Khan Inan
BI-GY 7633
Exercise 12

M12_v1_description) Bioinformatician's goals
M12_v1_1) What are three main services that Bioinformaticians provide?

The three services are BSA, SSS and MSA. BSA stands for biological sequence analysis, MSA is multiple sequence analysis and SSS sequence searching services.

M12_v2_description) Common plots part 1
M12_v2_1) What plot can be helpful in determining if your normalization process was successful? Explain what it shows

Usually a scatter plot or curve is best for displaying normalization. As long at the data points somewhat match the normalization line/curve, it can b considered that our normalization process was successful

M12_v3_description) Common plots part 2
M12_v3_1) What is the purpose of a Venn Diagram and what is its limitation?

A venn diagram is useful for comparing and contrasting processes and workflows. It's main limitation lies in its very arbitrary/subjective nature where its up to users to determine what aspects between two things are the same or different

M12_v4_description) Genomic Data Visualization part 1
M12_v4_1) How can we use Genome browsers to visualize RNA-seq data?

Heatmaps are the best way to visualize RNA-seq data, in regards to genome browsers. Genome browsers allow use to filter and sort our data using its menus and search options.

M12_v4_2) How do you create a dot plot?

A dot plot is essentially a representation of frequency. It involves first drawing a line and then placing a dot for the number of times that a value occurs in our data set, very similar to a bar graph.

M12_v5_description) Genomic Data Visualization part 2
M12_v5_1) How is the height in a Sequence Logo graph calculated?

The height in a specific log graph is directly proportional to the number of values that are at the specific position

M12_v5_2) How can you avoid making mistakes in interpreting dendrograms?

Since dendograms can get fairly complicated it's very important to always pay attention to what your branch of interest is stemming from and to zoom in to get a non-cluttered view of all the overlapping branches towards the bottom

M12_v6_description) Genomic Data Visualization part 3
M12_v6_1) What can you conclude by clustering samples instead of genes?

Clustering samples gives us a broader view of species-to-species relationships and general biological distances between specific samples. Genes are more specific and are more useful for determining genomic similarities

M12_v6_2) Why do we calculate the -log10() of the p-value when creating a volcano plot?

-log10() is a representation of the level of significance for each gene and first you calculate the p value from the test statistic and then you plug in the value into the log formula

M12_v6_3) How can we use heatmaps to interpret our data?

Heatmaps show concentrations and color change at key locations on the genome. The more sharp the color change is, the higher the similarity or the higher the concentration of values, compared to lighter portions of the heatmap

M12_v7_description) Principal Component Analysis
M12_v7_1) What is the relationship between principal components and the variation in the data?

The first principal component captures the highest amount of variation and each principal component after that captures sequentially less

M12_v8_description) Gene Networks
M12_v8_1) Give three examples of edges in a gene network and the nodes that connect them.

Certain examples of edges in gene networks are present in yeast studies, protein studies and any kind of pair-wise relationship in studies. The nodes that connect them essentially represent aspects of the network from which physical protein interactions stem from

M12_v8_2) What kind of edges can you draw using RNA-seq experimental data?

You can draw fairly complex edges using RNA-seq data and you can do this manually or using a software tools such as edgeR. These softwares use algorithing from a library of genes and calculates things such as distance and dispersion