Revista UNILUS Ensino e Pesquisa
v. 20, n. 58, jan./mar. 2023
ISSN 2318-2083 (eletrônico)

NIKAEL SOUZA DE OLIVEIRA
Universidade de Caxias do Sul, UCS, Caxias do Sul, RS, Brasil.

NATHALIA RECH
Universidade de Caxias do Sul, UCS, Caxias do Sul, RS, Brasil.

SCHEILA DE AVILA E SILVA
Universidade de Caxias do Sul, UCS, Caxias do Sul, RS, Brasil.

Recebido em março de 2023.
Aprovado em junho de 2023.

ANÁLISE DE GENES DIFERENCIALMENTE EXPRESSOS: UMA ABORDAGEM METODOLÓGICA IN SILICO

RESUMO

Tecidos normais e tumorais apresentam o mesmo genótipo, entretanto diferentes fenótipos. Assim, espera-se encontrar diferentes padrões de expressão quando estes são comparados. Devido ao grande volume de dados gerados na obtenção destes dados e o desafio da união do conhecimento molecular e computacional, bancos de dados públicos têm sido criados para que novas pesquisas sejam realizadas. Desta forma, este trabalho demonstrou uma abordagem metodológica na identificação de genes diferencialmente expressos (DEGs), utilizando repositórios públicos com o auxílio do ambiente R. Para isto, a metodologia se dividiu em quatro etapas: (i) Obtenção dos dados; (ii) Preparação dos dados; (iii) Análise estatística e (iv) Análise dos resultados. Sendo identificados 12 DEGs em câncer de tireoide. Esperamos que este trabalho sirva de guia a projetos futuros na identificação de potenciais DEGs.

Palavras-Chave: câncer. bancos de dados. repositórios públicos. genes diferencialmente expressos. abordagem metodológica.

DIFFERENTIALLY EXPRESSED GENES ANALYSIS: AN IN SILICO METHODOLOGICAL APPROACH

ABSTRACT

Normal and tumor tissues present the same genotype, however different phenotypes. In this way, it's expected to find different expression patterns when those comparisons. Due to the large volume of data generated in obtaining these data and the challenge of uniting molecular and computational knowledge, public databases have been created so that new research can be carried out. In this way, this work demonstrated a methodological approach in the identification of differentially expressed genes (DEGs), using public repositories with the R environment. For this, the methodology was divided into four stages: (i) Obtaining data; (ii) Data preparation; (iii) Statistical analysis and (iv) Analysis of results. Being identified 12 DEGs in thyroid cancer. We hope that this work will serve as a guide for future projects in identifying potential DEGs.

Keywords: cancer. data bases. public repositories. differentially expressed genes. methodological approach.

Revista UNILUS Ensino e Pesquisa
Rua Dr. Armando de Salles Oliveira, 150
Boqueirão - Santos - São Paulo
11050-071
http://revista.lusiada.br/index.php/ruep
revista.unilus@lusiada.br
Fone: +55 (13) 3202-4100

INTRODUÇÃO

Câncer é um grupo de patologias de origem genética, ocasionada devido alterações no funcionamento de alguns genes presentes no organismo, os quais modificam o metabolismo celular normal, causando uma fuga dos padrões originais. Naturalmente as células se dividem, cumprem suas funções e morrem. Entretanto, devido a mutações que ocorrem durante os processos de divisão celular, influenciados também por fatores ambientais, os ciclos normais se alteram e uma única célula pode gerar uma massa que se multiplica descontroladamente, invade tecidos vizinhos e causa danos ao organismo, tornando-se, portanto, a patologia conhecida como câncer (BRUCE, et al 2017).

Todas as células de um mesmo organismo apresentam o mesmo genótipo, entretanto diferentes fenótipos, isso ocorre devido a mecanismos de regulação gênica os quais alteram os genes expressos em cada conjunto celular, formando os diferentes tecidos. Nesse sentido, se entende o câncer como uma patologia de origem genética, a qual foge dos parâmetros celulares normais, e, portanto, apresentam fenótipos distintos do seu tecido de origem (JORDE; CAREY; BAMSHAD, 2017). Desta forma, trabalhos tem demonstrado como diferentes padrões de expressões são encontrados nestes conjuntos, podendo ser estes utilizados como biomarcadores de atividades patológicas, exemplifica-se TULP3 no câncer colorretal (SARTOR; RECAMONDE-MENDONZA; ASHTON-PROLLA, 2019), AGT, SERPINH1 e MMP7 no câncer gástrico (LIU et al, 2021), CDC20 e ASPM no câncer de bexiga (XU et al, 2019).

Há dois grupos principais de genes relacionados ao câncer: (i) oncogenes e (ii) genes supressores de tumor. O primeiro, antes de se tornarem oncogenes propriamente ditos, são denominados de proto-oncogenes, e atuam nas funções normais das células, como divisão, metabolismo, regulação do ciclo celular e senescência. O segundo grupo compreende genes que atuam na regulação do ciclo celular impedindo a formação de tumores, atuando em etapas de controle e verificação, a qual como exemplo induzem a apoptose células mutadas com potencial de divisão descontrolada, entretanto, estes genes quando mutados perdem sua ação, e com as etapas de controle de qualidade não funcionando corretamente, abre-se espaço para a formação de neoplasias (PIERCE, 2017).

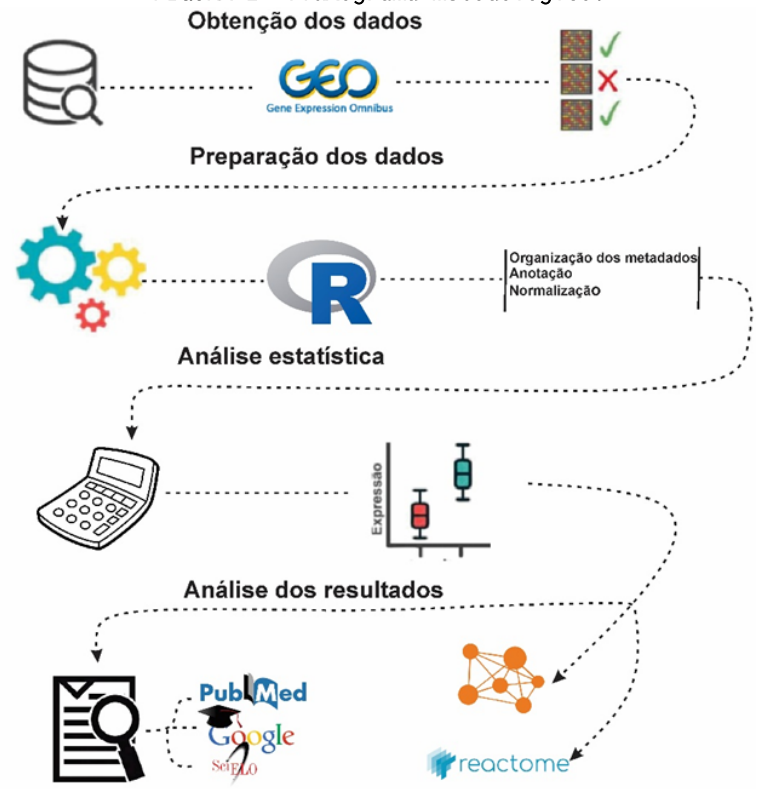
Com o advento das novas tecnologias moleculares, bem como o desenvolvimento das tecnologias computacionais, novas áreas de pesquisa vêm sendo desenvolvidas, e unindo estes dois tipos de conhecimento forma-se a bioinformática. Uma forma de estudo nesta jovem área é com análise de expressão de genes. Desta forma, visando pesquisas globais relacionadas a expressão gênica com intuitos de descobertas que possam auxiliar nos entendimentos destas patologias, bancos de dados têm sido criados permitindo o depósito de dados de expressão de genes. Como exemplo, cita-se: Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) (EDGAR; DOMRACHEV; LASH, 2002) e The cancer genoma atlas (https://portal.gdc.cancer.gov/) (Cancer Genome Atlas Research Network, 2014). Em relação a esses repositórios, o primeiro é voltado a dados de expressão obtidos a partir de análises de microarray e o segundo a partir de dados de RNA-seq.

O grande volume de dados gerados a partir das técnicas moleculares que identificam a expressão gênica, o formato dos dados e análises estatísticas adequadas requerem a utilização de ferramentas computacionais de análises, sendo essencial a integração entre o conhecimento molecular e o conhecimento computacional. Está é uma dificuldade encontrada por diversos pesquisadores e programas específicos têm sido desenvolvidos para facilitar estas análises como GEOR2 e GEPIA (TANG et al, 2017). Ainda assim, devido a manipulação para análises específicas são utilizadas linguagens de programação como R (R Development Core Team, 2021) e python (ROSSU; DRAKE, 1995). Nesse sentido, o presente trabalho possui como objetivo demonstrar um passo a passo para realização da coleta de dados e análise de expressão diferencial de genes em câncer utilizando o banco de dados GEO, bem como sua manipulação de dados no ambiente R.

PROCEDIMENTOS PARA ANÁLISE IN SILICO DE EXPRESSÃO DIFERENCIAL DE GENES

Para realização da análise de genes diferencialmente expressos, este trabalho irá se dividir em 4 etapas: (i) Obtenção dos dados pelos bancos de dados GEO; (ii) Manipulação e preparação dos dados; (iii) Análise estatística para obtenção dos genes diferencialmente expressos; (iv) Analise dos resultados obtidos, identificando os genes, suas funções e interações, bem como relatos na literatura associados a neoplasias. As etapas II e III são realizadas com auxílio do ambiente R. A figura 1 demonstra um fluxograma metodológico visualmente.

FIGURA 1- fluxograma metodológico.



Fonte: Elaborado pelos autores.

OBTENÇÃO DOS DADOS

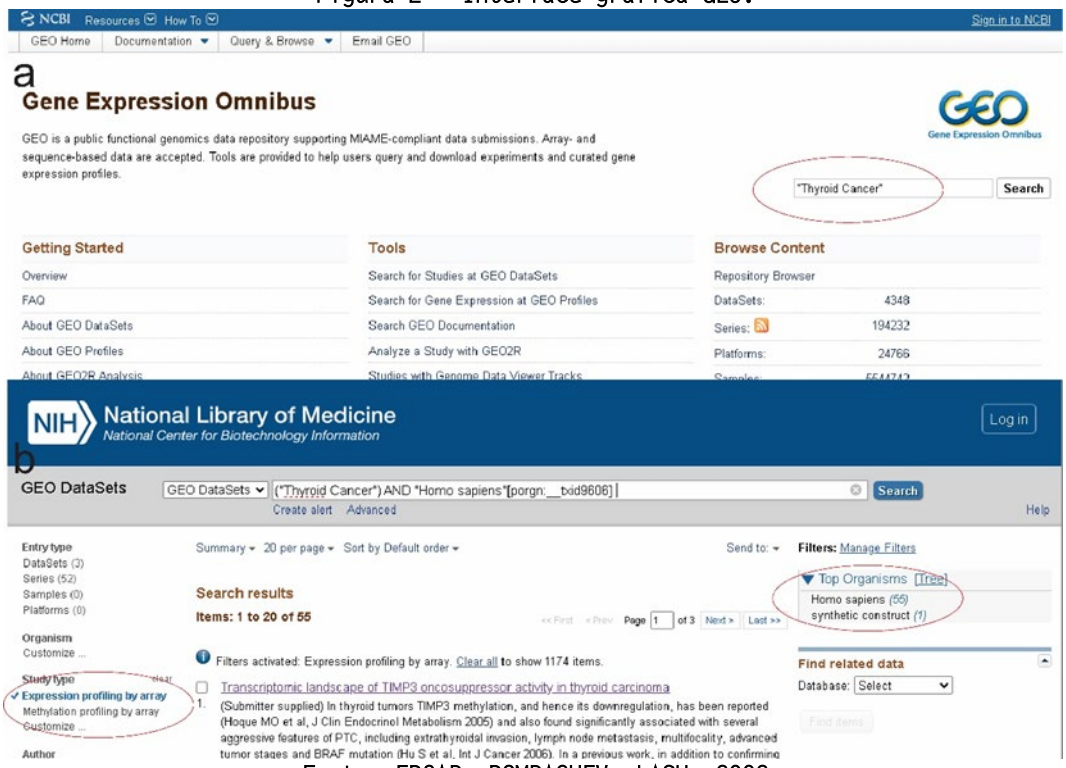
O GEO é um repositório público internacional que armazena e disponibiliza dados de microarray. Este possui três principais objetivos; (i) fornecer um banco de dados robusto e versátil para armazenar dados de alto rendimento; (ii) oferecer procedimentos de apresentação simples e formatos que suportam depósitos de dados completos da comunidade de pesquisa; (iii) fornecer mecanismos de fácil utilização que permita que o usuário pesquise, localize, revise e faça download de estudos e perfis de expressões dos genes (EDGAR; DOMRACHEV; LASH, 2002).

O tipo de câncer escolhido para utilização neste trabalho é o câncer de tireoide. Sendo necessária esta definição para demonstrar a forma de busca no banco. A pesquisa no banco de dados GEO se dá pela interface gráfica do website, utilizando de uma string de busca, a qual para a presente demonstração será: “Thyroid Cancer”.

O banco de dados GEO pode ser encontrado pelo link https://www.ncbi.nlm.nih.gov/geo/ e a string de busca deve ser digitada no campo “search” entre aspas (figura 2a). Em caso de busca de mais de um tipo de câncer este pode ser realizado em uma mesma pesquisa, separados por seus diferentes conjuntos de aspas e com a palavra OR no meio. Como o GEO apresenta uma vasta coleção de dados de

diferentes espécies e diferentes tipos de análises filtros devem ser selecionados na plataforma. No trabalho em questão estes serão tipo de organismo: Homo sapiens e tipo de estudo: Expression profillin by array (figura 2b).

Figura 2 - Interface gráfica GEO.



Fonte: EDGAR; DOMRACHEV; LASH, 2002

A busca apresenta diversas amostras depositadas, e estas devem ser analisadas individualmente e selecionadas para análise, para isto deve-se definir critérios de inclusão e exclusão, os quais devem ser baseados no objetivo do trabalho do pesquisador. Cada conjunto de dados apresenta um código de acesso determinado por GSE seguido por um numeral, estes devem ser anotados para posterior utilização. Ao mesmo tempo, cada GSE apresenta um GPL também seguido por um numeral, a qual é a plataforma utilizada na realização do experimento de microarray, e contém a marca e o modelo do chip. Este também deve ser anotado, pois será necessário durante o processo de obtenção dos dados utilizando o ambiente R.

Visando a obtenção de genes diferencialmente expressos in vivo que possam ser utilizados como potenciais biomarcadores os critérios de inclusão e exclusão para cada conjunto de dados devem ser definidos. Como exemplo, os critérios utilizados para este trabalho foram: (i) Cada conjunto de amostra deve possuir amostras normais e tumorais; (ii) Cada conjunto de amostras deve conter apenas amostras de tecidos biopsiados; (iii) Cada conjunto de amostras deve conter no mínimo 10 amostras. Após a seleção dos conjuntos de dados com seus respectivos GSEs anotados deve-se realizar o download das amostras com os dados de expressão e os respectivos metadados, utilizando como auxílio o pacote GEOquery (DAVIS; MELTZER, 2007), no ambiente R (Material suplementar 1).

PREPARAÇÃO DOS DADOS

Os dados baixados devem ser preparados utilizando o ambiente R (neste artigo foi utilizada a versão 4.1.1). Cada gene é representado por uma sonda, sendo um dos processos a anotação gênica a partir desta sonda. Os conjuntos de dados de expressão são organizados em arquivos diferentes do seu conjunto de metadados com as informações clínicas dos pacientes. Este conjunto de metadados carregam informações importantes, as

quais podem ser utilizadas com bases em novas pesquisas relacionadas aos aspectos clínicos dos pacientes. Quando estes apresentam-se nos objetivos da pesquisa e serão utilizados, faz-se importante a associação das informações dos diferentes arquivos. Os quais podem ser assimilados, utilizando o ambiente R, em um mesmo dado para realização de análises posteriores.

Para realização da análise estatística deve-se realizar a normalização dos dados (material suplementar 2), a qual é uma prática de organização e transformação dos valores de modo que estes possam ser comparados. Existem diversos tipos de normalização de dados, não havendo um consenso claro de qual o melhor método a ser utilizado. A transformação logarítmica auxilia na modelação do erro visto que deste modo apresenta alterações proporcionais e não aditivas, da mesma forma auxilia nas comparações de pequenas alterações. Ao mesmo tempo, pode-se pensar na amplificação do cDNA, utilizado nas técnicas moleculares de análise, uma vez que há resultados exponenciais sem seguir o padrão de distribuição normal, e a transformação logarítmica na base 2 é a que mais a aproxima (AMBROISE et al, 2011). Além disso, a transformação em LOG2 é amplamente citado para trabalhos com microarrray (SINHA et al, 2022; STEKEL, 2003; VALIZADEH et al, 2022). O pacote limma no ambiente R possui ferramentas que auxiliam neste processo (RITCHIE et al, 2015).

ANÁLISE ESTATÍSTICA

As diversas metodologias de comparação de médias são amplamente discutidas referente as análises relacionadas a comparação de dados de expressão gênica, não havendo um consenso claro de qual a melhor abordagem analítica. Entende-se que o fator de probabilidade associado, presente na estatística bayesiana se adapta melhor a incerteza dos dados biológicos e permite responder de forma mais apropriada as questões biológicas (JIMENEZ et al, 2021).

Para comparação entre amostras tumorais e normais é recomendada a estatística bayesina com correção da falsa taxa de descoberta de Benjamini e Hocberg para o valor p. (SINHA et al, 2023). Os pontos de cortes definidos para este trabalho são valor $p<0,05$ e Log2 Fold Change $>|1|$. Sendo o valor p a significância estatística, e log2 Fold Change o quanto a expressão do gene em câncer varia da normal, sendo estes valores negativos para os genes menos expressos e os valores positivos genes mais expressos. A definição dos pontos de corte cabe ao pesquisador, sendo que para o Log2 Fold Change varia entre 1 e 2 na maior parte dos trabalhos encontrados na literatura (SINHA et al, 2023; SONG et al, 2023; VALIZADEH et al, 2022; LIU et al 2023; ALAM; SULTANA et al, 2022). Definiu-se neste trabalho o valor de 1 para obtenção de um maior número de resultados (material suplementar 3).

ANÁLISE DOS RESULTADOS

Os diferentes GSEs analisados resultam em diferentes conjuntos de genes diferencialmente expressos. Deste modo deve ser realizado a intersecção dos conjuntos de resultados utilizando o diagrama de venn (Material suplementar 4). A análise dos resultados se dá por uma vasta pesquisa na literatura, utilizando-se de bancos de dados de artigos como PubMed (https://pubmed.ncbi.nlm.nih.gov/), e bancos de dados de exploração gênica como UNIPROT (https://www.uniprot.org/) ou CARDGENE (https://www.genecards.org/).

A partir disso, pode-se inferir a participação destes genes em processos biológicos normais, suas alterações relacionadas a atividades neoplásicas, bem como comparar se os resultados encontrados em um determinado tipo de câncer também já foram descritos em outros tecidos. Vale ressaltar que alguns destes genes diferencialmente expressos também são utilizados como alvos terapêuticos em determinadas neoplasias, podendo ser assimilados este tratamento já em uso a um novo tipo de câncer.

Uma forma de compreender o contexto metabólico destes genes é relacionando os processos biológicos e as vias metabólicas que estes atuam, sendo uma ferramenta disponível para análise a plataforma Reactome (https://reactome.org/). Esta é uma ferramenta de código aberto, curada manualmente e revisada por pares, a qual tem como objetivo fornecer caminhos em bioinformática para a visualização de dados, interpretação e análise de conhecimento. A plataforma permite a busca por lista de genes e indica os processos biológicos que estes participam, bem como demonstram em que etapa das rotas eles influenciam (GRISS et al, 2020).

A busca pode ser realizada na interface gráfica do website, selecionando o item ferramentas e após o item Analyse gene list (figura 3a). Em seguida deve-se realizar a busca utilizando o nome dos genes, sendo demonstrados os resultados referentes as vias metabólicas e os processos biológicos em que estes genes estão envolvidos (figura 3b). Os resultados obtidos a partir desta plataforma é dado em forma de interface gráfica demonstrando os processos biológicos relacionados, e em forma de relatório final, onde apresenta uma breve descrição literária do envolvimento dos genes pesquisados nas atividades celulares.

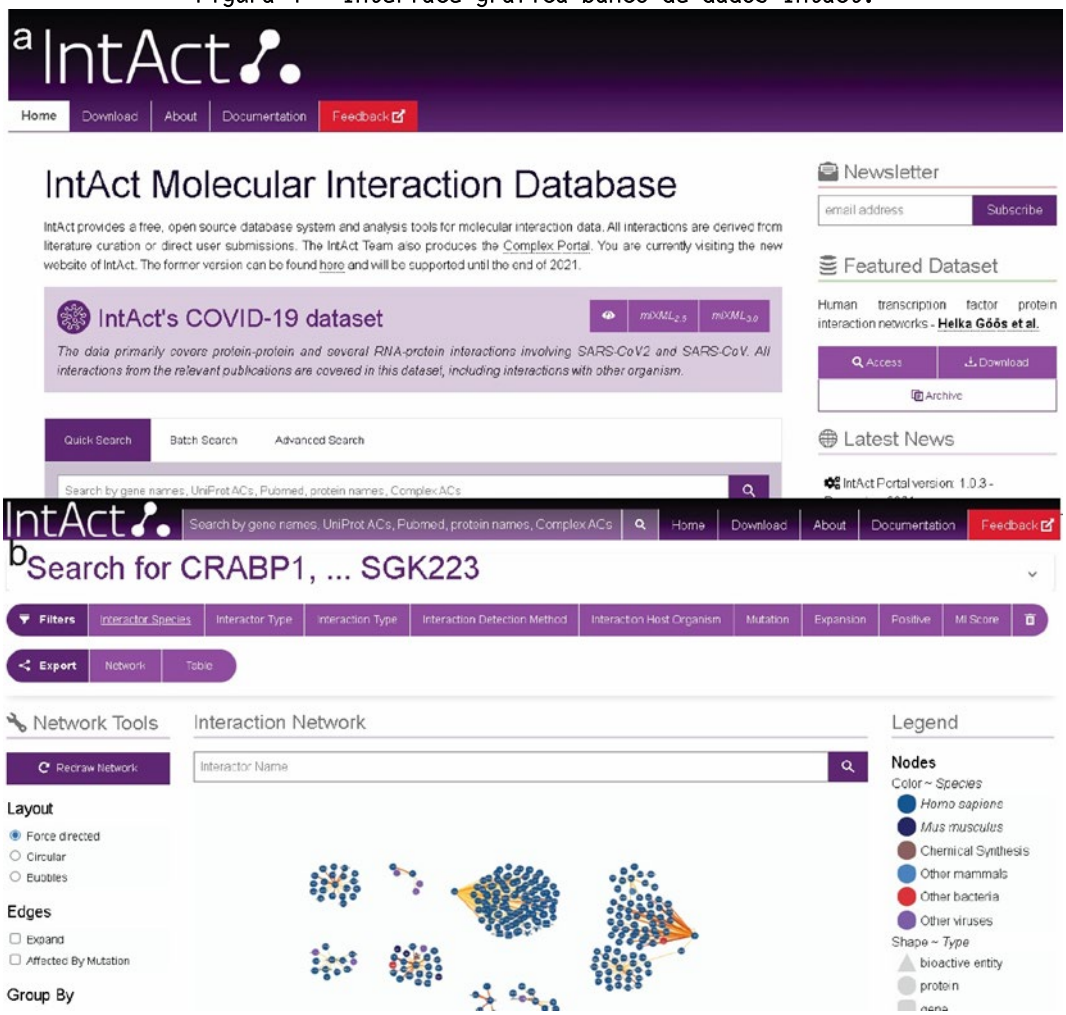
Figura 3 - Interface gráfica do banco de dados Reactome.

Fonte: GRISS et al, 2020.

Outra forma de verificar os conceitos biológicos dos genes identificados é pela construção de redes de interação proteína-proteína. Uma análise importante deste tipo de rede está relacionada a lista de genes, por vezes os genes diferencialmente expressos atuam em processos biológicos semelhantes, e interagem com os mesmos genes, os quais nem sempre são diferencialmente expressos, entretanto tem papéis importantes no câncer. Pode¬-se entender, de forma a exemplificar, a mutação de um gene de fator transcricional, o qual não apresentará alteração na expressão por si só, mas causará a diferença de expressão em algum outro gene, ou em um conjunto de genes. Sendo importante a identificação neste contexto de interação.

Desta forma, pode-se inferir importantes interações biológicas. Uma ferramenta com esta finalidade é IntAct (https://www.ebi.ac.uk/intact/home), que permite a busca na interface gráfica por lista de genes (figura 4a), retornando suas respectivas interações (figura4b) (ORCHAD et al, 2013). IntAct é um banco de dados público, o qual apresenta um amplo repositório relacionado a interações moleculares. Possui como principal objetivo aproveitar ao máximo os dados de interação pública e facilitar a adoção de padrões em biologia molecular. Além da interface gráfica, intAct pode ser trabalhado como extensão dentro da ferramenta cytoescape, permitindo a visualização de redes maiores e com maio nível de complexidade.

Figura 4 - Interface gráfica banco de dados Intact.



Fonte: ORCHAD et al, 2013

Vale ressaltar que existe uma ampla gama de ferramentas e bancos de dados disponíveis na internet e descritas na literatura. Cabe ao pesquisador realizar a busca pelas plataformas que lhe interessam e atendem seus objetivos. Aqui, foi demonstrado apenas alguns exemplos de ferramentas que podem auxiliar na pesquisa e na compreensão do aspecto biológico dos genes identificados como diferencialmente expressos. Ao mesmo tempo, ressalta-se a importância da busca de trabalhos relacionados já descritos em atividades neoplásicas.

ESTUDO DE CASO USANDO COMO EXEMPLO O CÂNCER DE TIREOIDE

Ao realizar a busca no banco de dados GEO, a partir da string “Thyroid Cancer”, se obteu como resultado preliminar 55 conjuntos de dados. Estes foram submetidos a

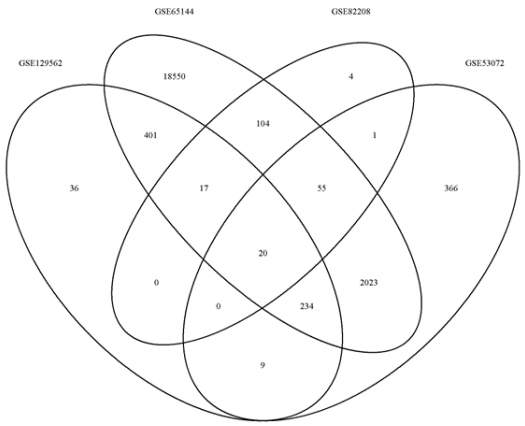
investigação manual e aplicação dos critérios de exclusão e inclusão, retornando em 4 conjuntos de dados, os quais são identificados pelo número de acesso: GSE82208, GSE65144, GSE53072 e GSE129562. Cada GSE foi submetido individualmente as etapas de obtenção dos dados no ambiente R, preparação dos dados e análise estatística, processos descritos na seção anterior. A partir disso, para cada GSE obteve-se diferentes conjuntos de genes diferencialmente expressos, demonstrados na tabela 1.

Tabela 1 - Número de genes diferencialmente expressos encontrado em cada conjunto de dados.

IDENTIFICAÇÃO	DEGS
GSE82208	257
GSE65144	6922
GSE53072	3206
GSE129562	884

Nota-se diferença no número de genes diferencialmente expressos encontrados nos diferentes conjuntos de dados. Esse resultado é esperado, visto que os dados de expressão são elaborados por diferentes pesquisadores em diferentes laboratórios com diferentes chips de microarray. Ao mesmo tempo, entende-se o organismo biológico em um estado dinâmico sobre as diferentes condições ambientais, o que pode alterar aspectos fenotípicos de expressão por si só. Desta forma, demonstra-se a necessidade dos cruzamentos dos dados para identificação dos DEGs presentes em todos os conjuntos. Para isso, com auxílio do R, foi realizado a intersecção dos resultados obtidos para cada GSE em um diagrama de venn (Figura 5). A intersecção dos 4 estudos resultou em 20 genes: CRABP1, GLT8D2, TNFRSF11B, FOS, SLC25A15, MT1F, LRP1B, DNALI1, WSCD2, SDPR, CSGALNACT1, CPQ, SDC2, PAX8, ITM2A, SELENBP1, SGK223, IER2, ID1, ID3.

Figura 5 - Diagrama de venn.



Fonte: Elaborado pelos autores.

Visto que o presente trabalho tem como objetivo uma abordagem metodológica referente a expressão diferencial de genes utilizando bancos de dados públicos, optou-se por 2 genes aleatórios, para a realização de uma breve discussão: ID3 e LRP1B. Ressalta-se que ambos os genes apresentaram menor valor de expressão no tecido neoplásico quando comparado ao tecido normal.

ID3 é um inibidor de proteína de ligação ao DNA. A qual atua como regulador transcricional que regula negativamente os fatores de transcrição básicos hélice-alfa-hélice (HLH), formando heterodímeros que inibem sua ligação ao DNA e atividade transcricional, implicando em processos celulares como crescimento, senescência, diferenciação e apoptose (DEED e t al, 1993). Sendo já relatado na literatura sua menor

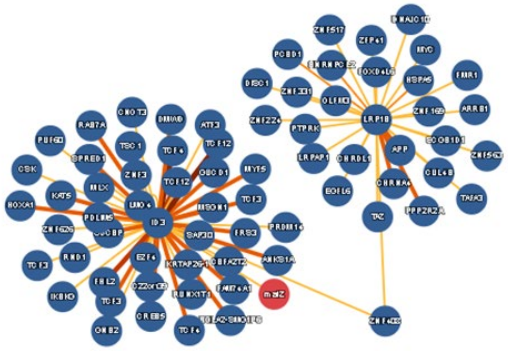
expressão em leucemia mieloide aguda (ZHAO et al, 2022) e relacionado a um circuito de regulação transcricional sobre a proteína dedo de zinco ZNF148 e ao proto-oncogene MYC em cânceres de mama agressivos (KIM et al, 2022). A busca no banco de dados reatome demonstrou ID3 envolvido em processos nucleares de transcrição e sinalização.

LRP1B é receptor de lipoproteína de baixa densidade localizado na membrana celular. Desempenhando diversos papeis na função e desenvolvimento celular normal, como transporte e sinalização celular, sendo considerado também um suposto supressor tumoral (PRINCIPE et al, 2021). Trabalhos relacionaram a frequente mutação deste gene associada a câncer gástrico (KUNG et al, 2023) e adenocarcinoma de pulmão (RAO et al, 2023). Além disso, da mesma forma que este trabalho, sua menor expressão já havia sido relatada em câncer gástrico (MIAO et al, 2022) e câncer de bexiga (CONCONI et al, 2022), demonstrando seu potencial como biomarcador ao diagnóstico desta patologia.

A busca no bando de dados Reactome resultou na identificação de processos biológicos relacionados a ID3, entretanto não foram encontrados resultados relacionados ao LRP1B. ID3 foi descrito como indutor transcricional de fator de crescimento nervoso, relacionado a eventos nucleares como quinase e ativador de fatores de transcrição, sinalização a partir de neurotrofinas tirosina quinase e sinalização por receptores de tirosina quinase. De modo geral, ID3 atua em processos de sinalização essencial ao metabolismo normal da célula, geralmente em células do sistema nervoso, entretanto também relacionado a outros tipos celulares, desta forma demonstrando que sua alteração implica em nesses processos básicos, sendo possível relacionar as alterações demonstradas em câncer (GRISS et al, 2020).

A busca no banco de dados intAct demonstrou uma interação mutua entre ID3 e LRP1B com uma proteína dedo de zinco ZNF408, (figura 6) porém não foi encontrado trabalhos relacionando este gene e câncer, apenas com outras patologias relacionadas a retina (AVILA-FERNANDEZ et al, 2015; CHEN et al, 2022). Entretanto, entende-se ZNF408 como uma proteína pertencente à família PRDM, a qual regulam a expressão gênica por modificação de histonas ou por interação de proteínas (KARJOSUKARSO et al, 2020). Permitindo a especulação de que esta proteína dedo de zinco, de alguma maneira, pode afetar a expressão destes genes e estar relacionada a suas diferenças de expressão. Desta forma, se sugere a necessidade de mais estudos moleculares in vivo e in vitro para compreensão do mecanismo moleculares de ZNF408 e suas interações com ID3, LRP1B e com outros genes, identificando talvez no futuro uma associação entre estes genes e câncer de tireoide.

Figura 6 – Rede de interação gerada a partir da busca no banco de dados IntAct.



Fonte: ORCHAD et al, 2013.

CONCLUSÃO

Câncer é uma patologia que assola a humanidade causando óbitos e prejuízos a qualidade de vida dos portadores. Estudos relacionados ao câncer, se fazem de suma

importante em um aspecto que busca processos terapêuticos que possibilitem a cura e preserve a dignidade dos pacientes. Diversas pesquisas têm-se voltado a esta área auxiliando na compreensão desta patologia. Atualmente, entende-se cada câncer como uma patologia ímpar, de origem genética com expressões gênicas diferentes dos tecidos normais. Desta forma, como já demonstrado em trabalhos descritos na literatura, se busca, com base nessas diferenças de expressões identificar padrões moleculares que sirvam como potenciais biomarcadores a diagnóstico e prognóstico, bem como potenciais alvos terapêuticos.

Bancos de dados públicos como o GEO e TCGA possuem conjuntos de dados, com enorme potencial de análise para novas descobertas relacionadas ao câncer. Estes permitem a deposição de dados gênicos, os quais poderão ser analisados por outros pesquisadores em novas pesquisas, permitindo maiores descobertas. Ao mesmo tempo a vasta literatura e repositórios biológicos possibilitam novas inferências moleculares, os quais podem apontar biomoléculas potenciais a realização de novos estudos in vitro e in vivo.

Uma grande dificuldade observada no cenário atual é a integração do conhecimento gnômico e do conhecimento computacional, bem como das análises estatísticas complexas realizadas em dados de expressão. Nota-se o crescimento de bancos de dados e de ferramentas de análises deste material. Entretanto, em grande maioria, mantém-se a utilização da linguagem de programação em R e em alguns casos python, visto que estas ferramentas permitem uma manipulação de dados mais dinâmica e individual. Desta forma, o presente trabalho aborda um contexto metodológico que visa exemplificar a análise de expressão diferencial de genes em câncer utilizando o banco de dados GEO e o ambiente R.

As etapas descritas no corpo do texto são divididas em obtenção dos dados; preparação dos dados, análise estatística e análise dos resultados. Os códigos utilizados para a realização dos mesmos se encontram no material suplementar. Para exemplificar este trabalho foi utilizado o câncer de tireoide, resultado em 20 genes diferencialmente expressos encontrados em todos os conjuntos analisados. Foram selecionados 2 genes para uma breve descrição e demonstrado com base na interação molecular, sua relação com a proteína dedo de zinco ZNF408.

O presente trabalho apresentou uma abordagem metodológica relacionada a identificação de genes diferencialmente expressos utilizando o banco de dado público GEO, bem como uma breve discussão acerca de dois genes selecionados, demonstrando seus aspectos biológicos e identificações prévias em outros estudos sobre câncer. Espera-se que este trabalho possa servir de auxílio a pesquisas futuras, e que seja utilizado como guia a identificação de novos genes diferencialmente expressos.

REFERÊNCIAS

ALAM, S. et al. Gene expression. profile analysis to discover molecular signatures for early diagnosis and therapies of triple-negative breast cancer. Frontiers. 9. 2022.

AMBROISE, J. et al. Impact of the spotted microarray preprocessing method on fold-change compression and variance stability. 12:413. 2011.

AVILA-FERNANDEZ, A; PEREZ-CARRO, R.; CORTON, M. Whole-exome sequencing reveals ZNF408 as a new gene associated with autosomal recessive retinitis pigmentosa with vitreal alteration. Hum Mol Genet. 24. 4037-4048. 2015.

BRUCE, A. et al. Fundamentos de Biologia Celular. 4.ed. Porto Alegre: Artmed, 2017.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. Nature. 513. 202-209. 2014.

CHEN, C. et al. Long-term clinical prognosis of 335 infant single-gene positive FEVR cases. 22(1). 2022.

CONCONI, D. et al. Analysis of copy number alteration in bladder cancer stem cells revealed a prognostic role of LRP1B. World J Urol. 40. 2267-2273. 2022.

DAVIS, S.; MELTZER, P. GEOQUERY: a bridge between the Gene Expression Omnibus (GEO) and Bioconductor. Bioinformatics. 23. 1846-11847. 2007.

DEED, R. W. et al. An immediate early human gene encodes an id-like helix-loop-helix protein and is regulated by protein kinase C activation in diverse cell types. 8(3): 599-607. 1993.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research. 30. 207-210. 2002.

GRISS, J. et al. Efficient Multi-Omics Comparative Pathways Analysis. Mol cell protemics. 2020.

JIMÉNEZ, V.; G.; MARTÍ-GOMEZ, C.; ANGEL, M. Bayesian inference of gene expresson. IN: HELDER, I. N, editor. Bionformatics. Brisbane(AU): Exon publcations. 2021.

JORDE, L. B.; CAREY, J.C.; BAMSHAD, M.J. Genética Médica. 5.ed. Rio de Janeiro: GEN. 2021.

KARJOSULKARO, D. W. et al. Modeling ZNF408-associated FEVR in Zebrafish Results in abnormal retinal vasculature. Biochemistry and molecular biology. 2020.

KIM, M. et al. A MYC-ZNF148-ID1/3 regulatory axis modulating cancer stem cell traits in aggressive breast cancer. Oncogenesis. 60. 2022.

LIU, W. et al. Bioinformatics analysis of key biomarkers for bladder cancer. Biomedical reports. 18. 2023.

MIAO, J. et al. Integrative analysis of the proteome and transcriptome in gastric cancer indetified LRP1b as a potential biomarker. Biomark med. 15. 1101-1111. 2022.

ORCHAD, S.; AMMARI, M.; ARANDA, B. et al. (2013). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction database. Nucleic Acids Res. 42.

PIERCE, B. A. (2017). Genética um enfoque conceitual. Rio de Janeiro: Guanabara Koogan.

PRINCIPE, C. et al. LRP1B: A giant lost in cancer translation. Pharmaceuticals. 14. 836. 2021.

R Core Team. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2020.

RAO, W. et al. Frequently mutated genes in predicting the relapse of stage I lung adenocarcinoma. Clin transl oncol. 2023.

RITCHIE, M. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nuclei Acids Research. 43. 2015.

ROSSUM, V.; DRAKE, J. F. L. Python reference manual. Centrum voor Wiskunde em informática Amsterdam. 1995.

SINHA, B. K. et al. (2022). Gene expression. profiling elucidates cellular responses to NCX3030 in human ovarian tumor cells: Implications in the mechanisms of action of NCX4040. Cancers Basel. 15. 2022.

SONG, Y. et al. Comprehensive molecular analyses of notch pathway-related genes to predict prognosis and immunotherapy response in patients with gastric cancer. Journal of oncology. 2023.

STEKEL, D. Microarray bioinformatics. New York: Cambridge. 2003.

TANG, Z. et al. Epigenetic Biomarkers of Breast Cancer Risk, Across the breast cancer prevention continuum. Advances in experimental medicine and biology. 2017.

VALIZADEH, S. et al. Upregulation of miR-142 in papillary thyroid carcinoma tisssues: a report based on in silico and in vitro analysis. Molecular Biology Research Communications. 11. 133-141. 2002.

ZHAO, Q.et al. Comprehensive analysis of ID genes reveals the clinical and prognostic value of ID3 expression in acute myeloid leukemia using bioinformatics identification and experimental validation. BMC Cancer. 22(1). 2022.