UNIVERSIDADE FEDERAL DE SANTA CATARINA

CENTRO DE CIÊNCIAS DA SAÚDE

DEPARTAMENTO DE MICROBIOLOGIA, IMUNOLOGIA E PARASITOLOGIA

CURSO DE GRADUAÇÃO EM FARMÁCIA

CÁSSIO RAMOS FERNANDES

ANÁLISE DE DADOS DE TRANSCRIPTOMA DE PACIENTES COM LEUCEMIA AGUDA DE FENÓTIPO MISTO.

FLORIANÓPOLIS – SC

2022

CÁSSIO RAMOS FERNANDES

ANÁLISE DE DADOS DE TRANSCRIPTOMA DE PACIENTES COM LEUCEMIA AGUDA DE FENÓTIPO MISTO.

Projeto apresentado na disciplina CIF5351 ao curso de Farmácia da Universidade Federal de Santa Catarina como requisito inicial para elaboração e apresentação do trabalho de conclusão de curso.

Orientador: Prof. Dr. Edroaldo Lummertz da Rocha

FLORIANÓPOLIS – SC

2022

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

Fernandes, Cássio
Análise de Dados de Transcriptoma de Pacientes com Leucemia Aguda de Fenótipo Misto / Cássio Fernandes ; orientador, Edroaldo Lummertz da Rocha, 2021.
31 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro de Ciências da Saúde, Graduação em Farmácia, Florianópolis, 2021.

Inclui referências.

1. Farmácia. 2. Biologia Computacional. 3. scRNA-seq. 4. Transcriptoma. I. Lummertz da Rocha, Edroaldo . II. Universidade Federal de Santa Catarina. Graduação em Farmácia. III. Título.

RESUMO

A Leucemia Aguda de Fenótipo Misto representa uma classe de leucemias altamente heterogêneas e de difícil tratamento. Portanto, uma melhor compreensão dos programas de expressão gênicos alterados em células leucêmicas com relação aos tipos celulares correspondentes na medula óssea saudável pode levar a uma melhor compreensão da patofisiologia da doença. Neste trabalho, foram reanalizados dados publicamente disponíveis da medula óssea e amostras provenientes de amostras leucêmicas com o intuito de determinar a similaridade molecular ou diferenças transcricionais entre as células leucêmicas e saudáveis. Utilizando o software FUSCA e o algoritmos Harmony e Symphony, foi elaborado um atlas de referência que representa a diversidade celular da medula óssea saudável. Então, as células leucêmicas foram mapeadas para este atlas de referência utilizando o algoritmo Symphony, ou classificadas utilizando o algoritmo singleCellNet, levando a conclusão que a uma fração substancial das células leucêmicas são transcricionalmente similares às células-tronco e progenitoras sanguíneas, consistente com a possibilidade de diferenciação multilinhagem das células leucêmicas e explicando, ao menos parcialmente, o fenótipo mielóide/linfóide misto desta classe de leucemias.

Palavras chave: scRNA-seq, biologia computacional, transcriptoma

ABSTRACT

Mixed-phenotype acute leukemia represents a class of highly heterogeneous and difficult to treat leukemias. Therefore, a better understanding of altered gene expression programs in leukemic cells relative to the corresponding cell types in healthy bone marrow may lead to a better understanding of the pathophysiology of the disease. In this work, publicly available data from bone marrow and samples from leukemic samples were reanalyzed in order to determine molecular similarity or transcriptional differences between leukemic and healthy cells. Using the FUSCA and Symphony algorithms, a reference atlas was created that represents the cellular diversity of healthy bone marrow. Then, the leukemic cells were mapped to this reference atlas using the Symphony algorithm, or classified using the singleCellNet algorithm, leading to the conclusion that a substantial fraction of the leukemic cells are transcriptionally similar to blood stem and progenitor cells, consistent with the possibility of multilineage differentiation of leukemic cells and explaining, at least partially, the myeloid/mixed lymphoid phenotype of this class of leukemias.

Keywords: scRNA-seq, computational biology, transcriptome

SUMÁRIO:

1. INTRODUÇÃO:

1.1. O Câncer:

Também conhecido como neoplasia e tumor, o câncer é um grande grupo de doenças que podem se apresentar em quase qualquer órgão ou tecido do corpo, quando células anormais crescem descontroladamente, ultrapassam seus limites habituais para invadir partes adjacentes do corpo e/ou se espalhar para outros órgãos, esse último processo é chamado de metástase e é uma das principais causas de morte por câncer (WHO, 2022).

O primeiro registro da doença é um papiro egípcio do século VII a.C. No entanto, são raras as referências ao câncer antes do século XIX, já que as pessoas morriam de flagelos para os quais existe tratamento hoje, como tuberculose, cólera, varíola, peste ou pneumonia. Com o prolongamento da vida humana, o câncer foi levado para o primeiro plano, tornando-se uma das doença de maior mortalidade da civilização moderna, o que justificou, na segunda metade do século XX, a intensificação da luta da medicina contra esse mal (MUKHERJEE, 2011).

A carga do câncer continua a crescer globalmente, exercendo uma enorme pressão física, emocional e financeira sobre indivíduos, famílias, comunidades e sistemas de saúde. Muitos sistemas de saúde em países de baixa e média renda estão menos preparados para gerenciar esse fardo, e um grande número de pacientes com câncer em todo o mundo não tem acesso a diagnóstico e tratamento de qualidade em tempo hábil.

1.2. Dados e informações sobre Câncer e Leucemias:

Segundo dados recentes, o câncer é uma das principais causas de mortalidade, potencialmente reduzindo a expectativa de vida humana no planeta. De acordo com estimativas da Organização Mundial da Saúde (WHO, 2021), dados do último censo, realizado em 2019, apontam o câncer como a primeira ou a segunda

principal causa de morte de pessoas com idade inferior a 70 anos em 112 de 183 países e ocupa o terceiro ou quarto lugar em outros 23 países (SUNG et al., 2021).

No Brasil, a doença possui uma incidência estimada de 309.750 casos novos em homens e 316.280 em mulheres, com um total de óbitos de 121.686 homens e 110.344 mulheres, de acordo com o levantamento feito em 2020, disponibilizado no site do INCA (Instituto Nacional de Câncer). O que reflete em uma urgente necessidade de explorar novas alternativas para combatê-lo.

Entre todos os tipos de câncer, existe a leucemia, que representa uma coleção heterogênea de neoplasias com origem no sistema sanguíneo, também conhecido como sistema hematopoiético, e nos órgãos formadores de sangue (HAO et al., 2019). Ela é responsável, de acordo com a estimativa do INCA, por um total de 10.810 novos casos no ano de 2020, sendo 5.920 em homens e 4.890 em mulheres. Entre os dados de mortalidade, de acordo com o “Atlas de Mortalidade por Câncer” do ano de 2019, um total de 7.370 pessoas vieram a óbito, sendo 4.014 homens e 3.356 mulheres (INCA, 2022).

Com base na origem do tipo de célula predominante (mieloide ou linfoide) e na taxa de progressão da doença (aguda ou crônica), a leucemia é categorizada em quatro subtipos principais: leucemia mieloide aguda (LMA), leucemia mieloide crônica (LMC), leucemia linfoide aguda (LLA) e leucemia linfoide crônica (LLC) (HAO et al., 2019). Entre alguns desses subtipos existe um outro, caracterizado recentemente, chamado de leucemia aguda de fenótipo misto.

1.3. Leucemia Aguda de Fenótipo Misto (LAFM):

A leucemia aguda de fenótipo misto (LAFM), também conhecida como leucemia aguda de linhagem mista, é um subtipo de leucemia aguda de linhagem ambígua. São formas de leucemias combinadas: leucemia linfoide aguda (LLA) e leucemia mieloide aguda (LMA).

O diagnóstico de LAFM é atribuído a um caso de leucemia aguda que apresenta expressão de uma combinação de antígenos de diferentes linhagens, de

modo que não é possível atribuir uma única linhagem. Uma referência específica se dá através de parâmetros de classificação de categorias de casos por exclusão, seja por características genéticas ou por características clínicas (BOROWITZ, M. J., 2008).

Mas também duas alterações genéticas foram relatadas com frequência em LAFM para agora serem consideradas como entidades separadas. A primeira é LAFM com rearranjo t(9;22)(q34;q11.2) ou *BCR-ABL1*. As características clínicas associadas a essa translocação são semelhantes às de outros pacientes com LAFM, mas é importante não fazer esse diagnóstico em pacientes com LMC. Tal ação colocaria um caso de transformação de blasto de LMC em vez de LAFM. As leucemias positivas para o cromossomo Filadélfia são geralmente mais frequentes em pacientes mais velhos. Embora a maioria dos estudos tenha encontrado a frequência de LAFM com t(9;22) de 28 a 35%, estudos pediátricos relatam que é muito menor em 3% (AL-SERAIHY et al., 2009). Muitos desses casos mostram uma população dimórfica de blastos, com a maioria apresentando linhagem B e mielóide. Essa translocação em leucemias com fenótipo misto foi associada a um pior prognóstico em alguns estudos (KILLICK et al., 1999). A segunda alteração genética mais frequente na LAFM são as translocações envolvendo os genes *MLL* e *AF4* no cromossomo 4, banda q21 (OWAIDAH et al., 2006). Isso tende a ocorrer mais comumente em crianças e é mais frequente na infância (XU et al., 2009). Esses casos também tendem a apresentar uma população blástica dimórfica e os linfoblastos têm um imunofenótipo CD19-positivo, CD10-negativo, B-precursor e são frequentemente positivos para CD15. O prognóstico de pacientes com LAFM com rearranjo *MLL* é ruim (OWAIDAH et al., 2006).

Se os blastos não possuem as anormalidades genéticas mencionadas acima, os casos de LAFM são categorizados pela linhagem dos blastos. Assim, existem três outras categorias menos específicas, que incluem B/mielóide, T/mielóide e outros tipos raros de LAFM listados sem outra especificação. Reconhece-se que os casos com uma combinação de casos de linhagem B e T e casos de trilinhagem são muito raros. Por exemplo, em casos de leucemia de células T, CD79a e CD10 isoladamente não devem ser considerados como evidência de diferenciação de

células B, pois esses marcadores são relativamente comuns na LLA-T (WEINBERG; ARBER, 2010).

1.4. Primeiros casos relatados de LAFM:

Os primeiros estudos publicados sobre leucemia aguda de fenótipo misto ocorreram na década de 1980, quando os anticorpos monoclonais foram usados pela primeira vez para caracterizar células leucêmicas (BENE, 2009). Uma dessas publicações anteriores observou a co-expressão da mieloperoxidase e desoxinucleotidil transferase terminal (TdT) em LMA (MCGRAW et al., 1981). Este e outros relatos sugeriram que a célula de origem deste subtipo de leucemia reside no compartimento das células-tronco e progenitoras hematopoiéticas, as quais são capazes de se diferenciar em células mieloides e linfoides (WEINBERG; ARBER, 2010).

Um dos primeiros grandes casos de leucemia aguda de linhagem mista foi descrito por MIRRO et al., 1985, onde observaram a frequência e o significado da leucemia aguda exibindo características linfoides e mieloides em 123 crianças. Neste estudo, a definição de leucemia aguda de linhagem mista incluiu blastos individuais com mais de uma linhagem. Isso foi determinado usando anticorpos associados a linfoides, como anti-CALLA (CD10), T-11 (CD2) ou T101 (CD5), e os anticorpos associados a mielóides eram compostos por MY-1 (CD15), MCS.2 (CD13) e Mo1 (CD11b). No entanto, nenhum desses marcadores é agora considerado específico de linhagem. Com base nesses marcadores, a leucemia de linhagem mista aguda compreendeu 20% do número total dos 123 casos. A partir dos dados clínicos disponíveis, os autores observaram que a maioria dos pacientes com LLA e marcadores mieloides entrou em remissão completa, enquanto naqueles com LMA e marcadores linfoides a resposta clínica foi mais heterogênea (MIRRO et al., 1985).

À medida que mais anticorpos se tornavam disponíveis, ficou claro que uma porcentagem significativa de casos de LMA e LLA demonstrou imunofenótipos anormais e critérios mais específicos foram necessários para diagnosticar uma leucemia fenotípica mista verdadeira (KHALIDI et al., 1999).

Existem diversas formas de definir esse transtorno, como os critérios de pontuação propostos pelo Grupo Europeu para a Caracterização Imunológica das Leucemias. No entanto, o peso relativo dado a alguns marcadores e a falta de especificidade de linhagem da maioria dos marcadores, levantaram questões sobre a importância desta abordagem.

A leucemia aguda com fenótipo misto é uma doença rara e compreende 2–5% de todas as leucemias agudas, acometem pacientes de todas as idades e compreende vários subtipos diferentes. Vários estudos sugeriram que os pacientes com leucemia aguda de fenótipo misto têm um pior resultado clínico quando comparados com controles pareados com leucemia mieloide aguda ou leucemia linfoide aguda. Mais estudos são necessários para determinar uma abordagem de tratamento padronizada e para melhor compreender os aspectos biológicos e clínicos desta doença (WEINBERG; ARBER, 2010).

1.5. História da Biologia do Câncer:

Em um panorama histórico com o objetivo de entender a história da doença e apresentar uma forma de auxílio para o tratamento desta, podemos resumir a biologia do câncer em duas fases, a “era da biologia celular” e a “era da biologia molecular”. Tudo começou com o nascimento da “era da biologia celular” em 1665 por Robert Hooke, em seu livro *Micrographia.* Através de suas observações, descreveu pela primeira vez o que hoje nós conhecemos como "células", estruturas quadradas, que lembravam as “células” às quais os monges residem.(LEWIS et al., 2021) Essa descoberta foi fundamental para a compreensão não só das estruturas biológicas em detalhes microscópicos, mas também das diferenças e mudanças associadas às doenças.

Mas foi em meados do século XIX que Johannes Müller observou os tumores pela primeira vez em um microscópio de luz. Desde então, os cientistas vêm procurando compreender a natureza e mais precisamente, a composição celular atípica do câncer (HAJDU, 2012). Em 1941, com a imunohistoquímica (COONS; CREECH; JONES, 1941), foi possível caracterizar tecidos e alvos proteicos e assim foram aparecendo cada vez mais melhorias nas instrumentações e técnicas

histológicas, sustentando inúmeras descobertas no assunto e que são utilizadas até hoje como diagnóstico e prognóstico em pacientes com câncer, como os exames histopatológicos de lesões biopsiadas.

A segunda fase surgiu em meados do século XX com a descoberta do DNA e do RNA e logo, o desenvolvimento do sequenciamento destes, da engenharia genética e dos oncogenes (COBB, 2015), levando ao nosso entendimento atual que alterações no genoma (como polimorfismos de nucleotídeos, variação no número de cópias e anomalias cromossômicas) e epigenoma (metilação do DNA, modificações na cromatina) são a base do estabelecimento e propagação dos tumores (BERGER; MARDIS, 2018).

As tecnologias de sequenciamento de segunda geração levaram a um aumento substancial de escala, possibilitando sequenciamento e análise de genomas, transcriptomas e outras propriedades moleculares da célula. Tais análises revelaram novas informações sobre a base molecular dos tumores e propiciaram o desenvolvimento de novas tecnologias que possibilitam o sequenciamento de RNA em células únicas (*single cell* RNA sequencing/scRNA-seq), revolucionando a forma como se avalia a heterogeneidade celular (STUART; SATIJA, 2019).

1.6. scRNA-seq:

Em um único organismo, a maioria das células possuem o mesmo genoma, mas a expressão gênica específica varia em diferentes tecidos e tipos de células. Qualquer tecido ou tipo de célula expressa um número grande de genes, uma vez que o genoma humano codifica mais de 20 mil genes codificadores de proteínas , dos quais, um subconjunto possui um padrão de expressão específico para cada tipo de célula, enquanto os genes restantes são expressos de forma relativamente uniforme(RAMSKÖLD et al., 2009).

Esses padrões únicos de expressão gênica se traduzem em diferenças no nível de proteína entre diferentes tipos de células e resultam na vasta gama de fenótipos celulares encontrados em todo o corpo. Portanto, um perfil instantâneo de

expressão gênica de uma única célula pode ser indicativo de seu fenótipo. Devido à quantidade limitada de RNA presente em cada célula, o perfil de expressão gênica foi historicamente realizado em células agrupadas, mas essa abordagem de sequenciamento em massa obscureceu a potencial heterogeneidade celular em uma amostra ou tecido. Por exemplo, em um conjunto de células progenitoras em desenvolvimento, células distintas podem ser preparadas para tomar decisões com diferentes destinos, mas esses programas de transcrição são indistinguíveis em uma análise em massa da expressão gênica média no conjunto progenitor.

O desenvolvimento de tecnologias que podem isolar milhares a dezenas de milhares de células e avaliar seus perfis de expressão gênica em nível de célula única permitiu aos pesquisadores dissecar essa heterogeneidade celular e trabalhar para uma melhor compreensão da fisiologia, desenvolvimento biológico e de doenças a partir desses estudos (POTTER, 2018).

O surgimento da genômica de células únicas possibilitou novas estratégias para identificar os mecanismos celulares e moleculares dos fenômenos biológicos. Os padrões na expressão de mRNA quantificados por scRNA-seq podem ser utilizados para descobrir tipos de células, estados e circuitos distintos dentro de populações de células e tecidos (SHALEK et al., 2013). Os métodos de scRNA-seq permitem estudos de alto rendimento de fenótipos celulares, particularmente para amostras de *low-input* ($\leq 10^4$ células), como espécimes clínicos, para aumentar nossa compreensão de comportamentos saudáveis e alterados e, assim, orientar diagnósticos e terapias com precisão (GIERAHN et al., 2017).

O scRNA-seq envolve o isolamento e a lise de células individuais e, em seguida, a transcrição reversa e a amplificação independente de seus mRNAs antes de gerar bibliotecas com código de barras que são agrupadas para sequenciamento. Métodos desenvolvidos recentemente atribuíram códigos de barras exclusivos aos mRNAs de cada célula durante a transcrição reversa, permitindo o processamento de diversas amostras, mantendo a resolução de uma única célula. Essas técnicas normalmente geram bibliotecas de célula única de baixa complexidade, mas o alto rendimento reduz o impacto do ruído técnico associado a cada célula nas análises (KLEIN et al., 2015) (MACOSKO et al., 2015).

Através de métodos abrangidos pela biologia computacional à partir de dados brutos gerados pelo sequenciamento de RNA de célula única (scRNA-seq), os quais contêm todas as leituras de DNA complementares sequenciadas, inicia-se a primeira etapa da análise, que consiste em atribuir leituras individuais à sua célula de origem para gerar uma matriz de contagem de células únicas.

O próximo passo envolve filtrar células e genes de acordo com as métricas de controle de qualidade (CQ). A normalização de dados, escalonamento e estabilização de variância são usados para lidar com vieses técnicos e facilitar a seleção dos genes biologicamente mais relevantes, garantindo que a análise seja conduzida por fenômenos biológicos relevantes e não por ruído técnico. A comparação de conjuntos de dados adquiridos de diferentes experimentos também requer a correção de efeitos de *batch* para permitir a integração de dados adequada. A redução de dimensionalidade resume os padrões de expressão de milhares de genes em dimensões menores, que são usados para criar aglomerados de células com padrões semelhantes de expressão gênica. Em conjuntos de dados de desenvolvimento, as células geralmente não se agrupam em clusters discretos, mas seguem trajetórias contínuas, exigindo um modelo contínuo de estados celulares. A inferência de trajetória visa identificar a localização das células ao longo do desenvolvimento contínuo.

Por fim, o conjunto de dados pode ser visualizado em duas dimensões e analisado para identificar os principais genes marcadores em cada cluster. Quaisquer tipos ou estados de células desconhecidas são então anotados usando esses genes marcadores-chave ou por meio de comparações com conjuntos de dados de referência existentes (WU; ZHANG, 2020).

2. OBJETIVOS:

2.1. Objetivo geral:

Identificar programas de expressão gênicos associados a heterogeneidade e plasticidade de células leucêmicas.

2.2. Objetivos específicos:

I. Analisar dados de sequenciamento do transcriptoma de células únicas em amostras de medula óssea saudável e definir a diversidade de tipos celulares nas amostras utilizando o software FUSCA.
II. Mapear dados de sequenciamento do transcriptoma de células leucêmicas no atlas celular de referência da medula óssea saudável utilizando o algoritmo Symphony
III. Classificar os tipos celulares presentes nas amostras utilizando o algoritmo singleCellNet.

3. METODOLOGIA:

3.1. Obtenção de Dados:

Serão selecionadas pesquisas com contagens de scRNA-seq publicamente disponíveis (Anexo 1) da medula óssea de indivíduos saudáveis e de indivíduos acometidos com Leucemia Aguda de Fenótipo Misto. A matriz de contagens, a qual quantifica a expressão de cada gene em cada célula, será obtida a partir do número de acesso GEO GSE139369. Esta matriz é o resultado do processamento dos dados brutos de sequenciamento gerados pelo *pipeline* CellRanger, da 10X genomics, e representa o ponto de partida do projeto proposto.

3.2 Tratamento dos Dados:

A análise dos dados de sequenciamento do transcriptoma de células únicas será realizada utilizando o Software FUSCA (Framework for Unified Single-Cell Analysis), desenvolvido no Laboratório de Biologia de Sistemas da Universidade Federal de Santa Catarina. O FUSCA disponibiliza um conjunto de métodos analíticos para interpretar e priorizar programas de expressão de genes, assim como redes de regulação gênicas, trajetórias de diferenciação e comunicação celular, sendo uma plataforma de biologia de sistemas altamente versátil para a compreensão de dados biológicos complexos. Os dados a serem analisados são provenientes de 6 doadores saudáveis, e 6 pacientes com LAFM. No total, os dados contêm em torno de 20 mil células saudáveis e 17 mil células leucêmicas. Os dados são provenientes do artigo original por GRANJA et al., 2019.

3.2.1. Controle de Qualidade e Análise Exploratória:

Inicialmente, a matriz de contagens será importada pelo FUSCA para o ambiente de programação e desenvolvimento R. Genes com níveis de expressão baixos em menos do que 10% das células, e células com alto percentual de expressão de genes mitocondriais (tipicamente acima de 10%) serão removidos, pois representam genes baixamente expressos, ou células estressadas e de baixa qualidade devido ao processamento da amostra, respectivamente. Em seguida, será realizada uma análise de redução de dimensionalidade utilizando os algoritmos do FUSCA para a análise utilizando Uniform Manifold Approximation and Projection (UMAP) para

visualizar os dados e investigar a presença de efeitos de *batch* ou possíveis fontes de variação técnica associadas aos doadores. Se um efeito de *batch* for observado, o algoritmo Harmony (https://github.com/immunogenomics/harmony) será utilizado para integrar os dados dos diversos doadores e pacientes pela identificação de genes diferencialmente expressos (KORSUNSKY et al., 2019).

3.2.2 Clusterização não-supervisionada e identificação de tipos celulares:

Após os dados serem apropriadamente integrados, os algoritmos de clusterização e análise de expressão diferencial (utilizando o teste estatístico de Wilcoxon) do FUSCA, serão utilizados para determinar a composição celular das amostras de medula óssea saudáveis, assim como das amostras leucêmicas. Para isto, clusters transcricionais serão verificados em termos da expressão de genes marcadores de diversas populações hematopoiéticas, o que permitirá determinar uma correspondência entre clusters transcricionais e tipos celulares nas amostras saudáveis.

3.2.3. Mapeamento de populações leucêmicas utilizando singleCellNet e Symphony:

Como a leucemia de fenótipo misto possui contribuições de tipos celulares provenientes de diversas linhagens hematopoiéticas, o próximo objetivo é determinar a equivalência de tipos celulares entre as amostras saudáveis e leucêmicas. Para isto, modelos de aprendizado de máquina da identidade celular de cada célula hematopoiética proveniente das amostras saudáveis serão desenvolvidos utilizando o algoritmo SingleCellNet. Com os modelos computacionais apropriadamente treinados, os dados de células leucêmicas serão quantitativamente comparados a estes modelos, o que determinará a similaridade transcricional de cada célula leucêmica com relação às linhagens hematopoiéticas saudáveis. Esta análise possibilitará uma determinação de alta resolução da conservação entre tipos celulares nas amostras saudáveis (e suas possíveis diferenças transcricionais), e também entre células hematopoiéticas unicamente presentes nas amostras saudáveis ou leucêmicas.

O Symphony (https://github.com/immunogenomics/symphony) é um algoritmo com o propósito de construir um atlas de referência integrado em larga escala com um

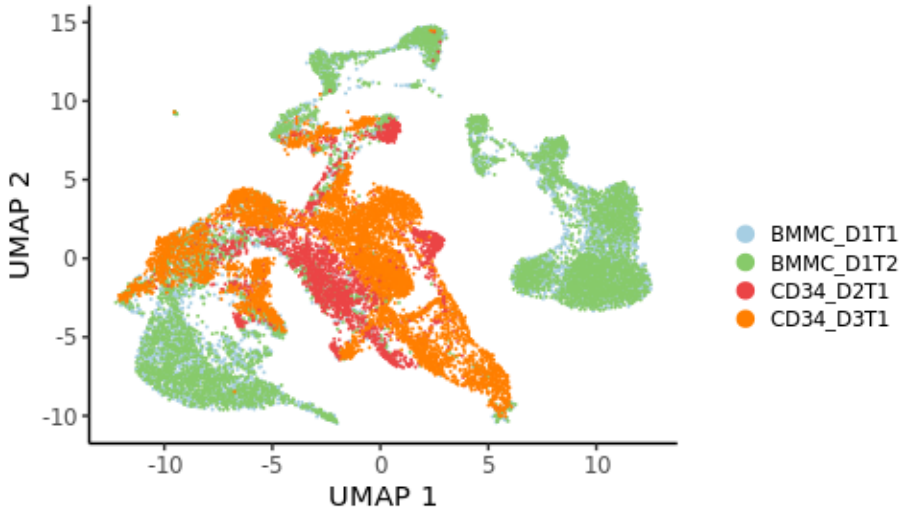
formato conveniente e portátil que permite mapeamento de consulta eficiente em segundos. O Symphony localiza *query cells* em uma incorporação de referência estável de baixa dimensão, facilitando a transferência reprodutível downstream de anotações definidas por referência para a consulta.

4. RESULTADOS:

Para identificar as características patológicas das células mononucleares neoplásicas da medula óssea, primeiramente criamos um atlas de referência da diversidade celular presente na medula óssea durante a homeostase. Como as Leucemias Agudas de Fenótipo Misto apresentam características de múltiplas linhagens hematopoiéticas, mapas imunofenotípicos e transcriptômicos do desenvolvimento normal das células hematopoiéticas foram construídos como parte de uma re-análise dos dados provenientes do artigo original por GRANJA et al., 2019.

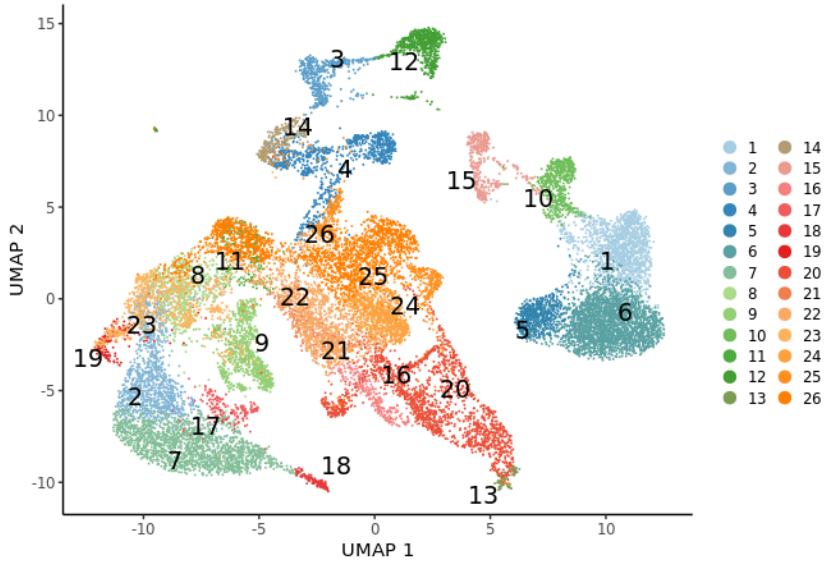
No total, foram analisadas em torno de 20 mil células da medula óssea, incluindo populações celulares terminalmente diferenciadas, e células progenitoras hematopoiéticas, as quais são caracterizadas pela expressão da proteína de superfície CD34. Estes dados foram obtidos de um banco de dados disponível no GEO (visualizar **Anexo 1)**. Inicialmente foi realizada uma análise UMAP para definir similaridades e diferenças entre as células presentes na amostra. As células foram coloridas de acordo com a sua amostra de origem, e as análises demonstraram que as células provenientes da medula óssea (BMMCs) se sobrepõe no mapa celular UMAP, enquanto que as populações progenitoras CD34+ apresentaram uma sobreposição menor, consistente com a heterogeneidade associadas às células progenitoras hematopoiéticas (**Figura 1**).

Figura 1. UMAP colorido de acordo com a amostra de origem.



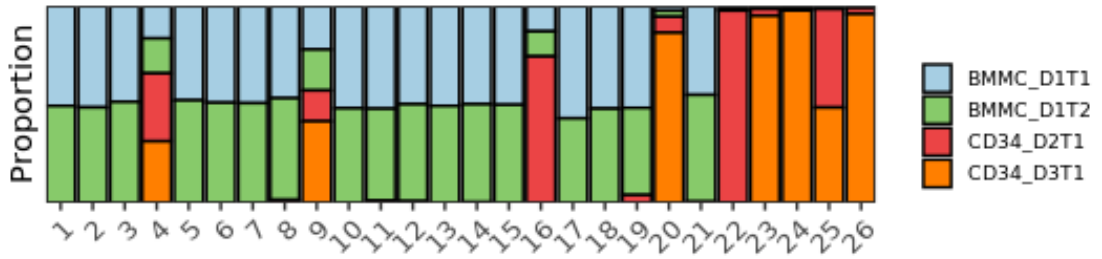
Já para a **Figura 2**, mostra uma análise de clusterização não-supervisionada dos dados transcriptômicos, a qual foi utilizada para definir a composição celular das amostras com base na expressão de genes específicos para diferentes subtipos de células tronco e progenitoras hematopoiéticas.

Figura 2. UMAP colorido de acordo com os clusters transcricionais identificados pelo FUSCA.



Os clusters foram sobrepostos no UMAP para visualização. Uma análise da proporção de células provenientes de cada amostra em cada cluster transcricional indica pouca sobreposição de células provenientes da medula óssea e os progenitores CD34+, embora clusters 4, 9, 16 e 20 contenham células de diversas amostras. Esta análise indica a apresentação, em potencial, de células diferenciadas na população de células progenitoras CD34+ (**Figura 3**).

Figura 3. Proporção das amostras em cada cluster transcricional identificado pelo FUSCA.



Observando os padrões de expressão gênicas dos genes selecionados (**Figura 4 e Figura 5**), os clusters foram então anotados com base no tipo celular correspondente (**Figura 6**). CD14 é um marcador de Monócitos, CD19 é um marcador de linfócitos B, CD8A, CD4 e TIGIT são marcadores de linfócitos T, ELANE é um marcador progenitores de neutrófilos, KLF1 e GATA1 são marcadores de eritrócitos, enquanto que RUNX1, MECOM e CD34 são marcadores de células progenitoras hematopoiéticas.

Figura 4. UMAPs onde cada célula é colorida de acordo com a expressão dos genes indicados.

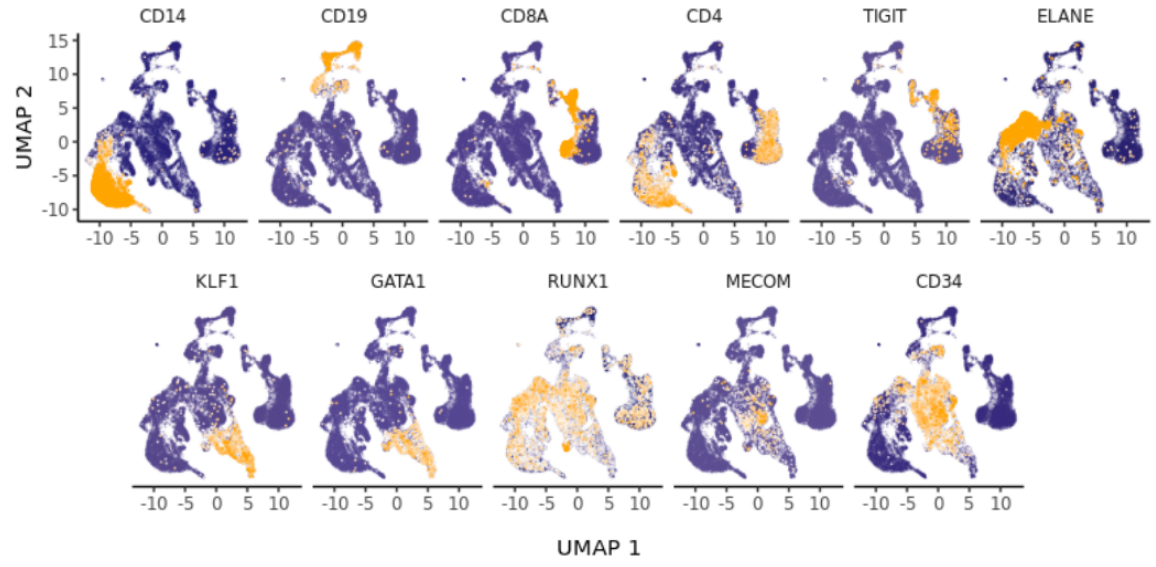
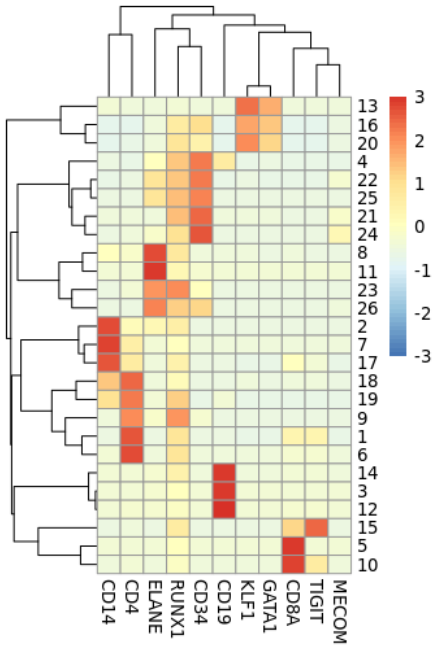
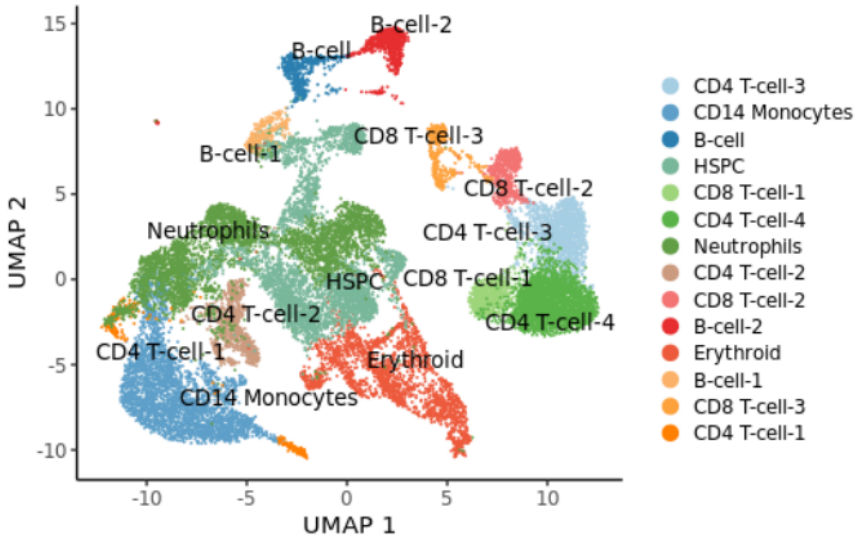


Figura 5. Heatmap mostrando a expressão média dos genes marcadores de linguagem em cada cluster transcricional.



Com base nos padrões de expressão destes genes selecionados, o atlas celular de referência da medula óssea foi anotado com os tipos celulares correspondentes que compõem a amostra (**Figura 6**).

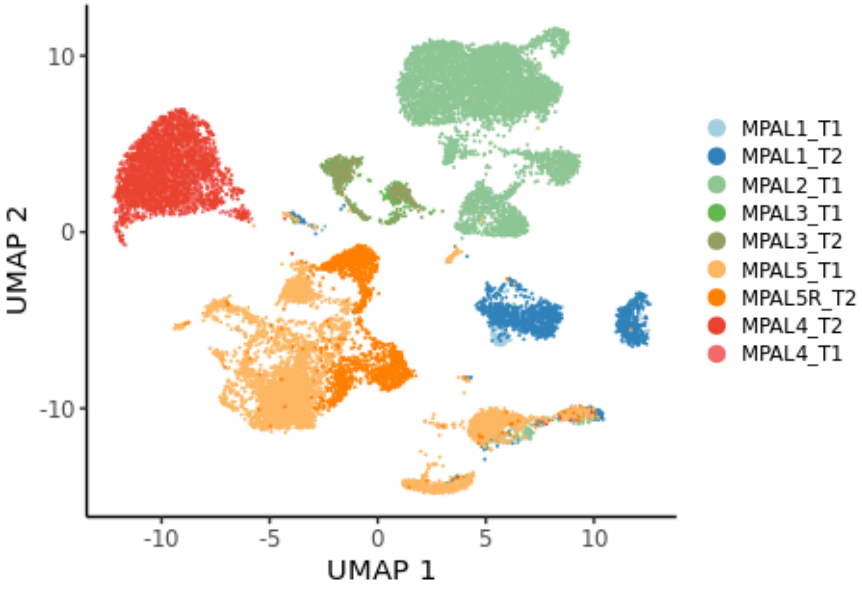
Figura 6. UMAP colorido de acordo com os tipos celulares identificados.



Para explorar a natureza da disfunção regulatória e fenotípica das células leucêmicas, foram analisada seis amostras de LAFM, incluindo três LAFMs

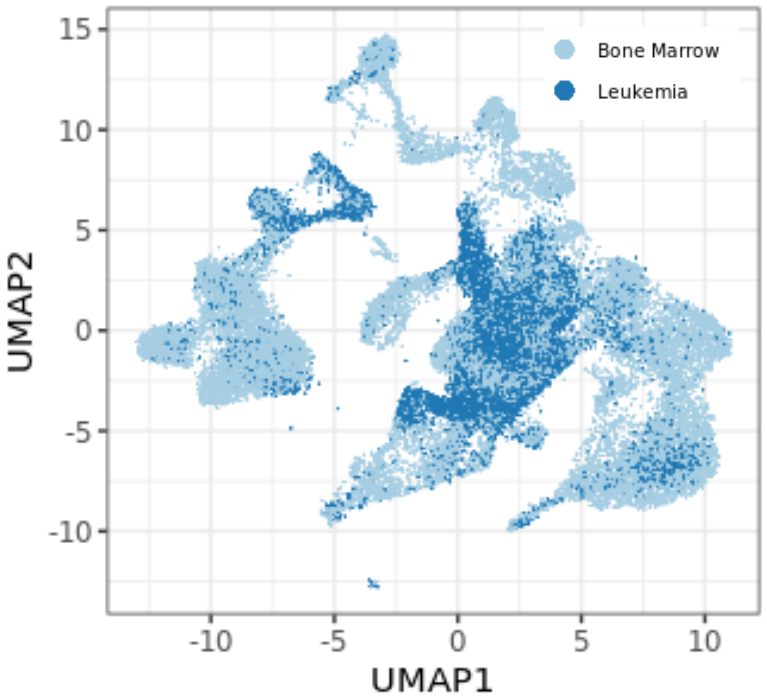
T-mielóide (LAFM1-LAFM3), 1 B-LAFM mielóide (LAFM4) e um LAFM T-mielóide amostrado antes da quimioterapia CALGB (LAFM5) e após recidiva pós-tratamento (LAFM5R). No total, foram re-analisadas aproximadamente 17,500 células leucêmicas (**Figura 7**).

Figura 7. UMAP das amostras leucêmicas coloridas de acordo com a amostra de origem.



Subsequentemente, as células leucêmicas foram mapeadas no atlas de referência da medula óssea saudável utilizado o algoritmo Symphony, e visualizadas utilizando UMAP, unificando e delineando os padrões celulares das amostras, como apresentado na **Figura 8**. Esta análise mostra que a maioria da células leucêmicas é transcricionalmente similar às células progenitoras hematopoiéticas CD34+, enquanto que mapeamentos também foram observados em regiões do UMAP onde células linfóides e/ou mielóides estão presentes no atlas de referência (**Figura 6 e 8**).

Figura 8. UMAP contendo os dados de medula óssea saudáveis e leucêmicas sobrepostos após a integração com o algoritmo Symphony.



É interessante notar que as células leucêmicas são altamente heterogêneas, com algumas amostras possuindo características transcricionais mais relativamente exclusivas das células progenitoras CD34+ (como as amostras MPAL4_T1 e MPAL4_T2), enquanto que outras amostras leucêmicas apresentaram similaridades com progenitores hematopoiéticas, células linfóides e mielóides (como MPAL1_T2 ou MPAL5_T1) (**Figura 9 e 10**).

Figura 9. UMAPs individuais demonstrando a contribuição de cada amostra leucêmica com o atlas celular integrado da medula óssea.

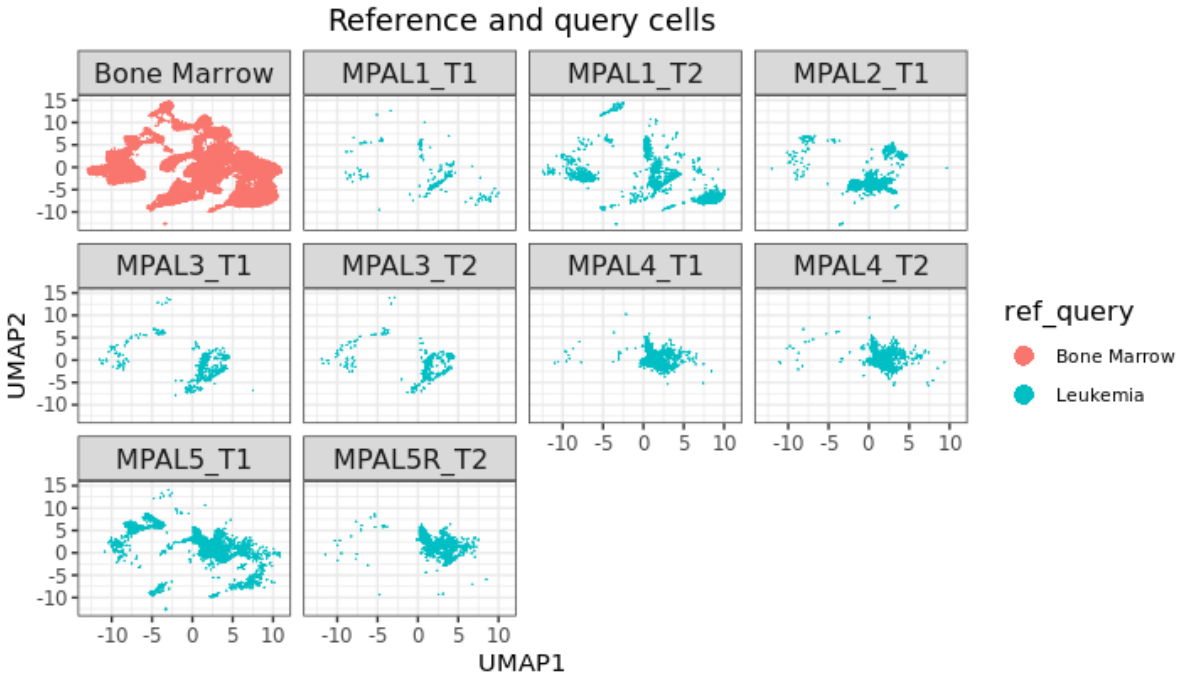
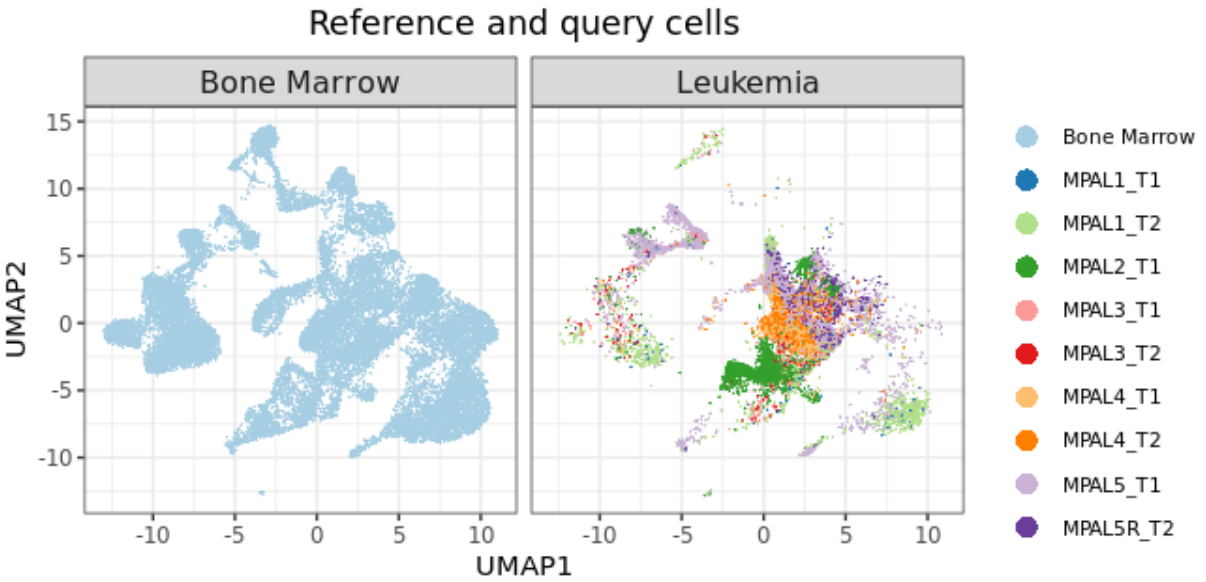
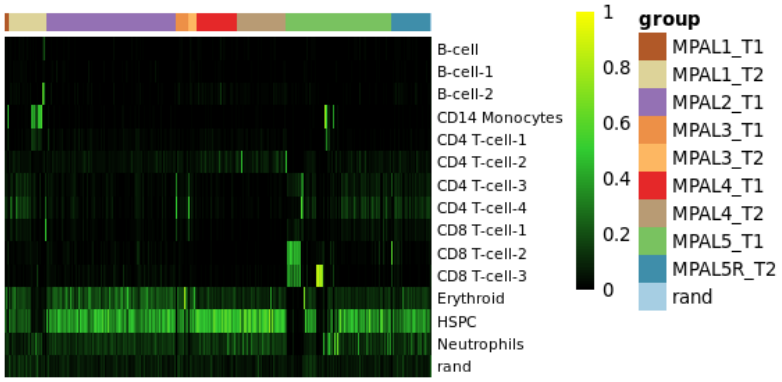


Figura 10. UMAPs mostrando como as amostras individuais são mapeadas no atlas de referência da medula óssea saudável utilizando o algoritmo Symphony.



Finalmente, o algoritmo singleCellNet foi utilizado para classificar células leucêmicas de acordo com a sua similaridade transcricional com células saudáveis da medula óssea. Em conformidade com as análises do algoritmo Symphony, a maioridade das células leucêmicas foi classificada como *Hematopoietic stem and progenitor cell* (HSPC), o que pode explicar o fenótipo misto observado nestas leucemias, pois as HSPCs podem se diferenciar em diferentes subtipos de células hematopoiéticas (**Figura 11**).

Figura 11. Heatmap dos scores de classificação celular gerados pelo algoritmo singleCellNet.



5. DISCUSSÃO:

Através de conhecimentos sobre o câncer, leucemias e scRNA-seq, este último sendo um método recente que utiliza ferramentas de análise computacional recebendo méritos em revistas renomadas e se apresentando tão relevante na pesquisa de tratamento à diversos tipos de câncer. Este projeto foi realizado com objetivo de promover uma análise alternativa aos dados apresentados no artigo de GRANJA et al., 2019, partindo do pressuposto de que a reanálise traria novas formas de se avaliar e apresentar novos mapas transcriptômicos para o trabalho.

O que se concretizou com o auxílio de novos algoritmos que foram utilizados para o tratamento dos dados, tais como o Harmony e o Symphony. Foram criados novos mapas de comparação entre os dados de transcriptoma de células mononucleares de medula óssea para indivíduos saudáveis e indivíduos acometidos com Leucemia Aguda de Fenótipo Misto, assim como uma proporção, também inédita no artigo original, entre os tipos celulares presentes nas amostras de medulas saudáveis (Figura 3).

Novos mapeamentos exclusivos utilizando um atlas de referência com dados de indivíduos saudáveis em comparação com pacientes com LAFM foram construídos, permitindo visualizar individualmente os seus respectivos clusters de expressão (Figura 9), como também de forma unificada (Figura 10). Possibilitando a visualização de uma aparente heterogeneidade entre as amostras disponíveis de LAFM. Mapeamentos específicos de certas amostras, como em MPAL4_T1,T2 e MPAL5R, quando comparados ao atlas de referência, apresentaram uma similaridade maior com progenitores hematopoiéticos CD34+, por meio da análise pelo Symphony e uma maior classificação como *Hematopoietic stem and progenitor cells* (HSPCs) pelo algoritmo singleCellNet. O que chama atenção pois a amostra MPAL5R se trata de um paciente pós-tratamento, trazendo a informação de que mesmo após recidiva, o organismo ainda se apresenta com manifestações recorrentes de leucemia, com uma expressão majoritária de células progenitoras com o potencial de se diferenciar em diferentes subtipos de células hematopoiéticas. Essa descoberta poderia gerar um novo trabalho.

Coletivamente, este trabalho estabelece uma abordagem analítica para avaliar características específicas da doença usando a análise integrativa de célula única. Foi observado que os programas malignos de LAFM são amplamente conservados em células fenotipicamente heterogêneas em pacientes individuais, essa observação é consistente com um relato anterior (ALEXANDER *et al.,* 2018) de que as células de LAFM provavelmente se originam de uma célula progenitora multipotente, compartilhando assim uma paisagem mutacional comum enquanto povoam diferentes regiões da árvore hematopoiética.

Como perspectivas futuras, poderiam agregar ao trabalho novas análises utilizando o sangue periférico para verificação da expressão à nível de célula-única circulante e novos métodos como scATAC-seq e scADT-seq.

Por fim, é possível prever que abordagens semelhantes serão usadas em estudos futuros para identificar o status de diferenciação de diferentes tipos de tumor e avaliar a disfunção molecular em subtipos celulares patogênicos, com o objetivo final de identificar alvos terapêuticos personalizados por meio da caracterização integrativa molecular de célula-única.

6. REFERÊNCIAS:

MS / SVS/DASIS/CGIAE/Sistema de Informação sobre Mortalidade, 2021.

MS / INCA / Coordenação de Prevenção e Vigilância / Divisão de Vigilância e Análise de Situação, 2021.

BOROWITZ, M. J. Acute leukemias of ambiguous line-age. **WHO** classification of tumors of haematopoietic and lymphoid tissues, p. 150-155, 2008.

AL-SERAIHY, A. S. et al. Clinical characteristics and outcome of children with biphenotypic acute leukemia. **Haematologica**, v. 94, n. 12, p. 1682–1690, 1 dez. 2009.

KILLICK, S. et al. Outcome of biphenotypic acute leukemia. **Haematologica**, v. 84, n. 8, p. 699–706, ago. 1999.

OWAIDAH, T. M. et al. Cytogenetics, molecular and ultrastructural characteristics of biphenotypic acute leukemia identified by the EGIL scoring system. **Leukemia**, v. 20, n. 4, p. 620–626, 26 abr. 2006.

POTTER, S. S. Single-cell RNA sequencing for the study of development, physiology and disease. **Nature Reviews Nephrology**, v. 14, n. 8, p. 479–492, 22 ago. 2018.

RAMSKÖLD, D. et al. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. **PLoS Computational Biology**, v. 5, n. 12, p. e1000598, 11 dez. 2009.

WEINBERG, O. K.; ARBER, D. A. Mixed-phenotype acute leukemia: historical overview and a new definition. **Leukemia**, v. 24, n. 11, 16 nov. 2010.

WU, Y.; ZHANG, K. Tools for the analysis of high-dimensional single-cell RNA sequencing data. **Nature Reviews Nephrology**, v. 16, n. 7, p. 408–421, 27 jul.

2020.

XU, X.-Q. et al. Clinical and biological characteristics of adult biphenotypic acute leukemia in comparison with that of acute myeloid leukemia and acute lymphoblastic leukemia: a case series of a Chinese population. **Haematologica**, v. 94, n. 7, p. 919–927, 1 jul. 2009.

ALEXANDER, T.B., Gu, Z., Iacobucci, I. *et al.* The genetic basis and cell of origin of mixed phenotype acute leukaemia. *Nature* 562, 373–379 (2018).

BENE, M. C. Biphenotypic, bilineal, ambiguous or mixed lineage: strange leukemias! **Haematologica**, v. 94, n. 7, 1 jul. 2009.

BERGER, M. F.; MARDIS, E. R. The emerging clinical relevance of genomics in cancer medicine. **Nature Reviews Clinical Oncology**, v. 15, n. 6, 29 jun. 2018.

COBB, M. Who discovered messenger RNA? **Current Biology**, v. 25, n. 13, jun. 2015.

COONS, A. H.; CREECH, H. J.; JONES, R. N. Immunological Properties of an Antibody Containing a Fluorescent Group. **Experimental Biology and Medicine**, v. 47, n. 2, 1 jun. 1941.

GIERAHN, T. M. et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. **Nature Methods**, v. 14, n. 4, 13 abr. 2017.

GRANJA, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. **Nature Biotechnology**, v. 37, n. 12, p. 1458–1465, 2 dez. 2019.

HAJDU, S. I. A note from history: Landmarks in history of cancer, part 3. **Cancer**, v. 118, n. 4, 15 fev. 2012.

HAO, T. et al. An emerging trend of rapid increase of leukemia but not all cancers in the aging population in the United States. **Scientific Reports**, v. 9, n. 1, 19 dez.

2019.

KANG, J. B. et al. Efficient and precise single-cell reference atlas mapping with Symphony. **Nature Communications**, v. 12, n. 1, p. 5890, 7 dez. 2021.
KILLICK, S. et al. Outcome of biphenotypic acute leukemia. **Haematologica**, v. 84, n. 8, p. 699–706, ago. 1999.

KHALIDI, H. S. et al. Acute Lymphoblastic Leukemia: Survey of Immunophenotype, French-American-British Classification, Frequency of Myeloid Antigen Expression, and Karyotypic Abnormalities in 210 Pediatric and Adult Cases. **American Journal of Clinical Pathology**, v. 111, n. 4, 1 abr. 1999.

KLEIN, A. M. et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. **Cell**, v. 161, n. 5, maio 2015.

KORSUNSKY, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. **Nature Methods**, v. 16, n. 12, p. 1289–1296, 18 dez. 2019.

LEWIS, S. M. et al. Spatial omics and multiplexed imaging to explore cancer biology. **Nature Methods**, 2 ago. 2021.

MACOSKO, E. Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. **Cell**, v. 161, n. 5, maio 2015.

MCGRAW, T. P. et al. Terminal deoxynucleotidyl transferase-positive acute myeloblastic leukemia. **American Journal of Hematology**, v. 10, n. 3, maio 1981.

MIRRO J, ZIPF TF, PUI CH, KITCHINGMAN G, WILLIAMS D, MELVIN S, MURPHY SB, STASS S. Acute mixed lineage leukemia: clinicopathologic correlations and prognostic significance. **Blood**. 1985 Nov;66(5):1115-23. PMID: 3931724.

SHALEK, A. K. et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. **Nature**, v. 498, n. 7453, 13 jun. 2013.

STUART, T.; SATIJA, R. Integrative single-cell analysis. **Nature Reviews Genetics**,

v. 20, n. 5, 29 maio 2019.

SUNG, H. et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. **CA: A Cancer Journal for Clinicians**, v. 71, n. 3, 4 maio 2021.

INSTITUTO NACIONAL DO CÂNCER (Brasil). Leucemia. *In*: INSTITUTO NACIONAL DO CÂNCER (Brasil). Tipos de câncer. RJ: **Instituto Nacional do Câncer**, 2021. Disponível em: https://www.inca.gov.br/tipos-de-cancer/leucemia Acesso em: 4 set. 2021.

World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. **WHO**; 2021. Acessado em 4 de Setembro de 2021. https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death

7. ANEXOS

7.1. Anexo 1: banco de dados públicos citados e respectivos endereços eletrônicos

● GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE139369);