

University of Alabama in Huntsville

**LOUIS**

---

Theses

UAH Electronic Theses and Dissertations

---

2019

## **An analysis on utilizing the CDR3 transcriptome in the detection of posttraumatic stress disorder**

Jake Brouwer

Follow this and additional works at: <https://louis.uah.edu/uah-theses>

---

### **Recommended Citation**

Brouwer, Jake, "An analysis on utilizing the CDR3 transcriptome in the detection of posttraumatic stress disorder" (2019). *Theses*. 628.

<https://louis.uah.edu/uah-theses/628>

This Thesis is brought to you for free and open access by the UAH Electronic Theses and Dissertations at LOUIS. It has been accepted for inclusion in Theses by an authorized administrator of LOUIS.

**AN ANALYSIS ON UTILIZING THE CDR3  
TRANSCRIPTOME IN THE DETECTION OF  
POSTTRAUMATIC STRESS DISORDER**

by

**JAKE BROUWER**

**A THESIS**

Submitted in partial fulfillment of the requirements  
for the degree of Master of Science  
in  
The Department of Biology  
to  
The School of Graduate Studies  
of  
The University of Alabama in Huntsville

**HUNTSVILLE, ALABAMA**

**2019**

In presenting this thesis in partial fulfillment of the requirements for a master's degree from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department or the Dean of the School of Graduate Studies. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

  
\_\_\_\_\_  
Jake Brouwer

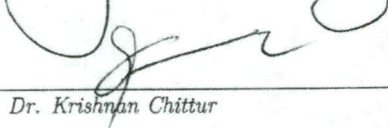
6-19-19  
(date)

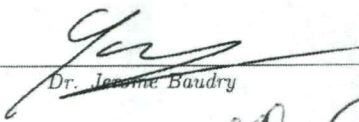
## THESIS APPROVAL FORM

Submitted by Jake Brouwer in partial fulfillment of the requirements for the degree of Master of Science in Biological Sciences and accepted on behalf of the Faculty of the School of Graduate Studies by the thesis committee.


We, the undersigned members of the Graduate Faculty of The University of Alabama in Huntsville, certify that we have advised and/or supervised the candidate of the work described in this thesis. We further certify that we have reviewed the thesis manuscript and approve it in partial fulfillment of the requirements for the degree of Master of Science in Biological Sciences.


  
\_\_\_\_\_  
Dr. Joe Ng (Date) 6-11-19 Committee Chair

  
\_\_\_\_\_  
Dr. Krishnan Chittur (Date) 6-11-19

  
\_\_\_\_\_  
Dr. Jasmine Baudry (Date) 6-11-19

  
\_\_\_\_\_  
Dr. Bruce Stallsmith (Date) 6-25-19 Department Chair

  
\_\_\_\_\_  
Dr. John Christy (Date) 07.03.2019 College Dean

  
\_\_\_\_\_  
Dr. David Berkowitz (Date) 7/10/19 Graduate Dean



## ABSTRACT

School of Graduate Studies  
The University of Alabama in Huntsville

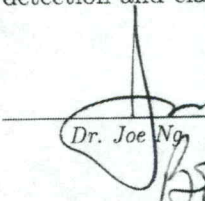
Degree Masters of Science College/Dept. Science/Biological Sciences

Name of Candidate Jake Brouwer


Title An analysis on utilizing the CDR3 transcriptome  
in the detection of posttraumatic stress disorder

An estimated 20% of individuals with occupations such as war fighters and first responders, or individuals that are subject to traumatic circumstance such as violence or abuse, will develop PTSD due to exposure to trauma. PTSD is known for its impact on brain chemistry but has also been shown to be co-morbid with other diseases and disorders such as cardio-vascular disease or fibromyalgia. This study analyzes the repertoire of CDR3 proteins utilized by the T-cells of the adaptive immune system and shows how the CDR3 transcriptome may be a useful alternative in the detection of PTSD. This study implements multiple different statistical, variational, and clustering analyses to show that the CDR3 transcriptome does contain information regarding the bodies response to trauma. The findings presented here provide a basis for the continued study of the relationship between the immune system and trauma response as well as potential methodologies for the detection and classification of PTSD.


Abstract Approval: Committee Chair

  
Dr. Joe Ng

Department Chair

  
Dr. Bruce Stallsmith

Graduate Dean

  
Dr. David Berkowitz

## ACKNOWLEDGMENTS

This work would not have been possible without the support of my family, nor the support of Dr. Ng and my committee. Thank you all for supporting my in this venture.

## TABLE OF CONTENTS

List of Figures	x
<b>Chapter</b>	
<b>1 Introduction</b>	<b>1</b>
1.1 Background . . . . .	1
1.2 Objective . . . . .	7
<b>2 Methods</b>	<b>10</b>
2.1 Wrangling the Raw Data . . . . .	10
2.2 Basic Statistical Analysis . . . . .	13
2.3 Numerical Comparisons with Heatmaps . . . . .	13
2.4 Variational Analysis . . . . .	16
2.5 Clustering Analysis . . . . .	19
<b>3 Results</b>	<b>21</b>
3.1 Basic Statistical Analysis . . . . .	21
3.2 Numerical Comparisons with Heatmaps . . . . .	23
3.3 Variational Analysis . . . . .	26
3.4 Identification of key features . . . . .	33
3.5 Clustering Analysis . . . . .	38

4	Discussion	46
5	Conclusion	50
6	References	51
	APPENDIX A: Individual Heatmaps	57

## LIST OF FIGURES

FIGURE	PAGE
2.1 Chart provided by Dr. Ng linking individual ID labels to their assigned designation . . . . .	11
3.1 Subtracting the average TESP heatmap from the average NTE heatmap	25
3.2 Subtracting the average TESN heatmap from the average NTE heatmap	25
3.3 Subtracting the average TESP heatmap from the average TESN heatmap	26
3.4 Percent explained variance of first three principal components . . . . .	27
3.5 Shows the centroid of centroid positionings for each feature space. Also shows the distances between each point. . . . .	30
3.6 Shows the centroid of centroid positionings for each feature space. Also shows the distances between each point. . . . .	32
3.7 Shows the centroid of centroid positionings for each feature space. Also shows the distances between each point. Centroids calculated from select or key features are labeled with a lowercase 's' i.e. sNTE. The selected or key features are indicated in the upper right corner. . . . .	36
3.8 The results after running the data of each individual of each group through a PCA and consequent normalization. The use of the V features give clear separation of each group while the use of the J features yielded tighter, less distinguishable clusters. . . . .	40
3.9 Result of applying the Agglomerative Clustering algorithm to the data. This algorithm clearly failed to recognize the existing clusters. . . . .	42
3.10 Result of applying the KMeans Clustering algorithm to the data. This algorithm also could not identify the exists clusters. . . . .	43
3.11 Result of applying the Spectral Clustering algorithm to the data. All three clusters were clearly identified with no obvious errors. . . . .	43

3.12	The Spectral Clustering algorithm failed to properly identify the centroids of any of the groups largely because of a lack of data from which the algorithm can draw conclusions from. . . . .	45
A.1	NTE 1 . . . . .	57
A.2	NTE 2 . . . . .	58
A.3	NTE 3 . . . . .	58
A.4	NTE 4 . . . . .	59
A.5	NTE 5 . . . . .	59
A.6	NTE 6 . . . . .	60
A.7	NTE 7 . . . . .	60
A.8	NTE 8 . . . . .	61
A.9	NTE 9 . . . . .	61
A.10	NTE 10 . . . . .	62
A.11	NTE 11 . . . . .	62
A.12	TESN 1 . . . . .	63
A.13	TESN 2 . . . . .	63
A.14	TESN 3 . . . . .	64
A.15	TESN 4 . . . . .	64
A.16	TESN 5 . . . . .	65
A.17	TESN 6 . . . . .	65
A.18	TESN 7 . . . . .	66
A.19	TESN 8 . . . . .	66
A.20	TESN 9 . . . . .	67

A.21 TESN 10 . . . . .	67
A.22 TESN 11 . . . . .	68
A.23 TESN 12 . . . . .	68
A.24 TESN 13 . . . . .	69
A.25 TESN 14 . . . . .	69
A.26 TESN 15 . . . . .	70
A.27 TESN 16 . . . . .	70
A.28 TESN 17 . . . . .	71
A.29 TESN 18 . . . . .	71
A.30 TESN 19 . . . . .	72
A.31 TESN 20 . . . . .	72
A.32 TESN 21 . . . . .	73
A.33 TESN 22 . . . . .	73
A.34 TESN 23 . . . . .	74
A.35 TESN 24 . . . . .	74
A.36 TESN 25 . . . . .	75
A.37 TESN 26 . . . . .	75
A.38 TESN 27 . . . . .	76
A.39 TESP 1 . . . . .	76
A.40 TESP 2 . . . . .	77
A.41 TESP 3 . . . . .	77
A.42 TESP 4 . . . . .	78
A.43 TESP 5 . . . . .	78



A.44 TESP 6 . . . . .	79
A.45 TESP 7 . . . . .	79
A.46 TESP 8 . . . . .	80
A.47 TESP 9 . . . . .	80
A.48 TESP 10 . . . . .	81
A.49 TESP 11 . . . . .	81

## CHAPTER 1

### INTRODUCTION

#### 1.1 Background

Posttraumatic stress disorder (PTSD) can be disastrously consequential to those that it impacts. Symptoms of PTSD can be varied with some being more severe than others. These can include undesired flashback or bad dreams in which the trauma is re-experienced, avoiding any thoughts, feelings, or places remotely associated with the trauma, being hyper-aroused and easily upset, as well as having trouble recalling key pieces or events related to the trauma, negative inward thoughts, distorted feelings of guilt or blame, and a loss of interest in previously enjoyable activities (National Institute of Mental Health 2016; American Psychological Association 2013). The current standard for PTSD diagnosis is laid out in the Diagnostic and Statistical Manual of Mental Disorders, 5th Edition, (DSM-5) published by the American Psychiatric Association (APA) in 2013. The DSM-5 Criteria for PTSD outlines eight criterion for detecting and diagnosing PTSD which include the following:

Criteria 1 - A stressor (at least one required for diagnosis): "The person was exposed to: death, threatened death, actual or threatened serious injury, or actual or threatened sexual violence, in the following way(s): Direct exposure to trauma,

witnessing the trauma, learning that a relative or close friend was exposed to trauma, indirect exposure to aversive details of the trauma usually in the course of professional duties (e.g. first responders” or combat medics or emergency service personnel). Criteria 2 - intrusion of symptoms (at least one required for diagnosis): ”The traumatic event is persistently re-experienced in the following way(s): Unwanted upsetting memories, nightmares, flashbacks, emotional distress after exposure to traumatic reminders, Physical reactivity after exposure to traumatic reminders”. Criteria 3 - avoidance (at least one required for diagnosis): ”Avoidance of trauma-related stimuli after the trauma, in the following way(s): Trauma-related thoughts or feelings, Trauma related external reminders”. Criteria 4 - negative alterations in cognitions and mood (at least two required for diagnosis): ”Negative thoughts or feelings that began and worsened after the trauma, in the following way(s): Inability to recall key features of the trauma, Overly negative thoughts and assumptions about oneself or the world, Exaggerated blame of self or others causing the trauma, Negative affect, Decreased interest in activities, Feeling isolated, Difficulty experiencing positive affect”. Criteria 5 - alterations in arousal and reactivity: ”Trauma-related arousal and reactivity that began or worsened after the trauma in the following way(s): Irritability or aggression, risky or destructive behavior, hypervigilance, heightened startle reaction, difficulty concentrating, difficulty sleeping”. Criteria 6 - duration (required for diagnosis): ”Symptoms last for more than one month”. Criteria 7 - functional significance (required for diagnosis): ”Symptoms create distress or functional impairment (e.g. social, occupational)”. Criteria 8 - exclusion (required for diagnosis): ”Symptoms are not due to medication, substance use, or other illness”.

Two specifications are also required for diagnosis: Dissociative Specification "In addition to meeting criteria for diagnosis, an individual experiences high levels of either of the following in reaction to trauma-related stimuli:" Depersonalization - Experience of being an outside observer of or detached from oneself (e.g. feeling as if "this is not happening to me" or as if in a dream); Derealization - Experience of unreality, distance, or distortion (e.g. "things are not real"). Delayed Specification - Full diagnostic criteria are not met until at least six months after the trauma(s), although onset of symptoms may occur immediately (all quoted text from Brainline 2018, American Psychiatric Association (APA) 2013).

The main mechanisms for ascertaining if an individual exposed to trauma meets the criterion and specifications as described above is through interviews and self-assessment and have been used as the primary way of detecting and diagnosing PTSD since their inception in 1980 (REF, Friedman 2018). The first criterion for diagnosing PTSD were outlined in the DSM-3 from the APA and these criterion were not revised until 1987 in the Diagnostic and Statistical Manual of Mental Disorders, revision 3 (DSM-3R), which was then revised again in the fourth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-4) which was published in 1994. Many of the initial criterion did not consider the etiology of the disorder, did not accurately capture the desired affect or were misinterpreted, were constrained by the way they were written and conveyed, and did not take multiple scenarios into consideration, such as the differentiation of symptom expression in children verses adults and the potential for fluctuation in the severity of expressed symptoms (Solomon et al. 1990; Brett et al. 1988; Green et al. 1985).



There are currently no widely used qualitative analyses implementing the measurement of biological markers or mechanisms. The need for such an analysis to at least aide in the assessment of PTSD is apparent as not only are the above criterion subjective, the choice of interview and self screen must be considered by clinicians as their scores are dependent on the interview being applied to the proper demographics. Additionally the timing of the interview or self screen and the level of experience of the clinician can have an impact on the scoring and consequent interpretation of the interview/screen by the clinician (Steel et al. 2011). This can lead to misdiagnosis and potential mistreatment of symptoms. As for the treatment of PTSD, current common treatments include, but are not limited to: trauma-focused psychotherapies, which can be prolonged exposure, cognitive processing, or eye movement desensitization and reprocessing therapies, as well as brief eclectic psychotherapy, narrative exposure therapy, written exposure therapy, in addition to medications such as selective serotonin reuptake inhibitors (SSRIs) and serotonin-norepinephrine reuptake inhibitors (SNRIs) which serve as antidepressants (Lancaster et al. 2016; U.S. Department of Veterans Affairs 2019). Current pharmacological PTSD treatments set their focus on mitigate chemical imbalances that are thought to be a result of exposure to trauma.

The focus of this study, however, is the impact trauma has on the body beyond the blood brain barrier. Any individual living in the United States has a high chance, 50-90%, of being exposed to a traumatic event at some point in their lifetime. However only roughly 8% of these individuals exposed to trauma will go on to develop PTSD (Vieweg et al. 2006). This suggests that those individuals who develop PTSD have a specific phenotype associated with the failure to recover from the

effects of trauma (Yehuda and LeDoux 2007). Numerous studies have attempted to accurately represent the percentage of individuals, combat veterans or otherwise, who are at risk for developing PTSD. These prevalence reports unsurprisingly vary due to differences in target demographics, the use of different screening methods which in and of themselves are far from perfect, the timing of the screenings, and differences in environments to which individuals were exposed. The range of prevalence estimated by such studies, which pertain particularly to combat veterans, suggests between 5-30% of combat veterans will develop PTSD with the consensus hovering more closely around 10-20% (Blake et al. 1990; Richardson et al. 2010; Ramchand et al. 2010; Gradus 2018). Again it must be noted that these results varied based on deployment, the time frame within which screens were done, and which screens were used. Additionally not all veterans who develop PTSD will report their symptoms or respond to such studies (Ramchand et al. 2010) thus adding in yet another bias which may suggest that the numbers these studies are rather conservative estimates.

It is well known that trauma can and will cause other extenuating health issue beyond brain chemistry imbalance. It has been shown that exposure to various types of stressors can lead to hematopoietic stem cell (HSC) activation and proliferation (Heidt et al. 2014), elevations in circulating blood platelet counts (Lindqvist et al. 2017), and the up-regulation of inflammatory gene expression in mouse models (Powell et al. 2013), all of which are tied in some form or fashion to the immune system. HSCs will proliferate into cells that comprise the adaptive immune system, including T cells. Platelets carry inflammatory inducers and are involved in the inflammatory process and are also associated with increased risk of cardiovascular

disease at elevated levels (Vinholt et al. 2016). It is then unsurprising to see that PTSD and related trauma is shown to play a role in the development of conditions such as cardiovascular disease, diabetes, gastrointestinal disease, fibromyalgia, chronic fatigue syndrome, musculoskeletal disorders, major depressive disorder, and more (Boscarino 2004; Coughlin 2011; Boyko et al. 2010; Maguen et al. 2014; Neumann and Buskila 2003; Danise et al 2013; Kelsall et al. 2014; Flory and Yehuda 2015). One of the features of these diseases comorbid with PTSD is that they involve abnormal immune system functionality and abnormal inflammatory processes (Boscarino 2004; von Kanel 2006; Kessler 1995; Early et al. 2014).

Furthermore a review of studies concerning blood gene expression and glucocorticoid (part of the inflammatory signaling network of the immune system) activity indicates that PTSD is more prevalent in individuals who exhibit abnormal expression of genes responsible that control and cause inflammation, respectively (Heinzleemann and Gill 2013). Another study examined the gene expression of blood leukocytes in Marines before and after they were deployed (Breen et al. 2015). It was found that Marines who were resilient to PTSD, or did not develop PTSD despite exposure to trauma, were more likely to have increased expression of genes related to hemostasis and immune system response to superficial wounds. Conversely Marines who were developed PTSD or were deemed at risk for PTSD had increased expression of genes linked to interferon signaling (in part responsible for adaptive immune system activation and response), thus these Marines exhibited a more active adaptive immune system.



## 1.2 Objective

The link between PTSD and inflammation and consequently the innate and adaptive immune system are key to this study. Inflammation and proper regulation of inflammatory genes is an integral part innate immune system, which is responsible for being the first line of defense against unwanted non-self (with the host or primary organism being the 'self') entities. The innate immune system is also responsible for initiating signal cascade pathways that lead to the recruitment and activation of the adaptive immune system. Cellular immune system responses, both innate and adaptive, are triggered by the presence of inflammatory inducers (any self or non-self molecule that triggers an inflammatory response). The adaptive immune system gets brought into the fold when inflammatory inducers increase the flow of lymph, which carries microbes or cells bearing antigens, to nearby lymphoid tissues. Once the lymph reaches the lymphoid tissue the antigens contained in the lymph are presented to lymphocytes, a class of cells composed of B lymphocytes (B cells) and T lymphocytes (T cells). Both are equally important to the immune response but it is the T cell receptor that is the focus of this study. When a T cell encounters and antigen that its receptor can bind it will proliferate and differentiate into one of several functional types of effector T lymphocytes. These different types of effector T cells which include Cytotoxic, Helper, and Regulatory T cells, all of which have unique roles (Murphy and Weaver 2017).

A T cell receptor (TCR) is a very unique bundle of proteins in both form and function. The TCR is a quaternary protein structure composed of four tertiary

proteins. These tertiary proteins are dubbed complementarity determining regions, or CDRs, because the surface which they form (the TCR) is complementary to that of the antigen that they bind. The three CDRs that comprise the TCR are aptly named CDR1, CDR2, CDR3, and CDR4. The CDR3 region is of particular interest because this CDR3 region is the piece of the TCR largely responsible for antigen recognition. This is believed to be due to its orientation in three dimensional space which puts the CDR3 in direct contact with the presented antigen. If the CDR3 is complementary in size, shape, and structure to the antigen it will proceed to bind the antigen and initiate the appropriate signal cascade (Murphy and Weaver 2017).

The four CDRs which comprise the TCR are formed from various different gene regions which include joining (J) and variable (V) regions. When the genome is being translated and these CDRs are being produced they undergo a quite remarkable process called gene recombination. This gene recombination is what allows for the generation of a vast library, or repertoire, of CDRs and is what causes each TCR to have such high specificity. This study encompassed CDR3s observed to be generated from 48 distinct V gene regions and 13 distinct J gene regions. These regions were chosen based on their "functional" designation through the IMGT database, which is the global reference in immunogenetics and immunoinformatics. There are no other genes related to CDR generation with the functional designation meaning that the 48 and 13 V and J regions encompassed by this study represent the entirety of the genome responsible for CDR generation.

Thus, given the relationship between trauma, inflammation, and consequently the immune system, the question that this study seeks to answer, at least in part, is

whether or not the CDR3 repertoire of the immune system can indicate the presence of PTSD. Data provided by iXpressgenes (iXG) at the HudsonAlpha Institute for Biotechnology on the CDR3 repertoire of 49 veterans was used to study the potential role of the CDR3 region of the TCR in the identification of the presence PTSD. Each of the 49 veterans was assigned one of three designations based on their responses to an interview and a survey which was conducted prior to the analysis herein. If it was determined that the individual was exposed to combat trauma and had developed PTSD then that individual was given a Trauma Exposed Symptom Positive (TESP) designation. If the individual was exposed to combat trauma but did not develop PTSD then they were given a Trauma Exposed Symptom Negative (TESN) designation. If the individual was not exposed to combat trauma they were given a No Trauma Exposed (NTE) designation. was analyzed with the hopes of uncovering distinctions between the different groups of veterans.

## CHAPTER 2

### METHODS

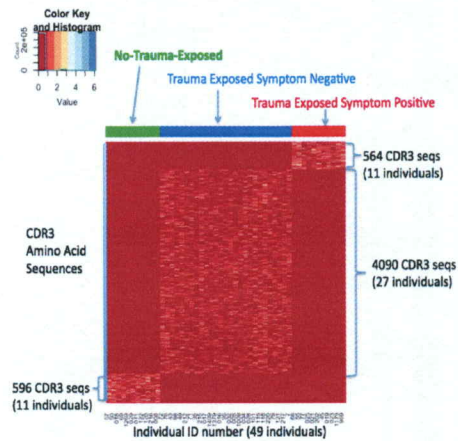
The goal of this research was to ultimately find distinctions between the CDR3 repertoires of the TESP, TESN, and NTE groups that would in some form or fashion set them apart from one another. To attempt to find such distinctions the data had to first be wrangled into a more usable and processed state. The wrangled data was then run through multiple tests and algorithms in attempts to identify such factors or markers that might including basic statistical analyses, comparisons via heatmap visualizations, variational analyses, and clustering analyses.

#### 2.1 Wrangling the Raw Data

All data handling and wrangling of the iXG data was done with the Python 3.6 language and the accompanying SciPy ecosystem which includes the pandas and numpy libraries. The pandas (McKinney 2010) and scikit-learn (Pedregosa et al. 2011) libraries were the main libraries used on this project, along with numpy and other native python libraries which were used to do additional computations. The matplotlib library, native to python 3.6 was used to generate the majority of the plots seen in this paper.



The raw data received from iXG were formatted as .csv files with number labels. Each .csv file contained all of the raw data collected from a single individual. The number labels of each .csv file corresponded to separate ID numbers which were given to each individual by iXG. These ID numbers were matched with ID numbers on a chart which showed which group (TESP, TESN, or NTE) each ID number belonged to. This chart, provided by Dr. Joe Ng, is shown in figure 1.



**Figure 2.1:** Chart provided by Dr. Ng linking individual ID labels to their assigned designation

Once the labeling and sorting of group designations was accomplished, the data needed to be cleaned and assembled into a usable fashion. There were a number of columns present in each .csv file that were unnecessary or whose information did not pertain to the present study. The relevant columns containing information on the variable and joining regions of the CDR3 repertoire were kept and unwanted columns were removed. Additionally, there were occasional rows whose data was not read into the .csv file correctly. These rows were designated with an asterisk in the raw data file

and needed to be removed before moving on with the analysis. Another aspect of the data which needed to be considered was the handling of the CDR3s themselves. Each .csv file contained upwards of 100,000 CDR3s and each CDR3 had a copy number associated with it. This copy number is a representation of how prevalent or expressed the mRNA responsible for producing a CDR3 is in the given sample. Unsurprisingly, the expression level of different CDR3s are inconsistent between different individuals. However each CDR3 comes from some finite combination and/or recombination of 48 V and 13 J gene regions.

Given that these V and J gene regions are present in all individuals it was determined that grouping the CDR3s based on the V and J regions from which they originated would not only provide a convenient method of analysis but would also allow for insight to be gained as to the level of CDR3 expression relative to each V and J combination. Additionally, this allowed for the copy numbers of all the CDR3s associated with each V and J combination to be summed together. This gave each V and J combination a quantifiable representation of how allows for comparison to other VJ pairings. Thus, group-wise comparisons of CDR3 repertoires is possible via VJ groupings. Furthermore, no information is lost in this reduction, simply consolidated.

The result of this initial wrangling process is simple and reduced data files of only the information kept from the original datasets and a data file representing the groupings of CDR3s. These data files are what will be used in the analyses described in the following sections. This initial data wrangling was handled by the `PTSD_preprocessing.py` script. All of the scripts related to this paper can be found in the Scripts Appendix.

## 2.2 Basic Statistical Analysis

The first analysis that was performed was a basic top level statistical analysis on the CDR3 repertoires of each of the three groups. This included basic notation and calculations including the total count of all CDR3s present within each group, total count of all unique CDR3s present within each group, the average number of CDR3s within the repertoire of any individual from any given group, and the average unique CDR3s within the repertoire of an individual from any given group. The total or global number of CDR3s, total unique CDR3s, and average CDR3 count of all the groups was also calculated. From there each group was compared to these global totals and averages in order obtain observations that might help to better explain what was contained within the data of each group such as a percentage based comparison of how the averages of each group compared to the global average. The results from each group were also compared to the results of other groups in an attempt to find any obvious differences between any of the groups.

## 2.3 Numerical Comparisons with Heatmaps

After the basic analyses were completed additional comparisons and visualizations of the data was needed. One of the ways which was decided upon to effectively compare each group to one another was through the use of heatmaps. The idea being that visualizing V and J combinations with high CDR3 copy, or expression, numbers will make for easy comparisons. Additionally, assigning values representative of the CDR3 expression levels of each pair would make direct comparisons possible. How-



ever in order to do direct comparisons a way to standardize CDR3 values across all of the groups was required. A process called binning was utilized in order to achieve this. With this data especially it could be hard to say how different expression levels of 6785 and 11526 are or determine how significant it is that a majority of the CDR3s expression is occurring at a particular interval for a V and J combination. This binning process attempts to help clarify which is really being seen with the data.

Binning involves taking values which fall within a given interval and assigning values that fall within this interval a new identifier representative of that interval. Each interval designates a bin and the values that fall within that interval are contained within said bin. Intervals are designed to be consistent in length and their spacing is determined by the number of bins being used. For example the range of numbers between 1 and 100 can be represented with 5 bins with an interval of 20 (so the first bin contains the interval 1 through 20), or 2 bins with an interval of 50 (the second bin would contain the interval 51 through 100). In this case 100 bins were used and an interval 3785 was decided upon. This interval was arrived at by taking the smallest maximum CDR3 expression number of any of the groups, which was 378614, dividing by 100, and rounding down. An exception was made with the interval for the last bin as the last bin was designed to capture the extreme values of the group. Thus the last bin contained any CDR3 expression values which ranged from the end of the 99th interval to the maximum CDR3 expression value.

The same interval was used for each group and value assignment for each interval was also kept consistent across each group. This value assignment was kept simple, assigning a zero to the expression value any CDR3 whose real expression

value fell within the 0 to 3785 interval, assigning a 1 to CDR3s whose real expression value that fell within the 3786 to 7570 interval, etc. In this way the qualification of CDR3 expression becomes easier and also allows for the development of heatmaps that represent each group that can then be compared to heatmaps of other groups.

This was done by first generating heatmaps of each individual of each group. This is accomplished by summing all of the CDR3s and their binned expression levels that belong to each V and J combination for a single individual. A heatmap can then be generated where each V and J combination is represented by plotting each of the possible 48 V genes on the X-axis and each of the possible 13 J genes on the Y-axis. The result is a plot where every possible V and J combination is represented in a 48 by 13 grid, or plot. This grid can be easily turned into a heatmap with the matplotlib python library that shades each V and J combination a particular color based on the binned expression value associated with that combination.

Once this has been done with every individual of a group, these heatmaps can be averaged together to get a single representation of each group in the form of a heatmap. This allows V and J gene combinations responsible for producing CDR3s with little to no expression or CDR3s with extremely high expression values to stand out. These heatmaps of each group can then be directly compared to one another by taking the values associated with the heatmap of one group and subtracting the values associated with the heatmap of another group.

The values of the resulting V and J gene regions reveal differences in CDR3 expression between groups. If the value of a V and J regions is close to zero, it indicates that individuals of both groups exhibited similar expression of CDR3s associated with

that V and J region. If the value is highly positive or highly negative it indicates that individuals (or potentially one single individual) exhibited much higher expression of CDR3s associated with that V and J region than the individuals of the other group. The sign of the value simply indicates which group exhibited higher expression.

## 2.4 Variational Analysis

Having observed the raw data of the CDR3 expression levels of each V and J combination the next step was an analysis on the variation of the CDR3 repertoire within each group and a subsequent group-wise comparison. This concept of leveraging the variation of a dataset to achieve a desired outcome is utilized in many areas of engineering including data compression, mechanical prognostics, and image reduction. Here it is utilized to measure the inherent deviations present within the V and J combinations of the CDR3 repertoires of each group. For example, if all of the V and J combinations of an individual had the same exact expression values, the variation within of that dataset would be zero as there is no change in the data between V and J combinations.

The variational analysis performed in this instance was a principal component analysis (PCA). In brief A PCA utilizes a linear transformation to convert a set of features into a set of linearly independent features or principal components. The transform is defined such that the first principal component contains largest possible variance, with each successive principal component containing the highest possible variance under the constraint that it is orthogonal to the preceding components. The result then is a set of independent features, or principal components, that define the

variation of the original dataset. There can be as many principal components as there are features in a dataset, features here being the columns of data that the PCA performs its transformation on.

In long form the process to compute a PCA is as follows. Given a dataset  $\mathbf{X}$ : take the mean of the data as:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

where  $n$  is the number of samples in the dataset, and subtract it from  $\mathbf{X}$ . Then calculate the covariance ( $cov$ ) and covariance matrix ( $\mathbf{C}$ ) of  $\mathbf{X}$ , given that  $A$  and  $B$  are features of  $\mathbf{X}$ , as:

$$cov(A, B) = \frac{\sum_{i=1}^n (A_i - \bar{A})(B_i - \bar{B})}{n - 1}, \mathbf{C} = \begin{pmatrix} cov(A, A) & cov(A, B) \\ cov(B, A) & cov(B, B) \end{pmatrix}$$

Another way to define  $\mathbf{C}$  is

$$\mathbf{C} = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T}{n - 1}$$

and following from this equation, the covariance can thus be stated as

$$\mathbf{C} = \frac{XX^T}{n - 1}$$

which more easily shows that the covariance matrix is a square matrix. Thus the eigenvalues ( $\lambda$ ) and eigenvectors ( $\mathbf{v}$ ) of  $\mathbf{C}$  can be calculated given  $\mathbf{C} \times \mathbf{v} = \lambda \times \mathbf{v}$ . This so called eigenvalue problem can be rewritten as  $(\mathbf{C} - \lambda \times \mathbf{I})\mathbf{v} = 0$  where  $\mathbf{I}$  is an



identity matrix. Thus, if  $\mathbf{v}$  is non-zero, this equation will only have a solution if

$$|\mathbf{C} - \lambda \times \mathbf{I}| = 0$$

These eigenvectors then represent the of each of the principal components of the PCA, the first eigenvector representing the first principal component, the second eigenvector represents the second principal component, and so on. The eigenvalues then represent how much variance is present within each principal component. Each component of the PCA can then be mapped into the variational space with the values given by each eigenvector.

The implementation of the PCA was done using the sklearn Python library which comes with a very robust PCA module that handles all of the necessary calculations and can return them to you upon request which is quite useful. The slightly problematic part of using PCA is that visualization can be tricky or misleading if a large majority of the variation of a data set is not captured in the first three principal components. Visualizing the first three components can be done in three dimensional space, with each principal component representing one axis, but visualizing more than the three components cannot be done. Thankfully the sklearn PCA module calculates the variation captured within each principal component each time the PCA is run.

It must be noted that none of the data was scaled or normalized before implementing the PCA as the objective was to pick out features with the highest variation. Any scaling or normalization would reduce the significance of features with high variation, which in most other instances is beneficial to the analysis, here however it would

be counter-productive. Once the PCA was performed the resulting variational data was then scaled from zero to one as to make it more logical and easier to interpret graphically.

## 2.5 Clustering Analysis

With the variational analysis done, a clustering analysis was then performed in an attempt to see if there existed any distinct or discernible clusters representative of any of the groups in the highest dimension that can be easily realized visually, which is three dimensional space. First just the raw data of every individual of each group was plotted in three dimensional space, then the same plot was generated but with data of every individual of each group scaled between zero and one. Once this initial plotting was completed, the data was then run through multiple different clustering algorithms from the sklearn.cluster python library. These different clustering algorithms included KMeans Clustering, Affinity Propagation, Mean Shift, Spectral Clustering, Agglomerative Clustering, DBSCAN, and Birch clustering. These different algorithms represent all of the clustering algorithms available in the sklearn.cluster library.

Once the data is input and fit to the algorithms, the algorithms outputs its best estimation as to where any possible clusters exist. These results can vary wildly due to different constraints and calculations made by each different algorithm. The original input data can then be compared to the output data of the clustering algorithm and each algorithm can be scored based on its accuracy. These comparisons and scores can include simply noting how many clusters the algorithms identified as well as the

calculation of the centroid of each of the clusters and the use of different scoring mechanisms. The closer the centroids of the output of the clustering algorithm are to the centroids of the raw data can give an indication of the accuracy of the clustering algorithm. Other scoring methodologies such as Adjusted Rand Score, Adjusted Mutual Info Score, and the Fowlkes Mallows Score give more objective quantification of the accuracy of the clustering algorithm. Additionally, each algorithm will give a label to each point to signify the cluster that that point belongs to. An accuracy score for that algorithm can be calculated by simply calculating how many points an algorithm labeled correctly.



## CHAPTER 3

### RESULTS

#### 3.1 Basic Statistical Analysis

With the data wrangled into a usable format, some initial statistical analyses were done to see if there was anything glaringly distinct about any of the three veteran groups. One of these tests included calculating, on average, what the total CDR3 copy (or expression) number within the data of each individual of each group. Taking the average was necessary due to having an unbalanced dataset with 27 TESN individuals compared to 11 TESP and 11 NTE individuals. The TESN group had a CDR3 copy number total of 61,548,305, with 23,172,689 for the NTE copy number total, and 18,365,492 for the TESP copy number total. This gives an average TESN individual with a CDR3 copy number total of 2,279,567, the average NTE individual with 2,106,608, and the average TESP individual with 1,669,590.

What immediately stands out is that the TESP average is only 73.2% of the TESN average. Beyond the margin being greater than 25% between the CDR3 copy number totals of these two groups, this result was proven significant via a t-test and p value approximation. Given an alpha of 0.05, the p value between these groups was calculated to be between 0.05 and 0.025 which gives statistical significance to these

findings. Other tests were done on the average total number of just the CDR3s and unique CDR3s of each group but there was little to suggest a significant or discernible difference between any of the groups.

The results of the initial testing show that the data does contain information which separates the TESP group from the other two groups. However, to verify the implications of these initial observations that the CDR3 repertoire does play a role in the expression of symptoms related to PTSD, further tests were conducted. The script detailing these initial observations is the `PTSD_Initial_Observations_T.py` script.

Due to the nature of the problem and the given sample size doing comparisons between individuals would not have been appropriate, at least initially. Before attempting to identify the discrete details which distinguished one individual from another, the three groups were first compared to one another. Given that the data provided only captures a singular snapshot of an individuals CDR3 repertoire at one specific time it is possible that any one individual could have been sick or had some other issues impacting their CDR3 repertoire, unrelated to PTSD, during the initial data collection. The assumption was made that the CDR3 repertoires of each individual were not perfect isolated case studies for studying solely the impact of PTSD on the CDR3 repertoire. Thus, to attempt to minimize the impact of unknown variables, each group was observed as a whole and then compared to the other groups.

### 3.2 Numerical Comparisons with Heatmaps

Using the methodology described above to generate the heatmaps generated for every individual of each group, it quickly becomes apparent how varied the heatmaps between individuals, even individuals of the same group, can be. In order to make comparisons between each group the binned data sets of each individual had to be averaged together. Then a heatmap representative of the CDR3 repertoire of each group can be generated. With these heatmaps of the average CDR3 repertoire of each group generated it is possible to subtract one group dataset from another to more directly compare the differences between each group. The subtraction reveals similarities between the CDR3 expression levels of each group, as the values of similar V J pairs go to zero, and differences between the CDR3 expression levels as the values of V and J pairs that are elevated in only one group will either remain elevated. For example, subtracting the TESP and TESN dataset may yield a V J pair whose value is close to zero. This indicates that both datasets had similar values for the V J pair. If a V J pair is highly negative this indicates that only that V J pair was elevated in only the TESN dataset whereas a highly positive number indicates that V J pair was elevated in only the TESP dataset.

The results of this numerical comparison via heatmaps revealed many promising and intriguing results. The NTE group had relatively higher levels of CDR3 expression compared to the TESP group except for roughly six V and J combinations ([V2 , J1-2] ; [V25-1 , J2-5] ; [V27 , J1-1] ; [V28 , J1-4] ; [V28 , J2-3] ; [V6-3 , J1-1]) where the TESP group presented higher expression levels (Figure 2). In comparison,

the TESN group had higher CDR3 expression numbers across the entire board, except for one or two low expression regions, when compared to the NTE group (Figure 3). Compared to the TESP group the TESN numbers were again higher across the board except for four V and J combinations ([V2 , J1-2] ; [V25-1 , J2-5] ; [V28 , J1-4] ; [V28 , J2-3]) where the TESP group exhibited generally higher expression levels (Figure 4). There is a notable overall elevation in CDR3 expression in the TESN group which suggests that having higher CDR3 expression may play a key role in the immune system being able to tolerate the stresses of trauma. Conversely, this data may suggest that the TESP group is more susceptible to developing symptoms of PTSD due to a lower levels of CDR3 expression compared to even the NTE groups who serves as the general control group in this study. Either way it is hard to dispute that there does in fact seem to be a connection between the level of CDR3 expression and the presence of symptoms of PTSD.

For all heatmaps, positive (red) values indicate greater CDR3 counts in the first group and negative (blue) values indicate greater CDR3 counts in the second group being that the second group is subtracted from the first. As shown in Figure 2 there are only a few VJ regions where the TESP group had higher CDR3 counts than the NTE group, the (V2, J1-2) (V28, J1-4) (V28, J2-3) and (V6-3, J1-1) regions. Additionally, Figure 4 shows that the only regions with higher CDR3 counts in the TESP group compared to the TESN group are the (V2, J1-2) (V28, J1-4) and (V28, J 2-3) regions. The script used to generate the heatmaps is the `PTSD_heatmap_variations_T.py` script. All of the heatmaps for each individual can be found in Appendix 2.



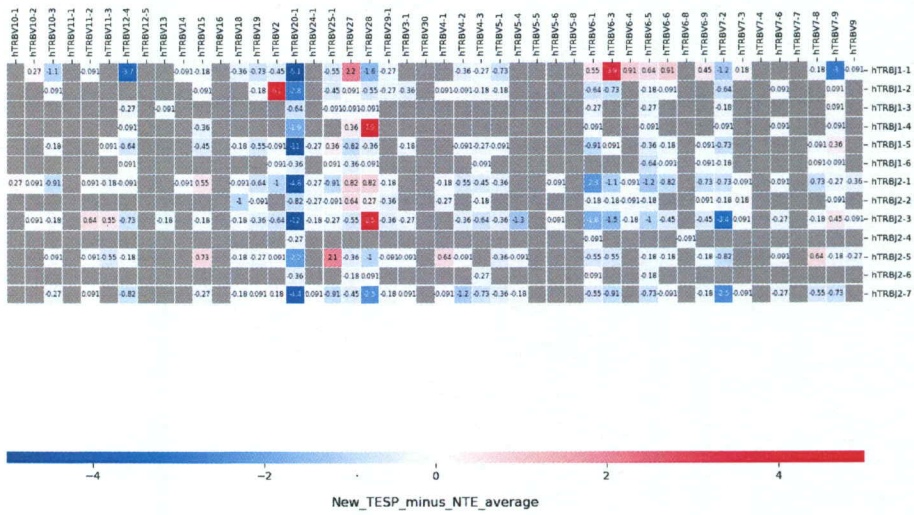


Figure 3.1: Subtracting the average TESP heatmap from the average NTE heatmap

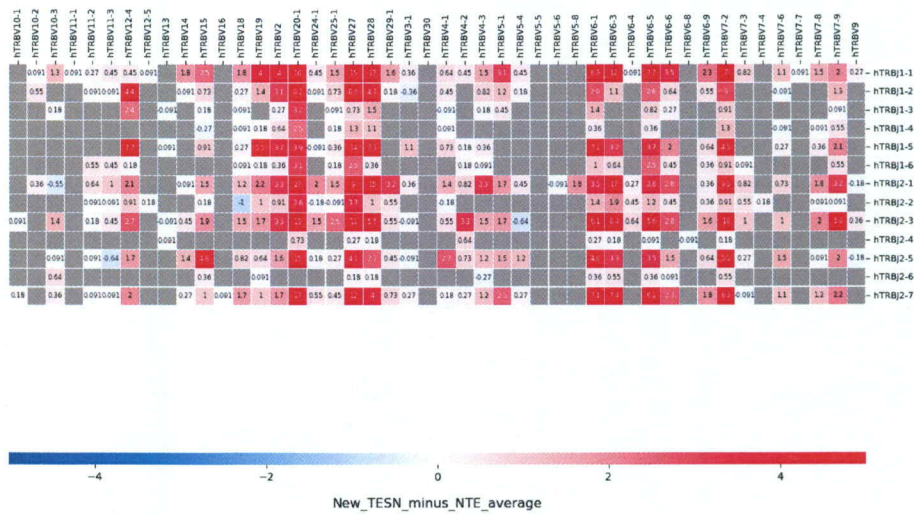
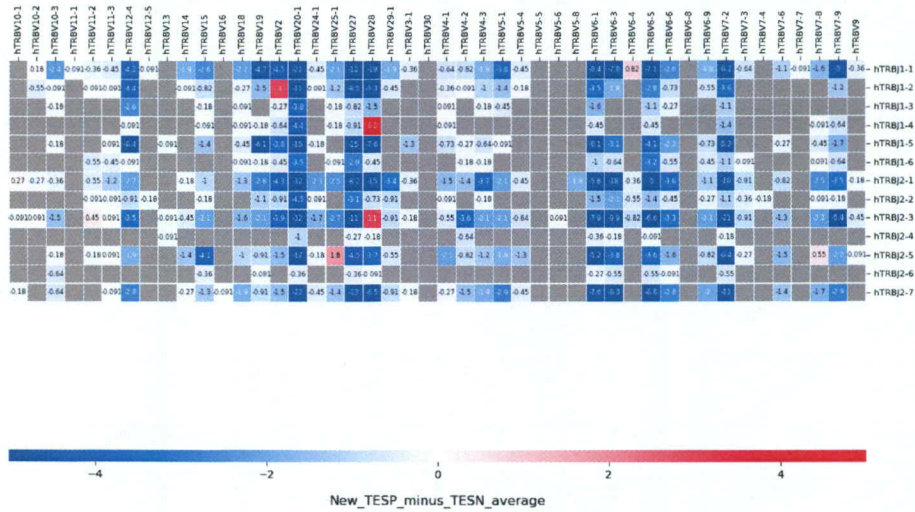


Figure 3.2: Subtracting the average TESN heatmap from the average NTE heatmap



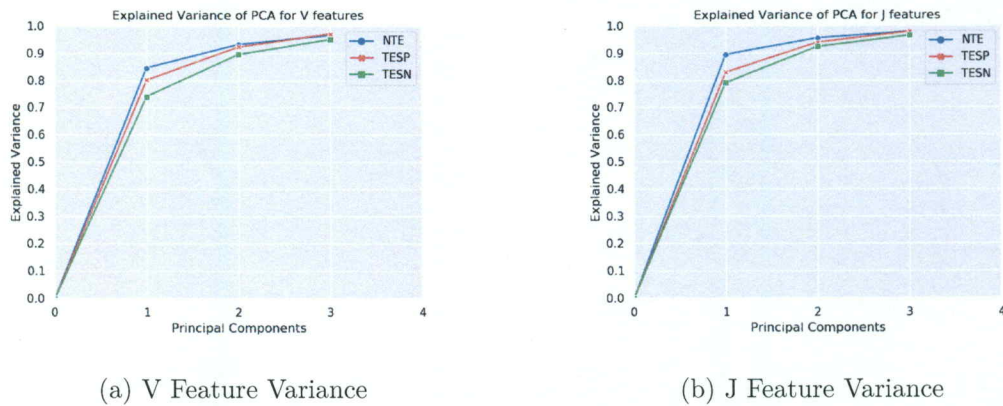


**Figure 3.3:** Subtracting the average TESP heatmap from the average TESN heatmap

### 3.3 Variational Analysis

With the results of the initial testing and the heatmap comparisons, it seemed apparent that there was indeed a discrepancy between the TESP group and the NTE and TESN groups. A variational analysis was then conducted to find any differences in the variation of the CDR3 repertoires of the three different veteran groups. A Principal Component Analysis (PCA) was used to implement this type of analysis. A PCA was done on all of the data of all the individuals of each group and the average total percent variation captured by the first two principal components using the V columns as features was as follows: NTE - 93.19%, TESP - 92.19%, TESN - 89.53%. The average total percent variation captured by the first two principal components

using the J rows as features was as follows: NTE - 95.39%, TESP - 93.73%, TESN - 92.21%. Figure 5 shows the total explained variance over the first three principal components. A PCA, by definition defines, the first principal component as having the largest variation within the dataset, the second principal component having the second highest variation, and so on. Figure 5 clearly shows that the majority of the variation related to the CDR3 repertoires of each group is contained within the first two principal components with a significant increase in information gained from the third principal component.



**Figure 3.4:** Percent explained variance of first three principal components

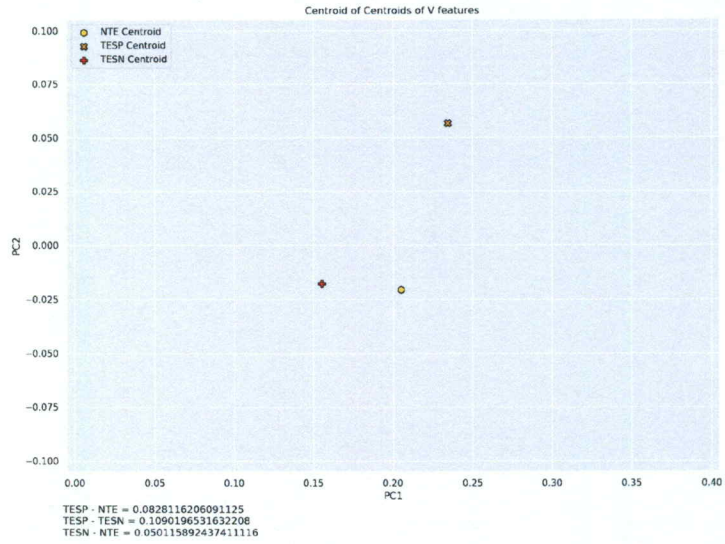
Based on these numbers it was deemed acceptable to use only the first two principal components for the analysis to follow, as upwards of 90% of the information contained within the data deemed to be captured by the first two principal components. Once again, the goal of using the PCA was to be able to compare the variation of the CDR3 repertoire of each group to the variation of the other groups. In order

to analyze and visualize the variation of each group a value representative of the combined variation of all of the individuals of a group was required. To accomplish this the centroid of all of the data of the individuals of each group was calculated. This was done by finding the average value of the first principal component (which can be thought of as the x-coordinate) and the average value of the second principal component (which can be thought of as the y-coordinate) of all of the features of each individuals data. This centroid thus represents the individual as a mean value of all of the data of that particular individual. Now each group can be represented by group of centroids, with these centroids representing the data of the individuals of that group. To find a singular point to represent the data of the entire group another centroid was calculated. This so called centroid of centroids (CoC) is calculated by taking the average of the X-coordinates of all of the individuals in the group, and the average of all of the Y-coordinates of all of the individuals in the group. The result is a singular point which represents the variational data of the first two principal components of one entire group. This process was done on all three groups yielding three distinct points which could then be plotted and compared.

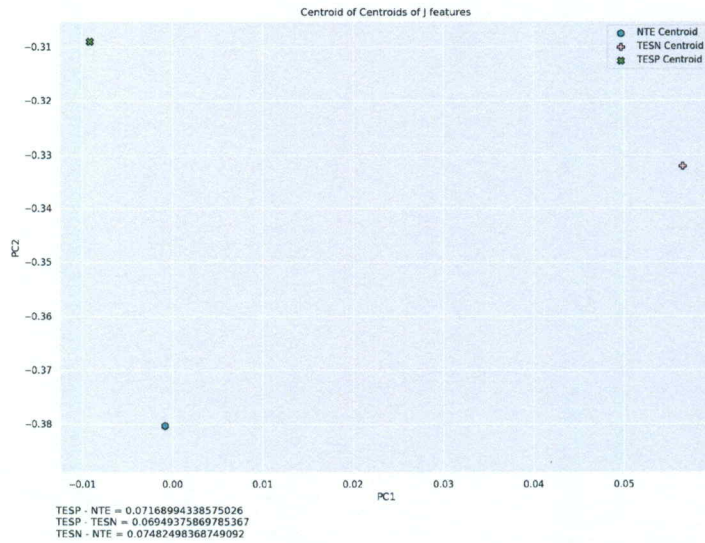
For the V features the CoC of the TESP group was a large distance removed from the CoC of the NTE group (0.0828 units) and the CoC of the TESN group (0.1090 units) which is highly suggestive of a variational disparity between the CDR3 repertoire of TESP group and the repertoire of the other groups. Additionally, the CoC's of the NTE and TESN groups were within a very small proximity of one another ( ) suggesting that the CDR3 repertoire of these two groups are variationally similar (Figure 6). These results reinforce what was found with the heatmap comparisons

and provides further evidence that the CDR3 repertoire plays a role in the way the body responds to trauma. For the J features, the CoCs of each group were almost equidistant from one another with the largest distance between any of the CoCs being only .00533 units longer than the smallest distance, which works out to be less than an eight percent difference. It is worth noting that the NTE and TESP centroids differ significantly on the y-axis but are much closer together on the x-axis than they are to the TESN centroid.





(a) Positioning of the Centroids of Centroids of the V features.

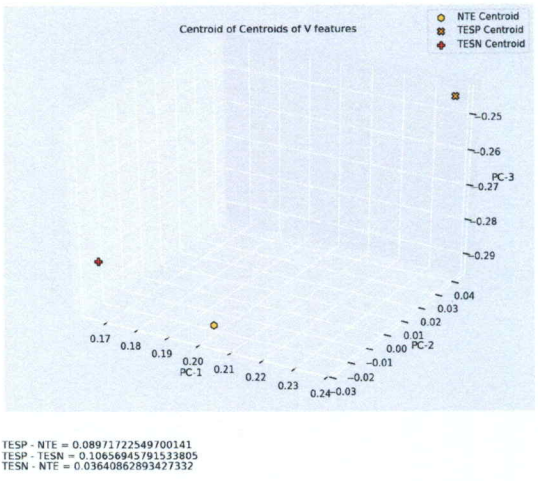


(b) Positioning of the Centroids of Centroids of the J features.

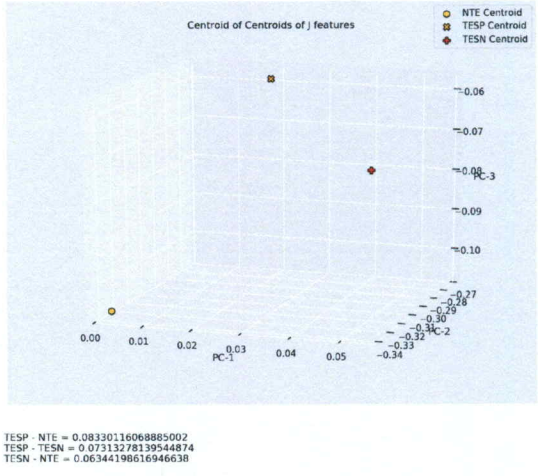
**Figure 3.5:** Shows the centroid of centroid positionings for each feature space. Also shows the distances between each point.



For completeness these centroid of centroids were also plotted in three dimensions using the third principal component as the z-axis. While there is not much variational information added to the total explained variance by using the third principal component (Figure 5), there is still practicality for plotting in three dimensions. For the V features there was noticeable movement of the TESP centroid along the z-axis away from the NTE and TESN centroids, further reinforcing what was uncovered in the two dimensional space. For the J features the NTE centroid was shown to be a noticeable distance below the TESP and TESN centroids, however the distances between each centroid remained relatively consistent to the two dimensional plot. These three dimensional plots, shown in Figure 7 confirm the findings of the initial two dimensional plots while also providing a basis for the clustering analyses which will be discussed in a forthcoming section.



(a) Positioning of the Centroids of Centroids of the V features in three dimensions.



(b) Positioning of the Centroids of Centroids of the J features in three dimensions.

**Figure 3.6:** Shows the centroid of centroid positionings for each feature space. Also shows the distances between each point.

These results also lead to the formulation of another hypothesis that restoring the variation of the CDR3 repertoire of the individuals who comprise the TESP group can be driven to levels more near the variational levels of the TESH/NTE group may lead to the alleviation of symptoms of PTSD. It must be noted that repeated random sampling of eleven of the twenty-seven total TESH individuals, subsequent repeated calculation of the centroid of the given eleven individuals and the CoC of the eleven individuals, followed by the calculation of the centroid of these CoCs showed that the average centroid of centroids was in roughly the same location as the CoC of all twenty-seven TESH individuals. What this proves is that the original CoC of the TESH group is in fact a good representation of the group despite the small number of individuals within the group. The script used to run this variational analysis is the `pca_playground.py` script. Other kernels besides the standard or normal PCA were examined to see how their utilization would affect the outcome of this variational analysis. A linear, polynomial, radial-basis function, sigmoid, and cosine kernel were all tested but did not yield satisfactory results or results that enhanced the understanding of the data beyond what was shown in the initial tests.

### **3.4 Identification of key features**

Having this CoC comparison proved useful in another way, as it allowed for the identification of key features that proved to contain most of the information in regards to the observed variation. It was first noted from the heatmaps that a large number of features contained little to no information. Logically then these features would have little impact on the location of the centroids and CoCs within the variational

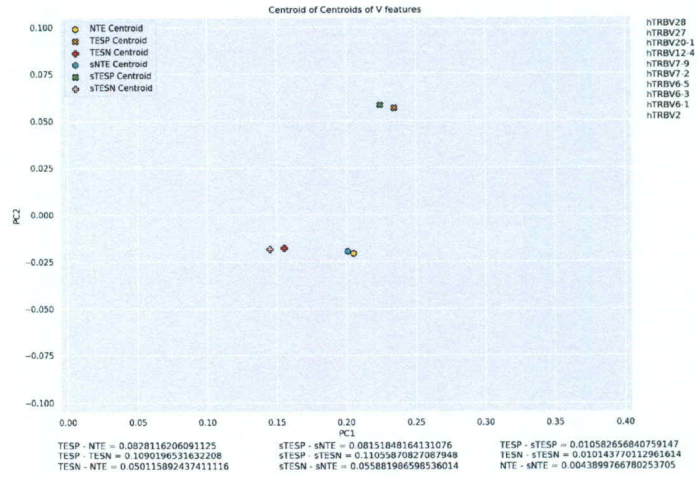
space. These features containing little information were identified and cut out of the original datasets. Then the same variational analysis described in the above section was performed on these reduced datasets and resulted in plots extremely similar to the plots of the original datasets with every feature included.

After the identification and removal of the most obviously non-contributing features from the original datasets, other features were removed via trial and error, by removing one or two features, re-running the variational analysis and measuring the distances between the CoC's produced from the original datasets to those produced from the reduced datasets. Eventually a point was reached where no additional features could be removed from the reduced datasets without the result of the variational analysis becoming significantly different from the result of the original analysis. It was then concluded that these features which comprised the reduced dataset were the features responsible for driving the location of the CoC's of the original datasets and thus contained the majority of the information related to the variation of the CDR3 repertoire. The V features identified as the driving features were: ['hTRBV28', 'hTRBV27', 'hTRBV20-1', 'hTRBV12-4', 'hTRBV7-9', 'hTRBV7-2', 'hTRBV6-5', 'hTRBV6-3', 'hTRBV6-1', and 'hTRBV2']. The J features identified as the driving features were: ['hTRBJ1-1', 'hTRBJ1-2', 'hTRBJ1-5', 'hTRBJ1-4', 'hTRBJ2-1', 'hTRBJ2-3', 'hTRBJ2-5', and 'hTRBJ2-7'].

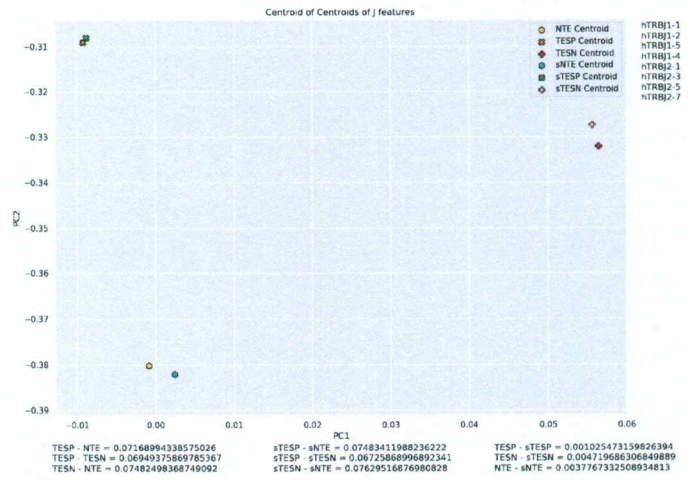
Figure 8 shows the same centroid of centroids seen in Figure 6 plotted alongside the centroid of centroids of the PCA data of only those key features that were identified. This reduction of the data lost almost no information related to the variation of each data set, cutting down the original forty-eight V features to just ten

cutting the original thirteen J features to just eight. Because such little variational information was lost cutting out the other features, these thirteen and eight features that were identified were deemed to be key or driving features, heavily influencing the variation contained within any particular dataset.





(a) Positioning of the Centroids of Centroids of the V features.



(b) Positioning of the Centroids of Centroids of the J features.

**Figure 3.7:** Shows the centroid of centroid positionings for each feature space. Also shows the distances between each point. Centroids calculated from select or key features are labeled with a lowercase 's' i.e. sNTE. The selected or key features are indicated in the upper right corner.

These identified features were shown to play critical role in driving the location of the CoC in the two-dimensional variational space of the first two principal components, which comprise upwards of ninety percent of the explained variance. This indicates that these features likely contain a majority of the information regarding the variation of the CDR3 repertoire. For the three dimensional case, the addition of the third principal component and its associated variation made it so that the 10 V features that could model the data in two dimensions could no longer accurately model the data for the three dimensional case. However, the addition of six more features ['hTRBV25-1', 'hTRBV18', 'hTRBV4-1', 'hTRBV4-2', 'hTRBV4-3', 'hTRBV5-1'] to the original ten that were first selected produced a plot that resembled the plot of all of the V features (Figure 9). It was difficult to confidently say which features were the best fit for the dataset as the addition of any one feature could cause the movement of any or all of the centroids in numerous directions. Those additional six were selected based on the shape that they provided to the data. Of course, the more features that are used the more explained variance is captured and the more accurate the fit to the original data will be.

These observations led to the formulation of the hypothesis that in order to most effectively manipulate the location of a CoC in variational space, the manipulation of the CDR3 content of these selected or key features identified here will likely have the most affect on the variation of the CDR3 repertoire. However, it also stands to reason that if the objective is to restore or "even out" the variation within the CDR3 repertoire, then every feature except these key features identified here should be targeted as to promote the expression of CDR3s within the repertoire that are

not already heavily expressed. The script used to identify these key features is the PTSD\_PCA\_T.py script.

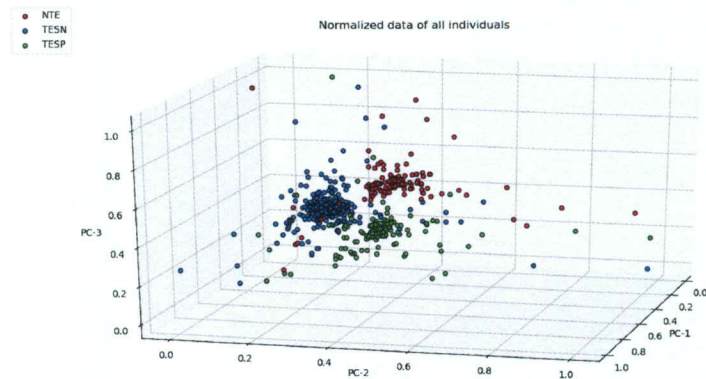
### 3.5 Clustering Analysis

Having shown that the the variation of the CDR3 repertoire provides a distinct separation of the three groups, the next step was to determine if it would be possible to build a classifier which could predict the diagnosis for PTSD of a previously un-diagnosed individual who had been exposed to some form of trauma. Having already established that there is a distinct differentiation in the variation of the CDR3 repertoires of each veteran group, a clustering analysis was done on the data of all of the individuals of each group. The goal of this clustering analysis was to determine if the data of all of the individuals of each group formed discernible clusters and if so, if the clusters could be identified by a clustering algorithm. A clustering algorithm itself could serve as a classifier or be used in tandem with other classification methodologies.

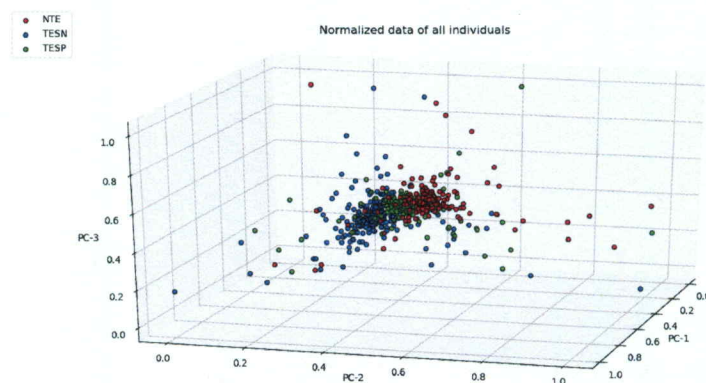
Clustering algorithms were chosen from the sci-kit learn libraries of classifiers which included the KMeans, Spectral, and Agglomerative Clustering algorithms. Other algorithms were available to choose from but these were selected because they look for a user defined, discrete number of clusters. The other algorithms in the scikit-learn library, such as DBSCAN and Birch, will look for and identify as many clusters as they see that fits the data, which could end up being too many or too few clusters.

Figure 9 shows the initial three dimensional mapping of normalized PCA data of both the J and V features of all the individuals of each group. The data was normalized Both mappings show that clear and distinct separations of each group, indicating that the use of clustering algorithms is in fact viable. The V features give more discernible and separable clusters of each group. The J features still show three distinct groupings however they show less separation between each cluster than the V features. The less separation that exists between each cluster the less accurate a clustering algorithm will be at correctly grouping test data or the data of individuals with an unknown PTSD diagnosis.





(a) Plot of the normalized PCA data of the V features.



(b) Plot of the normalized PCA data of the J features.

**Figure 3.8:** The results after running the data of each individual of each group through a PCA and consequent normalization. The use of the V features give clear separation of each group while the use of the J features yielded tighter, less distinguishable clusters.

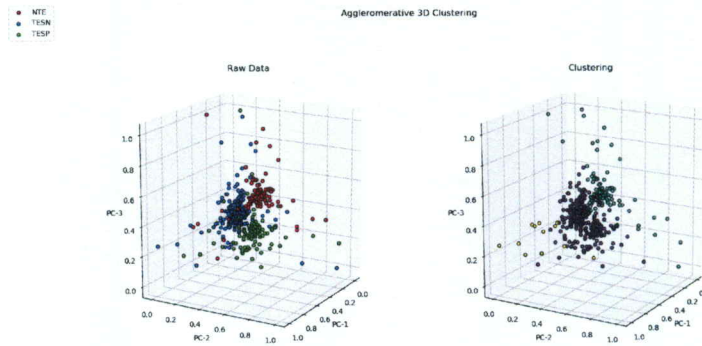


Knowing that the V features do indeed create distinct clusters in the variational space, the three previously selected clustering algorithms can now be tested to judge their performance at picking out the clusters that are known to exist. Figures 10, 11, and 12 show the performance of each of the clusters compared to the natural clustering of the data. The spectral clustering algorithm, which is clearly the only one that got even close to identifying the correct clusters, uses a similarity matrix, a matrix of eigenvectors of the Laplacian transform of the data, and a k-means algorithm (another type of clustering algorithm) in order to derive and define its clusters (Von Luxberg 2007). The other two algorithms could distinguish between the TESN and TESP clusters, instead keying in on outliers as the third cluster. To attempt to quantify the performance of the spectral clustering algorithm on the data the Fowlkes-Mallows index was used as it allows for the evaluation of clusterings when the ground-truth is known. In other words when it is already known which class or group each individual belongs to. The Fowlkes-Mallows index is the geometric mean of the pairwise precision and recall as given by

$$\frac{TP}{\sqrt{(TP + FP)(TP + FN)}}$$

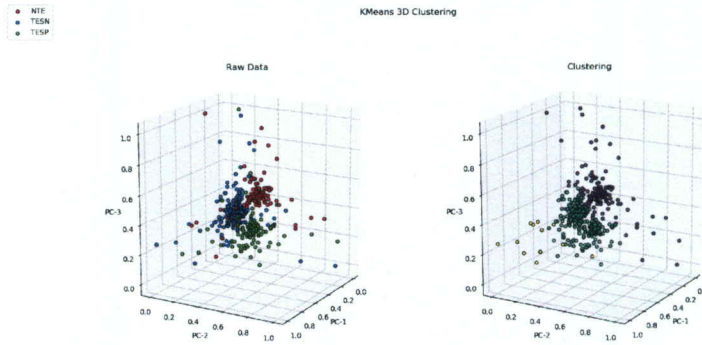
where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives (Fowlkes, Mallows 2007). A higher Fowlkes-Mallows score indicates greater similarity between the clusters and the ground truth. For the spectral clustering of the normalized V features, the Fowlkes-Mallows score for the NTE clusters (ground truth vs algorithm) was calculated to be 0.93768 (on a scale

of 0 to 1), the score for the TESN clusters was 0.86071, and the score for the TESP clusters was 0.92413. These scores indicate that the spectral clustering did in fact do quite well at classifying each group and performed exceptionally at identifying the NTE and TESP clusters. Scoring the spectral clustering algorithm based strictly on accuracy revealed that the algorithm was 88.85% accurate, meaning that it labeled 566 of the 637 points correctly.

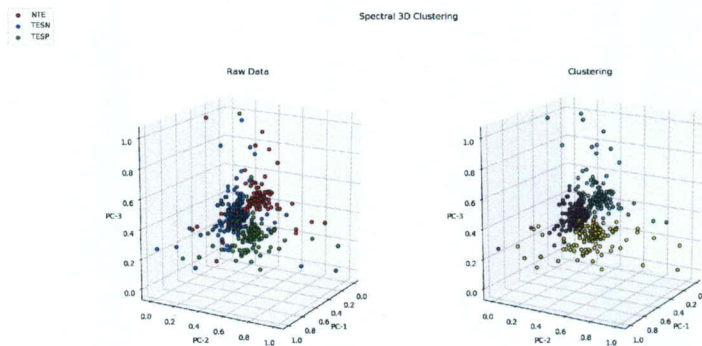


**Figure 3.9:** Result of applying the Agglomerative Clustering algorithm to the data.

This algorithm clearly failed to recognize the existing clusters.



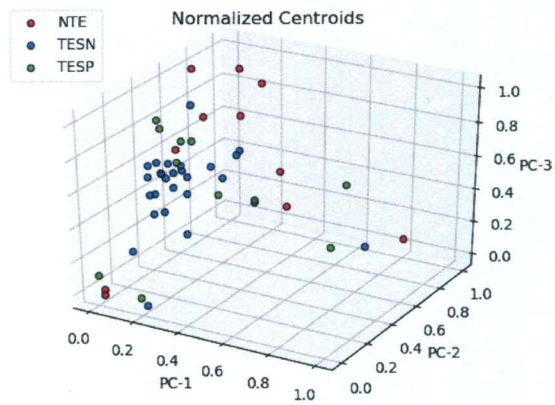
**Figure 3.10:** Result of applying the KMeans Clustering algorithm to the data. This algorithm also could not identify the exists clusters.



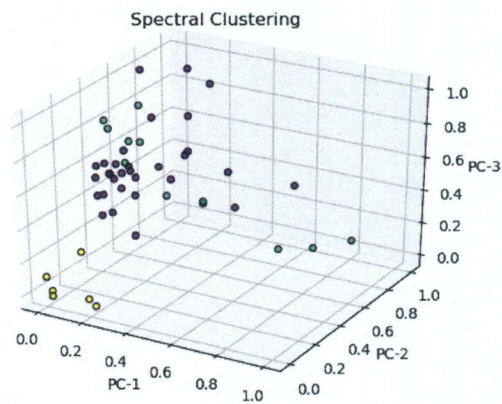
**Figure 3.11:** Result of applying the Spectral Clustering algorithm to the data. All three clusters were clearly identified with no obvious errors.

It must be remembered that all of the PCA data of every individual was used in these calculations, hence the 637 data points. While this methodology revealed the existence of the clusters and allowed the use of the clustering algorithm, this data

cannot be used for outright classification as each data point is not representative of an individual but rather a piece of an individuals data. Because of this some of the 71 points miss classified by the algorithm may be bits and pieces of data from multiple individuals. In order for accurate classification via clustering to occur a method to represent each individual as a single point, such as a centroid, is necessary or a standard is needed stating that it is acceptable to give an individual a certain label if some percentile of an individuals data points are classified as a that label. Due to the small sample size of data in this study neither of these methodologies are possible to implement. Figure 13 shows the lack of obvious clustering of the centroids of the data of individuals of each group and the failing of the spectral clustering to properly classify the centroids into their proper groups.



(a) Plot of the normalized PCA centroids of the V features.



(b) Plot of the normalized PCA centroids of the V features with the Spectral Clustering algorithm applied.

**Figure 3.12:** The Spectral Clustering algorithm failed to properly identify the centroids of any of the groups largely because of a lack of data from which the algorithm can draw conclusions from.



## CHAPTER 4

### DISCUSSION

Through the analysis of the basic statistics, numerical heatmaps, variation, driving features, and clustering of this data on the CDR3 repertoire of forty-nine individuals, it has been shown that there is strong evidence that the CDR3 of the TCR plays a role in the response of an individual to trauma. The initial basic statistical analysis showed that the TESP group had a significantly lower number of CDR3s that comprised their CDR3 repertoire than the NTE and TESN group, with the average TESP CDR3 repertoire being only 73% the size of the average TESN repertoire. The TESN group was also seen to possess the highest number of CDR3s within their repertoire which may suggest that the more robust an individual's CDR3 repertoire, the more resilient they are to trauma with a lower risk of developing PTSD. The numerical heatmaps confirmed the results of the statistical analysis, showing the almost every V and J region (which are largely responsible for the diversity of the CDR3 repertoire) produced more CDR3s in the TESN group compared to the TESP group. Unsurprisingly this was also the case for the TESN group compared to the NTE group. However, when comparing the heatmaps of the NTE group and the TESP group there are multiple different V and J regions that are highly represented

in one group but not the other. The results of the heatmap comparisons confirm that the TESN group has the greatest CDR3 diversity and expression of any of the groups providing more evidence for the hypothesis that the more robust an individual's CDR3 repertoire the greater their resilience to the effects of being exposed to trauma.

The analysis on the variation within the V region of each group indicated that, on average, the variation of the NTE and TESN groups were quite similar, even out to three dimensions with more than ninety percent of the variation captured by three principal components. What this suggests is that, due to the NTE and TESN groups being variationally similar to one another with neither group suffering from symptoms of PTSD caused by exposure to trauma, it is possible that altering the variation of the CDR3 repertoire of a TESP individual to more closely match that of an average TESN individual could lead to a reduction in their symptoms related to PTSD. The question is then how might the variation of the CDR3 repertoire be targeted to achieve such an effect? With the identification of the key V and J regions that are the main drivers behind the positioning of each centroid in the variational space, it can be said that these V and J regions then heavily contribute to the variation of each CDR3 repertoire. Thus in order to alter the variation of the CDR3 repertoire of an individual either the expression of the CDR3s that are a product of these key V and J features must be altered or the expression of CDR3s produced by every other V and J region must be altered. Given that the TESP individuals had far fewer CDR3s within their repertoire compared to TESN individuals, the most obvious first step would be to observe the effect on the variation of the CDR3 repertoire of a TESP individual caused by up-regulating the expression of CDR3s produced by V and J

regions outside of the identified key regions. This could be done via simulation as to avoid any possible unforeseen side effects on human subjects, as the mechanism by which the adaptive immune system may help regulate the bodies response to trauma is largely not understood.

This study also provides an example for the possible classification of PTSD for un-diagnosed individuals such as those who have been exposed to trauma but have not been properly screened or interviewed. At present there are no formal objective methods for diagnosing PTSD, the screenings and interviews are the only methods accepted in the process of clinically diagnosing PTSD (Steel et al. 2011). Clustering the variational data of the V features of each individual in three dimensions showed that each group formed a distinct cluster and consequently each cluster could be identified by a spectral clustering algorithm. With a larger dataset and further model definition and development it would be possible to use such an algorithm as a classifier for data from an individual who has been exposed to trauma but has not been screened for PTSD. This model definition and development would include determining the most efficient utilization of such a classifier and the methods of data acquisition and processing that would be most effective for use with the classifier. For example the model definition could involve setting a threshold above which, if the required percentile of the data points that comprise the whole of an individuals dataset were accurately classified as TESN or TESP, it would be acceptable to give the individual that same classification. For example, if an individual exposed to trauma were tested and the clustering algorithm labeled 90% of the individuals data points as TESN, it is then likely that the individual does not have PTSD despite

their exposure to trauma. Of course, additional research is required to help validate and improve upon this analysis in addition to all previously discussed analyses.

There are inherent issues with this study which may impact the presented results, the most notable being the small sample size. A larger study with a significantly larger sample size needs to be done in order to help validate the results presented in this study. A larger sample size will provide a much better representation of the the results or confirm the results found in the present study. However, the significance of the given results must not be understated. Future studies of the effect of trauma on the CDR3 repertoire should be conducted with much larger sample sizes to validate the results of the current study. If these results are in fact validated then it may be possible to design pharmaceuticals to target specific CDR3s in order to increase or decrease their expression levels. In this way it would be possible to alter the variation of the CDR3 repertoire of an individual. This would allow for the testing of the hypothesis that altering the variation of the CDR3 repertoire of an individual could lead to the alleviation of symptoms related to PTSD. It should be noted that although this data only provides information on the effect of combat trauma on the CDR3 repertoire of an individual, it is very likely that very similar results would be found on a study on the affect of trauma on the CDR3 repertoire of first responders and victims of abuse.



## CHAPTER 5

### CONCLUSION

Posttraumatic stress disorder is unfortunately an inevitability for roughly eight percent of the general population of Americans, with that number increasing to upwards of twenty percent for individuals who have certain occupations or a victim to devastating circumstance. The current landscape of PTSD detection, screening, diagnosis, and treatment lacks qualitative assessments that can aide in such diagnoses. This study provides a basis from which the adaptive immune system, specifically the CDR3 protein segment of the T cell receptor, can be tentatively linked to the bodies response to exposure to trauma. These results provide support for the continued research into the validation of these methods and testing of the proposed hypotheses: the CDR3 repertoire plays a role in the bodies response to trauma, the variation of the CDR3 repertoire has a role in the expression of symptoms related to PTSD, and that the CDR3 repertoire may be used in the classification of PTSD.



## CHAPTER 6

### REFERENCES

- American Psychological Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.
- American Psychological Association. (2019). *PTSD Assessment Instruments*. Retrieved from <https://www.apa.org/ptsd-guideline/assessment>
- Blake, D. D., Keane, T. M., Wine, P. R., Mora, C., Taylor, K. L., & Lyons, J. A. (1990). Prevalence of PTSD symptoms in combat veterans seeking medical treatment. *Journal of Traumatic Stress, 3*(1), 15–27. doi: 10.1007/BF00975133
- Boscarino, J. A. (2004). Posttraumatic stress disorder and physical illness: Results from clinical and epidemiologic studies. *Annals of the New York Academy of Sciences, 1032*, 141–153. doi: 10.1196/annals.1314.011
- Boyko, E. J., Jacobson, I. G., Smith, B., Ryan, M. A., Hooper, T. I., Amoroso, P. J., ... Smith, T. C. (2010). Risk of diabetes in U.S. military service members in relation to combat deployment and mental health. *Diabetes Care, 33*(8), 1771–1777. doi: 10.2337/dc10-0296
- Brainline. (2018). *DSM-5 Criteria for PTSD*. Retrieved from <https://www.brainline.org/article/dsm-5-criteria-ptsd>
- Breen, M. S., Maihofer, A. X., Glatt, S. J., Tylee, D. S., Chandler, S. D., Tsuang, M. T., ... Woelk, C. H. (2015). Gene networks specific for innate immunity define post-traumatic stress disorder. *Molecular Psychiatry, 20*(12), 1538–1545. doi: 10.1038/mp.2015.9
- Brett, E. A., Spitzer, R. L., & Williams, J. B. (1988). DSM-III-R Criteria for Posttraumatic Stress Disorder. *American Journal of Psychiatry, 145*, 1232–1236.
- Caspani, G., Corbet Burcher, G., Garralda, M. E., Cooper, M., Pierce, C. M., Als, L. C., & Nadel, S. (2018). Inflammation and psychopathology in children following PICU admission: An exploratory study. *Evidence-Based Mental Health, 21*(4), 139–144. doi: 10.1136/ebmental-2018-300027

- Coughlin, S. S. (2013). Chapter 7 Post-Traumatic Stress Disorder and Cardiovascular Disease . *Post-Traumatic Stress Disorder and Chronic Health Conditions*, 164–170. doi: 10.2105/9780875530161ch07
- Danise, E., Heppner, P., Furberg, H., Goldberg, J., Buchwald, D., & Afari, N. (2013). The Comorbidity of Self-Reported Chronic Fatigue Syndrome, Posttraumatic Stress Disorder, and Traumatic Symptoms. *Journal of Psychosomatics*, 53(3), 250–257. doi: 10.1038/mp.2011.182.doi
- Eraly, S. A., Nievergelt, C. M., Maihofer, A. X., Barkauskas, D. A., Biswas, N., Agorastos, A., ... Baker, D. G. (2014). Assessment of plasma C-Reactive protein as a biomarker of posttraumatic stress disorder risk. *JAMA Psychiatry*, 71(4), 423–431. doi: 10.1001/jamapsychiatry.2013.4374
- Flory, J. D., & Yehuda, R. (2015). Comorbidity between post-traumatic stress disorder and major depressive disorder: alternative explanations and treatment considerations. *Dialogues in clinical neuroscience*, 17(2), 141–50. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/26246789>{\%}0A<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4518698>
- Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383), 553–569. Retrieved from <https://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10478008> doi: 10.1080/01621459.1983.10478008
- Friedman, M. J. (2004). Acknowledging the Psychiatric Cost of War. *New England Journal of Medicine*, 351(1), 75–77. doi: 10.1056/nejme048129
- Friedman, M. J. (2018). *PTSD History and Overview*. Retrieved from [https://www.ptsd.va.gov/professional/treat/essentials/history{\\\_}ptsd.asp](https://www.ptsd.va.gov/professional/treat/essentials/history{\_}ptsd.asp)
- Gradus, J. L. (2018). *Epidemiology of PTSD*. Retrieved from <https://www.ptsd.va.gov/professional/treat/essentials/epidemiology.asp{\#}three>
- Green, B., Lindy, J., & Grace, M. (1985). Posttraumatic stress disorder: toward DSM-IV. *Journal of Nervous and Mental Disease*, 173, 406–411.
- Heinzelmann, M., & Gill, J. (2013). Epigenetic Mechanisms Shape the Biological Response to Trauma and Risk for PTSD: A Critical Review. *Nursing Research and Practice*, 2013, 1–10. doi: 10.1155/2013/417010
- Kelsall, H. L., McKenzie, D. P., Forbes, A. B., Roberts, M. H., Urquhart, D. M., & Sim, M. R. (2014). Pain-related musculoskeletal disorders, psychological comorbidity, and the relationship with physical and mental well-being in Gulf War veterans. *Pain*, 155(4), 685–692. Retrieved from <http://dx.doi.org/10.1016/j.pain.2013.12.025> doi: 10.1016/j.pain.2013.12.025



- Klikauer, T. (2016). Scikit-learn: Machine Learning in Python. *TripleC*, 14(1), 260–264. Retrieved from <http://dl.acm.org/citation.cfm?id=2078195> doi: 10.1007/s13398-014-0173-7.2
- Lancaster, C., Teeters, J., Gros, D., & Back, S. (2016). Posttraumatic Stress Disorder: Overview of Evidence-Based Assessment and Treatment. *Journal of Clinical Medicine*, 5(11), 105. doi: 10.3390/jcm5110105
- Lindqvist, D., Mellon, S. H., Dhabhar, F. S., Yehuda, R., Grenon, S. M., Flory, J. D., ... Wolkowitz, O. M. (2017). Increased circulating blood cell counts in combat-related PTSD: Associations with inflammation and PTSD severity. *Psychiatry Research*, 258, 330–336. Retrieved from <http://dx.doi.org/10.1016/j.psychres.2017.08.052> doi: 10.1016/j.psychres.2017.08.052
- Maguen, S., Madden, E., Cohen, B., Bertenthal, D., & Seal, K. (2014). Association of mental health problems with gastrointestinal disorders in Iraq and Afghanistan veterans. *Depression and Anxiety*, 31(2), 160–165. doi: 10.1002/da.22072
- McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th python in science conference* (pp. 51–56).
- Miller, G. E., Chen, E., Cole, S. W., Bailey, M. T., Powell, N. D., Kobor, M. S., ... Arevalo, J. M. G. (2013). Social stress up-regulates inflammatory gene expression in the leukocyte transcriptome via  $\alpha$ -adrenergic induction of myelopoiesis. *Proceedings of the National Academy of Sciences*, 110(41), 16574–16579. doi: 10.1073/pnas.1310655110
- Murphy, K., & Weaver, C. (2017). *Janeway Immunobiology*.
- National Institute of Mental Health. (2016). *Post-Traumatic Stress Disorder*. Retrieved from <https://www.nimh.nih.gov/health/topics/post-traumatic-stress-disorder-ptsd/index.shtml#part145373>
- Neumann, L., & Buskila, D. (2003). Epidemiology of fibromyalgia. *Current pain and headache reports*, 7(5), 362–8. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/12946289>
- Nievergelt, C. M., Baker, D. G., Agorastos, A., Barkauskas, D. A., Maihofer, A. X., Eraly, S. A., ... Biswas, N. (2014). Assessment of Plasma C-Reactive Protein as a Biomarker of Posttraumatic Stress Disorder Risk. *JAMA Psychiatry*, 71(4), 423. doi: 10.1001/jamapsychiatry.2013.4374
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., & Thirion, B. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

- Ramchad, R., Schell, T. L., Karney, B. R., Osiall, K. C., Burns, R. M., & Caldarone, L. B. (2010). Disparate Prevalence Estimates of PTSD Among Service Members Who Served in Iraq and Afghanistan: Possible Explanations. *Journal of Traumatic Stress, 23*(1), 59–68. doi: 10.1002/jts
- Richardson, L. K., Frueh, B. C., & Acierno, R. (2010). Prevalence Estimates of Combat-Related PTSD: A Critical Review. *The Australian and New Zealand journal of psychiatry, 44*(1), 4–19. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2891773/> doi: 10.3109/00048670903393597
- Sager, H. B., Zaltsman, A., Vinegoni, C., Lin, C. P., Fricchione, G. L., Courties, G., ... Bode, C. (2014). Chronic variable stress activates hematopoietic stem cells. *Nature Medicine, 20*(7), 754–758. Retrieved from <http://dx.doi.org/10.1038/nm.3589> doi: 10.1038/nm.3589
- Solomon, S. D., & Canino, G. J. (1990). Appropriateness of DSM-III-R criteria for posttraumatic stress disorder. *Comprehensive Psychiatry, 31*(3), 227–237. doi: 10.1016/0010-440X(90)90006-E
- Steel, J. L., Dunlavy, A. C., Stillman, J., & Pape, H. C. (2011). Measuring depression and PTSD after trauma: Common scales and checklists. *Injury, 42*(3), 288–300. Retrieved from <http://dx.doi.org/10.1016/j.injury.2010.11.045> doi: 10.1016/j.injury.2010.11.045
- U.S. Department of Veterans Affairs. (2019). *PTSD Treatment Basics*. Retrieved from [https://www.ptsd.va.gov/understand/tx/tx\\_basics.asp](https://www.ptsd.va.gov/understand/tx/tx_basics.asp)
- Van der walt, S., Colbert, C., & Varoquaux, G. (2011). The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering, 13*, 22–30. doi: 10.1109/MCSE.2011.37
- Vieweg, W. V. R., Julius, D. A., Fernandez, A., Beatty-Brooks, M., Hettema, J. M., & Pandurangi, A. K. (2006). Posttraumatic Stress Disorder: Clinical Features, Pathophysiology, and Treatment. *American Journal of Medicine, 119*(5), 383–390. doi: 10.1016/j.amjmed.2005.09.027
- Vinholt, P. J., Hvas, A. M., Frederiksen, H., Bathum, L., Jørgensen, M. K., & Nybo, M. (2016). Platelet count is associated with cardiovascular disease, cancer and mortality: A population-based cohort study. *Thrombosis Research, 148*, 136–142. Retrieved from <http://dx.doi.org/10.1016/j.thromres.2016.08.012> doi: 10.1016/j.thromres.2016.08.012
- Von Luxburg, U. (2007). *A Tutorial on Spectral Clustering* (Tech. Rep.). doi: 10.1.1.165.9323

von Känel, R., Hepp, U., Kraemer, B., Traber, R., Keel, M., Mica, L., & Schnyder, U. (2007). Evidence for low-grade systemic proinflammatory activity in patients with posttraumatic stress disorder. *Journal of Psychiatric Research*, *41*(9), 744–752. doi: 10.1016/j.jpsychires.2006.06.009

Yehuda, R., & LeDoux, J. (2007). Response Variation following Trauma: A Translational Neuroscience Approach to Understanding PTSD. *Neuron*, *56*(1), 19–32. doi: 10.1016/j.neuron.2007.09.006



## APPENDICES

# APPENDIX A

## INDIVIDUAL HEATMAPS

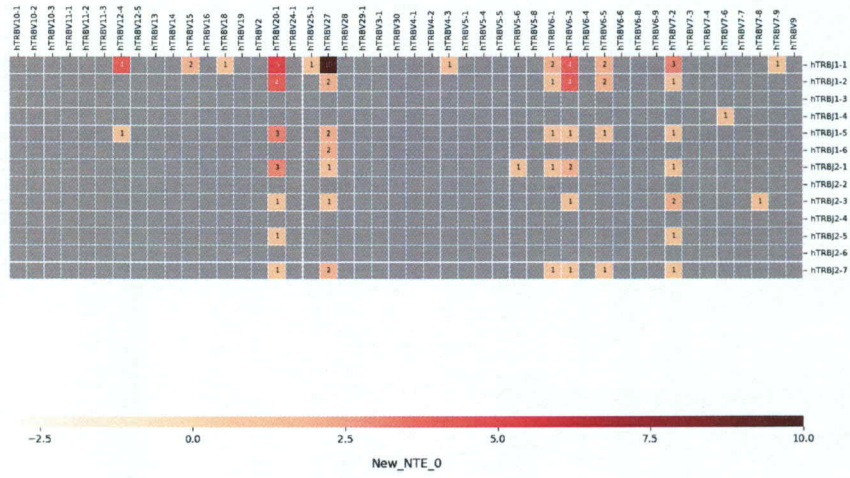


Figure A.1: NTE 1

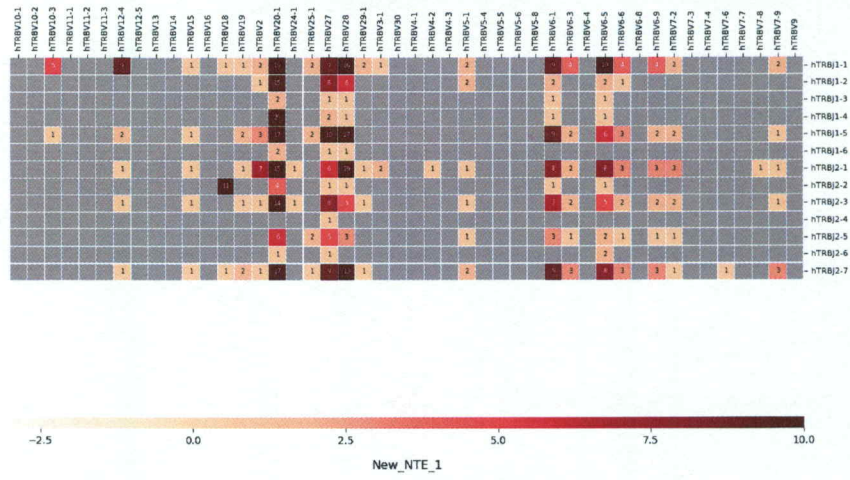


Figure A.2: NTE 2

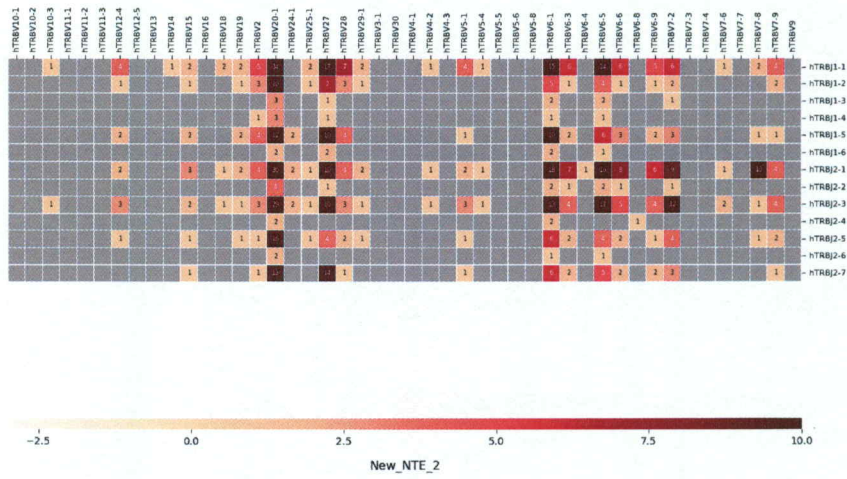


Figure A.3: NTE 3





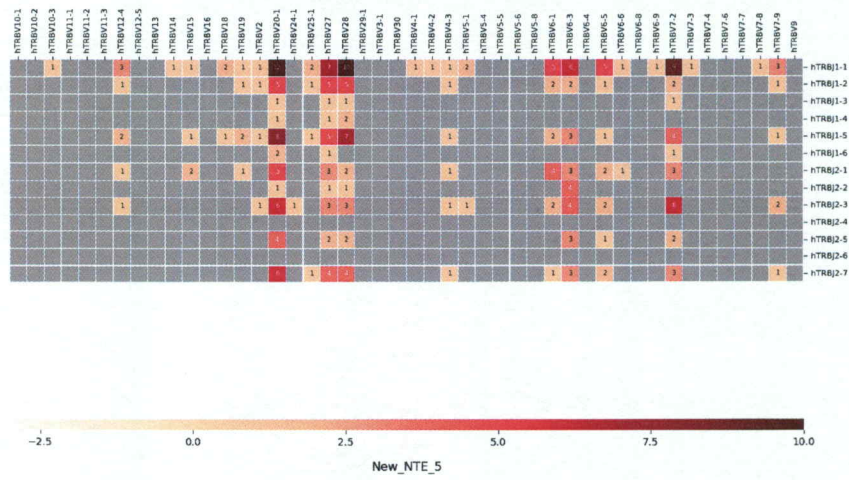


Figure A.6: NTE 6

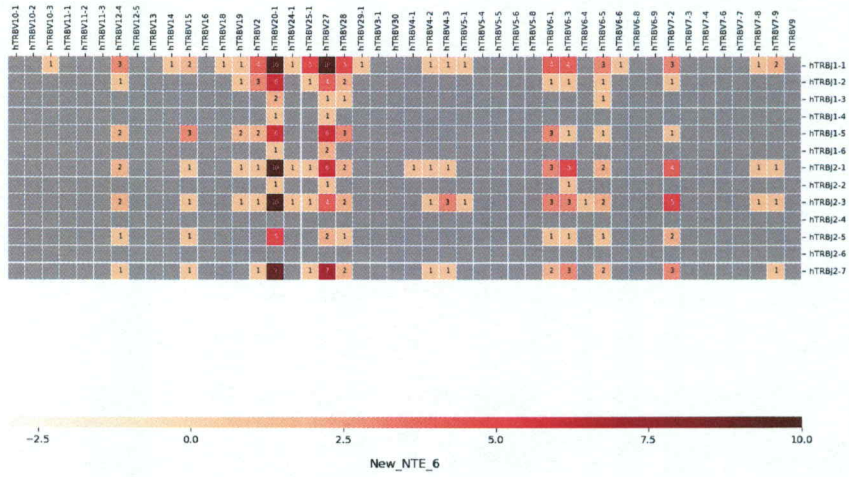


Figure A.7: NTE 7



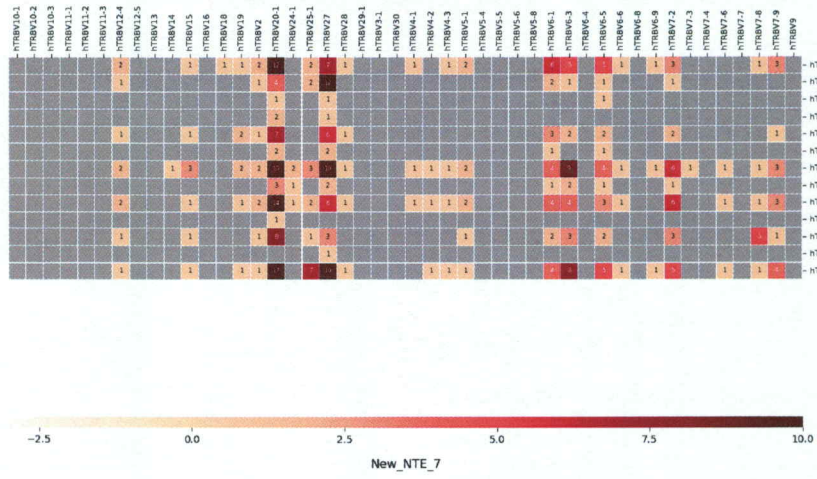


Figure A.8: NTE 8

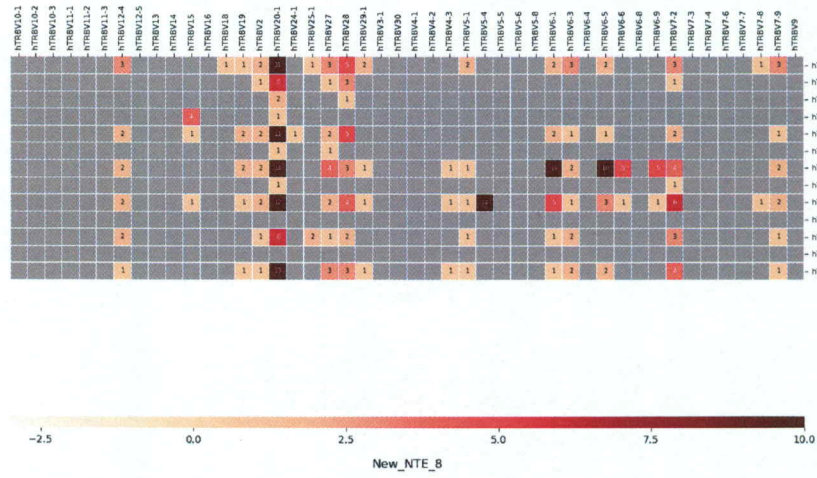


Figure A.9: NTE 9



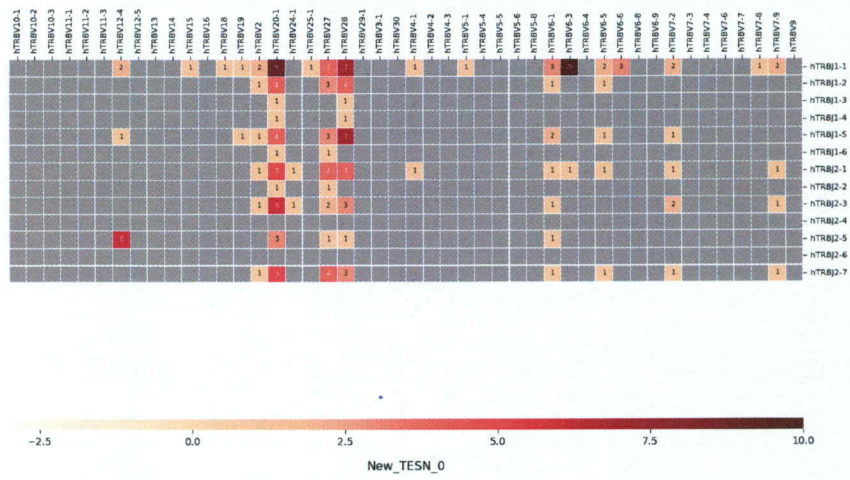


Figure A.12: TESN 1

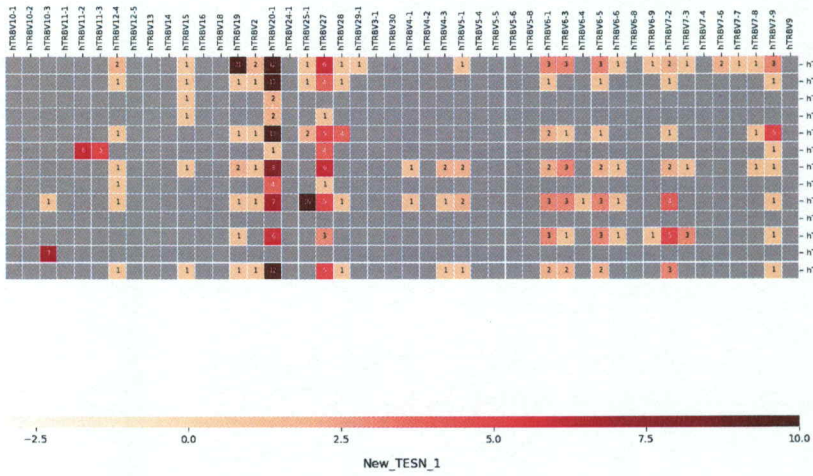


Figure A.13: TESN 2



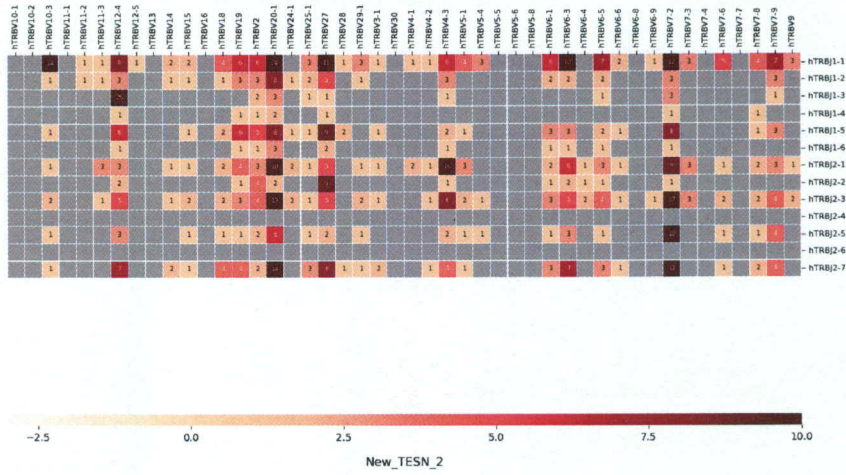


Figure A.14: TESN 3

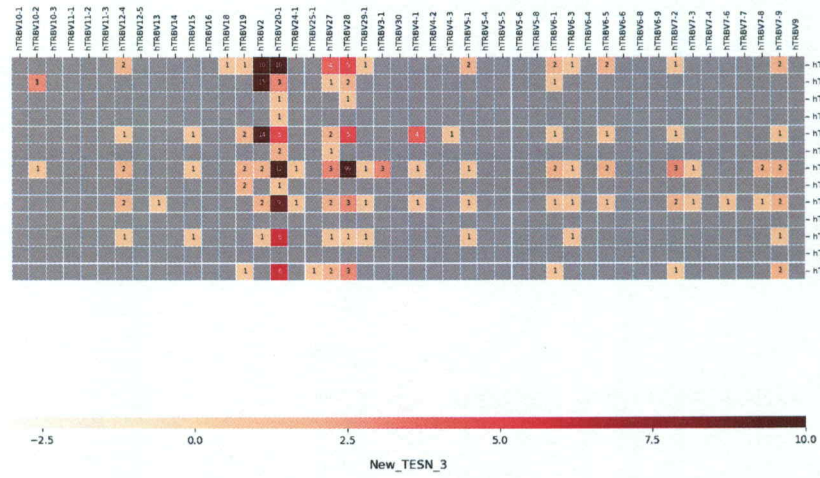


Figure A.15: TESN 4



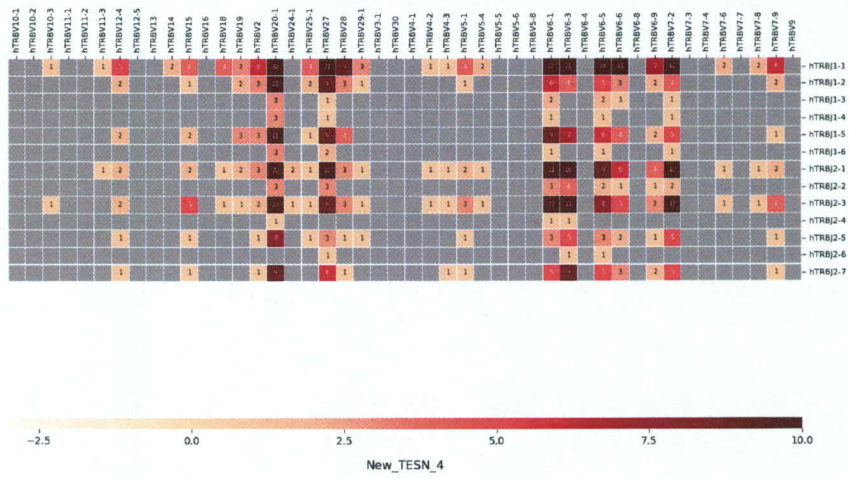


Figure A.16: TESN 5

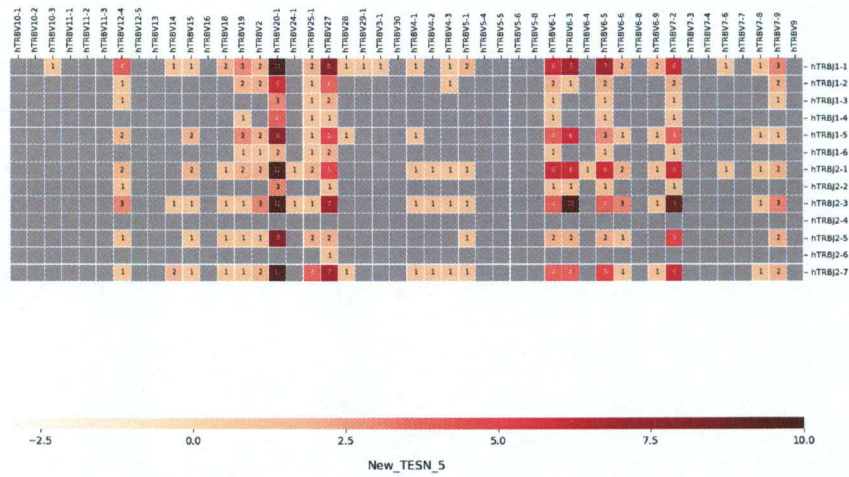


Figure A.17: TESN 6

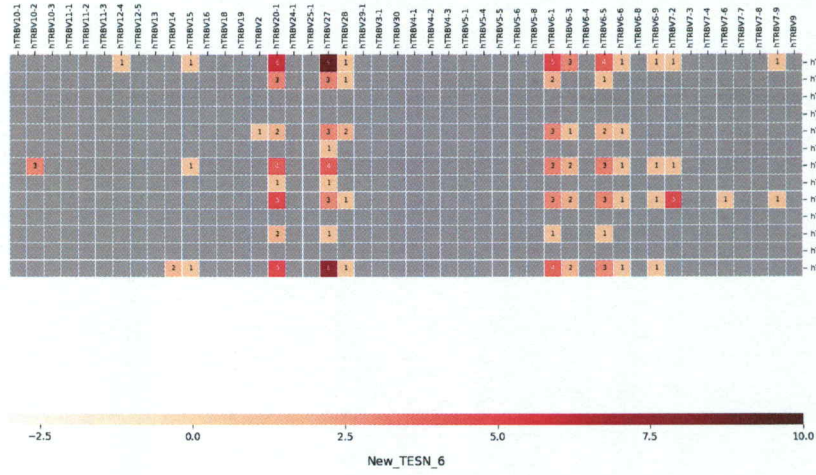


Figure A.18: TESN 7

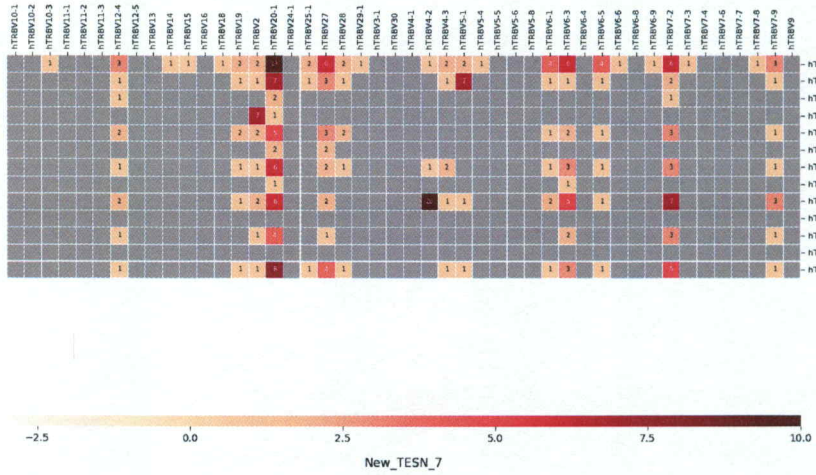


Figure A.19: TESN 8

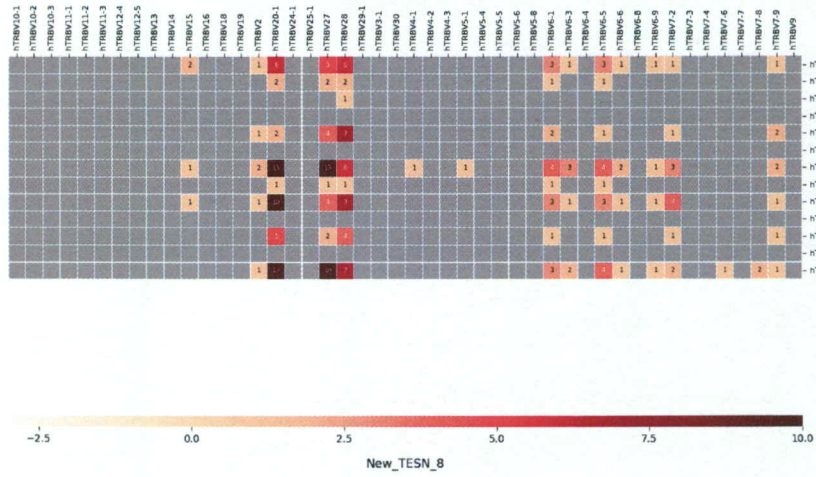


Figure A.20: TESN 9

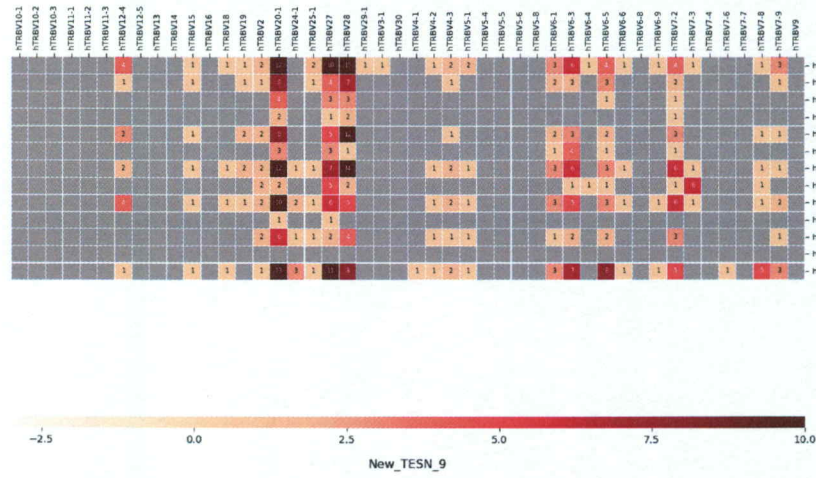


Figure A.21: TESN 10



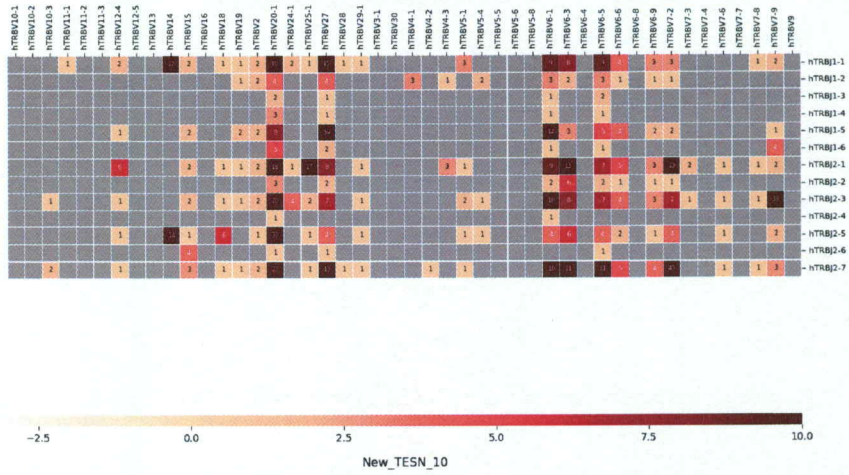


Figure A.22: TESN 11

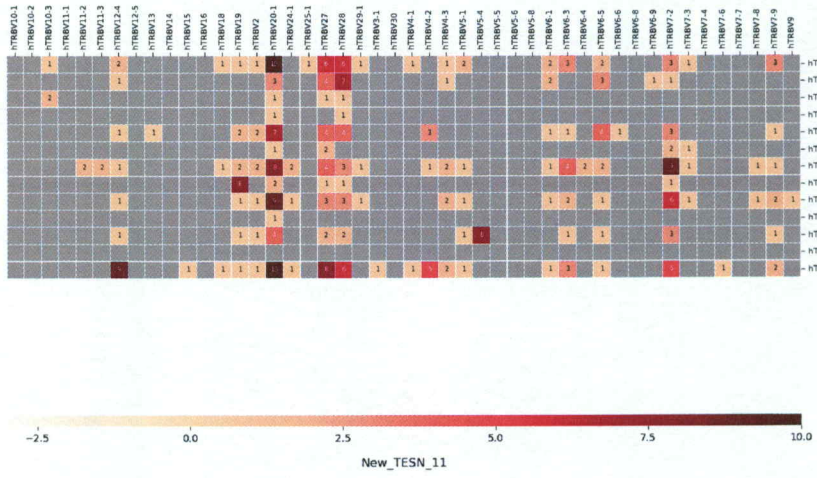
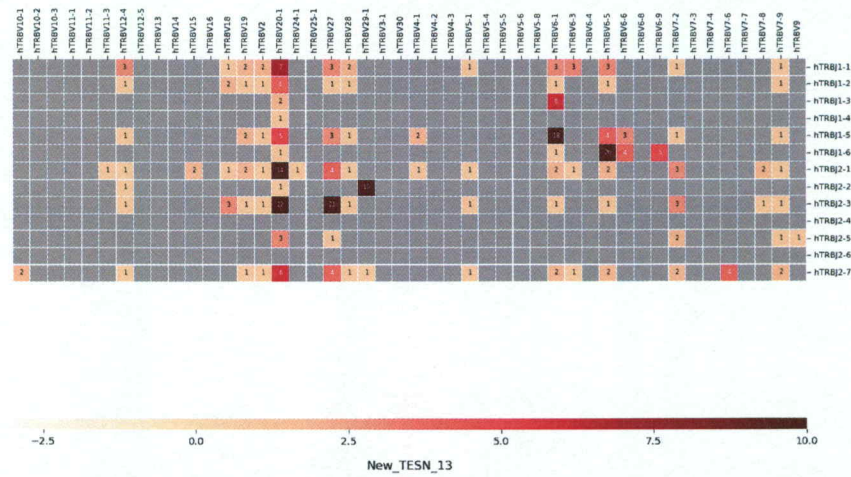
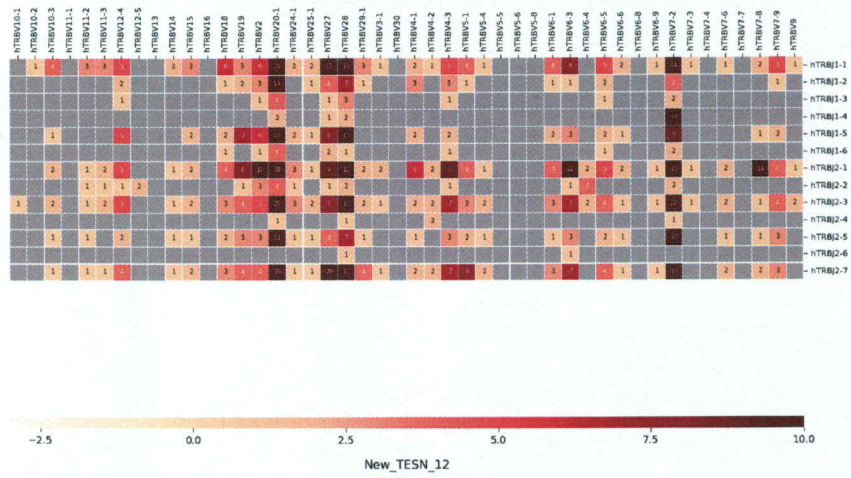


Figure A.23: TESN 12





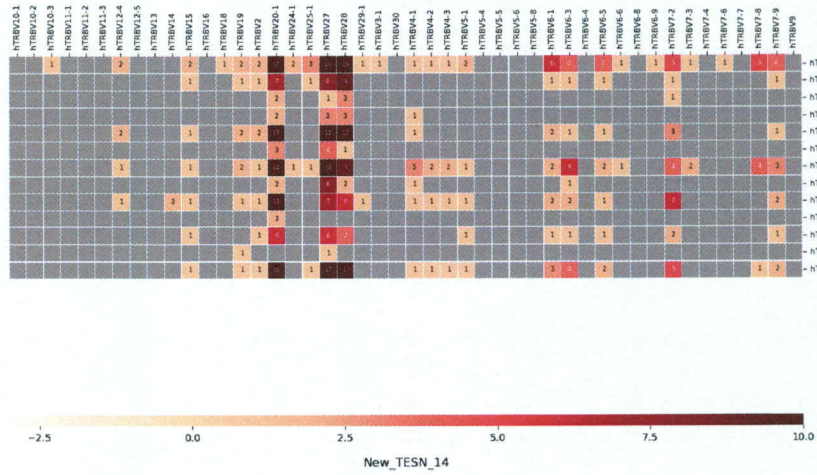


Figure A.26: TESN 14

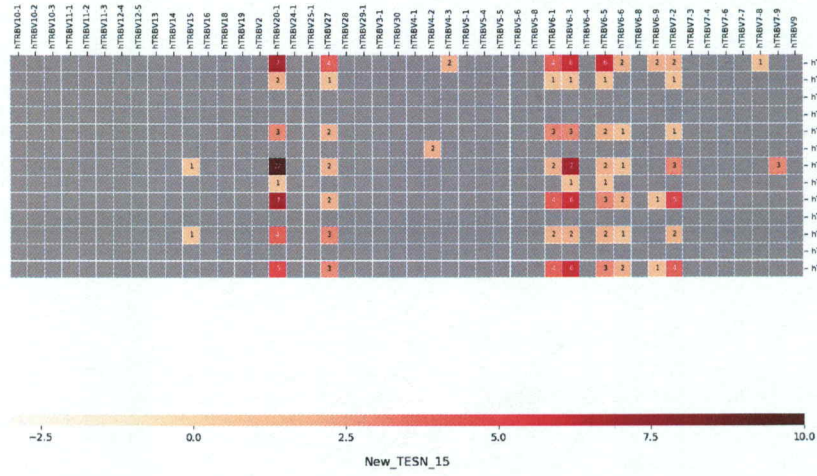


Figure A.27: TESN 15

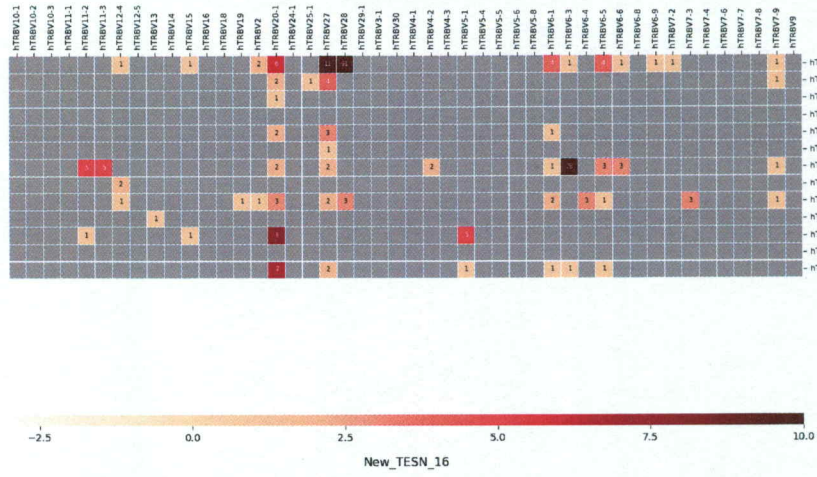


Figure A.28: TESN 17

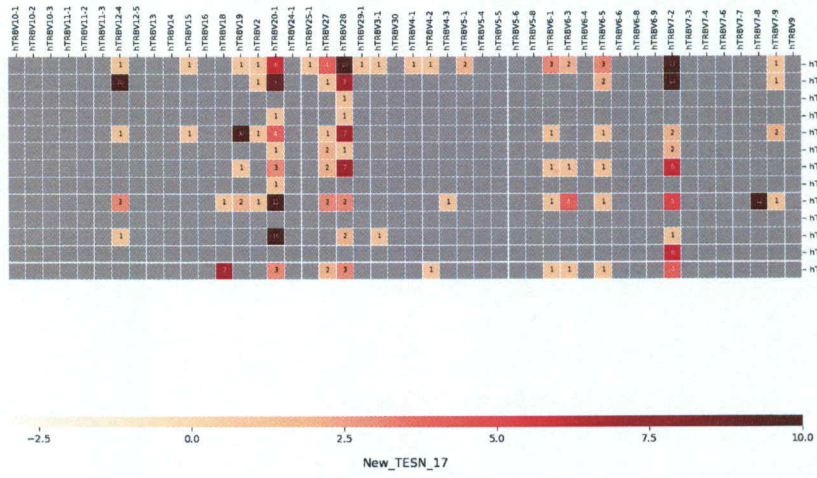


Figure A.29: TESN 18



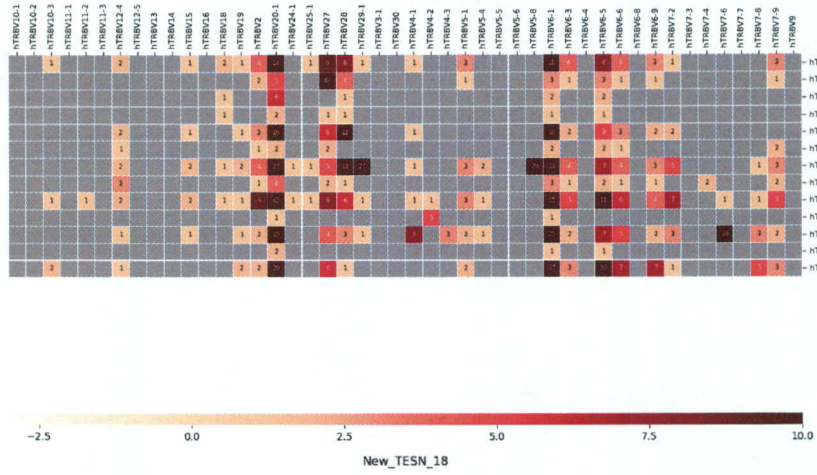


Figure A.30: TESN 19

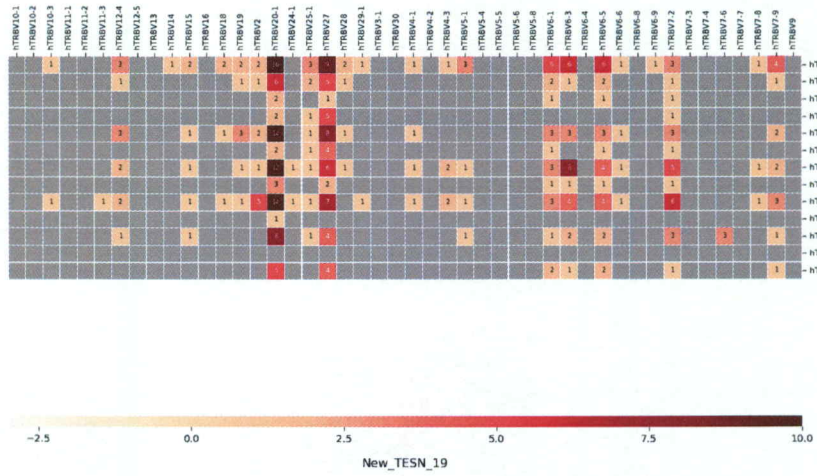


Figure A.31: TESN 20



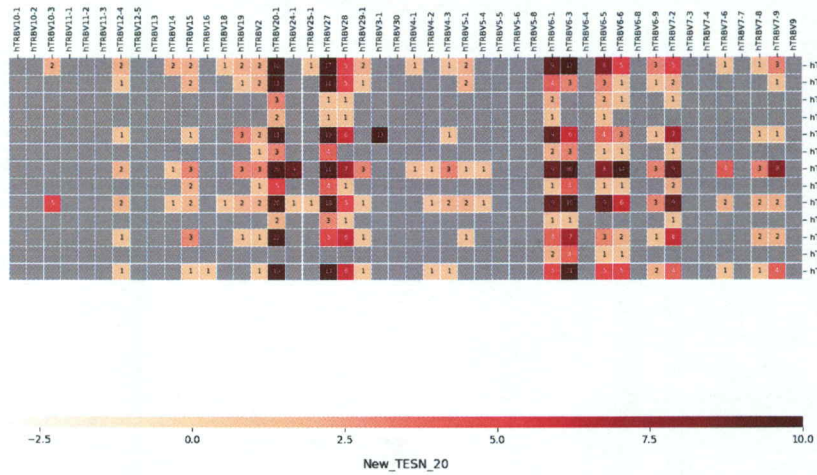


Figure A.32: TESN 21

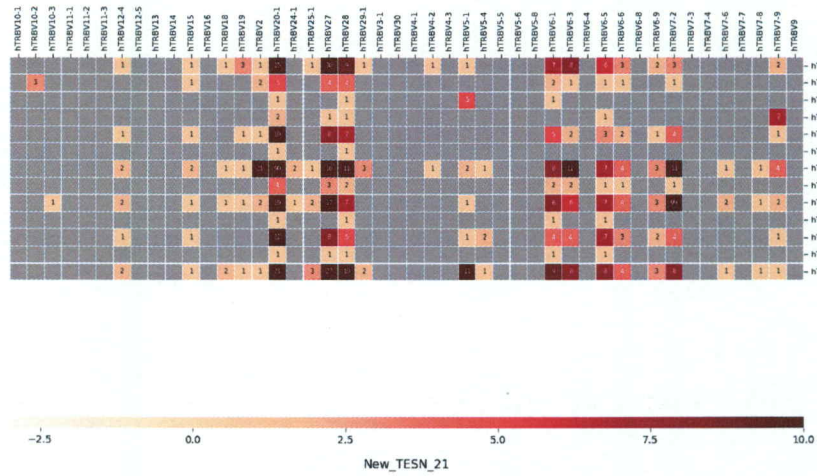


Figure A.33: TESN 22

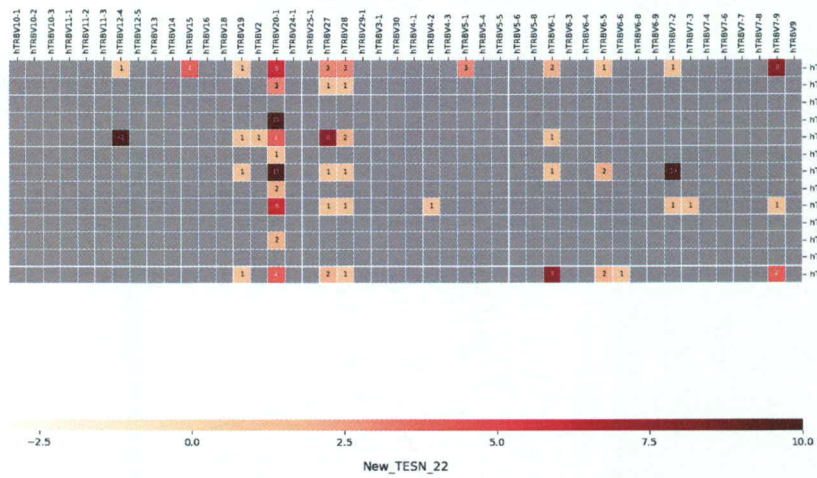


Figure A.34: TESN 23

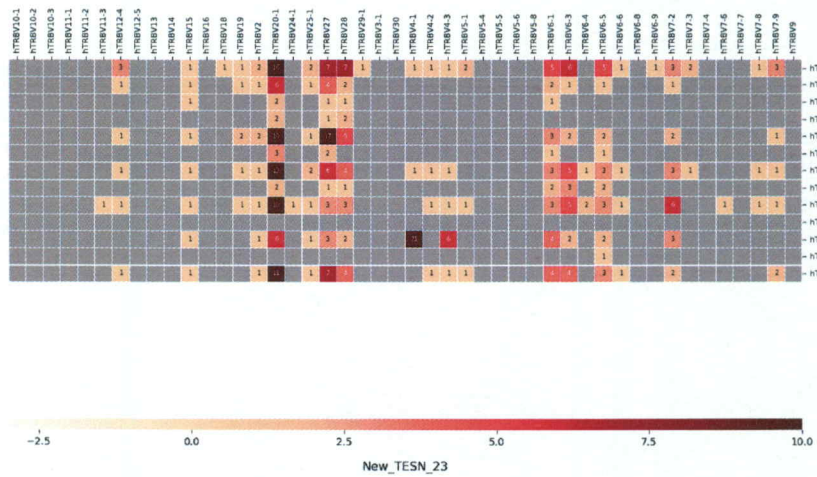


Figure A.35: TESN 24

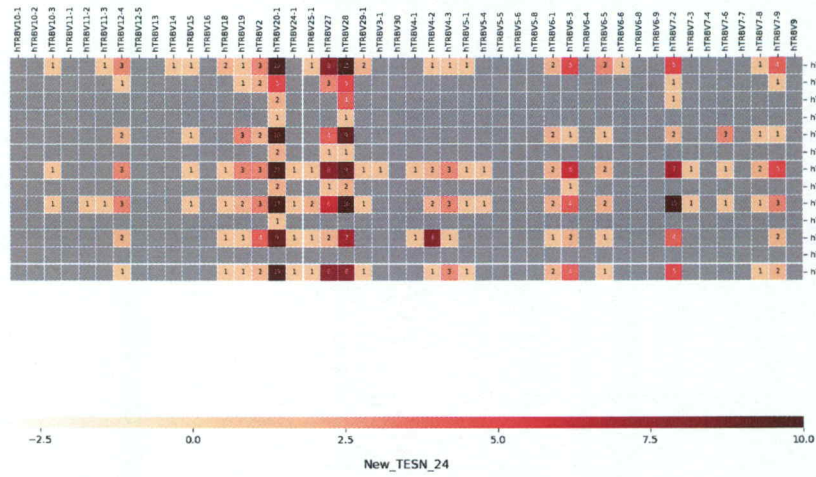


Figure A.36: TESN 25

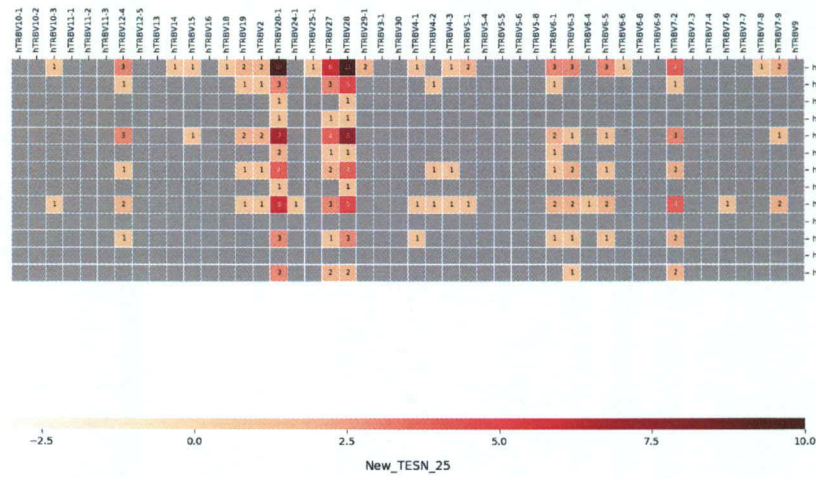


Figure A.37: TESN 26

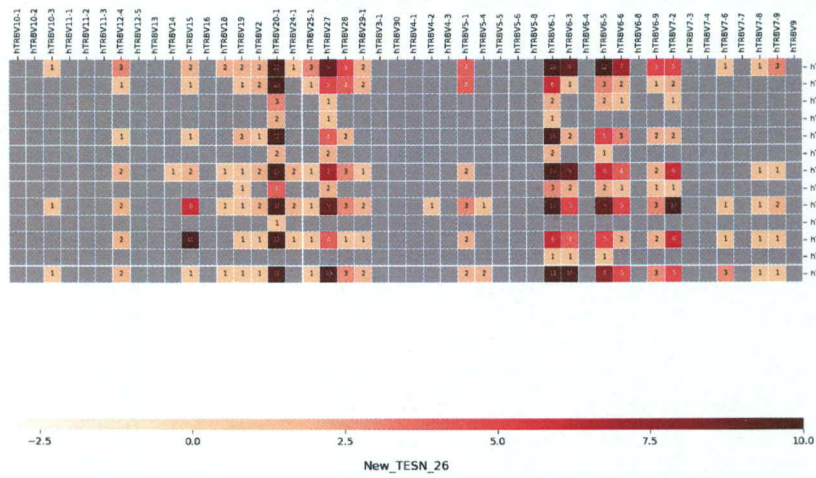


Figure A.38: TESN 27

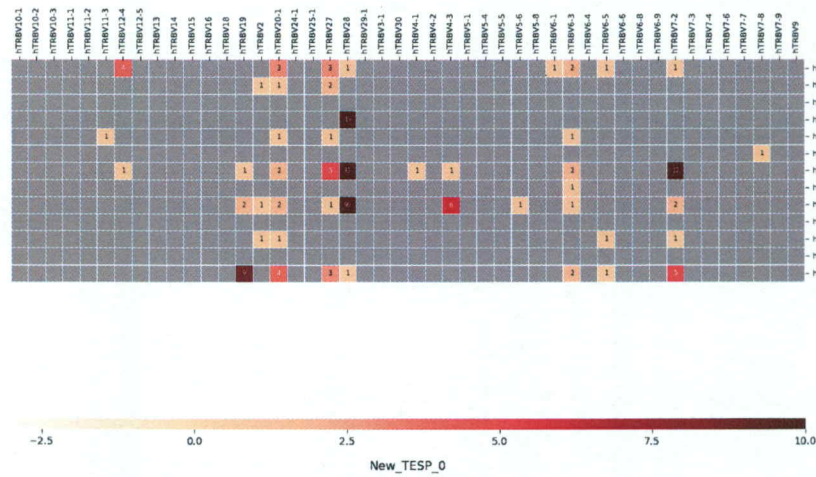


Figure A.39: TESP 1



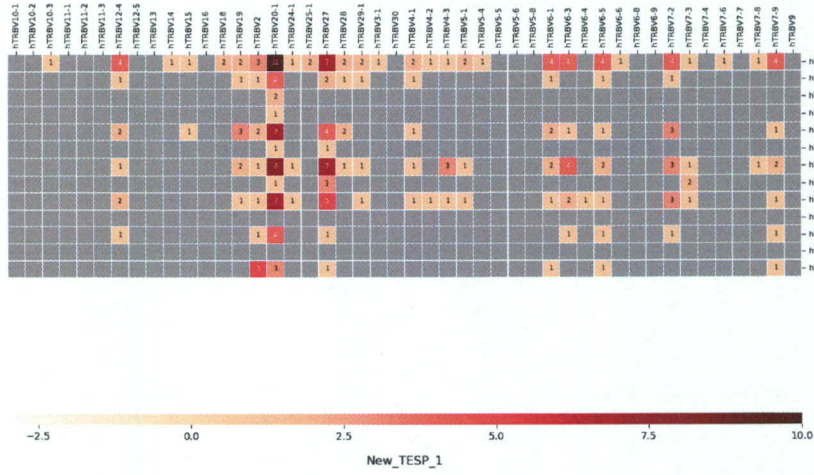


Figure A.40: TESP 2

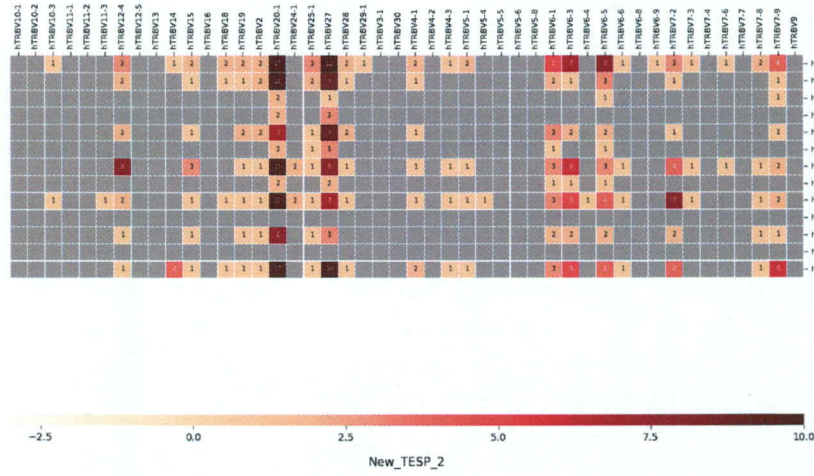


Figure A.41: TESP 3

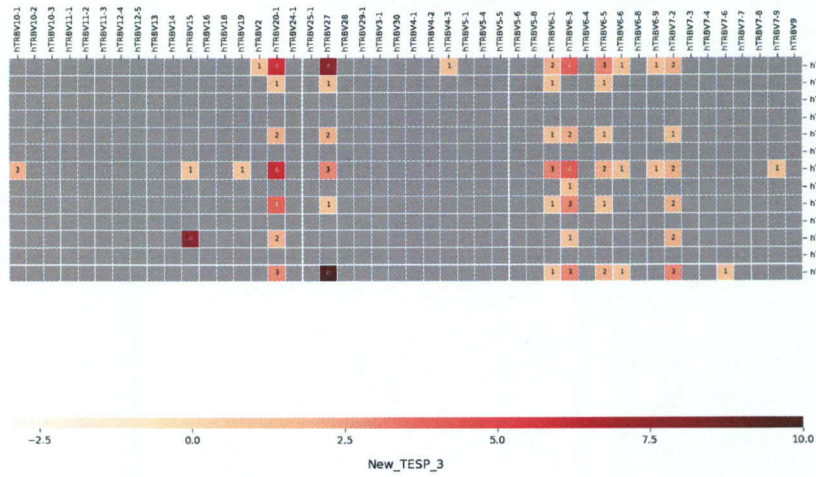


Figure A.42: TESP 4

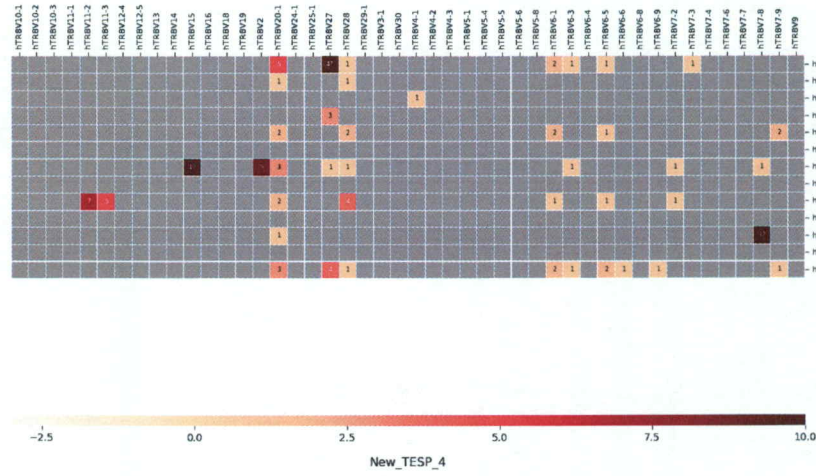


Figure A.43: TESP 5

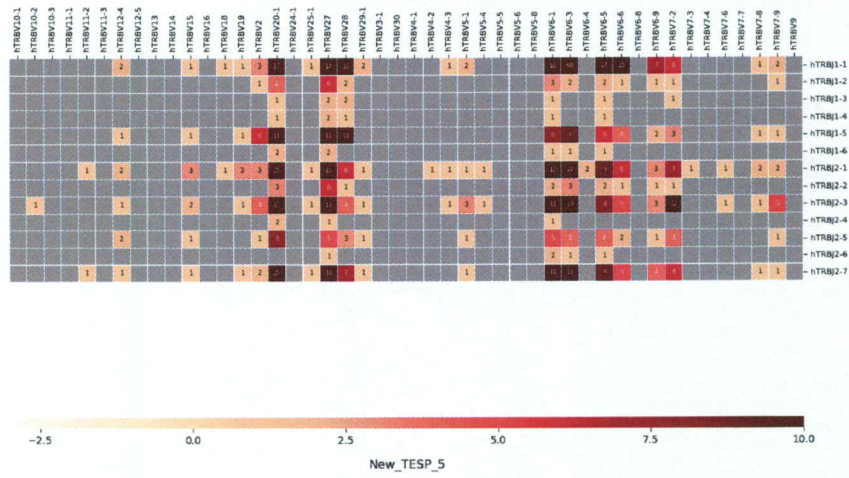


Figure A.44: TESP 6

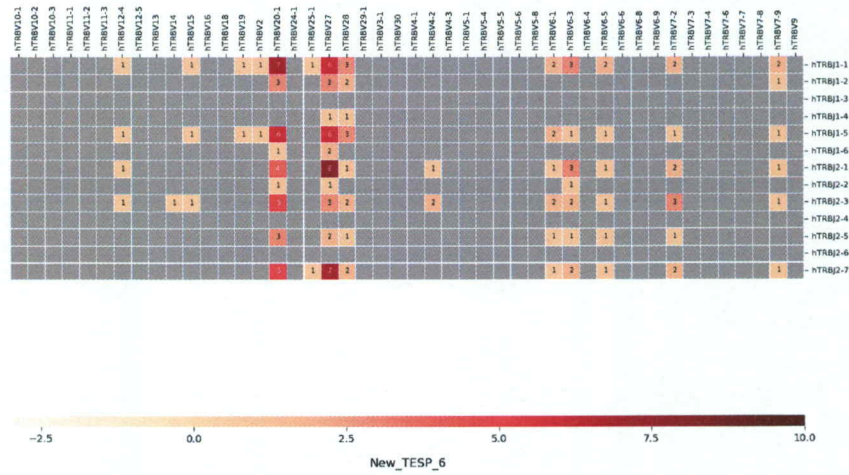


Figure A.45: TESP 7

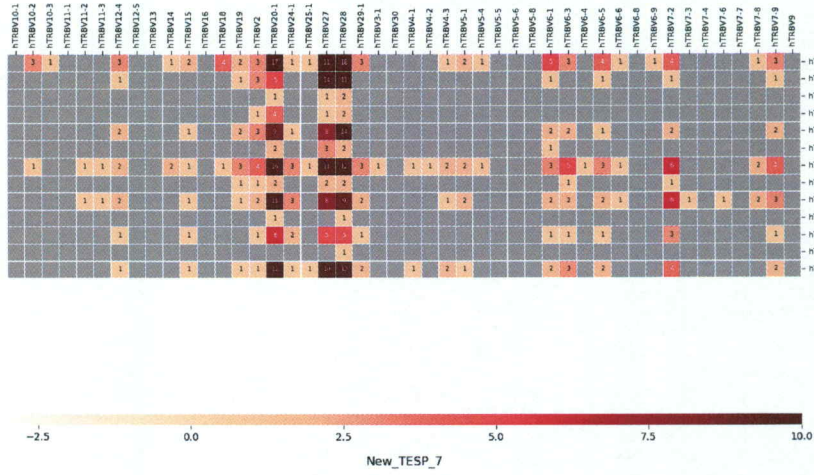


Figure A.46: TESP 8

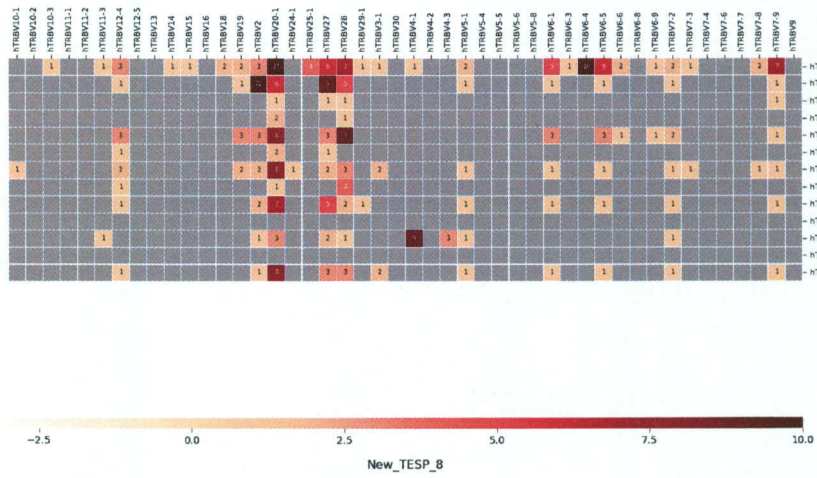


Figure A.47: TESP 9



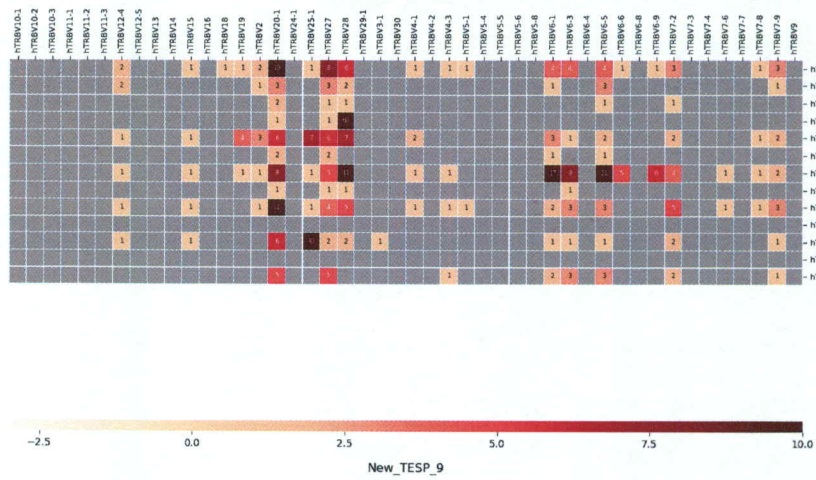


Figure A.48: TESP 10

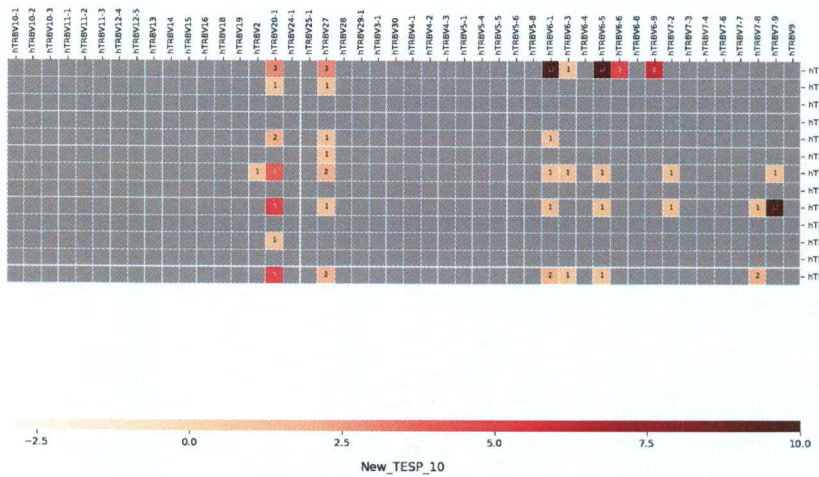


Figure A.49: TESP 11