

# Bulk RNA-seq analysis with the Cancer Genomics Cloud, powered by Seven Bridges



Phillip Webster, Genomics Scientist  
March 22, 2022



## Class Overview

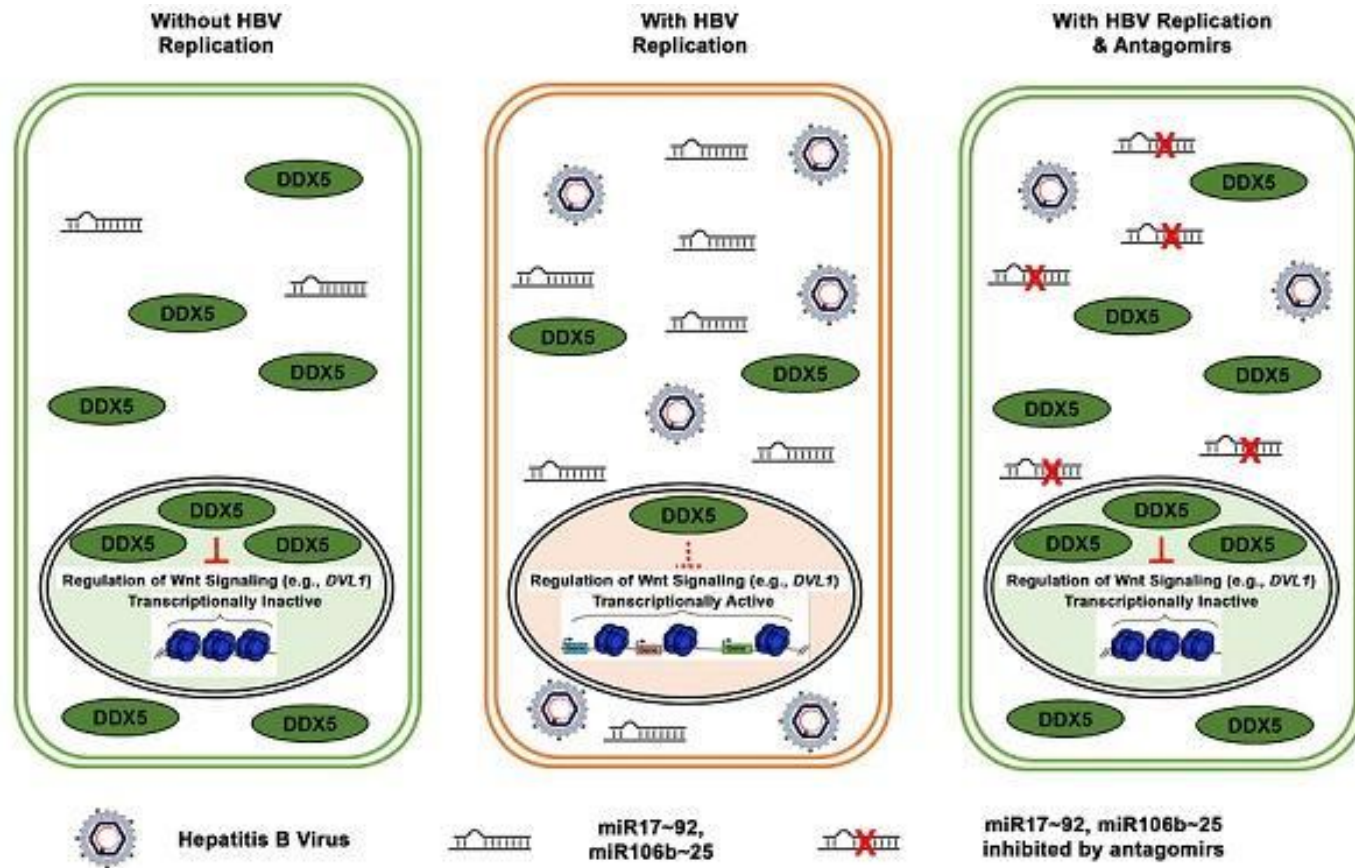
- Lecture 1 - Overview  
Hands-on demo: your first project on the CGC
- Lecture 2 - Bulk RNA-seq workflow(s) and sequence alignment  
Hands-on demo: your CGC workflow
- **Lecture 3 - Differential expression and visualizations**  
Hands-on demo: your CGC visualization
- Lecture 4 - Single-cell RNA-seq overview and workflow  
Hands-on demo: Seurat



# Lecture Overview

- Interactive/downstream interactive QC and data analysis for bulk RNAseq
- Post-alignment QC considerations
- Comparing samples, making visualizations
- Interpret results
- Find your favorite gene

# Overview of data we will use for bulk RNA-seq analysis

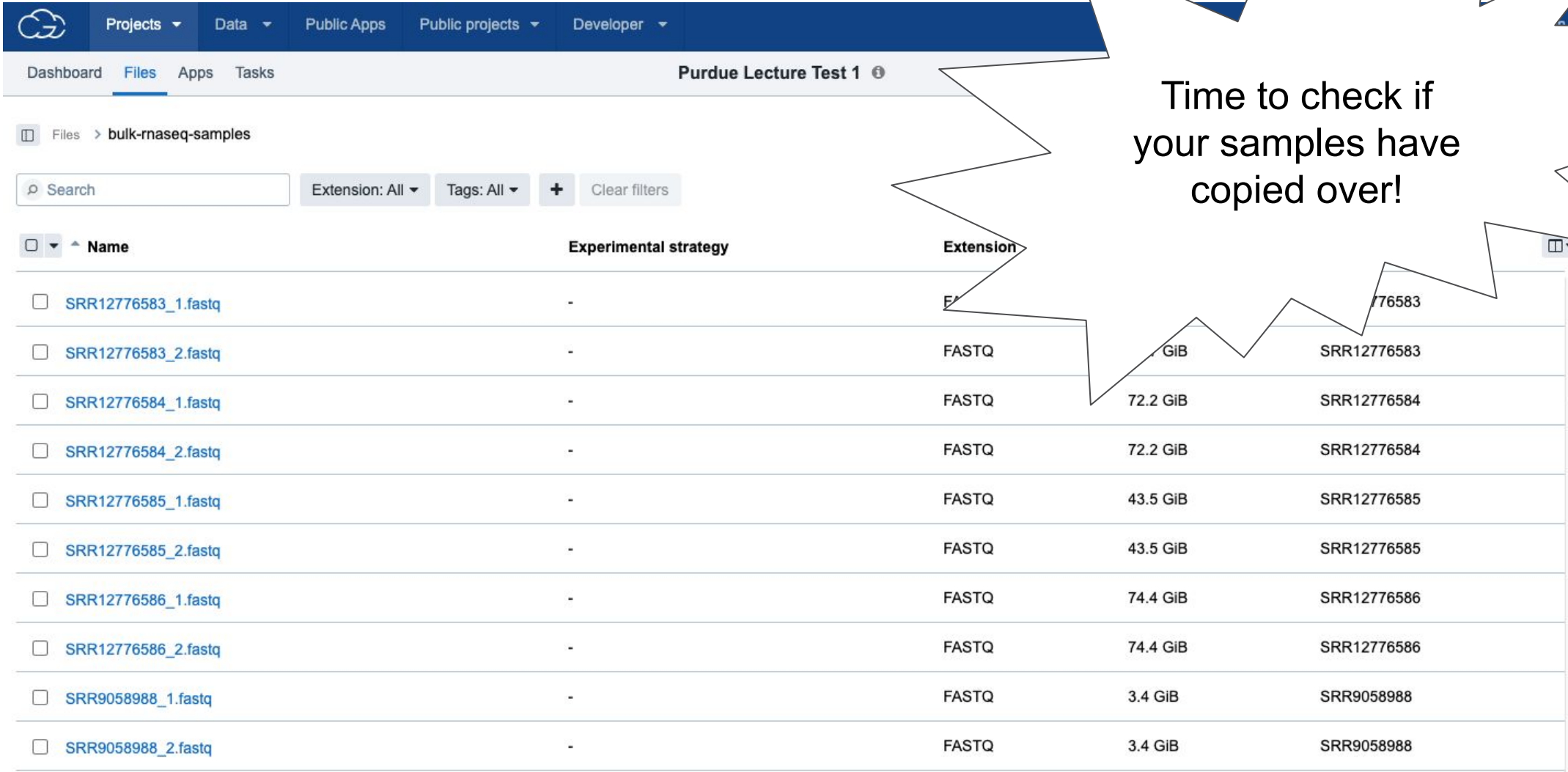


Mani SKK, et al. Restoration of RNA helicase DDX5 suppresses hepatitis B virus (HBV) biosynthesis and Wnt signaling in HBV-related hepatocellular carcinoma. *Theranostics* 2020; 10(24):10957-10972. doi:10.7150/thno.49629. <https://www.thno.org/v10p10957.htm>

## Data availability

All sequencing data are available through the NCBI Gene Expression Omnibus (GEO) database (accession number **GSE131257**).

# RNaseq data transferred from SRA to CGC



Projects ▾ Data ▾ Public Apps Public projects ▾ Developer ▾

Dashboard Files Apps Tasks **Purdue Lecture Test 1** ⓘ

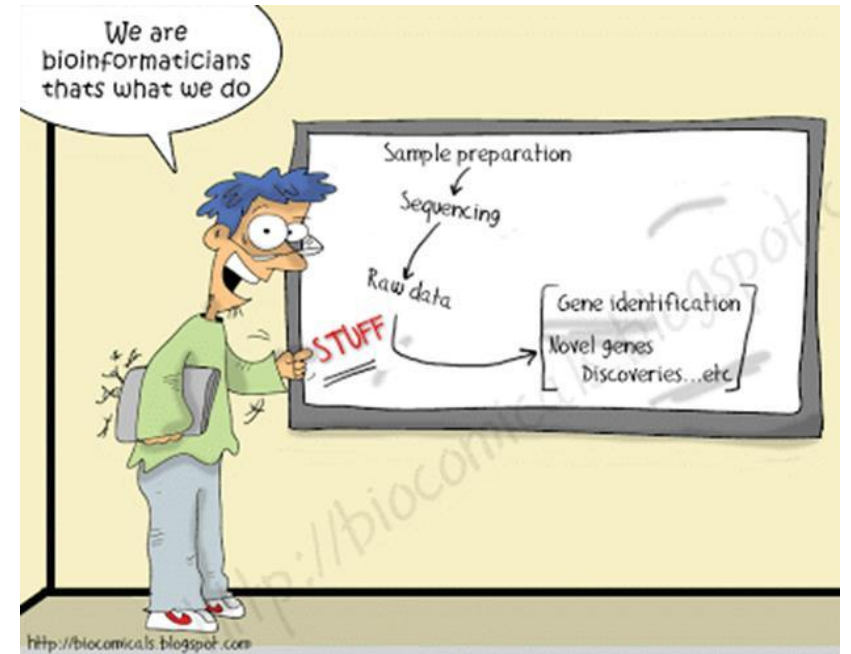
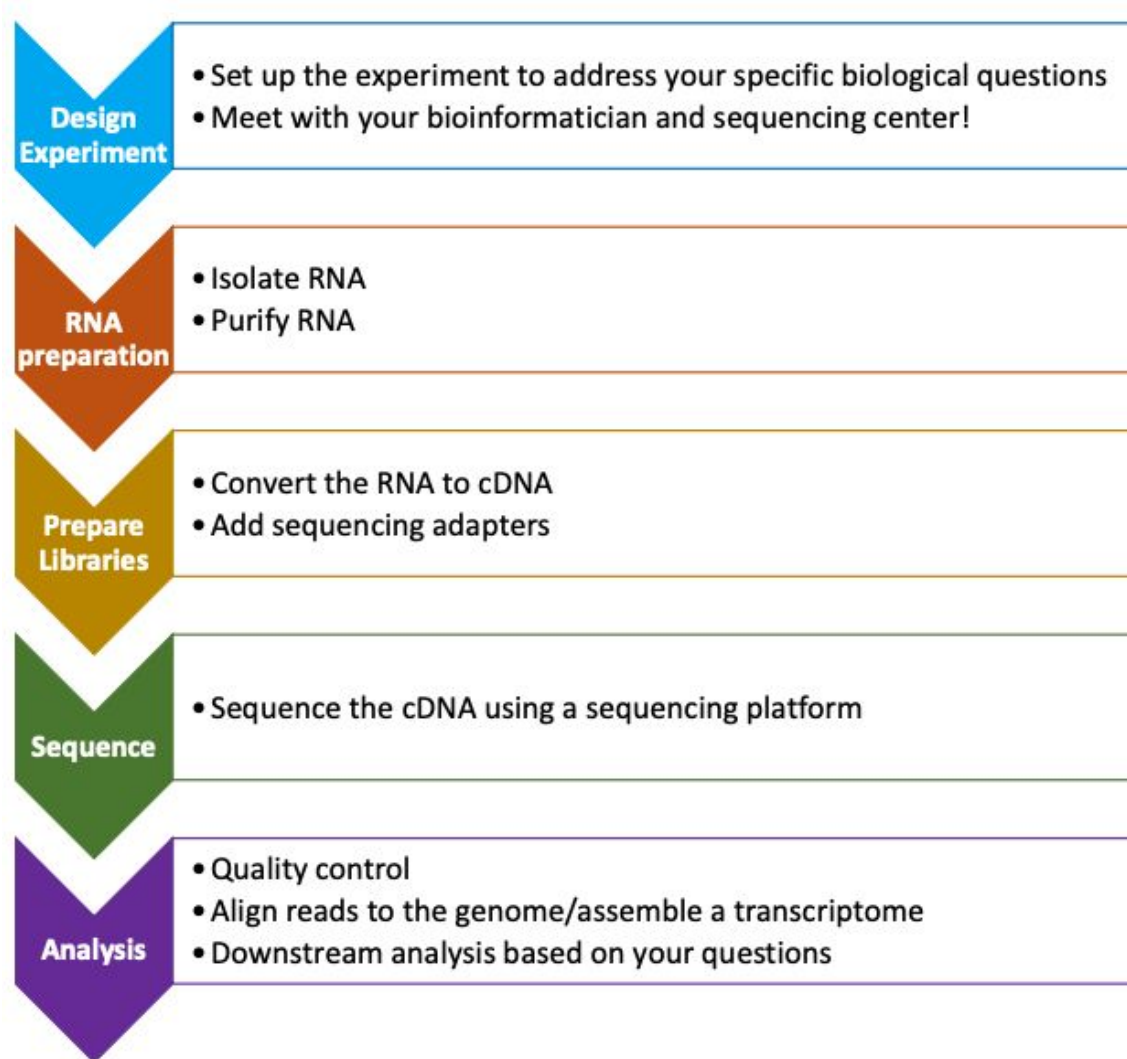
Files > bulk-rnaseq-samples

Search Extension: All ▾ Tags: All ▾ + Clear filters

<input type="checkbox"/> ▾ ^ Name	Experimental strategy	Extension		
<input type="checkbox"/> SRR12776583_1.fastq	-	FASTQ	72.2 GiB	SRR12776583
<input type="checkbox"/> SRR12776583_2.fastq	-	FASTQ	72.2 GiB	SRR12776583
<input type="checkbox"/> SRR12776584_1.fastq	-	FASTQ	72.2 GiB	SRR12776584
<input type="checkbox"/> SRR12776584_2.fastq	-	FASTQ	72.2 GiB	SRR12776584
<input type="checkbox"/> SRR12776585_1.fastq	-	FASTQ	43.5 GiB	SRR12776585
<input type="checkbox"/> SRR12776585_2.fastq	-	FASTQ	43.5 GiB	SRR12776585
<input type="checkbox"/> SRR12776586_1.fastq	-	FASTQ	74.4 GiB	SRR12776586
<input type="checkbox"/> SRR12776586_2.fastq	-	FASTQ	74.4 GiB	SRR12776586
<input type="checkbox"/> SRR9058988_1.fastq	-	FASTQ	3.4 GiB	SRR9058988
<input type="checkbox"/> SRR9058988_2.fastq	-	FASTQ	3.4 GiB	SRR9058988

Time to check if  
your samples have  
copied over!

# RNA-seq workflow overview



We are here!

Slide adapted from Dr. Nadia Atallah, Purdue University

# RNA-seq User Flow



DRAFT **rnaseq\_test run - 03-01-22 17:08:41 - Genotype**

Get support Discard Run

Last update by phil\_webster on Mar. 4, 2022 12:46

App: maseq\_test - Revision: 18

Task Inputs Execution Settings

## Inputs

Batching Off

FASTQ read files \* Change selection

- SRR9058988\_1.fastq
- SRR9058988\_2.fastq
- SRR9058990\_1.fastq
- SRR9058990\_2.fastq
- SRR9058993\_1.fastq

...and 25 more items

GTF annotation \* Change selection

- GRCh38ERCC.ensembl95.gtf

Genome FASTA Select file(s)

No files selected

Transcript FASTA or Salmon Index \* Change selection

- GRCh38ERCC.ensembl95.transcriptome.gentrome.salmon-1.2.0-index-arch...

## App Settings

Edit parameters Show editable

DESeq2 (#deseq2\_1\_26\_0)

Covariate of interest \*

Genotype

Factor level - reference

WT

Factor level - test

KD

## Output Settings

DESeq2 analysis results.	No value
Expression matrix genes	No value
Expression matrix transcripts	No value
Gene-level quantification	No value
HTML report	No value
HTML reports	No value
Normalized counts	No value
RData file	No value
Report zip	No value
Salmon Quant archive	No value
Salmon quant log	No value
Transcript-level quantification	No value
pheno_out	No value

# Okay, we have the output files. What is next?

**COMPLETED Differential Expression - Salmon + DESeq2 run - 03-09-22 20:07:51** [Get support](#) [View stats & logs](#) [Edit and rerun](#)

Executed on Mar. 9, 2022 15:21 by phil\_webster

Spot Instances: On | Memoization (WorkReuse): On | Price: \$0.81 | Duration: 1 hour, 7 minutes

App: Differential Expression - Salmon + DESeq2 - Revision: 1

### Inputs

- FASTQ read files**
  - SRR9059002\_2.fastq
  - SRR9059002\_1.fastq
  - SRR9059001\_2.fastq
  - SRR9059001\_1.fastq
  - SRR9059000\_2.fastq
  - ...and 25 more items
- GTF annotation**
  - GRCh38ERCC.ensembl95.gtf
- Genome FASTA**
  - No files selected
- Transcript FASTA or Salmon Index**
  - GRCh38ERCC.ensembl95.transcriptome.gentrome.salmon-1.2.0-index-archi...

### App Settings

DESeq2 (#deseq2\_1\_26\_0)

- Covariate of interest
- Factor level - reference
- Factor level - test

Genotype: WT, KD

### Output Settings

- DESeq2 analysis results**
  - SRR905.DEAnalysis.out.csv
- Expression matrix genes**
  - expression.matrix.gene.numreads.tsv
- Expression matrix transcripts**
  - expression.matrix.tx.numreads.tsv
- Gene-level quantification**
  - SRR9059002.salmon\_quant.genes.sf
  - SRR9059001.salmon\_quant.genes.sf
  - SRR9059000.salmon\_quant.genes.sf
  - SRR9058999.salmon\_quant.genes.sf
  - SRR9058998.salmon\_quant.genes.sf
  - ...and 10 more items
- HTML report**
  - SRR905.DEAnalysis.deseq2.1.26.0.summary\_report.b64html
- HTML reports**
  - SRR9058988\_1\_fastqc.html

# Outputs of the RNA-seq User Flow

The screenshot displays the 'Output Settings' section of a web application. At the top right, there is a user profile 'zworman' and a notification bell. Below this are navigation tabs for 'Interactive Analysis', 'Settings', and 'Notes'. A row of action buttons includes 'Get support', 'View stats & logs', and 'Edit and rerun'. A 'Show non-default' dropdown is visible on the left. The main content area is titled 'Output Settings' and lists several categories of outputs, each with a dropdown arrow, a help icon, and a share icon. The 'DESeq2 analysis results' category shows a link to 'SRR905.DEAnalysis.out.csv'. The 'Expression matrix genes' category shows a link to 'expression.matrix.gene.numreads.tsv'. The 'Expression matrix transcripts' category shows a link to 'expression.matrix.tx.numreads.tsv'. The 'Gene-level quantification' category lists five salmon quantification files: 'SRR9059002.salmon\_quant.genes.sf', 'SRR9059001.salmon\_quant.genes.sf', 'SRR9059000.salmon\_quant.genes.sf', 'SRR9058999.salmon\_quant.genes.sf', and 'SRR9058998.salmon\_quant.genes.sf', followed by '...and 10 more items'. The 'HTML report' category shows a link to 'SRR905.DEAnalysis.deseq2.1.26.0.summary\_report.b64html', which is highlighted by a red arrow. The 'HTML reports' category shows a link to 'SRR9058988\_1\_factor.html'. A blue question mark icon is located at the bottom right of the settings area.

zworman

Interactive Analysis Settings Notes

Get support View stats & logs Edit and rerun

Show non-default

**Output Settings**

- ▼ DESeq2 analysis results
  - Genotype [SRR905.DEAnalysis.out.csv](#)
- ▼ Expression matrix genes
  - WT [expression.matrix.gene.numreads.tsv](#)
  - KD [expression.matrix.gene.numreads.tsv](#)
- ▼ Expression matrix transcripts
  - [expression.matrix.tx.numreads.tsv](#)
- ▼ Gene-level quantification
  - [SRR9059002.salmon\\_quant.genes.sf](#)
  - [SRR9059001.salmon\\_quant.genes.sf](#)
  - [SRR9059000.salmon\\_quant.genes.sf](#)
  - [SRR9058999.salmon\\_quant.genes.sf](#)
  - [SRR9058998.salmon\\_quant.genes.sf](#)
  - ...and 10 more items
- ▼ HTML report
  - [SRR905.DEAnalysis.deseq2.1.26.0.summary\\_report.b64html](#)
- ▼ HTML reports
  - [SRR9058988\\_1\\_factor.html](#)

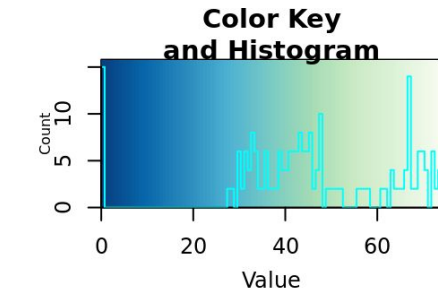


# Sample Distances

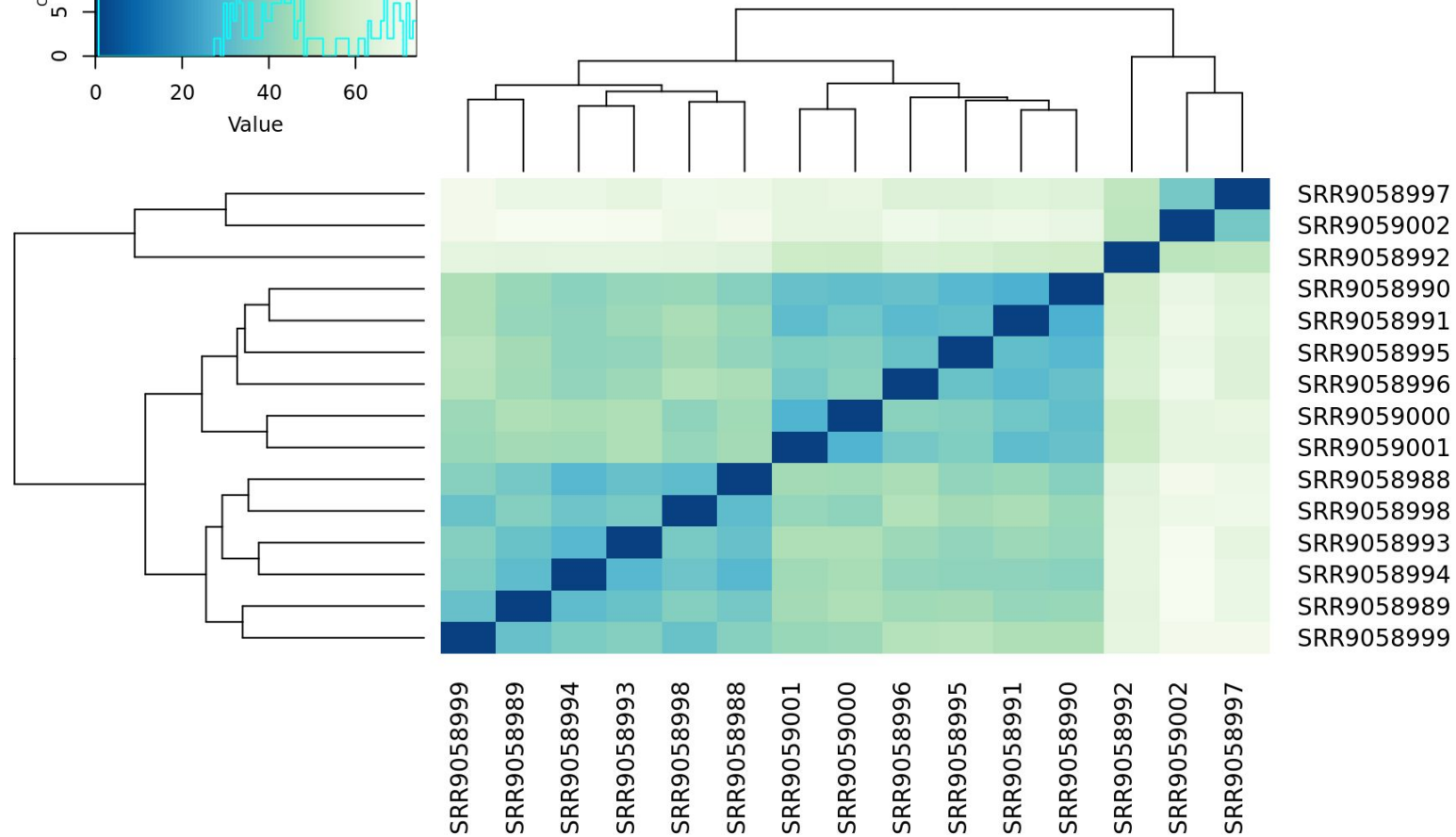
**A useful first step in an RNA-seq analysis is often to assess overall similarity between samples: Which samples are similar to each other, which are different? Does this fit to the expectation from the experiment's design?**

**Two methods to estimate distance**

# Sample Distance Heat Map

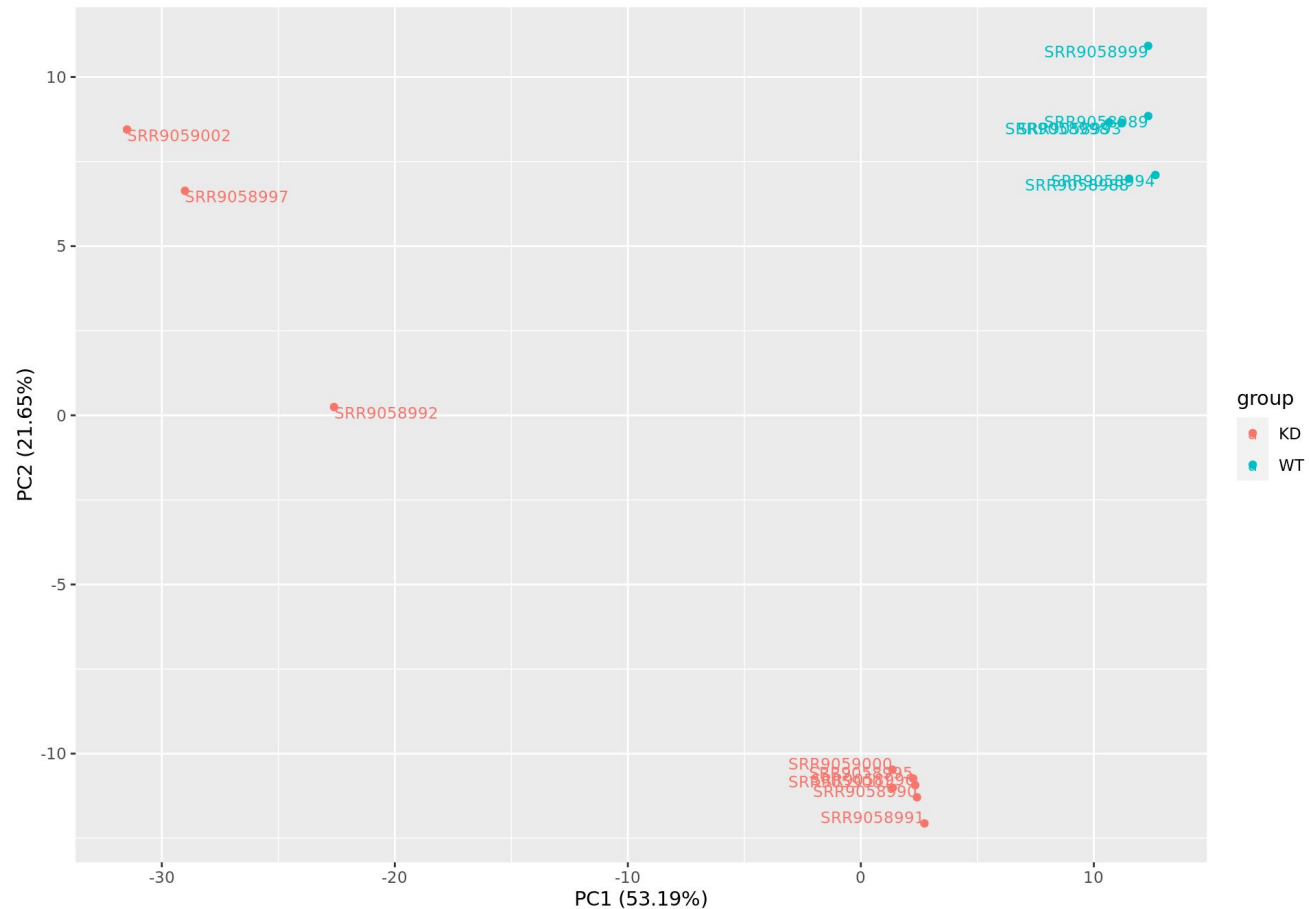


## Sample-to-sample distances



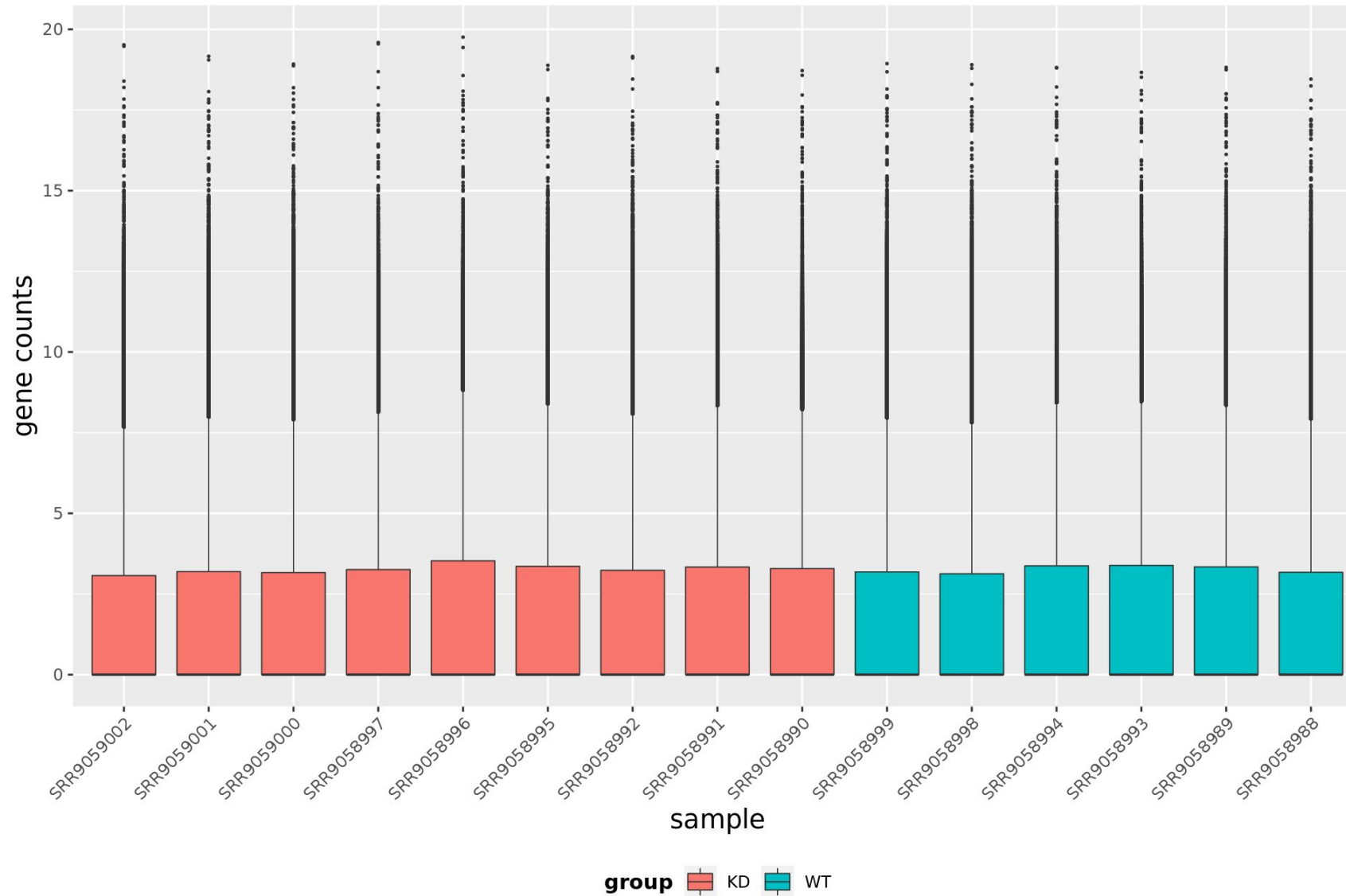
# What is a PCA and what can I use it for?

- Principal component analysis (PCA) is a statistical procedure that can be used for exploratory data analysis. PCA uses linear combinations of the original data (e.g. gene expression values) to define a new set of unrelated variables (principal components). T
- Thus, PCA can be used to reduce the dimensions of a data set, allowing the description of data sets and their variance with a reduced number of variables.
- PCA can also be used to identify outliers with respect to the principal components.



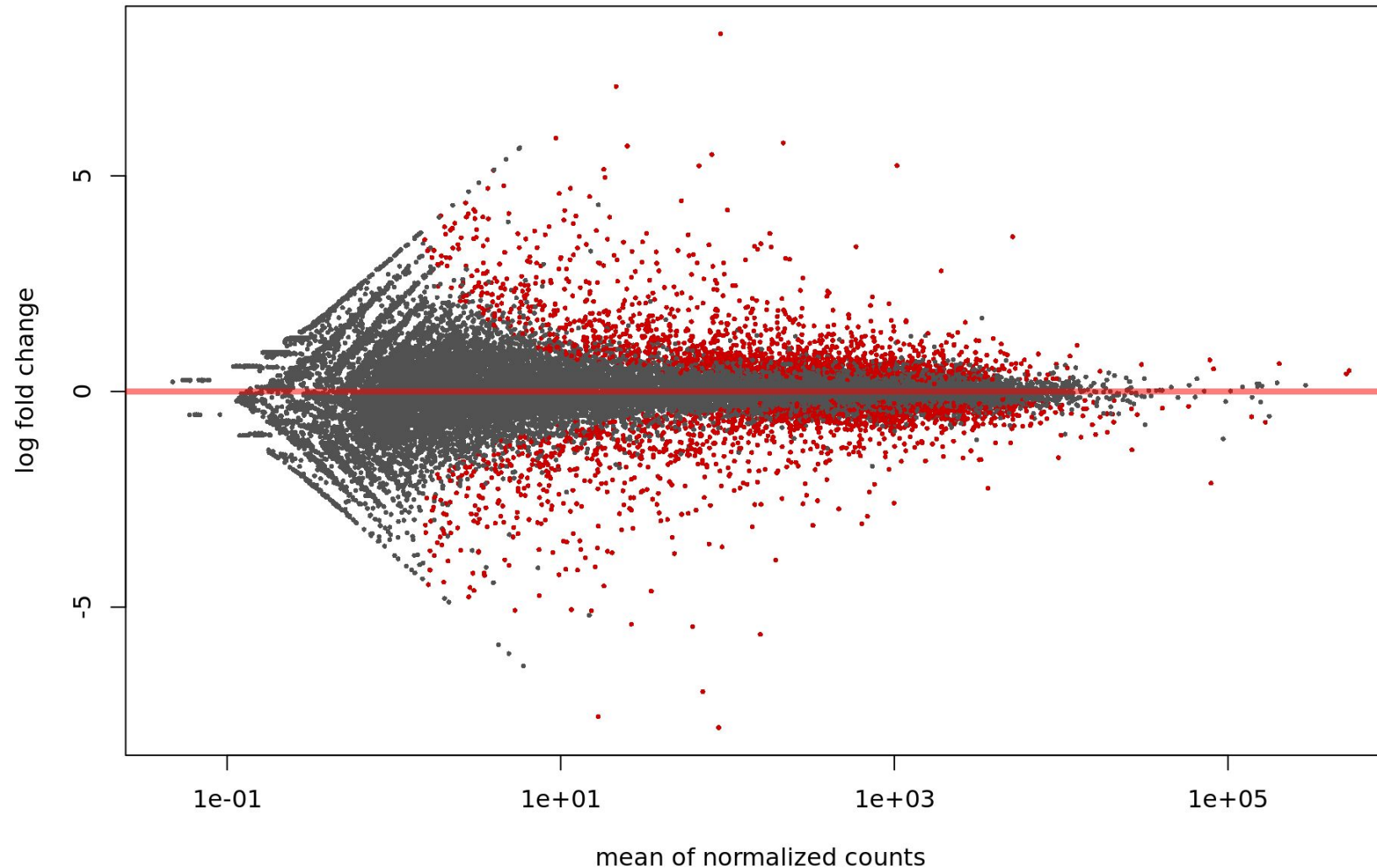
# Gene expression normalized counts

log2 transformed normalized counts



# MA plot: what is it and what can I use it for?

MA-plot for Genotype: KD vs WT

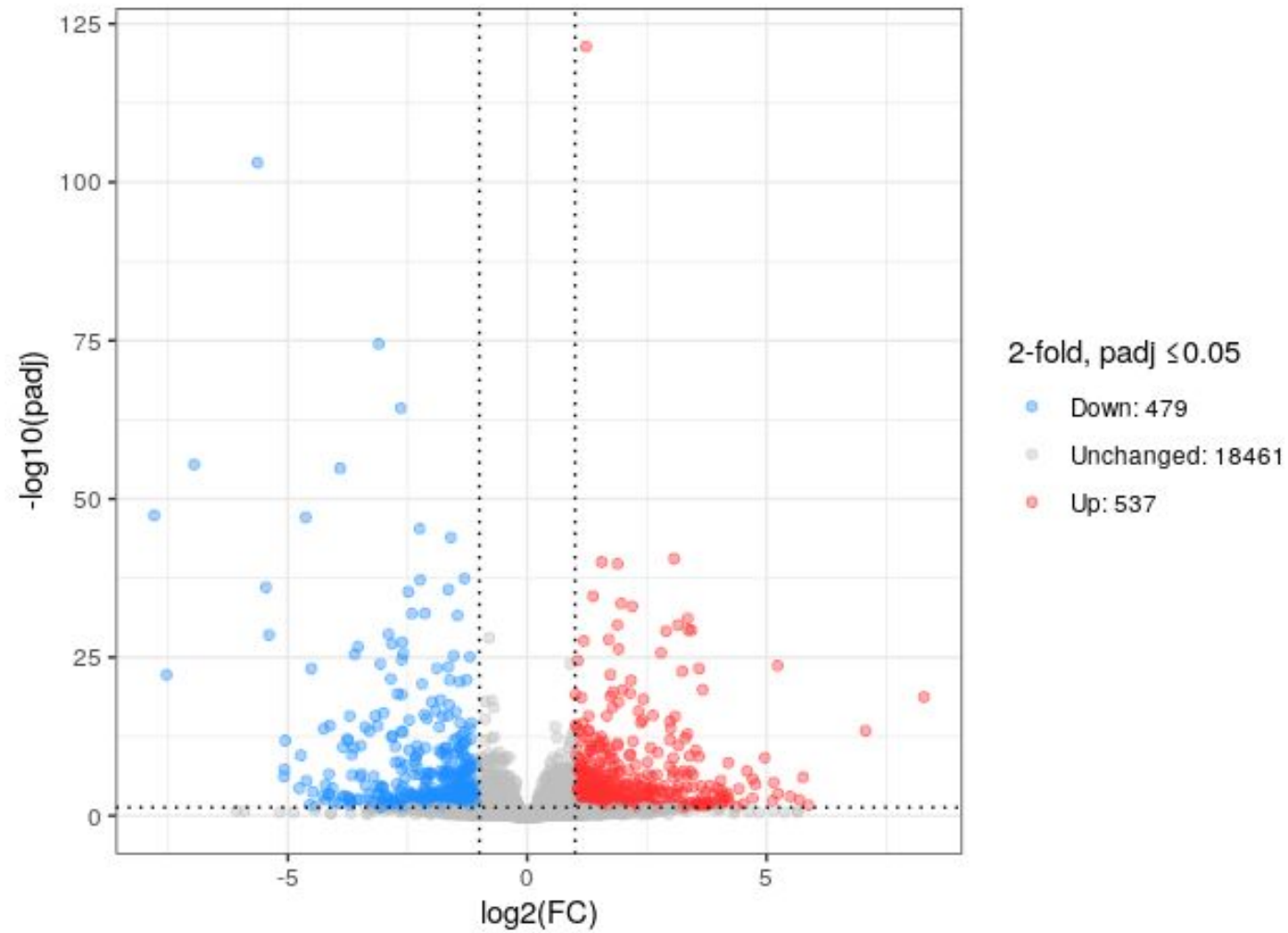


Bayesian procedure to moderate (or “shrink”) log<sub>2</sub> fold changes from genes with very low counts and highly variable counts

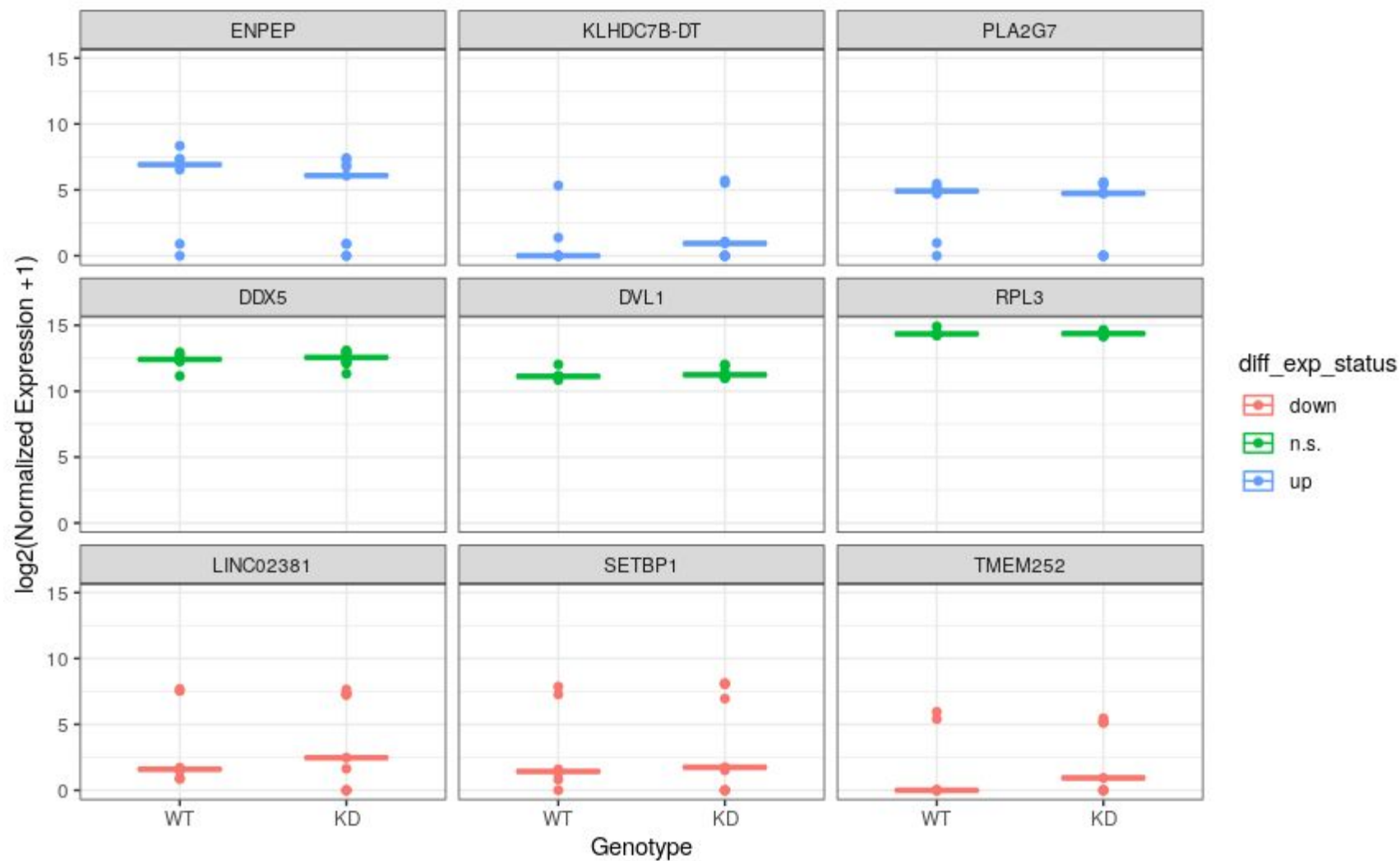
-RNA-seq workflow: gene-level exploratory analysis and differential expression

# What do our gene counts look like?

## Volcano Plot

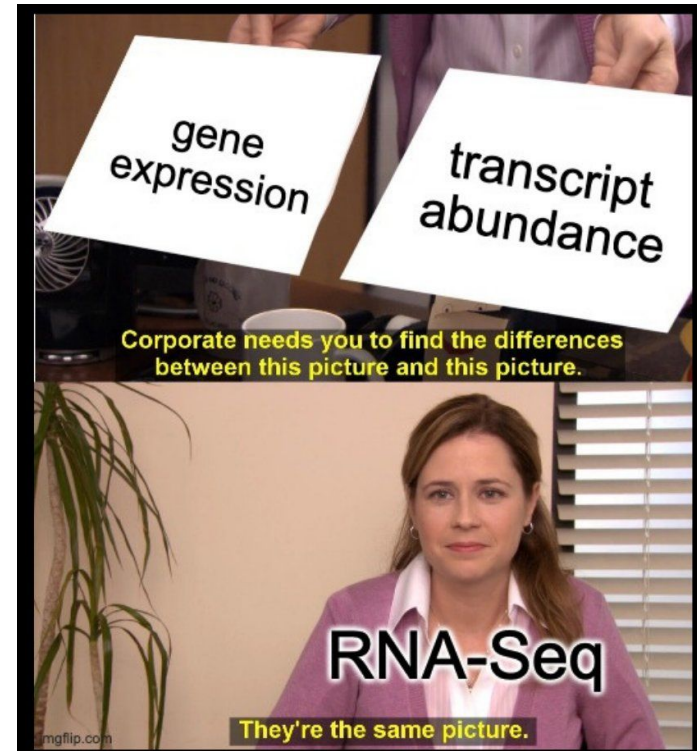


# What does our normalized expression look like?



# Differential expression live demo

Ask Questions!



# Have questions? Contact us via email or attend office hours



Attend Office Hours every week:

- 10:00 am ET Tuesday
- 2:00 pm ET Thursday



**Manisha Ray**

[manisha.ray@sevenbridges.com](mailto:manisha.ray@sevenbridges.com)



**Zelia Worman**

[zelia.worman@sevenbridges.com](mailto:zelia.worman@sevenbridges.com)



**Phil Webster**

[phil.webster@sevenbridges.com](mailto:phil.webster@sevenbridges.com)



<https://www.cancergenomicscloud.org>

# Q&A and Discussion



GGTGGGATAC  
TAATAATTT  
ACCCATGAC  
AATCATTAG  
CTTCAACGA  
AATTGGAAT  
TACTATATTT  
TACTCAAA  
GGGACTATA  
AATATTTTC  
CCATGACCC  
AATCATTAG  
ACTTCAACGA  
AATTGGAAT  
GGGACTATA  
AATCATTAG

ACCCTATGACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGTCTGTACAGCGGGTGATCTCAATGGCTAAGGCTTACGCCGTACTACCTCAGCAGTAGTAAGA AAAATATTTTTCACCCACCCCTA  
SACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGCTGAAGTTGCTACCTCATTAAAGAACGGAGAAGTATCCATTACGAAAGACGGGATCGCAGTCTATTATGATTCATAGATAATTTTTCACCC  
CATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGTATAAAAAGTGGTGGGATACGSSAAITGGAAITAGTAATCAGITTTATGTGTATGCCACCTACCGGGCATATGGCTATCGACATCGAGAAATATTTTTCACCC  
CAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGATAATTTTTCACCCACCCCTATGACCCCTAACCTTAATCATTAGTCAAATACACATACCCGTGGATGGCCCGTATACCGATAGCTGTAGCTTTGTAATGGGTGTAATTTCTTAAACATAGTCAA  
GCCTCTTCATAGGTAATGCTTTCACATAGAGACAGTCCGCCACTAGAGTTACCGAATCCGAATGCGGCATGATGGAGTCTGTATCTATTGTTAATTCGTTAACTGTTGGAGGAAGAATAAGAAATCTTTGTTATATATTTTTCACCCACCCCTATGACCCCTAACCT  
GGTGGGATACTATAATTTTTCACCCACCCCTATGACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGATAATTTTTCACCCACCCCTATGACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACG  
ATGACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGATAATTTTTCACCCACCCCTATGACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTA  
GGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGATAATTTTTCACCCACCCCTATGACCCCTAACCTTAATCATTAGTCAAATAGACTTCAACGATGGAGTAATCTTGCCTCTTCATAGGTAATGCTTTCACATAGCTGGATGGCCCGTATACCGATAGC