Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise de biclusterização em células únicas

Pedro Henrique Duarte

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Análise de biclusterização em células únicas

Pedro Henrique Duarte
Orientadora: Daiane Aparecida Zuanetti

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Julho de 2026

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

Single-cell bicluster analysis

Pedro Henrique Duarte
Advisor: Daiane Aparecida Zuanetti

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
July 2026

Pedro Henrique Duarte

Análise de biclusterização em células únicas

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por nome do Pedro Henrique Duarte e aprovado pela banca examinadora.

Aprovado em 26 de Junho de 2026

Banca Examinadora:

- Daiane Aparecida Zuanetti
- Ricardo Felipe Ferreira
- Márcio Luis Lanfredi Viola

Resumo

A análise de células únicas tem ganhado destaque nos últimos anos, impulsionada pelo avanço de tecnologias que tornaram o sequenciamento de RNA mais acessível e com menor perda de informação. Inicialmente, devido a limitações técnicas, o sequenciamento era realizado em massa: o RNA era extraído de uma amostra composta por várias células, resultando em um perfil médio de expressão gênica. Esse método permitia comparar tecidos saudáveis e doentes, ou diferentes condições experimentais, com base na média da expressão dos genes. No entanto, essa abordagem não capturava a heterogeneidade celular presente em um mesmo tecido.

Com o advento do sequenciamento de RNA em célula única, tornou-se possível analisar individualmente cada célula da amostra, o que permite identificar subpopulações celulares e caracterizar variações sutis entre células da mesma região, algo especialmente relevante em tecidos com grande diversidade celular, como o cérebro.

Apesar de suas vantagens, os dados de células únicas apresentam desafios específicos, como alto nível de ruído, grande proporção de zeros (esparsidade) e alta dimensionalidade, uma vez que cada célula é descrita por milhares de genes. Neste trabalho, foram exploradas técnicas de normalização para tratar os dados e reduzir o impacto de efeitos técnicos, e métodos de biclusterização para agrupar simultaneamente células com perfis semelhantes e os genes mais expressivos de cada grupo.

Para isso, uma base de dados de expressão gênica composta por 18.928 genes e 734 células provenientes do cérebro fetal humano e de organoides cerebrais obtidos em diferentes estágios de desenvolvimento foi analisada. A normalização foi capaz de reduzir os efeitos de variações técnicas ao detectar genes relevantes com baixo nível de expressão e a biclusterização, enquanto a biclusterização possibilitou a identificação simultânea de grupos de células com perfis de expressão semelhantes e dos genes mais representativos de cada grupo. Os resultados indicaram que organoides mais maduros apresentam maior similaridade com células fetais, embora ainda não reproduzam completamente todas as características do cérebro fetal humano.

Palavras-chave: *Linnorm*; *FABIA*; análise fatorial; organóides cerebrais; expressão gênica.

Abstract

The single-cell analysis has gained prominence in the last few years, propelled by technological advances that made RNA sequencing more accessible and reduced information loss. Initially, due to technical limitations, sequencing was performed in bulk: RNA was extracted from a sample composed of multiple cells, resulting in an average gene expression profile. This method allowed comparisons between healthy and diseased tissues, or between different experimental conditions, based on average gene expression. However, this approach didn't capture the cellular heterogeneity within the same tissue.

With the advent of single-cell RNA sequencing, it became possible to analyze each cell of the sample individually, enabling the identification of cell subpopulations and characterization of subtle variations among cells from the same region, something specially relevant in tissues with high cellular diversity, like the brain.

Despite its advantages, single-cells data presents specific challenges, such as high levels of noise, a large proportion of zeros (sparsity) and high dimensionality, since each cell is described by thousands of genes. In this study, normalization techniques were explored to process the data and reduce the impact of technical effects, as well as biclustering methods to simultaneously group cell with similar profiles and the most expressive genes in each group.

To this end, a gene expression dataset composed of 18,928 genes and 734 cells derived from the human fetal brain and cerebral organoids obtained at different developmental stages was analyzed. The normalization was able to reduce the effects of technical variation by detecting relevant genes with low expression levels, while biclustering enabled the simultaneous identification of groups of cells with similar expression profiles and the most representative genes within each group. The results indicated that more mature organoids exhibit greater similarity to fetal cells, although they still do not fully reproduce all the characteristics of the human fetal brain.

Keywords: *Linnorm*; *FABIA*; factor analysis; cerebral organoids; gene expression.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

A análise de células únicas tem como objetivo investigar a expressão gênica em nível individual, ou seja, em diferentes células de uma mesma amostra. A expressão gênica refere-se à quantidade relativa de RNA mensageiro de um gene presente em uma célula, o que indica o quanto ela está direcionando seus recursos para produzir a proteína codificada por esse gene (Saraiva, 2006). Uma expressão gênica com valor zero quer dizer que o gene não está expresso naquela célula, o que também fornece informação importante sobre seu estado funcional. Dessa forma, através do sequenciamento de RNA, é possível quantificar o grau de expressão de cada gene em células específicas, revelando padrões de atividade gênica e variações entre células que seriam ocultadas em análises em massa.

A expressão gênica é medida por meio do sequenciamento das moléculas de RNA presentes em uma amostra. Inicialmente, o RNA é convertido em DNA complementar (cDNA), que é sequenciado para gerar milhares de pequenas sequências, chamadas leituras. Em seguida, essas leituras são alinhadas a um genoma ou transcriptoma de referência para identificar o gene de origem de cada uma. A expressão de cada gene é então estimada pela quantidade de leituras associadas a ele, de modo que genes mais expressos apresentam um maior número de leituras (Wang *et al.*, 2009).

Antigamente, devido à elevada perda de informação das tecnologias disponíveis, era necessário recorrer ao sequenciamento de RNA em massa, que analisa o RNA extraído de um conjunto de células e gera um perfil genético médio da amostra. Esse procedimento permite comparar os perfis genéticos de tecidos saudáveis e doentes, casos controle e tratamento, entre outros, permitindo a identificação de genes diferencialmente expressos em cada caso. No entanto, por se basear em uma média amostral, o sequenciamento em massa não capta a heterogeneidade gênica entre células individuais da amostra (Predeus

et al., 2022).

Com o avanço das tecnologias, surgiram técnicas de sequenciamento de RNA mais acessíveis e com menor perda de informação, possibilitando a análise do perfil gênico de células individualmente. Isso permite estimar a expressão de cada gene em nível celular e estudar as diferenças entre células presentes em um mesmo tecido, dentro de uma única amostra (Adil *et al.*, 2021).

Atualmente, a análise de células únicas possui alguns desafios técnicos e computacionais, sendo um dos principais o alto nível de ruído e a grande quantidade de zeros presentes no conjunto de dados. Para solucionar esses problemas, pode-se aplicar técnicas de normalização para tratar o conjunto de dados antes das análises e métodos de redução de dimensionalidade para trabalhar com as variáveis mais relevantes (ou combinações lineares delas) no conjunto de dados (Adil *et al.*, 2021).

Outra abordagem para lidar com a alta dimensionalidade dos dados consiste na utilização de métodos de biclusterização, uma técnica de aprendizado não supervisionado que realiza simultaneamente o agrupamento de linhas e colunas de uma matriz de dados. Dessa forma, é possível identificar subconjuntos de observações que apresentam padrões semelhantes apenas em um subconjunto específico de variáveis, revelando estruturas locais que podem não ser detectadas por métodos tradicionais de clusterização (Castanho *et al.*, 2024).

A análise de células únicas tem sido realizada principalmente para análise de células cerebrais e, consequentemente, na neurociência. Como ilustração, segundo Gong *et al.* (2024), o Alzheimer é uma doença neurodegenerativa complexa que tem sido objeto de pesquisa e atenção nos últimos anos. A patogênese desta doença envolve alterações nas redes gênicas que envolvem vários tipos de células. Portanto, para analisá-la, foi necessário desenvolver uma abordagem que capturasse a complexidade das interações gênicas e a heterogeneidade celular simultaneamente. Assim, para realizar a análise, foi utilizado um método Bayesiano de biclusterização de células únicas que foi capaz de identificar as perturbações gênicas em diferentes grupos celulares e as correlações entre os genes e células durante o progresso da doença.

O objetivo deste trabalho, portanto, é estudar, aplicar e analisar o desempenho de técnicas de normalização e biclusterização na análise da expressão gênica de células únicas. Inicialmente, os dados foram tratados por meio de métodos de normalização, a fim de reduzir o ruído e as distorções geradas por fatores técnicos no processo de coleta dos dados.

Em seguida, foram aplicadas técnicas de biclusterização para agrupar simultaneamente células com perfis de expressão semelhantes e os genes mais expressivos de cada grupo, reduzindo a dimensionalidade do conjunto de dados e destacando os genes mais relevantes para a caracterização das subpopulações celulares.

Esse relatório está organizado da seguinte maneira. No Capítulo 2, é apresentado e descrito o método de normalização *Linnorm* que foi utilizado para tratar os dados. No Capítulo 3 é apresentado e descrito o método de biclusterização *FABIA*, que foi utilizado na análise de dados e para comparação de resultados. No Capítulo 4 a base de dados utilizada na análise foi descrita e brevemente analisada, e os métodos de normalização *Linnorm* e biclusterização *FABIA* foram aplicados. Por fim, o Capítulo 5 traz a conclusão e os próximos passos deste trabalho.

Capítulo 2

Normalização

O sequenciamento de RNA mensageiro introduz diversas fontes de variação que podem mascarar sinais biológicos relevantes. Por isso, em dados de células únicas, a normalização é uma etapa essencial da análise. No entanto, grande parte dos métodos de normalização, que funcionam em dados de sequenciamento em massa, como métodos baseados em fatores de escala globais, não se aplica bem aos dados de células únicas por conta da grande quantidade de zeros e do elevado ruído técnico (Bacher *et al.*, 2017).

Segundo Bacher *et al.* (2017), a normalização tem o papel de ajustar essas variações, tornando os níveis de expressão gênica comparáveis, seja entre genes de uma mesma célula, seja entre diferentes células ou amostras. De forma geral, podemos dividir os métodos de normalização em dois tipos:

- Normalização dentro da amostra: busca tornar os valores de expressão comparáveis entre os genes de uma mesma célula. Um exemplo simples é dividir a expressão de cada gene pela soma total das expressões da célula, obtendo proporções. Isso permite identificar quais genes estão mais ou menos expressos dentro daquela célula;

- Normalização entre amostras: corrige diferenças globais entre células, como variações na profundidade de sequenciamento para permitir comparações válidas entre diferentes células.

Atualmente, diversos métodos de normalização são utilizados em dados de células únicas. Entre eles, os métodos *SAMstrt*, *scran* e *SCnorm* utilizam abordagens semelhantes, baseadas na utilização de multiplicadores aplicados às expressões gênicas de cada célula, denominados fatores de escala. O método *Seurat* incorpora um passo adicional de imputação de dados para substituir os zeros. Já o *BASiCS* adota uma estratégia de

correção baseada em um modelo Bayesiano. Por sua vez, o método *NODES* utiliza uma abordagem não paramétrica, convertendo os dados em pseudocontagens com o objetivo de reduzir a variância de genes estáveis e mitigar a heterogeneidade do conjunto de dados (Yip *et al.*, 2017).

O *Linnorm*, em especial, é melhor descrito a seguir. Ele é um método de transformação e normalização baseado em um modelo linear e na suposição de normalidade dos dados de expressão gênica, desenvolvido para possibilitar uma análise estatística mais precisa de dados de células únicas. Esse método é capaz de resolver simultaneamente os principais problemas relacionados à variabilidade e ao viés nos dados. O algoritmo se assemelha aos métodos baseados em fatores de escala, nos quais a expressão gênica é ajustada por um fator multiplicativo interpretado como a inclinação de um modelo linear que passa pela origem. No entanto, o *Linnorm* aplica uma transformação logarítmica aos dados e estima um modelo linear que não necessariamente intercepta a origem. Isso permite que a expressão gênica seja ajustada de forma linear ou exponencial, proporcionando um ajuste mais adequado para cada célula e removendo mais ruído do que os métodos de fatores de escala (Yip *et al.*, 2017).

O estudo de Yip *et al.* (2017) comparou o método *Linnorm* com outras abordagens de normalização e transformação de dados, como *SCnorm*, *scran*, *Seurat*, *DESeq-sc*, *BASiCS*, *SAMstrt* e *NODES*. Os autores concluíram que o *Linnorm* apresentou vantagens significativas na remoção de ruído e na preservação das diferenças intercelulares, que possibilitaram melhora em análises de dados de células únicas.

2.1 *Linnorm*

O *Linnorm*, um método de normalização, assume que existe um subconjunto de genes cuja expressão é homogênea (ou estável) entre as diferentes células. A partir da identificação desses genes por meio de uma etapa de filtragem, o algoritmo estima os parâmetros de normalização e transformação com base neles, e então aplica esses parâmetros a todo o conjunto de dados. Esse procedimento é melhor detalhado nas Subseções 2.1.1 e 2.1.2.

2.1.1 Normalização e transformação inicial

Para identificar um subconjunto de genes estáveis de forma precisa, é realizada uma etapa inicial de normalização e transformação, na qual os dados são convertidos em pro-

porções. Seja E_{ij} a expressão gênica do gene i na célula j, $1 \leq i \leq m$, $1 \leq j \leq n$. As expressões gênicas são convertidas em proporções da expressão total de cada célula através da transformação:

$$R_{ij} = \frac{E_{ij}}{\sum_{i=1}^{m} E_{ij}}. \tag{2.1}$$

2.1.2 Filtragem

A filtragem ocorre em duas etapas. Primeiramente, é feita a filtragem dos genes com poucas contagens e alta quantidade de zeros e, em seguida, são filtrados os genes com alto nível de ruído técnico, com base no desvio padrão e na assimetria.

Na primeira etapa, busca-se remover genes pouco expressivos. Para isso, são eliminados os genes cuja proporção de células com expressão não nula é inferior a um valor limiar z, com $0 < z \leq 1$. Em conjuntos de dados com baixo ruído técnico, espera-se que o gráfico da média pelo desvio padrão da expressão gênica apresente uma inclinação negativa, isto é, genes mais expressos tendem a ter menor variabilidade relativa. No entanto, genes com expressão extremamente baixa (valores próximos de zero ou um) frequentemente apresentam média e desvio padrão igualmente baixos, distorcendo o gráfico e formando um padrão em forma de colina, ao conectar a origem $(0, 0)$ ao ponto mais alto da curva.

O Gráfico A da Figura 2.1, obtida a partir de Yip *et al.* (2017), apresenta os dados de amostras replicadas do conjunto SEQC[1], caracterizados por baixo nível de ruído. Nota-se uma inclinação negativa acompanhada de uma colina quando os pontos se aproximam da origem. Já o Gráfico B, referente ao conjunto Yan[2], evidencia, em amarelo, os genes com alto ruído, que seriam filtrados, e, em vermelho, os genes estáveis, cuja inclinação negativa é semelhante à observada no Gráfico A.

Para identificar um limiar de filtragem adequado, o algoritmo inicia com um valor z inicial, filtra os genes, e gera o gráfico da média por desvio padrão utilizando o terço menos expressivo dos genes restantes. O processo é repetido, aumentando gradualmente o valor de z, até que esses genes exibam uma tendência de inclinação negativa no gráfico, indicando que os principais genes ruidosos foram removidos e que os genes restantes apresentam um padrão estatístico mais confiável.

Na segunda etapa, são removidos genes que apresentam variabilidade excessiva com base no desvio padrão e na assimetria. Para identificá-los ajusta-se a curva *LOWESS*,

[1] O conjunto de dados SEQC pode ser obtido por meio do pacote seqc do Bioconductor.

[2] O conjunto de dados Yan pode ser obtido no site http://www.ncbi.nlm.nih.gov/geo utilizando o número de acesso GSE36552.

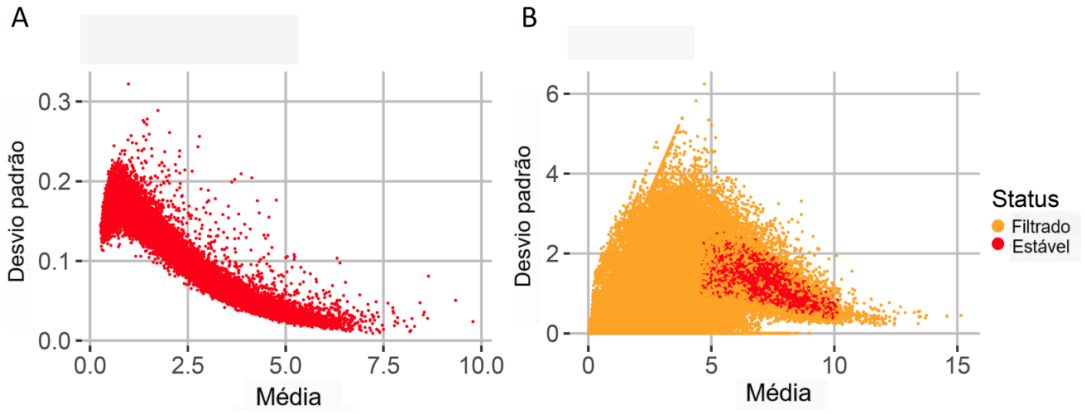

Figura 2.1: Gráfico da média versus desvio padrão dos dados dos conjuntos SEQC e Yan.

uma curva de suavização ponderada localmente, ao gráfico da média versus desvio padrão da expressão gênica. Em seguida, calcula-se a variação logarítmica (*log-fold change*) entre o desvio padrão observado de cada gene e seu valor estimado, obtido a partir da curva ajustada e calculado como:

$$LFC_i = ln\left(\frac{DP_i}{\widehat{DP}_i}\right),\ i = 1, ..., m_1, \tag{2.2}$$

em que m_1 é o número de genes restantes após a primeira etapa da filtragem e DP_i e $\widehat{DP}_i$ são os desvios padrão observado e esperado do gene i, respectivamente.

Essa métrica quantifica o quanto a variabilidade de um gene se desvia do comportamento esperado para o seu nível médio de expressão. Como sua magnitude pode não ser distribuída de forma homogênea (homocedástica) ao longo dos diferentes níveis de expressão, é ajustada uma regressão linear entre a média da expressão gênica e os *log-fold changes*. Assim, pode-se quantificar o quanto o *log-fold change* observado se desvia do esperado, obtido pela regressão linear, através do resíduo studentizado, dado por:

$$t_i = \frac{\sqrt{m_1 - 2}\left(\widehat{LFC}_i - LFC_i\right)}{s},\ i = 1, ..., m_1, \tag{2.3}$$

em que s é o desvio padrão do LFC e LFC_i e $\widehat{LFC}_i$ são os valores de *log-fold changes* observado e esperado do gene i, respectivamente.

Após a remoção de *outliers* (identificados por meio de boxplot), calcula-se, para cada gene restante (o gene i, por exemplo), a probabilidade de que uma variável aleatória T com distribuição t de Student e $m_1 - 2$ graus de liberdade assuma valores mais extremos do que o seu resíduo studentizado (t_i). Esse procedimento permite identificar genes cuja

variabilidade é maior ou menor do que a esperada. Assim, seja $T \sim t_{m_1-2}$, calculamos para cada gene i:

$$p_i = 2\mathbb{P}\left(T \geq |t_i|\right). \tag{2.4}$$

O mesmo procedimento é aplicado para identificar genes com assimetria maior ou menor que a esperada. Primeiramente, ajusta-se uma curva *LOWESS* ao gráfico da média de expressão versus a assimetria dos genes. Em seguida, calcula-se, para cada gene, o *log-fold change* entre a assimetria observada e esperada. É ajustada uma regressão linear entre a média de expressão gênica e os *log-fold changes* e são calculados os resíduos studentizados, dados por:

$$t'_i = \frac{\sqrt{m_1 - 2}\left(\widehat{LFC}_i - LFC_i\right)}{s},\ i = 1, ..., m_1. \tag{2.5}$$

Após a remoção de *outliers*, calcula-se também a probabilidade dada por:

$$p'_i = 2\mathbb{P}\left(T \geq |t'_i|\right). \tag{2.6}$$

Por fim, as probabilidades p_i e p'_i são combinadas usando uma adaptação empírica do método de Brown e os genes com baixa probabilidade são filtrados. O método de Brown não será detalhado neste trabalho, podendo ser consultado em Poole *et al.* (2016).

2.1.3 Cálculo do parâmetro de transformação dos dados

Para determinar o parâmetro da melhor transformação dos dados e, considerando apenas os genes que não foram excluídos pelos filtros anteriores, o *Linnorm* transforma o conjunto de dados usando uma transformação modificada *log-plus-one*. Seja T_{ij} a expressão gênica transformada do i-ésimo gene na j-ésima célula e $\lambda > 0$ o parâmetro de transformação. Então,

$$T_{ij} = ln\left(\lambda R_{ij} + 1\right). \tag{2.7}$$

Para encontrar o parâmetro λ^* que aplica a melhor transformação no conjunto de dados, considerando a homocedasticidade e a normalidade dos dados, é utilizado o coeficiente de desvio $F(\lambda)$, que mensura o desvio das expressões gênicas transformadas destas propriedades. Desta forma, a melhor transformação é obtida pelo parâmetro λ^* que minimiza

o coeficiente de desvio, dado por:

$$F(\lambda) = V(\lambda)^2 + S(\lambda)^2, \tag{2.8}$$

em que $V(\lambda)$ representa a heterocedasticidade e $S(\lambda)$ representa a assimetria dos dados, cujas são dadas nas Subseções 2.1.4 e 2.1.5.

2.1.4 Cálculo de $V(\lambda)$

$V(\lambda)$ quantifica o desvio do conjunto T_{ij} em relação à homocedasticidade. Sejam M_i e D_i a média e o desvio padrão da expressão gênica do gene i em T_{ij}, respectivamente. Se os dados forem homocedásticos, o desvio padrão deve permanecer constante independentemente da média, ou seja,

$$D_i = c, \ \forall i, \tag{2.9}$$

em que c é uma constante positiva.

Dado λ e R_{ij}, estima-se a relação entre D_i e M_i por meio de uma regressão linear dada por:

$$\hat{D}_i = a_d M_i + b_d. \tag{2.10}$$

Nesse contexto, os dados são considerados homocedásticos quando o coeficiente angular $a_d = 0$, pois, desse modo, o desvio padrão é constante, não depende de M_i. Portanto, a_d indica o grau de heterocedasticidade, quanto mais ele se afasta de zero, maior é o desvio da homocedasticidade.

Para transformar essa inclinação em uma métrica comparável com outras, aplica-se uma normalização logarítmica dada por:

$$V(\lambda) = \log\left(|a_d| + 1\right) + 1. \tag{2.11}$$

2.1.5 Cálculo de $S(\lambda)$

Um conjunto de dados normalmente distribuído possui assimetria nula e, por isso, $S(\lambda)$ quantifica a assimetria presente em T_{ij}. Seja s_i o coeficiente de assimetria de Pearson do gene i em T_{ij}; para que os dados sigam uma distribuição normal, é necessário que:

$$s_i = 0, \ \forall i. \tag{2.12}$$

Dado λ e R_{ij}, estima-se a relação entre s_i e a média M_i por uma regressão linear como:

$$\hat{s}_i = a_s M_i + b_s. \tag{2.13}$$

Sejam M_1 e M_m os valores mínimo e máximo da média, respectivamente, o desvio da assimetria estimada em relação a zero é quantificado pela integral:

$$S^{\text{raw}}(\lambda) = \int_{M_1}^{M_m} \frac{|a_s M + b_s|}{M_m - M_1}\, dM. \tag{2.14}$$

Para tornar essa medida comparável a outras métricas, aplica-se uma transformação logarítmica dada por:

$$S(\lambda) = \log\left(S^{\text{raw}}(\lambda) + 1\right) + 1. \tag{2.15}$$

2.1.6 Otimização de λ

Para que o conjunto de dados seja o mais homocedástico e normalmente distribuído possível, é necessário encontrar o valor λ^* que minimiza a função $F(\lambda)$ definida na Equação (2.8). Para isso, será definido um intervalo inicial de busca. Segundo estudos anteriores (Love *et al.*, 2014; Law *et al.*, 2014), observa-se que:

$$a_d \begin{cases} > 0, & \text{se } \lambda = 1, \\ < 0, & \text{se } \lambda = u, \end{cases} \tag{2.16}$$

em que u representa um limite superior para λ. Se os dados estão no formato de contagens brutas, u é definido como a maior soma das contagens entre todas as células. Caso contrário, ou seja, se os dados já foram transformados, estima-se u como:

$$u = \frac{1}{m^{\min}}, \tag{2.17}$$

em que $m^{\min}$ é a média dos valores diferentes de zero dos 5% dos genes com menor média de expressão.

Como o coeficiente a_d que produz o conjunto de dados mais homocedástico é $a_d = 0$, pela Equação (2.16) obtém-se o seguinte intervalo para os valores de λ:

$$1 < \lambda < u. \tag{2.18}$$

Dessa forma, o algoritmo restringe a busca aos valores dentro desse intervalo e aplica um algoritmo iterado de busca local para encontrar um mínimo local da função $F(\lambda)$, que representa o grau de heterocedasticidade e assimetria dos dados após a normalização.

2.1.7 Normalização

Ainda, considerando apenas os genes filtrados no processo descrito anteriormente, calcula-se o valor médio da expressão gênica para cada gene. Então, para cada célula j, $j = 1, ..., n$, é ajustada uma regressão linear da forma $\bar{R}_i = m_j R_{ij} + c_j + \epsilon_i$, em que a variável resposta $\bar{R}_i$ representa a média da expressão gênica do gene i e a covariável R_{ij} é o valor da expressão gênica no gene i e na célula j.

Cada um desses modelos é deslocado em direção à linha de identidade com base no coeficiente de normalização μ, $0 \leq \mu \leq 1$. Para isso, os coeficientes dos modelos são atualizados pelas equações:

$$\begin{cases} m_j^{atualizado} = \mu(m_j - 1) + 1, \\ c_j^{atualizado} = c_j \mu. \end{cases} \tag{2.19}$$

Quando $\mu = 0$, a normalização equivale à escala relativa de expressão gênica descrita na Equação (2.1). Por outro lado, quando $\mu = 1$, o método busca maximizar a remoção de ruído técnico. Por padrão, define-se $\mu = 0.5$, o que proporciona um nível moderado de normalização. No entanto, é importante otimizar este parâmetro com base no conjunto de dados que está sendo utilizado, pois cada conjunto de dados possui um nível diferente de ruído que precisa ser removido.

A seguir, os coeficientes $m_j^{atualizado}$ e $c_j^{atualizado}$ são aplicados a todos os genes do conjunto de dados pela fórmula:

$$B_{ij} = exp\{m_j^{atualizado} G_{ij} + c_j^{atualizado}\}, \tag{2.20}$$

em que $G_{ij} = ln\,(\lambda R_{ij})$, $j = 1, ..., n$ e $i = 1, ..., m$, e λ é o melhor valor encontrado para a transformação, descrito anteriormente.

Essa etapa reverte parcialmente a transformação logarítmica aplicada em G, incorporando os ajustes obtidos na normalização. Por fim, para completar o procedimento de transformação e normalização do algoritmo *Linnorm*, aplica-se novamente a trans-

formação logarítmica, obtendo-se:

$$T_{ij} = ln(B_{ij} + 1). \tag{2.21}$$

O resultado final, T_{ij}, representa a matriz de expressão normalizada e transformada, pronta para análises posteriores, como visualização, clusterização ou modelagem estatística.

Capítulo 3

Biclusterização

A biclusterização é uma técnica de aprendizado de máquina não supervisionado que realiza simultaneamente a clusterização das linhas e colunas de um conjunto de dados. Originalmente desenvolvida como uma ferramenta central na análise de dados de expressão gênica, tornou-se uma das abordagens mais utilizadas para a descoberta de padrões, com aplicações em tarefas descritivas e preditivas (Castanho *et al.*, 2024).

Um bicluster pode ser definido como uma submatriz do conjunto de dados original. Seja B um bicluster que consiste de um conjunto de m_1 genes e n_1 células, em que o elemento b_{ij} representa a expressão gênica do gene i na célula j, $i = 1, ..., m_1, j = 1, ...n_1$. Então o bicluster B pode ser representado por:

$$B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n_1} \\ b_{21} & b_{22} & \cdots & b_{2n_1} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m_1 1} & b_{m_1 2} & \cdots & b_{m_1 n_1} \end{pmatrix}. \tag{3.1}$$

Em biclusterização, coerência significa que os elementos de um bicluster seguem algum padrão consistente entre linhas e colunas. Tipicamente, biclusters possuem três padrões diferentes: constantes, *shifting* e *scaling*. Um bicluster segue um padrão constante se seus valores forem todos iguais, sendo este o padrão mais simples. Um bicluster B segue um padrão *shifting* se seus valores puderem ser obtidos pela soma de um efeito π_i do gene i com um efeito β_j da célula j. Neste caso, os valores de expressão gênica neste bicluster são dados por:

$$b_{ij} = \pi_i + \beta_j. \tag{3.2}$$

De forma similar, um bicluster segue um padrão *scaling* se seus valores puderem ser obtidos pela multiplicação de um efeito π_i do gene i com um efeito α_j da célula j. Nesse caso, os valores do bicluster são dados por:

$$b_{ij} = \pi_i \alpha_j. \tag{3.3}$$

Além disso, biclusters podem apresentar os padrões *shifting* e *scaling* simultaneamente, sendo o caso mais comum quando utilizamos dados reais. Nesse caso, os valores do bicluster são dados por:

$$b_{ij} = \pi_i \alpha_j + \beta_j. \tag{3.4}$$

A Figura 3.1 apresenta exemplos dos tipos de biclusters, facilitando a visualização dos biclusters constantes, caracterizados por possuirem todos os valores iguais, dos biclusters *shifting*, caracterizados por um efeito aditivo entre as linhas e colunas, e dos biclusters *scaling*, caracterizados por um efeito multiplicativo entre as linhas e colunas.

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 10 & 0 \\ 2 & 3 & 11 & 1 \\ 4 & 5 & 13 & 3 \\ 7 & 8 & 16 & 6 \end{pmatrix} \qquad \begin{pmatrix} 1 & 2 & 4 & 1.5 \\ 2 & 4 & 8 & 3 \\ 4 & 8 & 16 & 6 \\ 7 & 14 & 28 & 10.5 \end{pmatrix}$$

(a) Bicluster constante (b) Bicluster *shifting* (c) Bicluster *scaling*

Figura 3.1: Exemplos de biclusters.

Em comparação com métodos tradicionais de clusterização, a biclusterização apresenta três vantagens conceituais. A primeira é que a biclusterização reconhece que um subconjunto de linhas pode ser semelhante apenas em relação a um subconjunto específico de colunas, possibilitando a identificação de padrões locais, como um subconjunto de genes similar em um subconjunto de células (amostras ou observações). A segunda é que esta técnica possibilita que uma mesma observação ou um atributo pertença a diferentes grupos, refletindo, por exemplo, a participação de um mesmo gene em diferentes processos biológicos. A terceira é que, por captar padrões locais, é mais flexível em detectar relações complexas entre as observações.

Devido à grande quantidade de algoritmos de biclusterização, a seleção de um algoritmo adequado aos dados em análise torna-se uma etapa crucial, devendo considerar

aspectos como o tipo dos dados, a eficiência computacional, a robustez a distorções (como valores faltantes ou ruído) e a qualidade dos resultados obtidos (Castanho *et al.*, 2024).

Na análise de dados de expressão gênica, destaca-se o algoritmo *FABIA* (*Factor Analysis for Bicluster Acquisition*), um algoritmo de análise fatorial baseado em um modelo multiplicativo que considera dependência linear entre a expressão gênica e as observações (Hochreiter *et al.*, 2010).

No trabalho de Hochreiter *et al.* (2010), o *FABIA* foi comparado com 11 outros algoritmos de biclusterização em 100 conjuntos de dados simulados, demonstrando desempenho superior ao separar biclusters verdadeiros de biclusters espúrios, por meio da classificação baseada em seu conteúdo.

3.1 *FABIA*

O *FABIA* (do inglês, *Factor Analysis for Bicluster Acquisition*) é um algoritmo de biclusterização baseado em um modelo multiplicativo generativo, projetado para lidar com as particularidades dos dados de expressão gênica, como a presença de dados com distribuição de caudas pesadas. O método utiliza análise fatorial para identificar biclusters no conjunto de dados, definidos como subconjuntos de linhas e colunas, cujas linhas apresentam padrões semelhantes entre si em relação às colunas e vice-versa (Hochreiter *et al.*, 2010).

3.1.1 Modelo geral

Em um modelo multiplicativo, dois vetores são similares se o ângulo entre eles for pequeno (próximo de 0). Esta dependência linear pode ser expressa pelo produto externo $\boldsymbol{\lambda}\mathbf{z}^t$, em que $\boldsymbol{\lambda}$ é um vetor coluna protótipo que vale 0 para genes que não participam do bicluster e $\mathbf{z}$ é um vetor de fatores que vale 0 para as células (amostras ou observações) que não participam do bicluster. Por conterem muitos valores iguais a zero (ou valores próximos de zero), estes vetores são denominados vetores esparsos. Assim, o modelo geral para p biclusters e ruído aditivo é dado por:

$$\mathbf{X} = \sum_{b=1}^{p} \boldsymbol{\lambda}_b \mathbf{z}_b^t + \boldsymbol{\Upsilon} = \boldsymbol{\Lambda}\mathbf{Z} + \boldsymbol{\Upsilon}, \tag{3.5}$$

em que $\mathbf{X} \in \mathbb{R}^{m \times n}$ é o conjunto de dados (no caso deste estudo é a matriz $\mathbf{T}$, composta pela expressão gênica normalizada de n células e m genes), $\boldsymbol{\Upsilon} \in \mathbb{R}^{m \times n}$ é a matriz de ruídos ou erros aleatórios, $\boldsymbol{\lambda}_b \in \mathbb{R}^m$ é o vetor protótipo esparso, $\mathbf{z}_b \in \mathbb{R}^n$ é o vetor de fatores esparso, $\boldsymbol{\Lambda} \in \mathbb{R}^{m \times p}$ é a matriz protótipa esparsa, que contém os vetores protótipos $\boldsymbol{\lambda}_b$ como colunas e $\mathbf{Z} \in \mathbb{R}^{p \times n}$ é a matriz de fatores esparsa, que contém os vetores de fatores esparsos transpostos $\mathbf{z}_b^t$ como linhas.

A Figura 3.2, obtida a partir de Hochreiter *et al.* (2010), traz uma representação visual deste modelo multiplicativo.

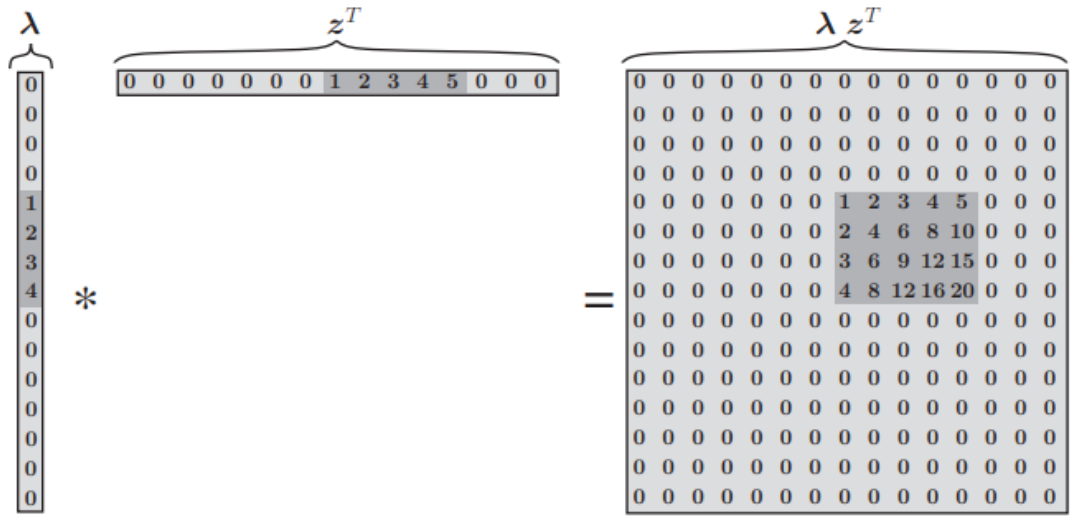

Figura 3.2: O produto externo dos vetores esparsos $\boldsymbol{\lambda}\mathbf{z}^t$ resulta em uma matriz com um bicluster.

Segundo a Equação (3.5), a j-ésima amostra $\mathbf{x}_j$ é definida como:

$$\mathbf{x}_j = \sum_{b=1}^{p} \boldsymbol{\lambda}_b z_{bj} + \boldsymbol{\epsilon}_j = \boldsymbol{\Lambda}\tilde{\mathbf{z}}_j + \boldsymbol{\epsilon}_j, \tag{3.6}$$

em que $\boldsymbol{\epsilon}_j$ é a j-ésima coluna da matriz de ruídos $\boldsymbol{\Upsilon}$ e $\tilde{\mathbf{z}}_j$ é a j-ésima coluna da matriz $\mathbf{Z}$.

A formulação na Equação (3.6) facilita uma interpretação generativa por meio de um modelo de análise fatorial com p fatores, dado por:

$$\mathbf{x} = \sum_{b=1}^{p} \boldsymbol{\lambda}_b \tilde{z}_b + \boldsymbol{\epsilon} = \boldsymbol{\Lambda}\tilde{\mathbf{z}} + \boldsymbol{\epsilon}, \tag{3.7}$$

em que $\mathbf{x}$ é a observação, $\boldsymbol{\Lambda}$ é a matriz de cargas fatorias, $\tilde{z}_b$ é o b-ésimo fator, $\tilde{\mathbf{z}}$ é o vetor de fatores e $\boldsymbol{\epsilon}$ é o ruído aditivo.

A análise fatorial assume que o ruído é independente de $\tilde{\mathbf{z}}$, $\tilde{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ e $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$.

Os parâmetros $\mathbf{\Lambda}$ explicam a variabilidade compartilhada entre as observações, enquanto os parâmetros $\mathbf{\Psi}$ explicam a variabilidade específica de cada variável.

Estas suposições implicam que um bicluster não será subdividido em biclusters menores com correlação alta, pois a matriz de correlação de $\tilde{\mathbf{z}}$ ser a identidade garante que os biclusters não sejam correlacionados.

Entretanto, a análise fatorial padrão não considera fatores e cargas fatorias esparsas, que são essenciais para representar os biclusters nesta formulação. Para contornar este problema, é utilizada a distribuição de Laplace ao invés da Gaussiana como distribuição para os fatores, dada por:

$$p(\tilde{\mathbf{z}}_j) = \left(\frac{1}{\sqrt{2}}\right)^p \prod_{b=1}^{p} e^{-\sqrt{2}|z_{bj}|}, j = 1, ..., n. \tag{3.8}$$

Além disso, a distribuição das cargas fatorias também será alterada. No modelo *FABIA*, é utilizada a distribuição de Laplace como distribuição para as cargas fatoriais (assim como para os fatores), dada por:

$$p(\boldsymbol{\lambda}_b) = \left(\frac{1}{\sqrt{2}}\right)^m \prod_{i=1}^{m} e^{-\sqrt{2}|\lambda_{ib}|}. \tag{3.9}$$

Este modelo contém o produto de variáveis aleatórias Laplacianas, que possui distribuição proporcional à função de Bessel modificada de ordem 0 do segundo tipo. Ao combinar as caudas desta distribuição (mais pesadas que a Laplaciana) com as caudas da distribuição Gaussiana do ruído (mais leves que a Laplaciana), o modelo resulta em uma distribuição com caudas de peso intermediário, próximas às da Laplace, o que é uma característica desejável.

3.1.2 Estimação dos parâmetros

Para identificar os biclusters, é necessário estimar os parâmetros $\mathbf{\Lambda}$ e $\mathbf{\Psi}$ que melhor explicam os dados. Usualmente, é utilizado o método de máxima verossimilhança para estimar os parâmetros, porém, neste caso, este método não possui solução analítica por considerar variáveis latentes, os fatores. Para contornar este problema, é utilizado o algoritmo EM variacional para estimar os parâmetros do modelo.

Segundo Girolami (2001), o método variacional é baseado em representar uma função convexa em termos de uma otimização da forma dual sobre um conjunto de parâmetros,

denominados parâmetros variacionais, de forma que a representação da função na forma dual seja mais adequada na aplicação de um método específico (no caso, o método EM). Assim, utilizando os parâmetros variacionais $\boldsymbol{\xi}$, é possível escrever uma função de verossimilhança como:

$$p(\mathbf{x}) = \max_{\boldsymbol{\xi}} p(\mathbf{x}|\boldsymbol{\xi}). \tag{3.10}$$

Por fim, o algoritmo EM variacional maximiza um limite inferior da função de verossimilhança em termos dos parâmetros $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ e $\boldsymbol{\xi}$ para obter os parâmetros estimados. Os detalhes do algoritmo EM variacional não serão apresentados neste trabalho, podendo ser consultados em Girolami (2001).

O algoritmo EM variacional, por ser iterativo, precisa de valores iniciais para os parâmetros $\boldsymbol{\Lambda}$, $\boldsymbol{\Psi}$ e $\boldsymbol{\xi}$. Hochreiter *et al.* (2010) sugere alguns valores iniciais para este algoritmo. Além disso, é necessário que a média, a mediana ou a moda dos dados seja 0.

Assim como na análise de componentes principais, em que os componentes são ranqueados de acordo com a variabilidade dos dados que eles explicam, os biclusters são rankeados de acordo com a informação que eles contém sobre os dados. Essa informação é uma medida que quantifica o quanto os genes e as células estão associados a cada bicluster.

Em seguida, são identificados os membros de cada bicluster $b, 1 \leq b \leq p$ (um bicluster para cada fator $\mathbf{z}_b$). Para identificar os membros do b-ésimo bicluster, selecionamos os valores de λ_{ib} e z_{bj} que são maiores em valor absoluto que os pontos de corte t_l e t_z, respectivamente. Neste processo, é necessário normalizar a variância de cada fator, resultando na matriz de fator $\hat{\mathbf{Z}}$. Consequentemente, a matriz $\boldsymbol{\Lambda}$ é reescalada para $\hat{\boldsymbol{\Lambda}}$ de forma que $\boldsymbol{\Lambda}\mathbf{Z} = \hat{\boldsymbol{\Lambda}}\hat{\mathbf{Z}}$ e, então, o ponto de corte t_z deve ser escolhido para determinar, em média, a proporção de células (amostras ou observações) que pertencerá ao bicluster. Como o b-ésimo fator da j-ésima célula z_{bj} segue uma distribuição de Laplace com média 0 e variância 1, a probabilidade de que z_{bj} exceda o ponto de corte t_z é

$$\mathbb{P}\left(z_{bj} > t_z\right) = 1 - \mathbb{P}\left(z_{bj} \leq t_z\right) = \begin{cases} 1 - \frac{1}{2}e^{\sqrt{2}t_z} = p_0, & \text{se } t_z \leq 0 \\ \frac{1}{2}e^{-\sqrt{2}t_z} = p_0, & \text{se } t_z > 0, \end{cases} \tag{3.11}$$

em que p_0 denota a proporção esperada de células que será atribuída ao bicluster. Mesmo que esse ponto de corte seja atribuído pensando em z_{bj} ele será aplicado em $\hat{z}_{bj}$.

Como um gene pode estar apenas presente (expressão positiva) ou ausente (expressão negativa), o b-ésimo bicluster é determinado pelos valores positivos ou negativos de $\hat{z}_{bj}$. A escolha entre esses dois grupos é feita com base naquele que apresenta a maior soma dos valores cujos módulos satisfazem $|\hat{z}_{bj}| > t_z$.

O ponto de corte t_l é definido como $t_l = \frac{sd_{lz}}{t_z}$, em que:

$$sd_{lz} = \sqrt{\frac{1}{pnm} \sum_{(b,j,i)=(1,1,1)}^{(p,n,m)} \left(\hat{\lambda}_{ib}\hat{z}_{bj}\right)^2}. \tag{3.12}$$

Isso corresponde a extrair para cada bicluster os genes cujas cargas fatoriais, em valor absoluto, são maiores do que o t_l (Hochreiter *et al.*, 2010).

3.1.3 Definição dos hiperparâmetros

No modelo *FABIA* é necessário definir previamente o número de biclusters (ou fatores) p. Para isso, pode-se ajustar o modelo para diferentes valores de p e selecionar aquele que apresentar o melhor desempenho segundo um critério de qualidade dos biclusters.

Foram propostas várias medidas para medir a qualidade de um bicluster e direcionar objetivamente a escolha do p, contudo nenhuma é capaz de detectar biclusters com padrões *shifting* e *scaling* simultaneamente, que representam o cenário mais comum ao trabalhar com dados de expressão gênica (Pontes *et al.*, 2015). Diante dessa limitação e visando a identificação de biclusters de maior qualidade, foram selecionadas quatro medidas para auxiliar a escolha dos hiperparâmetros, que serão definidas adiante. No entanto, não é garantia que os biclusters selecionados tenham total sentido biológico, sendo necessário analisar também descritivamente os resultados obtidos.

Considere B um bicluster composto por um conjunto de m_1 genes e n_1 células. Seja b_{ij} a expressão gênica do gene i na célula j; $b_{\cdot j}$ a média das expressões gênicas na célula j; $b_{i\cdot}$ a média das expressões gênicas no gene i e $b_{\cdot\cdot}$ a média global das expressões gênicas no bicluster B.

Uma métrica muito utilizada para avaliar a qualidade de um bicluster é a *Variance* (VAR). Essa métrica é utilizada para medir a variância de um bicluster com o objetivo de encontrar biclusters cujos valores sejam similares (baixa variância), portanto ela detecta biclusters constantes. Ela é definida por:

$$VAR(B) = \sum_{i=1}^{m_1}\sum_{j=1}^{n_1}\left(b_{ij} - b_{..}\right)^2. \tag{3.13}$$

Uma segunda métrica é o *Mean Squared Residue* (MSR): Esse indicador utiliza a coerência dos genes e células para avaliar a qualidade de um bicluster, e é dada por:

$$MSR(B) = \frac{1}{m_1 n_1}\sum_{i=1}^{m_1}\sum_{j=1}^{n_1}\left(b_{ij} - b_{i.} - b_{.j} + b_{..}\right)^2. \tag{3.14}$$

Quanto menor a medida, maior a coerência do bicluster e maior sua qualidade. Ela detecta biclusters com padrões *shifting*, mas é ineficiente para detectar biclusters com padrões *scaling*.

Outra métrica é o *Scaling Mean Squared Residue* ($SMSR$): Essa métrica é similar à anterior, contudo, ela detecta biclusters com padrões *scaling* ao invés de biclusters com padrões *shifting*, e é definida por:

$$SMSR(B) = \frac{1}{m_1 n_1}\sum_{i=1}^{m_1}\sum_{j=1}^{n_1}\frac{\left(b_{i.}b_{.j} - b_{ij}b_{..}\right)^2}{b_{i.}^2 b_{.j}^2}. \tag{3.15}$$

Nela, quanto menores os valores, mais o bicluster se aproxima de um padrão *scaling* ideal, e melhor sua qualidade.

Uma quarta métrica é o *Virtual Error* (VE): Essa métrica avalia a qualidade dos biclusters com base na coerência de seus valores após um processo de padronização (Xu *et al.*, 2024). Ela é capaz de detectar tanto padrões *shifting* como *scaling* (não simultaneamente) e é dada por:

$$VE(B) = \frac{1}{m_1 n_1}\sum_{i=1}^{m_1}\sum_{j=1}^{n_1}\left|\hat{b}_{ij} - \hat{\rho}_j\right|, \tag{3.16}$$

em que $\hat{b}_{ij} = \frac{b_{ij} - b_{i.}}{\sigma_i}$, $\hat{\rho}_j = \sum_{i=1}^{m_1}\frac{\hat{b}_{ij}}{m_1}$, e σ_i é o desvio padrão da linha i do bicluster B.

Assim como as demais métricas, quanto menor o valor do *Virtual Error*, maior a qualidade do bicluster, que se aproxima de um padrão *shifting* ou *scaling*.

Essas métricas avaliam a qualidade de um único bicluster. Para avaliar a qualidade de um modelo ajustado com um conjunto de hiperparâmetros, serão utilizados os valores médios destas, calculadas em cada bicluster.

Capítulo 4

Aplicação dos métodos em dados cerebrais

Estudar os genes relacionados ao desenvolvimento do neocórtex humano é importante para compreender as habilidades cognitivas da nossa espécie e nossa susceptibilidade a doenças relacionadas ao neurodesenvolvimento (Camp *et al.*, 2015).

Devido às limitações éticas e práticas em experimentações com humanos, é comum o uso de ratos como modelo para estudar as relações genéticas no neocórtex. No entanto, por conta das diferenças evolutivas e fisiológicas entre as espécies, a relevância das conclusões obtidas a partir desses modelos é questionável. Por isso, torna-se fundamental investigar sistemas que reproduzam, de forma mais fiel, as características do cérebro humano.

Recentemente, estruturas auto-organizadas que se assemelham ao cérebro humano em desenvolvimento, chamadas organoides cerebrais, foram geradas a partir de células-tronco pluripotentes. Esses organoides permitem a formação de várias regiões semelhantes ao cérebro, como o córtex e o hipocampo.

Por serem modelos que buscam simular o desenvolvimento do cérebro humano, os organoides cerebrais apresentam uma ampla diversidade de tipos celulares. No entanto, a análise gênica desses organoides por meio de sequenciamento em massa fornece apenas uma visão geral da expressão gênica, sem conseguir distinguir quais genes estão associados a cada tipo celular. Para superar essa limitação, utiliza-se a análise de células únicas, que permite identificar diferentes tipos celulares com base em sua expressão gênica, incluindo células intermediárias em processos de diferenciação.

Assim, a análise de células únicas foi empregada, neste estudo, para identificar os genes mais expressos em cada tipo celular, com o objetivo de avaliar se os organoides conseguem,

de fato, reproduzir com fidelidade as características genéticas do cérebro humano.

A base de dados original é composta por 18.928 genes e 734 células, obtidass por sequenciamento de RNA de células únicas (scRNA-seq). As células tem origem no neocórtex fetal humano ou em organoides cerebrais humanos. Como 2.136 genes possuem valores iguais a 0 em todas as observações, eles foram removidos e a base de dados passou a ter 16.792 genes e 734 células.

Do neocórtex fetal humano, foram analisadas 164 células de fetos com 12 semanas após a concepção (81 do chip 1 e 83 do chip 2) e 62 células de fetos com 13 semanas. Já as células provenientes de organoides cerebrais foram derivadas de células-tronco pluripotentes induzidas (iPSC, linha 409B2) coletadas em 33 dias (40 células), 35 dias (68 células), 37 dias (71 células), 41 dias (74 células) e 65 dias (80 células) após o início da cultura. Além disso, foram incluídas regiões corticais microdissecadas de organoides derivados de células-tronco embrionárias H9 com 53 dias (48 células da região 1 e 48 da região 2) e de organoides 409B2 com 58 dias (43 células da região 3 e 36 da região 4) (Camp *et al.*, 2015).

Todos os códigos utilizados encontram-se disponíveis em um repositório público no GitHub, no link https://github.com/pedrimbal/TCC.

4.1 Análise descritiva dos dados

A seguir foi feita uma análise descritiva da base de dados, antes e depois de aplicar o método de normalização *Linnorm*, utilizando o pacote "Linnorm" do R (Yip *et al.*, 2017).

Pela Figura 4.1, observa-se que a maioria dos valores de expressão gênica são iguais a 0 (aproximadamente 77,58%), o que é esperado para dados genéticos. Os valores diferentes de 0, melhor visualizados na Figura 4.2, possuem uma distribuição assimétrica à direita.

Após aplicar o método de normalização *Linnorm* à base de dados a proporção de zeros permaneceu igual, pois esses valores não são afetados pelo procedimento de normalização. Além disso, observa-se, pelas Figuras 4.3 e 4.4 e pela Tabela 4.1, que a expressão gênica foi reescalada para valores entre 0 e 1.13, o desvio-padrão dos dados não mudou, e a assimetria da distribuição aumentou e mudou de sentido (de assimetria direita para assimetria esquerda). Esse comportamento sugere que o método aumentou a expressão gênica de genes que, antes da normalização, exibiam valores de expressão menores.

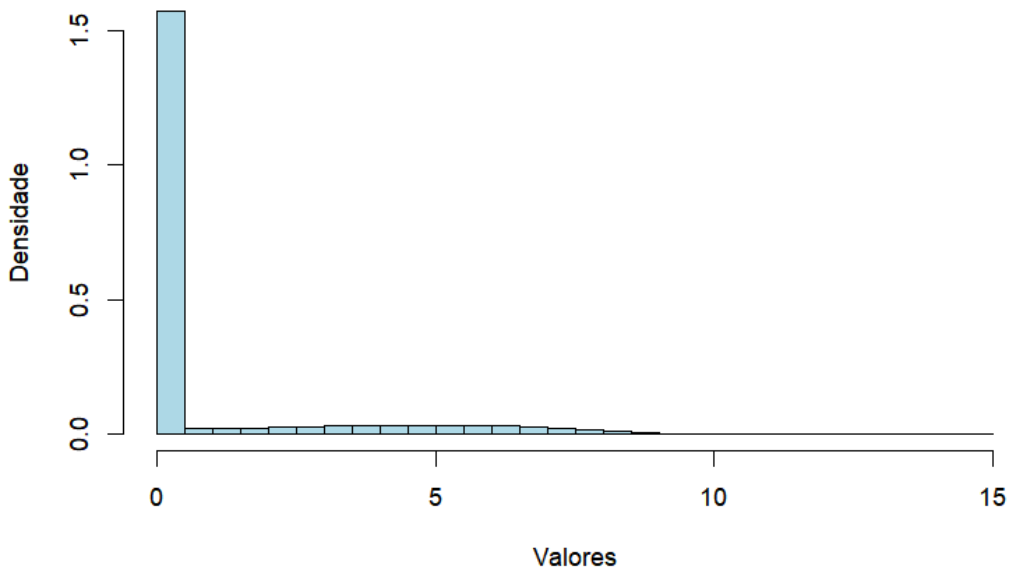

Figura 4.1: Histograma da expressão gênica.

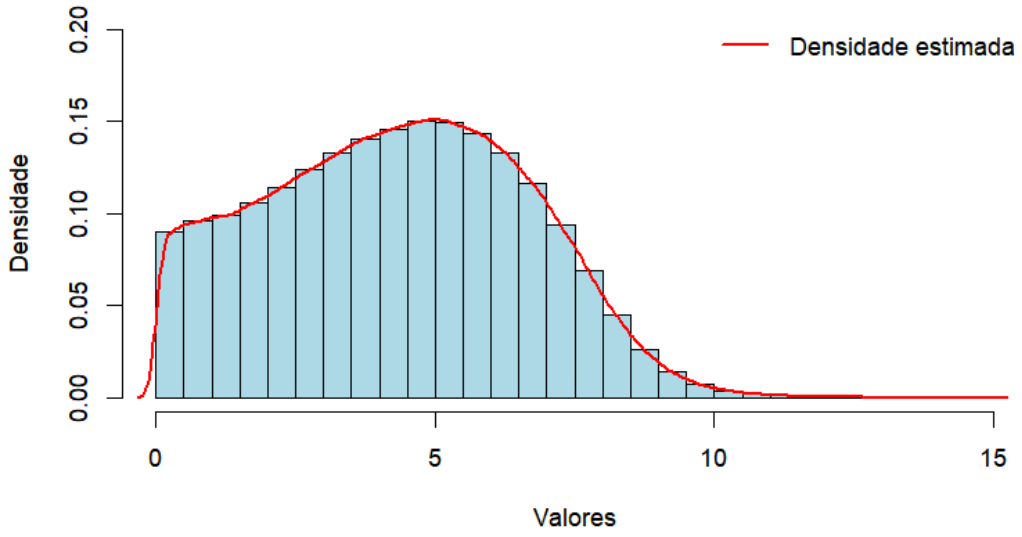

Figura 4.2: Histograma da expressão gênica sem valores iguais a 0.

Tabela 4.1: Estatísticas calculadas antes e depois da normalização (zeros removidos).

	Min	1°q	2°q	3°q	Max	Média	Desvio-padrão	Assimetria
Original	0.00001	2.48	4.34	6.04	14.92	4.3	0.17	0.08
Normalizado	0.00016	0.4	0.53	0.62	1.13	0.5	0.17	-0.56

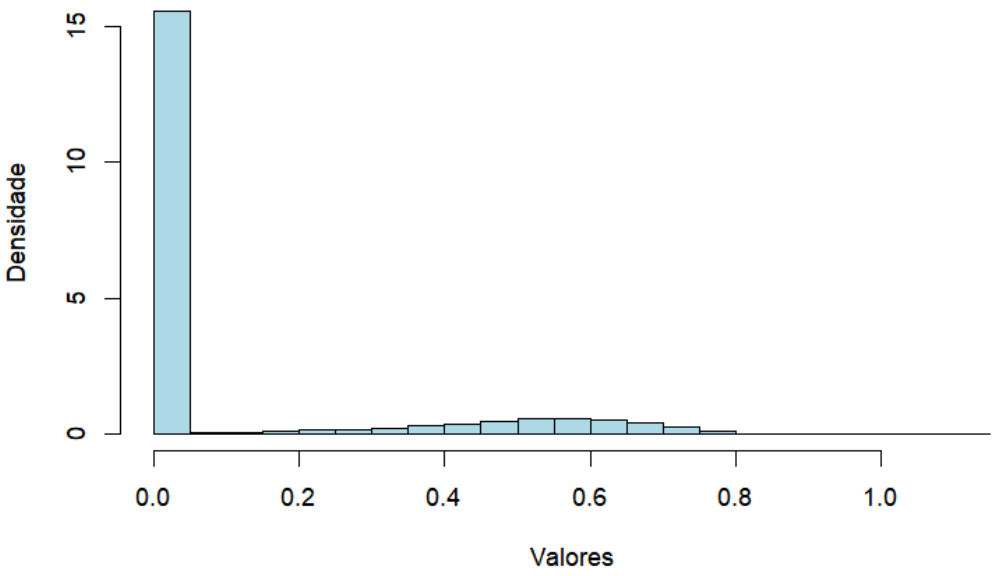

Figura 4.3: Histograma da expressão gênica normalizada.

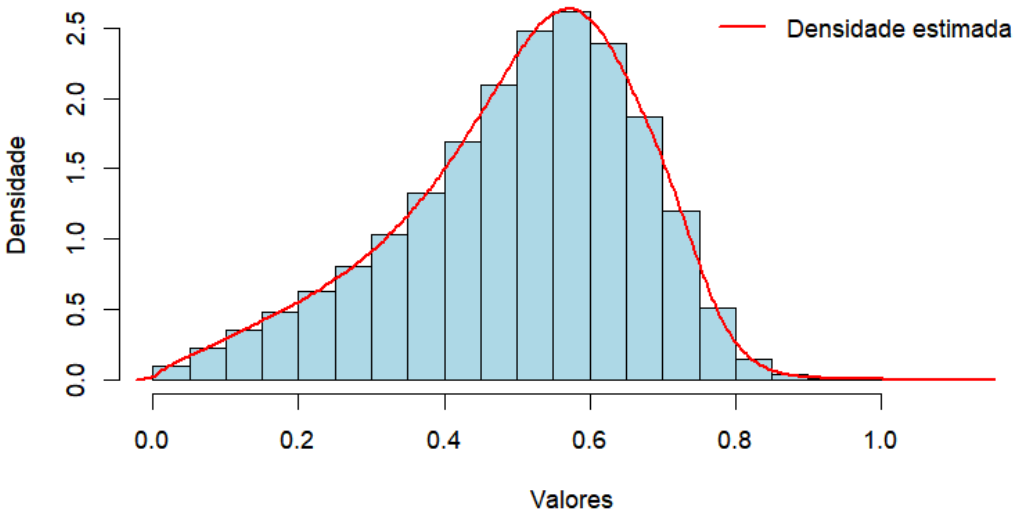

Figura 4.4: Histograma da expressão gênica normalizada sem valores iguais a 0.

A Figura 4.5 apresenta os histogramas da expressão gênica dos genes AAGAB, ABHD16A, ACTR1A, C9orf171 (linhas 11, 80, 200 e 2000 do banco de dados, respectivamente), selecionados de forma aleatória, antes e depois da normalização (excluindo os valores iguais a 0 para facilitar a visualização). Estes gráficos reforçam que a normalização aumenta os valores de baixa expressão, tornando a distribuição mais assimétrica à esquerda.

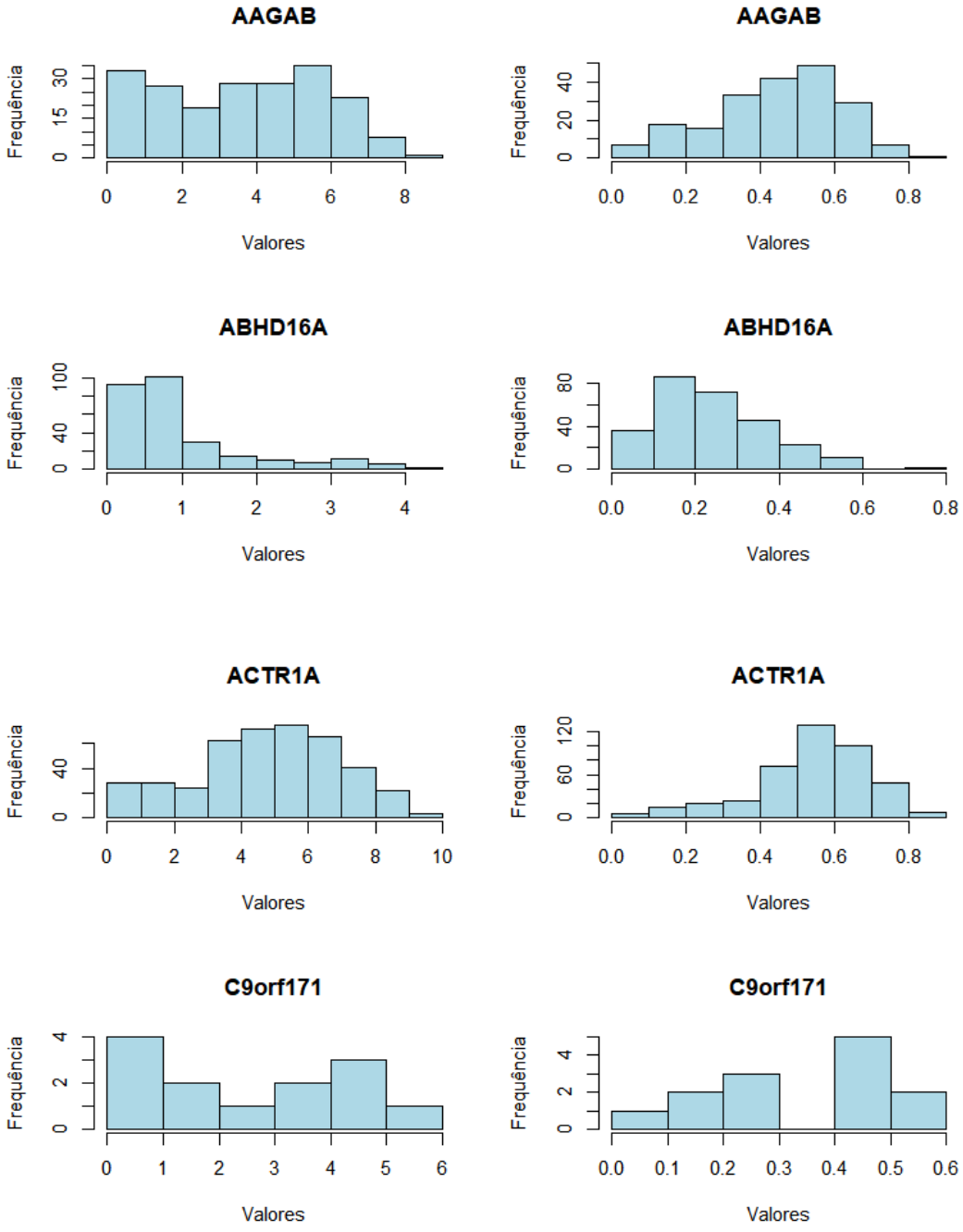

Figura 4.5: Histograma da expressão gênica de 4 genes, antes (à esquerda) e depois (à direita) da normalização.

Assim, conclui-se que o método de normalização *Linnorm* foi capaz de detectar genes pouco expressos, aumentando sua expressão gênica e preservando a variabilidade do conjunto de dados original. Por conta disso, as análises subsequentes foram feitas com o conjunto de dados normalizado.

4.2 Análise de biclusterização

No modelo *FABIA* é necessário definir alguns hiperparâmetros previamente, como o número de biclusters e a esparsidade, um fator de penalização para a estimação das cargas fatoriais utilizado no passo M do algoritmo EM variacional (Hochreiter *et al.*, 2010). Para isso, foram comparados modelos ajustados com diferentes combinações destes hiperparâmetros e foi selecionado aquele que apresentou melhor desempenho, segundo as métricas definidas anteriormente.

Foram comparados modelos variando o número de biclusters e o nível de esparsidade, em que o número de biclusters variou de 2 a 15 e a esparsidade assumiu os valores de 1%, 5%, 10% e 20%.

Pelas Figuras 4.6, 4.7, 4.8 e 4.9, observa-se que as métricas *VAR*, *MSR* e *VE* apresentam comportamentos semelhantes. Para um número reduzido de biclusters, os valores dessas métricas são elevados, aumentando conforme cresce o nível de esparsidade do modelo. À medida que o número de biclusters aumenta, os valores das métricas diminuem, estabilizando-se aproximadamente a partir de 6 biclusters. Nesse cenário, a relação com a esparsidade se inverte, de modo que maiores níveis de esparsidade passam a produzir menores valores para as métricas.

Por outro lado, o *SMSR* apresenta um comportamento contrário. Com poucos biclusters os valores são baixos, aumentando conforme o número de biclusters aumenta. Além disso, essa métrica estabiliza apenas para esparsidade de 1% e 5%, em torno de 6 biclusters.

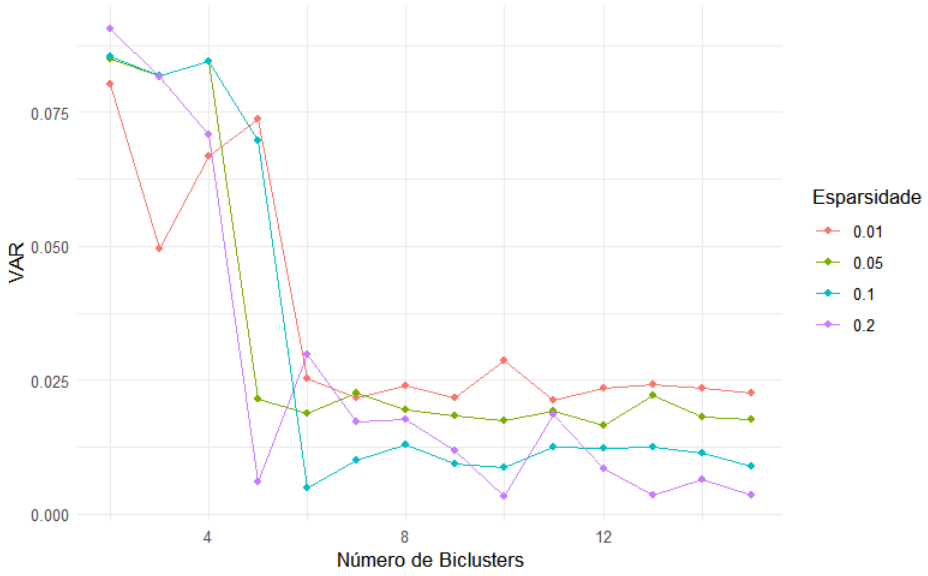

Figura 4.6: *VAR* média em cada modelo.

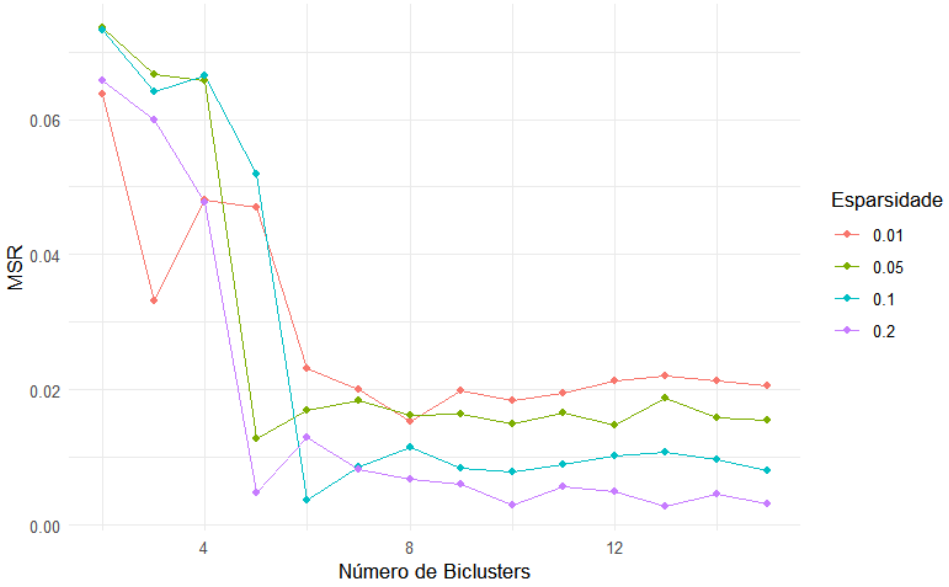

Figura 4.7: *MSR* médio em cada modelo.

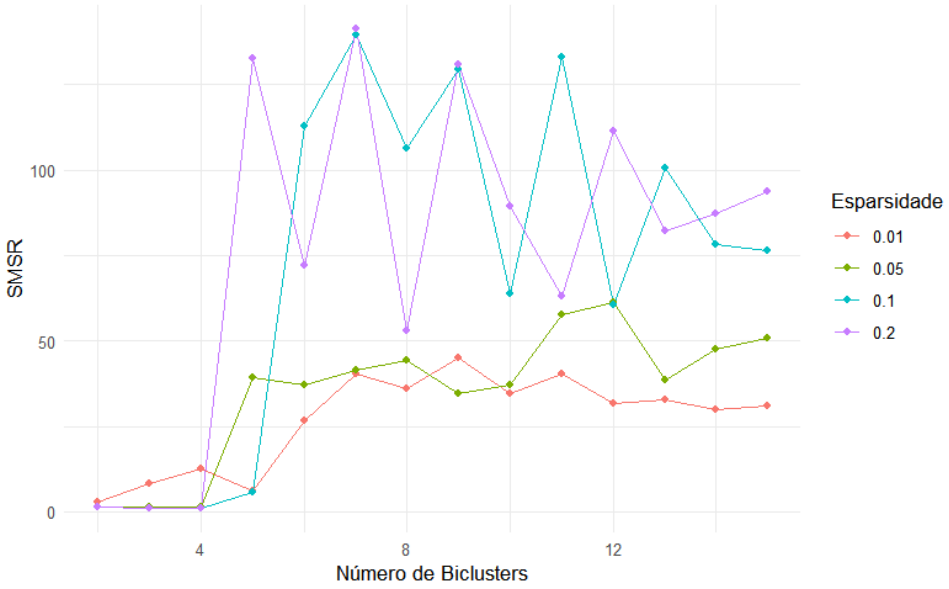

Figura 4.8: *SMSR* médio em cada modelo.

Também foram avaliados modelos ajustados aos dados originais, sem a aplicação de normalização. Entretanto, esses modelos apresentaram desempenho inferior em três das quatro métricas consideradas e, por esse motivo, foram descartados das análises subsequentes.

De acordo com Hochreiter *et al.* (2010), os dados devem ser centralizados de modo que apresentem média, moda ou mediana igual a zero. Como os dados analisados são inflados em zero, a moda e a mediana já assumem esse valor. Ainda assim, foram ajustados modelos utilizando os dados centralizados para média zero, a fim de avaliar o impacto desse

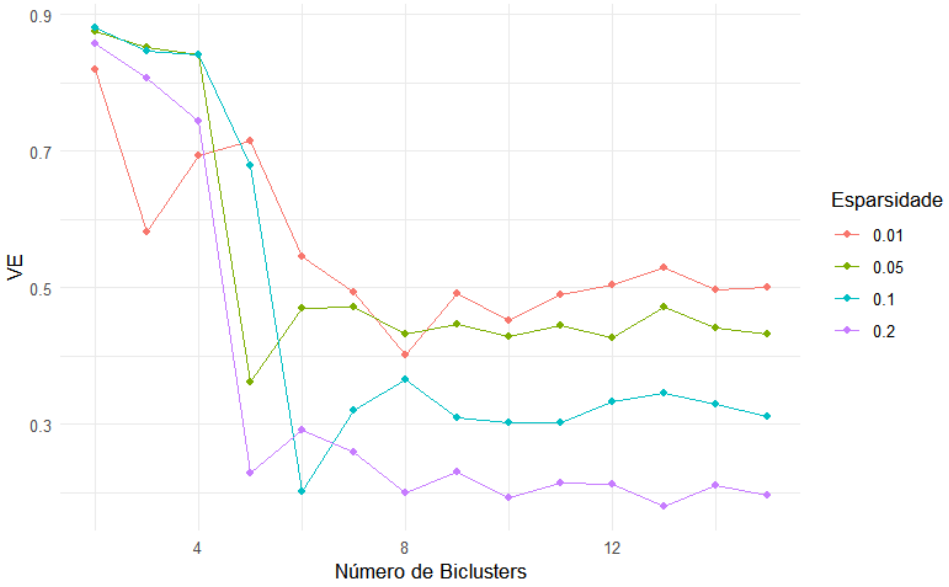

Figura 4.9: *VE* médio em cada modelo.

procedimento. Os resultados obtidos mostraram desempenho semelhante ao dos modelos ajustados sem centralização, indicando que essa etapa não trouxe ganhos relevantes para a qualidade dos biclusters identificados.

Com base nessas informações, foram analisados os modelos *FABIA* com 6 biclusters e esparsidade de 1% e 5%, ajustados nos dados normalizados e não centralizados, pois estes modelos possuem valores baixos e estáveis em todas as métricas observadas e, ao nosso ver, os melhores resultados de agrupamentos quando comparados com 7 ou 8 biclusters, também brevemente analisados, mas resultados não apresentados aqui.

4.2.1 Modelo com 6 biclusters e esparsidade de 1%

A Figura 4.10 mostra a informação que cada bicluster contém sobre os dados. Nela, destacam-se os biclusters 1, 2, 3 e 6, que explicam grande parte dos dados comparados aos biclusters 4 e 5. Essa informação é obtida através de conceitos da teoria da informação e melhor descrita em Hochreiter *et al.* (2010).

A Figura 4.11 apresenta o valor estimado das cargas fatoriais dos biclusters. Nota-se que, em todos os biclusters, o 1°, 2° e 3° quartil é igual ou muito próximo de 0, mudando apenas a distribuição dos *outliers* de um bicluster para o outro. Nos biclusters 2 e 3 eles estão concentrados em valores negativos, nos biclusters 1 e 6 em valores positivos e nos biclusters 4 e 5 estão concentrados em torno de 0.

A Figura 4.12 apresenta os valores absolutos das cargas fatoriais estimadas em cada

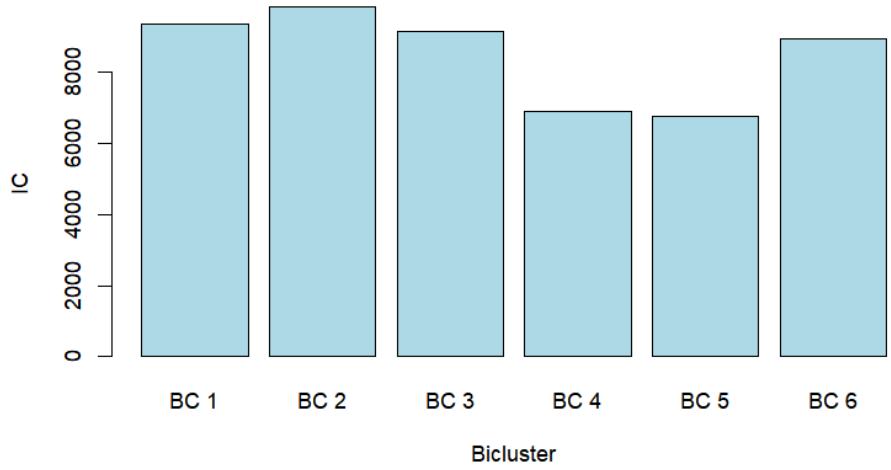

Figura 4.10: Informação de cada bicluster (esp. 1%).

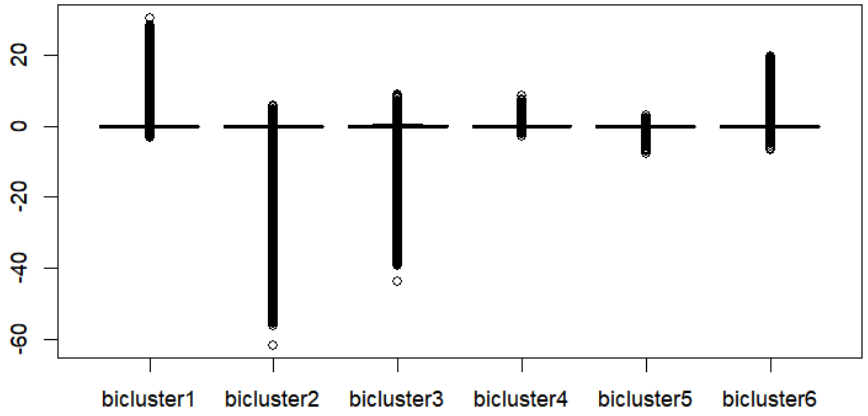

Figura 4.11: Boxplot das cargas fatoriais estimadas de cada bicluster (esp. 1%).

bicluster, os quais indicam os genes ativos em cada grupo. Observa-se que os genes com cargas diferentes de zero são praticamente os mesmos em todos os biclusters, variando principalmente em intensidade. Esse comportamento sugere que os biclusters compartilham conjuntos semelhantes de genes, diferenciando-se mais pela magnitude da contribuição de cada gene do que pela presença ou ausência deles.

A Figura 4.13 mostra a correlação entre as cargas fatoriais dos biclusters. Neste modelo, as cargas fatoriais possuem alta correlação (acima de 0.8 em valor absoluto) em todos os biclusters, reforçando que as cargas fatoriais estimadas em cada bicluster são associadas (negativamente ou positivamente), não necessariamente iguais ou muito

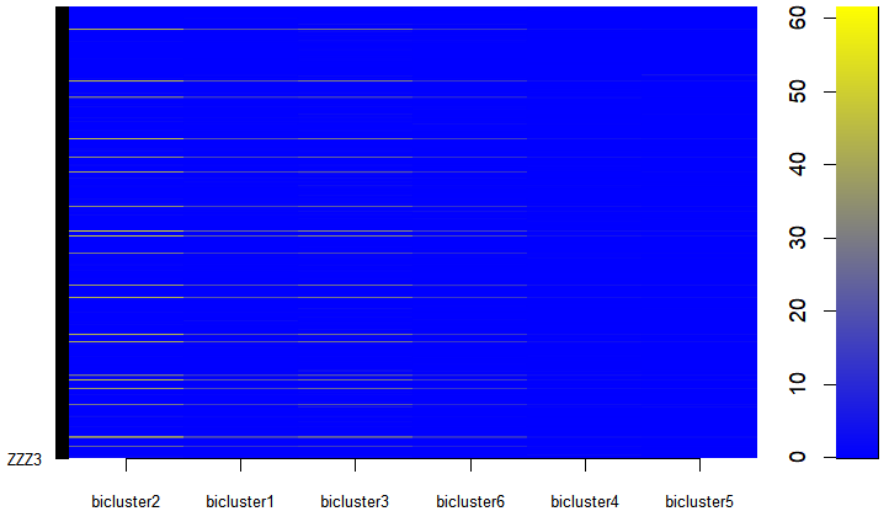

Figura 4.12: Valores absolutos das cargas fatoriais estimadas em cada bicluster (esp. 1%).

similares.

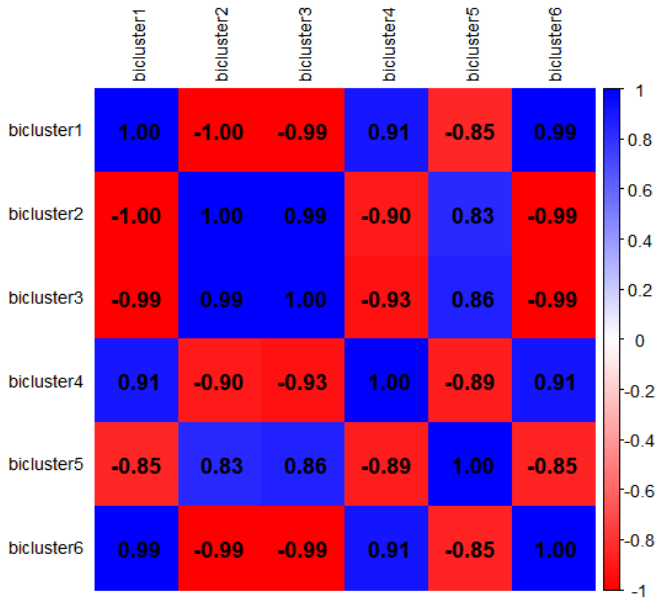

Figura 4.13: Correlação das cargas fatoriais dos biclusters (esp. 1%).

Pela Figura 4.14, podem ser observadas as distribuições dos fatores estimados para os biclusters. Nota-se que os biclusters 1 e 5 apresentam distribuições semelhantes, caracterizadas por mediana próxima de 0 e assimetria à esquerda. Nos demais biclusters, as distribuições são aproximadamente simétricas, o bicluster 2 possui mediana próxima de -1, os biclusters 3 e 4 apresentam medianas em torno de 0.5, enquanto o bicluster 6 possui mediana próxima de 0 e elevada presença de *outliers*.

A Figura 4.15 apresenta os valores absolutos dos fatores estimados em cada bicluster,

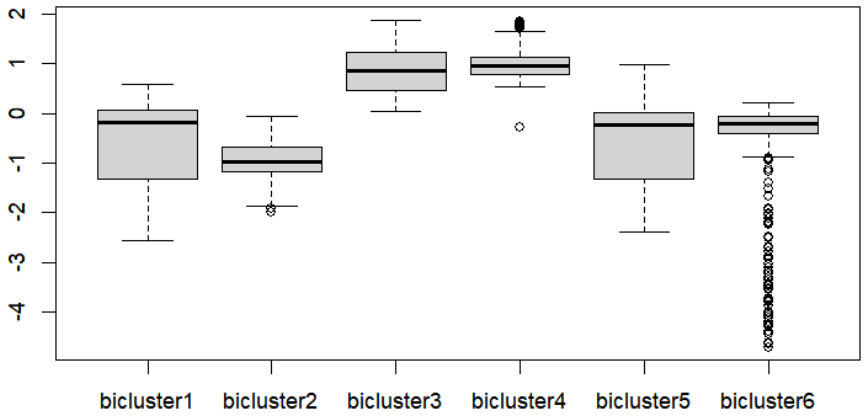

Figura 4.14: Boxplot dos fatores estimados de cada bicluster (esp. 1%).

que indicam as células pertencentes a cada grupo. Observa-se que os padrões dos fatores diferem consideravelmente entre os biclusters, indicando que cada um deles é composto por conjuntos distintos de células.

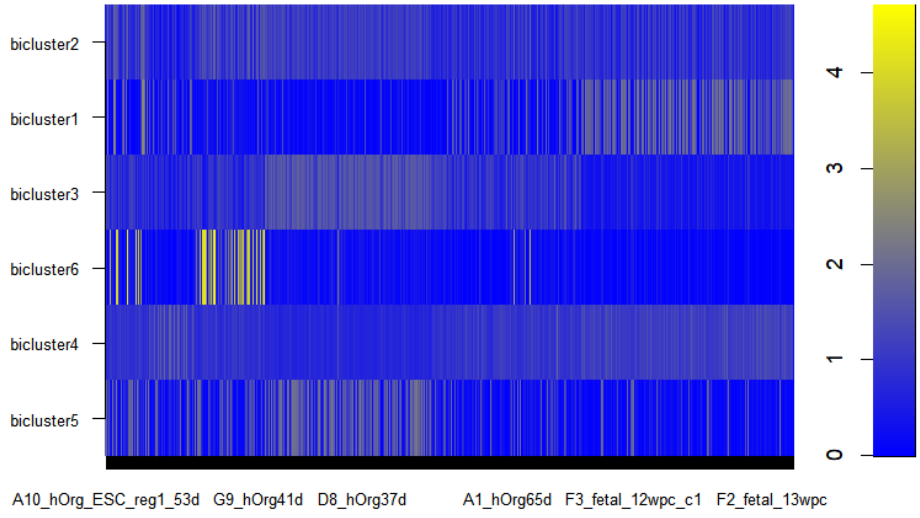

Figura 4.15: Valores absolutos dos fatores estimados em cada bicluster (esp. 1%).

A Figura 4.16 apresenta a correlação dos fatores. Nota-se que há uma correlação forte entre os biclusters 3, 4 e 5, que é positiva entre os biclusters 4 e 5 e negativa entre os demais. Por conta disso, as células presentes nestes biclusters podem ser similares.

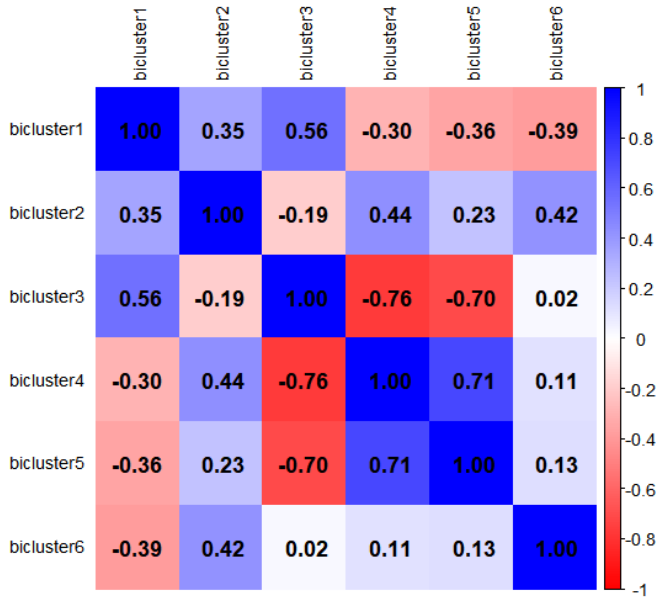

Figura 4.16: Correlação dos fatores dos biclusters (esp. 1%).

A Tabela 4.2 apresenta a composição dos biclusters segundo os tipos celulares descritos anteriormente. Observa-se que os biclusters 2 e 4 contêm no total de fetos (fetos_total), total de organoides pluripotentes (org_pluri_total) e total de organoides microdissecados (org_micro_total), mais de 80% da frequência relativa destas células, caracterizando-se como fatores gerais dos dados.

O bicluster 1 é composto majoritariamente por fetos (fetos_total, fetos_12s_c1, fetos_12s_c2 e fetos_13s), organoides derivados de células-tronco pluripotentes coletados após 65 dias de cultura (org_65d) e organoides provenientes de regiões corticais microdissecadas (org_micro_total, org_53_r1, org_53_r2, org_58_r3, org_58_r4). Esse resultado sugere que organoides de regiões microdissecadas e organoides pluripotentes maduros (mais dias) passam a compartilhar características com células fetais, diferenciando-se dos organoides pluripotentes coletados em períodos iniciais.

O bicluster 3 é formado predominantemente por células de organoides, distinguindo-as das células fetais e indicando que ainda existem características dos fetos que não são totalmente reproduzidas pelos organoides.

O bicluster 5 é composto principalmente por organoides pluripotentes coletados até 37 dias (org_33d_c1, org_33d_c2, org_35d e org_37d), diferenciando-os de organoides com 41 e 65 dias (org_41d e org_65d).

Já o bicluster 6 é composto, principalmente, por organoides pluripotentes coletados após 41 e 65 dias de cultura (org_41d e org_65d) e por organoides microdissecados da região 1 com 53 dias (org_53_r1). Esse padrão sugere que os organoides pluripotentes com 41 dias representam um estágio intermediário de transição entre organoides mais jovens e organoides mais maduros.

Tabela 4.2: Frequência e proporção de células pertencentes a cada bicluster (esp. 1%).

	BC 1	BC 2	BC 3	BC 4	BC 5	BC 6
fetos_total	178 (78.76%)	191 (84.51%)	60 (26.55%)	225 (99.56%)	30 (13.27%)	6 (2.65%)
fetos_12s_c1	60 (73.17%)	66 (80.49%)	25 (30.49%)	81 (98.78%)	13 (15.85%)	2 (2.44%)
fetos_12s_c2	63 (75.90%)	67 (80.72%)	24 (28.92%)	83 (100.00%)	11 (13.25%)	3 (3.61%)
fetos_13s	55 (90.16%)	58 (95.08%)	11 (18.03%)	61 (100.00%)	6 (9.84%)	1 (1.64%)
org_pluri_total	57 (17.12%)	322 (96.70%)	322 (96.70%)	333 (100.00%)	198 (59.46%)	94 (28.23%)
org_33d_c1	3 (12.50%)	24 (100.00%)	24 (100.00%)	24 (100.00%)	20 (83.33%)	1 (4.17%)
org_33d_c2	1 (6.25%)	16 (100.00%)	16 (100.00%)	16 (100.00%)	12 (75.00%)	2 (12.50%)
org_35d	9 (13.24%)	68 (100.00%)	68 (100.00%)	68 (100.00%)	55 (80.88%)	4 (5.88%)
org_37d	7 (9.86%)	71 (100.00%)	71 (100.00%)	71 (100.00%)	57 (80.28%)	9 (12.68%)
org_41d	8 (10.81%)	70 (94.59%)	67 (90.54%)	74 (100.00%)	35 (47.30%)	60 (81.08%)
org_65d	29 (36.25%)	73 (91.25%)	76 (95.00%)	80 (100.00%)	19 (23.75%)	18 (22.50%)
org_micro_total	60 (34.29%)	147 (84.00%)	152 (86.86%)	175 (100.00%)	65 (37.14%)	30 (17.14%)
org_53_r1	13 (27.08%)	45 (93.75%)	41 (85.42%)	48 (100.00%)	31 (64.58%)	25 (52.08%)
org_53_r2	15 (31.25%)	24 (50.00%)	36 (75.00%)	48 (100.00%)	15 (31.25%)	3 (6.25%)
org_58_r3	11 (25.58%)	42 (97.67%)	41 (95.35%)	43 (100.00%)	16 (37.21%)	1 (2.33%)
org_58_r4	21 (58.33%)	36 (100.00%)	34 (94.44%)	36 (100.00%)	3 (8.33%)	1 (2.78%)

A Tabela 4.3 mostra o número de genes, células e o percentual de cargas fatoriais não nulas em cada bicluster. Nela, destacam-se os biclusters 4 e 5, que não possuem nenhum gene identificado como relevante. Isso ocorre porque as cargas fatoriais estimadas destes genes (que estão mais concentradas em torno de 0) são todas menores que o ponto de corte t_l e, portanto, nenhum gene foi selecionado. Apesar disso, estes biclusters não possuem muitas cargas fatoriais iguais a 0 em comparação aos demais e, se necessário, um pódio dos genes com cargas fatoriais mais distantes de 0 pode ser passado aos geneticistas.

Nota-se também que o bicluster 2 é o que possui maior percentual de cargas fatoriais nulas, apesar de possuir o maior número de genes. Isso ocorre porque suas cargas fatoriais não nulas estão mais afastadas de zero, indicando maior relevância desses genes para a composição do bicluster.

Tabela 4.3: Número de genes, células e cargas fatoriais não nulas por bicluster (esp. 1%).

Bicluster	Número de genes	Número de células	Cargas fatoriais não nulas (%)
BC 1	974	295	69,8%
BC 2	1104	660	54,8%
BC 3	1041	534	74,2%
BC 4	0	733	65,4%
BC 5	0	293	66,8%
BC 6	744	130	67,5%

4.2.2 Modelo com 6 biclusters e esparsidade de 5%

A Figura 4.17 mostra a informação que cada bicluster contém sobre os dados. Nota-se que os biclusters 1, 4, 5 e 6 explicam grande parte dos dados em comparação aos biclusters 2 e 3.

Pela Figura 4.18, observa-se que as cargas fatoriais dos biclusters estão concentradas em torno de 0 e os *outliers* variam para valores maiores e menores que 0 em todos os biclusters, que se diferenciam pela variabilidade destes *outliers*. Além disso, os biclusters 2 e 3, que possuem menos informação que os demais, são biclusters cujas cargas fatoriais estão mais concentradas em torno de 0.

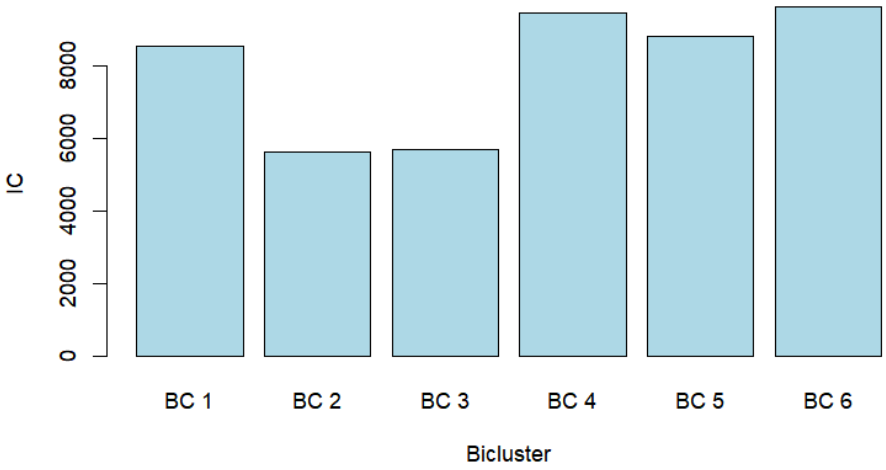

Figura 4.17: Informação de cada bicluster (esp. 5%).

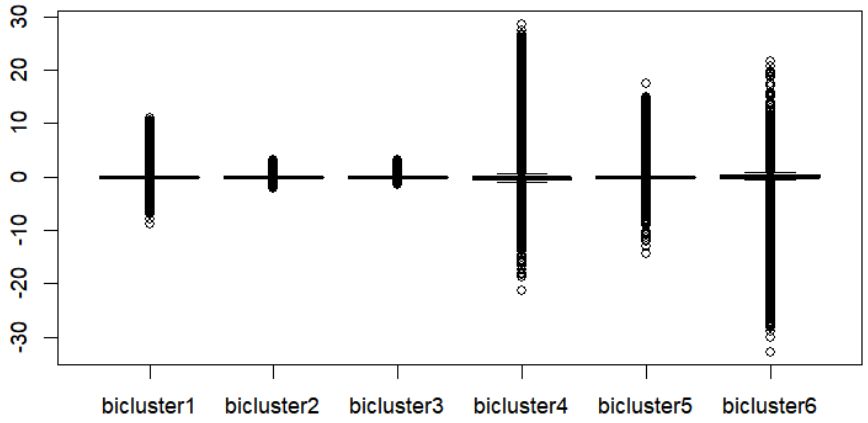

Figura 4.18: Boxplot das cargas fatoriais estimadas de cada bicluster (esp. 5%).

Pelas Figuras 4.19 e 4.20, observa-se que, diferentemente do modelo anterior, as cargas fatoriais não apresentam correlação alta entre todos os biclusters. Em particular, os biclusters 1, 4, 5 e 6 exibem fortes correlações entre si, enquanto os biclusters 2 e 3 formam um segundo grupo, também caracterizado por alta correlação interna. Por outro lado, as correlações entre esses dois grupos são substancialmente menores. Esse comportamento sugere a existência de dois conjuntos principais de genes representados pelo modelo, um associado aos biclusters 1, 4, 5 e 6, e outro associado aos biclusters 2 e 3.

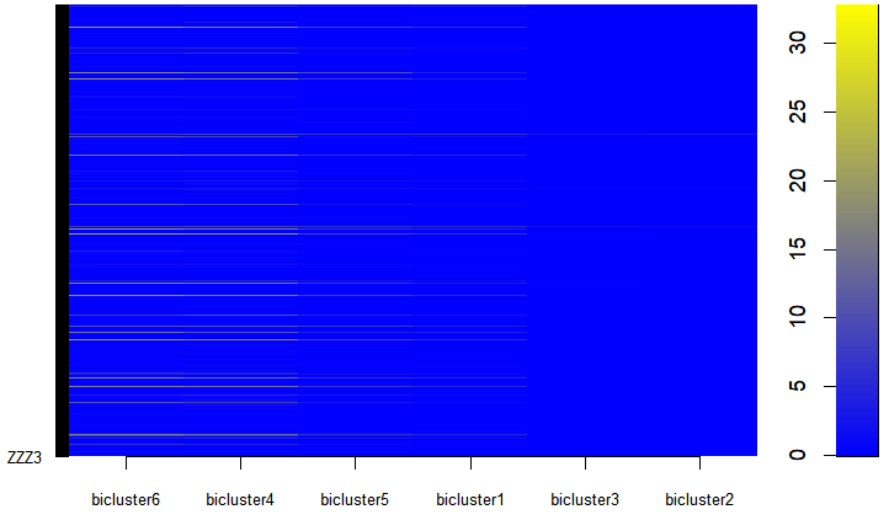

Figura 4.19: Valores absolutos das cargas fatoriais estimadas em cada bicluster (esp. 5%).

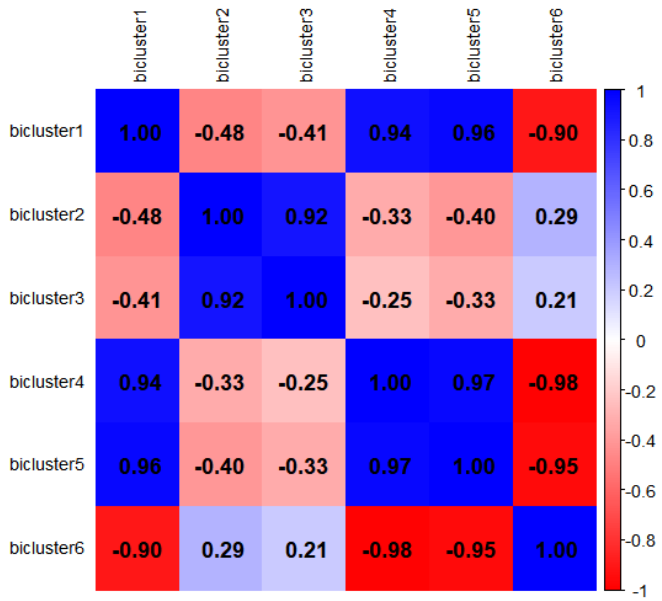

Figura 4.20: Correlação das cargas fatoriais dos biclusters (esp. 5%).

Pela Figura 4.21, observa-se que os fatores dos biclusters 2, 4, 5 e 6 possuem distribuições semelhantes, concentradas abaixo de 0, enquanto os demais possuem distribuições concentradas acima de 0, sendo o bicluster 3 aquele que possui maior variabilidade.

Pela Figura 4.22 notam-se três pares de biclusters (1 e 5, 2 e 3 e 4 e 6) cujos fatores (em valor absoluto) são parecidos, indicando que as células presentes nestes biclusters são semelhantes.

Pela Figura 4.23, observa-se a formação de dois grupos principais de fatores. O primeiro é composto pelos biclusters 1 e 5, que apresentam forte correlação negativa, indicando padrões opostos de ativação. O segundo é formado pelos biclusters 3, 4 e 6, que

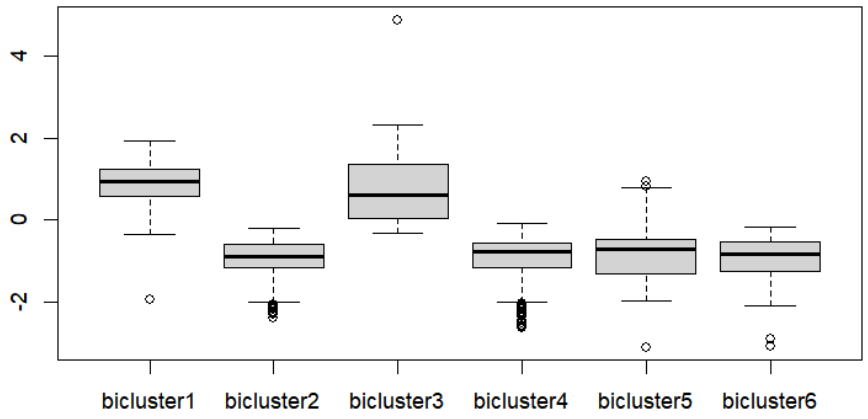

Figura 4.21: Boxplot dos fatores estimados de cada bicluster (esp. 5%).

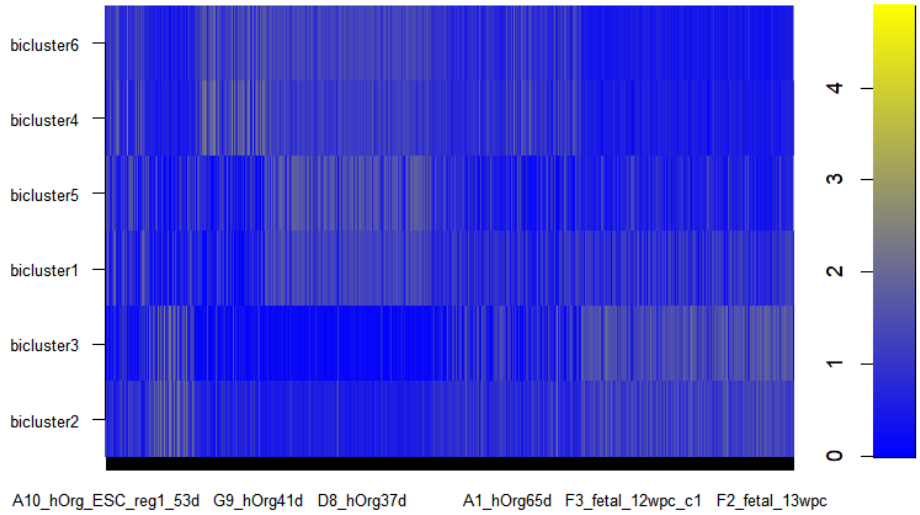

Figura 4.22: Valores absolutos dos fatores estimados em cada bicluster (esp. 5%).

exibem altas correlações entre si. Já o bicluster 2 apresenta correlações com ambos os grupos, sugerindo uma posição intermediária entre eles.

Pela Tabela 4.4, nota-se que os biclusters 1, 2 e 5 são fatores gerais dos dados, pois são compostos por quase todas as células, sem apresentar um agrupamento claro.

O bicluster 3, por outro lado, é composto por células de fetos (fetos_total, fetos_12s_c1, fetos_12s_c2 e fetos_13s), organoides pluripotentes com 65 dias (org_65d) e organoides de regiões microdissecadas (org_micro_total, org_53_r1, org_53_r2, org_58_r3, org_58_r4), reforçando que esses grupos possuem características semelhantes que os diferenciam de organoides pluripotentes mais jovens.

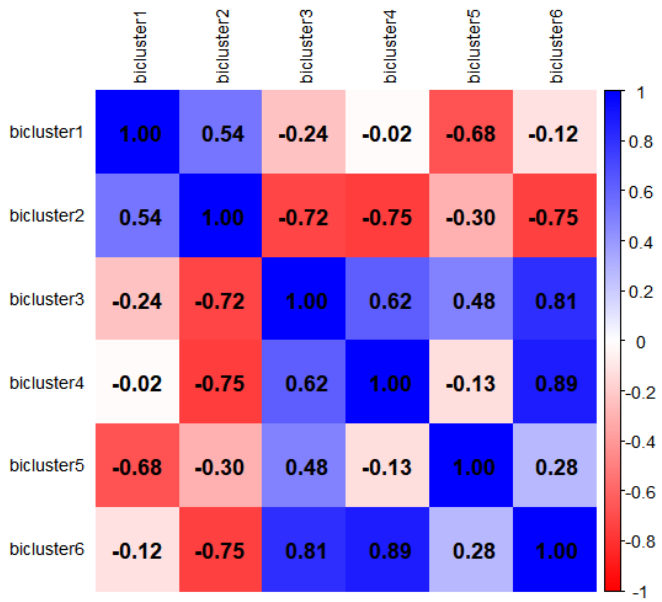

Figura 4.23: Correlação dos fatores dos biclusters (esp. 5%).

Além disso, os biclusters 4 e 6 são compostos majoritariamente por organoides, separando-os de fetos. Isso reforça a ideia de que existem características de fetos que não são capturadas pelos organoides.

Tabela 4.4: Frequência e proporção de células pertencentes a cada bicluster (esp. 5%).

	BC 1	BC 2	BC 3	BC 4	BC 5	BC 6
fetos_total	188 (83.19%)	226 (100.00%)	222 (98.23%)	129 (57.08%)	158 (69.91%)	84 (37.17%)
fetos_12s_c1	66 (80.49%)	82 (100.00%)	80 (97.56%)	47 (57.32%)	61 (74.39%)	38 (46.34%)
fetos_12s_c2	65 (78.31%)	83 (100.00%)	81 (97.59%)	49 (59.04%)	60 (72.29%)	31 (37.35%)
fetos_13s	57 (93.44%)	61 (100.00%)	61 (100.00%)	33 (54.10%)	37 (60.66%)	15 (24.59%)
org_pluri_total	273 (81.98%)	240 (72.07%)	69 (20.72%)	327 (98.20%)	249 (74.77%)	330 (99.10%)
org_33d_c1	23 (95.83%)	17 (70.83%)	2 (8.33%)	24 (100.00%)	23 (95.83%)	24 (100.00%)
org_33d_c2	15 (93.75%)	12 (75.00%)	2 (12.50%)	16 (100.00%)	15 (93.75%)	16 (100.00%)
org_35d	68 (100.00%)	46 (67.65%)	3 (4.41%)	68 (100.00%)	67 (98.53%)	68 (100.00%)
org_37d	69 (97.18%)	48 (67.61%)	3 (4.23%)	71 (100.00%)	67 (94.37%)	71 (100.00%)
org_41d	36 (48.65%)	48 (64.86%)	11 (14.86%)	71 (95.95%)	31 (41.89%)	74 (100.00%)
org_65d	62 (77.50%)	69 (86.25%)	48 (60.00%)	77 (96.25%)	46 (57.50%)	77 (96.25%)
org_micro_total	132 (75.43%)	166 (94.86%)	104 (59.43%)	153 (87.43%)	120 (68.57%)	154 (88.00%)
org_53_r1	38 (79.17%)	41 (85.42%)	20 (41.67%)	43 (89.58%)	36 (75.00%)	47 (97.92%)
org_53_r2	22 (45.83%)	48 (100.00%)	39 (81.25%)	32 (66.67%)	29 (60.42%)	33 (68.75%)
org_58_r3	39 (90.70%)	43 (100.00%)	19 (44.19%)	42 (97.67%)	37 (86.05%)	41 (95.35%)
org_58_r4	33 (91.67%)	34 (94.44%)	26 (72.22%)	36 (100.00%)	18 (50.00%)	33 (91.67%)

Pela Tabela 4.5 destacam-se os biclusters 2 e 3, que não possuem nenhum gene identificado como relevante neste modelo e possuem muitas cargas fatoriais iguais a 0 em comparação aos outros biclusters. Isso ocorre porque eles possuem cargas fatoriais muito concentradas em torno de 0, que são menores (em valor absoluto) que o ponto de corte t_l. No entanto, como o bicluster 3 se destaca pela baixa presença de organoides muito jovens, um pódio dos genes com cargas mais distantes de zero pode ser realizado e passado aos geneticistas.

Tabela 4.5: Número de genes, células e cargas fatoriais não nulas em cada bicluster (esp. 5%).

Bicluster	Número de genes	Número de células	Cargas fatoriais não nulas (%)
BC 1	390	593	47,6%
BC 2	0	632	28,1%
BC 3	0	395	26,1%
BC 4	1229	609	59,7%
BC 5	665	527	52,9%
BC 6	1235	568	60,2%

4.2.3 Comparação dos modelos

Entre estes dois modelos, o *FABIA* com 6 biclusters e esparsidade 1% se mostrou mais adequado, pois as cargas fatorias e os fatores estimados conseguem se diferenciar mais entre os biclusters, impactando em um modelo com maior discriminação dos genes e células pertencentes a cada bicluster.

Capítulo 5

Conclusão e estudos futuros

O objetivo deste trabalho é estudar, aplicar e analisar técnicas de normalização e biclusterização na análise de células únicas. Nele, foram destacadas as diferenças entre a análise de dados de células únicas e a de sequenciamento em massa, a importância de utilizar métodos de normalização para tratar o alto nível de ruído presente nesse tipo de dado e tornar a expressão gênica comparável entre diferentes genes de uma mesma célula, além do uso de técnicas de biclusterização para identificar, simultaneamente, genes relevantes no conjunto de dados e as células associadas a eles.

Também foi apresentada a teoria dos algoritmos de normalização *Linnorm* e de biclusterização *FABIA*, incluindo suas propriedades e vantagens em relação a outros métodos.

Por fim, uma base contendo dados genéticos do cérebro de fetos e organoides celulares foi descrita, analisada e utilizada para aplicação dos métodos estudados, com o objetivo de investigar se os organoides conseguem reproduzir as características do cérebro humano. Pela análise dos dados, concluiu-se que organoides maduros (coletados após 65 dias de cultura) provenientes de células-tronco pluripotentes e organoides provenientes de regiões corticais microdissecadas com, pelo menos, 53 dias, possuem características semelhantes à células de fetais. Contudo, a análise também indica que ainda existem caracteristicas de fetos que não são totalmente reproduzidas pelos organoides.

Como estudos futuros, pretendemos aplicar o modelo de biclusterização *FABIA* com outros métodos de centralização dos dados, avaliando o desempenho do modelo. Ademais, planejamos utilizar métodos de redução de dimensão dos dados (como o ACP, *UMAP* e *t-SNE*) para aplicar o *FABIA* utilizando menos variáveis, visto que a maior parte delas não foi relevante.

Referências Bibliográficas

Adil, A., Kumar, V., Jan, A. T. e Asger, M. (2021). Single-cell transcriptomics: current methods and challenges in data acquisition and analysis. *Frontiers in Neuroscience*, **15**, 591122.

Bacher, R., Chu, L.-F., Leng, N., Gasch, A. P., Thomson, J. A., Stewart, R. M., Newton, M. e Kendziorski, C. (2017). Scnorm: robust normalization of single-cell RNA-seq data. *Nature Methods*, **14**(6), 584–586.

Camp, J. G., Badsha, F., Florio, M., Kanton, S., Gerber, T., Wilsch-Bräuninger, M., Lewitus, E., Sykes, A., Hevers, W., Lancaster, M. *et al.* (2015). Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. *Proceedings of the National Academy of Sciences*, **112**(51), 15672–15677.

Castanho, E. N., Aidos, H. e Madeira, S. C. (2024). Biclustering data analysis: a comprehensive survey. *Briefings in Bioinformatics*, **25**(4), bbae342.

Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural computation*, **13**(11), 2517–2532.

Gong, Y., Xu, J., Wu, M., Gao, R., Sun, J., Yu, Z. e Zhang, Y. (2024). Single-cell biclustering for cell-specific transcriptomic perturbation detection in AD progression. *Cell Reports Methods*, **4**(4), 100742.

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W. *et al.* (2010). FABIA: factor analysis for bicluster acquisition. *Bioinformatics*, **26**(12), 1520–1527.

Law, C. W., Chen, Y., Shi, W. e Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, **15**, 1–17.

Love, M. I., Huber, W. e Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 1–21.

Pontes, B., Girldez, R. e Aguilar-Ruiz, J. S. (2015). Quality measures for gene expression biclusters. *PloS one*, **10**(3), e0115497.

Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B. e Knijnenburg, T. A. (2016). Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics*, **32**(17), i430–i436.

Predeus, A., Tavares, H., Murray, S., Kiselev, V., Andrews, T., Westoby, J., McCarthy, D., Büttner, M., Lee, J., Polanski, K., Müller, S. Y., Madissoon, E., Ballereau, S., Primo, M. D. N. L., Nunez, R. M. e Hemberg, M. (2022). Analysis of single cell RNA-seq data.

Saraiva, E. F. (2006). *Métodos estatísticos aplicados à análise da expressão gênica*. Dissertação de mestrado, Programa de Pós-Graduação em Estatística da UFSCar.

Wang, Z., Gerstein, M. e Snyder, M. (2009). Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, **10**(1), 57–63.

Xu, X., Zhang, S., Guo, J. e Xin, T. (2024). Biclustering of log data: Insights from a computer-based complex problem solving assessment. *Journal of Intelligence*, **12**(1), 10.

Yip, S. H., Wang, P., Kocher, J.-P. A., Sham, P. C. e Wang, J. (2017). Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Research*, **45**(22), e179–e179.