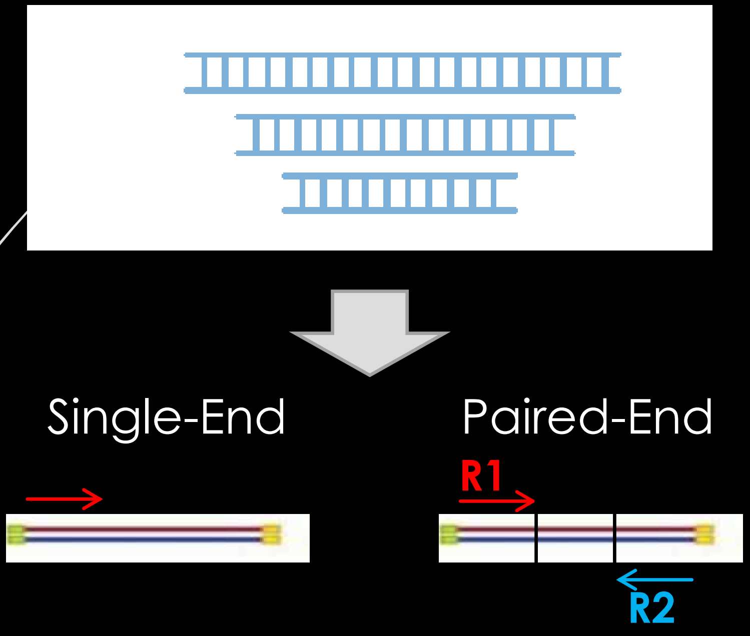
Biological sample or clinical specimen → *Millions of reads!*

Which factors to consider if I want to initiate an RNAseq experiment?

- Sample source (fresh cells, frozen, FFPE)
- Sample quality (RNA quality, DNA contamination, tumor purity)
- RNA concentration (e.g. 100ng total RNA)
- Ribosome RNA depletion (accounts for 80% of total RNA)
- **Library type (single-end or paired-end reads)**
- **Library strandness (unstranded or stranded/directional)**
- **Sequencing depth (20 million, 50 million, or >100 million reads)**

Library type: Single-end vs paired-end reads

DNA fragment

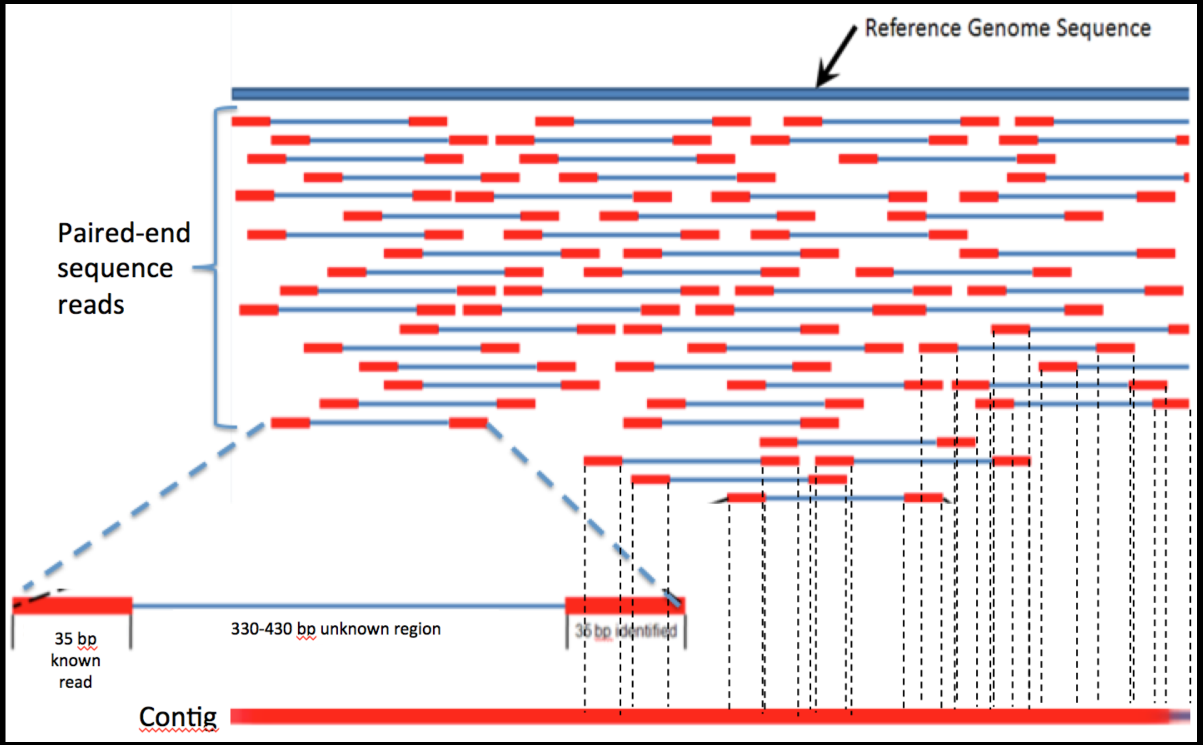


Single-End

Paired-End

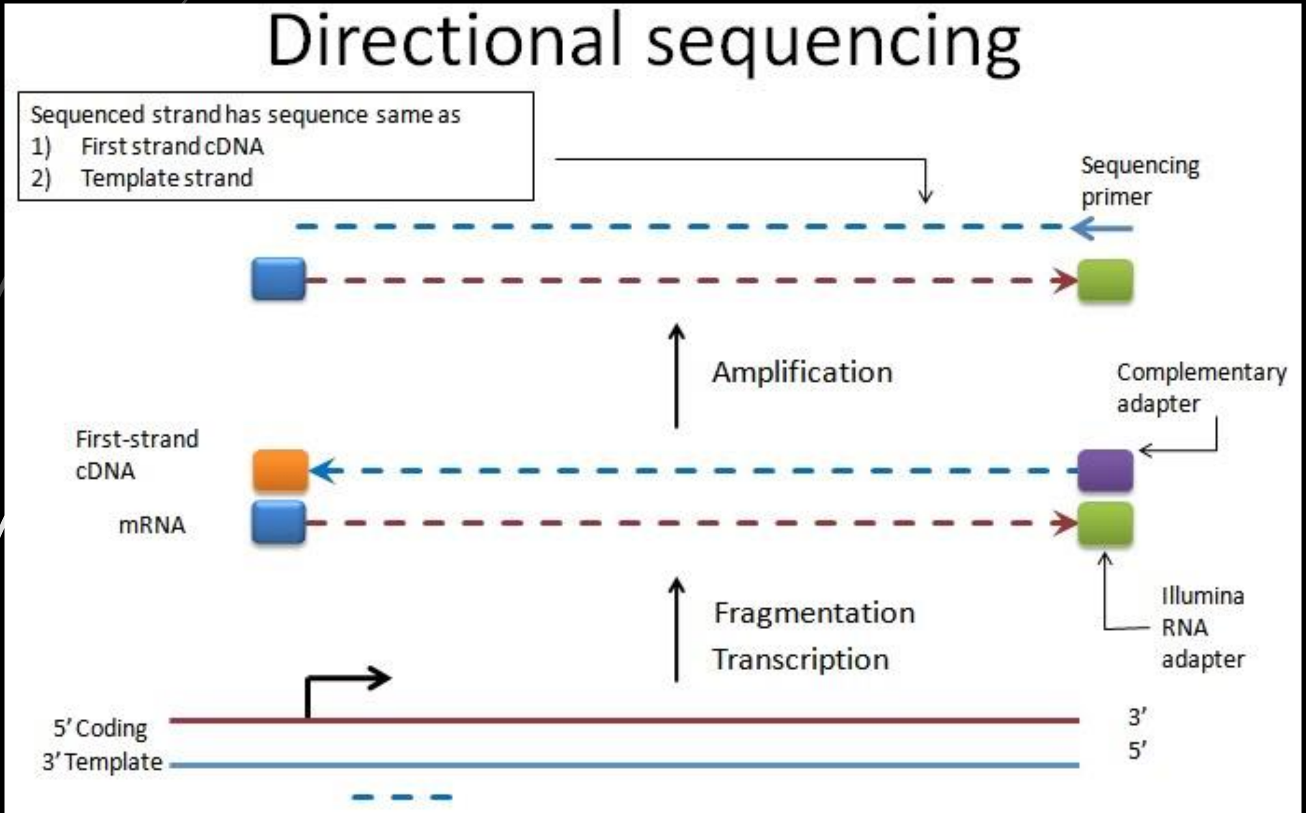
Fragment size = R1+R2+insert size

PE reads: Higher alignment rate and accuracy


Reference Genome Sequence
Paired-end sequence reads
35 bp known read
330-430 bp unknown region
35 bp identified
Contig

2 x 50bp PE reads >> 1 x 100 bp SE reads!

Library strandness: stranded-protocol is always recommended!

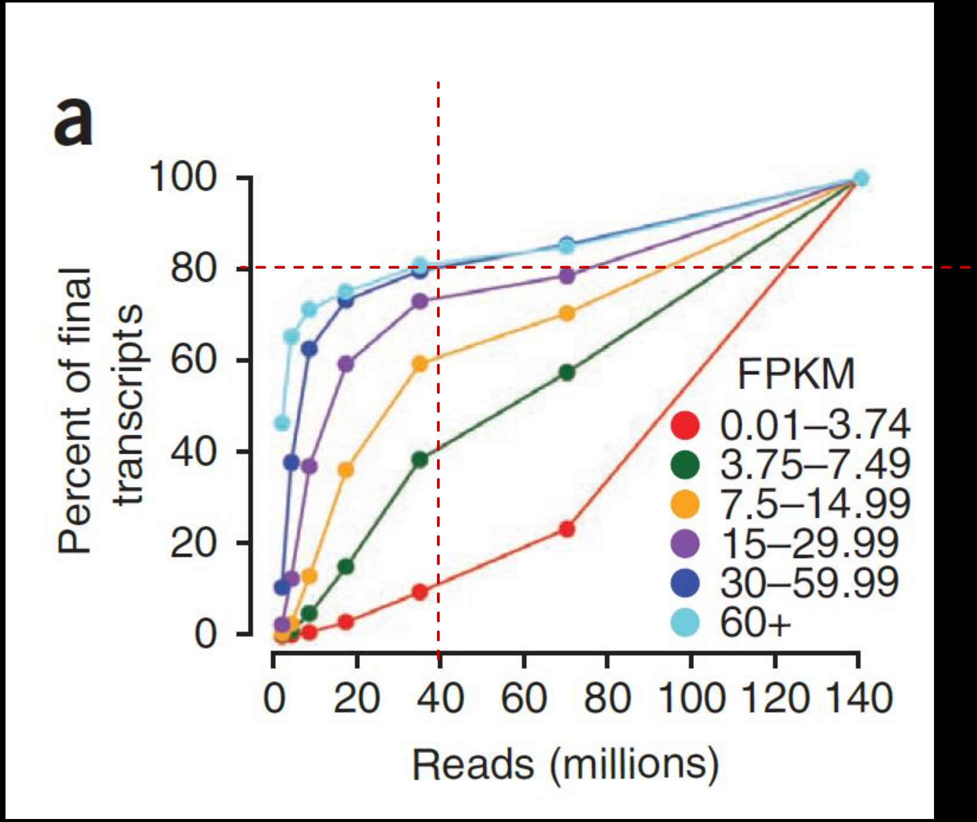


- Non-stranded/Non-directional
- Forward, also known as second-stranded (coding strand)
- Reverse, also known as first-stranded (template strand)

Sequencing depth: How many reads do I need?

Coverage affects downstream gene discovery, expression estimation, and power of statistical analysis

The more, the better!



Trapnell et al. Nature Protocol 2012

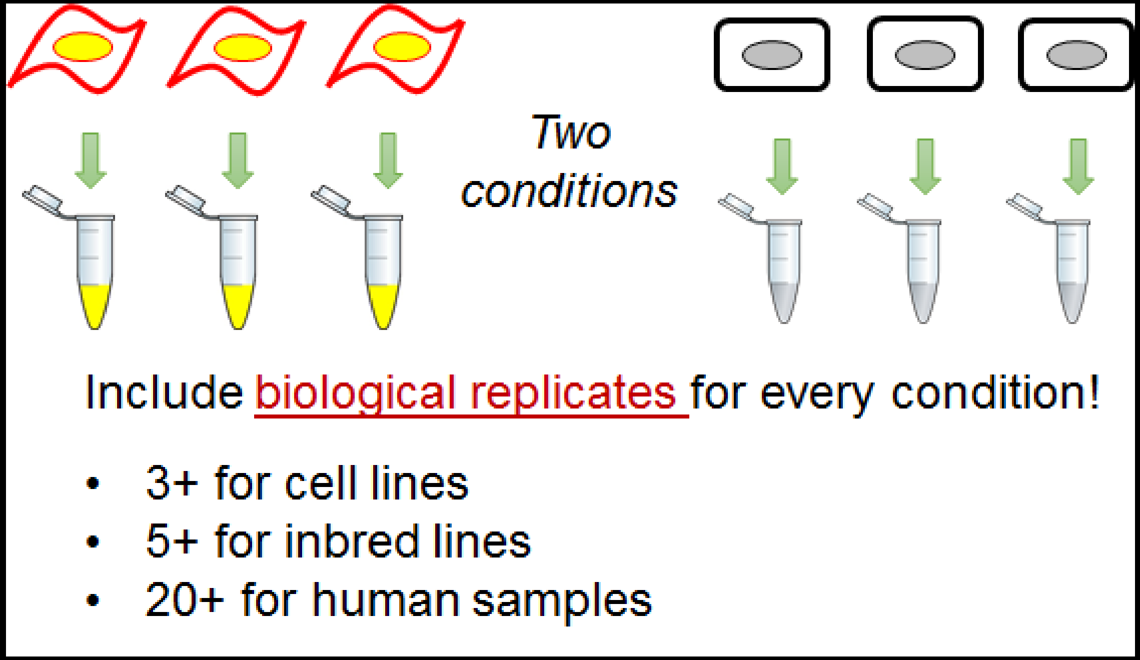
ENCODE saturation analysis

- 214 million 2x100bp PE reads
- H1 human embryonic stem cells
- 80% of the genes with FPKM≥10 are detected by ~36 million mapped reads per sample
- Genes with FPKM<10: ~80 million mapped reads per sample

Sims et al., 2014 Nature Reviews

Experimental design: Biological replicates

Include biological replicates for **increased discovery power** and reduced false positives/negatives!



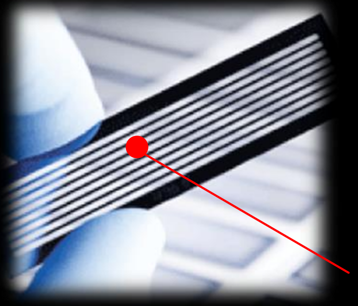
Include biological replicates for every condition!

- 3+ for cell lines
- 5+ for inbred lines
- 20+ for human samples

Experimental design: Multiplexing and Randomization

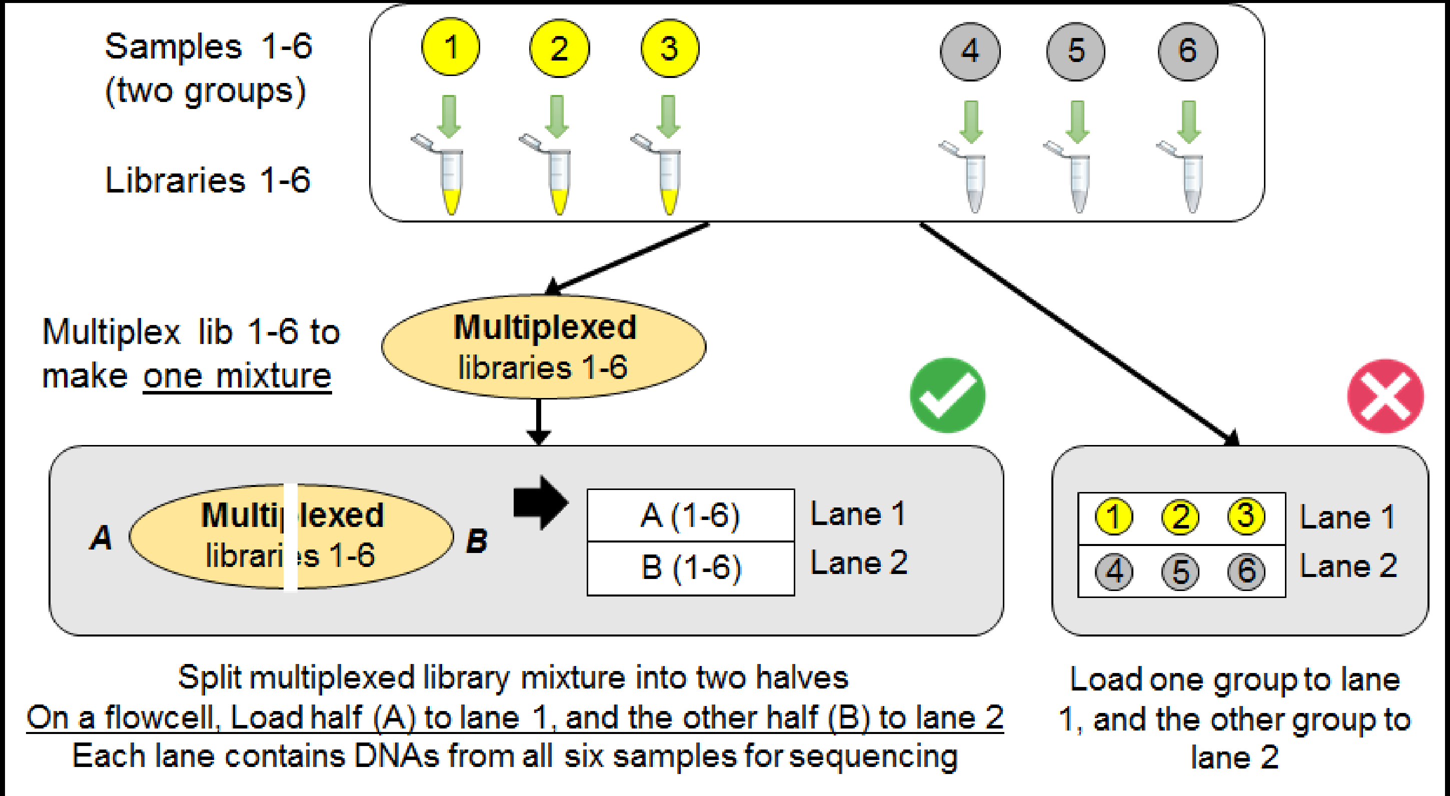
- **Multiplexing**: simultaneously measures multiple libraries in one sequencing lane. Unique barcodes are added to label DNA molecules from each library

- **Randomization**: Avoid loading samples from the same biological group in the same sequencing lane. Minimizes technical bias and lane-specific effects.



Sequencing lanes

Experimental design: Multiplexing and Randomization

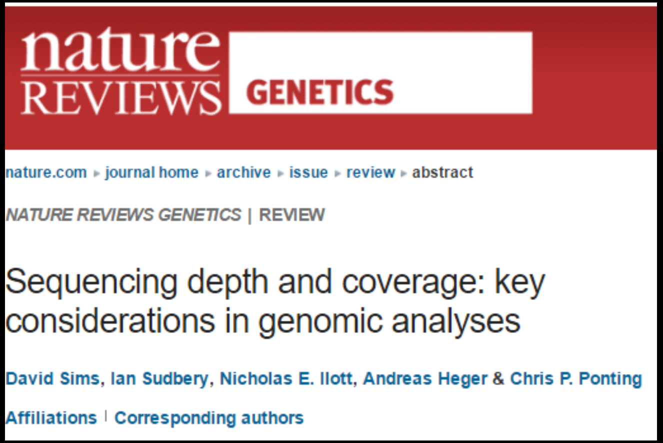


Split multiplexed library mixture into two halves
On a flowcell, Load half (A) to lane 1, and the other half (B) to lane 2
Each lane contains DNAs from all six samples for sequencing

Load one group to lane 1, and the other group to lane 2

Challenges and limitations

- Relatively poor RNA quality for tumor FFPE samples
- Contamination from adjacent normal tissue
- Still more expensive than targeted-panel sequencing such as NanoString
- 40 million mapped reads are usually sufficient for gene profiling, but > 80 million are required to detect bottom 1% lowly expressed genes



nature REVIEWS GENETICS

nature.com ▸ journal home ▸ archive ▸ issue ▸ review ▸ abstract

NATURE REVIEWS GENETICS | REVIEW

Sequencing depth and coverage: key considerations in genomic analyses

David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger & Chris P. Ponting

Affiliations | Corresponding authors

Whom to contact if I want to initiate an NGS (e.g. RNAseq) experiment?

- The University of Chicago Genomics Facility

Pieter W. Faber, Ph.D.
Technical Director
pfaber@bsd.uchicago.edu
773.834.8420

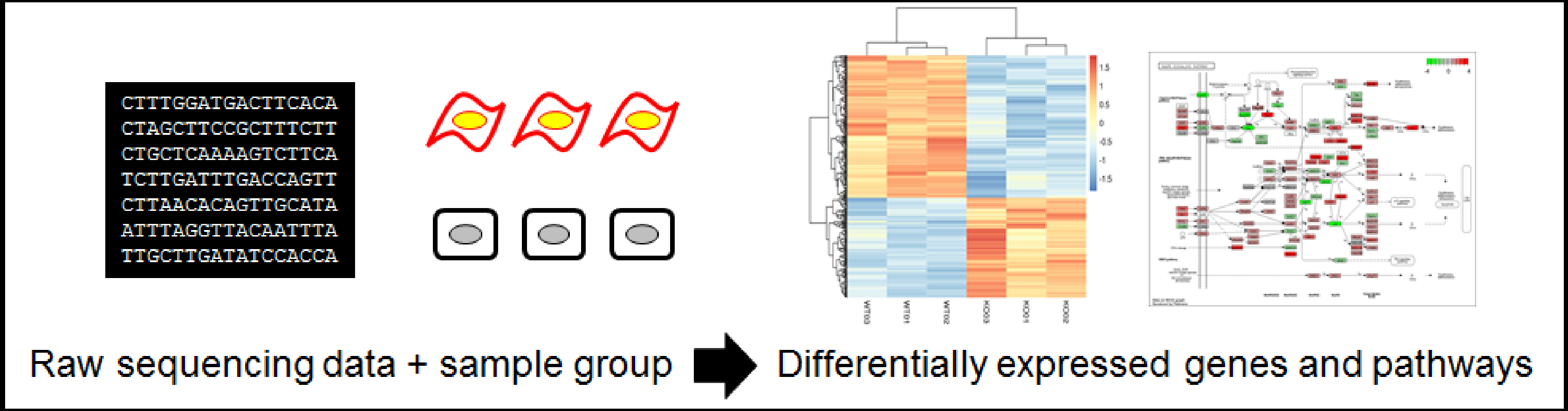
- The Center for Research Informatics Bioinformatics Core

Jorge Andrade, Ph.D.
Director of Bioinformatics
jandrade@bsd.uchicago.edu

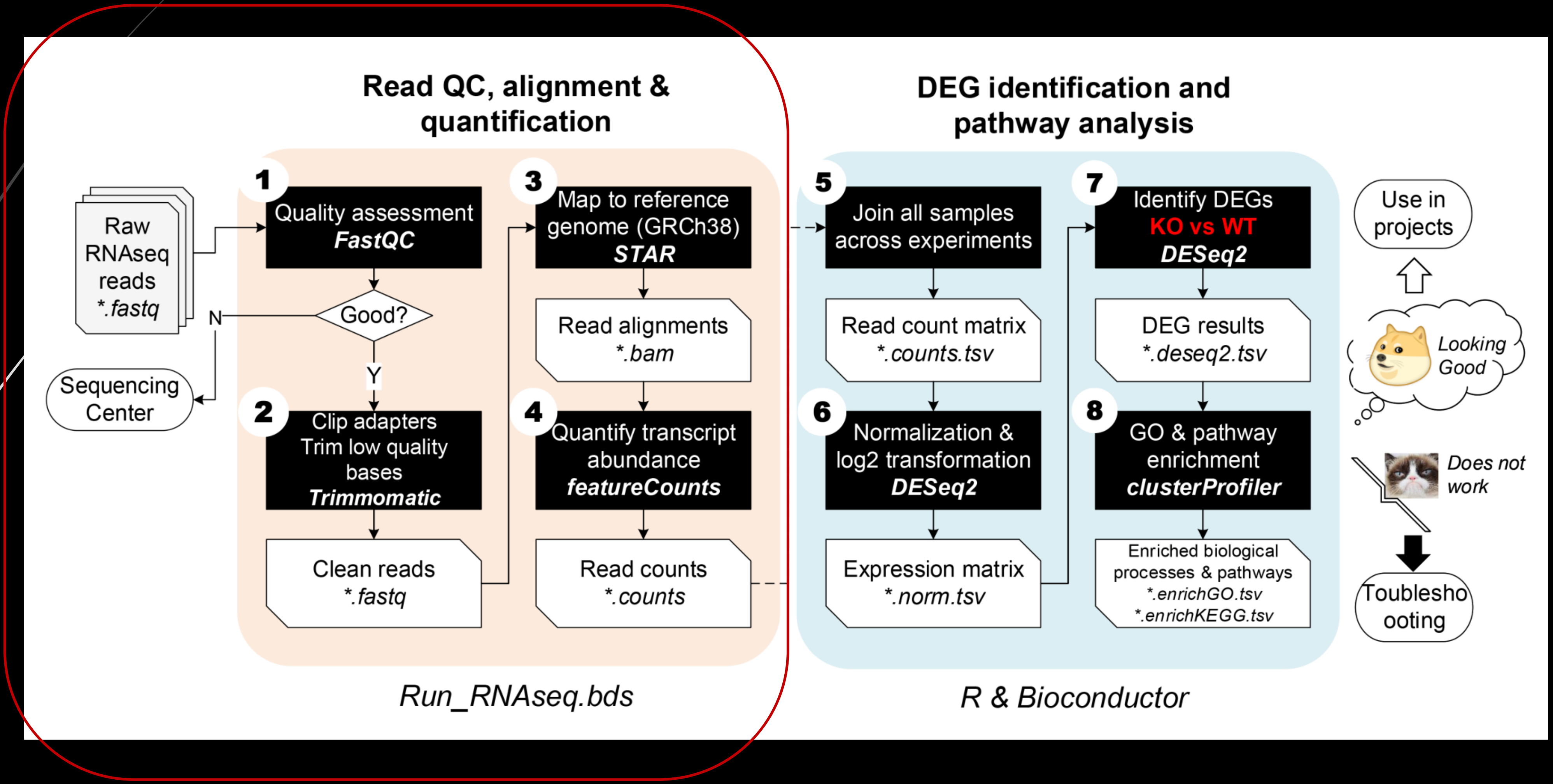
How to perform RNAseq analysis

The good-practice analysis protocol takes 8 major steps.

- **01-04**: From raw sequencing to transcript quantification
- **05-08**: DEG and pathway analysis (06/03, part II)



How to perform RNAseq analysis



Read QC, alignment & quantification

Raw RNAseq reads **.fastq*

1 Quality assessment ***FastQC***

Good?

N

Sequencing Center

Y

2 Clip adapters Trim low quality bases ***Trimmomatic***

Clean reads **.fastq*

3 Map to reference genome (GRCh38) ***STAR***

Read alignments **.bam*

4 Quantify transcript abundance ***featureCounts***

Read counts **.counts*

Run_RNAseq.bds

DEG identification and pathway analysis

5 Join all samples across experiments

Read count matrix **.counts.tsv*

6 Normalization & log2 transformation ***DESeq2***

Expression matrix **.norm.tsv*

7 Identify DEGs **KO vs WT** ***DESeq2***

DEG results **.deseq2.tsv*

8 GO & pathway enrichment ***clusterProfiler***

Enriched biological processes & pathways **.enrichGO.tsv* **.enrichKEGG.tsv*

R & Bioconductor

Use in projects

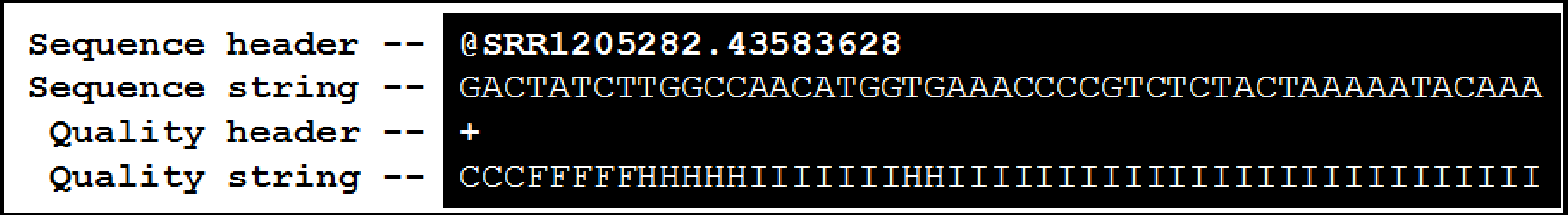
Looking Good

Does not work

Toublesho ooting

01 Quality assessment of raw sequencing reads: FastQC

- Raw sequencing reads are stored in **FastQ** format (e.g. *KO01.fastq.gz*), where each read is presented by 4 lines



```
Sequence header --  @SRR1205282.43583628
Sequence string --  GACTATCTTGGCCAACATGGTGAAACCCCGTCTCTACTAAAAATACAAA
 Quality header --  +
 Quality string --  CCCFFFFFHHHHHIIIIIIIHHIIIIIIIIIIIIIIIIIIIIIIIIIII
```

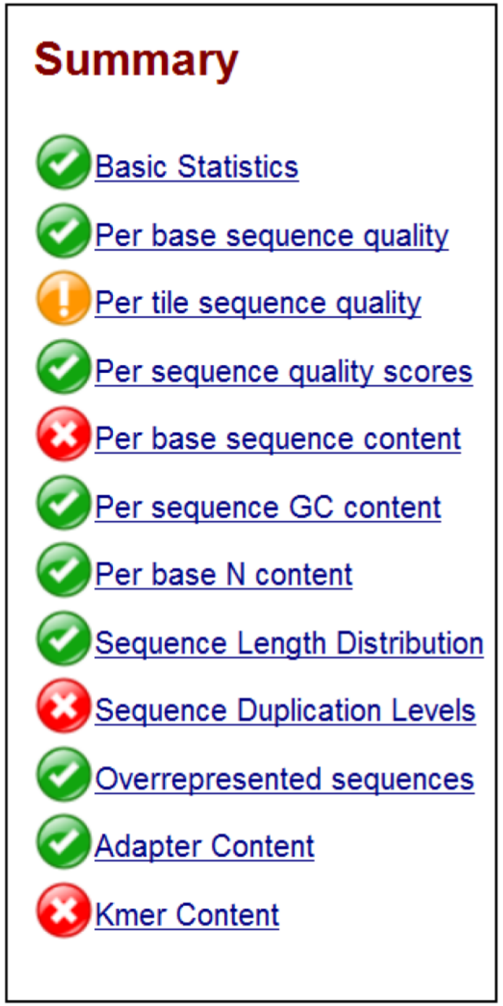
- Check if the reads are of high quality
- Check if any preprocessing step is required (e.g. base trimming, adapter clipping, read filtering)

01 Quality assessment of raw sequencing reads: FastQC

- Method
  - FastQC version 0.11.5
  - Scan raw sequencing reads and produce QC reports for evaluation
- Our data
  - Read quality pretty good (baseQ >= 30 in all base positions)
  - Preprocessing is optional

```
fastqc --extract -o $out.dir -t 2 --nogroup $r1.fastq.gz
```

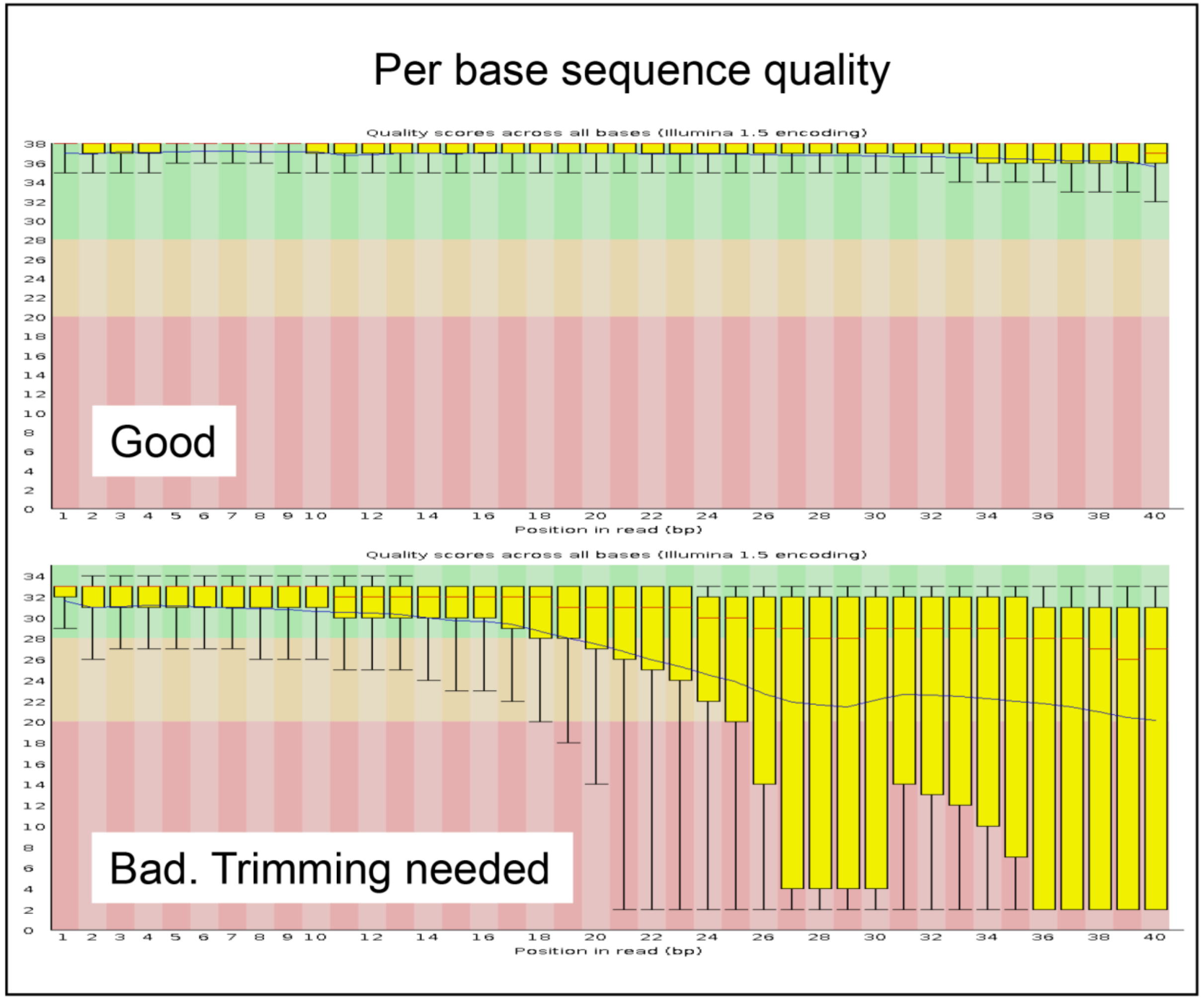
01 Quality assessment of raw sequencing reads: FastQC



Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ! Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

Per base sequence quality

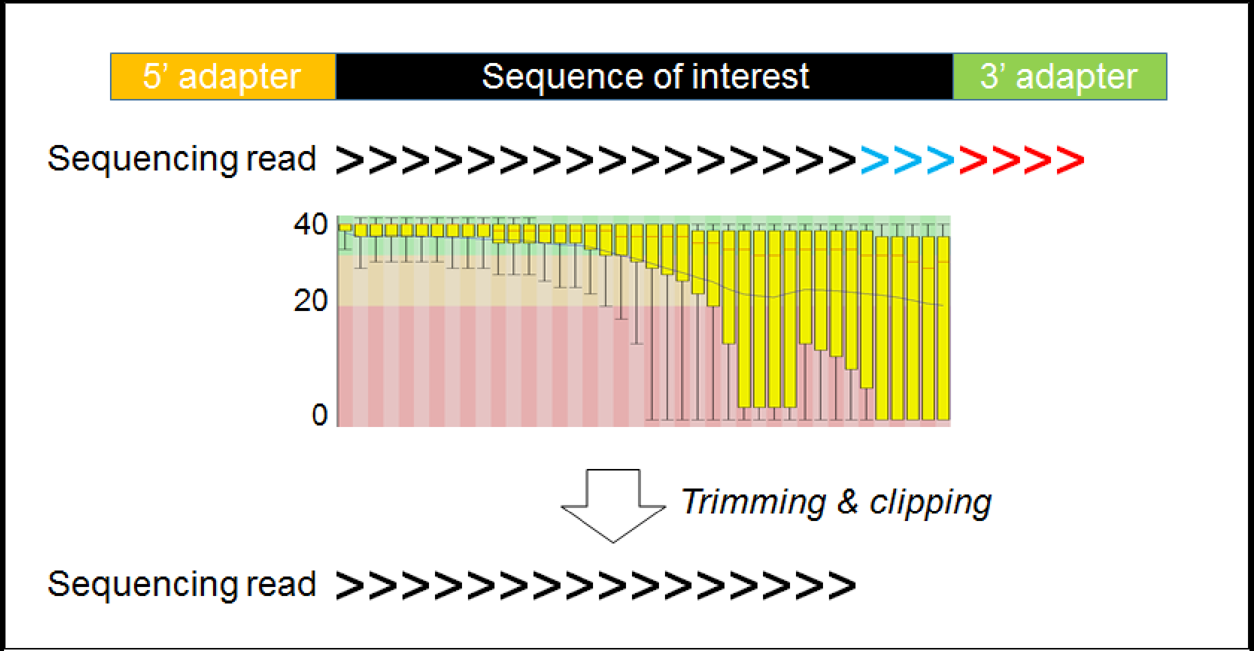


Good

Bad. Trimming needed

02 Preprocessing: Trimmomatic

- Preprocess reads to improve mapping rate and accuracy
  - Trim low-quality bases, clip adapters, etc.
  - Avoid over-trimming in RNAseq!
- Clean up reads for improved alignment rate and accuracy (for the next step)


5’ adapter
Sequence of interest
3’ adapter
Sequencing read >>>>>>>>>>>>>>>>>>>>>>>
40
20
0
Trimming & clipping
Sequencing read >>>>>>>>>>>>>>>>>

02 Preprocessing: Trimmomatic

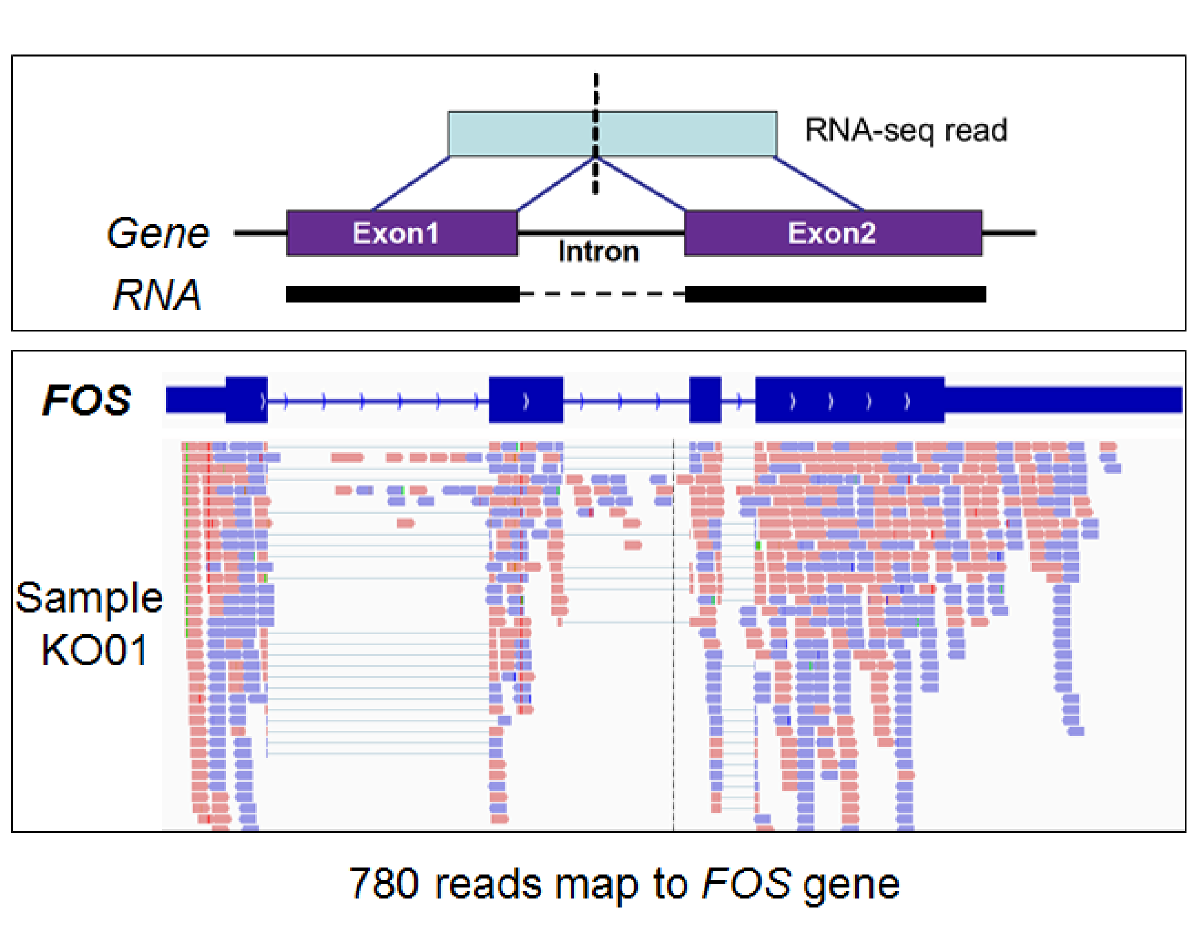
- Method
  - Trimmomatic version 0.36
  - Clip adapters
  - Trim leading/trailing low quality or N bases
  - Trim reads to a specific length
  - Filter out reads of low average quality / of specific length
  - Convert base quality scores between Phred33 and Phred64 FastQ format
- Our data (KO01 as an example)
  - 74,126,025 reads total. Survived: 73,636,793 (99.34%) Dropped: 489,232 (0.66%)

```
java -Xmx4G -jar trimmomatic-0.36.jar SE -threads 4 -phred33
$r1.fastq.gz $r1.trim.fastq.gz ILLUMINACLIP:TruSeq3-
SE.fa:2:30:10 LEADING:5 TRAILING:5 MINLEN:36 SLIDINGWINDOW:4:15
```

03 Map reads to reference genome (GRCh38): STAR

- Read mapping identifies the location in the genome where a sequencing read comes from
- **Splice-aware aligner** (e.g. *STAR*)

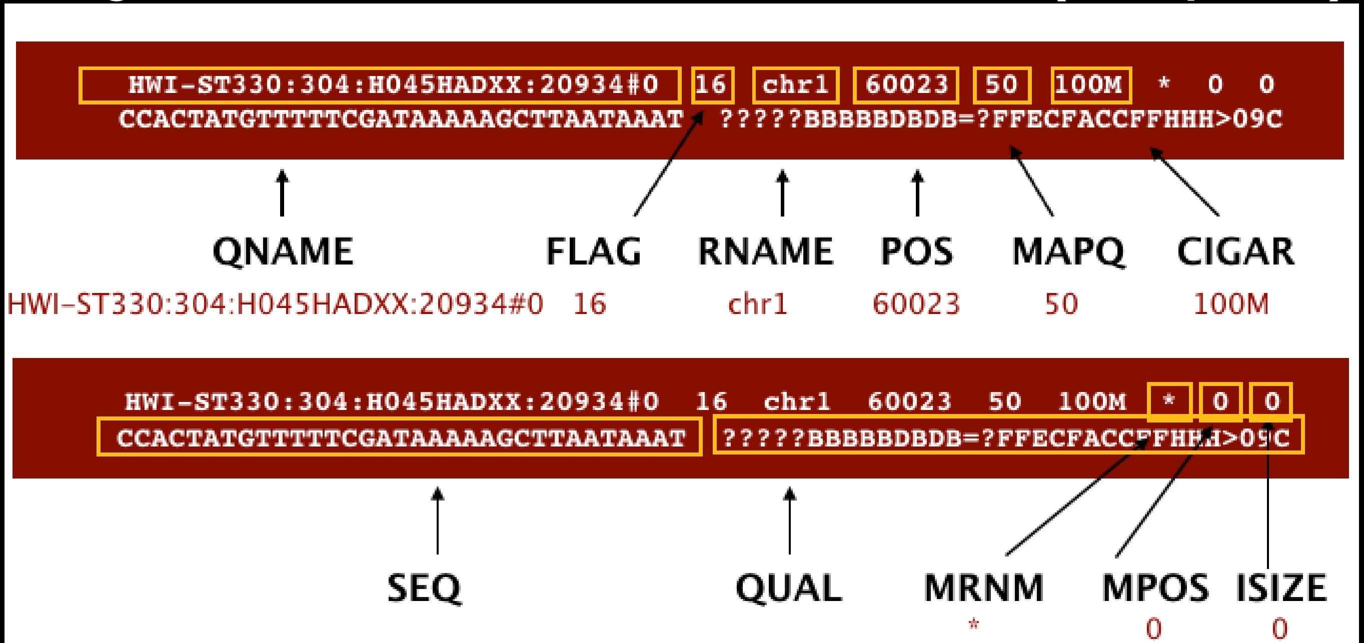
Each horizontal bar represents one read. Red/blue indicates reads aligned to plus/minus strand on the genome, respectively.



780 reads map to *FOS* gene

03 Map reads to reference genome (GRCh38): STAR

- A read may map to one, or multiple genomic locations (Mapping quality: MAPQ)
- Accurate mapping result is the key for downstream differential gene expression identification
- Read alignment results are stored in **SAM or BAM (Binary SAM)** format



```
HWI-ST330:304:H045HADXX:20934#0  16  chr1  60023  50  100M  *  0  0
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT  ?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

```
HWI-ST330:304:H045HADXX:20934#0  16  chr1  60023  50  100M  *  0  0
CCACTATGTTTTTCGATAAAAAGCTTAATAAAT  ?????BBBBBDBDB=?FFECFACCFFHHH>09C
```

03 Map reads to reference genome (GRCh38): STAR

- Method
  - STAR version 2.5.3a
  - splice-aware aligner
  - Ultrafast (~15 minutes for 50m reads), require lots of memory (~36GB for human genome)
  - Flexible options to allow canonical/non-canonical junctions, with/wo known gene annotations, etc.
- Different aligners may generate very different results! (Engström et al. *Nature methods* 2013. Systematic evaluation of spliced alignment programs for RNA-seq data)

```
STAR --runMode alignReads --genomeLoad NoSharedMemory --
outFileNamePrefix $out.prefix --readFilesCommand zcat --
genomeDir $refgenome.dir --readFilesIn $r1.trim.fastq.gz --
runThreadN 2 --outSAMstrandField intronMotif --
outFilterIntronMotifs RemoveNoncanonicalUnannotated --
outSAMtype BAM SortedByCoordinate
```

Reference genome

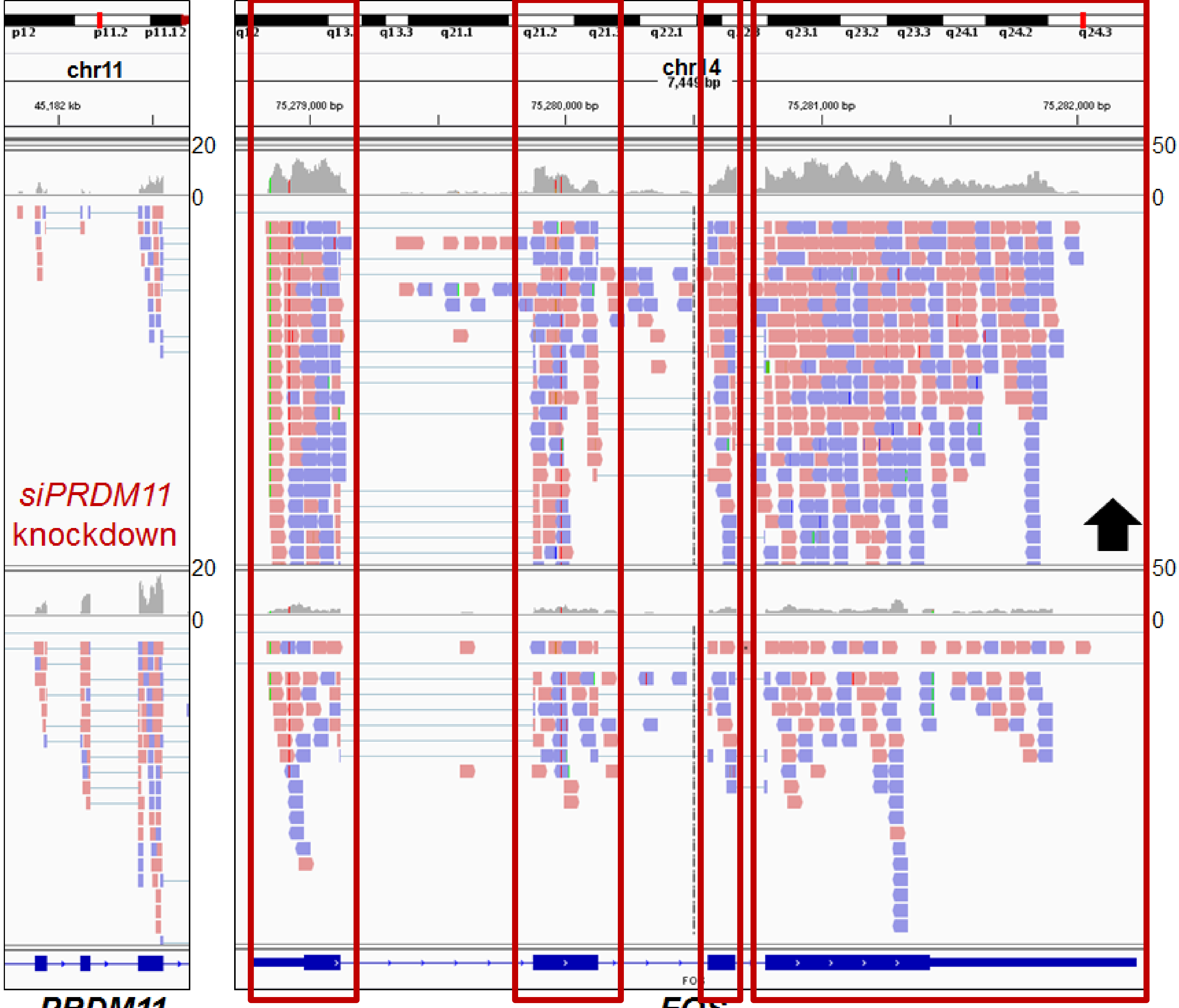
KO01 coverage

KO01 alignment

WT01 coverage

WT01 alignment

Gene



siPRDM11 knockdown

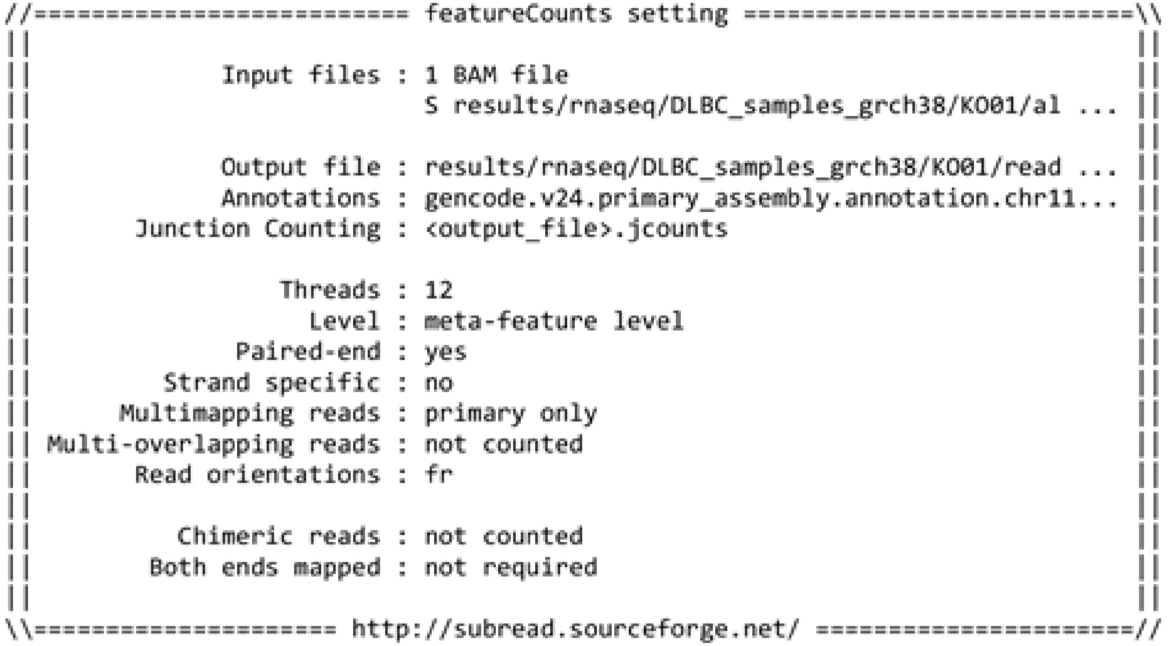
PRDM11
Exons 4,5,6

FOS
Full gene

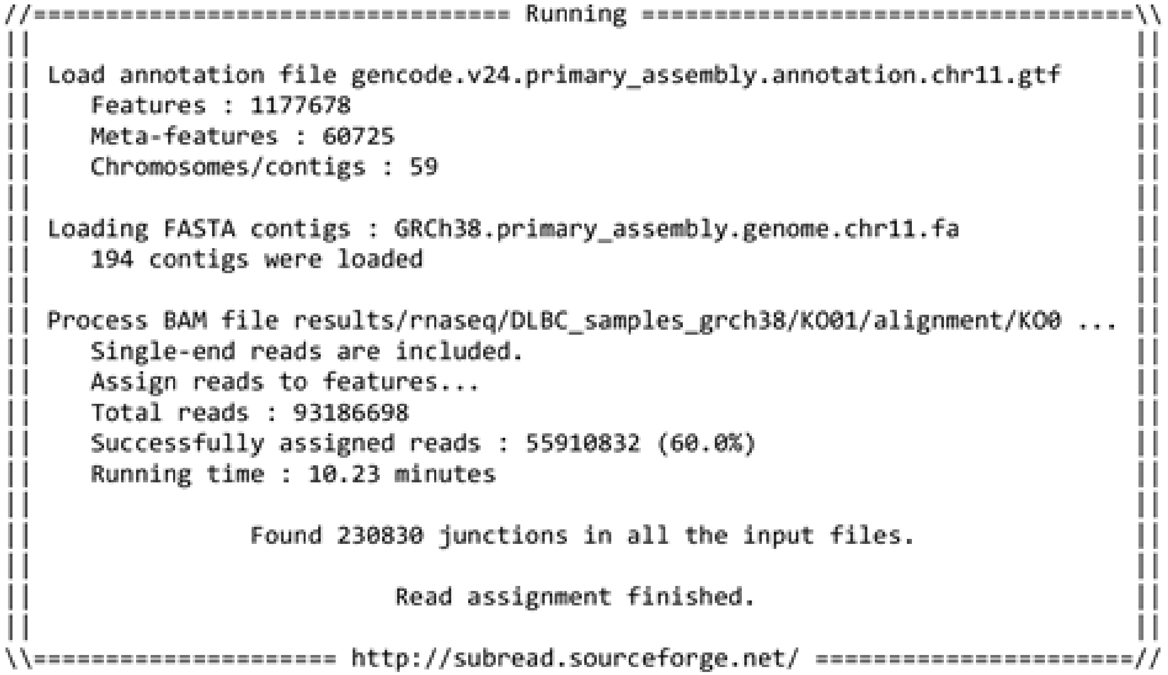
04 Quantify transcript abundance: featureCounts

- Estimate number of reads mapped to gene features (e.g. gene, exon, etc.)
- Method
  - featureCounts version 2.5.2b
  - Ultrafast (~10 minutes for 50m reads), require low amount of memory (~4GB for human genome)
  - Flexible options to count the reads based on specific mapping criteria or study purposes
    - Gene-level or exon-level (for isoforms); Uniquely mapped reads; Primary alignment; Properly paired reads (if reads are paired-end); Mapping quality thresholds
- Choose the option accordingly based on your experimental design!

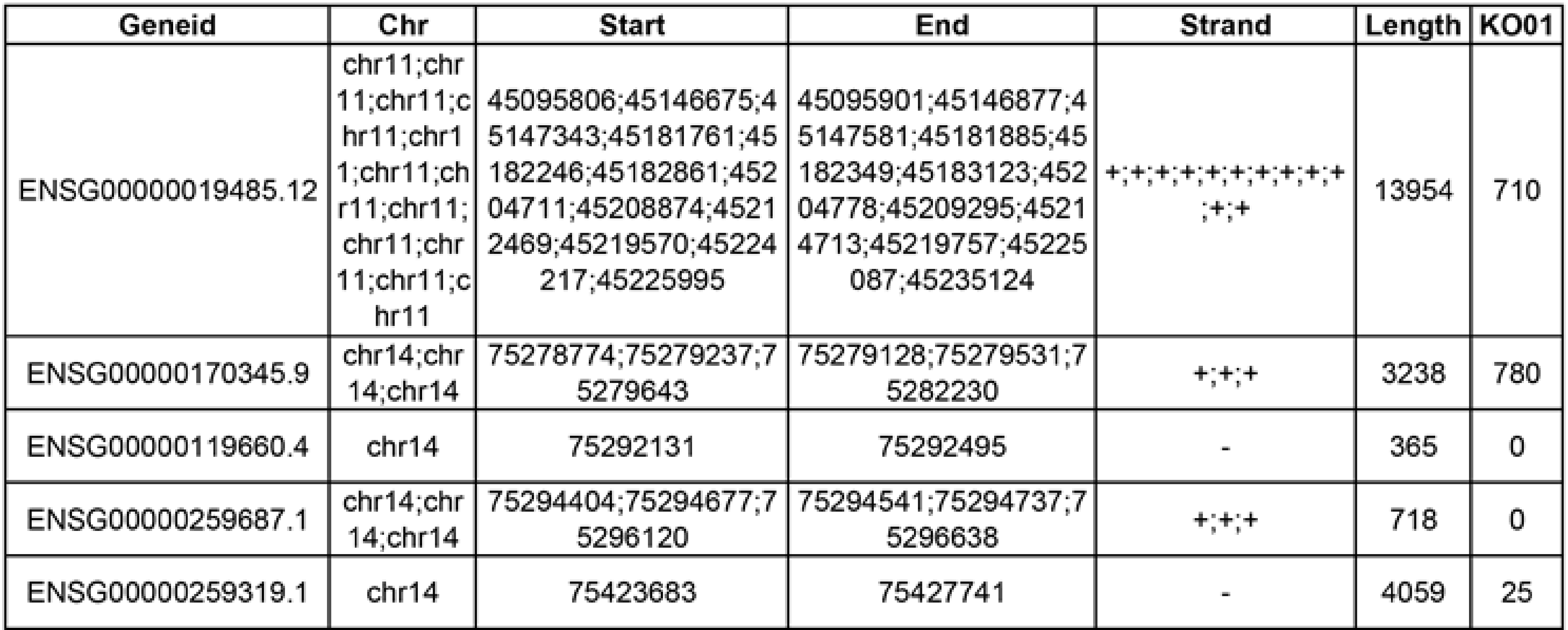
```
featureCounts -s 0 -F GTF -t exon -g gene_id -Q 255 -J --primary -a
$refgeneanno.gtf -T 2 -G $refgenome.fa -o
$sample.star.featurecounts.raw_counts.txt $sample.star.merged.bam
```



```
//========================== featureCounts setting ===========================\\
||                                                                            ||
||             Input files : 1 BAM file                                       ||
||                           S results/rnaseq/DLBC_samples_grch38/KO01/al ... ||
||                                                                            ||
||             Output file : results/rnaseq/DLBC_samples_grch38/KO01/read ... ||
||             Annotations : gencode.v24.primary_assembly.annotation.chr11... ||
||       Junction Counting : <output_file>.jcounts                            ||
||                                                                            ||
||                 Threads : 12                                               ||
||                   Level : meta-feature level                               ||
||              Paired-end : yes                                              ||
||         Strand specific : no                                               ||
||      Multimapping reads : primary only                                     ||
|| Multi-overlapping reads : not counted                                      ||
||       Read orientations : fr                                               ||
||                                                                            ||
||          Chimeric reads : not counted                                      ||
||        Both ends mapped : not required                                     ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//
```



```
//================================= Running ==================================\\
||                                                                            ||
|| Load annotation file gencode.v24.primary_assembly.annotation.chr11.gtf     ||
||    Features : 1177678                                                      ||
||    Meta-features : 60725                                                   ||
||    Chromosomes/contigs : 59                                                ||
||                                                                            ||
|| Loading FASTA contigs : GRCh38.primary_assembly.genome.chr11.fa            ||
||    194 contigs were loaded                                                 ||
||                                                                            ||
|| Process BAM file results/rnaseq/DLBC_samples_grch38/KO01/alignment/KO0 ... ||
||    Single-end reads are included.                                          ||
||    Assign reads to features...                                             ||
||    Total reads : 93186698                                                  ||
||    Successfully assigned reads : 55910832 (60.0%)                          ||
||    Running time : 10.23 minutes                                            ||
||                                                                            ||
||               Found 230830 junctions in all the input files.               ||
||                                                                            ||
||                          Read assignment finished.                         ||
||                                                                            ||
\\===================== http://subread.sourceforge.net/ ======================//
```



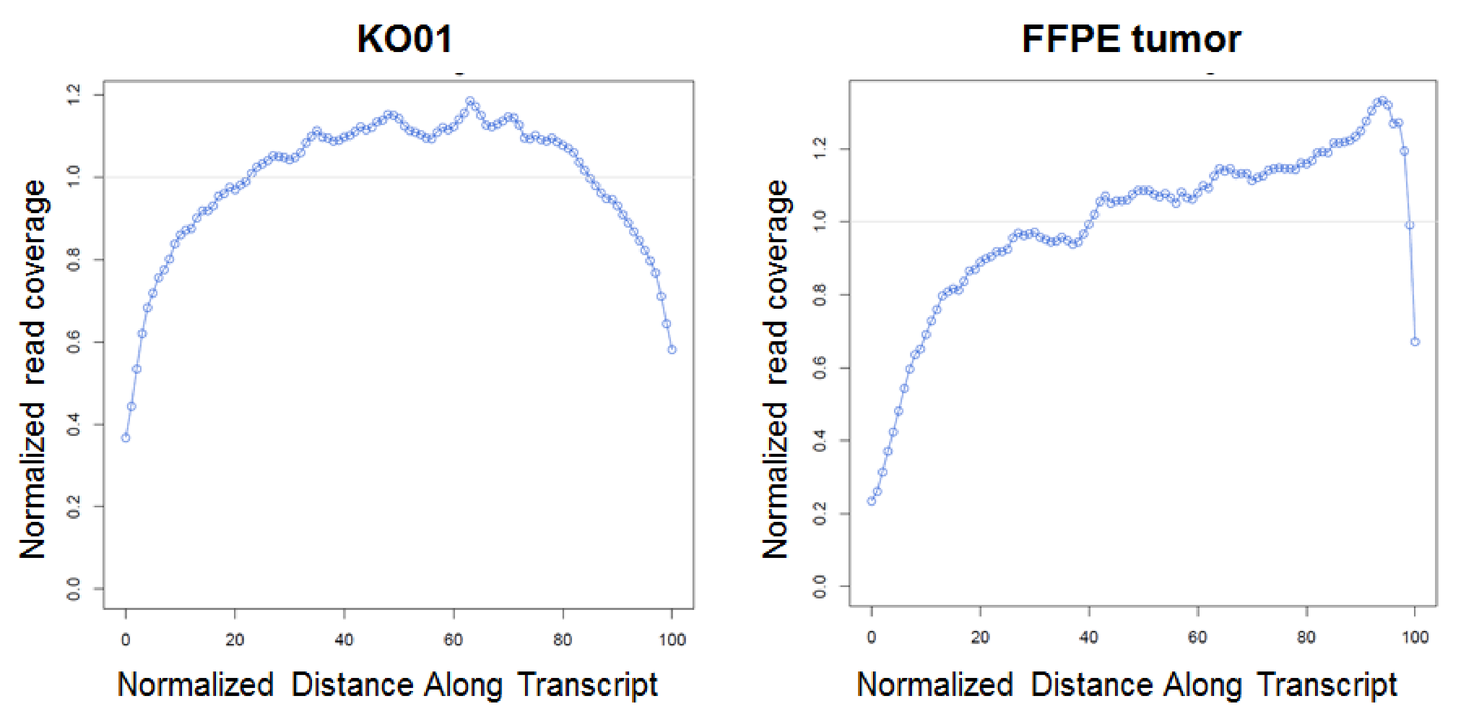
Geneid	Chr	Start	End	Strand	Length	KO01
ENSG00000019485.12	chr11;chr11;chr11;chr11;chr11;chr11;chr11;chr11;chr11;chr11;chr11;chr11	45095806;45146675;45147343;45181761;45182246;45182861;45204711;45208874;45212469;45219570;45224217;45225995	45095901;45146877;45147581;45181885;45182349;45183123;45204778;45209295;45214713;45219757;45225087;45235124	+;+;+;+;+;+;+;+;+;+;+;+	13954	710
ENSG00000170345.9	chr14;chr14;chr14	75278774;75279237;75279643	75279128;75279531;75282230	+;+;+	3238	780
ENSG00000119660.4	chr14	75292131	75292495	-	365	0
ENSG00000259687.1	chr14;chr14;chr14	75294404;75294677;75296120	75294541;75294737;75296638	+;+;+	718	0
ENSG00000259319.1	chr14	75423683	75427741	-	4059	25

RNAseq quality metrics & coverage: Picardtools, bedtools, RSeQC

- Evaluate the quality of reads and alignments
- Identify potential problems regarding (RNA) sample quality
  - Is the RNA highly degraded?
  - Is there high-level genomic DNA contamination?
  - Was ribosome RNA successfully depleted?
  - How do reads distribute on the genome? (exons, introns, intergenic, etc.)
  - Is the strandness of read alignment consistent with library type? non-stranded or forward/reverse strand-specific libs
  - Does the target gene that was knocked down in KO samples have reduced expression, compared to WT?

a) RNA degradation

- RNA quality: Fresh samples (e.g. cell line) > frozen samples (e.g. mouse tissues) >> FFPE samples (e.g. human tumors)
- During lib prep, RNA quality is inferred by The RNA integrity number (RIN) evaluated using the 28S to 18S rRNA ratio (e.g. RIN > 4.5)
- However, studies have shown that RIN can be quite inaccurate for FFPE samples
- *Gene body coverage plot* (Picardtools). FFPE samples often has 5' degradation



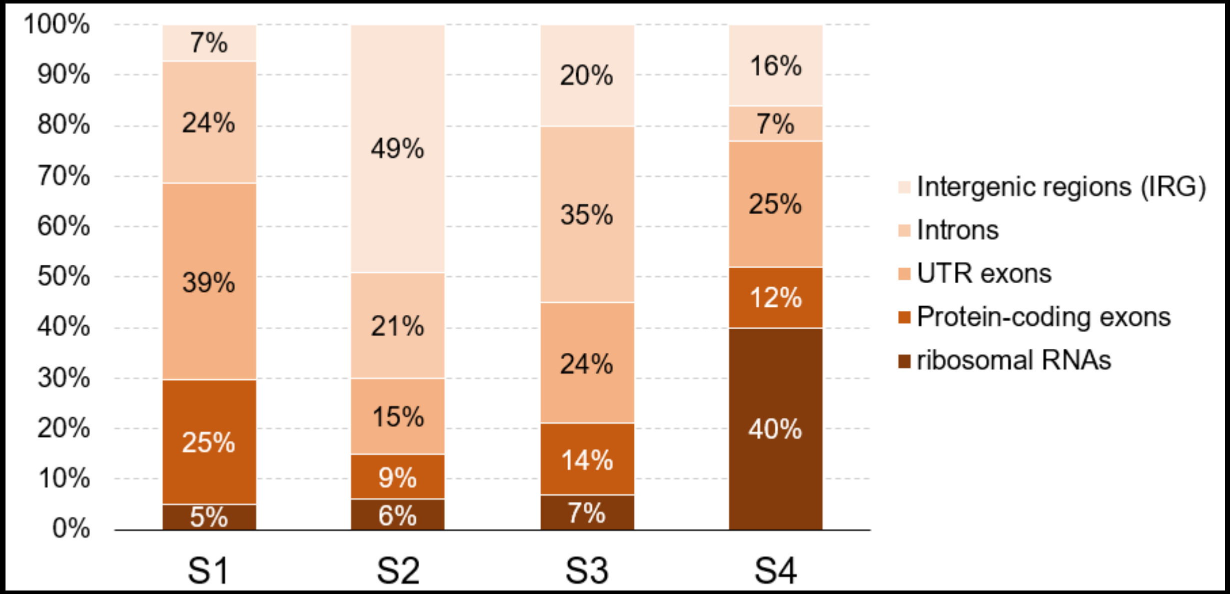
b) Genomic DNA contamination

- In lib prep, DNase digestion removes genomic DNA
- While fresh samples often has good-quality RNAs, highly-degraded samples (e.g. FFPE tumors) often has a higher degree of genomic DNA contamination
- Sometimes DNA contamination could occupy 70% of sequencing reads, greatly reducing the discovery power of DEG analysis
- Good assessment to identify and estimate genomic DNA contamination includes
  - Read distribution in genomic features: high fraction of intergenic reads indicates DNA contamination
  - Visualize intergenic region in genome browser (e.g. IGV)

b) Genomic DNA contamination

Q1: Which sample (S1-4) has the most severe genomic DNA contamination?

Hint: higher percentage of intergenic reads indicates more severe DNA contamination in RNA samples



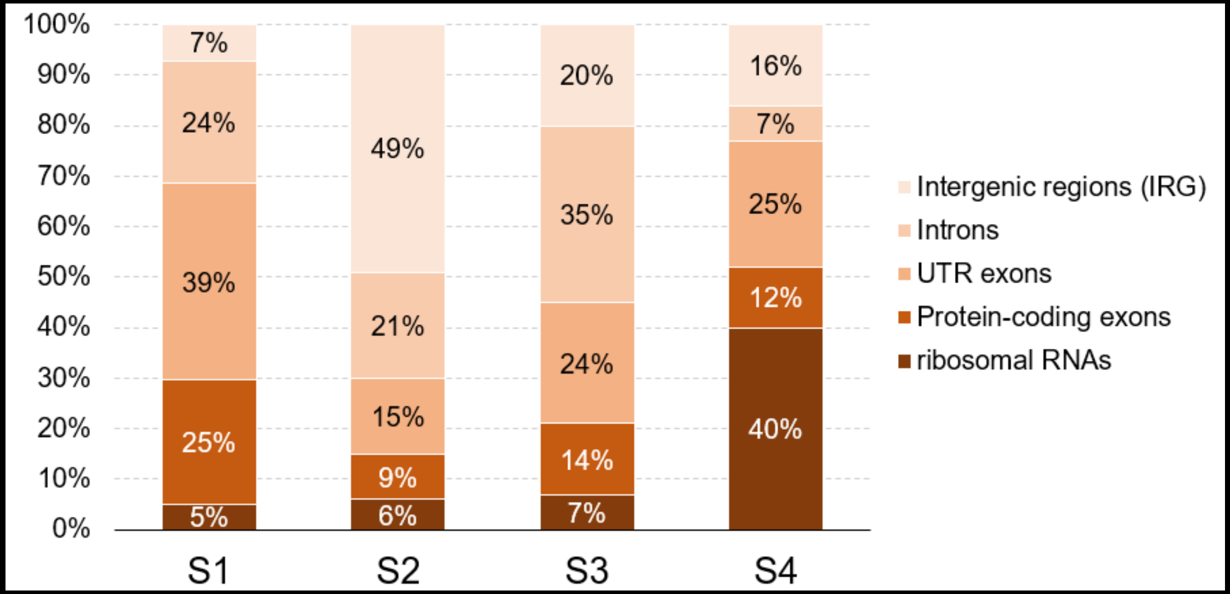
c) Ribosome RNA fraction

- rRNA accounts for > 80% of the transcriptome
- Mmost RNAseq lib prep protocol includes an ribosomal RNA depletion step
- However, if RNA quality is relatively poor (e.g. FFPE tumors), rRNA depletion often is efficient
- Accessing if the depletion step is successful through read distribution in genomic features
- How many reads map to ribosome RNA regions on the genome?

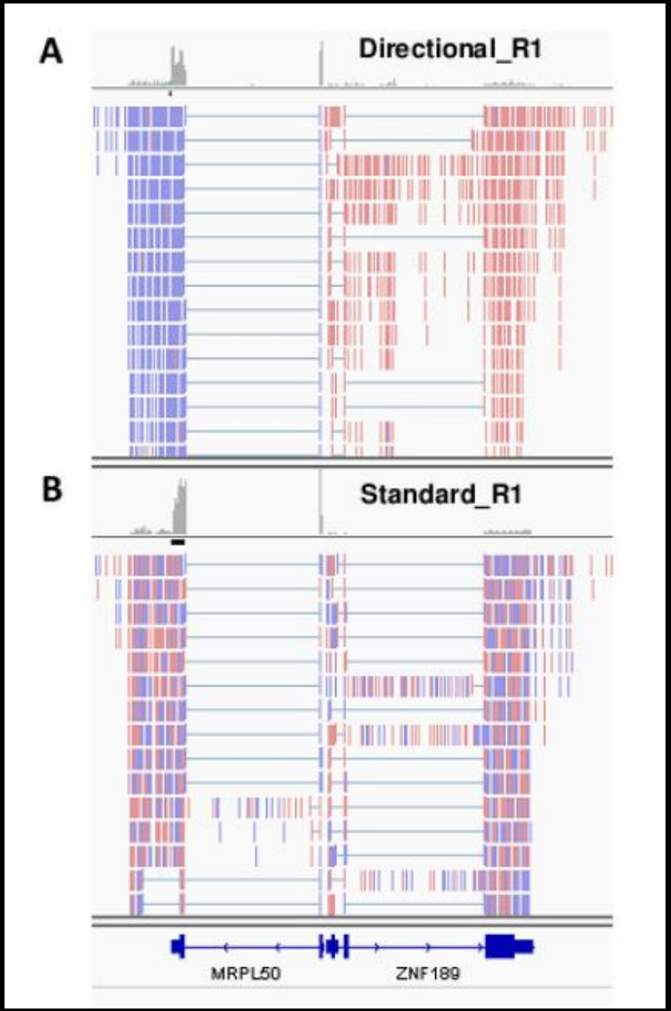
c) Ribosome RNA fraction

Q2: Which sample (S1-4) has the least efficient depletion of ribosome RNAs?

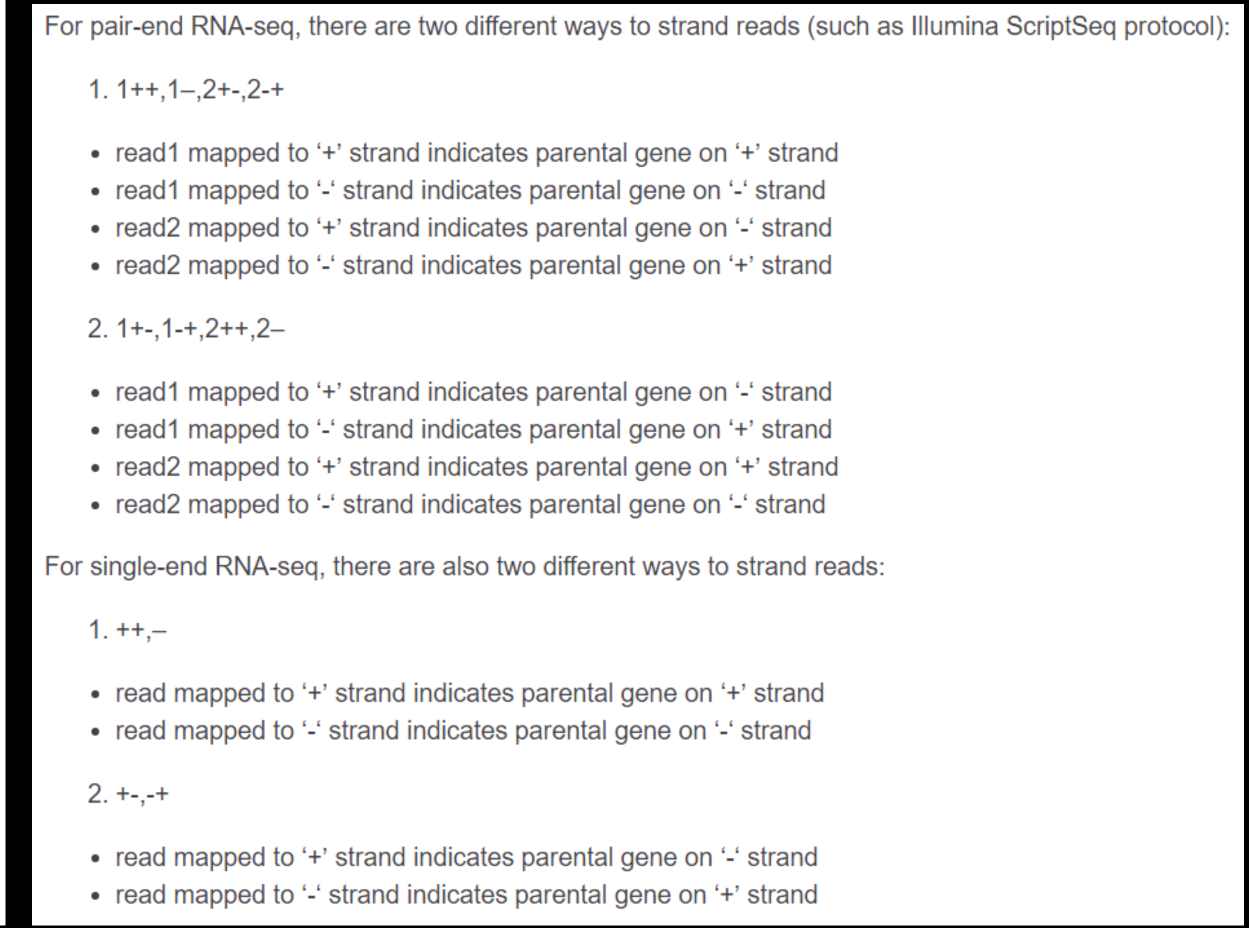
Hint: rRNAs account for > 80% of the whole transcriptome. If not removed, the majority of the sequencing reads will be derived from rRNA



c) Library standness: fraction of correctly oriented reads



RSeQC: infer_experiment.py



For pair-end RNA-seq, there are two different ways to strand reads (such as Illumina ScriptSeq protocol):

1. 1++,1–,2+-,2-+

- read1 mapped to ‘+’ strand indicates parental gene on ‘+’ strand
- read1 mapped to ‘-‘ strand indicates parental gene on ‘-‘ strand
- read2 mapped to ‘+’ strand indicates parental gene on ‘-‘ strand
- read2 mapped to ‘-‘ strand indicates parental gene on ‘+’ strand

2. 1+-,1-+,2++,2–

- read1 mapped to ‘+’ strand indicates parental gene on ‘-‘ strand
- read1 mapped to ‘-‘ strand indicates parental gene on ‘+’ strand
- read2 mapped to ‘+’ strand indicates parental gene on ‘+’ strand
- read2 mapped to ‘-‘ strand indicates parental gene on ‘-‘ strand

For single-end RNA-seq, there are also two different ways to strand reads:

1. ++,–

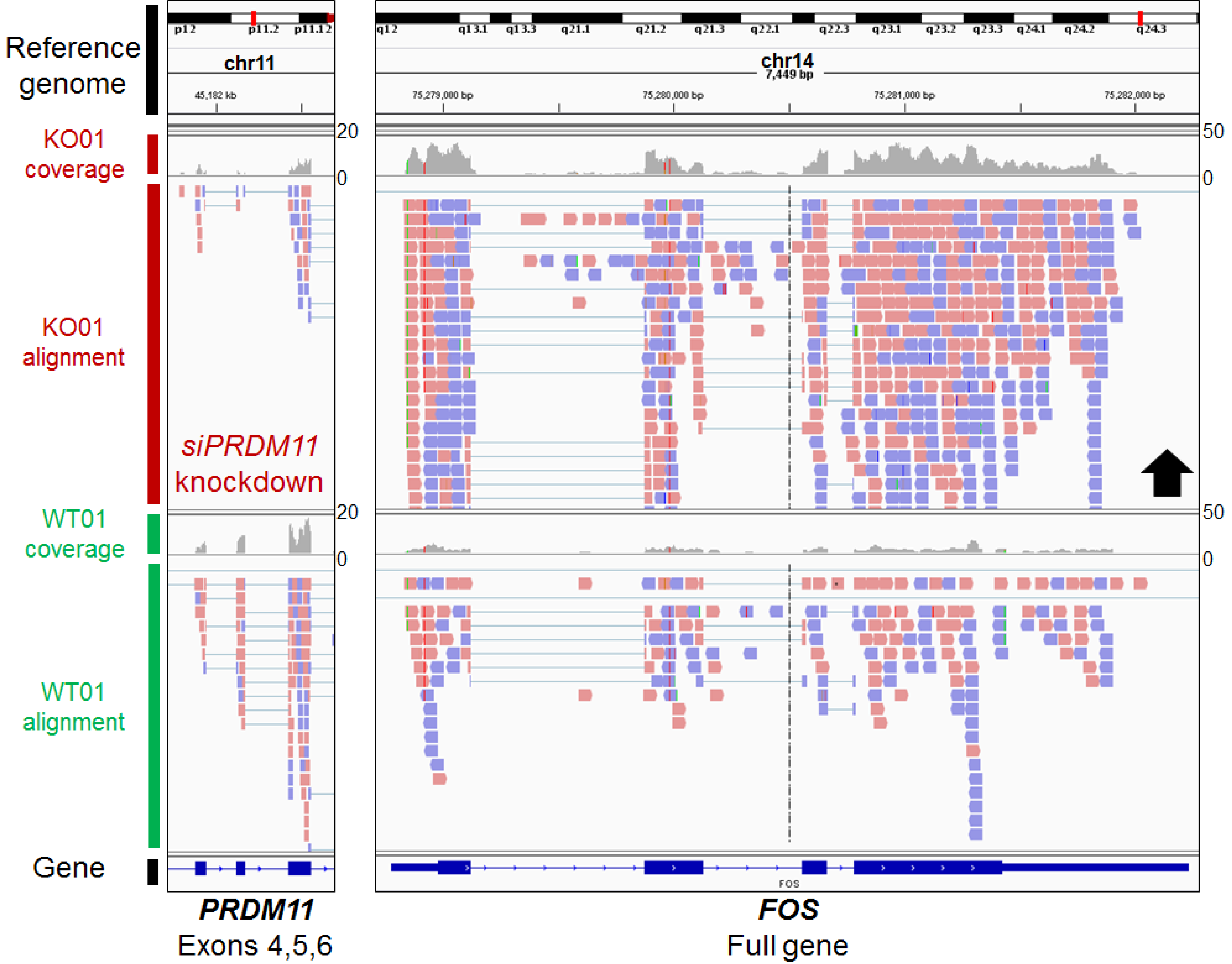
- read mapped to ‘+’ strand indicates parental gene on ‘+’ strand
- read mapped to ‘-‘ strand indicates parental gene on ‘-‘ strand

2. +-,-+

- read mapped to ‘+’ strand indicates parental gene on ‘-‘ strand
- read mapped to ‘-‘ strand indicates parental gene on ‘+’ strand

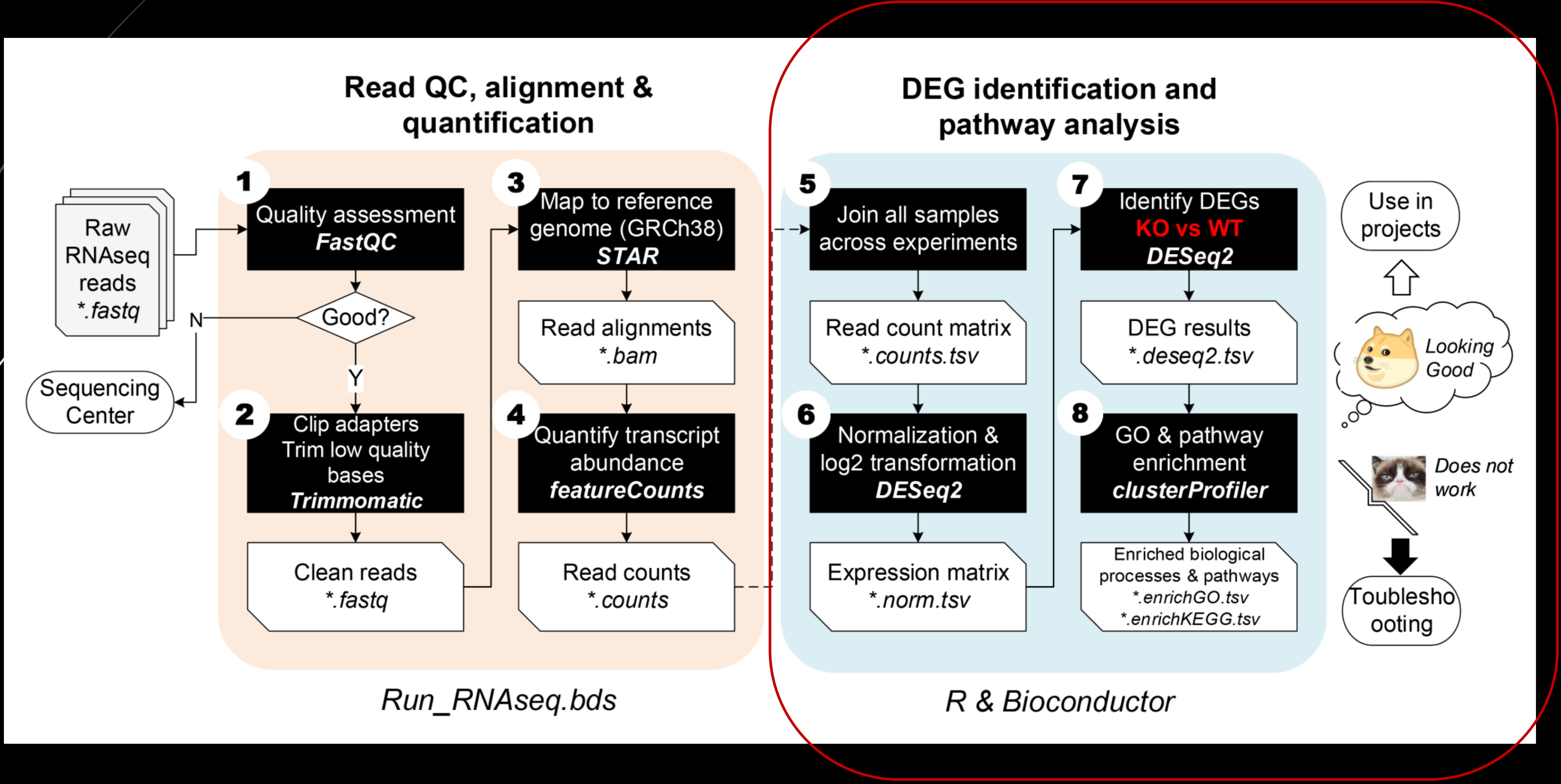
e) Confirmation of reduced/increased expression for knockdown/overexpressed genes in KO/OE samples

- Is there a gene that is expected/known to be repressed or overexpressed?
- Does the RNAseq result reflect expression change of this gene?
- In our data, *PRDM11* expression is expected to be down since this is the gene knocked down in U2932 cells
- NOT all exons are affected!
- Only those targeted by siRNA (small RNA interference) will be affected



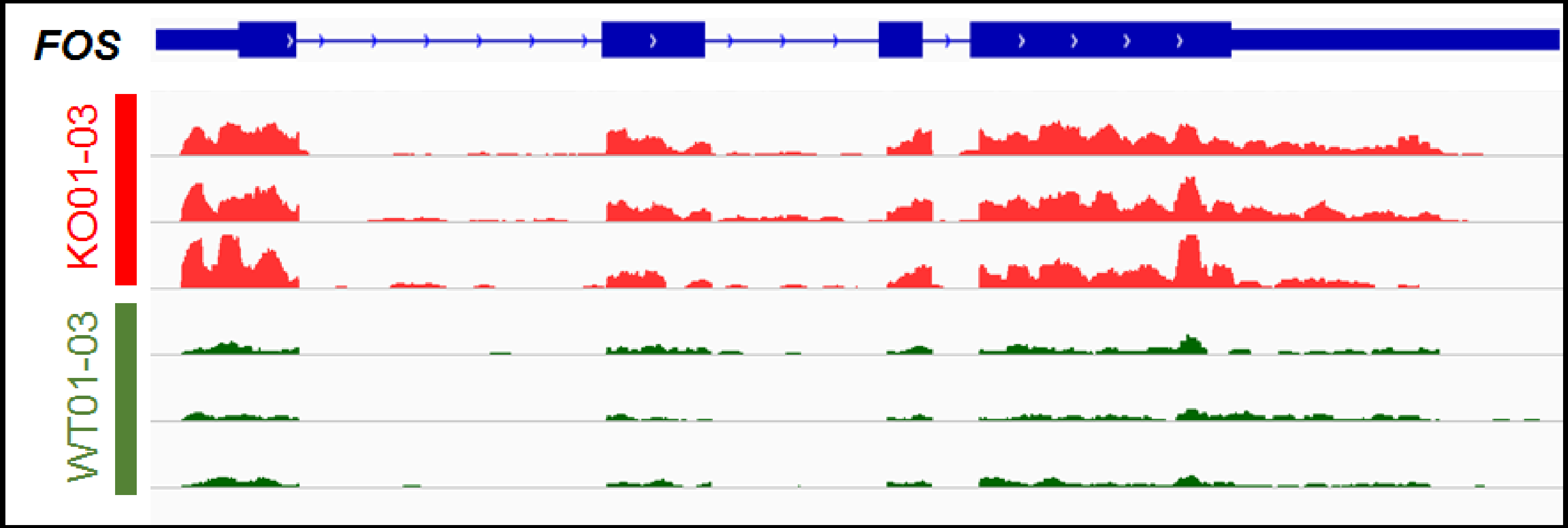
- Our data
  - *PRDM11* knockdown U2932 cells in triplicates (KO01-03 vs WT01-03)
  - NOT the full PRDM11 gene is knockdown!
  - siRNAs target exons 4,5,6 of PRDM11, thus only those three exons show a reduction of expression in the KO samples; other exons are not affected
  - PRDM11 knockdown leads to upregulation of FOS expression

How to perform RNAseq analysis

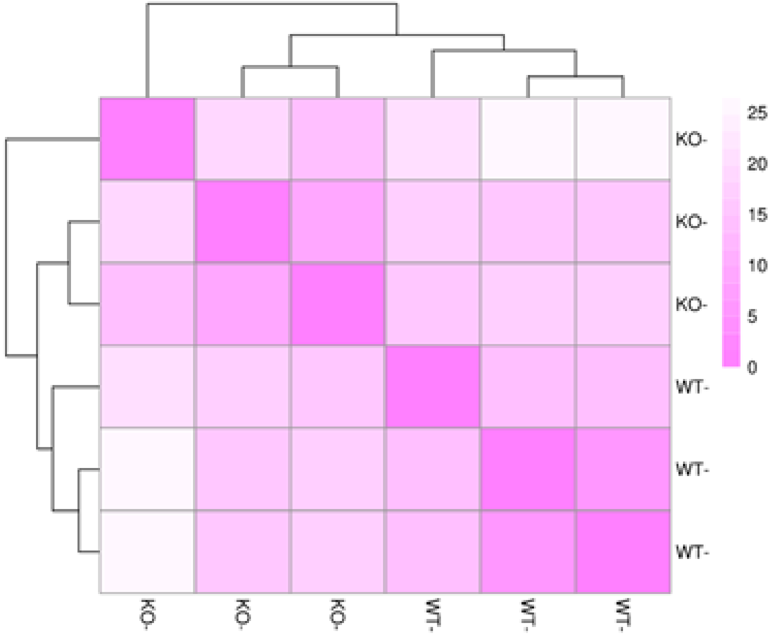

Read QC, alignment & quantification
DEG identification and pathway analysis
Raw RNAseq reads *.fastq
Sequencing Center
1
Quality assessment FastQC
Good?
N
Y
2
Clip adapters Trim low quality bases Trimmomatic
Clean reads *.fastq
3
Map to reference genome (GRCh38) STAR
Read alignments *.bam
4
Quantify transcript abundance featureCounts
Read counts *.counts
5
Join all samples across experiments
Read count matrix *.counts.tsv
6
Normalization & log2 transformation DESeq2
Expression matrix *.norm.tsv
7
Identify DEGs KO vs WT DESeq2
DEG results *.deseq2.tsv
8
GO & pathway enrichment clusterProfiler
Enriched biological processes & pathways *.enrichGO.tsv *.enrichKEGG.tsv
Use in projects
Looking Good
Does not work
Toubleshooting
Run_RNAseq.bds
R & Bioconductor

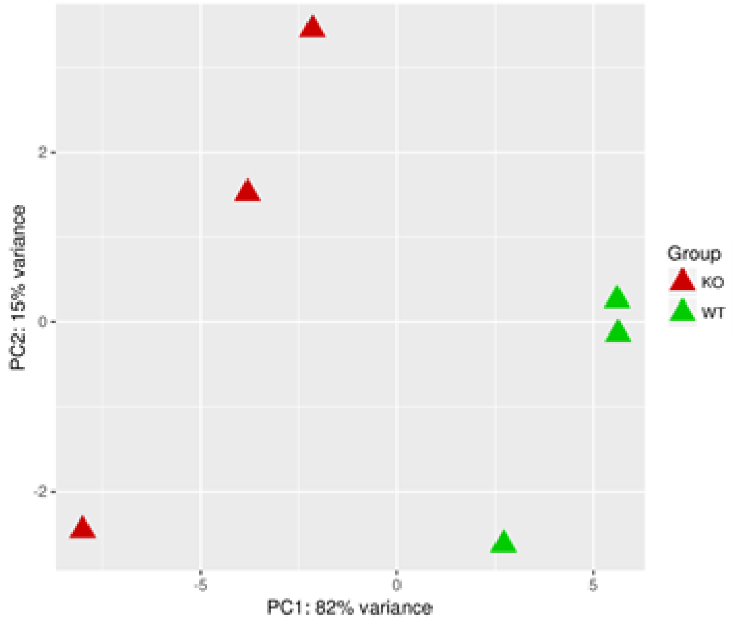
05-08: Identify differentially expressed genes and pathways: DESeq2, clusterProfiler

- After steps 01-04, we have generated read alignment and counts for every annotated gene on the genome
- The next step is to utilize the read counts data to detect DEGs
- For example, if we visualize *FOS* gene across 6 samples in genome browser

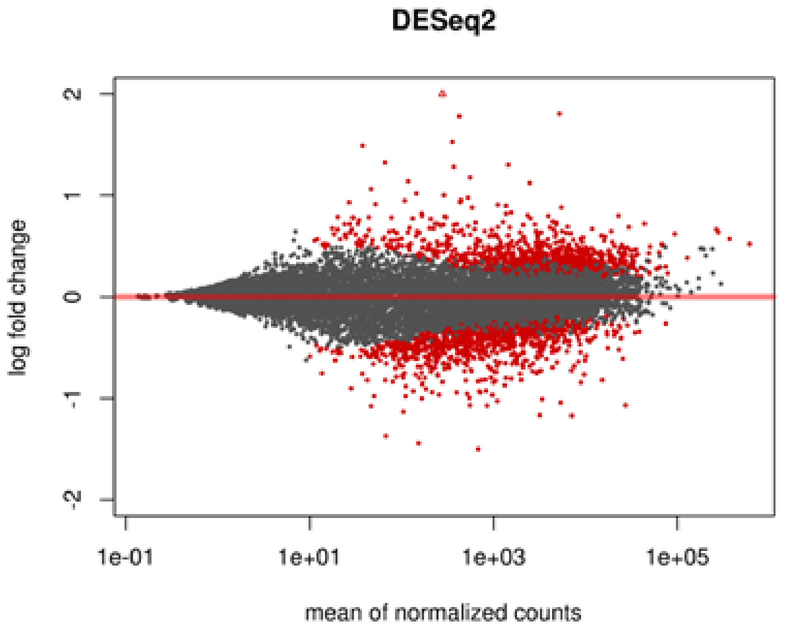


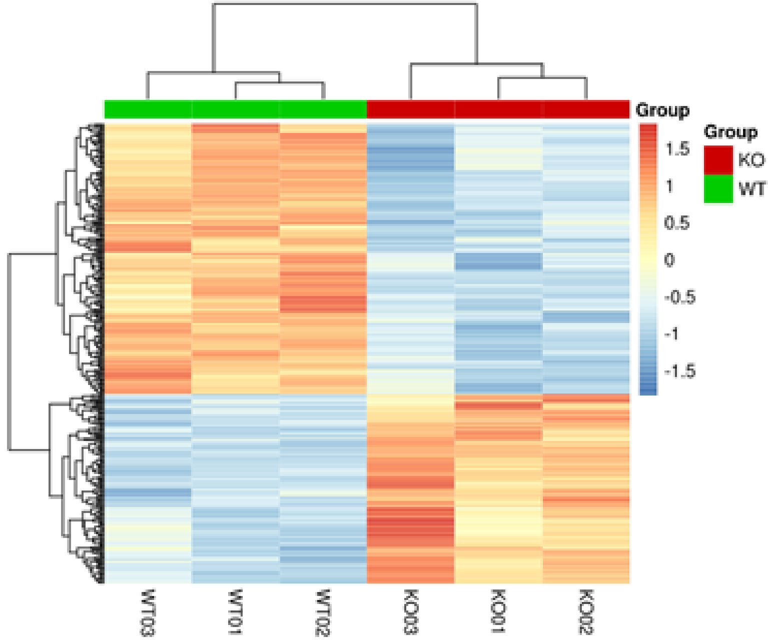
FOS = Fos proto-oncogene, AP-1 transcription factor subunit

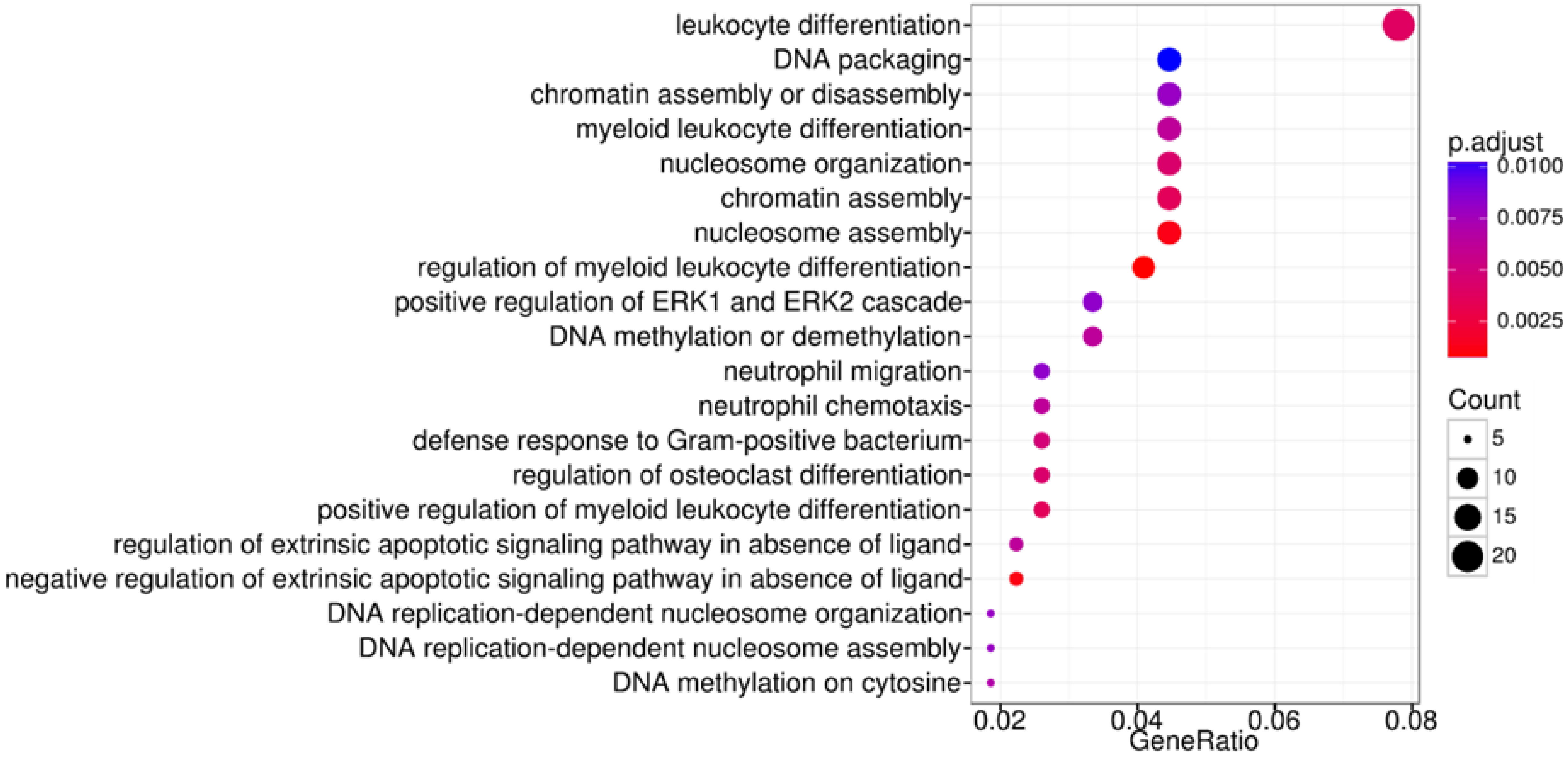


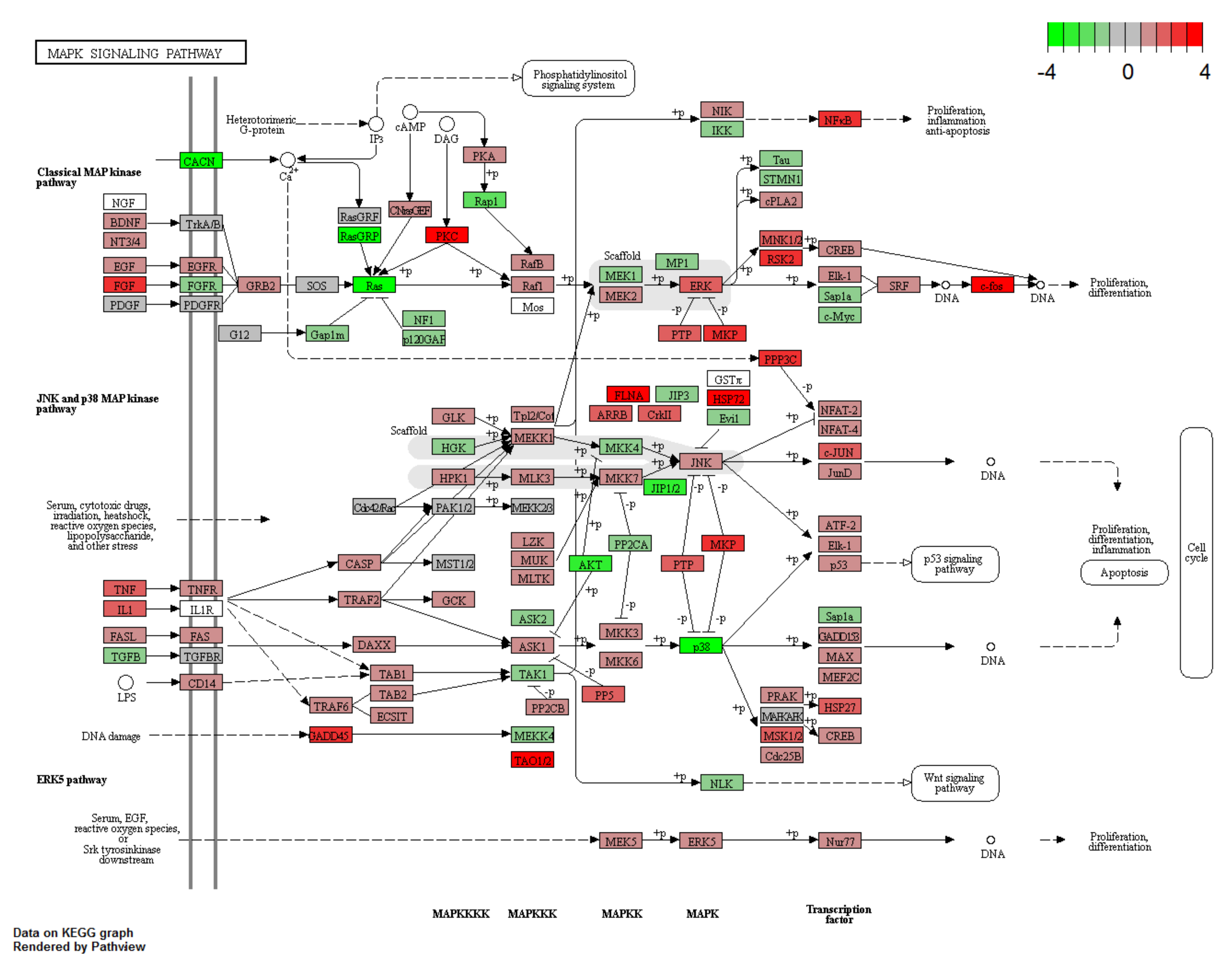


```
[1] "Genes significant = 296 (fc, 1.5, fdr 0.05)"
[1] "Heatmap = 296 genes on the row, 6 samples on the column"
```






leukocyte differentiation
DNA packaging
chromatin assembly or disassembly
myeloid leukocyte differentiation
nucleosome organization
chromatin assembly
nucleosome assembly
regulation of myeloid leukocyte differentiation
positive regulation of ERK1 and ERK2 cascade
DNA methylation or demethylation
neutrophil migration
neutrophil chemotaxis
defense response to Gram-positive bacterium
regulation of osteoclast differentiation
positive regulation of myeloid leukocyte differentiation
regulation of extrinsic apoptotic signaling pathway in absence of ligand
negative regulation of extrinsic apoptotic signaling pathway in absence of ligand
DNA replication-dependent nucleosome organization
DNA replication-dependent nucleosome assembly
DNA methylation on cytosine
0.02
0.04
0.06
0.08
GeneRatio
p.adjust
0.0100
0.0075
0.0050
0.0025
Count
5
10
15
20


MAPK SIGNALING PATHWAY
-4
0
4
Classical MAP kinase pathway
JNK and p38 MAP kinase pathway
ERK5 pathway
Heterotorimeric G-protein
IP3
cAMP
DAG
Ca2+
Phosphatidylinositol signaling system
CACN
NGF
BDNF
NT3/4
EGF
FGF
PDGF
TrkA/B
EGFR
FGFR
PDGFR
GRB2
SOS
Ras
RasGRF
RasGRP
CNrasGEF
PKC
PKA
Rap1
RafB
Raf1
Mos
G12
Gap1m
NF1
p120GAP
Scaffold
MP1
MEK1
MEK2
ERK
PTP
MKP
NIK
IKK
NFκB
Proliferation, inflammation anti-apoptosis
Tau
STMN1
cPLA2
MNK1/2
RSK2
CREB
Elk-1
Sap1a
c-Myc
SRF
c-fos
DNA
Proliferation, differentiation
PPP3C
GSTπ
FLNA
JIP3
HSP72
ARRB
CrkII
Evi1
NFAT-2
NFAT-4
c-JUN
JunD
GLK
HGK
HPK1
Tpl2/Cot
MEKK1
MLK3
MEKK2/3
Cdc42/Rac
PAK1/2
MKK4
MKK7
JNK
JIP1/2
Serum, cytotoxic drugs, irradiation, heatshock, reactive oxygen species, lipopolysaccharide, and other stress
LZK
MUK
MLTK
PP2CA
AKT
MKP
PTP
ATF-2
Elk-1
p53
p53 signaling pathway
Proliferation, differentiation, inflammation
Apoptosis
Cell cycle
TNF
IL1
FASL
TGFB
LPS
TNFR
IL1R
FAS
TGFBR
CD14
CASP
TRAF2
MST1/2
GCK
DAXX
ASK2
ASK1
MKK3
MKK6
p38
TAB1
TAB2
TAK1
TRAF6
ECSIT
PP2CB
PP5
Sap1a
GADD153
MAX
MEF2C
PRAK
MAPKAPK
MSK1/2
Cdc25B
HSP27
CREB
DNA damage
GADD45
MEKK4
TAO1/2
NLK
Wnt signaling pathway
Serum, EGF, reactive oxygen species, or Srk tyrosinkinase downstream
MEK5
ERK5
Nur77
+p
-p
MAPKKKK
MAPKKK
MAPKK
MAPK
Transcription factor
Data on KEGG graph
Rendered by Pathview

Thank you!

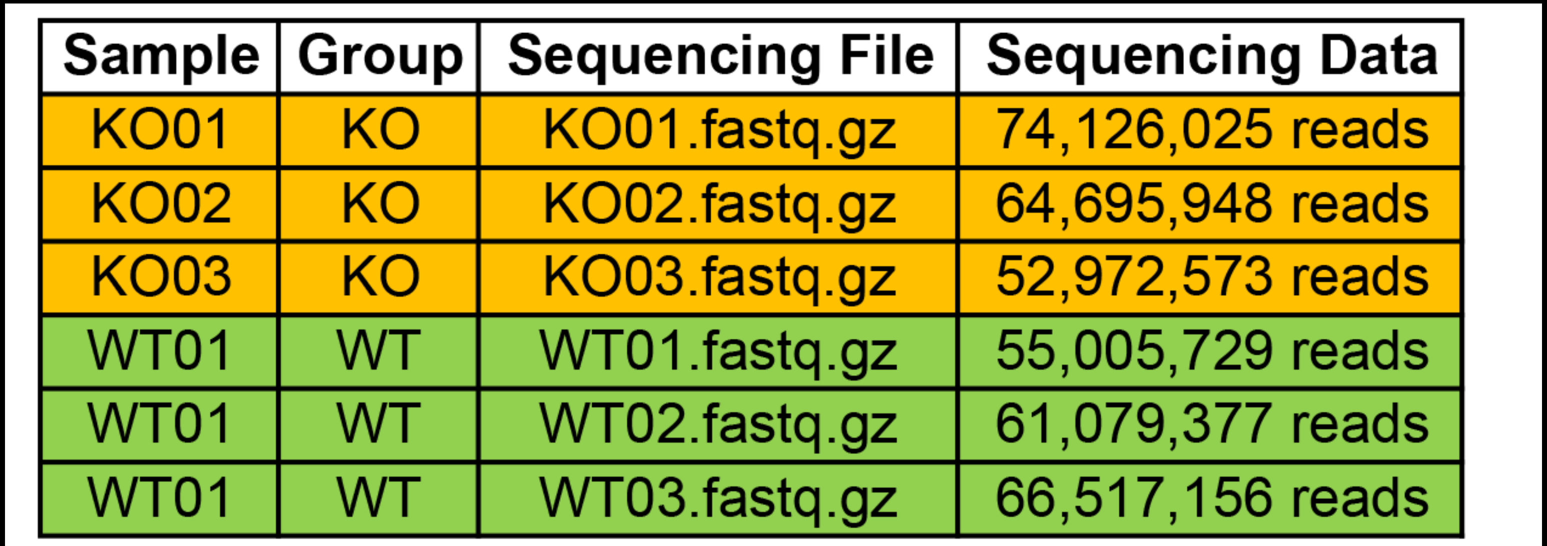
© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

DNA ANALYSIS

Questions

Hands-on practice START

- **Open your handson.Rmd on the Github**
- *Dataset: two groups (PRDM11 KO vs WT, human U2932 cells), 6 samples*
- *Single-end reads, unstranded libraries*

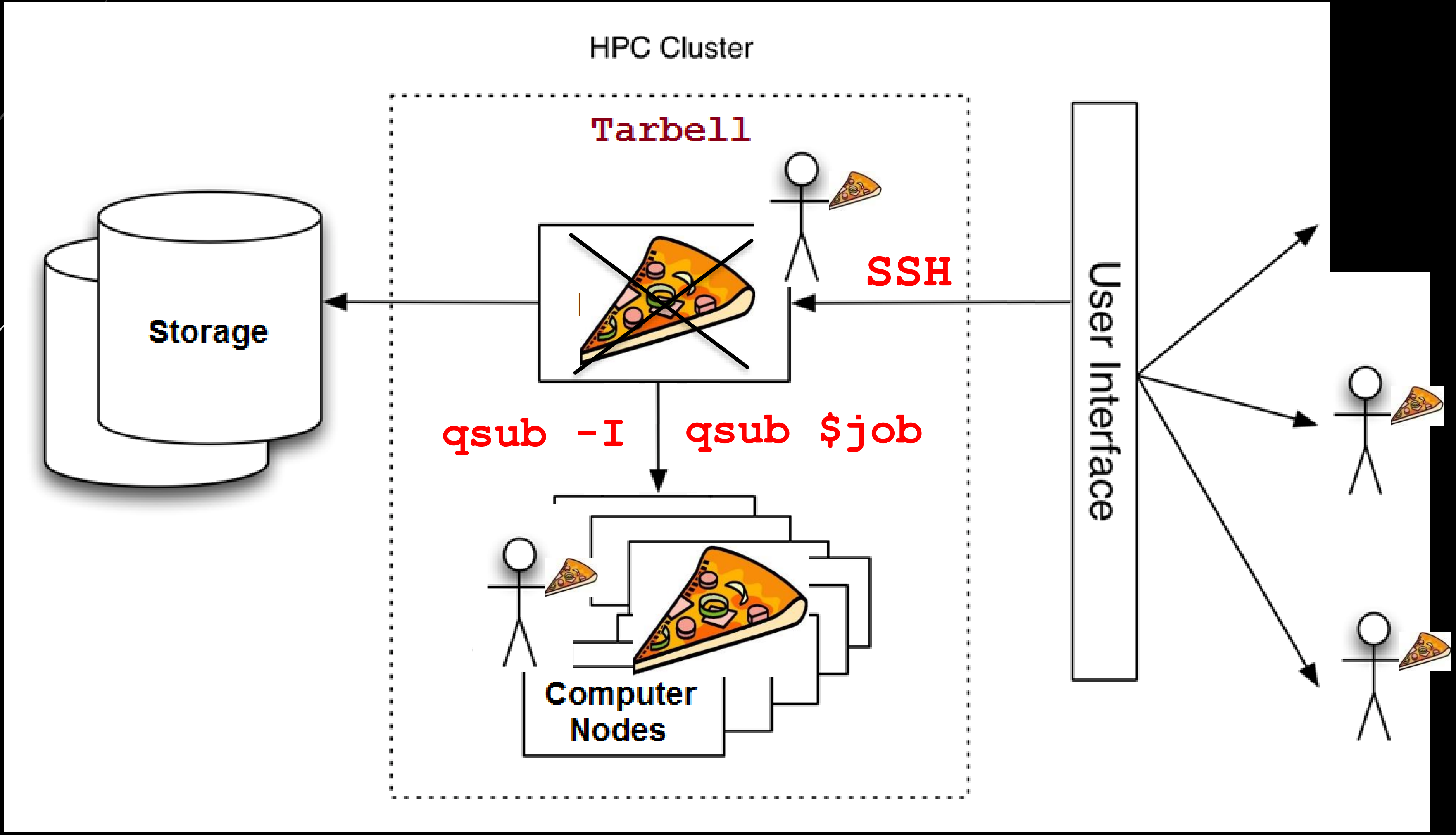


Sample	Group	Sequencing File	Sequencing Data
KO01	KO	KO01.fastq.gz	74,126,025 reads
KO02	KO	KO02.fastq.gz	64,695,948 reads
KO03	KO	KO03.fastq.gz	52,972,573 reads
WT01	WT	WT01.fastq.gz	55,005,729 reads
WT01	WT	WT02.fastq.gz	61,079,377 reads
WT01	WT	WT03.fastq.gz	66,517,156 reads

Fog. et al. 2015. Loss of *PRDM11* promotes MYC-driven lymphomagenesis. Blood 125(8):1272-81

PRDM11 = PR/SET domain 11

Job running is prohibited on the head node!



Ten essential Linux commands

cd	Change directory
ls	List contents
cp	Copy
mv	Move/rename
rm	Delete
pwd	Print the current path
head	Show the first few lines of a file
more	View a file by page
mkdir	Create a new directory
man	Show help menu for a command

Five useful HPC/Shell commands

qsub	Submit a job
qstat	List submitted jobs from the user
qdel	Delete a job
showq	Show all jobs on the cluster
ssh	Log into a server using Secure Shell (SSH)

INDEX
EXTRA QUALITY
90