SUPREME: a cancer subtype prediction methodology by integrating various data types

Jeanne Su with Ziynet Nesibe Kesimoglu and Dr. Serdar Bozdag
Department of Mathematics, Statistics, and Computer Science, Marquette University

Introduction

Cancer

- ➢ The second leading cause of death in the US
- ➢ Heterogeneous
  - The subtype a cancer patient has is essential for accurate diagnosis & prognosis.
- ➢ Caused by genetic & epigenetic changes
  - There is a lack of reliable biomarkers, (traits that indicate a certain biological state), for cancer diagnostic & prognostic purposes [1].

Background

- ➢ Many biological datasets have been created from normal and tumor tissue samples.
- ➢ Some studies use one type of dataset.
  - However, each data type impacts the body differently.
- ➢ Thus, other studies integrate data types by preselecting important data and combining them.
  - The combined data is reliant on preselection → important information could be lost.

Research Question

We want to develop a method that predicts cancer subtype by combining multiple data types without losing important information.

Terms

microRNA

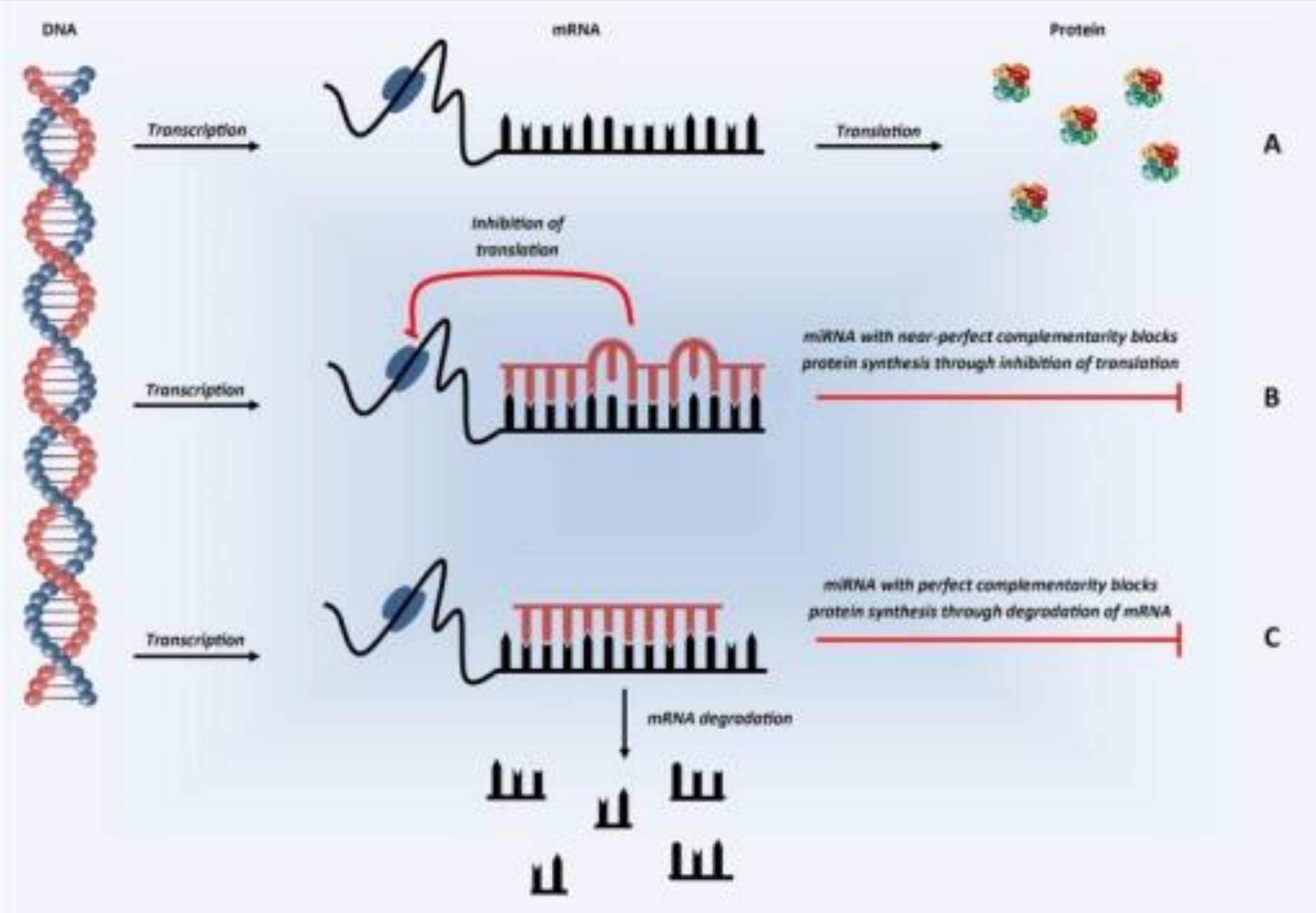
DNA
mRNA
Protein
Transcription
Translation
A
Inhibition of translation
Transcription
miRNA with near-perfect complementarity blocks protein synthesis through inhibition of translation
B
Transcription
miRNA with perfect complementarity blocks protein synthesis through degradation of mRNA
C
mRNA degradation

Figure 1. The mechanism of microRNA [2]

microRNA is a non-coding RNA involved in regulation of gene expression:

- ➢ if microRNA binds to mRNA in almost perfect match → protein synthesis is blocked
- ➢ if microRNA binds to mRNA in perfect match → mRNA is degraded

DNA methylation

A methyl (CH_3) group is added to DNA
↓
DNA is packed tightly together so nothing can bind to it
↓
The genes are not expressed

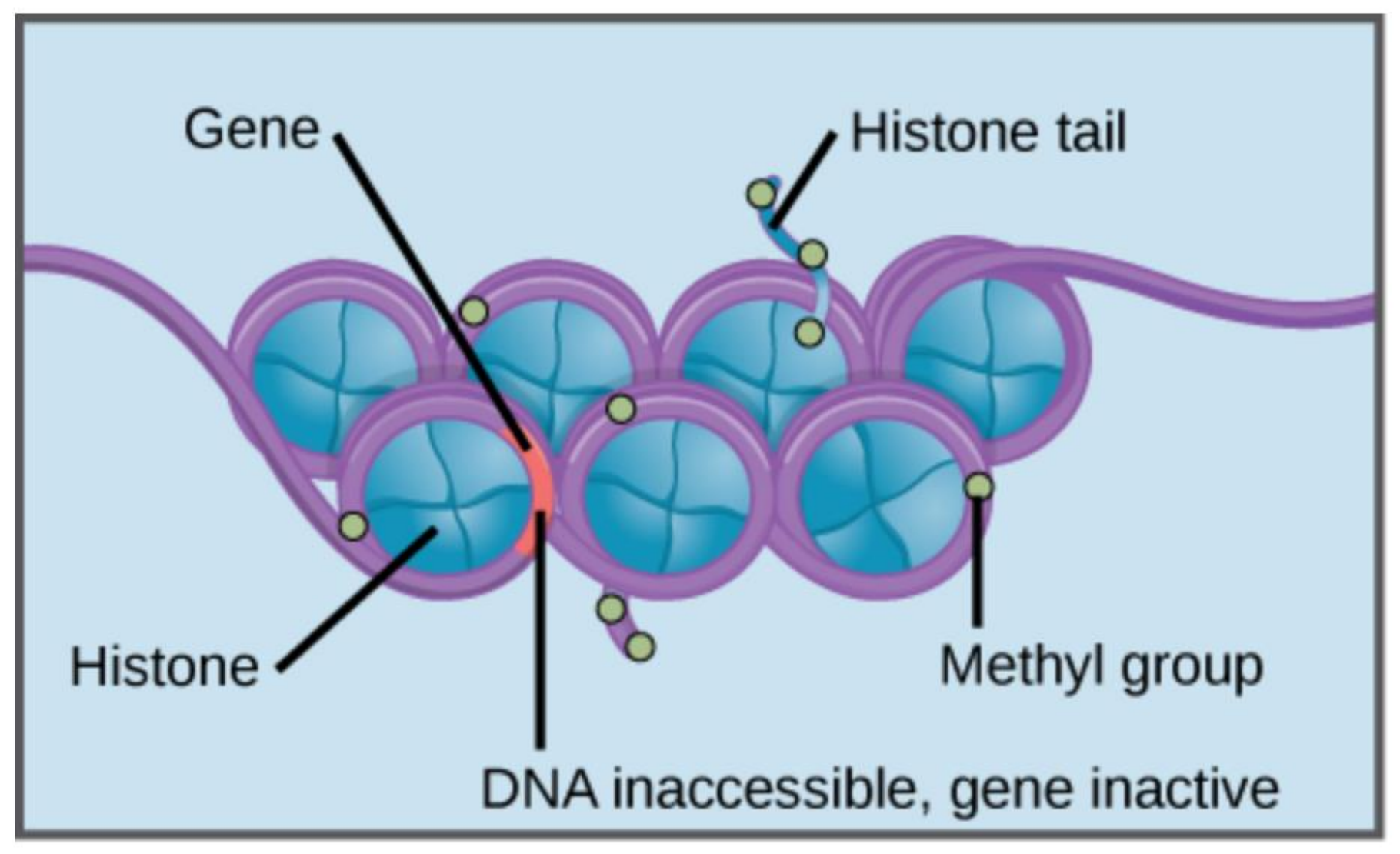

Figure 2. Effect of DNA methylation [3]

Method Outline

Data

Breast cancer tumor dataset from The Cancer Genome Atlas (TCGA)

- ➢ Data types for integration (numbers correspond with the datasets in Fig. 3):
  1. Gene expression
  2. microRNA expression
  3. DNA methylation
  4. Copy number variation (CNV): the number of times a DNA section is repeated
  5. Somatic mutation
- ➢ The main intrinsic breast cancer subtypes:
  - Luminal A
  - Luminal B
  - Basal-like
  - HER2-enriched
  - Normal-like

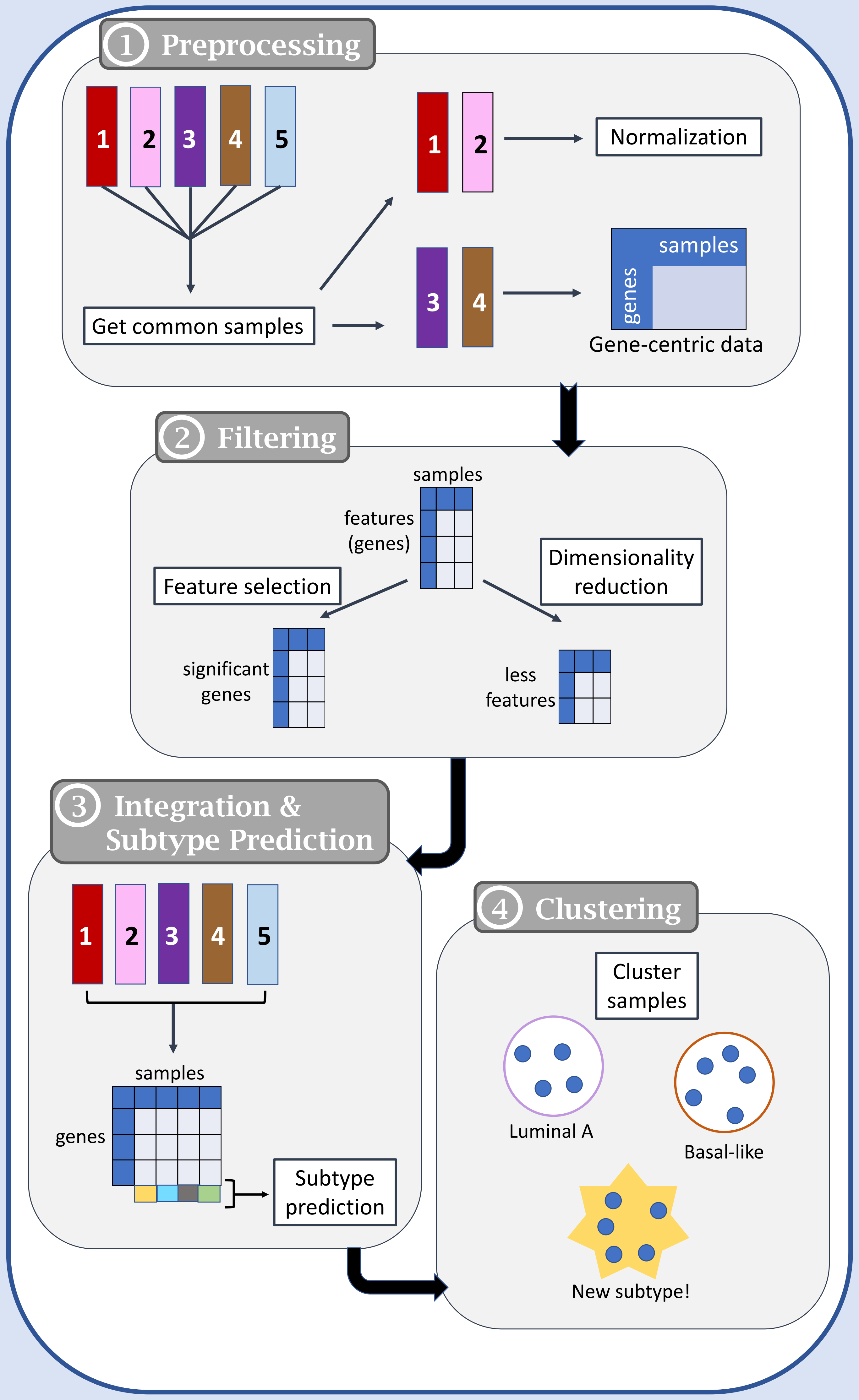

Figure 3. Diagram of method outline.

Methods

1. **Preprocessing:**
   - Get common samples from all data types
   - Normalize gene expression & microRNA expression data
   - Convert DNA methylation and CNV data to gene-centric data
2. **Filtering:**
   - Feature selection → significant genes as biomarkers
   - Dimensionality reduction → reduce number of features
3. **Integration & Subtype Prediction:**
   - Combine the data types together
   - Build classification model that predicts breast cancer subtype of the sample
4. **Clustering:**
   - Cluster the samples to potentially find new breast cancer subtypes

Conclusion

- ➢ We have completed the first step, preprocessing, and are now working on the remaining steps.
- ➢ Next steps:
  - ➢ Determine what the features should be (keep genes as features or use something else as features)
  - ➢ Determine how to combine the data types together
  - ➢ Determine how to classify the subtype of each sample

References

[1] Chakraborty, S., & Rahman, T. (2012). The difficulties in cancer treatment. *Ecancermedicalscience*, *6*, ed16. http://doi.org/10.3332/ecancer.2012.ed16

[2] Romaine, S. P. R., Tomaszewski, M., Condorelli, G., & Samani, N. J. (2015). MicroRNAs in cardiovascular disease: an introduction for clinicians. *Heart*, *101*(12), 921–928. http://doi.org/10.1136/heartjnl-2013-305402

[3] Libretexts. (2018, June 06). 16.3: Eukaryotic Epigenetic Gene Regulation. Retrieved June 25, 2018, from https://bio.libretexts.org/TextMaps/Introductory_and_General_Biology/Book:_General_Biology_(OpenStax)/3:_Genetics/16:_Gene_Expression/16.3:_Eukaryotic_Epigenetic_Gene_Regulation

Acknowledgments

This research was sponsored in part by an REU grant from Marquette University's Department of Mathematics, Statistics and Computer Science. Many thanks to Dr. Serdar Bozdag and Ziynet Nesibe Kesimoglu for their guidance and mentorship, and to Dr. Dennis Brylow and Dr. Petra Brylow for their efforts in running this REU as smoothly and successfully as possible.