Final Deliverables

Kristen Kelly

Pipeline of Paper

This paper focused on determining possible prognostic markers of colorectal cancer. The pipeline included screening for hub genes, then creating a ceRNA network to discover associations between miRNAs and lncRNAs.

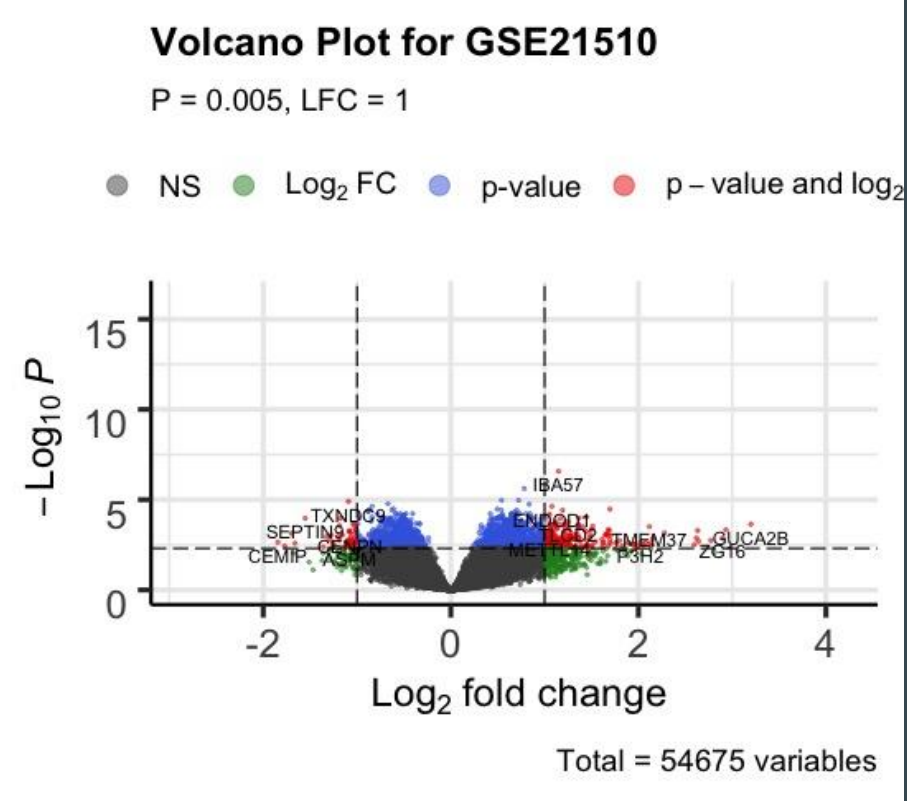
In our session, we learned that the researchers initially screened for differentially expressed genes. Afterwards, potential hub genes were analyzed using David, KEGG and STRING. For week 5 deliverables, groups used GO analysis, KEGG, WikiPathways, EnrichR, David and/or Metascape.

Parts of the pipeline that were new to me: “TCGAbiolinks” package for specific information about the patients’ cancers, GEPIA & Starbase databases for the prognostic results of candidate genes, analysis of the ceRNA network to find regulatory relationships, statistical analyses using DESeq2 and other statistical tests.

What's Missing about the Pipeline

- Quality control information
  - I used the array quality metrics package, which focuses on the relative quality of the different arrays in the dataset. The assessment of raw data provides information about the experimental procedures and the assessment of the normalized data gives insight on how to use it in future analyses.
- Normalization method
  - I used MAS5 normalization, which normalizes each array independently and sequentially. Other methods, such as RMA and GCRMA, could give different results. There are pros and cons to different types of normalization, but the pipeline did not specify which normalization method was used.
  - We used PCA plots to further determine the quality of our normalized data set. However, there are no PCA plots present in the paper.

Volcano Plot

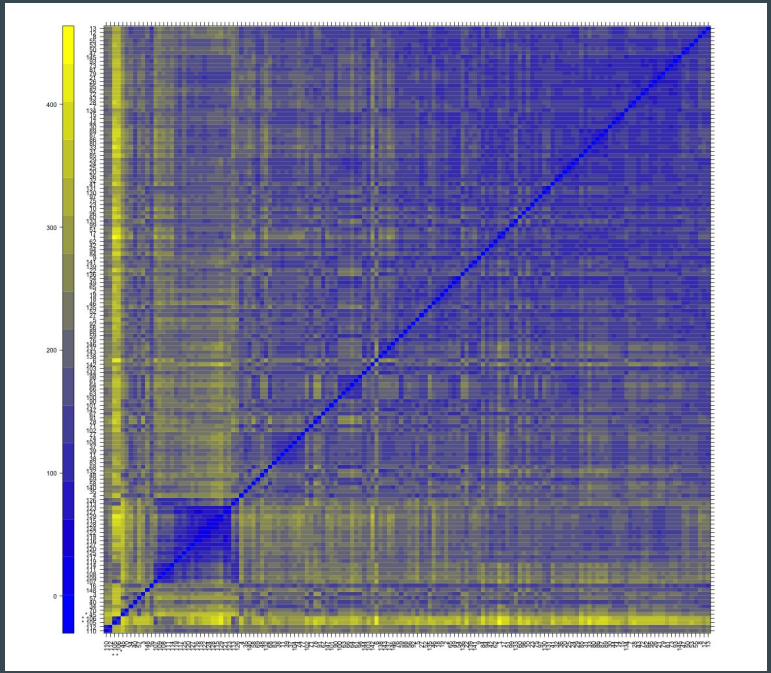


- Mas5 Normalization
- Significant upregulated genes: GUCA2B, ZG16, IBA57
- Significant downregulated genes: CEMIP, SEPTIN9, TXNDG9

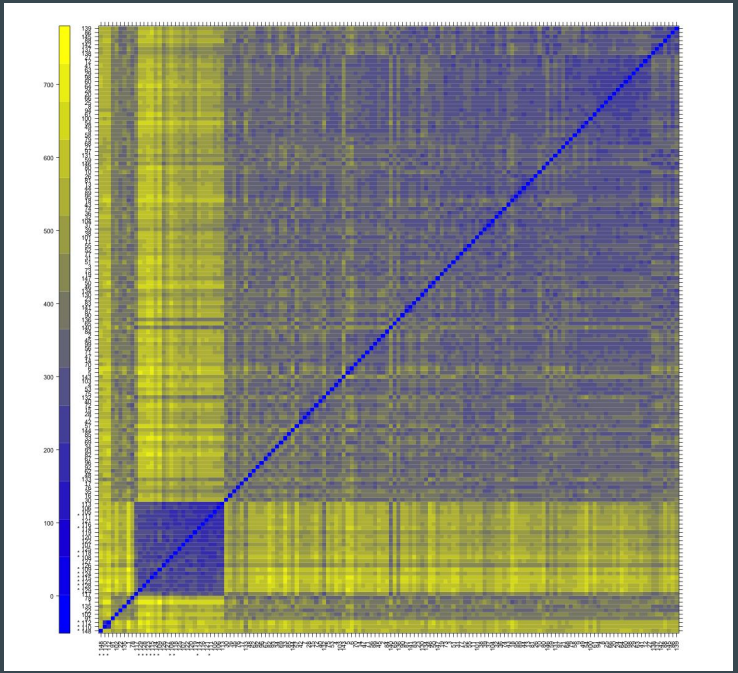
```
volano_plot <- EnhancedVolcano(t1,
                               x = "logFC",
                               y = "adj.P.Val",
                               xlim = c(min(t1$logFC, na.rm = T) -1,
                                        max(t1$logFC, na.rm = T) + 1 ),
                               ylim = c(0, max(-log10(t1$P.Value), na.rm = T) + 5),
                               pCutoff = 0.005,
                               FCcutoff = 1,
                               pointSize = 0.5,
                               lab = genes$SYMBOL,
                               labSize = 3.0,
                               title = "Volcano Plot for GSE21510",
                               subtitle = "P = 0.005, LFC = 1")
volano_plot
```

Heat Map

- Used array quality metrics to get heatmaps for raw and normalized data



Raw Data Heat Map



Mas5 Normalized Heat Map

Things I Learned

Throughout this internship, I learned:

- Bioinformatics is not just about coding
- Tips on how to read a research paper
- Different steps of analyzing a large data set such as quality control and normalization
- Interpretation of different plots and forming new hypotheses from results
- Divergent thinking and how to come up with more creative ideas
- How to play Secret Hitler at Gene Team Happy Hour!

Tools I learned:

- STEM-Away Forums
- Slack
- Asana
- R
- Python
- GitHub

Soft skills I learned:

- LinkedIn and networking tips
- Communication with my group
- Presentation and public speaking advice

Challenges

- I struggled at points throughout the internship because I had very little coding experience. The deliverables would sometimes take me a long time to complete and I would have to ask the leads and mentors many questions.
- In the beginning, I had some trouble following along during the trainings. I rewatched the recordings and looked at the notes many times.
- Time zone differences within my group made it difficult to communicate find meeting times
- When working alone, I faced the challenge of not being able to ask my group quick questions or bounce ideas off of them

Project Proposal

- A project based on the next steps of the paper using the hub genes found from the first analysis to construct the ceRNA network. Students could follow the same steps we did in the June session, compare hub genes, then create a network to determine biological interactions for these specific genes.
- Also, I think that having different groups for varying coding experience could be beneficial. At times, the more experienced group could be paired with the beginner group to give advice.
- Another idea could be to have different groups based on interest. One team could focus on coding and another team could focus on the biological side of bioinformatics.

Achievement Highlights

1. I was able to create the volcano plot and heatmaps all by myself for this presentation. I didn't have to reach out for help from mentors or leads, which was a big step for me.
2. Completing the second set of Python Exercises was an accomplishment for me because I spent a lot of time on it and it gave me confidence using Python.
3. I found the STEM-Away forums very confusing at the start, but now I am completely comfortable navigating the website.

Bioinformatics Excitement!

At the start of this internship, I knew very little about coding. I was nervous about keeping up with everyone else. I became very excited about bioinformatics after the webinar with Ali. After attending this webinar, I realized that I do not need to be a coding expert in order to participate in bioinformatics. I enjoy the biology side of bioinformatics, particularly interpreting the results and forming new hypotheses. I am happy I did this internship to redefine my idea of bioinformatics and to find my niche in this field.