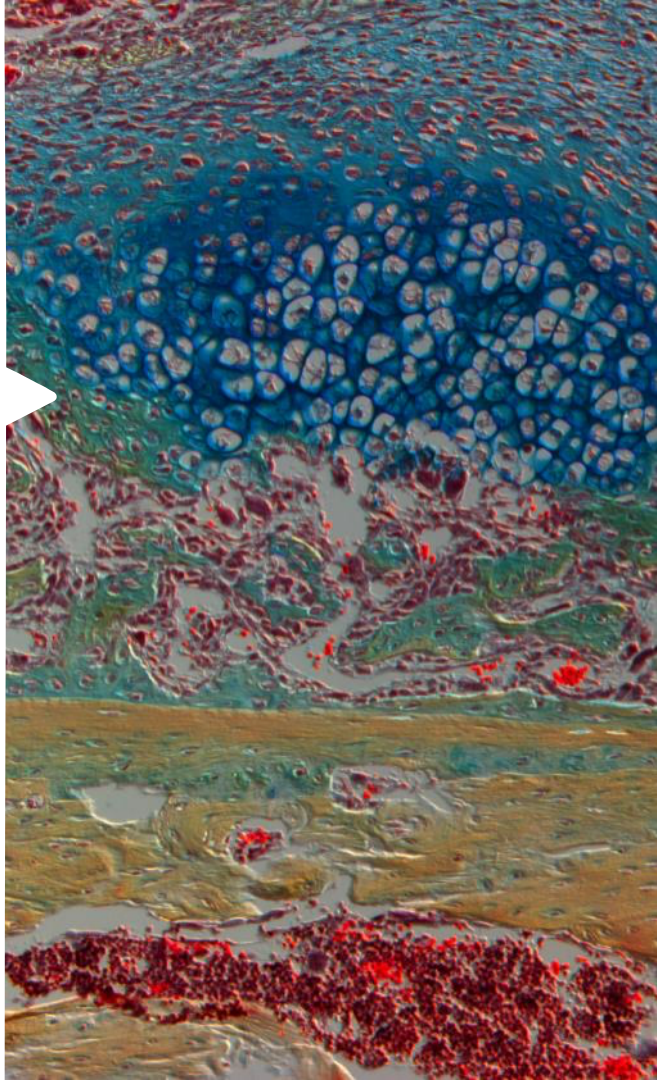
NYU Langone Health

Microarrays and the Golub AML/ALL Paper

Mark Grivainis

September 21st 2020



Outline

- Microarrays
- Data Processing Techniques
- Break
- Golub Paper
  - 20 min breakout session
  - 20 min review

- I need to make sure everyone has Kerberos Ids so that I can apply for VPN Access
  - I would like to apply for everyone at the same time, so I need a complete list
- To get a more up to date list of who has access I added a column for VPN Access
  - Please fill this out if you have already added your Kerberos ID
- Kerberos Ids
  - https://docs.google.com/spreadsheets/d/13KwtOaoLQf66Oaow_LUKoqhPMod2wbLdzroW7hwv1Ck/edit?usp=sharing

NYU Langone Health

The History of Microarrays

- 1975 | Grunstein and Hogness randomly cloned e-coli DNA onto agar plates
- 1979 | Gergan et al. adapted to the process to produce arrays
- 1980-1990's
  - The process is automated
  - Robotics technologies are incorporated
  - Analysis of multiple hybridizations
  - Increased efficiency and reduced errors
- 1995 | The term "microarray" is used for the first time; cDNA printed onto glass
- 1997 | First whole genome microarray study using yeast

Microarray Protocol

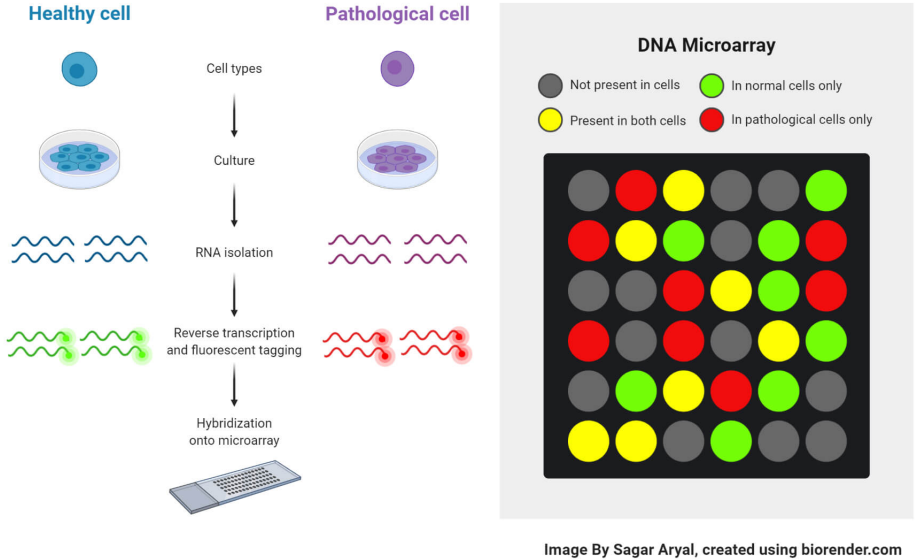


Image By Sagar Aryal, created using biorender.com

https://microbenotes.com/dna-microarray/

NYU Langone Health

The Applications of Microarrays

- Identification of single-nucleotide polymorphisms (SNPs) and mutations
- Classification of tumors
- Identification of target genes of tumor suppressors
- Identification of cancer biomarkers
- Identification of genes associated with chemoresistance
- Drug discovery
- Comparative genomic hybridization
- Antibiotic treatment
- Early detection of oral precancerous lesions
- And more

Microarray Files

- The raw output differs by platform
- Generally there will be some header information followed by some tabular information.
- Many file types have header information which describe the file or data within the file
- Header

```
!Series_title\t"Gene expression profile of human colorectal carcinoma"\n
!Series_geo_accession\t"GSE113513"\n
…
```

- Body

```
"ID_REF"\t"GSM3108231"\t"GSM3108232"\t"GSM3108233"\t"GSM3108234"\t"GSM3108235"\t”
"11715100_at"\t20.55044\t27.800337\t16.569454\t16.01692\t18.516483\t19.59903\t
…
```

Data Processing Techniques

Data Processing – Normalization

- Adjusting the values within a dataset to a common scale
- Two common approaches:
  - Min-Max feature scaling
  - Z Score Normalization (Standardization)
- Between sample normalization
  - Allows comparisons to be made between samples
- Within sample normalization
  - Using positive and negative controls to adjust values within a sample

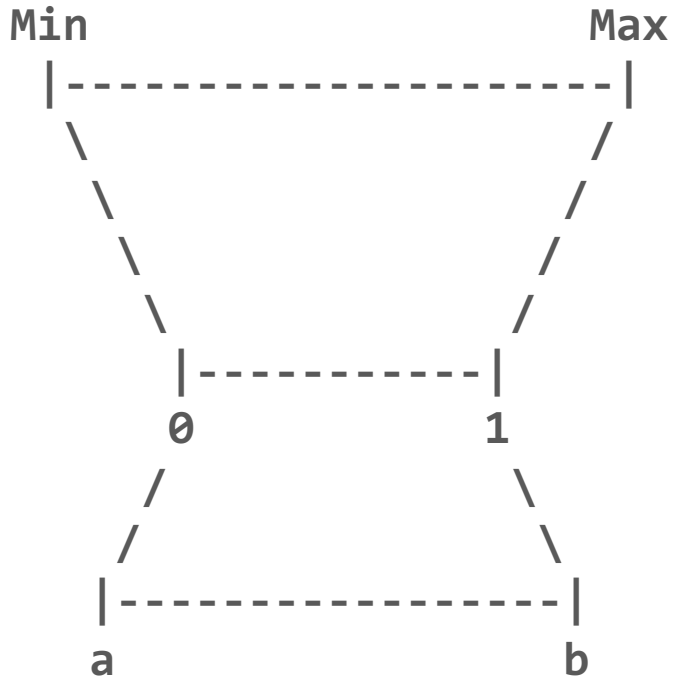
Data Processing – Min / Max Scaling

- Formula:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Formula to rescale the values to fall within [a, b]:

$$x' = (b - a)\frac{x - X_{min}}{X_{max} - X_{min}} + a$$



NYU Langone Health

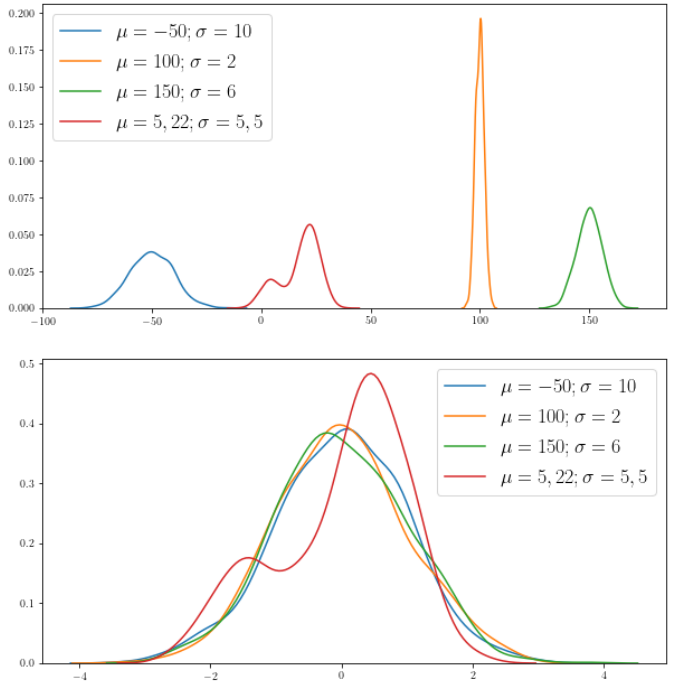
Data Processing - Standardization

- Formula:

$$Z = \frac{X - \mu}{\sigma}$$

- μ: mean
- σ: standard deviation

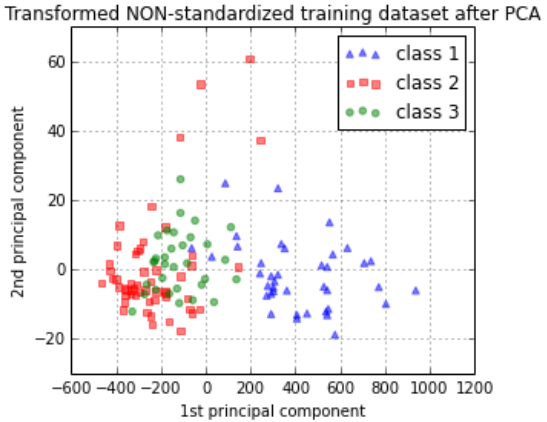
- Standardization will maintain the shape of each distribution

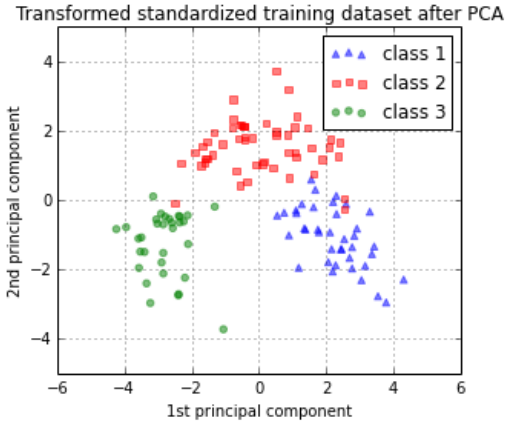


NYU Langone Health

Data Processing – Normalization Applications

- Use the right normalization strategy for the task
  - Principal Component Analysis (PCA)
    - Finds the components that maximize variance
    - Works best of standardized data
  - Image Analysis (values in range 0-255)
    - Min Max Scaling is often applied before using images for machine learning
    - By multiplying the normalized image by 255 you can revert back to the original image



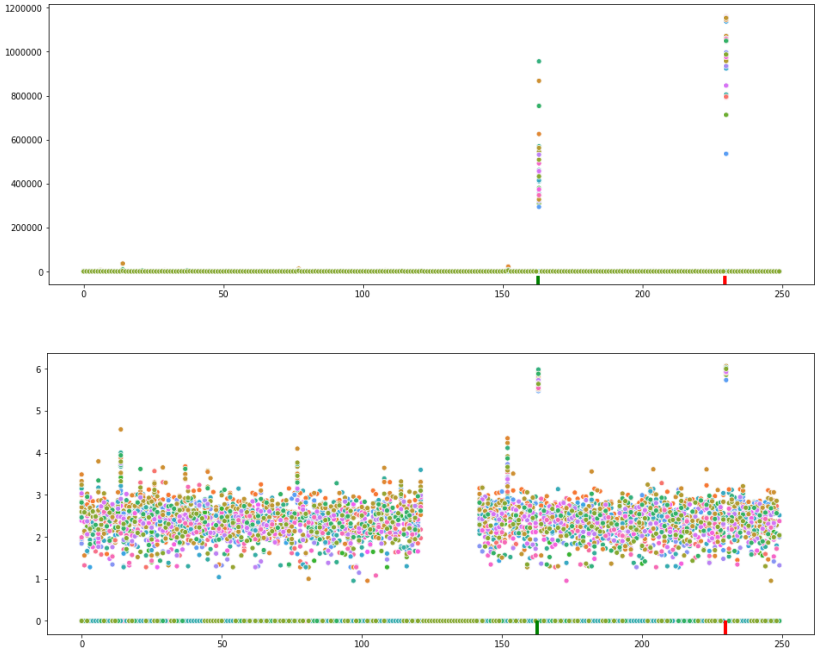


https://sebastianraschka.com/Articles/2014_about_feature_scaling.html

NYU Langone Health

Data Processing – Log Transformations

- Helpful when there are multiple orders of magnitude between values in the data
- Instead of viewing the distance between each sample you view the order of magnitude of the distance
- Values must be greater than zero
  - `Log10( 0.1) -> -1`
  - `log10(   1) -> 0`
  - `log10(  10) -> 1`
  - `log10( 100) -> 2`
  - `log10(1000) -> 3`
  - `...`

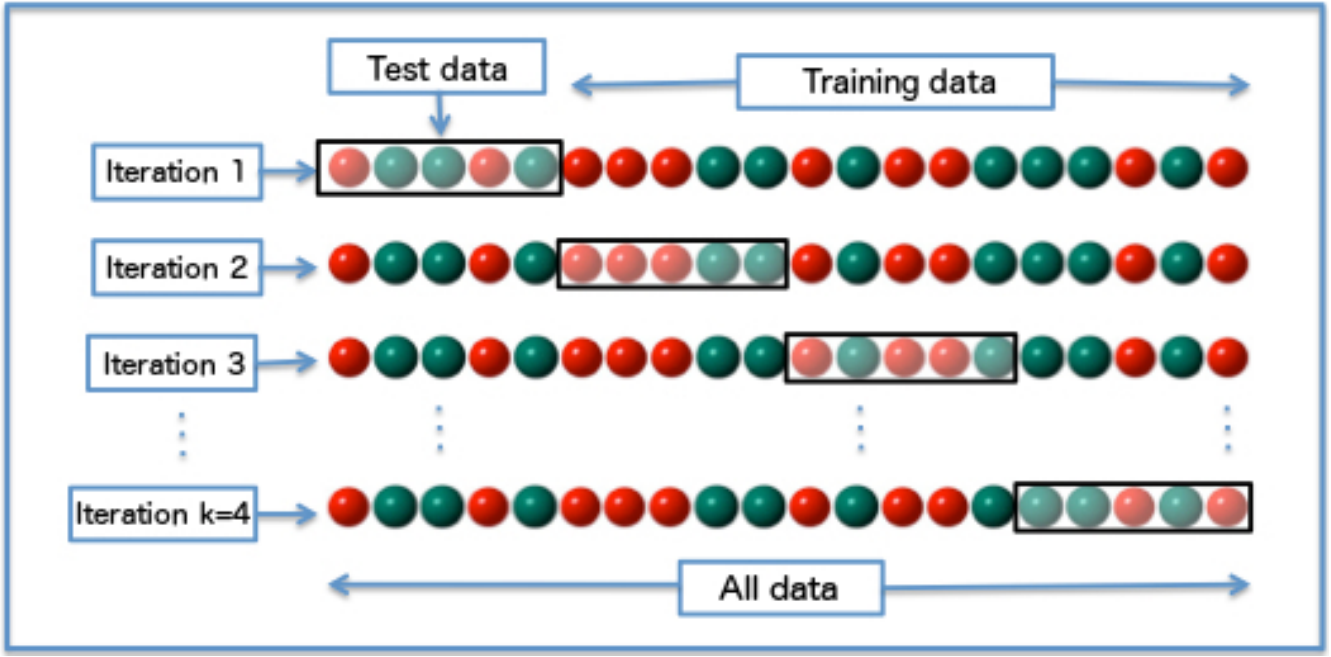


NYU Langone Health

Data Processing – Predictive Modeling

- Using statistics to predict outcomes
- For the Golub data the model will make a categorical prediction
- Scikit-Learn includes several commonly used models
  - Logistic Regression, Neural Networks, K Means Clustering, …
- A model is trained by feeding it labeled data
  - It will then adjust weights so that when shown that data again it will predict the labeled class
- General process for predictive modeling
  1. Split your data into a training and test set
  2. Train the model using the training set
  3. Test the performance of the model using the test set

Data Processing –Cross Validation



https://en.wikipedia.org/wiki/Cross-validation_(statistics)#/media/File:K-fold_cross_validation_EN.jpg

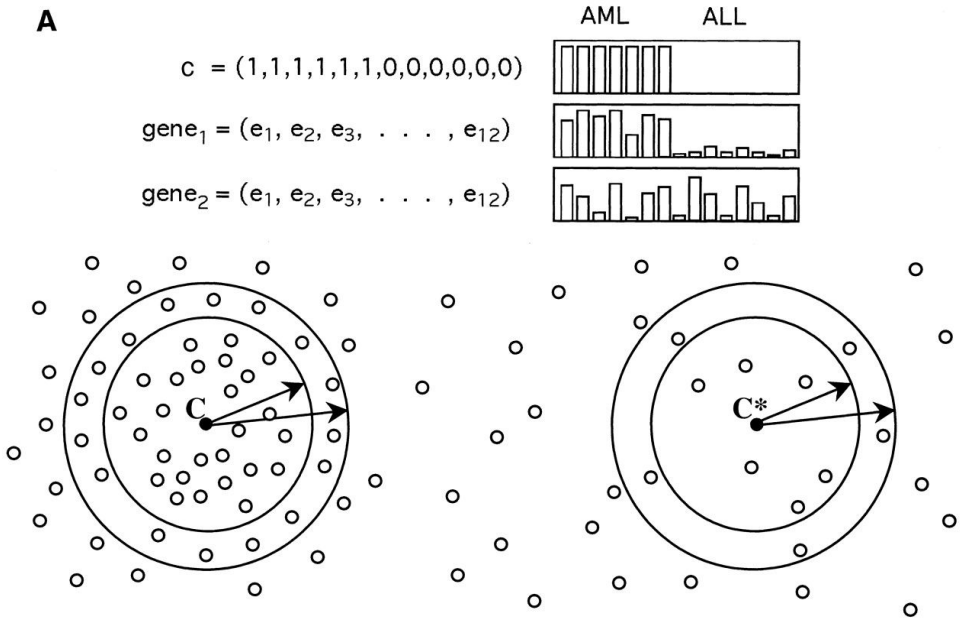
NYU Langone Health

Break / Questions

Golub Paper – Breakout Sessions

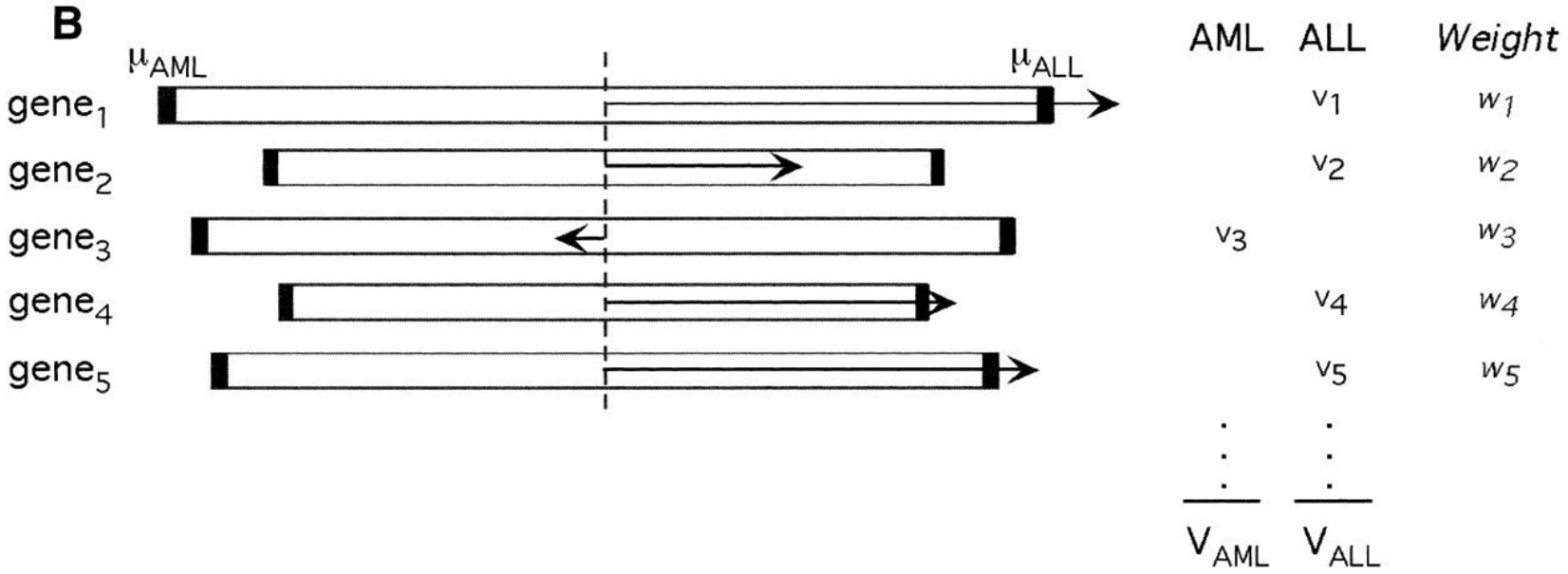
- You will be divided into random breakout rooms
- The purpose of these sessions is to discuss the paper
  - Some points that you might want to discuss were posted on Brightspace
    - I will also post them on Slack now in the “breakout-discussions” channel
- We will regroup in 15 – 20 minutes

Golub Paper – Figure 1A



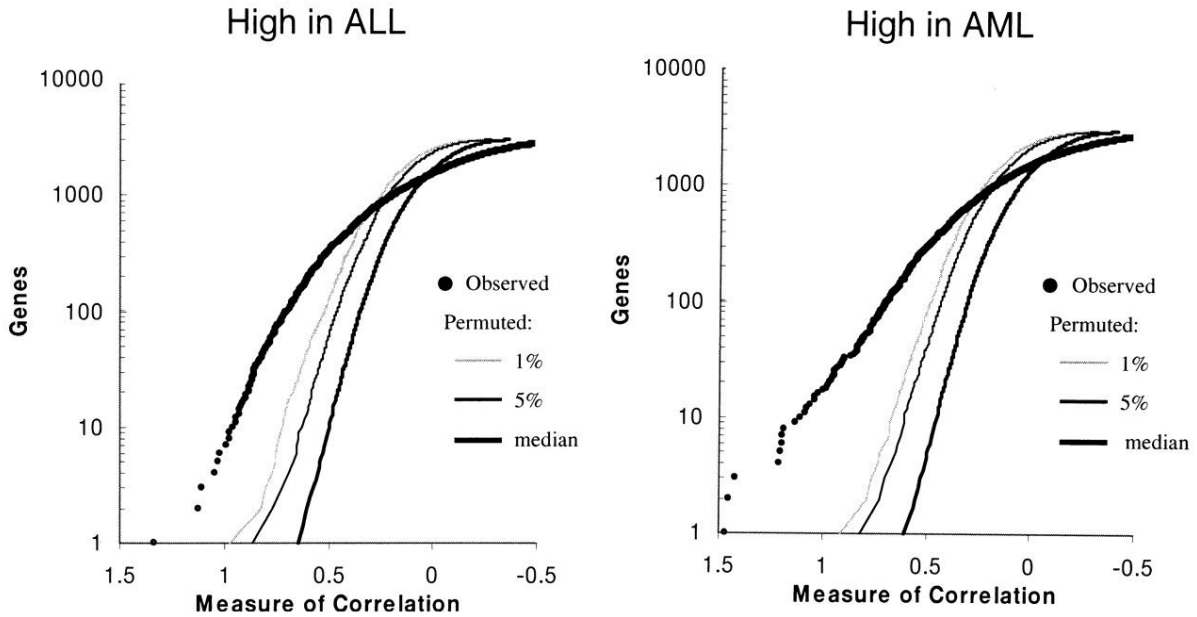
NYU Langone Health

Golub Paper – Figure 1B



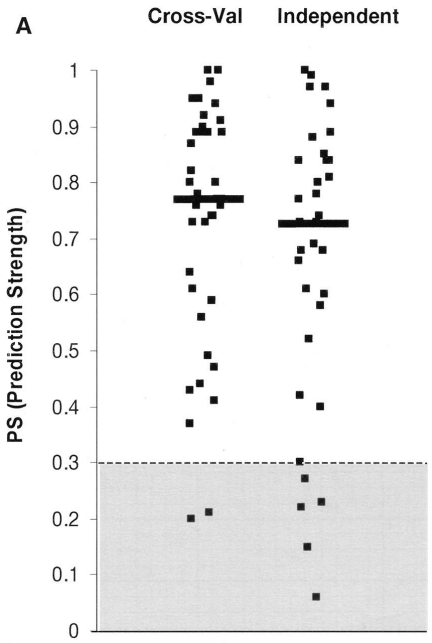
NYU Langone Health

Golub Paper – Figure 2



NYU Langone Health

Golub Paper – Figure 3A



NYU Langone Health

Golub Paper – Feature Selection

- For each gene
  - Take the mean and standard deviation for each of the classes
  - P(g,c) uses those metrics to measure the correlation
  - This metric emphasizes the “signal-to-noise” ratio
  - This metric is not commonly used
- For your analysis use Welch’s t-test
  - This tests the hypothesis that two populations have equal means

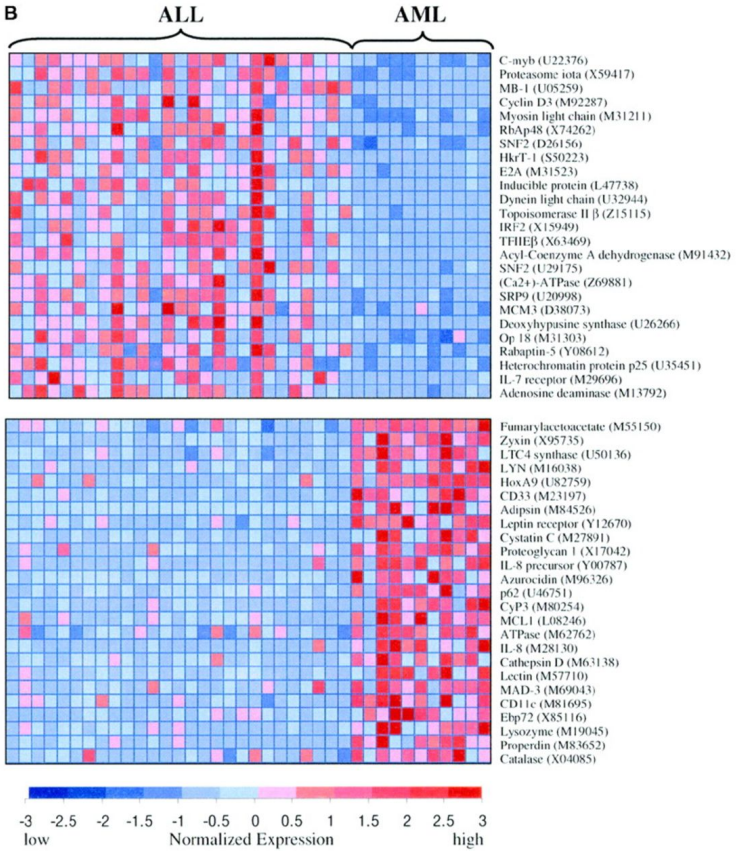
- Golub corelation metric:

$$P(g,c) = \frac{\mu_1 - \mu_2}{\sigma_1 + \sigma_2}$$

- Welch’s t-test:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Golub Paper – Figure 3B



NYU Langone Health

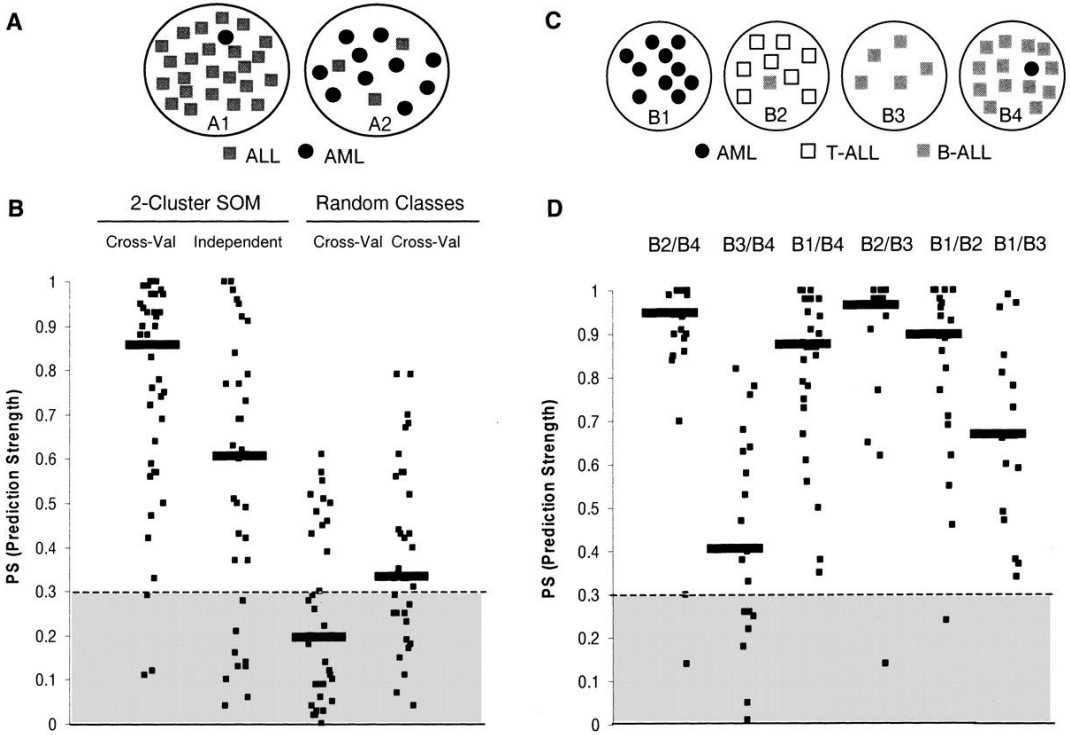
Golub Paper - Results

- Their model used the 50 top genes most correlated with the AML / ALL distinctions
- Train Data
  - 36/38 samples assigned correctly
  - 2 samples with predictive strength < 0.3
- Independent Data
  - 29/34 correct predictions
  - 5 samples with predictive strength < 0.3
  - Predictive strength is lower in AML samples from St-Jude (difference in protocol)

Golub Paper - Class Discovery

- Using only the gene expression to classify the samples
- Their method used Self Organizing Maps (SOMs)
  - This method requires the number of classes to assign to be specified
  - This method is not commonly used anymore
- When using 2 classes:
  - Class 1 contained 24/25 ALL
  - Class 2 contained 10/13 AML
- When using 4 classes
  - Distinction between ALL and AML
  - Distinction between T-Cell and B-Cell ALL

Golub Paper – Figure 4



NYU Langone Health

Assignment

1. Recreate Figure 3b - See how many overlaps you have with their top correlated genes
2. Train a model using the 50 most highly correlated genes (25 AML and 25 ALL) to predict the class of each sample using cross validation.
3. Try differing amounts of genes and compare the results e.g.: (top 2, 10, 100, 1000)
4. Cluster the data using any method you like (K-Means, Hierarchical). What are the defining features of the clusters.

- Optional: Does sample type matter (bone marrow vs peripheral blood)?

Groups

- Groups will be randomly assigned for each topic
- I will send out an email later today listing each group
  - All the groups members should be on Slack
  - Contact me if you have any concerns
- Only enrolled students will be assigned a group
- Auditors can join any group that they like
- Currently I am planning on having groups of 4-5 students
  - I recommend that you all do the analysis independently share your progress within your group

Topics Covered in Thursdays Lab

1. Loading a dataset from into Pandas
    1. Cleaning the dataset
    2. Feature selection
2. Generating Heat Maps from the Pandas Dataframe
    1. Shaping our data
    2. Using Seaborn to create a heatmap
    3. Using Seaborn to create a cluster map
3. Shaping data for use with Scikit-learn
    1. Splitting the data into a feature set and a label set
    2. Leave one out cross validation
4. Measuring model performance
    1. Accuracy
    2. Generating a confusion matrix

References

- https://sebastianraschka.com/Articles/2014_about_feature_scaling.html
- Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring BY T. R. GOLUB, D. K. SLONIM, P. TAMAYO, C. HUARD, M. GAASENBEEK, J. P. MESIROV, H. COLLER, M. L. LOH, J. R. DOWNING, M. A. CALIGIURI, C. D. BLOOMFIELD, E. S. LANDER SCIENCE15 OCT 1999 : 531-537
- https://www.news-medical.net/life-sciences/History-of-Microarrays.aspx#targetText=Introduction%20of%20miniaturized%20microarrays,printing%20of%20cDNA%20on%20glass.
- Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and its applications. *J Pharm Bioallied Sci*. 2012;4(Suppl 2):S310-S312. doi:10.4103/0975-7406.100283
- https://en.wikipedia.org/wiki/Welch%27s_t-test

NYU Langone Health