UNIVERSIDADE DE SÃO PAULO
FACULDADE DE MEDICINA DE RIBEIRÃO PRETO
GRADUAÇÃO EM INFORMÁTICA BIOMÉDICA

ISABELA DIAS ERTHAL

IDENTIFICAÇÃO DE GENES MARCADORES DE PROGRESSÃO EM CÂNCER DE BEXIGA NÃO-MÚSCULO INVASIVO

RIBEIRÃO PRETO, SP
2022

ISABELA DIAS ERTHAL

IDENTIFICAÇÃO DE GENES MARCADORES DE PROGRESSÃO EM CÂNCER DE BEXIGA NÃO-MÚSCULO INVASIVO

Monografia do Trabalho de Conclusão de Curso apresentada à Faculdade de Medicina de Ribeirão Preto como requisito para conclusão do curso de graduação em Informática Biomédica.

Orientadora: Dra. Tathiane Maistro Malta
Coorientador: Maycon Marção

Este trabalho foi apresentado e aprovado pela Comissão Coordenadora do Curso deInformática Biomédica em 15/03/2023.

RIBEIRÃO PRETO, SP
2022

AGRADECIMENTOS

Primeiramente, agradeço a minha família. Todo o apoio que recebo desde o meu primeiro minuto como universitária foi essencial para a conclusão desta etapa da minha vida. Obrigada por não deixarem eu me sentir sozinha, mesmo estando a quilômetros de distância de casa, e obrigada pela companhia ao vivo e em cores no isolamento durante a pandemia, que se estendeu por metade da minha graduação.

Agradeço a minha orientadora Tathi, que me acompanha desde o meu primeiro projeto como aluna de iniciação científica. Sou muito sortuda de estar sendo orientada por uma pesquisadora tão dedicada, competente e generosa quanto você. Obrigada por todo o aprendizado e por me incentivar a fazer ciência.

Também agradeço a todos meus colegas de laboratório, por me auxiliarem na minha evolução durante esses anos. Fico muito feliz em fazer parte desse grupo de pessoas incríveis, dedicadas e divertidas.

Aos amigos que fiz na faculdade, obrigada pela companhia, crises, reclamações, comemorações, tudo de bom e duvidoso que passamos juntos até aqui. Agradeço em especial aos meus amigos Larissa, Pedro Emílio, Gabriel, Felipe, Carlos, Bruno, Karen e Ninna.

Obrigada também aos meus melhores amigos de Goiânia, por sempre me receberem de braços abertos, semestralmente, em suas casas e seguirem até hoje como meus amigos de longa data apesar da distância. Em especial, à Sarah, Pedro Melo e Pedro Lopes.

Agradeço a minha namorada, por estar ao meu lado desde o meu primeiro ano e me ajudando a passar por tudo até os dias de hoje. Ter você comigo durante esse momento fez tudo se tornar mais leve.

E por fim, agradeço à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo financiamento do projeto.

RESUMO

O câncer de bexiga pode ser classificado entre dois tipos: câncer de bexiga não-músculo invasivo (CBNMI) e o câncer de bexiga músculo invasivo (CBMI). A invasão do câncer de bexiga para o músculo é um fator grave e gera um tumor mais agressivo comparado ao anterior. Atualmente, com a ascensão da bioinformática, há uma variedade de dados disponíveis *online* dos mais diversos contextos biológicos, além de ferramentas computacionais próprias para realização de análises relevantes para a área e, principalmente, para a oncologia. Este projeto teve como proposta trabalhar na identificação de genes que possam predizer a progressão do tumor de bexiga não-músculo invasivo para o tipo músculo invasivo. Isso foi feito com dados de expressão gênica, através de análise de expressão gênica diferencial e análise de enriquecimento de conjunto de genes, a partir de metodologias de bioinformática, como ferramentas computacionais e códigos na linguagem de programação R. Com o alcance do objetivo do trabalho, foram identificados 11 genes diferencialmente expressos que possuem relação com o fuso mitótico, a resposta inflamatória e a transição epitélio-mesenquimal no CBNMI, em destaque os genes *INHBA*, *PLAUR* e *FLNA*. Estudos adicionais podem ser realizados para avaliar o potencial destes genes em predizer a progressão do CBNMI.

Palavras-chave: Bioinformática; Transcriptômica; Câncer de bexiga.

ABSTRACT

Bladder cancer can be classified into two types: non-muscle invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). The invasion of bladder cancer into the muscle is a severe factor and leads to a more aggressive tumor than the previous one. Currently, with the rise of bioinformatics, there is a variety of data available online from the most diverse biological contexts, in addition to computational tools suitable for carrying out relevant analyzes for the area and, mainly, for oncology. This project aimed to work on the identification of genes that can predict the progression of the non-muscle invasive bladder tumor to the muscle invasive type. This was done with gene expression data, through differential gene expression analysis and gene set enrichment analysis, from bioinformatics methodologies, such as computational tools and scripts in the R programming language. When the purpose of this work was achieved, 11 differentially expressed genes were identified as being related to the mitotic spindle, inflammatory response, and epithelial-mesenchymal transition in NMIBC, highlighting the *INHBA*, *PLAUR,* and *FLNA* genes. Additional studies can be performed to evaluate the potential of these genes in predicting the progression of NMIBC.

Keywords: Bioinformatics; Transcriptomics; Bladder cancer.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CBMI	Câncer de bexiga músculo invasivo
CBNMI	Câncer de bexiga não-músculo invasivo
CIS	Carcinoma *in situ*
EGA	*European Genome-Phenome Archive*
ES	*Enrichment Score*
GSEA	*Gene set enrichment analysis*
INCA	Instituto Nacional de Câncer
LEA	*Leading Edge Analysis*
MSigDB	*Molecular Signature Database*
NGS	*Next Generation Sequencing*
PCA	*Principal Component Analysis*
EMT	Transição epitélio-mesenquimal

Sumário

1 Introdução

1.1 Câncer de bexiga

O câncer é uma doença genética caracterizada principalmente pelo crescimento descontrolado de células que levam à formação de tumores malignos. O câncer de bexiga atinge as células do órgão, que tem a função de armazenar a urina produzida pelos rins, e tem como principais sintomas a dor durante o ato de urinar, necessidade frequente de urinar e presença de sangue na urina (hematúria) (FARLING, 2017).

O câncer de bexiga não-músculo invasivo (CBNMI) é o tipo mais comum de câncer de bexiga. É um tumor bastante heterogêneo (LINDSKROG et al., 2021) e consequentemente possui diversos prognósticos, sendo um deles, a progressão para o tipo músculo invasivo.

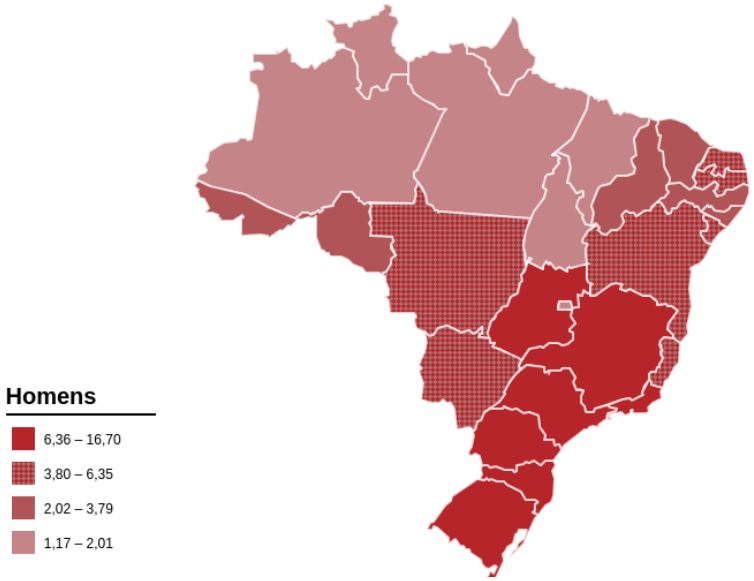
O câncer de bexiga músculo invasivo (CBMI) representa de 10% a 20% dos casos de CBNMI que progrediram e é um tumor mais agressivo, com cerca de 60% dos pacientes apresentando um índice de sobrevivência de apenas 5 anos. Em nível molecular, já foi caracterizado por sua instabilidade genômica e uma alta taxa de mutações (KAMOUN et al., 2020).

1.1.2 Epidemiologia

De acordo com INCA, o câncer de bexiga é o 7º câncer mais incidente entre os homens e o 14º em mulheres. Também foi apontado que homens brancos e de idade avançada são o grupo de maior risco de desenvolver o câncer de bexiga (INSTITUTO NACIONAL DE CÂNCER, 2022)..

No Brasil, os estados com as maiores taxas de incidência estimadas para o ano de 2020 foram Goiás, Minas Gerais, Rio de Janeiro, São Paulo, Paraná, Santa Catarina e Rio Grande do Sul (Figura 1), com taxas entre 6% e 16%.

Figura 1 - Representação espacial das taxas brutas de incidência por 100 mil homens, estimadas para o ano de 2020, segundo Unidade da Federação



Fonte: INCA, 2022

1.1.3 Estágios tumorais

O estágio tumoral é uma forma de medir e descrever como o câncer cresce. Essa informação é importante para compreender a progressão do câncer de bexiga para a camada muscular, por exemplo. Os estágios do tumor estão descritos no Quadro 1, junto com a informação do tipo do câncer baseado na presença ou não da progressão ao músculo da bexiga.

Quadro 1 - Estágios tumorais do câncer de bexiga e seus respectivos tipos baseados na progressão

Estágio	Descrição	Tipo
Ta	Tumor no revestimento da bexiga que não invade nenhuma das camadas da bexiga	CBNMI
Tis	Carcinoma *in Situ* (CIS) - Um câncer de alto grau. Parece um pedaço avermelhado e aveludado no revestimento da bexiga	
T1	O tumor atravessa o revestimento da bexiga, mas não atinge a camada muscular	
T2	O tumor cresce na camada muscular da bexiga	CBMI

T3	O tumor passa através da camada muscular para os tecidos que cercam a bexiga	
T4	O tumor se espalhou para estruturas próximas	

Fonte: Adaptado de Urology Care Foundation, 2017

Dessa forma, observa-se que a progressão do câncer de bexiga é algo gradual, passando por 3 estágios tumorais antes de se tornar CBMI definitivamente. Esses estágios devem ser melhor explorados molecularmente, para que novas terapias sejam direcionadas para precaver esse tipo de progressão.

Ainda dentro desse contexto de subgrupos, um estudo que analisou o transcriptoma de pacientes com câncer de bexiga dividiu suas amostras em 4 perfis diferentes baseados na biologia do tumor e sua agressividade. Os perfis foram nomeados de classe 1, 2a, 2b e 3, sendo que as classes 2a e 2b são as que mais progridem para o CBMI (LINDSKROG et al., 2021).

1.2 Expressão gênica

A transcrição é uma das duas etapas do dogma central da biologia molecular. Ela é definida pela transformação do DNA em RNA, que posteriormente será traduzido para proteína e exercerá sua função biológica, passando pela etapa de tradução. Essas etapas são importantes para compreender como funciona o processo de expressão gênica.

Entender o controle da expressão gênica é crucial para interpretar a relação entre genótipo e fenótipo. A expressão gênica é caracterizada pela constituição das características fisiológicas e fenotípicas de um ser vivo, isto é, as informações genéticas são interpretadas pela expressão gênica e o produto disso é o fenótipo. Esse fenótipo pode ser variável conforme o nível de expressão desses genes e, com isso, pode levar a uma variedade de doenças no ser humano, sendo o câncer uma delas.

A desregulação da expressão gênica é algo bastante observado em estudos de tumores, isso porque essas alterações estão diretamente relacionadas a um dos marcos do câncer, a desregulação epigenética (HANAHAN, 2022).

1.2.1 Dados de expressão gênica

Para a análise desses níveis de expressão gênica, surgiram alguns tipos de dados que são processados para essa finalidade, sendo os de *microarray* e RNA-seq os mais conhecidos e difundidos na área de transcriptômica atualmente. Ambos possuem suas próprias tecnologias e suas particularidades. As informações de expressão de genes em *microarrays* são retiradas de uma abordagem baseada em hibridização, e as de RNA-seq são baseadas em sequenciamentos de nova geração (NGS), englobando os métodos mais atuais de sequenciamento (STEFANO, 2014). No Quadro 2, observa-se as principais diferenças entre dados de *microarrays* e RNA-seq, realçando suas vantagens e desvantagens.

Quadro 2 - Diferenças entre *microarray* e RNA-seq

	Microarrays	RNA-seq
Abordagem	Hibridização	NGS
Ruído nos dados	Alto	Baixo
Capacidade de diferenciar isoformas diferentes	Não	Sim
Quantidade de RNA requerido	Alto	Baixo
Custo para mapear transcriptomas de genomas grandes	Alto	Relativamente baixo

Fonte: Adaptado de WANG; GERSTEIN; SNYDER (2009)

1.2.2 Análise de dados de expressão gênica diferencial

Hoje, na bioinformática, a análise de expressão gênica diferencial já é consolidada e bem utilizada. Esse tipo de análise consiste em comparar níveis de expressão de genes entre

dois grupos fenotípicos diferentes, por exemplo, um grupo saudável e outro patológico, e isso pode auxiliar em novas descobertas de biomarcadores para o desenvolvimento de tratamentos personalizados aos pacientes com aquela patologia.

Para realizar esse tipo de análise, é majoritariamente utilizada a linguagem de programação R. Com o R, pesquisadores e programadores criaram pacotes de funções que performam testes estatísticos necessários para a análise e disponibilizam através do Bioconductor (GENTLEMAN et al., 2004), um *software* aberto para biologia computacional e bioinformática que armazena mais de 2000 pacotes. Atualmente, os pacotes mais conhecidos para a análise de expressão gênica diferencial são o *limma* (RITCHIE et al., 2015), EdgeR (ROBINSON; MCCARTHY; SMYTH, 2009) e o DESeq2 (LOVE; HUBER; ANDERS, 2014).

1.2.2.1 Análise de enriquecimento

Uma outra forma de analisar dados de expressão gênica é identificar grupos de genes com funções similares, os quais têm sua expressão enriquecida em algum fenótipo quando comparado com outro. A ferramenta mais conhecida para esse tipo de análise é o GSEA (SUBRAMANIAN et al., 2005), que indica conjuntos de genes pertencentes a um processo biológico através da expressão diferencial de todo o transcriptoma entre duas situações fenotípicas (SHI; WALKER, 2007). A ferramenta também oferece a *Leading Edge Analysis* (LEA), que identifica os genes mais significativos em cada grupo gênico enriquecido.

A hipótese inicial para a realização deste trabalho é que há genes relacionados à progressão do CBNMI e que é possível identificá-los a partir de sua expressão gênica e predizer se o tumor irá progredir para o tipo músculo invasivo.

2 Objetivo

O objetivo principal deste trabalho foi investigar e identificar genes e grupos gênicos relacionados à progressão do CBNMI ao músculo invasivo a partir de dados de transcriptoma e ferramentas de bioinformática.

2.1 Objetivos específicos

1. Realizar análise de expressão gênica diferencial entre tumores que progrediram e não progrediram, a fim de identificar possíveis genes biomarcadores;
2. Realizar análise de enriquecimento de conjunto de genes entre as mesmas amostras do item 1, a fim de encontrar grupos de genes com funções similares que estão enriquecidos para o fenótipo de progressão tumoral;
3. Realizar uma análise comparativa entre os genes encontrados no item 1 e item 2 a fim de checar a confluência dos achados e filtrar os resultados significativos.

3 Materiais e Métodos

3.1 Base de dados

Para esse trabalho, foram utilizados dados de expressão gênica e seus respectivos dados clínicos (*phenoData*) de amostras de CBNMI, onde foram encontradas informações sobre a progressão do tumor.

Esse material de CBNMI foi acessado em um estudo que realizou uma análise integrativa multi-ômica desse tumor, armazenou os dados utilizados na base de dados EGA através do ID EGAS00001004693 e nos materiais suplementares do artigo (LINDSKROG et al., 2021). Para a matriz de expressão gênica bruta, foi necessário solicitar o acesso por *e-mail* para o responsável pelo dado.

Mais detalhes sobre o *dataset* que foi utilizado estão na Tabela 1. É válido mencionar que das 535 amostras de CBNMI, tem-se a informação no *phenoData* que 465 não progrediram em CBMI e 65 progrediram para CBMI. Nesse contexto, existem 5 amostras que não possuem informação a respeito de sua progressão, e portanto, não foram incluídas no desenvolvimento do trabalho.

Tabela 1 - Informações adicionais sobre o *dataset* selecionado para o trabalho

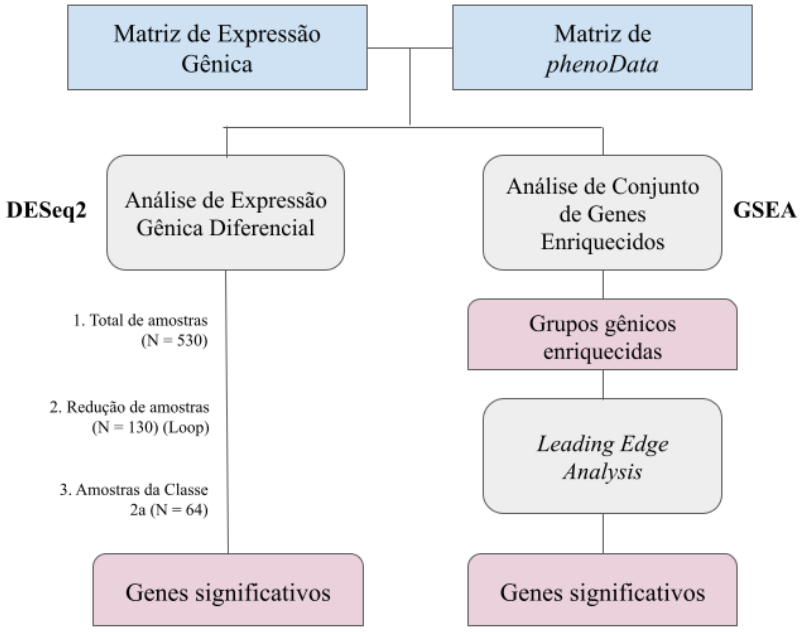
EGA ID	Tipo	Plataforma de Sequenciamento	Número de Amostras	*phenoData* disponível
EGAS00001004693	RNA-seq	Illumina HiSeq 2000, Illumina MiSeq, Illumina NovaSeq 6000, NextSeq 550	530	Sim

Fonte: Autoria própria

3.2 Fluxo de trabalho

Como ilustrado na Figura 2, a abordagem aplicada neste trabalho se baseou na utilização do pacote de análise de expressão gênica diferencial DESeq2 (LOVE; HUBER; ANDERS, 2014) e do *software* de análise de enriquecimento de conjunto de genes, o GSEA (SHI; WALKER, 2007).

Figura 2 - Fluxo de trabalho dos métodos utilizados



Fonte: Autoria própria

Primeiramente, foi feita a análise de expressão gênica diferencial, através do DESeq2. As amostras de CBNMI utilizadas nesta etapa foram divididas em dois grupos, principalmente: com e sem progressão. Essa informação é retirada da matriz de *phenoData* e relacionada com a matriz de expressão gênica, e foram feitas 3 abordagens de análises, como mostra a Figura 2. O resultado final esperado desta fase é uma lista de genes diferencialmente expressos.

Em paralelo, foi utilizado o *software* GSEA para identificar os conjuntos de genes que mais estão enriquecidos em um dos dois grupos da etapa anterior. O *input* para execução do *software* foi a matriz de expressão gênica normalizada e um arquivo de *phenoData*, como

indicado na documentação da ferramenta. O resultado esperado desta etapa são os conjuntos gênicos com maior *enrichment score* (ES), ou seja, que mais estão enriquecidos entre os grupos de estudo.

Além da análise de enriquecimento de conjunto de genes, também foi feita, nesse mesmo *software*, a *Leading Edge Analysis* (LEA), que permite identificar quais genes em um determinado conjunto gênico enriquecido mais colaboraram para o valor do ES nesse conjunto. Portanto, para essa etapa, foram utilizados os grupos com maiores ES.

3.3 Análise de Expressão Gênica Diferencial (DESeq2)

O DESeq2 oferece uma análise de expressão gênica diferencial baseada na distribuição binomial (LOVE; HUBER; ANDERS, 2014). É um pacote indicado para dados do tipo RNA-seq, como é o caso do projeto, e deve ser executado através da linguagem de programação R.

Nesta etapa, foi utilizada a matriz de expressão gênica com os níveis brutos (não normalizados) e a tabela com a informação adicional acerca da progressão de cada amostra presente na matriz de expressão. Essas informações foram importantes para a execução do *pipeline* do pacote, em que a função *DESeqDataSetFromMatrix()* recebe a matriz de expressão bruta e uma tabela especificando as duas classes que serão comparadas na análise (tumor progressivo e não-progressivo). Dessa forma, a dimensão inicial da matriz de expressão gênica é de 530 colunas (amostras) e 57165 linhas (genes/transcritos), e a dimensão da tabela de informação adicional é de 3 colunas (Progressão, Classe e Grau) e 530 linhas (amostras).

Antes de rodar a análise do pacote, a matriz de expressão gênica passou por um filtro para retirada de genes com níveis muito baixos de expressão, chamados *low counts*. Para isso, foi calculado a soma dos níveis de expressão de cada gene em todas as amostras e retirado os

genes em que essa soma era menor que a média desse valor para todos os genes. Após esse filtro, o número de genes caiu para 8568.

Como foi mostrado na Figura 2, foram realizadas 3 abordagens de análises. A primeira abordagem foi feita com o número total de amostras disponíveis com a informação acerca de sua progressão, portanto, 530 amostras (sendo 465 não-progressivas e 65 progressivas). Entretanto, essa comparação apresenta um desbalanceamento no número de amostras entre os dois grupos, já que o grupo de amostras não-progressivas é aproximadamente 7 vezes maior que o grupo de progressivas.

A segunda abordagem teve como objetivo solucionar o desbalanceamento. Foi feita uma redução no número de amostras não-progressivas para 65, selecionando-se aleatoriamente essas amostras através da função *stratified()* do pacote Splitstackshape. Com essa redução, o número de amostras total foi reduzido para 130, sendo 65 não-progressivas (escolhidas aleatoriamente) e 65 progressivas, e assim foi feita a análise com o DESeq2. No entanto, para que a análise não fosse feita em cima apenas de um único subgrupo escolhido aleatoriamente pela função, foi criado um laço de repetição (*loop*) com 1000 iterações, para que houvesse a chance de todas as amostras aparecerem em ao menos uma das reduções amostrais feitas. Para obter o resultado dessa abordagem, foram escolhidos os genes que mais apareceram nos resultados de cada iteração, portanto, os genes que estiveram presentes nos resultados de mais de 60% das iterações foram considerados como resultado dessa segunda análise.

Por fim, foi feita uma última abordagem de análise de expressão gênica diferencial, utilizando apenas amostras da Classe 2a do tumor. Essa classe é a que mais possui amostras que progrediram e por isso foi escolhida dentre as 4 classes citadas anteriormente. O número total de amostras nessa análise foi de 64, sendo 32 não-progressivas e 32 progressivas.

Para obter os resultados de todas as abordagens, os genes diferencialmente expressos foram selecionados a partir de seus valores de log2 *fold change*, p-*value* ajustado e *baseMean* (média). Os valores de *cut-off* estão descritos no Quadro 3.

Quadro 3 - Valores de *cut-off* para seleção de genes diferencialmente expressos

	p-*value* ajustado	*Log2foldchange*	*baseMean* (média)
Abordagem 1	0,05	1,5	1000
Abordagem 2	0,05	2	1000
Abordagem 3	0,05	2	1000

Fonte: Autoria própria

No fim dessas etapas, o resultado esperado é uma lista de genes diferencialmente expressos, para que em seguida, seja possível convergir com os resultados das outras análises que também foram realizadas neste trabalho.

3.4 Análise de Conjunto de Genes Enriquecidos (GSEA)

O *software* GSEA, sigla para *Gene Set Enrichment Analysis*, é um tipo de método computacional de bioinformática que determina se alguns conjuntos de genes, já existentes e organizados em coleções no banco de dados MSigDB (*Molecular Signature Database*) (LIBERZON et al., 2011), têm ou não relevância estatística em sua expressão entre dois grupos biológicos (SUBRAMANIAN et al., 2005), a partir de uma matriz de expressão gênica e um arquivo de *phenoData*. No caso do projeto, os grupos são as amostras de CBNMI com e sem progressão para o CBMI.

Esse *software* foi escolhido para este trabalho pois ele permite observar todas as mudanças no perfil de expressão do *dataset*, por menores que sejam. Isso é interessante, uma vez que, por mais que um gene esteja pouco diferencialmente expresso, ele pode ter grande influência dentro de um grupo gênico relevante para o grupo de estudo (YU, 2019).

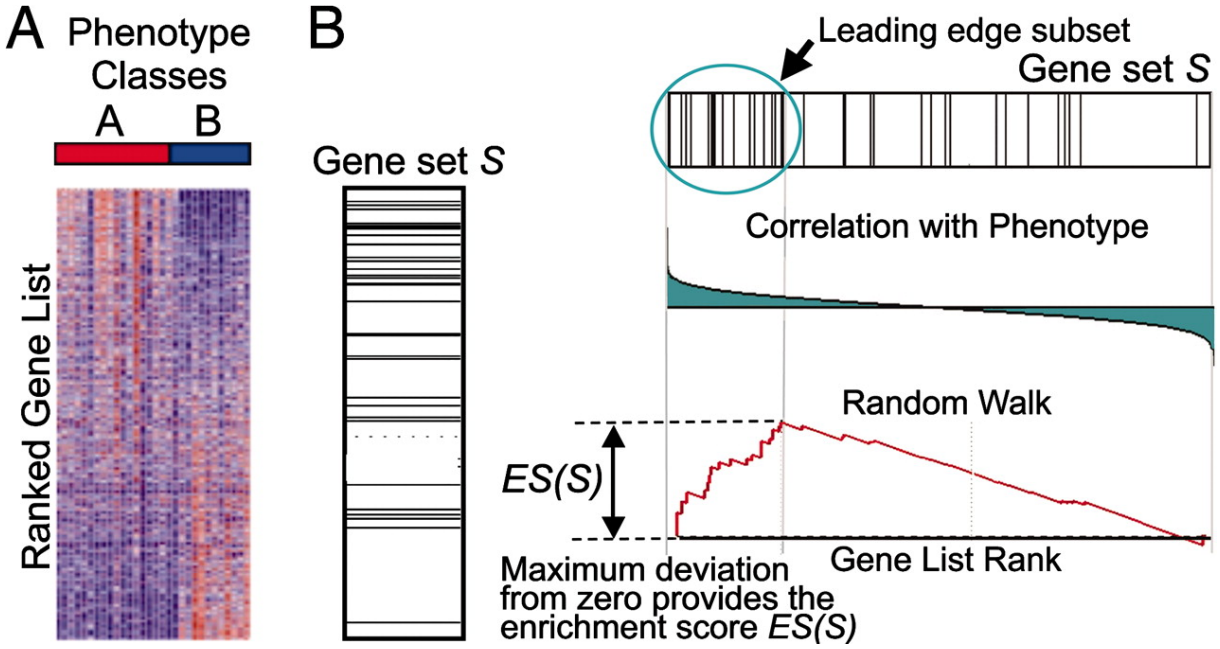
Tendo isso em vista, foi utilizada a matriz de expressão gênica original, com 57165 transcritos, com os valores de expressão normalizados através do pacote DESEq2 e as

mesmas 530 amostras da análise anterior. Além disso, a coleção de grupo de genes escolhida para realizar essa análise foi a *Hallmarks Gene Sets*, disponível no banco de dados MSigDB, a qual possui 50 grupos que representam estados biológicos bem definidos e que mostram ter perfis de expressão coerentes entre si.

Os parâmetros utilizados para performar o GSEA foram os padrões do próprio *software*, e o resultado principal são os grupos de genes com maior valor de *enrichment score* (ES) em módulo.

Durante a análise no GSEA, os genes da matriz de expressão são ranqueados de acordo com sua diferença de expressão média entre o grupo não-progressivo e progressivo. Portanto, os genes que ficam no topo do ranque estão superexpressos em um dos grupos, enquanto os genes que ficam por último estão hipoexpressos. Isso é representado na Figura 3.A, pelos grupos A e B respectivamente. Em seguida, essa lista ranqueada de genes é comparada com cada um dos 50 grupos de genes da coleção *Hallmarks Gene Sets*, e assim é calculado o valor do *enrichment score* (ES). O ES é calculado para cada grupo de genes do *Hallmarks Gene Sets* de acordo com a quantidade de genes que são encontrados em comum entre a lista ranqueada e o grupo. O gráfico de barras da Figura 3.B representa quais genes de um dado grupo de genes, representado por "*Gene set S*", foram identificados na lista ranqueada. A cada gene identificado, é colocada uma barra, e a densidade de barras no gráfico em alguma das extremidades representa o subgrupo de genes que levaram ao crescimento do valor de ES, citado como "*Leading edge subset*" na Figura 3. Para entender o ES de forma visual, o *software* também cria o gráfico "*Random Walk"*, em que é mostrado o crescimento do valor conforme eram encontrados os genes em comum entre as listas de genes.

Figura 3 - Representação geral do método do GSEA. (A) Uma matriz de expressão gênica ranqueada de acordo com a diferença de expressão média dos genes entre os grupos A e B, representada pelo *heatmap*. (B) Gráfico de barras correspondente aos genes que o grupo de gene "*Gene set S*" tem em comum com a matriz de expressão ranqueada, o subgrupo de genes "*Leading edge Subset*" responsável pelo crescimento do gráfico *Random Walk*, que representa visualmente o crescimento do valor do *enrichment score*



Fonte: (SUBRAMANIAN et al., 2005)

A *Leading Edge Analysis* (LEA) é oferecida também pelo *software* e permite identificar quais foram os genes que mais contribuíram para o enriquecimento de cada um dos grupos de genes. Essa etapa foi feita com os grupos de genes que obtiveram um valor de ES maior ou igual a 0.4, em módulo.

A partir dessas análises, o objetivo foi obter os conjuntos de genes que mais estão sendo relatados como significantes no grupo de amostras progressivas e, consequentemente, os genes que mais influenciam o enriquecimento em cada um dos conjuntos.

4 Resultados e Discussão

4.1 Análise de Expressão Gênica Diferencial (DESeq2)

Primeiramente, foi feita uma análise de PCA (*Principal Component Analysis*), para observar o comportamento das 530 amostras, baseada em todos os genes da matriz de expressão (*genome-wide*). Como é possível observar na Figura 4, não há nenhum agrupamento das amostras de acordo com as duas classes, progressivas e não-progressivas, demonstrando então que há pouca variância entre a expressão gênica das classes a nível global.

Figura 4 - PCA da expressão *genome-wide* das 530 amostras de CBNMI divididas em progressivas e não-progressivas

Fonte: Autoria própria

Na matriz de expressão, os genes estavam nomeados de acordo com seu código de identificação do Ensembl, por exemplo, ENSG00000198695. Para melhor visualização dos resultados, esses genes foram convertidos para seus respectivos símbolos, por exemplo, *ND6*. No entanto, alguns dos genes resultantes não possuem um símbolo associado a seu código

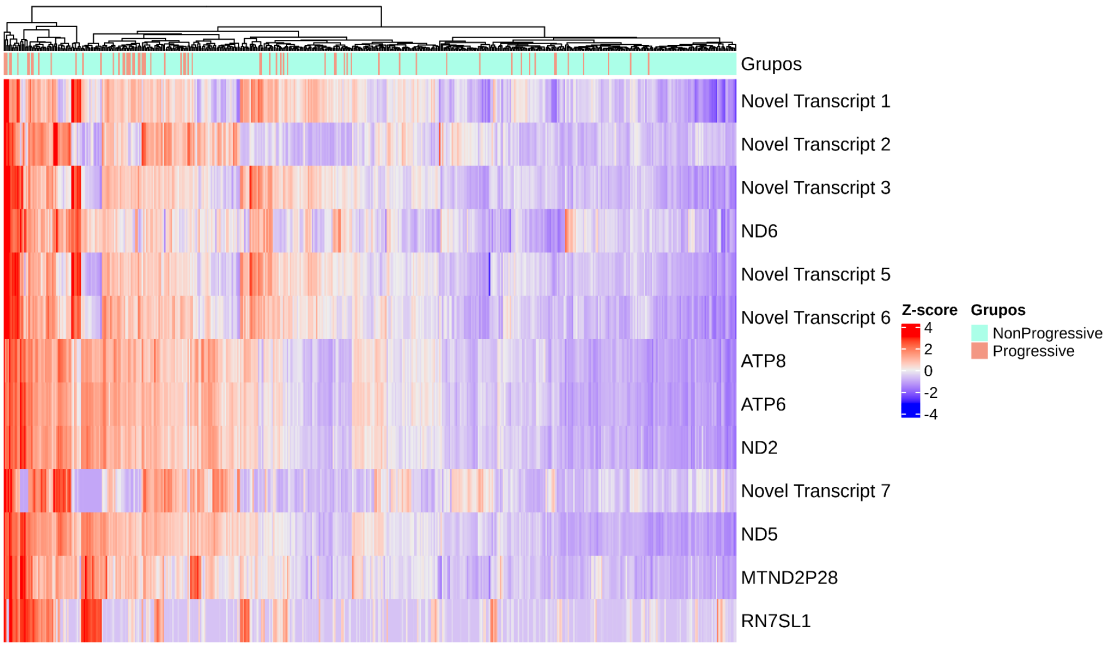
pois ainda não foram mapeados, portanto, foram nomeados de *novel transcripts* neste trabalho.

4.1.1 Análise de Expressão Gênica Diferencial com o total de amostras (N=530)

Nesta primeira abordagem com o pacote DESeq2, em que foi utilizado o número total de amostras, sendo 465 não-progressivas e 65 progressivas, foram identificados 13 genes diferencialmente expressos.

Como é possível observar na Figura 5, por mais que esses genes foram identificados como superexpressos no grupo progressivo, a visualização do resultado através do *heatmap* e do dendrograma mostrou que as amostras progressivas não se agrupam na anotação no topo do gráfico e estão distribuídas sem padrão ao longo da figura.

Figura 5 - *Heatmap* da expressão gênica dos 13 genes diferencialmente expressos encontrados na primeira abordagem de análise do trabalho

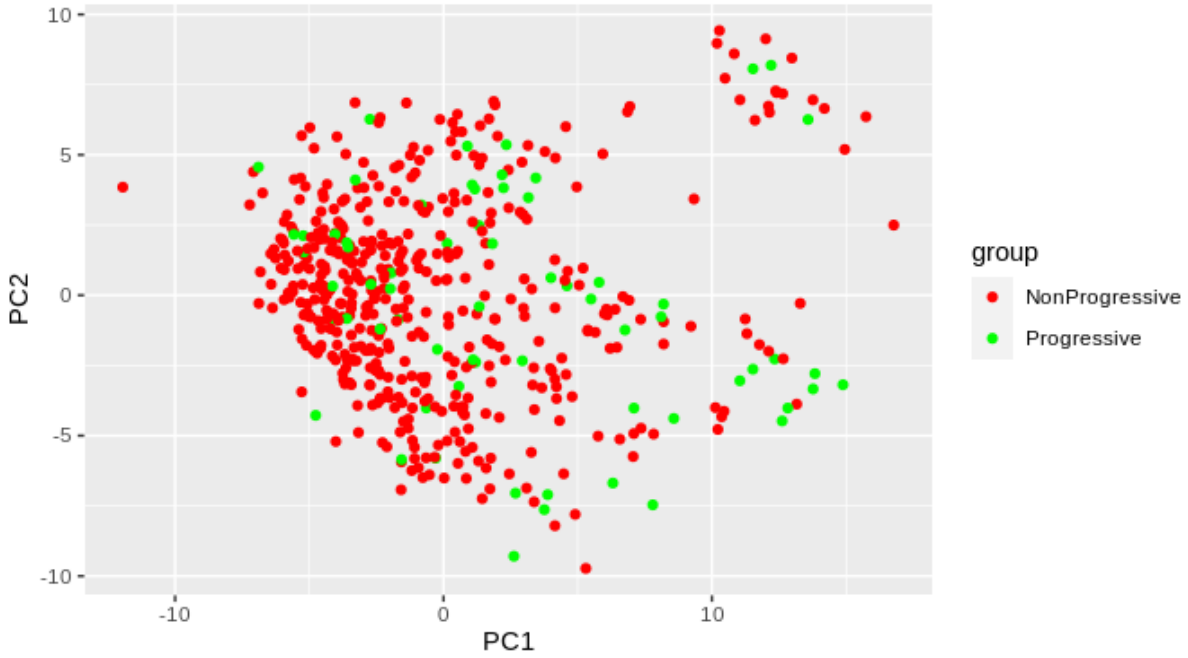


Fonte: Autoria própria

Isso pode ter acontecido pelo fato de que há pouca variância na expressão dos dois grupos, como foi demonstrado na Figura 4. Para confirmar essa hipótese, foi feita uma análise de PCA das amostras tendo como base apenas a expressão dos 13 genes encontrados nesta etapa. Como demonstrado na Figura 6, ainda assim as duas classes não se agrupam e se

mostram realmente semelhantes entre si. Em outras palavras, os genes diferencialmente selecionados não foram capazes de separar as duas classes.

Figura 6 - PCA da expressão dos 13 genes nas 530 amostras de CBNMI divididas em progressivas e não-progressivas



Fonte: Autoria própria

4.1.2 Análise de Expressão Gênica Diferencial com redução de amostras (N=130) (*Loop*)

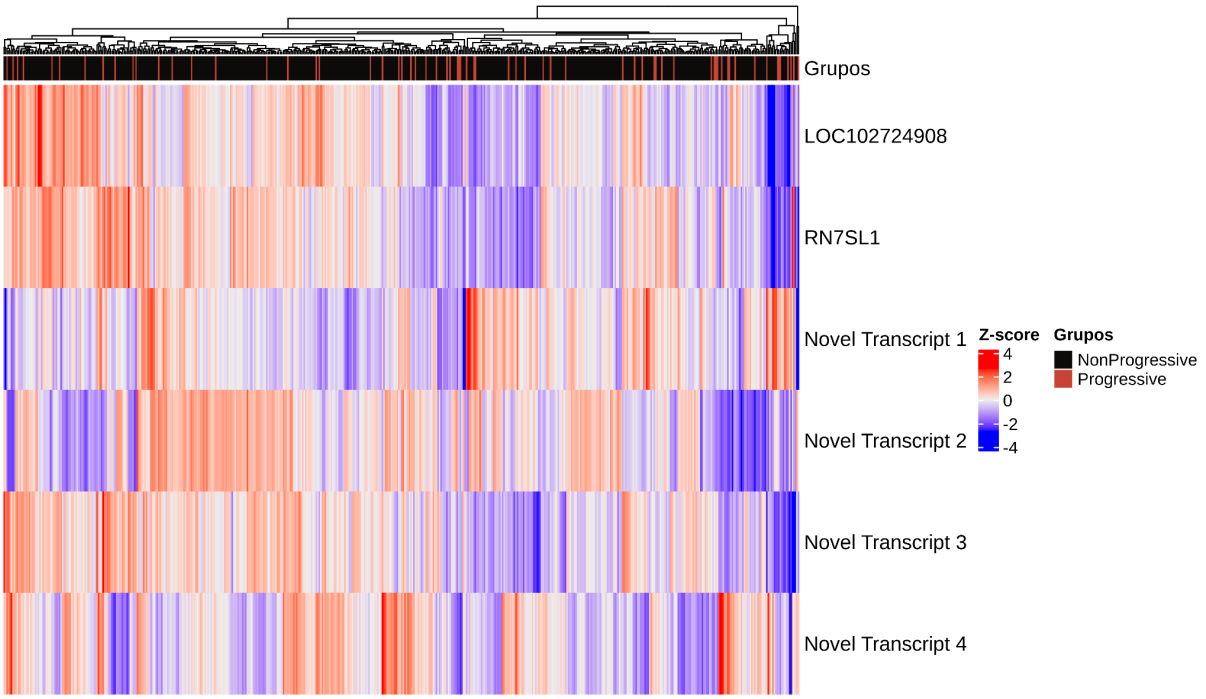
Nesta segunda abordagem, como foi descrito no tópico 3.3, foi realizada uma redução de amostras não-progressivas para resolver o problema de desbalanceamento no número de amostras na abordagem anterior, com a pretensão de obter melhores resultados. Para isso, foi construído um laço de repetição (*loop*) de 1000 iterações, para que todas as amostras estejam presentes em alguma redução amostral para análise. Em cada iteração foi realizada uma análise pelo DESeq2 com as 130 amostras, sendo 65 não-progressivas escolhidas aleatoriamente pela função *stratified()*.

Cada uma das 1000 análises realizadas retornou como resultado uma lista de genes diferencialmente expressos. Ao todo, foram 40 genes únicos presentes ao longo de todos esses

resultados. Para convergir e identificar os genes mais significativos para todas as amostras, foram selecionados os genes com frequência em 600 ou mais resultados (ou seja, genes identificados em pelo menos 60% das iterações). Dessa forma, foram 6 genes resultantes a partir dessa abordagem.

As amostras progressivas ainda não se agrupam no dendrograma do topo do gráfico, como demonstrado no *heatmap* da Figura 7, e o comportamento geral da imagem é parecido com o anterior, em que não fica claro os grupos com genes diferencialmente expressos.

Figura 7 - *Heatmap* da expressão gênica dos 6 genes diferencialmente expressos encontrados na segunda abordagem de análise do trabalho

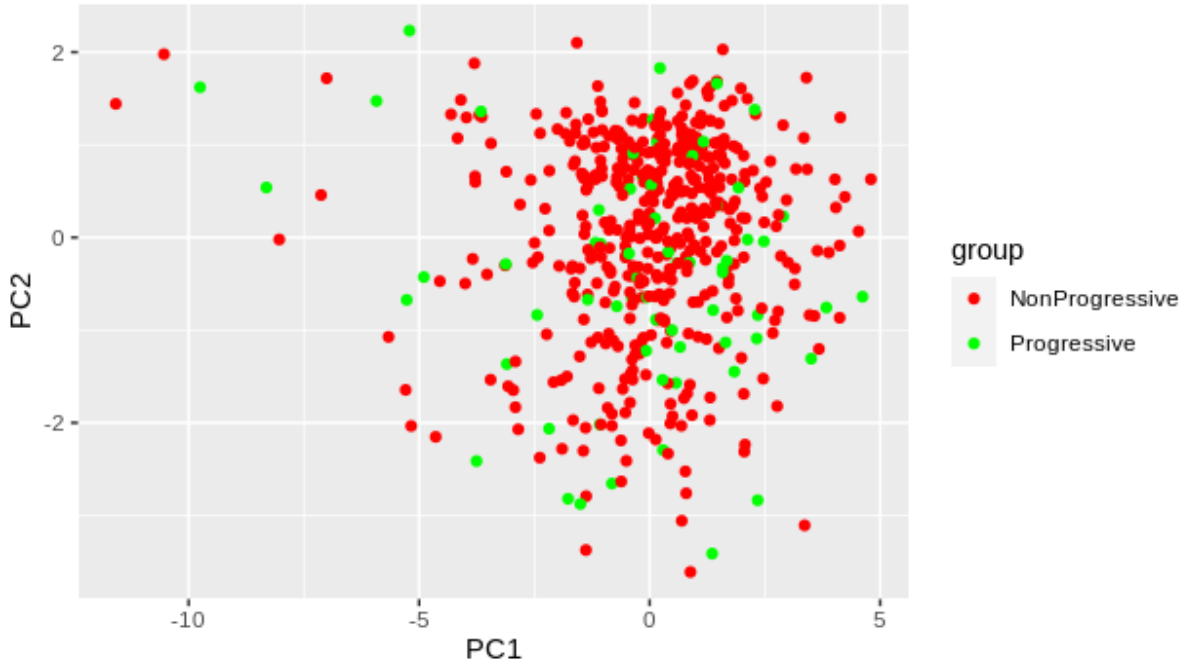


Fonte: Autoria própria

Além disso, também foi feita uma análise de PCA para essa abordagem, para que seja possível observar como as amostras se comportam baseadas na expressão dos 6 genes resultantes desta etapa (Figura 8). E, da mesma forma que nos outros casos, não houve agrupamento das amostras progressivas que possa indicar alguma diferença relevante entre os grupos.

Por mais que a redução amostral tenha sido feita para resolver o problema de desbalanceamento de amostras, isso não foi o suficiente para selecionarmos genes capazes de separar os grupos de amostras que progrediram do grupo que não progrediu.

Figura 8 - PCA da expressão dos 6 genes nas 530 amostras de CBNMI divididas em progressivas e não-progressivas



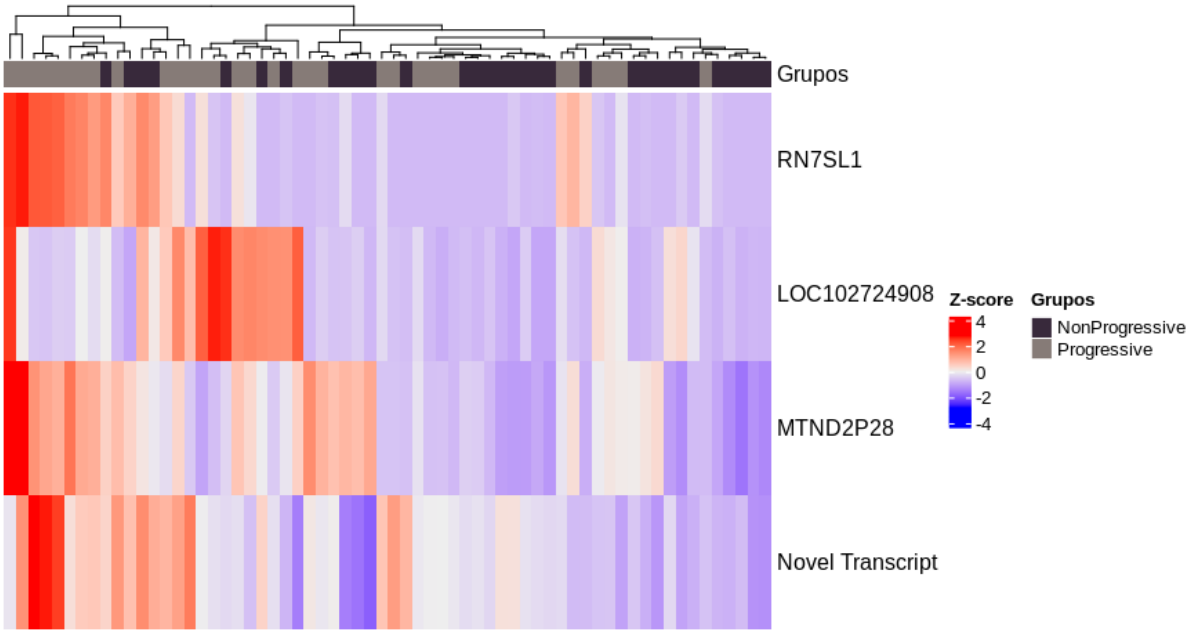
Fonte: Autoria própria

4.1.3 Análise de Expressão Gênica Diferencial com amostras da Classe 2a (N=64)

Por fim, uma última tentativa de análise de expressão gênica diferencial com o DESeq2 foi realizada utilizando apenas amostras do subtipo tumoral 2a. O número amostral nesta etapa foi de 64, sendo 32 amostras progressivas e 32 não-progressivas.

O resultado foi 4 genes superexpressos nas amostras progressivas, e como é possível visualizar na Figura 9, também não houve um agrupamento claro entre os grupos de interesse do trabalho.

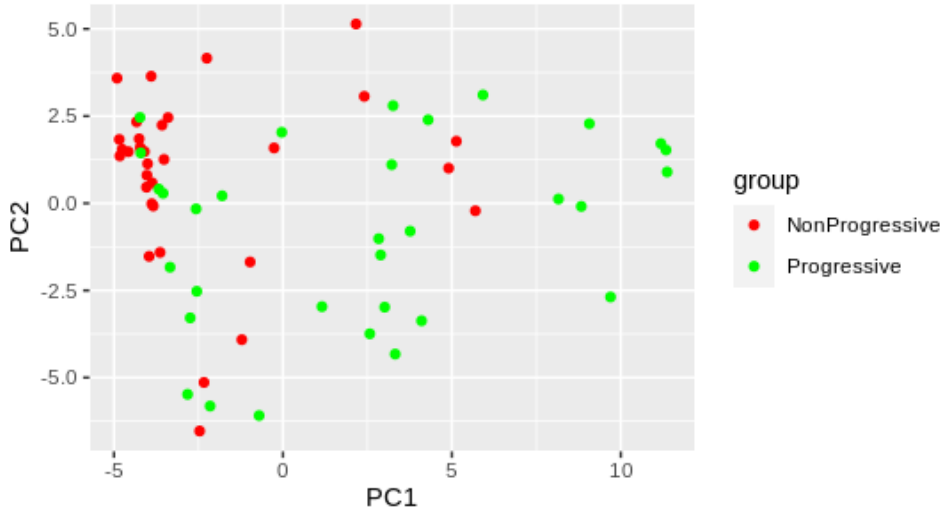
Figura 9 - *Heatmap* da expressão gênica dos 4 genes diferencialmente expressos encontrados na terceira abordagem de análise do trabalho, utilizando apenas amostras do subtipo tumoral 2a



Fonte: Autoria própria

No entanto, ao fazer a análise de PCA com as amostras dessa abordagem baseada na expressão dos 4 genes resultantes, o grupo progressivo ficou mais disperso ao longo do gráfico, diferente do grupo não-progressivo, que ficou majoritariamente concentrado em uma região específica (Figura 10). Estes genes podem ser investigados no futuro como candidatos a marcadores de progressão para amostras deste subtipo.

Figura 10 - PCA da expressão dos 4 genes nas 64 amostras de CBNMI do subtipo tumoral 2a divididas em progressivas e não-progressivas



Fonte: Autoria própria

4.2 Análise de Conjunto de Genes Enriquecidos (GSEA)

Como foi previamente dito, a análise com o *software* GSEA consistiu em identificar os grupos de genes mais enriquecidos nas 530 amostras de CBNMI, entre as classes progressiva e não-progressiva, de acordo com seus valores de *enrichment score* (ES). Após rodar o software, três conjuntos de genes foram identificados como super-representados no grupo de amostras progressivo. Esse resultado está descrito na Tabela 2, junto com as estatísticas para cada grupo de genes.

Tabela 2 - Conjuntos de genes resultantes da análise com o *software* GSEA e suas respectivas estatísticas acerca do nível de enriquecimento (ES) e p-*value*

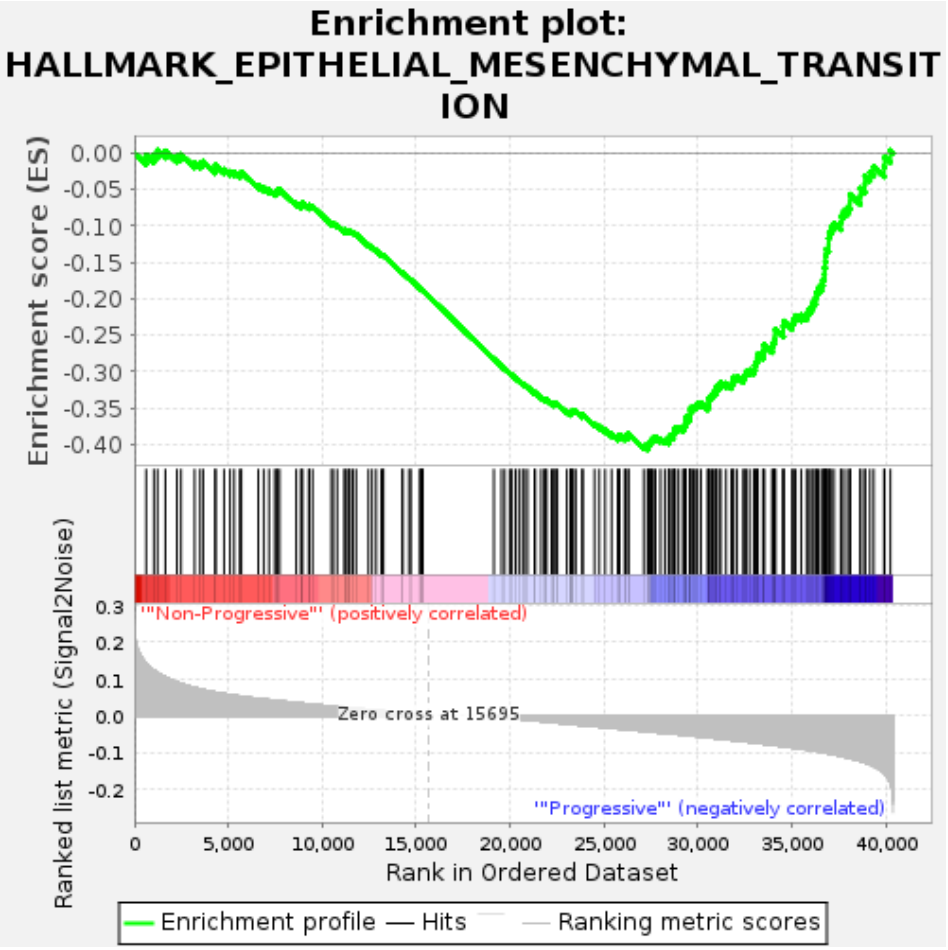
Conjunto de genes	ES	p-*value*
Fuso mitótico	-0,47	0,000
Resposta inflamatória	-0,46	0,000
Transição epitélio-mesenquimal	-0,41	0,000

Fonte: Resultado adaptado do *software* GSEA

Para cada conjunto de gene resultante, foi gerada uma imagem para a visualização do gráfico de barras e do gráfico do "*random walk*" do valor de ES, citados anteriormente. Na Figura 11, observa-se o crescimento modular do ES do conjunto de genes relacionado à transição epitélio-mesenquimal (EMT). Houve uma densidade maior de genes em comum entre a lista ranqueada de genes do *dataset* do trabalho e esse conjunto de genes, como é possível observar ainda na Figura 11, do lado direito do gráfico de barras.

Como a concentração de genes aconteceu na extremidade direita da lista, isso significa que os genes responsáveis por esse enriquecimento estão superexpressos no último grupo dentre os dois analisados, ou seja, no grupo de amostras progressivas. Por esse motivo, o crescimento do ES se dá de forma modular e neste caso, retorna um valor negativo de 0,47.

Figura 11 - Representação visual do valor de ES para o conjunto de genes relacionados à EMT



Fonte: Produzido pelo software GSEA

O processo de EMT já foi relacionado na literatura com a progressão, recorrência e metástase do câncer de bexiga (YUN; KIM, 2013), corroborando com o resultado encontrado neste trabalho.

4.2.1 *Leading Edge Analysis* (LEA)

Como anteriormente citado, a LEA tem como objetivo identificar os genes responsáveis pelo enriquecimento de um determinado conjunto de genes em um dos grupos de estudo do trabalho. Os conjuntos selecionados nesta etapa foram os três citados no tópico anterior, relacionados ao fuso mitótico, resposta inflamatória e EMT. Ao todo, foram identificados 11 genes, listados no Quadro 4.

Quadro 4 - Genes resultantes da análise LEA e seus respectivos conjuntos de genes

	Fuso mitótico	Resposta inflamatória	Transição epitélio-mesenquimal
SERPINE1			
TIMP1			
IL15			
INHBA			
CXCL6			
ITGB3			
PLAUR			
ITGA5			
KIF1B			
FLNA			
NOTCH2			

Fonte: Adaptado do resultado do *software* GSEA

Três desses genes se destacam por suas descrições mais recentes na literatura. O gene *INHBA* já foi descrito como um biomarcador de progressão tumoral e metástase em pacientes com câncer de bexiga, em um estudo que buscou relacionar a hipometilação em células tumorais da bexiga com a superexpressão desse gene e seu impacto na progressão (KAO et al., 2022). Um estudo sobre um subtipo de câncer renal já citou o gene *PLAUR* como envolvido no processo de progressão desse tumor, em que ele também foi identificado como parte de uma via biológica relacionada à resposta inflamatória (WANG et al., 2022b), assim como no resultado deste trabalho. Ademais, já foi citado que o gene *FLNA* promove a transição epitélio-mesenquimal ao ser regulado pelo composto Cromo Hexavalente (Cr VI), um fator carcinogênico que promove a proliferação celular no câncer de bexiga, quando presente na urina (SHENG et al., 2021).

Além disso, os genes *TIMP1*, *ITGA5*, *NOTCH2* e *SERPINE1* já foram relacionados com proliferação celular e progressão tumoral em câncer de ovário epitelial (SONEGO et al., 2019), câncer gástrico (WANG et al., 2022a), carcinoma cístico adenóide salivar (QU et al., 2016) e câncer de células escamosas do esôfago (WANG et al., 2020), respectivamente.

5 Conclusão

Os resultados obtidos na primeira etapa do trabalho, que consistiu em análises de expressão gênica diferencial utilizando o pacote DESeq2, mostraram que não há uma evidência forte de diferença de expressão gênica entre os grupos não-progressivo e progressivo do *dataset* deste trabalho. Dessa forma, por mais que alguns genes foram considerados superexpressos no grupo progressivo de acordo com seus valores de *cut-off*, ao fazer uma representação visual desse resultado, não é mostrado nenhum agrupamento coerente entre as amostras dos diferentes grupos. Por esta abordagem, não foi possível identificar genes que possam predizer a progressão tumoral.

Por outro lado, na segunda etapa do trabalho, o método realizado pelo *software* GSEA em cima do mesmo *dataset* foi capaz de retornar resultados congruentes com a hipótese do trabalho e que estão de acordo com o que vêm sido citado na literatura ultimamente acerca da progressão tumoral, seja no próprio câncer de bexiga ou em tipos de câncer diversos.

Os genes *INHBA*, *PLAUR* e *FLNA* podem ser melhor explorados no contexto da progressão do CBNMI para o músculo da bexiga, dadas as evidências encontradas ao utilizar o GSEA, e tendo em vista os grupos de genes evidenciados também pelo *software*, relacionados ao fuso mitótico, à resposta inflamatória e à transição epitélio-mesenquimal. É interessante que esse resultado seja validado com amostras de CBNMI de *datasets* diferentes, a fim de confirmar que o que foi encontrado faz parte do perfil desse tipo tumoral.

Por fim, pode-se concluir que existem diferentes metodologias para realizar análise de expressão gênica, que se distinguem por fatores como tipo de processamento do dado para *input*, testes estatísticos, entre outros parâmetros. No caso deste trabalho, a metodologia oferecida pelo pacote DESeq2 não foi capaz de retornar bons resultados com o *dataset* utilizado. Por outro lado, o método do GSEA realizado no mesmo *dataset* mostrou ser eficaz e retornou resultados coerentes com a literatura e com a hipótese do trabalho.

6 Referências Bibliográficas

FARLING, K. B. Bladder cancer: Risk factors, diagnosis, and management. **The Nurse Practitioner**, v. 42, n. 3, p. 26–33, mar. 2017.

GENTLEMAN, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. **Genome Biology**, v. 5, n. 10, p. R80, 2004.

HANAHAN, D. Hallmarks of Cancer: New Dimensions. **Cancer Discovery**, v. 12, n. 1, p. 31–46, jan. 2022.

INSTITUTO NACIONAL DE CÂNCER. **Câncer de bexiga**. Disponível em: <https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/bexiga>. Acesso em: 21 nov. 2022.

KAMOUN, A. et al. A Consensus Molecular Classification of Muscle-invasive Bladder Cancer. **European Urology**, v. 77, n. 4, p. 420–433, 1 abr. 2020.

KAO, C.-C. et al. DNA Hypomethylation Is Associated with the Overexpression of INHBA in Upper Tract Urothelial Carcinoma. **International Journal of Molecular Sciences**, v. 23, n. 4, p. 2072, 13 fev. 2022.

LIBERZON, A. et al. Molecular signatures database (MSigDB) 3.0. **Bioinformatics (Oxford, England)**, v. 27, n. 12, p. 1739–40, 2011.

LINDSKROG, S. V. et al. An integrated multi-omics analysis identifies prognostic molecular subtypes of non-muscle-invasive bladder cancer. **Nature Communications**, v. 12, n. 1, p. 2301, 16 abr. 2021.

LOVE, M. I.; HUBER, W.; ANDERS, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. **Genome Biology**, v. 15, n. 12, dez. 2014.

QU, J. et al. Notch2 signaling contributes to cell growth, invasion, and migration in salivary adenoid cystic carcinoma. **Molecular and Cellular Biochemistry**, v. 411, n. 1-2, p. 135–141, 1 jan. 2016.

RITCHIE, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Research**, v. 43, n. 7, p. e47–e47, 20 jan. 2015.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, n. 1, p. 139–140, 11 nov. 2009.

SHENG, F. et al. Chromium (VI) promotes EMT by regulating FLNA in BLCA. **Environmental Toxicology**, v. 36, n. 8, p. 1694–1701, 12 maio 2021.

SONEGO, M. et al. TIMP-1 is Overexpressed and Secreted by Platinum Resistant Epithelial Ovarian Cancer Cells. **Cells**, v. 9, n. 1, p. 6, 18 dez. 2019.

STEFANO, G. B. Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. **Medical Science Monitor Basic Research**, v. 20, p. 138–142, 2014.

SHI, J.; WALKER, M. Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles. **Current Bioinformatics**, v. 2, n. 2, p. 133–137, 1 maio 2007.

SUBRAMANIAN, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences**, v. 102, n. 43, p. 15545–15550, 30 set. 2005.

YU, G. **Chapter 5 Overview of enrichment analysis | Biomedical Knowledge Mining using GOSemSim and clusterProfiler**. Disponível em: <https://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html>. Acesso em: 12 jul. 2022.

WANG, D. et al. PAI-1 overexpression promotes invasion and migration of esophageal squamous carcinoma cells. **Yi Chuan = Hereditas**, v. 42, n. 3, p. 287–295, 20 mar. 2020.

WANG, J.-F. et al. ITGA5 Promotes Tumor Progression through the Activation of the FAK/AKT Signaling Pathway in Human Gastric Cancer. **Oxidative Medicine and Cellular Longevity**, v. 2022, p. 8611306, 2022a.

WANG, Z. et al. Comprehensive analysis of the importance of PLAUR in the progression and immune microenvironment of renal clear cell carcinoma. **PloS One**, v. 17, n. 6, p. e0269595, 2022b.

WANG, Z.; GERSTEIN, M.; SNYDER, M. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, v. 10, n. 1, p. 57–63, jan. 2009.

YUN, S. J.; KIM, W.-J. Role of the Epithelial-Mesenchymal Transition in Bladder Cancer: From Prognosis to Therapeutic Target. **Korean Journal of Urology**, v. 54, n. 10, p. 645, 2013.