Potential Effects of DNA Demethylation Drugs on Alzheimer's Disease

Haley Abramson and Priya Pillai

Abstract

According to data collected in 2018, approximately 5.7 million Americans suffer from Alzheimer's Disease (AD). Due to the symmetry between changing methylation patterns in the epigenome and AD, insight into the epigenetic changes that occur could shed light on potential drugs to combat the neurodegenerative disease. As research is rapidly progressing on a cure, we seek to advance current research by identifying targets for demethylation drugs that would slow or stop the effects of AD. This identification could predict which drugs might succeed as well as which might harm the patient. We located CpG islands that exist in the methylated state in AD patients and in the unmethylated state in control, healthy patients. After identifying the sequences of these CpG islands, we used Gibbs sampling to find motifs methylated uniquely in AD patients. In addition, we looked into which genes would be affected by a demethylase drug targeting the motifs we identified. We selected 4 common motifs found in AD patient-only methylated CpG islands and identified 59 functional genes that contained these motifs.

Introduction

DNA hypermethylation has been associated with cancer and AD for decades and therefore has been a key target in drug development. Because methylation serves as an indication of disease states, demethylases have been under development since the 1980s. However, methylation is an important part of epigenetics. It determines which genes are expressed; for example, a methylated promoter represses its respective gene(s). Therefore, a drug that demethylates DNA would allow expression of genes that might otherwise not be transcribed. These drugs often work by inhibiting a methylator. Because genome methylation occurs naturally as we age [3], demethylases could act not only on disease-associated genes, but also genes that should stay methylated. In order to prevent off-target effects, there is a significant need to find specific targets where the demethylases should work. For this project, we seek to recommend locations in the genome of microglia cells in AD patients for researchers designing demethylases to take into account as targets when developing drugs against the disease.

In order to study potential demethylation sites, we looked at data from Illumina's Infinium HumanMethylation450 BeadChip array for normal brain tissue and for brain tissue of AD patients collected from The Religious Orders Study and Memory and Aging Project

(ROSMAP) [14]. We received this data from Phuong Pham, Yongjin Park, Matthew Eaton, and Manolis Kellis from their previous work with de Jager. There were 455 AD patients, 280 normal patients, and 31814 CpG island probes in total, some of which showed methylation in only one set of patients or another. We looked at brain tissue because as both an immune disease and a neurodegenerative disease, AD affects methylation states most prominently in microglia, the immune cells of the brain. An ideal location for a demethylase to act might be evident in a gene that is methylated in tissue from patients with AD and demethylated in normal tissue. We looked for complementary regions of methylation in the normal versus AD patient data and visualized AD via changes in methylation patterns to find regions to be targeted. Demethylases typically act by inhibiting one of two types of methylases: ones that methylate CpG islands or ones that inhabit the replication fork, keeping the epigenome constant during mitosis. For the purposes of our project, we focused on identifying CpG islands in order to recognize differences in methylation states of AD patients. Each drug has a slightly different targeting pattern in order to demethylate an area—they can require specific DNA sequences, methylation patterns, or simply CpG islands. The drugs we focused on are DNMTIs, which affect specific methylases.

For these drugs, we looked at the sequences in the genome the methylase targets and simulated the passive demethylation of these areas [6]. We located CpG islands that were methylated in AD patients and not in normal patients in the brain tissue for each chromosome, looked into the gene functions at each of the CpG island locations, and then recommended specific locations for drug targets. Functional data for the affected genes was provided by NCBI, which allowed us to determine the functional risks of the drugs.

Current State of the Field and Our Innovation
Previous literature has looked into the effect of methylation at specific CpG sites in relation to AD [5], the correlation between methylation and aging or cancer [2,3], and the potential for demethylases as drugs against these changes in methylation state [1]. Significant research has been done on potential causes of AD as an immune and neurodegenerative disease, and beyond methylation states, it has been found that histone modifications and non-coding RNAs both contribute to a patient's diagnosis [5]. We also utilized previously determined definitions of CpG islands [8] and methylation data of neural cells in healthy patients by the ROSMAP study [7]. While demethylase drugs have been proposed as treatments for various diseases like AD and cancer, the risks of using these drugs and the optimal target genes for these drugs are not yet known. Our work sought identify potential target sequences for drugs without testing on a human patient, meaning the eventual human trials can be safer. This general approach can be used for other drugs. Algorithmically, we searched for demethylase drug targets in the genome and epigenome by looking for entire CpG islands rather than solely CpG sites, as in other studies. We then searched for motifs among all CpG islands that were methylated uniquely in AD patients. Our modified algorithms could be implemented in a broader software package to do similar analyses in the future.

Methods

All data comes from the ROSMAP Study [7]. We first separated all patients based on AD diagnosis, then located methylated CpG islands on the hg19 genome. If the average beta value was greater than 0.6 for AD patients or normal patients, we considered the CpG site to be methylated [9]. Given a site was methylated, we found the surrounding CpG island on the hg19 genome using the Gardiner-Garden method [8]. According to this method, a CpG island is defined as a region of the genome at least 200 base pairs long composed more than half of G's and C's and satisfying the equation $\frac{\#\ CpG's}{\#C's\ *\ \#G's} * 200\ >\ 0.6$. Utilizing these three constraints, we found CpG islands surrounding and extending each methylated CpG site that passed the initial methylation (beta value) threshold for both AD patients and normal patients.

Following identification of CpG islands, we visualized differences in methylated CpG island locations. We did this first by motif finding. To account for variations in CpG island length spanning from 200 base pairs (bp) to 1,703bp, we aligned CpG island sequences in two separate categories: subsequences of CpG islands found only in AD patients and subsequences of CpG islands found only in normal patients. Multi-sequence alignment was obtained using Clustal Omega in order to obtain a PWM diagram and suggest motifs [10]. We performed motif finding using Gibbs sampling with an epsilon value of $0.045 * (len(motif))$ and tested with multiple motif lengths. Common motifs found specifically in CpG islands methylated in AD patients and not normal patients suggest target sequences for demethylase drugs.

In addition, we sought to characterize the genes found to be methylated only in AD patients to view where we suggest the drugs act upon. To do so, we found the most common motifs suggested by running Gibbs sampling 16 times with a threshold of a motif appearing at least 40 times among all 267 unique subsequences of AD-only CpG islands and all Gibbs runs. We repeated this for all 140 unique subsequences of control patient-only CpG islands and all Gibbs runs and used a threshold of appearing greater than 20 times in the Gibbs sampling results to select potential unique motifs. By storing the motif sequence and location in the genome (chromosome and base pair site), we then wrote a script to search for the genes on which each of these instances of the most common motifs were located. This script accessed Ensembl GRChr37 release 94 to identify genes in the hg19 genome [11]. A subsequent script mapped gene function to genes by accessing NCBI gene database [13].

Results and Discussion
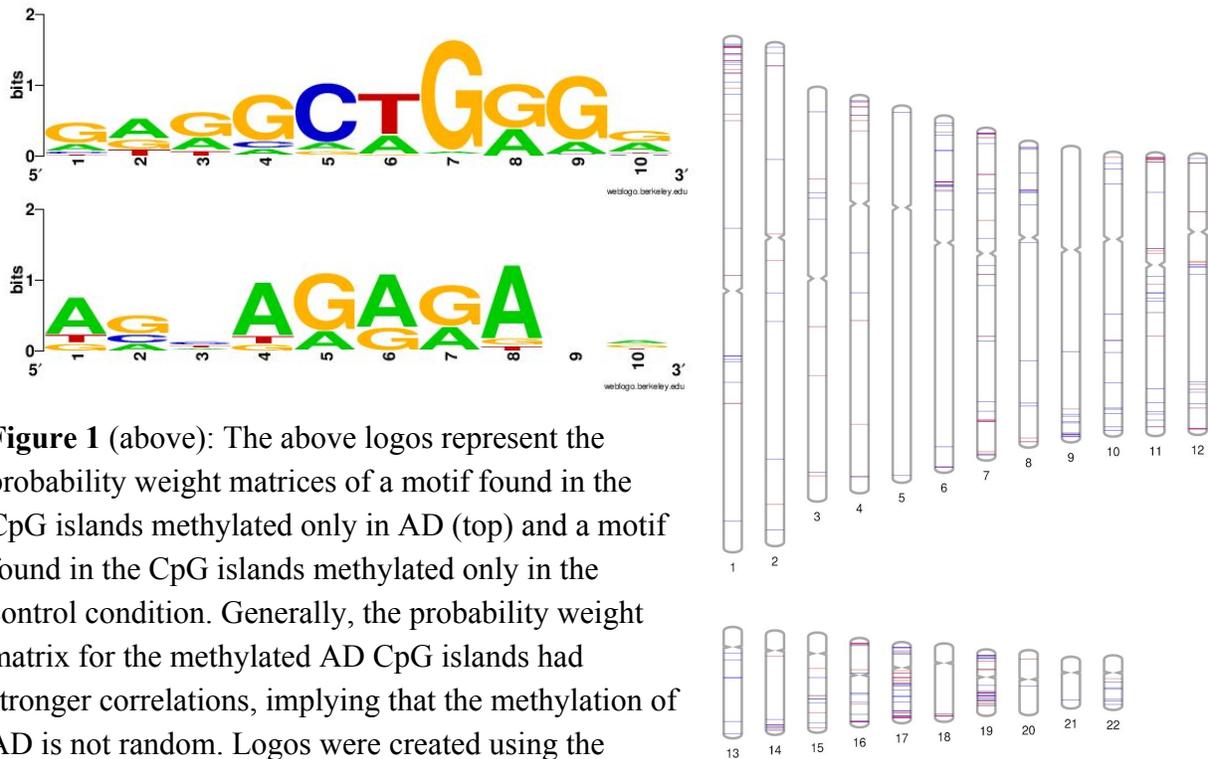
Figure 1 (above): The above logos represent the probability weight matrices of a motif found in the CpG islands methylated only in AD (top) and a motif found in the CpG islands methylated only in the control condition. Generally, the probability weight matrix for the methylated AD CpG islands had stronger correlations, implying that the methylation of AD is not random. Logos were created using the WebLogo web tool by University of California, Berkeley.

Figure 2 (right): The PhenoGram represents the locations of the methylated CpG islands that are exclusively in AD (blue) and those that are exclusively in controls (red). There are significantly more locations exclusively in AD, consistent with the fact that AD is associated with hypermethylation. However, there are still some locations in the control only, indicative of the complex nature of the epigenetic changes of AD (and potentially caused by random noise between individuals). Information on exact counts of methylated CpG islands can be found in Supplemental Table 1. The PhenoGram was created using the PhenoGram Plot web tool by the Ritchie Lab at the University of Pennsylvania.

In order to progress current research that has focused on the causes of AD, this project sought to suggest potential treatments for it. We narrowed down from the idea of using demethylases as drugs to reverse the effects of AD to finding CpG islands that were methylated in brain tissue of AD patients and not in normal patients. Using these unique CpG islands, we performed motif finding in order to locate specific sequences that a demethylase drug could target and checked if these sequences appeared in a relatively high percentage in AD-only CpG islands and in a low percentage in overlapping or control-only CpG islands.

Using the Gardiner-Garden method to determine methylated CpG islands in the AD and control samples was effective. As mentioned previously, AD is correlated with hypermethylation, and accordingly there were more methylated AD CpG islands than methylated control CpG islands. To determine the similarities within the unique AD methylated CpG islands, we started by attempting to align the sequences with Clustal Omega. Unfortunately, there was too much variance in the sequences to allow for an alignment to produce meaningful results-it simply did not align many sequences. The alignment is in Supplemental Figure 3.

Our next attempt to find similarities was using Gibbs sampling to find the motifs within each island. We recorded the number of iterations of Gibbs sampling for each run on the methylated CpG islands, and found that the AD data took significantly fewer iterations (around 20 to 30) to run Gibbs sampling versus the normal data (anywhere from 20 to 100). We interpreted this to mean that the AD data had clearer motifs than the normal data or greater repetition within it compared to the normal data. Importantly, the distinction implies that there is an identifiable difference between the two sets of CpG islands, and this difference could be used as a drug target. Further evidence of this distinction came from searching for the most common motifs in all of the results of Gibbs sampling for AD versus control data. We found that the most common motifs in AD were very common, with 4 AD motifs having significantly more than 40 appearances and no control motifs having greater than 20 appearances.

We not only wanted to know what motifs were common on AD methylated CpG islands, but also what genes those motifs were a part of, and therefore what they controlled. By using the Ensembl database, we identified 59 genes that were related to the 4 AD island motifs. Interestingly, when testing the top 12 AD motifs, we found 78 genes, a relatively small increase. We postulate that this is because there are similar motifs near each other and perhaps multiple motifs working in concert to cause the methylation during AD. For the 59 genes from the 4 AD island motifs, we used the NCBI database to determine the most common functions they were involved in. We found that many were protein-binding, many were DNA transcription factors, and many localized nuclearly (see Supplemental Table 2). This makes sense as AD is an age related disease and likely related to the levels of expression of various genes.

When we investigated the results of 4 AD motifs found from Gibbs sampling in greater detail, we did find some concerning data. We wanted to confirm the uniqueness of the 10 bp motifs that Gibbs sampling identified in the AD CpG islands to those islands. However, when we calculated the number of times our AD motifs were found in the control specific islands and overlapping islands, we found the AD motifs were equally likely to be found in any of these categories. This implies that our motifs may have been focused on too repetitive regions--a particularly difficult issue for CpG islands, which are known to be repetitive. While this might seem to be an overwhelming problem for our goal of finding specific targets, we do not believe it will be, as the differences in the behavior of Gibbs sampling imply there is something fundamentally different about these two sets of sequences, whether or not we can find it via motif finding specifically.

As a consequence of our modifications of standard methods, our work has additional limitations and constraints. For example, the data from the Illumina's Infinium HumanMethylation450 BeadChip array only covers 1.5% of CpG sites in the human genome and looks mainly only at promoter regions [12]. While the methylases we sought to target as inhibitors act mainly upon CpG islands at promoter regions, our data and results lack completeness by excluding parts of the genome. Since promoter regions tend to be repeated or constant across genes, our work to find uniquely methylated regions was constrained by only looking at generally repetitive regions. In addition, we were constrained computationally by our motif-finding algorithm. We were accurate in finding motifs that exist in high percentages in AD-only CpG islands and in very low percentages in overlapping or normal-only islands, but we need to increase the difference between these percentages. To do so, we should look for longer motifs rather than 10bp motifs. Therefore, we should investigate using a different motif finding algorithm or a different manner of searching for long, repeated sequences. While we believe our method of finding the most common motifs is valid, because it is not standard, we do not have a significance value for the results of motif-finding and therefore are limited in the definitiveness our conclusions.

We followed a series of innovative methods to find suggested drug target motifs. Our use of Gibbs sampling as a motif-finding algorithm within CpG islands provided unique data in the context of AD. From this data, we looked for the most common motifs and searched for which genes they appeared in using Ensembl. Our methods directly return the functions of genes that contain each of these selected motifs, translating a chromosome number and base pair into a gene name, and from the gene name into the gene function as provided by NCBI. This streamlined approach holds potential for use in other research to identify drug targets in the genome and possible off-target effects.

In designing our experiments and methods, we made a few choices to lead us towards unique motifs that exist in AD-specific methylated CpG islands. We focused only on AD-specific islands because we did not want a drug to have off-target effects and demethylate a part of the genome that should be methylated. Our Gibbs sampling method implemented an epsilon value of $0.045 * (len(motif))$ to account for the difficulty of convergence when searching for longer motifs. In the end, however, we selected 10bp motifs because PWM logos generated from motifs longer than 10bp appeared incohesive and random. Additionally, when searching for the most common motifs, we chose a threshold that eliminated the possibility of any control-specific motifs appearing often enough to be significant. Due to the commonness of the motifs and the short size of 10 bp, we did not use a probability weight matrix to identify motifs, but rather looked for exact matches. Because the maximum number of appearances throughout all 16 runs of Gibbs sampling among all control-specific CpG islands was 20, we looked at motifs that appeared in this context for AD patients greater than 40 times. This factor of 2 to determine the threshold is based on the fact that there were approximately twice as many AD-specific CpG islands as control-specific islands. As explained above, this resulted in 4

AD-specific motifs. We decided to exclude any control motifs because we assumed these motifs were likely random, as explained by the significantly higher number of iterations until convergence of the control-specific data.

Our project, although unable to provide specific results for drug targets, lays the groundwork for identifying not only drugs to combat AD but rather any genetically linked disease. The methods we developed to identify specific genes and gene functions directly from microarray data have many further applications and can be expanded and modified to improve our results.

Future Steps

Despite the incredible advances in knowledge surrounding AD thanks to collaborative research over the past few decades, it is clear further progress is required before translating our results and the results of other scientists into clinical contexts. Specific to our project, we suggest the following steps to reach a positive conclusion and recommend a drug target for demethylases in AD patients.

First, we need to determine a significance threshold for our results, especially the motifs we look into as being the most common. As explained, because our experiments went beyond standard methods, we do not currently have a means of identifying a motif as being an integral part of AD-specific methylated CpG islands other than recognizing it as a repeated sequence. One advancement that would help with finding a significance threshold would be selecting a method other than Gibbs sampling for motif finding. Sequences of 10bp motifs appear too often in parts of the genome, indicating the necessity for longer motifs as drug targets. However, we also saw that motifs of length 20 appeared random and inconclusive from the PWM logos. Therefore, we suggest searching for highly probable sequences of longer lengths rather than specific sequences. Because proteins typically work by binding the minor (less specific) groove of DNA, locating a highly probable sequence as a drug target would work to combat AD. Another option for a variant of Gibbs sampling would be an algorithm that finds sequences more common to AD-specific methylated CpG islands than overlapping or control-specific islands. We could also look into using suffix trees to identify such sequences.

While we implemented the commonly used Gardiner-Garden definition for CpG islands, there are other ways to distinguish regions of the genome that differ between AD and control patients. Many papers use Hidden Markov Models (HMMs) to identify CpG islands. We could then view the differences in these CpG islands between our two classes of patients differently than simply by base pair, for example using another HMM to identify methylated CpG regions as being AD-associated, control-associated, or both. In summary, we need to refine our algorithms and ensure sequence specificity to AD methylation before confidently suggesting drug targets to combat the disease.

Sources

[1] Delgado-Morales, Raúl, et al. "Epigenetic Mechanisms during Ageing and Neurogenesis as Novel Therapeutic Avenues in Human Brain Disorders." *Clinical Epigenetics*, BioMed Central, 29 June 2017, www.ncbi.nlm.nih.gov/pmc/articles/PMC5493012/.

[2] Horvath, Steve. "DNA Methylation Age of Human Tissues and Cell Types." *Genome Biology*, BioMed Central, 21 Oct. 2013, www.ncbi.nlm.nih.gov/pmc/articles/PMC4015143/.

[3] Oberdoerffer, Philipp, and David A. Sinclair. "The Role of Nuclear Architecture in Genomic Instability and Ageing." *Nature News*, Nature Publishing Group, 1 Sept. 2007, www.nature.com/articles/nrm2238.

[4] Texas Department of State Health Services, Alzheimer's Association. "Alzheimer's Disease - Questions and Answers." *Texas Department of State Health Services*, 22 Mar. 2018, dshs.texas.gov/alzheimers/qanda.shtm.

[5] Jager, Philip L De, et al. "Alzheimer's Disease: Early Alterations in Brain DNA Methylation at ANK1, BIN1, RHBDF2 and Other Loci." *Nature News*, Nature Publishing Group, 17 Aug. 2014, www.nature.com/articles/nn.3786.

[6] Bhutani, Nidhi, et al. "DNA Demethylation Dynamics." *Cell*, vol. 146, no. 6, 16 Sept. 2011, pp. 866–872., doi:10.1016/j.cell.2011.08.042.

[7] Bennett, David A et al. "Overview and findings from the religious orders study" *Current Alzheimer research* vol. 9,6 (2012): 628-45.

[8] Bock C, Walter J, Paulsen M, Lengauer T (2007) CpG Island Mapping by Epigenome Prediction. PLOS Computational Biology 3(6): e110. https://doi.org/10.1371/journal.pcbi.0030110

[9] "GEO Accession Viewer." *Current Neurology and Neuroscience Reports.*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40699.

[10] EMBL-EBI. "Clustal Omega." *The European Bioinformatics Institute < EMBL-EBI*, www.ebi.ac.uk/Tools/msa/clustalo/.

[11] Daniel R Zerbino, et al.; Ensembl 2018, *Nucleic Acids Research*, Volume 46, Issue D1, 4 January 2018, Pages D754–D761, https://doi.org/10.1093/nar/gkx1098.

[12] Yokoyama, Amy S et al. "DNA methylation alterations in Alzheimer's disease"*Environmental epigenetics* vol. 3,2 dvx008. 6 Jun. 2017, doi:10.1093/eep/dvx008.

[13] Home - Gene - NCBI." *Current Neurology and Neuroscience Reports.*, U.S. National Library of Medicine, www.ncbi.nlm.nih.gov/gene.

[14] Bionetworks, info@sagebase.org Sage. "Synapse | Sage Bionetworks." *Synapse | Sage Bionetworks*, www.synapse.org/#!Synapse:syn3219045. ROSMAP Data.

Supplemental Material

Supplemental Table 1:

| Chromosome Number	AD Islands	Control Islands	Overlapping Islands	AD Only Islands	Control Only Islands
1	6139	6127	6113	26	14
2	3297	3293	3289	8	4
3	2155	2154	2150	5	4
4	1940	1945	1933	8	12
5	513	511	511	2	0
6	2914	2903	2896	19	7
7	3617	3604	3591	27	13
8	2267	2257	2253	15	4
9	1674	1664	1662	12	2
10	2769	2759	2755	14	4
11	3543	3534	3521	23	13
12	2589	2586	2576	13	10
13	1155	1149	1149	6	0
14	1640	1638	1633	8	5
15	1627	1628	1623	5	5
16	3976	3970	3962	14	8
17	4515	4504	4487	29	17
18	491	491	489	3	2
19	4657	4650	4636	22	14
20	1359	1358	1358	1	1
21	700	699	699	1	0
22	1444	1440	1439	6	1

Supplemental Figure 1:

A set of 4 AD motif logos from Gibbs sampling. There were 16 iterations of Gibbs sampling done.
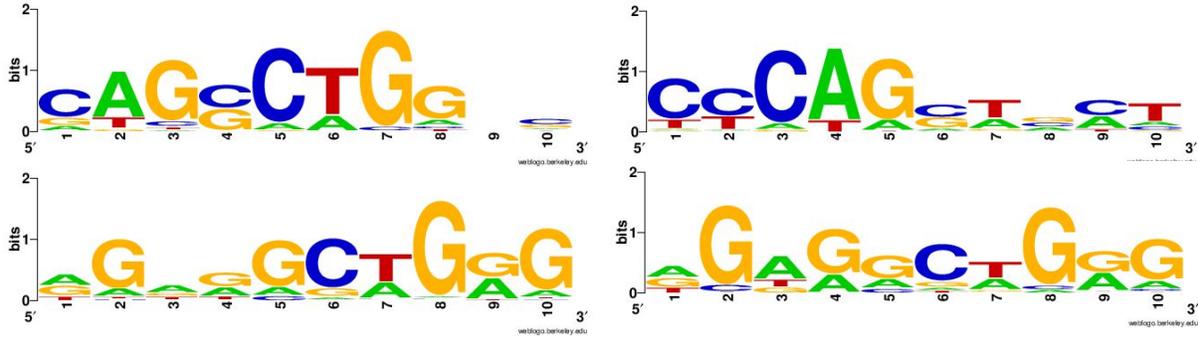
Supplemental Figure 2:

A set of 4 control motif logos from Gibbs sampling. There were 16 iterations of Gibbs sampling done.
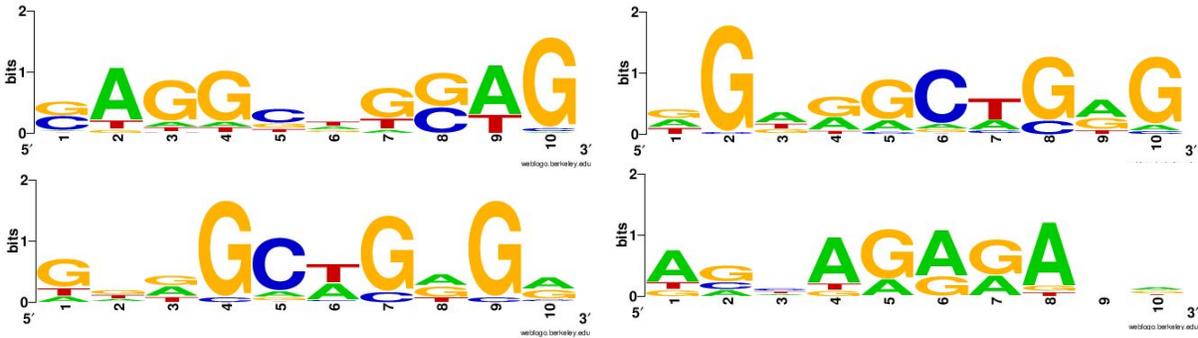
Supplemental Figure 3:
Left: Logo of Clustal Omega alignment of AD methylated CpG Islands
Right: Logo of Clustal Omega alignment of Control methylated CpG Islands

Supplemental Table 2: Top 10 Functions

Function	Number of Genes	Genes
protein binding	24	['ENSG00000162105', 'ENSG00000095370', 'ENSG00000119185', 'ENSG00000109906', 'ENSG00000006194', 'ENSG00000197081', 'ENSG00000172613', 'ENSG00000144476', 'ENSG00000234127', 'ENSG00000167840', 'ENSG00000153944', 'ENSG00000129204', 'ENSG00000184697', 'ENSG00000064999', 'ENSG00000171603', 'ENSG00000171606', 'ENSG00000132704', 'ENSG00000164828', 'ENSG00000069702', 'ENSG00000177045', 'ENSG00000166736', 'ENSG00000173540', 'ENSG00000168310', 'ENSG00000169515']
cytoplasm	17	['ENSG00000141519', 'ENSG00000095370', 'ENSG00000119185', 'ENSG00000185559', 'ENSG00000177045', 'ENSG00000164828', 'ENSG00000153944', 'ENSG00000172613', 'ENSG00000006283', 'ENSG00000173540', 'ENSG00000129204', 'ENSG00000164068', 'ENSG00000166736', 'ENSG00000069702', 'ENSG00000171606', 'ENSG00000165495', 'ENSG00000169515']
plasma membrane	14	['ENSG00000162105', 'ENSG00000119185', 'ENSG00000109906', 'ENSG00000132704', 'ENSG00000166736', 'ENSG00000197081', 'ENSG00000006283', 'ENSG00000144476', 'ENSG00000129204', 'ENSG00000184697', 'ENSG00000074660', 'ENSG00000127472', 'ENSG00000169515', 'ENSG00000159842']
nucleus	14	['ENSG00000172613', 'ENSG00000119185', 'ENSG00000109906', 'ENSG00000006194', 'ENSG00000189319', 'ENSG00000153944', 'ENSG00000144476', 'ENSG00000234127', 'ENSG00000132530', 'ENSG00000177045', 'ENSG00000171603', 'ENSG00000167840', 'ENSG00000165495', 'ENSG00000169515']
cytosol	13	['ENSG00000162105', 'ENSG00000119185', 'ENSG00000109906', 'ENSG00000159842', 'ENSG00000169515', 'ENSG00000234127', 'ENSG00000161040', 'ENSG00000066735', 'ENSG00000064999', 'ENSG00000132530', 'ENSG00000167840', 'ENSG00000168310', 'ENSG00000164068']
integral component of membrane	8	['ENSG00000185559', 'ENSG00000132704', 'ENSG00000176020', 'ENSG00000109066', 'ENSG00000184697', 'ENSG00000074660', 'ENSG00000090554', 'ENSG00000144476']
DNA-binding transcription factor activit...	7	['ENSG00000006194', 'ENSG00000109906', 'ENSG00000177045', 'ENSG00000168310', 'ENSG00000167840', 'ENSG00000171606', 'ENSG00000165495']
cell surface	7	['ENSG00000132704', 'ENSG00000127472', 'ENSG00000197081', 'ENSG00000144476', 'ENSG00000171603', 'ENSG00000090554', 'ENSG00000069702']
metal ion	6	['ENSG00000109906', 'ENSG00000006194', 'ENSG00000234127',

binding		'ENSG00000164068', 'ENSG00000167840', 'ENSG00000171606']	
nucleoplasm	6	['ENSG00000172613', 'ENSG00000119185', 'ENSG00000085760', 'ENSG00000168310', 'ENSG00000064999', 'ENSG00000165495']	

Supplemental Table 3:

Script	Description	Input	Output
analyze_alzheimers_data.py	Gets the methylated CpG islands from the experimental data	Experimental data-probes, list of methylated CpGs	List of methylated CpG islands for AD and for Normal
compare_alz_norm.py	Produces AD, Normal, and overlapping methylated CpG islands	List of methylated CpG islands for AD and for Normal	List of methylated CpG islands that are unique to AD, unique to Normal, and in both
gather_all_sequences.py	Convert a list of start and stop locations for CpG islands to their base pair sequence	List of starts and stops for CpG islands	Fasta file of sequences for both AD-specific and control-specific CpG islands
gibbs.py	Run Gibbs sampling on a list of sequences (CpG islands)	List of sequences Number of iterations	List of most common motifs and their PWMs
locate_motifs.py	Find instances of most common motifs in the AD-specific CpG islands	CSV file of CpG island data and list of motifs	Dictionary of motif sequence mapped to instances in genome that are AD specific, methylated CpG islands
download_gene_data.py	Download gene names from Ensembl	Dictionary of motifs to instances in genome	Dictionary of genes mapped to motif sequences on those genes
download_gene_function.py	Download gene functions from NCBI	Dictionary of genes mapped to motif sequences on those genes	CSV file of gene function and genes that have those functions
motif_signal_noise.py	Check the amount a list of motifs appears in each class of methylated CpG island	Lists of sequences for each class of CpG island List of motifs	Counts and percentages of each motif in each class of CpG island
motifs_in_genome.py	Allow the motif data to be read by STAMP	List of AD motifs from Gibbs sampling	File that can be exported to STAMP