Week 8 (Final) Deliverables

•••

By: Shanya Sanof

Guo Paper Pipeline

- 9 hub genes, 13 related miRNAs, and 29 candidate lncRNAs screened in colorectal cancer
    - Used to create ceRNA network
- Determined MFAP5, miR-200b-3p and AC005154.6 may all have potential prognostic value for CRC patient population

Data Set Used:

- GSE21510
- **44 homogenized (25 normal and 19 cancer),** 104 LCM

Guo Paper Comparison

Similarities:

- Quality Control
- Screened Differentially Expressed Genes

Differences:

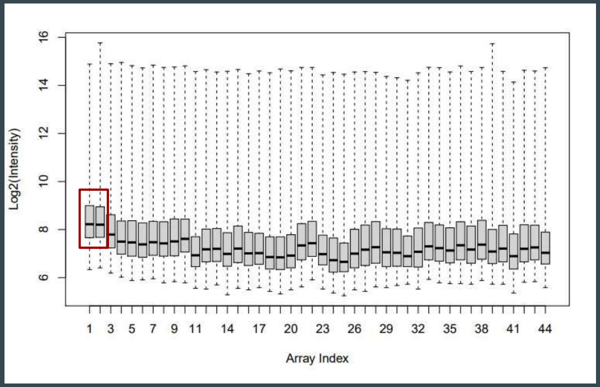
- Potential prognostic values of candidate genes (TCGAbiolinks, GEPIA, StarBase)
- Screening related miRNAs and lncRNAs
- Construction and in depth analysis of the ceRNA network
- Statistical analysis and network visualization

QC Method (affyQCReport)

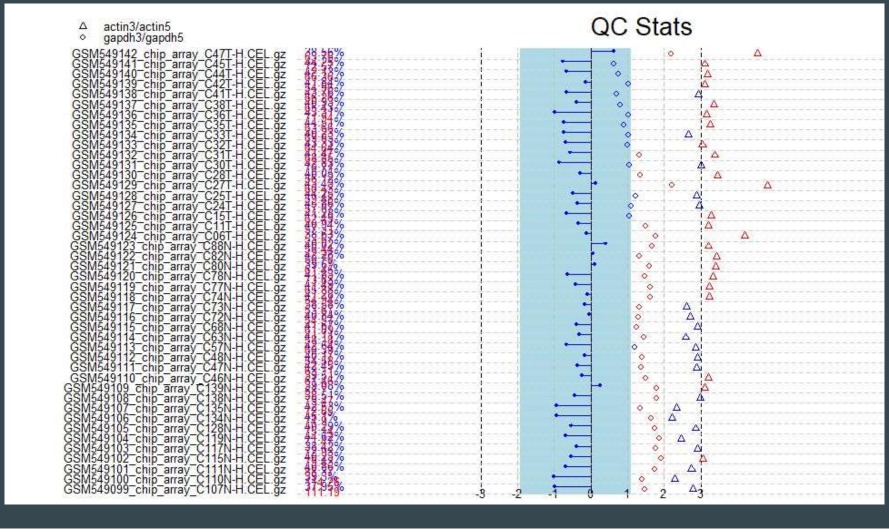
- Generated a report with many plots

- Intensity boxplot helped identify outliers
    - GSM549099 and GSM549100 (Sample 1 and 2) were identified as outliers

Intensity Plot

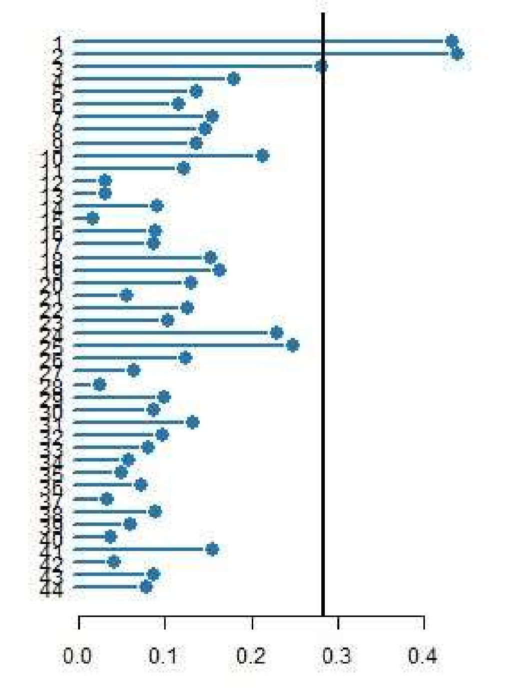


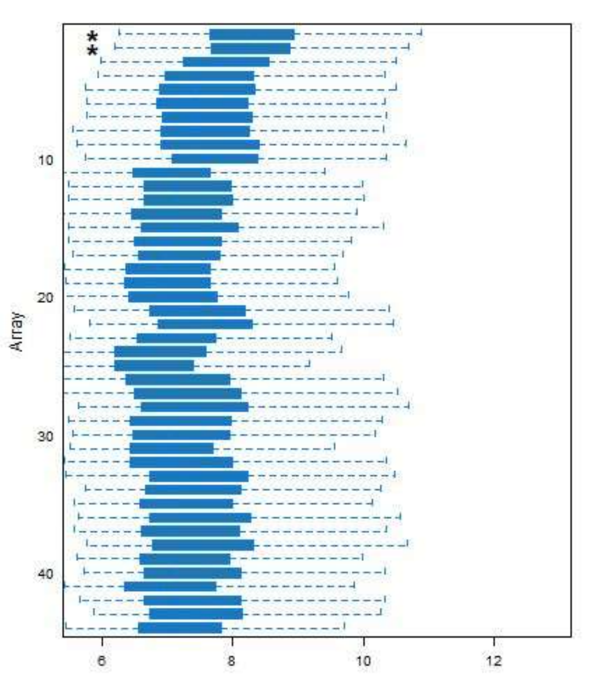
simpleAffy QC Plot



QC Method (arrayQualityMetrics)

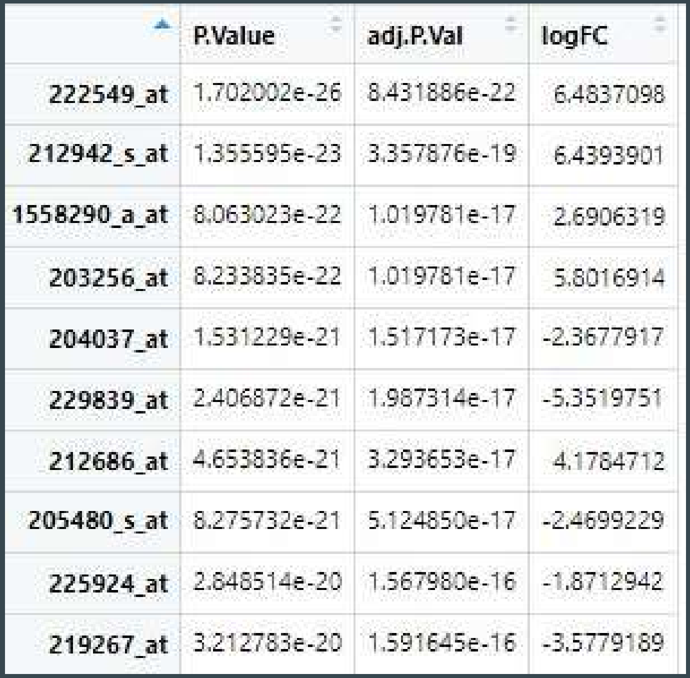
- Outlier Detection
  - ○ Threshold of 0.283 (indicated by vertical line)
  - ○ Values that exceed threshold = outliers
- Used to identify/confirm outliers found using affyQCReport
  - ○ Confirmed GSM549099 and GSM549100 are outliers





Differentially Expressed Gene Table

- Decreasing p-value as threshold decreased the amount of statistically significant DE genes



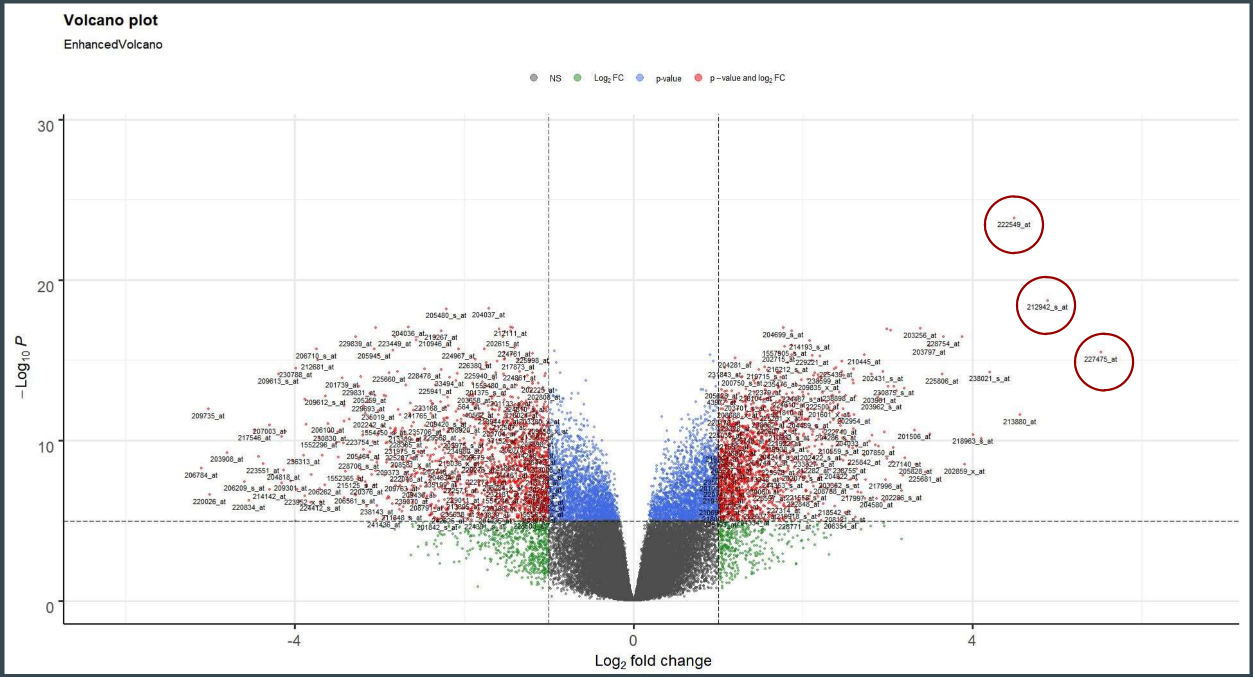
	P.Value	adj.P.Val	logFC
222549_at	1.702002e-26	8.431886e-22	6.4837098
212942_s_at	1.355595e-23	3.357876e-19	6.4393901
1558290_a_at	8.063023e-22	1.019781e-17	2.6906319
203256_at	8.233835e-22	1.019781e-17	5.8016914
204037_at	1.531229e-21	1.517173e-17	-2.3677917
229839_at	2.406872e-21	1.987314e-17	-5.3519751
212686_at	4.653836e-21	3.293653e-17	4.1784712
205480_s_at	8.275732e-21	5.124850e-17	-2.4699229
225924_at	2.848514e-20	1.567980e-16	-1.8712942
219267_at	3.212783e-20	1.591645e-16	-3.5779189

Volcano Plot

- The most statistically significant genes are the ones with the highest -log P values.

- The most relevant genes are the ones that have both the highest log2 FC and log P values

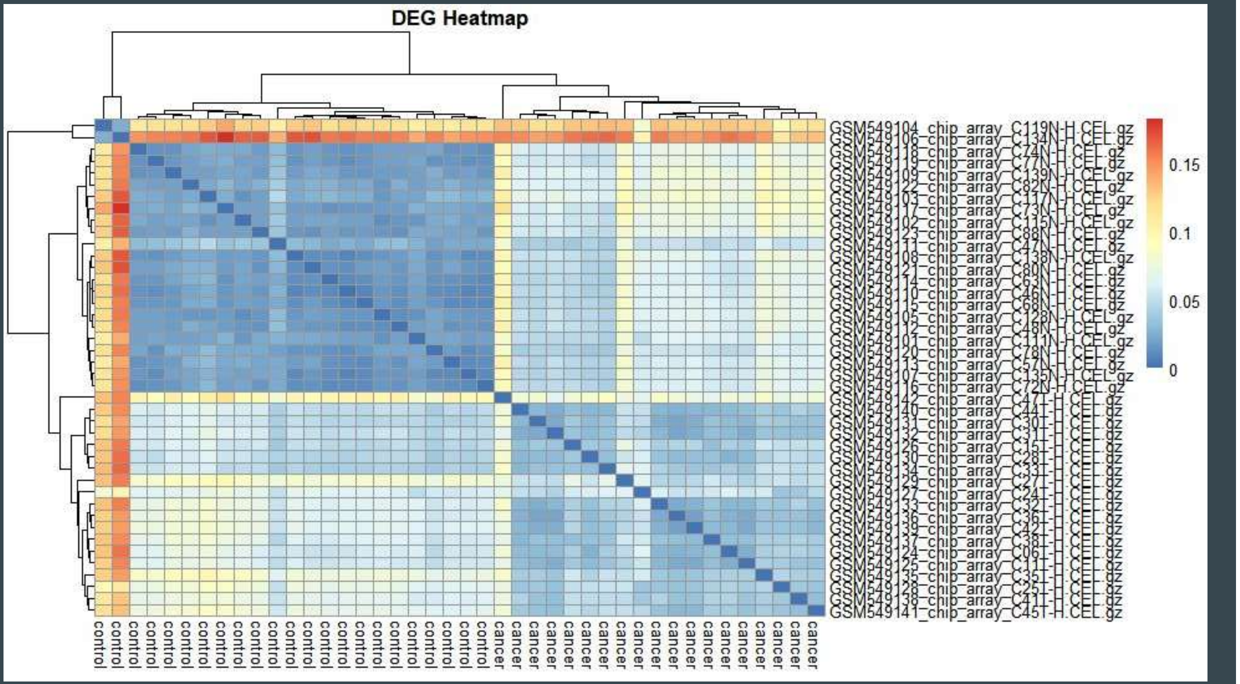
- Most relevant genes
  - **CLDN1**
  - CEMIP
  - FOXQ1



Heatmap

- Red indicates increased interaction between genes
- Blue indicates least amount of interaction

- **First two control genes had highest amount of interaction**



Things Learned

- How to individually interpret the different plots
    - Better understanding on how to interpret each plot
- How to troubleshoot on my own

Challenges Faced

- Confusion with wording of paper and determining which part of the dataset to use
- Working individually without the help of my teammates
- Determining outliers using different QC methods
- Limma code produced many errors

Things Learned in this Session

- How to read a scientific paper
- Coding languages: R and Python
- How to work effectively in a virtual setting
- Working with people in time zones

Thank you!