UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
PROGRAMA DE PÓS-GRADUAÇÃO EM GENÉTICA

CARLOS ALBERTO OLIVEIRA DE BIAGI JUNIOR

Sequenciamento de *RNA* de células únicas como ferramenta para a compreensão do microambiente intratumoral

Ribeirão Preto
2022

CARLOS ALBERTO OLIVEIRA DE BIAGI JUNIOR

Sequenciamento de *RNA* de células únicas como ferramenta para a compreensão do microambiente intratumoral

Tese apresentada ao Programa de Pós-Graduação em Genética da Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, para obtenção do título de Doutor em Ciências.

Área de concentração: Genética

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 14 de Abril de 2022. A versão original encontra-se disponível tanto na Biblioteca da Unidade que aloja o Programa, quanto na Biblioteca Digital de Teses e Dissertações da USP (BDTD).

Orientador: Prof. Dr. Wilson Araújo da Silva Junior

Ribeirão Preto

2022

Biagi-Jr, Carlos Alberto Oliveira de

Sequenciamento de *RNA* de células únicas como ferramenta para a compreensão do microambiente intratumoral. Ribeirão Preto, 2022.

179 p. : il. ; 30 cm

Tese de Doutorado, apresentada à Faculdade de Medicina de Ribeirão Preto/USP. Área de concentração: Genética.

Orientador: Araújo da Silva Junior, Wilson.

1. Sequenciamento de *RNA* de células únicas. 2. Bioinformática. 3. Melanoma. 4. Câncer de pulmão de pequenas células.

Tese de autoria de Carlos Alberto Oliveira de Biagi Junior, sob o título **"Sequenciamento de *RNA* de células únicas como ferramenta para a compreensão do micro-ambiente intratumoral"**, apresentada à Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, para obtenção do título de Doutor em Ciências pelo Programa de Pós-graduação em Genética, na área de concentração Genética, aprovada em 14 de Abril de 2022 pela comissão julgadora constituída pelos doutores:

Prof. Dr. Wilson Araújo da Silva Junior
FMRP - USP
Presidente

Prof. Dr. Miguel Luiz Batista Junior
UMC

Prof. Dr. Rodrigo Alexandre Panepucci
FUNDHERP (FMRP)

Profa. Dra. Tathiane Maistro Malta Pereira
FCFRP (FMRP)

Apoio e suporte financeiro

Este trabalho foi desenvolvido no Laboratório de Genética Molecular e Bioinformática (LGMB), localizado no Hemocentro de Ribeirão Preto, da Faculdade de Medicina de Ribeirão Preto (FMRP), da Universidade de São Paulo (USP), e contou com o apoio ou suporte financeiro das seguintes instituições:

- Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES). Código de Financiamento: 001;
- Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP). Código de Financiamento: 13/08135-2;
- Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Código de Financiamento: 14/50947-7;
- Fundação de Apoio ao Ensino, Pesquisa e Assistência (FAEPA);
- Faculdade de Medicina de Ribeirão Preto (FMRP);
- Fundação Hemocentro de Ribeirão Preto.

Este trabalho é todo dedicado aos meus pais, pois é graças ao seus esforços que hoje sou quem sou.

Agradecimentos

Agradeço:

A Deus, por tudo!

E graças a Ele, muitas pessoas boas fizeram parte desta caminhada.

À minha preciosa família, a eles dedico esse trabalho. Aos meus pais, Carlos e Selma, por serem modelos de coragem, pelo apoio incondicional, incentivo, amizade e paciência demonstrados e total ajuda na superação dos obstáculos que ao longo desta caminhada foram surgindo. À minha irmã Natália e meu cunhado Filipe, por estarem sempre ao meu lado me apoiando e encorajando em todos os momentos. Agradeço a Laurinha por ter sido um presente na minha vida e por ter a oportunidade de ser seu titio. À minha esposa Karen, que esteve o tempo todo ao meu lado, incondicionalmente, nos momentos mais difíceis sempre me fazendo acreditar que chegaria ao final desta difícil, porém gratificante etapa.

Ao Prof. Dr. Wilson Araújo da Silva Junior por ter-me deixado fazer parte do seu grupo de trabalho e, ter acreditado em mim e nas minhas capacidades. Agradeço ainda por todas as oportunidades proporcionadas, caminhos abertos e portas escancaradas. Nunca vou me esquecer do curso de verão em bioinformática de 2017, onde toda essa história começou. Agradeço ainda a confiança em mim depositada para trabalhar com o temática deste trabalho. Em seu nome agradeço a secretaria da pós-graduação por todo o apoio e agilidade em resolver as minhas inquietações e desafios, especial agradecimento à Susie Nalon.

Ao Programa de Pós-Graduação em Genética, Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo (USP) e Hemocentro de Ribeirão Preto pelo acolhimento nestes anos e por proporcionar a oportunidade de estudar e desenvolver pesquisa em um dos melhores locais do Brasil.

Ao meu grande amigo Cleidson pelas conversas iniciais sobre *single-cell* em 2017 que tornou-se uma das etapas mais especiais da minha vida com a oportunidade de passar 10 meses na Universidade de Colônia. Em especial agradeço ao Prof. Dr. Martin Peifer pelo acolhimento em seu laboratório nesse período e por me mostrar que tudo é possível quando trabalhamos em colaboração. Aos bons amigos que fiz e que me ensinaram e ensinam até hoje: Julie George, Miloš Nikolić, Marcel Schmiel e Nima Abedpour.

Aos meus amigos e colegas do BiT: Raul, Ricardo, Patrícia, Marcelo, João, Rafael, Jéssica, entre outros que não menciono o nome mas que sabem quem são, amigos que estiveram ao meu lado durante esta fase, pelo companheirismo, força e apoio em certos momentos difíceis.

Aos meus colegas do LGMB (Laboratório de Genética Molecular e Bioinformática), gostaria de agradecer pelos momentos que passamos. Agradeço o bom convívio, as boas discussões e, a alegria que por vezes se instalava.

Aos meus bons amigos do IPEC (Instituto de Pesquisa para o Câncer): David, Lara, Isabela, Madá, Bárbara Luísa, Cadu, Manu, Kamila, Daiane, Fernanda, Kati, Martinha e Bárbara Paz. Obrigado pelo acolhimento, incentivo e amizade. Vocês são muito queridos.

Aos servidores(as) das instituições citadas acima. Muito obrigado pelas boas conversas, risadas e conselhos.

Por fim, aos professores e amigos que me encorajaram e me inspiraram a trilhar essa caminhada.

“Believe!”

(Ted Lasso)

Resumo

Biagi-Jr, Carlos Alberto Oliveira de. **Sequenciamento de RNA de células únicas como ferramenta para a compreensão do microambiente intratumoral**. 2022. 179 f. Tese (Doutorado em Genética) – Faculdade de Medicina de Ribeirão Preto da Universidade de São Paulo, Ribeirão Preto, 2022.

O sequenciamento de RNA de célula única (*scRNA-seq*) é uma técnica revolucionária que permite caracterizar os transcriptomas de muitas células individuais em uma amostra. A aplicação desta abordagem é imprescindível para conhecer os mecanismos moleculares que regulam a progressão tumoral, na busca de novos alvos elegíveis para o desenvolvimento de novos biomarcadores e aprimoramento de terapias já existentes. Neste trabalho, exploramos dados públicos de melanoma e dados gerados *in house* de Câncer de Pulmão de Pequenas Células (CPPC) para buscar um melhor entendimento destas doenças. Utilizamos ferramentas para avaliar as subpopulações de células, tipos celulares, vias enriquecidas, trajetórias e comunicação célula-célula. Em melanoma, identificamos um *lncRNA* chamado *TRHDE-AS1* que parece atuar na via transição epitélio-mesênquima (*EMT*) em parceria com o *HOTAIR*. Com isso foi possível propor um circuito gênico regulado por *feedback* negativo, onde o *HOTAIR* regula positivamente o *TRHDE-AS1*, e o *TRHDE-AS1*, ao atingir um certo nível de expressão, passa a regular negativamente o *HOTAIR*. A partir da análise de inferência de trajetória entre os tipos celulares foi possível identificar um ponto de diferenciação onde há a ativação da via de *EMT* a partir da presença de células endoteliais e fibroblastos associados ao câncer (*CAF*). Para CPPC avaliou-se o perfil transcricional das células únicas e concluiu-se que os dados estão de acordo com a classificação atual de subtipos moleculares, expressando *NEUROD1*, *POU2F3* e *ASCL1*. Não foram identificadas amostras *YAP1* nesse estudo. Por fim, a via de sinalização *NOTCH* é demonstrada como um papel importante para a plasticidade transcricional em CPPC além das vias de sinalização *MK*, *JAM*, *PTN*, *CD99*, *NMU*, *CADM* e *NCAM*, identificadas a partir da comunicação célula-célula, que possuem um papel importante para o desenvolvimento de terapias e biomarcadores. Este estudo fornece abordagens para a identificação de novos biomarcadores e desenvolvimento de terapias a partir de dados de sequenciamento de RNA de células únicas.

Palavras-chaves: Sequenciamento de RNA de células únicas. Bioinformática. Melanoma. Câncer de pulmão de pequenas células.

Abstract

Biagi-Jr, Carlos Alberto Oliveira de. **Single cell RNA sequencing as a tool for understanding the intratumoral microenvironment**. 2022. 179 f. Thesis (PhD in Genetics) – School of Medicine of Ribeirao Preto from University of Sao Paulo, Ribeirao Preto, 2022.

Single-cell RNA sequencing (scRNA-seq) is a revolutionary technology that allows the characterization of the transcriptomes of many individual cells in one sample. The application of this approach is indispensable to know the molecular mechanisms regulating tumor progression in the search for eligible new targets for the development of new biomarkers and enhancement of existing therapies. In this work, we exploit public melanoma data and in-house generated Small Cell Lung Cancer (SCLC) data to seek a better understanding of these diseases. We use tools to assess cell subpopulations, cell types, enriched pathways, trajectories, and cell-cell communication. In melanoma, we identified a lncRNA called *TRHDE-AS1* that appears to act in the epithelium-mesenchymal transition (EMT) pathway in partnership with *HOTAIR*. With this, it was possible to propose a negative feedback regulated gene circuit, where *HOTAIR* positively regulates *TRHDE-AS1*, and *TRHDE-AS1*, upon reaching a certain expression level, negatively regulates *HOTAIR*. From the pathway inference analysis between cell types, it was possible to identify a point of differentiation where there is the activation of the EMT pathway from the presence of endothelial cells and cancer-associated fibroblasts (CAF). For SCLC, we evaluated the transcriptional profile of single cells and concluded that the data agree with the current classification of molecular subtypes, expressing *NEUROD1*, *POU2F3*, and *ASCL1*. No *YAP1* samples were identified in this study. Finally, the *NOTCH* signaling pathway is shown to play an essential role for transcriptional plasticity in SCLC in addition to the *MK*, *JAM*, *PTN*, *CD99*, *NMU*, *CADM*, and *NCAM* signaling pathways identified from cell-cell communication, which have an essential role in the development of therapies and biomarkers. This study provides approaches for identifying novel biomarkers and developing therapies from single-cell RNA sequencing data.

Keywords: Single cell RNA sequencing. Bioinformatics. Melanoma. Small cell lung cancer

Lista de figuras

Lista de tabelas

Lista de abreviaturas e siglas

AUC	*Area Under the Curve* (Área sob a curva)
BC	*Barcodes*
CAF	*Cancer Associated Fibroblasts* (Fibroblastos Associados ao Câncer)
CDX	*Cell-Derived Xenografts*
CPPC	Câncer de Pulmão de Pequenas Células
EMT	*Epithelial-Mesenchymal Transition* (Transição Epitélio Mesênquima)
FCR	Fase de Crescimento Radial
FCV	Fase de Crescimento Vertical
GSVA	*Gene Set Variation Analysis*
HGV	*High Variable Genes* (Genes Altamente Variáveis)
ITH	*Intratumor Heterogeneity* (Heterogeneidade Intratumoral)
lncRNAs	*Long non-coding RNAs* (RNAs longos não codificantes)
MEC	Matriz Extracelular
mRNA	RNA Mensageiro
ncRNAs	*Non-Coding RNAs* (RNAs não codificantes)
NK	*Natural Killers* (Exterminadoras Naturais)
PCA	*Principal Component Analysis* (Análise de Componente Principal)
PCIT	Análise da Correlação Parcial com Teoria da Informação
PCR	*Polymerase Chain Reaction* (Proteína C Reativa)
PDX	*Patient-Derived Xenografts*
RIF	Fatores de Impactos Regulatórios
rRNA	RNA Ribossomal

scRNA-seq	*Single-Cell RNA-Seq* (Sequenciamento de Células Únicas)
siRNA	*Small Infering RNA*
SNP	*Single Nucleotide Polimorphism* (Polimorfismo de nucleotídeo único)
TF	*Transcription Factor* (Fator de Transcrição)
TPM	Transcrito Por Milhão
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
UMAP	*Uniform Manifold Approximation and Projection*
UMI	Identificadores Moleculares Únicos

Sumário

1 Introdução

Nesta seção são abordados os conceitos iniciais para o entendimento do trabalho, como uma breve introdução sobre tecnologia de sequenciamento de células únicas, melanoma, câncer de pulmão de pequenas células (CPPC), *RNAs* não codificantes e a via de transição epitélio-mesenquimal.

1.1 Sequenciamento de RNA de células únicas (scRNA-seq)

Antes do ano 2000 era muito utilizada a técnica do microarranjo para estudar a expressão gênica. Após o ano 2000 ocorreu um grande avanço nas pesquisas e tecnologia, o que ocasionou o surgimento da técnica de *RNA-seq* que tem sido usada desde então. O *RNA-seq* mensura o nível de expressão para cada gene em uma grande população de células (*pool* de células). Esta técnica, apesar de trazer muitas informações relevantes para o estudo da expressão gênica, é insuficiente para estudar sistemas heterogêneos, como por exemplo estudos iniciais de desenvolvimento, tecidos complexos (cérebro), além de não fornecer *insights* sobre a natureza estocástica da expressão gênica. Para suprir tal deficiência, em 2009 foi publicado o primeiro artigo com a técnica de sequenciamento de células únicas por Tang *et al.* (2009).

O sequenciamento do RNA de células únicas (*scRNA-seq*) influenciou drasticamente os campos de pesquisa que vão da biologia do câncer, da biologia das células-tronco à imunologia. Em comparação com o *RNA-seq* do *bulk* de tecidos com milhões de células, o *scRNA-seq* oferece uma oportunidade para analisar a composição de tecidos/órgãos e a diversidade de estados celulares, bem como para detectar tipos de células raras. Com o aperfeiçoamento das tecnologias de sequenciamento, o sequenciamento do RNA de células únicas (*scRNA-seq*) está se tornando robusto e acessível para análise de transcriptomas.

O *scRNA-seq* permite a comparação dos transcriptomas de células individuais. Portanto, um dos principais usos de *scRNA-seq* tem sido avaliar semelhanças e diferenças transicionais dentro de uma população de células, com relatos iniciais revelando níveis de heterogeneidade anteriormente não reconhecidos, por exemplo, em células embrionárias e imunes. Assim, a análise de heterogeneidade genética continua sendo uma das principais razões para o início dos estudos de *scRNA-seq* (HAQUE *et al.*, 2017). Apesar de ter sido

apresentada pela primeira vez em 2009, esta técnica ganhou mais popularidade a partir de 2014, quando novos protocolos e menores custos de sequenciamento a tornaram mais acessível.

A técnica de *scRNA-seq* também permite estudar novas questões biológicas em que as alterações específicas no transcriptoma de uma célula são importantes, como por exemplo a identificação dos tipos celulares, heterogeneidade celular, estocasticidade de expressão gênica, inferência de redes reguladoras de genes através das células e inferência da velocidade do RNA. Os conjuntos de dados variam de 10^2 a 10^6 células e aumentam de tamanho a cada ano.

A análise da diversidade de expressão dentro de tumores (ou outros tecidos) geralmente revela duas camadas de informações: agrupamentos altamente distintos que podem ser considerados como "tipos celulares" (por exemplo, células malignas, células T e fibroblastos), e maior diversidade dentro de cada um desses agrupamentos celulares pode ser considerado como "estados funcionais", por exemplo a progressão ao longo do ciclo celular, diferentes estados metabólicos e outros programas genéticos dinâmicos. Um tema emergente dos estudos com abordagens de *scRNA-seq* é que a heterogeneidade intertumoral é maior nas células malignas do que para qualquer outro tipo de células não malignas, ou seja, para células malignas a heterogeneidade intertumoral é muito maior do que a heterogeneidade intratumoral. Com o poder dos métodos computacionais existentes para análise de dados de *scRNA-seq* possibilita-se que sejam interrogados os padrões de heterogeneidade intratumoral, que têm sido difíceis de avaliar com abordagens computacionais anteriores (FILBIN *et al.*, 2018; JERBY-ARNON *et al.*, 2018; TIROSH *et al.*, 2016a; TIROSH *et al.*, 2016b; VENTEICHER *et al.*, 2017).

1.1.1 Protocolo metodológico do sequenciamento do *RNA* de células únicas

A compreensão de sistemas biológicos complexos requer análises da expressão e da regulação das células. O sequenciamento de células únicas pode revelar os tipos de celulares dentro de um sistema, quais funções estão ocorrendo em cada célula e como essas células estão interagindo umas com as outras e seu microambiente. Com a análise de *bulk*, a expressão gênica é calculada como média entre as células. O sequenciamento de uma única célula mostra a expressão gênica por células individuais, proporcionando uma visão

muito mais profunda e específica da variação de célula para célula. É possível examinar os genomas, epigenomas, transcriptomas ou proteínas de células individuais.

O *workflow* (Figura 1) começa com o preparo inicial do tecido, que envolve o isolamento das células de seu ambiente nativo através do isolamento mecânico, digestão enzimática, ou uma combinação dos dois. O método para esta etapa depende dos tipos de células que precisam ser preparadas para uma suspensão monocelular viável. O enriquecimento é uma parte opcional desta etapa do processo. Duas formas comuns de enriquecer a amostra são a separação de células ativadas por fluorescência e a separação com base em esferas (*beads*) magnéticas. Isto remove células indesejadas, particularmente células mortas.

A partir daí, as células individuais precisam ser isoladas e compartimentadas. Há vários métodos disponíveis para isolar as células únicas, e sua abordagem depende de suas prioridades experimentais e de seu rendimento. As tecnologias de microfluidos permitem a caracterização de células únicas de alto rendimento de até dezenas de milhares de células por experimento. Os micropoços também podem capturar células individuais com alto rendimento.

Uma vez isoladas e compartimentadas as células, elas são lisadas para preparar sua biblioteca de alvos para o sequenciamento. Aqui, o alvo a ser interrogado é codificado por *barcode* de uma forma que permite identificar de qual célula ele veio. Dessa forma, a preparação da biblioteca começa. O método dependerá se você está estudando no genoma, epigenoma, transcriptoma, ou expressão de proteínas. Independentemente disso, seu alvo escolhido será amplificado e preparado para o sequenciamento. Em seguida, bibliotecas preparadas são carregadas em seu sequenciador de escolha.

A etapa final é a análise e visualização dos dados que será possível elucidar as complexidades por trás dos sistemas biológicos.

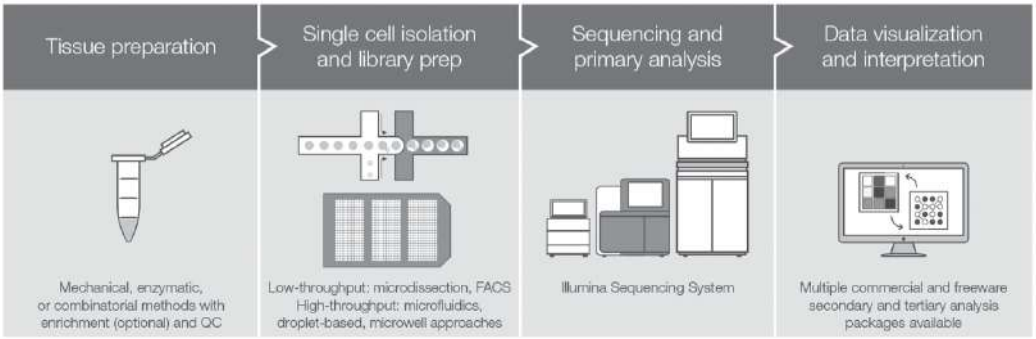

Figura 1 – *Workflow* de sequenciamento de células únicas procede do preparo inicial do tecido através do isolamento de células únicas e preparação da biblioteca, do sequenciamento e análise primária, e da visualização e interpretação dos dados.

Fonte – Retirada de ⟨http://tiny.cc/bv8muz⟩

1.1.2 Plataformas de sequenciamento de *scRNA-seq*

Smart-seq2 (PICELLI *et al.*, 2013) e *10X Genomics Chromium* são duas das plataformas de *scRNA-seq* mais utilizadas. *Smart-seq2* é baseado em placas de microtitulação (GRÜN; OUDENAARDEN, 2015; PICELLI *et al.*, 2014), onde o *mRNA* é separado e transcrito reversamente para *cDNA* para cada célula. As leituras mapeadas para um gene são usadas para quantificar sua abundância em cada célula, e transcritos por milhões (*TPM*) é uma métrica comum de normalização de expressão. Em contraste, *10X* é uma tecnologia baseada em gotículas, permitindo obter o perfil de expressão em todo o genoma para milhares de células de uma só vez. O número de identificadores moleculares únicos (*UMIs*) é considerado como uma apresentação direta do nível de expressão gênica. Tanto *TPM* (*Smart-seq2*) quanto *UMI* normalizado (*10X*) são analisados para detectar genes altamente variáveis (*HVGs*), que são frequentemente utilizados tanto para a classificação do fenótipo celular quanto para a identificação de novas subpopulações (WANG *et al.*, 2021).

O *Smart-seq2* é um dos métodos mais bem-sucedidos para detectar a expressão gênica de células únicas com alta robustez e confiabilidade. Por outro lado, a plataforma mais comumente utilizada atualmente é a da *10X Genomics*. Ambas as plataformas possuem suas próprias vantagens e desvantagens. Cabe ao pesquisador, no design do experimento, ter saber escolher a melhor plataforma de acordo com a sua necessidade (WANG *et al.*, 2021).

As tecnologias baseadas em gotículas (*10x Genomics*) frequentemente encontram o problema dos dupletos, onde uma gotícula pode conter duas ou mais células com o mesmo barcode durante a etapa de isolamento das células únicas. Dessa forma a gotícula de dupletos é contada como uma única célula no dado (WOLOCK; LOPEZ; KLEIN, 2019). De acordo com a composição dos dupletos, eles podem ser divididos em duas classes principais: os dupletos homotípicos, que se originam do mesmo tipo de célula, e os dupletos heterotípicos que surgem de células transcricionais distintas gerando um transcriptoma híbrido artificial (MCGINNIS; MURROW; GARTNER, 2019; WOLOCK; LOPEZ; KLEIN, 2019). Em comparação com os dupletos homotípicos, os dupletos heterotípicos são considerados como tendo mais impacto nas análises posteriores, incluindo redução de dimensão, clusterização, expressão diferencial e trajetórias (BERNSTEIN *et al.*, 2020; XI; LI, 2021). Para reduzir o número de dupletos em experimentos, a diminuição da concentração de células carregadas é uma medida de controle eficaz para obter uma taxa menor de dupletos, mas esta abordagem também reduz o número de células capturadas e aumenta dramaticamente o custo por amostra (BERNSTEIN *et al.*, 2020; ZHENG *et al.*, 2017).

1.1.3 Modelos *CDX*/*PDX*

Entre as novas abordagens que têm sido estudadas no desenvolvimento da medicina personalizada em câncer, os ensaios derivados de pacientes (*PDX* do inglês, *Patient-Derived Xenografts*) surgiram como uma das mais promissoras. O *PDX* é o modelo pré-clínico que representa com maior fidelidade a individualidade de tumores humanos. Este modelo consiste na implantação de fragmentos frescos de tumores em camundongos imunodeficientes, permissivos ao crescimento do tumor. Os *PDXs* podem ser abordados para diferentes estratégias terapêuticas em função de uma assinatura gênica individualizada do tumor humano e os resultados podem informar quais estratégias podem ser clinicamente relevantes para o paciente. Há também o modelo de explantes derivados de células tumorais circulantes (*CDX*) que podem ser facilmente coletados em uma coleta de sangue. Este modelo não é invasivo para o paciente e as amostras podem ser coletadas independentemente do estágio da doença.

1.1.4 Abordagem Computacional para análise de dados de *scRNA-seq*

A Figura 2 descreve em forma de fluxograma a abordagem computacional realizada em três etapas a partir dos dados obtidos pelos experimentos de *scRNA-seq*. A primeira etapa (representada pela cor verde) é executada para qualquer dado de sequenciamento *high throughput*, onde é realizado o controle de qualidade das *reads* obtidas, alinhamento e mapeamento contra o genoma de referência e o controle de qualidade do mapeamento. Para esta etapa há uma ferramenta mais conhecida que é comumente utilizada chamada *CellRanger* ⟨http://software.10xgenomics.com/single-cell/overview/welcome⟩. Esta ferramenta, que é utilizada apenas para dados gerados pela plataforma *Chromium da 10x Genomics*, realiza todas as etapas destacadas em verde (controle de qualidade das *reads*, alinhamento e controle de qualidade do mapeamento) com apenas um comando, gerando assim a matriz de contagem. Há outras ferramentas que também podem ser utilizadas, como é o caso do *M3K*, que é uma ferramenta que está sendo desenvolvida em parceria com pesquisadores da Universidade de Colônia na Alemanha, e foi utilizada neste trabalho para o processamento do dado proveniente do sequenciador. Maiores detalhes serão comentados no Capítulo 3.

A segunda etapa (representada pela cor azul) faz o uso de algumas técnicas já utilizadas para a análise de *RNA-seq* convencional, mas também há novas ferramentas específicas para dados de *scRNA-seq*. Nesta etapa é realizado controle de qualidade das células e a normalização dos dados.

Por fim, a terceira etapa (representada pela cor vermelha) compreende expressão diferencial, clusterização, identificação dos tipos celulares etc. Esta etapa tem como principal objetivo extrair informações relevantes dos dados (STEGLE; TEICHMANN; MARIONI, 2015).

Para as etapas azuis (controle de qualidade das células e normalização) e vermelhas (expressão diferencial, clusterização e identificação dos tipos celulares) há diversas opções em diferentes linguagens de programação, sendo o conhecimento do usuário o fator determinante para a escolha da ferramenta a ser usada. Em *Python* há o módulo *SCANPY* (WOLF; ANGERER; THEIS, 2018) e em R há o pacote *Seurat* (STUART *et al.*, 2019).

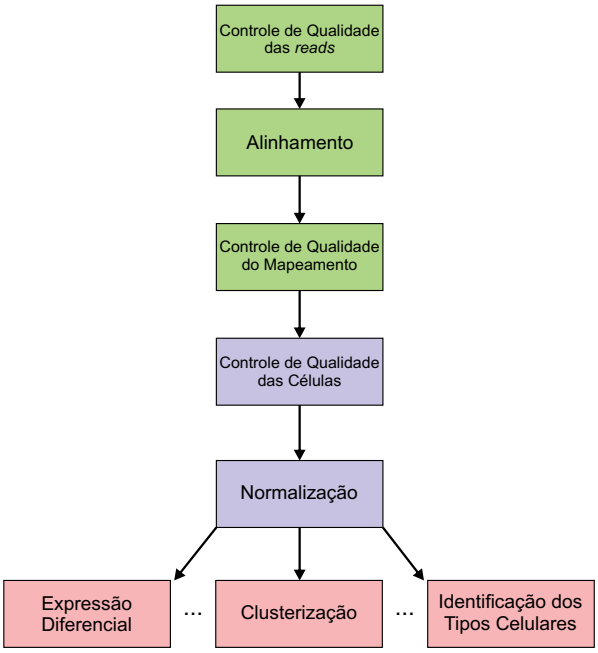

Figura 2 – *Pipeline* da análise computacional de *scRNA-seq* iniciando-se com o controle de qualidade das *reads*, alinhamento e controle de qualidade do mapeamento. Em seguida é feito o controle de qualidade das células e normalização. Por fim, é realizada a expressão diferencial, clusterização, identificação dos tipos celulares etc.

O número de ferramentas existentes para análises de dados de *scRNA-seq* cresce exponencialmente como pode ser observado na Figura 3. Zappia, Phipson e Oshlack (2018) discutem que novas ferramentas continuarão a ser produzidas, tornando-se cada vez mais sofisticadas para explorar os dados de *scRNA-seq*. Como a tecnologia de captura e sequenciamento de células únicas continuam a melhorar, as ferramentas de análise terão que se adaptar a um número maior de dados (milhões de células), o que pode exigir estruturas e algoritmos mais especializados (ZAPPIA; THEIS, 2021). Luecken e Theis (2019) discorrem sobre a importância de ter boas práticas para analisar dados de células únicas, discutindo todas as possibilidades de análises e sugerindo ferramentas para cada etapa.

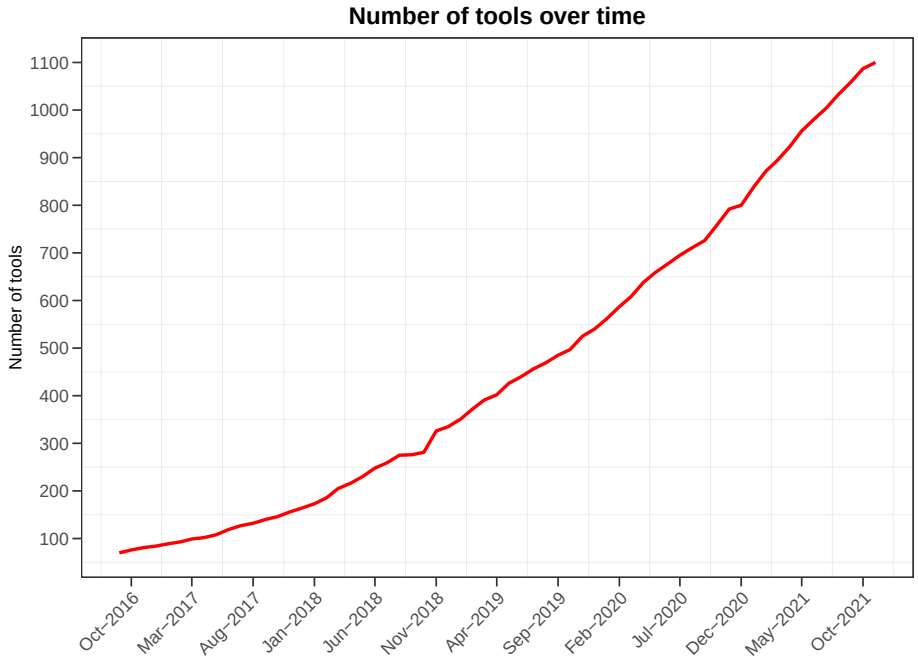

Figura 3 – Crescimento da quantidade de ferramentas para análise de *scRNA-seq* (eixo y) existentes ao decorrer do tempo (eixo x). Em 2021 foi atingida a marca de 1100 ferramentas desenvolvidas para análise de dados de *scRNA-seq*.

1.1.5 Identificação de subpopulações raras

A progressão tumoral é um processo que ocorre por evolução clonal e, portanto, apenas poucas células são suficientes para que um novo fenótipo surja gradualmente por acúmulo de mutações. Consequentemente, subpopulações críticas, como células-tronco tumorais, células resistentes a drogas e células migratórias subjacentes à metástase, são conjuntos de poucas células (por exemplo, menos de 0,1%) que não são detectadas pela maioria dos protocolos de *NGS*, incluindo o de *scRNA-seq*. Inicialmente, os estudos de *scRNA-seq* eram realizados utilizando plataformas que mensuravam um número limitado de células (dezenas a centenas). No entanto, mais recentemente com a entrada da tecnologia de microfluídos de gotículas (*droplet*), entre outras tecnologias (CAO *et al.*, 2017; GIERAHN *et al.*, 2017; JERBY-ARNON *et al.*, 2018), foi possível sequenciar um grande número de células (na casa dos milhares). Além disso, o desenvolvimento de novas tecnologias tendem a aumentar ainda mais essa capacidade em dezenas de milhares ou até mesmo milhões de células, o que ajudaria na descoberta de subpopulações muito raras.

Após a identificação de subpopulações únicas, sua importância funcional e clínica continuam sendo difíceis de ser investigadas. Duas abordagens principais permitem uma análise mais aprofundada destas subpopulações. Primeiro, estudos de grandes coortes podem revelar associações estatísticas entre a presença e/ou frequência de subpopulações e características clínicas, como sobrevida, metástase e respostas a medicamentos. Uma segunda abordagem depende da caracterização detalhada de tais subpopulações por *scRNA-seq* para identificar assinaturas gênicas que permitam o isolamento dessas subpopulações e estudos funcionais adicionais em modelos animais e de cultura (SUVÀ; TIROSH, 2019).

Vale ressaltar que o estudo de *scRNA-seq* de um tumor não é capaz de identificar a célula na qual a mutação ocorreu primeiro ("célula de mutação"), que pode ser distinta da célula que dá origem ao tumor.

Uma das vantagens de utilizar a técnica de *scRNA-seq* é a possibilidade de identificar novos transcritos e transcritos pouco expressos ou considerados raros. Os *RNAs* não codificadores (*ncRNAs*) são exemplos de transcritos normalmente pouco expressos, e exercem um papel importante na regulação de todos os processos celulares.

Os tumores são doenças complexas e heterogêneas que são compostas por diferentes populações de células com perfis moleculares e respostas a tratamentos distintos. A maioria das pesquisas básicas são realizadas *in vitro*, a partir do cultivo isolado de células em monoculturas em placas de cultura plásticas, sob condições que não refletem com precisão o microambiente tumoral apropriado. O estudo do transcriptoma de células únicas possibilita melhor avaliação *in vitro* do comportamento *in vivo* do tumor. Usando esta abordagem é possível conhecer a heterogeneidade tumoral, e dessa forma, avançar na compreensão da biologia dos tumores. A seguir serão apresentados dois tipos tumorais, o Melanoma e o Câncer de Pulmão de Pequenas Células (CPPC), onde serão utilizadas diferentes abordagens para estudar o papel do microambiente intratumoral na progressão tumoral.

1.2 Melanoma

O câncer de pele é o mais comum, sendo o melanoma, o carcinoma basocelular e o carcinoma espinocelular os tipos mais frequentes (KOH; GELLER; LEW, 2021). O câncer de pele pode ser classificado de duas formas, como melanoma e não melanoma. O melanoma é o 19º câncer mais frequente no mundo, com prevalência de 3,3 casos por 100

mil habitantes. Em 2012, foi estimado 232.130 novos casos, sendo que, entre os homens, foram 120.649 casos novos (3,4/100 mil) ocupando a 15ª posição; e 111.481 casos novos (3,2/100 mil) em mulheres, atingindo a 16ª posição entre todos os cânceres (FERLAY *et al.*, 2015). Ainda em 2012, um total de 55.488 óbitos puderam ser atribuídos diretamente ao melanoma, sendo que para os homens o número de óbitos por câncer de pele melanoma foi de 31.390; e para as mulheres, 24.098 (FERLAY *et al.*, 2015). As maiores incidências são observadas em países como Austrália e Nova Zelândia onde há uma predominância da cor de pele mais clara (WILD; STEWART; WILD, 2014).

No Brasil, em 2015, ocorreram 1.012 óbitos por câncer de pele melanoma em homens e 782 em mulheres (SANTOS; SOUZA, 2019). O principal fator de risco para os cânceres de pele melanoma e não melanoma é a exposição excessiva à radiação solar ultravioleta (UV). Outros fatores, como cor de pele, olhos e cabelos claros, histórico familiar ou pessoal de câncer de pele, podem aumentar o risco desenvolver câncer de pele (KOH; GELLER; LEW, 2021).

Para o desenvolvimento do melanoma é necessário que ocorra etapas específicas que envolvem fatores ambientais e genéticos. O melanoma tem origem nos melanócitos, que são embriologicamente derivados da crista neural. Normalmente os melanócitos situam-se na membrana basal e nos folículos capilares, próximos aos queratinócitos, que possuem o comportamento e crescimento direcionados através de um sistema de fatores de crescimento e moléculas de adesão celular (HAASS *et al.*, 2005). O processo de transformação dos melanócitos em melanoma ocorre de forma gradual. Inicia-se no começo da proliferação que leva ao desenvolvimento de nevos benignos, que representa a lesão melanocítica hiperplásica mais precoce, seguida de crescimento desenfreado e displasia. O primeiro estágio maligno é a fase de crescimento radial (FCR), em que as células tumorais atingem a capacidade de proliferar no interior da epiderme. Na fase de crescimento vertical (FCV) as células tumorais adquirem a capacidade de crescer verticalmente e invadir a derme e o tecido subcutâneo. Por fim, as células tumorais adquirem a capacidade de metatizar, representando estágios mais avançados da tumorigênese.

Os eventos moleculares que levam à transformação de melanócitos normais em melanoma não são compreendidos totalmente, embora mutações em genes críticos para o crescimento e sobrevivência tenham sido associados à iniciação e progressão do melanoma. Mutações na proteína quinase *BRAF* são as mais comuns, correspondendo a 40% a 70% de incidência em pacientes com melanoma. A mutação mais comum no gene *BRAF* resulta de

um ácido glutâmico para substituição de valina na posição 600 (*V600E*) dentro do domínio quinase de proteína, levando a ativação constitutiva da via da *MAPK* e estimulação do crescimento, sobrevivência e angiogênese.

Dentro da pele, os queratinócitos exercem a função de proliferação de melanócitos através de fatores de crescimento parácrinos e comunicação intercelular via moléculas de adesão celular. A proliferação desregulada ocorre quando os melanócitos escapam ao controle imposto pelos queratinócitos por meio da regulação negativa das moléculas de adesão celular, como *E-caderina*, *P-caderina*, desmogleína e conexinas (HAASS *et al.*, 2005; LI; FUKUNAGA; HERLYN, 2004). As caderinas (*E-caderina*, *P-caderina* e desmogleína) são uma família de proteínas transmembrana que promovem a adesão celular enquanto que as conexinas são moléculas-chave para a comunicação celular direta e também são consideradas importantes para a liberação de moléculas sinalizadoras das células para o microambiente (BRANDNER; HAASS, 2013).

A superexpressão da *E-caderina* em células de melanoma restaura o controle dos queratinócitos sobre a proliferação e impede a invasão, enquanto a adesão célula-célula contribui para o controle da proliferação e supressão tumoral mediada por queratinócitos. A *E-caderina* também pode prevenir a progressão do tumor pela regulação negativa da sinalização mediada pela β-catenina (GOTTARDI; WONG; GUMBINER, 2001; LI; SATYAMOORTHY; HERLYN, 2001), pois a β-catenina promove a proliferação celular induzindo a transcrição de genes reguladores de crescimento e sobrevivência, tais como c-*Myc*, ciclina D1 e *MITF* (LARUE; DELMAS, 2006; WIDLUND *et al.*, 2002).

A *E-caderina* pode ser regulada negativamente por vários mecanismos. Os repressores *SLUG* e *SNAI1* inibem a expressão da *E-caderina* ao nível da transcrição em células de melanoma (BOLÓS *et al.*, 2003; CONACCI-SORRELL *et al.*, 2003; POSER *et al.*, 2001). Durante a progressão do melanoma, a secreção autócrina do fator de crescimento de hepatócitos/fator de células tronco também promove a regulação negativa da *E-caderina* e desmogleína I (LI; SATYAMOORTHY; HERLYN, 2001).

Concomitantemente com a perda de *E-caderina* e a progressão do melanoma para um fenótipo invasivo, alterações na expressão de integrinas promovem o desprendimento das células do melanoma do local primário através do estroma contíguo e eventualmente se disseminam através de vasos linfáticos ou vasculares para órgãos distantes. As integrinas são proteínas transmembranas heterodiméricas que não apenas atuam na adesão celular à

matriz extracelular (MEC), mas também regulam processos importantes como: proliferação celular, migração, invasão, angiogênese e sobrevivência da célula (MOSCHOS *et al.*, 2007).

O microambiente tumoral é um participante ativo na tumorigênese. Além das células tumorais há as células estromais, que possuem grande importância. O estroma tumoral é complexo e inclui a MEC, fatores de crescimento e citocinas, a microvasculatura, células inflamatórias infiltrantes e fibroblastos (RUITER *et al.*, 2002). No estroma do melanoma decorrem alguns processos como a proteólise de colágeno e elastina na borda invasiva dos tumores, bem como infiltração de linfócitos e angiogênese variável (LABROUSSE *et al.*, 2004).

As células do melanoma interagem ativamente células do estroma. Durante os estágios invasivos iniciais, as células do melanoma precisam ativar um mecanismo que lhes permita migrar, invadir e sobreviver fora de seu nicho original sob novas condições microambientais e estabelecer residência com sucesso em um novo local.

Os melanomas também secretam *TGF-*β. A sinalização de *TGF-*β desempenha um papel importante na tumorigênese e na metástase, exercendo efeitos diretos ou indiretos sobre a própria célula do tumor ou sobre o microambiente do tumor (BIERIE; MOSES, 2006).

Embora o melanoma seja às vezes considerado um processo celular autônomo que resulta apenas de alterações genéticas e epigenéticas nas células transformadas, muitas etapas do processo de transformação (proliferação, invasão, angiogênese e metástase) são moduladas por fatores de crescimento e enzimas proteolíticas produzidas por células estromais. É a interação dinâmica entre células tumorais e estromais que determina o resultado do processo de transformação. Além disso, as células estromais participam dos mecanismos de evasão imunológica do câncer. As células estromais do tumor e seus produtos são alvos promissores para a terapia do câncer.

Embora muitos estudos tenham levado a um melhor entendimento da biologia e genética do melanoma, não há tratamento efetivo disponível atualmente. O tratamento mais eficaz continua sendo a detecção precoce e a ressecção cirúrgica, considerando que é uma doença metastática altamente refratária ao tratamento e a taxa de sobrevida em 5 anos permanece em 15% (TAWBI; KIRKWOOD, 2007).

1.3 Câncer de Pulmão de Pequenas Células (CPPC)

O câncer de pulmão é a principal causa de mortalidade por câncer em todo o mundo, com uma estimativa de 2,2 milhões de novos casos e 1,8 milhões de mortes em 2020 (SUNG *et al.*, 2021). O CPPC compreende uma estimativa de 250.000 novos casos e pelo menos 200.000 mortes no mundo todo a cada ano (MATSUDA; OKUYAMA, 2018). O câncer de pulmão, incluindo todos os subtipos histológicos, é mais prevalente em países/regiões de alta renda, refletindo os níveis de consumo de tabaco (SUNG *et al.*, 2021). Entretanto, a incidência específica do CPPC em diferentes países/regiões ou continentes não está bem descrita. Como no câncer de pulmão em geral, o CPPC é mais prevalente nos homens, mas a proporção de casos em mulheres em comparação com homens aumentou no mundo inteiro nos últimos 50 anos, refletindo novamente as tendências de consumo de tabaco (GOVINDAN *et al.*, 2006).

O CPPC é um carcinoma neuroendócrino de alto grau que surge predominantemente em fumantes ou ex-fumantes e tem um prognóstico excepcionalmente ruim (HELLMAN; ROSENBERG, 2001). O CPPC representa cerca de 15% dos casos de câncer de pulmão. Os pacientes com CPPC normalmente apresentam sintomas respiratórios, incluindo tosse, dispneia (respiração laborativa) ou hemoptise (tosse de sangue). Os locais mais comuns de metástase incluem o pulmão contralateral, o cérebro, o fígado, as glândulas suprarrenais e o osso. Espelhando sua alta predileção metastática, a concentração de células tumorais circulantes (CTCs) em CPPC está entre as mais altas de qualquer tumor sólido (HOU *et al.*, 2012).

Nos raros pacientes que apresentam doença em estágio muito precoce no diagnóstico, o tratamento pode incluir cirurgia e quimioterapia adjuvante à base de platina, embora, mais tipicamente, pacientes com doença em estágio precoce ou localmente avançada sejam tratados com radiação concorrente e quimioterapia à base de platina. Pacientes com doença metastática são tratados com quimioterapia sistêmica com ou sem imunoterapia.

O CPPC inicialmente responde excepcionalmente às terapias citotóxicas - até 25% dos pacientes com CPPC em estágio inicial alcançam o controle a longo prazo da doença com quimioradioterapia concomitante e as taxas de resposta são consistentemente superiores a 60%, mesmo em pacientes com doença metastática. Entretanto, na grande maioria dos pacientes, estas respostas são transitórias, resultando em uma duração média

de sobrevivência de < 2 anos para pacientes com doença em estágio inicial e de ≈ 1 ano para pacientes com doença metastática.

1.3.1 Via de sinalização *NOTCH*

No trabalho de George *et al.* (2015) foram sequenciados 110 genomas completos de pacientes a fim de entender mais sobre as alterações genômicas nestes casos. *TP53* e *RB1* estão mutados em praticamente todas as amostras. É interessante notar que a perda biológica de ambos os genes é necessária para o aparecimento desta doença. Além disso, foram identificadas muitas outras alterações (vias de sinalização) que são afetadas no CPPC, como: regulação do ciclo celular, receptor quinase/sinalização *PI3K*, regulação transcricional e sinalização *Notch*/diferenciação neuroendócrina.

Foram realizados estudos funcionais que mostraram que o perfil de mutação geral é bastante prejudicial e é principalmente enriquecido em *NOTCH1*, que possui um papel supressor de tumores. Como pode-se observar na Figura 4, o *NOTCH* é necessário para regular o *Has1* e o *Hey1*, eles contra-atacam a expressão neuroendócrina, portanto, quando temos mutações inativadoras ou a desregulação da *NOTCH*, temos uma expressão alta de marcadores neuroendócrinos que marcam esses tumores. O *DLL3*, que está downregulado, possui um papel interessante por causa de alvos terapêuticos que vão contra o receptor de superfície celular e têm sido acompanhados em vários ensaios clínicos nos últimos anos (RUDIN *et al.*, 2017; HIPP *et al.*, 2020; GIFFIN *et al.*, 2021). A expressão dos marcadores neuroendócrinos é sempre regulada pelo *Ascl1*, um dos fatores de transcrição importantes no CPPC. Desta forma, foi mostrado que em CPPC temos uma downregulação desta via, por isso temos um tumor neuroendócrino e uma alta expressão de marcadores neuroendócrinos (GEORGE *et al.*, 2015).

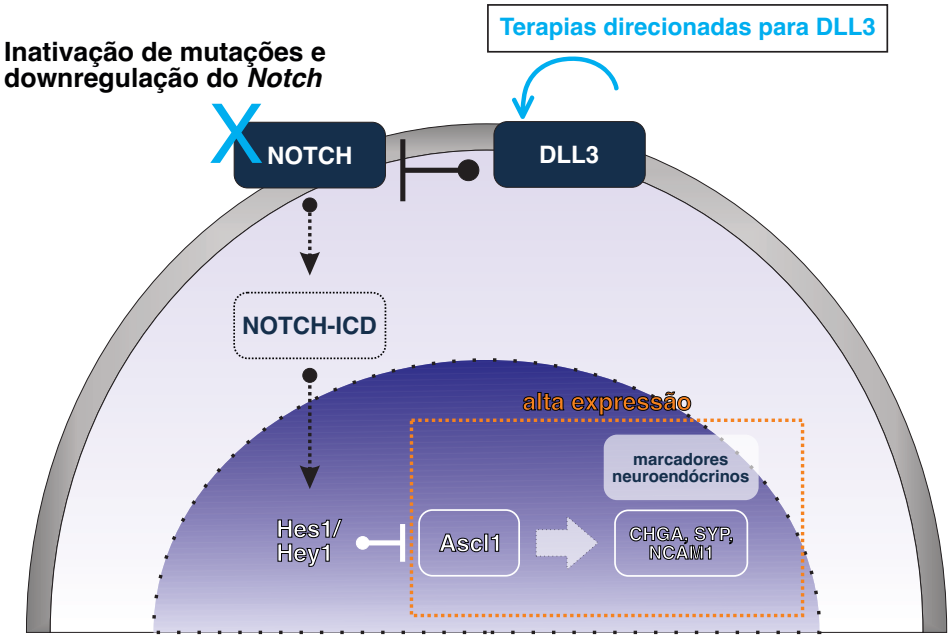

Figura 4 – Via de sinalização *NOTCH* no CPPC.

1.3.2 Subtipos moleculares

O CPPC possui uma malignidade excepcionalmente letal para a qual são urgentemente necessárias terapias mais eficazes. Durante o tempo, mais especificamente nos últimos 10 anos com o avanço das ômicas, o CPPC vem passando por mudanças em sua classificação de subtipos moleculares. Na Figura 5 podemos observar todas as classificações já consideradas para o CPPC, sendo a mais atual (2019) definida a partir da expressão de 2 fatores de transcrição (*TFs*) neuroendócrinos (*NEUROD1* e *ASCL1*) e 2 TFs não neuroendócrinos (*POU2F3* e *YAP1*).

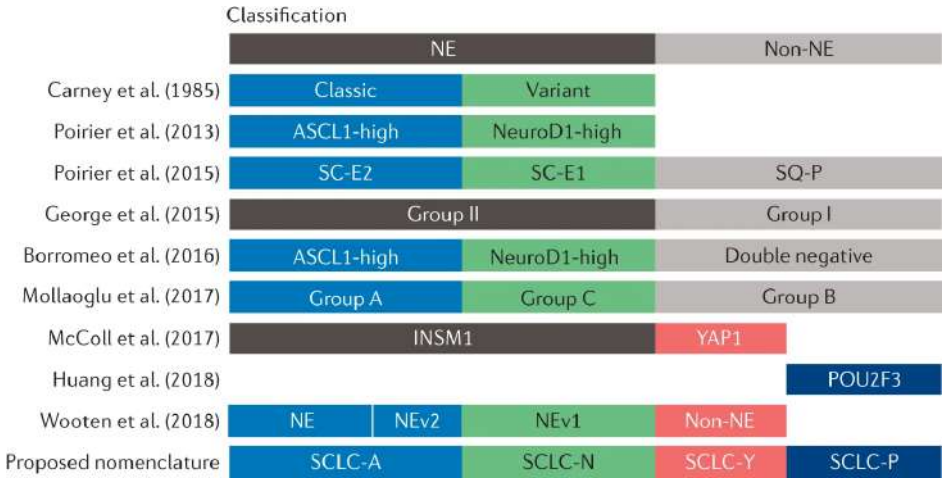

Figura 5 – Classificações dos subtipos moleculares do CPPC durante o tempo, sendo a dos 4 *TFs* (*ASCL1*, *NEUROD1*, *YAP1* e *POU2F3*) a mais atual.

Fonte – Nomenclatura e figura retirada de Rudin *et al.* (2019)

Estes subtipos moleculares foram definidos a partir da expressão diferencial dos reguladores discriminantes de transcrição, distinguindo claramente estes subtipos em ambas as linhagens celulares e em tumores humanos, como pode ser observado na Figura 6. A definição dos subtipos moleculares do CPPC fornece novas alternativas para o desenvolvimento de medicamentos. Como novas opções terapêuticas são avaliadas prospectivamente dentro do contexto de subtipos identificados de CPPC, existe uma ótima oportunidade para definir melhor os biomarcadores preditivos. Por exemplo, a expressão de *POU2F3* pode ser interessante na identificação de tumores suscetíveis à inibição de *PARP* (KNELSON; PATEL; SANDS, 2021).

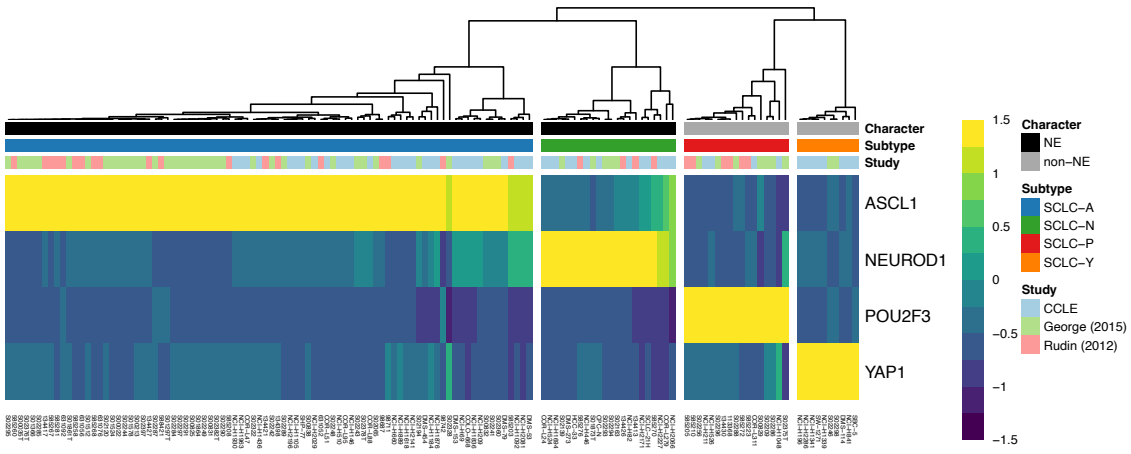

Figura 6 – Clusterização hierárquica da expressão gênica relativa dos 4 reguladores-chave de transcrição (A = *ASCL1*, N = *NEUROD1*, P = *POU2F3* e Y = *YAP1*) que definem os subtipos no CPPC.

Fonte – Figura retirada de Rudin *et al.* (2019)

1.4 O microambiente tumoral

Um tumor não é apenas um grupo de células cancerígenas, mas sim um conjunto heterogêneo de células hospedeiras infiltrantes e residentes, fatores secretados e matriz extracelular. As células tumorais estimulam mudanças moleculares, celulares e físicas significativas dentro de seus tecidos hospedeiros para apoiar o crescimento e a progressão do tumor e a resistência ao tratamento (ANDERSON; SIMON, 2020).

A composição do microambiente tumoral varia de acordo com o tipo de tumor, mas as características marcantes incluem células imunes, células estromais, vasos sanguíneos e matriz extracelular (Figura 7). Acredita-se que o "microambiente tumoral não é apenas um espectador silencioso, mas um promotor ativo da progressão do câncer"(TRUFFI; SORRENTINO; CORSI, 2020).

No início do crescimento tumoral, desenvolve-se uma relação dinâmica e recíproca entre as células cancerígenas e os componentes do microambiente tumoral que suporta a sobrevivência das células cancerígenas, a invasão local e a disseminação metastática. Os tumores se infiltram com diversas células imunes adaptativas e inatas que podem desempenhar tanto funções pró e anti tumorigênicas. O microambiente tumoral vem sendo relatado de suma importância para a identificação de novos alvos para intervenção terapêutica (XIAO; YU, 2021).

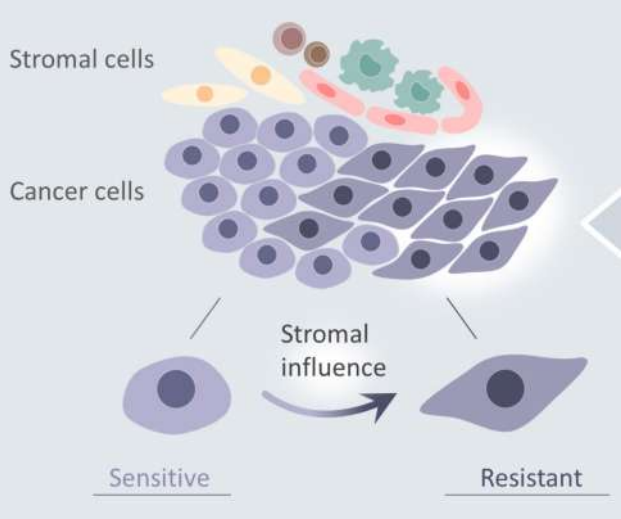

Figura 7 – Microambiente tumoral, que é um participante ativo na tumorigênces e promissor para encontrar novos biomarcadores, é composto por células estromais, células cancerosas, etc.

Fonte – Figura retirada e adaptada de ⟨ous-research.no/home/prasmickaite/home/13609⟩

1.5 Os fatores de transcrição (TFs)

Os *TFs* são proteínas envolvidas no processo de conversão, ou transcrição, do *DNA* em *RNA*. Os *TFs* incluem um grande número de proteínas, excluindo a *RNA* polimerase, que iniciam e regulam a transcrição de genes. Uma característica distinta dos *TFs* é que eles têm domínios de ligação de *DNA* que lhes dão a capacidade de se ligar a sequências específicas de *DNA* chamadas sequências de realçador ou promotor. Alguns *TFs* ligam-se a uma sequência promotora de *DNA* próxima ao local de início da transcrição e ajudam a formar o complexo de iniciação da transcrição. Outros *TFs* ligam-se a sequências reguladoras, tais como sequências de realçadores, e podem estimular ou reprimir a transcrição do gene relacionado. Estas sequências reguladoras podem ser milhares de pares de bases a montante ou a jusante do gene que está sendo transcrito. A regulamentação da transcrição é a forma mais comum de controle do gene. A ação dos *TFs* permite a expressão única de cada gene em diferentes tipos de células e durante o desenvolvimento (LAMBERT *et al.*, 2018).

1.6 Os RNAs não codificantes

O corpo humano é composto por aproximadamente 100 trilhões de células, organizadas em tecidos e órgãos molecularmente distintos que surgem através do processo de diferenciação. Com o avanço da tecnologia de análise genômica de células únicas, ficou evidente que as populações de células somáticas (normais e tumorais) possuíam uma grande variação molecular. A capacidade de avaliar transcriptomas a partir de células individuais levou a um maior número de tipos celulares identificados que possuem variação significativa de célula para célula, mesmo dentro de linhagens celulares clonais. Tal fato pode ser explicado se considerarmos duas visões distintas (GAWRONSKI; KIM, 2017):

1. **Funcionalmente Neutra:** podem existir múltiplos estados moleculares "equifenotípicos" que podem existir porque diferentes combinações de expressão gênica podem levar ao mesmo resultado (KIM; EBERWINE, 2010);
2. **Funcionalmente Adaptativa:** existem várias razões adaptativas possíveis para a variabilidade célula a célula, incluindo perturbações ambientais, composição com-

binatória para função ao nível dos tecidos, interações ecológicas e sinalizadoras multicelulares etc (DUECK; EBERWINE; KIM, 2016).

Ambas as visões citadas acima, a respeito da variação de célula única, não são mutuamente exclusivas. O entendimento dessas duas visões para a variabilidade celular é muito importante pois é o problema chave na biologia das células únicas, com efeito sobre a transdiferenciação e medicina regenerativa. Com isso, uma abordagem possível para identificar a variabilidade funcional de células únicas pode estar no estudo da variabilidade de ncRNAs, pois tendem a estar predominantemente envolvidos na função reguladora do genoma (MORRIS; MATTICK, 2014).

1.6.1 Os *RNAs* longos não codificantes

Os *RNAs* longos não codificantes (*lncRNAs*) são moléculas que possuem ao menos 200 nucleotídeos e geralmente não codificam proteínas. Eles exercem funções muito importantes como o *imprinting*, tradução, *splicing*, diferenciação celular e controle do ciclo celular, porém a relevância biológica da grande maioria permanece incerta. Sua complexidade depende não apenas da mudança funcional, mas também da sua capacidade de ser tecido/célula específico (CALLE *et al.*, 2018). Como os *lncRNAs* estão envolvidos com a regulação da expressão gênica e ciclo celular, eles podem desempenhar um papel muito importante no desenvolvimento de tumores e outras doenças. Uma das principais características dos *lncRNAs* é o seu padrão de expressão específica de tecidos e tipos celulares (LORENZEN; THUM, 2016; WAPINSKI; CHANG, 2011).

Os *locis* associados com o surgimento e progressão do câncer transcrevem genes codificadores e não codificadores. que são comumente desregulados no câncer (CALIN *et al.*, 2007). Por exemplo, alguns *SNPs* estão entre as alterações de alto risco associadas à ocorrência de câncer; curiosamente, 85% dos *SNPs* são anotados em regiões reguladoras (*enhancers* e promotor) ou de genes não codificadores ligadas ao desenvolvimento de doenças (FREEDMAN *et al.*, 2011). Essas alterações têm impacto sobre a função e regulação dos *lncRNAs*, que exibem expressão alterada e afetam a regulação de seus alvos.

Um dos desafios de se trabalhar com os dados de células únicas e, mais especificamente, com *lncRNAs*, é o fato de possuírem níveis de expressão menores que os genes codificadores de proteínas (CABILI *et al.*, 2011; GUTTMAN *et al.*, 2010; IYER *et al.*,

2015; DERRIEN *et al.*, 2012). Uma das explicações pode ser pelo fato de que os *lncRNAs* demonstram um padrão de expressão tecido-específico (CABILI *et al.*, 2015), na maioria das vezes restrito a uma subpopulação específica.

Sabendo-se que os *lncRNAs* são expressos em tipos celulares específicos, usando dados de *scRNA-seq* é possível identificar grupos de *lncRNAs* enriquecidos em determinados tipos celulares, enquanto que outros estão ausentes ou em níveis muito baixos de expressão (LIU *et al.*, 2016). A análise desses transcritos é de extrema importância, pois pode contribuir para definir o grau de heterogeneidade de um determinado tumor, ajudando na priorização de populações celulares como alvos terapêuticos (LV *et al.*, 2016; SHALEK; BENSON, 2017). Além disso é possível, a partir do dado bruto, predizer e identificar novos *lncRNAs* e tipos de *lncRNAs*.

Os *lncRNAs* possuem um papel muito importante na área da oncologia. Análises recentes têm demonstrado que tumores classificados histologicamente como de um mesmo tipo são altamente heterogêneos. Essa heterogeneidade ocorre em dois níveis: tanto inter-tumoral, ou seja, entre tumores de diferentes indivíduos, quanto intratumoral, refletindo a presença de diferentes subpopulações de células tumorais (perfil genético e transcricional) dentro de um mesmo tumor (MARUSYK; ALMENDRO; POLYAK, 2012). A heterogeneidade intratumoral, que tem sérias implicações no diagnóstico e no desenvolvimento de resistência terapêutica, vem sendo explorada em crescente detalhamento com as inovações no campo de sequenciamento de nova geração, especialmente com o avanço das técnicas de sequenciamento de células únicas ou individuais (NAVIN *et al.*, 2011; TIROSH *et al.*, 2016b; TIROSH *et al.*, 2016a). Embora a evolução clonal das células tumorais seja um fator determinante na heterogeneidade intratumoral, existem evidências de que fatores não-genéticos também contribuem para esse cenário. A perspectiva da presença e atuação de células tronco-tumorais, o que levaria a uma hierarquia de diferenciação (diferentes programas transcricionais), é considerada uma das fontes não-genéticas de heterogeneidade fenotípica dentro dos tumores (MARUSYK; ALMENDRO; POLYAK, 2012). Por definição, as células-tronco representam uma população rara e de alta plasticidade fenotípica (são células capazes de se autorrenovarem e se diferenciarem, além de se dividir e se transformar em diferentes tipos celulares).

Os *lncRNAs* podem ser localizados genomicamente entre dois genes codificadores de proteínas (*lncRNA* intergênico) (Figura 8A), transcritos de um promotor de um gene codificador de proteínas, ainda na direção oposta (*lncRNA* bidirecional) (Figura 8B),

originados da fita *antisense* do *RNA* de um gene codificador de proteína (*lncRNA antisense*) (Figura 8C), ou sobreposição com um ou mais *introns/exons* de diferentes genes codificadores de proteínas fina *sense* de *RNA* (*lncRNAs* com sobreposição de sentido) (Figura 8D) (BALAS; JOHNSON, 2018).

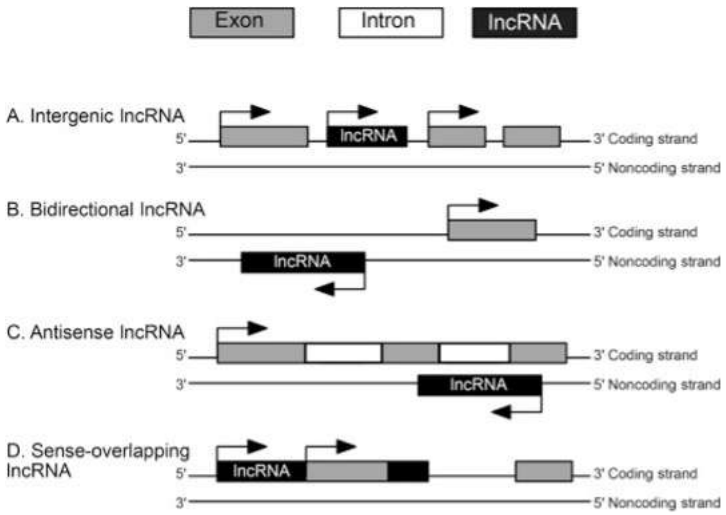

Figura 8 – Classificação dos *lncRNAs* baseada na localização genômica. A) Os *lncRNAs* intergênicos estão localizados entre genes codificadores de proteínas. B) Os *lncRNAs* bidirecionais são transcritos do mesmo promotor como um gene codificador de proteínas, mas na direção oposta. C) Os *lncRNAs antisense* originam-se da fita de *RNA antisense* de um gene codificador de proteína. D) Os *lncRNAs* com sobreposição de sentidos se sobrepõem a um ou mais *introns* e/ou *exons* de um gene codificador de proteína na direção *sense* da fita de *RNA*.

Fonte – Figura retirada de Balas e Johnson (2018)

Para ajudar na identificação de populações raras de células-tronco há uma subcategoria de *lncRNAs* que são os *RNAs* longos intergênicos não codificantes (*lincRNAs*), cujo *HOTAIR* faz parte.

1.6.2 *HOTAIR*

O *HOTAIR* desempenha um papel muito importante na pluripotência e diferenciação, influenciando a expressão gênica global e reprimindo programas de linhagens (GUTTMAN *et al.*, 2011).

Há estudos que mostram que o *HOTAIR* atua reprogramando os estados da cromatina para promover a metástase de câncer (GUPTA *et al.*, 2010), levando assim à supressão de vários genes. Geralmente ele está associado com metástase e mau diagnóstico

em diferentes tipos tumorais. Foi o primeiro *lncRNA* implicado na modulação do estado da cromatina em trans (BERGMANN; SPECTOR, 2014). Ele é transcrito a partir da fita *antisense* do gene *HoxC* e é capaz de interagir com o complexo repressivo policombe 2 (*PRC2*) – uma histona metiltransferase que gera silenciamento epigenético – guiando-o para diferentes sítios no genoma (TANG; HANN, 2018).

De acordo com Alves *et al.* (2013), em estudo realizado pelo nosso grupo, o *HOTAIR* funciona como um gatilho que ativa a via Transição Epitélio Mesênquima (*EMT*) e participa da manutenção das células tronco em linhagens celulares cancerosas. A superexpressão do *HOTAIR* promove alterações no perfil de expressão de genes que regulam a *EMT* e genes relacionados à *stemness* (Figura 9). No trabalho de Alves *et al.* (2013) foram analisados os resultados das células tumorais de mama *MDA-MB-231* expressando *HOTAIR*, para uma seleção de 35 genes (GUPTA *et al.*, 2010) relacionados com *EMT* e/ou *stemness*. O perfil de expressão gênica foi realizado em células que expressam ectopicamente *HOTAIR* com ou sem depleção concomitante de *PRC2* pelo uso de *shRNA* contra *SUZ12*.

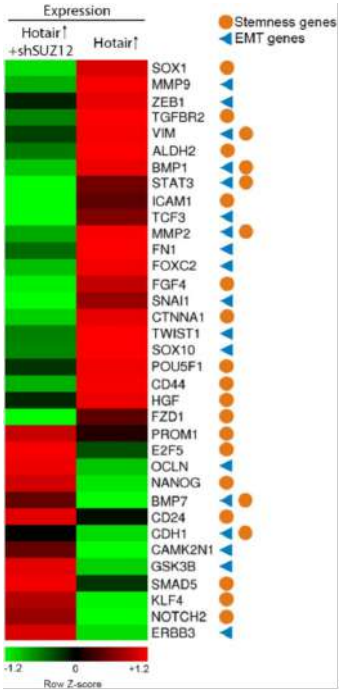

Figura 9 – Superexpressão do *HOTAIR* em células tumorais de mama *MDA-MB-231* expressando *HOTAIR*, para uma seleção de 35 genes relacionados com *EMT* e/ou *stemness*. O perfil de expressão gênica foi realizado em células que expressam ectopicamente *HOTAIR* com ou sem depleção concomitante de *PRC2* pelo uso de *shRNA* contra *SUZ12*.

Fonte – Figura retirada de Alves *et al.* (2013)

Além disso, o *HOTAIR* também interage com os complexos *PRC2* e *LSD1/coREST* através dos seus domínios de ligação 5' e 3', respectivamente (Figura 10). O *knockdown* do *HOTAIR* diminui a ocupação do complexo *LSD1* e, consequentemente, a uma perda de *H3K27* e o ganho de *H3K4me2/3* nos promotores mais próximos à região do gene *SNAI1* e, por fim, a metilação de *H3K27* e a desmetilação *H3K4me2/3* acaba influenciando o silenciamento epigenético do gene *CDH1* e, finalmente, a ativação da via de *EMT*. Essas alterações epigenéticas resultam na modificação e alteração de perfil de uma série de genes de regulação positiva ou negativa de proliferação, invasão, apoptose e migração de células cancerosas. De acordo com Tang *et al.* (2013), Wu *et al.* (2014), em melanoma o *HOTAIR* leva a uma superexpressão em tecido metastático e promove a motilidade e invasão celular.

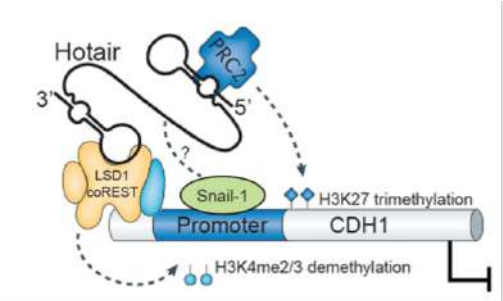

Figura 10 – Esquema de como o *HOTAIR* age em meio aos complexos *LSD1/coREST* e *PRC2* junto com o *SNAI1* a fim de inibir a expressão de *CDH1*.

Fonte – Figura retirada de Alves *et al.* (2013)

1.7 Transição epitélio-mesenquimal (EMT)

A *EMT* é o processo pelo qual as células epiteliais iniciam perdendo as suas junções e suas polaridades apico-basais, reorganizam seus citoesqueletos, sofrem mudanças nos seus programas de sinalização que definem seus formatos e reprogramam sua expressão gênica, aumentando a sua motilidade e desenvolvendo um fenótipo invasivo (LAMOUILLE; XU; DERYNCK, 2014). Esse processo simula o evento de gastrulação normal do desenvolvimento e pode ser classificado em três subtipos: o primeiro está associado com o desenvolvimento embrionário original e também ocorre durante o crescimento pós-natal; o segundo subtipo é iniciado em resposta a ferimentos e gera fibroblastos para reconstruir tecidos feridos; o terceiro é justamente o subtipo oncogênico, que recapitula características típicas da *EMT* do desenvolvimento, porém de forma menos ordenada e coordenada (SAMATOV; TONEVITSKY; SCHUMACHER, 2013).

A perda da expressão de *E-caderina* é considerado um evento fundamental da *EMT*, assim como a perda da polaridade celular (THIERY *et al.*, 2009). A mudança na expressão de genes relacionados às junções aderentes previne que haja nova formação de junções célula-célula e resulte na perda da função da barreira epitelial (LAMOUILLE; XU; DERYNCK, 2014). A reorganização do citoesqueleto, por sua vez, permite que a célula adquira uma motilidade direcional, gerando projeções de membrana que facilitam o movimento celular e que possuem função proteolítica, o que leva à degradação da matriz extracelular e facilita a invasão (LAMOUILLE; XU; DERYNCK, 2014).

A ativação do programa *EMT* nas células cancerosas sinaliza o início de processos invasivos e antiapoptóticos que levam à metástase; gerando células estromais ativadas que levam à progressão do câncer por alterações bioquímicas e estruturais no microambiente tumoral; e estimulando uma maior malignidade associada ao fenótipo de célula-tronco cancerosa (NISTICÒ; BISSELL; RADISKY, 2012)(BISSELL; RADISKY; NISTICO, 2016). A regulação da *EMT* conta ainda com a ativação de diversos fatores de transcrição, dentre os quais o *ZEB1*, *ZEB2*, *Snail*, *Slug* e *Twist* são os mais bem caracterizados (SAMATOV; TONEVITSKY; SCHUMACHER, 2013).

Um fato interessante, é que os tumores metastáticos podem apresentar um aumento na expressão de *E-caderina*, quando comparada com a expressão aberrante ou perda da sua expressão nos tumores primários, levando-os a se assemelhar ao fenótipo epitelial do tumor de origem (CHAO; SHEPARD; WELLS, 2010). Tal característica deve-se em grande parte ao processo inverso da *EMT*, a transição mesenquimal-epitelial (*MET*), que é crítica para os últimos estágios da cascata metastática (YAO; DAI; PENG, 2011).

2 Objetivos

2.1 Objetivo Geral

Avaliação integrada da plasticidade das células e do microambiente tumoral em câncer de melanoma e câncer de pulmão de pequenas células (CPPC) em nível de células únicas.

2.2 Objetivos específicos

2.2.1 Para dados de Melanoma

- Identificar populações de células enriquecidas com as assinaturas de EMT;
- Identificação de circuitos gênicos envolvidos na ativação da via de EMT;

2.2.2 Para dados de Câncer de Pulmão de Pequenas Células (CPPC)

- Avaliar o perfil transcricional das células únicas;
- Verificar subtipos moleculares e marcadores neuroendócrinos;
- Verificar a heterogeneidade intratumoral;
- Identificar mecanismos e vias envolvidas na progressão tumoral.

3 Materiais e Métodos

3.1 Melanoma

Para as análises do dado de melanoma foi utilizado um *pipeline* metodológico a seguir mostrado na Figura 11. Inicialmente o dado em *TPM* é convertido em contagem (*counts*). A partir da matriz de contagem é utilizado o *Seurat* (BUTLER *et al.*, 2018; STUART *et al.*, 2019) para uma análise mais geral e abrangente do dado, seguido pelo *MAGIC* (DIJK *et al.*, 2018), onde observa-se o comportamento de até três diferentes genes a partir da imputação de dados. O enriquecimento das células para assinaturas específicas foi realizado a partir do *GSVA* (HÄNZELMANN; CASTELO; GUINNEY, 2013) e *AUCell* (AIBAR *et al.*, 2017), seguido pela inferência de trajetória utilizando-se a ferramenta *STREAM* (CHEN *et al.*, 2019). Por fim, foi utilizado o pacote *CeTF* (BIAGI *et al.*, 2021) para identificar os TFs que possuem papéis importantes entre as células malignas e não malignas.

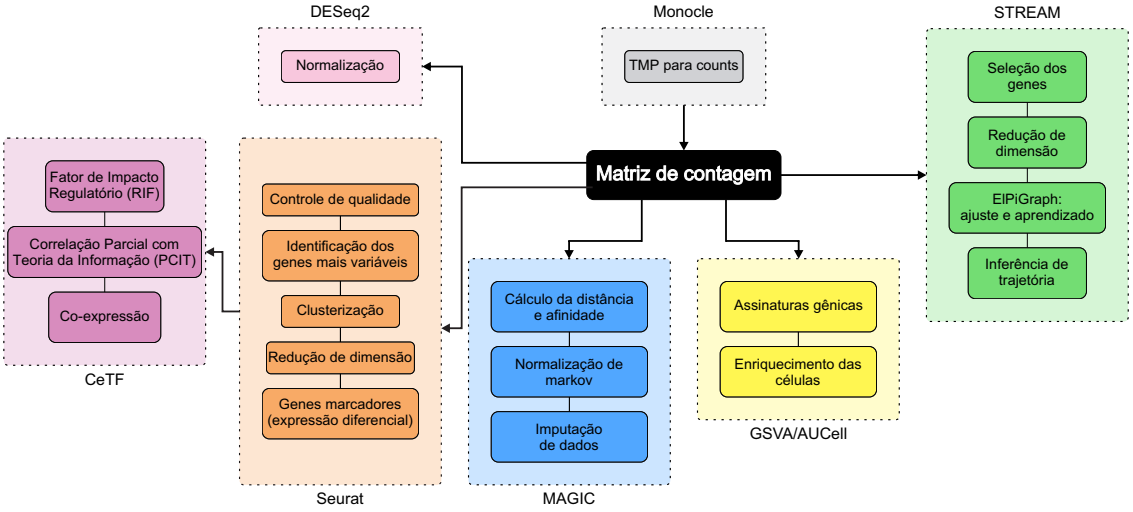

Figura 11 – *Workflow* dos métodos utilizados neste trabalho, passando pelas ferramentas *Seurat*, *MAGIC*, *GSVA*, *AUCell*, *STREAM* e *CeTF*.

3.1.1 Infraestrutura computacional

Os dados de melanoma foram analisados utilizando a estrutura do Equipamento MultiUsuário (EMU) localizado no Laboratório de Genética Molecular e Bioinformática (LGMB) da Faculdade de Medicina de Ribeirão Preto. A EMU é um sistema de lotes portátil (*Portable Batch System - PBS*), que permite o escalonamento dos processos,

o *PBS* distribui os processos entre os diversos recursos computacionais disponíveis. A EMU conta com 11 computadores, todos com o sistema operacional *Linux*, utilizando a distribuição *CentOS Linux* versão 7.6.1810 em todos os nós deste *cluster*, onde 10 dos 11 computadores possui processadores *AMD Opteron* 6174, com um total de 20 *cores* e $64GB$ de memória *RAM* em cada nó, o único nó que possui um hardware diferente possui processadores Intel *Xeon E7540*, com um total de 20 cores e $120GB$ de memória *RAM*, totalizando 220 *cores* e $760GB$ de *RAM*. Para o armazenamento, a EMU conta com três *storages* sendo dois deles de $20TB$ e um de $32TB$ de armazenamento, totalizando $72TB$ de espaço.

3.1.2 Dado de *scRNA-seq* de Melanoma

Os dados de melanoma de *scRNA-seq* são provenientes de Tirosh *et al.* (2016a) que estão disponíveis nos bancos de dados do *GEO* ou *ArrayExpress* através do número de acesso *GSE72056* ou *E-GEOD-72056*, respectivamente. O dado original possui 23.684 genes e 4.645 células, sendo 3.256 células não malignas, 1.257 células malignas e 132 não resolvidas.

3.1.3 *Monocle*

O dado disponível está processado em $\log_2(TPM+1)$, mas para realizar as análises seguintes é recomendado utilizar o dado no formato de contagem pelo fato das ferramentas disponíveis possuírem métodos estatísticos mais acurados para ajustar ao modelo binomial negativo, que será utilizado para estes dados. Então, para converter o dado para contagem foi utilizado o pacote em R chamado *Monocle v2.22.0*(QIU *et al.*, 2017a; QIU *et al.*, 2017b; CAO *et al.*, 2019). O *monocle* é capaz de transformar uma matriz de expressão relativa em uma matriz de transcritos absolutos com base nos parâmetros de regressão linear inferidos a partir do valor de expressão relativa da isoforma mais abundante.

3.1.4 *DESeq2*

Com o dado de contagem é necessário ajustá-lo a um modelo específico que consiga tratar os dados de *scRNA-seq*. Para isso foi utilizado um modelo binomial negativo

generalizado e flexível com zero inflado chamado *ZINB-WaVe* e que está implementado através do pacote em R chamado *zinbwave v1.16.0* (RISSO *et al.*, 2018).

Com o dado ajustado ao modelo, foi utilizado o pacote em R chamado *DESeq2 v1.34.0* (LOVE; HUBER; ANDERS, 2014) para a realização da normalização dos dados. A escolha do *DESeq2* é baseada de acordo com Wang *et al.* (2019) , onde foram realizados testes de sensibilidade e especificidades mostrando que o *DESeq2* possui um ótimo desempenho em relação a outras ferramentas.

Tanto o *zinbwave* e *DESeq2* fazem uso do pacote *BiocParallel v1.28.0* para realizar a computação paralela. Pelo fato do dado de *scRNA-seq* possuir uma alta quantidade de amostras (células) e genes em relação a outras tecnologias (por exemplo, *RNA-seq*, Microarranjo etc.), faz-se necessário utilizar computação paralela para que o processo de modelagem e normalização seja realizado de uma maneira mais eficiente.

3.1.5 *Seurat*

Para a análise e exploração dos dados de *scRNA-seq* foi utilizado o pacote em R chamado *Seurat v2.3.4* (BUTLER *et al.*, 2018; STUART *et al.*, 2019). A partir dele foi possível a realização do controle de qualidade das células e genes, identificação dos genes mais variáveis, identificação dos *clusters* (tipos celulares) originalmente determinados no estudo, clusterização, redução de dimensão e identificação dos genes marcadores (expressão diferencial).

3.1.6 *MAGIC*

As tecnologias de *scRNA-seq* sofrem de muitas fontes de ruído técnico. Para isso foi utilizada a ferramenta *MAGIC v3.0.0* (DIJK *et al.*, 2018) em *python*, que é um método de imputação de células baseada na afinidade de *Markov*. A partir do *MAGIC* é possível visualizar e inferir como a expressão de até três genes estão se comportando ao mesmo tempo. Inicialmente é calculada a matriz de distância de célula por célula, em seguida a matriz de distância é convertida em matriz de afinidade usando a função gaussiana, dessa forma gerando uma curva de distância x afinidade. As afinidades são normalizadas, resultando em uma matriz de *Markov*. Para realizar a difusão, a matriz de *Markov* é

exponenciada a uma potência escolhida t. Por fim, multiplica-se a matriz exponencial de *Markov* pela matriz dos dados originais obtendo assim uma matriz de dados sem ruídos e imputada.

3.1.7 Enriquecimento funcional das células

Foram utilizados dois tipos de enriquecimentos para as células neste trabalho. O primeiro é focado na identificação de células com conjuntos de genes ativos (por exemplo, assinaturas, módulos de genes etc.). Essa análise foi possibilitada pelo pacote em R chamado *AUCell v1.16.0* (AIBAR *et al.*, 2017), que usa a "área sob a curva" (AUC) para calcular se um subconjunto do conjunto de genes de entrada é enriquecido dentro dos genes expressos para cada célula. A distribuição do score da AUC em todas as células permite explorar a expressão relativa da assinatura. Uma vez que o método de score é baseado em ranking, o *AUCell* é independente das unidades de expressão gênica e do procedimento de normalização. Além disso, como as células são avaliadas individualmente, a ferramenta pode ser facilmente aplicada a conjuntos maiores de dados.

A segunda forma de enriquecimento é o *GSVA v1.42.0* (*Gene Set Variation Analysis*) (HÄNZELMANN; CASTELO; GUINNEY, 2013). Esta ferramenta é disponibilizada na linguagem de programação R. Este é um método não paramétrico e não supervisionado utilizado para estimar o enriquecimento de um conjunto de genes através das amostras de um conjunto de dados de expressão. É realizado uma alteração nos sistemas de coordenadas, transformando os dados de um gene por matriz de amostra em um conjunto de genes por matriz de amostra, permitindo assim a avaliação do enriquecimento da via para cada amostra.

Para ambos os tipos de enriquecimento foram utilizadas assinaturas obtidas nesse trabalho e também assinaturas públicas disponibilizadas pelo banco de dados *MSigDB* (LIBERZON *et al.*, 2011; LIBERZON *et al.*, 2015). Neste banco há assinaturas relacionadas aos *hallmarks*, assinaturas computacionais, *GO*, *KEGG*, *Reactome*, imunológicas, oncogênicas etc.

3.1.8 Inferência de Trajetória

A análise de inferência de trajetória (ou análise pseudo temporal) foi realizada utilizando a ferramenta *STREAM v1.0* (CHEN *et al.*, 2019) disponibilizada em *python*. O objetivo dessa análise é verificar as trajetórias de diferenciação das diferentes linhagens, caracterizar a heterogeneidade celular e identificar as transições de estado. Isto é possibilitado a partir da mensuração da distância entre uma célula e o início da trajetória ao longo do caminho mais curto. O comprimento total da trajetória é definido em relação à quantidade total de mudança transcricional que uma célula sofre quando se move do estado inicial para o estado final. O *STREAM* é capaz de desemaranhar e visualizar trajetórias complexas a partir de dados de *scRNA-seq*.

A aplicação do *STREAM* inicia-se com uma matriz de expressão gênica de *scRNA-seq*, realizando três principais etapas: seleção de genes informativos, redução de dimensão e aprendizagem simultânea de estrutura de árvore e ajuste por *ElPiGraph*. A estrutura ótima é selecionada com base na minimização de energia elástica entre um conjunto de estruturas candidatas que são construídas toda vez que um nó de árvore é adicionado. A árvore final é interpretada como um conjunto de curvas conectadas representando diferentes trajetórias.

Vale ressaltar que o desenvolvimento/entendimento dos *scripts* e a interpretação dos resultados para essa ferramenta foram feitos em colaboração com is desenvolvedores da ferramenta, o Prof. Dr. Luca Pinello e o Prof. Dr. Huidong Chen, ambos do Instituto e Escola de Medicina de *Harvard* associados ao *Instituto Broad* do *MIT* e *Harvard*.

Além disso foi utilizada a ferramenta *psupertime* (MACNAIR; CLAASSEN, 2019) que é uma abordagem supervisionada de *pseudotime* que faz uso dos *labels* sequenciais como *input*. Esta ferramenta usa um modelo simples, baseado em regressão, que ao reconhecer os *labels* assegura que genes relevantes para o processo sejam encontrados. Esta metodologia complementa os resultados da ferramenta *STREAM*, pois a partir da identificação sequencial da trajetória com seus respectivos tipos celulares, é possível utilizar o *psupertime* para entender melhor a trajetória.

3.1.9 Coexpressão para Fatores de Transcrição usando Fatores de Impactos Regulatórios (RIF) e Correlação Parcial com Teoria da Informação (PCIT)

A análise transcriptômica tornou-se crucial para identificar os circuitos genéticos envolvidos na regulação dos *hallmarks* do câncer (HANAHAN; WEINBERG, 2011). Uma das formas inteligentes de explorar este tipo de dados e obter informações biologicamente relevantes sobre os mecanismos envolvidos na modulação dos circuitos genéticos é a inferência das redes regulatórias de genes (*GRNs*). Conceitualmente, podemos definir *GRN* como a reconstrução de redes de genes a partir de dados de expressão gênica, revelando a conexão dos fatores de transcrição (*TFs*) com seus alvos (HU *et al.*, 2020), com o objetivo de destacar quais interações gênicas são as mais relevantes para o estudo. Apesar da infinidade de ferramentas, novos métodos são necessários para avaliar todas as interações possíveis e sua significância (YU *et al.*, 2013). Além disso, a presença de *TFs* nas interações entre genes é funcionalmente crucial, pois elas podem estar desempenhando um papel regulador essencial nos processos biológicos (FARNHAM, 2009). Os *TFs* são considerados moléculas importantes que podem regular a expressão de um ou mais genes em um sistema biológico, determinando assim como as células funcionam e se comunicam com os ambientes celulares (VAQUERIZAS *et al.*, 2009).

Para melhor entendimento dos passos a seguir são necessários alguns conceitos como o da Análise dos Fatores de Impactos Regulatórios (*RIF*). O *RIF* visa identificar Fatores de Transcrição críticos calculando para cada condição a correlação de co-expressão entre os genes *TFs* e os Genes Diferencialmente Expressos. O resultado são as métricas *RIF1* e *RIF2* que permitem a identificação de *TFs* críticos. A métrica *RIF1* classifica os *TFs* como sendo os mais diferencialmente co-expressos com os genes altamente abundantes e altamente diferencialmente expressos, e a métrica *RIF2* classifica o *TF* com a capacidade mais alterada de agir como preditores da abundância de genes diferencialmente expressos (REVERTER *et al.*, 2010).

Outro conceito importante é o da Análise da Correlação Parcial com Teoria da Informação (*PCIT*). Essa metodologia tem sido utilizada para a reconstrução das Redes de Co-expressão Gênica (*GCN*). O *GCN* combina o conceito do coeficiente de correlação parcial com a Teoria da Informação para identificar associações significativas entre gene-gene. Nesta fase, a correlação pareada de três genes é realizada simultaneamente, fazendo assim a inferência dos genes coexpressos. Esta abordagem é mais sensível do que outros métodos

e permite a detecção de interações gene-gene validadas funcionalmente (REVERTER; CHAN, 2008).

Após a apresentação dos conceitos de *RIF* e *PCIT* foi-se utilizado o pacote em R chamado *CeTF v1.6.0* (BIAGI *et al.*, 2021) disponível a partir do repositório *Bioconductor* (https://bioconductor.org/packages/CeTF) comparando as condições de células Malignas x Não Malignas. Nessa etapa foram utilizados os parâmetros $|lfc| > 1$ e $padj < 0.05$. Em seguida é possível identificar os *keyTFs*, ou seja, os *TFs* que foram considerados como mais relevantes para ambas as condições. Além dos *keyTFs* foram adicionados outros genes relacionados a via de *EMT* e *Stemness* (ALVES *et al.*, 2013) para os próximos passos. Dentro do ambiente R foi utilizado o pacote *RCy3 v2.14.0* (GUSTAVSEN *et al.*, 2019) para fazer a interface com o Cytoscape (SHANNON *et al.*, 2003), dessa forma utilizando os dados gerados pelo pacote *CeTF* dentro do R para gerar as redes usando do *Cytoscape*. O próximo passo foi selecionar cada gene individualmente e aplicar o algoritmo de difusão (CARLIN *et al.*, 2017) a fim de gerar sub redes menores focadas especificamente no gene em questão. A partir dessas sub redes foi possível identificar *clusters* (comunidades) que possuem similaridades entre si. Para isso foi utilizado o algoritmo *louvain* que é baseado na medida de modularidade e numa abordagem hierárquica. Por fim, para cada comunidade gerada foi realizado o enriquecimento funcional utilizando os bancos de dados do *KEGG* (*Kyoto Encyclopedia of Genes and Genomes*) e do *GO* (*Gene Ontology*) para Processos Biológicos (*BP*), Funções Moleculares (*MF*) e Componentes Celulares (*CC*)

3.1.10 Validação funcional da expressão do *lncRNA TRHDE-AS1* e *HOTAIR*

A avaliação dos níveis de expressão gênica foi verificada em cultivos celulares que representam os principais grupos de progressão do melanoma: linhagens celulares de melanoma primário (A375) e linhagens celulares de melanoma metastático (SKMEL147). Para verificar os níveis de expressão gênica foi utilizada a técnica de *PCR* quantitativa em tempo real (*RT-qPCR*). Assim, foi realizada a extração do *RNA* total das diferentes células utilizando o *miRNeasy Mini Kit* (*Qiagen*), que em seguida foi convertido para *cDNA* por meio do *High-Capacity cDNA Reverse Transcription Kit* (*Applied Biosystems*) e armazenado em -20°C. Foram desenhados *primers* gene-específicos que possam abranger as extremidades de dois *exons* diferentes, para evitar amplificação inespecífica (*DNA*

contaminante) e que possuam tamanho de 80-150 *nt* (tamanho ideal para a detecção). O reagente utilizado para detectar a amplificação pela *PCR* foi o *SYBR Green PCR Master Mix* (*Applied Biosystems*) e este contém todos os componentes necessários, como enzima *DNA* polimerase e reagente fluorescente. Assim, após o preparo seguindo as normas do fabricante, a reação será levada ao equipamento 7500 *Fast Real-Time PCR* (*Applied Biosystems*) para realização do experimento. Os valores de expressão obtidos serão analisados e comparados com valores de genes endógenos para verificar variações nos níveis de expressão.

Foi utilizada a metodologia de *siRNA* (*Small Infering RNA*) para fazer o silenciamento do *HOTAIR* e *TRHDE-AS1* nas duas diferentes linhagens celulares. Os *siRNAs* são altamente específicos e geralmente sintetizados para reduzir a tradução de *RNAs* de mensageiros específicos (*mRNAs*). Isto é feito para reduzir a síntese de proteínas particulares. Eles se formam a partir de *RNA* de cadeia dupla transcritos e depois cortados à medida no núcleo antes de serem liberados no citoplasma. A interferência do *RNA* (*RNAi*) é um fenômeno no qual a introdução do *RNA* de dupla cadeia (*dsRNA*) em uma gama diversificada de organismos e tipos celulares causa a degradação do *mRNA* complementar. Na célula, os *dsRNAs* longos são clivados em pequenos nucleotídeos de tamanho 21-25 de *RNAs* pequenos interferentes, ou *siRNAs*, por um ribonuclease conhecido como *Dicer*. Os *siRNAs* são posteriormente montados com componentes proteicos em um complexo de silenciamento induzido por *RNA* (*RISC*), desenrolando-se no processo. O *RISC* ativado se liga então à transcrição complementar por interações de emparelhamento de base entre o fio *siRNA antisense* e o *mRNA*. O *mRNA* ligado é clivado e a degradação específica da sequência do *mRNA* resulta no silenciamento do gene. Nesta validação, para ambas as linhagens (SKMEL147 e A375) foram utilizados 25 pmol/ul de *TRHDE-AS1 siRNA* ou *HOTAIR siRNA*. Os experimentos foram feitos em triplicatas.

3.1.11 *Docker*

Para todos os passos e análises realizados para os dados de melanoma foi utilizado o *Docker*, que é uma plataforma *open source* com a finalidade de facilitar a criação e administração de ambientes isolados. O *Docker* possibilita que uma aplicação ou um ambiente de trabalho completo esteja disponível dentro de um *container*, a partir disso o

ambiente todo torna-se portável para qualquer outro *host* que tenha o *Docker* instalado. Isso reduz drasticamente o tempo de implementação de alguma infraestrutura ou até mesmo aplicação, pois não há a necessidade de ajustar o ambiente para o funcionamento correto do serviço, o ambiente é sempre o mesmo, basta configurar uma única vez e é possível replicá-lo quantas vezes quiser. Na Figura 12 é possível observar como funciona o *Docker* resumidamente.

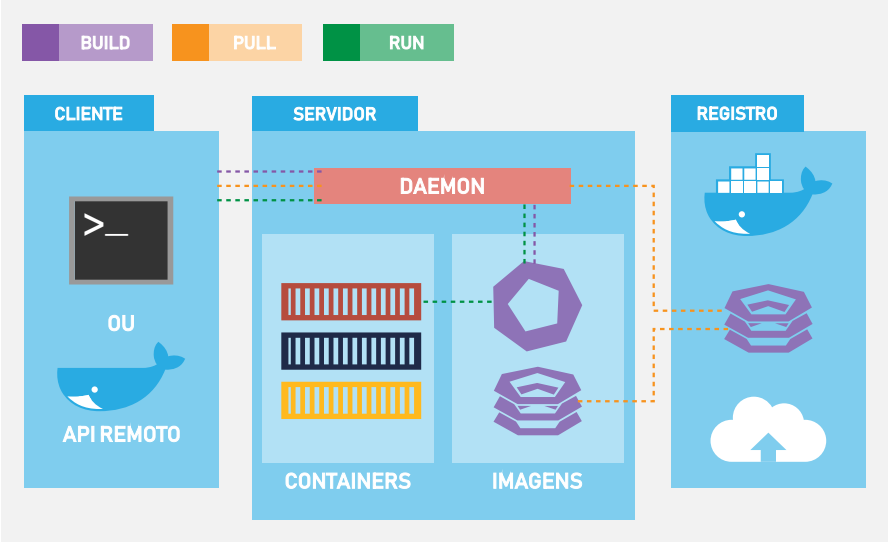

Figura 12 – Fluxograma do funcionamento do *Docker*.

Foram utilizadas três imagens, uma contendo a interface gráfica do R (*Rstudio*) com todos os pacotes e dependências necessárias para as análises (*Monocle*, *Seurat*, *GSVA* e *AUCell*), outra contendo o software *STREAM*, e por fim uma imagem contendo o software *MAGIC*. As imagens estão disponíveis para download a partir do *dockerhub* nos seguintes links:

- ⟨https://hub.docker.com/r/biagii/scrnaseq-bit⟩;
- ⟨https://hub.docker.com/r/biagii/magic⟩;
- ⟨https://hub.docker.com/r/pinellolab/stream⟩;

Além das imagens disponíveis em *docker*, há o repositório do *GitHub* onde pode ser encontrado o arquivo usado para a construção das imagens: ⟨https://github.com/cbiagii⟩.

3.2 Câncer de Pulmão de Pequenas Células (CPPC)

Para a análise de dados de CPPC foi utilizado o *pipeline* descrito a seguir. O dado gerado pelo sequenciador, chamado de dado bruto, é submetido à ferramenta *M3K* para controle de qualidade das amostras, mapeamento contra o genoma de referência, remoção e quantificação das *reads* duplicadas, remoção das células contaminadas e determinação do número de células viáveis. Para cada amostra é gerado um arquivo de contagem dos genes (linhas) por células (colunas). A partir da contagem é utilizado o *Seurat* (HAO *et al.*, 2021) para o pré-processamento do dado, incluindo normalização, identificação dos genes mais variáveis, clusterização e redução de dimensão. A remoção dos dupletos foi feita utilizando a ferramenta *Chord* (XIONG *et al.*, 2021). Em seguida foi calculado o *score* de heterogeneidade intratumoral (*ITH*) (STEWART *et al.*, 2020) e então todas as amostras foram integradas com a correção do *batch* pela química do sequenciamento. Para estimar a velocidade do RNA nas amostras foram utilizadas as ferramentas *velocyto* (MANNO *et al.*, 2018) e *scVelo* (BERGEN *et al.*, 2020). Por fim, para inferir as comunicações célula a célula foi utilizada a ferramenta *CellChat* (JIN *et al.*, 2021). O *workflow* da metodologia pode ser observado na Figura 13.

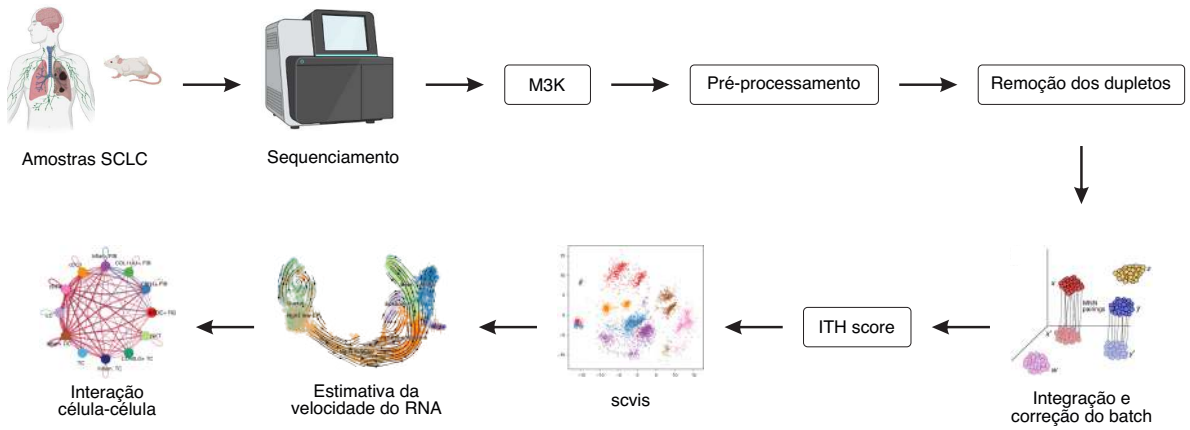

Figura 13 – *Workflow* da metodologia usado para os dados de CPPC. Iniciando com a coleta das amostras e em seguida o sequenciamento. Os dados brutos provenientes do sequenciamento são processados pelo software *M3K* e as matrizes de contagem são geradas. A partir das matrizes de contagem é realizado o pré-processamento utilizando o *Seurat*. A remoção dos dupletos é feita utilizando o pacote *Chord*. A integração de todas as amostras é feita e a correção do *batch* baseado na diferente química usada no sequenciamento. O *ITH* é calculado seguido pela redução de dimensão usando o *scvis*. Por fim é calculada a estimativa da velocidade do *RNA* e a interação célula-célula.

3.2.1 Infraestrutura computacional

Todos as etapas de processamento para os dados de CPPC utilizaram o *cluster CHEOPS* (*Cologne High Efficiency Operating Platform for Science*). O acesso ao *CHEOPS* só foi possível pela parceria que estabeleci na ocasião do meu estágio sanduíche na Universidade de Colônia, na Alemanha. Este *cluster* já foi considerado um dos 500 computadores de alto desempenho mais rápidos do mundo. Além das especificações observadas na Figura 14 vale a pena ressaltar que a infraestrutura também conta com máquinas equipadas com unidade de processamento gráfico (*GPU*), o que otimizou o tempo das análises.

Figura 14 – Especificações do *CHEOPS*, servidor utilizado para rodar as análises de CPPC. Destacando 500 *TB* de armazenamento, 1.730 *CPUs* e 9.712 *CPU-Cores*.

3.2.2 Dados de *scRNA-seq* de CPPC

Foram utilizadas 54 amostras totais, sendo 14 amostras de pacientes diretos e 40 amostras de tumores gerados por xenotransplante (*xenograft*). Dessas 40 amostras, 33 são de ensaios derivados de pacientes (*PDX*) e 7 são de explantes derivados de células tumorais circulantes (*CDX*). Todas as amostras totalizam 213.874 células, sendo 67.384 clínicas, 121.385 *CDX* e 25.105 *PDX*. O dado não está disponível publicamente pois está em processo de publicação.

Essas amostras foram extraídas de diferentes fontes como metástase cerebral, tumor primário, metástase intrapulmonar, metástase dos linfonodos, metástase pleural e células tumorais circulantes. Em seguida foi realizada a suspensão das células, formação das *GEMs* (*Gel Beads in Emulsion*) e por fim o sequenciamento utilizando o equipamento *Illumina NovaSeq 6000*. O *workflow* do sequenciamento pode ser observado na Figura 15.

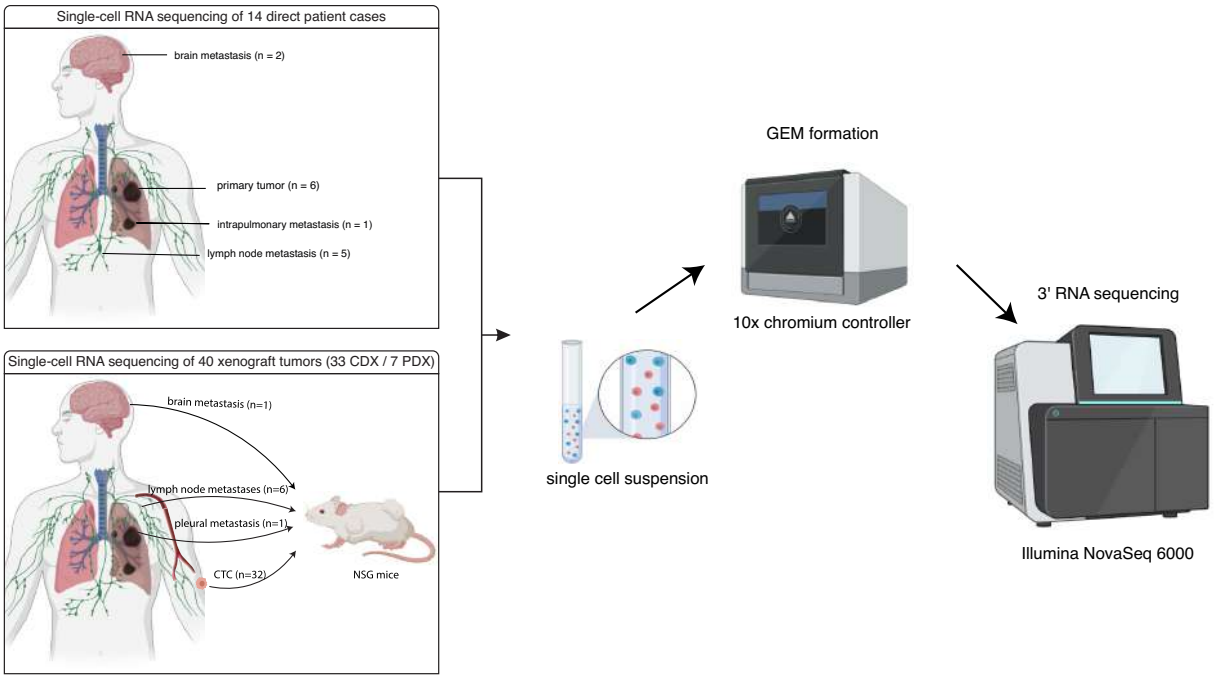

Figura 15 – *Workflow* do processo de coleta das amostras, tanto para casos de pacientes diretos e para os tumores de xenotransplantes, mostrando o local de onde as amostras foram extraídas. Em seguida é realizada a suspensão das células únicas, formação das *GEMs* e, por fim, o sequenciamento.

3.2.3 *M3K*

A ferramenta *M3K* está sendo desenvolvida em parceria com o Dr. Miloš Nikolić e Dr. Martin Peifer da Universidade de Colônia, na Alemanha. O artigo está em processo final de escrita e submissão. A ferramenta possui 7 principais módulos que serão abordados a seguir:

Módulo 1: Controle de qualidade

Este módulo serve como um *check-up* inicial da qualidade da amostra e consiste em duas partes, o software *FastQC* que gera estatísticas básicas da amostra e *check-ups* de qualidade (*Babraham Institute*, ⟨http://www.bioinformatics.babraham.ac.uk/projects/fastqc/⟩), e um programa adicional que calcula as frequências de nucleotídeos das primeiras 30 posições de nucleotídeos a partir do arquivo *Read 2*. O último programa fornece informações sobre possíveis problemas de sequenciamento.

Módulo 2: Fragmentação dos arquivos *FASTQ*

Para melhorar a eficácia do software, os arquivos *FASTQ* são divididos em vários pedaços de igual tamanho.

Módulo 3: Extração e correção dos *barcodes*

Os *barcodes* (*BC*) e o Identificador Molecular Único (*UMI*) são extraídos de cada linha no arquivo *Read 1*; o comprimento de um *BC* é de 16 nucleotídeos, enquanto o comprimento do *UMI* pode ser de 10 ou 12 nucleotídeos, dependendo da versão da química usada no experimento (v.2 e >v.3, respectivamente). Como algumas dessas sequências inevitavelmente conterão erros de sequenciamento, e os *BCs* corretos são identificados por comparação com uma lista de referência correspondente de *BCs*, que são então adicionados com seus correspondentes *UMIs* aos nomes das sequências na *Read 2*. Para os *BCs* restantes, uma correção da sequência daqueles com uma distância de um a qualquer um dos *BCs* da lista de referência é tentada, enquanto as leituras dos *BCs* restantes contendo erros de sequenciamento em múltiplos nucleotídeos são excluídas de análises posteriores.

Módulo 4: Mapeamento das sequências

O mapeamento dos arquivos *FASTQ* da *Read 2* contra um genoma de referência é realizado independentemente em cada pedaço (ver **Módulo 3**) usando o software de alinhamento *STAR* (DOBIN *et al.*, 2013) retendo apenas *reads* mapeadas de forma única.

Módulo 5: Junção dos arquivos *BAM*

A junção dos pedaços contendo sequências mapeadas como descrito anteriormente é realizada usando a ferramenta *Samtools* (LI *et al.*, 2009).

Módulo 6: Procedimento de deduplicação e quantificação das contagens

As leituras duplicadas são definidas como sequências da mesma célula (*BC*), tendo o mesmo *UMI* e as mesmas posições iniciais ou ligeiramente diferentes (entre 1 e 3 nucleotídeos) (SENA *et al.*, 2018), que são excluídas da análise. O procedimento de quantificação das *reads* restantes é realizado pela sobreposição de cada leitura com regiões codificadoras de um transcriptoma ligeiramente alterado (regiões codificadoras sobrepostas pertencentes aos mesmos genes são previamente fundidas, e sua união é utilizada em seu lugar).

Módulo 7: Descontaminação das amostras e chamada de células

No módulo final, o número de células viáveis é determinado a partir de uma amostra. Caso uma amostra seja derivada de um modelo *PDX* ou de um modelo de rato *CDX* (*xenografts* derivados de pacientes ou *xenografts* derivados de células tumorais circulantes, respectivamente) (HIDALGO *et al.*, 2014; HODGKINSON *et al.*, 2014), um passo prévio é necessário, ou seja, a identificação e remoção de células pertencentes a um organismo hospedeiro (tipicamente um rato). A este respeito, uma amostra *PDX/CDX* é primeiramente mapeada independentemente tanto para o genoma humano (*hg19*) quanto para um genoma de camundongo (*mm10*), as leituras são quantificadas para produzir matrizes de contagem, que são então usadas para calcular uma pontuação de mapeamento para cada célula.

Em seguida, as pontuações são classificadas em ordem ascendente e plotadas como uma curva suavizada para calcular a primeira derivada em cada ponto, após isso a curva é novamente suavizada e a segunda derivada é calculada. O ponto extremo local da curva é usado como um corte, e todas as células com uma pontuação inferior ao corte são consideradas como células hospedeiras ou mortas, e são, portanto, removidas da análise.

Finalmente, o procedimento para determinar o número total de células viáveis é similar ao procedimento acima mencionado para remover células hospedeiras indesejadas. Em resumo, as células são classificadas por seu número de leituras quantificadas em ordem decrescente e plotadas. Em seguida, a curva é suavizada e a primeira derivada é calculada, representando um corte para as células viáveis.

3.2.4 *Seurat*

Inicialmente cada amostra, individualmente, é submetida a um pré-processamento utilizando o pacote em R *Seurat* v4.0.4 (HAO *et al.*, 2021). A partir da matriz de contagem com as células viáveis obtidas pelo *M3K* é realizado um controle de qualidade removendo as células com baixa qualidade, como fragmentos de células ou células apoptóticas que possuem poucos transcritos ou conteúdo mitocondrial elevado. Em seguida é realizada a normalização, identificação dos genes mais variáveis, escalonamento dos dados corrigindo os efeitos do ciclo celular e genes mitocondriais, clusterização, identificação dos genes marcadores e, por fim, a redução de dimensão (*t-SNE* e *UMAP*).

Vale a pena ressaltar a correção do efeito da heterogeneidade do ciclo celular. Esse efeito é corrigido calculando um *score* de fase do ciclo celular baseado em marcadores canônicos (TIROSH *et al.*, 2016a) e aplicando um modelo de regressão linear para descontar esse viés. Como primeira etapa atribuímos a cada célula um *score* baseado na expressão de marcadores da fase *G2/M* e *S*. No caso das células que não expressam nenhum dos marcadores anteriores essas células são classificadas na fase *G1*. Após a correção é gerada uma matriz de expressão A de dimensões m x 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. Essa matriz de expressão corrigida será usada para calcular a redução de dimensão.

3.2.5 Remoção de dupletos

Se duas células se aglomeram e receberem o mesmo *BC*, os transcritos serão misturados, gerando dupletos. A fração de dupletos é, em geral baixa, e podemos aplicar métodos computacionais para removê-los. Para isso utilizamos o pacote em R *Chord* v1.0 (XIONG *et al.*, 2021) que é um algoritmo para identificação dos dupletos que utiliza métodos de classificação de aprendizado de máquina para integrar múltiplas ferramentas e resultados diferentes.

3.2.6 Integração, normalização de quantis e correção de *batch*

As amostras individuais foram fusionadas em uma única matriz. Para isso foi necessário realizar uma normalização entre essas amostras. Utilizamos a normalização de quantis, que é uma técnica que transforma duas ou mais distribuições idênticas em propriedades estatísticas. Para normalizar por quantil duas ou mais distribuições uma para a outra, sem uma distribuição de referência, ordenar como antes, depois definir a média (geralmente, média aritmética) das distribuições. Assim, o valor mais alto em todos os casos torna-se a média dos valores mais altos, o segundo valor mais alto torna-se a média dos segundos valores mais altos, e assim por diante (vide Figura 16).

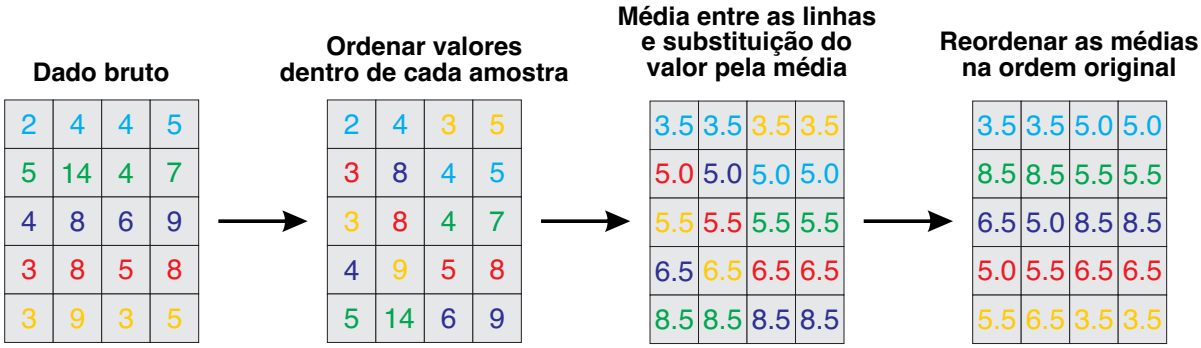

Figura 16 – Exemplificação esquemática de como funciona a normalização de quantil. Inicia-se com uma matriz, a partir dessa matriz as amostras (colunas) são ordenadas com seus valores em ordem crescente. Em seguida é calculada a média para cada linha e os valores que foram ordenados são substituídos pela média por linha. Por fim, as médias são reordenadas na ordem inicial.

As amostras foram sequenciadas com dois tipos diferentes de versões de química, v2 e v3. Para corrigir esse efeito, que pode interferir no dado, foi realizada uma correção de *batch*, considerando a versão como o efeito. Para isso usamos o pacote em R *batchelor* v1.10.0 (HAGHVERDI *et al.*, 2018) que utiliza o método *MNN* (*Mutual Nearest Neighbors*). Esse método é baseado na detecção de vizinhos mútuos mais próximos (*MNNs*) no espaço de expressão de alta dimensão. Essa abordagem não se baseia em composições populacionais pré-definidas ou iguais entre *batches*; em vez disso, requer apenas que um subconjunto da população seja compartilhado entre os *batches*.

3.2.7 *Score* de heterogeneidade intratumoral (*ITH*)

O *score* de heterogeneidade intratumoral (*ITH*) foi definido como a distância Euclidiana média entre as células individuais e todas as outras células, em termos dos primeiros 20 componentes principais derivados dos níveis de expressão normalizados de genes altamente variáveis. Os genes altamente variáveis foram identificados a partir do *Seurat*, usando a função *FindVariableGenes()*. Esse processo foi feito realizando um *bootstrap* de 1.000 interações considerando 500 células aleatórias por vez. Esse procedimento foi realizado pelo fato de haver uma variação considerável de número de células entre as diferentes amostras. Dessa forma é possível balizar o *score ITH* e torná-lo comparável entre as amostras.

Além de calcular o *ITH* para cada uma das amostras de CPPC, foi calculado também, a título de comparação, para um estudo de adenocarcinoma pulmonar com número de acesso *GSE131907*. Para este estudo foram selecionadas apenas as células tumorais (epiteliais usando o expressão positiva do gene marcador *EPCAM*). E por fim, foi calculado o *ITH* para uma amostra de células PC9, que são derivadas de um adenocarcinoma humano de tecido pulmonar que permanece indiferenciado.

3.2.8 *scvis*

Os algoritmos de redução de dimensão existentes ou não são capazes de descobrir as estruturas de agrupamento nos dados ou perdem informações globais, tais como grupos de agrupamentos que estão próximos uns dos outros. Para isso, usamos o pacote em *python* *scvis* v0.1.0 (DING; CONDON; SHAH, 2018) que é um modelo estatístico robusto para capturar e visualizar estruturas de baixa dimensão em dados de expressão gênica de células únicas. Com o *scvis* é possível preservar tanto as estruturas vizinhas locais quanto globais.

A ferramenta consiste em duas etapas. Uma primeira etapa de treinamento, onde o algoritmo aprende o mapeamento paramétrico probabilístico dos dados. E por fim a segunda etapa que é a de mapeamento. Após aprender o mapeamento paramétrico probabilístico, é adicionado os resultados a uma nova dimensão.

3.2.9 Estimativa da velocidade do *RNA*

Para determinar a estimativa de velocidade do *RNA* é necessário o arquivo *BAM* que contém o alinhamento gerado pela ferramenta *M3K*. A partir do arquivo *BAM* foi utilizada a ferramenta *velocyto* v0.17.15 (MANNO *et al.*, 2018) e gerado um arquivo com extensão *loom* que contém duas matrizes de contagem de abundâncias pré-maduras (*unspliced*) e maduras (*spliced*).

A partir do arquivo *loom* é possível usar a ferramenta em *python scVelo* v0.2.4 (BERGEN *et al.*, 2020) a fim de, finalmente, estimar as velocidades. A partir dessa ferramenta também foi possível projetar as velocidades estimadas em uma redução de dimensão já calculada, identificar genes mais importantes, verificar a dinâmica de genes específicos, entre outras funcionalidades.

3.2.10 Interação célula-célula

Usamos o pacote em R *CellChat* v1.1.3 (JIN *et al.*, 2021) para inferir e analisar quantitativamente as redes de comunicação intercelular usando seu repositório de interações entre ligantes, receptores e seus cofatores que representam com precisão complexos moleculares heteroméricos conhecidos. Além disso, o *CellChat* é capaz de quantificar a semelhança entre todas as vias de sinalização significativas e depois agrupá-las com base na semelhança de sua rede de comunicação celular. Uma das formas de fazer o agrupamento é com base na similaridade funcional. O alto grau de similaridade funcional indica que os principais ligantes e receptores são semelhantes, e pode ser interpretado como as duas vias de sinalização ou dois pares ligante-receptor que exibem papéis semelhantes e/ou redundantes.

4 Resultados e Discussão

Nesta seção são abordados os resultados utilizando os dados de células únicas de melanoma e câncer de pulmão de pequenas células a partir das metodologias propostas acima.

4.1 Melanoma

4.1.1 Visão geral dos dados

Inicialmente foram aplicados dois filtros a fim de melhorar a qualidade do dado. O primeiro filtro foi para remover as células não resolvidas e não classificadas, e o segundo filtro foi aplicado para manter apenas os genes que possuem o valor de contagem maior que 1. Após a aplicação dos filtros para os genes e células restaram 22.606 genes e 4.097 células, com 2.840 não malignas e 1.257 malignas. As células não malignas possuem seis subtipos: 512 células B, 56 fibroblastos associados ao câncer (*CAF*), 62 células endoteliais, 199 macrófagos, 51 células exterminadoras naturais (*NK*) e 2040 células T.

A fim de representar visualmente a distribuição das células, foi utilizado o método *UMAP – Uniform Manifold Approximation and Projection for Dimension Reduction* (MCINNES; HEALY; MELVILLE, 2020) que é uma técnica de aprendizado de máquina para redução de dimensão. A partir do conhecimento da divisão dos tipos e subtipos celulares presentes nesse estudo fez-se o uso da técnica *UMAP* para verificar a distribuição espacial dos dados. Essa redução foi baseada em todos os genes disponíveis após a realização dos filtros nas células e genes. É possível observar na Figura 17 que as células malignas estão bem separadas em relação às não malignas e seus respectivos subtipos. A Figura 18 mostra a distribuição de cada tipo celular individualmente. Nota-se que as células B, *CAF*, endoteliais e macrófagos possuem uma separação bem definida, enquanto as células *NK* e T estão sobrepostas e as células malignas possuem alguns outliers de seu grupo maior. Para a realização do *UMAP* é necessário definir quantas componentes da *PCA* (*Principal Component Analysis*) serão utilizadas. Para encontrar o melhor número, verificou-se, dentre as 50 componentes principais calculadas, quantas são necessárias para explicar 80% ou mais da variância. O número ideal é de 11 componentes principais.

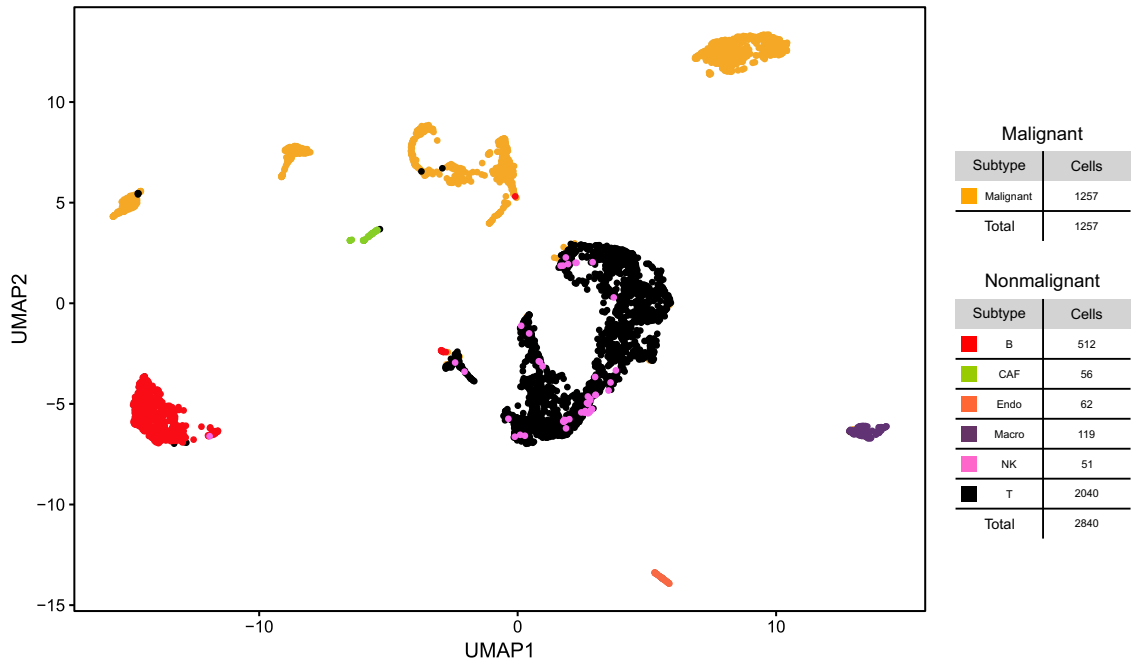

Figura 17 – Redução da dimensão utilizando a técnica *UMAP* com a anotação das células original para a visualização de todas as células e genes após a realização do filtro.

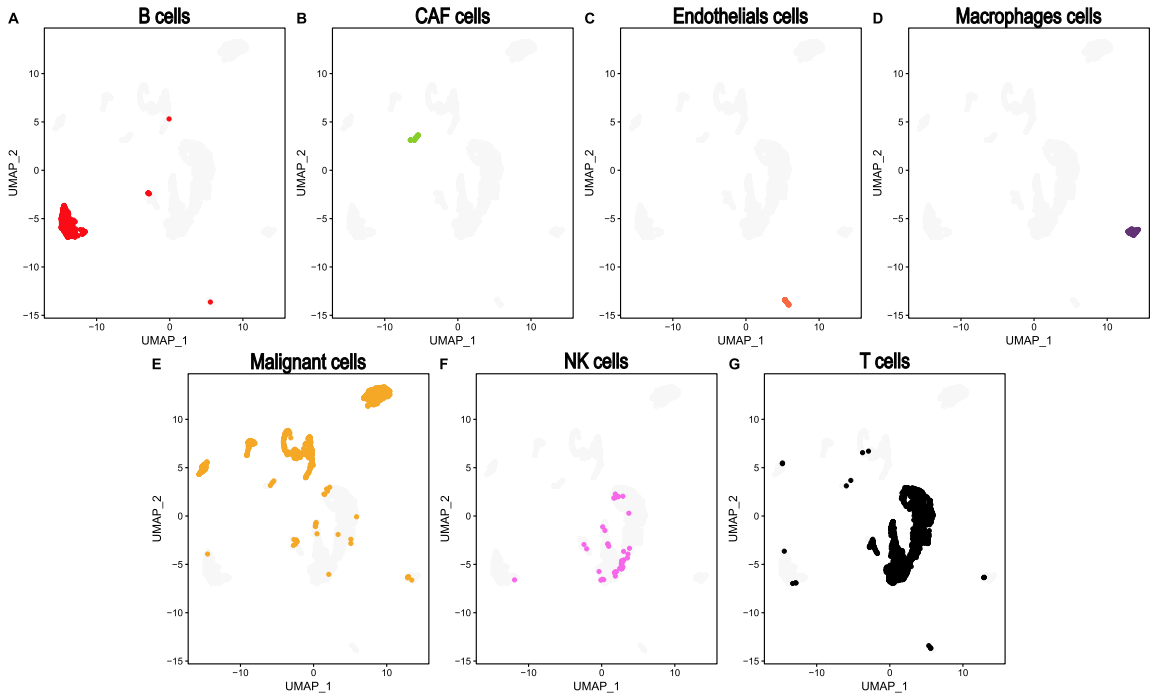

Figura 18 – Identificação individual de cada tipo celular, sendo: (A) células B, (B) células *CAF*, (C) células endoteliais, (D) macrófagos, (E) células malignas, (F) células *NK* e (G) células T.

Para cada tipo/subtipo celular foi possível identificar os genes marcadores (diferencialmente expressos). Para isso foi utilizado o teste de *Wilcox* onde comparou-se a

expressão de cada grupo individualmente em relação aos demais. Na Tabela 1 podemos observar os *top* 4 marcadores exclusivos identificados para cada grupo. Para a seleção dos marcadores foram utilizados valores de cortes para módulo de $logFC > 1$ e $pvalue\_adj < 0.05$. Dentre os marcadores identificados há os já conhecidos na literatura, como *IGJ* e *MS4A1* (HYSTAD *et al.*, 2007) para células B, *IFI30* (NGUYEN *et al.*, 2016) para células Macrofagiais, *PMEL* e *SERPINE2* (JUNGBLUTH *et al.*, 1999; YANG *et al.*, 2018) para células malignas, *IL32* (DAHL *et al.*, 1992) para células T, entre outros.

Tabela 1 – *Top* 4 genes marcadores (diferencialmente expressos) para cada grupo quando comparados contra todos os outros utilizando *Wilcox*. A coluna *pct.1* e *pct.2* representam, respectivamente, a fração de expressão das células que o determinado marcador possui para o grupo em questão e a porcentagem de expressão das células que o marcador possui para os outros grupos. Na coluna grupo temos entre parênteses a quantidade total de marcadores únicos para o determinado grupo.

Markers	avg_logFC	pct.1	pct.2	p_val_adj	Grupo
IGLL5	3,2670495	0,691	0,037	0	B (44)
ELK2AP	3,2534135	0,443	0,018	4,40E-269	
IGJ	2,721187	0,264	0,014	7,70E-143	
MS4A1	2,6169791	0,957	0,034	0	
COL3A1	3,93611	1	0,027	0	CAF (117)
CCL19	3,8255565	0,232	0,004	4,97E-84	
DCN	3,8223357	0,964	0,01	0	
COL1A2	3,80332	1	0,063	1,38E-166	
CCL14	3,3981564	0,597	0,004	0	Endo (134)
DARC	3,2731436	0,258	0,006	1,60E-81	
IGFBP7	3,2153989	1	0,18	5,06E-79	
EFEMP1	3,0107769	0,532	0,013	4,90E-172	
C1QB	3,81853	0,681	0,058	2,52E-152	Macro (135)
LYZ	3,7074272	0,992	0,746	1,25E-67	
IFI30	3,5572226	0,975	0,176	2,52E-140	
C1QA	3,4199387	0,697	0,016	0	
PMEL	3,4731665	0,831	0,129	0	Malignant (87)
SERPINE2	3,3914864	0,947	0,061	0	
APOD	3,3094289	0,891	0,061	0	
S100B	3,073743	0,949	0,064	0	
GNLY	3,7755032	0,706	0,043	8,01E-103	NK (40)
CGR3A	2,3171837	0,647	0,047	1,21E-76	
KLRB1	2,2939385	0,804	0,078	7,31E-75	
KLRC1	2,184789	0,549	0,029	1,47E-87	
CD3D	2,8570077	0,907	0,019	0	T (77)
IL32	2,4413885	0,899	0,088	0	
CD8A	2,4091536	0,543	0,011	0	
GZMK	2,3231764	0,495	0,016	4,83E-265	

Observa-se na Figura 17 e Figura 18 que há subdivisões dentro de cada tipo e subtipo celular. A fim de identificar as subdivisões dentro dos tipos celulares aplicamos uma clusterização não supervisionada nos dados para que fossem identificados *clusters*

independente da anotação original. A clusterização retornou 18 *clusters* (Figura 19), sendo que os *clusters* 4, 5, 8, 9, 10, 12 e 17 correspondem às células malignas, enquanto os demais *clusters* correspondem às células não malignas e seus respectivos subtipos. Com essa nova divisão é possível identificar populações mais específicas e seus respectivos marcadores. Na Figura 20 é mostrado individualmente a distribuição das células entre os 18 *clusters* gerados. Da mesma maneira que foi feita anteriormente, foram identificados os genes marcadores para cada *cluster*. Na Tabela 2 é possível visualizar o principal marcador para cada *cluster*.

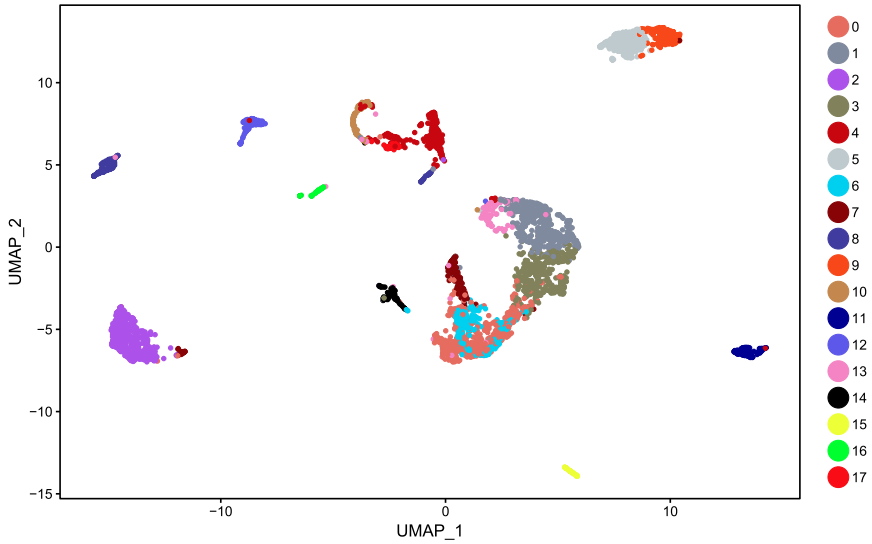

Figura 19 – Redução da dimensão utilizando a técnica *UMAP* após a clusterização não supervisionada para a visualização de todas as células e genes.

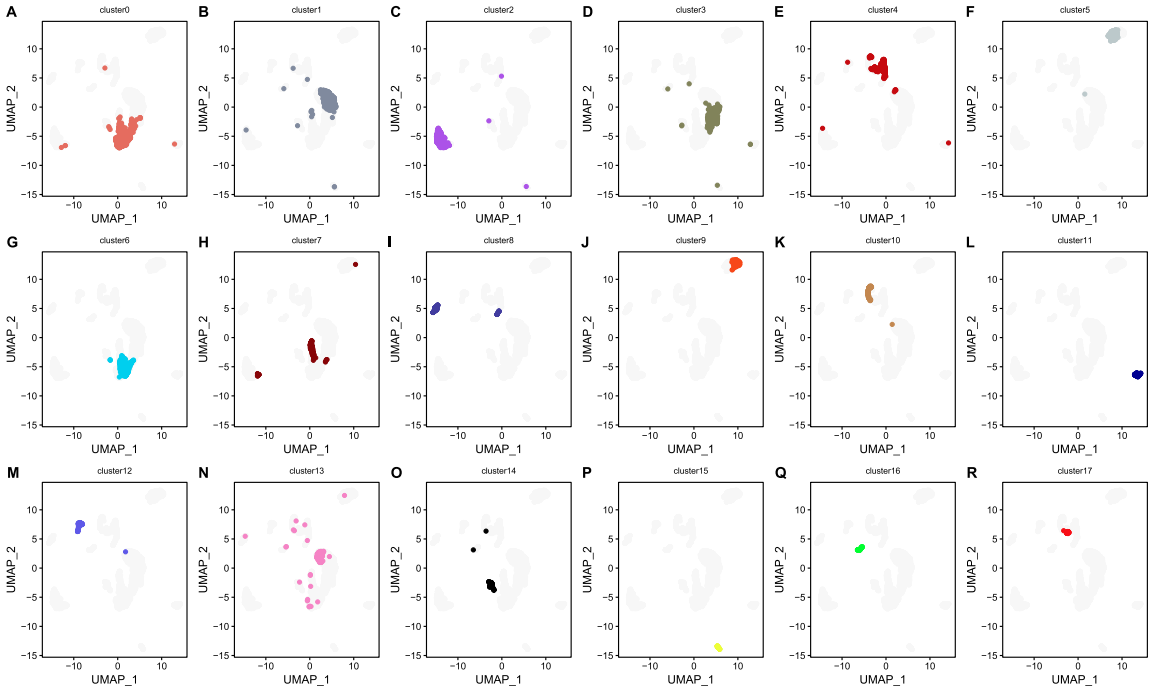

Figura 20 – Identificação individual de cada *cluster*, sendo: (*A*) *cluster0* até (*R*) *cluster17*

Tabela 2 – Principais genes marcadores (diferencialmente expressos) para cada *cluster* quando comparados contra todos os outros utilizando o teste de *Wilcox*. A coluna *pct.1* e *pct.2* representam, respectivamente, a fração de expressão das células que o determinado marcador possui para o grupo em questão e a porcentagem de expressão das células que o marcador possui para os outros grupos. Na coluna grupo temos entre parênteses a quantidade total de marcadores únicos para o determinado grupo.

Markers	avg_logFC	pct.1	pct.2	p_val_adj	Grupo
XCL2	1,1896624	0,254	0,054	1,16E-54	cluster0 (4)
IL7R	1,7560677	0,848	0,144	0,00E+00	cluster1 (7)
IGLL5	3,1681814	0,685	0,044	0,00E+00	cluster2 (43)
IL32	1,071815	0,943	0,443	1,05E-88	cluster3 (3)
BCAN	1,7486594	0,612	0,046	3,40E-249	cluster4 (7)
RDH5	1,1642526	0,633	0,051	1,12E-248	cluster5 (5)
CD8A	1,7083146	0,943	0,227	2,72E-178	cluster6 (30)
UGDH-AS1	2,5623522	1	0,994	5,88E-99	cluster7 (49)
SAA1	4,3550331	0,825	0,005	0	cluster8 (42)
SERPINA3	2,3942934	0,963	0,146	9,68E-175	cluster9 (31)
TYRP1	2,8102121	0,788	0,052	3,90E-226	cluster10 (35)
C1QB	3,8451649	0,683	0,057	1,09E-158	cluster11 (138)
CAPG	2,25736	0,992	0,271	3,49E-105	cluster12 (63)
GNLY	1,4909425	0,148	0,049	1,10E-02	cluster13 (5)
HMGB2	1,9650862	0,99	0,282	2,55E-76	cluster14 (48)
CCL14	3,4107996	0,603	0,004	0	cluster15(111)
CCL21	4,8562757	0,196	0,014	4,68E-22	cluster16 (116)
SLC45A2	3,2281766	0,98	0,152	5,46E-70	cluster17 (65)

4.1.2 Visualização dos dados usando os lncRNAs

Como abordado no tópico anterior, observa-se que há uma clara separação entre os diferentes tipos celulares quando utilizados todos os genes para a visualização. Foi verificado, a seguir, se esse mesmo comportamento de separação entre os diferentes tipos celulares também se mantém quando o foco está apenas nos *lncRNAs*. Para a realização do *UMAP* é necessário definir, inicialmente, quantas componentes principais da *PCA* serão utilizadas. Para encontrar o melhor número, verificou-se, dentre as 50 componentes principais calculadas, quantas são necessárias para explicar 80% ou mais da variância. O número ideal é de 14 componentes principais. Para isso foram identificados 929 *lncRNAs* presentes no estudo e foi gerado o *UMAP* para verificar a distribuição das diferentes células. Na Figura 21 observa-se que o comportamento continua parecido quando comparado com todos os genes (Figura 17; as células malignas aparecem bem diferenciadas em relação às não malignas, enquanto que os subtipos das não malignas ficam todas homogêneas, sendo possível diferenciar visualmente apenas a subpopulação de células *CAF* representadas na cor verde.

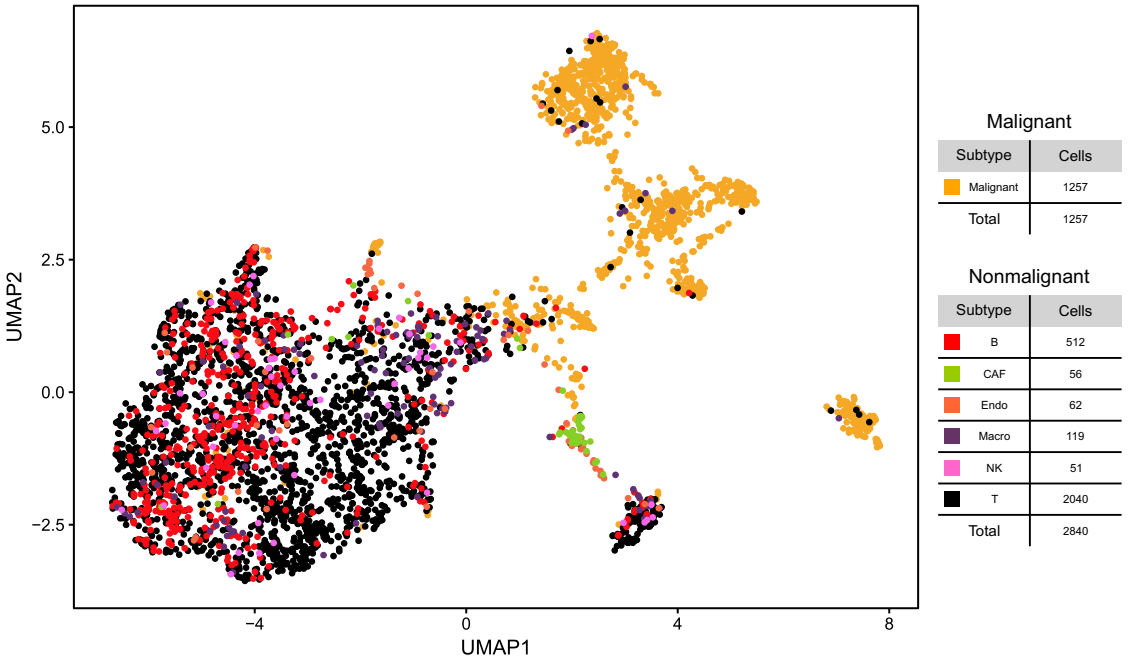

Figura 21 – Redução da dimensão utilizando a técnica *UMAP* com a anotação das células original para a visualização de todas as células e somente os *lncRNAs* (929 totais identificados).

Foram identificados também os *lncRNAs* marcadores para cada grupo conforme mostrados na Tabela 3. O principal marcador para as células malignas é o *SEMA3B* que quando comparado com dados de *bulk* de *RNA-seq* mostram que realmente há um maior valor de expressão nos tecidos tumorais em relação aos normais em melanoma (*SKCM*), além de colangiocarcinoma (*CHOL*), câncer de cabeça e pescoço (*HNSC*) e timoma (*THYM*) (Figura 22); os demais tumores possuem maior expressão de tecidos normais em relação aos tumorais ou expressões equivalentes de ambos os tecidos. De acordo com Rolny *et al.* (2008), este gene é comumente expresso em células cancerígenas humanas, principalmente em melanoma.

Tabela 3 – Principais *lncRNAs* marcadores (diferencialmente expressos) para cada tipo e subtipo celular quando comparados contra todos os outros utilizando o teste de *Wilcox*. A coluna *pct.1* e *pct.2* representam, respectivamente, a fração de expressão das células que o determinado marcador possui.

Markers	avg_logFC	pct.1	pct.2	p_val_adj	Grupo
VAV3-AS1	3,2349881	0,107	0,002	8,74E-72	B (4)
MEG3	4,9888451	0,75	0,01	0,00E+00	CAF (14)
LINC00636	3,5910745	0,194	0,002	3,37E-105	Endo (6)
FAM157B	3,2885554	0,218	0,02	2,32E-39	Macro (6)
SEMA3B	4,4967642	0,443	0,007	9,65E-297	Malignant (24)
LINC00299	2,5880598	0,137	0,005	5,49E-26	NK (1)
MIAT	2,9236567	0,281	0,018	3,77E-121	T (8)

Figura 22 – Expressão do *lncRNA SEMA3B* em todos os tipos tumorais disponíveis no projeto *TCGA*.

A fim de identificar os subgrupos presentes nas células malignas quando se é observado apenas os *lncRNAs*, foi realizada uma clusterização não supervisionada obtendo-se 9 *clusters*. Essa nova divisão pode ser observada na Figura 23. Os *clusters* 3, 4, 6 e 7 corresponde às células malignas, enquanto os demais às células não malignas. Evidentemente, pelo fato de termos populações bem restritas e com um número pequeno de *lncRNAs* os marcadores aparecem em menor quantidade. Apesar disso, a maioria dos marcadores identificados (Tabela 4) não são descritos na literatura como marcadores da progressão ou supressão de melanoma, bem como para tipos e subtipos celulares específicos. No entanto há um *cluster* específico que chama a atenção, que é o *cluster* 6.

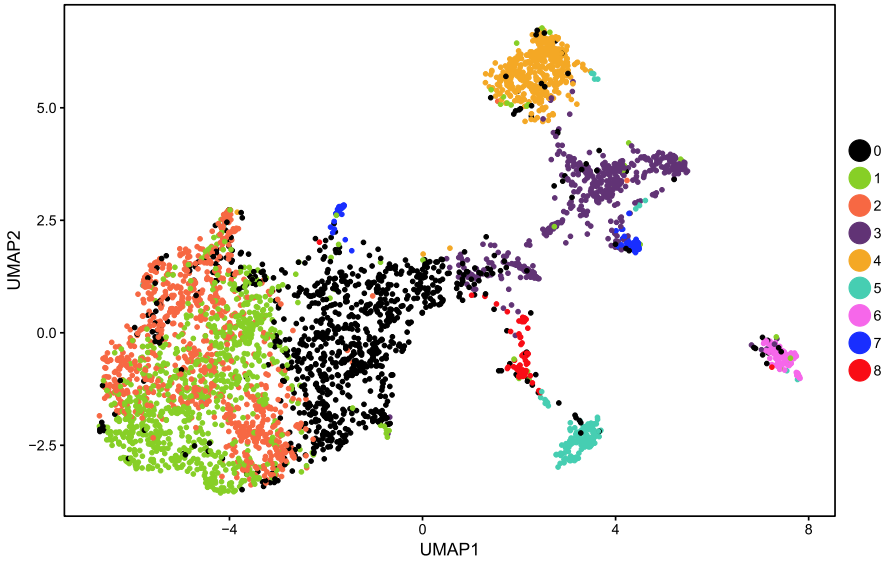

Figura 23 – Redução dimensional utilizando a técnica *UMAP* após a clusterização não supervisionada para a visualização de todas as células e somente os *lncRNAs*.

Tabela 4 – Principais *lncRNAs* marcadores (diferencialmente expressos) para cada grupo quando comparados contra todos os outros utilizando o teste de *Wilcox*. A coluna *pct.1* e *pct.2* representam, respectivamente, a fração de expressão das células que o determinado marcador possui para o grupo em questão e a porcentagem de expressão das células que o marcador possui para os outros grupos. Na coluna grupo temos entre parênteses a quantidade total de marcadores únicos para o determinado grupo.

Markers	avg_logFC	pct.1	pct.2	p_val_adj	Grupo
MIR155HG	1,7556994	0,271	0,083	8,55E-54	cluster0 (1)
PRKCQ-AS1	1,1176551	0,164	0,059	1,03E-21	cluster1 (1)
STAU2-AS1	1,669302	0,311	0,034	9,18E-132	cluster2 (35)
WDR11-AS1	1.623837	0,22	0,017	2,96E-107	cluster2 (35)
ZBED3-AS1	1,9777519	0,174	0,009	5,44E-92	cluster3 (5)
PHKA2-AS1	1,8109262	0,119	0,012	3,39E-44	cluster3 (5)
FAM224A	2,7285421	0,185	0,008	1,77E-101	cluster4 (21)
LINC00511	2,3866470	0,235	0,025	2,98E-82	cluster4 (21)
SLX1A-SULT1A3	3,942692	1	0,018	0,00E+00	cluster5 (3)
SLX1A-SULT1A4	3,942692	1	0,018	0,00E+00	cluster5 (3)
HOTAIR	3,1318278	0,396	0,002	1,68E-284	cluster6 (14)
TRHDE-AS1	2,4485587	0,297	0,006	2,00E-144	cluster6 (14)
LINC00479	4,0475652	0,626	0,002	0	cluster7 (41)
LINC00112	3,0316180	0,232	0	2,19E-193	cluster7 (41)
MEG3	4,7822417	0,623	0,008	0	cluster8 (26)
H19	4,2677202	0,286	0,005	1,21E-130	cluster8 (26)

4.1.3 O *RNA* longo não codificante *HOTAIR* e a Transição Epitélio-mesenquimal (*EMT*)

O *cluster* 6, observado na Figura 23, possui como melhor marcador o gene *HOTAIR*. De acordo com a Figura 24 é possível identificar que o *HOTAIR* possui uma alta expressão em algumas células que estão circuladas em destaque. Na Figura 24A é possível observar a distribuição da expressão do *HOTAIR* quando levado em consideração todos os genes. A subpopulação em que há uma superexpressão é equivalente ao *cluster* 8 que é identificado na Figura 19. Este *cluster* possui 166 células somente malignas, sendo que 45 dessas possuem a alta expressão de *HOTAIR*. Em contrapartida, na Figura 24B observa-se que a subpopulação que possui uma superexpressão de *HOTAIR* é equivalente ao *cluster* 6 identificado na Figura 23. Este *cluster* possui 111 células também apenas malignas, sendo que 44 dessas possuem alta expressão do *HOTAIR*. Verifica-se que há uma intersecção de 40 células ($\approx$ 0,98% das células totais) para ambos os *clusters* que possuem alta expressão de *HOTAIR*. Nota-se na Figura 25 a expressão do *HOTAIR* nas células malignas (em *ALL Original* e *lncRNAs Original*) e nas células dos *clusters* 8 e 6, *ALL Clustering* e *lncRNAs Clustering*, respectivamente. Nas comparações originais a expressão de *HOTAIR* se mantém baixa com alguns *outliers* com uma expressão elevada. É interessante notar, mais uma vez, que quando ocorre a identificação dos *clusters* esse cenário é alterado provendo assim uma alta expressão de *HOTAIR* como observado em *ALL Clustering* e *lncRNAs Clustering*.

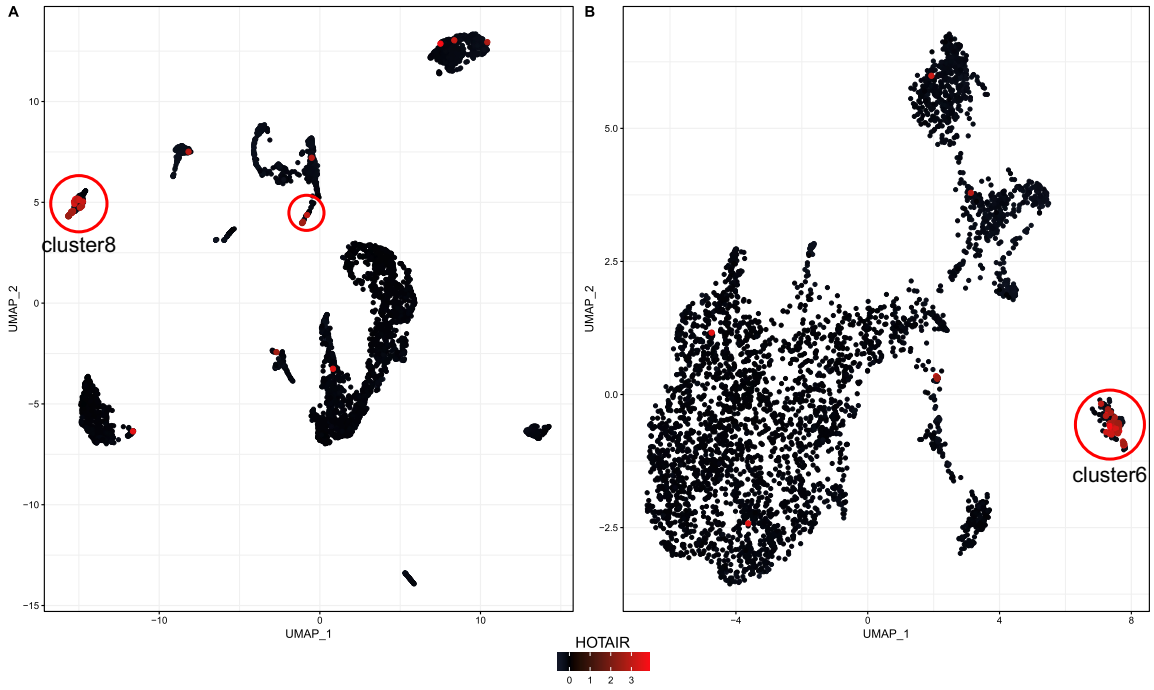

Figura 24 – *UMAP* da distribuição das células para todos os genes (*A*) e somente para os *lncRNAs* (*B*) mostrando a expressão do *HOTAIR*. As subpopulações que possuem alta expressão de *HOTAIR* estão circuladas em destaque juntamente com a indicação do *cluster* correspondente.

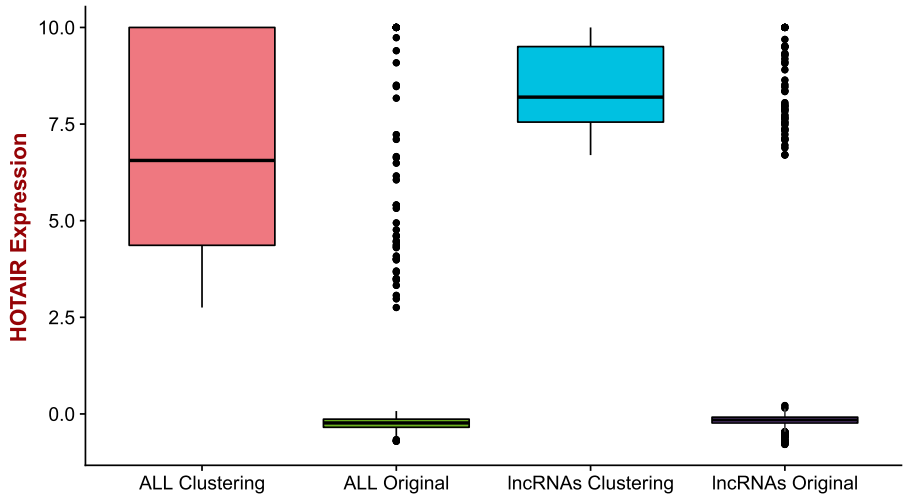

Figura 25 – *Boxplot* mostrando a expressão de *HOTAIR* quando usado todos os genes e apenas os *lncRNAs*, com anotação original ou clusterizando.

O que foi identificado para o *HOTAIR* em relação a outros genes propostos por Alves *et al.* (2013) foi para *bulk de RNA*. Dessa forma, é importante verificar se o comportamento se mantém ou se é possível refiná-lo para dados de células únicas (*scRNA-seq*) no caso do Melanoma. Como relatado anteriormente, a alta expressão de *HOTAIR* ocasiona, após vários processos, o silenciamento de *CDH1*. A partir da ferramenta *MAGIC* foi possível

gerar gráficos a fim de mostrar o comportamento da expressão do *HOTAIR* e *CDH1* quando comparado contra cada um dos genes relacionados à *EMT* e *stemness* presentes na Figura 9. Interpreta-se a Figura 26 e Figura 27 fixando no eixo X a expressão do gene *CDH1* e a coloração dos pontos a expressão do gene *HOTAIR*. O perfil de alta expressão de *HOTAIR* e baixa expressão de *CDH1* faz com que ocorra a alta expressão de alguns genes, como por exemplo o *VIM*, *SNAI1*, *FN1* e *BMP1* (Figura 26) e a baixa expressão de alguns genes, como por exemplo o *NANOG*, *POU5F1*, *ERBB3* e *GSK3B* (Figura 27).

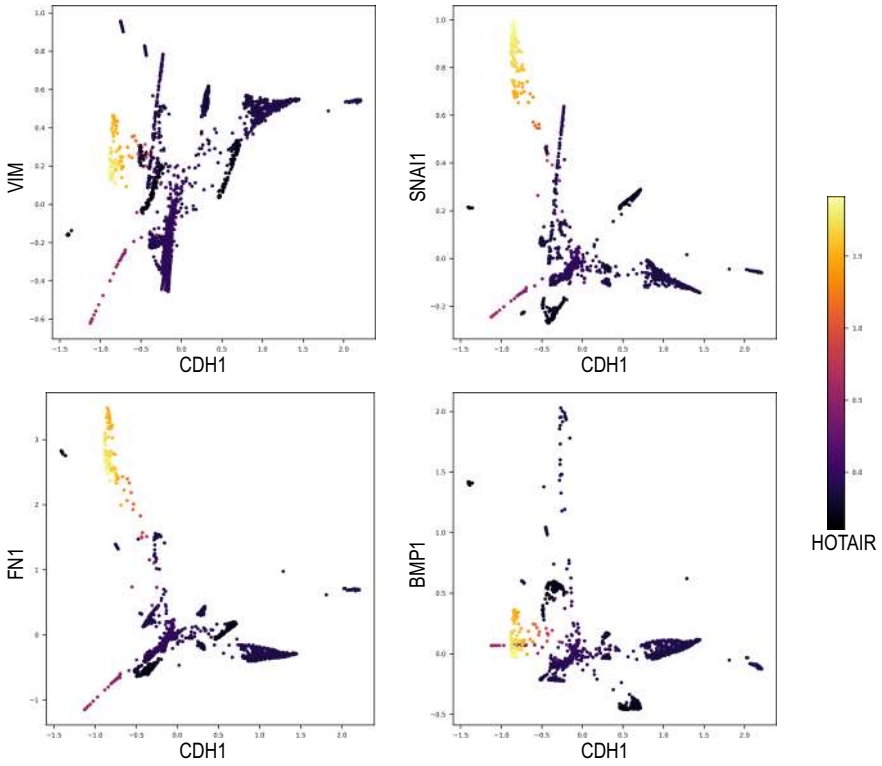

Figura 26 – Visualização da expressão do *HOTAIR* (coloração de cada ponto), *CDH1* (eixo x) e de alguns genes que possuem alta expressão, como *VIM*, *SNAI1*, *FN1* e *BMP1* no eixo y.

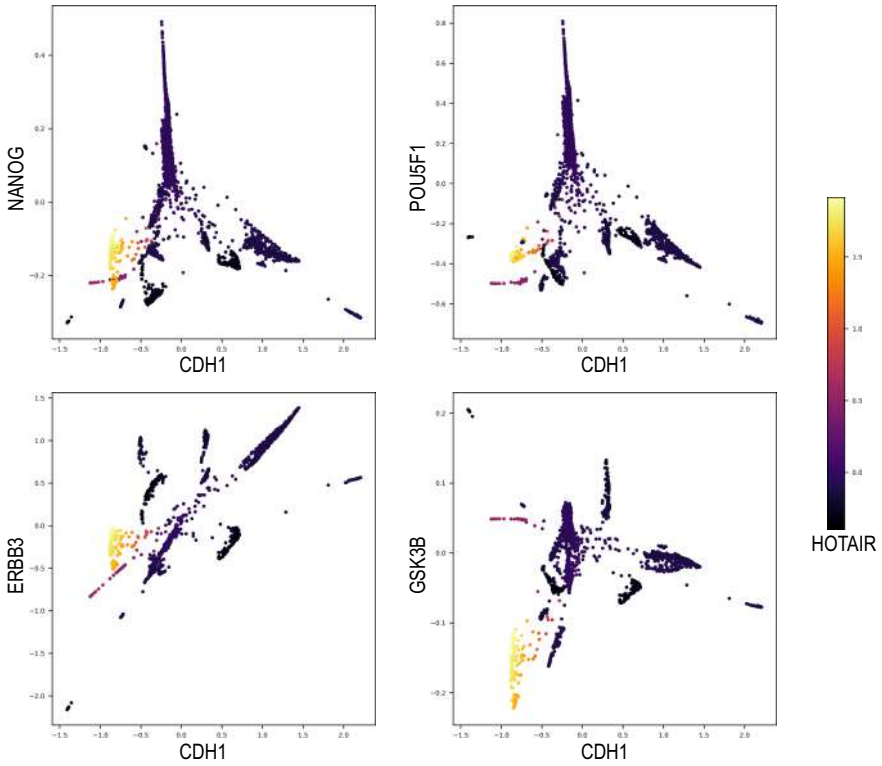

Figura 27 – Visualização da expressão do *HOTAIR* (coloração de cada ponto), *CDH1* (eixo x) e de alguns genes que possuem baixa expressão, como *NANOG*, *POU5F1*, *ERBB3* e *GSK3B* no eixo y.

4.1.4 O *RNA* longo não codificante *TRHDE-AS1*

Verifica-se que há também outro *RNA* longo não codificante com o perfil de expressão muito parecido com o do *HOTAIR* para a mesma subpopulação de células no *cluster* 6 (vide Tabela 4). Trata-se do *TRHDE-AS1*, que é um *RNA antisense* do gene *TRHDE* e não possui descrição da sua função específica. Na Figura 28 observa-se a expressão do *HOTAIR* e do *TRHDE-AS1* na mesma subpopulação. A Figura 28A e Figura 28B mostram a expressão de *HOTAIR* e *TRHDE-AS1*, respectivamente, usando a conformação do *UMAP* para todos os genes. Enquanto a Figura 28C e Figura 28D mostram também a expressão de *HOTAIR* e *TRHDE-AS1*, respectivamente, usando a conformação do *UMAP* apenas para os *lncRNAs*.

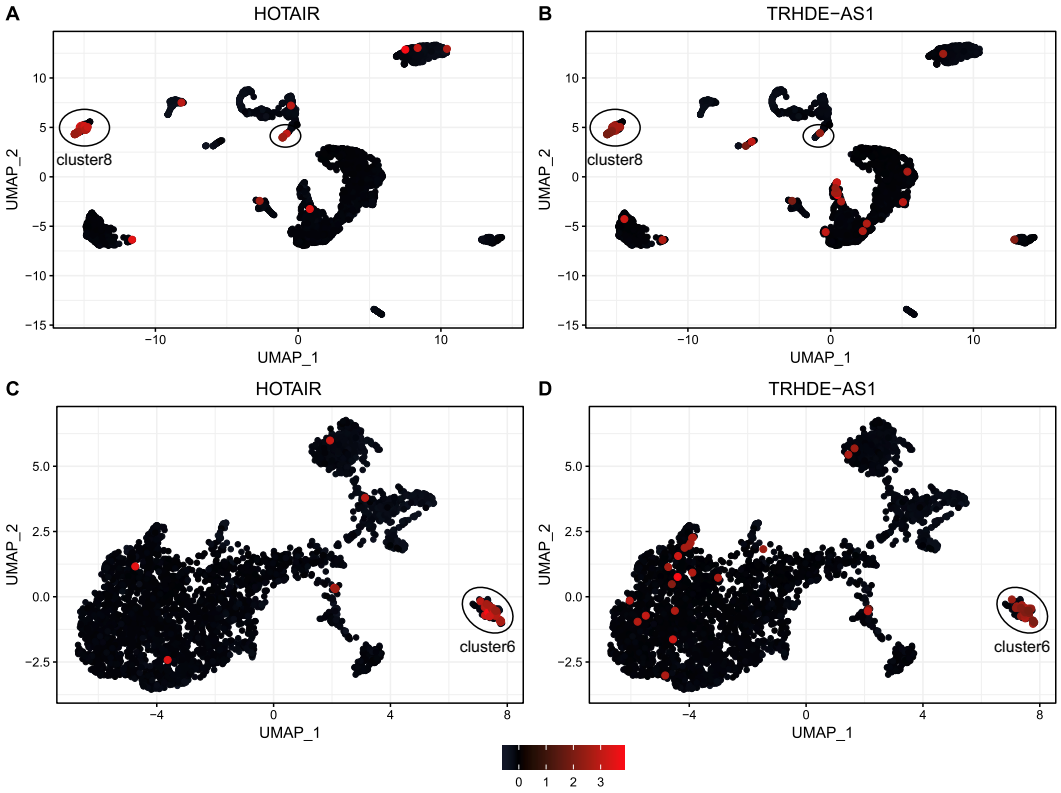

Figura 28 – Comparação da expressão dos *lncRNAs HOTAIR* e *TRHDE-AS1*. As figuras *A* e *B* mostram a expressão de *HOTAIR* e *TRHDE-AS1*, respectivamente, usando a conformação estrutural do *UMAP* para todos os genes. Enquanto as figuras *C* e *D* mostram também a expressão de *HOTAIR* e *TRHDE-AS1*, respectivamente, usando a conformação estrutural do *UMAP* apenas para os *lncRNAs*.

A fim de detalhar a subpopulação que possui alta expressão de *HOTAIR* e *TRHDE-AS1*, isolamos essas células e realizamos uma nova clusterização. Com isso foi possível identificar 3 novos sub *clusters* que possuem alta expressão de *HOTAIR* / baixa expressão de *TRDE-AS1*, alta expressão de *HOTAIR* / alta expressão de *TRHDE-AS1* e outro *cluster* que possui baixa expressão de *HOTAIR* / baixa expressão de *TRHDE-AS1*, como pode ser observado na Figura 29, sendo a correlação de 0,953 entre as células que coexpressam *HOTAIR* e *TRHDE-AS1* (Figura 30). Tais fatos evidenciam que o perfil de expressão do *HOTAIR* e *TRHDE-AS1* são muito parecidos e que eles podem estar agindo em conjunto para a ativação da via *EMT*.

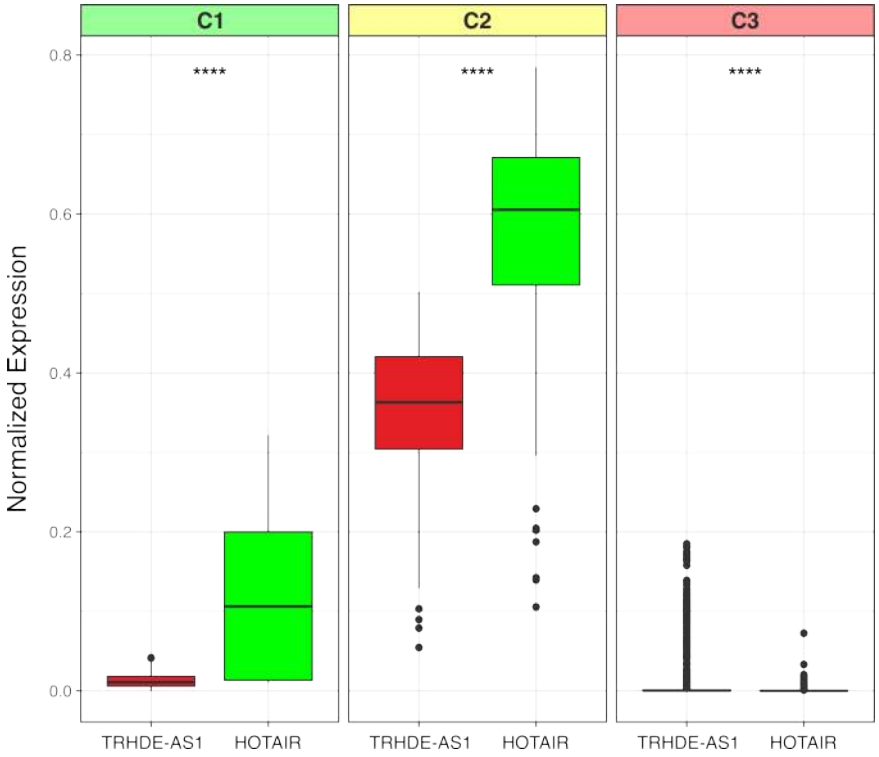

Figura 29 – *Boxpot* com a expressão de *HOTAIR* e *TRHDE-AS1* nos 3 sub *clusters* (C1, C2 e C3). Em vermelho encontra-se a expressão do *TRHDE-AS1* e em verde a expressão do *HOTAIR*.

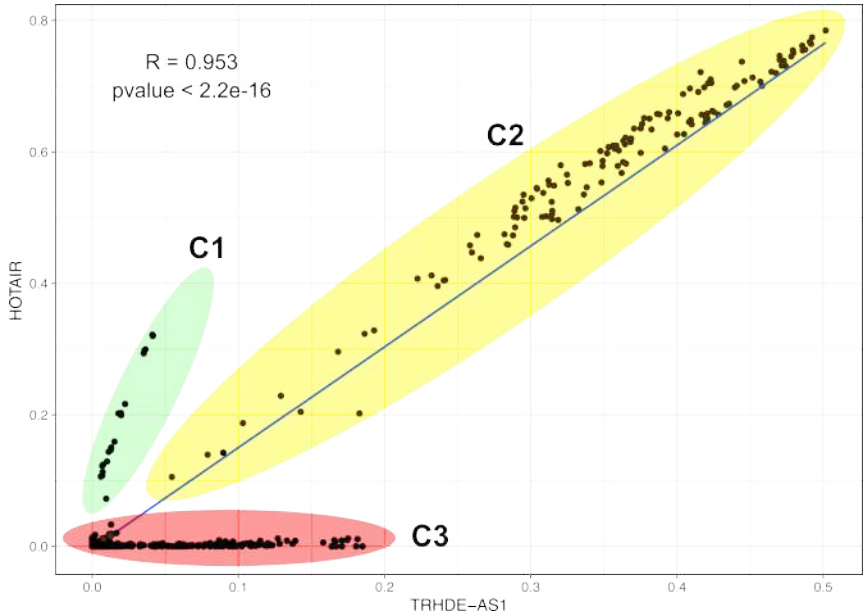

Figura 30 – Gráfico de correlação entre a expressão do gene *TRHDE-AS1* (eixo x) em relação a expressão do gene *HOTAIR* (eixo y). Os sub *clusters* estão discriminados em C1, C2 e C3, sendo uma correlação de 0,953 entre as células coexpressando HOTAIR e TRHDE-AS1.

Baseado nos dados descritos na Figura 29 e Figura 30, propomos um circuito de *feedback* negativo. Hipotetizamos que a *up* regulação do *HOTAIR* (*cluster* C2 na Figura 30) ativa o *TRHDE-AS1*. No entanto, quando o *TRHDE-AS1* atinge um certo nível de ativação, ele começa a regular negativamente o *HOTAIR* (*cluster* C3 na Figura 30). O *cluster* C1 (Figura 30) é um estado em que o *HOTAIR* inicia a ativação do *TRHDE-AS1*. Em resumo, o *HOTAIR* regula positivamente o *TRHDE-AS1*, e o *TRHDE-AS1* regula negativamente o *HOTAIR*. A Figura 31 mostra uma representação esquemática do mecanismo de *feedback* negativo entro os *lncRNAs HOTAIR* e *TRHDE-AS1*.

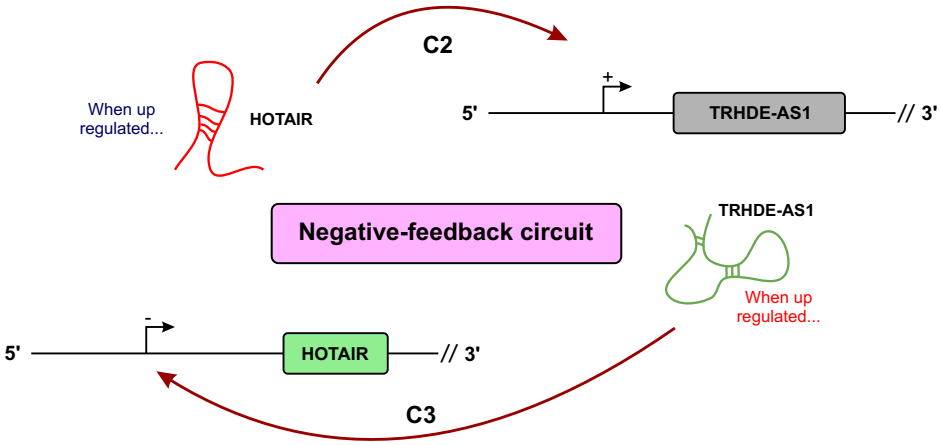

Figura 31 – Circuito de *feedback* negativo onde o *HOTAIR* regula positivamente o *TRHDE-AS1*, e o *TRHDE-AS1* regula negativamente o *HOTAIR*

A fim de validar este mecanismo proposto foi realizado o silenciamento tanto de *TRHDE-AS1* quanto do *HOTAIR* pela metodologia *siRNA* e foram mensurados os níveis de expressão de ambos os genes. Na Figura 32 é possível observar os resultados da validação. Quando o *TRHDE-AS1* é silenciado a expressão de *HOTAIR* aumenta em ambas as linhagens de melanoma primário (Figura 32A) e melanoma metastático (Figura 32C). Por outro lado, quando o *HOTAIR* é silenciado a expressão de *TRHDE-AS1* diminui em ambas as linhagens de melanoma primário (Figura 32B) e melanoma metastático (Figura 32D). Estes dados trazem suporte a hipótese de *feedback* negativo entre *HOTAIR* e *TRHDE-AS1* representado na Figura 31.

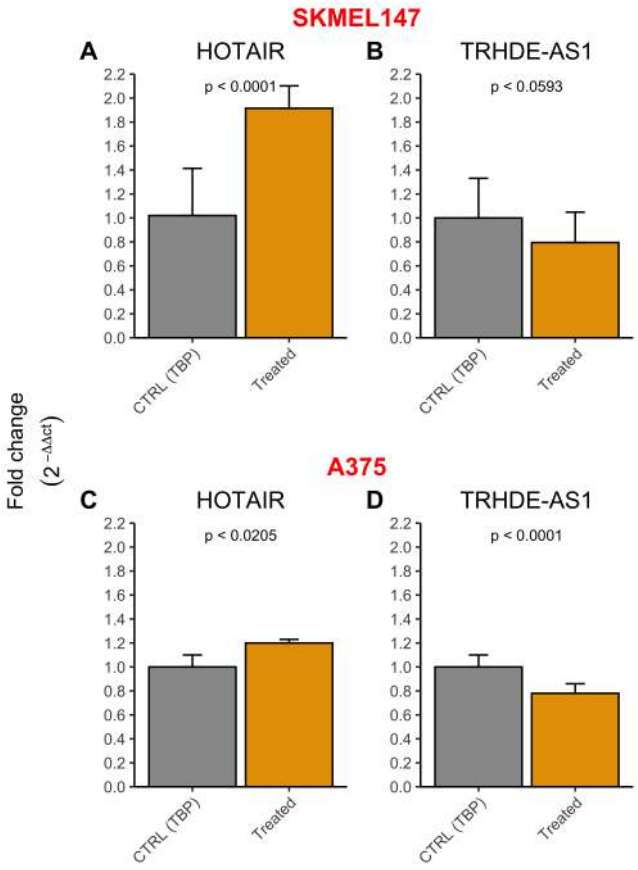

Figura 32 – Resultados do silenciamento dos genes *HOTAIR* e *TRHDE-AS1*. Os valores representam o *fold change* ($2^{-\Delta\Delta ct}$) e o desvio padrão é mostrado acima das barras. (A) e (C) correspondem à expressão do *HOTAIR* com o *TRHDE-AS1* silenciado, enquato (B)e (D) correspondem à expressão do *TRHDE-AS1* com o *HOTAIR* silenciado.

Quando visualizamos o *TRHDE-AS1* utilizando a plataforma online *Genome Browser* (KENT *et al.*, 2002) identificamos que ele possui 3 isoformas (Figura 33). A presença do sítio de ligação do fator de transcrição *EZH2* sugere que o *HOTAIR*, através do complexo repressivo *policomb* 2 (*PCR2*), esteja regulando o *TRHDE-AS1*. O *HOTAIR* não codifica proteína, mas foi demonstrado que está associado ao complexo repressivo *polycomb* 2 (*PRC2*) que é constituído pela metilase *H3K27*, *EZH2*, *SUZ12* e *EED* (GUPTA *et al.*, 2010; SIMON; KINGSTON, 2009; TSAI *et al.*, 2010). As proteínas que constituem o grupo *polycomb* medeiam a repressão da transcrição de muitos genes que controlam as vias de diferenciação durante o desenvolvimento, e têm papel importante na pluripotência das células tronco e câncer (GIBB; BROWN; LAM, 2011). Isso evidencia, mais uma vez, que o *TRHDE-AS1* possui uma forte relação com o *HOTAIR*.

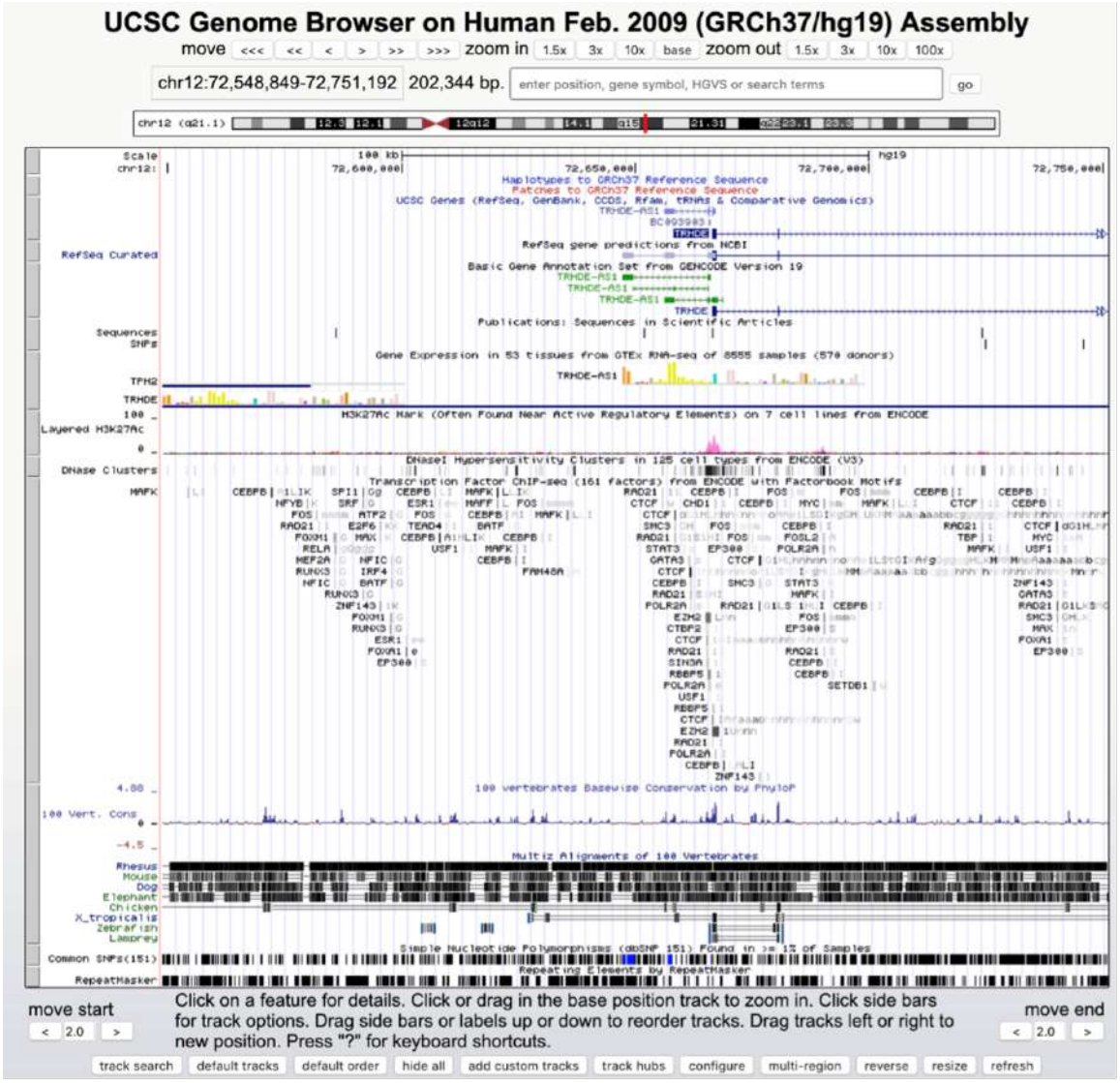

Figura 33 – Visualização do *TRHDE-AS1* no *Genome Browser*.

Também foi realizada a predição de alvos dos *lncRNAs TRHDE-AS1* e *HOTAIR* utilizando o banco de dados rtools (TERAI *et al.*, 2016). Foram considerados como alvos os genes que possuem a menor soma de energia, e com isso foi possível identificar 387 alvos em comum, 93 alvos exclusivos do *HOTAIR* e 99 exclusivos do *TRHDE-AS1*. Com essa informação geramos uma rede de interações representada na Figura 34. Juntamente, foi realizado o enriquecimento funcional utilizando o *KEGG*, e foi identificada algumas vias interessantes (Figura 35). Por exemplo, para os alvos da intersecção está enriquecida a via de sinalização que regula a pluripotência das células-tronco, podendo estar relacionada com as características das células-troco tumorais que iniciam a *EMT*. Para os alvos exclusivos de *HOTAIR* ou *TRHDE-AS1* há vias relacionadas à adesão focal, câncer endometrial, câncer de próstata, câncer de mama etc.

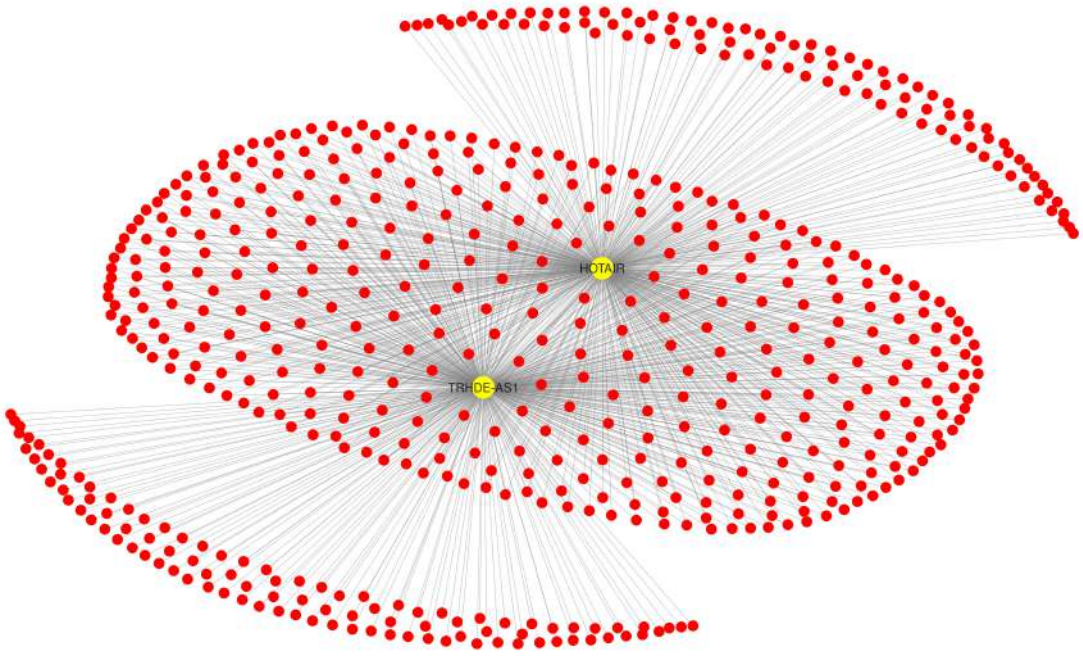

Figura 34 – Rede de interações entre os alvos dos *lncRNAs HOTAIR* e *TRHDE-AS1*.

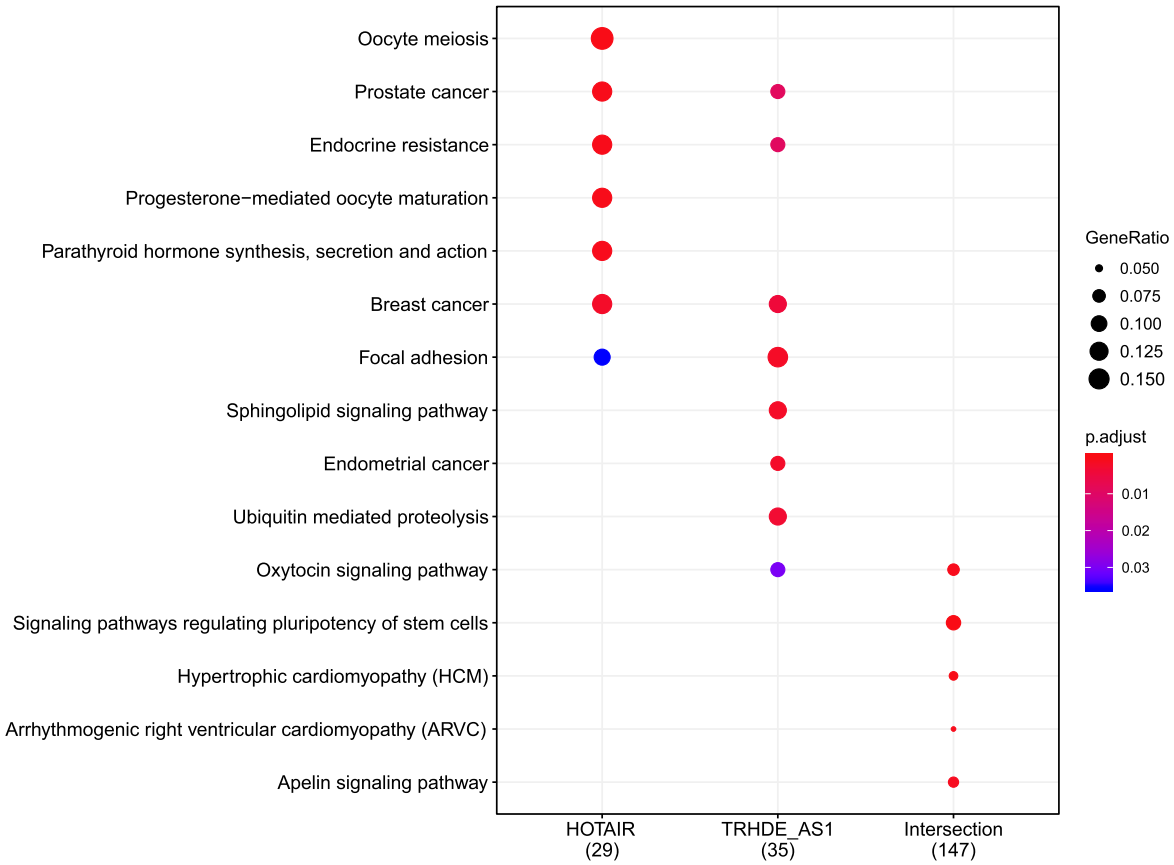

Figura 35 – Enriquecimento dos alvos exclusivos do *lncRNA HOTAIR*, exclusivos do *lncRNA TRHDE-AS1* e alvos da intersecção.

Por fim, uma revisão na literatura foi capaz de mostrar alguns trabalhos que demonstram a relevância do *TRHDE-AS1* em diferentes tipos tumorais. O trabalho de Zhuan *et al.* (2019) menciona diretamente o *TRHDE-AS1*, onde demonstra que uma super expressão do *TRHDE-AS1* inibe a progressão do câncer de pulmão através do eixo *miRNA-103/KLF4*. Wu *et al.* (2021) identificou o *TRHDE-AS1* como potencial biomarcador de prognóstico no câncer gástrico, onde a sua upregulação do inibe o crescimento do carcinoma pulmonar através da combinação competitiva com o eixo *miRNA-103-KLF4*. Hu, Zheng e Jin (2021) mostra que a baixa expressão de *TRHDE-AS1* está associada a prognósticos negativos em pacientes com câncer de mama e contribui potencialmente para a biologia tumoral agressiva do câncer de mama. Yang *et al.* (2016) mostra que em câncer de mama, em dados do *TCGA*, o *HOTAIR* possui uma alta expressão em tecidos cancerosos, enquanto o *TRHDE-AS1* possui maior expressão nos tecidos que não são cancerosos. Tal fato é muito interessante, pois pode evidenciar um comportamento exclusivo para o câncer de mama em dados de *RNA-seq* ou que há um perfil antagônico entre *HOTAIR* e *TRHDE-AS1*. Outros estudos foram encontrados onde não havia um foco principal no *TRHDE-AS1*, mas que mostram como o *TRHDE-AS1* está expresso em diferentes tecidos utilizando diferentes tecnologias (AHN *et al.*, 2016; DONG *et al.*, 2019; GUPTA *et al.*, 2016; LIU *et al.*, 2017; PANG *et al.*, 2019; SHANG *et al.*, 2016; XING *et al.*, 2018; WEI *et al.*, 2021).

4.1.5 Enriquecimento de assinaturas de subpopulações de células únicas de melanoma

Foram utilizadas algumas assinaturas a fim de verificar quais células possuem maior enriquecimento. Essas assinaturas estão relacionadas a células tronco, *stemness*, invasão, proliferação etc. A lista de assinaturas utilizadas pode ser observada na Tabela 5 com suas respectivas informações de código, nome, número de genes e a referência.

Tabela 5 – Lista de assinaturas utilizadas com seus respectivos códigos, nomes, número de genes e referência.

Code	Name	Genes	Reference
ES1	Embryonic Stem Cell 1	342	Ben-Porath *et al.* (2008)
ES2	Embryonic Stem Cell 2	30	Ben-Porath *et al.* (2008)
NANOG	NANOG targets	872	Ben-Porath *et al.* (2008)
OCT4	OCT4 targets	265	Ben-Porath *et al.* (2008)
SOX2	SOX2 targets	654	Ben-Porath *et al.* (2008)
NOS	NOS targets	161	Ben-Porath *et al.* (2008)
NOS_TFs	NOS Transcrit Factors	36	Ben-Porath *et al.* (2008)
SUZ12	SUZ12 targets	945	Ben-Porath *et al.* (2008)
EED	EED targets	971	Ben-Porath *et al.* (2008)
H3K27	H3K27	1022	Ben-Porath *et al.* (2008)
PCR2	PCR2 targets	596	Ben-Porath *et al.* (2008)
MYC1	MYC targets 1	221	Ben-Porath *et al.* (2008)
MYC2	MYC targets 2	725	Ben-Porath *et al.* (2008)
Invasive	Invasive	45	Widmer *et al.* (2012)
Proliferative	Proliferative	51	Widmer *et al.* (2012)
EMT1	Epithelial–mesenchymal transition 1	18	-
EMT2	Epithelial–mesenchymal transition 2	15	-
EMT3	Epithelial–mesenchymal transition 3	82	-
SCP	Stem Cel Pathway	67	-
AME_UP	Alonso Metastasis EMT UP	36	Alonso *et al.* (2007)
AMI	Anastassiou Multicancer Invasiveness	64	Anastassiou *et al.* (2011)
CLD_DN	Cleidson DOWN	13	Alves *et al.* (2013)
CLD_UP	Cleidson UP	20	Alves *et al.* (2013)
HEMT	Hallmark EMT	200	Liberzon *et al.* (2015)
SCTFP	Stem Cell Transcription Factor Pathway	5	Ben-Porath *et al.* (2008)

A partir das assinaturas foi realizado o enriquecimento das 4097 células de melanoma. O *GSVA* atribui um *score* para cada célula individualmente baseado em uma assinatura específica. Na Figura 36 observa-se o *heatmap* com as 25 assinaturas citadas na Tabela 5 e seus respectivos enriquecimentos para cada célula e cada subtipo celular.

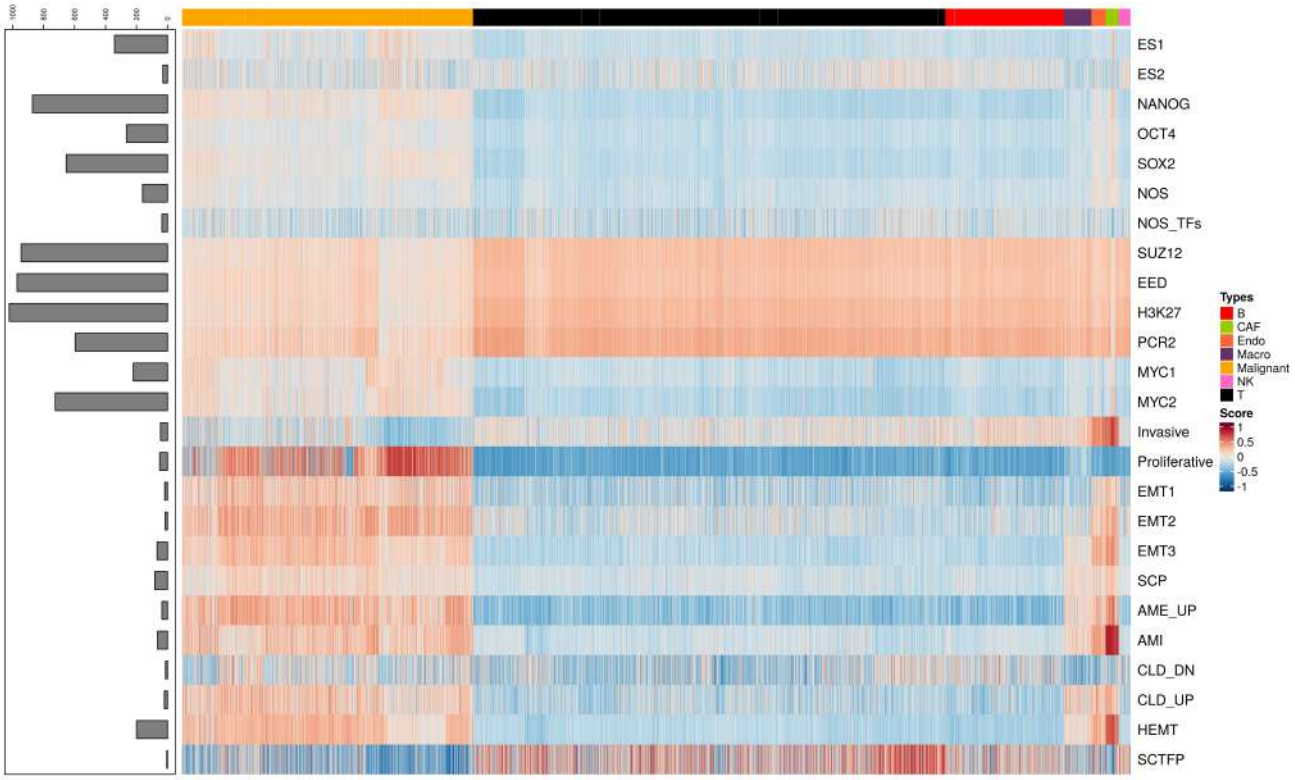

Figura 36 – *Heatmap* representando o resultado do enriquecimento utilizando o *GSVA*, onde cada linha é uma assinatura, cada coluna é uma célula com sua respectiva anotação acima, e o *barplot* à esquerda mostra o número de genes que cada assinatura possui. Os códigos das assinaturas podem ser obtidos através da Tabela 5.

O perfil geral das assinaturas é representado com um maior *score* de enriquecimento (tons de vermelho) nas células malignas e menor enriquecimento nas células não malignas (tons de azul). Destacando-se:

- Assinaturas *NANOG*, *OCt4*, *SOXC2*, *NOS_TFs*: ão constituídas de genes cujos promotores são ligados e ativados em células tronco embrionárias humanas e por genes alvos de ativação de *NOS* que codificam reguladores de transcrição, possuem um leve enriquecimento das células malignas em relação as células não malignas;
- Assinaturas *SUZ12*, *EED*, *H3K27* e *PCR2*: representam alvos de *polycomb* e possuem genes ligados pelo complexo repressivo de *polycomb* 2 em células embrionárias humanas, possuem um perfil onde a maioria das células estão com um enriquecimento alto, tanto para as células malignas e não malignas;
- Assinaturas de Invasão e Proliferação: genes associados à invasão e proliferação em melanoma metastático;

- As demais assinaturas estão relacionadas à *EMT* ou *HOTAIR*, e seus respectivos enriquecimentos estão com scores mais altos nas células malignas;
- Algumas assinaturas como as de invasão, proliferação e as demais relacionados a *EMT*, além de possuírem maior enriquecimento para as células malignas possuem também o grupo de células *CAF* e Endoteliais muito enriquecidas.

A fim de corroborar o enriquecimento das assinaturas feito pelo *GSVA*, foi utilizada outra metodologia para enriquecer as assinaturas baseada na estatística da área sob a curva. Foram utilizados os mesmos exemplos de assinaturas e os resultados foram muito consistentes com os mostrados anteriormente.

Quando verificamos o enriquecimento de algumas assinaturas relacionadas à via de *EMT* (Figura 37), tais como *AME_UP*, *AMI*, *EMT3* e *HEMT*, observamos que a população de células que possui alta expressão de *HOTAIR* também possui alto enriquecimento para *EMT*.

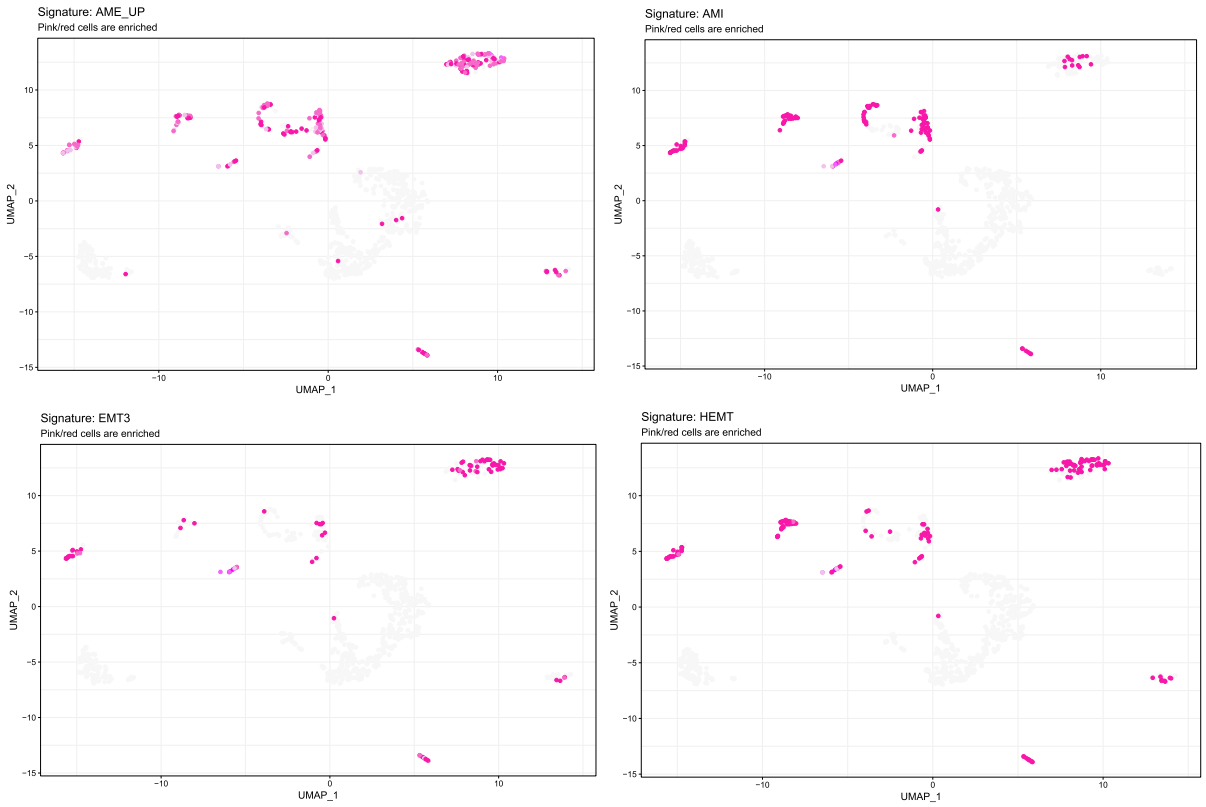

Figura 37 – Enriquecimento de algumas vias relacionadas a *EMT* verificando que há um alto enriquecimento na subpopulação de células com alta expresão de *HOTAIR*, como mostrado anteriormente.

4.1.6 Proposta de assinaturas para tipos celulares específicos e via *EMT*

A partir da análise de expressão diferencial utilizando o pacote *Seurat*, foi possível identificar genes marcadores para cada tipo celular, e com isso definir assinaturas. Foram geradas assinaturas para as células malignas, B, T, *CAF*, *NK*, Endoteliais, Macrofagiais e duas para a identificação da via *EMT* (uma utilizando os marcadores e outra utilizando a correlação entre os *lncRNAs HOTAIR* e *TRHDE-AS1*). A lista de genes que compõe cada assinatura pode ser observada no Apêndice A.

Observa-se na Figura 38 as assinaturas e sua distribuição de enriquecimento, utilizando a metodologia *GSVA*, para todas as células de melanoma. As assinaturas para células B, T, Macrofagiais, Malignas e *EMT* estão bem definidas de acordo com a metodologia do *GSVA*, enquanto as assinaturas para as células endoteliais, NK e CAF possuem alto enriquecimento para outros tipos celulares também.

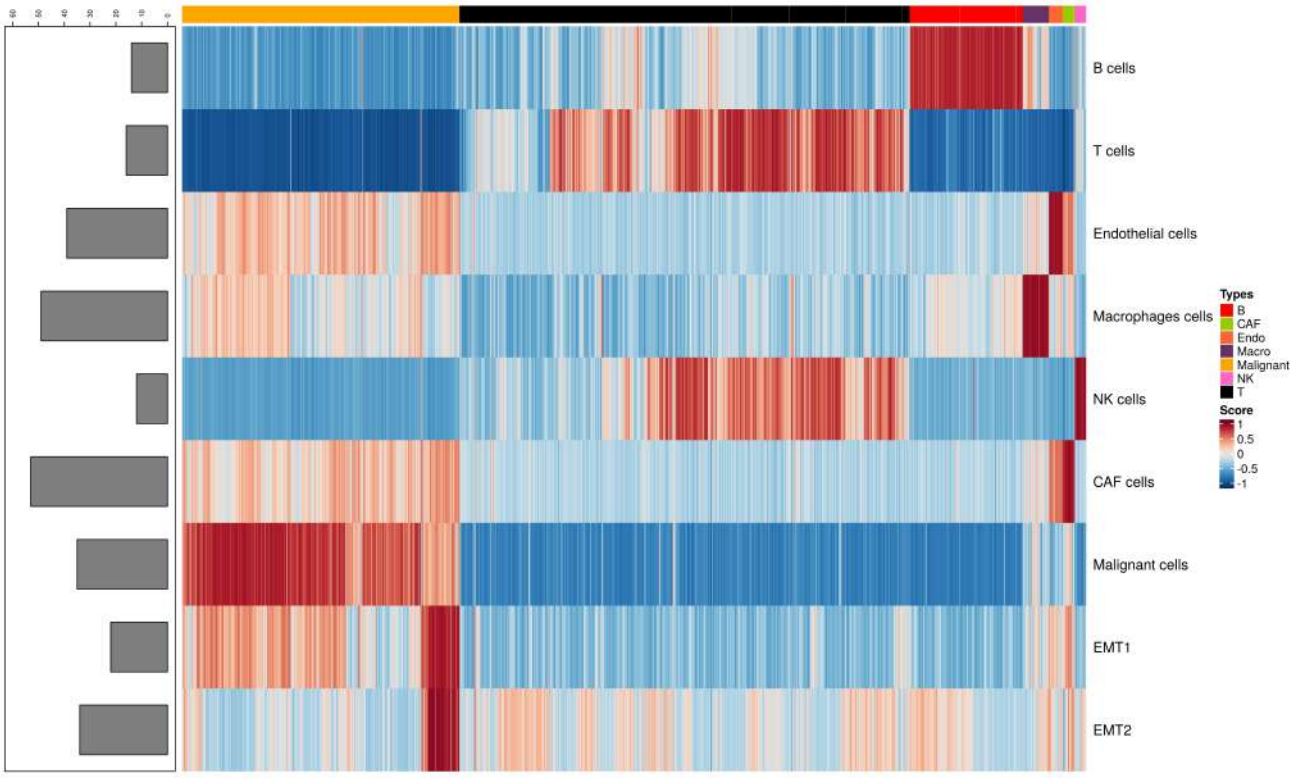

Figura 38 – *Heatmap* com as assinaturas propostas para os tipos celulares e via *EMT*.

Enriquecendo essas assinaturas utilizando a metodologia *AUCell*, foi possível identificar, a partir da visualização *UMAP*, se realmente as assinaturas são eficientes ou não. Observa-se que as assinaturas para as células B (Figura 39A), T (Figura 39B), macrofagiais

(Figura 39D), malignas (Figura 39G), *EMT1* (Figura 39H) e *EMT2* (Figura 39I) possuem ótimos marcadores para identificar os respectivos tipos celulares e/ou a via de *EMT* para melanoma, enquanto que as assinaturas para as células endoteliais (Figura 39C), *NK* (Figura 39E) e *CAF* (Figura 39F) não são exclusivas, e precisam de uma otimização a fim de ser possível identificar apenas o tipo celular em questão. É interessante notar que a Figura 38 e Figura 39 apesar de se utilizarem de métodos diferentes, trazem informações complementares.

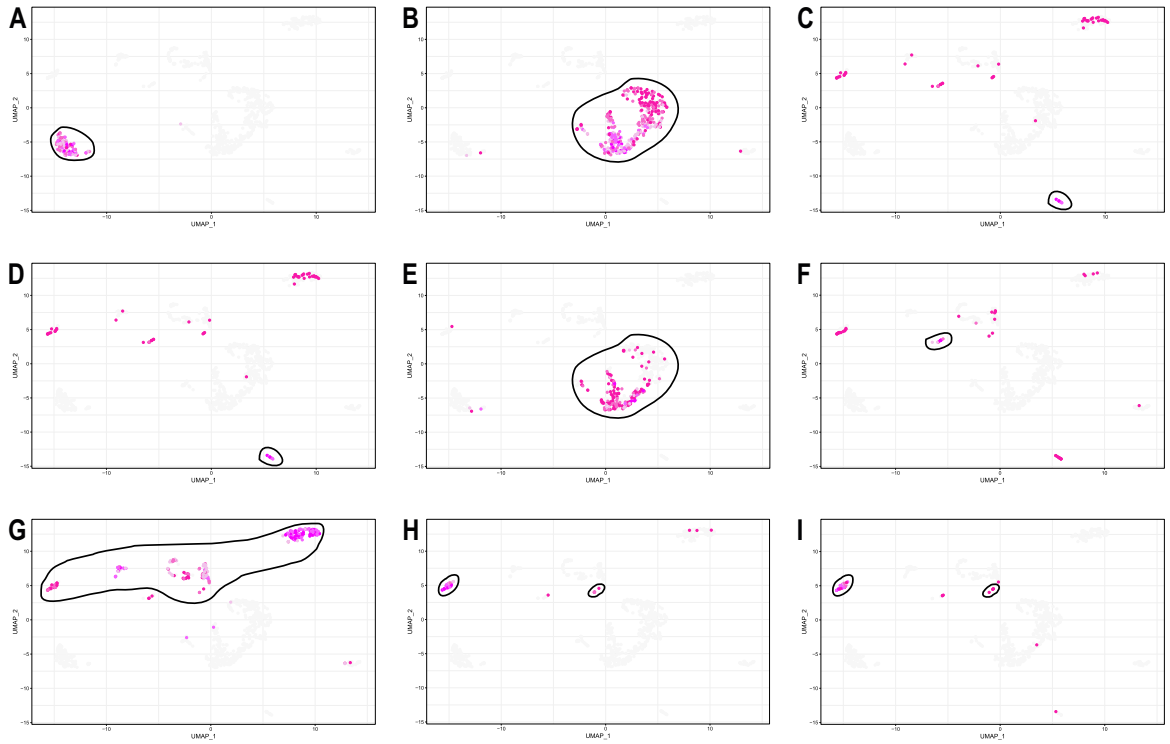

Figura 39 – Enriquecimento das assinaturas propostas para cada tipo celular e *EMT* utilizando a metodologia *AUCell*. As assinaturas são para células B (*A*), T (*B*), endoteliais (*C*), macrofagiais (*D*), *NK* (*E*), *CAF* (*F*), malignas (G), *EMT1* (*H*) e *EMT2* (*I*). Para cada *UMAP* estão circuladas as células que compõe o respectivo grupo. Tal visualização pode ser observada em detalhes na Figura 17.

4.1.7 Co-expressão para fatores de transcrição

Embora os *TFs* desempenham um papel regulador central na biologia celular, a detecção de sua expressão nas análises de sequência de *RNA* é limitada devido a sua baixa, e muitas vezes esparsa expressão. A análise *PCIT* e a métrica *RIF* foram usadas para identificar fatores de transcrição chaves (*KeyTF*) a partir de dados de expressão

gênica. A análise de co-expressão gênica, realizada pelo pacote *CeTF* (BIAGI *et al.*, 2021) calculou *RIF1*, que captura *TFs* mostrando conectividade entre genes diferencialmente expressos encontrados em contraste entre células malignas, e *RIF2*, que se concentra em *TFs* mostrando evidências como preditores de mudança em abundância de genes com expressão diferencial entre células não malignas.

Na Figura 40 é possível observar a rede de interações gerada com dados de células malignas (*A*) com 1.983 genes e 62.581 interações entre si, e não malignas (B) com 1.982 genes e 33.859 interações entre si. Na Figura 41 é possível observar a distribuição de expressão no formato de um gráfico (*smear plot*). Neste gráfico os *TFs* diferencialmente expressos para as condições malignas e não malignas estão coloridos de azul e vermelho, respectivamente. Os *TFs* destacados são chamados de *TFs* chave/principais (*key TFs*). Esses principais *TFs* podem ser observados na Tabela 6.

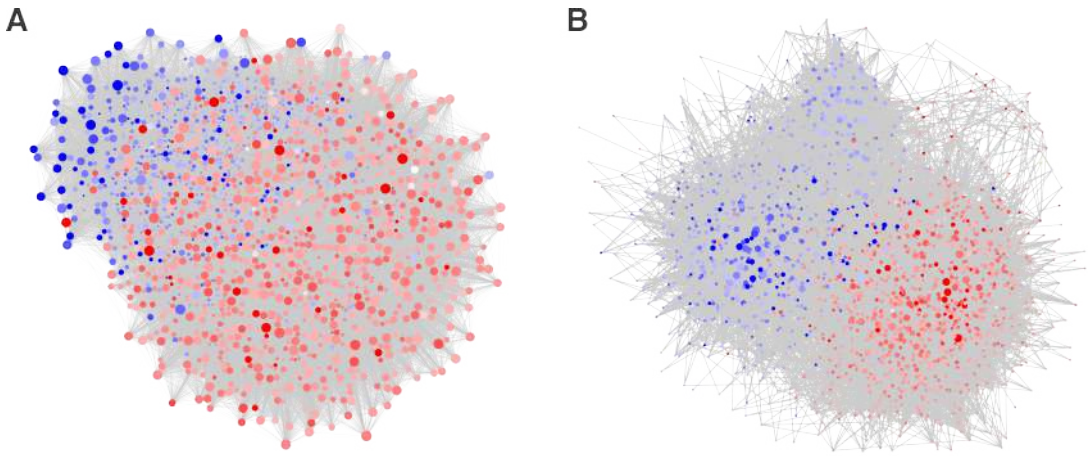

Figura 40 – Rede de interação gênica colorindo pelo valor de diferença de expressão e o tamanho de cada ponto representa o grau (quantidade de interações que um gene tem que outro/outros). Em (*A*) podemos observar a rede de interações para a condição de células malignas, e em (*B*) podemos observar a rede de interações para a condição de células não malignas.

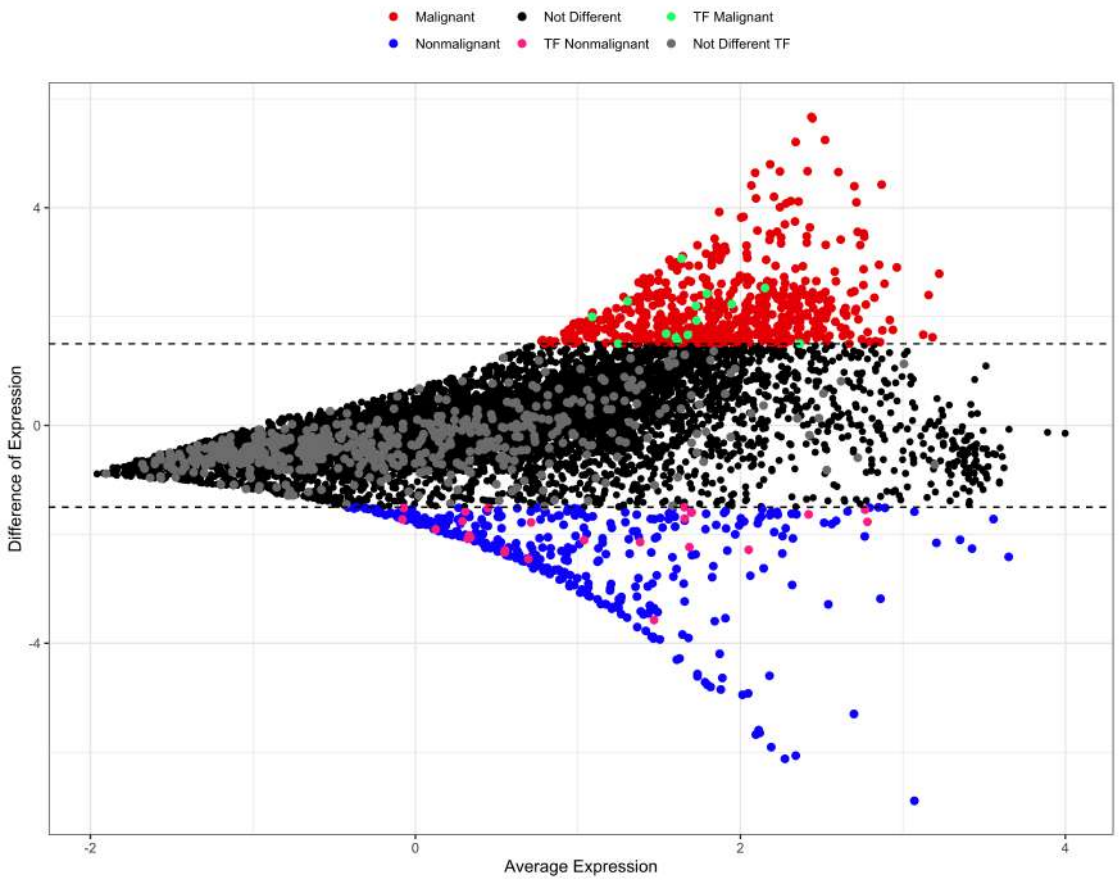

Figura 41 – Gráfico mostrando a diferença de expressão para 8.560 genes, os quais 610 são regulados positivamente (cor vermelha), 374 são regulados negativamente (cor azul), e os pontos na cor preta não são diferencialmente expressos com base em um corte de módulo de 1,5 na diferença de expressão. Há 14 *TFs* regulados positivamente (cor verde), 24 *TFs* regulados negativamente (cor rosa), e 553 *TFs* não expressos diferencialmente (cor cinza).

Tabela 6 – *TFs* encontrados como desempenhando um papel importante na comparação entre células malignas e não malignas. Também é mostrado a média de expressão (*AvgExpr*) para cada *TF*, além dos valores das métricas *RIF1* e *RIF2*. Finalmente, as colunas *Freq. Malignant* e *Freq. Nonmalignant* representam a frequência de aparecimento do *TF* dado em cada condição, sendo *Freq. Diff* a diferença entre estas frequências. Uma diferença positiva significa que o *TF* desempenha um papel importante na condição de referência no caso nas células malignas, enquanto uma diferença negativa significa que o *TF* desempenha um papel importante na condição das células não malignas.

TF	AvgExpr	RIF1	RIF2	Freq. Malignant	Freq. Nonmalignant	Freq. Diff
GTF3A	3,949	2,495	4,244	142	7	135
IKZF1	1,715	-2,758	-0,453	126	9	117
HMG20B	2,464	1,197	3,737	136	22	114
TSC22D3	7,244	9,532	4,583	127	14	113
ELOF1	1,994	2,477	3,831	125	16	109
MLX	2,168	0,978	2,578	112	11	101
CCDC124	2,116	2,143	3,922	124	26	98
ETV4	0,904	3,417	2,865	95	2	93
SNAI2	0,800	7,395	3,910	152	87	65
MITF	2,002	2,388	2,890	68	15	53
ETS2	0,566	4,394	0,731	32	41	-9
PYHIN1	2,195	-0,508	-2,144	39	72	-33
PAX5	0,462	-1,350	-2,187	36	88	-52
IRF8	1,702	-2,755	-2,370	24	154	-130
BCL11A	0,814	-1,625	-3,077	11	193	-182

Entre os *TFs* chave para a condição das células malignas estão: *GTF3A*, *IKZF1*, *HMG20B*, *TSC22D3*, *ELOF1*, *MLX*, *CCDC124*, *ETV4*, *SNAI2* e *MITF*. Vale ressaltar o papel do *IKZF1* que é parte de um *hub* de genes envolvidos na resposta imune e desenvolvimento de células tumorais na tumorigênese (WANG; LI; CHEN, 2018). Além disso, a superexpressão do *IKZF1* aumentou a eficácia da imunoterapia *PD-1* e *CTLA-4* em um estudo pré-clínico (CHEN *et al.*, 2018). O gene induzido, que está associado à citotoxicidade em distúrbios autoimunes, inibiu o crescimento do câncer em modelos de camundongos e aumentou a suscetibilidade imunológica de linhagens celulares derivadas de tumores. Além disso, de acordo com Rivero *et al.* (2015), o fato de o *HMG20A* ser *upregulado* em três estudo de melanoma maligno e a forte correlação positiva entre o *HMG20A* e os níveis de expressão dos marcadores mesenquimais, em vários tipos de tumores, abre a possibilidade de investigar o papel desta proteína na metástase cancerígena. Por fim, o *MITF* atua como um regulador mestre do desenvolvimento, função e sobrevivência dos melanócitos, modulando vários genes de diferenciação e progressão do ciclo celular. É demonstrado que o *MITF* é um oncogene amplificado em uma fração dos melanomas humanos e que também tem um papel oncogênico no sarcoma de células claras humanas.

Entretanto, o *MITF* também modula o estado de diferenciação dos melanócitos. Ele é colocado entre a instrução dos melanócitos para a diferenciação e/ou pigmentação terminal e, alternativamente, a promoção do comportamento maligno (LEVY; KHALED; FISHER, 2006; BALLOTTI; CHELI; BERTOLOTTO, 2020; HARTMAN; CZYZ, 2015).

Entre os *TFs* chave para a condição das células não malignas estão: *ETS2*, *PYHIN1*, *PAX5*, *IRF8*, *BCL11A*. Vale ressaltar o papel do *BCL11A* que é importante na regulação da barreira de permeabilidade epidérmica, que evita a desidratação e infecção dos organismos (LI *et al.*, 2017). O *IRF8* também desempenha um papel importante de controle da progressão do melanoma a partir da regulação da conversa cruzada entre o câncer e suas células imunes dentro do microambiente tumoral (MATTEI *et al.*, 2012). Além disso, o *IRF8* foi identificado como um modulador crucial da progressão do melanoma operando na interface entre as células malignas e o infiltrado imunológico. As células efetoras imunológicas infiltram ativamente as lesões neoplásicas que sustentam a expressão do *IRF8* pelas células do melanoma e, por sua vez, os níveis intratumorais basais do *IRF8* são indispensáveis para manter um microrganismo imunológico adequado, em parte através da modulação dos fatores antitumorais solúveis, favorecendo assim o controle de doenças. Este mecanismo de interação mútua entre células malignas e imunológicas pode também se aplicar a outros fatores oncosupressores, como proposto recentemente para células *IRF1* e *NK* em um modelo murino de metástase pulmonar (KSIENZYK *et al.*, 2011). A proteína oncosupressora *IRF8* pode ser reexaminada como biomarcador clinicamente relevante para o tratamento precoce de melanoma com abordagens que combinam agentes anti-neoplásicos direcionados e imunoterapia (SCHIAVONI; GABRIELE; MATTEI, 2013).

O *TF* com maior valor de frequência para a condição maligna foi o *GTF3A*. Não foi encontrado na literatura indícios que está relacionado com melanoma, o que pode trazer novos *insights* do entendimento desta doença a partir dele. A rede de interação para o *GTF3A* possui 261 genes correlatos além de 3.512 interações entre eles. Foram encontrados 6 *clusters* que podem ser observados na Figura 42. Os genes de cada *cluster* foram enriquecidos a partir do *Gene Ontology* e as principais vias são destacadas na Tabela 7. As principais vias mais significantes em comum aos *clusters* estão relacionadas a membrana interna mitocondrial (*GO:0005743*) e matriz mitocondrial (*GO:0005759*). As mitocôndrias retardam significativamente o crescimento do tumor e aumentam os dias de sobrevivência dos animais. O efeito antitumoral das mitocôndrias está relacionado à interferência nos metabolismos das células tumorais, como a redução da glicólise e a

produção de um ambiente intracelular oxidativo, que não são todos adequados para a proliferação de células tumorais. Além disso, as mitocôndrias aumentam a apoptose celular, a necrose e a mitofagia (FU *et al.*, 2019). Além disso, as proteases de matriz mitocondrial são sugeridas como novos alvos terapêuticos na malignidade (GOARD; SCHIMMER, 2014).

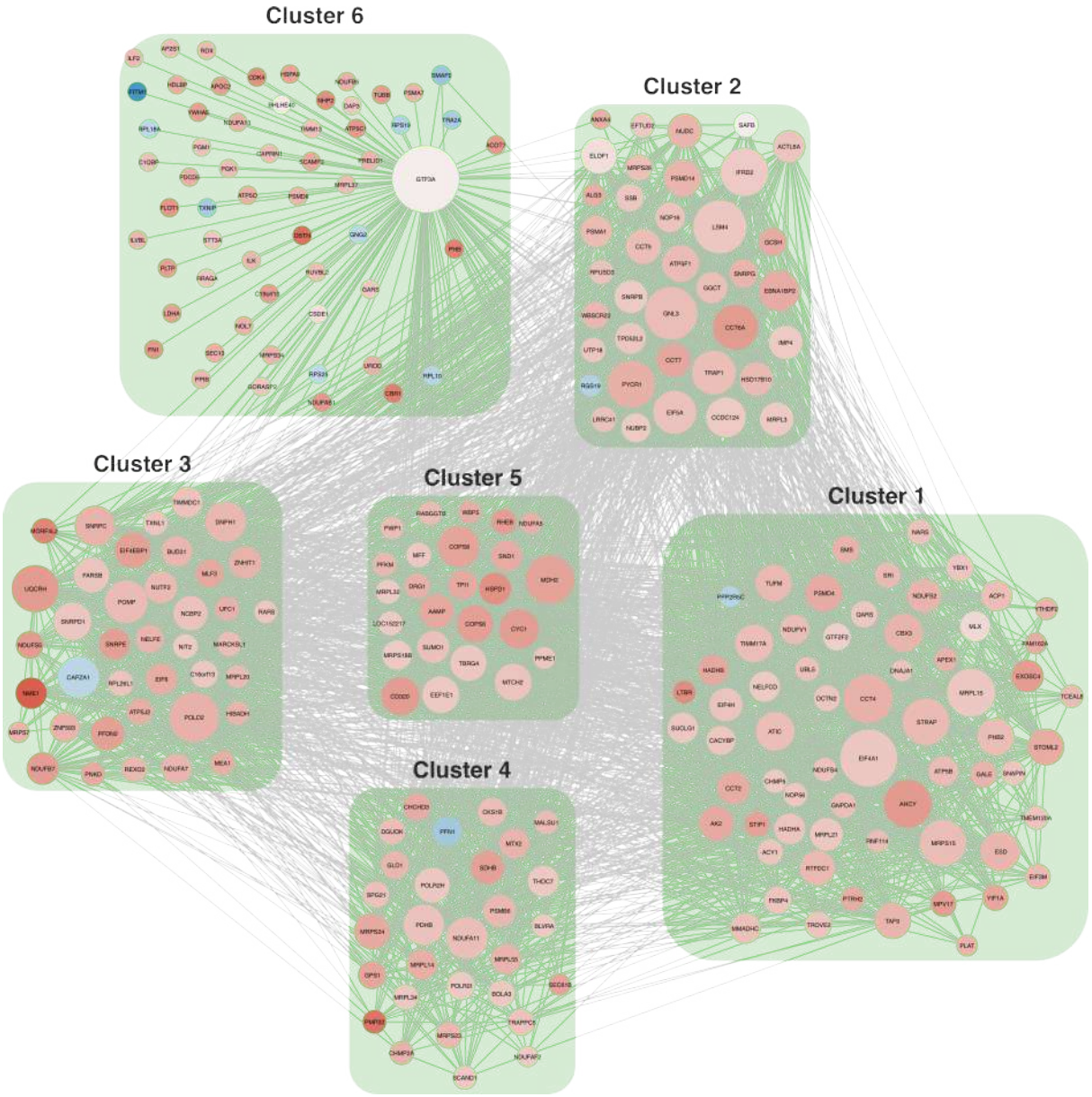

Figura 42 – Rede de interações para o fator de transcrição *GTF3A*. Essa rede possui 261 genes correlatos totalizando 3.512 interações entre si. Após a aplicação do algoritmo *louvain*, foi possível dividir a rede em 6 *clusters*.

Tabela 7 – *Top* 3 vias enriquecidas por *cluster* da rede do fator de transcrição *GTF3A* usando o banco de dados do *Gene Ontolology*. A coluna *ID* contém a codificação da via e *ONT* representa o tipo de ontologia que a via pertence (*CC* = componente celular, *BP* = processos biológicos e *MF* = funções moleculares). Há uma coluna contendo a descrição da via, o valor de p ajustado e, por fim, a qual *cluster* a via pertence.

ID	ONT	Description	p.adjust	Cluster
GO:0098798	CC	mitochondrial protein-containing complex	5,355E-05	Cluster 1
GO:0005743	CC	mitochondrial inner membrane	1,083E-04	Cluster 1
GO:0098800	CC	inner mitochondrial membrane protein complex	4,513E-04	Cluster 1
GO:0043021	MF	ribonucleoprotein complex binding	6,090E-03	Cluster 2
GO:0005743	CC	mitochondrial inner membrane	1,141E-02	Cluster 2
GO:0042773	BP	ATP synthesis coupled electron transport	1,141E-02	Cluster 2
GO:0005743	CC	mitochondrial inner membrane	6,041E-03	Cluster 3
GO:0032048	BP	cardiolipin metabolic process	7,648E-03	Cluster 3
GO:0044282	BP	small molecule catabolic process	1,850E-02	Cluster 3
GO:1904816	BP	positive regulation of protein localization to chromosome, telomeric region	4,129E-04	Cluster 4
GO:0005759	CC	mitochondrial matrix	4,129E-04	Cluster 4
GO:1904814	BP	regulation of protein localization to chromosome, telomeric region	4,129E-04	Cluster 4
GO:1904816	BP	positive regulation of protein localization to chromosome, telomeric region	4,129E-04	Cluster 5
GO:0005759	CC	mitochondrial matrix	4,129E-04	Cluster 5
GO:1904814	BP	regulation of protein localization to chromosome, telomeric region	4,129E-04	Cluster 5
GO:1904816	BP	positive regulation of protein localization to chromosome, telomeric region	4,129E-04	Cluster 6
GO:0005759	CC	mitochondrial matrix	4,129E-04	Cluster 6
GO:1904814	BP	regulation of protein localization to chromosome, telomeric region	4,129E-04	Cluster 6

O *TF* com maior valor de frequência para a condição não maligna foi o *IRF8*. Apesar de ter sido encontrado informações relativas a esse *TF* na literatura, podemos trazer diferentes *insights* para o entendimento desta doença a partir dele. A rede de interação para o *IRF8* possui 205 genes correlatos além de 3.394 interações entre eles. Foram encontrados 4 *clusters* que podem ser observados na Figura 43. Os genes de cada *cluster* foram enriquecidos a partir do *Gene Ontology* e as principais vias são destacadas na Tabela 8. Os *clusters* 1 e 4 possuem vias de regulação positiva da ativação de células T (*GO:0050870*, *GO:2000318*, *GO:0050852*) e ligação *FK506* (*GO:0005528*). Foi mostrado o papel da modulação da ativação de células T por células iniciadoras de melanoma maligno (TONG; JIANG, 2016). Além disso, as proteínas de ligação *FK506* (*FKBPs*), que são uma grande família de proteínas, nos melanomas, a expressão do gene *FKBP51* se correlaciona com a invasividade e a agressividade do câncer. Foi demonstrado que o *FKBP51* exibe atividade anti-apoptose e protege as células cancerosas da morte celular induzida pela irradiação. Esta função pró-sobrevivência do *FKBP51* é mediada por *NF-κB*. Em resposta à irradiação, *FKBP51* estimula a ativação da NF-κB, que suprime a apoptose. O *FKBP51* também pode promover a migração e invasão das células do melanoma aumentando a sinalização TGF-β e a ativação dos genes de *EMT* (SCHATTON *et al.*, 2010). O *cluster* 2

possui vias de receptor e ativação de células B (*GO:0050853*, *GO:0042113*, *GO:0050855*). Estudos mostram que as células B associadas ao tumor são vitais para a inflamação associada ao melanoma (GRISS *et al.*, 2019), e também a partir da crescente valorização do papel da imunidade celular B há a necessidade de novas estratégias terapêuticas e biomarcadores para serem explorados e traduzidos na clínica para otimizar a imunoterapia de inibidores de ponto de verificação no melanoma (WILLSMORE *et al.*, 2020).

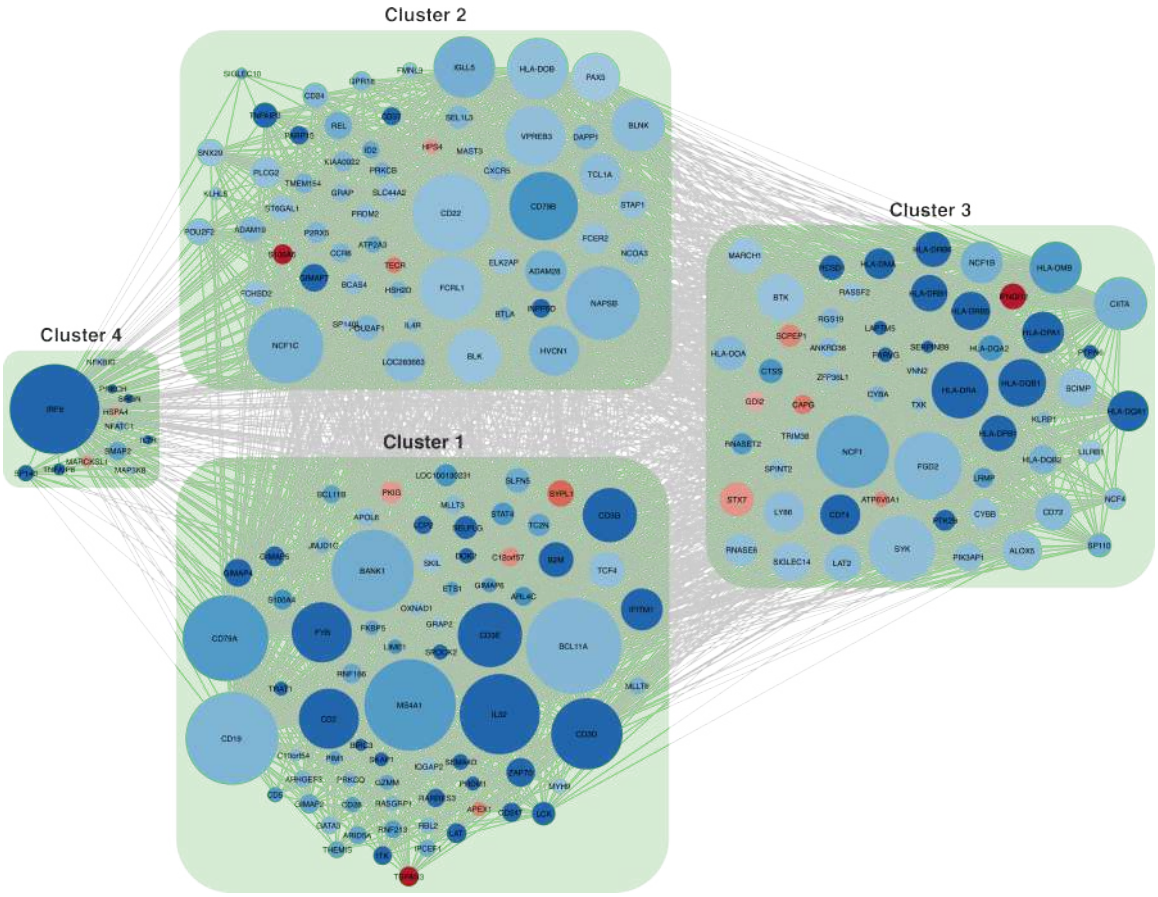

Figura 43 – Rede de interações para o fator de transcrição *IRF8*. Essa rede possui 205 genes correlatos totalizando 3.394 interações entre si. Após a aplicação do algoritmo *louvain*, foi possível dividir a rede em 4 *clusters*.

Tabela 8 – *Top* 3 vias enriquecidas por *cluster* da rede do fator de transcrição *IRF8* usando o banco de dados do *Gene Ontolology*. A coluna *ID* contém a codificação da via e *ONT* representa o tipo de ontologia que a via pertence (CC = componente celular, BP = processos biológicos e MF = funções moleculares). Há uma coluna contendo a descrição da via, o valor de p ajustado e, por fim, a qual *cluster* a via pertence.

ID	ONT	Description	p.adjust	Cluster
GO:0050870	BP	positive regulation of T cell activation	8,116E-02	Cluster 1
GO:0005528	MF	FK506 binding	8,116E-02	Cluster 1
GO:2000318	BP	positive regulation of T-helper 17 type immune response	8,116E-02	Cluster 1
GO:0050853	BP	B cell receptor signaling pathway	3,703E-11	Cluster 2
GO:0042113	BP	B cell activation	7,381E-10	Cluster 2
GO:0050855	BP	regulation of B cell receptor signaling pathway	3,110E-08	Cluster 2
GO:0042613	CC	MHC class II protein complex	1,992E-28	Cluster 3
GO:0042611	CC	MHC protein complex	9,078E-25	Cluster 3
GO:0002478	BP	antigen processing and presentation of exogenous peptide antigen	7,175E-21	Cluster 3
GO:0050852	BP	T cell receptor signaling pathway	8,203E-17	Cluster 4
GO:0050851	BP	antigen receptor-mediated signaling pathway	1,494E-16	Cluster 4
GO:0002429	BP	immune response-activating cell surface receptor signaling pathway	1,229E-13	Cluster 4

4.1.8 Inferência da trajetória em células únicas de melanoma

Após a normalização com base no tamanho da biblioteca, transformação desse resultado em escala logarítmica, remoção dos genes mitocondriais e filtro dos genes e células, foi possível selecionar os genes mais variáveis. Essa seleção é um passo muito importante pois a partir dela é serão escolhidos os genes para auxiliar na redução da dimensão. O ideal é que os genes mais variáveis sigam a curva gaussiana quando comparado o valor médio de expressão em relação ao desvio padrão. Na Figura 44 é possível observar o comportamento do dado seguindo a gaussiana para os 998 genes mais variáveis selecionados.

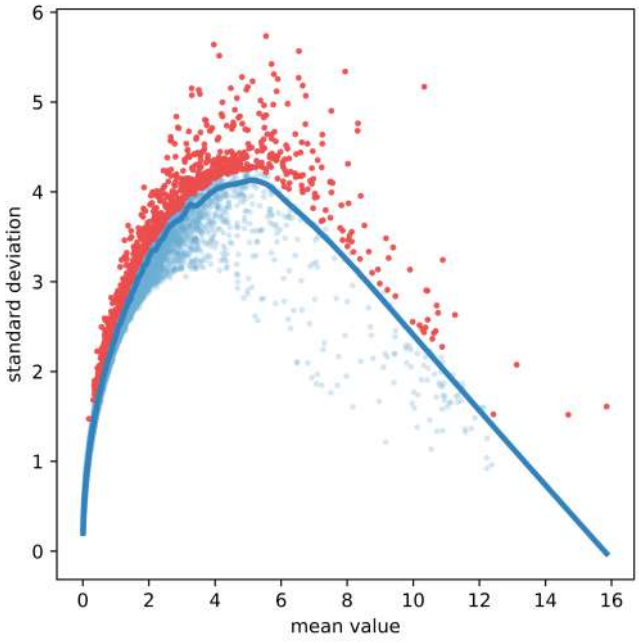

Figura 44 – Gráfico da média de expressão gênica em relação ao desvio padrão. Os pontos vermelhos mostram os genes mais variáveis, enquanto os pontos azuis representam os genes restantes. A linha mais espessa em azul mostra a tendência gaussiana dos dados.

O próximo passo foi a redução de dimensão dos dados. Para isso foi utilizado o método *Spectral Embedding* com um total de 7 componentes, cobrindo os 7 tipos celulares disponíveis no estudo. Como resultado (Figura 45 foi possível observar a separação entre os tipos celulares, o que será importante para os próximos passos.

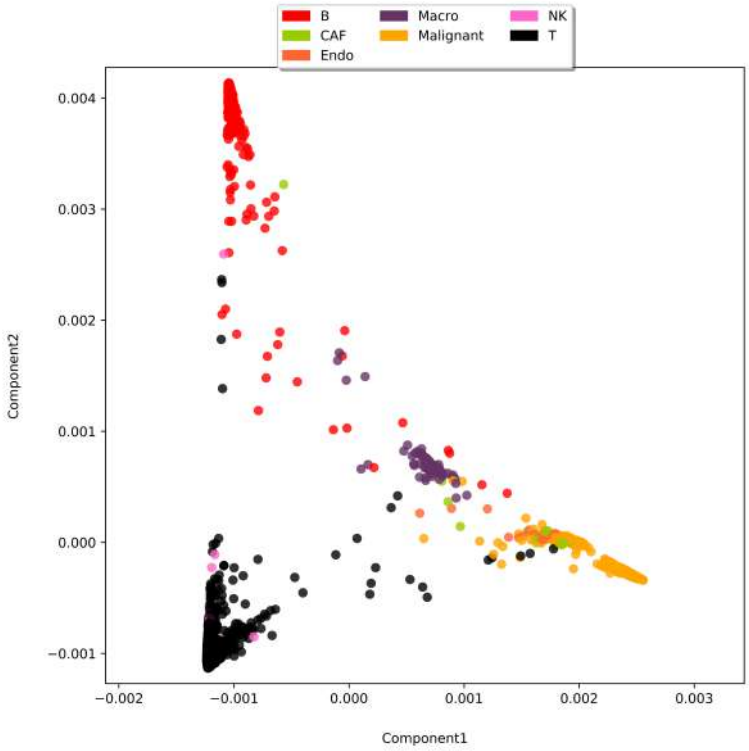

Figura 45 – Redução da dimensão utilizando o método *Spectral Embedding* e um total de 7 componentes.

Finalmente é calculada a inferência da trajetória, gerando assim os dois gráficos da Figura 46. Nos gráficos A e B observa-se as trajetórias obtidas a partir do *STREAM*, sem e com as células, respectivamente. Para uma melhor visualização foi gerado o gráfico de fluxo (Figura 47). A partir do gráfico de fluxo observa-se a diferenciação dos tipos celulares, sendo que a trajetória se inicia em 1, onde estão situadas as células macrofagiais. Com o "passar do tempo" elas se diferenciam em seis braços, que são representadas pelas células B (2), células T e *NK* (3), e por fim nas células malignas (4, 5, 6 e 7), sendo que há a diferenciação das células Endoteliais e *CAF* no braço 6.

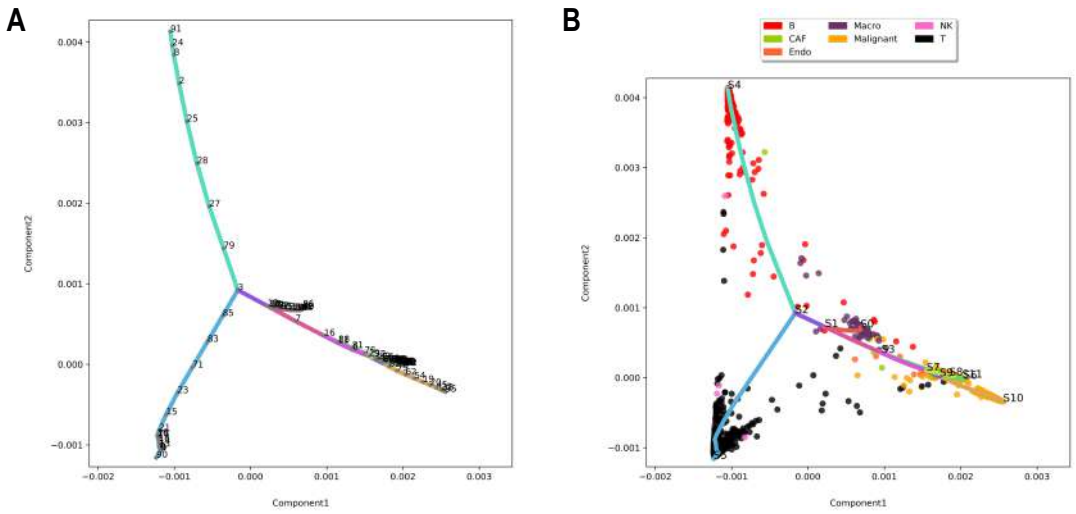

Figura 46 – Distribuição da inferência de trajetória realizada pelo *STREAM* mostrando apenas a trajetória (A) e mostrando a trajetória com a localização das células (B).

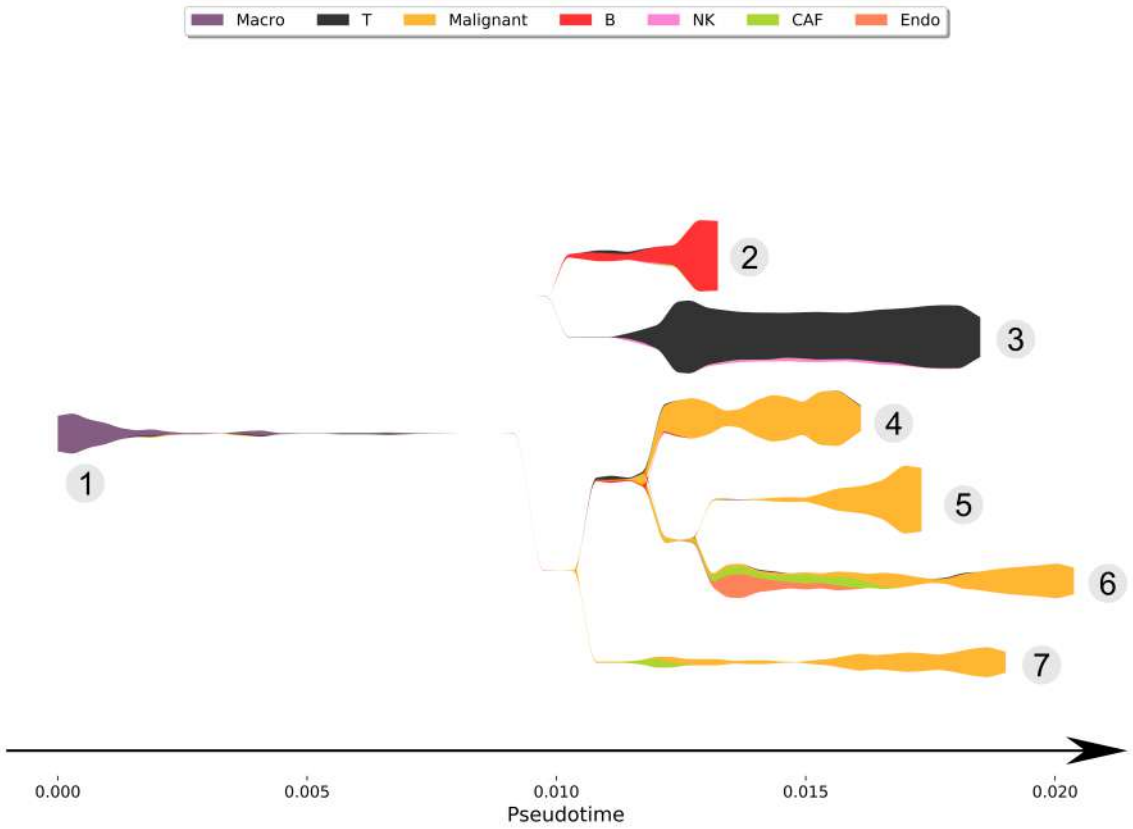

Figura 47 – Inferência de trajetória dos 7 tipos celulares representada pelo gráfico de fluxo (*stream plot*).

Nota-se que o braço de número 6, identificado na Figura 47, inicia-se com células endoteliais, que logo em seguida são diferenciadas em *CAF* e, por fim, em células malignas.

Quando verificado os valores de expressão dos genes longos não codificantes *HOTAIR* e *CDH1*, que são fundamentais na ativação da via *EMT* e foram abordados com mais detalhes nos itens anteriores, percebe-se que o *HOTAIR* possui expressão elevada no braço 6 (Figura 48) enquanto os outros braços são completamente zerados em nível de expressão. Em contrapartida, o *CDH1* possui expressão elevada nos braços 4, 5 e 7 (Figura 49). Mais uma vez demonstrando que tais transcritos possuem expressão inversa indicando ativação ou não da via de *EMT*, o que valida nossa análise.

O *lncRNA TRHDE-AS1*, como abordado anteriormente, possui níveis de expressão muito parecidos com o *HOTAIR*. Verificando-se a expressão do *TRHDE-AS1* na trajetória (Figura 50) nota-se que o braço 6 possui alta expressão, muito parecido com a expressão do *HOTAIR* mostrado na Figura 48 e provavelmente ativas nas mesmas células.

O fato do braço 6 possuir o comportamento de expressão de ativação da via *EMT* levanta alguns questionamentos do porquê células endoteliais e *CAF* são diferenciadas em células malignas (células com alta expressão de *HOTAIR* e *TRHDE-AS1* e baixa expressão do *CDH1*). Essa topologia pode reforçar as evidências do papel das células estromais na manutenção e progressão do clone tumoral. Os tumores sólidos podem ser considerados como órgãos aberrantes, que sofreram reprogramação molecular e celular, promovendo um nicho proliferativo e invasivo na progressão tumoral. Os tumores contêm componentes celulares e não celulares, que juntos formam o microambiente tumoral (FIORI *et al.*, 2019). Os tipos celulares que compõe o microambiente tumoral incluem: células neuroendócrinas, adiposas, endoteliais, mesenquimais, imuno-inflamatórias, bem como fibroblastos. Os *CAFs* podem derivar de diferentes tipos de células como as células epiteliais após *EMT*, células endoteliais via transição endotelial-mesenquimal (*EndMT*), etc (PRAKASH, 2016). As células endoteliais são apontadas como precursoras das *CAFs* quando associadas à via *EMT*. Outro fato interessante é que o estroma tumoral não é mais visto apenas como suporte físico para células epiteliais mutantes, mas como um modulador importante e até mesmo um impulsionador da tumorigenicidade. Dentro do meio estromal do tumor, populações heterogêneas de *CAF* são peça fundamental na indução clonal, dependentes do estroma, que contribuem para a iniciação e progressão maligna (GASCARD; TLSTY, 2016). Desta forma, as *CAFs*, juntamente com as células endoteliais, junto com outras células do estroma formam o nicho tumoral que podem desempenhar um papel crucial na ativação da via *EMT*, progressão tumoral e metástase (SASAKI *et al.*, 2018). A detecção

de células tumorais com expressão concomitante de *HOTAIR* e *TRHDE-AS1* pode ajudar na identificação de células metastática ou de células com potencial de metastização.

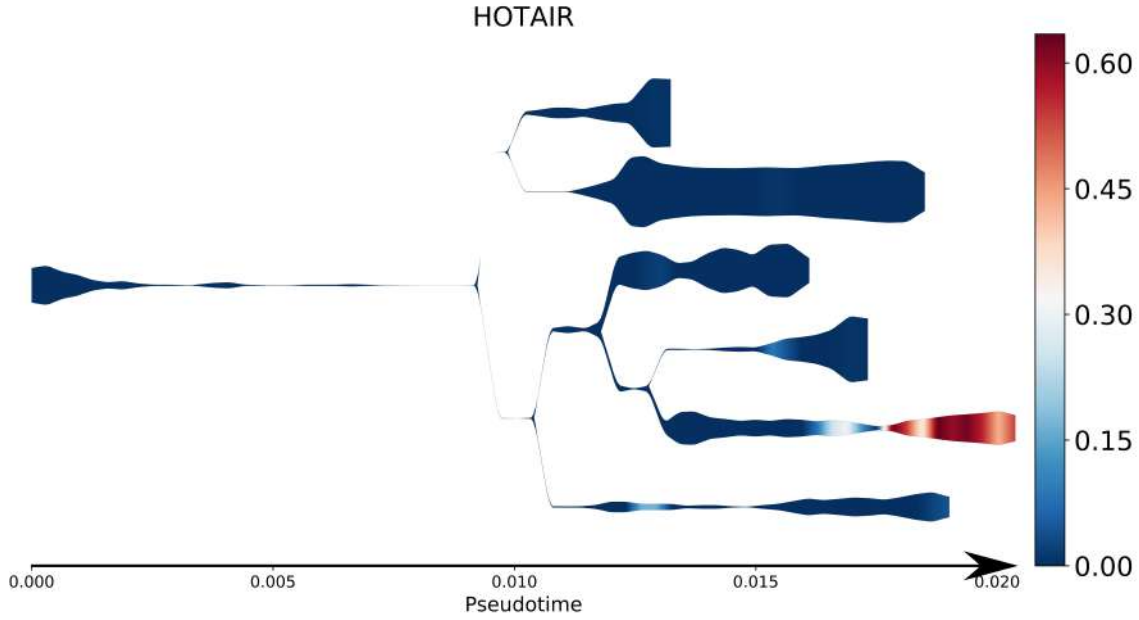

Figura 48 – *Stream plot* mostrando a expressão do gene longo não codificante *HOTAIR*.

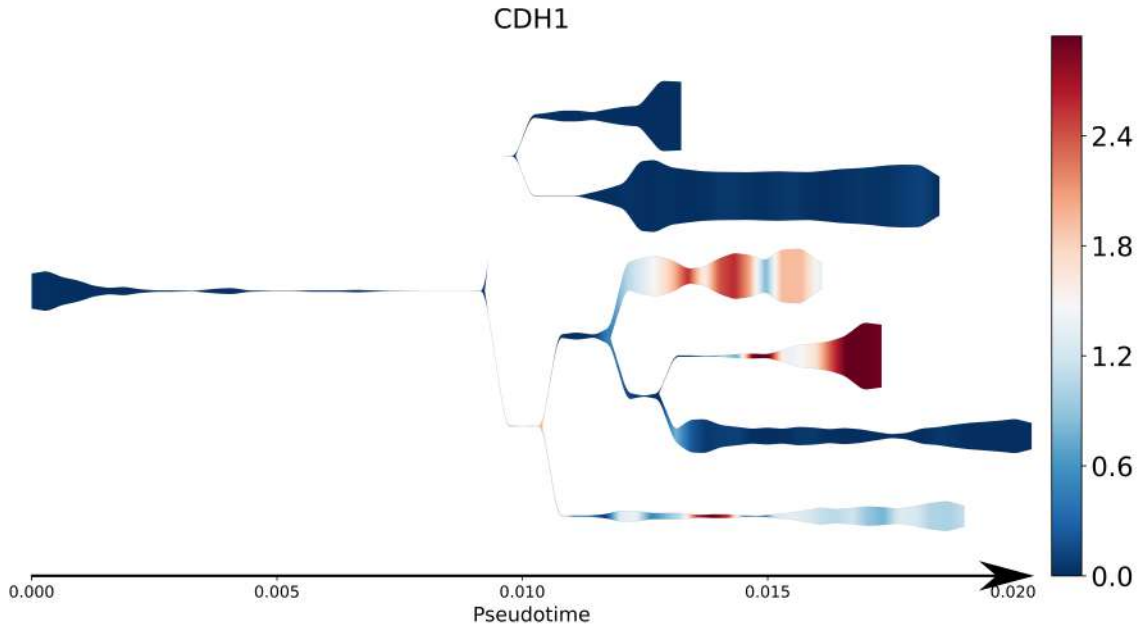

Figura 49 – *Stream plot* mostrando a expressão do gene longo não codificante *CDH1*.

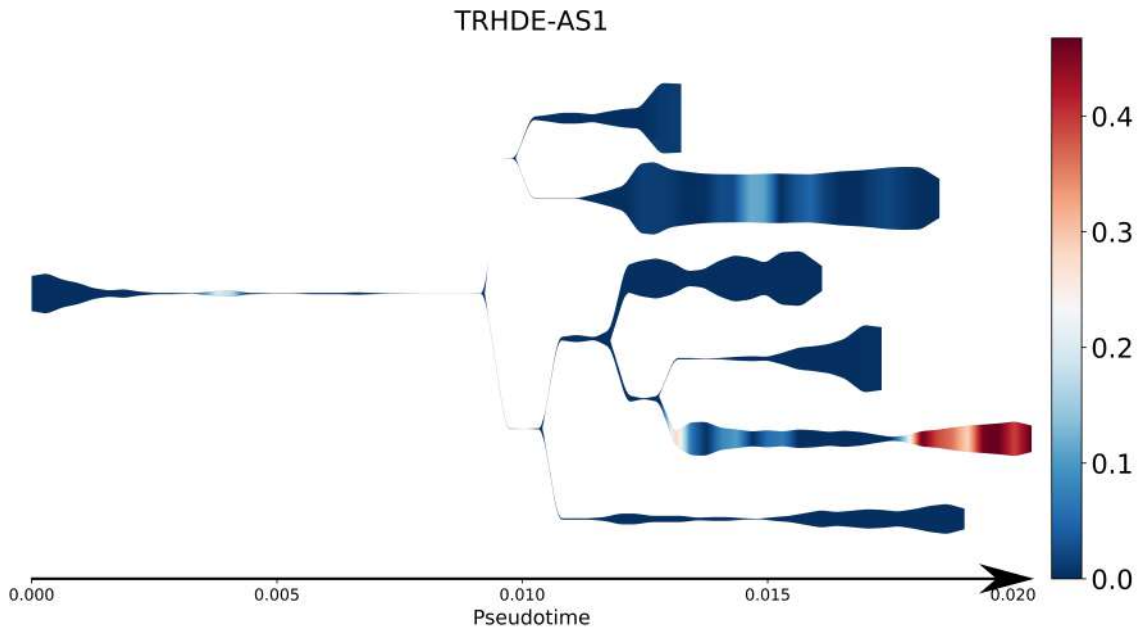

Figura 50 – *Stream plot* mostrando a expressão do gene longo não codificante *TRHDE-AS1*.

Como descrito anteriormente, a partir da Figura 47 foi possível identificar e demonstrar que o braço 6 possui um perfil de ativação ou não da via de *EMT* baseado na expressão de genes característicos. Sabe-se que o braço 6 inicia-se com células endoteliais, que logo em seguida são diferenciadas em *CAF* e, por fim, em células malignas. Baseada nesta ordem, foi possível utilizar a ferramenta *psupertime* para calcular o *pseudotime* e verificar a expressão de genes característicos da via *EMT*. É possível observar na Figura 51 a expressão de 8 genes (*HOTAIR*, *TRHDE-AS1*, *CDH1*, *TWIST1*, *FN1*, *VIM*, *TGFB1*, *NANOG* e *ZEB1*). Observa-se, novamente, que para o braço 6 os genes *HOTAIR* e *TRHDE-AS1* possuem o mesmo perfil de expressão, comportamento inverso é encontrado para o gene *CDH1*. Além destes genes há outros genes que compõe a via de *EMT* e correspondem ao perfil proposto por Alves *et al.* (2013). Essa análise possibilita, mais uma vez, reafirmar a hipótese de que o braço 6 possui um perfil de *EMT*.

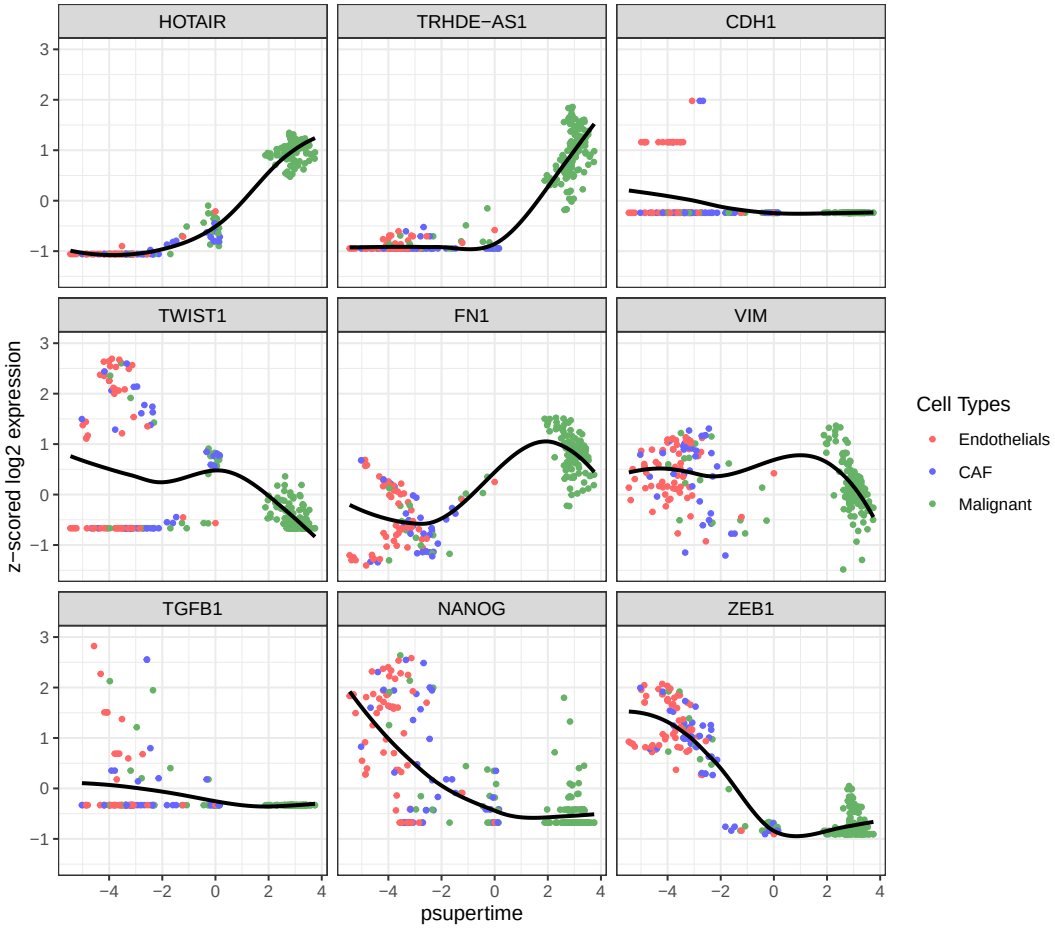

Figura 51 – *Pseudotime* calculado utilizando a ferramenta *psupertime* mostrando no eixo x o valor de *pseudotime* e no eixo y a expressão em log_2 do *z-score*. Os pontos com diferentes cores representam os tipos celulares (endoteliais, CAF e malignas) presentes no braço 6. Os genes representados nesta figuras possuem relação com a via de *EMT*.

4.2 Câncer de Pulmão de Pequenas Células (CPPC)

4.2.1 Visão geral dos dados

Após o processamento das amostras utilizando a ferramenta *M3K* foi possível identificar o número de células, genes, *reads* e *reads*/células para cada amostra (Tabela 9). Em média foram obtidas 4.173 células, além de 23.8778 genes e 45.274.168 *reads* por amostra. Na Figura 52 é possível observar em um gráfico de barras o número de células e de *reads* por amostra. Nota-se que tanto a distribuição de células e principalmente de *reads* por amostra não são uniformes, dessa forma sugerindo que é necessário um sequenciamento com melhor cobertura e cálculo de *reads* por amostra para haver um balanceamento.

Tabela 9 – Número de células, genes, *reads* e *reads*/células por amostra.

Sample	Cells	Genes	Reads	Reads/Cell
Patient 01	6.412	25.425	58.607.892	9.140
Patient 02	5.634	25.915	52.941.771	9.397
Patient 03	4.267	24.434	45.803.348	10.734
Patient 04	6.292	25.180	80.882.604	12.855
Patient 05	5.084	25.465	70.888.348	13.943
Patient 06	3.006	22.501	32.337.522	10.758
Patient 07	2.532	25.213	36.888.624	14.569
Patient 08	3.369	22.260	27.536.670	8.174
Patient 09	2.743	23.474	40.029.305	14.593
Patient 10	4.702	22.689	33.819.954	7.193
Patient 11	5.112	25.297	52.073.633	10.187
Patient 12	1.937	23.390	31.523.642	16.274
Patient 13	2.880	22.754	37.351.833	12.969
Patient 14	4.744	25.057	60.246.584	12.700
Patient 15	634	20.509	7.111.827	11.217
Patient 16	4.169	24.979	87.381.521	20.960
Patient 17	1.518	21.685	33.237.795	21.896
Patient 18	901	21.733	25.528.754	28.334
Patient 19	2.330	25.772	37.011.494	15.885
Patient 20	11.656	26.769	92.325.079	7.921
Patient 21	3.161	24.609	22.545.213	7.132
Patient 22	3.051	23.838	25.094.714	8.225
Patient 23	456	19.650	5.771.073	12.656
Patient 24	5.633	23.073	27.614.111	4.902
Patient 25	4.766	24.151	81.812.339	17.166
Patient 26	3.514	24.600	55.995.646	15.935
Patient 27	3.049	23.772	33.635.197	11.032
Patient 28	2.654	24.156	30.722.097	11.576
Patient 29	8.105	26.726	69.879.544	8.622
Patient 30	837	16.445	3.933.987	4.700
Patient 31	5.858	27.289	38.527.475	6.577
Patient 32	2.411	26.169	62.942.982	26.107
Patient 33	4.044	18.792	6.997.355	1.730
Patient 34	5.467	26.330	84.826.282	15.516
Patient 35	915	18.859	17.146.976	18.740
Patient 36	3.881	25.546	79.414.838	20.462
Patient 37	1.915	25.873	44.970.161	23.483
Patient 38	1.346	20.361	9.564.418	7.106
Patient 39	9.955	25.731	138.526.665	13.915
Patient 40	3.065	29.447	50.467.968	16.466
Patient 41	5.259	26.461	115.394.142	21.942
Patient 42	8.646	24.465	36.109.709	4.176
Patient 43	6.326	27.486	70.638.413	11.166
Patient 44	8.601	26.557	73.900.993	8.592
Patient 45	2.967	27.729	68.019.453	22.925
Patient 46	4.981	27.785	86.579.942	17.382
Patient 47	4.260	22.028	16.185.656	3.799
Patient 48	4.930	18.014	7.348.942	1.491
Patient 49	3.599	19.138	9.122.205	2.535
Patient 50	1.046	13.237	1.235.092	1.181
Patient 51	1.950	23.184	23.375.367	11.987
Patient 52	8.181	25.479	60.434.059	7.387
Patient 53	6.895	27.729	28.640.604	4.154
Patient 54	3.693	24.160	13.903.276	3.765

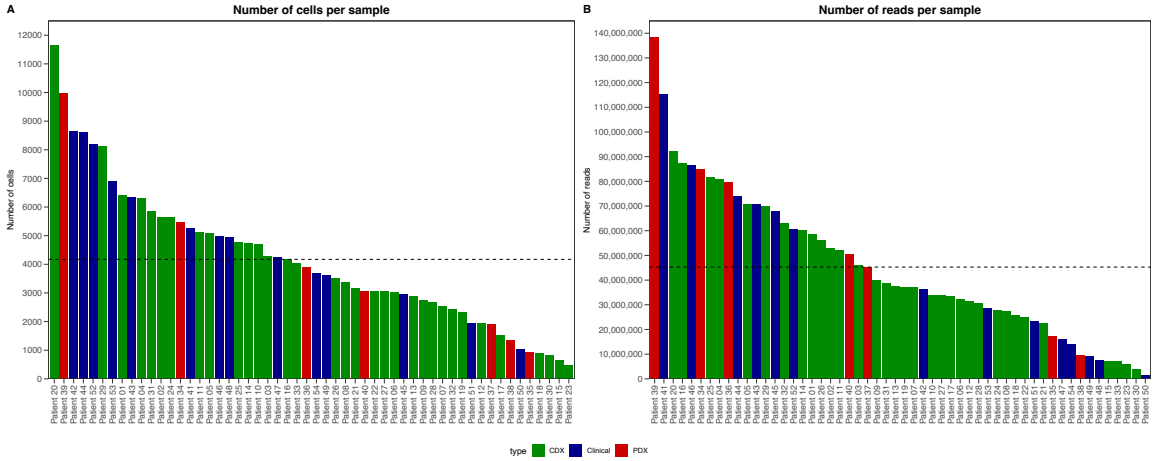

Figura 52 – Gráfico de barras representando em (*A*) o número de células por amostra e em (*B*) o número de *reads* por amostra. Em ambos o gráfico a linha tracejada em preto representa a média.

4.2.2 Ciclo celular

Determinamos o estado do ciclo celular para todos os casos a fim de mitigar os efeitos da heterogeneidade do ciclo celular, e determinamos a distribuição das células nas fases *G1*, *S* e *G2/M* (Figura 53). Além dos casos de CPPC, foram determinados os estados do ciclo celular para o dado de adenocarcinoma (Figura 54). De acordo com os exames histológicos (EYMIN; GAZZERI, 2010; STERLACCI; FIEGL; TZANKOV, 2012), o CPPC revela uma fração significativamente maior de células na fase *G2/M* em comparação com os adenocarcinomas, confirmando assim um estado celular altamente proliferativo para CPPC.

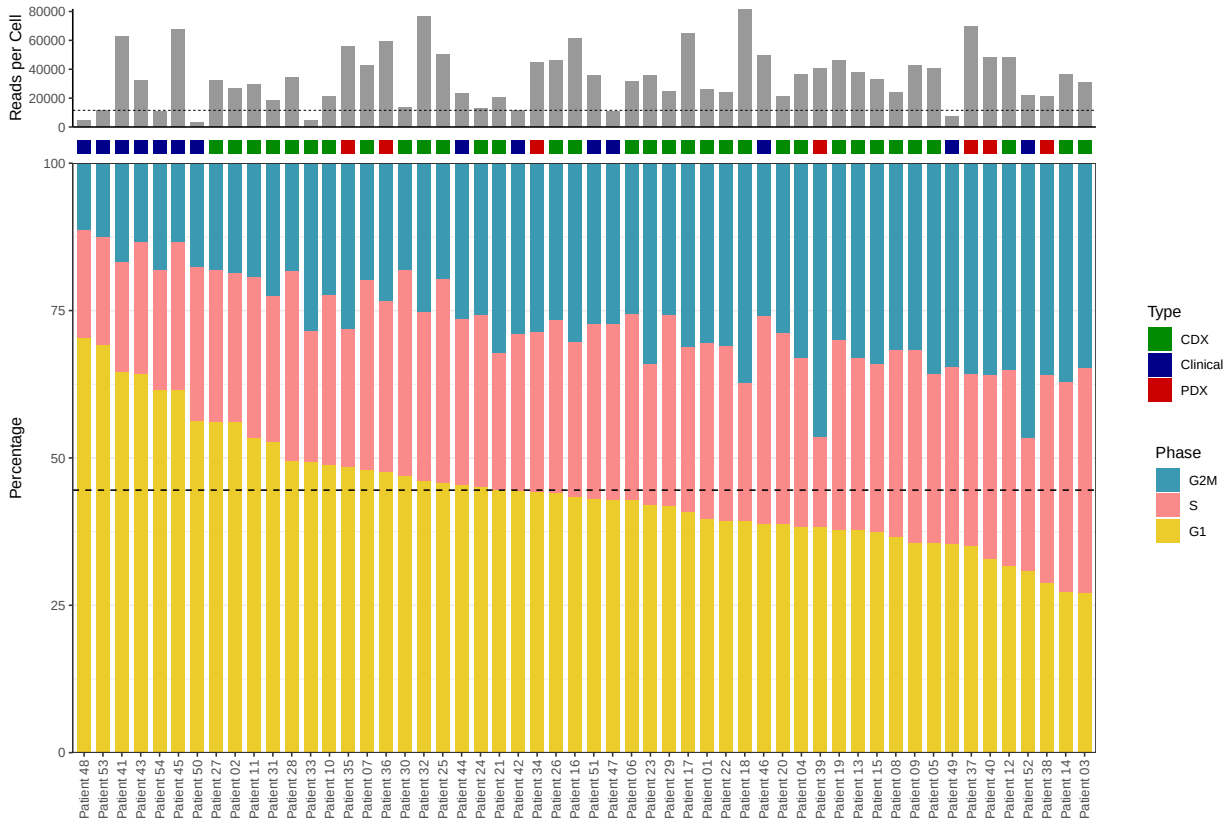

Figura 53 – Estados do ciclo celular para cada uma das 54 amostras de CPPC. Os estados do ciclo celular são divididos em *G2/M*, *S* e *G1*. A anotação em verde, azul e vermelho identificam se as amostras são *CDX*, *PDX* ou Clínicas, respectivamente. Além disso há um gráfico de barras na parte superior indicando o número de *reads* por célula para cada amostra.

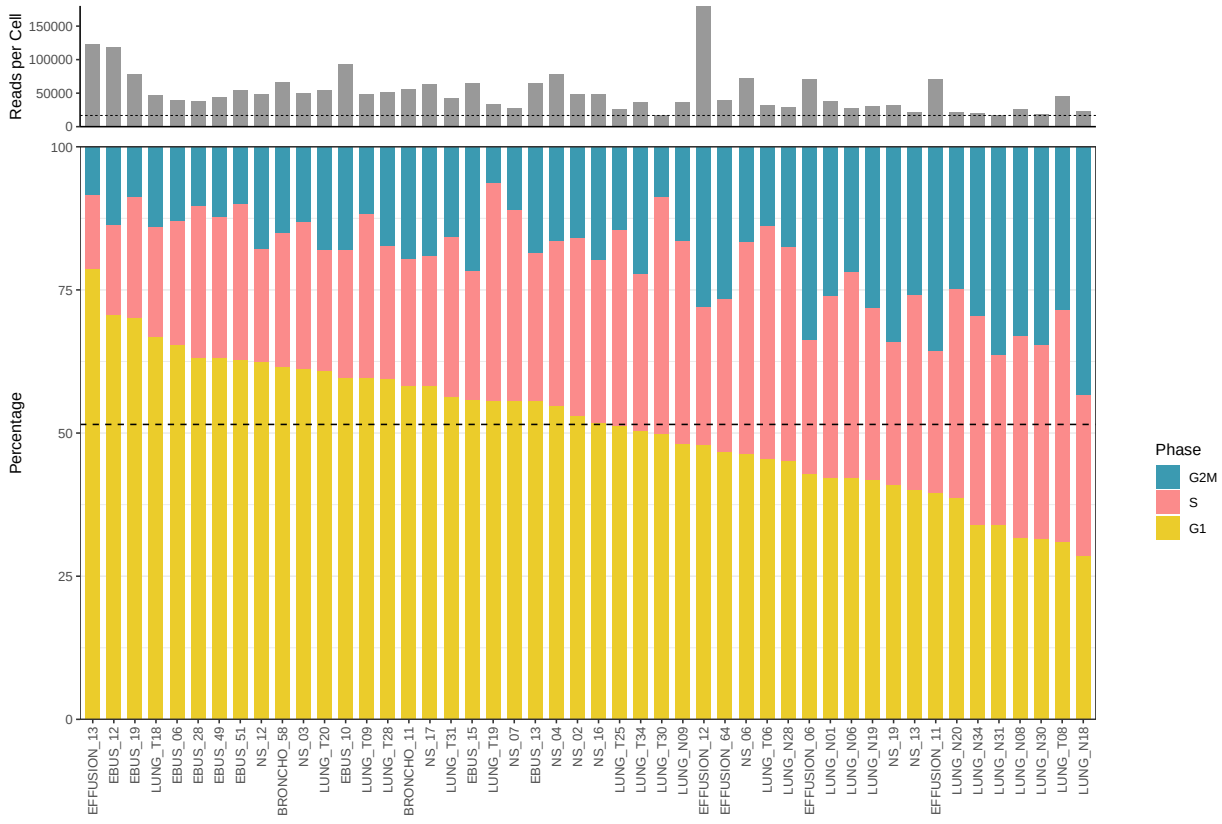

Figura 54 – Estados do ciclo celular para 48 amostras de adenocarcinoma. Os estados do ciclo celular são divididos em *G2/M*, *S* e *G1*. Há um gráfico de barras na parte superior indicando o número de *reads* por célula para cada amostra.

Após determinar o estado do ciclo celular para todos os casos, podemos usar o *PCA* colorindo pela expressão dos genes do ciclo celular. Se as células se agruparem por estado específico do ciclo celular no *PCA*, então devemos retirar o efeito da variação do ciclo celular a partir de uma regressão linear. Podemos observar na Figura 55A o *PCA* para as amostras clínicas, onde, na esquerda, inicialmente há uma clara separação de grupos *G1*, *G2/M* e *S* (sem correção do ciclo celular), sendo que a direita há o *PCA* com a correção do ciclo celular com uma melhor sobreposição dos três estados celulares, removendo a variação indesejada. Também se observa na Figura 55B o *PCA* para as amostras *CDX/PDX*. Do lado esquerdo sem a correção e do lado direito após a correção, seguindo o mesmo padrão das amostras clínicas descrito anteriormente. Esta etapa é muito importante pois a matriz dos dados normalizados corrigidos pelo ciclo celular será usada em todos os passos seguintes (redução de dimensão, clusterização etc.).

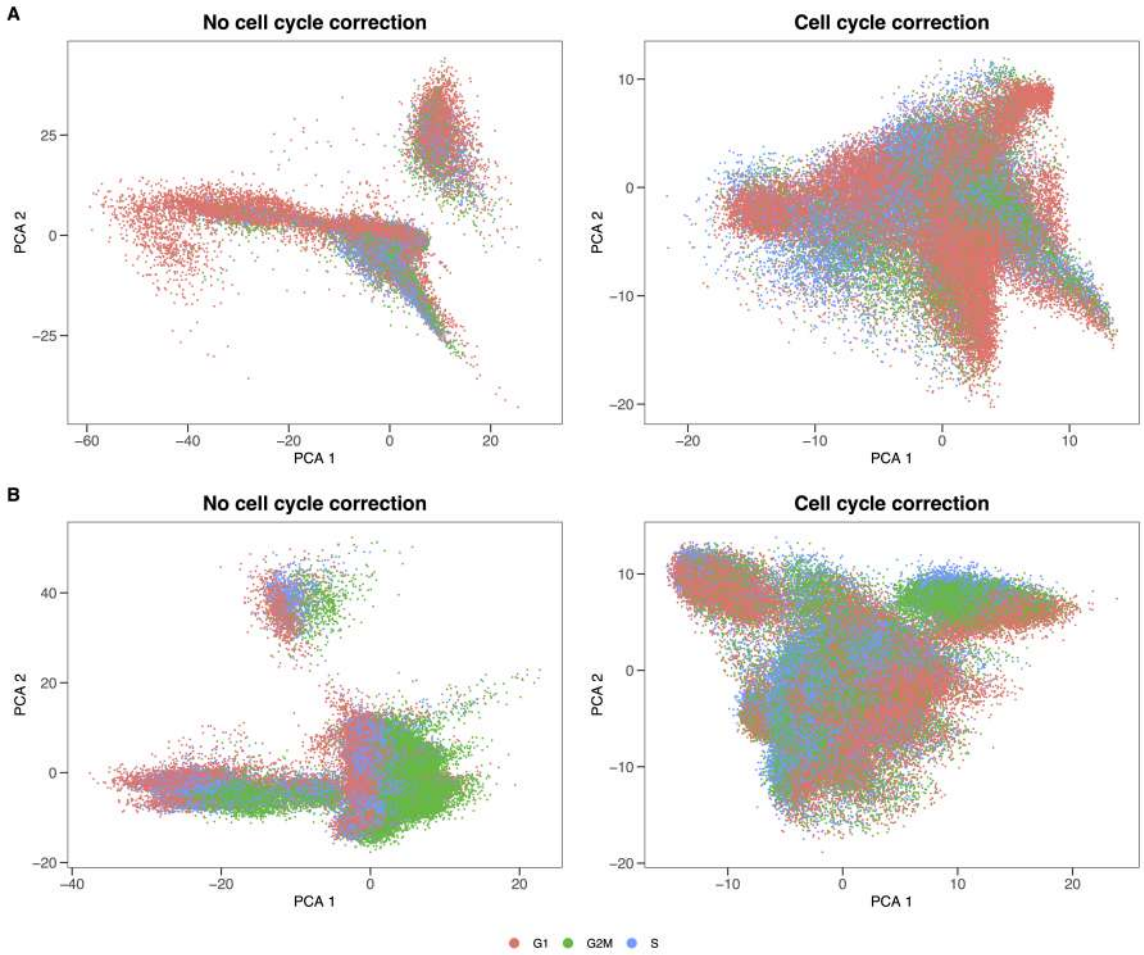

Figura 55 – *PCA* dos estados do ciclo celular para as amostras (*A*) clínicas e (*B*) *CDX/PDX*. À esquerda os *PCAs* sem a correção e na parte da direita os *PCAs* após a correção da variação do ciclo celular.

4.2.3 Identificação dos dupletos

Os dupletos foram identificados para as amostras clínicas e *CDX/PDX*. De um total de 71.334 células para as amostras clínicas foram detectados 3.950 dupletos, ficando assim com 67.384 células de amostras clínicas no total. Em contrapartida, de um total de 154.005 células para as amostras *CDX/PDX* foram detectados 7.515 dupletos, ficando assim com 146.490 células para as amostras *CDX/PDX* no total. Podemos observar graficamente esses valores a partir da Figura 56 onde são destacados na cor vermelha os dupletos para as amostras clínicas e *CDX/PDX*.

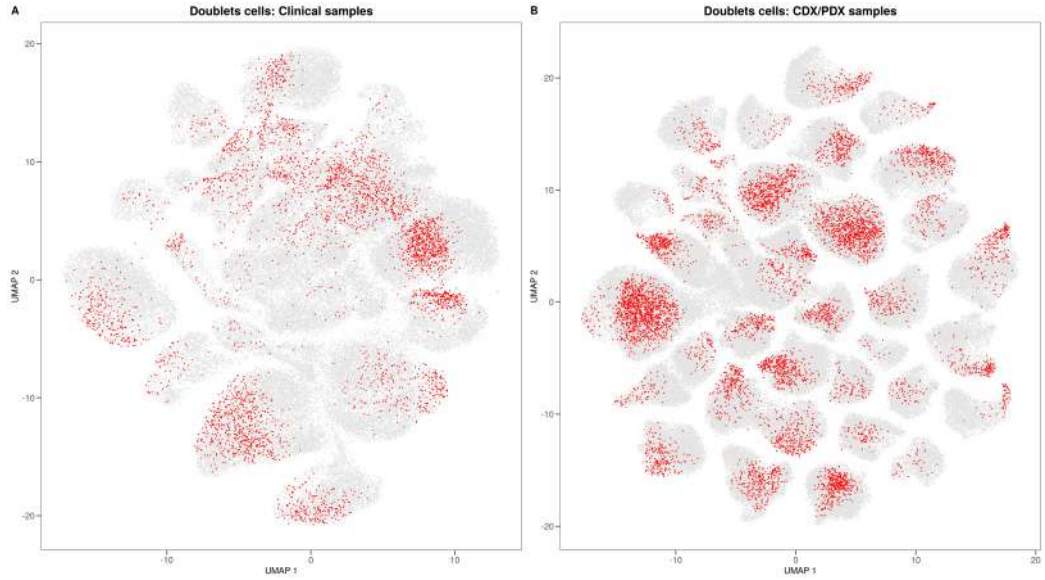

Figura 56 – *UMAP* representando os dupletos detectados (coloridos em vermelho) para as amostras (*A*) clínicas e (*B*) *CDX/PDX*.

4.2.4 *Score* de heterogeneidade intratumoral (*ITH*)

Verificamos o score de *ITH* neste conjunto de dados. Como esperado, amostras clínicas de pacientes (Figura 57) revelaram níveis mais altos de *ITH* principalmente devido à presença de células não tumorais. Entretanto, os tumores de pacientes de CPPC não diferiram significativamente em seu score *ITH* dos adenocarcinomas pulmonares. Foi também calculado o *ITH* para uma amostra de células *PC9*, que são derivadas de um adenocarcinoma humano de tecido pulmonar que permanece indiferenciado.

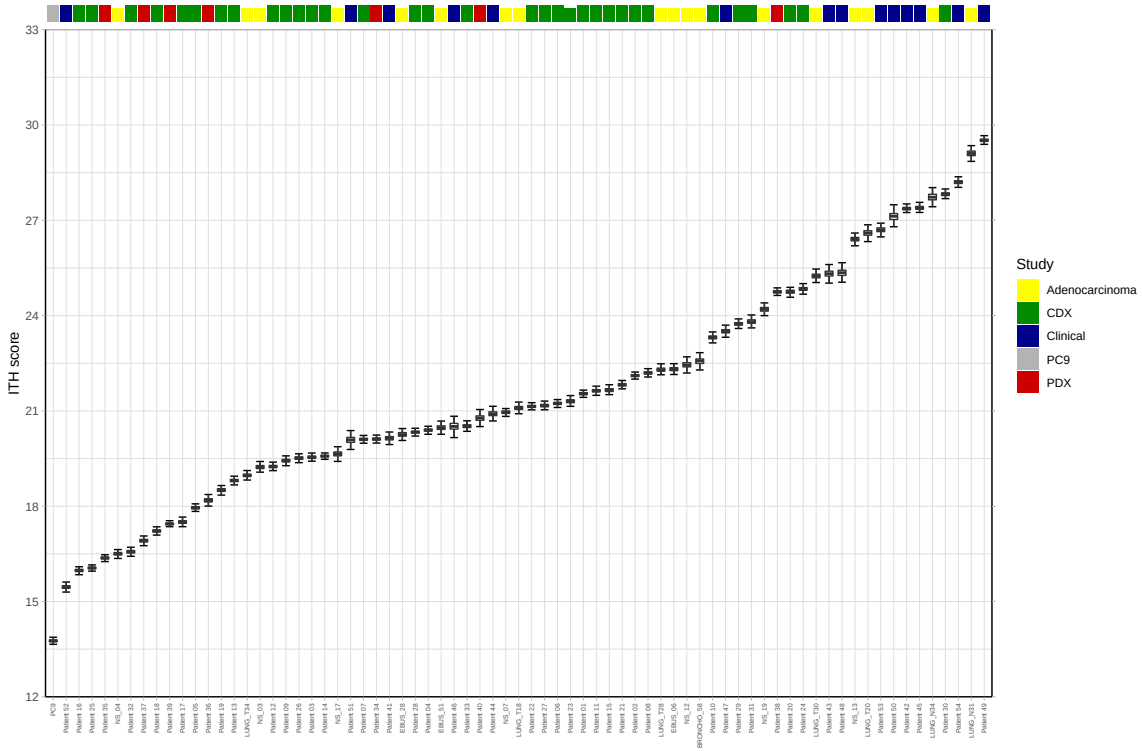

Figura 57 – Score de *ITH* para as amostras de CPPC (*CDX* em verde, *PDX* em vermelho e Clínicas em azul), adenocarcinoma pulmonar (em amarelo) e uma amostra PC9 (em cinza).

4.2.5 Correção do *batch*

A correção do batch, considerando as diferentes versões das químicas utilizadas para o sequenciamento, foi feita para as amostras clínicas e CDX/PDX. Nota-se, na Figura 58, que antes da correção do *batch* as células da versão 2 e da versão 3 estão em sua maioria separadas, agrupadas entre si. Após a correção do *batch* é possível observar que o efeito das diferentes químicas (v2 e v3) foi corrigido, dessa forma integrando e misturando melhor as células. O *batch* também foi corrigido quando integradas todas as 54 amostras como pode ser observado na Figura 59.

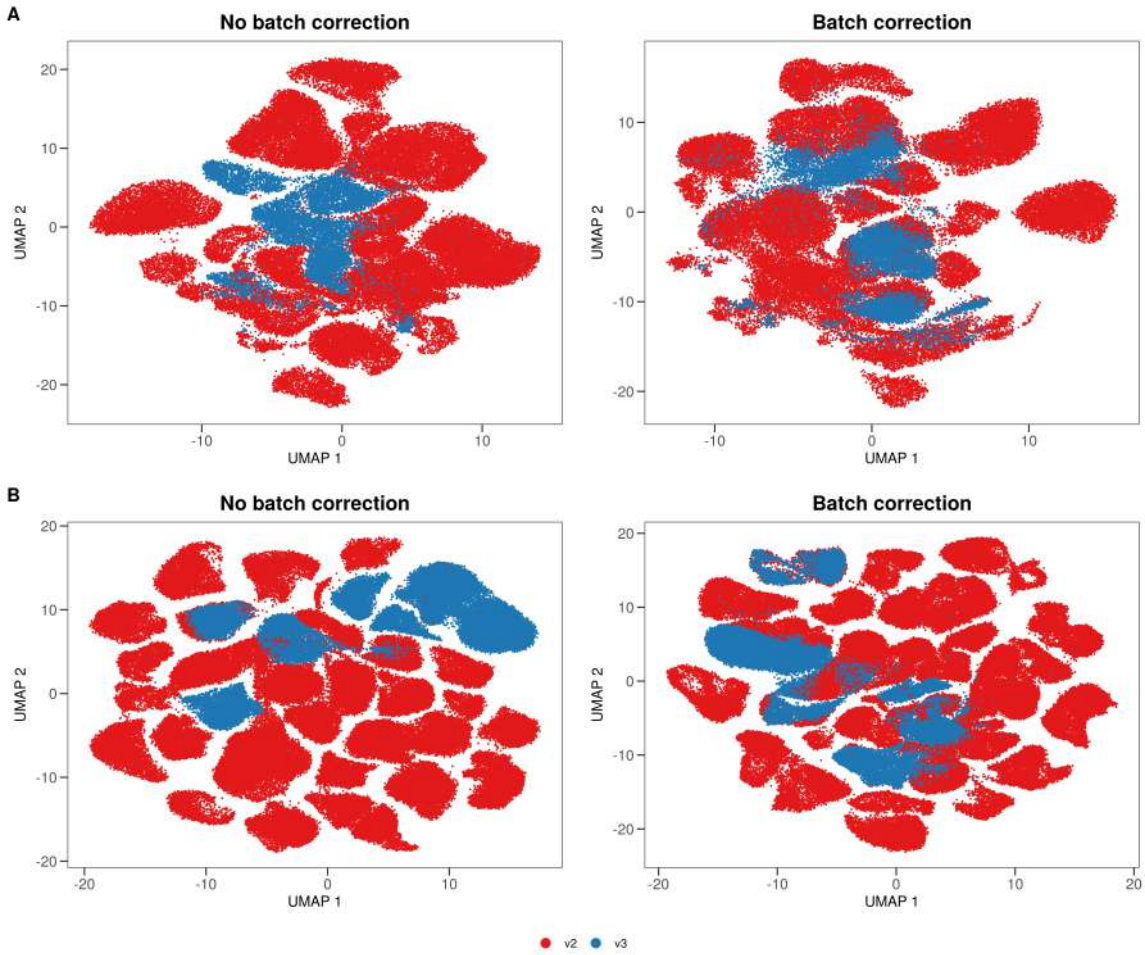

Figura 58 – *UMAP* da correção do *batch* considerando a versão da química do sequenciamento (v2 e v3) para as amostras (*A*) clínicas e (*B*) *CDX/PDX*. À esquerda os *UMAPs* sem a correção e na parte da direita os *UMAPs* após a correção do *batch*. Os pontos em vermelho representam as células correspondentes à química da versão v2 e os ponto em azul representam as células correspondentes à química da versão v3.

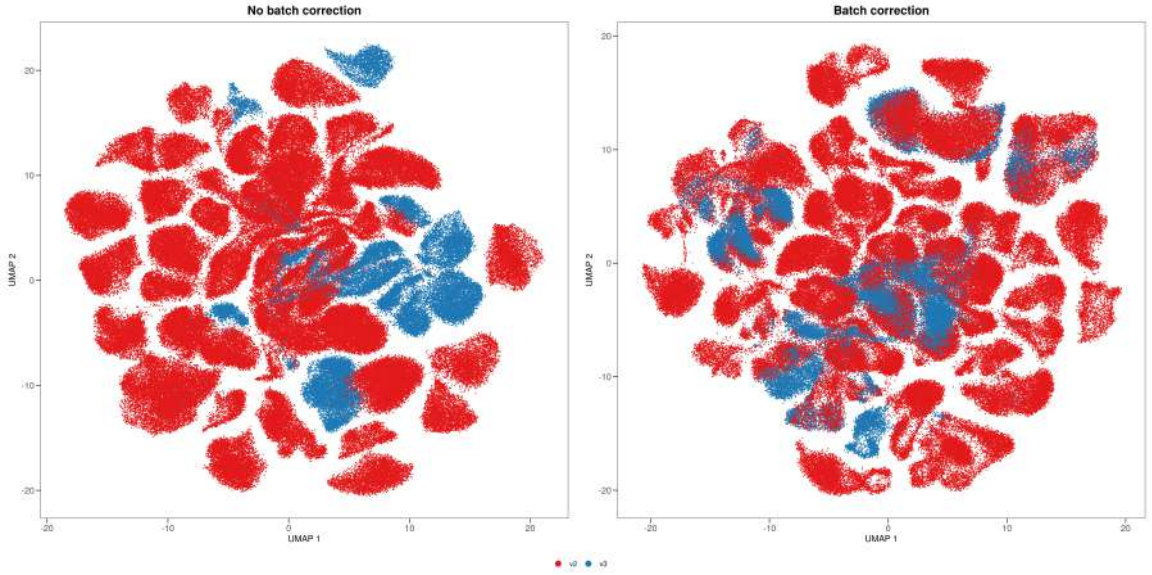

Figura 59 – *UMAP* da correção do *batch* considerando a versão da química do sequenciamento (v2 e v3) para todas as 54 amostras integradas. À esquerda o *UMAP* sem a correção e na parte da direita o *UMAP* após a correção do *batch*. Os pontos em vermelho representam as células correspondentes à química da versão v2 e os ponto em azul representam as células correspondentes à química da versão v3.

4.2.6 *scvis*

Como uma visualização alternativa ao *UMAP*, foi calculada a redução *scvis*. Na Figura 60 observamos as amostras clínicas, na Figura 61 as amostras *CDX/PDX* e na Figura 62 todas as 54 amostras. O *scvis* foi calculado pois tem a capacidade de preservar tanto as estruturas vizinhas locais quanto globais.

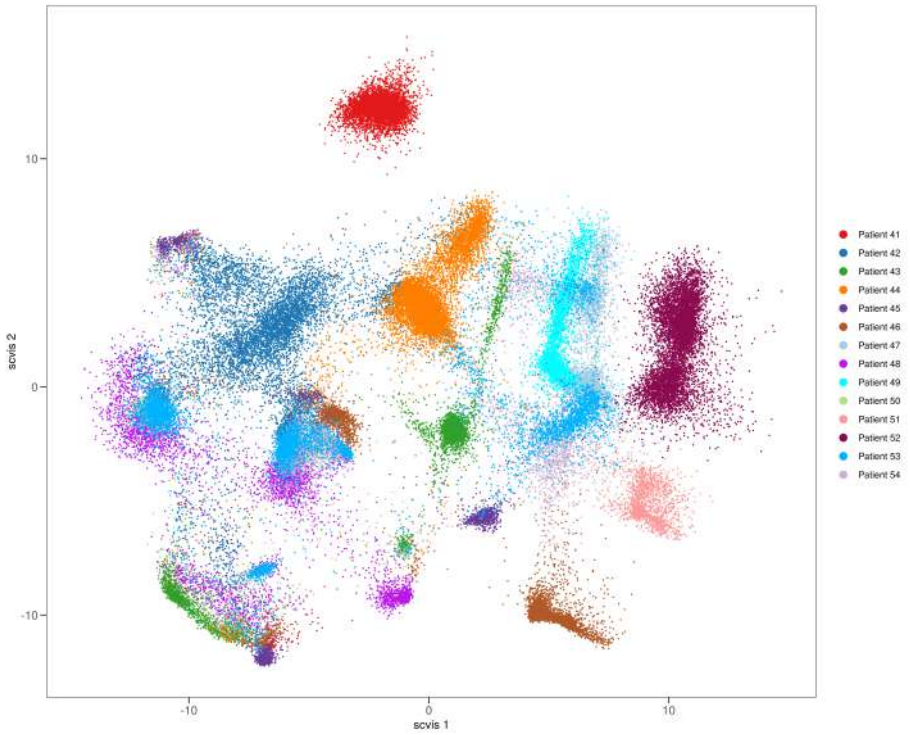

Figura 60 – *scvis* das 14 amostras clínicas coloridas de acordo com os nomes das amostras.

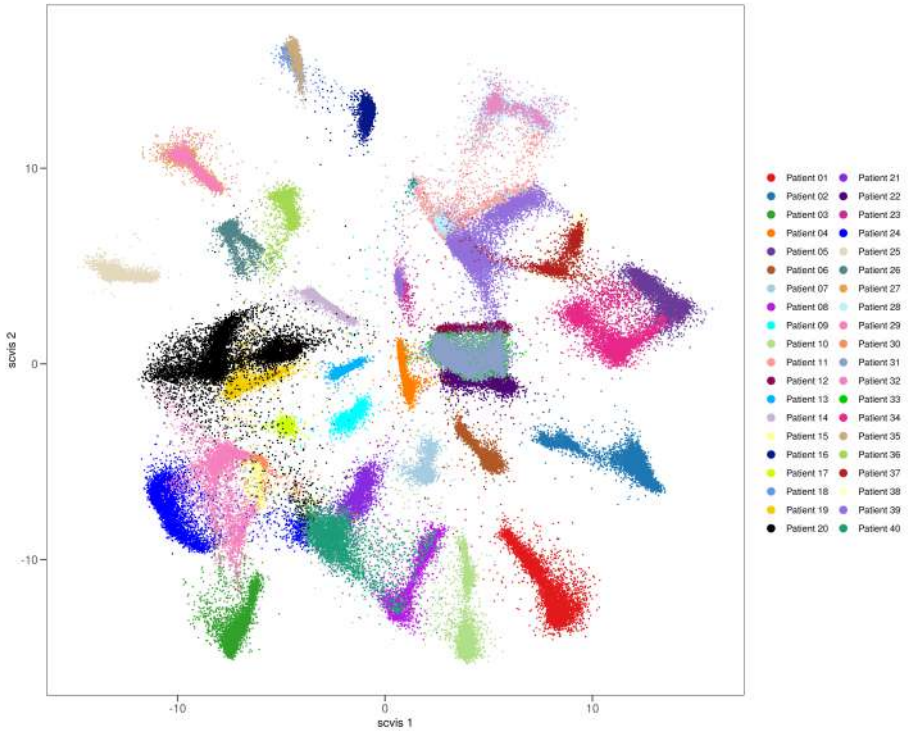

Figura 61 – *scvis* das 40 amostras *CDX/PDX* coloridas de acordo com os nomes das amostras.

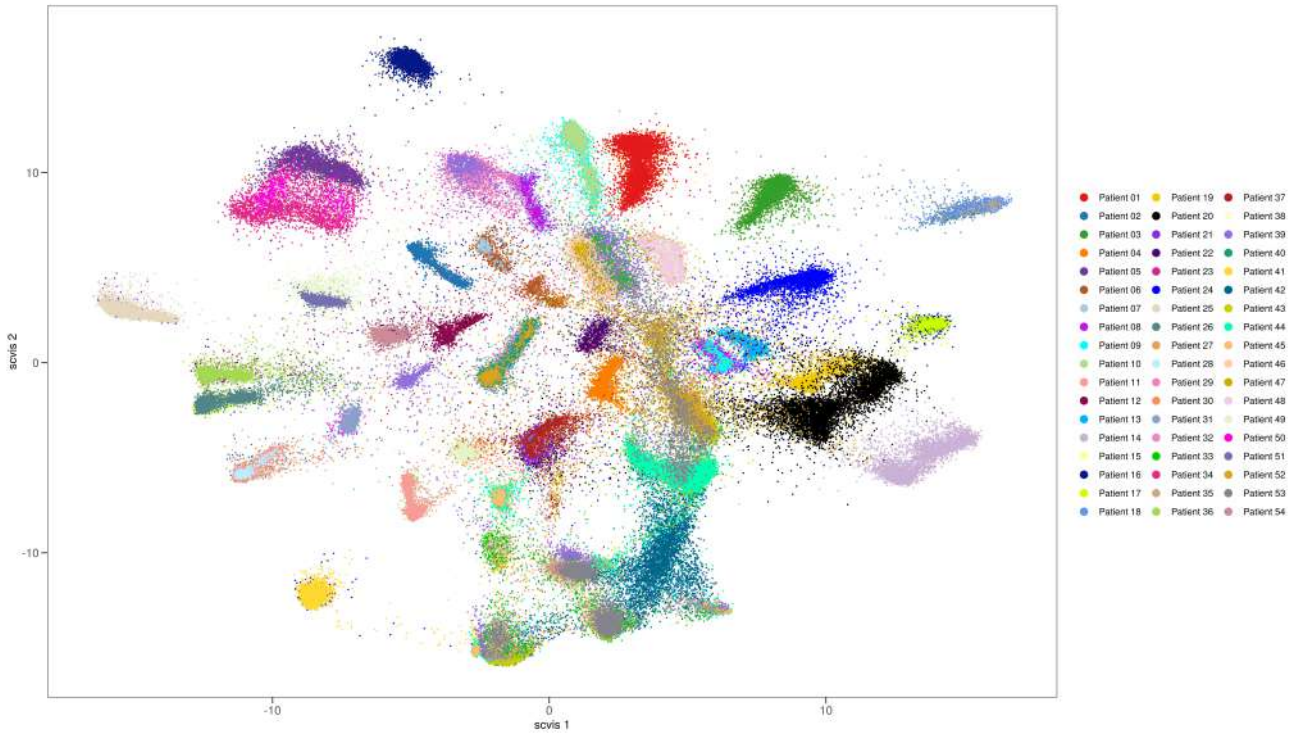

Figura 62 – *scvis* das de todas as 54 amostras coloridas de acordo com os nomes das amostras.

4.2.7 Classificação dos tipos celulares

Para a identificação dos tipos celulares primeiramente foi feita a clusterização dos dados quando integradas as amostras Clínicas e *CDX/PDX*, encontrando 57 *clusters* (Figura 63). Em seguida foram identificados os marcadores, a partir da análise de expressão diferencial, para cada *cluster*. Os marcadores encontrados ajudaram na identificação dos tipos celulares que foram divididos em células epiteliais (*EPCAM*, *TNNC2*, *HMGCS1*), imunes (*PTPRC*) e fibroblastos (*TNFRSF12A*). A expressão desses marcadores pode ser observada na Figura 64.

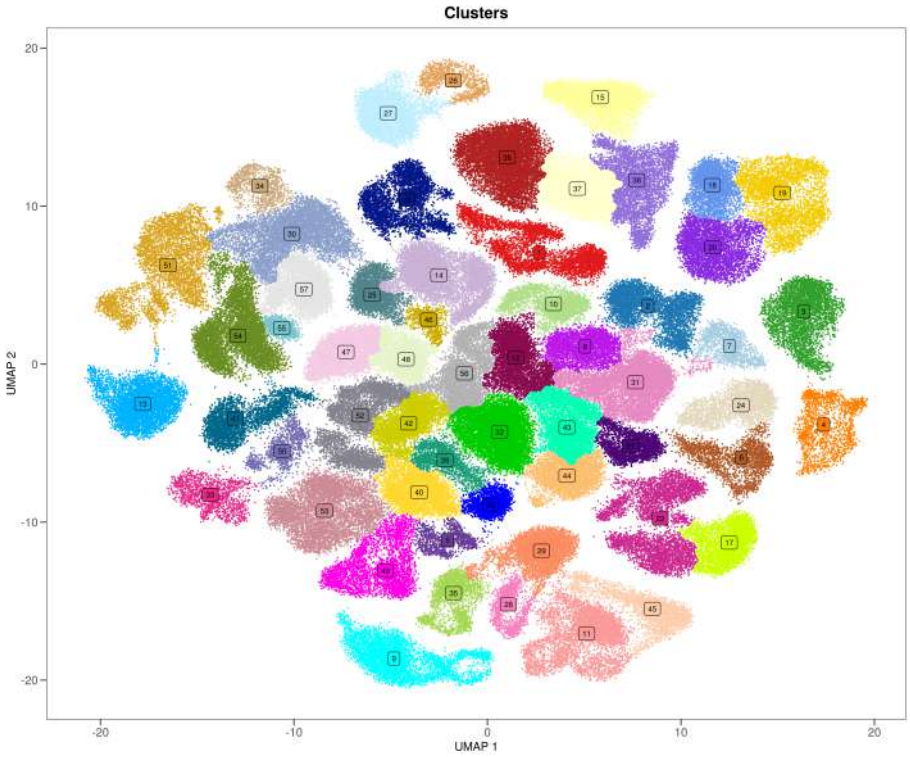

Figura 63 – *Clusters* identificados nas amostras *CDX/PDX* e clínicas integradas. Foram identificados 57 *clusters* e seus respectivos marcadores.

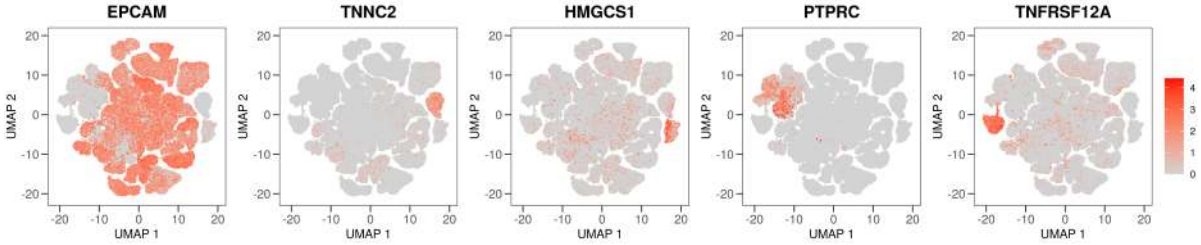

Figura 64 – Principais marcadores utilizados para identificação dos tipos celulares. Os marcadores *EPCAM*, *TNNC2* e *HMGCS1* são exclusivamente de células epiteliais, enquanto o marcador *PTPRC* identifica células imunes e, por fim, o marcador *TNFRSF12A* identifica os fibroblastos.

Com isso foi possível dividir todas as 54 amostras de CPPC em três tipos celulares mais abrangentes: células epiteliais, células imunes e fibroblastos. Na Figura 65 é possível observar que há a predominância de células epiteliais pelo fato de a grande maioria das amostras desse estudo serem provenientes do modelo *CDX/PDX*. Em contrapartida, as células imunes são provenientes diretamente dos tumores primários das amostras clínicas de pacientes.

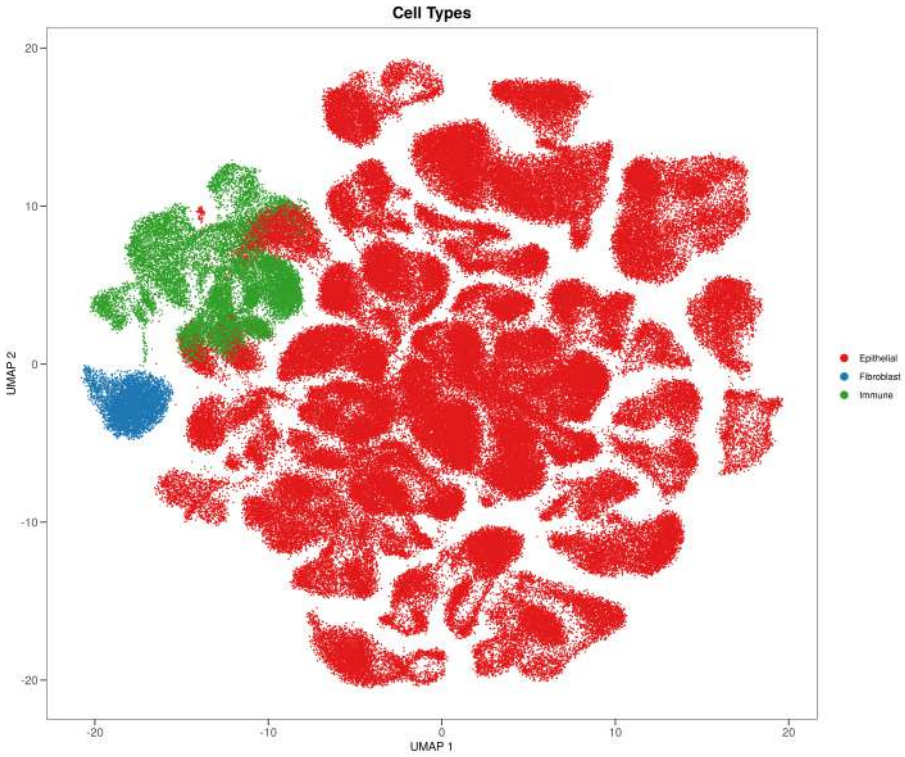

Figura 65 – Tipos celulares identificados após a clusterização e identificação dos marcadores. Foram identificados três principais tipos celulares: células epiteliais, imunes e fibroblastos.

4.2.8 Subtipos moleculares

Até agora, os estudos sobre a heterogeneidade transcricional se concentraram nos 4 fatores de transcrição *ASCL1*, *NEUROD1*, *POU2F3* e *YAP1* (RUDIN *et al.*, 2019). Dessa forma, determinamos inicialmente a presença principais *TFs* em nossos conjuntos de dados exclusivamente nas amostras *CDX/PDX* (Figura 66).

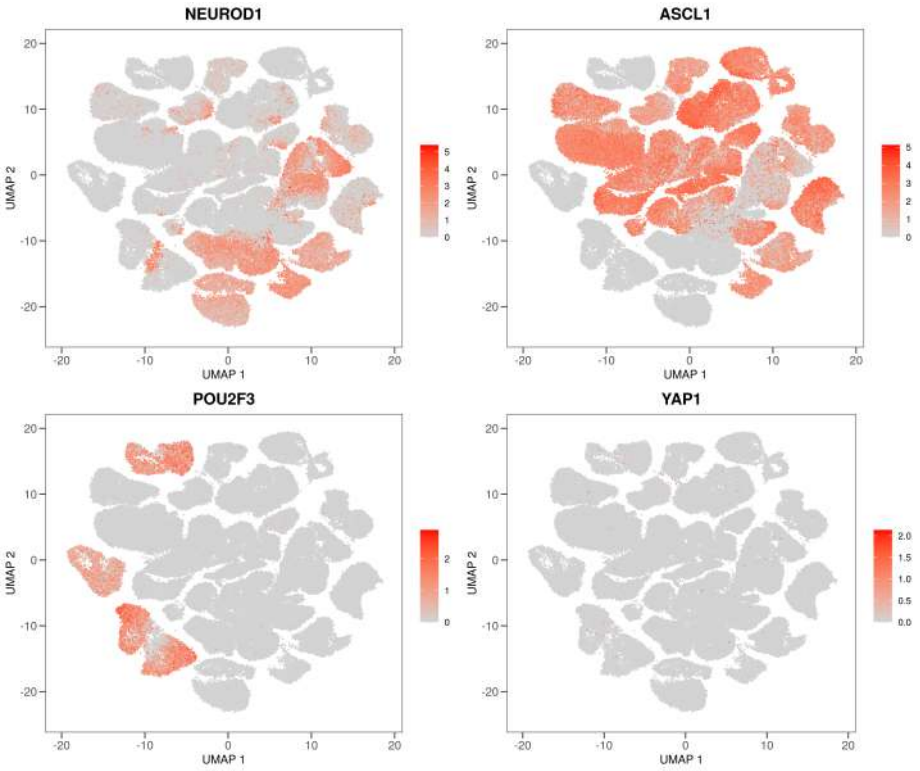

Figura 66 – Expressão dos 4 fatores de transcrição (*NEUROD1*, *ASCL1*, *POU2F3* e *YAP1*) nas amostras *CDX/PDX*.

Ainda que a atual classificação sugira uma classificação rigorosa baseada em um destes 4 *TFs*, os dados deste estudo também revelaram a expressão de mais de 1 *TF* em cada caso (Figura 67). Pode-se pontuar observando o gráfico de violino, por exemplo, que os pacientes 03 e 19 estão coexpressando os fatores de transcrição *NEUROD1* e *ASCL1*.

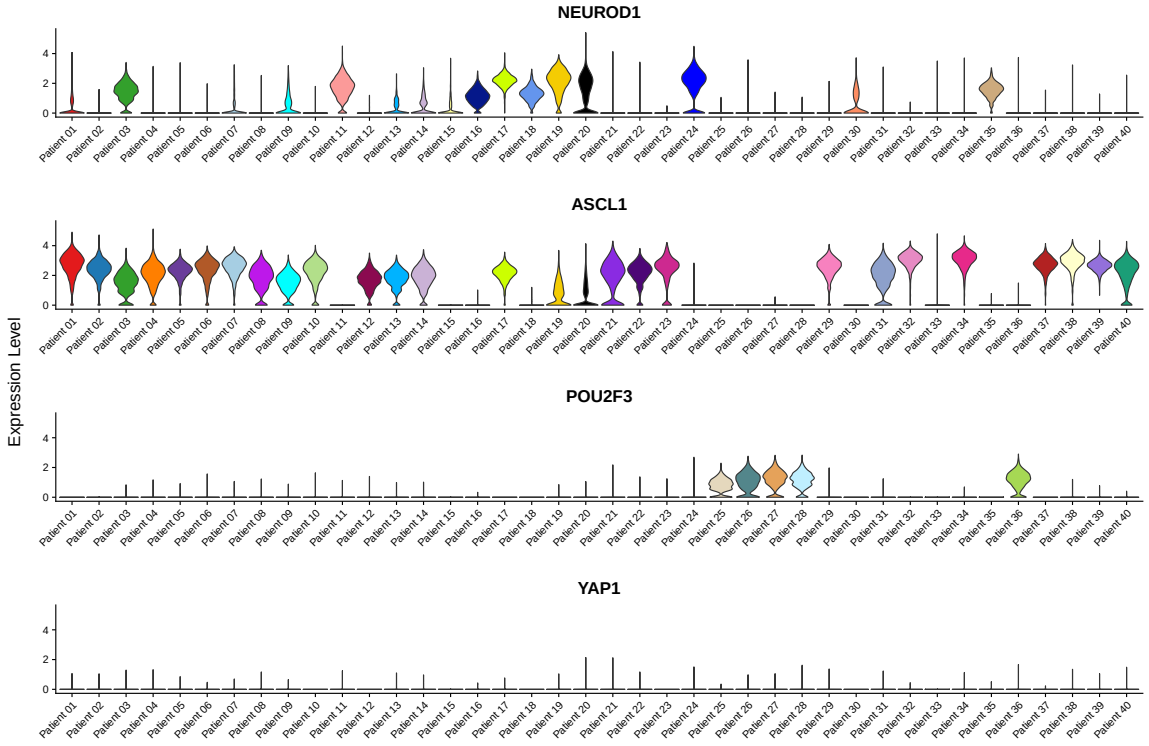

Figura 67 – Gráfico de violino mostrando a expressão detalhada para cada um dos 40 pacientes provenientes das amostras *CDX/PDX* para os quatro fatores de transcrição (*NEUROD1*, *ASCL1*, *POU2F3* e *YAP1*).

Em seguida, determinamos a quantidade de células com níveis substanciais de um dos *TFs* dominantes e identificamos que em todos os tumores, as células coexpressavam pelo menos 2 *TFs* com *ASCL1/NEUROD1* sendo a combinação mais abundante seguida por *NEUROD1/POU2F3* (Figura 68). A observação de que mais de 1 *TF* está presente em uma amostra e mesmo em células únicas sugere um estado de alta plasticidade entre as células tumorais. Além disso, os dados de células únicas (*CDX/PDX*) indicam uma fração celular que é negativa para todos os 4 *TFs*. Como as células foram inicialmente filtradas com base em sua qualidade, pode-se excluir que esta fração consiste em células de má qualidade.

Nossos dados não incluíram nenhum tumor com expressão *YAP1*; vale ressaltar que a expressão de *YAP1* está presente no CPPC primário em apenas 2,4% dos casos. E expressão de *YAP1*, entretanto, é alto em outros tipos de câncer de pulmão (LEE *et al.*, 2017; TSUJI *et al.*, 2020). Embora estes dados mostrem características transcritivamente distintas de *TF* dentro de um tumor, os mecanismos subjacentes que contribuem para isto não são bem compreendidos.

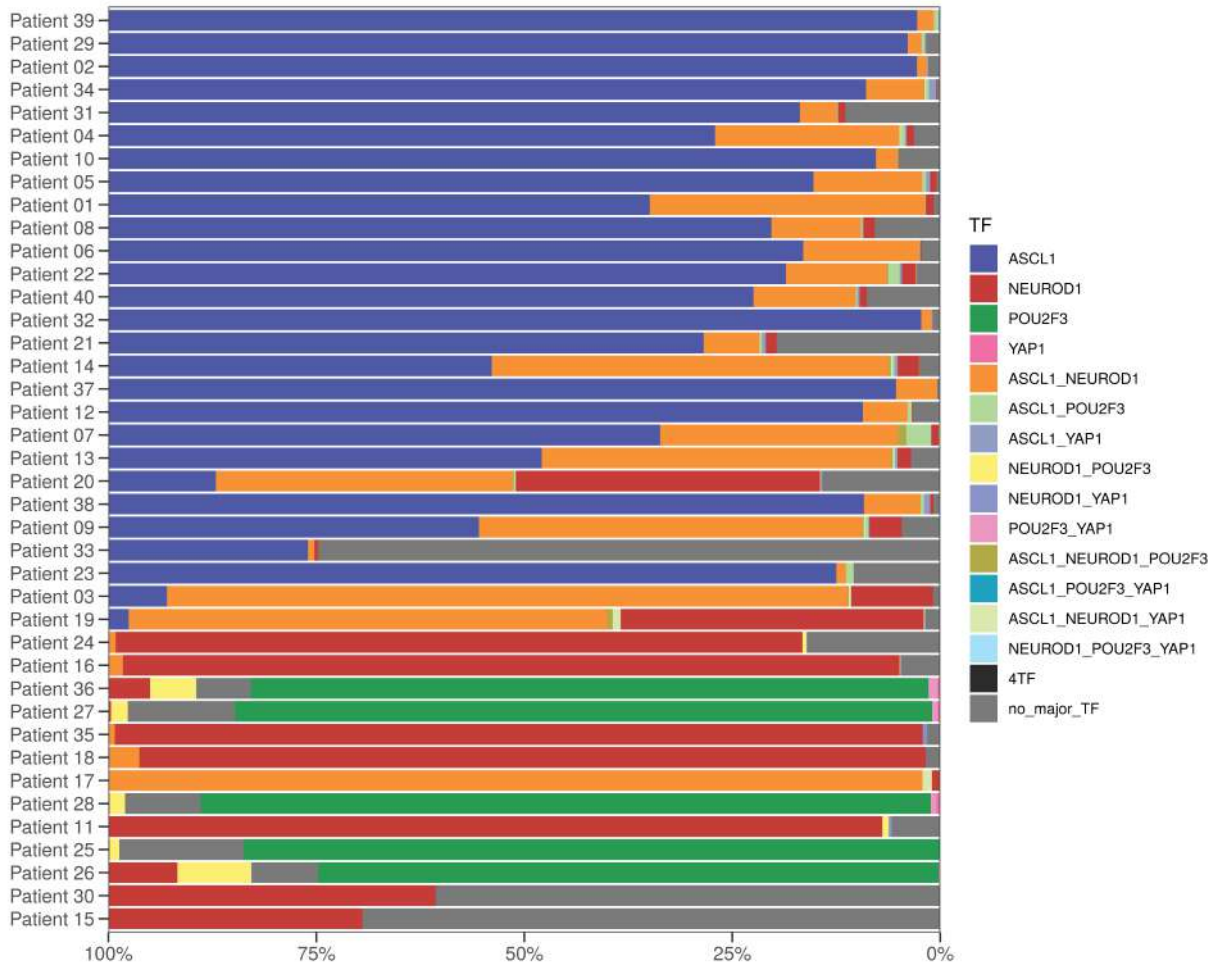

Figura 68 – Porcentagem de células expressando os quatro fatores de transcrição individualmente ou coexpressando pelo menos 2 *TFs*.

4.2.9 Estimativa da velocidade do *RNA* em diferentes casos

Para esse tópico iremos considerar 3 diferentes casos de amostras de tumores combinados de pacientes que existiam ou como análises de amostras múltiplas de pacientes sem tratamento, ou tumores longitudinais combinados durante todo o tratamento. Todos esses 3 casos são de *timepoints* distintos.

O primeiro caso corresponde aos pacientes 05 e 34 que são amostras de pacientes que não receberam tratamento. A proporção de *spliced* e *unspliced* para esse caso é uma média de 60% e 40%, respectivamente. A distribuição por *clusters* podem ser observadas na Figura 69. Os *UMAPs* coloridos pelos diferentes *clusters* (esquerda) e pacientes (direita), podem ser observados na Figura 70. Nota-se que a trajetória se inicia no *cluster* 7. De

acordo com a Figura 68 os pacientes 05 e 34 possuem predominantemente a expressão alta de *ASCL1*, com coexpressão de *ASCL1+NEUROD*. Na Figura 71 observa-se a relação da expressão entre *spliced*/*unspliced*, a velocidade (dinâmica) para essas amostras além da expressão para os *TFs* mais os genes *NOTCH* e *REST*. Apesar da expressão de *ASCL1* ser alta para essas amostras, a velocidade não se move da mesma maneira, mostrando assim uma baixa mudança na abundância de *RNA* mensageiro. No gráfico central da Figura 71 é possível observar, para *ASCL1*, uma velocidade maior apenas na parte do meio, sendo que o restante das células possui uma velocidade menor. Outro ponto importante é em relação aos genes *NOTCH1* e *REST*. Nesse caso em que os pacientes não estão recebendo tratamento a expressão e velocidade desses genes são baixas.

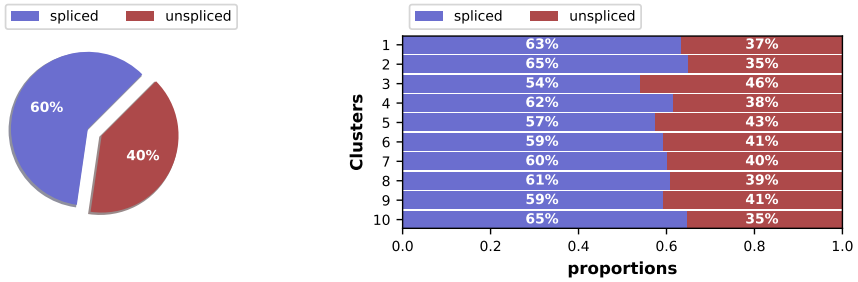

Figura 69 – Proporção média de *spliced* e *unspliced* (esquerda) e proporção para cada *cluster* (direita) identificado para os pacientes 05 e 34.

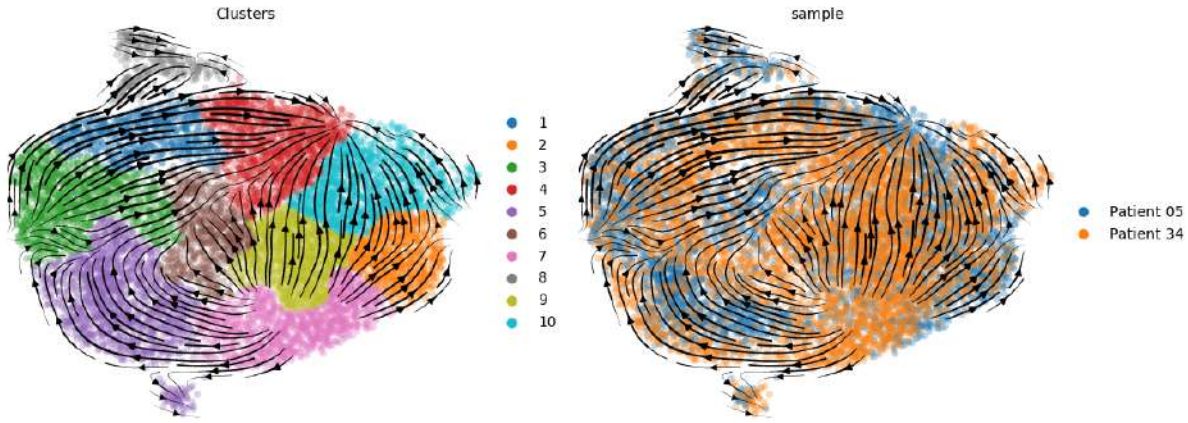

Figura 70 – *UMAP* mostrando a dinâmica do *RNA* para os pacientes 05 e 34 coloridos pelos *clusters* (esquerda) e pelos diferentes pacientes (direita).

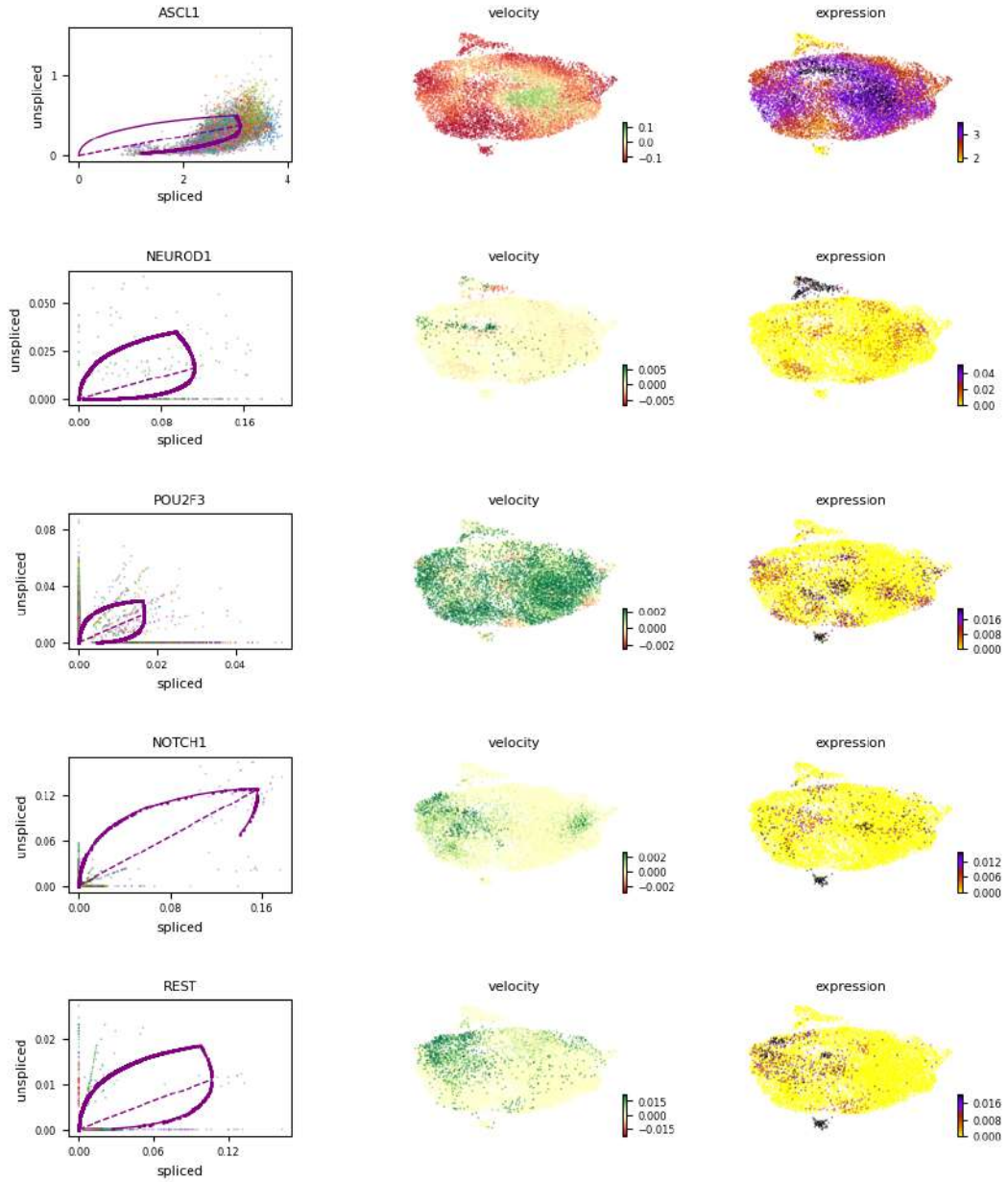

Figura 71 – Gráfico de relação entre a expressão do *spliced* e *unspliced* para os genes *ASCL1*, *NEUROD1*, *POU2F3*, *NOTCH1* e *REST*, além da velocidade que esses genes se movem e a expressão em todas as células. Esses gráficos são referentes às amostras dos pacientes 05 e 34.

O segundo caso corresponde aos pacientes 21 e 25 que são amostras de pacientes que estão em processo de tratamento com inibidores *PARP*. A enzima poli-(*ADP*)-sibose polimerase (*PARP*) funciona para reparar quebras de *DNA* e é um alvo terapêutico promissor com taxas de resposta modestas nos primeiros ensaios clínicos. A proporção de *spliced* e *unspliced* para esse caso é uma média de 70% e 30%, respectivamente. A distribuição por *clusters* podem ser observadas na Figura 72. Os *UMAPs* coloridos pelos diferentes *clusters* (esquerda) e pacientes (direita), podem ser observados na Figura 73.

Nota-se que a trajetória se inicia entre os *clusters* 3 e 7. Para confirmar, na Figura 74 observamos os valores de *pseudotime*, sendo os menores valores correspondentes ao início da trajetória. De acordo com a Figura 68 os pacientes 21 e 25 possuem predominantemente a expressão alta de *ASCL1* e *POU2F3*, respectivamente. Na Figura 75 observa-se a relação da expressão entre *spliced*/*unspliced*, a velocidade (dinâmica) para essas amostras além da expressão para os *TFs* mais os genes *NOTCH* e *REST*. Há uma alta expressão de *NOTCH* e *REST* para estes pacientes além de uma velocidade alta em diferentes partes da trajetória mostrando alta mudança de abundância de *RNA* mensageiro, tornando muito interessante os dados, pois de acordo com Marignol (2017), a interação *PARP* com a via de sinalização *Notch* na leucemia linfoblástica aguda de células B foi proposta para prejudicar a sinalização de *HES-1* e a indução de apoptose (KANNAN *et al.*, 2011), sendo justificada a avaliação dessas interações no CPPC.

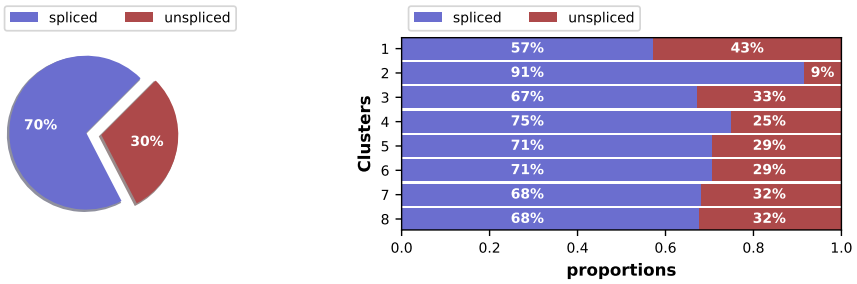

Figura 72 – Proporção média de *spliced* e *unspliced* (esquerda) e proporção para cada *cluster* (direita) identificado para os pacientes 21 e 25.

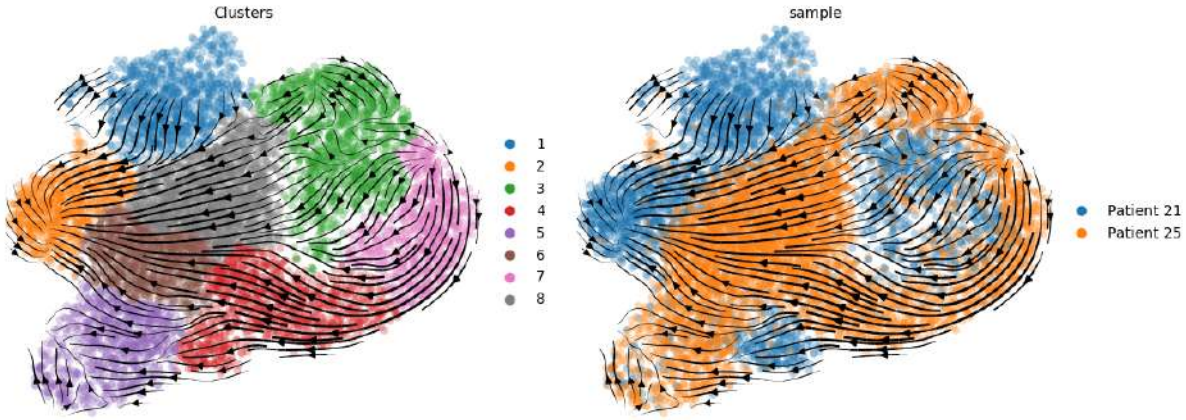

Figura 73 – *UMAP* mostrando a dinâmica do *RNA* para os pacientes 21 e 25 coloridos pelos *clusters* (esquerda) e pelos diferentes pacientes (direita).

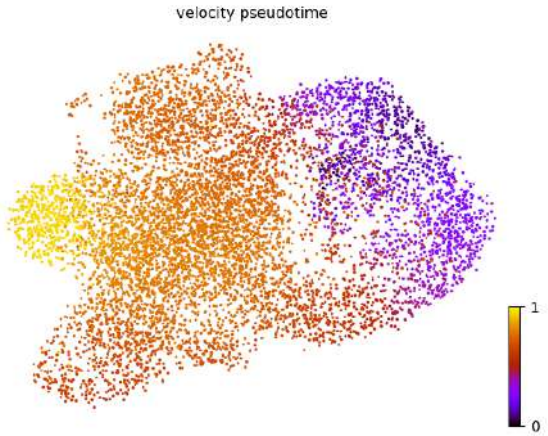

Figura 74 – *Pseudotime* mostrando o início da trajetória (menores valores da escala com cores escuras) até o final (maiores valores da escala com cores claras) projetadas no *UMAP*.

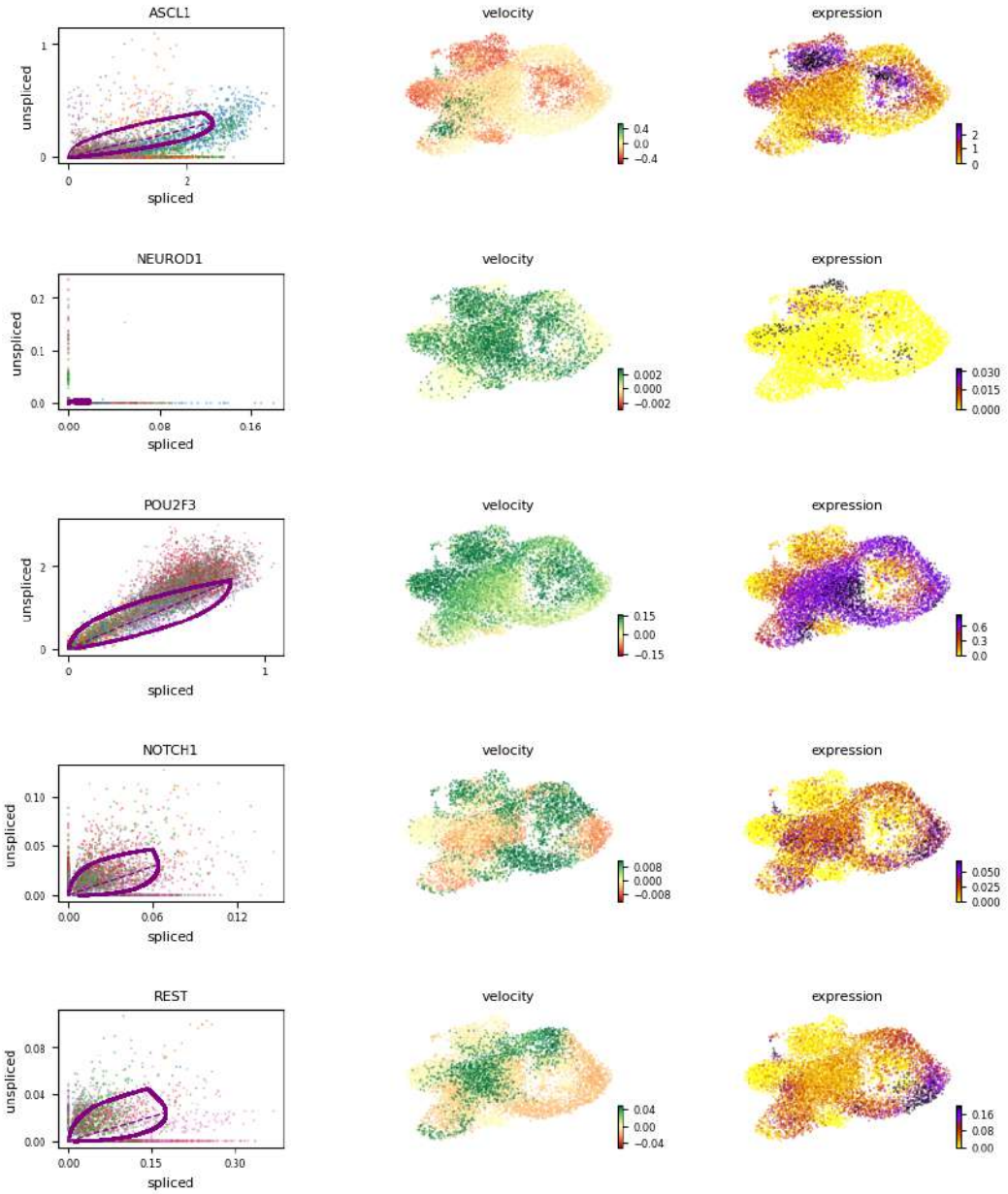

Figura 75 – Gráfico de relação entre a expressão do *spliced* e *unspliced* para os genes *ASCL1*, *NEUROD1*, *POU2F3*, *NOTCH1* e *REST*, além da velocidade que esses genes se movem e a expressão em todas as células. Esses gráficos são referentes às amostras dos pacientes 21 e 25.

Por fim, o terceiro caso corresponde aos pacientes 18 e 35 que são amostras de pacientes que estão em recidiva. A proporção de *spliced* e *unspliced* para esse caso é uma média de 47% e 53%, respectivamente. A distribuição por *clusters* podem ser observadas na Figura 76. Os *UMAPs* coloridos pelos diferentes clusters (esquerda) e pacientes (direita), podem ser observados na Figura 77. Nota-se que a trajetória se inicia no *cluster* 1. Para confirmar, na Figura 78 observamos os valores de *pseudotime*, sendo os menores valores correspondentes ao início da trajetória. De acordo com a Figura 68 os pacientes 18 e 35

possuem predominantemente a expressão alta de *NEUROD1*. Na Figura 79 observa-se a relação da expressão entre *spliced*/*unspliced*, a velocidade (dinâmica) para essas amostras além da expressão para os *TFs* mais os genes *NOTCH* e *REST*. Há uma expressão considerável de *NOTCH* e *REST* para estes pacientes além de uma velocidade alta em diferentes partes da trajetória mostrando alta mudança de abundância de *RNA* mensageiro. Tal fato pode sugerir uma trajetória complementar entre *REST* e *NOTCH*.

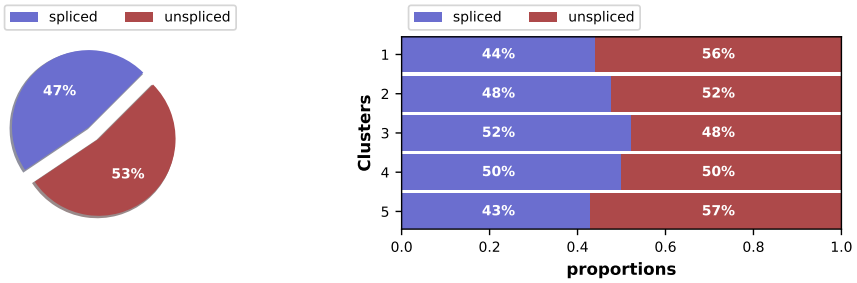

Figura 76 – Proporção média de *spliced* e *unspliced* (esquerda) e proporcão para cada cluster (direita) identificado para os pacientes 18 e 35.

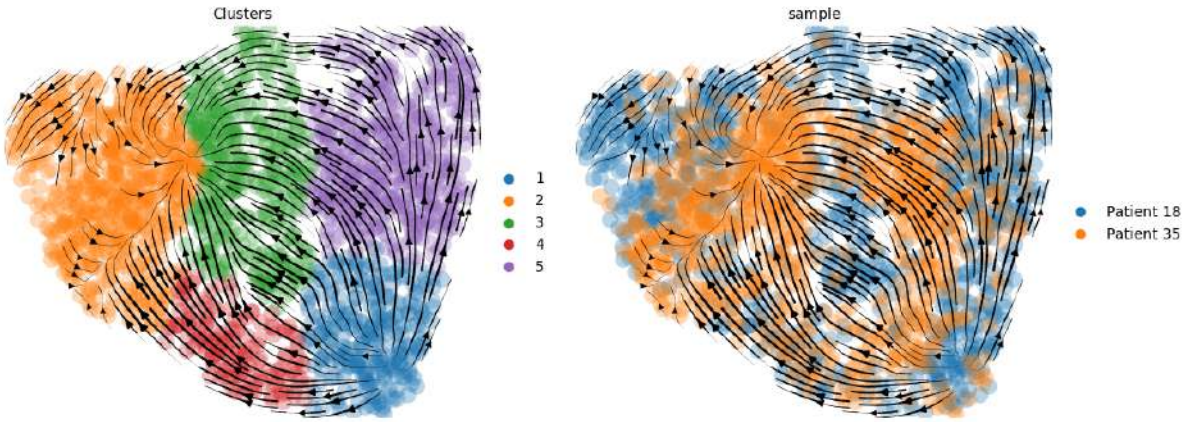

Figura 77 – *UMAP* mostrando a dinâmica do *RNA* para os pacientes 18 e 35 coloridos pelos *clusters* (esquerda) e pelos diferentes pacientes (direita).

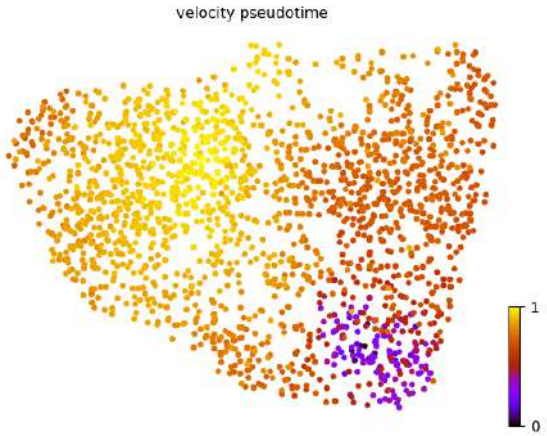

Figura 78 – *Pseudotime* mostrando o início da trajetória (menores valores da escala com cores escuras) até o final (maiores valores da escala com cores claras) projetadas no *UMAP*.

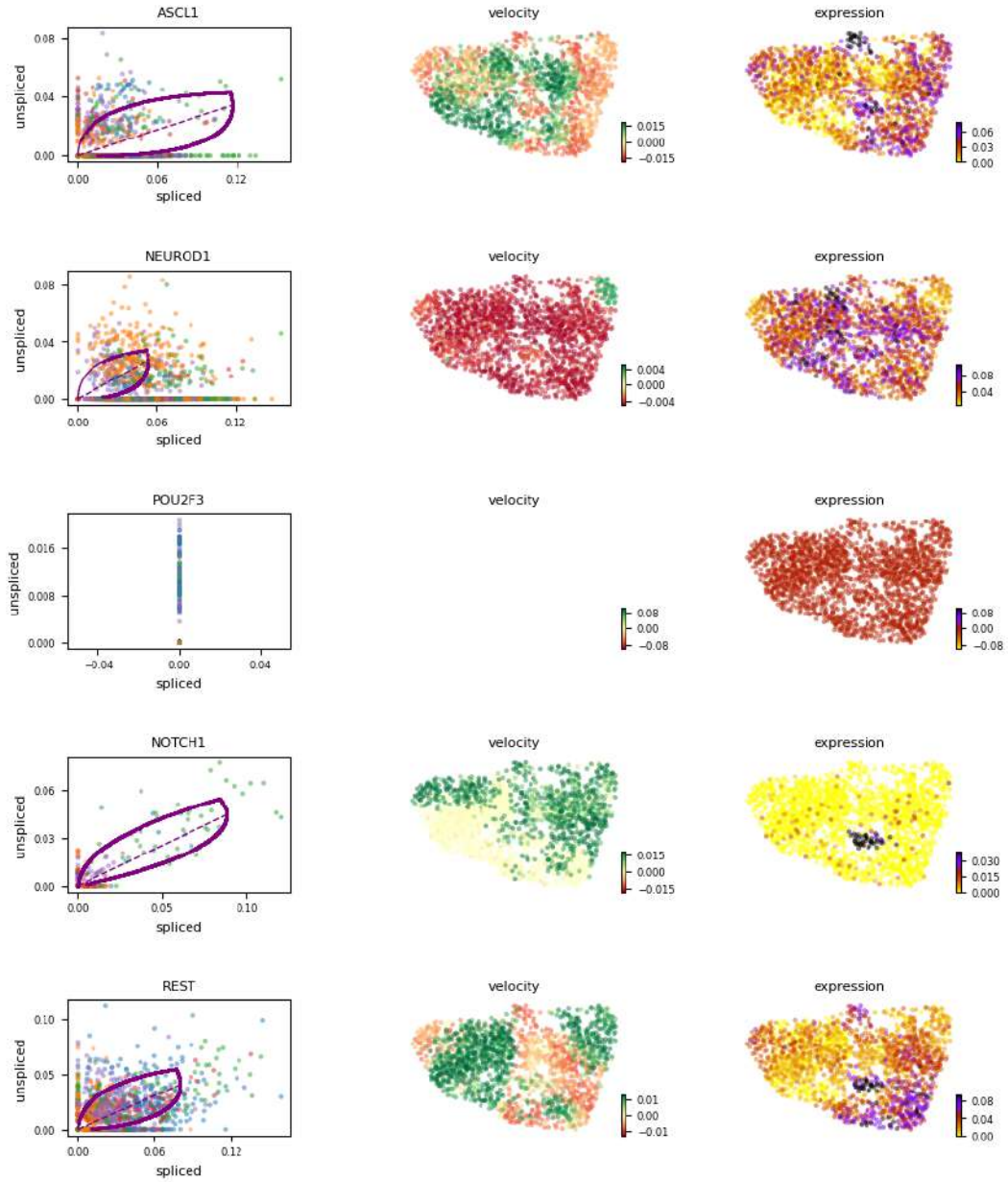

Figura 79 – Gráfico de relação entre a expressão do *spliced* e *unspliced* para os genes *ASCL1*, *NEUROD1*, *POU2F3*, *NOTCH1* e *REST*, além da velocidade que esses genes se movem e a expressão em todas as células. Esses gráficos são referentes às amostras dos pacientes 18 e 35.

4.2.10 Comunicação célula-célula

A análise de comunicação célula-célula foi feita para os mesmos 3 casos vistos anteriormente no tópico da estimativa da velocidade do RNA.

O primeiro caso é correspondente aos pacientes 05 e 34 que são amostras de pacientes que não receberam tratamento. Inicialmente essas duas amostras foram integradas e feita

a clusterização obtendo-se 10 *clusters*. Considerando os 10 *clusters* foi feita a análise de comunicação célula-célula. As vias de sinalização mais significantes foram as *MK*, *JAM* e *PTN*, como pode ser observado na Figura 80. Essas vias de sinalização mais significantes, quando visualizadas as similaridades funcionais (Figura 81), descobrimos que *MK* e *JAM* fazem parte do mesmo grupo (Grupo 2), enquanto *PTN* faz parte de um grupo separado (Grupo 3). Isso mostra que *MK* e *JAM* possuem papéis semelhantes e/ou redundantes nessas amostras. As interações entre os pares receptores-ligantes para as vias *MK*, *JAM* e *PTN* podem ser observadas na Figura 82. Observamos que os pares receptores-ligantes que possuem os maiores valores de probabilidade de comunicação têm o ligante *NCL* (*PTN-NCL* e *MDK-NCL*). Os receptores *MK* e *PTN* se ligam à *NCL*, embora com afinidade significativamente menor em relação a outros receptores de superfície celular (SAID *et al.*, 2002). A *NCL* é ubiquamente expressa e pode estar envolvida no reconhecimento de células apoptóticas precoces por macrófagos (HIRANO *et al.*, 2005). A interação entre *NCL* e *MK/PTN* promove a localização nuclear, a migração de células endoteliais e a sobrevivência celular (SHIBATA *et al.*, 2002). A compreensão da sinalização *MK/PTN* através de seus respectivos receptores é dificultada pelo fato de que *MK/PTN* interagem com numerosas proteínas de superfície celular, e seus supostos receptores também interagem com uma variedade de outros ligantes. Assim, elucidar a contribuição funcional da sinalização *MK/PTN* requer o conhecimento da expressão do receptor potencial (SORRELLE; DOMINGUEZ; BREKKEN, 2017). Além disso, a *MK* impulsiona o crescimento de tumores (HAO *et al.*, 2013).

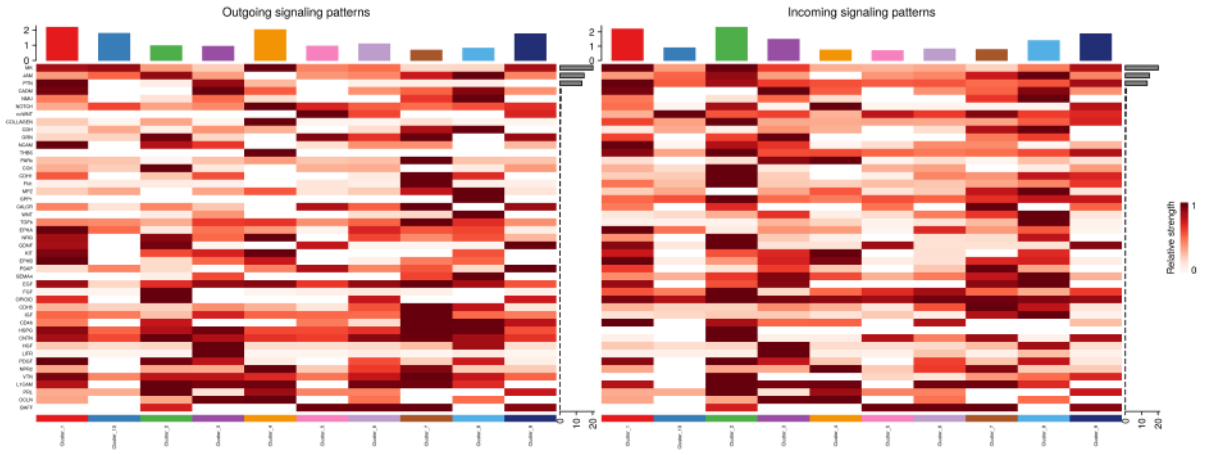

Figura 80 – *Heatmap* onde é representado pela barra de cores a força relativa de sinalização de uma via de sinalização. O gráfico de barra colorido na parte superior mostra a força total de sinalização, resumindo todas as vias de sinalização exibidas no *heatmap*. O gráfico da barra cinza na parte direita mostra a força total de sinalização de uma via de sinalização. Na esquerda o *heatmap* representando as vias ou pares de ligantes-receptores que saem, e na direita o *heatmap* representando as vias ou pares de ligantes-receptores de entrada. Esse *heatmap* corresponde às células dos pacientes 05 e 34.

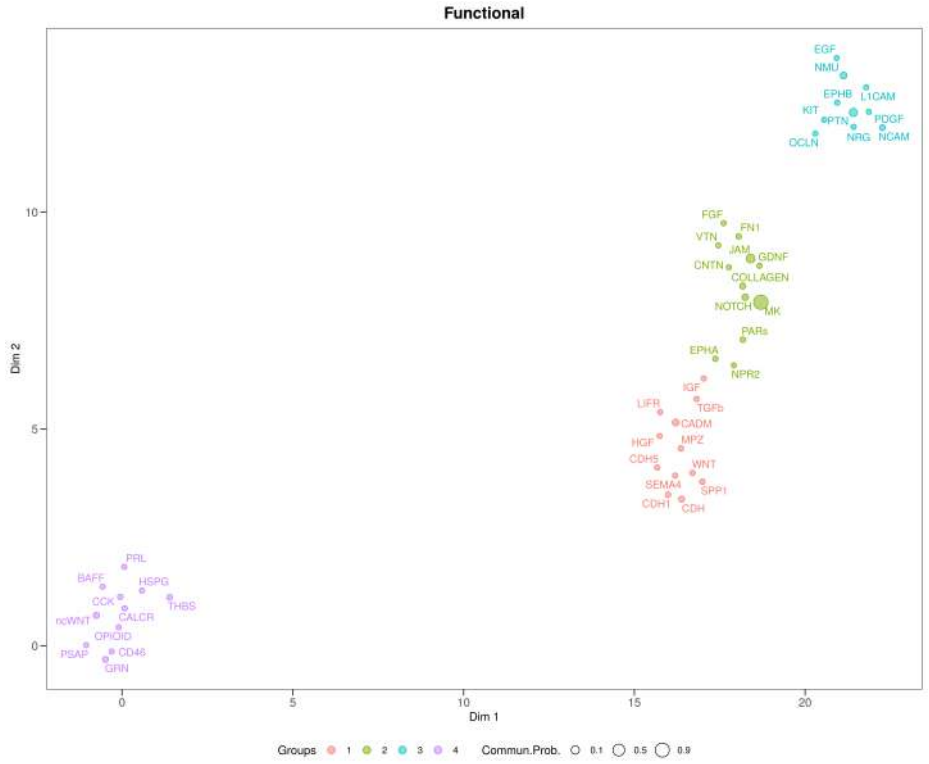

Figura 81 – Similaridade funcional para as vias significantes dos pacientes 05 e 34. Foram detectados 4 principais grupos de similaridade sendo o tamanho dos pontos a probabilidade de comunicação.

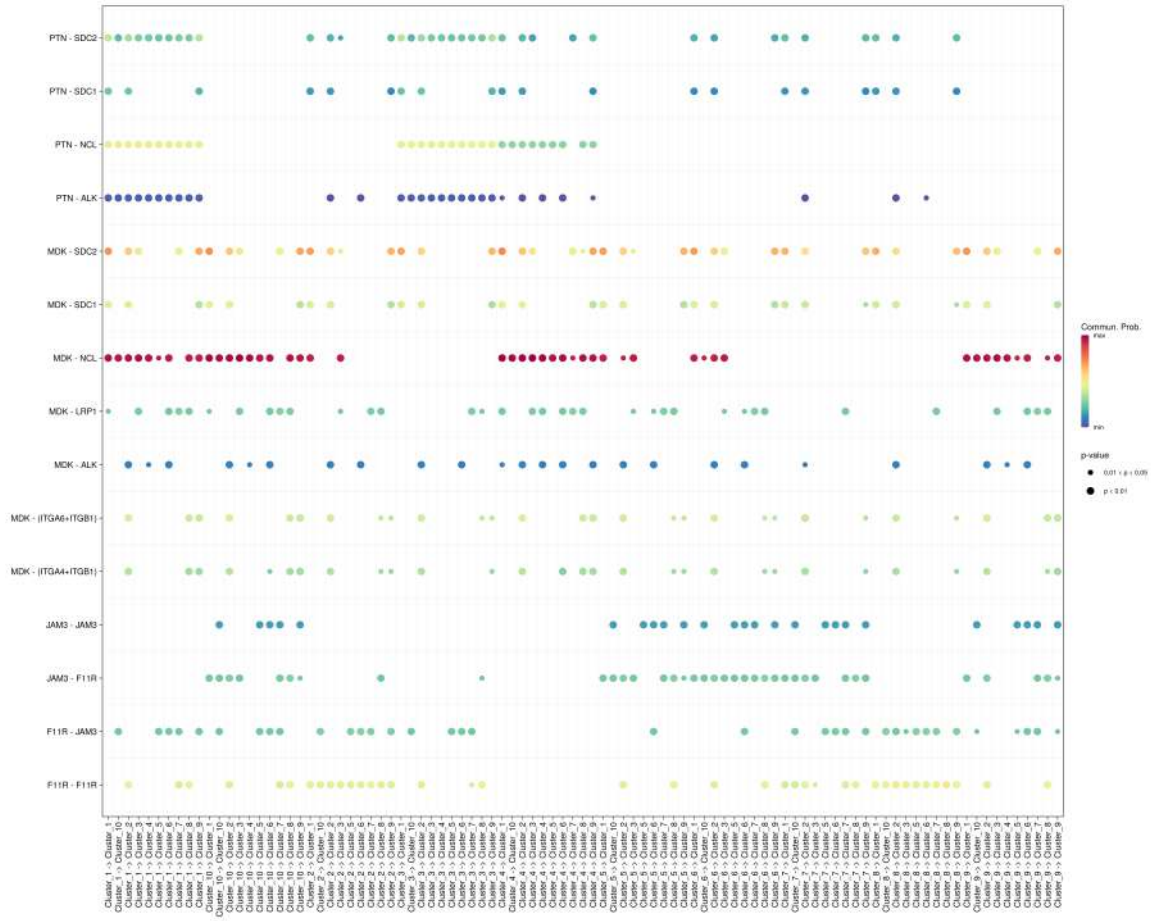

Figura 82 – Gráfico bolha mostrando as interações significantes (pares ligantes-receptores) para as vias *MK*, *PTN* e *JAM*. A cor dos pontos e o tamanho representam a probabilidade de comunicação e p-valor, respectivamente.

O segundo caso é correspondente aos pacientes 21 e 25 que são amostras de pacientes que estão em processo de tratamento com inibidores *PARP*. Inicialmente essas duas amostras foram integradas e feita a clusterização obtendo-se 9 *clusters*. Considerando os 9 *clusters* foi feita a análise de comunicação célula-célula. As vias de sinalização mais significantes foram as *MK*, *CD99* e *NMU*, como pode ser observado na Figura 83. Essas vias de sinalização mais significantes, quando visualizadas as similaridades funcionais (Figura 84), descobrimos que *CD99* e *NMU* fazem parte do mesmo grupo (Grupo 1), enquanto *MK* faz parte de um grupo separado (Grupo 2). Isso mostra que *CD99* e *NMU* possuem papéis semelhantes e/ou redundantes nessas amostras. As interações entre os pares receptores-ligantes para as vias *MK*, *CD99* e *NMU* podem ser observadas na Figura 85. Mais uma vez há um alto valor de probabilidade de comunicação entre o par *MDK-NCL*, levando à mesma discussão levantada para o caso 1. Adicionalmente, observamos altos valores de probabilidade de comunicação para as vias de sinalização

CD99 e *NMU*. Ambas são descritas na literatura por serem utilizadas como potenciais biomarcadores para tratamento de câncer de pulmão de células não pequenas (PELOSI *et al.*, 2006; YOU; GAO, 2018), que podem vir a ser potenciais biomarcadores para CPPC após verificações experimentais.

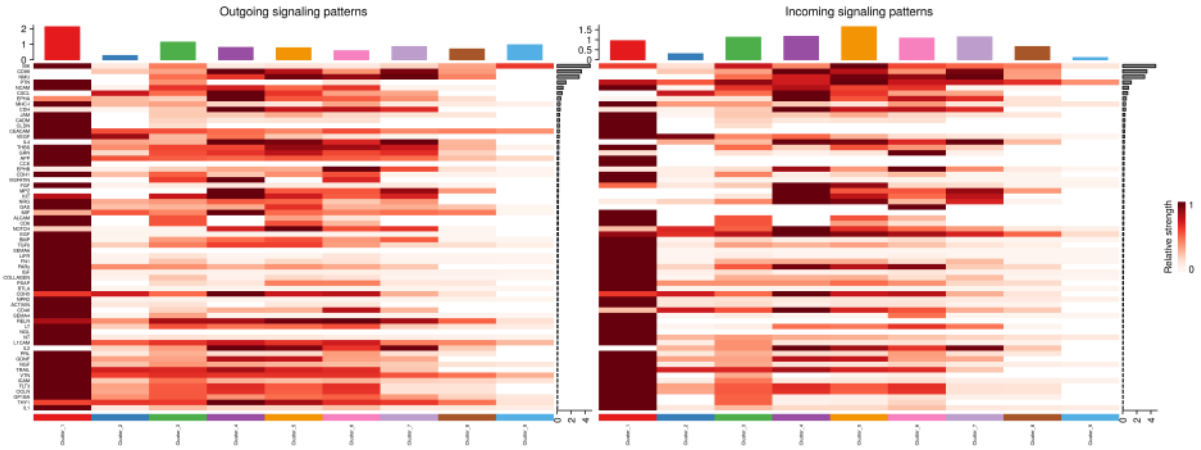

Figura 83 – *Heatmap* onde é representado pela barra de cores a força relativa de sinalização de uma via de sinalização. O gráfico de barra colorido na parte superior mostra a força total de sinalização, resumindo todas as vias de sinalização exibidas no *heatmap*. O gráfico da barra cinza na parte direita mostra a força total de sinalização de uma via de sinalização. Na esquerda o *heatmap* representando as vias ou pares de ligantes-receptores que saem, e na direita o *heatmap* representando as vias ou pares de ligantes-receptores de entrada. Esse *heatmap* corresponde às células dos pacientes 21 e 25.

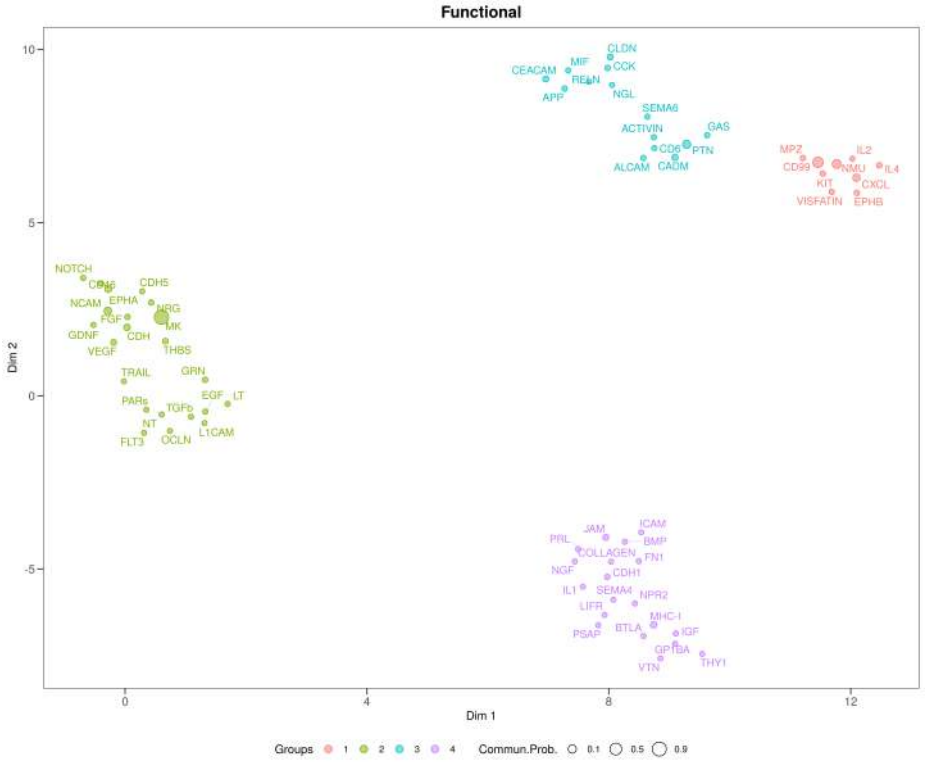

Figura 84 – Similaridade funcional para as vias significantes dos pacientes 21 e 25. Foram detectados 4 principais grupos de similaridade sendo o tamanho dos pontos a probabilidade de comunicação.

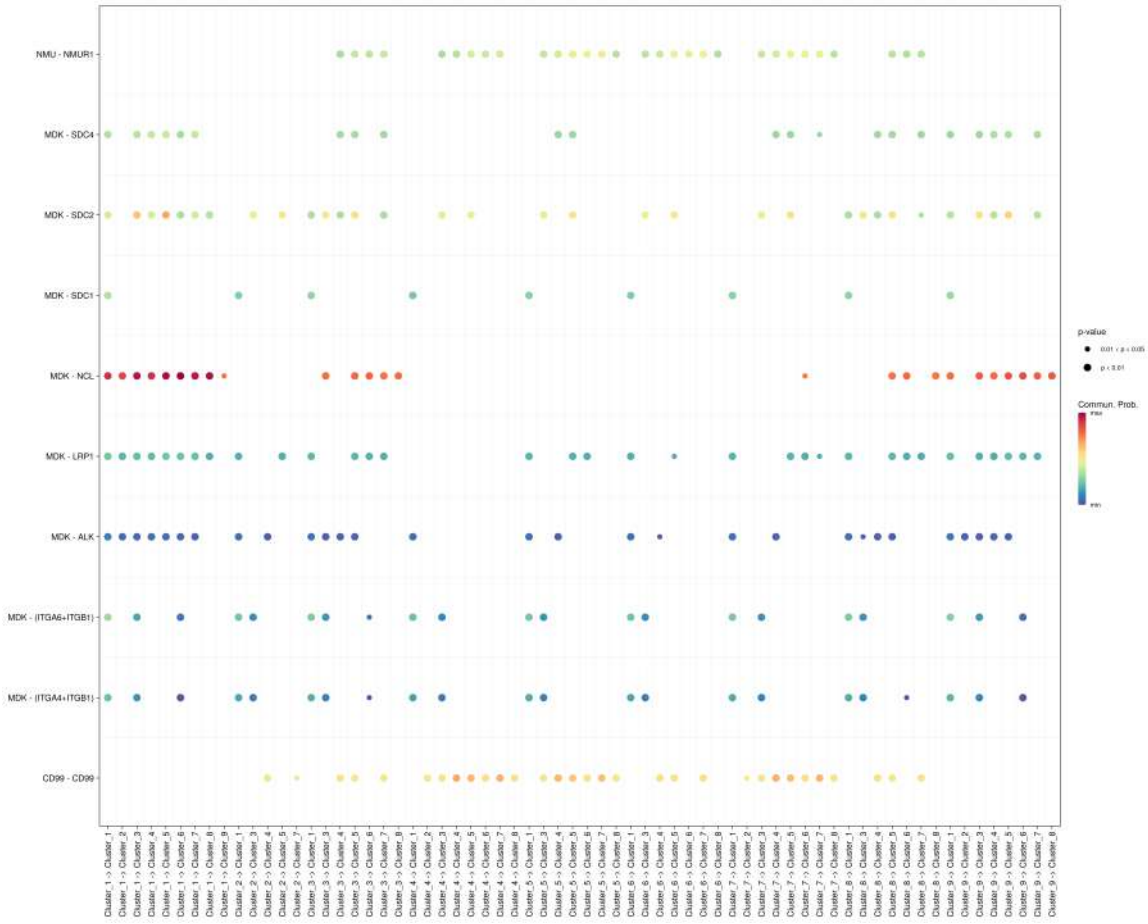

Figura 85 – Gráfico bolha mostrando as interações significantes (pares ligantes-receptores) para as vias *MK*, *CD99* e *NMU*. A cor dos pontos e o tamanho representam a probabilidade de comunicação e p-valor, respectivamente.

Por fim, o terceiro caso é correspondente aos pacientes 18 e 35 que são amostras de pacientes que estão em recidiva. Inicialmente essas duas amostras foram integradas e feita a clusterização obtendo-se 5 *clusters*. Considerando os 5 *clusters* foi feita a análise de comunicação célula-célula. As vias de sinalização mais significantes foram as *CADM*, *CD99* e *NCAM*, como pode ser observado na Figura 86. Essas vias de sinalização mais significantes, quando visualizadas as similaridades funcionais (Figura 87), descobrimos que *CADM* e *NCAM* fazem parte do mesmo grupo (Grupo 3), enquanto *CD99* faz parte de um grupo separado (Grupo 2). Isso mostra que *CADM* e *NCAM* possuem papéis semelhantes e/ou redundantes nessas amostras. As interações entre os pares receptores-ligantes para as vias *CADM*, *CD99* e *NCAM* podem ser observadas na Figura 88. As vias de sinalização *CADM* e *NCAM*, cujos principais genes (receptores-ligantes) *CADM1* e *NCAM1*, respectivamente, estão associados em células do CPPC. Ambos estão associados a características neuroendócrinas e são considerados marcadores neuroendócrinos de

superfície celular (RUDIN *et al.*, 2019), fornecendo um promissor marcador de diagnóstico de CPPC (FUNAKI *et al.*, 2021).

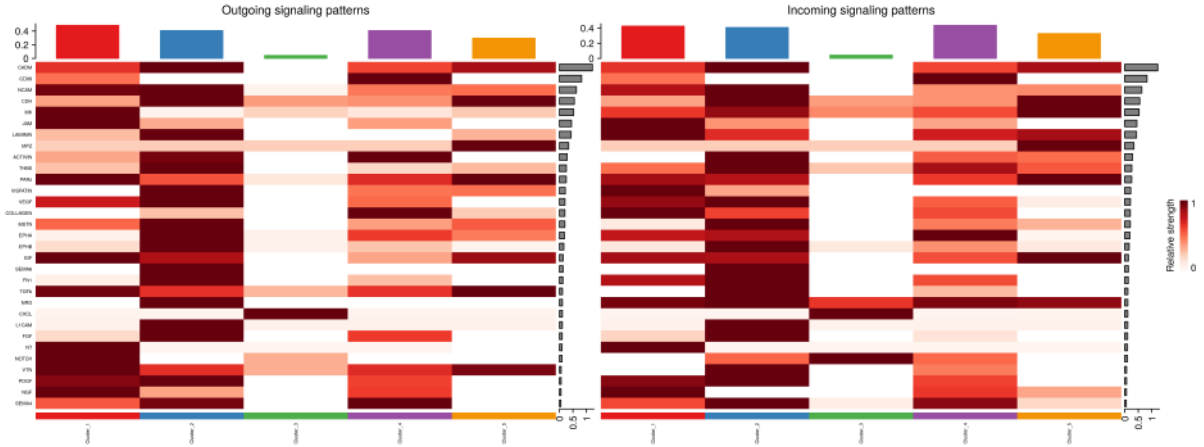

Figura 86 – *Heatmap* onde é representado pela barra de cores a força relativa de sinalização de uma via de sinalização. O gráfico de barra colorido na parte superior mostra a força total de sinalização, resumindo todas as vias de sinalização exibidas no *heatmap*. O gráfico da barra cinza na parte direita mostra a força total de sinalização de uma via de sinalização. Na esquerda o *heatmap* representando as vias ou pares de ligantes-receptores que saem, e na direita o *heatmap* representando as vias ou pares de ligantes-receptores de entrada. Esse *heatmap* corresponde às células dos pacientes 18 e 35.

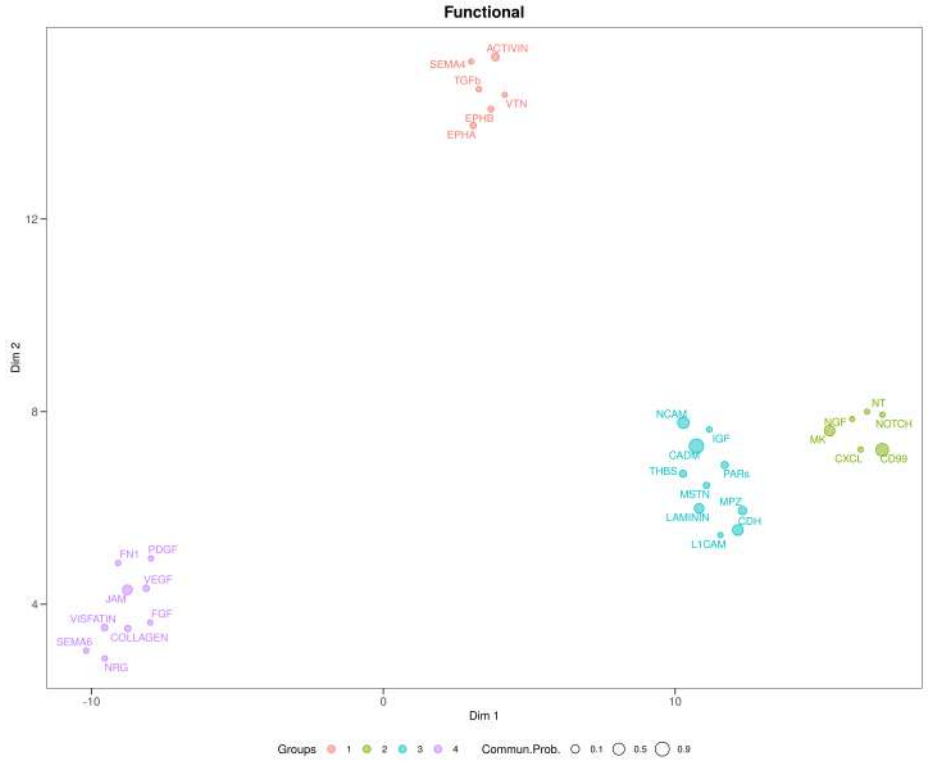

Figura 87 – Similaridade funcional para as vias significantes dos pacientes 18 e 35. Foram detectados 4 principais grupos de similaridade sendo o tamanho dos pontos a probabilidade de comunicação.

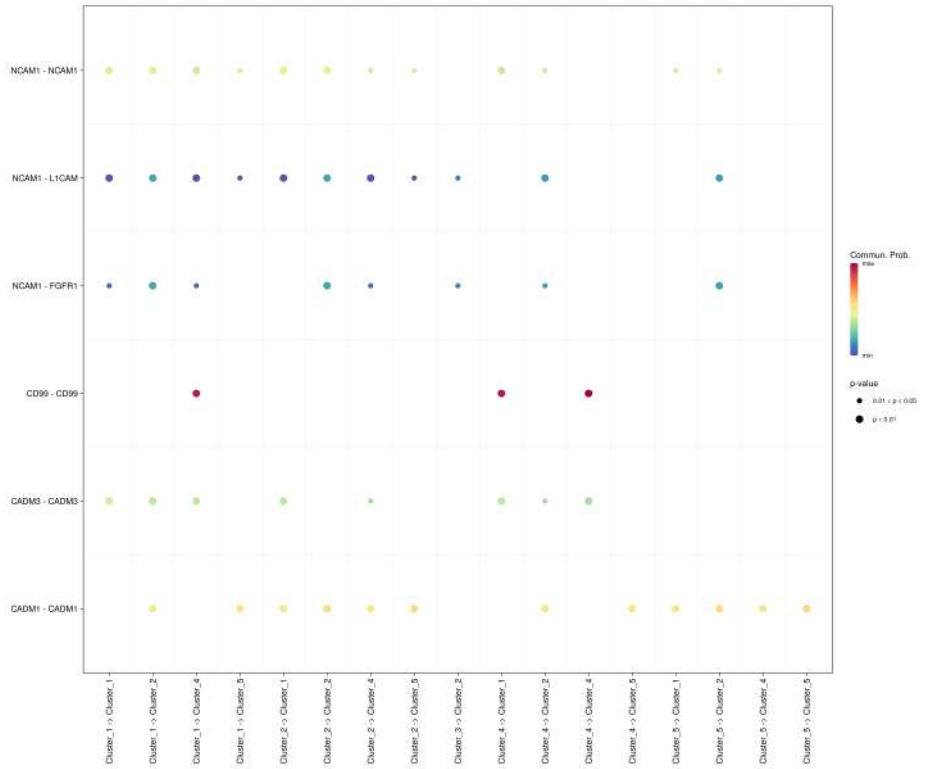

Figura 88 – Gráfico bolha mostrando as interações significantes (pares ligantes-receptores) para as vias *CADM*, *CD99* e *NCAM*. A cor dos pontos e o tamanho representam a probabilidade de comunicação e p-valor, respectivamente.

5 Conclusão

Este trabalho mostrou uma gama de possibilidades de análises para dados de sequenciamento de *RNA* de células únicas, tanto para Melanoma e para Câncer de Pulmão de Pequenas Células (CPPC). A avaliação integrada da evolução genômica e plasticidade das células tumorais contribuíram para o melhor entendimento desses tumores a nível de células únicas. Abaixo, em negrito, estão os objetivos específicos e quais as conclusões obtidas neste trabalho para cada um deles:

5.1 Para dados de Melanoma

Identificar populações de células enriquecidas com as assinaturas de *EMT*:

- Há subpopulação de células tronco tumorais nos dados de melanoma.

A clusterização dos dados possibilitou a identificação da subpopulação de células tronco tumorais.

- Identificação de um braço da trajetória que possui evidências de ser um ponto de diferenciação onde há a ativação da via de *EMT* em melanoma.

Há a presença de uma subpopulação de células tronco tumorais nas células de melanoma neste braço (Braço 6). Além disso a presença de células endoteliais e *CAF* demonstram que a ativação da via *EMT* em células únicas de melanoma estão condicionadas à presença dessas células estromais.

- Geração de assinaturas para tipos celulares específicos em melanoma.

Estas assinaturas irão facilitar a identificação de tipos celulares específicos, além de identificar possíveis subpopulações de células troncos.

Identificação de circuitos gênicos envolvidos na ativação da via de *EMT*:

- *TRHDE-AS1* possui um perfil de *EMT* em potencial, agindo juntamente com o *HOTAIR* no circuito de *feedback* negativo.

Foram identificados padrões interessantes de expressão do *HOTAIR* e genes associados em células de melanoma, e nos permitiu identificar o *lncRNA TRHDE-AS1* que possui um perfil de *EMT* em potencial. Foi proposto um circuito de *feedback* negativo onde o *HOTAIR* regula positivamente o *TRHDE-AS1*, e o *TRHDE-AS1* regula negativamente o *HOTAIR*. Foi possível confirmar este circuito a partir de uma validação funcional a aprtir da metodologia de silenciamento gênico. Estudos mostram que o *TRHDE-AS1* é um biomarcador para o tratamento de diferentes tipos tumorais, mas nada foi relatado para melanoma, trazendo assim a ideia que o *TRHDE-AS1*, juntamente com o *HOTAIR* e a via EMT, possam ajudar na identificação de novos biomarcadores no tratamento do melanoma.

- Os dados de melanoma obedecem o perfil proposto por Alves *et al.* (2013).

As ferramentas possibilitaram análises para a identificação do perfil proposto por Alves *et al.* (2013) em dados de células únicas de melanoma, onde o *HOTAIR* ativado promove um aumento na expressão de uma série de genes enquanto que a diminuição de outros genes, como a *CDH1*.

5.2 Para dados de Câncer de Pulmão de Pequenas Células (CPPC)

Avaliar o perfil transcricional das células únicas:

- Avaliação transcricional e conformidade dos subtipos moleculares para CPPC.

A partir dos dados de CPPC foi possível avaliar o perfil transcricional, além de verificar que os subtipos moleculares estão de acordo com a classificação proposta por Rudin *et al.* (2019). Há expressão de *ASCL1*, *NEUROD1* e *POU2F3*, sendo que não há amostras do subgrupo *YAP1*, pois são muito raras.

Verificar subtipos moleculares e marcadores neuroendócrinos:

- A classificação atual de subtipos proposta por Rudin *et al.* (2019) é obedecida para os dados deste estudo.

Avaliou-se o perfil transcricional das células únicas e concluiu-se que os dados estão de acordo com a classificação atual de subtipos moleculares, expressando *NEUROD1*, *POU2F3* e *ASCL1*. Não foram identificadas amostras *YAP1* neste estudo (expresso em apenas 2,4%

dos casos). Além disso foi constatada a expressão de mais de 1 *TF* em cada caso, trazendo uma diferente visão de que os subtipos moleculares são compostos por apenas 1 *TF*.

Verificar a heterogeneidade intratumoral:

- Maior heterogeneidade intratumoral em amostras clínicas de CPPC.

Foi observado maior heterogeneidade intratumoral em amostras clínicas de pacientes devido à presença de células não tumorais em relação às amostras *CDX/PDX*.

Identificar mecanismos e vias envolvidas na progressão tumoral:

- A via de sinalização *NOTCH* contribui para a plasticidade transcricional em CPPC.

A heterogeneidade intratumoral gerada pela via de sinalização *Notch* promove o câncer de pulmão de pequenas células. Dessa forma, foram identificados mecanismos e vias envolvidas na plasticidade transcricional para pacientes em diferentes condições. Ficou evidente o papel importante dos genes *NOTCH1* e *REST* no desenvolvimento do CPPC nas diferentes condições.

- As vias de sinalização *MK*, *JAM*, *PTN*, *CD99*, *NMU*, *CADM* e *NCAM* são importantes para o desenvolvimento de terapias e biomarcadores em CPPC.

A identificação dos ligantes e receptores que se comunicam entre as células trazem um novo caminho para as terapias de CPPC. A partir das análises desenvolvidas nessse trabalho foi possível identificar ligantes e receptores importantes para o desenvolvimento do câncer, e principalmente, para a elucidação de possíveis novos biomarcadores.

Considerações gerais

Os dados gerados pelo presente trabalho fornecem novas informações e *insights* sobre a progressão tumoral do Melanoma e do CPPC. Tais informações podem ser úteis para o avanço no conhecimento destes tumores e no desenvolvimento de novos biomarcadores para terapias. Além disso o trabalho também descreve um conjunto de ferramentas que podem ser utilizadas para explorar dados de células únicas de diferentes tipos tumorais.

Referências

AHN, R.; GUPTA, R.; LAI, K.; CHOPRA, N.; ARRON, S. T.; LIAO, W. Network analysis of psoriasis reveals biological pathways and roles for coding and long non-coding rnas. *BMC genomics*, Springer, v. 17, n. 1, p. 1–10, 2016. Citado na página 91.

AIBAR, S.; GONZÁLEZ-BLAS, C. B.; MOERMAN, T.; HUYNH-THU, V. A.; IMRICHOVA, H.; HULSELMANS, G.; RAMBOW, F.; MARINE, J.-C.; GEURTS, P.; AERTS, J. *et al.* Scenic: single-cell regulatory network inference and clustering. *Nature methods*, Nature Publishing Group, v. 14, n. 11, p. 1083–1086, 2017. Citado 2 vezes nas páginas 54 e 57.

ALONSO, S. R.; TRACEY, L.; ORTIZ, P.; PÉREZ-GÓMEZ, B.; PALACIOS, J.; POLLÁN, M.; LINARES, J.; SERRANO, S.; SÁEZ-CASTILLO, A. I.; SÁNCHEZ, L. *et al.* A high-throughput study in melanoma identifies epithelial-mesenchymal transition as a major determinant of metastasis. *Cancer research*, AACR, v. 67, n. 7, p. 3450–3460, 2007. Citado na página 92.

ALVES, C. P.; FONSECA, A. S.; MUYS, B. R.; BUENO, R. de Barros e L.; BUERGER, M. C.; SOUZA, J. E. de; VALENTE, V.; ZAGO, M. A.; JR, W. A. S. Brief report: the lincrna hotair is required for epithelial-to-mesenchymal transition and stemness maintenance of cancer cell lines. *Stem cells*, Wiley Online Library, v. 31, n. 12, p. 2827–2832, 2013. Citado 7 vezes nas páginas 50, 51, 60, 82, 92, 110 e 146.

ANASTASSIOU, D.; RUMJANTSEVA, V.; CHENG, W.; HUANG, J.; CANOLL, P. D.; YAMASHIRO, D. J.; KANDEL, J. J. Human cancer cells express slug-based epithelial-mesenchymal transition gene expression signature obtained in vivo. *BMC cancer*, BioMed Central, v. 11, n. 1, p. 1–9, 2011. Citado na página 92.

ANDERSON, N. M.; SIMON, M. C. The tumor microenvironment. *Current Biology*, Elsevier, v. 30, n. 16, p. R921–R925, 2020. Citado na página 45.

BALAS, M. M.; JOHNSON, A. M. Exploring the mechanisms behind long noncoding rnas and cancer. *Non-coding RNA research*, Elsevier, v. 3, n. 3, p. 108–117, 2018. Citado na página 49.

BALLOTTI, R.; CHELI, Y.; BERTOLOTTO, C. The complex relationship between mitf and the immune system: a melanoma immunotherapy (response) factor? *Molecular cancer*, BioMed Central, v. 19, n. 1, p. 1–12, 2020. Citado na página 100.

BEN-PORATH, I.; THOMSON, M. W.; CAREY, V. J.; GE, R.; BELL, G. W.; REGEV, A.; WEINBERG, R. A. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. *Nature genetics*, Nature Publishing Group, v. 40, n. 5, p. 499–507, 2008. Citado na página 92.

BERGEN, V.; LANGE, M.; PEIDLI, S.; WOLF, F. A.; THEIS, F. J. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, Nature Publishing Group, v. 38, n. 12, p. 1408–1414, 2020. Citado 2 vezes nas páginas 63 e 71.

BERGMANN, J. H.; SPECTOR, D. L. Long non-coding rnas: modulators of nuclear structure and function. *Current opinion in cell biology*, Elsevier, v. 26, p. 10–18, 2014. Citado na página 50.

BERNSTEIN, N. J.; FONG, N. L.; LAM, I.; ROY, M. A.; HENDRICKSON, D. G.; KELLEY, D. R. Solo: doublet identification in single-cell rna-seq via semi-supervised deep learning. *Cell Systems*, Elsevier, v. 11, n. 1, p. 95–101, 2020. Citado na página 33.

BIAGI, C. A. Oliveira de; NOCITI, R. P.; BROTTO, D. B.; FUNICHELI, B. O.; RUY, P. d. C.; XIMENEZ, J. P. B.; FIGUEIREDO, D. L. A.; SILVA, W. A. *et al.* Cetf: an r/bioconductor package for transcription factor co-expression networks using regulatory impact factors (rif) and partial correlation and information (pcit) analysis. *BMC genomics*, BioMed Central, v. 22, n. 1, p. 1–8, 2021. Citado 3 vezes nas páginas 54, 60 e 97.

BIERIE, B.; MOSES, H. L. Tgfβ: the molecular jekyll and hyde of cancer. *Nature Reviews Cancer*, Nature Publishing Group, v. 6, n. 7, p. 506–520, 2006. Citado na página 40.

BOLÓS, V.; PEINADO, H.; PÉREZ-MORENO, M. A.; FRAGA, M. F.; ESTELLER, M.; CANO, A. The transcription factor slug represses e-cadherin expression and induces epithelial to mesenchymal transitions: a comparison with snail and e47 repressors. *Journal of cell science*, Company of Biologists, v. 116, n. 3, p. 499–511, 2003. Citado na página 39.

BRANDNER, J. M.; HAASS, N. K. Melanoma's connections to the tumour microenvironment. *Pathology*, Elsevier, v. 45, n. 5, p. 443–452, 2013. Citado na página 39.

BUTLER, A.; HOFFMAN, P.; SMIBERT, P.; PAPALEXI, E.; SATIJA, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, Nature Publishing Group, v. 36, n. 5, p. 411–420, 2018. Citado 2 vezes nas páginas 54 e 56.

CABILI, M. N.; DUNAGIN, M. C.; MCCLANAHAN, P. D.; BIAESCH, A.; PADOVAN-MERHAR, O.; REGEV, A.; RINN, J. L.; RAJ, A. Localization and abundance analysis of human lncrnas at single-cell and single-molecule resolution. *Genome biology*, Springer, v. 16, n. 1, p. 1–16, 2015. Citado na página 48.

CABILI, M. N.; TRAPNELL, C.; GOFF, L.; KOZIOL, M.; TAZON-VEGA, B.; REGEV, A.; RINN, J. L. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes & development*, Cold Spring Harbor Lab, v. 25, n. 18, p. 1915–1927, 2011. Citado 2 vezes nas páginas 47 e 48.

CALIN, G. A.; LIU, C.-g.; FERRACIN, M.; HYSLOP, T.; SPIZZO, R.; SEVIGNANI, C.; FABBRI, M.; CIMMINO, A.; LEE, E. J.; WOJCIK, S. E. *et al.* Ultraconserved regions encoding ncrnas are altered in human leukemias and carcinomas. *Cancer cell*, Elsevier, v. 12, n. 3, p. 215–229, 2007. Citado na página 47.

CALLE, A. S.; KAWAMURA, Y.; YAMAMOTO, Y.; TAKESHITA, F.; OCHIYA, T. Emerging roles of long non-coding rna in cancer. *Cancer science*, Wiley Online Library, v. 109, n. 7, p. 2093–2100, 2018. Citado na página 47.

CAO, J.; PACKER, J. S.; RAMANI, V.; CUSANOVICH, D. A.; HUYNH, C.; DAZA, R.; QIU, X.; LEE, C.; FURLAN, S. N.; STEEMERS, F. J. *et al.* Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, American Association for the Advancement of Science, v. 357, n. 6352, p. 661–667, 2017. Citado na página 36.

CAO, J.; SPIELMANN, M.; QIU, X.; HUANG, X.; IBRAHIM, D. M.; HILL, A. J.; ZHANG, F.; MUNDLOS, S.; CHRISTIANSEN, L.; STEEMERS, F. J. *et al.* The

single-cell transcriptional landscape of mammalian organogenesis. *Nature*, Nature Publishing Group, v. 566, n. 7745, p. 496–502, 2019. Citado na página 55.

CARLIN, D. E.; DEMCHAK, B.; PRATT, D.; SAGE, E.; IDEKER, T. Network propagation in the cytoscape cyberinfrastructure. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 13, n. 10, p. e1005598, 2017. Citado na página 60.

CHAO, Y. L.; SHEPARD, C. R.; WELLS, A. Breast carcinoma cells re-express e-cadherin during mesenchymal to epithelial reverting transition. *Molecular cancer*, BioMed Central, v. 9, n. 1, p. 1–18, 2010. Citado na página 52.

CHEN, H.; ALBERGANTE, L.; HSU, J. Y.; LAREAU, C. A.; BOSCO, G. L.; GUAN, J.; ZHOU, S.; GORBAN, A. N.; BAUER, D. E.; ARYEE, M. J. *et al.* Single-cell trajectories reconstruction, exploration and mapping of omics data with stream. *Nature communications*, Nature Publishing Group, v. 10, n. 1, p. 1–14, 2019. Citado 2 vezes nas páginas 54 e 58.

CHEN, J. C.; PEREZ-LORENZO, R.; SAENGER, Y. M.; DRAKE, C. G.; CHRISTIANO, A. M. Ikzf1 enhances immune infiltrate recruitment in solid tumors and susceptibility to immunotherapy. *Cell systems*, Elsevier, v. 7, n. 1, p. 92–103, 2018. Citado na página 99.

CONACCI-SORRELL, M.; SIMCHA, I.; BEN-YEDIDIA, T.; BLECHMAN, J.; SAVAGNER, P.; BEN-ZE'EV, A. Autoregulation of e-cadherin expression by cadherin–cadherin interactions: the roles of β-catenin signaling, slug, and mapk. *The Journal of cell biology*, Rockefeller University Press, v. 163, n. 4, p. 847–857, 2003. Citado na página 39.

DAHL, C.; SCHALL, R.; HE, H.; CAIRNS, J. Identification of a novel gene expressed in activated natural killer cells and t cells. *The Journal of Immunology*, Am Assoc Immnol, v. 148, n. 2, p. 597–603, 1992. Citado na página 74.

DERRIEN, T.; JOHNSON, R.; BUSSOTTI, G.; TANZER, A.; DJEBALI, S.; TILGNER, H.; GUERNEC, G.; MARTIN, D.; MERKEL, A.; KNOWLES, D. G. *et al.* The gencode v7 catalog of human long noncoding rnas: analysis of their gene structure, evolution, and expression. *Genome research*, Cold Spring Harbor Lab, v. 22, n. 9, p. 1775–1789, 2012. Citado 2 vezes nas páginas 47 e 48.

DIJK, D. V.; SHARMA, R.; NAINYS, J.; YIM, K.; KATHAIL, P.; CARR, A. J.; BURDZIAK, C.; MOON, K. R.; CHAFFER, C. L.; PATTABIRAMAN, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell*, Elsevier, v. 174, n. 3, p. 716–729, 2018. Citado 2 vezes nas páginas 54 e 56.

DING, J.; CONDON, A.; SHAH, S. P. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications*, Nature Publishing Group, v. 9, n. 1, p. 1–13, 2018. Citado na página 70.

DOBIN, A.; DAVIS, C. A.; SCHLESINGER, F.; DRENKOW, J.; ZALESKI, C.; JHA, S.; BATUT, P.; CHAISSON, M.; GINGERAS, T. R. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, Oxford University Press, v. 29, n. 1, p. 15–21, 2013. Citado na página 66.

DONG, S.; WANG, R.; WANG, H.; DING, Q.; ZHOU, X.; WANG, J.; ZHANG, K.; LONG, Y.; LU, S.; HONG, T. *et al.* Hoxd-as1 promotes the epithelial to mesenchymal transition of ovarian cancer cells by regulating mir-186-5p and pik3r3. *Journal of Experimental & Clinical Cancer Research*, BioMed Central, v. 38, n. 1, p. 1–13, 2019. Citado na página 91.

DUECK, H.; EBERWINE, J.; KIM, J. Variation is function: are single cell differences functionally important? testing the hypothesis that single cell variation is required for aggregate function. *Bioessays*, Wiley Online Library, v. 38, n. 2, p. 172–180, 2016. Citado na página 47.

EYMIN, B.; GAZZERI, S. Role of cell cycle regulators in lung carcinogenesis. *Cell adhesion & migration*, Taylor & Francis, v. 4, n. 1, p. 114–123, 2010. Citado na página 113.

FARNHAM, P. J. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, Nature Publishing Group, v. 10, n. 9, p. 605–616, 2009. Citado na página 59.

FERLAY, J.; SOERJOMATARAM, I.; DIKSHIT, R.; ESER, S.; MATHERS, C.; REBELO, M.; PARKIN, D. M.; FORMAN, D.; BRAY, F. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, Wiley Online Library, v. 136, n. 5, p. E359–E386, 2015. Citado na página 38.

FILBIN, M. G.; TIROSH, I.; HOVESTADT, V.; SHAW, M. L.; ESCALANTE, L. E.; MATHEWSON, N. D.; NEFTEL, C.; FRANK, N.; PELTON, K.; HEBERT, C. M. *et al.* Developmental and oncogenic programs in h3k27m gliomas dissected by single-cell rna-seq. *Science*, American Association for the Advancement of Science, v. 360, n. 6386, p. 331–335, 2018. Citado na página 30.

FIORI, M. E.; FRANCO, S. D.; VILLANOVA, L.; BIANCA, P.; STASSI, G.; MARIA, R. D. Cancer-associated fibroblasts as abettors of tumor progression at the crossroads of emt and therapy resistance. *Molecular Cancer*, BioMed Central, v. 18, n. 1, p. 1–16, 2019. Citado na página 108.

FREEDMAN, M. L.; MONTEIRO, A. N.; GAYTHER, S. A.; COETZEE, G. A.; RISCH, A.; PLASS, C.; CASEY, G.; BIASI, M. D.; CARLSON, C.; DUGGAN, D. *et al.* Principles for the post-gwas functional characterization of cancer risk loci. *Nature genetics*, Nature Publishing Group, v. 43, n. 6, p. 513–518, 2011. Citado na página 47.

FU, A.; HOU, Y.; YU, Z.; ZHAO, Z.; LIU, Z. Healthy mitochondria inhibit the metastatic melanoma in lungs. *International journal of biological sciences*, Ivyspring International Publisher, v. 15, n. 12, p. 2707, 2019. Citado na página 101.

FUNAKI, T.; ITO, T.; TANEI, Z.-i.; GOTO, A.; NIKI, T.; MATSUBARA, D.; MURAKAMI, Y. Cadm1 promotes malignant features of small-cell lung cancer by recruiting 4.1 r to the plasma membrane. *Biochemical and Biophysical Research Communications*, Elsevier, v. 534, p. 172–178, 2021. Citado na página 142.

GASCARD, P.; TLSTY, T. D. Carcinoma-associated fibroblasts: orchestrating the composition of malignancy. *Genes & development*, Cold Spring Harbor Lab, v. 30, n. 9, p. 1002–1019, 2016. Citado na página 108.

GAWRONSKI, K. A.; KIM, J. Single cell transcriptomics of noncoding rnas and their cell-specificity. *Wiley Interdisciplinary Reviews: RNA*, Wiley Online Library, v. 8, n. 6, p. e1433, 2017. Citado na página 46.

GEORGE, J.; LIM, J. S.; JANG, S. J.; CUN, Y.; OZRETIĆ, L.; KONG, G.; LEENDERS, F.; LU, X.; FERNÁNDEZ-CUESTA, L.; BOSCO, G. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature*, Nature Publishing Group, v. 524, n. 7563, p. 47–53, 2015. Citado na página 42.

GIBB, E. A.; BROWN, C. J.; LAM, W. L. The functional role of long non-coding rna in human carcinomas. *Molecular cancer*, BioMed Central, v. 10, n. 1, p. 1–17, 2011. Citado na página 88.

GIERAHN, T. M.; WADSWORTH, M. H.; HUGHES, T. K.; BRYSON, B. D.; BUTLER, A.; SATIJA, R.; FORTUNE, S.; LOVE, J. C.; SHALEK, A. K. Seq-well: portable, low-cost rna sequencing of single cells at high throughput. *Nature methods*, Nature Publishing Group, v. 14, n. 4, p. 395–398, 2017. Citado na página 36.

GIFFIN, M. J.; COOKE, K.; LOBENHOFER, E. K.; ESTRADA, J.; ZHAN, J.; DEEGEN, P.; THOMAS, M.; MURAWSKY, C. M.; WERNER, J.; LIU, S. *et al.* Amg 757, a half-life extended, dll3-targeted bispecific t-cell engager, shows high potency and sensitivity in preclinical models of small-cell lung cancer. *Clinical Cancer Research*, AACR, v. 27, n. 5, p. 1526–1537, 2021. Citado na página 42.

GOARD, C.; SCHIMMER, A. Mitochondrial matrix proteases as novel therapeutic targets in malignancy. *Oncogene*, Nature Publishing Group, v. 33, n. 21, p. 2690–2699, 2014. Citado na página 101.

GOTTARDI, C. J.; WONG, E.; GUMBINER, B. M. E-cadherin suppresses cellular transformation by inhibiting β-catenin signaling in an adhesion-independent manner. *The Journal of cell biology*, The Rockefeller University Press, v. 153, n. 5, p. 1049–1060, 2001. Citado na página 39.

GOVINDAN, R.; PAGE, N.; MORGENSZTERN, D.; READ, W.; TIERNEY, R.; VLAHIOTIS, A.; SPITZNAGEL, E. L.; PICCIRILLO, J. Changing epidemiology of small-cell lung cancer in the united states over the last 30 years: analysis of the surveillance, epidemiologic, and end results database. *Journal of clinical oncology*, American Society of Clinical Oncology, v. 24, n. 28, p. 4539–4544, 2006. Citado na página 41.

GRISS, J.; BAUER, W.; WAGNER, C.; SIMON, M.; CHEN, M.; GRABMEIER-PFISTERSHAMMER, K.; MAURER-GRANOFSZKY, M.; ROKA, F.; PENZ, T.; BOCK, C. *et al.* B cells sustain inflammation and predict response to immune checkpoint blockade in human melanoma. *Nature communications*, Nature Publishing Group, v. 10, n. 1, p. 1–14, 2019. Citado na página 103.

GRÜN, D.; OUDENAARDEN, A. van. Design and analysis of single-cell sequencing experiments. *Cell*, Elsevier, v. 163, n. 4, p. 799–810, 2015. Citado na página 32.

GUPTA, R.; AHN, R.; LAI, K.; MULLINS, E.; DEBBANEH, M.; DIMON, M.; ARRON, S.; LIAO, W. Landscape of long noncoding rnas in psoriatic and healthy skin. *Journal of Investigative Dermatology*, Elsevier, v. 136, n. 3, p. 603–609, 2016. Citado na página 91.

GUPTA, R. A.; SHAH, N.; WANG, K. C.; KIM, J.; HORLINGS, H. M.; WONG, D. J.; TSAI, M.-C.; HUNG, T.; ARGANI, P.; RINN, J. L. *et al.* Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, Nature Publishing Group, v. 464, n. 7291, p. 1071–1076, 2010. Citado 3 vezes nas páginas 49, 50 e 88.

GUSTAVSEN, J. A.; PAI, S.; ISSERLIN, R.; DEMCHAK, B.; PICO, A. R. Rcy3: Network biology using cytoscape from within r. *F1000Research*, Faculty of 1000 Ltd, v. 8, 2019. Citado na página 60.

GUTTMAN, M.; DONAGHEY, J.; CAREY, B. W.; GARBER, M.; GRENIER, J. K.; MUNSON, G.; YOUNG, G.; LUCAS, A. B.; ACH, R.; BRUHN, L. *et al.* lincrnas act in the circuitry controlling pluripotency and differentiation. *Nature*, Nature Publishing Group, v. 477, n. 7364, p. 295–300, 2011. Citado na página 49.

GUTTMAN, M.; GARBER, M.; LEVIN, J. Z.; DONAGHEY, J.; ROBINSON, J.; ADICONIS, X.; FAN, L.; KOZIOL, M. J.; GNIRKE, A.; NUSBAUM, C. *et al.* Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nature biotechnology*, Nature Publishing Group, v. 28, n. 5, p. 503–510, 2010. Citado 2 vezes nas páginas 47 e 48.

HAASS, N. K.; SMALLEY, K. S.; LI, L.; HERLYN, M. Adhesion, migration and communication in melanocytes and melanoma. *Pigment cell research*, Wiley Online Library, v. 18, n. 3, p. 150–159, 2005. Citado 2 vezes nas páginas 38 e 39.

HAGHVERDI, L.; LUN, A. T.; MORGAN, M. D.; MARIONI, J. C. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, Nature Publishing Group, v. 36, n. 5, p. 421–427, 2018. Citado na página 69.

HANAHAN, D.; WEINBERG, R. A. Hallmarks of cancer: the next generation. *cell*, Elsevier, v. 144, n. 5, p. 646–674, 2011. Citado na página 59.

HÄNZELMANN, S.; CASTELO, R.; GUINNEY, J. Gsva: gene set variation analysis for microarray and rna-seq data. *BMC bioinformatics*, Springer, v. 14, n. 1, p. 1–15, 2013. Citado 2 vezes nas páginas 54 e 57.

HAO, H.; MAEDA, Y.; FUKAZAWA, T.; YAMATSUJI, T.; TAKAOKA, M.; BAO, X.-H.; MATSUOKA, J.; OKUI, T.; SHIMO, T.; TAKIGAWA, N. *et al.* Inhibition of the growth factor mdk/midkine by a novel small molecule compound to treat non-small cell lung cancer. *PloS one*, Public Library of Science San Francisco, USA, v. 8, n. 8, p. e71093, 2013. Citado na página 136.

HAO, Y.; HAO, S.; ANDERSEN-NISSEN, E.; III, W. M. M.; ZHENG, S.; BUTLER, A.; LEE, M. J.; WILK, A. J.; DARBY, C.; ZAGER, M. *et al.* Integrated analysis of multimodal single-cell data. *Cell*, Elsevier, 2021. Citado 2 vezes nas páginas 63 e 68.

HAQUE, A.; ENGEL, J.; TEICHMANN, S. A.; LÖNNBERG, T. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, BioMed Central, v. 9, n. 1, p. 1–12, 2017. Citado na página 29.

HARTMAN, M. L.; CZYZ, M. Mitf in melanoma: mechanisms behind its expression and activity. *Cellular and Molecular Life Sciences*, Springer, v. 72, n. 7, p. 1249–1260, 2015. Citado na página 100.

HELLMAN, V. T. D. J. S.; ROSENBERG, S. A. *Cancer Principles and Practice of oncology.: 2001*. [S.l.]: Lippincott, 2001. Citado na página 41.

HIDALGO, M.; AMANT, F.; BIANKIN, A. V.; BUDINSKÁ, E.; BYRNE, A. T.; CALDAS, C.; CLARKE, R. B.; JONG, S. de; JONKERS, J.; MÆLANDSMO, G. M. *et al.* Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer discovery*, AACR, v. 4, n. 9, p. 998–1013, 2014. Citado na página 67.

HIPP, S.; VOYNOV, V.; DROBITS-HANDL, B.; GIRAGOSSIAN, C.; TRAPANI, F.; NIXON, A. E.; SCHEER, J. M.; ADAM, P. J. A bispecific dll3/cd3 igg-like t-cell engaging antibody induces antitumor responses in small cell lung cancer. *Clinical Cancer Research*, AACR, v. 26, n. 19, p. 5258–5268, 2020. Citado na página 42.

HIRANO, K.; MIKI, Y.; HIRAI, Y.; SATO, R.; ITOH, T.; HAYASHI, A.; YAMANAKA, M.; EDA, S.; BEPPU, M. A multifunctional shuttling protein nucleolin is a macrophage receptor for apoptotic cells. *Journal of Biological Chemistry*, ASBMB, v. 280, n. 47, p. 39284–39293, 2005. Citado na página 136.

HODGKINSON, C. L.; MORROW, C. J.; LI, Y.; METCALF, R. L.; ROTHWELL, D. G.; TRAPANI, F.; POLANSKI, R.; BURT, D. J.; SIMPSON, K. L.; MORRIS, K. *et al.* Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. *Nature medicine*, Nature Publishing Group, v. 20, n. 8, p. 897–903, 2014. Citado na página 67.

HOU, J.-M.; KREBS, M. G.; LANCASHIRE, L.; SLOANE, R.; BACKEN, A.; SWAIN, R. K.; PRIEST, L.; GREYSTOKE, A.; ZHOU, C.; MORRIS, K. *et al.* Clinical significance and molecular characteristics of circulating tumor cells and circulating tumor microemboli in patients with small-cell lung cancer. *J Clin Oncol*, v. 30, n. 5, p. 525–532, 2012. Citado na página 41.

HU, S.; ZHENG, W.; JIN, L. Astragaloside iv inhibits cell proliferation and metastasis of breast cancer via promoting the long noncoding rna trhde-as1. *Journal of natural medicines*, Springer, v. 75, n. 1, p. 156–166, 2021. Citado na página 91.

HU, X.; HU, Y.; WU, F.; LEUNG, R. W. T.; QIN, J. Integration of single-cell multi-omics for gene regulatory network inference. *Computational and structural biotechnology journal*, Elsevier, 2020. Citado na página 59.

HYSTAD, M. E.; MYKLEBUST, J. H.; BØ, T. H.; SIVERTSEN, E. A.; RIAN, E.; FORFANG, L.; MUNTHE, E.; ROSENWALD, A.; CHIORAZZI, M.; JONASSEN, I. *et al.* Characterization of early stages of human b cell development by gene expression profiling. *The Journal of Immunology*, Am Assoc Immnol, v. 179, n. 6, p. 3662–3671, 2007. Citado na página 74.

IYER, M. K.; NIKNAFS, Y. S.; MALIK, R.; SINGHAL, U.; SAHU, A.; HOSONO, Y.; BARRETTE, T. R.; PRENSNER, J. R.; EVANS, J. R.; ZHAO, S. *et al.* The landscape of long noncoding rnas in the human transcriptome. *Nature genetics*, Nature Publishing Group, v. 47, n. 3, p. 199–208, 2015. Citado 2 vezes nas páginas 47 e 48.

JERBY-ARNON, L.; SHAH, P.; CUOCO, M. S.; RODMAN, C.; SU, M.-J.; MELMS, J. C.; LEESON, R.; KANODIA, A.; MEI, S.; LIN, J.-R. *et al.* A cancer cell program promotes t cell exclusion and resistance to checkpoint blockade. *Cell*, Elsevier, v. 175, n. 4, p. 984–997, 2018. Citado 2 vezes nas páginas 30 e 36.

JIN, S.; GUERRERO-JUAREZ, C. F.; ZHANG, L.; CHANG, I.; RAMOS, R.; KUAN, C.-H.; MYUNG, P.; PLIKUS, M. V.; NIE, Q. Inference and analysis of cell-cell communication using cellchat. *Nature communications*, Nature Publishing Group, v. 12, n. 1, p. 1–20, 2021. Citado 2 vezes nas páginas 63 e 71.

JUNGBLUTH, A.; IVERSEN, K.; COPLAN, K.; WILLIAMSON, B.; CHEN, Y.-T.; STOCKERT, E.; OLD, L. J.; BUSAM, K. J. Expression of melanocyte-associated markers gp-100 and melan-a/mart-1 in angiomyolipomas. *Virchows Archiv*, Springer, v. 434, n. 5, p. 429–435, 1999. Citado na página 74.

KANNAN, S.; FANG, W.; SONG, G.; MULLIGHAN, C. G.; HAMMITT, R.; MCMURRAY, J.; ZWEIDLER-MCKAY, P. A. Notch/hes1-mediated parp1 activation: a cell type–specific mechanism for tumor suppression. *Blood, The Journal of the American Society of Hematology*, American Society of Hematology Washington, DC, v. 117, n. 10, p. 2891–2900, 2011. Citado na página 130.

KENT, W. J.; SUGNET, C. W.; FUREY, T. S.; ROSKIN, K. M.; PRINGLE, T. H.; ZAHLER, A. M.; HAUSSLER, D. The human genome browser at ucsc. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 6, p. 996–1006, 2002. Citado na página 88.

KIM, J.; EBERWINE, J. Rna: state memory and mediator of cellular phenotype. *Trends in cell biology*, Elsevier, v. 20, n. 6, p. 311–318, 2010. Citado na página 46.

KNELSON, E. H.; PATEL, S. A.; SANDS, J. M. Parp inhibitors in small-cell lung cancer: Rational combinations to improve responses. *Cancers*, Multidisciplinary Digital Publishing Institute, v. 13, n. 4, p. 727, 2021. Citado na página 44.

KOH, H. K.; GELLER, A. C.; LEW, R. A. *Melanoma*. [S.l.]: CRC Press, 2021. Citado 2 vezes nas páginas 37 e 38.

KSIENZYK, A.; NEUMANN, B.; NANDAKUMAR, R.; FINSTERBUSCH, K.; GRASHOFF, M.; ZAWATZKY, R.; BERNHARDT, G.; HAUSER, H.; KRÖGER, A. Irf-1 expression is essential for natural killer cells to suppress metastasis. *Cancer research*, AACR, v. 71, n. 20, p. 6410–6418, 2011. Citado na página 100.

LABROUSSE, A.-L.; NTAYI, C.; HORNEBECK, W.; BERNARD, P. Stromal reaction in cutaneous melanoma. *Critical reviews in oncology/hematology*, Elsevier, v. 49, n. 3, p. 269–275, 2004. Citado na página 40.

LAMBERT, S. A.; JOLMA, A.; CAMPITELLI, L. F.; DAS, P. K.; YIN, Y.; ALBU, M.; CHEN, X.; TAIPALE, J.; HUGHES, T. R.; WEIRAUCH, M. T. The human transcription factors. *Cell*, Elsevier, v. 172, n. 4, p. 650–665, 2018. Citado na página 46.

LAMOUILLE, S.; XU, J.; DERYNCK, R. Molecular mechanisms of epithelial–mesenchymal transition. *Nature reviews Molecular cell biology*, Nature Publishing Group, v. 15, n. 3, p. 178–196, 2014. Citado 2 vezes nas páginas 51 e 52.

LARUE, L.; DELMAS, V. The wnt/beta-catenin pathway in melanoma. *Front Biosci*, v. 11, n. 1, p. 733–742, 2006. Citado na página 39.

LEE, T.-F.; TSENG, Y.-C.; CHANG, W.-C.; CHEN, Y.-C.; KAO, Y.-R.; CHOU, T.-Y.; HO, C.-C.; WU, C.-W. Yap1 is essential for tumor growth and is a potential therapeutic target for egfr-dependent lung adenocarcinomas. *Oncotarget*, Impact Journals, LLC, v. 8, n. 52, p. 89539, 2017. Citado na página 126.

LEVY, C.; KHALED, M.; FISHER, D. E. Mitf: master regulator of melanocyte development and melanoma oncogene. *Trends in molecular medicine*, Elsevier, v. 12, n. 9, p. 406–414, 2006. Citado na página 100.

LI, G.; FUKUNAGA, M.; HERLYN, M. Reversal of melanocytic malignancy by keratinocytes is an e-cadherin-mediated process overriding β-catenin signaling. *Experimental cell research*, Elsevier, v. 297, n. 1, p. 142–151, 2004. Citado na página 39.

LI, G.; SATYAMOORTHY, K.; HERLYN, M. N-cadherin-mediated intercellular interactions promote survival and migration of melanoma cells. *Cancer research*, AACR, v. 61, n. 9, p. 3819–3825, 2001. Citado na página 39.

LI, H.; HANDSAKER, B.; WYSOKER, A.; FENNELL, T.; RUAN, J.; HOMER, N.; MARTH, G.; ABECASIS, G.; DURBIN, R. The sequence alignment/map format and samtools. *Bioinformatics*, Oxford University Press, v. 25, n. 16, p. 2078–2079, 2009. Citado na página 66.

LI, S.; TEEGARDEN, A.; BAUER, E. M.; CHOI, J.; MESSADDEQ, N.; HENDRIX, D. A.; GANGULI-INDRA, G.; LEID, M.; INDRA, A. K. Transcription factor ctip1/bcl11a regulates epidermal differentiation and lipid metabolism during skin development. *Scientific reports*, Nature Publishing Group, v. 7, n. 1, p. 1–16, 2017. Citado na página 100.

LIBERZON, A.; BIRGER, C.; THORVALDSDÓTTIR, H.; GHANDI, M.; MESIROV, J. P.; TAMAYO, P. The molecular signatures database hallmark gene set collection. *Cell systems*, Elsevier, v. 1, n. 6, p. 417–425, 2015. Citado 2 vezes nas páginas 57 e 92.

LIBERZON, A.; SUBRAMANIAN, A.; PINCHBACK, R.; THORVALDSDÓTTIR, H.; TAMAYO, P.; MESIROV, J. P. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, Oxford University Press, v. 27, n. 12, p. 1739–1740, 2011. Citado na página 57.

LIU, S.-H.; ZHU, J.-W.; XU, H.-H.; ZHANG, G.-Q.; WANG, Y.; LIU, Y.-M.; LIANG, J.-B.; WANG, Y.-X.; WU, Y.; GUO, Q.-F. A novel antisense long non-coding rna satb2-as1 overexpresses in osteosarcoma and increases cell proliferation and growth. *Molecular and cellular biochemistry*, Springer Nature BV, v. 430, n. 1-2, p. 47, 2017. Citado na página 91.

LIU, S. J.; NOWAKOWSKI, T. J.; POLLEN, A. A.; LUI, J. H.; HORLBECK, M. A.; ATTENELLO, F. J.; HE, D.; WEISSMAN, J. S.; KRIEGSTEIN, A. R.; DIAZ, A. A. *et al.* Single-cell analysis of long non-coding rnas in the developing human neocortex. *Genome biology*, Springer, v. 17, n. 1, p. 1–17, 2016. Citado na página 48.

LORENZEN, J. M.; THUM, T. Long noncoding rnas in kidney and cardiovascular diseases. *Nature Reviews Nephrology*, Nature Publishing Group, v. 12, n. 6, p. 360–373, 2016. Citado na página 47.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, BioMed Central, v. 15, n. 12, p. 1–21, 2014. Citado na página 56.

LUECKEN, M. D.; THEIS, F. J. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, v. 15, n. 6, p. e8746, 2019. Citado na página 35.

LV, D.; WANG, X.; DONG, J.; ZHUANG, Y.; HUANG, S.; MA, B.; CHEN, P.; LI, X.; ZHANG, B.; LI, Z. *et al.* Systematic characterization of lncrnas' cell-to-cell expression heterogeneity in glioblastoma cells. *Oncotarget*, Impact Journals, LLC, v. 7, n. 14, p. 18403, 2016. Citado na página 48.

MACNAIR, W.; CLAASSEN, M. psupertime: supervised pseudotime inference for single cell rna-seq data with sequential labels. *bioRxiv*, Cold Spring Harbor Laboratory, p. 622001, 2019. Citado na página 58.

MANNO, G. L.; SOLDATOV, R.; ZEISEL, A.; BRAUN, E.; HOCHGERNER, H.; PETUKHOV, V.; LIDSCHREIBER, K.; KASTRITI, M. E.; LÖNNERBERG, P.; FURLAN, A. *et al.* Rna velocity of single cells. *Nature*, Nature Publishing Group, v. 560, n. 7719, p. 494–498, 2018. Citado 2 vezes nas páginas 63 e 71.

MARIGNOL, L. Notch signalling: the true driver of small cell lung cancer? *Nature*, v. 545, p. 360–4, 2017. Citado na página 130.

MARUSYK, A.; ALMENDRO, V.; POLYAK, K. Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, Nature Publishing Group, v. 12, n. 5, p. 323–334, 2012. Citado na página 48.

MATSUDA, T.; OKUYAMA, A. Cancer incidence rates in the world from the cancer incidence in five continents xi. *Japanese journal of clinical oncology*, Oxford University Press, v. 48, n. 2, p. 202–203, 2018. Citado na página 41.

MATTEI, F.; SCHIAVONI, G.; SESTILI, P.; SPADARO, F.; FRAGALE, A.; SISTIGU, A.; LUCARINI, V.; SPADA, M.; SANCHEZ, M.; SCALA, S. *et al.* Irf-8 controls melanoma progression by regulating the cross talk between cancer and immune cells within the tumor microenvironment. *Neoplasia*, Elsevier, v. 14, n. 12, p. 1223–IN43, 2012. Citado na página 100.

MCGINNIS, C. S.; MURROW, L. M.; GARTNER, Z. J. Doubletfinder: doublet detection in single-cell rna sequencing data using artificial nearest neighbors. *Cell systems*, Elsevier, v. 8, n. 4, p. 329–337, 2019. Citado na página 33.

MCINNES, L.; HEALY, J.; MELVILLE, J. Umap: uniform manifold approximation and projection for dimension reduction. 2020. Citado na página 72.

MORRIS, K. V.; MATTICK, J. S. The rise of regulatory rna. *Nature Reviews Genetics*, Nature Publishing Group, v. 15, n. 6, p. 423–437, 2014. Citado na página 47.

MOSCHOS, S. J.; DROGOWSKI, L. M.; REPPERT, S. L.; KIRKWOOD, J. M. Integrins and cancer. *Oncology*, MultiMedia Healthcare Inc., v. 21, n. 9, p. 13, 2007. Citado na página 40.

NAVIN, N.; KENDALL, J.; TROGE, J.; ANDREWS, P.; RODGERS, L.; MCINDOO, J.; COOK, K.; STEPANSKY, A.; LEVY, D.; ESPOSITO, D. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature*, Nature Publishing Group, v. 472, n. 7341, p. 90–94, 2011. Citado na página 48.

NGUYEN, J.; BERNERT, R.; IN, K.; KANG, P.; SEBASTIAO, N.; HU, C.; HASTINGS, K. T. Gamma-interferon-inducible lysosomal thiol reductase is upregulated in human melanoma. *Melanoma research*, NIH Public Access, v. 26, n. 2, p. 125, 2016. Citado na página 74.

NISTICÒ, P.; BISSELL, M. J.; RADISKY, D. C. Epithelial-mesenchymal transition: general principles and pathological relevance with special emphasis on the role of matrix metalloproteinases. *Cold Spring Harbor perspectives in biology*, Cold Spring Harbor Lab, v. 4, n. 2, p. a011908, 2012. Citado na página 52.

PANG, B.; WANG, Q.; NING, S.; WU, J.; ZHANG, X.; CHEN, Y.; XU, S. Landscape of tumor suppressor long noncoding rnas in breast cancer. *Journal of Experimental & Clinical Cancer Research*, Springer, v. 38, n. 1, p. 1–18, 2019. Citado na página 91.

PELOSI, G.; LEON, M. E.; VERONESI, G.; SPAGGIARI, L.; PASINI, F.; VIALE, G. Decreased immunoreactivity of cd99 is an independent predictor of regional lymph node metastases in pulmonary carcinoid tumors. *Journal of Thoracic Oncology*, Elsevier, v. 1, n. 5, p. 468–477, 2006. Citado na página 139.

PICELLI, S.; BJÖRKLUND, Å. K.; FARIDANI, O. R.; SAGASSER, S.; WINBERG, G.; SANDBERG, R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods*, Nature Publishing Group, v. 10, n. 11, p. 1096–1098, 2013. Citado na página 32.

PICELLI, S.; FARIDANI, O. R.; BJÖRKLUND, Å. K.; WINBERG, G.; SAGASSER, S.; SANDBERG, R. Full-length rna-seq from single cells using smart-seq2. *Nature protocols*, Nature Publishing Group, v. 9, n. 1, p. 171–181, 2014. Citado na página 32.

POSER, I.; DOMINGUEZ, D.; HERREROS, A. G. de; VARNAI, A.; BUETTNER, R.; BOSSERHOFF, A. K. Loss of e-cadherin expression in melanoma cells involves up-regulation of the transcriptional repressor snail. *Journal of Biological Chemistry*, ASBMB, v. 276, n. 27, p. 24661–24666, 2001. Citado na página 39.

PRAKASH, J. Cancer-associated fibroblasts: perspectives in cancer therapy. *Trends in cancer*, Elsevier, v. 2, n. 6, p. 277–279, 2016. Citado na página 108.

QIU, X.; HILL, A.; PACKER, J.; LIN, D.; MA, Y.-A.; TRAPNELL, C. Single-cell mrna quantification and differential analysis with census. *Nature methods*, Nature Publishing Group, v. 14, n. 3, p. 309–315, 2017. Citado na página 55.

QIU, X.; MAO, Q.; TANG, Y.; WANG, L.; CHAWLA, R.; PLINER, H. A.; TRAPNELL, C. Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, Nature Publishing Group, v. 14, n. 10, p. 979–982, 2017. Citado na página 55.

REVERTER, A.; CHAN, E. K. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. *Bioinformatics*, Oxford University Press, v. 24, n. 21, p. 2491–2497, 2008. Citado na página 60.

REVERTER, A.; HUDSON, N. J.; NAGARAJ, S. H.; PÉREZ-ENCISO, M.; DALRYMPLE, B. P. Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. *Bioinformatics*, Oxford Academic, v. 26, n. 7, p. 896–904, 2010. Citado na página 59.

RISSO, D.; PERRAUDEAU, F.; GRIBKOVA, S.; DUDOIT, S.; VERT, J.-P. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, Nature Publishing Group, v. 9, n. 1, p. 1–17, 2018. Citado na página 56.

RIVERO, S.; CEBALLOS-CHAVEZ, M.; BHATTACHARYA, S.; REYES, J. Hmg20a is required for snai1-mediated epithelial to mesenchymal transition. *Oncogene*, Nature Publishing Group, v. 34, n. 41, p. 5264–5276, 2015. Citado na página 99.

ROLNY, C.; CAPPARUCCIA, L.; CASAZZA, A.; MAZZONE, M.; VALLARIO, A.; CIGNETTI, A.; MEDICO, E.; CARMELIET, P.; COMOGLIO, P. M.; TAMAGNONE, L. The tumor suppressor semaphorin 3b triggers a prometastatic program mediated by interleukin 8 and the tumor microenvironment. *The Journal of experimental medicine*, Rockefeller University Press, v. 205, n. 5, p. 1155–1171, 2008. Citado na página 78.

RUDIN, C. M.; PIETANZA, M. C.; BAUER, T. M.; READY, N.; MORGENSZTERN, D.; GLISSON, B. S.; BYERS, L. A.; JOHNSON, M. L.; III, H. A. B.; ROBERT, F. *et al.* Rovalpituzumab tesirine, a dll3-targeted antibody-drug conjugate, in recurrent small-cell lung cancer: a first-in-human, first-in-class, open-label, phase 1 study. *The Lancet Oncology*, Elsevier, v. 18, n. 1, p. 42–51, 2017. Citado na página 42.

RUDIN, C. M.; POIRIER, J. T.; BYERS, L. A.; DIVE, C.; DOWLATI, A.; GEORGE, J.; HEYMACH, J. V.; JOHNSON, J. E.; LEHMAN, J. M.; MACPHERSON, D. *et al.* Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer*, Nature Publishing Group, v. 19, n. 5, p. 289–297, 2019. Citado 4 vezes nas páginas 44, 124, 142 e 146.

RUITER, D.; BOGENRIEDER, T.; ELDER, D.; HERLYN, M. Melanoma–stroma interactions: structural and functional aspects. *The lancet oncology*, Elsevier, v. 3, n. 1, p. 35–43, 2002. Citado na página 40.

SAID, E. A.; KRUST, B.; NISOLE, S.; SVAB, J.; BRIAND, J.-P.; HOVANESSIAN, A. G. The anti-hiv cytokine midkine binds the cell surface-expressed nucleolin as a low affinity receptor. *Journal of Biological Chemistry*, ASBMB, v. 277, n. 40, p. 37492–37502, 2002. Citado na página 136.

SAMATOV, T. R.; TONEVITSKY, A. G.; SCHUMACHER, U. Epithelial-mesenchymal transition: focus on metastatic cascade, alternative splicing, non-coding rnas and modulating compounds. *Molecular cancer*, Springer, v. 12, n. 1, p. 1–12, 2013. Citado 2 vezes nas páginas 51 e 52.

SANTOS, C. A. d.; SOUZA, D. L. B. Melanoma mortality in brazil: trends and projections (1998-2032). *Ciencia & saude coletiva*, SciELO Public Health, v. 24, p. 1551–1561, 2019. Citado na página 38.

SASAKI, K.; SUGAI, T.; ISHIDA, K.; OSAKABE, M.; AMANO, H.; KIMURA, H.; SAKURABA, M.; KASHIWA, K.; KOBAYASHI, S. Analysis of cancer-associated fibroblasts and the epithelial-mesenchymal transition in cutaneous basal cell carcinoma, squamous cell carcinoma, and malignant melanoma. *Human pathology*, Elsevier, v. 79, p. 1–8, 2018. Citado na página 108.

SCHATTON, T.; SCHÜTTE, U.; FRANK, N. Y.; ZHAN, Q.; HOERNING, A.; ROBLES, S. C.; ZHOU, J.; HODI, F. S.; SPAGNOLI, G. C.; MURPHY, G. F. *et al.* Modulation of t-cell activation by malignant melanoma initiating cells. *Cancer research*, AACR, v. 70, n. 2, p. 697–708, 2010. Citado na página 102.

SCHIAVONI, G.; GABRIELE, L.; MATTEI, F. The dual role of irf8 in cancer immunosurveillance. *Oncoimmunology*, Taylor & Francis, v. 2, n. 8, p. e25476, 2013. Citado na página 100.

SENA, J. A.; GALOTTO, G.; DEVITT, N. P.; CONNICK, M. C.; JACOBI, J. L.; UMALE, P. E.; VIDALI, L.; BELL, C. J. Unique molecular identifiers reveal a novel sequencing artefact with implications for rna-seq based gene expression analysis. *Scientific reports*, Nature Publishing Group, v. 8, n. 1, p. 1–13, 2018. Citado na página 67.

SHALEK, A. K.; BENSON, M. Single-cell analyses to tailor treatments. *Science translational medicine*, American Association for the Advancement of Science, v. 9, n. 408, 2017. Citado na página 48.

SHANG, D.; ZHENG, T.; ZHANG, J.; TIAN, Y.; LIU, Y. Profiling of mrna and long non-coding rna of urothelial cancer in recipients after renal transplantation. *Tumor Biology*, Springer, v. 37, n. 9, p. 12673–12684, 2016. Citado na página 91.

SHANNON, P.; MARKIEL, A.; OZIER, O.; BALIGA, N. S.; WANG, J. T.; RAMAGE, D.; AMIN, N.; SCHWIKOWSKI, B.; IDEKER, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, Cold Spring Harbor Lab, v. 13, n. 11, p. 2498–2504, 2003. Citado na página 60.

SHIBATA, Y.; MURAMATSU, T.; HIRAI, M.; INUI, T.; KIMURA, T.; SAITO, H.; MCCORMICK, L. M.; BU, G.; KADOMATSU, K. Nuclear targeting by the growth factor midkine. *Molecular and cellular biology*, Am Soc Microbiol, v. 22, n. 19, p. 6788–6796, 2002. Citado na página 136.

SIMON, J. A.; KINGSTON, R. E. Mechanisms of polycomb gene silencing: knowns and unknowns. *Nature reviews Molecular cell biology*, Nature Publishing Group, v. 10, n. 10, p. 697–708, 2009. Citado na página 88.

SORRELLE, N.; DOMINGUEZ, A. T.; BREKKEN, R. A. From top to bottom: midkine and pleiotrophin as emerging players in immune regulation. *Journal of leukocyte biology*, Wiley Online Library, v. 102, n. 2, p. 277–286, 2017. Citado na página 136.

STEGLE, O.; TEICHMANN, S. A.; MARIONI, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, Nature Publishing Group, v. 16, n. 3, p. 133–145, 2015. Citado na página 34.

STERLACCI, W.; FIEGL, M.; TZANKOV, A. Prognostic and predictive value of cell cycle deregulation in non-small-cell lung cancer. *Pathobiology*, Karger Publishers, v. 79, n. 4, p. 175–194, 2012. Citado na página 113.

STEWART, C. A.; GAY, C. M.; XI, Y.; SIVAJOTHI, S.; SIVAKAMASUNDARI, V.; FUJIMOTO, J.; BOLISETTY, M.; HARTSFIELD, P. M.; BALASUBRAMANIYAN, V.; CHALISHAZAR, M. D. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nature Cancer*, Nature Publishing Group, v. 1, n. 4, p. 423–436, 2020. Citado na página 63.

STUART, T.; BUTLER, A.; HOFFMAN, P.; HAFEMEISTER, C.; PAPALEXI, E.; III, W. M. M.; HAO, Y.; STOECKIUS, M.; SMIBERT, P.; SATIJA, R. Comprehensive integration of single-cell data. *Cell*, Elsevier, v. 177, n. 7, p. 1888–1902, 2019. Citado 3 vezes nas páginas 34, 54 e 56.

SUNG, H.; FERLAY, J.; SIEGEL, R. L.; LAVERSANNE, M.; SOERJOMATARAM, I.; JEMAL, A.; BRAY, F. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 71, n. 3, p. 209–249, 2021. Citado na página 41.

SUVÀ, M. L.; TIROSH, I. Single-cell rna sequencing in cancer: lessons learned and emerging challenges. *Molecular cell*, Elsevier, v. 75, n. 1, p. 7–12, 2019. Citado na página 37.

TANG, F.; BARBACIORU, C.; WANG, Y.; NORDMAN, E.; LEE, C.; XU, N.; WANG, X.; BODEAU, J.; TUCH, B. B.; SIDDIQUI, A. *et al.* mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, Nature Publishing Group, v. 6, n. 5, p. 377–382, 2009. Citado na página 29.

TANG, L.; ZHANG, W.; SU, B.; YU, B. Long noncoding rna hotair is associated with motility, invasion, and metastatic potential of metastatic melanoma. *BioMed research international*, Hindawi, v. 2013, 2013. Citado na página 51.

TANG, Q.; HANN, S. S. HOTAIR: An oncogenic long non-coding RNA in human cancer. *Cellular Physiology and Biochemistry*, S. Karger AG, v. 47, n. 3, p. 893–913, 2018. Citado na página 50.

TAWBI, H. A.; KIRKWOOD, J. M. Management of metastatic melanoma. In: ELSEVIER. *Seminars in oncology*. [S.l.], 2007. v. 34, n. 6, p. 532–545. Citado na página 40.

TERAI, G.; IWAKIRI, J.; KAMEDA, T.; HAMADA, M.; ASAI, K. Comprehensive prediction of lncrna–rna interactions in human transcriptome. In: SPRINGER. *BMC genomics*. [S.l.], 2016. v. 17, n. 1, p. 153–164. Citado na página 89.

THIERY, J. P.; ACLOQUE, H.; HUANG, R. Y.; NIETO, M. A. Epithelial-mesenchymal transitions in development and disease. *cell*, Elsevier, v. 139, n. 5, p. 871–890, 2009. Citado na página 52.

TIROSH, I.; IZAR, B.; PRAKADAN, S. M.; WADSWORTH, M. H.; TREACY, D.; TROMBETTA, J. J.; ROTEM, A.; RODMAN, C.; LIAN, C.; MURPHY, G. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, American Association for the Advancement of Science, v. 352, n. 6282, p. 189–196, 2016. Citado 4 vezes nas páginas 30, 48, 55 e 68.

TIROSH, I.; VENTEICHER, A. S.; HEBERT, C.; ESCALANTE, L. E.; PATEL, A. P.; YIZHAK, K.; FISHER, J. M.; RODMAN, C.; MOUNT, C.; FILBIN, M. G. *et al.* Single-cell rna-seq supports a developmental hierarchy in human oligodendroglioma. *Nature*, Nature Publishing Group, v. 539, n. 7628, p. 309–313, 2016. Citado 2 vezes nas páginas 30 e 48.

TONG, M.; JIANG, Y. Fk506-binding proteins and their diverse functions. *Current molecular pharmacology*, Bentham Science Publishers, v. 9, n. 1, p. 48–65, 2016. Citado na página 102.

TRUFFI, M.; SORRENTINO, L.; CORSI, F. Fibroblasts in the tumor microenvironment. *Tumor Microenvironment*, Springer, p. 15–29, 2020. Citado na página 45.

TSAI, M.-C.; MANOR, O.; WAN, Y.; MOSAMMAPARAST, N.; WANG, J. K.; LAN, F.; SHI, Y.; SEGAL, E.; CHANG, H. Y. Long noncoding rna as modular scaffold of histone modification complexes. *Science*, American Association for the Advancement of Science, v. 329, n. 5992, p. 689–693, 2010. Citado na página 88.

TSUJI, T.; OZASA, H.; AOKI, W.; ABURAYA, S.; FUNAZO, T. Y.; FURUGAKI, K.; YOSHIMURA, Y.; YAMAZOE, M.; AJIMIZU, H.; YASUDA, Y. *et al.* Yap1 mediates survival of alk-rearranged lung cancer cells treated with alectinib via pro-apoptotic protein regulation. *Nature communications*, Nature Publishing Group, v. 11, n. 1, p. 1–16, 2020. Citado na página 126.

VAQUERIZAS, J. M.; KUMMERFELD, S. K.; TEICHMANN, S. A.; LUSCOMBE, N. M. A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, Nature Publishing Group, v. 10, n. 4, p. 252–263, 2009. Citado na página 59.

VENTEICHER, A. S.; TIROSH, I.; HEBERT, C.; YIZHAK, K.; NEFTEL, C.; FILBIN, M. G.; HOVESTADT, V.; ESCALANTE, L. E.; SHAW, M. L.; RODMAN, C. *et al.* Decoupling genetics, lineages, and microenvironment in idh-mutant gliomas by single-cell rna-seq. *Science*, American Association for the Advancement of Science, v. 355, n. 6332, 2017. Citado na página 30.

WANG, L.-x.; LI, Y.; CHEN, G.-z. Network-based co-expression analysis for exploring the potential diagnostic biomarkers of metastatic melanoma. *PloS one*, Public Library of Science San Francisco, CA USA, v. 13, n. 1, p. e0190447, 2018. Citado na página 99.

WANG, T.; LI, B.; NELSON, C. E.; NABAVI, S. Comparative analysis of differential gene expression analysis tools for single-cell rna sequencing data. *BMC bioinformatics*, BioMed Central, v. 20, n. 1, p. 1–16, 2019. Citado na página 56.

WANG, X.; HE, Y.; ZHANG, Q.; REN, X.; ZHANG, Z. Direct comparative analyses of 10x genomics chromium and smart-seq2. *Genomics, Proteomics & Bioinformatics*, Elsevier, 2021. Citado na página 32.

WAPINSKI, O.; CHANG, H. Y. Long noncoding rnas and human disease. *Trends in cell biology*, Elsevier, v. 21, n. 6, p. 354–361, 2011. Citado na página 47.

WEI, Y.; WANG, T.; ZHANG, N.; MA, Y.; SHI, S.; ZHANG, R.; ZHENG, X.; ZHAO, L. Lncrna trhde-as1 inhibit the scar fibroblasts proliferation via mir-181a-5p/pten axis. *Journal of Molecular Histology*, Springer, v. 52, n. 2, p. 419–426, 2021. Citado na página 91.

WIDLUND, H. R.; HORSTMANN, M. A.; PRICE, E. R.; CUI, J.; LESSNICK, S. L.; WU, M.; HE, X.; FISHER, D. E. β-catenin–induced melanoma growth requires the downstream target microphthalmia-associated transcription factor. *The Journal of cell biology*, The Rockefeller University Press, v. 158, n. 6, p. 1079–1087, 2002. Citado na página 39.

WIDMER, D. S.; CHENG, P. F.; EICHHOFF, O. M.; BELLONI, B. C.; ZIPSER, M. C.; SCHLEGEL, N. C.; JAVELAUD, D.; MAUVIEL, A.; DUMMER, R.; HOEK, K. S. Systematic classification of melanoma cells by phenotype-specific gene expression mapping. *Pigment cell & melanoma research*, Wiley Online Library, v. 25, n. 3, p. 343–353, 2012. Citado na página 92.

WILD, C. P.; STEWART, B. W.; WILD, C. *World cancer report 2014*. [S.l.]: World Health Organization Geneva, Switzerland, 2014. Citado na página 38.

WILLSMORE, Z. N.; HARRIS, R. J.; CRESCIOLI, S.; HUSSEIN, K.; KAKKASSERY, H.; THAT, D.; CHEUNG, A. K.; CHAUHAN, J.; BAX, H. J.; CHENOWETH, A. *et al.* B cells in patients with melanoma: implications for treatment with checkpoint inhibitor antibodies. *Frontiers in Immunology*, Frontiers, v. 11, p. 3560, 2020. Citado na página 103.

WOLF, F. A.; ANGERER, P.; THEIS, F. J. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, Springer, v. 19, n. 1, p. 1–5, 2018. Citado na página 34.

WOLOCK, S. L.; LOPEZ, R.; KLEIN, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, Elsevier, v. 8, n. 4, p. 281–291, 2019. Citado na página 33.

WU, Y.; DENG, J.; LAI, S.; YOU, Y.; WU, J. A risk score model with five long non-coding rnas for predicting prognosis in gastric cancer: an integrated analysis combining tcga and geo datasets. *PeerJ*, PeerJ Inc., v. 9, p. e10556, 2021. Citado na página 91.

WU, Y.; ZHANG, L.; WANG, Y.; LI, H.; REN, X.; WEI, F.; YU, W.; WANG, X.; ZHANG, L.; YU, J. *et al.* Long noncoding rna hotair involvement in cancer. *Tumor Biology*, Springer, v. 35, n. 10, p. 9531–9538, 2014. Citado na página 51.

XI, N. M.; LI, J. J. Benchmarking computational doublet-detection methods for single-cell rna sequencing data. *Cell systems*, Elsevier, v. 12, n. 2, p. 176–194, 2021. Citado na página 33.

XIAO, Y.; YU, D. Tumor microenvironment as a therapeutic target in cancer. *Pharmacology & Therapeutics*, Elsevier, v. 221, p. 107753, 2021. Citado na página 45.

XING, Q.; HUANG, Y.; WU, Y.; MA, L.; CAI, B. Integrated analysis of differentially expressed profiles and construction of a competing endogenous long non-coding rna network in renal cell carcinoma. *PeerJ*, PeerJ Inc., v. 6, p. e5124, 2018. Citado na página 91.

XIONG, K.-X.; ZHOU, H.-L.; YIN, J.-H.; KRISTIANSEN, K.; YANG, H.-M.; LI, G.-B. Chord: Identifying doublets in single-cell rna sequencing data by an ensemble machine learning algorithm. *bioRxiv*, Cold Spring Harbor Laboratory, 2021. Citado 2 vezes nas páginas 63 e 68.

YANG, F.; LV, S.-X.; LV, L.; LIU, Y.-H.; DONG, S.-Y.; YAO, Z.-H.; DAI, X.-x.; ZHANG, X.-H.; WANG, O.-C. Identification of lncrna fam83h-as1 as a novel prognostic marker in luminal subtype breast cancer. *OncoTargets and therapy*, Dove Press, v. 9, p. 7039, 2016. Citado na página 91.

YANG, Y.; XIN, X.; FU, X.; XU, D. Expression pattern of human serpine2 in a variety of human tumors. *Oncology letters*, Spandidos Publications, v. 15, n. 4, p. 4523–4530, 2018. Citado na página 74.

YAO, D.; DAI, C.; PENG, S. Mechanism of the mesenchymal–epithelial transition and its relationship with metastatic tumor formation. *Molecular cancer research*, AACR, v. 9, n. 12, p. 1608–1620, 2011. Citado na página 52.

YOU, S.; GAO, L. Identification of nmu as a potential gene conferring alectinib resistance in non-small cell lung cancer based on bioinformatics analyses. *Gene*, Elsevier, v. 678, p. 137–142, 2018. Citado na página 139.

YU, D.; KIM, M.; XIAO, G.; HWANG, T. H. Review of biological network data and its applications. *Genomics & informatics*, Korea Genome Organization, v. 11, n. 4, p. 200, 2013. Citado na página 59.

ZAPPIA, L.; PHIPSON, B.; OSHLACK, A. Exploring the single-cell rna-seq analysis landscape with the scrna-tools database. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 14, n. 6, p. e1006245, 2018. Citado na página 35.

ZAPPIA, L.; THEIS, F. J. Over 1000 tools reveal trends in the single-cell rna-seq analysis landscape. *Genome biology*, Springer, v. 22, n. 1, p. 1–18, 2021. Citado na página 35.

ZHENG, G. X.; TERRY, J. M.; BELGRADER, P.; RYVKIN, P.; BENT, Z. W.; WILSON, R.; ZIRALDO, S. B.; WHEELER, T. D.; MCDERMOTT, G. P.; ZHU, J. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications*, Nature Publishing Group, v. 8, n. 1, p. 1–12, 2017. Citado na página 33.

ZHUAN, B.; LU, Y.; CHEN, Q.; ZHAO, X.; LI, P.; YUAN, Q.; YANG, Z. Overexpression of the long noncoding rna trhde-as1 inhibits the progression of lung cancer via the mirna-103/klf4 axis. *Journal of cellular biochemistry*, Wiley Online Library, v. 120, n. 10, p. 17616–17624, 2019. Citado na página 91.

Apêndice A – Lista das assinaturas propostas

Tabela 10 – Assinatura a partir dos marcadores (expressão diferencial) para as células B.

ID	Ensembl ID	Gene Symbol	Gene Name
1	ENSG00000153064	BANK1	B-cell scaffold protein with ankyrin repeats 1
2	ENSG00000177455	CD19	CD19 molecule
3	ENSG00000104894	CD37	CD37 molecule
4	ENSG00000019582	CD74	CD74 molecule, major histocompatibility complex, class II invariant chain
5	ENSG00000105369	CD79A	CD79a molecule, immunoglobulin-associated alpha
6	ENSG00000007312	CD79B	CD79b molecule, immunoglobulin-associated beta
7	ENSG00000112149	CD83	CD83 molecule
8	ENSG00000241674	HLA-DMB	major histocompatibility complex, class II, DM beta
9	ENSG00000243496	HLA-DOB	major histocompatibility complex, class II, DO beta
10	ENSG00000168384	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1
11	ENSG00000206302	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1
12	ENSG00000206308	HLA-DRA	major histocompatibility complex, class II, DR alpha
13	ENSG00000254709	IGLL5	immunoglobulin lambda-like polypeptide 5
14	ENSG00000140968	IRF8	interferon regulatory factor 8
15	ENSG00000222041	LINC00152	long intergenic non-protein coding RNA 152
16	ENSG00000156738	MS4A1	membrane-spanning 4-domains, subfamily A, member 1
17	ENSG00000131401	NAPSB	napsin B aspartic peptidase, pseudogene
18	ENSG00000158517	NCF1	neutrophil cytosolic factor 1
19	ENSG00000128218	VPREB3	pre-B lymphocyte 3

Tabela 11 – Assinatura a partir dos marcadores (expressão diferencial) para as células *CAF*.

1	ENSG00000107796	ACTA2	actin, alpha 2, smooth muscle, aorta
2	ENSG00000154734	ADAMTS1	ADAM metallopeptidase with thrombospondin type 1 motif, 1
3	ENSG00000182492	BGN	biglycan
4	ENSG00000165507	C10orf10	chromosome 10 open reading frame 10
5	ENSG00000159403	C1R	complement component 1, r subcomponent
6	ENSG00000182326	C1S	complement component 1, s subcomponent
7	ENSG00000125730	C3	complement component 3
8	ENSG00000122786	CALD1	caldesmon 1
9	ENSG00000091986	CCDC80	coiled-coil domain containing 80
10	ENSG00000174807	CD248	CD248 molecule, endosialin
11	ENSG00000000971	CFH	complement factor H
12	ENSG00000120885	CLU	clusterin
13	ENSG00000108821	COL1A1	collagen, type I, alpha 1
14	ENSG00000164692	COL1A2	collagen, type I, alpha 2
15	ENSG00000168542	COL3A1	collagen, type III, alpha 1
16	ENSG00000142156	COL6A1	collagen, type VI, alpha 1
17	ENSG00000163359	COL6A3	collagen, type VI, alpha 3
18	ENSG00000118523	CTGF	connective tissue growth factor
19	ENSG00000164932	CTHRC1	collagen triple helix repeat containing 1
20	ENSG00000143387	CTSK	cathepsin K
21	ENSG00000107562	CXCL12	chemokine (C-X-C motif) ligand 12
22	ENSG00000145824	CXCL14	chemokine (C-X-C motif) ligand 14
23	ENSG00000142871	CYR61	cysteine-rich, angiogenic inducer, 61
24	ENSG00000011465	DCN	decorin
25	ENSG00000143196	DPT	dermatopontin
26	ENSG00000115380	EFEMP1	EGF containing fibulin-like extracellular matrix protein 1
27	ENSG00000120738	EGR1	early growth response 1
28	ENSG00000077942	FBLN1	fibulin 1
29	ENSG00000163520	FBLN2	fibulin 2
30	ENSG00000115414	FN1	fibronectin 1
31	ENSG00000163430	FSTL1	follistatin-like 1
32	ENSG00000142089	IFITM3	interferon induced transmembrane protein 3
33	ENSG00000141753	IGFBP4	insulin-like growth factor binding protein 4
34	ENSG00000167779	IGFBP6	insulin-like growth factor binding protein 6
35	ENSG00000163453	IGFBP7	insulin-like growth factor binding protein 7
36	ENSG00000129009	ISLR	immunoglobulin superfamily containing leucine-rich repeat
37	ENSG00000183722	LHFP	lipoma HMGIC fusion partner
38	ENSG00000139329	LUM	lumican
39	ENSG00000234456	MAGI2-AS3	MAGI2 antisense RNA 3
40	ENSG00000214548	MEG3	maternally expressed 3 (non-protein coding)
41	ENSG00000166482	MFAP4	microfibrillar-associated protein 4
42	ENSG00000197614	MFAP5	microfibrillar associated protein 5
43	ENSG00000111341	MGP	matrix Gla protein
44	ENSG00000087245	MMP2	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
45	ENSG00000162576	MXRA8	matrix-remodelling associated 8
46	ENSG00000166741	NNMT	nicotinamide N-methyltransferase
47	ENSG00000116774	OLFML3	olfactomedin-like 3
48	ENSG00000106333	PCOLCE	procollagen C-endopeptidase enhancer
49	ENSG00000174348	PODN	podocan
50	ENSG00000106538	RARRES2	retinoic acid receptor responder (tazarotene induced) 2
51	ENSG00000106366	SERPINE1	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
52	ENSG00000132386	SERPINF1	serpin peptidase inhibitor, clade F (alpha-2 antiplasmin, pigment epithelium derived factor), member 1
53	ENSG00000149131	SERPING1	serpin peptidase inhibitor, clade G (C1 inhibitor), member 1
54	ENSG00000113140	SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
55	ENSG00000149591	TAGLN	transgelin
56	ENSG00000120708	TGFBI	transforming growth factor, beta-induced, 68kDa
57	ENSG00000137801	THBS1	thrombospondin 1
58	ENSG00000154096	THY1	Thy-1 cell surface antigen
59	ENSG00000102265	TIMP1	TIMP metallopeptidase inhibitor 1
60	ENSG00000140416	TPM1	tropomyosin 1 (alpha)
61	ENSG00000198467	TPM2	tropomyosin 2 (beta)
62	ENSG00000176014	TUBB6	tubulin, beta 6 class V

Tabela 12 – Assinatura a partir dos marcadores (expressão diferencial) para as células endoteliais.

1	ENSG00000142192	APP	amyloid beta (A4) precursor protein
2	ENSG00000213494	CCL14	chemokine (C-C motif) ligand 14
3	ENSG00000137077	CCL21	chemokine (C-C motif) ligand 21
4	ENSG00000174059	CD34	CD34 molecule
5	ENSG00000010278	CD9	CD9 molecule
6	ENSG00000179776	CDH5	cadherin 5, type 2 (vascular endothelium)
7	ENSG00000184113	CLDN5	claudin 5
8	ENSG00000176435	CLEC14A	C-type lectin domain family 14, member A
9	ENSG00000120885	CLU	clusterin
10	ENSG00000187498	COL4A1	collagen, type IV, alpha 1
11	ENSG00000134871	COL4A2	collagen, type IV, alpha 2
12	ENSG00000182809	CRIP2	cysteine-rich protein 2
13	ENSG00000144476	CXCR7	C-X-C Chemokine Receptor Type 7
14	ENSG00000142871	CYR61	cysteine-rich, angiogenic inducer, 61
15	ENSG00000249751	ECSCR	endothelial cell surface expressed chemotaxis and apoptosis regulator
16	ENSG00000115380	EFEMP1	EGF containing fibulin-like extracellular matrix protein 1
17	ENSG00000172889	EGFL7	EGF-like-domain, multiple 7
18	ENSG00000106991	ENG	endoglin
19	ENSG00000170323	FABP4	fatty acid binding protein 4, adipocyte
20	ENSG00000164687	FABP5	fatty acid binding protein 5 (psoriasis-associated)
21	ENSG00000127920	GNG11	guanine nucleotide binding protein (G protein), gamma 11
22	ENSG00000261921	HYAL2	hyaluronoglucosaminidase 2
23	ENSG00000125968	ID1	inhibitor of DNA binding 1, dominant negative helix-loop-helix protein
24	ENSG00000142089	IFITM3	interferon induced transmembrane protein 3
25	ENSG00000163453	IGFBP7	insulin-like growth factor binding protein 7
26	ENSG00000270408	JAG1	Jagged Canonical Notch Ligand 1
27	ENSG00000133800	LYVE1	lymphatic vessel endothelial hyaluronan receptor 1
28	ENSG00000111341	MGP	matrix Gla protein
29	ENSG00000138722	MMRN1	multimerin 1
30	ENSG00000166741	NNMT	nicotinamide N-methyltransferase
31	ENSG00000107281	NPDC1	neural proliferation, differentiation and control, 1
32	ENSG00000107438	PDLIM1	PDZ and LIM domain 1
33	ENSG00000130300	PLVAP	plasmalemma vesicle associated protein
34	ENSG00000067113	PPAP2A	phosphatidic acid phosphatase type 2A
35	ENSG00000137509	PRCP	prolylcarboxypeptidase (angiotensinase C)
36	ENSG00000131477	RAMP2	receptor (G protein-coupled) activity modifying protein 2
37	ENSG00000122679	RAMP3	receptor (G protein-coupled) activity modifying protein 3
38	ENSG00000188643	S100A16	S100 calcium binding protein A16
39	ENSG00000168497	SDPR	serum deprivation response
40	ENSG00000174640	SLCO2A1	solute carrier organic anion transporter family, member 2A1
41	ENSG00000003436	TFPI	tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)
42	ENSG00000198959	TGM2	transglutaminase 2
43	ENSG00000100234	TIMP3	TIMP metallopeptidase inhibitor 3
44	ENSG00000169908	TM4SF1	transmembrane 4 L six family member 1
45	ENSG00000156298	TSPAN7	tetraspanin 7
46	ENSG00000110799	VWF	von Willebrand factor

Tabela 13 – Assinatura a partir dos marcadores (expressão diferencial) para as células macrofagiais.

1	ENSG00000237727	AIF1	allograft inflammatory factor 1
2	ENSG00000130208	APOC1	apolipoprotein C-I
3	ENSG00000173372	C1QA	complement component 1, q subcomponent, A chain
4	ENSG00000173369	C1QB	complement component 1, q subcomponent, B chain
5	ENSG00000159189	C1QC	complement component 1, q subcomponent, C chain
6	ENSG00000006075	CCL3	chemokine (C-C motif) ligand 3
7	ENSG00000170458	CD14	CD14 molecule
8	ENSG00000177575	CD163	CD163 molecule
9	ENSG00000129226	CD68	CD68 molecule
10	ENSG00000019582	CD74	CD74 molecule, major histocompatibility complex, class II invariant chain
11	ENSG00000106066	CPVL	carboxypeptidase, vitellogenic-like
12	ENSG00000182578	CSF1R	colony stimulating factor 1 receptor
13	ENSG00000101439	CST3	cystatin C
14	ENSG00000117984	CTSD	cathepsin D
15	ENSG00000163131	CTSS	cathepsin S
16	ENSG00000161921	CXCL16	chemokine (C-X-C motif) ligand 16
17	ENSG00000188820	FAM26F	family with sequence similarity 26, member F
18	ENSG00000158869	FCER1G	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide
19	ENSG00000143226	FCGR2A	Fc fragment of IgG, low affinity IIa, receptor (CD32)
20	ENSG00000203747	FCGR3A	Fc fragment of IgG, low affinity IIIa, receptor (CD16a)
21	ENSG00000104870	FCGRT	Fc fragment of IgG, receptor, transporter, alpha
22	ENSG00000127951	FGL2	fibrinogen-like 2
23	ENSG00000171051	FPR1	formyl peptide receptor 1
24	ENSG00000087086	FTL	ferritin, light polypeptide
25	ENSG00000135821	GLUL	glutamate-ammonia ligase
26	ENSG00000233276	GPX1	glutathione peroxidase 1
27	ENSG00000030582	GRN	granulin
28	ENSG00000101336	HCK	hemopoietic cell kinase
29	ENSG00000242685	HLA-DMA	major histocompatibility complex, class II, DM alpha
30	ENSG00000241674	HLA-DMB	major histocompatibility complex, class II, DM beta
31	ENSG00000168384	HLA-DPA1	major histocompatibility complex, class II, DP alpha 1
32	ENSG00000230763	HLA-DPB1	major histocompatibility complex, class II, DP beta 1
33	ENSG00000206305	HLA-DQA1	major histocompatibility complex, class II, DQ alpha 1
34	ENSG00000206302	HLA-DQB1	major histocompatibility complex, class II, DQ beta 1
35	ENSG00000206308	HLA-DRA	major histocompatibility complex, class II, DR alpha
36	ENSG00000206306	HLA-DRB1	major histocompatibility complex, class II, DR beta 1
37	ENSG00000198502	HLA-DRB5	major histocompatibility complex, class II, DR beta 5
38	ENSG00000227231	IER3	immediate early response 3
39	ENSG00000216490	IFI30	interferon, gamma-inducible protein 30
40	ENSG00000140749	IGSF6	immunoglobulin superfamily, member 6
41	ENSG00000104951	IL4I1	interleukin 4 induced 1
42	ENSG00000100600	LGMN	legumain
43	ENSG00000090382	LYZ	lysozyme
44	ENSG00000110079	MS4A4A	membrane-spanning 4-domains, subfamily A, member 4A
45	ENSG00000110077	MS4A6A	membrane-spanning 4-domains, subfamily A, member 6A
46	ENSG00000131669	NINJ1	ninjurin 1
47	ENSG00000119655	NPC2	Niemann-Pick disease, type C2
48	ENSG00000085514	PILRA	paired immunoglobin-like type 2 receptor alpha
49	ENSG00000011422	PLAUR	plasminogen activator, urokinase receptor
50	ENSG00000197746	PSAP	prosaposin
51	ENSG00000163191	S100A11	S100 calcium binding protein A11
52	ENSG00000163220	S100A9	S100 calcium binding protein A9
53	ENSG00000130066	SAT1	spermidine/spermine N1-acetyltransferase 1
54	ENSG00000197249	SERPINA1	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 1
55	ENSG00000112096	SOD2	superoxide dismutase 2, mitochondrial
56	ENSG00000120708	TGFBI	transforming growth factor, beta-induced, 68kDa
57	ENSG00000106565	TMEM176B	transmembrane protein 176B
58	ENSG00000011600	TYROBP	TYRO protein tyrosine kinase binding protein
59	ENSG00000155659	VSIG4	V-set and immunoglobulin domain containing 4

Tabela 14 – Assinatura a partir dos marcadores (expressão diferencial) para as células malignas.

1	ENSG00000175899	A2M	alpha-2-macroglobulin
2	ENSG00000182287	AP1S2	adaptor-related protein complex 1, sigma 2 subunit
3	ENSG00000189058	APOD	apolipoprotein D
4	ENSG00000163399	ATP1A1	ATPase, Na+/K+ transporting, alpha 1 polypeptide
5	ENSG00000092529	CAPN3	calpain 3, (p94)
6	ENSG00000135404	CD63	CD63 molecule
7	ENSG00000245694	CRNDE	colorectal neoplasia differentially expressed (non-protein coding)
8	ENSG00000109846	CRYAB	crystallin, alpha B
9	ENSG00000164733	CTSB	cathepsin B
10	ENSG00000130513	GDF15	growth differentiation factor 15
11	ENSG00000136235	GPNMB	glycoprotein (transmembrane) nmb
12	ENSG00000148180	GSN	gelsolin
13	ENSG00000084207	GSTP1	glutathione S-transferase pi 1
14	ENSG00000165949	IFI27	interferon, alpha-inducible protein 27
15	ENSG00000100097	LGALS1	lectin, galactoside-binding, soluble, 1
16	ENSG00000131981	LGALS3	lectin, galactoside-binding, soluble, 3
17	ENSG00000108679	LGALS3BP	lectin, galactoside-binding, soluble, 3 binding protein
18	ENSG00000223414	LINC00473	long intergenic non-protein coding RNA 473
19	ENSG00000160789	LMNA	lamin A/C
20	ENSG00000140545	MFGE8	milk fat globule-EGF factor 8 protein
21	ENSG00000261857	MIA	melanoma inhibitory activity
22	ENSG00000120215	MLANA	melan-A
23	ENSG00000125148	MT2A	metallothionein 2A
24	ENSG00000123560	PLP1	proteolipid protein 1
25	ENSG00000185664	PMEL	premelanosome protein
26	ENSG00000185686	PRAME	preferentially expressed antigen in melanoma
27	ENSG00000197956	S100A6	S100 calcium binding protein A6
28	ENSG00000160307	S100B	S100 calcium binding protein B
29	ENSG00000012171	SEMA3B	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3B
30	ENSG00000196136	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3
31	ENSG00000135919	SERPINE2	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2
32	ENSG00000163597	SNHG16	small nucleolar RNA host gene 16 (non-protein coding)
33	ENSG00000102265	TIMP1	TIMP metallopeptidase inhibitor 1
34	ENSG00000077498	TYR	tyrosinase
35	ENSG00000108828	VAT1	vesicle amine transport 1
36	ENSG00000026025	VIM	vimentin

Tabela 15 – Assinatura a partir dos marcadores (expressão diferencial) para as células malignas.

1	ENSG00000006075	CCL3	chemokine (C-C motif) ligand 3
2	ENSG00000129277	CCL4	chemokine (C-C motif) ligand 4
3	ENSG00000198821	CD247	CD247 molecule
4	ENSG00000172543	CTSW	cathepsin W
5	ENSG00000158869	FCER1G	Fc fragment of IgE, high affinity I, receptor for; gamma polypeptide
6	ENSG00000203747	FCGR3A	Fc fragment of IgG, low affinity IIIa, receptor (CD16a)
7	ENSG00000115523	GNLY	granulysin
8	ENSG00000100453	GZMB	granzyme B
9	ENSG00000171476	HOPX	HOP homeobox
10	ENSG00000100385	IL2RB	interleukin 2 receptor, beta
11	ENSG00000111796	KLRB1	killer cell lectin-like receptor subfamily B, member 1
12	ENSG00000134545	KLRC1	killer cell lectin-like receptor subfamily C, member 1
13	ENSG00000150045	KLRF1	killer cell lectin-like receptor subfamily F, member 1
14	ENSG00000105374	NKG7	natural killer cell group 7 sequence
15	ENSG00000180644	PRF1	perforin 1 (pore forming protein)
16	ENSG00000198574	SH2D1B	SH2 domain containing 1B
17	ENSG00000011600	TYROBP	TYRO protein tyrosine kinase binding protein

Tabela 16 – Assinatura a partir dos marcadores (expressão diferencial) para as células T.

1	ENSG00000161570	CCL5	chemokine (C-C motif) ligand 5
2	ENSG00000116824	CD2	CD2 molecule
3	ENSG00000139193	CD27	CD27 molecule
4	ENSG00000167286	CD3D	CD3d molecule, delta (CD3-TCR complex)
5	ENSG00000160654	CD3G	CD3g molecule, gamma (CD3-TCR complex)
6	ENSG00000153563	CD8A	CD8a molecule
7	ENSG00000077984	CST7	cystatin F (leukocystatin)
8	ENSG00000158050	DUSP2	dual specificity phosphatase 2
9	ENSG00000147168	IL2RG	interleukin 2 receptor, gamma
10	ENSG00000008517	IL32	interleukin 32
11	ENSG00000078596	ITM2A	integral membrane protein 2A
12	ENSG00000213809	KLRK1	killer cell lectin-like receptor subfamily K, member 1
13	ENSG00000182866	LCK	lymphocyte-specific protein tyrosine kinase
14	ENSG00000105374	NKG7	natural killer cell group 7 sequence
15	ENSG00000090104	RGS1	regulator of G-protein signaling 1
16	ENSG00000089012	SIRPG	signal-regulatory protein gamma
17	ENSG00000181847	TIGIT	T cell immunoreceptor with Ig and ITIM domains

Tabela 17 – Assinatura a partir dos marcadores (expressão diferencial) para populações *EMT*.

1	ENSG00000184254	ALDH1A3	aldehyde dehydrogenase 1 family, member A3
2	ENSG00000167772	ANGPTL4	angiopoietin-like 4
3	ENSG00000240583	AQP1	aquaporin 1 (Colton blood group)
4	ENSG00000184324	CSAG2	CSAG family, member 2
5	ENSG00000101439	CST3	cystatin C
6	ENSG00000115414	FN1	fibronectin 1
7	ENSG00000228630	HOTAIR	HOX transcript antisense RNA
8	ENSG00000215533	LINC00189	long intergenic non-protein coding RNA 189
9	ENSG00000215417	MIR17HG	miR-17-92 cluster host gene (non-protein coding)
10	ENSG00000171421	MRPL36	mitochondrial ribosomal protein L36
11	ENSG00000187193	MT1X	metallothionein 1X
12	ENSG00000125148	MT2A	metallothionein 2A
13	ENSG00000249915	PDCD6	programmed cell death 6
14	ENSG00000004799	PDK4	pyruvate dehydrogenase kinase, isozyme 4
15	ENSG00000170955	PRKCDBP	protein kinase C, delta binding protein
16	ENSG00000143248	RGS5	regulator of G-protein signaling 5
17	ENSG00000173432	SAA1	serum amyloid A1
18	ENSG00000134339	SAA2	serum amyloid A2
19	ENSG00000196136	SERPINA3	serpin peptidase inhibitor, clade A (alpha-1 antiproteinase, antitrypsin), member 3
20	ENSG00000104332	SFRP1	secreted frizzled-related protein 1
21	ENSG00000113140	SPARC	secreted protein, acidic, cysteine-rich (osteonectin)
22	ENSG00000091513	TF	transferrin

Tabela 18 – Assinatura a partir da correlação entre *HOTAIR* e *TRHDE-AS1* para populações *EMT*.

1	ENSG00000141338	ABCA8	ATP-binding cassette, sub-family A (ABC1), member 8
2	ENSG00000168306	ACOX2	acyl-CoA oxidase 2, branched chain
3	ENSG00000078295	ADCY2	adenylate cyclase 2 (brain)
4	ENSG00000170425	ADORA2B	adenosine A2b receptor
5	-	BAGE4	BAGE Family Member 4
6	ENSG00000082196	C1QTNF3	C1q and tumor necrosis factor related protein 3
7	ENSG00000182600	C2orf82	chromosome 2 open reading frame 82
8	ENSG00000164047	CAMP	cathelicidin antimicrobial peptide
9	ENSG00000047457	CP	ceruloplasmin (ferroxidase)
10	ENSG00000186377	CYP4X1	cytochrome P450, family 4, subfamily X, polypeptide 1
11	ENSG00000122176	FMOD	fibromodulin
12	ENSG00000151834	GABRA2	gamma-aminobutyric acid (GABA) A receptor, alpha 2
13	ENSG00000170775	GPR37	G protein-coupled receptor 37 (endothelin receptor type B-like)
14	ENSG00000145681	HAPLN1	hyaluronan and proteoglycan link protein 1
15	ENSG00000228630	HOTAIR	HOX transcript antisense RNA
16	ENSG00000132470	ITGB4	integrin, beta 4
17	ENSG00000166159	LRTM2	leucine-rich repeats and transmembrane domains 2
18	ENSG00000147381	MAGEA4	melanoma antigen family A, 4
19	ENSG00000117983	MUC5B	mucin 5B, oligomeric mucus/gel-forming
20	ENSG00000066248	NGEF	neuronal guanine nucleotide exchange factor
21	ENSG00000234068	PAGE2	P antigen family, member 2 (prostate associated)
22	ENSG00000112852	PCDHB2	protocadherin beta 2
23	ENSG00000196604	POTEF	POTE ankyrin domain family, member F
24	ENSG00000131771	PPP1R1B	protein phosphatase 1, regulatory (inhibitor) subunit 1B
25	ENSG00000088320	REM1	RAS (RAD and GEM)-like GTP-binding 1
26	ENSG00000143248	RGS5	regulator of G-protein signaling 5
27	ENSG00000104332	SFRP1	secreted frizzled-related protein 1
28	ENSG00000196542	SPTSSB	serine palmitoyltransferase, small subunit B
29	ENSG00000164362	TERT	telomerase reverse transcriptase
30	ENSG00000091513	TF	transferrin
31	ENSG00000008196	TFAP2B	transcription factor AP-2 beta (activating enhancer binding protein 2 beta)
32	ENSG00000087510	TFAP2C	transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma)
33	ENSG00000164761	TNFRSF11B	tumor necrosis factor receptor superfamily, member 11b
34	ENSG00000236333	TRHDE-AS1	TRHDE antisense RNA 1

Apêndice B – Artigo: CeTF: an R/Bioconductor package for transcription factor co expression networks using regulatory impact factors (RIF) and partial correlation and information (PCIT) analysis

Biagi *et al. BMC Genomics* (2021) 22:624
https://doi.org/10.1186/s12864-021-07918-2

BMC Genomics

SOFTWARE Open Access

CeTF: an R/Bioconductor package for transcription factor co-expression networks using regulatory impact factors (RIF) and partial correlation and information (PCIT) analysis

Carlos Alberto Oliveira de Biagi Jr[1,2,3], Ricardo Perecin Nociti[2,4], Danielle Barbosa Brotto[1,2], Breno Osvaldo Funicheli[2], Patrícia de Cássia Ruy[2,5], João Paulo Bianchi Ximenez[2], David Livingstone Alves Figueiredo[3,6] and Wilson Araújo Silva Jr[1,2,3,7*]

Abstract

Background: Finding meaningful gene-gene interaction and the main Transcription Factors (TFs) in co-expression networks is one of the most important challenges in gene expression data mining.

Results: Here, we developed the R package "CeTF" that integrates the Partial Correlation with Information Theory (PCIT) and Regulatory Impact Factors (RIF) algorithms applied to gene expression data from microarray, RNA-seq, or single-cell RNA-seq platforms. This approach allows identifying the transcription factors most likely to regulate a given network in different biological systems — for example, regulation of gene pathways in tumor stromal cells and tumor cells of the same tumor. This pipeline can be easily integrated into the high-throughput analysis. To demonstrate the CeTF package application, we analyzed gastric cancer RNA-seq data obtained from TCGA (The Cancer Genome Atlas) and found the HOXB3 gene as the second most relevant TFs with a high regulatory impact (TFs-HRi) regulating gene pathways in the cell cycle.

Conclusion: This preliminary finding shows the potential of CeTF to list master regulators of gene networks. CeTF was designed as a user-friendly tool that provides many highly automated functions without requiring the user to perform many complicated processes. It is available on Bioconductor (http://bioconductor.org/packages/CeTF) and GitHub (http://github.com/cbiagii/CeTF).

Keywords: Bioinformatics, R package, R, Transcript factors, Network

*Correspondence: wilsonjr@usp.br
[1]Department of Genetics at Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil
[2]Center for Cell-Based Therapy (CEPID/FAPESP), National Institute of Science and Technology in Stem Cell and Cell Therapy (INCTC/CNPq), Regional Blood Center of Ribeirão Preto, Ribeirão Preto, Brazil
Full list of author information is available at the end of the article

Background

Transcriptome analysis has become crucial to identify gene circuits involved in regulating cancer hallmarks [1]. One of the intelligent ways to explore this type of data and obtain biologically relevant information about the mechanisms involved in modulating gene circuits is the inference of gene regulatory networks (GRNs). Conceptually, we can define GRN as the reconstruction of gene networks from gene expression data, revealing the connection of transcription factors (TFs) with their targets [2], aiming to highlight which gene interactions are the most relevant to the study. Despite the plethora of tools, new methods are needed to assess all possible interactions and their significance [3]. Besides, the presence of TFs in interactions for gene-to-gene is functionally crucial because they may be playing an essential regulatory role in biological processes [4]. TFs are considered key molecules that can regulate the expression of one or more genes in a biological system, thus determining how cells function and communicate with cellular environments [5]. Furthermore, integrating genome-scale and network generation with the identification of main TFs provides new insights into their data function. In this article, we provide an R package that enables performing the Regulatory Impact Factors (RIF) and Partial Correlation with Information Theory (PCIT) analysis separately, or by applying the full pipeline.

We, therefore, developed an R package called CeTF, which would not only apply the RIF and PCIT analysis, but would also perform network diffusion analysis, generate circos plots for specifics TFs/genes, functional enrichment for network conditions, and others features. The biggest advantage is that the package is intuitive to use, and the main functions are written in C/C++, which provides faster analysis for large data.

Implementation

CeTF is an C/C++ implementation in R for PCIT [6] and RIF [7] algorithms, which initially were made in FORTRAN language. From these two algorithms, it was possible to integrate them in order to increase performance and Results. Input data may come from microarray, RNA-seq, or single-cell RNA-seq. The input data can be read counts or expressions (TPM, FPKM, normalized values, etc.). The main pipeline (Fig. 1) consists of the following steps.

Data adjustment

If the input data is a count table, data will be converted to TPM by each column (x) as follows:

$$TPM = \frac{10^6 x}{sum(x)} \tag{1}$$

The mean for TPM values different than zero and the mean values for each gene are used as a threshold to filter the genes. Genes with values above half of the previous averages will be considered for subsequent analyses. Then, the TPM data is normalized using:

$$Norm = \frac{log(x+1)}{log(2)} \tag{2}$$

If the input already has normalized expression data (TPM, FPKM, etc), the only step will be the same filter for genes that consider half of the means.

Differential expression analysis

There are two options for differential analysis of the gene expression, the Reverter method [8] and DESeq2 [9]. In both methods, two conditions are required (i.e., control *vs.* tumor samples). In the Reverter method, the mean between samples of each condition for each gene is calculated. Then, subtraction is made between the mean of one condition concerning the other conditions. The variance of the subtraction is performed, then is calculated the difference of expression using the following formula, where s is the result of subtraction and var is the variance:

$$diff = \frac{s - \frac{sum(s)}{length(s)}}{\sqrt{var}} \tag{3}$$

The DESeq2 method applies the Differential expression analysis based on the negative binomial distribution. Although both methods can be used on count data, it is strongly recommended to use only the Reverter method on expression input data.

Regulatory impact factors (RIF) analysis

The RIF algorithm is well described in the original paper [7]. This step aims to identify critical Transcription Factors calculating for each condition the co-expression correlation between the TFs and the Differentially Expressed (DE) genes (from previously item). The result is RIF1 and RIF2 metrics that allow the identification of critical TFs. The RIF1 metric classifies the TFs as most differentially co-expressed with the highly abundant and highly DE genes, and the RIF2 metric classifies the TF with the most altered ability to act as predictors of the abundance of DE genes. The main TF is defined if:

$$\sqrt{RIF1^2} \quad or \quad \sqrt{RIF2^2} \quad > 1.96 \tag{4}$$

Partial correlation and information theory (PCIT) analysis

The PCIT algorithm is also well described in the original paper from Reverter and Chan [6]. Moreover, it has been used for the reconstruction of Gene Co-expression Networks (GCN). The GCN combines the concept of the Partial Correlation coefficient with Information Theory to identify significant gene-to-gene associations defining edges in the reconstruction of the network. At this stage, the paired correlation of three genes is performed simultaneously, thus making the inference of co-expressed genes.

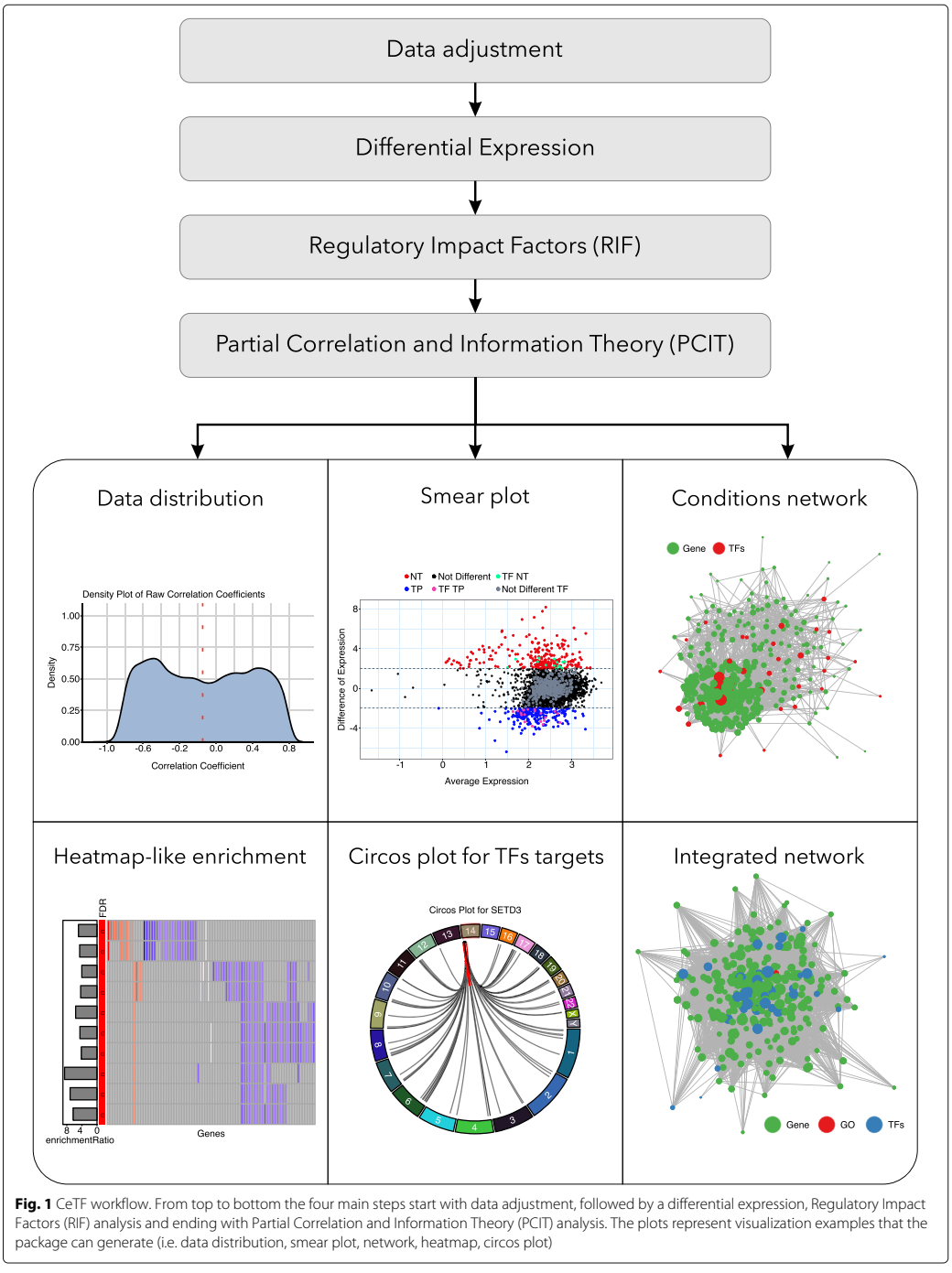

Fig. 1 CeTF workflow. From top to bottom the four main steps start with data adjustment, followed by a differential expression, Regulatory Impact Factors (RIF) analysis and ending with Partial Correlation and Information Theory (PCIT) analysis. The plots represent visualization examples that the package can generate (i.e. data distribution, smear plot, network, heatmap, circos plot)

This approach is more sensitive than other methods and allows the detection of functionally validated gene-gene interactions. First, is calculated for every trio of genes x, y, and z the partial correlation coefficients:

$$r_{xy,z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \quad (5)$$

And similarly, for $r_{xz,y}$ and $r_{yz,x}$. After that, for each trio of genes is calculated the tolerance level (ε) to be used as a threshold for capturing significant associations. The average ratio of partial to direct correlation is computed as follows:

$$\varepsilon = \frac{1}{3}\left(\frac{r_{xy,z}}{r_{xy}} + \frac{r_{xz,y}}{r_{xz}} + \frac{r_{yz,x}}{r_{yz}}\right) \quad (6)$$

The association between the genes x and y is discarded if:

$$|r_{xz}| \leq |\varepsilon r_{xz}| \quad and \quad |r_{xy}| \leq |\varepsilon r_{yz}| \quad (7)$$

Otherwise, the association is defined as significant, and the interaction between the genes x and y is used in the reconstruction of the GCN. The final output includes the network with gene-gene and gene-TF interactions for both conditions, besides generating the main TFs identified in the network.

Functions of the package

There are 28 functions and 5 example datasets available in CeTF, which are described in Table 1. A working example for each of these functions is given in the package documentation in the Supplementary Material. The package allows the integration with many other packages and different types of genomics/transcriptomics analysis.

Additional functionalities

The CeTF package also includes additional features in order to visualize the results. After running PCIT and RIF analysis, it is possible to plot the data distribution, the distribution of differentially expressed genes/TFs that shows the average expression (in log2) by the difference of expression, the network for both conditions and the integrated network with genes, TFs and enriched pathways. Besides, it is possible to visualize the targets for specific TFs as a circos plot. It is also possible to perform the grouping of ontologies [10] without statistical inference and functional enrichment for several databases with the statistical inference of many organisms using WebGestalt database [11]. Finally, it is possible to save all tables that include interaction networks, enrichment, differential expression, main TFs, and others.

Table 1 Functions available in CeTF

Function	Description
bivar.awk	Summary statistics from two variables
CircosTargets	Circos plot for the Transcription Factors/genes targets
clustCoef	Calculate the clustering coefficient
clustCoefPercentage	Calculate the clustering coefficient as a percentage
densityPlot	Density distribution of correlation coefficients and significant PCIT values
diffusion	Network diffusion analysis
enrichdemo	Enrichment data
enrichPlot	Plots to visualize the enrichment analysis results
expDiff	Differential expression analysis
getData	Data accessor for a CeTF class object
getDE	Differential Expression accessor for a CeTF class object
getEnrich	Enrichment analysis for genes of network
getGroupGO	Functional Profile of a gene set at specific GO level
heatPlot	Heatmap-like functional classification
histPlot	Histogram of connectivity distribution
InputData	Input data accessor for a CeTF class object
netConditionsPlot	Network plot of gene-gene/gene-TFs interactions
netGOTFPlot	Plot a network for Ontologies, genes and TFs
NetworkData	Networks data accessor for a CeTF class object
normExp	Normalized expression transformation
OutputData	Output data accessor for a CeTF class object
PCIT	Partial Correlation and Information Theory (PCIT) analysis
pcitC	A helper to calculate PCIT implemented in C/C++
refGenes	List of reference genes for 5 different organisms to perform enrichment
RIF_input	Regulatory Impact Factors (RIF) input
RIF	Regulatory Impact Factors (RIF) analysis
RIFPlot	Relationship plots between RIF1, RIF2 and DE genes
runAnalysis	Whole analysis of RIF and PCIT
simCounts	Simulated counts data
simNorm	Simulated normalized data
SmearPlot	Smear plot for Differentially Expressed genes and TFs
TFs	Transcripition Factors data
Tolerance	Tolerance level between 3 pairwise correlations implemented in C/C++

Software construction

CeTF is an R-based toolkit, and most of the code is written in R language. PCIT and tolerance functions were written in C/C++ using Rcpp (v1.0.5) [12] and RcppArmadillo (v0.10.1.2.2) [13] for better performance. The main R packages used for analysis and visualization of the results were the circlize (v0.4.10) [14], ComplexHeatmap (v2.6.0) [15], DESeq2 (v1.30.0) [9], ggplot2 (v3.3.2) [16], RCy3 (v2.10.0) [17], and others listed in the Supplementary Material.

Results

To demonstrate the tool's utility, we used stomach adenocarcinoma RNA-seq data from The Cancer Genome Atlas (TCGA) project [18] and applied all analyzes available in the CeTF package. Here, we compared samples from normal tissue (NT=36) and primary tumor (PT=408) of Stomach adenocarcinoma (STAD). The TFs-HRi are shown in Table 2 and the analysis of partial results in Fig. 2A.

Table 2 describes a list of 37 TFs-HRi. Among the main TFs-HRi identified, we highlight four TFs (SETD3, HOX3B, FOXA1, and SOX4) for being widely reported in association with stomach adenocarcinoma. Some studies show that high expression of the SETD3 gene is associated with poor survival in triple-negative breast cancer [19], while HOXB3 and FOXA1 were identified as indicators of better prognosis [20–22]. Interestingly, the elevated expression of the SOX4 gene has been described to regulate the epithelial-mesenchymal transition (EMT) mechanism mediated by TGF-beta [23]. The Results presented below will be centered on the HOXB3 gene, as it is one of the HOX genes studied by our group [24, 25].

After filtering data, a total of 8,037 genes remained in the analysis and are represented in Fig. 2A, with 151 up-regulated genes (red dots) and 118 down-regulated genes (blue dots). On this set of genes, 7 TFs are up-regulated (green dots), 9 TFs are down-regulated (pink dots) and 504 are not differentially expressed. Figure 2B places the HOXB3 gene as a central hub and its 2520 gene-to-gene interactions obtained with the CeTF package. Seventy-six up-regulated targets, and 58 down-regulated targets were found.

Figure 2C shows the heatmap with all 163 HOXB3 targets, which revealed no correlation with the two main groups of samples with clinical and histopathological data. A graph with the enrichment of gene pathways only with HOXB3 targets (Fig. 2D) shows that only one biological process (muscle system process) was enriched with overexpressed HOXB3 targets. Nine other biological processes were enriched with downregulated targets associated with the cell cycle, corroborating with the biology of normal tissues (Fig. 2D). Furthermore, the Chip-seq data from one of our studies (unpublished data) were used to

Table 2 List of TFs-HRi from TCGA-STAD analysis. Here we have the Transcript Factors (*TF*) found as playing an important role in the given comparison. Also shown is the mean of expression (*avgexpr*) for each TF, in addition to the values of the metrics RIF1 and RIF2. Finally, *freq.NT* and *frep.TP* columns represent the frequency of appearance of the given TF in each condition, with *freq.diff* being the difference between these frequencies. A positive difference means that TF plays an important role in the reference condition in the NT case, whereas a negative difference means that TF plays an important role in the condition TP

TF	avgexpr	RIF1	RIF2	freq.NT	freq.TP	freq.diff
SETD3	5.854	1.409	2.189	162	13	149
HOXB3	4.309	0.517	2.282	159	14	145
RNF115	4.96	-2.324	1.64	153	19	134
TOX4	6.183	2.345	1.63	139	9	130
ASCL2	3.96	2.179	0.678	147	18	129
FOXA1	5.597	-0.801	2.022	159	34	125
SOX4	7.281	3.554	1.072	149	29	120
CSDE1	8.816	-0.069	2.153	172	53	119
TEAD3	5.903	-0.225	2.031	157	46	111
VEZF1	6.211	-0.385	2.243	157	47	110
TERF1	4.853	-2.475	0.902	123	17	106
RBBP7	7.086	2.393	1.872	147	42	105
BBX	6.22	-0.314	2.09	154	55	99
ECD	5.17	3.16	0.778	115	20	95
SPDEF	3.749	-2.081	1.078	114	20	94
TULP3	5.059	0.698	2.012	152	58	94
TRIM16	5.74	-2.266	0.721	127	35	92
ZBTB7C	3.999	-3.093	0.824	122	30	92
NFX1	5.733	3.149	0.852	96	13	83
TP53	6.305	-2.016	-0.005	89	8	81
NFE2L3	6.01	2.484	1.068	175	112	63
TSC22D4	5.989	-1.976	-0.106	72	9	63
AFF4	7.147	2.486	0.539	89	27	62
ELF1	6.758	-2.384	-0.09	76	16	60
VTN	1.905	-2.277	0.02	66	13	53
ADNP2	5.251	2.319	0.311	79	29	50
KLF4	6.519	-3.313	-0.297	73	24	49
CDC5L	5.794	2.845	-0.058	69	36	33
KLF6	7.737	-2.584	-1.222	31	10	21
PER1	5.694	2.051	0.866	127	115	12
MYC	7.064	2.127	-0.714	35	29	6
LYAR	4.775	2.242	-1.301	45	75	-30
HMGB2	6.67	-0.737	-2.214	3	66	-63
MAFB	5.29	-2.433	-1.844	32	95	-63
E2F3	4.673	0.238	-2.131	7	80	-73
SSRP1	7.323	1.431	-2.081	44	128	-84
MAF	5.527	0.495	-2.282	20	124	-104

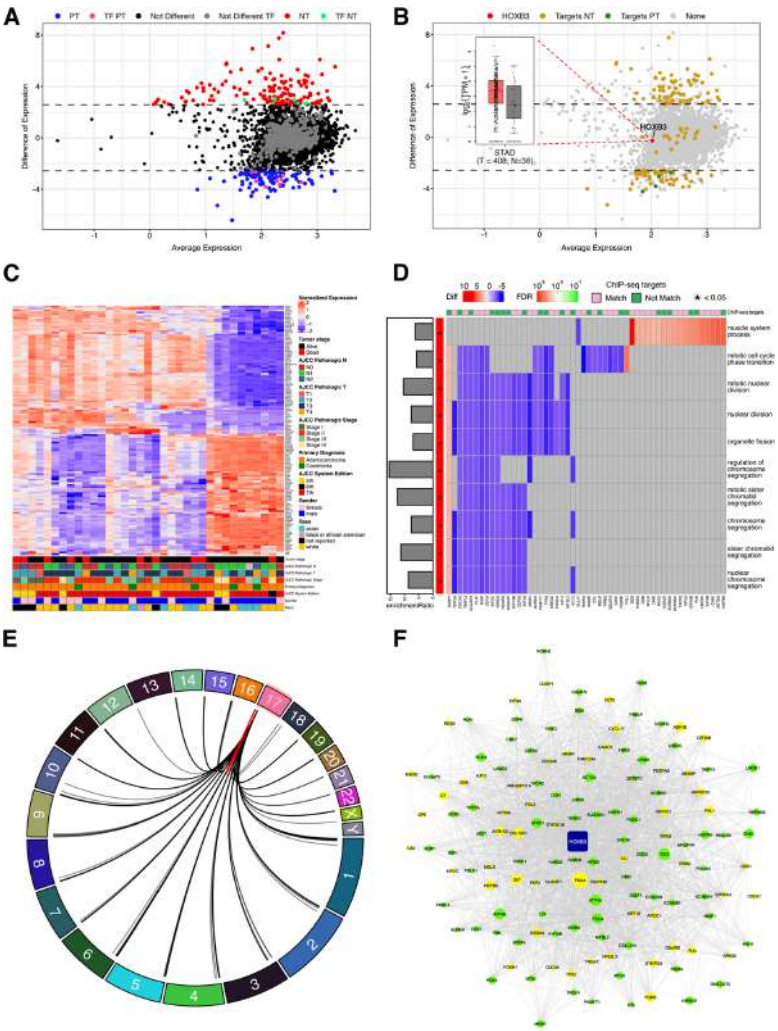

Fig. 2 TCGA-STAD data results comparing normal versus tumor samples using CeTF. (A) Smear plot showing the difference of expression for 8,037 genes, which 151 are up-regulated (colored in red), 118 are down-regulated (blue color), and the dots in black color are not differentially expressed based in a difference of expression module cutoff of 2.57. There are 7 TFs up-regulated (green color), 9 TFs down-regulated (pink color), and 504 not differentially expressed TFs (grey color). (B) Smear plot showing the 163 HOXB3 targets. Of these, 76 are up-regulated, 58 are down-regulated, and 29 are not differentially expressed. The yellow dots represent the 149 targets associated with NT samples, and the green dots represent the 4 targets associated with TP samples. (C) Heatmap with 163 HOXB3 targets in NT samples. The bottom annotation has clinical information as tumor stage, AJCC pathologic N, AJCC pathologic T, AJCC pathologic stage, primary diagnosis, AJCC system edition, gender, and race. (D) Enrichment of 163 HOXB3 targets with Gene Ontology Biological Process showing which genes are enriched with the pathways and their expression difference. The bar plot on the left side shows the enrichment ratio. The left sidebar shows the enriched pathway significance with an asterisk if significant, a p-value less than 0.05. Finally, the top annotation shows the match between HOXB3 targets from CeTF and ChIP-seq. (E) Circos plot representing the HOXB3 targets and their chromosome position. HOXB3 is located on chromosome 17. The red line shows the 10 cis interactions (the target is located at the same chromosome HOXB3), and the black lines indicate a trans interaction (the target is locatedon a different chromosome than HOXB3). (F) Network with 134 down and up-regulated HOXB3 targets. The network has 135 nodes and 2520 edges. HOXB3 is represented in the center of the network in blue color. The green nodes represent the 79 targets found in CeTF that match with ChIP-seq targets for HOXB3 and the yellow nodes represent the 55 targets that don't fit with them

validate the 163 targets predicted. Although the CHIP-seq data were generated from placental tissue, 54% of the targets predicted by the CeTF package have been validated (Fig. 2D). In addition to the negative control of the cell cycle, the DUSP1 gene, which is upregulated in all cell cycle biological processes, is related to the negative regulation of cellular proliferation [26]. A representation of the genomic distribution of the HOXB3 targets (located on chromosome 17) shows that the vast majority of targets are in different chromosomes. Ten targets are located on chromosome 17 (Fig. 2E). Finally, we built the network for HOXB3 and their targets (Fig. 2F). The targets validated by Chip-seq are highlighted in green color.

Conclusions

CeTF is a tool that assists the identification of meaningful gene-gene associations and the main TFs in co-expression networks, as demonstrated previously. It offers functions for a complete and customizable workflow from count or expression data to networks and visualizations in a freely available R package. We expect that CeTF will be widely used by the genomics and transcriptomics community and scientists who work with high-throughput data to understand how main TFs are working in a co-expression network and what are the pathways involved in this context. We employ RNA-seq data of stomach adenocarcinoma from the TCGA project to demonstrate all the CeTF package analyses. We believe that the present study will help researchers either identify transcription factors with a critical role in regulating gene pathways involved with tumorigenesis or other biological systems of interest.

Availability and requirements

Project name: CeTF
Project home page: http://bioconductor.org/packages/CeTF and http://github.com/cbiagii/CeTF
Operating system: platform independent
Programming language: R
Other requirements: R 4.0 or higher
License: GPL-3
Any restrictions to use by non-academics: no licence needed

Abbreviations

CeTF: Coexpression for Transcription Factors; RIF: Regulatory Impact Factors; PCIT: Partial Correlation with Information Theory; TFs: Transcription Factors; TCGA: The Cancer Genome Atlas; TFs-HRi: Transcription Factors with a High Regulatory impact; GRNs: Gene Regulatory Networks; TPM: Transcripts Per Million; FPKM: Fragments Per Kilobase Million; DE: Differentially Expressed; STAD: Stomach adenocarcinoma; EMT: Epithelial-Mesenchymal Transition

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07918-2.

Additional file 1: Detailed tutorial for CeTF package. This file is an tutorial showing step-by-step how to use CeTF package.

Acknowledgements

We thank Regional Blood Center of Ribeirão Preto for all support. We appreciate Bioconductor reviewers and the GitHub community for constructive discussions about software usage, statistics and better performance.

Authors' contributions

CAOBJ implemented the package and drafted the manuscript. RPN and BOF made substantial contributions to the bioinformatics analysis, manuscript draft and editing. CAOBJ, RPN,BOF, PCR, JPBX, DBB, and DLAF participated in study design and result interpretation. WASJ supervised the project and critically revised the manuscript. All authors read and approved the final manuscript.

Funding

(CAPES), grant #88882.378695/2019-01; São Paulo Research Foundation (FAPESP), #2013/08135-2, and by Research Support of the University of Sao Paulo, CISBi-NAP/USP Grant #12.1.25441.01.2.

Availability of data and materials

CeTF is a publicly available Bioconductor package available from http://bioconductor.org/packages/CeTF. Documentation is available on the Bioconductor website, and we provide vignettes describing more example analyses. We also maintain a public github repository (http://github.com/cbiagii/CeTF), and invite the community to submit or request additional functionality to incorporate into this package. This package requires R ≥4.0.0 and depends on several R/Bioconductor packages including circlize, ComplexHeatmap, clusterProfiler, DESeq2, GenomicTools, GenomicTools.fileHandler, ggnetwork, GGally, ggplot2, ggpubr, ggrepel, graphics, grid, igraph, Matrix, network, Rcpp, RCy3, S4Vectors, stats, SummarizedExperiment, utils and WebGestaltR. A web page is also available with tutorials and additional information: http://cbiagii.github.io/CeTF/. A docker image with the latest version is available in https://hub.docker.com/r/biagii/cetf.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Genetics at Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil. [2]Center for Cell-Based Therapy (CEPID/FAPESP), National Institute of Science and Technology in Stem Cell and Cell Therapy (INCTC/CNPq), Regional Blood Center of Ribeirão Preto, Ribeirão Preto, Brazil. [3]Institute for Cancer Research, IPEC, Guarapuava, Brazil. [4]Laboratory of Molecular Morphophysiology and Development, Department of Veterinary Medicine, Faculty of Animal Science and Food Engineering, University of São Paulo, Pirassununga, Brazil. [5]Center for Medical Genomics, HCFMRP/USP, Ribeirão Preto, Brazil. [6]Department of Medicine, Midwest State University of Paraná-UNICENTRO, Guarapuava, Brazil. [7]Center for Integrative Systems Biology (CISBi) - NAP/USP, University of São Paulo, Ribeirão Preto, Brazil.

Received: 5 January 2021 Accepted: 30 July 2021
Published online: 20 August 2021

References

1. Hanahan D, Weinberg R. Hallmarks of cancer: the next generation. cell. 2011;144(5):646–74.
2. Hu X, Hu Y, Wu F, Leung RWT, Qin J. Integration of single-cell multi-omics for gene regulatory network inference. Comput Struct Biotechnol J. 2020;18:1925–38.
3. Yu D, Kim M, Xiao G, Hwang T. Review of biological network data and its applications. Genomics Inform. 2013;11(4):200.
4. Farnham P. Insights from genomic profiling of transcription factors. Nat Rev Genet. 2009;10(9):605–16.

5. Vaquerizas J, Kummerfeld S, Teichmann S, Luscombe N. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10(4):252–63.
6. Reverter A, Chan E. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics. 2008;24(21):2491–7.
7. Reverter A, Hudson N, Nagaraj S, Pérez-Enciso M, Dalrymple B. Regulatory impact factors: unraveling the transcriptional regulation of complex traits from expression data. Bioinformatics. 2010;26(7):896–904.
8. Reverter A, Ingham A, Lehnert S, Tan S-H, Wang Y, Ratnakumar A, Dalrymple B. Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer. Bioinformatics. 2006;22(19):2396–404.
9. Love M, Huber W, Anders S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. Genome Biol. 2014;15(12):550.
10. Consortium G. The gene ontology resource: 20 years and still going strong. Nucleic Acids Res. 2019;47(D1):330–8.
11. Liao Y, Wang J, Jaehnig E, Shi Z, Zhang B. Webgestalt 2019: gene set analysis toolkit with revamped uis and apis. Nucleic Acids Res. 2019;47(W1):199–205.
12. Eddelbuettel D, François R, Allaire J, Ushey K, Kou Q, Russel N, Chambers J, Bates D. Rcpp: Seamless r and c++ integration. J Stat Softw. 2011;40(8):1–18.
13. Eddelbuettel D, Sanderson C. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. Comput Stat Data Anal. 2014;71:1054–63.
14. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in r. Bioinformatics. 2014;30(19):2811–2.
15. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016;32(18):2847–9.
16. Wickham H. Elegant graphics for data analysis (ggplot2); 2009. https://ggplot2-book.org. Accessed 18 Nov 2020.
17. Gustavsen JA, Pai S, Isserlin R, Demchak B, Pico AR. RCy3: network biology using cytoscape from within R. F1000Research. 2019;8:1774.
18. Weinstein J, Collisson E, Mills G, Shaw K, Ozenberger B, Ellrott K, Shmulevich I, Sander C, Stuart J, Network C, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113.
19. Hassan N, Rutsch N, Győrffy B, Espinoza-Sánchez N, Götte M. Setd3 acts as a prognostic marker in breast cancer patients and modulates the viability and invasion of breast cancer cells. Sci Rep. 2020;10(1):1–16.
20. Tomioka N, Morita K, Kobayashi N, Tada M, Itoh T, Saitoh S, Kondo M, Takahashi N, Kataoka A, Nakanishi K, et al. Array comparative genomic hybridization analysis revealed four genomic prognostic biomarkers for primary gastric cancers. Cancer Genet Cytogenet. 2010;201(1):6–14.
21. Ren H, Zhang P, Tang Y, Wu M, Zhang W. Forkhead box protein a1 is a prognostic predictor and promotes tumor growth of gastric cancer. OncoTargets Ther. 2015;8:3029.
22. Camolotto S, Pattabiraman S, Mosbruger T, Jones A, Belova V, Orstad G, Streiff M, Salmond L, Stubben C, Kaestner K, et al. Foxa1 and foxa2 drive gastric differentiation and suppress squamous identity in nkx2-1-negative lung cancer. Elife. 2018;7:38579.
23. Peng X, Liu G, Peng H, Chen A, Zha L, Wang Z. Sox4 contributes to tgf-β-induced epithelial–mesenchymal transition and stem cell characteristics of gastric cancer cells. Genes Dis. 2018;5(1):49–61.
24. Brotto D, Siena ADD, de Barros I, Carvalho SdCeS, Muys B, Goedert L, Cardoso C, Plaça J, Ramão A, Squire J, et al. Contributions of hox genes to cancer hallmarks: Enrichment pathway analysis and review. Tumor Biol. 2020;42(5):1010428320918050.
25. Ramão A, Pinheiro D, Alves C, Kannen V, Jungbluth A, de Araújo LF, Muys B, Fonseca A, Plaça J, Panepucci R, et al. Hox genes: potential candidates for the progression of laryngeal squamous cell carcinoma. Tumor Biol. 2016;37(11):15087–96.
26. Cheng C, Liu F, Li J, Song Q. Dusp1 promotes senescence of retinoblastoma cell line so-rb5 cells by activating akt signaling pathway. Eur Rev Med Pharmacol Sci. 2018;22(22):7628–32.

Publisher's Note