STATE UNIVERSITY OF NEW YORK AT OSWEGO

MASTERS THESIS PROPOSAL

Biomedical and Health Informatics

DNA Methylation for the Diagnosis and Characterization of Four Psychiatric Disorders

---

Author:
Christopher Bartlett

Supervisor:
Isabelle Bichindaritz

Committee:

Abstract

Psychiatric disorder diagnoses are heavily reliant on the Diagnostic and Statistical Manual of Mental Disorders' listing of observable symptoms and clinical traits, the skill level of the physician, and the patient's ability to verbalize experienced events. Therefore, a body of researchers have sought to identify physical biomarkers which accurately differentiate mental disorder subtypes from psychiatrically normal controls. One such biomarker, DNA methylation, has recently become more prevalent in genetic research studies, with one particular study finding that DNA methylation analysis can predict cancer versus normal tissue with more than 95% accuracy. This paper proposes to apply these findings in a study of the diagnostic accuracy of DNA methylation signatures for classifying schizophrenia, bipolar disorder, posttraumatic stress disorder and major depressive disorder. Additionally, we intend to investigate which genes and biological pathways are associated with each of these four disorders.

Contents

Introduction

Epigenetics, the study of genetic alterations that do not affect the sequence of DNA, is a set of environmentally influenced processes that can switch genes on or off (Marlow, 2010). The epigenetic process that's been studied the most, DNA methylation, is the addition of a methyl group to cytosine or adenine and it has been associated with development (Ladd-Acosta et al., 2007), aging (Jung and Pfeifer, 2015), and the onset of cancers (Klutstein et al., 2016). As Ptak and Petronis, 2010 explain, the epigenetic model of disease assumes that an epigenetic disruption occurs during maturation of the germline which increases the organism's risk for disease. While small disruptions may not cause the disease, they say, these minute changes compound on one another until a threshold is crossed. They continue by saying that "relapse" and "remission" can be defined as fluctuating epimutation severity, or epimutations regressing back to their normal state (Ptak and Petronis, 2010).

Another key area of study in epigenetics is the role that DNA methylation has in individuals who have been diagnosed with a psychiatric disorder. Recent studies have investigated how individuals clinically diagnosed with schizophrenia (Liu et al., 2014; Wockner et al., 2014), bipolar disorder (Strauss et al., 2013), posttraumatic stress disorder (Kuan et al., 2017), and major depressive disorder (Davies et al., 2014), differ from psychiatrically normal individuals. These studies typically examine differential methylation in areas where the cytosine nucleotide is followed by a guanine nucleotide, known as a CpG site. Recently, differentially methylated CpG sites were used to differentiate tumor samples from four common cancers (breast, tumor, liver and lung) with that of normal tissue. Hao et al., 2017 used whole-genome methylation data from The Cancer Genome Atlas (TCGA) to construct a training cohort of 1,619 tumor samples and 173 matched adjacent normal tissue samples, and a validation cohort of 791 tumor samples and 93 matched adjacent normal tissue samples. The correct diagnosis rate for their training set was 98.4%, which was then replicated in the validation cohort for a statistically similar rate of 97.1%. A third, independent cohort of Chinese cancer samples (394 tumor samples and 324 matched adjacent normal tissue samples) resulted in a correct diagnosis rate of 95.0%. Methylation patterns were also able to correctly identify 29 of 30 colorectal cancer metastases to liver and 32 of 34 colorectal cancer metastases to lung (Hao et al., 2017). These findings suggest a potential usage of DNA methylation profiles for the diagnosis of primary and metastatic cancers, and it would be interesting to assess its replicability for psychiatric disorders.

Technical Proposal

2.1 Significance

The current method of diagnosing a psychiatric disease relies on the Diagnostic and Statistical Manual of Mental Disorders' predefined lists of symptoms and observed clinical traits. As stated by Demkow & Wolańczyk (2017), the patient's ability to consistently verbalize their experiences coupled with varying degrees of perceptive awareness in the health professional inflate complications in proper diagnosis. This sentiment is echoed in the mission of the National Institute of Mental Health's (NIMH) Research Domain Criteria (RDoC) initiative. In commentary for the initiative, Insel (2014) suggests that "While we can improve psychiatric diagnostics by more precise clustering of symptoms, diagnosis based only on symptoms may never yield the kind of specificity that we have begun to expect in the rest of medicine." While diagnostic rates are not expected to be as high in psychiatric disorders as they are in the cancers listed above, as disorder versus normal is not as black-and-white as cancer versus normal, the significance of this endeavor is in determining the diagnostic accuracy of DNA methylation. Further, a subset of accurately identified individuals can be compared to misidentified individuals to further determine their degrees of separation.

2.2 Aims

Four goals comprise this project; a primary, secondary, tertiary, and quaternary. They are listed as follows:

Primary: The primary aim of this project is to investigate whether one can correctly classify four mental disorder from DNA methylation profiles. These four mental disorders are: schizophrenia (SZ), bipolar disorder (BD), posttraumatic stress disorder (PTSD) and major depressive disorder (MDD). Specifically, a classification system will be constructed that uses DNA methylation levels found in whole blood and post-mortem brain samples to attempt to accurately classify which subjects have been clinically diagnosed with a disorder, which disorder they've been diagnosed with, or if they belong to a psychiatrically-normal control group. This goal includes comparing accurately identified individuals to misidentified individuals to determine which factors provide correct classification.

Secondary: Determine a methylation signature associated with each mental disorder and whether feature selection improves on classification performance.

Tertiary: Using the CpG site coordinates and a subset of differentially methylated regions, construct a list of associated genes for each of the four disorders and investigate the classification effectiveness. This will most likely be handled through a Bioconductor R package.

Quaternary: Map these associated genes to their associated pathways using pathway analysis software and examine the classification effectiveness for the four mental disorders.

2.3 Approach

The following steps will need to be conducted in the pursuit of the goals listed above:

1. Acquisition of pre-processed Ilumina HumanMethylation450 BeadChip datasets for the four disorders and their matched controls.
2. Merging of the datasets into one composite dataset for ease of comparison and analysis.
3. Distinguishing a global methylation score amongst all four psychiatric disorders, and a global methylation score among the control group.
4. Determining the accuracy for assigning any given subject to one of these two groups.
5. Distinguishing a global methylation score for each psychiatric disorder, and with the score of the control group.
6. Determining the accuracy of assigning any given subject to one of these five groups.
7. Isolating the differentially methylated regions between disorder and control for each of the four disorders.
8. Using the subsets from step 7, determine the new accuracy of assigning any given subject to one of these five groups.
9. Determining the genes associated with the differentially methylated regions.
10. Using the associated genes, determine the new accuracy.
11. Identifying biological pathways that characterize the four disorders.
12. From a listing of the biological pathways, determine the new accuracy.

2.3.1 Acquisition

The acquisition of relevant data has already begun, with information being gleaned from ArrayExpress, the Psychiatric Disorder Next-Generation Sequencing Atlas (PD_NGSAtlas), and the Psychiatric Disorder specific Methylation database (PDMeth). Table 2.1 displays current sample totals for each of the disorders as well as the number of psychiatrically normal control samples in datasets investigating the disorder. These samples used the Ilumina HumanMethylation450 (HM450) BeadChip array. The purpose behind using the HM450 array is to control for the tool used in data collection. As different institutions will have different protocols, different machines and different people running the equipment however, some data cleaning is expected. An important note is that 129 individuals with whole blood samples in Table 2.1 were treated for depression, which does not necessarily indicate major depressive disorder. Additionally, 33 MDD blood samples in Table 2.1 were dexamethasone stimulated while 46 MDD control blood samples were dexamethasone stimulated. It is presently unknown whether these samples can, and will be, included. Table 2.2 displays current sample totals for each of the disorders and their controls which have been sequenced using Methylated DNA

Class	Sample Location	Number of Disorder Samples	Number of Control Samples
Schizophrenia	Blood	786	736
	Brain	42	39
Major Depressive Disorder	Blood	195	322
	Brain	-	-
Bipolar Disorder	Blood	33	227
	Brain	-	-
Posttraumatic Stress Disorder	Blood	31	227
	Brain	-	-

TABLE 2.1: The number of disorder and matched control HM450 methylation samples obtained, whether these samples were derived from blood or brain tissue, and which disorder they correspond to.

MeDIP Sequencing

Class	Sample Location	Number of Disorder Samples	Number of Control Samples
Schizophrenia	Blood	12	2
	Brain	11	17
Major Depressive Disorder	Blood	-	-
	Brain	-	-
Bipolar Disorder	Blood	6	2
	Brain	12	17
Posttraumatic Stress Disorder	Blood	-	-
	Brain	-	-

TABLE 2.2: The listing of obtained samples which have been sequenced with Methylated DNA immunoprecipitation (MeDIP). It is currently anticipated that these samples will be used to construct a validation set.

immunoprecipitation (MeDIP). Samples in Table 2.2 are currently isolated for their potential usage as a validation set.

2.3.2 Software

RStudio and the R programming language is expected to perform most of the work to make the merging and analysis as streamlined as possible. However, Microsoft Excel, SPSS, SAS and SAS Enterprise Miner may also be used.

Related Work

Aside from the aforementioned study utilizing differentially methylated CpG sites to differentiate tumor samples from four common cancers (Hao et al., 2017), methylation-based classifiers have also been used to classify bone sarcomas (Cooper, Killian, and Pisapia, 2017), pediatric brain tumors (Danielsson et al., 2015), thyroid carcinomas (Reis et al., 2017), and prostate cancer (Mundbjerg et al., 2017). Cooper, Killian, and Pisapia, 2017 and Danielsson et al., 2015 developed a random forest classifier after isolating the most differentially methylated CpG sites (400 sites and 100 sites respectively) while Mundbjerg et al., 2017 used 25 probes, and hierarchical clustering with Euclidean distance. Reis et al., 2017 used 21 probes and unsupervised hierarchical clustering. Sample size and accuracy varied with Cooper, Killian, and Pisapia, 2017 classifying 10 of 10 synovial sarcomas, 85 of 86 osteosarcomas, and 14 of 15 Ewing sarcoma samples. Danielsson et al., 2015 achieved high accuracy (k = 0.98) for 28 pediatric tumors. Mundbjerg et al., 2017 tested 496 prostate samples (tumor and adjacent-normal) and received 97.4% specificity and 96.2% cancer sensitivity while Reis et al., 2017 had 63% sensitivity and 92% specificity for 141 thyroid samples. Of note, each of these studies used the Ilumina HumanMethylation450 BeadChip array.

One issue when comparing these results to a diagnostic classifier for the four investigated mental disorders is that these studies use tissue samples straight from the source of the disease while DNA methylation datasets for psychiatric diagnoses are typically whole blood samples with a smaller portion being derived from post-mortem brain tissue. Discrepancy exists regarding the predictability of brain methylation levels through whole blood samples. Walton et al., 2016 obtained temporal lobe biopsy samples from 12 epilepsy patients and compared them to paired blood samples and a gene set enrichment analyses of peripheral blood DNA methylation data from 111 schizophrenia patients and 122 controls and found that of the 227,428 variable, brain-associated CpG sites investigated, only 7.9% had a significant correlation between blood and brain tissue. Hannon et al., 2015 similarly found that interindividual variation in whole blood was not a strong predictor of interindividual variation in the brain when quantifying DNA methylation in matched whole blood and brain samples from 122 individuals. In contrast, however, Davies et al., 2014 found that canonical DNA methylation profiles do not differ across tissue or sample types and, alongside Wockner et al., 2014, found that methylation rates in leukocytes are correlated with rates in the brain. Similarly, Tylee, Kawaguchi, and Glatt, 2013 reviewed epigenomic literature which suggests that CpG-island methylation levels are highly correlated between blood and brain. They state that one can confidently assume that genetic sequences detected in peripheral blood samples will be identical to those found in the brain, with somatic mutations and other causes of mosaicism being the exceptions (Tylee, Kawaguchi, and Glatt, 2013). These disparate findings urge some degree of caution when using blood and brain tissue samples, though we intend to strive for clinical diagnostic validity with blood samples and will fall back on scientific relevance if post-mortem brain samples are found to be significantly more descriptive in the classification of a mental disorder.

Deliverables

The following deliverables are anticipated:

1. A conclusory paper that outlines the research methods, important findings, scientific and clinical relevance stemming from the project.
2. The R packages that were used for merging and analysis, as well as any modifications or newly-scripted code.
3. Source information for the utilized datasets.
4. The final, merged dataset in either comma-separated or tab-delimited format (or both).
5. An oral presentation along with any materials that aided the presentation.

Educational Statement

This project will most heavily draw upon experiences from a graduate assistantship position in which clinical, gene expression, copy number variation, and somatic mutation data was compiled to determine the genetic markers of breast cancer. This position provided experience in investigating the genetic components of a disease, as well as using R packages to analyze data on these components. Additionally, data preprocessing, classification and clustering techniques learned in the Data Analytics course will be useful, while the Big Data, Genes and Medicine course instructed on how to use these techniques for genetic analysis.

Criteria for Success

Successful completion of the project will be contingent on accomplishing the goals outlined above, strict adherence to the guidelines for accessing the data within each of the datasets, and utilizing knowledge from the Biomedical and Health Informatics curriculum to fulfill a research project.

Timeline

Semester	Month	Associated Steps	Task
Fall 2017	November	Step 1	Acquisition of relevant datasets.
	December	Step 2.	Merging data into one composite dataset.
Spring 2018	February	Steps 3 through 6	Determining classification accuracy for normal versus disorder and normal versus SZ, BD, MDD or PTSD.
	March	Steps 7 through 8	Determining differentially methylated regions for each disorder and re-calculating classification accuracy.
	April	Steps 9 through 11	Determining associated genes, re-classifying, performing pathway analysis and re-calculating classification accuracy.
	May		Concluding write-up and presenting results.

Bibliography

Cooper, Benjamin T, J Keith Killian, and David J Pisapia (2017). "DNA Methylation – Based Classifier for Accurate Molecular Diagnosis of Bone Sarcomas original report". In: pp. 1–11.

Danielsson, Anna et al. (2015). "MethPed: a DNA methylation classifier tool for the identification of pediatric brain tumor subtypes". In: *Clinical Epigenetics* 7.1, p. 62. ISSN: 1868-7075. DOI: 10.1186/s13148-015-0103-3. URL: http://www.clinicalepigeneticsjournal.com/content/7/1/62.

Davies, Matthew N et al. (2014). "Hypermethylation in the ZBTB20 gene is associated with major depressive disorder". In: *Genome Biology* 15.4, R56. ISSN: 1465-6906. DOI: 10.1186/gb-2014-15-4-r56. URL: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-4-r56.

Hannon, Eilis et al. (2015). "Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes". In: *Epigenetics* 10.11, pp. 1024–1032. ISSN: 15592308. DOI: 10.1080/15592294.2015.1100786.

Hao, Xiaoke et al. (2017). "DNA methylation markers for diagnosis and prognosis of common cancers". In: *PNAS* 114.28, pp. 7414–7419. DOI:

10.1073/pnas.1703577114/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1703577114.

Jung, Marc and Gerd P Pfeifer (2015). "Aging and DNA methylation". In: *BMC Biology* 13.1, p. 7. ISSN: 1741-7007. DOI: 10.1186/s12915-015-0118-4. URL: http://www.biomedcentral.com/1741-7007/13/7.

Klutstein, Michael et al. (2016). "DNA methylation in cancer and aging". In: *Cancer Research* 76.12, pp. 3446–3450. ISSN: 15387445. DOI: 10.1158/0008-5472.CAN-15-3278.

Kuan, P-F et al. (2017). "An epigenome-wide DNA methylation study of PTSD and depression in World Trade Center responders". In: *Translational Psychiatry* 7.6, e1158. ISSN: 2158-3188. DOI: 10.1038/tp.2017.130. URL: http://www.nature.com/doifinder/10.1038/tp.2017.130.

Ladd-Acosta, Christine et al. (2007). "DNA Methylation Signatures within the Human Brain". In: *The American Journal of Human Genetics* 81.6, pp. 1304–1315. ISSN: 00029297. DOI: 10.1086/524110. URL: http://linkinghub.elsevier.com/retrieve/pii/S0002929707637793.

Liu, Jingyu et al. (2014). "Methylation patterns in whole blood correlate with symptoms in Schizophrenia patients". In: *Schizophrenia Bulletin* 40.4, pp. 769–776. ISSN: 17451701. DOI: 10.1093/schbul/sbt080.

Mundbjerg, Kamilla et al. (2017). "Identifying aggressive prostate cancer foci using a DNA methylation classifier". In: *Genome Biology* 18.1, p. 3. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1129-3. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1129-3.

Ptak, Carolyn and Arturas Petronis (2010). "Epigenetic approaches to psychiatric disorders." In: *Dialogues in clinical neuroscience* 12.1, pp. 25–35. ISSN: 12948322.

Reis, Mariana Bisarro dos et al. (2017). "Prognostic classifier based on genome-wide DNA methylation profiling in well-differentiated thyroid tumors". In: *The Journal of Clinical Endocrinology & Metabolism* 102.November, pp. 4089–4099. ISSN: 0021-972X. DOI: 10.1210/jc.2017-00881. URL: http://academic.oup.com/jcem/article/doi/10.1210/jc.2017-00881/4082866/Prognostic-classifier-based-on-genomewide-DNA.

Strauss, John S et al. (2013). "Quantitative leukocyte BDNF promoter methylation analysis in bipolar disorder." In: *International journal of bipolar disorders* 1, p. 28. ISSN: 2194-7511. DOI: 10.1186/2194-7511-1-28. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4215812{\&}tool=pmcentrez{\&}rendertype=abstract.

Tylee, Daniel S., Daniel M. Kawaguchi, and Stephen J. Glatt (2013). "On the outside, looking in: A review and evaluation of the comparability of blood and brain "-omes"". In: *American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics* 162.7, pp. 595–603. ISSN: 15524841. DOI: 10.1002/ajmg.b.32150.

Walton, Esther et al. (2016). "Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research". In: *Schizophrenia Bulletin* 42.2, pp. 406–414. ISSN: 17451701. DOI: 10.1093/schbul/sbv074.

Wockner, L F et al. (2014). "Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients". In: *Translational Psychiatry* 4.1, e339. ISSN: 2158-3188. DOI: 10.1038/tp.2013.111. URL: http://www.nature.com/doifinder/10.1038/tp.2013.111.