

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

Text Import Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.
If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

- Delimited - Characters such as commas or tabs separate each field.
- Fixed width - Fields are aligned in columns with spaces between each field.

Start import at row: 1 File origin: 862 : Hebrew (DOS)

My data has headers.

Preview of file E:\NGS\Anat_OrenLab\desec_invitro_de_wt.txt

1	D1YAP.raw	E3YAP.raw	AL2WT.raw	ASWT.raw	ATWT.raw	D1YAP.normalized	E3YAP.
2	I11r1	283	479	720	1371	904	316.4488955
3	Map2	167	116	366	362	543	186.7383942
4	Ikzf2	6	14	92	81	132	6.709163862
5	Tns1	5	4	130	70	98	5.590969885

Buttons: Cancel, < Back, Next >

Text Import Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters:

- Tab
- Semicolon
- Comma
- Space
- Other: []

Treat consecutive delimiters as one

Text qualifier: []

Data preview

	D1YAP.raw	E3YAP.raw	AL2WT.raw	ASWT.raw	ATWT.raw	D1YAP.normalized	E3YA
I11r1	283	479	720	1371	904	316.4488955	540.
Map2	167	116	366	362	543	186.7383942	130.
Ikzf2	6	14	92	81	132	6.709163862	15.8
Tns1	5	4	130	70	98	5.590969885	4.51

Buttons: Cancel, < Back, Next >

Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

- General
- Text
- Date MDY
- Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Advanced...

Data preview

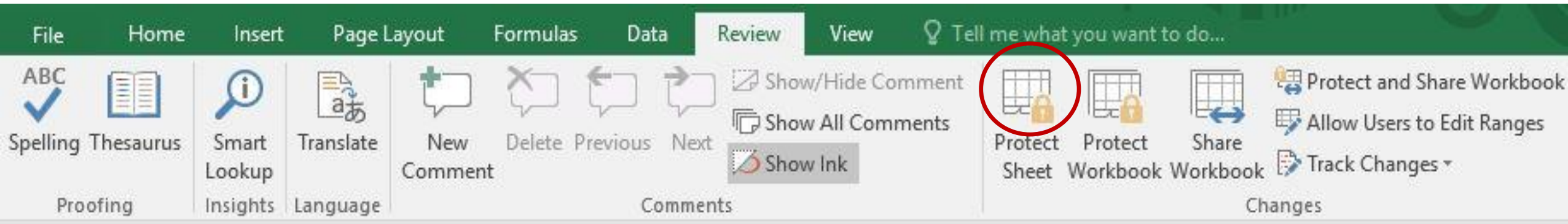
Text	General	General	General	General	General	General	Gene
D1YAP.raw	E3YAP.raw	AL2WT.raw	ASWT.raw	ATWT.raw	D1YAP.normalized	E3YA	
I11r1	283	479	720	1371	904	316.4488955	540.
Map2	167	116	366	362	543	186.7383942	130.
Ikzf2	6	14	92	81	132	6.709163862	15.8
Tns1	5	4	130	70	98	5.590969885	4.51

Buttons: Cancel, < Back, Next >, Finish

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

EXCEL Tips – data protection

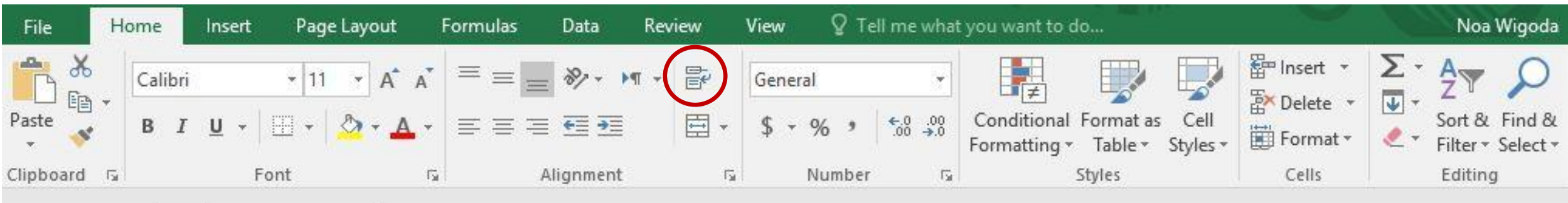
- **Protect** your data from unwanted changes



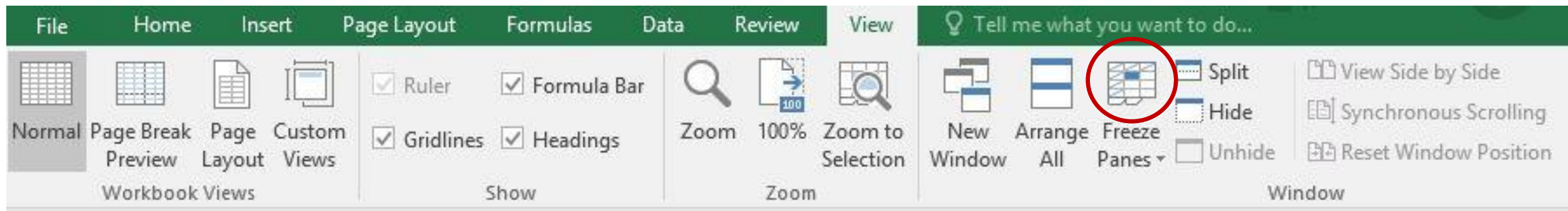
- Keep a copy of your original file, or duplicate the first tab and work only on the duplicated version.

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.

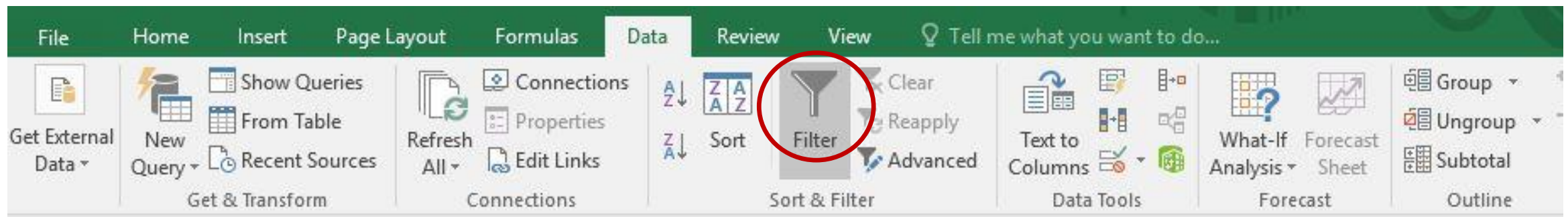
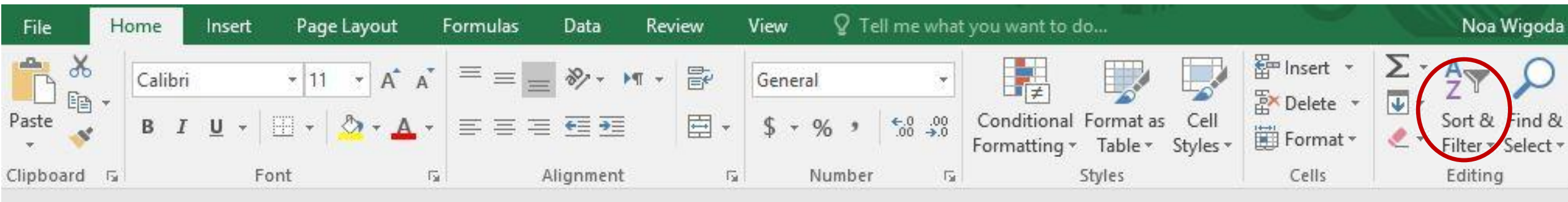
EXCEL Tips – file content (cont.)

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.log₂FoldChange – Fold change is a measure of the ratio of means of two populations (i.e control and treatment). $\text{Log}_2(2)=1$
 - 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: $\text{baseMean} \geq 5$, $|\text{log}_2\text{FoldChange}| \geq 1$ and $\text{padj} \leq 0.05$
 - 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

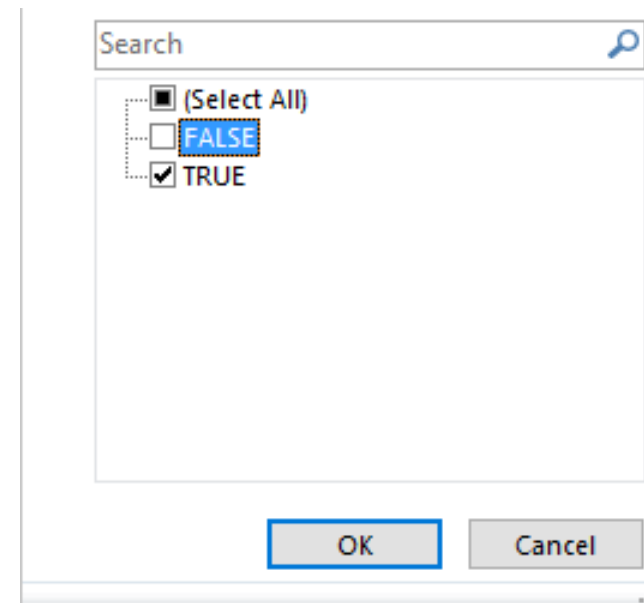
EXCEL Tips – add a query

- 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

baseMean	.log2FoldChange	pvalue	.padj	pass	.Direction
170.9628766	0.251266771	0.564424794	0.906485796	no	down
431.4356073			0.08136645	no	down
276.4899787			50504224	no	up
5.40704255			97		
225.8101762			83		
2670.803231			15		
570.1068855			20		

Sort Smallest to Largest
Sort Largest to Smallest
Sort by Color
Clear Filter From "control_miR23a.pv..."
Filter by Color
Number Filters

Search

- (Select All)
- 1.35E-16
- 1.14E-14
- 1.33E-13
- 1.37E-12
- 4.61E-12
- 9.27E-12
- 1.97E-11
- 2.19E-11
- 1.45E-10

Not all items showing

OK Cancel

Equals...
Does Not Equal...
Greater Than...
Greater Than Or Equal To...
Less Than...
Less Than Or Equal To...
Between...
Top 10...
Above Average
Below Average
Custom Filter...

Custom AutoFilter

Show rows where:
Control_Treatment.padj

is less than or equal to

And Or

Use ? to represent any single character
Use * to represent any series of characters

OK Cancel

Venn Diagrams using Venny

VENNY 2.0 By Juan Carlos Oliveros
Computational Genomics, CNB-CSIC

Try **Venny 2.1.0!** (with gray scales and percentages)

1. Paste up to four lists. One element per row ([example](#)).
2. Click the numbers to see the results.
3. Right-click the figure to view and save it
(actual size in pixels: 1280x1280)

UPPERCASE lowercase ←cannot be undone!

control vs A 44 **Control vs B** 89

Tmem14a
Gsta3
Gm4956
Khdc1a
Khdc1b

clear

Control vs B 89

I118r1
I118rap
Slc9a2
Mfsd9
Mrps9

clear

List 3 0 **List 4** 0

clear

clear

Results:

19 common elements in "control vs A" and "Control vs B"

Trpa1
Terf1
Sbspon
Rp17
Rdh10
Stau2

Thank you for using Venny! (please [cite](#))

Venny was first inspired by [this visionary tool](#) by Chris Seidel

Color Background Line: - 3 + Font: - 28 + Serif Sans-Serif Monospace

control vs A **Control vs B**

Category	Count
control vs A (exclusive)	25
Intersection	19
Control vs B (exclusive)	70

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!A1:A12,1,FALSE)`

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

#N/A

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

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)