Jen & Emily :D
EEB5300
Final Project Write-up
7 May 2021

Overview and Motivation:

Type 1 Diabetes (T1D) is an autoimmune disease caused by genetic and non-genetic factors, such as the environment, diet, and epigenetics. Beta islet cells of the pancreas are responsible for taking in glucose and secreting insulin. With T1D, these beta cells are destroyed by the immune system, meaning the individual cannot produce insulin and will have increased blood sugar levels. While T1D is mainly diagnosed in children, teens, and young adults, it can actually develop at any age (https://www.cdc.gov/diabetes/basics/what-is-type-1-diabetes.html).

Similarly to T1D, Type 2 diabetes (T2D) is caused by both genetic and non-genetic factors. With T2D, muscle, fat, and liver cells become resistant to insulin and do not take in enough sugar. Additionally, the pancreas is unable to produce enough insulin, leading to dysregulated blood sugar levels. T2D is usually diagnosed in older adults, but like T1D, it can also be diagnosed at any age (https://www.mayoclinic.org/diseases-conditions/type-2-diabetes/symptoms-causes/syc-20351193).

Motivation for this project came from a shared interest in exploring methylation differences and whether these differences contributed to the development and/or progression of a certain disease. Our group chose to examine methylation differences in Type 1 and 2 diabetes. Since these two diseases are similar in nature, yet still different enough to be classified separately, we thought it would be interesting to see how both diseases compared to each other as well as to individuals without the disease. Another motivational factor is that the Lynes lab, which Jen is a part of, has studied how specific antibodies can diminish the severity of both T1D and T2D.

Related Work/Citations: If you are using public data, you would want to include relevant descriptions of how the data was generated and cite this information. Your analysis should be different from anything published on the data. A full description of how this data was generated (study design, libraries, sequencing, etc) should be included.

Data from the study "Identification of Type 1 Diabetes-Associated DNA Methylation Variable Positions That Precede Disease Diagnosis" was used for analysis in R. The subjects in this study were monozygotic (MZ) twins; the twins were either normal MZ or T1D discordant, where one twin had T1D and the other did not. The authors of this study conducted genome-wide DNA methylation analysis of CD4 tissue and CD14+ monocytes from both types of MZ twins enrolled in the study. This analysis was done with Illumina 27K, which allows for DNA methylation measurements at 27,458 different CpG sites across the genome. The main objective of this study was to examine methylation differences between T1D and Unaffected MZ twins to identify T1D methylation variable positions (Rakyan et. al. 2011).

For this project, two different comparisons were made based on the samples present in the dataset. The first comparison made was between T1D/unaffected discordant twins CD4 tissue and normal MZ twins CD4 tissue. The results presented in the study mainly focused on the CD14+ monocytes, which is why we chose to analyze the CD4 tissue samples in more depth ourselves. We also made a comparison between T1D patients CD4 tissue samples and CD14+ monocytes samples. For both, the group ran the R script through GEO2R, which was available on the NCBI website. The R script was modified to reflect the specific samples we were selecting for both comparisons. An attempt was made to run quality control on the selected samples in R, however, the issues we had were not resolved in time.

Data from the study “Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the FTO Type 2 Diabetes and Obesity Susceptibility Locus” was also used for analysis in R. The data did not provide too much information about the subjects, such as age or sex, and the paper mentions an all-female set was excluded from analysis to remove any effects relating to sex. A total of 60 samples from Caucasian individuals were included in the study -- 30 control and 30 patients with T2D. The authors of this study utilized both the NimbleGen human 385k tiling array and MeDIP. The tiling array is used to probe intensively for sequences that are known to exist in a contiguous region, and the 385k array can tile up to 385k 50- to 75-mer probes. MeDIP, or methylation DNA immunoprecipitation, is performed to quantify DNA methylation (Bell et. al. 2010).

Due to the limitation of variability in the data, all 60 samples were analyzed together. All of the samples appeared to be from whole blood DNA. A comparison was made between the

control and T2D samples, with the R script being run through GEO2R, available through the NCBI. Quality control was attempted in R, however, as with the T1D data the issues were not resolved in time.

References:

1. Rakyan, V. K., Beyan, H., Down, T. A., Hawa, M. I., Maslau, S., Aden, D., Daunay, A., Busato, F., Mein, C. A., Manfras, B., Dias, K. R., Bell, C. G., Tost, J., Boehm, B. O., Beck, S., & Leslie, R. D. (2011). Identification of type 1 diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS genetics*, *7*(9), e1002300. https://doi.org/10.1371/journal.pgen.1002300
2. Bell, C. G., Finer, S., Lindgren, C. M., Wilson, G. A., Rakyan, V. K., Teschendorff, A. E., Akan, P., Stupka, E., Down, T. A., Prokopenko, I., Morison, I. M., Mill, J., Pidsley, R., International Type 2 Diabetes 1q Consortium, Deloukas, P., Frayling, T. M., Hattersley, A. T., McCarthy, M. I., Beck, S., & Hitman, G. A. (2010). Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PloS one*, *5*(11), e14040. https://doi.org/10.1371/journal.pone.0014040

Initial Questions: What questions are you trying to answer? How did these questions evolve over the course of the project? What new questions did you consider in the course of your analysis?

The initial questions our group had are as follows:

1. What are the differences in DNA methylation patterns in subjects with T1D, or T2D, and healthy individuals (control)?
2. What are the differences in methylation patterns between subjects with T1D versus T2D?
3. Are there differences in the accumulation of alternative methylation that are age-dependent? Does this change with age of onset / diagnosis?
4. Do specific CpG islands accumulate methylation changes faster than others? If so, what genes are nearby?
5. Is there a correlation between increased methylation and decreased insulin production?
6. How can we assess the accuracy and/or completeness of the data using tools we learned in class?

Over the course of the project, there were certain questions we had to abandon due to lack of data. For example, the studies used for both types did not include any data on insulin

production, meaning we could not determine if there was a correlation between increased methylation and decreased insulin production. Another question we could not answer was whether methylation differences existed between both types because the T2D datasets were limited. Any questions relating to age, at least for T2D, could not be investigated, again due to the data limitations. Lastly, the quality control issue we were having in R was not resolved.

For the T1D comparisons done between T1D/Unaffected MZ twins and healthy twins, there were methylation differences seen when looking at the C4 tissue samples. Comparison between the T1D discordant twins and normal MZ twins showed that nine of the ten top differentially expressed genes had CpG sites present and that there was hypermethylation and hypomethylation present at these CpG sites. This relates to one of our initial questions; we wanted to look at CpG methylation differences between T1D and control subjects. This result did not quite answer our initial question regarding CpG sites. There were no samples present at various time points for each individual included in the study, therefore we could not determine if certain CpG sites accumulated methylation faster than others. However, we did find the genes where methylation differences were present, and a majority were present within CpG islands.

One question we had after this analysis was whether this methylation difference seen would be the same if the study was not done in MZ twins. The T2D comparisons done, which did not use twins, were not as substantial as the T1D. The other difference between the two datasets was tissue type. The T1D data was from CD4 tissue and the T2D data was from whole blood. It would be interesting to do a future study with T2D discordant twins and / or with different sample types to determine if either of these conditions have an effect on the results.

Data: Raw data uploaded in T1D and T2D folders. Datasets obtained through the NCBI BioProject.
Links:

- T1D: https://www.ncbi.nlm.nih.gov/gds?Db=gds&DbFrom=bioproject&Cmd=Link&LinkName=bioproject_gds&LinkReadableName=GEO+DataSets&ordinalpos=1&IdsFromResult=244074

- T2D: https://www.ncbi.nlm.nih.gov/gds?Db=gds&DbFrom=bioproject&Cmd=Link&LinkName=bioproject_gds&LinkReadableName=GEO+DataSets&ordinalpos=1&IdsFromResult=125195

Workflow/Analysis: What visualizations did you use to look at your data in different ways? What are the different statistical methods you considered? Why did you select the packages you did?

Our group used R to analyze the selected samples for both the T1D and the T2D datasets. For both types, a similar workflow was followed. The samples we wanted to compare were placed into separate groups and the R script was run using GEO2R. The R script was modified based on which samples were selected for comparison. The R script generated various graphs and plots including a volcano plot, boxplot, and expression density plot. The R libraries we used were GEOquery, limma, umap, and minfi. GEOquery was used as a connection between the GEO public repository and BioConductor. Limma was used to analyze the microarray data samples present in the dataset. Umap helped with data visualization. Lastly, minfi was used to analyze and visualize the Illumina methylation array data.

Overall, we thought the volcano plot provided us with the most information for the T1D data. The volcano plot showed us differences in DNA methylation between the groups being analyzed/compared. Each point on this plot represents a CpG site, which was exciting for us because this related back to one of our group's initial questions. Black dots represented significantly differentially methylated CpGs, blue dots indicated hypomethylation, and red dots indicated hypermethylation.

On the other hand, for the T2D data, none of the plots generated in R were all that telling. This contradicts the paper from which the data came from, which reported that the *FTO* gene locus was significant in terms of methylation. However, the volcano plot in R showed no significant up- or down-regulated genes between the control and T2D subjects. The limma plot indicated that of 387,835 chromosomal loci tested, 0 were significant between control versus T2D subjects. The adjusted p-value plot showed that none of the p-values were below 0.80; it is to our understanding that a p value <= 0.05 is considered statistically significant.

Discussion: What did you learn? What additional sequencing and/or analysis is necessary? What questions remain? What new insights were gained?

This project was helpful for us to learn where to access public data as well as how to download and utilize the data to answer a question of our choosing. One of the bigger obstacles we had to overcome initially was how to download/access the data. The data was not in a format we could easily view on our laptops and we had to unzip files in order to open certain data in notepad/excel. Our group also learned more about R and how to use it to generate different graphs and plots to visualize the samples we wanted to compare. Being able to select the samples we wanted was very helpful.

The last workflow step our group tried to do was quality control on the samples in R. We attempted to follow along with a Bioconductor methylation array tutorial, however, were unable to properly tell R where to find the directory that had our data in it. If this issue were resolved, we would have been able to use the detP function in minfi to calculate the p-values for our samples. Aside from this, one question our group still has is why the T1D comparison appeared more significant than the T2D comparison. We wonder if it could be due to twins vs no twins or sample tissue type. We also wonder if this could be attributed to one set of data including information such as subject age, and / or the fact that one data set was strictly from Caucasians. More diverse data would, in theory, aid us in resolving some of these lingering questions.