Gene Expression Analysis of Classic Hodgkin's Lymphoma

Relapsed/Refractory vs. Non-Relapsed Patients
Male vs. Female

Gene Expression Analysis and Visualization / Final Project

November 30, 2017

Julie Garcia

Introduction

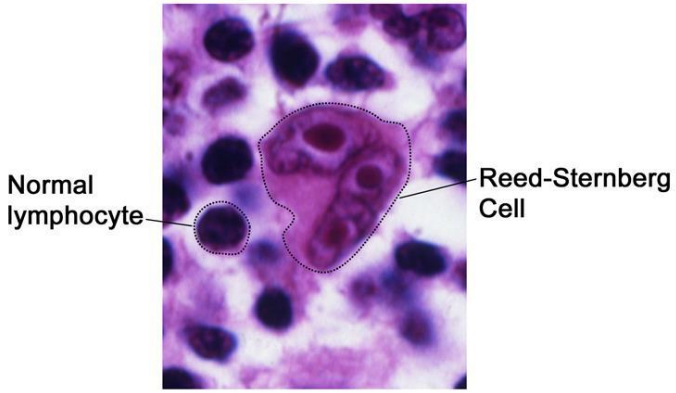
- Dataset
  - Source = Human Classic Hodgkin's lymphoma diagnostic lymph-node biopsies (pre-treatment with ABVD chemotherapy)
  - 130 samples
  - 54,675 features
  - Affymetrix Human Genome U133 Plus 2.0 Array
  - In-situ oligonucleotide array
  - RNA sample type
  - Tissue = Lymph Node
  - Downloaded from GEO

Introduction

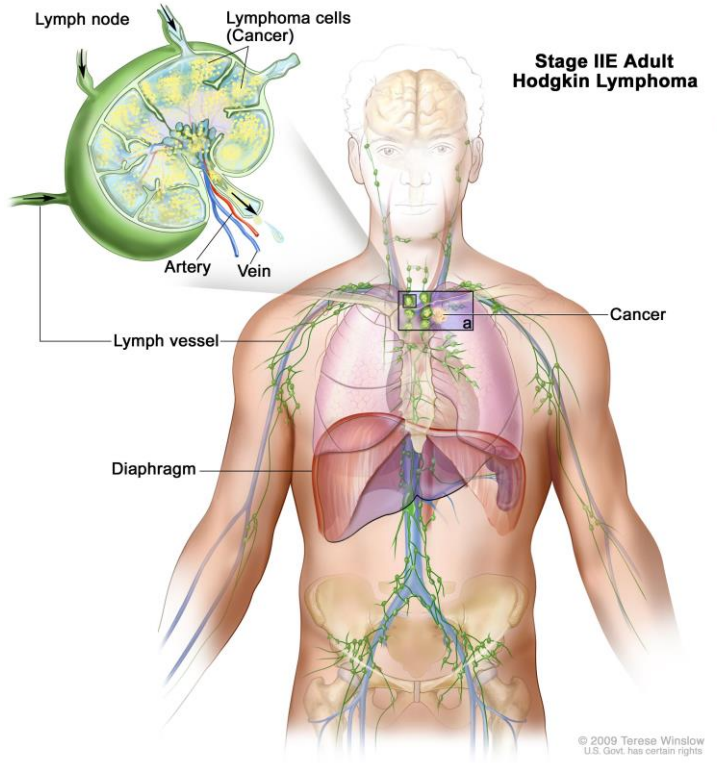
- Gene Expression Analysis
  - Data was separated into many class including gender, prognostic scores, disease stage and treatment outcomes
  - I focused on analysis of gene expression pre-treatment as it is related to treatment outcomes
    - Main question = How is gene expression different in patients for whom first line ABVD chemotherapy failed?
- Main Factor – Treatment Outcome
  - Success = Non-Relapse after treatment w/ ABVD
  - Failure = Early Relapse, Late Relapse, or Refractory Disease after treatment w/ ABVD
- Secondary Factor – Gender
- Secondary Factor – Disease State
  - Early = Stage 1 or Stage 2
  - Late = Stage 3 or Stage 4

Classic Hodgkin's Lymphoma

- Cancer of the lymphatic system
- Majority of people have classic type
- Enlarged lymphocytes (Reed-Sternberg Cells)
- Exact cause unknown
  - Related to exposure to viral infections (Epstein-Barr), familial factors, and immunosuppression
  - Common among HIV patients



http://www.dana-farber.org/hodgkin-lymphoma/



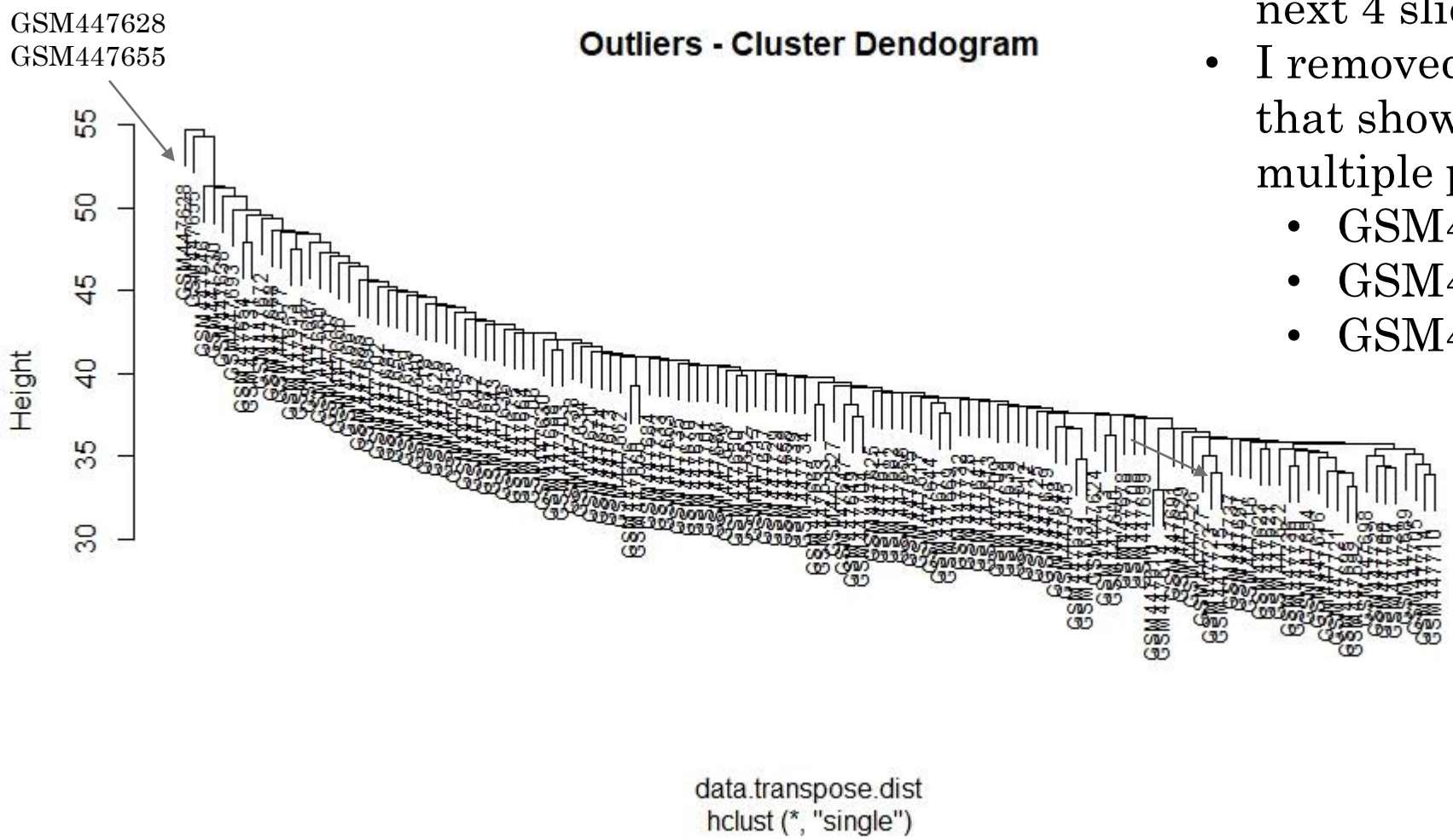
http://www.dana-farber.org/hodgkin-lymphoma/

Treatment Outcomes

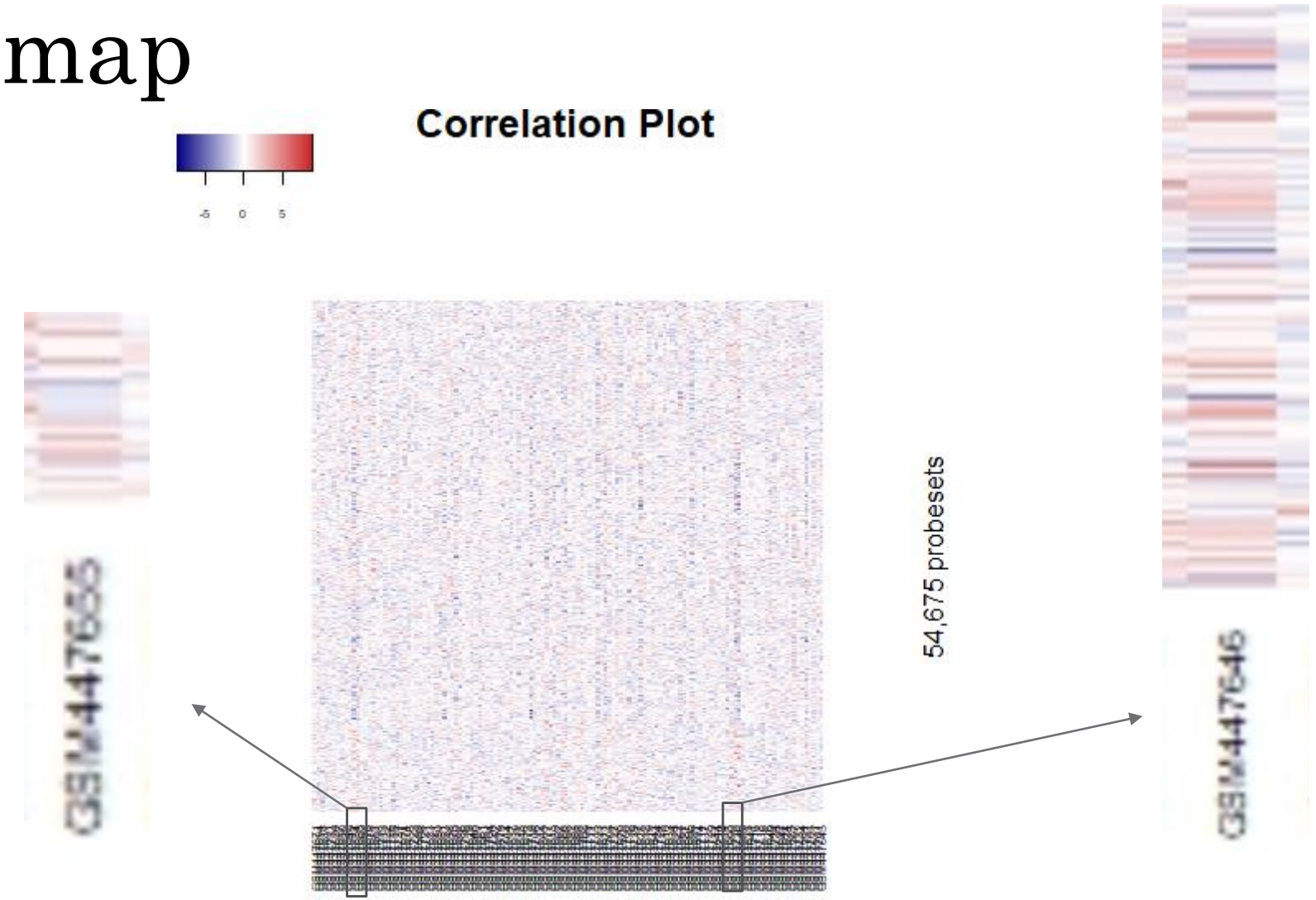
- Classic Hodgkin's Lymphoma is considered one of the most curable forms of cancer
- ~9,000 new cases per year in the US
- More common in males, aged 45 or older
- 80% of patients are cured by first line treatment with ABVD
  - ABVD = doxorubicin, bleomycin, vinblastine, dacarbazine
  - Radiation therapy depending on stage of disease
- 20% of patients still have poor outcome and relapse or have refractory disease
  - 50% are rescued by autologous stem cell transplant
- Novel biomarkers are needed to predict favorable or unfavorable outcomes with first line treatment, so that more appropriate treatments can be prescribed in this group that is resistant to ABVD

Normalization & Identify Outliers

- First I log2'd the data to normalize it
- I identified potential outlier samples with charts on the next 4 slides
- I removed only the samples that showed up as outliers in multiple plots
  - GSM447646
  - GSM447628
  - GSM447655

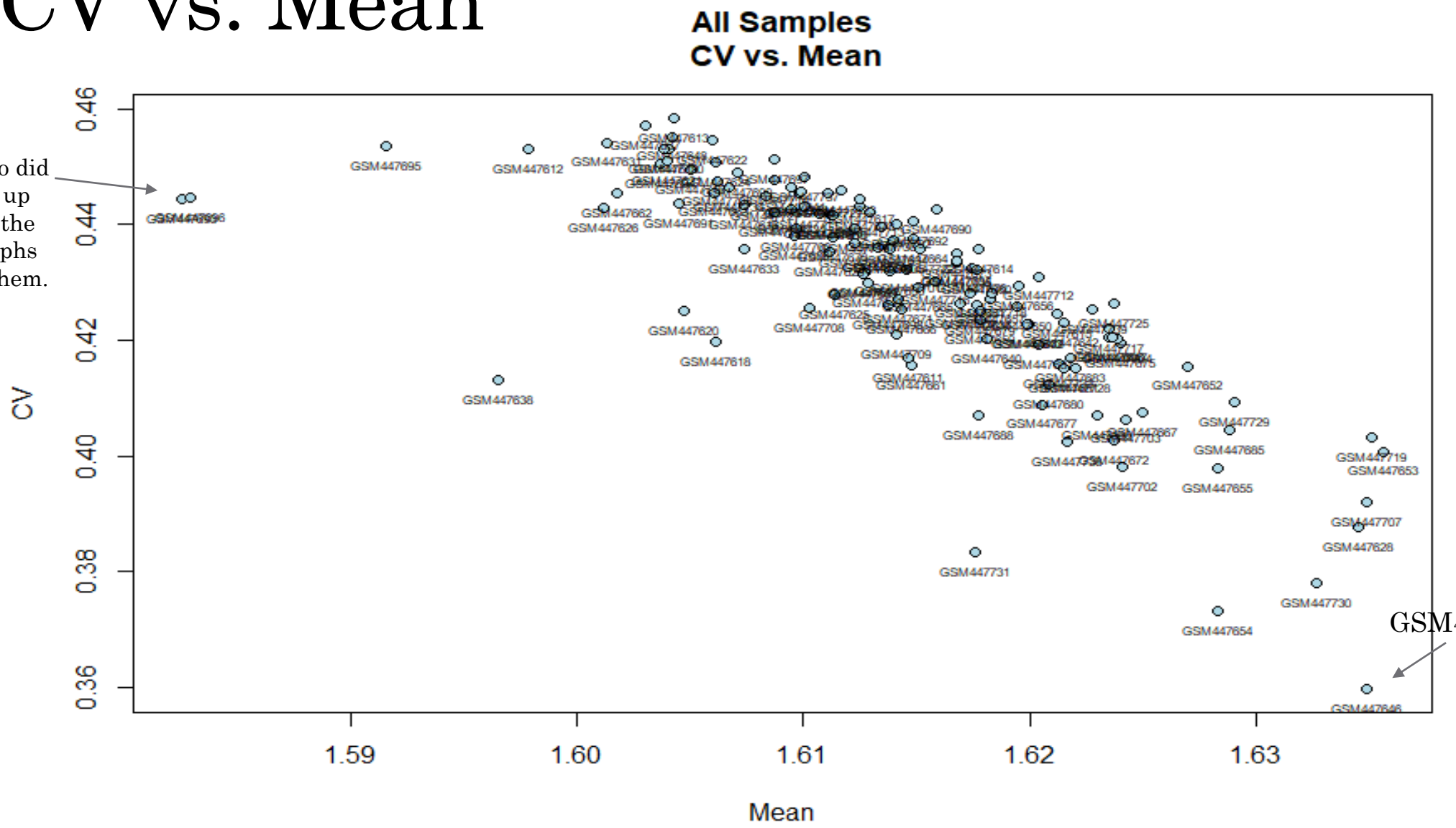


Heatmap


Correlation Plot
-5
0
5
GSM447635
54,675 probesets
GSM447646

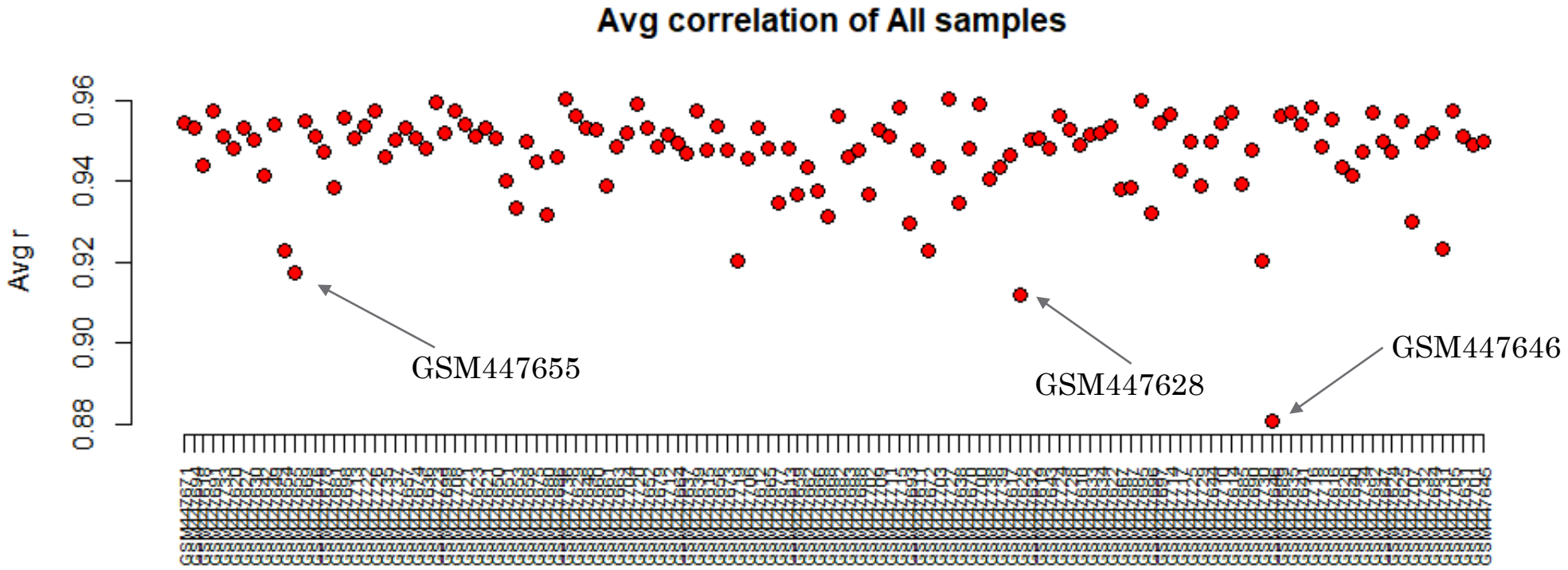
CV vs. Mean

These two did not show up in any of the other graphs so I left them.



GSM447646

Average Correlation



Gene filtering low expression

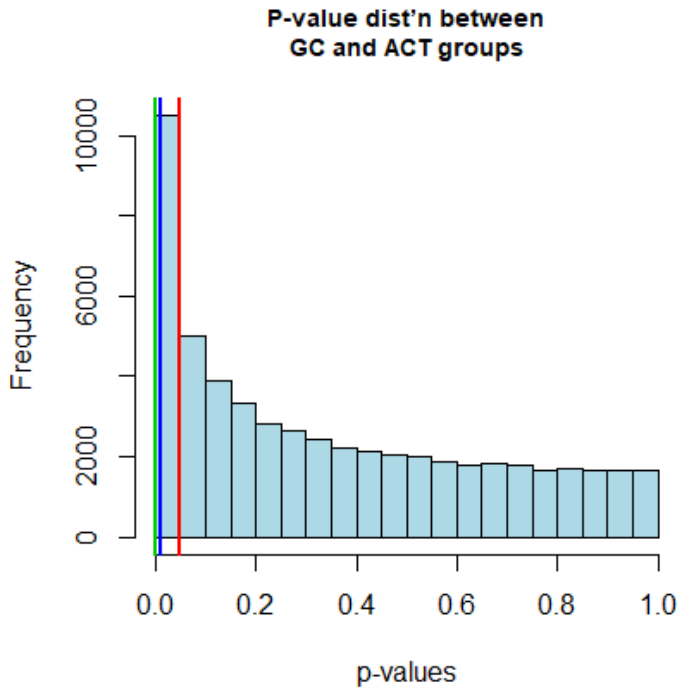
- The lowest 5% mean expression of genes were filtered out
- Method
  - Calculated the mean of each gene across samples
  - Sorted the list of means
  - Removed lowest 5% of mean gene expression from the dataset
- 51941 genes remaining
- 127 samples (after outlier removal)

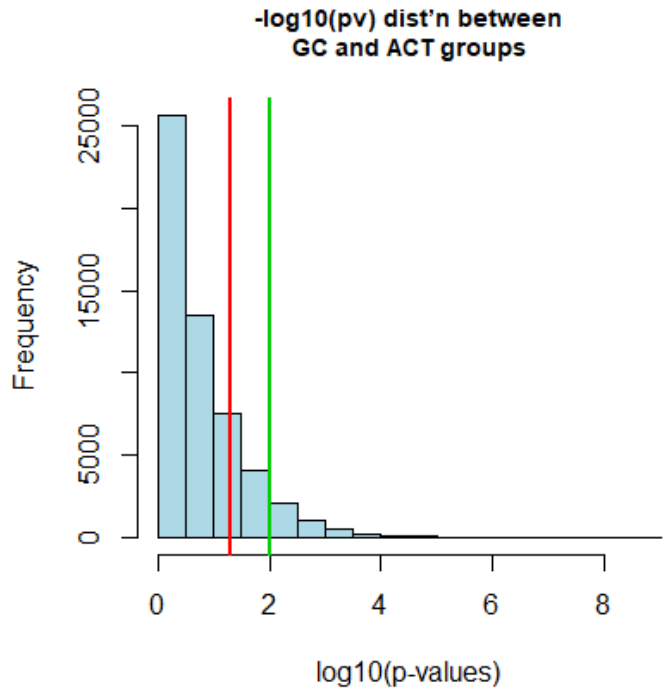
Feature selection overview

- Samples were divided into two sets
  - Success (gc) = No Relapse after ABVD
  - Failure (act) = Early or Late Relapse or Refractory Disease
- Calculated Student's t-test on all genes (univariate approach)
  - NA's were removed for each test set
  - p-values were output
  - Significant p-values were selected ($p < 0.0001$)
- Fold change was calculated (data in log2 so I subtracted)
  - fold = success.m - failure.m
  - Fold change of 1.1 was used as cutoff (using 2 did not reveal any significantly expressed genes)
  - 639 genes found
- Intersection of significant p-values and significant fold change was found and revealed
  - 36 genes were found in common and fed into next step (dimensionality reduction)

Feature Selection (p-values)

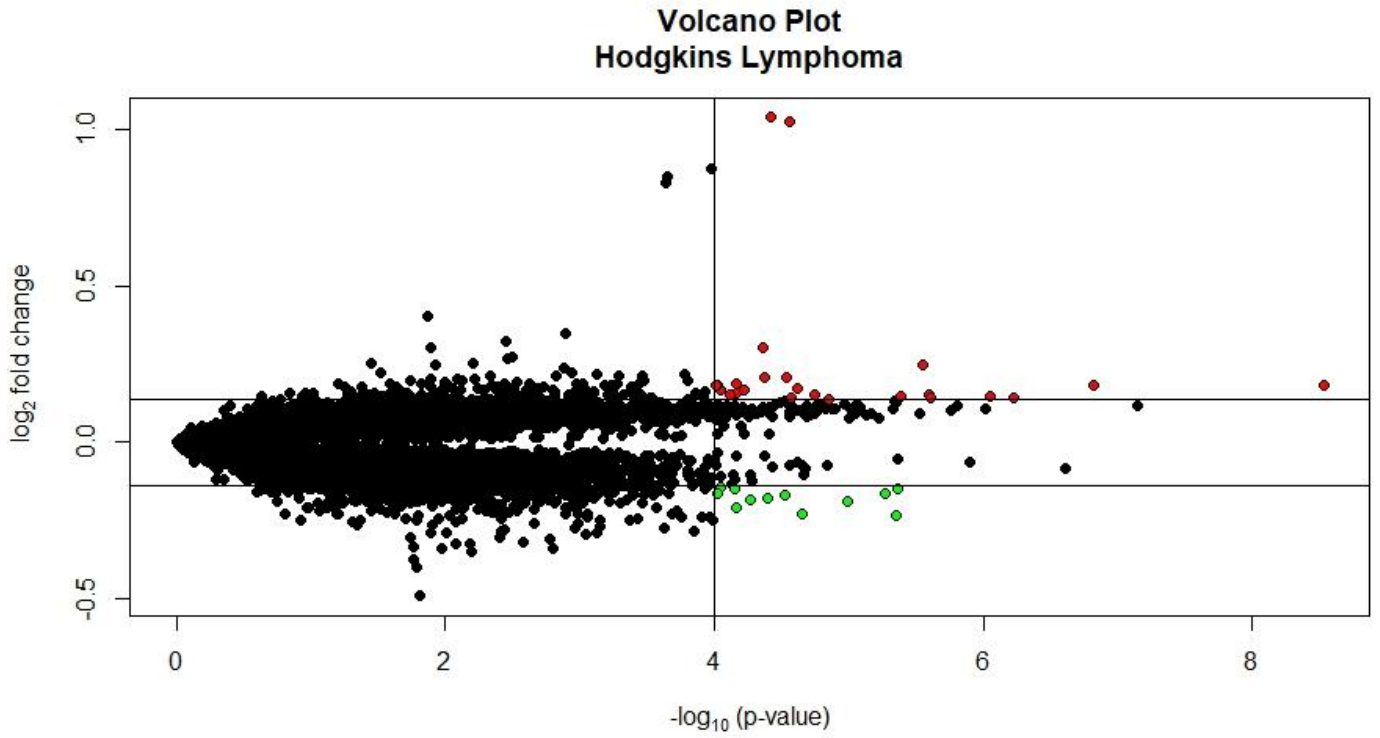
- Tried three different p-value thresholds to determine the most significantly expressed genes
  - Threshold = .05 there were 9732 significant p-values (red line)
  - Threshold = .01 there were 3600 significant p-values (blue line)
  - Threshold = 0.00001 there were 146 probesets with significant p-values (green line)





Feature Selection (fold change)

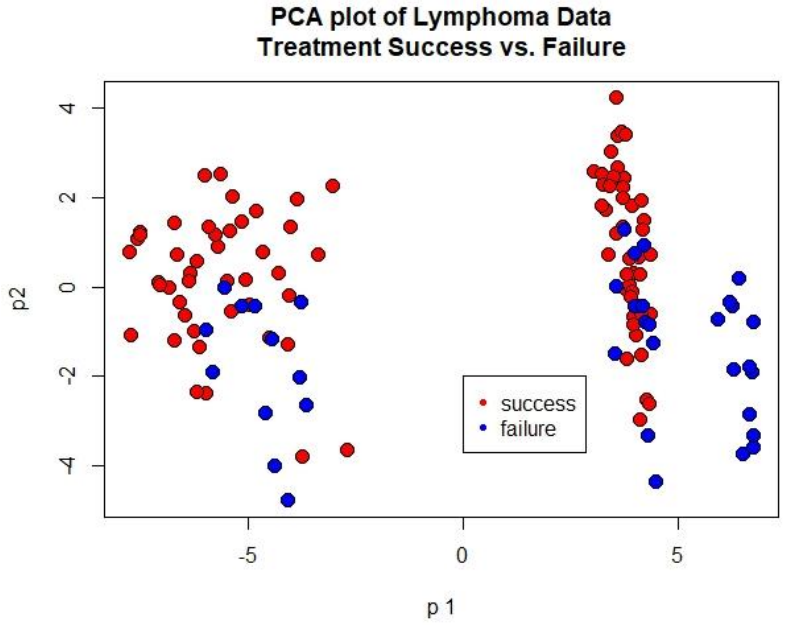
- Fold changes
  - In order to find overlap I had to reduce the fold change value to 1.1 which indicates that there are not a lot of largely significantly expressed genes between the two classes
- This plot shows the genes overexpressed in red and underexpressed in green
- 36 genes with significant over- or under- expression found

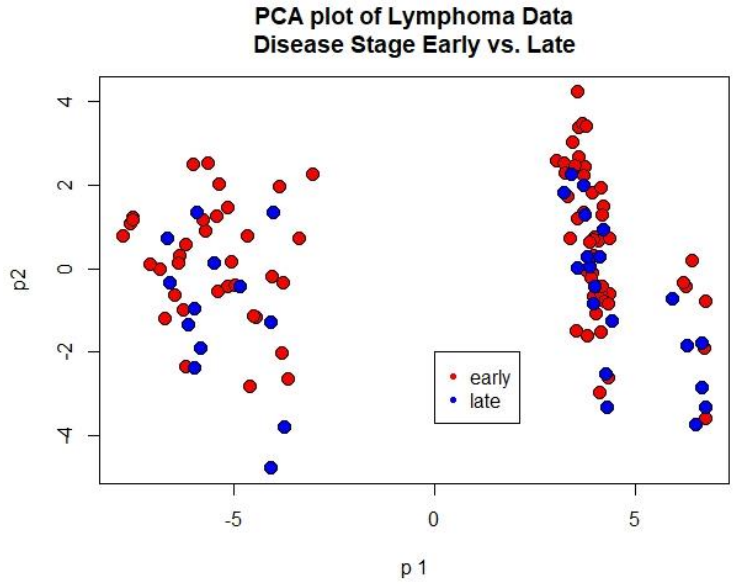


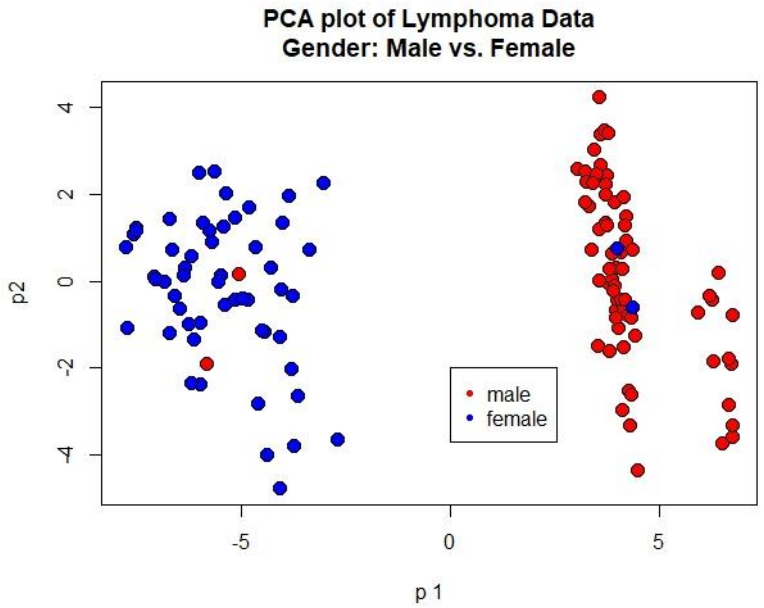
Dimensionality Reduction

- Method
  - 36 genes, 127 samples fed into algorithm
  - This dataset was transposed
  - Ran PCA
  - Charted PCA with three different factors
    - Treatement outcome
    - Gender
    - Disease State
  - Gender showed greatest division between groups, which makes sense because me and women's gene expression is different
  - There does seem to be a not as clear line between treatment outcomes perpendicular to the Y-axis that separates successes vs. failures in treatment outcome
  - See next slide for PCA plots

Dimensionality Reduction





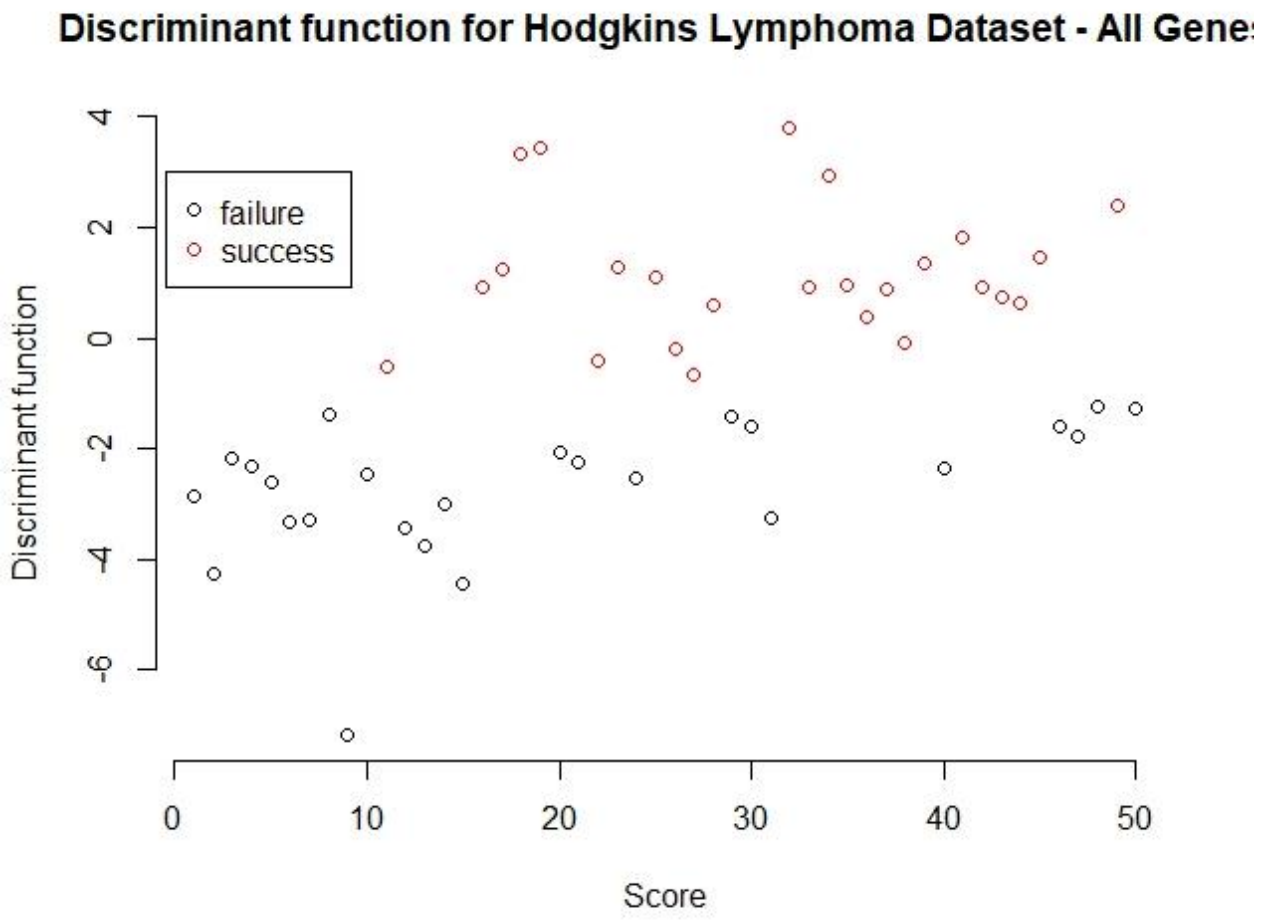


Classifying samples

- Split the data into Training and Test sets
  - Training = 60%
  - Test = 40%
- Ran LDA with training set
- Ran prediction algorithm with test set
- Confusion Matrix shows 12 misclassified

	Failure	Success	Total
Failure	14	11	25
Success	1	24	25
Total	15	35	50

Classification



Top 10 Significant Gene Information

Overexpressed

AFFYMETRIX_3PRIME_IVT_ID	Gene Name	Related Genes	Species
1569652_at	MLLT3, super elongation complex subunit(MLLT3)	RG	Homo sapiens
208576_s_at	histone cluster 1 H3 family member b(HIST1H3B)	RG	Homo sapiens
211753_s_at	relaxin 1(RLN1)	RG	Homo sapiens
214218_s_at	X inactive specific transcript (non-protein coding) (XIST)	RG	Homo sapiens
224588_at	X inactive specific transcript (non-protein coding) (XIST)	RG	Homo sapiens

Underexpressed

Download File

AFFYMETRIX_3PRIME_IVT_ID	Gene Name	Related Genes	Species
206157_at	pentraxin 3(PTX3)	RG	Homo sapiens
208607_s_at	SAA2-SAA4 readthrough(SAA2-SAA4)	RG	Homo sapiens
222883_at	cytochrome c oxidase assembly factor 7 (putative) (COA7)	RG	Homo sapiens
233587_s_at	signal induced proliferation associated 1 like 2 (SIPA1L2)	RG	Homo sapiens
202990_at			

Conclusions

- There was not a huge difference in gene expression between the two groups that I set out to study (successes and failures in first line treatment with ABVD). The genes that I found to be the most significantly expressed only had a slight greater than 1 fold change. Some of them had a fold change of between 0 and 1, which I would consider to not be signficant
- The most significantly overexpressed genes were two non-protein coding genes that were two different variants of XIST. XIST controls X-inactivation in women and is only minimally expressed in men in the testis. If XIST is overexpressed in males with lymphoma who do not respond to treatment, this could be a pathway worth exploring in the future.
- The most significantly underexpressed genes had fold changes of again between 0 and 1, so I do not consider them to be that significant. The most significantly underexpressed gene was PTX3, which is known to be enduced by the presence of inflammatory cytokines, which is a path which may be worth exploring.
- I think based on analysis of this dataset, there doesn't seem to be a huge difference in gene expression between the two cohorts. However, I'd like to further explore the male vs. female path.

References

1. Ansell, S. M. (2015). Hodgkin Lymphoma: Diagnosis and Treatment. *Mayo Clinic Proceedings, 90*(11), 1574-1583. doi:10.1016/j.mayocp.2015.07.005

2. Steidl, C., Lee, T., Shah, S. P., Farinha, P., Han, G., Nayar, T., … Gascoyne, R. D. (2010). Tumor-Associated Macrophages and Survival in Classic Hodgkin's Lymphoma. *The New England Journal of Medicine*, *362*(10), 875–885. http://doi.org/10.1056/NEJMoa0905680

3. http://www.genecards.org/cgi-bin/carddisp.pl?gene=PTX3&keywords=PTX3

4. http://www.genecards.org/cgi-bin/carddisp.pl?gene=XIST&keywords=XIST

5. https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS4222#details