מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

LIFE SCIENCE
CORE FACILITIES

Introducing UTAP: User-friendly Transcriptome Analysis Pipeline

EXCEL Tips for Deseq_all_results.txt

Noa Wigoda
Bioinformatics Unit, LSCF

EXCEL Tips – for Deseq_all_results.txt

From the report:

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

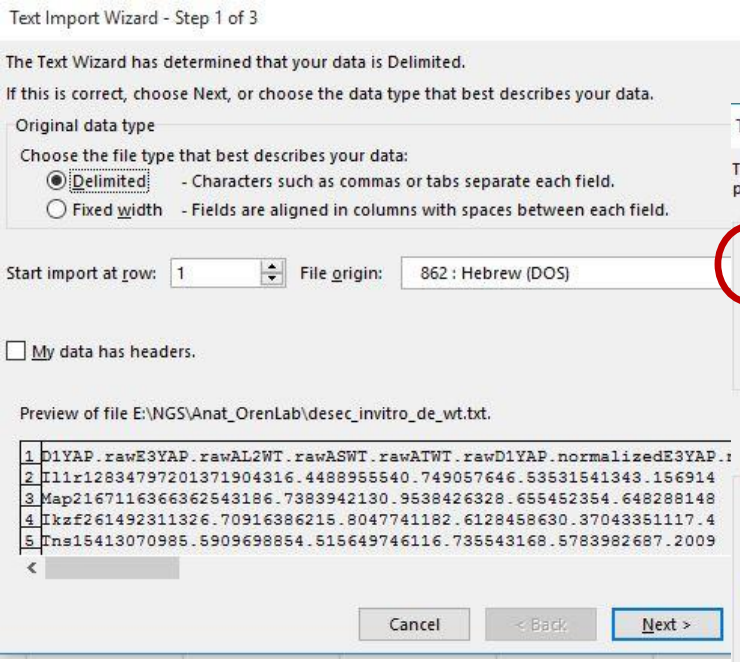
Output folder: /home/labs/mosheoren/Collaboration/anat_rna_seq/180320_D00257_0307_BCBYFNANXX/20.8.18_anatger1/20180820_140229_20_8_18_anatger1_transcriptome_RNA-seq/

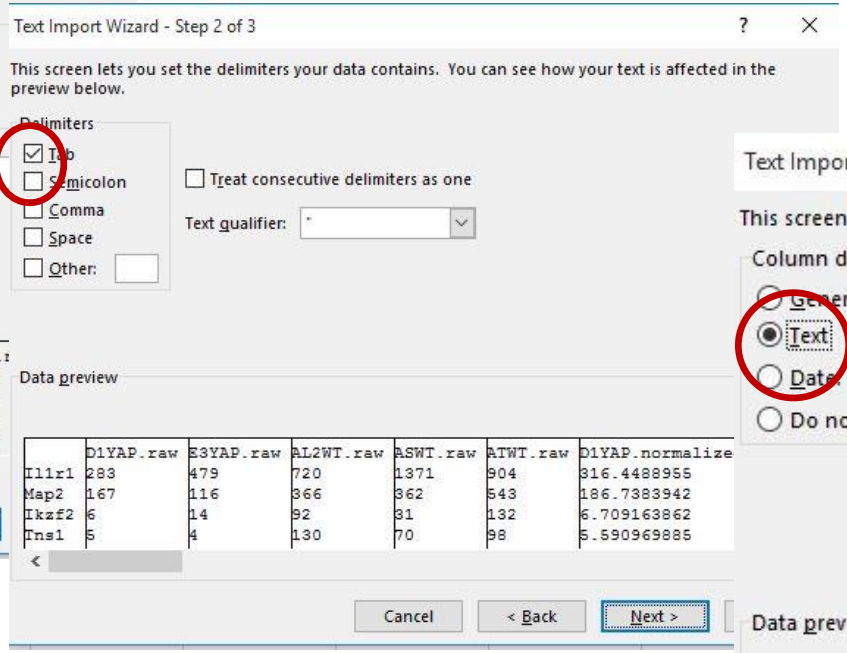
Statistics regarding the number of reads for each sample for various steps of the pipeline can be downloaded from here.

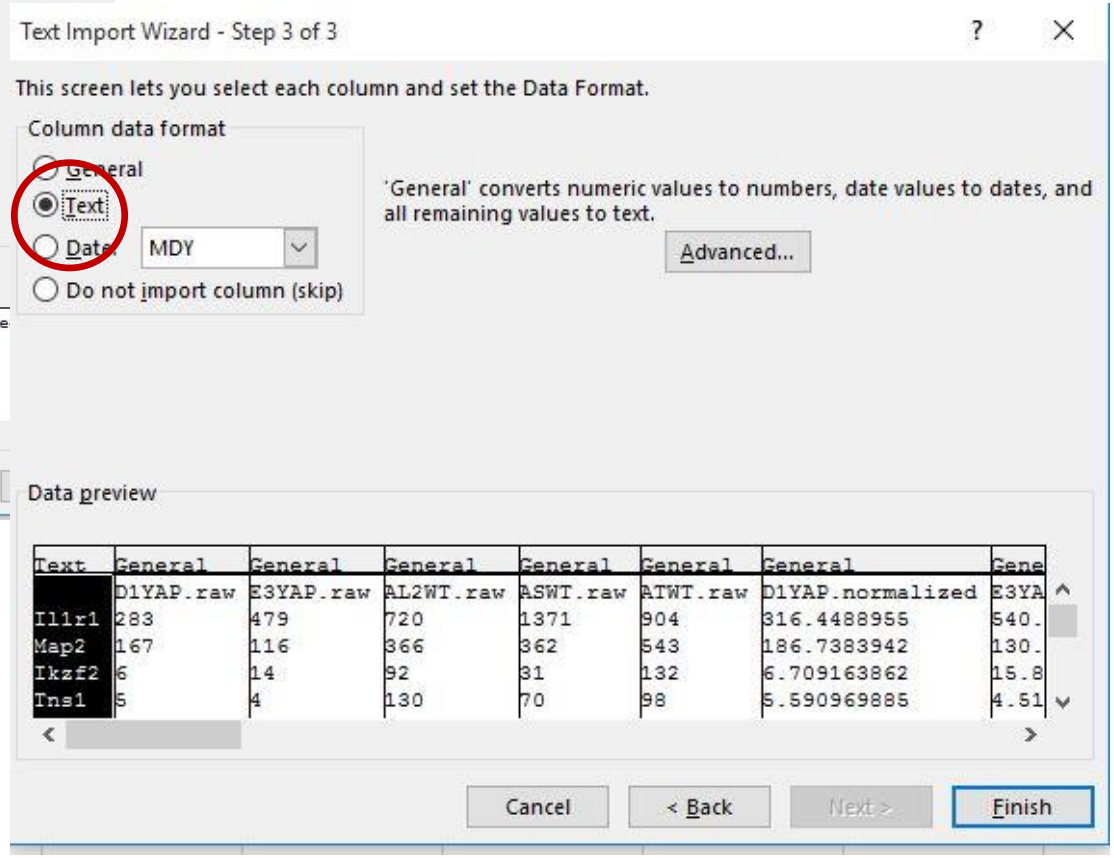
Raw counts can be downloaded from here.

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

EXCEL Tips – Import



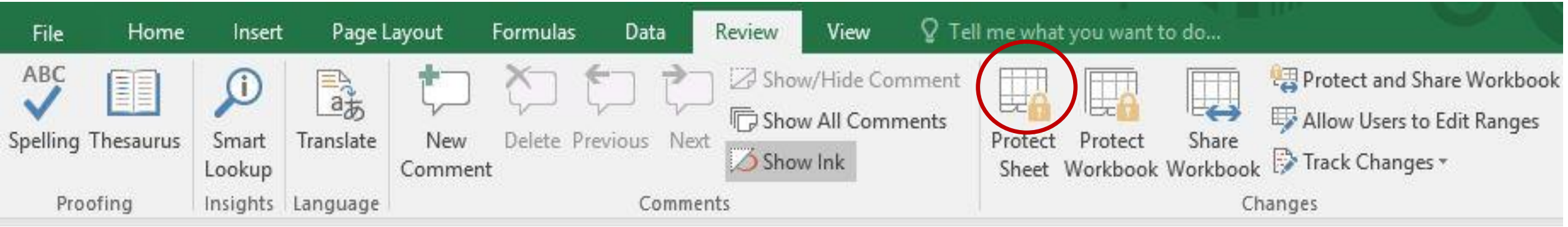




Note: **Text** format will enable gene names to be displayed correctly (not as dates).

EXCEL Tips – data protection

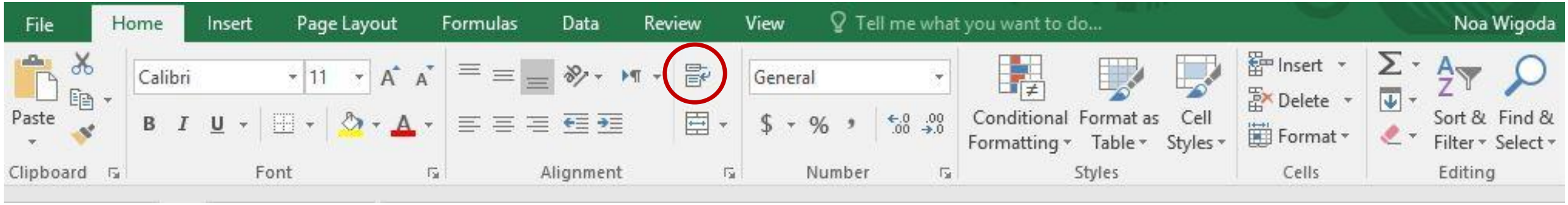
- **Protect** your data from unwanted changes



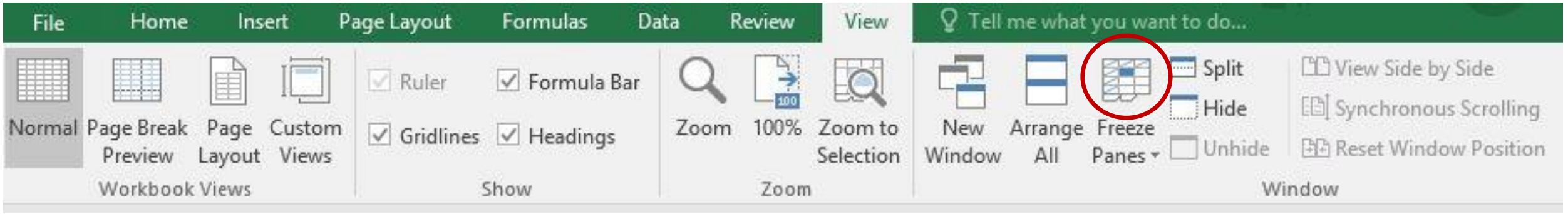
- Keep a copy of your original file, or duplicate the first tab and work only on the duplicated version.

EXCEL Tips – Header view

- **Wrap** text of header so it appears on multiple lines



- **Freeze** panes – keep gene names (first column) and header (first line) visible when you scroll through the rest of the sheet.



EXCEL Tips – 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.

EXCEL Tips – file content (cont.)

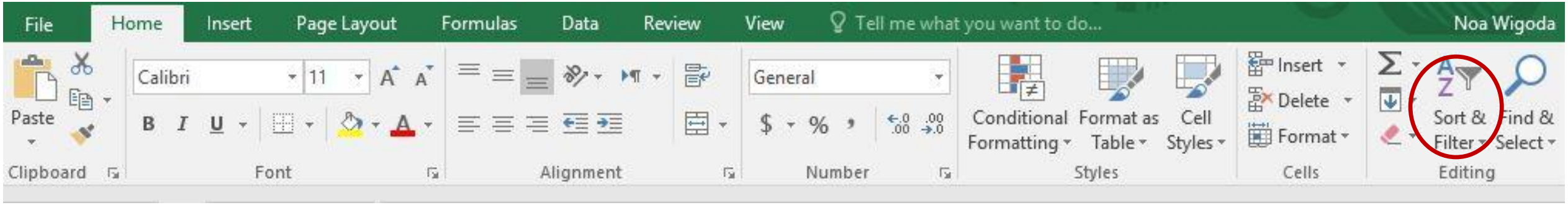
Control_Treatment. baseMean	Control_Treatment .log2FoldChange	Control_Treatment .pvalue	Control_Treatment. padj	Control_Treatment .pass	Control_Treatment. Direction

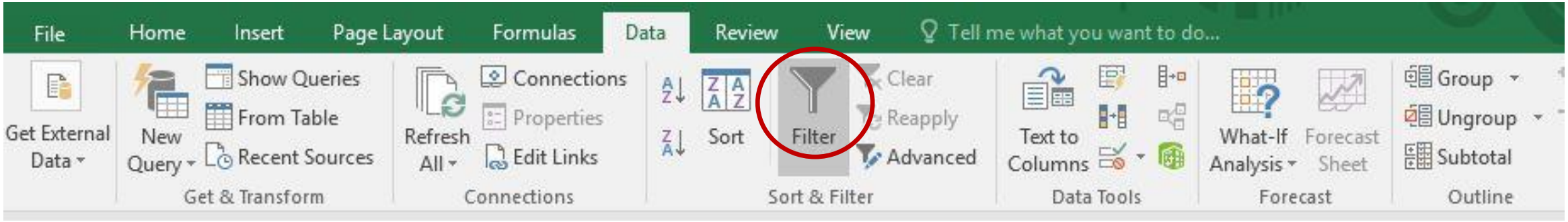
- For each **comparison** (for example control vs. treatment), 6 columns:
  - control_treatment.baseMean – mean (average) normalized read counts of all samples
  - control_treatment.log_2FoldChange – Fold change is a measure of the ratio of means of two populations (i.e control and treatment). *$Log_2(2)=1$*
  - 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, $|log_2FoldChange| \geq 1$ and padj ≤ 0.05
  - control_treatment.Direction – Up or Down

EXCEL Tips – filter

- Filter - easy way to find and work with a subset of data
- Turn on filtering, then click the arrow in the header row to narrow down the data





EXCEL Tips – filter

- Comparison of two conditions (for example: control vs. treatment)
- Turn on filtering, then click the arrow in the header row to narrow down the data

baseMean	.log2FoldChang	pvalue	.padj	.pass	.Direction
170.9628766	-0.251366771	0.564424794	0.806485796	no	down
431.4356073	-0.0993643	0.740899787	0.908136645	no	down
276.4899787	0.440414494	0.238138563	0.550504224	no	up
5.40704255	0.014185193	0.993908607	0.997530196	no	up
225.8101762	-0.02491451	0.947755404	0.983293568	no	down
2670.803231	1.144508531	5.64E-04	0.015037284	yes	up
570.1068855	0.405872901	0.212529583	0.520350531	no	up

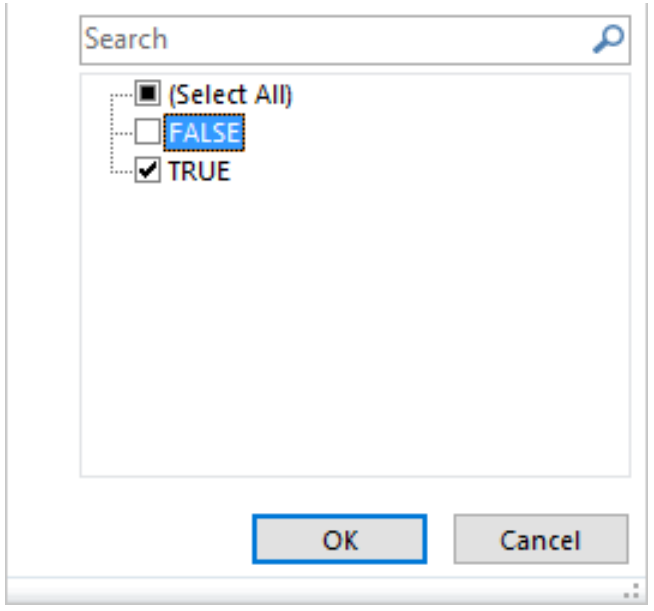
EXCEL Tips – add a query

- Comparison of a few conditions (for example: control vs. treatment A or treatment B) – based on pass column
- Add column (last) for your query:

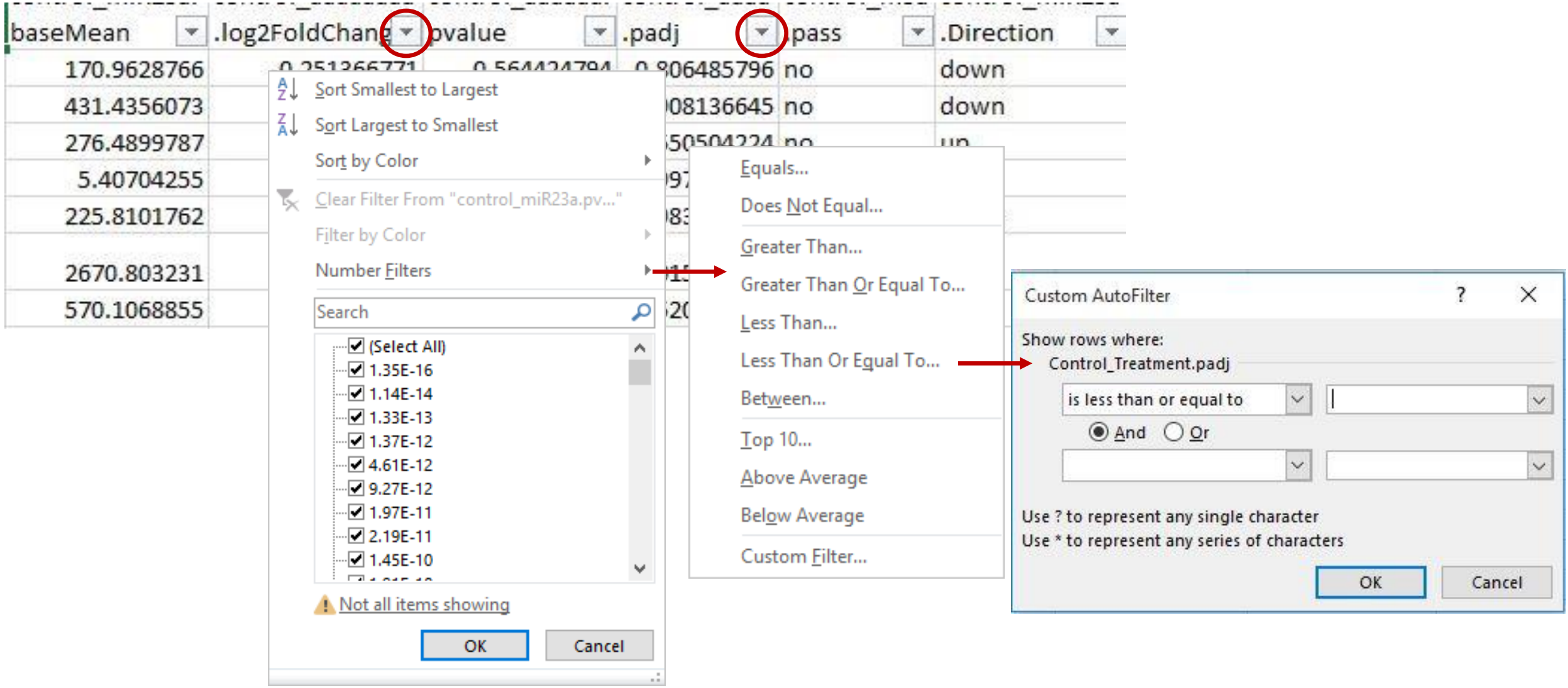
  =OR(AD2="yes",AJ2="yes") *in this example AD and AJ are the column indices*.

  It means: control vs. treatment A pass column ="yes“ **OR** control vs. treatment B pass column ="yes"

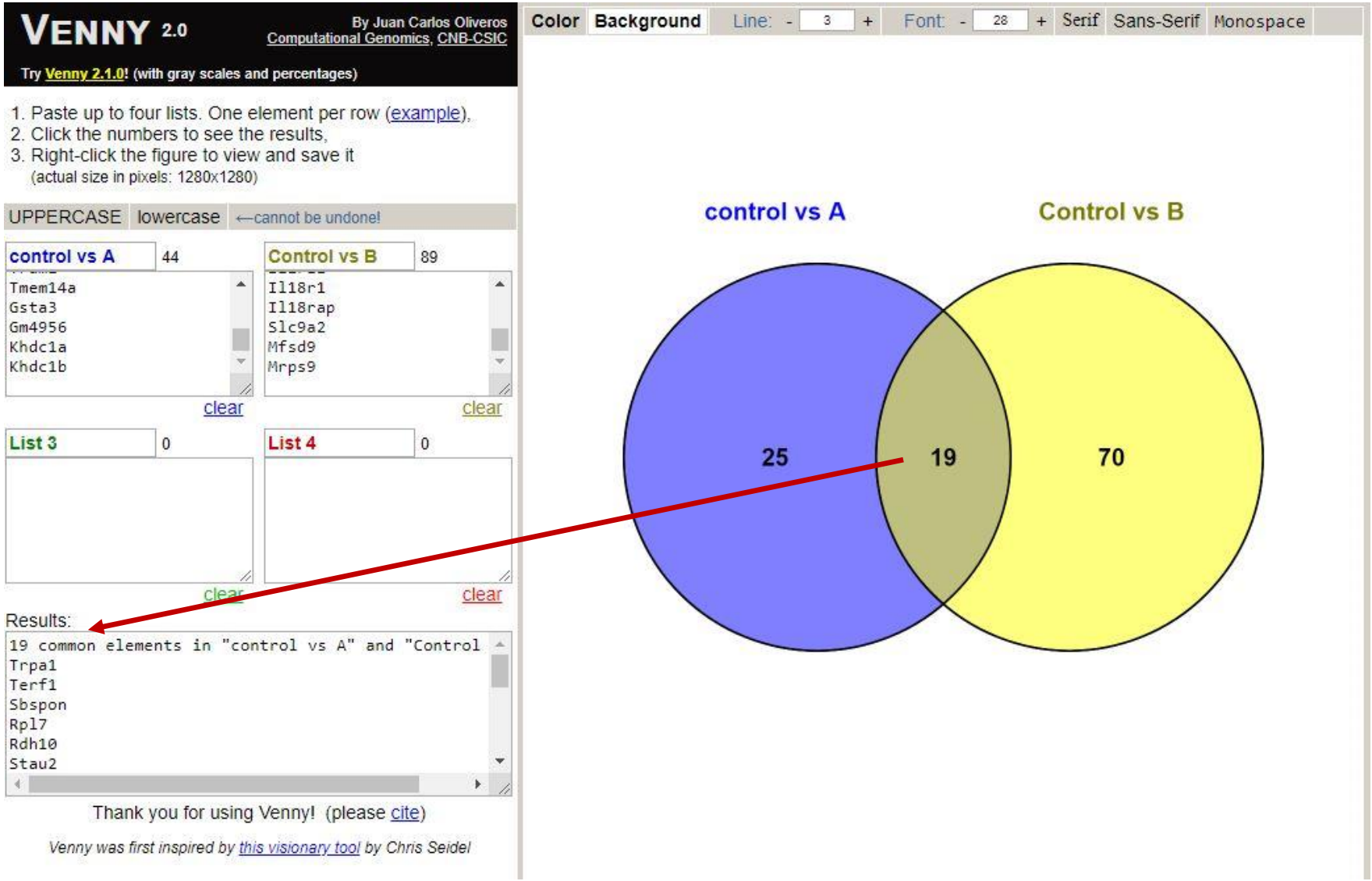
- Filter the new query column – select all cells which are TRUE.



EXCEL Tips Change Filter - of fold change or adjusted p value thresholds



Venn Diagrams using Venny



http://bioinfogp.cnb.csic.es/tools/venny/index2.0.2.html

Advanced tip

EXCEL tips – Get the Deseq2 results of a specific list of genes (use the function **vlookup**)

- Look for a value in the leftmost column of a table (gene ID), and return the value in the same row from a column you specify.
- List of genes – sorted ascending (in a separate tab or file)
- Deseq results file – sorted ascending by gene name
- Add column (last) for your request:

  =VLOOKUP(A2,[Book1]Sheet1!$A$1:$A$12,1,FALSE)

- If the gene ID is not in the specified list you will get:

מכון ויצמן למדע
WEIZMANN INSTITUTE OF SCIENCE

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

Contact Dena...

Support of 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)