TÉCNICO LISBOA

Mapping the Spatial Distribution of Neurotransmission Markers using Post-Mortem Cohorts

Miguel da Mata Moreira

Thesis to obtain the Master of Science Degree in

Computer Science and Engineering

Supervisors: Prof. Rui Miguel Carrasqueiro Henriques
Prof. Guilherme Henrique Caçador Ramos

Examination Committee

Chairperson: Prof. Miguel Ângelo Marques de Matos
Supervisor: Prof. Rui Miguel Carrasqueiro Henriques
Member of the Committee: Prof. Rafael Sousa Costa

November 2024

Declaration
I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

Gostaria de começar por agradecer aos meus dois fantásticos orientadores, Rui Henriques e Guilherme Ramos, pela oportunidade incrível de poder trabalhar com eles, pelo apoio e força que me deram ao longo do último ano. Sem eles e sem a sua constante boa disposição, este trabalho não teria sido possível.

Quero agradecer a todos os meus amigos, por estarem ao meu lado em qualquer circunstância e por me darem a força e a vontade para alcançar a melhor versão de mim.

Por último, guardo o maior agradecimento à minha família, aos meus pais e avós em particular, pelo apoio e amor incondicional que sempre me providenciaram ao longo de toda a minha vida.

A todos, um enorme obrigado por me fazerem feliz.

Abstract

Mapping the spatial distribution of transcriptomic neuromarkers involves identifying differential regulatory activity in specific human brain tissues, highlighting tissue-specific transcriptomic signatures in homeostatic states or misregulated in disease. In this work, we primarily focus on protein-coding genes with important roles in neurotransmission systems, such as receptor-coding genes of the dopaminergic system. In this context, we are interested not only in studying how gene expression varies across tissues and individuals but also in capturing covariate effects, including those driven by demographic factors and varying agonal states.

To accomplish this, we introduce a statistical framework to assess pairwise and multiwise differential expression by comparing brain-to-brain tissues, brain-to-peripheral tissues, and brain tissues under different conditions. Furthermore, we also propose an approach for pattern discovery in gene expression data using biclustering and triclustering algorithms for two-way and three-way data, respectively. Both approaches are used to retrieve tissue-specific and cross-tissue putative regulatory modules (patterns) with guarantees of statistical significance.

As major outcomes, this thesis provides a tool for a parameterizable discovery of transcriptomic markers associated with specific neurotransmission systems and visual maps to guide the exploration of both individual and modular markers. The acquired results, validated against available knowledge using well-established biological networks and knowledge bases, confirm the relevance of the proposed tool. Particular attention is drawn to novel putative associations for the acquisition of actionable knowledge.

Keywords

Neurotransmission markers, Brain transcriptomics, Gene expression data analysis, Pattern discovery

Resumo

Mapear a distribuição espacial de marcadores de transcrição envolve identificar representações da atividade reguladora diferencial em tecidos cerebrais humanos específicos, destacando assinaturas transcriptómicas específicas dos tecidos em estados homeostáticos ou mal reguladas na doença. Neste trabalho, concentramo-nos principalmente em genes que codificam proteínas com papéis importantes na neurotransmissão. Neste contexto, estamos interessados não só em estudar a forma como a expressão genética varia entre tecidos e indivíduos, mas também em estudar diversos tecidos do cérebro na presença de cofatores de interesse, incluindo fatores demográficos e estados agonais variáveis.

Neste contexto, propomos uma metodologia para sumarizar a expressão diferencial entre tecidos do cérebro, tecidos do cérebro contra tecidos periféricos, e tecido cerebral em diferentes condições. Além disso, propomos também uma abordagem para a descoberta de padrões em dados de expressão genética utilizando algoritmos de biclustering e triclustering. Ambos os algoritmos são fundamentais para identificar módulos reguladores putativos (padrões) específicos dos tecidos.

Assim, esta tese disponibiliza uma ferramenta para a identificação automática de marcadores transcriptómicos associados a sistemas de neurotransmissão específicos e mapas visuais para orientar a exploração de marcadores individuais e modulares. Os resultados obtidos, validados utilizando redes biológicas e bases de conhecimento bem estabelecidas, confirmam a relevância da ferramenta proposta.

Palavras Chave

Marcadores de neurotransmissão, Transcriptómica cerebral, Análise de dados de expressão genética, Descoberta de padrões

Contents

1 Introduction **1**

1.1 Context and Motivation . . . 2

1.2 Problem Description . . . 3

1.3 Contributions . . . 4

1.4 Document Structure . . . 4

2 Background **5**

2.1 Computational Background . . . 6

2.1.1 Statistical Tests . . . 6

2.1.2 Information Retrieval . . . 8

2.1.3 Biclustering . . . 9

2.1.4 Triclustering . . . 12

2.1.5 Evaluation . . . 13

2.2 Neurobiological Background . . . 14

3 Related Work **19**

3.1 Differential Gene Expression Data Analysis . . . 20

3.1.1 Imaging Transcriptomics of Brain Disorders . . . 20

3.1.2 Best Practices for Imaging Transcriptomics . . . 21

3.1.3 Molecular Maps with Spatial Transcriptomics . . . 22

3.1.4 Human Brain Proteome . . . 23

3.2 Pattern Discovery in Omics Data . . . 23

3.2.1 Biclustering Algorithms . . . 23

3.2.2 Triclustering Algorithms . . . 25

4 Data Profiling **27**

5 Solution **31**

5.1 Data Preprocessing . . . 32

5.1.1 Integrative Cross-Tissue Scores of Differential Expression . . . 32

5.1.2 Identification of Neurotransmission-Related Genes . . . 33

5.1.3 Extension to Co-Factors of Interest . . . 34
5.2 Differential Gene Expression for Multi-Tissue Analysis . . . 35
5.3 Biclustering . . . 38
5.3.1 Problem Mapping . . . 38
5.3.2 Algorithm Selection . . . 39
5.3.3 Hyperparametrization . . . 40
5.4 Triclustering . . . 41
5.4.1 Problem Mapping . . . 41
5.4.2 Algorithm Selection . . . 42
5.4.3 Hyperparametrization . . . 42
6 Visualization Tool **45**
6.1 Gene-Centric Tab . . . 46
6.2 Pattern-Centric Tab . . . 48
7 Results **51**
7.1 Gene-Centric Results . . . 52
7.1.1 General Overview . . . 52
7.1.2 Dopaminergic System . . . 54
7.2 Pattern-Centric Results . . . 62
7.2.1 Biclustering . . . 62
7.2.2 Triclustering . . . 64
8 Conclusion **69**
8.1 Concluding Remarks . . . 70
8.2 Software Availability . . . 70
8.3 Future Work . . . 70
Bibliography **73**
A Additional Code **81**
B Rodents Case Study **83**

List of Figures

2.1 Different types of biclusters [1]: (a) constant overall; (b) constant rows; (c) constant columns; (d) coherent values (addictive model); (e) coherent values (multiplicative model); and (f)-(i) order-preserving in the absence and presence of symmetries. 10
2.2 Different structures of biclusters [1]. 12
2.3 Central dogma of molecular biology [2]. 15
2.4 Brain tissues visually represented [3]. 17

4.1 Number of samples for each tissue. 28
4.2 Sex distribution for all the individuals. 29
4.3 Age distribution for all the individuals. 29
4.4 Agonal state distribution for all the individuals. 29
4.5 Distribution of GABBR2 across all the tissues and samples. 30

5.1 One input tissue (A) against multiple output tissues (B and C). 36
5.2 Multiple input tissues (A and B) against one output tissue (C). 37
5.3 Multiple input tissues (A and B) against multiple output tissues (C and D). 37
5.4 Biclustering problem mapping dimensions. 39
5.5 Triclustering problem mapping dimensions. 42

6.1 Gene-centric tab. 46
6.2 View for one gene selected. 47
6.3 View for more than one gene selected. 47
6.4 Pattern-centric tab for biclustering algorithms from scikit-learn. 48
6.5 Pattern-centric tab for BicPAMS algorithm. 49
6.6 Pattern-centric tab for TriCluster algorithm. 49
6.7 View displaying an illustrative BicPAMS solution. 50

7.1 Hippocampus and hypothalamus versus whole blood and adipose (visceral) tissue. 53

7.2 Distribution of the expression of the gene DBH. 55
7.3 Distribution of the expression of the gene DRD2. 56
7.4 Dopamine-related genes differentially expressed depending on the sex. 59
7.5 Expression values for each gene across the different tissues within the Bicluster1. 63
7.6 Expression values for each gene across the different tissues within the order-preserving bicluster. 64
7.7 Expression values for each gene across the different tissues within the tricluster for the individual GTEX-1GN73. 65
7.8 Expression values for each gene across the different tissues within the tricluster for the individual GTEX-1B8L1. 66

B.1 Mean wild-type (left) and knockout (right) sample expression of GABBR2 gene across tissues. 84

List of Tables

2.1 Principal constituents of brain tissue [4, 5]. 15
2.2 Examples of neurotransmitters [6–8]. 16
2.3 Brain tissues and their functions [9–22]. 16

4.1 Example row of the Genotype-Tissue Expression (GTEx) dataset for the amygdala tissue. 28

5.1 Illustrative gene description from the UniProt knowledge base. 34
5.2 Example row of the dataset created from the metadata. 35

7.1 Number of differentially expressed genes for all $56,200$ genes. 52
7.2 Distribution of the 196 neurologically related genes across various p-value ranges. 53
7.3 Number of differentially expressed genes for the 196 automatically retrieved genes with known brain-specific functions. 54
7.4 Biological functions of the genes from the dopaminergic system (from the Gene Ontology (GO) initiative). 54
7.5 Scores and regulation conclusion for amygdala versus frontal cortex (BA9). 55
7.6 Mean and standard deviation of the gene DBH. 56
7.7 Mean and standard deviation of the gene DRD2. 56
7.8 Enriched terms associated with the up-regulated (activated) dopamine-related genes in the amygdala (DRD2, SLC6A3, SLC6A2). 57
7.9 Enriched terms associated with the up-regulated (activated) dopamine-related genes in the frontal cortex (DBH, DRD5, DRD3, DRD4). 57
7.10 Scores and regulation conclusion for amygdala versus frontal cortex (BA9) on males. . . 58
7.11 Scores and regulation conclusion for amygdala versus frontal cortex (BA9) on females. . 58
7.12 Enriched terms associated with the up-regulated (activated) dopamine-related genes in the amygdala for males (SLC6A3, DRD2). 59
7.13 Enriched terms associated with the up-regulated (activated) dopamine-related genes in the frontal cortex (BA9) for males (DBH, DRD5, DRD3, DRD4, SLC18A2). 59

7.14 Enriched terms associated with the up-regulated (activated) dopamine-related genes in the frontal cortex (BA9) for females (DBH, DRD5). 60
7.15 Scores and regulation conclusion for dopamine-related genes in the frontal cortex (BA9) versus the whole blood. 61
7.16 List of the constant biclusters retrieved from BicPAMS. 62
7.17 Enriched terms associated with the up-regulated (activated) genes from the Bicluster1. . 63
7.18 List of the triclusters retrieved from TriCluster (first parameterization). 65
7.19 List of the triclusters retrieved from TriCluster (second parameterization). 67

B.1 Example row of the Gene Expression Omnibus (GEO) dataset for the first wild-type sample. 84

Acronyms

3DMSR	3D Mean Squared Residue
AHBA	Allen Human Brain Atlas
ALS	Amyotrophic Lateral Sclerosis
AMBA	Allen Mouse Brain Atlas
ANOVA	Analysis of Variance
CGE	Correlated Gene Expression
DNA	Deoxyribonucleic Acid
DTI	Diffusion Tensor Imaging
GCEA	Gene Category Enrichment Analysis
GEO	Gene Expression Omnibus
GO	Gene Ontology
GTEx	Genotype-Tissue Expression
GWASs	Genome-Wide Association Studies
ICA	Independent Component Analysis
IDPs	Imaging-Derived Phenotypes
ISH	In Situ Hybridization
MRI	Magnetic Resonance Imaging
MSR	Mean Squared Residue
PCA	Principal Component Analysis
PCM1	Pericentriolar Material 1
RNA	Ribonucleic Acid
SVD	Singular Value Decomposition
TPM	Transcripts Per Million

1

Introduction

Contents

1.1 Context and Motivation 2
1.2 Problem Description 3
1.3 Contributions 4
1.4 Document Structure 4

Accessing the regulatory activity of the human brain is crucial for neuropsychiatric studies, nevertheless, it remains limited. At this light, this work explores the value contained in large-scale post-mortem cohort initiatives to automatically identify regulatory markers associated with the functioning of neurotransmission systems. Section 1.1 provides the context and motivation for this thesis. Section 1.2 describes the problem at hand. Section 1.3 outlines the main contributions of this work. Finally, Section 1.4 details the organization of the document.

1.1 Context and Motivation

In the context of this work, mapping the spatial distribution of neurotransmission markers involves identifying actionable representations of differential regulatory activity in specific brain tissues of the human brain. This mapping supports the analysis of tissue-specific transcriptional activity, taking into account the individual psycho-demographic traits and agonal state at death.

Despite the relevance of neurotranscriptomic maps for knowledge acquisition, several factors challenge their learning: (i) the variability of expression; (ii) the need for a highly usable visualization and navigation of information; (iii) the need to account for multiwise gene interactions; (iv) the lack of *in vivo* data in humans due to the difficulty of accessing the brain; and (v) the batch effects and variability associated with co-factors, such as agonal state and demographic profile [23].

It is noteworthy that the need to model how critical molecular fingerprints vary across the different regions of the brain for the study of neurotransmission systems, providing a visual representation of the tissue-specific regulation, is not recent. In particular, we can observe a notable development and progress in this field due to advances in technology and automation of laboratory procedures [24].

Amongst the incursion of studies on this field, a significant landmark was the creation of the Allen Brain Atlas project[1] [25] by the Allen Institute for Brain Science[2]. In 2004, the Allen Mouse Brain Atlas (AMBA)[3] [25] was released with the objective of providing an encompassing map detailing gene expression throughout the entirety of the mouse brain. This venture represented a notable mark in comprehending the molecular organization of the brain. More recently, in 2010, the Allen Human Brain Atlas (AHBA)[4] [26] was created for humans. This atlas combines anatomical and genome information, utilizing various data modalities such as Magnetic Resonance Imaging (MRI), Diffusion Tensor Imaging (DTI), histology, and gene expression data obtained through both microarray and In Situ Hybridization (ISH) methods [27].

Despite the incursion of studies, none of them consider the inclusion of contrast tissues, e.g., peripheral tissues, to isolate specific regulatory aspects of the human brain. This approach enables a deeper

[1] Allen Brain Atlas website.
[2] Allen Institute for Brain Science website.
[3] Allen Mouse Brain Atlas website.
[4] Allen Human Brain Atlas website.

understanding of the multiwise differences in spatial distribution of expression between different groups of tissues. Such comparisons will contribute to the development of a brain model that could relate distinct gene expressions to acquire information about phenotypes and psycho-demographic conditions, accessing relevant regulatory patterns for the acquisition of novel knowledge in the data.

1.2 Problem Description

The problem we aim to address is the discovery modeling of transcriptomic markers that regulate neurotransmission systems, particularly, focusing on the regulation of neurotransmitter receptors, including GABA and dopamine receptors, based on epitranscriptomic screening. Our goal is to develop interactive maps that may provide organized views of neurotransmission markers aided by known functions present in pathway repositories and ontologies, as well as putative functions derived from functional enrichment analysis of co-expression patterns.

To establish these neurotranscriptomic maps, it is necessary to explore human brain tissues to assess the spatial distribution of neurotransmission markers. These maps will enable direct comparisons of gene-centric and multiwise transcriptional activity at different brain tissues, between brain tissues and peripheral tissues, and with the presence of various co-factors of interest, such as psycho-demographic profiles, i.e., sex and age, and agonal states, i.e., circumstances of death.

To this purpose, we are primarily interested in modeling neurotransmission markers, i.e., sets of genes that are differentially expressed in nervous tissues to support the regulation of neurotransmission systems. Identifying the existence or absence of specific gene transcripts can act as a marker in describing a particular medical condition. To this end, neurotransmission markers can offer a better understanding of the chemical framework of signal processing and transmission, aiding the monitoring of developmental or pathological advancements, and uncovering crucial details about possible targets for medical interventions [28].

To these ends, this work uses data derived from post-mortem human cohorts associated with the Genotype-Tissue Expression (GTEx)[5] initiative [29]. This initiative provides a public resource of tissue-specific gene expression for humans that involves samples collected from 54 non-diseased tissue sites from almost 1000 individuals.

Ultimately, studying these neurotransmission systems is paramount to understand the functioning of the brain, neurological and psychiatric diseases, and to supporting advances in psychopharmacology with the introduction of new and improved therapies, including psychotropic drugs. The use of neurotransmission markers reveals the brain patterning at different tissues, thereby potentiating the standardization of the diagnosis of clinical conditions and diseases like schizophrenia [30].

[5] Genotype-Tissue Expression website.

Additionally, the interactive maps illustrating gene expression across different brain tissues will facilitate the identification of specific differences related to psycho-demographic profiles, sex, age, and agonal states. At this light, cross-relating gene expression with different phenotypes is crucial in the context of this work.

1.3 Contributions

The main contributions of this thesis are fourfold:

- We develop a methodology for gene-centric multi-tissue differential expression analysis, ensuring sensitivity to individual psycho-demographic profiles, satisfying guarantees of usability, statistical significance, and completeness to medical annotations;
- We discover putative regulatory modules (patterns) using biclustering algorithms (for two-way data) and triclustering algorithms (for three-way data), ensuring statistical significance and discriminative power;
- We create a novel web-based application that enables the exploration and analysis of interactive neurotranscriptomic maps for gene-centric multi-tissue differential expression analysis and pattern-centric putative regulatory modules (patterns);
- We evaluate the acquired maps against available knowledge using, e.g., biological networks, knowledge bases like Gene Ontology (GO)[6], and Enrichr[7], to assess their potential for actionable insights.

1.4 Document Structure

Chapter 2 presents the background with the context and essential concepts required to understand the proposed work. Chapter 3 surveys contributions to gene-centric modelling, i.e., gene expression data analysis, and pattern-centric modelling, i.e., pattern discovery. Chapter 4 provides an exploratory analysis of the available data. Chapter 5 establishes the solution for this work. Chapter 6 details the visualization tool developed. Chapter 7 presents and discusses the acquired results. Finally, concluding remarks and future work are synthesized in Chapter 8.

[6] Gene Ontology website.
[7] Enrichr website.

2

Background

Contents

2.1 Computational Background 6
2.2 Neurobiological Background 14

Essential background is presented in two major sections. Section 2.1 introduces technical concepts for this work, more specifically, statistical testing, information retrieval, biclustering, triclustering, and evaluation. Section 2.2 introduces biomedical concepts, including differential gene expression and neurotransmission systems.

2.1 Computational Background

This section presents nuclear concepts for gene-centric and pattern-centric modeling of transcriptomic data. These concepts will enable a comprehensive mapping of the spatial distribution of gene expression across brain tissues and the detection of putative regulatory modules (patterns) of neurotransmission in different tissues.

2.1.1 Statistical Tests

In this work, we will employ statistical tests to analyse data samples from various genes and tissues of interest. Specifically, we will consider four tests: (i) the Shapiro-Wilk normality test; (ii) the Student's t-test; (iii) the Wilcoxon signed-rank test; and (iv) the Analysis of Variance (ANOVA).

A Shapiro-Wilk normality test is a statistical hypothesis test to determine how well the data sample fits into a normal distribution [31]. We have as the null hypothesis that the data sample follows a normal distribution,

$$H_0 : X \sim N(\mu, \sigma^2), \tag{2.1}$$

where X represents the random variable of interest with the distribution described by a sample of observations, and N represents the normal distribution with mean μ and variance σ^2. With the calculation of the p-value and subsequently comparison with α, for a given confidence level c, $\alpha = 1 - c$, and $\alpha \in [0, 1]$, we reject the null hypothesis in favor of the alternative one,

$$H_1 : X \not\sim N(\mu, \sigma^2), \tag{2.2}$$

whenever the p-value is lower than the subtraction, stating that the data sample does not follow a normal distribution. We can calculate the t-value in the following way,

$$t = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} \left(x_i - \mu\right)^2}, \tag{2.3}$$

where $x_{(i)}$ represents the ordered sample values. Besides that, a_i are constants, and m is a vector with the expected values of the ordered statistics that are independent and identically distributed random

variables that follow the standard normal distribution, respectively defined by,

$$(a_1, a_2, \ldots, a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} m)^{1/2}}, \text{ and} \tag{2.4}$$

$$m = (m_1, m_2, \ldots, m_n)^\intercal, \tag{2.5}$$

where V is the covariance matrix of the order statistics [32].

If we want to compare two data samples and both follow a normal distribution, as determined by the Shapiro-Wilk normality test, we can use the Student's t-test. A Student's t-test is a statistical test used to examine if sample values from two populations have the same mean. Usually, this hypothesis testing considers a null hypothesis that the samples have the same mean,

$$H_0 : \mu_1 = \mu_2, \tag{2.6}$$

where μ_1 is the mean for the first sample, and μ_2 is the mean for the second sample. If the p-value is lower than α, for a given confidence level c, we reject the null hypothesis in favor of the alternative one,

$$H_1 : \mu_1 \neq \mu_2, \tag{2.7}$$

stating that the difference between the means of the two samples is statistically significant. If we assume equal variances in the two samples (paired samples), we can further calculate the t-value,

$$t = \frac{\mu_1 - \mu_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{2.8}$$

where μ_1 and μ_2 are the means of the first and second sample, respectively, s is the pooled standard error of the two groups, and n_1 and n_2 are the sizes of the first and second samples, respectively. In this sense, larger values of t indicate a larger difference between samples. With the t-value, we can calculate the p-value, only depending on the aspect of the region of rejection of H_0, thus representing the higher level of confidence that leads to the non-rejection of H_0 [33].

Conversely, if the data samples do not follow a normal distribution, we will consider using the Wilcoxon signed-rank test, which does not require the assumption of normality. A Wilcoxon signed-rank test is the non-parametric equivalent of the paired Student's t-test [34]. The test assesses whether the mean values between paired samples is different from zero.

Moreover, if we have more than two samples, we can use the ANOVA statistical test. An ANOVA is a statistical test that allows one to inspect if two or more samples have identical means. We use ANOVA

in hypothesis testing, where the null hypothesis states that all the samples have the same mean,

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k, \tag{2.9}$$

where k is the number of samples. At the time of the p-value computation, if the score is lower than α, for a given confidence level, we reject the null hypothesis in favor of the alternative one,

$$H_1 : \exists_{m,n \in \{1,...,k\}}, \mu_m \neq \mu_n, \tag{2.10}$$

where μ_m and μ_n are two specific means that are different in favor of the alternative hypothesis. A significant p-value in this test indicates that at least for a pair (m, n), the mean difference is statistically significant.

2.1.2 Information Retrieval

In the context of this work, information retrieval techniques will be crucial for identifying genes whose descriptions are associated with the regulation of neurotransmission systems. Techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) will assist in retrieving textual information about the function of each gene. TF-IDF is a text vectorization scoring scheme that measures the importance of a query of terms to a document in a collection of documents. The importance of a term is higher when it appears more times in a given document and when it appears less times in the remaining documents [35]. We can define a TF-IDF statistic as follows,

$$TF\text{-}IDF = TF \times IDF. \tag{2.11}$$

The element TF represents the term frequency of a given term in a specific document, and it is represented as follows,

$$TF = \frac{n}{t}, \tag{2.12}$$

where n represents the number of times a term appears in the document, and t is the total number of different terms. On a different note, the element IDF represents the inverse document frequency and defines the frequency of a given term in the whole collection. Its expression is given by the following,

$$IDF = \log\left(\frac{N}{DF}\right), \tag{2.13}$$

where N represents the number of documents in the collection, and DF the number of documents in the corpus containing the term [36].

2.1.3 Biclustering

In this work, we will utilize clustering approaches to analyze gene expression data [1]. Clustering is a task that groups observations based on their similarity [37]. It uses unsupervised machine-learning techniques to group points without relying on their true labels. The grouping can be either distance-based or frequency-based, depending on the measure used, distance or probability.

Biclustering is a subspace clustering approach that aims to find sub-matrices, subsets of rows and columns, where the rows exhibit a strong correlation for every column. Assuming genes correspond to rows, conditions to columns, and each element typically represents the logarithmic expression level of a specific gene under a given condition, the ultimate goal is to find sub-matrices, that is, subgroups of genes and conditions that are highly correlated [1]. The difference between clustering and biclustering remains in the number of dimensions considered. We can apply clustering to either rows or columns separately, while biclustering addresses rows and columns simultaneously. To this end, we are concerned with studying the type of bicluster, the structure produced by it, and the algorithms used for each one of the approaches [1].

Consider the matrix $A \in \mathbb{R}^{n \times m}$, with n rows and m columns. We define A by its set of rows $X = \{x_1, ..., x_n\}$, and its set of columns $Y = \{y_1, ..., y_m\}$. We can denote the matrix A as $A = (X, Y)$. If we take I as a subset of X, $I \subseteq X$, and J as a subset of Y, $J \subseteq Y$, the resulting sub-matrix is, $A_{IJ} = (I, J)$, which corresponds to a bicluster of dimensions $k \times s$. We can also get a cluster of rows, $A_{IY} = (I, Y)$, or a cluster of columns, $A_{XJ} = (X, J)$. Given a matrix A, biclustering aims at identifying the set of biclusters $\{B_k\}_k$ such that each B_k satisfies a specific criteria of homogeneity and statistical significance [1]. Moreover, we define B_k as follows,

$$B_k = (I_k, J_k). \tag{2.14}$$

We can visualize a data matrix as a weighted bipartite graph. We define a graph G as a pair, $G = (V, E)$, for V nodes and $E \subset V \times V$ edges. If we can divide the set of nodes V into two sets, the graph is bipartite, those sets being L and R, where $L \cap R = \emptyset$ and $L \cup R = V$, and every edge of E has one end in L and another end in R, i.e., $E \subset L \times R$. Making the visual association with the data matrix, the set L corresponds to the rows, $n_i \in L$, the set R corresponds to the columns, $n_j \in R$, and the edges have the values a_{ij}.

Finding a maximum size bicluster is an NP-complete problem. The term “NP” stands for the class of decision problems that can be solved with nondeterministic polynomial time, which is a class of decision problems that can be efficiently solved by a nondeterministic Turing machine [38]. An NP-complete problem is a problem that is both NP-hard and NP. An NP-hard problem is a problem that is at least as hard as any NP problem [39].

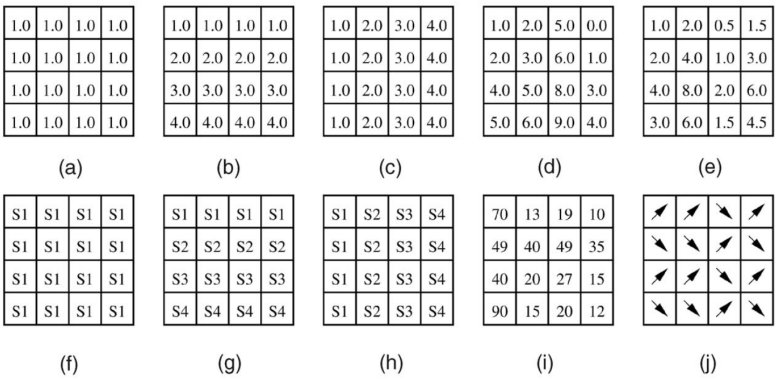


Figure 2.1: Different types of biclusters [1]: (a) constant overall; (b) constant rows; (c) constant columns; (d) coherent values (addictive model); (e) coherent values (multiplicative model); and (f)-(i) order-preserving in the absence and presence of symmetries.

There are four major classes of biclusters types, which are illustrated in Figure 2.1. Firstly, biclusters with constant values, Figure 2.1(a), are relevant for this thesis as they can reveal sets of genes with similar expression within a subset of conditions. For that being, it is only natural to consider reordering the data matrix to (better) understand and observe those similar rows and columns that form a bicluster [1]. A perfect constant bicluster is a sub-matrix (I, J) where all the values are equal, for all $x_i \in I$ and $y_j \in J$, and defined as follows,

$$a_{ij} = \mu, \tag{2.15}$$

and where μ represents the real value of a_{ij}. In real-valued data contexts, a noise term is necessary to accommodate deviations from pattern expectations, given by,

$$a_{ij} = \mu + \eta_{ij}. \tag{2.16}$$

Hartigan [40] introduced an algorithm named Block Clustering that splits the original data into a set of biclusters and evaluates the quality of each one with the variance formula. The assumption is that there are K biclusters in the data matrix, $(I, J)_k$. According to this approach, a perfect bicluster has zero variance. The term a_{IJ}, essential to compute the variance, represents the mean of all the elements in a bicluster. To this end, we define the mean and variance with the following expressions, respectively,

$$a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij}, \tag{2.17}$$

$$VAR(I, J)_k = \sum_{k=1}^{K} \sum_{i \in I, j \in J} \left(a_{ij} - a_{IJ}\right)^2. \tag{2.18}$$

Secondly, biclusters with constant values on rows or columns, Figure 2.1(b) and Figure 2.1(c), disclose a substantial amount of information since a bicluster with constant values in the rows identifies a subset of genes that have the same expression across a subset of conditions, or the other way around, expressing a subset of conditions in the same way through a set of genes [1]. A perfect bicluster with constant rows is a sub-matrix (I, J) where we obtain all the values using an additive model, or a multiplicative model. In this case, μ represents the entry value, and α_i represents the adjustment for row x_i. We can define those adjustments as follows,

$$a_{ij} = \mu + \alpha_i, \tag{2.19}$$

$$a_{ij} = \mu \times \alpha_i. \tag{2.20}$$

Similarly, a perfect bicluster with constant columns is a sub-matrix (I, J) where we can obtain all the values using an additive model, or a multiplicative model. This time, β_j functions as an adjustment for column y_j. We can define those adjustments as follows,

$$a_{ij} = \mu + \beta_j, \tag{2.21}$$

$$a_{ij} = \mu \times \beta_j. \tag{2.22}$$

A simplistic approach to identify a bicluster of this type is to normalize the row and column values [1]. Researchers have contemplated other ways to achieve this effect. Getz et al. [41] performed a similar approach of normalization with the addition of a complexity pre-step. Califano et al. [42] aimed at finding δ-valid ks-patterns, i.e., the difference between the maximum and minimum value of each row, and compare that result with a given δ.

Thirdly, one can model biclusters with coherent values, as depicted in Figure 2.1(d) and Figure 2.1(e), using either an additive model or a multiplicative model, depending on the biclustering strategies assumed. With that idea, we can define a perfect bicluster with coherent values (I, J), with two adjustments, the row adjustment, and the column adjustment. We define those adjustments as follows,

$$a_{ij} = \mu + \alpha_i + \beta_j, \tag{2.23}$$

$$a_{ij} = \mu \times \alpha_i \times \beta_j. \tag{2.24}$$

Lastly, biclusters with coherent evolutions, Figures 2.1(f)-(j), show order-preserving permutations evolutions across the rows and/or columns. Their research concentrates on the order of the columns in the bicluster rather than the homogeneity of the data matrix values [1].

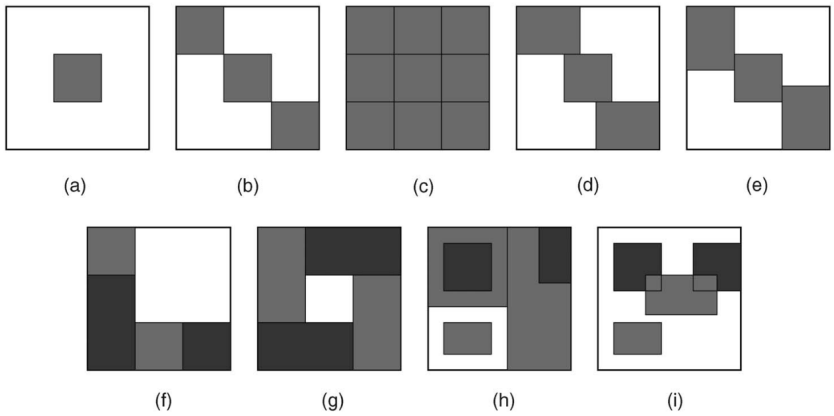


Figure 2.2: Different structures of biclusters [1].

Regarding bicluster structure, in Figure 2.2(a), we can observe a single bicluster in the matrix. An ideal reordering of the matrix would produce visually perceptible biclusters like the ones represented in Figure 2.2(b), with some number K of rectangular biclusters in the diagonal of the matrix. The next natural step is to assume that each set of rows and columns can belong to more than one bicluster at the same time, therefore assuming a checkerboard structure, like the one in Figure 2.2(c), where we can notoriously identify K non-overlapping and non-exclusive biclusters. There are other approaches, namely, exclusive rows biclusters, where each row can only belong to one bicluster while columns can belong to a number of them, proposed by Sheng et al. [43] and Tang et al. [44] and portrayed in Figure 2.2(d). The opposite, exclusive columns biclusters in Figure 2.2(e) can also happen.

Up to this point, we only observed exhaustive representations, i.e., every row and column belongs to at least one bicluster, but in non-exhaustive representations that is not a mandatory condition. By observing Figure 2.2(f) and Figure 2.2(g), we can understand that concept. However, in real data, it is more likely to have rows and columns that do not belong to any bicluster, and biclusters that overlap others spatially. We can note that specific situation in Figure 2.2(h) and Figure 2.2(i).

2.1.4 Triclustering

When a third dimension needs to be considered in our data, biclustering techniques become inadequate. In this context, triclustering algorithms provide a method for identifying genes with similar expression patterns across a range of contexts within specific conditions. Triclustering is a subspace clustering technique specialized in finding patterns in three-way data, i.e., data with three dimensions, first proposed by Zhao and Zaki [45].

Having a three-way dataset D, let $X = \{x_1, ..., x_n\}$ be the set of objects, $Y = \{y_1, ..., x_m\}$ be the set of features, and $Z = \{z_1, ..., z_p\}$ be the set of contexts, we can define a tricluster $\mathscr{T}$ with the following,

$$\mathscr{T} = (I, J, K), \tag{2.25}$$

where $I \subseteq X$, $J \subseteq Y$, and now $K \subseteq Z$. Each element a_{ijk} corresponds to the element of row i, column j and context k.

We can visualize a tricluster as a set of S biclusters, i.e., all concordant biclusters along a certain dimension [46]. In this case, each B_k represents a bicluster, and we can define it with the following,

$$\mathscr{T} = (B_1, B_2, ..., B_S). \tag{2.26}$$

A tricluster should satisfy specific homogeneity and statistical significance criteria. We can access homogeneity using merit functions to evaluate the quality of the triclustering solution. In terms of statistical significance, a tricluster must be statistically significant, i.e., their probability to occur should be small [46].

2.1.5 Evaluation

We are interested in evaluating the relevance of neurotransmission markers that capture gene expression in specific human brain tissues. This analysis provides valuable insights into the regulatory processes unique to each tissue.

In bivariate stances, statistical tests are used to evaluate gene expression in specific human brain tissues. However, in multivariate stances, i.e., putative regulatory modules (patterns), we must consider four paramount aspects: (i) accuracy against reference solutions; (ii) homogeneity; (iii) statistical significance; and (iv) discriminative power.

Firstly, accuracy metrics are typically similarity measures, and we can emulate them when we know *a priori* the true solution. In biclustering some measures comprehend Relative Nonintersecting Area (RNAI) [47], Subspace Clustering Error [48] and Jaccard-based Scores (MS) [49]. In triclustering, we can extend the Jaccard-based Scores for three dimensions. Additionally, the 3D Revised Match Score (RMS3) [46] is a measure that considers the non-matched volume between triclusters, offering a penalization for that case, being an accurate measure for this end. Assuming that $\mathcal{B} = \{B_1, ..., B_S\}$ represents the learned triclusters, and $\mathcal{H} = \{H_1, ..., H_R\}$ denotes the true triclusters, we can define the RMS3 measure as follows,

$$\mathrm{RMS3}(\mathcal{B}, \mathcal{H}) = \frac{1}{|\mathcal{B}|} \sum_{B_1 \in \mathcal{B}, \mathrm{maxarg}\{\mathrm{Jac}(B_1, B_2) | B_2 \in \mathcal{H}\}} \sqrt[3]{\frac{|I_1 \cap I_2|}{|I_1 \cup I_2|} \frac{|J_1 \cap J_2|}{|J_1 \cup J_2|} \frac{|K_1 \cap K_2|}{|K_1 \cup K_2|}}. \tag{2.27}$$

Secondly, we can access homogeneity using merit functions to evaluate the quality of the clustering solution. A merit function is a function that, based on the values of its elements, evaluates how good a

tricluster is, i.e., variance in triclustering. Each merit function is partial to certain homogeneity criteria and can offer unique perspectives. Hence, the combination of merit functions can be beneficial for fair evaluations of the solution [46]. In biclustering, we can assess clustering quality using the Mean Squared Residue (MSR) score [50] as follows,

$$\mathrm{MSR}_{(I,J)} = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} \eta_{ij}^2, \quad \eta_{ij} = a_{ij} - a_{iJ} - a_{Ij} + a_{IJ}, \tag{2.28}$$

where a_{ij} is the element of row i and column j, a_{iJ} the mean value of row i, a_{Ij} the mean value of column j, and a_{IJ} the mean value in the sub-matrix (I, J). Alternatively, in triclustering, we can calculate the 3D Mean Squared Residue (3DMSR) score [46] as follows,

$$3\mathrm{DMSR}_{(I,J,K)} = \frac{1}{|I||J||K|} \sum_{i \in I} \sum_{j \in J} \sum_{k \in K} \eta_{ijk}, \quad \eta_{ijk} = a_{ijk} - (a_{iJK} + a_{IjK} + a_{IJk} - 2a_{IJK}). \tag{2.29}$$

Thirdly, statistical tests are another approach to consider for evaluating the statistical significance of biclustering solutions, limiting the occurrence of false-positive and false-negative cases [51].

Lastly, we can also employ discriminative power metrics to evaluate the quality of our solution. The discriminative power of a given subspace cluster (pattern) defines its probability to only occur on a subset of classes, $C \subset \mathcal{C}$, where $\mathcal{C} = \{c_1, \ldots, c_i\}$ represents the set of classes (conditions of interest) [52]. For a given association rule R, $\varphi_B \rightarrow c$, we can calculate the confidence and lift measures, respectively, as follows,

$$confidence(R) = \frac{P(\varphi_B, c)}{P(c)}, \text{ and} \tag{2.30}$$

$$lift(R) = \frac{P(\varphi_B, c)}{P(\varphi_B)P(c)}, \tag{2.31}$$

where φ_B represents the discriminative pattern of expression, and c represents the condition of interest, e.g., clinical or demographical condition.

Apart from the statistical significance and the discriminative power, we can also evaluate the functional enrichment of the found patterns associated with neurotransmission systems and medical conditions.

2.2 Neurobiological Background

Neurobiology is a dedicated branch of biology mainly focused on studying the nervous system and how it affects human behaviour. Three principal constituents compose the brain tissue, presented with the respective function in Table 2.1.

Table 2.1: Principal constituents of brain tissue [4, 5].

Component	Function
Neurons	Process and transmission of signals between neurons
Glial cells	Provide protection and stability for neurons
Extracellular matrix	Offers structural support to neurons and glial cells

Deoxyribonucleic Acid (DNA) is a molecule that contains genetic information and that exists mostly in the nucleus of the cells. A DNA molecule comprises two lengthy polynucleotide chains made up of four nucleotide subunits. These chains, named DNA strands, maintain their connection through hydrogen bonds (H-bonds) between the nucleotide base portions [53].

Ribonucleic Acid (RNA) is a molecule responsible for the assembly of proteins. To form the RNA, cells copy the DNA to messenger RNA (mRNA) through a process named transcription [53], shown in Figure 2.3. The mature mRNA is formed with the removal of the introns of the precursor mRNA with splicing, only leaving the exons since they are the only ones coding proteins [54].

Proteins are large molecules formed by the combination of amino acids. To create a protein, mRNA leaves the nucleus for the cytoplasm to the ribosome, where it guides protein synthesis that involves transfer RNA (tRNA) in a process known as translation, shown in Figure 2.3. The ribosome divides mRNA into three base segments called codons, with each codon representing a specific amino acid. As the mRNA moves through the ribosome, codons match with anti-codons on tRNA through base pairing, facilitating the incorporation of amino acids into the growing protein chain [54].

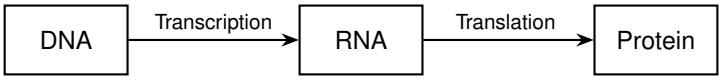


Figure 2.3: Central dogma of molecular biology [2].

The proteome is the set of proteins expressed by a specific organism at a specific time under certain conditions. The proteome is dynamic and can change with the developmental stage of the organism, and both internal and external conditions [55].

Neurotransmitters are molecules that amplify, transmit, and convert signals in cells to transport information throughout the nervous system [56]. Excitatory neurotransmitters lead to the initiation of an action potential in the target neuron, while inhibitory neurotransmitters avoid the generation of an action potential in the target neuron. Moreover, we can also divide neurotransmitter receptors into two principal groups: (i) ligand-gated ion channels, these receptors are transmembrane protein ion channels that directly open upon ligand binding [57]; and (ii) metabotropic receptors, neurotransmitter binding initiates a signaling pathway that can indirectly influence the opening or closing of channels [58]. In this context, some neurotransmitters are explicitly described in Table 2.2.

Table 2.2: Examples of neurotransmitters [6–8].

Neurotransmitter	Effect	Type of receptor	Function
Dopamine	Inhibitory/Excitatory	Metabotropic	Pleasure
GABA	Inhibitory	Ligand-gated	Calming
Adrenaline	Excitatory	Metabotropic	Fight or flight
Noradrenaline	Inhibitory/Excitatory	Metabotropic	Concentration
Serotonin	Inhibitory/Excitatory	Metabotropic	Mood
Glutamate	Excitatory	Ligand-gated	Memory
Acetylcholine	Inhibitory/Excitatory	Ligand-gated/Metabotropic	Learning
Endorphins	Inhibitory	Metabotropic	Euphoria
Oxytocin	Excitatory	Ligand-Gated	Reproduction
Melatonin	Inhibitory	Metabotropic	Sleep-wake regulation

The brain is the most complex organ in the human body [59]. The natural definition of brain tissue or region maintains an opaque definition, meaning that the differences in brain structure between individuals complicate the establishment of strict anatomical definitions [23]. Nevertheless, and in view of this work, we can divide the brain into a set of natural tissues to analyse in the future, presented in Table 2.3 and in Figure 2.4.

Table 2.3: Brain tissues and their functions [9–22].

Tissue	Function
Amygdala	Emotional processing and memory formation
Anterior Cingulate Cortex (BA24)	Cognitive functions, emotional regulation, and decision-making
Caudate (basal ganglia)	Motor control and learning
Cerebellar Hemisphere	Coordination of voluntary movements
Cerebellum	Motor control and balance
Cortex	Higher cognitive functions, including thinking and perception
Frontal Cortex (BA9)	Executive functions, reasoning, and problem-solving
Hippocampus	Formation of new memories and spatial navigation
Hypothalamus	Regulation of basic survival functions, such as hunger and thirst
Nucleus accumbens (basal ganglia)	Reward and pleasure center, involved in motivation
Putamen (basal ganglia)	Motor control and learning
Spinal cord (cervical c-1)	Transmitting signals between the brain and the body
Substantia nigra	Motor control and coordination
Pituitary	Growth and reproductive development and function

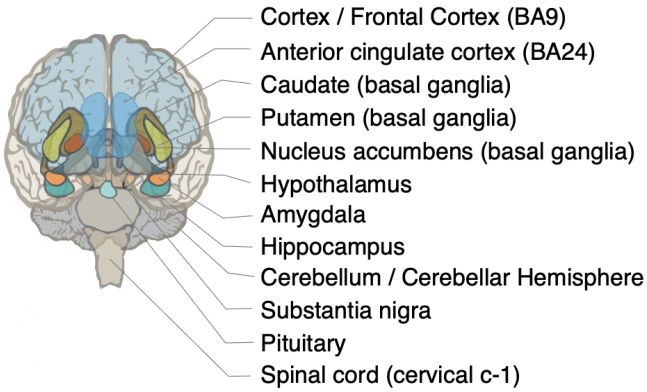


Figure 2.4: Brain tissues visually represented [3].

A synapse is an inter-cellular junction specialized for the communication between a pre-synaptic and a post-synaptic neuron [60]. It starts when an electrical pulse arrives in the pre-synaptic neuron, making the synaptic vesicles fuse with the membrane. That fusion leads to the release of neurotransmitters to the synaptic gap where the receptors of the post-synaptic neuron receive them, leading to a neurobiological response [61].

3

Related Work

Contents

3.1 Differential Gene Expression Data Analysis 20
3.2 Pattern Discovery in Omics Data 23

The existing literature is organized into two major parts. First, in Section 3.1, we overview the works related to differential gene expression analysis in the brain. Then, in Section 3.2, we present a collection of insights on the advances of pattern discovery in omics data.

3.1 Differential Gene Expression Data Analysis

To address our first objective, we explored several works within four topics related to the analysis of gene expression data: (i) imaging transcriptomics of brain disorders, which is pertinent to understanding neural changes associated with numerous brain disorders using neuroimaging data; (ii) best practices for imaging transcriptomics of the human brain, to optimally integrate neuroimaging data with transcriptomic measures; (iii) molecular maps of the mammalian brain with spatial transcriptomics, which are relevant due to the fine-grained spatial resolution of molecular signatures; and (iv) human brain proteome description to visit complementary regulatory stances in the brain and their association with distinct cell types and functions.

3.1.1 Imaging Transcriptomics of Brain Disorders

Imaging transcriptomics involves using imaging techniques to visualize and quantify gene expression within tissue samples. The ultimate goal of imaging transcriptomics is to understand if there is a pattern that connects the changes of gene expression in multiple tissues with the changes of function and structure in the tissues themselves.

Arnatkeviciute et al. [62] explored the value of imaging transcriptomics in understanding psychiatric disorders by studying gene expression data at subcellular resolution and cross-validating it with neuroimaging information. To this end, the authors explored how imaging transcriptomics played a role in understanding the molecular basis of Imaging-Derived Phenotypes (IDPs) [62]. Typically, this type of study uses atlases that refer to the gene expression of each gene in different locations/tissues of the brain, cross-validating those values with the contents of IDPs. Arnatkeviciute et al. [62] focused on human brain data, deriving it from the AHBA initiative. The data was collected from 6 adult brain donors whose ages ranged from 24 to 57 years old [62].

Three major approaches have been pursued to perform imaging transcriptomics: (i) regional gene expression, which is a spatial correlation study across brain tissues for a selected sample of genes, i.e., fixing the tissues; (ii) Correlated Gene Expression (CGE), which is the comparison between selected pairs of tissues across the set of genes, i.e., fixing the genes; and (iii) gene co-expression, which involves analyzing spatial correlations between the expressions presented by pairs of genes across different brain regions [26].

To analyse the connection between the different expression levels and neuroimaging data, we can take two different approaches. The first approach centers on hypothesis-driven analysis, where we start with a set of genes proposed as a hypothesis to take into consideration. The second approach involves data-driven analysis, aiming to identify patterns across thousands of genes, without any gene used as a hypothesis.

We can point out relevant conclusions taken from Arnatkeviciute et al. [62] about psychiatric diseases. Mainly, regional cortical thickness differences are typically related to cell-type marker genes. Moreover, genes that have any bond with the medical condition schizophrenia through Genome-Wide Association Studies (GWASs) correlate with deviations in regional grey matter volume. Lastly, the research achieved a conclusion about the dysconnectivity in schizophrenia and bipolar disorder, hinting that dysconnectivity has an association with risk genes and structural variants, linking gene expression to macro-scale alterations [62].

3.1.2 Best Practices for Imaging Transcriptomics

Processing transcriptional data involves various choices that can affect the resulting spatial expression maps, leading to latent errors in research outcomes [26]. One needs to consider those choices in a moderate and careful manner. It is here that the best practices for imaging transcriptomics come into play.

The quantification of the expression of each gene is highly dependent on several conditional factors that could include the organism studied, the overall availability for each tissue, and the required spatial distribution [26]. It is essential for clinical contexts to apply and deploy imaging transcriptomics to investigate associations at the level of brain regions. Typically, regions that are physically closer to each other have more correlated gene expression values than other region pairs that are distant [26].

In the context of best practices, Arnatkeviciute et al. [26] emphasises the need for careful consideration in the analysis phases of the data, highlighting the relevance of selecting an appropriate matched gene set to evaluate the specificity of the findings [62]. A statistical gene-by-gene analysis Gene Category Enrichment Analysis (GCEA) can also be considered. A GCEA is a method that assesses if a defined set of genes shows any statistical differences between two different phenotypes [63].

Another topic is the importance of controlling the spatial auto-correlation in gene-phenotype associations, recommending specific null models to avoid biased results in GCEA, mitigate false positives, and reduce overestimation [62]. Since AHBA data is from a small set of post-mortem measurements, if we find a new relationship between the expression of a gene and an imaging phenotype, then this is purely correlational and does not directly reveal causality or causal mechanisms [64, 65]. Furthermore, the data retrieved from AHBA has only gene expression values for adult human brains. Thus, we need more information to get the full scope and a better understanding [66–68]. Moreover, the AHBA relies

on microarray analysis of bulk tissue samples, but since regional variants in cellular composition may affect this analysis, it is paramount and advised to cross-reference those results with *single-cell RNA sequencing* (single-cell RNAseq) [69, 70].

3.1.3 Molecular Maps with Spatial Transcriptomics

In contrast with imaging transcriptomics, spatial transcriptomics is a technique to study the effect of gene expression within the context of spatial organization of each tissue. It counts the number of transcripts of a gene at distinct spatial locations. Transcriptomic analysis only accounts for information about gene expression values, not considering a fine-grained spatial distribution of gene expression within the tissue. In this sense, spatial transcriptomics produces the data used to classify brain regions based on gene expression profiles, leading to the creation of reference frameworks that contain information about cell types and connectivity [23].

The definition of a brain region is largely opaque. To that end, Ortiz et al. [23] suggested four characteristics to better define a brain region: function; architecture; connectivity; and topography. Ortiz et al. [23] referred to many methods used to map gene expression in cells. The first one, named *bulk RNA sequencing* (bulk RNAseq), aims to reveal the expression of all the genes considered of the tissue cells. The second one is *single-cell RNA sequencing* (single-cell RNAseq), to capture the expression profile of individual cells. However, we lose the spatial organization of the cells in both methods [23].

In order to better analyse the data organized from the applied spatial transcriptomics techniques, there are standardized computational pipelines such as Seurat [71] and Scanpy [72], expanded from analysing and explaining only single-cell RNAseq data to analysing and visualizing spatial transcriptomics data [23]. The frameworks are useful because the datasets collected from spatial transcriptomics experiments can be extensively large [23].

To better analyse sequencing data from spatial transcriptomics experiments, we can take into account some key steps, mainly: (i) data normalization; (ii) dimensionality reduction; and (iii) clustering.

Regarding data normalization, it is fundamental to transform the gene expression values obtained in the whole dataset to have normalized expression values across the dataset. Moreover, there are a series of biases that could occur and that we must consider.

The first source of bias is related to differences in sequencing depth between samples. These differences lead to variations in the number of reads per gene. To address this, we can divide each gene count by the total number of reads per region. The second source of bias can originate from differences in gene expression between batches of processed tissue sections. To avoid this noise, different batch normalization algorithms have been suggested [73, 74]. The third origin of bias can be related to the number of reads detected for a gene depending on the length of the transcripts of that gene. For this third source of bias, transforming the values from gene counts to Transcripts Per Million (TPM) can be

a good option [75]. Lastly, a fourth source of bias can come from the sparsity of some genes when expressed. Genes that were not detected, despite being present, can result in a high number of missing values. Algorithms like Scran [76] or SCnorm [77] implement strategies to help distinguish between selective expression, missing data, and low sensitivity.

Upon completion of data normalization, we can use dimensionality reduction algorithms to compact the data representation while preserving important biological information. Principal Component Analysis (PCA), capturing most of the data variance with a compact set of components, and Independent Component Analysis (ICA) are commonly used [23]. Additionally, we can apply dimensionality reduction techniques to project high-dimensional data into 2D/3D space for visualization [23].

After reducing the data representation with dimensionality reduction techniques, we can identify the brain regions that share common molecular signatures [23].

3.1.4 Human Brain Proteome

Sjöstedt et al. [78] explored the expression of protein-coding and non-coding brain-enriched genes. The authors resorted to deep sequencing transcriptomics to analyse 95 human samples from 27 different tissue types. Consequently, the researchers classified the genes into various categories based on their expression levels across each tissue. Moreover, Sjöstedt et al. [78] explored the pattern of expression of long non-coding RNAs and their respective relationship with neighboring protein-coding genes. This approach permits the identification of coding and non-coding brain-enriched genes, with a focus on regulatory cross-paths. This study introduces the concept of cross-validating each gene with its corresponding gene ontology information, thereby enhancing the understanding of a specific gene expression value within a particular context.

3.2 Pattern Discovery in Omics Data

To address our second objective, we explored two major approaches for pattern discovery in omics data: (i) biclustering algorithms for two-way biological data analysis; and (ii) triclustering algorithms for three-way biological data analysis. Both biclustering and triclustering algorithms are essential for acquiring knowledge from gene expression data and will aid in identifying tissue-specific putative regulatory modules (patterns) [1, 46].

3.2.1 Biclustering Algorithms

Researchers commonly use biclustering for the discovery of putative regulatory modules (patterns) in both expression data, i.e., subsets of genes correlated with subsets of conditions, and network data, i.e.,

groups of coherently interconnected biological entities [79].

Cheng and Church [50] proposed an algorithm that defines a bicluster as a subset of samples and variables with a high similarity score. Moreover, the authors introduced the concept of residue to quantify the deviation between the actual and expected values, accounting for potential noise in the data [50]. In this approach, the algorithm identifies large and maximal biclusters with residue scores below a specified threshold, δ.

Prelić et al. [49] introduced a fast divide-and-conquer algorithm called Bimax, which they claim can find optimal biclusters in a reasonable time. Notwithstanding, the algorithm remains limited, as it imposes binarization of the data and does not handle noise well. In terms of implementation, Bimax requires repeating two main tasks multiple times with different initial configurations for better results: (i) rearranging the matrix with samples and variables, placing the entries with values equal to one in the upper right corner of the matrix; and (ii) divide the matrix into two submatrices, returning the submatrices with only values of one in the entries [49].

Kluger et al. [80] proposed the Spectral algorithm with the primary objective of identifying coherent value coclusters in the data. The algorithm can detect checkerboard patterns in the data matrix by integrating both samples and variables with a specialized normalization process that highlights potential coclusters within the data. Moreover, the Spectral algorithm is based on the concept that eigenvectors corresponding to characteristic expression patterns across genes or conditions can identify checkerboard structures. To identify these eigenvectors, researchers employ methods such as Singular Value Decomposition (SVD) [80].

Henriques et al. [79] developed BicPAMS, a software that facilitates the use of pattern-based biclustering algorithms with three major contributions: (i) it exhaustively mines non-constant types of biclusters, such as additive and multiplicative coherencies in the presence or absence of symmetries; (ii) it provides strategies to compose different biclustering structures; and (iii) it includes graphical and programmatic interfaces [79]. BicPAMS consists of three steps: (i) mapping, i.e., data preparation; (ii) mining, i.e., pattern discovery; and (iii) closing, i.e., post-processing [79].

Li et al. [81] introduced the QUBIC algorithm, which solves the biclustering problem in a more general form by employing a combination of qualitative and semi-quantitative measures of gene expression data and a combinatorial optimization technique. The authors claim that the algorithm can also solve the biclustering problem efficiently, e.g., a desktop computer required only a few minutes to process tens of thousands of genes with thousands of conditions [81].

Later, Xie et al. [82] proposed the QUBIC2 algorithm, inheriting the qualitative representation from the QUBIC algorithm. This new method uses a statistical test framework for testing the significance of the found biclusters in any organism, even if the organism does not have substantial functional annotations [82]. Moreover, the QUBIC2 algorithm consists of three sequential steps: (i) discretization of gene

expression data, followed by a split function; (ii) seeding and core biclustering, with the creation of cores; and (iii) dual biclustering, expanding the cores to recruit more genes and conditions under a preset consistency level [82].

3.2.2 Triclustering Algorithms

When analyzing gene expression data, researchers commonly use clustering and biclustering techniques to identify similar behavior patterns, which can indicate similar regulatory processes [83]. However, these techniques are sometimes insufficient when we need to consider a third dimension in our data, e.g., the effect of time or tissue spatial location on gene behavior.

In this context, triclustering algorithms act as a method for identifying genes with similar expression patterns over a segment of contexts within a subset of conditions. We can define a coherent tricluster as a group of genes that either: (i) show similar numerical values across the contexts and conditions, i.e., coherent values [1]; or (ii) demonstrate similar trends, regardless of the exact numerical values, by exhibiting correlated positive and negative changes in expression levels, i.e., coherent behavior [1].

The first triclustering approach, first proposed by Zhao and Zaki [45], comprises three essential steps: (i) constructing a multi-graph[1] with similar value ranges between all pairs of features; (ii) mining maximal biclusters from the multi-graphs constructed for each time point; and (iii) extracting the triclusters by merging similar biclusters from different context points. This triclustering approach is able to find different types of subspace clusters (patterns) encompassing the ones with constant or similar values along each dimension, and the ones with scaling and shifting expression patterns [45]. The authors claimed that the algorithm could find significant triclusters in real microarray datasets, being able to merge and delete some clusters that have large overlaps [45].

Yin et al. [84] proposed a new subspace clustering algorithm for time series gene expression data, denominated ts-cluster, which identifies all significant ts-clusters across different subsets of dimensions. This approach allows the genes to follow a certain time-shifting relationship, and tolerates noise by using relative expression magnitude instead of the absolute one [84].

Xu et al. [85] introduced an approach called LagMiner to find time-lagged 3D clusters. The authors stated that this new approach could identify triclusters that satisfied constraints of coherence, regulation, minimum gene number, sample subspace size, and period length.

Sim et al. [86] presented a new strategy to mine 3D clusters in real-valued data. At that time, the authors had concerns about discovering subspaces with a significant number of items and experimented with this new strategy in real-world case scenarios.

Gutiérrez-Avilés et al. [83] proposed the Triclustering-Genetic based (TriGen) algorithm, which can find pattern similarities for genes in a three-dimensional space, considering gene, conditions, and time

[1] A multi-graph is a graph that has more than one edge connecting a pair of vertices [45].

factors [83]. TriGen relies on a genetic algorithm and consists of four essential steps: (i) initialization, where the initial population is generated with consideration for overlap with previously identified triclusters; (ii) evaluation, for assessing the quality of each individual; (iii) selection, for determining which individuals will survive to the next generation; and (iv) mutation, where connections between pairs of individuals are established to share new genetic material [83]. Notably, TriGen has demonstrated the capability to extract putative regulatory modules related by their functional annotations from the knowledge base GO.

Soares et al. [87] proposed a new predictive model to achieve prognostic prediction over time, a triclustering-based classifier that learns prognostic models from three-way clinical data, taking advantage of the temporal dependence between the monitored features. Soares et al. [87] present this classifier in the context of Amyotrophic Lateral Sclerosis (ALS), and attempt to predict whether a specific patient will evolve to one of five clinical endpoints within a certain time window. The algorithm takes three-way data as input. The next natural step is preprocessing the data with the tasks of handling missing values and class imbalance. The third step is learning the triclustering best parameters and the fourth is learning the triclustering-based classifier. Finally, we can apply the trained predictor to forecast patient outcomes [87].

In the context of this work, we consider triclustering techniques to explore the three dimensions of our data: genes, tissues, and individuals. We are interested in finding relevant transcriptional signatures in different tissues to aid the learning of the target spatial maps.

4

Data Profiling

To establish the target maps of neurotranscriptomic activity, the GTEx dataset[1], capturing tissue-specific gene expression levels from human post-mortem cohorts, is hereby introduced. The dataset refers to data obtained from RNA sequencing experiments. In this thesis, we utilize 18 datasets from the GTEx initiative, version 8, one for each brain tissue present in Table 2.3, along with four peripheral tissues for comparison: (i) whole blood; (ii) adipose tissue (visceral); (iii) skin (not sun exposed); and (iv) skin (sun exposed).

Table 4.1: Example row of the GTEx dataset for the amygdala tissue.

Identifier	Name	Sample 1	Sample 2	Sample 3	...
ENSG00000227232.5	WASH7P	42	46	101	...

In Table 4.1, we simplified the name of each sample for visualization purposes. Every dataset contains exactly $56,200$ rows, while the number of columns varies depending on the tissue, present in Figure 4.1. The entries represent the expression value in TPM of a given gene (row) in a specific sample (column).

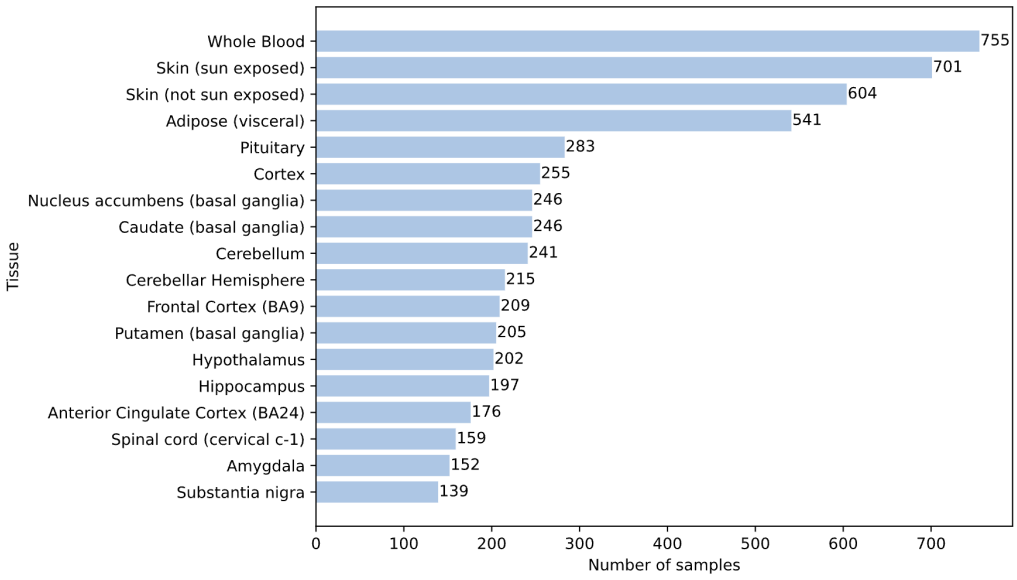


Figure 4.1: Number of samples for each tissue.

Among the 18 tissues, we have data from 931 different individuals. To further describe our population, we present the distributions of these individuals by sex, age, and agonal state. These distributions are illustrated in Figure 4.2, Figure 4.3, and Figure 4.4, respectively.

[1] GTEx dataset website.

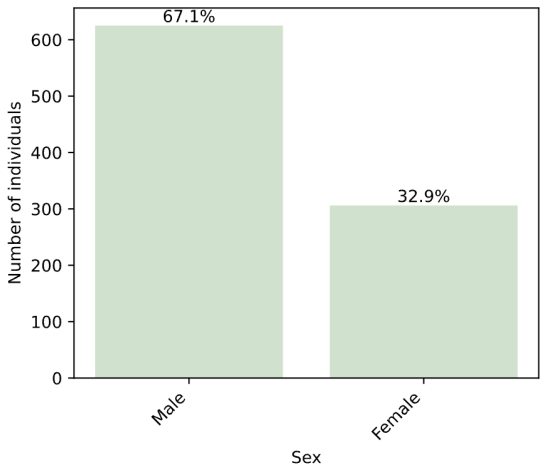


Figure 4.2: Sex distribution for all the individuals.

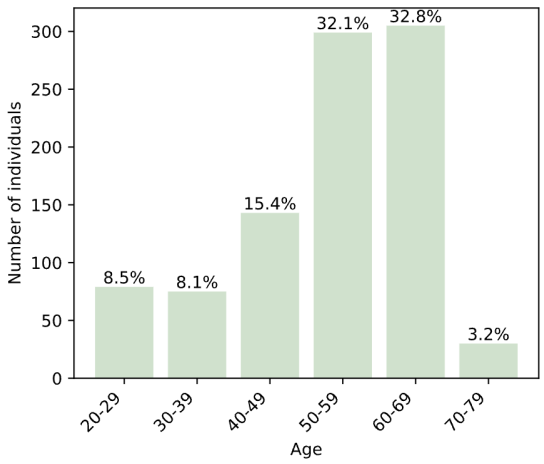


Figure 4.3: Age distribution for all the individuals.

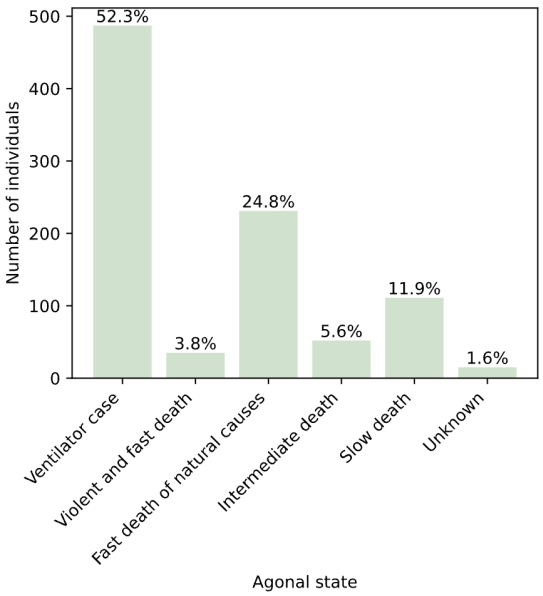


Figure 4.4: Agonal state distribution for all the individuals.

To better illustrate the information from the GTEx initiative, we select a gene and examine its expression across samples for each of the 18 tissues considered. In this context, we place special attention on a gene responsible for enabling the neurotransmitter GABA receptor activity, named GABBR2, according to the National Library of Medicine[2]. Figure 4.5 shows the expression of the GABBR2 gene across all samples in the 14 brain tissues (depicted in red) and the four peripheral tissues (depicted in orange) following logarithmic normalization.

[2]National Library of Medicine website.

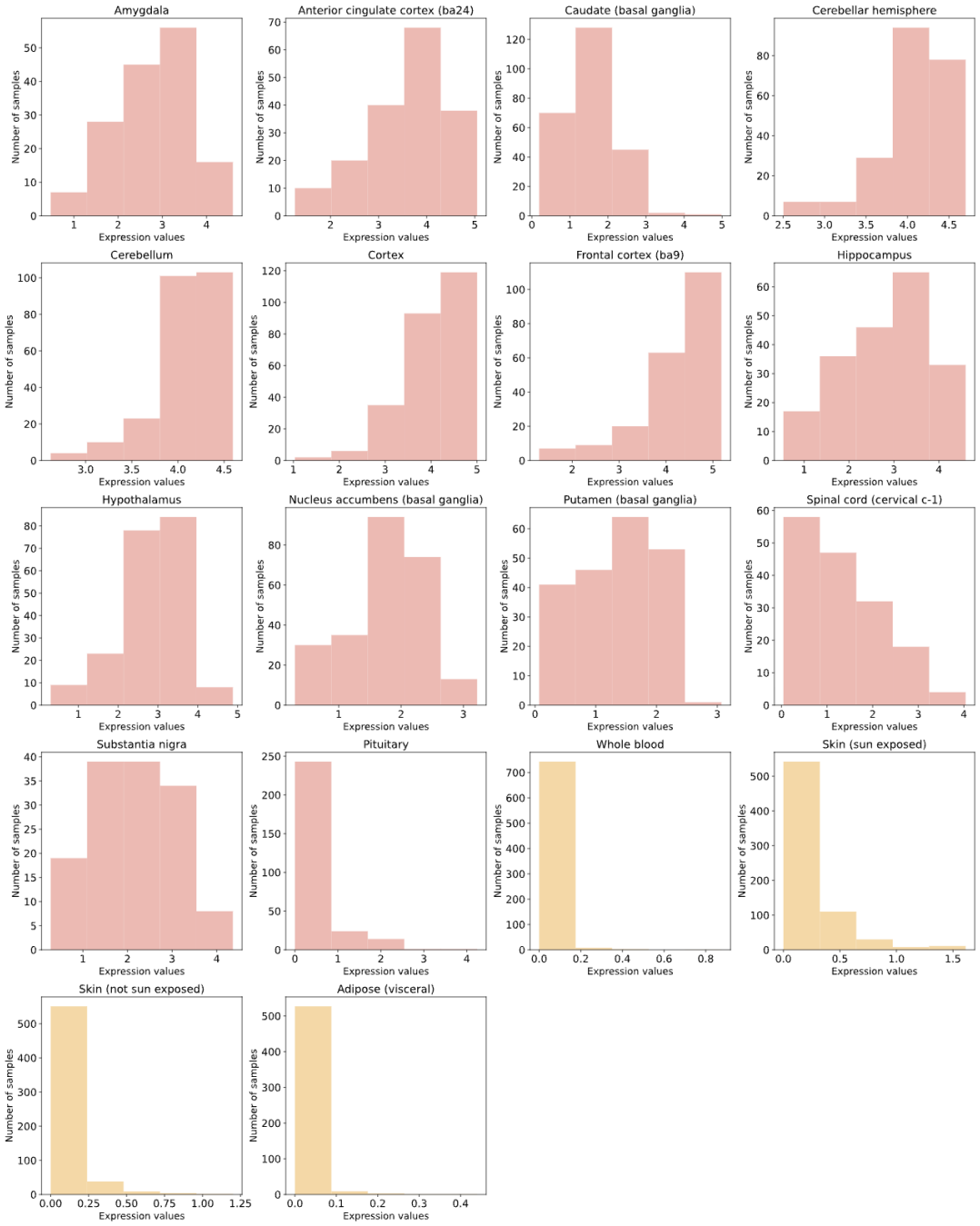


Figure 4.5: Distribution of GABBR2 across all the tissues and samples.

We observe relatively low expression levels of GABBR2 in both the caudate (basal ganglia) and the spinal cord (cervical c-1) across all samples. Conversely, tissues like the cerebellum exhibit an up-regulation (activation) of the GABBR2 gene. Furthermore, peripheral tissues such as the whole blood and the adipose tissue (visceral) predominantly show low expression levels for the GABBR2 gene, exhibiting down-regulation (inactivation).

5

Solution

Contents

5.1 Data Preprocessing 32

5.2 Differential Gene Expression for Multi-Tissue Analysis 35

5.3 Biclustering 38

5.4 Triclustering 41

The main goal of this work is to develop an application that enables the exploration of interactive neurotranscriptomic maps. This application will provide organized views of neurotransmission markers based on established functions using ontologies, as well as putative functions derived from functional enrichment through co-expression patterns.

To establish the desired neurotranscriptomic maps, we need to consider several aspects. Section 5.1 introduces a series of steps for gene-centric modelling of transcriptomic signatures between two different tissues. Section 5.2 extends this modelling by enabling the comparison of a set of input tissues with a different set of output tissues. The remainder sections describe the pattern-centric modelling steps for the complementary analysis of transcriptomic signatures under modular views. In particularly, Section 5.3 addresses the implementation of biclustering approaches to this end, while Section 5.4 focuses on triclustering approaches.

5.1 Data Preprocessing

A major stride in the context of this work is the preprocessing of gene expression data from the 18 tissues obtained from the GTEx initiative through a series of transformations. The entailed transformations are essential for facilitating the subsequent modeling of transcriptomic signatures between two distinct tissues, while also allowing for a focused analysis of specific genes and co-factors of interest.

5.1.1 Integrative Cross-Tissue Scores of Differential Expression

It is necessary to analyze the statistically significant differences in gene expression across various human tissues. To achieve this, we first need to normalize the data. Since we are dealing with RNA counts, we represent all gene expression values as TPM, followed by logarithmic function normalization,

$$\log_{10}(v+1), \tag{5.1}$$

where v represents the TPM ratio. The purpose of the logarithmic function is to normalize the distribution of counts, thereby mitigating skewness.

To determine if two different tissues exhibit statistically significant differences in gene expression for a particular gene, we must conduct a statistical assessment on the list of samples from each tissue. For this purpose, we first assess the normality of each list of samples using the Shapiro-Wilk normality test. Considering a 95% confidence interval, if at least one list of samples has a p-value below 0.05, we reject the null hypothesis, indicating that at least one list of samples does not follow a normal distribution. In this case, we apply a paired Wilcoxon signed-rank test, provided that the samples are paired (same number of individuals).

Conversely, if normality is not rejected under a significance threshold of 0.05, we cannot reject the null hypothesis. In this case, we can apply a Student's t-test between the two lists of samples.

Lastly, once we determine which statistical test to employ, we can proceed with the calculation of the p-value. Given a specific gene of interest, this value will indicate whether the gene expression exhibits statistically significant differences between two tissues. If the p-value is lower than 0.05, we reject the null hypothesis and conclude that the gene expression on the two tissues is statistically different. However, if the p-value is equal to or above 0.05, we cannot draw a definitive conclusion about the differential expression. We can observe the pseudocode of this statistical assessment in Algorithm 5.1.

Algorithm 5.1: Statistical Assessment

begin
- $\alpha \longleftarrow 0.05$
- $p1 \longleftarrow shapiro_wilk(sample1, \alpha)$
- $p2 \longleftarrow shapiro_wilk(sample2, \alpha)$
- **if** $p1 < \alpha$ **or** $p2 < \alpha$ **then**
  - $min_length \longleftarrow min(len(sample1), len(sample2))$
  - $p_value \longleftarrow wilcoxon_signed_rank(sample1[: min_length], sample2[: min_length], \alpha)$
- **else if** $p1 \geq \alpha$ **and** $p2 \geq \alpha$ **then**
  - $p_value \longleftarrow students_t_test(sample1, sample2, \alpha)$
- **return** p_value

5.1.2 Identification of Neurotransmission-Related Genes

To establish the target neurotranscriptomic maps, we focus on genes that are differentially expressed in nervous tissues, thereby indicating involvement in the regulation of neurotransmission systems. Our approach involves identifying these genes through automatic categorization based on their biological functions and pathways. Specifically, we aim to categorize genes according to their roles in key neurological processes, particularly those associated with neurotransmitters and their corresponding receptors.

With this end in mind, we retrieved a dataset containing the complete set of genes from humans (Homo sapiens) and their respective functions from the UniProt[1] initiative [88], which provides comprehensive protein sequence and functional information. For demonstration purposes, we present an example row retrieved from the dataset in Table 5.1.

[1] UniProt website.

Table 5.1: Illustrative gene description from the UniProt knowledge base.

Gene names	Function
DRD2	Dopamine receptor whose activity is mediated by G proteins which inhibit adenylyl cyclase, positively regulates postnatal regression of retinal hyaloid vessels via suppression of VEGFR2/KDR activity

Having the necessary information about each gene, we can then utilize Term Frequency-Inverse Document Frequency (TF-IDF), an information retrieval technique, to categorize the genes. In this context, we consider the set of functions of all the genes in the dataset to be the collection of documents, and we aim to identify the genes that have a higher TF-IDF score for a query with the terms: “receptor”, “brain”, “neuro”, “gaba”, and “dopamine”. We chose the first three terms because we want to identify the genes responsible for the reception of neurotransmitters in the brain. Additionally, we incorporated the last two to capture genes involved in the dopaminergic and GABAergic systems.

Before the automatic categorization of the genes, we normalize our documents by removing english stop words and applying a lemmatizer to reduce inflectional and variant forms to their base forms. Herein, lemmatization was observed to be preferable to stemming since it is more sensitive to context, helping us obtain more reliable results in the end. After this preprocessing, we can then calculate the final score for each gene with the following expression,

$$score = \alpha_1 \times score_{\text{“receptor”}} + \alpha_2 \times score_{\text{“brain”}} + \alpha_3 \times score_{\text{“neuro”}} + \alpha_4 \times score_{\text{“gaba”}} + \alpha_5 \times score_{\text{“dopamine”}},$$

where the score is the sum of the individual scores for the five query terms associated with each gene. Equal importance is assigned to each term, meaning the weights of the α parameters are uniformly distributed. In the end, to obtain the best results, we establish a threshold of 0.3 based on the analysis of the TF-IDF outcomes, retrieving only the genes with scores equal to or higher than this value. We obtain 220 genes, of which 196 are represented in the GTEx initiative datasets.

To test the automatically retrieved genes, we will allow the user to choose to utilize pre-established genes related to neurotransmitters of interest, such as those involved in dopaminergic and GABAergic systems, retrieved from the STRING[2] initiative [89], a tool that integrates protein-protein interactions.

5.1.3 Extension to Co-Factors of Interest

At this stage, we are equipped to explore the distribution of the previously selected genes when comparing two distinct tissues. Building on this, we can extend our analysis to filter samples based on

[2] STRING website.

co-factors such as psycho-demographic profiles, e.g., sex, age, and agonal states. This approach allows us to compare two distinct tissues in the context of specific co-factors of interest.

In the context of this work, and to have all the necessary information about each sample, we will resort to two different metadata files[3] present in the GTEx initiative: (i) a subject phenotypes dataset that contains, for each subject, their sex, age and agonal state; and (ii) a sample attributes dataset that contains relevant information about each sample. Together, we have a dataset that will allow us to explore the spatial distribution of neurotransmission markers in the presence of co-factors of interest. An example row of the dataset is shown in Table 5.2.

Table 5.2: Example row of the dataset created from the metadata.

Sample	Tissue	Sex	Age	Agonal state
GTEX-1117F-0003-SM-58Q7G	Whole Blood	Female	60-69	Slow death

It is essential to recognize that the sample identifiers in the GTEx initiative follow a specific structure designed to convey information about the individual (donor), and the sample itself. The structure of each identifier is given by the following,

`GTEX-[INDIVIDUAL_ID]-[TISSUE_ID]-SM-[SAMPLE_ID]`,

where `[INDIVIDUAL_ID]` refers to the individual's unique identifier (conferring anonymity), `[TISSUE_ID]` indicates the specific tissue sample from the donor (in the example row from Table 5.2, "0003" is the identifier of the whole blood), and `[SAMPLE_ID]` specifies the unique sample identifier for this particular tissue sample to distinguish it from other samples of the same tissue type.

In this work, we incorporate three distinct co-factors of interest, as presented in Table 5.2, aimed at facilitating the acquisition of information regarding gene expression across various phenotypes:

- Sex: Categorized into "Male" and "Female";
- Age: Segmented into the ranges 20-29, 30-39, 40-49, 50-59, 60-69, and 70-79;
- Agonal state: Classified as "Ventilator case", "Violent and fast death", "Fast death of natural causes", "Intermediate death", and "Slow death".

5.2 Differential Gene Expression for Multi-Tissue Analysis

At this stage, we can explore the spatial distribution of neurotransmission markers across a set of relevant genes in the presence of specific co-factors of interest. Nonetheless, we have only established

[3] GTEx metadata website.

a pairwise comparison between two different tissues, one input tissue and one output tissue, through a statistical assessment that culminates in a p-value calculation, and its subsequent conclusion about the statistical differences in expression.

In the context of this work, we want to perform multiwise comparisons between one or more tissues in the input set, and one or more tissues in the output set, to examine specific brain markers, i.e., genes with differential expression in brain tissues compared to peripheral tissues. This approach will enable us to gain knowledge pertaining to multiwise differences in spatial distribution of expression between groups of tissues. Typically, the input tissues correspond to tissues of interest, e.g., brain tissues, while the output tissues correspond to contrast tissues, e.g., peripheral tissues. In this case, we will also use the statistical assessment from Algorithm 5.1 applied in the pairwise comparison, now applying it to each pairwise combination of input and output tissue.

First, we consider the case where there is only one tissue in the input set, and more than one tissue in the output set. For demonstration purposes, we present Figure 5.1.

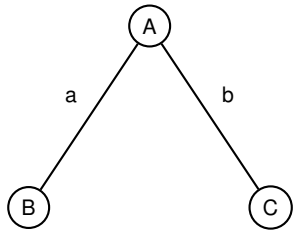


Figure 5.1: One input tissue (A) against multiple output tissues (B and C).

In the example depicted in Figure 5.1, a represents the p-value between tissues `A` and `B`, while b represents the p-value between tissues `A` and `C`. To conclude if the two sets of tissues are differently expressed, i.e., p-value < 0.05, we are going to take the maximum of the calculated p-values,

$$p\text{-}value_{\text{1v2}} = \max(a, b). \tag{5.2}$$

Selecting the maximum p-value as the final p-value allows us to take the most conservative stance regarding the differences in expression between tissues, which helps mitigate the risk of falsely declaring significance. If either comparison yields a significant result, i.e., a p-value below 0.05, then we can confidently conclude that at least one of the comparisons is significant.

Secondly, we face the scenario where we want to compare multiple tissues from the input set with only one tissue from the output set. To illustrate this scenario, we present Figure 5.2.

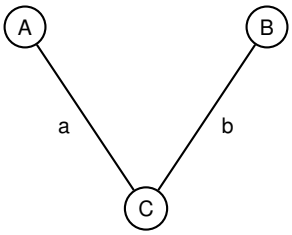


Figure 5.2: Multiple input tissues (A and B) against one output tissue (C).

In the example provided in Figure 5.2, a represents the p-value between tissues `A` and `C`, whereas b represents the p-value between tissues `B` and `C`. To conclude if the two sets of tissues are significantly differently expressed, we calculate the minimum between the calculated p-values,

$$p\text{-}value_{\text{2v1}} = \min(a, b). \tag{5.3}$$

By taking the minimum p-value, we are looking for the most significant difference between any of the input tissues and the single output tissue, meaning, genes with specific functions in the brain may be expressed only in certain regions of the brain. If either comparison yields a significant result, i.e., a p-value below 0.05, then the overall conclusion is that at least one of the input tissues is significantly different from the output tissue.

Lastly, we consider the situation where there are multiple tissues in the input set, and multiple tissues as well in the output set. To showcase this situation, we present Figure 5.3.

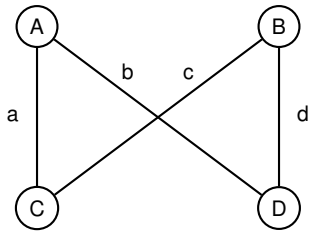


Figure 5.3: Multiple input tissues (A and B) against multiple output tissues (C and D).

In the situation shown in Figure 5.3, a represents the p-value between tissues `A` and `C`, b between tissues `A` and `D`, c between tissues `B` and `C`, and d between tissues `B` and `D`. To conclude if the two sets of tissues are significantly differently expressed, we are going to calculate the maximum for each input tissue list of p-values, and then retrieve the minimum of the resulting p-values,

$$p\text{-}value_{\text{2v2}} = \min(\max(a, b), \max(c, d)). \tag{5.4}$$

By applying the minimum of the maximums, we merge the methodologies used in the two cases above. When we calculate the maximum p-value for each input tissue list, we are identifying the largest

uncertainty observed within each group of comparisons. Then, we take the minimum of these maximum p-values to ensure that we select the most conservative estimate of significance across all comparisons.

Ultimately, given a list of tissues (input and output), genes, and conditions of interest, our solution filters candidate genes for brain-specific markers, ranking them according to their level of differential expression. It is worth mentioning that the proposed estimators can be altered to others, e.g., the median or percentiles. Additionally, the use of ANOVA would be insufficient to analyze differential expression, since we are only comparing two sets of populations at a time.

5.3 Biclustering

Moving from gene-centric to modular views of neurotranscriptomic signatures, an essential aspect of this work involves establishing biclustering algorithms to identify and explore potential patterns within our two-way data, uncovering subsets of genes with coherent expression patterns across specific tissues. This step is crucial for our pattern-centric modeling approach.

5.3.1 Problem Mapping

In the context of this work, we aim to identify putative regulatory modules (patterns) alongside sets of genes and tissues. To achieve this, we construct a matrix with genes as rows, and tissues as columns. This organization helps us find subsets of genes that exhibit coherent expression patterns across specific subsets of tissues. Moreover, in gene expression studies, it is a common convention to arrange data with genes as rows, and tissues as columns. This standardization facilitates the comparison and integration of different datasets.

The matrix entries correspond to the mean gene expression value in a given tissue across multiple samples. This approach offers several advantages, specifically: (i) noise reduction, averaging the expression values helps mitigate noise, resulting in clearer and more stable data; (ii) comparability, using the mean ensures that expression values are comparable across different tissues and genes, facilitating accurate analysis; and (iii) integration, having a single value for each gene and tissue enables the application of biclustering algorithms by eliminating the third dimension, which represents the expression levels for each sample.

Furthermore, we will consider the genes collected from the automatic gene identification described in Section 5.1.2, and we can construct the matrix using data from any number of tissues greater than one, sourced from the GTEx initiative. Figure 5.4 depicts the matrix representation, illustrating the arrangement of genes in the rows, and tissues in the columns.

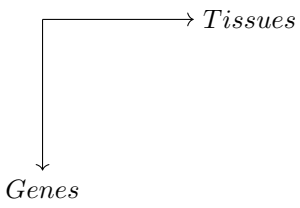


Figure 5.4: Biclustering problem mapping dimensions.

5.3.2 Algorithm Selection

With the data matrix constructed, we can now proceed with the application of the biclustering algorithms. In this work, we will focus on three distinct biclustering algorithms: the Spectral Coclustering [90], the Spectral Biclustering [80], and the BicPAMS algorithm proposed by Henriques et al. [79].

First, we consider the Spectral Coclustering algorithm[4] from the *Python* library *scikit-learn*. This algorithm aims to identify sets of genes that are co-regulated across subsets of conditions. It simultaneously clusters rows and columns of a data matrix to identify subgroups that exhibit coherent patterns. It achieves this by transforming the data into a graph representation and then applying spectral clustering techniques [90].

Second, we explore the Spectral Biclustering algorithm[5], also from the *Python* library *scikit-learn*. Unlike Spectral Coclustering, which focuses on co-regulated gene sets, Spectral Biclustering identifies subgroups of rows and columns that reveal coherent patterns. It achieves this by partitioning rows and columns under the assumption of an underlying checkerboard structure in the data. Spectral Biclustering transforms the data matrix into a bipartite graph and then applies spectral techniques to identify cohesive biclusters, which may not necessarily correspond to co-regulated genes [80].

Finally, we examine the BicPAMS algorithm, which facilitates the effective utilization of pattern-based biclustering. This algorithm extensively explores various types of biclusters, including additive and multiplicative coherencies, while considering both symmetrical and asymmetrical patterns within the data. Furthermore, BicPAMS provides strategies for constructing diverse bicluster structures efficiently, enabling the exploration of complex relationships within the data [79].

In this work, Spectral Coclustering and Spectral Biclustering serve as foundational tools to understand biclustering patterns before delving into a more complex method. The last approach allows for a comprehensive exploration of the data and deeper insights into the underlying structures. BicPAMS provides users with the flexibility to specify the type of bicluster they want to find, whether they are constant or order-preserving biclusters, enabling targeted analysis based on specific requirements. In contrast, simpler algorithms like Spectral Coclustering and Spectral Biclustering do not offer direct con-

[4] Spectral Coclustering website.
[5] Spectral Biclustering website.

trol over the type of bicluster to find. Instead, their objective is to identify coherent patterns within the data, regardless of whether they are constant or order-preserving biclusters. Moreover, BicPAMS offers additional flexibility in other steps of the algorithm, enhancing its utility for comprehensive analysis.

5.3.3 Hyperparametrization

In the biclustering algorithms, Spectral Coclustering and Spectral Biclustering, described in Section 5.3.2, we focus on a single parameter:

- `Number of clusters`: Maximum number of clusters the algorithm can identify. This value cannot exceed the number of columns (tissues) in our data. In the proposed methodology, we allow the user to customize it, selecting a value between two and the total number of tissues selected.

Contrarily, for the BicPAMS algorithm, we need to control a set of different hyperparameters related to the mining, mapping, and closing steps of the algorithm:

- `Pattern type`: In this work, we aim to identify both constant patterns and order-preserving patterns in our data. Users can choose the pattern type they wish to find;
- `Transpose`: For pattern orientation, we ensure that our matrix is not transposed (value equals false), as we aim to identify genes that are co-expressed in different tissues. This approach keeps the genes in the rows, and the conditions in the columns;
- `Mapper strategy`: To capture various possible results and enable comparison, users have the option to perform normalization and discretization by row, column, or at the entire dataset level. Additionally, discretization can be applied using equal-width, equal-frequency, or Gaussian cut-off points;
- `Number of bins`: In constant patterns, it is crucial to differentiate distinct expression levels without encountering issues related to discretization limits, such as expression values being too close to a frontier. To address this, we select five bins as the default option. For order-preserving patterns, our focus is solely on the value order, so we selected 20 as the number of bins.
- `Number of iterations`: We fix the number of iterations of the algorithm to two to explore regions that may be underexplored;
- `Stop conditions`: Our objective is to identify sets of genes exhibiting coherent expression across a customizable number of tissues. Depending on the selected number of tissues, we define the minimum number of columns in our patterns to be the maximum between two and the floor value of the square root of the selected number of tissues. Additionally, we set the minimum number of

biclusters to be retrieved from the algorithm to five, ensuring the extraction of a moderate number of patterns;

- `Merge overlap`: We aim to identify patterns with a quality of 70%, allowing for the presence of 30% noisy elements. To achieve this, we set this value to 0.70;

- `Filter overlap`: By fixing this value to 0.50, we ensure that patterns with a similarity higher than 50% to a more relevant pattern are removed to reduce redundancy. This approach will provide solutions with more diverse patterns.

5.4 Triclustering

To fully leverage pattern-centric modeling, a crucial step involves the application of triclustering algorithms. These algorithms allow us to delve into our three-way data, revealing subsets of genes that exhibit co-expression patterns specific to certain individuals and tissues.

5.4.1 Problem Mapping

Triclustering is a clustering technique specialized in finding patterns in three-way data, i.e., data with three dimensions. To perform triclustering, the available expression data is first mapped into a data cube with individuals as rows, genes as columns, and tissues as contexts (through the z-axis). Individuals will be placed on the x-axis since each individual represents a different observation. Genes, where each gene is a different variable, will be represented on the y-axis to allow the tricluster to effectively highlight patterns and correlations in gene expression across various individuals and tissues. Tissues will be situated on the z-axis to capture interactions between genes and individuals in different spatial contexts.

We will organize our information in this manner because we are interested in studying the effect of each gene through each individual, within various tissue contexts. Notwithstanding, visualizing individuals, genes, and tissues in this structured manner simplifies the interpretation of triclustering analysis results. This organization allows us to identify biologically meaningful clusters that can further our understanding of regulatory mechanisms and functional relationships.

With that in view, we consider for the cube entries the gene expression value for a given individual, gene, and tissue. Additionally, we consider the genes collected from the automatic gene identification described in Section 5.1.2. The cube is then constructed with data from any number of tissues greater than one, sourced from the GTEx initiative. Figure 5.5 demonstrates the representation of the cube, illustrating the arrangement of individuals in the rows, genes in the columns, and tissues along the z-axis.

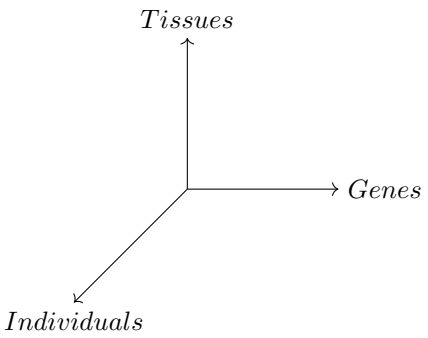


Figure 5.5: Triclustering problem mapping dimensions.

5.4.2 Algorithm Selection

For our triclustering approach, we select the state-of-the-art TriCluster algorithm proposed by Zhao and Zaki [45]. This robust and flexible algorithm can identify clusters of varying sizes and shapes while simultaneously clustering individuals, genes, and tissues (contexts).

Moreover, our choice is supported by the demonstrated effectiveness and reliability of the algorithm across numerous research contexts, making it a valuable tool for discovering coherent patterns of gene expression. By analyzing subsets of individuals, genes, and tissues, the TriCluster algorithm uncovers biologically meaningful insights crucial for our study, and for the discovery of putative regulatory modules (patterns) in our data.

5.4.3 Hyperparametrization

As part of the triclustering approach described in Section 5.4.2, our application requires several distinct hyperparameters to be specified and controlled during execution. Although the default value of the δ parameter was maintained throughout the analysis, a set of hyperparameters plays a crucial role in shaping the behavior and outcomes of our analysis:

- `Minimum number of individuals`: This hyperparameter defines the minimum size of our triclusters in terms of rows, representing the number of individuals included. The number of individuals considered depends on the tissues selected by the user prior to the execution of the triclustering algorithm. For each case, we intersect the individuals from each selected tissue to determine the number of individual data points available for the entire set of selected tissues. The minimum size of individuals is then set as a selectable option for the user, ranging between two and maximum number of individuals obtained from the intersection;

- `Minimum number of genes`: This hyperparameter details the minimum number of columns, representing the number of genes included. The retrieved set of 196 genes with brain-related functions,

like presented in Section 5.1.2, is herein suggested for the triclustering analysis. Therefore, we will use that value as the maximum number for the selectable option, with two being set as the minimum;

- `Minimum number of tissues`: This hyperparameter, also adaptable by the user, denotes the depth, representing the number of tissues included. This number is defined as a range between two and the number of tissues selected by the user.

6

Visualization Tool

Contents

6.1 Gene-Centric Tab 46
6.2 Pattern-Centric Tab 48

To facilitate the spatial mapping of transcriptomic neuromarkers, and the detection of putative regulatory modules (patterns) of neurotransmission in various tissues, we developed a web-based visualization tool using the *Dash Plotly*[1] *Python* library. This tool has two distinct components, both organized into different tabs within the application: (i) gene-centric modeling, referred to in Section 6.1; and (ii) pattern-centric modeling, mentioned in Section 6.2.

6.1 Gene-Centric Tab

In Figure 6.1, we present the display of the gene-centric tab of the application, where each parameterizable value is marked as `[P1]` through `[P5]`.

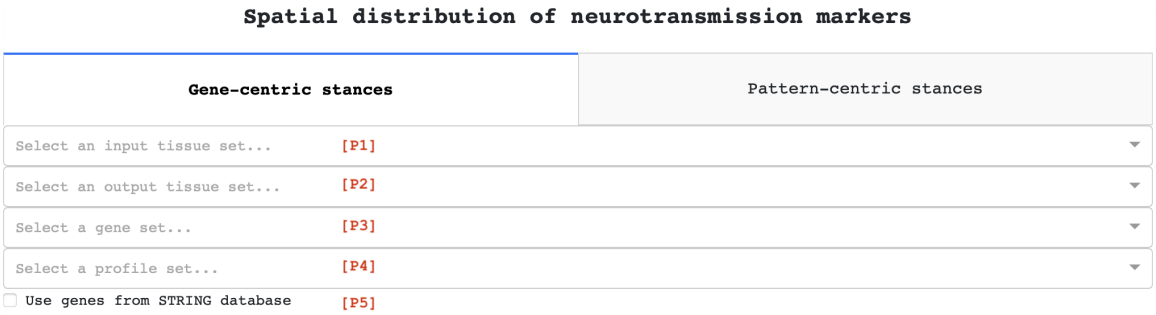


Figure 6.1: Gene-centric tab.

We now proceed to explain each parameter in detail: (i) `[P1]` dropdown denotes the tissues in the input set, offering 18 options with no limit on the number of selections; (ii) `[P2]` dropdown refers to the tissues in the output set, similar to the input set, with 18 possible tissue options, ensuring that each tissue is mutually exclusive and cannot be used in both the input and output sets; (ii) `[P3]` dropdown marks the genes retrieved from automatic gene identification, allowing the user to select one or more genes; (iv) `[P4]` dropdown represents different psycho-demographic profiles, e.g., sex, age, and agonal state, with the ability to apply multiple filters; and (v) `[P5]` checkbox serves as an alternative for the automatically retrieved genes, enabling the user to utilize pre-established genes related to neurotransmitters of interest in the `[P2]` dropdown, such as those involved in dopaminergic and GABAergic systems retrieved from the STRING initiative.

In this tab, we present two different views, namely, one for when only a single gene is selected, and another for when more than one gene is selected. We provide a simple example in Figure 6.2 and Figure 6.3, respectively.

[1] Dash Plotly website.

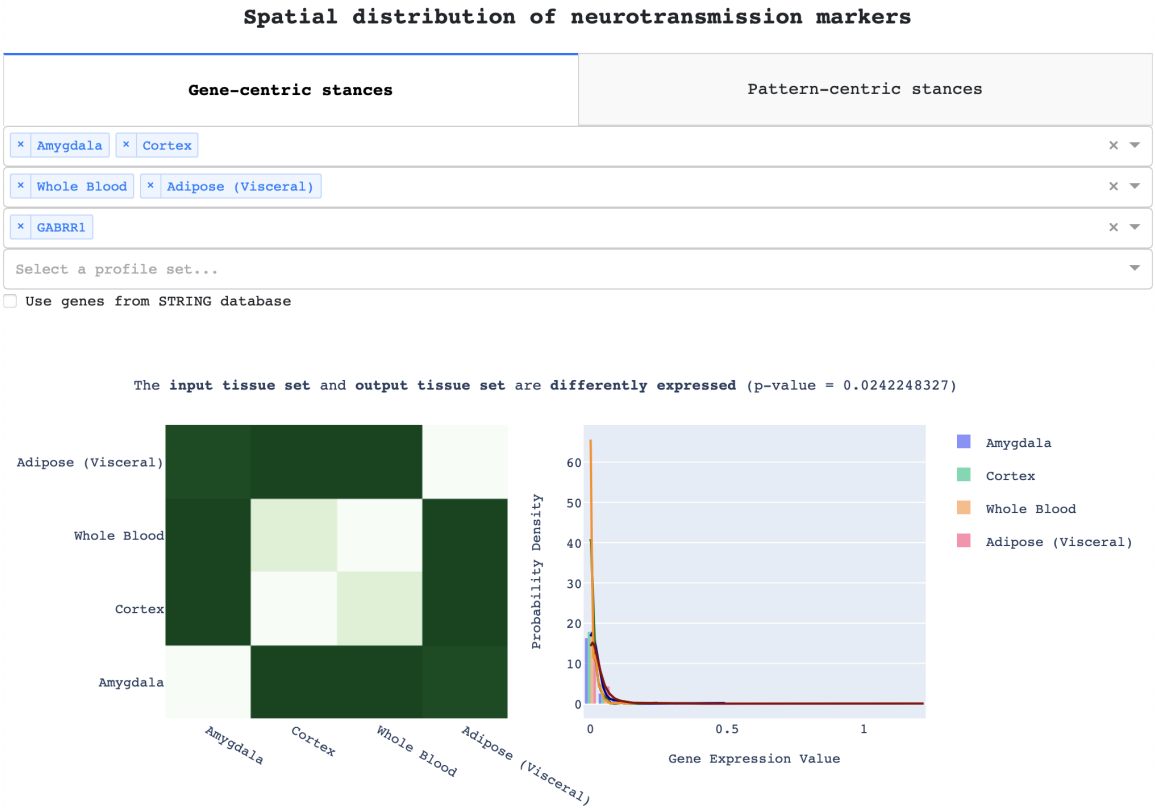
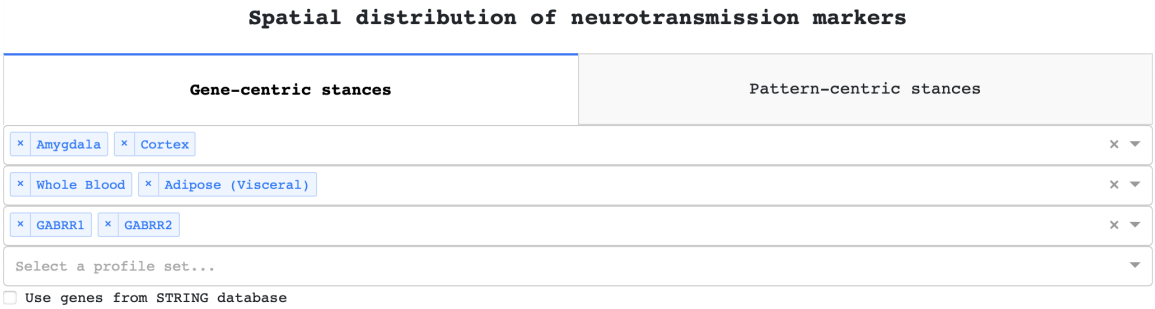


Figure 6.2: View for one gene selected.



Selected Genes	Occurrences	Scores (p-values)	Tissue Expressions
GABRR2	4	0	More details here...
GABRR1	3	0.0242248327	More details here...

Input - Output	Whole Blood	Adipose (Visceral)
Amygdala	2	2
Cortex	1	2

Figure 6.3: View for more than one gene selected.

In Figure 6.2, we selected two tissues in the input set (amygdala and cortex), two tissues in the output set (blood and adipose tissue), and a single gene of interest (GABRR1). The application provides an interactive heatmap representing the p-values calculated for each combination of tissues, where darker green colors indicate lower p-values, highlighting statistical differences in expression. Additionally, the application includes an interactive kernel density estimator to show the distribution of the expression values of the selected gene across the samples in each chosen tissue. Finally, we provide a textual conclusion on whether the input and output sets exhibit differential expression, calculated as demonstrated in Section 5.2.

Regarding Figure 6.3, we maintained the selected tissues while adding a new gene (GABRR2). This figure presents a first table, ordered by p-value, that summarizes information for each gene, including the number of times it is differentially expressed for each combination of tissues, the final calculated p-value for the gene, and the mean and standard deviation of the gene expression for each tissue. The table provides additional details upon clicking each row: (i) the first column displays a kernel density estimator, similar to the view with only one gene; (ii) the third column presents a boxplot with all the p-values used in the calculation of the final p-value, if three or more p-values are available; and (iii) the fourth column offers textual information. Furthermore, each row in the table is colour-coded based on the p-value conclusion, ranging from dark green for the lowest p-value to red for the highest.

Lastly, we also present a second table (non-clickable) that indicates the number of genes exhibiting differential expression between each combination of input and output tissues.

6.2 Pattern-Centric Tab

In Figure 6.4, we present the display of biclustering algorithms from the *scikit-learn* library in the pattern-centric tab of the application, where each parameterizable value is labeled from `[P6]` to `[P8]`.

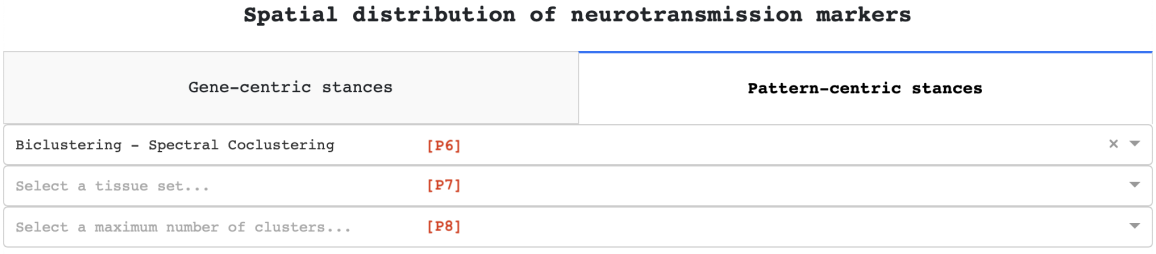


Figure 6.4: Pattern-centric tab for biclustering algorithms from scikit-learn.

We now explain the parameters: (i) `[P6]` dropdown refers to the algorithm chosen by the user, with options including “Spectral Coclustering”, “Spectral Biclustering”, “BicPAMS”, and “TriCluster”; (ii) `[P7]` dropdown denotes the tissues selected for the algorithm, where at least two tissues must be chosen to

run any algorithm; and (iii) [P8] dropdown represents the maximum number of clusters the algorithm should find.

In Figure 6.5, we showcase the display of the BicPAMS algorithm in the pattern-centric tab of the application, where the new parameterizable values are labeled as [P9] and [P10].

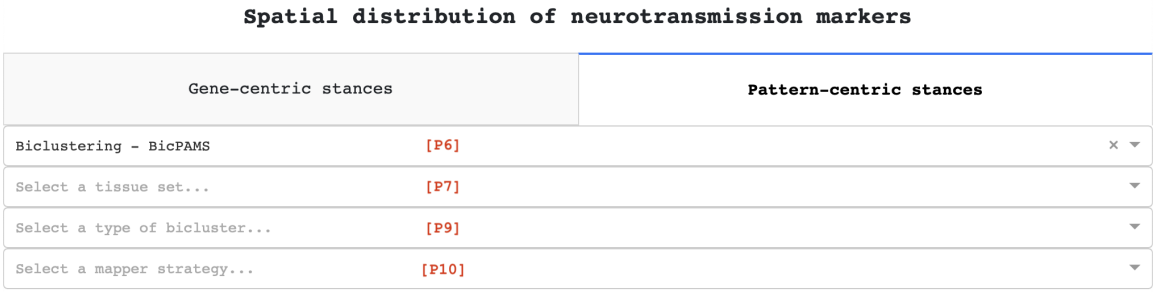


Figure 6.5: Pattern-centric tab for BicPAMS algorithm.

In this case, we have the following parameters to consider: (i) [P9] dropdown represents the type of biclusters to find in the algorithm, either constant or order-preserving biclusters; and (ii) [P10] refers to the mapper strategy to be considered.

In Figure 6.6, we illustrate the display of the TriCluster algorithm in the pattern-centric tab of the application, where each parameterizable value is labeled from [P11] to [P13].

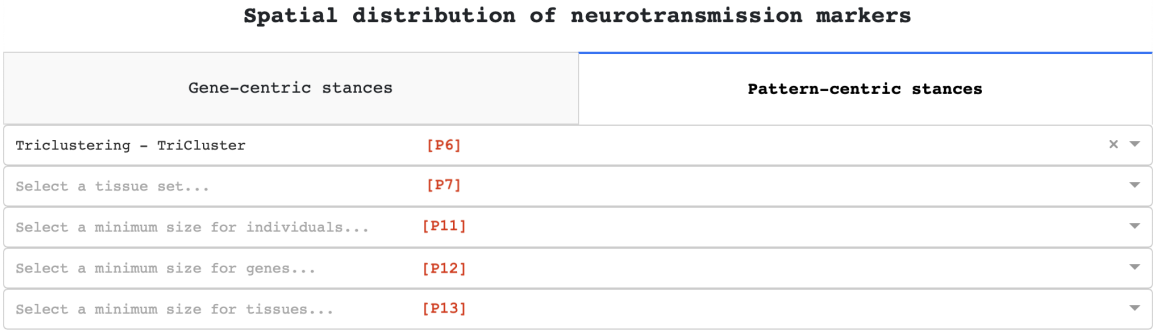


Figure 6.6: Pattern-centric tab for TriCluster algorithm.

For this algorithm, we consider the following parameters: (i) [P11] dropdown refers to the minimum size of the triclusters in terms of individuals, with a value between two and the total number of individuals present for the selected tissues; (ii) [P12] dropdown represents the minimum size of the triclusters in terms of genes, with a value between two and the maximum number of genes, which is 196; and (iii) [P13] dropdown refers to the minimum size of the triclusters in terms of tissues, with a value between two and the number of selected tissues.

For illustration purposes, we show an example of the application's display after executing the BicPAMS biclustering algorithm in Figure 6.7.

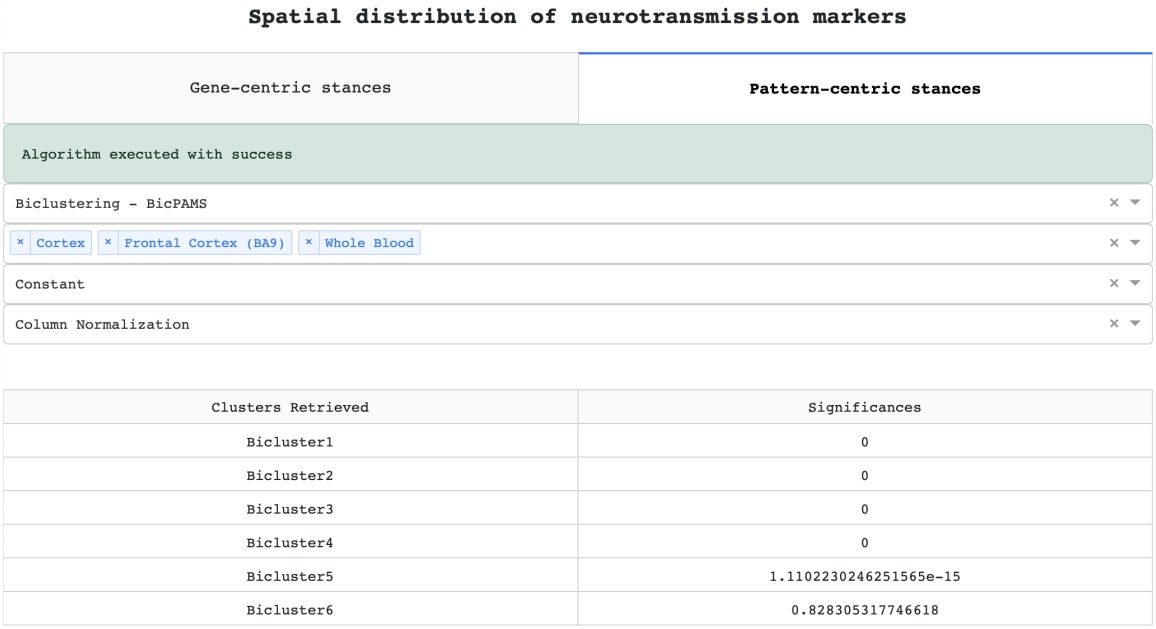


Figure 6.7: View displaying an illustrative BicPAMS solution.

In Figure 6.7, we selected the BicPAMS algorithm and specified three tissues (cortex, frontal cortex, and blood). We aimed to find constant biclusters after performing column normalization on our data. The application provides an interactive table that showcases all the biclusters found, ordered by significance. By clicking on each bicluster, a pop-up opens, displaying the dimensions, row and column labels, and details of the selected bicluster.

7

Results

Contents

7.1 Gene-Centric Results 52
7.2 Pattern-Centric Results 62

This chapter provides an assessment of the proposed methodology, with a focus on the relevance of the learned visual maps for knowledge acquisition. Section 7.1 presents results at the gene level, starting with a general overview and then analyzing a particular neurological system of interest, specifically, the one related to the dopamine neurotransmitter. Section 7.2 details results at the pattern level, encompassing the identification and functional analysis of putative regulatory modules (patterns).

7.1 Gene-Centric Results

The gene-centric assessment of neurotranscriptomic markers is conducted in three stages, corresponding to coarse, intermediary, and fine grained views. First, we initiate our testing with a broad analysis encompassing all $56,200$ genes supplied by the GTEx initiative. Subsequently, we refine our scope to focus on the 196 automatically retrieved genes and provide a brief results description. Finally, we narrow down to an in-depth analysis of 11 pre-established genes associated with the dopaminergic system. We will examine the differential expression between various configurations of brain and peripheral tissues throughout the following two subsections.

7.1.1 General Overview

To gain a comprehensive overview of the results for all $56,200$ genes, we assess the number of genes exhibiting differential expression between two different groups of tissues, in this case we will consider a gene differentially expressed between two tissues if the acquired p-value is below 0.01. Specifically, the first group contains four brain tissues: the amygdala, the frontal cortex (BA9), the hippocampus, and the hypothalamus. Conversely, the second group contains three peripheral tissues: the whole blood, the adipose (visceral) tissue, and the skin (not sun exposed). The results are depicted in Table 7.1.

Table 7.1: Number of differentially expressed genes for all $56,200$ genes.

	Whole Blood	Adipose (Visceral)	Skin (Not Sun Exposed)
Amygdala	$37,264$	$36,718$	$38,079$
Frontal Cortex (BA9)	$39,864$	$38,206$	$39,623$
Hippocampus	$38,706$	$37,620$	$38,950$
Hypothalamus	$39,143$	$37,731$	$39,156$

For demonstration purposes, we select two brain tissues, two peripheral tissues, and one gene to test the multiwise comparisons between two distinct sets of tissues. From the above table, we select the hippocampus and the hypothalamus as the input tissue set, and the whole blood and the adipose (visceral) tissue as the output tissue set, for the gene COMT. Figure 7.1 presents the results.

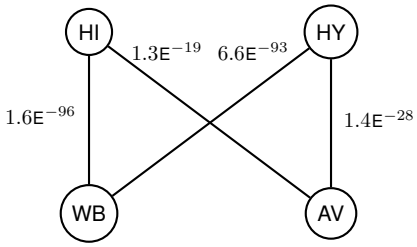


Figure 7.1: Hippocampus and hypothalamus versus whole blood and adipose (visceral) tissue.

$$p\text{-}value_{2\text{v}2} = \min(\max(1.6\text{E}^{-96}, 1.3\text{E}^{-19}), \max(6.6\text{E}^{-93}, 1.4\text{E}^{-28})) = 1.4\text{E}^{-28}. \tag{7.1}$$

As the final p-value is below 0.05 significance threshold, COMT gene can be said to be differentially expressed on the brain tissues when considering peripheral tissues as the reference contrast tissues.

By narrowing the scope to the 196 automatically retrieved genes, we focus on those specifically implicated in the regulation of neurotransmission systems. We now assess the number of genes exhibiting differential expression between two different groups, the amygdala and the frontal cortex (BA9) versus the whole blood. We categorize the results based on the level of their expression differences. As previously mentioned, the p-value is determined as the maximum of the p-value obtained from comparing the amygdala with the whole blood, and the p-value obtained from comparing the frontal cortex (BA9) with the whole blood. We can visualize the results mentioned in Table 7.2.

Table 7.2: Distribution of the 196 neurologically related genes across various p-value ranges.

	Number of genes
Genes with p-value < 0.001	182
Genes with $0.001 \leq p$-value < 0.01	4
Genes with $0.01 \leq p$-value < 0.05	3
Genes with $0.05 \leq p$-value < 0.10	0
Genes with p-value ≥ 0.10	4

Observing Table 7.2, we can conclude that most of the considered genes have a very high level of differential expression between the two sets of tissues, corresponding to a p-value lower than 0.001. This observation underscores the tissue-specific nature of gene expression between these groups. Such differential expression is expected, as genes that vary between brain regions and peripheral tissues typically serve pivotal roles in specialized biological processes unique to each tissue type. Furthermore, three genes were excluded as they were not sufficiently representative in any of the three tissues.

Additionally, the number of genes that exhibit differential expression between pairs of tissues is presented in Table 7.3.

Table 7.3: Number of differentially expressed genes for the 196 automatically retrieved genes with known brain-specific functions.

	Whole Blood
Amygdala	185
Frontal Cortex (BA9)	182

Upon observing Table 7.3, it is evident that a significant number of genes exhibit differential expression between each combination of input and output tissues. As previously mentioned, this is in line with expectations, given the comparison between two distinct types of tissues, brain versus peripheral.

7.1.2 Dopaminergic System

We now turn our focus to the dopaminergic system, specifically analyzing 11 established genes identified by the STRING initiative as being directly involved in dopaminergic processes. The biological functions of each established gene within the dopaminergic system are firstly synthesized to broaden our understanding of its overall function. This information is present in Table 7.4, and it was obtained by selecting the lowest adjusted p-value from the GO initiative.

Table 7.4: Biological functions of the genes from the dopaminergic system (from the GO initiative).

Gene	Function
SLC6A2	Dopaminergic synaptic transmission
	Dopamine uptake
SLC6A3	Dopaminergic synaptic transmission
	Dopamine uptake
	Dopamine transport
SLC18A2	Dopamine transport
TH	Dopamine metabolic process
	Dopaminergic synaptic transmission
DBH, LRTOMT	Dopamine catabolic process
	Dopamine metabolic process
COMT	Dopamine metabolic process
DRD2, DRD3, DRD4, DRD5	Dopamine receptor

We compare one brain tissue against a second brain tissue, specifically, the amygdala versus the frontal cortex (BA9), to identify differentially expressed genes within the set of 11 pre-established genes. We present the results in Table 7.5.

Table 7.5: Scores and regulation conclusion for amygdala versus frontal cortex (BA9).

Gene	Score	Regulation
DBH	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD5	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD2	4.5E^{-9}	Up-regulated (activated) on the amygdala
SLC6A3	9.55E^{-8}	Up-regulated (activated) on the amygdala
DRD3	3.253E^{-5}	Up-regulated (activated) on the frontal cortex (BA9)
SLC6A2	2.015E^{-4}	Up-regulated (activated) on the amygdala
DRD4	2.572E^{-4}	Up-regulated (activated) on the frontal cortex (BA9)
SLC18A2	0.154	No conclusion
LRTOMT	0.379	No conclusion
COMT	0.738	No conclusion
TH	0.933	No conclusion

As depicted in Table 7.5, we conclude that 7 out of the 11 genes have a final p-value lower than 0.001 (represented in green), indicating significant differential expression between the amygdala and the frontal cortex (BA9) in this gene set. The remaining four genes (represented in red) have a p-value higher than 0.05, which prevents a definitive conclusion about their differential between these brain tissues.

The table also indicates which tissue is up-regulated (activated) relative to the other for each gene. We derived this conclusion from the distribution of expression levels for each sample and gene. Figures 7.2 and 7.3 illustrate this point by showing examples of a gene up-regulated (activated) in the frontal cortex (BA9) and a gene up-regulated (activated) in the amygdala, respectively.

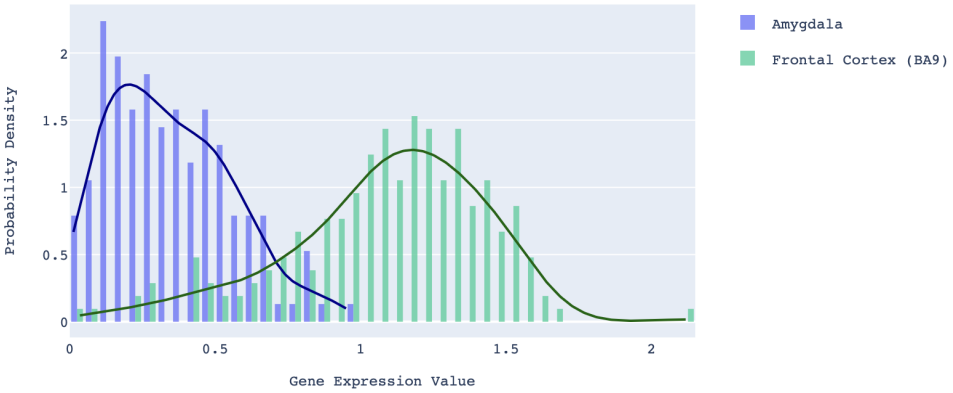


Figure 7.2: Distribution of the expression of the gene DBH.

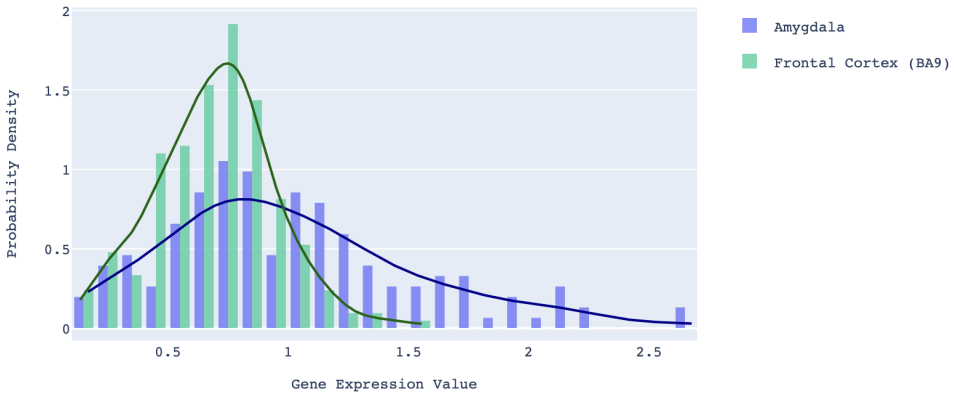


Figure 7.3: Distribution of the expression of the gene DRD2.

Figure 7.2 illustrates the distribution of expression levels for the gene DBH in both the amygdala and the frontal cortex (BA9). As shown, the mean expression level in the frontal cortex (BA9) is higher than in the amygdala, indicating that the gene DBH is up-regulated (activated) in the frontal cortex (BA9). Using the visualization tool, we also obtained, for each gene, the mean and standard deviation values to confirm this observation. Those values are present in Table 7.6.

Table 7.6: Mean and standard deviation of the gene DBH.

	Mean	Standard deviation
Amygdala	0.346	0.208
Frontal Cortex (BA9)	1.091	0.333

In contrast, Figure 7.3 shows the distribution of expression levels for the gene DRD2 in both the amygdala and the frontal cortex (BA9). Here, the mean expression level is slightly higher in the amygdala compared to the frontal cortex (BA9), despite the higher variance in the amygdala. To confirm this observation, we can refer, for each gene, to the mean and standard deviation values for each gene, as presented in Table 7.7.

Table 7.7: Mean and standard deviation of the gene DRD2.

	Mean	Standard deviation
Amygdala	1.018	0.520
Frontal Cortex (BA9)	0.697	0.246

By identifying which genes are up-regulated (activated) in each tissue, we can acquire knowledge of pathway enrichment, ontologies, and disease associations for each set of up-regulated (activated) genes. This analysis will help us understand the functional significance of these genes in each tissue

and draw comprehensive conclusions about their roles. That knowledge is present in Table 7.8 and Table 7.9.

Table 7.8: Enriched terms associated with the up-regulated (activated) dopamine-related genes in the amygdala (DRD2, SLC6A3, SLC6A2).

Category	Description	Adjusted p-value
Pathways	Na+/Cl- Dependent Neurotransmitter Transporters R-HSA-442660 (Reactome 2022)	4.102E^{-5}
	Dopaminergic synapse (KEGG 2021 Human)	4.735E^{-4}
Ontologies	Dopaminergic Synaptic Transmission and Uptake (GO Biological Process 2023)	7.621E^{-6}
	Monoamine Transmembrane Transporter Activity (GO Molecular Function 2023)	4.678E^{-6}
Diseases	Social Phobia (DisGeNET)	1.961E^{-7}
	Impulsive Character (DisGeNET)	2.485E^{-6}
	Schizophrenia (DisGeNET)	0.03

Table 7.9: Enriched terms associated with the up-regulated (activated) dopamine-related genes in the frontal cortex (DBH, DRD5, DRD3, DRD4).

Category	Description	Adjusted p-value
Pathways	Dopamine Receptors R-HSA-390651 (Reactome 2022)	5.099E^{-10}
	Dopaminergic synapse (KEGG 2021 Human)	5.593E^{-6}
Ontologies	Dopamine Metabolic Process (GO Biological Process 2023)	3.050E^{-7}
	Postsynaptic Neurotransmitter Receptor Activity (GO Molecular Function 2023)	5.124E^{-5}
Diseases	Nicotine Dependence (DisGeNET)	4.025E^{-7}
	Conduct Disorder (DisGeNET)	6.602E^{-7}
	Schizophrenia (DisGeNET)	5.186E^{-4}

The enrichment analysis reveals that the identified genes are associated with distinct biological processes and pathways specific to each brain tissue. Up-regulated (activated) genes in the amygdala are primarily involved in neurotransmitter transport and dopaminergic synaptic transmission. These genes have a low adjusted p-value for disorders such as social phobia and impulsivity, suggesting a crucial role in emotional regulation and behavioral responses mediated by the amygdala. Conversely, in the frontal cortex (BA9), up-regulated (activated) genes are associated with dopamine receptor signaling, metabolic processes, and diseases affecting cognitive and behavioral functions. The enrichment of schizophrenia-related pathways in both tissues implies that these genes may contribute to the disease through distinct mechanisms in each brain region.

We now aim to investigate sex-specific functions within the dopaminergic system by analyzing differences in expression across 11 genes between distinct sexes. Using the visualization tool, we applied the sex-specific filter, resulting in separate results for males and females. Table 7.10 presents the results for males, while Table 7.11 shows the findings for females.

Table 7.10: Scores and regulation conclusion for amygdala versus frontal cortex (BA9) on males.

Gene	Score	Regulation
DBH	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD5	5E^{-10}	Up-regulated (activated) on the frontal cortex (BA9)
SLC6A3	7.043E^{-6}	Up-regulated (activated) on the amygdala
DRD2	7.487E^{-6}	Up-regulated (activated) on the amygdala
DRD3	2.068E^{-4}	Up-regulated (activated) on the frontal cortex (BA9)
DRD4	0.003	Up-regulated (activated) on the frontal cortex (BA9)
SLC18A2	0.018	Up-regulated (activated) on the frontal cortex (BA9)
SLC6A2	0.084	No conclusion
LRTOMT	0.512	No conclusion
TH	0.816	No conclusion
COMT	0.851	No conclusion

Table 7.11: Scores and regulation conclusion for amygdala versus frontal cortex (BA9) on females.

Gene	Score	Regulation
DBH	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD5	3.514E^{-6}	Up-regulated (activated) on the frontal cortex (BA9)
DRD2	3.232E^{-5}	Up-regulated (activated) on the amygdala
SLC6A3	5.506E^{-5}	Up-regulated (activated) on the amygdala
SLC6A2	0.004	Up-regulated (activated) on the amygdala
DRD3	0.064	No conclusion
TH	0.267	No conclusion
DRD4	0.335	No conclusion
LRTOMT	0.446	No conclusion
COMT	0.763	No conclusion
SLC18A2	0.806	No conclusion

As shown in Table 7.10 and Table 7.11, there are 8 out of the 11 genes differentially expressed between the two brain tissues in males (represented in green, lighter green, and yellow), while only five genes in females (represented in green and lighter green). Moreover, we can conclude that in males, more genes are differentially expressed and up-regulated (activated) in the frontal cortex (BA9) compared to the amygdala. Conversely, in females, more genes are differentially expressed and up-regulated (activated) in the amygdala. We illustrate the intersection and exclusive differentially expressed genes between these two tissues in the Venn diagram in Figure 7.4.

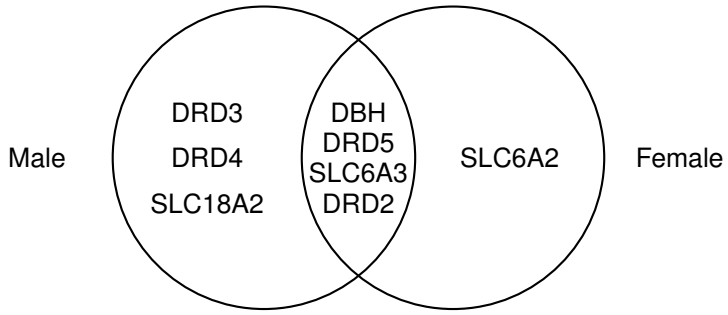


Figure 7.4: Dopamine-related genes differentially expressed depending on the sex.

By identifying which genes are up-regulated (activated) in each tissue for each sex, we gain insights similar to those discussed without the sex filter. The insights for males are present in Table 7.12 for the amygdala, and Table 7.13 for the frontal cortex (BA9).

Table 7.12: Enriched terms associated with the up-regulated (activated) dopamine-related genes in the amygdala for males (SLC6A3, DRD2).

Category	Description	Adjusted p-value
Pathways	Dopamine Receptors R-HSA-390651 (Reactome 2022)	0.008
	Dopaminergic synapse (KEGG 2021 Human)	2.378E^{-4}
Ontologies	Dopamine Transport (GO Biological Process 2023)	0.005
	Dopamine Uptake (GO Biological Process 2023)	0.005
	Monoamine Transmembrane Transporter Activity (GO Molecular Function 2023)	0.002
Diseases	Disruptive Behavior Disorder (DisGeNET)	3.944E^{-5}
	Social Phobia (DisGeNET)	3.944E^{-5}
	Schizophrenia (DisGeNET)	0.013

Table 7.13: Enriched terms associated with the up-regulated (activated) dopamine-related genes in the frontal cortex (BA9) for males (DBH, DRD5, DRD3, DRD4, SLC18A2).

Category	Description	Adjusted p-value
Pathways	Dopamine Receptors R-HSA-390651 (Reactome 2022)	2.024E^{-9}
	Dopaminergic synapse (KEGG 2021 Human)	9.919E^{-8}
Ontologies	Dopamine Metabolic Process (GO Biological Process 2023)	8.783E^{-7}
	Adenylate Cyclase-Activating Dopamine Receptor Signaling Pathway (GO Biological Process 2023)	1.512E^{-5}
	Postsynaptic Neurotransmitter Receptor Activity (GO Molecular Function 2023)	1.280E^{-4}
Diseases	Nicotine Dependence (DisGeNET)	2.913E^{-9}
	Depressive Disorder (DisGeNET)	3.394E^{-6}
	Schizophrenia (DisGeNET)	7.838E^{-5}

Overall, in males, both regions are heavily involved in dopaminergic signaling. The amygdala is more linked to dopamine transport and uptake, whereas the frontal cortex (BA9) shows a more pronounced role in dopamine metabolism and receptor activity. The amygdala has a low adjusted p-value for diseases involved in emotional regulation and psychiatric conditions, while the frontal cortex (BA9) has a low adjusted p-value for diseases related to cognitive functions. Notably, both regions exhibit a lower adjusted p-value for the disease schizophrenia.

The insights for females are present in Table 7.8 for the amygdala, and Table 7.14 for the frontal cortex (BA9). In females, the genes differentially expressed and up-regulated (activated) in the amygdala correspond to those differentially expressed and up-regulated (activated) in the entire sample set. This relation suggests no significant differences between the female and the overall sample sets.

Table 7.14: Enriched terms associated with the up-regulated (activated) dopamine-related genes in the frontal cortex (BA9) for females (DBH, DRD5).

Category	Description	Adjusted p-value
Pathways	Dopamine Receptors R-HSA-390651 (Reactome 2022)	0.008
	Dopaminergic synapse (KEGG 2021 Human)	0.030
Ontologies	Adenylate Cyclase-Activating Dopamine Receptor Signaling Pathway (GO Biological Process 2023)	0.002
	Dopamine Catabolic Process (GO Biological Process 2023)	0.002
	Postsynaptic Neurotransmitter Receptor Activity (GO Molecular Function 2023)	0.008
Diseases	Conduct Disorder (DisGeNET)	1.490E^{-4}
	Endogenous Depression (DisGeNET)	3.793E^{-4}
	Schizophrenia (DisGeNET)	0.017

In essence, both brain regions are involved in dopaminergic signaling. The amygdala is more involved in dopamine transport and uptake, emphasizing its role in emotional regulation and immediate emotional responses, while the frontal cortex (BA9) in females is more focused on dopamine metabolism and receptor activity, which is crucial for cognitive and mood regulation. Both regions share an association with schizophrenia, indicating a common link to this psychiatric condition.

In both sexes, the frontal cortex (BA9) exhibits significant up-regulation (activation) in pathways and processes related to dopamine receptor activity and synaptic transmission, with associated diseases emphasizing its role in cognitive functions and mood regulation. Specifically, in males, the frontal cortex (BA9) is linked to diseases such as nicotine dependence, and depressive disorder, while in females, it is associated with conduct disorder, and endogenous depression.

In the amygdala, both sexes show enrichment in pathways related to neurotransmitter transport and synaptic transmission. However, the specific pathways and associated diseases differ slightly. In males, the amygdala is associated with diseases such as disruptive behavior disorder, and social phobia,

reflecting its role in emotional regulation. In females, the amygdala is linked to social phobia, and impulsive character, indicating its involvement in emotional responses and psychiatric conditions.

Lastly, we compare a brain tissue, the frontal cortex (BA9), with a peripheral tissue, the whole blood, to determine which genes are differentially activated in each tissue. We present the results in Table 7.15.

Table 7.15: Scores and regulation conclusion for dopamine-related genes in the frontal cortex (BA9) versus the whole blood.

Gene	Score	Regulation
COMT	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DBH	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD2	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD4	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD5	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
LRTOMT	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
SLC18A2	$< 1\text{E}^{-10}$	Up-regulated (activated) on the whole blood
SLC6A2	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
TH	$< 1\text{E}^{-10}$	Up-regulated (activated) on the frontal cortex (BA9)
DRD3	3.580E^{-6}	Up-regulated (activated) on the frontal cortex (BA9)
TH	0.330	No conclusion

Table 7.15 illustrates that 10 out of the 11 genes from the dopaminergic system are differentially expressed between the two tissues (represented in green), with only the gene TH not showing differential expression (represented in red). The TH gene plays a central role in the physiology of adrenergic neurons, encoding a protein involved in the conversion of tyrosine to dopamine.

Of the 10 differentially expressed genes, 9 of them are up-regulated (activated) in the frontal cortex (BA9), and only one gene is up-regulated (activated) in the whole blood, the gene SLC18A2. This gene encodes a transmembrane protein, VMAT2, that functions as an ATP-dependent transporter of monoamines, such as dopamine. Polymorphisms in this gene are associated with schizophrenia and bipolar disorder. The up-regulation (activation) of SLC18A2 in the whole blood compared to the frontal cortex (BA9) indicates higher gene expression and likely higher protein levels of VMAT2 in blood cells compared to the neurons in the frontal cortex (BA9). The elevated expression of SLC18A2 in blood cells suggests that the SLC18A2 gene functions as a dopamine transporter, with its corresponding protein, VMAT2, being more abundantly produced in peripheral tissues (whole blood).

7.2 Pattern-Centric Results

The pattern-centric stances incorporated in the visualization tool assist in the identification of putative regulatory modules (patterns). They facilitate a better understanding of gene regulation, unraveling new biological insights and guiding the development of novel therapeutics. In this subsection, we focus on two different approaches for pattern identification: (i) biclustering, for two-dimensional data (gene-tissue); and (ii) triclustering, for three-dimensional data (individual-gene-tissue).

7.2.1 Biclustering

For the pattern analysis of brain patterning using biclustering stances, we utilize the BicPAMS algorithm [79]. We consider three distinct tissues: the cortex, the frontal cortex (BA9), and the whole blood. We aim to find both types of biclusters, constant and order-preserving, using a column normalization mapping technique. In Table 7.16, we present the constant biclusters retrieved from the BicPAMS algorithm using the proposed visualization tool.

Table 7.16: List of the constant biclusters retrieved from BicPAMS.

Clusters	Statistical significance (p-value)
Bicluster1, Bicluster2, Bicluster3, Bicluster4	$< 1\text{E}^{-30}$
Bicluster5	1.110E^{-15}
Bicluster6	0.828

As shown in Table 7.16, we found 6 constant biclusters under the introduced parameterization (different parameterizations could yield distinct results). All, except Bicluster6, exhibit a meaningful significance level, as no significance value exceeds 0.05. This observation leads us to conclude that the algorithm is prepared towards the task at hands. For demonstration purposes, we now examine one of the constant biclusters with the lowest significance, Bicluster1, described bellow:

```
Dimensions:  (62 genes, 3 tissues);
X = (ACVR1C, ADGRF3, CCDC62, CCKAR, CLEC4M, DBH, DDC, DRD2, DRD5, GABRA6, GABRB1,
GABRE, GABRG3, GABRQ, GALR1, GLP2R, GPR148, GPR149, GPR45, GPR52, GPR63, HTR1A, HTR1B,
KLHL1, MC5R, P2RY6, PCDHA1, PCDHA10, PCDHA11, PCDHA12, PCDHA13, PCDHA2, PCDHA3,
PCDHA5, PCDHA6, PCDHA8, PCDHA9, PCDHAC1, PCDHB11, PCDHB12, PCDHB13, PCDHB15, PCDHB16,
PCDHB2, PCDHB3, PCDHB7, PCDHB9, PCDHGA1, PCDHGA10, PCDHGA11, PCDHGA2, PCDHGA4,
PCDHGA5, PCDHGA6, PCDHGA7, PCDHGB1, PCDHGB3, PCDHGB4, PCDHGB5, PCDHGB6, RANBP17,
RGS17);
Y = (Cortex, Frontal Cortex (BA9), Whole Blood).
```

The bicluster shown contains 62 genes marked by X, and three tissues marked by Y. By examining the bicluster, we confirm that it is, indeed, a constant bicluster, as the values have low variation to each other. Therefore, we conclude that the mean sample expression in these three tissues is consistently similar across all genes. Moreover, the bicluster has an MSR score of 0.024, indicating that the model accurately captures the data patterns within the bicluster. We visualize the bicluster in Figure 7.5, where each line represents a different gene from the bicluster.

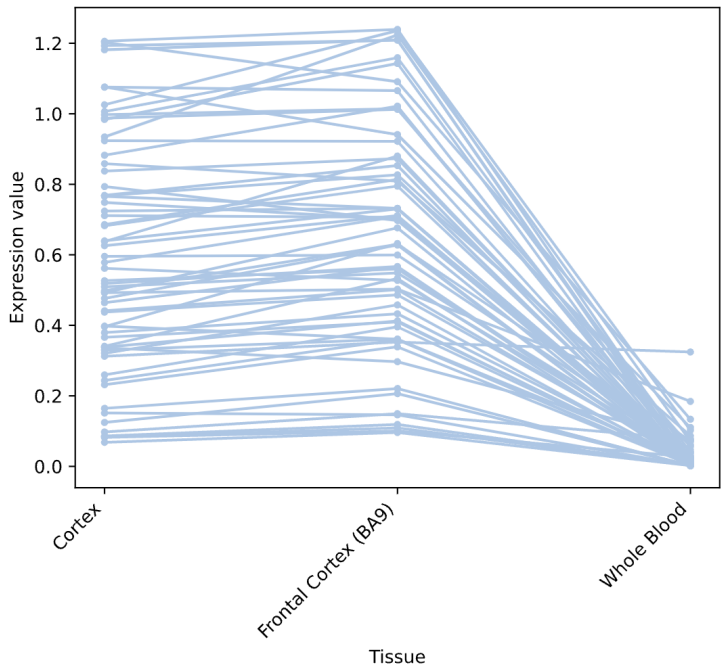


Figure 7.5: Expression values for each gene across the different tissues within the Bicluster1.

For this set of 62 genes, we performed a functional enrichment analysis to gain insights into pathway enrichment, ontologies, and disease associations. We present the main results retrieved from the Enrichr platform in Table 7.17.

Table 7.17: Enriched terms associated with the up-regulated (activated) genes from the Bicluster1.

Category	Description	Adjusted p-value
Pathways	Amine Ligand-Binding Receptors R-HSA-375280 (Reactome 2022)	8.619E^{-5}
	Neuroactive ligand-receptor interaction (KEGG 2021 Human)	5.232E^{-11}
Ontologies	Chemical Synaptic Transmission (GO Biological Process 2023)	2.883E^{-14}
	GABA-A Receptor Activity (GO Molecular Function 2023)	6.657E^{-8}
Diseases	Neutrophil count (procedure) (DisGeNET)	2.286E^{-11}
	Blood basophil count (lab test) (DisGeNET)	2.367E^{-8}

The functional enrichment analysis of the 62 genes set reveals significant associations with various biological processes and pathways. The genes are notably enriched in pathways related to amine ligand-binding receptors and neuroactive ligand-receptor interactions, highlighting their crucial role in neurotransmission and receptor signaling. This role is further supported by their strong involvement in chemical synaptic transmission and GABA-A receptor activity, indicating a significant contribution to synaptic communication and inhibition. These findings underscore the prominent roles of the genes in both neurobiological and immune functions, emphasizing their importance in neurotransmission and immune regulation.

Additionally, an order-preserving bicluster identified by the BicPAMS algorithm, with an MSR score of 0.099, is further highlighted in Figure 7.6, indicating a strong alignment between the biclustering solution and the observed data.

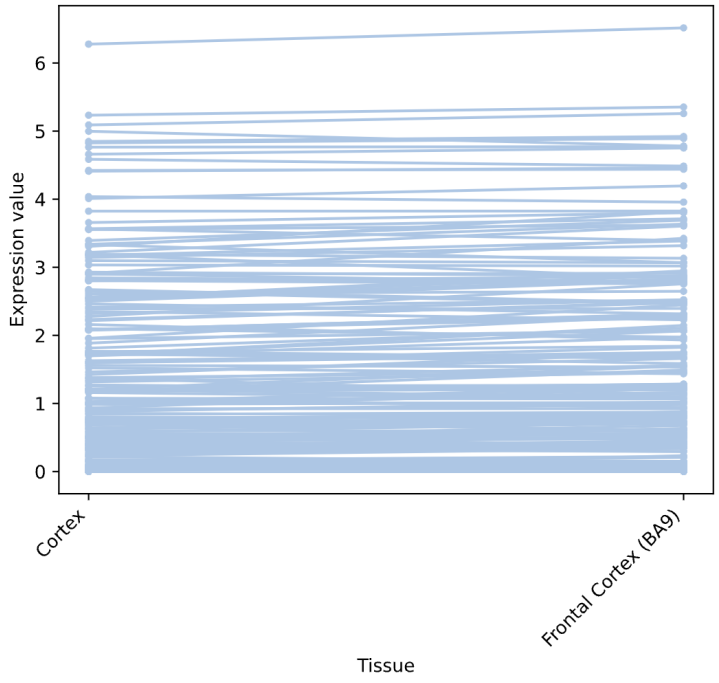


Figure 7.6: Expression values for each gene across the different tissues within the order-preserving bicluster.

7.2.2 Triclustering

According to the introduced principles in Section 5.4, the state-of-the-art TriCluster algorithm [45] is applied over the preprocessed individuals, genes, and tissues tensor. We start by considering all tissues as input. Table 7.18 provides an illustrative set of triclusters discovered under the following parameterization: (i) a minimum of two individuals at the row level; (ii) a minimum of 10 genes at the column level; and (iii) a minimum of 9 tissues at the context level.

Table 7.18: List of the triclusters retrieved from TriCluster (first parameterization).

Clusters	Statistical significance (p-value)
Tricluster1	2.914E^{-222}
Tricluster2	6.985E^{-210}
Tricluster3	6.985E^{-210}
Tricluster4	5.563E^{-183}

As shown in Table 7.18, the triclustering algorithm identified four distinct triclusters, all with very low p-values. For exemplification purposes, we will focus on a single tricluster, Tricluster4, described below:

```
Dimensions:  (2 individuals, 10 genes, 9 tissues);
X = (GTEX-1GN73, GTEX-1B8L1);
Y = (ARRB2, CPE, DBI, GPR162, GPRASP2, L1CAM, NCOA5, SLC7A14, SYNGR3, THY1);
Z = (Anterior cingulate cortex (BA24), Caudate (basal ganglia), Cerebellar Hemisphere,
Cerebellum, Cortex, Frontal Cortex (BA9), Hypothalamus, Nucleus accumbens (basal
ganglia), Putamen (basal ganglia)).
```

The retrieved tricluster contains two individuals, 10 genes, and 9 tissues, with an 3DMSR score of 7.796E^{-16}. We visualize the two individual biclusters in Figures 7.7 and 7.8.

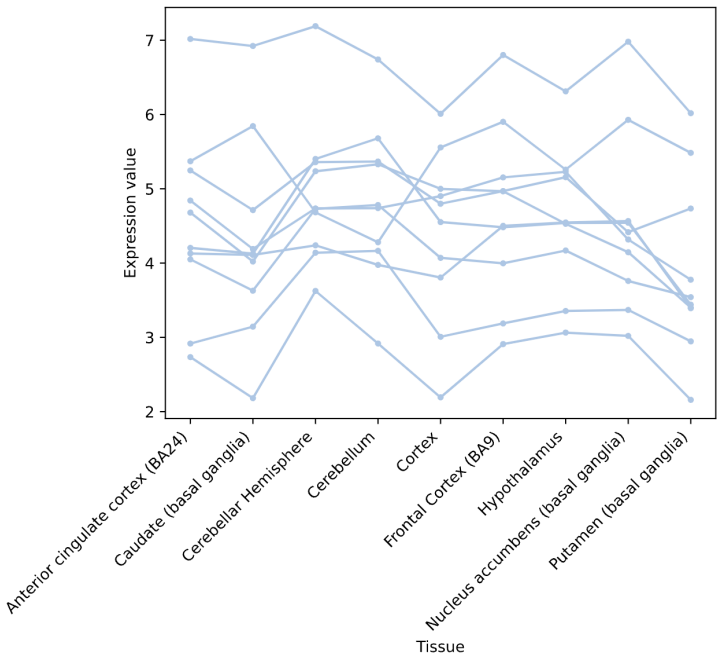


Figure 7.7: Expression values for each gene across the different tissues within the tricluster for the individual GTEX-1GN73.

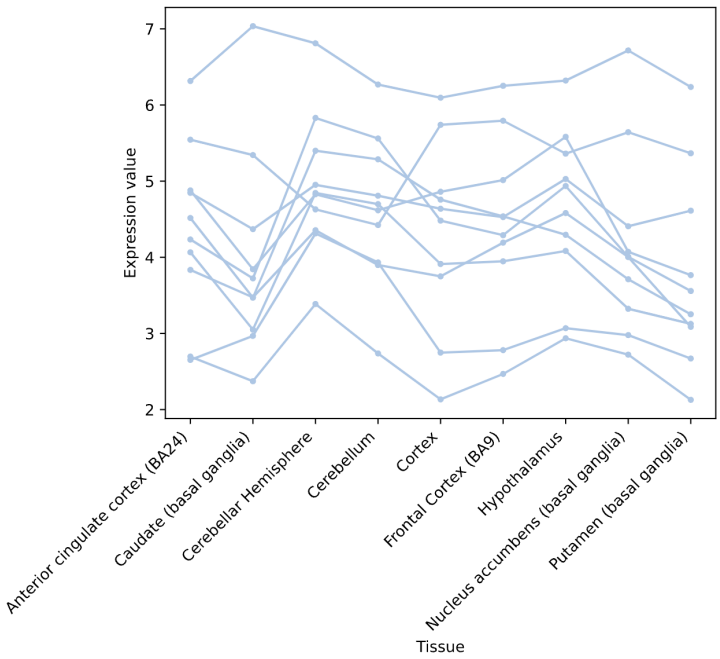


Figure 7.8: Expression values for each gene across the different tissues within the tricluster for the individual GTEX-1B8L1.

Regarding the two individuals present in the tricluster, marked by X, GTEX-1GN73 and GTEX-1B8L1, both correspond to distinct male individuals with ages in the range 60-69, and agonal state of fast death from natural causes.

In this context, we calculate the lift to assess the strength of the association between the tricluster and a covariate of interest. This measure provides insight into how specific covariate combinations influence the likelihood of observing the tricluster. Assuming that φ_B denotes the discriminative pattern of expression, i.e., the tricluster, and c represents the covariate of interest, i.e., males aged 60-69 with an agonal state of fast death from natural causes, we have the following calculation:

$$lift(\varphi_B \to c) = \frac{P(\varphi_B, c)}{P(\varphi_B)P(c)} = \frac{\frac{2}{931}}{\frac{2}{931} \times \frac{84}{931}} = 11.083. \tag{7.2}$$

This result indicates that the occurrence of the tricluster φ_B is 11 times more likely to be associated with the covariate c than would be expected by random chance.

Pertaining to the genes found in the tricluster, marked by Y, we performed a functional enrichment analysis on the set of 10 genes, revealing associations with several key biological pathways and processes. Notably, these genes are associated with dopaminergic synapses, which implies a potential role in synaptic function. This set of genes also show connections with pathways related to cell-matrix adhesion and G protein-coupled receptor binding, highlighting their relevance to cellular interactions and receptor activity.

Moreover, we consider a second parameterization, which includes the following tissues: amygdala, cortex, frontal cortex (BA9), hippocampus, hypothalamus, pituitary, whole blood, adipose (visceral), skin (not sun exposed), and skin (sun exposed). Table 7.19 presents the top three triclusters from the 29 identified under this parameterization, based on statistical significance. The parameterization criteria are: (i) a minimum of 10 individuals at the row level; (ii) a minimum of four genes at the column level; and (iii) a minimum of four tissues at the context level. In this case, we calculate the 3DMSR for each identified tricluster and assess the lift with respect to two covariates of interest: $s1$, representing males, and $s2$, representing females.

Table 7.19: List of the triclusters retrieved from TriCluster (second parameterization).

Clusters	Statistical significance (p-value)	3DMSR	$lift(\varphi_B \rightarrow s1)$	$lift(\varphi_B \rightarrow s2)$
Tricluster1	4.727E^{-290}	5.181E^{-16}	1.192	0.608
Tricluster2	1.466E^{-246}	1.599E^{-16}	1.043	0.913
Tricluster3	1.097E^{-239}	7.905E^{-16}	1.043	0.913

As shown in Table 7.19, the three triclusters exhibit relatively low p-values, indicating statistical significance. The 3DMSR values for all triclusters demonstrate a good fit to the data, reflecting their high quality. Additionally, lift values are considered interesting if they exceed one, as this suggests a substantial association between the tricluster and the covariates $s1$ and $s2$. Notably, the three triclusters described are more likely to be associated with the covariate $s1$ than would be expected by random chance.

In essence, all the putative regulatory modules (patterns) retrieved and analyzed were evaluated based on four distinct criteria: (i) statistical significance, determined by calculating the respective p-values for each bicluster/tricluster found; (ii) homogeneity, assessed by calculating the MSR for biclusters and the 3DMSR score for triclusters; (iii) discriminative power, evaluated by calculating the lift to measure the quality of the solution; and (iv) biological relevance, ensured through functional enrichment analysis to confirm the value of the patterns retrieved for knowledge acquisition.

8

Conclusion

Contents

8.1 Concluding Remarks 70

8.2 Software Availability 70

8.3 Future Work 70

This chapter closes the work with brief concluding remarks in Section 8.1, software availability in Section 8.2, and suggestions for future work in Section 8.3.

8.1 Concluding Remarks

This work proposes the creation of a web-based visualization tool for learning neurotranscriptomic maps. These maps allow the exploration of the spatial distribution of neurotransmission markers in the brain using a statistical framework grounded on pairwise and multiwise differential expression views. In this context, users can contrast brain-to-brain and brain-to-peripheral tissues in the presence of various co-factors of interest, such as psycho-demographic profiles, i.e., sex and age, and agonal states, i.e., circumstances of death.

The proposed solution can allow a focus towards specific families of protein-coding genes, including genes that contribute to well-studied neurological processes, particularly those associated with neurotransmitters and their corresponding receptors. Moreover, and using data from the GTEx initiative, the proposed solution finds putative regulatory modules (patterns) using both biclustering algorithms to find co-expressed genes across tissues, and triclustering algorithms to find groups of individuals with coherent multi-tissue gene expression.

The results show the possibility for acquiring novel knowledge through the use of the proposed web-based visualization tool. By selecting a set of genes, tissues, and conditions, we conducted functional enrichment analyses that confirm the critical role of the genes highlighted by the tool in target neurotransmission systems, identifying their associations with specific biological pathways and diseases. Furthermore, the role of biclustering and triclustering in finding relevant and discriminative putative regulatory modules (patterns) of brain activity is further highlighted through a joint analysis of statistical significance, homogeneity, discriminative power, and functional enrichment.

8.2 Software Availability

The proposed web-based application can be accessed through the following link:

- `https://github.com/miguelmatamoreira/sdnm`.

8.3 Future Work

The following directions of future work are suggested:

- Augment the neurotranscriptomic maps with a second organism of study, rodents, to combine the neurobiological knowledge acquired through post-mortem cohort studies with critical animal

models. Similarly to the undertaken analysis, this direction should encompass the examination of gene expression across different rodent tissues in the application both gene-centric and pattern-centric stances;

- Apply the same methodology to the remaining neurotransmitter systems for further analysis and knowledge acquisition, beyond the already implemented dopaminergic and GABAergic systems;
- Study of specific clinical conditions by incorporating individual clinical annotations, similar to the co-factors of interest considered in this thesis, and evaluating the discriminative power of the obtained genes and patterns towards specific clinical outcomes;
- Include a second triclustering algorithm, namely, the TriGen algorithm [83], to establish further comparisons with the existing state-of-the-art TriCluster algorithm [45] in the deployed application.

Bibliography

[1] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: A survey,” *IEEE Transactions on Computational Biology and Bioinformatics, Vol. 1, No. 1*, 2004.

[2] H. Y. Chang, “Reversing the central dogma: Rna-guided control of dna in epigenetics and genome editing,” *Review, Volume 83, Issue 3, P442-451*, 2023.

[3] G. Portal. Tissue sampling sites. (accessed in: 22 May 2024). [Online]. Available: https://gtexportal.org/home/samplingSitePage

[4] C. D. Luca, A. M. Colangelo, A. Virtuoso, L. Alberghina, and M. Papa, “Neurons, glia, extracellular matrix and neurovascular unit: A systems biology approach to the complexity of synaptic plasticity in health and disease,” *Int J Mol Sci. 2020 Feb; 21(4): 1539*, 2020.

[5] Z. Sherrell. How do glial cells function in the body. (accessed in: 22 May 2024). [Online]. Available: https://www.medicalnewstoday.com/articles/glial-cells-function

[6] C. Clinic. Neurotransmitters. (accessed in: 22 May 2024). [Online]. Available: https://my.clevelandclinic.org/health/articles/22513-neurotransmitters

[7] K. Cherry. What are neurotransmitters. (accessed in: 22 May 2024). [Online]. Available: https://www.verywellmind.com/what-is-a-neurotransmitter-2795394

[8] C. Clinic. Melatonin. (accessed in: 22 May 2024). [Online]. Available: https://my.clevelandclinic.org/health/articles/23411-melatonin

[9] M. Gallagher and A. A. Chiba, “The amygdala and emotion,” *Curr Opin Neurobiol. 1996 Apr;6(2):221-7*, 1996.

[10] E. T. Rolls, “The cingulate cortex and limbic systems for emotion, action, and memory,” *Brain Struct Funct. 2019 Dec;224(9):3001-3018*, 2019.

[11] M. T. Herrero, C. Barcia, and J. M. Navarro, “Functional anatomy of thalamus and basal ganglia,” *Childs Nerv Syst. 2002 Aug;18(8):386-404*, 2002.

[12] H. Mitoma, A. Buffo, F. Gelfo, X. Guell, E. Fucà, S. Kakei, J. Lee, M. Manto, L. Petrosini, A. G. Shaikh, and J. D. Schmahmann, "Cerebellar reserve: From cerebellar physiology to cerebellar disorders," *Cerebellum. 2020 Feb;19(1):131-153*, 2020.

[13] J. F. Marsden, "Cerebellar ataxia," *Handb Clin Neurol. 2018;159:261-281*, 2018.

[14] N. P. Friedman and R. T. W., "The role of prefrontal cortex in cognitive control and executive function," *Neuropsychopharmacology. 2022 Jan;47(1):72-89*, 2022.

[15] D. T. Jones and J. Graff-Radford, "Executive dysfunction and the prefrontal cortex," *Continuum (Minneap Minn). 2021 Dec 1;27(6):1586-1601*, 2021.

[16] J. Lisman, G. Buzsáki, H. Eichenbaum, L. Nadel, C. Ranganath, and A. D. Redish, "Viewpoints: how the hippocampus contributes to memory, navigation and cognition," *Nat Neurosci. 2017 Oct 26;20(11):1434-1447*, 2017.

[17] H. Fong, J. Zheng, and D. Kurrasch, "The structural and functional complexity of the integrative hypothalamus," *Science. 2023 Oct 27;382(6669):388-394*, 2023.

[18] H. N. Harris and Y. B. Peng, "Evidence and explanation for the involvement of the nucleus accumbens in pain processing," *Neural Regen Res. 2020 Apr; 15(4): 597–605*, 2020.

[19] J. L. Lanciego, N. Luquin, and J. A. Obeso, "Functional neuroanatomy of the basal ganglia," *Cold Spring Harb Perspect Med. 2012 Dec; 2(12): a009621*, 2012.

[20] M. Harrow-Mortelliti, V. Reddy, and G. Jimsheleishvili. Physiology, spinal cord. (accessed in: 22 May 2024). [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK544267/

[21] J. Sonne, V. Reddy, and M. R. Beato. Neuroanatomy, substantia nigra. (accessed in: 22 May 2024). [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK536995

[22] B. N. Institute. About the pituitary gland. (accessed in: 23 May 2024). [Online]. Available: https://www.barrowneuro.org/resource/about-the-pituitary-gland/#:~:text=Function%20of%20the%20Pituitary%20Gland,%2C%20adrenal%20glands%2C%20and%20gonads)

[23] C. Ortiz, M. Carlén, and K. Meletis, "Spatial transcriptomics: Molecular maps of the mammalian brain," *Annual Review of Neuroscience*, pp. 547–562, 2021.

[24] A. Mahfouz, S. M. H. Huisman, B. P. F. Lelieveldt, and M. J. T. Reinders, "Brain transcriptome atlases: a computational perspective," *Brain Struct Funct. 2017; 222(4): 1557–1580*, 2017.

[25] S. M. Sunkin, L. Ng, and C. e. a. Lau, "Allen brain atlas: an integrated spatio-temporal portal for exploring the central nervous system," *Nucleic Acids Res. 2013 Jan; 41(Database issue): D996–D1008*, 2012.

[26] A. Arnatkeviciute, R. D. Markello, B. D. Fulcher, B. Misic, and A. Fornito, “Toward best practices for imaging transcriptomics of the human brain,” *Biological Psychiatry*, pp. 391–404, 2022.

[27] T. L. Gilbert, “The allen brain atlas as a resource for teaching undergraduate neuroscience,” *J Undergrad Neurosci Educ. 2018 Summer; 16(3): A261–A267*, 2018.

[28] C. L. Rees, C. M. White, and G. A. Ascolia, “Neurochemical markers in the mammalian brain: Structure, roles in synaptic communication, and pharmacological relevance,” *Curr Med Chem. 2017; 24(28): 3077–3103*, 2017.

[29] T. G. CONSORTIUM, “The gtex consortium atlas of genetic regulatory effects across human tissues,” *SCIENCE, Vol 369, Issue 6509, pp. 1318-1330*, 2020.

[30] L. Iversen, “Neurotransmitter transporters and their impact on the development of psychopharmacology,” *Br J Pharmacol. 2006 Jan; 147(Suppl 1): S82–S88*, 2006.

[31] A. P. King and R. J. Eckersley, *Statistics for Biomedical Engineers and Scientists, Chapter 7, pp. 147-171*. Academic Press, 2019.

[32] K. M. Ramachandran and C. P. Tsokos, *Mathematical Statistics with Applications in R (Third Edition), Chapter 11, pp. 461-490*. Academic Press, 2020.

[33] M. C. Morais, *Probabilidades e Estatística: Teoria, Exemplos & Exercícios*. IST Press, pp. 325-423, 2020.

[34] M. D. Riina, C. Stambaugh, N. Stambaugh, and K. E. Huber, *Translational Radiation Oncology, Chapter 28, pp. 153-163*. Academic Press, 2023.

[35] P. Mishra, U. Singh, C. M. Pandey, P. Mishra, and G. Pandey, “Application of student’s t-test, analysis of variance, and covariance,” *PubMed Central*, p. 407–411, 2019.

[36] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval, the concepts and technology behind search - 2nd edition*. Addison-Wesley Professional, 2011.

[37] A. E. Ezugwu and A. M. e. a. Ikotun, “A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects,” *Engineering Applications of Artificial Intelligence, Volume 110, 104743*, 2022.

[38] S. Arora and B. Barak, *Computational complexity: a modern approach*. Cambridge University Press, 2009.

[39] S. Cook, “The p versus np problem,” *Clay Mathematics Institute*, 2000.

[40] J. A. Hartigan, “Direct clustering of a data matrix,” *J. Am. Statistical Assoc. (JASA), vol. 67, no. 337, pp. 123-129*, 1972.

[41] G. Getz, E. Levine, and E. Domany, “Coupled two-way clustering analysis of gene microarray data,” *Proc. Natural Academy of Sciences US, pp. 12079-12084*, 2000.

[42] A. Califano, G. Stolovitzky, and Y. Tu, “Analysis of gene expression microarays for phenotype classification,” *Proc. Int’l Conf. Computacional Molecular Biology, pp. 75-85*, 2000.

[43] Q. Sheng, Y. Moreau, and B. De Moor, “Biclustering microarray data by gibbs sampling,” *Bioinformatics, vol. 19, pp. ii196-ii205*, 2003.

[44] C. Tang, L. Zhang, I. Zhang, and M. Ramanathan, “Interrelated two-way clustering: An unsupervised approach for gene expression data analysis,” *Proc. Second IEEE Int’l Symp. Bioinformatics and Bioeng., pp. 41-48*, 2001.

[45] L. Zhao and M. J. Zaki, “Tricluster: an effective algorithm for mining coherent clusters in 3d microarray data,” *In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, SIGMOD ’05, 694–705*, 2005.

[46] R. Henriques and S. C. Madeira, “Triclustering algorithms for three-dimensional data analysis: A comprehensive survey,” *ACM Computing Surveys, Vol. 51, No. 5, Article 95*, 2018.

[47] D. Bozdağ, A. S. Kumar, and U. V. Catalyurek, “Comparative analysis of biclustering algorithms,” *BCB. ACM, New York, 265–274*, 2010.

[48] A. Patrikainen and M. Meila, “Comparing subspace clusterings,” *IEEE TKDE 18, 7 (2006), 902–916*, 2006.

[49] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Henning, L. Thiele, and E. Zitzler, “A systematic comparison and evaluation of biclustering methods for gene expression data,” *Bioinformatics 22, 9 (2006), 1122–1129*, 2006.

[50] Y. Cheng and G. M. Church, “Biclustering of expression data,” *International Conference on Intelligent Systems for Molecular Biology 8:93-103*, 2000.

[51] R. Henriques and S. C. Madeira, “Bsig: evaluating the statistical significance of biclustering solutions,” *Data Min. Knowl. Discov., 124-161*, 2018.

[52] ——, “Flebic: Learning classifiers from high-dimensional biomedical data using discriminative biclusters with non-constant patterns,” *Pattern Recognit., 115 (2021), p. 107900*, 2021.

[53] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of the Cell, The Structure and Function of DNA*. Garland Science, 2002.

[54] ATDBio. Dna, rna and protein synthesis. (accessed in: 30 Oct 2023). [Online]. Available: https://atdbio.com/nucleic-acids-book/Transcription-Translation-and-Replication

[55] Nature. Proteome. (accessed in: 19 Dec 2023). [Online]. Available: https://www.nature.com/scitable/definition/proteome-297/

[56] R. I. Teleanu, A.-G. Niculescu, E. Roza, O. Vladâcenco, A. M. Grumezescu, and D. M. Teleanu, "Neurotransmitters - key factors in neurological and neurodegenerative disorders of the central nervous system," *Int J Mol Sci. 2022 Jun; 23(11): 5954*, 2022.

[57] S. Li, A. H. C. Wong, and F. Liu, "Ligand-gated ion channel interacting proteins and their role in neuroprotection," *Front Cell Neurosci. 2014 May 9:8:125*, 2014.

[58] D. E. Haines and G. A. Mihailoff, *Fundamental Neuroscience for Basic and Clinical Applications, pp. 54-71*. Elsevier, 2018.

[59] N. I. of Neurological Disorders and Stroke. Brain basics: Know your brain. (accessed in: 27 Oct 2023). [Online]. Available: https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-know-your-brain

[60] T. C. Südhof, "Towards an understanding of synapse formation," *CellPress, Neuron 100*, pp. 276–293, 2018.

[61] T. E. o. E. Britannica. Synapse. (accessed in: 30 Oct 2023). [Online]. Available: https://www.britannica.com/science/synapse

[62] A. Arnatkeviciute, B. D. Fulcher, M. A. Bellgrove, and A. Fornito, "Imaging transcriptomics of brain disorders," *Biological Psychiatry: GOS*, pp. 319–331, 2021.

[63] B. D. Fulcher, A. Arnatkeviciute, and A. Fornito, "Overcoming false-positive gene-category enrichment in the analysis of spatially resolved transcriptomic brain atlas data," *Nature Communications volume 12, Article number: 2669*, 2021.

[64] F. Hormozdiari, G. Kichaev, W.-Y. Yang, B. Pasaniuc, and E. Eskin, "Identification of causal genes for complex traits," *Bioinformatics. 2015 Jun 15; 31(12): i206–i213*, 2015.

[65] D. J. Schaid, W. Chen, and N. B. Larson, "From genome-wide associations to candidate causal variants by statistical fine-mapping," *Nature Reviews Genetics volume 19, pp. 491–504*, 2018.

[66] H. J. Kang, Y. I. Kawasawa, and F. e. a. Cheng, “Spatio-temporal transcriptome of the human brain,” *Nature, 26 Oct 2011, 478(7370):483-489*, 2011.

[67] N. Chaudhari and W. E. Hahn, “Genetic expression in the developing brain,” *Science, pp. 924-928*, 1983.

[68] M. Somel, S. Guo, and Z. e. a. Yan, “Microrna, mrna, and protein expression link development and aging in human and macaque brain,” *Genome Res. 2010 Sep; 20(9): 1207–1218*, 2010.

[69] U. C. Eze, A. Bhaduri, T. J. N. Haeussler, and A. R. Kriegstein, “Single-cell atlas of early human brain development highlights heterogeneity of human neuroepithelial cells and early radial glia,” *Nature Neuroscience volume 24, pp. 584–594*, 2021.

[70] E. J. Armand, J. Li, F. Xie, C. Luo, and E. A. Mukamel, “Single-cell sequencing of brain cell transcriptomes and epigenomes,” *Neuron. 2021 Jan 6; 109(1): 11–26*, 2021.

[71] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, “Spatial reconstruction of single-cell gene expression data,” *Nature Biotechnology volume 33, pp. 495–502*, 2015.

[72] F. A. Wolf, P. Angerer, and F. J. Theis, “Scanpy: large-scale single-cell gene expression data analysis,” *Genome Biology volume 19, Article number: 15*, 2018.

[73] L. Haghverdi, A. T. L. Lun, M. D. Morgan, and J. C. Marioni, “Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors,” *Nat Biotechnol. 2018 Jun;36(5):421-427*, 2018.

[74] K. Polański, M. D. Young, Z. Miao, K. B. Meyer, S. A. Teichmann, and J.-E. Park, “Bbknn: fast batch alignment of single cell transcriptomes,” *Bioinformatics. 2020 Feb 1;36(3):964-965*, 2020.

[75] G. P. Wagner, K. Kin, and V. J. Lynch, “Measurement of mrna abundance using rna-seq data: Rpkm measure is inconsistent among samples,” *Theory Biosci. 2012 Dec;131(4):281-5*, 2012.

[76] A. T. L. Lun, K. Bach, and J. C. Marioni, “Pooling across cells to normalize single-cell rna sequencing data with many zero counts,” *Genome Biology volume 17, Article number: 75*, 2016.

[77] R. Bacher, L.-F. Chu, and N. e. a. Leng, “Scnorm: robust normalization of single-cell rna-seq data,” *Nat. Methods*, 2017.

[78] E. Sjöstedt, L. Fagerberg, B. M. Hallström, A. Häggmark, N. Mitsios, P. Nilsson, F. Pontén, T. Hökfelt, M. Uhlén, and J. Mulder, “Defining the human brain proteome using transcriptomics and antibody-based profiling with a focus on the cerebral cortex,” *PLOS ONE*, pp. 1–20, 2015.

[79] R. Henriques, F. L. Ferreira, and S. C. Madeira, “Bicpams: software for biological data analysis with pattern-based biclustering,” *BMC Bioinformatics (2017) 18:82*, 2017.

[80] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, “Spectral biclustering of microarray data: coclustering genes and conditions,” *Genome research, 13(4):703–716*, 2003.

[81] G. Li, Q. Ma, H. Tang, A. H. Paterson, and Y. Xu, “Qubic: a qualitative biclustering algorithm for analyses of gene expression data,” *Nucleic Acids Research, Volume 37, Issue 15, Page e101*, 2009.

[82] J. Xie, A. Ma, Y. Zhang, B. Liu, S. Cao, C. Wang, J. Xu, C. Zhang, and Q. Ma, “Qubic2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale rna-seq data,” *Bioinformatics, Volume 36, Issue 4, Pages 1143–1149*, 2020.

[83] D. Gutiérrez-Avilés, C. Rubio-Escudero, F. Martínez-Álvarez, and J. C. Riquelme, “Trigen: A genetic algorithm to mine triclusters in temporal gene expression data,” *Neurocomputing 132*, 2014.

[84] Y. Yin, Y. Zhao, B. Zhang, and G. Wang, “Mining time-shifting co-regulation patterns from gene expression data,” *Lecture Notes in Computer Science, vol. 4505, 2007, pp. 62–73*, 2007.

[85] X. Xu, Y. Lu, K. L. Tan, and A. Tung, “Finding time-lagged 3d clusters,” *IEEE International Conference on Data Engineering, 2009, pp. 445–456*, 2008.

[86] K. Sim and Z. Aung, “Discovering correlated subspace clusters in 3d continuous-valued data,” *IEEE International Conference on Data Mining, 2010, pp. 471–480*, 2010.

[87] D. F. Soares, R. Henriques, M. Gromicho, M. de Carvalho, and S. C. Madeira, “Triclustering-based classification of longitudinal data for prognostic prediction: targeting relevant clinical endpoints in amyotrophic lateral sclerosis,” *Scientific Reports*, 2023.

[88] T. U. Consortium, “Uniprot: the universal protein knowledgebase in 2023,” *Nucleic Acids Research, Volume 51, Issue D1, Pages D523–D531*, 2023.

[89] D. Szklarczyk, R. Kirsch, M. Koutrouli, K. Nastou, F. Mehryary, R. Hachilif, A. L. Gable, T. Fang, N. T. Doncheva, S. Pyysalo, P. Bork, L. J. Jensen, and C. von Mering, “The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest,” *Nucleic Acids Research, Volume 51, Issue D1, Pages D638–D646*, 2023.

[90] I. S. Dhillon, “Co-clustering documents and words using bipartite spectral graph partitioning,” *KDD ’01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, Pages 269–274*, 2001.

[91] T. O. Monroe, M. E. Garrett, and M. e. a. Kousi, “Pcm1 is necessary for focal ciliary integrity and is a candidate for severe schizophrenia,” *Nat Commun 11, 5903 (2020)*, 2020.

[92] M.-H. Kim, C. Radaelli, and E. R. e. a. Thomsen, “Target cell-specific synaptic dynamics of excitatory to inhibitory neuron connections in supragranular layers of human neocortex,” *eLife. 2023; 12: e81863*, 2023.

[93] E. Clough and T. Barrett, “The gene expression omnibus database,” *Methods Mol Biol. 2016; 1418: 93–110.*, 2016.

A

Additional Code

Below we present the code used to construct the results matrix present in Subsection 7.1.1. This matrix shows how many genes from the GTEx initiative exhibit differential expression between two different groups of tissues.

Listing A.1: PYTHON Code

```python
# define the tissues
brain_tissues = ["amygdala", "frontal_cortex", "hippocampus", "hypothalamus"]
peripheral_tissues = ["blood", "adipose", "skin"]

# read all the datasets
dfs = {}
for tissue in brain_tissues + peripheral_tissues:
    file_path = f"gene_tpm_{tissue}.gct"
    dfs[tissue] = pd.read_csv(file_path, sep='\t', skiprows=2, index_col=0)
    dfs[tissue].iloc[:, 2:] = np.log1p(dfs[tissue].iloc[:, 2:])

# define the genes
genes = dfs["amygdala"]["Description"].tolist()
```

```
# construct the results matrix
results = np.zeros((len(brain_tissues), len(peripheral_tissues)))
for gene in genes:
    for bt_i, bt in enumerate(brain_tissues):
        s1 = dfs[bt].loc[dfs[bt][dfs[bt]["Description"] == gene].index.tolist
            ()[0]].tolist()[2:]
        for pt_i, pt in enumerate(peripheral_tissues):
            s2 = dfs[pt].loc[dfs[pt][dfs[pt]["Description"] == gene].index.
                tolist()[0]].tolist()[2:]
            _, p = ttest_ind(s1, s2)
            if p < 0.01:
                results[bt_i, pt_i] += 1
```

B

Rodents Case Study

Studies have found that the protein encoded by the Pericentriolar Material 1 (PCM1) gene is a relevant cellular structure involved in cell division and ciliary function [91], being essential for maintaining the integrity of the primary cilium. Some individuals with severe forms of schizophrenia have variants in the PCM1 gene, suggesting a potential link between the two, making the case that disruptions in ciliary function contribute to the making of the disease [91].

Analyzing rodents is an advantageous approach for several reasons. Firstly, rodents provide a consistent and controlled environment, making it easier to reproduce similar conditions across different individuals. Additionally, the structural and functional similarities between rodent and human brains are particularly noteworthy. For instance, Kim et al. [92] demonstrated that inhibitory circuit motifs remain conserved across rodents and humans. This conservation underscores the potential of rodents as a model system for expanding our understanding of neurotranscriptomic maps.

For illustrative purposes in future studies, we employed the Gene Expression Omnibus (GEO) dataset GSE145915[1] [93] to retrieve information about RNA sequencing of different brain tissues in rodents, gaining insights into the gene expression changes associated with the absence of the Pericentriolar Material 1 (PCM1) protein. The consolidated dataset is the union of 18 distinct transcriptomic assays that correspond to three different samples from three distinct brain areas (striatum, frontal cortex, and

[1] GSE145915 dataset website.

hippocampus), and under two different possible conditions (PCM1 knockout and wild-type).

In this context, PCM1 knockout (PCM1-/-) mice correspond to mice that have been genetically modified to have a knockout (KO) or a deletion (-/-) of the PCM1 gene. Conversely, wild-type (WT) littermates are mice from the same breeding with an intact PCM1 gene that functions as a control group. Information about this genetic modification is relevant to access insights about mental dysfunctions associated with schizophrenia.

Table B.1: Example row of the GEO dataset for the first wild-type sample.

Gene	Striatum	Frontal Cortex	Hippocampus
ENSMUSG00000000001	1136	967	1019

As we can observe in Table B.1, the dataset has four columns: one indicating the name of the gene, and the other three representing the expression of that same gene in each brain region. The entries of the table represent the expression value of a given gene (row) in a specific tissue (column).

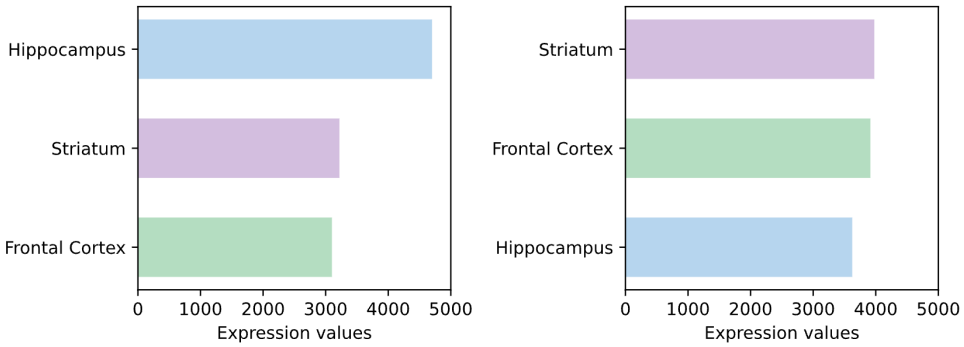


Figure B.1: Mean wild-type (left) and knockout (right) sample expression of GABBR2 gene across tissues.

Figure B.1 shows the expression of the GABBR2 gene across the 18 samples, grouping each sample by type (wild-type/knockout) and by tissue. In the samples with wild-type, the region exhibiting the highest mean expression value of the GABBR2 gene is the hippocampus, followed by the striatum and the frontal cortex. In the knockout samples, the three regions have closer mean expression values for the GABBR2 gene, but the striatum and the frontal cortex still present higher expression values.