

Excel Tips and Tricks

[Bingbing Yuan]

January 14, 2021
Bioinformatics and Research Computing
Whitehead Institute

http://barc.wi.mit.edu/hot_topics/



WHITEHEAD INSTITUTE



1

You will Learn:

- General tips and tricks in navigation, filtering, and selection
 - Keep input data types
 - Keep headers visible as you scroll
 - Keyboard navigation shortcuts
 - Quickly identify patterns within a column
 - Split items within single column
 - Highlight cells by their values
- Using formulas to work with data
- Merge multiple tables with vlookup
- Summarize data with PivotTable



WHITEHEAD INSTITUTE

2



Problem: Excel changes my input data

Month: JAN to DEC

a number contains the letter E or
number contains more characters
than can be displayed

	A1	ID	
	A	B	
1	ID	Default format	
2	SEP2	2-Sep	
3	2310009E13	2.31E+19	
4	0223345	223345	
5			
6			

a number contains leading zeros

Note: Copy from HTML page: use paste special -> Unicode text

3

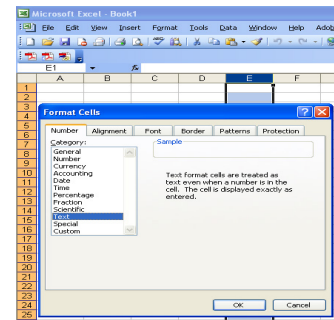


WHITEHEAD INSTITUTE



Solution: Change formatting to "Text"

1. Open file with plain text
2. Before you populate the data into Excel, select the cells that will contain the gene symbols, right-click and choose **Format Cells > Text > OK**
3. Copy and paste data



WHITEHEAD INSTITUTE

4



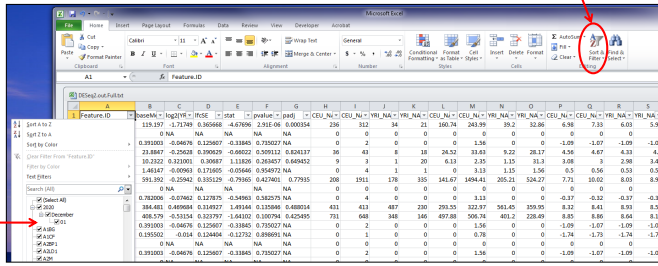
3

4

Sort & Filter button: Easy to spot patterns

- Temporarily hide repetitive elements
- Sort the rest of the elements in order

Sort & Filter

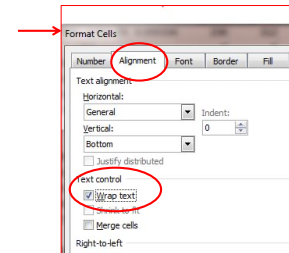


5

Problem: End of cell contents is missing Solution: Wrap text

	A	B	C
1	Feature	baseMea	log2(YR/
2	A1BG	119.197	-1.71749
3	A1CF	0 NA	
4	A2BP1	0.391003	-0.04676
5	A2LD1	23.8847	-0.25628

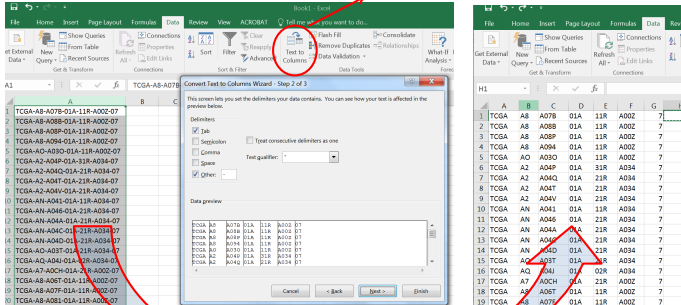
	A	B	C
1	Feature	baseMea	log2(YR/
2	A1BG	119.197	-1.71749
3	A1CF	0 NA	
4	A2BP1	0.391003	-0.04676
5	A2LD1	23.8847	-0.25628



6

Split items within single column: example 1

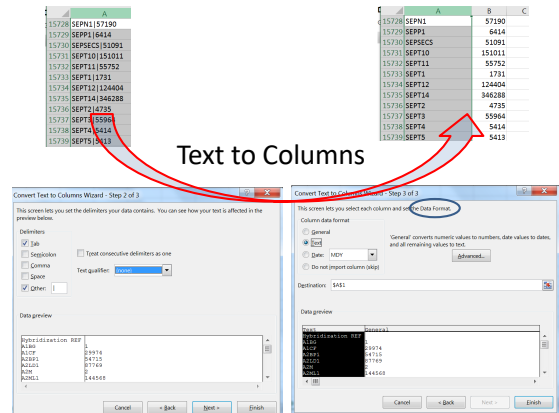
Text to Columns



7

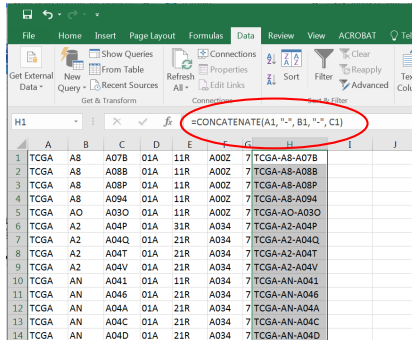
Split items within single column: example 2

Text to Columns



8

Merge multiple columns

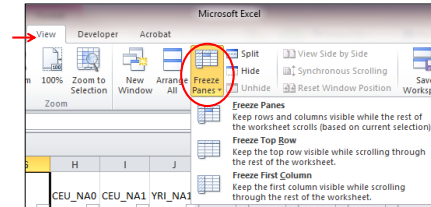


9



Keep headers visible as you scroll

- Choose **View > Freeze Panes > Freeze top Row**
- To unfreeze frozen rows, choose **View > Freeze Panes > Unfreeze Panes**



Freeze both row and column



10



Keyboard navigation shortcuts

	PC	MAC
First cell in the active column	Control/Ctrl	Command
Last cell in the active column	Control/Ctrl	Command
First cell in the active row	Control/Ctrl	Command
Last cell in the active row	Control/Ctrl	Command
The very first cell (A1)	Control/Ctrl	Control/Ctrl
Bottom right corner	Control/Ctrl	Control/Ctrl



11



Keyboard selection shortcuts

- Selecting a large range of cells can be a pain. Here's one trick that could help.
 - To start the selection of a large matrix, **start by selecting the bottom right cell** of the desired matrix (**Ctrl + End**)
 - Use "**Ctrl + Shift + up arrow**" to select everything above the original cell.
 - Use "**Ctrl + Shift + left arrow**" to select everything to the left of the original cell.
 - Use "**Shift + down arrow**" to move down one row (to avoid the header line, if present).
 - Use "**Shift + right arrow**" to move to the right one column (to avoid any ID column(s), if present)



12



The Cancer Genome Atlas (TCGA)

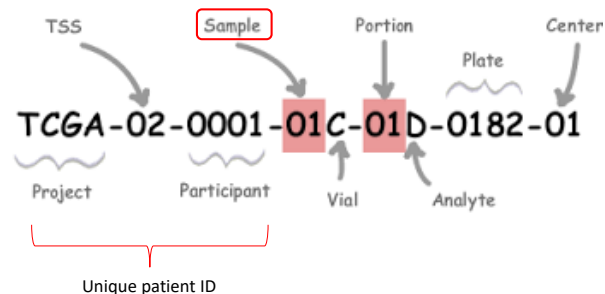
- NCI / NHGRI project to molecularly characterize
 - Over 20k primary cancer samples
 - Matched normal samples
 - 33 cancer types
- Genomic, epigenomic, transcriptomic, and proteomic assays
- Data we'll use for exercises:
 - Small subset of BRCA RNA-seq data from TCGA
 - Unit: normalized gene expression levels



13



TCGA barcode IDs



https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/



14



13

14

TCGA Sample codes

	Code	Definition	Short Letter Code
Tumor (01-09)	01	Primary Solid Tumor	TP
	02	Recurrent Solid Tumor	TR
	03	Primary Blood Derived Cancer - Peripheral Blood	TB
	04	Recurrent Blood Derived Cancer - Bone Marrow	TRBM
	05	Additional - New Primary	TAP
	06	Metastatic	TM
	07	Additional Metastatic	TAM
	08	Human Tumor Original Cells	THOC
	09	Primary Blood Derived Cancer - Bone Marrow	TBM
Normal	10	Blood Derived Normal	NB
	11	Solid Tissue Normal	NT
	12	Buccal Cell Normal	NBC
	13	EBV Immortalized Normal	NEBV
	14	Bone Marrow Normal	NBM
	15	sample type 15	15SH
	16	sample type 16	16SH
	20	Control Analyte	CELLC

Complete sample codes can be found at:
<https://cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes>



15



Ex.1 (TCGA_data.xlsx)

- Freeze the portion of the sheet BRCA_from_TCGA, so you can always see the column names (TCGA barcodes) and gene names no matter how you scroll.
 - hint: View -> Freeze panes / Freeze Top Row
- Split the TCGA barcodes and check the sample codes. How many sample types?
 - hint: copy and paste barcodes to another sheet with Paste special -> Transpose, then split barcodes with Data -> Text to Columns -> View sample codes by turning on Sort & Filter button under Home panel)
- Replace sample barcodes with meaningful names:
 - 01 -> Primary
 - 06 -> Metastasis
 - 11 -> Normal
- Add sample codes as additional row to top of the BRCA_from_TCGA sheet.
 - hint: Copy -> Paste special -> Transpose
- Sort the samples by sample codes
 - hint: Data -> Sort -> Options -> Sort left to right



16



15

16

Math formulas

- Absolute value:
=ABS(B4)
- Average:
=AVERAGE(B4:D4)
- Log2 ratio:
=LOG((O4+1)/(Q4+1),2)

Add pseudocount 1 to avoid denominator being 0



17

17

Statistics Formulas

- Example: T test:

=TTEST(B4:D4, I4:N4, 2, 3)

- **Tails** Required. Specifies the number of distribution tails. If tails = 1, T.TEST uses the one-tailed distribution. If tails = 2, T.TEST uses the two-tailed distribution.
- **Type** Required. The kind of t-Test to perform.

Parameters

If type equals	This test is performed
1	Paired
2	Two-sample equal variance (homoscedastic)
3	Two-sample unequal variance (heteroscedastic)



18

18

Logical formulas

- Examples:

- If, else
=IF(U4<0.05, 1, 0)

True

False

- and or
=IF (AND (T4<0.05, R4>=1), "yes", "no")



19

19

Formula details

- Begin with equal sign: =
- Dollar sign (\$): absolute cell references
- Convert formula to real value:
 - Copy → Paste Special → Values



20

20

Apply formula to multiple rows at once

Copy equation (optional)

2. Arrow appear

3. Double clicks

Note: double check several cells to make sure

21

Remove extra spaces within a cell

=TRIM(A1)

22

How to get a unique list of items within a field?

Methods	Sort & Filter	Data -> Remove Duplicates
Is output list in sorted order?	Yes	No
Can the list be saved?	No	Yes
Can all items in a super long list be displayed?	No	Yes

23

Color cells with Conditional Formatting Example 1

Conditional Formatting

Highlight Cells Rules

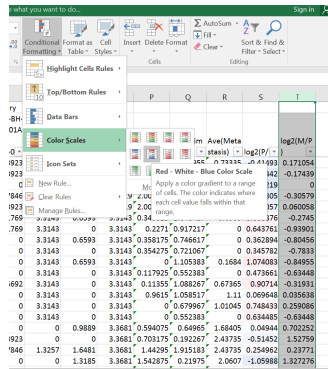
Greater Than...

24

Color cells with Conditional Formatting

Example 2

Highlight extreme values: Color Scales



WHITEHEAD INSTITUTE

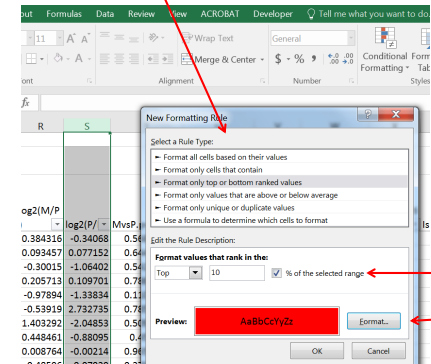
25



Color cells with Conditional Formatting

Example 3

Highlight extreme values: Color Scales -> More Rules



MAC:
Excel version 16:
It's "Top/Bottom Rules"

Top/bottom



WHITEHEAD INSTITUTE

26



Ex.2 (TCGA_data.xlsx)

- Drop genes with no expression in all samples
 - Hint: Create another column and name it as "Max" -> use formula to calculate maximum value for each gene -> Sort data by "Max"
 - How do you easily find out where the non-zero value start?
- Compare primary tumor vs normal, and metastatic cancer vs primary tumor:
 - Convert values to log2 format:
 - To avoid dividing by 0, you can add 1 to each value
 - Start to get the log2 expression for the top gene in the first patient sample, use our tricks (see slides) to fill out the rest of the cells
 - Calculate log2 average for each gene in normal samples, and do the same for primary tumor and metastasis samples.
 - Calculate log2 fold changes (log2FC) between primary tumor and normal, and between metastatic cancer and primary tumor, and highlight the log2FC with Conditional Formatting
 - Calculate p-values with ttest, delete the genes with no valid pvalues, find out FDR adjusted p-value with the BarC web tool (<http://bioinfo.wi.mit.edu/bio/tools/fdr/>), and highlight FDR adjusted p-values < 0.05 with Conditional Formatting.
- We consider a gene is "differentially expressed" if meeting the cutoffs (FDR adjusted p-value < 0.05 + at least 2 fold difference). Add two columns named as Sig(PvsN) and "Sig(MvsP)", fill in with "y" and "n" to indicate if a gene is "differentially expressed" or not.

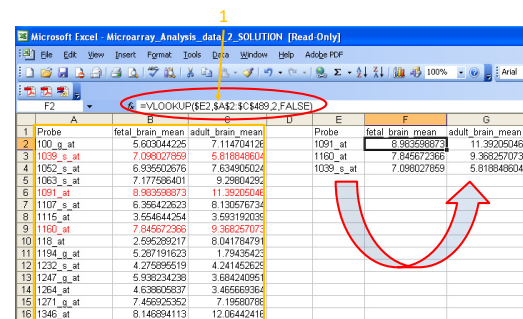
Hint: with "if" plus "and" formula



27



Lookup and retrieve data



Too complicated?

Try BarC Submatrix Selector

http://bioinfo.wi.mit.edu/bio/tools/submatrix_selector/



WHITEHEAD INSTITUTE

28



27

Look up values in data list: vlookup

• VLOOKUP(lookup_value,table_array,col_index_num,range_lookup)

- **lookup_value**: The value to search in the first column of the table array
- **table_array**: The table to search (containing the value to search for in the first column)
 - Note: use "\$" to fix the table
- **col_index_num**: the column number from which the matching value is returned
- **range_lookup**:
 - True/omitted, an exact/approximate match
 - False: an exact match

Dollar sign (\$): absolute cell references

29

Ex.3 (TCGA_data.xlsx)

1. Extract gene symbols from the cells under column "Hybridization REF". How do you check gene symbols to make sure none of them is converted to date format?
Hint:
Copy cells in this column and Paste them next to the rightmost column -> use "Text to Columns" to split cells
2. Use vlookup to get the expression levels of a list of genes, such as genes in the KEGG_ECM_RECEPTOR_INTERACTION gene set from [Broad's MSigDB](#).
3. Are any of these genes "differentially expressed" as defined by
 - 1) adjusted p-value < 0.05
 - 2) absolute log2 fold change >= 1
 Hint: Use "Sort & Filter" button to quickly narrow down the genes

30

29

30

Summarize data with PivotTable

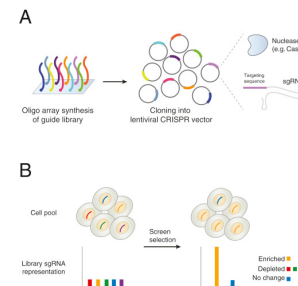
The screenshot shows the Excel interface with a PivotTable being created. The PivotTable Fields task pane on the right shows 'Sample3' as the source and 'Average of Sample3' as the field. The PivotTable is set to 'Sum' and 'Average of Sample3'. The PivotTable is located in the 'Insert' tab, and the 'PivotTable' button is highlighted. The PivotTable is set to 'Sum' and 'Average of Sample3'.

In this example, it grouped the values by gene and calculated the average

31

Pooled CRISPR screening

- Simultaneous targets of a large number of genes or regulatory elements.
- To prevent off-target, library includes multiple single guide RNAs (sgRNAs) for each gene.
- The entire cellular population is subjected to selective pressure.
- The genes driving the phenotype of interest are identified by comparing the relative abundance of corresponding sgRNAs in both fractions.



Anal Biochem. 2017 532:95

32

31

32

Ex.4 (CRISPR_data.xlsx)

- CRISPR_data.xlsx includes the percentage of guide RNA expression in a subset of a pseudo pooled CRISPR screen experiment. Each gene has 4 single guide RNAs.
- Could you calculate the percentage of expression at gene level with PivotTable? You can summarize expression of the sgRNAs by taking their average percentage.



33

33

References

- Excel Annoyances - Curtis Frye
- Excel Hacks - O'Reilly
- <http://www.exceltip.com> (Joseph Rubin)
- <http://excel.tips.net> (Allen Wyatt)
- Some Excel Basics as well as formula basics:
 - http://barc.wi.mit.edu/education/bioinfo2005/arrays/Excel_help.html
- Zeeberg BR, Riss J, Kane DW, Bussey KJ, Uchio E, Linehan WM, Barrett JC, Weinstein JN. BMC Bioinformatics 5:80 2004



Whitehead Institute

34

- Next Hot Topics:
 - Enrichment Analysis - January 28
- List of Winter/Spring Hot Topics Sessions:
 - http://barc.wi.mit.edu/education/hot_topics/upcoming/



35

35