TECHNICAL REPORTS

VISUALIZATION OF RNA-SEQ RESULTS WITH VOLCANO PLOT

Disusun untuk memenuhi tugas UTS mata kuliah Bioinformatika Lanjut

Dosen : Dr. Risman Adnan, S.Si., M.Si.

Muhammad Ridho (2206130776)

PROGRAM MAGISTER MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS INDONESIA

DEPOK

2023

Contents

1. Introduction

Volcano plots are commonly used to display the results of RNA-seq or other omics experiments. A volcano plot is a type of scatterplot that shows statistical significance (P value) versus magnitude of change (fold change). It enables quick visual identification of genes with large fold changes that are also statistically significant. These may be the most biologically significant genes. In a volcano plot, the most upregulated genes are towards the right, the most downregulated genes are towards the left, and the most statistically significant genes are towards the top.

To generate a volcano plot of RNA-seq results, we need a file of differentially expressed results which is provided for you here. To generate this file yourself, see the RNA-seq counts to genes tutorial. The file used here was generated from limma-voom but you could use a file from any RNA-seq differential expression tool, such as edgeR or DESeq2, as long as it has the required columns (see below).

The data for this tutorial comes from Fu et al. 2015. This study examined the expression profiles of basal and luminal cells in the mammary gland of virgin, pregnant and lactating mice. Here we will visualize the results of the luminal pregnant vs lactating comparison.

2. Import Data

We will use two files for this analysis:

- **Differentially expressed results file** (genes in rows, and 4 required columns: raw P values, adjusted P values (FDR), log fold change and gene labels)
- **Genes of interest file** (list of genes to be plotted in volcano)

Hands-on: Data Upload

1. Create and name a new history for this tutorial.
   Click the + icon at the top of the history panel.
   If the + is missing:
   - Click on the ⚙ icon (**History options**) on the top of the history panel.
   - Select the option **Create New** from the menu.
2. Import the FASTQ file pairs from Zenodo or a data library:
   https://zenodo.org/record/2529117/files/limma-voom_luminalpregnant-luminallactate
   https://zenodo.org/record/2529117/files/volcano_genes
   - Copy the link location.
   - Click ⤒ **Upload Data** at the top of the tool panel.
   - Select ✎ **Paste/Fetch Data**.
   - Paste the link(s) into the text field
   - Select *"Type":* **tabular**
   - Press Start
   - Close the window

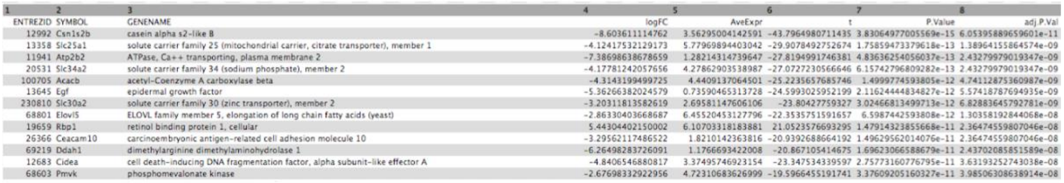
   As an alternative to uploading the data from a URL or your computer, the files may also have been made available from a *shared data library*:

- Go into **Shared data** (top panel) then **Data libraries**
- Navigate to the correct folder as indicated by your instructor. On most Galaxies tutorial data will be provided in a folder named **GTN - Material –> Topic Name -> Tutorial Name**.
- Select the desired files
- Click on **Add to History** near the top and select **as Datasets** from the dropdown menu
- In the pop-up window, choose *“Select history”:* the history you want to import the data to (or create a new one)
- Click on Import

3. Change the datatype from **fastqsanger** to **fastq.**

- Click on the **pencil icon** for the dataset to edit its attributes
- In the central panel, click on the **Convert** tab on the top
- In the lower part **Datatypes**, select **tabular**
- Click the **Save** button

4. Click on the (eye) icon and take a look at the limma-voom file. It should look like below, with 8 columns.



1	2	3	4	5	6	7	8
ENTREZID	SYMBOL	GENENAME	logFC	AveExpr	t	P.Value	adj.P.Val
12992	Csn1s2b	casein alpha s2-like B	-8.603611114762	3.56295004142591	-43.7964980711435	3.83064977005569e-15	6.05395889659601e-11
13358	Slc25a1	solute carrier family 25 (mitochondrial carrier, citrate transporter), member 1	-4.12417532129173	5.77969894403042	-29.9078492752674	1.75859473379618e-13	1.38964155864574e-09
11941	Atp2b2	ATPase, Ca++ transporting, plasma membrane 2	-7.38698638678659	1.28214314739647	-27.8194991746381	4.83636254056037e-13	2.43279979019347e-09
20531	Slc34a2	solute carrier family 34 (sodium phosphate), member 2	-4.17781242057656	4.27862903538987	-27.0727230566646	6.15742796809282e-13	2.43279979019347e-09
100705	Acacb	acetyl-Coenzyme A carboxylase beta	-4.3143199499725	4.4409137064501	-25.2235657685746	1.4999774593805e-12	4.74112875360987e-09
13645	Egf	epidermal growth factor	-5.36266382024579	0.73590465313728	-24.5993025952199	2.11624444834827e-12	5.57418787694935e-09
230810	Slc30a2	solute carrier family 30 (zinc transporter), member 2	-3.20311813582619	2.69581147606106	-23.80427759327	3.02466813499713e-12	6.82883645792781e-09
68801	Elovl5	ELOVL family member 5, elongation of long chain fatty acids (yeast)	-2.86330403668687	6.45520453127796	-22.3535751591657	6.5987442593808e-12	1.30358192844068e-08
19659	Rbp1	retinol binding protein 1, cellular	5.44304402150002	6.10703318183881	21.0523576693295	1.47914323855668e-11	2.36474559807046e-08
26366	Ceacam10	carcinoembryonic antigen-related cell adhesion molecule 10	-3.29562117486522	1.8210142363816	-20.9392688664192	1.49629562014076e-11	2.36474559807046e-08
69219	Ddah1	dimethylarginine dimethylaminohydrolase 1	-6.26498283726091	1.1766693422008	-20.867105414675	1.69623066588679e-11	2.43702085851589e-08
12683	Cidea	cell death-inducing DNA fragmentation factor, alpha subunit-like effector A	-4.8406546880817	3.37495746923154	-23.347534339597	2.75773160776795e-11	3.63193252743038e-08
68603	Pmvk	phosphomevalonate kinase	-2.67698332922956	4.72310683626999	-19.5966455191741	3.37609205160327e-11	3.98506308638914e-08

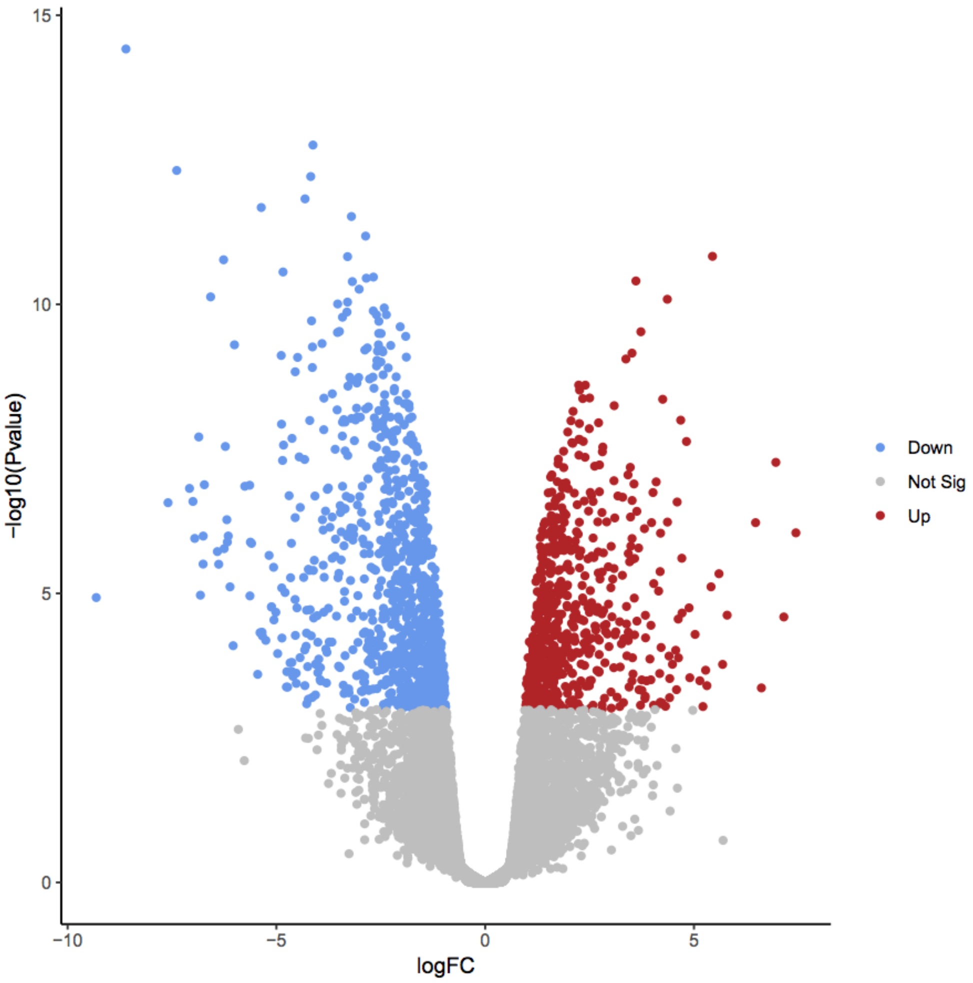
3. Create Volcano Plot Highlighting Significant Genes

First we will create a volcano plot highlighting all significant genes. We will call genes significant here if they have FDR < 0.01 and a log fold change of 0.58 (equivalent to a fold-change of 1.5). These were the values used in the original paper for this dataset.

Hands-on: Create a Volcano plot

1. **FastQC** (Galaxy version 0.0.5) to create a volcano plot:

- “*Specify an input file*”: limma-voom file
- “*FDR (adjusted P value)*”: **Column 8**
- “*P value (raw)*”: **Column 7**
- “*Log Fold Change*”: **Column 4**
- “*Labels*”: **Column 2**
- “*Significance threshold*”: **0.01**
- “*LogFC threshold to colour*”: **0.58**
- “*Points to label*”: **None**



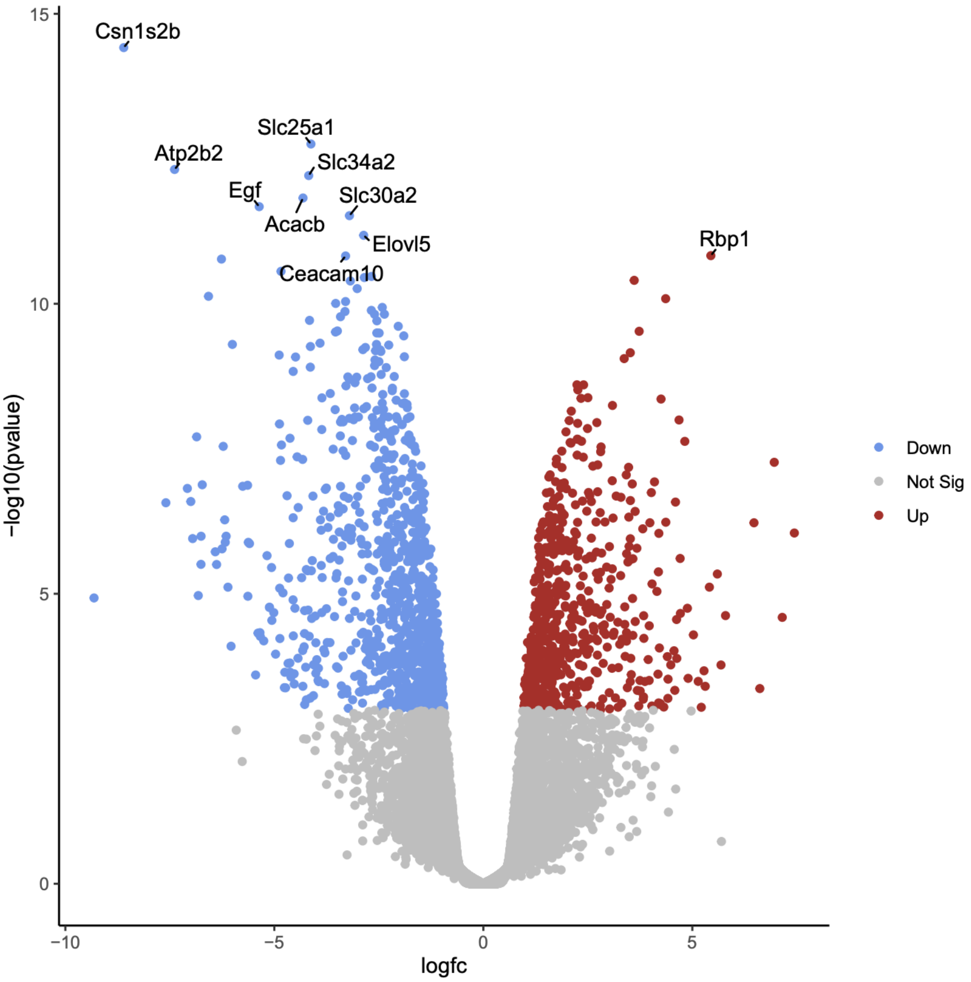
In the plot above the genes are coloured if they pass the thresholds for FDR and Log Fold Change, red if they are upregulated and blue if they are downregulated. You can see in this plot that there are many (hundreds) of significant genes in this dataset.

4. Create Volcano Plot Labelling Top Significant Genes

You can also choose to show the labels (e.g. Gene Symbols) for the significant genes with this volcano plot tool. You can select to label all significant or just the top genes. The top genes are those that pass the FDR and logFC thresholds that have the smallest P values. As there are hundreds of significant genes here, too many to sensibly label, let’s label the top 10 genes.

Hands-on: Create a Volcano plot labelling top genes

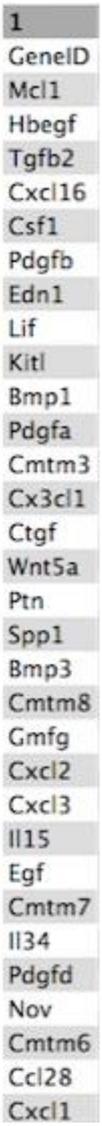
1. Use the **Rerun** ⟳ button in the History to rerun **Volcano Plot** 🔧 with the same parameters as before except:
   - ▼ “*Points to label*”: **Significant**
     - ✎ “*Only label top most significant*”: **10**



As in the previous plot, genes are coloured if they pass the thresholds for FDR and Log Fold Change, (red for upregulated and blue for downregulated) and the top genes by P value are labelled. Note that in the plot above we can now easily see what the top genes are by P value and also which of them have bigger fold changes.

5. Create Volcano Plot Labelling Genes of Interest

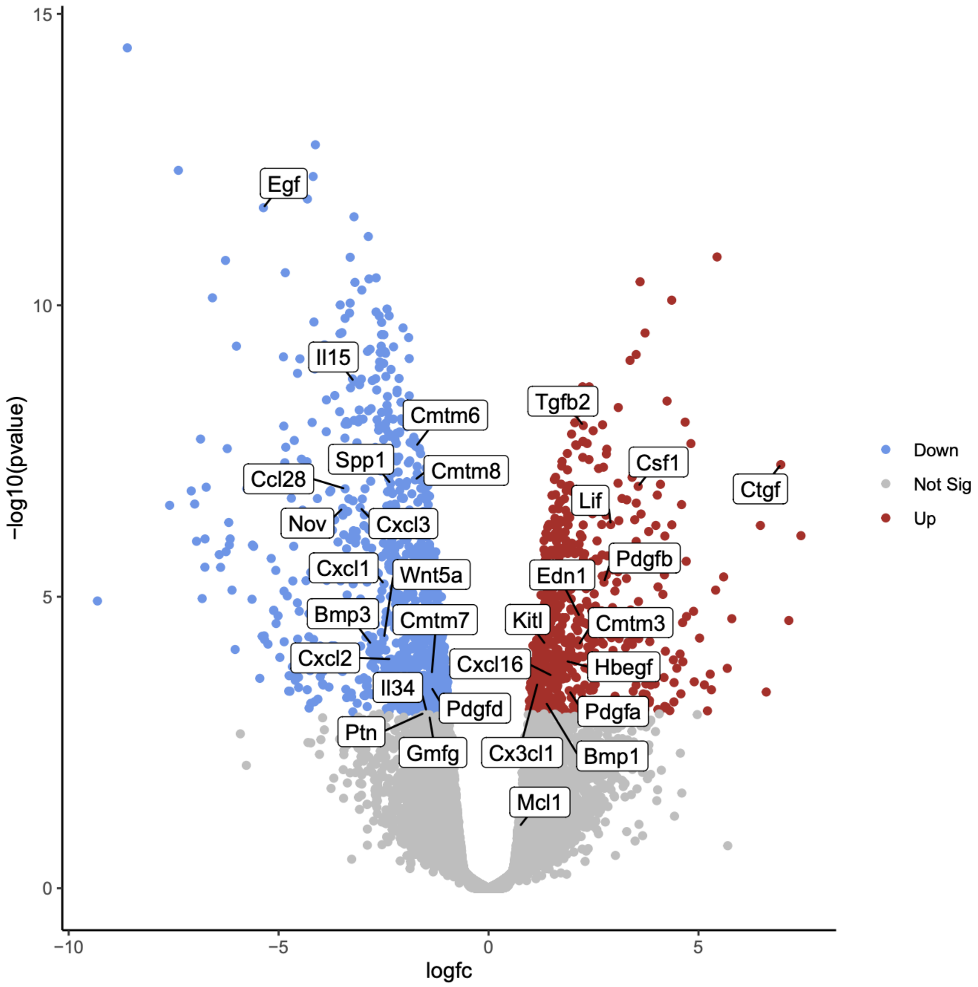
We can also label one or more genes of interest in a volcano plot. This enables us to visualize where these genes are in terms of significance and in comparison to the other genes. In the original paper using this dataset, there is a heatmap of 31 genes in Figure 6b (see the tutorial here if you would like to see how to generate the heatmap). These genes are a set of 30 cytokines/growth factor identified as differentially expressed, and the authors' main gene of interest, Mcl1. These genes are provided in the **volcano_genes** file and shown below. We will label these genes in the volcano plot. We'll add boxes around the labels to highlight the gene names.



1
GeneID
Mcl1
Hbegf
Tgfb2
Cxcl16
Csf1
Pdgfb
Edn1
Lif
Kitl
Bmp1
Pdgfa
Cmtm3
Cx3cl1
Ctgf
Wnt5a
Ptn
Spp1
Bmp3
Cmtm8
Gmfg
Cxcl2
Cxcl3
Il15
Egf
Cmtm7
Il34
Pdgfd
Nov
Cmtm6
Ccl28
Cxcl1

Hands-on: Create a Volcano plot labelling genes of interest

1. Use the **Rerun** button in the History to rerun **Volcano Plot** with the same parameters as before except:
   - “*Points to label*”: **Input from file**
     - “*File of labels*”: volcano gene file
   - In “*Plot Options*”:
     - ☑ “*Label Boxes*”: **Yes**



6. Alignment to a Reference Genome

The above function will import the files selected in RStudio and it will return the path where they will be stored. We will execute it multiple times, one for every distinct file (6 in our occasion). And now every time we are going to need one of the files imported we are going to use the path returned by the function.

7. Conclusion

In this tutorial, we learned about the generation and utilization of volcano plots in the context of RNA-seq data analysis. A volcano plot provides a powerful visual representation that allows for the rapid identification and visualization of significant genes within the dataset. This tool serves as an essential component in the analysis of RNA-seq results, offering a straightforward and efficient means of identifying genes that exhibit notable changes in expression levels under different experimental conditions.