Cool Volcano Plot – App

Daheng He[1] and Chi Wang

Biostatistics & Bioinformatics Shared Resource Facility (BB SRF)

Markey Cancer Center, University of Kentucky

Lexington, Kentucky 40536

USA

(November 2022)

[1] For technical supports, please contact me by dhe2@g.uky.edu

Relax! Let's first blabber a little

Volcano plot is a very useful tool to visualize the genetic alterations between two treatment conditions in bioinformatics studies. Volcano plot is essentially a dot plot with X-axis represents $\log_2$(fold change) and Y-axis $-\log_{10}$(p-value) (here the p-value refers to the raw p-values without any adjustment) that are computed for each gene based on its expressions between two groups of interest. Each dot represents a gene on volcano plot.

Nowadays, you can surely find many free online tools that can generate volcano plot for you, "Easy-peasy!", you say. But the situation may get a bit more challenging for you when you want to highlight a couple of genes of interest by labelling their names on their volcano plots – several labels are often terribly overlapped if their X and Y coordinates are too close, resulting in a messy region of texts entanglement, have you met such an annoying mess?

When this is happening, a non-programming researcher may roll up his sleeves, and open the volcano plot with messy overlapped labels with some Microsoft graphical editors like Paint, or some much fancier tools, such as Photoshop, to manually erase those mess and re-type each gene name on the plot at his favorite position. If this is what you used to do, the pro and con in doing so is:

Pro: it is admirable – you are a diligent researcher willing to face the challenge with the maximum physical effort! I solute you!

Con: it is meaningless – as an excellent researcher, you are wasting your valuable time, which you could have spent somewhere else and gain more terrific achievements! In some extreme cases, you may face a way more desperate situation, untangling over 30 terribly overlapped labels may take forever! "Mission impossible!", now you shout!

On the other hand, we programming researchers, will also roll up our sleeves and start coding, trying to build a tool of our own, that can employ certain algorithms to do the overlapping untangling automatically instead of manually. There is also pro and con in doing so:

Pro: it is fast and reusable – it doesn't matter how many labels you want to untangle, for us, all it takes is simply changing one or two lines of our source code, which takes like 20 seconds. Also, the data input may change, but the code is always the same. We only need to write the code once, once and for all!
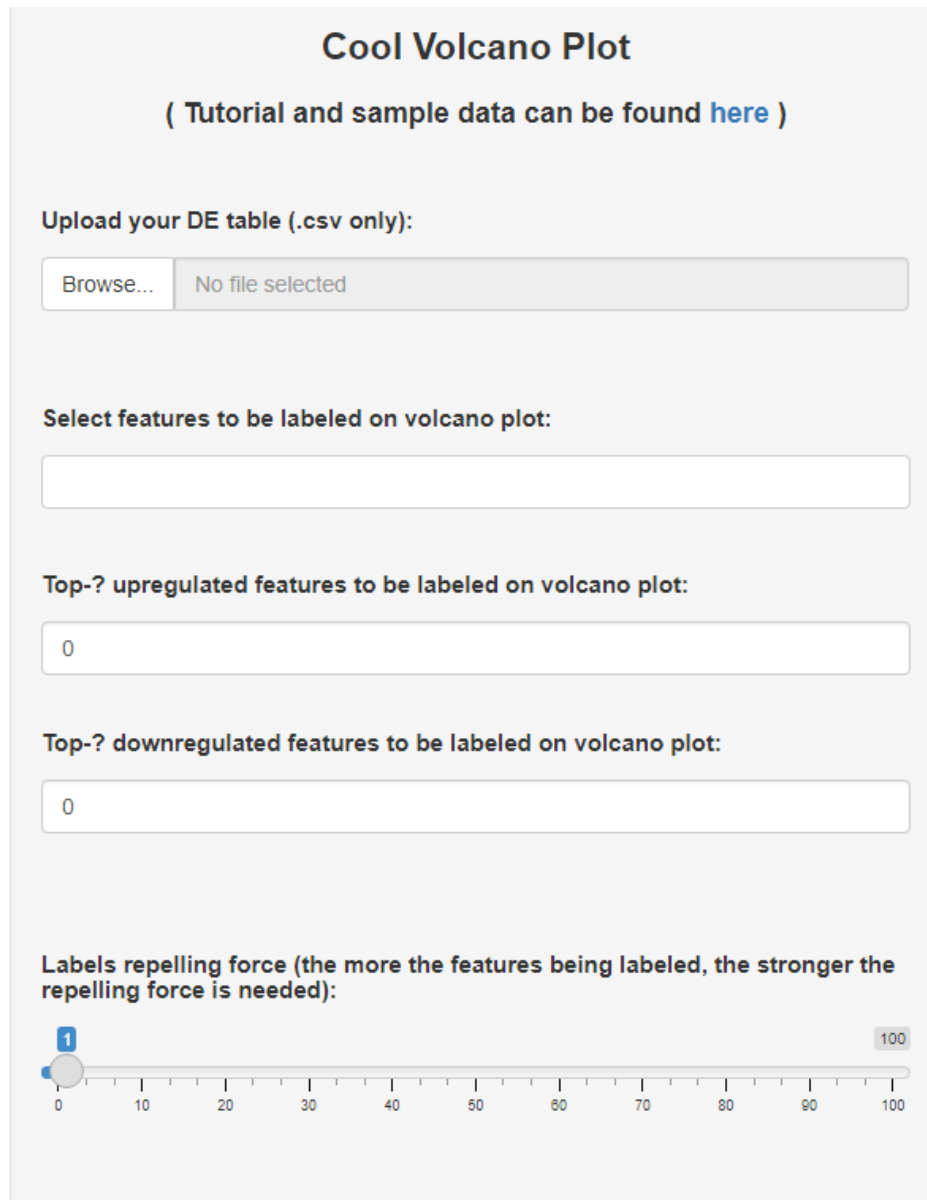
Con: due to educational background difference, the source code, which looks crystal clear to ourselves, may look kind of like "Martian language" to you. Simply emailing you our source code may not be any helpful to your situation. And we programmers have the similar feeling when we are watching you working with tubes, tweezers, and microscopes in your lab – handing over your lab apparatus to us without long-term professional training does not make us a master of the lab.

To smooth out the gap between these pros and cons, I built this R-Shiny free online App, ***Cool Volcano Plot***, with all of the confusing "Martian language" wrapped up in a "Blackbox", so that you don't need to worry about them, all you see is a simple UI allowing you to talk to our source code in human, instead of Martian, language – all you need to do is some super simple stuff like typing those gene names you are hoping to show on your volcano plot, and dragging around a slider bar button to let the background source code know that the current parameter choice is still not yielding an output cool enough to you, you are wanting something cooler! At the very end, after you have obtained a satisfactory volcano plot with all genes properly labeled at reasonable non-overlapping places, you can download the plot as an editable pptx file to your own computer and open it with PowerPoint, which allows you to easily further decorate your volcano plot until it reaches the level of art in your own eyes. "Cool volcano plot!", you now cheer!

OK! Now let's get serious! It's time to work

As you have seen, the web App *Cool Volcano Plot* has a very simple layout – a input panel on the left-hand side of the webpage, and the results viewing panels, **Your DE Table** and **Volcano Plot**, on the right.

Let's start with the input panel on the left. Upon opening the App for the first time without any input, the empty panel looks like this:

As a starting point, you must upload your own DE table, which may have been generated by some popular DE analysis programs, such as DESeq2, limma, edgeR,

etc. These programs may have their own unique components reported in the output DE tables. But for a typical DE table, they all share four common types of output:

1. a column containing gene names
2. a column containing the estimated $\log_2$(fold change) value of each gene
3. a column containing the estimated unadjusted (raw) p-value of each gene
4. a column containing the estimated adjusted p -value of each gene

These four columns in the DE table are the key components to draw a volcano plot. Therefore, you need to make sure that the four columns are simultaneously available in your DE table, and these columns must be named exactly in the following required case-sensitive way, as the App will ignore all other irrelevant contents of your DE table, but search for the four key contents with the required names.:

1. a column containing gene names – name this column as **Gene_Name**
2. a column containing the estimated $\log_2$(fold change) value of each gene – name this column as **logFC**
3. a column containing the estimated unadjusted (raw) p-value of each gene – name this column as **PValue**
4. a column containing the estimated adjusted p -value of each gene – name this column as **FDR**

Please note that the **FDR** in 4 above stands for "False Discovery Rate", which is a commonly adopted multi-comparison p-value adjustment method. Even if the adjusted p-values in your own DE table were not actually estimated by this method, you still need to temporarily name it as **FDR**, and easily change it to anything you want afterwards, as the output figure is in editable pptx format. The DE table to be uploaded to the App must be saved as csv format and when you open it by Excel, it should look something like this in Excel:

	A	B	C	D	E
1	Gene_Name	logFC	PValue	FDR	
2	COL9A3	5.420403939	4.76E-27	5.81E-23	
3	PEX5	7.260999603	9.81E-26	5.98E-22	
4	CCL5	5.854684227	1.63E-25	6.62E-22	
5	PGM1	5.292329551	3.91E-25	1.04E-21	
6	KIAA1109	4.463149255	4.27E-25	1.04E-21	
7	SPATA33	3.602536111	6.60E-25	1.34E-21	
8	RPL26L1	4.026842227	7.92E-25	1.38E-21	
9	ARF5	-5.78433262	4.94E-24	7.53E-21	
10	G3BP2	4.439590482	7.34E-24	9.44E-21	

The App will assume that the first row of the entries are column names, and check if the four necessary column names, **Gene_Name**, **logFC**, **PValue**, **FDR**, are all present in the uploaded DE table. If some of these four columns is not found, the App will return an error message in the input panel like this:

At this stage, the App will not allow you to go further, until you have fixed the problem and uploaded the corrected DE table. Please also note that the DE tables in genetic studies usually contain quite many genes, typically 10,000 ~ 20,000, the data uploading and parsing may take a while, depending on your internet speed,

please give it like 5-10 seconds before the table uploading and parsing are completed.

After loading the correct DE table, the App automatically scan and parse the gene names in the column of Gene_Name, and make them available for your selection in a form of drop-down menu -- **Select features to be labeled on volcano plot**, where you can make multiple choices of gene names, i.e., features, to be labelled on the volcano plot. Since the gene list is usually very long, you don't want to browse the list by scrolling your mouse wheel, which would take forever to find your genes. Instead, you should simply type the first two or three letters of each gene name in the text box, the App will narrow down the gene list that matches the letters you have typed. In most cases, after typing the first three letters, the gene list would be quickly narrowed down to just a few matching cases, then you may easily use your mouse to click on the one you want. Here, using the sample data that is available on line, I am randomly choosing five genes to be labelled on the volcano plot: PLK1, MYC, BRAF, CD274, CD2AP, just for demonstration purpose. In addition to these genes that you specifically want to show on your plot, it is often the case that you may want to show certain number of top up/down-regulated (ranked by p-value) genes, as these genes are the most statistically significantly differentially expressed ones, therefore may be worth sufficient attentions. If these top genes are not really what you care about, just skip the numerical inputs of **Top-? upregulated features to be labeled on volcano plot** and **Top-? downregulated features to be labeled on volcano plot**, and leave their default values of 0 be, then none top genes shall show. In fact, you may even skip the step of selecting genes from the drop-down menu, that way, you will get just a clean volcano plot without any label. Here just as a demonstration, I am randomly choosing to label the top-10 most significantly up-regulated as well as top-10 most significantly down-regulated genes on my plot. So, by now I have in total 5 + 10 + 10 = 25 genes to be shown on the plot. The scree shot of the input panel now looks something like this (see next page):

Cool Volcano Plot

(Tutorial and sample data can be found here)

Upload your DE table (.csv only):

| Browse... | sample_data.csv |

Upload complete

Select features to be labeled on volcano plot:

PLK1 MYC BRAF CD274 CD2AP

Top-? upregulated features to be labeled on volcano plot:

10

Top-? downregulated features to be labeled on volcano plot:

10|

Labels repelling force (the more the features being labeled, the stronger the repelling force is needed):

1   100

0   10   20   30   40   50   60   70   80   90   100

Once you are seeing the input panel on the left of webpage looks something like this, you are ready to view your output on the results panels. In the Your DE Table panel, you can view the DE table you have just uploaded. Of course, this may seem a little redundant, but the main purpose here is to help you make sure that you are submitting the correct data to generate your volcano plot.

The core output is of course the second result panel, **Volcano Plot**, once it is open, you will see that the prototype of the volcano plot is already there, and it already looks pretty cool. With the sample data and the choices of 25 genes above, the App calls the background "Martian language" that I hide in a Blackbox, and computes
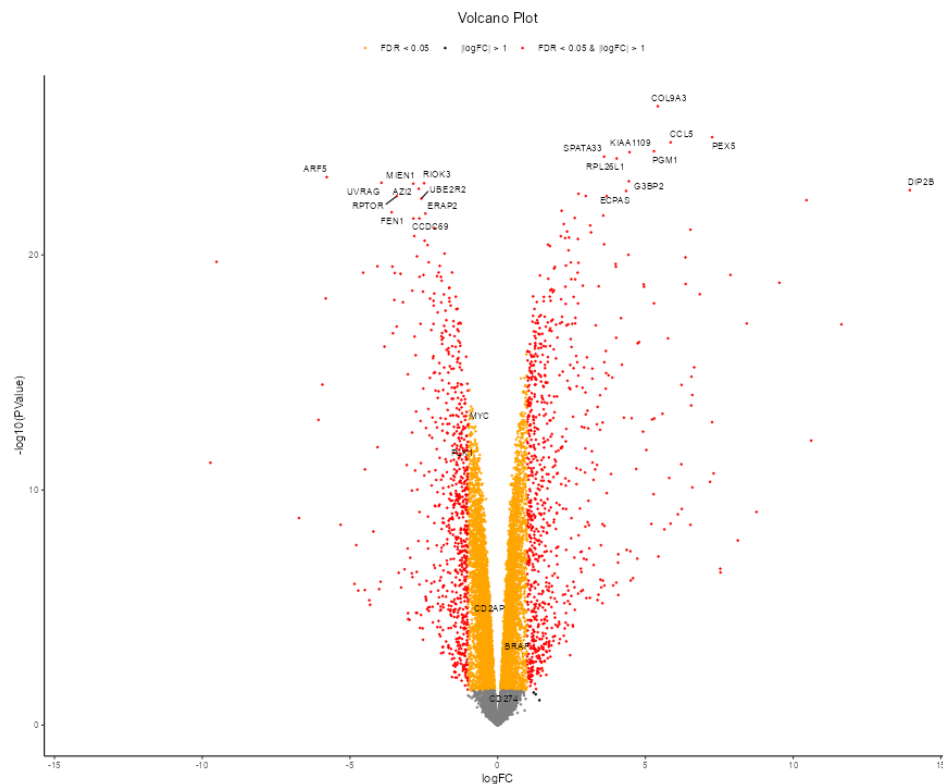
all the relative distances of labels and finds out an optimal labels arrangement which minimizes the overlapping of the labels based on a genius "overlapping repelling" algorithm. The details of the algorithm are mathematically formidable (and admirable!) for non-mathematicians, but on your side, Fear Not! All you need to know is that it only needs a single parameter, I personally call it **"labels repelling force"** – the more labels tending to entangle together (or, some labels are deeply "immersed" in a huge number of uninterested dots "ocean", the stronger force is needed to push them away from each other (or pull the deeply "immersed" labels out of the "ocean"). The default setting of **"labels repelling force"**, as can be seen on the slider bar button at the very bottom of the input panel, is 1, which is already a pretty strong repelling force – it is already strong enough to allow the simultaneous appearance of all the 25 gene labels on the volcano plot without any worrisome overlapping. Under the default repelling force (=1), the App returns me a cool volcano plot looking like this:

Pretty cool, isn't it? "Yes, it is!", you smile and nod.

The repelling force parameter has a default setting at 1, which is already good enough for many situations. And to be honest, in my real practice, I have never run into any situation that the App cannot handle when we increase the repelling force parameter to around 30. But I am giving you the option all the way up to 100, which is such a huge force, you may never need such a large value in your whole life, I am giving it to you to make you feel brave and confident, like our superhero Tony Stark, who died of a hero saving the whole universe from Thanos, and his only real-world equal, Elon Musk, usually feel about themselves every day! There is an option of 0 on the left end of the slider bar button, you may just try it for fun, and see what happens!

Downloading the figure

After obtaining the figure, you can download the figure by clicking the **Download the volcano plot as editable pptx** button. As indicated in the button label, the initiation of the downloading after clicking on the button may take a while, this is because after you click the download button, the background source code will start working on a series of steps to load large amount of information of your volcano plot into a PowerPoint object, these steps are quite computationally costly, therefore needs quite a bit of time. Give it like 5-10 seconds, then you will get the cool volcano plot, once and for all, with minimum physical effort.

Last but not least

The labels repelling algorithm depends dynamically on the real-time pixel of the graph, this means that the arrangement of labels on the volcano plot you are seeing on the screen of the App may NOT look identical to the one on the downloaded pptx file when you view it with Microsoft PowerPoint. This is because the one shown on the App webpage and the one you download to your own computer are not generated with the same pixel. Worry Not! The layouts may look a little bit different, but they have the same level of repelling force, therefore the labels on both of them are free from overlapping, and therefore look equally cool!