

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 requires 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

The screenshot shows the 'Cut' tool interface in Galaxy. The 'File to cut' field contains '188: Samples Salmon Gene Quantification'. The 'Operation' dropdown is set to 'Keep'. The 'Delimited by' dropdown is set to 'Tab'. The 'Cut by' dropdown is set to 'fields'. The 'List of Fields' section has a checkbox for 'Select/Unselect all' and two selected fields: 'Column: 1' and 'Column: 5'. Annotations with arrows point to the file name and the selected columns.

Select the collection with the gene quantification

Select the first and fifth columns

Column Join on Collections

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

The screenshot shows the 'Column Join on Collections' tool interface in Galaxy. The 'Tabular files' section lists several collections, with '220: Cut on collection 214' selected. The 'Identifier column' field is set to '1'. The 'Number of header lines in each input file' field is set to '1'. The 'Add column name to header' section has 'Yes' and 'No' buttons, with 'No' selected. Annotations with arrows point to the selected collection, the identifier column, the number of header lines, and the 'No' button.

Select the collection with the NumReads columns extracted

Indicate which column has the identifiers

Skip the header line

Do not copy the name of the header

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 Perform differential expression with limma-voom or limma-trend (Galaxy Version 3.38.3+galaxy3)

Differential Expression Method

limma-voom

Select the limma-voom or limma-trend method. See Help section below for more information. Default: limma-voom

Apply voom with sample quality weights?

Yes No

Apply weights if outliers are present (voomWithQualityWeights). Default: False.

Count Files or Matrix?

Single Count Matrix

You can choose to input either separate count files (one per sample) or a single count matrix

Count Matrix

196: Sample Gene Raw Counts Matrix

- 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

Input factor information from file?

No

You can choose to input the factor and group information for the samples from a file or manually enter below. NOTE: Please only use letters, numbers or underscores (case sensitive), the group names MUST not contain hyphens.

Factor

1: Factor

Factor Name

Genotype

Name of experiment factor of interest (e.g. Genotype). One factor must be entered and the name must be unique. Additional factors (e.g. Batch) can be entered using the Insert Factor button below, see Help section for more information.

Groups

WT,WT,WT,Nr1KO,Nr1KO,Nr1KO

Enter the group names for the samples separated with commas e.g. WT,WT,WT,Mut,Mut,Mut. The group names MUST not contain hyphens.

Insert Factor

1	2
Sample	Genotype
SRR358714	WT
SRR358715	WT
SRR358716	WT
SRR358717	Nr1KO
SRR358718	Nr1KO
SRR358719	Nr1KO

Input factor information from file?

Yes

You can choose to input the factor and group information for the samples from a file or manually enter below. NOTE: Please only use letters, numbers or underscores (case sensitive), the group names MUST not contain hyphens.

Factor File

198: Nr1KODesign.txt

Use Gene Annotations?

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

Use Gene Annotations?

Yes

If you provide an annotation file, annotations will be added to the table(s) of differential expression results to provide descriptions for each gene, and used to label the top genes in the Volcano plot. Interactive Glimma Volcano and MD plots will also be generated. See Help section below.

Gene Annotations

146: geneNames.txt

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

Input Contrast information from file?

No

You can choose to input the contrast information for the samples from a file or manually enter below. NOTE: Please only use letters, numbers or underscores (case sensitive), the group names MUST not contain hyphens. Use a hyphen to separate the groups you want to compare, as shown in the Help section below.

Contrast

1: Contrast

Contrast of Interest

NrIKO-WT

Names of two groups to compare separated by a hyphen e.g. Mut-WT. If the order is Mut-WT the fold changes in the results will be up/down in Mut relative to WT. If you have more than one contrast enter each separately using the Insert Contrast button below. For differences between contrasts use e.g. (Mut1-Ctrl1)-(Mut2-Ctrl2). For more info, see Chapter 8 in the limma User's guide: <https://www.bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf>

+ Insert Contrast

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

If you
provide a
file, this
should be
the format

1

NrIKO-WT

Input Contrast information from file?

Yes

You can choose to input the contrast information for the samples from a file or manually enter below. NOTE: Please only use letters, numbers or underscores (case sensitive), the group names MUST not contain hyphens. Use a hyphen to separate the groups you want to compare, as shown in the Help section below.

Contrasts File

54: ContrastList.csv

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

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 biological 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.

Filter Low Counts

Filter lowly expressed genes?

Yes

Treat genes with very low expression as unexpressed and filter out. See the Filter Low Counts section below for more information. Default: No

Filter on CPM or Count values?

CPM

It is slightly better to base the filtering on count-per-million (CPM) rather than the raw count values so as to avoid favoring genes expressed in samples sequenced to a higher depth.

Minimum CPM

1

Treat genes with CPM below this value as unexpressed and filter out. See the Filter Low Counts section below for more information.

Minimum Samples

3

Filter out all genes that do not meet the Minimum CPM in at least this many samples. See the Filter Low Counts section below for more information.

As we have 6 samples we will set this to 3 samples

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

Advanced Options

Minimum Log2 Fold Change

2

Genes above this threshold and below the p-value threshold are considered significant and highlighted in the MD plot.

P-Value Adjusted Threshold

0.05

Genes below this threshold are considered significant and highlighted in the MD plot. If either BH(1995) or control. If Holm(1979) is selected then this is an adjusted p-value for family-wise error rate. Default: 0.05.

P-Value Adjustment Method

Benjamini and Hochberg (1995)

Default: BH

Test significance relative to a fold-change threshold (TREAT)

Yes

If you want to apply a cut-off on a fold change the TREAT function can be used, see Help section below.

Number of genes to highlight in Volcano plot, Heatmap and Stripcharts

10

The top DE genes will be highlighted in the Volcano plot for each contrast and can be output in heatmap and stripchart PDFs (max 100). Default: 10.

Normalisation Method

TMM

Default: TMM

Use Robust Settings?

Yes

This will indicate which genes will be highlighted in the plot

This will indicate how to calculate the adjusted P value

Leave the normalization method as TMM (default)

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

Output Options

Additional Plots

☒ Select/Unselect all

☒ Glimma Interactive Plots

☒ Density Plots (if filtering)

☒ CpmVsCounts Plots (if filtering on cpm)

☒ Box Plots (if normalising)

☒ MDS Extra (Dims 2vs3 and 3vs4)

☒ MD Plots for individual samples

☒ Heatmaps (top DE genes)

☒ Stripcharts (top DE genes)

Output Filtered Counts Table?

Yes

No

Output a file containing the raw filtered counts if Filter Low Counts is selected. Default: No

Output Normalised Counts Table?

Yes

No

Output a file containing the normalised counts, these are in log2 counts per million (logCPM). Default: No

Output Library information file?

Yes

No

Output a tabular file showing the library sizes, normalisation factors and effective library sizes for the samples. Default: No

Output Rscript?

Yes

No

If this option is set to Yes, the Rscript used will be provided as a text file in the output. Default: No

Output RData file?

Yes

No

- Limma will produce several outputs depending on our selection.
 - **Library Information:** Several stats about each sample
 - **Rscript:** The R script used in galaxy to perform this analysis
 - **Normalised Counts:** Normalised to $\log_2(\text{CPMs}+0.5)$
 - **Filtered Counts:** Raw Counts with only the genes filtered considering the 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 performed.

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 enriched 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 provide 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

Remember that limma produces a collection of DE Tables because you can specify several contrasts to perform

Filter data on any column using simple expressions (Galaxy)

Filter

154: limma on data 80, data 129, and data 128: DE tables

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Dataset missing? See TIP below.

With following condition

$c7 < 0.05$ and $(c3 > 1$ or $c3 < -1)$

Double equal signs, ==, must be used as shown above. To filter for an exact value.

Number of header lines to skip

1

Indicate that the dataset has 1 header line

Select the collection from limma with the DE Tables

Filter by the 7th column (adj.P.val) and 3rd column

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	
ENSMUSG00000034159	Slc24a1	-6.8785881328893	7.11186181719338	-45.8498781858881	1.1888658118731e-13	1.83881388917388e-10	

Cut: The 2nd column with the gene names

Advanced Cut columns from a table (cut) (Galaxy Version 1.1.0)

File to cut

138: Filter on collection 131

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Operation

Keep

Delimited by

Tab

Cut by

fields

List of Fields

Select/Unselect all

Column: 2

Select the filtered data table from the previous step. Remember it is a collection.

Select the 2nd column

1

Gene.name

Rho

Gnat1

Gnb1

Nrl

Slc24a1

Pde6b

Kcnj14

Cngb1

Opn1sw

Reep6

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:

The screenshot shows the 'Replace parts of text (Galaxy Version 1.1.3)' tool interface. Annotations with arrows point to specific fields:

- File to process:** A dropdown menu showing '150: Advanced Cut on collection 144'. An annotation box says: 'Select the dataset with the column of gene names we have obtained from the previous step. It's a collection'.
- Find pattern:** A text input field containing the regular expression `\n`. An annotation box says: 'Indicate we want to find all newlines of the text file by using the regular expression “\n”'.
- Replace with:** A text input field containing a single space character. An annotation box says: 'Indicate we want to replace them with a space by typing here a space'.
- Find-Pattern is a regular expression:** Radio buttons for 'Yes' and 'No'. The 'Yes' button is selected. An annotation box says: 'Indicate we are using a regular expression in our Find pattern field'.
- Replace all occurrences of the pattern:** Radio buttons for 'Yes' and 'No'. The 'Yes' button is selected. An annotation box says: 'Indicate we want to replace all occurrences of the pattern'.

Below the tool interface, a preview of the output is shown:

- We should obtain something like this:

1
Gene.name Rho Gnat1 Gnb1 Nrl Slc24a1 Pde6b Kcnj14 Cngb1 Opm1sw Reep6 Samd11 Esrrb Glb1l2 Ccdc136

RNA-Seq Analysis: Functional Analysis

gProfiler GOST: performs functional enrichment analysis of gene lists

gProfiler GOST performs functional enrichment analysis of gene lists Options

Input is whitespace-separated list of genes, proteins, probes, term IDs

152: Replace on collection 150

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Organism

Common organisms

Common organisms

Mus musculus (Mouse)

Ordered query

Yes No

When input gene lists are ranked this option may be used to get GSEA style p-values.

[Advanced options](#)

[Data sources](#)

[Tool settings](#)

Base URL

http://biit.cs.ut.ee/gprofiler

Useful for overriding the default URL (http://biit.cs.ut.ee/gprofiler) (available starting from the version e94_eg41_p11, e.g. https://biit.cs.ut.ee/gprofiler/eg41_p11)

Export plot

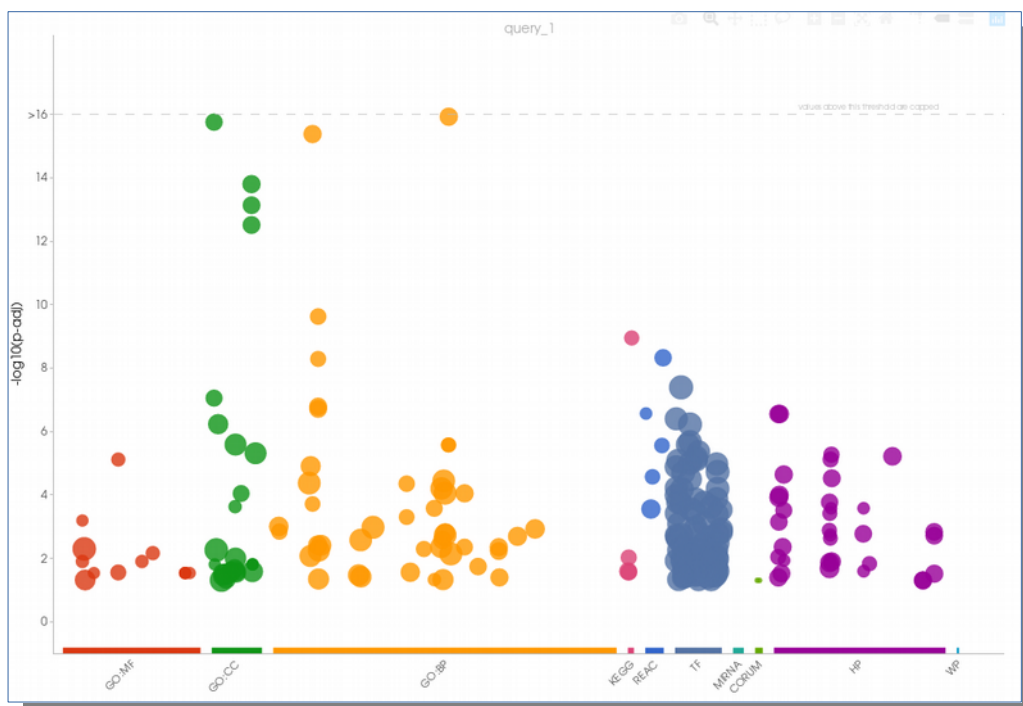
Yes No

Select the list of gene names separated by spaces from the previous step.

Select the species used in the experiment. Mouset

Select the interactive output plot

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 only with the data from the filtered genes.

Cut: cut the 1st column from filtered DE Table

The screenshot shows the 'Advanced Cut columns from a table (cut)' tool in Galaxy Version 1.1.0. The 'File to cut' section has a dropdown menu showing '7: Filter on collection 89'. The 'Operation' is set to 'Keep'. The 'Delimited by' is set to 'Tab'. The 'Cut by' is set to 'fields'. The 'List of Fields' section has a checkbox for 'Select/Unselect all' and a text input field containing 'x Column: 1'. Annotations with arrows point to the 'File to cut' dropdown, the 'List of Fields' section, and the 'x Column: 1' input field.

Use the filtered Table from previous step. Remember that it is a collection

Select the first column with the gene IDs

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

The screenshot shows the 'Join two Datasets side by side on a specified field' tool in Galaxy Version 2.1.3. The 'Join' section has a dropdown menu showing '189: Advanced Cut on collection 166'. The 'using column' is set to '1'. The 'with' section has a dropdown menu showing '134: limma on data 80, data 129, and data 128: Normalised counts'. The 'and column' is set to 'Column: 1'. The 'Keep lines of first input that do not join with second input' is set to 'No'. The 'Keep lines of first input that are incomplete' is set to 'No'. The 'Fill empty columns' is set to 'No'. The 'Keep the header lines' is set to 'Yes'. Annotations with arrows point to the 'Join' dropdown, the 'using column' input, the 'with' dropdown, the 'and column' input, the 'Keep lines of first input that do not join with second input' dropdown, the 'Keep lines of first input that are incomplete' dropdown, the 'Fill empty columns' dropdown, and the 'Keep the header lines' dropdown.

Use the the table with Gene ID. Remember it is a collection

We will use column 1 with the ensembl gene IDs to match the normalised counts matrix

Select the Normalised Counts matrix dataset from limma and indicate the column with the IDs (column 1)

We don't want the lines that do not match between them.

Keep the header

RNA-Seq Analysis: Heatmap Visualisation

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

Advanced Cut columns from a table (cut) (Galaxy Version 1.1.0) Favorite Versions Options

File to cut

109: Join two Datasets on collection 107

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Operation

Discard

Indicate you want to “Discard” the columns

Delimited by

Tab

Cut by

fields

List of Fields

Select/Unselect all

Column: 1

Column: 2

Indicate the columns

Melt: transform the matrix into a 3 column table format

melt collapse combinations of variables/values to single lines (Galaxy Version 1.4.2) Favorite Options

Input should have column headers - these will be the variable IDs that are summarized

105: melt on collection 103

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Select the collection from the previous step

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

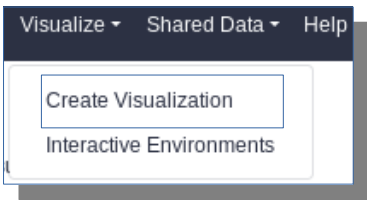
This is the three column format

1	2	3	4	5	6	7
Gene name	wildtype_rep1	wildtype_rep2	wildtype_rep3	NiKO_rep1	NiKO_rep2	NiKO_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
Esrnb	8.90048944466671	8.64881136001425	8.64361165647582	3.21798515438428	3.1644439100697	3.28540711334356
Tnfrsf3	8.49209080439342	8.56488063937743	8.4350532999062	4.87597070298405	4.88223965373634	5.02164900358552


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




Clustered Heatmap
Applies hierarchical clustering to a matrix using R. The data has to be provided in 3-column format. The result is

Select a dataset to visualize:

melt on data 104

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

✓ Create Visualization



Clustered Heatmap
Applies hierarchical clustering to a matrix using R. The data has to be provided in 3-column format. The result is displayed as clustered heatmap.

Provide a title

New Chart

This title will appear in the list of 'Saved Visualizations'.

Color scheme

Jet

Select a color scheme for your heatmap.

Uri template

http://someurl.com?id=__LABEL__

Enter a url to link the labels with external sources. Use __LABEL__ as placeholder.

X-Axis label

Samples

Provide a label for the axis.

X-Axis value type

Auto

Select the value type of the axis.

Y-Axis label

Genes

1: Data series

Provide a label

KOvsWT

Column labels

Column: 2

Row labels

Column: 1

Observation

Column: 3

- 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