Análise de marcadores celulares em células da medula óssea por meio do sequenciamento de Single Cell RNA

Lucas Kaoru Kobo Ferreira

Orientador: Helder Takashi Imoto Nakaya

Faculdade de Ciências Farmacêuticas

Universidade de São Paulo

Lucas.kaoru@usp.br

Objetivos

Este projeto tem por objetivo identificar e classificar diferentes tipos celulares encontrados na medula óssea, bem como acompanhar a diferenciação celular, a fim de encontrar possíveis marcações que levam ao câncer.

Objetivos específicos

Acesso e busca de dados scRNA-seq, os quais sejam relevantes e no correto formato para a análise.

Seleção e filtragem das células relevantes para o estudo, por meio de diferentes métodos de normalização, controles de qualidade e dimensionamento de dados.

Utilização dos dados pré-processados para separação das diferentes células em distintos grupos, com base na variação de suas expressões gênicas.

utilização de diferentes marcadores a fim de identificar quais são mais expressos e atribuir uma pontuação com base no seu nível de expressão em cada célula pré selecionada.

Métodos e Procedimentos

Para a obtenção de dados scRNA-seq foi utilizado o banco de dados público GEO (*Gene Expression Omnibus*, www.ncbi.nlm.nih.gov/geo/). Os dados analisados são encontrados por meio do código de acesso GSE120446, publicado em 26 de setembro de 2018 sob o título de "Human Bone Marrow Assessment by Single Cell RNA Sequencing, Mass Cytometry and Flow Cytometry". Os dados incluem sequenciamentos de scRNA, citometria de massa e citometria de fluxo de medulas ósseas de 20 doadores saudáveis e são utilizados em "Human bone marrow assessment by single-cell RNA sequencing, mass cytometry, and flow cytometry". Os dados de scRNA-seq utilizados neste estudo foram baixados em formato TSV e MTX, nos quais os barcodes e os genes são arquivos TSV e a matrix MTX.

Controle de qualidade: Antes de realizar a análise da expressão dos genes, é necessário garantir que todos os “barcodes” (sequências de DNA utilizadas para identificação) correspondam a células viáveis, assim, o pré-processamentos dos dados brutos foi realizado por meio do pacote Seurat em R. Foram utilizados como parâmetros: o número de genes únicos em cada célula ou nFeatures_RNA, no qual um um valor elevado indica a presença de “doublet” ou “multiplet”, enquanto um valor baixo indica uma célula morta ou um “droplet” vazio; e as frações de genes mitocondriais, que podem indicar uma contaminação de DNA mitocondrial. Assim, foram utilizados os parâmetros nFeatures_RNA > 200 e < 2500 e percent.mt < 8.

Normalização: Após a remoção das células indesejadas, foi realizada a normalização dos dados a fim de reduzir a

redundância dos dados e melhorar a integridade dos dados, seguida da transformação em log por meio da função “LogNormalize” e da modificação linear com a função “ScaleData” que escala e centraliza os conjuntos de dados ao alterar a expressão de cada gene para que a expressão média das células seja 0, enquanto escala a expressão para que a variação nas células seja 1, assim, os dados são dimensionados a fim de permitir um melhor trabalho sem perder desempenho.

Visualização: Para a visualização da expressão dos genes, foi utilizada a redução dimensional linear em um gráfico de *Principal Component Analysis* (PCA), um processo que converte um conjunto de diversos dados relacionados em um conjunto de valores variáveis linearmente não correlacionados nos dados pré escalados. Para definir quantos componentes devem ser incluídos no conjunto de dados foi utilizada a função “JackStrawPlot” que compara os valores de p para cada PC e permitirá a visualização da queda de significância, sugerindo quais PCs capturam o sinal verdadeiro, sendo selecionado 20 PCs.

Integração: A integração dos dados de sequenciamento de RNA das medulas ósseas foi realizada por meio do pacote Seurat com as funções “FindIntegrationAnchors” e “IntegrateData”, as quais criam “âncoras” que são conjuntos de células (uma em cada dataset) que formam a base para a integração dos dados realizada a seguir

Regressão do ciclo celular: As etapas de controle de qualidade e normalização não conseguem remover vieses biológicos como os efeitos do ciclo celular, assim, foi realizada a remoção desses efeitos no transcriptoma por meio de uma simples regressão linear com o pacote Seurat. A partir dos dados gerados, é realizada uma pontuação e classificação da fase celular (G2M, S or G1) de cada uma das células por meio da função “CellCycleScoring”, que utiliza uma lista de marcadores de ciclo celular. Em seguida, é realizada a regressão linear por meio da função “ScaleData”, que modela a relação entre a expressão do gene e a pontuação das fases G2M e S.

Agrupamento de células: Para realizar o agrupamento de células, ou clustering, empregou-se a função “FindNeighbors”, que determina os vizinhos mais próximos de cada célula gerando um gráfico KNN e então constrói um gráfico SNN, calculando a sobreposição da vizinhança entre cada célula. Foi usada, ainda, a função “FindClusters” que implementa o procedimento realizando a mesma sequência de cálculo dos vizinhos mais próximos e construção de um gráfico SNN, porém esta função contém um parâmetro de resolução que define a “granularidade” do cluster, aumentando o número de clusters. Por fim, a redução dimensional foi realizada com a utilização de técnicas como UMAP e tSNE, os quais são algoritmos que realizam uma redução de dimensões e geram uma projeção.

Expressão e identificação de células: Para encontrar marcadores de expressão diferencial foi utilizada a função “FindMarkers”, a qual compara as células e automatiza o processo de clusterização. Ela retorna os genes marcadores, ou seja, os genes mais expressos do cluster analisado em comparação com os demais clusters e assim, com base nesses genes, é possível inferir o tipo celular em questão. Também foi utilizada a visualização de marcadores celulares já conhecidos por meio de diversos plots que o pacote Seurat oferece, tais como o “VInplot”, que mostra distribuições de expressão entre os clusters, e “FeaturePlot” que colore as células de acordo com um recurso selecionado, como um gene marcador.

Foi pretendido, ainda, o treino e utilização de um classificador de células por meio do pacote Garnett em R. O pacote parte de um arquivo com os tipos celulares e seus respectivos marcadores, e em seguida, treina o classificador ao realizar uma comparação com células por meio da função “train_cell_classifier”.

Resultados

As visualizações e plots gerados revelam uma clara diferenciação ou “amadurecimento” celulares em todos os tipos presentes na medula óssea. Além de também mostrar as grandes semelhanças entre células T CD8 e NKs, o que torna sua distinção muito difícil.

A realização da regressão dos efeitos de ciclo celular gerou um impacto negativo nas

análises de downstream, o que pode estar associado com a importância dos genes envolvidos no ciclo celular com a identificação de populações de células em proliferação.

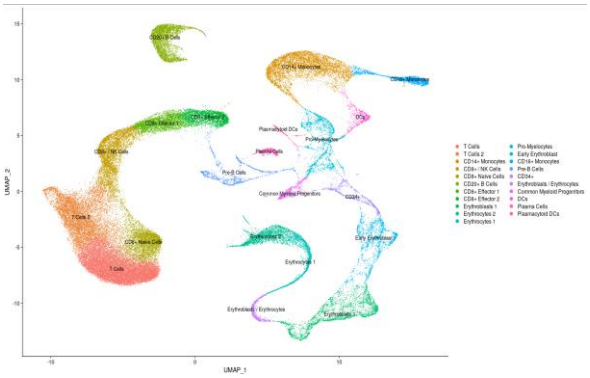

Figura 1: Visualização dos dados integrados das amostras de medulas ósseas por meio da redução dimensional UMAP

Conclusões

A análise por scRNA-seq das medulas ósseas nos revela que células T, assim como os demais tipos celulares encontrados, sofrem um intenso processo de diferenciação, em que há expressão de genes diferentes ao longo de seu amadurecimento. Assim, as células presentes no microambiente tumoral podem apresentar expressão diferencial e fatores de transcrição incomuns, o que nos auxilia no melhor entendimento do desenvolvimento das células T CD8, além de abrir portas para novas terapias ao reconhecer esses alvos.

Entretanto, essas análises também requerem uma grande adaptabilidade para identificar e reconhecer os diferentes processamentos de dados e aplicação conforme o tipo celular estudado, por exemplo, a não utilização de regressão por ciclo celular.

Por fim, o reconhecimento de diferentes marcadores celulares pode auxiliar na identificação de mais tipos celulares, além de ampliar o conhecimento de marcadores relacionados com a exaustão celular.

Referências Bibliográficas

van den Broek, T., Borghans, J. A. M. and van Wijk, F. (2018) **'The full spectrum of human naive T cells'**, *Nature reviews. Immunology*, 18(6), pp. 363–373.

Hewitt, E. W. (2003) **'The MHC class I antigen presentation pathway: strategies for viral immune evasion'**, *Immunology*, 110(2), pp. 163–169.

Hwang, B., Lee, J. H. and Bang, D. (2018) **'Single-cell RNA sequencing technologies and bioinformatics pipelines'**, *Experimental & molecular medicine*, 50(8), p. 96.

Lambert, M. *et al.* (2018) **'Targeting Transcription Factors for Cancer Treatment'**, *Molecules*, p. 1479. doi: 10.3390/molecules23061479.

Langermans, J. A. M., Hazenbos, W. L. W. and van Furth, R. (1994) **'Antimicrobial functions of mononuclear phagocytes'**, *Journal of Immunological Methods*, pp. 185–194. doi: 10.1016/0022-1759(94)90021-3.

Larbi, A. and Fulop, T. (2014) **'From "truly naïve" to "exhausted senescent" T cells: When markers predict functionality'**, *Cytometry Part A*, pp. 25–35. doi: 10.1002/cyto.a.22351.

Latchman, D. S. (1997) **'Transcription factors: An overview'**, *The International Journal of Biochemistry & Cell Biology*, pp. 1305–1312. doi: 10.1016/s1357-2725(97)00085-x.

Leal, A. M., Kumeda, C. A. and Elvira D R (2009) **'Características genéticas da leucemia promielocítica aguda de novo'**, *Revista Brasileira de Hematologia e Hemoterapia*, pp. 454–462. doi: 10.1590/s1516-84842009005000088.

Libermann, T. A. and Zerbini, L. F. (2006) **'Targeting transcription factors for cancer gene therapy'**, *Current gene therapy*, 6(1), pp. 17–33.

Lopes-Ramos, Camila M., Cho-Yi Chen, Marieke L. Kuijjer, Joseph N. Paulson, Abhijeet R. Sonawane, Maud Fagny, John Platig, Kimberly Glass, John Quackenbush, and Dawn

L. DeMeo. 2020. “**Sex Differences in Gene Expression and Regulatory Networks across 29 Human Tissues**.” *Cell Reports* 31 (12): 107795.

Luecken, Malte D., and Fabian J. Theis. 2019. “**Current Best Practices in Single-Cell RNA-Seq Analysis: A Tutorial.**” *Molecular Systems Biology* 15 (6): e8746.

Man, K. *et al.* (2017) **‘Transcription Factor IRF4 Promotes CD8 T Cell Exhaustion and Limits the Development of Memory-like T Cells during Chronic Infection’**, *Immunity*, pp. 1129–1141.e5. doi: 10.1016/j.immuni.2017.11.021.

Medzhitov, R. (2007) **‘Recognition of microorganisms and activation of the immune response’**, *Nature*, pp. 819–826. doi: 10.1038/nature06246.

Nestorowa, Sonia, Fiona K. Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K. Wilson, David G. Kent, and Berthold Göttgens. 2016. “**A Single-Cell Resolution Map of Mouse Hematopoietic Stem and Progenitor Cell Differentiation**.” *Blood* 128 (8): e20–31.

Pliner, Hannah A., Jay Shendure, and Cole Trapnell. n.d. “**Supervised Classification Enables Rapid Annotation of Cell Atlases**.” https://doi.org/10.1101/538652.

Sallusto, F., Geginat, J. and Lanzavecchia, A. (2004) **‘Central memory and effector memory T cell subsets: function, generation, and maintenance’**, *Annual review of immunology*, 22, pp. 745–763.

Single-Cell RNA-Seq: An Introductory Overview and Tools for Getting Started, 10x Genomics, Aug 24, 2017. Disponível em: <https://community.10xgenomics.com/t5/10x-Blog/Single-Cell-RNA-Seq-An-Introductory-Overview-and-Tools-for/ba-p/547>

Schiavinato, J. L. dos S(2011) **‘Papel de Notch e NF-kB na regulação de fatores de transcrição durante a diferenciação in vitro de células T a partir de células progenitoras hematopoéticas CD34’**. doi: 10.11606/d.17.2011.tde-30102014-115459.