Shayla Ghose

HYRS Notes-July 15th

Volcano Plot:

-Powerful visualization tool used in high-throughput experiments to identify significant changes between two conditions – like controlled vs treated samples.

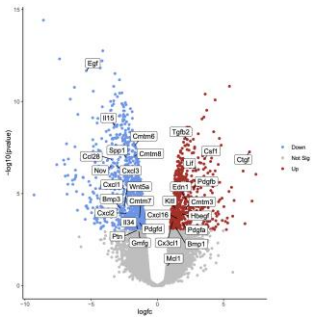
-Type of scatterplot used in DGEA to visualize the relationship between magnitude of gene expression and their statistical significance.

-Helps quickly identify genes that are significantly up or down regulated, and those that are biologically impactful.

-Useful when dealing with many genes as they focus on those that are most impactful/significant.

-A legend is always included to explain the color scale.

Example of Volcano Plot:



Components of Volcano Plot:

What variables are included in the volcano plot?

-In the volcano plot, gene symbols, p-value, and the log2 fold change are included. The dots represent genes and are labeled based on their significance.

-Top right represents upregulated, and top left represents downregulated. Non-significant genes are those that are unlabeled and uncolored.

What is the x- and y-axis in the volcano plot?

-The x-axis represents the log2fold change. It measures the magnitude of change in expression. If pointing far to the right, it is upregulated. If pointing far to the left, it is downregulated.

-The y-axis represents the –log10 p-value. The higher up a dot (representing a gene) is, the more statistically significant it is.

-Genes in the top corners of the plot are the most interesting as they show large changes and are significant. Those that are near the center bottom are not significant.

Why are log2FC and p-values taken into account for the volcano plot?

-log2FC values are crucial in a volcano plot because they indicate the magnitude of change in gene expression between two conditions. A positive log2FC value means the gene is upregulated, while a negative log2FC value means the gene is downregulated. This helps identify genes that have undergone significant changes in their expression levels.

-P-values are crucial in volcano plots because they determine the statistical significance of observed changes in gene expression levels. By plotting the negative logarithm of the p-values on the y-axis, the most significant genes are highlighted at the top of the plot.

Why are B values typically not included in the volcano plot?

-B values (log odds score) are from the limma package (used for microarray and RNAseq analysis) and is the log-odd that a gene is differentially expressed. It assures greater confidence that a gene is truly significant the higher the B-value is.

-B values are skipped in the volcano plot because the plot mainly focuses on the magnitude of change (log2FC) and stastical significance (p-value). Although B values are useful, they are less easy to visualize than p-values and are mostly only relevant in limma analyses.

Why are t values typically not included in the volcano plot?

-The t value tells you how different two groups are in a statistical test (usually a t-test), comparing the difference between groups, variations, and sample sizes. It basically measures how far your sample result is from what you would expect if there were no difference (the null hypothesis).

-The larger the t value, the more likely it is that the two groups compared have real differences. When the t value is large, it means the result is far from the null hypothesis, resulting in a small p value (stronger evidence against the null hypothesis). -T-values are not included in volcano plots because they are intermediate statistics from tests like the t-test and are not intuitive for visual representation. P-values, derived from t values, give a clearer sense of significance.

When are B and t values included in the volcano plot?

-B values are not typically included in volcano plots, but researchers may use them by substituting the –log10(p-value) with the B-value on the y-axis. This reflects posterior probability instead of stastical significance. Posterior probability relies on a prior belief and using experimental data to update/confirm this belief, adding confidence to whether a gene is actually differentially expressed.

-T values are harder to interpret visually and do not have a specific threshold but can be used to reflect stastical strength instead of significance, and can bridge the gap between raw test output and derived significance.

Theory behind volcano plot creation:

-Purpose is to identify genes that are significant and biologically meaningful.

-Resembles a volcano because non-significant genes cluster in the middle, with fewer genes showing extreme changes.

-Emerges from trade-offs: in bioinformatics, between fold change and statistical significance.

Heatmaps:

-Visual representation of data where values are shown using colors.

-Each region on the map corresponds to a data point.

-Color intensity reflects the magnitude of the value: warm colors (like red and orange) indicate high values or activity (upregulation), while cool colors (like blue and green) indicate low values or activity (downregulation).

-Nonsignificant genes (high p-value and low log2FC) will appear as white or gray (middle of color scale).

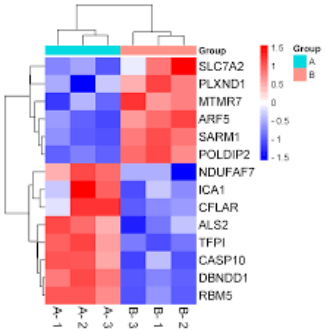
-Used to show gene expression levels in bioinformatics through RNA-seq and microarray heatmaps.

-Highlights patterns and outliers, and makes large datasets easy to understand.

-Data is mapped to a spatial layout, and then a color scale is applied to represent values.

-A legend is always included to explain the color scale.

Example of a heatmap:



Components of Heatmap

What are the variables a heatmap uses?

-Gene symbols are used as row labels (y-axis), samples/conditions are the columns (x-axis), average expression data is the main data shown by color intensity, and regulation (up or down regulated) is implied by the color.

-The color scale represents the log2FC for each gene expressed in different samples/conditions, and distinguishes between different groups.

What is not included in the heatmap?

-P-values are better suited for significant plots and are not color scaled, so they are not generally included in heatmaps.

-Probe ID is hidden by gene symbols for more clarity and is not directly shown on the heatmap, t-values or intermediate stats and are not intuitive for color mapping, and B-values are rarely shown in heatmaps.

What do the x- and y-axis represent in a heatmap?

-The x-axis (columns) could be different depending on the context, but it can represent samples (patients) or experimental groups/conditions (control vs treated).

-The y-axis (rows) typically represents the features being analyzed (genes); each row corresponds to a specific gene.

-Looking down a column, you can see which genes are "on" or "off" relative to other genes. Looking across a row, you can see how its expression varies between different conditions and samples.

Theory behind heatmap creation:

-Each little square shows a value and is color coded to represent gene expression levels.

-The x-axis represents samples or conditions, while the y-axis represents features like genes.

-The clustering tree (dendrogram) groups similar samples or genes so it is easier to recognize the patterns.

-The color key or legend helps describe what each color represents and how to interpret the results visually.

-Shows how clusters of genes act alike in different conditions.

Other types of Graphs for DGEA:

Boxplot/Violin plot – distinctly highlights median, range, and outliers of a gene's expression between control vs treated groups, making it easy to visually compare variability and detect significant differential expression.

Bar plot – uses bar height to represent gene counts and often color to indicate statistical significance; highlights the number of up and downregulated genes across different conditions.

MA plot - visualizes each gene's fold change against its average expression, highlighting the up and downregulated genes between two conditions.

Hierarchical Clustering Dendrogram - tree-like diagram visualizing how clusters of genes group together based on similarity.

Gene Ontology Enrichment plot - highlights biological themes that are overrepresented in a set of genes; instead of just knowing which genes are up and down regulated, this plot tells you what these genes do and where they are active.