Excel Tips and Tricks

March 12, 2020
Bioinformatics and Research Computing
Whitehead Institute

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

Whitehead Institute

BaRC Hot Topics

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

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

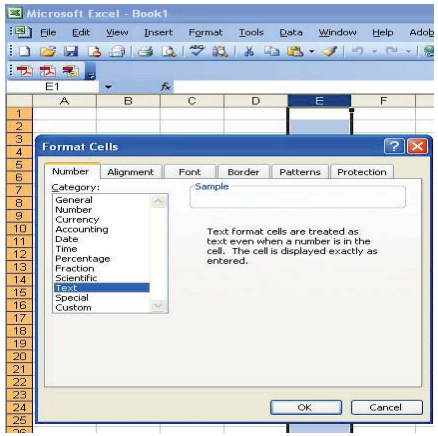
	A	B
1	ID	Default format
2	SEP2	2-Sep
3	2310009E13	2.31E+19
4	0223345	223345
5		
6		

a number contains leading zeros

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

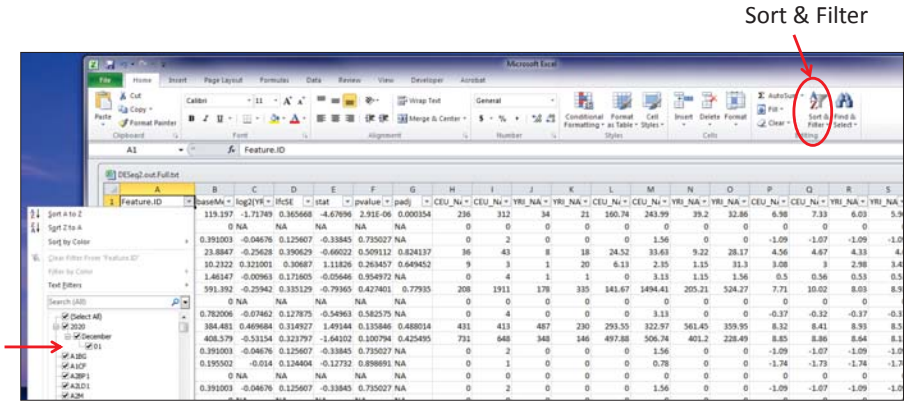
Solution: Change formatting to "Text"

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

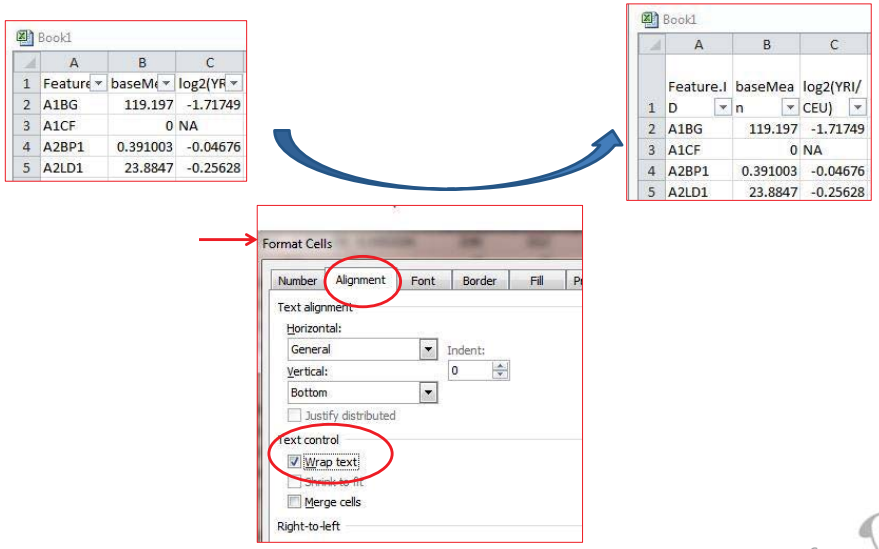


Sort & Filter button: Easy to spot patterns

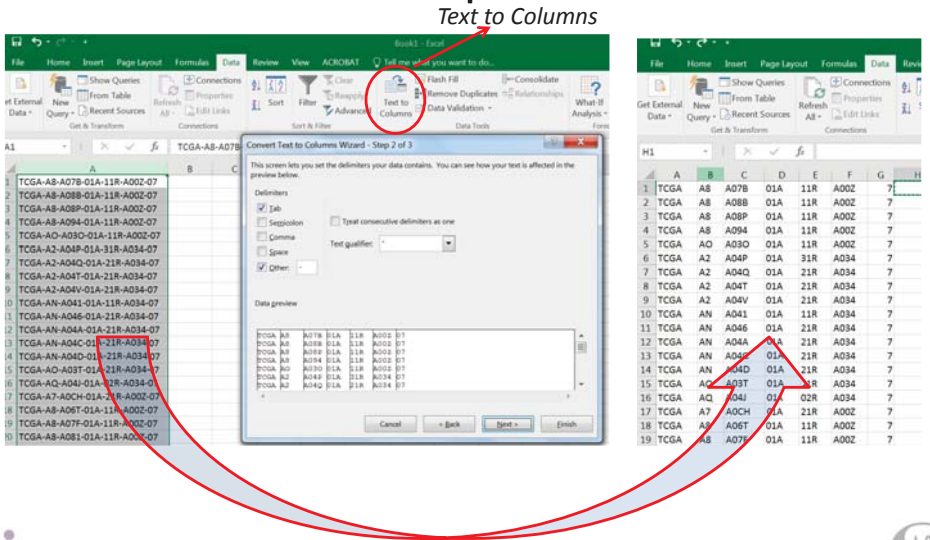
- Temporarily hide repetitive elements
- Sort the rest of the elements in order



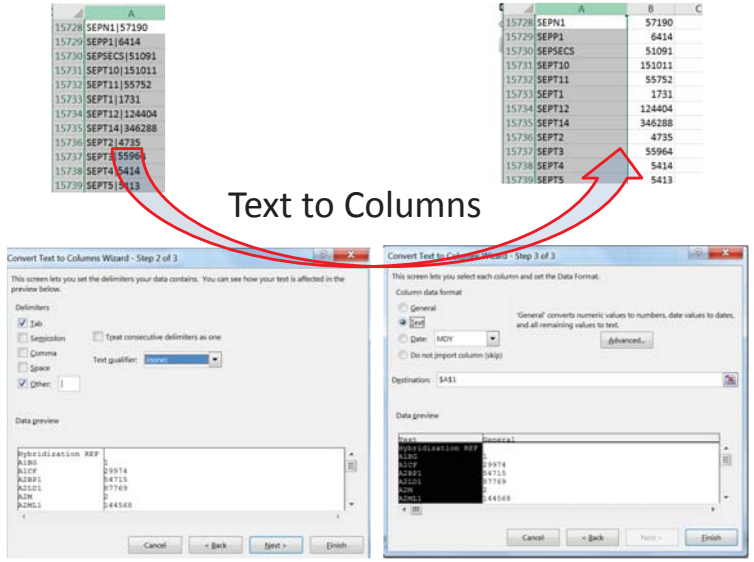
Problem: End of cell contents is missing
Solution: Wrap text



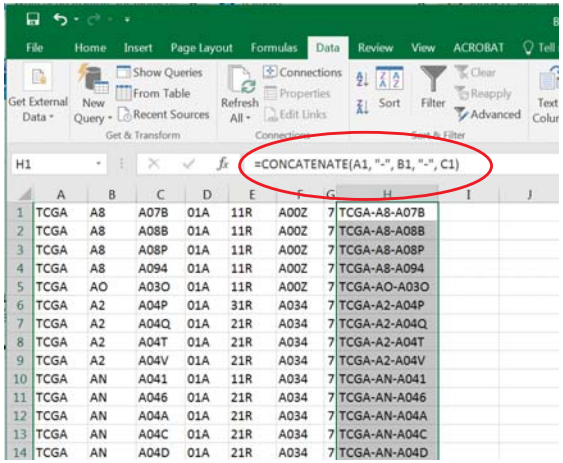
Split items within single column: example 1



Split items within single column: example 2

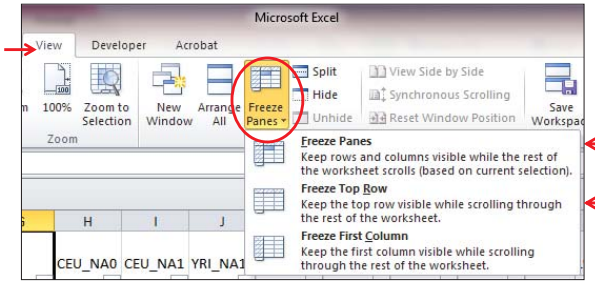


Merge multiple columns



Keep headers visible as you scroll

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



Keyboard navigation shortcuts

- PC:
  - First cell in the active column: **Ctrl + up arrow**
  - Last cell in the active column: **Ctrl + down arrow**
  - First cell in the active row: **Ctrl + left arrow**
  - Last cell in the active row: **Ctrl + right arrow**
  - The very first cell (A1): **Ctrl+ Home**
  - Bottom right corner: **Ctrl+ End**
- MAC:
  - use either **Command** or **Control**

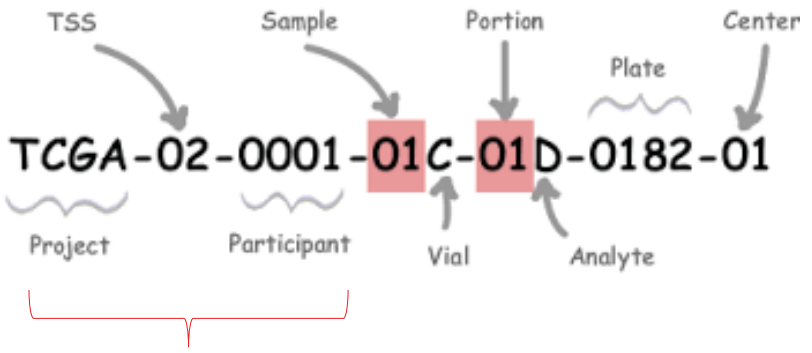
Keyboard selection shortcuts

- 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)

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

TCGA barcode IDs



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

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://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/sample-type-codes

Ex. 1

- 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
- 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

Math formulas

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

Add 1 to avoid denominator being 0

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)

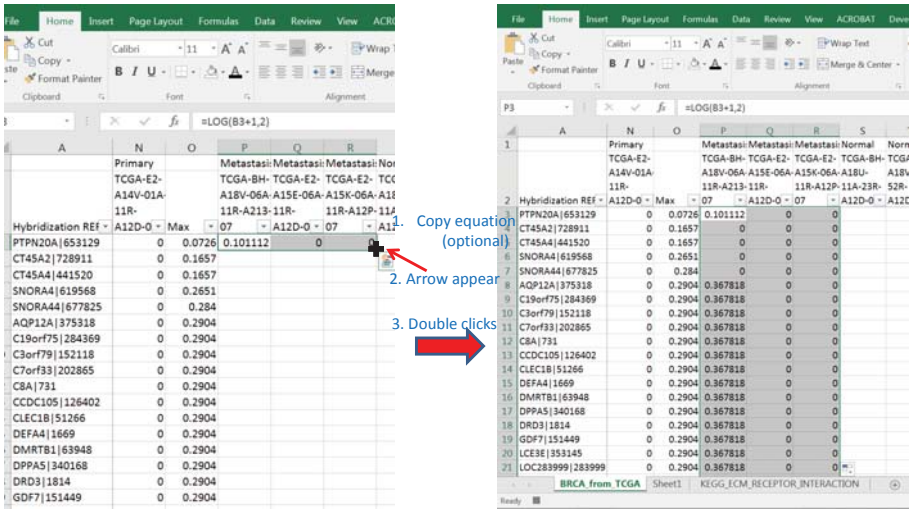
Logical formulas

- Examples:
  - If, else
    =IF(U4<0.05, 1, 0)
    True
    False
  - and or
    =IF (AND (T4<0.05, R4>=1), ”yes”, ”no”)

Formula details

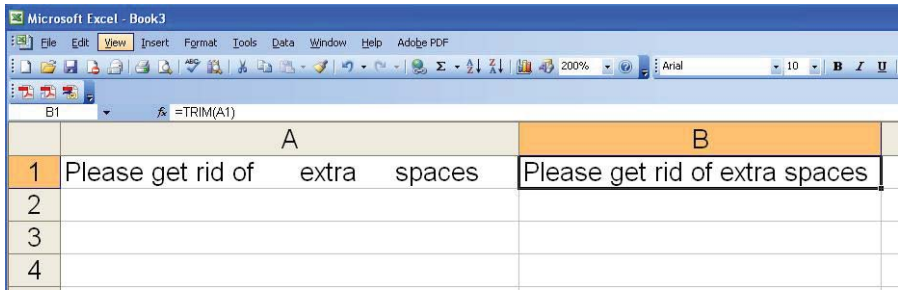
- Begin with equal sign: =
- Dollar sign ($): absolute cell references
- Copy → Paste Special → Values

Apply formula to multiple rows at once

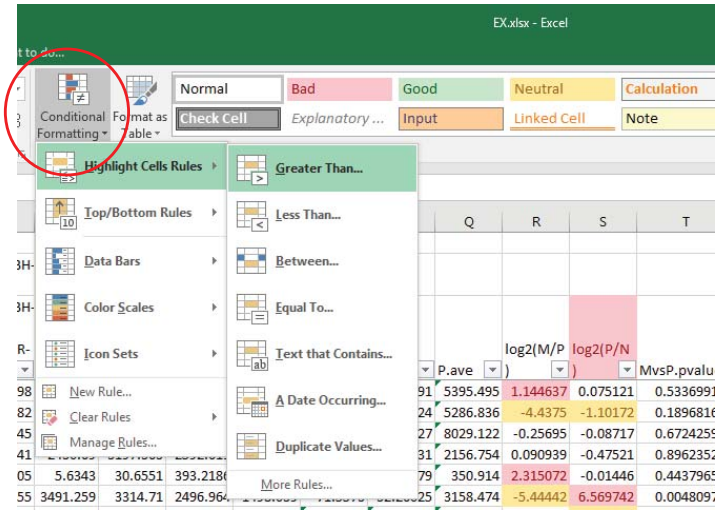


Note: double check several cells to make sure

Remove extra spaces within a cell

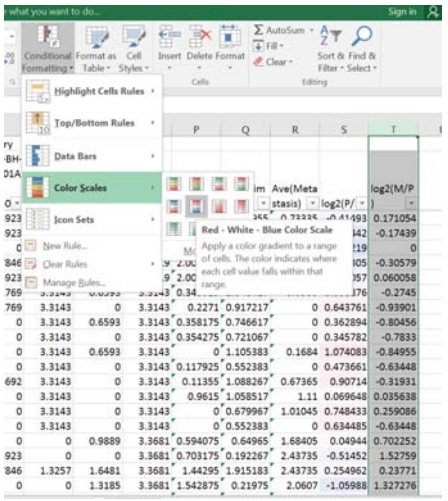


Color cells with Conditional Formatting Example 1



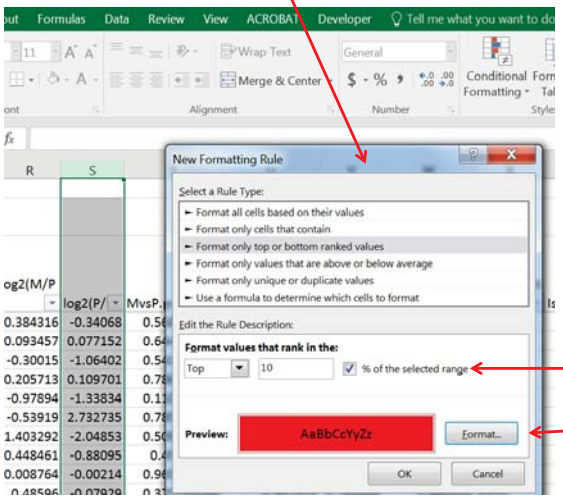
Color cells with Conditional Formatting Example 2

Highlight extreme values: Color Scales



Color cells with Conditional Formatting Example 3

Highlight extreme values: Color Scales -> More Rules



Top/bottom

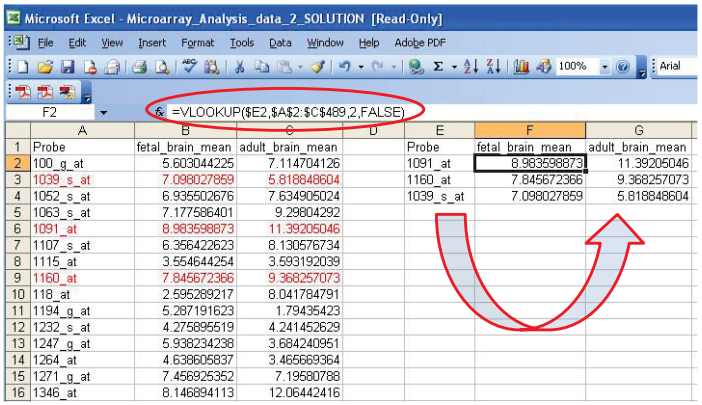
Ex. 2

1. Drop genes with no expression in all samples
   1) Hint: Create another column and name it as "Max" -> use formula to calculate maximum value for each gene -> Sort data by "Max"
   2) How do you easily find out where the non-zero value start?

      Hint:
      a) Convert values in "Max" column from formula format to value with Copy -> Paste Special -> Values
      b) After sorting, with Sort & Filter button to find out the value right above 0 -> Search for this value under "Max" column
2. Compare primary tumor vs normal, and metastatic cancer vs primary tumor:
   1. Convert values to log2 format:
      1) To avoid dividing by 0, you can add 1 to each value
      2) 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
   2. Calculate log2 average for each gene in normal samples, and do the same for primary tumor and metastasis samples.
   3. Calculate log2 fold changes (log2FC) between primary tumor and normal, and between metastatic cancer and primary tumor, and highlight the log2FC with Conditional Formatting
   4. Calculate p-values, 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.
3. 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

Lookup and retrieve data



	A	B	C	D	E	F	G
1	Probe	fetal_brain_mean	adult_brain_mean		Probe	fetal_brain_mean	adult_brain_mean
2	100_g_at	5.603044225	7.114704126		1091_at	8.983598873	11.39205046
3	1039_s_at	7.098027859	5.818848604		1160_at	7.845672366	9.368257073
4	1052_s_at	6.935502676	7.634905024		1039_s_at	7.098027859	5.818848604
5	1063_s_at	7.177586401	9.29804292				
6	1091_at	8.983598873	11.39205046				
7	1107_s_at	6.356422623	8.130576734				
8	1115_at	3.554644254	3.593192039				
9	1160_at	7.845672366	9.368257073				
10	118_at	2.595289217	8.041784791				
11	1194_g_at	5.287191623	1.79435423				
12	1232_s_at	4.275895519	4.241452629				
13	1247_g_at	5.938234238	3.684240951				
14	1264_at	4.638605837	3.465669364				
15	1271_g_at	7.456925352	7.19580788				
16	1346_at	8.146894113	12.06442416				

Too complicated?

Try BaRC Submatrix Selector

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

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

Ex.3

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 data format?

   Hint:

   Copy cells in this column and Paste them next to the rightest 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

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