Example data sheet for

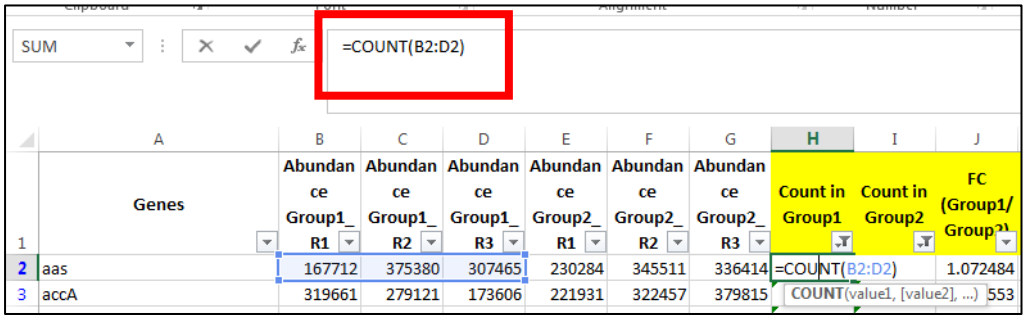
QUANTITATIVE DATA INTERPRETATION

1. The facility provides the quantitative data in the excel sheet with name of the protein identified or genes and its abundance value in the respective column as per sample. The most general format would look like the below mentioned table. In this table there was two group of sample (control vs Treated) in triplicate and facility had acquired the data using DIA (Data independent workflow) and reported the protein abundance values for each of the samples from different groups. If user has data in some other format, then user need to transform its data in the following format for analysis using this pipeline of analysis.
2. (Note: absolute abundance values required, not log transformed)

A	B	C	D	E	F	G
Genes	Abundance Group1_R1	Abundance Group1_R2	Abundance Group1_R3	Abundance Group2_R1	Abundance Group2_R2	Abundance Group2_R3
rsmF		75861.7			59158.6	75152.6
rsmG	81262.2	542694	262755	173445	266169	489787
rsmH	61023.4	120446	74140.9	54994.1	180622	230509
rsmI		292793	216863		295803	256072
rsmJ		150699	102968		153065	165143
rssB				49036.3		187844
rstA	174472	1.06E+06	518260	108080	410738	803927
rsuA	225422	290118	130653	80319.9	259841	237177
rsxC	198574	442890	254043	288612		148187
rsxG	118480	72081.7	54304.6	89666.1		60048.2
ruvA	57107.1	71061.3	61080.6		110591	52039.5
ruvB						44963
sad	231802	168840	224339		355740	261521
sapA		47088.4			115030	54220.1
sbcB	114832	125472	74786.9			139984
sbcD		54945.7	43998		89298.8	103192
sbmC	144756	163906	130291	51549.4	211601	58729.5
sbp	300425	1.10E+06	886680	492698	432399	156461
sdaA	332156	677762	524046	314457	409767	1.37E+06
sdaB	461096	264817	414467	482938	420396	614893
sdhA	1.19E+06	1.11E+06	1.99E+06	231627	315826	184849
sdhB	531509	626629	1.34E+06	117018	167540	98029.5
sdhE	518351	312523	272510	254527	101190	147702

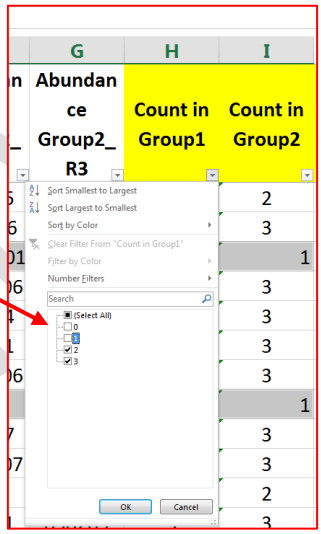
3. **Filtering of the data table:** It might be possible that all the replicate might not have the abundance value in every sample of the group so we always recommend that user should accept the proteins which have abundance value in at least 70% (2/3) of the samples in each group. To filter out those proteins/ genes (eg. rsmF, rssB) which are not having the abundance value in replicate or groups, the numbers of data points in each of the group have to be counted to filter the genes with no replicate data points by using the **COUNT** function in Excel:

4.

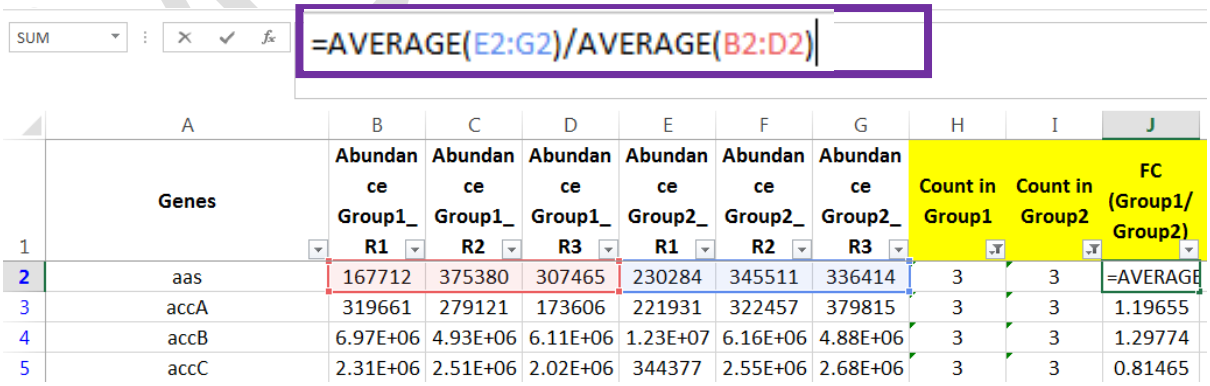


Such examples should be filtered

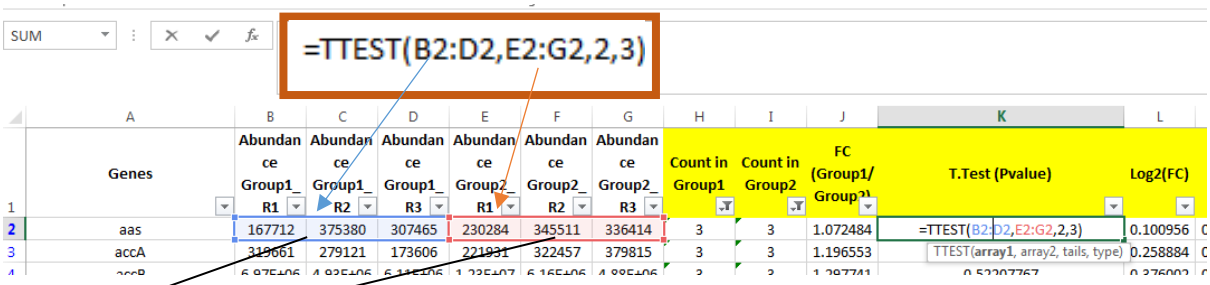
Genes	Abundance Group1_ R1	Abundance Group1_ R2	Abundance Group1_ R3	Abundance Group2_ R1	Abundance Group2_ R2	Abundance Group2_ R3	Count in Group1	Count in Group2
acpS		185495	168046		147245	160310	2	2
acrA	87987.9	46440.6	70558	49577.7	86254.6	63247.9	3	3
acs	87849.7	35622	51822.4		19701		3	1
acuI	987995	986137	1.08E+06	812161	1.15E+06	1.45E+06	3	3
add	859218	817237	813385	94681.4	585724	923337	3	3
ade	65055	117547	94185.9	24582	130191	169395	3	3
adhE	1.78E+06	1.23E+06	1.33E+06	1.43E+06	1.22E+06	1.07E+06	3	3
adhP	73996.4		55953.3	35927.6			2	1
adiA	1.32E+06	927853	1.35E+06	1.66E+06	657797	609483	3	3
adk	1.39E+07	2.03E+07	1.13E+07	1.14E+07	1.64E+07	1.62E+07	3	3
agaR	42462.6	40631.6	44860.9	110522		76164.5	3	2



5. **Fold change Calculation/ Differential expression of Protein (DEP):** To calculate the fold change expression of the protein the protein abundance values in each groups should be averaged and then the average abundance value need to be used for differential expression of Gene. In order to calculate the DEP use average abundance value of one protein in each group and divide the average abundance value from the remaining group of samples as mentioned below.

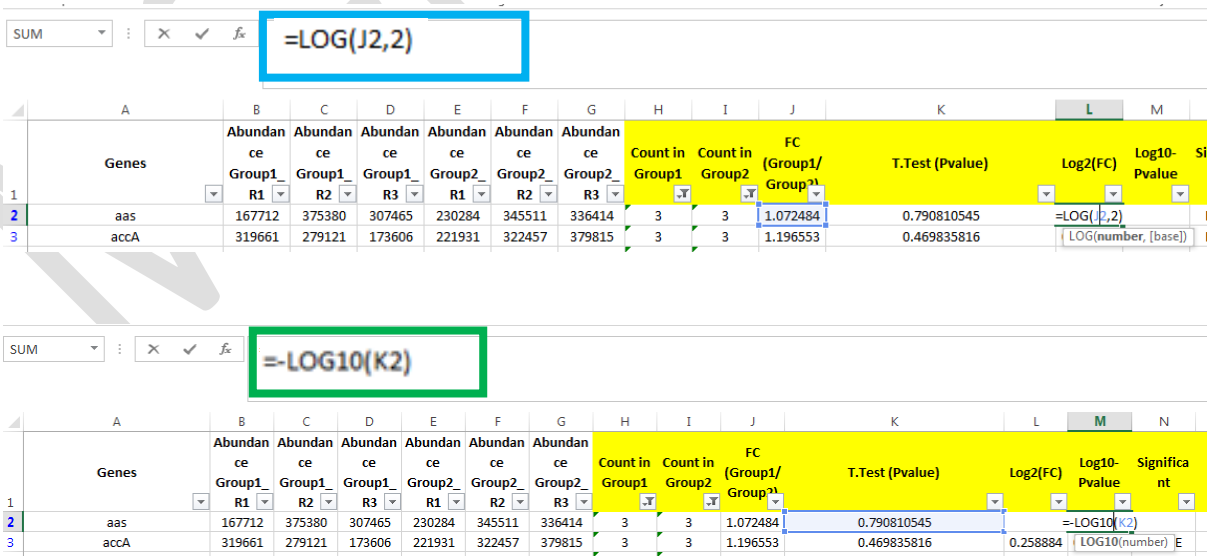


6. **Statistical Significance (P value) calculation using student t-test:** P-values were calculated using T-Test after selecting the replicate in the group and then using the formula in the formula tab.



- **Array1** – The first group data set.
- **Array2**– The second group data set.
- **Tails** – Specifies if it is a one-tailed or two-tailed test.
  - If tails = 1, T-TEST uses the one-tailed distribution.
  - If tails = 2, T-TEST uses the two-tailed distribution.
- **Type**– The type of t-test to perform:
  - Type 1: Performs a paired t-test
  - Type 2: Two-sample equal variance t-test Unpaired.
  - Type 3: Two-sample unequal variance t-test Unpaired.

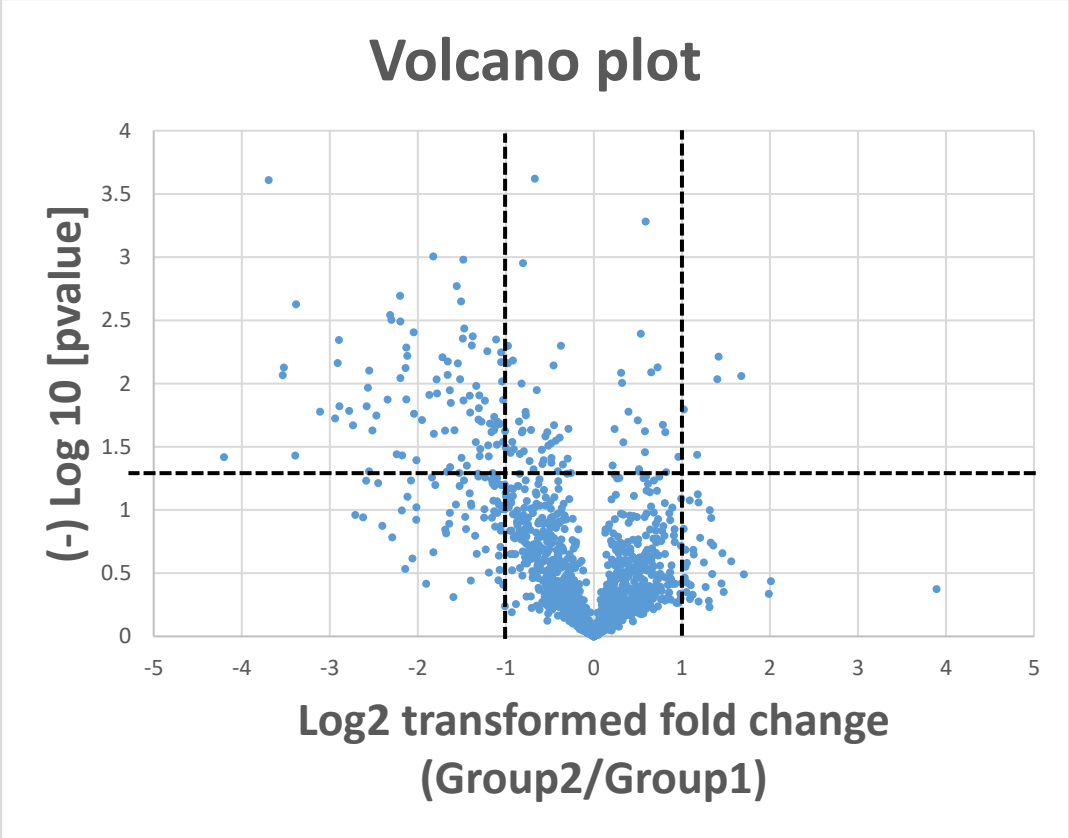
7. **Log transformation of differential expression of proteins and p-value** Log transformation of **fold change** and **p-values** were done to the base of **2** and **10** respectively. [Note: for p value (negative logarithm to base 10 is to be calculated)]. The above log transformed values were used for the generation of Volcano Plot as mentioned below. **P value cut-off of less than 0.05 was used to determine the significant proteins and log 2 fold change of ≥1 was considered as up-regulated proteins whereas log2 fold change of ≤ -1 is considered as down regulated**.



Volcano Plot for Differentially Expressed Proteins

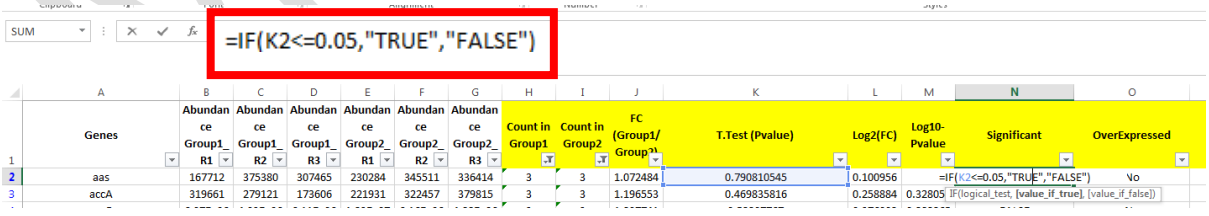
X-axis: Log2 (Fold Change) and **Y-axis: (-) Log10 [P-values]** are needed for making a volcano plot.

P value cut-off of less than 0.05 was used to determine the significant proteins and log 2 fold change of ≥1 was considered as up-regulated proteins whereas log2 fold change of ≤ -1 is considered as down regulated.



Horizontal line: pvalue cut-off ≤ 0.05, Vertical lines: Log 2 fold change cut-off ≥1 or ≤ -1 are added manually from shapes

8. The **significance** was concluded from p-value.



If **p-value<= 0.05**, **SIGNIFICANT** (**TRUE**)
If **p-value >0.05**, **NON-SIGNIFICANT** (**FALSE**)

Similar function can also be used for the UP and Down for up regulated and down regulated proteins, if the same formula is used on log2 (FC) column and cut off for up regulated and down regulated should be used as ≥1 or ≤ -1 for up regulated and down regulated proteins respectively then separate list of the DEP (UP and Down) can be generated for the GO (Gene Ontology)
GO and Heat map analysis may be performed on DEP using Shiny GO and Morpheus.

Resources

1. **ShinyGO –** For Gene Ontology and enrichment analysis
   http://bioinformatics.sdstate.edu/go/

2. **Morpheus –** For heat map and clustering analysis
   https://software.broadinstitute.org/morpheus/

3. **BoxPlotr –** For boxplot representation of data
   http://shiny.chemgrid.org/boxplotr/

4. **ClustVis –** For principal component analysis
   https://biit.cs.ut.ee/clustvis/

5. **VolcaNoseR –** For volcano plot
   https://huygens.science.uva.nl/VolcaNoseR2/

6. **ggVolcanoR –** for volcano and upset plot
   https://ggvolcanor.erc.monash.edu/

7. **Venny 2.0 –** for Venn Diagrams
   https://csbg.cnb.csic.es/BioinfoGP/venny.html