RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

ANÁLISE DE GENES DIFERENCIALMENTE EXPRESSOS EM AMOSTRAS DE CÂNCER DE MAMA DO *SEQUENCE READ ARCHIVE* (SRA)

ANALYSIS OF DIFFERENTIALLY EXPRESSED GENES IN BREAST CANCER SAMPLES FROM THE SEQUENCE READ ARCHIVE (SRA)

ANÁLISIS DE GENES EXPRESADOS DIFERENCIALMENTE EN MUESTRAS DE CÁNCER DE MAMA DEL SEQUENCE READ ARCHIVE (SRA)

Matheus Correia Casotti[1], Giulia Maria Giacinti[1], Aléxia Stefani Siqueira Zetum[1], Camilly Victória Campanharo[1], Karen Ruth Michio Barbosa[1], Flavia de Paula[1], Débora Dummer Meira[1], Iúri Drumond Louro[1]

e534955
https://doi.org/10.47820/recima21.v5i3.4955

PUBLICADO: 03/2024

RESUMO

O câncer/neoplasias de mama (CM) é uma doença altamente incidente em mulheres com milhões de novos casos a cada ano. Dentre os avanços tecnológicos destaca-se a tecnologia de RNA-seq que permitiu compreender melhor a expressão gênica, possibilitando desvendar as interações proteicas entre tumores de mama em estágio inicial e recorrente (pós-mastectomia). Novas ferramentas baseadas em bioinformática surgiram para acompanhar o avanço dos sequenciamentos, e tem-se como principais exemplos as plataformas *online* de análise *Galaxy* e WebGestalt. Além disso, foi estabelecido o *Sequence Read Archive* (SRA) como um repositório público para os dados de sequência de nova geração, assim como foi estabelecido o uso do repositório de dados genômicos funcionais o *Gene Expression Omnibus* (GEO). Neste trabalho, utilizando análise de sequenciamento de RNA total, foi possível demonstrar comparações generalizadas do CM em um estágio inicial com CM recorrente. Além disso, utilizou-se Ontologia Genética (GO), KEGG e Reactome para avaliar as relações funcionais e vias aprimoradas entre CM em um estágio inicial e CM recorrente pós-mastectomia. Em conclusão, através do desenvolvimento deste estudo foi possível descobrir novos biomarcadores que poderão ser utilizados como futuros alvos terapêuticos, possibilitando um melhor diagnóstico e prognóstico no CM visando à melhoria da sobrevida global das pacientes.

PALAVRAS-CHAVE: Perfilação da Expressão Gênica. Biologia Computacional. Neoplasias da Mama. RNA-seq.

ABSTRACT

Breast neoplasms/cancer (BC) is a highly prevalent disease in women with millions of new cases each year. Among the technological advances, RNA-seq technology stands out, which has allowed us to better understand gene expression, making it possible to unveil protein interactions between early and recurrent (post-mastectomy) breast tumors. New tools based on bioinformatics have emerged to follow the advancement of sequencing, with the main examples being the online analysis platforms Galaxy and WebGestalt. Additionally, the Sequence Read Archive (SRA) was established as a public repository for next-generation sequence data, as was the use of the Gene Expression Omnibus (GEO) functional genomic data repository. In this work, using total RNA sequencing analysis, it was possible to demonstrate generalized comparisons of early-stage CM with recurrent CM. Furthermore, Gene Ontology (GO), KEGG and Reactome were used to evaluate the functional relationships and improved pathways between early-stage CM and post-mastectomy recurrent CM. In conclusion, through the development of this study it was possible to discover new biomarkers that could be used as future therapeutic targets, enabling a better diagnosis and prognosis in BC aiming to improve the overall survival of patients.

KEYWORDS: *Gene Expression Profiling. Computational Biology. Breast Neoplasms. RNA-seq.*

[1] Universidade Federal do Espírito Santo.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

RESUMEN
El cáncer/neoplasias de mama (CM) es una enfermedad altamente prevalente en las mujeres con millones de casos nuevos cada año. Entre los avances tecnológicos destaca la tecnología RNA-seq, que ha permitido comprender mejor la expresión génica, permitiendo desvelar interacciones proteicas entre tumores de mama tempranos y recurrentes (posmastectomía). Han surgido nuevas herramientas basadas en bioinformática para seguir el avance de la secuenciación, siendo los principales ejemplos las plataformas de análisis online Galaxy y WebGestalt. Además, se estableció el Archivo de lectura de secuencias (SRA) como un depósito público para datos de secuencias de próxima generación, al igual que el uso del depósito de datos genómicos funcionales Gene Expression Omnibus (GEO). En este trabajo, utilizando el análisis de secuenciación de ARN total, fue posible demostrar comparaciones generalizadas de CM en etapa temprana con CM recurrente. Además, se utilizaron Gene Ontology (GO), KEGG y Reactome para evaluar las relaciones funcionales y las vías mejoradas entre la CM en etapa temprana y la CM recurrente posmastectomía. En conclusión, gracias al desarrollo de este estudio fue posible descubrir nuevos biomarcadores que podrían ser utilizados como futuras dianas terapéuticas, permitiendo un mejor diagnóstico y pronóstico en CM con el objetivo de mejorar la supervivencia global de los pacientes.

PALABRAS CLAVE: *Perfilación de la Expresión Génica. Biología Computacional. Neoplasias de la Mama. RNA-seq.*

1 INTRODUÇÃO

A neoplasia maligna da mama é o segundo tipo de câncer mais incidente no mundo e o mais comum entre as mulheres, sendo registrados a cada ano, mais de 1,6 milhão de casos novos e pouco mais de 521 mil óbitos pela doença (Campêlo de Sousa, M; Campêlo de Sousa, C, 2020).

O câncer de mama (CM) é uma doença heterogênea que abrange uma ampla variedade de entidades patológicas e uma gama de comportamentos clínicos. Eles são sustentados no nível molecular por um completo conjunto de alterações genéticas que afetam os processos celulares (Simpson *et al*., 2005).

No CM, vários mecanismos prejudicam progressivamente o sistema de crescimento e morte celular. Entre esses mecanismos pode-se observar o acúmulo de mutações, instabilidades cromossômicas e alterações epigenéticas capazes de alterar a taxa de proliferação celular e danificar o sistema de reparo do DNA (Campêlo de Sousa, M; Campêlo de Sousa, C, 2020).

Diversas técnicas surgem no sentido de compreender como esses tumores interagem via expressão gênica. Dentre elas, vale destacar o RNA-seq (sequenciamento do RNA total), que é um sequenciamento de alto rendimento e que se tornou a principal opção para medir os níveis de expressão. Essa técnica pode ser realizada sem o conhecimento prévio da referência ou sequência de interesse e permite uma ampla variedade de aplicações (Costa-Silva; Domingues; Lopes, 2017). Além disso, a compreensão dos dados de RNA-seq depende da questão científica de interesse (Oshlack; Robinson; Young, 2010).

Devido a essa grande quantidade de dados circulante de sequenciamento de última geração, criou-se o *Sequence Read Archive* (SRA). O SRA foi estabelecido como um repositório público para os dados de sequência da próxima geração e é operado pela *International Nucleotide Sequence*

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Database Collaboration (INSDC) (Leinonen; Sugawara; Shumway, 2011). Logo, esse banco de dados permite um acesso a dados para estudos futuros relacionados com a técnica de RNA-seq.

Em geral, a tecnologia RNA-Seq é muito útil para análise de expressão diferencial envolvendo algumas condições específicas, sendo que, a relevância dos dados produzidos é finalmente avaliada a partir de um contexto biológico. Com a crescente popularidade da tecnologia RNA-Seq, muitos softwares e pipelines foram desenvolvidos para análise diferencial da expressão gênica a partir desses dados (Costa-Silva; Domingues; Lopes, 2017).

Dentre os *softwares* desenvolvidos, vale destacar, a Plataforma *Galaxy*. Iniciado em 2005, o Projeto *Galaxy* mantém o foco em possibilitar uma melhoria de estudos para a ciência biomédica orientada a dados, perseguindo três objetivos: (a) análise acessível de dados, servindo a todos os cientistas, independentemente de seus conhecimentos; (b) análises reproduzíveis, independentemente da plataforma específica e (c) comunicação transparente das análises, que permitirão entender as análises feitas para toda a comunidade acadêmica, principalmente na parte prática de ensino (Afgan; *et al*., 2018).

Após uma análise da expressão diferencial dos genes, diversos estudos destacam uma grande importância das pesquisas realizadas para uma compreensão da interação dos genes em redes por meio de um enriquecimento de vias. Dentre os programas mais comumente utilizados, destaca-se o *WebGestalt* (*Gene SeT AnaLysis Toolkit* baseado na *Web*), que é uma ferramenta popular para a interpretação de listas de genes derivadas de estudos de dinâmica em larga escala, logo, ajuda os usuários a extrair insights biológicos de genes de interesse (Liao *et al*., 2019).

Dessa maneira, compreende-se como um ponto importante a realização de uma correlação entre dados de sequenciamento de amostras de CM e a realização de uma análise de genes diferencialmente expressos, provindos do *Sequence Read Archive* (SRA) (disponíveis através de vários provedores de nuvem e servidores NCBI), utilizando-se da Plataforma *Galaxy*, *WebGestalt* e do *Gene Expression Omnibus* (GEO), visando determinar genes com potencial ação moduladora que possam ser utilizados como biomarcadores e futuros alvos terapêuticos. Como hipótese geradora, entende-se que há uma ligação entre genes diferencialmente expressos e sua respectiva atuação sobre pressões evolutivas para um estadiamento de pior prognóstico, servindo como uma base inicial para futuros estudos de associação entre genes diferencialmente expressos, filogenética tumoral e simulação por modelagem de redes regulatórias gênicas no câncer.

2 CÂNCER DE MAMA

O CM é o câncer mais comum e a segunda principal causa de morte relacionada ao câncer entre mulheres em todo o mundo. Logo, tem a maior incidência entre as doenças malignas do sexo feminino, e o prognóstico para essas pacientes permanece ruim (Tian *et al*., 2020). O CM é amplamente classificado em carcinoma ductal não invasivo in situ (DCIS) e carcinoma ductal invasivo (IDC).

Compreender o mecanismo da carcinogênese da mama em nível genético e transcricional

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

pode ajudar na caracterização de CDIS ou de tumores IDC em estágio inicial. As assinaturas de expressão gênica são usadas para classificar subtipos de câncer de mama (Perou *et al*., 2000; Sorlie *et al*., 2001; Rao *et al*., 2019).

A incidência de CM é influenciada por vários fatores que incluem idade, genética e história reprodutiva. Um entendimento da biologia do tecido normal e sua heterogeneidade inerente é um passo importante para dissecar os mecanismos que levam à oncogênese. O tecido mamário normal compreende um sistema ductal epitelial complexo embutido em uma matriz estromal que é composta de fibroblastos, adipócitos, células endoteliais e imunológicas. As mudanças dinâmicas que ocorrem no epitélio da mama durante a puberdade, gravidez e lactação são impulsionadas pela ação combinada de hormônios sistêmicos e fatores de crescimento. Ao longo da vida de uma mulher, a exposição sustentada aos hormônios esteróides ovarianos é um fator de risco bem estabelecido para CM, com uma correlação clara entre o número de ciclos menstruais e o risco de câncer de mama (Pal *et al*., 2021).

Ademais, o CM abarca um conjunto diversificado de doenças caracterizadas pela heterogeneidade que influenciam sobremaneira na resposta ao tratamento pelo paciente. Contudo, essa heterogeneidade dificilmente pode ser definida com precisão por meio dos parâmetros clássicos de histopatologia, grau do tumor e envolvimento nodal (Pal *et al*., 2021). Assim, mesmo o CM sendo atualmente tratado com cirurgia, radioterapia, quimioterapia citotóxica e/ou terapias direcionadas para erradicar células cancerosas viáveis, ainda apresenta dois grandes desafios sobre o seu tratamento, diante da resistência terapêutica e a formação de metástases em locais secundários (pulmão, osso, nódulos linfáticos, cérebro e fígado) levando inevitavelmente à mortalidade do paciente (Minn *et al*., 2005; Parsons; Francavilla, 2020).

E nesses casos de resistência e metástases em situações de recidiva, destaca-se um aspecto muito característico, sendo a uma grande heterogeneidade intertumoral e intratumoral dentro de cada um desses subtipos moleculares. Além disso, os tumores apresentam heterogeneidade intratumoral significativa gerada por mecanismos genéticos e epigenéticos, o que ocasiona um desenvolvimento hierárquico de células tumorais a partir das células-tronco cancerígenas precursoras (CSCs) e até mesmo por meio de células poliploides cancerígenas gigantes (*Polyploid Giant Cancer Cells*, PGCCs), que conduzem a tumorigênese e metástase. Essas CSCs e as PGCCs também contribuem para a resistência terapêutica por meio de vários mecanismos. E essa heterogeneidade em níveis celulares e transcriptômicos permane sendo um desafio para a pesquisa e terapia do CM (Wu *et al*., 2020).

Tal ponto explicitado, demonstra que, mesmo com as terapias mais agressivas e modernas, cerca de 15% dos pacientes podem desenvolver recorrências locorregionais (LRR) que, por sua vez, aumentam o risco de doença à distância e morte, ou seja, há a possibilidade de recorrência mesmo depois de uma ação mais forte como na pós- mastectomia. A população que permanece com risco aumentado de um LRR pode se beneficiar de uma terapia mais intensiva, como uma dose de reforço

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

de radiação aumentada, adição de um sensibilizador de radiação, um aumento no volume irradiado, ou terapia sistêmica direcionada adicional, porém resultantes tão benéficas podem ser originadas nessas aplicações como a geração de CSCs e/ou pelas PGCCs (Keene *et al*., 2018).

2.1 RNA-SEQ, bioinformática e câncer de mama

Métodos tradicionais para o tratamento ou detecção precoce do CM, tornam-se ineficazes para serem aplicados em pacientes com suas singularidades. Assim, a abordagem de novas tecnologias que possam especificar melhor a expressão de células tumorais em comparação com células normais, vem se tornando uma boa abordagem para descobrir como a expressão gênica diferencial pode influenciar uma maior resistência ou um aprimoramento do tumor ao ponto de ocasionar recidivas. E dentre diversas tecnologias, vale destacar a recente tecnologia de RNA-Seq que revolucionou, enormemente o caminho para a identificação e quantificação de transcritos, logo, suas respectivas expressões gênicas (Zhang *et al*., 2020).

Em somatório, usar da expressão gênica significa compreender melhor o mRNA (RNA mensageiro), que, por ter uma alta taxa de decomposição, sua composição momentânea reflete, ou pode ser considerada, um instantâneo da atividade gênica. Dessa forma, a análise do transcriptoma de uma célula em diferentes condições ou pontos de tempo pode, portanto, fornecer percepções valiosas nas diferenças em um nível molecular entre tecidos saudáveis ou doentes ou a resposta a estímulos externos como drogas ou estresse (Kloet *et al*., 2020).

Por definição, o RNA-seq é um método de sequenciamento de alto rendimento (HTS) que mede os transcritos de cDNA (DNA complementar). E os transcritos são mapeados para um gene/isoforma e sua abundância deve se correlacionar com a expressão. O RNA-seq é uma tecnologia relativamente nova que foi rapidamente adotada na pesquisa clínica, desenvolvendo a medicina de precisão, mas é necessário ainda uma otimização adicional com certa urgência. Ademais, o RNA-seq é mais frequentemente analisado para investigar os níveis de expressão de genes/transcritos entre duas ou mais condições (isto é, grupos de contraste) em uma análise de Expressão Gênica Diferencial (DGE). Isto porque, na pesquisa do câncer, o DGE tem sido essencial na avaliação da função biológica, patogênese e descoberta de biomarcadores (Stupnikov *et al*., 2021).

A análise de DGE permite identificar as diferenças entre o estado doente e saudável para auxiliar no entendimento da patologia das doenças, como no CM, por meio do conhecimento de genes que (1) são superexpressos em vez de subexpressos; (2) são superexpressos em várias doenças; e (3) são populares na literatura biomédica em geral. A DGE, vem sendo muito aplicada para alcançar a elucidação de prováveis biomarcadores candidatos, alvos terapêuticos e assinaturas de genes para diagnósticos. Assim, mesmo que mudanças específicas de expressão gênica nem sempre se traduzem em atividade biológica, esses dados podem ser agrupados para criar análises integradas, como construir o cenário-alvo de uma doença. E todo o processo pode ter o auxílio virtual, como por meio da Plataforma *Galaxy* (Rodriguez-Esteban; Jiang, 2017).

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Para as análises funcionais acerca da DGE, utiliza-se muito da ontologia genética (GO) por meio de outras plataformas como o *WebGestalt*, a qual é amplamente utilizada no campo da bioinformática para anotar genes e seus produtos gênicos em grande escala (CHEN; *et al*., 2020). O GO é atualmente a maior fonte de informações abrangente a nível mundial sobre a função de genes e produtos gênicos (proteínas e RNAs não codificantes). E são informações utilizáveis por humanos e máquinas, resultando em um papel crítico na análise computacional de dados ômicos e biomédicos. Ademais, existem um compilado de informações derivadas diretamente de experimentos, assim como dados inferidos por meio de homologia de sequência ou de outros métodos de inferência, no caso de organismos menos estudados (The Gene Ontology Consortium, 2021).

As novidades promovidas pelo GO, inovaram a aquisição de anotações funcionais de produtos gênicos, promoveram o progresso do desenvolvimento de drogas, análise de doenças e dentre muitas outras aplicações que vem surgindo. Caracterizando o GO, há uma divisão em três subcategorias: função molecular (*molecular function*, MF), processo biológico (*biological process*, BP) e componente celular (*cellular component*, CC). A MF descreve as atividades elementares de um produto gênico no nível molecular. A BP captura o início e o fim, pertinentes ao funcionamento das unidades vivas integradas. E o CC descreve as partes das células e seus ambientes extracelulares. Cada ontologia consiste em um conjunto de termos ontológicos (termos GO), que são organizados em uma hierarquia, ou em um grafo (rede) acíclico direcionado (DAG) (Zhao *et al*., 2020).

3 MÉTODO

A pesquisa foi realizada segundo uma abordagem quantitativa e qualitativa, com uma natureza básica e aplicada sobre a análise de banco de dados públicos de acesso gratuito, em seguida, realizou-se a avaliação, montagem e expressão diferencial de amostras de câncer de mama de estágio inicial por meio da Plataforma *Galaxy* e o enriquecimento dos dados por meio do *WebGestalt*. Em seguida, comparou-se com dados de expressão de amostra de câncer de mama recorrentes pós-mastectomia.

4 AMOSTRAS

As amostras referentes ao câncer de mama em estágio inicial foram coletadas do banco de dados SRA-NCBI, com um consentimento público, assim como demonstrado no número de identificação das amostras. As amostras do banco de tumor foram sequenciadas pela plataforma *Illumina*, utilizando o sistema *Illumina HiSeq* 2500, tipo *Paired-end*, para fins de transcriptômica. Os valores de expressão gênica referentes ao câncer de mama recorrentes pós-mastectomia foram coletadas do banco de dados de expressão gênica GEO, com um consentimento público, assim como demonstrado no número de identificação das amostras, visando comparar o resultado alcançado com todo o processo de análise de RNA-seq das amostras de câncer de mama em estágio inicial. As amostras do banco de tumor foram sequenciadas pela plataforma *Illumina*, utilizando o sistema *Illumina HiSeq* 2000, tipo *Paired-end*, para fins de transcriptômica. Tendo em vista a extensa quantidade de

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

dados e um limite de 15gb de espaço de análise na Plataforma *Galaxy*, foram incluídos dados de uma amostra e sua respectiva replicata de cada amostra analisada.

4.1 Plataforma *Galaxy*

Na Plataforma *Galaxy*, o conjunto de dados obtidos pelo SRA foram mapeados no genoma humano (hg38), com o arquivo de anotação da mesma versão, ambos do projeto GENCODE. Para o mapeamento, alinhamento e quantificação, foram utilizados programas/*plugins* da Plataforma *Galaxy*. No caso do alinhamento, foram utilizados os *softwares: FATSQC* (que visa fornecer uma maneira simples de verificar o controle de qualidade nos dados brutos da sequência provenientes de pipelines do RNA-seq, demonstrando, assim, uma rápida impressão se os dados têm algum problema que deve ser reconhecido antes de se fazer qualquer análise adicional) e *Trimmomatic* (que realiza uma variedade de tarefas de corte úteis para dados de extremidade emparelhada e extremidade simples da Illumina).

Para o mapeamento, foi feito o uso do software *HISAT2* de mapeamento o qual é um alinhador universal de RNA-seq ultrarrápido e possui uma enorme quantidade de opções para filtrar alinhamentos e configurar o formato exato de sua saída. Na quantificação, utilizou-se o software *featureCounts* que gerou a matriz de contagem ao utilizar um arquivo de alinhamento no formato SAM ou BAM, resultando em um arquivo no formato GFF, e calculou o número de mapeamentos de leituras para cada recurso, além de outros parâmetros padrão. As matrizes de contagem foram utilizadas como entrada para os métodos de expressão diferencial. Neste trabalho, foi utilizado o *software limma* para a identificação das transcrições diferencialmente expressas. Esse programa usa as tabelas de contagem geradas a partir de *featureCounts* como entrada. O *limma* foi capaz de lidar com vários fatores, como o que afeta as expressões gênicas.

4.2 *Webgestalt*

Para a análise de enriquecimento funcional de vias, utilizou-se a ferramenta *WebGestalt* que permitiu interpretar os resultados obtidos do *limma*, ou seja, os genes diferencialmente expressos com as suas interações em redes ou vias distintas, podendo assim compreender uma lista de genes ou proteínas interessantes para futuros trabalhos como alvos terapêuticos (figuras 1 e 2).

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

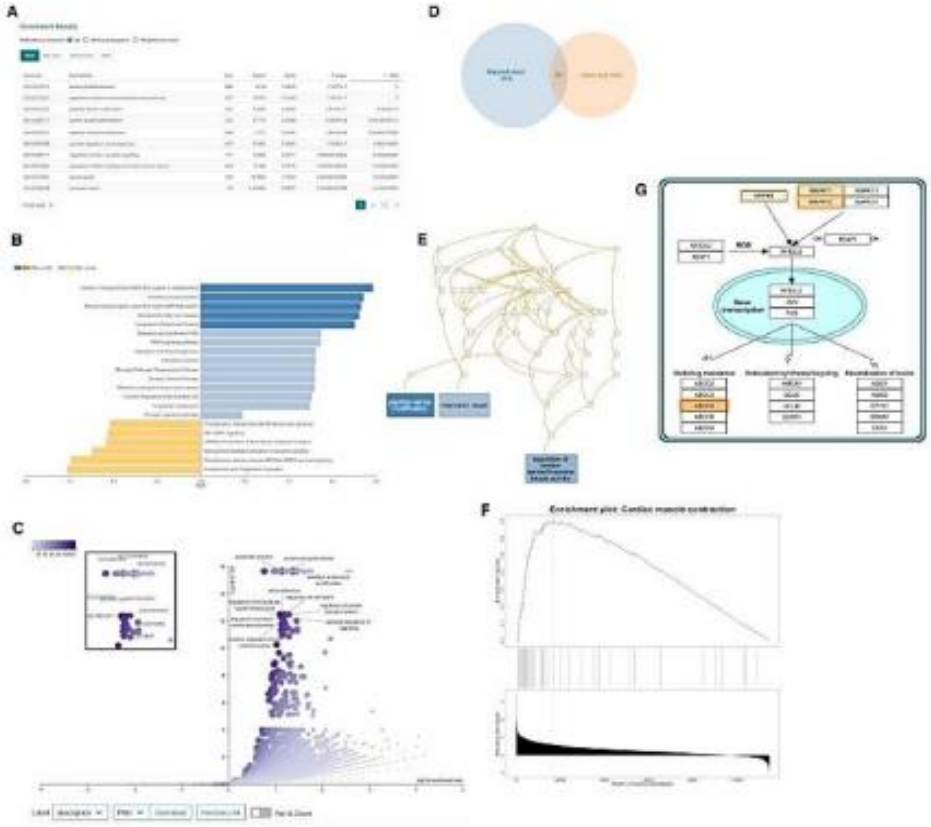

Figura 1. Exemplo de visualizações obtidas na página de resultados do WebGestalt. (A) Resumo da tabela de resultados significativos. (B) O gráfico de barras mostra a taxa de enriquecimento ou NES dos resultados com a direção. (C) Parcela do gráfico de vulcão personalizável. (D) DAG destacando nós enriquecidos. (E) O diagrama de Venn mostra a sobreposição entre o conjunto de genes na entrada e na referência. (F) Gráfico de enriquecimento da GSEA. (G) Visualização de caminho dos WikiPathways, destacando os genes de ponta com base na pontuação.
Fonte: Liao *et al*., 2019.

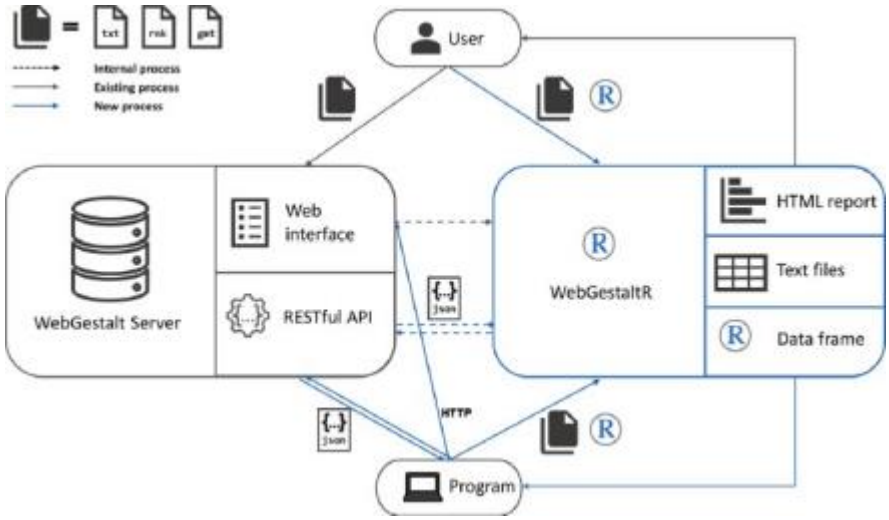

Figura 2. O sistema *WebGestalt* permite fácil acesso de usuários e programas.
Fonte: Liao *et al*., 2019.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

5 RESULTADOS

Neste estudo foram feitos e alcançados os resultados analisados na Plataforma *Galaxy* das amostras referenciadas na figura 3, sendo que as amostras com os códigos SRR7645075 e SRR7645076, representam as amostras iniciais do câncer de mama em estágio inicial, enquanto as amostras com códigos SRR7645077 e SRR7645078, representam uma replicata técnica, ou seja, analisou-se a mesma amostra biológica independentemente mais de uma vez. Por meio dessas amostras, separou-se o que eram amostras normais das tumorais e foram baixados seus sequenciamentos de RNA total na plataforma *Galaxy*, visando realizar os procedimentos abarcados na metodologia.

SampleID	Genotipos (Primary Factor)
SRR7645075	No
SRR7645076	Tu
SRR7645077	No
SRR7645078	Tu

Figura 3. Amostras Coletadas do SRA, com a respectiva identificação do repositório, com os respectivos genótipos associados a cada amostra, definindo como amostra No (Normal) e Tu (Tumoral).
Fonte: Produção do próprio autor pela Plataforma *Galaxy*.

Após todos os procedimentos na Plataforma *Galaxy*, foram filtrados 7974 genes do total de 28395 genes obtidos da análise de expressão gênica diferencial, e utilizou-se uma contagem por milhão (CPM) de 0.2 em, no mínimo, 2 amostras, ou seja, pelo menos 20% das CPMs em no mínimo 2 amostras foram filtrados como sendo insignificantes. Além disso, o método TMM foi utilizado para normalizar o tamanho das bibliotecas de *reads* (leituras) com os respectivos *contigs* (conjunto da maior sobreposição da maior quantidade de leituras) formados. Isto porque, o TMM é um método simples que fornece uma boa estimativa dos níveis relativos da produção de RNA a partir da análise de RNA-seq feita. Para além, utilizou-se também nos resultados o valor de Log2Fold Change ou logFC (mudança de dobra em duas vezes), que descreve, neste trabalho, o quanto a quantidade da expressão gênica muda entre uma amostra A e uma amostra B, sendo a razão das duas. Logo, alteração de 30 para 60, significa que houve uma alteração de 2 vezes, sendo um aumento de 2 vezes.

A seguir, a figura 4 representa a identificação, segundo o NCBI, das amostras (em azul) e o aumento ou diminuição da expressão gênica frente a comparação de “No” com “Tu”, intercalando entre um valor de Log2 Fold Change de -5 (subexpressão) a um valor de Log2 Fold Change de 5 (superexpressão).

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

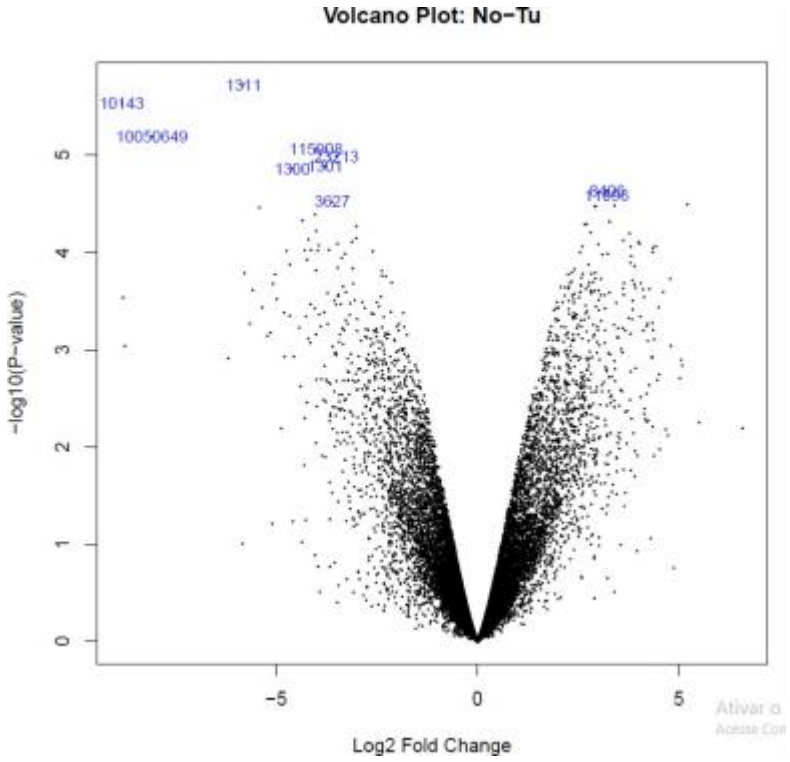

Figura 4. "Gráfico de Vulcão" ("Volcano Plot") representando o padrão de expressão (gênica) entre a razão de amostras normais (No) pelas (/) amostras tumorais (Tu).
Fonte: Produção do próprio autor pela Plataforma *Galaxy*.

Por conseguinte, foi obtido um MD plot (MA plot) que é um gráfico que está representando o logFC (log Fold Change) versus a expressão gênica média entre as amostras "No" para com as "Tu", sendo um gráfico de dispersão onde os valores extremos ao longo do eixo das ordenadas (eixo y) representa o nível de expressão gênica diferencial, sendo vermelho os superexpressos e, azul, os subexpressos, como evidencia a figura 5.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

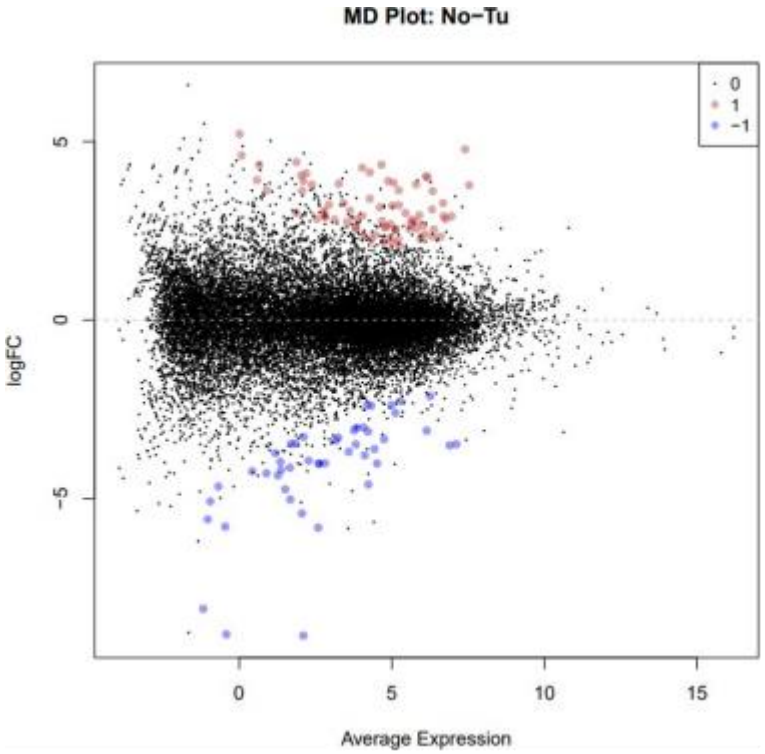

Figura 5. Gráfico de Diferença Média do contraste No e Tu, representando no eixo das ordenadas o logFC (log Fold Change) e no eixo das abscissas a expressão (gênica) média (average expression). Utilizou-se um FDR de 0.05 sobre a avaliação estatística do gráfico e seus dados respectivos.
Fonte: Produção do próprio autor pela Plataforma *Galaxy*.

Com base nos resultados da figura 4 e 5, obteve-se uma lista de genes identificados pela nomenclatura de identificação gênica do NCBI, que ao ser submetida no *Webgestalt*, obteve-se as estruturas de ontologia genética evidenciadas para a totalidade dos genes alcançados na pesquisa para as amostras de tumores de mama em estágio inicial, como demonstrado na figura 6. E, em seguida, após a utilização dos logFC e da identificação dos genes no *Webgestalt* para obter as análises de ontologia genética, aplicou-se os dados para verificar pelo KEGG e Reactome, prezando pela aquisição das informações acerca dos genes regulados “para cima” (*Upregulated*/superexpressos) e os genes regulados “para baixo” (*Downregulated*/subexpressos), como destaca a Tabela 1.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

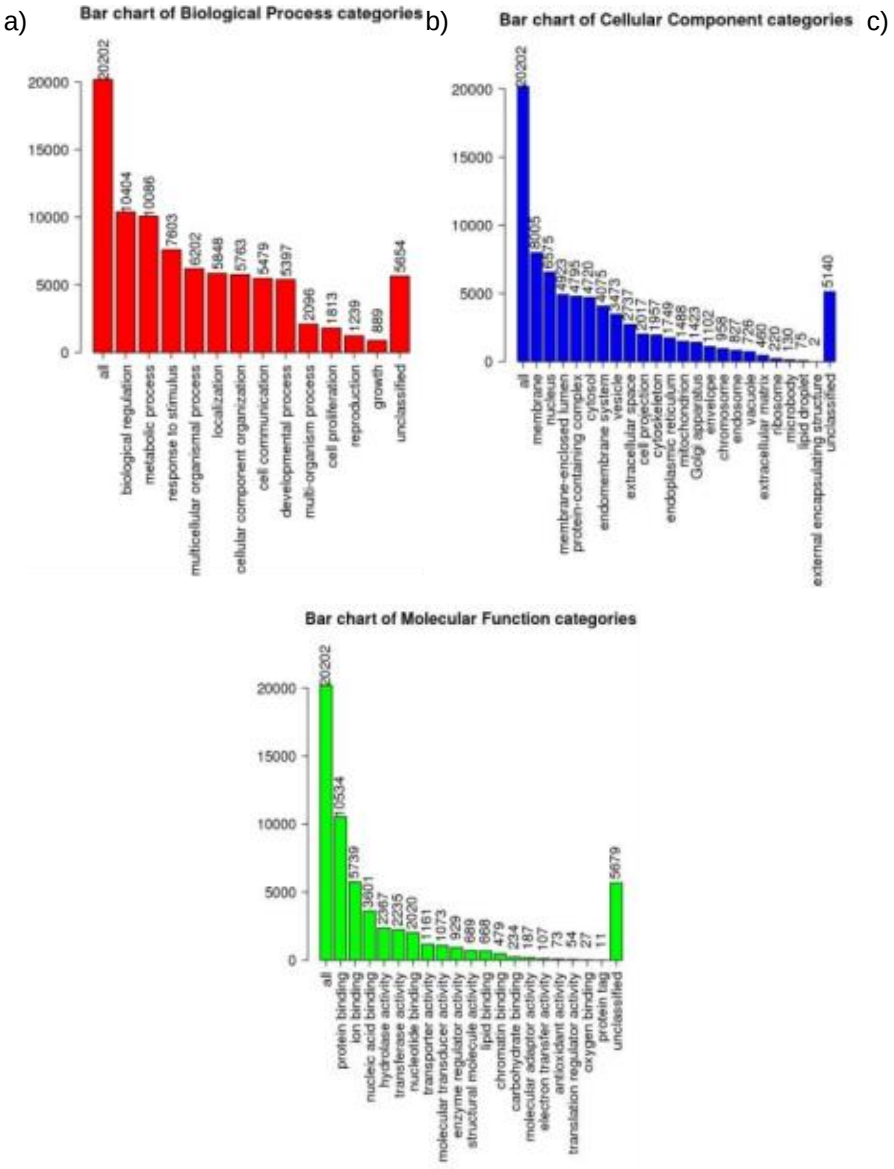

Figura 6. Gráfico de barras resumindo ontologia genética segundo Processo Biológico, em 'a', Componente Celular, em 'b', e Função Molecular, em 'c', frente ao percentual total de genes alcançados no RNA-Seq.
Fonte: Produção do próprio autor pelo *Webgestalt.*

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Descrição de GO, KEGG e Reactome	Expressão Gênica Diferencial	
	Upregulated (valor de P ajustado)	*Upregulated* (valor de P ajustado)
Processo Biológico	Morfogênese de Estrutura Anatômica (6.1 E-26)	Processo do Sistema Imune (1.1 E-31)
Componente Celular	Região Extracelular (2.8 E-18)	Região Extracelular (1.7 E-19)
Função Molecular	Ligação de Íons de Cálcio (2.0 E-08)	Atividade de Heterodimerização Proteica (7.1 E-06)
KEGG	Proteoglicanos no Câncer (5.1 E-04)	Lúpus Eritematoso Sistêmico (2.7 E-15)
Reactome	Contração Muscular (2.1 E-06)	HDACs desacetilase de histonas (3.8 E-14)

Tabela 1. Formatação dos dados do *Webgestalt* referente aos genes regulados “para cima” *(Upregulated)* e os genes regulados “para baixo” (c) provindos seus resultados da análise feita na Plataforma *Galaxy*.
Fonte: Produção do próprio autor segundo dados do *Webgestalt*

Ao final, realizou-se a análise dos genes alcançados no estudo de expressão gênica diferencial no estudo identificado no experimento com acesso pelo GEO (GSE119937) referente ao estudo sobre Câncer de Mama Recorrente Pós-mastectomia, de um estudo publicado por Keene *et al*. (2018), que publicou os dados no GEO para acesso gratuito. Assim, obteve-se uma descrição de GO, KEGG e *Reactome*, representando as cinco principais descrições de forma generalizada, para um FDR < 0.05, para processo biológico, componente celular, função molecular, KEGG e *Reactome*, como demonstra a Tabela 2.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Descrição de GO, KEGG e *Reactome*	Dados Obtidos no *Webgestalt*
Processo Biológico	a) Morfogênese de um epitélio ramificado; b) Cascata STAT; c) Resposta ao organofosforado; d) Desenvolvimento do fígado; e) Regulação da proliferação de células musculares lisas.
Componente Celular	a) Complexo de fator de transcrição nuclear; b) Periferia nuclear; c) complexo de polimerase de RNA dirigido por DNA; d) Grânulo de pigmento; e) Borda em escova.
Função Molecular	a) Atividade do transportador transmembranar de ácido orgânico; b) Atividade de canal dependente de ligante; c) Soluto: atividade simportadora de cátions; d) Atividade do transportador transmembrana do fármaco; e) Atividade do receptor de citocinas.
KEGG	a) Câncer gástrico; b) Via de sinalização do TNF; c) Doença de Chagas; d) Síntese e secreção de aldosterona; e) Via de sinalização de GnRH.
Reactome	a) Transporte em canal de íons; b) ER para Transporte Anterógrado de Golgi; c) PPARA ativa a expressão gênica; d) Hidrólise de GTP e junção da subunidade ribossomal 60S; e) Sinalização por NTRKs.

Tabela 2. Formatação dos dados segundo as cinco principais descrições do *Webgestalt* referente aos genes identificados no experimento com identificação de acesso pelo GEO (GSE119937) referente ao estudo sobre Câncer de Mama Recorrente Pós-mastectomia.
Fonte: Produção do próprio autor segundo dados do *Webgestalt*

6 DISCUSSÃO

Neste estudo, constata-se o grande impacto e benefícios frente à utilização de bancos de dados públicos, como o SRA e o GEO, as vantagens para com os trabalhos feitos em plataformas *online* como a Plataforma *Galaxy* e o *Webgestalt*, benefícios esses que perpassam pelo acesso gratuito a dados valorosos de sequenciamento de DNA e RNA total que são bem custosos e de difícil acesso para certos grupos de pesquisa iniciantes ou em uma situação de renda e de infraestrutura complexas. Além disso, com o uso de plataformas online, diminui o gasto de espaço no computador e de memória do próprio computador sobre a demanda de tratamento das amostras.

Desta forma, com a abordagem em web deste estudo, isto significa um modelo para trabalhos mais diversos para compreender inúmeras doenças, como no caso do câncer, assim é possível avançar no entendimento da expressão gênica diferencial e como ela atua na elucidação molecular e celular melhorada de um dado momento situacional comparativo entre estado saudável e de doença.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

Perante a tal ponto, há um aumento sobre a reprodutibilidade de dados em bioinformática e uma maior divulgação entre grupos de pesquisa em todas as áreas ao ponto de criar redes de colaborações e novas divulgações necessárias em bioinformática.

Frente aos resultados obtidos (figuras 3 a 5), destacam-se os genes que são superexpressos e subexpressos segundo os métodos seguidos na Plataforma *Galaxy*, destacam-se genes com suas respectivas expressões que passam a servir como uma fonte de dados atrelados ao entendimento do CM em estágio inicial versus células normais, possibilitando identificar o mecanismo de regulação da expressão gênica diferenciada, podendo fornecer possíveis biomarcadores prognósticos úteis, sendo abarcada tal ideia central, segundo os estudos de Keene *et al*. (2018), Zhang *et al*. (2020), Kloet *et al*. (2020), Stupnikov *et al*. (2021) e Rodriguez-Esteban & Jiang (2017), que, em síntese, demonstram o grande impacto do RNA-seq e da DGE para criar análises cada vez mais integradas e aplicadas à clínica de maneira a melhor entender e tratar o CM.

Por conseguinte, com os dados de expressão do CM em estágio inicial analisados de maneira mais generalizadas, assim como, com a quantidade de genes analisados, de forma geral também, referentes ao CM recorrente pós- mastectomia, foram alcançadas as avaliações de GO, do BP, CC e da MF, destacando-se no resultado da figura 6, somando-se com as vias metabólicas do KEGG e do Reactome, visando estabelecer uma ponte do entendimento funcional do aspecto biológico traduzido dos genes evidenciados de estudos focados em DGE, assim como feito por Chen *et al*. (2020). Isto porque, o GO, dividindo em BP, CC e MF, junto com o KEGG e o Reactome, há uma organização hierárquica baseada em uma análise estatística de FDR, que integra diversas atividades e vias metabólicas de um produto gênico a nível molecular que resulta em associações comparativas entre estados e momentos diferentes do CM como no estágio inicial, mas também nos casos recorrentes pós-mastectomia, e tal destaque de GO também é abarcado por Zhao *et al*. (2020).

Quando se analisa os dados obtidos pelo *Webgestalt* em uma generalização ampliada, segundo as tabelas 1 e 2, observa-se que o CM em estágio inicial tem um grande impacto de desregulação sobre a morfogênese, sistema imune, aspecto epigenéticos e das questões celulares atreladas aos passos básicos, cruciais para tumores de mama, poderem alterar o seu microambiente tanto a nível celular quanto molecular, ligando-se com a capacidade de se esconder do sistema imune e também poder alterar seu próprio material genético frente a esses novos microambientes, sendo pontos cruciais para iniciar, desenvolver e progredir a estágios mais avançados do CM.

Em paralelo, quando se observa essa desregulação segundo o CM recorrente pós-mastectomia, quando se destacam as cinco descrições principais do GO, KEGG e Reactome na tabela 2, pode-se perceber um avanço cada vez mais robusto sobre a capacidade de morfogênese mais ampliada frente a regiões de novos nichos tumorais e um avanço sobre a capacidade de sobrevivência obtendo-se características peculiares de células com um alto desenvolvimento como células-tronco ou PGCCs, uma diversidade de expressão proteica em novos locais celulares, logo evidencia também em consonância ao BP, um avanço da capacidade de resistência do CM recorrente pós-mastectomia e por

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

um destaque final, há uma observação do grande impacto sobre a sinalização mais robusta frente a fatores de crescimento, tirosina quinase, transporte celular mais “veloz” e característico, visando uma entrega rápida de vesículas para um novo fenótipo tumoral.

Contudo, mesmo com todas as possíveis generalizações supracitadas e sustentáveis por Keene *et al*. (2018) e RAO; *et al*. (2019), há uma necessidade de uma melhoria do estudo a nível de realização de um sequenciamento de RNA total que seja longitudinal, ou seja, seja feito ao decorrer de todo o processo de iniciação, progressão e metástase do CM em estágio inicial para um estágio de recidiva de um CM recorrente pós-mastectomia, logo, podendo inserir avaliações estatísticas que sustentam e esclarecem de uma forma integrada a avaliação molecular e celular do CM de um estágio até o outro de uma forma com constante avaliações em um sequenciamento mais profundo ao ponto de etapas avaliadas. Logo, solucionando os erros atrelados a uma comparação generalizada e com pouca evidência de aproximação dos sequenciamentos, mas mantendo uma avaliação estatística pontual e não constante e uniforme de um estudo único em contínua avaliação.

7 CONSIDERAÇÕES

Através deste estudo, utilizando dados do sequenciamento de RNA total e também da expressão gênica diferencial, foi possível demonstrar comparações generalizadas frente ao processo característico que o CM possui em um estágio inicial e depois associar com outro estudo com CM recorrente pós-mastectomia, que evidenciam, primeiramente em estágio inicial, características de GO, KEGG e Reactome que abrangem e perpassam as formas de melhorar a progressão, proliferação e morfogênese ampliada a diversos nichos tumorais com um microambiente mais diverso, e, em paralelo, destaca no estágio recorrente, um aprimoramento sobre a resistência, sobrevivência, morfogênese específica, microambiente com nichos mais pontuais.

Assim, mesmo sendo generalizado, pode-se sugerir possíveis candidatos que podem ser destacados como biomarcadores sugestivos para casos de CM recorrente pós-mastectomia que possa reduzir a possibilidade de uso de tratamentos tradicionais que venham a desenvolver células mais resistentes com características de células-tronco tumorais ou de PGCCs, assim como biomarcadores possíveis para demonstrar o estágio inicial do CM. Em conclusão, através do desenvolvimento deste estudo foi possível sugerir a descoberta frente a novos biomarcadores que poderão ser utilizados como futuros alvos terapêuticos, possibilitando um melhor diagnóstico e prognóstico no CM visando à melhoria da sobrevida global das pacientes.

REFERÊNCIAS

AFGAN, Enis et al. The *Galaxy* platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. **Nucleic Acids Research**, v. 46, n. W1, p. W537–W544, 2018.

CAMPÊLO DE SOUSA, Maisa; CAMPÊLO DE SOUSA, Camila. Diagnóstico de câncer de mama por exames genéticos: uma revisão de literatura (Diagnosis of breast cancer by genetic exams: a literature

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

review). **Brazilian Journal of health Review Braz. J. Hea. Rev**, Teresina & Codó, n. 2, p. 1786–1797, 2020.

CHEN, Jiarui et al. KEGG-expressed genes and pathways in triple negative breast cancer. **Medicine (Baltimore)**, v. 99, n. 18, e19986, 2020. Doi: 10.1097 / MD.0000000000019986. PMCID: PMC7440132. PMID: 32358373.

COSTA-SILVA, Juliana; DOMINGUES, Douglas; LOPES, Fabricio Martins. RNA-Seq differential expression analysis: An extended review and a software tool. **PLoS ONE**, New Jersey (EUA), 21 dec. 2017.

KEENE, Kimberly S. *et al*. Molecular determinants of post-mastectomy breast cancer recurrence. **NPJ Breast Cancer**, v. 4, n. 34, 2018. Doi: 10.1038 / s41523-018-0089-z. PMCID: PMC6185974. PMID: 30345349.

KLOET, Frans M. van der; *et al*. Increased comparability between RNA-Seq and microarray data by utilization of gene sets. **PLoS Comput Biol.**, v. 16, n. 9, e1008295, 2020. Doi: 10.1371 / journal.pcbi.1008295. PMCID: PMC7549825. PMID: 32997685.

LEINONEN, Rasko; SUGAWARA, Hideaki; SHUMWAY, Martin. The sequence read archive. **Nucleic Acids Research**, v. 39, n. 1, p. 3, 2011.

LIAO, Yuxing et al. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. **Nucleic acids research**, v. 47, n. W1, p. W199–W205, 2019.

OSHLACK, Alicia; ROBINSON, Mark; YOUNG, Matthew. From RNA-seq Reads to Differential. **Genome Biology**, Parkville, Australia, p. 10, 2010.

PAL, Bhupinder et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. **EMBO J.**, v. 40, n. 11, e107333, 2021. Doi: 10.15252 / embj.2020107333. PMCID: PMC8167363. PMID: 33950524.

PARSONS, Joseph; FRANCAVILLA, Chiara. 'Omics Approaches to Explore the Breast Cancer Landscape. **Front Cell Dev Biol.**, v. 7, n. 395, 2020. Doi: 10.3389 / fcell.2019.00395. PMCID: PMC6987401. PMID: 32039208.

RAO, Arunagiri Kuha Deva Magendhra; *et al*. Identification of lncRNAs associated with early-stage breast cancer and their prognostic implications. **Mol Oncol.**, v 13, n. 6, p. 1342–1355, 2019. Doi: 10.1002 / 1878- 0261.12489. PMCID: PMC6547626. PMID: 30959550.

RODRIGUEZ-ESTEBAN, Raul; JIANG, Xiaoyu. Differential gene expression in disease: a comparison between high-throughput studies and the literature. **BMC Medical Genomics**, v. 10, n. 59, 2017.

SIMPSON, Peter T. *et al*. Molecular evolution of breast cancer. **Journal of Pathology**, 2005.

STUPNIKOV, A. *et al*. Robustness of differential gene expression analysis of RNA-seq. **Comput Struct Biotechnol J.**, v. 19, p. 3470–3481, 2021. Doi: 10.1016/j.csbj.2021.05.040. PMCID: PMC8214188. PMID: 34188784.

THE GENE ONTOLOGY CONSORTIUM. The Gene Ontology resource: enriching a GOld mine. **Nucleic Acids Research**, v. 49, n. D1, p. D325-D334, 2021. https://doi.org/10.1093/nar/gkaa1113.

RECIMA21 - REVISTA CIENTÍFICA MULTIDISCIPLINAR
ISSN 2675-6218

TIAN, Zelin et al. Identification of Important Modules and Biomarkers in Breast Cancer Based on WGCNA. **Onco Targets Ther.**, v. 13, p. 6805–6817, 2020. Doi: 10.2147 / OTT.S258439. PMCID: PMC7367932. PMID: 32764968.

WU, Shaocheng et al. Cellular, transcriptomic and isoform heterogeneity of breast cancer cell line revealed by full-length single-cell RNA sequencing. **Comput Struct Biotechnol J.**, v. 18, p. 676–685, 2020. Doi: 10.1016 / j.csbj.2020.03.005. PMCID: PMC7114460. PMID: 32257051.

ZHANG, Fan et al. Identification of novel alternative splicing biomarkers for breast cancer with LC/MS/MS and RNA-Seq. **BMC Bioinformatics**, v. 21, n. 541, 2020. Doi: 10.1186 / s12859-020-03824-8. PMCID: PMC7713335. PMID: 33272210.

ZHAO, Yingwen *et al*. A Literature Review of Gene Function Prediction by Modeling Gene Ontology. **Front Genet.**, v. 11, n. 400, 2020. Doi: 10.3389 / fgene.2020.00400. PMCID: PMC7193026. PMID: 32391061.