UNIVERSIDADE DE SÃO PAULO

Instituto de Matemática e Estatística

Programa de Pós-Graduação Interunidades em Bioinformática

**Deteccção de populações em dados de sequenciamento de RNA de
células individuais por meio de módulos de coexpressão**

Tiago Lubiana Alves

DISSERTAÇÃO PARA OBTENÇÃO DO TÍTULO DE
MESTRE EM CIÊNCIAS

Orientador: Prof. Dr. Helder Takashi Imoto Nakaya

São Paulo

2020

Detecção de populações em dados de sequenciamento de RNA de células individuais por meio de módulos de coexpressão

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 31/03/2020. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Prof. Dr. Helder Takashi Imoto (orientador) - FCF-USP
- Prof. Dr. André Fujita - IME-USP
- Dr. Israel Tojal da Silva - A.C. Camargo
- Prof. Dr. Eduardo Lani Volpi da Silveira - FCF-USP

Agradecimentos

A parte dos agradecimentos é, provavelmente, a parte mais poética das dissertações. O texto a seguir não contém ironia. São alguns dos agradecimentos de coração inteiro que senti justo destacar. Adiciono uma observação, para constar, que não há distinção de gênero nos agradecimentos a seguir.

Agradeço à minha mãe Edilcéia, ao meu pai Luiz Carlos, ao meu irmão André e a minha irmã Luíza e à minha família de forma geral pelo apoio perpétuo à minha vida, incluindo à minha vida acadêmica.. Agradeço enormemente ao Prof. Helder pelos direcionamentos acadêmicos, por me acolher em seu laboratório, e sempre zelar pelo bem estar emocional e profissional meu e de todos sob sua tutela. Agradeço também de forma especial os colegas do laboratório e profissão, atuais e já formados, pelas ajudas, correções, incontáveis momentos de parceria, apoio e tolerância.

Agradeço à sorte de estar vivo. Ao conjunto de acasos que levou ao surgimento do ser humano como entidade física e, muitos anos depois, à criação da ciência, da vida acadêmica e dos programas de mestrado e doutorado. Agradeço à uma série de antepassados que não desistiram e conseguiram se reproduzir, dando origem a todos que estão por aqui.

Esse trabalho não teria sido possível também sem a contribuição dos bilhões de seres humanos durante séculos para refrear o obscurantismo e dar espaço à razão. Sem a busca por conhecimento, que por sorte (ou quem sabe por seleção natural) está entranhada na natureza animal. Sem as pessoas que defendem, e avançam o pensamento humano em prol do coletivo. Agradeço, em especial, a duas materializações do espírito da busca pelo conhecimento: a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) e a Universidade de São Paulo. Esse estudo foi financiado principalmente pelo processo da FAPESP 2018/10257-2, e por essa oportunidade e voto de confiança sou profundamente grato. Agradeço também à Fundação Instituto de Pesquisas Farmacêuticas - FIP-FARMA projeto nº 87/Fundação Bill e Melinda Gates por apoiar este mestrado. Adicionalmente, o presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Agradeço aos produtores rurais que cultivaram os grãos, hortaliças e vegetais dos restaurantes universitários, aos trabalhadores da cozinha, aos que servem os pratos, aos trabalhadores que limpam, aos que ficam nas portas, aos nutricionistas que preparam o cardápio. Agradeço aos funcionários do centro de prática esportiva da USP. Agradeço aos demais funcionários da USP, até os cortadores de grama que volta e meia acertam pedras nos transeuntes, e podadores de árvore que cortam os galhos das árvores frutíferas. Até eles, pois contribuem para a harmonia de um campus que abraça o verde, e permite pequenas alegrias. Agradeço à quem plantou amoras, pitangas, goiabas, jambos-brancos e jamelões pelo campus. Agradeço aos vigias e funcionários da limpeza do INOVA-USP, que trabalham para o prédio funcionar tão bem. Agradeço também aos vigias da portaria da Vila Indiana, sempre alertas nas voltas noturnas do laboratório. É imensa a importância dessas pessoas na qualidade de vida e na produtividade da USP.

Agradeço aos professores e professoras que tive, da matemática ao violão. Nominalmente, agradeço aos professores Frederico e Fabiano, que lecionam ciências da vida no Ensino Médio, além de biologia e química, respectivamente. Aos professores das Ciências Biomédicas, pelas aulas e pela dedicação em montar o curso em que me formei cientista e me prepararam para essa etapa.

Agradeço ao apoio de tantos orientadores e orientadoras pelos últimos anos. Ao finado Prof. Moacélio, que me iniciou na vida acadêmica na FioCruz. Aos Profs. Sérgio Ferreira, Fernanda Felice e Julia Clarke por me receberem e orientarem na UFRJ. Ao Prof. Jean Pierre pela orientação no primeiro semestre de USP e tantos ensinamentos ao longo dos últimos seis anos. Ao Prof. Luiz Roberto Britto e ao Adilson; ao Prof. Roberto Pasquale; à Prof. Erica Kague; ao Prof. João Carlos Monteiro de Carvalho e ao João Vitor Dutra Molino e ao Prof. Chuck Farah e ao William Cenens por participarem na construção de minha maturidade científica na Universidade de São Paulo. À Prof. Michal Schwartz e à Aleksandra Deczkowska, do Instituto Weizman; ao Prof. Stephan Neuhauss e à Stella Glasauer, da Universidade de Zurique e a Prof. Andrea Pauli e à Jessica Stock, pelas oportunidades e ensinamentos em verões passados.

Agradeço aos amigos e amigas de uma vida. Aos do Clube de Biologia Sintética e do iGEM, aos do laboratório (sim, agradeço como colega e como amigo) aos do futsal, do basquete, do futebol de campo, de toda a Atlético Doze de Fevereiro, aos do CA da Biologia. Aos do Rio de Janeiro, que sempre estiveram presentes, apoiando, mesmo que eu não fosse capaz de explicar exatamente o que estive fazendo. E aos demais amigos e amigas pelo Brasil e pelo mundo, pois cada um, de sua forma, contribuiu para esse trabalho.

Por fim, agradeço à quem teve paciência para ler esses agradecimentos. Gostaria de dizer que não há espaço suficiente para agradecer a cada um. Todavia, garanto que todos foram lembrados e tem uma parte indelével nestes escritos.

Resumo

LUBIANA, T. **Detecção de populações em dados de sequenciamento de RNA de células individuais por meio de módulos de coexpressão**. 2010. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

O advento do sequenciamento de RNA de células únicas trouxe consigo avanços científicos e desafios técnicos. O desenvolvimento de novas metodologias de análise é uma etapa crucial para maximizar a extração de conhecimento desses dados. Neste trabalho, exploramos o reposicionamento de um algoritmo de seleção de características (*feature selection*) para tratar desafios analíticos de dados públicos de RNA de células únicas. Adaptamos o método FCBF (Filtro Rápido Baseado em Correlação, *Fast Correlation-Based Filter*) para selecionar genes relevantes para distinguir tipos celulares, e, à partir desta lista de genes, encontrar módulos de genes coexpressos. Em dados de células de sangue, notamos que os módulos encontrados correspondiam a programas transcricionais característicos de grupos celulares específicos. Em decorrência disso, implementamos um pipeline capaz de utilizar os módulos de coexpressão para inferir os tipos celulares presentes no conjunto de dados de forma multinível, evadindo os limites das rotulações únicas tradicionais. Processamos conjuntos de dados de células mononucleares de sangue humano e células embrionárias de peixe-zebra, observando que os módulos e populações encontradas traziam a luz informações biologicamente relevantes. Os algoritmos foram implementados em dois pacotes do Bioconductor, *FCBF* e *fcoex*, e estão disponíveis para comunidade, aumentando o arsenal para análise de dados de sequenciamento de RNA de células únicas.

Palavras-chave: sequenciamento de RNA de células únicas, coexpressão, tipo celular.

Abstract

LUBIANA, T. **Detection of populations in single-cell RNA sequencing data via coexpression modules..** 2010. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2010.

The advent of single-cell RNA sequencing has brought scientific advances and technical challenges. The development of new analysis methodologies is a crucial step to maximize the extraction of knowledge from this data. In this work, we explore the repositioning of a feature selection algorithm to address the analytical challenges of public single-cell RNA data. We adapted the FCBF method (Fast Correlation-Based Filter) to select relevant genes to distinguish cell types, and, from this list of genes, find modules of coexpressed genes. In blood cell data, we noted that the modules found corresponded to transcriptional programs characteristic of specific cell groups. As a result, we implemented a pipeline capable of using the coexpression modules to infer the cell types present in the data set in a multilevel way, avoiding the limits of traditional single labels. We processed data sets from human blood mononuclear cells and zebrafish embryonic cells, noting that the modules and populations found brought to light biologically relevant information. The algorithms were implemented in two Bioconductor packages, FCBF and fcoex, and are available to the community, increasing the arsenal for analyzing single-cell RNA sequencing data.

Keywords: single-cell RNA sequencing, coexpression, cell type.

Sumário

Lista de Abreviaturas	ix
Lista de Figuras	xi
Lista de Tabelas	xiii
1 Introdução	1
1.1 Transcritômica de células individuais	1
1.1.1 Pré-processamento	2
1.1.2 Agrupamento de células semelhantes	2
1.2 Seleção de características	3
1.2.1 Visão geral: embrulhadores, mergulhadores e filtros	4
1.3 Teoria da Informação	4
1.3.1 Entropia condicional ou equívoco: comparando variáveis	5
1.3.2 Ganho de informação e incerteza simétrica	6
1.3.3 FCBF: Filtro Rápido Baseado em Correlações	6
1.4 Redundância e coexpressão	7
1.4.1 Redes de coexpressão	7
1.4.2 Coexpressão em dados de células únicas	7
1.4.3 Módulos de coexpressão para encontrar populações	8
2 Objetivos	9
2.1 Objetivo principal	9
2.2 Objetivos adjacentes	9
3 Métodos	11
3.1 Dados usados	11
3.2 Categorização da expressão gênica	11
3.2.1 Binarização pela média ou mediana	11
3.2.2 Discretização por k-means	11
3.3 Identificação dos módulos <i>fcoex</i>	12
3.3.1 Filtrando genes por correlação com marcadores	12
3.3.2 Encontrando sementes dos módulos	12
3.3.3 Construindo os módulos / comunidades de coexpressão	12
3.4 Análise de sobre-representação	12
3.5 Identificação de módulos via WGCNA e <i>monocle3</i>	12

3.6	Análise de reagrupamento das células	12
3.7	Disponibilidade de código	13
3.8	Resumo do método desenvolvido	13
4	Resultados e discussão	15
4.1	FCBF como ferramenta de detecção de genes relevantes	15
4.1.1	Explorando discretizações alternativas	18
4.1.2	Observações de redundância como coexpressão	19
4.2	Módulos de coexpressão por <i>fcoex</i>	22
4.3	Análises comparativas aprofundadas	27
4.3.1	Comparação de pipelines para encontrar módulos de coexpressão	27
4.3.2	Comparação de incerteza simétrica e variância para seleção de bons genes . .	28
4.3.3	Comparação de algoritmos para agrupar as células com base nos módulos . .	29
5	Conclusões	31

Lista de Abreviaturas

FCBF	Filtro rápido baseado em correlação (<i>Fast Correlation-Based Filter</i>)
SU	Incerteza Simétrica (<i>Symmetrical Uncertainty</i>)
SP	Positivo para o gene semente (<i>Seed Positive</i>)
SN	Negativo para o gene semente (<i>Seed Negative</i>)
M(número)	Módulo de co-expressão e respectivo número identificador

Lista de Figuras

1.1	Entropia para variáveis binárias	5
3.1	Visão geral do método desenvolvido.	13
4.1	O conjunto de dados <i>pbmc3k</i>	15
4.2	Binarização baseada em limiar para dados de expressão de células individuais única.	16
4.3	O número de genes selecionados pelo FCBF depende do limiar de binarização	17
4.4	A binarização seguida pelo FCBF pode destacar genes relevantes.	17
4.5	Relação entre os genes encontrados pelo algoritmo de FCBF e os genes diferencialmente expressos	18
4.6	Métodos diferentes de discretização mudam o valor máximo das correlações.	18
4.7	CD3D e genes redundantes segundo o FCBF	20
4.8	CD79A e genes redundantes segundo o FCBF	21
4.9	Genes semente são robustos a variação do número inicial de grupos.	22
4.10	Módulos de coexpressão e enriquecimento funcional em PBMCs.	23
4.11	Os módulos gerados pelo <i>fcoex</i> agrupam funcionalmente populações de PBMCs.	24
4.12	Os módulos do <i>fcoex</i> permitem a identificação de tipos diferentes de monócitos.	25
4.13	Células divididas originalmente em 9 grupos, se reposicionam nos reagrupamentos do <i>fcoex</i>	25
4.14	Fcoex revela dinâmicas das células embrionárias do paulistinha (peixe-zebra).	26
4.15	Módulos gerados por outros pacotes agrupam funcionalmente populações de PBMCs.	28
4.16	A pontuação de enriquecimento para marcadores celulares é maior para incerteza simétrica em comparação com variância.	29
4.17	Agrupamento baseado em módulos leva a combinação puras das populações originais	30

Lista de Tabelas

4.1	Descrição dos módulos de coexpressão encontrados para cada pacote	27
4.2	Enriquecimento via GSEA para cada tipo celular no conjunto de dados <i>pbmc3k</i> . . .	29

Capítulo 1

Introdução

1.1 Transcritômica de células individuais

O nome “transcritômica” refere-se às técnicas que avaliam a expressão de diversos genes simultaneamente em tecidos ou células. Para isso, tecidos ou populações de células são lisados, o RNA total é extraído e a expressão é quantificada. Para quantificação, usam-se tecnologias com alta processividade, como o sequenciamento de RNA e os microarranjos, que geram milhares de pontos de dados por amostra. Por fim, os dados obtidos são analisados e, via de regra, depositados como um recurso público. Os dados de transcritômica, incluindo sua reanálise pela comunidade científica, são parte integral do arsenal de ferramentas da ciência atual.

Os dados da área de transcritômica são quantificações de RNA geradas a partir de populações. No caso de amostras de tecidos, em especial de tecido sanguíneo, muito das variações de expressão gênica encontradas advém da variação da proporção de tipos celulares nas amostras (MICHLMAYR et al., 2018; NEWMAN et al., 2015). Esse fator técnico acaba por esconder as flutuações de expressão gênica de células individuais. Apesar de haver métodos para deconvoluir ¹ os diferentes componentes celulares, como o CIBERSORT (MICHLMAYR et al., 2018) (NEWMAN et al., 2015), a resolução em nível de células únicas foi marcada, de fato, por evoluções técnicas na aquisição do dado, criando a subárea da transcritômica de células individuais.

O nome “sequenciamento de células individuais” (do inglês, *single-cell RNA seq*) refere-se a uma série de formas diferentes para quantificar transcritos célula a célula. A importância da técnica levou a revista Science a considerar a técnica a revolução (*breakthrough*) do ano de 2018 (BERG, 2018). De forma geral, a transcrição reversa do RNA em cDNA (DNA complementar) é feita de forma compartimentalizada, e cada célula é confinada com “códigos de barras”, sequências únicas adicionadas ao cDNA. Depois da marcação individual, o RNA das diversas células é aglomerado e o RNA é enviado para sequenciadores de nova geração. Com os “códigos de barras” únicos, é possível atribuir computacionalmente cada transcrito à cada célula. Por exemplo, tecnologias baseadas em gotas permitem sequenciamentos de centenas de milhares de células em um único estudo, separando de forma automatizada cada célula em uma gotícula individual contendo os reagentes para transcrição reversa e um “código de barras” único.

Os dados de RNA-seq de células individuais têm características gerais que os diferenciam dos de RNA-seq tradicional (LUECKEN; THEIS, 2019). Uma das diferenças é que o RNA-seq de células individuais apresenta maior multimodalidade, ou seja, o histograma da expressão de cada gene apresenta mais de um pico. Por exemplo, uma expressão bimodal para um gene pode apresentar um pico em “0” leituras e outro pico no nível de expressão mais comum. Isso contrasta com a distribuição “binomial negativa” da expressão gênica nos dados de RNA-seq tradicional (KORTHAUER et al., 2016).

Os dados de RNA-seq individual são também esparsos, ou seja, com grande número de zeros. Cada zero pode significar tanto um *drop-out*, um gene que estava expresso na célula mas não foi

¹Recuperar matematicamente uma informação que está presente na amostra, mas que não é clara dada a natureza do dado em questão.

identificado pelo sequenciamento, como um gene que não estava expresso nessa célula.

1.1.1 Pré-processamento

O pré-processamento de dados de RNA-seq individual consiste no alinhamento e mapeamento dos fragmentos de leitura a um genoma de referência e na normalização da tabela de expressão gênica. As metodologias de alinhamento são compartilhadas pelo processamento de células únicas e o tradicional, com a pequena diferença que cada transcrito de RNA-seq individual tem de ser mapeado pelo seu código de barras às células individuais.

O alinhamento produz uma tabela de expressão que passa por etapas de controle de qualidade. Se duas células aglomerarem-se e receberem um mesmo código de barras, os transcritos serão misturados, gerando “dupletos”. A fração de dupletos é, em geral baixa, e podemos aplicar métodos computacionais para removê-los (WOLOCK; LOPEZ; KLEIN, 2019). Além disso, células com baixa qualidade, como fragmentos de células ou células apoptóticas, introduzem ruído e podem ser removidas por filtros que retiram eventos² com poucos transcritos ou conteúdo mitocondrial elevado.

Após os filtros iniciais, o procedimento padrão é a conversão de “contagens” em “transcritos por milhão”, corrigindo por vieses de sequenciamento, seguida da transformação logarítmica dos dados. É relevante pontuar que existem críticas a essa normalização, pois o número absoluto de “contagens” para cada gene é correlacionado também com o tamanho da célula, com células maiores produzindo mais RNA total (MARGUERAT; BÄHLER, 2012) e a expressão de determinados genes está correlacionada com o tamanho da células (VARGAS-GARCIA; GHUSINGA; SINGH, 2018). De toda forma, ainda que subótima, essa normalização permite que os dados sejam analisados com sucesso para a maioria das aplicações.

Para as etapas seguintes, a tabela de expressão pode ser transformada por modelos de regressão linear que descontam vieses (como etapa do ciclo celular ou efeitos de lote). A seleção das técnicas apropriadas para cada conjunto de dados cabe ao analista. Ao final, obtém-se uma matriz de expressão A de dimensões $m \times n$ contendo m genes/transcritos e n eventos (células), e com cada item a_{ij} correspondendo à expressão normalizada do gene/transcrito i na célula j .

1.1.2 Agrupamento de células semelhantes

Após o processamento, uma tarefa fundamental é a anotação de cada evento em seus tipos celulares. No campo da citometria de fluxo, os tipos celulares são separados com base em janelas (no inglês, gates) de marcadores celulares conhecidos. Esse método é pouco eficaz para o sequenciamento de células, já que as janelas usadas não permitem a descoberta de novas populações (BUTLER et al., 2018) (SAEYS; VAN GASSEN; LAMBRECHT, 2016). O recurso mais usado é o agrupamento não supervisionado, determinando conjuntos de eventos (células) que são parecidos entre si para identificação de um rótulo ad hoc.

O grande número de genes sequenciados (na casa de dezenas de milhares) dificulta computações e inclui diversas fontes de “ruído”. Sendo assim, é comum que os dados sejam transformados por técnicas de redução dimensional, como a análise de componentes principais (PCA, do inglês *Principal Component Analysis*). Cada componente principal é uma combinação linear dos valores de expressão que reflete uma faceta das distâncias entre as células nesse espaço multidimensional de milhares de genes. A tabela de p componentes principais ($p \ll m$ genes/transcritos) por n eventos guarda grande parte da informação inicial, com menos custos computacionais envolvidos. Essa tabela, então, é usada como ponto de partida para os algoritmos de agrupamento.

O algoritmo de agrupamento mais usado constrói redes de proximidade célula-célula e aplica técnicas de detecção de subcomunidades. Uma distância (ex. distância euclidiana) é calculada e cada célula é ligada às suas k vizinha mais próximas. Ao grafo de k vizinhas mais próximas (k) é

²Eventos, aqui, são as observações atribuídas a células individuais pelo método de análise. Idealmente, cada evento corresponde a uma célula única. Os termos “evento” e “célula” são usados, em geral, de forma intercambiável na literatura da área, apesar das diferenças conceituais.

aplicado o algoritmo de detecção de módulos de Louvain. Essa abordagem foi descrita no programa PhenoGraph (LEVINE et al., 2015) e é a base dos algoritmos de agrupamento usados no popular pacote em R *Seurat* (BUTLER et al., 2018). O número de agrupamentos encontrado varia com parâmetros do modelo, incluindo parâmetros internos do método de Louvain e o número k de células vizinhas. No pacote *Seurat*, há apenas um parâmetro chamado de resolução (*resolution*), que é inversamente proporcional ao número de agrupamentos encontrados.

Em resumo, as células são agrupadas baseadas em todos os genes do transcrito conjunto, mas condensados na forma de componentes principais. O número de grupos encontrados, então, é um parâmetro rígido determinado pelo analista. Após a determinação dos grupos, cabe ao analista determinar os tipos celulares na amostra. Apesar de agrupamentos não serem tipos celulares, de forma geral, células de tipos parecidos agrupam-se (LUECKEN; THEIS, 2019). A discussão do que define um tipo celular é profunda, mas nas análises de transcritomas de células únicas, os termos “população”, “agrupamento” (*cluster*) e “tipo celular” são frequentemente usados de forma intercambiável. Genes que diferenciam um grupo dos outros (“marcadores”), são usados como guia para mapear os grupos aos tipos celulares da literatura. Por exemplo, um grupo que expresse, em média, muito mais CD3D que os demais grupos na amostra será denominado como grupo de células T.

Em muitos estudos, as combinações de marcadores são usadas para purificação por citometria de fluxo, permitindo análises bioquímicas (VILLANI et al., 2017), ou para identificação *in situ* por imuno-histoquímica. A identificação dos tipos também serve a objetivos “cartográficos”, mapeando as células presentes em um tecido. Isso é importante para a construção de modelos teóricos de dinâmicas celulares e para evidenciar os níveis fisiológicos de cada tipo celular. Conhecer níveis fisiológicos permite que anomalias sejam utilizadas como diagnóstico de problemas, como no caso clínico do hemograma, usado para aferir a concentração de plaquetas, linfócitos e outros glóbulos brancos.

Concluimos que a identificação das classes de células das amostras é essencial para o aproveitamento completo da transcritômica de células únicas. Como mencionado anteriormente, um problema para o agrupamento de células é a presença de milhares de genes na amostra, o que aumenta consideravelmente o tempo de processamento computacional dos dados. Uma das formas de acelerar esse processo é pela seleção de genes importantes e, para isso, é possível usar ferramentas de seleção de características. Tais ferramentas serão discutidas na próxima seção.

1.2 Seleção de características

Devido ao enorme número de genes detectados em cada célula (em torno de vinte mil por experimento), a análise de dados transcritômicos entra no contexto de “big data”, dando origem à ciência de dados de células únicas (*single-cell data science* (LÄHNEMANN et al., 2020)). A seleção de genes de interesse é rotineira tanto para dados de RNA-seq individual quanto para a transcritômica clássica. Como um exemplo, a ferramenta de redes de coexpressão *WGCNA*, que seleciona os genes com maior variância do conjunto de dados. Nesse caso, a premissa usada é que os genes que variam mais fazem isso em grande parte por causa de variação real, biológica, e por isso são mais interessantes.

O campo de seleção de características (do inglês, *feature selection*) é uma área do aprendizado de máquinas que dedica-se a pensar nas melhores formas de escolher variáveis. Nesse campo, os autores buscam formas rigorosas de determinar importância de variáveis, visando otimizar alguma tarefa de classificação, seja em termos de desempenho, seja em termos de entendimento do modelo treinado.

A remoção de características não-informativas ou redundantes reduz o ruído e o custo computacional de processamento, frequentemente levando a classificadores com melhor performance. Neste trabalho, adaptamos um algoritmo de seleção de características (o filtro rápido baseado em correlação, ou *fast correlation-based filter*, FCBF) para o contexto do RNA seq de células únicas. Partimos do princípio que os fatores capazes de separar os grupos celulares uns dos outros na amostras são

os mais interessantes biologicamente. Para isso, discutimos aqui os diferentes algoritmos do tipo e onde o algoritmo que abordamos se enquadra nesse contexto.

1.2.1 Visão geral: embrulhadores, mergulhadores e filtros

Os métodos de seleção de características dividem o objetivo de otimização da classificação por algoritmos de aprendizado estatístico. Além dessa faceta comum, eles podem ser subdivididos em três grandes tipos (tradução livre): embrulhadores (*wrappers*), mergulhadores (*embedders*) e filtros (*filters*). Os métodos embrulhadores selecionam características com base na performance de classificadores específicos, mantendo as características que resultam, por exemplo, em maior acurácia, precisão ou outra métrica de qualidade de classificação. Um exemplo dessa classe são as regressões passo-a-passo. Quando são feita “para frente” (*forward selection*), a característica que melhor consegue separar as classes sozinha é adicionada. Depois a segunda, a terceira e assim, sucessivamente, até o número escolhido pelo analista. Quando feita para trás, todas as características são usadas, e remove-se aquela que, ao sair, leva ao melhor desempenho de classificação, processo que é repetido também de forma iterativa. Os mergulhadores (*embedders*) inserem ferramentas matemáticas dentro do algoritmo classificador para remover características menos úteis. Eles baseiam-se no princípio de regularização, uma abordagem que penaliza os pesos de cada variável, tentando sempre reduzir os valores. O mecanismo mais conhecido, a penalização linear, também chamada de L1 ou Lasso (TIBSHIRANI, 1996), quando aplicado, pode levar o peso de algumas variáveis a 0. Isso faz que elas sejam removidas, pois não contribuem mais para o modelo final. No campo da transcritômica, um exemplo conhecido de penalização por lasso está no algoritmo de sPLS ou quadrados mínimos parciais esparsos (*sparse Partial Least Squares*) (LÊ CAO; BOITARD; BESSE, 2011), que forma a base do pacote em R mixomics para integração de dados multiômicos (ROHART et al., 2017). Por fim, os filtros (*filters*) extraem características independentemente do algoritmo de classificação. Por exemplo, a seleção dos genes com maior variância é um passo de filtro de diversos fluxos de trabalho em transcritômica. É possível também ordenar as características pela na correlação com um alvo, como a condição biológica de interesse. O algoritmo do Filtro Rápido Baseado em Correlação, objeto de estudo deste trabalho, pertence a essa classe. É um método baseado em conceitos da teoria da informação, que serão discutidos na seção a seguir

1.3 Teoria da Informação

A área da Teoria da Informação, que fundamenta o algoritmo base deste projeto, foi introduzida em um trabalho seminal de Claude Shannon, em 1948, intitulado “*A Mathematical Theory of Communication*” (“Uma Teoria Matemática de Comunicação”) (SHANNON, 1948). Esse estudo apresentou a base para a análise quantitativa de informação, usada em engenharia da computação. A definição de “bits”, uma popular unidade de quantificação de informação, deriva da sua definição de entropia H para eventos binários (dado que o logaritmo utilizado esteja em base 2):

$$H(p) = -p \log p + (1 - p) \log(1 - p) \quad (1.1)$$

Na equação acima, H é a função que define entropia baseada em p , a probabilidade de um acontecimento binário acontecer. H , então, é a função de entropia de Shannon ou entropia na teoria da informação. É importante ressaltar que a entropia da teoria de comunicação não é equivalente, física ou matematicamente, à entropia descrita pela segunda lei da termodinâmica, sendo apenas homônimos. A entropia, na teoria da informação, serve para medir incerteza. Ao jogarmos uma moeda honesta, cada opção (cara ou coroa) tem probabilidade de 50% ($p = 0,5$). Isso representa uma entropia de 1 bit. Se tivermos uma moeda viciada, enviesada mais para cara ou coroa, a incerteza sobre o resultado é menor, sendo mais fácil prever o que vai acontecer em cada lançamento. No caso extremo, no qual uma moeda cai sempre “cara” ou sempre “coroa”, a entropia vai ser nula. Isso é perceptível pela equação 1: na primeira metade, o logaritmo de $p = 1$ vai ser 0 e, na segunda,

Figura 1.1: Exemplo de como a entropia de um resultado binário varia de acordo com a probabilidade. A Figura foi gerada implementando a equação 1 em R e aplicando-a a diferentes valores de p entre 0 e 1, com intervalos de 0,25. A probabilidade de um evento encontra-se, por definição, entre 0 e 1.

$1-p$ vai ser também igual a 0. Como podemos ver na figura 1.1, a incerteza aumenta conforme a probabilidade se aproxima de 0,5.

A entropia também é bem definida para categorias com mais de duas classes. A equação abaixo (1.2) apresenta essa generalização, na qual i representa cada classe diferente da variável.

$$H(p) = - \sum p_i \log p_i \quad (1.2)$$

1.3.1 Entropia condicional ou equívoco: comparando variáveis

Para fazer a seleção de características para algoritmos classificadores, muitas vezes almeja-se remover variáveis redundantes entre si. Mas como medir redundância?

A entropia condicional fornece uma solução para esse problema, representando a incerteza de uma determinada medida, sendo uma outra conhecida. Tomando-se por exemplo duas moedas honestas grudadas, de forma que quando uma cai “cara”, a outra cai “cara”. A entropia individual de cada moeda é 1 bit. Porém, se o resultado de uma for conhecido, o resultado da outra também o será. Logo, a entropia condicional de uma, dado o conhecimento da outra, é de 0 bit.

Supondo-se que, uma vez em cada 10, ao lançar as duas moedas, elas desgrudam. Sabendo-se que o resultado de uma é “cara”, não é possível dizer com certeza o resultado da outra, mas é possível afirmar que a chance de ser cara também é maior que 0,5. Logo, a entropia condicional estaria entre 0 e 1. Por fim, para duas moedas independentes, a entropia condicional é exatamente igual à entropia, já que a incerteza não é alterada por se saber o resultado da outra moeda. Isso vale para todos os eventos independentes. Shannon (SHANNON, 1948) define a entropia condicional da seguinte forma:

“Definimos a entropia condicional de y [...] como a média da entropia de y para cada valor de x , ponderada de acordo com a probabilidade de obter esse x específico. [...] Essa quantidade mede quão incertos somos de y em média quando sabemos x ”.

$$H(Y|X) = \sum p_x H(Y|X = x) \quad (1.3)$$

Numa equação na qual Y e X são duas variáveis aleatórias e $H(Y|X)$ representa a entropia condicional de Y, dado que o valor de X é conhecido.

1.3.2 Ganho de informação e incerteza simétrica

Uma variação do conceito de entropia condicional é o ganho de informação ou informação mútua, que trata de quantificar o quanto de certeza ganha-se sobre Y quando se descobre o valor de X. Esse aumento de confiança sobre o valor de Y é refletido pela redução do valor da entropia, que quantifica a incerteza. A informação mútua é definida como a entropia de X menos a entropia condicional de X quando os valores de Y são conhecidos:

$$IG(X; Y) = H(X) - H(X|Y) \quad (4) \quad (1.4)$$

Na equação acima, IG representa o ganho de informação (também conhecido como informação mútua), enquanto X e Y são variáveis e $H(X|Y)$ é a entropia de X dado Y. É possível facilitar mais as análises, normalizando os resultados entre 0 e 1 usando uma métrica chamada de “incerteza simétrica”.

$$SU(X, Y) = 2 \frac{IG(X; Y)}{H(X) + H(Y)} \quad (5) \quad (1.5)$$

Na equação acima, SU representa a incerteza simétrica, que divide o ganho de informação (IG, definido em (4)) pela soma das entropias das variáveis X e Y e multiplica por 2. Essa transformação também mitiga variações que decorrem de uma dependência do valor da entropia do número de classes de X. O desempenho com número de classes distintos é importante neste trabalho, pois compara-se a incerteza simétrica de genes contra genes (binarizados nas classes “ligado” e “desligado”, logo com 2 classes) com a incerteza simétrica de genes contra os tipos celulares (neste trabalho, em torno de 10 classes em cada conjunto de dados analisados).

Exceto quando explicitado, neste texto “correlação”, “correlação por incerteza simétrica”, “SUs de X contra Y” e “correlação por SU” serão usados sempre em relação ao resultado da comparação de duas variáveis pelo cálculo na equação 5.

1.3.3 FCBF: Filtro Rápido Baseado em Correlações

A comparação de SUs de genes contra genes com os SUs de genes contra a classe é a base matemática do algoritmo do Filtro Rápido Baseado em Correlação (FCBF; *Fast Correlation-Based Filter*) (YU; LIU, 2003). Esse algoritmo possui mais de 2000 citações na base de dados acadêmica Google Scholar, mas teve poucas aplicações para transcritômica até então. A parte inicial deste trabalho constituiu na adaptação desse algoritmo para seleção de genes em dados de sequenciamento de RNA, e por isso ele será discutido em seguida. O algoritmo de FCBF (na figura 2) consiste em um ranking das variáveis por incerteza simétrica em relação à classe, mantendo apenas aquelas com maior correlação com a classe. As variáveis que sobram são processadas por uma heurística de redundância para selecionar as variáveis com “correlação predominante”. Ordenadas pela SU contra a classe, são comparadas entre si. Quando a correlação entre duas variáveis é numericamente menor que a correlação com a classe, a variável que explica menos a classe é removida. Após comparar a variável no topo da lista com todas as outras, segue-se uma segunda iteração usando apenas aquelas que sobreviveram à primeira rodada. A seguir, exemplificamos no contexto de transcritômica. Neste exemplo, considera-se uma tabela na qual a expressão de cada gene é medido para diversas células. Essas células têm uma classe, que indica o tipo celular. Calcula-se, então, o SU de cada gene em relação à classe, é feito um ranking e são selecionados genes nomeados de A-Z. O gene "A" tem uma correlação de 0,5 com a classe. O gene "B" tem uma correlação de 0,4 com a classe e de 0,7 com o gene "A". Nesse caso, o gene "B" será considerado redundante por ser:

- mais correlacionado com o gene "A" do que com a classe alvo;
- abaixo de "A" no ranking de SU em relação à classe.

O gene A é comparado dessa mesma forma com todos os genes da lista. Qualquer gene marcado para redundância com A é removido, evitando o cálculo de todos versus todos. Supondo que o primeiro "sobrevivente" não redundante seja o gene "H", o processo é repetido para este. No fim, será obtida uma lista reduzida de genes ("A", "H" e outros mais) ditos predominantemente correlacionados.

1.4 Redundância e coexpressão

O cálculo do algoritmo FCBF é baseado em no conceito de redundância de variáveis em relação a uma classe alvo. Pode-se entender redundância também como coexpressão, já que para serem redundantes dois genes devem ser expressos pelas mesmas células. Logo, uma simples reversão do FCBF (e, por analogia, de outros algoritmos baseados em redundância) pode adaptá-lo para identificar redes de coexpressão. Nas próximas seções, serão abordados o campo de estudo das redes de coexpressão e módulos dessas redes, e as possíveis implicações para novas análises de sequenciamento de RNA.

1.4.1 Redes de coexpressão

O campo das redes de coexpressão é central para a biologia de sistemas (DAM et al., 2017). Tais redes são estruturas matemáticas que relacionam os genes entre si por um valor de correlação. Para esse cálculo, são usadas tabelas com valores normalizados de expressão gênica. No caso de microarranjos, os valores correspondem à luminosidade das sondas para cada gene. No caso do sequenciamento de RNA, esse valor vem das contagens de transcritos na etapa de mapeamento.

A montagem de redes de coexpressão começa pelo cálculo de correlação entre os genes. As medidas de correlação usadas são diversas (como correlação de Spearman, correlação de Pearson, informação mútua e incerteza simétrica). Dois genes são conectados na rede se a correlação entre eles no conjunto de dados ultrapassa um determinado limiar. Acredita-se que essas redes são modulares (LANGFELDER; HORVATH, 2008), pois as diferentes vias celulares e programas transcricionais criam blocos de genes que variam de forma conjunta (covariam).

Algoritmos como o usado pelo WGCNA (LANGFELDER; HORVATH, 2008), implementado na linguagem R impulsionam estudos descritivos de sistemas biológicos. Esses estudos abrangem desde o desenvolvimento de plantas (SERIN et al., 2016) até infecções humanas (JOCHEMS et al., 2018)). Os módulos de coexpressão resumem os dados analisados, permitindo a geração de hipóteses sobre quais produtos gênicos interagem. Variações da estrutura de coexpressão (como módulos com composições diferentes ou métricas de redes alteradas) proveem insights sobre variações globais do sistema, uma abordagem para investigar doenças complexas, multifatoriais (DAM et al., 2017).

Por sua vez, ferramentas, como o ARACNE (MARGOLIN et al., 2006) foram criadas com a intenção de desenvolver modelos causais de como os produtos gênicos interagem. Ao podar redes de coexpressão, ARACNE remove ligações entre genes inferidas como indiretas, enriquecendo a rede para relações causais. Essa classe de algoritmos visa usar a estrutura de coexpressão para inferir redes gênicas regulatórias (GRNs, do inglês *gene regulatory networks*) e também é usada em uma gama de aplicações na área biológica.

1.4.2 Coexpressão em dados de células únicas

O advento do sequenciamento de células únicas está sendo acompanhado pelas análises de coexpressão. Diversos artigos incluem identificação e aplicação de módulos de coexpressão, ainda que sem usar ferramentas específicas. Por exemplo, Farrel et al. visualizaram como a média da expressão dos genes de módulos de coexpressão varia no desenvolvimento do paulistinha (peixe-zebra) (FARRELL

et al., 2018). Shalek et al. usam módulos gênicos para caracterizar a resposta de células dendríticas a diferentes estímulos externos, vendo a expressão agrupada dos genes do módulo (SHALEK et al., 2014). Treutlein et al. usam a média de genes coexpressos (‘módulos’) para caracterizar o desenvolvimento embrionário de células alveolares de camundongo (TREUTLEIN et al., 2014). Além disso, a coexpressão está presente de forma implícita em todo estudo de sequenciamento de células únicas, pois a estrutura de quais genes co-variam influencia as etapas de agrupamento de células (CROW; GILLIS, 2018).

Apesar de aplicações diretas e indiretas, há uma carência de ferramentas especializadas para identificação de coexpressão em dados de células únicas. De fato, muitas dos principais conjuntos de ferramentas para análise de transcrito de células únicas (Seurat (BUTLER et al., 2018), Scanpy (WOLF; ANGERER; THEIS, 2018) e OSCA (AMEZQUITA et al., 2020)) não incluem identificação de redes ou módulos de coexpressão. Outra suíte popular, monocle (<http://cole-trapnell-lab.github.io/monocle-release/>) incluiu apenas recentemente, em sua versão 3 (em fase beta de desenvolvimento) a possibilidade de encontrar módulos gênicos. Logo, a criação de ferramentas para explorar essas possibilidades de análise é um tema em aberto para pesquisa.

A adaptação simples das ferramentas usadas para as outras modalidades não basta. A estrutura dos dados de células únicas é bem diferente daquela dos dados obtidos com a transcriptômica clássica e métodos para inferência de redes gênicas regulatórias não performam tão bem nos novos dados (CHEN; MAR, 2018). Até mesmo as métricas de correlação entre genes - como correlação de Pearson - são colocadas em cheque no contexto das células únicas (SKINNIDER; SQUAIR; FOSTER, 2019). Apesar de diversas abordagens notáveis para inferir redes de coexpressão para células únicas (<https://cole-trapnell-lab.github.io/monocle3/docs/differential/#gene-modules>, (NAZZICARI et al., 2019)), pode-se dizer que não há um “padrão-ouro” no campo, sendo um tema aberto para desenvolvimento teórico e prático.

1.4.3 Módulos de coexpressão para encontrar populações

Neste trabalho, explora-se a ideia que os módulos gênicos de coexpressão guardam informação sobre os diferentes tipos celulares no sistema de interesse. Essa ideia, já presente na literatura (CROW; GILLIS, 2018), não havia sido trabalhada de forma explícita até a execução deste trabalho. A possibilidade de utilizar diferentes partições do transcrito permite fugir da obrigação de atribuir um, e apenas um, tipo celular a cada célula. Isso torna as análises mais completas, aumentando o arsenal do analista e proporcionando melhor aproveitamento do “big data”.

Devido à ausência de um “padrão-ouro”, desenvolvemos um novo método para identificação de coexpressão em células únicas, resignificando um algoritmo de aprendizado de máquinas. Implementamos novidades teóricas relacionadas à forma de filtrar os “melhores” genes, de medir relações de similaridade e de inferir módulos de coexpressão. Além disso, introduzimos a ideia de reagrupamento usando apenas genes co-expressos. No capítulo a seguir, de métodos, serão explicados os procedimentos que fundamentam os pacotes do R/Bioconductor FCBF e fcoex, publicados como resultado deste projeto, e usados para obtenção dos resultados apresentados.

Capítulo 2

Objetivos

2.1 Objetivo principal

- Repropor uma ferramenta de seleção de características (feature selection) para uso na transcritômica de células únicas.

2.2 Objetivos adjacentes

- Adaptar a ferramenta de seleção de características FCBF para encontrar módulos de coexpressão gênica.
- Testar o uso de módulos de coexpressão gênica como passo intermediário inferir rotulações multiníveis de eventos (células).
- Comparar as ferramentas desenvolvidas com o estado da arte em transcritômica de células únicas.

Capítulo 3

Métodos

Neste trabalho, almejamos desenvolver novas formas de analisar dados de células únicas transpondo ferramentas do campo do aprendizado de máquinas. Sendo assim, os métodos aqui descritos referem-se a como os conjuntos de dados públicos utilizados para teste e validação foram obtidos, às ferramentas utilizadas para processamento desses dados e a detalhes metodológicos acerca dos algoritmos implementados nos pacotes FCBF e *fcoex*.

3.1 Dados usados

O conjunto de dados pbmc3k, contendo dados de células mononucleares de sangue humano, versão 3.0.0 foi obtido por meio do pacote *SeuratData* (<https://github.com/satijalab/seurat-data>). O conjunto de dados de desenvolvimento do peixe-zebra no estágio de 75% de epibolia foi retirado do portal Broad Single Cell (https://singlecell.broadinstitute.org/single_cell/study/SCP162).

3.2 Categorização da expressão gênica

Como o algoritmo original de filtro baseado em correlação rápida (FCBF) foi construído para dados discretos, foi preciso discretizar os valores de expressão gênica, passando de contagens (“counts”) para categorias, como “ligado” e “desligado”. A métrica de discretização escolhida foi uma abordagem min-max-percent. Para cada gene, pegamos o menor e o maior valor normalizado nas células. Definimos um limite em 25% da faixa max-min. Todos os valores abaixo desse limite foram considerados “desligados” e todos os acima foram “ligados”. Também implementamos um conjunto de alternativas para comparação:

3.2.1 Binarização pela média ou mediana

Calculamos a média ou a mediana do nível de expressão para cada gene. Células com um nível de expressão maior que a média (ou mediana) foram consideradas “ligadas” para aquele gene, e células com um nível igual ou abaixo da média (ou mediana) foram consideradas “desligadas”.

3.2.2 Discretização por k-means

A discretização por k-means foi usada para particionar a expressão de cada gene em k grupos, usando a função *k-means* do pacote *stats* em R. O número de centros (k) foi escolhido como 2, 3 ou 4. As configurações padrão foram usadas: algoritmo de Hartigan-Wong (HARTIGAN; WONG, 1979) e 10 iterações.

3.3 Identificação dos módulos *fcoex*

3.3.1 Filtrando genes por correlação com marcadores

Após a etapa de discretização, os genes foram classificados por sua correlação com os agrupamentos definidos pela função *FindClusters* de *Seurat*. Usamos a métrica de correlação da incerteza simétrica (SU, do inglês *symmetrical uncertainty*), uma variação do mais conhecido ganho de informação (*information gain*). A correlação por SU varia entre 0 (pior) e 1 (melhor) e, diferente do ganho de informação, corrige para diferenças de entropia oriundas de um número diferente de classes em cada variável.

O filtro de correlação por SU foi usado para escolher n genes, sendo o número exato n escolhido para cada análise. O pacote *fcoex* sugere como default o n igual a 100 genes, anedoticamente observado como suficiente para a análise em diversos casos. As etapas abaixo foram realizadas com os genes filtrados.

3.3.2 Encontrando sementes dos módulos

Os módulos foram construídos de baixo para cima, primeiro selecionando genes predominantemente relacionados aos marcadores para servirem de sementes. Por predominantemente correlacionado, intencionamos dizer que esses genes possuem maior correlação por SU com os agrupamentos do *Seurat* do que em relação a qualquer outro gene. Essa seleção é feita de forma heurística, sendo as sementes o resultado da seleção de características do algoritmo do filtro baseado em correlação rápida (FCBF).

3.3.3 Construindo os módulos / comunidades de coexpressão

O algoritmo constrói uma matriz de incerteza simétrica (SU) de todos os genes contra todos os genes: a matriz de adjacência da rede de coexpressão. Essa matriz de adjacência é podada pelo pacote *fcoex* e as arestas entre os genes Y_i e Y_j são removidas da rede se $SU(Y_i, Y_j) < SU(Y_i, L)$ ou $SU(Y_i, Y_j) < SU(Y_j, L)$, sendo L o vetor de agrupamentos determinados via pacote *Seurat* ($(L$, do inglês *label*).) Ou seja, são removidas todas as correlações intergênicas menores que a correlação de quaisquer um dos genes com as etiquetas. No fim, os módulos mantidos correspondem a cada um dos genes sementes (ver acima) e todos os genes conectados a eles na rede final podada.

3.4 Análise de sobre-representação

Realizamos uma análise de sobre-representação no conjunto de dados de PBMC humano pelos conjuntos de genes do Reactome Pathway processados localmente antes da análise dos dados (FABREGAT et al., 2016). As visualizações no pacote *fcoex* foram adaptadas do pacote *CEMiTool* (RUSSO et al., 2018).

3.5 Identificação de módulos via WGCNA e *monocle3*

Identificamos os módulos usando duas outras ferramentas, além do *fcoex*: o pacote para R *monocle3*, em teste beta, e o pacote para R *CEMiTool* (RUSSO et al., 2018), um embrulhador facilitador de uso do pacote WGCNA (LANGFELDER; HORVATH, 2008). Os módulos identificados com as configurações padrão do pacote foram carregados em um objeto *fcoex* para a análise de reagrupamento.

3.6 Análise de reagrupamento das células

Usamos a função "recluster" do pacote *fcoex* para reagrupar as células com base em cada módulo. Essa função usa os conjuntos de genes de cada módulo de coexpressão para selecionar uma

parte da tabela de expressão original, que passa a conter apenas os genes dentro do módulo. As distâncias entre as células nessa tabela filtrada são calculadas pela distância de Manhattan (a soma das diferenças absolutas na expressão de cada gene) e agrupadas por agrupamento hierárquico. A métrica usada para calcular a distância de ligação entre os grupos foi a “ward.D2”, conforme implementada na função `hclust` do pacote de estatísticas no R 3.6.1. Para cada módulo, recuperamos dois grupos de células (i.e o parâmetro k da função `hclust` foi 2). O grupo com uma expressão mais alta do gene semente foi rotulado como SP (positivo para a semente, do inglês Seed Positive). O outro grupo, complementar, recebeu, então, o rótulo SN (negativo para a semente, do inglês Seed Negative). Os gráficos de reagrupamento foram gerados pela função `DimPlot` do pacote *Seurat*, substituindo os rótulos do objeto *Seurat* pelos rótulos derivados de cada módulo.

3.7 Disponibilidade de código

O pacote *fcoex*, que realiza a análise de coexpressão, está disponível no Bioconductor (<http://bioconductor.org/packages/FCBF/>). Os algoritmos de discretização e seleção de recursos estão disponíveis em um segundo pacote, o *FCBF* (<http://bioconductor.org/packages/FCBF/>).

3.8 Resumo do método desenvolvido

Na figura 3.1 está resumido o método desenvolvido neste trabalho. O pacote *fcoex* gera módulos de coexpressão de genes e clusters derivados de módulos. As células únicas são agrupadas anteriormente - via *Seurat*, por exemplo, e o *fcoex* encontra módulos discretizando os dados, selecionando os genes e usando os rótulos de cada célula no algoritmo. Módulos são usados para subdividir os genes do transcriptoma para encontrar diferentes aglomerados celulares.

Figura 3.1: Visão geral do método desenvolvido.

Capítulo 4

Resultados e discussão

4.1 FCBF como ferramenta de detecção de genes relevantes

Com o objetivo de explorar as possibilidades do FCBF para a análise de dados de células individuais, começamos desenvolvendo testes piloto para o primeiro pacote de R/Bioconductor que desenvolvemos: FCBF. Esse conjunto piloto, de prova de conceito, foi feito para explorar as possibilidades de discretização de dados e analisar qualitativamente os resultados (*outputs*) dos métodos implementados. Começamos por testar diversos parâmetros para discretização no conjunto de dados *pbmc3k* (figura 4.1), usado para demonstrações de pacotes de análise de células únicas .

Figura 4.1: O conjunto de dados *pbmc3k* . Cada ponto representa uma célula, marcada no espaço dimensionalmente reduzido por *tSNE*. O *tSNE* é um algoritmo que permite a visualização das similaridades no espaço milhares de transcritos em apenas duas dimensões (assim como o algoritmo de PCA, do inglês *Principal Component Analysis*). Os rótulos de um dos 8 agrupamentos coloridos acima foram dados pela equipe de analistas do pacote *Seurat* , com base nos marcadores de cada agrupamento encontrado.

Para usar os algoritmos de teoria da informação que almejamos, foi preciso categorizar os dados de expressão, ou seja, binarizar, trinarizar ou agrupar tais dados de alguma outra maneira. Para essa tarefa, o primeiro passo foi o desenvolvimento de um algoritmo para binarização inspirado na categorização de largura igual (*equal-width discretization*), que classifica cada observação com base em intervalos iguais do valor mínimo ao máximo.

Para cada análise, usamos o mesmo limiar percentual para todos os genes diferentes. A partir

desse limiar, decidimos considerar apenas dois estados possíveis de expressão gênica, ligado ou desligado (acima ou abaixo do limiar). A **figura 4.2** explica o processo de forma gráfica.

Figura 4.2: Binarização baseada em limiar para dados de expressão de células individuais única. Histogramas dos genes S100A9 e NKG7 no conjunto de dados pbmc3k. O mesmo limite é aplicado a cada gene. Vermelho: o gene está ativado. Azul: o gene está desativado.

A seleção de diferentes limiares, claro, muda quais genes são considerados ligados e desligados. Logo, decidimos observar o comportamento do algoritmo do filtro rápido baseado em correlação (FCBF) para os diferentes limiares. A primeira característica que observamos foi o número de genes recuperados pelo algoritmo quando o único parâmetro alterado era o limiar (**figura 4.3**).

Usamos 20 limiares diferentes e observamos que, para o conjunto de dados de pbmc3k, mais genes eram recuperados em torno do limiar de 25%, diminuindo para valores mais extremos. O número maior de genes recuperados pode refletir que a discretização foi mais bem sucedida em capturar diferenças biológicas. O número maior de genes recuperados é decorrente de maior correlação por SU com a classe, com mais genes passando o filtro de qualidade inicial.

A diferença no número de genes selecionados pode ter um impacto baixo, caso os genes escolhidos sejam os mesmos ou ao menos envolvidos nos mesmos processos. Por isso, fizemos mais duas análises comparando, para diferentes limiares, a constância dos genes selecionados e das vias enriquecidas (**Figura 4.4**).

Os genes selecionados pelo algoritmo de FCBF são relativamente constantes, com 42 genes selecionados em 8 ou mais limiares diferentes (**figura 4.4A**) para o conjunto de dados pbmc3k. O enriquecimento funcional para vias do banco de dados KEGG 2016 mostrou que o conjunto de genes selecionados para os diversos limiares apresentaram enriquecimento para vias relacionadas ao sistema imune, como “imunodeficiência primária” (Primary immunodeficiency) e “via de sinalização do receptor de células T” (T-cell receptor signalling pathway) (**figura 4.4B**). Curiosamente, os limiares mais altos levaram à seleção de diversos genes ribossomais como separadores de tipos celulares imunes. Essa observação está de acordo com trabalhos anteriores, que mostraram que o nível de expressão de genes ribossomais distingue linfócitos de granulócitos no sangue (PALMER

Figura 4.3: O número de genes selecionados pelo FCBF depende do limiar de binarização. Contagem de genes selecionados usando diferentes limiares para discretização de largura variável, aplicada aos dados pbmc3k. O valor mínimo de incerteza simétrica de 0,02 foi usado para o filtro dos genes¹. A linha verde corresponde à regressão loess de médias “amaciadas” (smoothed means), conforme implementado no ggplot2. FCBF: filtro rápido baseado em correlação.

Figura 4.4: A binarização seguida pelo FCBF pode destacar genes relevantes. A) Genes selecionados 8 ou mais vezes para 20 limites de discretização de largura variável. B) Enriquecimento no banco de dados KEGG 2016 para cada um dos limites da binarização de largura variável. FCBF: filtro rápido baseado em correlação.

et al., 2006).

Com esse resultado, concluímos que a discretização afeta tanto qualitativamente quanto quantitativamente o funcionamento do algoritmo de FCBF. Apesar das variações, os resultados foram pertinentes para uma ampla gama de discretizações simples. Exemplificando, muitos dos genes selecionados para 8 ou mais limiares são essenciais para a função imune, como: CD3D, (parte essencial do complexo TCR nas células T), CCR7 (receptor de quimiocinas) e GZMK (uma granzima secretada por células citotóxicas efectoras, como células T CD8 + e células NK).

Esses genes, além de relevantes, são diferentes daqueles apontados como marcadores pelo pacote

Figura 4.5: Diagrama de Venn mostrando a relação entre os genes encontrados pelo algoritmo de FCBF e os genes diferencialmente expressos (DEGs, do inglês Differentially Expressed Genes).

Seurat. Recuperamos os top 4 marcadores do *Seurat* para cada um dos oito grupos (totalizando 32) para comparação com os 30 genes selecionados pelo FCBF, encontrando tanto genes em comum como genes específicos para cada técnica (figura 4.4A). Obtivemos indícios que os genes selecionados pelo FCBF correspondiam à combinação dos rótulos originais (figura 4.5B), como o gene CD3D, expresso tanto pelos grupos de células CD4+ como células CD8+. Dessa forma, postulamos que o FCBF poderia contribuir de forma complementar aos padrões da literatura atual, destacando genes interessantes.

4.1.1 Explorando discretizações alternativas

Figura 4.6: Métodos diferentes de discretização mudam o valor máximo das correlações. Para cada discretização, a figura mostra o boxplot dos 30 maiores valores de incerteza simétrica em relação aos 8 rótulos do conjunto de dados pbmc3k. O método mais à direita (“supervised”) refere-se a selecionar a discretização que maximiza a incerteza simétrica para cada gene.

Exploramos também diferentes formas de discretizar a expressão gênica. Há inúmeras formas de transformar a expressão numérica (counts normalizados) em categórica. Para avaliar qual delas seria melhor postulamos que a discretização que permitir melhor captura da diferença entre os grupos celulares será a mais relevante. A partir disso, aplicamos 9 diferentes combinações de funções e parâmetros ao conjunto de genes pbmck3k e avaliamos quão bem os genes estavam correlacionados com as classes. Como a análise foi preliminar, escolhemos apenas os top 30 genes de cada combinação e comparamos na figura 4.6.

Podemos observar que a performance foi semelhante para os diversos métodos, sendo pior para *slope jump*, e o *min-%max* com corte no 75% entre o valor mínimo e o máximo. No primeiro caso, o método havia sido desenvolvido para dados de microarranjo total (SHMULEVICH; ZHANG, 2002) ressaltando que a natureza diferente dos dados preclui o uso cego de ferramentas antigas. No segundo caso, o corte de 75% entre o mínimo e o máximo levou a muitos genes serem considerados “desligados” de forma equivocada.

Para outros conjuntos de dados, não mostrados aqui, o *ranking* de performance dos diferentes discretizadores se alterou bastante, com exceção dos 3 casos mencionados acima, sendo limiares altos e *slope jump* consistentemente ruins e o supervisionado, consistentemente o melhor. Como o supervisionado requer a rotulação das células como um tipo celular, para as análises seguintes preferimos utilizar o método *min-%max* com o corte de 0,25. A escolha desse conjunto de parâmetros foi arbitrária, e outros parâmetros poderiam ter sido usados. Nota-se, contudo, que essa arbitrariedade não invalida as observações seguintes.

4.1.2 Observações de redundância como coexpressão

Além de verificar que os genes *selecionados* pelo FCBF são relevantes biologicamente, decidimos observar se os genes *removidos* eram redundantes aos selecionados. A hipótese subjacente é que a redundância estatística nos dados, determinada pela comparação de correlações, reflete uma redundância de funções biológicas. Ou seja, os genes removidos teriam envolvimento em funções parecidas, correspondendo a programas específicos dos tipos celulares na amostra.

Para esse conjunto de dados e combinação de parâmetros, 30 genes foram selecionados no conjunto de dados de células mononucleares de sangue periférico (PBMCs). Os cinco genes melhor ranqueados pelo algoritmo foram: *CST3*, *CD3D*, *HLA-DRA*, *NKG7*, e *CD79A*. Abaixo apresentaremos as discussões preliminares acerca de dois genes: *CD3D* e *CD79A*. Quanto aos outros, falaremos brevemente, apenas, que os genes redundantes a *CST3* são relacionados à funções de células fagocíticas (como *LYZ*, que codifica lisozima, e *S100A9*, que codifica uma proteína de transdução de sinal do receptor *TLR4*), os redundantes a *HLA-DRA* codificam proteínas do MHC de classe II e os redundantes a *NKG7* se relacionam a citotoxicidade (como *PRF1*, da perforina, e *GZMA*, da granzima A).

CD3D e genes correlatos determinados pelo FCBF

O primeiro gene da lista de genes selecionados pelo FCBF que vamos discutir é o *CD3D*. Esse gene codifica a cadeia delta da proteína CD3, que forma um complexo que transduz o sinal do receptor de células T. Esperamos, então, que esse gene seja característico das células T, CD4 e CD8, e é isso que vemos na figura 4.7.

A lista de genes removidos pelo FCBF em contraste com *CD3D* é composta pelos seguintes genes:

- *CD3E*, que codifica uma cadeia do CD3, e parece ser expresso pelas mesmas células que *CD3D*.
- *IL7R*, que codifica o receptor de IL7. A IL7 é uma citocina que possui como uma de suas funções, a promoção da sobrevivência de células T. Além disso, a sinalização do receptor IL7R está relacionada à do receptor TCR, com mediadores comuns a ambos (CARRETTE; SURH, 2012).

Figura 4.7: A-D) gráficos *t*-SNE da expressão de *CD3D* (A), *CD3E* (B) *IL7R* (C) e *CD2* (D). A expressão é plotada em uma escala de amarelo (baixa expressão) a vermelho (alta expressão). A Figura 4.1 apresenta uma referência para os clusters e seus rótulos. A expressão maior dos 4 genes ocorre nos grupos de células *T CD4* e *T CD8*. FCBF: filtro rápido baseado em correlação.

- *CD2*, que codifica a proteína de mesmo nome. Essa proteína associa-se a *CD58*, e essa associação participa do reconhecimento de outras células pelas células *T* (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=CD2>).
- *LCK*, que codifica uma tirosina quinase da família *Src* específica de células *T*. Está envolvido na transdução de sinal pelo *TCR*, fosforilando os motivos de ativação à base de tirosina do imunorreceptor (*ITAM*) (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=LCK>)

O gene *IL32* não foi removido em comparação com *CD3D* por que já tinha sido removido durante a comparação com *HLA-DRA*. O gene da *IL-32* é expresso por linfócitos *T* e células *NK*. Ele havia sido removido por ser expresso nas células que *não* expressam *HLA-DRA*. A correlação de incerteza simétrica não diferencia correlações negativas e positivas por si. De qualquer forma, podemos notar que os genes considerados redundantes ao *CD3D* pelo FCBF estão envolvidos de forma geral na mesma função: sinalização do *TCR* e demais atividades características de células *T*.

CD79A e genes correlatos determinados pelo FCBF

Outro gene selecionado pelo FCBF foi o *CD79A*, um marcador de células *B*. A proteína *CD79A*, ao lado de *CD79B*, forma um complexo que medeia a sinalização de células *B* pelo receptor de células *B* (*BCR*). (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=CD79A>). A lista de genes removidos em comparação com *CD79A* inclui os seguintes genes:

- *CD79B*, que codifica a proteína que compõe o heterodímero com *CD79A* à jusante na resposta do *BCR*.
- *MS4A1*, também chamado de *CD20*, que codifica uma proteína de superfície específica das células *B* e alvo de quimioterapia, mas cujos mecanismos principais de funcionamento são desconhecidos (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=MS4A1>).

Figura 4.8: *CD79A* e genes redundantes segundo o FCBF. A-D) gráficos *t*-SNE da expressão de *CD79A* (A), *MS4A1* (B) *LINC00926* (C) e *VPREB3* (D). A expressão é plotada em uma escala de amarelo (baixa expressão) a vermelho (alta expressão). E) Visualização do navegador de genomas da UCSC do transcrito *LINC00926*. O asterisco verde indica o local de início do gene *TCF12*, relacionado ao desenvolvimento de células B. O asterisco rosa mostra o site inicial do *LINC00926*. A Figura 4.1 apresenta uma referência para os clusters e seus rótulos. FCBF: filtro rápido baseado em correlação.

- *LINC00926* é um RNA longo não codificador intergênico sem função descrita. Sua associação a *MS4A1* e *CD79A* implica, por culpa por associação (do inglês, *guilt by association*), que esse gene pode estar envolvido na função das células B e, talvez até na sinalização do BCR. (<https://www.genecards.org/cgi-bin/carddisp.pl?gene=LINC00926>).
- *VPREB3*, que codifica uma proteína que é expressa durante a diferenciação de células B (RODIG et al., 2010).

O padrão de expressão desses quatro genes pode ser visto nas células do conjunto de dados *pbmc3k* na figura 4.8. Apesar de nem todas as células que expressam *CD79A* expressarem *LINC0096* ou *VPREB3*, há um grau considerável de coexpressão. Essas associações de coexpressão, então, podem servir de ponte para direcionar hipóteses e caminhos de pesquisa. Por exemplo, olhando para localização genômica, vê-se que o *LINC00926* está próximo ao gene que codifica a proteína *TCF12* (figura 4.8B), um fator de transcrição que desempenha um papel no desenvolvimento das células B (outro nome para *TCF12* é *HEB*; (ZHUANG; CHENG; WEINTRAUB, 1996)).

Essa análise preliminar do algoritmo FCBF, aponta que ele traz à luz associações de genes e tipos de células, selecionando genes de alta relevância para entender a biologia das células. Além disso, os genes excluídos parecem apontar para as mesmas funções, com casos de redundância biológica

(por exemplo, a remoção do *CD79B* em comparação ao *CD79A*). Isso motivou a etapa seguinte do trabalho: a modificação do algoritmo FCBF para identificação de módulos de coexpressão.

4.2 Módulos de coexpressão por *fcoex*

Figura 4.9: Genes semente são robustos a variação do número inicial de grupos. A) Top-3 sementes selecionadas pelo algoritmo de FCBF para números diferentes de grupos determinados pelo Seurat. Combinações dos mesmos 8 genes são selecionadas nas 18 situações testadas.

Dadas as observações anteriores, avançamos para usar o filtro rápido baseado em correlação (FCBF) para detectar módulos de coexpressão. Para isso, implementamos mais um pacote em R, o *fcoex*. Esse pacote agrega métodos para discretizar dados de single-cell, identificar genes importantes via FCBF e, em seguida, inferir módulos de coexpressão baseados nesses genes.

Analizamos, mais uma vez, o conjunto de dados *pbmc3k*. Por inovações metodológicas que surgiram no decorrer do trabalho, alguns resultados obtidos diferem em detalhes dos descritos acima. A mudança mais marcante é a forma de dispor os dados, que acima foram mostrados via tSNE (*t Stochastic Neighbour Embedding*) que agora foi trocada por UMAP (*Uniform Manifold Approximation and Projection*). Esse algoritmo recentemente despontou como um novo padrão de visualização para dados de células únicas, e os autores originais afirmam que o agrupamento por UMAP preserva melhor a estrutura global (BECHT et al., 2018), apesar de controvérsias (KOBAK; LINDERMAN, 2019). Outra mudança significativa ocorreu no pacote *Seurat*, que mudou a normalização padrão e o número de grupos fornecidos a priori (de 8 no *Seurat 2* para 9 no *Seurat 3*). Essas mudanças não afetam as conclusões finais, e os genes selecionados pelo FCBF são em grande parte os mesmos da fase anterior.

Testamos o impacto da variação do número de grupos a priori do *Seurat* nos genes selecionados pelo FCBF. Na **figura 4.9** podemos ver que os genes selecionados pelo FCBF são muito semelhantes, mesmo variando o parâmetro inicial de 3 para 26 grupos. O gene *TYROBP*, por exemplo, é escolhido em todas as 18 situações testadas, enquanto *CD3D* é encontrado em 16/18 casos. Pela estrutura do algoritmo de agrupamento do *Seurat*, que é baseado em modularidade de grafos, o número de grupos não é um parâmetro controlável de forma direta, levando aos pulos observáveis no gráfico (entre #15 e #18, por exemplo).

Dentre os módulos no conjunto de dados *pbmc3k*, destacamos dois na **figura 4.10**. A visualização, adaptada do pacote *CEMiTool* (RUSSO et al., 2018) mostra os genes como nós da rede e as

Figura 4.10: Módulos de coexpressão e enriquecimento funcional em PBMCs. A) Dois dos nove módulos de coexpressão determinados pelo pacote *fcoex* para o conjunto de dados *pbmc3k*. B) Análise de sobre-representação para o conjunto de genes dos módulos em A em comparação com vias curadas pela base Reactome.

correlações como arestas entre eles. Eles são conectados na rede se algoritmo do FCBF considerá-los redundantes. Na **figura 4.10a**, vemos o módulo para *CD3D*, para o qual alguns membros foram discutidos na sessão anterior, e o módulo *HLA-DRB1*, que é composto majoritariamente por genes que codificam proteínas do MHC de classe II. Na **figura 4.10b** vemos o enriquecimento funcional por análise de sobre-representação de cada módulo, confirmando que o módulo *CD3D* está relacionado com funções de células T e o módulo *HLA-DRB1* está relacionado à apresentação de antígenos.

Para etapa seguinte do projeto, dois fatores serviram de inspiração. Em primeiro lugar, observamos mais uma vez que os genes selecionados pelo FCBF estavam expressos em mais de um dos grupos originais do *Seurat*. Por exemplo, o *CD3D*, citado em ambas sessões, está expresso em células T CD4 naïve, T CD4 de memória e CD8. Em seguida, observamos que os módulos de genes correlacionados tinham coerência biológica com genes funcionalmente relacionados. Sendo assim, conjecturamos que os módulos de coexpressão poderiam servir para encontrar populações de células biologicamente relacionadas no conjunto de dados. De início, então, aplicamos o algoritmo para tentar recuperar populações usando os módulos no conjunto de dados *pbmc3k*.

Na **figura 4.11a**, podemos ver as 9 populações determinadas a priori pela equipe do pacote *Seurat*. O conjunto de dados inclui células citotóxicas e, no módulo retratado na **figura 4.11b**, identificamos diversos genes de citotoxicidade (como os codificadores de perforina, *PRF1*, e granzimas, *GZMA*, *GZMB*, *GZMM* e *GZMH*). Utilizando somente os genes presentes neste módulo para subdividir o conjunto de dados em duas populações vemos que o pacote *fcoex* agrupa células NK e T CD8 (em azul, marcado como SP, *seed positive*) em conjunto (**figura 4.11c**). Isso é uma indicação que as populações recuperadas com base em módulos de coexpressão fazem sentido biologicamente.

O mesmo padrão pôde ser observado para os outros módulos recuperados. Em 4.11d vemos o caso do módulo *HLA-DRB1*, descrito acima, que levou ao agrupamento de células B com monócitos e células dendríticas, coerente com as funções semelhantes desses grupos no que se refere à apresentação especializada de antígenos para células T CD4. O módulo *CD79A*, entretanto, com genes relacionados à funções específicas do receptor de células B, separou essa população das demais no dataset. Cada módulo recuperado pelo *fcoex* gerou populações de células biologicamente coerentes, agrupando diferentes partições originais. As diferentes divisões podem ser vistas na figura 4.11e.

Figura 4.11: Os módulos gerados pelo *fcoex* agrupam funcionalmente populações de PBMCs. A) Visualização UMAP (Uniform Manifold Approximation and Projection) do conjunto de dados *pbmc3k*. Nota: devido à atualizações no pacote *Seurat*, o padrão usado a partir desse ponto foi a divisão em nove grupos, e não oito como anteriormente. B) Módulo M8, liderado pelo gene *CST7*. Os nós representam genes, as arestas representam correlações por incerteza simétrica (ver Métodos). C) Divisão em 2 grupos baseada no módulo M8, em B. D) Clusters derivados dos módulos M5 (*HLA-DRB1*) e M7 (*CD79A*). E) Tabela de atribuições, mostrando a porcentagem de células em cada cluster *Seurat* que estão contidas nos agrupamentos “SP”. PBMC: células mononucleares do sangue periférico; SP: seed positive; SN: seed negative

Em seguida, separamos as células do conjunto de dados *pbmc3k* que correspondiam à monócitos e re-subdividimos com o algoritmo do pacote *Seurat* (figura 4.12a) em 10 subgrupos. Essa divisão ingênua (*naive*) é necessária para o funcionamento de algoritmo de filtro do FCBF. Vale mencionar que a divisão é arbitrária - mas não aleatória - agrupando células que são parecidas no espaço multidimensional. Dois dos módulos recuperados eram encabeçados pelos genes *FCGR3A* (que codifica a proteína CD16) e *CD14* (figura 4.12b). De fato, os monócitos sanguíneos são divididos tradicionalmente com base nesses marcadores proteicos, sendo chamados de clássicos, não-clássicos e intermediários (BROOKS; MOORE, 1988). Vemos também que nem todas as células pertencem a um ou outro grupo azul. Há diversos fatores que podem contribuir para essas diferenças, de dropouts técnicos e transcrição estocástica, até presença de células semelhantes a monócitos tradicionais que não expressam nenhum dos dois marcadores.

Um outro módulo gênico identificado só nessa população era caracterizado por genes relaciona-

Figura 4.12: Os módulos do *fcoex* permitem a identificação de tipos diferentes de monócitos. A: Células marcadas com monócitos, subdivididas em 9 grupos pela função *FindClusters* de *Seurat*. B: Subpopulações de monócitos dos módulos M1 e M3 (subpopulações canônicas, CD16 e CD14 positivas, respectivamente). C) Grupos derivados do módulo M5 separam em duas populações com maior ou menor expressão de certos genes do MHC de classe II. D) Módulo M5, que deu origem à subdivisão em C.

dos ao MHC de classe II, mostrado na **figura 4.12d**. O tema da expressão em níveis distintos de proteínas do MHC II entre diferentes subtipos de monócito é presente na literatura (BUNBURY et al., 2009). Além disso, macrófagos associados à borda do cérebro (que são derivados de monócitos) se subdividem em grupos fenotipicamente distintos por nível de expressão de MHC de classe II (MRDJEN et al., 2018). O grupo observado em decorrência do módulo HLA-DRB1, contudo, não corresponde aparentemente a nenhum grupo clássico definido anteriormente, constituindo uma forma nova de enxergar as diferentes subpopulações de monócitos. É possível observar como as 10 subdivisões do *Seurat* fluem para os grupos derivados dos módulos na **figura 4.13**.

Figura 4.13: Células divididas originalmente em 9 grupos, se reposicionam nos reagrupamentos do *fcoex*. Diagramas de Sankey, ou aluviais, mostrado para três módulos, M1 (FCGR3A), M3 (CD14) e M5 (HLA-DRB1).

As populações celulares presentes no conjunto de dados *pbmck3* são fenotipicamente bastante distintas, com grupos celulares relativamente homogêneos, de funções conhecidas e bem definidas. Almejamos então testar o algoritmo do *fcoex* em um sistema diferente, com menor grau de separação fenotípica entre as células. Para isso, usamos um conjunto de dados de gastrulação de peixe zebra no estado de 75% de epibolia. Nesse momento do desenvolvimento, as células do envelope externo de células do peixe zebra estão migrando rumo ao polo vegetal e invaginando, começando a diferenciarem-se entre endoderma, mesoderma e ectoderma.

Assim como no conjunto de dados de monócitos, dividimos as células de peixe-zebra de forma ingênua em 10 grupos utilizando o pacote *Seurat* (figura 4.14a). O *fcoex* identificou 8 agrupamentos baseados em módulos de coexpressão, que podemos ver na figura 4.14b. Notamos que os módulos e os conjuntos originais de maneiras distintas, fornecendo uma visão multicamada das identidades celulares.

Figura 4.14: *Fcoex* revela dinâmicas das células embrionárias do paulistinha (peixe-zebra). A) 10 grupos identificados por *Seurat* no conjunto de dados de gastrulação de peixe zebra, no estágio de epibolia a 75%. B) Diferentes atribuições para as células com base em cada módulo. C) Genes no módulo M2 (*SOX19A*) expressos majoritariamente no grupo SP (em azul) ou SN (em vermelho). D) Padrões de expressão para o gene *APELA*, para os genes que codificam receptores de seu produto, *APLNRB* e *APLNRA*, e um fator à jusante, *mespab*.

Diversos genes selecionados como sementes pelo FCBF tem um papel no desenvolvimento do peixe-zebra. *msgn1*, que encabeça o primeiro módulo, é importante para o desenvolvimento do mesoderma pré-somítico, que da origem aos somitos e grande parte dos músculos do animal (YABE; TAKADA, 2012). Já o *sp5l* é um gene da via de Wnt/beta-catenina que responde à sinalização por proteínas da família Fgf na formação do mesoderma e do neuroectoderma (WEIDINGER et al., 2005). O gene *rippy1* está também relacionado à formação do mesoderma paraxial e dos músculos esqueléticos (WEIDINGER et al., 2005), enquanto *ta* é conhecido por mediar tanto a formação do endoderma quanto do mesoderma (NELSON et al., 2017), 2017). É possível notar que os agrupamentos originários dos módulos 3 e 5 (*sp5l* e *ta*) são bem semelhantes, podendo isso ser atribuído à um overlap dos genes incluídos nos dois módulos.

Tabela 4.1: Descrição dos módulos de coexpressão encontrados para cada pacote

Pacote	Número de módulos	Tamanho médio dos módulos
monocle3	9	596
WGCNA	4	309
fcoex	9	23

Apesar do gene *sox19a* ser pouco discutido na literatura sobre o peixe-zebra, o módulo M2 contém uma série de genes interessantes (**figura 4.14c**). Dentre seus 18 genes, 4 são conhecidamente relacionados a vias importantes do desenvolvimento embrionário. Como os módulos do *fcoex* contêm genes correlacionados negativamente, detectamos o gene *apela* e os genes *aplnra* e *aplnrb*. O produto do gene *apela* é secretado e sinaliza através dos receptores codificados por *aplnrb* e *aplnra*. Além disso, esse módulo contém o gene *mespab*, cuja expressão é reduzida em peixes nocaute para *aplnrb*, e é um possível fator à jusante desta cascata de sinalização. O padrão de expressão desses genes pode ser visto na **figura 4.14d**, que deixa claro a separação dos genes em duas populações relevantes: a de transmissores do sinal e de receptores do sinal.

4.3 Análises comparativas aprofundadas

Após termos desenvolvido essa metodologia para encontrar populações a partir de módulos de coexpressão, diversas questões surgiram. Em retrospectiva, certos experimentos computacionais encaixariam melhor no encadeamento lógico em partes anteriores desse texto. Contudo, como esses resultados vieram após as escolhas e conclusões que motivaram a primeira parte. Logo, o fato desses resultados serem expostos ao fim visa evitar a aparência que este conjunto de análises influenciou no anterior.

4.3.1 Comparação de pipelines para encontrar módulos de coexpressão

Após os resultados positivos com *fcoex*, hipotetizamos que conseguiríamos usar módulos de coexpressão para identificar populações usando outras ferramentas. Fizemos, então, a comparação para o conjunto de dados *pbmc3k* dos resultados do *fcoex* nativo com a combinação de funções de reagrupamento do *fcoex* com funções de localizar módulos de dois pacotes em R: *monocle3* e *WGCNA* (com otimização de parâmetros pelo pacote *CEMiTool*). Esses resultados podem ser vistos na **figura 4.15**.

Enquanto o *WGCNA* foi capaz apenas de encontrar 4 módulos de coexpressão (**figura 4.15a-b**) (mais um módulo de genes não correlacionados), o *monocle3* encontrou 9 módulos (**figura 4.15c-d**). O uso da função de reagrupamento do *fcoex* levou a observação que as populações encontradas usando os módulos do *monocle3* correspondiam, de fato, aos grupos identificados pelo *fcoex* (**figura 4.15e-f**) e mais 2: linfócitos (B + T) e plaquetas. O *monocle3*, contudo, não recuperou um módulo exclusivo para células B, encontrado no *WGCNA* e *fcoex*. Logo, as diferentes abordagens para localizar módulos de coexpressão apresentam sobreposições marcantes, destacando a robustez da abordagem, e são complementares entre si, justificando o uso e desenvolvimento de novas ferramentas.

Além de observar as populações resultantes, outro detalhe é importante para essa comparação de ferramentas: o tamanho dos módulos gerados (**figura 4.1**). Enquanto o *monocle3* gerou módulos com, em média, 593 genes, os módulos do *fcoex* tem média de 23 genes. Isso se dá ao filtro eficiente de genes relevantes, e permite ao analista explorar os dados manualmente com facilidade. Esse fator foi crucial para as conclusões obtidas com os dados de peixe-zebra que exploramos na seção anterior.

Figura 4.15: Módulos gerados por outros pacotes agrupam funcionalmente populações de PBMCs. Na esquerda, tabelas de atribuições, mostrando quanto de cada agrupamento original (Seurat) está contido nos agrupamentos derivados de cada módulo. Note: 2 agrupamentos são feitos por módulo, logo 0% em relação ao grupo azul (SP) significa 100% em relação ao grupo vermelho (SN). Na direita, reagrupamentos plotados no espaço de do UMAP (Uniform Manifold Approximation and Projection). Resultados para módulos encontrados com os pacotes de R *monocle3* (A-B), *WGCNA* (chamado via pacote *CEMiTool*) (C-D) e *fcoex* (E-F). O M5 do *WGCNA* corresponde a genes não pertencentes a nenhum módulo propriamente dito.

4.3.2 Comparação de incerteza simétrica e variância para seleção de bons genes

Além de comparar as pipelines por completo, decidimos comparar pontos específicos que propusemos com os padrões atuais da literatura. Uma novidade teórica que apresentamos é o uso da combinação de discretização e incerteza simétrica para selecionar genes relevantes para análises seguintes. Os métodos usados atualmente frequentemente usam os genes mais variáveis como *proxy* para relevância (LUECKEN; THEIS, 2019).

Apesar de bons resultados empíricos, é possível que outras alternativas de seleção de genes sejam melhores para extrair relevância. Precisamos, então, para uma função objectiva que decida qual ranqueamento de genes é mais relevante para compreensão biológica. Decidimos utilizar GSEA (*Gene Set Enrichment Analysis*, análise de enriquecimento de conjuntos gênicos) (SUBRAMANIAN et al., 2005), por sua popularidade na área e aplicabilidade. Dessa forma, um ranqueamento de genes que destacasse genes mais relevantes iria ter uma pontuação (NES, Normalized Enrichment Score) mais alta no que se refere ao enriquecimento dos marcadores das populações de interesse no dataset.

Usando os marcadores da base de dados PanglaoDB (FRANZÉN; GAN; BJÖRKEGREN, 2019) para células T, células NK, células B, monócitos, plaquetas e células dendríticas, testamos o ranqueamento por incerteza simétrica e genes variáveis no conjunto de dados *pbmc3k*. Notavelmente, em todos os 6 casos, a pontuação encontrada para incerteza simétrica superou a encontrada para alta variância (tabela 4.2 e figura 4.16). Por um teste binomial, o p-valor para a incerteza simétrica

Tabela 4.2: *Enriquecimento via GSEA para cada tipo celular no conjunto de dados pbmc3k*

Tipo celular	Método de ordenamento	
	Incerteza Simétrica	Variância
células T	1,95	1,58
células B	2,06	1,63
monócitos	2,09	1,81
plaquetas	2,16	2,01
células dendríticas	1,81	0,73
células NK	2,11	2,02

aparecer sempre ao acaso é de 0,15625 (0,5 elevado à sexta potência), o que indica que, ao menos para o conjunto de dados *pbmc3k*, nossa abordagem seleciona genes mais relevantes pelos critérios de relevância decididos anteriormente.

Figura 4.16: A pontuação de enriquecimento para marcadores celulares é maior para incerteza simétrica em comparação com variância. Os valores de enriquecimento para 6 conjuntos de marcadores obtidos da literatura para de células presentes no conjunto de dados *pbmc3k* para os ranqueamentos por incerteza simétrica (esquerda) e variância (direita). O enriquecimento foi calculado com o pacote para R *fgsea*.

4.3.3 Comparação de algoritmos para agrupar as células com base nos módulos

Para agrupar as células com base nos módulos, utilizamos originalmente o algoritmo de clusterização hierárquica, calculando a distância entre células por uma métrica denominada *manhattan* e um método de linkage de grupos chamado *ward.D2*. A escolha inicial desses parâmetros foi arbitrária, com base em observações anedóticas de como o agrupamento se comportava. Para comparar nossa escolha com as outras abordagens possíveis, precisávamos criar uma forma de comparar a qualidade de algoritmos de agrupamento não supervisionado. Em todos os casos, o resultado final destes algoritmos seriam a divisão do conjunto de dados em 2 grupos de células relativos aos módulos.

Partindo da observação que os módulos encontrados via *WGCNA*, *fcoex* e *monocle3* correspondem a reagrupamentos dos diferentes grupos originais (determinados pela equipe do *Seurat*), postulamos que métodos de agrupamento (*clustering*) que capturem com mais pureza os grupos biológicos originais seriam mais interessantes. Por exemplo, o agrupamento ideal derivado do módulo do gene *CD3D* capturaria todas as células T CD4 de memória, T CD4 imaturas e T CD8, e não incluiria nenhuma célula dos outros grupos. Caso o agrupamento pegasse somente uma parcela das células T, a pureza seria menor. No caso extremo de um agrupamento aleatório, a pureza seria mínima.

Para comparar a pureza obtida por diferentes algoritmos de agrupamento, utilizamos a medida de entropia. Para cada grupo do *Seurat* original, calculamos o percentual das células incluídas nos agrupamentos do *fcoex* para o conjunto de dados *pbmc3k* e, a esse valor entre 0 e 1, aplicamos a fórmula da entropia, descrita na introdução (figura 1.1). Valores mais extremos demonstram pureza, enquanto valores intermediários sinalizam que a atribuição foi mais solta, devido à ruídos e vieses. Por exemplo, a proporção de células T CD4 naive pertencentes ao grupo SP (*seed positive*) para *CD3D* é 1, logo a entropia é 0. A proporção de plaquetas pertencentes ao mesmo grupo é 0, logo a entropia também é 0. Se células fossem distribuídas aleatoriamente dentro ou fora do grupo SP e a proporção fosse por volta de 0,5, a entropia seria por volta de 1.

Como prova de conceito, testamos dois algoritmos, *k-means* e agrupamento hierárquico, e comparamos com a atribuição aleatória das células a dois grupos postíços (figura 4.17). A entropia relaciona-se inversamente com a pureza. Ambos algoritmos tiveram performance muito superior ao acaso, como esperado pela abordagem proposta. Porém, não identificamos diferenças significativas entre a qualidade da atribuição originada por ambos algoritmos.

Figura 4.17: Agrupamento baseado em módulos leva a combinação puras das populações originais. Entropia para quão puro cada grupo original do conjunto de dados *pbmc3k* está em relação aos agrupamentos “SP” derivados dos módulos *fcoex*. Agrupamento feito via dois algoritmos: agrupamento hierárquico e *kmeans*. Cada boxplot contém 9 grupos do *Seurat* x 9 módulos *fcoex* = 81 observações.

Capítulo 5

Conclusões

O estudo como um todo abordou e analisou cinco aspectos teóricos com implicações práticas na área:

- Os dados de sequenciamento de RNA de células únicas podem ser discretizados.

Apesar de diversos estudos binarizarem a expressão gênica, a literatura carece de abordagens explícitas sobre a discretização de expressão gênica de células únicas. Aqui mostramos para os conjuntos de dados pbmc3k e de células de paulistinha que a discretização, mesmo que ingênua e subótima, ainda permite a recuperação de informações biológicas.

- Os métodos de seleção de características podem ser reaproveitados para coexpressão.

A literatura de aprendizado de máquinas é recheada de soluções elegantes para encontrar redundância. Esses algoritmos são uma fonte rica de possíveis abordagens metodológicas para inferir relações entre genes. Esse trabalho é, até onde sabemos, o primeiro a reverter um algoritmo de remoção de redundância para captura de redes de coexpressão. Ou seja, há um campo de possibilidades a ser explorado.

- Módulos de coexpressão com poucos genes facilitam a exploração manual dos dados.

A literatura de coexpressão gênica, de forma geral, utiliza módulos com grande número de genes, na casa de centenas. Em vez de considerar os genes individuais, os módulos são resumidos por análises de sobre-representação de vias, abstraindo os genes na relação módulo-via. A análise do módulo do gene *sox19a*, no conjunto de dados de células de peixe-zebra, é uma demonstração de que o uso de módulos de coexpressão pequenos pode ser vantajoso, dependendo da pergunta biológica em questão. Para isso, torna-se necessária uma boa etapa de filtro de genes relevantes, o que, no nosso caso, é feita pelo algoritmo de FCBF.

- Comunidades de coexpressão podem ser aproveitadas para inferir tipos celulares.

Como mencionado no texto, o panorama de coexpressão gênica já influencia, de forma implícita, os algoritmos de agrupamento. Neste trabalho, construímos a ideia que os módulos de coexpressão podem ser usados diretamente para inferência de grupos celulares no conjunto de dados (“tipos celulares”). Os genes coexpressos têm, de forma geral, função semelhante e, quando usados para dividir o conjunto de dados, são capazes de fazer separações funcionais. Aos arquétipos celulares funcionais, damos os nomes de tipos celulares. Logo, aqui mostramos que o uso de módulos de coexpressão pode dar uma faceta funcional ao agrupamento não supervisionado de células.

- Agrupamento de múltiplos níveis permitem uma análise mais completa da diversidade de tipos celulares.

Em contraste com as rotulações clássicas de uma célula, um tipo celular, trabalhamos a ideia que cada célula pertence a diversos tipos celulares. Esses tipos podem ser aninhados (como no caso de linfócitos “T CD4” e o supergrupo “linfócitos T”) ou sobrepostos (como no caso de “linfócitos” e “células apresentadoras de antígeno”, que se sobrepõem nas “células B”). O uso de módulos de coexpressão como pontos de partida para atribuir tipos celulares permite, de forma natural, revelar as múltiplas facetas de tipos celulares.

Em suma, esperamos que os pacotes do R/Bioconductor desenvolvidos e as ponderações teóricas detalhadas acima contribuam para o avanço das análises no campo da transcritômica de células únicas.

Referências

- AMEZQUITA, Robert A et al. Orchestrating single-cell analysis with Bioconductor. en. **Nat. Methods**, v. 17, n. 2, p. 137–145, fev. 2020.
- BECHT, Etienne et al. Dimensionality reduction for visualizing single-cell data using UMAP. en. **Nat. Biotechnol.**, dez. 2018.
- BERG, Jeremy. **Exploring organisms cell by cell**. [S.l.]: American Association for the Advancement of Science, 2018.
- BROOKS, C F; MOORE, M. Differential MHC class II expression on human peripheral blood monocytes and dendritic cells. en. **Immunology**, v. 63, n. 2, p. 303–311, fev. 1988.
- BUNBURY, Allyson et al. Functional analysis of monocyte MHC class II compartments. **The FASEB Journal**, Federation of American Societies for Experimental Biology, v. 23, n. 1, p. 164–171, 2009.
- BUTLER, Andrew et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. en. **Nat. Biotechnol.**, v. 36, n. 5, p. 411–420, jun. 2018.
- CARRETTE, Florent; SURH, Charles D. IL-7 signaling and CD127 receptor regulation in the control of T cell homeostasis. In: ELSEVIER, 3. SEMINARS in immunology. [S.l.: s.n.], 2012. v. 24, p. 209–217.
- CHEN, Shuonan; MAR, Jessica C. Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. **BMC bioinformatics**, BioMed Central, v. 19, n. 1, p. 232, 2018.
- CROW, Megan; GILLIS, Jesse. Co-expression in single-cell analysis: Saving grace or original sin? **Trends in Genetics**, Elsevier, v. 34, n. 11, p. 823–831, 2018.
- DAM, Sipko van et al. Gene co-expression analysis for functional classification and gene–disease predictions. **Briefings in Bioinformatics**, bbw139, 2017.
- FABREGAT, Antonio et al. The reactome pathway knowledgebase. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. d481–d487, 2016.
- FARRELL, Jeffrey A et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. en. **Science**, American Association for the Advancement of Science, v. 360, n. 6392, jun. 2018.
- FRANZÉN, Oscar; GAN, Li-Ming; BJÖRKEGREN, Johan L M. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. en. **Database**, v. 2019, jan. 2019.
- HARTIGAN, J A; WONG, M A. Algorithm AS 136: A K-Means Clustering Algorithm. **Applied Statistics**, v. 28, n. 1, p. 100, 1979.
- JOCHEMS, Simon P et al. Inflammation induced by influenza virus impairs human innate immune control of pneumococcus. en. **Nat. Immunol.**, v. 19, n. 12, p. 1299–1308, dez. 2018.
- KOBAK, Dmitry; LINDERMAN, George C. UMAP does not preserve global structure any better than t-SNE when using the same initialization. **bioRxiv**, Cold Spring Harbor Laboratory, 2019.

- KORTHAUER, Keegan D et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. en. **Genome Biol.**, v. 17, n. 1, p. 222, out. 2016.
- LÄHNEMANN, David et al. Eleven grand challenges in single-cell data science. en. **Genome Biol.**, v. 21, n. 1, p. 31, fev. 2020.
- LANGFELDER, Peter; HORVATH, Steve. WGCNA: an R package for weighted correlation network analysis. **BMC bioinformatics**, Springer, v. 9, n. 1, p. 559, 2008.
- LÊ CAO, Kim-Anh; BOITARD, Simon; BESSE, Philippe. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. en. **BMC Bioinformatics**, v. 12, p. 253, jun. 2011.
- LEVINE, Jacob H et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. **Cell**, Elsevier, v. 162, n. 1, p. 184–197, 2015.
- LUECKEN, Malte D; THEIS, Fabian J. Current best practices in single-cell RNA-seq analysis: a tutorial. en. **Mol. Syst. Biol.**, v. 15, n. 6, e8746, jun. 2019.
- MARGOLIN, Adam A et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. v. 7, S1, s7, 2006.
- MARGUERAT, Samuel; BÄHLER, Jürg. Coordinating genome expression with cell size. en. **Trends Genet.**, v. 28, n. 11, p. 560–565, nov. 2012.
- MICHLMAYR, Daniela et al. Comprehensive innate immune profiling of chikungunya virus infection in pediatric cases. en. **Mol. Syst. Biol.**, v. 14, n. 8, e7862, ago. 2018.
- MRDJEN, Dunja et al. High-Dimensional Single-Cell Mapping of Central Nervous System Immune Cells Reveals Distinct Myeloid Subsets in Health, Aging, and Disease. en. **Immunity**, v. 48, n. 3, p. 599, mar. 2018.
- NAZZICARI, Nelson et al. MTGO-SC, A Tool to Explore Gene Modules in Single-Cell RNA Sequencing Data. en. **Front. Genet.**, v. 10, p. 953, out. 2019.
- NELSON, Andrew C et al. In Vivo Regulation of the Zebrafish Endoderm Progenitor Niche by T-Box Transcription Factors. en. **Cell Rep.**, v. 19, n. 13, p. 2782–2795, jun. 2017.
- NEWMAN, Aaron M et al. Robust enumeration of cell subsets from tissue expression profiles. en. **Nat. Methods**, v. 12, n. 5, p. 453–457, mai. 2015.
- PALMER, Chana et al. Cell-type specific gene expression profiles of leukocytes in human peripheral blood. en. **BMC Genomics**, BioMed Central, v. 7, p. 115, 2006.
- RODIG, Scott J et al. The pre-B-cell receptor associated protein VpreB3 is a useful diagnostic marker for identifying c-MYC translocated lymphomas. **haematologica**, Haematologica, v. 95, n. 12, p. 2056–2062, 2010.
- ROHART, Florian et al. mixOmics: An R package for 'omics feature selection and multiple data integration. en. **PLoS Comput. Biol.**, v. 13, n. 11, e1005752, nov. 2017.
- RUSSO, Pedro S T et al. CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses. en. **BMC Bioinformatics**, v. 19, n. 1, p. 56, fev. 2018.
- SAEYS, Yvan; VAN GASSEN, Sofie; LAMBRECHT, Bart N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. en. **Nat. Rev. Immunol.**, v. 16, n. 7, p. 449–462, jul. 2016.
- SERIN, Elise A R et al. Learning from Co-expression Networks: Possibilities and Challenges. en. **Front. Plant Sci.**, v. 7, p. 444, abr. 2016.
- SHALEK, Alex K et al. Single cell RNA Seq reveals dynamic paracrine control of cellular variation. en. **Nature**, NIH Public Access, v. 510, n. 7505, p. 363, jun. 2014.
- SHANNON, Claude E. A mathematical theory of communication. **Bell system technical journal**, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948.

- SHMULEVICH, Ilya; ZHANG, Wei. Binary analysis and optimization-based normalization of gene expression data. en. **Bioinformatics**, v. 18, n. 4, p. 555–565, abr. 2002.
- SKINNIDER, Michael A; SQUAIR, Jordan W; FOSTER, Leonard J. Evaluating measures of association for single-cell transcriptomics. en. **Nat. Methods**, v. 16, n. 5, p. 381–386, mai. 2019.
- SUBRAMANIAN, Aravind et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. en. **Proc. Natl. Acad. Sci. U. S. A.**, v. 102, n. 43, p. 15545–15550, out. 2005.
- TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 58, n. 1, p. 267–288, 1996.
- TREUTLEIN, Barbara et al. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. en. **Nature**, v. 509, n. 7500, p. 371–375, mai. 2014.
- VARGAS-GARCIA, Cesar A; GHUSINGA, Khem Raj; SINGH, Abhyudai. Cell size control and gene expression homeostasis in single-cells. en. **Curr Opin Syst Biol**, v. 8, p. 109–116, abr. 2018.
- VILLANI, A C et al. Single-cell RNA-seq Reveals New Types of Human Blood Dendritic Cells, Monocytes, and Progenitors. **Science**, Science, v. 356, n. 6335, abr. 2017.
- WEIDINGER, Gilbert et al. **The Sp1-Related Transcription Factors sp5 and sp5-like Act Downstream of Wnt/ β -Catenin Signaling in Mesoderm and Neuroectoderm Patterning**. v. 15. [S.l.: s.n.], 2005. p. 489–500.
- WOLF, F Alexander; ANGERER, Philipp; THEIS, Fabian J. SCANPY: large-scale single-cell gene expression data analysis. en. **Genome Biol.**, v. 19, n. 1, p. 15, fev. 2018.
- WOLOCK, Samuel L; LOPEZ, Romain; KLEIN, Allon M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. en. **Cell Syst**, v. 8, n. 4, 281–291.e9, abr. 2019.
- YABE, Taijiro; TAKADA, Shinji. Mesogenin causes embryonic mesoderm progenitors to differentiate during development of zebrafish tail somites. en. **Dev. Biol.**, v. 370, n. 2, p. 213–222, out. 2012.
- YU, Lei; LIU, Huan. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. **Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)**, 2003.
- ZHUANG, Yuan; CHENG, Peifeng; WEINTRAUB, Harold. B-lymphocyte development is regulated by the combined dosage of three basic helix-loop-helix genes, E2A, E2-2, and HEB. **Molecular and cellular biology**, Am Soc Microbiol, v. 16, n. 6, p. 2898–2905, 1996.