UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS E NATUREZA
BACHARELADO EM CIÊNCIAS BIOLÓGICAS

ANNA BEATTRIZ MARQUES CARVALHO

Doença de Parkinson: análise exploratória de dados disponíveis no Gene Expression Omnibus (GEO)

JOÃO PESSOA - PB
2021

ANNA BEATTRIZ MARQUES CARVALHO

Doença de Parkinson: análise exploratória de dados disponíveis no Gene Expression Omnibus (GEO)

Trabalho de conclusão de curso apresentado ao curso de Bacharelado em Ciências Biológicas do Centro de Ciências Exatas e Natureza, da Universidade Federal da Paraíba, como requisito para a obtenção do grau de Bacharel em Ciências Biológicas.

Orientador: Prof. Dr. Sávio Torres de Farias
Co-orientador:Prof.Dr.Vinicius Maracaja-Coutinho

JOÃO PESSOA - PB

2021

Catalogação na publicação
Seção de Catalogação e Classificação

C331d Carvalho, Anna Beattriz Marques.
Doença de Parkinson : análise exploratória de dados disponíveis no Gene Expression Omnibus (GEO) / Anna Beattriz Marques Carvalho. - João Pessoa, 2021.
54 f. : il.

Orientação: Sávio Torres de Farias.
Vinicius Maracaja-Coutinho.
TCC (Graduação/Bacharelado em Ciências Biológicas) - UFPB/CCEN.

1. Doença de Parkinson. 2. Bioinformática. 3. MicroRNAs. 4. Neurônios dopaminérgicos. I. Farias, Sávio Torres de. II. Maracaja-Coutinho, Vinicius. III. Título.

UFPB/CCEN CDU 616.858(043.2)

Elaborado por Josélia Maria Oliveira da Silva - CRB-15/113

FOLHA DE APROVAÇÃO

Anna Beattriz Marques Carvalho

Doença de Parkinson: análise exploratória de dados disponíveis no Gene Expression Omnibus (GEO).

Trabalho de conclusão de curso apresentado ao Curso de Ciências Biológicas do Departamento de Biologia Molecular, do Centro de Ciências Exatas e da Natureza, da Universidade Federal da Paraíba, como requisito para obtenção do título de Bacharelado.

Aprovado em: __22_____/___julho__ ____/_____2021_____.

Banca Examinadora

_____

Prof. Doutor Savio Torres de Farias
Orientador
(UFPB/CCEN/Departamento de Biologia Molecular)

_____

Profa. Dra. Thaís Gaudencio do Rêgo
Membro Interno
(UFPB/CI/Departamento de Informática)

_____

Ma. Thaís de Almeida Ratis Ramos
Membro Externo
(UFRN)

Dedico este trabalho à minha família, meu maior bem.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por tudo que vive e respira, todos os fenômenos da natureza e pela vida e capacidade cognitiva de contemplá-los. Segundamente a minha família, razão de todo esforço, por todo apoio e incentivo nos meus melhores e piores momentos, especialmente minha mãe, a quem devo tudo e seria incapaz de expressar em palavras o quanto. Também à minha irmã, por ter me ajudado neste presente trabalho nas horas de desespero, ansiedade e doenças.

Agradeço também aos meus amigos e irmãos de crisma pelo companheirismo, ensinamentos e constante motivação e inspiração na vida humana, de estudos e profissional. Aos meus colegas e amigos de caminhada na trajetória da biologia, assim como aos colegas do laboratório do ARIA, pela constante ajuda e aprendizado neste universo multidisciplinar da bioinfo, ao qual eu serei eternamente grata pelo aprendizado, em especial aos meus primos Annie e Flávio, sem os quais nenhum trabalho com informática seria possível.

Um especial agradecimento à minha orientadora quase angélica Thaís, por me receber nesse laboratório, mesmo sem prévia bagagem em informática e me proporcionar uma das maiores e mais importantes experiências acadêmicas. Por último, ao magnânimo Hans Zimmer, cujas obras de arte me ajudaram a enfrentar não só a pandemia, mas a edição de cada tupla e escrita de cada página deste trabalho.

“Mas, cidadãos atenienses, parece-me que também os artífices tinham o mesmo defeito dos poetas: pelo fato de exercitar bem a própria arte, cada um pretendia ser sapientíssimo também nas outras coisas de maior importância, e esse erro obscurecia o seu saber."

Sócrates

RESUMO

A Doença de Parkinson (DP) é a segunda doença neurodegenerativa de maior prevalência na população humana, com maior incidência em indivíduos do sexo masculino. Apesar dos avanços diagnósticos e terapêuticos, ainda há divergências na literatura, sobre a classificação e diferenciação clínica da doença,além da multifatorialidade dos fatores que contribuem para esta doença, cuja classificação de “idiopática” tem sido desafiada em trabalhos recentes, que também contribuem com os inúmeros desafios tangentes ao desenvolvimento de potenciais terapias. Com o envelhecimento da população mundial, doenças como essa têm sido foco de inúmeros tipos de pesquisas. Aqui, nós fizemos uma análise exploratória da base de dados pública integrada ao GEO (Gene Expression Omnibus), chamada BioProject,na qual contém dados “ômicos” referentes à Doença de Parkinson, de animais modelos a pacientes humanos,. Análises gráficas com geração de diagramas em linguagem Python (Versão 3.7.4) e com ferramentas do google sheets foram utilizadas para visualização dos principais dados ômicos, assim como a descrição das demais informações relevantes para pesquisas relacionadas à DP. Deste modo, foram encontradas as principais espécies, tecidos e moléculas utilizadas em 365 projetos. Foi encontrada uma maior frequência de projetos com a espécie humana, com tecidos nervosos e com moléculas de RNA (sendo a extração mais comum de RNA total).Além disso, também foram levantadas informações sobre o sexo, idade e etnia dos organismos, e seus possíveis impactos também serão discutidos neste trabalho.Por fim concluisse que a prevalência de dados oriundos de espécimes humanas e roedores não variou muito do encontrado em trabalhos anteriores, assim como as principais tecnologias e tecidos em geral,bem como uma falta de informação sobre gênero, iade e etnia que somados a baixa quantidade de amostras por condição em alguns projetos e falta de padronização em nomenclaturas tem impacto na qualidade e reprodutibilidade correta dos dados bem como na perspectiva de contribuição para futuras análises.

Palavras-chaves: Doença de Parkinson; Bioinformática; microRNAs; Neurônios dopaminérgicos

ABSTRACT

Parkinson's Disease (PD) is the second most prevalent neurodegenerative disease in the human population, with the highest incidence in males. Despite diagnostic and therapeutic advances, there are still divergences in the literature on the clinical classification and differentiation of the disease, in addition to the multifactorial nature of the factors that contribute to this disease, whose classification as "idiopathic" has been challenged in recent studies, which also contribute to the numerous challenges involved in the development of potential therapies. With the aging of the world population, diseases like this one have been the focus of countless types of research. Here, we did an exploratory analysis of the public database integrated with GEO (Gene Expression Omnibus), called BioProject, which contains “omic” data regarding Parkinson's Disease, from animal models to human patients. Graphical analysis with diagram generation in Python language (Version 3.7.4) and with google sheets tools was used to visualize the main omic data, as well as the description of other relevant information for researches related to PD. Thus, the main species, tissues, and molecules used in 365 projects were found. A higher frequency of projects with the human species was found, with nervous tissues and with RNA molecules (the most common extraction being total RNA). In addition, information was also collected on the sex, age, and ethnicity of the organisms, and their possible impacts will also be discussed in this work. Finally, it was concluded that the prevalence of data from human and rodent specimens did not vary much from what was found in previous works, as well as the main technologies and tissues in general, as well as a lack of information on gender, age, and ethnicity that, added to the low amount of samples per condition in some projects and the lack of standardization in nomenclatures, has an impact on the quality and correct reproducibility of the data, as well as on the perspective of contribution to future analyses.

Keywords: Parkinson disease; Bioinformatics; microRNAs; Dopaminergic neurons; omics datasets

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

DP- Doença de Parkinson

DA- Neurônios dopaminérgicos

SRA -The Sequence Read Archive

GWAS- Genome-wide association study

NSCs - Neural stem cells

hESCs - Human embryonic stem cells

iPSC- Induced pluripotent stem cells

NCBI - National Center for Biotechnology Information

INSDC - International Nucleotide SequenceDatabase Consortium

DDBJ - DNA DataBank do Japão

EMBL- European Molecular Biology Laboratory

GEO - Gene Expression Omnibus

SRA -The Sequence Read Archive

PMID - PubMed Unique Identifier

SUMÁRIO:

1 INTRODUÇÃO

Em meados de 1817, James Parkinson em seu ensaio "Paralisia agitante" (*do inglês, “Essay on the shaking palsy”*) descreveu uma síndrome comumente referida como 'doença de Parkinson' (DP). A síndrome é caracterizada pelos processos de tremor, bradicinesia, rigidez e instabilidade postural, além de uma variedade de outros sintomas motores e não motores (JANKOVIC, 2008).

A DP é uma das doenças relacionadas ao avanço da idade, ao envelhecimento e ao aumento da expectativa de vida da população global, por isso, vem cada vez mais chamando a atenção da comunidade científica.Com os avanços tecnológicos, computacionais e das técnicas genéticas e de estudos populacionais, que incluem desde estudos de associação ampla do genoma (do inglês *Genome-wide association study* - GWAS) à estudos epigenômicos e proteômicos, cerca de 20 formas monogênicas de DP foram descritas e mais de 100 loci foram identificados como fatores de risco para DP (JANKOVIC e TAN, 2020),

Para além destes fatos, diversos estudos e informações amostrais têm sido disponibilizados em diferentes bancos de dados pela comunidade científica ao redor do mundo. A quantidade e disponibilidade desses dados tem se mostrado simultaneamente um desafio e uma oportunidade no desenvolvimento de pesquisas científicas. Hoje, há uma grande variedade de bancos de dados que proporcionam informações e dados brutos de genes, proteínas e pequenas moléculas, além de ferramentas que ajudam na mineração e análise dos mesmos,como a ferramenta GEO2R, uma aplicação web baseada em R que ajuda os usuários a analisarem os dados do GEO descrita em Barrett *et al.* (2012) .A evolução constante de recursos computacionais, bem como sua popularização, facilidade de acesso e disponibilidade de recursos dentro da internet podem estar entre as grandes responsáveis pelo crescente volume de informação, tanto na web, quanto nas organizações, proporcionando uma nova dinâmica no processo de coleta e análise de dados.

Nesse contexto, entender a variabilidade, características e disponibilidade dos tipos de dados, relacionados a DP, disponíveis em banco de dados genômicos,

constituem uma poderosa ferramenta para compreensão e atualização sobre novos mecanismos moleculares e patológicos, que podem ser integrados e visualizados usando várias meta-análises e abordagens estatísticas.

Um dos objetivos de estudos como esse, é levantar informações que possam fornecer novas pistas para a compreensão de diferentes aspectos da DP, incluindo o potencial para auxiliar na identificação de vias metabólicas e biomarcadores. Sendo assim, à partir do acervo de conjuntos de dados ômicos disponíveis, foi realizada uma análise exploratória de um conjunto de dados de 365 projetos relacionados desde epigenômica, genômica, e transcriptômica à proteômica.

2 OBJETIVOS

2.1 Objetivo Geral

Realizar uma análise exploratória das informações relevantes acessíveis (tais como identificadores de projetos,componentes moleculares e tecidos) sobre a Doença de Parkinson no banco de dados “GEO DataSets” do “National Center for Biotechnology Information”.

2.2 Objetivos Específicos

Realizar extração das informações sobre o número de projetos encontrados no GEO sobre Doença de Parkinson (DP), que foram publicados, assim como o número de projetos que contém dados de alta qualidade (isto é, projeto com pelo menos 3 amostras para cada condição estudada).Identificar as espécies mais utilizadas em projetos com DP e também os tipos de dados e tecnologias mais utilizadas nesses projetos.Revelar os principais tecidos encontrados nos projetos com DP, explorando as 3 principais espécies (Homo sapiens, Mus musculus e Rattus norvegicus) presentes nos estudos com DP em relação aos principais tecidos encontrados, tipos moleculares, e presença de informações como gênero, idade e etnia.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Doença de Parkinson

A Doença de Parkinson (DP) é uma doença neurodegenerativa progressiva, caracterizada pelo aparecimento de sintomas como tremores, rigidez, bradicinesia e, em certos casos, de instabilidade postural. Além disso, também está associada a outras disfunções não-motoras, como distúrbios do humor, disfunção do sono, déficits cognitivos, demência e outros sintomas neuropsiquiátricos (Figura 1) (JANKOVIC e TAN, 2020).

Os principais mecanismos patogênicos moleculares desta doença incluem desdobramento e agregação de α-sinucleína, disfunção mitocondrial, comprometimento da depuração de proteínas (associada com deficiência de ubiquitina-proteassoma e sistemas de autofagia-lisossomal), neuroinflamação e estresse oxidativo. O envolvimento das vias dopaminérgicas, bem como noradrenérgicas, glutamatérgicas, serotonérgicas e da adenosina fornecem percepções sobre a rica e variável fenomenologia clínica associada à DP e a possibilidade de abordagens terapêuticas alternativas, além das terapias tradicionais de reposição de dopamina (JANKOVIC e TAN, 2020).

A patologia da Doença de Parkinson se caracteriza pela perda de inervação dopaminérgica da porção do mesencéfalo, conhecida como substância nigra, muito embora a neurodegeneração não se restrinja a mesma, mas envolve células localizadas em outras regiões da rede neural. Enquanto a síndrome clínica era inicialmente atribuída a essa disfunção gânglio-basal, estudos em animais modelo e paciente pós-morte têm mostrado o envolvimento de neurônios não-dopaminérgicos de outras regiões do cérebro. (STOKER, 2018).

Hoje, já se reconhece o envolvimento de vias não-dopaminérgicas na evolução dos sintomas não motores, reconhecidos clinicamente como fatores de impacto na qualidade de vida de pacientes com DP, porém, os critérios diagnósticos atuais ainda são em grande parte baseados nos critérios do *UK Parkinson's Disease Society Brain Bank*,que se concentram na sintomatologia motora e excluem os sintomas não motores (JANKOVIC e TAN, 2020).

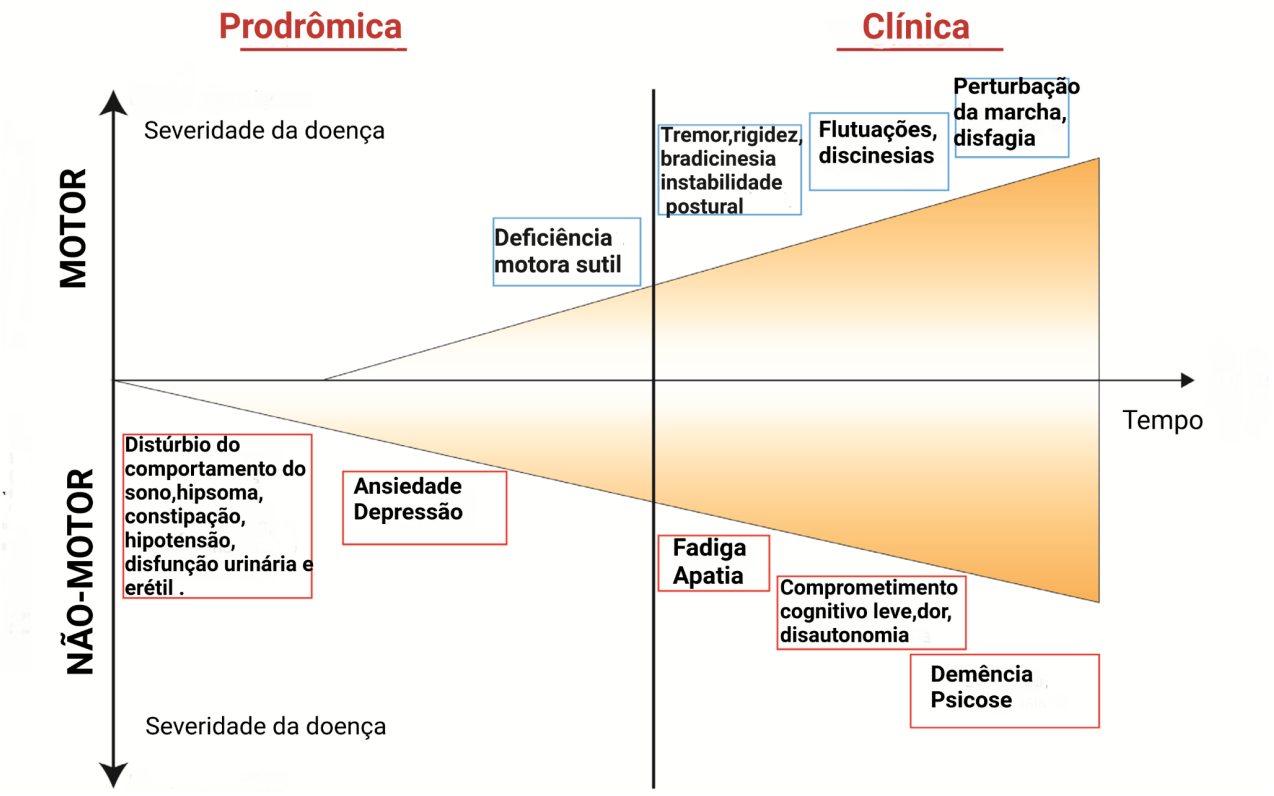


Figura 1: Curso da DP da fase prodrômica (fase anterior aos sintomas motores, mas que já apresenta neurodegeneração) à fase clínica.

FONTE: JANKOVIC e TAN (2020)

A DP é a segunda doença neurodegenerativa de maior prevalência na população humana (cerca de 0,5-1% entre aqueles com 65-69 anos e 1-3% entre aqueles com 80 anos ou mais) e, com o envelhecimento da população, é esperado um aumento gradativo nos próximos 20 anos na prevalência e índices de DP em mais de 30%, uma vez que o avanço da idade é um dos fatores que contribuem para o progresso da doença (STOKER, 2018).

3.1.1 Envelhecimento

O envelhecimento está associado a diversos processos moleculares que promovem danos genéticos e celulares, não obstante também está entre um dos

maiores fatores de risco a doenças neurodegenerativas, como Parkinson. Já houveram demonstrações em estudos anteriores demonstrando que, a quantidade de neurônios se mantém relativamente estável ao longo do processo de envelhecimento em diferentes áreas do cérebro, tais como: hipocampo, putâmen, núcleo mamilar medial, hipotálamo e núcleo basal de Meynert (PAKKENBERG e GUNDERSEN, 1997).

Nas populações neuronais dopaminérgicas, a perda celular é significativamente maior, podendo chegar a 50%, mostrando uma maior vulnerabilidade à perda neuronal comparada às demais regiões cerebrais, o que pode estar associado aos inúmeros processos primordiais para o funcionamento regular dos neurônios da substância nigra, incluindo o metabolismo da dopamina (Figura 2), o número de cópias de DNA mitocondrial de tipo selvagem e o declínio da degradação de proteínas que são afetados com o passar dos anos (REEVE, SIMCOX e TURNBULL, 2014).

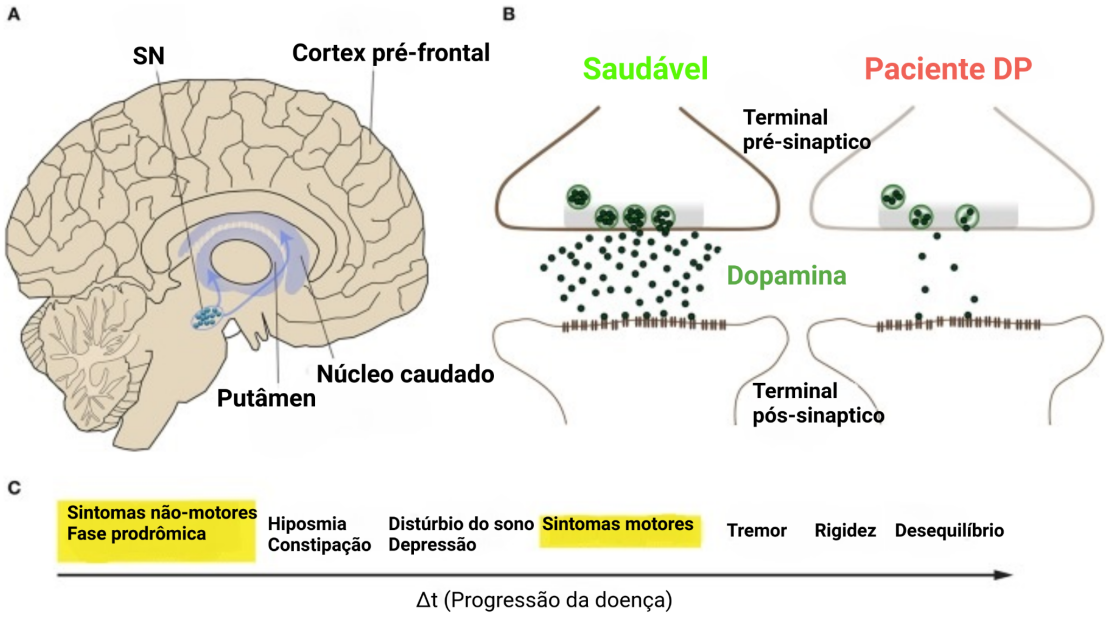


Figura 2:Os neurônios dopaminérgicos do mesencéfalo são especificamente vulneráveis na doença de Parkinson.(A) Os sintomas predominantes da doença de Parkinson (DP) são causados pela perda de neurônios dopaminérgicos (DA) na substância negra (SN), de acordo com a hipótese

em que a degeneração dos neurônios DA é precedida por disfunção e, por sua vez, degeneração da via nigroestriatal, que inerva o núcleo caudado e o putâmen que juntos formam o estriado. (B) Em comparação com controles saudáveis (esquerda), a degeneração nigroestriatal resulta na depleção e perda final do neurotransmissor dopamina nos terminais sinápticos dos neurônios do estriado (direita). (C) Os sintomas motores resultantes, entre outros, geralmente são diagnosticados quando aproximadamente 30-60% dos neurônios estriados DA já estão perdidos.

Fonte: BRIDI e HIRTH (2018)

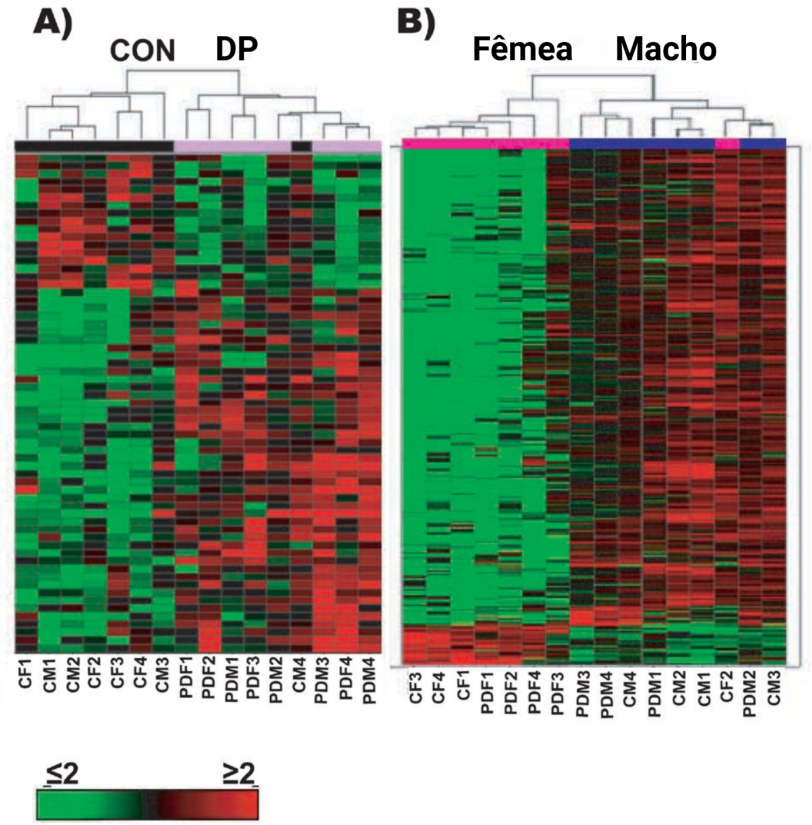
Em animais modelos, tais como macacos e ratos, uma disparidade no gênero pôde ser observada, com animais fêmeas mostrando maior número de células dopaminérgicas na substância nigra, enquanto embriões de ambos ratos machos e fêmeas, cultivados em ambientes hormonais idênticos, mostraram dimorfismo relacionado ao gênero em seus neurônios dopaminérgicos, levando à crescente discussão sobre o impacto dos efeitos do gênero nas doenças neurodegenerativas (CANTUTI-CASTELVETRI *et al.*, 2007).

3.1.2 Diferenças de gênero

Através de diversas pesquisas, inclusive epidemiológicas, o fator que mais vem sendo apontado, recentemente, como fator de susceptibilidade ao risco de desenvolvimento da DP, depois da idade, é o gênero, sendo o masculino o mais propenso. Relatórios de proporção na incidência de DP entre populações masculinas e femininas têm taxas variando de 1,37 a 3,7 (TAYLOR, COOK e COUNSELL, 2007).

O trabalho de CANTUTI-CASTELVETRI et al., 2007 mostrou impactos relativos ao sexo no padrão de expressão para diversos genes em neurônios dopaminérgicos, assim como o efeito da DP nesses mesmos neurônios.Neste trabalho, foram identificados conjuntos de cerca de 120 genes supra regulados nas mulheres em relação aos homens, e aproximadamente mais de 2.000 genes supra regulados nos homens, em comparação às mulheres (Figura 3). Não surpreendentemente, os genes com a maior diferença de magnitude na expressão estavam ligados aos cromossomos X e Y, muito embora a maioria dos genes

especificamente relacionados com o fator de gênero identificados, não estejam ligados a cromossomos sexuais e pouco implicados em estudos à priori, em diferenças de gênero no cérebro humano (CANTUTI-CASTELVETRI *et al.*, 2007).



w

Figura 3: Agrupamento hierárquico em amostras normalizadas por medianas usando correlação de cosseno com ligação completa realizado em todas as amostras para determinar o agrupamento de genes e para melhor visualizar diferenças nos perfis de expressão entre os pacientes com DP e os controles saudáveis, independentemente de seu gênero (painel A, Doença); e entre os sujeitos femininos e masculinos independentemente do seu estado de doença (painel B, Sexo).

Fonte: CANTUTI-CASTELVETRI *et al.* (2007)

Outro grande fator de impacto nessa discrepância são os hormônios gonadais, uma vez que, além de desempenhar um papel de suma importância no desenvolvimento estrutural e funcional do cérebro, tem se mostrado objeto de estudo para compreensão de seu papel em doenças neurodegenerativas e, até

mesmo, com recentes utilizações de terapias baseadas em estrógeno, para alívio dos sintomas iniciais da DP (GILLIES *et al.*, 2014).

O impacto dos hormônios gonadais vem sendo verificado ainda em estudos de casos clínicos em pacientes, como nos principais animais modelos utilizados em pesquisas relacionadas à DP, tais como primatas não humanos, ratos e camundongos, onde foram mostradas diferenças, tanto na perda de neurônios, como na sua morfologia, consideradas as devidas limitações dos respectivos modelos para a compreensão do quadro mais geral da DP em populações de ambos os sexos (LERANTH *et al.*, 2000).

3.1.3 Animais modelo

De acordo com STOKER (2018), os 3 principais grupos mais utilizados como modelos de doenças humanas são: roedores, primatas não humanos e espécies não mamíferas (Figura 4). Cada um deles tem vantagens e limitações, com relação a especificidades dos mais variados estudos, tais como o fato de roedores serem amplamente estudados em campos biomédicos, porque são convenientes para cuidar em condições de laboratório e têm protocolos experimentais robustos associados, incluindo diferentes formas de administração de drogas, geração de cepas transgênicas e avaliações comportamentais (STOKER, 2018).

Em comparação, os primatas não humanos são maiores, têm uma vida útil mais longa, requerem cuidados mais exigentes, incorrem em custos mais elevados e envolvem considerações éticas mais complexas, fatores estes que pesam na determinando do uso destas espécies. Nenhum desses modelos reproduz perfeitamente a neuropatologia da DP, bem como não replica com exatidão o quadro clínico de pacientes, no entanto, a heterogeneidade desses modelos pode ser vista como uma vantagem, uma vez que um diagnóstico clínico de DP reflete um grupo heterogêneo de pacientes com diferenças no início, progressão, sintomas e neuropatologia (STOKER, 2018).

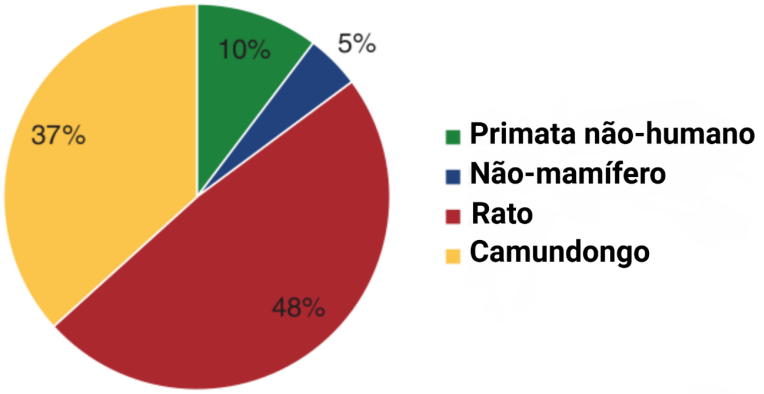


Figura 4: Proporção de animais modelos usados em 23,000 artigos de pesquisa sobre Doença de Parkinson de Janeiro de 1990 à Junho de 2018.

Fonte: STOKER (2018)

A genética é um fator que desempenha um papel importante na patogênese da DP, onde mutações causadoras de doenças foram identificadas por meio de análises em DP familiar, assim como fatores de risco genéticos para DP idiopática. Por meio de análises de associação em pacientes e controles, observa-se que a maioria dos modelos genéticos só foi eficaz na reprodução de algumas das marcas registradas da DP, como por exemplo, o modelo para a mutação LRRK2, onde ratos e camundongos transgênicos apresentam pouca neurodegeneração dopaminérgica e ausência de deficiências motoras. O desenvolvimento de novas ferramentas genéticas, no entanto, permite a geração de novos modelos baseados na genética (STOKER, 2018).

Apesar de novas descobertas relacionadas aos fatores genéticos na DP, os fatores epigenéticos não podem ser menosprezados, uma vez que o quadro de Parkinson se desenvolve também com interações de uma série de fatores ambientais. Até mesmo diferenças de gênero são preservadas também nas interações entre fatores ambientais, onde observou-se que a exposição pré-natal, por exemplo, a pesticidas, aumenta a vulnerabilidade dos homens, mas não das

mulheres, a neurotóxicos dopaminérgicos mais tarde na vida (BARLOW *et al.*, 2004).

3.1.4 Influência ambiental

Sendo uma doença com causas multifatoriais, a doença de Parkinson tem tido suas causas exploradas em estudos com metais pesados, herbicidas, águas de poços e diversos outros componentes ambientais, além da associação de hábitos, como ingestão de cafeína e fumo, entre medidas protetivas, com potencial de diminuição de risco de desenvolvimento de DP (Figura 5). Ademais, foi apontada uma associação entre pesticidas e DP, onde demonstrou-se uma associação aumentada com a exposição profissional a pesticidas em homens e DP de início tardio, hipótese reforçada por numerosos relatórios subsequentes, sugerindo que tais agentes ambientais produzem perda de neurônios em animais. (ELBAZ *et al.*, 2009)

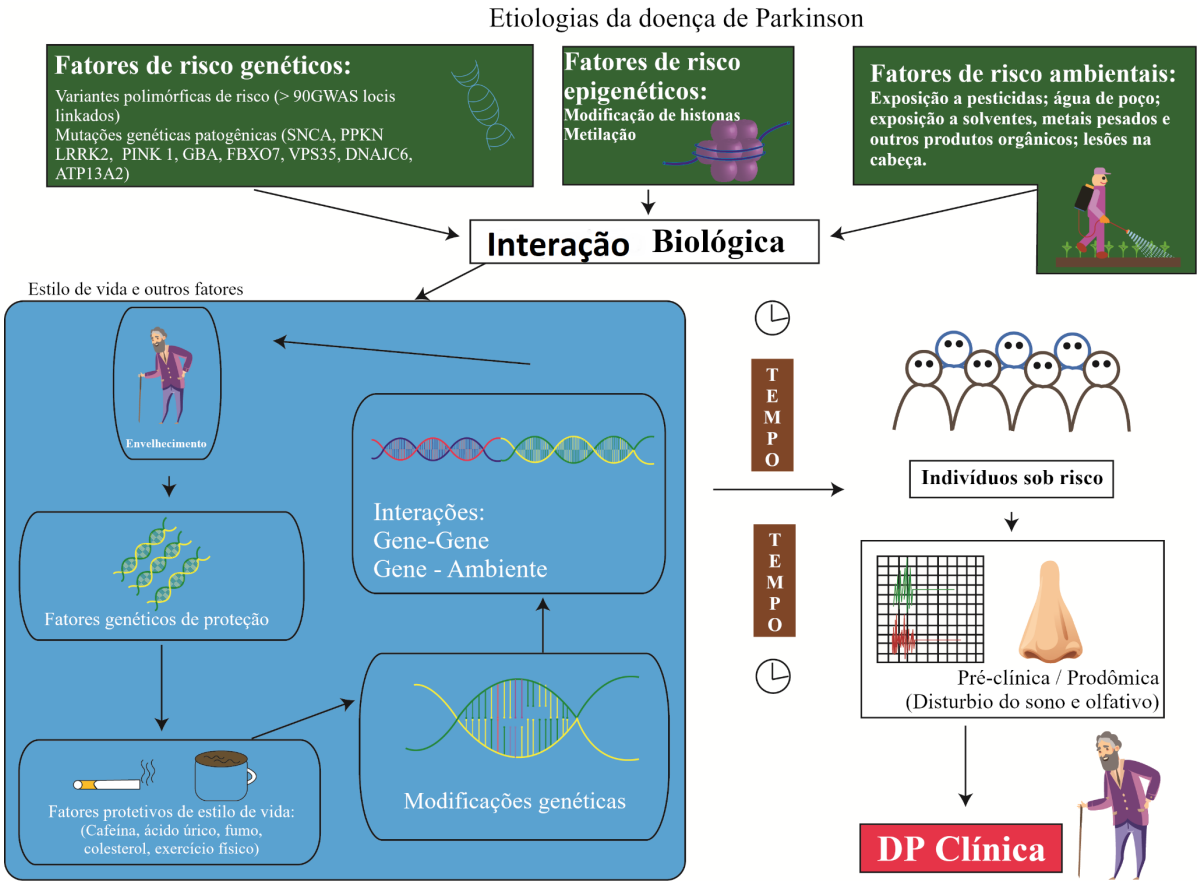


Figura 5: Etiologias da DP: interação biológica entre fatores genéticos, epigenéticos e ambientais.

Fonte: JANKOVIC e TAN (2020)

Estas interações, entre essas diversas variáveis, contribuem para a "*multiple hit hypothesis*", que argumenta a interação de vários fatores de risco, incluindo predisposição genética, exposição a toxinas, envelhecimento e outros fatores potencialmente desconhecidos, interagindo para produzir a síndrome conhecida como Doença de Parkinson. (CARVEY, PUNATI e NEWMAN, 2006)

3.2 Bancos de Dados

Com uma etiologia diversa e complexa, esforços e colaborações de diversos pesquisadores têm buscado criar recursos de pesquisa relacionados à DP. Em meados de 2009, criou-se um banco de dados de genes e variações genéticas relacionados à DP chamado PDbase, utilizando a substância nigra (SN) em DP e tecidos normais (Yang, 2009),. Existe também o ParkDB, um banco de expressão gênica que contém um conjunto completo de dados de *microarray* com curadoria e anotados. Estes recursos permitem que os pesquisadores identifiquem e comparem as assinaturas de expressão envolvidas na DP e na diferenciação de neurônios dopaminérgicos, em diferentes condições biológicas e entre espécies (TACCIOLI *et al.*, 2011).

Fundações como a *The Michael J. Fox Foundation também* tem disponibilizado ferramentas e bases de dados gratuitamente para pesquisas científicas. Dentro dessa perspectiva, o presente trabalho se ateve à análise dos dados dispostos no GEOdatabase, uma vez que as bases de dados estavam disponíveis publicamente, gratuitamente, com facilidade de acesso via web em browsers comuns, como *Google Chrome,* e sem a necessidade de qualquer tipo de cadastro.

3.2.1 NCBI e GEO

O *National Center for Biotechnology Information* (NCBI) possui diferentes bancos de dados integrados, entre eles, o BioProjects, que constitue um meta-recurso para dados biológicos depositados em repositórios de arquivos mantidos por membros da *International Nucleotide SequenceDatabase Consortium* (INSDC), que inclui o *DNA DataBank* do Japão (DDBJ), o European Nucleotide Archive (ENA) do European Molecular Biology Laboratory (EMBL) e o GenBank do NCBI (Lopez-Crisosto, 2021).

O NCBI também possui um banco de dados de bioinformática que fornece um repositório público para dados de sequenciamento de DNA e RNA, como um armazenamento para dados brutos de sequenciamento gerados por tecnologias de próxima geração, incluindo Ilumina, IonTorrent, Complete Genomics, entre outras plataformas de sequenciamento. Atualmente o NCBI armazena mais de 3 milhões de estudos, cobrindo mais de 8,6 milhões de amostras, incluindo 2.597.223 experimentos de acesso público e outros 674.522 estudos controlados.

O GEO é um repositório de dados genômicos funcional, internacional e público que arquiva e distribui gratuitamente *microarrays*, sequenciamentos de próxima geração e outras formas de dados genômicos funcionais de alto desempenho, oferecendo suporte a envios de dados em conformidade com o padrão “Informações mínimas sobre um experimento de *microarray*” (do inglês ‘*Minimum Information About a Microarray Experiment* - MIAME’) (BARRETT *et al.*, 2008)

Dados baseados em matriz e sequência também são aceitos, além de serem fornecidas ferramentas para ajudar os usuários a consultar, analisar e fazer o download de experimentos e perfis de expressão gênica para fins como curadoria. Hoje, o repositório conta com mais de 5 milhões de amostras, oriundas de diversos organismos, das quais 4.510.937 são públicas, dentro de uma variedade de tecnologias distribuídas entre variados estudos, tais como perfil de expressão gênica, perfil de metilação por *array*, entre outros (BARRETT *et al.*, 2012).

3.2.2 Tipos de tecnologías arquivadas

Tecnologias de sequenciamento têm proporcionado desenvolvimento de plataformas de sequenciamento de alto rendimento e baixo custo além de novas abordagens de sequenciamento rápido e de baixo custo que não apenas causam mudança no cenário de projetos de sequenciamento de genomas, mas também trazem oportunidades para sequenciamento em uma variedade de formas inovadoras para aplicação dessas tecnologias que estão sendo desenvolvidas.Ao longo dos anos, viu-se tecnologias de próxima geração sendo aplicadas em uma diversidade de contextos, incluindo sequenciamento de todo o genoma, re-sequenciamento de genomas por variações, perfilamento de mRNAs e outros RNAs pequenos e não codificantes, avaliação de proteínas de ligação ao DNA e estruturas de cromatina e detecção de padrões de metilação. Técnicas tradicionais como microarray em aplicações de genômica funcional, vem sendo desafiadas pelas técnicas de sequenciamento de próxima geração de alto rendimento.(ZHOU et al., 2010)

O RNA-Seq, por exemplo, é uma abordagem desenvolvida recentemente para o perfilamento de transcriptoma que usa tecnologias de sequenciamento profundo.Utilizando este método alguns estudos já alteraram nossa visão da extensão e complexidade dos transcriptomas eucarióticos. Essa abordagem também fornece uma medição muito mais precisa dos níveis de transcritos e suas isoformas do que outros métodos.Tendo em vista a importância do transcriptoma para interpretar os elementos funcionais do genoma e revelar constituintes moleculares das células, tecidos além da importância para compreender o desenvolvimento e doenças (como DP,por exemplo), abordagens transcriptômicas têm como alguns de seus objetivos: “catalogar todas as espécies de transcritos, incluindo mRNAs, RNAs não codificantes e pequenos RNAs; para determinar a estrutura transcricional dos genes, em termos de seus locais de início, extremidades 5 'e 3', padrões de splicing e outras modificações pós-transcricionais; e quantificar os níveis de expressão variáveis de cada transcrição durante o desenvolvimento e sob diferentes condições”.Em contraste com os métodos de microarray, as abordagens baseadas em sequência determinam diretamente a sequência do cDNA.(WANG; GERSTEIN; SNYDER, 2009)

Como descreve Dopazo et al. (2001): “a utilização de tecnologias de array de DNA de alta densidade para monitorar a transcrição de genes têm tido impacto nas mudanças de paradigma na biologia. Grande parte dos grupos de pesquisa agora têm a capacidade de medir a expressão de uma significativa proporção do genoma humano em um único experimento,como resultado um volume sem precedentes de dados vem sendo disponibilizado para a comunidade científica. Como consequência, armazenamento, análise e interpretação desta informação apresentam um grande desafio”.Bancos de dados como o GEO descrevem que continuam a ter o perfil de expressão por array como tipo de estudo submetido mais comum,em uma ordem de magnitude, embora sua taxa de crescimento esteja diminuindo. As taxas de envio de sequência de próxima geração têm aumentado rapidamente desde 2008, sendo que curiosamente, métodos como a imunoprecipitação da cromatina por sequenciamento estão aumentando de modo que agora são submetidos a uma frequência mais alta do que o chip ChIP de sua contraparte baseado em array.(BARRETT et al., 2008)"

4 MATERIAL E MÉTODOS

4.1 Levantamento das bases de dados e escolha do GEO

Durante a pesquisa sobre bases de dados com buscadores usuais, como Google e Bing, foram encontradas bases de dados, tais como o PDbase: um banco de dados de genes relacionados à doença de Parkinson e variação genética usando substantia nigra (YANG *et al.*, 2009). O PDbase está disponível gratuitamente em http://bioportal.kobic.re.kr/PDbase/ e as listas de genes relacionados à DP geradas podem ser encontradas em http://diseasome.kobic.re.kr/, no entanto, nenhum desses links foi acessado com sucesso em browsers comuns neste presente trabalho. Fato semelhante ocorreu para a base ParkDB (TACCIOLI et al., 2011), uma base de dados de expressão gênica da doença de Parkinson, com a URL: http://www2.cancer.ucl.ac.uk/Parkinson_Db2/.

As bases de dados e repositórios de arquivos disponíveis e integradas do *National Center for Biotechnology Information* (NCBI), tais como *Gene Expression*

Omnibus (GEO), *BioProject*, *Sequence Read Archive* (SRA), além de públicas e de fácil acesso, oferecem suporte ao armazenamento de dados brutos, dados processados e metadados que são indexados e reticulados. Todos estes dados estão disponíveis gratuitamente para *download* em uma variedade de formatos. O GEO também fornece várias ferramentas e estratégias baseadas na web, além de ferramentas e aplicações baseadas em R, que auxiliam os usuários a consultar, visualizar e analisar os GEOdatasets. Por esta facilidade e quantidade de informações, o presente trabalho concentrou-se em bases de dados disponíveis em https://www.ncbi.nlm.nih.gov/gds/.

4.2 Extração manual de dados disponíveis

Após a escolha de uma série de palavras chaves, tais como “Parkinson” e “Parkinson disease”, todos os projetos apresentados entre os resultados da pesquisa no repositório *GEO DataSets* do NCBI foram selecionados e checados quanto ao conteúdo de suas informações disponíveis e relevantes para o presente trabalho.

Um levantamento manual foi realizado inicialmente com todos os trabalhos elencados no resultado da pesquisa no *GEO DataSets*, considerando dados como: identificadores do repositório BioProject, publicações (caso houvessem) e respectivos identificadores PMID, “highlights” destacados dos projetos, acessos do GEO, identificadores do SRA, organismos dos projetos, linhagens celulares, tipos de dados, tipos de tecnologia, plataformas utilizadas nos experimentos, quantidades de amostras e de condições, tipos de tecidos e células, quantidades de amostras femininas e masculinas, caso presentes e ausência, ou presença de informação de idade, etnia e informações sobre identificação de controle e paciente em amostras, além de suas respectivas quantidades.

Para garantir a relevância dos projetos de DP, foi realizada uma nova curadoria manual através de verificação de título e sumário de trabalhos levantados. Em seguida, foi elaborada uma tabela atribuindo os trabalhos e suas respectivas informações, utilizando uma filtragem com critérios baseados nos principais fatores relacionados para visualizar as informações de projetos diretamente relacionados

com a DP (como os que apresentam genes, substâncias, tecidos entre outros fatores correlacionados) e submetê-los a uma análise exploratória de dados disponíveis.

4.3 Montagem de planilha e análise exploratória de dados

Estatística Descritiva é uma das etapas iniciais de análises de dados e constitui-se do ramo da Estatística que implementa técnicas para descrição, organização e resumo de um conjunto de dados. Assim sendo, explora-se, dentre outros assuntos, medidas como as de posição (média, moda e mediana) e de dispersão (variância, desvio padrão,amplitude, etc.), assim como avaliações de ordem qualitativa. Para tal fim, foi realizada uma montagem de tabela manualmente, através do Planilhas Google, um programa de planilhas gratuito e online incluído como parte do pacote de Editores do Google Docs, oferecido pelo Google. As análises exploratórias de dados e geração de gráficos foram realizadas por programas e ferramentas gráficas do próprio Google Sheets e utilizando a linguagem Python (Versão 3.7.4) de programação, com utilização da biblioteca *pandas* via Google Colab e com utilização do BioRender para finalização das ilustrações.

4.4 Resultado das buscas e visualização das informações disponíveis

Na utilização de diferentes palavras-chaves foram obtidas quantidades diferentes de trabalhos e dados de amostras associadas à pesquisa no banco de dados do GEO, onde foram obtidos os seguintes resultados para: "Parkinson", 298; "Parkinson RNA", 124; " Parkinson RNA seq", 32; "Parkinson ncRNA", 28; "Parkinson neuron", 30; "Parkinson DNA", 27 e “Parkinson disease’, 3916.

A consulta com a palavra-chave “Parkinson disease” retornou o maior número de resultados e abrange trabalhos encontrados nas demais. Dessa forma, todos eles foram utilizados para composição da tabela, excluindo daí aqueles que não tinham relação com Parkinson, mas que existiam pela relação com o nome do autor, de trabalhos diferentes, mas com amostras de mesmos repositórios, ou por

qualquer outra anomalia, por exemplo, a do projeto com identificador PRJNA513339 (publicado por, Camunas-Soler J,2020), que trata de disfunções fisiológicas em diabetes. Para isso, foi realizada uma busca pelo termo “Parkinson disease” nos títulos e sumários dos trabalhos..

Foram encontrados 458 resultados para o termo “Parkinson Disease” (Título+sumário) selecionados para tabela, com 384 no Sumário e 74 trabalhos com “Parkinson disease” no título. Aplicando o filtro final, foi realizada a seleção dos trabalhos onde a DP era apenas mencionada como um impacto secundário da pesquisa descrita, e não constava nenhum dos fatores de impacto e relevância para DP, previamente descritos neste trabalho, restando um total de 365 projetos.

5 RESULTADOS E DISCUSSÃO

5.1. Análise das informações gerais extraídas

Dos 365 projetos da curadoria final, 274 (75,1%) encontravam-se com alguma publicação associada, enquanto 91 (24,9%) encontravam-se sem qualquer publicação (Figura 6). Segundo o critério previamente estabelecido em Lopez-Crisosto *et al.* (2021), a qualidade dos dados foi considerada alta em 240 (65,8%) projetos e baixa em 125 (34,2%), uma vez que satisfaziam o critério mínimo de pelo menos 3 amostras para cada condição estudada (Figura 7). Todos os 365 trabalhos apresentavam os respectivos acessos do GEO linkados em seus projetos e 183 apresentavam link para os respectivos “arquivos de *reads* de sequência” (do inglês, *Sequence Read Archive* - SRA).

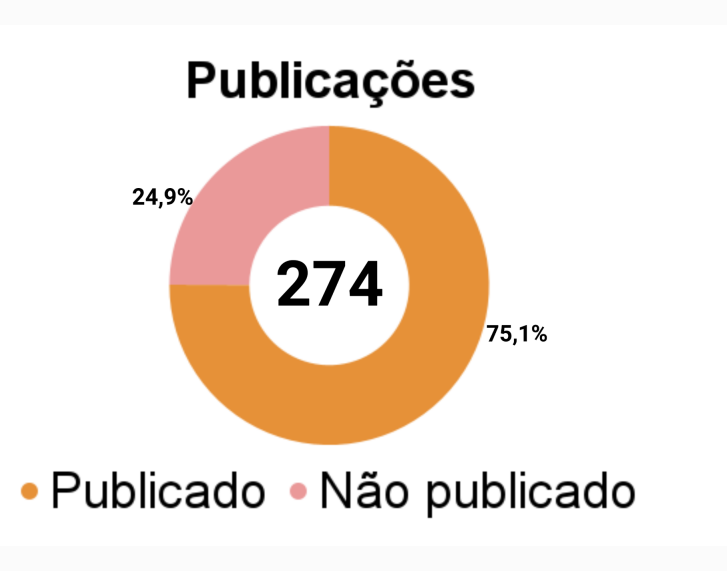


Figura 6: Porcentagem de BioProjects com publicações associadas e sem publicação associada

Fonte: AUTORIA PRÓPRIA

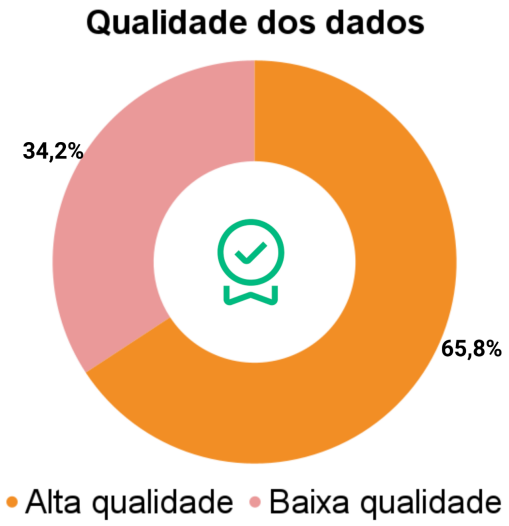


Figura 7: Porcentagem de BioProjects com dados de alta e baixa qualidade

Fonte: AUTORIA PRÓPRIA

Inicialmente foram encontradas 20 espécies diferentes entre os resultados da pesquisa com “Parkinson disease”, sendo estas: *Homo sapiens* (3626), *Mus musculus* (213), *Drosophila melanogaster* (42), *Rattus norvegicus* (29), *Danio rerio* (5), *Caenorhabditis elegans* (5), synthetic construct (Constructo sintético amostral de *Homo sapiens* e *Mus Musculus*) (5), *Microcebus murinus* (4), *Callithrix jacchus* (3), *Saccharomyces cerevisiae* (3), *Macaca fascicularis* (2), *Mesocricetus auratus* (2),

Ovis aries (2), *Gallus gallus* (2), *Macaca mulatta* (1), *Chlorocebus sabaeus* (1), *Rattus rattus* (1), *Nannospalax galili* (1)*,Tupaia belangeri* (1) e *Bos indicus* (1).

Com a curadoria final, onde apenas os 365 trabalhos com “Parkinson disease” mencionado ao menos no sumário, e cujo impacto não fosse secundário ou não envolvesse nenhum dos fatores de impacto para DP, obtivemos uma diversidade menor, onde encontramos uma maior ocorrência de *Homo sapiens* com 233, representando 63,5% do total dentre a espécies encontradas; *Mus musculus*, com 94 ocorrências, representando 25,6%; *Rattus norvegicus*, com 17 (4,6%); *Drosophila melanogaster*, com 6 ocorrências (1,5%) e uma variedade menor de outras espécies, com poucas ocorrências, representando 4,6% do total de 367 espécies descritas (Figura 8). Ressalta-se que mais de uma espécie pôde ser observada em um mesmo projeto, como observado no BioProject de identificar PRJNA101815 (publicado em BRUG *et al.*, 2008).

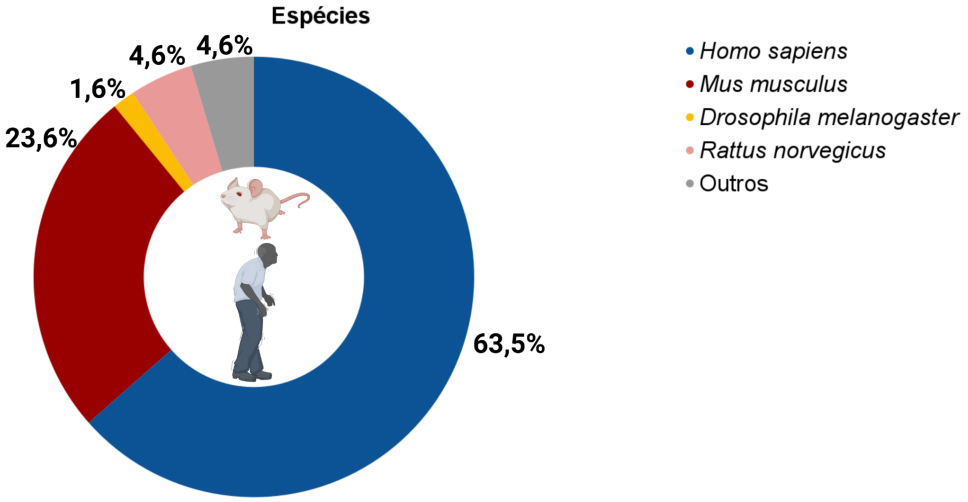


Figura 8: Principais espécies das 367 encontradas nos projetos extraídos do GEO dataset.
Fonte: AUTORIA PRÓPRIA

A proporção de espécies encontradas não divergiu muito das descritas em trabalhos anteriores e em outras bases de dados como o ParkDB, no entanto, nota-se uma proporção inferior de primatas não-humanos (presentes em apenas 4, dos 365 projetos), que são comumente utilizados para avaliação pré-clínica de

terapias, além de fornecerem informações valiosas sobre a patologia da DP, devido à sua semelhança anatômica e genética com os humanos, como demonstrado em Grow, McCarrey e Navara (2016).

Muito se argumenta que diversificar os modelos animais pode ajudar a compreender vários subtipos de DP e desenvolver tratamentos personalizados, no entanto, trabalhos como o descrito no BioProject de identificador PRJNA698999 (publicado em MONZÓN-SANDOVAL *et al.*, 2020), argumentam que para comparação da variação da expressão gênica em populações humana e de camundongo de substância nigra pars compacta (SNpc), há uma grande necessidade de estudos ômicos com foco humano.

Os tipos de dados "ômicos" encontrados foram: 269 de Transcriptoma, representando 73,7%; 34 de Epigenoma (9,3%); 9 de Proteoma (2,5%); 7 de Genoma (1,9%) e 34, representando 12,6%, de uma variação de outros, que podiam incluir amostras de diferentes tipos dos anteriormente citados (Figura 9).

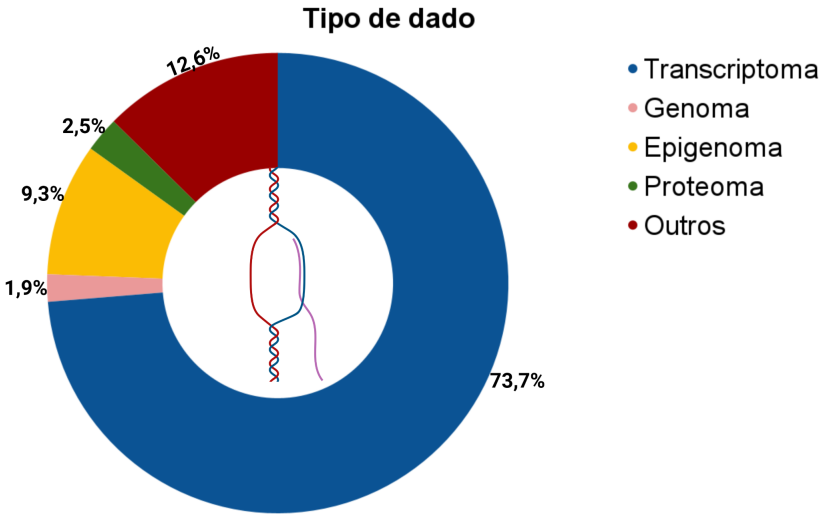


Figura 9: Principais tipos de dados "ômicos" encontrados entre os 365 projetos

Fonte: AUTORIA PRÓPRIA

Dados de transcriptoma tem contribuído para identificação de diferenças significativas entre tipos de células e tecidos celulares, como observado no BioProject de identificador PRJNA434419 (publicado em SCHULZE *et al.*, 2018), bem como na compreensão das marcas patológicas da DP, identificação de vias

moleculares (PRJNA132861, publicado em ZHENG *et al.*, 2010) e possíveis biomarcadores (PRJNA556869).

Um enfoque crescente no estudo do papel de RNAs não codificantes em diferentes aspectos da DP também contribui para essa maior representação, muito embora uma maior variabilidade na disponibilidade de dados de qualidade dos demais tipos de dados “ômicos” seja desejável, uma vez que dados proteômicos, por exemplo, que representam apenas 2,5% do total encontrado entre os projetos, têm demonstrado potencial para identificação de biomarcadores e no diagnóstico da doença de Parkinson em estágio inicial usando autoanticorpos (PRJNA263674, publicado em DEMARSHALL *et al.*, 2015).

Embora nem todas as técnicas estivessem detalhadas nas páginas de acesso utilizadas, foram encontrados diversos tipos de *arrays*, tais como perfil de expressão, perfil de RNA não codificante e perfil de metilação. Deste modo, dentre os projetos da curadoria final, as técnicas com maior ocorrência foram de *array* com 163 ocorrências, representando 52,9%; RNA-Seq com 103 (33,4%), Bisulfite-Seq com 19 (6,1%); ChIP-Seq com 8 (2,6%); e uma série de ocorrências menores de outros, como por exemplo, ATAC-seq e MeDIP-Seq, representando conjuntamente 4,8% do total (Figura 10).

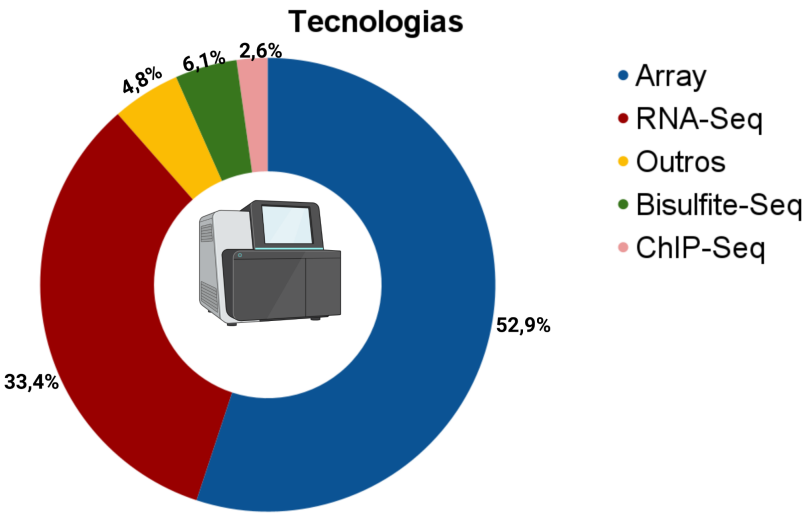


Figura 10: Principais tecnologias utilizadas nas experimentações das 308 descritas entre os projetos

Fonte: AUTORIA PRÓPRIA

O GEO aceita dados de sequência para estudos que examinam expressão gênica com uso de RNA-Seq, regulação gênica e epigenômica com uso, por exemplo, de ChIP-Seq e metil-Seq, ou outros estudos em que se deseja medir alguma forma de abundância ou caracterização de sequência. Apesar do crescente número de estudos selecionados e liberados pelo GEO, ainda assim observa-se uma significativa ausência de dados brutos contendo as leituras de sequência original.

O GEO hospeda os arquivos de dados processados junto com amostras e metadados de estudo onde são intermediados e vinculados ao banco de dados *Sequence Read Archive* (SRA) do NCBI. Os respectivos identificadores que concedem acesso ao SRA só estavam presentes em 183, ou seja, quase metade deles não tinham esses dados brutos acessíveis, o que pode ser considerado um ponto negativo, uma vez que impede a verificação e comparação destes dados com demais experimentos semelhantes ou de interesse para fins de esclarecimento da DP, além de que, como previamente mencionado, novas tecnologias com resultados acessíveis podem contribuir para o melhoramento de organismos modelo para estudos de DP.

Nem todos os projetos apresentavam os tecidos e seus respectivos tipos na descrição, assim como, nem sempre com precisão na classificação histológica ou mesmo padronização na nomenclatura, onde pode-se encontrar desde descrições referentes apenas ao cérebro, como as regiões do encéfalo (cerebelo, substantia nigra, cortex, entre outras) ou tipos celulares e cortes histológicos. Sendo assim, para caráter de levantamento geral dos diferentes tipos de tecidos, uma nomenclatura geral foi utilizada para os tecidos de mesma fonte histológica.

No geral, dentre os projetos que descreviam seus respectivos tecidos, obteve-se uma grande variedade de tecidos e nomenclaturas, apesar de alguns desses terem uma frequência muito baixa de ocorrências, por isso, foram agrupados como “outros”. Das principais ocorrências, observou-se o tecido cerebral com 158 ocorrências, representando 73,0%; sanguíneo com 22 (10,00%); oriundo de células tronco pluripotente induzidas com 7 (3,2%); demais tecidos, como por exemplo

testículos, bulbo olfativo e rim, tiveram baixa frequência (1 ou 2 projetos) e foram agregados. Esses demais tecidos representaram 12,8% do total verificado nos projetos (Figura 11).

Sendo uma doença neurodegenerativa, é esperado que a maioria dos dados sejam oriundos de diferentes células do tecido nervoso, ressaltando mais uma vez que a falta de padronização na nomenclatura, uma vez que nem sempre classificações histológicas são apresentadas, pode gerar uma dificuldade para levantamento de dados em pesquisas relacionadas, que visam esclarecer pontos sobre o funcionamento dos mesmos tecidos na DP.

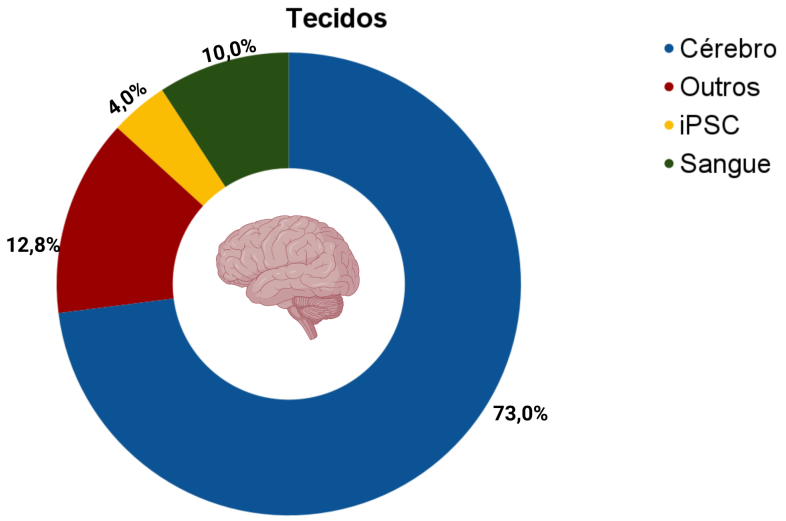


Figura 11: Principais tecidos do total de 219 encontrados nas descrições entre os 365 BioProjects

Fonte: AUTORIA PRÓPRIA

Mesmo com a maioria oriunda de tecidos cerebrais, pesquisas recentes têm se voltado, tanto para aspectos imunogenéticos da DP, como para para potenciais terapias com células tronco que têm se mostrado promissoras (STOKER, 2018). Tal fato ressalta a importância de acesso a dados de qualidade de demais tecidos que aqui representaram menos de 30% dos projetos (dos quais células de origem sanguínea e células tronco representaram menos de 15%).

5.2. Análise das 3 principais espécies presentes

Sendo a população humana a população de maior interesse em estudos com DP, para fins de melhor compreensão diagnóstica, desenvolvimento de tratamentos e medicamentos, encontra-se, como esperado, esta espécie com a maior representação entre os 365 projetos (63,5%), seguida de camundongos (Mus musculus, 23,6%) e ratos (Rattus novergicus, 4,6%), que são também alguns dos principais organismos modelo utilizados para essa doença.Estes roedores foram os mamíferos não humanos mais utilizados em pesquisas sobre DP nos últimos 28 anos, representando 85% dos 23.000 trabalhos descritos entre 1990 e 2018, como descreve a literatura em Stoker (2018).

Nesta seção analisamos os dados presentes nessas 3 principais espécies, uma vez que representam as mais utilizadas nos experimentos, com mais de 90% dos 365 projetos curados no presente trabalho.

5.2.1 *Homo sapiens*

Com a maior frequência entre as espécies verificadas nos projetos (63,5%) e sendo a principal espécie alvo de estudos, diagnóstico e tratamento, verificamos os principais tecidos e tipos moleculares utilizadas nos projetos, assim como a presença ou ausência de informação referente a gênero, idade e etnia de espécimes elencadas nos respectivos projeto.

Nos tecidos presentes para os projetos com amostras oriundas de seres humanos, observou-se, dentre os que disponibilizaram a informação na descrição, 75 projetos contendo tecidos cerebrais, representando 61,0%; 21 apresentando tecidos sanguíneos (17,1%); 7 oriundos de células tronco pluripotente induzidas (5,7%) e um conjunto de outros 20, com frequência menor e oriundos, por exemplo, de tecidos musculares (16,2%) (Figura 12).

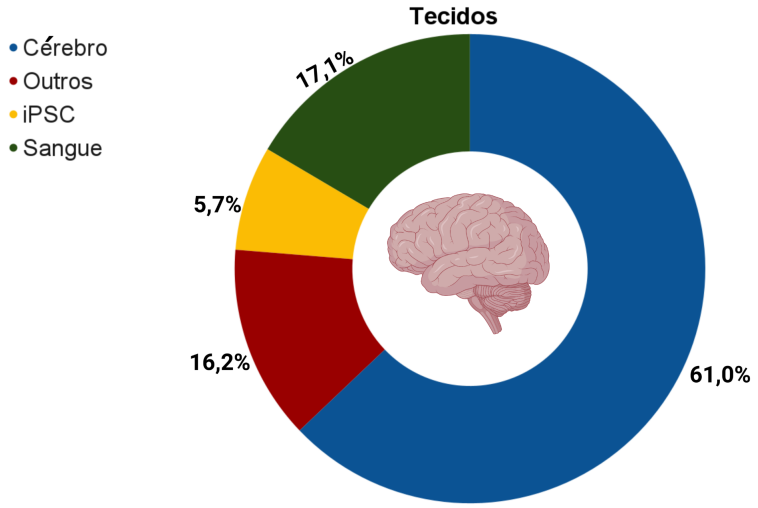


Figura 12: Principais tecidos encontrados do total de 123 descritos entre os projetos de experimentos com *Homo sapiens*

Fonte: AUTORIA PRÓPRIA

Como anteriormente abordado, observa-se aqui também a predominância de tecidos cerebrais, mas também um aumento na proporção de experimentos com amostras teciduais sanguíneas e oriundas de células tronco, os aspectos moleculares e imunogenéticos têm se tornando alvo de estudos para fins, inclusive, de avaliação terapêutica, como observado no BioProject de identificador PRJNA167759, onde foi avaliado o perfil de expressão de microRNA (miRNA) de leucócitos sanguíneos de pacientes com doença de Parkinson (DP) pré e pós-tratamento de estimulação cerebral profunda e de voluntários saudáveis. Dados de projetos como esse, por exemplo, não encontravam-se publicados ou ligados a repositórios de dados brutos como SRA. Tais circunstâncias dificultam a proposta de integralidade e utilização de resultados destes experimentos com as demais pesquisas relacionadas à DP.

Diante da diversidade de tipos moleculares extraídos para experimentação, os principais encontrados entre os projetos, com utilização de amostras de espécimes humanos, foram de RNA, totalizando 55 ocorrências (64,3%); DNA genômico, com 56 (23,2%); RNA polyA, com 19 (7,9%); proteínas, com 5 (2,1%) e um conjunto de outros tipos moleculares com ocorrências menores, representando

2,5% do total encontrado entre os 233 projetos, com espécimes humanos (Figura 13).

RNAs não codificantes têm tido crescente foco no estudo de diversas patologias,como observado com relação aos micro RNAs em Yuan, Ma, Wei, Wu, Hu e Liu (2013), uma vez que suas diversas funções regulatórias podem ajudar a contribuir para esclarecimento, tanto do funcionamento, como de possíveis alvos terapêuticos para doenças como DP, provavelmente uma das razões para a extração de RNA representar mais de 60% de todas as extrações para experimentação dos projetos aqui levantados.

Foi relatado que microRNAs (miRNAs) contribuem para a fisiopatologia da Doença de Parkinson (DP), são uma classe de pequenas moléculas de RNA implicadas na regulação pós-transcricional da expressão gênica durante o desenvolvimento. Potenciais terapias baseadas em miRNA podem ser uma ferramenta poderosa para estudar a função do gene, investigar o mecanismo da doença e validar alvos de medicamentos logo, o acesso a dados de qualidade destes é de extrema importância (YUAN *et al*., 2013). Infelizmente, nem todos os trabalhos detalham em suas descrições e acessos linkados, os tipos de RNAs não codificantes. Foram encontrados miRNAs mencionados nos títulos de 7 projetos com humanos, onde apenas um apresentava o respectivo identificador e link de acesso ao SRA.

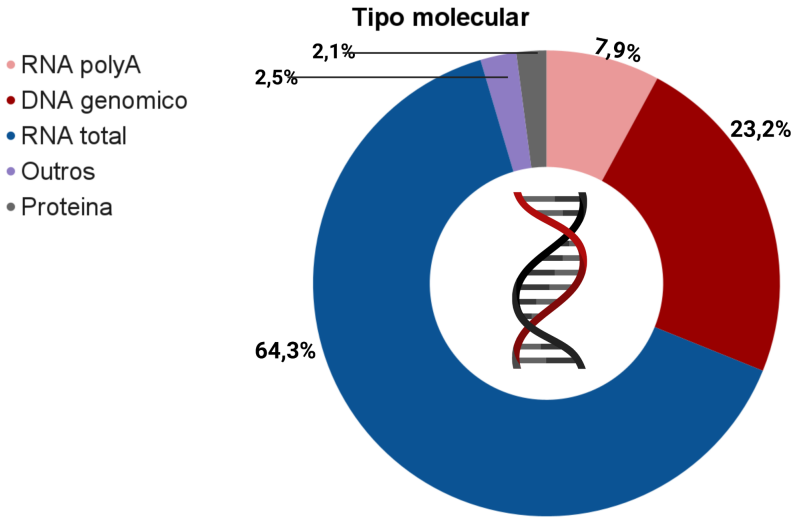


Figura 13: Principais moléculas extraídas das 241 descritas entre os projetos experimentos com *Homo sapiens*

Fonte: AUTORIA PRÓPRIA

Dentre as informações disponíveis e visualizáveis através da página web do banco de dados sobre o gênero, idade ou etnia dos espécimes, foram encontradas 118 projetos cujo gênero era uma informação ausente e 116, que apresentavam essa informação, representando 50,47% e 49,57%,respectivamente, do total encontrado entre os 233 projetos com espécimes humanos. A idade foi uma informação ausente em amostras de 142 projetos e presente em 91, representando 60,94% e 30,05%, respectivamente. Para a informação sobre a etnia, observou-se a maior diferença entre presença e ausência de todos os levantamentos, com apenas 7 projetos que informaram os respectivos grupos étnicos envolvidos na amostragem, logo, com uma ausência de informação representando 96,98% e apenas 3,01% apresentando-a (Figura 14).

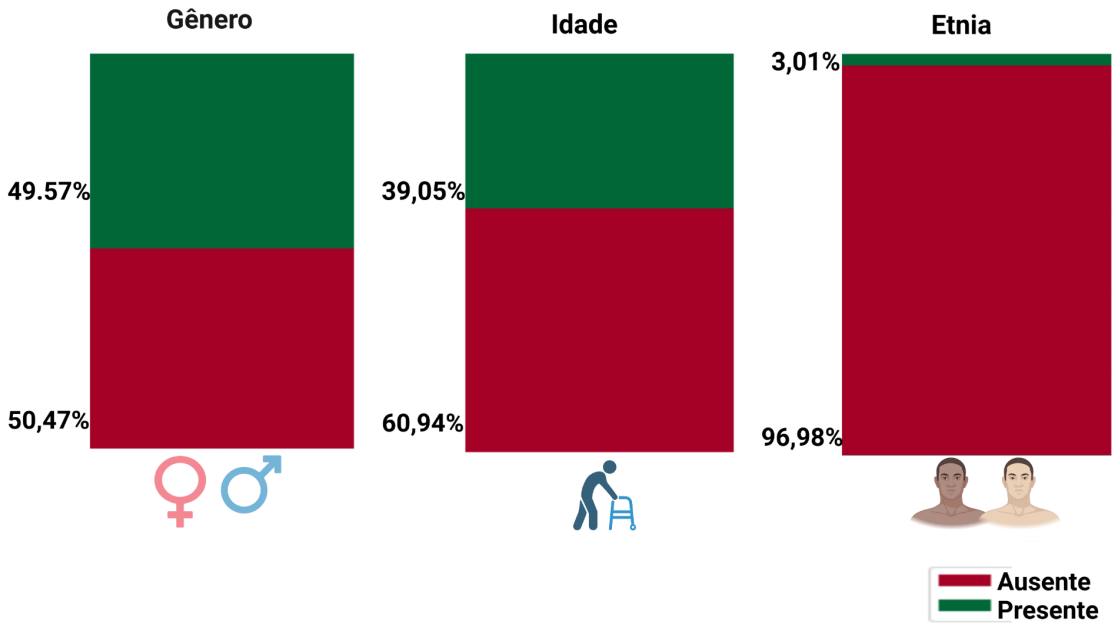


Figura 14: Presença (verde) ou ausência (vermelho) de informação sobre gênero,idade e etnia respectivamente em amostras de *Homo sapiens.*

Fonte: AUTORIA PRÓPRIA

A ausência desses dados em conjuntos de dados públicos sobre DP é de destaque importante, uma vez que características como gênero e idade são reconhecidos como fatores de impacto no risco de desenvolvimento e no progresso da DP e a etnia, mesmo que para censo demográfico, ou descarte de hipótese num impacto maior ou menor em populações de determinados grupos étnicos, também seja uma característica amostral de importante destaque. Tamanho é o impacto do sexo, por exemplo, que estudos recentes tem levantado o efeito do gênero na qualidade do RNA e na quantidade total de RNA extraído *post mortem* (CANTUTI-CASTELVETRI *et al.*, 2007). Logo, certas particularidades moleculares específicas do sexo podem influenciar as conclusões de análises originais e posteriores, ressaltando a importância dessa informação.

5.2.2 *Mus musculus*

Mus musculus é a espécie mamífera não-humana mais utilizada nos trabalhos visualizados na base dados do GEO, representando 23,6% dos 365

projetos. Também verificamos os principais tecidos, tipos de moléculas extraídas para experimentação e a presença ou ausência de informações referentes ao gênero, a idade e raça destes roedores.

Nos camundongos, os principais tecidos também foram oriundos do cérebro, com a diferença principal em comparação às amostras humanas de uma representação maior de amostras renais. No geral, foram visualizados 65 projetos com descrição de tecidos cerebrais, representando 89% do total de projetos com tecidos descritos e amostras oriundas de camundongos; com células renais, verificou-se a ocorrência de 3 projetos (4,1%), enquanto nos humanos, representavam apenas 1 projeto; e apenas 1 projeto (1,4%) contendo amostras oriundas de tecido sanguíneo, além de outros 4 tecidos representando, em conjunto, 5,5% do total de projetos (Figura 15).

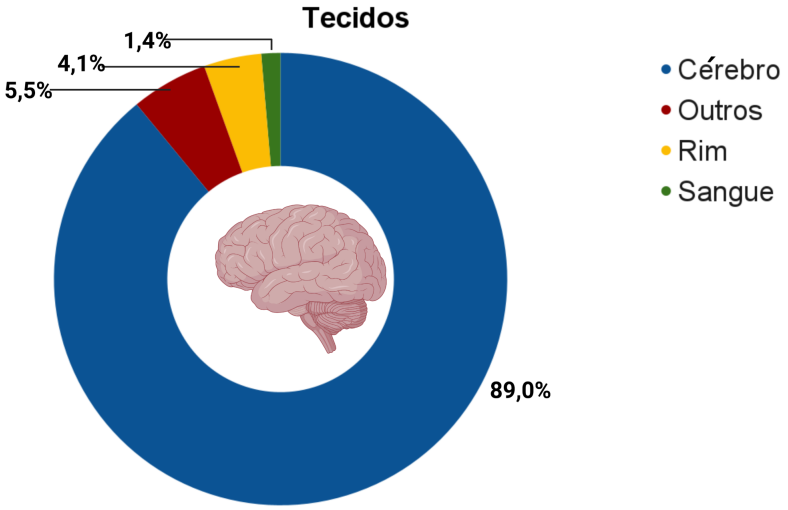


Figura 15: Principais tecidos encontrados dos 73 descritos entre os BioProjects de experimentos com *Mus musculus*

Fonte: AUTORIA PRÓPRIA

Os principais tipos moleculares extraídos para experimentação encontrados entre os projetos, com utilização de amostras de espécimes de camundongos foram: RNA total, com 80, representando 85,1%; DNA genômico com 7 (7,4%); RNA polyA com 6 (6,4%) e 1 projeto com RNA nuclear, (1,1%) do total de tecidos descritos em projetos levantados para esta espécie (Figura 16).

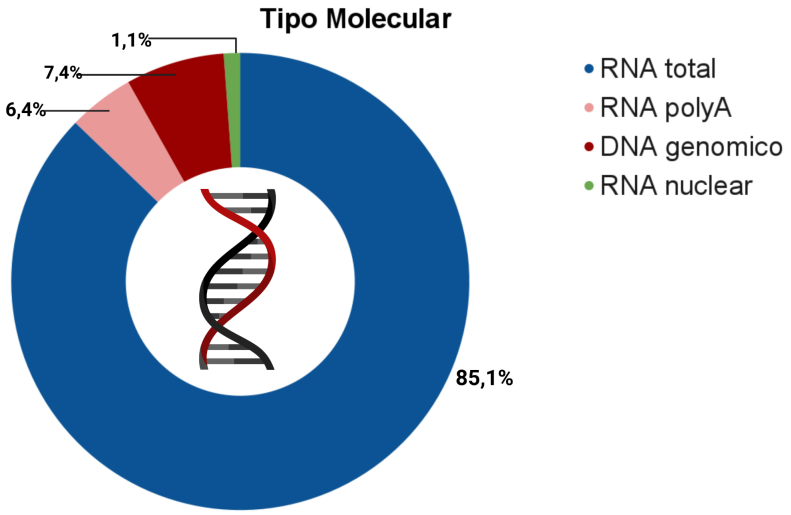


Figura 16: Principais moléculas extraídas das 94 descritas entre os projetos de experimentos com *Mus musculus*

Fonte: AUTORIA PRÓPRIA

Observa-se também uma super representação da extração do RNA total, no entanto, informações mais específicas sobre os tipos de RNA são ainda mais escassas até na titulação dos trabalhos com esta espécie, sendo mencionados 1 microRNA e 1 RNA longos não codificantes.

Dentre as informações disponíveis e visualizáveis através da página web do banco de dados sobre o gênero, idade ou etnia dos espécimes, cujas amostras são originárias de camundongos, foram encontradas 66, cujo gênero era uma informação ausente e 28, que apresentavam essa informação, representando 69,56% e 30,43%, respectivamente, do total de 94 projetos que utilizavam esta espécie. A idade foi uma informação ausente em amostras de 56 projetos e presente em 38, representando 59,13% e 40,86%, respectivamente. Para a informação sobre a raça dos camundongos, observou-se 64 projetos que não informaram os respectivos grupos envolvidos na amostragem e 30 que apresentavam, representando 67,74% e 32,35% do total, respectivamente (Figura 17).

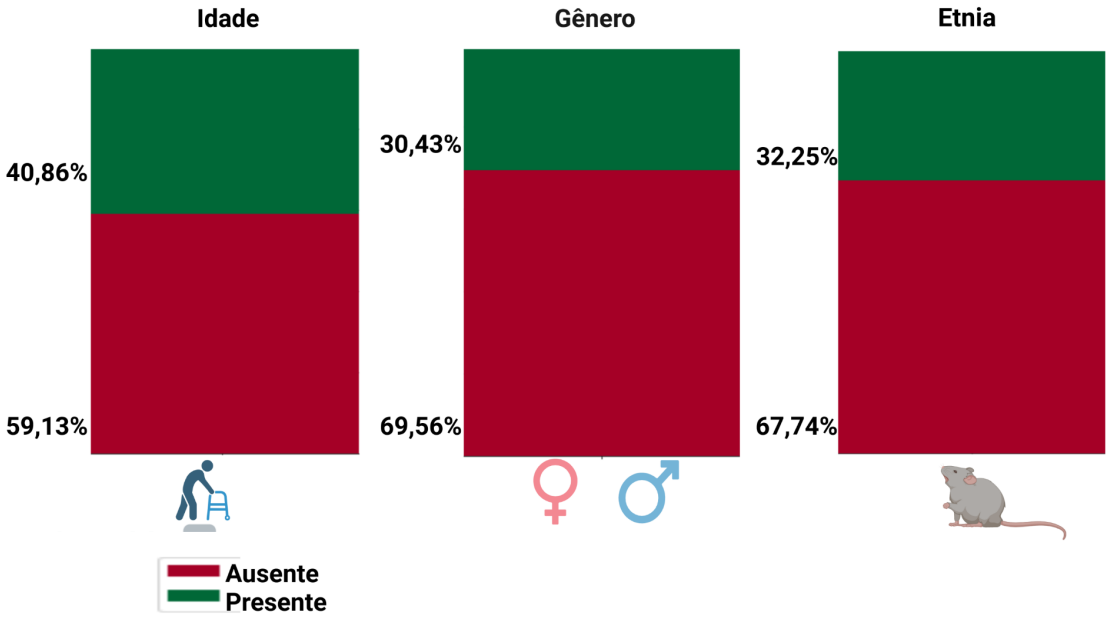


Figura 17: Presença (verde) ou ausência (vermelho) de informação sobre idade, gênero e etnia respectivamente em amostras de *Mus musculus*.

Fonte: AUTORIA PRÓPRIA

Devido às limitações já mencionadas destes animais modelos, informações como a linhagem (raça/etnia) são importantes para melhor compreensão e desenvolvimentos destes modelos. Assim como a disponibilização de dados brutos linkados com respectivo identificador SRA, que aqui estavam ausentes em 51, dos mais de 90 projetos em que esta espécie foi utilizada, também têm suma importância para melhoramento de análises em organismos modelos para fins de estudos parkinsonianos.

5.2.3 *Rattus norvegicus*

Todos os tipos de tecidos visualizados nos projetos que descreviam tal informação em seus acessos para os BioProjects desta espécie continham tecidos de diferentes regiões do cérebro (tais como *striatum,* com 7 ocorrências e córtex frontal, com apenas 1), onde demais projetos, quando mencionaram o tecido, referiam-se apenas como oriundo do "cérebro", tornando desnecessária uma representação gŕafica. Aqui mais uma vez ressalta-se a importância de uma

padronização quanto a exigência de nomenclaturas de caráter histológico, para facilitação de busca, curadoria e seleção amostral para futuras pesquisas e colaborações com estas bases de dados.

Quanto aos tipos de moléculas extraídas de espécimes de ratos, as principais não divergiram das demais espécies, sendo RNA total também o de maior ocorrência com 14 dos 17 projetos, representando 82,4%; DNA genômico com 2 ocorrências (11,8%) e RNA polyA, com apenas 1 uma ocorrência (5,9%) (Figura 18).

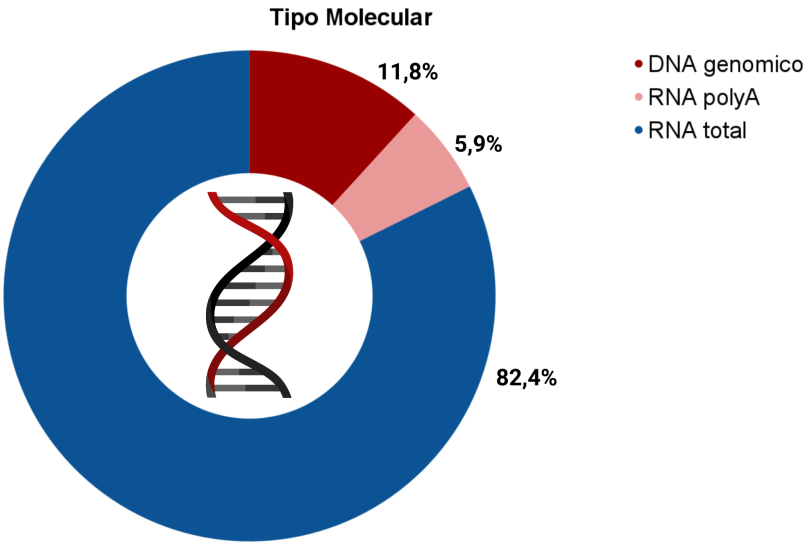


Figura 18: Principais moléculas extraídas das 17 descritas entre projetos de experimentos com *Rattus novergicus.*
Fonte: AUTORIA PRÓPRIA

Dentre as informações disponíveis e visualizáveis através da página web do banco de dados sobre o gênero, idade ou etnia dos espécimes, cujas amostras são originárias ratos, foram encontradas 9, cujo gênero era uma informação ausente e 8 que apresentavam essa informação, representando 52,94% e 47,05%, do total de 17, respectivamente.

A idade foi uma informação ausente em amostras de 13 projetos e presente em 4, representando 76,47% e 23,52%, respectivamente. Para a informação sobre a raça dos camundongos, observou-se em 9 projetos que não informaram os respectivos grupos envolvidos na amostragem e 8 que apresentavam, representando 52,94% e 47,05% do total, respectivamente (Figura 19).

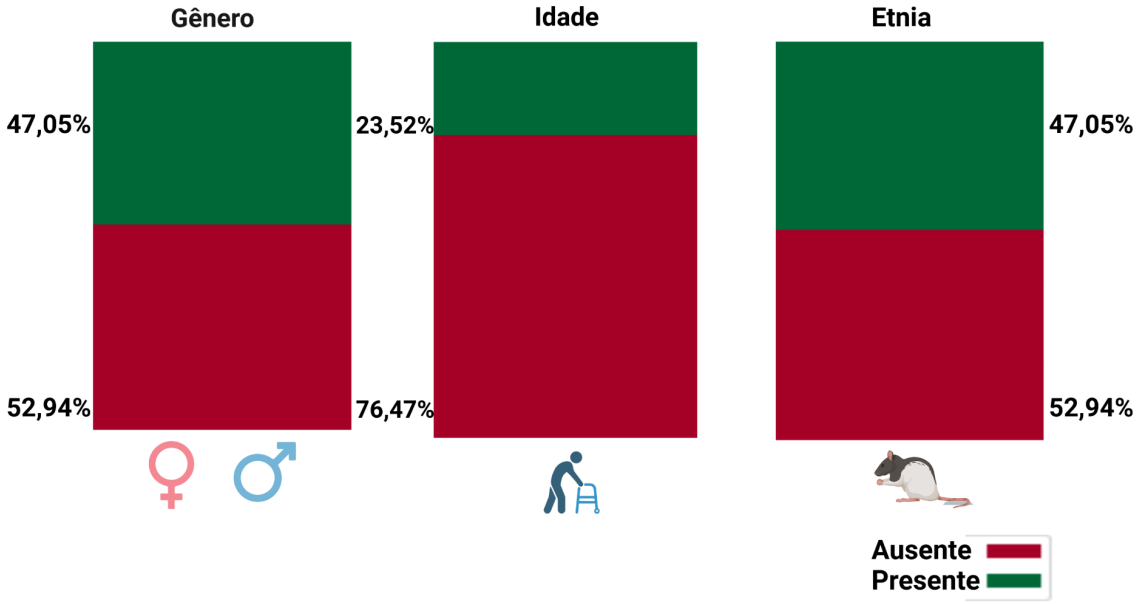


Figura 19: Presença (verde) ou ausência (vermelho) de informação sobre idade, gênero e etnia respectivamente em amostras de *Rattus norvegicus.*

Fonte: AUTORIA PRÓPRIA

Por fim ressalta-se novamente a importância da identificação das respectivas linhagens de animais modelos e a disponibilização de dados brutos de sequenciamento, onde aqui, dos 17 projetos apenas 3 encontravam-se com SRA presente nas respectivas páginas web.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Para auxiliar na compreensão da patologia e da complexidade de causas multifatoriais da DP, diversas ferramentas, tais como bases de dados e recursos de informações integrados têm sido utilizados com mais frequência nessas pesquisas. Uma vez que estas ferramentas são disponibilizadas, compreender a extensão e funcionamento destas, a fim de inclusive tratar de seu aprimoramento, torna-se de suma importância.

As informações disponíveis no GEO tem diversidade quanto aos dados como tecidos, espécies, tipos moleculares, plataformas e tecnologias utilizadas para fins de experimentação. Foi observado, no entanto, uma predominância de espécies tais como humanos e roedores, não muito divergente do encontrado nos trabalhos dos

últimos anos, bem como a predominância de dados de array e de tecidos nervosos, como descrito em trabalhos prévios, além disso mais de 70% dos trabalhos levantados contém dados transcriptômicos oriundos de tipos moleculares com extração predominante de RNA.Os conjuntos de dados vêm de uma variedade de origens genéticas e plataformas com diferentes níveis moleculares, tal variabilidade pode ser considerada um ponto negativo para possíveis meta-análises, especialmente quando não há uma padronização de nomenclatura, onde um mesmo tipo celular ou tecidual encontra-se registrado de diferentes formas.

O mesmo observa-se para a variabilidade na disposição de informações nas páginas web correspondentes aos projetos, o que pode vir a dificultar na acurácia das respectivas classificações e análises, como descreve Lopez-Crisosto *et al.* (2021): “Atualmente, a maioria das revistas científicas precisa carregar os dados brutos em um banco de dados público. Ainda assim, na maioria dos casos, não há uma revisão completa dos metadados da amostra carregada, o que não permite a reprodutibilidade correta e deixa de fora dados primários cruciais, como idade, etnia, gênero, entre outros.”.

Tal observação também ocorre no presente trabalho, uma vez que mais da metade dos trabalhos não apresentavam informação de gênero, idade ou etnia para nenhuma das respectivas espécies descritas. Sendo características como idade e gênero os maiores fatores de impacto para DP, atualmente ressaltados nas recentes pesquisas, a descrição destas nos respectivos projetos e amostras torna-se importante para proporcionar uma acesso amostral de melhor qualidade para demais experimentos e outros fins de pesquisas e colaborações.

Quanto aos trabalhos futuros pretende-se:

- Extrair as respectivas quantidades de amostras femininas e masculinas dos projetos que as tiverem disponíveis para analisá-las;
- Extrair as respectivas quantidades de amostras de controles e casos clínicos de DP dos projetos que as tiverem disponíveis e analisá-las;
- Levantar os principais RNAs não codificantes (ncRNA) envolvidos nos recentes trabalhos em repositórios de dados de Parkinson disponíveis no GEO.

REFERÊNCIAS

BARLOW, Brian K *et al*. A fetal risk factor for Parkinson's disease. **Developmental Neuroscience**, [*s. l*], v. 26, n. 1, p. 11-23, 2004.

BARRETT, Tanya *et al*. NCBI GEO: archive for high-throughput functional genomic data. **Nucleic Acids Research**, [S.L.], v. 37, p. 885-890, 21 out. 2008. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkn764.

BARRETT, Tanya *et al*. NCBI GEO: archive for functional genomics data sets - update. **Nucleic Acids Research**, [S.L.], v. 41, p. 991-995, 26 nov. 2012. Oxford University Press (OUP). https://doi.org/10.1093/nar/gks1193

BONAT, Wagner H.; KRAINSKI, Elias T.; MAYER, Fernando P.. **Introdução à análise exploratória de dados**. Universidade Federal do Paraná: Laboratório de Estatística e Geoinformação, 2018. 34 slides, color.

BRIDI, Jessika C.; HIRTH, Frank. Mechanisms of α-Synuclein Induced Synaptopathy in Parkinson's Disease. **Frontiers In Neuroscience**, [S.L.], v. 12, p. 1-18, 19 fev. 2018. Frontiers Media SA. http://dx.doi.org/10.3389/fnins.2018.00080.

BRUG, M. P. van Der *et al*. RNA binding activity of the recessive parkinsonism protein DJ-1 supports involvement in multiple cellular pathways. **Proceedings Of The National Academy Of Sciences**, [S.L.], v. 105, n. 29, p. 10244-10249, 14 jul. 2008. Proceedings of the National Academy of Sciences. http://dx.doi.org/10.1073/pnas.0708518105.

CANTUTI-CASTELVETRI, Ippolita *et al*. Effects of gender on nigral gene expression and parkinson disease. **Neurobiology Of Disease**, [S.L.], v. 26, n. 3, p. 606-614, jun. 2007. Elsevier BV. http://dx.doi.org/10.1016/j.nbd.2007.02.009.

CARVEY, Paul M.; PUNATI, Ashok; NEWMAN, Mary B.. Progressive Dopamine Neuron Loss in Parkinson's Disease: the multiple hit hypothesis. **Cell Transplantation**, [S.L.], v. 15, n. 3, p. 239-250, mar. 2006. SAGE Publications. http://dx.doi.org/10.3727/000000006783981990.

CHAUDHURI, K. Ray; SAUERBIER, Anna. Unravelling the nonmotor mysteries of Parkinson disease. **Nature Reviews Neurology**, [S.L.], v. 12, n. 1, p. 10-11, 30 dez. 2015. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrneurol.2015.236.

DEMARSHALL, Cassandra A. *et al*. Potential utility of autoantibodies as blood-based biomarkers for early detection and diagnosis of Parkinson’s disease. **Immunology Letters**, [S.L.], v. 168, n. 1, p. 80-88, nov. 2015. Elsevier BV. http://dx.doi.org/10.1016/j.imlet.2015.09.010.

DENG, Hao; WANG, Peng; JANKOVIC, Joseph. The genetics of Parkinson disease. **Ageing Research Reviews**, [S.L.], v. 42, p. 72-85, mar. 2018. Elsevier BV. http://dx.doi.org/10.1016/j.arr.2017.12.007.

DOPAZO, Joaquin *et al*. Methods and approaches in the analysis of gene expression data. **Journal Of Immunological Methods**, [S.L.], v. 250, n. 1-2, p. 93-112, abr. 2001. Elsevier BV. http://dx.doi.org/10.1016/s0022-1759(01)00307-6.

ELBAZ, Alexis *et al*. Professional exposure to pesticides and Parkinson disease. **Annals Of Neurology**, [S.L.], v. 66, n. 4, p. 494-504, 13 abr. 2009. Wiley. http://dx.doi.org/10.1002/ana.21717.

GILLIES, Glenda E. *et al*. Sex differences in Parkinson’s disease. **Frontiers In Neuroendocrinology**, [S.L.], v. 35, n. 3, p. 370-384, ago. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.yfrne.2014.02.002.

GROW, Douglas A.; MCCARREY, John R.; NAVARA, Christopher S.. Advantages of nonhuman primates as preclinical models for evaluating stem cell-based therapies for Parkinson's disease. **Stem Cell Research**, [S.L.], v. 17, n. 2, p. 352-366, set. 2016. Elsevier BV. http://dx.doi.org/10.1016/j.scr.2016.08.013.

JANKOVIC, Joseph; TAN, Eng King. Parkinson’s disease: etiopathogenesis and treatment. **Journal Of Neurology, Neurosurgery & Psychiatry**, [S.L.], v. 91, n. 8, p. 795-808, 23 jun. 2020. BMJ. http://dx.doi.org/10.1136/jnnp-2019-322338.

JANKOVIC, J. Parkinson's disease: clinical features and diagnosis. **Journal Of Neurology, Neurosurgery & Psychiatry**, [S.L.], v. 79, n. 4, p. 368-376, 1 abr. 2008. BMJ. http://dx.doi.org/10.1136/jnnp.2007.131045.

KODAMA, Y.; SHUMWAY, M.; LEINONEN, R.. The sequence read archive: explosive growth of sequencing data. **Nucleic Acids Research**, [S.L.], v. 40, n. 1, p. 54-56, 18 out. 2011. Oxford University Press (OUP). http://dx.doi.org/10.1093/nar/gkr854.

LERANTH, Csaba *et al*. Estrogen Is Essential for Maintaining Nigrostriatal Dopamine Neurons in Primates: implications for parkinson's disease and memory. **The Journal Of Neuroscience**, [S.L.], v. 20, n. 23, p. 8604-8609, 1 dez. 2000. Society for Neuroscience. http://dx.doi.org/10.1523/jneurosci.20-23-08604.2000.

LOPEZ-CRISOSTO, Camila *et al*. Novel molecular insights and public omics data in pulmonary hypertension. **Biochimica Et Biophysica Acta (Bba) - Molecular Basis Of Disease**, [S.L.], v. 1867, n. 10, p. 166200, out. 2021. Elsevier BV. http://dx.doi.org/10.1016/j.bbadis.2021.166200.

MOCELLIN, Simone *et al*. DNA Array-Based Gene Profiling. **Annals Of Surgery**, [S.L.], v. 241, n. 1, p. 16-26, jan. 2005. Ovid Technologies (Wolters Kluwer Health). http://dx.doi.org/10.1097/01.sla.0000150157.83537.53.

MONZÓN-SANDOVAL, Jimena *et al*. Human-Specific Transcriptome of Ventral and

Dorsal Midbrain Dopamine Neurons. **Annals Of Neurology**, [S.L.], v. 87, n. 6, p. 853-868, 30 mar. 2020. Wiley. http://dx.doi.org/10.1002/ana.25719.

PAKKENBERG, Bente; GUNDERSEN, Hans Jorgen G.. Neocortical neuron number in humans: effect of sex and age. **The Journal Of Comparative Neurology**, [S.L.], v. 384, n. 2, p. 312-320, 28 jul. 1997. Wiley. http://dx.doi.org/10.1002/(sici)1096-9861(19970728)384:23.0.co;2-k.

PREECE, Paul; CAIRNS, Nigel J. Quantifying mRNA in postmortem human brain: influence of gender, age at death, postmortem interval, brain ph, agonal state and inter-lobe mrna variance. **Molecular Brain Research**, [S.L.], v. 118, n. 1-2, p. 60-71, out. 2003. Elsevier BV. http://dx.doi.org/10.1016/s0169-328x(03)00337-1.

REEVE, Amy; SIMCOX, Eve; TURNBULL, Doug. Ageing and Parkinson's disease: why is advancing age the biggest risk factor?. **Ageing Research Reviews**, [S.L.], v. 14, p. 19-30, mar. 2014. Elsevier BV. http://dx.doi.org/10.1016/j.arr.2014.01.004.

SCHULZE, Markus *et al*. Sporadic Parkinson’s disease derived neuronal cells show disease-specific mRNA and small RNA signatures with abundant deregulation of piRNAs. **Acta Neuropathologica Communications**, [S.L.], v. 6, n. 1, p. 1-18, 10 jul. 2018. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/s40478-018-0561-x.

STOKERS, T.B., GREENLAND, J.C., editors. (2018) Parkinson’s Disease: Pathogenesis and Clinical Aspects [Internet]. Brisbane (AU): **Codon Publications**. https://doi.org/10.15586/codonpublications.parkinsonsdisease.2018

TACCIOLI, C. *et al*. ParkDB: a parkinson's disease gene expression database. **Database**, [S.L.], v. 2011, p. 1-6, 18 maio 2011. Oxford University Press (OUP). http://dx.doi.org/10.1093/database/bar007.

TARCA, Adi L.; ROMERO, Roberto; DRAGHICI, Sorin. Analysis of microarray experiments of gene expression profiling. **American Journal Of Obstetrics And Gynecology**, [S.L.], v. 195, n. 2, p. 373-388, ago. 2006. Elsevier BV. http://dx.doi.org/10.1016/j.ajog.2006.07.001.

TAYLOR, K s M; A COOK, J; COUNSELL, C e. Heterogeneity in male to female risk for Parkinson's disease. **Journal Of Neurology, Neurosurgery & Psychiatry**, [S.L.], v. 78, n. 8, p. 905-906, 1 ago. 2007. BMJ. http://dx.doi.org/10.1136/jnnp.2006.104695.

WANG, Zhong; GERSTEIN, Mark; SNYDER, Michael. RNA-Seq: a revolutionary tool for transcriptomics. **Nature Reviews Genetics**, [S.L.], v. 10, n. 1, p. 57-63, jan. 2009. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/nrg2484.

YANG, Jin Ok *et al*. PDbase: a database of parkinson's disease-related genes and genetic variation using substantia nigra ests. **Bmc Genomics**, [S.L.], v. 10, n. 3, p. 1-7, 2009. Springer Science and Business Media LLC. http://dx.doi.org/10.1186/1471-2164-10-s3-s32.

YUAN, Weien; MA, Liuqing; WEI, Liangming Liangming; WU, Fei; HU, Zhenhua; LIU, Zhenguo. Advances with microRNAs in Parkinson’s disease research. **Drug Design, Development And Therapy**, [S.L.], p. 1103, out. 2013. Informa UK Limited. http://dx.doi.org/10.2147/dddt.s48500.

ZHENG, B. *et al*. PGC-1 , A Potential Therapeutic Target for Early Intervention in Parkinson's Disease. **Science Translational Medicine**, [S.L.], v. 2, n. 52, p. 1-29, 6 out. 2010. American Association for the Advancement of Science (AAAS). http://dx.doi.org/10.1126/scitranslmed.3001059.

ZHOU, Xiaoguang *et al*. The next-generation sequencing technology and application. **Protein & Cell**, [S.L.], v. 1, n. 6, p. 520-536, jun. 2010. Springer Science and Business Media LLC. http://dx.doi.org/10.1007/s13238-010-0065-3..