RNA-Seq Analysis: Contrast Conditions

- We are going to use “limma” as a tool to contrast expression profiles between conditions.
- Limma needs a matrix of gene counts with genes in rows and samples in columns and a table describing the conditions of each sample.
- We will use the Gene quantification level. In the transcript quantification level, the variables are not independent for the statistical analysis.
- We will get the counts matrix from the salmon quantification steps.
- Limma requieres de Counts (NumReads).
- As we have a collection of salmon outputs and we need a matrix of counts, we will process this collection to produce a matrix.

Gene Level Quantification

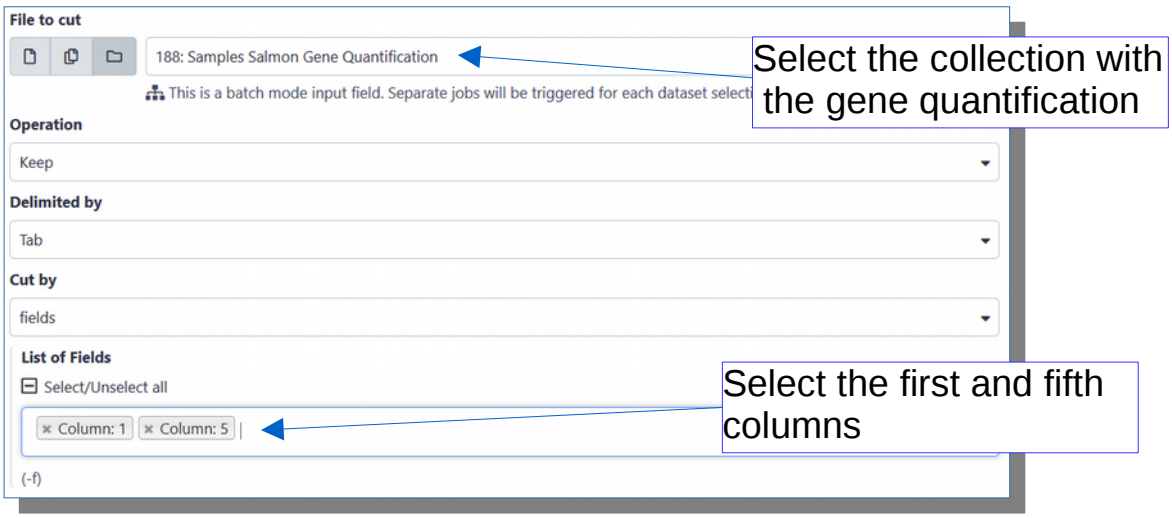
1	2	3	4	5
Name	Length	EffectiveLength	TPM	NumReads
ENSMUSG00000096768	1405.7	909.741	45.1347	1048.45
ENSMUSG00000095366	653.069	401.597	10.4436	107.093
ENSMUSG00000099399	444.5	195	0	0
ENSMUSG00000096178	844	594	0	0
ENSMUSG00000102011	437	188	0	0
ENSMUSG00000100608	931	681	0	0
ENSMUSG00000101402	967.5	718	0	0
ENSMUSG00000100492	963.5	714	0	0
ENSMUSG00000100067	967	717.5	0	0
ENSMUSG00000101984	802.5	553	0	0
ENSMUSG00000099649	397	148	0	0
ENSMUSG00000100440	968	718.5	0	0
ENSMUSG00000100930	2167	1917	0	0

RNA-Seq Analysis: Contrast Conditions

Process the Slamon Gene Counts Collection with the following tools to obtain a matrix of counts.

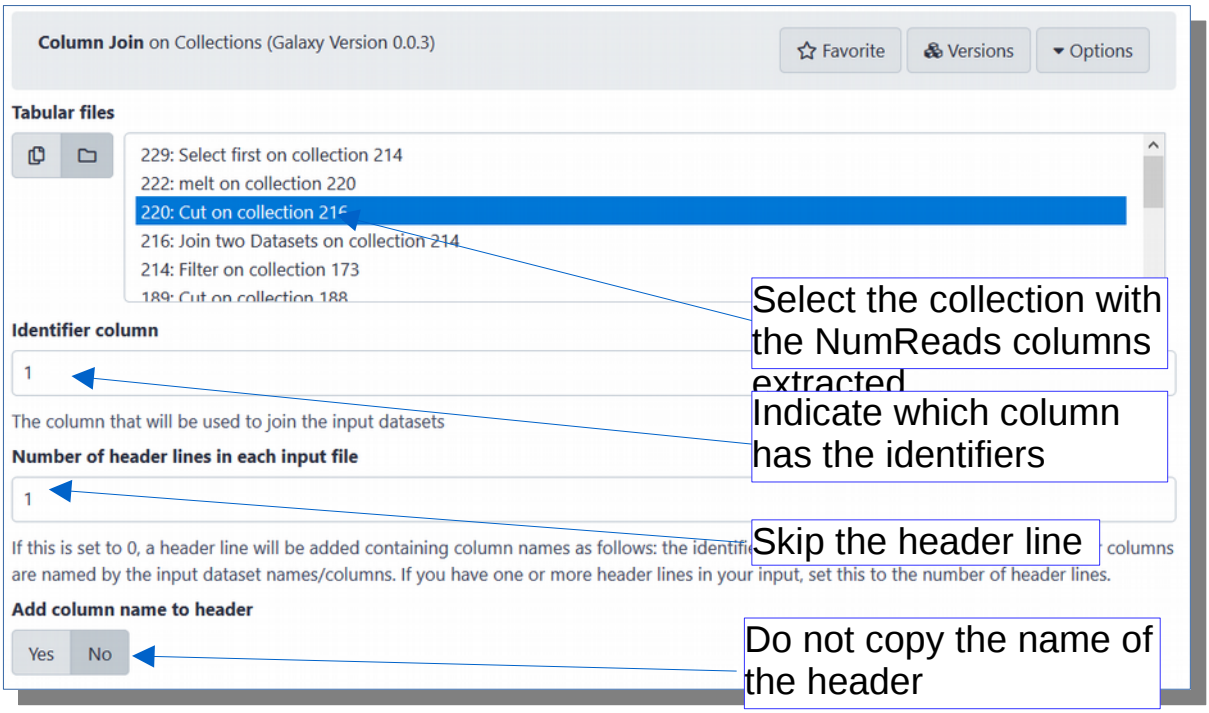
Cut: cut the 1st and 5th column

We will get the first column with the gene identifiers and the fifth column with the counts



Column Join on Collections

By indicating not to copy column names, it will add the sample name to the column

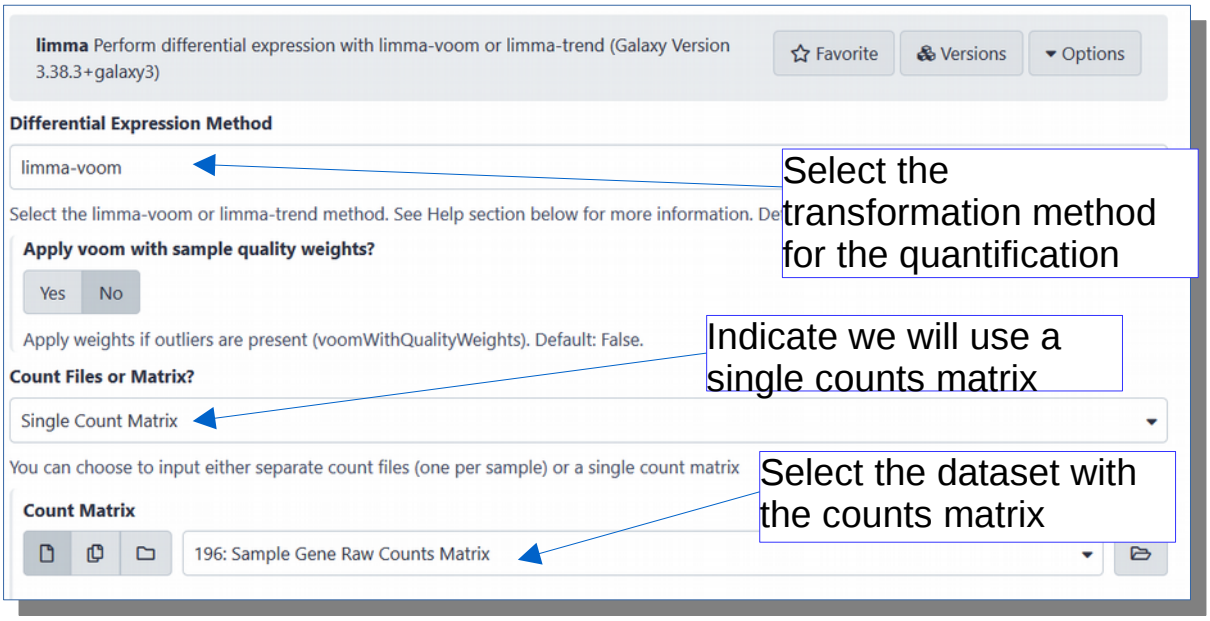


Repeat this steps for the other set if you have splitted your reads into two sets

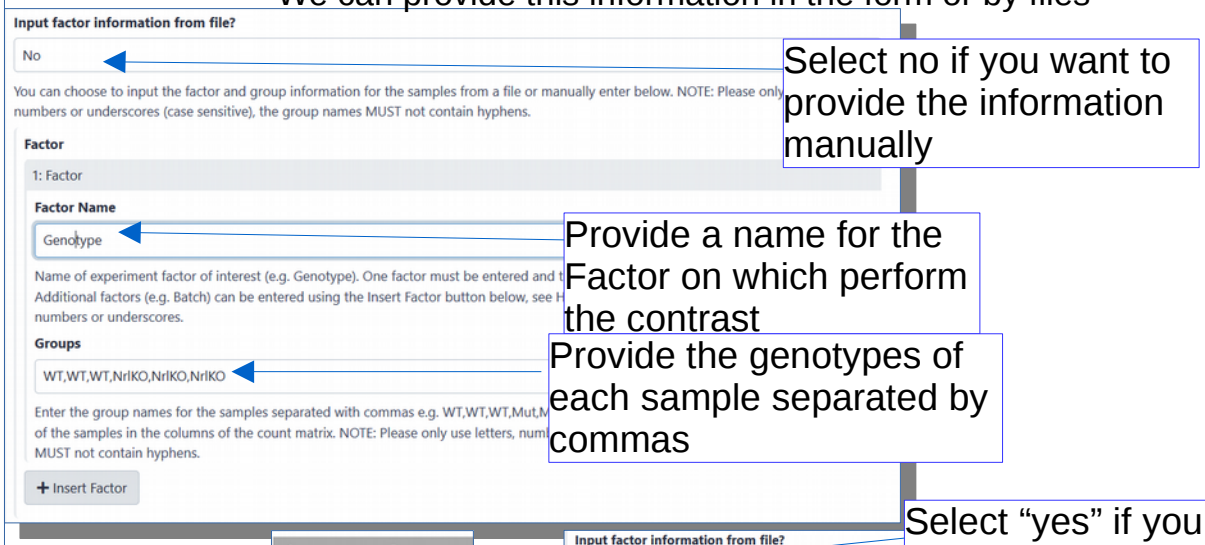
RNA-Seq Analysis: Contrast Conditions

limma: contrast between
RNASeq samples

- Limma can be used either to contrast quantifications produced by microarrays or by RNA-Seq.
- In order to normalize and transform the RNASeq data for the limma model we need to apply the voom transformation.

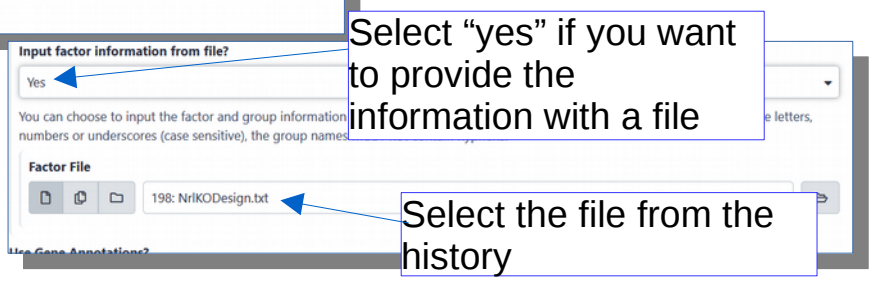


- Limma also needs the experiment design: a table with the definition of which condition is each sample.
- Also the contrasts to perform.
- We can provide this information in the form or by files



If you provide a file, this should be the format

1	2
Sample	Genotype
SRR358714	WT
SRR358715	WT
SRR358716	WT
SRR358717	NrlKO
SRR358718	NrlKO
SRR358719	NrlKO



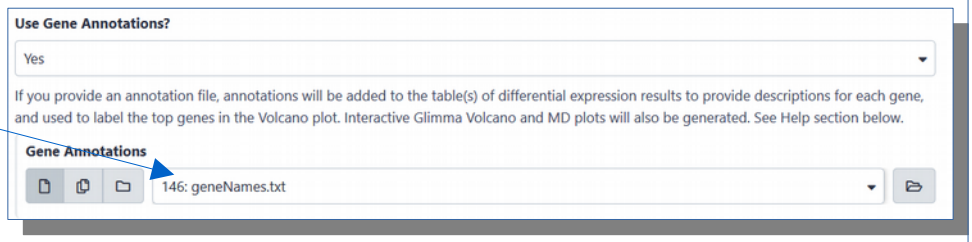
RNA-Seq Analysis: Contrast Conditions

limma: contrast between
RNASeq samples

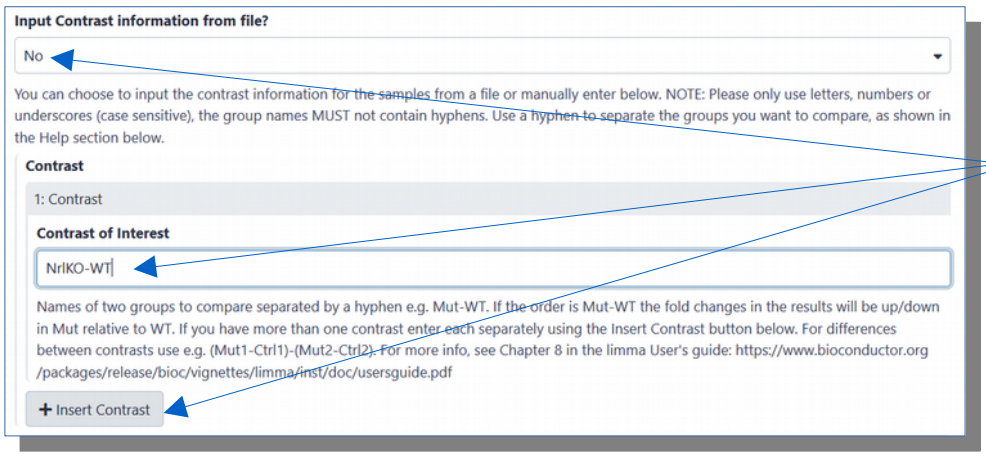
- You can provide to limma a mapping table to associate the gene IDs to their names

Gene stable ID	Gene name
ENSMUSG00000064372	mt-Tp
ENSMUSG00000064371	mt-Tt
ENSMUSG00000064370	mt-Cytb
ENSMUSG00000064369	mt-Te
ENSMUSG00000064368	mt-Nd6
ENSMUSG00000064367	mt-Nd5
ENSMUSG00000064366	mt-Tl2
ENSMUSG00000064365	mt-Ts2
ENSMUSG00000064364	mt-Th

This should be the format of the file



- With the contrast information is the same, we can provide it manually or with a file

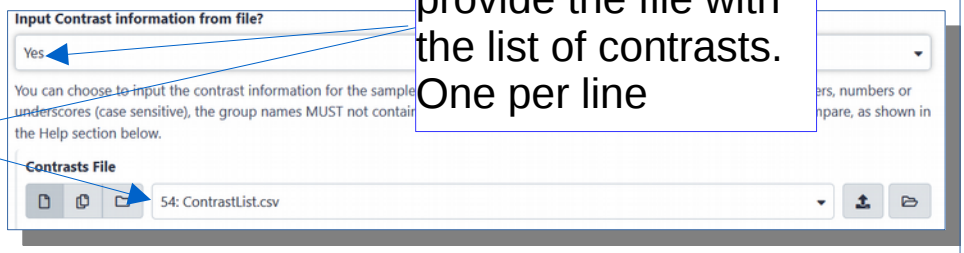


If we say no file, we will provide the contrast we want manually. We can add more contrasts if we want

If we say yes, just provide the file with the list of contrasts. One per line

If you provide a file, this should be the format

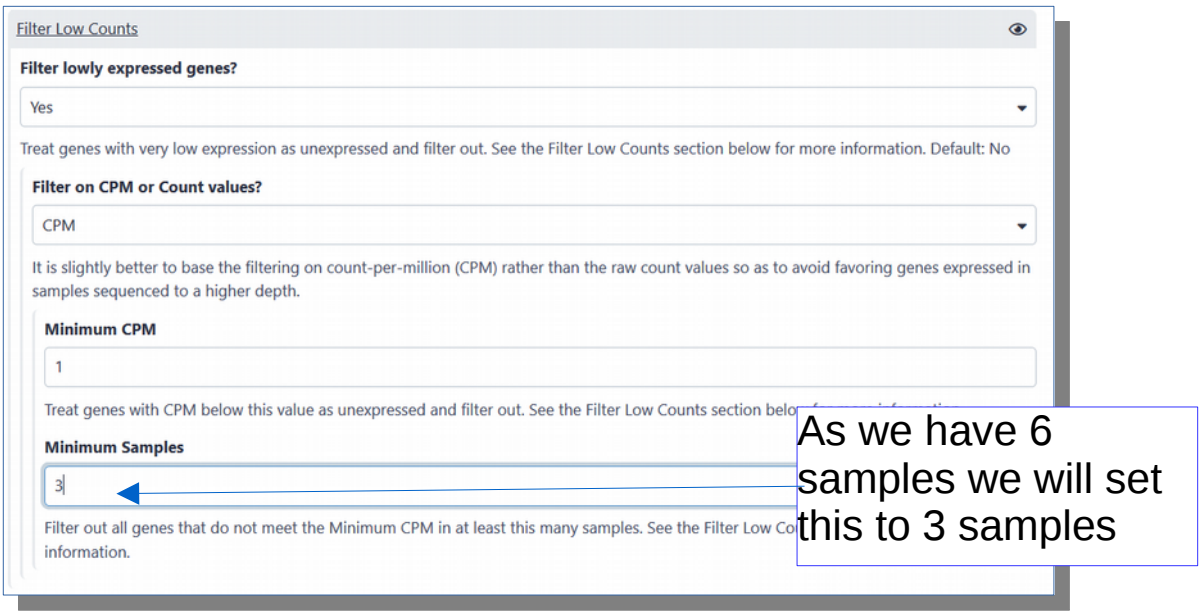
1
NrlKO-WT



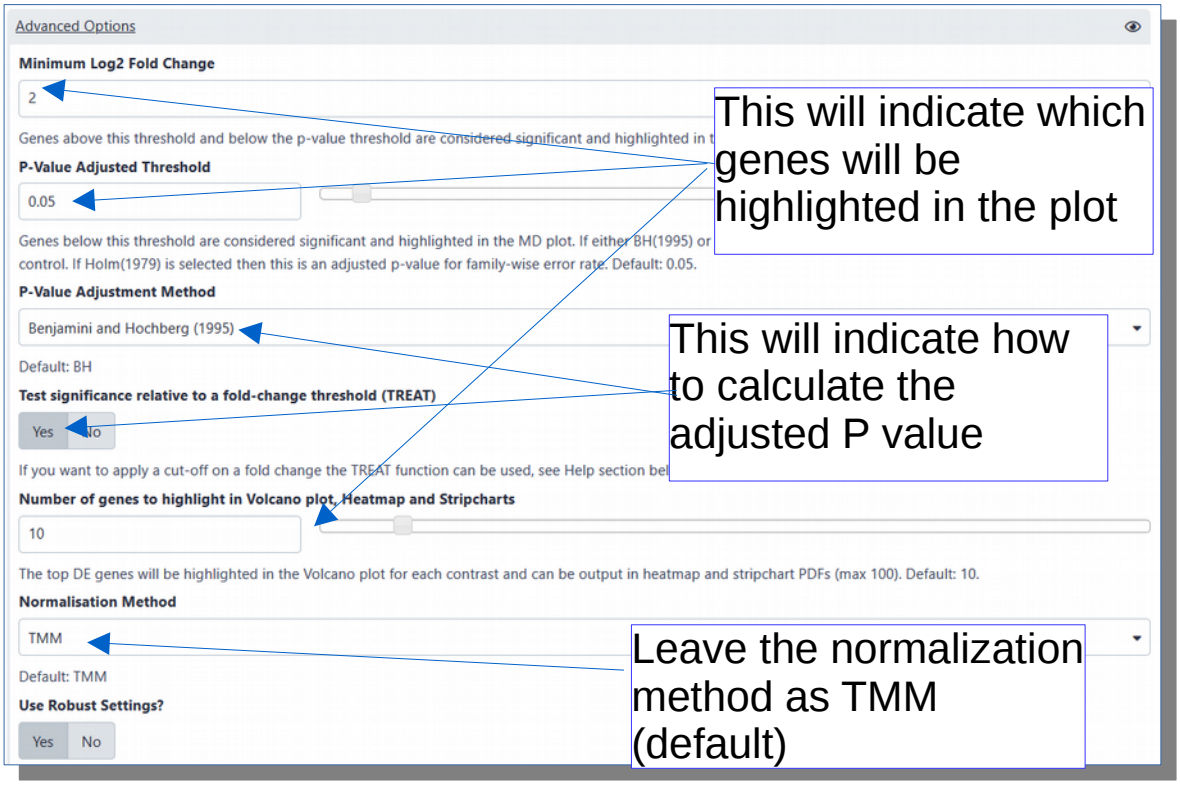
RNA-Seq Analysis: Contrast Conditions

limma: contrast between
RNASeq samples

- We can configure some aspects of the analysis:
- Usually is better to reduce our genes of interest to a list that is robustly expressed. Otherwise we will have many low expressed genes adding noise without biologial relevance.
- A standard to filter genes is to select those with at least 1 count per million reads sequenced in at least half of the samples of the experiment.



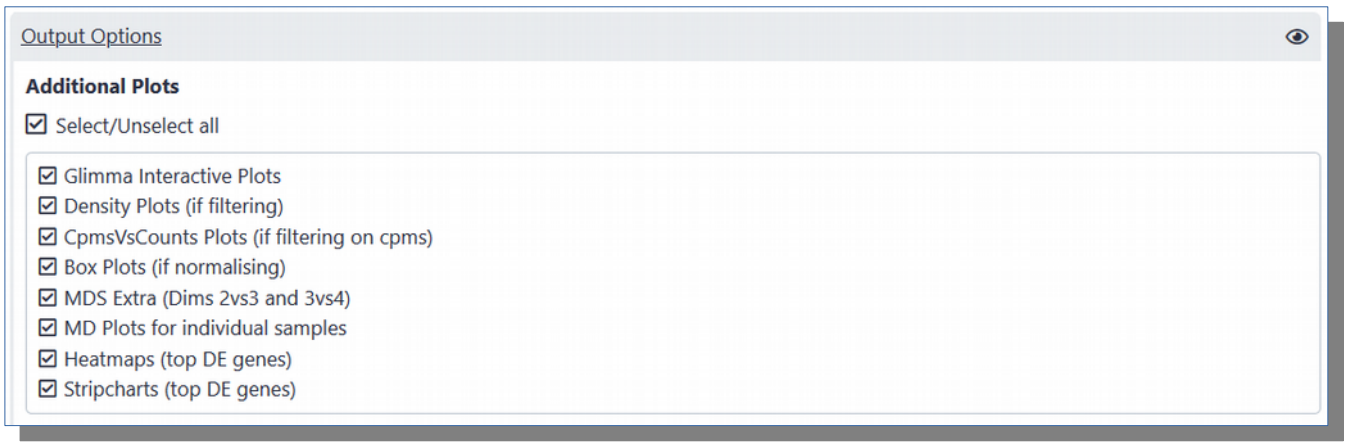
- Also we can configure how the adjusted pvalue is calculated and add some filtering parameters for the plots

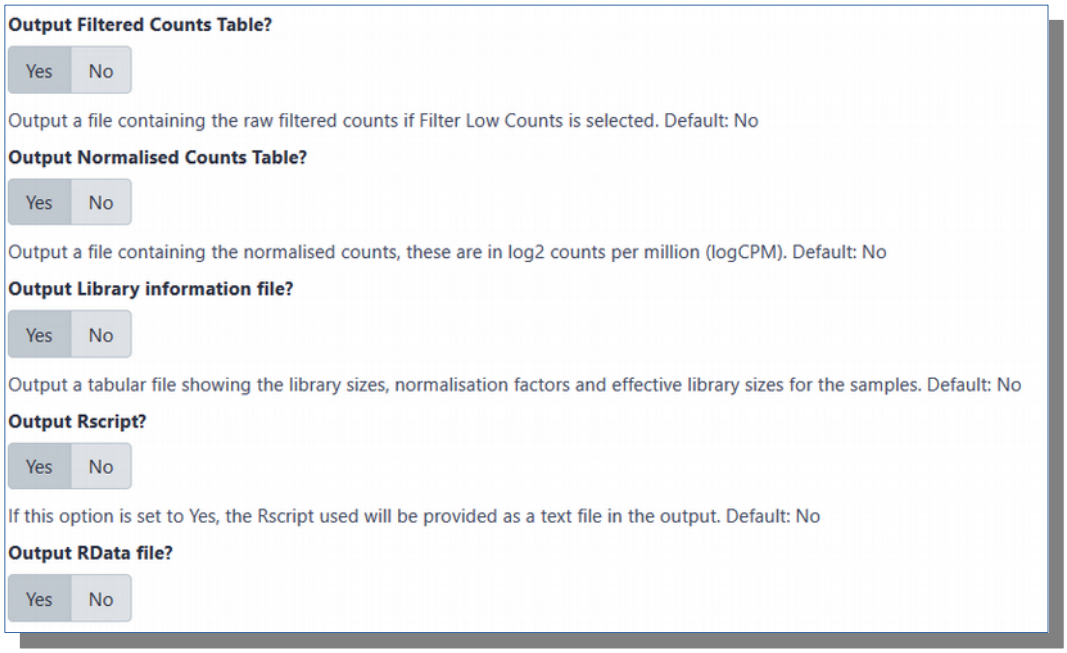


RNA-Seq Analysis: Contrast Conditions

limma: contrast between
RNASeq samples

- We can configure the outputs we want limma to provide us.
- Set everything on to explore later the results





- Limma will produce several outputs depending on our selection.
  - **Library Information**: Severa stats about each sample
  - **Rscript**: The R script used in galaxy to perform this analysis
  - **Normalised Counts**: Normalised to log2(CPMs+0.5)
  - **Filtered Counts**: Raw Counts with only the genes filtered considering de CPMs >=1 in at least 3 samples
  - **Report**: A Web page with access to all plots and stats of the analysis#
  - **DE Tables**: A collection or list of tables of the different contrasts perform.

RNA-Seq Analysis: Functional Analysis

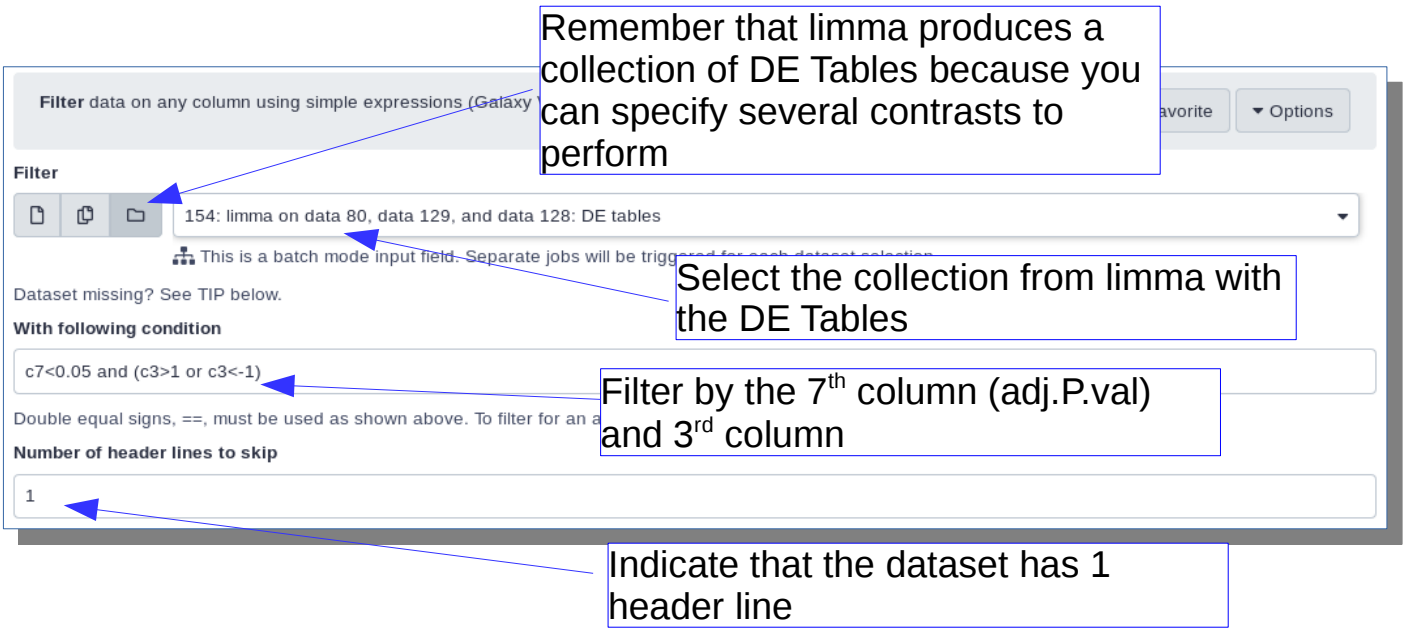
- In a functional analysis we evaluate which cellular functions are mainly in our experiment.
- Cellular functions are categories in which genes are classified.
- The main functional classification of genes is the one provided by the Gene Ontology Consortium. The categories are classified also in three main groups: Biological Processes, Molecular Functions and Cellular localization.
- We will test whether there is any functional category which is enrich in genes that we have identified as significantly changed.
- We would say a functional category is affected in our experiment if a significant number of genes (more than expected by chance) in that category are changed in our experiment.
- There are many tools and web sites that provides this type of analysis.
- We will use **GOst**, a simple tool to provide Functional enrichments.
- **GOst** needs a list of genes separated by spaces like:

  “Myc Notch1 Gbx3 sHh Tgfb Myh7 ...”
- We will need to process our DE Table from limma to get a list like that one.

RNA-Seq Analysis: Filter Results

Filter: DE genes

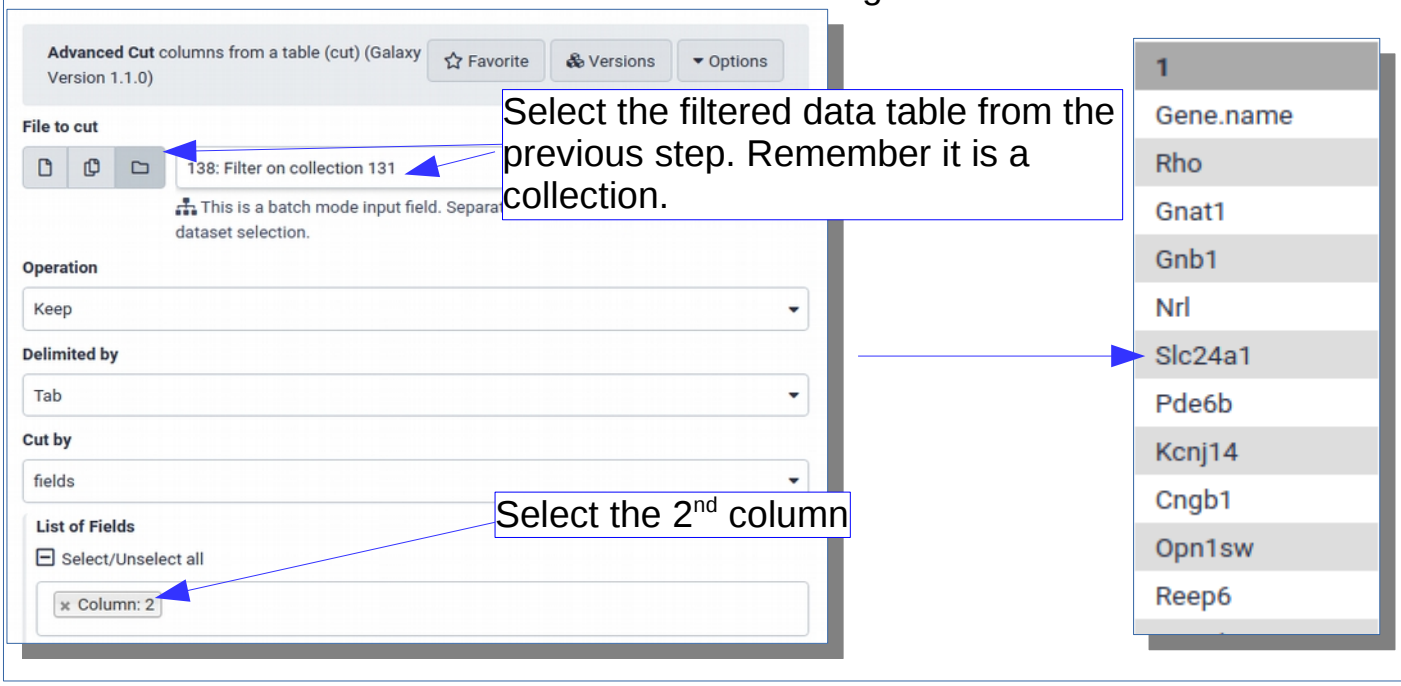
- First we are going to get our set of differentially expressed genes
- Among the results from limma there is a collection of datasets that contains DE Tables (Differentially Expressed Tables). This table contains the differentially expression test for all the genes in our experiment.
- We need to filter this table by its adj.P.val column and select only those genes with values below 0.05 and a logFC over 1 or below -1



DE Table from limma output

1	2	3	4	5	6	7
Gene.stable.ID	Gene.name	logFC	AveExpr	t	P.Value	adj.P.Val
ENSMUSG00000030324	Rho	-13.3504158270078	8.37470856912848	-102.142726750361	3.19779935757912e-17	5.43146220884814e-13
ENSMUSG00000034837	Gnat1	-11.9416955250676	7.52865763493883	-76.3274524050987	6.48749189540026e-16	5.50950249216867e-12
ENSMUSG00000040632	Nrl	-8.38851459017494	6.24345071308252	-57.4459899798456	1.2065132329481e-14	6.83087575387448e-11
ENSMUSG00000029064	Gnb1	-4.51141949661918	10.0864062101838	-47.8046262110053	8.03246897075605e-14	3.41078713670729e-10

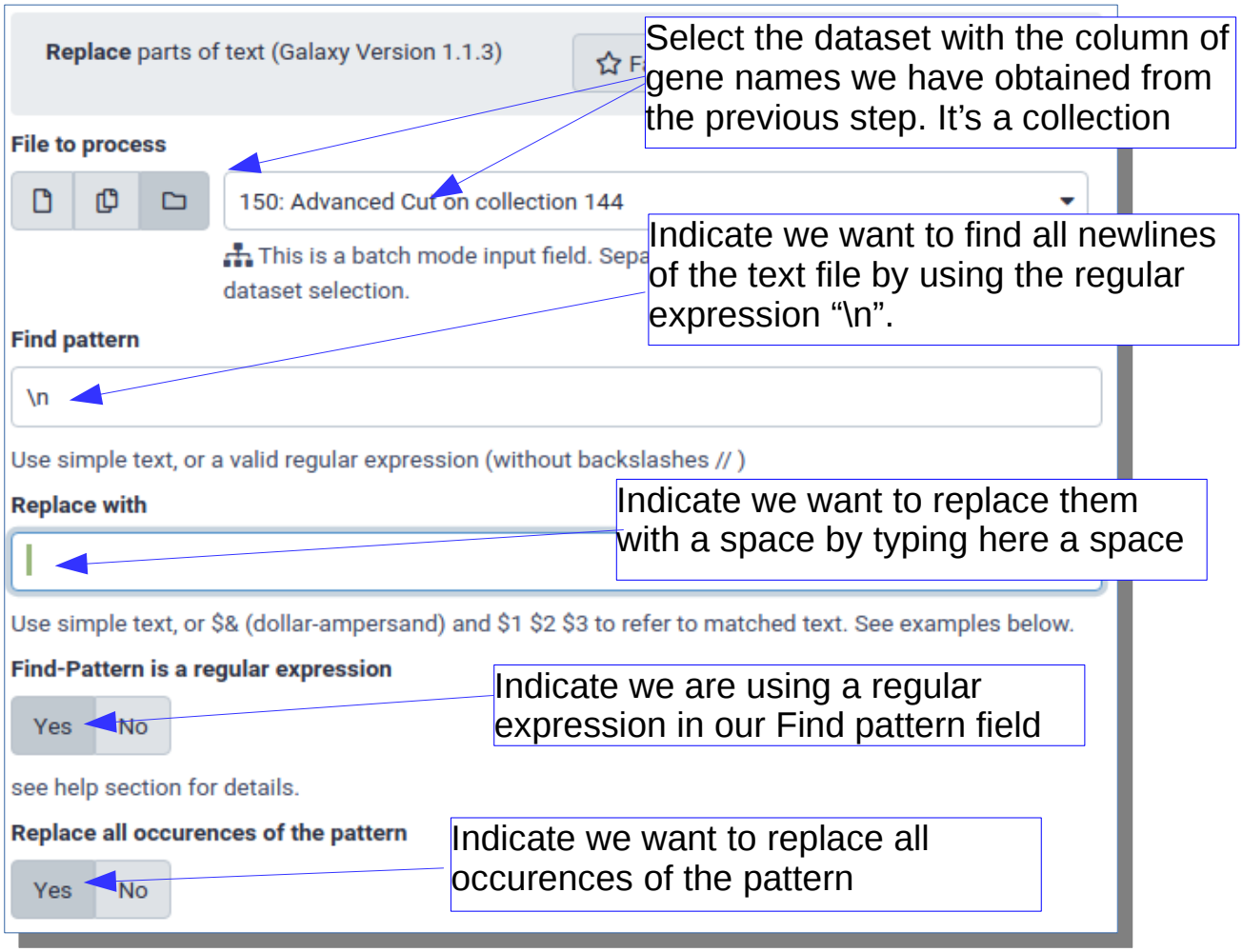
Cut: The 2nd column with the gene names



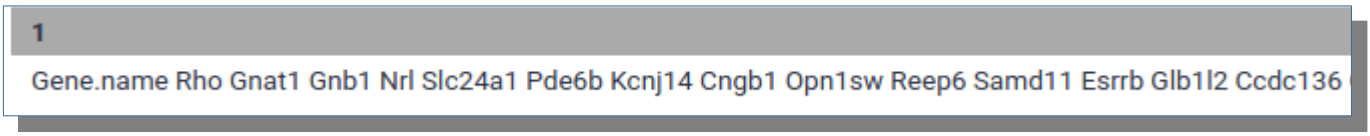
RNA-Seq Analysis: Functional Analysis

Replace: parts of text

- With this tool we can replace the “newline” symbol of a text file (defined as **\n** in regular expression format) by a space:

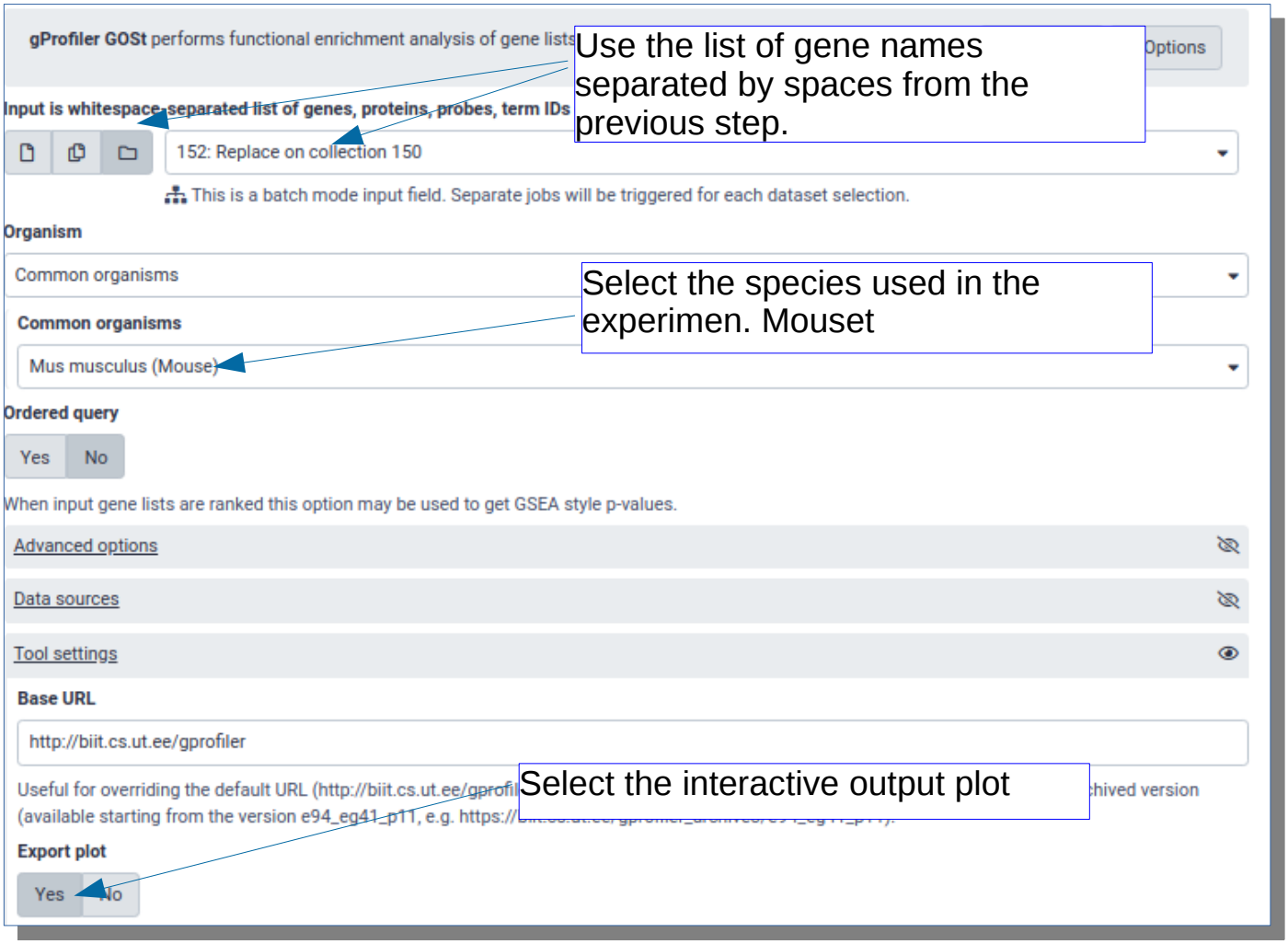


- We should obtain something like this:

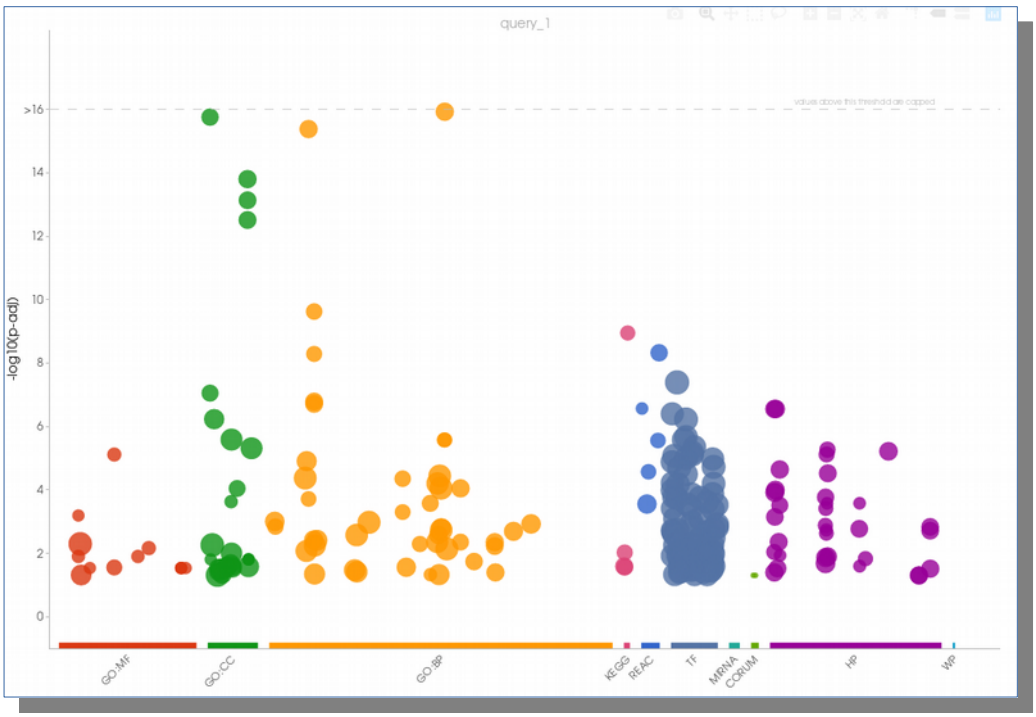


RNA-Seq Analysis: Functional Analysis

gProfiler GOSt: performs functional enrichment analysis of gene lists



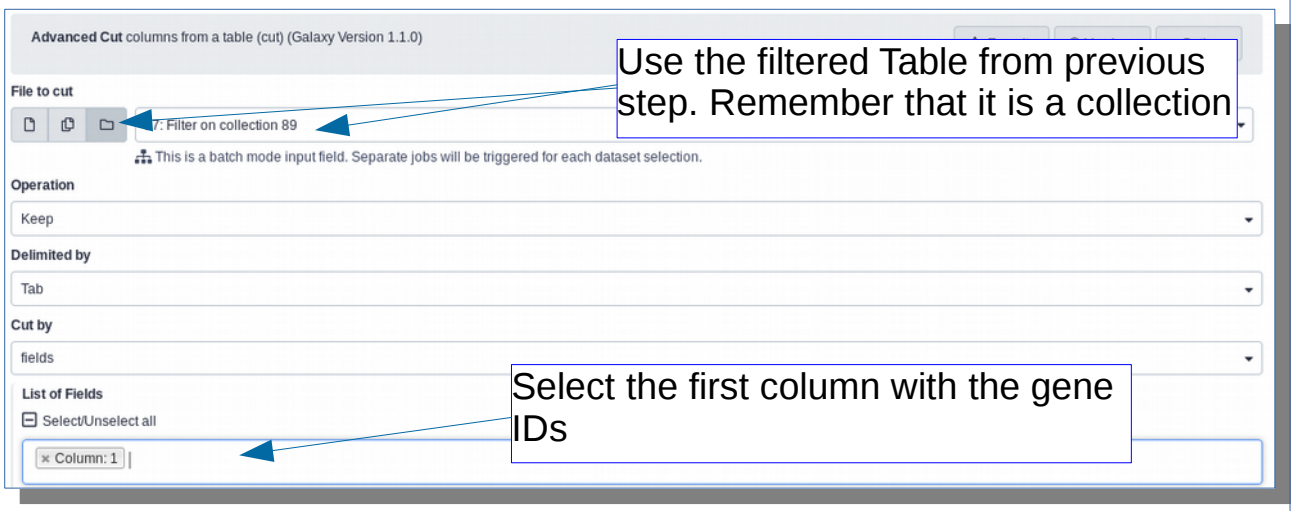
The graphical output. It is interactive and very intuitive.



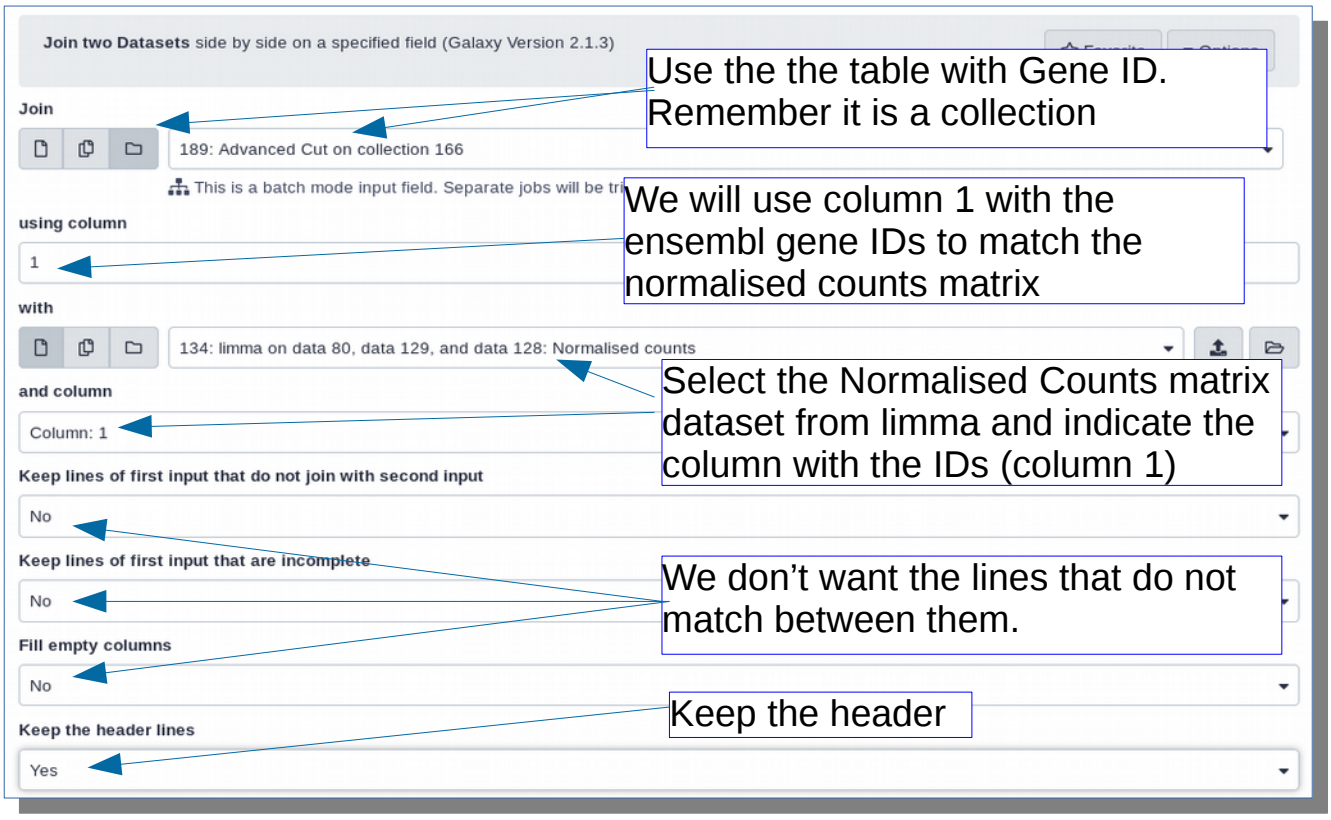
RNA-Seq Analysis: Heatmap Visualisation

- We are going to generate a clustered Heatmap with the filtered genes using the Galaxy visualisation tools.
- We need the table of Normalised Counts, but ony with the data from the filtered genes.

Cut: cut the 1st column from filtered DE Table

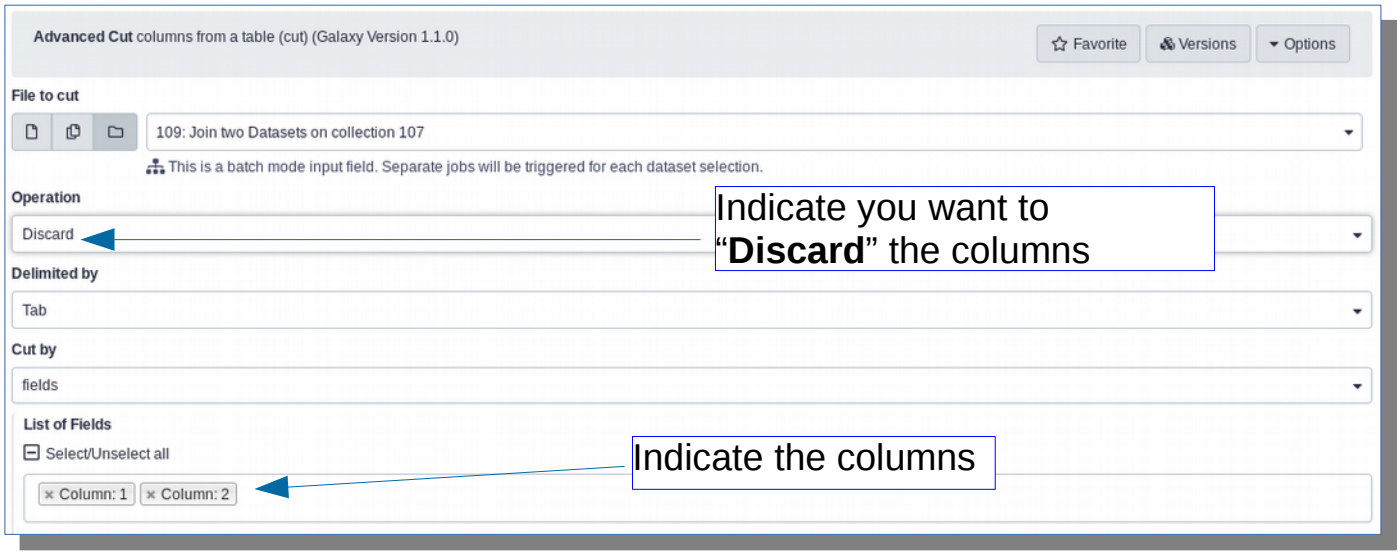


Column Join on datatsets: join filtered gene IDs with the Normalised Counts Table

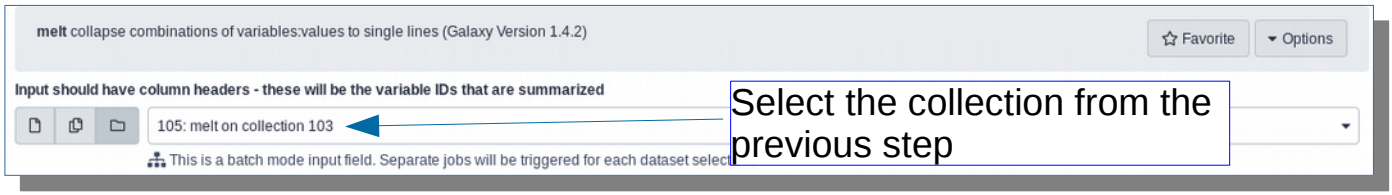


RNA-Seq Analysis: Heatmap Visualisation

Cut: cut out the 1st and 2nd column with the ensembl IDs and leave the Gene name column



Melt: transform the matrix into a 3 column table format



This is the matrix format. Genes in rows and samples in columns

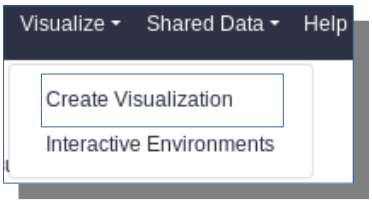
1	2	3	4	5	6	7
Gene.name	wildtype_rep1	wildtype_rep2	wildtype_rep3	NrlKO_rep1	NrlKO_rep2	NrlKO_rep3
Rho	15.0269132037375	14.9891507041218	15.1530640050294	1.57572398416494	1.74679519733432	1.71487168325769
Gnat1	13.5552440611003	13.5790980820855	13.3816890219821	1.25329763507097	1.06844448363237	1.43738120760187
Gnb1	12.3227246290013	12.3787502347082	12.3530989893968	7.8436818443039	7.8673919645391	7.80633785853636
Nrl	10.4463455756131	10.5401898123918	10.3436534274396	1.94695782161335	1.59415696145178	1.6969896103198
Pde6b	10.948750281503	11.1226820979823	10.9081569742748	6.19902952985531	6.28542932391721	6.34023523397144
Slc24a1	10.6361734902942	10.6827971055365	10.4798270097812	3.74653434620017	3.75548092520054	3.31884856189749
Ccdc136	7.26316558207711	7.1712791943573	7.22280858870968	10.5053077238051	10.4987466987438	10.4500125682061
Kcnj14	8.78548357670807	8.76400198347875	8.60792263350913	1.68320841083134	1.64576527710307	1.51503110391109
Opn1sw	8.84642925368084	8.65972113283757	8.64401135706438	12.8519994389787	12.8206812600419	12.7278401216635
Glb1l2	8.16224881015989	8.15546735302724	8.30141934479404	3.9671362598704	3.82094979741832	3.94086480705465
Cngb1	9.87753629384738	9.85879768868545	9.57921589192208	4.53916745131089	4.40269121128461	4.34296855818985
Gnat2	7.5246024159466	7.59386605577822	7.37656199482425	11.0663255821866	11.0130272361536	11.0592669697282
Reep6	9.84666550931731	9.56543928218824	9.52548969001634	4.28167134647889	4.17095134122408	4.11774315922341
Samd11	9.15327873783849	9.06402457271404	9.04477900014121	3.84087902709257	3.48781971285977	3.36956712602407
Esrrb	8.90048944466671	8.64881136001425	8.64361165647582	3.21798515438428	3.1644439100697	3.28540711334356
Tnfaip3	8.49209080439342	8.56488063937743	8.43505329996062	4.87597070298405	4.88223965373634	5.02164900358552

This is the three column format

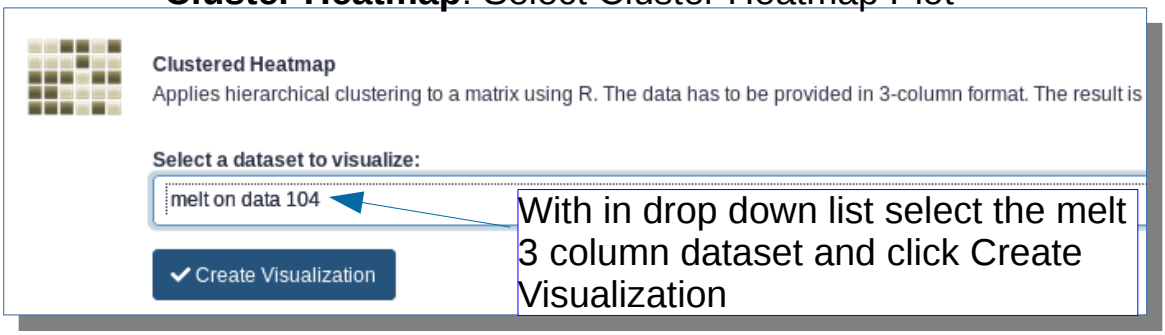
1	2	3
Gene.name	variable	value
Rho	wildtype_rep1	15.0269132037375
Gnat1	wildtype_rep1	13.5552440611003
Gnb1	wildtype_rep1	12.3227246290013
Nrl	wildtype_rep1	10.4463455756131
Pde6b	wildtype_rep1	10.948750281503
Slc24a1	wildtype_rep1	10.6361734902942
Ccdc136	wildtype_rep1	7.26316558207711
Kcnj14	wildtype_rep1	8.78548357670807
Opn1sw	wildtype_rep1	8.84642925368084
Glb1l2	wildtype_rep1	8.16224881015989
Cngb1	wildtype_rep1	9.87753629384738

RNA-Seq Analysis: Heatmap Visualisation

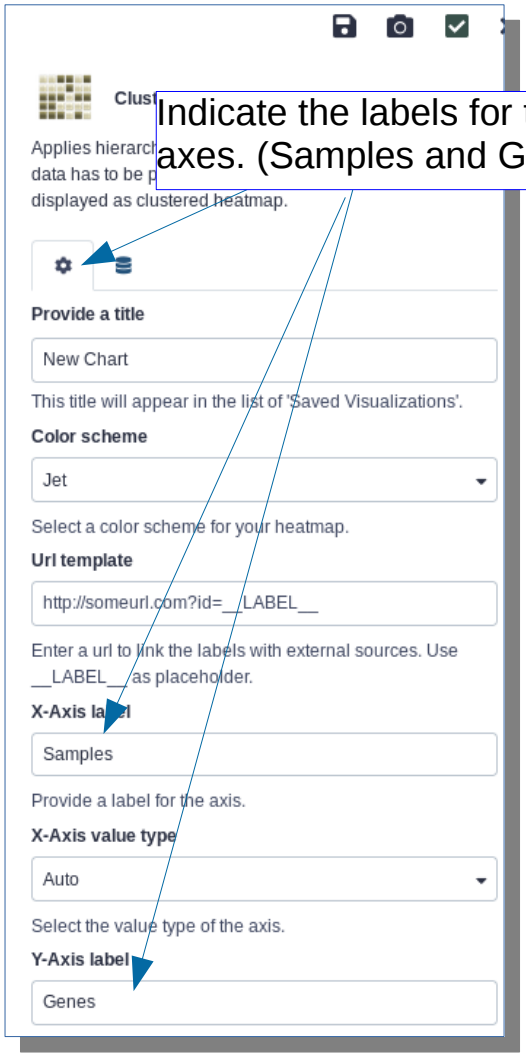
Visualization: Select Create Visualization from Visualize menu



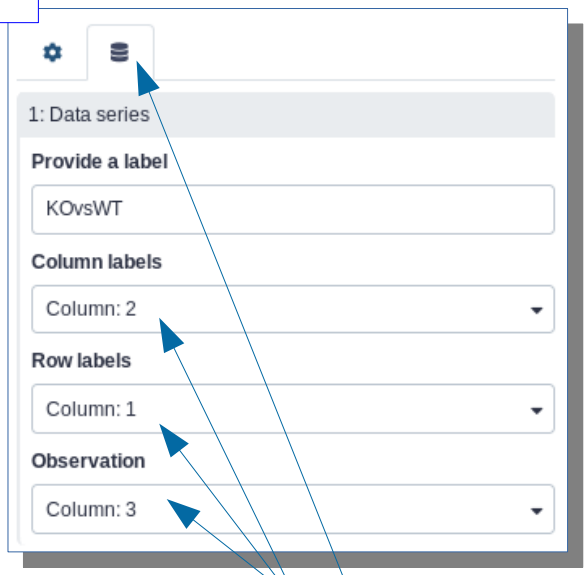
Cluster Heatmap: Select Cluster Heatmap Plot



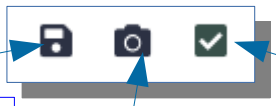
With in drop down list select the melt 3 column dataset and click Create Visualization



Indicate the labels for the X and Y axes. (Samples and Genes)



- Indicate a title for the dataset
- Which column contains the genes to place them in rows.
- Which the samples to place them in columns.
- Column 3 contains the values or observations



Save it to your visualizations.

Save it to your visualizations.

Finally click on this icon to select the settings