UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE MICROBIOLOGIA, IMUNOLOGIA E PARASITOLOGIA
CURSO DE CIÊNCIAS BIOLÓGICAS

Giovanni Alberto Crestani

UM FLUXO DE ANÁLISE QUANTITATIVA DE DADOS DE TRANSCRIPTÔMICA DE CÉLULAS ÚNICAS NO CONTEXTO DE CÉLULAS-TRONCO PLURIPOTENTES INDUZIDAS

Florianópolis
2022

Giovanni Alberto Crestani

UM FLUXO DE ANÁLISE QUANTITATIVA DE DADOS DE TRANSCRIPTÔMICA DE CÉLULAS ÚNICAS NO CONTEXTO DE CÉLULAS-TRONCO PLURIPOTENTES INDUZIDAS

Trabalho de Conclusão de Curso do curso de Graduação em Ciências Biológicas do Centro de Ciências Biológicas da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Licenciado em Ciências Biológicas.

Orientador: Prof. Dr. Edroaldo Lummertz da Rocha

Florianópolis

2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Crestani, Giovanni Alberto
Um fluxo de análise quantitativa de dados de transcriptômica de células únicas no contexto de células tronco pluripotentes induzidas / Giovanni Alberto Crestani ; orientador, Edroaldo Lummertz da Rocha, 2022.
65 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Graduação em Ciências Biológicas, Florianópolis, 2022.

Inclui referências.

1. Ciências Biológicas. 2. Células-tronco. 3. Célula única. 4. Pipeline. 5. Bioinformática. I. Lummertz da Rocha, Edroaldo. II. Universidade Federal de Santa Catarina. Graduação em Ciências Biológicas. III. Título.

Giovanni Alberto Crestani

UM FLUXO DE ANÁLISE QUANTITATIVA DE DADOS DE TRANSCRIPTÔMICA DE CÉLULAS ÚNICAS NO CONTEXTO DE CÉLULAS-TRONCO PLURIPOTENTES INDUZIDAS

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Licenciado em Ciências Biológicas e aprovado em sua forma final pelo Curso de Ciências Biológicas.

Florianópolis, 09 de março de 2022.

Documento assinado digitalmente
Daniela Cristina de Toni
Data: 17/03/2022 12:03:12-0300
CPF: 774.080.129-15
Verifique as assinaturas em https://v.ufsc.br

_____

Profª. Daniela Cristina de Toni, Drª.

Coordenadora do Curso

Banca Examinadora:

Documento assinado digitalmente
Edroaldo Lummertz da Rocha
Data: 16/03/2022 09:46:57-0300
CPF: 010.282.360-00
Verifique as assinaturas em https://v.ufsc.br

_____

Prof. Edroaldo Lummertz da Rocha, Dr.

Orientador

Universidade Federal de Santa Catarina

Documento assinado digitalmente
Alfeu Zanotto Filho
Data: 15/03/2022 17:03:25-0300
CPF: 007.553.300-64
Verifique as assinaturas em https://v.ufsc.br

_____

Prof. Alfeu Zanotto-Filho, Dr.

Avaliador

Universidade Federal de Santa Catarina

Documento assinado digitalmente
MARCELO LUIZ BRUNATTO FALCHETTI
Data: 17/03/2022 11:27:02-0300
CPF: 084.676.289-79
Verifique as assinaturas em https://v.ufsc.br

Marcelo Luiz Brunatto Falchetti. MSc.

Avaliador

Universidade Federal de Santa Catarina

Dedico este trabalho à toda minha família, que me mostrou o valor do conhecimento e do amor.

AGRADECIMENTOS

Agradeço ao meu orientador. Dr. Edroaldo Lummertz da Rocha, pela paciência, mentoria e amizade.

Agradeço a todos os professores, pós-doutorandos e estudantes de graduação que me mentoraram em todos os laboratórios que participei durante o curso.

Agradeço a todos os meus colegas de laboratório e de curso que de alguma forma cooperaram com a minha formação pessoal e profissional.

Agradeço à minha família por todo o suporte. Todos vocês são fundamentais para qualquer sucesso meu.

Agradeço à minha noiva, Maryane, pela paciência e animação. Tudo é melhor contigo.

Agradeço à OWT. Com vocês, posso ser eu mesmo.

Agradeço à UFSC como um todo, pela estrutura, acolhimento e instrução.

RESUMO

Células-tronco pluripotentes induzidas são células reprogramadas a partir de células somáticas de modo a adquirir pluripotência – a capacidade de se diferenciar em qualquer tipo de célula. Com um protocolo de diferenciação adequado, podemos transformá-las em diversas outras células do organismo. Desde sua criação, diversos avanços em protocolos e técnicas laboratoriais permitem seu uso em pesquisa e terapias celulares. Contudo, o processo de diferenciação é falho e nem todas as células se transformam nas células alvo intencionadas. Nesse contexto, o sequenciamento de transcriptômica de células únicas se mostra uma poderosa ferramenta para a obtenção de informações. Ferramentas de bioinformática são fundamentais nesse processo, nos permitindo analisar a expressão gênica de uma célula e inferir seu tipo celular. Diversas ferramentas são utilizadas em diferentes passos do processo de análise. De modo geral, essas ferramentas são reprodutíveis. No entanto, é comum que o usuário tenha dificuldades em instalar a ferramenta e utilizar *scripts* fora do contexto onde foram escritos. Para que não ocorram situações como essas, estruturamos o uso dessas ferramentas em uma *pipeline* de análise. Boas práticas de construção de *pipeline* mostram a necessidade de desenvolvê-la de forma modular, reprodutível e compartimentalizada. Para tal, é necessário o uso de ferramentas de gerenciamento de fluxo de trabalho e containers de dependências dos pacotes. Este trabalho buscou construir uma *pipeline* de análise de dados de transcriptômica de células únicas no contexto de células-tronco pluripotentes induzidas. Além disso, visou criar um *score* que avalia a importância que determinado gene teve na classificação de uma amostra. As ferramentas de análise utilizadas na *pipeline* foram FUSCA, singleCellNet, Seurat e Symphony. Os recursos utilizados para a construção da estrutura da *pipeline* foram o gerenciador de fluxo de trabalho Snakemake e o container Singularity. A avaliação de eficácia da *pipeline* foi medida com sua aplicação em dados de células únicas de neurônios dopaminérgicos derivados de células-tronco pluripotentes induzidas, utilizando um conjunto de dados de células da região ventral do mesencéfalo de embriões humanos. A *pipeline* foi capaz de identificar os tipos celulares das células em questão e esses foram compatíveis com a tipagem feita pelos autores. As figuras geradas são acessíveis e podem ser utilizadas para a construção de um relatório ou trabalho científico. Por fim,

a *pipeline* está disponível para acesso e uso público em https://github.com/gacrestani/ipsc-pipeline.
Palavras-chave: células-tronco; células únicas; transcriptômica; pipeline; bioinformática;

ABSTRACT

Induced pluripotent stem cells (iPSCs) are cells reprogrammed from somatic cells to acquire pluripotency – the ability to differentiate into any cell type of an organism. With a differentiation protocol, one can transform them into those several other cells. Since their creation, several advances in laboratory protocols and techniques allow their use in biomedical research and cell therapies. However, the differentiation process is flawed and not all cells turn into the intended target cells. In this context, single cell transcriptomics sequencing proves to be a powerful tool for obtaining information. Bioinformatics tools are fundamental in this process, allowing us to analyze the gene expression of a cell and, by it, infer its cell type. Several tools are used in different steps of the analysis process. In general, these tools are reproducible. However, it is common for the user to have difficulties installing the tool and using scripts outside the context where they were written. To minimize those situations, we have structured the use of these tools in an analysis pipeline. Good pipeline construction practices state the need to develop it in a modular, reproducible and compartmentalized way. To do so, it is necessary to use workflow management tools and package dependency containers. This work aimed to build a pipeline for analyzing single cell transcriptomics data in the context of induced pluripotent stem cells. In addition, it aimed to create a score that assesses the importance that a given gene had in the classification of a sample. The analysis tools used in the pipeline were FUSCA, singleCellNet, Seurat and Symphony. The resources used to build the pipeline structure were the workflow manager Snakemake and the container manager Singularity. The evaluation of the effectiveness of the pipeline was measured with its application to single cell data from dopaminergic neurons derived from induced pluripotent stem cells, using a dataset of cells from the ventral region of the midbrain of human embryos. The pipeline was able to identify the cell types of the cells in question and these were compatible with the types found by the authors. The generated figures are accessible and can be used to build a report or scientific work. Finally, the pipeline is available for public use on https://github.com/gacrestani/ipsc-pipeline.

Keywords: stem cells; single cell; transcriptomics; pipeline; bioinformatics;

LISTA DE FIGURAS

LISTA DE QUADROS

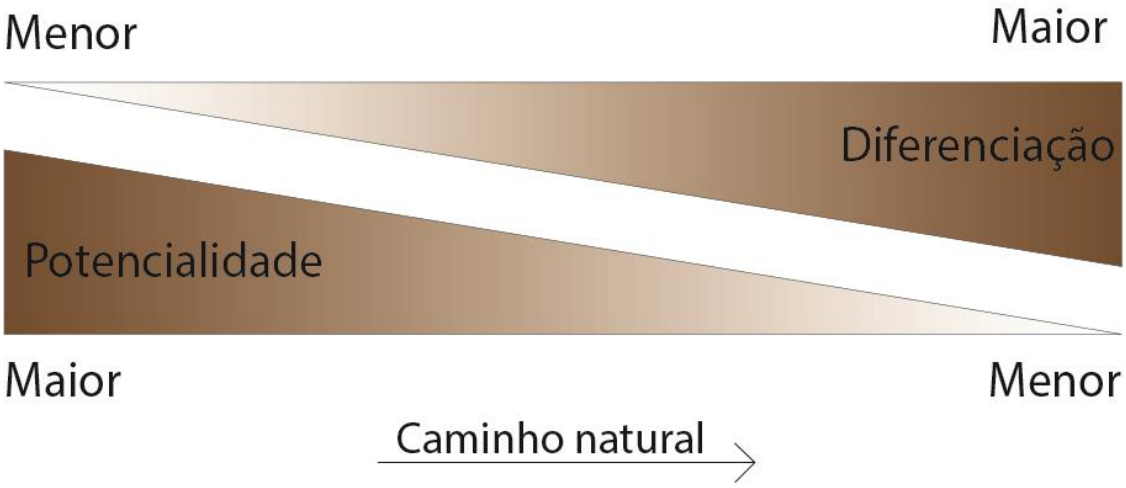
SUMÁRIO

1 INTRODUÇÃO

1.1 CÉLULAS-TRONCO

Células-tronco são células não especializadas que possuem a capacidade de se diferenciar em qualquer célula de um determinado organismo, além da habilidade de auto renovação por tempo indeterminado. Essas células estão presentes em todas as fases de desenvolvimento de um indivíduo, desde o embrião, até o adulto. As células-tronco possuem um gradiente de diferenciação – quanto mais especializada é a célula, menor será o seu potencial de diferenciação (ZAKRZEWSKI et al., 2019). A Figura 1 ilustra esse gradiente.

Figura 1 – Gradiente de potencialidade e diferenciação de células-tronco.



Quanto mais à direita do gradiente, maior a potencialidade e menor o grau de diferenciação. Quanto mais à esquerda do gradiente, maior o grau de diferenciação e menor a potencialidade.

Fonte: Adaptado de Junqueira, 2012.

As células-tronco podem ser classificadas de acordo com o seu potencial de diferenciação. As classes, em ordem decrescente de potencialidade são totipotentes, pluripotentes, multipotentes, oligopotentes e onipotentes. Cada uma destas classes tem impactos diferentes e atua em de forma diferente em diferentes etapas da vida do organismo. Além de serem o primeiro estágio de quaisquer células de um indivíduo, as células-tronco também são responsáveis por reparos de tecidos, como um

processo teoricamente ilimitado ao longo da vida do organismo. (ZAKRZEWSKI et al., 2019).

Com o avanço do conhecimento acerca de células-tronco, surgem propostas para utilizá-las como ferramentas em tratamentos para doenças e síndromes. Temos como exemplos terapias ortopédicas (ELBULUK; EINHORN; IORIO, 2017), terapias estéticas (TREMOLADA; COLOMBO; VENTURA, 2016) e até mesmo terapias celulares para tratamento de COVID-19 e síndrome do desconforto respiratório agudo (ZANIRATI et al., 2021), dentre outras inúmeras possibilidades do campo da medicina celular.

Contudo, diversos aspectos do uso de células-tronco para terapias devem ser considerados. Em primeiro lugar, devemos considerar a origem destas células: células pluripotentes podem ser obtidas a partir de embriões, chamadas células-tronco embrionárias e possuem a capacidade de se diferenciar em células de quaisquer folhetos embrionários, dando a elas um extraordinário potencial para terapias (ZAKRZEWSKI et al., 2019). No entanto, aspectos éticos devem ser considerados ao se trabalhar com embriões humanos, mesmo que em cultura (LO; PARHAM, 2009). Para a obtenção destas células, há um processo complexo onde o embrião é destruído ao final. Em alguns países, embriões pré-implantação não são considerados seres humanos; em outros, a legislação permite a criação ou uso de embriões para pesquisa e propósitos terapêuticos. Contudo, a maioria das nações não concorda com estas diretrizes (MORADI et al., 2019). Desta forma, a pesquisa e aplicação de células-tronco embrionárias é fortemente freada por problemas éticos.

Como alternativa, podemos considerar as células-tronco originadas a partir de um indivíduo adulto. Essas células não são pluripotentes, mas sim multipotentes, reduzindo a sua gama de aplicabilidades na medicina e na pesquisa. Um exemplo são as células-tronco mesenquimais, que são originadas em vários tecidos, como cordão umbilical, pólipos endometriais, medula óssea, tecido adiposo, pele, dentre outros. Essas fontes de células-tronco abrem a possibilidade de tratamentos autólogos, com o uso de células de um indivíduo para realizar seu o próprio tratamento. Como outra vantagem, temos também a redução da taxa de rejeição de terapia desta natureza (DING; SHYU; LIN, 2011).

Essa área da biologia celular ainda está em sua infância, com muitas possibilidades e desafios a frente. Apesar disso, já existem alternativas inovadoras em desenvolvimento. Um dos principais avanços atualmente é a possibilidade de

desdiferenciar uma célula; isto é, fazê-la traçar o caminho contrário ao natural, aumentando seu potencial e reduzindo sua especialização.

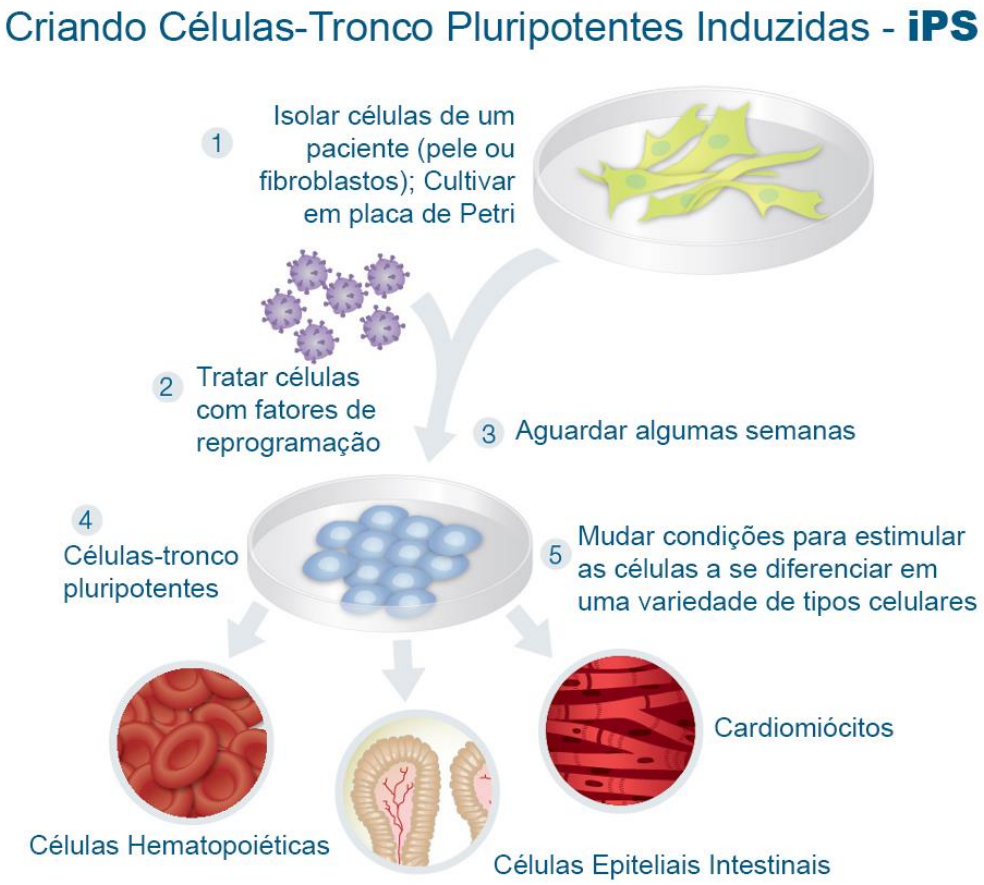
1.1.1 Células-tronco induzidas

A criação de células-tronco pluripotentes induzidas tem suas raízes nos trabalhos Sir John B. Gurdon. O pesquisador foi o primeiro a mostrar que as células, apesar de estarem no ápice de sua especialização, ainda possuem toda a informação gênica e são capazes de serem reprogramadas em células-tronco novamente. Isso foi provado ao se substituir o núcleo de células do ovo pelo núcleo de células epiteliais do intestino de *Xenopus laevis laevis* (rã-de-unhas-africana), dando origem a girinos saudáveis (GURDON, 1962). Sessenta anos depois, Yamanaka descreveu como células maduras de camundongo podem ser reprogramadas para tornarem-se células-tronco pluripotentes (TAKAHASHI; YAMANAKA, 2006). Por esses trabalhos, Gurdon e Yamanaka receberam o Prêmio Nobel de Fisiologia e Medicina em 2012.

Pouco tempo depois, Takahashi e seus colaboradores conseguiram realizar o mesmo feito com células humanas. Tais células eram similares às células embrionárias nos aspectos morfologia, proliferação, antígenos de superfície, expressão genética, condição epigenética de genes específicos para células pluripotentes, e atividade de telomerase – tais informações foram coletadas com técnicas moleculares e de microscopia. As células desdiferenciadas podem ser mantidas em cultura indefinidamente e produzir células dos três folhetos embrionários. (TAKAHASHI et al., 2007).

Os protocolos de geração de células-tronco induzidas envolvem a exposição de células como fibroblastos dermais (células de fácil obtenção) a meios de cultura específicos para a desdiferenciação (CHAMBERS et al., 2009) e fatores de transcrição que controlam redes de regulação gênicas responsáveis pela pluripotência (HOCHEDLINGER; PLATH, 2009), além da transfecção de genes com expressão importante para essas células. Tal procedimento tem o sucesso confirmado a partir da expressão de alguns antígenos marcadores para células-tronco (como TRA-1-81, TRA-1-60, SSEA-4 e NANOG) por citometria de fluxo e imunofluorescência (SCHOPPERLE; DEWOLF, 2007). A Figura 2 ilustra o procedimento de geração de células-tronco induzidas a partir de fibroblastos.

Figura 2 – Esquema ilustrativo de geração de células-tronco induzidas



Protocolo de criação de células tronco e diferenciação.

Fonte: Traduzido de IPSC21 (2022).

Desde a descoberta das células desdiferenciadas, avanços na ciência expandiram os tipos celulares-alvo para os quais podemos reprogramar células (MADRID et al., 2021). Hoje, existem protocolos para diferenciação de células para células-alvo de diversos tipos, como neuronais (KRIKS et al., 2011), renais (TAKASATO et al., 2015), cardíacas (OU et al., 2021), dentre outras.

Essas células abrem novas fronteiras na pesquisa em medicina regenerativa e biologia de células-tronco. Devido a elas, vemos novas fronteiras na modelagem de doenças, desenvolvimento de novas drogas, descoberta de mecanismos patológicos, dentre outras áreas. Mais recentemente, o desenvolvimento de organoides – pequenos órgãos tridimensionais, abrem ainda novas possibilidades de pesquisa nessas áreas (SHI et al., 2017). Ainda assim, implicações éticas e legais continuam presentes. O código de Nuremberg (1947) e a Declaração de Helsinki (1964), criados para restringir a pesquisa não-ética com seres humanos, se mostram desatualizados no contexto de células-tronco, visto que pesquisa com embriões não eram levadas em

consideração naquela época. É importante a criação de novas diretrizes e regulamentações, mantendo a ética no mesmo passo do avanço tecnológico (MORADI et al., 2019).

Apesar de todas as vantagens, os desafios para a reprogramação celular são abundantes. Por exemplo, um problema frequente é a baixa eficácia dos protocolos de diferenciação – no final do procedimento, nem todas as células se diferenciam nas células-alvo desejadas (EBRAHIMI, 2015). Isso se mostra um obstáculo para a aplicação clínica destas terapias, com ferramentas de análise genômica tendo papel fundamental nesta situação.

1.2 SEQUENCIAMENTO DE ÁCIDOS NUCLEICOS

Em 1953, Watson e Crick resolveram a estrutura tridimensional do ácido desoxirribonucleico (DNA), trabalhando com dados de cristalografia gerados por Rosalind Franklin e Maurice Wilkins (WATSON; CRICK, 1953). Contudo, muito tempo se passou até conseguirmos verificar a sequência de uma molécula de DNA. Esforços iniciais focavam em desvendar o genoma de vírus de fita simples de ácido ribonucleico (RNA) e outras espécies de RNA de isolamento relativamente simples. No entanto, esses esforços produziam apenas medidas quantitativas dos nucleotídeos, mas não sua sequência (HOLLEY et al., 1961). Eventualmente, em 1965, Robert Holley e seus colaboradores foram capazes de criar a primeira sequência de ácidos nucleicos, que codificava para o RNA transportador de alanina de *Saccharomyces cerevisiae* (HOLLEY et al., 1965). Em paralelo, Frederick Sanger e seus colaboradores trabalhavam em um método relacionado, baseado na fragmentação bidimensional de sequências nucleicas (SANGER; BROWNLEE; BARRELL, 1965). Com o passar do tempo, a técnica de Sanger foi sendo aprimorada, dando origem ao Método de Sequenciamento de Sanger, tecnologia amplamente e comumente usada para sequenciar DNA durante anos (HEATHER; CHAIN, 2016).

Esses métodos compõem a primeira geração de sequenciamento. Passamos pela segunda geração de métodos de sequenciamento e durante esse período grandes empresas foram criadas e desenvolvidas no ramo de equipamentos e serviços de sequenciamento. Hoje, a ciência se encontra na terceira geração de métodos de sequenciamentos. O avanço tecnológico foi rampante, chegando até no desenvolvimento de aparelhos do tamanho de um dispositivo USB que são capazes de gerar dados de sequenciamento fidedignos (LOMAN; QUICK; SIMPSON, 2015).

Acima disso, outras técnicas de sequenciamento permitem a extração de outro tipos de informações, como RNA (RNA-Seq), interações proteína-DNA (ChIP-Seq) (GONG et al., 2018), acessibilidade de cromatina (ATAC-Seq) (BUENROSTRO et al., 2015), dentre outros. Para este trabalho, os dados utilizados foram gerados a partir de RNA sequenciado de células únicas.

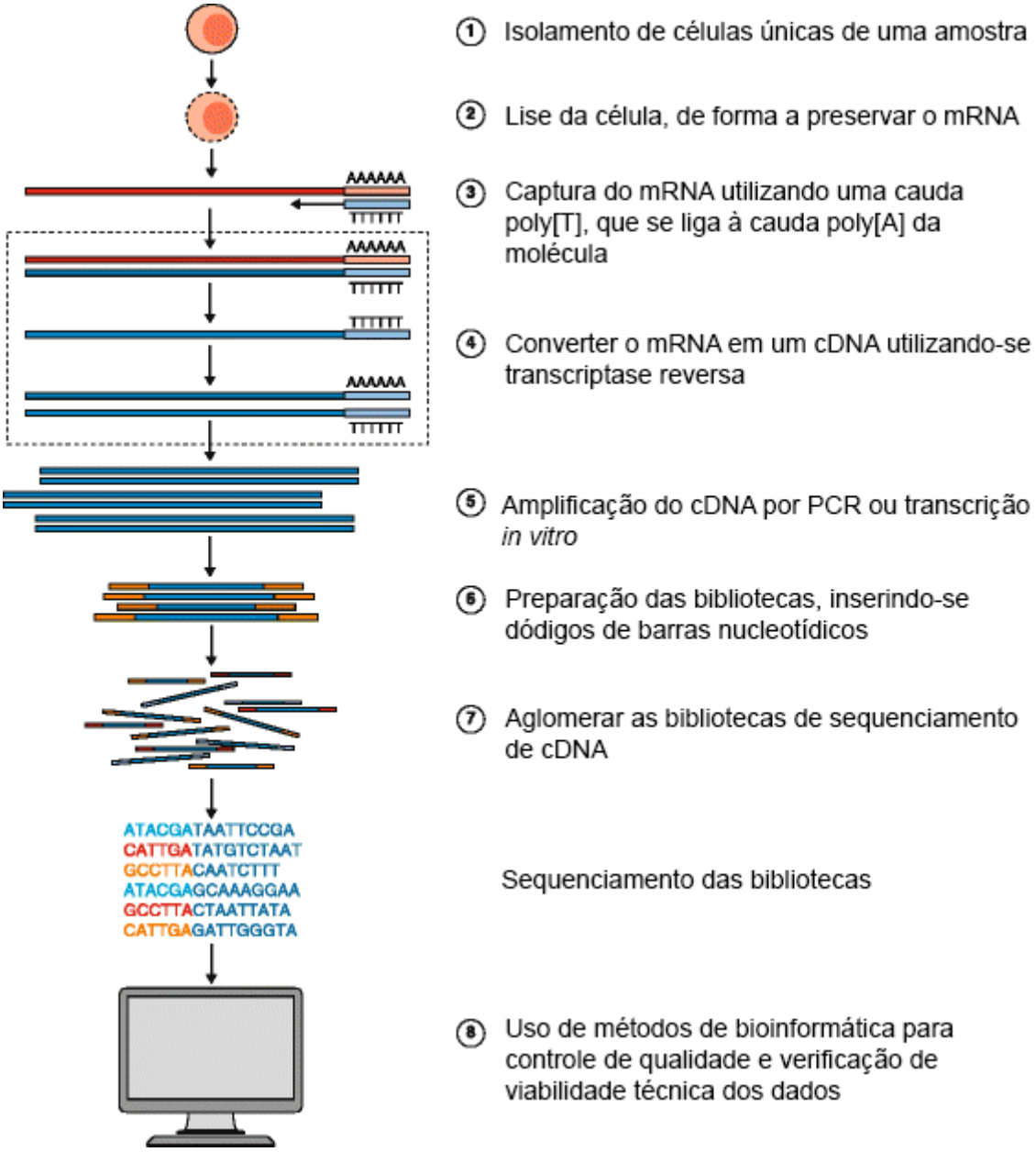
1.2.1 Sequenciamento de RNA de células únicas

O primeiro trabalho que produziu dados transcriptômicos completos de uma única célula foi publicado por Tang et al, em 2009 (TANG et al., 2009). Desde então, o sequenciamento de RNA de células únicas se tornou um dos métodos mais utilizados para a geração de dados de transcriptômica em alta resolução. Contudo, devido a limitações técnicas e fatores biológicos, os dados de sequenciamento de RNA de células únicas são mais complexos do que os de sequenciamento tradicional de RNA – a natureza esparsa de expressão gênica os torna computacionalmente e estatisticamente desafiadores de se analisar. Por exemplo, dados *outliers* muitas vezes podem ser confundidos com artefatos do método (JIANG; THOMSON; STEWART, 2016) (CHEN; NING; SHI, 2019).

Para a geração de dados de RNA de células únicas, é necessário o isolamento de células individuais. Diversos métodos foram desenvolvidos para esse isolamento, incluindo diluição limitante, micromanipulação, dentre outros. Cada protocolo de isolamento possui suas vantagens e desvantagens (HWANG; LEE; BANG, 2018), sendo apropriados em diferentes situações.

Após o isolamento, é feita a lise da célula, captura do RNA mensageiro (geralmente usando uma cauda *poly[A]*). Com esse material, utilizamos uma enzima de transcriptase reversa para geração de DNA complementar (cDNA). Depois é feita a amplificação do cDNA, formação das bibliotecas, controles de qualidade das amostras e, por fim, o sequenciamento do material. Então, já *in silico*, utilizam-se métodos de bioinformática para verificar a qualidade dos dados gerados, filtrando aqueles que não atingem critérios quantitativos (HAQUE et al., 2017). A Figura 3 exemplifica um fluxo de trabalho clássico para a geração de dados de transcriptômica de células únicas.

Figura 3 – Fluxo de trabalho clássico para geração de dados de transcriptômica de células únicas



Este fluxo é o modelo simplificado dos passos de se sequenciar RNA de células únicas. Este trabalho se enquadra no passo 8.

Fonte: Adaptado de HAQUE et al., 2017.

O sequenciamento de células únicas é aplicado em diversas áreas da biologia, principalmente na medicina de precisão. Por exemplo, avanços na área de pesquisas com câncer permitem a descoberta de diversos perfis de expressão em um único tumor (DARMANIS et al., 2017). Previamente, com dados de microarranjo ou sequenciamento de RNA clássico, essas particularidades eram perdidas. Avanços na área da imunologia demonstram melhor as dinâmicas de formação do sistema

hematopoiético (CHEN et al., 2020), dinâmicas imunes em doenças infecciosas (DELGOBO et al., 2019); além de outras inúmeras descobertas em diversas outras áreas da biologia.

Dentro desse cenário, a geração de dados vem crescendo de forma exponencial. Graças o barateamento de técnicas, hoje pequenos grupos conseguem gerar seus próprios dados. A análise desses dados é possível por conta do desenvolvimento de *softwares* por grupos de bioinformática especializados. Estas ferramentas são geralmente de código aberto e empregadas em diferentes passos do complexo fluxo de trabalho das análises (WRATTEN; WILM; GÖKE, 2021).

Entretanto, bioinformatas frequentemente encontram obstáculos no uso dessas ferramentas. Diferentes sistemas operacionais, recursos computacionais e versionamento/documentação de ferramentas criam um problema sério de reprodutibilidade dos resultados e experimentos em biologia computacional (GRÜNING et al., 2018). Em resposta a essas situações, surgem as *pipelines* de análise de dados.

1.3 *PIPELINES* DE ANÁLISE DE DADOS

Pipeline é um termo do inglês que, em tradução literal, significa “encanamento”. Contudo, no contexto da informática (onde não há tradução adequada), significa o uso sucedido de *softwares* e ferramentas, onde o *output* de uma ferramenta serve como *input* da próxima, até se chegar no resultado final (WRATTEN; WILM; GÖKE, 2021).

Uma *pipeline* tem um propósito definido, com um *output* padronizado (que pode ser uma figura, tabela, arquivo, relatório, etc), gerado a partir do *input* que a *pipeline* receber. Diversas *pipelines* estão disponíveis em repositórios de *software* como o Github (https://github.com/pditommaso/awesome-pipeline), e seus acessos são permitidos e abertos para todos.

Historicamente, tais *pipelines* eram desenvolvidas em *scripts* customizados ou do tipo *Make* (LEIPZIG, 2017). Esses métodos tradicionais apresentam diversas desvantagens, como o alto acoplamento à arquitetura do sistema onde foi desenvolvida, incapacidade de retomar um trabalho interrompido, falta de documentação, falta de rastreamento de parâmetros, dentre outros (WRATTEN; WILM; GÖKE, 2021).

Para tal, surgem as ferramentas de gerenciamento de fluxo de trabalho, capazes de auxiliar no desenvolvimento, uso e modificação de *pipelines*.

1.3.1 Ferramentas de gerenciamento de fluxo de trabalho

Ferramentas de gerenciamento de fluxo de trabalho (em inglês, *workflow managers*) são *softwares* desenvolvidos com o intuito de facilitar a criação, customização e reprodução de *pipelines*, auxiliando na solução do problema de reprodutibilidade (WRATTEN; WILM; GÖKE, 2021).

Empresas privadas que trabalham com uma grande quantidade de dados, como AirBNB e Netflix, desenvolveram suas ferramentas de gerenciamento de fluxo de trabalho dedicadas, como por exemplo *Airflow* (https://airbnb.io/projects/airflow/) e *Metaflow* (https://github.com/Netflix/metaflow). Tais ferramentas possibilitam a criação de *pipelines* para diversos tipos de dados. Isso serve para ilustrar a importância desta padronização, bem como da capacidade de reproduzir resultados.

Recentemente, diversas ferramentas de gerenciamento de fluxo de trabalho foram criadas especificamente para dados biomédicos (LEIPZIG, 2017). Essas ferramentas oferecem integração com containers, servidores de alta capacidade computacional, gerenciamento de pacotes, dentre outros benefícios (PERKEL, 2019). Uma *pipeline* implementada dentro de uma ferramenta de gerenciamento de fluxo de trabalho é facilmente desenvolvida, mantida, usada e editada; além de ter melhor portabilidade e reprodutibilidade (WRATTEN; WILM; GÖKE, 2021).

1.4 OBJETIVOS

1.4.1 Objetivo geral

Desenvolvimento de uma *pipeline* de análise quantitativa de dados transcriptômicos de células únicas, voltada para a avaliação de experimentos de diferenciação a partir de células-tronco pluripotentes induzidas.

1.4.2 Objetivos Específicos

- Construir uma *pipeline* que considere diferentes ferramentas e métodos estatísticos de classificação do tipo celular de conjuntos de dados de células únicas;
- Implementar a *pipeline* dentro de uma ferramenta de gerenciamento de fluxo de trabalho, facilitando sua reprodutibilidade;
- Criação de um container capaz de executar a *pipeline* de forma compartimentalizada;
- Criação de um *score* para avaliar a importância que determinados genes tem para a classificação de células únicas em seus respectivos tipos celulares.

2 DESENVOLVIMENTO

2.1 MATERIAL E MÉTODOS

O presente trabalho foi elaborado a partir de dados abertos e de ferramentas de código aberto, publicados em artigo de periódicos da área, construídas com o propósito de analisar dados transcriptômicos de células únicas. Desta forma, os únicos materiais físicos utilizados foram computador conectado à internet, dentro da rede da UFSC (ou em VPN) para se ter acesso à bases de dados de artigos científicos e se utilizar a capacidade computacional do computador do laboratório.

Vale ressaltar que a análise contida nessa *pipeline* parte de uma tabela de expressão de genes por células, na normalização de *counts* – quantidade de *reads* para cada gene. O controle de qualidade de *reads*, alinhamento e outras análises prévias não são abordadas.

2.1.1 Ferramentas de análise de dados

Na *pipeline* desenvolvida, foram adicionadas quatro ferramentas de análise de dados transcriptômicos de células únicas. Estas ferramentas são amplamente utilizadas pela comunidade científica para análises (DING et al., 2021). Cada uma utiliza um algoritmo estatístico e computacional diferente, resultando em uma convergência de informações acerca dos dados. Todos os *scripts* das ferramentas em questão foram escritos baseados em suas *vignettes*, ou tutoriais - portanto, são otimizadas para seus propósitos. São publicadas (ou aguardando publicação), de código aberto e implementadas na linguagem de programação R (R Core Team, 2017).

2.1.1.1 singleCellNet

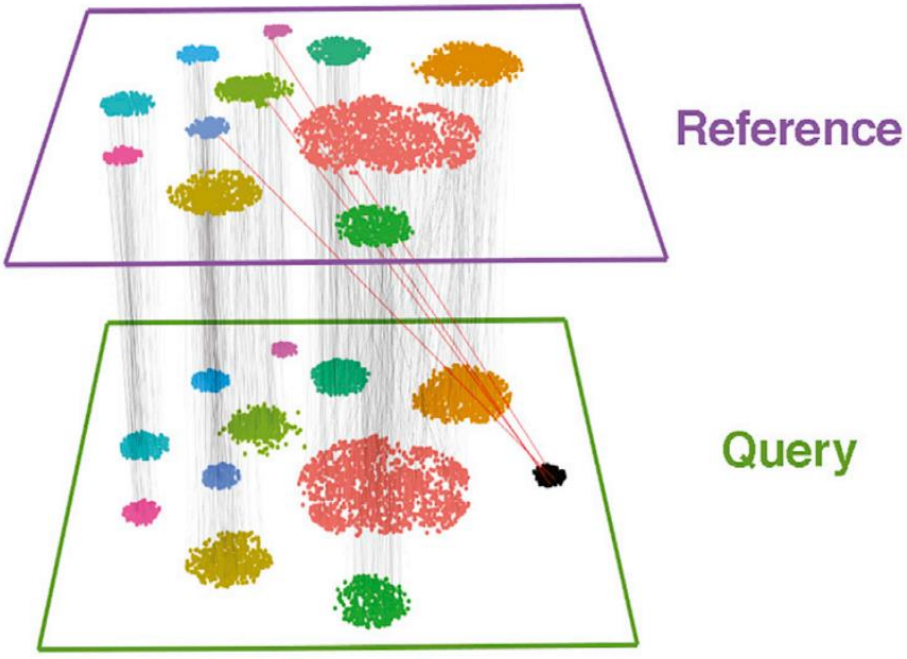
A ferramenta singleCellNet (TAN; CAHAN, 2019) foi desenvolvida especificamente para análise de dados de transcriptoma de células únicas. Ela auxilia na classificação do tipo celular de cada célula. Esta ferramenta constrói um classificador a partir de um conjunto de dados com os tipos celulares conhecidos, para depois classificar um conjunto de dados com tipos celulares desconhecidos (nosso conjunto de dados questão. É importante que ambos estejam no mesmo contexto, isto é, pertençam a situações amostrais semelhantes (mesmo órgão e mesma condição biológica). O algoritmo de aprendizado de máquina, chamado *Random Forest*

(BREIMAN, 2001), utiliza pares de genes como características levadas em conta pelo classificador. A informação gerada para esses pares, a partir do valor da tabela de expressão, é qual dos dois genes foi mais expresso naquela célula. Assim, a matriz de decisão não é alimentada com valores de expressão, mas apenas com `0` se o segundo gene for mais expresso que o primeiro, e `1` para o caso contrário (GEMAN et al., 2004). Ao final, a ferramenta retorna informações e métricas que representam a identidade celular das células em questão, bem como a qualidade do classificador.

2.1.1.2 Seurat

Seurat (HAO et al., 2021) está em sua quarta versão e é uma das mais utilizadas nesse tipo de análise (DING et al., 2021). Sua funcionalidade é baseada em âncoras (que podem ser visualizadas na Figura 4), criadas por um algoritmo de correlação canónica, com a possibilidade de se integrar conjuntos de dados que servem como modelo, bem como o mapeamento de um conjunto de dados com tipos celulares conhecidos, que então terá suas células classificadas.

Figura 4 – Ancoras do Seurat



O classificador do Seurat se baseia em âncoras que são projetadas no conjunto de dados questão. O agrupamento de células em preto no conjunto de dados foi criado erroneamente, sendo um artefato do algoritmo.

Fonte: STUART et al., 2019.

Além disso, essa ferramenta permite a integração de conjuntos de dados de naturezas diferentes – processo chamado de integração multimodal. Por exemplo, podemos integrar informações de ATAC-Seq (mencionado no item 1.2) com dados de expressão de RNA de células únicas. No entanto, para essa *pipeline*, utilizaremos apenas sua funcionalidade de mapeamento de células únicas.

2.1.1.3 Symphony

A ferramenta Symphony (KANG et al., 2021) é uma ferramenta de construção de atlas celulares. Assim como as demais, o *software* cria um mapa de redução de dimensionalidade (como por exemplo um UMAP) a partir de um conjunto de dados fornecido. Esse mapa é utilizado para classificar células com tipos celulares desconhecidos, de forma rápida, com o método estatístico *k Nearest Neighbours* (ou KNN) (FIX; HODGES, 1952). Ele faz o mapeamento de novas células sem integrá-las ao mapa, dando agilidade à análise, o que se mostra o principal diferencial do *Symphony.*

2.1.1.4 FUSCA

O FUSCA: *Framework for Unified Single Cell Analysis*, traduzido aproximadamente para estrutura para análise unificada de células únicas, é uma ferramenta desenvolvida pelo Laboratório de Biologia de Sistemas da UFSC, grupo do qual faço parte. Esta ferramenta ainda está em desenvolvimento e aguarda publicação. Portanto, no momento, seu código está somente disponível no GitHub (https://github.com/edroaldo/fusca/). Sua funcionalidade tem similaridades com outra ferramenta desenvolvida pelo grupo, Cellrouter (LUMMERTZ DA ROCHA et al., 2018), gerando informações de comunicação celular e trajetória em pseudotempo. Mesmo que esse não seja o escopo desta *pipeline*, a ferramenta apresenta funcionalidades de organização dos dados que, se integradas com outras ferramentas de classificação, a tornam útil e relevante no contexto deste trabalho.

2.1.2 Score de importância de genes na classificação celular

Um dos objetivos deste trabalho é a criação de um *score* de importância para cada característica utilizada na classificação dos tipos celulares pelo *singleCellNet*. Em outras palavras, uma métrica que busca auxiliar na resolução de um problema das

ferramentas de classificação - a impossibilidade de se saber quais genes foram responsáveis pela classificação de um determinado tipo celular.

A construção do *score* começou ao se criar uma tabela T, onde as colunas são os tipos celulares obtidos e as linhas são os genes do conjunto de dados questão. O valor de expressão dos genes para cada um dos tipos celulares foi calculado ao se obter a média do valor de *counts* de todas as amostras classificadas naquele tipo celular no conjunto de dados questão. Em seguida, escala-se os dados em uma distribuição normal centrada, onde para cada gene, o valor de expressão médio se torna 0 e os valores de expressão são escore-z, obtidos a partir do desvio padrão da distribuição (LARSON; FARBER; RUNGER, 2004). O mesmo foi feito para os dados de treino, obtendo-se duas tabelas de igual dimensão. A tabela T pode ser visualizada como um mapa de calor, que pode ser encontrado na Figura 18, seção 2.2.2.1.

A partir delas, obtemos o nosso *score* S. A equação a seguir é utilizada para gerar o *score* para uma determinada amostra*:*

$$S = log_2(1 + importance(s\_exp_1 + s\_exp_2))$$

Onde S é o valor do *score;* $importance$ representa o coeficiente *mean decrease GINI; e* s_exp_1 e s_exp_2 são os valores de expressão escalados para cada gene do par de genes. O código da equação e da geração das tabelas está disponível no repositório da *pipeline*.

Utiliza-se *mean decrease GINI* pois essa métrica indica a importância de cada uma das variáveis (ou par de genes) para o modelo de *Random Forest*. Note que a importância para o modelo pode diferir da importância para uma classificação específica que aquele modelo faz. Portanto, faz-se necessária o ajuste pela proporção de expressão dos genes. Como o algoritmo utiliza pares de genes como características para o classificador, fazemos a soma do produto da expressão desses dois genes pelo valor de importância de Gini. Ao final, soma-se um e tira-se o logaritmo para facilitar na visualização em um mapa de calor.

2.1.3 Snakemake

Uma das principais ferramentas de gerenciamento de fluxo de trabalho, o Snakemake (KÖSTER; RAHMANN, 2012) é a linguagem de desenvolvimento da *pipeline*. Com uma integração com *scripts* em R, a ferramenta permite o fácil

desenvolvimento de uma *pipeline* de análise unificada. Além disso, seus *scripts* são facilmente editáveis e sua estrutura permite a criação de um container Singularity que contém todas as ferramentas e pacotes de R necessários para o funcionamento das análises. Ainda, o Snakemake permite o usuário escolher quais ferramentas disponíveis utilizar, bem como a adição de novos *softwares*. Por fim, a adição de novos conjuntos de dados para análise é fácil e requer apenas a formatação dos dados em um padrão pré-estabelecido.

2.1.4 Singularity

Singularity (SYLABS, 2018), desenvolvido pela Sylabs, é uma ferramenta de criação de containers para desenvolvimento de *pipelines* de analises de ciência de dados. Focado especialmente em analises biomédicas, seu foco é colocar toda e qualquer dependência que determinada *pipeline* pode exigir em um arquivo que serve como estrutura onde aquela *pipeline* irá ser executada. Isso resolve diversos problemas de versionamento de dependências, exigência de sistemas operacionais e generalização de especificações próprias de um único usuário. Dessa forma, o Singularity permite rodar *pipelines* em *clusters* de computação de alto desempenho, sendo compatível com outras ferramentas de alocação e otimização de recursos computacionais.

2.1.5 Dados de La Manno et al

Para testar a *pipeline*, utilizamos dados de disponibilidade pública. Escolhemos a publicação de LA MANNO et al., 2016, que contém dados de sequenciamento de transcriptômica de células únicas de células humanas e de células derivadas de células pluripotentes induzidas. As células humanas foram coletadas em embriões, na região ventral do mesencéfalo e servirão como nossos dados de referência; isto é, aqueles dos quais temos o conhecimento do tipo celular. Sua tipagem celular foi feita de acordo com a expressão de determinados genes, como demonstrado no artigo original já citado.

Os tipos celulares encontrados, bem como seu código de referência utilizado nas figuras, podem ser visualizados no Quadro 1. Os códigos de referência utilizados foram adaptados a partir dos *scripts* do artigo.

Quadro 1 – Códigos e tipos celulares do conjunto de dados de referência

Código de referência	Tipo celular
DA	Neurônio dopaminérgico
Endo	Célula endotelial
Gaba	Neurônio gabaérgico
Mgl	Célula da microglia
NbM	Neuroblasto medial
NbML1	Neuroblasto mediolateral
NProg	Progenitor neuronal
OMTN	Núcleo oculomotor e troclear
OPC	Percursor de oligodendrócito
Peric	Pericitos
ProgBP	Progenitor de placa basal
ProgFP	Progenitor de placa neural
RN	Núcleos vermelhos
Sert	Neurônios serotonérgicos
Rgl	Célula glial radial
rand	Célula randômica (gerada *in silico*)
Unk	Célula desconhecida

Os tipos celulares foram encontrados de acordo com a morfologia e expressão de genes-chaves definidos pelos autores.

Fonte: Elaborado a partir de La Manno et al. (2016).

As células que servirão como nosso conjunto de dados em questão são neurônios dopaminérgicos (responsáveis pela síntese da dopamina) derivados de células pluripotentes induzidas. O contexto do estudo diz respeito ao grande potencial dos neurônios dopaminérgicos no tratamento de Parkinson (BARKER; DROUIN-OUELLET; PARMAR, 2015). A diferenciação foi feita a partir do protocolo proposto por KRIKS et al., 2011. Diversas avaliações foram feitas durante todo o processo de diferenciação e os autores tratam o procedimento de diferenciação como bem sucedido. A estas também foi atribuída uma classificação baseada na expressão de alguns genes-chave. Vale deixar claro que o experimento também utilizou células-tronco embrionárias, mas estas não foram utilizadas neste trabalho.

Ao final do processo de diferenciação, algumas células não atingiram o alvo de diferenciação. La Manno et al. (2011) relata:

> “As células-tronco pluripotentes induzidas geraram dois tipos de neuroblastos, dois tipos celulares semelhantes a neurônios motores, um neurônio com um núcleo vermelho bem definido, e três neurônios dopaminérgicos com características de neurônios dopaminérgicos fetais. O tipo de neurônio dopaminérgico mais maduro (iDAc) expressou genes-chave como NR4A2, KLHL1, PBX1, SLC18A2, TH, DDC, GFRA1 ou EN1. Nós concluímos que estas preparações contém uma diversidade celular muito maior do que conhecido previamente, incluindo células TH+ em um estágio

> de diferenciação similar ao tecido usado atualmente para transplantes celulares em doença de Parkinson.” (LA MANNO et al. 2011, tradução nossa)

Os tipos celulares identificados após a diferenciação de células-tronco pluripotentes induzidas em direção à linhagem neuronal não são os mesmos do conjunto de dados de embriões humanos (utilizado na construção dos classificadores). Portanto, essa tipagem nos serve como um norte, mas não como base de uma métrica para medir a acurácia da *pipeline*. O código de acesso dos dados, no banco de dados NCBI GEO (BARRETT et al., 2012) é GSE76381. O *script* que foi utilizado para a obtenção dos dados está presente na *pipeline*. Os dados obtidos são matrizes de expressão e conjunto de metadados. Neles, encontramos 337 amostras (com cada amostra sendo uma única célula), com valores de expressão para 14.726 genes. O Quadro 2 demonstra um exemplo de matriz de expressão, enquanto o Quadro 3 demonstra um exemplo de metadados. Esses quadros servem apenas como ilustração da formatação dos dados; as tabelas utilizadas no trabalho podem ser facilmente acessadas no repositório da *pipeline*. Os dados foram manuseados no R, criando-se matrizes esparsas para os dados de expressão (em *counts*) e *data frames* para os metadados.

Quadro 2 - Exemplo de conjunto de dados de expressão

	Cell01	Cell02	Cell03
NCOAD3	0	0	1
SURF2	0	3	0
LINC94	0	0	0
PREX11	3	0	4
ARFGEF2	0	0	0
CSE1L-AS1	0	0	0
CSE1L	0	2	0
STAU1	0	0	1
DDX27	1	0	0
ZNFX1	0	0	0
ZFAST1	4	4	1
SNORD12C	0	0	0
SNORD12B	0	1	0

Fonte: Elaborado pelo autor (2022).

Quadro 3 – Exemplo de metadados

	Cell_ID	Cell_type
Cell01	Cell01	Rgl2a
Cell02	Cell02	hOMTN
Cell03	Cell03	hNbM
Cell04	Cell04	Rgl2a
Cell05	Cell05	hNProg
Cell06	Cell06	hNProg
Cell07	Cell07	hNbM
Cell08	Cell08	Rgl2a
Cell09	Cell09	Rgl2a
Cell10	Cell10	Unk
Cell11	Cell11	hProgBP
Cell12	Cell12	Unk
Cell13	Cell13	hProgBP

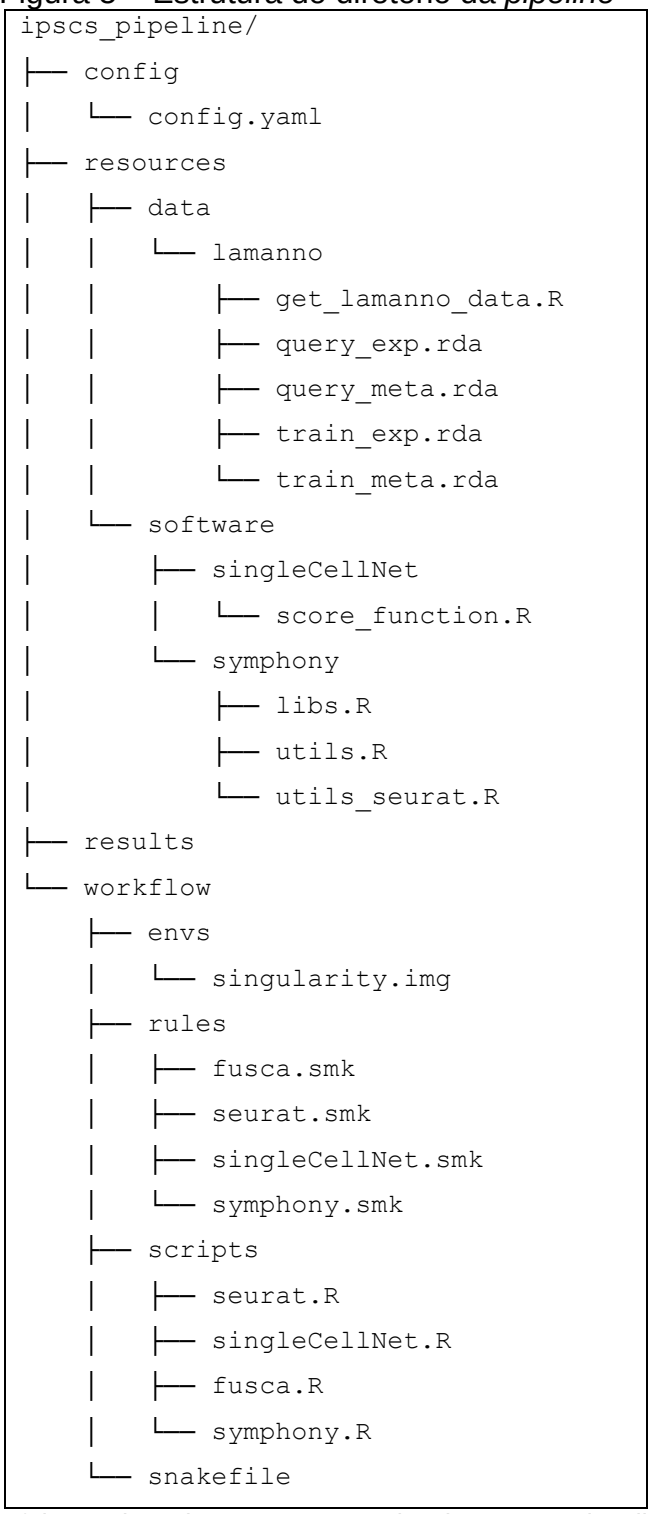
Fonte: Elaborado pelo autor (2022).

2.2 RESULTADOS E DISCUSSÕES

2.2.1 Pipeline

A *pipeline* construída em Snakemake tem uma estrutura recomendada por padrão (KÖSTER; RAHMANN, 2012). A estrutura da *pipeline* deste trabalho foi construída seguindo tal recomendação e pode ser visualizada na Figura 5.

Figura 5 – Estrutura do diretório da *pipeline*



A estrutura foi gerada pelo pacote `tree` do Linux. Sua visualização também é possível no repositório do trabalho.

Fonte: Elaborado pelo autor (2022).

O primeiro diretório é o `config`, com um arquivo chamado `config.yaml`. Os arquivos no formato YAML (https://yaml.org/) são de fácil leitura e servem para

organização de dados. Nesse caso, é aqui onde indicamos os parâmetros necessários para o uso da *pipeline*, como os dados que foram utilizados, o nome de algumas colunas específicas, dentre outros.

O segundo diretório é o `resources`. Nele, encontramos uma pasta `data` que terá uma subpasta para cada conjunto de dados que será analisado. Para o bom funcionamento da *pipeline*, suas estruturas devem seguir o padrão do exemplo de dados fornecido.

Outro diretório dentro de `resources` é o `software`. Aqui colocamos quaisquer programas, pacotes ou *scripts* que sejam dependência das ferramentas utilizadas pela *pipeline*. Nesse caso, as ferramentas singleCellNet, Symphony e FUSCA necessitam de *softwares* adicionais.

O próximo diretório é o `results`. Aqui serão inseridas todos as figuras geradas pela *pipeline*, organizadas em pastas de conjunto de dados e ferramenta, nesta ordem. A Figura 6 demonstra um exemplo de resultados gerados por esta *pipeline*, que serão explanados adiante no texto.

Figura 6 – Estrutura do diretório *results*

```
results/
└── lamanno
    ├── seurat
    │   ├── refAndQuery_umap.png
    │   └── transfdata_umap.png
    ├── singleCellNet
    │   ├── assessment-curves.png
    │   ├── assessment_metrics.png
    │   ├── atrr-plot.png
    │   ├── average-top-genepair.png
    │   ├── query_classification-heatmap.png
    │   ├── query_umap.png
    │   ├── query-violin-plot.png
    │   ├── score_heatmap.png
    │   └── test-data_heatmap.png
    ├── symphony
    │   ├── reference_umap.png
    │   ├── umap_mixed_queryAndReference.png
    │   └── umap_query.png
    └── fusca
        └── umap_queryAndReference.png
```

O repositório com as figuras geradas é o principal produto da pipeline. Com mais conjuntos de dados, outros diretórios de resultados seriam gerados.

Fonte: Elaborado pelo autor (2022).

O último diretório é o `workflow`. Nele estão inseridos os arquivos centrais da *pipeline*. O arquivo que inicializa a *pipeline* é o `snakefile`. Ele é escrito na linguagem

Python (VAN ROSSUM; DRAKE, 2009) e é composto por uma série de *rules* (ou regras, na tradução literal) que evocam *outputs* – as figuras que desejamos gerar para nossos dados. O Snakemake é inteligente em sua reprodução: apenas os *outputs* evocados terão seus *scripts* executados e se determinado *output* já estiver presente no diretório *results*, seu respectivo *script* não será executado. Isso é importante para economia de uso da capacidade computacional. As *rules* podem ser encontradas no diretório `rules`. A Figura 7 ilustra uma *rule* da *pipeline*.

Figura 7 – *Rule* `seurat.smk`

```
rule seurat:

  input:
    metaQuery = "resources/data/{dataset}/query_meta.rda",
    expQuery = "resources/data/{dataset}/query_exp.rda",
    metaTrain = "resources/data/{dataset}/train_meta.rda",
    expTrain = "resources/data/{dataset}/train_exp.rda"

  output:
    transfdata_umap =
      "results/{dataset}/seurat/transfdata_umap.png",
    refAndQuery_umap =
      "results/{dataset}/seurat/refAndQuery_umap.png"

  params:
    Cell_type_colname = config["Cell_type_colname"],
    Cell_ID_colname = config["Cell_ID_colname"]

  script:
    "scripts/seurat.R"
```

Este código serve de exemplo de *rule* do Snakemake.

Fonte: Elaborado pelo autor (2022).

Além disso, temos o diretório `envs`, onde são inseridos os arquivos containers que são utilizados para as análises. Nesse caso, o único arquivo é o `singularity.img`, imagem de um container Singularity que contém todas as dependências da *pipeline*.

Por fim, a pasta `scripts` contém todos os *scripts* requisitados pelas *rules*. Há um para cada ferramenta.

A *pipeline* está disponível no GitHub, pelo link de acesso https://github.com/gacrestani/ipsc-pipeline. Sua instrução de download, instalação e uso estão disponíveis em seu arquivo `README.md`. Sua execução é feita pelo terminal,

navegando-se até o diretório da *pipeline* e executando o seguinte comando demonstrado no Quadro 4.

Quadro 4 – Uso da *pipeline* no terminal

```
snakemake -c4 -use--singularity
```

Fonte: Elaborado pelo autor (2022).

O comando chama o *software* Snakemake, que está instalado no sistema; o parâmetro `-c4` passa ao programa o número de núcleos de processamento a serem utilizados. Este valor varia de acordo com a capacidade computacional de onde se utiliza a ferramenta. O parâmetro `-use--singularity` indica ao programa que ele deve utilizar a imagem Singularity disponível (previamente configurada no arquivo `config.yaml`).

Com o crescente valor de mercado da terapia celular (GRAND VIEW RESEARCH, 2021), impulsionado pela grande quantidade de dados gerados, *pipelines* desta natureza se mostram ferramentas de trabalho fundamentais de empresas do ramo. Seu desenvolvimento é, muitas vezes, acoplado à grandes experimentos, estruturas laboratoriais, núcleos computacionais, e fluxo de dados (LUO et al., 2016). A presente *pipeline* foi construída de forma eficiente, adaptável e escalonável, utilizando-se de ferramentas de ponta, tornando-a apta a ser utilizada de forma comercial e/ou servir como base para a construção de novas ferramentas.

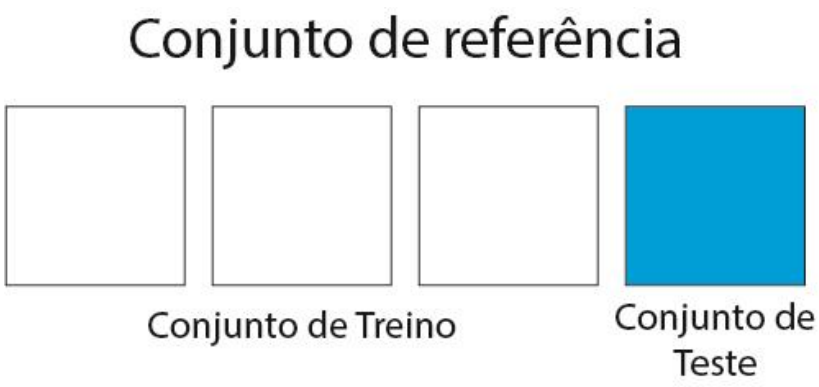
2.2.2 Ferramentas

Aqui serão listadas as figuras geradas ao se aplicar as ferramentas no conjunto de dados de La Manno et al (2016), utilizando-se a *pipeline*.

2.2.2.1 singleCellNet

O primeiro passo na execução do singleCellNet é a criação de um classificador baseado no conjunto de dados de referência. Esse classificador é criado a partir de um subconjunto do conjunto de dados de referência, chamado de subconjunto treino. O restante do conjunto de dados de referência é alocado para um subconjunto chamado teste. A divisão é ilustrada na Figura 8. Então, classificamos o conjunto teste e verificamos sua acurácia.

Figura 8 – Conjunto de referência, treino e teste

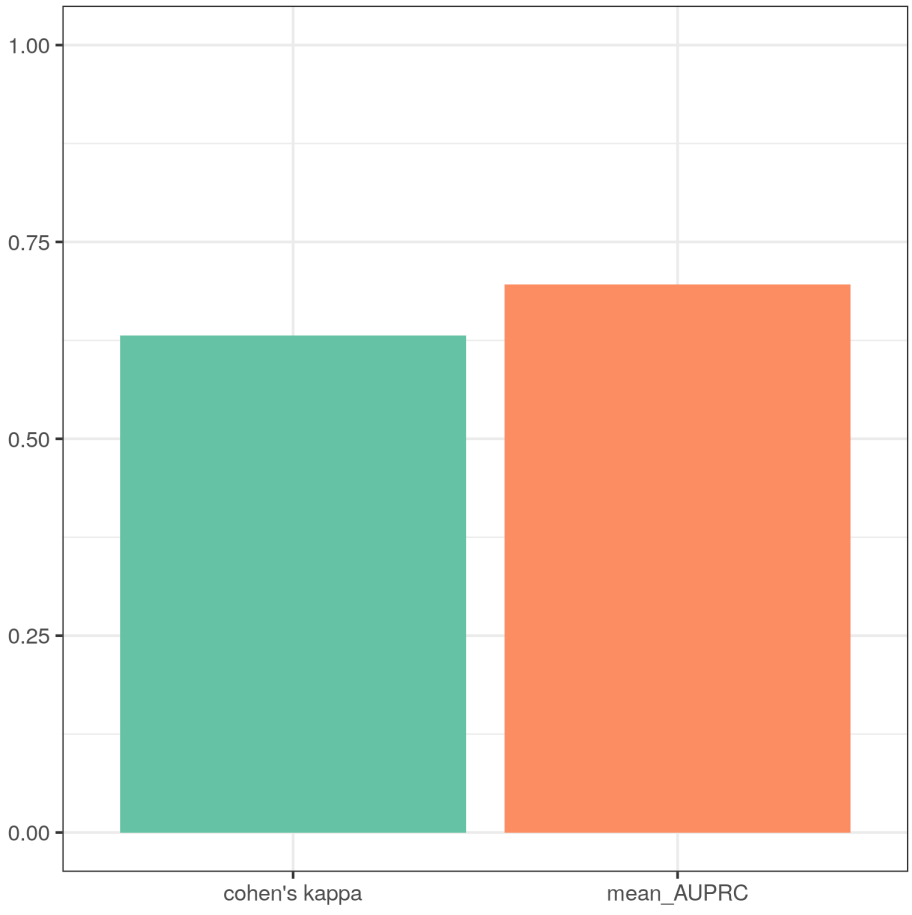


O conjunto de referência se divide em 4 partes. Destas, três são utilizadas pelo classificador como conjunto de treino e uma é utilizada como conjunto de teste.

Fonte: Elaborado pelo autor (2022).

As métricas de avaliação geram as figuras `refAndQuery_umap.png` (Figura 8) e `transfdata_umap.png` (Figura 9).

Figura 9 – Métricas de avaliação do classificador de singleCellNet



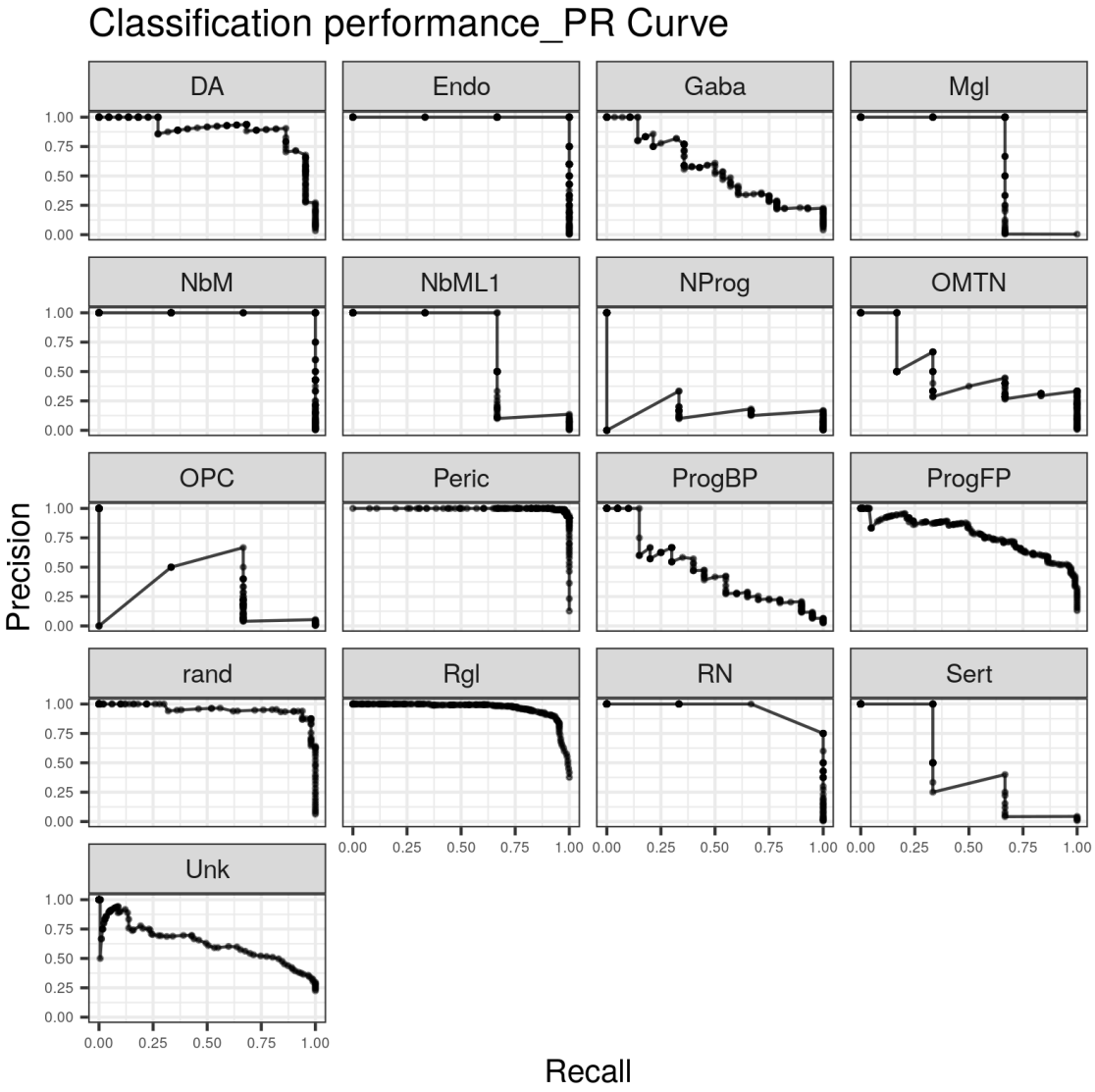
Fonte: Elaborado pelo autor (2022).

A Figura 9 representa duas métricas de avaliação do classificador – o coeficiente Kappa de Cohen (COHEN, 1960), mede a concordância de classificação entre o classificador recém criado e as informações das quais já se tinha conhecimento. Um classificador perfeito teria coeficiente 1,00. O classificador em questão resultou em um coeficiente de 0,67, sendo considerado moderado, segundo MCHUGH, 2012.

A outra métrica é o AUPRC (*Area under precision recall curve*, ou área abaixo da curva de precisão/revocação) (BOYD; ENG; PAGE, 2013) mede basicamente a mesma informação do coeficiente de Kappa, mas com uma estatística diferente – seu cálculo é feito baseado em uma curva de precisão/revocação, que mede quão corretos (precisão) e completos (revocação) são os resultados da classificação. Nosso valor de AUPRC foi de 0,71.

É importante deixar claro que o processo de criação do classificador não é determinístico – em outras palavras, há uma variação nos algoritmos que produz resultados diferentes a cada rodada. Esses valores de coeficiente Kappa de Cohen e AUPRC são exemplos de informações que variam. Além disso, para que se obtenham coeficientes maiores, nosso conjunto de dados deveria ser maior, com mais exemplares de cada célula. No entanto, há um limite para quantos dados devemos utilizar ao alimentar um classificador de aprendizado de máquina. Se utilizarmos muitos dados, corremos o risco de sofrer de um fenômeno chamado *over fitting* (ou sobreajuste) (SALMAN; LIU, 2019).

Figura 10 – Curvas de avaliação do classificador de singleCellNet



As curvas de precisão e revocação de cada tipo celular mostra quão completos eram os dados utilizados como referência pelo classificador.

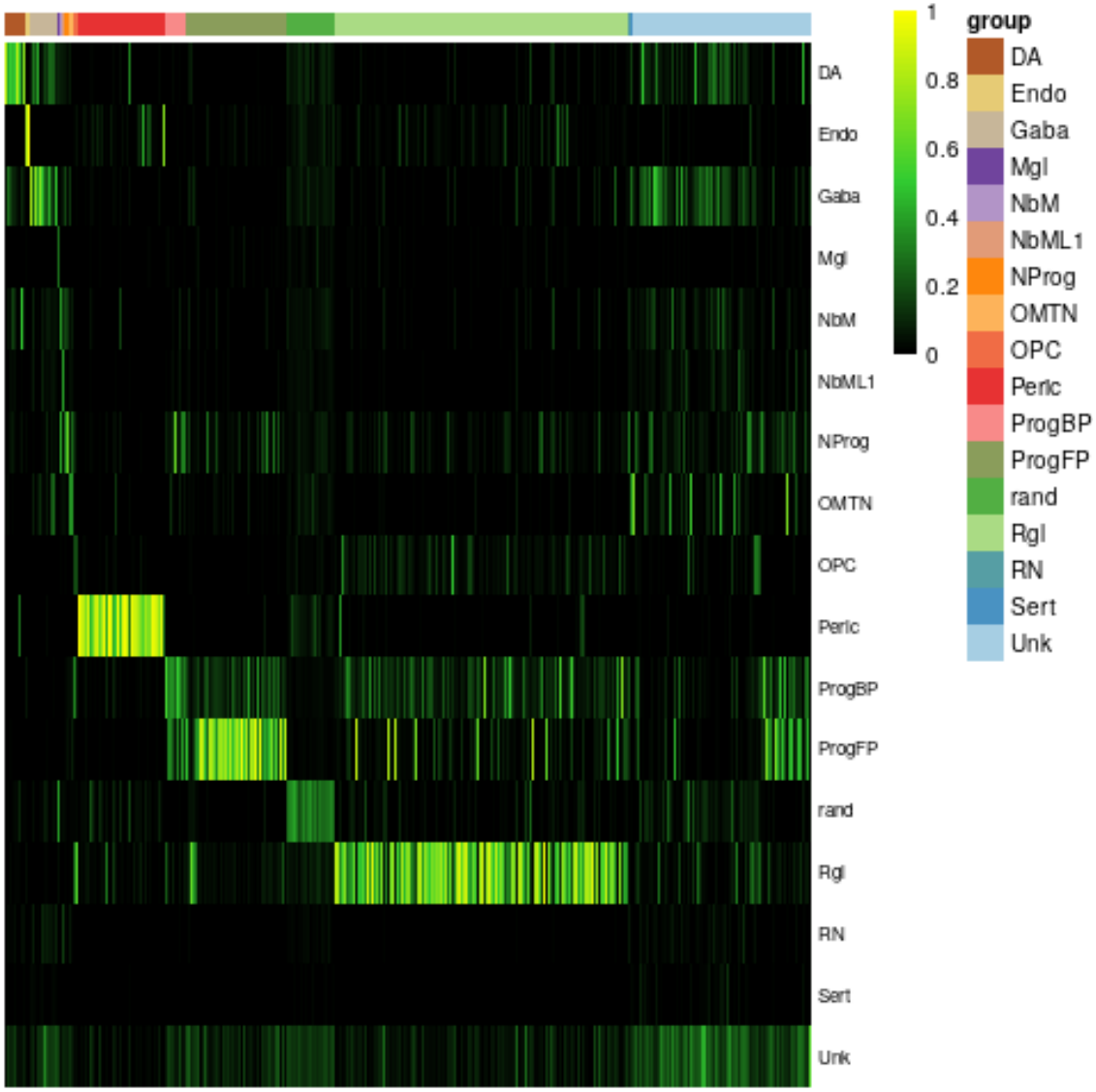
Fonte: Elaborado pelo autor (2022).

A Figura 10 demonstra a curva de precisão/revocação para cada um dos tipos celulares usados na classificação. Alguns tipos celulares, como o *Endo* e *Rgl* obtiveram uma curva melhor do que outros, como *OPC* e *ProgBP*. Em suma, isso significa que a classificação daqueles tipos celulares será mais precisa e completa do que a destes.

Mais adiante, o *script* gera um mapa de calor da classificação das células, de nome `test-data_heatmap.png`, demonstrado na Figura 11. Nesse mapa, cada valor é a probabilidade de determinada amostra pertencer à determinado tipo celular, enquanto a barra de classificação acima do mapa determina o tipo de classificação

previamente conhecido. O tipo celular `rand` é gerado pelo *script* com amostras de valores de expressão randômicos e serve como um grupo controle (TAN; CAHAN, 2019).

Figura 11 – Mapa de calor de conjunto de dados teste do singleCellNet



A barra colorida em cima indica os grupos previamente conhecidos do conjunto de dados de teste. As linhas são as classificações recentes feitas pelo classificador recém treinado, com a finalidade de verificar sua acurácia.

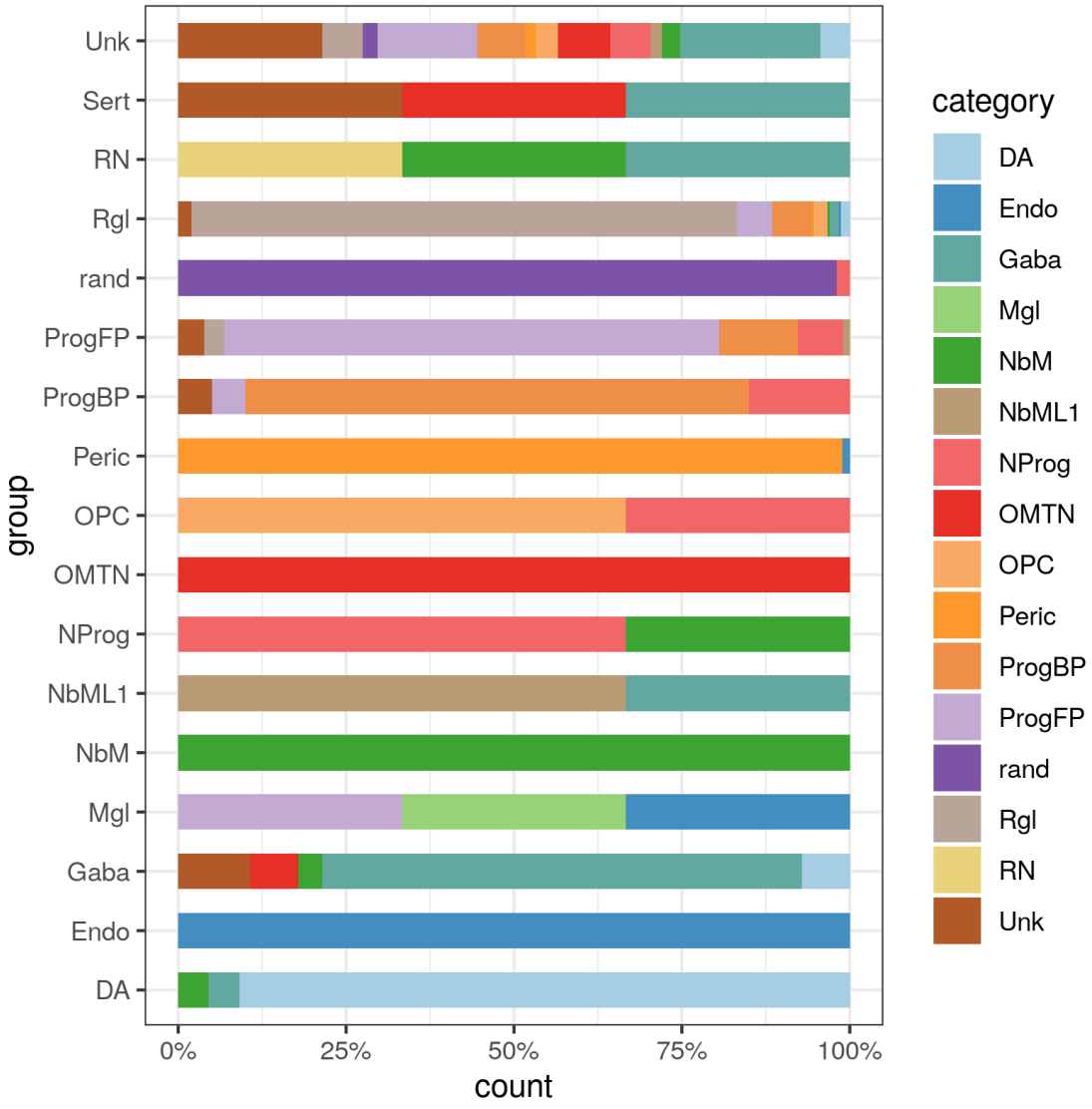
Fonte: Elaborado pelo autor (2022).

Outro resultado gerado acerca do classificador é um gráfico de atribuição – `atrr-plot.png`. Esse gráfico, mostrado na Figura 12, representa a porcentagem de tipos celulares classificados para cada tipo celular previamente conhecido. Isto é, se olharmos a linha de *group* Rgl, vemos uma grande porcentagem de células

classificada como Rgl. No entanto, algumas células do tipo Rgl foram erroneamente classificadas como ProgFP, ProgBP, OPC, NbM e Unk. Os tipos celulares estão descritos no Quadro 1.

Ainda sobre o classificador, podemos ver a importância dos pares de genes para a classificação de cada tipo celular. Para tal, são considerados os valores de expressão desses genes no conjunto de dados treino e seu valor de importância para o algoritmo de *Random Forest*. O resultado em questão é o `average-top-genepair.png` e podemos vê-lo na Figura 13.

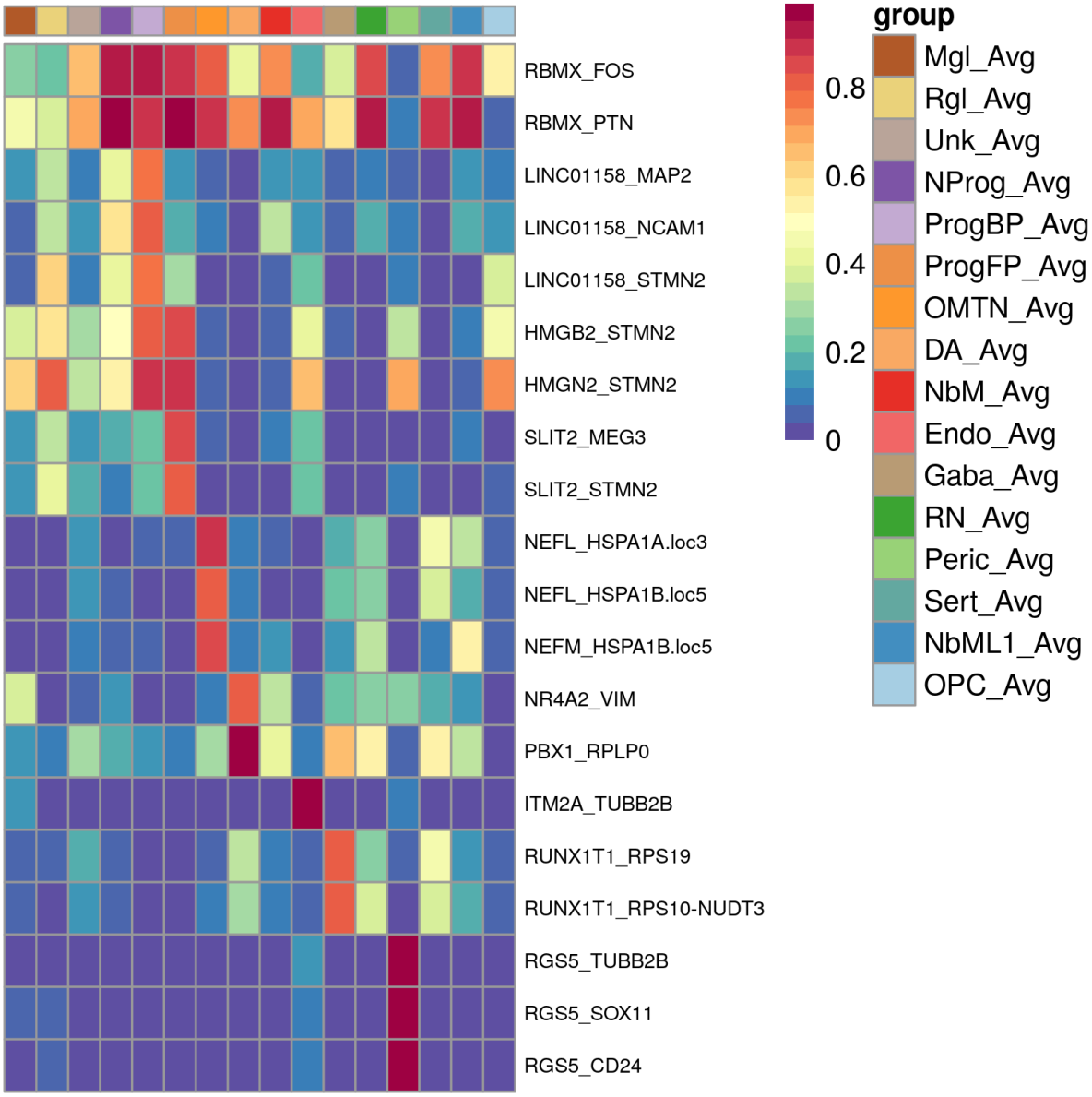
Figura 12 – Gráfico de atribuição do conjunto de dados teste do singleCellNet



A barra colorida indica os grupos previamente conhecidos do conjunto de dados de teste. As linhas são as classificações recentes feitas pelo classificador.

Fonte: Elaborado pelo autor (2022).

Figura 13 – Pares de genes e sua importância para cada tipo cellular do conjunto de dados teste do singleCellNet



Os pares de genes são as características utilizadas pelo classificador do singleCellNet. Cada coluna representa um tipo celular encontrado, e a importância dos pares de genes no momento da criação do classificador. Note que essa importância é diferente da encontrada pelo score criado para este trabalho – a primeira é sobre o classificador, a segunda é sobre as células classificadas.

Fonte: Elaborado pelo autor (2022).

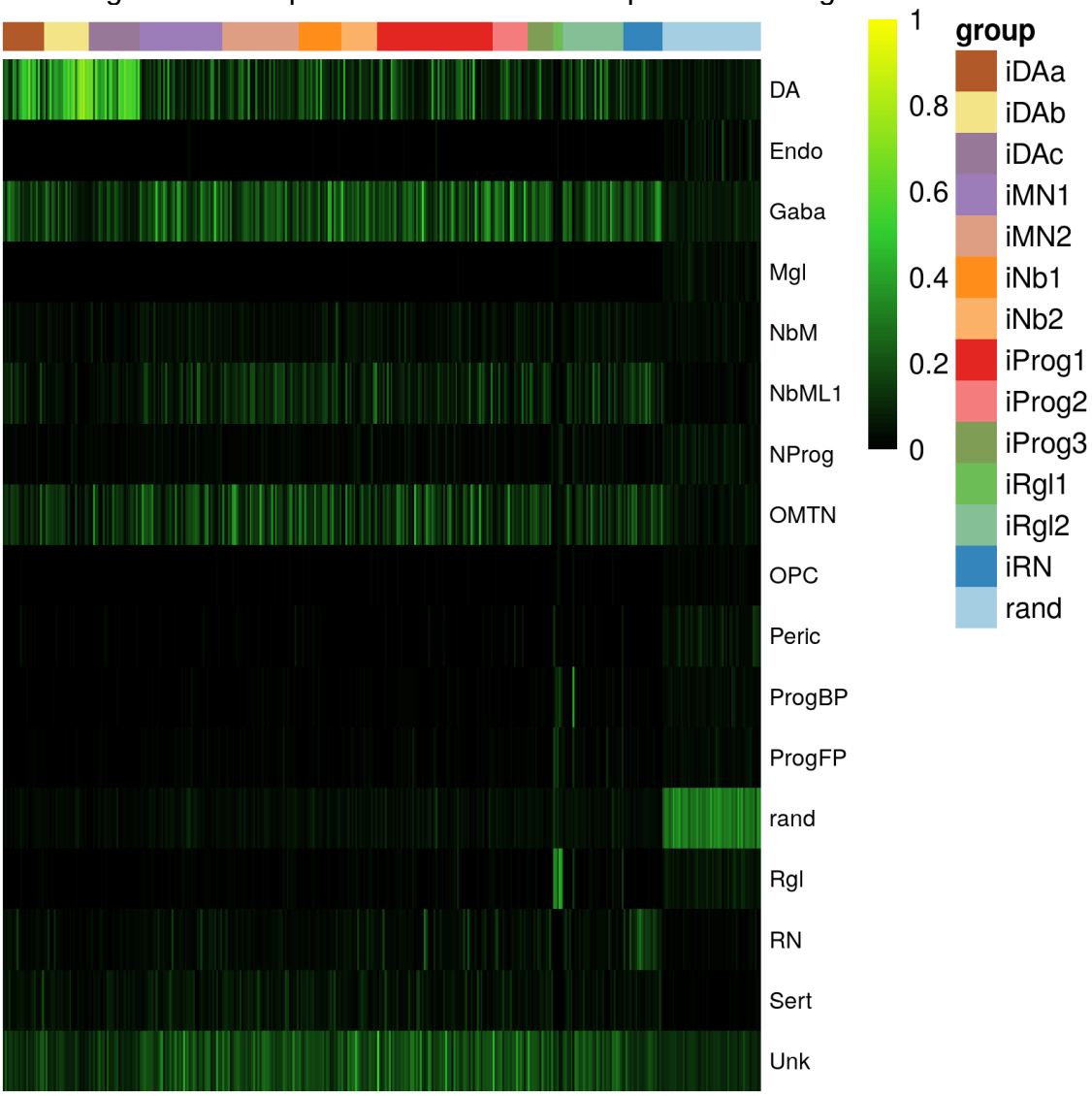
Agora, analisaremos as figuras geradas a partir do conjunto de dados que se está sendo classificado – nesse caso, as células derivadas de células pluripotentes induzidas. A primeira figura gerada é o mapa de calor da classificação, que pode ser

encontrado em `query_classification-heatmap.png` e visualizado na Figura 14.

No caso do conjunto de dados questão utilizado neste trabalho, temos o conhecimento dos tipos celulares e usaremos esta informação para agrupar as amostras nas figuras – a tipagem das células foi feita por La Manno em seu artigo. No entanto, outra informação também poderia ser usada, como a origem da amostra, data da coleta, dentre outras. Cada conjunto de dados terá uma informação que poderá ser utilizada. Caso não haja tal informação, o *script* gerará um mapa de calor utilizando a própria classificação como fator de agrupamento.

Podemos ver que as células foram classificadas, em sua maioria, como DA (neurônios dopaminérgicos), principalmente dentro dos grupos representantes desse tipo de célula. Esse resultado se mostra satisfatório visto que, como descrito na seção 2.1.5, as células passaram por um protocolo de diferenciação para neurônios dopaminérgicos. Como esperado, há outros tipos celulares presentes na amostra, uma heterogeneidade intrínseca à diferenciação direcionada de células-tronco pluripotentes induzidas em células-alvo. Assim como anteriormente, foi criado um tipo celular fictício `rand` com valores aleatórios, servindo como controle. Além disso, esse grupo serve como um artifício para classificar aquelas células do conjunto de dados questão que não tem tipo celular correspondente no conjunto de dados de referência. Ainda, nesse grupo, são classificados novos grupos celulares ou células que completamente falharam em se diferenciar.

Figura 14 – Mapa de calor das células-questão do singleCellNet

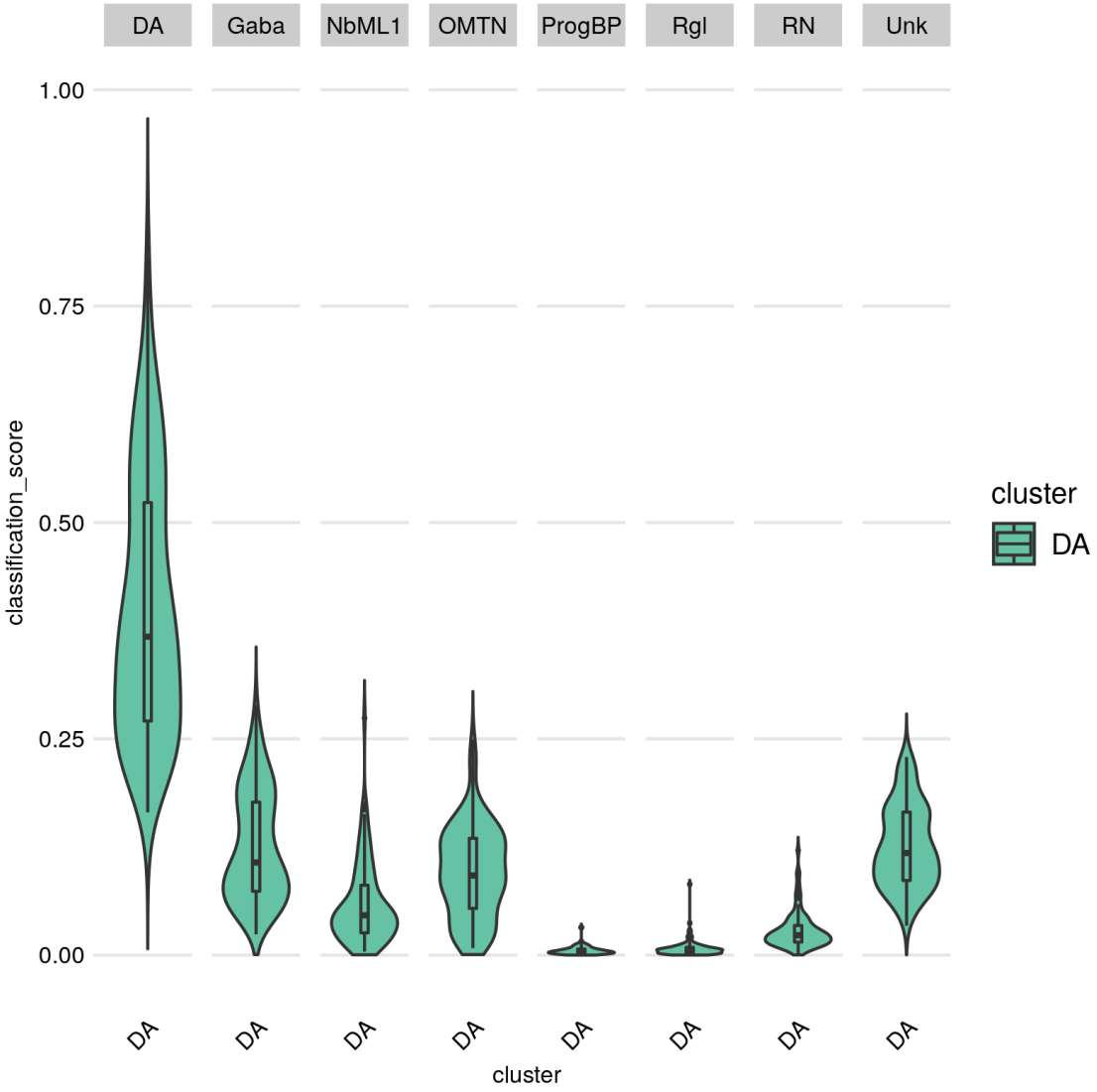


A barra colorida em cima indica os grupos previamente conhecidos do conjunto de dados questão. As linhas são as classificações recentes feitas pelo classificador recém treinado.

Fonte: Elaborado pelo autor (2022).

A Figura 15 representa um gráfico chamado *violin plot*, que pode ser acessado no arquivo `query-subcluster-violin-plot.png`. Ele representa os mesmos quartis que um *box plot*, além de mostrar a densidade dos dados (DATAPLOT, 2003). Podemos visualizar os *scores* de classificação para cada tipo celular. Esse *score* criado pelo singleCellNet diz a probabilidade de determinada amostra pertencer a um determinado tipo celular. Nesse gráfico, estamos observando os dados apenas para as células que foram classificadas como neurônios dopaminérgicos. O mesmo tipo de gráfico pode ser criado considerando-se todos os tipos celulares. No entanto, nele há muita informação, tornando-o de difícil visualização.

Figura 15 – *Violin plot* de neurônios dopaminérgicos do singleCellNet



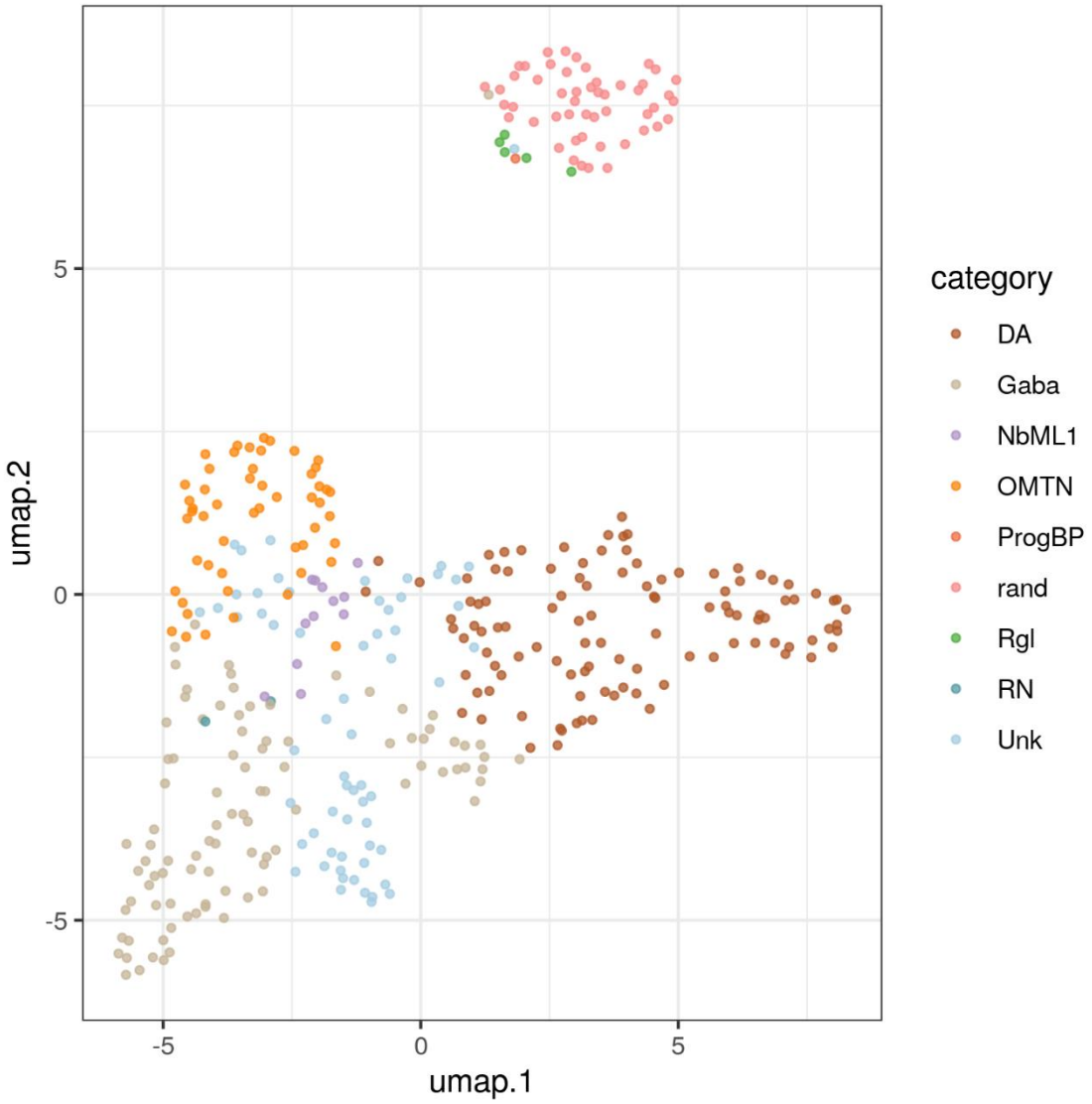
O classificador atribui, a cada amostra e para cada tipo celular, um valor que representa quão provável é que aquela a mostra seja daquele tipo celular. Este gráfico representa a distribuição destes valores para o tipo DA por entre todas as amostras. Cada coluna representa o tipo celular para qual aquela a mostra recebeu a probabilidade mais alta.

Fonte: Elaborado pelo autor (2022).

Mais uma vez, encontramos a maioria dos neurônios sendo classificados como dopaminérgicos. Como dito anteriormente, isto é esperado e foi relatado por LA MANNO et al., 2016. Um eventual erro na classificação também pode se dar pela baixa acurácia do classificador. Isso acontece, geralmente, devido ao uso de um conjunto de dados muito pequeno para servir como treino. No entanto, esse assunto é alvo de livros e artigos inteiros da área de ciência de dados e aprendizado de máquina (BISHOP, 2011).

Na Figura 16, outro gráfico gerado é de redução de dimensionalidade, que transforma o espaço multidimensional de expressão de genes em um gráfico bidimensional de fácil visualização. O método de redução de dimensionalidade é o UMAP (MCINNES; HEALY; MELVILLE, 2020). As amostras se agrupam no espaço do gráfico de acordo com sua expressão gênica e a cor de seus pontos representa o tipo celular recebido por aquela amostra. Em determinados conjuntos de dados, agrupamentos bem definidos são formados. No caso deste trabalho, há um grupo claro de células `rand` e outro grupo espalhado que abriga diferentes tipos celulares em diferentes regiões. A classificação próxima é esperada, visto que o protocolo de diferenciação é para neurônios dopaminérgicos.

Figura 16 – UMAP do conjunto de dados-questão do singleCellNet

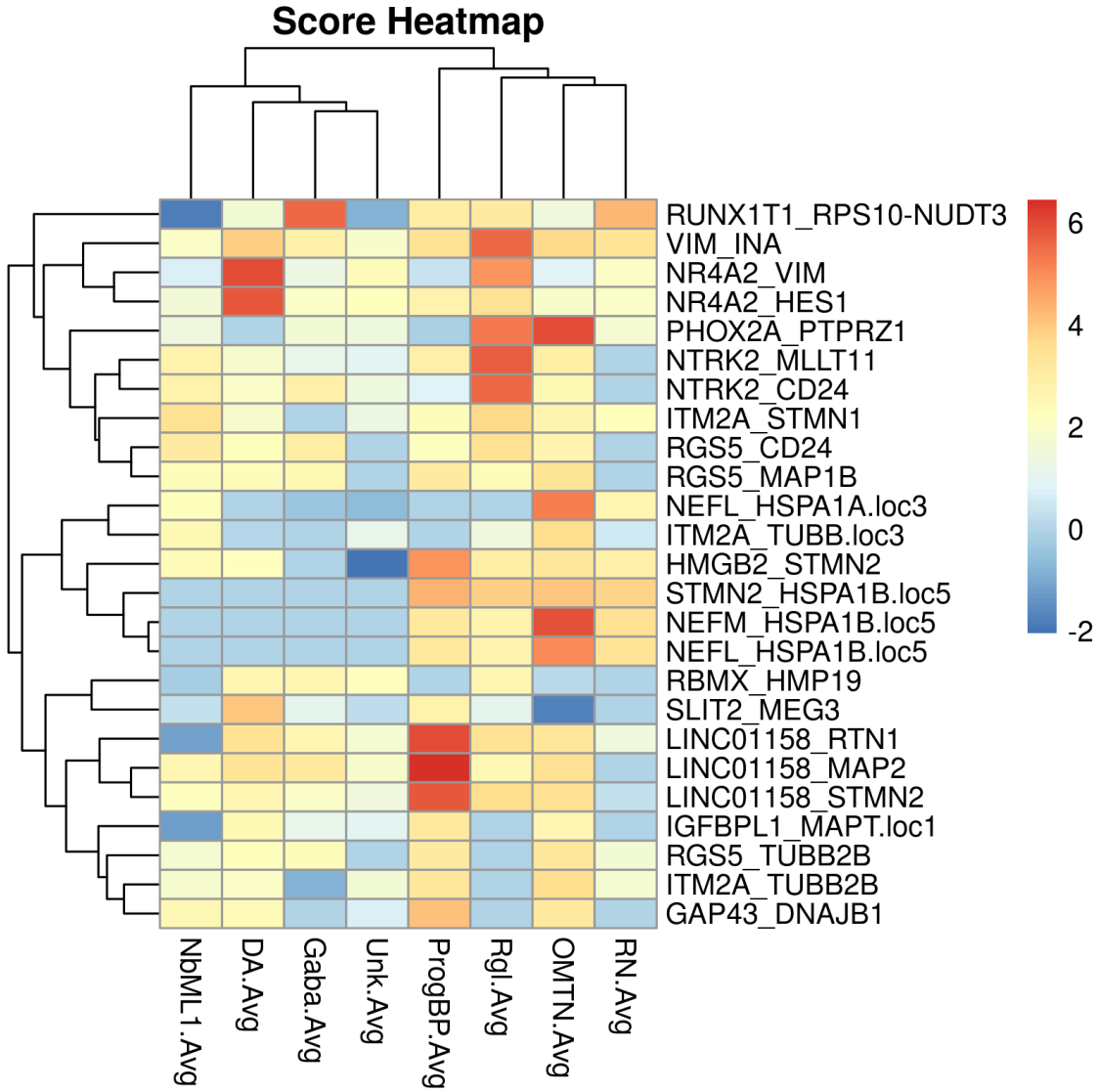


UMAP do conjunto de dados questão. Sua cor representa a classificação atribuída pelo singleCellNet.

Fonte: Elaborado pelo autor (2022).

Por fim, a Figura 17 é o mapa de calor dos *scores* de importância para a classificação dos tipos celulares, como explicado na seção 2.1.2. Seu arquivo é o `score_heatmap.png`.

Figura 17 – *Score* de importância para cada par de gene na classificação celular



O mapa de calor representa a importância que os pares de genes tiveram para a classificação daquele tipo celular.

Fonte: Elaborado pelo autor (2022).

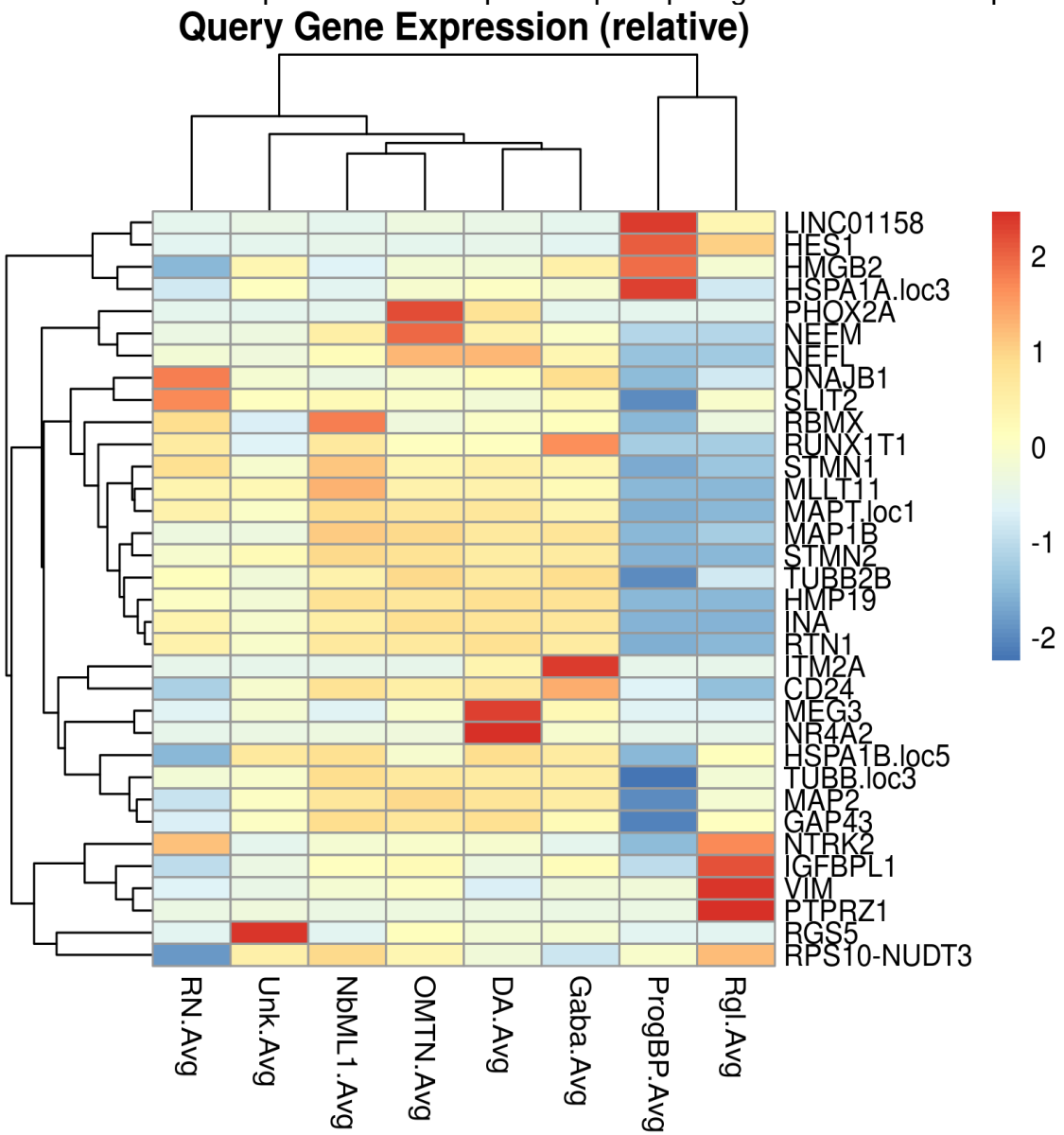
O uso de ferramentas de classificação muitas vezes traz informações cruciais para o andamento da pesquisa. Contudo, tais ferramentas manejam os dados de forma que muitas vezes a propriedade biológica estudada é perdida ou ocultada. Por exemplo, no caso do singleCellNet, o classificador é montado sob pares de genes; saber quais desses pares foram importantes para a classificação de determinado tipo

celular é uma informação que pode auxiliar a extração de conhecimento biológico das ferramentas computacionais. Por esse motivo, a tabela de *score* foi criada, utilizando-se o conjunto de dados questão – a partir dela, pode-se ter uma ideia de quais pares de genes foram importantes para determinado tipo celular naquele conjunto questão, baseando-se na expressão relativa desses genes no conjunto de dados treino, além do *mean decrease GINI* (método já explicado na seção 2.1.2).

Podemos ver que, nesse exemplo, o par de genes `NR4A2_HES1` teve grande importância na classificação dos neurônios dopaminérgicos. Segundo La Manno et al. (2016), *NR4A2* é um gene-chave, cuja expressão é utilizada para se identificar um neurônio como dopaminérgico. Da mesma forma, o gene *HES1* é expresso em neurônios progenitores. Assim, a partir do método de classificação por pares de genes (explicado na seção 2.1.1.1) podemos inferir que esse par de gene foi importante para que o classificador diferenciasse neurônios dopaminérgicos de neurônios progenitores.

O mapa de calor da figura 18 é feito a partir dos genes que compõem os pares de genes com pontuação de *score* mais alta, para cada tipo celular. Neste gráfico, podemos ver que a expressão desses genes é, de fato, mais alta nos respectivos tipos celulares:

Figura 18 – Valor de expressão relativo para os principais genes nas células questão



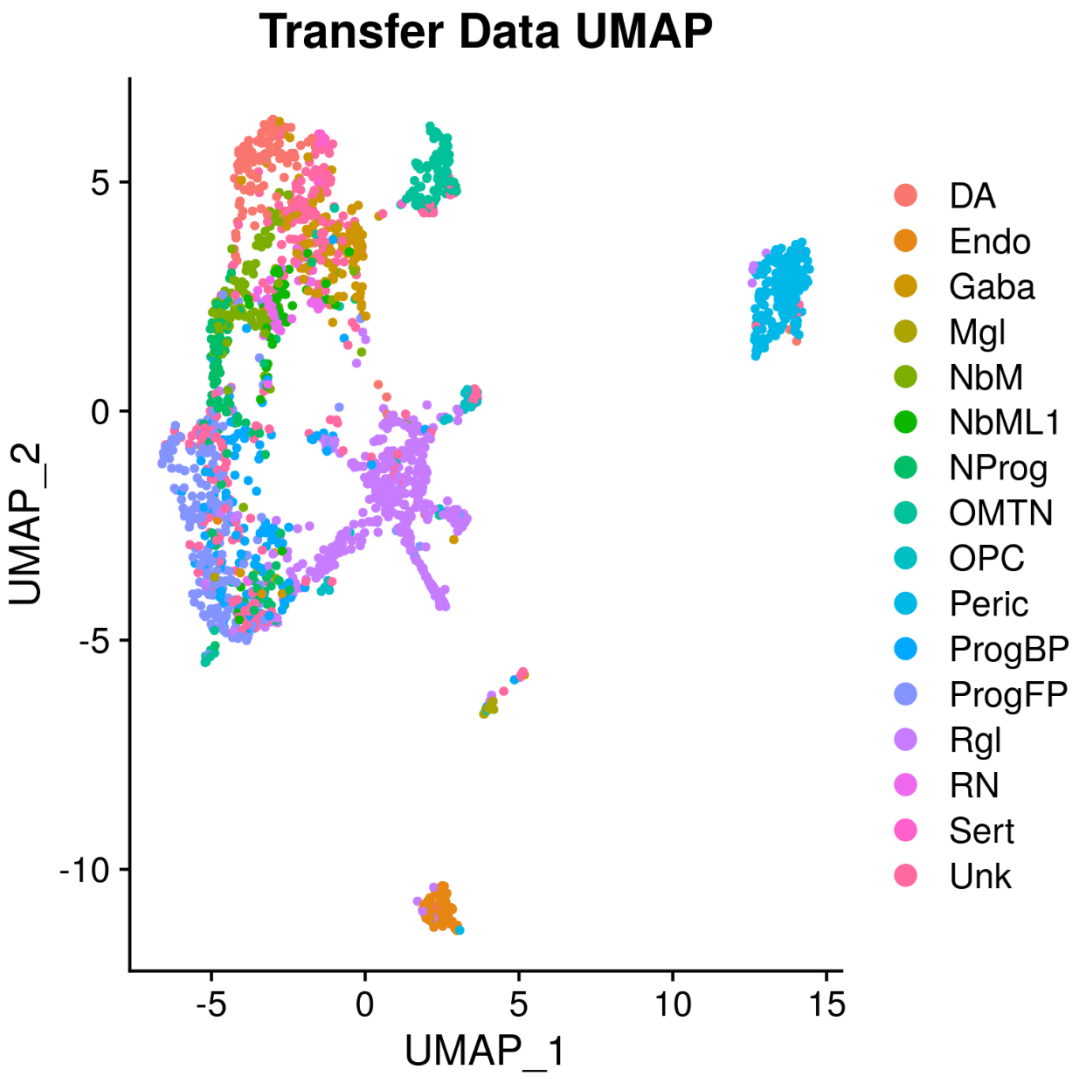
Este mapa de calor representa a expressão dos genes naqueles tipos celulares, seguindo a distribuição em um escore-z.

Fonte: Elaborado pelo autor (2022).

2.2.2.2 Seurat

As figuras geradas pelo Seurat são de redução de dimensionalidade utilizando UMAP, como feito no singleCellNet. A Figura 19 representa os dados utilizados para treinar o classificador, enquanto a Figura 20 representa os dados do conjunto em questão.

Figura 19 – UMAP do conjunto de dados de treino do Seurat

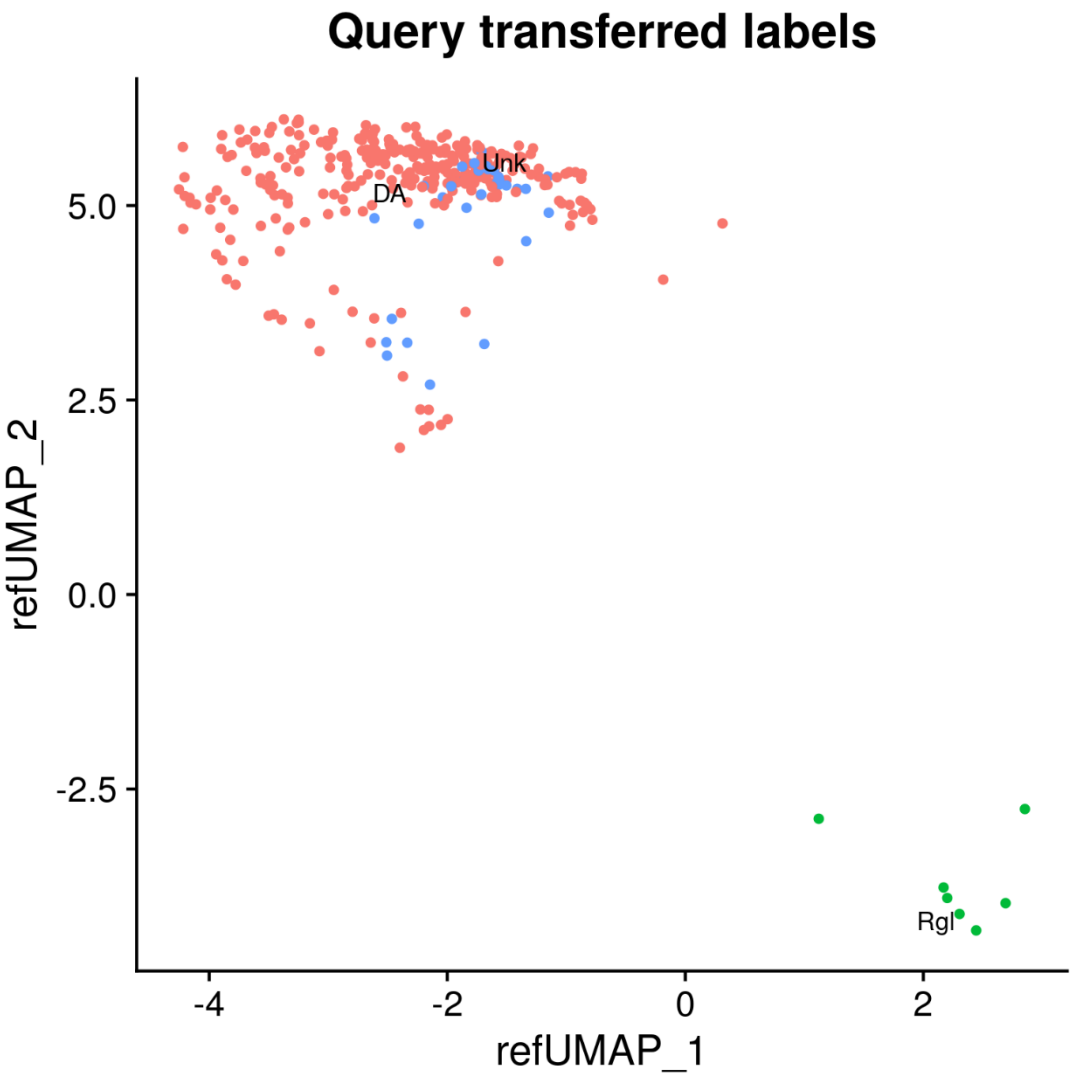


UMAP do conjunto de dados treino utilizado pelo Seurat. As cores representam os tipos celulares previamente conhecidos.

Fonte: Elaborado pelo autor (2022).

Ao se criar e plotar um UMAP, vemos a ordenação das células de treino em alguns agrupamentos isolados, com o maior deles contendo vários tipos celulares. Já as células questão não formaram grupos isolados, mas uma região com neurônios dopaminérgicos e outras células e outra com algumas poucas células do tipo *Rgl*.

Figura 20 – UMAP do conjunto de dados questão do Seurat



UMAP dos dados questão classificados pelo Seurat. Diferente das outras ferramentas, o Seurat encontrou muito menos tipos celulares nos dados questão.

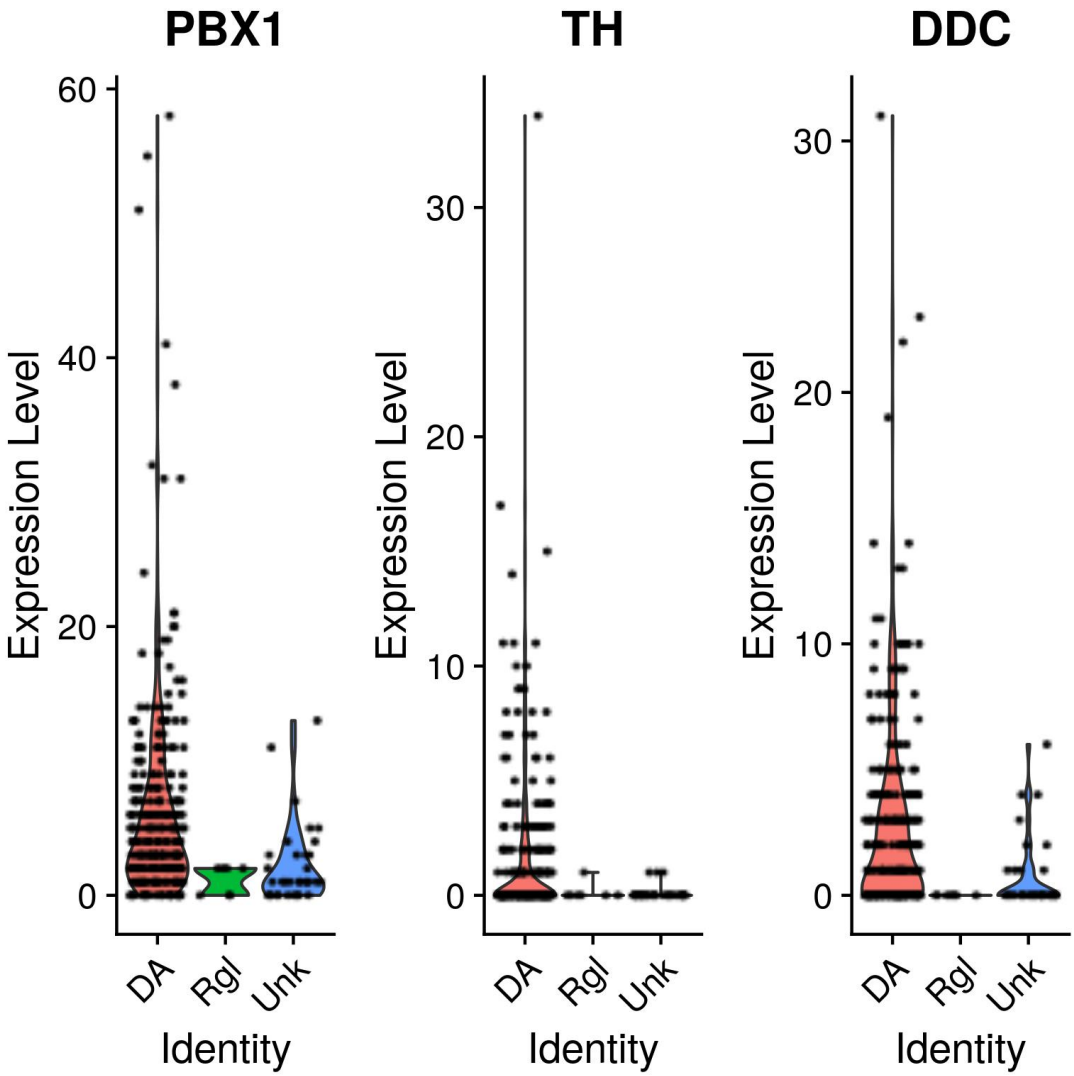
Fonte: Elaborado pelo autor (2022).

Podemos observar, na Figura 20, que houve uma eficácia de reprogramação satisfatória no nosso conjunto de dados questão – 87,8% das células foram reprogramadas em neurônios dopaminérgicos.

O Seurat é uma poderosa ferramenta para se analisar a expressão de alguns genes de interesse do conjunto de dados questão. Esses genes variam de acordo com o experimento e a natureza dos dados. Para tais análises, o usuário deve saber de antemão quais genes de interesse merecem visualização. Esses genes são inseridos no arquivo `config.yaml` como um parâmetro que é passado à *rule* do Seurat. A Figura 21 e Figura 22 representam um *violin plot* e um mapa de calor,

respectivamente. Ambas demonstram a expressão de três genes de interesse para esse conjunto de dados em específico. Os genes foram retirados da lista de genes dada por La Manno et al (2011). Elas são acessadas pelos arquivos `query_violin.png` e `query_heatmap.png`, respectivamente. Podemos ver que a alta expressão desses genes é, de fato, um indicador do tipo celular neurônio dopaminérgico.

Figura 21 – *Violin plot* dos genes de interesse do Seurat



Expressão de três genes de interesse definidos por La Manno et al. (2016). Os níveis de expressão estão agrupados pelos tipos celulares previamente classificados pelo Seurat.

Fonte: Elaborado pelo autor (2022).

Figura 22 – Mapa de calor da expressão dos genes de interesse do Seurat

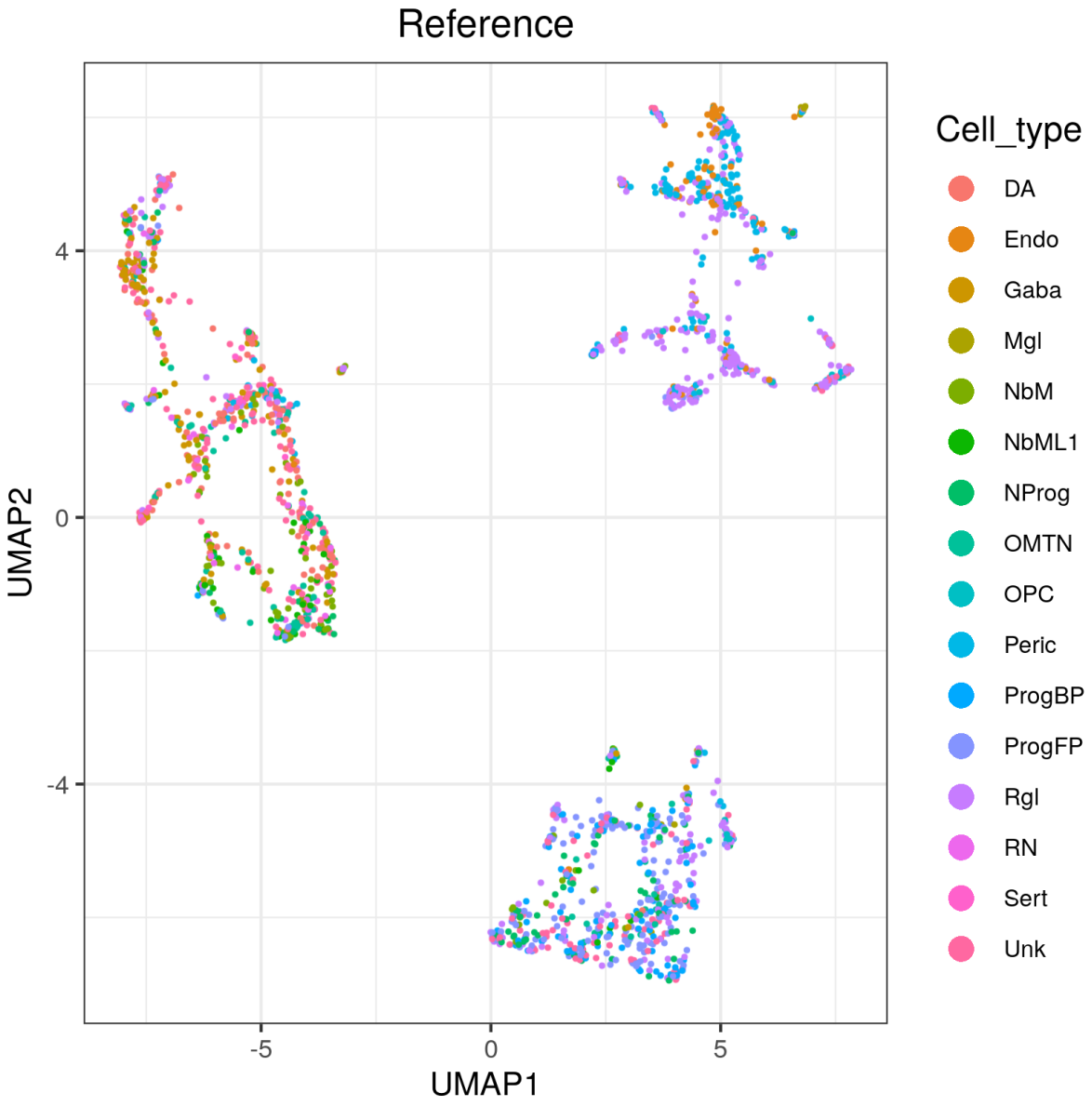
Mapa de calor da expressão (*counts*) dos genes de interesse. As linhas são os genes, enquanto as colunas são amostras agrupadas pelo tipo celular atribuído pelo Seurat.

Fonte: Elaborado pelo autor (2022).

2.2.2.3 Symphony

A ferramenta Symphony também gera figuras de redução de dimensionalidade. A Figura 23 representa o UMAP do conjunto de dados de referência, mostrando três grupos de células claramente divididos. O arquivo desta figura é o `reference_umap.png`.

Figura 23 – UMAP do conjunto de dados de referência do Symphony

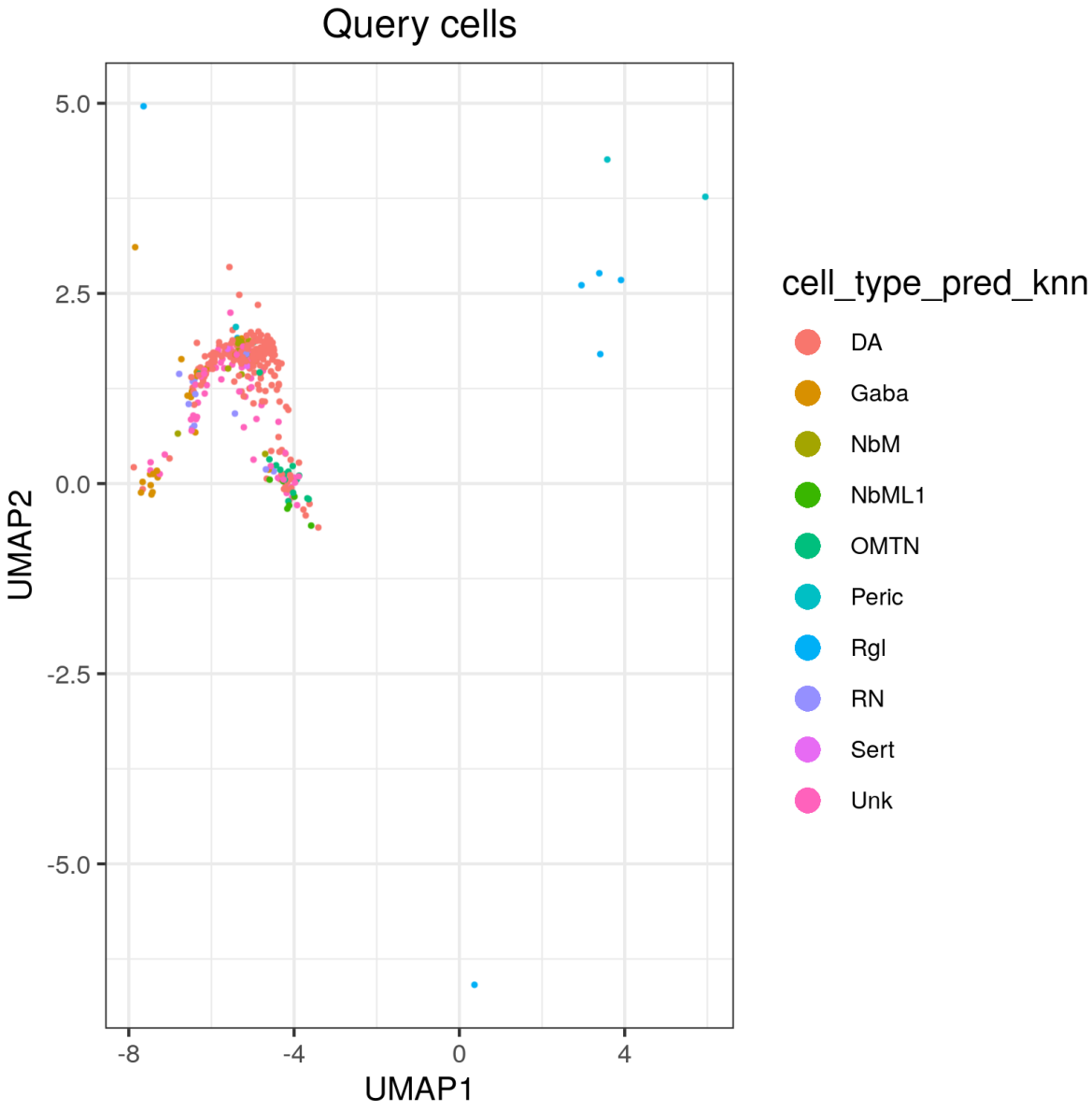


UMAP dos dados de treino e teste utilizados pelo Symphony.

Fonte: Elaborado pelo autor (2022).

A Figura 24 representa o UMAP do conjunto de dados questão. O arquivo desta figura é o `umap_query.png`. Suas células foram alocadas próximo do agrupamento em que se encontram as células de neurônios dopaminérgicos.

Figura 24 – UMAP do conjunto de dados questão do Symphony

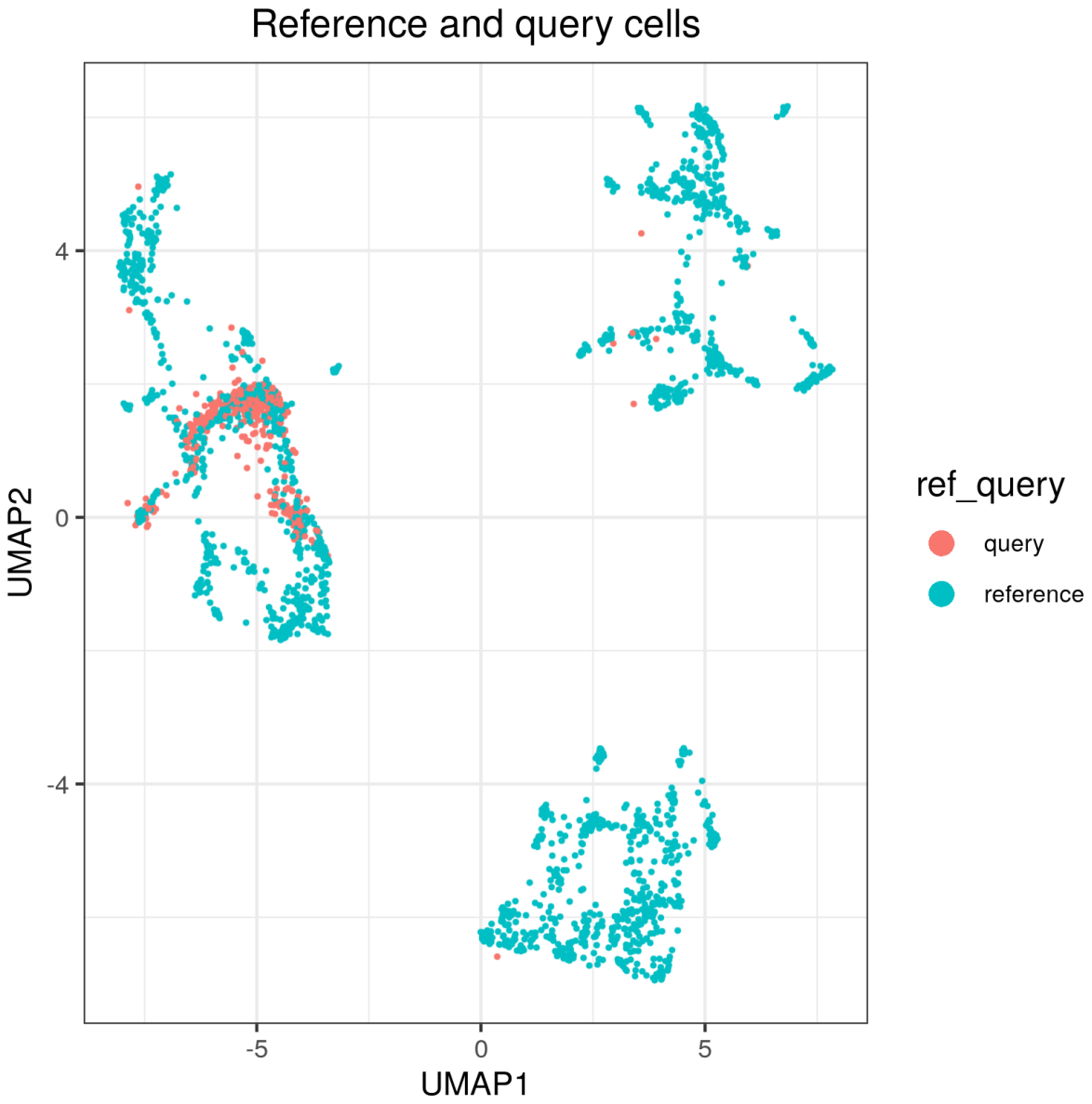


UMAP dos dados questão classificados pelo Symphony.

Fonte: Elaborado pelo autor (2022).

A Figura 25 representa o UMAP do conjunto de dados de referência juntamente com os questão. Se utilizarmos a Figura 19 para auxiliar a visualização, vemos que os dados questão são alocados próximos de onde os neurônios dopaminérgicos se encontram. O arquivo desta figura é o `umap_mixed_queryAndReference.png`.

Figura 25 – UMAP do conjunto de dados de referência e questão do Symphony



UMAP com todas as células dos conjuntos de dados referência e questão agrupadas. A cor identifica seu conjunto de dados de origem.

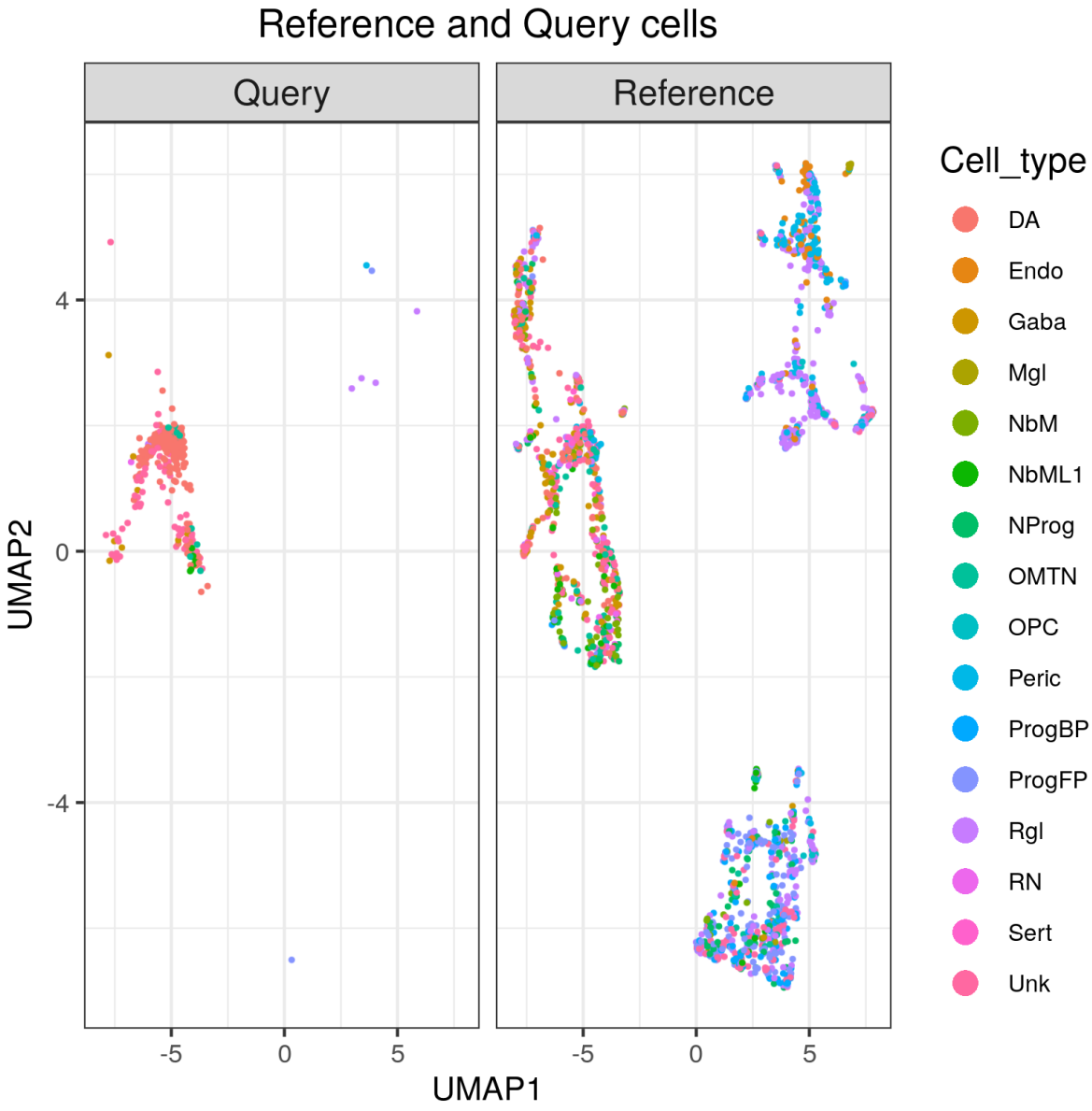
Fonte: Elaborado pelo autor (2022).

2.2.2.4 FUSCA

A ferramenta FUSCA, assim como sua predecessora Cellrouter (LUMMERTZ DA ROCHA et al., 2018) é focada em análises de trajetória celular entre dois estados. Embora essas análises sejam extremamente pertinentes no contexto em que estamos trabalhando, elas fogem do escopo desta *pipeline*. Em uma futura versão, a adição dessas funcionalidades seria uma ótima melhoria para esta *pipeline*. O FUSCA também permite a criação de diversas tabelas de redução de dimensionalidade, como

PCA, UMAP (que será utilizada) e tSNE. Dessa forma, também é possível utilizar o pacote FUSCA para tipagem celular do conjunto de dados questão, com o auxílio da ferramenta já citada Symphony. Essa sinergia de ferramentas é uma poderosa metodologia para se obter melhores resultados de ambas. A Figura 26 é um gráfico de redução de dimensionalidade utilizando UMAP (já citado) gerado pelo FUSCA e Symphony. Ela pode ser acessada no arquivo `umap_queryAndReference.png`. Como resultado, podemos ver que as células questão estão alocadas na mesma região de espaço bidimensional dos neurônios dopaminérgicos no conjunto de referência. Mais uma vez, isso indica uma concordância dos tipos celulares classificados com os descrito por La Manno et al, 2011.

Figura 26 – UMAP do conjunto de dados referência e questão do FUSCA



UMAP dos conjuntos de dados questão e referência, construídos pelo conjunto de ferramentas Symphony e FUSCA.

Fonte: Elaborado pelo autor (2022).

3 CONCLUSÃO

Este trabalho criou uma *pipeline* de análise de dados de transcriptômica de células únicas no contexto de células-tronco pluripotentes induzidas, utilizando ferramentas bem estabelecidas pela comunidade científica (DING et al., 2021). A *pipeline* foi feita de forma modular, utilizando o gerenciador de fluxo de trabalho Snakemake e compartimentalizando as dependências em uma imagem Singularity, para torna-la reprodutível em qualquer arquitetura de sistema. Utilizamos um conjunto

de dados de células cerebrais como referência para classificar um conjunto de dados de células neuronais dopaminérgicas provindas de células-tronco pluripotentes induzidas. Aplicamos quatro ferramentas, cada uma com um método de classificação diferente e com sucesso determinamos o tipo celular das células questão, por meio de figuras que poderiam ser utilizadas em um artigo ou relatório. Além disso, criamos uma métrica para medir a importância que determinada característica tem para a classificação nos tipos celulares dentro do singleCellNet. A métrica foi eficaz na demonstração das características que foram importantes para classificar um determinado tipo celular, trazendo significância biológica à uma informação que até então era apenas computacional. É importante deixar claro que seu desenvolvimento está em estágios iniciais e a informação por ela gerada deve ser averiguada. Ao se realizar mais testes e estudos, com diferentes conjuntos de dados de diferentes tecidos, seu desenvolvimento será melhorado, afinando as informações geradas e ampliando seu potencial de aplicação.

3.1 PERSPECTIVAS FUTURAS

Com o avanço do conhecimento acerca das células-tronco, veremos grandes descobertas na área de medicina de precisão. Com melhorias nos protocolos de reprogramação celulares, as possibilidades de terapia são vastas. Protocolos novos de diferenciação direta de um tipo celular a outro (sem passar pela fase de pluripotência) também reservam novas descobertas (COHEN; MELTON, 2011). No futuro, a cura para doenças de fertilidade e neurodegenerativas, criação de novos órgãos saudáveis e terapias de rejuvenescimento pode ser possíveis, com o simples acesso a bancos de células-tronco (ZAKRZEWSKI et al., 2019).

Contudo, ainda encontramos grandes barreiras que precisam ser vencidas. Por exemplo, a taxa de formação de teratomas em células induzidas é muito maior do que a de células-tronco embrionárias (NARSINH et al., 2011). Além disso, isolar células-tronco de um paciente também não é uma tarefa trivial.

A bioinformática terá, sem dúvida, um papel chave nesse futuro. O barateamento da geração de dados, juntamente dos avanços na ciência da computação, guarda um futuro brilhante à esta área tão recente da biologia.

REFERÊNCIAS

BARKER, Roger A.; DROUIN-OUELLET, Janelle; PARMAR, Malin. Cell-based therapies for Parkinson disease—past insights and future potential. **Nature Reviews. Neurology**, *[S. l.]*, v. 11, n. 9, p. 492–503, 2015. DOI: 10.1038/nrneurol.2015.123.

BARRETT, Tanya et al. NCBI GEO: archive for functional genomics data sets—update. **Nucleic Acids Research**, *[S. l.]*, v. 41, n. D1, p. D991–D995, 2012. DOI: 10.1093/nar/gks1193.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. [s.l: s.n.]. Disponível em: https://link.springer.com/book/9780387310732. Acesso em: 2 fev. 2022.

BOYD, Kendrick; ENG, Kevin H.; PAGE, C. David. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. *In*: (Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, Filip Železný, Org.)MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES 2013, Berlin, Heidelberg. **Anais** [...]. Berlin, Heidelberg: Springer, 2013. p. 451–466. DOI: 10.1007/978-3-642-40994-3_29.

BREIMAN, Leo. Random Forests. **Machine Learning**, *[S. l.]*, v. 45, n. 1, p. 5–32, 2001. DOI: 10.1023/A:1010933404324.

BUENROSTRO, Jason; WU, Beijing; CHANG, Howard; GREENLEAF, William. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. **Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]**, *[S. l.]*, v. 109, p. 21.29.1-21.29.9, 2015. DOI: 10.1002/0471142727.mb2129s109.

CARNEIRO, José; JUNQUEIRA, Luiz Carlos. **Biologia Celular e Molecular**. 9. ed. Rio de Janeiro: Guanabara Koogan, 2012.

CHAMBERS, Stuart M.; FASANO, Christopher A.; PAPAPETROU, Eirini P.; TOMISHIMA, Mark; SADELAIN, Michel; STUDER, Lorenz. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. **Nature Biotechnology**, *[S. l.]*, v. 27, n. 3, p. 275–280, 2009. DOI: 10.1038/nbt.1529.

CHEN, Geng; NING, Baitang; SHI, Tieliu. Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. **Frontiers in Genetics**, *[S. l.]*, v. 10, 2019. Disponível em: https://www.frontiersin.org/article/10.3389/fgene.2019.00317. Acesso em: 31 jan. 2022.

CHEN, Michael J. et al. Transcriptome Dynamics of Hematopoietic Stem Cell Formation Revealed Using a Combinatorial Runx1 and Ly6a Reporter System. **Stem Cell Reports**, *[S. l.]*, v. 14, n. 5, p. 956–971, 2020. DOI: 10.1016/j.stemcr.2020.03.020.

COHEN, Dena E.; MELTON, Douglas. Turning straw into gold: directing cell fate for regenerative medicine. **Nature Reviews Genetics**, *[S. l.]*, v. 12, n. 4, p. 243–252, 2011. DOI: 10.1038/nrg2938.

COHEN, Jacob. A Coefficient of Agreement for Nominal Scales. **Educational and Psychological Measurement**, *[S. l.]*, v. 20, n. 1, p. 37–46, 1960. DOI: 10.1177/001316446002000104.

DARMANIS, Spyros et al. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. **Cell Reports**, *[S. l.]*, v. 21, n. 5, p. 1399–1410, 2017. DOI: 10.1016/j.celrep.2017.10.030.

DELGOBO, Murilo et al. An evolutionary recent IFN/IL-6/CEBP axis is linked to monocyte expansion and tuberculosis severity in humans. **eLife**, *[S. l.]*, v. 8, p. e47013, 2019. DOI: 10.7554/eLife.47013.

DING, Dah-Ching; SHYU, Woei-Cherng; LIN, Shinn-Zong. Mesenchymal stem cells. **Cell Transplantation**, *[S. l.]*, v. 20, n. 1, p. 5–14, 2011. DOI: 10.3727/096368910X.

DING, Jun; ALAVI, Amir; EBRAHIMKHANI, Mo R.; BAR-JOSEPH, Ziv. Computational tools for analyzing single-cell data in pluripotent cell differentiation studies. **Cell Reports Methods**, *[S. l.]*, v. 1, n. 6, p. 100087, 2021. DOI: 10.1016/j.crmeth.2021.100087.

EBRAHIMI, Behnam. Reprogramming barriers and enhancers: strategies to enhance the efficiency and kinetics of induced pluripotency. **Cell Regeneration**, *[S. l.]*, v. 4, p. 10, 2015. DOI: 10.1186/s13619-015-0024-9.

ELBULUK, Ameer; EINHORN, Thomas A.; IORIO, Richard. A Comprehensive Review of Stem-Cell Therapy. **JBJS Reviews**, *[S. l.]*, v. 5, n. 8, p. e15–e15, 2017. DOI: 10.2106/JBJS.RVW.17.00002.

FIX, Evelyn; HODGES, Jr. **Discriminatory Analysis - Nonparametric Discrimination: Small Sample Performance**. [s.l.] : CALIFORNIA UNIV BERKELEY, 1952. Disponível em: https://apps.dtic.mil/sti/citations/ADA800391. Acesso em: 2 fev. 2022.

GEMAN, Donald; D'AVIGNON, Christian; NAIMAN, Daniel Q.; WINSLOW, Raimond L. Classifying Gene Expression Profiles from Pairwise mRNA Comparisons. **Statistical Applications in Genetics and Molecular Biology**, *[S. l.]*, v. 3, n. 1, 2004. DOI: 10.2202/1544-6115.1071. Disponível em: https://www.degruyter.com/document/doi/10.2202/1544-6115.1071/html. Acesso em: 16 fev. 2022.

GONG, Zhen; LIU, Jianyun; XIE, Xin; XU, Xiaoyuan; WU, Ping; LI, Huimin; WANG, Yaqin; LI, Weidong; XIONG, Jianjun. Identification of potential target genes of USP22 via ChIP-seq and RNA-seq analysis in HeLa cells. **Genetics and Molecular Biology**, *[S. l.]*, v. 41, p. 488–495, 2018. DOI: 10.1590/1678-4685-GMB-2017-0164.

GRAND VIEW RESEARCH. Cell Therapy Market Size, Share & Trends Analysis Report By Use-type, By Therapy Type (Autologous, Allogenic), By Region (North America, Europe, Asia Pacific, Latin America, MEA), And Segment Forecasts, 2021 - 2028. *[S. l.]*, p. 120, 2021.

GRÜNING, Björn et al. Practical Computational Reproducibility in the Life Sciences. **Cell Systems**, *[S. l.]*, v. 6, n. 6, p. 631–635, 2018. DOI: 10.1016/j.cels.2018.03.014.

GURDON, J. B. The Developmental Capacity of Nuclei taken from Intestinal Epithelium Cells of Feeding Tadpoles. **Development**, *[S. l.]*, v. 10, n. 4, p. 622–640, 1962. DOI: 10.1242/dev.10.4.622.

HAO, Yuhan et al. Integrated analysis of multimodal single-cell data. **Cell**, *[S. l.]*, v. 184, n. 13, p. 3573- 3587.e29, 2021. DOI: 10.1016/j.cell.2021.04.048.

HAQUE, Ashraful; ENGEL, Jessica; TEICHMANN, Sarah A.; LÖNNBERG, Tapio. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. **Genome Medicine**, *[S. l.]*, v. 9, n. 1, p. 75, 2017. DOI: 10.1186/s13073-017-0467-4.

HEATHER, James M.; CHAIN, Benjamin. The sequence of sequencers: The history of sequencing DNA. **Genomics**, *[S. l.]*, v. 107, n. 1, p. 1–8, 2016. DOI: 10.1016/j.ygeno.2015.11.003.

HOCHEDLINGER, Konrad; PLATH, Kathrin. Epigenetic reprogramming and induced pluripotency. **Development**, *[S. l.]*, v. 136, n. 4, p. 509–523, 2009. DOI: 10.1242/dev.020867.

HOLLEY, R. W.; APGAR, J.; EVERETT, G. A.; MADISON, J. T.; MARQUISEE, M.; MERRILL, S. H.; PENSWICK, J. R.; ZAMIR, A. STRUCTURE OF A RIBONUCLEIC ACID. **Science (New York, N.Y.)**, *[S. l.]*, v. 147, n. 3664, p. 1462–1465, 1965. DOI: 10.1126/science.147.3664.1462.

HOLLEY, Robert W.; APGAR, Jean; MERRILL, Susan H.; ZUBKOFF, Paul L. NUCLEOTIDE AND OLIGONUCLEOTIDE COMPOSITIONS OF THE ALANINE-, VALINE-, AND TYROSINE-ACCEPTOR “SOLUBLE” RIBONUCLEIC ACIDS OF YEAST. **Journal of the American Chemical Society**, *[S. l.]*, v. 83, n. 23, p. 4861–4862, 1961. DOI: 10.1021/ja01484a040.

HWANG, Byungjin; LEE, Ji Hyun; BANG, Duhee. Single-cell RNA sequencing technologies and bioinformatics pipelines. **Experimental & Molecular Medicine**, *[S. l.]*, v. 50, n. 8, p. 1–14, 2018. DOI: 10.1038/s12276-018-0071-8.

JIANG, Peng; THOMSON, James A.; STEWART, Ron. Quality control of single-cell RNA-seq by SinQC. **Bioinformatics**, *[S. l.]*, v. 32, n. 16, p. 2514–2516, 2016. DOI: 10.1093/bioinformatics/btw176.

KANG, Joyce B.; NATHAN, Aparna; WEINAND, Kathryn; ZHANG, Fan; MILLARD, Nghia; RUMKER, Laurie; MOODY, D. Branch; KORSUNSKY, Ilya; RAYCHAUDHURI, Soumya. Efficient and precise single-cell reference atlas mapping with Symphony. **Nature Communications**, *[S. l.]*, v. 12, n. 1, p. 5890, 2021. DOI: 10.1038/s41467-021-25957-x.

KÖSTER, Johannes; RAHMANN, Sven. Snakemake—a scalable bioinformatics workflow engine. **Bioinformatics**, *[S. l.]*, v. 28, n. 19, p. 2520–2522, 2012. DOI: 10.1093/bioinformatics/bts480.

KRIKS, Sonja et al. Dopamine neurons derived from human ES cells efficiently engraft in animal models of Parkinson’s disease. **Nature**, *[S. l.]*, v. 480, n. 7378, p. 547–551, 2011. DOI: 10.1038/nature10648.

LA MANNO, Gioele et al. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. **Cell**, *[S. l.]*, v. 167, n. 2, p. 566- 580.e19, 2016. DOI: 10.1016/j.cell.2016.09.027.

LARSON, Ron; FARBER, Betsy; RUNGER, George C. **Estatística aplicada**. 2. ed. São Paulo: Pearson Prentice Hall, 2004.

LEIPZIG, Jeremy. A review of bioinformatic pipeline frameworks. **Briefings in Bioinformatics**, *[S. l.]*, v. 18, n. 3, p. 530–536, 2017. DOI: 10.1093/bib/bbw020.

LO, Bernard; PARHAM, Lindsay. Ethical Issues in Stem Cell Research. **Endocrine Reviews**, *[S. l.]*, v. 30, n. 3, p. 204–213, 2009. DOI: 10.1210/er.2008-0031.

LOMAN, Nicholas J.; QUICK, Joshua; SIMPSON, Jared T. A complete bacterial genome assembled de novo using only nanopore sequencing data. **Nature Methods**, *[S. l.]*, v. 12, n. 8, p. 733–735, 2015. DOI: 10.1038/nmeth.3444.

LUMMERTZ DA ROCHA, Edroaldo et al. Reconstruction of complex single-cell trajectories using CellRouter. **Nature Communications**, *[S. l.]*, v. 9, n. 1, p. 892, 2018. DOI: 10.1038/s41467-018-03214-y.

LUO, Jake; WU, Min; GOPUKUMAR, Deepika; ZHAO, Yiqing. Big Data Application in Biomedical Research and Health Care: A Literature Review. **Biomedical Informatics Insights**, *[S. l.]*, v. 8, p. 1–10, 2016. DOI: 10.4137/BII.S31559.

MADRID, Marinna; SUMEN, Cenk; AIVIO, Suvi; SAKLAYEN, Nabiha. Autologous Induced Pluripotent Stem Cell–Based Cell Therapies: Promise, Progress, and Challenges. **Current Protocols**, *[S. l.]*, v. 1, n. 3, p. e88, 2021. DOI: 10.1002/cpz1.88.

MCHUGH, Mary L. Interrater reliability: the kappa statistic. **Biochemia Medica**, *[S. l.]*, v. 22, n. 3, p. 276–282, 2012.

MCINNES, Leland; HEALY, John; MELVILLE, James. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. **arXiv:1802.03426 [cs, stat]**, *[S. l.]*, 2020. Disponível em: http://arxiv.org/abs/1802.03426. Acesso em: 2 fev. 2022.

MORADI, Sharif; MAHDIZADEH, Hamid; ŠARIĆ, Tomo; KIM, Johnny; HARATI, Javad; SHAHSAVARANI, Hosein; GREBER, Boris; MOORE, Joseph B. Research and therapy with induced pluripotent stem cells (iPSCs): social, legal, and ethical considerations. **Stem Cell Research & Therapy**, *[S. l.]*, v. 10, n. 1, p. 341, 2019. DOI: 10.1186/s13287-019-1455-y.

NARSINH, Kazim H. et al. Single cell transcriptional profiling reveals heterogeneity of human induced pluripotent stem cells. **The Journal of Clinical Investigation**, *[S. l.]*, v. 121, n. 3, p. 1217–1221, 2011. DOI: 10.1172/JCI44635.

OU, Minglin et al. Single-cell sequencing reveals the potential oncogenic expression atlas of human iPSC-derived cardiomyocytes. **Biology Open**, *[S. l.]*, v. 10, n. 2, p. bio053348, 2021. DOI: 10.1242/bio.053348.

PERKEL, Jeffrey M. Workflow systems turn raw data into scientific knowledge. **Nature**, *[S. l.]*, v. 573, n. 7772, p. 149–150, 2019. DOI: 10.1038/d41586-019-02619-z.

SALMAN, Shaeke; LIU, Xiuwen. Overfitting Mechanism and Avoidance in Deep Neural Networks. **arXiv:1901.06566 [cs, stat]**, *[S. l.]*, 2019. Disponível em: http://arxiv.org/abs/1901.06566. Acesso em: 14 mar. 2022.

SANGER, F.; BROWNLEE, G. G.; BARRELL, B. G. A two-dimensional fractionation procedure for radioactive nucleotides. **Journal of Molecular Biology**, *[S. l.]*, v. 13, n. 2, p. 373–398, 1965. DOI: 10.1016/s0022-2836(65)80104-8.

SCHOPPERLE, William M.; DEWOLF, William C. The TRA-1-60 and TRA-1-81 human pluripotent stem cell markers are expressed on podocalyxin in embryonal carcinoma. **Stem Cells (Dayton, Ohio)**, *[S. l.]*, v. 25, n. 3, p. 723–730, 2007. DOI: 10.1634/stemcells.2005-0597.

SHI, Yanhong; INOUE, Haruhisa; WU, Joseph C.; YAMANAKA, Shinya. Induced pluripotent stem cell technology: a decade of progress. **Nature Reviews Drug Discovery**, *[S. l.]*, v. 16, n. 2, p. 115–130, 2017. DOI: 10.1038/nrd.2016.245.

STUART, Tim et al. Comprehensive Integration of Single-Cell Data. **Cell**, *[S. l.]*, v. 177, n. 7, p. 1888- 1902.e21, 2019. DOI: 10.1016/j.cell.2019.05.031.

TAKAHASHI, Kazutoshi; TANABE, Koji; OHNUKI, Mari; NARITA, Megumi; ICHISAKA, Tomoko; TOMODA, Kiichiro; YAMANAKA, Shinya. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. **Cell**, *[S. l.]*, v. 131, n. 5, p. 861–872, 2007. DOI: 10.1016/j.cell.2007.11.019.

TAKAHASHI, Kazutoshi; YAMANAKA, Shinya. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. **Cell**, *[S. l.]*, v. 126, n. 4, p. 663–676, 2006. DOI: 10.1016/j.cell.2006.07.024.

TAKASATO, Minoru et al. Kidney organoids from human iPS cells contain multiple lineages and model human nephrogenesis. **Nature**, *[S. l.]*, v. 526, n. 7574, p. 564–568, 2015. DOI: 10.1038/nature15695.

TAN, Yuqi; CAHAN, Patrick. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. **Cell Systems**, *[S. l.]*, v. 9, n. 2, p. 207- 213.e2, 2019. DOI: 10.1016/j.cels.2019.06.004.

TANG, Fuchou et al. mRNA-Seq whole-transcriptome analysis of a single cell. **Nature Methods**, *[S. l.]*, v. 6, n. 5, p. 377–382, 2009. DOI: 10.1038/nmeth.1315.

TREMOLADA, Carlo; COLOMBO, Valeria; VENTURA, Carlo. Adipose Tissue and Mesenchymal Stem Cells: State of the Art and Lipogems® Technology Development. **Current Stem Cell Reports**, *[S. l.]*, v. 2, n. 3, p. 304–312, 2016. DOI: 10.1007/s40778-016-0053-5.

VAN ROSSUM, Guid; DRAKE, Fred L. **Python 3 Reference Manual**. Scotts Valley, CA: CreateSpace, 2009.

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. **Nature**, *[S. l.]*, v. 171, n. 4356, p. 737–738, 1953. DOI: 10.1038/171737a0.

WRATTEN, Laura; WILM, Andreas; GÖKE, Jonathan. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. **Nature Methods**, *[S. l.]*, v. 18, n. 10, p. 1161–1168, 2021. DOI: 10.1038/s41592-021-01254-9.

ZAKRZEWSKI, Wojciech; DOBRZYŃSKI, Maciej; SZYMONOWICZ, Maria; RYBAK, Zbigniew. Stem cells: past, present, and future. **Stem Cell Research & Therapy**, *[S. l.]*, v. 10, n. 1, p. 68, 2019. DOI: 10.1186/s13287-019-1165-5.

ZANIRATI, Gabriele; PROVENZI, Laura; LIBERMANN, Lucas Lobraico; BIZOTTO, Sabrina Comin; GHILARDI, Isadora Machado; MARINOWIC, Daniel Rodrigo; SHETTY, Ashok K.; DA COSTA, Jaderson Costa. Stem cell-based therapy for COVID-19 and ARDS: a systematic review. **npj Regenerative Medicine**, *[S. l.]*, v. 6, n. 1, p. 1–15, 2021. DOI: 10.1038/s41536-021-00181-9.