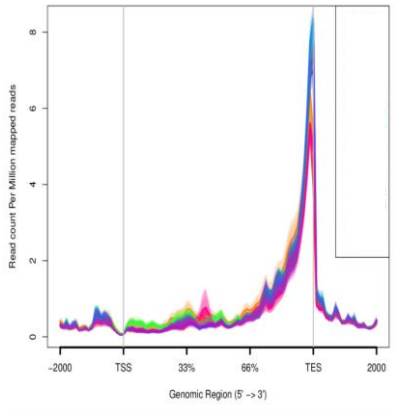
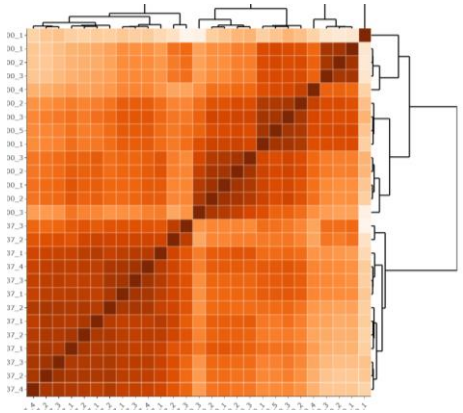
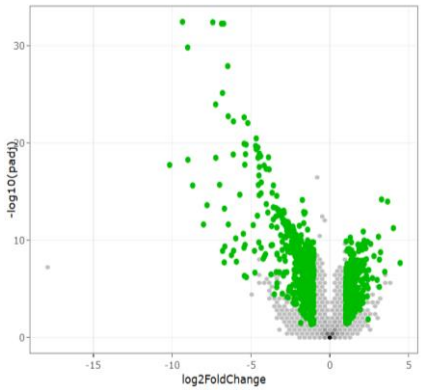
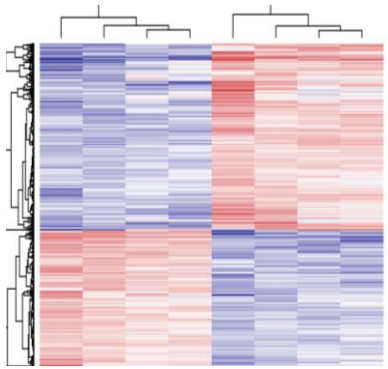
Hands-on: User-friendly Transcriptome Analysis Pipeline

1. From SandBox samples to sequenced reads
2. Setting up a new transcriptome analysis
3. How to read the MARS-seq report









Bareket Dassa, June 2020

Before we start….

Ronald A. Fisher
(1890-1962)

"To consult the statistician **after** an experiment is finished
is often merely to ask him to conduct a post mortem examination.
He can perhaps say what the experiment died of." (1938)

Think about your experimental design:

1. Biological question
2. Quality of the input material (method of RNA extraction)
3. Replicates
4. Batch effects
5. Coverage: how many reads?
6. Select library protocol, Paired- or single-end sequencing?

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Come and consult us **before** performing your experiment

Bioinformatics unit
NGS data analysis

Noa Wigoda

Ester Feldmesser

Dena Leshkowitz
Unit head

Bareket Dassa

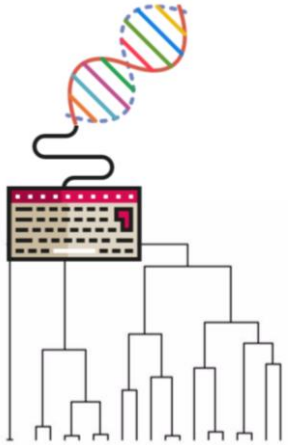
Gil Stelzer

This training does not cover the bioinformatics theory behind RNA-seq analysis

Learn more with our:

Workshop: Introducing UTAP: User-friendly Transcriptome Analysis Pipeline **(July 6th 2020)**
Course: An Introduction to deep-sequencing analysis for biologists 20203331
e-learning tool (to be released)

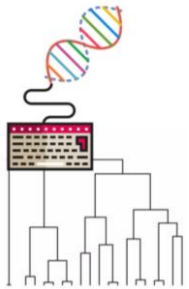
Part 1:



From SandBox samples to sequenced reads

where are my files?

From sequencing to analyzed data



Developed by the Bioinformatics unit (LSCF)

Start a sequencing run

Prepare and test your Sample Sheet

Login to susanc3 server

Select your NGS machine

Sample demultiplexing

No *Yes*

Select experimental protocol

Retrieve your sequences

Set up a transcriptome analysis using UTAP

First QC report

Detailed guidelines:

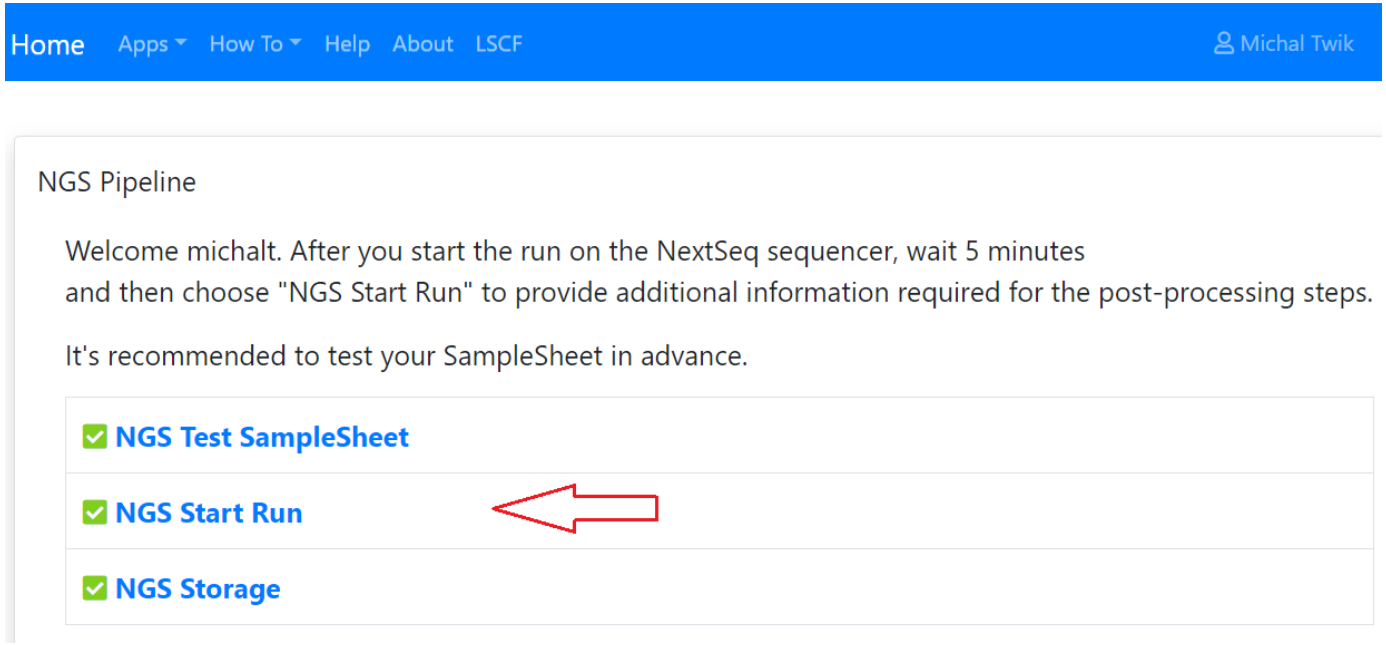
https://bbcunit.atlassian.net/wiki/spaces/BP/pages/104431617/From+samples+to+analyzed+NGS+data+UTAP

Once a sequencing run starts
you must log in to the NGS Pipeline (susanc3) website and register the run

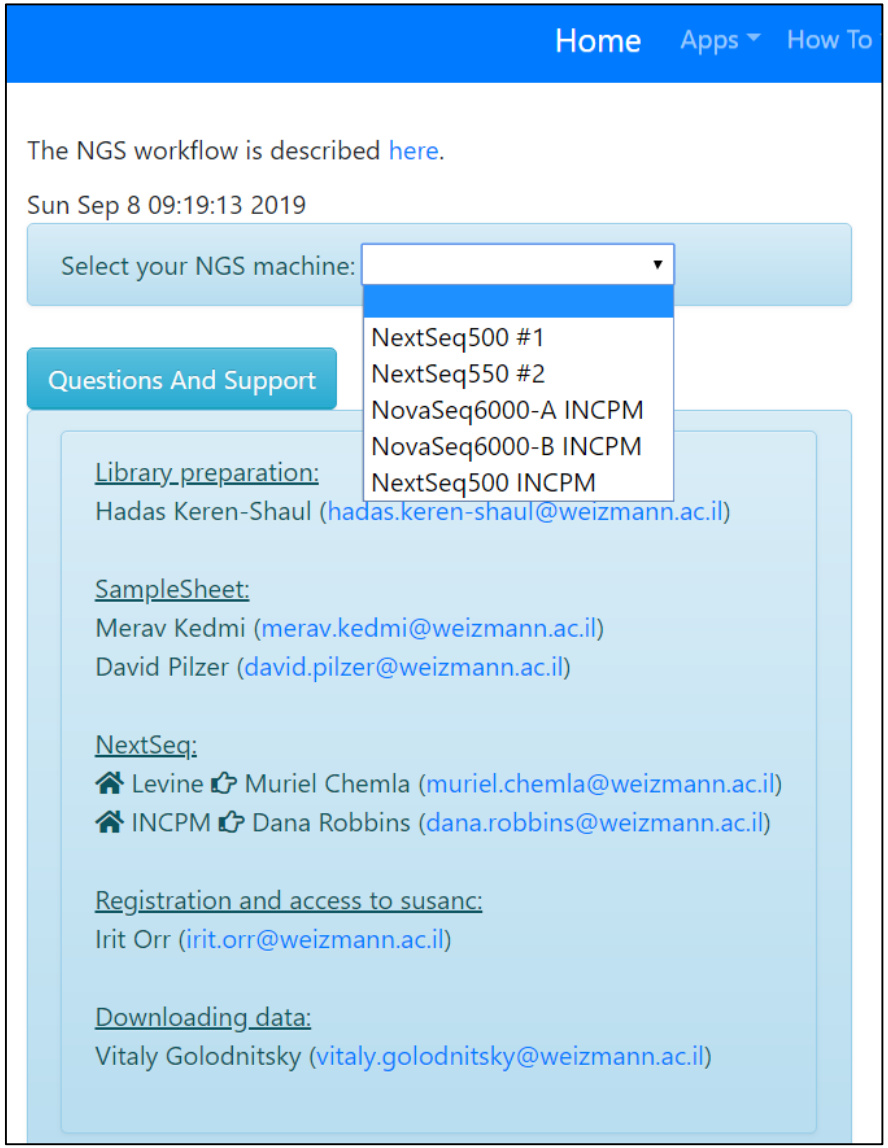
Start
a sequencing run

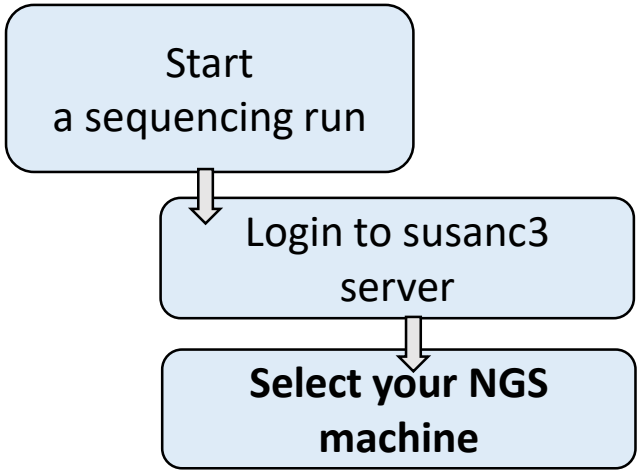
Login to susanc3 server

Obtain an account (userID and password)
on https://ngs-pipeline.weizmann.ac.il/
from Irit Orr at 08-934-2470, irit.orr@weizmann.ac.il



Select your sequencing machine

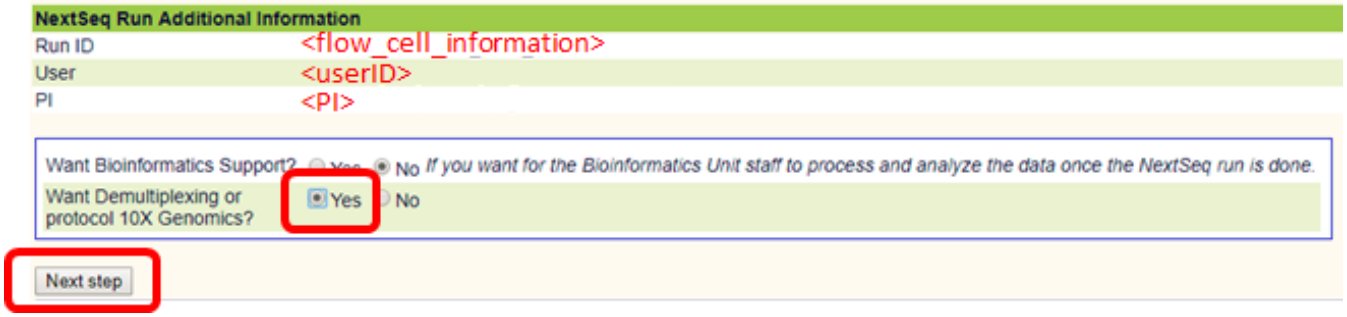


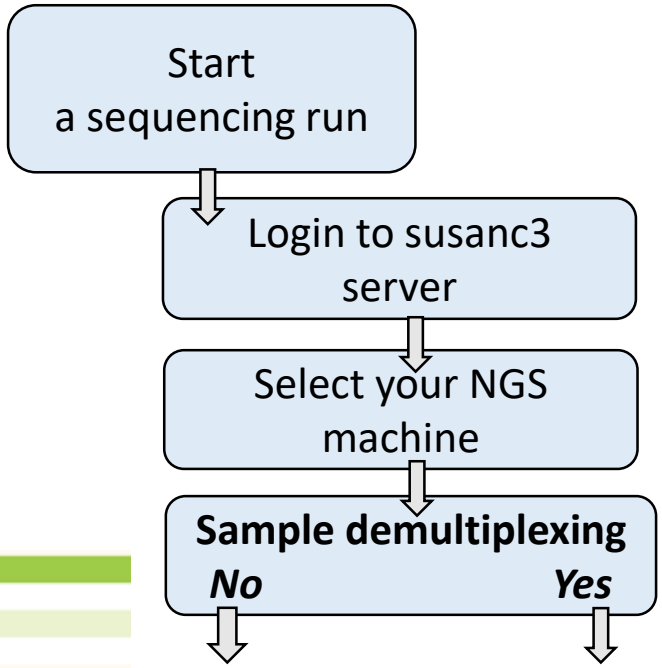


Warning: Use NovaSeq only for unique **dual** indexing pooling combinations (unique i5 and i7 **indexes**). Index hopping or index switching is a known phenomenon in NovaSeq. It causes incorrect assignment of libraries from the expected index to a different index (in the multiplexed pool).

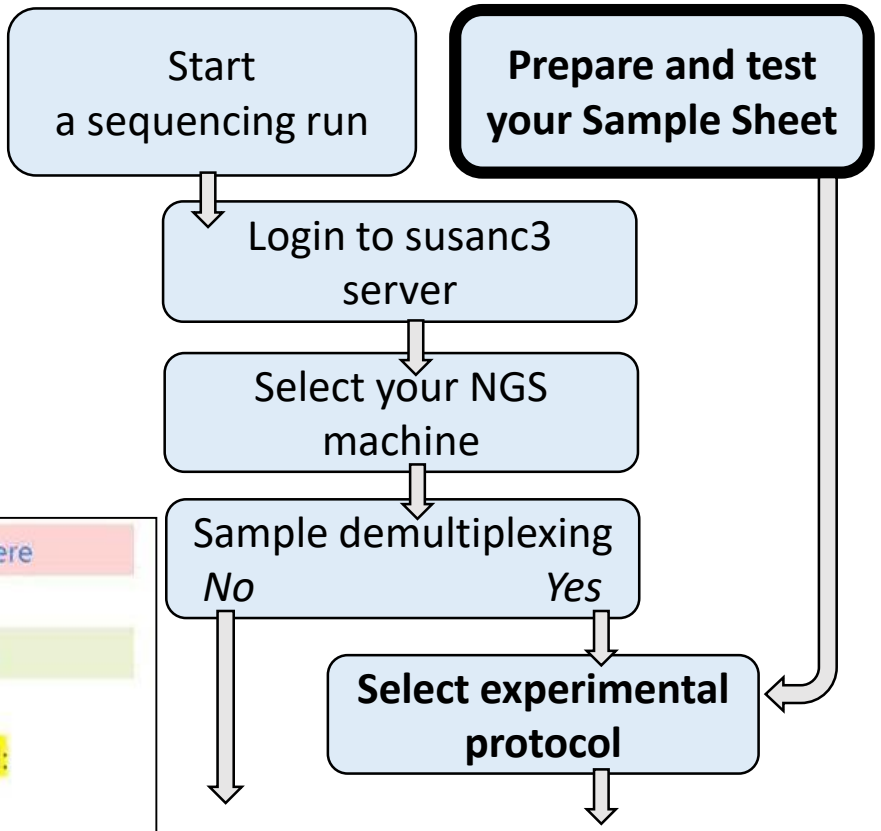
Select for demultiplexing

to get your reads separated into samples (according to sample barcodes provides in your SampleSheet)





Prepare your Sample Sheet according to your sequencing protocol template



It's recommended to **test in advance** your SampleSheet or MARSseq file here

Please select one of these options by uploading the corresponding file

MARS-Seq if you prepared your samples following the MARS-Seq protocol:
Please upload the MARS-Seq excel file (sample)
Choose File No file chosen

Illumina if you prepared your samples following an Illumina protocol:
Please upload the SampleSheet (sample) Choose File No file chosen
You may generate a SampleSheet with the Illumina Experiment Manager

10X Genomics if you prepared your samples following the 10X Genomics protocol:
Please upload the SampleSheet (sample) Choose File No file chosen

Example of MARS-seq Sample Sheet

A	B	C	D	E	F	G	H	I	J	K
#	Sample ID	rd1 barcode	rd1 barcode name	rd2 barcode name	rd2 barcode	i7 index barcode	i7 index name	i5 index barcode	i5 index name	userid
1	CTR_1			v3_gr1C2	GCAGTAG					sorek
2	CTR_2			v3_gr1B1	ACTCAGG					sorek
3	CTR_3			v3_gr1C1	ACACGTG					sorek
4	CTR_4			v3_gr1D1	TCTTCGA					sorek
5	CTR_5			v3_gr1E1	AGCACTG					sorek
6	KO_1			v3_gr1F1	AGTGCGT					sorek
7	KO_2			v3_gr1G1	AACGTTC					sorek
8	KO_3			v3_gr1H1	ATCTGTC					sorek
9	KO_4			v3_gr1A2	GTATGCG					sorek
10	KO_5			v3_gr1B2	ATCTGCA					sorek
11	WT_1			v3_gr1C2	GCAGTAG					hadask
12	WT_2			v3_gr1D2	CAGTACG					hadask
13	WT_3			v3_gr1E2	TGCACAA					hadask

Give meaningful names for your samples!

Note: Valid characters for sample names are **A-Z a-z 0-9 . _ -**

Don't use special characters such as " ' ` ? , ; + = @ # $ % ^ & () [] {} <> / \ in sample names.

Don't use Hebrew, Arabic, Chinese or any character-set other than English (Roman alphabet)

Test your SampleSheet!

Example of MARS-seq Sample Sheet

It's recommended to **test in advance** your SampleSheet or MARSseq file CLICK HERE

A	B	C	D	E	F	G	H	I	J	K
#	Sample ID	rd1 barcode	rd1 barcode name	rd2 barcode name	rd2 barcode	i7 index barcode	i7 index name	i5 index barcode	i5 index name	userid
1	CTR_1			v3_gr1C2	GCAGTAG					sorek
2	CTR_2			v3_gr1B1	ACTCAGG					sorek
3	CTR_3			v3_gr1C1	ACACGTG					sorek
4	CTR_4			v3_gr1D1	TCTTCGA					sorek
5	CTR_5			v3_gr1E1	AGCACTG					sorek
6	KO_1			v3_gr1F1	AGTGCGT					sorek
7	KO_2			v3_gr1G1	AACGTTC					sorek
8	KO_3			v3_gr1H1	ATCTGTC					sorek
9	KO_4			v3_gr1A2	GTATGCG					sorek
10	KO_5			v3_gr1B2	ATCTGCA					sorek
11	WT_1			v3_gr1C2	GCAGTAG					hadask
12	WT_2			v3_gr1D2	CAGTACG					hadask
13	WT_3			v3_gr1E2	TGCACAA					
14	WT+treatment_1			v3_gr1F2	CGTAACT					
15	WT+treatment_2			v3_gr1G2	AGCTCAA					
16	WT+treatment_3			v3_gr1H2	ATAACCG					
17	mutant_1			v3_gr1A3	TGTCACG					
18	mutant_2			v3_gr1B3	TTCCTGA					
19	mutant_3			v3_gr1C3	GGATCTA					
20	Mutant-tratment_1			v3_gr1D3	TACCAGT					
21	Mutant-tratment_1			v3_gr1E3	GGAGACT					
22	Mutant-tratment_1			v3_gr1F3	AGCTAGT					

Test here your NextSeq SampleSheet

Please fix and resubmit
Illegal characters Sample 'WT+treatment_1'
Illegal characters Sample 'WT+treatment_2'
Illegal characters Sample 'WT+treatment_3'
Duplicated barcode GCAGTAG on CTR_1, WT_1

File to test Choose File Mars-seq_users.xlsx

Submit

When sharing a flow cell with other users:

- **Use unique sample barcodes**
- **Do not mix sequencing protocols in one run**
- **Do not overload**

Retrieving your NGS data:

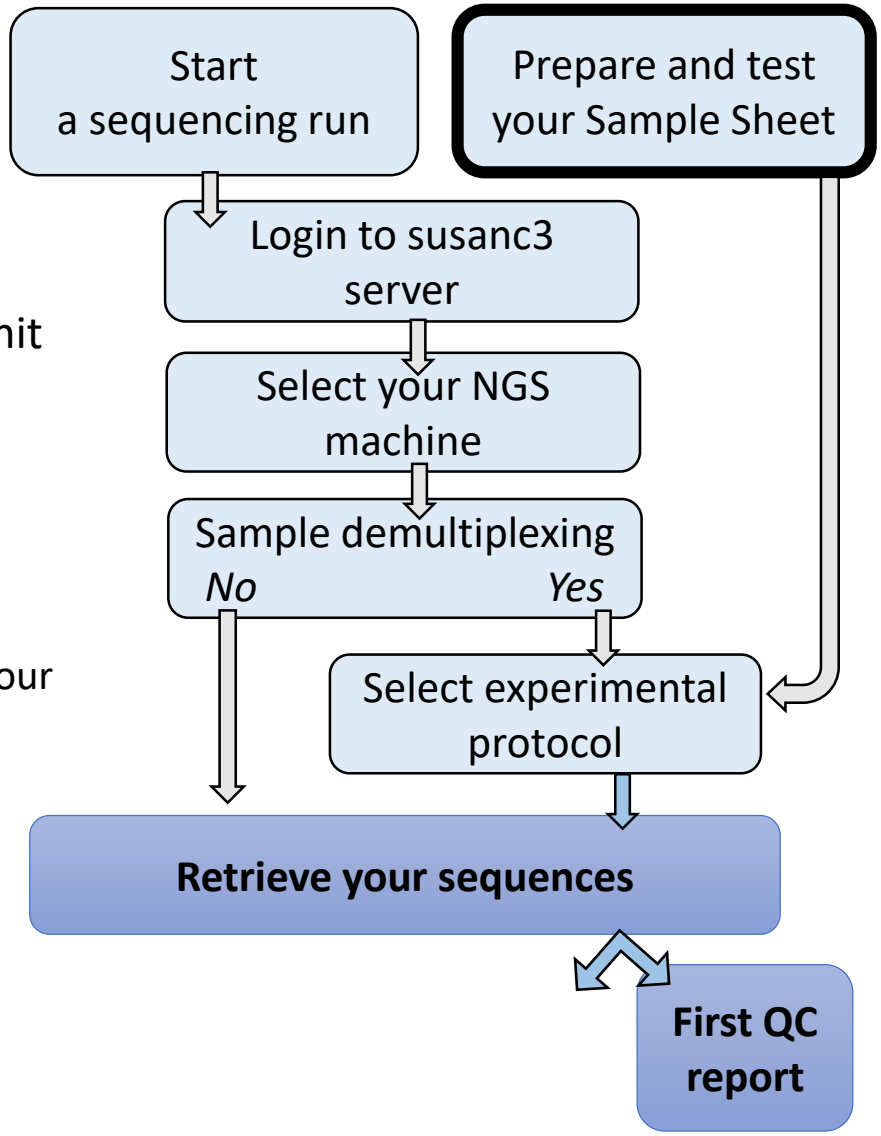
You will receive an email from the Bioinformatic-Unit with:

1. A link to your Raw (Bcl and Fastq) output files
http://stefan.weizmann.ac.il/fqc/**RUN_ID**

Data is temporarily stored only for 3 months

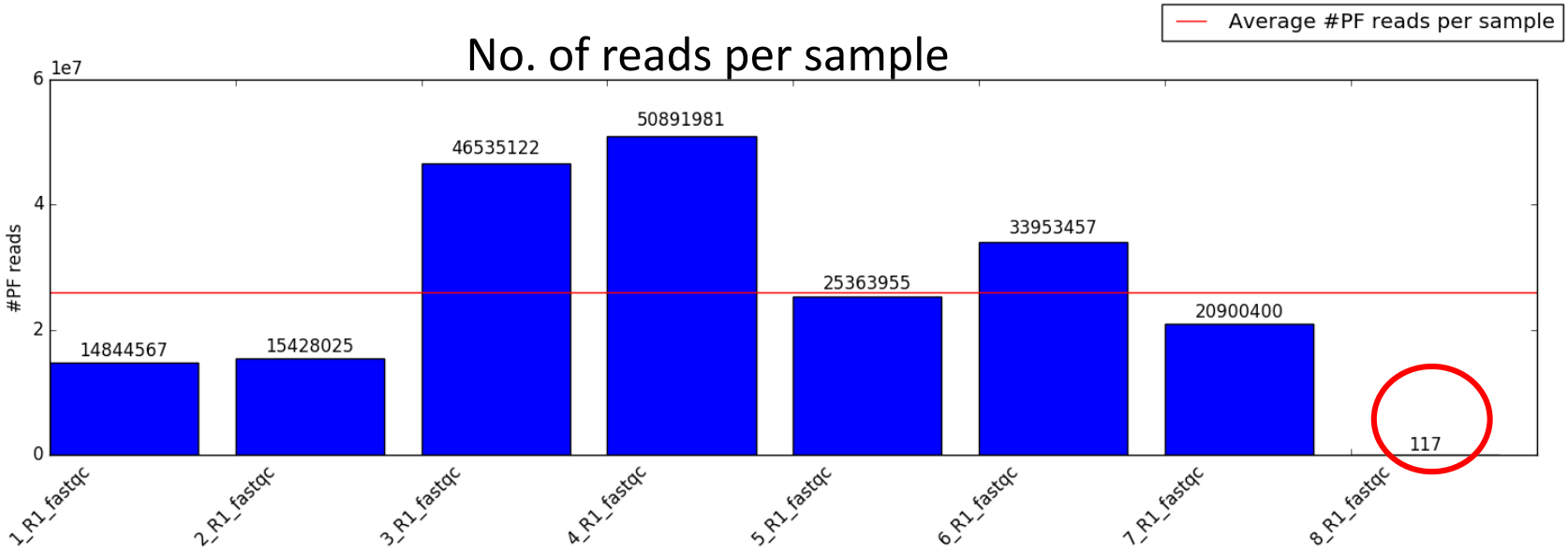
Download your data via WGET option on UNIX, or from your web browser.

2. A link to the **first QC report**



A sequencing run typically takes a few hours or more to complete, depending on the number of samples

Explore the first QC report



Flowcell Summary

Clusters (Raw)	Clusters (PF)	Yield (MBases)
269,921,898	262,607,862	19,696

Basic parameters per sample

Sample	Index	# PF Clusters	% Clusters per sample	Yield (MBases)	%≥Q30	FastQC Analysis
1	CAAGGCGA	14,844,567	5.65	1,113	99.07	R1
2	GACGCTAT	15,428,025	5.88	1,157	99.08	R1
3	ACTTCTTC	46,535,122	17.72	3,490	99.08	R1
4	CCTAGAAT	50,891,981	19.38	3,816	99.03	R1
5	TGGTAACG	25,363,955	9.66	1,902	99.13	R1
6	CATCAGAC	33,953,457	12.93	2,547	99.02	R1
7	GTGCGTAA	20,900,400	7.96	1,569	99.06	R1
8	CTATTCAA	117	0.00	0	97.44	R1
Undetermined Indices	Undetermined	54,690,238	20.82	4,102		

Explore the first QC report

General QC for run

Sequence protocol: **Single-read**

Quick Navigation

- Sequence quality
- #PF reads
- Flowcell Summary
- Basic parameters per sample

See here a more comprehensive report of MultiQC software

MultiQC

A modular tool to aggregate results from bioinformatics analyses across many samples into a single report.

Report generated on 2019-02-28, 16:17 based on data in: `/data/fastq/190219_NB501465_0472_AHT7WWBGX9/FastQC`

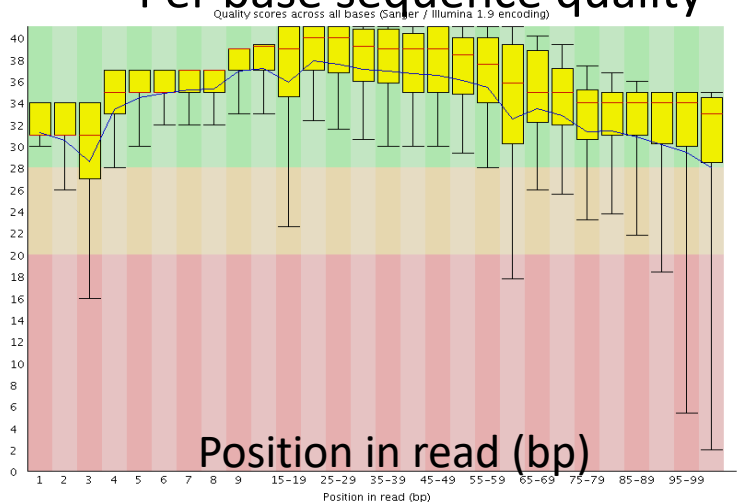
Welcome! Not sure where to start? Watch a tutorial video *(6:06)*

General Statistics

Copy table | Configure Columns | Plot | Showing 84/84 rows and 4/5 columns.

Sample Name	% Dups	% GC	Length	M Seqs
MB38-005_AllPrep_R1	9.0%	49%	75 bp	3.0
MB38-005_AllPrep_R2	97.8%	47%	15 bp	3.0
MB38-005_Rneasy_R1	5.4%	53%	75 bp	0.7
MB38-005_Rneasy_R2	91.7%	51%	15 bp	0.7

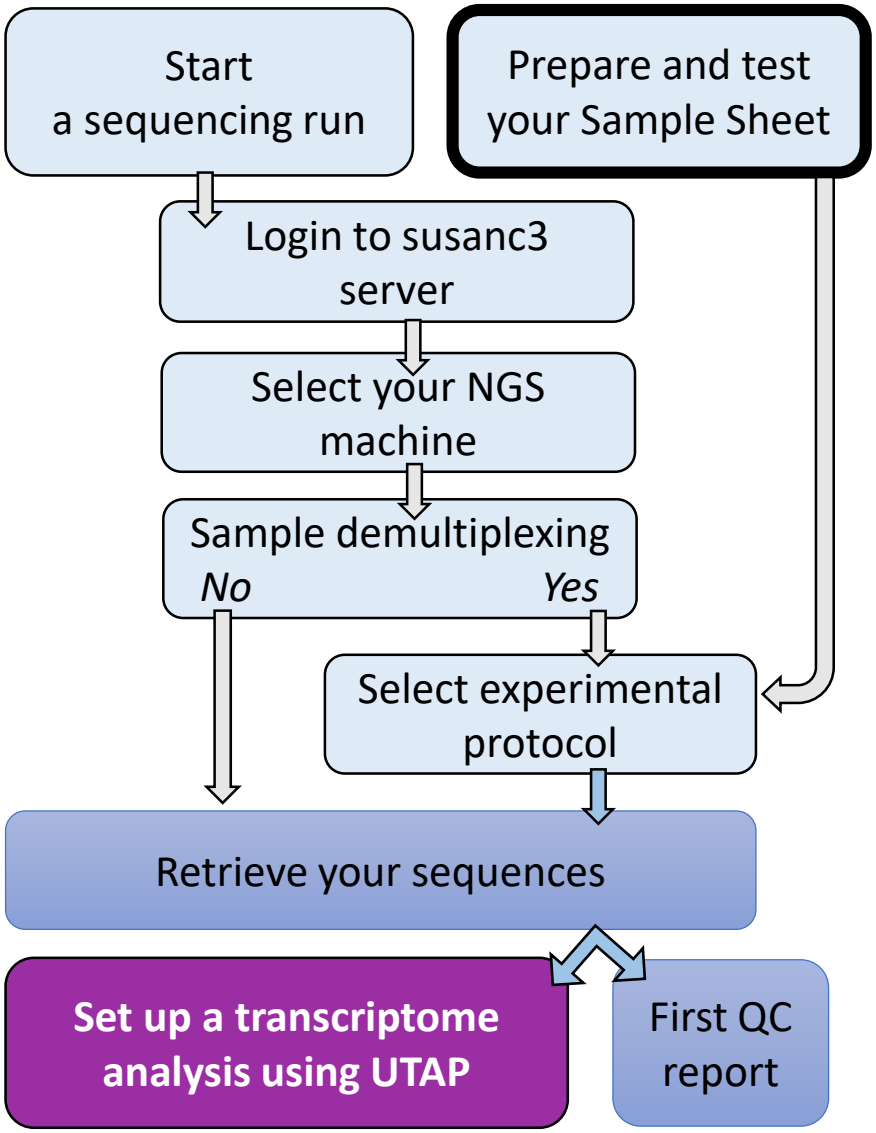
Per base sequence quality



First QC report

Demo: http://stefan.weizmann.ac.il/fqc/180708_NB551168_0156_AH2F77BGX7/

From sequencing to analyzed data



Detailed guidelines:

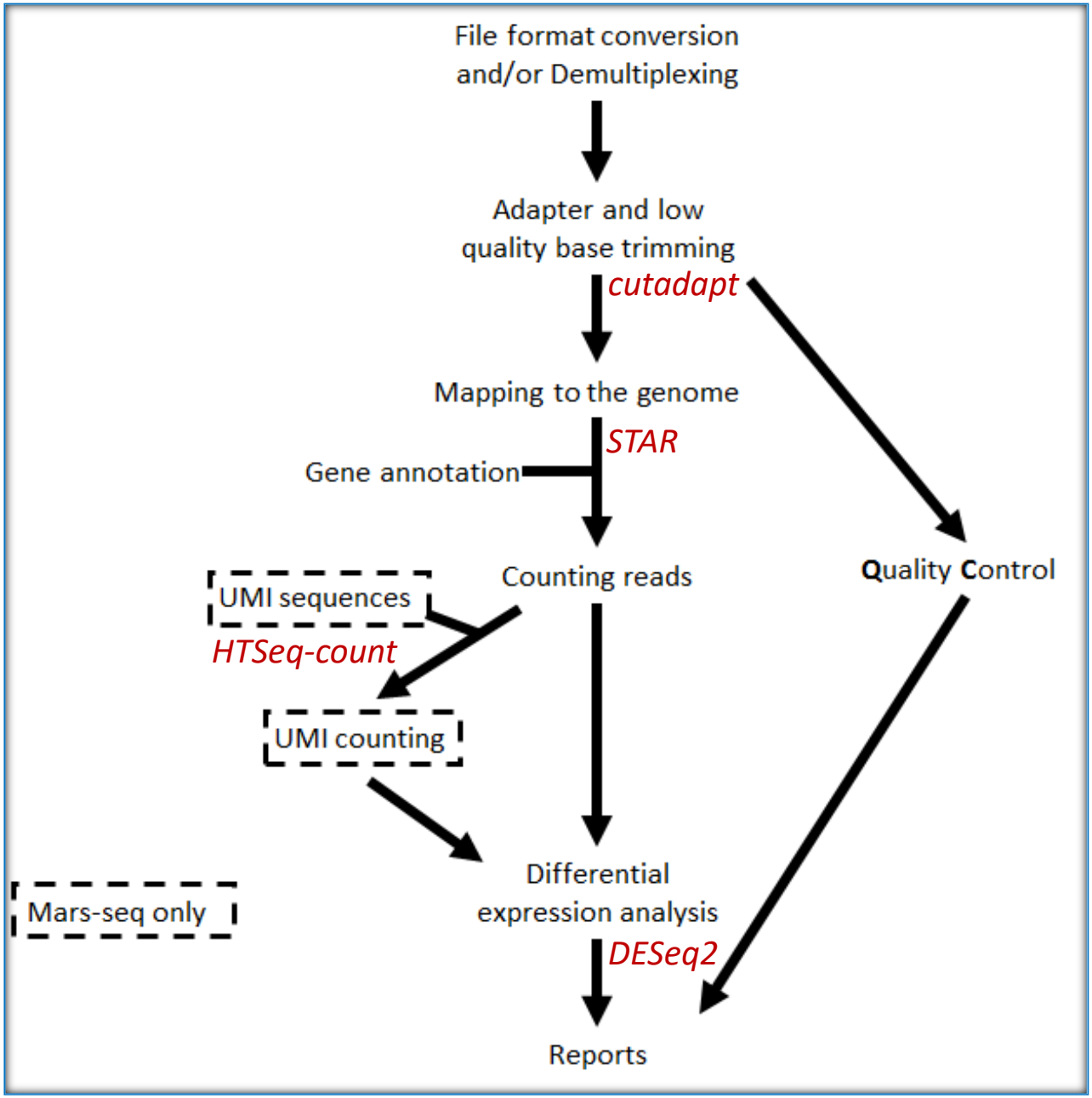
https://bbcunit.atlassian.net/wiki/spaces/BP/pages/104431617/From+samples+to+analyzed+NGS+data+UTAP

Part 2:
Setting up a transcriptome analysis with UTAP

UTAP: User-friendly Transcriptome Analysis Pipeline

Developed by the Bioinformatics unit (LSCF)

BMC Bioinformatics 2019



1. Setting up a new transcriptome analysis

Before you start- please prepare in advance:

1. An **account** (userID) on Wexac computer cluster
2. A "**Collaboration**" folder with read and write permissions for the Bioinformatics unit
3. Sufficient free **storage** space on Wexac (> 400Gb).

Transfer demultiplexed sequencing data (**fastq** files) to your Wexac Collaboration folder

- ***UTAP Pipeline website:*** ***http://ngsbio.wexac.weizmann.ac.il***

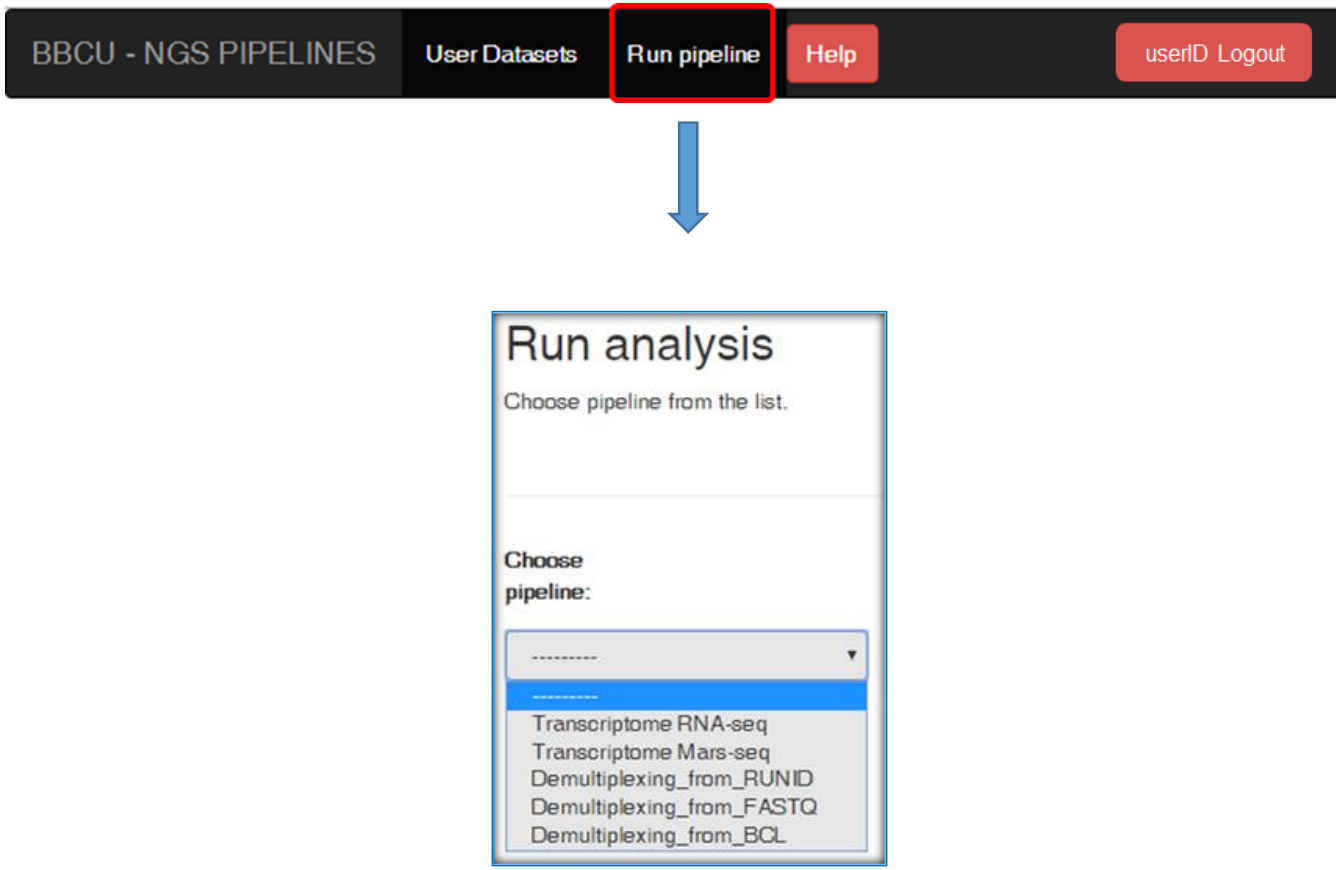
  OPEN WITH CHROME BROWSER

- UTAP Manual

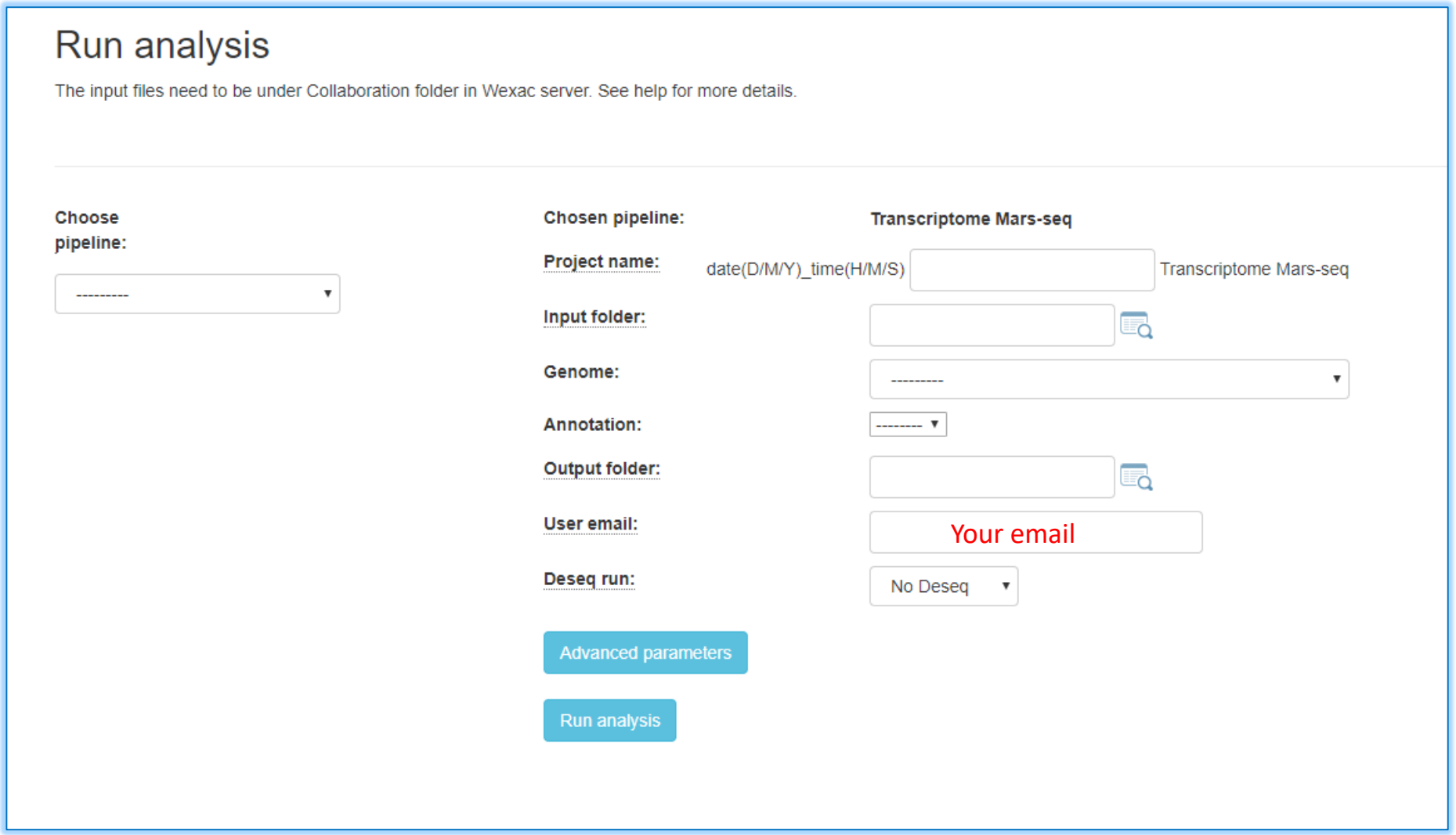
  https://bbcunit.atlassian.net/wiki/spaces/BP/pages/509214758/UTAP+guidelines-+User-friendly+Transcriptome+Analysis+Pipeline

Setting up a new transcriptome analysis

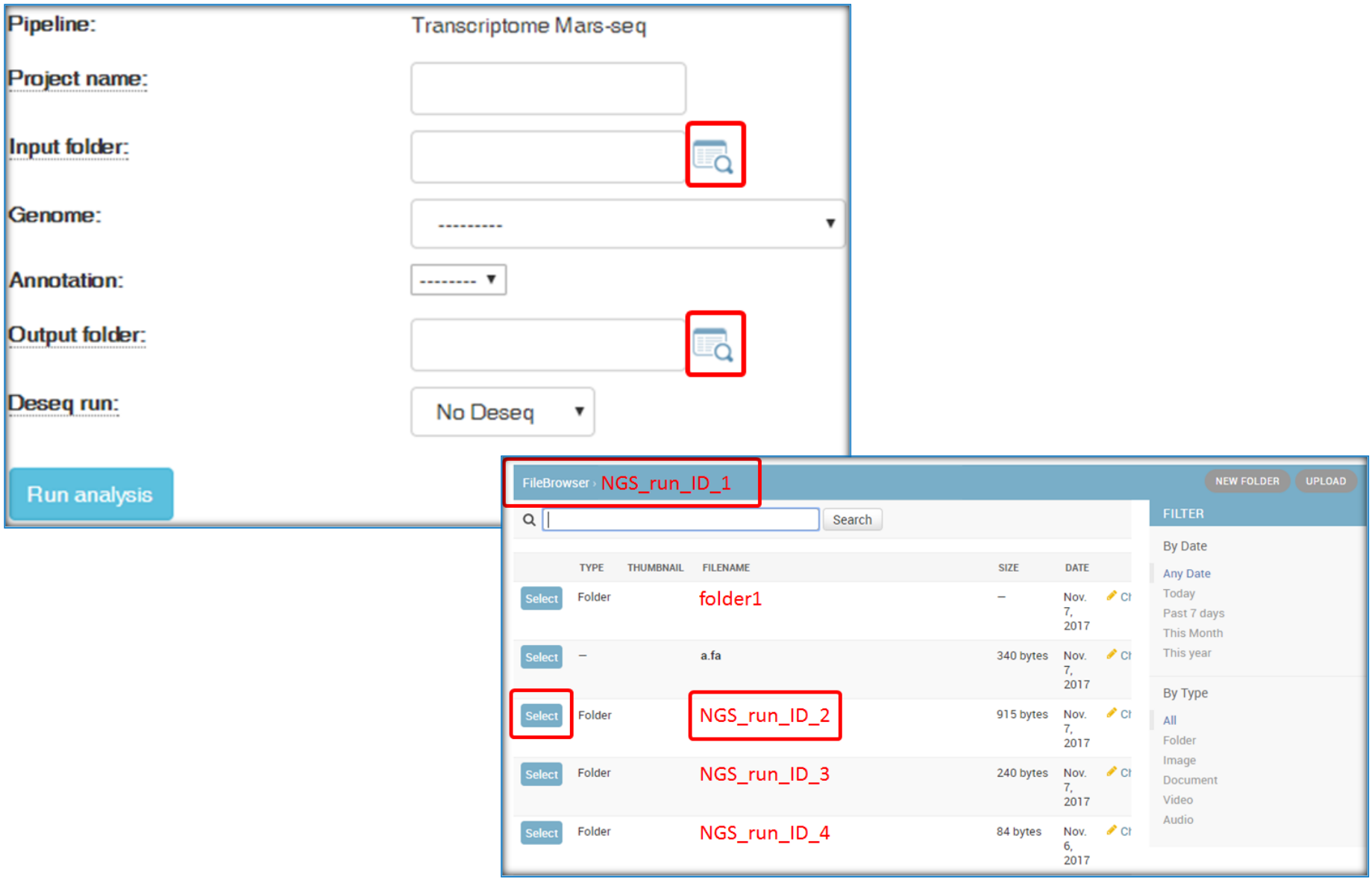
Select the type of analysis



Fill in a **project name**, select the **reference genome** and **annotation** for which the reads will be aligned to

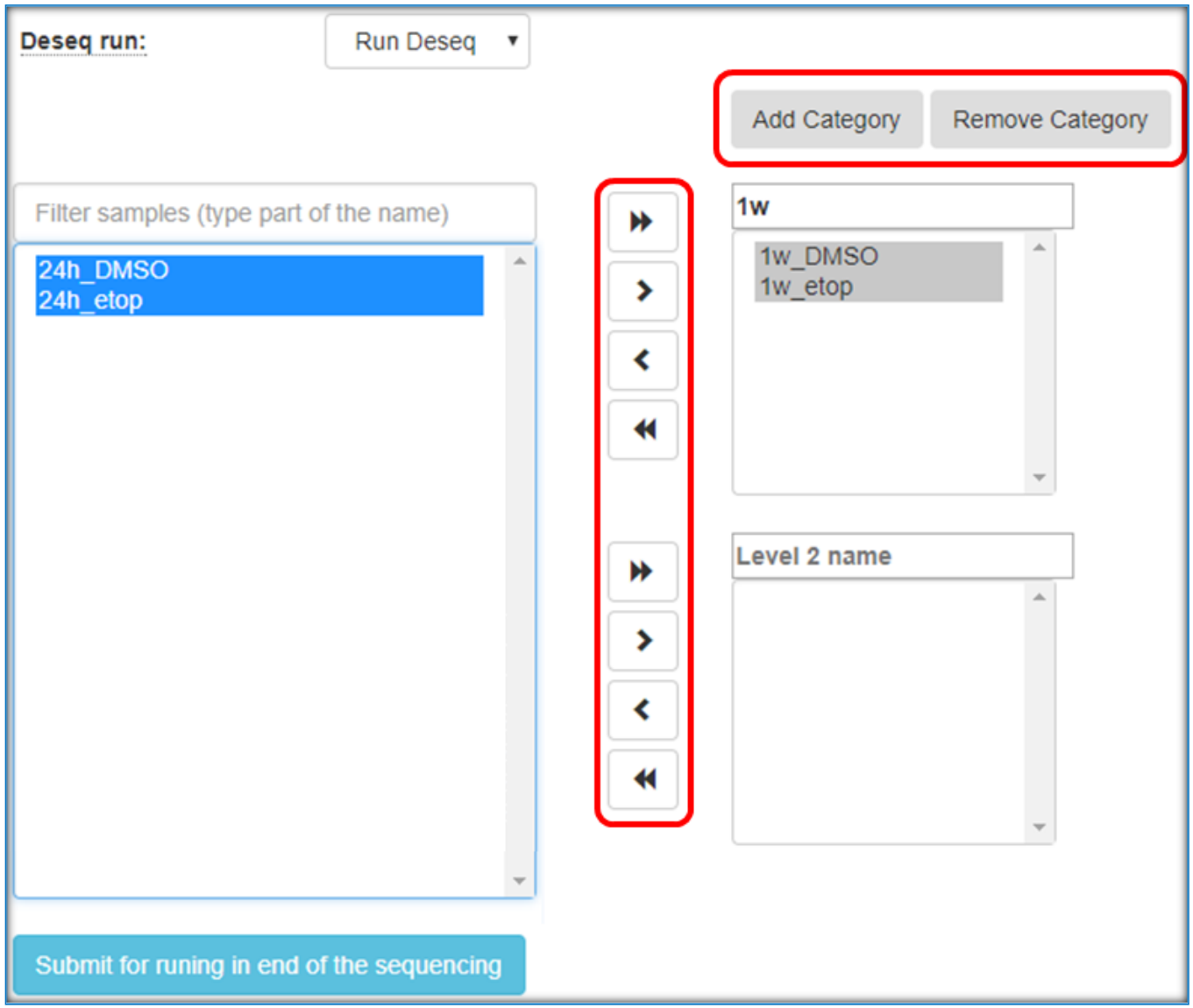


Select the input folder



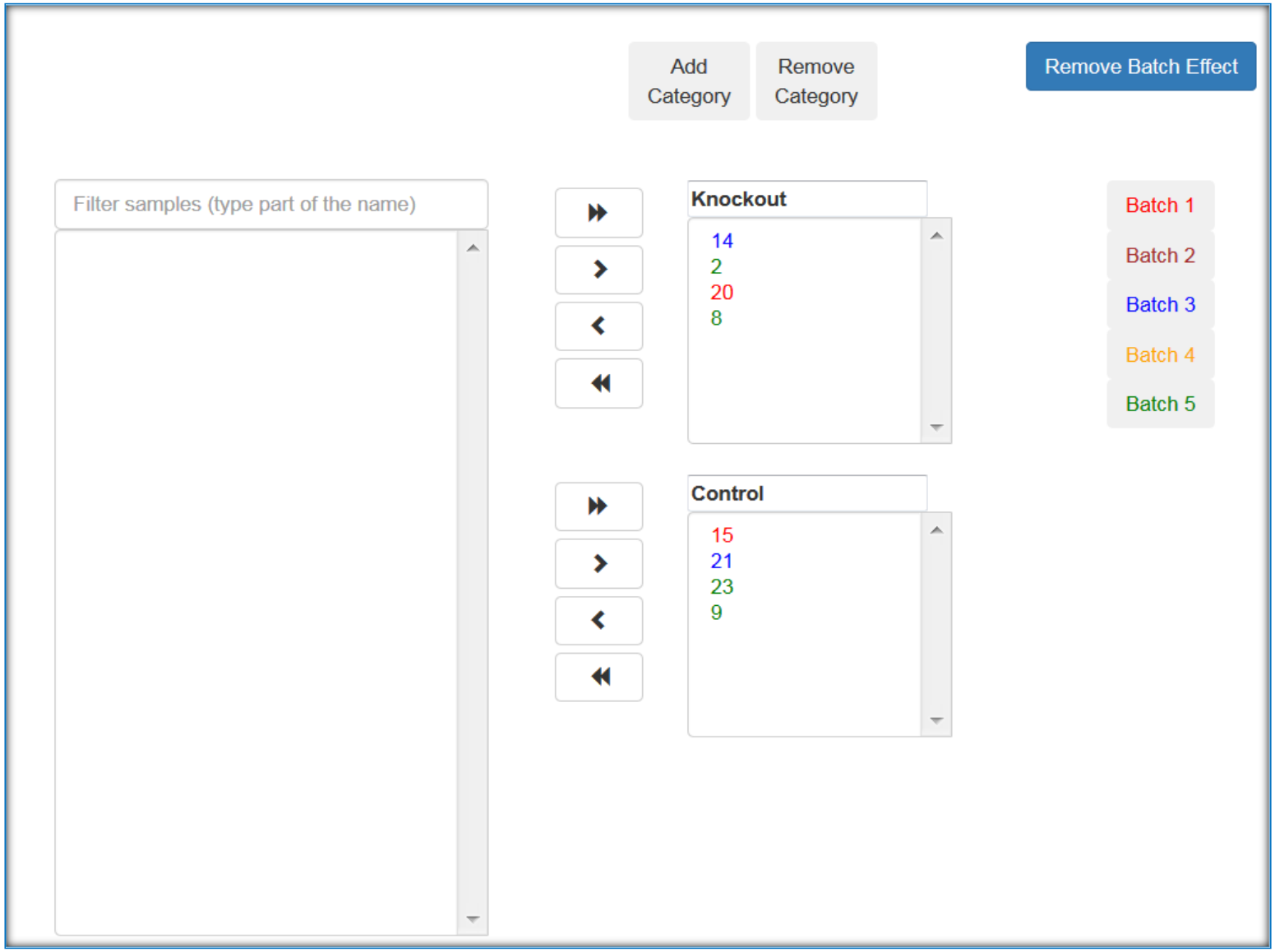
Differential gene expression analysis

with DESeq2 package

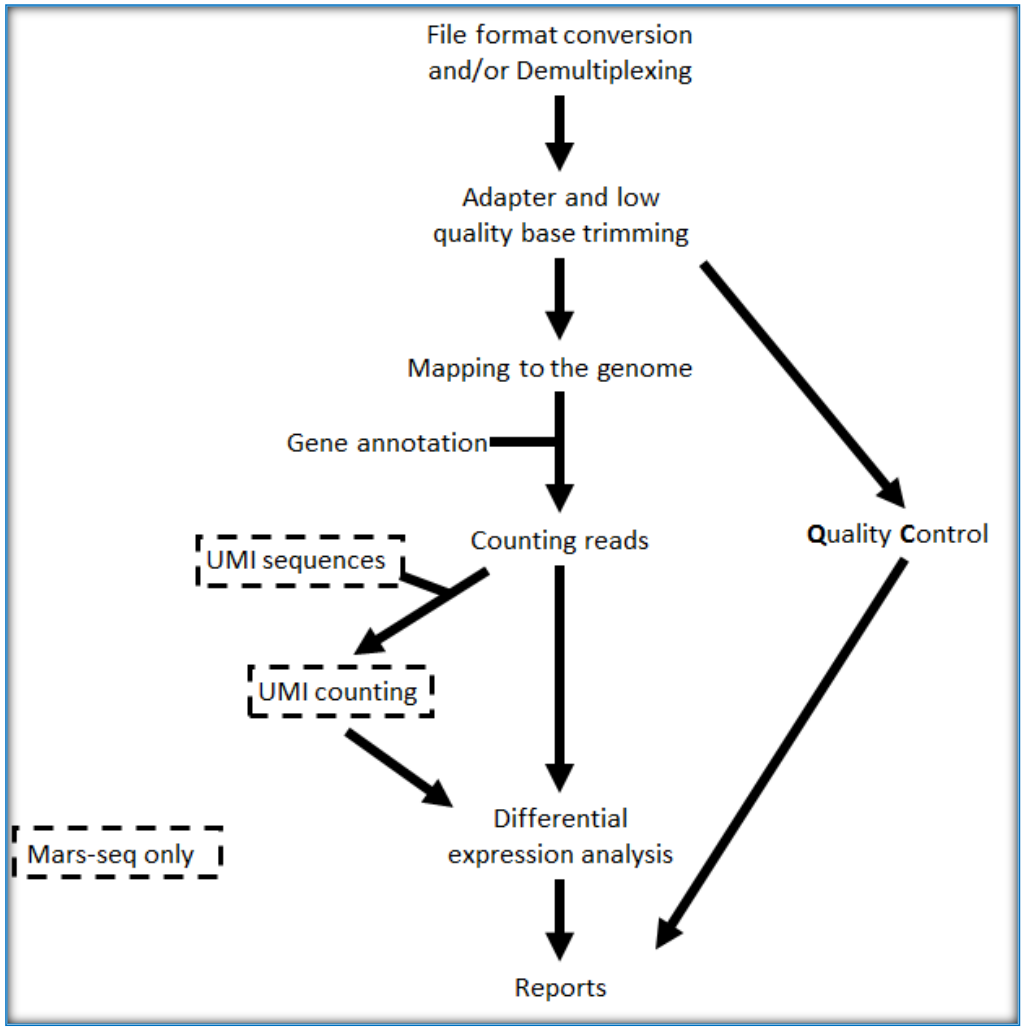


Create categories for the treatments that you would like to compare

Relate to batch effects



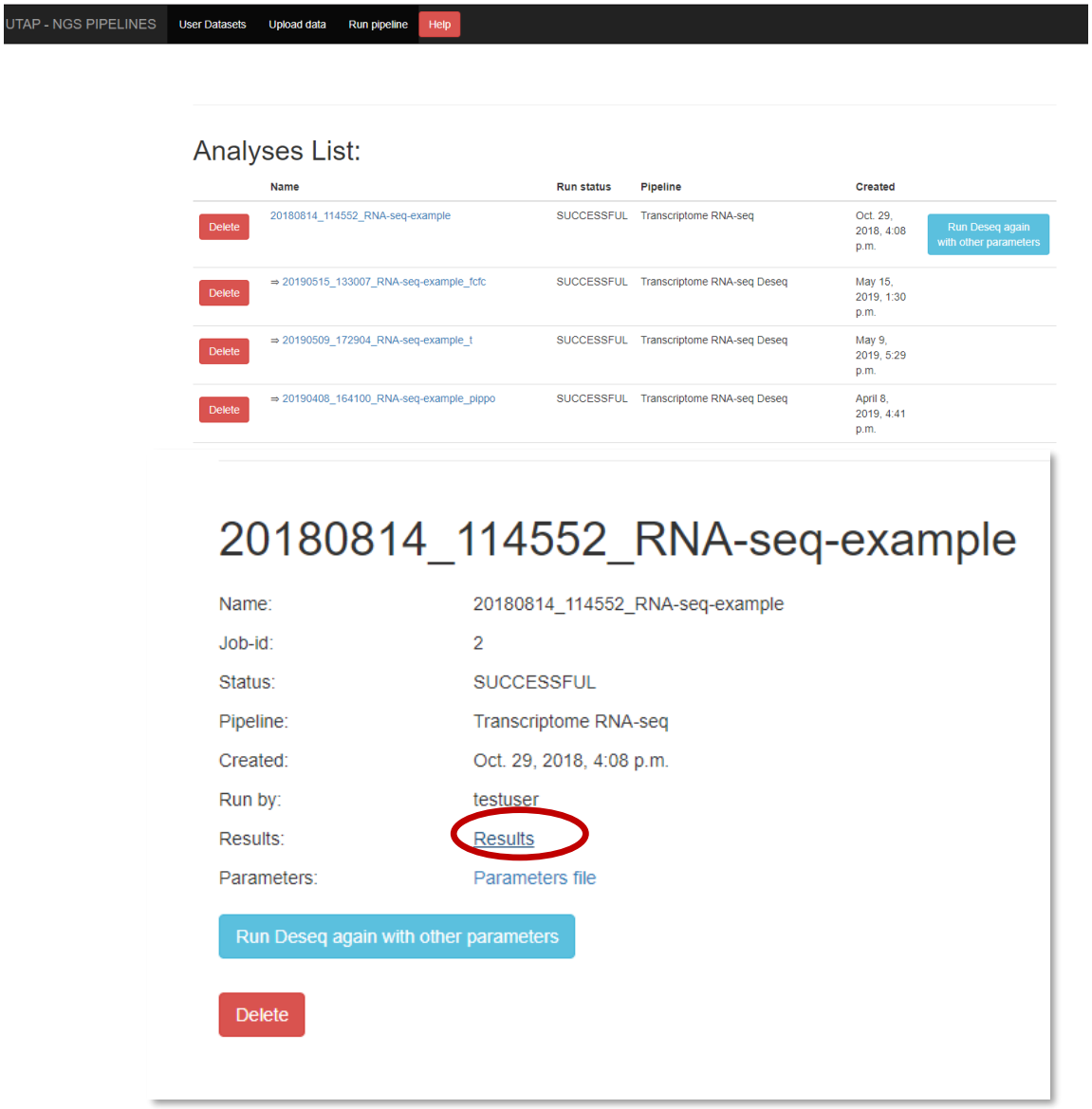
Run the pipeline



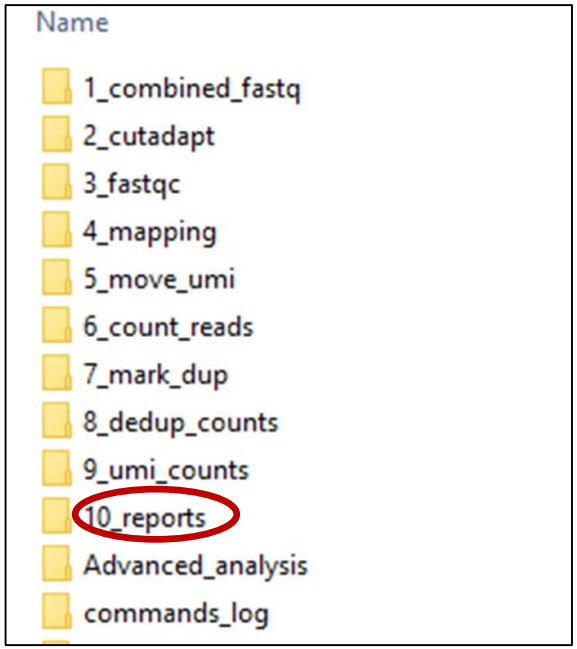
UTAP outputs are available at:

1. A link to the report sent by mail

2. UTAP pipeline website

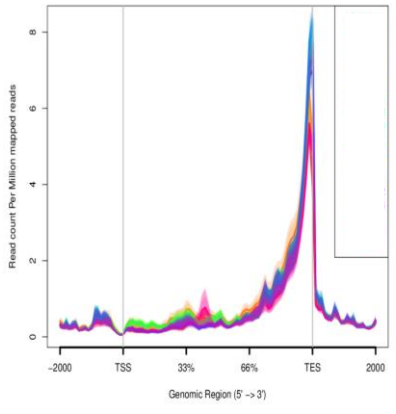


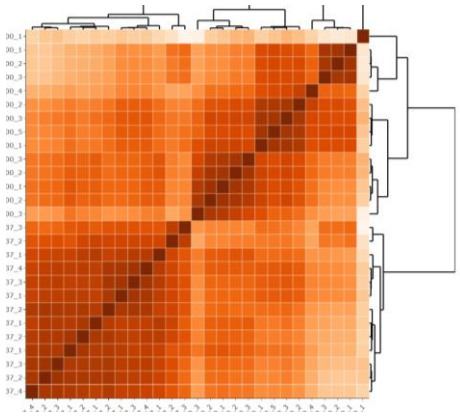
3. Collaboration folder on WEXAC:

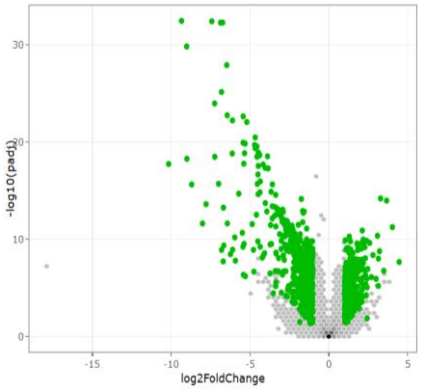


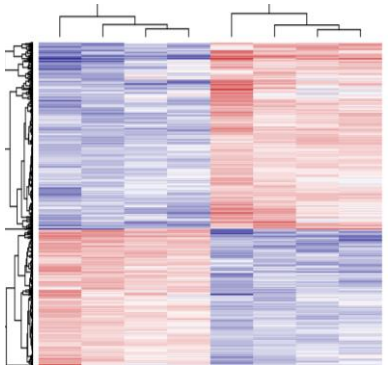
Part 3:
How to read the MARS-seq report

Please regard this analysis as a good starting point and not an end result!





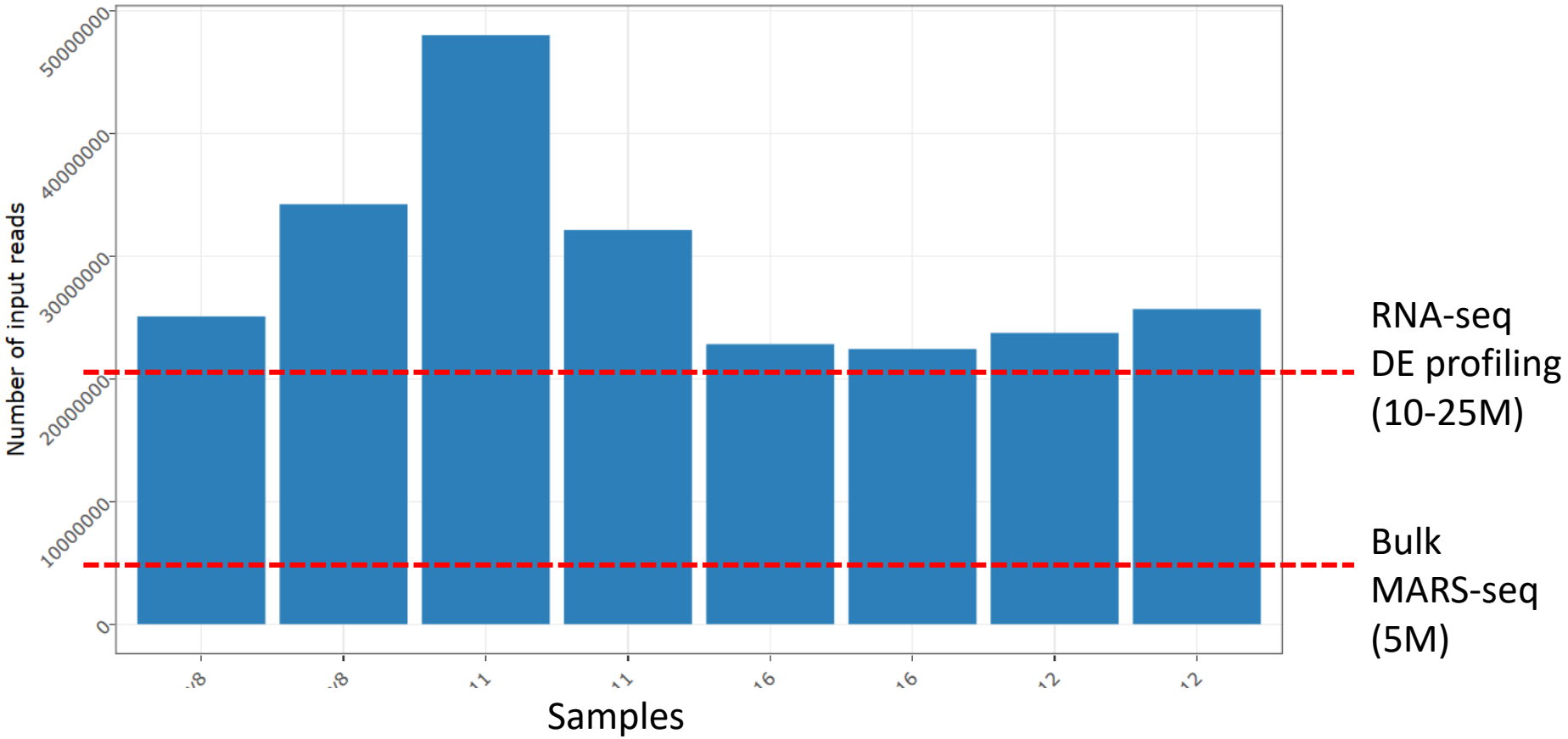




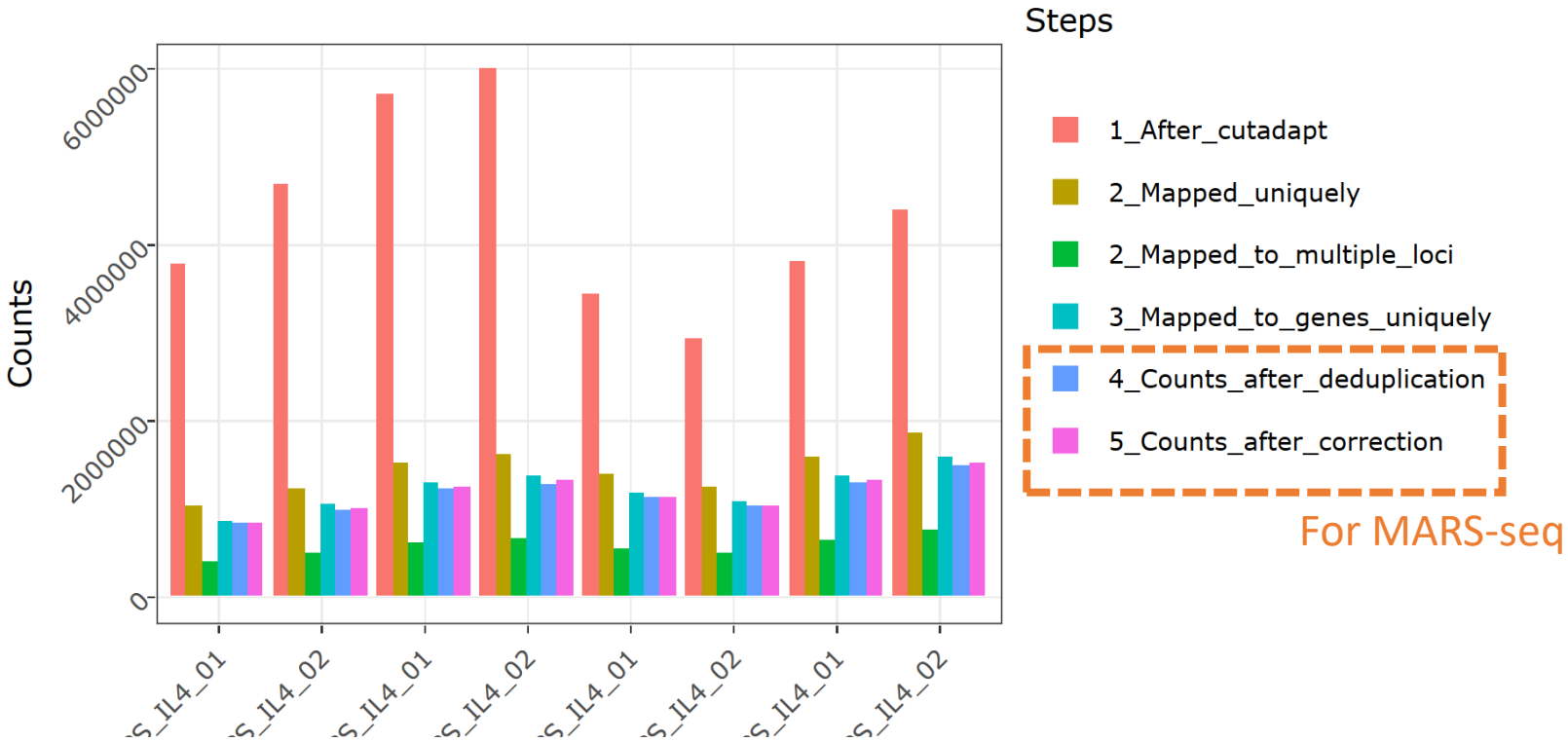
MARS-seq example report:
https://bip.weizmann.ac.il/mars-seq

Total number of reads for each sample in raw data

There is **always** an unequal distribution of reads per samples on a flowcell



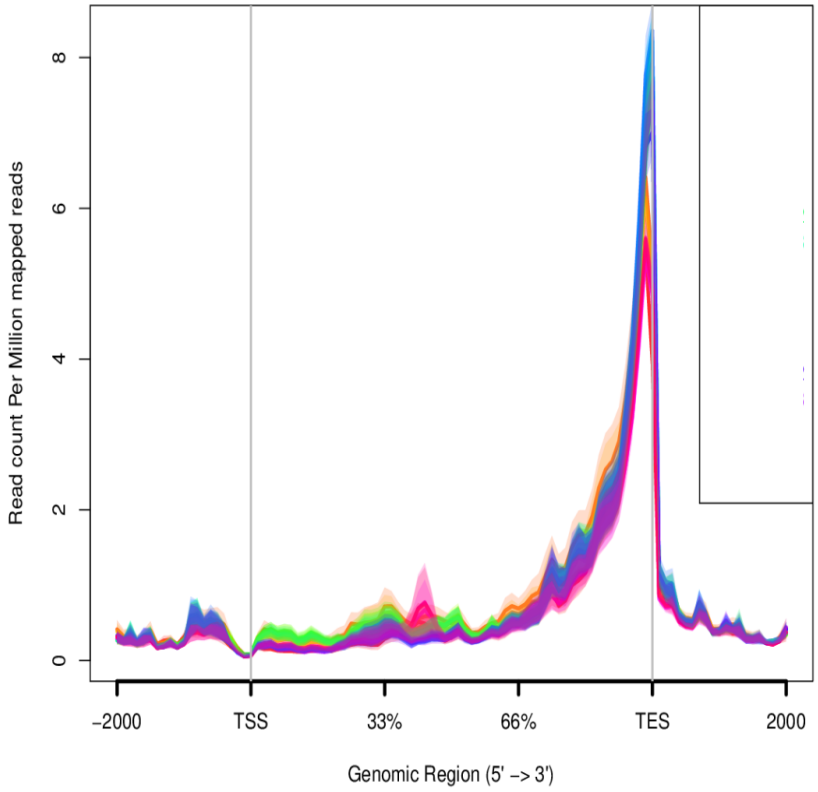
Summary of the number of reads for each sample in each step of the pipeline



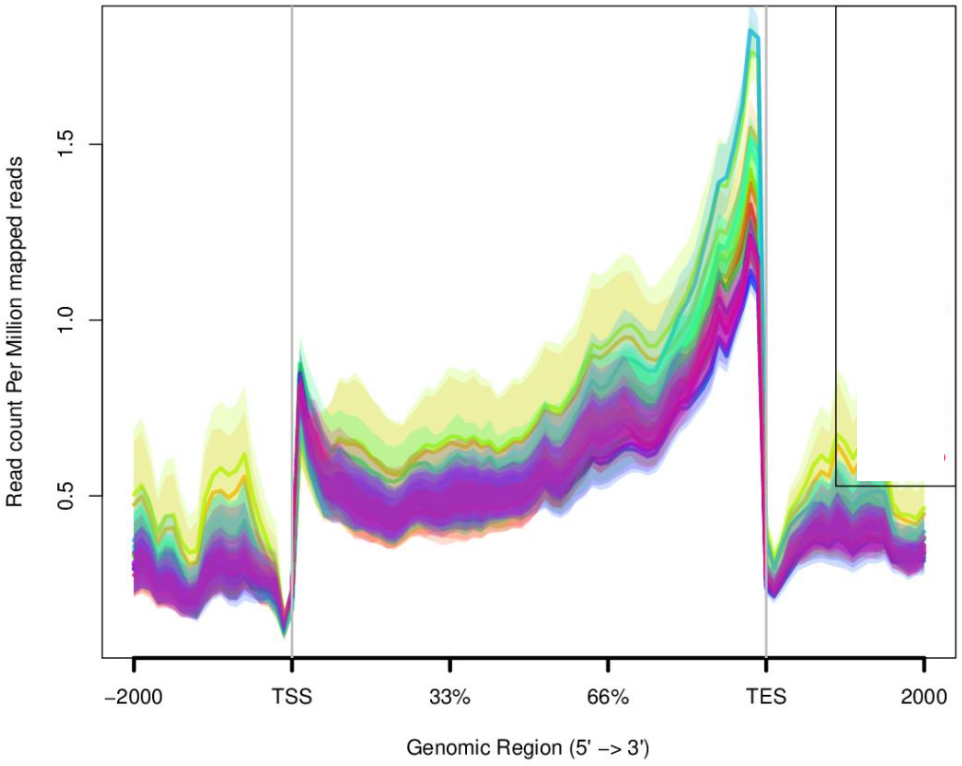
For MARS-seq

Genomics regions to which the reads (raw data) are mapped

MARS-seq



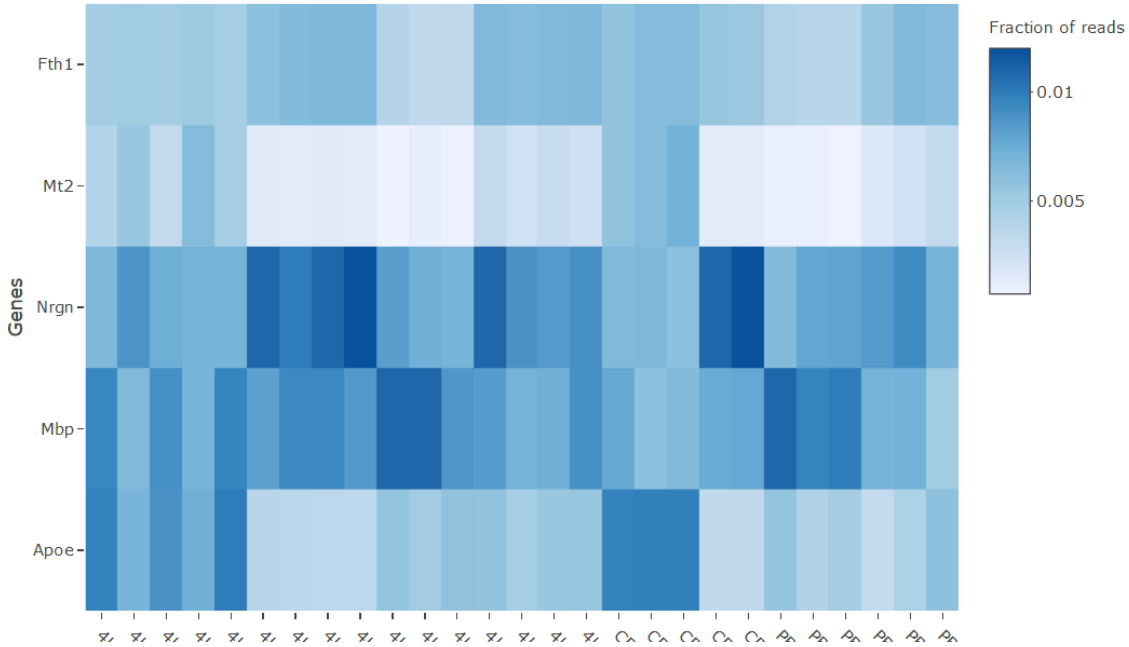
RNA-seq



Top highly-expressed genes

(above 5% of total expression)

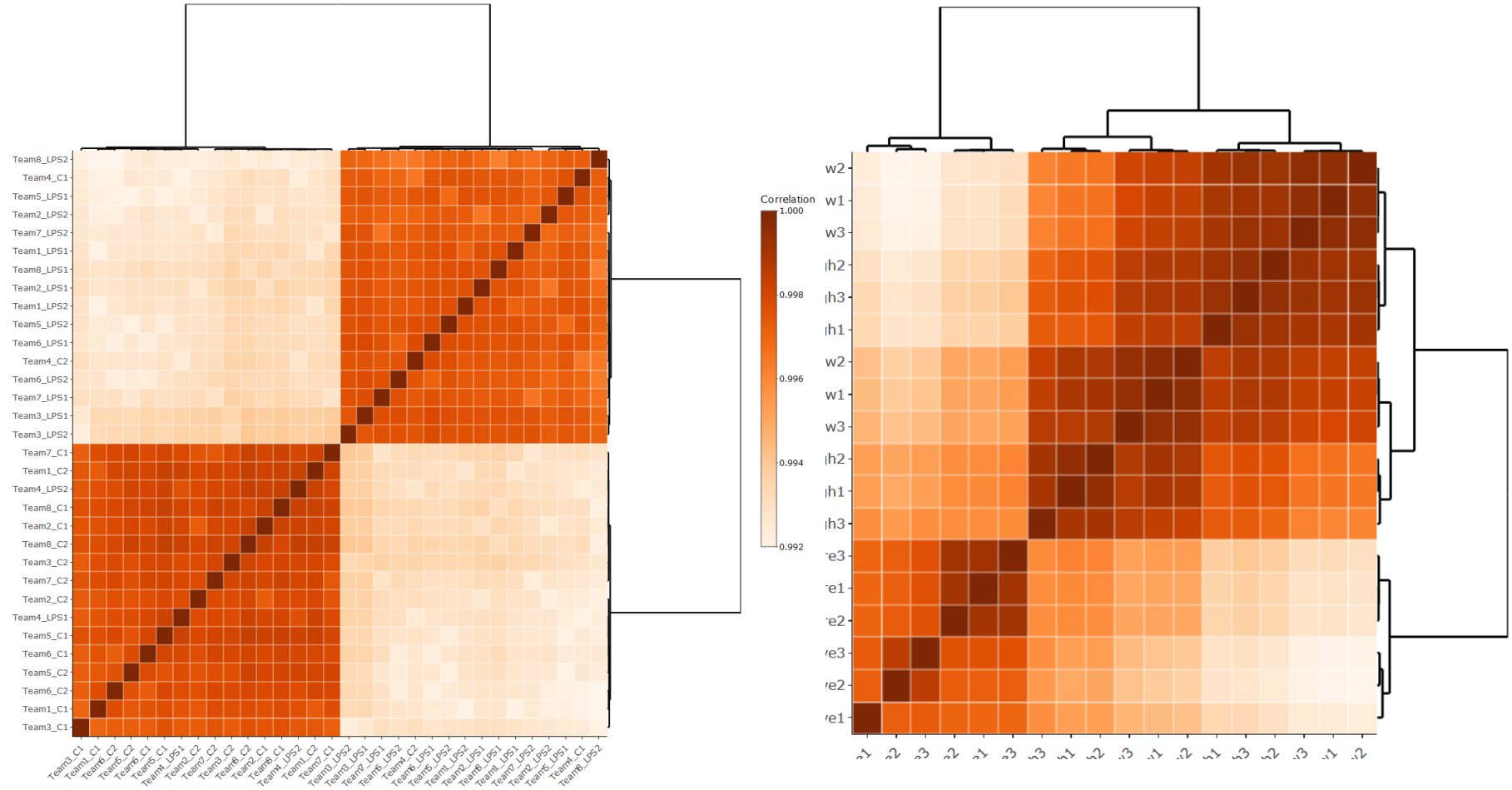
The fraction of reads from the genes with the most counts



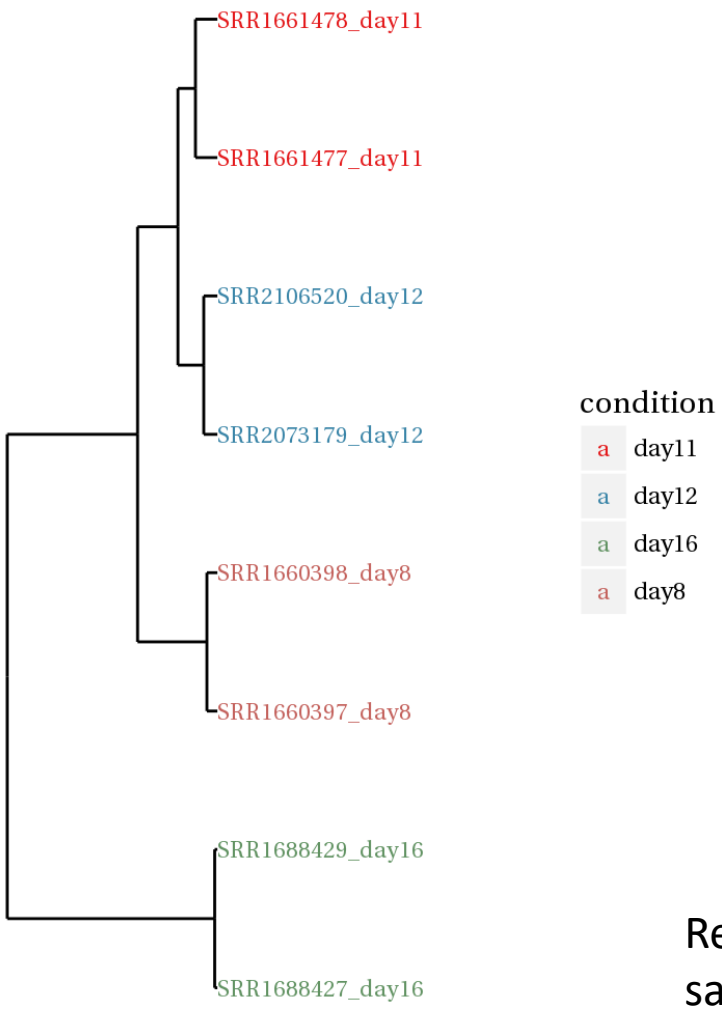
very highly expressed genes can dominate the total lane count and skew the expression analysis

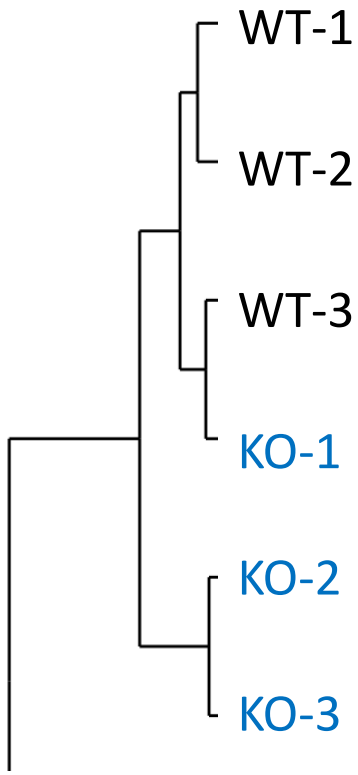
Explore samples Correlation

Heatmap of Pearson correlation coefficients between gene expression values of each sample



Explore samples Correlation in a dendrogram

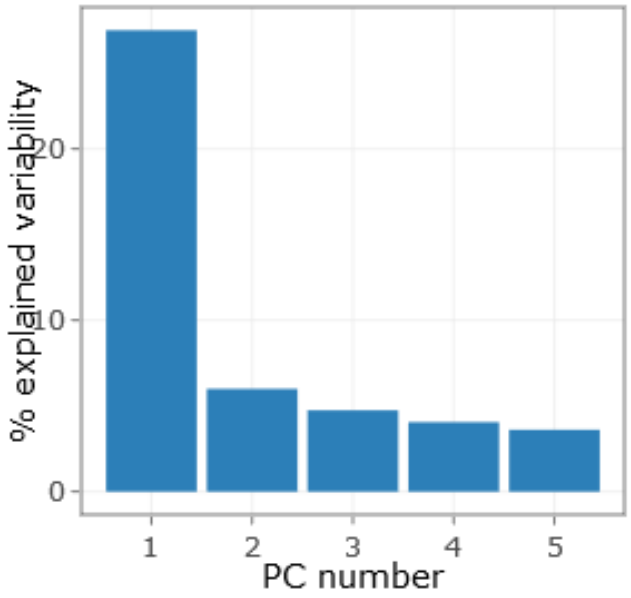


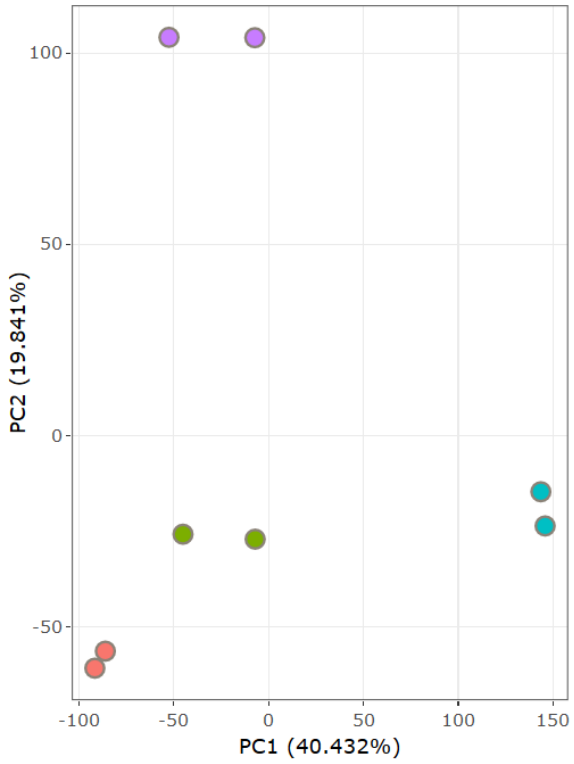


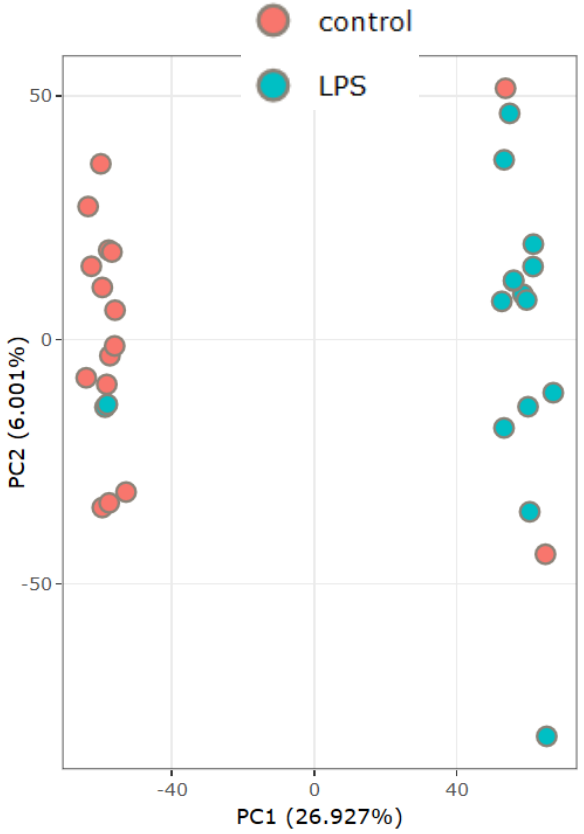
Remember - write meaningful names for your samples!

PCA- Principal component analysis

Dimensionality reduction to assess overall similarity between samples







Differential Expression Analysis

Differential expression analysis is performed using DESeq2.

Thresholds for **significant differential expressed genes** for each comparison:

- $|\log_2 \text{FoldChange}| >= 1$
- padj <= 0.05
- baseMean >= 5

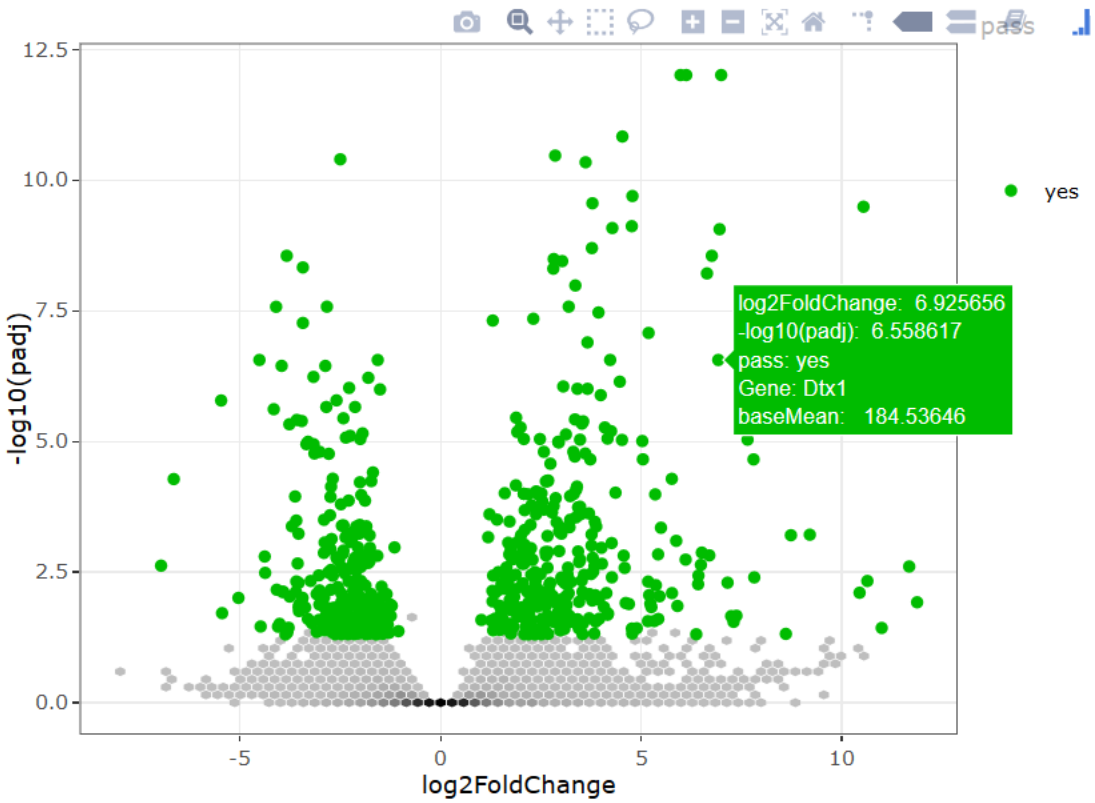
Comparison	Factor	A	B	Formula	Padj corrected by fdrtool	Plots	DE Genes
control_vs_LPS	control_or_LPS	control	LPS	control_or_LPS	FALSE	link	link

Comparison wt-rest_vs_dko-rest

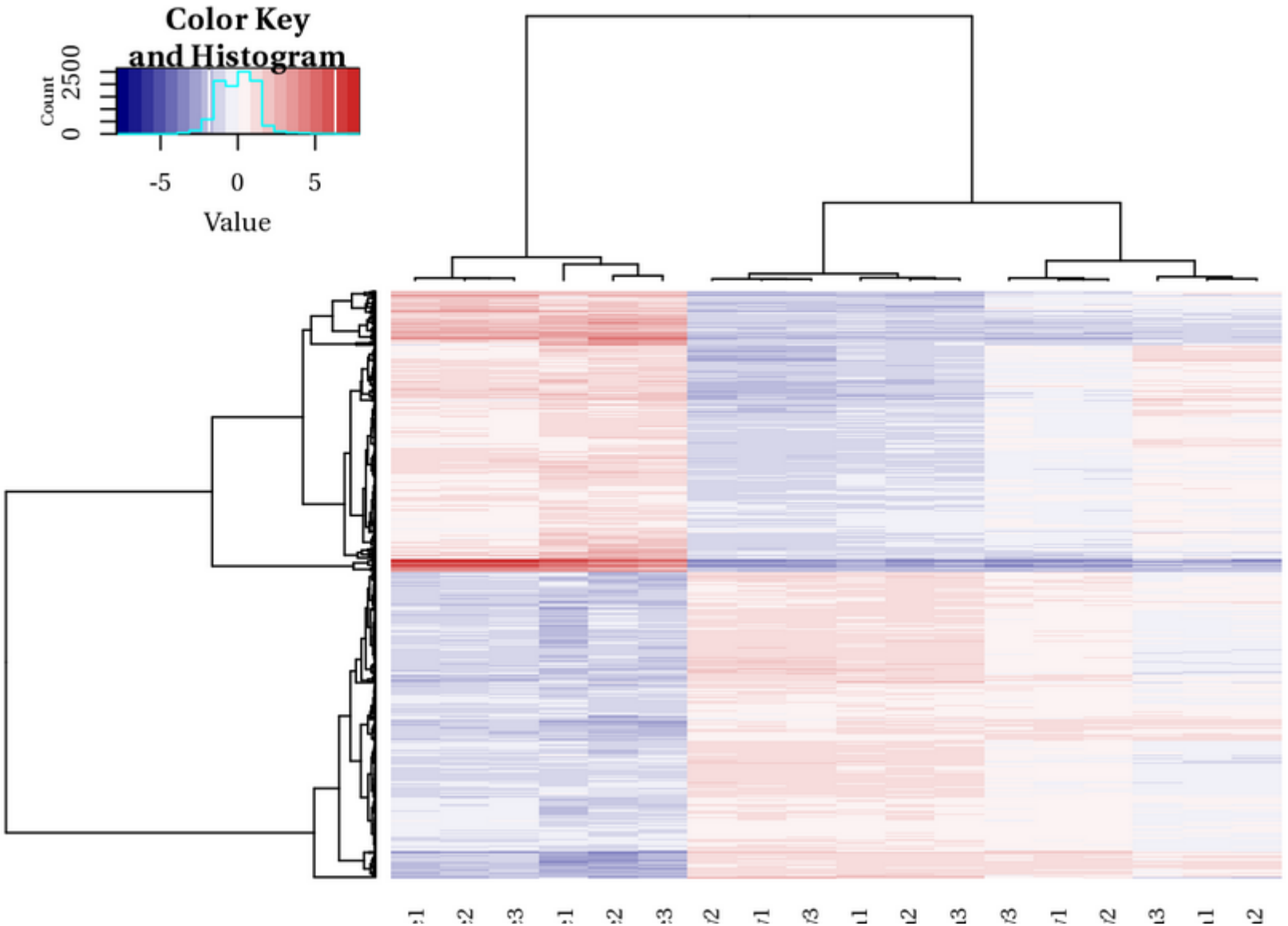
Show 20 entries Search:

	Comparison	Gene	baseMean	log2FoldChange	linearFoldChange	pvalue	padj	pass	Direction	Plot
	All		All	All	All					
1	wt-rest_vs_dko-rest	BC018473	77.8	-8.042	-263.562261359103	2.1e-13	9.6e-13	yes	down	
2	wt-rest_vs_dko-rest	Icam1	200	4.165	17.9386572484561	4e-23	9.6e-13	yes	up	
3	wt-rest_vs_dko-rest	St6galnac2	64.2	-5.115	-34.6552014568296	3.5e-17	9.6e-13	yes	down	
4	wt-rest_vs_dko-rest	Ddx3y	61.2	-3.699	-12.9870332956365	1e-11	5.4e-11	yes	down	
5	wt-rest_vs_dko-rest	Xist	49	4.404	21.1707430385299	4.2e-9	8.7e-8	yes	up	

Volcano plots
An interactive scatter plot of significance versus fold-change



Hierarchical clustering heat map of differentially expressed genes
using the genes expression values rld (log2 normalized)



Links to functional enrichments analysis

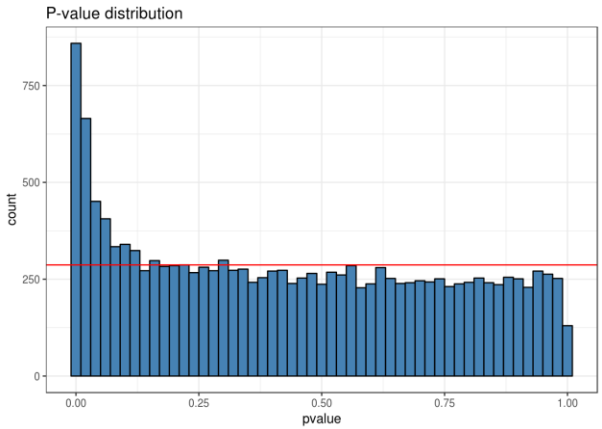
To perform functional enrichments, you can try one or more of the following websites: Intermine, Reactome, GeneAnalytics from GeneCards(R) or STRING. You can also use the links below to send the differentially expressed genes directly to Intermine (**In the first time click on the button twice to get the correct page.**:

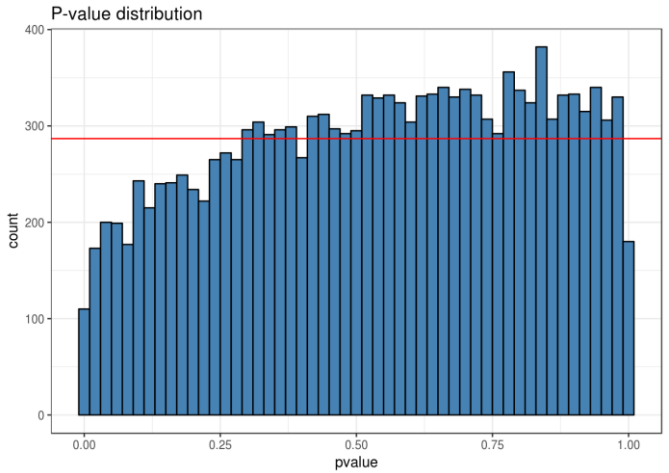
Show 10 **entries** **Search:**

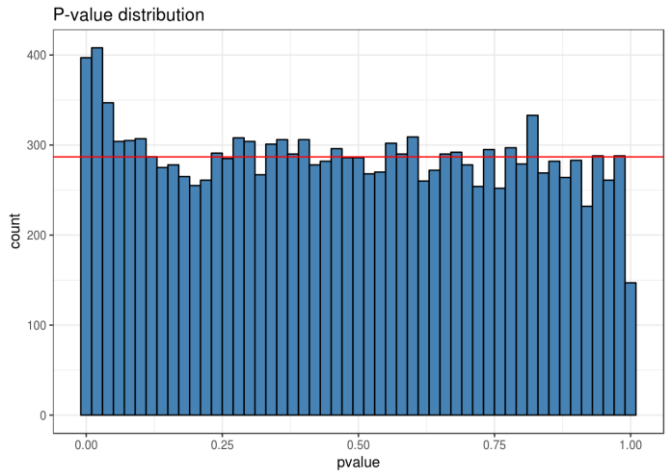
	Comparison	pass	Direction	Number of genes	InterMine
1	control_vs_LPS	yes	up	449	Send
2	control_vs_LPS	yes	down	596	Send

A few words on interpretation of p Value histograms

Comparison	Factor	A	B	Formula	Padj corrected by fdrtool	Plots	DE Genes
control_vs_LPS	control_or_LPS	control	LPS	control_or_LPS	FALSE	link	link





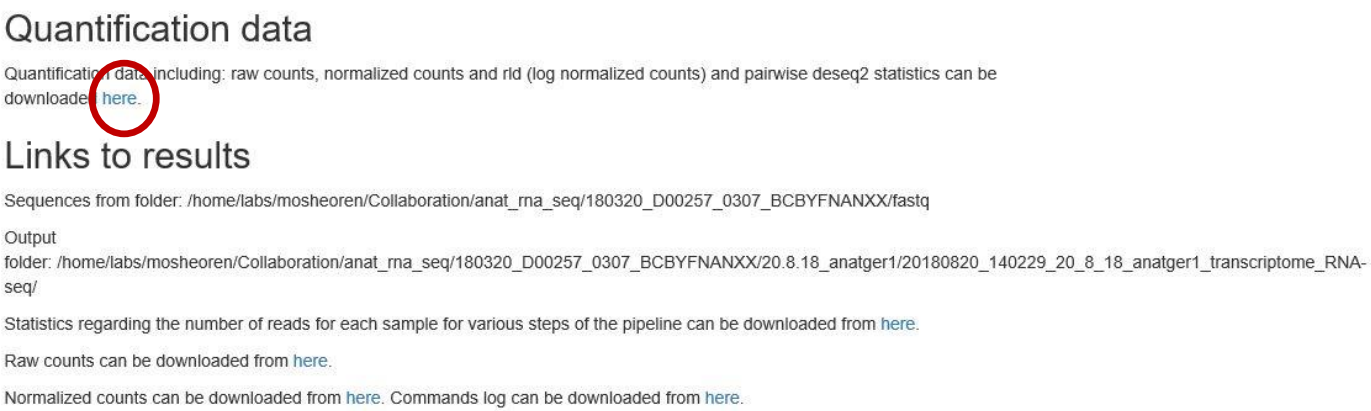


If you are not sure –
Stop whatever you're doing, and find a statistician!

Additional documentations in the report

- **Bioinformatics pipeline methods**
- **Links to results**
- **Quantification of data**

See excel



Quantification data

Quantification data including: raw counts, normalized counts and rld (log normalized counts) and pairwise deseq2 statistics can be downloaded here.

Links to results

Sequences from folder: /home/labs/mosheoren/Collaboration/anat_rna_seq/180320_D00257_0307_BCBYFNANXX/fastq

Output folder: /home/labs/mosheoren/Collaboration/anat_rna_seq/180320_D00257_0307_BCBYFNANXX/20.8.18_anatger1/20180820_140229_20_8_18_anatger1_transcriptome_RNA-seq/

Statistics regarding the number of reads for each sample for various steps of the pipeline can be downloaded from here.

Raw counts can be downloaded from here.

Normalized counts can be downloaded from here. Commands log can be downloaded from here.

Results file content

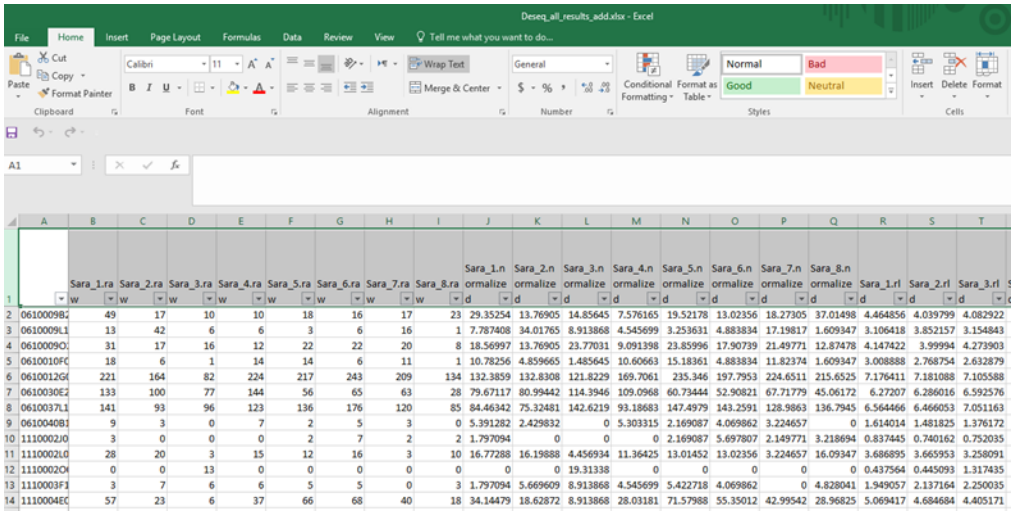
- For each sample, 3 columns:
  - Raw – read count
  - Normalized – normalized read counts to the sample library size
  - Rld – transformed normalized read counts

Count values for a gene can be zero in some conditions (and non-zero in others).

For other downstream analyses – e.g. for visualization or clustering – it is useful to work with transformed versions of the count data:

$$y=\log_2(N+C)$$

Where:
N represents the count values
C is a positive constant.



	A	Sara_1.raw	Sara_2.raw	Sara_3.raw	Sara_4.raw	Sara_5.raw	Sara_6.raw	Sara_7.raw	Sara_8.raw	Sara_1.normalized	Sara_2.normalized	Sara_3.normalized	Sara_4.normalized	Sara_5.normalized	Sara_6.normalized	Sara_7.normalized	Sara_8.normalized	Sara_1.rld	Sara_2.rld	Sara_3.rld
2	0610009B2	49	17	10	10	18	16	17	23	29.35254	13.76905	14.85645	7.576165	19.52178	13.02356	18.27305	37.01498	4.464856	4.039799	4.082922
3	0610009L1	13	42	6	6	3	6	16	1	7.787408	34.01765	8.913868	4.545699	3.253631	4.883834	17.19817	1.609347	3.106418	3.852157	3.154843
4	0610009O1	31	17	16	12	22	22	20	8	18.56997	13.76905	23.77031	9.091398	23.85996	17.90739	21.49771	12.87478	4.147422	3.99994	4.273903
5	0610010F0	18	6	1	14	14	6	11	1	10.78256	4.859665	1.485645	10.60663	15.18361	4.883834	11.82374	1.609347	3.008888	2.768754	2.632879
6	0610012G0	221	164	82	224	217	243	209	134	132.3859	132.8308	121.8229	169.7061	235.346	197.7953	224.6511	215.6525	7.176411	7.181088	7.105588
7	0610030E2	133	100	77	144	56	65	63	28	79.67117	80.99442	114.3946	109.0968	60.73444	52.90821	67.71779	45.06172	6.27207	6.286016	6.592576
8	0610037L1	141	93	96	123	136	176	120	85	84.46342	75.32481	142.6219	93.18683	147.4979	143.2591	128.9863	136.7945	6.564466	6.466053	7.051163
9	0610040B1	9	3	0	7	2	5	3	0	5.391282	2.429832	0	5.303315	2.169087	4.069862	3.224657	0	1.614014	1.481825	1.376172
10	1110002J0	3	0	0	0	2	7	2	2	1.797094	0	0	0	2.169087	5.697807	2.149771	3.218694	0.837445	0.740162	0.752035
11	1110002L0	28	20	3	15	12	16	3	10	16.77288	16.19888	4.456934	11.36425	13.01452	13.02356	3.224657	16.09347	3.686895	3.665953	3.258091
12	1110002O0	0	0	13	0	0	0	0	0	0	0	19.31338	0	0	0	0	0	0.437564	0.445093	1.317435
13	1110003F1	3	7	6	6	5	5	0	3	1.797094	5.669609	8.913868	4.545699	5.422718	4.069862	0	4.828041	1.949057	2.137164	2.250035
14	1110004E0	57	23	6	37	66	68	40	18	34.14479	18.62872	8.913868	28.03181	71.57988	55.35012	42.99542	28.96825	5.069417	4.684684	4.405171

Results file: statistics **for each comparison**

Control_Treatment.baseMean	Control_Treatment.log2FoldChange	Control_Treatment.pvalue	Control_Treatment.padj	Control_Treatment.pass	Control_Treatment.Direction

- Control_Treatment.baseMean – mean (average) normalized read counts of all samples
- Control_Treatment.log2FoldChange – Fold change is a measure of the ratio of means of two populations (i.e control and treatment). Log2(2)=1
- Control_Treatment.pvalue – measures statistical significance of the difference of the two populations.

  If you have, 10,000 genes, then

  you expect 500 of them to have $p<0.05$ just by chance.
- Control_Treatment.padj - statistically significant as part of the multiple comparison testing (many genes).
- Control_Treatment.pass – Yes or No –

  passing threshold: baseMean >= 5, $|\text{log2FoldChange}| \geq 1$ and $\text{padj} \leq 0.05$
- Control_Treatment.Direction – Up or Down

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Bioinformatics unit
NGS data analysis


Noa Wigoda


Ester Feldmesser


Dena Leshkowitz
Unit head


Bareket Dassa


Gil Stelzer

Contact us:
Dena.Leshkowitz@weizmann.ac.il

Support for other issues:

Library preparation:
Hadas Keren-Shaul (hadas.keren-shaul@weizmann.ac.il)

SampleSheet:
Merav Kedmi (merav.kedmi@weizmann.ac.il)
David Pilzer (david.pilzer@weizmann.ac.il)

NextSeq:
Levine Muriel Chemla (muriel.chemla@weizmann.ac.il)
INCPM Dana Robbins (dana.robbins@weizmann.ac.il)

Registration and access to susanc:
Irit Orr (irit.orr@weizmann.ac.il)

Downloading data:
Vitaly Golodnitsky (vitaly.golodnitsky@weizmann.ac.il)

Learn more with our:

Workshop: Introducing UTAP: User-friendly Transcriptome Analysis Pipeline **(July 6th 2020)**

Course: An Introduction to deep-sequencing analysis for biologists 20203331

e-learning tool (to be released)