



**PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PÓS-GRADUAÇÃO EM MEIO AMBIENTE E
DESENVOLVIMENTO REGIONAL**

VICTOR HUGO GARCIA DE OLIVEIRA

**AVALIAÇÃO *IN SÍLICO* DO PERFIL DE EXPRESSÃO GENICA DOS TIPOS E
SUBTIPOS DE CÂNCER DE MAMA**

Presidente Prudente - SP
2025



**PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PÓS-GRADUAÇÃO EM MEIO AMBIENTE E
DESENVOLVIMENTO REGIONAL**

VICTOR HUGO GARCIA DE OLIVEIRA

**AVALIAÇÃO *IN SÍLICO* DO PERFIL DE EXPRESSÃO GENICA DOS TIPOS E
SUBTIPOS DE CÂNCER DE MAMA**

Dissertação apresentada à Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade do Oeste Paulista, como parte dos requisitos para obtenção do título de mestre em Meio Ambiente e Desenvolvimento Regional.

Área de concentração: Meio Ambiente e Desenvolvimento

Linha de pesquisa: Avaliação e Análise de Impacto Ambiental

Orientador: Prof. Dr. Edson Assunção Mareco

616.994 O48a	Oliveira, Victor Hugo Garcia de. Avaliação <i>in silico</i> do perfil de expressão genica dos tipos e subtipos de câncer de mama / Victor Hugo Garcia de Oliveira. -Presidente Prudente, 2025. 47 f.: il. Dissertação (Mestrado em Meio Ambiente e Desenvolvimento Regional) - Universidade do Oeste Paulista - Unoeste, Presidente Prudente, SP, 2025. Bibliografia. Orientador: Dr. Edson Assunção Mareco 1.Co-expressão gênica. 2.TNBC. 3.WGCNA. 4.Biomarcadores. 5.Transcriptoma. I. Título.
-----------------	--

Catlogação na Fonte: Maria Letícia Silva Vila Real - CRB 8/10699

VICTOR HUGO GARCIA DE OLIVEIRA

**AVALIAÇÃO *IN SÍLICO* DO PERFIL DE EXPRESSÃO GENICA DOS TIPOS E
SUBTIPOS DE CÂNCER DE MAMA**

Dissertação apresentada à Pró-Reitoria de Pesquisa e Pós-Graduação, Universidade do Oeste Paulista, como parte dos requisitos para obtenção do título de mestre em Meio Ambiente e Desenvolvimento Regional.

Presidente Prudente, 31 de janeiro de 2025.

BANCA EXAMINADORA

Prof. Dr. Edson Assunção Mareco (Orientador)
Universidade do Oeste Paulista – Unoeste
Presidente Prudente-SP

Prof. Dr. Rondinelle Arthur Simões Salomão
Faculdade de Tecnologia de Presidente Prudente - Fatec
Presidente Prudente-SP

Profa. Dra. Renata Calciolari Rossi
Universidade do Oeste Paulista – Unoeste
Presidente Prudente-SP

DEDICATÓRIA

Ao meus pais, que desde o início investiram seus recursos mais valiosos em mim, o amor, a atenção e o conhecimento.

A minha companheira de aventuras, assim como ela gosta de falar, Jhulya Zanfolin Andrade, por estar comigo em todos os momentos, desde os mais fáceis até os mais difíceis, tornando esses últimos mais leves.

Ao meu “patrão” durante quase todo o mestrado, Fabricio Fernando da Silva Rosário, que sem sua compreensão e incentivo este trabalho não teria sido realizado.

Ao meu tio, Fernando “sapatinho”, com suas caronas para faculdade encurtou as distancias para realização deste projeto.

A todos os outros que me apoiaram, verbalmente ou em pensamento, muito obrigado!

AGRADECIMENTOS

Ao meu orientador, Dr. Edson Assunção Mareco, por despertar em mim o anseio pela pesquisa e pelo conhecimento.

*"Se eu vi mais longe,
foi por estar sobre ombros de gigantes." (Isaac Newton)*

RESUMO

Avaliação *in silico* do perfil de expressão genica dos tipos e subtipos de câncer de mama

O câncer de mama é a neoplasia mais comum entre as mulheres no Brasil, com destaque para as taxas de incidência e mortalidade para as regiões Sul e Sudeste, sendo o subtipo triplo negativo (TNBC) um dos mais agressivos e de difícil tratamento devido à ausência de receptores hormonais. Este estudo tem como objetivo identificar módulos de co-expressão gênica associados ao TNBC, utilizando a metodologia Weighted Gene Co-Expression Network Analysis (WGCNA), a partir de dados públicos do The Cancer Genome Atlas (TCGA-BRCA). A pesquisa foi desenvolvida com abordagem quantitativa e exploratória, utilizando análise bioinformática de transcriptomas de 1.032 amostras tumorais, das quais 115 pertencem ao TNBC e 917 aos subtipos hormonais. Os métodos aplicados incluíram a filtragem e normalização dos dados por meio do pacote DESeq2, análise diferencial de expressão e identificação de redes de co-expressão gênica. Foram identificados 21 módulos de co-expressão, destacando-se os módulos "MEblack", associado aos subtipos hormonais, e "MEcyan", correlacionado ao TNBC. A análise de enriquecimento funcional demonstrou que o módulo MEblack está relacionado a processos de sinalização hormonal e desenvolvimento glandular, enquanto o módulo MEcyan está associado à organização celular e proliferação. Redes de interação proteína-proteína (PPI) revelaram genes-chave envolvidos na tumorigênese, como ESR1, ERBB4, PGR e GATA3, que desempenham papel crítico na regulação do crescimento celular e na diferenciação tecidual. O estudo reforça a importância da análise de redes gênicas para a compreensão das diferenças moleculares entre os subtipos de câncer de mama, possibilitando a identificação de novos biomarcadores e alvos terapêuticos. A aplicação de abordagens de bioinformática permite avançar no desenvolvimento de estratégias personalizadas para o diagnóstico e tratamento do TNBC, além de contribuir para a formulação de políticas públicas voltadas à saúde oncológica.

Palavras-chave: Co-expressão gênica; TNBC; WGCNA; Biomarcadores; Transcriptoma.

ABSTRACT

***In Silico* Evaluation of Gene Expression Profiles of Breast Cancer Types and Subtypes**

Breast cancer is the most common neoplasm among women in Brazil, with notable incidence and mortality rates in the South and Southeast regions. The triple-negative subtype (TNBC) is one of the most aggressive and difficult to treat due to the absence of hormone receptors. This study aims to identify gene co-expression modules associated with TNBC using the Weighted Gene Co-Expression Network Analysis (WGCNA) methodology, based on publicly available data from The Cancer Genome Atlas (TCGA-BRCA). The research was conducted with a quantitative and exploratory approach, utilizing bioinformatics analysis of transcriptomes from 1,032 tumor samples, including 115 TNBC cases and 917 hormonal subtypes. The applied methods involved data filtering and normalization using the DESeq2 package, differential gene expression analysis, and gene co-expression network identification. A total of 21 co-expression modules were identified, with two standing out: "MEblack", associated with hormonal subtypes, and "MEcyan", correlated with TNBC. Functional enrichment analysis revealed that the MEblack module is related to hormonal signaling and glandular development, whereas the MEcyan module is associated with cellular organization and proliferation. Protein-protein interaction (PPI) networks identified key genes involved in tumorigenesis, such as ESR1, ERBB4, PGR, and GATA3, which play a critical role in cell growth regulation and tissue differentiation. This study highlights the importance of gene network analysis in understanding the molecular differences between hormonal and non-hormonal breast cancer subtypes, providing insights into novel biomarkers and therapeutic targets. The application of bioinformatics approaches advances the development of personalized diagnostic and treatment strategies for TNBC, while also contributing to public health policies focused on oncology.

Keywords: Gene co-expression; TNBC; WGCNA; Biomarkers; Transcriptome.

LISTA DE SIGLAS

DEGs	—	<i>Differentially Expressed Genes</i> (Genes diferencialmente expressos)
PPI	TCGA	— <i>Protein-Protein Interactions</i> (Interações proteína-proteína)
TNBC		— <i>The Cancer Genome Atlas Program</i> (Atlas genômico do câncer)
WGCNA		— <i>Triple Negative Breast Cancer</i> (Câncer de mama triplo negativo)— <i>Weighted Correlation Network Analysis</i> (Análise de rede de correlação ponderada)

LISTA DE FIGURAS

Figura 1—	Volcano plot representando genes diferencialmente expressos entre os grupos triplo negativo e não triplo negativo.....	32
Figura 2—	Teste de conectividade segundo o modelo de rede de livre escala	33
Figura 3—	Matriz de correlações entre os módulos e as variáveis clínicas.....	35
Figura 4—	Processos Biológicos identificados para o modulo “MEblack”	36
Figura 5—	Processos Biológicos identificados para o modulo “MEcyan”	37
Figura 6—	Rede de interações proteína-proteína (PPI) do “MEblack”	38
Figura 7—	Rede de interações proteína-proteína (PPI) do “MEcyan”	39

SUMÁRIO

PRÓLOGO	11
1. INTRODUÇÃO	13
1.1. Problema de pesquisa e Hipótese.....	14
1.2. Objetivos	15
1.2.1. Objetivo geral	15
1.2.2. Objetivos específicos.....	15
1.3. Metodologia da Pesquisa.....	15
1.3.1. Classificação da Pesquisa.....	16
1.3.2. Procedimentos Metodológicos	16
1.3.3. Procedimentos técnicos utilizados	16
1.4. Considerações Éticas e Retorno à Sociedade	17
2. REFERENCIAL TEÓRICO.....	19
2.1. Epidemiologia e Ciências Ômicas no Estudo do Câncer.....	19
2.2. Tecido Mamário e Neoplasias.....	20
2.2.1. Neoplasias Mamárias	21
2.3. Meio ambiente e as doenças neoplásicas.....	22
3. ARTIGO - ANÁLISE IN SILICO DO PERFIL TRANSCRIPTÔMICO DO CÂNCER DE MAMA: APLICAÇÃO DA WGCNA NA IDENTIFICAÇÃO DE GENES-CHAVE	25
4. CONCLUSÃO.....	45
REFÊRENCIAS	46

PRÓLOGO

Acredito que a biologia molecular tem avançado de forma exponencial, revelando a complexidade dos processos biológicos e a interconectividade dos sistemas vivos. No entanto, compreendo que a busca por desvendar o câncer, especialmente em subtipos agressivos como o câncer de mama triplo negativo, exige mais do que uma visão fragmentada ou mecanicista. Durante anos, o pensamento científico predominante tratou os sistemas biológicos como máquinas, em que cada componente poderia ser compreendido de forma isolada. Contudo, essa visão reducionista, embora poderosa para explicar mecanismos específicos, é insuficiente para abarcar a complexidade de sistemas vivos como redes gênicas ou o comportamento dinâmico dos tumores.

Acredito que a vida, em sua essência, é profundamente dinâmica, não apenas em sua dimensão biológica, mas também em seu caráter filosófico. Assim como somos continuamente moldados pelas experiências que vivemos e pelas interações que estabelecemos com o mundo ao nosso redor, os sistemas biológicos refletem esse mesmo dinamismo intrínseco. Redes gênicas, não são estruturas estáticas; elas reagem, se adaptam e se reorganizam em resposta a estímulos internos e externos, criando padrões de interação que determinam a funcionalidade ou a disfunção do organismo. Esse dinamismo, gera um estado constante de transformação, as vezes de difícil previsão, onde os desequilíbrios e as adaptações podem dar origem a fenômenos emergentes – desde a regeneração de tecidos até o desenvolvimento de tumores.

Adotar uma visão holística, para mim, significa reconhecer as limitações de análises exclusivamente mecanicistas e incluir os contextos ambientais e sistêmicos na investigação biológica. Fatores externos como disruptores endócrinos e poluentes não podem ser ignorados, pois adicionam camadas de complexidade que influenciam diretamente os desfechos biológicos e, em especial, as disparidades regionais na incidência e prognóstico do câncer. A biologia molecular, quando aliada a essa visão sistêmica, transcende o estudo dos genes e proteínas isolados para abraçar a compreensão do dinamismo dos sistemas vivos.

Estou convencido de que desafios tão amplos só podem ser enfrentados com abordagens igualmente abrangentes. Por isso, esse trabalho nasce da necessidade de compreender a biologia como um sistema integrado. Tenho como objetivo investigar como redes moleculares, que sustentam a vida em equilíbrio, podem, sob

a influência de múltiplos fatores, transformar-se no terreno fértil para o desenvolvimento de tumores. Acredito que somente ao olharmos nessa interseção entre a biologia molecular e a visão sistêmica, conseguimos de fato observar o potencial para identificar novos biomarcadores, desenvolver terapias mais eficazes e propor estratégias inovadoras para mitigar riscos. Afinal, como bem apontou Aristóteles em sua obra "Metafísica", "o todo é maior do que a soma das partes".

1. INTRODUÇÃO

O câncer de mama é o tipo de câncer que mais acomete as mulheres no Brasil, perdendo apenas para os tumores de pele não melanoma, e é também o que apresenta maior mortalidade. Foram estimados 73.610 novos casos para cada ano do triênio de 2023 a 2025, com maior concentração nas regiões Sul e Sudeste. As mesmas regiões também apresentam as maiores taxas de mortalidade por câncer em mulheres no país, com 13,60 e 13,16 óbitos por 100 mil mulheres, respectivamente, seguidas do Centro-Oeste (12,06 óbitos por 100 mil mulheres), Nordeste (10,78 óbitos por 100 mil mulheres) e Norte (9,73 óbitos por 100 mil mulheres) (Costa et al., 2024).

O câncer de mama é uma doença heterogênea, composta por múltiplos subtipos que se dividem em dois grandes grupos: aqueles nos quais as células tumorais expressam receptores hormonais, como receptores de estrogênio, progesterona e HER2 (fator de crescimento epidérmico tipo 2), e aqueles onde as células não expressam esses receptores, como no caso do subtipo triplo negativo (TNBC) (Russnes et al., 2017). O TNBC corresponde a aproximadamente 15-20% dos casos de câncer de mama e se caracteriza por uma alta taxa de recorrência e um prognóstico desfavorável (Dent et al., 2007). A ausência de receptores hormonais impede a utilização de terapias-alvo específicas, como as baseadas em tamoxifeno ou trastuzumabe, tornando seu manejo clínico um desafio. Além disso, o TNBC apresenta características agressivas com maior frequência, incluindo rápido crescimento tumoral, maior taxa de reincidência, maior probabilidade de metástases à distância e resistência a muitos regimes terapêuticos (Kim et al., 2012).

Sob o prisma dos *hallmarks of cancer* descritos por Hanahan e Weinberg, o TNBC exemplifica processos fundamentais do câncer, como a sustentação da proliferação celular descontrolada, evasão da apoptose, indução de angiogênese e capacidade de invadir tecidos e metastatizar. Esses processos são orquestrados por redes gênicas dinâmicas que interagem continuamente com o microambiente tumoral e são moduladas por fatores epigenéticos. Modificações como a metilação do DNA, alterações nas histonas e regulação por RNA não codificante (como miRNAs) contribuem para reprogramar a expressão gênica.

Essa soma de fatores é o grande responsável pelas dificuldades clínicas observadas no manejo do TNBC. Pesquisas atuais têm focado em desenvolver intervenções direcionadas para esse subtipo, com base na recente descoberta de genes relacionados a ele (Agostinetto et al., 2021). A análise de expressão diferencial

tem sido amplamente utilizada e se mostra uma ferramenta importante para identificar genes que influenciam o processo de desenvolvimento do TNBC (Saddiq et al., 2024). Abordagens que utilizam a co-expressão para identificar redes gênicas baseiam-se na premissa de que genes com perfis de expressão similares podem compartilhar funções biológicas (Eisen et al., 1998) e, portanto, participar de processos biológicos comuns. Uma dessas abordagens é a análise de redes de correlação ponderada para módulos de co-expressão gênica, conhecida como *Weighted Gene Co-Expression Network Analysis* (WGCNA) (Li et al., 2018), que permite identificar essas redes de interação gênica de maneira abrangente. Essa técnica possibilita a identificação de módulos de genes co-expressos, oferecendo uma visão integrada e aprofundada dos padrões moleculares (Qiu et al. 2019; Farhadian et al., 2021).

A complexidade fenotípica dos tumores pode ser mais bem elucidada com abordagens holísticas, que interpretam as alterações nos níveis de transcritos específicos como parte de uma grande rede de interações responsável pela manifestação de determinadas características (Bertrand et al., 2015). Desse modo, a WGCNA contribui para um entendimento mais completo da dinâmica molecular dos tumores, auxiliando diretamente nas decisões clínicas relacionadas ao tratamento e prognóstico da doença. A identificação de redes e módulos específicos no TNBC pode revelar novos alvos terapêuticos e biomarcadores prognósticos, melhorando potencialmente as abordagens de tratamento para pacientes com esse subtipo (Ahmadi et al., 2024). Além disso, os resultados obtidos com essa abordagem, podem vir a fomentar políticas públicas nacionais de ampliação do rastreamento precoce da doença, por meio da criação de bases de dados de sequenciamento de transcriptomas, tendo em vista que sua ausência, inviabiliza o emprego da análise.

1.1. Problema de pesquisa e Hipótese

A relação entre o câncer e o meio ambiente tem sido amplamente investigada, destacando-se o impacto de fatores ambientais na carcinogênese (Hafezi et al., 2019). Disruptores Endócrinos presentes no ambiente, como bisfenol A (BPA), ftalatos e pesticidas, podem mimetizar ou interferir na ação dos hormônios (Daronch et al., 2020; Albring et al., 2023; Soriano et al., 2023). Esses compostos estão associados ao desenvolvimento de subtipos hormonais de câncer de mama, devido à sua

capacidade de alterar os receptores hormonais, como os de estrogênio e progesterona.

A exposição a poluentes está associada ao aumento do risco de câncer (Cathey et al., 2023). Estudos apontam que a distribuição desigual de contaminantes ambientais, particularmente em áreas urbanas densamente povoadas, exacerba desigualdades na incidência de câncer, reforçando a importância de estratégias de mitigação ambiental como ferramenta de saúde pública. Neste sentido, o conhecimento dos padrões e das características dos cânceres permite a identificação de possíveis marcadores precoces para o desenvolvimento destas doenças.

1.2. Objetivos

1.2.1. Objetivo geral

Identificar módulos de co-expressão gênica relacionados ao câncer de mama triplo negativo (TNBC) por meio da análise de redes de co-expressão gênica (WGCNA), visando caracterizar vias moleculares correlacionadas ao subtipo em comparação aos subtipos hormonais.

1.2.2. Objetivos específicos

- Correlacionar módulos de co-expressão identificados à subtipos hormonais e não hormonais do câncer de mama;
- Caracterizar as vias metabólicas presentes dentro de cada módulo;
- Identificar interações proteína-proteína (PPI) relevantes dentro dos módulos de interesse;
- Identificar genes chaves (hubs);

1.3. Metodologia da Pesquisa

A pesquisa foi desenvolvida com base em uma abordagem quantitativa e exploratória, utilizando análise bioinformática de dados públicos de expressão gênica relacionados ao câncer de mama.

1.3.1. Classificação da Pesquisa

A pesquisa pode ser caracterizada sob diferentes perspectivas metodológicas. Podemos classificá-la como uma pesquisa quantitativa, pois envolve a análise numérica e estatística dos dados de expressão gênica, utilizando métodos computacionais para a identificação de padrões de co-expressão. Nossa pesquisa também pode ser classificada como uma pesquisa exploratória e descritiva, pois investiga novos padrões de expressão gênica e redes de co-expressão gênica, utilizando dados de transcriptomas de amostras tumorais. Além disso, é descritiva, pois caracteriza as diferenças de expressão gênica entre os subtipos de câncer de mama e apresenta a correlação de módulos gênicos com características clínicas.

1.3.2. Procedimentos Metodológicos

A pesquisa foi conduzida por meio de *análise in silico*, utilizando bases de dados públicas de dados genéticos de câncer de mama. Foi realizada uma abordagem de bioinformática para normalização dos dados, análise de expressão diferencial e identificação de redes de co-expressão gênica. Os procedimentos metodológicos utilizados na pesquisa foram estruturados em diferentes etapas, conforme descrito a seguir:

1.3.3. Procedimentos técnicos utilizados

Realizamos uma pesquisa bibliográfica e documental utilizando bases de dados sobre os temas câncer de mama, técnicas de análise de expressão gênica e métodos de bioinformáticas, para embasar a metodologia e interpretar os resultados. A pesquisa documental forneceu dados que foram obtidos do *The Cancer Genome Atlas* (TCGA-BRCA), um repositório público contendo perfis de transcriptomas de amostras tumorais de câncer de mama.

a. Obtenção dos dados:

- Foram acessados dados do TCGA-BRCA, utilizando a ferramenta TCGAbiolinks no ambiente de programação R.
- As amostras foram classificadas em dois grupos: subtipos hormonais e triplo negativo (TNBC), conforme descrito nos metadados clínicos do TCGA.

b. Processamento dos dados:

- Foi realizada a filtragem e normalização dos dados de expressão gênica utilizando o pacote DESeq2.
- Foram excluídos transcritos com baixa expressão para garantir a qualidade dos dados.

c. Análise de expressão diferencial:

- A expressão gênica diferencial foi avaliada entre os subtipos de câncer de mama utilizando a abordagem $\log_2$ Fold Change $\geq |1,5|$ e p-valor ajustado $\leq 0,05$.
- Os genes diferencialmente expressos foram visualizados utilizando gráficos Volcano Plot.

d. Construção das redes de co-expressão gênica:

- Aplicação da metodologia WGCNA para identificar módulos de genes co-expressos.
- Seleção do power adequado para a rede, garantindo que os padrões de conectividade seguissem uma distribuição de rede de escala livre.

e. Correlação dos módulos gênicos com características clínicas:

- Testes de correlação de Pearson foram aplicados para associar os módulos de co-expressão com variáveis clínicas das amostras.

f. Análise funcional dos módulos identificados:

- Foram realizadas análises de enriquecimento funcional de Gene Ontology (GO) para identificar vias metabólicas associadas a cada módulo de co-expressão.
- As redes de interação proteína-proteína (PPI) foram exploradas utilizando a base de dados STRING e visualizadas no software Cytoscape.

1.4. Considerações Éticas e Retorno à Sociedade

A pesquisa utilizou exclusivamente dados de domínio público, garantindo a conformidade com normas éticas e regulatórias para o uso de dados de expressão

gênica. Os resultados obtidos podem contribuir para o entendimento da biologia do câncer de mama, auxiliando na identificação de potenciais biomarcadores e possíveis alvos terapêuticos.

2. REFERENCIAL TEÓRICO

2.1. Epidemiologia e Ciências Ômicas no Estudo do Câncer

A epidemiologia é uma disciplina relativamente recente que evoluiu em resposta às transformações sociais e ao surgimento de novas doenças (FRÉROT et al., 2018). De maneira geral, essa área do conhecimento tem como principal objetivo avaliar os determinantes de saúde e sua distribuição em populações específicas (LIMA-COSTA; BARRETO, 2003). No contexto oncológico, a epidemiologia do câncer se dedica ao estudo da ocorrência e da distribuição da doença, bem como à análise dos fatores de risco envolvidos e das estratégias preventivas mais eficazes.

A interação entre fatores ambientais e genéticos desempenha um papel crucial no surgimento e na progressão do câncer. Nesse sentido, a epidemiologia genética surge como um campo interdisciplinar que investiga a influência dos fatores hereditários na etiologia do câncer, considerando a distribuição desses fatores na população (FEITOSA; KRIEGER, 2002).

Com o avanço das técnicas de biologia molecular, a epidemiologia passou por uma nova transformação, impulsionada pelo desenvolvimento das tecnologias de sequenciamento de nova geração (NGS). A incorporação de dados moleculares na análise epidemiológica resultou na emergência da epidemiologia molecular, um campo que integra abordagens clássicas da epidemiologia com dados genômicos, transcriptômicos e proteômicos, ampliando a capacidade de identificar biomarcadores e entender os mecanismos subjacentes à carcinogênese (BILAL; KLUTTS, 2022).

A base operacional para a identificação de marcadores moleculares na epidemiologia molecular é herdada da epidemiologia convencional: obter evidências de exposição a carcinógenos e detectar indivíduos com maior susceptibilidade quando expostos a esses agentes, seja por fatores genéticos ou adquiridos (WÜNSCH FILHO; GATTÁS, 2001). Além disso, essa abordagem possibilita a identificação de fatores que influenciam a presença, manifestação e progressão da doença (TÜMMLER, 2020).

A obtenção e análise de dados moleculares dependem de metodologias robustas de sequenciamento e armazenamento de dados. Atualmente, existem diversas bases de dados especializadas que organizam e disponibilizam informações sobre genomas, transcritos e proteínas, permitindo a realização de estudos de

montagem genômica, análise de expressão gênica, anotação de genes, modelagem estrutural, entre outros (RANGWALA et al., 2021).

Um dos repositórios mais amplamente utilizados para análise transcriptômica é o Gene Expression Omnibus (GEO), um banco de dados público internacional criado em 2000, que armazena informações sobre perfis de expressão gênica e metadados de estudos transcriptômicos. Além de oferecer acesso gratuito aos dados, a plataforma GEO disponibiliza ferramentas analíticas e de visualização que permitem explorar os resultados de forma interativa e contextualizada (CLOUGH; BARRETT, 2016). Atualmente, o GEO conta com aproximadamente 194.327 séries de dados, sendo que a maior parte delas foi gerada por tecnologias de sequenciamento de nova geração e microarranjos. No total, há cerca de 5.737.182 amostras depositadas, abrangendo diversos organismos. Entre elas, as amostras de *Homo sapiens* e *Mus musculus* representam cerca de 70% do total disponível (AGARWALA et al., 2016).

A estrutura do GEO é segmentada em duas sub-bases principais: GEODatasets, que armazena metadados sobre os estudos responsáveis pela geração dos dados, e GEOProfiles, que contém os perfis de expressão gênica e transcriptômica associados a diferentes condições experimentais. Essas informações permitem a identificação de genes diferencialmente expressos e sua posterior validação com base nos metadados disponíveis (CLOUGH; BARRETT, 2016).

2.2. Tecido Mamário e Neoplasias

O tecido mamário apresenta uma organização estrutural e funcional complexa, sendo constituído por quatro tipos principais de tecidos: conjuntivo, epitelial, muscular e nervoso. O conhecimento da anatomia e da composição tecidual normal da mama é essencial para a compreensão das suas patologias. Estruturalmente, a mama pode ser dividida em dois componentes principais: estrutural e funcional (CONCEIÇÃO, 2008). O componente estrutural é composto predominantemente por tecido conjuntivo e tecido adiposo, sendo este último um tipo especializado de tecido conjuntivo responsável pelo armazenamento de lipídios. O tecido conjuntivo, por sua vez, desempenha funções de suporte e proteção às demais estruturas da mama e é conhecido como estroma mamário, podendo ser subdividido em estroma interlobular e estroma intralobular. No tecido mamário jovem, o estroma interlobular é rico em tecido conjuntivo denso, enquanto com o envelhecimento há um aumento na

quantidade de tecido adiposo em detrimento do tecido conjuntivo (WEIL LARA; PÉREZ MARTÍNEZ; ROMERO MADRID, 2020).

O componente funcional da mama é representado pelo tecido epitelial glandular e pelo tecido muscular, sendo responsável pela produção e secreção do leite. Essas estruturas são denominadas parênquima mamário, que compreende os lóbulos mamários – formados por agrupamentos de ácinos e ductos – envolvidos pelo estroma interlobular. Os ácinos, compostos por uma camada interna de células epiteliais e uma camada externa de células mioepiteliais, desempenham a função secretora da glândula mamária, enquanto os ductos são responsáveis pela drenagem do leite (WEIL LARA; PÉREZ MARTÍNEZ; ROMERO MADRID, 2020).

2.2.1. Neoplasias Mamárias

As neoplasias são caracterizadas pelo crescimento anormal e descontrolado de células, podendo ser benignas ou malignas, conforme seu comportamento biológico. A distinção entre esses tipos baseia-se em critérios como capacidade de invasão tecidual, taxa mitótica, grau de diferenciação celular e potencial metastático (TAVASSOLI, 2008).

A nomenclatura das neoplasias segue padrões estabelecidos, levando em consideração a origem embriológica e tecidual do tumor. Em geral, as neoplasias benignas são designadas pelo sufixo "-oma", anexado ao nome da célula de origem (ex.: fibroma para tumores de fibroblastos e condroma para tumores de células cartilaginosas). Já as neoplasias malignas de origem mesenquimal recebem a denominação de sarcomas, enquanto aquelas de origem epitelial são classificadas como carcinomas, podendo ser subdivididas em adenocarcinomas (crescimento glandular) e carcinomas escamocelulares (crescimento a partir do epitélio escamoso) (DAI et al., 2017). Entre as principais neoplasias benignas da mama, destaca-se a adenose, uma condição caracterizada pelo crescimento anômalo dos ácinos, que pode resultar no aumento do volume lobular e na presença de fibrose intralobular. Outra lesão benigna comumente observada é o fibroadenoma, o tumor benigno mais frequente em mulheres em idade fértil, originado a partir da proliferação do estroma intralobular. Além disso, há as alterações fibrocísticas, que consistem em hiperplasias do estroma e do epitélio mamário, frequentemente associadas à fibrose interlobular e intralobular, podendo provocar alterações na textura e na sensibilidade do tecido mamário.

As neoplasias malignas da mama são predominantemente epiteliais e podem se manifestar de diferentes formas. O carcinoma ductal *in situ* e o carcinoma lobular *in situ* são considerados estágios pré-cancerígenos, nos quais o crescimento neoplásico permanece restrito aos ductos ou lóbulos mamários, sem invadir os tecidos adjacentes. Em contraste, o carcinoma ductal/lobular invasivo representa uma forma mais agressiva da doença, caracterizando-se pela infiltração no estroma e pela invasão de estruturas vizinhas, como adipócitos, vasos linfáticos e nervos (BADVE; GÖKMEN-POLAR, 2019). Diante da complexidade dessas neoplasias, o estudo das características histopatológicas do tecido mamário, aliado à análise de biomarcadores moleculares, desempenha um papel fundamental no diagnóstico, prognóstico e definição da abordagem terapêutica mais adequada para cada paciente.

2.3. Meio ambiente e as doenças neoplásicas

A interação entre fatores genéticos e ambientais é um dos principais determinantes do risco e da progressão do câncer, refletindo a complexidade da carcinogênese e os múltiplos mecanismos envolvidos nesse processo (Zahir et al., 2020). Estudos demonstram que, enquanto os fatores hereditários desempenham um papel importante na predisposição ao câncer (Tommasi et al., 2011), os fatores ambientais exercem um impacto substancial no desenvolvimento da maioria dos cânceres esporádicos (Khaledi & Ghasemi, 2021). No caso do câncer de mama, aspectos como idade da menarca, idade ao primeiro parto, menopausa, terapia de reposição hormonal, além de predisposição genética e status socioeconômico, explicam apenas parte da sua incidência. Assim, acredita-se que a exposição a contaminantes ambientais, combinada com fatores genéticos e hormonais, contribua significativamente para o risco de desenvolvimento da doença (Coyle, 2004).

Os micropoluentes ambientais emergem como uma preocupação crescente, especialmente devido ao seu potencial de afetar a saúde humana e contribuir para o aumento da incidência de doenças neoplásicas (Bila & Dezotti, 2007). Entre esses compostos, destacam-se os desreguladores endócrinos (EDCs), substâncias químicas amplamente distribuídas no ambiente e conhecidas por sua capacidade de interferir nos mecanismos hormonais e metabólicos do organismo. Os EDCs podem atuar de diversas formas, incluindo a mimetização ou bloqueio de hormônios naturais, a modulação da expressão gênica e a interferência nos processos de diferenciação celular (Diamanti-Kandarakis et al., 2009).

A exposição a EDCs pode ocorrer por diversas vias, incluindo ingestão de alimentos e água contaminados, inalação de compostos voláteis no ar e absorção cutânea. Entre os compostos mais estudados estão os ftalatos, bisfenol A (BPA), pesticidas organoclorados, poluentes industriais, metais pesados, entre outros. Estudos demonstram que a exposição crônica a essas substâncias está associada a um maior risco para diferentes tipos de câncer, incluindo câncer de mama, próstata, testículo e tireoide (Ibarluzea et al., 2004; Gore et al., 2015).

Um dos principais mecanismos pelo qual os EDCs contribuem para a carcinogênese é por meio da disfunção do sistema endócrino, que pode resultar no aumento da proliferação celular descontrolada e na perda de apoptose, processos fundamentais para o equilíbrio homeostático dos tecidos. A ativação do receptor de estrogênio por compostos como o bisfenol A tem sido implicada no desenvolvimento de tumores de mama hormônio-dependentes, uma vez que pode estimular a proliferação de células epiteliais mamárias e induzir alterações na expressão de genes associados ao ciclo celular e apoptose (Soto & Sonnenschein, 2010).

Além dos compostos sintéticos, metais pesados como o cádmio, chumbo e mercúrio também têm sido investigados por seu papel como potenciais desreguladores endócrinos. Evidências experimentais sugerem que o cádmio, mesmo em doses muito baixas, pode atuar como um imitador de estrogênio, ativando o receptor de estrogênio e promovendo proliferação celular descontrolada, aumentando o risco de câncer de mama (Johnson et al., 2003).

Outro aspecto relevante da influência ambiental no desenvolvimento do câncer de mama é a relação entre poluição atmosférica e incidência da doença. Estudos epidemiológicos sugerem que a exposição prolongada a hidrocarbonetos aromáticos policíclicos (PAHs) e a partículas finas (PM_{2.5}) está associada a um aumento na incidência de câncer de mama, especialmente em áreas urbanas altamente poluídas (White et al., 2019). Esses poluentes podem induzir mutações no DNA, promover estresse oxidativo e inflamação crônica, fatores que contribuem para a transformação maligna das células epiteliais mamárias (Loomis et al., 2013).

A disfunção do ritmo circadiano é outro fator ambiental que tem sido associado ao desenvolvimento de câncer de mama. Profissionais que trabalham em turnos noturnos apresentam um risco aumentado para esse tipo de câncer, possivelmente devido à supressão da melatonina, hormônio envolvido na regulação do ciclo

circadiano e na inibição da proliferação celular em tecidos mamários (Papantoniou et al., 2015).

Por outro lado, alguns fatores ambientais podem ter um efeito protetor contra o câncer de mama. Evidências indicam que a exposição à radiação solar e os níveis adequados de vitamina D estão associados a um menor risco da doença (Yin et al., 2010). A vitamina D atua como reguladora do crescimento celular e pode reduzir a proliferação descontrolada de células epiteliais mamárias, além de induzir a apoptose em células cancerígenas (Garland et al., 2006).

Dado o crescente acúmulo de evidências sobre a relação entre fatores ambientais e câncer, torna-se essencial adotar medidas preventivas para reduzir a exposição a compostos carcinogênicos. Entre as estratégias mais eficazes estão o uso de produtos livres de bisfenol A e ftalatos, a redução do consumo de alimentos contaminados por agrotóxicos e metais pesados, o incentivo ao uso de filtros de ar e purificadores de água, além do desenvolvimento de regulamentações mais rigorosas para substâncias químicas industriais (Gore et al., 2015).

Em suma, o impacto dos fatores ambientais na incidência de câncer de mama é um campo de pesquisa em expansão, com evidências crescentes de que a exposição a micropoluentes, desreguladores endócrinos e poluentes atmosféricos pode desempenhar um papel significativo na carcinogênese. Embora a genética seja um fator de risco importante, a interação entre predisposição genética e exposição ambiental parece ser determinante na maioria dos casos de câncer esporádico. Dessa forma, além de estratégias terapêuticas avançadas, a adoção de medidas preventivas e políticas públicas eficazes são essenciais para reduzir a carga global do câncer de mama e proteger a saúde das populações expostas a agentes carcinogênicos ambientais.

3. ARTIGO - ANÁLISE *IN SILICO* DO PERFIL TRANSCRIPTÔMICO DO CÂNCER DE MAMA: APLICAÇÃO DA WGCNA NA IDENTIFICAÇÃO DE GENES-CHAVE

RESUMO

O câncer de mama é a neoplasia maligna mais prevalente entre as mulheres em todo o mundo, sendo o subtipo triplo-negativo (TNBC) um dos mais agressivos e de difícil tratamento. Caracterizado pela ausência de receptores hormonais e HER2, o TNBC apresenta alta taxa de recorrência e baixa resposta a terapias convencionais, tornando essencial a identificação de biomarcadores prognósticos e alvos terapêuticos específicos. Neste contexto, abordagens transcriptômicas têm sido amplamente utilizadas para a compreensão dos mecanismos moleculares da doença. Neste estudo buscamos realizar uma análise *in silico* da expressão gênica de diferentes subtipos de câncer de mama, com ênfase no TNBC, utilizando a metodologia Weighted Gene Co-expression Network Analysis (WGCNA). Foram analisados dados públicos de transcriptoma oriundos do The Cancer Genome Atlas (TCGA), sendo identificados módulos de co-expressão gênica correlacionados a características clínicas e biológicas. A análise de interação proteína-proteína (PPI) foi integrada para compreender melhor as redes de regulação molecular envolvidas na progressão tumoral. Os resultados revelaram módulos de genes significativamente associados ao TNBC, destacando genes *hub* envolvidos em processos biológicos críticos, como proliferação celular, regulação do ciclo celular e vias de sinalização tumorais. A análise funcional indicou o envolvimento de vias metabólicas e inflamatórias associadas à progressão da doença. Além disso, foram identificados potenciais biomarcadores que podem contribuir para novas estratégias terapêuticas. Portanto, conclui-se que a abordagem baseada em redes de co-expressão gênica permitiu a identificação de genes chave e vias moleculares potencialmente relevantes para o TNBC. Esses achados reforçam a importância de análises *in silico* na compreensão dos mecanismos tumorais e na busca por novos alvos terapêuticos, podendo contribuir para o desenvolvimento de estratégias mais eficazes no tratamento do câncer de mama triplo-negativo.

Palavras-chave: Câncer de mama triplo-negativo; WGCNA; Biomarcadores; Transcriptoma; Análise *in silico*.

ABSTRACT

Breast cancer is the most prevalent malignant neoplasm among women worldwide, with the triple-negative subtype (TNBC) being one of the most aggressive and difficult to treat. Characterized by the absence of hormone receptors and HER2, TNBC presents a high recurrence rate and low response to conventional therapies, making the identification of prognostic biomarkers and specific therapeutic targets essential. In this context, transcriptomic approaches have been widely used to understand the molecular mechanisms of the disease. This study conducted an *in silico* analysis of gene expression in different breast cancer subtypes, with an emphasis on TNBC, using the Weighted Gene Co-expression Network Analysis (WGCNA) methodology. Public transcriptome data from The Cancer Genome Atlas (TCGA) were analyzed, identifying gene co-expression modules correlated with clinical and biological characteristics.

Protein-protein interaction (PPI) analysis was integrated to better understand the molecular regulatory networks involved in tumor progression. The results revealed gene modules significantly associated with TNBC, highlighting hub genes involved in critical biological processes such as cell proliferation, cell cycle regulation, and tumor signaling pathways. Functional analysis indicated the involvement of metabolic and inflammatory pathways associated with disease progression. Additionally, potential biomarkers were identified that may contribute to new therapeutic strategies. It is concluded that the gene co-expression network-based approach allowed the identification of key genes and molecular pathways potentially relevant to TNBC. These findings reinforce the importance of *in silico* analyses in understanding tumor mechanisms and in the search for new therapeutic targets, which may contribute to the development of more effective strategies for the treatment of triple-negative breast cancer.

Keywords: Triple-negative breast cancer; WGCNA; Biomarkers; Transcriptome; *In silico* analysis.

INTRODUÇÃO

O câncer de mama é a neoplasia maligna mais prevalente entre as mulheres em todo o mundo e representa um importante problema de saúde pública (BRAY et al., 2018). No Brasil, estima-se que o câncer de mama seja responsável por aproximadamente 29% dos casos de câncer em mulheres, sendo também a principal causa de mortalidade oncológica feminina (INCA, 2023). É uma doença heterogênea, composta por múltiplos subtipos moleculares, que diferem em termos de prognóstico, resposta terapêutica e alterações genéticas subjacentes (PEROU et al., 2000).

Dentre os subtipos de câncer de mama, destaca-se o câncer de mama triplo-negativo (TNBC, do inglês Triple-Negative Breast Cancer), que representa cerca de 15-20% dos casos diagnosticados e se caracteriza pela ausência de expressão dos receptores de estrogênio (α ER), progesterona (α PR) e HER2 (fator de crescimento epidérmico humano tipo 2) (LEHMANN et al., 2011). Este subtipo apresenta comportamento agressivo, maior risco de metástase e pior prognóstico em comparação a outros subtipos moleculares (DENT et al., 2007). A ausência de alvos terapêuticos clássicos limita as opções de tratamento a regimes baseados em quimioterapia, tornando essencial a identificação de novos biomarcadores e alvos terapêuticos para o manejo adequado dessa condição (TUTT et al., 2021).

No contexto da investigação biomolecular, a análise da expressão gênica tem sido amplamente utilizada para identificar diferenças moleculares entre os subtipos de

câncer de mama. Abordagens baseadas em sequenciamento de RNA (RNA-Seq) e outras técnicas transcriptômicas possibilitam a caracterização detalhada dos perfis de expressão gênica, contribuindo para a compreensão das alterações biológicas associadas à progressão tumoral (VANDANJOO et al., 2022). Dentre as ferramentas bioinformáticas empregadas para a análise de co-expressão gênica, destaca-se a metodologia Weighted Gene Co-expression Network Analysis (WGCNA), que permite a identificação de módulos de genes co-expressos e sua relação com características clínicas e biológicas (LANGFELDER; HORVATH, 2008).

A aplicação da WGCNA no estudo do TNBC pode revelar redes de regulação gênica fundamentais para a compreensão da agressividade tumoral e da resistência a tratamentos. Estudos recentes têm demonstrado que esse método é eficaz na identificação de genes centrais (hub genes) envolvidos na progressão da doença e na resposta terapêutica (ZHANG et al., 2018).

Paralelamente à caracterização molecular, fatores ambientais também desempenham um papel crucial na gênese do câncer de mama. Exposição a desreguladores endócrinos, como bisfenol A (BPA) e ftalatos, tem sido associada a alterações na expressão de genes relacionados à carcinogênese mamária (RUSSO; RUSSO, 2006). Da mesma forma, poluentes ambientais, incluindo metais pesados e hidrocarbonetos aromáticos policíclicos, podem atuar como carcinógenos ao promoverem estresse oxidativo e mutações no DNA (WHITE et al., 2019). Assim, uma abordagem integrativa que combine dados moleculares e exposição ambiental pode fornecer novas perspectivas sobre os mecanismos que impulsionam a progressão do TNBC.

Diante desse cenário, este estudo tem como objetivo realizar uma análise in silico do perfil de expressão gênica dos diferentes subtipos de câncer de mama, com ênfase no TNBC, utilizando WGCNA para identificar módulos de co-expressão gênica e suas correlações com características clínicas e biológicas. A identificação de assinaturas moleculares específicas pode contribuir para a elaboração de novas estratégias terapêuticas e biomarcadores prognósticos, auxiliando no desenvolvimento de terapias mais eficazes para o tratamento dessa neoplasia de alto risco.

MATERIAL E METODOS

Ambiente de análise

Todos os dados foram obtidos de base de dados públicos. Às análises foram realizadas por meio do uso da linguagem R, versão 4.2.3. Alguns pacotes foram utilizados em momentos distintos, como “TCGAbiolinks” para o acesso e filtragem dos dados do portal GDC, “DESeq2” para normalização da matriz de expressão gênica, e o pacote “WGCNA” para realização das análises

referentes a identificação dos módulos de co-expressão gênica, dentre outros, utilizados para manipulação de dados. Também foi utilizado o software “Cytoscape”, versão 3.9.1, para realização da análise de enriquecimento de vias.

Foram utilizados dados de expressão gênica do projeto “TCGA-BRCA”. As amostras foram separadas em dois grupos, o primeiro apresenta expressão positiva para receptores de membrana de estrógeno, progesterona e HER2 (No TPN) e o segundo é caracterizado pela ausência de todos os receptores de membrana, sendo classificado como triplo negativo (TPN-). As classificações foram confirmadas por imuno-histoquímica, depositadas na base de dados, no conjunto de informações clínicas de cada amostra.

Dados em bancos de dados públicos

Nas bases de dados tais como: TCGA, ENA, NCBI, estão disponibilizados diversos dados “brutos” de estudos de sequenciamento de câncer de mama em diferentes estágios. Neste sentido, estes dados foram acessados, e realizado novas avaliações de expressão gênica diferencial. Para isso, foram realizados *downloads* dos dados de sequenciamento das amostras.

Nessa primeira etapa, de acesso e filtragem dos dados, foi utilizado o pacote “TCGAbiolinks”, do repositório “Bioconductor”. A função “GDCquery” foi a primeira a ser utilizada, com os parâmetros “project” e “data.category” selecionando o projeto “TCGA-BRCA” e os dados clínicos, utilizando o comando “Clinical”, respectivamente. Com a busca realizada, a função “GDCdownload” foi utilizada para acessar os dados encontrados com o uso da primeira função, de acordo com os parâmetros utilizados. Em seguida, a função “GDCprepare” foi utilizada para formatar e orientar os dados baixados a um objeto, para realização da análise exploratória.

Com os dados clínicos, foi realizada a filtragem e classificação das amostras correspondentes a cada um dos grupos, obtendo o código de acesso de cada uma das amostras. Com as amostras filtradas e separadas em grupos, foi realizada a busca e posterior download das matrizes de expressão gênica para cada uma das amostras, com os comandos “GDCquery” e “GDCdownload” modificando o parâmetro “data.category” para “Transcriptome Profiling”.

A natureza dos dados permite classificá-los como uma série atemporal, visto que não foram reunidos em intervalos regulares e não são ordenados cronologicamente. A obtenção, e posterior análise, desses dados classifica nosso estudo como um estudo transversal.

Processamento dos dados

Os dados foram processados com o objetivo de obter a formatação correta para execução da normalização da matriz de expressão gênica, realizada pelo uso do pacote “DESeq2”, com a função “dds” e posteriormente com uso da função “vst”.

Com uso do pacote “tidyverse” os dados foram formatados no sentido de que todas as colunas da matriz contendo os valores de expressão gênica, sendo cada coluna, uma amostra, estivessem presentes, e na mesma ordem, que as linhas da matriz contendo os dados clínicos, que também representam uma amostra cada.

Quantificação e análise de expressão diferencial

O algoritmo DESeq2 do pacote Bioconductor / R FOI utilizado para estimar as diferenças globais na expressão gênica digital (GDE) entre os fenótipos. A análise resultou em valores de log2 Fold Change e valor *p* ajustado para cada transcrito detectado. Os transcritos foram considerados diferencialmente expressos quando $\log_2 \text{Fold Change} \geq |1,5|$ e o valor *p ajustado* (*padj*) ≤ 0.05 .

A representação da intensidade e o padrão de expressão foi realizada utilizando o pacote “EnhancedVolcano” para R.

Identificação dos módulos de co-expressão

A matriz de expressão gênica normalizada passou por uma filtragem para detectar transcritos outliers, utilizando a função ‘goodSamplesGenes’ do pacote ‘WGCNA’. Em seguida, aplicamos o mesmo procedimento às amostras para identificar

aquelas que apresentavam padrões de expressão fora da média. Nesta etapa, construímos um dendrograma usando o método de distância entre as médias dos valores de expressão e realizamos uma análise de componentes principais, utilizando a variância e o desvio padrão dos valores de expressão de cada amostra.

Os dados filtrados foram selecionados em relação a representatividade biológica de cada transcrito, sendo mantidos os counts que somaram ao menos quinze contagens em pelo menos setenta e cinco por cento das amostras, etapa recomendada pelo desenvolvedor do algoritmo. Foram obtidos os genes que possuem ao menos o valor mínimo de quinze contagens em setenta e cinco por cento das amostras.

O poder de conectividade entre os módulos foi selecionado mediante o uso da função “pickSoftThreshold”. A matriz de expressão foi posteriormente transformada em uma matriz de adjacência, e em seguida, a matriz de adjacência foi transformada em uma matriz de sobreposição topológica (TOM). Com base na TOM, utilizamos o método de agrupamento hierárquico de ligação média para agrupar genes, e definimos o número mínimo de genes para cada módulo. Em seguida, realizamos análise de cluster nos módulos e mesclamos os módulos mais próximos em um novo módulo

Análise de enriquecimento

Com os módulos de co-expressão gênica identificados, eles foram correlacionados com algumas variáveis clínicas disponibilizadas pela base de dados. A análise de correlação foi utilizada para identificar a magnitude da relação entre as variáveis, foi escolhido método de correlação de Pearson por se tratar de método paramétrico. As variáveis clínicas são fornecidas no formato categórico, para realização da análise de correlação, que correlaciona variáveis quantitativas contínuas, esses dados clínicos passaram por um processo de dicotomização, tornando-se variáveis quantitativas. Foram calculados os valores de correlação de cada modulo em relação a três variáveis clínicas, subtipo molecular da doença, origem étnica e presença de casos de câncer na família de cada paciente.

Com o uso do software “Cytoscape” Os transcritos presentes em cada módulo foram extraídos, e construída uma rede de vias metabólicas de acordo com a correlação de cada subtipo.

RESULTADOS

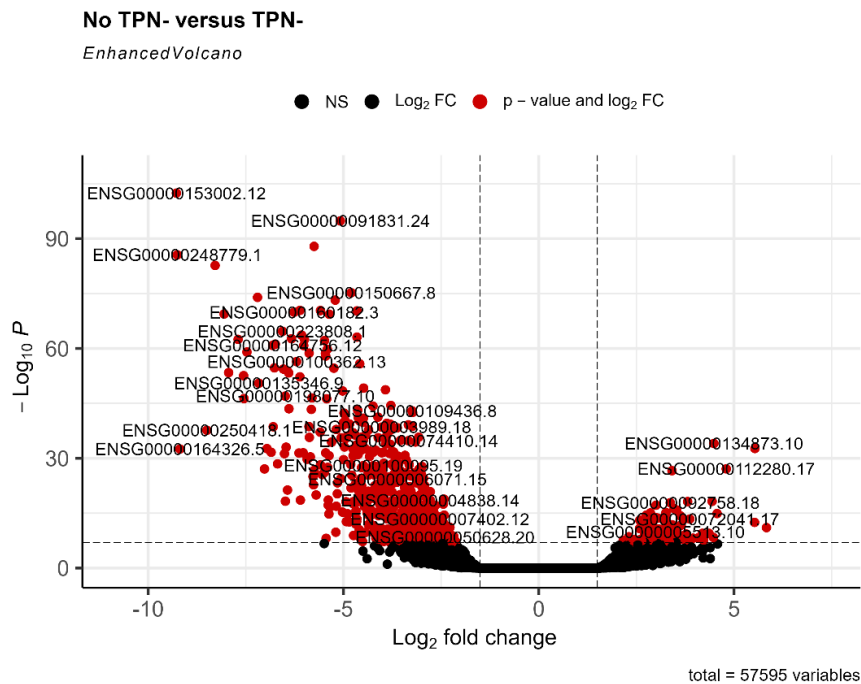
Obtenção dos dados

Com a utilização do pacote “TCGABiolinks”, desenvolvido para facilitar o acesso aos dados da base TCGA (The Cancer Genome Atlas), foi possível acessar 1.032 casos com perfis de transcriptoma, após a filtragem das amostras que apresentaram resultados inconclusivos na imuno-histoquímica. Com o objetivo de identificar e comparar módulos de coexpressão gênica entre o subtipo triplo negativo e os subtipos que expressam receptores hormonais, as amostras foram classificadas com base nas informações moleculares disponíveis na categoria “patients” do conjunto de dados clínicos da TCGA. Foram identificados 115 casos do subtipo triplo negativo e 917 casos classificados como subtipo não triplo negativo, ou seja, contendo receptores hormonais, de acordo com o protocolo de imuno-histoquímica fornecido. Após a filtragem de genes sem leituras ou com baixa contagem, foram obtidos 57.595 genes.

Análise de genes diferencialmente expressos

Com as amostras filtradas e categoricamente separadas, foi utilizado o pacote “DESeq2” para obtenção de genes diferencialmente expressos. Com os parâmetros de significância estatística e de *foldchange* definidos em $p < 0,05$ e $|1,5|$, respectivamente, foram identificados 1425 genes diferencialmente expressos, sendo 517 *up* regulados, e 908 *down* regulados, tendo o grupo “NO TPN-” como referência, representado pela Figura 1. Os primeiros dez genes com maiores valores de *foldchange* estão representados na tabela 1.

Figura 1— *Volcano plot* representando genes diferencialmente expressos entre os grupos triplo negativo e não triplo negativo



Fonte: Autor, 2024.

Tabela 1 - Top 10 genes com maiores valores de FC triplo negativo e não triplonegativo

Categoria	Gene Name	Log FC
DOWN regulados	CPB1	-9.276652
	ENSG00000248779	-9.255597
	CARTPT	-9.188109
	ENSG00000250418	-8.497308
	LINC02224	-8.285713
UP Regulados	PRR27	5.833356
	RANBP20P	5.537014
	LINC03061	5.528270
	COL9A1	4.803041
	PAGE1	4.580452

Fonte: Autor, 2024.

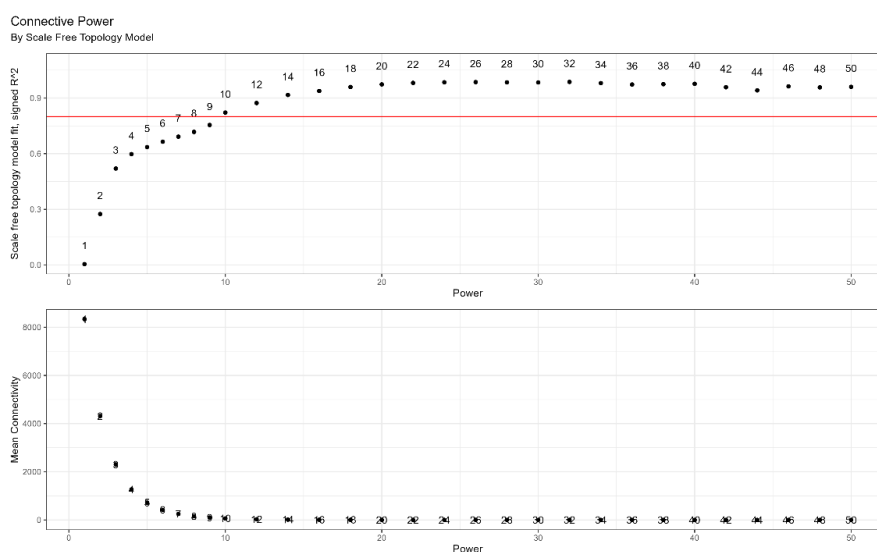
Identificação dos módulos de co-expressão gênica

A matriz de counts foi filtrada em relação representatividade biológica de cada transcrito. Transcritos que apresentaram contagem superior a quinze em pelo menos setenta e cinco por cento das amostras, ou seja, 774 do total de 1032, foram mantidos, os demais, descartados. Toda essa etapa de filtragem foi realizada por meio linha de comando na linguagem R, de acordo com o protocolo da análise.

Após a filtragem foram obtidos 16.641 transcritos, que foram submetidos a normalização “vst” e posteriormente a clusterização para identificação dos módulos.

Para que a clusterização seja eficiente, ou seja, para que as interações detectadas entre os transcritos sejam relevantes, foi selecionado um valor de conectividade, identificado como “*power*” dentro da metodologia WGCNA. Esse valor é calculado testando-se vários valores de “*power*” distintos dentro do conjunto de dados. É escolhido aquele valor que mais aproxima a distribuição das interações entre os transcritos como uma rede de escala livre. Para esse conjunto de dados, o valor de “*power*” selecionado foi de dez, como demonstrado pela Figura 2. Esse parâmetro foi utilizado para ponderar a conectividade entre os transcritos. Interações entre transcritos que apresentaram maiores valores de conectividade foram destacadas, enquanto interações de menor conectividade foram suprimidas.

Figura 2— Teste de conectividade segundo o modelo de rede de livre escala



Fonte: Autor, 2024.

Após identificar e ponderar a similaridade entre os transcritos, realizou-se o cálculo da matriz de sobreposição topológica (TOM), que mede o grau de compartilhamento de vizinhos comuns entre pares de transcritos. Esse procedimento vai além da correlação direta entre dois transcritos, pois considera também a relação de cada um com transcritos vizinhos. Assim, ainda que dois transcritos (A e B) não apresentem uma correlação direta expressiva, a TOM possibilita que esses transcritos tenham uma sobreposição topológica elevada caso ambos compartilhem muitos vizinhos em comum. Tal abordagem permite detectar agrupamentos de transcritos que, mesmo sem uma correlação direta entre si, podem fazer parte de um mesmo processo biológico ao apresentarem fortes correlações com um terceiro transcrito em comum. Em seguida, os valores obtidos foram convertidos em uma matriz de dissimilaridade ($1 - \text{TOM}$), onde transcritos com alta sobreposição topológica apresentam valores baixos de dissimilaridade, enquanto aqueles que compartilham poucos vizinhos possuem valores de dissimilaridade altos. Essa matriz de dissimilaridade foi utilizada para a aglomeração hierárquica dos transcritos, culminando na identificação de módulos de co-expressão.

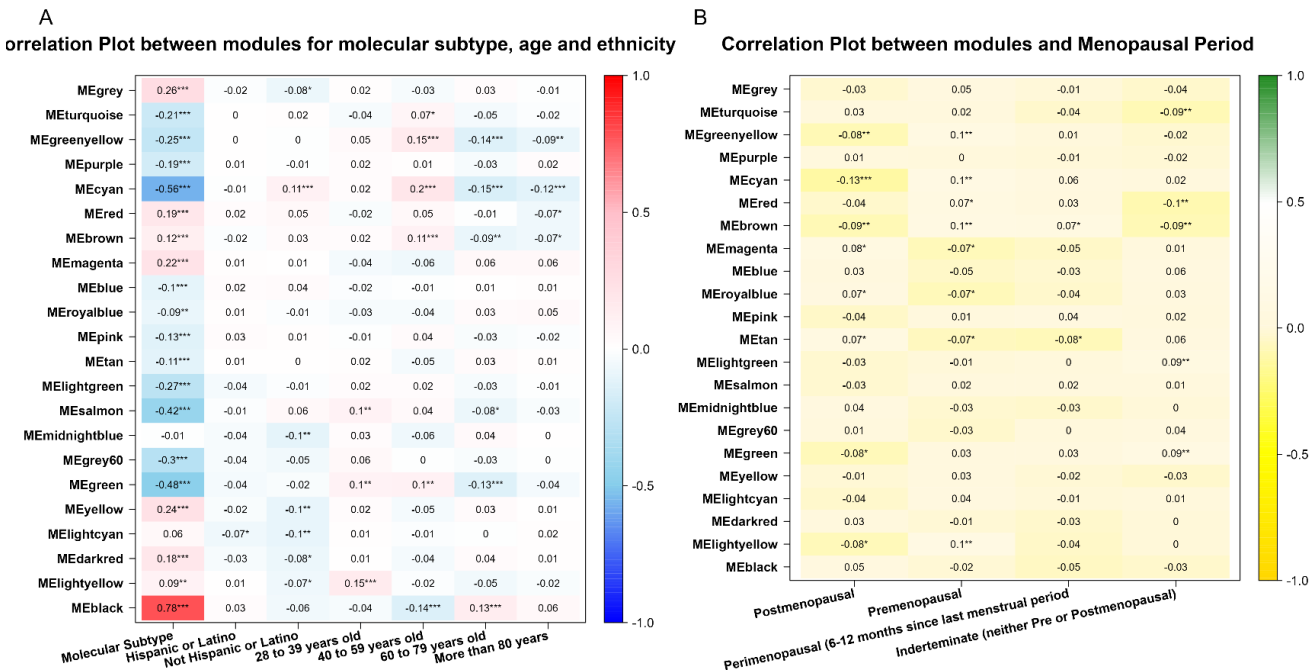
Foram identificados 21 módulos de co-expressão, para um total de 16641 transcritos. Mencionar quantos transcritos por modulo.

Correlacionando módulos as variáveis clínicas

Os módulos identificados foram submetidos a testes de correlações de Pearson, com algumas variáveis clínicas disponibilizados pelo banco de dados. Dentre as variáveis, foram testadas as correlações entre os módulos e o subtipo molecular, origem étnica, idade e período de menopausa.

O módulo preto, “MEblack”, apresentou correlação com o subtipo molecular que expressa receptores de membrana, NO TPN-. Enquanto o módulo ciano, “MEcyan”, apresentou correlação com o subtipo que não expressa receptor de membrana, TPN-. As outras variáveis correlacionadas não apresentaram grandes valores de correlação, aproximando-se de valores em torno de 0, demonstrado na Figura 3.

Figura 3— Matriz de correlações entre os módulos e as variáveis clínicas

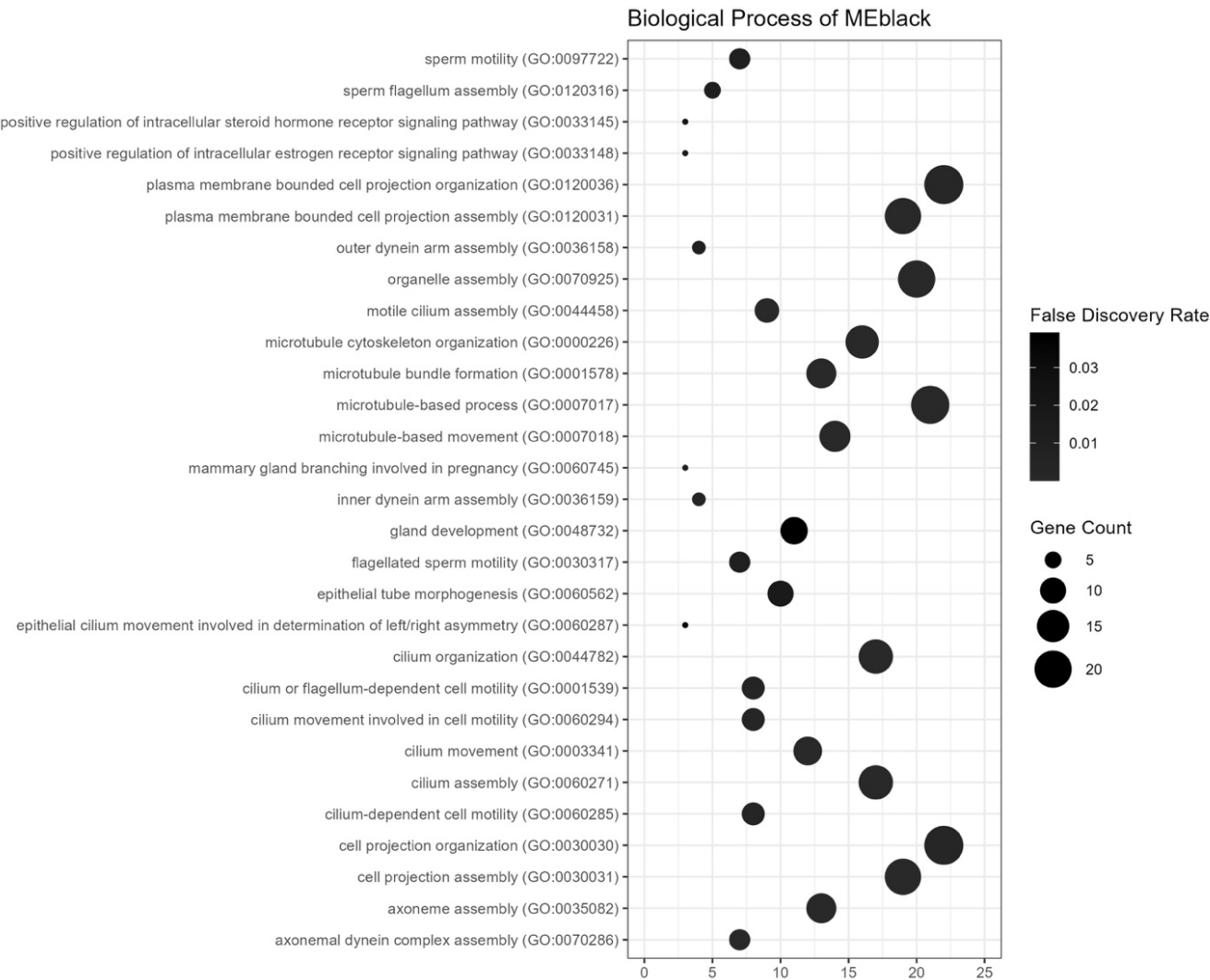


Fonte: Autor, 2024.

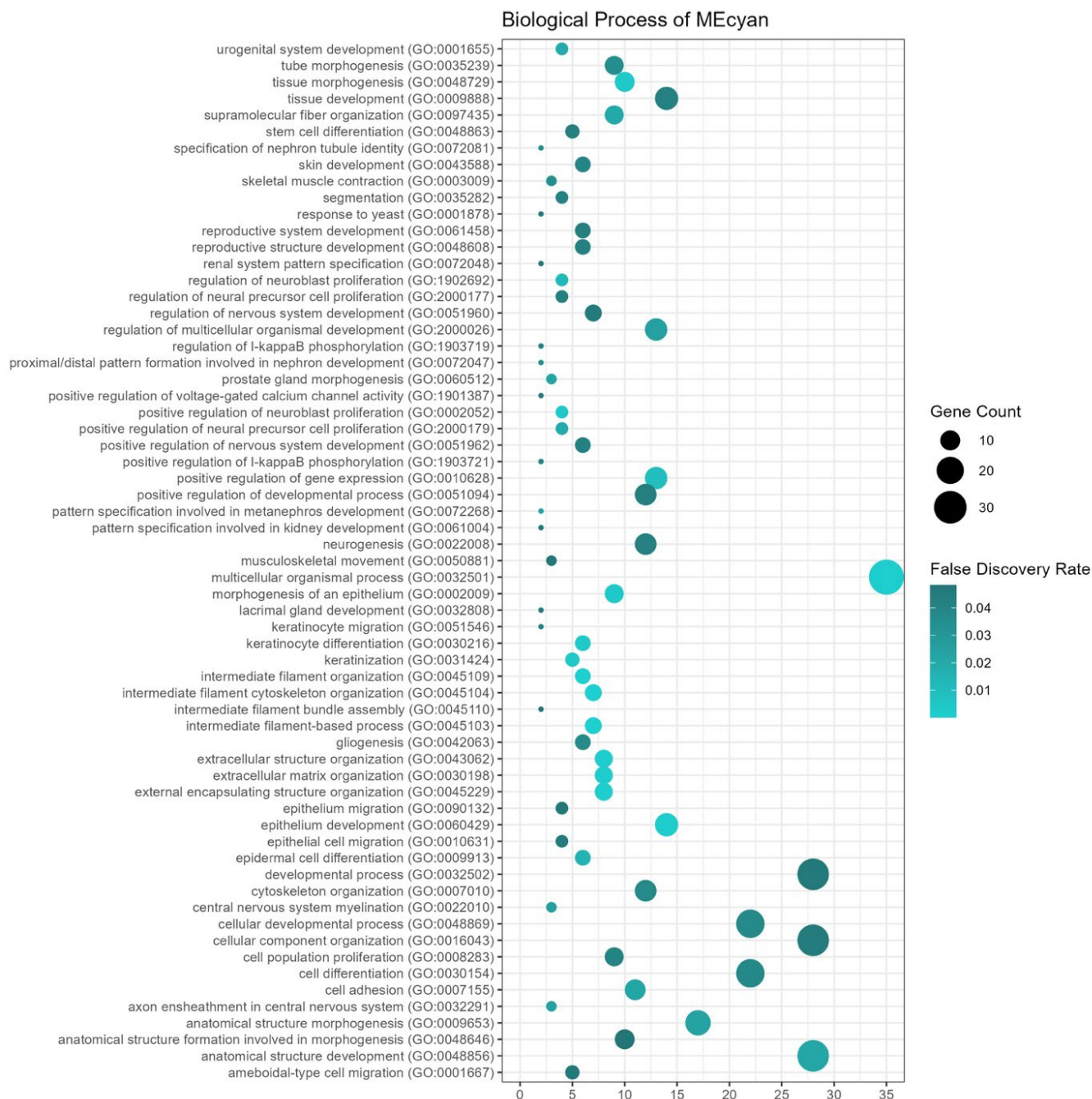
Análise de enriquecimento

Os módulos de interesse, “MEblack” e “MEcyan” foram submetidos a análises de enriquecimento (Gene Ontology). Os módulos foram enriquecidos segundo processos biológicos, como demonstram as figuras 4 e 5.

Figura 4— Processos Biológicos identificados para o modulo “MEblack”



Fonte: Autor, 2024.

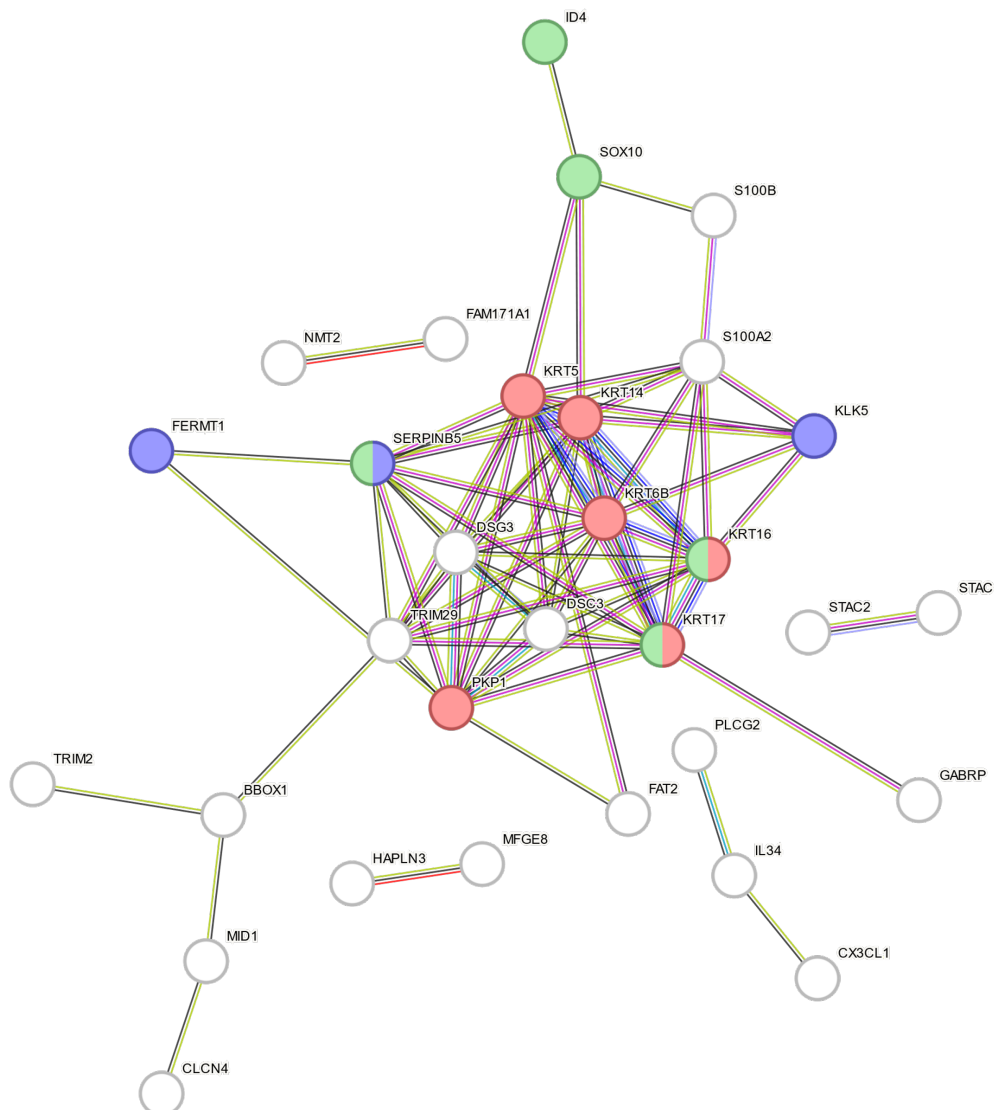
Figura 5— Processos Biológicos identificados para o módulo “MEcyan”

Fonte: Autor, 2024.

Redes de interação proteína-proteína

Redes de interação proteína-proteína (PPI) foram construídas para os módulos “MEblack” e “MEcyan”, utilizando a base de dados STRING. Foram identificadas 73 interações entre 54 proteínas presentes, com um grau médio de conexões de 2,70 no “MEcyan”, enquanto o “MEblack” apresentou 180 interações

Figura 7— Rede de interações proteína-proteína (PPI) do “MEcyan”



Fonte: Autor, 2024.

Legenda: Círculos marcados em vermelho indicam a participação delas nas vias metabólicas envolvidas com a organização das proteínas do citoesqueleto. Círculos marcados em azul indicam a participação delas em vias metabólicas envolvidas com a proliferação celular. Círculos marcados em verde indicam a participação delas em vias metabólicas envolvidas com a organização da matriz extracelular.

DISCUSSÃO

O câncer de mama triplo negativo (TNBC) abrange um grupo heterogêneo de doenças fundamentalmente diferentes, com diferentes perfis histológicos, genômicos e imunológicos, que são agregados sob este termo devido à falta de expressão do receptor de estrogênio, receptor de progesterona e receptor do fator de crescimento epidérmico humano do tipo 2 (Derakhshan & Reis-Filho, 2022). Entretanto, por mais que essa classificação em torno da ausência dos três principais receptores hormonais se mantenha como parâmetro central na classificação do TNBC, com base nas evidências mais recentes, o TNBC seria mais bem considerado como um termo que abrange uma variedade de tumores com diferenças genéticas, transcricionais, histológicas e clínicas marcadas (Zhao et al., 2020).

As análises de genes diferencialmente expressos (DEGs) permitem comparar os níveis de expressão de milhares de transcritos entre amostras submetidas a diferentes condições (Costa-Silva et al., 2017). Através de nossas análises pudemos identificar 1425 genes diferencialmente expressos, sendo 517 *up* regulados, e 908 *down* regulados. Embora a identificação de DEGs tenha proporcionado avanços significativos na busca por genes associados à manifestação de determinados fenótipos, essa abordagem, ao considerar os genes isoladamente e como entidades independentes, desconsidera a complexa rede de interações que regula a expressão gênica. A regulação da expressão gênica é um processo altamente coordenado, resultado da interação de múltiplos genes e vias metabólicas. Assim, para compreender de forma mais abrangente a manifestação de fenótipos, como a desdiferenciação celular ou a transição epitélio-mesenquimal—processos frequentemente associados a células tumorais avançadas—, é fundamental realizar análises que identifiquem vias metabólicas alteradas, indo além da detecção de genes diferencialmente expressos (Carthew, 2020).

A alta taxa de mortalidade registrada para o TNBC faz com que o diagnóstico molecular precoce desempenhe um papel crítico na obtenção de um prognóstico favorável (Kim et al., 2012). Em nossas análises, o módulo preto, denominado “MEblack”, apresentou a maior correlação com o subtipo hormonal do câncer de mama, caracterizado pela expressão de receptores de membrana clássicos, como os receptores de estrogênio, progesterona e o receptor do fator de crescimento epidérmico tipo 2 (HER2). Segundo Hoeflerlin et al. (2013) os cânceres de mama triplo-

negativos (TNBCs) são altamente heterogêneos e por essa razão apresentam uma sobreposição considerável com outros tipos de cânceres, incluindo o HER2.

Nossa análise de enriquecimento de vias, identificou processos biológicos relacionados à recepção de estímulos hormonais e ao desenvolvimento do tecido mamário. Os processos biológicos encontrados puderam ser organizados em categorias como: “resposta ao hormônio”, “desenvolvimento de glândulas”, “processo maternal envolvido na gravidez feminina” e “regulação intracelular positiva da via de sinalização do receptor de estrogênio”. Além de sua relação direta com a atividade hormonal, esses processos refletem outro aspecto importante: o estímulo à proliferação celular do tecido mamário. Considerando que TNBCs não apresentam receptores em sua constituição, podemos considerar a existência de um mecanismo autócrino de estimulação do crescimento (Köster et al., 2008).

Diversos estudos (Zhang et al., 2020; Xia et al., 2021; Singh et al., 2023) têm destacado a importância dos RNAs não codificantes nos cânceres de tecido mamário. Em nossas análises, identificamos outros tipos de RNAs, que podem desempenhar papéis importantes na regulação pós-transcricional. Dos 150 transcritos agrupados no módulo “MEblack”, 132 foram classificados como RNAs codificantes e 18 como RNAs não codificantes. Entre estes, destaca-se o longo RNA intergênico não codificante LINC03061, que apresentou níveis de expressão aumentados nos cânceres não triplo-negativos (NO-TNBC), conforme relatado por Tang et al. (2017).

A análise da rede de interações proteína-proteína (PPI) revelou uma diversidade de processos relacionados ao desenvolvimento glandular, recrutamento de proteínas do citoesqueleto e motilidade celular. Algumas proteínas destacaram-se pelo alto número de conexões com outras proteínas. Dentre elas, a ESR1 (receptor de estrogênio alfa) foi identificada como um elemento central do módulo. É amplamente reconhecido na literatura científica que essa proteína desempenha um papel crucial nas vias de sinalização hormonal, particularmente na via de sinalização do estrogênio. Essa via regula genes envolvidos no controle do ciclo celular. Além disso, a análise revelou interações de co-expressão com proteínas como ERBB4, PGR e GATA3.

A proteína tirosina quinase ERBB4 e o receptor de progesterona PGR também estão associadas à sinalização hormonal e ao desenvolvimento do tecido mamário. Já a GATA3, um fator de transcrição específico de células T, participa de processos de diferenciação e ativação celular, frequentemente mediados por respostas

imunológicas inflamatórias. Essa interação sugere que o aumento da expressão de genes pró-proliferativos pode desencadear o recrutamento de linfócitos para o microambiente tumoral, como resposta inflamatória induzida pela proliferação de células anaplásicas. Além disso, as proteínas CCDC40 e DNAAF1 desempenharam papéis centrais no módulo “MEblack”, estando diretamente associadas à motilidade celular. Por outro lado, o módulo “MEcyan” apresentou a maior correlação com o subtipo não triplo-negativo (TNBC). A análise de agrupamento identificou 56 transcritos neste módulo, dos quais 54 foram classificados como RNAs codificantes e 2 como RNAs não codificantes. Entre as proteínas identificadas, a família KRT destacou-se pela alta conectividade com outras proteínas. Esta família está amplamente associada a processos relacionados à estrutura celular e à regulação de funções epiteliais.

Todos esses resultados implicam na necessidade de testes de validação, correlações entre genes e proteínas aos subtipos hormonais ou não hormonais, foram obtidos através de análises *in silico*. Essa abordagem reduz o caminho entre os testes de bancada, *in situ* ou *ex situ*, e as descobertas de novos biomarcadores, tendo em vista que serão testados os genes, ou proteínas, que apresentam maior correlação, ou seja, restringe a lista de candidatos moleculares.

A abordagem holística permite a interpretação de um determinado contexto de forma complexa, considerando não somente as partes, mas também a interação entre elas. Dentro do contexto tumoral, fatores ambientais, como desreguladores endócrinos, maior tempo de exposição à radiação solar, maior contato com substâncias carcinogênicas, dentre outros, são partes que também devem ser analisadas suas interações e seu papel dentro do fenótipo tumoral. Nesse sentido, a depender da origem dos dados de sequenciamento, os módulos identificados podem vir a serem diferentes, devido ao contexto ambiental distinto no qual as amostras estão inseridas. Por esse motivo, esse tipo de abordagem pode apresentar resultados ligeiramente distintos a depender do conjunto de dados analisados.

CONCLUSÃO

O presente estudo buscou identificar módulos de co-expressão gênica associados ao câncer de mama triplo negativo (TNBC) utilizando a metodologia WGCNA. Os achados destacaram diferenças significativas entre os módulos relacionados ao TNBC e aos subtipos hormonais, particularmente no que tange às vias metabólicas e padrões de expressão gênica. Essas descobertas contribuem para a ampliação do conhecimento sobre as bases moleculares do TNBC, potencialmente auxiliando na identificação de biomarcadores e no desenvolvimento de terapias direcionadas. Estudos futuros deverão explorar mais profundamente as interações moleculares e validar os alvos terapêuticos identificados em contextos clínicos, visando melhorar o diagnóstico e o tratamento desta condição desafiadora.

REFERÊNCIAS

- ALIZADEH, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. **Nature**, v. 403, n. 6769, p. 503–511, 2000. DOI: 10.1038/35000501.
- LANGFELDER, P.; HORVATH, S. WGCNA: an R package for weighted correlation network analysis. **BMC Bioinformatics**, v. 9, p. 559, 2008. DOI: 10.1186/1471-2105-9-559.
- MILLER, J. A. *et al.* Strategies for aggregating gene expression data: the collapseRows R function. **BMC Bioinformatics**, v. 12, p. 322, 2011. DOI: 10.1186/1471-2105-12-322.
- RITCHIE, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. **Nucleic Acids Research**, v. 43, n. 7, e47, 2015. DOI: 10.1093/nar/gkv007.
- SUTTORP, N. *et al.* Gene expression analysis of tumor and normal tissues identifies key genes and pathways associated with breast cancer subtypes. **Frontiers in Oncology**, v. 10, p. 118, 2020. DOI: 10.3389/fonc.2020.00118.
- YANG, Y. *et al.* Systematic network analysis reveals distinct gene modules and pathways associated with breast cancer subtypes. **Frontiers in Genetics**, v. 12, p. 726791, 2021. DOI: 10.3389/fgene.2021.726791.
- ZHANG, B.; HORVATH, S. A general framework for weighted gene co-expression network analysis. **Statistical Applications in Genetics and Molecular Biology**, v. 4, n. 1, p. Article17, 2005. DOI: 10.2202/1544-6115.1128.

4. CONCLUSÃO

O presente trabalho buscou compreender, sob uma perspectiva sistêmica, os aspectos moleculares e ambientais associados ao câncer de mama, com ênfase no subtipo triplo negativo (TNBC). A dissertação integrou elementos conceituais e metodológicos de diferentes naturezas — da revisão teórica sobre as interações entre meio ambiente e doenças neoplásicas até a aplicação de abordagens bioinformáticas de alta complexidade, como a análise de redes de co-expressão gênica (WGCNA).

A investigação *in silico* permitiu identificar módulos de co-expressão gênica significativamente associados ao TNBC e aos subtipos hormonais, revelando processos biológicos relacionados à sinalização hormonal, organização celular e proliferação tumoral. Esses resultados reforçam o potencial da análise de redes para a descoberta de genes-chave e vias moleculares relevantes, contribuindo para a identificação de novos biomarcadores e alvos terapêuticos.

O estudo dialoga com a necessidade de compreender como fatores ambientais — como desreguladores endócrinos, poluentes e desigualdades regionais — podem modular a biologia tumoral e influenciar os padrões epidemiológicos do câncer de mama no Brasil. Assim, o trabalho contribui não apenas para o avanço do conhecimento em biologia molecular e bioinformática, mas também para a formulação de políticas públicas de prevenção e manejo oncológico com base em evidências científicas.

Por fim, a dissertação demonstra a importância das abordagens integrativas na pesquisa biomédica, evidenciando que a compreensão do câncer demanda o diálogo entre genética, ambiente e tecnologia. Estudos futuros, incluindo validações experimentais e análises com dados clínicos nacionais, poderão consolidar os achados aqui apresentados e ampliar sua aplicabilidade na medicina personalizada e na saúde pública.

REFERÊNCIAS

- AGARWALA, R. *et al.* Database resources of the National Center for Biotechnology Information. **Nucleic Acids Research**, v. 44, n. D1, p. D7–D19, 2016. DOI: 10.1093/nar/gkv1290.
- BADVE, S.; GÖKMEN-POLAR, Y. HER2-negative breast cancer: the relevance of hormone receptor expression. **Breast Cancer Research**, v. 21, n. 1, p. 200, 2019. DOI: 10.1186/s13058-019-1244-5.
- BILA, D. M.; DEZOTTI, M. Micropoluentes em águas residuais domésticas: ocorrência e tratamento. **Química Nova**, v. 30, n. 3, p. 651–658, 2007. DOI: 10.1590/S0100-40422007000300027.
- CLOUGH, E.; BARRETT, T. The Gene Expression Omnibus database. **Methods in Molecular Biology**, v. 1418, p. 93–110, 2016. DOI: 10.1007/978-1-4939-3578-9_5.
- CONCEIÇÃO, G. M. **Anatomia e fisiologia da mama: estrutura e função da glândula mamária**. São Paulo: Atheneu, 2008.
- COYLE, Y. M. The effect of environment on breast cancer risk. **Breast Cancer Research and Treatment**, v. 84, n. 3, p. 273–288, 2004. DOI: 10.1023/B:BREA.0000019969.56994.4f.
- DAI, X. *et al.* Classification of breast cancer based on proteomics. **Clinical Proteomics**, v. 14, n. 5, p. 1–12, 2017. DOI: 10.1186/s12014-017-9133-5.
- DIAMANTI-KANDARAKIS, E. *et al.* Endocrine-disrupting chemicals: an Endocrine Society scientific statement. **Endocrine Reviews**, v. 30, n. 4, p. 293–342, 2009. DOI: 10.1210/er.2009-0002.
- FEITOSA, M. F.; KRIEGER, J. E. Epidemiologia genética: conceitos e aplicações. **Revista de Saúde Pública**, v. 36, n. 3, p. 325–331, 2002. DOI: 10.1590/S0034-89102002000300012.
- FRÉROT, M. *et al.* What is epidemiology? Changing definitions of epidemiology 1978–2017. **PLoS ONE**, v. 13, n. 12, e0208442, 2018. DOI: 10.1371/journal.pone.0208442.
- GORE, A. C. *et al.* EDC-2: The Endocrine Society's Second Scientific Statement on Endocrine-Disrupting Chemicals. **Endocrine Reviews**, v. 36, n. 6, p. E1–E150, 2015. DOI: 10.1210/er.2015-1010.
- IBARLUZEA, J. M. *et al.* Breast cancer risk and the combined effect of environmental estrogens. **Cancer Causes & Control**, v. 15, n. 6, p. 591–600, 2004. DOI: 10.1023/B:CACO.0000036155.60851.09.
- JOHNSON, M. D. *et al.* Cadmium mimics the in vivo effects of estrogen in the uterus and mammary gland. **Nature medicine**, v. 9, n. 8, p. 1081–1084, 2003. DOI: 10.1038/nm902.

LIMA-COSTA, M. F.; BARRETO, S. M. Tipos de estudos epidemiológicos: conceitos básicos e aplicações na área do envelhecimento. **Epidemiologia e Serviços de Saúde**, v. 12, n. 4, p. 189–201, 2003. DOI: 10.5123/S1679-49742003000400003.

LOOMIS, D. et al. The carcinogenicity of outdoor air pollution. **The Lancet Oncology**, v. 14, n. 13, p. 1262–1263, 2013. DOI: 10.1016/S1470-2045(13)70487-X.

PAPANTONIOU, K. *et al.* Night shift work and breast cancer risk: what do the data say? **European Journal of Cancer**, v. 50, n. 9, p. 1578–1589, 2015. DOI: 10.1016/j.ejca.2014.12.015.

SOTO, A. M.; SONNENSCHN, C. Environmental causes of cancer: endocrine disruptors as carcinogens. **Nature Reviews Endocrinology**, v. 6, n. 7, p. 363–370, 2010. DOI: 10.1038/nrendo.2010.87.

TAVASSOLI, F. A. **Pathology of the breast**. 3. ed. New York: Springer, 2008.

WHITE, A. J. *et al.* Air pollution and breast cancer risk in the Sister Study cohort. **Environmental Health Perspectives**, v. 127, n. 10, p. 107002, 2019. DOI: 10.1289/EHP5137.

WÜNSCH FILHO, V.; GATTÁS, G. J. F. Epidemiologia genética do câncer. **Revista Brasileira de Epidemiologia**, v. 4, n. 2, p. 130–142, 2001. DOI: 10.1590/S1415-790X2001000200005.