UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

HUDSON PEREIRA

DISSERTAÇÃO DE MESTRADO EM BIOINFORMÁTICA

ANÁLISE DE DADOS PÚBLICOS DE EXPRESSÃO GÊNICA DE DISTÚRBIOS DO ESPECTRO DO AUTISMO

CORNÉLIO PROCÓPIO

2022

HUDSON PEREIRA

ANÁLISE DE DADOS PÚBLICOS DE EXPRESSÃO GÊNICA DE DISTÚRBIOS DO ESPECTRO DO AUTISMO

ANALYSIS OF PUBLIC DATA OF GENE EXPRESSION OF AUTISM SPECTRUM DISORDERS

Dissertação apresentada como requisito para obtenção do título de Mestre em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR).
Orientador(a): Alexandre Rossi Paschoal
Coorientador(a): Artur Trancoso Lopo de Queiroz

CORNÉLIO PROCÓPIO

2022

Ministério da Educação
Universidade Tecnológica Federal do Paraná
Campus Cornélio Procópio

HUDSON PEREIRA

ANÁLISE DE DADOS PÚBLICOS DE EXPRESSÃO GÊNICA DE DISTÚRBIOS DO ESPECTRO DO AUTISMO

Trabalho de pesquisa de mestrado apresentado como requisito para obtenção do título de Mestre Em Bioinformática da Universidade Tecnológica Federal do Paraná (UTFPR). Área de concentração: Bioinformática.

Data de aprovação: 08 de Abril de 2022

Dr. Alexandre Rossi Paschoal, Doutorado - Universidade Tecnológica Federal do Paraná

Dr. Danilo Sipoli Sanches, Doutorado - Universidade Tecnológica Federal do Paraná

Dra. Flavia Cristina De Paula Freitas, Doutorado - Instituto Carlos Chagas

Marcella Scoczynski Ribeiro Martins, - Universidade Tecnológica Federal do Paraná

Documento gerado pelo Sistema Acadêmico da UTFPR a partir dos dados da Ata de Defesa em 07/06/2022.

Dedico este trabalho à minha Vó Dione Striker, esposa Evelyn A. Pereira e meus filhos, Thais e Guilherme Striker Pereira, pelos momentos de ausência.

AGRADECIMENTOS

Primeiramente sou grato a DEUS que de forma singular me conduziu até aqui, me capacitando, dando saúde e criando sonhos. Chegar até aqui não foi fácil, realmente foram dias de muita luta e alguns momentos sem motivação, porém tudo isso me proporcionou novas metas, resiliência e perseverança.

Ao querido professor e orientador, Prof. Dr Alexandre Rossi Paschoal, pela orientação, paciência e confiança. Muitas vezes fazendo o papel não só de orientador mas também de amigo, gerando motivação nos momentos em que eu estava sem energia para continuar. Não posso deixar de comentar que em muitos dos pedidos de ajuda, em horários de descanso dele, sempre que solicitado ajuda, lá estava meu orientador, pronto a responder minhas dúvidas; foram noites, finais de semana e sempre esteve disponível, deixo meu registro de gratidão por esse respeito e companheirismo.

Ao Prof. Dr. Artur Trancoso Lopo de Queiroz, obrigado pela ajuda e pelas opiniões e comentários sobre o trabalho.

Agradeço a Minha esposa Evelyn, minha filha Thais e meu filho Guilherme, que sempre entenderam a ausência da minha parte, pois foi necessário para desenvolvimento deste trabalho, ao Guilherme, meu filho caçula, portador do espectro autista, que DEUS nos presenteou, sendo o meu motivador em buscar conhecimento sobre o tema e continuar estudando sobre o autismo.

A minha avó Dione, que sempre motivou, me ajudou e contribuiu com incentivos, motivações e conversas sobre o trabalho, muitas vezes sonhando junto comigo os resultados desse trabalho.

Dedico esse trabalho a Profª Drª Maria Berenice Reynaud Steffens (in memoriam) que mesmo eu não tendo a oportunidade de conhecer pessoalmente em tempos de pandemia, tive o prazer de receber valiosos comentários e sugestões sobre esse trabalho e futuros projetos, infelizmente ela nós deixou de forma precoce devido a covid-19, porém deixou seu legado de conhecimento que era transmitido com o tom de voz calmo e sábios comentários sobre o tema deste trabalho.

Quando se perde a riqueza, nada se perde, quando se perde a saúde, algo se perde; quando se perde o caráter, perde-se tudo. (Billy Graham).

RESUMO

A síndrome do Transtorno do Espectro do Autismo (TEA) é caracterizada por dificuldades de interação, desvio na comunicação e comportamentos repetitivos. Essa síndrome também é definida como perda de contato com a realidade, causada por impossibilidade ou grande dificuldade na comunicação interpessoal. O TEA pode ser classificado de acordo com a gravidade em: leve, moderado e grave. O diagnóstico precoce do autismo é essencial para um tratamento eficaz. As análises transcriptômicas são um meio de obter informações regulatórias para entender o TEA. Nesse sentido, este trabalho apresenta o resultado de uma meta-análise em dados públicos de expressão gênica disponíveis do TEA em estudos associados. A a metodologia aplicada consistiu em utilizarmos dados de expressão obtidos após uma revisão da literatura sobre a TEA, Sendo, três conjuntos de dados selecionados, coletados no portal NCBI GEO em Dezembro/19, e analisados via dados RNA-Seq os genes chaves relativos à TEA. O pipeline de análise de RNA-Seq foi utilizado para: (i) extração dos dados em SRA utilizando o fastq-dump, no Rstudio; (ii) avaliação e controle de qualidade via programa Trimmomatic, no qual foi feito o corte de qualidade das sequências; (iii) em seguida, os dados foram alinhados com o genoma de referência (GRCh38) utilizando o Salmon e aplicado a estimativa de quantificação e nível de transcrição; e (iv) o txtimport foi utilizado para a montagem da matriz de contagem, por fim, utilizamos o DESeq para análise de expressão diferencial. A análise da dispersão dos dados de expressão foram exibidos graficamente usando o Vulcano. Em seguida, a técnica PCA (do inglês Principal component analysis) para análise de grupos, junto com a análise de genes enriquecidos, utilizando os termos do GO, identificamos potenciais, grupos e funções dos genes analisados sendo possível identificar um total de dez genes diferencialmente expressos, sendo três genes altamente expressos e sete genes com baixa expressão. Destes genes, oito são codificadores de proteínas, e dois RNAs pequenos. Além disso, foi observado que alguns genes apresentam relação com outra doença genética, no caso a esquizofrenia.

Palavras-chaves: Transcriptoma, RNA-Seq, TEA, meta-análise, Expressão gênica, Bioinformática.

ABSTRACT

Autism Spectrum Disorder (ASD) syndrome is characterized by interaction difficulties, communication deviation and repetitive behaviors. This syndrome is also defined as loss of contact with reality, caused by impossibility or great difficulty in interpersonal communication. ASD can be classified according to severity into: mild, moderate and severe. Early diagnosis of autism is essential for effective treatment. Transcriptomic analyzes are a means of obtaining regulatory information to understand ASD. In this sense, this work presents the result of a meta-analysis on publicly available gene expression data from ASD in associated studies. The methodology applied consisted of using expression data obtained after a review of the literature on ASD, being, three sets of selected data, collected in the NCBI GEO portal in December/19, and analyzed via RNA-Seq data the key genes related to TEA The RNA-Seq analysis pipeline was used to: (i) extract data in SRA using fastq-dump, in Rstudio; (ii) evaluation and quality control via the Trimmomatic program, in which the quality cut of the sequences was performed; (iii) then, the data were aligned with the reference genome (GRCh38) using Salmon and applied to estimate quantification and transcription level; and (iv) txtimport was used to assemble the counting matrix, finally, we used DESeq for differential expression analysis. The scatter analysis of expression data was displayed graphically using Vulcan. Then, the PCA (Principal component analysis) technique for analysis of groups, together with the analysis of enriched genes, using the terms of the GO, we identified potentials, groups and functions of the analyzed genes, being possible to identify a total of ten genes differentially. expressed, being three genes highly expressed and seven genes with low expression. Of these genes, eight are protein-coding, and two are small RNAs. In addition, it was observed that some genes are related to another genetic disease, in this case schizophrenia.

Keywords: Transcriptome, RNA-Seq, TEA, meta-analysis, Gene expression, Bioinformatics.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

TEA	Transtorno do Espectro Autista
GEO	Gene Expression Omnibus (Base de dados de expressão gênica Omnibus)
DEGs	Genes Diferencialmente expressos
RNA-Seq	Sequenciamento de nova geração aplicado para quantificação de ácido ribonucleico
IPSC	Universidade Tecnológica Federal do Paraná
NCBI	Centro Nacional de Biotecnologia

SUMÁRIO

1 INTRODUÇÃO

O transtorno do espectro do autismo (TEA) é definido como um conjunto de distúrbios do neurodesenvolvimento caracterizado por falta de comunicação social, movimentos repetitivos e hiper-atenção (Ornoy, at al., 2015). Estudos revelam que, na última década, o número de pessoas com diagnóstico de TEA aumentou e a taxa de prevalência foi excedida de 1 pessoa em 150 no mundo inteiro (Posar A at al., 2015). TEA é uma doença relacionada a fatores genéticos e, neste sentido, um grande desafio é identificar as alterações genéticas que são responsáveis pelo TEA (De Rubeis, S at al.,2015). O número de descobertas de novos genes relacionados a doenças complexas tem aumentado (Posar A at al., 2015). Informações advindas de estudos de associação por todo o genoma (do inglês genome-wide association studies - GWAS) e de transcriptoma têm mostrado que, tanto a variabilidade genômica (polimorfismos), quanto a variabilidade transcriptômica (variação da expressão gênica) têm influência na susceptibilidade de doenças (Need et al., 2010). No entanto, a maneira pela qual estes genes se relacionam ainda não está completamente elucidada.

Nesse cenário, na atual era do "big data", é crescente o número de estudos relacionados ao TEA, principalmente referente à dados de expressão gênica que são depositados em repositórios públicos, como o GEO NCBI. Por exemplo, a revisão neste trabalho identificou 79 estudos com dados sobre autismo depositados no Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/) até a data de 2017. A consequência natural é uma gama de grupos independentes produzindo estudos em diversos aspectos, tecidos e condições em TEA, mas individualizados, sem uma análise comparativa entre esses dados. Como aplicação da bioinformática, no contexto deste projeto, a análise de dados pode contribuir para elucidar divergências e similaridades nesta síndrome. Nesse sentido, abre uma oportunidade de aplicar a análise de dados in silico em diversos estudos independentes para investigar a seguinte questão: quais são e quem são os potenciais genes chaves nos dados públicos de TEA.

Deste modo, este dissertação partiu de uma análise dos dados públicos de expressão disponíveis no GEO, e investigou, de forma comparativa, três conjuntos de dados de RNA-Seq, para buscar responder essa pergunta. Após análise de qualidade,

controle, expressão diferencial, e técnicas complementares (e.g., PCA, termos GO), foi possível elucidar 10 genes potencialmente diferencialmente expressos, sendo 7 codificadores e 3 não-codificadores.

A seguir, são pontuados de forma clara e objetiva, os principais conceitos necessários para compreensão das técnicas utilizadas neste projeto, em seguida os objetivos, e um resumo sobre a organização do restante do texto.

1.1 Transtorno do Espectro Autista

O Transtorno do Espectro Autismo (TEA) é uma condição complexa do desenvolvimento que envolve desafios persistentes na interação social, fala, comunicação não-verbal e comportamentos restritos e repetitivos. O grau de severidade do TEA e a gravidade dos sintomas são diferentes para cada pessoa. O grau de comprometimento que os indivíduos apresentam para cada um desses sintomas é diverso, indo desde formas mais leves, em que os pacientes têm uma vida independente, até os mais graves, o que os impossibilitam de ter uma vida social, podendo apresentar comportamento agressivo, distúrbio do sono e alimentar, problemas gastrointestinais, hiperatividade e ansiedade. (*American Psychiatric Association, 2019*)

O diagnóstico, na maioria das vezes, é feito pela primeira vez na infância, utilizando como parâmetros de análise os sinais mais comuns, classificado como sinais estereótipos do TEA, com cerca de 2-3 anos de idade. TEA também é até quatro vezes mais comum em meninos do que em meninas, e meninas com TEA tendem a apresentar menores sinais estereótipos em relação aos meninos (*Linhas de cuidado, Ministério da Saúde 2021*).

O TEA apresenta heterogeneidade genética e pode ser enquadrado em diferentes modelos de herança, o que torna difícil apontar os fatores genéticos associados ao transtorno (*Linhas de cuidado, Ministério da Saúde 2022*).

1.2 Delineamento e desafios da proposta desta pesquisa

Com a prevalência dos TEA na população (Posar A at al., 2015) a pesquisa sobre esse tema é importante para entendimento das causas e potenciais medidas de prevenção. As pesquisas biológicas relacionadas ao TEA também são importantes

para a compreensão da classe de distúrbios do desenvolvimento neurológico. As deficiências da infância são cada vez mais classificadas como comportamentais/neurológicas e é provável que haja pontos em comum nas etiologias e tratamentos das condições (Neal Halfon et al., 2012).

A literatura apresenta descrição das estruturas do sistema nervoso central anatomicamente alteradas em indivíduos diretamente afetados pelo TEA, por exemplo o cerebelo, a amígdala e o córtex frontal (Donovan et al., 2017). Porém, essas alterações não são observadas em todos os pacientes com TEA. Contudo, uma grande dificuldade em estudos relacionados ao sistema nervoso central e a obtenção de tecido ou amostras para a condução de experimentos celulares e moleculares.

O Autismo é uma condição relacionada ao desenvolvimento do cérebro, sendo ele o órgão-chave para estudos e compreensão da doença. Por se tratar de uma região complexa, por muito tempo houve a dificuldade da retirada de amostras através de biópsia ou na aquisição de amostras post-mortem, sendo um grande desafio no avanço dos estudos do TEA (Neuroimagem. Brazilian Journal of Psychiatry, 2006).

Com o avanço das pesquisas relacionadas às células-tronco pluripotentes induzidas (iPSC, do inglês induced pluripotent stem cells), as quais são resultados de um processo em que as células somáticas são reprogramadas a um estado pluripotente a partir da superexpressão de genes específicos (Takahashi et al., 2007), tornou-se possível a geração de modelos de doenças in vitro e in vivo, o que viabiliza o estudo de células neurais por diferentes doenças.

A análise de expressão gênica em dados de iPSC de células neurais revelou a regulação gênica em um subconjunto de genes, sugerindo um padrão de expressão que pode ser usado como biomarcador do TEA que em outras pesquisas; o padrão não se repetiu com dados de post-mortem (Griesi-Oliveira et al., 2020). Embora estudos de perfil transcricional em TEA utilizando conjunto de dados em microarray tenham identificado mudanças na expressão gênica, pouca informação foi identificada para células específicas, o que dificulta a elucidação de novos biomarcadores ou expressão gênica relacionada ao TEA. (Parikshak et al., Nature, 2016).

Será apresentado no próximo capítulo, uma diversidade de dados públicos de expressão em TEA disponível em banco como o GEO. A grande questão, é alvo de estudo deste projeto, é ver a convergência e divergência entre diversos estudos de TEA realizados por grupos independentes, e que podem assim, elucidar potenciais

novos genes relacionados a TEA. Desta forma, esta proposta teve como objetivo elucidar genes chaves ligados ao TEA, baseado na análise de bioinformática e mineração de dados de modo a investigar esses dados públicos de TEA. associados ao transtorno (*Linhas de cuidado, Ministério da Saúde 2022*).

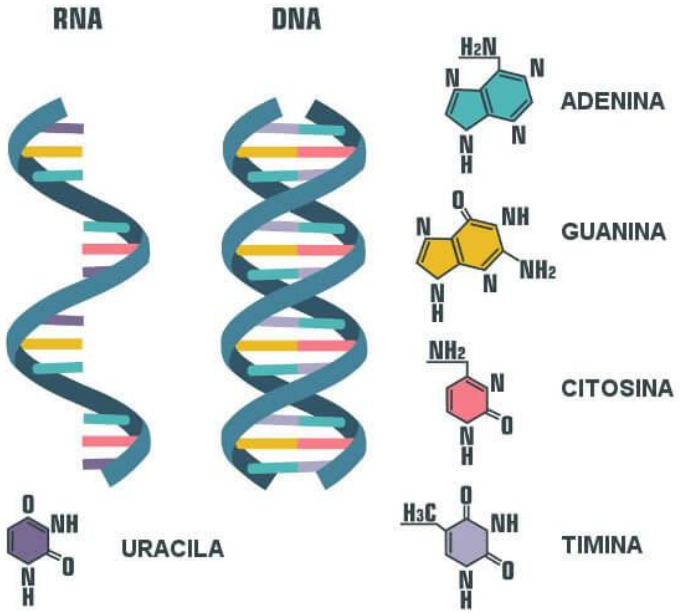
1.3 Análise de expressão Gênica

A expressão gênica é o processo no qual as instruções contidas no DNA são convertidas em um produto funcional (Figura 2). Para executar esse processo, a célula interpreta o código genético e para cada três letras, adiciona um dos 20 aminoácidos diferentes que são as unidades básicas necessárias para construir proteínas. Como a expressão gênica varia de acordo com as condições na qual o organismo se encontra, é possível se utilizar de análise de RNA-Seq para mensurar os efeitos de diferentes tratamentos ou condições na expressão de diferentes genes simultaneamente. Esse processo de análise é conhecido como expressão gênica diferencial, ele permite entender como o perfil de expressão de um determinado organismo é alterado a ser submetido a uma determinada condição. No processo de análise de expressão gênica diferencial, é possível identificar genes com expressão aumentada (up-regulated) ou diminuída (down-regulated) em determinada situação, ou tecido (HAZEN et al., 2003).

1.3.1 A molécula de RNA

O RNA (também conhecido por ácido ribonucleico)é constituídopor uma pentosee um fosfato, tendo como bases nitrogenadas a adenina, guanina, citosina e uracila. O RNA, ao contrário do DNA, é composto apenas por uma fita e ela é produzida no núcleo celular a partir de uma das fitas de uma molécula de DNA. Depois de pronto, o RNA segue para o citoplasma celular, onde desempenha sua principal função, que é controlar a síntese de proteínas(Fleischmann, et al., 1995).

O Sequenciamento de RNA engloba um conjunto de técnicas experimentais e computacionais que possibilitam identificar a quantidade de sequências de RNA em amostras biológicas em um determinado estágio de desenvolvimento (KORPELAINEN et al., 2015).



Fonte: Socientifica, 2021

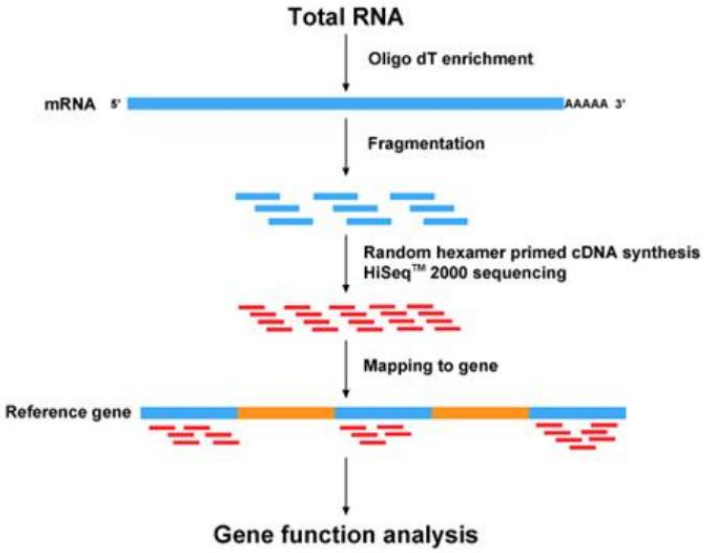
Figure 1 - Estrutura molecular de DNA e RNA. Na figura a molécula de DNA é uma dupla hélice, enquanto o RNA apresenta uma cadeia mais simples. As moléculas de DNA possuem dois polinucleotídeos que se espiralam, formando a estrutura conhecida como dupla hélice. A parte externa da hélice é formada pelas cadeias principais de açúcar-fosfato, quando as bases nitrogenadas estão pareadas no interior da hélice. Os dois polinucleotídeos estão unidos por ligações estabelecidas entre os pares de bases.

1.3.2 RNA-Seq

O RNA-Seq (do inglês – RNA sequencing) é uma abordagem baseada no sequenciamento de nova geração ou (do inglês – Next-generation sequencing NGS) apresenta e quantidade de RNA em um transcriptoma (ZONG et al., 2014). Essa abordagem pode dizer quais genes estão ativos em uma célula, e qual os genes que estão mais expressos (*up-regulated*), ou com menor expressão (*down-regulated*) Isso permite o entendimento do perfil transcricional, identificação de SNP (Polimorfismo de Nucleotídeo único) e análise diferencial de expressão gênica, o que pode fornecer informações sobre a função dos genes, ou destacar quais genes estão sendo expressos em tecidos do corpo (ZONG et al., 2014) e analisados estatisticamente.

Análises estatísticas são realizadas para identificar os transcritos diferencialmente expressos (Figura 2). São utilizados programas de bioinformática para agrupar os genes, quantificar e mapear as reads (Leituras de sequenciamento)

com base em amostras, e então os dados são exibidos graficamente, apresentando os genes mais expressos (*up-regulated*) ou menos expressos (*down-regulated*).



Fonte: Bio Lundberg, 2021.

Figure 2 - Fluxograma do RNA-Seq. As amostras de RNA são convertidas em bibliotecas de cDNA, que são separadas por adaptadores e os fragmentos são sequenciados. O sequenciamento gera centenas de milhares de leituras (reads) que são analisadas estatisticamente para identificar os transcritos diferencialmente expressos.

1.3.3 Vantagens do RNA-Seq sobre Microarray

A análise de expressão gênica é uma ferramenta útil para se investigar doenças. Um trabalho publicado por Mahan e colaboradores (2019) teve como objetivo comparar as duas tecnologias de análise de transcriptômicas (RNA-Seq e Microarray) para determinar se o RNA-Seq oferece vantagens sobre o Microarray em estudos toxicogenomicos.

Houve um resultado com maior expressão diferencial dos genes quando utilizamos a plataforma de RNA-Seq comparado com microarray (*Mahan et al, fontiers 2019*) . Ambas as plataformas identificaram um número maior de genes diferencialmente expressos (DEGs), porém, o RNA-Seq mostrou DEGs adicionais.

O RNA-Seq fornece um nível mais alto de sensibilidade e precisão(SCHULZE; DOWNWARD, 2001), bem como a identificação de novas

expressões, comparado com o microarray, o RNA-Seq demonstra ser uma ferramenta valiosa para estudos de expressão gênica.

2 OBJETIVOS

Este trabalho teve como objetivo analisar dados públicos de estudos de RNA-Seq diretamente relacionadosao Espectro do Autismo disponíveis no banco de dados Públicos (GEO).

2.1 OBJETIVO ESPECÍFICOS

- Revisar a literatura, coleta e definição dos dados públicos de expressão relativos a TEA para realizar posterior análise neste projetO.
- Analisar dados de expressão gênica diferencial de TEA.
- Analisar o perfil de expressão dos dados coletados.
- Identificar genes diferencialmente expressos que possam estar unicamente relacionados ao TEA.
- Disponibilizar todos os dados deste estudo à comunidade científica, na forma de uma manuscrito a ser publicado em periódico Q1.
- Contribuir para formação de recursos humanos no PPGBIOINFO.

2.1.1 Organização desta proposta

O primeiro capítulo deste trabalho faz uma apresentação da pesquisa, apresentando o que é o autismo, desafios e oportunidades relacionados à pesquisa e os objetivos. O segundo capítulo contém os trabalhos (Figura 4) relacionados que foram identificados junto com os dados (Figura 3) a serem utilizados na pesquisa. No terceiro capítulo, é apresentada a metodologia e as ferramentas propostas para análise dos dados de Autismo utilizados nesta pesquisa.

3 REVISÃO DE LITERATURA DOS TRABALHOS RELACIONADOS

Neste trabalho, foi feita uma revisão da literatura com dois objetivos (i) levantamento dos dados públicos de transcriptoma referente a TEA; e (ii) evidenciar os estudos de meta-análise na temática.

O levantamento dos dados de expressão sobre autismo foi feito no banco de dados *Gene Expression Omnibus* (GEO) (*https://www.ncbi.nlm.nih.gov/geo/*). Para tanto, a pesquisa para coleta foi feita com a sintaxe de busca: ((autism) AND "Homo sapiens"[porgn:__txid9606]). Em seguida, foram aplicados os filtros *Homo sapiens em* **organismo**, e o **tipo de estudo** apenas de RNA-Seq (estudo de expressão baseado no sequenciamento de nova geração - *technology-based sequencing next generation sequencing*). O resultado após o filtro foi inspecionado de forma manual, de modo a manter apenas os registros com os seguintes critérios: (i) Dados biológicos de humanos com TEA; (ii) dados de experimentos de RNA-Seq; e (iii) Tipo de amostra, tecido e cultura biológica do conjunto de dados. A revisão dos trabalhos foi feita no período de Novembro/2019 a Dezembro/2019 e os demais dados de TEA que não se enquadraram nesses critérios foram descartados. Os resultados dos estudos selecionados estão resumidos na Tabela 1.

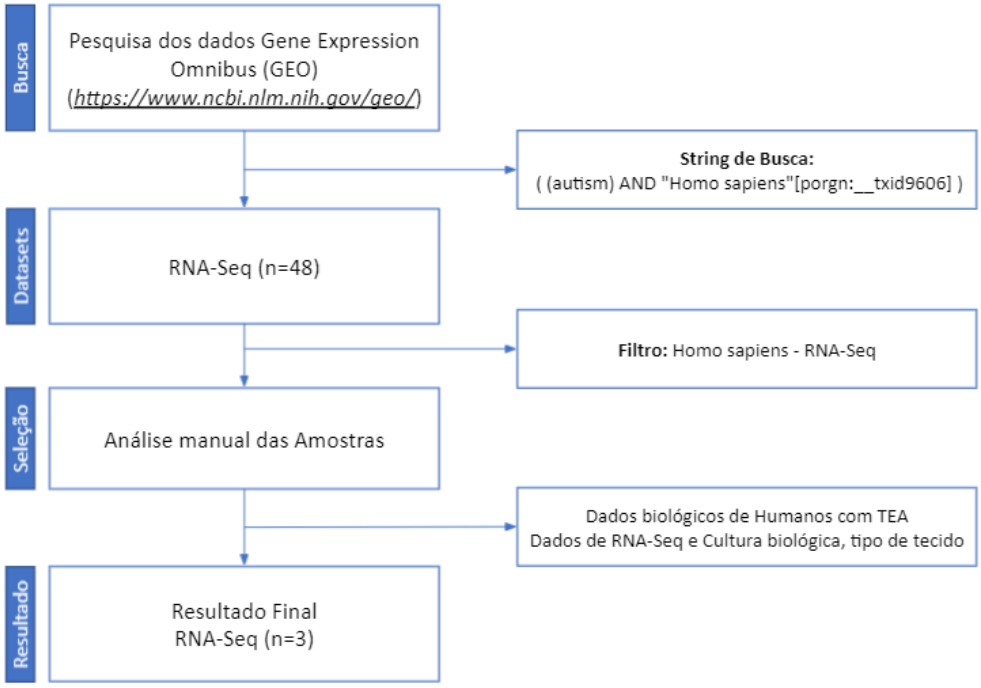


Figure 3 - Revisão dos dados públicos de expressão de TEA: Os dados foram obtidos no repositório do GEO NCBI, sendo selecionados apenas dados de RNA-Seq, ao total 48 conjunto de dados, na sequência utilizado critério de seleção (Homo Sapiens) e por fim aplicado o filtro de seleção para manter apenas dados, biológicos e tecido neural. Resultando em três conjuntos de dados para análise.

ID GEO	PMID	Tecido	Amostras	Plataforma
GSE129808	31540669	IPSC Neurônios	7	Illumina HiSeq 2500
GSE67528	27378147	IPSC Neurônios	83	Illumina HiSeq 2000
GSE61476	26186191	IPSC Neurônios	45	Illumina HiSeq 2000

Tabela 1 Tabela do conjunto de dados de RNA-Seq utilizados na análise de expressão diferencial.

3.1 CONJUNTO DE DADOS

A partir da revisão (figura 3), foram usados três conjuntos de dados de RNA-Seq publicados e disponíveis no GEO sobre o número de acesso GSE129808, GSE67528 GSE61476. O conjunto de dados de Ross e colaboradores (2020) (GSE129808) relata o estudo utilizando IPSC (Células-tronco pluripotentes induzidas), de pessoas com TEA. O conjunto de dados GSE67528 (Marchetto MC et al., 2017) relata a geração de células IPSC de oito pacientes com TEA que foram usadas para derivar células progenitoras neurais e neurônios em cultura. E por último, o conjunto de dados GSE61476 (Mariani et al., 2015) que foi utilizado para células iPSC derivadas de culturas neurais tridimensionais (organóides) em pacientes com TEA e

microcefalia para investigar alterações no desenvolvimento neurológico que causam essa forma de TEA.

Por fim, como forma de certificar-se do estado da arte sobre estudos de meta-análise similar ou igual ao desenvolvido (figura 4), foi feito uma busca na literatura via o sítio PubMed (https://pubmed.ncbi.nlm.nih.gov/). Para a revisão, foi usado a sintaxe de busca ((autism[Title]) AND (meta-analysis[Title])) apenas no título dos artigos. Em seguida, foram aplicados os filtros de: (i) artigos nos últimos 10 anos; (ii) espécie Humano; e (iii) idioma Inglês. A revisão dos trabalhos foi feita no período de Novembro/2019 a Dezembro/2019. Em seguida, uma inspeção manual dos trabalhos revisados considerando os critérios de inclusão e exclusão dos artigos. Os critérios de inclusão utilizados foram: (i) artigos provenientes de meta-análise relacionados ao autismo; (ii) artigos publicados nos últimos 10 anos; (iii) Trabalhos em língua inglesa; e (iv) Trabalhos utilizando humanos. Já para os critérios de exclusão foram definidos: (i) artigos não relacionados com genética ou biologia; (ii) Trabalhos utilizando animais que não humanos (e.g. camundongo); (iii) artigos utilizando técnicas de imagem, texto e leitura de vídeo. Todos os trabalhos foram lidos na íntegra. A busca inicial sem filtro retornou um total de 213 artigos. Ao ser aplicado o critério de exclusão, restaram o total de 23 artigos. Por último, foram selecionados os artigos com análise de expressão gênica em base de dados com autismo, contabilizando um total de 3 artigos(figura 4).

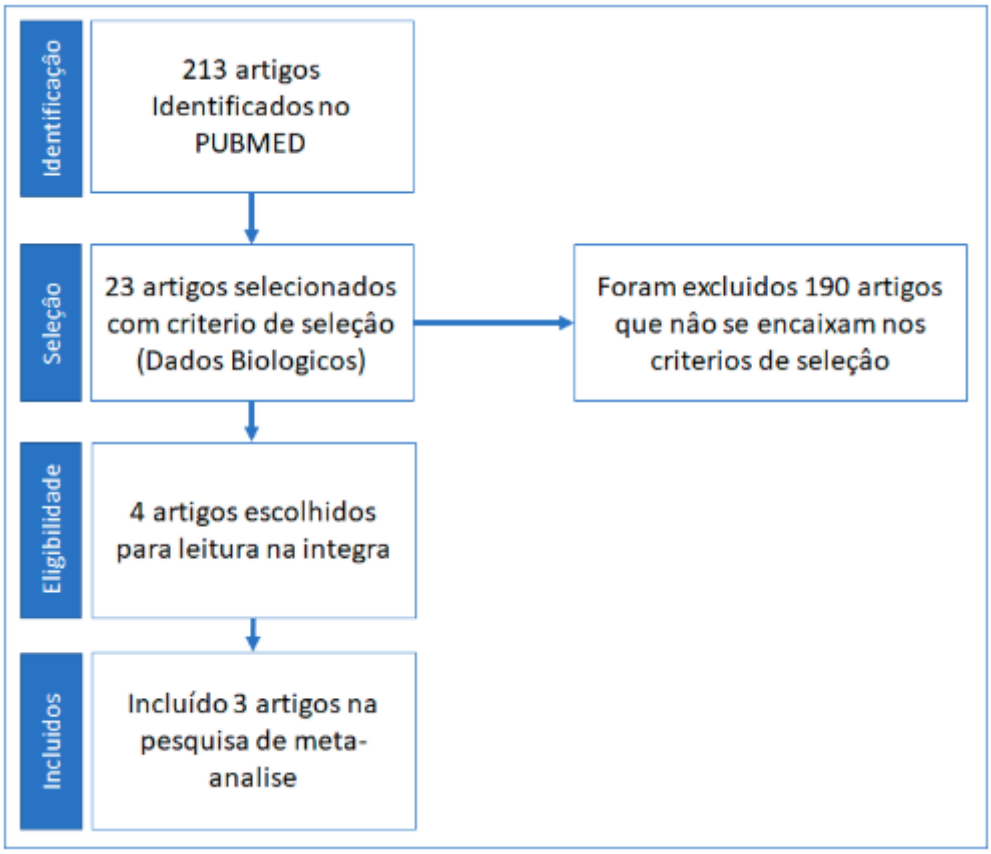


Figure 4 - Revisão de literatura para atualização quanto ao estado da arte no tema do projeto: Os trabalhos científicos foram obtidos no repositório público do PUBMED, sendo selecionados os artigos relacionados a dados biológicos de humanos, na sequência os artigos selecionados foram lidos na íntegra e por fim selecionados apenas três artigos.

O trabalho do banco de dados dbMDEGA (Schena et al.,1995) apresenta um banco de dados biológicos relacionados ao Autismo. Foram utilizados modelos de expressão de tecido cerebral (Córtex e cerebelo) de seres humanos e modelos de camundongos com sintomas relacionados ao Autismo; dados analisados em microarray.

Já Ning e colaboradores (Ning et al., 2015) realizaram uma meta-análise usando dez conjuntos de dados públicos disponíveis no GEO, sendo nove conjuntos dados de microarray e um conjunto de dados de RNA-Seq de portadores do TEA. Os tecidos utilizados foram amostras de sangue e amostras de neurônios pós-mortem.

Por fim, no último estudo foram utilizados doze conjuntos de dados, sendo nove conjuntos de amostras de sangue e três amostras de tecido cerebral. Este estudo teve como objetivo a meta-análise de dados com intuito de identificar semelhanças moleculares entre os grupos de amostras.(Carolyn et al., 2016).

Nota-se que, apesar de apenas três trabalhos, estes são focados em perguntas específicas como tecidos específicos, o que contribuiu para o objetivo deste projeto que foi a análise comparativa entre três estudos de TEA.

3.2 BACKGROUND SOBRE ANÁLISES DE EXPRESSÃO DIFERENCIAL

Neste pipeline foi utilizado o Rstudio na versão R-4.1.2, sendo utilizado as ferramentas abaixo para cada etapa.

3.2.1 Download GEO NCBI

Utilizamos o programa *wget* no linux para fazer download das amostras (SRA) no repositório do NCBI GEO. a busca e download foi executada através do código (*bioproject*) e salvos no formato .csv.

3.2.2 Fastq-dump

Após baixar as leituras de sequenciamento (SRA) do repositorio do GEO NCBI, utilizamos o programa Fastq-dump para (i) comprimir as saídas das leituras, (ii) descartar as leituras técnicas e seleção de leituras biológicas, (ii) Anexar o ID de em cada leitura para diferenciar as leituras entre pares ou não, (iii) filtrar leituras que estão de acordo com a filtragem e seleção (iv) dividir as leituras do FASTQ em dois arquivos, sendo que um terceiro arquivo é gerado sem leituras repetidas, (v) remoção das sequências de SRA com tags não conforme e por fim, (vi) Formata as sequências, no qual os pares de bases são representados por números).

3.2.3 Trimmomatic

A limpeza das leituras e pares de base de baixa qualidade foram removidas usando o TrimmomaticSE, seguindo as etapas: (i) remoção dos adaptadores Illumina (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10), (ii) remoção das bases principais de baixa qualidade (abaixo de qualidade 3) (LEADING:3), (iii) remoção das bases finais de baixa qualidade (abaixo de qualidade 3) (TRAILING:3), (iv) leitura de 4 bases e retirando quando a qualidade média por base ficar abaixo de 15

(SLIDINGWINDOW:4:15), e por fim, (v) retira as leituras inteiras que estão abaixo do comprimento específico (MINLEN:36).

3.2.4 Salmon

O Salmon foi utilizado para explorar a qualidade das leituras brutas e quantificação. O objetivo desta etapa é identificar de qual transcrição cada uma das leituras se originou e o número total de leituras associadas a cada transcrição. utilizamos transcriptoma de referência (no formato FASTA) e as leituras de sequenciamento bruto (no formato FASTQ) como entrada para realizar o mapeamento e a quantificação das leituras.

O Transcriptoma de referência humano (Homo Sapiens GRCh38) foi baixado do repositório do Ensembl (https://ensembl.org/), sendo um conjunto de arquivos *cdna* e outro *ncrna*, depois mesclamos os dois conjuntos transformando em um único arquivo, com intuito de analisar regiões não codificantes.

A execução do Salmon foi executado em duas fases; Quantificação da abundância em nível de gene e indexação. por fim, utilizando os dados extraídos do Salmon, montamos a matriz de contagem que será utilizado na fase posterior de análise de expressão diferencial.

3.2.1 Combat

Foi ajustado o efeito em lote do conjunto de dados utilizando o pacote do Combat (Johnson et al. 2007). sendo que os dados de entrada, foram retirados o efeito em lote e normalizados. O combat utiliza a metodologia de Bayes para ajudar o efeito em lote.

3.2.2 Deseq2

Na etapa final do fluxo de trabalho, o pacote Deseq2 foi utilizado para análise de expressão diferencial. Para essa análise as contagens brutas são ajustadas ao modelo Naive Bayses e realizado o teste estatístico para genes expressos diferencialmente entre dois grupos (Autismo) e (Controle).

Os metadados foram importados para o Deseq, (i) os dados da matriz foi ajudado para o p-valor 0,05, (ii) conforme o p-valor ajustado os dados foram transformados em dataframe, (iii) os dados foram sumarizados, sendo p-valor ajustado <0,05 para genes (up-regulated) e Log2FC (down-regulated). Por fim, plotamos esses dados em um gráfico de Vulcano.

4 ANÁLISE DE DADOS DE EXPRESSÃO DO AUTISMO

Este capítulo apresenta a metodologia, bem como a análise e os resultados da dissertação. Trazemos para esta seção a versão formatada, em língua portuguesa, do artigo científico que será submetido

Análise de dados públicos de Distúrbios do Espectro do Autismo

Hudson Pereira[1], Eduardo Fukutani[2], Artur Trancoso Lopo de Queiroz[2], Alexandre Rossi Paschoal[1]

[1]Departamento de Computação, Programa de Pós-Graduação em Bioinformatica(PPGBIOINFO),Universidade Tecnológica Federal do Paraná (UTFPR), Cornélio Procópio, Paraná.
[2]Instituto Gonçalo Moniz – FIOCRUZ/BA, Salvador, Bahia

RESUMO

A síndrome do Transtorno do Espectro Autista (TEA) é caracterizada por dificuldades de interação, desvio na comunicação e comportamentos repetitivos. Essa síndrome também é definida como perda de contato com a realidade, causada por impossibilidade ou grande dificuldade na comunicação interpessoal. O TEA é classificado em três graus de gravidade: leve, moderado e grave. O diagnóstico precoce do autismo é essencial para um tratamento eficaz. As análises transcriptômicas fornecem informações importantes para entender o TEA do ponto de vista de expressão e regulação gênica. Ainda, é crescente o número de estudos que disponibilizam dados públicos de transcriptoma, inclusive em doenças. Nesse sentido, este trabalho apresenta uma meta-análise de dados públicos de expressão gênica disponíveis do TEA em estudos associados. Deste modo, dados de três estudos de expressão gênica de RNA-Seq de TEA foram investigados, com o objetivo de elucidar potenciais genes chaves relativos à TEA. Em seguida foi aplicado um pipeline para análise de expressão contendo (i) coleta dos dados, remoção de adaptadores e controle de qualidade, montagem da matriz de contagem das reads e por fim foi utilizado o DESeq para análise de expressão diferencial. Por fim, foram identificados 10 genes (7 codificantes de proteínas e 3 não codificantes).

Palavras-chaves: Transcriptoma, RNA-Seq, TEA, meta-análise, Expressão gênica, Bioinformática.

INTRODUÇÃO

O transtorno do espectro do autismo (TEA) é definido como um conjunto de distúrbios do neurodesenvolvimento caracterizado por falta de comunicação social, comportamentos repetitivos e hiper-atenção (*Ornoy et al., 2015*). Estudos revelam que na última década, o número de pessoas com

diagnóstico de TEA aumentou e a taxa de prevalência foi excedida de 1 pessoa em 150 no mundo inteiro (*Posar A at al., 2015*). TEA é uma doença relacionada a fatores genéticos e, neste sentido, um grande desafio é identificar as alterações genéticas que são responsáveis pelo TEA (*De Rubeis, S at al.,2015*). O número de descobertas de novos genes relacionados a doenças complexas tem aumentado. Informações advindas de estudos de associação por todo o genoma (do inglês *genome-wide association studies-GWAS*) e de transcriptoma têm mostrado que, tanto a variabilidade genômica (polimorfismos), quanto a variabilidade transcriptômica (variação da expressão gênica) têm influência na susceptibilidade de doenças (*Need et al., 2010*). No entanto, a maneira pela qual estes genes se relacionam ainda não está completamente elucidada.

Com o grande número de estudos relacionados ao TEA, é crescente o número de dados de expressão gênica que são depositados em repositórios públicos, como o GEO NCBI. Por exemplo, a revisão aqui apresentada identificou 79 estudos com dados sobre autismo depositados no Gene Expression Omnibus (GEO: https://www.ncbi.nlm.nih.gov/geo/ (GEO NCBI Database)), seja RNA-Seq ou Microarray. Isso implica que temos uma gama de grupos independentes analisando diversos tecidos/condições em TEA, o que abre a possibilidade de aplicar a análise de dados in silico para buscar responder quais são os genes chaves nos dados públicos de TEA e que possam ajudar a elucidar divergências e similaridades nesta síndrome. Por fim, a revisão realizada identificou apenas três artigos que aplicaram meta-análise em autismo, o que demonstra a oportunidade de investigação com esta técnica e nesta temática.

Por fim, este estudo teve como objetivo analisar conjuntos de dados de RNA-Seq que foram depositados no repositório público GEO (Gene Expression Omnibus), aplicando análise de DE (Expressão Diferencial) com intuito de identificar e selecionar genes correlatos ao autismo.

MATERIAIS E MÉTODOS

Conjunto de dados utilizado

Em resumo, a partir dos artigos selecionados (Tabela 1 e 2), os dados foram coletados no portal NCBI GEO em Dezembro/19, e separados em dados de expressão de RNA-Seq e tecido neural (Figura 1 - Parte 1). Sendo selecionados apenas dados de RNA-Seq, ao total 48 conjunto de dados, na sequência utilizado critério de seleção (*Homo sapiens*) e por fim aplicado o filtro de seleção para manter apenas dados, biológicos e tecido neural, sendo dados de IPSC. Resultando em três conjunto de dados para análise que foi realizada. A partir da extração dos dados de cada GSE, foram analisados os seguintes dados: (i) o GSE129808 tem-se sete amostras, sendo quatro controle e três TEA; (ii) GSE67528 tivemos de 83 amostras sendo 28 controles e 55 TEA; e o (iii) GSE61476 contendo 45 amostras de pacientes TEA e não-TEA.

ID GEO	PMID	Tecido	Amostras	Plataforma
GSE129808	31540669	IPSC Neurônios	7	Illumina HiSeq 2500
GSE67528	27378147	IPSC Neurônios	83	Illumina HiSeq 2000
GSE61476	26186191	IPSC Neurônios	45	Illumina HiSeq 2000

Tabela 1. Tabela do conjunto de dados de RNA-Seq utilizados na análise de expressão diferencial.

Workflow de análise dos dados de expressão

Na Workflow do pipeline é detalhado a metodologia que foi utilizada neste projeto, a partir dos dados de expressão obtidos na revisão da literatura sobre a TEA (Figura 1) . Os dados dos três estudos selecionados foram extraídos para o servidor da Bioinformática da UTFPR-CP do site do SRA utilizando o fastq-dump (NCBI SRA database), via programa Rstudio. O corte de qualidade e retirada dos adaptadores das sequências, foi realizado usando o Trimmomatic. Em seguida, os dados foram mapeados contra o genoma de referência (GRCh38) utilizando o Salmon. o por fim, o txtimport foi utilizado para a montagem da matriz de contagem. Para análise da expressão diferencial foi utilizado o pacote DEseq (Huber et al., 2015) no Rstudio. Ainda, foram feitos as anotações funcional dos genes diferencialmente expressos utilizando o banco de dados do Gene Ontology (GO), bem como a análise de PCA foi aplicada para compreender a inter-relação dos genes.

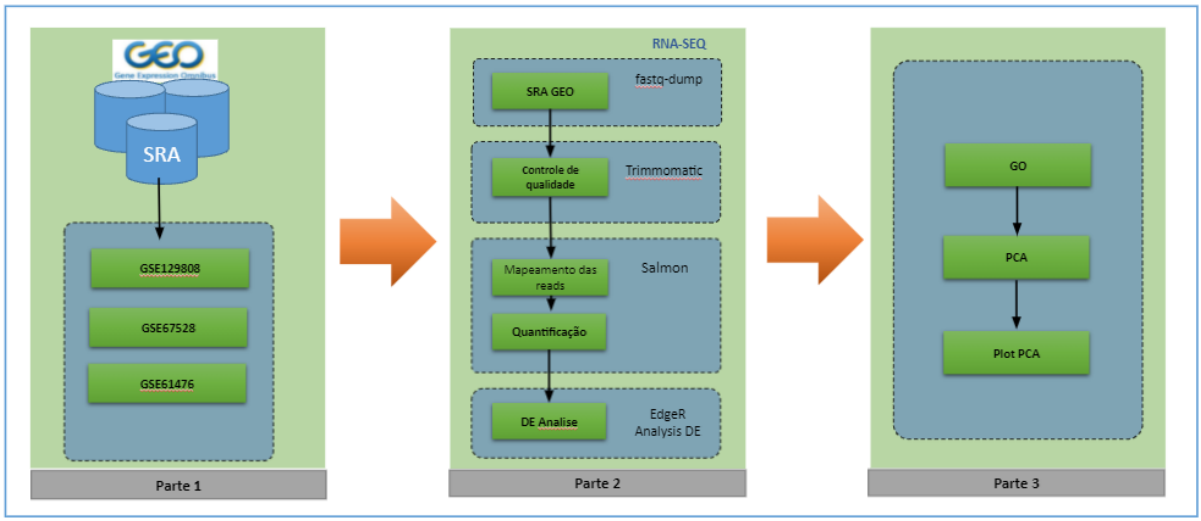


Figure 5 - Workflow Pipeline. O trabalho foi dividido em três fases de análise, sendo; Parte 1 - Download dos dados no repositório do NCBI, Parte 2 - Processamento e tratamento dos dados que foram baixados no repositório e por fim na, Parte 3 - Análise dos dados e plotagem dos dados de expressão.

Resultados

Após a seleção dos dados, quantificação, controle de qualidade das reads foram obtidos um total de 29374 transcritos de cDNA e ncRNA. Em seguida, estes transcritos foram analisados usando o pacote DEseq para identificar diferencialmente (mais ou menos) expressos em relação ao grupo controle. Nesse sentido, foi aplicado o vulcano plot para a visualização dessa dispersão dos dados de expressão (Figura 2).

O gráfico é composto pelo procedimento de duas etapas: primeiro, a alteração para base logarítmica devido a abundância do gene no grupo ASD para o grupo de controle, seguido por uma transformação log2 para obter uma distribuição normal ou quase normal. Valores maiores que 0 são considerados genes regulados positivamente, em quando o valor é menor que zero são regulados negativamente. Em segundo lugar, um valor de p ajustado (ou q-values), corrigido para múltiplas correções, é usado para calcular se a expressão do gene sofre alterações entre os grupos de ASD e controle são significativamente diferentes. Por fim, é realizada uma transformação do valor P ajustado para o formato -log10

No gráfico do vulcano plot foram configurados e definidos o limite de significância estatística, ou seja (pcutoff e Fccutoff) para que apenas os genes que ultrapassarem os limites definidos tenham significância estatisticamente. Sendo pCutoff = 0.05 e FCcutoff = 1.

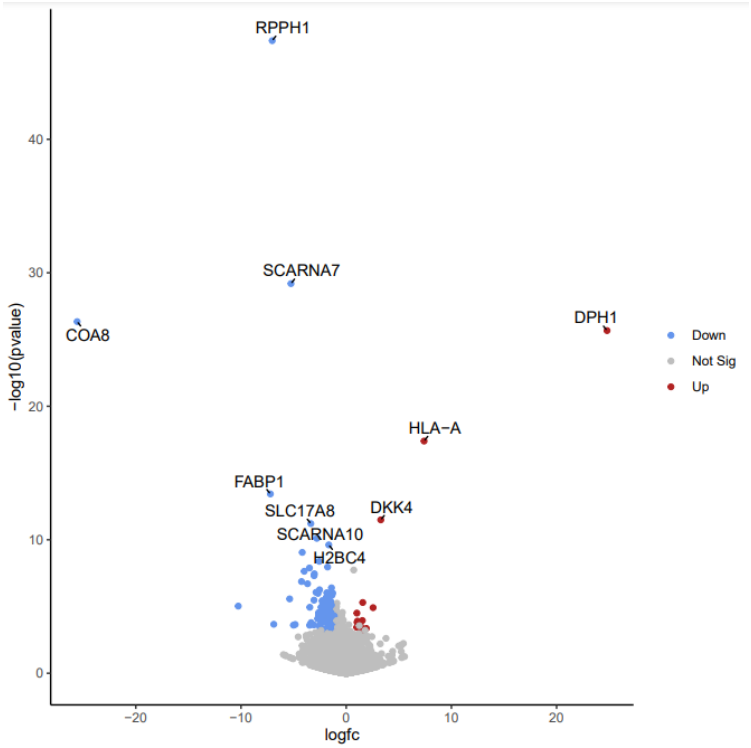


Figure 6 - Análise do Volcano Plot - O gráfico de vulcão apresenta os genes mais expressos, sendo três genes em vermelho, e sete genes menos expressos, ao total são dez genes que foram identificados como sendo os mais expressos entre a análise dos conjuntos de dados estudados.

Deste modo a partir deste limiar de corte foi possível identificar dez genes diferencialmente expressos (Figura 2), sendo que três genes (SACARNA7, SCARNA10, RPPH1) não-codificantes, e sete genes (DPH1, FABP1, SLC17A8, DKK4, HLA-A, H2BC4, COA8) são genes codificantes.

Análise Funcional Usando os Termos GO

A análise de enriquecimento de gene (GO) foi utilizada para avaliar o processo biológico dos genes e a localização celular da expressão gênica dos dados utilizados neste trabalho. O gráfico do enriquecido com o GO sugere que que vias que têm função de reconstituição das extremidades cromossômicas, e que estão relacionadas com a manutenção da capacidade de réplica estão mais expressas nos dados analisados neste trabalho (Figura 3).

Figure 7 - Análise do termo GO - As vias enriquecidas foram selecionadas de genes diferencialmente expressos, para os genes significativos, sendo 10 genes significativos de expressão diferencial. Os pontos em vermelho indicam alto enriquecimento, em azul indicam baixo enriquecimento. Finalmente, usamos o pacote clusterProfiler do R para comparar esses agrupamentos de genes por seus processos biológicos enriquecidos, com o corte estrito de valores de (p-adjust <0,01 e FDR <0,05) relacionados ao Autismo.

Análise PCA

A PCA com VST (Transformação de estabilização de variância) foi utilizada para verificar como as amostras se agrupam no projeto experimental. O gráfico dos três conjuntos de dados estudados (GSE129808, GSE67528 e GSE61476) foram separadas em dois grupos (Autismo e Controle). Como esperado, os dados analisados apresentam variação de expressão em genes relacionados ao autismo, sendo que o conjunto de dados GSE67528 apresentou uma maior variação

em relação aos demais genes ligados ao autismo. Nessa conjunto em particular (GSE67528) foi possível observar a separação em dois grupos, mas como ele foi o único a ter esse resultado, a análise comparativa via PCA esperada não foi possível de ser feita.A grande variação dos dados de (ASD e Controle), sugere que mais genes são expressos diferencialmente nestes conjunto de dados públicos que foram selecionados para análise. Para entender a variação genética apresentada pelo PCA, plotamos os dados em um gráfico de Volcano Plot, com intuito de identificar os genes expressos e estudar a funcionalidade de cada gene expresso.

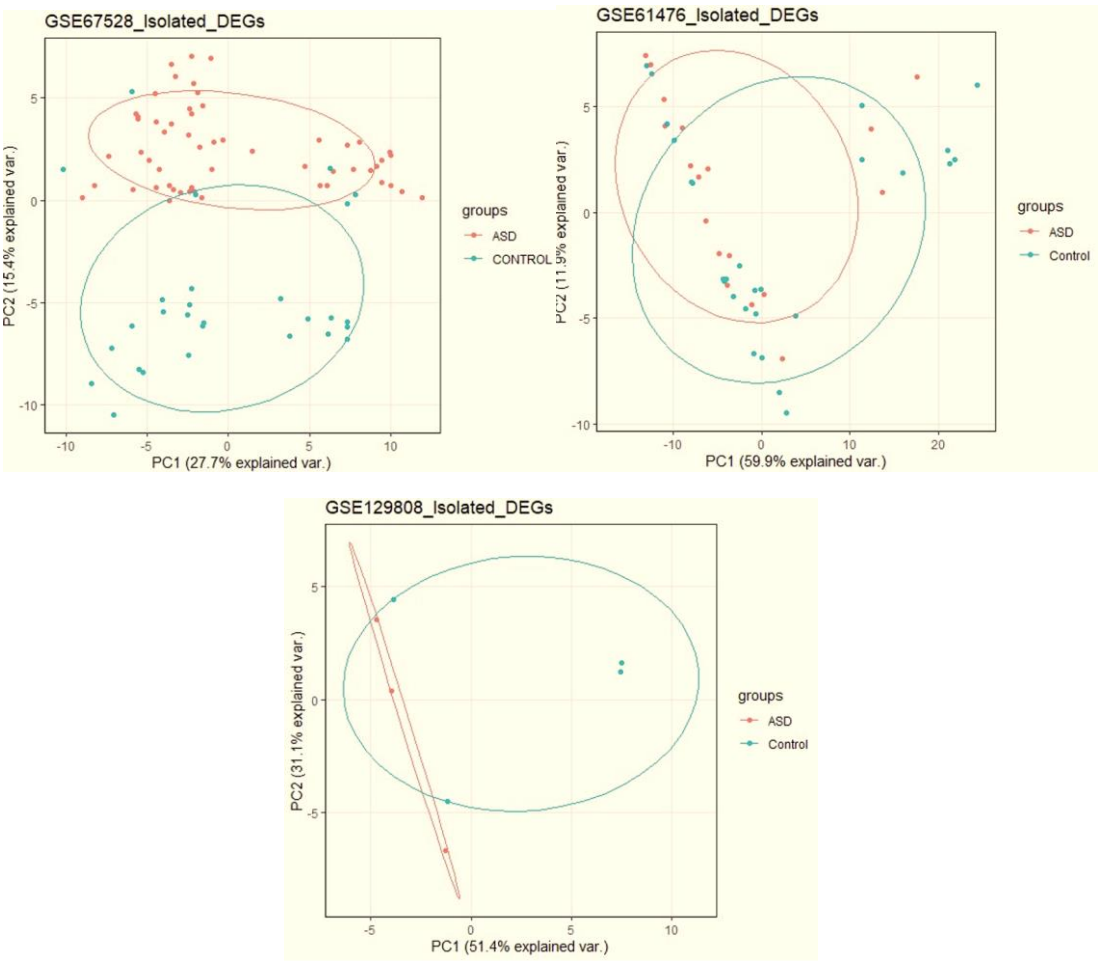


Figure 8 - PCA - Para cada conjunto de dados fizemos o gráfico exploratório dos dados estabilizados usando a análise de componentes principais (Figura 4). O que demonstrou que o conjunto de dados GSE67528 é o conjunto que apresenta maior variabilidade dos dados comparado com os outros dois conjuntos de dados utilizados neste experimento.

CONCLUSÃO

O trabalho foi capaz de identificar variações em dados disponíveis em repositório público conforme proposto. O conjunto de dados públicos utilizando a meta-análise apresenta variações para o entendimento do Autismo. A análise de PCA apresentou dispersão em apenas um conjunto de dados, dos três que foram utilizados, era esperado a possibilidade de ver isso em todos os conjuntos de dados, o que não foi o caso. No final, dez genes apresentaram expressão diferencial, sendo essa a grande contribuição do trabalho.

Na tabela 1 resume informações sobre os 10 genes.Fizemos uma busca rápida no repositório (disgenet.org) consultandos os três genes Up regulados, sendo os genes DKK4 e HLA-A que apresentaram associação à Esquizofrenia e o gene DPH1 está associado ao atraso no desenvolvimento e deficiência intelectual.

Dos dez genes identificados, sete genes codificantes de proteinas e três RNA não-codificantes, sendo três genes UP regulados (os genes DPH1 - DKK4 – HLA-A), enquanto que os outros setes foram Down regulados (H2BC4 - SCARNA10 - SLC17A8 - FABP1 - SCARNA7 - COA8 - RPPH1 – Tabela 2).

A próxima etapa futura é a análise mais refinada do papel regulatório, região gênica, e em particular dos RNAs não-codificadores.

Gene ID	Symbol	Gene name	Gene type	Scientific name	Ensembl Gene ID
1801	DPH1	diphthamide biosynthesis 1	protein-coding	Homo sapiens	ENSG00000108963
2168	FABP1	fatty acid binding protein 1	protein-coding	Homo sapiens	ENSG00000163586
246213	SLC17A8	solute carrier family 17 member 8	protein-coding	Homo sapiens	ENSG00000179520
27121	DKK4	dickkopf WNT signaling pathway inhibitor 4	protein-coding	Homo sapiens	ENSG00000104371
3105	HLA-A	major histocompatibility complex, class I, A	protein-coding	Homo sapiens	ENSG00000206503
677767	SCARNA7	small Cajal body-specific RNA 7	ncRNA	Homo sapiens	ENSG00000238741
692148	SCARNA10	small Cajal body-specific RNA 10	ncRNA	Homo sapiens	ENSG00000239002
8347	H2BC4	H2B clustered histone 4	protein-coding	Homo sapiens	ENSG00000180596
84334	COA8	cytochrome c oxidase assembly factor 8	protein-coding	Homo sapiens	ENSG00000256053
85495	RPPH1	ribonuclease P RNA component H1	ncRNA	Homo sapiens	ENSG00000277209

Tabela 2 Na tabela temos as informações referente aos genes expressos diferencialmente neste trabalho, sendo ao total dez genes, sete codificantes de proteínas e três não codificantes.

5 CONCLUSÃO FINAL DO TRABALHO

Este trabalho apresentou uma análise em cima de três conjuntos de dados públicos de expressão e publicados em artigos relacionados ao autismo. Este trabalho teve como objetivo comparar esses dados individuais numa só análise de modo a buscar os genes potencialmente chaves nesta condição estudada.

Foi aplicado análises tradicionais de RNA-Seq para obter os resultados de expressão, o qual foram analisados os dados de RNA-Seq, e foram identificados 11 genes diferencialmente expressos relacionados ao TEA.

Por fim, este estudo abre oportunidade para novas discussões sobre a interação dos genes mais expressos, sendo que eles apresentam ligação com outra doença neurológica e também demonstra que os dados públicos de RNA-Seq podem conter e apresentar novas descobertas quando utilizando a metodologia da meta-análise dos dados.

6 REFERÊNCIAS

1. Ornoy A, Weinstein-Fudim L, Ergaz Z. Prenatal factors associated with autism spectrum disorder (ASD). Reprod Toxicol. (2015) 56:155–69. doi: 10.1016/j.reprotox.2015.05.007

2. Posar A, Resca F, Visconti P. Autism according to diagnostic and statistical manual of mental disorders 5(th) edition: the need for further improvements. J Pediatr Neurosci. (2015) 10:146–8. doi: 10.4103/1817-1745.1 59195

3. Autism Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators and Centers for Disease Control and Prevention. Prevalence of autism spectrum disorders–autism and developmental disabilities monitoring network, 14 sites, United States, 2008. MMWR Surveill Summ. (2012) 61:1–19.

4. De La Torre-Ubieta L, Won H, Stein JL, Geschwind DH. Advancing the understanding of autism disease mechanisms through genetics. Nat Med. (2016) 22:345–61. doi: 10.1038/nm.4071

5.https://www.psychiatry.org/patients-families/autism/what-is-autism-spectrum-disorder

6. Christense, D. L., Baio, J., Braum, K. V. N., Bilder, D., Charles, J., Constantino, J. N. Yeargin-Allsopp, M (2016). Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years - Autism and Developmental Disabilities Monitoring Netwok, 11 Sites, United States, 2012. Morbidity and mortality Weekly Report. Surveillance Sumaries, 65(3), 1-23 https://doi.org/10.15585/mmwr.ss6503a1

7. Halfon N., Houtrow A., Larson K., Newacheck PW. The changing landscape of disability in childhood. Future Child. 2012;22:13–42.

8.Donovan, A. P. A., & Basson, M.A(2017). The neuroanatomy of autism -a developmental perspective. Journal of Anatomy, 230(1) 230, 4-15. https://doi.org/1.111/joa.12542

9.Takahashi, K., Tanabe, K., Oknuki, M., Narita, M., Ichisaka, T., Tomoda, K., & Yamanaka S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts By Denined Factors. Cell, 131(5), 861-872. https://doi.org/10/1016/j.cell.2007.11.0019

10. Han, G., Sun, J., Wang, J., Bai, Z., Song, F., & Lei, H (2014). Genomics in Neurilogica Disorders. Genomics, Proteomics and Bioinformatics, 12(4), 156-163. https://doi.org/10.1186/1471-2164-14-778

11. GLASS, G. Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, v. 5, n. 10,p. 3-8, 1976

12. EGGER, M.; SMITH, G.D. Meta-analysis: potentials and promise. British Journal of Medical, v.315, 1371-1374, 1997.

13. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014; 30(15): 2014-20.

14. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77.

15. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017. 14(4): 417-19.

16. Oksanen J, Guillaume FB, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Vegan: community ecology package. 2017; 1(2): 1-12.

17. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77.

18. Gavery MR, Roberts SB (2012) Characterizing short read sequencing for gene discovery and RNA-Seq analysis in Crassostrea gigas . Comp Biochem Phys D Genomics Proteomics 7: 94–99. 10.1016/j.cbd.2011.12.003

19. Zhang, S., Deng, L., Jia, Q., Huang, S., Gu, J., Zhou, F., Gao, M., Sun, X., Feng, C. and Fan, G. (2017) dbMDEGA: a database for meta-analysis of differentially expressed genes in autism spectrum disorder. BMC Bioinformatics, 18,494.

20. L. F. Ning, Y. Q. Yu, E. T. Guoji et al., “Meta-analysis of differentially expressed genes in autism based on gene expression data,” Genetics and Molecular Research, vol. 14, no. 1, pp. 2146–2155, 2015.

21.Ch'ng, C., Kwok, W., Rogic, S., & Pavlidis, P. (2015). Meta-Analysis of Gene Expression in Autism Spectrum Disorder. Autism Research.

22. De Rubeis, S., &Buxbaum, J. D. (2015). Genetics and genomics of autism spectrum disorder: embracing complexity. Human molecular genetics, 24(R1), R24-R31.

23. Need AC , Goldstein DB 2010 Whole genome association studies in complex diseases: where do we stand? Dialogues Clin Neurosci 12:37–46

24. J. M. Cubillos-Angulo, E. R. Fukutani, L. A. B. Cruz et al., “Systems biology analysis of publicly available transcriptomic data reveals a critical link between AKR1B10 gene expression, smoking and occurrence of lung cancer,” PLoS One, vol. 15, no. 2, Article ID e0222552, 2020.

25. Croft D, O’Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, et al. Reactome: a database of reactions, pathways and biological processes, Nucleic Acids Res., 2011, vol. 39 (pg. D691-D697)

26. Hall M, Frank E, Holmes G, et al. The WEKA data mining software: an update. SIGKDD Explor 2009;11(1):10–18.

27. Schena, M., et al. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science, 270, 467-470.

28. Rao, M.S., Van Vleet, T.R., Ciurlionis, R., Buck, W.R., Mittelstadt, S.W., Blomme, E.A.G., Liguori, M.J., 2019. Comparison of RNA-Seq and microarray gene expression platforms for the Toxicogenomic evaluation of liver from short-term rat toxicity studies. Front. Genet. 9.

29. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences. 2003;100(16):9440–9445. doi:10.1073/pnas.1530509100.

30. Rogic S, Pavlidis P. Meta-analysis of kindling-induced gene expression changes in the rat hippocampus. Front Neurosci. 2009;3:53. doi:10.3389/neuro.15.001.2009.

31. Ramoz N, Reichert JG, Smith CJ, Silverman JM, Bespalova IN, Davis KL, Buxbaum JD. Linkage and Association of the Mitochondrial Aspartate/Glutamate Carrier SLC25A12 Gene With Autism. American Journal of Psychiatry. 2004;161(4):662–669. doi:10.1176/appi.ajp.161.4.662.

32. Mistry M, Pavlidis P. A cross-laboratory comparison of expression profiling data from normal human postmortem brain. Neuroscience. 2010;167:384–95. doi:S0306-4522(10)00017-5 [pii] 10.1016/j.neuroscience.2010.01.016.

33. Lord C, Rutter M, Goode S, Heemsbergen J, Jordan H, Mawhood L, Schopler E. Autism diagnostic observation schedule: a standardized observation of communicative and social behavior. Journal of Autism and Developmental Disorders. 1989;19(2):185–212.

34. Hu VW, Sarachana T, Kim KS, Nguyen A, Kulkarni S, Steinberg ME, Lee NH. Gene expression profiling differentiates autism case-controls and phenotypic variants of autism spectrum disorders: evidence for circadian rhythm dysfunction in severe autism. Autism Research: Official Journal of the International Society for Autism Research. 2009;2(2):78–97. doi:10.1002/aur.73.

35. Hu Q, Kukull WA, Bressler SL, Gray MD, Cam JA, Larson EB, Deeb SS. The human FE65 gene: genomic structure and an intronic biallelic polymorphism associated with sporadic dementia of the Alzheimer type. Human Genetics. 1998;103(3):295–303.

36. Zilbovicius, Mônica, Meresse, Isabelle e Boddaert, NathalieAutismo: neuroimagem. Brazilian Journal of Psychiatry [online]. 2006, v. 28, suppl 1 [Acessado 12 Março 2022] , pp. s21-s28. Disponível em: <https://doi.org/10.1590/S1516-44462006000500004>. Epub 12 Jun 2006. ISSN 1809-452X. https://doi.org/10.1590/S1516-44462006000500004.

37. HAZEN, S. P.; WU, Y.; KREPS, J. A. Gene expression profiling of plant responses to abiotic stress. Funcional & Integrative Genomics, v. 3, p. 105–111, 2003

38. Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, et al. (August 2014). "A comparative study of techniques for differential expression analysis on RNA-Seq data". PLOS ONE. 9 (8): e103207.

39. Basic RNA-seq processing: unix tools and IGV. bio.lundberg, 2021. Disponivél em : http://bio.lundberg.gu.se/courses/vt13/rnaseq.html

40. Qual a diferença entre DNA e RNA. Sociencia, 2021. Disponivél em :https://socientifica.com.br/o-que-sao-o-dna-e-o-rna/.

34. SCHULZE, A.; DOWNWARD, J. Navigating gene expression using microarrays–a technology review. Nature cell biology, Nature Publishing Group, v. 3, n. 8, p. E190, 2001.