Levantamento da correlação molecular entre Doença de Alzheimer e Diabetes tipo 2 por estudos in sílico.

Palavras-Chave: Alzheimer, Diabetes tipo 2, Bioinformática.

Autores(as):

Jennifer Stefani Ribeiro Dos Santos– UNICAMP

MSc. Kaira Cristina Peralis Tomaz –UNICAMP

MSc. Rafaela Ferreira Pacheco – UNICAMP

Prof(ª). Dr(ª). Marcelo Mendes Brandão – UNICAMP

INTRODUÇÃO:

A doença de Alzheimer (DA) destaca-se como uma das desordens neurodegenerativas mais prevalentes mundialmente, com incidência em contínua elevação, reflexo do envelhecimento global da população. Caracteriza-se pela deposição progressiva de placas de beta-amiloide e emaranhados neurofibrilares no cérebro, culminando em perda neuronal e declínio cognitivo. No Brasil, cerca de 1,5 milhão de pessoas—predominantemente idosos a partir de 65 anos—são acometidas pela DA, com projeções que apontam para uma possível quadruplicação desse número nas próximas décadas (SBGG, 2023).

Paralelamente, observa-se no país um aumento expressivo na incidência de distúrbios metabólicos, notadamente o diabetes tipo 2 (DM2), que afeta parcela significativa da população adulta. O DM2 apresenta-se como uma condição metabólica multifatorial, caracterizada por resistência à insulina, hiperglicemia, hipertensão arterial, dislipidemia e obesidade abdominal (Saklayen, 2018; Fathi, 2018). Importante notar que essas comorbidades compartilham vias bioquímicas relacionadas, potencializando-se mutuamente em sua gravidade e consequências clínicas (Kahn et al., 2005), além de elevarem substancialmente o risco de complicações cardiovasculares e demais agravos à saúde (Brant et al., 2022).

Nos últimos anos, evidências científicas vêm ressaltando a complexa interação entre o DM2 e o desenvolvimento e progressão da DA, indicando uma relevante sobreposição epidemiológica entre essas doenças (Alzheimer's Association, 2022). Por exemplo, o receptor ativado pelo proliferador de peroxissoma gama (PPARG), um gene central no metabolismo da glicose e processos inflamatórios, surge nesse contexto: variantes genéticas no PPARG têm sido associadas ao aumento do risco de síndrome metabólica, DM2 e DA (CIRELLI, T. et al., 2023; SARHANGI,N. et al., 2020; HENEKA, M. T. et al., 2011).

Essas condições metabólicas ativam mecanismos críticos como estresse oxidativo, inflamação crônica e disfunção vascular, os quais são responsáveis pelo aumento na produção de espécies reativas de oxigênio, ativação de vias inflamatórias e alterações na microcirculação cerebral, contribuindo assim para o acúmulo de placas de beta-amiloide e a formação de emaranhados neurofibrilares, marcas patológicas da DA (Willette et al., 2015; Reaven, 1988; Kahn et al., 2005; de la Monte & Wands, 2008). Além disso, a resistência à insulina pode influenciar negativamente a função cerebral ao perturbar a sinalização de insulina e glicose, cruciais para a manutenção da função neuronal e plasticidade sináptica, emergindo como um contribuinte significativo para os processos neurodegenerativos observados na doença de Alzheimer (Craft, 2007; Alzheimer's Association, 2022).

Com base nesse panorama, este projeto tem como objetivo avaliar a força e significância das correlações entre variantes genéticas, buscando estabelecer relações causais entre tais variantes e o desenvolvimento dessas patologias. Almeja-se, assim, identificar biomarcadores genéticos úteis para diagnóstico, prognóstico e acompanhamento terapêutico, além de permitir a identificação de clusters de genes ou proteínas funcionalmente relacionados, contribuindo para a descoberta de novos alvos terapêuticos e o desenvolvimento de estratégias terapêuticas inovadoras e personalizadas. Particularmente, a identificação de mutações frameshift apresenta elevada relevância, já que tais alterações podem comprometer a estrutura das proteínas, resultando em perda funcional ou produção de proteínas aberrantes.

METODOLOGIA:

ANÁLISE COMPUTACIONAL E DISPONIBILIDADE DOS RESULTADOS

Todo o desenvolvimento deste projeto será conduzido em ambiente computacional, utilizando os sistemas e recursos instalados no Laboratório de Bioinformática e Sistemas (LaBIS) da Unicamp. Os dados gerados ao longo da pesquisa serão organizados e disponibilizados publicamente por meio do Repositório de Dados da Unicamp (REDU), garantindo acesso aberto e transparência dos resultados. Além disso, todos os scripts e pipelines desenvolvidos para as análises computacionais serão rigorosamente documentados e depositados no repositório institucional do LaBIS no GitHub (https://github.com/labis-unicamp), assegurando reprodutibilidade, rastreabilidade e contribuição para a comunidade científica.

AQUISIÇÃO DE DADOS

A etapa de aquisição de dados foi realizada a partir do banco de dados dbGaP (Database of Genotypes and Phenotypes - NCBI, 2023), reconhecido internacionalmente por sua relevância no arquivamento, curadoria e disseminação de grandes conjuntos de dados genômicos e fenotípicos. O dbGaP desempenha papel central na promoção do avanço científico ao disponibilizar informações integradas sobre genótipos e fenótipos para estudos epidemiológicos, de associação genômica ampla (GWAS), sequenciamento médico e investigações sobre a relação entre variantes genéticas e características clínicas.

Ressalta-se que todos os dados a serem utilizados neste projeto serão obtidos de coleções públicas e controladas do dbGaP, em estrita concordância com os princípios FAIR (Findable, Accessible, Interoperable, Reusable), o que garante que os conjuntos sejam facilmente localizáveis, acessíveis por protocolos padronizados, interoperáveis com diferentes plataformas e reusáveis em novos contextos científicos. Este compromisso com práticas de ciência aberta e integração de dados facilita a colaboração científica, a transparência nos processos e o aumento da reprodutibilidade dos resultados.

O LaBIS já possui autorização para acesso a projetos específicos, incluindo o T2D-GENES Multi-Ethnic Exome Sequencing Study: PGR, o Alzheimer's Disease Sequencing Project (ADSP), bem como os estudos The Mount Sinai Brain Bank (MSBB) e The Accelerating Medicines Partnership Alzheimer's Disease Diverse Cohorts Study (AMP-AD_DiverseCohorts), todos eles disponibilizados via dbGaP. O processo de acesso observa as diretrizes de segurança e confidencialidade estabelecidas pelo NCBI, preservando a integridade dos dados e a privacidade dos participantes dos estudos

ANOTAÇÃO DE VARIANTES COM ANNOVAR

Para a anotação funcional das variantes genéticas provenientes dos dados obtidos no dbGaP, será utilizado o software ANNOVAR, ferramenta amplamente reconhecida por sua eficiência e precisão na caracterização de variantes genômicas (Wang & Hakonarson, 2010). Considerando as informações estruturadas desses dados, tais como cromossomo, posições cromossômicas, nucleotídeos de referência e variantes observadas, o ANNOVAR oferece múltiplas camadas de anotação funcional que são essenciais para a interpretação biológica das variantes.

Dentre as funcionalidades do ANNOVAR que foram exploradas neste estudo, destacam-se três abordagens centrais: a anotação baseada em genes, que permitiu identificar se as variantes — como SNPs ou CNVs — provocam alterações em regiões codificadoras de proteínas, detalhando inclusive os aminoácidos envolvidos; a anotação centrada em regiões genômicas específicas, como promotores, elementos intrônicos ou regiões intergênicas, o que possibilita maior compreensão das possíveis repercussões regulatórias das variantes; e, por fim, a anotação por filtros, que viabiliza o cruzamento das variantes identificadas com bancos de dados especializados, facilitando a catalogação de variantes previamente documentadas.

O uso integrado dessas ferramentas será fundamental para se atingir os objetivos propostos neste projeto, especialmente no que diz respeito à identificação de biomarcadores genéticos, ao estabelecimento de relações causais entre variantes e doenças neurodegenerativas ou metabólicas, e à descoberta de potenciais novos alvos terapêuticos.

ANÁLISE ESTATÍSTICA

A metodologia de análise estatística adotada neste projeto está estruturada em etapas sistemáticas que visam garantir a robustez e a clareza na interpretação dos resultados. Inicialmente, os dados coletados são organizados em tabelas de contingência, permitindo sumarizar e visualizar de maneira eficiente a distribuição das variáveis de interesse, como a presença ou ausência das doenças analisadas. Essa abordagem facilita a identificação preliminar de padrões, tendências e potenciais associações entre as variáveis envolvidas (UCLA).

Na sequência, aplico o teste de Qui-Quadrado para avaliar a independência entre as variáveis categóricas mapeadas nas tabelas de contingência. Este teste estatístico é fundamental para determinar se existe associação significativa entre as doenças e variantes analisadas, fornecendo assim subsídios quantitativos para interpretações subsequentes (Khan Academy).

Os resultados obtidos orientam a análise da força e da significância das correlações identificadas entre genes e doenças, bem como a detecção de clusters de genes com funções relacionadas. Para ampliar a visualização dessas relações, utilizo ferramentas especializadas, como o VennMaker ou o BioVenn, capazes de gerar diagramas de Venn interativos. Esses diagramas ilustram de modo claro as interseções e distinções entre conjuntos de genes associados a cada doença, permitindo explorar a sobreposição funcional e biológica dos grupos analizados (Baur & Schönhuth, 2010; Hulsen & Alkema, 2008).

Por meio dessas análises detalhadas e contextualizadas, será possível avançar na compreensão dos mecanismos moleculares subjacentes aos processos patológicos abordados neste projeto.

RESULTADOS E DISCUSSÃO:

Os dados foram organizados em categorias específicas para análise e visualização, sendo separados em grupos de Alzheimer (AD), controle AD, Diabetes (DIA) e controle DIA. Importante destacar que os datasets Hong Kong No, BioMe No e Promis No não apresentavam dados de controle disponíveis. Assim, o universo amostral compreendeu 13 arquivos de diabetes e 23 arquivos de Alzheimer.

Após o processamento e curadoria dos dados, foram obtidos 6.773 participantes com diabetes positivo, totalizando 659 GB de dados, enquanto o grupo controle de diabetes apresentou 1.660 participantes com 115 GB. Para os dados de Alzheimer, identificaram-se 5.996 participantes positivos e 698 controles, correspondendo a 4,4 TB e 86 GB, respectivamente.

O processamento dos dados incluiu filtragem rigorosa quanto à qualidade das amostras, utilizando scripts desenvolvidos em bash. Os dados foram convertidos para formato texto (txt) para análise no ANNOVAR e VEP (Variant Effect Predictor). Paralelamente, scripts em Python permitiram a extração de informações sobre anotação estrutural dos genomas, incluindo cromossomo, posição, gene referência, gene alterado, genótipo do participante e ID para cada variante, todos referenciados à montagem do genoma humano HG19. Todos os scripts estão disponíveis em https://github.com/labis-unicamp.

A análise dos dados brutos revelou heterogeneidade considerável nas alterações genéticas identificadas. Observaram-se desde substituições simples de bases nitrogenadas (p.ex A para C) até alterações complexas envolvendo múltiplas bases, como substituição de A por sequências como G:T:C:C:T. Similarmente, os padrões alélicos variaram substancialmente do padrão de referência '1/1;0/0;0/1', apresentando variações como '0/7;6/1', refletindo diferenças na frequência e combinação de alelos entre os indivíduos analisados.

Análise de Variantes e Identificação de Genes

grafico_upset.png

(Imagem 1. Correlação das variantes de Alzheimer e diabetes tipo 2).

A análise por meio de gráficos UpSet revelou a distribuição das variantes identificadas: 3.616.701 variantes exclusivas em pacientes com Alzheimer, 1.487.419 exclusivas em diabetes, e 99.986 variantes compartilhadas entre ambas as condições na região de interseção. Essas variantes compartilhadas são particularmente relevantes para compreender os mecanismos moleculares comuns entre as duas patologias.

Utilizando o Variant Effect Predictor (VEP) (https://www.ensembl.org/Homo_sapiens/Tools/VEP/), foi possível identificar sete genes associados à interseção das doenças, 84 genes específicos para Alzheimer e 79 genes para diabetes. Esses genes foram posteriormente filtrados para remover variantes benignas e variantes étnicas não relacionadas aos objetivos do estudo, utilizando scripts Python customizados.

Análise de Ontologia Gênica

A interpretação funcional dos genes identificados, foi realizada em três níveis da ontologia gênica: processo biológico, componente celular e função molecular. A visualização dessas anotações foi facilitada pelo uso do REVIGO(http://revigo.irb.hr/) ,

ferramenta que permite sumarizar e visualizar listas extensas de termos de Gene Ontology, reduzindo redundâncias e destacando as principais categorias funcionais afetadas.

Os resultados das análises de ontologia gênica, demonstraram padrões distintos de envolvimento funcional para cada condição. Para diabetes, observaram-se enriquecimentos em processos relacionados ao metabolismo de glicose e resposta inflamatória. Para Alzheimer, destacaram-se processos neurodegenerativos e de morte celular. Importante notar que a análise da interseção revelou funções compartilhadas relacionadas a estresse oxidativo e disfunção mitocondrial, corroborando a literatura sobre os mecanismos comuns entre essas patologias.

CONCLUSÕES:

Este estudo demonstrou a aplicabilidade e relevância da análise genômica integrada na investigação dos mecanismos moleculares que conectam o diabetes tipo 2 e a doença de Alzheimer. A análise abrangente das variantes genéticas revelou não apenas a existência de 99.986 variantes compartilhadas entre ambas as patologias, mas também a identificação de sete genes na região de interseção, evidenciando bases moleculares comuns entre essas condições neurodegenerativas e metabólicas.

A diversidade genética observada, desde substituições simples de nucleotídeos até alterações complexas envolvendo múltiplas bases, reflete a complexidade dos mecanismos patológicos subjacentes. Importante destacar que o splicing alternativo representa um mecanismo fundamental que permite a um único gene gerar múltiplas isoformas proteicas, expandindo significativamente a diversidade do proteoma. Esta diversidade proteica, resultante do processamento diferencial de éxons, pode influenciar diretamente a função celular e contribuir para a susceptibilidade a doenças, conforme evidenciado pela literatura científica atual.

A utilização do REVIGO para análise de ontologia gênica permitiu mapear as principais vias funcionais afetadas, revelando enriquecimentos em processos relacionados ao estresse oxidativo, disfunção mitocondrial e inflamação crônica—mecanismos centrais na patogênese tanto do diabetes quanto do Alzheimer. Esses achados corroboram estudos prévios que demonstram a interconexão metabólica entre essas doenças, particularmente através de vias relacionadas à resistência à insulina e à neurodegeneração.

A metodologia empregada, incluindo o uso de ferramentas como bcftools para processamento de variantes e análises de ontologia gênica, estabeleceu um framework robusto para identificação de biomarcadores genéticos potenciais. Os resultados obtidos fornecem subsídios importantes para futuras investigações direcionadas à descoberta de alvos terapêuticos e ao desenvolvimento de estratégias de medicina personalizada, contribuindo para o avanço do conhecimento sobre a base molecular das doenças neurodegenerativas associadas a distúrbios metabólicos.

Para uma análise detalhada e abrangente, é imprescindível considerar todas as variantes genéticas contidas nos arquivos, e, por vezes, explorar o splicing alternativo que visa explorar os mesmo exons, mas são unidos em diversas combinações diferentes que, por sua vez contribui para diferentes formações de proteínas, por meio de um único gene. Dessa forma, talvez podendo comparar as variantes filtradas no projeto e ampliar o para encontrar novos marcadores genéticos. As informações adquiridas, embora inicialmente centradas em um conjunto restrito a seleção do VEP, são significativas para um primeiro parecer e podem oferecer diretrizes fundamentais para o direcionamento da investigação a ser empreendida.

A análise de variantes genéticas é um processo complexo que demanda a avaliação de um grande volume de dados. Inicialmente, a análise dos dados brutos revelou a existência de um número significativo de variantes, apresentando desafios em relação ao processamento e à interpretação dos resultados.

Nesse sentido, o Revigo (http://revigo.irb.hr/) foi importante para obtivemos um mapa das funções das vias genéticas onde as variantes encontradas estão envolvidas e assim, classificando genes e produtos gênicos em termos de suas funções biológicas, processos celulares e localizações celulares. Os quais se relacionam com a literatura para evidenciar de forma geral quais são as principais áreas afetadas do genes com alteração associadas ao Alzheimer e a diabetes tipo 2.

Vale a pena ressaltar que o processo de junção dos dados estão sendo otimizados com a ferramenta bcftools (https://samtools.github.io/bcftools/bcftools.html) a qual pretendemos utilizá-la também para a obtenção da função alélica garantindo uma melhor acurácia dos dados. Finalizando esta etapa, daremos prosseguimento ao restante do projeto.

BIBLIOGRAFIA

- Alzheimer's Association. (2022). Alzheimer's Disease Facts and Figures. Alzheimer's & Dementia, 18(4).
- Baur, M., & Schönhuth, M. (2010). VennMaker for Historians: Sources, Social Networks and Software. Revista Redes, 17, 377-403.
- Brant, L.C.C. et al. (2022). Burden of Cardiovascular diseases attributable to risk factors in Brazil: data from the "Global Burden of Disease 2019" study. Rev Soc Bras Med Trop. 28;55 (suppl 1):e0263. DOI: 10.1590/0037-8682-0263-2021. PMID: 35107526; PMCID: PMC9009428.
- Butterfield, D. A., & Halliwell, B. (2019). Oxidative stress, dysfunctional glucose metabolism and Alzheimer disease. Nature Reviews Neuroscience, 20, 148-160. DOI: 10.1038/s41583-019-0132-6
- Cirelli, T., et al. (2023). "Evidence Linking PPARG Genetic Variants with Periodontitis and Type 2 Diabetes Mellitus in a Brazilian Population" International Journal of Molecular Sciences 24, no. 7: 6760. https://doi.org/10.3390/ijms24076760
- Craft, S. (2007). Insulin resistance and Alzheimer's disease pathogenesis: Potential mechanisms and implications for treatment. Current Alzheimer Research, 4(2), 147-152.
- dbGaP/database of Genotypes and Phenotypes/ National Center for Biotechnology Information, National Library of Medicine
- REVIGO, Reduce + visualize Gene Ontology; acesso em 26 de fevereiro de 2025; disponível em:
- Centro Nacional de Processamento de Alto Desempenho em São Paulo acesso em 26 de fevereiro de 2025: disponível em: <http://revigo.irb.hr/ >
- VEP, Variant Effect predictor acesso em 26 de fevereiro de 2025; disponível em: <https://www.ensembl.org/Homo_sapiens/Tools/VR?tl=k2dBMwVpFcRHNgnD-10806356>
- dbGAP, acesso em 16 de março de 2025; disponível em: <https://www.ncbi.nlm.nih.gov/gap/>
- NCBI, acesso em 16 de março de 2025; disponível em: <https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.25/>
- BCFTOOLS, acesso em 16 de março de 2025; disponível em: <https://samtools.github.io/bcftools/bcftools.html>
- labis-unicamp: repositório Project-genetic-relationship-between-alzheimer-s-and-type-2-diabetes , acesso em 17 de março de 2025; disponível em: <https://github.com/labis-unicamp/Project-genetic-relationship-between-alzheimer-s-and-type-2-diabetes >
- STRING, acesso em 16 de março de 2025; disponível em: <https://string-db.org/ >