For this demo, we will identify biological functions that are enriched in the differential expression analysis from our previous demo

Getting the data

1. We used the output from **sleuth** analysis of paired-end RNA-seq of Saccharomyces cerevisiae under aerobic and anaerobic conditions, each with 2 replicates (r1 and r2).
2. Those that do not have the output from the previous demo can get the files from https://drive.google.com/drive/folders/1i8MgMHK56I-TJbfNiH2SeMBTERS8SIDw?usp=sharing

Setting up software

We will do some data manipulation in Excel. But all enrichment analyses will be performed using online tools.

Running the demo

1. Open the differential expression analysis file, "sleuth_differential_expression.txt", in excel. The content should look like this.

A	B	C	D	E	F	G	H	I
	target_id	pval	qval	test_stat	rss	degrees_free	mean_obs	var_obs
1	NM_001181124.1	2.59E-10	1.27E-06	39.96539563	15.36339695	1	8.417310722	5.121132316
2	NR_132186.1	6.55E-10	1.27E-06	38.14979375	3.640905722	1	9.879766586	1.213635241
3	NR_132187.1	6.55E-10	1.27E-06	38.14979375	3.640905722	1	9.879766586	1.213635241
4	NM_001178902.1	4.30E-09	1.45E-06	34.48167964	6.678866411	1	8.185847672	2.226288804
5	NM_001179305.1	6.06E-09	1.45E-06	33.81376634	0.431345958	1	11.06920134	0.143781986
6	NM_001179347.3	6.59E-09	1.45E-06	33.65244067	14.28124796	1	8.201656196	4.760415988
7	NM_001180385.3	5.58E-09	1.45E-06	33.97442784	1.862161499	1	9.003896352	0.6207205
8	NM_001180810.3	6.39E-09	1.45E-06	33.71281635	2.818466062	1	8.519619279	0.939488687

2. Open the gene expression file in TPM unit, "kallisto_expression.txt", in excel. The content should look like this.

A	B	C	D	E
	aerobic1	aerobic2	anaerobic1	anaerobic2
NM_001178148.1	9.924504945	9.865041182	10.84427431	9.033759097
NM_001178149.1	9.138502413	8.571050881	9.582161084	10.74528304
NM_001178150.1	1171.672132	1170.681103	2148.017091	2154.089911
NM_001178151.1	190.2430374	196.6867042	89.27467731	91.84976857
NM_001178152.1	167.4525834	167.4358872	120.3110809	118.6026502
NM_001178153.1	31.80449311	40.69292227	28.71252687	33.11786185

3. We will merge the two table as follow

target_id	qval	log fold change
NR_132186.1	1.27E-06	-0.846798174
NR_132187.1	1.27E-06	-0.846798174
NM_001181124.1	1.27E-06	-1.720923198
NM_001181996.1	1.45E-06	1.067893183
NM_001180810.3	1.45E-06	0.724496369
NM_001180385.3	1.45E-06	0.577292558
NM_001181943.1	1.45E-06	0.459029393
NM_001184177.1	1.45E-06	-0.235505878
NM_001179305.1	1.45E-06	-0.291933475
NM_001182136.1	1.45E-06	-0.308702533
NM_001181781.1	1.45E-06	-0.327914251
NM_001181121.1	1.45E-06	-0.40426956

	aerobic1	aerobic2	anaerobic1	anaerobic2	fold change	log fold change
NM_001178389.1	35.56092439	37.85897267	25.87062603	23.72409111	1.480397536	0.170378354
NM_001178390.1	28.25988918	26.27723287	25.84739755	29.70966684	0.981641536	-0.008047073
NM_001178391.2	17.92996967	17.79542328	22.76090259	22.12903716	0.795844083	-0.099172008
NM_001178392.2	22.024952	18.78142542	16.03879598	14.32088656	1.344097632	0.128430816
NM_001178393.2	3.449614777	2.736569409	7.238449644	8.280980499	0.398608978	-0.399452924
NM_001178394.1	35.27046457	33.64590215	18.63186869	21.2124899	1.729639255	0.237955533
NM_001178395.1	25.71360899	27.22618424	28.61011762	28.40657691	0.928496358	-0.032219795
NM_001178396.1	936.81646	906.9509346	1875.80298	1858.019585	0.493801557	-0.306447545
NM_001178397.1	28.16718595	27.12251261	32.19316705	33.90564981	0.836470321	-0.077549464
NM_001178398.1	3.163446961	3.008737023	2.22773366	1.829762007	1.521180671	0.182180798
NM_001178400.1	90.57250569	90.09467014	47.45437341	41.86977887	2.022601628	0.305910353
NM_001178401.1	70.83396203	68.98677087	11.13402392	13.24232197	5.735918483	0.758602971

   a. Copy target_id and qval columns from the differential expression table
   b. Copy all columns from the TPM expression table
   c. Calculate log fold change between aerobic and anaerobic samples
      i. AVERAGE(F2:G2)/AVERAGE(H2:I2)
   d. Use VLOOKUP to map the log fold change to the column next to the qval

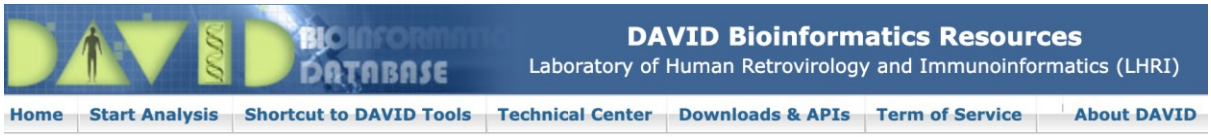
4. Notice that some log fold change is an error #N/A

NM_001181383.3	4.35E-06	1.482790675
NM_001178855.3	4.35E-06	-3.183932774
NM_001183977.1	4.38E-06	4.485657183
NM_001178379.1	4.59	#N/A
NM_001180073.1	4.60E-06	-2.322672209

   a. This is because some transcripts were completely absent in some condition (TPM = 0)
   b. For simplicity, let us remove these transcripts from consideration
5. Use the Text to Columns tool in the Data tab to extract transcript ID without .version. See image below if you cannot find the tool. Specify "." as the delimiters.
   a. In NM_001180073.1, only NM_001180073 is the transcript ID, .1 is the version number
   b. When performing analysis, only the NM_001180073 is needed

target_id	qval	log fold change
NR_132186	1.27E-06	-0.846798174
NR_132187	1.27E-06	-0.846798174
NM_001181124	1.27E-06	-1.720923198
NM_001181996	1.45E-06	1.067893183
NM_001180810	1.45E-06	0.724496369
NM_001180385	1.45E-06	0.577292558
NM_001181943	1.45E-06	0.459029393
NM_001184177	1.45E-06	-0.235505878
NM_001179305	1.45E-06	-0.291933475
NM_001182136	1.45E-06	-0.308702533
NM_001181781	1.45E-06	-0.327914251
NM_001181121	1.45E-06	-0.40426956
NM_001184001	1.45E-06	-0.42096849

6. We will extract data from this formatted table to perform functional enrichment analysis in DAVID (https://david.ncifcrf.gov/home.jsp) and WebGestalt (http://www.webgestalt.org/)



7. Just in case you have trouble using Excel. The final table can be downloaded from https://www.dropbox.com/s/prfs583aewvu3za/3000788_Fall2022_L12_demo_092222.xlsx?dl=0