



Universidade de Brasília
Departamento de Estatística

Agrupamento por Misturas Finitas de Normais com Aplicação a Dados de Expressão Gênica

Louise Barbosa dos Santos

Prof. George von Borries

Brasília
2021

Louise Barbosa dos Santos

**Agrupamento por Misturas Finitas de Normais com Aplicação a Dados de
Expressão Gênica**

Orientador:

Prof. George von Borries

Colaboradora:

Prof(a). Joanlise Marco de Leon Andrade

Relatório apresentado para o Departamento de Estatística, Instituto de Ciências Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

**Brasília
2021**

Agradecimentos

A Deus por até aqui ter me sustentado, me capacitado e me dado forças para enfrentar todos os desafios.

Aos meus pais, Rildo e Deyse Santos, que são meus exemplos e meu irmão, Gabriel Barbosa dos Santos, que é meu melhor amigo por serem sempre presentes e apoiarem todos os meus sonhos. Por terem vivido comigo cada nova fase da minha vida e me mostrado que, com persistência, podemos alcançar nossos sonhos.

Ao meu noivo, Matheus Soares Machado, por estar junto comigo na conclusão de mais uma etapa. Por sempre apoiar meus estudos e meu crescimento pessoal, acadêmico e profissional. Por sempre me incentivar a seguir em frente e sempre me trazer tranquilidade para todos os momentos.

Ao meu padrinho, Alfredo Moreira Salgado, e ao meu professor do Ensino Médio, Eduardo Marques, por terem me apresentado à área de Estatística e acreditado no meu potencial para esta profissão.

Aos professores do Departamento de Estatística da UnB por estarem sempre dispostos a tirar dúvidas e a nos ensinar sobre vida acadêmica e profissional (além das matérias). Em especial, à professora Juliana Betini Fachini por ter feito eu me apaixonar pela Estatística desde o primeiro semestre por meio de seu exemplo, ao professor Luís Gustavo do Amaral Vinha por ter me orientado no PIBIC, à professora Maria Teresa Leão por incentivar os estudos de SAS e por estar sempre disposta a me ajudar e orientar.

Ao professor e orientador George Freitas von Borries por todas as quatro disciplinas ministradas que pude aprender mais com sua experiência. Por não se limitar a passar o conteúdo, mas por ampliar meu olhar crítico aos problemas, pesquisas e vida profissional e estudantil.

Aos amigos que a UnB me deu, que sempre me apoiaram e sempre acreditaram que eu poderia chegar até aqui. Em especial, ao Filipe Oliveira do Vale Ribeiro, Richard Wallan Paulino de Sousa, à Julia Verly Klin e ao Rodrigo Dantas Berçott por terem enfrentado comigo os semestres de curso mais desafiadores.

Àqueles que tive o prazer de trabalhar junto na ESTAT por terem ajudado a melhorar meus relacionamentos interpessoais e me ensinarem a como liderar um grupo de trabalho. À Karin Kawano Matuda e à Juliana Paula Degani por terem sido grandes amigas me ajudando a equilibrar os estudos e o trabalho, sempre me trazendo alegria e me motivando.

Ao Rafael Lins que esteve auxiliando durante o período de estruturação deste trabalho. Por ter sanado dúvidas que surgiam sobre o tema e compartilhado seus conhecimentos e experiências para aprimorar as análises aqui descritas.

Por fim, a todos aqueles que participaram da minha jornada e me fizeram crescer com seus ensinamentos.

Resumo

Introdução: As técnicas de agrupamento e classificação de dados foram sendo refinadas e surgiram algoritmos que agrupam objetos por meio de características comuns de acordo com cada critério. Diversos algoritmos têm sido utilizados na área da genética para a identificação de padrões das doenças e como, futuramente, uma nova forma de diagnóstico. Neste trabalho, foi utilizado o algoritmo de agrupamento de misturas finitas de normais com o objetivo de analisar dados de expressão gênica de pacientes com diagnóstico de doença de Alzheimer.

Metodologia: Foram analisadas as expressões gênicas diferenciadas obtidas por meio da diferença entre as expressões de pacientes com a DA e a mediana das expressões de idosos saudáveis, coletadas em seis regiões do cérebro. Para cada região, havia uma quantidade diferente de casos, controles e de transcritos a seres avaliados. O agrupamento por misturas finitas de normais foi aplicado dentro de cada região e, após identificar os grupos de transcritos mais extremos, foram identificadas as funções gênicas correspondentes.

Resultados: Para cada região do cérebro, foram identificados, utilizando o critério ICL, ao menos 2 grupos de transcritos considerados mais super-expressos, seja em pacientes com a DA, seja em idosos saudáveis. Algumas funções, como a *AUF 1 (hnRNP D0) binds and destabilizes mRNA*, foram encontradas em mais de uma região em estudo.

Conclusão: O modelo de misturas finitas de normais, utilizando o critério de seleção ICL, foi eficiente para identificar os grupos de transcritos considerados mais expressos e para obter os respectivos processos biológicos. Foi possível identificar funções específicas de cada grupo final selecionado dentro de determinada região, além de observar algumas funções comuns em mais de duas regiões.

Palavras-chave: agrupamento por misturas finitas de normais, técnicas de agrupamento, critérios de seleção de modelos, ICL, doença de Alzheimer, dados de expressão gênica.

Abstract

Introduction: Data clustering and classification techniques were refined and algorithms emerged that cluster objects by common characteristics according to each criterion. These algorithms have been used in the field of genetics to identify disease patterns and as, in the future, a new form of diagnosis. In this paper, the model-based clustering algorithm by finite mixtures of normal distribution was used to analyze gene expression data from patients diagnosed with Alzheimer's disease.

Methods: The differentiated gene expressions obtained through the difference between the expressions of patients with AD and the median expression of healthy elderly people, collected in six brain regions, were analyzed. For each region, there were a different number of cases, controls and transcripts to be evaluated. The model-based clustering by finite mixtures of normal distributions was applied within each region and, after the identification of the most extreme groups of transcripts, the corresponding gene functions were identified.

Results: For each brain region, at least 2 groups of transcripts considered to be more over-expressed were identified, using the ICL criterion, either in patients with AD or in healthy elderly people. Some functions, such as AUF 1 (hnRNP D0) binds and destabilizes mRNA, were found in more than one region under study.

Conclusions: The model-based clustering by finite mixtures of normal distribution, using the ICL selection criterion, was efficient to identify the groups of transcripts considered more expressed and to obtain the respective biological processes. It was possible to identify specific functions of each final group selected within a given region, in addition to observing some common functions in more than two regions.

Keywords: model-based clustering by finite mixtures of normal distributions, clustering techniques, model selection criteria, ICL, Alzheimer's disease, gene expression data.

Lista de Ilustrações

1	Possíveis modelos obtidos por meio das misturas finitas de normais multivariadas utilizando <i>iris data</i> (conjunto de dados explorado em exemplos de análises multivariadas) disponível no <i>software R</i> . (Ver código R no Apêndice).	21
2	GeneChip fabricado pela empresa Affymetrix. <i>Fonte: https://www.thermofisher.com/order/catalog/product/900471#/900471</i> .	27
3	<i>Dotplot</i> e <i>cnetplot</i> da região do Córtex Entorrinal.	29
4	Figura ilustrativa de um paciente saudável (à esquerda) e com a DA (à direita) retratando as alterações nos neurônios e presença de placas amiloides. <i>Fonte: Boletim Científico do Centro de Simulação e Pesquisa São Camilo, 2018</i> .	31
6	Regiões do cérebro analisadas no estudo (córtex entorrinal, hipocampo, giro temporal médio, córtex cingulado posterior, giro frontal superior e córtex visual primário). <i>Fonte: https://www.brainfacts.org/3d-brainintro=true</i> .	32
7	Representação do Córtex Entorrinal. <i>Fonte: Society for Neuroscience (2017)</i> . Disponível em: <i>https://www.brainfacts.org/3d-brainintro=truefocus=Brain-limbic_system-entorhinal_cortex</i> .	34
8	Gráficos correspondentes ao comportamento dos transcritos em cada amostra biológica dentro dos grupos obtidos pelo modelo de misturas finitas.	35
9	Boxplot de cada amostra biológica, excluindo a amostra 2, com a distribuição das expressões gênicas diferenciadas dos 21.857 transcritos no Córtex Entorrinal.	36
10	Heatmap dos 21.857 transcritos da região Córtex Entorrinal resultante do primeiro agrupamento.	37
11	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 501 transcritos no Córtex Entorrinal.	38
12	Heatmap dos 501 transcritos da região Córtex Entorrinal resultante do último agrupamento.	39
13	Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Córtex Entorrinal representado na Figura 12.	40
14	Processos biológicos da Figura 13 presentes no Córtex Entorrinal com os respectivos genes relacionados.	41
15	Representação do Hipocampo. <i>Fonte: Society for Neuroscience (2017)</i> . Disponível em: <i>https://www.brainfacts.org/3d-brainintro=truefocus=Brain-limbic_system-hippocampus-hippocampus</i> .	41
16	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 18.042 transcritos no Hipocampo.	42
17	Heatmap dos 18.042 transcritos da região Hipocampo resultante do primeiro agrupamento.	43
18	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 1.063 transcritos no Hipocampo.	45
19	Heatmap dos 1.063 transcritos da região Hipocampo resultante do último agrupamento.	46

20	Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Hipocampo representado na Figura 19.	47
21	Processos biológicos da Figura 20 presentes no Hipocampo com os respectivos genes relacionados.	48
22	Representação do lobo temporal em que encontra-se, no centro, o Giro Temporal Médio. <i>Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=truefocus=Brain-cerebral_hemisphere-temporal_lobe</i>	48
23	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 19.891 transcritos no Giro Temporal Médio.	49
24	Heatmap dos 19.891 transcritos da região Giro Temporal Médio resultante do primeiro agrupamento.	50
25	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 4.646 transcritos no Giro Temporal Médio.	52
26	Heatmap dos 4.646 transcritos da região Giro Temporal Médio resultante do último agrupamento.	53
27	Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Giro Temporal Médio representado na Figura 26. . . .	54
28	Processos biológicos da Figura 27 presentes no Giro Temporal Médio com os respectivos genes relacionados.	55
29	Representação do Córtex Cingulado Posterior. <i>Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=truefocus=Brain-cerebral_hemisphere-temporal_lobe-cingulate_cortex</i>	55
30	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 18.297 transcritos no Córtex Cingulado Posterior.	56
31	Heatmap dos 18.297 transcritos da região Córtex Cingulado Posterior resultante do primeiro agrupamento.	57
32	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 515 transcritos no Córtex Cingulado Posterior.	58
33	Heatmap dos 515 transcritos da região Córtex Cingulado Posterior resultante do último agrupamento.	59
34	Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Córtex Cingulado Posterior representado na Figura 33. . . .	61
35	Processos biológicos da Figura 34 presentes no Córtex Cingulado Posterior com os respectivos genes relacionados.	61
36	Representação do lobo frontal no qual o Giro Frontal Superior está localizado em sua parte superior, próximo à linha central do cérebro. <i>Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=truefocus=Brain-cerebral_hemisphere-frontal_lobe</i>	62
37	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 20.643 transcritos no Giro Frontal Superior.	63
38	Heatmap dos 20.643 transcritos da região Giro Frontal Superior resultante do primeiro agrupamento.	64
39	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 642 transcritos no Giro Frontal Superior.	65
40	Heatmap dos 642 transcritos da região Giro Frontal Superior resultante do último agrupamento.	66

41	Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Giro Frontal Superior representado na Figura 40. . . .	67
42	Processos biológicos da Figura 41 presentes no Giro Frontal Superior com os respectivos genes relacionados.	68
43	Representação do Córtex Visual Primário. <i>Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=truefocus=Brain-cerebral_hemisphere-occipital_lobe-primary_visual_cortex</i>	68
44	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 18.664 transcritos no Córtex Visual Primário.	69
45	Heatmap dos 18.664 transcritos da região Córtex Visual Primário resultante do primeiro agrupamento.	70
46	Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 332 transcritos no Córtex Visual Primário.	71
47	Heatmap dos 332 transcritos da região Córtex Visual Primário resultante do último agrupamento.	72
48	Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Córtex Visual Primário representado na Figura 47. . .	73
49	Processos biológicos da Figura 48 presentes no Córtex Visual Primário com os respectivos genes relacionados.	73
50	Passo a passo para obter os dados pré-processados pelos autores Liang et. al. (2008)	83

Lista de Tabelas e Quadros

Lista de Tabelas

1	Número de indivíduos e transcritos que compuseram os dados finais	33
2	Resultados do ajuste do agrupamento inicial da região Córtex Entorrinal após a retirada da amostra 2.	37
3	Divisão em grupos dos 21.857 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.	37
4	Resultados do ajuste do agrupamento final da região Córtex Entorrinal. . .	39
5	Divisão em grupos dos 501 transcritos, obtidos pelo modelo final, e respectivas medianas.	40
6	Resultados do ajuste do agrupamento inicial da região Hipocampo.	43
7	Divisão em grupos dos 18.042 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.	44
8	Resultados do ajuste do agrupamento intermediário da região Hipocampo. . . .	44
9	Divisão em grupos dos 7.016 transcritos, obtidos pelo modelo anterior, e respectivas medianas.	44
10	Resultados do ajuste dos agrupamento finais da região Hipocampo.	46
11	Divisão em grupos dos 1.063 transcritos, obtidos pelo modelo final, e respectivas medianas.	46
12	Resultados do ajuste dos agrupamentos inicial e finais da região Giro Temporal Médio.	50
13	Divisão em grupos dos 19.891 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.	51
14	Resultados do ajuste do agrupamento intermediário da região Giro Temporal Médio.	51
15	Divisão em grupos dos 6.915 transcritos, obtidos pelo modelo anterior, e respectivas medianas.	51
16	Resultados do ajuste dos agrupamentos finais da região Giro Temporal Médio.	53
17	Divisão em grupos dos 4.646 transcritos, obtidos pelo modelo final, e respectivas medianas.	53
18	Resultados do ajuste do agrupamento inicial da região Córtex Cingulado Posterior.	57
19	Divisão em grupos dos 18.297 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.	58
20	Resultados do ajuste do agrupamento final da região Córtex Cingulado Posterior.	59
21	Divisão em grupos dos 515 transcritos, obtidos pelo modelo final, e respectivas medianas.	60
22	Resultados do ajuste do agrupamento inicial da região Giro Frontal Superior. . . .	63
23	Divisão em grupos dos 20.643 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.	64
24	Resultados do ajuste do agrupamento final da região Giro Frontal Superior. . . .	66

25	Divisão em grupos dos 642 transcritos, obtidos pelo modelo final, e respectivas medianas.	66
26	Resultados do ajuste do agrupamento inicial da região Córtex Visual Primário.	69
27	Divisão em grupos dos 18.664 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.	70
28	Resultados do ajuste do agrupamento final da região Córtex Visual Primário.	71
29	Divisão em grupos dos 332 transcritos, obtidos pelo modelo final, e respectivas medianas.	72

Lista de abreviaturas e siglas

VSO	<i>Volume-Shape-Orientation</i>
EM	<i>Expectation-Maximization</i>
BIC	Critério de informação bayesiana
ICL	Critério de verossimilhança de dados completos integrados
LRT	Teste de razão de verossimilhança
LRTS	Estatística do teste de razão de verossimilhança
DNA	Ácido Desoxirribonucleico
RNA	Ácido Ribonucleico
mRNA	RNA mensageiro
FDR	<i>False Discovery Rate</i>
DA	Doença de Alzheimer
SNP	Polimorfismo de Nucleotídeo Único

Sumário

1 INTRODUÇÃO	17
2 METODOLOGIA	19
2.1 MODELOS DE MISTURAS FINITAS.	19
2.1.1 Estimação por meio do método de máxima verossimilhança	22
2.1.2 Seleção do modelo e do número de grupos	24
2.2 DADOS DE EXPRESSÃO GÊNICA	26
2.2.1 Função Gênica	27
3 APLICAÇÃO	30
3.1 DOENÇA DE ALZHEIMER	30
3.2 ESTUDO DE CASO	31
3.2.1 Estudo das Regiões do Cérebro	33
4 DISCUSSÃO E CONCLUSÃO	75
REFERÊNCIAS	78
5 APÊNDICE	80
5.1 LEITURA DOS DADOS PRÉ-PROCESSADOS POR LIANG ET. AL. (2008)	80
5.2 Código para gerar os gráficos da Figura 1	83

1 INTRODUÇÃO

O estudo de técnicas de agrupamento e classificação de dados vem sendo cada vez mais importante por conta do advento de grandes bases de dados envolvendo muitas observações (“*big data*”) e/ou muitas variáveis (dados superdimensionados). Dessa forma, novas técnicas têm surgido para tratamento desses dados, por meio de algoritmos para agrupamento e classificação de objetos de estudo (indivíduos, plantas, animais) de acordo com as características, sem depender apenas do senso comum (BOUVEYRON et al., 2019).

Um dos algoritmos desenvolvidos envolve modelos de misturas finitas, em que a base de dados é formada pela mistura de densidades de probabilidades que irão determinar grupos com características específicas. No início do processo, cada um dos objetos constitui um grupo e vão se unindo até que todos os objetos estejam juntos em um único grupo (SCRUCCA, 2016). O objetivo é encontrar a etapa do processo que melhor representa a base de dados, determinando o número de grupos, a melhor mistura para ajustar os dados e qual a incerteza associada ao agrupamento (BOUVEYRON et al., 2019).

As técnicas de agrupamento e classificação de dados possuem aplicações em diversas áreas, como o agrupamento de genes e de pessoas (usando dados genéticos), agrupamento e padrões em dados de código de barras de varejo, formação de grupos de usuários em páginas de *Internet*, como redes sociais e lojas, análise e compressão de imagens (BOUVEYRON et al., 2019). Entre os muitos exemplos de aplicações com mistura de normais, pode-se destacar SILVA, 2012 que realiza o agrupamento de dados espectrais de café via misturas finitas, e o estudo que compara o modelo de mistura de normais e diversos outros algoritmos de agrupamento com um desenvolvido especificamente para dados super dimensionados com pequenas amostras e aplicam no agrupamento de dados de expressão genética (von Borries; WANG, 2009). SOUZA et al., 2017 apresentam ainda um interessante estudo de astroestatística para classificação de galáxias por meio da mistura de normais.

O uso de técnicas de agrupamento em genética tem se mostrado uma importante área de investigação. Neste trabalho, estuda-se o agrupamento de dados de expressão genética de indivíduos diagnosticados com a doença de Alzheimer (DA). Esta doença está associada à idade e à perda progressiva das funções intelectuais, sendo neurodegenerativa que afeta, principalmente, a memória recente dos indivíduos (eles se recordam dos acontecimentos a longo prazo, porém não se lembram dos recentes), além de perda de funções da linguagem, funções cognitivas e alterações de humor que podem ocorrer com a evolução da doença (SERENIKI; VITAL, 2008). O diagnóstico desta doença é feito, atualmente, por imagens ou análise do líquido cerebrospinal, porém essa forma de diagnóstico pode ser

ineficiente e imprecisa quando o paciente está em um estágio inicial da doença, afetando, assim, o início dos tratamentos (WANG; LIU, 2019).

No trabalho de LIANG et al., 2008 é destacada a importância de interpretar as diferenças entre as expressões gênicas de pacientes saudáveis e diagnosticados com DA. Este tipo de estudo pode ser utilizado para compreender os mecanismos moleculares associados à patologia da doença, como a formação de placas amiloides e os emaranhados neurofibrilares que são características desenvolvidas ao longo da evolução da doença.

Assim, tomando como base o trabalho de WANG; LIU, 2019 e LINS, 2021 deseja-se explorar os genes expressos em pacientes com diagnóstico de Alzheimer por meio de técnicas de agrupamento com base em modelos de misturas finitas de normais para dados superdimensionados. Para isso, são analisados os dados dos genes de seis regiões do cérebro a fim de identificar padrões nas expressões gênicas em cada uma das regiões.

2 METODOLOGIA

A técnica de agrupamento baseada em modelos probabilísticos consiste em agrupar os dados de acordo com a probabilidade de pertencerem a determinada distribuição. Neste trabalho, será abordada a família de modelos de misturas finitas de distribuições normais multivariadas. Além disso, serão utilizados dados de expressão gênica que possuem algumas particularidades para realizar a aplicação.

2.1 MODELOS DE MISTURAS FINITAS

Os algoritmos hierárquicos aglomerativos consistem no agrupamento das observações similares utilizando algum critério de similaridade, como ligação simples ou completa. Esta forma de agrupamento se inicia com cada observação em um grupo distinto e une os pares de observação mais similares pelo critério estabelecido até que, ao final, todos estejam em um único grupo. A cada iteração, dois grupos são mesclados de maneira que minimizem o critério de similaridade escolhido ou maximizem a verossimilhança da classificação (do inglês *classification likelihood*) sobre todos os possíveis pares a serem agrupados e, portanto, quando há n observações, haverá $n - 1$ estágios no algoritmo de agrupamento.

Entre os critérios de similaridade, pode-se destacar os mais comuns que são (JOHNSON; WICHERN et al., 2007 e BOUVEYRON et al., 2019):

- Ligação simples: une as observações mais próximas; menor dissimilaridade entre a observação de um grupo e a de outro grupo.
- Ligação média: une pontos médios; dissimilaridade média entre os grupos.
- Ligação completa: une os pontos mais distantes; maior dissimilaridade entre os grupos.
- Soma de quadrados: soma de quadrados total dentro do grupo.

As técnicas de agrupamento a serem estudadas são baseadas em modelos probabilísticos e utilizam o princípio de mistura de distribuição, além de métodos-padrão para realizar a inferência. Essa forma de agrupamento pode ser classificada como algoritmo hierárquico aglomerativo e é explorada em estudos que envolvem muitas variáveis e poucas observações. Além disso, os agrupamentos baseados em modelos permitem verificar a incerteza ao agrupar as observações nos grupos.

O principal conceito da técnica baseada em modelo de misturas finitas é o princípio da mistura de distribuições. Assim, a modelagem representa a distribuição de probabilidade de uma observação multivariada na forma de uma mistura finita ou uma média ponderada das densidades de probabilidade, como mostrado na equação (2.1.1) abaixo (BOUYEYRON et al., 2019):

$$p(y_i; \Phi) = \sum_{g=1}^G \tau_g f_g(y_i | \theta_g), \quad (2.1.1)$$

em que:

- Φ é o vetor dos parâmetros μ e Σ do modelo de mistura;
- G é a quantidade de componentes da mistura;
- τ_g é a probabilidade de determinada observação ser gerada pela g -ésima componente;
- $f_g(y_i | \theta_g)$ é a densidade da g -ésima componente dado o parâmetro θ_g .

De maneira geral, a densidade f_g segue uma distribuição Normal Uni ou Multivariada, dependendo do tipo de estudo. Dessa forma, no caso univariado, f_g tem como parâmetros a média μ_g e o desvio padrão σ_g correspondentes a cada grupo da mistura; e, no caso multivariado, ela é parametrizada por um vetor μ_g de médias e uma matriz Σ_g de variância-covariâncias e estes elementos auxiliam na formação dos grupos que são centralizados em torno do vetor de médias. Por meio desta matriz, pode-se determinar a superfície de cada grupo, como volume, forma e orientação (SCRUCCA et al., 2016). Esta distribuição pode ser expressa por meio da seguinte fórmula:

$$\phi_g(y_i | \mu_g, \Sigma_g) = |2\pi \Sigma_g|^{-\frac{1}{2}} e^{-\frac{1}{2}(y_i - \mu_g)^T \Sigma_g^{-1} (y_i - \mu_g)}. \quad (2.1.2)$$

Ressalta-se que, apesar do agrupamento por misturas finitas ser explorado em estudos que envolvem muitas variáveis como mencionado anteriormente, o modelo é sensível ao número de parâmetros. Isso significa que, quando esta quantidade aumenta, podem surgir dificuldades na interpretação dos resultados e nas estimações, devido ao aumento da dimensão da matriz de variância-covariâncias. Uma maneira proposta para corrigir esse problema é a decomposição da matriz de covariância do grupo g (Σ_g) em autovalores, conforme a equação (2.1.3), e é conhecida, em alguns casos, de decomposição geométrica ou VSO (*Volume-Shape-Orientation*). Esta maneira de decomposição foi proposta porque, por meio dela, são obtidas possíveis estruturas para a matriz Σ_g e auxilia no estudo da geometria dos grupos, como mostrado na Figura 1.

$$\Sigma_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^T. \tag{2.1.3}$$

Recebe este nome pois cada componente, presente na equação (2.1.3), é responsável por uma “propriedade geométrica” de volume, forma e orientação: λ_g é uma constante de proporcionalidade que determina o volume que a g -ésima componente ocupa, enquanto \mathbf{D}_g (matriz de autovetores) define a orientação e a matriz diagonal de autovalores em escala \mathbf{A}_g estabelece o formato.

Portanto, com auxílio da decomposição VSO, é possível obter 14 diferentes modelos por meio das misturas finitas de normais multivariadas, envolvendo a forma da estrutura da distribuição, volume, forma e orientação da matriz de covariância (SCRUCCA et al., 2016). Para nomear cada modelo, foi utilizada a ordem VSO (*Volume-Shape-Orientation*) e as letras E, V e I (*Equal, Varying, Identity*) para representar igualdade, variabilidade e modelo com representação esférica, respectivamente. O “volume” da matriz de covariância pode ser igual (E), quando existe a restrição dos grupos possuírem o mesmo volume, ou variável (V) quando não há essa restrição; a forma da matriz diagonal pode ser igual (E), variável (V), pelos mesmos motivos do volume, e ainda possui uma terceira opção I que representa grupos esféricos. Por fim, a orientação dos grupos pode ser classificada da mesma maneira que a forma da matriz diagonal. A Figura 1 ilustra todos os 14 possíveis modelos.

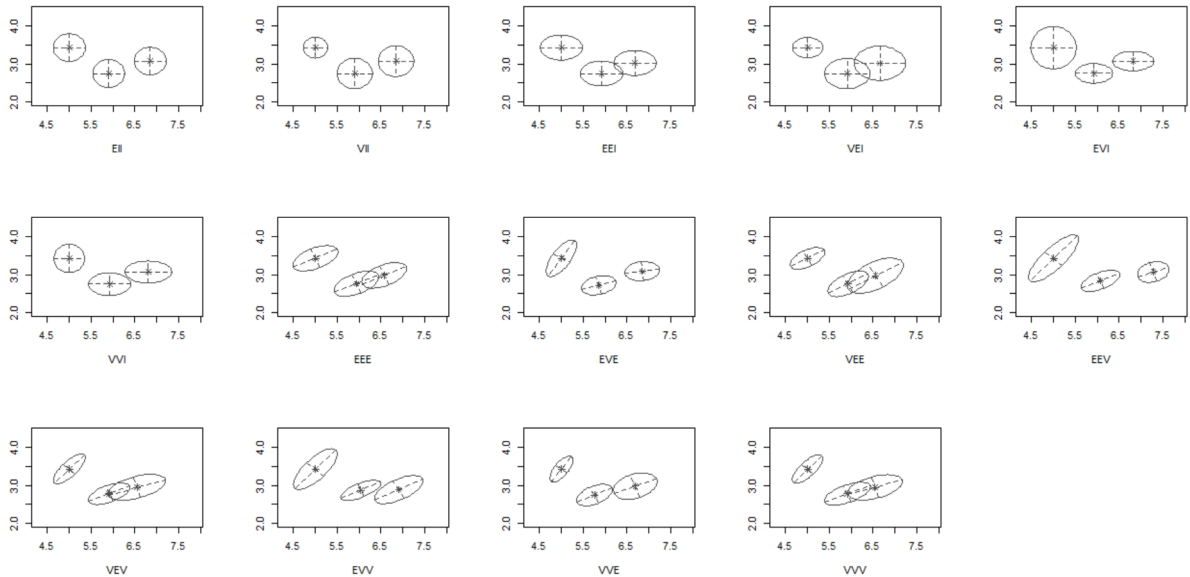


Figura 1: Possíveis modelos obtidos por meio das misturas finitas de normais multivariadas utilizando *iris data* (conjunto de dados explorado em exemplos de análises multivariadas) disponível no *software R*. (Ver código R no Apêndice).

2.1.1 Estimação por meio do método de máxima verossimilhança

Os modelos de misturas finitas de normais multivariadas podem ser estimados por meio do critério de máxima verossimilhança. Existem duas verossimilhanças obtidas em modelos de misturas finitas (BOUYEYRON et al., 2019): verossimilhança dos dados completos (L_c - *complete-data likelihood*) e verossimilhança dos dados observados (L_O - *observed data likelihood*) que também pode ser chamada de verossimilhança da mistura (*mixture likelihood*). Considerando que o conjunto de dados é formado por n observações multivariadas (y_i, z_i) , em que y_i e z_i são, respectivamente, os valores observados e não-observados, $\mathbf{y} = (y_1, \dots, y_n)$ e $\mathbf{z} = (z_1, \dots, z_n)$, então é possível escrever as duas verossimilhanças da seguinte forma:

$$L_C(\mathbf{y}, \mathbf{z} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \prod_{i=1}^n \phi(y_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad (2.1.4)$$

e

$$L_O(\mathbf{y} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \int L_C(\mathbf{y}, \mathbf{z} | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) dz = \prod_{i=1}^n \sum_{g=1}^G \tau_g \phi(y_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad (2.1.5)$$

em que

- $\phi(y_i | \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ é a distribuição Normal Multivariada representada na equação 2.1.2;
- $\mathbf{y} = (y_1, \dots, y_n)$ é o vetor formado pelos valores observados;
- $\mathbf{z} = (z_1, \dots, z_n)$ é o vetor dos valores não-observados no qual cada z_i pode assumir valores 0 ou 1, indicando a população de onde a observação y_i pertence;
- $\boldsymbol{\mu}_g$ é o vetor de médias e $\boldsymbol{\Sigma}_g$ é a matriz de variância-covariâncias da g -ésima componente da mistura.

Assim, para estimar os parâmetros da mistura ($\boldsymbol{\theta} = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$), deve-se maximizar a verossimilhança dos dados observados (Equação 2.1.5) com respeito a $\boldsymbol{\theta}$. Para isto, é aplicado o algoritmo EM que é a maneira mais comum de se obter esta estimativa.

2.1.1.1 Algoritmo EM (*Expectation-Maximization*)

O algoritmo EM é numericamente estável com convergência global sob condições gerais, porém nem sempre converge para o máximo global, porque seus resultados são sensíveis aos valores iniciais. O algoritmo EM é dividido em dois passos: *Expectation* (E) e *Maximization* (M). No primeiro passo (E), é calculada a probabilidade condicional

da log-verossimilhança dos dados completos, considerando que há uma parte dos dados que já foi observada e que os parâmetros estimados foram computados. No segundo passo (M), são determinados os valores dos parâmetros que maximizam a log-verossimilhança do passo anterior. Os passos são repetidos até ocorrer convergência, i.e., os valores estimados no passo M não sofrem alteração ou a variação é desprezível (SCRUCCA et al., 2016).

Dessa forma, pode-se escrever os dois passos do algoritmo em duas fórmulas gerais que resumem os cálculos das estimativas e classificações das observações em grupos, considerando um modelo de misturas finitas de normais multivariadas. As etapas consistem em obter uma estimativa para $\hat{z}_{i,g}^{(s)}$ (equação 2.1.6) e para os parâmetros que devem maximizar a log-verossimilhança l_C da equação (2.1.7) (BOUVEYRON et al., 2019).

$$\hat{z}_{i,g}^{(s)} = \frac{\hat{\tau}_g^{(s-1)} \phi_g(y_i | \hat{\mu}_g^{(s-1)}, \hat{\Sigma}_g^{(s-1)})}{\sum_{h=1}^G \hat{\tau}_h^{(s-1)} \phi_h(y_i | \hat{\mu}_h^{(s-1)}, \hat{\Sigma}_h^{(s-1)})}, \quad (2.1.6)$$

na qual $\hat{z}_{i,g}^{(s)}$ é a esperança condicional de $z_{i,g}^{(s)}$ dados os valores dos parâmetros e dos valores observados \mathbf{y} e estima a probabilidade condicional da observação i pertencer ao grupo g no passo s .

$$l_C = \sum_{i=1}^n \sum_{g=1}^G z_{i,g} \log[\tau_g \phi_g(y_i | \mu_g, \Sigma_g)]. \quad (2.1.7)$$

Como o agrupamento consiste em misturas de normais multivariadas, as estimativas para os parâmetros possuem fórmulas fechadas dadas pelo seguinte cálculo:

$$\hat{\tau}_g = \frac{\hat{n}_g^{(s-1)}}{n}; \quad \hat{\mu}_g^{(s)} = \frac{\sum_{i=1}^n \hat{z}_{i,g}^{(s-1)} y_i}{\hat{n}_g^{(s-1)}}; \quad \hat{n}_g^{(s-1)} = \sum_{i=1}^n \hat{z}_{i,g}^{(s-1)}. \quad (2.1.8)$$

Com isso, é possível mensurar a incerteza associada à alocação de uma observação em um determinado grupo (Equação 2.1.9). De forma geral, a incerteza (do inglês, *uncertainty*) é maior quando *Uncer* se aproxima de $1/G$ e é menor no caso em que *Uncer* se aproxima de zero (ou seja, os $\hat{z}_{i,g}$'s são próximos a 1 que significaria uma alocação quase perfeita).

$$Uncer_i = 1 - \max_{g=1, \dots, G} \hat{z}_{i,g}. \quad (2.1.9)$$

Como mencionado no início desta seção, os resultados do algoritmo EM são sensíveis aos valores iniciais escolhidos. Isso porque a verossimilhança é sensível aos valores dos parâmetros e pode, nos casos de misturas finitas de normais, tender a infinito caso

os valores iniciais estejam próximos aos limites (por exemplo, $\sigma \rightarrow 1$ e $\mu \rightarrow y_i$), sendo necessário aplicar outra solução envolvendo ponto de máximo local.

Para selecionar os valores iniciais do algoritmo, existem dois métodos: agrupamento baseado em modelos hierárquicos e o método *smallem*. O primeiro tem como característica principal o agrupamento de grupos por meio da minimização de algum critério definido e o segundo é implementado separadamente para cada modelo e quantidade de grupos. De maneira geral, ele é iniciado com valores aleatórios para os parâmetros, enquanto os modelos hierárquicos utilizam o número de grupos como pontos iniciais. O pacote *mclust* utiliza o método de agrupamento baseado em modelos hierárquicos para iniciar a estimação EM, nos casos multivariados, e, em casos univariados, utiliza os quantis da distribuição.

Atualmente, o critério utilizado nos agrupamentos hierárquicos para decidir quais serão as observações que devem ser agrupadas é o de maximização da verossimilhança da classificação. Este critério se inicia com cada observação em um grupo distinto e, a cada iteração, os pares de observações são mesclados de maneira que gerem o maior aumento da verossimilhança que está representada na equação (2.1.10) abaixo. Por meio deste critério, é possível obter não somente a estimação de quais observações pertencem a cada grupo, mas também as estimativas dos parâmetros do modelo.

$$L_{CL}(\theta, z|y) = \prod_{i=1}^n f_{z_i}(y_i|\theta_{z_i}) \quad (2.1.10)$$

2.1.2 Seleção do modelo e do número de grupos

Uma das dificuldades encontradas nos primeiros agrupamentos é saber quantos grupos serão formados com a base de dados e qual o melhor modelo para o ajuste. Com a técnica de misturas finitas, é possível obter essas respostas, porém essas escolhas entram, de certa forma, em conflito. Isso porque um modelo mais simples requer uma maior quantidade de grupos para poder chegar em melhores resultados, enquanto um modelo mais complexo não necessita de uma grande quantidade de grupos; ocorre um *trade-off* entre complexidade do modelo e número de grupos.

As medidas mais utilizadas para a escolha do modelo são critérios de informação que são baseados na penalização da log-verossimilhança. A mais utilizada é o BIC (critério de informação bayesiana, do inglês *bayesian information criterion*) e uma segunda opção é o ICL (critério de verossimilhança de dados completos integrados, do inglês *integrated complete-data likelihood criterion*). De forma geral, a escolha do modelo é feita, em ambos os casos, considerando o par (modelo, número de componentes) que maximiza o valor da informação bayesiana.

O BIC tende a escolher o número de componentes (número de grupos que irá dividir as observações) necessário para aproximar o modelo da respectiva densidade, ao invés de selecionar o número de grupos dentro dos próprios dados (SCRUCCA et al., 2016). Este critério é calculado conforme a equação abaixo (2.1.11) e está dividido em dois termos: o primeiro é a log-verossimilhança do modelo M com G componentes ($l_{M,G}(\mathbf{x}|\Phi)$) e o segundo é o termo de penalização para o número de parâmetros estimado (v), considerando o tamanho da amostra (n).

$$BIC_{M,G} = 2 \cdot l_{M,G}(\mathbf{x}|\Phi) - v \cdot \log(n). \quad (2.1.11)$$

Já o ICL, apresentado na equação (2.1.12), tem melhor desempenho na seleção do número de grupos. Como ele penaliza o BIC por meio de um termo que mede a sobreposição de grupos, ele é melhor, principalmente, quando os grupos estão bem definidos no conjunto de dados (SCRUCCA et al., 2016). Dessa forma, é mais utilizado quando o objetivo principal é o agrupamento das informações ao invés de encontrar o melhor modelo de misturas que se ajuste aos dados (BOUYEYRON et al., 2019).

$$ICL_{M,G} = BIC_{M,G} + 2 \cdot \sum_{i=1}^n \sum_{g=1}^G c_{i,g} \log(z_{i,g}), \quad (2.1.12)$$

em que $c_{i,g}$ é uma constante que assume valores 1 se a observação i pertence ao grupo g e 0 caso contrário, e $z_{i,g}$ é a probabilidade condicional da observação i pertencer ao grupo g , estimada pela Equação 2.1.6.

Além dessas medidas, é possível utilizar o teste de razão de verossimilhança (em inglês, *likelihood ratio test* - LRT) para verificar qual a melhor “ordem” do modelo de misturas. Isso significa que o teste irá comparar o modelo (selecionado entre os 14 apresentados na Figura 1) com duas opções para a quantidade de grupos (G_0 é o número de grupos da hipótese nula e G_1 da hipótese alternativa) e, por meio da estatística do teste de razão de verossimilhança (LRTS - likelihood ratio test statistic) (2.1.13), calcula-se o p-valor para verificar se G_0 é melhor que G_1 .

$$LRTS = 2[l(\hat{\Phi}_{G_1}) - l(\hat{\Phi}_{G_0})], \quad (2.1.13)$$

em que $l(\hat{\Phi}_{G_j})$ é o log função de verossimilhança e $\hat{\Phi}$ é o estimador de máxima verossimilhança do vetor de parâmetros da distribuição 2.1.1.

Dessa forma, valores elevados da estatística LRTS evidenciam a rejeição da hipótese nula. O p-valor do teste pode ser obtido por meio de um algoritmo *bootstrap* que estima a distribuição da estatística do teste dada a hipótese nula (como a distribuição

testada é uma mistura de normais, não é possível afirmar que a distribuição de LRTS, sob H_0 , é uma qui-quadrado e, portanto, utiliza-se de outros métodos para obtê-la).

Apesar das três técnicas serem as mais comuns, o algoritmo *mclust* utiliza como *default* o critério de informação bayesiana (BIC). No presente trabalho, como o objetivo envolve escolher o número de componentes em um modelo de mistura, optou-se pelo critério de verossimilhança de dados completos integrados (ICL).

2.2 DADOS DE EXPRESSÃO GÊNICA

A expressão gênica é um processo em que a sequência de DNA de um gene pode ser transcrita em RNA ou em um polipéptido, caso o gene em estudo seja codificante de proteínas (GRIFFITHS et al., 2019). Ela pode ser dividida em duas etapas (transcrição e tradução), nas quais ocorre a formação do RNA mensageiro (mRNA) por meio do DNA e a formação da proteína por meio da decodificação da informação armazenada no mRNA. Nesse processo, a quantidade de proteína produzida é estimada com base na quantidade de mRNA.

Existem diversas técnicas que permitem a obtenção desses dados e a mais utilizada é a de *microarrays* que permite medir o nível de expressão de muitos genes em paralelo (por exemplo, em diferentes tecidos). O chip de *microarrays* é formado por vários transcritos¹ e cada um deles é composto por um grupo de sondas idênticas que, no caso deste trabalho, são de oligonucleotídeos². No mercado, os microarranjos de sondas oligonucleotídeas são fabricados por duas empresas - Affymetrix e Illumina (DUDOIT et al., 2002).

Para medir a expressão gênica, utiliza-se um processo no qual cada sonda se liga a sua sequência complementar de nucleotídeos. Esse processo é chamado de hibridização e ocorre entre a informação genética contida no mRNA e a sonda oligonucleotídea, formando, assim, um transcrito da amostra. Após esse processo, a quantidade de DNA complementar³ ligada à sonda é medida por meio da intensidade de fluorescência, permitindo obter uma estimativa do nível de expressão gênica na amostra.

Esta expressão coletada na autópsia necessita passar por um pré-processamento para que esteja em um formato adequado para análise. Esse procedimento possui algumas técnicas específicas que consistem em obter estimativas dos níveis de expressão dos transcritos que, nos dados brutos, são valores resultantes da intensidade fluorecente das sondas de oligonucleotídeos. Entre elas, destaca-se a MAS5.0 (*Affymetrix Microarray Suite 5 method*) que foi utilizada pelos autores do estudo que deu origem aos dados que serão

¹Quantidade de RNA mensageiro produzido no DNA (GRIFFITHS et al., 2019).

²São pequenos segmentos de pares de bases retirados de uma parte do DNA ou RNA (MCLACHLAN; DO; AMBROISE, 2004).

³“DNA sintetizado a partir de um molde de mRNA”(MOREIRA, 2014).

aqui utilizados (LIANG et al., 2008). Esta técnica (MAS5.0), implementada no *software* GCOS da empresa *Affymetrix*, consiste em subtrair o ruído de fundo das intensidades das sondas de linha e realizar um processo de normalização para obter valores da expressão gênica normalizados e em uma única escala.

Os dados que serão analisados, neste trabalho, foram obtidos por meio do chip de microarranjos de oligonucleotídios desenvolvido pela empresa *Affymetrix*. Ele contém, aproximadamente, 55.000 transcritos e recebe o nome de *GeneChip Human Genome U133A 2.0 Array* (Figura 2). As expressões gênicas medidas foram coletadas de tecidos seccionados (serão chamados de amostras) de seis regiões do cérebro que estão relacionadas com a doença de Alzheimer e o envelhecimento do encéfalo, tanto por conta da histologia quanto pelo metabolismo.



Figura 2: GeneChip fabricado pela empresa *Affymetrix*. Fonte: <https://www.thermofisher.com/order/catalog/product/900471#/900471>.

2.2.1 Função Gênica

Após a realização dos agrupamentos, foi decidido analisar as funções gênicas dos transcritos considerados mais “diferencialmente” expressos por critérios pré-definidos. Para isso, foram utilizados os procedimentos abordados por PAGÈS et al., 2020 e YU; HE, 2016 que consistem em encontrar os genes correspondentes aos transcritos e obter as respectivas funções que são consideradas “super representadas” (do inglês, *over-represented*) por eles, respectivamente.

Primeiro, foi necessário obter o nome dos genes correspondentes a cada transcrito (ou conjunto de transcritos), assim como o respectivo ID que seria utilizado no passo seguinte. Essas informações foram extraídas do *hgu133plus2.db* (CARLSON, 2016) que é um pacote que contém as informações correspondentes ao chip da *Affymetrix* utilizado para coletar os dados apresentados neste trabalho e, por meio dele e com auxílio do pacote *AnnotationDbi* (PAGÈS et al., 2020), os nomes e IDs dos genes são armazenados.

Após esta etapa, utilizou-se o pacote *ReactomePA* (YU; HE, 2016) para obter as respectivas funções “super-representadas” por estes genes. Esta etapa consiste na análise

chamada de “Análise de Super-Representação” (do inglês, *Over Representation Analysis* (ORA)) que determina se funções ou processos biológicos conhecidos estão presentes em uma determinada lista de genes, neste caso, diferencialmente expressos por meio de um teste estatístico que utiliza a distribuição hipergeométrica para obter o p-valor (Equação 2.2.1).

$$\text{P-valor} = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}}, \quad (2.2.1)$$

em que:

- N representa o número total de genes na distribuição de fundo;
- M é o número de genes dentro dessa distribuição que são anotados para o conjunto de genes de interesse;
- n é o tamanho da lista de genes de interesse;
- k representa o número de genes dentro desta lista de interesse que são anotados no conjunto de genes;

O p-valor obtido deve ser ajustado por algum método de correção para testes de comparação múltipla. Optou-se por utilizar a correção *False Discovery Ratio* (opção *default* do pacote usado nas análises) que consiste em controlar a proporção esperada de hipóteses erroneamente rejeitadas entre todas que foram rejeitadas (BENJAMINI; HOCHBERG, 1995). Dessa forma, define-se um valor de corte para rejeitar a hipótese de ausência de super-representação e compara-se com o p-valor ajustado. Neste trabalho, foi utilizado como corte 0,05.

Dotplot e Cnetplot

O resultado do teste descrito anteriormente pode ser representado em dois gráficos distintos: *dotplot* e *cnetplot*. O *dotplot* tem como objetivo ilustrar o p-valor obtido na análise ORA para cada função gênica classificada como “super-representada”, além de inserir a informação da razão/proporção de genes ($GeneRatio = \frac{k}{n}$). O *cnetplot* mostra as funções encontradas no conjunto de genes de interesse e os respectivos genes responsáveis por determinado processo biológico, sendo possível, assim, saber quais funções possuem genes em comum.

Um exemplo destes gráficos está apresentado na Figura 3 abaixo e será discutido na aplicação presente no Capítulo 3 para dados de expressão gênica de pacientes com a doença de Alzheimer.

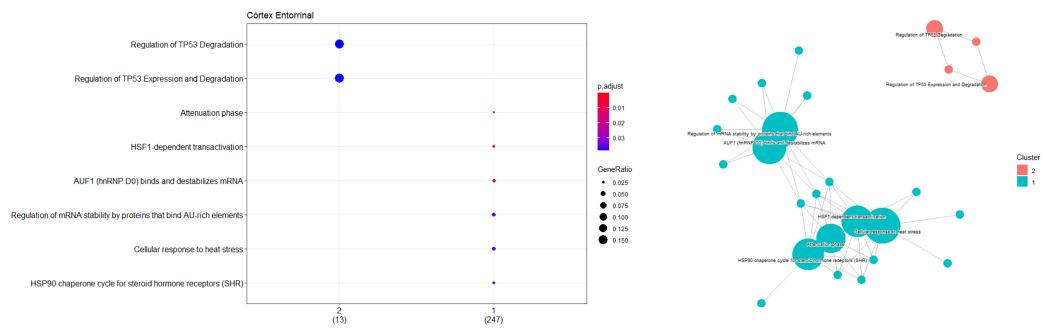


Figura 3: Dotplot e cnetplot da região do Córtex Entorrinal.

3 APLICAÇÃO

3.1 DOENÇA DE ALZHEIMER

A expectativa de vida tem aumentado ao longo dos anos e, com isso, também aumenta a quantidade de pessoas com alterações cognitivas. Essas modificações podem não impactar na qualidade de vida do idoso, sendo apenas alguns lapsos de memória, ou podem resultar em prejuízos mais graves, causando perda da capacidade de raciocínio, por exemplo. Atividades motoras, alteração na memória e no padrão de sono são outras alterações relacionadas à idade avançada e que são resultados de mudanças nas funções e estruturas do encéfalo. Os prejuízos cognitivos relacionados à idade são conhecidos como demência senil e podem, na maioria dos casos, evoluir para a doença de Alzheimer⁴ (DA). (KANDEL et al., 2000).

A DA é uma doença neurodegenerativa⁵ e é caracterizada, principalmente, pela amnésia anterógrada na qual o indivíduo se recorda dos acontecimentos a longo prazo, porém não se lembra nos recentes (SANAR, 2018). Os primeiros sintomas costumam aparecer a partir dos 70 anos (KANDEL et al., 2000) e envolvem alterações cognitivas relacionadas às “funções da linguagem, redução da velocidade de pensamento, dificuldade na orientação espaço-temporal” (SANAR, 2018) e os pacientes também podem desenvolver perdas nas funções executivas, tornando-os debilitados e dependentes. Além disso, pessoas com essa doença podem apresentar alterações comportamentais devido à deficiência na memória declarativa e à perda das capacidades cognitivas.

Além de alterações psicológicas, também são observados prejuízos no encéfalo de pacientes com esta comorbidade. Atrofiação, peso diminuído, ventrículos aumentados, neurônios com anormalidades no citoesqueleto e presença de placas amiloides⁶ são algumas das mudanças mais comuns desenvolvidas ao longo da doença. As regiões encefálicas mais vulneráveis são: área entorrinal, núcleo basal, hipocampo e neocórtex (estas últimas regiões possuem alterações que são, possivelmente, associadas às causas estruturais para o problema de memória desenvolvido na DA, diferentemente do que ocorre devido à idade que as alterações estão presentes nos circuitos fronto estriais).

⁴Segundo Kandel (2000), ela se tornou um dos maiores problemas de saúde pública nos EUA por conta do aumento no número de idosos do país.

⁵Doença que apresenta ampla morte de neurônios.

⁶Segundo Kandel (2000), “são depósitos extracelulares de polímeros de peptídeos β -amilóide”.

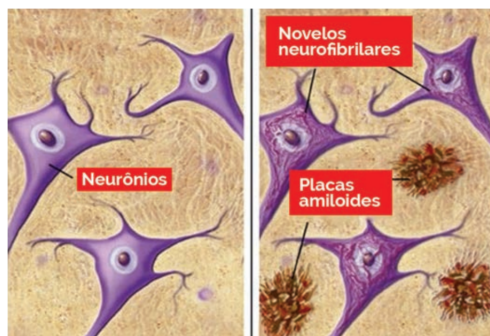


Figura 4: Figura ilustrativa de um paciente saudável (à esquerda) e com a DA (à direita) retratando as alterações nos neurônios e presença de placas amiloides. *Fonte: Boletim Científico do Centro de Simulação e Pesquisa São Camilo, 2018.*

Atualmente, não se conhece claramente a causa para o desenvolvimento da doença de Alzheimer. Porém, existem alguns genes que parecem estar associados ao surgimento e desenvolvimento da doença e todos eles estão envolvidos na produção do peptídeo⁷ β -amiloide. Eles estão presentes nas proteínas APP (proteína precursora de amiloide), PS1 e PS2 (pré-selina 1 e 2) que estão relacionadas com o surgimento precoce da doença e ApoE (Apolipoproteína-E) que está associada ao surgimento tardio (SANAR, 2018). Portanto, poucos indivíduos desenvolvem a DA por terem alelos mutantes para os genes da APP e da pré-selina, mas o alelo ApoE4 é um dos fatores de risco genético desta comorbidade.

Apesar das características elencadas acima, o diagnóstico precoce da doença (nos estágios iniciais) é difícil porque os primeiros sintomas são semelhantes às alterações cognitivas relacionadas à idade. Porém, por meio da ressonância magnética⁸, é possível analisar quais pacientes com alterações cognitivas moderadas desenvolverão a comorbidade. Ademais, a tomografia, por meio da emissão de pósitrons (PET), tornou-se uma outra forma de diagnosticar a DA, permitindo a visualização das placas amiloides. Por outro lado, estudos começaram a ser desenvolvidos para utilizar marcadores moleculares que podem auxiliar no diagnóstico precoce da doença mesmo que não haja a presença de sintomas clínicos (KANDEL et al., 2000).

3.2 ESTUDO DE CASO

O estudo apresenta um delineamento do tipo caso-controle, em que são selecionados dois grupos de indivíduos: um com a característica de interesse (no caso deste trabalho, diagnosticados com a doença de Alzheimer) e outro sem a característica. Assim, os dados são compostos por tecidos das seguintes regiões do cérebro: Córtex Entorrinal, Hipocampo, Giro Temporal Médio, Córtex Cingulado Posterior, Giro Frontal Superior

⁷Biomoléculas formadas pela ligação de aminoácidos

⁸Segundo Kandel (2000), “revela alterações iniciais no córtex temporal medial e no hipocampo”.

e Córtex Visual Primário (LIANG et al., 2008) que estão ilustradas na Figura 6. Estes tecidos foram retirados de 34 pacientes diagnosticados com a doença e 14 indivíduos saudáveis, coletados em três centros da doença de Alzheimer (ADCs): Arizona, Universidade de Dukey e Universidade de Washington. Além disso, as amostras⁹ finais disponibilizadas pelos autores são aquelas que apresentaram a expressão de GFAP (proteína fibrilar ácida, do inglês *glial fibrillary acidic protein*) maior do que um desvio-padrão da média. Segundo LIANG et al. (2008), a expressão da proteína GFAP foi avaliada para garantir a pureza das células neurais nas amostras.



Figura 6: Regiões do cérebro analisadas no estudo (córtex entorrinal, hipocampo, giro temporal médio, córtex cingulado posterior, giro frontal superior e córtex visual primário). *Fonte:* <https://www.brainfacts.org/3d-brainintro=true>.

Ademais, antes de realizar os agrupamentos, foi necessário reduzir a quantidade de transcritos que seriam analisados a fim de eliminar previamente aqueles que não trariam informações biologicamente relevantes para o entendimento da DA. No caso deste trabalho, considerou-se que um transcrito tinha “pouca informação relevante” quando sua média, na escala original (2^x , sendo x a expressão gênica do transcrito), fosse maior do que 100 considerando conjuntamente os casos e os controles. Logo, considerou-se para as análises subsequentes um número menor observações (amostras de tecido dos indivíduos estudados) e uma quantidade menor de transcritos.

Para se obter o valor da expressão gênica diferenciada que será utilizada para o agrupamento por misturas finitas, calcula-se a mediana dos controles, na escala logarítmica na base 2 (\log_2), e tal medida é calculada conforme a equação (5.1.1) abaixo.

$$\text{Expr}_{dif} = \text{expr}_{casos} - \text{mediana}_{controle} \quad (3.2.1)$$

A composição final dos dados é a apresentada na Tabela 1 abaixo.

⁹Amostra é o tecido coletado de uma das seis regiões de um dos cérebros.

Tabela 1: Número de indivíduos e transcritos que compuseram os dados finais

Região do Cérebro	Casos	Controles	Número de Transcritos
Córtex entorrinal (EC)	10	13	21.857
Hipocampo (HIP)	10	13	18.042
Giro temporal médio (MTG)	16	12	19.891
Cingulado posterior (PC)	9	13	18.297
Giro frontal superior (SFG)	23	11	20.643
Córtex visual primário (VCX)	19	12	18.664

3.2.1 Estudo das Regiões do Cérebro

Como mostrado na Tabela 1, os dados analisados foram extraídos de seis regiões do cérebro de pacientes com a doença de Alzheimer e idosos saudáveis. Cada região, conforme a tabela, apresenta uma composição diferente de casos, controles e quantidade de transcritos analisados. Assim, para cada uma, seguiram-se as seguintes etapas:

1. Agrupamento por misturas finitas de normais para as quantidades totais de transcritos mostradas na Tabela 1, obtenção do modelo inicial e da quantidade de grupos, utilizando o critério do ICL.
2. Identificação dos grupos extremos de transcritos. Para distinguir quais foram os transcritos mais expressos, foi usado como ponto de comparação a mediana dos grupos obtidos por meio do agrupamento realizado na etapa anterior.
 - (a) No agrupamento inicial de cada região, foram avaliadas as distribuições das expressões diferenciadas de cada amostra nos grupos extremos, por meio de *boxplots*, para verificar se todas as amostras apresentavam comportamento semelhante nos valores extremos.
 - (b) Caso alguma amostra apresentasse comportamento oposto às demais, esta seria retirada do estudo e um novo agrupamento seria feito, obtendo novo modelo e nova divisão de grupos. Isso foi feito para evitar que comportamentos muito diferentes nos extremos alterassem o agrupamento dos transcritos.
3. Seleção dos grupos de transcritos com maiores medianas, em valor absoluto. Assim, permaneceram no estudo aqueles grupos que fossem mais extremos independentemente de serem positiva ou negativamente expressos. Como o objetivo do trabalho é identificar as funções gênicas dos transcritos mais expressos e o algoritmo para obtenção destes processos biológicos é limitado para uma menor quantidade de transcritos, esta etapa foi necessária para retirar do estudo transcritos considerados como pouco expressos.

4. Realização de um novo agrupamento por misturas finitas com os grupos selecionados anteriormente, obtendo um novo modelo com diferentes quantidades de grupos.
 - (a) Identificação e seleção dos grupos de transcritos com maior mediana, em valor absoluto.
 - (b) Novo agrupamento usando o modelo de misturas finitas de normais com os grupos selecionados no item anterior.

Esta etapa de novos agrupamentos e seleções de grupos de transcritos foi repetida até que o algoritmo não encontrasse mais divisões entre os transcritos, desde que a quantidade final de transcritos fosse próxima ou inferior a 1.000 por conta da “limitação” do algoritmo para obtenção da função gênica (é mais eficiente para a análise ter menos transcritos para serem avaliados).

5. Obtenção da função gênica (YU; HE, 2016 e CARLSON, 2016) com aqueles grupos de transcritos considerados como altamente expressos.

Córtex Entorrinal

A região do Córtex Entorrinal (Figura 7) está ligada às memórias declarativa, espacial e emocional. Ela auxilia na orientação de “espaço-tempo”, se conecta diretamente ao Hipocampo (região que será analisada na próxima seção) e é responsável por formar a identidade de cada indivíduo (CÓRTEX..., 2020). Como mencionado anteriormente (na seção 3.1), a doença de Alzheimer é caracterizada pela perda da memória recente do indivíduo (SANAR, 2018), ou seja, o paciente começa a ter deficiências nas memórias declarativa e espacial. Estudos mostram que esta região do cérebro é uma das primeiras a ser afetada pela doença porque é nela que a proteína TAU mais se acumula juntamente com os emaranhados neurofibrilares. Portanto, espera-se observar processos biológicos que sejam relacionados ou afetados pela DA.

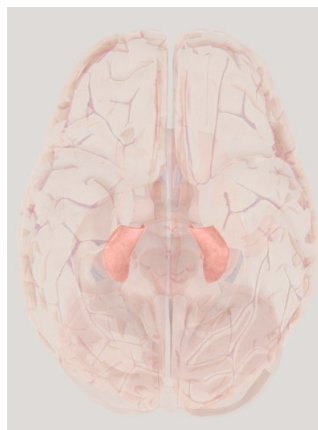


Figura 7: Representação do Córtex Entorrinal. *Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=true&focus=Brain-limbic_system-entorhinal_cortex.*

Nesta região, ao realizar as etapas 1 e 2 citadas anteriormente, observou-se que a expressão diferenciada da amostra 2 apresentou comportamento oposto ao esperado e diferente das outras (Figura 8). Quando foi analisado o grupo cuja mediana era menor que zero, notou-se que a amostra apresentou um comportamento diferente das demais, com expressões gênicas com valores mais próximos ou maiores que zero; já ao observar o grupo com maior mediana positiva, vê-se que, além do comportamento ser diferente dos demais, ele é oposto ao que se espera deste grupo (esperava-se transcritos com valores positivos, na maioria, e esta amostra apresentou grande parte dos transcritos com valores negativos). Com isso, ela foi retirada do estudo e os agrupamentos foram refeitos sem esta observação.

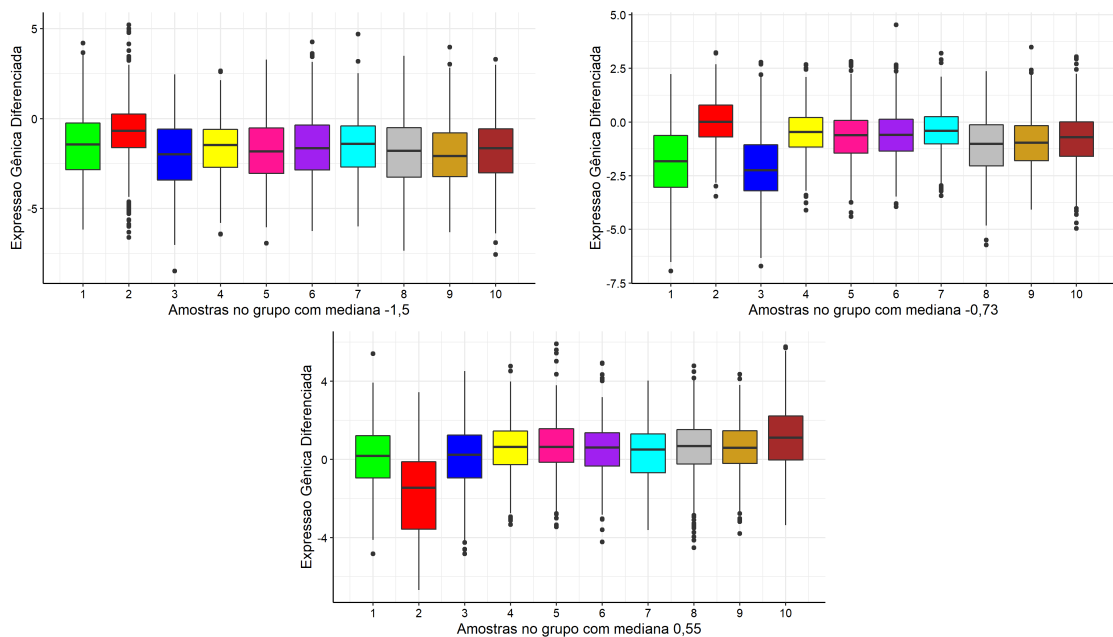


Figura 8: Gráficos correspondentes ao comportamento dos transcritos em cada amostra biológica dentro dos grupos obtidos pelo modelo de misturas finitas.

Os dados da região do Córtex Entorrinal (EC) apresentavam, aproximadamente, 21.000 transcritos, 9 amostras (porque a amostra 2 foi retirada do estudo como mencionado anteriormente) e expressões gênicas estimadas, por amostra, segundo o gráfico seguinte (Figura 9). Nota-se que a maior parte das expressões diferenciadas encontram-se em torno do zero e que, em todas as amostras, existem casos extremos tanto para valores positivos quanto para os negativos. As amostras 3 e 5 apresentaram os valores extremos positivos mais elevados do que as demais, assim como a 3 e a 7 tiveram os valores extremos negativos mais baixos.

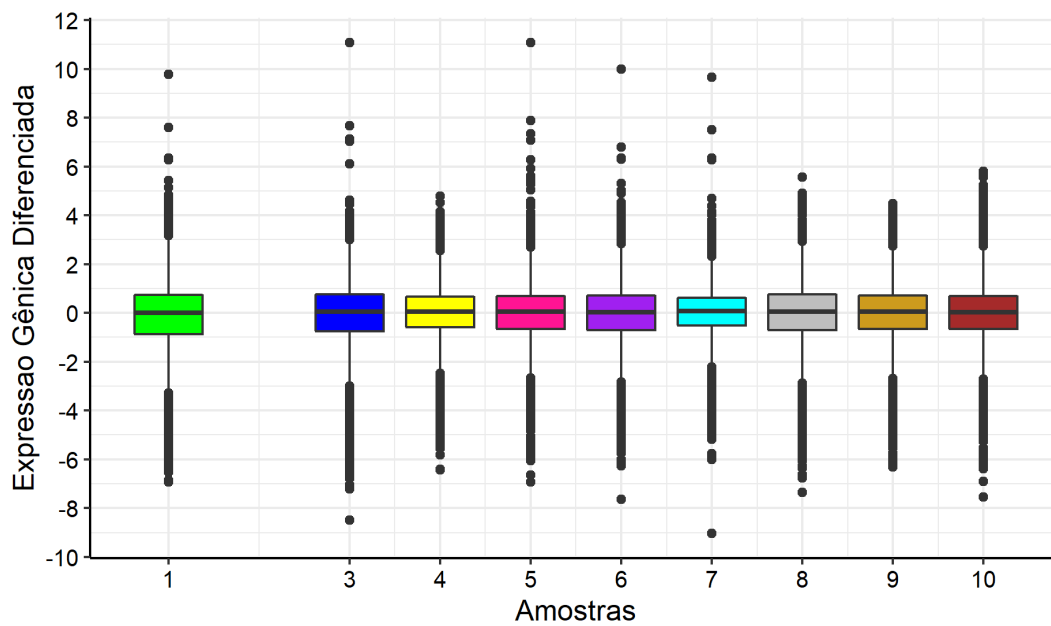


Figura 9: Boxplot de cada amostra biológica, excluindo a amostra 2, com a distribuição das expressões gênicas diferenciadas dos 21.857 transcritos no Córtex Entorrinal.

Após realizar o agrupamento descrito na etapa 2.b apresentada no início desta seção, foi obtido um modelo VVV com 2 componentes para os 21.857 transcritos desta região (Tabela 2). O modelo encontrado é o mais complexo entre todos os 14 possíveis modelos VSO, por apresentar volume, forma e orientação variados entre as componentes, porém foi possível obter poucos grupos, mostrando o *trade-off* que existe entre complexidade do modelo e quantidade de componentes.

Além disso, utilizando o *heatmap* (Figura 10) que é um mapa de calor, é possível ter a ilustração dos valores das expressões gênicas por meio da intensidade das cores. Com isso, pode-se observar que o grupo 1 tem maior quantidade de genes em sua formação e que grande parte deles possui expressão positiva. Isso significa que este grupo é formado, principalmente, por transcritos que estão mais expressos em pacientes com a DA em comparação com idosos saudáveis; e o grupo 2 é formado predominantemente por transcritos sub-expressos em pacientes com a doença.

Tabela 2: Resultados do ajuste do agrupamento inicial da região Córtex Entorrinal após a retirada da amostra 2.

Estatísticas	Agrupamento Inicial Modelo VVV com 2 componentes
Log-verossimilhança	-190.106,90
N	21.857
Graus de liberdade	109
BIC	-381.303,10
ICL	-383.664,60



Figura 10: Heatmap dos 21.857 transcritos da região Córtex Entorrinal resultante do primeiro agrupamento.

A Tabela 3 apresenta o resultado da divisão em grupo dos 21.857 transcritos com as respectivas medianas que foram utilizadas para a escolha do grupo que foi considerado como extremo. Nota-se que o grupo que apresentou maior quantidade de transcritos com expressão negativa foi o que obteve maior mediana (em valor absoluto) e este foi o selecionado para prosseguir aos próximos agrupamentos.

Tabela 3: Divisão em grupos dos 21.857 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	17.702	0,15
2	4.155	-0,36

Depois de repetir a etapa 4 três vezes, foram obtidos 501 transcritos considerados como mais expressos. Observa-se, na Figura 11, que a maioria dos transcritos selecionados apresentam expressão diferenciada abaixo de zero, mostrando que foram selecionados, em grande parte, aqueles que são menos expressos em pacientes com a DA em comparação com os saudáveis. Vê-se também que não há mais tantos valores extremos dentro de cada amostra; apenas as amostras 4, 6 e 10 que apresentaram alguns *outliers*.

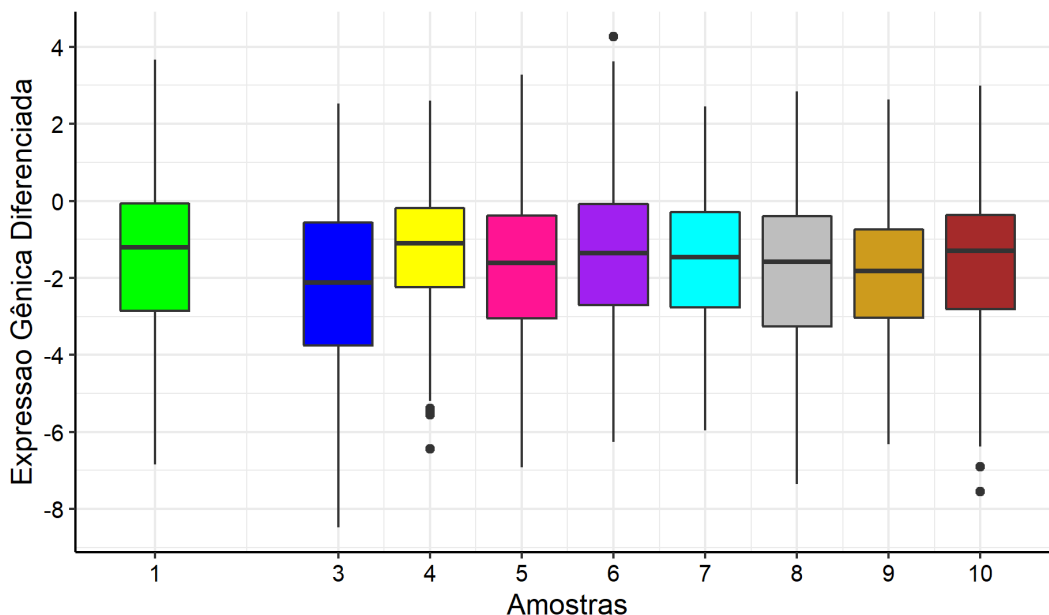


Figura 11: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 501 transcritos no Córtex Entorrinal.

O modelo final obtido após realizar todas as etapas descritas no início desta seção apresentou volume variado, forma e orientação iguais (modelo VEE) e foi composto por 2 componentes (Tabela 4). Este modelo, comparado ao primeiro apresentado na Tabela 2, teve redução no valor das estatísticas como BIC e ICL e também é considerado um modelo mais simples por ter forma e orientação iguais nas 2 componentes do agrupamento. Assim, ao reduzir a dimensão dos dados, também foi reduzida a complexidade do modelo.

No *heatmap* a seguir (Figura 12), nota-se que, ao contrário do observado na Figura 10, o maior grupo de transcritos (472, conforme a Tabela 5) é formado por aqueles com expressão negativa (sub-expressos nos pacientes com a doença de Alzheimer), mas ainda foram selecionados 29 que são mais expressos nos pacientes. Ademais, destaca-se que a amostra 9 (penúltima linha horizontal do gráfico) apresentou resultado diferente das demais porque, no grupo com transcritos de expressão positiva, ela apresentou transcritos negativos e mais próximos do zero do que as outras.

Tabela 4: Resultados do ajuste do agrupamento final da região Córtex Entorrinal.

Estatísticas	Agrupamento Final
	Modelo VEE com 2 componentes
Log-verossimilhança	-7.888,54
N	501
Graus de liberdade	65
BIC	-16.181,15
ICL	-16.182,53

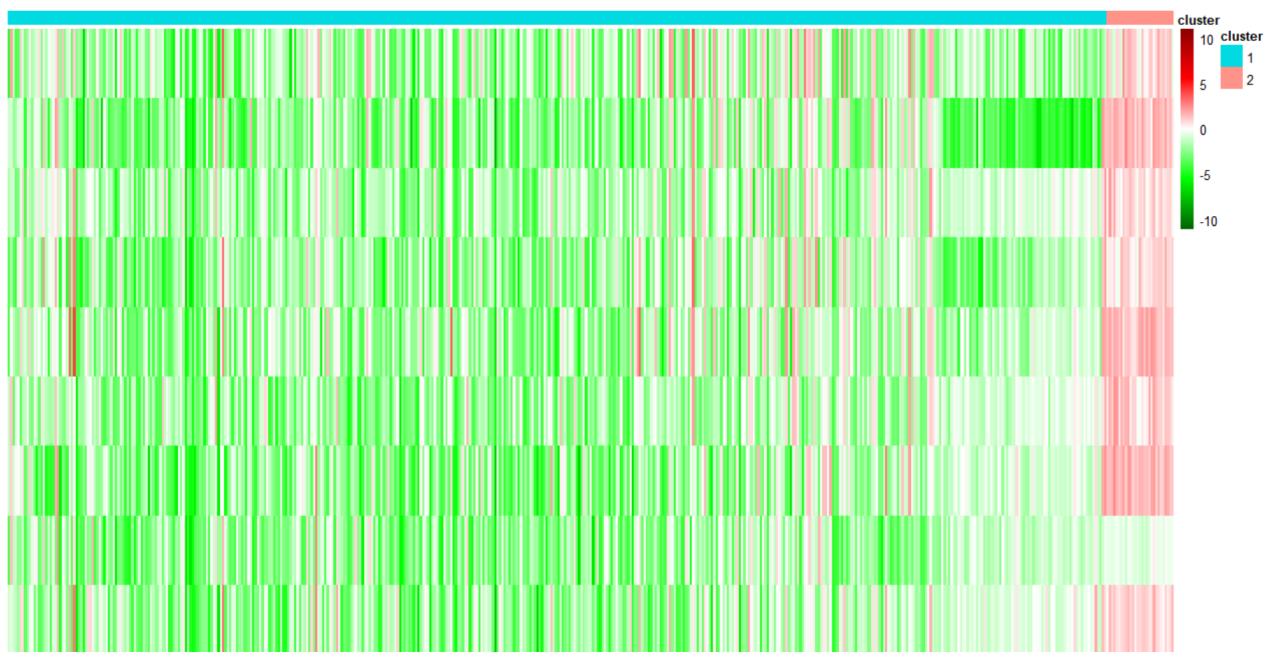


Figura 12: Heatmap dos 501 transcritos da região Córtex Entorrinal resultante do último agrupamento.

No último agrupamento, ilustrado nos gráficos acima, foi possível obter dois grupos de transcritos com medianas mais elevadas, em valor absoluto (Tabela 5). Como esperado pelo *heatmap* da Figura 12, o grupo 1 teve mediana negativa de -1,53, o que significa que 50% dos transcritos têm expressão gênica diferenciada com valores mais baixos (“mais negativos”) do que este, e o grupo 2 obteve mediana de 0,97, aproximadamente, indicando que metade dos transcritos tem expressões maiores que este valor. Dessa forma, nota-se que o objetivo de obter os transcritos com expressões mais fortes, independentemente do “sinal” da expressão (seja negativa, seja positiva) foi atingido.

Tabela 5: Divisão em grupos dos 501 transcritos, obtidos pelo modelo final, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	472	-1,53
2	29	0,97

Com esses transcritos selecionados, aplicou-se a técnica ORA para verificar quais foram as funções gênicas super-representadas em cada grupo deste conjunto final. A Figura 13 a seguir mostra as funções encontradas em cada grupo considerado como negativo e positivo (grupos 1 e 2 da Figura 12 e da Tabela 5, respectivamente). Nota-se que estas funções, apesar de serem apresentadas como significativas por apresentarem P-valor ajustado menor que 0,05, possuem um P-valor muito próximo à região de rejeição. Os três primeiros processos biológicos foram os que, dentro deste conjunto, apresentaram valores mais distantes da região de rejeição e, portanto, trazem maior segurança na decisão.

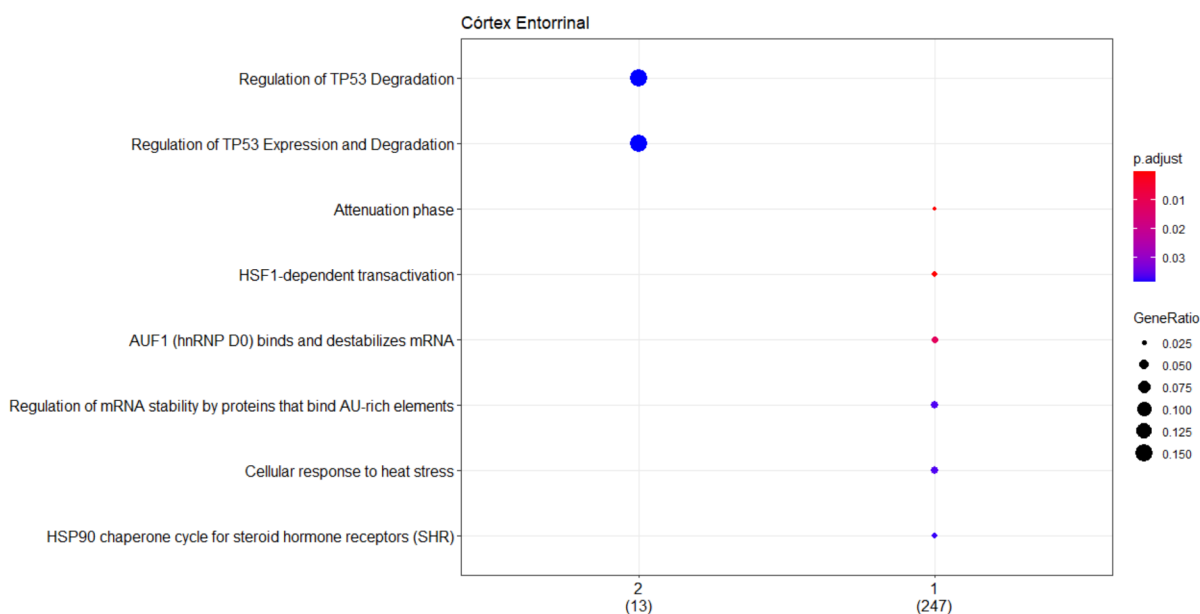


Figura 13: Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Córtex Entorrinal representado na Figura 12.

Por meio do seguinte gráfico (Figura 14), é possível ver, em formato de rede, como são as ligações de genes e os processos biológicos. As menores bolas, sem marcação, representam os genes e as maiores são as funções especificadas no gráfico. Nota-se que alguns genes estão associados a mais de um processo: no grupo 1, vê-se que 3 genes estão ligados a todas as funções encontradas e que existem 4 funções/processos que têm vários genes em comum. Já no grupo 2, observa-se que apenas 2 genes encontrados nos dados estão envolvidos com os processos “super-representados” neste grupo e que ambos estão ligados com eles.



Figura 14: Processos biológicos da Figura 13 presentes no Córtex Entorrinal com os respectivos genes relacionados.

Hipocampo

O Hipocampo (Figura 15) é uma região do cérebro presente no lobo temporal e está relacionado com audição, aprendizado, memória e emoção (KANDEL et al., 2000). Assim, ela está diretamente relacionada com as principais características observadas em pacientes com Alzheimer (como, segundo SANAR, 2018, perda de memória recente, disfunções de linguagem, redução da velocidade de pensamento e alterações de comportamento, humor e personalidade) e levanta-se a hipótese de que podem ser encontrados genes com processos biológicos bem definidos e relacionados à doença.



Figura 15: Representação do Hipocampo. Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=true?focus=Brain-limbic_system-hippocampus-hippocampus.

Inicialmente, foram analisados 18.042 transcritos coletados em 10 pacientes com DA. O gráfico da Figura 16 mostra a distribuição das expressões gênicas diferenciadas,

obtidas por meio da Equação 5.1.1, nestas amostras. Nota-se, por meio dele, que a maioria concentra-se em torno de valores próximos a zero e que foram observados muitos valores extremos, principalmente valores positivos.

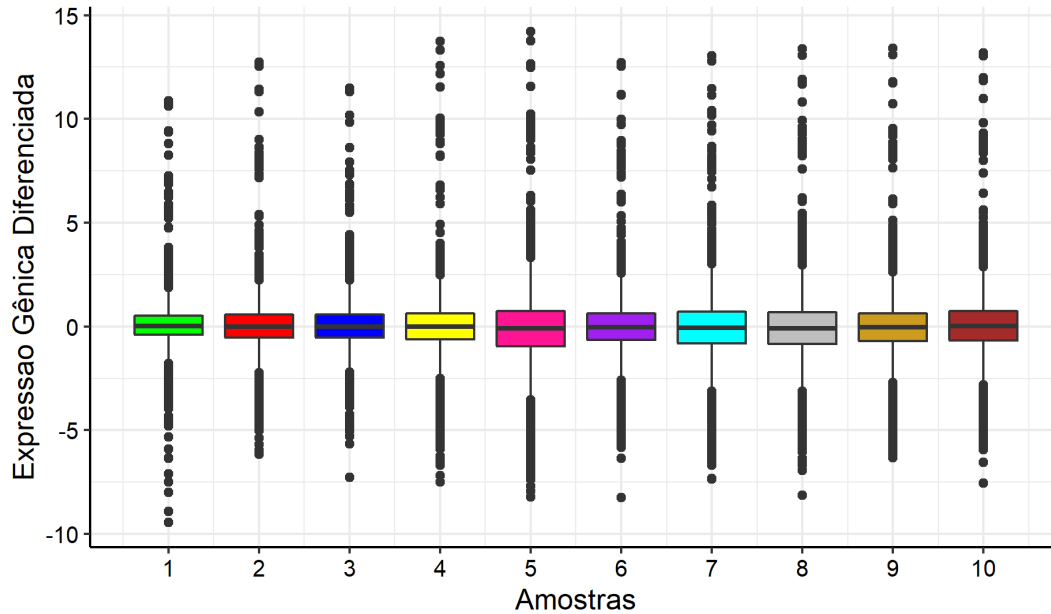


Figura 16: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 18.042 transcritos no Hipocampo.

Sob esses transcritos, um modelo de misturas finitas de normais foi ajustado com o objetivo de identificar grupos de transcritos mais expressos (etapa 1). Dessa forma, os 18.042 transcritos foram divididos, inicialmente, em 5 grupos seguindo um modelo VVV, em que todas as componentes (volume, forma e orientação) são variadas. As estatísticas obtidas por meio deste primeiro ajuste do modelo podem ser observadas na Tabela 6.

Analisando-se a Figura 17, vê-se que eles foram divididos em 5 grupos e os maiores foram os grupos 1 e 3, respectivamente. O grupo 3 apresentou, em sua maioria, transcritos com expressão diferenciada positiva, ou seja, são aqueles transcritos com maior expressão em pacientes com a doença em relação à mediana da expressão de idosos saudáveis. O grupo 2, por sua vez, possui menos transcritos que o mencionado anteriormente e eles possuem expressão diferenciada negativa, indicando baixa atividade destes no grupo de pacientes.

Tabela 6: Resultados do ajuste do agrupamento inicial da região Hipocampo.

Estatísticas	Agrupamento Inicial Modelo VVV com 5 componentes
Log-verossimilhança	-190.410,60
N	18.042
Graus de liberdade	329
BIC	-384.045,60
ICL	-393.211,90

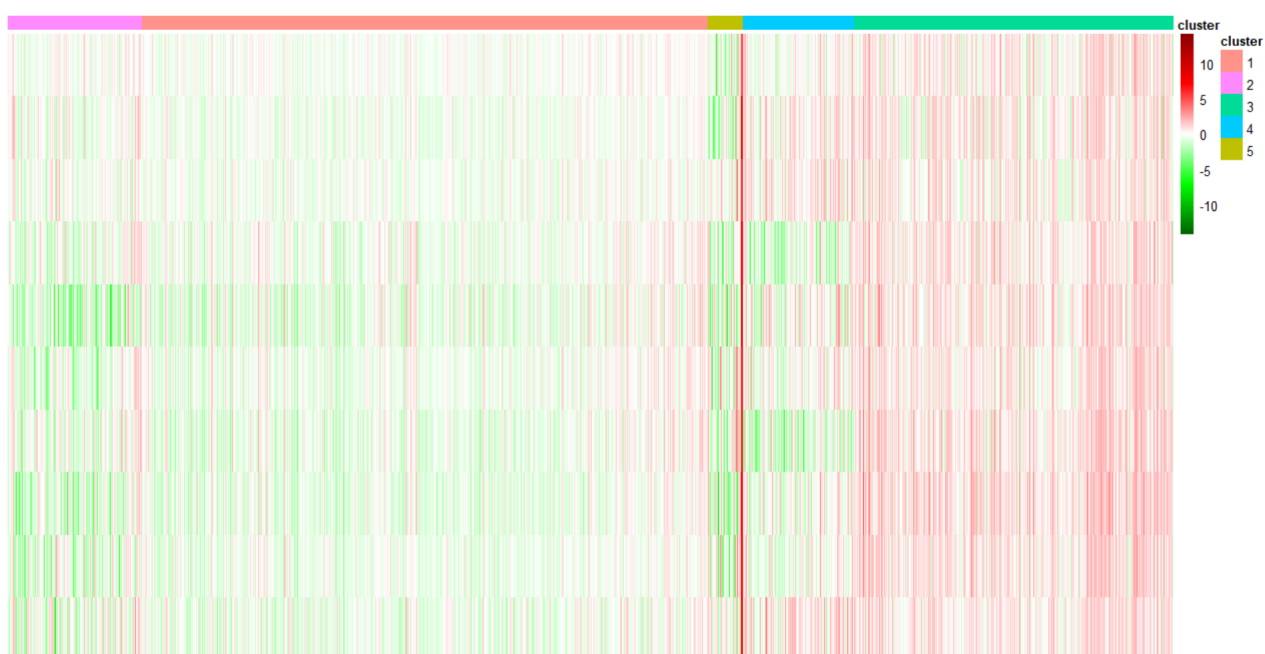


Figura 17: Heatmap dos 18.042 transcritos da região Hipocampo resultante do primeiro agrupamento.

No agrupamento ilustrado no gráfico acima, foram obtidos 5 grupos e a respectiva mediana de cada um está descrita na Tabela 7. Nota-se que o grupo com maior mediana positiva foi o grupo 3 que possui a segunda maior quantidade de transcritos, e o grupo com menor mediana (mediana negativa mais distante do zero) é o grupo 2 que possui 2.069 transcritos em sua formação. Portanto, estes foram os grupos selecionados para as próximas etapas dos agrupamentos.

Tabela 7: Divisão em grupos dos 18.042 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	8.762	-0,22
2	2.069	-0,32
3	4.947	0,66
4	1.729	-0,000075
5	535	-0,17

Ao aplicar os procedimentos semelhantes aos da região anterior e realizar a etapa 4 apenas uma vez, o algoritmo “estabilizou”, ou seja, não encontrou mais divisões entre os 7.016 transcritos restantes. Como esta quantidade de transcritos é elevada, as etapas descritas no “Estudo das Regiões do Cérebro” foram aplicadas em cada um dos dois grupos obtidos neste agrupamento mostrado na Tabela 8 com o intuito de encontrar grupos de transcritos mais extremos e, assim, poder obter por volta de 1.000 transcritos para analisar as respectivas funções.

Tabela 8: Resultados do ajuste do agrupamento intermediário da região Hipocampo.

Estatísticas	Agrupamento Inicial Modelo VVV com 2 componentes
Log-verossimilhança	-81.394,34
N	7.016
Graus de liberdade	131
BIC	-163.948,80
ICL	-164.820,30

Tabela 9: Divisão em grupos dos 7.016 transcritos, obtidos pelo modelo anterior, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	4.867	-0,67
2	2.149	-0,32

Após fazer os agrupamentos separados para cada um dos grupos apresentados acima, foram obtidos 1.063 transcritos que têm suas distribuições representadas na Figura 18 a seguir. Percebe-se que a maioria dos valores extremos da Figura 16 não aparecem mais em destaque; as observações com valores muito elevados foram removidas, pelo critério de exclusão dos grupos pela mediana. Além disso, destaca-se que, apesar de ainda ser observado que as expressões estão em torno de zero, algumas amostras têm, em

sua maioria, transcritos com expressões negativas, principalmente, a amostra 5 que tem mais de 75% de expressões negativas.

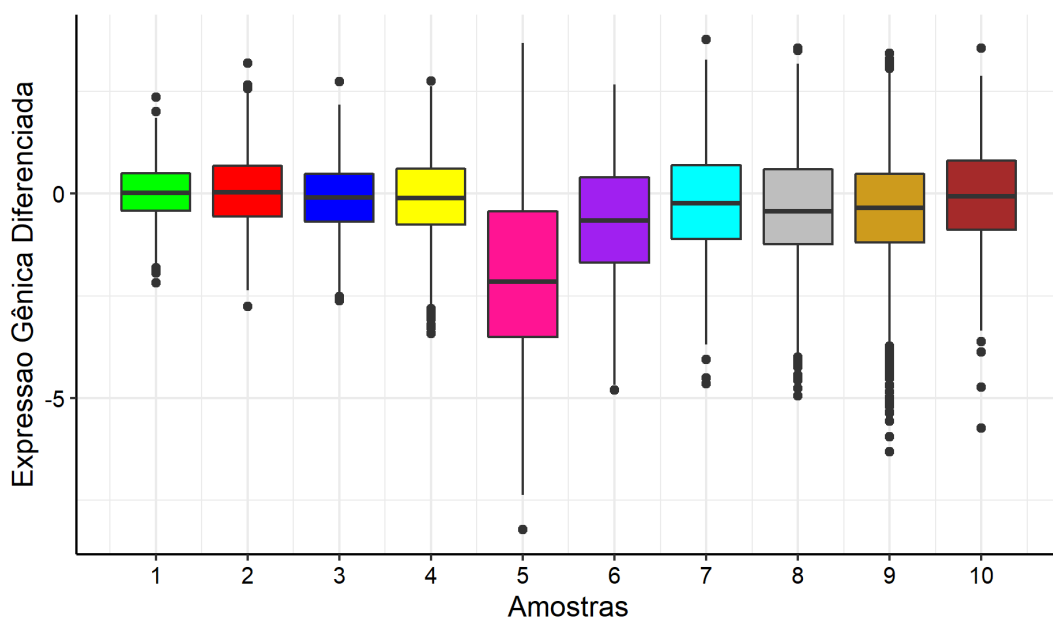


Figura 18: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 1.063 transcritos no Hipocampo.

Estes transcritos mencionados anteriormente, foram divididos em três grupos em que dois deles são formados, em sua maioria, por transcritos com expressão negativa, e um por transcritos com expressão positiva. Os grupos 1 e 2 foram obtidos por meio dos agrupamentos realizados no grupo 2 da Tabela 9 e têm seus resultados apresentados na Tabela 10, e o grupo 3 foi resultado dos agrupamentos realizados no grupo 1 da Tabela 9 e têm as respectivas estatísticas mostradas também na Tabela 10. Nota-se que ambos os modelos possuem volume, forma e orientação iguais entre as componentes, sendo a forma mais simples de modelagem de misturas finitas de normais.

O *heatmap* a seguir (Figura 19) mostra, por meio da intensidade das cores, o valor das expressões gênicas. É possível observar que a amostra 5 foi a que teve transcritos mais sub-expressos (com forte expressão diferenciada negativa, ilustrada no gráfico em verde) em relação às outras amostras. Além disso, nota-se que o grupo 3, apesar de não ser muito grande, foi formado por transcritos mais expressos em pacientes com a doença e com expressões relativamente fortes.

Tabela 10: Resultados do ajuste dos agrupamento finais da região Hipocampo.

Estatísticas	Agrupamento Final Negativo	Agrupamento Final Positivo
	Modelo EEE com 2 componentes	Modelo EEE com 1 componente
Log-verossimilhança	-11.622,21	-247,06
N	941	122
Graus de liberdade	76	65
BIC	-23.764,78	-806,38
ICL	-23.765,70	-806,38

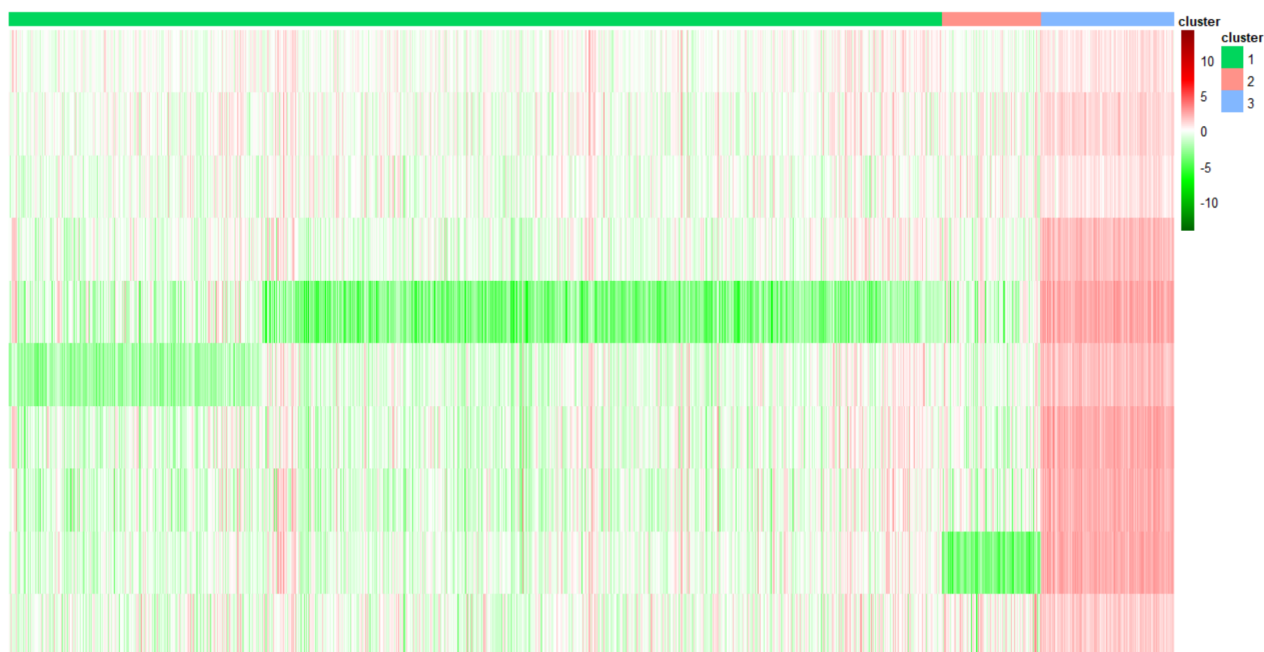


Figura 19: Heatmap dos 1.063 transcritos da região Hipocampo resultante do último agrupamento.

Na Tabela 11, verifica-se a formação final dos grupos apresentados acima. É possível notar que, comparado com o resultado apresentado na Tabela 7, o objetivo de selecionar transcritos mais extremos por meio da mediana dos grupos foi atingido devido à mediana positiva final ser 1,76 que é maior do que a obtida anteriormente. Já nos transcritos do grupo com mediana negativa, observa-se que, no último agrupamento, foram divididos em 2 grupos e que as medianas permaneceram, aproximadamente, iguais à do grupo 2 do primeiro modelo.

Tabela 11: Divisão em grupos dos 1.063 transcritos, obtidos pelo modelo final, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	851	-0,38
2	90	-0,33
3	122	1,76

Dessa forma, após serem selecionados estes transcritos e aplicando a técnica ORA, foram encontradas as funções gênicas apresentadas na Figura 20 abaixo. O gráfico mostra as 20 funções mais “super-representadas” em cada grupo, totalizando em 40 funções presentes nesta região. É possível ver que o grupo 2 não está representado no gráfico; isto porque, pelo teste aplicado, não foi identificado nenhum processo biológico “super-representado” nos genes deste grupo. Além disso, nota-se que foram encontrados mais processos biológicos super-representados do que apresentado na região anterior (Córtex Entorrinal). Os p-valores dos testes foram, na maioria, abaixo de 0,01, o que já mostra uma forte indicação de “super-representação” dessas funções, principalmente no grupo 3.

Destaca-se, entre as funções apresentadas abaixo, que duas delas também foram encontradas no Córtex Entorrinal. São elas: *AUF1 (hnRNP D0) binds and destabilizes mRNA* (AUF1 (hnRNP D0) se liga e desestabiliza mRNA) e *Regulation of mRNA stability by proteins that bind AU-rich elements* (regulação da estabilidade do mRNA por proteínas que se ligam a elementos ricos em AU) que estão entre os processos super-representados pelos transcritos do grupo 1 que é aquele formado por transcritos sub-expressos em pacientes com a DA.

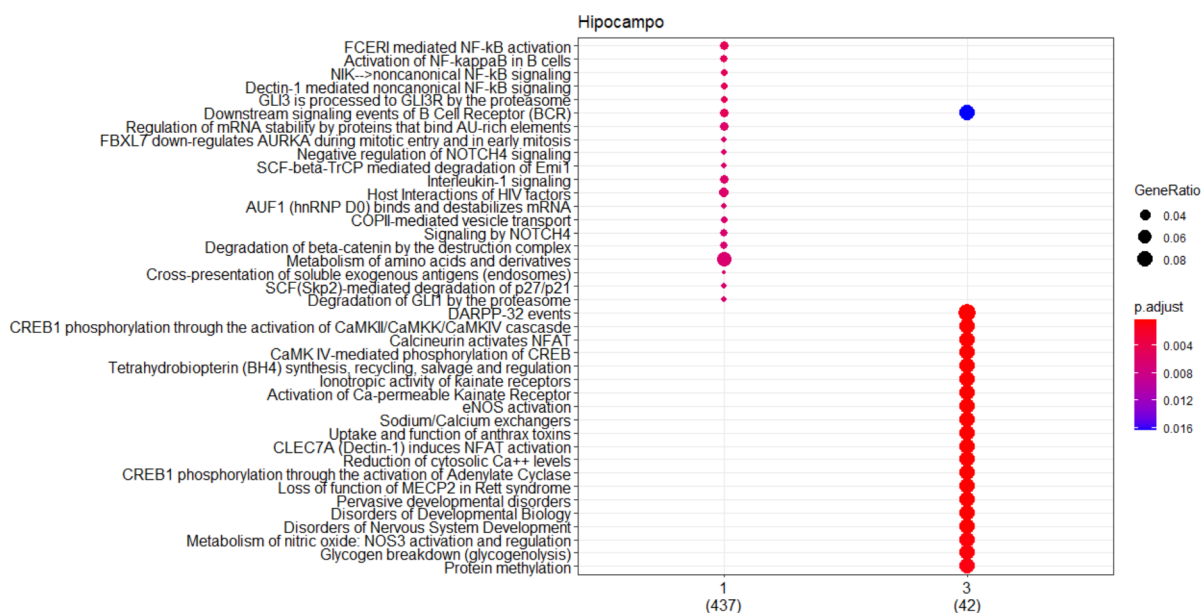


Figura 20: Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Hipocampo representado na Figura 19.

No gráfico a seguir (Figura 21), é possível notar muitos genes associados a cada função e parte deles está ligado a apenas uma delas. Os processos como *metabolism of amino acids and derivatives* (metabolismo de aminoácidos e derivados) e *COPII-mediated vesicle transport* (transporte de vesículas mediado por COPII) se destacam por terem muitos genes particulares em relação aos demais.



Figura 21: Processos biológicos da Figura 20 presentes no Hipocampo com os respectivos genes relacionados.

Giro Temporal Médio

O Giro Temporal Médio, localizado no lobo temporal, se encontra entre os sulcos temporais inferior e superior e sua principal função está associada às funções visuais secundárias, como o reconhecimento facial.

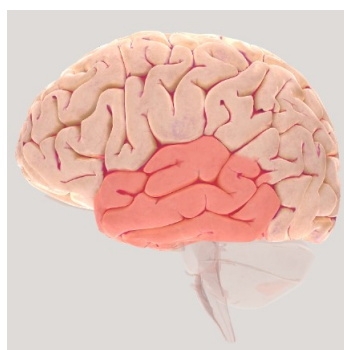


Figura 22: Representação do lobo temporal em que encontra-se, no centro, o Giro Temporal Médio.

Fonte: Society for Neuroscience (2017). Disponível em:

https://www.brainfacts.org/3d-brainintro=true&focus=Brain-cerebral_hemisphere-temporal_lobe

À princípio, os dados desta região eram compostos por 19.891 transcritos em 16 amostras de pacientes com a doença de Alzheimer. Nota-se, na Figura 23, que existem transcritos com valores extremos positivos e negativos em maior quantidade do que foi observado nas regiões anteriores (Figuras 9 e 16). Além disso, o gráfico abaixo mostra

que as medianas das amostras estão, aparentemente, bem próximas de zero e que metade das expressões estão em torno desse valor.

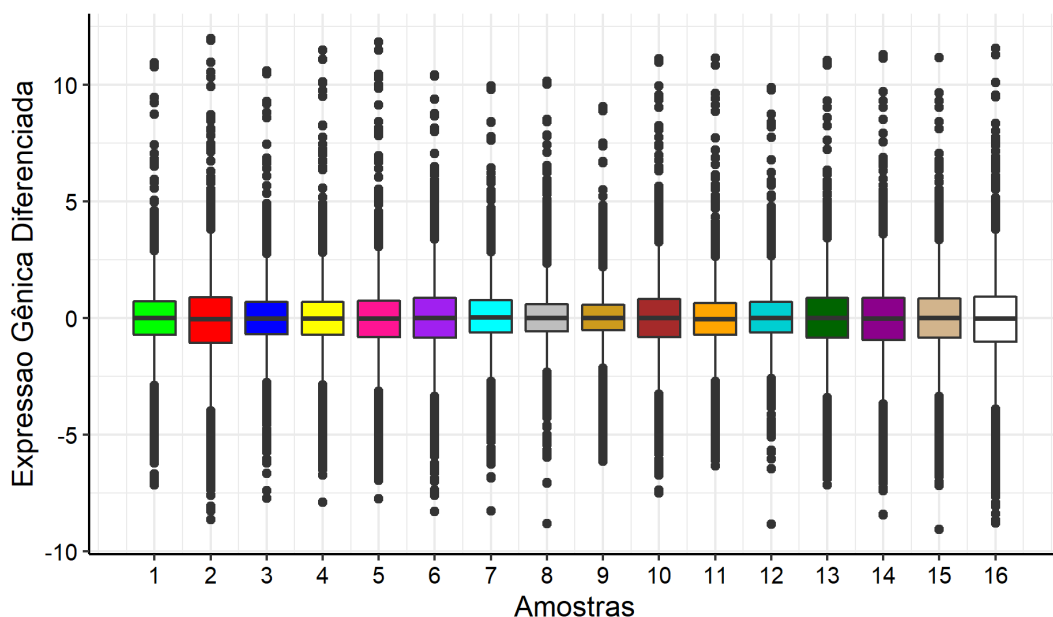


Figura 23: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 19.891 transcritos no Giro Temporal Médio.

O modelo inicial, obtido após aplicar a etapa 1 descrita no início da seção, foi formado por 7 componentes nas quais o volume, a forma e a orientação eram variados (modelo VVV). Nota-se que esta região foi a que apresentou maior quantidade de grupos nesta etapa inicial em relação às duas regiões anteriormente analisadas. Além disso, as estatísticas desta região (Tabela 12) foram as que apresentaram maiores resultados (log-verossimilhança, BIC e ICL em valor absoluto) comparadas com as do Córtex Entorrinal e do Hipocampo.

Pelo *heatmap* a seguir (Figura 24), pode-se ver que muitos transcritos apresentaram valores próximos de -5 e 5 que são aqueles com cores verde e vermelho claros, respectivamente. Os grupos que tiveram expressão mediana mais forte negativa e positivamente foram, nessa ordem, o 1 e 5, conforme indicado no gráfico. Ademais, destaca-se o grupo 4 pois, apesar de apresentar, em sua maioria, transcritos com expressão negativa (ou seja, mais expressos em pacientes com a DA), também observou-se alguns com a cor vermelha, representando uma expressão mais forte em idosos saudáveis.

Tabela 12: Resultados do ajuste dos agrupamentos inicial e finais da região Giro Temporal Médio.

Estatísticas	Agrupamento Inicial
	Modelo VVV com 7 componentes
Log-verossimilhança	-362.415,50
N	19.891
Graus de liberdade	1.070
BIC	-735.421,90
ICL	-747.342,50

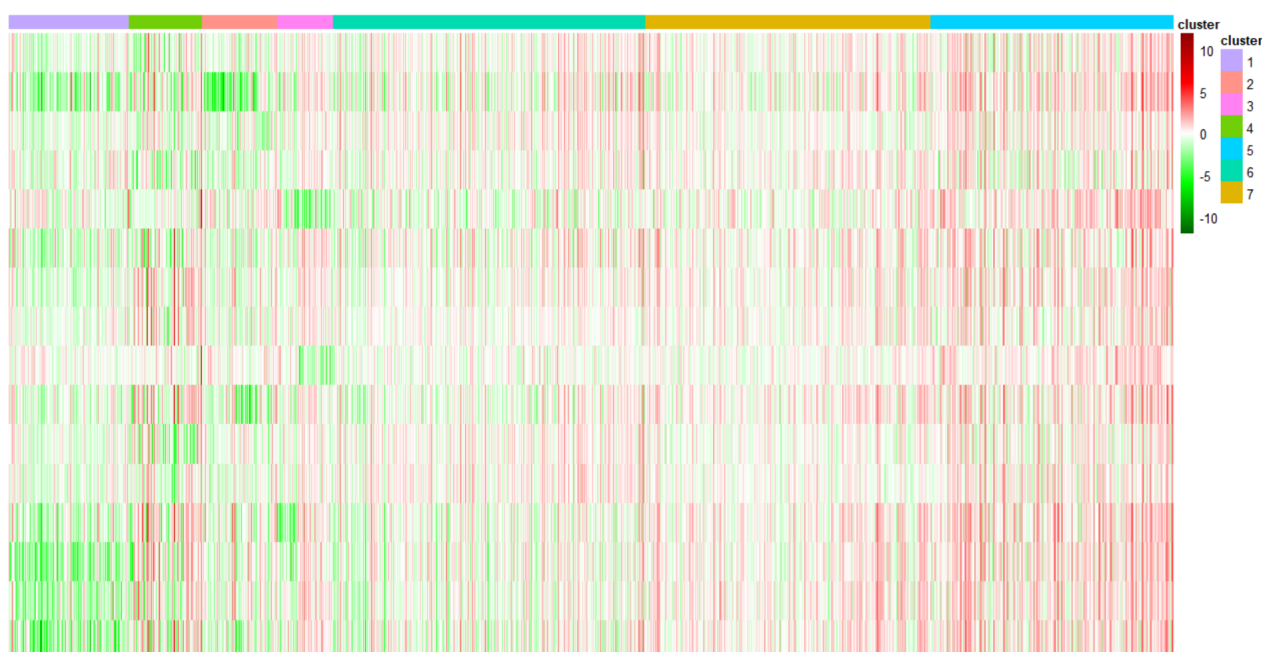


Figura 24: Heatmap dos 19.891 transcritos da região Giro Temporal Médio resultante do primeiro agrupamento.

A seguir, é possível observar os valores obtidos para as medianas de cada um dos grupos formados no modelo representado pela Figura 24. Os grupos 1 e 4 apresentaram os menores valores, ou seja, são as medianas negativas mais distantes do zero, e, por outro lado, o grupo 5 foi o que obteve maior mediana positiva. Portanto, como estes foram os grupos que tiveram as maiores medianas, em valor absoluto, os transcritos selecionados para os demais passos das análises foram os que compuseram estes.

Tabela 13: Divisão em grupos dos 19.891 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	2.052	-0,51
2	1.277	-0,24
3	949	-0,15
4	1.250	-0,37
5	4.150	0,42
6	5.340	-0,05
7	4.873	-0,10

Assim como no Hipocampo, para tentar reduzir mais a quantidade de transcritos para a análise das funções gênicas, foram realizados dois agrupamentos separados. O primeiro para aqueles grupos de transcritos com expressão gênica mediana mais baixa (grupos 1 e 3) e o segundo para o grupo 2 que tem expressão mediana mais alta, pois o algoritmo estabilizou após duas repetições da etapa 4 e ainda restaram muitos transcritos (6.915). O resultado deste agrupamento intermediário pode ser visualizado na Tabela 14 abaixo, assim como os valores das medianas de cada grupo (Tabela 15).

Tabela 14: Resultados do ajuste do agrupamento intermediário da região Giro Temporal Médio.

Estatísticas	Agrupamento Inicial Modelo VVV com 3 componentes
Log-verossimilhança	-146.682,40
N	6.915
Graus de liberdade	458
BIC	-297.414,20
ICL	-298.429

Tabela 15: Divisão em grupos dos 6.915 transcritos, obtidos pelo modelo anterior, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	2081	-0,49
2	3967	0,43
3	867	-0,61

Ao final dos procedimentos de reagrupamento e seleção de grupos com expressões medianas mais elevadas em valor absoluto, restaram 4.646 transcritos nas 16 amostras,

conforme ilustrado na Figura 25. Observa-se que ainda sobraram vários transcritos com valores extremos, principalmente negativos, como os observados nas amostras 2, 13, 14, 15 e 16. Além disso, pode-se destacar o aumento na quantidade de transcritos com expressões diferenciadas positivas: vê-se que as amostras 2, 5, 9, 13, 14, 15 e 16 tem mediana maior que zero e mais de 50% das expressões são positivas (isso pode ser observado pela posição em que a 'caixa' do gráfico se encontra).

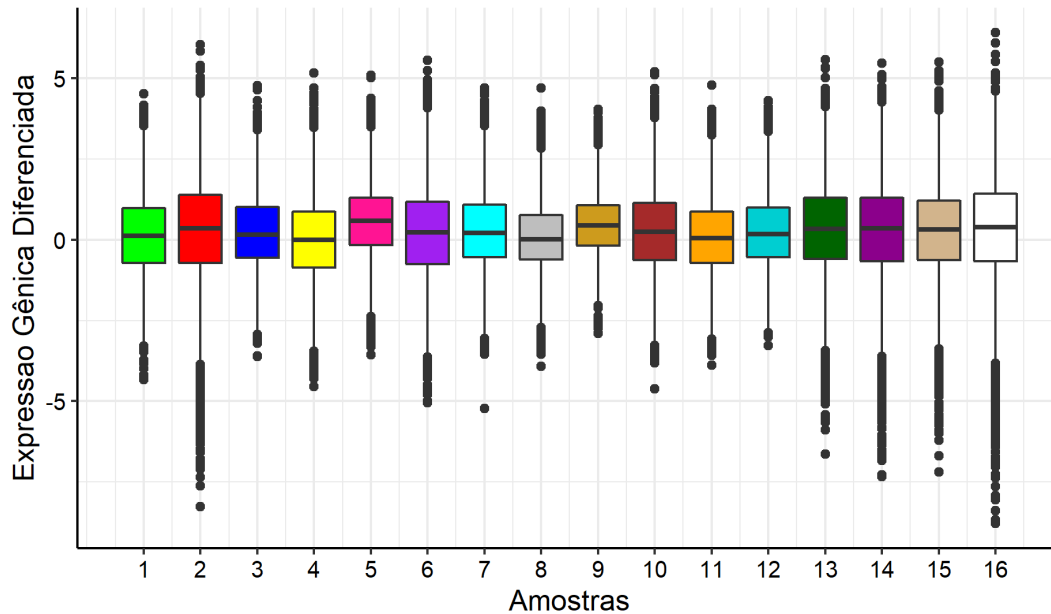


Figura 25: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 4.646 transcritos no Giro Temporal Médio.

Ademais, foram obtidos dois modelos EEE dentro de cada grupo extremo. Ou seja, o algoritmo de misturas de normais, utilizando o critério do ICL e após ser aplicado dentro dos grupos negativos e positivo separadamente, resultou na forma mais simples de modelo VSO que é com todas as componentes iguais e apenas em 1 grupo. Pela Figura 26, nota-se que foram selecionados muitos transcritos com expressão gênica mais forte em pacientes com a doença em comparação com a expressão mediana de idosos saudáveis (transcritos com expressão positiva, representada pela cor vermelha). É possível destacar também que algumas amostras apresentaram valores "opostos" para determinados transcritos: por exemplo, na amostra 3 vê-se que alguns transcritos que pertencem ao grupo 2 (que seria aquele que engloba transcritos de expressão positiva) possuem expressão negativa. Já nas amostras 4 e 8, isso ocorre de maneira contrária.

Tabela 16: Resultados do ajuste dos agrupamentos finais da região Giro Temporal Médio.

Estatísticas	Agrupamento Final Negativo	Agrupamento Final Positivo
	Modelo EEE com 1 componente	Modelo EEE com 1 componente
Log-verossimilhança	-12.685,55	-74.699,68
N	679	3.967
Graus de liberdade	152	152
BIC	-26.362.23	-26.362.23
ICL	-150.658,80	-150.658,80

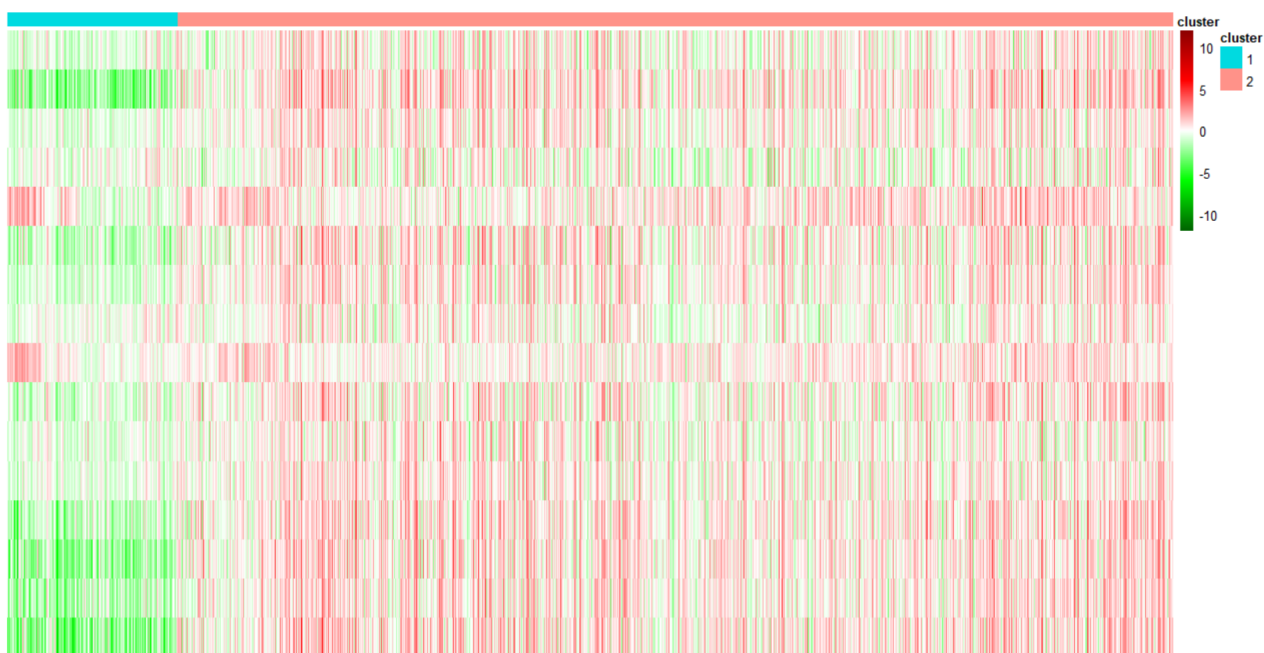


Figura 26: Heatmap dos 4.646 transcritos da região Giro Temporal Médio resultante do último agrupamento.

Os grupos finais apresentados acima tiveram medianas iguais a -0,79 (grupo 1) e 0,43 (grupo 2), como mostrado na Tabela 17 abaixo. Em comparação com as medianas apresentadas na Tabela 13, vê-se que as medianas, em valor absoluto, aumentaram, principalmente considerando a mediana negativa que antes era -0,51 e, após os devidos cortes, foi obtido um grupo com expressão mediana igual a -0,79, atingindo o objetivo encontrar transcritos extremos por meio da mediana dos grupos que foram alocados.

Tabela 17: Divisão em grupos dos 4.646 transcritos, obtidos pelo modelo final, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	679	-0,79
2	3.967	0,43

Com esses transcritos selecionados e após encontrar os genes correspondentes, foram achadas as funções gênicas mostradas no gráfico abaixo (Figura 27). Percebe-se que a maioria das funções resultou em um p-valor inferior a 1% indicando que essas são super-representadas, considerando um corte de 5% para os p-valores. Além disso, é possível notar que essas funções estão associadas a processos de neurotransmissão e que elas foram encontradas em genes originados por meio do grupo 1 que é o grupo de transcritos que apresentou mediana negativa, ou seja, menos expressos em pacientes com a DA.

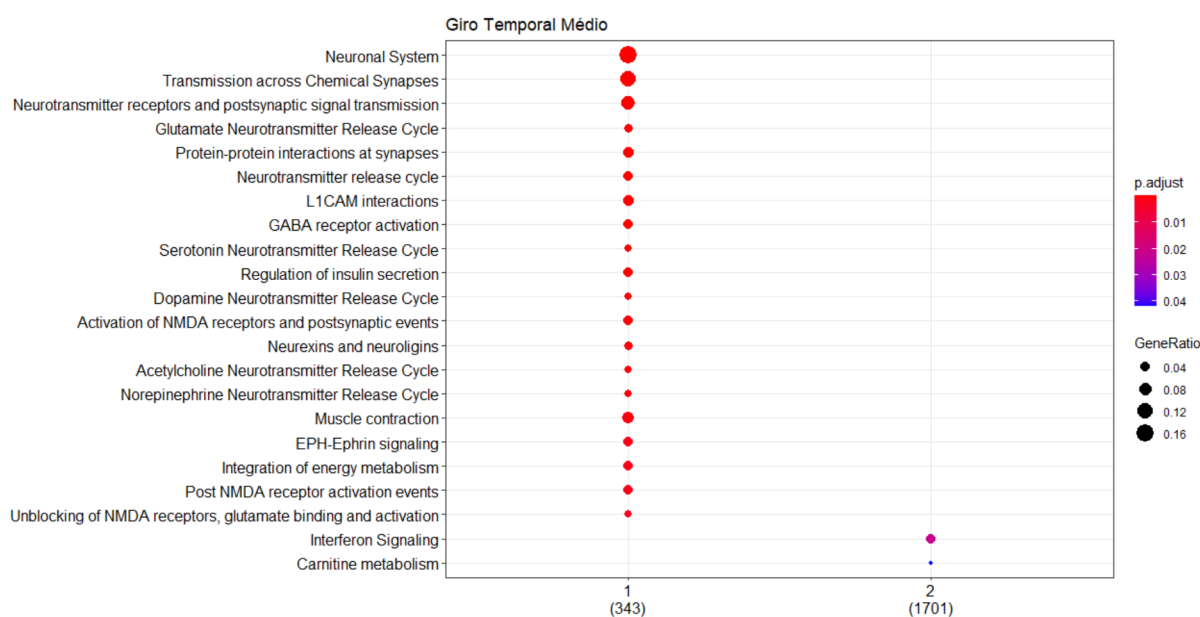


Figura 27: Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Giro Temporal Médio representado na Figura 26.

Com auxílio da Figura 28 abaixo, é possível ver a relação entre os processos biológicos citados no gráfico acima (Figura 27) por meio dos genes que possuem em comum. Nota-se que a maioria das funções associadas à neurotransmissão tem os mesmos genes responsáveis, exceto pela função *neurotransmitter receptors and postsynaptic signal transmission* (receptores de neurotransmissores e transmissão de sinal pós-sináptico).



Figura 28: Processos biológicos da Figura 27 presentes no Giro Temporal Médio com os respectivos genes relacionados.

Córtex Cingulado Posterior

O Córtex Cingulado Posterior está localizado no Giro Cingulado. Este giro é uma parte do cérebro que está relacionado, principalmente, com a expressão corporal e emocional por meio de movimentos, gestos e posturas. Além disso, essa região tem sido associada a vários distúrbios, como a doença de Alzheimer possivelmente pela sua relação com o estado emocional. Especificamente, o Cingulado Posterior está associado à memória topocinética que é responsável pela noção de espaço, armazenamento de movimentos e posições corporais (GIRO..., 2020).



Figura 29: Representação do Córtex Cingulado Posterior. Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=true?focus=Brain-cerebral_hemisphere-temporal_lobe-cingulate_cortex

No gráfico da Figura 30, são observadas as distribuições das expressões gênicas

dos transcritos em cada uma das 9 amostras do Córtex Cingulado Posterior. Percebe-se que, assim como as outras regiões, a mediana das amostras está em torno de zero e existem vários valores extremos. Por outro lado, pode-se notar que as expressões negativas atingem um valor máximo menor (por volta de -6) do que as positivas que ultrapassam de 10.

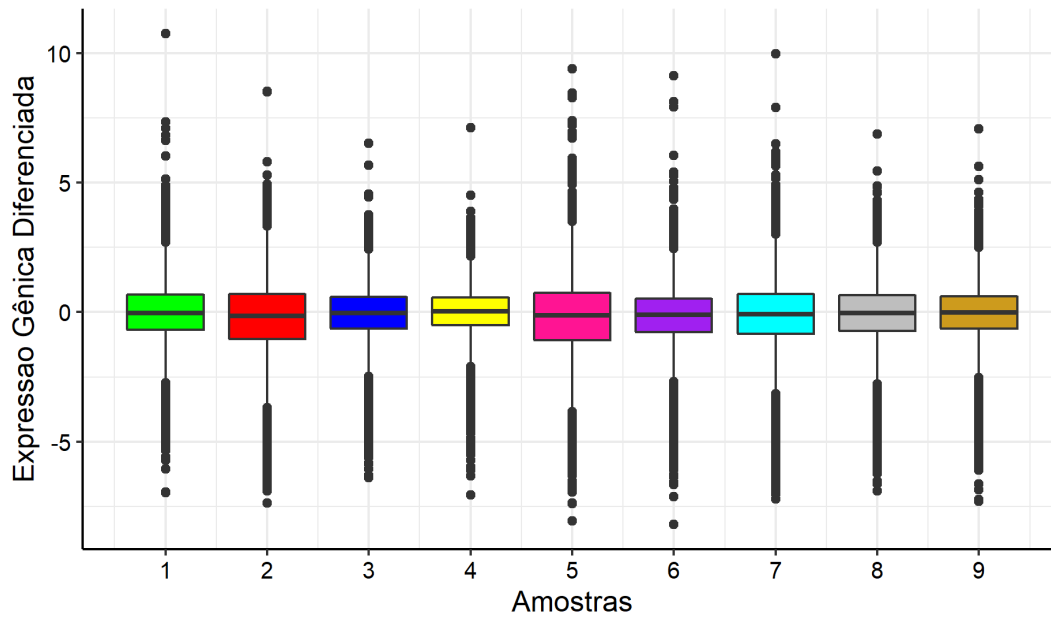


Figura 30: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 18.297 transcritos no Córtex Cingulado Posterior.

A Figura 31 e a Tabela 18 seguinte mostram os resultados do primeiro modelo de mistura, apresentado na etapa 1 da seção de Estudo das Regiões do Cérebro, ajustado para os transcritos anteriormente apresentados. De maneira semelhante às regiões anteriores, o modelo resultante foi um VVV com apenas 2 componentes. Destaca-se que, nos dois grupos formados utilizando o critério do ICL, existem características distintas de expressões gênicas: em um mesmo grupo é possível encontrar transcritos com expressão negativa e com expressão positiva, tornando-o heterogêneo. Isso pode ter ocasionado nas baixas medianas observadas na Tabela 19. Outro ponto a ser destacado é que, apesar desta divergência dentro de cada grupo, o grupo 1 foi formado, aparentemente, pelos transcritos com expressão gênica mais forte em relação aos transcritos do outro grupo.

Tabela 18: Resultados do ajuste do agrupamento inicial da região Córtex Cingulado Posterior.

Estatísticas	Agrupamento Inicial Modelo VVV com 2 componentes
Log-verossimilhança	-204.332,70
N	18.297
Graus de liberdade	109
BIC	-409.735,10
ICL	-412.207,90



Figura 31: Heatmap dos 18.297 transcritos da região Córtex Cingulado Posterior resultante do primeiro agrupamento.

Por meio do gráfico anterior e com auxílio da Tabela 19, observa-se que o grupo 2 foi formado pela maior quantidade de transcritos e que também foi o grupo com menor mediana em valor absoluto (quase nula). Como mencionado anteriormente, os grupos formados foram divergentes, tendo muitos transcritos opostos. Esse comportamento tem grande destaque neste grupo 2 que, além de ser maior, possui muitos transcritos com expressão positiva e negativa, resultando neste valor de mediana quase zero. Dessa forma, selecionou-se apenas o grupo 1 para seguir o restante dos procedimentos.

Tabela 19: Divisão em grupos dos 18.297 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	4.279	-0,17
2	14.018	0,0072

Diferentemente das duas regiões anteriores (Hipocampo e Giro Temporal Médio), não foi necessário realizar os procedimentos separados para cada grupo extremo. Esta região seguiu as etapas normalmente e, após realizar o passo 4 apenas uma vez, foram selecionados 515 para serem analisados. No próximo gráfico (Figura 32), é possível ver que os *outliers* anteriores foram, em sua maioria, excluídos ou deixaram de ser pontos extremos e que restaram transcritos com expressão tanto positiva quanto negativa. Isto resultou em medianas de cada amostra em torno de zero, como pode ser observado pela linha horizontal preta destacada na caixa de cada amostra.

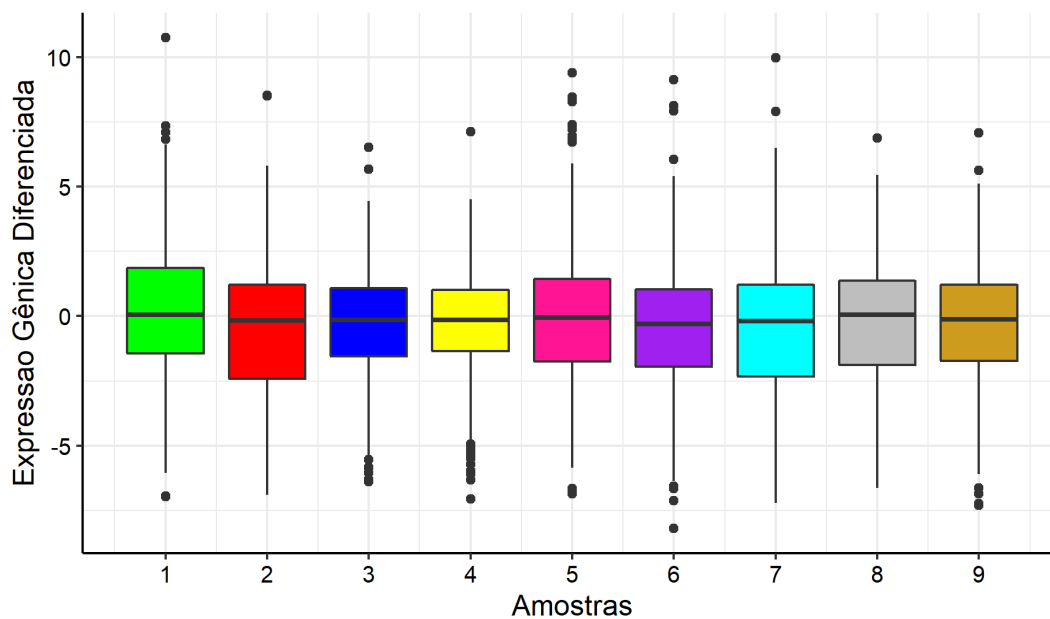


Figura 32: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 515 transcritos no Córtex Cingulado Posterior.

Com esta seleção de transcritos, foi possível obter aqueles que possuem forte expressão, seja ela negativa ou positiva, como observado na Figura 33 a seguir. Nota-se também, por meio da Tabela 20, que foi possível reduzir a complexidade do modelo em relação ao primeiro modelo (Tabela 18), pois o modelo final foi um VVE no qual o volume e a forma são variados, porém a orientação é igual. Além disso, apesar do baixo número de transcritos restantes, o algoritmo os separou em 5 grupos distintos e todos aparentam ter em sua formação transcritos fortemente expressos. O grupo 4 é o maior

grupo e aquele formado por transcritos sub-expressos em pacientes com a DA, enquanto o grupo 5 é aquele com transcritos super-expressos naqueles com a doença.

Tabela 20: Resultados do ajuste do agrupamento final da região Córtex Cingulado Posterior.

Estatísticas	Agrupamento Final Modelo VVE com 5 componentes
Log-verossimilhança	-9.135,61
N	515
Graus de liberdade	130
BIC	-19.082,96
ICL	-19.112,33

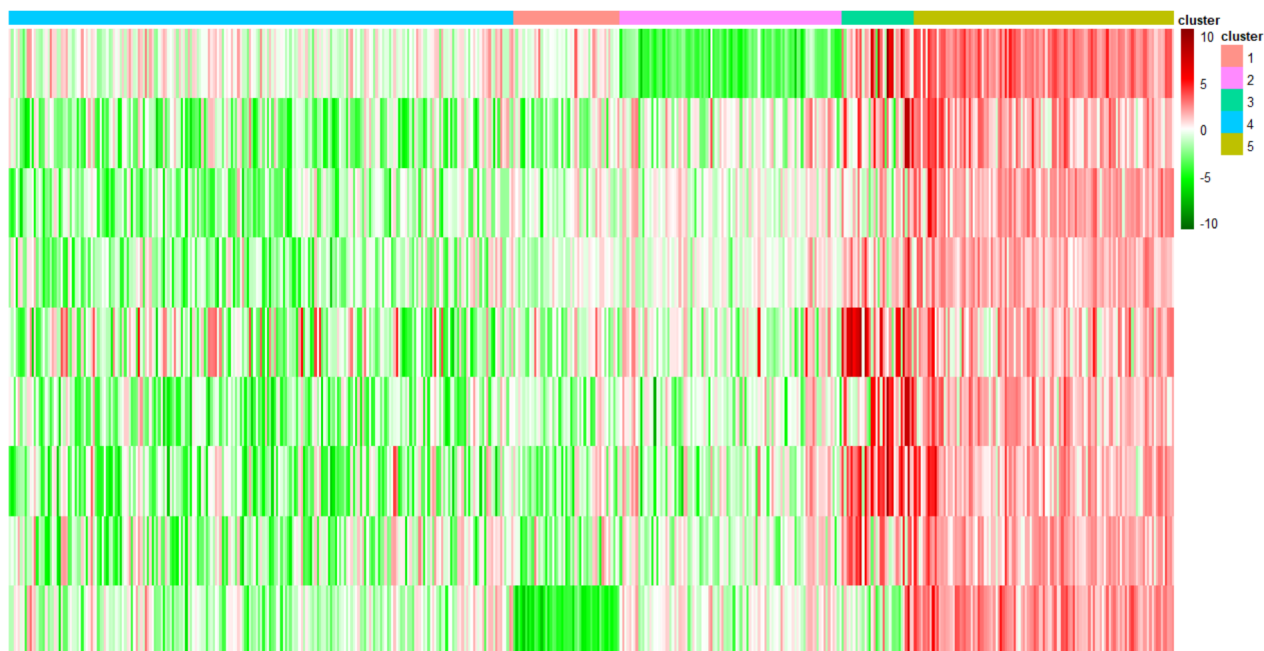


Figura 33: Heatmap dos 515 transcritos da região Córtex Cingulado Posterior resultante do último agrupamento.

Com auxílio da Tabela 21, pode-se melhor interpretar os resultados gráficos apresentados acima. Nela, tem-se as informações a respeito das medianas dos grupos e nota-se que, exceto pelo grupo 2 (grupo em rosa na Figura 33), todas as medianas foram próximas a -1 ou 1. As maiores medianas, em módulo, foram dos grupos 3 e 5 que são compostos, em sua maioria, por transcritos positivamente expressos, ou seja, são aqueles que têm maior expressão em idosos com a DA.

Tabela 21: Divisão em grupos dos 515 transcritos, obtidos pelo modelo final, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	47	-0,94
2	98	-0,37
3	32	1,38
4	223	-1,00
5	115	1,88

Apesar do grupo 5 ser o de maior mediana, observa-se que não foram encontrados processos biológicos “super-representados” neste, conforme apresentado pela Figura 34. Por outro lado, os grupos 1, 3 e, principalmente, 4 (por apresentar mais processos), que foram os demais que apresentaram mediana elevada, tiveram funções super-representadas pelos seus genes.

Destaca-se que, de maneira geral, cada grupo apresentou processos biológicos distintos em relação a outro grupo. A função do grupo 1 está associada às vias da anemia de Falconi que é uma doença hereditária caracterizada pela redução progressiva de leucócitos e plaquetas no sangue; já o grupo 3 apresenta funções relacionadas à contração muscular. No grupo 4, as primeiras funções apresentadas são semelhantes às funções vistas no Córtex Entorrinal (13) e no Hipocampo (Figura 20), são elas: *AUF1 (hnRNP D0) binds and destabilizes mRNA* (AUF1 (hnRNP D0) se liga e desestabiliza mRNA) e *Regulation of mRNA stability by proteins that bind AU-rich elements* (regulação da estabilidade do mRNA por proteínas que se ligam a elementos ricos em AU). As demais funções estão associadas, por exemplo, a algum tipo de degradação.

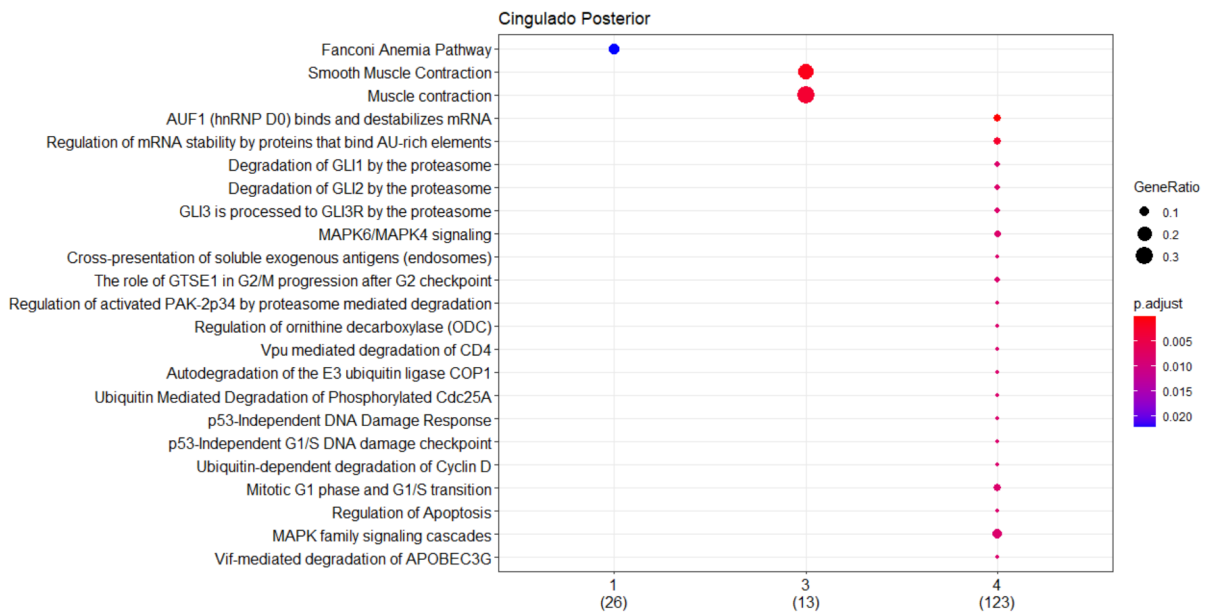


Figura 34: Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Córtex Cingulado Posterior representado na Figura 33.

As funções de cada grupo possuem genes em comum com aquelas encontradas no mesmo grupo, porém não têm relação com as de grupos diferentes (Figura 35). Observe-se que os processos do grupo 4 têm 5 genes comuns que são responsáveis por estes e se diferenciam por alguns genes particulares. Isso ocorre de maneira semelhante no grupo 3 que possui 2 genes comuns e um exclusivo da função de contração muscular (*muscle contraction*).

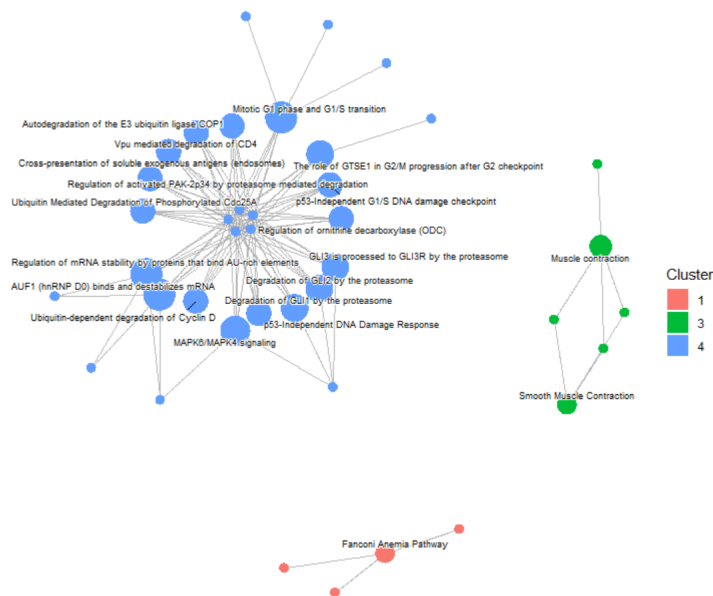


Figura 35: Processos biológicos da Figura 34 presentes no Córtex Cingulado Posterior com os respectivos genes relacionados.

Giro Frontal Superior

O Giro Frontal Superior (Figura 36), juntamente com o Giro Frontal Médio, é responsável por diversas funções no organismo, como as comportamentais e relacionadas ao movimento corporal. Ele está localizado no lobo frontal do cérebro, em que as partes pré-frontais estão associadas comandos de iniciativa e atenção e as outras partes controlam e programam o movimento.

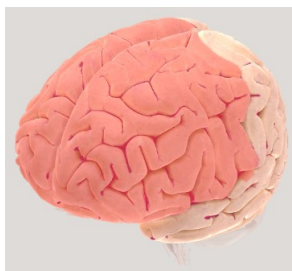


Figura 36: Representação do lobo frontal no qual o Giro Frontal Superior está localizado em sua parte superior, próximo à linha central do cérebro. *Fonte: Society for Neuroscience (2017). Disponível em: https://www.brainfacts.org/3d-brainintro=true?focus=Brain-cerebral_hemisphere-frontal_lobe.*

A Figura 37 mostra a distribuição das expressões gênicas de 20.643 transcritos coletados nesta região anteriormente mencionada. As 23 amostras representadas possuem, de maneira geral, comportamento semelhante: maioria das expressões abaixo de zero (pois o terceiro quartil que é o valor que deixa 75% dos dados abaixo dele, representado pela linha superior do quadrado, está próximo ou abaixo de zero) com muitos pontos extremos, tanto positivos, quanto negativos. Destaca-se que as amostras 7, 10, 13, 19 e 20 foram as que tiveram comportamentos mais diferenciados em relação as demais por apresentar um intervalo interquartil maior.

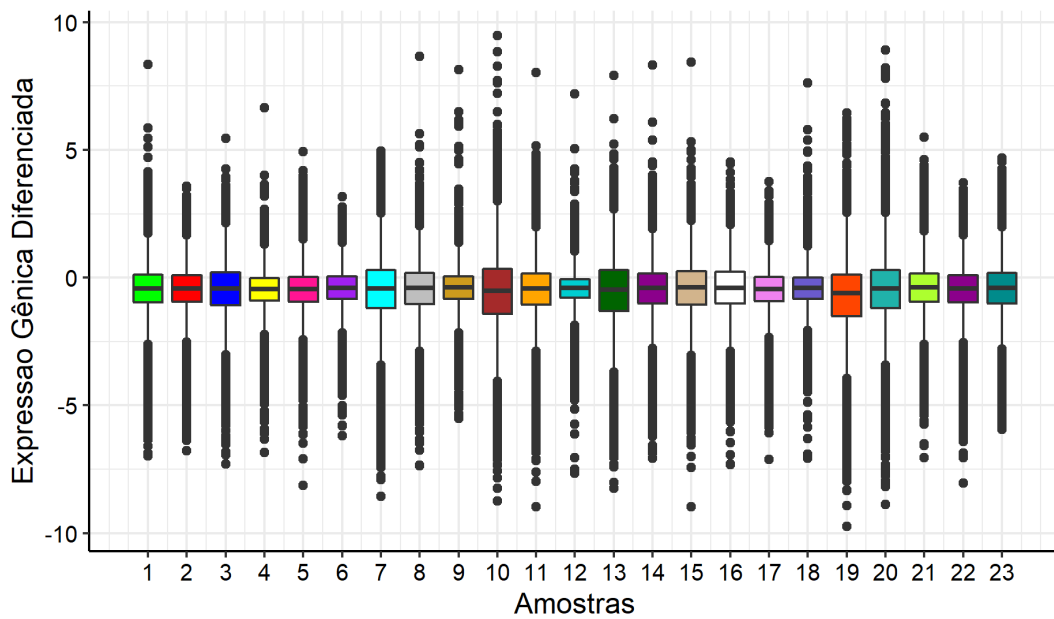


Figura 37: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 20.643 transcritos no Giro Frontal Superior.

Pelo *heatmap* a seguir (Figura 38), tem-se a representação do primeiro agrupamento realizado com os 20.643 transcritos do Giro Frontal Superior, correspondente à etapa 1 dos procedimentos adotados para todas as regiões. É possível perceber que grande parte dos transcritos é sub-expressa em pacientes com a doença de Alzheimer, pois apresentaram valores da expressão gênica diferenciada menor que zero. Além disso, nota-se que o modelo os separou em 7 grupos (Tabela 22), nos quais os volumes, as formas e as orientações são variadas, e que estes grupos, em geral, são homogêneos internamente, ou seja, a maioria dos transcritos que compõe cada um tem o mesmo “sinal” de expressão (positivo ou negativo). Os grupos 6 e 7 foram os que apresentaram uma maior variedade das expressões, tendo em sua composição aquelas menores e maiores que zero.

Tabela 22: Resultados do ajuste do agrupamento inicial da região Giro Frontal Superior.

Estatísticas	Agrupamento Inicial Modelo VVV com 7 componentes
Log-verossimilhança	-457.543,30
N	20.643
Graus de liberdade	2.099
BIC	-935.940,50
ICL	-945.881

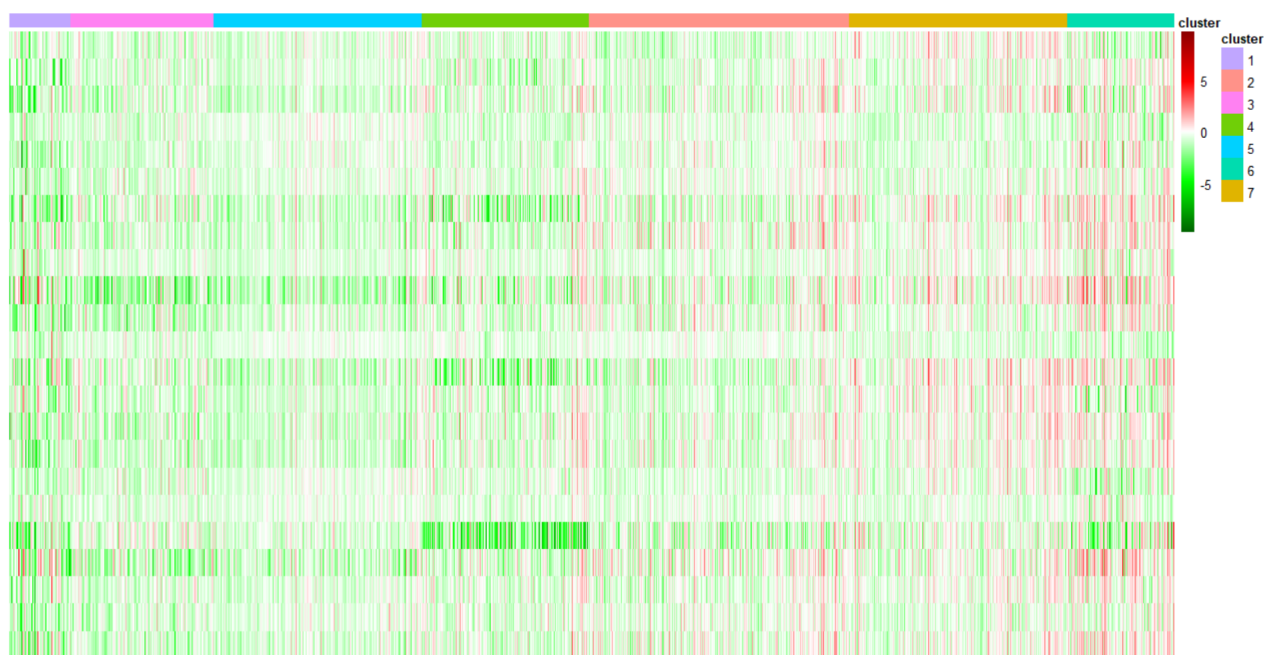


Figura 38: Heatmap dos 20.643 transcritos da região Giro Frontal Superior resultante do primeiro agrupamento.

Nota-se, na Tabela 23, que todas as medianas foram abaixo de zero, confirmando o que foi mostrado no gráfico de que a maioria dos transcritos tem expressão negativa (Figuras 37 e 38). Os grupos de menores medianas (grupos 6 e 7), em módulo, são aqueles anteriormente mencionados como mais heterogêneos por possuírem transcritos tanto sub-expressos como super-expressos em pacientes com a DA. Por outro lado, destacam-se os grupos 1 e 3 que têm, em sua formação, 50% dos transcritos abaixo de -0,79 e -0,65, respectivamente; estes foram os selecionados para as etapas seguintes do agrupamento.

Tabela 23: Divisão em grupos dos 20.643 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	1.081	-0,79
2	4.601	-0,35
3	2.539	-0,65
4	2.961	-0,48
5	3.686	-0,50
6	1.902	-0,10
7	3.873	-0,24

Após todos os cortes de transcritos, restaram 642 distribuídos nas 23 amostras em estudo. Observa-se que, como esperado, mais de 75% dos transcritos de cada amostra tiveram valores de expressão inferiores a zero e com alguns *outliers* principalmente na

parte negativa dos dados. Além disso, o gráfico da Figura 39 mostra que as amostras passaram a ter maiores diferenças na distribuição das expressões: algumas com intervalo interquartil maior que outras, com medianas variando mais e com distribuições variando entre simétricas e assimétricas.

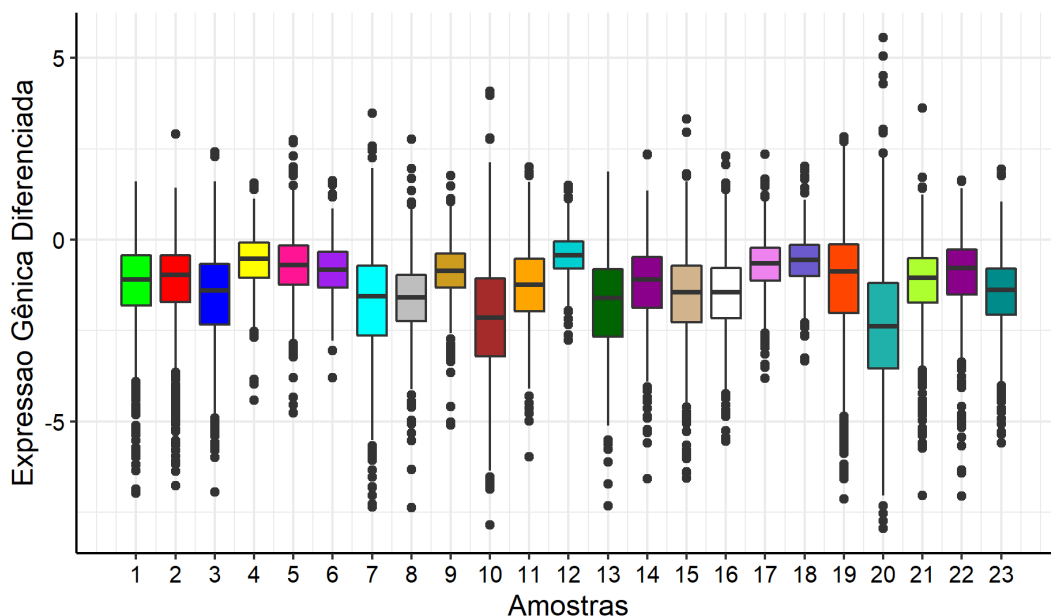


Figura 39: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 642 transcritos no Giro Frontal Superior.

No último agrupamento realizado, ou seja, após realizar a quarta etapa três vezes antes do algoritmo estabilizar, os 642 transcritos foram divididos em apenas 2 grupos com quase o mesmo tamanho (Tabela 25). Dessa forma, além de reduzir a complexidade do modelo, o número de componentes também foi diminuído (Tabela 24): o modelo, que antes tinha 7 componentes e volume, forma e orientação variadas entre essas componentes, passou a ter apenas duas componentes com orientação igual.

Na Figura 40, é quase imperceptível a presença de transcritos com expressão positiva; pode-se observá-los melhor no grupo 2 em que encontra-se alguns traços vermelhos. Por outro lado, os transcritos em verde estão bem representados no gráfico tanto no grupo 1, quanto no grupo 2. Analisando a imagem, é possível notar que algumas amostras possuem transcritos com expressões mais baixas que outras. Por exemplo, a amostra 4 está representada por cores claras e bem próximas do branco que é a representação do número zero. Já a amostra 20 tem a cor verde clara em grande destaque, mostrando que a maioria de seus transcritos está próximo a -5.

Tabela 24: Resultados do ajuste do agrupamento final da região Giro Frontal Superior.

Estatísticas	Agrupamento Final	
	Modelo VVE com 2 componentes	
Log-verossimilhança	-18.688,17	
N	642	
Graus de liberdade	346	
BIC	-39.613,09	
ICL	-39.622,24	

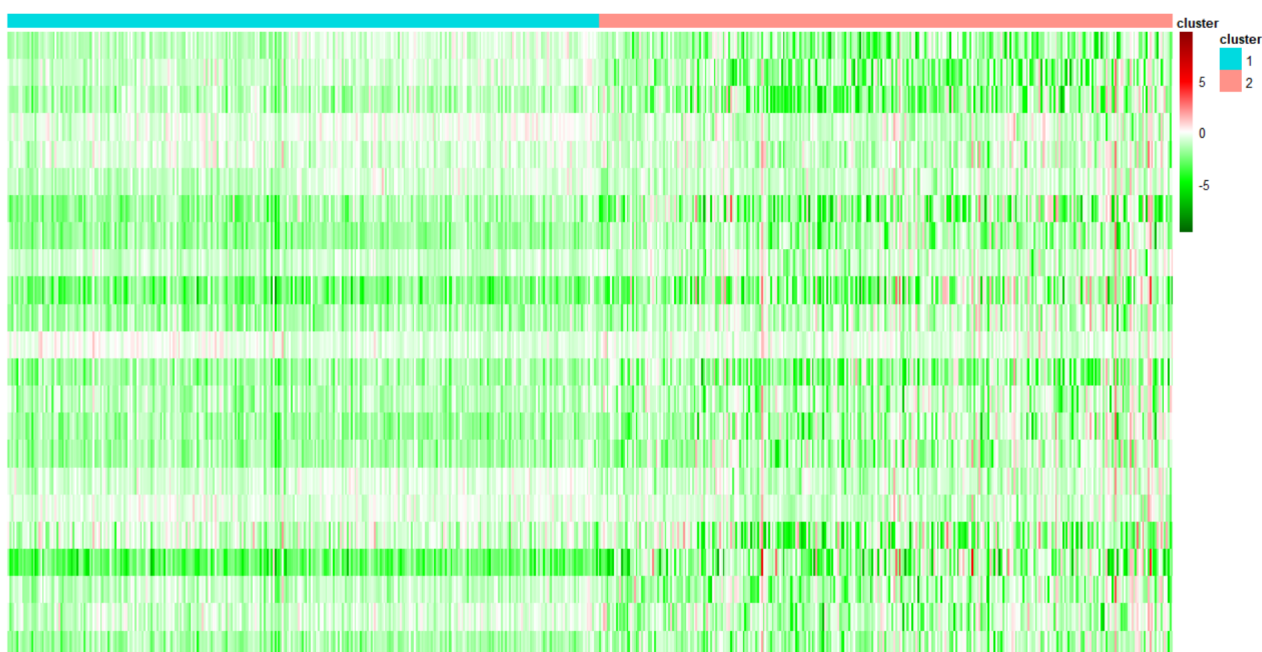


Figura 40: Heatmap dos 642 transcritos da região Giro Frontal Superior resultante do último agrupamento.

Por meio da próxima tabela (25), vê-se que a quantidade de transcritos em cada grupo é bem semelhante, tendo apenas 10 transcritos de diferença. Além disso, as medianas também foram muito semelhantes e ambas relativamente altas (em módulo): o grupo 1 apresentou mediana igual a -1,07, enquanto a do grupo 2 foi -1.

Tabela 25: Divisão em grupos dos 642 transcritos, obtidos pelo modelo final, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	326	-1,07
2	316	-1,00

Com estes transcritos selecionados, aplicou-se a técnica de Análise de Super-Representação para encontrar os genes correspondentes a cada transcrito e suas respectivas

funções. Pela Figura 41, vê-se que ambos os grupos tiveram funções gênicas consideradas como super-representadas, apesar de terem alguns p-valores muito próximos do nível de 5% de significância (principalmente os do grupo 2).

Algumas funções apresentadas estão presentes em outras regiões. Por exemplo, os seis primeiros processos biológicos apresentados no gráfico abaixo que são relacionados ao sistema neural, sinapses e neurotransmissão também foram vistos na região do Giro Temporal Médio.

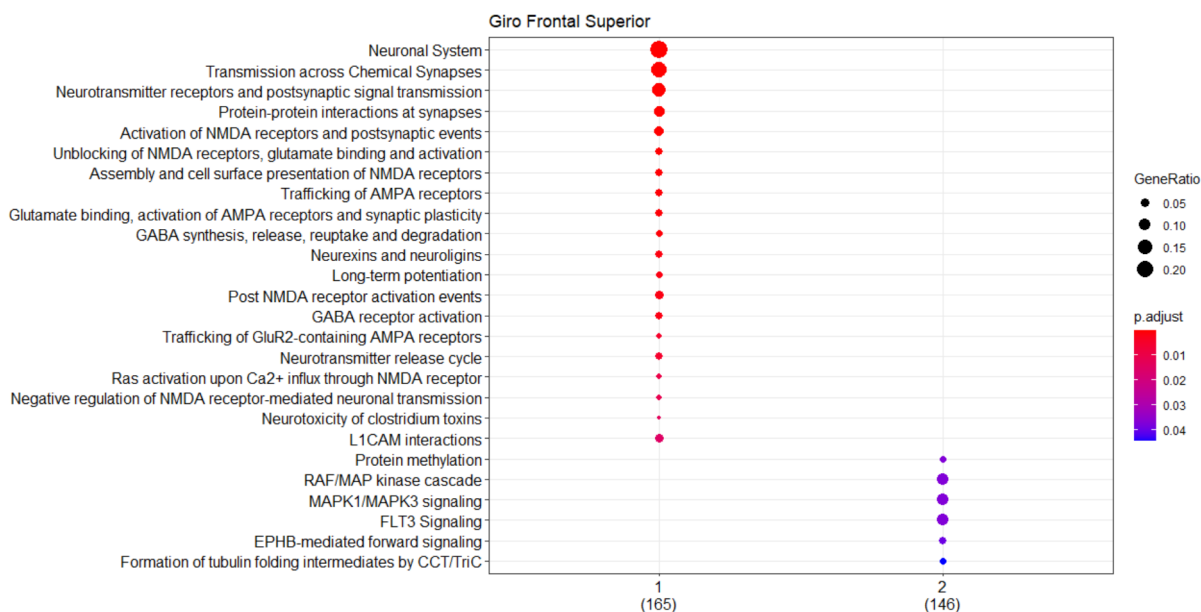


Figura 41: Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Giro Frontal Superior representado na Figura 40.

Diferentemente das regiões anteriores, esta apresentou funções todas interligadas (Figura 42). Isso significa que todos os processos biológicos apresentados possuem genes em comum que são responsáveis por eles. Outro ponto a se destacar é que, mesmo que as funções tenham sido encontradas em grupos distintos, elas se ligam por vários genes.



Figura 42: Processos biológicos da Figura 41 presentes no Giro Frontal Superior com os respectivos genes relacionados.

Córtex Visual Primário

O Córtex Visual Primário (Figura 43) recebe este nome por ser o primeiro local do cérebro a receber a informação visual. A informação interpretada pela região do córtex visual está relacionada a outras regiões do cérebro, como o lobo temporal que está diretamente associado à doença de Alzheimer.

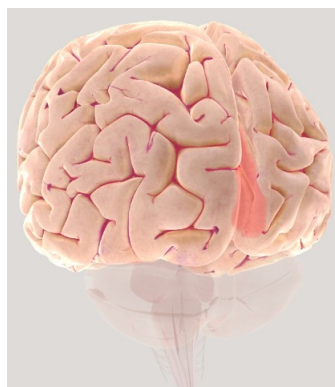


Figura 43: Representação do Córtex Visual Primário. *Fonte: Society for Neuroscience (2017).*
Disponível em: https://www.brainfacts.org/3d-brainintro=true&focus=Brain-cerebral_hemisphere-occipital_lobe-primary_visual_cortex.

Esta região foi analisada, inicialmente, com 18.664 transcritos em 19 amostras. O gráfico abaixo (Figura 44) mostra a distribuição destes transcritos em cada uma das amostras e, com isso, pode-se notar que as amplitudes são variadas, assim como o intervalo interquartil. As que se destacam com menor amplitude são as de número 2, 15, 16 e 17 e as de maior amplitude são 1, 7, 9, 18 e 19.

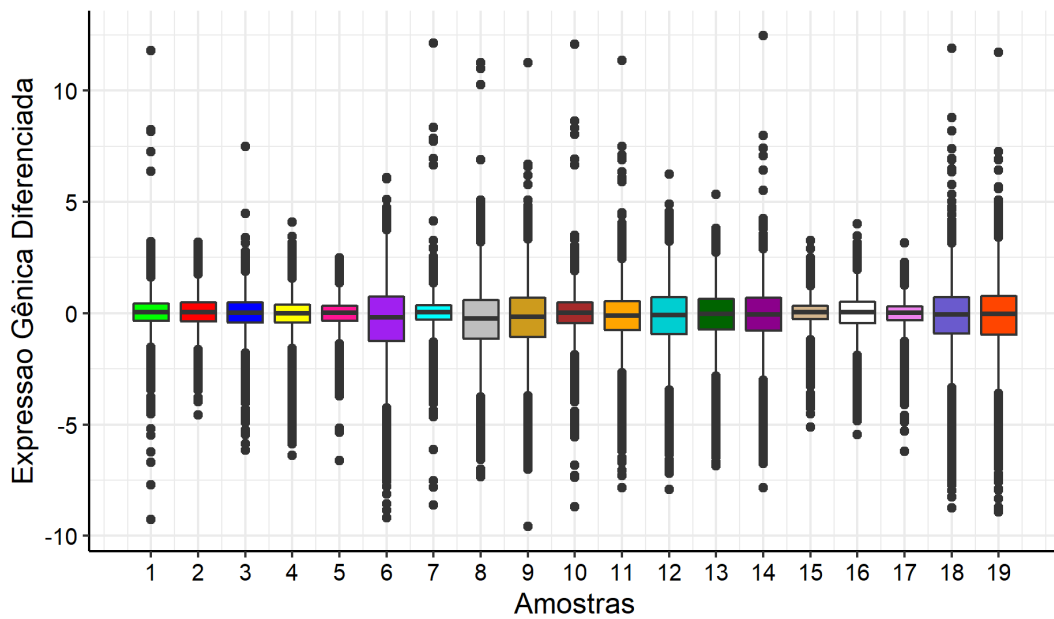


Figura 44: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 18.664 transcritos no Córtex Visual Primário.

Além disso, essa variedade pode ser visualizada no *heatmap* abaixo que mostra, em todos os grupos, transcritos com expressões gênicas positivas, negativas e muitas próximas de zero (devido ao gráfico estar bem claro). Apesar disso, é possível ver que esses valores estão bem divididos entre aqueles sub-expressos (em verde) em pacientes com a doença e aqueles super-expressos (em vermelho). Este gráfico é resultante da primeira etapa do procedimento descrito no início desta seção e o resultado deste agrupamento inicial encontra-se na Tabela 26 abaixo. Nota-se que os mais de 18.000 transcritos foram alocados em 7 grupos distintos, com tamanhos diferentes (Tabela 27), e que apresentaram volume, forma e orientação variados (modelo VVV).

Tabela 26: Resultados do ajuste do agrupamento inicial da região Córtex Visual Primário.

Estatísticas	Agrupamento Inicial
	Modelo VVV com 7 componentes
Log-verossimilhança	-363.116,70
N	18.664
Graus de liberdade	1.469
BIC	-740.680
ICL	-749.941,60

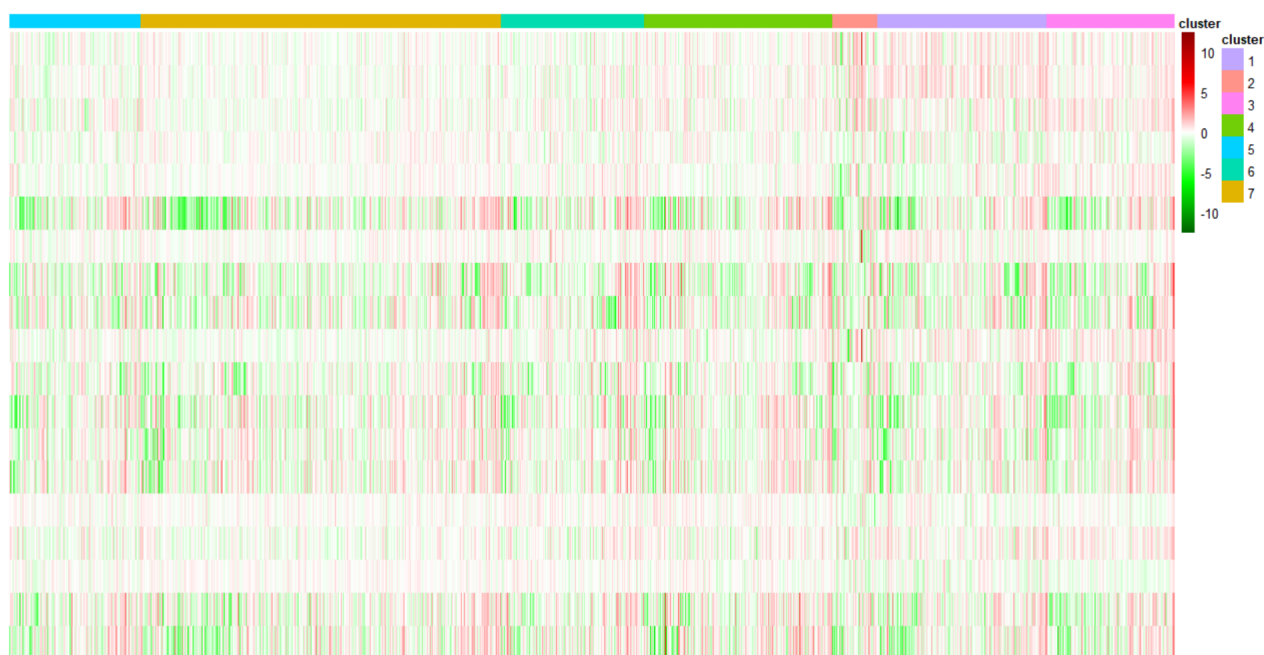


Figura 45: Heatmap dos 18.664 transcritos da região CórteX Visual Primário resultante do primeiro agrupamento.

Nota-se, por meio da Tabela 27, que os valores das medianas estão bem divididos entre positivos e negativos. A maior mediana positiva foi a do grupo 3, seguida pelos grupos 1 e 2, respectivamente, e, entre as negativas, observa-se que a maior mediana, em módulo, é a do grupo 5. Portanto, os transcritos destes grupos permaneceram para as outras etapas dos processos de reagrupamento.

Tabela 27: Divisão em grupos dos 18.664 transcritos, obtidos pelo primeiro modelo, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	2.708	0,28
2	724	0,23
3	2.062	0,40
4	3.002	0,03
5	2.105	-0,22
6	2.297	-0,08
7	5.766	-0,13

Esta região se destaca em relação às outras devido aos transcritos selecionados após os reagrupamentos da etapa 4 serem, em grande parte, positivamente expressos (Figura 46 e 47). Restaram, ao final da etapa 4, apenas 332 transcritos considerados como fortemente expressos por meio dos critérios adotados. Pelo gráfico abaixo, nota-se que as amostras ficaram com comportamentos muito diferentes umas em relação às outras,

principalmente as de número 6, 8, 9, 12, 14, 18 e 19.

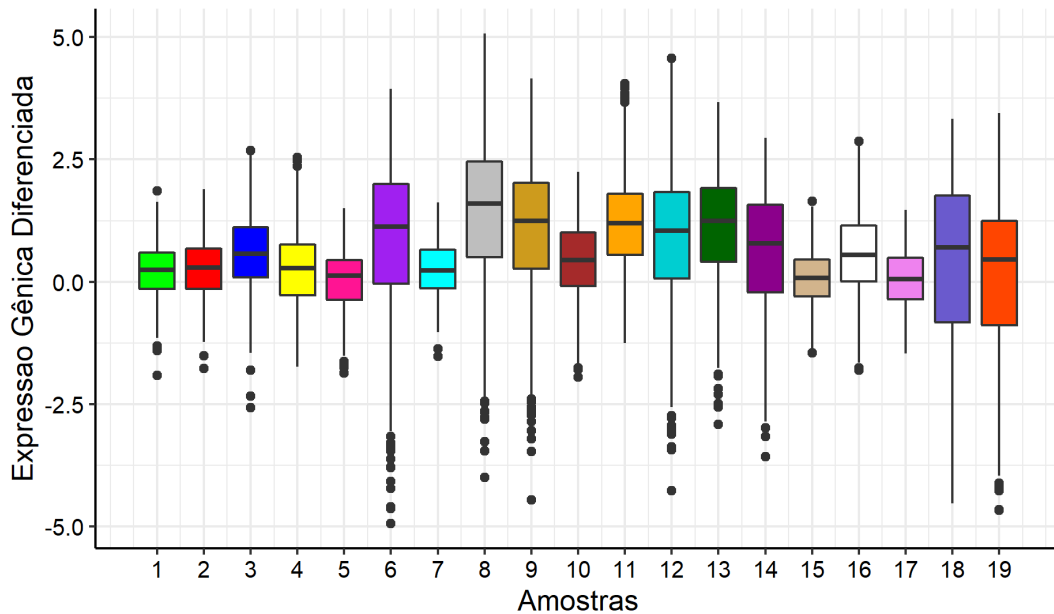


Figura 46: Boxplot de cada amostra com a distribuição das expressões gênicas diferenciadas dos 332 transcritos no Córtex Visual Primário.

Estes transcritos foram divididos em dois grupos, conforme a Figura 47. O primeiro englobou a maioria dos transcritos e é onde encontram-se a maior variabilidade das expressões (positivas, negativas e próximas de zero), enquanto o segundo é majoritariamente formado por transcritos super-expressos em pacientes com a doença. Além disso, pela Tabela 28, nota-se que os dois grupos apresentaram volume e forma variados, porém a orientação de ambos é igual, o que é resultado de um ajuste de modelo VVE de 2 componentes.

Tabela 28: Resultados do ajuste do agrupamento final da região Córtex Visual Primário.

Estatísticas	Agrupamento Final
	Modelo VVE com 2 componentes
Log-verossimilhança	-5.128,29
N	332
Graus de liberdade	248
BIC	-11.696,25
ICL	-11.696,25

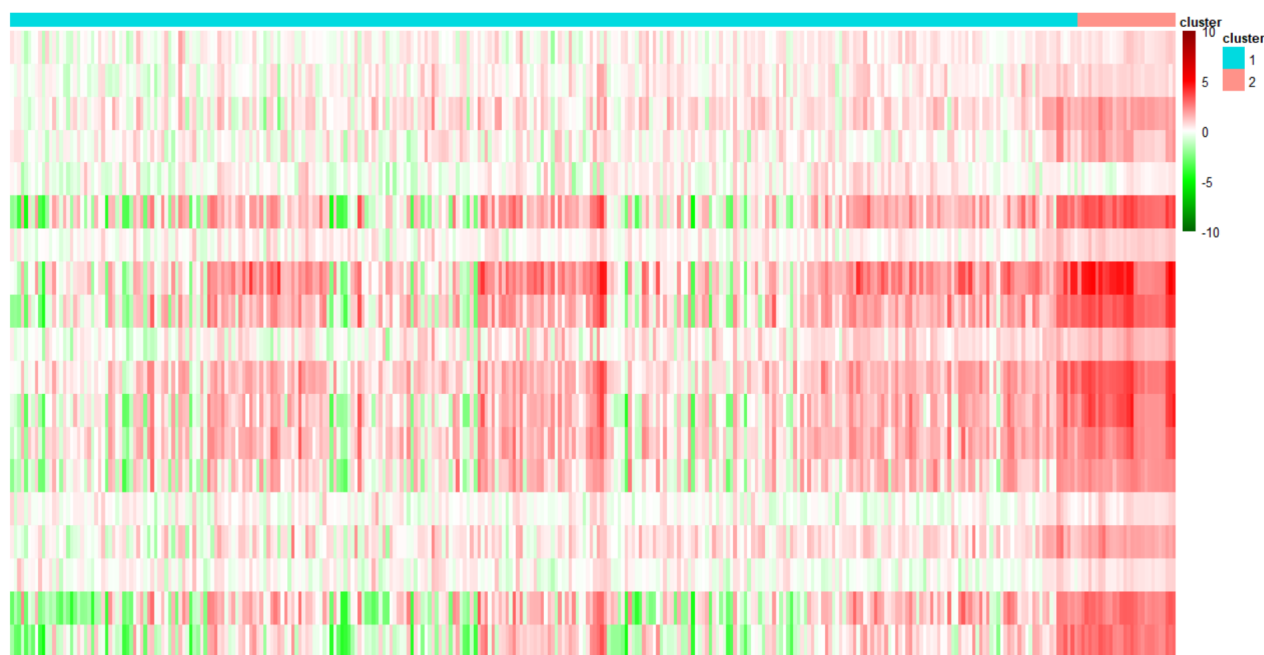


Figura 47: Heatmap dos 332 transcritos da região Córte Visual Primário resultante do último agrupamento.

Pela Tabela 29, pode-se ver a quantidade de transcritos que compuseram cada grupo e as respectivas medianas de cada grupo. O maior grupo é o que apresentou a menor mediana - igual a 0,46 -, já o menor grupo, formado apenas por 28 transcritos, teve mediana igual a 1,92 (uma das maiores encontradas em todo o estudo).

Tabela 29: Divisão em grupos dos 332 transcritos, obtidos pelo modelo final, e respectivas medianas.

Grupo	Quantidade de Transcritos	Mediana
1	304	0,46
2	28	1,92

O Córte Visual Primário apresentou apenas 4 funções gênicas super-representadas entre os genes observados por meio dos transcritos do grupo 1. Estas estão relacionadas com morte celular e neurodegeneração: o primeiro processo envolve a morte celular, como o próprio nome diz - *diseases of programmed cell death* (doença de morte celular programada); o segundo com o metabolismo de fosfatos que fornecem porções de carbono para reações biossintéticas (<https://www.wikipathways.org/index.php/Pathway:WP3806>); o terceiro está diretamente relacionado com a doença de Alzheimer, assim como o último que é um processo associado com uma doença degenerativa (característica que define a DA), porque é um processo associado à proteína CDK5 que, quando desregulada, desencadeia vias neurodegenerativas na DA.

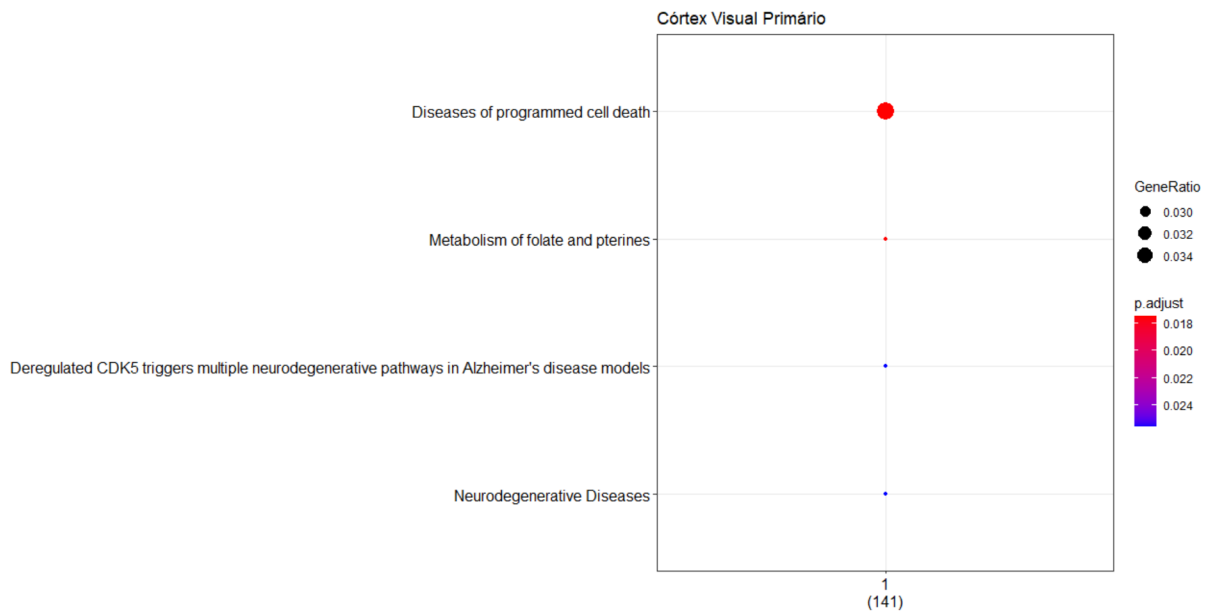


Figura 48: Processos biológicos super-representados em cada um dos grupos obtidos no agrupamento do Córtex Visual Primário representado na Figura 47.

Como citado, três das quatro funções estão relacionadas com neurodegeneração e morte celular. Portanto, é de se esperar que elas estejam interligadas e o gráfico da Figura 49 mostra que existem genes em comum entre os três processos e que o processo correspondente ao metabolismo de fosfolato é originado de 4 genes distintos aos das demais funções.

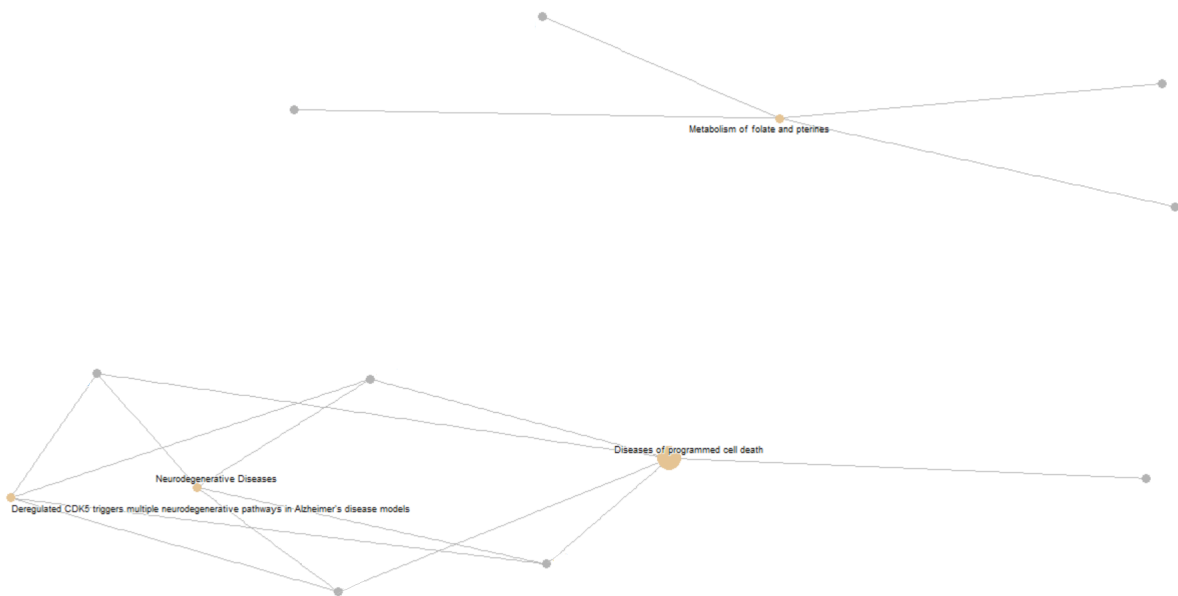


Figura 49: Processos biológicos da Figura 48 presentes no Córtex Visual Primário com os respectivos genes relacionados.

Os processos biológicos encontrados nesta região do cérebro se destacam em re-

lação aos achados nas outras regiões devido à relação direta com a neurodegeneração. Estudos mostram que a visão, que é interpretada primeiro pelo Córtex Visual Primário, pode ser utilizada como uma nova forma de diagnosticar a doença de Alzheimer. No artigo de CRISCUOLO et al., 2018, é abordado que a retina interna, quando comprometida, precede distúrbios neurais que podem ocorrer em outras regiões cerebrais e analisar este comprometimento pode ser uma forma de diagnosticar precocemente as disfunções da DA, assim como mencionado por KUSNE et al., 2017 Além disso, em CHIASSEU et al., 2017 são encontrados acúmulos da proteína tau¹⁰ no sistema visual de camundongos com DA, mostrando mais um indício de que esta área do cérebro pode ser estudada para novas formas de diagnóstico menos invasivo.

¹⁰Esta proteína está associada à formação dos emaranhados neurofibrilares que são estruturas que afetam o cérebro de pacientes com DA (SANAR, 2018).

4 DISCUSSÃO E CONCLUSÃO

O agrupamento por meio de misturas finitas de normais possibilita a divisão e classificação de objetos utilizando a probabilidade de determinada observação pertencer a uma densidade de probabilidade específica. Essa técnica auxiliou na melhoria dos critérios de agrupamento e classificação de dados que, por vezes, era realizado de maneira empírica, pelo censo comum, e, por isso, se restringiam a áreas de aplicação mais “simples”. O uso dessas técnicas de agrupamento para áreas de investigação, como a genética, têm se mostrado cada vez mais forte e eficiente.

Os modelos de misturas finitas de normais utiliza o princípio de mistura de distribuições normais, podendo ser univariadas ou multivariadas, e são classificados como algoritmos hierárquicos aglomerativos. Dessa forma, cada observação inicia em um grupo separado e, por meio da minimização de algum critério de dissimilaridade, são agrupadas até que todas pertençam a um grande grupo. A distribuição de probabilidade de uma observação multivariada, nesta forma de modelagem, consiste em uma média ponderada das densidades de probabilidade que, neste caso, seguirão a distribuição Normal com vetor de médias μ_g e matriz de variância-covariância Σ_g , nos casos multivariados.

Com esses parâmetros, pode-se utilizar a decomposição VSO para obter 14 possíveis modelos de misturas finitas de normais. Esta técnica de decomposição surgiu para corrigir a sensibilidade do modelo em relação ao número de parâmetros que, quando aumenta, pode gerar dificuldades na interpretação dos resultados e na estimação. Para a estimação dos parâmetros do modelo, é usado o algoritmo EM que é um método para se alcançar a convergência na maximização da verossimilhança dos dados observados.

Além disso, assim como em outros métodos de modelagem, existem critérios de seleção de modelos. Para o modelo de misturas finitas, as medidas mais comuns são o critério de informação bayesiana (BIC), o critério de verossimilhança de dados completos integrados (ICL) e o teste de razão de verossimilhança (LRT). A utilização do ICL se tornou, de maneira geral, eficiente para atingir o objetivo de escolher a melhor quantidade de grupos para separar os dados em estudo, além de ser um critério que é mais rápido, em comparação ao LRT, por exemplo. O teste da razão de verossimilhança se torna mais demorado e “pesado” computacionalmente por utilizar o algoritmo *bootstrap* para o cálculo do p-valor. O conceito desta técnica é comparar um mesmo modelo VSO com duas opções para a quantidade de grupos e verificar, por meio do teste, qual é o mais adequado, ou seja, aquele que mais acrescentou informação na verossimilhança. Portanto, deixa-se de sugestão para estudos futuros, a comparação dos modelos escolhidos pelo ICL e pelo LRT.

Neste trabalho, a técnica em estudo foi aplicada a dados de expressão gênica com

o objetivo de agrupar e identificar os transcritos mais expressos, tanto negativa, quanto positivamente, em pacientes com a doença de Alzheimer. Ademais, com essa seleção, foram identificadas as funções mais super-representadas nestes grupos de transcritos considerados extremos e quais eram semelhantes ao comparar os genes responsáveis por cada uma. Com isso, foi possível identificar grupos de transcritos extremos (de um lado transcritos mais expressos em pacientes com a DA e outro com aqueles mais expressos em idosos saudáveis), os genes correspondentes e suas respectivas funções. Nas regiões do Córtex Entorrinal (Figura 14) e Córtex Cingulado Posterior (Figura 35), verificou-se que cada grupo formado por meio da técnica de misturas finitas teve funções e genes específicos, ou seja, não havia semelhança entre os grupos, mostrando que o modelo conseguiu separar bem os transcritos com características diferentes entre grupos, mas comum dentro do grupo. Por outro lado, no Hipocampo (21), os grupos finais apresentaram muitos genes e funções em comum e também foi identificada uma função, dentro de um dos grupos, que só tinha a semelhança de um gene em relação às outras do mesmo grupo. Nas demais regiões, não foi possível identificar um padrão estabelecido dentro dos grupos finais selecionados.

Ademais, foi observado que algumas regiões apresentaram processos biológicos comuns. O Córtex Entorrinal, o Hipocampo e o Córtex Cingulado Posterior tiveram a presença das funções *AUF1 (hnRNP D0) binds and destabilizes mRNA* e *Regulation of mRNA stability by proteins that bind AU-rich elements* e os giros Temporal Médio e Frontal Superior destacaram-se por apresentarem funções relacionadas à neurotransmissão, às sinapses e ao sistema neural.

Outro ponto que se destaca em relação à diferença das regiões do cérebro é que o algoritmo utilizado para a seleção dos grupos extremos, juntamente com o agrupamento por modelos de misturas finitas de normais, não foi homogêneo na quantidade final de transcritos obtidos após a etapa 4. O Giro Temporal Médio, por exemplo, finalizou o algoritmo com mais de 4.000 transcritos em sua formação, enquanto a região do Córtex Visual Primário foi a que teve menos transcritos finais selecionados, sendo apenas 332 que foi bem distante do obtido nas demais regiões.

Por conseguinte, a técnica foi eficiente para atingir o objetivo inicial de separar os grupos de transcritos considerados mais expressos e de obter as respectivas funções gênicas super-representadas nestes grupos. Apesar de algumas divergências nos resultados de cada região do cérebro analisada, foram identificadas algumas funções comuns em determinadas regiões, o que sugere que elas compartilham transcritos de genes comuns e levanta-se a hipótese de que podem ter alguma relação com a doença de Alzheimer. Além disso, para trabalhos futuros, outros métodos de seleção de modelos, como o LRT, outras técnicas e critérios de análises de expressões gênicas, como utilização de média dos grupos ao invés da mediana para identificação de grupos extremos, poderiam ser utilizados para refinar

as técnicas que foram adotadas.

Referências

- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, Wiley Online Library, v. 57, n. 1, p. 289–300, 1995.
- BOUVEYRON, C. et al. *Model-based clustering and classification for data science: with applications in R*. Cambridge, UK: Cambridge University Press, 2019. v. 50.
- CARLSON, M. *hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2)*. [S.l.], 2016. R package version 3.2.3.
- CHIASSEU, M. et al. Tau accumulation in the retina promotes early neuronal dysfunction and precedes brain pathology in a mouse model of alzheimer’s disease. *Molecular neurodegeneration*, BioMed Central, v. 12, n. 1, p. 1–20, 2017.
- CÓRTEX Entorrinal, a área onde a memória se consolida. *A mente é maravilhosa*, 2020. Disponível em: <<https://amenteemaravilhosa.com.br/cortex-entorrinal/>>.
- CRISCUOLO, C. et al. The retina as a window to early dysfunctions of alzheimer’s disease following studies with a 5xfad mouse model. *Neurobiology of aging*, Elsevier, v. 67, p. 181–188, 2018.
- DUDOIT, S. et al. Introduction to dna microarray technologies. *Bioconductor short course, summer*, 2002. Disponível em: <<https://www.bioconductor.org/help/course-materials/2002/Seattle02/MarrayTech.pdf>>.
- GIRO cingulado: estrutura e principais funções. *A mente é maravilhosa*, 2020. Disponível em: <<https://amenteemaravilhosa.com.br/giro-cingulado/>>.
- GRIFFITHS, A. J. et al. *Introdução à genética*. 11. ed. Rio de Janeiro: Guanabara Koogan, 2019.
- JOHNSON, R. A.; WICHERN, D. W. et al. *Applied multivariate statistical analysis*. 6. ed. New Jersey: Prentice hall Upper Saddle River, 2007.
- KANDEL, E. R. et al. *Principles of neural science*. New York: McGraw-hill, 2000. v. 4.
- KUSNE, Y. et al. Visual system manifestations of alzheimer’s disease. *Acta Ophthalmologica*, Wiley Online Library, v. 95, n. 8, p. e668–e676, 2017.
- LIANG, W. S. et al. Altered neuronal gene expression in brain regions differentially affected by alzheimer’s disease: a reference data set. *Physiological genomics*, American Physiological Society, v. 33, n. 2, p. 240–256, 2008.
- LINS, R. d. S. *Aplicação de Modelos Gráficos em Dados Genômicos da Doença de Alzheimer*. Dissertação (Mestrado) — Universidade de Brasília, 2021.
- MCLACHLAN, G. J.; DO, K.-A.; AMBROISE, C. *Analyzing microarray gene expression data*. New Jersey: John Wiley & Sons, Inc., 2004.

MOREIRA, C. Dna complementar. *Revista de Ciência Elementar*, Casa das Ciências, v. 2, n. 1, 2014.

PAGÈS, H. et al. *AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor*. [S.l.], 2020. R package version 1.52.0. Disponível em: <<https://bioconductor.org/packages/AnnotationDbi>>.

SANAR. Doença de alzheimer. 2018.

SCRUCCA, L. et al. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, NIH Public Access, v. 8, n. 1, p. 289, 2016.

SERENIKI, A.; VITAL, M. A. B. F. A doença de alzheimer: aspectos fisiopatológicos e farmacológicos. *Revista de psiquiatria do Rio Grande do Sul*, SciELO Brasil, v. 30, n. 1, p. 0–0, 2008.

SILVA, P. H. D. da. Classificação de dados espectrais de café utilizando análise de discriminantes via mistura de distribuições. 2012.

SOUZA, R. D. et al. A probabilistic approach to emission-line galaxy classification. *Monthly Notices of the Royal Astronomical Society*, Oxford University Press, v. 472, n. 3, p. 2808–2822, 2017.

von Borries, G.; WANG, H. Partition clustering of high dimensional low sample size data based on p-values. *Computational statistics & data analysis*, Elsevier, v. 53, n. 12, p. 3987–3998, 2009.

WANG, L.; LIU, Z.-P. Detecting diagnostic biomarkers of alzheimer’s disease by integrating gene expression data in six brain regions. *Frontiers in genetics*, Frontiers, v. 10, p. 157, 2019.

YU, G.; HE, Q.-Y. Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, v. 12, n. 12, p. 477–479, 2016. Disponível em: <<http://pubs.rsc.org/en/Content/ArticleLanding/2015/MB/C5MB00663E>>.

5 APÊNDICE

5.1 LEITURA DOS DADOS PRÉ-PROCESSADOS POR LI-ANG ET. AL. (2008)

Neste trabalho, um dos conjuntos de dados utilizados foi disponibilizado no *website* da *Gene Expression Omnibus* (GEO). Para ter acesso a tais dados, é necessário realizar o *download* dos comandos a serem executados no *software* R que podem ser obtidos seguindo as etapas descritas a seguir e ilustradas na Figura 50.

1. Entre na seção “Analyse with GEO2R” presente no final da página;
2. Ao entrar nesta página, clique na opção “R script” que encontra-se no último quadro ao final da página e copie o código para ser executado no R;
3. No *script*, insira as linhas 4 a 7 para instalar os pacotes necessários para realizar a leitura dos dados;
4. Salve os dados gerados até a linha 20 em que é obtido o \log_2 da expressão gênica. Esses serão os dados utilizados posteriormente;
5. Após essas etapas, retorne à página inicial e baixe o arquivo em formato .xls com a informação das amostras coletadas;
6. Utilize essas informações para identificar quais são os tecidos provenientes de pacientes com a doença (casos) e sem a doença (controles) de cada região do cérebro;

ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281

Contact name: Winnie Liang
 E-mail(s): wliang@tgen.org
 Organization name: Translational Genomics
 Street address: 445 N. Fifth Street
 City: Phoenix
 State/province: AZ
 ZIP/Postal code: 85012
 Country: USA

Platforms (1): GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array

Samples (161): GSM119615 EC control 1
 # More... GSM119616 EC control 2
 GSM119617 EC control 3

Relations
 BioProject: PRJNA96387

Analyze with GEO2R

Download family | Format
 SOFT formatted family file(s) | SOFT
 MINIML formatted family file(s) | MINIML
 Series Matrix File(s) | TXT

Supplementary file	Size	Download	File type/resource
GSE5281_RAW.tar	965.9 Mb	(http)(custom)	TAR (of CEL, CHP)
GSE5281_sample_characteristics.xls	83.5 Kb	(ftp)(http)	XLS

Raw data provided as supplementary file
 Processed data included within Sample table

NLM | NIH | GEO Help | Disclaimer | Accessibility

(1)

ncbi.nlm.nih.gov/geo2r/?acc=GSE5281

GSM238963 VCX_affected_19 brain, Primary visual cortex

GEO2R | Options | Profile graph | R script

```
# Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
#####
# Data plots for selected GEO samples
library(GEOquery)
library(limma)
library(umap)

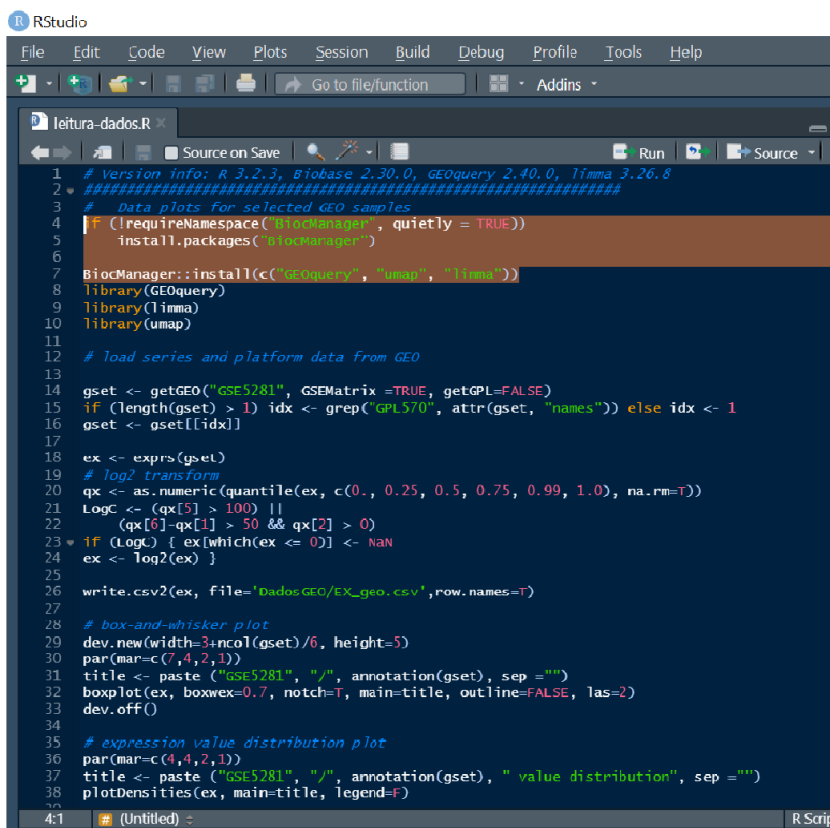
# load series and platform data from GEO
gset <- getGEO("GSE5281", GSEMatrix = TRUE, getGPL = FALSE)
if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
gset <- gset[[idx]]

ex <- exprs(gset)
# log2 transform
qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
LogC <- (qx[5] > 100) ||
        (qx[6]-qx[1] > 50 && qx[2] > 0)
if (LogC) { ex[which(ex <= 0)] <- NaN
  ex <- log2(ex) }

# box-and-whisker plot
dev.new(width=3*ncol(gset)/6, height=5)
par(mar=c(7,4,2,1))
title <- paste("GSE5281", "/", annotation(gset), sep = "")
boxplot(ex, boxwex=0.7, notch=T, main=title, outline=FALSE, las=2)
dev.off()
```

https://www.ncbi.nlm.nih.gov/geo2r/?acc=GSE5281#r_script

(2)

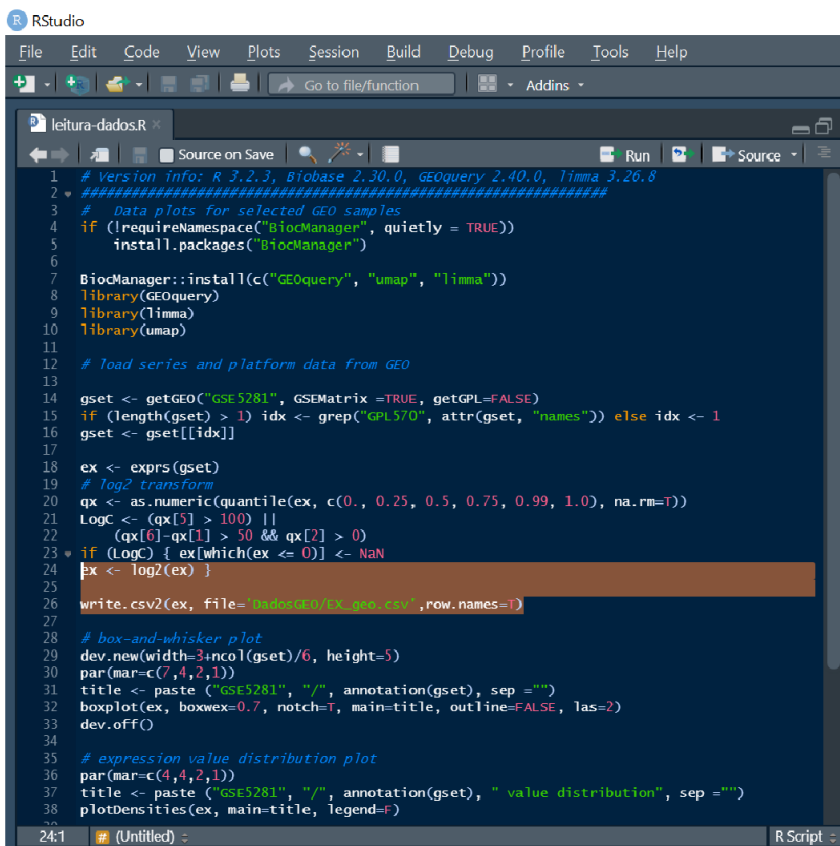


```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
leitura-dados.R
1 # Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
2 #####
3 # Data plots for selected GEO samples
4 if (!requireNamespace("BiocManager", quietly = TRUE))
5   install.packages("BiocManager")
6
7 BiocManager::install(c("GEOquery", "umap", "limma"))
8 library(GEOquery)
9 library(limma)
10 library(umap)
11
12 # load series and platform data from GEO
13
14 gset <- getGEO("GSE5281", GSEMatrix = TRUE, getGPL = FALSE)
15 if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
16 gset <- gset[[idx]]
17
18 ex <- exprs(gset)
19 # log2 transform
20 qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
21 LogC <- (qx[5] > 100) ||
22   (qx[6]-qx[1] > 50 && qx[2] > 0)
23 if (LogC) { ex[which(ex <= 0)] <- NaN
24 ex <- log2(ex) }
25
26 write.csv2(ex, file="DadosGEO/EX_geo.csv", row.names=T)
27
28 # box-and-whisker plot
29 dev.new(width=3+ncol(gset)/6, height=5)
30 par(mar=c(7,4,2,1))
31 title <- paste("GSE5281", "/", annotation(gset), sep = "")
32 boxplot(ex, boxwex=0.7, notch=T, main=title, outline=FALSE, las=2)
33 dev.off()
34
35 # expression value distribution plot
36 par(mar=c(4,4,2,1))
37 title <- paste("GSE5281", "/", annotation(gset), " value distribution", sep = "")
38 plotDensities(ex, main=title, legend=F)
39
4:1 # (Untitled)
R Script

```

(3)



```

RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
leitura-dados.R
1 # Version info: R 3.2.3, Biobase 2.30.0, GEOquery 2.40.0, limma 3.26.8
2 #####
3 # Data plots for selected GEO samples
4 if (!requireNamespace("BiocManager", quietly = TRUE))
5   install.packages("BiocManager")
6
7 BiocManager::install(c("GEOquery", "umap", "limma"))
8 library(GEOquery)
9 library(limma)
10 library(umap)
11
12 # load series and platform data from GEO
13
14 gset <- getGEO("GSE5281", GSEMatrix = TRUE, getGPL = FALSE)
15 if (length(gset) > 1) idx <- grep("GPL570", attr(gset, "names")) else idx <- 1
16 gset <- gset[[idx]]
17
18 ex <- exprs(gset)
19 # log2 transform
20 qx <- as.numeric(quantile(ex, c(0., 0.25, 0.5, 0.75, 0.99, 1.0), na.rm=T))
21 LogC <- (qx[5] > 100) ||
22   (qx[6]-qx[1] > 50 && qx[2] > 0)
23 if (LogC) { ex[which(ex <= 0)] <- NaN
24 ex <- log2(ex) }
25
26 write.csv2(ex, file="DadosGEO/EX_geo.csv", row.names=T)
27
28 # box-and-whisker plot
29 dev.new(width=3+ncol(gset)/6, height=5)
30 par(mar=c(7,4,2,1))
31 title <- paste("GSE5281", "/", annotation(gset), sep = "")
32 boxplot(ex, boxwex=0.7, notch=T, main=title, outline=FALSE, las=2)
33 dev.off()
34
35 # expression value distribution plot
36 par(mar=c(4,4,2,1))
37 title <- paste("GSE5281", "/", annotation(gset), " value distribution", sep = "")
38 plotDensities(ex, main=title, legend=F)
39
24:1 # (Untitled)
R Script

```

(4)

(2)

Supplementary file	Size	Download	File type/resource
GSE5281_RAW.tar	965.9 Mb	(http)(custom)	TAR (of CEL, CHP)
GSE5281_sample_characteristics.xls	83.5 Kb	(ftp)(http)	XLS

Raw data provided as supplementary file
Processed data included within Sample table

(5)

Figura 50: Passo a passo para obter os dados pré-processados pelos autores Liang et. al. (2008)

Em seguida, com os dados pré-processados salvos, é necessário separá-los, primeiro, nas seis regiões do cérebro e retornar os valores à escala original (2^{exp}). Dentro de cada região, calcula-se a média geral dos transcritos (sem a distinção de casos e controles) para que sejam removidos aqueles genes com expressão média menor ou igual a 100. Com esses dados filtrados, retorna-se à escala logarítmica na base 2 (\log_2), calcula-se a mediana dos controles e a expressão gênica diferenciada apresentada na equação 5.1.1 abaixo.

$$\text{Expr}_{dif} = \text{expr}_{casos} - \text{mediana}_{controle} \quad (5.1.1)$$

Após tal filtragem, foi realizado o agrupamento por meio de misturas finitas e as demais análises apresentadas neste relatório.

5.2 Código para gerar os gráficos da Figura 1

```

1 library(mclust)
2 library(Andrews)
3 library(ggplot2)
4 library(factoextra)
5
6 data(iris)
7 iris_quanti <- iris[,1:4]
8
9 ## MODELO EII ##
10 irisEII <- Mclust(iris_quanti, G = 3, modelNames = "EII")

```

```
11 summary(irisEII)
12
13 ## MODELO VII ##
14 irisVII <- Mclust(iris_quanti, G = 3, modelNames = "VII")
15 summary(irisVII)
16
17 ## MODELO EEI ##
18 irisEEI <- Mclust(iris_quanti, G = 3, modelNames = "EEI")
19 summary(irisEEI)
20
21 ## MODELO VEI ##
22 irisVEI <- Mclust(iris_quanti, G = 3, modelNames = "VEI")
23 summary(irisVEI)
24
25 ## MODELO EVI ##
26 irisEVI <- Mclust(iris_quanti, G = 3, modelNames = "EVI")
27 summary(irisEVI)
28
29 ## MODELO VVI ##
30 irisVVI <- Mclust(iris_quanti, G = 3, modelNames = "VVI")
31 summary(irisVVI)
32
33 ## MODELO EEE ##
34 irisEEE <- Mclust(iris_quanti, G = 3, modelNames = "EEE")
35 summary(irisEEE)
36
37 ## MODELO EVE ##
38 irisEVE <- Mclust(iris_quanti, G = 3, modelNames = "EVE")
39 summary(irisEVE)
40
41 ## MODELO VEE ##
42 irisVEE <- Mclust(iris_quanti, G = 3, modelNames = "VEE")
43 summary(irisVEE)
44
45 ## MODELO EEV ##
46 irisEEV <- Mclust(iris_quanti, G = 3, modelNames = "EEV")
47 summary(irisEEV)
48
49 ## MODELO VEV ##
50 irisVEV <- Mclust(iris_quanti, G = 3, modelNames = "VEV")
51 summary(irisVEV)
52
53 ## MODELO EVV ##
54 irisEVV <- Mclust(iris_quanti, G = 3, modelNames = "EVV")
55 summary(irisEVV)
56
57 ## MODELO VVE ##
58 irisVVE <- Mclust(iris_quanti, G = 3, modelNames = "VVE")
59 summary(irisVVE)
60
61 ## MODELO VVV ##
62 irisVVV <- Mclust(iris_quanti, G = 3, modelNames = "VVV")
63 summary(irisVVV)
64
65
66 ## GRAFICOS ##
67
68 par(mfrow = c(3, 5))
69
```

```
70 plot(irisEII , what = "classification", dim = 1:2, xlab = "EII", ylab = "", cex = 0)
71 plot(irisVII , what = "classification", dim = 1:2, xlab = "VII", ylab = "", cex = 0)
72 plot(irisEEI , what = "classification", dim = 1:2, xlab = "EEI", ylab = "", cex = 0)
73 plot(irisVEI , what = "classification", dim = 1:2, xlab = "VEI", ylab = "", cex = 0)
74 plot(irisEVI , what = "classification", dim = 1:2, xlab = "EVI", ylab = "", cex = 0)
75 plot(irisVVI , what = "classification", dim = 1:2, xlab = "VVI", ylab = "", cex = 0)
76 plot(irisEEE , what = "classification", dim = 1:2, xlab = "EEE", ylab = "", cex = 0)
77 plot(irisEVE , what = "classification", dim = 1:2, xlab = "EVE", ylab = "", cex = 0)
78 plot(irisVEE , what = "classification", dim = 1:2, xlab = "VEE", ylab = "", cex = 0)
79 plot(irisEEV , what = "classification", dim = 1:2, xlab = "EEV", ylab = "", cex = 0)
80 plot(irisVEV , what = "classification", dim = 1:2, xlab = "VEV", ylab = "", cex = 0)
81 plot(irisEVV , what = "classification", dim = 1:2, xlab = "EVV", ylab = "", cex = 0)
82 plot(irisVVE , what = "classification", dim = 1:2, xlab = "VVE", ylab = "", cex = 0)
83 plot(irisVVV , what = "classification", dim = 1:2, xlab = "VVV", ylab = "", cex = 0)
```