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

Phillip Webster, Genomics Scientist

March 22, 2022

CANCER GENOMICS CLOUD
SEVEN BRIDGES

CANCER GENOMICS CLOUD
SEVEN BRIDGES

Class Overview

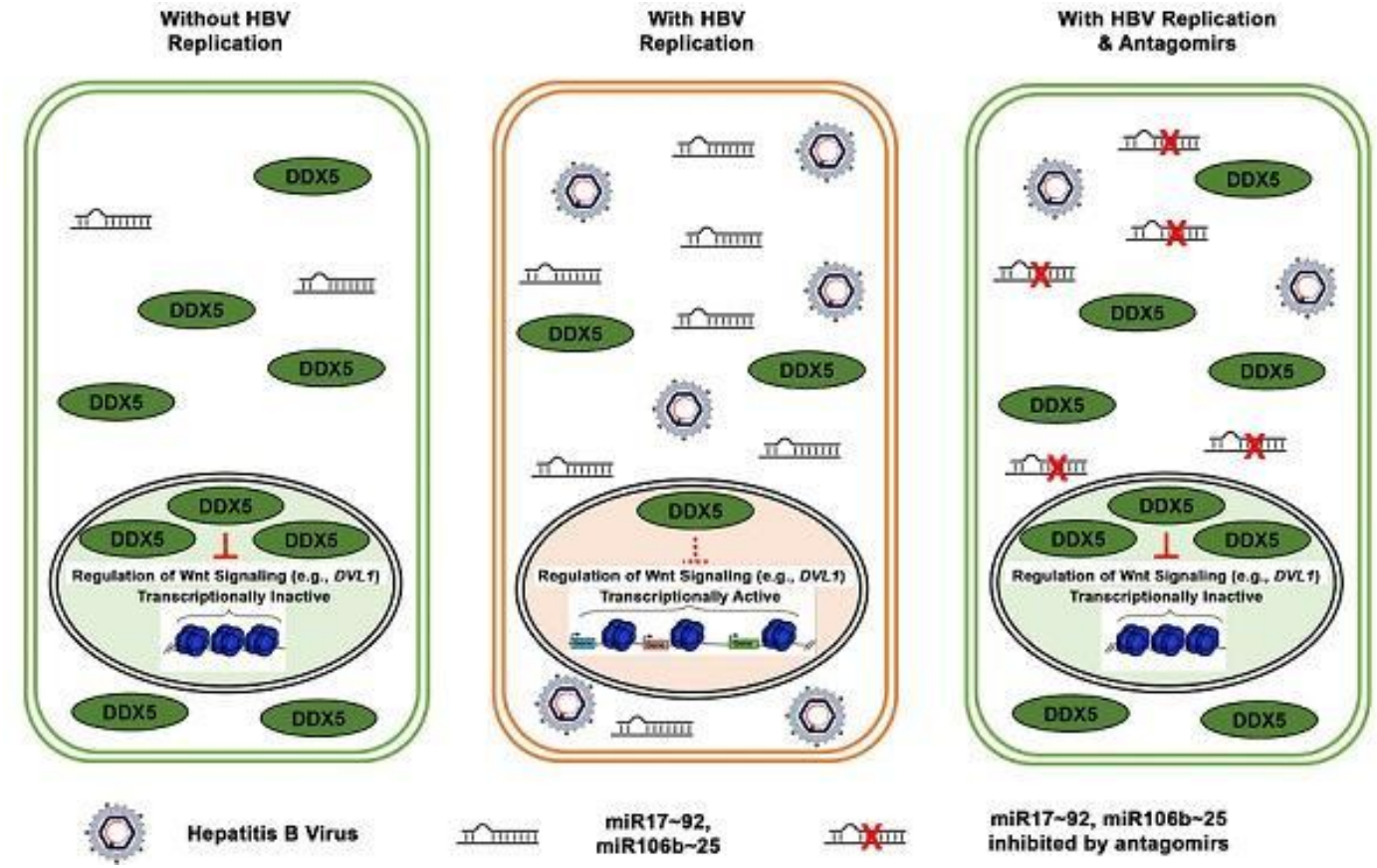
- Lecture 1 - Overview
  Hands-on demo: your first project on the CGC
- Lecture 2 - Bulk RNA-seq workflow(s) and sequence alignment
  Hands-on demo: your CGC workflow
- **Lecture 3 - Differential expression and visualizations**
  Hands-on demo: your CGC visualization
- Lecture 4 - Single-cell RNA-seq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

CANCER GENOMICS CLOUD
SEVEN BRIDGES

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

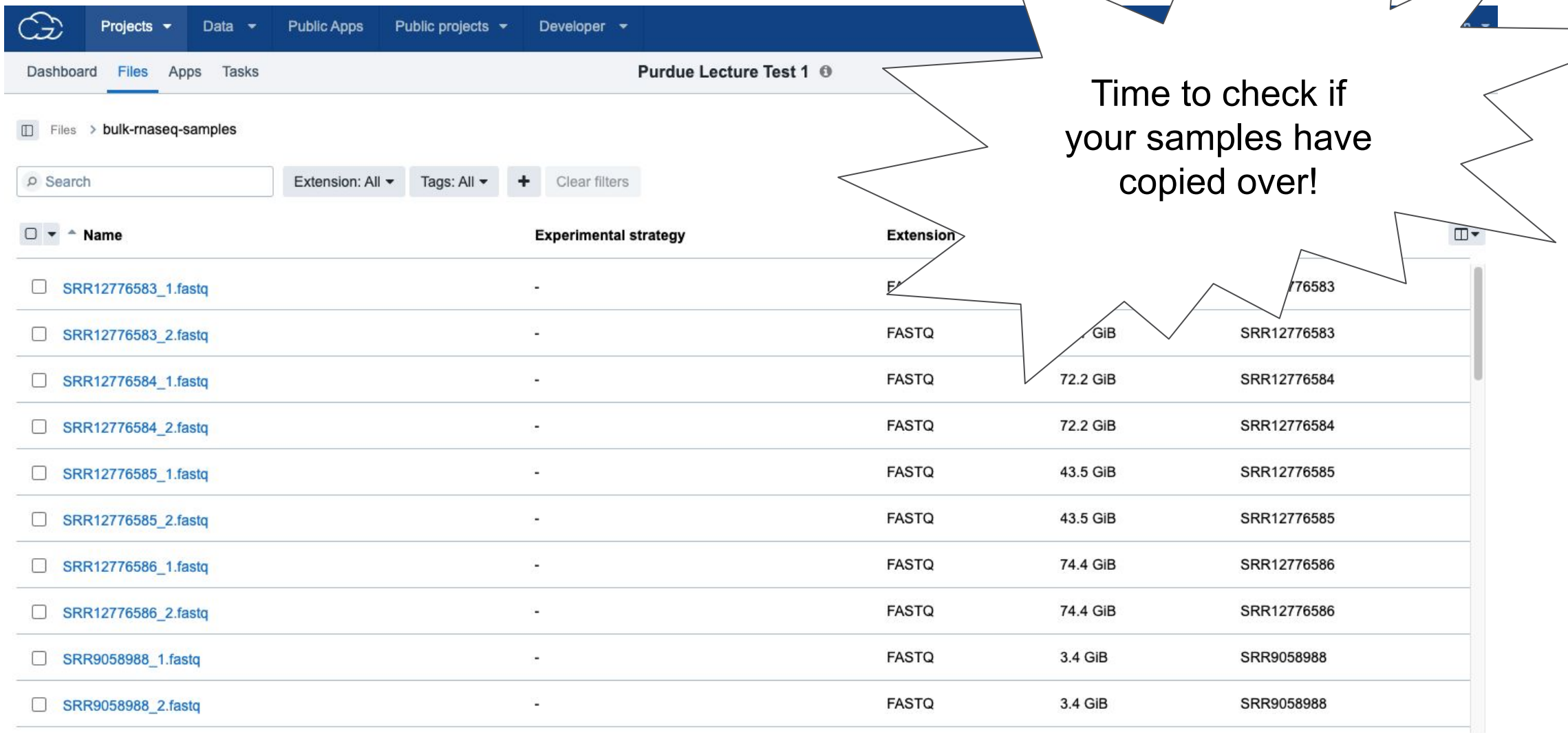


Mani SKK, et al. Restoration of RNA helicase DDX5 suppresses hepatitis B virus (HBV) biosynthesis and Wnt signaling in HBV-related hepatocellular carcinoma. *Theranostics* 2020; 10(24):10957-10972. doi:10.7150/thno.49629. https://www.thno.org/v10p10957.htm

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



Time to check if your samples have copied over!

RNA-seq workflow overview

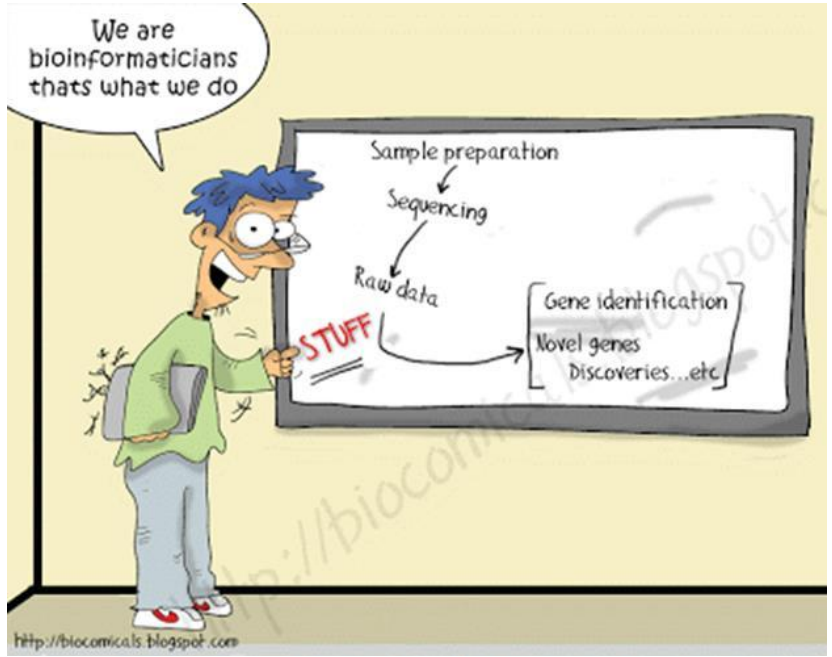
Design Experiment
- Set up the experiment to address your specific biological questions
- Meet with your bioinformatician and sequencing center!

RNA preparation
- Isolate RNA
- Purify RNA

Prepare Libraries
- Convert the RNA to cDNA
- Add sequencing adapters

Sequence
- Sequence the cDNA using a sequencing platform

Analysis
- Quality control
- Align reads to the genome/assemble a transcriptome
- Downstream analysis based on your questions



We are here!

Slide adapted from Dr. Nadia Atallah, Purdue University

RNA-seq User Flow

Create a Project → Select/access data → Select/create tools → Create and run analysis

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: rnaseq_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

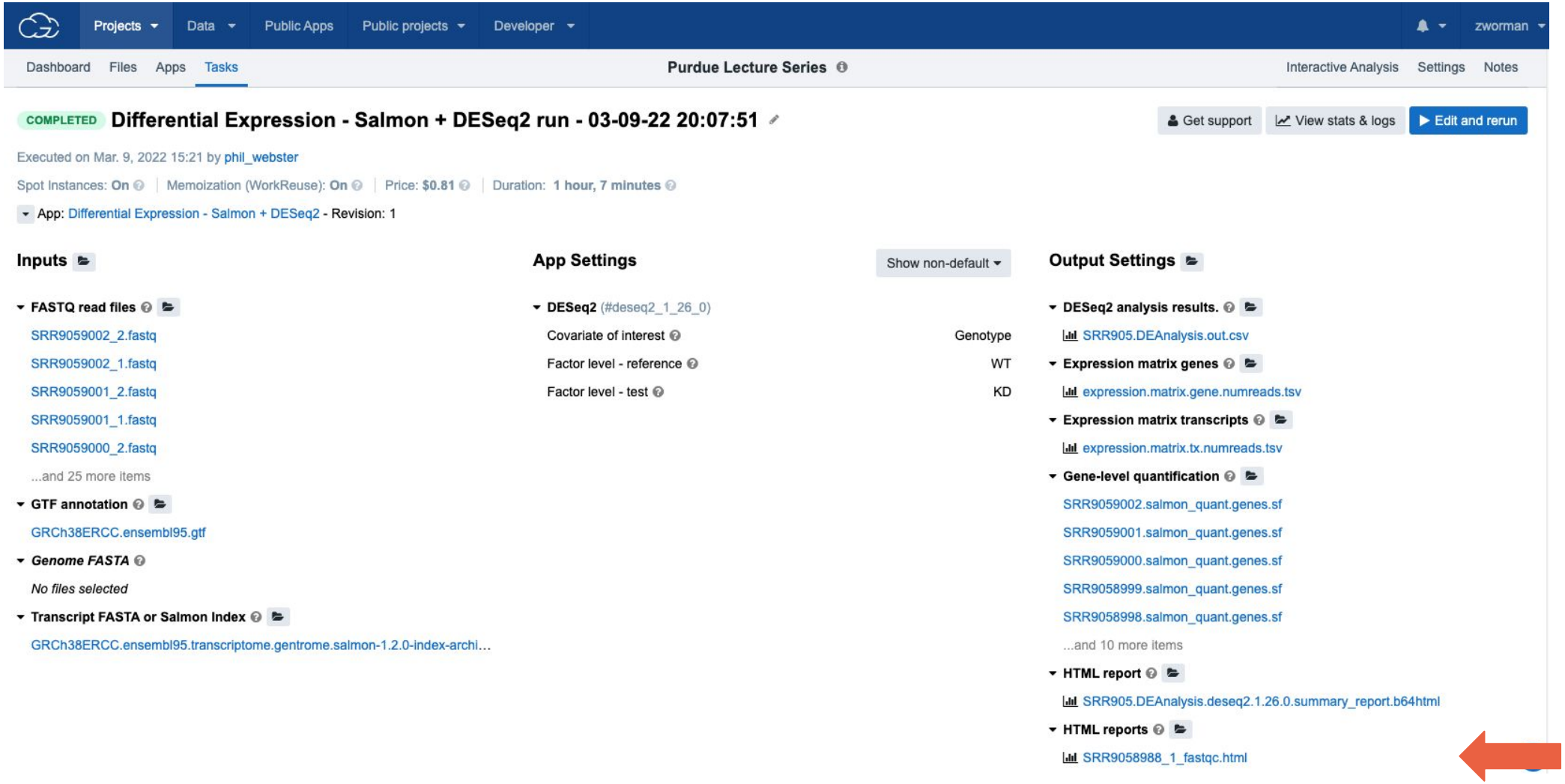
Factor level - reference: WT

Factor level - test: KD

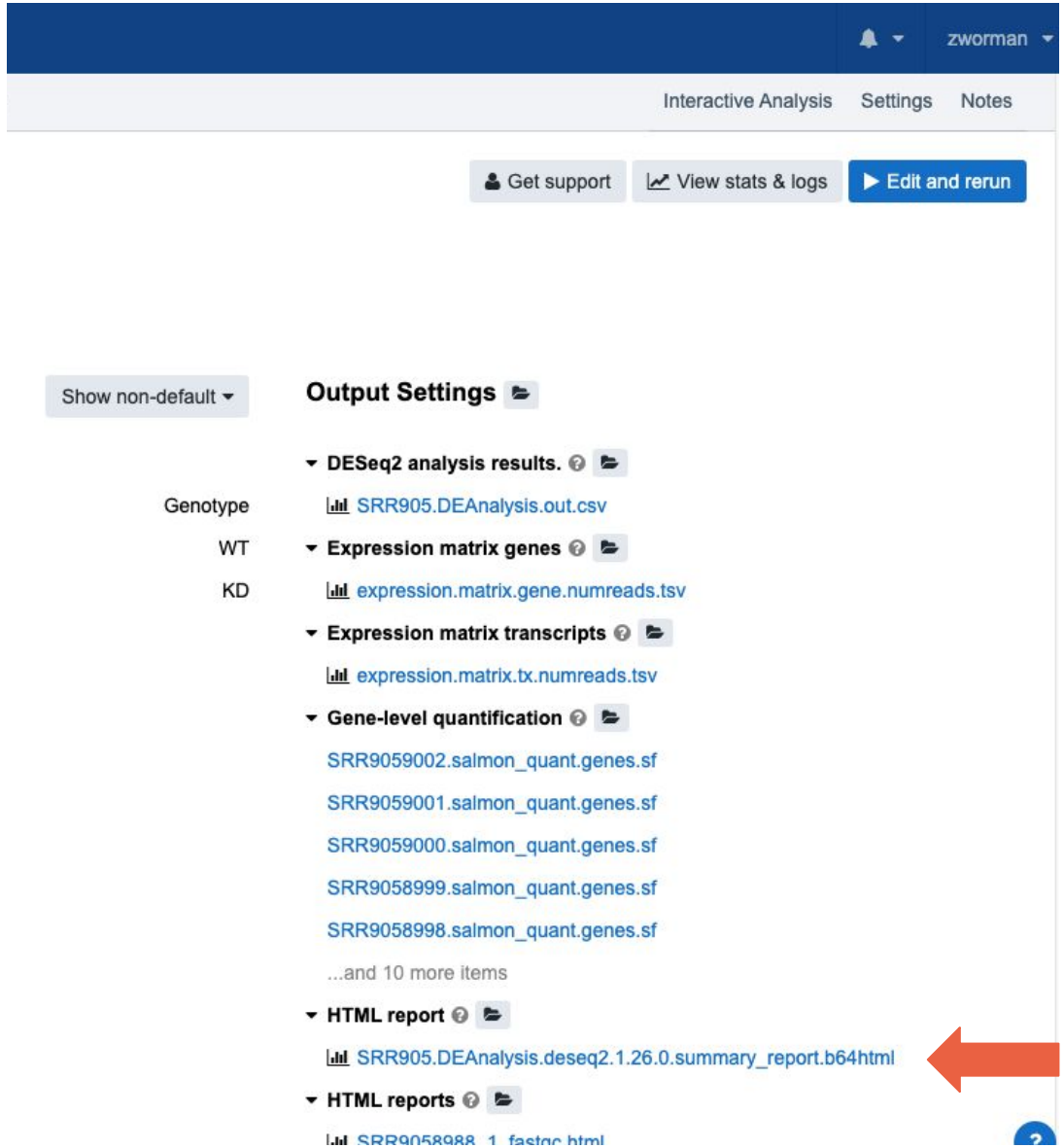
Output Settings

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?



Outputs of the RNA-seq User Flow



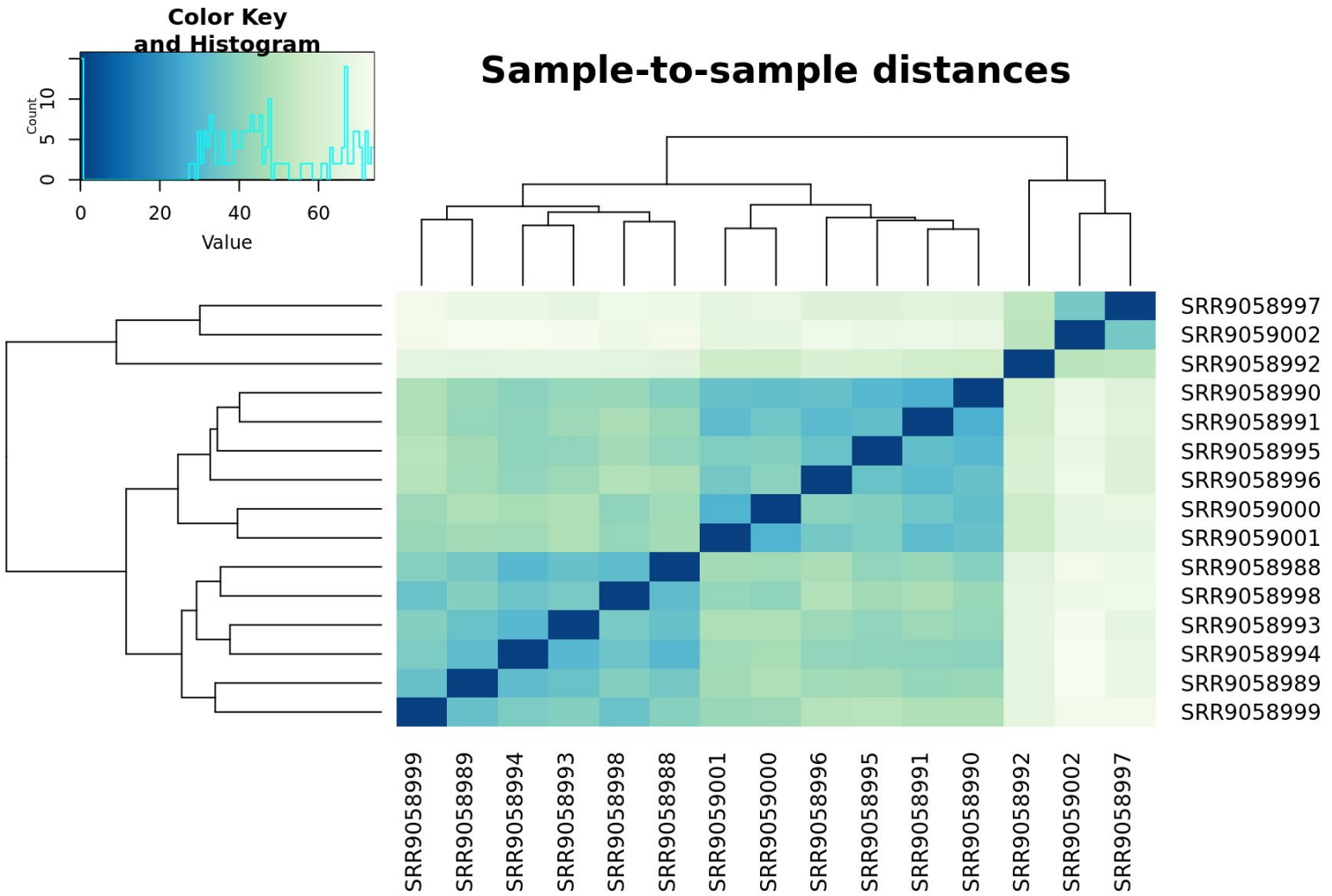
CANCER GENOMICS CLOUD
SEVEN BRIDGES

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

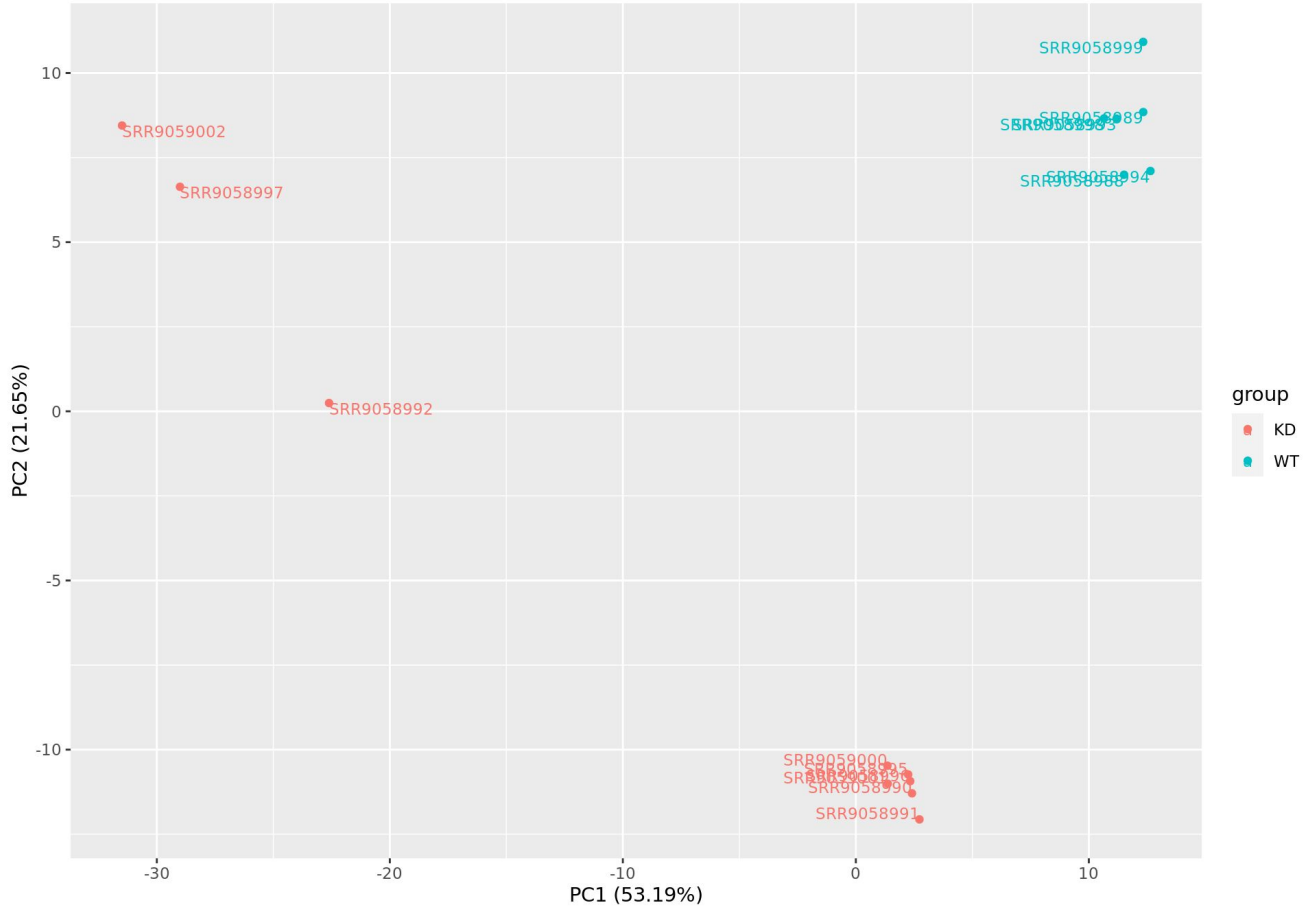
Sample Distance Heat Map



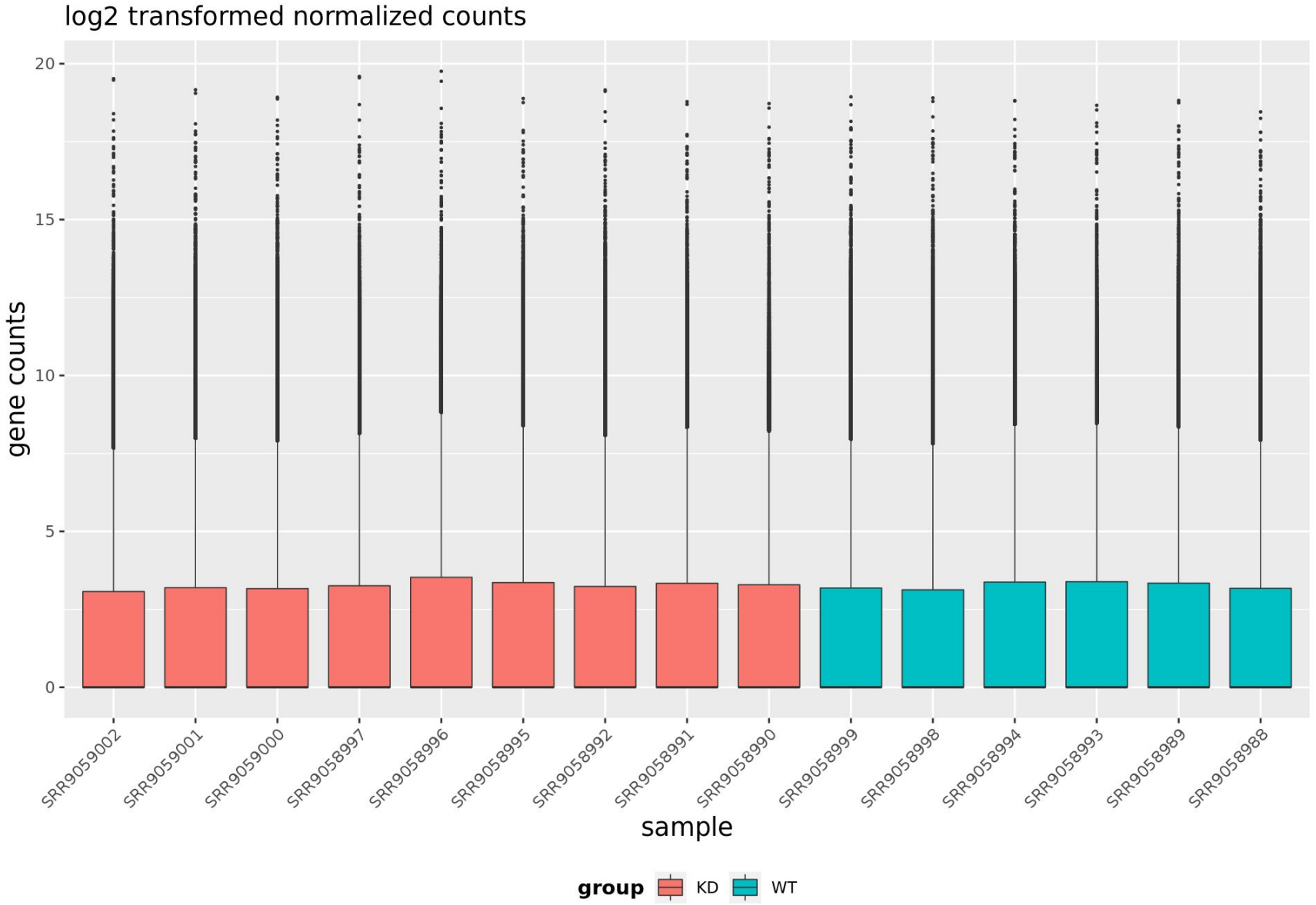
Normalized gene counts

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

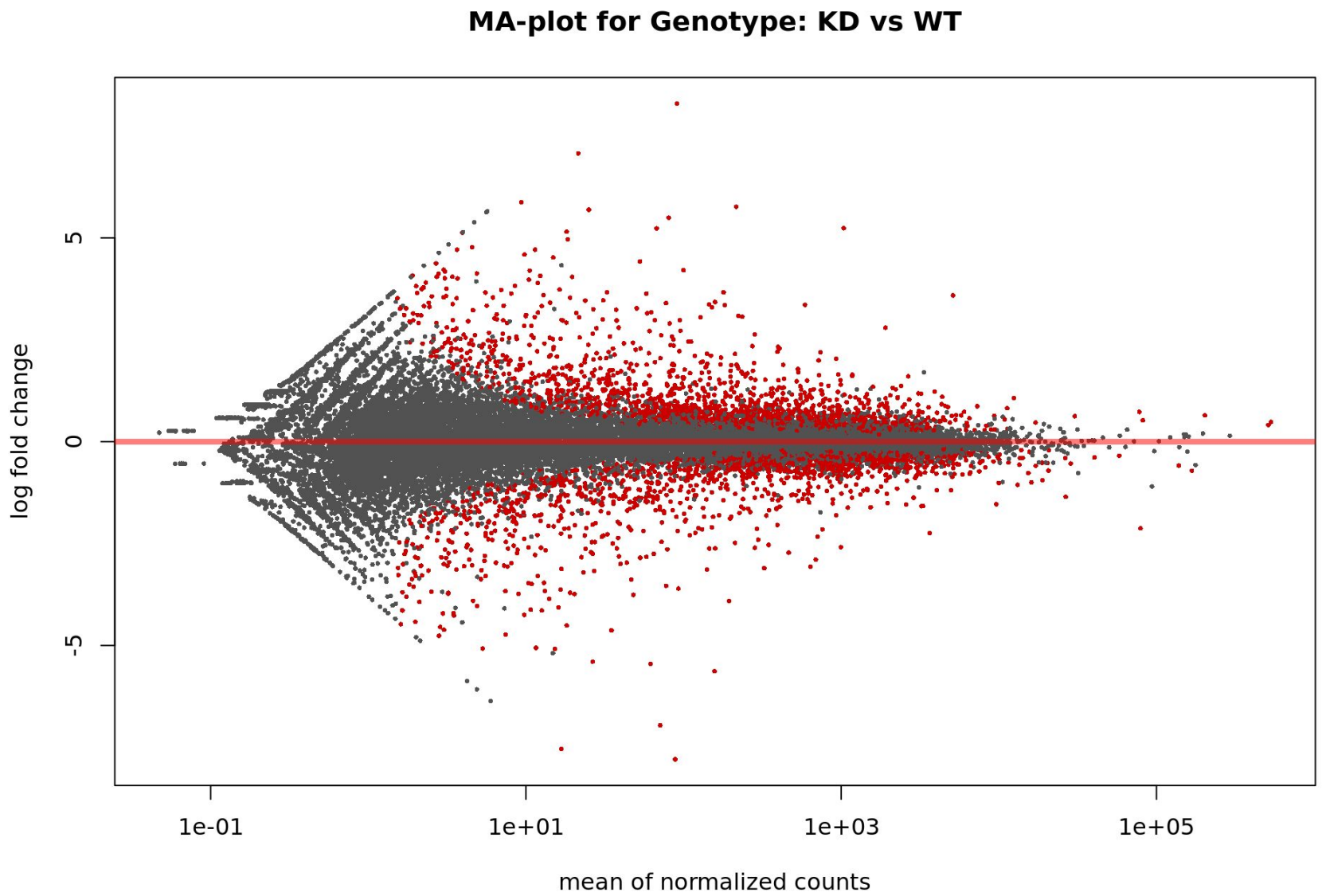
- Principal component analysis (PCA) is a statistical procedure that can be used for exploratory data analysis. PCA uses linear combinations of the original data (e.g. gene expression values) to define a new set of unrelated variables (principal components). T
- Thus, PCA can be used to reduce the dimensions 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



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

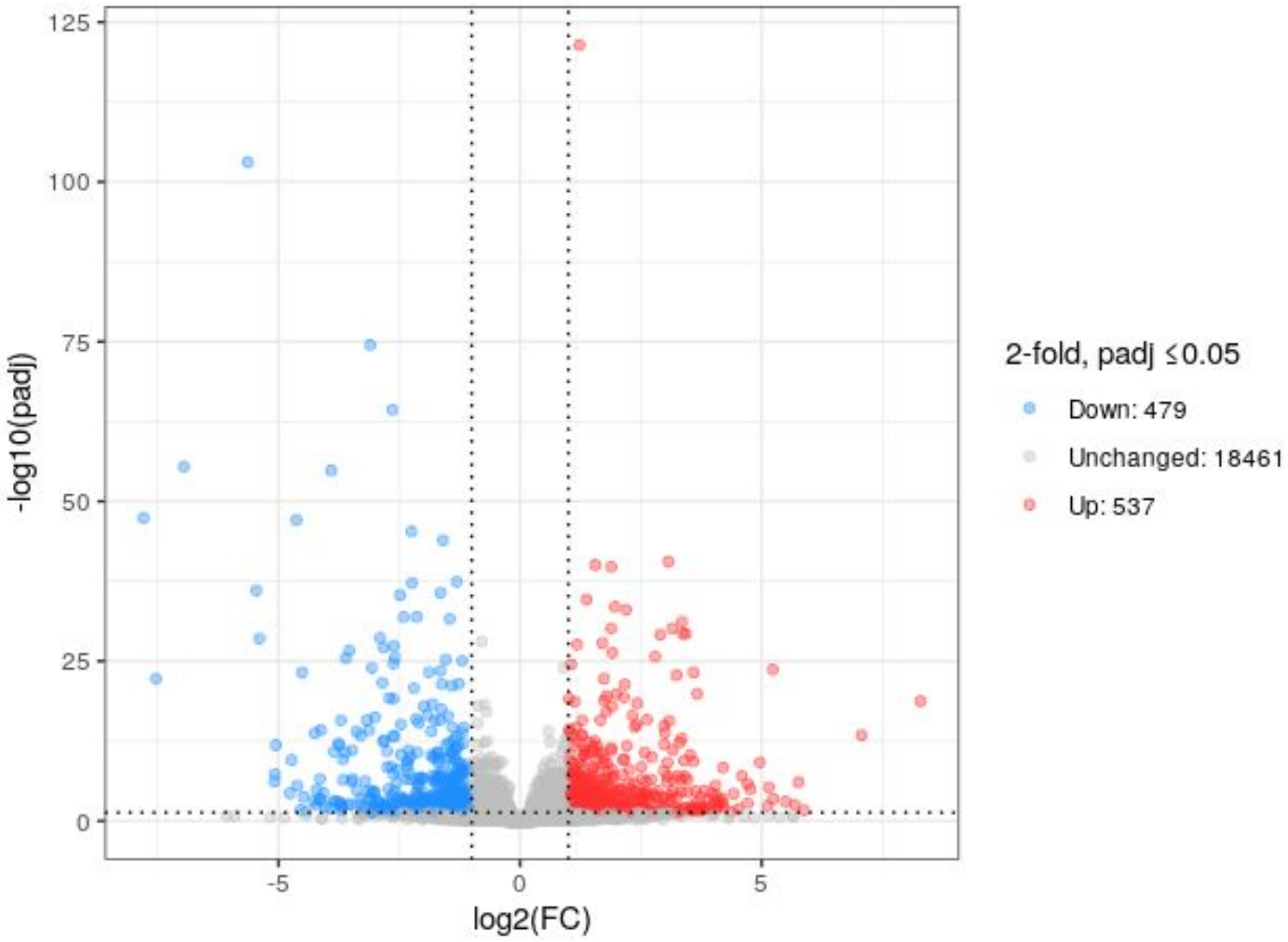


Bayesian procedure to moderate (or “shrink”) log2 fold changes from genes with very low counts and highly variable counts

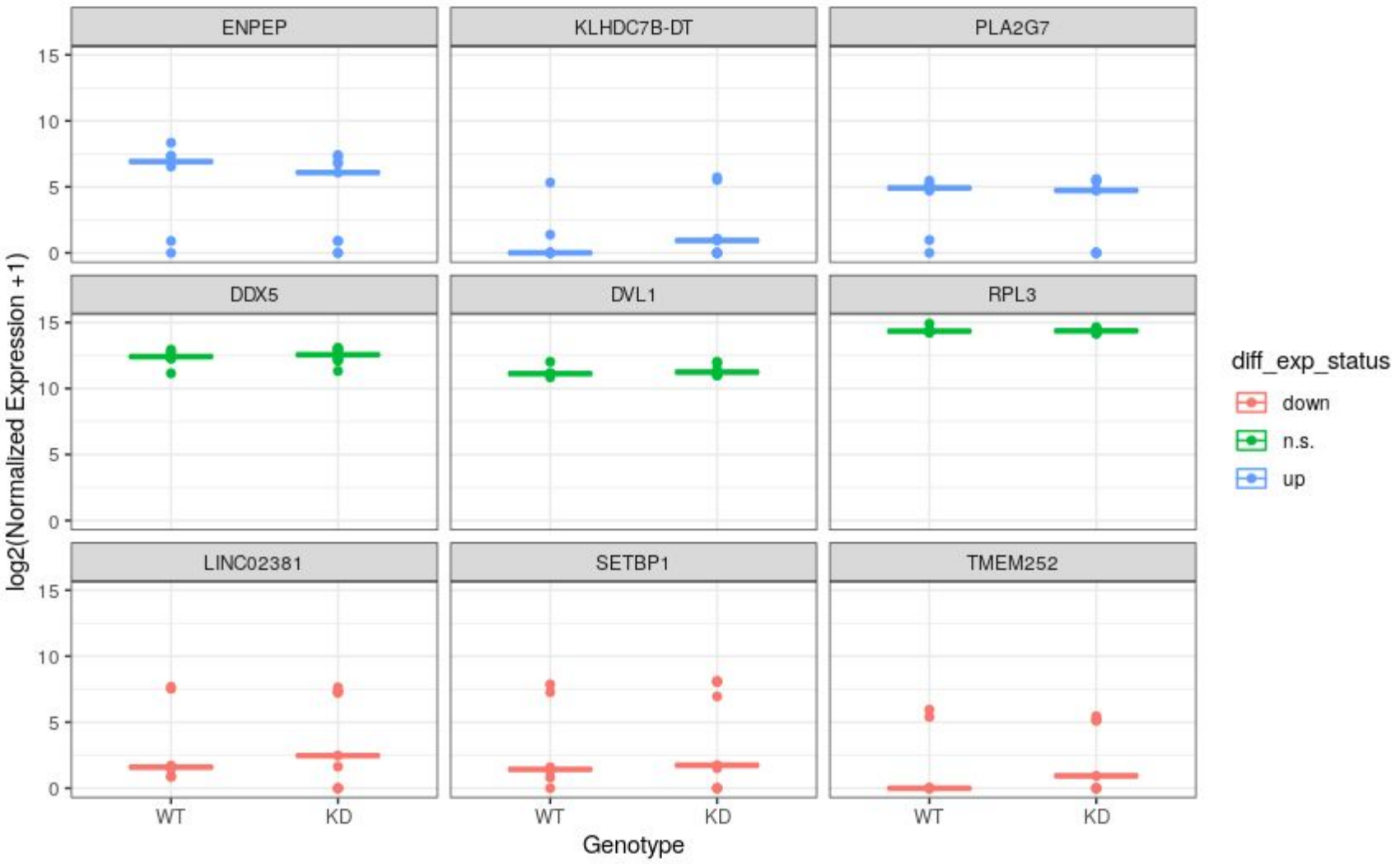
-RNA-seq workflow: gene-level exploratory analysis and differential expression

What do our gene counts look like?

Volcano Plot

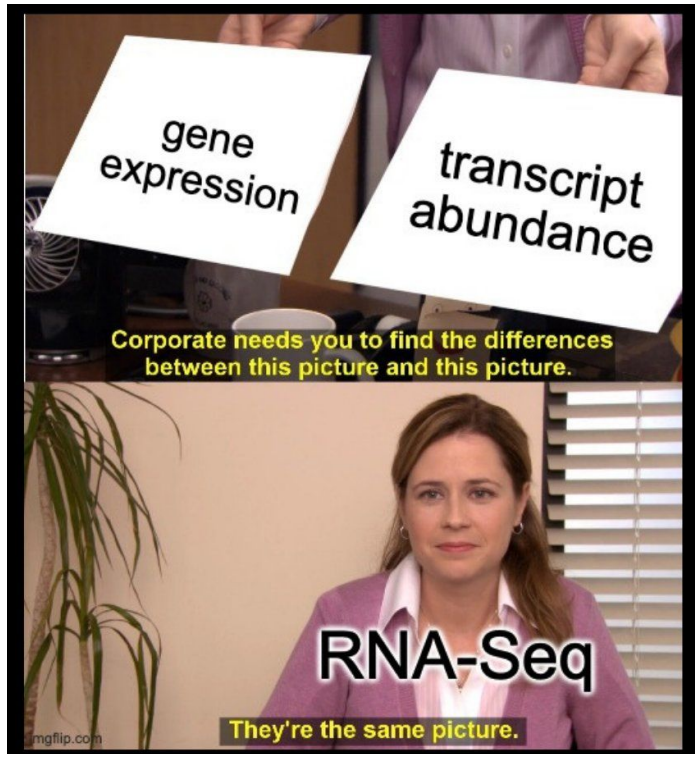


What does our normalized expression look like?



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

Zelia Worman

zelia.worman@sevenbridges.com

Phil Webster

phil.webster@sevenbridges.com

CANCER GENOMICS CLOUD
SEVEN BRIDGES

https://www.cancergenomicscloud.org

Q&A and Discussion

CANCER GENOMICS CLOUD