Multi-Omic Biomarker Identification and Characterization for Posttraumatic Stress Disorder

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029495">http://nrs.harvard.edu/urn-3:HUL.InstRepos:42029495</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Multi-omic biomarker identification and characterization for Posttraumatic Stress Disorder

A DISSERTATION PRESENTED
BY
KELSEY R. DEAN
TO
THE DEPARTMENT OF SYSTEMS BIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
SYSTEMS BIOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MARCH 2019
Multi-omic biomarker identification and characterization for Posttraumatic Stress Disorder

Abstract

Posttraumatic Stress Disorder (PTSD) is an anxiety disorder affecting 5-15% of the population that develops after exposure to a traumatic event. Diagnosis and treatment of PTSD is limited by the lack of understanding of the disease mechanism, inability to identify early cases of disease development, failure to determine prognosis, and poor determination of successful personalized treatment. Integrating multi-scale molecular data from blood can provide new insights into the disorder, and generate hypotheses regarding candidate diagnostic biomarkers, novel therapeutic targets, and disease subgroups.

Utilizing multiple cohorts of male veterans with and without PTSD, we integrated molecular datasets including genetics, epigenetics, proteomics, metabolomics, and others to address some of these core challenges in PTSD diagnosis, prognosis, and treatment. Primarily, our efforts were focused in identification of blood-based biomarkers for diagnosing PTSD, and characterization of disease biology. Using an integrated methodology, we consolidated candidate biomarkers across all molecular data types, and performed two stages of biomarker panel refinement to identify a final candidate biomarker panel for diagnosing PTSD. We evaluated the performance of this final set of 28 biomarkers in an independent validation cohort, and achieved accuracy of 81%, sensitivity of 85% and specificity of 77% in diagnosing warzone-related PTSD in a cohort of combat-exposed male veterans. This biomarker panel consisted of a heterogeneous set of molecular data types, including DNA methylation, miRNAs, proteins, metabolites, and small molecules. These features include some with direct links to known PTSD biology, including methylation of the PDE9A and CPT1B genes, which have previously been shown to have altered expression in PTSD, as well as novel biology, including miRNAs associated with obesity, diabetes, and inflammation.
Additionally, we have developed and implemented multiple novel methodologies for: (1) characterizing diseases from high-throughput datasets, (2) integrating incomplete multi-omic data, and (3) identifying biomarkers in clinical settings where false positive and false negative errors have unequal weights. We quantified changes in expression variance from high throughput datasets, and incorporated the differential expression variance genes into a network-based biomarker identification framework. These network-based biomarkers from gene expression variance resulted in larger disease-related subnetworks that can be used for classification. Next, we proposed a maximum probability integration strategy that allows for integration of incomplete multi-omic datasets. Using individual data type classifiers, we incorporated additional incomplete samples in the training dataset, and evaluated performance based on maximum probability in the test set, resulting in improvements in AUC. Finally, we used expression variance to generate a feature selection approach that improves either sensitivity or specificity of the prediction, without loss of accuracy.

Overall, we have applied multiple novel approaches to identifying biomarkers and characterizing disease signals from multi-omic PTSD datasets. These algorithms have generated candidate biomarkers for further evaluation, identified disease signals for future studies, and provided tools for analyses of new datasets, in PTSD or any disease of interest.
# Contents

1 **Introduction** ......................................................... 1
   1.1 Posttraumatic Stress Disorder ................................. 2
   1.2 Systems Biology and Multi-Omic datasets ...................... 2
   1.3 Machine learning and biomarker identification for understanding disease .................. 5
   1.4 Biomarkers for PTSD ........................................... 6
   1.5 PTSD Heterogeneity ............................................ 10
   1.6 Thesis Overview ............................................... 11

2 **Multi-omic Biomarker Identification for Diagnosing Posttraumatic Stress Disorder** 12
   2.1 Background .................................................. 12
   2.2 Methods ..................................................... 14
   2.3 Results ..................................................... 17
   2.4 Discussion .................................................. 28

3 **Disease Subnetwork Characterization through Mean, Variance, and Distribution Changes in Gene Expression and DNA Methylation** 32
   3.1 Background .................................................. 32
   3.2 Results ..................................................... 34
   3.3 Discussion .................................................. 47

4 **Advancing biomarker identification approaches: integrating incomplete multi-omic data and optimizing sensitivity vs specificity tradeoff** 49
   4.1 Incorporating incomplete datasets ............................ 50
   4.2 A biomarker identification approach for optimizing sensitivity or specificity ........... 55
   4.3 Discussion .................................................. 59

5 **Conclusions and Future Directions** ................................. 61
   5.1 Background .................................................. 61
   5.2 Preliminary Results ........................................... 63
   5.3 Incorporating PTSD subgroup-specific biomarkers .................. 66
   5.4 Understanding disease progression .............................. 68

Appendix A **PTSD diagnostic criteria** ................................. 71

References ............................................................ 87
## Listing of figures

1.1 Integration of multi-omic, multi-scale datasets ........................................ 5
2.1 Overview of PTSD biomarker identification approach ................................. 18
2.2 Overview of molecular datasets and cohort symptom severity ..................... 19
2.3 Two stage biomarker down-selection approach ........................................... 21
2.4 Validation of biomarker panels ................................................................. 23
2.5 Summary of potential confounders from PTSD biomarker validation ........... 27
2.6 Predicted probability of PTSD in MDD+ and MDD- populations .................. 28
3.1 Summary of breast cancer subnetworks ..................................................... 38
3.2 Summary of brain cancer subnetworks ....................................................... 39
3.3 Summary of PTSD gene expression subnetworks ....................................... 40
3.4 Summary of PTSD DNA methylation subnetworks .................................... 41
3.5 Largest identified breast cancer subnetworks from each metric ................. 43
3.6 Largest identified brain cancer subnetworks from each metric .................... 44
3.7 Breast cancer gene related subnetworks .................................................... 45
3.8 Largest identified PTSD subnetworks from each metric ............................. 46
3.9 Largest identified PTSD methylation subnetworks from each metric ........... 47
4.1 Classification performance vs training sample size in four publicly available cancer datasets .................................................................................. 50
4.2 Classification performance vs training sample size in PTSD datasets ........... 51
4.3 Overview of maximum probability approach .............................................. 53
4.4 Summary of PTSD multi-omic classification results ................................... 54
4.5 Overview of four disease gene expression distributions .............................. 57
4.6 Example ROC curve, with candidate operating points ............................... 58
4.7 Sensitivity and specificity optimization with PTSD and public cancer datasets .. 59
5.1 Proportion of Deviance (POD) for clustering PTSD subjects using iCluster .... 65
5.2 Proportion of Deviance (POD) for clustering control subjects using iCluster ... 65
5.3 Individual data type principal component plots ......................................... 67
5.4 PTSD progression network structure inferred from DNA methylation patterns in 83 PTSD+ subjects ................................................................. 69
## Listing of tables

1.1 PTSD Systems Biology Consortium members .................................................. 3  
1.2 Summary of molecular, physiological, and clinical datasets for three PTSD cohorts . 4  
1.3 Summary of neuroimaging findings in subjects with PTSD ............................. 10  

2.1 Summary of blood tubes and samples for molecular assays ............................. 16  
2.2 Summary of biomarker identification approaches ........................................... 17  
2.3 Summary of candidate biomarkers ................................................................. 18  
2.4 PTSD biomarker confusion matrix ................................................................. 23  
2.5 Predictive performance of nonlinear biomarkers ............................................ 25  
2.6 Summary of cohort demographics and clinical symptoms ............................... 26  
2.7 Overview of biomarker signals in each of the three cohorts ............................ 29  

3.1 Summary of datasets ....................................................................................... 37  
3.2 Average clustering coefficient, network density, and centralization for subnetworks . 42  
3.3 Average classification performance from gradient boosting with decision trees ... 42  

4.1 Summary of available clinical and molecular data for PTSD cohort ............... 52  
4.2 Size of public multi-omic datasets ................................................................. 55  
4.3 Summary of public multi-omic classification performance ............................. 55  

5.1 Summary of PTSD data included in subgroup identification ......................... 64
To my incredibly supportive family—my parents, grandparents, sisters, and fiancé. I could not have done this without you.
Acknowledgments

First, I would like thank my advisor, Frank Doyle, for his support throughout my PhD. His patience, guidance, and encouragement was instrumental to me during my PhD, and allowed me the opportunity to develop research, leadership, and communication skills that will be invaluable to me in the future.

Additionally, I would like to thank the faculty that served on my committees, both at UCSB and Harvard, including Dr. Linda Petzold, Dr. Michelle O’Malley, Dr. Sean Eddy, Dr. Kerry Ressler, Dr. Doug Lauffenburger, and Dr. Rasha Hammamieh. Their guidance and thoughtful feedback shaped my work over the years, and provided diverse perspectives that I may not have otherwise considered.

I would also like to thank my colleagues and mentors in the Doyle group. Over the years I was fortunate to learn from and collaborate with many talented researchers, including Dr. John Abel, Dr. Gunjan Thakur, Dr. Burook Misganaw, Dr. Pramod Rajaram, and many others.

Finally, this work would not have been possible without the many individuals involved in the PTSD Systems Biology consortium, including Dr. Charles Marmar and his team (Duna Abu-Amara, Dr. Meng Qian, Dr. Eugene Laska, Nicholas Milton, and others), Dr. Marti Jett and her team (Dr. Ruoting Yang, Dr. Duncan Donohue, Uma Mudunuri, and others), Dr. Owen Wolkowitz, Dr. Sindy Mellon and their team (Michelle Coy, Gwyneth Wu, and others), Dr. Rachel Yehuda and her team (Dr. Janine Flory, Alex Ropes, and others), Dr. Leroy Hood and his team (Dr. Kai Wang, Dr. Inyoul Lee, Tae-Kyu Kim, Dr. Min Young Lee, and others), Dr. Kerry Ressler and his team (Dr. Adriana Lori, Dr. Guia Guffanti, and others), Dr. Bernie Daigle Jr., Dr. Ron Hoover, and many others.

Lastly, I would like to acknowledge generous funding from the U.S. Army Research Office, though grant numbers W911NF-10-2-0111, W911NF-16-2-0015, and W911NF-17-2-0086.
Introduction*

In the past decade, high throughput technologies have enabled large-scale studies of disease by measuring thousands of unique data points per individual, from a small amount of biological samples (e.g. blood, urine, saliva, or biopsy samples). Integrated, multi-omic approaches to cancer and other physical health conditions have been able to identify diagnostic and prognostic markers, signals of disease heterogeneity and subtypes, and provide insights disease mechanism by combining genetics, epigenetics, transcriptomics, and other molecular and clinical datasets. To generate these multi-omic datasets, collaborative teams such as The Cancer Genome Atlas (TCGA) have been established. These collaborative groups consisting of biologists, clinicians, statisticians, engineers, and computer scientists have generated large molecular datasets for many diseases, leading to many novel findings related to disease diagnosis, disease mechanism, and disease heterogeneity.

While multi-omic approaches have had successes in cancer and many areas of physical health, biological ap-

CHAPTER 1. INTRODUCTION

... approaches to understanding mental health conditions continues to lag behind, likely due to the history of treating mental health conditions differently than physical health conditions. In order to bring quantitative, systems biology approaches to the study of Post-Traumatic Stress Disorder (PTSD), the PTSD Systems Biology Consortium was established. This consortium sought to formally investigate biological signals of PTSD by generating a wealth of multi-omic data on veterans with and without PTSD. Coupled with detailed clinical phenotyping, as well as extensive demographic, medical history, and neurocognitive function data, this data provided the ability to look for: (1) blood-based biomarkers for diagnosing PTSD, (2) evidence of relevant subtypes of PTSD, (3) therapeutic targets for treatment of PTSD, and (4) molecular signals giving insight into disease risk, development, and progression.

1.1 Posttraumatic Stress Disorder

PTSD is a psychiatric disorder that can develop following exposure to a traumatic event. PTSD has estimated lifetime and current prevalence rates ranging from 13-30% in military populations and approximately 8% in civilian populations. Diagnostic criteria for PTSD are outlined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5). The DSM-5 includes eight required criteria and two additional specifications. The eight required criteria include: (1) exposure to traumatic event, (2) intrusion symptoms, (3) avoidance symptoms, (4) negative alterations in cognition and mood, (5) alterations in arousal and reactivity, (6) symptoms persisting for at least one month, (7) impairment in social, occupational, or other areas of functioning, and (8) absence of substance use or other medical conditions causing symptoms. Additionally, a PTSD diagnosis may be accompanied by two specifications: (1) with dissociative symptoms, including those with symptoms of depersonalization or derealization and (2) with delayed expression, for those not meeting full diagnostic criteria until at least six months following the trauma.

The complete set of detailed diagnostic criteria, as written in the DSM-5, are listed in Appendix A.

1.2 Systems Biology and Multi-Omic datasets

While clinical diagnosis may work well for clearly distinguishing moderate or severe cases of PTSD from asymptomatic controls, a robust and unbiased molecular approach may offer additional advantages. Identifying robust molecular markers may assist in many aspects of diagnosing, treating, preventing, and understanding the disease.
CHAPTER 1. INTRODUCTION

This may include:

- more accurate diagnosis, to overcome bias due to self-reported symptoms (including both under-reporting and over-reporting of symptoms),
- characterization of PTSD development and disease progression,
- prediction of susceptibility risk,
- personalization of treatment strategies,
- identification of therapeutic targets for PTSD treatment, and
- stratification of PTSD into molecularly and clinically relevant subtypes.

Table 1.1: PTSD Systems Biology Consortium members. The PTSD Systems Biology Consortium consists of seven collaborating sites, working on the same cohorts of PTSD and control participants to generate multi-omic

<table>
<thead>
<tr>
<th>Institution</th>
<th>Principal Investigator(s)/Site Leads</th>
<th>Core</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institute for Systems Biology</td>
<td>Leroy Hood</td>
<td>Multi-omics</td>
</tr>
<tr>
<td>New York University</td>
<td>Charles Marmar</td>
<td>Clinical &amp; Brain Imaging</td>
</tr>
<tr>
<td>UCSF</td>
<td>Synthia Mellon &amp; Owen Wolkowitz</td>
<td>Metabolism &amp; Cell Aging</td>
</tr>
<tr>
<td>USACEHR</td>
<td>Marti Jett &amp; Rasha Hammamieh</td>
<td>Multi-omics</td>
</tr>
<tr>
<td>Emory/McLean</td>
<td>Kerry Ressler</td>
<td>Genetics</td>
</tr>
<tr>
<td>Harvard</td>
<td>Frank Doyle</td>
<td>Bioinformatics</td>
</tr>
<tr>
<td>Mt Sinai</td>
<td>Rachel Yehuda &amp; Janine Flory</td>
<td>Neuroendocrine</td>
</tr>
</tbody>
</table>

To search for robust, molecular markers related to PTSD development, progression, stratification, and recovery, multi-scale molecular data was collected from human participants (Figure 1.1). Integration of this molecular, cellular, and physiological datasets may provide a comprehensive picture of disease mechanism, and allows biological signals to be tracked through each layer of data. Through the PTSD Systems Biology Consortium (Table 1.1), this multi-scale data was collected for multiple cohorts of combat-exposed veterans. The molecular data included genetics, DNA methylation, miRNA expression, proteomics, metabolomics, neuroendocrine markers, routine clinical lab markers, immune cell counts, small molecules, physiological measurements, structural and functional brain imaging, neurocognitive testing, as well as extensive clinical, demographic, and medical history information (Table 1.2).
<table>
<thead>
<tr>
<th>Data Type</th>
<th>Assay/Details</th>
<th>~# of features measured</th>
<th>N, Cohort 1</th>
<th>N, Cohort 2</th>
<th>N, Cohort 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics</td>
<td>Illumina PsychArray BeadChip</td>
<td>~300k</td>
<td>162</td>
<td>59</td>
<td>56</td>
</tr>
<tr>
<td>DNA Methylation</td>
<td>Illumina 450K BeadChip array</td>
<td>~450k</td>
<td>162</td>
<td>59</td>
<td>53</td>
</tr>
<tr>
<td>miRNA</td>
<td>small RNA sequencing</td>
<td>335</td>
<td>164</td>
<td>59</td>
<td>56</td>
</tr>
<tr>
<td>Proteomics</td>
<td>Selected Reaction Monitoring</td>
<td>96</td>
<td>162</td>
<td>59</td>
<td>56</td>
</tr>
<tr>
<td>Metabolites</td>
<td>UHPLC/MS/MS² and GS/MS</td>
<td>147</td>
<td>165</td>
<td>59</td>
<td>57</td>
</tr>
<tr>
<td>Small Molecules</td>
<td>Neurosteroids, oxidation byproducts</td>
<td>11</td>
<td>165</td>
<td>59</td>
<td>57</td>
</tr>
<tr>
<td>Clinical Labs</td>
<td>cholesterol, C reactive protein, glucose, etc.</td>
<td>44</td>
<td>165</td>
<td>58</td>
<td>56</td>
</tr>
<tr>
<td>Neuroendocrine</td>
<td>cortisol, ACTH, etc.</td>
<td>15</td>
<td>165</td>
<td>59</td>
<td>57</td>
</tr>
<tr>
<td>Physiological</td>
<td>heart rate, BP, BMI, etc.</td>
<td>9</td>
<td>165</td>
<td>58</td>
<td>57</td>
</tr>
<tr>
<td>Clinical Symptoms</td>
<td>CAPS, SCID, etc.</td>
<td>236</td>
<td>165</td>
<td>59</td>
<td>57</td>
</tr>
</tbody>
</table>
CHAPTER 1. INTRODUCTION

Figure 1.1: Integration of multi-omic, multi-scale datasets. Figure adapted from ⁷.

1.3 Machine learning and biomarker identification for understanding disease

Disease can cause molecular, cellular, and tissue changes within an organism, in many cases resulting in observable phenotypic changes. The intital cause of these molecular, cellular or tissue changes may be due to genetic causes, environmental factors, or a combination of the two. Biological markers of a disease, biomarkers, may be used to track disease progression, status, or risk. These markers may include molecular markers (e.g. DNA mutations, gene expression, small molecules), imaging markers, electrical activity markers, or other quantifiable measures. High throughput assays provide a means to search for candidate biomarkers, particularly in diseases with little available prior knowledge for hypothesis-driven approaches.
CHAPTER 1. INTRODUCTION

1.4 Biomarkers for PTSD

Previous work has identified many molecular and other signals related to PTSD. These candidate biomarkers may be used for diagnosis, prognosis, or stratification. In the following sections, we provide an overview of previously identified PTSD signals, including genetics, epigenetics, transcriptomics, proteomics, small molecules, neuroendocrine markers, and brain imaging.

1.4.1 Genetics

Using identical twin studies, PTSD heritability has been estimated to be anywhere between 20 and 70%\textsuperscript{8,9,10,11}. To identify the specific genetic risks factors responsible for this heritability, many candidate gene studies were performed to search in previously identified PTSD-related genes, including genes related to the hypothalamic pituitary adrenal (HPA) axis and those involved in limbic-frontal neuronal circuitry, as implicated by other psychiatric disorders\textsuperscript{12}. Four single nucleotide polymorphisms (SNPs) in FKBP5 and three SNPs in GABARA2 significantly interacted with childhood abuse or trauma to predict PTSD symptoms in adults\textsuperscript{13,14}. In addition to candidate gene studies, data-driven genome-wide association studies (GWAS) were conducted on large cohorts to identify new genes using a hypothesis-neutral approach. In both the PTSD Systems Biology Consortium military cohort, and the Grady Trauma Project civilian cohort, a intergenic SNP, rs717947, was associated with PTSD diagnosis\textsuperscript{15}. However, in the civilian replication cohort, the identified SNP was only significantly associated in the female participants. The significant SNP, rs717947, is located upstream from DTHD1, a protein-coding gene involved in signaling and apoptosis. Similarly, in a cohort of 503 females and 295 males, a SNP in the ADCYAP1R1 receptor (rs2267735) was significantly associated with PTSD diagnosis in females, but not in males\textsuperscript{16}, again suggesting gender differences in PTSD risk, and potentially disease mechanism. In a large study of over 20,000 male and female participants, no genome-wide significant polymorphisms were identified, but the use of polygenic risk scores was suggested, to quantify accumulated risk over many loci\textsuperscript{17}. Further work into gender-specific SNPs, polygenic risk scores, and connections to genetic risk, heritability and mechanisms of similar psychiatric conditions is ongoing.
1.4.2 Epigenetics

Epigenetics, in particular DNA methylation, contains information related to changes in gene expression that are not due to mutations in DNA sequence. DNA methylation can occur on cytosine bases in cytosine-guanine dinucleotides (known as CpGs). Changes in these methylation patterns, especially those in promoter regions, may alter regulation of gene expression, leading to biological dysfunction in a particular disease. DNA methylation has been shown to carry long-term signals of early life stress, and identical twins show increasing epigenetic variance with age. In PTSD, many candidate gene studies and high throughput, hypothesis-neutral studies have identified altered methylation patterns in a number of genes. In PTSD, identified hypomethylated genes include: FKBP5, SLC6A4, genes related to innate immunity and immune function (TLR1, TLR3, IL8, LTA, and KLRG-1), NR3C1, and LINE-1. Interestingly, decreased methylation of FKBP5 and SLC6A4 was only associated with PTSD in the presence of childhood abuse or greater traumatic exposure, respectively. This link indicates that some methylation signals may not be related to PTSD diagnosis overall, but may be related to PTSD risk due to earlier trauma, or to a subset of PTSD cases who have experienced a large number of traumas.

Contrastly, hypermethylated genes in PTSD include: SLC6A3, MAN2C1, CLEC9A, ACP5, and TLR8. Similar to FKBP5 and SLC6A4, PTSD risk was only associated with hypermethylation of SLC6A3 in the presence of a specific genotype, while hypermethylation of MAN2C1 was only associated with PTSD in individuals with greater trauma exposure. Overall, many DNA methylation signals have been linked to PTSD, but the role of DNA methylation and other epigenetic modifications in prior childhood abuse, trauma exposure, PTSD risk, and PTSD development remains unclear.

1.4.3 Transcriptomics

Previous gene expression studies of PTSD have focused on two types of biomarkers: predictive and diagnostic signatures. Predictive biomarkers indicate risk of PTSD development after trauma exposure and require longitudinal studies. Multiple longitudinal studies have presented panels of potential risk biomarkers. These biomarkers include up-regulated genes involved in type-1 interferon signaling (IFI27, OAS1, OAS2, OAS3, XAF1, and USP18) and differentially expressed genes which associate with glucocorticoid receptors (GR), including low FKBP5 and high GILZ expression. Additionally, many studies have suggested potential mRNA diagnostic
biomarkers, including the GR-associated genes FKBP5 and STAT5B, among others. Early markers for PTSD development have been reported which may aid in developing early detection diagnostic tools. Gene expression data collected in emergency rooms immediately following trauma led to the discovery of 574 differentially expressed genes distinguishing those who would later develop PTSD from those who would recover. These transcripts were enriched in genes involved in immune activation, cell cycle, and signal transduction, among other biological processes. While these and other initial results are promising, no validation studies have shown that these biomarkers are valid in independent data sets. Additionally, the biomarkers were identified strictly by differential expression statistics or panels derived from classification algorithms (e.g., support vector machines) and do not incorporate known or inferred biological network information. This limits the ability of these gene panels to give biological insight into the development and progression of PTSD and does not provide a clear path for identification of therapeutic targets.

1.4.4 Proteomics

While proteomic signals of PTSD may remain poorly understood, multiple protein markers have been suggested for Traumatic Brain Injury (TBI) and other PTSD-related conditions. Serum levels changes in C-reactive protein (CRP), amyloid A, and retinol binding protein 4 (RBP4) were observed in patients with TBI and elevated intracranial pressure. Other candidate biomarkers for predicting TBI include S100B, NSE, GFAP, and inflammatory markers and cytokines (including IL-6, and TNF-α). In addition to blood-based biomarkers, protein markers for TBI have also been identified in cerebrospinal fluid (CSF) and brain tissues. An overview of these candidate proteins is reviewed by Dash et al. These and other candidate markers identified from PTSD-related conditions may overlap with signals of PTSD, or may be used to distinguish PTSD cases with and without common comorbidities.

1.4.5 Neuroendocrine

The major constituents of the neuroendocrine response to physical and emotional threat and stress are the sympathetic nervous system (SNS) and the HPA axis. The immediate SNS response mobilizes the acute fight-or-flight response and is followed by a response from the HPA axis that reinstates homeostasis. PTSD patients showed low cortisol levels, which is surprising for a disorder precipitated by extreme stress, and differs from that observed in
studies of acute and chronic stress and major depressive disorder (which have been associated with increased cortisol levels)\textsuperscript{41}. Moreover, PTSD patients exhibited increased levels of norepinephrine, corticotropin-releasing hormone (CRH) and proinflammatory cytokines, reflecting reduced glucocorticoid signaling. This profile of alterations has been associated with PTSD pathophysiology. However, emerging research indicates that these alterations may instead reflect pre-traumatic vulnerabilities to the later development of PTSD\textsuperscript{40}.

\subsection*{1.4.6 Brain Imaging}

To gain a better understanding of the disorder, many studies have focused on macroscopic features of the human brain by employing techniques such as structural (sMRI) and functional magnetic resonance imaging (fMRI). The brain regions related to symptoms of the disorder include the hippocampus, amygdala, ventromedial prefrontal cortex (vmPFC), dorsal anterior cingulate cortex (dACC) and insular cortex (see Table 1.3). sMRI studies have shown a decrease in the volume of hippocampus\textsuperscript{42,43}, vmPFC\textsuperscript{44}, dACC\textsuperscript{25}, and the insular cortex\textsuperscript{46,47} for patients with PTSD. In the case of changes to the hippocampal volume, it is not clear whether this decrease is caused by the extreme trauma or simply represents a risk factor; however, some studies point to the latter\textsuperscript{48}. We note that exceptions to the above have been reported in children with PTSD\textsuperscript{49}, possibly due to lack of neuronal maturation, and in cases where symptom severity was not very strong\textsuperscript{50}. Studies performed with combat-exposed PTSD twins using voxel-based morphometry suggest that the reduction in ACC volume may be an acquired feature due to extreme trauma\textsuperscript{44}. Functional MRI studies have shown increased activity in the amygdala\textsuperscript{51,52}, dACC\textsuperscript{53}, and insular cortex\textsuperscript{54}, and decreased activity in the vmPFC\textsuperscript{55} for PTSD patients subjected to trauma-related stimuli. Positron emission tomography, an fMRI technique quantifying local changes in cerebral blood flow, has been performed on identical twins to study resting dACC glucose metabolism and dACC activation during non-emotional inference tasks. These studies demonstrated that combat-exposed veterans with PTSD and their co-twins have greater metabolism\textsuperscript{56} and greater activation\textsuperscript{57} compared to veterans without PTSD and their co-twins. This suggests that increased activation of dACC is a PTSD risk factor rather than a symptom. Finally, activation in the amygdala\textsuperscript{58} and insular cortex\textsuperscript{59} does not appear to be unique to PTSD, as these responses are also seen in other anxiety disorders.

Functional MRI techniques are also used to measure resting state brain activity and functional connectivity. Yan et al.\textsuperscript{60} found increased activity in the amygdala and the anterior insula, and decreased activity in the thalamus to be
common features in both resting state brain activity and task-based fMRI studies, in male combat-exposed veterans with PTSD. Their results also highlight decreased activity in the precuneus region of the brain, which is responsible for integrating information from the past and future. Furthermore, Rabinak et al. found enhanced connectivity between the amygdala and insula in combat-related PTSD patients. However, as many studies of anxiety-related disorders have also reported fMRI-based amygdala and insula activation, these responses may not be unique to PTSD.

### 1.5 PTSD Heterogeneity

The heterogeneous nature of the symptoms and development of PTSD creates challenges in diagnosing and treating the disease. While multiple studies have emphasized clear divisions into subgroups of PTSD, very little work has focused on identifying unique biological signatures for these subtypes. Instead, recent work has focused on using symptoms and symptom trajectories to identify these distinct subgroups, or on correlating specific biological measures to these symptom classes. For example, longitudinal data from the Jerusalem Trauma Outreach and Prevention Study identified three trajectories of PTSD symptoms which they called rapid-remitting, slow-remitting and non-remitting. Interestingly, they note that only subjects classified as slow-remitting showed symptom improvements when given antidepressants. Similar responses to antidepressants were seen in another study where promoter methylation of FKBP5 and GR showed differences between responders and non-responders to treatment. These examples further emphasize the existence of subgroups within PTSD, but overall provide little biological in-
sight into the key differences between these groups. In order to progress in the identification and characterization of these subtypes in a manner useful for diagnostic and therapeutic development, specific biological signatures must be extracted from large-scale data. In particular, the biological features extracted from this type of analysis would provide a more comprehensive picture of subtype biology than these single gene correlations. We note that the identification of subgroups from high throughput data has previously been applied to the study of diseases such as cancer and fibromyalgia

1.6 Thesis Overview

In Chapter 2, we present a consortium-based approach to identifying blood-based biomarkers for PTSD from multi-omic datasets. Using various data-driven, hypothesis-driven, and biological network-based approaches, a candidate set of 343 markers was identified and refined to a final panel of 28 multi-omic biomarkers. This panel achieved good performance in an independent validation dataset, with 81% accuracy, 85% sensitivity, 77% specificity, and 0.80 Area Under the Receiver-Operator Curve (AUC). Translation of these robust biomarkers to a cost-effective, field-deployable technology may provide a strategy for assessing PTSD risk and development in military populations.

In Chapter 3, a novel approach to identifying dysregulated disease subnetworks is implemented in multiple high throughput datasets. Incorporating additional statistical measures, including expression variance and distribution-based metrics, identified both biologically validated and novel disease networks. These disease subnetworks outperformed individual gene biomarkers, and presented densely connected sets of disease-related genes for future studies in disease characterization and therapeutic target identification.

In Chapter 4, we provide additional details regarding novel approaches for multi-omic biomarker identification, including incorporating incomplete data records and optimization for either sensitivity and specificity performance in biomarker predictions.

Finally, in Chapter 5, we discuss preliminary work in identifying subtypes of PTSD from multi-omic data, and use spanning trees to characterize disease progression trajectories. We explore how these methods could be used or adapted further in additional cohorts, to characterize subgroups of PTSD, identify and evaluate subgroup-specific biomarkers, and understand progression of and recovery from PTSD.
2 Multi-omic Biomarker Identification for Diagnosing Posttraumatic Stress Disorder*

2.1 Background

Combat-related Post-Traumatic Stress Disorder (PTSD) has a lifetime prevalence between 10.1%–30.9% in U.S. veterans of the Vietnam and subsequent conflicts, including the Iraq and Afghanistan wars. PTSD is precipitated by experiencing or witnessing actual or threatened death, serious injury, or violence, and has symptoms that include re-experiencing, avoidance, negative thoughts or moods associated with the traumatic event and hyper-arousal (DSM 5). There is limited understanding of the biological processes underlying the core features of PTSD.

CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

and associated psychiatric and somatic comorbidity. Limited progress in the discovery of biological markers of PTSD has hampered accurate diagnosis, early identification of cases, staging and prognosis, stratification, personalized treatment, and new drug development. Additionally, individuals meeting diagnostic criteria for PTSD represent a heterogeneous group, as evidenced by differences in symptomatology, course and treatment response. Currently, case identification is limited by heavy reliance on self-reported symptoms for a disorder in which many trauma survivors under-report symptoms because of stigma, and some over-report symptoms for financial or other gains. Personalized treatment selection is limited by errors of omission (failing to identify individuals who would likely benefit from a specific behavioral or biological treatment) and errors of commission (treating individuals who are unlikely to benefit from a specific treatment), in part because of the lack of validated diagnostic markers.

Previous PTSD biomarker studies have primarily focused on using gene expression for predicting risk and diagnosis. These studies have demonstrated moderate success in identifying predictive and diagnostic markers, but have been limited due to small sample sizes, as well as the focus on an individual molecular data type. In cancer, multi-site, integrated multi-omic studies have shown great promise in generating novel insights into disease mechanism, diagnostic and predictive markers, and signals of progression and stratification. These studies have included high-throughput ‘omics data such as genomics, transcriptomics, proteomics, methylomics, lipidomics and metabolomics. By employing a systems biology framework, multi-omic datasets provide the ability to understand the underlying disease network-associated biological processes.

The systems biology approach aims to characterize a large and diverse set of molecules within an illness or individual by examining entire biological systems, not just individual components, allowing the assessment of interactions among levels of cellular pathology, ranging from DNA to circulating metabolites. This approach has the potential to provide a more comprehensive characterization of illnesses, to track underlying biological dysregulation before clinical symptoms develop or worsen, to lead to the identification of improved diagnostic markers, and to allow for the discovery of novel targets for treatment.

In 2012, the Department of Defense initiated a multi-site “PTSD Systems Biology Consortium”, which applied multiple ‘omics technologies to the same sample of combat-exposed PTSD and control participants. The goals of the PTSD Systems Biology Consortium included developing a reproducible panel of blood-based biomarkers with good sensitivity and specificity for PTSD diagnosis. Here, we present identification and validation of a set of
biomarkers for diagnosing warzone-related PTSD.

2.2 Methods

2.2.1 Participant Recruitment, Enrollment, and Clinical Assessment

Participants were recruited from the Mental Health Services of the Manhattan, Bronx and Brooklyn Veterans Affairs Medical Centers, other regional VA medical centers, Veterans Service Organizations, National Guard, reservist agencies and organizations from the general community. Recruitment methods included flyers, in-person presentations, media advertisements, internet postings (e.g. Craigslist) and referral from clinicians.

Participants were compensated for their participation. Each enrolled participant completed an initial phone screen to determine preliminary eligibility. Participants who appeared from the phone screen to be potential matches were then invited to come in person for the informed consent process, which included a complete description of the study. Participants that consented then completed a clinical assessment using the Clinician Administered PTSD Scale (CAPS) and Structured Clinical Interview for DSM-IV (SCID), to determine eligibility for the study. The CAPS was used to determine the presence versus absence of a DSM-IV diagnosis of PTSD related to combat trauma exposure and to assess symptom severity of the combat-related PTSD. The SCID was conducted to assess for mood, anxiety, psychosis, and substance related disorders. The SCID PTSD module was also used to determine whether participants had ever met DSM-IV diagnostic criteria for PTSD related to a non-combat traumatic exposure.

In addition, participants were asked to complete self-report questionnaires, which provide background information as well as assessments related to PTSD symptoms, head injuries, physical health, depression, anxiety, alcohol use, trauma exposure, and measures of potential covariates for use in biomarkers analyses.

General inclusion criteria included being an OEF/OIF male veteran between the age of 20 and 60 years old, being able to understand the protocol and sign written informed consent, and meeting criteria for either PTSD positive or PTSD negative groups. PTSD positive participants were defined as participants who met CAPS criteria for current war zone-related PTSD for at least 3 months duration, as indexed by the DSM-IV with CAPS total score >40, which was calculated by summing each symptom on frequency and intensity ratings. Full criteria for DSM-IV diagnosis of PTSD was also met for all PTSD positive participants. PTSD negative controls were combat
exposed veterans that were negative for lifetime combat or civilian PTSD and had a current CAPS total score <20. These distinct groups were chosen to ensure clinically significant levels of symptomatology in the cases, and to exclude subthreshold cases in the control group. All study participants experienced combat traumas as described in criterion A of DSM-IV PTSD diagnostic criteria.

Exclusion criteria included history of open head injury or closed head injury with loss of consciousness of more than 10 min, any head injury leading to a currently elevated score of post concussive symptoms, current drug abuse or dependence within the past year, current alcohol dependence or history of dependence within the past 8 months, lifetime history of any psychiatric disorder with psychotic features, bipolar disorder, or obsessive-compulsive disorder, currently exposed to recurrent trauma or exposed to a traumatic event within the past 3 months, participants with prominent suicidal or homicidal ideation, neurologic disorder or systemic illness affecting central nervous system function, anemia, recent blood donation in the past 2 months, and participants who were not stable for 2+ months on psychiatric medication, anticonvulsants, antihypertensive medication or sympathomimetic medication.

All clinicians who conducted the clinical interviews for this study were post-doctorate level psychologists who had several years of experience working with veterans and trauma. All interviews were recorded and calibrated weekly with the PTSD program clinical team across sites, including a senior clinician, which ensured a similar use of diagnostic measures and consistent application of inclusion and exclusion criteria.

The SCID was used to determine whether participants met DSM-IV diagnostic criteria for mood, anxiety, psychotic and substance use disorders. The CAPS was used to determine combat related PTSD status, as well as the severity of current PTSD symptoms (past month is the “CAPS current”) and the severity of the most severe lifetime episode of combat-related PTSD (“CAPS lifetime”).

### 2.2.2 Blood draws

Participants reported to the laboratory at James J. Peters VA Medical Center (JJP VAMC) or Icahn School of Medicine at Mount Sinai (ISMMS) between 7:30 and 8:00am after an overnight fast. Vital signs, weight, height and waist-hip ratio were measured and then approximately 160 cc of whole blood was collected and processed for subsequent assays, including the Dexamethasone Suppression Test (DST). Participants received a 0.50 mg tablet of Dexamethasone to ingest at 11:00pm and returned the following morning at 8 am (post-Dex) for collection of 10cc of blood.
CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

Table 2.1: Summary of blood tubes and samples for molecular assays.

<table>
<thead>
<tr>
<th>Markers</th>
<th>Tube Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA-Genomic Markers</td>
<td>Yellow Top ACD</td>
</tr>
<tr>
<td>RNA-Gene Expression</td>
<td>PAX-gene (RNA)</td>
</tr>
<tr>
<td>Endothelial Progenitor Cells</td>
<td>EDTA Lavender Top (LTT)</td>
</tr>
<tr>
<td>Cytokines and Brain Derived Neurotrophic Factors</td>
<td>SST Red/Gray Top</td>
</tr>
<tr>
<td>Neurosteroids and Metabolomics</td>
<td>EDTA Lavender Top (LTT)</td>
</tr>
<tr>
<td>Oxidative stress markers</td>
<td>EDTA Lavender Top (LTT)</td>
</tr>
<tr>
<td>miRNA, DNA Methylation and Metabolomics</td>
<td>EDTA Lavender Top (LTT)</td>
</tr>
<tr>
<td>DNA-Methylation</td>
<td>PAX-gene (DNA)</td>
</tr>
<tr>
<td>Peripheral Blood Mononuclear Cells</td>
<td>EDTA Lavender Top (LTT)</td>
</tr>
</tbody>
</table>

Blood samples were processed and aliquoted for storage into whole blood, serum, plasma, buffy coat or peripheral blood mononuclear cells (PBMCs) depending on the assay (Table 2.1). PAXgene tubes were used to collect samples for DNA and RNA. Samples were stored, inventoried and frozen at -80°C using Freezerworks software (Mountlake Terrace, WA). Samples were shipped in batches on dry ice to each of the respective assaying laboratories, with the following exceptions: one tube of whole blood (unfrozen) was shipped to UCSF for FACS sorting for cell type and activity on the day of blood collection; the IC-50DEX assay was performed on live (unfrozen) PBMCs beginning on the day of collection; blood samples were delivered to ISMMS or JJP VAMC CLIA-certified laboratories for same-day assays of standard clinical labs, including cell counts. Urine toxicology testing was also conducted at the time of blood collection.

2.2.3 Molecular Assays

Blood samples were assayed for many molecular species, including genetics, methylomics, proteomics, metabolomics, immune cell counts, cell aging, endocrine markers, miRNAs, cytokines, and more. Genotyping was performed using DNA extracted from blood, with the Illumina Infinium PyschArray BeadChip. DNA methylation was quantified using two approaches: a genome-wide unbiased approach, and a targeted sequencing-based approach. The genome-wide methylation approach quantified methylation using the Illumina Infinium HumanMethylation450K BeadChip array. Using targets generated from this genome-wide approach, as well as other hypotheses generated from literature, a smaller set of methylation sites were evaluated by targeted sequencing via Zymo Research. Plasma miRNAs were evaluated using small RNA sequencing, and processed using sRNAAnalyzer. Proteins were evaluated using three methods: peptide quantification using selected reaction monitoring (SRM), quantification of six neurodegenerative disease-related markers using the Human Neurodegenerative Disease Panel 1, and quantification of serum levels of BDNF using a BDNF ELISA assay. Non-targeted metabolomics analysis was conducted using three platforms: ultrahigh performance liquid chromatography/tandem mass spectrom-
CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

Table 2.2: Summary of biomarker identification approaches. Data-driven approaches used algorithms to identify biomarkers from the discovery datasets. Hypothesis-driven approaches selected specific biomarkers based on literature, additional datasets, or similar diseases.

<table>
<thead>
<tr>
<th>Approach</th>
<th>Classifier</th>
<th>Data-driven</th>
<th>Hypothesis-driven</th>
<th>Incorporates biological networks</th>
<th>Single-omic implementation</th>
<th>Multi-omic implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method A</td>
<td>SVM</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method B</td>
<td>SVM</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method C</td>
<td>SVM</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method D</td>
<td>logistic regression</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method E</td>
<td>Tree-based boosting</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method F</td>
<td>LDS</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method G</td>
<td>Polygenic Risk Score</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method H</td>
<td>Random Forest</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method I</td>
<td>Lasso</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Additional data types, including routine clinical lab values and physiological measurements, were collected using standard procedures.

2.2.4 Candidate Biomarker Identification

Using a “wisdom of crowds” approach, a diverse set of algorithms were used to identify candidate PTSD biomarkers. These strategies included traditional data-driven machine learning methods, including Recursive Feature Elimination (RFE) with Support Vector Machines (SVM), biological network driven approaches, including an adapted version of COMBINER \(^{102,103}\), as well as hypothesis-driven approaches based on previously identified candidate PTSD biomarkers. In total, nine approaches were used to generate 50 candidate biomarker panels, totaling 343 unique candidate markers for PTSD (Table 2.2). The 343 candidate markers are generated from 12 data types/assays (Table 2.3).

2.3 Results

2.3.1 Candidate Biomarker Identification and Panel Refinement

A set of three cohorts totaling 281 samples from male combat veterans from Operation Enduring Freedom (OEF) and/or Operation Iraqi Freedom (OIF) conflicts were recruited as part of a larger study designed to identify biomarkers for PTSD diagnosis using a combination of clinical, genetic, endocrine, multi-omic and imaging information (Fig. 2.1). Participants were recruited in three cohorts: discovery, recall, and validation. The discovery cohort (co-
CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

Table 2.3: Summary of candidate biomarkers.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Number of candidate biomarkers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biometrics</td>
<td>2</td>
</tr>
<tr>
<td>Clinical Lab</td>
<td>20</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>27</td>
</tr>
<tr>
<td>Methylation – Illumina array</td>
<td>123</td>
</tr>
<tr>
<td>Methylation – targeted sequencing</td>
<td>33</td>
</tr>
<tr>
<td>miRNA</td>
<td>81</td>
</tr>
<tr>
<td>Neuroendocrine</td>
<td>8</td>
</tr>
<tr>
<td>Peptides – SRM</td>
<td>38</td>
</tr>
<tr>
<td>Protein – multiplex</td>
<td>2</td>
</tr>
<tr>
<td>Protein – BDNF</td>
<td>1</td>
</tr>
<tr>
<td>Small Molecules</td>
<td>4</td>
</tr>
</tbody>
</table>

| Nonlinear feature combinations | 4 |

Total 343

Figure 2.1: Overview of PTSD biomarker identification approach—details of cohort recruitment, and biomarker identification, down-selection, and validation.

Cohort 1) consisted of 83 PTSD positive and 82 trauma-exposed PTSD negative participants who met the inclusion and exclusion criteria. All participants completed clinical interviews and blood draws. After assessment of data
Figure 2.2: Overview of molecular datasets and cohort symptom severity. (A) Flow diagram for participant recruitment and enrollment. Participant eligibility was determined through a phone pre-screen and a baseline diagnostic clinical interview. Eligible participants completed fasting blood draws for multi-omic molecular assays. Participants in the initial discovery cohort were invited to return for follow-up in the recall cohort. Some participants returned with symptom changes, including “subthreshold” PTSD symptoms (below original study inclusion criteria). (B) Trajectory of PTSD symptoms in recalled participants. CAPS total for current symptoms at baseline (T0) and follow-up (T1) for each participant are connected. Participants who remained in the PTSD+ group at both time points are shown in purple. Participants who remained in the PTSD- group are shown green. Participants with PTSD status changes are shown in gray, including participants who became “subthreshold” PTSD cases. (C) Distribution of molecular data types at three stages of biomarker identification: Full exploratory dataset (All Data), reduced set of 343 potential biomarkers (candidate set) and the final panel of 28 biomarker (final set). Methylation and GWAS data represents 99% of initial data screen due to high-throughput arrays. Other molecular data types are well represented in the second and final stages of biomarker identification and selection.

In quality, 77 PTSD positive and 74 trauma-exposed PTSD negative samples were available with all completed blood marker assays. This discovery cohort was used to generate an initial pool of candidate biomarkers. Participants from the discovery cohort were invited back for clinical re-evaluation and a blood draw approximately three years
CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

after their initial evaluation. This cohort of recalled subjects (recall cohort, cohort 2), included 55 participants from the initial discovery cohort. Some of these participants showed PTSD symptom and status changes based on clinical assessment. In addition, some participants no longer met the original inclusion/exclusion criteria for the study; these participants had symptoms intermediate between the PTSD positive and PTSD negative groups, in some cases meeting criteria for subthreshold PTSD. The 55 recall participants included 15 PTSD positive, 11 subthreshold PTSD, and 29 PTSD negative participants. The third cohort, an independent group of 26 PTSD positive and 26 PTSD negative participants, became the validation cohort (cohort 3), used for validating the final set of PTSD biomarkers. A summary of the cohort recruitment and final cohort sizes is shown in Fig. 2.2A.

To identify a minimally invasive PTSD diagnostic panel, blood-based multi-omics and other analytes were assayed for each individual (and during both visits for recalled participants), including DNA methylation, proteomics, metabolomics, miRNAs, small molecules, endocrine markers, and routine clinical lab panels. Additionally, physiological measures were recorded and non-linear marker combinations were computed. Using a strategy described in the next sections, a robust and diverse 28-member biomarker panel for diagnosing PTSD was identified from this pool of more than one million markers (Fig. 2.2A).

Using a variety of data-driven, hypothesis-driven, hybrid, and other approaches (Table 2.2), we identified a set of candidate diagnostic panels, totaling 343 unique potential biomarkers, from an exploratory search of almost one million markers in the discovery cohort (Step 2 from Fig. 2.1). To filter and refine the pool of candidate biomarkers, we used data from recalled participants (recall cohort, cohort 2). Many of these returning participants experienced symptom changes over the 3.3±0.9 years (mean±sd) between the initial and follow-up evaluation. Clinical Administered PTSD Scale (CAPS) totals for recalled participants at both time points are shown in Fig. 2.2B. The panel was refined using the recall cohort along with a two-stage down-selection approach to select the final set of PTSD biomarkers (Steps 4-5 from Fig. 2.1).

The two-stage down-selection process is based on the following methodology. Poor performance biomarkers were eliminated one-by-one based on biomarker performance (Fig. 2.1, Step 4). Specifically, 10 iterations of model training and validation were performed for each feature set, with a single biomarker removed. Bootstrapped versions of the discovery and recall cohort were used for model training and validation, respectively. Validation performance was determined based on Area Under the Receiver Operator Characteristic (ROC) curve (AUC). Candidate biomarkers were removed one-by-one based on the largest average AUC of the remaining biomarker set. After elimi-
Figure 2.3: Two stage biomarker down-selection approach. (A) Stage 1: one-by-one feature elimination using recalled participants for biomarker validation. Features are eliminated one at a time, starting with n=343, based on maximum average validation AUC. The selected local maximum validation AUC occurs at n=77 biomarkers. (B) Stage 2: random forest variable importance ranking. Remaining 77 biomarkers were sorted based on random forest variable importance in Cohort 2. Features with importance <30% of maximum importance are retained for final biomarker validation (n=28).

To reduce the number of features in the panel, we implemented a second stage of down-selection, based on Random Forest Variable Importance (Fig. 2.1, Step 5). Using the recall cohort, the remaining 77 biomarkers were sorted based on random forest variable importance (Fig. 2.3B). We retained biomarkers with importance greater than 30% of the maximum importance score for the final biomarker panel (n=28).
Table 2.4: PTSD biomarker confusion matrix. Distribution of predicted PTSD cases and controls vs clinically determined cases and controls in the validation cohort.

<table>
<thead>
<tr>
<th>Predicted PTSD</th>
<th>PTSD Status (assessed by clinical interview)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PTSD Positive</td>
<td>17 (True Positives) 3 (False Negatives)</td>
<td>20</td>
</tr>
<tr>
<td>PTSD Negative</td>
<td>9 (False Negatives) 23 (True Negatives)</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>26</td>
<td>52</td>
</tr>
</tbody>
</table>

2.3.2 Biomarker validation

After the two-stage feature reduction strategy, the final biomarker set consisted of 28 features, including methylation, metabolomics, miRNA, protein, and other data types. A random forest model trained on the combined cohorts 1 and 2 predicted PTSD status in an independent validation set (cohort 3) with an area under the ROC curve (AUC) of 0.80 (95% CI 0.66-0.93, Fig. 2.4A). Using the point closest to (0,1) on the ROC curve (shown in Fig. 2.4A), the model was validated with an accuracy of 81%, sensitivity of 85%, and specificity of 77%. The PTSD positive participants in the validation cohort had CAPS ranging from 47-114. We found that predicted...
PTSD scores from the random forest model for these cases were correlated with CAPS (r=0.59, p=0.001) indicating the current biomarker model predicts not only disease status, but potentially PTSD symptom severity of cases (Fig. 2.4B). In addition, predicted PTSD scores were moderately correlated with individual CAPS B (re-experiencing), C (avoidance), and D (hyperarousal) sub-scores (r=0.44-0.53, Supplementary Fig. 4), suggesting that the identified molecular markers are not specific to a single symptom cluster, but to overall symptoms.

Overall, the set of identified PTSD biomarkers contains many molecular data types (DNA methylation, miRNAs, proteins, metabolites, and others), with signals primarily including under-expressed proteins and miRNAs, and signatures of DNA hypermethylation with highly consistent signals. Of the 28 markers comprising the final panel, 17 markers had consistent fold change directions in all three cohorts (Table 2.7). Five of the final 28 markers were retained during panel refinement even though the fold change direction was inconsistent between the discovery and recall cohorts, indicating that these features may contain relevant PTSD signal that is not purely measured by group differences in mean. A post hoc analysis of the biomarker panel performance without these inconsistent features resulted in decreased validation performance (AUC=0.68 and 0.72 when using only markers with consistent fold change directions across the discovery and recall cohorts (23 markers), and all three cohorts (17 markers), respectively). While the level of fold change decreased overall between training and validation, the direction of PTSD signal was retained. Primarily, the signals comprising the final set of biomarkers included under-expressed proteins and miRNAs in PTSD participants, as well as signatures of DNA hypermethylation.

Using random forest variable importance, the top 10 biomarkers from the final 28 marker panel included five of the six molecular data types: DNA methylation, physiological, miRNAs, clinical lab measures, and metabolites (Fig. 2.4C). These data types contribute primarily uncorrelated signals, with only small clusters of moderate to highly correlated biomarkers from three data types: proteins, miRNAs, and DNA methylation (Fig. 2.4D).

Through the biomarker identification and down-selection process, two intermediate biomarker sets were identified, consisting of 343 and 77 candidate biomarkers. Trained random forest models on these biomarker sets validated with slightly lower AUCs than the final biomarker panel (AUCs of 0.74, 0.75, and 0.80 in the 343, 77, and 28 biomarker panels; Fig. 2.4E). The consistent validation AUC indicates robust signal in these sets of candidate biomarkers, without loss of signal during down-selection from 343 to 28 features. The final panel of 28 markers consisted of six different data types: routine clinical lab markers, metabolites, DNA methylation marks, miRNAs, proteins, and physiological measurements. The combined panel out-performed all six panels composed of each
Table 2.5: Predictive performance of nonlinear biomarkers. GABR and lactate/citrate both outperform their individual components in prediction AUCs and accuracy by 0.05-0.16.

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Arginine Bioavailability Ratio (GABR)</td>
<td>0.65</td>
<td>0.67</td>
</tr>
<tr>
<td>Arginine</td>
<td>0.50</td>
<td>0.51</td>
</tr>
<tr>
<td>Ornithine</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Citrulline</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lactate/Citrate</td>
<td>0.61</td>
<td>0.68</td>
</tr>
<tr>
<td>Lactate</td>
<td>0.56</td>
<td>0.53</td>
</tr>
<tr>
<td>Citrate</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table highlights the predictive performance of different biomarkers, showing that the combination of GABR and lactate/citrate yields higher AUC and accuracy compared to their individual components.

individual data type (Fig. 2.4E), demonstrating the power of combining different types of markers in a diverse biomarker panel, capable of capturing the complexities of PTSD.

2.3.3 **Nonlinear biomarkers**

Two biomarker features included in our final panel are computed, nonlinear metrics: Global Arginine Bioavailability Ratio (GABR), where

\[
GABR = \frac{\text{Arginine}}{\text{Ornithine} + \text{Citrulline}},
\]

and lactate/citrate. These computed ratios outperform their combined individual components in predictive performance (Table 2.5, Fig. 2.4F). In addition, these ratios begin to alleviate single-sample normalization issues that need to be addressed for clinical use of a biomarker diagnostic panel.

2.3.4 **Evaluating clinical and demographic factors**

The cohorts recruited for this study are diverse, but restricted to participants with CAPS less than 20 (PTSD negative) or greater than 40 (PTSD positive). The heterogeneity of the participants included in these three cohorts, including race, age, and clinical comorbidities, as well as PTSD severity are shown in Table 2.6.

To evaluate the performance of this biomarker panel in the context of participant demographics and other clinical factors, we computed biomarker performance in stratified subsets of the validation cohort. While biomarker performance was highest in Hispanic participants (AUC=0.95), we observed no statistically significant differences in AUC across ethnicities (Fig. 3F). Multiple studies have examined the increased prevalence and greater symptom
### Table 2.6: Summary of cohort demographics and clinical symptoms.

<table>
<thead>
<tr>
<th></th>
<th>Cohort 1</th>
<th>Cohort 2</th>
<th>Cohort 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PTSD+ (n=77)</td>
<td>PTSD- (n=74)</td>
<td>PTSD+ (n=15)</td>
</tr>
<tr>
<td><strong>Age, years [mean (sd)]</strong></td>
<td>32.8 (7.4)</td>
<td>32.6 (8.0)</td>
<td>33.7 (8.2)</td>
</tr>
<tr>
<td><strong>Race/ethnicity [n (%)]</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic</td>
<td>34 (44%)</td>
<td>24 (32%)</td>
<td>9 (60%)</td>
</tr>
<tr>
<td>Non-Hispanic Asian</td>
<td>1 (1%)</td>
<td>5 (7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Non-Hispanic black</td>
<td>21 (27%)</td>
<td>16 (22%)</td>
<td>4 (27%)</td>
</tr>
<tr>
<td>Non-Hispanic white</td>
<td>18 (23%)</td>
<td>24 (32%)</td>
<td>2 (13%)</td>
</tr>
<tr>
<td>Non-Hispanic other</td>
<td>3 (4%)</td>
<td>5 (7%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Education [n (%)]</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 12th grade</td>
<td>2 (3%)</td>
<td>0 (0%)</td>
<td>1 (7%)</td>
</tr>
<tr>
<td>HS diploma or GED</td>
<td>27 (35%)</td>
<td>13 (18%)</td>
<td>2 (13%)</td>
</tr>
<tr>
<td>2 years college, AA degree</td>
<td>23 (30%)</td>
<td>21 (28%)</td>
<td>6 (40%)</td>
</tr>
<tr>
<td>4 years college, BA degree</td>
<td>22 (29%)</td>
<td>28 (38%)</td>
<td>5 (33%)</td>
</tr>
<tr>
<td>Major's degree</td>
<td>3 (4%)</td>
<td>11 (15%)</td>
<td>1 (7%)</td>
</tr>
<tr>
<td>Doctoral degree</td>
<td>0 (0%)</td>
<td>1 (1%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td><strong>Body mass index [mean (sd)]</strong></td>
<td>30.1 (5.1)</td>
<td>28.1 (4.3)</td>
<td>32.2 (5.3)</td>
</tr>
<tr>
<td><strong>Cholesterol [mean (sd)]</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HDL Cholesterol</td>
<td>47.6 (11.9)</td>
<td>50.1 (13.3)</td>
<td>43.0 (8.0)</td>
</tr>
<tr>
<td>LDL Cholesterol</td>
<td>108.2 (25.4)</td>
<td>100.0 (26.4)</td>
<td>115.1 (26.4)</td>
</tr>
<tr>
<td><strong>hbA1c [mean (sd)]</strong></td>
<td>5.4 (0.9)</td>
<td>5.4 (0.4)</td>
<td>5.6 (0.8)</td>
</tr>
<tr>
<td><strong>PTSD Severity, Total CAPS [mean (sd)]</strong></td>
<td>68.0 (16.1)</td>
<td>3.6 (4.9)</td>
<td>69.5 (18.5)</td>
</tr>
<tr>
<td><strong>Early Trauma Exposure, ETISR total [mean (sd)]</strong></td>
<td>7.7 (5.8)</td>
<td>5.1 (4.0)</td>
<td>5.2 (4.6)</td>
</tr>
<tr>
<td><strong>Major Depressive Disorder [n (%)]</strong></td>
<td>42 (55%)</td>
<td>0 (0%)</td>
<td>9 (64%)</td>
</tr>
<tr>
<td><strong>Peritraumatic Dissociate Experience, Rater Version [mean (sd)]</strong></td>
<td>1.8 (0.4)</td>
<td>1.2 (0.2)</td>
<td>1.8 (0.5)</td>
</tr>
<tr>
<td><strong>Peritraumatic Distress Inventory, Rater Version [mean (sd)]</strong></td>
<td>2.1 (0.8)</td>
<td>1.1 (0.6)</td>
<td>2.2 (0.7)</td>
</tr>
<tr>
<td><strong>Sleep Quality, PSQI [mean (sd)]</strong></td>
<td>13.1 (3.3)</td>
<td>5.9 (3.6)</td>
<td>11.8 (3.5)</td>
</tr>
<tr>
<td><strong>Number of tours [mean (sd)]</strong></td>
<td>1.8 (0.9)</td>
<td>1.7 (0.8)</td>
<td>1.5 (0.6)</td>
</tr>
</tbody>
</table>
Figure 2.5: Summary of potential confounders from PTSD biomarker validation. Using one-way ANOVA, no significant differences were observed between True Positive, False Negative, True Negative, and False Positive groups for: (A) number of military service tours, (B) childhood trauma, using the Early Trauma Inventory Self Report, (C) age, (D) education level, or (E) Body Mass Index (BMI). (F) Sleep quality, based on Pittsburgh Sleep Quality Index (PSQI) was significantly different between True Positive cases and (i) True Negatives, and (ii) False Positives.

The severity of PTSD in Hispanic populations\textsuperscript{104,105} which may correspond to stronger biological signals, leading to the differences in AUC. Next, we evaluated the performance of this biomarker panel in the context of participant demographic and lifestyle factors. We observed no differences in biomarker validation performance based on ethnicity, smoking status, age, body mass index, or sleep quality (Fig. 2.5A-B, D-F). In the validation cohort, 35% of PTSD cases also met the criteria for Major Depressive Disorder (MDD). Using the identified biomarker panel and model, these PTSD+/MDD+ cases could be distinguished from all controls with an AUC of 0.92, while the PTSD+/MDD- could only be distinguished from controls with an AUC of 0.73 (Fig. 2.4F). Similarly, predicted PTSD scores were more strongly correlated with PTSD symptom severity in PTSD+/MDD+ participants than in
CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

Figure 2.6: Predicted probability of PTSD based on trained random forest model using a biomarker panel of 28 features in validation cohort (26 PTSD, 26 control). (A) In PTSD+/MDD- participants, predicted PTSD probability is only weakly correlated with PTSD symptom severity, measured by CAPS (r=0.37). (B) In PTSD+/MDD+ participants, predicted PTSD probability is more strongly correlated with PTSD symptom severity, measured by CAPS (r=0.64).

PTSD+/MDD- participants, with r=0.64 and r=0.37, respectively (Fig. 2.6). This decreased in prediction accuracy and correlation with PTSD symptoms in the absence of comorbid MDD indicates a potential overlap of biological signals for MDD and PTSD that should be explored further.

Overall, the set of identified PTSD biomarkers contain many molecular data types (methylation marks, miRNAs, proteins, metabolites, and others), with highly consistent signals. Of the 28 markers comprising the final panel, 17 markers had consistent fold change directions in all three cohorts (Table 2.7). While the level of fold change decreased overall between training and validation, the direction of PTSD signal was retained. Primarily, the signals comprising the final set of biomarkers included under-expressed proteins and miRNAs in PTSD patients, as well as signatures of DNA hypermethylation.

2.4 Discussion

This study presents the identification and validation of a biomarker panel for the diagnosis of PTSD. The panel consists of 28 features that perform well in identifying PTSD cases from combat-exposed controls in a male, veteran population (77% accuracy). These markers were identified based on predictive ability, feature stability, and
<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Data Type</th>
<th>Cohort 1</th>
<th>Cohort 2</th>
<th>Cohort 3</th>
<th>Literature findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>70672835 (SHANK2)</td>
<td>Methylation</td>
<td>↓</td>
<td>↑</td>
<td>↓</td>
<td>Genetic variants of SHANK2 associated with schizophrenia[109]</td>
</tr>
<tr>
<td>75938326 (C2orf3)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>AFM-LPN [106]</td>
</tr>
<tr>
<td>75938338 (C2orf3)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>Genetic variants of SHANK2 associated with schizophrenia[109]</td>
</tr>
<tr>
<td>AFM-LPN</td>
<td>Protein</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>cg01208318</td>
<td>Methylation</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>cg01409026 (MLH1)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>cg0343241</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>cg04121006 (CES2)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>cg15687973 (PDE9A)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>CPT1B expression has been associated with monoamine neurotransmitter regulation and depression[107]</td>
</tr>
<tr>
<td>cg17137457 (CPT1B)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>CPT1B expression has been associated with rodent stress and human PTSD[108]</td>
</tr>
<tr>
<td>cg20578780</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>cg26456901 (MDC1)</td>
<td>Methylation</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>CPN1-IVQ</td>
<td>Protein</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>CPN2-LN</td>
<td>Protein</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>CTSS-GID</td>
<td>Protein</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>F10-NCE</td>
<td>Protein</td>
<td>↑</td>
<td>↓</td>
<td>↓</td>
<td></td>
</tr>
<tr>
<td>GABR</td>
<td>Metabolite</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>Decreased coagulation in PTSD[90]</td>
</tr>
<tr>
<td>Gammaglutamyltyrosine</td>
<td>Metabolite</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>GABR is decreased in patients in MDD[109]; previous PTSD finding[110]</td>
</tr>
<tr>
<td>heart rate</td>
<td>Physiological</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>Gamma-glutamyltyrosine negatively correlated with leukocyte telomere length[111]</td>
</tr>
<tr>
<td>hsa-miR-133a-3p</td>
<td>miRNA</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>hsa-miR-192-5p</td>
<td>miRNA</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>Elevated heart rate following trauma associated with PTSD[112]</td>
</tr>
<tr>
<td>hsa-miR-424-3p</td>
<td>miRNA</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>hsa-miR-9-5p</td>
<td>miRNA</td>
<td>↓</td>
<td>↑</td>
<td>↑</td>
<td>Abundant in liver; associated with obesity and diabetes[113,114]</td>
</tr>
<tr>
<td>Insulin</td>
<td>Clinical Labs</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>miR-424-3p has been associated with inflammation[115]</td>
</tr>
<tr>
<td>ITIH2-VQF</td>
<td>Protein</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>Enriched in brain tissue and regulated neurogenesis[116]</td>
</tr>
<tr>
<td>Lactate/citrate</td>
<td>Metabolite</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>Increased insulin resistance in veterans with PTSD[117]</td>
</tr>
<tr>
<td>Mean platelet volume</td>
<td>Clinical Labs</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td></td>
</tr>
<tr>
<td>PTGDS-AQG</td>
<td>Protein</td>
<td>↑</td>
<td>↑</td>
<td>↑</td>
<td>Lactate has been considered panic-inducing in both panic disorder and PTSD[118]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Increased mean platelet volume associated with panic disorder[119] and major depression[120]</td>
</tr>
</tbody>
</table>

Table 2.7: Overview of biomarker signals in each of the three cohorts. Red and green cells indicate upregulated and downregulated signals, respectively. Dark colors indicate consistent signals across all three cohorts, while lighter colors indicate contradictory signal directions.
dynamics with changing PTSD symptoms over three years. Some of the biomarkers have been linked to PTSD previously, including elevated heart rate\(^{112}\) and decreased level of coagulation factors\(^9\), and also included markers linked to MDD, anxiety, and other co-morbid conditions, including platelet volume\(^{119,120}\), insulin resistance\(^{117,123}\) and alterations in the SHANK2 gene\(^{106}\) (Table 2.7).

In particular, the circulating miRNAs selected in the panel reflect the diverse pathology and comorbidities present in PTSD populations, including connections to metabolic diseases and cardiovascular conditions. The miR-133-3p, a member of myomiRs that are highly abundant in muscle, including cardiac muscle, has been implicated in cardiomyocyte differentiation and proliferation\(^{124}\). The circulating miR-133-3p level has been linked to various cardiovascular disorders including myocardial infarction, heart failure, and cardiac fibrosis\(^{125,126}\). The miR-9-5p is enriched in brain\(^{116}\) and known as a regulator for neurogenesis. It is also involved in heart development and heart hypertrophy\(^{43}\). The miR-192 is highly abundant in the liver and circulating miR-192-5p levels have been associated with various liver conditions as well metabolic diseases such as obesity and diabetes\(^{113,114}\). The circulating miR-192 level has also been used as a biomarker for ischemic heart failure\(^{127}\).

In addition to molecular markers, our approach selected heart rate as a contributor to the PTSD diagnostic panel. More than two decades ago, heart rate differences were observed between eventual PTSD cases and controls during emergency room visits and at one week follow-ups after trauma\(^{112}\). While these differences did not persist for longer time points in Shalev’s study, we observed significant mean group differences for heart rates in two of the three cohorts from this study, a number of years following trauma exposure (\(p<0.01\) for discovery and validation cohorts). Heart rate alone predicts diagnosis of PTSD in the validation cohort with 69% accuracy. Of note, removing heart rate from our biomarker panel did not result in significant decreased model performance (molecular-only panel without heart rate still achieves 77% accuracy).

To understand the clinical utility of the proposed biomarker panel, further validation is required in other PTSD populations, including active duty soldiers, populations with civilian trauma, female cohorts, and carefully phenotyped populations with and without many conditions commonly comorbid with PTSD. The cohorts recruited for this study were subject to strict inclusion and exclusion criteria, intentionally creating a pool of chronic and moderate to severe cases of PTSD to compare with asymptomatic controls. This study design may have allowed for the clearest and strongest signals of PTSD to emerge, but will need additional validation in patients early in PTSD development, during and after recovery, and those with intermediate PTSD symptoms (CAPS from 20-
CHAPTER 2. MULTI-OMIC BIOMARKER IDENTIFICATION FOR PTSD

40), where the current model performance may be decreased. Previous studies of gene expression-only predictors of PTSD risk have shown strong performance in younger, active duty populations (70-80% accuracy)\textsuperscript{28}, indicating that biomarker signals may exist pre-trauma and during early stages of PTSD development. These previous biomarkers also contain many immune-related genes, indicating strong overlap of identified signals for PTSD risk and diagnosis.

Additionally, many studies have emphasized the high rates of PTSD comorbidity with other conditions, including depression\textsuperscript{128}, anxiety\textsuperscript{129}, alcoholism and substance abuse\textsuperscript{130}, cardiovascular disease\textsuperscript{131}, diabetes\textsuperscript{132}, and others. A robust PTSD biomarker panel should be (i) specific to PTSD and not any of these or other comorbidities, and (ii) able to detect PTSD in both the presence and absence of these comorbid conditions. To further identify potential confounders, other samples including MDD without PTSD, diabetes with and without PTSD, and other conditions should be studied to evaluate the specificity of the panel further.

In an exploratory search of more than one million markers, we assayed a range of molecular data types including DNA methylation marks, proteins, miRNAs, and metabolites. Due to quality control and other limitations, several molecular data types were incomplete and therefore excluded from biomarker identification and refinement. These included gene expression, immune cell counts, and cytokine assays. Some of these assays were completed for the discovery cohort, and were included in early approaches for candidate biomarker selection. Any identified biomarker candidates from these assays were removed prior to down-selection and validation due to lack of data in recall and validation cohorts. The presence of these markers in the discovery phase may have influenced the selection of candidate biomarkers for some of the machine learning approaches. However, the exclusion of these datasets was not based on biomarker validation performance and therefore could not have affected the final accuracy and performance of the 28-marker panel.

In summary, we have presented a robust multi-omic panel for predicting PTSD diagnosis in male veteran populations. These 28 biomarkers include features from DNA methylation, proteins, miRNAs, metabolites, and other molecular and physiological measurements. In an independent validation cohort, we predicted PTSD diagnosis with 77% accuracy, 81% sensitivity and 73% specificity, indicating a blood-based screening or diagnostic tool is promising for identifying PTSD, particularly in males with warzone-related PTSD.
Disease Subnetwork Characterization through Mean, Variance, and Distribution Changes in Gene Expression and DNA Methylation

3.1 Background

The generation of public databases for high throughput gene expression studies, along with the curation of biological networks through protein-protein interactions and gene regulatory networks, provide an opportunity to uncover molecular and cellular changes related to biological functions for almost any disease of interest. Thousands of gene expression datasets are available via the Gene Expression Omnibus (GEO), The Cancer Genome Atlas (TCGA), and other online databases. Often these studies are used to identify individual differentially expressed genes (DEGs),

CHAPTER 3. DISEASE SUBNETWORK CHARACTERIZATION

a set of genes with differences in mean expression between the control and disease condition. These DEGs can later become candidate prognostic or diagnostic biomarkers, or can be used to search for enriched signals in biological pathways of interest.

3.1.1 Characterizing disease-related networks

Instead of searching for sets of individually significant, but unconnected, genes and relying on gene set enrichment to identify altered pathways, many network-based approaches have been implemented to identify disease-relevant subnetwork and candidate “driver genes.” These approaches can be used to identify biomarkers for diagnostics and prognostics, cluster subjects from a heterogeneous disease into subgroups, or explore disease signatures. To incorporate network topology information into a network-based framework, Chuang et al. searched for dysregulated disease subnetworks by incorporating neighbors through protein-protein interactions, and computing subnetwork statistical significance based on normalized subnetwork pathway activity. Additional algorithms for identifying connected sets of genetic mutations have shown promise in identifying genetic subtypes of disease and in generating a hierarchy of gene sets for genetic studies.

3.1.2 Subnetwork identification approaches

Primarily, gene expression analyses (including network-based approaches) use differential expression statistics to select candidate genes and subnetworks. However, changes in mean expression do not encompass all changes that may occur in a disease state, especially in cases of heterogeneous disease with multiple subtypes. In particular, Yu et al. incorporated genes with significant differences in expression variance, known as differential expression variance genes, or DEVGs, along with DEGs, into a pathway enrichment analysis framework. This approach was complementary to standard enrichment analysis, in which biological pathways are defined as gene sets, without explicit account of internal network topology. Overall, previous algorithms involve either (1) network-based biomarker identification and disease characterization using differential expression statistics, or (2) pathway enrichment analysis using gene sets (without network topology information) using hybrid differential expression and differential expression variance metrics.
3.1.3 **Statistical metrics**

Here, we implement a network-based approach to identifying and characterizing disease subnetworks using four previously-defined statistical metrics: differential expression (DEG), differential expression variance (DEVG), and two distribution-based metrics—Earth mover’s distance (EMD), and the Kolmogorov-Smirnov statistic (KS). We identified significant subnetworks using an approach adapted from Chuang et al.\textsuperscript{134} (an approach previously only incorporating differential expression), and compare the identified subnetworks across the four metrics using public expression data.

3.2 **Results**

3.2.1 **Approach**

The subnetwork identification and characterization approach consists of five phases: (1) subnetwork initialization, (2) neighbor generation, (3) subnetwork growth, (4) subnetwork termination, and (5) subnetwork characterization. Details for each phase are given in the following.

Network interaction data is gathered from the ‘CePa’ package in R. CePa contains pathway information from the Pathway Interaction Database (PID)\textsuperscript{141}, a curated database including pathways and interactions from the NCI-Nature, BioCarta, KEGG, and Reactome databases. Primarily, these databases include signaling, regulatory, and metabolic pathways. In total, interactions from 1004 biological pathways, including 13338 unique nodes, were included in the subnetwork search. To explore all candidate subnetworks independently, each node in the network was used to initialize a new subnetwork for growth. Each of these subnetworks began as a network of size one, using only the initialized starting node.

During each round of attempted subnetwork growth, the pool of neighboring nodes from the current subnetwork set were identified. Nodes connected via a single interaction were considered neighboring nodes. Each neighbor node was then in turn temporarily added to the subnetwork, in order to compute the resulting pathway activity p-value. Pathway activity vectors, $\bar{X}$ and $\bar{Y}$, were computed for each neighbor candidate by averaging the expression value of all subnetwork nodes, including the temporarily added node. $\bar{X}$ and $\bar{Y}$ designate pathway activity vectors for disease and control groups, respectively. The p-value and test statistic for the resulting pathway activity vector were computed according to the selected metric (DEG, DEVG, EMD, or KS). Differential expres-
CHAPTER 3. DISEASE SUBNETWORK CHARACTERIZATION

Expression p-values were based on moderated t-statistics using the ‘limma’ package. Statistics for differential expression variance were computed using Wilcoxon rank sum test on relative expression, following\textsuperscript{140}. Relative expression was computed as $|X - \mu_x|$ and $|Y - \mu_y|$, where $\mu_x$ and $\mu_y$ are the mean expression in disease and control classes, respectively.

The earth mover’s distance quantifies the minimum work (flow*distance) required to transform one distribution to another. Greater work indicates larger differences between gene expression distributions, while less work indicates similar distributions. Following\textsuperscript{142}, the work between two distributions $X$ and $Y$, divided into $m$ and $n$ histogram bins, with Euclidean distances $d_{ij}$ between $x_i$ and $y_j$ is:

$$\text{WORK}(X, Y, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij},$$

subject to feasibility constraints on the identified flow, $f_{ij}$:

$$f_{ij} \geq 0,$$

$$\sum_{i=1}^{m} f_{ij} \leq w_i,$$

$$\sum_{j=1}^{n} f_{ij} \leq u_j,$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_i, \sum_{j=1}^{n} u_j),$$

where $w_i$ and $u_j$ are the weights of the $i$th and $j$th histogram of $X$ and $Y$, respectively, and $x$ and $y$ are the histogram bins. EMD is then a normalized version of the minimum work,

$$\text{EMD}(X, Y) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}}.$$  

EMD computations for pathway activities are computed identically using $\bar{X}$ and $\bar{Y}$, instead of $X$ and $Y$.

The Kolmogorov-Smirnov metric quantifies differences in empirical distribution functions of $X$ and $Y$. The
largest distance between $X$ and $Y$, or $\bar{X}$ and $\bar{Y}$, is given by:

$$D = \max_{1 \leq i \leq m,n} (|E_x(i) - E_y(i)|),$$

(3.7)

where $D$ is the test statistic used to evaluate significance and $E_x$ and $E_y$ are the empirical distributions of $X$ and $Y$.

The EMD and KS metrics were computed using ‘EMDOmics’ in R. The candidate neighbor with the smallest resulting pathway activity p-value is selected and permanently added to the subnetwork, unless the new subnetwork satisfies the termination condition for termination.

Neighbor generation and subnetwork growth are repeated after the addition of each new node to the subnetwork. The algorithm terminates when the resulting p-value is greater than the previous step. The final set of identified subnetworks are further characterized based on size, significance, and overlap with other metric subnetworks.

To standardize p-value calculations across statistical metrics, nonparametric p-values were computed for each single gene and identified subnetwork. Nonparametric p-values were computed using 10,000 disease label permutations.

3.2.2 Overview of datasets

To evaluate our proposed approach, we applied this subnetwork identification algorithm to three public datasets, and a novel dataset: two cancer gene expression studies, a small gene expression study of Posttraumatic Stress Disorder (PTSD), and a novel PTSD DNA methylation dataset (Table 3.1). Gene expression and methylation data was matched with corresponding nodes in protein-protein interactions from CePa. DNA methylation data was transformed to gene-level data according the following methodology: methylation probes within 1500bp of the transcription start site (TSS), according to the manufacturer’s documentation, are averaged to generate a single overall promoter-region methylation value for each individual gene.

3.2.3 Breast Cancer subnetworks

Using a publicly available breast cancer gene expression dataset, we identified disease subnetworks associated with five-year survival. For each statistical metric (DEG, DEVG, EMD, KS), we seeded subnetwork growth individually from each node in the dataset. We were able to enrich the breast cancer signal for small p-values in subnetworks (as
CHAPTER 3. DISEASE SUBNETWORK CHARACTERIZATION

Table 3.1: Summary of datasets. Breast and brain cancer datasets from The Cancer Genome Atlas (TCGA) contained approximately 100 subjects each. Posttraumatic Stress Disorder (PTSD) data from the Gene Expression Omnibus (GEO) contained 33 total subjects.

<table>
<thead>
<tr>
<th>Disease</th>
<th>description</th>
<th>platform</th>
<th>availability</th>
<th># of probes</th>
<th># of pos. subj.</th>
<th># of neg. subj.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast Cancer</td>
<td>6-year survival</td>
<td>Agilent gene expression</td>
<td>TCGA-BRCA</td>
<td>17814</td>
<td>62</td>
<td>45</td>
</tr>
<tr>
<td>Brain Cancer</td>
<td>450-day survival</td>
<td>Agilent gene expression</td>
<td>TCGA-GBM</td>
<td>17809</td>
<td>46</td>
<td>55</td>
</tr>
<tr>
<td>PTSD1</td>
<td>PTSD and controls</td>
<td>Affymetrix gene expression</td>
<td>GEO, GSE860</td>
<td>12600</td>
<td>17</td>
<td>16</td>
</tr>
<tr>
<td>PTSD2</td>
<td>PTSD and controls</td>
<td>Illumina 450k Methylation</td>
<td>TCGA-BRCA</td>
<td>18187</td>
<td>80</td>
<td>82</td>
</tr>
</tbody>
</table>

compared to individual genes/nodes). In particular, subnetworks selected based on DEG, DEVG, and KS metrics showed the strongest increase in signal (Fig. 3.1A). These enriched p-values may lead to more reproducible findings and more accurate classification performance, as has been shown for differential expression in\(^\text{134}\). While the largest overlap of identified subnetworks were common among all four metrics (325 starting nodes, 15% of all significant networks), a large fraction of subnetworks were distinguished uniquely by a single metric (Fig. 3.1B). Evaluating subnetworks based on DEVG identified the largest number of metric-specific subnetworks (221 versus 100, 12, and 135 for DEG, EMD, and KS, respectively). Identified subnetworks ranged in size from 1-25 nodes, with significant subnetworks (p < 0.05), originating from 2148 starting nodes (Fig. 3.1C-F). Overall, we note that using differential expression variance identified the greatest number of subnetworks, largest size subnetworks, and most unique subnetworks in this breast cancer study.

3.2.4 Brain Cancer subnetworks

Analogous to breast cancer, we implemented the proposed algorithm in a brain cancer gene expression data from TCGA\(^\text{94}\). The gene expression dataset consisted of 46 patients surviving 450 days, and 55 patients who did not. Similar results were achieved, including: (1) enriched signal for identified subnetworks over single gene p-values, (2) a large number of unique subnetworks identified based on DEVG (107 networks, versus 68, 11, and 81 identified only in analyses using DEG, EMD, and KS, respectively), and (3) the largest size subnetworks were identified based on differential expression variance (Fig. 3.2).
Figure 3.1: Summary of breast cancer subnetworks. (A) Distribution of individual gene and subnetwork p-values using differential expression (DEG), differential expression variance (DEVG), Earth Mover’s Distance (EMD), and Kolmogorov-Smirnov (KS) metrics. (B) Overlap of significant subnetworks (p < 0.05) from the same starting node. (C)-(F) Histogram of significant subnetwork sizes for DEG (C), DEVG (D), EMD (E), and KS (F) metrics.

3.2.5 Posttraumatic Stress Disorder subnetworks – gene expression

In order to evaluate the feasibility of this approach for many applications and datasets, we also selected a small gene expression dataset from emergency room patients with and without Posttraumatic Stress Disorder (PTSD). This dataset consisted of a similar high throughput gene expression microarray study, with less than half of the total number of subjects from the breast cancer implementation. Again, we were able to significantly enrich the dataset for small p-values in identified subnetworks (Fig. 3.3A). Following the cancer findings, a large fraction of subnetworks were identified from the same starting node using all four metrics (552 subnetworks, 30% of total subnetworks). Similarly, DEVG identified 107 unique subnetworks, the largest number of unique networks across all four metrics (Fig. 3.3B). Larger networks were again from DEVGs, with 105 subnetworks with at least 10 nodes
CHAPTER 3. DISEASE SUBNETWORK CHARACTERIZATION

Figure 3.2: Summary of brain cancer subnetworks. (A) Distribution of individual gene and subnetwork p-values using differential expression (DEG), differential expression variance (DEVG), Earth Mover’s Distance (EMD), and Kolmogorov-Smirnov (KS) metrics. (B) Overlap of significant subnetworks (p < 0.05) from the same starting node. (C)-(F) Histogram of significant subnetwork sizes for DEG (C), DEVG (D), EMD (E), and KS (F) metrics.

(Fig. 3.3C-F). Other metrics each identified fewer than 30 subnetworks of at least that size.

3.2.6 Posttraumatic Stress Disorder 2 subnetworks – DNA Methylation

In addition to high throughput studies of gene expression, genome-wide DNA methylation patterns can contain signals of biological dysfunction in the presence of disease. Changes in methylation profiles of related gene sets may indicate changes in epigenetic regulation of gene expression. These methylation changes may occur due to early life stress\textsuperscript{18}, aging\textsuperscript{144}, diet—including the effects of a mother’s pre-conception diet on offspring methylation\textsuperscript{145} and the effects of an offspring’s diet on its own methylation\textsuperscript{146}—and other environmental factors. In the analyzed PTSD methylation dataset, similar results to the gene expression studies were observed. Subnetwork p-values were
enriched in more significant p-values compared to single gene statistics, though not as strongly as the three gene expression datasets (Fig. 3.4A). Similarly, the largest number of subnetworks were identified by all four metrics, and the differential expression variance identified the most unique subnetworks (Fig. 3.4B). Additionally, as seen in the gene expression results, large subnetworks (size>10) were primarily identified by differential expression variance (Fig. 3.4C-F).

3.2.7 Biological network properties

To understand how biological network structure is related to disease subnetworks, we used graph theory tools to compare network properties of the overall biological network search space and the identified subnetworks from
Figure 3.4: Summary of PTSD DNA methylation subnetworks. (A) Distribution of individual gene and subnetwork p-values using differential expression (DEG), differential expression variance (DEVG), Earth Mover’s Distance (EMD), and Kolmogorov-Smirnov (KS) metrics. (B) Overlap of significant subnetworks (p < 0.05) from the same starting node. (C)-(F) Histogram of significant subnetwork sizes for DEG (C), DEVG (D), EMD (E), and KS (F) metrics.

each of the four statistical approaches (Table 3.2). Overall network properties using only nodes with available gene expression or methylation data were computed for comparison with identified subnetworks. The average clustering coefficient remained fairly constant across all identified subnetworks in all datasets, although in brain cancer identified subnetworks had greater clustering coefficients than the full network search space. Identified subnetworks were all greater density than the full search space. Low density biological networks are thought to be an evolutionary advantage due to greater robustness\textsuperscript{147}. Disease perturbations may be more likely to persist over long time periods in these high density network regions. Additionally, we saw slight increases in subnetwork centralization for some subnetworks. Larger centralization values indicate a more “hub” or “star-like” topology, while lower centralization values indicates all network nodes have a similar node degree. Previous studies have suggested...
Table 3.2: Average clustering coefficient, network density, and centralization for subnetworks from each statistical metric. Network properties for overall network search space are shown for comparison.

<table>
<thead>
<tr>
<th>Network property</th>
<th>Breast Cancer</th>
<th>Brain Cancer</th>
<th>PTSD1</th>
<th>PTSD2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full networks</td>
<td>DEG</td>
<td>DEVG</td>
<td>EMD</td>
</tr>
<tr>
<td>clustering coeff.</td>
<td>0.14</td>
<td>0.12</td>
<td>0.15</td>
<td>0.07</td>
</tr>
<tr>
<td>density</td>
<td>0.06</td>
<td>0.34</td>
<td>0.57</td>
<td>0.61</td>
</tr>
<tr>
<td>centralization</td>
<td>0.48</td>
<td>0.52</td>
<td>0.57</td>
<td>0.75</td>
</tr>
</tbody>
</table>

Table 3.3: Average classification performance from gradient boosting with decision trees, based on mean (±s.d) Area Under the Receiver-Operator Characteristic Curve (AUC).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Individual gene AUCs</th>
<th>Random Subnetwork AUCs</th>
<th>Subnetwork AUCs</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRCA</td>
<td>0.64±0.03</td>
<td>0.58±0.12</td>
<td>0.68±0.05</td>
<td>0.66±0.03</td>
</tr>
<tr>
<td>GBM</td>
<td>0.65±0.04</td>
<td>0.69±0.12</td>
<td>0.69±0.05</td>
<td>0.70±0.05</td>
</tr>
<tr>
<td>PTSD1</td>
<td>0.75±0.05</td>
<td>0.75±0.18</td>
<td>0.81±0.06</td>
<td>0.79±0.06</td>
</tr>
<tr>
<td>PTSD2</td>
<td>0.64±0.09</td>
<td>0.58±0.08</td>
<td>0.67±0.09</td>
<td>0.64±0.09</td>
</tr>
</tbody>
</table>

that disease genes are more likely to be the “hub” node in these topologies\cite{148}. Multiple of these topologies exist in the identified disease subnetworks, including around NFATC1, RAC1, and SHC1 in breast cancer (Fig. 3.5), and around FURIN, RHOA, and MYC in PTSD (Fig. 3.8).

3.2.8 Subnetwork classification performance

Network-based biomarkers have been shown to be more robust than single gene markers\cite{134}. We evaluated biomarker performance for identified subnetworks from each statistical metric, as well as individual gene markers, and randomly generated subnetworks. To compare subnetwork biomarker performance, we evaluated three sets of biomarker types: (1) individual gene markers, using only genes with differential expression $p<0.05$, (2) randomly generated subnetworks, using genes with differential expression $p<0.05$, and (3) subnetworks identified using each of the four statistical metrics. Using ‘CMA’ in R, we evaluated each set of markers over 1000 iterations of 5-fold cross-validation using gradient boosting with decision trees for classification. In each round of cross-validation, param-
Figure 3.5: Largest identified breast cancer subnetworks from each metric, including overlapping nodes from differential expression (DEG), differential expression variance (DEVG), and Earth Mover’s Distance (EMD).

Parameters were tuned using 3-fold cross-validation in only the training portion of the dataset (using ‘tune’ with the default parameter search grid in CMA). Area Under the Receiver-Operator Curve (AUC) was computed in each round based on predicted probabilities in the testing portion of the dataset. Randomly generated subnetwork pathway activities were computed by averaging a random sampling of genes with \( p < 0.05 \). The number of randomly sampled genes was drawn from the distribution of subnetwork sizes in the combined set of subnetworks from all four metrics. Using the proposed algorithm, we observed average AUC improvements of 0.02-0.09 for individual subnetworks over single gene predictors, and even greater improvements over subnetworks randomly generated from top DEGs (Table 3.3). Further tuning, larger sample sizes, and additional network information may lead to additional improvements in AUCs, providing better tools for disease diagnostics and prognostics.
CHAPTER 3. DISEASE SUBNETWORK CHARACTERIZATION

Figure 3.6: Largest identified brain cancer subnetworks from each metric do not include overlapping nodes from differential expression (DEG), differential expression variance (DEVG), Earth Mover’s Distance (EMD), or Kolmogorov-Smirnov (KS) approaches.

3.2.9 DISEASE NETWORK BIOLOGY

On average, each identified subnetwork covered approximately 5% of the total connected network search space, indicating that relevant disease subnetworks make up only small parts of larger biological functions. In differential expression subnetworks, selected nodes typically showed same-direction disease fold changes, resulting in improved disease signals over single genes. Specifically, in 85% of breast cancer subnetworks and 84% of brain cancer subnetworks, all included nodes show same-direction fold change. In the PTSD gene expression dataset, 90% of subnetworks included identical fold change directions, indicating the robustness of the method even in smaller and lower datasets. Less consistency was observed in differential expression variance subnetworks. Unlike differential expression, only 50-60% of subnetworks identified based on DEVGs showed unanimous disease changes in expression variance. This inconsistency may be due to noise, disease subtypes, or data normalization procedures.
CHAPTER 3. DISEASE SUBNETWORK CHARACTERIZATION

Figure 3.7: Breast cancer gene related subnetworks. Subnetworks involving BRCA1 gene were identified using differential expression variance (DEVG) and Kolmogorov-Smirnov (KS) metrics, and included three overlapping nodes (BRCA1, ATM, and BARD1). No subnetworks identified using differential expression (DEG) and Earth Mover’s Distance (EMD) included BRCA1 or BRCA2.

In addition to validating the ability of the proposed algorithm to identify statistically significant disease subnetwork signals, we explored the agreement between the identified subnetworks and previously identified disease biology. The largest identified breast cancer subnetworks from DEG and EMD analyses contained multiple overlapping genes involved in the JAK-STAT pathway, including JAK2, STAT3, STAT5A, ERBB4 and others (Fig. 3.5). Additionally, through the connection in SDC2, the DEVG subnetwork overlaps with the DEG subnetworks. In addition to SDC2, the largest DEVG subnetwork includes EPHA2, SHC1, and RAC1, genes previously identified as breast cancer related signals.

Some of the most commonly studied signals of breast cancer are mutations, copy number alteration and gene expression changes in BRCA1 or BRCA2. Only DEVG and KS approaches identified significant subnetworks containing either BRCA1 or BRCA2. Both identified networks included expression changes in BRCA1, BARD1, and ATM. The DEVG subnetwork also included signals from CHEK1 and other connected genes, while the KS subnetwork included expression changes in CHEK2 (Fig. 3.7). Increased CHEK1 expression has been associated with poor prognosis in breast cancer. Here, we observed no significant mean difference in CHEK1 expression.
Figure 3.8: Largest identified PTSD subnetworks from each metric do not include overlapping nodes from differential expression (DEG), differential expression variance (DEVG), Earth Mover’s Distance (EMD), or Kolmogorov-Smirnov (KS) approaches. DEVG identified the largest PTSD subnetwork, involving MAPK signaling pathways, and cell cycle and division regulation.

level (p > 0.05). Instead, we identified network-based signatures of breast cancer involving CHEK1 using DEVG metrics, indicating expression variance changes may be associated with breast cancer survival, not differences in mean. Unlike breast cancer, the largest identified brain cancer and PTSD gene expression subnetworks from each statistical metric were disconnected from each other (Figs. 3.6 and 3.8). The largest brain cancer subnetworks included signals related to CASP genes, previously implicated by\textsuperscript{150}, as well as alterations in SMAD expression, an identified signal of brain cancer prognosis\textsuperscript{151}. In PTSD, DEVG statistics uncovered subnetworks related to MAPK signaling, and cell cycle and division regulation, while the differential expression subnetwork included multiple coagulation factors, including F2 and F10 (Fig. 3.8). Blood coagulation differences have been seen in PTSD population previously\textsuperscript{152,153}. Unlike the PTSD gene expression networks, the largest PTSD subnetworks identified from DNA methylation data included overlapping nodes in DEG, DEVG, and KS networks (Fig. 3.9). However,
overlapping biological functions appeared between PTSD gene expression and DNA methylation subnetworks, including MAPK signaling. In addition to corroborating previously identified cancer and PTSD signals, many identified subnetworks include novel candidate signals. These additional signals may improve understanding of disease biology, identification of disease subtypes, and biomarker performance.

### 3.3 Discussion

#### 3.3.1 Limitations

The proposed approach identified many novel disease-relevant subnetworks, connected through known protein-protein and gene regulatory interactions. Computed subnetwork scores are based on average gene expression over all subnetwork nodes. Using this approach, the algorithm poorly captures inhibitory relationships, due to signal
loss from averaging expression values. Larger subnetworks may be identified by incorporating both positive and negative interaction and should be explored in future implementations. Additionally, all approaches implemented here involved only a single exploratory gene expression dataset used for (1) subnetwork identification and growth, (2) significance computations, and (3) model tuning and biomarker performance evaluation. To independently evaluate the improvements and robustness of identified subnetworks, and to prevent over-fitting and AUC inflation, additional datasets are needed.

3.3.2 Conclusions

To identify and characterize disease signals other than purely changes in mean expression, we implemented a disease subnetwork identification approach using four statistical metrics characterizing gene expression changes in mean, variance, and overall expression. In all four explored datasets, differential expression variance accounted for the largest number of unique subnetworks, as well as the largest size subnetworks. Subnetworks included connections to known biology (including BRCA1 in breast cancer and blood coagulation in PTSD), as well as many unique disease subnetworks requiring further evaluation. Network-based biomarkers improved predictive performance for all statistical metrics, indicating greater diagnostic or prognostic power. Extensions of this work could incorporate additional data types (e.g. proteins, copy number) to further elucidate network-level disease alterations.
Advancing biomarker identification approaches: integrating incomplete multi-omic data and optimizing sensitivity vs specificity tradeoff

To improve performance of multi-omic biomarkers—like those discussed in Chapter 2—novel methods and approaches are needed to integrate many biological datasets of various sizes, characteristics, and strengths. In this chapter, we present strategies for tackling two challenges in multi-omic integration and biomarker use in clinical settings: (1) handling incomplete datasets in multi-omic, integrated studies, and (2) identifying and developing biomarker panels in clinical situations in which false positive and false negative predictions have unequal consequences.
Figure 4.1: Classification performance vs training sample size in four publicly available cancer datasets. Average classification performance increases with increasing training sample size (except in low signal datasets). Error bars indicate standard deviation over 50 iterations of classification, with 30% of overall datasets held out for testing in each iteration. Classification performance is computed over 100 rounds of model training and prediction, using various training datasets sizes. Test dataset was identical in all scenarios. (A) Predicting 2190 day survival using protein data from TCGA breast cancer dataset. (B) Predicting 450 day survival using miRNA data from TCGA brain cancer dataset (GBM). (C) Predicting 400 day survival using mRNA data from TCGA leukemia datasets (AML). (D) Predicting squamous versus ductal cell carcinoma from copy number data in CCLE dataset.

4.1 Incorporating incomplete datasets

Biomarker identification and machine learning algorithms results typically improve with increasing training sample size. For example, increasing the training sample size of a high throughput dataset improves classification performance (based on AUC) in a held-out test dataset. Using four publicly available cancer datasets of various molecular data types, we demonstrated this improvement by varying the training dataset size, while holding test dataset size constant (Fig. 4.1). The four public datasets included: (1) TCGA breast cancer (BRCA) proteomics data for predicting 2190 day survival, (2) TCGA brain cancer (GBM) miRNA data for predicting 450 day survival, (3) TCGA leukemia (AML) mRNA data for predicting 400 day survival, and (4) Cancer Cell Line Encyclopedia (CCLE) carcinoma copy number variation data for predicting squamous versus ductal cell carcinoma. The four datasets also varied in size, with the number of training data subjects ranging from approximately 50-350. In three of the
Figure 4.2: Classification performance vs training sample size in PTSD datasets. Average classification performance slightly increases with increasing training sample size. Error bars indicate standard deviation over 50 iterations of classification, with 30% of overall datasets held out for testing in each iteration. (A) Predicting PTSD status using DNA methylation data from Illumina’s Infinium Human Methylation 450k array. (B) Predicting PTSD status using high throughput metabolomics data (250 metabolites). (C) Predicting PTSD status using routine clinical lab values (50 markers). (D) Predicting PTSD status using protein/peptide markers from Selected Reaction Monitoring (SRM).

In multi-omic studies, methods for integration of multiple molecular datasets often require each sample to have complete data available for all data types. Due to poor sample quality, not enough total sample volume, limitation in cost, or other restrictions, often some samples are excluded from each molecular assay. During integration each sample excluded from any one particular assay must now be excluded from all integrative analyses, resulting in a significant decrease in sample size. For example, Table 4.1 shows the sizes for various multi-omics datasets collected...
CHAPTER 4. DEVELOPING MULTI-OMIC APPROACHES

**Table 4.1**: Summary of available clinical and molecular data for PTSD cohort, including number of subjects with complete data available.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>PTSD cases</th>
<th>Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>83</td>
<td>83</td>
</tr>
<tr>
<td>Clinical Lab</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>Endocrine</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>GWAS</td>
<td>81</td>
<td>81</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>Methylation (array)</td>
<td>80</td>
<td>82</td>
</tr>
<tr>
<td>Methylation (targeted sequencing)</td>
<td>79</td>
<td>76</td>
</tr>
<tr>
<td>miRNA</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Physiological</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>Protein (SRM)</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td>Protein (ELISA)</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>Small Molecules</td>
<td>83</td>
<td>82</td>
</tr>
<tr>
<td>Complete Data</td>
<td>76</td>
<td>73</td>
</tr>
</tbody>
</table>

in the discovery cohort of 83 PTSD cases and 83 age and ethnicity-matched controls. While each individual datasets ranges in size from 155 to 166 total subjects, the number of subjects with complete data available in only 149, a loss of 17 subjects (~10% of the cohort) with partial data available.

While the decrease in sample size due to incomplete multi-omic data may result in decreased performance, integrating multiple levels of molecular data may provide improvements in PTSD classification performance. Multiple studies have explored the performance improvements from multi-omic integration, with most finding prediction improvements from integrated approaches.\textsuperscript{154,155,156}

Due to the occasional missing samples for single data types and large difference in number of features across the available datasets (44-450k), combining all datasets would result in both a reduced sample size, and heavy weighting towards larger high-throughput features. Instead, we propose to integrate at the final prediction step (Fig. 4.3). Each single-omic dataset is analyzed independently, using a selected feature selection and classifier training methodology. Using the trained single-omic classifiers, disease status prediction will then be attempted for each subject using each of the available datasets, and predicted disease status probabilities will be recorded for each molecular data type. From all available predicted probabilities, the most confident prediction will be used for final prediction.

We implemented this methodology in the PTSD datasets and multiple public cancer datasets. Individual ‘omic predictions were generated for over 500 rounds of three-fold cross-validation. Within each round of cross-validation, model parameters were tuned with internal nested cross-validation (three-fold). Feature selection was performed
only in the training subset of each round, selecting the top 10 differentially expressed genes. Cross-validation, feature selection, parameter tuning, model generation, and prediction were done using the ‘CMA’ package in R.

4.1.1 Results

With application to PTSD, this approach resulted in a slightly better AUC of 0.68, compared to a maximum single-omic AUC of 0.67 (Figure 4.4A). Additionally, the data types with the highest performance (Clinical Labs and Metabolomics) were used most frequently to make prediction (22 and 45%, respectively). Further, many subjects appeared to have a specific data type with the strongest signal, with more than half of the subjects in the PTSD datasets using the same data type for prediction over 50% of the time. Specifically, 96 subjects used metabolomics data most often for prediction, while 36 subjects used Clinical Labs, 16 subjects each used proteomics and physiological data, and one subject used endocrine data most often for prediction. The incorrect predictions were concentrated within a consistent group of subjects, across all rounds of cross-validation, indicating the presence of subgroups or potential co-morbid diseases may alter the PTSD signal in specific subpopulations (Figure 4.4C). Increasing the total sample size may allow for the identification of more robust PTSD signals, which are unaffected by subgroup status or co-morbid diseases, or for the identification subtype-specific biomarker panels. Additionally, the highest performing individual data types showed the widest distribution in predicted PTSD probability, indicating that the most confident predictions consistently came from the best performing single-omic data type (Figure 4.4D).
Figure 4.4: Summary of PTSD multi-omic classification results. (A) Average performance of single-omic and integrated classification. Average AUC is indicated by height of bars. Integrated biomarker panels performs only slightly better than the best single-omic datasets (0.67, Clinical Labs vs 0.68, integrated panel). (B) Fraction of each datasets used for final maximum value prediction. (C) Distribution of subject-specific misclassification rates over 100 rounds of cross-validation, separated by data type. (D) Predicted probability distribution from cross-validation, by data type.

While performance improvements were small in the PTSD datasets, we observed greater improvements in AUC in larger, public cancer datasets. A summary of the available cancer data used is shown in Table 4.2. Integrating DNA methylation and gene expression data in the TCGA AML Leukemia dataset, resulted in an AUC improvement from 0.60 to 0.63 (Table 4.3). From these two molecular datasets, only 90% of the subjects had completed
CHAPTER 4. DEVELOPING MULTIOMIC APPROACHES

DNA methylation and gene expression data. By integrating the incomplete datasets, an additional 12 participants were included in model training. Using TCGA’s BRCA breast cancer dataset, the integrated multi-omic strategy resulted in a small AUC improvement (integrated AUC: 0.981, maximum single omic AUC: 0.977, Table 4.3). The minor improvements in AUC in a small PTSD multi-omic cohort, as well as larger improvements in public cancer datasets, suggest that this strategy may provide a solution for integrating incomplete subjects into a multi-omic classification framework.

Table 4.2: Size of public multi-omic datasets. For each molecular data type, the number of subjects with available data (positive/negative), as well as the number of subjects with complete data (positive/negative).

<table>
<thead>
<tr>
<th>Disease</th>
<th>Data Type 1</th>
<th>Data Type 2</th>
<th>Data Type 3</th>
<th>Complete Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA AML Leukemia</td>
<td>DNA methylation (65/55)</td>
<td>gene expression (69/49)</td>
<td>-</td>
<td>64/49</td>
</tr>
<tr>
<td>TCGA BRCA Breast Cancer</td>
<td>DNA methylation (238/69)</td>
<td>gene expression (403/117)</td>
<td>protein (638/198)</td>
<td>181/51</td>
</tr>
</tbody>
</table>

Table 4.3: Summary of public multi-omic classification performance.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Data Type 1</th>
<th>Data Type 2</th>
<th>Data Type 3</th>
<th>Integrated</th>
</tr>
</thead>
<tbody>
<tr>
<td>TCGA AML Leukemia</td>
<td>0.604</td>
<td>0.601</td>
<td>-</td>
<td>0.627</td>
</tr>
<tr>
<td>TCGA BRCA Breast Cancer</td>
<td>0.942</td>
<td>0.977</td>
<td>0.965</td>
<td>0.981</td>
</tr>
</tbody>
</table>

4.2 A biomarker identification approach for optimizing sensitivity or specificity

Often, evaluation of clinical predictions utilizes multiple metrics, including AUC, error rate, false positive rate, false negative rate, and others. Based on clinical considerations, a diagnostic biomarker test may be most useful if diagnostic errors are mostly False Positives (or False Negatives). For example, a biomarker prediction used to screen patients for further evaluation or follow-up may be able to tolerate many False Positives, that can be easily identified during more extensive evaluations, but should not miss patients who need to be evaluated further (False Negatives). A test with a low False Negative Rate, is a high sensitivity test. Conversely, a highly specific test (low False Positive Rate), can be used to rule out the presence of disease in an individual or population.

Multiple approaches have been proposed to control or balance the sensitivity or specificity of a classifier. These methods primarily fall into two categories: (1) under or oversampling training data for classifiers trained on unbalanced classes, or (2) class-weighted classifiers to penalize incorrect predictions from each class separately. Briefly, under and oversampling are used to generate balanced classes in the training dataset by oversampling the minority class, or undersampling the majority class. This under or oversampling can occur randomly, can be done by cre-
ating synthetic data, or can be done by using supervised learning to select the most informative data to use. An overview of over and undersampling techniques can be found in\(^\text{157}\).

Class-weighted techniques, including class-weighted SVM, are often used to re-define a decision boundary in scenarios with highly imbalanced classes. In traditional linear SVM, the regularization parameter \(C\) in the objective function is defined or tuned by the user. The SVM objective function is:

\[
\text{minimize} \left( \frac{||w||^2}{2} + C \sum_{i=1}^{n} \xi_i \right), \text{ subject to } \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \forall x_i \text{ in training data, and} \quad 0 \leq \xi_i \leq 1
\]  

(4.1)

(4.2)

(4.3)

where \(y_i\) is the response variable, \(x_i\) is the predictive feature set, \(\xi_i\) is a slack variable for linearly inseparable problems, and \(w\) is the width of the margin. To separately penalize misclassifications in each class, class-weighted SVM separates \(C\) into \(C_1\) and \(C_2\), creating a modified objective function,

\[
\text{minimize} \left( \frac{||w||^2}{2} + C_1 \sum_{i=1}^{n_1} \xi_i + C_2 \sum_{j=1}^{n_2} \xi_j \right), \text{ subject to } \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \forall x_i \text{ in training data, and} \quad 0 \leq \xi_i \leq 1
\]

(4.4)

(4.5)

where \(C_1\) and \(C_2\) can independently weight errors in class 1 and class 2. Similar approaches have been implemented for other classifiers, including Random Forest.

Using over or undersampling, or class-weighting, often does not directly impact feature selection. Instead, decision boundaries are re-computed using the previously identified features, to tune the sensitivity and specificity of the model. However, certain features may provide greater ability to tune this trade-off, in both balanced and unbalanced class datasets. With this in mind, we have developed a feature selection and classification scheme to carefully tune false positive and false negative errors. Instead of seeking to maximize accuracy alone, additional metrics can be prioritized during feature selection and classification. For simplicity, we focused on two metrics: sensitivity and specificity. The approach incorporates a feature elimination strategy that is metric-specific and can be combined with any classification strategy (SVM, Linear Discriminant Analysis, Nearest Shrunken Centroids, etc). We have
applied this methodology to public data as well as PTSD datasets to demonstrate the improvements in False Positive or False Negative Errors. The proposed feature selection strategy identifies features with differences in variance. The notion of disease group differences in variance has been previously proposed as Differential Expression Variance Genes (DEVGs)\textsuperscript{140}, and previously discussed in Chapter 3. An illustration of DEVG scenarios is show in Fig. 4.5.

In our proposed methodology, we have incorporated this idea of DEVGs, along with traditional Differentially Expressed Genes (DEGs) to identify a robust set of features for specificity and sensitivity-specific classification. We used the following equations to rank all features for feature selection:

\[
\frac{sd_{\text{control}}}{sd_{\text{PTSD}}} e^{-p}, \tag{4.6}
\]

\[
\frac{sd_{\text{PTSD}}}{sd_{\text{control}}} e^{-p}, \tag{4.7}
\]

where \( p \) is the differential expression p-value, and \( sd \) is the standard deviation of all PTSD or control samples for the gene of interest. Equation 4.6 is used to maximize the True Positive Rate (TPR), while Equation 4.7 is used to maximize the True Negative Rate (TNR). We used the proposed feature selection algorithm along with the Nearest Shrunken Centroids (NSC) classifier to improve sensitivity and specificity. The shrinkage parameter of the NSC classifier was tuned with three-fold nested cross-validation.
CHAPTER 4. DEVELOPING MULTI-OMIC APPROACHES

Figure 4.6: Example ROC curve, with candidate operating points. Black points indicate candidate points: points with minimum error. In a maximum sensitivity scenario, point A would be selected. In a maximum specificity scenario, point B would be chosen.

After feature selection and parameter tuning, the ROC curve was generated for each round of cross-validation. The points on the operating curve with maximum accuracy were selected as candidate operating points. When maximizing sensitivity, the maximum sensitivity candidate operating point was chosen. When maximizing specificity, the maximum specificity candidate operating point was chosen. An example ROC curve is shown in Fig. 4.6.

We compared the performance of a traditional classification strategy (using t-test p-values for feature selection and minimizing the error rate during centroid shrinking), with the proposed methodology using 100 rounds of 5-fold cross-validation. For comparison, we used the same strategy to select operating points on the ROC curve, identifying the maximum sensitivity and specificity, evaluating only the improvements in performance metrics due to the new feature ranking/selection algorithm.

We applied this methodology to multiple molecular PTSD datasets, including metabolomics, proteins, and miRNA from the 83-83 discovery cohort, as well as two public datasets (TCGA AML Leukemia DNA methylation and CCLE gene expression). The cross-validated results for the standard algorithm, and proposed sensitivity and specificity-specific algorithms are shown in Fig. 4.7. In the AML, CCLE, and PTSD protein and metabolomics datasets, the algorithm showed improvements in sensitivity and specificity using the TPR-maximization and TNR-maximization algorithms, respectively. However, in the PTSD miRNA data, our proposed algorithm resulted in only minor improvements in both sensitivity and specificity. In all datasets, the overall algorithm accuracy was
CHAPTER 4. DEVELOPING MULTI-OMIC APPROACHES

Figure 4.7: Sensitivity and specificity optimization with PTSD and public cancer datasets. (A) Average sensitivity indicated by height of bars for accuracy maximization and sensitivity maximization scheme, for each data type. (B) Average specificity indicated by height of bars for accuracy maximization and specificity maximization scheme, for each data type.

not significantly different between the maximum accuracy, sensitivity, and specificity implementations. We expect additional improvements can be seen with the incorporation of a TPR or TNR-specific decision boundary, or a more sophisticated classifier.

4.3 Discussion

Both multi-omic classification methods presented in this chapter showed moderate improvements in classification performance, and provide candidate solutions to overcoming challenges involved in multi-omic classification. By providing a solution for integrating incomplete subjects, the amount of usable data in large multi-omic studies will increase. In our PTSD discovery cohort, integrating all incomplete subjects resulted in a 2% larger training dataset, resulting in a 1% improvement in AUC. The probability distributions were widest for higher performance data types, and narrower for lower performance data types, indicating that within cross-validation of each data type, a tree-based classifier was better able to generate accurate predictions based on the signal strength of the dataset.
CHAPTER 4. DEVELOPING MULTI-OMIC APPROACHES

Since each molecular data type follows individual pre-processing and normalization steps, additional work should focus on understanding how these methods affect the distribution of predicted probabilities, particularly when comparing across data types.

Additionally, while this incomplete data integration strategy allows for the inclusion of additional data, the final prediction requires all available molecular data for each participant be used to generate a set of predicted probabilities. This strategy would require multiple molecular assays per person, in order to be implemented in a clinical setting. In independent validation, if only minor improvements are achieved, the extensive cost to generate multi-omic data for this algorithm may not be worthwhile.

In our sensitivity and specificity optimization approach, we observed improvements in the metric of interest, without overall loss of accuracy (or AUC). This improvement in sensitivity or specificity was achieved using a novel feature selection approach and providing the ability to choose alternate operating points on the ROC curve. However, in each of the datasets we explored, the positive and negatives classes were balanced. This allowed us to use accuracy to optimize classifier parameters, and select candidate operating points on the ROC curve for all implementations of the approach. In datasets with unbalanced classes, other metrics should be used for these steps, in order to weight predictions of each class individually.

Overall, in this chapter we presented two novel approaches to biomarker identification and classification. First, we developed an integrated multi-omic approach that can leverage signal contained in subjects with incomplete data, and demonstrated improvement in AUC using this maximum confidence approach. This approach can provide great benefits to large, collaborative studies with frequently incomplete data records. Second, we demonstrated how differential expression variance genes can be used to tune the prediction errors to favor one class or the other. This methodology can identify predictive biomarkers for scenarios where the test sensitivity and specificity are not valued equally.
Conclusions and Future Directions

Thus far, we have focused our efforts on PTSD biomarker identification and characterization, considering PTSD to consist of a single homogeneous disease. While we have incorporated methodologies that may lend themselves to capturing heterogeneity or disease progression (e.g. using differential expression variance), we have not focused on identifying or characterizing these PTSD subgroups for use in diagnostics or prognostics. In this chapter, we present some preliminary findings and outline some next steps for investigating the existence of these molecular or clinical subgroups.

5.1 Background

Previous literature has suggested the existence of PTSD subtypes based on clinical observations, therapeutic and behavioral treatment outcomes, hypothesis-driven molecular markers, and other strategies. These subtypes may inform diagnosis, prognosis or recommended treatment strategies, or even risk for common PTSD comorbidities. In the future, molecular and clinical data could be used to generate data-driven subgroups of PTSD cases. These
subgroups can be further characterized based on molecular signatures, clinical phenotypes, and demographics. Using the independent validation cohort, we can compare similarly defined subgroups for consistency. Finally, we propose using an approach for ordering time series biological data into trees to explore disease progression/stage, and distinct disease subtypes.

5.1.1 Heterogeneity of PTSD

Previous studies of PTSD subtypes have been based on distinct symptom trajectories over time\textsuperscript{80}, correlation between treatment response to medication and promoter region methylation of glucocorticoid related genes\textsuperscript{81}, as well as clinical and neurobiological signals related to the dissociative subtype from the DSM-V\textsuperscript{158}. Details regarding these candidate subtypes are detailed next.

**PTSD symptom trajectories**

In a study of emergency room participants, PTSD symptoms were evaluated using the PTSD Symptom Scale (PSS) at 10 days, one month, five months, seven months, and 15 months following exposure to a potentially traumatic event (PTE)\textsuperscript{80}. Using latent growth mixture modeling, and multiple model criteria, three distinct symptom trajectories were selected. These trajectories were referred to as Rapid Remitting (rapid decrease in symptoms from one to five months), Slow Remitting (a slow, linear decrease in symptoms from one to 15 months) and Non-Remitting (no improvement in symptoms over 15 months). Participants were randomly assigned to one of multiple treatment conditions, including multiple types of cognitive behavioral therapy (CBT) and a double-blind SSRI/placebo treatment. Only participants in the slow remitting group showed faster symptom decline from CBT. CBT in participants in the rapid remitting and non-remitting classes had non-significant treatment effects.

**FKBP5 and NR3C1 methylation and treatment response**

In a military veteran study of prolonged exposure (PE) therapy, —a form of CBT— promoter region methylation of the FKBP5 and NR3C1 gene were quantified to compare treatment responders and non-responders\textsuperscript{81}. Promoter region methylation of the NR3C1 gene (the gene encoding the glucocorticoid receptor) prior to treatment, was significantly different between treatment responders and non-responders, indicating a potential treatment response predictor. Contrastly, while promoter region methylation of FKBP5 did not contain significant group
differences pre-treatment, the methylation patterns showed changes at the follow-up time point, with treatment responders showing decreased methylation over time, while treatment non-responders showed increased methylation over time.

Dissociative subtype of PTSD

Based on clinical observations, the most recent version of the DSM (DSM-V), includes a dissociative subtypes of PTSD\textsuperscript{5}. The dissociative subtype of PTSD includes patients with symptoms of depersonalization and derealization. Lanius et al. reviewed evidence of the dissociative subtype, including differences in emotional modulation and brain activity\textsuperscript{158}.

5.1.2 Multi-omic clustering

Integrating multiple molecular datasets may provide novel insights into relevant disease subtypes. These subtypes may exist due to inherited genetic risk, environmental changes in DNA methylation, manifestations of clinical symptoms, molecular markers of co-morbid conditions, or a combination of these factors. Previously, PTSD subtype identification and characterization has focused on only a single data type contribution, with primary identification being either clinical symptom-based\textsuperscript{158,80}, or treatment response-based\textsuperscript{81}. These approaches have explored candidate signal differences between the clinically identified subtypes, including brain activity\textsuperscript{158}, methylation patterns\textsuperscript{81}, and endocrine markers\textsuperscript{81}. In contrast, we have implemented a purely data-driven approaches to PTSD subtype identification, integrating both clinical and molecular data to learn the strongest disease clusters.

Multi-omic clustering has previously been applied to other heterogeneous diseases—primarily cancers—including lung cancer\textsuperscript{159}, breast cancer\textsuperscript{160}, ovarian cancer\textsuperscript{92}, prostate cancer\textsuperscript{161}, and others. Methods for both multi-omic integration and clustering include MVDA\textsuperscript{161}, iCluster\textsuperscript{159}, multi-view NMF\textsuperscript{162}, and others.

5.2 Preliminary Results

To capture multiple levels of PTSD biology and symptoms, we integrated DNA methylation patterns, metabolites, endocrine markers, miRNA expression, clinical lab measures, physiological features, and clinical symptoms to generate candidate subgroups of PTSD participants. A summary of the included data types in shown in Table 5.1.
Table 5.1: Summary of PTSD data included in subgroup identification. Clinical, physiological, and molecular datasets were integrated to cluster PTSD participants into candidate subtypes. The included datasets ranged in size from 10’s of markers to close to half a million markers.

<table>
<thead>
<tr>
<th>Data Type</th>
<th># of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical</td>
<td>83</td>
</tr>
<tr>
<td>Clinical Lab</td>
<td>44</td>
</tr>
<tr>
<td>Endocrine</td>
<td>20</td>
</tr>
<tr>
<td>Metabolomics</td>
<td>230</td>
</tr>
<tr>
<td>Methylation</td>
<td>435491</td>
</tr>
<tr>
<td>miRNA</td>
<td>335</td>
</tr>
<tr>
<td>Physiological</td>
<td>12</td>
</tr>
</tbody>
</table>

5.2.1 Identification of PTSD subgroups

We have implemented a data-driven approach to PTSD subtype identification using multiple levels of molecular and clinical data. Due to the sample size and high signal-to-noise ratio in many of our high throughput PTSD discovery datasets, we sought to integrate multiple levels of molecular and clinical data in order to identify consistent cluster signals. To reduce the dimensionality of the large DNA methylation microarray dataset, we projected CpG probe-level data onto known biological pathways. Using pathifier\(^{82}\), pathway deregulation scores were computed to represent methylation pattern changes in related sets of genes. Briefly, for each pathway the ‘pathifier’ method considers a lower dimensional subspace containing only genes or components contained in the pathway. The principal curve\(^{163}\) through the data points in the subspace is generated, and each data point is projected onto this curve. A single scalar value representing the length along this curve can then be used as a measure of biological pathway activity for each individual, for each pathway.

We then clustered PTSD subjects from the discovery cohort into two distinct subgroups. These PTSD subgroups were defined using the iCluster framework\(^{159}\), based on PTSD symptom severity and signals from biometric, clinical lab, endocrine, metabolomic, methylation, and physiological datasets. Using \(k=2\) and \(\lambda=0.16\), iCluster identified two subgroups, containing 33 and 45 PTSD subjects in groups 1 and 2, respectively.

The number of clusters and \(\lambda\) value in the iCluster model were selected based on the Proportion of Deviance (POD) value from each subgrouping, over \(k=2\) to \(k=6\), with 50 \(\lambda\) values, equally spaced between 0 and 1 (Fig. 5.1). \(\lambda\) values of 1 result in sparse clustering solutions, while \(\lambda\) values of 0 result in a non-sparse solution. The minimum POD value occurs at \(k=2\) and \(\lambda=0.16\).
Figure 5.1: Proportion of Deviance (POD) for clustering PTSD subjects using iCluster. Red line indicates average POD over all lambda values; gray region indicates min/max. The minimum POD value occurs at \( k=2 \) and \( \lambda=0.16 \).

Figure 5.2: Proportion of Deviance (POD) for clustering control subjects using iCluster. Red line indicates average POD over all lambda values; gray region indicates min/max. The minimum POD value occurs at \( k=2 \) and \( \lambda=0.12 \).
Similar clustering can be performed on the control participants from the discovery cohort. In the control population, $k=2$ and $\lambda=0.12$ resulted in the smallest POD value (Fig. 5.2), though the minimum POD value was significantly larger in the control population vs PTSD population ($0.07$ vs $<10^{-18}$). The characteristics of these control subgroups should be explored further to understand how these clusters may be related to ethnicity, co-morbidities, or other demographic and clinical features.

While independent clustering on each data type does not result in identical clustering, the metabolomic, clinical labs, and clinical datasets show weak separation from the two iCluster-defined subgroups, while the methylation data shows strong separation (Fig. 5.3). These subgroups show stronger cluster separability than cluster of control subjects, based on proportion of deviance (POD) from a perfect block diagonal cluster matrix (Fig. 5.2). Further validation and characterization of these candidate subgroups should be explored in additional datasets. Additionally, identification of individual markers for stratification, as well as subtype-specific biomarkers may aid in improved diagnostic performance.

### 5.3 Incorporating PTSD subgroup-specific biomarkers

Previously identified molecular and clinical subtypes of PTSD, or those identified in this chapter, may require subtype-specific biomarkers in order to be used for diagnosis. These subtype-specific biomarkers may lead to improved accuracy of PTSD diagnosis and the ability to predict the most effective treatment strategy. Instead of using a set of biomarkers to build a binary classifier (control vs PTSD), a potentially new set of biomarkers could be identified to use in multi-class classification (control vs PTSD subtype 1 vs PTSD subtype 2, etc...). Multiple standard classification algorithms are able to predict more than two outcome groups, including decision trees/random forest, k-nearest neighbors, and neural networks. Other algorithms can be easily adapted to multi-class data, simply by reducing the problem to multiple binary predictions. In particular, this set of adapted binary classifiers may contain different biomarkers for each binary prediction, utilizing unique biology for each subtype of PTSD. Using additional cohorts, these multi-class strategies can be used to generate and evaluate subtype-specific biomarkers for PTSD.
Figure 5.3: Individual data type principal component plots. First two principal components for each data type (clinical labs, clinical, metabolomics, and methylation). Plotted points indicate individual PTSD participants, colored by multi-omic iCluster subgroups. iCluster-defined subgroups are primarily driven by DNA methylation signals.
5.4 Understanding disease progression

Our preliminary approaches to uncover disease progression trajectories and networks have relied on using pattern recognition algorithms to link similar subjects from cross-sectional studies, due to lack of longitudinal data. The linked subjects form a network that may be related to disease progression, including diseases stages and branch points, forming from disease subtypes. Specifically, we have applied the Sample Progression Discovery (SPD) algorithm\textsuperscript{164} to characterize subject similarity and disease progression based on DNA methylation patterns in the 83-83 PTSD discovery cohort.

SPD is an approach designed to recover temporal ordering and structure of data from time series experiments. Originally designed to understand complex biological processes including cell differentiation, this algorithm can also be applied to longer time-scale longitudinal data. SPD builds a minimum spanning tree based on the similarity of gene expression patterns within clustered gene expression modules. This tree represents the structure of progression during a biological process or disease progression. SPD has been applied to both short time-series datasets (tracking embryonic stem cell differentiation), and cross-sectional disease datasets (identifying stages of prostate cancer).

We applied SPD to the 83-83 PTSD discovery cohort. Using the average DNA methylation within 1500 bp of a transcription start site for each gene, we implemented the SPD algorithm. From only PTSD participants, we identified the minimum spanning tree from clustered methylation modules. Fig. 5.4 illustrates the disease progression map, with multiple branch points and potential disease trajectories. By incorporating longitudinal data with multiple time points, we can investigate short individual trajectories through this disease progression network to understand branch point decision-making and network directionality. In particular, this will provide insight into whether branches of this network indicate stages of disease progression, distinct subgroups, or regions of comorbidity.

Throughout this thesis, we have presented a variety of computational tools and analyses relevant to studying Posttraumatic Stress Disorder. We have presented a collaborative approach to identifying a candidate biomarker panel for diagnosing PTSD in veterans. This biomarker panel performed well in an independent validation cohort, with 81% accuracy, and should be explored further in larger, more complex cohorts. We also implemented a methodology to explore biological signals of disease, including PTSD, by comparing multiple statistical metrics.
CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

Figure 5.4: PTSD progression network structure inferred from DNA methylation patterns in 83 PTSD+ subjects. Each node represent a single subject, connected to its nearest neighbor based on similar methylation patterns. Multiple branch points exist within the tree, indicating potential for distinct disease subtypes that may diverge at specific points during disease progression.

for comparing disease and control populations. Using differential expression variance provided novel insight into relevant biological networks of PTSD, including signals from blood coagulation factors. Finally, we explored solutions to two common problems in high throughput biomarker identification projects: incomplete multi-omic datasets, and unequal consequences for prediction errors in positive vs negative classes. We implemented a maximum confidence-based approach to integrating multi-omic predictions, allowing for individual datasets to be missing throughout the cohort. Using the most confident predictor for each subject allowed for a complex disease to be captured by the clearest and most relevant data type for each individual. To generate a biomarker model with the ability to tune sensitivity and specificity, we developed a variance-based feature selection algorithm. This allowed for greater maximization of the metric of interest, without loss of overall accuracy, compared to traditional differential expression based feature selection. Finally, we discussed approaches that should be explored in the future, to further characterize the heterogeneity of PTSD, and model the development and progression of this complex disease.

The integration of data-driven computational tools for biomarker identification and disease characterization
provides a novel opportunity to improve diagnosis, treatment, and stratification of a complex disease such as PTSD. We have presented some promising candidate biomarkers and classification approaches, as well as approaches to deepen the knowledge of PTSD biology and hope this inspires further work that will make great impacts in the lives of those living with PTSD.
PTSD diagnostic criteria

The complete set of diagnostic criteria, exactly as written in the DSM-5, are listed below:

Criterion A– Exposure to actual or threatened death, serious injury, or sexual violence in one (or more) of the following ways:

- Directly experiencing the traumatic event(s).
- Witnessing, in person, the event(s) as it occurred to others.
- Learning that the traumatic event(s) occurred to a close family member or close friend. In cases of actual or threatened death of a family member or friend, the event(s) must have been violent or accidental.
- Experiencing repeated or extreme exposure to aversive details of the traumatic event(s).

Criterion B– Presence of one (or more) of the following intrusion symptoms associated with the traumatic event(s), beginning after the traumatic event(s) occurred:

- Recurrent, involuntary, and intrusive distressing memories of the traumatic event(s).
- Recurrent distressing dreams in which the content and/or affect of the dream are related to the traumatic event(s).
- Dissociative reactions (e.g., flashbacks) in which the individual feels or acts as if the traumatic event(s) were recurring.
- Intense or prolonged psychological distress at exposure to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).
- Marked physiological reactions to internal or external cues that symbolize or resemble an aspect of the traumatic event(s).
APPENDIX A. PTSD DIAGNOSTIC CRITERIA

Criterion C– Persistent avoidance of stimuli associated with the traumatic event(s), beginning after the traumatic event(s) occurred, as evidenced by one or both of the following:

• Avoidance of or efforts to avoid distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).

• Avoidance of or efforts to avoid external reminders (people, places, conversations, activities, objects, situations) that arouse distressing memories, thoughts, or feelings about or closely associated with the traumatic event(s).

Criterion D– Negative alterations in cognition and mood associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:

• Inability to remember an important aspect of the traumatic event(s) (typically due to dissociative amnesia and not to other factors such as head injury, alcohol, or drugs).

• Persistent and exaggerated negative beliefs or expectations about oneself, others, or the world.

• Persistent, distorted cognitions about the cause or consequences of the traumatic event(s) that lead the individual to blame himself/herself or others.

• Persistent negative emotional state.

• Markedly diminished interest or participation in significant activities.

• Feelings of detachment or estrangement from others.

• Persistent inability to experience positive emotion.

Criterion E– Marked alterations in arousal and reactivity associated with the traumatic event(s), beginning or worsening after the traumatic event(s) occurred, as evidenced by two (or more) of the following:

• Irritable behavior and angry outbursts (with little or no provocation) typically expressed as verbal or physical aggression towards people or objects.

• Reckless or self-destructive behavior.

• Hypervigilance.

• Exaggerated startle response.

• Problems with concentration.

• Sleep disturbance.
APPENDIX A. PTSD DIAGNOSTIC CRITERIA

Criterion F– Duration of the disturbance (Criteria B, C, D, and E) is more than one month.
Criterion G– The disturbance causes clinically significant distress or impairment in social, occupational, or other important areas of functioning.
Criterion H– The disturbance is not attributable to the physiological effects of a substance or another medical condition.
Specifications–

• With dissociative symptoms: The individual’s symptoms meet the criteria for posttraumatic stress disorder, and in addition, in response to the stressor, the individual experiences persistent or recurrent symptoms of either of the following:

  – Depersonalization– Persistent or recurrent experiences of feeling detached from, and as if one were an outside observer of, one’s mental processes or body.
  – Derealization– Persistent or recurrent experiences of unreality of surroundings.

• With delayed expression: If the full diagnostic criteria are not met until at least six months after the event.
References


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES


86