UNIVERSIDADE FEDERAL DO PARANÁ

GIULIA LEONEL PASCHOAL
JÚLIA FERNANDES DA SILVA
LARA RIBEIRO RAMPIM
LUCAS MÜLLER
NICOLLI DE OLIVEIRA ROSA

USO DO K-MÉDIAS PARA AGRUPAMENTO DE CÉLULAS MALIGNAS E BENIGNAS NO CÂNCER GÁSTRICO

CURITIBA
2024

Giulia Leonel Paschoal
Júlia Fernandes da Silva
Lara Ribeiro Rampim
Lucas Müller
Nicolli de Oliveira Rosa

USO DO K-MÉDIAS PARA AGRUPAMENTO DE CÉLULAS MALIGNAS E BENIGNAS NO CÂNCER GÁSTRICO

Trabalho apresentado à disciplina de Bioinformática, Setor de Ciências Exatas, Universidade Federal do Paraná.

Orientador: Prof. Dr. Eduardo Jaques Spinosa

CURITIBA
2024

SUMÁRIO

1 INTRODUÇÃO

O câncer gástrico é uma doença complexa e heterogênea com muitos fatores de risco, sendo a quinta malignidade mais comum e a quarta principal causa de morte relacionada ao câncer globalmente, indicando que o câncer gástrico ainda é um desafio de saúde mundial (YANG *et al*., 2023). Dentre os fatores que podem influenciar o desenvolvimento estão fatores genéticos, epigenéticos e ambientais, principalmente hábitos alimentares e comportamentos sociais (MACHLOWSKA *et al*., 2020). Por conta dos avanços em estratégias preventivas, de triagem e terapêuticas, a incidência e mortalidade do CG vêm diminuindo gradualmente em todo o mundo. No entanto, ainda existem certos desafios na gestão do CG, como as aplicações clínicas do tratamento cirúrgico e da quimioterapia (YANG *et al*., 2023).

Este artigo foi baseado no estudo de Zhao *et al*. (2023), no qual, a partir do scRNA-seq, foi identificado o receptor de quimiocina tipo C-X-C 4 (CXCR4) como um gene-chave no crescimento e metástase do câncer gástrico. O CXCR4 é identificado como um receptor de quimiocina envolvido em múltiplas condições patológicas, como doenças imunológicas e câncer. No processo de identificação do receptor CXCR4 foi utilizado o algoritmo K-médias, que foi aplicado para remover células não malignas de um conjunto de células epiteliais, permitindo a identificação e análise das células malignas restantes (ZHAO *et al*., 2023). O K-médias é um algoritmo de agrupamento que organiza dados em grupos (clusters) de forma que os elementos dentro de um grupo sejam mais semelhantes entre si do que com elementos de outros grupos (YU; DONG; YAO, 2022). O objetivo deste estudo é analisar o uso do algoritmo K-médias para dividir em clusters distintos células epiteliais malignas e não malignas, buscando reproduzir o experimento feito por Zhao *et al*. (2023).

O uso de algoritmos como o K-médias no processo de análise e identificação do receptor CXCR4 é importante para o tratamento de câncer gástrico porque ele desempenha um papel crucial no crescimento e na metástase das células-tronco cancerígenas (CSCs) associadas ao câncer gástrico de fenótipo maligno, que possui maior invasão e potencial metastático. Logo, utilizar algoritmos para melhorar a identificação deste e de outros receptores relacionados ao câncer gástrico pode auxiliar na busca e implementação de tratamentos, melhorando a qualidade de vida dos pacientes acometidos pela doença (ZHAO *et al*., 2023).

2 FUNDAMENTAÇÃO TEÓRICA

O câncer gástrico (CG) é uma doença altamente heterogênea, tanto molecular quanto fenotipicamente (ZHAO *et al*., 2023), tendo origem multifatorial e é caracterizada pela desordenada multiplicação de células da parede do órgão (LEE, 2019). O desenvolvimento rápido, a metástase distante e a resistência à quimioterapia são características marcantes do CG, em grande parte devido à heterogeneidade tumoral (REYA *et al*., 2001). As células-tronco cancerosas foram definidas como uma célula dentro de um tumor que possui a capacidade de se autorrenovar e de gerar as linhagens heterogêneas de células cancerosas que compõem o tumor (CLARKE *et al*., 2006).

As CSCs, por sua vez, são reconhecidas como impulsionadoras principais do crescimento e metástase do CG, indicando que estratégias terapêuticas que visam essas células podem ser cruciais para o controle da doença (BEKAII‐SAAB; EL‐RAYES, 2017). O microambiente tumoral (TME) também desempenha um papel crítico na progressão do câncer, influenciando a resposta imunológica e, consequentemente, o prognóstico dos pacientes. Estudos mostram que um TME rico em células estromais e imunológicas está associado a uma resposta imunológica mais robusta, o que pode impactar positivamente os resultados terapêuticos (WEN *et al*., 2022).

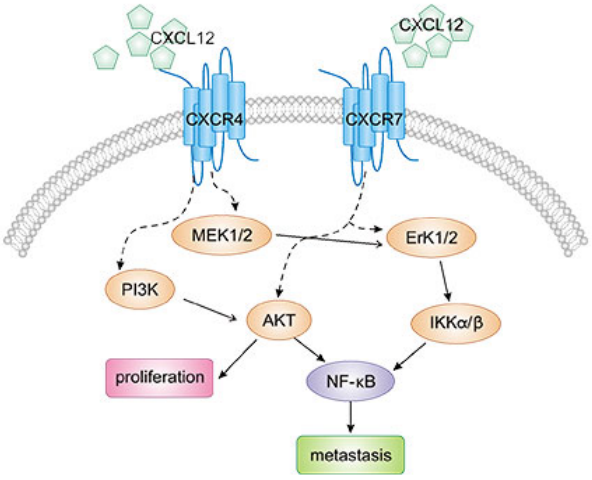
Portanto, estratégias terapêuticas para atingir as CSCs são imperativas para erradicar a doença residual e prevenir a recorrência não apenas no CG, mas também em outros tipos de câncer. Além disso, com os atuais insights sobre os mecanismos celulares da carcinogênese, o potencial metastático das células tumorais é atribuído às CSCs, que podem iniciar clonalmente a formação de tumores em locais distantes (FUJITA *et al*., 2015). Dessa forma, desempenham um papel crucial no diagnóstico, tratamento e prognóstico do câncer (BARBATO *et al*., 2019)

O gene CXCR4 tem emergido como um importante marcador molecular no contexto do câncer gástrico, devido à sua significativa associação com a progressão e prognóstico da doença. Estudos recentes sugerem que a expressão de CXCR4 está fortemente ligada à tumorigênese e à metástase hepática, e que a sua inibição pode suprimir fenótipos malignos das células-tronco cancerígenas (CSCs), retardando, assim, o avanço do tumor e a disseminação metastática (ZHAO *et al*., 2023). Além disso, o CXCR4 foi identificado como um marcador prognóstico importante para pacientes com CG, o que reforça sua relevância no desenvolvimento de terapias direcionadas (ZHAO *et al*., 2023).

O CXCR4 não só está relacionado à metástase e a uma menor sobrevida global (OS) em vários tipos de câncer, mas também é um alvo terapêutico promissor, especialmente no CG, onde sua expressão está associada ao TME e à infiltração de células imunológicas, como células B naïve, células T CD8+ e células NK (WEN *et al*., 2022). Inibidores de CXCR4 mostraram eficácia no aumento da atividade antitumoral, indicando um potencial para estratégias combinadas no tratamento de cânceres graves, incluindo o CG (WEN *et al*., 2022).

O CXCL12 é uma quimiocina, uma molécula sinalizadora que exerce um papel crucial na regulação da migração e localização de várias células no organismo. O CXCR4 é o receptor específico para o CXCL12 (DANIEL; SEO; PILLARISETTY, 2020). Quando o CXCL12 se liga ao CXCR4, ocorre a ativação de uma série de vias de sinalização intracelular que podem influenciar diversos processos biológicos, incluindo o crescimento celular, a sobrevivência, a invasão e a migração (DANIEL; SEO; PILLARISETTY, 2020).

Figura 1: CXCR4/CXCL12 participa de ativação de diversas vias em células cancerígenas.



Fonte: CUSABIO team (s.d.).

Finalmente, a introdução de terapias que visam o eixo CXCR4-CXCL12 mostrou promissores resultados pré-clínicos, incluindo a repressão do crescimento metastático e a regulação do microambiente imunológico, o que pode beneficiar significativamente os pacientes com câncer gástrico (WEN *et al*., 2022). A utilização de marcadores imunológicos prognósticos baseados em imunomoduladores relacionados ao CXCR4 pode proporcionar uma abordagem mais personalizada e eficaz para prever a sobrevida global e direcionar o tratamento dos pacientes com câncer gástrico (WEN *et al*., 2022).

3 DESCRIÇÃO DA ABORDAGEM COMPUTACIONAL

No contexto da distinção entre células epiteliais malignas e não malignas em amostras de câncer gástrico, Zhao *et al* (2023) utilizaram o algoritmo de K-médias como parte de uma análise mais ampla, com o objetivo de revelar mecanismos moleculares subjacentes ao crescimento e metástase do câncer gástrico. Esta abordagem permitiu a investigação das características malignas e suas influências no desenvolvimento e progressão do câncer gástrico e por isso, essa foi a estratégia iterativa escolhida para ser replicada no atual trabalho.

De acordo com Cristianini e Hahn (2006), uma das melhores maneiras de se reconhecer padrões, ou nesse caso, reconhecer malignidade ou não num grupo de células, é usando agrupamento. Ou seja, a identificação de subconjuntos de dados que possuem alta similaridade interna, assim como baixa similaridade entre clusters.

Ainda, esses autores demonstram que a definição de clusters nos dados depende de uma medida de distância específica, calculada pela Correlação de Pearson ou pela distância Euclidiana e que será usada entre dois perfis de expressão gênica.

Depois, também se faz necessário definir um critério pelo qual o conjunto de dados será particionado em clusters, que permite definir quais pontos devem ser considerados como pertencentes ao mesmo cluster. Uma forma simples de fazer isso é subtraindo as distâncias inter-clustering (desejando que sejam as maiores possíveis) da distância intra-clusters (desejando que saiam às menores possíveis).

Uma vez que a medida de distância tenha sido estabelecida e um critério de clustering tenha sido definido, ainda é necessário buscar entre todas as possíveis partições do conjunto de dados para encontrar aquela que maximize o critério de clustering. Para isso, Zhao *et al* (2023) escolheram o algoritmo de K-médias.

A abordagem do k-médias em si começa escolhendo quantos clusters queremos obter do nosso conjunto de dados e esse número é chamado de k. Os clusters são definidos implicitamente especificando k pontos chamados centros ou protótipos. Depois, cada centro é substituído pelas médias dos clusters definidos por eles. Os novos centros especificam novos clusters e, portanto, novas médias, e assim por diante, até que o algoritmo convirja. O resultado é um mínimo local do critério de agrupamento mencionado acima. Uma vantagem desse método é que também podemos plotar os centróides como uma forma de representar os clusters e resumir os dados (Cristianini; Hahn, 2006).

4 DESCRIÇÃO DAS FERRAMENTAS

Neste estudo foi desenvolvido um script em Python que utiliza várias ferramentas de bioinformática para manipulação e análise de sequências de RNA de célula única (scRNA-seq). As ferramentas utilizadas no script, bem como o seu propósito no contexto de processamento das sequências, são descritas conforme a ordem de uso.

4.1 BASE DE DADOS GEO

A base de dados Gene Expression Omnibus (GEO) foi utilizada para obter os dados de expressão gênica do estudo (JIANG et al., 2022). A amostra específica utilizada foi a GSE163558, que contém dados de RNA de célula única de tecidos tumorais gástricos.

4.1.1 dataset GSE163558

O dataset GSE163558 foi baixado da base de dados GEO. Este dataset contém dados de RNA de célula única de sete amostras de tumor gástrico, incluindo três amostras de tumor in situ, duas amostras de tumor metastático em linfonodo e duas amostras de tumor metastático em fígado.

4.2 BASE DE DADOS GENCODE

A base de dados GENCODE (GENCODE, 2024) foi utilizada para obter anotações abrangentes dos genes nos cromossomos de referência, o que nos permitiu classificar as células das diferentes amostras em diferentes tipos celulares: NK, T, B, Epitelial, Stromal e Myeloid.

4.3 SCANPY

A biblioteca scanpy (SCANPY, 2024) foi utilizada para análise de dados de RNA de célula única, incluindo normalização, filtragem, e redução de dimensionalidade.

4.4 K-MEANS

O algoritmo K-means foi aplicado para a clusterização dos dados de expressão gênica normalizados.

4.5 INFERCNVPY

A biblioteca infercnvpy (INFERCNVPY, 2024) foi utilizada para a análise de variação do número de cópias (CNV), identificando possíveis células malignas com base nas assinaturas de CNV.

4.6 MATPLOTLIB

A biblioteca matplotlib (MATPLOTLIB... 2024) foi utilizada para gerar visualizações dos dados, como gráficos UMAP e distribuições de clusters.

4.7 PANDAS

A biblioteca pandas (PANDAS... 2024) foi usada para manipulação e análise de dados tabulares, como a leitura de arquivos CSV contendo posições de genes e marcadores celulares.

4.8 SCIKIT-LEARN

A biblioteca scikit-learn (SCIKIT-LEARN, 2024) foi utilizada para aplicar o algoritmo K-means e outras técnicas de pré-processamento de dados, como a normalização.

4.9 OUTRAS CONSIDERAÇÕES

Os scripts e os gráficos resultantes podem ser encontrados no repositório GitHub associado a este projeto, conforme descrito no Apêndice A.

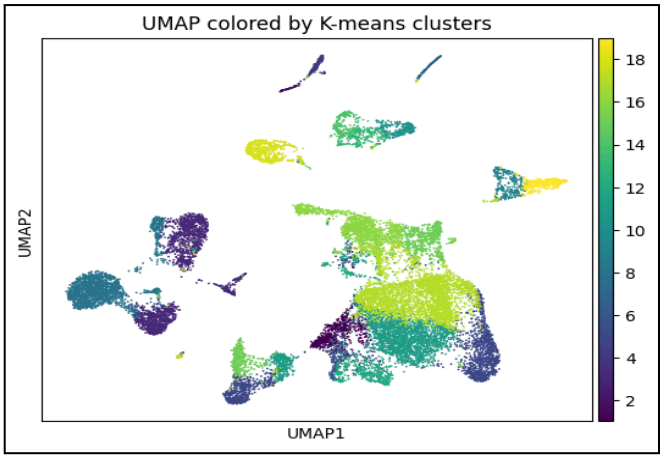
5 RESULTADOS

Nesta seção são apresentados os resultados obtidos a partir da análise dos dados de RNA de célula única (scRNA-seq) do dataset GSE163558. Conforme descrito na seção anterior, a análise foi realizada utilizando técnicas de clusterização com K-means e infercnvpy. Os principais resultados incluem a visualização dos dados em um espaço bidimensional utilizando UMAP e a distribuição dos clusters obtidos com K-means.

5.1 VISUALIZAÇÃO DOS DADOS COM UMAP

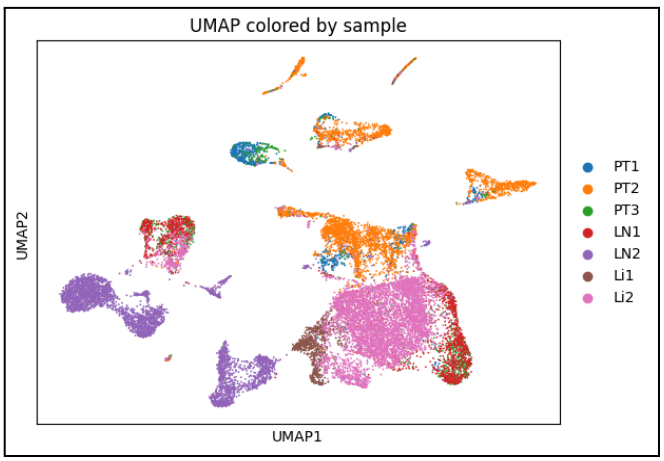
A visualização UMAP dos dados permite uma interpretação intuitiva da estrutura de alto nível dos dados de RNA de célula única. As figuras a seguir mostram a projeção UMAP dos dados colorida pelos clusters de K-means, pelas amostras de origem, pelos tipos celulares identificados e pela distinção entre células malignas e não-malignas (epiteliais).

Figura 2: UMAP colorido por clusters de K-means.



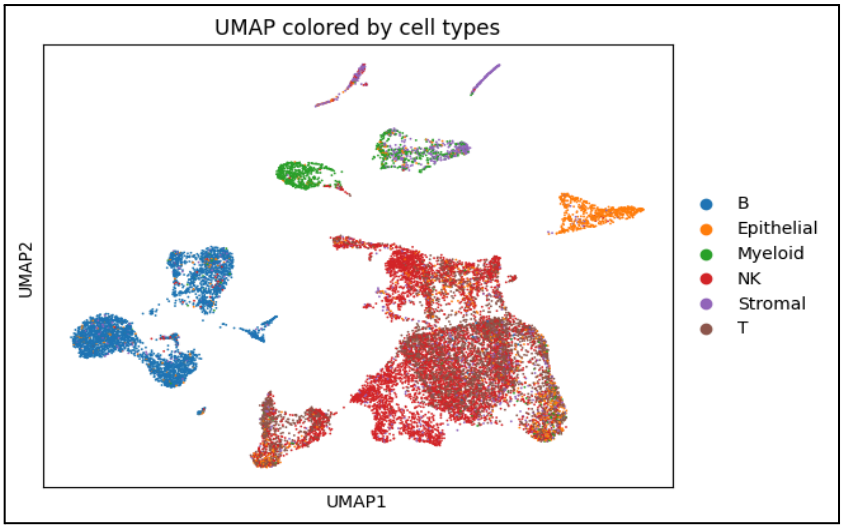
Fonte: Do autor (2024).

Figura 3: UMAP colorido por amostra.



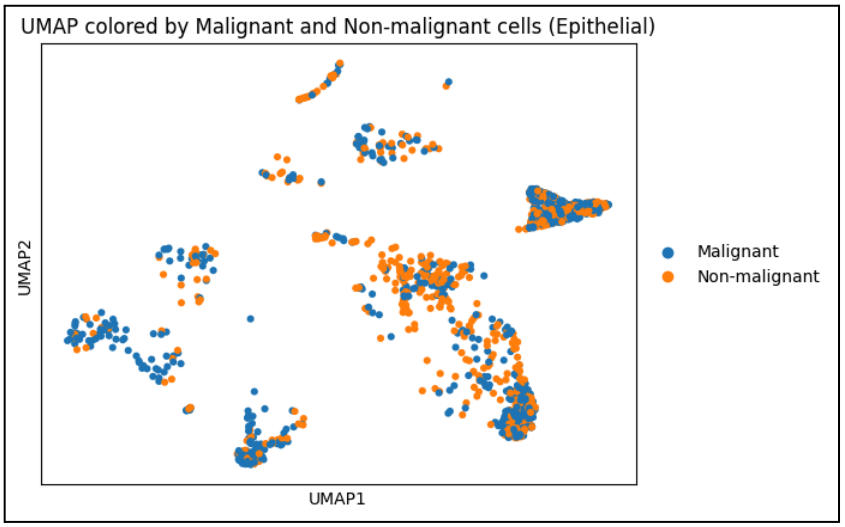
Fonte: Do autor (2024).

Figura 4: UMAP colorido por tipos celulares.



Fonte: Do autor (2024).

Figura 5: UMAP colorido pela distinção entre malignas e não malignas (epiteliais).



Fonte: Do autor (2024).

5.2 ANÁLISE DOS RESULTADOS

A análise de clusterização foi realizada assumindo 19 clusters com base na análise do artigo de referência. As principais análises foram realizadas utilizando visualizações UMAP e clusterização com o algoritmo K-means.

5.2.1 Visualização UMAP

A visualização UMAP foi utilizada para representar os dados em um espaço bidimensional, facilitando a interpretação das relações entre as células. Quatro gráficos UMAP foram gerados:

- **UMAP colorido pelos clusters de K-mean (Figura 2):** Demonstra a separação das células em 19 clusters distintos com base nos perfis de expressão gênica.
- **UMAP colorido por amostra (Figura 3):** Indica a origem das células de acordo com as amostras de tumor gástrico, linfonodo e fígado.
- **UMAP colorido pelos tipos celulares (Figura 4):** Mostra a distribuição de diferentes tipos celulares, incluindo células NK, T, B, epiteliais, estromais e mieloides.
- **UMAP colorido pela distinção entre células malignas e não malignas (epiteliais) (Figura 5):** Mostra a separação entre células epiteliais malignas e não malignas.

5.2.2 Distribuição dos tipos celulares

A análise revelou a presença de diversos tipos celulares no dataset, com as células NK sendo as mais abundantes. A distribuição dos tipos celulares é a seguinte:

```
cell_type
NK            6508
T             4041
B             3682
Epithelial    1563
Stromal       1454
Myeloid       1002
Name: count, dtype: int64
```

5.2.3 Células malignas e não malignas

A análise das células malignas e não malignas em cada tipo celular revelou as seguintes contagens e percentuais de células malignas:

malign	Malignant	Non-malignant	Total	Percent Malignant
cell_type				
B	2476	1206	3682	67.246062
Epithelial	818	745	1563	52.335253
Myeloid	439	563	1002	43.812375
NK	2211	4297	6508	33.973571
Stromal	678	776	1454	46.629986
T	1580	2461	4041	39.099233

As células B apresentaram o maior percentual de malignidade, seguidas por células epiteliais e mieloides.

5.2.4 Interpretação dos clusters

A análise de clusterização K-means indicou uma separação clara entre diferentes grupos de células, refletindo a heterogeneidade celular significativa no câncer gástrico. A maior parte das células foram agrupadas em grandes cluster, sugerindo tipos celulares predominantes.

6 CONCLUSÕES

Neste estudo, foi possível demonstrar a aplicação bem-sucedida das técnicas de bioinformática utilizadas para explorar a heterogeneidade celular no câncer gástrico (CG), a partir dos dados de *single-cell* RNA *sequencing* (scRNA-seq), fornecidos pelo artigo de base. A visualização UMAP e a clusterização pelo algoritmo K-médias, nos gráficos das Figuras 2, 3 e 4, como uma analogia aos gráficos A, C e E, respectivamente, da Figura 1 fornecida por Zhao *et al* (2023), foram utilizadas para analisar a diversidade celular tanto em termos de tipos celulares quanto em termos de origem das amostras (linfonodos ou tecidos primários) de CG.

A Figura 2 representa o agrupamento geral do *dataset* GSE163558 em 19 clusters, com base nos padrões de expressão gênica heterogêneos evidenciados no microambiente tumoral. Esses 19 clusters foram anotados, com base na expressão de genes específicos de cada tipo celular, demonstrando a presença de células T, células mieloides, células NK, células B, células epiteliais e células estromais no tumor (Figura 4). O scRNA-seq possibilitou a determinação de genes altamente variáveis dentro de células homogêneas.

Os resultados indicaram uma clara segregação das células em clusters distintos, refletindo a complexidade e heterogeneidade do CG. Isto se faz muito relevante, já que cada tipo celular desempenha um papel específico na progressão do câncer e na resposta e mecanismos de escape imune em relação ao tumor. Ademais, pode revelar subpopulações de células que são resistentes a terapias ou que possuem capacidades metastáticas elevadas, permitindo o desenvolvimento de estratégias terapêuticas mais eficazes (Zhao *et al,* 2023).

Além disso, conforme visualizado na Figura 3 gerada no trabalho, a origem do microambiente tumoral pode influenciar diretamente a heterogeneidade das células e nos níveis de expressão gênica de cada uma. Essa comparação pode revelar como as células tumorais evoluem e se adaptam ao longo da progressão do câncer, visto que, linfonodos são frequentemente caminhos para a metástase e no CG, o fígado é um importante destino das células malignas, podendo fornecer informações relevantes sobre os mecanismos de disseminação e as diferentes proporções entre os tipos de células envolvidos em cada tecido (Zhao *et al,* 2023).

Como destacado por Zhao *et al* (2023), o CG se origina principalmente do epitélio glandular da mucosa gástrica, e por isso, as células epiteliais do microambiente tumoral foram classificadas entre malignas e não malignas (Figura 5), tal qual foi demonstrado na Figura 3D do artigo de base. Apesar de não bem delimitadas as células malignas das não malignas, como

no artigo de base, devido à falta de informações fornecidas quanto à metodologia utilizada, foi possível ao menos estabelecer uma proporção a partir de K-médias.

Esse agrupamento se faz essencial para a avaliação de características e marcadores que as diferenciam, como a expressão de CXCR4, correlacionada com a alta capacidade metastática, resistência a quimioterápicos e pior prognóstico, além de marcar, entre as células malignas, a presença de células-tronco cancerosas, alvos importantes para o tratamento do CG (Zhao *et al,* 2023).

Vale ressaltar também, que os padrões dos clusters obtidos nos gráficos reproduzidos neste estudo são diferentes dos observados no artigo base, e isso foi atribuído ao fato de que foram utilizados ferramentas, biblioteca, linguagem e padrões de processamento e filtragem dos dados diferentes dos que possivelmente foram realizados no estudo de Zhao *et al* (2023), não fornecidos em detalhes. Mesmo assim, com exceção da divisão entre as células malignas e não malignas, foi possível estabelecer uma boa correlação entre os clusters reproduzidos e o artigo de base, e demonstrar que K-médias é uma ótima ferramenta para os fins apresentados, visto que, há agrupamentos bem coesos e divididos com base em suas assinaturas de expressão gênica, como indicativo do bom funcionamento do algoritmo.

REFERÊNCIAS BIBLIOGRÁFICAS

BARBATO, Luisa; BOCCHETTI, Marco; BIASE, Anna di; REGAD, Tarik. **Cancer Stem Cells and Targeting Strategies.** Cells, [S.L.], v. 8, n. 8, p. 926, 18 ago. 2019. MDPI AG. http://dx.doi.org/10.3390/cells8080926.

BEKAII‐SAAB, Tanios; EL‐RAYES, Bassel. **Identifying and targeting cancer stem cells in the treatment of gastric cancer.** Cancer, [S.L.], v. 123, n. 8, p. 1303-1312, 24 jan. 2017. Wiley. http://dx.doi.org/10.1002/cncr.30538.

CUSABIO team. **CXCR4: An Attractive Target of GPCR family Brings Promising New Drugs for Cancer Therapy**. Disponível em: https://www.cusabio.com/c-21062.html. Acesso em: 31 jul. 2024.

CLARKE, Michael F.; DICK, John E.; DIRKS, Peter B.; EAVES, Connie J.; JAMIESON, Catriona H.M.; JONES, D. Leanne; VISVADER, Jane; WEISSMAN, Irving L.; WAHL, Geoffrey M. **Cancer Stem Cells—Perspectives on Current Status and Future Directions: aacr workshop on cancer stem cells**. Cancer Research, [S.L.], v. 66, n. 19, p. 9339-9344, 1 out. 2006. American Association for Cancer Research (AACR). http://dx.doi.org/10.1158/0008-5472.can-06-3126.

DANIEL, Sara K.; SEO, Y. David; PILLARISETTY, Venu G. **The CXCL12-CXCR4/CXCR7 axis as a mechanism of immune resistance in gastrointestinal malignancies**. Seminars In Cancer Biology, [S.L.], v. 65, p. 176-188, out. 2020. Elsevier BV. http://dx.doi.org/10.1016/j.semcancer.2019.12.007.

FUJITA, Takeshi; CHIWAKI, Fumiko; TAKAHASHI, Ryou-U; AOYAGI, Kazuhiko; YANAGIHARA, Kazuyoshi; NISHIMURA, Takao; TAMAOKI, Masashi; KOMATSU, Masayuki; KOMATSUZAKI, Rie; MATSUSAKI, Keisuke. **Identification and Characterization of CXCR4-Positive Gastric Cancer Stem Cells.** Plos One, [S.L.], v. 10, n. 6, p. 1-19, 25 jun. 2015. Public Library of Science (PLoS). http://dx.doi.org/10.1371/journal.pone.0130808.

GENCODE. 2024. Disponível em: https://www.gencodegenes.org/. Acesso em: 07 ago. 2024.

INFERCNVPY: Scanpy plugin to infer copy number variation (CNV) from single-cell transcriptomics data. Scanpy plugin to infer copy number variation (CNV) from single-cell transcriptomics data. 2024. Disponível em: https://infercnvpy.readthedocs.io/en/latest/#contact. Acesso em: 07 ago. 2024.

JIANG H, YU D, YANG P, Guo R et al. **Revealing the transcriptional heterogeneity of organ-specific metastasis in human gastric cancer using single-cell RNA Sequencing**. *Clin Transl Med* 2022 Feb;12(2):e730. PMID: 35184420

LEE, Ohana Peres; CESARIO, Fabiana Copês. **Relação entre escolhas alimentares e o desenvolvimento de câncer gástrico: uma revisão sistemática**. Brazilian Journal of Health Review, v. 2, n. 4, p. 2640-2656, 2019.

MACHLOWSKA, Julita; BAJ, Jacek; SITARZ, Monika; MACIEJEWSKI, Ryszard; SITARZ, Robert. **Gastric Cancer: Epidemiology, Risk Factors, Classification, Genomic Characteristics and Treatment Strategies.** International Journal of Molecular Sciences, v. 21, n. 11, p. 4012, 4 jun. 2020.

MATPLOTLIB 3.9.1 documentation. 2024. Disponível em: https://matplotlib.org/stable/index.html. Acesso em: 07 ago. 2024.

PANDAS documentation. 2024. Disponível em: https://pandas.pydata.org/docs/. Acesso em: 07 ago. 2024.

REYA, Tannishtha et al. **Stem cells, cancer, and cancer stem cells**. Nature, [S.L.], v. 414, n. 6859, p. 105-111, nov. 2001. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/35102167.

SCANPY. 2024. Disponível em: https://scanpy.readthedocs.io/en/stable/references.html. Acesso em: 07 ago. 2024.

SCIKIT-LEARN: Machine Learning in Python. Machine Learning in Python. 2024. Disponível em: https://scikit-learn.org/stable/. Acesso em: 07 ago. 2024.

WEN, Fang; LU, Xiaona; HUANG, Wenjie; CHEN, Xiaoxue; RUAN, Shuai; GU, Suping; GU, Peixing; LI, Ye; LIU, Jiatong; LIU, Shenlin. **Characteristics of immunophenotypes and immunological in tumor microenvironment and analysis of immune implication of CXCR4 in gastric cancer**. Scientific Reports, [S.L.], v. 12, n. 1, p. 1-12, 6 abr. 2022. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-022-08622-1.

YANG, Wen-Juan; ZHAO, He-Ping; YU, Yan; WANG, Ji-Han; GUO, Lei; LIU, Jun-Ye; PU, Jie; LV, Jing. **Updates on global epidemiology, risk and prognostic factors of gastric cancer**. World Journal of Gastroenterology, v. 29, n. 16, p. 2452–2468, 28 abr. 2023.

YU, Donghua; DONG, Shuhua; YAO, Shuang. **Improvement of K-Means Algorithm and Its Application in Air Passenger Grouping.** Computational Intelligence and Neuroscience, v. 2022, p. 1–13, 12 set. 2022.

ZHAO, Hongying; JIANG, Rongke; ZHANG, Chunmei; FENG, Zhijing; WANG, Xue. **The regulatory role of cancer stem cell marker gene CXCR4 in the growth and metastasis of gastric cancer.** npj Precision Oncology, v. 7, n. 1, p. 86, 7 set. 2023.

APÊNDICE A - GitHub

Para uma melhor organização do trabalho, foi criado um repositório no GitHub com informações adicionais do projeto, este repositório contém os arquivos e o script utilizados para o experimento. O link está disponível a seguir:

https://github.com/lcsmuller/CI1169_Bioinformatics