Marcelo Luiz Brunatto Falchetti

IDENTIFICAÇÃO DE ASSINATURA GÊNICA PARA CLASSIFICAÇÃO DIAGNÓSTICA DA DOENÇA DE PARKINSON IDIOPÁTICA UTILIZANDO TRANSCRIPTOMAS DE SANGUE PERIFÉRICO E ALGORITMOS DE APRENDIZADO DE MÁQUINA

Dissertação submetida ao Programa de Pós-Graduação em Farmacologia da Universidade Federal de Santa Catarina para a obtenção do Título de Mestre em Farmacologia
Orientador: Prof. Dr. Rui Daniel Prediger
Coorientador: Prof. Dr. Alfeu Zanotto-Filho

Florianópolis
2019

Ficha de identificação da obra elaborada pelo autor através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Falchetti, Marcelo Luiz Brunatto
Identificação de assinatura gênica para classificação diagnóstica da doença de Parkinson idiopática utilizando transcriptomas de sangue periférico e algoritmos de aprendizado de máquina / Marcelo Luiz Brunatto Falchetti ; orientador, Rui Daniel Schröder Prediger, coorientador, Alfeu Zanotto-Filho, 2019.
175 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro de Ciências Biológicas, Programa de Pós-Graduação em Farmacologia, Florianópolis, 2019.

Inclui referências.

1. Farmacologia. 2. Doença de Parkinson. 3. Aprendizado de máquina. 4. Bioinformática. I. Prediger, Rui Daniel Schröder. II. Zanotto-Filho, Alfeu. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Farmacologia. IV. Título.

"Identificação de assinatura gênica para classificação diagnóstica da doença de Parkinson idiopática utilizando transcriptomas de sangue periférico e algoritmos de aprendizado de máquina"

Por

Marcelo Luiz Brunatto Falchetti

Dissertação julgada e aprovada em sua forma final pelos membros titulares da Banca Examinadora (002/2018/PPGFMC) do Programa de Pós-Graduação em Farmacologia - UFSC.

Prof. Dr. Rui Daniel Schröder Prediger
Coordenador(a) do Programa de Pós-Graduação em Farmacologia

Banca examinadora:

Dr. Alfeu Zanotto Filho (Universidade Federal de Santa Catarina)
Coorientador

Dr. Mauro Antônio Alves Castro (Universidade Federal do Paraná)

Dr. Geison de Souza Izídio (Universidade Federal de Santa Catarina)

Florianópolis, 15 de fevereiro de 2019.

Este trabalho é dedicado ao meu pai, Luiz Falchetti.

In memoriam.

AGRADECIMENTOS

Agradeço ao meu orientador Prof. Dr. Rui Daniel Schröder Prediger e coorientador Alfeu Zanotto-Filho pelo ensino, paciência, confiança e orientação durante os meus anos de mestrado.

À banca avaliadora, Prof. Dr. Mauro Antônio Alves Castro, Prof. Dr. Geison de Souza Izídio e a Prof.ª Dr.ª Gabriela Flávia Rodrigues Luiz pela disponibilizade de examinar esse trabalho. Agradeço por poder apresentar e defender esse trabalho para profissionais que tanto respeito.

Ao Programa de Pós-graduação em Farmacologia por ter me recebido e oportunizado tantas conjunturas de crescimento social e intelectual e à CAPES pela indispensável bolsa de estudos.

A todos os meus amigos do LEXDON e do LAB204 pelo carinho, atenção, aprendizado e companherismo nos momentos mais difíceis e mais prazerosos da minha formação no mestrado. Obrigado Ana Elisa Speck, Angela França, Gislaine Olescowicz, Josiel Mack, Katiane Roversi, Marina Magnini, Marissa Schamme, Nei Daniel de Souza Peres, Samantha Lopes e Tuane Sampaio do LEXDON e Jonathan Agnes, Marina Delgobo, Rosângela Gonçalves e Vitória Wibbelt do LAB204.

Obrigado meus amigos e irmãos científicos do mestrado por terem me auxiliado a continuar e chegar nesse ponto, sempre me alegrando e estimulando. Obrigado especialmente Adriane Fagundes, Kalil Campozana, Luan Silva Gomes e Marianna Schneider.

Aos professores, por incitarem minha admiração por essa linda área da ciência. Profissionais que sempre estiveram dispostos a ensinar e crescer juntos. Especialistas que, apaixonados, nos apaixonam diariamente.

Aos outros profissionais motivadores e também responsáveis a minha instrução, como os professores das plataformas Alura, DataCamp, de diversos canais do YouTube, colegas do Stack Overflow, de diversos *blogs* e tantos outros.

A todos os funcionários da Universidade Federal de Santa Catarina que auxiliam e possibilitam a preservação e o exercício dessa instituição. Em especial agradeço os servidores, os profissionais da área da limpeza e do Restaurante Universitário.

Ao meu companheiro Valdriano Polla pelo amor, calma e apoio em todas as minhas escolhas.

E principalmente à minha família pelo amor incondicional, por acreditarem e me estimularem a acreditar que sou capaz de realizar os meus sonhos. Amo todos vocês!

A todos aqueles que contribuiram e aos que ainda contribuem na minha formação como pessoa, como profissional e na minha saúde física e mental. A todos os presentes na minha história. Muito obrigado!

A ciência nos diz que somos bestas, porém não nos sentimos assim. Nos sentimos como anjos presos em corpos de bestas, sempre almejando transcendência
(Vilayanur S. Ramachandran, 2011)

RESUMO

No Brasil, anualmente, mais de 150 mil pessoas são diagnosticadas com a doença de Parkinson (DP). Apenas alguns casos da DP são hereditários e atribuídos a mutações em genes, enquanto a vasta maioria (~ 90%) é classificada como DP idiopática. O diagnóstico da DP ainda é um desafio na prática clínica, e a identificação de marcadores moleculares para diagnóstico e acompanhamento pode proporcionar um tratamento mais eficaz para os pacientes. Uma opção fácil e não invasiva é a biópsia líquida. A utilização de técnicas de biologia molecular de alto rendimento vinculadas as metodologias de meta-análise podem contribuir na busca de assinaturas transcricionais com potencial aplicação diagnóstica para a DP idiopática. Este estudo teve como objetivo caracterizar as diferenças no perfil transcricional do sangue periférico de pacientes com a DP e indivíduos sadios, e identificar genes capazes de auxiliar no diagnóstico da DP, utilizando algoritmos de aprendizado de máquina (AAM). Todas as análises foram realizadas em ambiente de análises estatísticas e gráficas R. Para isso, foram utilizados os dados de microarranjo de expressão obtidos em repositórios públicos (GEO e ArrayExpress). Foram selecionados 4 conjuntos de dados independentes incluindo 711 amostras no total, sendo 388 de sangue de pacientes com a DP idiopática e 323 de indivíduos sadios. Foram realizadas meta-análises de 17.712 genes calculando e combinando os seus tamanhos de efeito. Os 200 genes com os maiores valores de tamanho de efeito, ou seja, os de maior distinção em expressão entre os grupos analisados apresentam ontologias relacionadas ao sistema imune e imunidade mediada pelos leucócitos, poliubiquitinação de proteínas e organização do citoesqueleto. Os 200 genes não são capazes de separar as amostras de DP idiopático e indivíduos sadios quando individualmente avaliados por agrupamentos hierárquicos. Para caracterização de uma assinatura gênica da DP idiopática, foram utilizados algoritmos de seleção de preditores de reconhecimento de colinearidades e de eliminação de preditores recursivo. Para as análises de predição, as amostras foram separadas em dois grupos, o grupo-treino (GTr), utilizado para contruir e ajustar os modelos, e o grupo-teste (GTe), para estimar os desempenhos dos modelos. Para a correção do desbalanço amostral no GTr foram utilizados métodos de criação de amostras sintéticas para arquitetar mais 3 GTr equilibrados. Para criação dos modelos de classificação foram utilizados 9 AAM ajustados com um total de 18 hiperparâmetros. Por fim, foram eleitas as combinações de modelos e ajustes que dispunham dos maiores valores de áreas sob a curva ROC (AUC) para cada GTr. Dessa forma,

foram selecionados 30 modelos capazes de classificar as amostras com AUC de 0,72 a 0,80. Para otimização dos resultados, foram calculadas as probabilidades de classe de amostras do GTe nos modelos com os maiores valores de sensibilidade e especificidade criados com cada GTr, e seguidamente filtradas as 25% com as menores probabilidades. Dessa forma, foram estabelecidos os valores mínimos de probabilidade para classificação e selecionados os modelos com os maiores valores mínimos. Os modelos escolhidos pós-otimização apresentaram 0,84 de sensibilidade e 0,88 de especificidade. A média de probabilidade de amostras serem de DP no modelo de maior sensibilidade é diferente de grupos de amostras de indivíduos sadios, da doença de Huntington e de formas genéticas da DP causadas por mutações nos genes *LRRK2* e *PRKN*, porém análogas de grupos da doença de Alzheimer, da atrofia multissistêmica, da paralisia supranuclear progressiva e de formas genéticas da DP causadas por mutações nos genes *ATP13A2* e *PINK1*. Elaborou-se uma sequência de operações alternativa para classificações de imagens que busca valorizar dados colineares. O modelo gerado utilizando este processo apresentou 84% de acertos. Apesar da variação amostral quanto aos tratamentos farmacológicos, idades, severidade da doença, a metodologia aplicada neste trabalho foi capaz identificar uma assinatura e modelos confiáveis na classificação da DP idiopática em amostras de sangue, o que pode fornecer base bioinformática para futuras otimizações.

Palavras-chave: Bioinformática, Neurociências computacional.

ABSTRACT

In Brazil, over 150,000 people are diagnosed annually with Parkinson's disease (PD). Only a few cases of PD are inherited and attributed to mutations in genes, while the vast majority (~ 90%) are classified as idiopathic PD. The diagnosis of PD is still a challenge in clinical practice, and the identification of molecular markers for diagnosis and follow-up may provide a more effective treatment for patients. An easy and noninvasive option is liquid biopsy. The use of high-throughput molecular biology techniques related to meta-analysis methodologies may contribute to the search for transcriptional signatures with potential diagnostic application for idiopathic PD. Therefore, this study aimed to characterize the differences in the transcriptional profile of peripheral blood of patients with PD and healthy individuals, and to identify genes capable of assisting in the PD diagnosis using machine learning algorithms (MLA). All the analyzes were carried out in the environment of statistical and graphic analysis R. For this, we used the gene expression microarray data obtained in public repositories (GEO and ArrayExpress). We identified 4 independent datasets including 711 samples in total, 388 of them were blood from patients with idiopathic PD and 323 from healthy individuals. Meta-analyzes of 17,712 genes were performed calculating and combining their effect sizes. The 200 genes with the highest effect size values, that is, those with the highest distinction in expression between the analyzed groups present ontologies related to the immune system and leukocyte mediated immunity, protein polyubiquitination and cytoskeleton organization. The 200 genes are not able to separate idiopathic PD samples and healthy individuals when individually assessed by hierarchical clustering. Algorithms of collinearity recognition and recursive predictor elimination were used to to characterize a gene signature of idiopathic PD. For the prediction analyzes, the samples were separated into two groups, the training group (TrG), used to construct and adjust the models, and the test group (TeG), to estimate the performance of models. For the correction of the sample, unbalance in the TrG, synthetic sample creation methods were used to architect three more balanced TrG. To create the classification models, 9 MLA were tuned with a total of 18 hyperparameters. Finally, the combinations of models and tunings that had the highest values of areas under the ROC curve (AUC) were chosen for each TrG. Thus, 30 models capable of classifying samples with AUC from 0.72 to 0.80 were selected. In order to optimize the results, the TeG sample class probabilities were calculated in the models with the highest values of sensitivity and

specificity created with each TrG and then filtered the 25% with the lowest probabilities. This way, the minimum probability values for classification were established and the models with the highest minimum values were selected. The models chosen post-optimization presented 0.84 of sensitivity and 0.88 of specificity. The average probability of the samples being of PD in the highest sensitivity model is different from groups of samples from healthy individuals, from Huntington's disease and from PD genetic forms caused by mutations in the *LRRK2* and *PRKN* genes. However, they seem to be analogous to the disease groups of Alzheimer's disease, multiple systemic atrophy, progressive supranuclear palsy and genetic forms of PD caused by mutations in the *ATP13A2* and *PINK1* genes. An alternative sequence of operations for classifying images that seeks to value collinear data was developed. The model generated using this process presented 84% of correct answers. Despite the sample variation in pharmacological treatments, ages, and disease severity, the methodology applied in this study was able to identify a reliable signature models for the classification of idiopathic PD in blood samples, which may provide a bioinformatic basis for future optimizations.

Keywords: Bioinformatics. Computational neuroscience.

LISTA DE FIGURAS

Figura 1 Pirâmide etária do Brasil em 2018 e as projeções para 2040 e 2060. 32

Figura 2 Neurodegeneração na SN em pacientes com a doença de Parkinson. 33

Figura 3 Disposição de sondas em microarranjos de DNA. 40

Figura 4 *Grid* de pontos contendo sondas de um *chip* de oligonucleotídeos. 41

Figura 5 Organograma de cálculos de tamanhos de efeito para meta-análise. 58

Figura 6 Resultados das estratégias de balanceamento de dados implementadas no pacote *ROSE* utilizando os dados de *hacide*. 70

Figura 7 Validação cruzada de 10-*fold*. 71

Figura 8 Organograma da preparação de dados para as análises de aprendizado de máquina. 73

Figura 9 Comparação de diferentes médias e o efeito do limiar de corte de 25% das amostras. 81

Figura 10 Organograma do processo de busca por dados de microarranjo de DNA. 88

Figura 11 Gene ID únicos distintos identificados em conjuntos de dados representados neste trabalho. 90

Figura 12 Idades, sexos e severidade da doença de amostras em diferentes trabalhos. 91

Figura 13 Tamanhos de efeito e intervalo de confiança dos GSP e GSN com os 5 maiores valores de diferença de classes. 93

Figura 14 GDE resultantes das reanálises de expressão diferencial individuais de cada conjunto de dados. 100

Figura 15 *Volcano plots* de reanálises de expressão diferencial individuais e da dispersão de GSP e GSN em cada estudo. 102

Figura 16 Cladograma de amostras com base nos valores de expressão dos genes selecionados. 103

Figura 17 Total de amostras em cada GTr separado por fenótipo. 108

Figura 18 Total de genes selecionados contidos em cada conjunto de dados. 109

Figura 19 Correlação gênica de todos os genes selecionados presentes em mais de um conjunto de dados e destes após a eliminação de informações redundantes. 110

Figura 20 Organograma de criação de modelos de classificação e seleção de valores de hiperparâmetros. 114

Figura 21 Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr. 123

Figura 22 Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr superamostragem. 124

Figura 23 Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr subamostragem. 125

Figura 24 Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr combinados. 126

Figura 25 Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr. 127

Figura 26 Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr superamostragem. 128

Figura 27 Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr subamostragem. 129

Figura 28 Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr combinados. 130

Figura 29 Modelos escolhidos com os limites de maior valor de probabilidade de classe. 131

Figura 30 Probabilidades de predição na classe *Parkinson* de amostras de outras categorias. 132

Figura 31 Estruturas da rede de GSP. 135

Figura 32 Estruturas da rede de GSN. 136

Figura 33 Estruturas combinadas das redes de GSP e GSN. 137

Figura 34 Exemplos de topologias de redes geradas com o ViaComplex. 138

LISTA DE QUADROS

Quadro 1 Estudos utilizados na meta-análise. 89

Quadro 2 Genes selecionados com os 100 maiores valores de tamanhos de efeito positivos e negativos. 94

Quadro 3 Termos de ontologias enriquecidos com os GSP. 104

Quadro 4 Termos de ontologias enriquecidos com os GSN. 105

Quadro 5 Vias de interação molecular enriquecidas com os GSP. 106

Quadro 6 Vias de interação molecular enriquecidas com os GSN. 107

Quadro 7 Assinatura gênica. 111

Quadro 8 Hiperparâmetros selecionados. 114

LISTA DE TABELAS

Tabela 1 Comparação de limiares para as famílias *r* e *d* de tamanhos de efeito. 60

Tabela 2 Predições com os modelos criados com o GTr. 119

Tabela 3 Predições com o modelos criados com o GTr superamostragem. 120

Tabela 4 Predições com os modelos criados com o GTr subamostragem. 121

Tabela 5 Predições com os modelos criados com o GTr combinados. 122

LISTA DE ABREVIATURAS E DE SIGLAS

DCNT: Doenças Crônicas Não Transmissíveis
DA: Dopamina
DNA: Ácido desoxirribonucleico
DP: Doença de Parkinson
GEA: *Gene Enrichment Analysis*
GEO: *Gene Expression Omnibus*
GSE: Séries de genes do GEO
GSEA: *Gene Set Enrichment Analysis*
GSP: Genes selecionados positivos
GSN: Genes selecionados negativos
GTe: Grupo de teste
GTr: Grupo de treino
IBGE: Instituto Brasileiro de Geografia e Estatística
KEGG: *Kyoto Encyclopedia of Genes and Genomes*
L-DOPA: L-dihidroxifenilalanina
MIAME: Informações mínimas sobre os experimentos de microarranjo
mRNA: RNA mensageiro
ONU: Organização das Nações Unidas
PB: Processos biológicos
RMA: *Robust Multi-array Average*
RNA: Ácido ribonucleico
SN: *Substantia nigra*
SNC: Sistema Nervoso Central

SUMÁRIO

1 INTRODUÇÃO 31
1.1 A DOENÇA DE PARKINSON 31
1.1.1 Descrição e principais alterações neuropatológicas 31
1.1.2 Epidemiologia 34
1.1.3 Etiologia 35
1.1.4 Tratamento farmacológico 37
1.1.5 Diagnóstico diferencial 37
1.2 MICROARRANJOS DE DNA 39
1.2.1 A tecnologia 39
1.2.2 Medidas de expressão 41
1.2.3 Processamento analítico de microarranjos de DNA 42
1.2.3.1 Pré-processamento de dados 42
1.2.3.3 Processamento de dados 44
1.2.3.3.1 Comparação de classes 45
1.2.3.3.2 Análise de vias 46
1.2.3.3.3 Descoberta de classes 47
1.2.3.3.4 Predição de classes 48
1.2.3 Meta-análise de microarranjos de DNA 49
2 HIPÓTESE 51
3 OBJETIVOS 53
3.1 OBJETIVO GERAL 53
3.2 OBJETIVOS ESPECÍFICOS 53
4 MATERIAL E MÉTODOS 55
4.1 BUSCA E CRITÉRIOS DE INCLUSÃO E EXCLUSÃO DE CONJUNTOS DE DADOS 55
4.2 IMPORTAÇÃO E PRÉ-PROCESSAMENTO DE CONJUNTOS DE DADOS 55
4.3 META-ANÁLISE DE DADOS DE MICROARRANJO 57
4.3.1 Cálculos de tamanhos de efeito individuais 59

4.3.2 Cálculos de tamanhos de efeito agrupados 60

4.3.2.1 Cálculos de tamanhos de efeito médios 60

4.3.2.2 Cálculos das significância estatísticas dos tamanhos de efeito médio ponderado 61

4.3.2.3 Exame das variabilidades das distribuições dos estimados de tamanhos de efeito 63

4.4 CARACTERIZAÇÕES FUNCIONAIS UTILIZANDO TESTES HIPERGEOMÉTRICOS 64

4.5 REANÁLISE DE EXPRESSÃO DIFERENCIAL E ANÁLISE DE AGRUPAMENTO 65

4.6 ALGORITMOS DE APRENDIZADO DE MÁQUINA 67

4.6.1 Pré-processamento dos dados 67

4.6.2 Preparação dos dados 67

4.6.3 Algoritmos de aprendizado de máquina 72

4.6.3.1 *k-Nearest Neighbors* 72

4.6.3.2 *Naive Bayes* 74

4.6.3.3 *Decision Trees* 75

4.6.3.4 *Support Vector Machine* 76

4.6.3.5 *Bagging* 77

4.6.3.6 *Random Forest* 77

4.6.3.7 *Gradient Boosting Machine* 78

4.6.3.8 *eXtreme Gradient Boosting* 79

4.6.4 Avaliação e seleção de modelos criados 80

4.6.4 Testes para outras condições 82

4.7 ELABORAÇÃO DE REDES DE CORRELAÇÃO GÊNICA E OBTENÇÃO DE REDES DE CO-EXPRESSÃO GÊNICA 82

4.8 ELABORAÇÃO DE UM MODELO DE CLASSIFICAÇÃO BASEADO EM VALORES DE CORRELAÇÃO GÊNICA 84

5 RESULTADOS 87

5.1 BUSCA DE CONJUNTOS DE DADOS DE MICROARRANJOS DE DNA 87

5.2 DESCRIÇÃO DE CONJUNTOS DE DADOS 87

5.3 DADOS CLÍNICOS DE CONJUNTOS DE DADOS 90

5.4 META-ANÁLISE E SELEÇÃO DE GENES 92

5.5 REANÁLISES DE EXPRESSÕES DIFERENCIAIS DE CADA CONJUNTO DE DADOS 100

5.6 ANÁLISE FUNCIONAL DE GENES SELECIONADOS 104

5.7 PREDITORES PARA ASSINATURA GÊNICA 107

5.8 COMPETÊNCIA PREDITIVA DE MODELOS DE CLASSIFICAÇÃO EM AMOSTRAS DE SANGUE DE PACIENTES COM PARKINSON IDIOPÁTICA 113

5.9 REFINAMENTO PARA APLICAÇÃO DE MODELOS 115

5.10 APLICAÇÃO DE MODELOS EM EXEMPLARES DE OUTRAS CATEGORIAS 118

5.11 COMPETÊNCIA PREDITIVA DO MODELO DE CLASSIFICAÇÃO BASEADO EM IMAGENS E ENRIQUECIDO POR COLINEARIDADE 133

6 DISCUSSÃO 139

7 PRINCIPAIS RESULTADOS 155

8 CONCLUSÃO E PERSPECTIVAS 157

8.1 CONCLUSÃO 157

8.2 PERSPECTIVAS 157

REFERÊNCIAS 159

APÊNDICE A –ARTIGOS OBTIDOS NA BUSCA NO PUBMED

APÊNDICE B –CONJUNTOS DE DADOS OBTIDOS NO GEO

APÊNDICE C –CONJUNTOS DE DADOS OBTIDOS NO ARRAYEXPRESS

APÊNDICE D –GENE ID ÚNICOS NO CONJUNTOS DE DADOS GSE6613

APÊNDICE E –GENE ID ÚNICOS NO CONJUNTOS DE DADOS GSE57475

APÊNDICE F –GENE ID ÚNICOS NO CONJUNTOS DE DADOS GSE72267

APÊNDICE G –GENE ID ÚNICOS NO CONJUNTOS DE DADOS GSE99039

APÊNDICE H –GENE ID ÚNICOS DISTINTOS PRESENTES EM TODOS OS CONJUNTOS DE DADOS
APÊNDICE I –GENE ID PRESENTES EM MAIS DE UM CONJUNTO DE DADOS
APÊNDICE J –DADOS CLÍNICOS DO CONJUNTO DE AMOSTRAS DE GSE6613
APÊNDICE K –DADOS CLÍNICOS DO CONJUNTO DE AMOSTRAS DE GSE57475
APÊNDICE L –DADOS CLÍNICOS DO CONJUNTO DE AMOSTRAS DE GSE72267
APÊNDICE M –DADOS CLÍNICOS DO CONJUNTO DE AMOSTRAS DE GSE99039
APÊNDICE N –TAMANHOS DE EFEITO DE GENE ID ÚNICOS DO CONJUNTOS DE DADOS GSE6613
APÊNDICE O –TAMANHOS DE EFEITO DE GENE ID ÚNICOS DO CONJUNTOS DE DADOS GSE57475
APÊNDICE P –TAMANHOS DE EFEITO DE GENE ID ÚNICOS DO CONJUNTOS DE DADOS GSE72267
APÊNDICE Q –TAMANHOS DE EFEITO DE GENE ID ÚNICOS DO CONJUNTOS DE DADOS GSE99039
APÊNDICE R –TAMANHOS DE EFEITO AGRUPADOS DE GENE ID PRESENTES EM MAIS DE UM CONJUNTO DE DADOS
APÊNDICE S –TERMOS DE ONTOLOGIAS DE PROCESSOS BIOLÓGICOS ENRIQUECIDOS COM OS GSP UTILIZANDO O MÉTODO GOANA/LIMMA
APÊNDICE T –TERMOS DE ONTOLOGIAS DE PROCESSOS BIOLÓGICOS ENRIQUECIDOS COM OS GSP UTILIZANDO O MÉTODO DAVID
APÊNDICE U –TERMOS DE ONTOLOGIAS DE PROCESSOS BIOLÓGICOS ENRIQUECIDOS COM OS GSN UTILIZANDO O MÉTODO GOANA/LIMMA
APÊNDICE V –TERMOS DE ONTOLOGIAS DE PROCESSOS BIOLÓGICOS ENRIQUECIDOS COM OS GSN UTILIZANDO O MÉTODO DAVID
APÊNDICE W –VIAS DE INTERAÇÃO MOLECULAR ENRIQUECIDAS COM OS GSP UTILIZANDO O MÉTODO GOANA/LIMMA
APÊNDICE X –VIAS DE INTERAÇÃO MOLECULAR ENRIQUECIDAS COM OS GSP UTILIZANDO O MÉTODO DAVID

APÊNDICE Y –VIAS DE INTERAÇÃO MOLECULAR ENRIQUECIDAS COM OS GSN UTILIZANDO O MÉTODO GOANA/LIMMA
APÊNDICE Z –VIAS DE INTERAÇÃO MOLECULAR ENRIQUECIDAS COM OS GSN UTILIZANDO O MÉTODO DAVID
APÊNDICE AA –CORRELAÇÕES ENTRE OS PARES DE GSP DO GSE6613
APÊNDICE AB –CORRELAÇÕES ENTRE OS PARES DE GSP DO GSE57475
APÊNDICE AC –CORRELAÇÕES ENTRE OS PARES DE GSP DO GSE72267
APÊNDICE AD –CORRELAÇÕES ENTRE OS PARES DE GSP DO GSE99039
APÊNDICE AE –CORRELAÇÕES ENTRE OS PARES DE GSN DO GSE6613
APÊNDICE AF –CORRELAÇÕES ENTRE OS PARES DE GSN DO GSE57475
APÊNDICE AG –CORRELAÇÕES ENTRE OS PARES DE GSN DO GSE72267
APÊNDICE AH –CORRELAÇÕES ENTRE OS PARES DE GSN DO GSE99039
APÊNDICE AI –CORRELAÇÕES AGRUPADAS ENTRE OS PARES DE GSP DE TODOS OS CONJUNTOS DE DADOS
APÊNDICE AK –CORRELAÇÕES AGRUPADAS ENTRE OS PARES DE GSP ELEMENTOS DA REDE
APÊNDICE AL –CORRELAÇÕES AGRUPADAS ENTRE OS PARES DE GSN ELEMENTOS DA REDE

1 INTRODUÇÃO

A humanidade vem passando, desde a metade do século XX, por uma transição demográfica. Até 2050 o número de pessoas idosas chegará a 2 bilhões e este valor representará mais do que o dobro das 900 milhões de pessoas com essa mesma faixa etária registrada em 2015. Dessa forma, em 2050 20% da população mundial será formada por pessoas com idade igual ou superior a 60 anos (ONU, 2017). Esse envelhecimento populacional está diretamente relacionado com a diminuição das taxas de fecundidade e de natalidade e aos esforços em melhorar as condições de vida da população (IBGE, 2012).

Até pouco tempo o Brasil era considerado um país de jovens e o envelhecimento um fenômeno apenas de países desenvolvidos. Entretanto essa realidade está se transformando (DE BARROS; JUNIOR, 2013). No Brasil, espera-se uma população com 35 milhões de pessoas com idade superior a 60 anos no ano de 2025, colocando o país em 6º lugar no mundo no número de indivíduos idosos (IBGE, 2010).

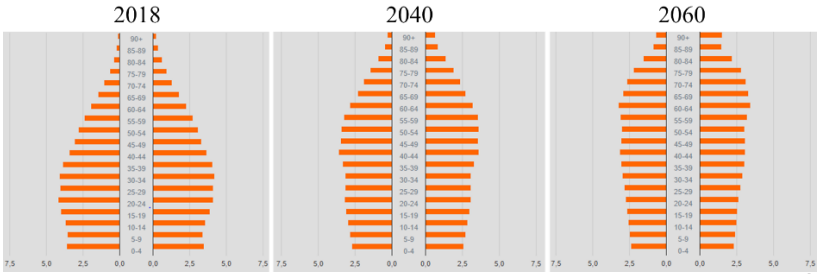
As pirâmides etárias ilustram o envelhecimento populacional, comparando a projeção entre as décadas de 2040 e 2060 com o ano de 2018 (IBGE, 2018) (**Figura 1**). Observa-se uma transição dos formatos das pirâmides, passando de piramidal, em 2018, com uma base larga para retangular devido ao aumento no número de idosos. Embora esse efeito demonstre um avanço na qualidade de vida de parte da população, é também um alerta para a possibilidade de que em um futuro próximo cresça significativamente o número de idosos acometidos pelos chamados "males da idade"(MCGEER; MCGEER, 2004), entre os quais destacam-se as doenças crônicas, sendo as de maior prevalência as classificadas como doenças crônicas não transmissíveis (DCNT). As DCNT estão associadas com as mudanças no estilo de vida da população e o aumento da expectativa de vida (UNWIN; ALBERTI, 2006). Portanto, o envelhecimento é o maior fator de risco para diversas doenças humanas, incluindo as doenças neurodegenerativas. Entre as doenças neurodegenerativas relacionadas a idade, encontra-se a doença de Parkinson (DP) (POEWE et al., 2017).

1.1 A DOENÇA DE PARKINSON

1.1.1 Descrição e principais alterações neuropatológicas

A DP foi descrita pela primeira vez como uma doença neuropatológica em 1817, pelo médico inglês James Parkinson na mono-

Figura 1: Pirâmide etária do Brasil em 2018 e as projeções para 2040 e 2060.



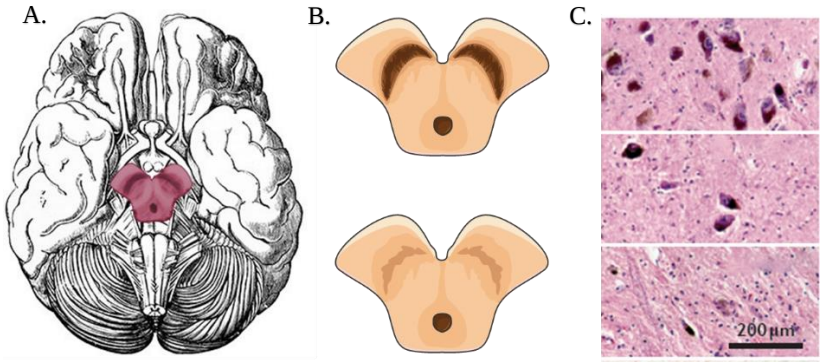
Fonte: Fundação Instituto Brasileiro de Geografia e Estatística – IBGE, 2018. À esquerda: homens. À direita: mulheres.

grafia denominada “An assay on the shaking palsy” (do inglês, “Um ensaio da paralisia agitante”) (PARKINSON, 1817). Neste estudo foram descritos os sintomas motores de 6 pacientes que apresentavam uma associação particular consistindo em tremores em repouso, lentidão de movimentos (bradicinesia), perda da capacidade de mover os músculos voluntariamente, postura curvada e marcha festinante, porém com o intelecto e os sentidos intactos.

Aproximadamente 50 anos depois, em 1872, o médico francês Jean-Martin Charcot adicionou importantes contribuições ao que se sabia sobre a DP, distinguindo-a de outras formas de parkinsonismo ou condições que causavam tremores, além de propor a primeira opção ao tratamento farmacológico da DP utilizando alcaloides derivados da *Atropa belladonna* (agentes anticolinérgicos, antagonistas de receptores muscarínicos para acetilcolina) (CHARCOT, 1872). Charcot também renomeou a doença em homenagem a James Parkinson, foi responsável por apontar a existência de prejuízos cognitivos e nos sentidos dos pacientes e por descrever os sinais cardinais da doença: a rigidez muscular, a bradicinesia e os tremores de repouso, os quais são utilizados até os dias de hoje no diagnóstico da doença (CHARCOT, 1872).

As alterações neuropatológicas associadas à *substantia nigra* (SN) na DP foram relatadas primeiramente em 1919 pelo neuropatologista Constantin Trétiakoff. Ele observou uma despigmentação na região da SN e a atribuiu a uma perda de neurônios ricos em neuromelanina (**Figura 2**); além de perceber microscopicamente a presença de estruturas atualmente nomeadas como corpos de Lewy (TRÉTIAKOFF, 1919). Os corpos de Lewy foram descritos em 1913 pelo neurologista Frederick Lewy como inclusões citoplasmáticas eosinofílicas em neurônios em degeneração (LEWY, 1913). Juntamente, as observações de Trétiakoff

Figura 2. Neurodegeneração na SN em pacientes com a doença de Parkinson.



A. Esquema do corte axial do encéfalo apresentando a localização da *substantia nigra* do mesencélo **B**. Esquema do corte axial do mesencéfalo de indivíduos sadios e de pacientes com a doença de Parkinson, evidenciando a degeneração de neurônios na *substantia nigra*. **C**. Coloração de hematoxilina e eosina da região ventrolateral da *substantia nigra* retratanto a distribuição normal de neurônios pigmentados em indivíduos sadios e pacientes com a doença Parkinson moderado e severo. Fonte: Parte A: retirado de http://bodytomy.com/. Parte C: retirada de Poewe e colaboradores (2017).

sobre a degeneração da SN e as inclusões dos corpos de Lewy são até hoje as principais alterações neuropatológicas da DP, utilizadas em análises de imagem e para diagnóstico *post-mortem* (HARTMANN, 2004; LOANE; POLITIS, 2011).

As evidências da correlação entre a despigmentação/perda neuronal na SN ressaltada por Trétiakoff e a apresentação de sintomas motores clássicos começaram a ser evidenciadas e publicadas na década de 1950. Nesse período, Arvid Carlsson e colaboradores publicaram um estudo relatando a evidência do papel da dopamina (DA) no sistema nervoso central (SNC) (CARLSSON; LINDQVIST; MAGNUSSON, 1957). Neste trabalho ficou demonstrado o papel de neurotransmissor da DA além da observação de que em coelhos administrados com reserpina (um depletor de monoaminas), a administração do precursor da DA, a L-dihidroxifenilalanina (L-DOPA) foi capaz de reverter as reduções de DA na região do estriado e melhorar a atividade motora desses animais. Esses achados evidenciaram o envolvimento de núcleos da base, principalmente da via da SN para o estriado (nigroestriatal), no controle motor por sinalização dopaminérgica (PRZEDBORSKI, 2017).

Os corpos de Lewy são compostos majoritariamente por agregados de proteínas α-sinucleína e ubiquitina. Estes acúmulos podem surgir devido a falhas em mecanismos de proteostase da α-sinucleina por uma superprodução desta, pela presença de mutações facilitadoras de oligomerizações, má formação da proteína ou degradação proteica (NALLS et al., 2014; SOLDNER et al., 2016; VEKRELLIS et al., 2011). Os mecanismos de degradação da α-sinucleina envolvem os sistemas ubiquitina-proteossoma e o sistema de autofagia lisossomal (BRUNDIN et al., 2008; XILOURI; BREKK; STEFANIS, 2013). Disfunções em qualquer um desses sistemas pode induzir um aumento da concentração de α-sinucleína. Entre os fatores de risco para tais disfunções está o envelhecimento e, de fato, encontra-se os corpos de Lewy em cérebros de idosos sadios (CHU; KORDOWER, 2007; KAUSHIK; CUERVO, 2015).

Outros mecanismos neuropatológicos contribuem para o progresso neurodegenerativo associado a DP, como o estresse oxidativo, a neuroinflamação e a excitotoxicidade glutamatérgica que são geralmente associados ao envelhecimento, mas que são acentuados na DP (LEE MOSLEY et al., 2006; SUBRAMANIAM; CHESSELET, 2013). O estresse oxidativo pode levar à morte neuronal por ativação da via das caspases e, consequentemente, ativar a micróglia, que por sua vez produz e libera mediadores imunes e citocinas inflamatórias, aumentando a neurotoxicidade e ampliando os processos neurodegenerativos (LEE MOSLEY et al., 2006). A excitotoxicidade é outro mecanismo envolvido com a fisiopatologia da DP e ocorre devido ao excesso de glutamato na fenda sináptica, desencadeando uma atividade neuronal excessiva pela ativação contínua e exacerbada de seus receptores, principalmente os do tipo N-metil-D-aspartato (NMDA) estimulando um processo neurotóxico. Neste cenário, a neurotransmissão excitatória acontece de modo descontrolado, induzindo a morte neuronal, alterando a excitabilidade de membrana, a homeostase iônica e disparando cascatas químicas associadas ao processo neurodegenerativo (STOLL et al., 2007).

1.1.2 Epidemiologia

A incidência da DP varia de 5 a 35 novos casos para cada 100 mil indivíduos por ano (TWELVES; PERKINS; COUNSELL, 2003). Essa variação pode ser decorrente das diferentes populações estudadas e dos desenhos dos estudos. Esses dados fazem com que a DP seja a segunda doença neurodegenerativa relacionada a idade mais comum na atualidade, ficando atrás somente da doença de Alzheimer (POEWE et al., 2017). A

DP é rara em indivíduos de idade inferior a 50 anos, porém a sua incidência cresce de 5 a 10 vezes da sexta para a nona década de vida (SAVICA et al., 2013; TWELVES; PERKINS; COUNSELL, 2003). A prevalência global da doença é de aproximadamente 0,3%, porém aumenta abruptamente para 3% em indivíduos com idade superior a 80 anos (PRINGSHEIM et al., 2014). No Brasil, anualmente, mais de 150 mil pessoas são diagnosticadas com a DP. Um estudo de 2006 realizado em Bambuí, Minas Gerais, por Barbosa e colaboradores, apontou que 3,4% dos indivíduos com idade superior a 60 anos possuíam diagnóstico da DP (BARBOSA et al., 2006).

A incidência parece variar em subgrupos definidos por gênero, genótipos, etnias e ambientes. Em diversas populações, a DP é duas vezes mais comum em homens do que em mulheres (BALDERESCHI et al., 2000; VAN DEN EEDEN et al., 2003). Essa característica epidemiológica parece estar menos relacionada a um efeito protetor de hormônios sexuais femininos ou genética do gênero, e mais a fatores socioambientais como exposição a agrotóxicos e traumatismo cranioencefálico (POEWE et al., 2017). Na população de judeus Ashkenazi, em Israel, a prevalência da DP é alta, possivelmente refletindo as altas taxas de prevalência de genes associados com a DP, como o *LRRK2* (que codifica a proteína serina/treonina repetida proteína kinase rica em leucina 2 ou *leucine-rich repeat serine/threonine-protein kinase 2* em inglês) e o *GBA* (que codifica a proteína glicocerebrosidade) (CHILLAG-TALMOR et al., 2011). Apesar de existirem diferenças relatadas na literatura quanto a incidência em alguns grupos étnicos, outras variáveis que podem explicar essas diferenças, como multirracialidade e fatores sociais geralmente são pouco explorados (VAN DEN EEDEN et al., 2003).

1.1.3 Etiologia

A etiologia da DP ainda é desconhecida, apesar de esforços de pesquisa que movimentam diversos grupos. O principal fator de risco para o desenvolvimento da DP é o envelhecimento, entretanto acredita-se que o desenvolvimento da DP seja resultante da combinação de fatores ambientais, genéticos e epigenéticos (GORELL et al., 2004; LILL, 2016).

Entre os fatores de risco ambientais, a DP é atribuída a exposição a contaminantes ambientais, os herbicidas, fungicidas e pesticidas como o paraquat, a rotenona e o maneb, sendo que estas substâncias são utilizadas para modelar a DP em roedores (CICCHETTI; DROUIN-OUELLET; GROSS, 2009; SAMPAIO et al., 2017). A exposição a

metais não metabolizáveis, reativos e bioacumuláveis como o ferro, alumínio, cobre, zinco e manganês também parece estar relacionada a etiologia da DP (BJORKLUND et al., 2018). A exposição a esses compostos e o traumatismo cranioencefálico estão entre os principais fatores etiológicos associados à DP idiopática (sem etiologia conhecida, neste caso sem base genética definida) (LEE et al., 2012).

Entre os fatores genéticos, ainda não são conhecidos os mecanismos exatos relacionados a perda de neurônios dopaminérgicos da SN na DP, embora seja reconhecida a associação do desenvolvimento da DP em famílias e populações (POEWE et al., 2017). Os últimos 20 anos de pesquisas genéticas demonstraram que certas sequências (variantes) nucleotídicas participam ativamente do desenvolvimento da DP (KLEIN; WESTENBERGER, 2012; LILL, 2016). Aproximadamente, de 5 a 10% de pacientes apresentam a forma monogênica da DP, enquanto a maior parte dos casos parece estar relacionada com uma interação de diversas variações na sequência do DNA (LESAGE; BRICE, 2009). O espectro de variantes genéticas relativas a etiologia da DP varia de sequências raras penetrantes a sequências comuns de efeito moderado. Atualmente, há 26 variantes genéticas relacionadas com a DP descritas, divididas entre (1) genes autossômicos dominantes como o *SNCA* (que codifica a proteína α-sinucleína), *LRRK2*, e *VPS35,* (2) genes autossômicos recessivos como o *PRKN*, *PARK7*, *PINK1*, *DNAJC6* e (3) genes predominantemente causadores da DP atípica, como o *ATP13A2*, *PLA2G6*, *FBX07* e outros (KLEIN; WESTENBERGER, 2012; LILL, 2016).

Fatores epigenéticos podem explicar as relações entre os genes e os fatores ambientais, e vêm surgindo como um conceito atraente de pesquisa da DP (MARSIT, 2015). A epigenética é a área que busca estudar variações hereditárias de expressão de genes que independem de alterações na sequência primária (como mutações) do ácido desoxirribonucleico (DNA, *deoxyribonucleic acid* em inglês). Essas alterações podem ser modificações de DNA, de histonas, ácidos ribonucleicos (RNA, *ribonucleic acid* em inglês) regulatórios, remodelamento da cromatina e outros. As alterações epigenéticas sobrevêm em resposta aos sinais recebidos pela célula (CHEN et al., 2017). A hipótese é que genes, não responsáveis por casos monogênicos da DP, contudo ligados a doença, poderiam sofrer mudanças epigenéticas através de fatores ambientais (FENG; JANKOVIC; WU, 2015; MARSIT, 2015).

1.1.4 Tratamento farmacológico

O tratamento farmacológico da DP é sintomático, visando restaurar a atividade dopaminérgica principalmente da via nigro-estriatal. Desde os anos 1960 o tratamento é feito com objetivo de restaurar os níveis de DA com o uso de sua molécula precursora, a L-DOPA (CONNOLLY; LANG, 2014).

Outros fármacos utilizados na DP como os agonistas de receptores dopaminérgicos para ativação de receptores pós-sinápticos tais como a bromocriptina e o pramipexol, inibidores de enzimas catecol-O-metiltransferase (COMT) e monoaminoxidase B (MAO-B), responsáveis pela metabolização da DA como o entacapone e a selegilina, respectivamente, inibidores da recaptação da DA como a amantadina, (atualmente mais relacionado com o bloqueio dos receptores NMDA) para aumentar os níveis de DA na fenda sináptica.

Alguns fármacos de ação não-dopaminérgica utilizados envolvem mecanismos de antagonismo de receptores muscarínicos (ex: biperideno) para estimular a liberação de DA no estriado, auxiliando o restabelecimento provisório da neurotransmissão dopaminérgica (CONNOLLY; LANG, 2014; GIROUX, 2007).

Embora o tratamento com a L-DOPA continue sendo o "padrão ouro" e promova uma melhora da sintomatologia motora da DP, o tratamento é sintomático, não sendo capaz de impedir a progressão da doença, além de apresentar diminuição de eficácia e surgimento de efeitos colaterais ao longo do tratamento (PARK; STACY, 2015). Entre os efeitos colaterais estão os movimentos automáticos anormais (discinesias), flutuações motoras de final de dose (*wearing off*, em inglês) ou aleatórias (efeito "liga-desliga", *on-off*, em inglês), psicoses e outros (AHLSKOG; MUENTER, 2001; POEWE et al., 2017).

Devido a estas dificuldades, novos tratamentos vêm sendo pesquisados na tentativa de encontrar alternativas terapêuticas mais eficazes, como antagonistas de receptores para adenosina, glutamato ou noradrenalina, agonistas serotoninérgicos ou de receptores para o peptídeo 1 semelhante ao glucagon (GPL-1), bloqueadores de canais de cálcio, quelantes de ferro, anti-inflamatórios, antioxidantes, anticorpos para α-sinucleína e outros (STAYTE; VISSEL, 2014).

1.1.5 Diagnóstico diferencial

O diagnóstico definitivo da DP é somente confirmado em um exame anatomopatológico *post-mortem*. Exames de neuroimagem são

inespecíficos e não existem marcadores biológicos que classifiquem a DP com a acurácia necessária para conclusão clínica. O diagnóstico clínico da DP é baseado no aparecimento da tríade de sintomas: tremor de repouso, rigidez muscular e bradicinesia, além de uma boa resposta ao tratamento farmacológico com a L-DOPA (JANKOVIC, 2008). Entretanto, um estudo de Hughes e colaboradores (1992), no Reino Unido, mostrou que 25% dos pacientes diagnosticados em vida possuíam outros diagnósticos no exame *post-mortem* (HUGHES et al., 1993).

Clinicamente, as manifestações da DP são iguais às demais síndromes parkinsonianas, excetuando-se algumas peculiaridades. O aumento da oleosidade da pele e do couro cabeludo, por vezes levando ao desenvolvimento de inflamação na pele do tipo seborreica e a presença de reflexos glabelares inesgotáveis, especialmente quando os demais reflexos axiais da face não estiverem exaltados, são considerados quadros típicos da DP (FLINT, 1977; VREELING et al., 1993). A depressão também parece estar associada em aproximadamente um terço dos pacientes com a DP idiopática e aparentemente é relacionada a doença e não reativa às dificuldades motoras (JASINSKA-MYGA et al., 2010; MARSH, 2013). Existem três subtipos clínicos de DP: (1) com predomínio de tremores, (2) com predomínio de acinesia, rigidez e distúrbios de equilíbrio de marcha e (3) uma forma mista. O segundo subtipo é o mais incomum, tende a iniciar tardiamente e possui rápida evolução (FERESHTEHNEJAD et al., 2015).

O diagnóstico da DP é uma das maiores limitações no tratamento e estudo da doença, dado que sua confirmação normalmente ocorre apenas em estágios avançados do processo neurodegenerativo, no momento no qual os pacientes começam a manifestar as alterações motoras. Essa característica pode explicar a limitada eficácia clínica de fármacos neuroprotetores e neuromoduladores testados para o tratamento da DP (JANKOVIC, 2008; POSTUMA; BERG, 2016).

As metodologias e experimentos de alto rendimento (ou *high throughput experiments*, em inglês) (FENG; WANG, 2017; MILLER; FEDEROFF, 2006) como o sequenciamento de RNA (RNA-seq ou RNA *sequencing*, do inglês) e os microarranjos de DNA (ou DNA *microarray*, do inglês) apresentam a capacidade de gerar dados abundantes, os quais podem ser analisados por intermédio das diversas ferramentas matemáticas, de modo a potencialmente permitir não apenas uma compreensão biológica através da biologia de sistemas, como também facultar as buscas de assinaturas moleculares com potencial diagnóstico e prognóstico de doenças complexas, que apresentam desafio clínico como a DP.

1.2 MICROARRANJOS DE DNA

1.2.1 A tecnologia

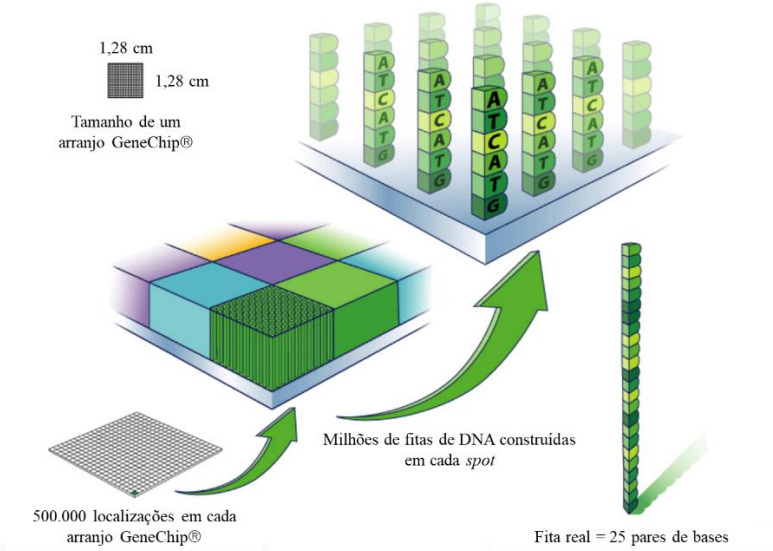
Os microarranjos de DNA, ou *DNA chips* são ferramentas que permitem a identificação e a quantificação de transcritos de mRNA presente em amostras biológicas. O total de moléculas de mRNA provindos da transcrição de um gene pode ser considerado uma aproximação do nível de expressão do mesmo (ALLISON et al., 2006). Entretanto, pode ser que haja variabilidade quanto a essa afirmação relacionada a ação de genes sobre a expressão de genes, atividade elevada em fator de baixa concentração de mRNA e outros. Entretanto, como regra geral, essa relação se mostra válida (BIER et al., 2007).

Um *slide* de microarranjo consiste numa superfície sólida onde fitas de polinucleotídeos, as sondas, foram anexadas ou sintetizadas em posições fixas (BIER et al., 2007) (**Figura 3**). Existem dois tipos principais de microarranjos de expressão, (1) os microarranjos de cDNA, ou *spotted*, e (2) os *chips* de oligonucleotídeos. Entre as suas principais diferenças está o modo como as sondas são aplicadas nos *slides*. Nos microarranjos de cDNA as sondas são sintetizadas à parte e impressas mecanicamente nos *slides*. O termo cDNA é utilizado porque as sondas são cópias complementares das sequências originais, e cada gene é então representado por uma sonda. Nos *chips* de oligonucleotídeos, representados principalmente pelos *chips* Genechip da Affymetrix, nome comercial da produtora, as sondas são diretamente sintetizadas nos *slides*. O termo "oligonucleotídeo" se refere ao fato de que durante a síntese se permite somente a criação de fragmentos pequenos de nucleotídeos e, sendo assim, os genes são representados por um conjunto de sondas (*probeset*, em inglês) (SÁNCHEZ; VILLA, 2008).

Inicialmente, em um experimento de microarranjo, o RNA total é extraído de amostras biológicas e algumas moléculas no mRNA são substituídas por outras contendo um corante fluorescente (LIPSHUTZ et al., 1999). Os transcritos resultantes marcados são chamados alvos (*targets*, em inglês). As amostras marcadas são depositadas sobre o arranjo e deixadas numa câmara de hibridização durante algumas horas. Os *targets* se ligam por hibridização nas sondas com a qual eles apresentam complementaridade suficiente de bases. O próximo passo é lavar os arranjos para eliminar os *targets* que não hibridizaram (ALLISON et al., 2006).

O modo com que as etapas anteriores são realizadas é a segunda importante diferença entre os tipos de microarranjos de expressão. Em

Figura 3: Disposição de sondas em microarranjos de DNA.



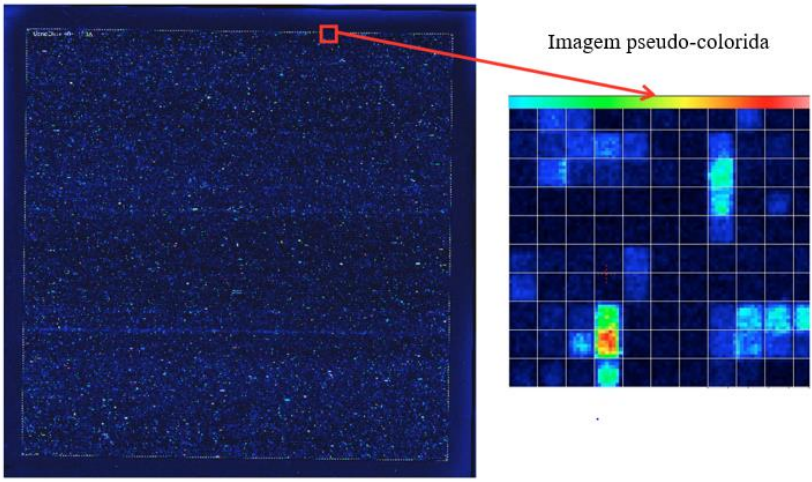
Canto superior esquerdo: Ilustração de um *slide* de microarranjo (modelo GeneChip®) e dimensões reais apontadas (1,28 por 1,28 cm). Canto inferior esquerdo: Ilustração do mesmo *slide* levemente aumentado indicando as 500.000 localizações (*spots*) em cada arranjo. As setas apontam as ampliações do *slide* original exibindo a presença de milhões de fitas de DNA (contendo geralmente 25 pares de bases) em cada *spot*. Fonte: Retirada de http://www.affymetrix.com/about_affymetrix/media/image-library.affx.

microarranjos de cDNA dois tecidos, de duas condições de interesse, marcados com corantes fluorescentes diferentes (geralmente vermelho e verde), acham-se hibridizados em um único *chip*. Por essa razão, os microarranjos de cDNA são conhecidos por microarranjos de canal duplo ou microarranjos de duas cores (*two-channel microarrays* e *two-color microarrays*, em inglês, respectivamente). Em *chips* de oligonucleotídeos se hibridiza somente uma amostra por *chip*. Esse fato faz com que sejam necessários mais *slides* por experimento, assim como não aproveita as vantagens de utilizar a hibridização por competição, entretanto simplifica o desenho experimental e é fundamentado numa tecnologia mais sensível (SÁNCHEZ; VILLA, 2008).

Para determinar a quantidade de amostra hibridizada, e consequentemente uma proporção do nível de expressão dos genes representados, o microarranjo é iluminado com uma luz *laser* que causa

excitação das moléculas marcadas, fazendo-as emitir uma fluorescência proporcional a sua quantidade. Essa fluorescência é captada via *scanner* capaz de produzir uma imagem que consiste em um *grid* de pontos, correspondentes às sondas (**Figura 4**). Essa imagem e esses pontos serão transformados em valores numéricos que serão as bases da análise (ALLISON et al., 2006; WHITWORTH, 2010).

Figura 4. *Grid* de pontos contendo sondas de um *chip* de oligonucleotídeos.



A esquerda: Imagem captada via *scanner* de um *slide* inteiro de microarranjo. A direita: Ampliação mostrando alguns poucos *spots*. O espectro de cores na parte superior da imagem indica o espectro de intensidade de fluorescência captada pelo *scanner* e por tal razão dita como imagem pseudo-colorida. Fonte: Retirada de http://www.affymetrix.com/about_affymetrix/media/image-library.affx.

1.2.2 Medidas de expressão

As duas tecnologias de microarranjos geram tipos diferentes de imagens e consequentemente diferentes dimensões que devem ser adequadamente operadas para fornecer os estimados da variável estudada, a expressão gênica (SÁNCHEZ; VILLA, 2008).

A imagem obtida de um microarranjo de cDNA é analisada, ou seja, quantificada e algumas medidas são geradas para cada ponto. Embora as medidas sejam dependentes do *software* utilizado, geralmente elas consistem de medidas de sinal, vermelho e verde para cada canal, medidas de ruído de fundo (ou *background*, em inglês), visando medir a

fluorescência não resultante das hibridizações e medidas de qualidade do ponto.

As medidas obtidas de expressão comparada são comumente transformadas em logaritmo na base 2 por aproximarem os dados de expressão de uma distribuição log-normal e por simetrizar as diferenças, facilitando a interpretação (DO; CHOI, 2006).

Os *chips* de oligonucleotídeos representam cada gene como um conjunto de sondas, ou cadeias de oligonucleotídeos. Cada "sonda" dos *chips* é composta por um par de sondas, uma sonda com *match* (compatível, em português) perfeito, que corresponde a sequência nativa de DNA e uma sonda *mismatch* (incompatível, em português), cujo nucleotídeo central da sequência nativa de DNA foi modificado. Essa abordagem visa representar a "expressão real", excluindo e considerando como ruído de fundo toda hibridização com uma sonda *mismatch* (DO; CHOI, 2006; IRIZARRY et al., 2003).

As principais diferenças entre as medidas de expressão envolvem no fato de *chips* de oligonucleotídeos terem um único valor de expressão para cada condição enquanto em um microarranjo de cDNA se obtém medidas de expressão relativas entre as duas condições. Embora os *chips* de oligonucleotídeos produzam estimados mais acertados, medidas de expressão relativas salientam uma interpretação mais intuitiva (SÁNCHEZ; VILLA, 2008).

1.2.3 Processamento analítico de microarranjos de DNA

O processamento analítico de microarranjos é provavelmente a etapa de maior dificuldade desse experimento. Para utilizar o enorme volume de dados gerado em um experimento de microarranjo é importante direcionar a análise para os objetivos específicos do experimento (SIMON, 2009). Microarranjos e outros experimentos genômicos são em sua natureza diferentes de outros experimentos tradicionais cuja maior parte das técnicas estatísticas foram elaboradas. Devido a este fato, foram essenciais as adaptações de técnicas existentes e o desenvolvimento de novos métodos que se ajustassem as circunstâncias (SÁNCHEZ; VILLA, 2008).

1.2.3.1 Pré-processamento de dados

O controle de qualidade busca determinar se os experimentos foram corretamente realizados e se os dados estão adequadamente expressos para que seja considerado confiável. Não existem métodos

padrões para essa etapa de pré-processamento de microarranjos (BEISSBARTH et al., 2000). A maior parte dos controles de qualidade é baseada na análise de imagens e de gráficos, embora determinadas sumarizações numéricas já tenham sido desenvolvidas (ALLISON et al., 2006).

O controle de qualidade de microarranjos de cDNA é predominantemente baseado em imagens e gráficos como a inspeção de imagens na detecção de irregularidades como arranhões, bolhas ou um elevado ruído de fundo. Para este último, histogramas de sinal ou de sinal por ruído são examinados na detecção de anormalidades. Além do mais, a maior parte dos *softwares* de análise de imagens de microarranjo conseguem gerar *flags* (bandeiras em inglês) de pontos indicando quanto os pontos podem ser considerados nas análises. Esses valores podem ser utilizados eventualmente na filtragem de alguns desses pontos (KAUFFMANN; GENTLEMAN; HUBER, 2009; SÁNCHEZ; VILLA, 2008).

O controle de qualidade de *chips* de oligonucleotídeos, por sua vez, exibe algumas diferenças. Os histogramas e os gráficos de degradação são úteis para uma primeira inspeção visual e pode ajudar na detecção de grandes irregularidades no *chip*. *Softwares* de marcas comerciais como Affymetrix proveem resumos numéricos de ruído de fundo, chamadas de presença e outros que podem ser comparados para determinar a qualidade do experimento (KAUFFMANN; GENTLEMAN; HUBER, 2009; SÁNCHEZ; VILLA, 2008).

Após a avaliação do controle de qualidade, alguns passos do pré-processamento ainda são necessários. Dependendo da tecnologia do microarranjo isso envolve três passos: (1) um ajuste do ruído de fundo para remoção do sinal resultante de hibridizações inespecíficas, (2) uma normalização dos dados para correção de vieses sistemáticos devido a absorção diferencial de corantes, heterogeneidade espacial do *chip* e outros e (3) em *chips* de oligonucleotídeos é necessário concentrar os diferentes sinais adquiridos de todas as sondas representantes de um gene em um único valor (ALLISON et al., 2006).

Parte do sinal observado é resultado de ligações não específicas, uma parcela da amostra pode se ligar a fitas não complementares ou mesmo a materiais inorgânicos. Além do mais, parte do sinal pode ser devido a fontes não biológicas. De forma geral, existe a necessidade de se estimar e remover esse ruído para que a hibridização específica seja a utilizada (WHITWORTH, 2010). Diferentes métodos foram desenvolvidos para essa etapa e algumas comparações foram previamente publicadas (FREUDENBERG, 2005; RITCHIE et al., 2007). Como

conclusão geral, esses estudos mostram que técnicas baseadas no modelo são aquelas que melhor executam seus empregos. Entre as metodologias normalmente utilizadas estão o normexp (SHI; OSHLACK; SMYTH, 2010) para microarranjos de cDNA, VSN (HUBER et al., 2002) para ambos os tipos de microarranjo de DNA e RMA para *chips* de oligonucleotídeos (IRIZARRY, 2003).

A normalização é um passo chave em processos de análise de microarranjos e há muito trabalho devotado no desenvolvimento de metodologias que testem diferentes artefatos técnicos que devem ser corrigidos. Entretanto, nenhum método consegue lidar com todas estas dificuldades (CHEN et al., 2003). De forma geral, os métodos de normalização são baseados nos seguintes princípios: a maior parte dos genes no experimento não são expressos ou são equitativamente expressos em qualquer condição. Somente uma pequena parcela de genes devem mostrar modificações de expressão entre condições experimentais (DO; CHOI, 2006).

Diferentes métodos foram desenvolvidos para essa etapa como o método lowess (YANG et al., 2012) para microarranjos de cDNA, RMA para *chips* de oligonucleotídeos (IRIZARRY, 2003) e outros. O método lowess normaliza os valores de expressão para tornar consistentes as intensidades em cada arranjo (YANG et al., 2012). Em diversos cenários é necessária a obtenção de consistência entre os arranjos, obtidas com métodos como a normalização de escala ou de quantil (SÁNCHEZ; VILLA, 2008). Os *chips* de oligonucleotídeos apresentam artefatos técnicos diferentes, requerendo métodos de normalização particulares. O método comumente utilizado é o RMA e ele é executado com 3 passos: um ajuste de ruído de fundo baseado em modelos com valores de sondas, uma normalização de quantil e uma sumarização integrando os valores das sondas que correspondem ao gene (IRIZARRY, 2003). Algumas comparações foram previamente publicadas sobre esses e outros métodos (BOLSTAD et al., 2003; FREUDENBERG, 2005). Os estudos mostram que técnicas baseadas no modelo são aquelas que melhor executam seus empregos, porém ainda se requerem discussões.

1.2.3.2 Processamento de dados

Em experimentos de microarranjo, o objetivo da análise estatística é determinar o resultado da variável resposta, as medidas de expressão (BIER et al., 2007). Entretanto, dependendo da tecnologia manipulada, essa medida pode apresentar diferentes formas. Em microarranjos de cDNA a abordagem é utilizar a expressão relativa, a variável resposta

relação-log das intensidades. Em *chips* de oligonucleotídeos a abordagem é utilizar a expressão absoluta, a variável resposta valor de intensidade de cada arranjo sozinho medido em escala logarítmica (SÁNCHEZ; VILLA, 2008).

A análise de dados de microarranjos é útil para identificação de genes que estão apresentando mudanças significativas em seus níveis de expressão entre condições. Os limites de significância são determinados por vários testes estatísticos executados dependendo da pergunta do experimento (CAPALDI, 2010). A técnica de microarranjos de DNA é aplicável a múltiplas situações experimentais. Em nível molecular, o desenvolvimento de doenças, como a DP, ocorre devido a alterações na transcrição/tradução de genes. Os microarranjos permitem comparar amostras de tecido normal e anormal para determinar as alterações decorrentes da condição e assim identificar alvos moleculares de estudo.

Geralmente, um experimento de microarranjo busca os genes diferencialmente expressos (GDE) entre condições, esses experimentos são conhecidos por *comparação de classes*, entretanto outros propósitos são também estudados. Os estudos de *análise de vias* buscam reconhecer co-regulações de genes e as relacione em processos bioquímicos comuns. Os estudos de *descoberta de classes* buscam reconhecer novos subtipos de indivíduos em uma população, por exemplo, separar o que era conhecido como um fenótipo em subconjuntos se baseando no perfil transcricional. Os estudos de *predição de classes* buscam desenvolver modelos matemáticos capazes de indicar a classe de um novo indivíduo, por exemplo, predizer a resposta a um tratamento (como "responsivo" ou "não responsivo"), diagnóstico ("indivíduo sadio" ou "paciente com a DP"), e outros. (SÁNCHEZ; VILLA, 2008; TSENG; GHOSH; FEINGOLD, 2012)

1.2.3.3.1 Comparação de classes

Os estudos de *comparação de classes* buscam selecionar os genes cujas expressões são significativamente diferentes entre condições experimentais, os chamados GDE (CAPALDI, 2010).

Diversos métodos foram criados para esse tipo de análise discutidos na literatura (GUI et al., 2005; PAN, 2002). Em geral, separam-se os métodos em dois grupos, os baseados no modelo e os globais.

Entre as metodologias baseadas no modelo, o processo habitual utiliza uma análise de variância (ANOVA) específica para microarranjos de cDNA onde as intensidades relativas a expressão de cada corante, geralmente cianina 3 e 5, são abordadas separadamente (WU et al., 2003).

Dessa forma, esse tipo de método utiliza os valores absolutos de expressão, porém podem ser modificados e utilizados para dados de *chips* de oligonucleotídeos. A diferença de expressão é resultado da interação de "tratamento pelo gene" quando exibem significância estatística (SÁNCHEZ; VILLA, 2008).

Os métodos globais, apesar do seu nome, analisam um gene por vez visualizando um experimento de microarranjo como um experimento simples de tal modo que se pode realizar as análises utilizando um tratamento padrão de modelo linear e testes *t* para todos os genes (STAFFORD, 2008). Entretanto, esse conceito é considerado ineficiente devido a dois fatores comuns em experimentos de microarranjo: (1) geralmente a dimensão amostral é pequena complicando a estimação de variância e (2) as variâncias em si podem apresentar muita variação entre os genes (TONG; WANG, 2007). Para enfrentar tais limitações, uma estratégia geralmente empregada é o encolhimento de variância (ou *variance shrinkage*, em inglês) que consiste na dependência de estimados de variância melhorados $\tilde{S}$, onde essa melhoria provém do empréstimo de informação de todos os genes do arranjo (SÁNCHEZ; VILLA, 2008).

Dois métodos globais comumente empregues, o método *SAM* (de *Significance Analysis of Microarrays* ou análise de significância de microarranjos em inglês) (TUSHER; TIBSHIRANI; CHU, 2001) e o *limma* (de *Linear Models for Microarray Data* ou modelo linear para dados de microarranjo em inglês) (RITCHIE et al., 2015) utilizam diferentes alternativas de encolhimento de variância. O método *SAM* calcula os estimados aplicando métodos de permutação enquanto o método *limma* calcula aplicando métodos bayesianos empíricos (SÁNCHEZ; VILLA, 2008).

1.2.3.3.2 Análise de vias

Os experimentos de microarranjo de DNA de comparação de classes podem derivar em longas listas de genes que foram selecionados utilizando um critério para atribuir significâncias estatísticas. Com essa lista é possível seguir diversos caminhos analíticos, como a interpretação biológica (ASHBURNER et al., 2000; LI; BECICH; GILBERTSON, 2004).

Um tratamento comum para interpretação biológica é o reprocessamento da lista relacionando os genes selecionados com alguns bancos de dados de anotações funcionais como o *Gene Ontology* (GO, ontologia gênica em inglês), *Kyoto Encyclopedia of Genes and Genomes* (KEGG, enciclopédia de Kyoto de genes e genomas em inglês) e outros

(TSENG; GHOSH; FEINGOLD, 2012). Há muitos métodos e modelos para realizar essa tarefa, em forma geral separam-se os métodos em dois grupos, (1) o *Gene Enrichment Analysis* (GEA, análises de enriquecimento gênico em inglês) e (2) o *Gene Set Enrichment Analysis* (GSEA, análises de enriquecimento de conjuntos de genes em inglês) (SUBRAMANIAN et al., 2005).

As análises de GEA almejam estabelecer se uma dada categoria de ontologia gênica ou de vias de interação molecular aparece com maior (enriquecida) ou menor (empobrecida) frequência na lista de genes selecionados do que na população de onde eles foram obtidos, por exemplo o próprio arranjo, um genoma ou simplesmente os genes que foram selecionados para esse teste. A significância estatística do potencial enriquecimento/empobrecimento gênico é estabelecida utilizando testes hipergeométricos (LI; BECICH; GILBERTSON, 2004).

As análises de GSEA diferem das GEA pelo fato de requererem além da lista de genes, uma variável numérica de ranqueamento, geralmente o valor de *p* de um teste estatístico de expressão diferencial, como os testes *t* ou ANOVA. Utilizando as listas ranqueadas são computados escores cumulativos (para enriquecimento) baseados na presença ou ausência de cada gene ou conjuntos de genes em cada categoria funcional. O teste estatístico de Kolmogorov-Sminov é utilizado para verificar e comparar a distribuição de escores na categoria funcional com a distribuição empírica da variável numérica para decidir se os genes do topo ou da base da lista estão supra-representados (SUBRAMANIAN et al., 2005).

O conceito das análises de ontologias gênicas desempenha um papel chave na indagação biomédica atual, sendo referenciado em milhares de artigos científicos. Somente na análise de dados de expressão, 34.225 artigos utilizaram os termos de ontologias para criar e validar hipóteses (http://www.geneontology.org/page/publications, acessado em 05 de janeiro de 2019).

1.2.3.3.3 Descoberta de classes

Os métodos de *descoberta de classes* ou agrupamento (*clustering*, em inglês) são atualmente utilizados como uma das primeiras inspeções da matriz de expressão gênica buscando associar amostras e genes similarmente expressos para correlacionar os resultados e as interpretações biológicas (SÁNCHEZ; VILLA, 2008). As técnicas ou algoritmos de agrupamento como a análise de componente principal (PCA, *Principal component analysis*, em inglês) (PEARSON, 1901) e o

dimensionamento multidimensional (MDS, *Multidimensional scaling*, em inglês) (TORGERSON, 1958) reduzem a dimensionalidade do sistema e permite a gestão de modo facilitado do conjunto de dados.

O conceito é que genes que são corregulados e/ou relacionados funcionalmente devem ser genes com perfil de expressão semelhantes e que podem ser agrupados (O'CONNELL, 2003). As técnicas de agrupamento são utilizadas na construção de arranjos de classificação de genes, amostras ou ambos. Quando aplicadas em grupos de genes podem auxiliar na identificação de grupos de elementos corregulados, na identificação espacial e temporal de padrões de expressão, na redução da redundância de preditores em algoritmos de aprendizado e outros. Quando aplicadas em grupos de amostras podem auxiliar na identificação de novas classes ou fenótipos biológicos (por exemplo novas classes de tumores), na detecção de produtos experimentais ou para propósitos de exibição (SÁNCHEZ; VILLA, 2008).

1.2.3.3.4 Predição de classes

As análises de *predição de classes* almejam desenvolver funções multivariadas para prever a classificação de um novo indivíduo em um fenótipo, utilizando variáveis preditoras para isso, como os perfis de expressão (KUHN; JOHNSON, 2013).

É importante distinguir as *predições de classes* e *de prognóstico*. A primeira é aplicada na designação de uma nova amostra às categorias existentes, como em um diagnóstico, e a segunda é aplicada na predição do progresso da condição. Um exemplo de *predições de classes* é a atribuição de tumores a uma das classes predefinidas (GOLUB et al., 1999), enquanto um exemplo de *predição de prognóstico* é a construção de preditores que determinem quais tumores podem evoluir para metástase (VAN'T VEER et al., 2002).

Uma das formas de construir as funções de predições é com a utilização de algoritmos de aprendizado de máquina. O aprendizado de máquina explora a construção e a utilização de algoritmos que aprendem com os dados de que são providos (LANTZ, 2015). Essa área do conhecimento surgiu da intersecção da estatística, que busca estudar as relações entre os dados e a ciência da computação, com sua ênfase em algoritmos de computação eficientes. Essa relação entre áreas foi impulsionada por desafios de construção de modelos estatísticos para as quantidades crescentes de dados gerados (KUHN; JOHNSON, 2013). O objetivo atual dessa área do conhecimento é o auxílio na interpretação de dados criados e armazenados em rápida velocidade, como os resultados

de experimentos de alta taxa de transferência, como os microarranjos de DNA (LANTZ, 2015).

Os algoritmos de aprendizado de máquina são corretamente aplicados quando aumentam, e não substituem, o conhecimento do especialista. São atualmente utilizados em associação ao serviço de médicos no desafio da erradicação do câncer (KOUROU et al., 2015), com os engenheiros e os programadores no desenvolvimento de instrumentos inteligentes (KASSAHUN et al., 2016), com os cientistas sociais para entender como suas dinâmicas funcionam (MULLAINATHAN; SPIESS, 2017), e outros. Os algoritmos estão sendo utilizados em diversas empresas, laboratórios científicos, hospitais, organizações governamentais e, até mesmo, redes sociais.

1.2.4 Meta-análise de microarranjos de DNA

Os estudos de microarranjo apresentam diversas utilidades, sendo esperado que diversos resultados ainda surjam dessa tecnologia. Entretanto, muitos relatos de achados não são reprodutíveis ou robustos a mais leve perturbação de dados. Entre as causas comuns para isso estão a análise ou validação inadequadas, o controle insuficiente de falsos positivos e os relatos inadequados de métodos. A situação é exacerbada pelos estudos com pequeno tamanho amostral, onde geralmente são analisados milhares de sondas em apenas dezenas de amostras. Além do mais, é necessária a generalização de resultados antes da consideração de aplicação prática, levando em consideração fatores geográficos, ambientais, genéticos, e outros (RAMASAMY et al., 2008; TSENG; GHOSH; FEINGOLD, 2012).

A combinação de informação de vários trabalhos pode adicionar confiabilidade e generalização de resultados de microarranjos. A execução de técnicas estatísticas na combinação de resultados de estudos independentes é chamada meta-análise (ELLIS, 2010). Através da meta-análise de microarranjos de DNA é possível alcançar estimativas mais precisas do perfil de expressão diferencial e avaliar a heterogeneidade do conjunto estimado (RAMASAMY et al., 2008).

Apesar das vantagens das meta-análises de expressão gênica serem utilizadas em diferentes campos da pesquisa científica, numerosos processos foram propostos para análises no contexto de microarranjos (RHODES et al., 2002, 2004; CHOI et al., 2003 e outros). Entretanto, não existe uma estrutura inclusiva e abrangente de como realizar essas análises (RAMASAMY et al., 2008; TSENG; GHOSH; FEINGOLD, 2012).

As técnicas de alto rendimento estão sendo rapidamente desenvolvidas e amplamente utilizadas. Metodologias voltadas as análises destas se tornarão cada vez mais essenciais para utilização de informações contidas nas extensas bases de dados, produzindo resultados cada vez mais reais e reprodutíveis (SWEENEY et al., 2017).

As melhorias no sistema de saúde levam a uma maior sobrevida, exacerbando a prevalência da DP. Espera-se que os números de pacientes com a DP dobrem de 2005 até 2030. Esses fatos apontam um aumento no número de pacientes e de anos vividos com a DP (POEWE et al., 2017). Torna-se cada vez mais necessário o desenvolvimento de metodologias de prognóstico e diagnóstico para o descobrimento de tratamentos e para o aumento do período de eficácia do tratamento atual, resultando em qualidade de vida para os pacientes com a DP (SHARMA et al., 2013). A utilização de técnicas de alto rendimento vinculadas as metodologias de meta-análise apresentam resultados mais confiáveis e susceptíveis de generalização do que em experimentações independentes e podem ser a resposta na criação de assinaturas gênicas diagnósticas baseadas no perfil de expressão de pacientes (RAMASAMY et al., 2008). Para que uma assinatura gênica seja útil clinicamente é recomendado que ela seja detectável no sangue periférico, fluidos ou tecidos facilmente coletáveis. Recentemente, pesquisas de expressão gênica identificaram modificações no sangue relacionadas a neurodegeneração na DP (MUTEZ et al., 2011; RILEY et al., 2014; SCHERZER et al., 2007; SOREQ et al., 2008). Dessa forma, faz-se necessária a caracterização da expressão gênica do sangue periférico de pacientes com a DP para elaboração de agrupamentos de genes que sejam capazes de diagnosticar a DP.

2. HIPÓTESE

É possível classificar amostras de pacientes com a DP idiopática quando comparadas com amostras de indivíduos sadios se baseando em perfis de expressão gênica do sangue periférico analisados com algoritmos de aprendizado de máquina.

3. OBJETIVOS

3.1 OBJETIVO GERAL

Identificar uma assinatura gênica eficaz na classificação de pacientes com a DP idiopática utilizando algoritmos de aprendizado de máquina.

3.2 OBJETIVOS ESPECÍFICOS

– Construir um banco de dados de microarranjo de amostras de sangue periférico de pacientes com a DP idiopática.

– Calcular os tamanhos de efeito de cada gene em cada trabalho e calcular os tamanhos de efeito agrupados destes.

– Selecionar os genes de maior distinção de expressão entre pacientes com a DP idiopática e indivíduos sadios.

– Caracterizar funcionalmente os genes escolhidos utilizando ontologias gênicas de processos biológicos e de vias de interação molecular.

– Verificar a competência de classificação na análise de agrupamento em estudos independentes aproveitando os genes selecionados.

– Identificar uma assinatura gênica eficiente para discernir amostras de pacientes com a DP idiopática de indivíduos sadios.

– Elaborar e otimizar modelos de classificação utilizando algoritmos de aprendizado de máquina e ajustes de hiperparâmetros.

– Selecionar os modelos de melhor classificação para as amostras de pacientes com a DP idiopática e para as amostras de indivíduos sadios e descrever os limites numéricos de probabilidade de classe para categorização/aplicação.

– Testar o modelo de melhor classificação para as amostras de pacientes com a DP idiopática em amostras de outras categorias/doenças e avaliar a probabilidade de serem classificadas como DP idiopática.

– Construir mapas de paisagem de redes de correlação para um modelo de classificação baseado em imagens e enriquecido por colinearidade.

4. MATERIAL E MÉTODOS

4.1 BUSCA E CRITÉRIOS DE INCLUSÃO E EXCLUSÃO DE CONJUNTOS DE DADOS

Para localizar os conjuntos de dados (*datasets*, do inglês) de microarranjo de DNA contendo os dados individuais de pacientes foram utilizadas duas metodologias de busca distintas. A primeira envolveu buscar em um banco de dados de resumos, o MEDLINE (*Medical Literature Analysis and Retrieval System Online* ou sistema online de busca e análise de literatura médica, em inglês), utilizando o motor de busca PubMed, artigos científicos que contivessem no corpo do texto identificadores de dados de microarranjo depositados. A segunda buscou diretamente em repositórios públicos de dados de microarranjo recomendados pelo requerimento MIAME (*Minimum Information About a Microarray Experiment* ou informações mínimas sobre um experimento de microarranjo, em inglês) (BRAZMA et al., 2001), o Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) (EDGAR; DOMRACHEV; LASH, 2002) e o ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) (BRAZMA et al., 2003).

Para as duas metodologias foram utilizadas as mesmas palavras-chave de busca: "*Parkinson*", "*Blood*" e "*Microarray*" (Parkinson, sangue e microarranjo, respectivamente, em inglês). Todos os artigos científicos ou conjuntos de dados encontrados foram analisados. Os critérios de inclusão do trabalho foram: estudos contendo amostras de sangue periférico de humanos; amostras de pacientes com a DP idiopática e indivíduos do grupo controle; análises realizadas em microarranjo de expressão; estudos contendo mais de 10 amostras no conjunto de dados, com pelo menos 5 amostras por grupo (minimizando os possíveis efeitos de lote); estudos contendo dados de expressão individuais de pacientes (DIP). Por sua vez, os critérios de exclusão foram: estudos contendo amostras de outros tecidos ou amostras de outros seres não-humanos; amostras de pacientes com outras patologias que não a DP idiopática; análises realizadas em outras tecnologias; estudos contendo menos de 10 amostras no conjunto de dados; estudos que não provem DIP.

4.2 IMPORTAÇÃO E PRÉ-PROCESSAMENTO DE CONJUNTOS DE DADOS

Os conjuntos de dados selecionados foram importados para o ambiente de análises estatísticas e gráficas R versão 3.5.0, disponível em

http://cran.r-project.org/, com auxílio do ambiente de desenvolvimento integrado RStudio, disponível em http://www.rstudio.com/. A importação dos dados foi feita com a função *getGEO*, presente no pacote *GEOquery* (DAVIS; MELTZER, 2007), utilizando como parâmetros os códigos de identificação "GSE" (*Gene Series*) de cada conjunto de dados, obtidos no GEO, contendo um conjunto de amostras de um estudo.

Para cada conjunto de dados importados (como objetos de classes *ExpressionSet*), 3 matrizes foram analisadas: a matriz de expressão, a matriz de *features* e a matriz de fenótipos. As matrizes de expressão possuem as medidas de expressão e é organizada por sondas nas linhas e amostras nas colunas. As matrizes de *features* contêm as informações indicativas as sondas, geralmente tidas como dados de anotação. As matrizes de fenótipo descrevem as amostras do estudo, trazendo aspectos demográficos como sexo, idade, diagnóstico clínico (*Parkinson* ou *Controle*), etc. Essas matrizes são organizadas por amostras nas linhas e características nas colunas, sendo assim, os números de linhas em matrizes de fenótipos é o mesmo de colunas em matrizes de expressão.

As matrizes de expressão continham os valores de intensidade de sinal em escala linear ou dados escalonados em diferentes intervalos. Para possibilitar a integração e comparação entre os dados de diferentes estudos e plataformas, nos dados com escala linear foi feita uma transformação logarítmica de base 2 (log2), obtendo-se valores entre 0 e 16. Os dados previamente escalonados (em log10) foram reescalonados para este mesmo intervalo aplicando a fórmula: log2(10^(intensidade log10-transformada)). Para visualizar as distribuições de níveis de expressão gênica em todas as amostras de cada conjunto e diagnosticar os dados sobre a necessidade de transformação numérica, foi empregue a função *plotDensities*, presente no pacote *limma* (*Linear Models for Microarray Data* ou Modelo linear para dados de microarranjo, em inglês) (RITCHIE et al., 2015).

As matrizes de *features* foram atualizadas utilizando os pacotes de anotação presentes no projeto *Bioconductor*, plataformas comerciais específicas acessíveis em http://www.bioconductor.org/packages/2.8/data/annotation/. Para cada sonda foram anexados os identificadores Entrez Gene ID (MAGLOTT et al., 2011) e símbolos de gene (Gene Symbol, em inglês), pequenos identificadores do nome do gene (ex: Beta Actina, Entrez Gene ID: 60, Gene symbol: *ACTB*). Para as anotações empregou-se os pacotes *hgu133a.db*, *hgu133a2.db*, *hgu133plus2.db* para plataforma Affymetrix e *illuminaHumanv3.db* para plataforma Illumina.

Diferentes parâmetros de controle de qualidade, separados em três sessões, foram avaliados nos conjuntos estudados, como as distâncias entre os arranjos, a distribuição de intensidades de arranjos e as análises de componentes principais. As métricas de qualidade foram obtidas com a função *arrayQualityMetrics*, presente no pacote igualmente denominado (KAUFFMANN; GENTLEMAN; HUBER, 2009) e depositadas em um relatório do tipo HTML gerado pela função. Amostras desviantes em mais de uma das sessões analisadas foram tidas como amostras de má qualidade e retiradas das próximas análises.

As técnicas de pré-processamento, a correção de *background* (ruído de fundo, em português) e a normalização foram selecionadas para cada tipo de plataforma, escolhendo as técnicas padrão na literatura para cada. Para os arranjos da plataforma Affymetrix e Illumina o pré-processamento foi feito com as funções *rma* (Robust Multi-Array Average expression measure, ou Medida de expressão média robusta multi-arranjo, em português) (IRIZARRY, 2003), presente no pacote *affy* (GAUTIER et al., 2004) e *neqc*(NormExp Background Correction and Normalization Using Control Probes, ou Correção de fundo NormExp e normalização usando sondas de controle) (SHI; OSHLACK; SMYTH, 2010), presente no pacote *limma*, respectivamente. As duas funções aplicam normalização de quantis, porém utilizam características próprias de cada plataforma para obtenção de melhores resultados.

4.3 META-ANÁLISE DE DADOS DE MICROARRANJO

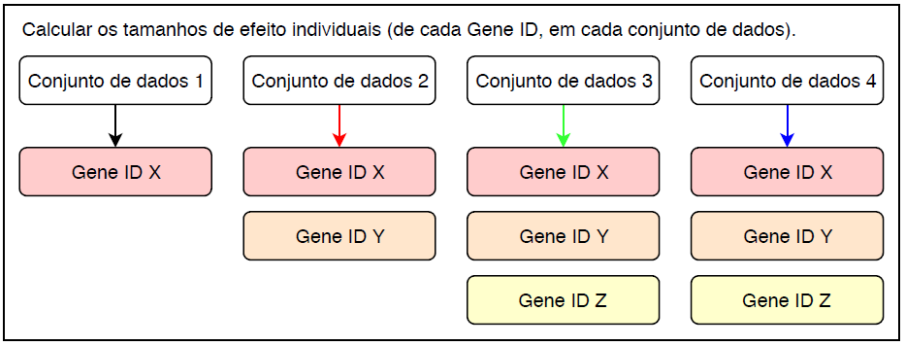
Para meta-análise de dados de microarranjo diversas técnicas podem ser utilizadas dependendo da natureza da resposta (do tipo binária, contínua, de sobrevivência, etc.) e do objetivo do estudo. Para comparação de duas classes, como feito aqui (resposta sendo os fenótipos *Parkinson* ou *Controle*), há quatro maneiras gerais para combinação de informação (RAMASAMY et al., 2008).

Nenhum dos métodos descritos poderia ser realizado sem uma reanálise dos dados de microarranjo de todos os estudos, porque nenhum trabalho com esses dados fornece uma lista completa (contendo todos os genes) com (1) GDE, (2) suas classificações/ranqueamentos, (3) valores de *p* ou (4) tamanhos de efeito, exigências essenciais para tais métodos (RAMASAMY et al., 2008). Adicionalmente, quando presentes, os dados foram obtidos por metodologias distintas de seleção de GDE, utilizando algoritmos que têm objetivos distintos e limiares de corte (*thresholds*, em inglês) diversos. Sendo assim, foi optado por calcular os tamanhos de

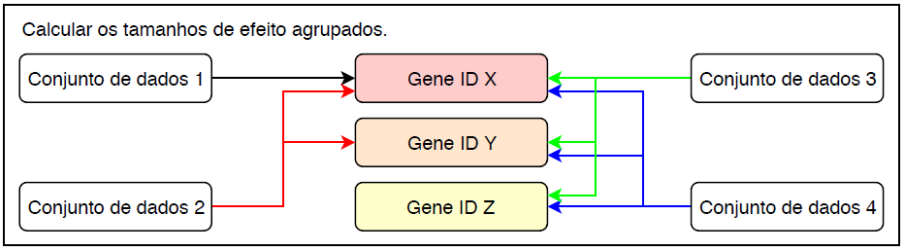
efeito de todos os genes em todos os conjuntos de dados e posteriormente realizar a meta-análise utilizando a combinação de tamanhos de efeito (**Figura 5**).

Figura 5. Organograma de cálculos de tamanhos de efeito para meta-análise.

A



B



A: Esquema do cálculo do tamanho de efeito de cada Gene ID de cada conjunto de dados atingido. Devido às diferenças de tecnologias e plataformas utilizadas, a presença de Gene ID pode ser diferente nos conjuntos analisados. No conjunto de dados 1 está presente apenas o Gene ID X, enquanto no conjunto de dados 3 estão presentes os Gene ID X, Y e Z. **B**: Esquema do cálculo do tamanho de efeito agrupado de cada Gene ID. Para as operações são utilizados os valores de tamanho de efeito de cada conjunto que este Gene ID está incluído. Para o Gene ID X são utilizados os valores de tamanho de efeito deste gene nos conjuntos 1, 2, 3 e 4, enquanto para o Gene ID Z, são utilizados dos conjuntos 3 e 4.

Existem diferentes tipos de tamanhos de efeito, comumente separados em duas famílias: (1) diferenças entre os grupos (família *d*) e (2) medidas de associação (família *r*) (ELLIS, 2010). A primeira foi empregue inicialmente por oferecer diversas maneiras de correção de dados. Além disso, qualquer valor pode ser transformado em um valor de

outra família (*d* em *r*, e vice-versa), porém converter um valor *r* em *d* geralmente resulta em perda de informação (COHEN, 1983).

4.3.1 Cálculos de tamanhos de efeito individuais

Para calcular os tamanhos de efeito individuais foram empregues as matrizes de expressão, descritas no item 4.2.

Os cálculos de tamanhos de efeito baseados em diferenças de dois grupos foram realizados utilizando a subtração da média de um grupo pela do outro ($M_A - M_B$) e sua divisão por o desvio padrão da população amostrada. Como o desvio padrão populacional é desconhecido, há três formas clássicas para solucionar esse obstáculo: utilizando as equações de (1) Cohen's *d* (COHEN, 1962), (2) Glass's Δ (GLASS; MCGAW; SMITH, 1981), ou (3) Hedges' *g* (HEDGES, 1981). A última (*g*) é recomendada em grupos de tamanhos diferentes e por isso foi optada nesse estudo.

As equações operadas, utilizando as matrizes de expressão de valores normalizados estão representadas nas **Equações 1** e **2**, onde M_A e M_B são as médias, n_A e n_B são os tamanhos das amostras, DP_A e DP_B são os desvios padrão dos grupos com fenótipos *Parkinson* e *Controle*, respectivamente.

Equação 1.

$$\text{Hedges}'g = \frac{M_A - M_B}{DP^*_{agrupado}}$$

Equação 2.

$$SD^*_{agrupado} = \sqrt{\frac{(n_A - 1)DP_A^2 + (n_B - 1)DP_B^2}{n_A + n_B - 2}}$$

Foi feita uma correção sobre os valores de *g* para calcular o estimado não-enviesado dos tamanhos de efeito (g^*) (**Equação 3**) (HEDGES; OLKIN, 1985). Este fator visa não superestimar diferenças normalizadas de tamanhos de efeito em estudos com pequeno tamanho amostral.

Equação 3.

$$g^* \cong d\left(1 - \frac{3}{4(n_A + n_B) - 9}\right)$$

Os cálculos da meta-análise de microarranjos foram realizados utilizando os valores da família de tamanhos de efeito *r*, por ser o procedimento com melhor descrição encontrado (ELLIS, 2010). A transformação do valor *g* em *r* (coeficiente de correlação produto-momento de Pearson) foi realizada sobre o valor estimado não-enviesado (g^*) e, portanto, a correção foi mantida (**Equação 4**). Neste cálculo, *d* se refere ao valor da fórmula de tamanhos de efeito da família *d* utilizada (aqui Hedges' *g* corrigido, g^*).

Equação 4.

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

Para interpretação de valores de tamanho de efeito, foram utilizados os valores de limiares publicados por Rosenthal e Rosnow (1984) (**Tabela 1**), que demonstra uma inconsistência entre os tamanhos de efeito de famílias *r* e *d* quando os valores obtidos de *r* são valores de *d* transformados, como feito acima.

Tabela 1. Comparação de limiares para as famílias *r* e *d* de tamanhos de efeito.

	d	***r***	***r* equivalente a *d***
Pequeno	,20	,10	,10
Médio	,50	,30	,24
Grande	,80	,50	,37

4.3.2 Cálculos de tamanhos de efeito agrupados

4.3.2.1 Cálculos de tamanhos de efeito médios

Para responder as perguntas referentes aos tamanhos dos efeitos que cada gene possui no sangue de pacientes com a DP idiopática, deve-se combinar os tamanhos de efeito calculados em cada estudo. Existem muitas formas de calcular esse valor "médio", sendo o cálculo da média desses valores a menos adequada. Uma das alternativas, que foi a utilizada neste trabalho, é calcular os tamanhos de efeito médio ponderado (*weighted mean size effect*, do inglês) (ELLIS, 2010; RAMASAMY et al., 2008). Para esse cálculo, as ponderações (ou pesos) são estimadas dos tamanhos das amostras, ou seja, são colocados pesos

maiores em decorrência de amostras maiores por essas serem menos enviesadas por erros amostrais.

Para os cálculos de tamanhos de efeito médio ponderado ($r_{médio}$) (**Equação 5**) foram somados todos os tamanhos de efeito de cada gene em cada conjunto de dados (r_i) multiplicado pelo tamanho amostral do conjunto de dados correspondente (n_i) e dividido pelo tamanho total da amostra (Σn_i).

Equação 5.

$$r_{médio} = \frac{\sum n_i r_i}{\sum n_i}$$

4.3.2.2 Cálculos das significâncias estatísticas dos tamanhos de efeito médio ponderado

Foram utilizadas duas maneiras para calcular a significância estatística dos tamanhos de efeito médio ponderado ($r_{médio}$), (1) pela conversão do resultado em um escore *z* para determinar se a probabilidade de obter um valor deste tamanho é menor que 0.05, significância estatística escolhida e (2) pelo cálculo do intervalo de confiança de 95% (IC 95%), posteriormente checando se esse intervalo tange o valor nulo de zero. Para ambos os casos, foi necessário calcular a variância dos tamanhos de efeito médio ponderado de cada gene.

Para os cálculos de variância da amostra de correlações ($v._r$) (**Equação 6**) foram somados todos os valores da multiplicação de quadrados da diferença entre o valor de correlação estimado (tamanhos de efeito, r_i) e o valor de correlação médio (tamanhos de efeito médio ponderado, $r_{médio}$) pelo tamanho da amostra (n_i), e então os dividindo pelo tamanho total da amostra (n_i).

Equação 6.

$$v._r = \frac{\sum n_i (r_i - r_{médio})^2}{\sum n_i}$$

Tamanhos de efeito reais podem ser maiores ou menores em diferentes amostras. Para calcular essa correção na variância, as possíveis distribuições foram contabilizadas no erro padrão do valor de correlação médio (EP$r_{médio}$). Dessa forma, os testes de significância estatística foram menos suscetíveis a erros de falso-positivo.

Para os cálculos do erro padrão do valor de correlação médio (EP$r_{médio}$) (**Equação 7**) foi calculado, para cada gene, a raiz quadrada da divisão da variância da amostra de correlações ($v._r$) pelo total de trabalhos em que está contido cada gene examinado (k).

Equação 7.

$$\text{EP}r_{médio} = \sqrt{\frac{v._r}{k}}$$

Com esse erro padrão do valor de correlação médio (EP$r_{médio}$) converteu-se os tamanhos de efeito médio ponderado ($r_{médio}$) de cada gene em um equivalente de padrão normal, ou um escore padrão (escore z). O escore z traduz a magnitude do efeito em unidades de desvio padrão.

Para essa conversão de escore r em escore z correspondente (**Equação 8**) foi dividido o valor absoluto de $r_{médio}$ por seu erro padrão (EP$r_{\text{médio}}$).

Equação 8.

$$z = \frac{|r_{médio}|}{EPr_{médio}}$$

O resultado z foi comparado com o valor z da significância estatística escolhida, aqui $\alpha = 0.05$. Para este valor de α, o valor z crítico (ou $z_{\alpha/2}$) é de ±1.96. Sempre que o valor z de um gene excedesse o valor $z_{\alpha/2}$ em um teste bicaudal era aceita a hipótese alternativa (H1) e, consequentemente, a significância estatística.

A segunda maneira de avaliação da significância estatística utilizada foi o cálculo do intervalo de confiança de 95% (IC 95%). Esse valor para o intervalo é análogo ao critério de significância de $p < 0.05$. A equação do cálculo da amplitude do intervalo de confiança é descrita como os tamanhos de efeito médio ponderado ($r_{médio}$) mais (para o limite superior)/menos (para o limite inferior) (**Equação 9** e **10**, respectivamente) a multiplicação do valor z crítico pelo erro padrão do valor de correlação médio (EP$r_{médio}$).

Equação 9.

$$\text{IC } 95\%_{superior} = r_{médio} + z_{\alpha/2}\ \text{SE}r_{médio}$$

Equação 10.

$$\text{IC } 95\%_{inferior} = r_{médio} - z_{\alpha/2}\ \text{SE}r_{médio}$$

Se o intervalo excluísse o valor nulo de zero, era aceita a hipótese alternativa (H1) e, consequentemente, a significância estatística.

4.3.2.3 Exame das variabilidades das distribuições dos estimados de tamanhos de efeito

Para analisar se os tamanhos de efeito estavam concentrados em um único valor de média populacional, sendo assim provindos de uma população, testou-se as hipóteses de distribuição homogênea de médias. Tal avaliação foi realizada pelo cálculo da estatística *Q*, responsável por quantificar os graus de diferença entre os tamanhos de efeito observados e os expectados. Os resultados foram comparados a uma distribuição de qui-quadrado (ou χ^2) com *k*-1 graus de liberdade, no qual *k* é o total de estudos da meta-análise em que está contido cada gene examinado.

Para os cálculos da estatística *Q* (**Equação 11**), foram multiplicadas as diferenças entre os tamanhos de efeito observados (r_i) e os expectados ($r_{médio}$) para cada estudo pelo tamanho da amostra do estudo menos o valor um (como um peso relevante) e somados todos os resultados, para cada gene.

Equação 11.

$$Q = \sum (n_i - 1)(r_i - r_{médio})^2$$

Se o valor de *Q* ultrapassasse o valor crítico do χ^2, era aceita heterogeneidade em tamanhos de efeito.

As estimativas críticas do χ^2 foram obtidas da tabela de distribuição de χ^2 (α = 0.05, graus de liberdade = *k*-1). Amostras de tamanhos de efeito tidos como heterogêneos se tornam candidatos às análises de moderador, uma variável que explique a heterogeneidade dos valores, contudo estas não foram analisadas neste trabalho.

Os genes que apresentavam $r_{\text{médio}} > |0{,}1|$ (limite inferior de efeito pequeno), valor de *p* ajustado < 0,01 e distribuição homogênea de médias (valor de *Q* < valor crítico de χ^2) foram filtrados. Os 100 genes com os maiores tamanhos de efeito positivos e negativos foram selecionados para as análises posteriores (GSP ou genes selecionados positivos e GSN ou genes selecionados negativos). No contexto da meta-análise, esses genes foram tratados como os GDE de estudos individuais obtidos em análise de expressão diferencial.

4.4 CARACTERIZAÇÕES FUNCIONAIS UTILIZANDO TESTES HIPERGEOMÉTRICOS

Para as caracterizações funcionais foram utilizados os GSP e os GSN, separadamente. As análises de enriquecimentos de ontologias foram fundamentadas em testes hipergeométricos.

Os significados biológicos/funcionais dos genes selecionados foram analisados em ontologias de processos biológicos (PB) e em vias de interação molecular. Os testes hipergeométricos para identificação de termos referentes a estes significados foram realizados com a utilização de funções do pacote *limma*, em ambiente R e do banco de dados DAVID disponível em https://david.ncifcrf.gov/ (*Database for Annotation, Visualization and Integrated Discovery*, ou banco de dados para anotação, visualização e descoberta integrada, do inglês) (DENNIS et al., 2003).

Para a área da estatística, as distribuições hipergeométricas descrevem as probabilidades de *k* sucessos em *n* retiradas sem reposição de populações de tamanhos *N* que contém exatamente *K* sucessos, onde cada retirada é tida como um sucesso/fracasso. Calcula-se sobre a distribuição hipergeométrica o teste hipergeométrico de super-representação, para obtenção da significância estatística (valor de *p* hipergeométrico) da obtenção de um número específico *k* ou mais de sucessos a partir da população em um total de *n* retiradas.

Para estudos de ontologias gênicas, os testes hipergeométricos realizados utilizam as seguintes variáveis: *N* o total de genes ou o universo de genes sendo um referencial para os testes, *M* o total de genes anotados na ontologia ou na via analisada, *n* o total de genes submetidos/utilizados como *input* (entrada, do inglês) e *m* os genes da lista submetida pertencentes a ontologia ou a via analisada. O enriquecimento e a significância estatística de testes hipergeométricos utilizados nestas duas metodologias foram calculados utilizando as **Equações 12** e **13**.

Equação 12.

$$\text{enriquecimento} = \frac{m/n}{M/N}$$

Equação 13.

$$\text{valor de } p = \sum_{m}^{\min(K,n)} \frac{\binom{M}{m}\binom{N-M}{n-m}}{\binom{N}{n}}$$

Para as análises de ontologias e de vias de interação molecular foram utilizadas as funções *goana* e *kegga* (YOUNG et al., 2010), do pacote *limma*, respectivamente. Como universo (referente a variável *N*) para execução das funções foi utilizada uma lista contendo todos os genes presentes em todos os conjuntos de dados (não duplicados), empregando seus marcadores Entrez Gene ID. Os resultados de ambas as funções foram posteriormente filtrados para selecionar os resultados mais informativos e fidedignos. Para os resultados da função *goana* foram filtrados os termos representados por menos de 3 genes e para os resultados de ambas as funções foram selecionados os termos que apresentaram valores de $p \leq 0.05$, não ajustados para múltiplos testes.

Para as análises utilizando o banco de dados DAVID, foi selecionado como universo os genes anotados de *homo sapiens*. Os resultados obtidos também foram filtrados, sendo que foram selecionados os termos que apresentaram valores de $p \leq 0.05$ em testes de valor de *p* exato de Fisher modificado, calculados automaticamente no banco de dados.

Para as duas buscas os termos usados e os genes pertencentes a estes foram selecionados de anotações de *homo sapiens.*

Em ambas as metodologias de obtenção de termos enriquecidos, os termos que foram obtidos com os dois grupos, GSP e GSN, foram anulados por serem inespecíficos. Esse processo de anulação foi feito intra e inter-metodologias.

4.5 REANÁLISE DE EXPRESSÃO DIFERENCIAL E ANÁLISE DE AGRUPAMENTO

As reanálises de todos os conjuntos de dados foram realizadas com a utilização das 3 matrizes importadas no ponto 4.2, considerando as matrizes de expressão normalizadas. As reanálises foram realizadas com métodos globais de ajustes de modelos lineares específicos ao experimento e encolhimento de variância, utilizando estimadores bayesianos empíricos, empregando as funções do pacote *limma*. Foram utilizados os contrastes “pacientes com a DP idiopática – amostras de indivíduos do grupo controle” para cada conjunto. Utilizando os ajustes de modelos lineares, foram computadas as estatísticas *t*, *F* e o *log-odds* da expressão diferencial aplicando moderação de Bayes empírica de erro padrão em direção de um valor comum. Em todas as reanálises foram utilizados os mesmos critérios para obtenção de GDE, *p*-valor $\leq 0{,}05$ para

o teste-*t* e valor absoluto positivo de *fold change* ≥ 1,2. Esses critérios foram selecionados de outros trabalhos que aplicaram análise de expressão diferencial ou meta-análise em sangue de pacientes com a DP idiopática (SANTIAGO; LITTLEFIELD; POTASHKIN, 2016; SANTIAGO; POTASHKIN, 2017; SCHERZER et al., 2007) e outras doenças do SNC e detectadas no sangue periférico como a doença de Alzheimer (BAI et al., 2014), acidente vascular cerebral isquemático (DYKSTRA-AIELLO et al., 2016) e nos pacientes com malformação arteriovenosa cerebral (WEINSHEIMER et al., 2011)

Foram observados o total de GDE em cada conjunto de dados supra- e infra-regulados e o total de GSP e de GSN que fazem parte dos GDE de cada conjunto de dados. Foram filtrados os GSP e os GSN de todas as reanálises para observar as expressões de genes selecionados na meta-análise em cada conjunto de dados. Foram realizados agrupamentos hierárquicos de amostras de cada estudo utilizando os GSP e os GSN para observar a distribuição das amostras com esses genes/preditores em estudos separados. Para os agrupamentos hierárquicos foram utilizadas as distâncias de 1-valores de correlação de Pearson e os métodos de aglomeração de Ward (WARD, 1963) utilizando as funções *cor* (*Correlation, Variance and Covariance* ou Correlação, variância e covariância, em inglês), *as.dist* (*Distance Matrix Computation* ou Computação em matriz de distância, em inglês) e *hclust* (*Hierarchical Clustering* ou Agrupamento hierárquico, em inglês) do pacote *stats*. A fórmula gráfica de representação foi obtida com a função *heatmap.2* (*Enhanced Heat Map* ou Mapa de calor aprimorado, em inglês) do pacote *Hmisc* (*Harrell Miscellaneous*, ou Miscelânea de Harrell, em inglês) (HARRELL JR; DUPONT, 2018).

Os grupos criados pela divisão seguida do primeiro nó, a primeira partição, foram avaliados para separação de amostras em dois grupos conforme o fenótipo de pacientes com a DP idiopática e indivíduos do grupo controle. Para as análises de pureza dos grupos foram calculadas as porcentagens de presença de cada fenótipo em cada grupo. Sendo assim, a acurácia foi prevista com base nos grupos com maiores valores de acerto de classe. O grupo com maior porcentagem de algum fenótipo foi o primeiro rotulado quanto a classe. Subsequentemente, o outro grupo foi rotulado com a classe remanescente. Os cálculos de acurácia foram realizados pela soma do total de verdadeiros positivos e verdadeiros negativos dividido pelo total de amostras do grupo. O maior valor de acurácia entre os dois grupos é o que foi explorado. A taxa de não-informação assinala a acurácia que pode ser obtida sem modelo e foi

obtida da proporção de classe com a maioria dos membros (KUHN; JOHNSON, 2013).

4.6 ALGORITMOS DE APRENDIZADO DE MÁQUINA

4.6.1 Pré-processamento dos dados

Para execução de alguns algoritmos de aprendizado de máquina, principalmente os baseados em dimensões de distâncias, é necessária uma transformação dos dados. Para isso, as matrizes normalizadas, obtidas em 4.2, tiveram seus valores reescalonados (LANTZ, 2015). Para o reescalonamento de preditores, as variáveis dependentes do modelamento, os intervalos necessitam ser comprimidos ou expandidos para que cada gene contribua de forma relativamente igual nas análises. A metodologia selecionada para transformação foi o escore *z*. Os cálculos de escore *z* (**Equação 14**) foram realizados para cada observação (cada gene de cada amostra) de cada estudo. O cálculo envolve subtrair o valor da observação pelo valor da média do preditor e dividir o resultado pelo desvio padrão do preditor.

Equação 14.

$$z = \frac{\mathrm{X} - \mu}{\sigma} = \frac{\mathrm{X} - \text{média(X)}}{\text{desvio padrão(X)}}$$

Essa fórmula é baseada em propriedades da distribuição normal, reescalonando os valores em termos de quantos desvios padrão eles apresentam de distanciamento do valor da média. O escore *z* não possui um limite de números positivos ou negativos e não apresenta um mínimo ou máximo predefinido (MENDENHALL, WILLIAM M., SINCICH, TERRY L., S. BOUDREAU, 2016).

As matrizes reescalonadas de cada estudo foram compiladas em uma megamatriz com os dados de todas as amostras da meta-análise.

4.6.2 Preparação dos dados

Uma das questões mais críticas da análise de aprendizados de máquina é a seleção dos preditores que serão envolvidos na elaboração do modelo, principalmente com o advento de métodos de alto rendimento. Os experimentos de microarranjos de DNA e as tecnologias de sequenciamento podem medir numerosos valores de expressão de RNA

em diversas amostras, produzindo uma variedade de preditores numéricos (LANTZ, 2015). Do ponto de vista prático, modelos que utilizam menos preditores podem ser facilmente explicados e são menos custosos, especialmente se há um custo na mensuração de preditores, como no caso de microarranjos, reação em cadeia da polimerase (PCR) e outros. Estatisticamente, é mais atrativo calcular poucos parâmetros, além do fato de algumas técnicas serem afetadas negativamente por preditores não/pouco informativos (KUHN; JOHNSON, 2013).

Uma das maneiras de selecionar os preditores é eliminar as informações redundantes. O nome técnico para a situação em que um par de parâmetros preditores possuem correlação substancial é colinearidade, ou multicolinearidade no caso de múltiplos preditores. Para lidar com estas situações, uma das soluções (bastante heurística) é eliminar o número mínimo de preditores cujas correlações estão acima de um limiar. Esse método identifica colinearidades em duas dimensões e possui um efeito significativamente positivo na execução de alguns algoritmos. Esse método calcula matrizes de correlação de preditores, determina os preditores associados as maiores correlações em números absolutos e elimina todas as correlações acima do limiar estabelecido (KUHN; JOHNSON, 2013). Para filtragem das colinearidades foi empregada a função *findCorrelation* do pacote *caret* (*Classification and Regression Training* ou Treinamento de classificação e regressão, em inglês) (KUHN et al., 2018). A função aponta a posição das colunas com preditores recomendados a deleção. Para examinação das estruturas de correlação de dados, foi empregada a função *corrplot* do pacote igualmente denominado (*Visualization of a Correlation Matrix* ou Visualização de matriz de correlação, em inglês) (WEI; SINKO, 2017). O limiar estabelecido foi de 0,75, utilizando os valores de correlação de Pearson.

Outra forma de eliminação de preditores comumente utilizada é o algoritmo de eliminação de preditores recursivo, ou algoritmo de seleção "para trás" (*backward selection algorithm*, em inglês) (GUYON; ELISSEEFF, 2003). Em cada etapa essa operação ranqueia os preditores e os com menor relevância são iterativamente excluídos antes da modelagem. Essa prática elimina a montagem de diversos modelos em cada etapa de busca. Os cálculos de importância podem ser baseados no modelo (como os cálculos de importância obtidos com o algoritmo *random forest*) ou empregar diferentes abordagens (KUHN; JOHNSON, 2013). Para realizar essa eliminação de preditores, foram empregadas as funções *rfeControl* (*Controlling the Feature Selection Algorithms*, ou controlando os algoritmos de seleção de preditores, em inglês) e *rfe* (*backward selection algorithm*) do pacote *caret*. Os cálculos de

importância de preditores utilizados foram baseados em *random forest*. O algoritmo de eliminação de preditores recursivo foi utilizado com o método de reamostragem de validação cruzada de 10-*fold* e com todas as possibilidades de "número de preditores que devem ser retidos", de 1 a *n*, sendo *n* o total de preditores.

Os genes selecionados nessas duas etapas de filtragem foram definidos como os partícipes da assinatura gênica.

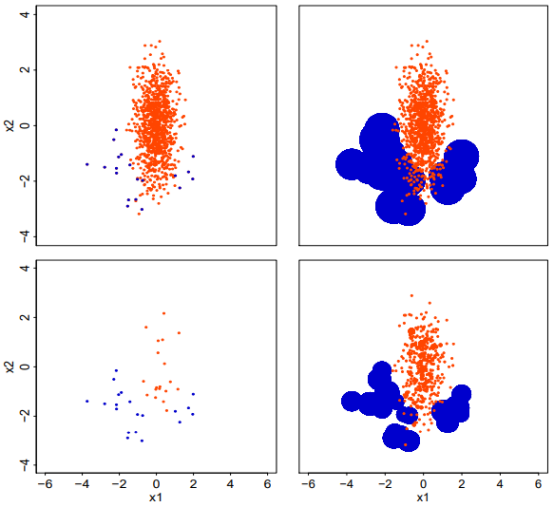
A construção de modelos de aprendizado de máquina é limitada pelos dados. Em diversos cenários os dados podem ser limitados pelo número de amostras, que podem ser de baixa qualidade a desejável e/ou podem não ser representativas da população. Nestes pressupostos, deve-se utilizar as informações disponíveis para conseguir o melhor modelo preditivo. Quase todos as técnicas de modelagem preditiva apresentam parâmetros de ajuste, ou *hiperparâmetros*, que auxiliam no encontro da estrutura entre os dados. Dessa forma, deve-se utilizar os dados existentes para esclarecer as configurações de hiperparâmetros que geram modelos realistas e de alto desempenho (KUHN; JOHNSON, 2013). Para isso, os dados existentes devem ser divididos em dois grupos, o grupo-treino (GTr) e o grupo-teste (GTe). O GTr é utilizado para construir e ajustar os modelos e o GTe é utilizado para estimar o desempenho de predição do modelo. Idealmente, as predições não devem ser realizadas em amostras utilizadas na construção do modelo para que elas forneçam imparcialidade da eficácia do modelo (LANTZ, 2015). Para as análises, os dados foram separados em dois grupos sendo que o GTr utilizou 80% das amostras para afinação de modelos e o GTe utilizou os 20% restante de amostras para avaliação das predições.

Outra questão a ser abordada são os problemas de classificação binária na presença de classes desbalanceadas nos GTr, ou seja, conter no estudo mais amostras de um grupo experimental do que de outro, fazendo com a modelagem seja voltada para uma das classes, principalmente (KUHN; JOHNSON, 2013). Para lidar com estas questões, pode-se utilizar um pacote intitulado *ROSE* (*Random Over-Sampling Examples* ou Exemplos aleatórios de amostragem excessiva, em inglês) (LUNARDON; MENARDI; TORELLI, 2014). Esse pacote contém funções capazes de criar amostras sintéticas balanceadas e assim melhorar as estimativas de classificadores binários. A técnica é baseada em *bootstrap* e pode auxiliar na presença de classes raras (como em doenças de baixa incidência). Essa técnica lida com dados contínuos e categóricos gerando exemplos (amostras) sintéticos da estimativa de densidade condicional das duas classes. Diferentes métricas são utilizadas com objetivo de analisar a precisão da função. Para correção do

desbalanceamento das classes foi empregue a função *ovun.sample* (*over-sampling, under-sampling, combination of over- and under-sampling* ou superamostragem, subamostragem, combinação de super e subamostragem em inglês) do pacote *ROSE*, fabricando três grupos adicionais de treino, (1) um grupo-treino com superamostragem do grupo-amostral com menos amostras, o *GTr superamostragem*, (2) um grupo-treino com subamostragem do grupo-amostral maior, o *GTr subamotragem*, e (3) um grupo-treino buscando equilíbrio entre os grupos supra e infrarepresentados, o *GTr combinados*.

A **Figura 6**, retirada de Lunardon, Menardi e Torelli (2014), busca exemplificar as metodologias de balanceamento do pacote *ROSE* utilizando os dados de *hacide* presentes no mesmo. Estes dados simulam a classificação binária desbalanceada, sendo que uma das classes é representada por 2% dos dados, somente. As cores laranja e azul denotam as classes majoritárias e minoritárias, respectivamente. Há 980 amostras no grupo maior e 20 amostras do grupo menor (**Figura 6**, canto superior esquerdo).

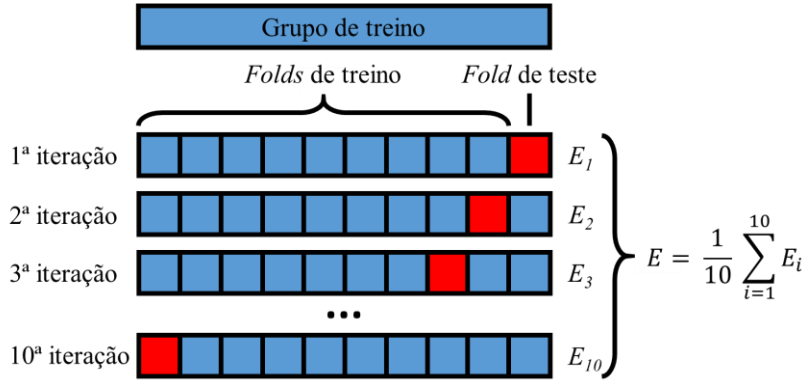
Figura 6. Resultados das estratégias de balanceamento de dados implementadas no pacote *ROSE* utilizando os dados de *hacide*.



Em laranja: Classe majoritária. Em azul: Classe minoritária. O painel superior esquerdo mostra os dados de treino, o painel superior direito mostra um exemplo de superamostragem do grupo-amostral com menos amostras, o painel inferior esquerdo mostra um exemplo de subamostragem do grupo-amostral maior e o painel inferior direito mostras a combinação de super e subamostragem. Extraído de Lunardon, Menardi e Torelli (2014).

Para estimar os desempenhos futuros de modelos durante os ajustes de hiperparâmetros, estes foram avaliados em dados ainda não apresentados, aproveitando as técnicas de validação cruzada de 10-*fold* (*10-fold cross validation*, em inglês) (**Figura 7**). Esse método separa os GTr em 10 partes de iguais grandezas, ou *folds*, randomicamente (10 partes contendo 10% dos GTr). Para cada um dos 10 *folds* um modelo é criado utilizando os 90% dos demais elementos e o *fold* correspondente ao 10% amostral é utilizado na avaliação do modelo. Após esse processo de formação de 10 modelos/avaliações, utilizando 10 divergentes combinações para treinamentos e testes, o desempenho médio de todos os *folds* é o relatado. Como essa técnica destaca randomicamente os 10 *folds*, em cada utilização ela foi repetida 3 vezes e, sendo assim, o desempenho médio declarado é resultado de 30 combinações para treinamentos e testes distintas.

Figura 7. Validação cruzada de 10-*fold*.



Em cada iteração: Em azul: *folds* utilizados para elaboração de modelos. Em vermelho: *fold* utilizado para avaliação do modelo. E_i: desempenho de cada iteração. E: desempenho médio de todas as iterações.

Para as análises de predição utilizando os GTr e os GTe foram calculadas em tabulações cruzadas (*Cross-tabulation*, em inglês) as classes observadas (classes reais) e as preditas, com estatísticas relacionadas utilizando a função *confusionMatrix* do pacote *caret*. Para as questões de duas categorias foram selecionados dados de valores de verdadeiro positivo (predição correta positiva, aqui pacientes com a DP idiopática), verdadeiro negativo (predição correta negativa, aqui

indivíduos do grupo controle), falso positivo (predição incorreta positiva), falso negativo (predição incorreta negativa), as acurácias, as taxas com não informação (*No Information Rate* ou NIR, em inglês), as significâncias estatísticas de valores de acurácia sobre os valores de NIR, as sensibilidades (predição correta positiva sobre o total de positivos, ou seja, as proporções de sucessos em amostras de pacientes com Parkinson idiopático) e as especificidades (predição correta negativa sobre o total de negativos, ou seja, as proporções de sucessos em amostras de pacientes do grupo controle). As áreas sob a curva ROC (*Area under the ROC curve* ou AUC, em inglês) foram calculadas utilizando a função *predict* do *caret* empregando o parâmetro *type = "prob"*, para se conseguir as probabilidades de classes de cada amostra e as utilizar em funções *roc* do pacote *pROC*.

A **Figura 8** representa um organograma da metodologia de preparação dos dados, envolvendo os algoritmos de seleção de preditores, elaboração dos 4 GTr e do GTe, indica os passos do modelamento, utilizando diferentes algoritmos e múltiplas combinações de hiperparâmetros, seleção dos modelos, a combinação algoritmos/hiperparâmetros com maiores valores de AUC, e a avaliação das predições obtidas com estes modelos.

4.6.3 Algoritmos de aprendizado de máquina

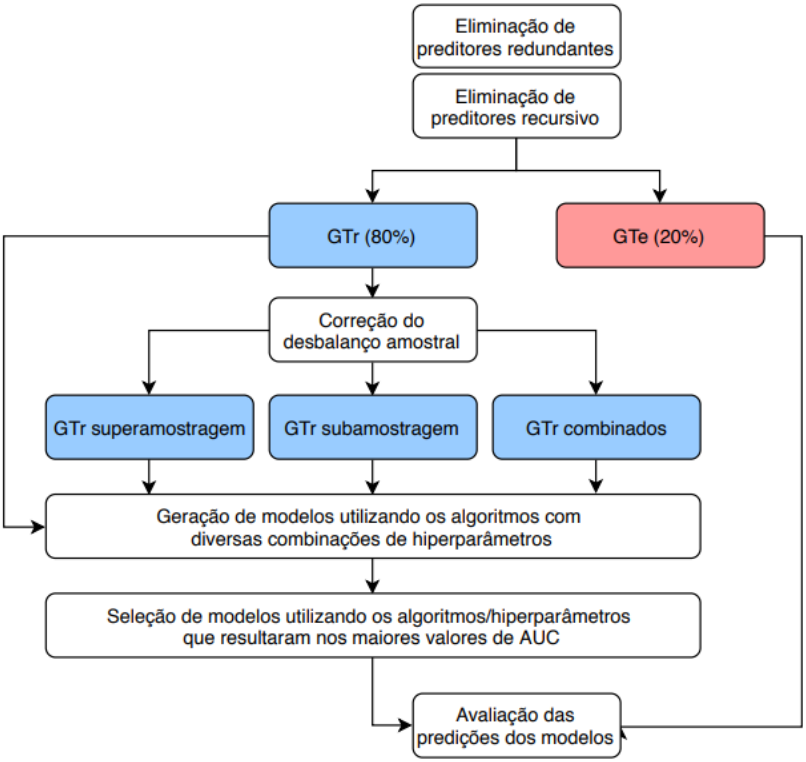
4.6.3.1 *k-Nearest Neighbors*

Os classificadores de *nearest neighbors* (ou vizinhos mais próximos em inglês) são métodos supervisionados utilizados para classificação e regressão. São definidos por classificar amostras não marcadas (sem uma classe, por exemplo sem um diagnóstico) por imputar classes de exemplos similares marcados (COVER; HART, 1967).

De forma geral, os classificadores de *nearest neighbors* são bem adequados às tarefas de classificação em que as relações entre os preditores e as classes-alvo são abundantes, complicadas ou extremamente difíceis de se entender, porém com itens da mesma classe sendo bastante homogêneos. Entretanto, se os dados forem ruidosos e, portanto, não houver uma distinção explícita de grupos amostrais, os algoritmos de *nearest neighbors* podem encontrar dificuldade na identificação de limites entre as classes (LANTZ, 2015).

O algoritmo *k-Nearest Neighbors* é um dos mais simples e básicos aprendizados de máquina, e é extensivamente utilizado. O algoritmo apresenta esse nome por utilizar a informação dos *k* vizinhos mais próxi-

Figura 8. Organograma da preparação de dados para as análises de aprendizado de máquina.



Em azul: diferentes GTr utilizados para modelamento. Em vermelho: GTe utilizado para predições.

mos para classificar exemplos não marcados (ALTMAN, 1992). A letra *k* é uma variável que indica que qualquer valor pode ser utilizado. Após escolhido o valor *k*, o algoritmo manipula um GTr feito de exemplos marcados, classificados em diferentes categorias, situando-os em um espaço bidimensional conforme os seus valores de variáveis preditoras. Para cada amostra não marcada do GTe, o algoritmo *k-Nearest Neighbors* identifica as *k* amostras do grupo-treino mais próximas (os *k* vizinhos mais próximos) e rotula o registro conforme a classe da maior parte de *k* vizinhos (COVER; HART, 1967).

A similaridade entre as amostras é mensurada pela distância e, tradicionalmente, o algoritmo utiliza a distância euclidiana para essa análise. A distância euclidiana se refere rota direta mais curta, como

conectar dois pontos com uma linha (LANTZ, 2015). Essa medida de distância é calculada segundo a **Equação 15** onde *p* e *q* são referentes aos exemplos a serem comparados, cada qual contendo *n* preditores. O termo p_1 se refere ao valor do primeiro preditor do exemplo *p*, enquanto q_1 se refere ao valor do primeiro preditor do exemplo *q*.

Equação 15

$$\text{distância euclidiana}(p,q) = \sqrt{(p_1 - q_1)^2 + \cdots + (p_n - q_n)^2}$$

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "knn", referente a *k-Nearest Neighbors*. O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 vezes e foram testadas todas as possibilidades de "número de vizinhos mais próximos" de 1 a *n*, sendo *n* o total de preditores (hiperparâmetro "k"). A métrica para seleção do melhor modelo escolhida foi a " ROC", referente a área sob a curva ROC.

4.6.3.2 *Naive Bayes*

A algoritmo de aprendizado *Naive Bayes* (ou Bayes ingênuo em inglês) são métodos supervisionados utilizados para classificação e regressão. O algoritmo descende do trabalho do século XVIII do matemático Thomas Bayes que descreveu os princípios da probabilidade de eventos e como elas deveriam ser revistas à luz de novas informações. Esse princípio fundamentou o que é conhecido como técnicas bayesianas (LANTZ, 2015).

Probabilidades são apresentadas por números entre 0 e 1 (referentes a 0 e 100%) que capturam as chances de eventos ocorrerem a luz das evidências existentes. Quanto menor for a probabilidade, menor é a chance do evento ocorrer. Classificadores baseados em técnicas bayesianas utilizam os dados do GTr para calcular probabilidades observadas de cada resultado baseado em evidências providas por valores dos preditores. Para cada amostra não marcada, o classificador utiliza as probabilidades observadas para predizer a provável classe se baseando em valores de preditores (MURTY; DEVI, 2012).

Os classificadores bayesianos são bem adequados as tarefas de classificação na qual as informações de numerosos atributos devem ser juntamente consideradas para calcular a probabilidade de um resultado. Enquanto diversos algoritmos ignoram preditores menos relevantes, os métodos bayesianos aplicam todas as evidências existentes para

modificar sutilmente as predições. A ideia é que um grande número de preditores menos relevantes podem possuir um impacto combinado expressivo (WEBB; BOUGHTON; WANG, 2005). O algoritmo *Naive Bayes* é assim chamado porque ele fez suposições "ingênuas" sobre os dados. O algoritmo pressupõe que todos os preditores apresentam independência e igual relevância. Essas suposições são raramente verdadeiras nas aplicações reais (LANTZ, 2015).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "nb", referente a *Naive Bayes*. O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 vezes e foram testados diversos valores de correção de Laplace (hiperparâmetro "fL") e as estimações de densidades normal e de kernel (hiperparâmetro "usekernel"). A métrica para seleção do melhor modelo escolhida foi a " ROC", referente a área sob a curva ROC.

4.6.3.3 *Decision Trees*

Os algoritmos de aprendizado *Decision Trees* (ou àrvores de decisão em inglês) são métodos supervisionados utilizados para classificação e regressão. São potentes classificadores que utilizam a estrutura de árvores e ramificações de decisão para modelar a relação entre os preditores e os potenciais resultados (LANTZ, 2015). Na estrutura de árvore todos os dados começam no nó raiz (ou *root node*, em inglês) e cada nó representa um valor de preditor que separará as amostras em dois grupos. No caso de uma decisão final ser feita, a árvore de decisão termina em "folhas", também conhecidos como nós terminais que designam a resposta da série de decisões. Um benefício de algoritmos de árvores de decisão é a sua disposição em forma de fluxograma que não é necessariamente exclusivo do algoritmo, mas que resultam em um formato legível para humanos (JAMES et al., 2013).

As árvores de decisão são construídas utilizando um algoritmo designado particionamento recursivo (ou *recursive partitioning*, em inglês). Essa abordagem, também denominada "dividir e conquistar", divide seguidamente os dados em subconjuntos até que o processo pare quando o algoritmo determina que os dados estão suficientemente homogêneos ou até que outro critério de pausa tenha sido obtido. Para escolher o melhor ponto de cisão ou divisão a operação deve primeiro identificar o melhor preditor que decompõe as amostras em dois grupos cujos elementos sejam puros ou o mais puro possível. Subgrupos puros são aqueles compostos por elementos de uma só categoria. Os algoritmos

devem fazer isso continuamente até que os nós terminais contenham somente grupos puros (STROBL; MALLEY; TUTZ, 2009).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "dt", referente a *Decision Trees*. O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 vezes e foram testadas diversas opções do menor valor de observações em um nó para que ele possa ser dividido (hiperparâmetro "minsplit") e de profundidade máxima da árvore, indicada pelo tamanho do maior caminho entre o nó raiz e as folhas (hiperparâmetro "maxdepth"). A métrica para seleção do melhor modelo escolhida foi a " ROC", referente a área sob a curva ROC.

4.6.3.4 *Support Vector Machine*

Os algoritmos de *Support Vector Machine* (ou Máquina de suporte de vetores, em inglês) são métodos supervisionados utilizados para classificação e regressão. Situam as amostras em um espaço *n*-dimensional, onde *n* é o número de preditores e as coordenadas são relativas ao valor de amostras para cada preditor (LANTZ, 2015). Um *Support Vector Machine* é uma superfície que cria uma fronteira entre os dados. Os dados foram classificados pela posição de um hiperplano (fronteira) que diferencia as duas classes. Em algoritmos de máquina de suporte de vetores, um hiperplano ótimo é o que maximiza os espaços entre os elementos de ambas as classes e o suporte. Os algoritmos de *Support Vector Machine* combinam aspectos de aprendizados do tipo de redes neurais baseado em instâncias e de regressões lineares e essa combinação capaz e estável permite que o algoritmo modele relações altamente complexas (CORTES; VAPNIK, 1995).

Os dados podem ser chamados de lineares, caso um hiperplano em linha reta possa dividir as classes e não lineares, quando isso é tido como impossível. Em casos de dados não lineares se adiciona uma terceira dimensão (ou dimensão *z*), geralmente utilizando uma função kernel, o kernel radial para mudar o espaço de busca de um hiperplano (KUHN; JOHNSON, 2013).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "svmLinear", referente a *Support Vector Machine* com função de kernel linear e "svmRadial", referente a *Support Vector Machine* com função de kernel de base radial, que utiliza funções kernel. O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 vezes e foram testadas diversas possibilidades de Custo para os modelos linear e radial, um modo

de ponderar as classificações introduzindo um custo em classificações erradas (hiperparâmetro “C” e Sigma para os modelos radiais criando classificadores locais ou mais generalistas, utilizando os diferentes valores (hiperparâmetro “sigma”). A métrica para seleção do melhor modelo escolhida foi a “ ROC”, referente a área sob a curva ROC.

4.6.3.5 *Bagging*

Os algoritmos de *Bagging*, ou *Bootstrap aggregating* (ou agregação por *bootstrap*, em inglês) são métodos supervisionados utilizados para classificação e regressão. Foram uma das primeiras técnicas *emsemble* (de conjunto, em inglês) a ser difundida. Técnicas de conjunto agregam múltiplos modelos de aprendizado de máquina, permitindo investigar a melhor performance global. A lógica se baseia na utilização de modelos individuais fracos, porém fortes quando analisados conjuntamente. O *Bagging* gera um grande número de GTr que são grupos de amostras de treino preparados por amostragem. Esses GTr são utilizados na geração de um grupo de modelos empregando um único algoritmo de aprendizado. Para classificação, as predições são combinadas por voto (BREIMAN, 1996).

Embora os algoritmos de *Bagging* sejam técnicas de conjunto relativamente simples, eles tendem a apresentar bons resultados quando usados com algoritmos de aprendizado relativamente instáveis, ou seja, algoritmos que geram modelos que tendem a variar substancialmente quando os dados mudam ligeiramente. Por essa razão, o *Bagging* é frequentemente utilizado com os algoritmos de árvores de decisão, que tendem a variar drasticamente com mudanças nos dados de entrada (KUHN; JOHNSON, 2013).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como “treebag”, referente a *Bagged Classification and Regression Trees* (ou Árvores de classificação e regressão por “ensacamento”, em inglês), que utiliza árvores de decisão como os algoritmos de construção de modelos. O método de reamostragem escolhido foi o “repeatedcv”, referente a validação cruzada de 10-*fold* repetida 3. A métrica para seleção do melhor modelo escolhida foi a “ ROC”, referente a área sob a curva ROC.

4.6.3.6 *Random Forest*

Os algoritmos de *Random Forest* (floresta randômica ou floresta aleatória, em inglês) são métodos supervisionados utilizados para

classificação e regressão. É mais uma das técnicas de conjunto, porém focada unicamente em conjuntos de árvores de decisão. O *Random Forest* combina os princípios básicos do *Bagging* com o adicional da geração de um grande número de conjuntos de preditores preparados por amostragem para adicionar uma maior diversidade aos modelos de árvores de decisão (KUHN; JOHNSON, 2013). São gerados GTr com amostras e preditores diferentes para elaboração de diversas árvores de decisão randômicas. Esses GTr são utilizados na geração de um grupo de modelos empregando um único algoritmo de aprendizado. Para classificação, as predições são combinadas por voto (BREIMAN, 2001).

A principal distinção entre o algoritmo de *Random Forest* e as árvores de decisão é que o procedimento de encontrar os nós "raiz", a primeira secessão, é feita randomicamente. Os algoritmos de *Random Forest* combinam poder e versatilidade. Como os conjuntos manipulam apenas uma porção randômica do total de preditores, as florestas podem lidar com dados grandes, onde outros modelos falham pela "maldição da dimensionalidade" (LANTZ, 2015).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "rf", referente a *Random Forest*. O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 e foram testadas diversas possibilidades de do hiperparâmetro "mtry" e "ntree" primeiramente de forma randômica e ulteriormente direcionada aos resultados superiores. O hiperparâmetro "mtry" é relacionado ao número de preditores selecionados aleatoriamente e "ntree" é relacionado ao número de árvores geradas para comparação. A métrica para seleção do melhor modelo escolhida foi a "ROC", referente a área sob a curva ROC.

4.6.3.7 *Gradiente Boosting Machine*

Algoritmos de *Boosting* (otimização, em inglês) fazem parte de algoritmos de *emsemble* como o *Bagging* e o *Random Forest*. Entretanto a utilização de muitos algoritmos não ocorre em paralelo, com modelos sendo criados ao mesmo tempo, e sim sequencialmente, onde os esquemas subsequentes aprendem com erros dos modelos prévios. Dessa forma, as amostras têm uma probabilidade desigual de aparecer em modelos subsequentes, aparecendo mais as de maior complexidade de catalogação (SCHAPIRE, 2003).

Os algoritmos de *Gradient Boosting Machine* (Máquina de otimização de gradientes, em inglês) são métodos supervisionados utilizados para classificação e regressão. Esses algoritmos criam modelos

seriais e identifica suas deficiências na forma de gradientes em funções de perda. Funções de perda são modos de avaliar a performance do modelo. Se os modelos são capazes de acertar muitas predições, o resultado é um valor baixo. O *Gradient Boosting Machine* busca esses resultados para suas avaliações e sequenciamento. Os algoritmos escolhem iterativamente as funções que apontam na direção do gradiente negativo de valores resultantes de funções de perda (MASON et al., 2000).

Uma das causas da popularidade do *Gradient Boosting Machine* é o fato de que se bem ajustado, ele atuará melhor que algoritmos estado da arte em *deep learning* (aprendizagem profunda, em inglês). Entre as principais desvantagens da utilização desses algoritmos é a sua tendência ao super-ajustamento (*overfitting* em inglês) que é a falta de generalização do modelo, criando modelos aptos a lidar somente com as amostras de grupos-treino e por tal motivo é necessário utilizar pontos de parada antecipada dos gradientes (LANTZ, 2015).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "gbm", referente a *Stochastic Gradient Boosting Machine* (ou máquina de otimização de gradientes estocástico, em inglês). O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 e foram testadas diversas possibilidades de hiperparâmetros "n.trees", "interaction.depth", "shrinkage" e "n.minobsinnode" primeiramente de forma randômica e posteriormente direcionada aos resultados superiores. O hiperparâmetro "n.trees" é relacionado a abundância de iterações de otimização, "interaction.depth" é relacionado a profundidade máxima da árvore, "shrinkage" é relacionado a contribuição (ou ponderação) de cada árvore aos resultados das etapas e "n.minobsinnode" é relacionado ao volume mínima do nó terminal. A métrica para seleção do melhor modelo escolhida foi a "ROC", referente a área sob a curva ROC.

4.6.3.8 *eXtreme Gradient Boosting*

Os algoritmos de *eXtreme Gradient Boosting* (Otimização de gradientes extremo, em inglês) são métodos supervisionados utilizados para classificação e regressão. O algoritmo é congênere ao *Gradient Boosting Machine* cujo nome se refere ao objetivo da engenharia de elevar o limite de recursos de computação para algoritmos de árvores de decisão otimizadas (CHEN; GUESTRIN, 2016).

A popularidade desses algoritmos deve-se as suas velocidades e performances. Devido ao fato do núcleo do *eXtreme Gradient Boosting*

ser paralelizável ele pode se aproveitar do poder de processamento de computadores de múltiplos núcleos, do paralelizamento em unidades de processamento gráficas (GPU ou *Graphics Processing Unit*, em inglês) e entre redes de computadores, tornando possível criar modelos de GTr enormes. Além do mais, em performance os algoritmos de *eXtreme Gradient Boosting* superam continuamente os outros algoritmos em competições de aprendizado de máquina (NIELSEN, 2016).

Para modelagem foi utilizada a função *train* do pacote *caret* especificando o método como "xgbTree", referente a *eXtreme Gradient Boosting*. O método de reamostragem escolhido foi o "repeatedcv", referente a validação cruzada de 10-*fold* repetida 3 e foram testadas diversas possibilidades de hiperparâmetros "nrounds", "max_depth", "eta" e "colsample_bytree" primeiramente de forma randômica e posteriormente direcionada aos resultados superiores. O hiperparâmetro "nrounds" é relacionado a abundância de iterações de otimização, "max_depth" é relacionado a profundidade máxima da árvore, "eta" é relacionado a contribuição (ou ponderação) de cada árvore aos resultados das etapas e "colsample_bytree" é relacionado a proporção de conjuntos de amostras ao produzir as árvores. A métrica para seleção do melhor modelo escolhida foi a "ROC", referente a área sob a curva ROC.

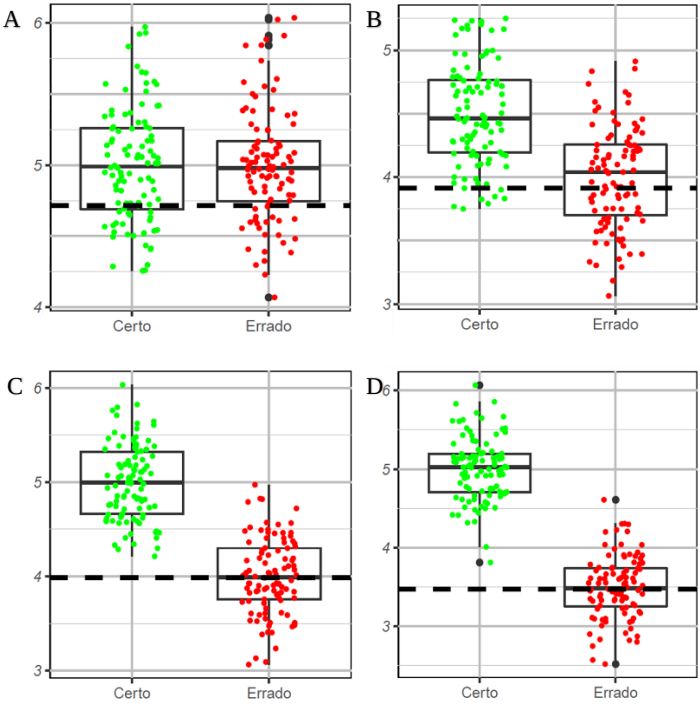
4.6.4 Avaliação e seleção de modelos criados

Para as análises de predição foram calculadas em tabulações cruzadas com estatísticas relacionadas utilizando a função *confusionMatrix* do pacote *caret*. Dessa forma, foram escolhidos os algoritmos com os maiores valores de sensibilidade e especificidade treinados e ajustados com cada GTr. Destes, suas competências de indicação foram avaliadas quanto aos estudos e quanto aos fenótipos das amostras em cada estudo.

Os modelos selecionados com cada GTr foram otimizados utilizando a remoção dos 25% das amostras do GTe com os menores valores de probabilidade da classe *Parkinson* em modelos de maior sensibilidade e dos 25% das amostras do GTe com os menores valores de probabilidade de classe *Controle* em modelos de maior especificidade. Essa metodologia foi considerada pela possibilidade de haver diferenças entre as distribuições de valores de probabilidade de classes nas predições corretas e incorretas dos modelos criados. As diferenças de distribuições foram mensuradas por meio das médias, utilizando um teste de normalidade de Shapiro-Wilk seguido de testes *t* ou Wilcoxon-Mann-Whitney, dependendo das distribuições dos valores. A **Figura 9**

representa diferentes distribuições de valores (representando probabilidade de classes), hipotéticos, criados randomicamente no R, em amostras de predição corretas e incorretas e o efeito do limiar, eliminação dos 25% com menores valores, no corte de amostras incorretas.

Figura 9. Comparações de diferentes médias e o efeito do limiar de corte de 25% das amostras.



Cada quadro contém 200 amostras, 100 “corretas” e 100 “incorretas”. A linha tracejada determina os valores de separação, os limites, entre as 25% de amostras de menor valor. Em todos os quadros o grupo *Certo* representa 100 amostras em uma distribuição com média = 5 e desvio padrão = 0,4, já as distribuições do grupo *Errado* apresentam médias de 5 (**A**), 4,5 (**B**), 4 (**C**) e 3,5 (**D**) e desvio padrão = 0,4.

Sendo assim, os modelos são capazes de predizer um grupo de amostras menor, por ter que passar de um limiar de probabilidade de classe, porém com maior certeza. As acurácias de GTe foram recalculadas utilizando os resultados de modelos otimizados. Foram escolhidos os

modelos com o maior valor de probabilidade de classe acima do corte pós-otimização para composição do par de modelos final, o par ótimo.

4.6.5 Testes para outras condições

Para as análises da predição com o par ótimo utilizando exemplares de outras categorias, os conjuntos obtidos foram examinados na presença de grupos que não fossem de DP idiopática ou do grupo controle, com pelo menos 5 amostras. Cada amostra alusiva a outro grupo foi extraída de sua matriz originária de expressões com os valores normalizados para todos os genes da assinatura e adicionada em uma matriz com as amostras de dois grupos de interesse, produzindo matrizes de amostras de Parkinson e amostras do grupo controle mais um (a amostra adicionada). Essa matriz teve seus valores de expressão escalonados em *z*. A amostra anexada foi retirada e aplicada na predição de modelos escolhidos do par ótimo. Os valores de probabilidades de classes foram comparados com os valores do limiar de sensibilidade e sensitividade dos modelos com remoção dos 25% das amostras do GTe com os menores valores de probabilidade da classe. A amostra somente foi classificada se os valores de probabilidades de classes fossem maiores do que os valores limites.

4.7 ELABORAÇÃO DE REDES DE CORRELAÇÃO GÊNICA E OBTENÇÃO DE REDES DE CO-EXPRESSÃO GÊNICA

As redes de correlação de genes foram elaboradas utilizando um ponderamento, ou peso, em seus conectores para que sua forma retratasse a predição sobre as correlações. Para essa elaboração foram calculados os valores de correlação de genes de cada estudo e as combinações de valores de *p* utilizando o método de Fisher (FISHER, 1932).

Foram elaboradas duas matrizes de valores de expressão normalizados para cada estudo, contendo os GSP e GSN separadamente. Para os cálculos de obtenção de coeficientes de correlação de Pearson e de valores de significância estatística de cada gene em cada estudo foram empregues a função rcorr, presente no pacote Hmisc. Essa função produz duas matrizes, ambas com genes nos eixos, uma matriz de coeficientes de correlação e uma com os valores de significância estatística para cada par gênico. Os nomes dos pares de genes (ex. “Gene01-Gene02”) e os dois valores obtidos pela função, retirados de cada matriz criada, foram utilizados na elaboração de tabelas individuais de cada estudo.

Os valores de significância estatística de cada gene foram combinados entre os conjuntos de dados utilizando o método de

combinação de valores de *p* pela soma de logaritmos, conhecido como método de Fisher ou método do qui-quadrado. Esse método combina os valores extremos de testes de probabilidade, os valores de *p* em um teste estatístico de χ^2. Para os cálculos do método de Fisher (**Equação X**) foram somados os valores de *p* transformados em logaritmo de cada estudo contendo o gene e o resultado multiplicado -2. O valor final é um χ^2 com 2 *k* de graus de liberdade (FISHER, 1932), onde *k* é o número de conjuntos abarcando o gene. A relação entre os valores de *p* e o do teste estatístico de χ^2 são inversamente proporcionais, então genes estatisticamente significantes devem apresentar um valor elevado de χ^2 propondo que a hipótese nula não é verdadeira.

$$\chi^2_{2k} \sim -2\sum_{i=1}^{k} p_i$$

Para obtenção de redes de co-expressão gênica dos GSP e GSN foi utilizado um banco de dados de redes descritas, o GeneMANIA (http://www.genemania.org/) e foram procuradas as interações significativas. As redes podem ser elaboradas utilizando diferentes caracteres de conectores, para este experimento foram utilizados os dados de co-expressão, onde dois genes estão ligados se seus valores de expressão forem similares entre condições em estudos de expressão. Muitos desses dados foram coletados de estudos de transcriptomas em repositórios de dados públicos. O processo de ponderação dependente de consulta (ou *Query-dependent weighting*, em inglês) empregue foi o método de ponderação selecionado automaticamente, método padrão do GeneMANIA em que se atribui pesos visando maximizar a conectividade entre todos os genes de entrada.

As redes de correlação de genes foram montadas utilizando o programa e plataforma Cytoscape para visualização de redes complexas e integração de dados. O *layout* direcionado por força com conectores ponderados (ou *Edge weighted force directed layout*, em inglês) para criar as redes com uma forma modelada pelos valores de χ^2. Esses valores foram previamente reescalonados em um intervalo de 0 a 1, para utilização do *layout*. As 500 conexões com os maiores valores de χ^2 foram elegidas para obtenção de uma rede de conectores robustos. Os conectores presentes nas redes de correlação e nas redes de co-expressão gênica foram destacados.

4.8 ELABORAÇÃO DE UM MODELOS DE CLASSIFICAÇÃO BASEADO EM VALORES DE CORRELAÇÃO GÊNICA

Para elaboração de um modelo de classificação capaz de valorizar as colinearidades, foram elaborados *heatmaps* de redes montadas utilizando os valores de expressão e a estrutura formada pelas duas redes de 500 conexões com os maiores valores de χ^2.

As redes de GSP e de GSN foram postas lado a lado em um espaço utilizando o programa e plataforma Cytoscape. Dessa construção foram obtidos os valores de coordenadas dos nós. As coordenadas e os valores de expressão dos conjuntos foram posteriormente transformados em valores de um intervalo 0 – 1. Foram geradas imagens topográficas de cada amostra utilizando o programa ViaComplex (CASTRO et al., 2009) utilizando as opções de execução de resolução, contraste, suavidade e zoom em 50%, as configurações padrão. As imagens tiveram que ser renomeadas para: "nome de amostra" + "_park" para amostras do grupo de DP idiopática ou "_ctrl" para indivíduos sadios, respectivamente.

As imagens foram levadas ao ambiente R utilizando a função *readImage* do pacote *EBImage* e redimensionadas para 100x100. As imagens foram transformadas em vetores de tamanho 10.000, onde cada elemento representa um valor de *pixel*. Para classificar cada imagem de acordo com seu grupo amostral, foram utilizadas funções envolvendo correspondência de padrões (*pattern matching*, em inglês) para identificação de *park* ou *ctrl* (para amostras do grupo de pacientes com a DP idiopática e indivíduos sadios, respectivamente) no nome da imagem.

Para as análises, os dados foram separados em dois grupos sendo que o GTr utilizou 80% das amostras para criação de modelos e o GTe utilizou os 20% restante de amostras para avaliação das predições, conforme efetuado para os modelos de análise de valores de expressão. Este modelo foi treinado com o algoritmo de *deep learning* (aprendizagem profunda em inglês) *MXNET* (CHEN et al., 2015).

Deep learning é um tipo de aprendizado de máquina que em vez de organizar os dados para serem executados mediante equações predefinidas, ele representa parâmetros básicos sobre os dados e treina o computador para aprender sozinho através do reconhecimento de padrões em camadas de processamento. Esses algoritmos aspiram modelar abstrações de alto nível de dados usando grafos (redes) com múltiplas camadas compostas de transformações lineares e não-lineares (PATTERSON; GIBSON, 2017).

O *deep learning* é uma das bases da inteligência artificial e ele tem aprimorado a capacidade de computadores de classificar, reconhecer,

descrever e detectar, ou simplesmente compreender, os dados. O *deep learning* é utilizado hoje em reconhecimento de fala como empregados pelo Xbox, Skype, Google Now e outros, em sistemas de recomendação como feitos pela Amazon e Netflix e em reconhecimento de imagens como empregados para legendação automática e em estudos de carros autônomos. Com base nisso, o *deep learning* foi indicado nesta classificação de imagens (NAJAFABADI et al., 2015).

Este modelo foi treinado com redes neurais do tipo *feed-forward*, redes cujas informações se movem em somente uma direção, para frente, dos nós de *input*, para os nós das camadas intermediárias e, finalmente, para os nós de saída (LIPPMANN, 1987).

Nesta modelagem foi utilizada a função *mx.model.FeedForward.create* do pacote *mxnet*. Foram empregues 30 iterações sobre o GTr para treinar os modelos. Os valores selecionados de hiperparâmetros nesta análise foram: "learning rate" de 0,05 e "momentum" de 0,9. O "learning rate" é o controle de ajuste de pesos, ou ponderações, na rede em relação as funções do gradiente de perda. Dessa forma, foi utilizado um valor baixo de forma a não haver perda de valores no *"*mínimo local*"*, ou seja, atingir os maiores valores de acurácia. O "momentum" é a "força" que permite que os valores cheguem ao "mínimo local" sem que fiquem trancados em valores subótimos.

5 RESULTADOS

5.1 BUSCA DE CONJUNTOS DE DADOS DE MICROARRANJOS DE DNA

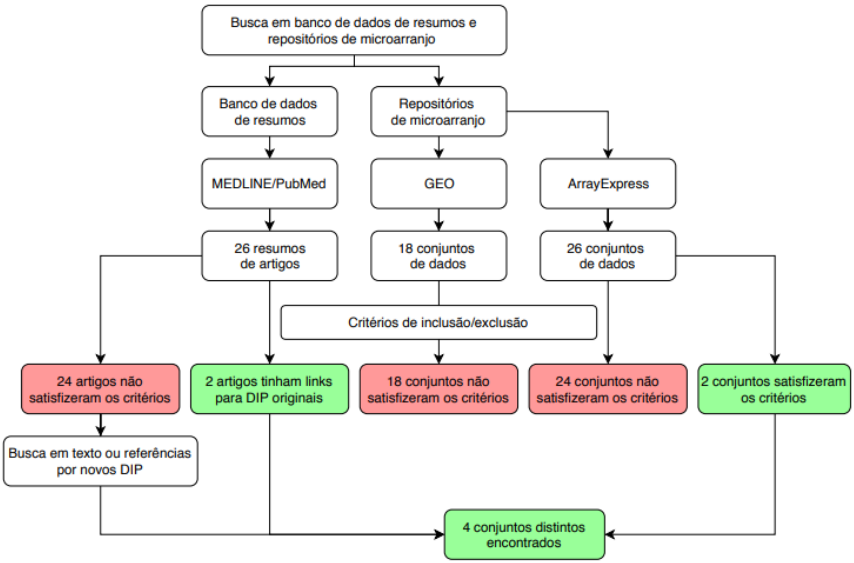
A identificação de conjuntos de dados de sangue periférico de pacientes com a DP idiopática e de indivíduos do grupo controle avaliados por microarranjo foi realizada por buscas em bancos de dados públicos. Foram procurados resumos de estudos científicos na base MEDLINE, utilizando o motor de busca PubMed e dados de microarranjo em repositórios recomendados pelo requerimento MIAME, o GEO e o ArrayExpress (**Figura 10**).

As buscas no PubMed resultaram em 26 resumos de artigos científicos publicados entre 2003 e 2017. Destes, somente 2 artigos, Locascio e colaboradores (2015) e Shamir e colaboradores (2017) continham no corpo do texto identificadores de dados de microarranjo depositados que satisfaziam os critérios de inclusão e exclusão, o GSE57475 e o GSE99039 (referência ao conjunto de dados depositado no GEO). As buscas em repositórios de dados de microarranjo resultaram em 18 conjuntos de dados no GEO e 26 no ArrayExpress. Destes, nenhum conjunto de dados do GEO satisfez os critérios de inclusão e exclusão, enquanto no ArrayExpress 2 satisfizeram, o E-GEOD-6613 e o E-GEOD-72267 (referência ao conjunto de dados depositado no ArrayExpress). Para padronização, foram utilizados os identificadores de conjunto de dados do GEO, para esses dois sendo GSE6613 e GSE72267, respectivamente. Estes estudos foram publicados por Scherzer e colaboradores (2007) e Calligaris e colaboradores (2015), respectivamente. As listas contendo os títulos dos artigos, os códigos de todos os conjuntos de dados e informações suplementares estão contidos nos **Apêndices de A a C**.

5.2 DESCRIÇÃO DE CONJUNTOS DE DADOS

Os 4 conjuntos de dados utilizados nesse estudo possuem no total 711 amostras, sendo 323 do grupo Controle e 388 do grupo de pacientes com a DP idiopática (**Quadro 1**). Os conjuntos de dados GSE6613, GSE57475, GSE72267 e GSE99039 representaram aproximadamente 10, 20, 8 e 61% do total de amostras, respectivamente. Quanto as porcentagens de fenótipos, os conjuntos de dados apresentaram aproximadamente 69, 65, 69 e 47% de amostras de pacientes com a DP idiopática.

Figura 10. Organograma do processo de busca por dados de microarranjo de DNA.



Em verde: conjuntos de dados aptos para essa análise identificados. Em vermelho: conjuntos de dados não aptos identificados. Conjuntos: conjuntos de dados. DIP: Dados individuais de pacientes. GEO: Gene Expression Omnibus.

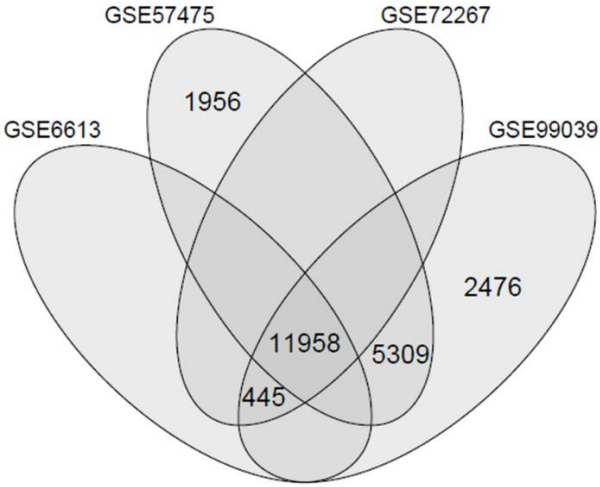
Os estudos foram todos analisados em plataformas de microarranjo diferentes. Os conjuntos de dados GSE6613, GSE72267 e GSE99039 foram analisados em plataformas da marca Affymetrix sendo todos *chips* de oligonucleotídeos de diferentes modelos, e o conjunto GSE57475 em plataforma da marca Illumina, sendo um microarranjo de cDNA. Neste trabalho foram analisadas 148.001 sondas. Destas, 114.832 (77.5%) possuíam anotação, segundo os pacotes de anotação utilizados. Para as sumarizações de informação foram utilizados os valores de *p*, obtidos de análises de expressão diferencial. Os conjuntos GSE6613 e GSE72267 apresentaram 12.418 identificadores de gene (Gene ID) únicos (**Apêndice D** e **F**), o GSE57475 19.223 (**Apêndice E**) e o GSE99039 20188 (**Apêndice G**), totalizando 64.217 Gene ID em todos os conjuntos de dados sujeitos às análises individuais. Destes, foram representados 22.144 Gene ID únicos distintos (**Apêndice H**), sendo que 17.712 Gene ID únicos estavam contidos em mais de um conjunto de dados, passíveis de meta-análise (**Figura 11** e **Apêndice I**).

Quadro 1. Estudos utilizados na meta-análise.

Informação do estudo		**Amostras**	
ID no GEO	**Citação**	**Controle**	**DP idiopática**
GSE6613	Scherzer et al. (2007)	22	50
GSE57475	Locascio et al. (2015)	49	93
GSE72267	Calligaris et al. (2015)	19	40
GSE99039	Shamir et al. (2017)	233	205
Total		**323**	**388**

	Informações do arranjo e de anotações			
ID no GEO	**Plataforma**	**Número de sondas**	**Sondas anotadas**	**Sondas após sumarização**
GSE6613	HG-U133A	22283	19812	12403
GSE57475	Illumina HumanHT-12 V3.0	48766	29426	19223
GSE72267	HG-U133A_2	22277	19812	12403
GSE99039	HG-U133_Plus_2	54675	45782	20188
Total		**148001**	**114832**	**64217**

Figura 11. Gene ID únicos distintos identificados em conjuntos de dados representados neste trabalho.



5.3 DADOS CLÍNICOS DE CONJUNTOS DE DADOS

Os dados clínicos dos pacientes dos trabalhos analisados foram obtidos nos artigos originais, em materiais suplementares e em arquivos de informações de fenótipos (phenoData) obtidos com o auxílio da função *getGEO* do pacote *GEOquery* em ambiente R. As informações contidas nesses arquivos, os metadados, eram bastante variados entre os estudos. As únicas informações, ou metainformações, contidas em todos os conjuntos de dados foram a idade, as porcentagens de homens e de mulheres para os dois grupos analisados e a severidade da doença, utilizando a escala Hoehn-Yahr das amostras com a DP (**Figura 12**). Os pacientes do conjunto de dados GSE72267 não faziam uso de medicações para tratamento da DP, pois suas amostras foram coletadas logo após o diagnóstico clínico. Quanto aos demais metadados, alguns conjuntos de dados apresentaram as variáveis hematológicas, porcentagens de caracteres clínicos como tremor em repouso, bradicinesia e rigidez, porcentagem de usuários de medicações além de outras medidas e escalas de severidade da doença. Estes metadados estão contidos nos **Apêndices de J a M**.

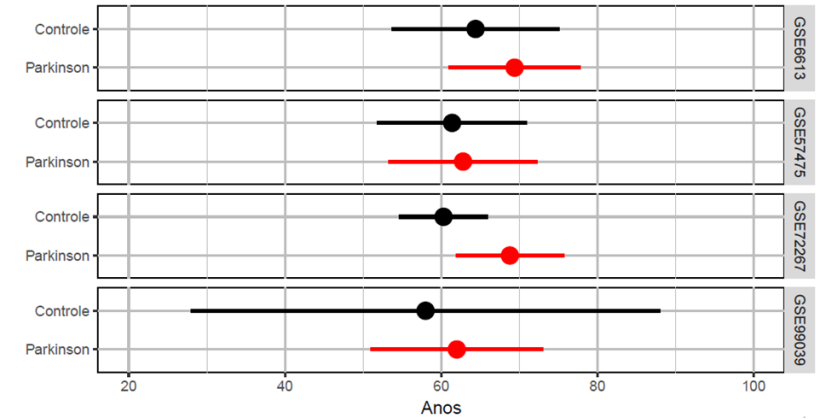
As idades de indivíduos do grupo controle do GSE6613 tiveram média de 64,4 e desvio padrão de 10,7, as do GSE57475 tiveram média de 61,4 e desvio padrão de 9,6, as do GSE72267 tiveram média de 60,3 e desvio padrão de 5,7 e as do GSE99039 tiveram média de 58,0 e desvio padrão de 30,0. As idades de pacientes com a DP idiopática do GSE6613 tiveram média de 69,4 e desvio padrão de 8,4, as do GSE57475 tiveram média de 62,8 e desvio padrão de 9,5, as do GSE72267 tiveram média de 68,8 e desvio padrão de 6,9 e as do GSE99039 tiveram média de 62,0 e desvio padrão de 11,0 (**Figura 12 A**).

As severidades da doença, utilizando a escala Hoehn-Yahr, de pacientes com a DP idiopática no GSE6613 tiveram média de 2,3 e desvio padrão de 0,7, as do GSE57475 tiveram média de 2,0 e desvio padrão de 0,5, as do GSE72267 tiveram média de 1,4 e desvio padrão de 0,5 e as do GSE99039 tiveram média de 1,8 e desvio padrão de 0,9 (**Figura 12 B**).

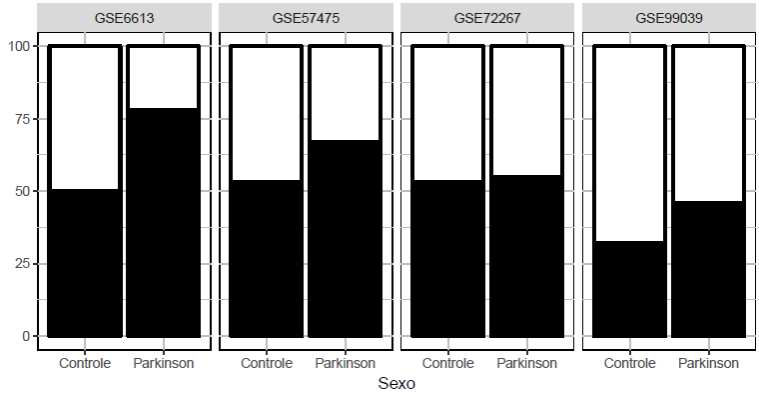
O GSE6613 apresentou 50% de homens nas amostras de indivíduos do grupo controle e 78% nas amostras com a DP idiopática, no GSE57475 as relações foram de 53 e 67%, no GSE72267 as relações foram de 53 e 55% e no GSE99039 as relações foram de 32 e 46% (**Figura 12 C**).

Figura 12. Idades, sexos e severidade da doença de amostras em diferentes trabalhos.

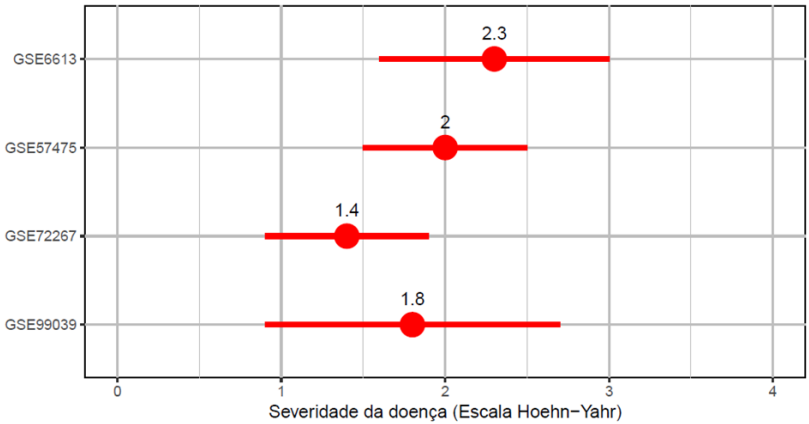
A



B



C



A: idades, em anos, representadas por média ± desvio padrão. Em preto: indivíduos do grupo controle. Em vermelho: pacientes com a DP idiopática. **B**: proporção de homens e mulheres em diferentes grupos. Em preto: homens. Em branco: mulheres. **C**: severidade da doença em pacientes com a DP idiopática, na escala Hoehn-Yahr, representadas por média ± desvio padrão. Valor apresentado sobre o círculo: média.

5.4 META-ANÁLISE E SELEÇÃO DE GENES

Os tamanhos de efeito individuais de cada gene para cada conjunto de dados foram calculados utilizando o estimador não-enviesado de Hedges' g (g*) e transformados em coeficiente de correlação produto-

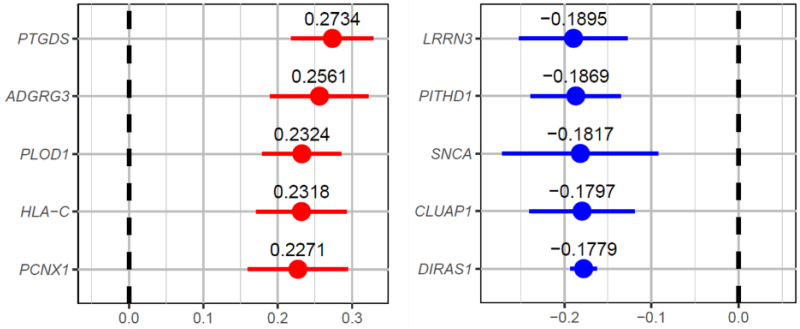
momento de Pearson (*r*). No total foram calculados os tamanhos de efeito para os 64.217 Gene ID únicos concentrados de todos os conjuntos de dados. Todos os dados estão contidos nos **Apêndices de N a Q**.

Para todos os 17.712 Gene ID únicos contidos em mais de um estudo, foram calculados os tamanhos de efeito médio ponderado ($r_{médio}$) (**Apêndice R**). Destes, na escala de Rosenthal e Rosnow (1984), 2 genes apresentaram um efeito "médio", estes detendo valores de tamanho de efeito positivos, e 2343 genes apresentaram um efeito "pequeno", 1667 com valores de tamanho de efeito positivos e 676 negativos. Os 15.367 genes remanescentes não possuíam tamanhos de efeito relevantes.

Os 100 genes com os maiores valores de tamanhos de efeito positivos e negativos foram selecionados (GSP = 100 Genes Selecionados através dos maiores valores Positivos e GSN = 100 Genes Selecionados através dos maiores valores Negativos) para as análises funcionais e para as seleções de preditores de algoritmos de classificação.

Os 100 GSP apresentaram um intervalo de $r_{médio}$ de 0,273 a 0,182, escore *z* de 232,4 a 3,6 e valores de *p* ajustados de 2,2E-19 a 2,0E-03. Os 100 GSN, por sua vez, apresentaram um intervalo de $r_{médio}$ de -0,189 a -0,136, escore *z* de 144,4 a 3,1 e valores de *p* ajustados de 2.2E-19 a 8.0E-03. Nenhum valor do intervalo de confiança dos GSP e dos GSN cruzaram o valor nulo de zero (**Figura 13**).

Figura 13 Tamanhos de efeito e intervalo de confiança dos GSP e GSN com os 5 maiores valores de diferença de classes.



Em vermelho: GSP. Em azul: GSN. Linha tracejada: valor nulo de zero. Valor sobre o círculo: tamanho de efeito.

Os símbolos, nomes completos e os tamanhos de efeito dos GSP e dos GSN estão listados no **Quadro 2**.

Quadro 2. Genes selecionados com os 100 maiores valores de tamanhos de efeito positivos e negativos.

GS com tamanhos de efeito positivos (GSP)			**GS com tamanhos de efeito negativos (GSN)**		
Gene Symbol	**Gene name**	$r_{médio}$	**Gene Symbol**	**Gene name**	$r_{médio}$
PTGDS	prostaglandin D2 synthase	0.273	*LRRN3*	leucine rich repeat neuronal 3	-0.190
ADGRG3	adhesion G protein-coupled receptor G3	0.256	*PITHD1*	PITH domain containing 1	-0.187
PLOD1	procollagen-lysine,2-oxoglutarate 5-dioxygenase 1	0.232	*SNCA*	synuclein alpha	-0.182
HLA-C	major histocompatibility complex, class I, C	0.232	*CLUAP1*	clusterin associated protein 1	-0.180
PCNX1	pecanex 1	0.227	*DIRAS1*	DIRAS family GTPase 1	-0.178
LILRB3	leukocyte immunoglobulin like receptor B3	0.227	*EBF1*	early B cell factor 1	-0.172
KIR2DL3	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 3	0.223	*CCDC50*	coiled-coil domain containing 50	-0.171
KIAA0319L	KIAA0319 like	0.222	*BCL2*	BCL2, apoptosis regulator	-0.170
PILRA	paired immunoglobin like type 2 receptor alpha	0.221	*XIST*	X inactive specific transcript	-0.170
STK40	serine/threonine kinase 40	0.216	*LINC00242*	long intergenic non-protein coding RNA 242	-0.169
KIR2DL4	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 4	0.215	*CYB5A*	cytochrome b5 type A	-0.169
CA4	carbonic anhydrase 4	0.212	*TCF3*	transcription factor 3	-0.169
MMP9	matrix metallopeptidase 9	0.212	*LTBP3*	latent transforming growth factor beta binding protein 3	-0.168
NLRP12	NLR family pyrin domain containing 12	0.210	*RBM38*	RNA binding motif protein 38	-0.165
HLA-E	major histocompatibility complex, class I, E	0.210	*NELL2*	neural EGFL like 2	-0.165
P2RX1	purinergic receptor P2X 1	0.210	*FAM102A*	family with sequence similarity 102 member A	-0.165

SELPLG	selectin P ligand	0.209	*UBE3D*	ubiquitin protein ligase E3D	-0.165
RAB37	RAB37, member RAS oncogene family	0.209	*HBD*	hemoglobin subunit delta	-0.164
ABTB1	ankyrin repeat and BTB domain containing 1	0.208	*ENOSF1*	enolase superfamily member 1	-0.163
LRRC25	leucine rich repeat containing 25	0.207	*GYPB*	glycophorin B (MNS blood group)	-0.163
PLAUR	plasminogen activator, urokinase receptor	0.207	*GLRX5*	glutaredoxin 5	-0.163
PDLIM7	PDZ and LIM domain 7	0.205	*MCF2L*	MCF.2 cell line derived transforming sequence like	-0.162
ITPRIP	inositol 1,4,5-trisphosphate receptor interacting protein	0.204	*ACTL10*	actin like 10	-0.162
TNFSF14	TNF superfamily member 14	0.204	*DKK3*	dickkopf WNT signaling pathway inhibitor 3	-0.162
CXCL16	C-X-C motif chemokine ligand 16	0.203	*ORC5*	origin recognition complex subunit 5	-0.162
SLC44A2	solute carrier family 44 member 2	0.202	*FCRL2*	Fc receptor like 2	-0.161
TFE3	transcription factor binding to IGHM enhancer 3	0.202	*TOB1-AS1*	TOB1 antisense RNA 1	-0.161
TOP3A	DNA topoisomerase III alpha	0.202	*SYNPO2*	synaptopodin 2	-0.160
ATG2A	autophagy related 2A	0.202	*RGCC*	regulator of cell cycle	-0.159
SRA1	steroid receptor RNA activator 1	0.202	*NUDT4*	nudix hydrolase 4	-0.159
GAB2	GRB2 associated binding protein 2	0.201	*C21orf2*	chromosome 21 open reading frame 2	-0.159
KIR2DL1	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 1	0.200	*LTBP4*	latent transforming growth factor beta binding protein 4	-0.158
LILRA2	leukocyte immunoglobulin like receptor A2	0.200	*DAAM1*	dishevelled associated activator of morphogenesis 1	-0.157
TMX4	thioredoxin related transmembrane protein 4	0.199	*ATP8B2*	ATPase phospholipid transporting 8B2	-0.156
MOB3A	MOB kinase activator 3A	0.199	*LINC01405*	long intergenic non-protein coding RNA 1405	-0.156
NLRX1	NLR family member X1	0.199	*PSMF1*	proteasome inhibitor subunit 1	-0.155

PSD4	pleckstrin and Sec7 domain containing 4	0.199	*LRRC27*	leucine rich repeat containing 27	-0.155
DAPK2	death associated protein kinase 2	0.198	*MARCH8*	membrane associated ring-CH-type finger 8	-0.155
PHC2	polyhomeotic homolog 2	0.198	*SMC1A*	structural maintenance of chromosomes 1A	-0.155
HSH2D	hematopoietic SH2 domain containing	0.197	*MATN2*	matrilin 2	-0.154
ZYX	zyxin	0.197	*DCAF12*	DDB1 and CUL4 associated factor 12	-0.153
DLGAP4	DLG associated protein 4	0.196	*TUBB2A*	tubulin beta 2A class IIa	-0.152
IGF2R	insulin like growth factor 2 receptor	0.196	*ALAS2*	5'-aminolevulinate synthase 2	-0.152
ARHGAP26	Rho GTPase activating protein 26	0.196	*NF2*	neurofibromin 2	-0.152
GBAP1	glucosylceramidase beta pseudogene 1	0.196	*CALR3*	calreticulin 3	-0.152
CSF3R	colony stimulating factor 3 receptor	0.194	*TPP2*	tripeptidyl peptidase 2	-0.152
GPX3	glutathione peroxidase 3	0.194	*ISCA1*	iron-sulfur cluster assembly 1	-0.151
RGS2	regulator of G protein signaling 2	0.193	*FGFR1OP2*	FGFR1 oncogene partner 2	-0.150
IFITM2	interferon induced transmembrane protein 2	0.193	*SKP2*	S-phase kinase associated protein 2	-0.150
APOBR	apolipoprotein B receptor	0.193	*RHOH*	ras homolog family member H	-0.150
HLA-A	major histocompatibility complex, class I, A	0.193	*COX17*	cytochrome c oxidase copper chaperone COX17	-0.150
EFHD2	EF-hand domain family member D2	0.192	*GPT*	glutamic--pyruvic transaminase	-0.149
UBN1	ubinuclein 1	0.192	*KRT16*	keratin 16	-0.149
VSIR	V-set immunoregulatory receptor	0.192	*IL23A*	interleukin 23 subunit alpha	-0.148
CXCR1	C-X-C motif chemokine receptor 1	0.192	*TTC25*	tetratricopeptide repeat domain 25	-0.148
CFLAR	CASP8 and FADD like apoptosis regulator	0.192	*LRRC34*	leucine rich repeat containing 34	-0.148
SYTL3	synaptotagmin like 3	0.192	*OLMALINC*	oligodendrocyte maturation-associated long intergenic non-coding RNA	-0.148

ACVR1B	activin A receptor type 1B	0.192	*COPS8*	COP9 signalosome subunit 8	-0.147
PRR14	proline rich 14	0.192	*EPB42*	erythrocyte membrane protein band 4.2	-0.147
FCAR	Fc fragment of IgA receptor	0.191	*ATM*	ATM serine/threonine kinase	-0.147
FTH1P5	ferritin heavy chain 1 pseudogene 5	0.190	*SLC4A1*	solute carrier family 4 member 1 (Diego blood group)	-0.147
SLC11A1	solute carrier family 11 member 1	0.190	*INHBE*	inhibin subunit beta E	-0.146
MTG2	mitochondrial ribosome associated GTPase 2	0.189	*PDZRN3*	PDZ domain containing ring finger 3	-0.146
RAB11FIP1	RAB11 family interacting protein 1	0.189	*FMR1NB*	FMR1 neighbor	-0.146
CEBPD	CCAAT enhancer binding protein delta	0.189	*RPS28*	ribosomal protein S28	-0.146
C15orf39	chromosome 15 open reading frame 39	0.188	*SELENBP1*	selenium binding protein 1	-0.145
TMEM120A	transmembrane protein 120A	0.188	*MKRN1*	makorin ring finger protein 1	-0.145
RARA	retinoic acid receptor alpha	0.188	*UROD*	uroporphyrinogen decarboxylase	-0.145
SSH2	slingshot protein phosphatase 2	0.188	*ZNF254*	zinc finger protein 254	-0.144
CST7	cystatin F	0.188	*RPLP1*	ribosomal protein lateral stalk subunit P1	-0.144
OSCAR	osteoclast associated, immunoglobulin-like receptor	0.188	*TNS1*	tensin 1	-0.144
VNN3	vanin 3	0.188	*MTMR2*	myotubularin related protein 2	-0.144
TYK2	tyrosine kinase 2	0.187	*DCAF12L2*	DDB1 and CUL4 associated factor 12 like 2	-0.143
CEACAM4	carcinoembryonic antigen related cell adhesion molecule 4	0.187	*ENTPD6*	ectonucleoside triphosphate diphosphohydrolase 6 (putative)	-0.143
METRNL	meteorin like, glial cell differentiation regulator	0.187	*NBEAP1*	neurobeachin pseudogene 1	-0.143
PADI4	peptidyl arginine deiminase 4	0.187	*PCDH15*	protocadherin related 15	-0.142
HELZ2	helicase with zinc finger 2	0.186	*RPS2*	ribosomal protein S2	-0.142

TSEN34	tRNA splicing endonuclease subunit 34	0.186	*EIF3B*	eukaryotic translation initiation factor 3 subunit B	-0.141
NINJ1	ninjurin 1	0.186	*PASK*	PAS domain containing serine/threonine kinase	-0.141
CTBP2	C-terminal binding protein 2	0.186	*KRT1*	keratin 1	-0.140
LMO4	LIM domain only 4	0.186	*TLE2*	transducin like enhancer of split 2	-0.140
ABCG1	ATP binding cassette subfamily G member 1	0.186	*RAB2B*	RAB2B, member RAS oncogene family	-0.140
ABCA1	ATP binding cassette subfamily A member 1	0.185	*FAM117A*	family with sequence similarity 117 member A	-0.140
ARAP1	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 1	0.185	*MRPS5*	mitochondrial ribosomal protein S5	-0.140
CMIP	c-Maf inducing protein	0.185	*GART*	phosphoribosylglycinamide formyltransferase, phosphoribosylglycinamide synthetase, phosphoribosylaminoimidazole synthetase	-0.139
CKAP4	cytoskeleton associated protein 4	0.185	*KRAS*	KRAS proto-oncogene, GTPase	-0.139
SERPINA1	serpin family A member 1	0.185	*PLCG1*	phospholipase C gamma 1	-0.139
TTC38	tetratricopeptide repeat domain 38	0.185	*ZNF555*	zinc finger protein 555	-0.139
TNFRSF9	TNF receptor superfamily member 9	0.185	*PPIE*	peptidylprolyl isomerase E	-0.139
RNF24	ring finger protein 24	0.185	*PKLR*	pyruvate kinase L/R	-0.139
TLE3	transducin like enhancer of split 3	0.184	*FAM207A*	family with sequence similarity 207 member A	-0.139
HSPA6	heat shock protein family A (Hsp70) member 6	0.184	*NSG1*	neuronal vesicle trafficking associated 1	-0.139
ALOX5AP	arachidonate 5-lipoxygenase activating protein	0.184	*IL21R*	interleukin 21 receptor	-0.138

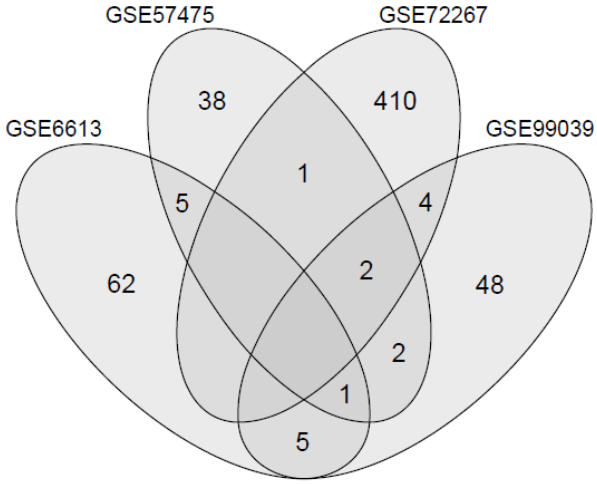
PLEKHG3	pleckstrin homology and RhoGEF domain containing G3	0.184	*TULP4*	tubby like protein 4	-0.138
KIR3DL1	killer cell immunoglobulin like receptor, three Ig domains and long cytoplasmic tail 1	0.183	*DNPH1*	2'-deoxynucleoside 5'-phosphate N-hydrolase 1	-0.138
ST6GALNAC2	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 2	0.183	*ZNF678*	zinc finger protein 678	-0.138
MTMR14	myotubularin related protein 14	0.183	*SNRPA1*	small nuclear ribonucleoprotein polypeptide A'	-0.137
G6PD	glucose-6-phosphate dehydrogenase	0.183	*SIPA1L3*	signal induced proliferation associated 1 like 3	-0.137
SLC15A4	solute carrier family 15 member 4	0.182	*BANK1*	B cell scaffold protein with ankyrin repeats 1	-0.137
ARHGEF40	Rho guanine nucleotide exchange factor 40	0.182	*MRPL30*	mitochondrial ribosomal protein L30	-0.137

5.5 REANÁLISES DE EXPRESSÕES DIFERENCIAIS DE CADA CONJUNTO DE DADOS

A maior parte de experimentos de microarranjo visa identificar os GDE, os genes diferencialmente expressos entre condições experimentais, porém muitas vezes utilizando metodologias extremamente diferentes de pré-processamento, processamento e limiares de apuração gênica. Com base nisso, foram realizadas reanálises de expressão diferencial de cada conjunto utilizando os mesmos limiares para obtenção de GDE, *p*-valor $\leq 0{,}05$ para o teste-*t* e valor absoluto positivo de *fold change* $\geq 1{,}2$.

As reanálises individuais identificaram 73 GDE no conjunto de dados GSE6613 (16 supra- e 57 infra-regulados), 49 no GSE57475 (21 supra- e 28 infra-regulados), 417 no GSE72267 (171 supra- e 246 infra-regulados) e 62 no GSE99039 (51 supra- e 11 infra-regulados) e apenas 20 desses (3,4%) estavam presentes em dois ou mais conjuntos de dados. Os totais de GDE de cada conjunto de dados individual e os GDE que estavam presentes em mais de um conjunto estão exibidos na **Figura 14**.

Figura 14. GDE resultados das reanálises de expressão diferencial individuais de cada conjunto de dados.



Os genes *EPB42*, *GYPB*, *HBD*, *MARCH8* e *PIP4K2A* estavam presentes na intersecção de GSE6613 e GSE57475, os genes *KDM5D*, *MMP9*, *RPS4Y1*, *TRAK2* e *XIST* na intersecção de GSE6613 e

GSE990039, o gene *HLA-DQA1* na intersecção de GSE57475 e GSE72267, os genes *FAM46C* e *PTGDS* na intersecção de GSE57475 e GSE99039, os genes *ARG1*, *BMX*, *ENC1* e *LRRN3* na intersecção de GSE72267 e GSE99039, o gene *SNCA* na intersecção de GSE6613, GSE57475 e GSE99039 e os genes *ADGRG3* e *LILRA5* na intersecção de GSE57475, GSE72267 e GSE99039.

Dentre os GSP e GSN, os Gene ID selecionados na análise de tamanho de efeito agrupado, positivos e negativos, 151 estavam representados nas plataformas dos conjuntos de dados GSE6613 e GSE72267, 195 na plataforma do GSE57475 e 200 na plataforma do GSE99039. Entretanto, somente 17 Gene ID selecionados estavam representados entre os GDE do GSE6613, 16 entre os GDE do GSE57475, 18 entre os GDE do GSE72267 e 17 entre os GDE do GSE99039. Dessa forma, os genes selecionados configuraram aproximadamente 23%, 33%, 4% e 27% dos GDE em reanálises individuais/independentes de conjuntos de dados utilizados neste trabalho (**Figura 15**).

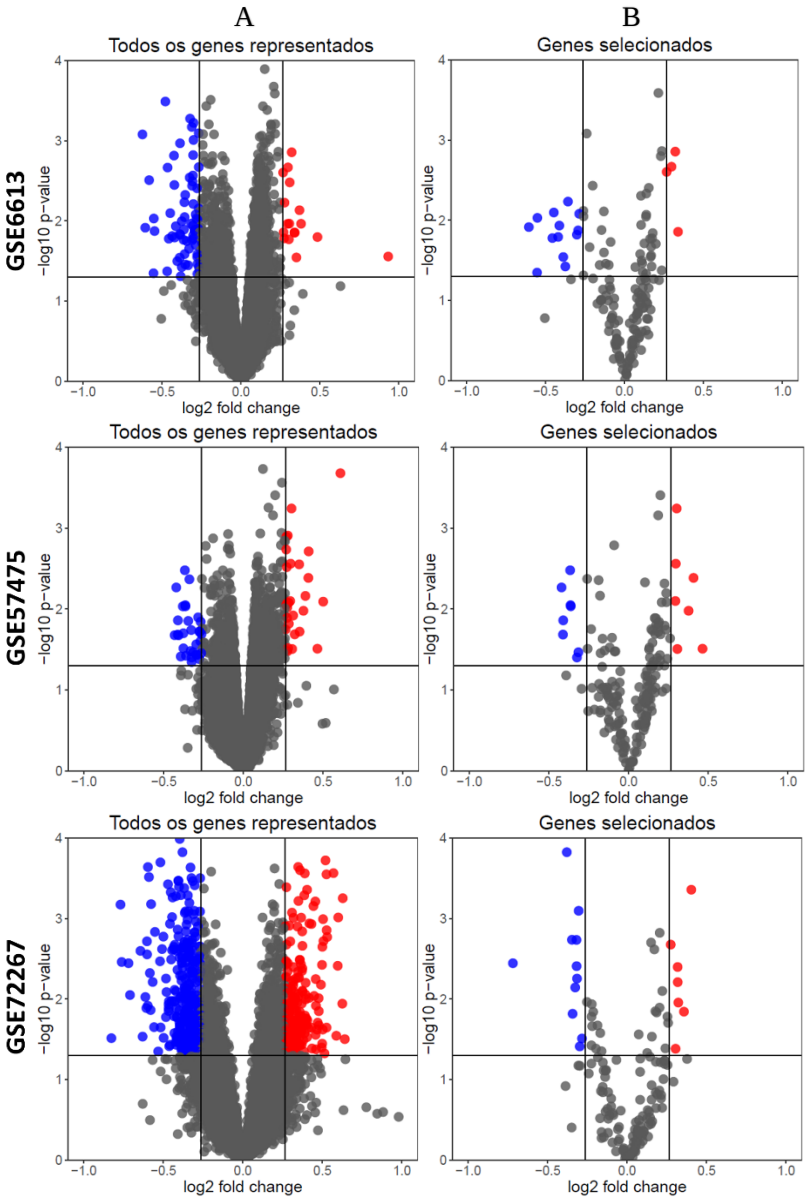
Entre os genes selecionados apresentados como GDE, o gene *SNCA* estava presente na intersecção de GSE6613, GSE57475 e GSE99039, os genes *EPB42*, *GYPB*, *HBD* e *MARCH8* na intersecção de GSE6613 e GSE57475, os genes *MMP9* e *XIST* na intersecção de GSE6613 e GSE99039, o gene *PTGDS* na intersecção de GSE57475 e GSE99039 e os genes *LRRN3* e *ADGRG3* de GSE72267 e GSE99039.

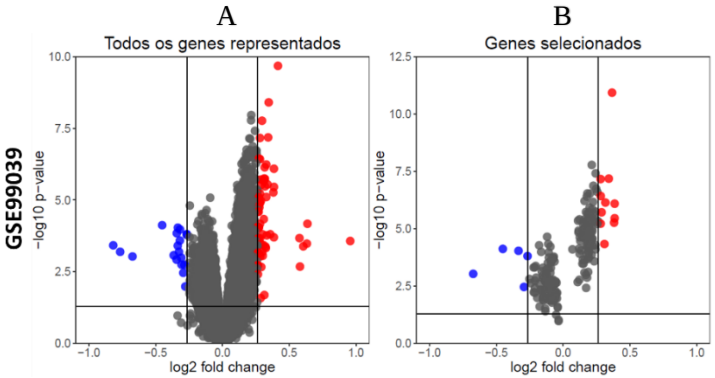
As amostras de cada conjunto foram reunidas utilizando os valores de expressão de GSP e de GSN para realização de agrupamentos hierárquicos utilizando a correlação de Pearson e a medida de distância de Ward (**Figura 16**).

A primeira partição foi avaliada por motivação de entender se as amostras estariam separadas em dois grupos conforme o fenótipo de pacientes com a DP idiopática e indivíduos do grupo controle. A acurácia e a taxa de não-informação do conjunto GSE6613 foram ambas de 0,69., enquanto do conjunto GSE57475 foi 0,55 e 0,65, do conjunto GSE72267 foi 0,52 e 0,67 e do conjunto GSE99039 foi 0,54 e 0,53, respectivamente.

Sendo assim, os genes selecionados na meta-análise foram pouco identificados (menos de 33%) e não são capazes de separar as amostras de pacientes com a DP idiopática e de indíviduos sadios utilizando as reanálises de expressão diferencial.

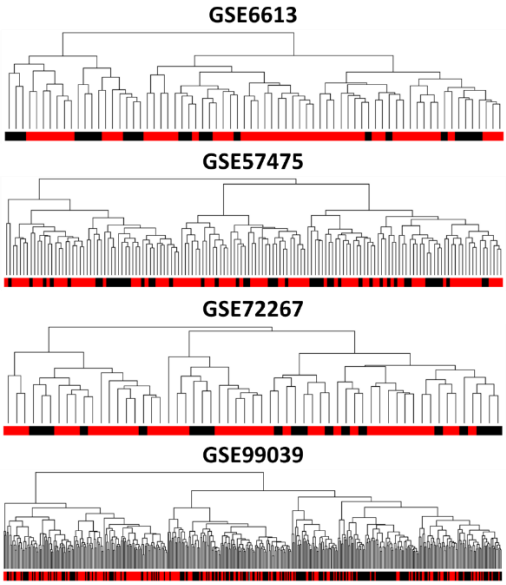
Figura 15. *Volcano plots* de reanálises de expressão diferencial individuais e da dispersão de GSP edGSN em cada estudo.





Coluna A: gráficos de *volcano* contendo todos os genes representados contidos em cada conjunto de dados. Coluna B: gráficos de *volcano* contendo os GSP e os GSN contidos em cada conjunto de dados. Em vermelho: genes supra-regulados obtidos com limiares de *fold change* absoluto > 1,2 e valor de $p < 0,05$. Em azul: genes infra-regulados obtidos com limiares de *fold change* absoluto < -1,2 e valor de $p < 0,05$.

Figura 16. Cladograma de amostras com base nos valores de expressão dos genes selecionados.



Nós terminais pretos: amostras de indivíduos do grupo controle. Nós terminais vermelhos: amostras de pacientes com a DP idiopática.

5.6 ANÁLISE FUNCIONAL DE GENES SELECIONADOS

As caracterizações funcionais ou significados biológicos foram realizados utilizando as funções *goana* e *kegga* do pacote *limma* e o banco de dados DAVID. Em ambas as metodologias foram procurados os 10 termos de PB e as 10 vias de interação molecular enriquecidas em testes hipergeométricos com (1) os menores valores de *p*, não ajustados para múltiplos testes, em funções de *limma*, e (2) os menores valores de *p* exato de Fisher modificado em buscas no banco de dados DAVID.

Os 10 termos enriquecidos na análise de enriquecimento funcional dos GSP utilizando o método goana/limma evidenciaram a participação dos genes obtidos na meta-análise em ontologias relacionadas a resposta imune, imunidade mediada por leucócitos, ativação leucocitária e degranulação leucocitária. Utilizando o método de busca no banco de dados DAVID, os termos evidenciaram ontologias relacionadas a resposta imune, processamento e apresentação de antígeno peptídico exógeno via MHC (*Major Histocompatibility Complex* ou Complexo principal de histocompatibilidade, em inglês) de classe I, transdução de sinal e processos virais. Os 10 primeiros termos de ambas as metodologias nos GSP estão apresentados no **Quadro 3**. Todos os termos encontrados para os GSP estão apresentados nos **Apêndices S e T**.

Quadro 3. Termos de ontologias enriquecidos com os GSP.

GSP - GOANA

GO ID	Termo	valor de *p*
GO:0006955	immune response	1.41E-14
GO:0002275	myeloid cell activation involved in immune response	1.05E-10
GO:0002283	neutrophil activation involved in immune response	1.88E-10
GO:0042119	neutrophil activation	2.50E-10
GO:0036230	granulocyte activation	3.02E-10
GO:0002366	leukocyte activation involved in immune response	6.85E-10
GO:0043299	leukocyte degranulation	6.99E-10
GO:0002263	cell activation involved in immune response	7.38E-10
GO:0002444	myeloid leukocyte mediated immunity	1.02E-09
GO:0002274	myeloid leukocyte activation	1.14E-09

GSP - DAVID

GO ID	Termo	valor de *p*
GO:0006955	immune response	1.09E-05

GO:0050776	regulation of immune response	3.83E-05
GO:0007165	signal transduction	2.73E-04
GO:0060337	type I interferon signaling pathway	3.37E-04
GO:0002480	antigen processing and presentation of exogenous peptide antigen via MHC class I, TAP-independent	9.33E-04
GO:0016032	viral process	4.58E-03
GO:0034341	response to interferon-gamma	6.80E-03
GO:0045742	positive regulation of epidermal growth factor receptor signaling pathway	6.80E-03

Os 10 termos enriquecidos na análise de enriquecimento funcional dos GSN utilizando o método goana/limma evidenciaram a participação dos genes obtidos na meta-análise em ontologias menos associadas entre si, porém relacionadas a ubiquitinação proteica, processos biossintéticos de pigmentos, iniciação da tradução na expressão gênica, processo metabólico do álcool e regulação positiva de atividade catalítica. Utilizando o método de busca no banco de dados DAVID, 2 termos passaram pela significância estatística e evidenciaram ontologias relacionadas a ubiquitinação proteica e organização do citoesqueleto. Os 10 primeiros termos da metodologia goana/limma e os 2 termos da metodologia de busca no DAVID nos GSN estão apresentados na **Quadro 4**. Todos os termos encontrados para os GSN estão apresentados nos **Apêndices U e V**.

Quadro 4. Termos de ontologias enriquecidos com os GSN.

GSN - GOANA		
GO ID	**Termo**	**valor de *p***
GO:0042440	pigment metabolic process	3.37E-04
GO:0043647	inositol phosphate metabolic process	3.37E-04
GO:0000209	protein polyubiquitination	6.51E-04
GO:0051353	positive regulation of oxidoreductase activity	1.52E-03
GO:0016567	protein ubiquitination	2.12E-03
GO:0046148	pigment biosynthetic process	2.16E-03
GO:0019751	polyol metabolic process	2.54E-03
GO:0006413	translational initiation	2.58E-03
GO:0032233	positive regulation of actin filament bundle assembly	2.81E-03
GO:0034655	nucleobase-containing compound catabolic process	2.93E-03

GSN - DAVID

GO ID	Termo	valor de *p*
GO:0000209	protein polyubiquitination	2.18E-03
GO:0007010	cytoskeleton organization	8.32E-03

As 10 vias de interação molecular enriquecidas na análise de enriquecimento funcional dos GSN utilizando o método kegga/limma evidenciaram a participação dos genes obtidos na meta-análise em ontologias relacionadas ao próprio sistema imune, resposta imune a aloenxertos e a vírus, endocitose, diabetes mellitus do tipo I, e outros. Utilizando o método de busca no banco de dados DAVID, apenas 2 vias passaram pela significância estatística e evidenciaram interações moleculares relacionadas ao processamento e exibições de antígenos e a endocitose. As 10 primeiras vias da metodologia kegga/limma e as 2 vias da metodologia de busca no DAVID nos GSN estão apresentados no **Quadro 5**. Todos os termos encontrados para os GSN estão apresentados nos **Apêndices W e X**.

Quadro 5. Vias de interação molecular enriquecidas com os GSP.

GSP - KEGGA

KEGG ID	Via	valor de *p*
hsa04612	Antigen processing and presentation	3.74E-09
hsa05332	Graft-versus-host disease	4.13E-08
hsa04144	Endocytosis	2.28E-06
hsa04650	Natural killer cell mediated cytotoxicity	3.31E-06
hsa04218	Cellular senescence	1.37E-04
hsa05168	Herpes simplex infection	2.85E-04
hsa04380	Osteoclast differentiation	4.31E-04
hsa05330	Allograft rejection	7.37E-04
hsa04940	Type I diabetes mellitus	1.16E-03
hsa05320	Autoimmune thyroid disease	2.04E-03

GSP - DAVID KEGG

KEGG ID	Via	valor de *p*
hsa04612	Antigen processing and presentation	1.01E-06
hsa04144	Endocytosis	4.66E-04

Utilizando o método kegga/limma apenas 4 vias passaram pela significância estatística e evidenciaram interações relacionadas a

ribossomos, ciclo celular, resistência a inibidores de tirosina-quinases do receptor do fator de crecimento epidérmico (EGFR ou *Epidermal Growth Factor Receptor*, em inglês). Utilizando o método de busca no banco de dados DAVID, apenas 1 via passou pela significância estatística e evidenciou interações relacionadas a ribossomos. As 4 vias da metodologia kegga/limma e a via da metodologia de busca no DAVID nos GSN estão apresentados na **Quadro 6**. Todos os termos encontrados para os GSN estão apresentados nos **Apêndices Y e Z**.

Quadro 6. Vias de interação molecular enriquecidas com os GSN.

GSN - KEGGA

KEGG ID	Via	valor de *p*
hsa03010	Ribosome	4.79E-04
hsa04110	Cell cycle	3.22E-03
hsa01521	EGFR tyrosine kinase inhibitor resistance	7.01E-03
hsa04064	NF-kappa B signaling pathway	9.32E-03

GSN - DAVID KEGG

KEGG ID	Via	valor de *p*
hsa03010	Ribosome	5.19E-03

5.7 PREDITORES PARA ASSINATURA GÊNICA

Para criação de modelos e predição da capacidade de classificação, primeiramente os dados de todos os conjuntos precisaram ser preparados para as análises. Os dados de expressão de cada conjunto foram reescalonados em escore *z* para que cada gene contribuísse igualmente em cada análise. Os dados transformados foram agrupados em uma megamatriz com todas as amostras da meta-análise.

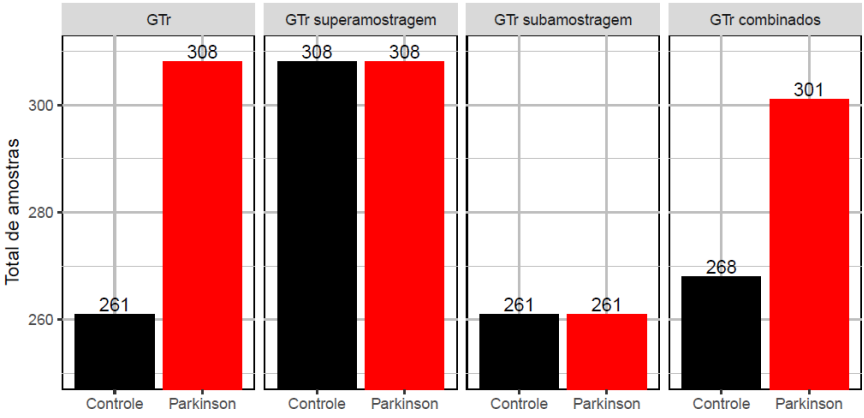
A megamatriz conta com 711 observações, ou amostras, sendo 388 de pacientes com a DP idiopática, representando 54,6% do total de amostras e 323 de indivíduos do grupo controle, representando os 45,4% restante.

Estas foram divididas em GTr, utilizado para concepção e ajuste de modelos e GTe, utilizado para estimação do desempenho de modelos. O GTr foi construído com 80% do total de amostras, escolhidas por amostragem. O grupo continha 303 (53,2%) amostras de pacientes com Parkinson idiopático e 266 (46,8%) amostras do grupo controle. O grupo

era composto por 54 amostras do conjunto de dados GSE6613 (75% do total do conjunto), 124 do GSE57475 (87,3%), 46 do GSE72267 (77,9%) e 345 do GSE99039 (78,7%). Quanto ao fenótipo das amostras de cada conjunto de dados no GTr, 38 amostras (70,3%) do GSE6613 eram de pacientes com a DP idiopática e 16 eram de indivíduos sadios, no GSE57475 a relação foi de 78(62,9%)/46, no GSE72267 foi de 29(63%)/17 e no GSE99039 foi de 158(45,8%)/187.

Os métodos de criação de amostras sintéticas do pacote ROSE foram empregados para minimizar possíveis dificuldades derivadas do desbalanceamento de dados (total de amostras) entre os grupos *Parkinson* e *Controle* no GTr base. Os métodos criaram 3 GTr: um GTr com superamostragem do grupo amostral com menos amostras, o GTr superamostragem, um GTr com subamostragem do maior grupo amostral, o GTr subamostragem e um GTr que busca equilíbrio entre os grupos supra e infrarrepresentados combinando os métodos de super e subamostragem, o GTr combinados. O total de amostras de cada um dos GTr está representado na **Figura 17**, onde se identifica o equilíbrio acarretado da super e da subamostragem e a suavização das diferenças entre os grupos no GTr de métodos combinados. Todos os algoritmos, com todos os hiperparâmetros foram testados nos 4 GTr (O GTr base e os 3 criados pelos métodos de ROSE).

Figura 17. Total de amostras em cada GTr separado por fenótipo.

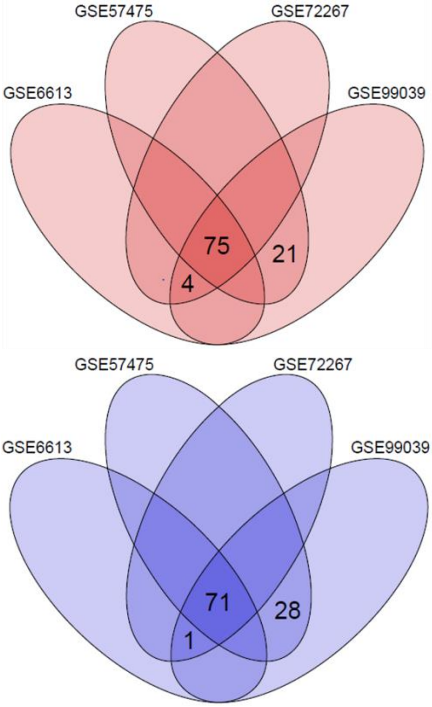


Os preditores utilizados em algoritmos de aprendizado de máquina devem estar presentes em todas as amostras para via de comparação. Dos

100 GSP e GSN, 75 GSP e 71 GSN estavam presentes em todos os conjuntos de dados, ou seja, em todas as plataformas examinadas (**Figura 18**), resultando em 146 possíveis variáveis, ou preditores, para as análises dentre os genes selecionados.

A primeira forma de selecionar os preditores foi com a eliminação de informações redundantes utilizando os métodos de recognição e exclusão de colinearidade. Essa exclusão eliminou 78 genes colineares, resultando em 68 genes mantidos nas análises. Na **Figura 19** estão demonstradas as matrizes de correlações significativas antes (A) e depois (B) da filtragem de valores maiores que 0,75 de correlação de Pearson. Os nomes dos genes de linhas/colunas foram omitidos. Percebe-se na segunda matriz a ausência de pontos de cores quentes superiores ao laranja-claro (tom da barra na posição entre 0,25 – 0,5), excetuando-se os pontos de cruzamento próprio na linha diagonal decrescente, e cores frias inferiores ao azul-claro (tom da barra na posição entre -0,25 e 0,5).

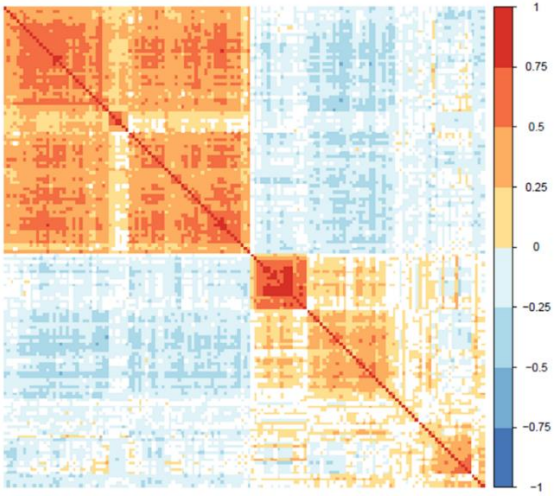
Figura 18. Total de genes selecionados contidos em cada conjunto de dados.



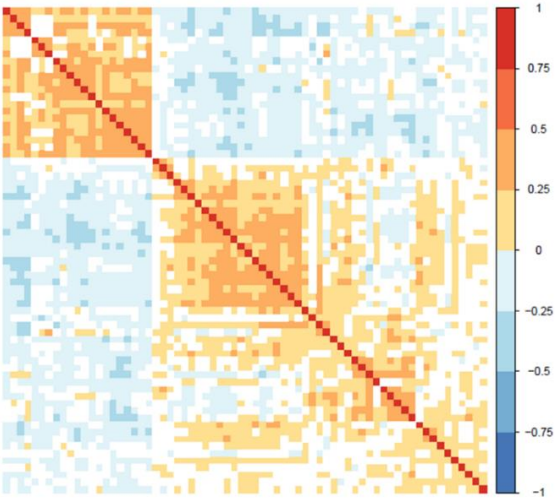
A: diagrama de Venn contendo os GSP. **B**: diagrama de Venn contendo os GSN.

Figura 19. Correlação gênica de todos os genes selecionados presentes em mais de um conjunto de dados e destes após a eliminação de informações redundantes.

A



B



A: Matriz de correlação de genes selecionados presentes em mais de um conjunto de dados. **B**: Matriz de correlação de genes selecionados presentes em mais de um conjunto de dados após eliminação de informações redundantes. As barras a direita relacionam as cores e os valores de correlação entre os pares de genes.

A segunda ferramenta de seleção de preditores operada foi o algoritmo de eliminação de preditores recursivo, utilizando todas as possibilidades de elementos retidos após o ranqueamento de importância. O valor ótimo do número de preditores escolhido pelo algoritmo foi de 58 alcançando acurácia de 0,71 utilizando o método de reamostragem de validação cruzada, e assim os 10 preditores de menor relevância no procedimento de classificação foram eliminados.

Dentre os 58 preditores utilizados nos algoritmos de classificação como assinatura gênica 19 eram GSP e 39 eram GSN (**Quadro 7**). Análises de enriquecimento funcional mostraram que 5 dos GSP da assinatura eram relacionados ao termo de "resposta imune" (*CST7*, *FCAR*, *HLA-A*, *KIR2DL1* e *TNFRSF9*), 4 eram relacionados ao termo de "processo apoptótico" e "regulação do processo apoptótico" (*CFLAR*, *DAPK2*, *P2RX1* e *TNFRSF9*) e 2 participam da via de interação molecular do "câncer de bexiga" (*DAPK2* e *MMP9*). Não houve termo ou via de interação molecular enriquecida com os GSN da assinatura gênica (dados não apresentados).

Quadro 7. Assinatura gênica.

GSP na assinatura gênica

Gene Symbol	Gene name
ABCA1	ATP binding cassette subfamily A member 1
CFLAR	CASP8 and FADD like apoptosis regulator
CST7	cystatin F
CTBP2	C-terminal binding protein 2
DAPK2	death associated protein kinase 2
DLGAP4	DLG associated protein 4
FCAR	Fc fragment of IgA receptor
GPX3	glutathione peroxidase 3
HLA-A	major histocompatibility complex, class I, A
KIR2DL1	killer cell immunoglobulin like receptor, two Ig domains and long cytoplasmic tail 1
MMP9	matrix metallopeptidase 9
P2RX1	purinergic receptor P2X 1
PLEKHG3	pleckstrin homology and RhoGEF domain containing G3
PRR14	proline rich 14

PTGDS	prostaglandin D2 synthase
RGS2	regulator of G protein signaling 2
TLE3	transducin like enhancer of split 3
TNFRSF9	TNF receptor superfamily member 9
TTC38	tetratricopeptide repeat domain 38

GSN na assinatura gênica

Gene Symbol	Gene name
ATM	ATM serine/threonine kinase
ATP8B2	ATPase phospholipid transporting 8B2
BANK1	B cell scaffold protein with ankyrin repeats 1
CLUAP1	clusterin associated protein 1
COPS8	COP9 signalosome subunit 8
CYB5A	cytochrome b5 type A
DAAM1	dishevelled associated activator of morphogenesis 1
DKK3	dickkopf WNT signaling pathway inhibitor 3
EIF3B	eukaryotic translation initiation factor 3 subunit B
ENOSF1	enolase superfamily member 1
ENTPD6	ectonucleoside triphosphate diphosphohydrolase 6 (putative)
FAM102A	family with sequence similarity 102 member A
FAM117A	family with sequence similarity 117 member A
FCRL2	Fc receptor like 2
IL21R	interleukin 21 receptor
IL23A	interleukin 23 subunit alpha
INHBE	inhibin subunit beta E
KRAS	KRAS proto-oncogene, GTPase
KRT1	keratin 1
KRT16	keratin 16
LRRN3	leucine rich repeat neuronal 3
LTBP3	latent transforming growth factor beta binding protein 3
MATN2	matrilin 2
NUDT4	nudix hydrolase 4
ORC5	origin recognition complex subunit 5

PDZRN3	PDZ domain containing ring finger 3
PKLR	pyruvate kinase L/R
PLCG1	phospholipase C gamma 1
RHOH	ras homolog family member H
RPLP1	ribosomal protein lateral stalk subunit P1
RPS28	ribosomal protein S28
SIPA1L3	signal induced proliferation associated 1 like 3
SKP2	S-phase kinase associated protein 2
SMC1A	structural maintenance of chromosomes 1A
TPP2	tripeptidyl peptidase 2
TUBB2A	tubulin beta 2A class IIa
TULP4	tubby like protein 4
UROD	uroporphyrinogen decarboxylase
XIST	X inactive specific transcript

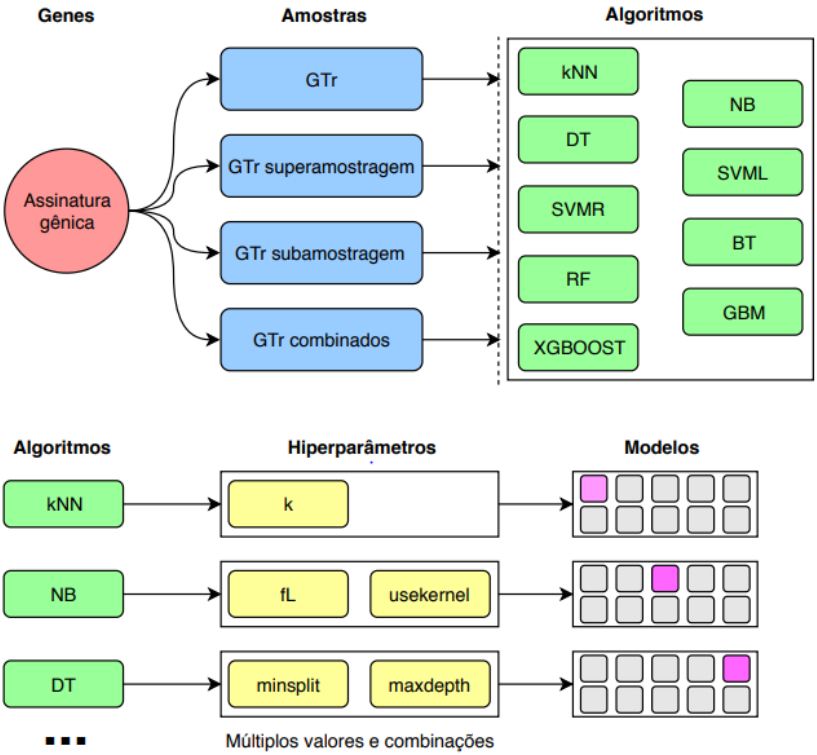
5.8 COMPETÊNCIA PREDITIVA DE MODELOS DE CLASSIFICAÇÃO EM AMOSTRAS DE SANGUE DE PACIENTES COM A DOENÇA DE PARKINSON IDIOPÁTICA

Utilizando os genes da assinatura e os diferentes GTr, foram criados modelos empregando algoritmos de classificação e ponderando suas predições.

A **Figura 20** representa um organograma da metodologia de criação dos modelos de classificação, envolvendo a utilização de genes da assinatura identificada em amostras de diferentes GTr. Essa matriz gerada de genes/amostras é aplicada em diferentes algoritmos de classificação que, por sua vez, formam vários modelos resultados de algoritmos e de múltiplos valores e de múltiplas combinações de hiperparâmetros. Os modelos resultantes da combinação algoritmos/hiperparâmetros com os maiores valores de AUC foram selecionados para análises de predição.

Os hiperparâmetros testados para cada algoritmo e os valores (dos hiperparâmetros) que geraram os modelos com as maiores capacidades de predição (medida pelas áreas sob a curva ROC) estão expostos no **Quadro 8**. Utilizando os valores selecionados de hiperparâmetros foram criados 4 modelos finais, um para cada GTr, para cada algoritmo.

Figura 20. Organograma de criação de modelos de classificação e seleção de valores de hiperparâmetros.



Em vermelho: assinatura gênica. Em azul: GTr utilizados. Em verde: algoritmos de aprendizado de máquina experimentados. Em amarelo: hiperparâmetros modulados na criação dos modelos. Em cinza: modelos criados. Em violeta: modelos com os menores valores de AUC.

Quadro 8. Hiperparâmetros selecionados.

Modelo	**Hiperparâmetro escolhido**
k-Nearest Neighbors	`"k" = 18`
Naive Bayes	`"fL" = 0` `"usekernel" = TRUE`
Decision Trees	`"minsplit" = 20` `"maxdepth" = 30`
Support Vector Machine Linear	`"C" = 1`

Support Vector Machine Radial	"sigma" = 0.0102 "C" = 1
Bagged Trees	–
Random Forest	"mtry" = 2 "ntree" = 1000
Gradient Boost Machine	"n.trees" = 150 "interaction.depth" = 3 "shrinkage" = .1 "n.minobsinnode" = 10
eXtreme Gradient Boosting	"nrounds" = 150 "max_depth" = 3 "eta" = .3 "colsample_bytree" = 0.8

O desempenho dos modelos foi estimado com a utilização do GTe. O GTe foi composto por 142 amostras não utilizadas no GTr, sendo 85 (59,9%) amostras de pacientes com a DP idiopática e 57 (40,1%) amostras de indivíduos sadios. O GTe foi composto por 18 amostras do GSE6613, 18 do GSE57475, 13 do GSE72267 e 93 do GSE99039, representando 25%, 12,7%, 22,1% e 21,3% do total de amostras de cada conjunto de dados, respectivamente. Quanto ao fenótipo das amostras no GTe, no GSE6613 12 amostras (66,6%) eram relativas a pacientes com a DP idiopática e 6 eram de indivíduos sadios; no GSE57475 a relação foi de 15 (83,3%)/3; no GSE72267 foi de 11 (84,6%)/2, e no GSE99039 foi de 47 (50,5%)/46. As **Tabelas 2** a **5** contém os valores relativos as predições efetuadas de modelos gerados pelos diferentes GTr.

Os valores de AUC dos algoritmos, que indicam uma acurácia aonde o valor 1 representa um teste perfeito e 0.5 um teste sem valor, variaram de 0,742 a 0,794 nos modelos treinados com o GTr, de 0,744 a 0,792 nos treinados com o GTr superamostragem, de 0,738 a 0,800 nos treinados com o GTr subamostragem e de 0,671 a 0,803 nos treinados com o GTr combinados. Os algoritmos que não contém resultados de AUC não apresentaram significância estatística na diferença da acurácia e da taxa de não-informação. Dessa forma, em todos os GTr houve algoritmos capazes de separar de forma razoável/boa os pacientes pelos fenótipos.

5.9 REFINAMENTO PARA APLICAÇÃO DE MODELOS

Os modelos responsáveis às melhores classificações de cada classe (sensibilidade e especificidade) foram investigados para um

aperfeiçoamento e para um modo de utilização de modelos combinados explorando suas aptidões.

As **Figuras 21** a **24** contém as porcentagens de acerto (predições exatas) para cada conjunto e para os fenótipos de amostras dos algoritmos que apresentaram as melhores relações de sensibilidade (correta predição de amostras de pacientes com a DP idiopática) e especificidade (correta predição de indivíduos sadios).

O maior valor de sensibilidade dentre os algoritmos treinados com o GTr foi de 0,835 com o *Support Vector Machine Linear*, dentre os treinados com o GTr superamostragem foi de 0,882 com o *Random Forest*, dentre os treinados com o GTr subamostragem foi de 0,776 com o *Gradient Boosting Machine* e dentre os treinados com o GTr combinados foi de 0,764 com o *Gradient Boosting Machine*.

Destes, o *Support Vector Machine Linear* treinado com o GTr apresentou de 69,89 a 88,89% de acertos nos estudos e de 75 a 100% de acertos em amostras de pacientes com a DP idiopática (**Figura 21 A**). O *Random Forest* treinado com o GTr superamostragem apresentou de 64,52 a 100% de acertos nos estudos e de 80,85 a 100% de acertos em amostras de pacientes com a DP idiopática (**Figura 22 A**). O *Gradient Boosting Machine* treinado com o GTr subamostragem apresentou de 66,67 a 100% de acertos nos estudos e de 72,34 a 100% de acertos em amostras de pacientes com a DP idiopática (**Figura 23 A**). O *Gradient Boosting Machine* treinado com o GTr combinados apresentou de 66,67 a 100% de acertos nos estudos e de 80,85 a 100% de acertos em amostras de pacientes com a DP idiopática (**Figura 24 A**).

O maior valor de especificidade dentre os algoritmos treinados com o GTr foi de 0,719 com o *k-Nearest Neighbors*¸ dentre os treinados com o GTr superamostragem foi de 0,771 com o *Support Vector Machine Linear*, dentre os treinados com o GTr subamostragem foi de 0,789 com o *k-Nearest Neighbors* e dentre os treinados com o GTr combinados foi de 0,771 com o *k-Nearest Neighbors*.

Destes, o *k-Nearest Neighbors* treinado com o GTr apresentou de 67,74 a 88,89% de acertos nos estudos e de 67,39 a 100% de acertos em amostras do grupo controle (**Figura 21 B**). O *Support Vector Machine Linear* treinado com o GTr superamostragem apresentou de 61,54 a 88,89% de acertos nos estudos e de 71,74 a 100% de acertos em amostras do grupo controle (**Figura 22 B**). O *k-Nearest Neighbors* treinado com o GTr subamostragem apresentou de 64,52 a 83,33% de acertos nos estudos

e de 69,57 a 100% de acertos em amostras do grupo controle (**Figura 23 B**). O *k-Nearest Neighbors* treinado com o GTr combinados apresentou de 60,22 a 77,78% de acertos nos estudos e de 73,91 a 100% de acertos em amostras do grupo controle (**Figura 24 B**).

As **Figuras 25** a **28** apresentam as probabilidades de classe de cada amostra para os dois modelos, de maior sensibilidade e maior especificidade treinados com cada GTr. Os gráficos estão partidos em duas facetas, *Controle* e *Parkinson* referentes as classificações verídicas de cada uma das amostras e subdividido em duas colunas *Certo* e *Errado* referentes as suas classificações pelo modelo, como corretas ou incorretas. Testes de Wilcoxon-Mann-Whitney para amostras independentes (após os testes de normalidade de Shapiro-Wilk) foram utilizados na comparação entre as classificações corretas e incorretas.

Dentre os algoritmos de maior sensibilidade para cada GTr, o *Support Vector Machine Linear* treinado com o GTr e o *Gradient Boosting Machine* treinado com o GTr combinados apresentaram diferença estatística entre as médias de probabilidade de classe no grupo de amostras de pacientes com a DP idiopática (p = 0,01 e 0,04, respectivamente). O algoritmo *Gradient Boosting Machine* treinado com o GTr subamostragem apresentou diferença estatística entre as médias de probabilidade de classe no grupo de amostras de indivíduos sadios (p = 0,03).

Dentre os algoritmos de maior especificidade para cada GTr, o *k-Nearest Neighbors* treinado com o GTr, pelo GTr subamostragem e pelo GTr combinados apresentaram diferença estatística entre as médias de probabilidade de classe no grupo de amostras de pacientes sadios (p = $4x10^{-3}$, $5x10^{-4}$ e $6x10^{-3}$, respectivamente). Nenhum dos algoritmos apresentou diferença estatística entre as médias de probabilidade de classe no grupo de amostras de pacientes com a DP idiopática.

Tendo em vista as diferenças estatísticas entre as predições corretas e incorretas, a otimização dos modelos foi finalizada com o corte dos 25% das amostras com os menores valores para classe *Parkinson* em modelos de maior sensibilidade e dos 25% das amostras com os menores valores para classe *Controle* em modelos de maior especificidade.

A sensibilidade do *Support Vector Machine Linear* treinado com o GTr subiu de 0,835 para 0,904, do *Random Forest* treinado com o GTr superamostragem subiu de 0,882 para 0,920, do *Gradient Boosting Machine* treinado com o GTr subamostragem subiu de 0,776 para 0,793

e do *Gradient Boosting Machine* treinado com o GTr combinados subiu de 0,764 para 0,841.

O menor valor de probabilidade de classe acima do corte do *Support Vector Machine Linear* treinado com o GTr foi 0,583, do *Random Forest* treinado com o GTr superamostragem foi 0,654, do *Gradient Boosting Machine* treinado com o GTr subamostragem foi 0,630 e do *Gradient Boosting Machine* treinado com o GTr combinados foi 0,709.

A especificidade do *k-Nearest Neighbors* treinado com o GTr subiu de 0,719 para 0,881, do *Support Vector Machine Linear* treinado com o GTr superamostragem subiu de 0,772 para 0,785, do *k-Nearest Neighbors* treinado com o GTr subamostragem subiu de 0,789 para 0,904 e do *k-Nearest Neighbors* treinado com o GTr combinados subiu de 0,772 para 0,833.

O menor valor de probabilidade de classe acima do corte do *k-Nearest Neighbors* treinado com o GTr foi 0,611, do *Support Vector Machine Linear* treinado com o GTr superamostragem foi 0,563, do *k-Nearest Neighbors* treinado com o GTr subamostragem foi 0,611 e do *k-Nearest Neighbors* treinado com o GTr combinados foi 0,611.

Foram escolhidos os modelos com o maior valor de probabilidade de classe acima do corte para compor o par ótimo, por serem os valores de maior dificuldade de obtenção. Dessa forma, foram selecionados o *Gradient Boosting Machine* treinado com o GTr combinados e o *k-Nearest Neighbors* treinado com o GTr (**Figura 29**).

5.10 APLICAÇÃO DE MODELOS EM EXEMPLARES DE OUTRAS CATEGORIAS

Utilizando amostras de outras categorias contidas nos conjuntos avaliados, aplicou-se o algoritmo de maior sensibilidade escolhido, o *Gradient Boosting Machine* treinado com o GTr combinados. Os grupos de amostras da doença de Huntington, da DP genética por mutações nos genes *LRRK2* e *PRKN* apresentaram diferenças estatísticas entre as médias dos grupos e a média do grupo *Parkinson* (pacientes com a DP idiopática) ($p = 0{,}0063$, $0{,}0271$ e 0,0003, respectivamente), enquanto as da doença de Alzheimer, da atrofia multissistêmica, da paralisia supranuclear progressiva e da DP genética por mutações nos genes ATP13A2 e PINK1 não apresentaram (**Figura 30**).

Tabela 2. Predições com os modelos criados com o GTr.

	kNN	**NB**	**DT**	**SVML**	**SVMR**	**BT**	**RF**	**GBM**	**XGBOOST**
VP	62	63	59	71	69	65	70	68	64
VN	41	38	32	36	37	34	36	33	39
FP	23	22	26	14	16	20	15	17	21
FN	16	19	25	21	20	23	21	24	18
Acurácia	0.7254	0.7113	0.6408	0.7535	0.7465	0.6972	0.7465	0.7113	0.7254
NIR	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986
Valor de *p* [Acu > NIR]	1.11E-03	3.47E-03	1.73E-01	7.60E-05	1.56E-04	9.56E-03	1.56E-04	3.47E-03	1.11E-03
Sensitividade	0.7294	0.7412	0.6941	0.8353	0.8118	0.7647	0.8235	0.8000	0.7529
Especificidade	0.7193	0.6667	0.5614	0.6316	0.6491	0.5965	0.6316	0.5789	0.6842
AUC	0.7943	0.7697	-	0.7903	0.7756	0.7427	0.7760	0.7575	0.7802

kNN = *k-Nearest Neigbors*, NB = *Naive Bayes*, DT = *Decision Trees*, SVML = *Support Vector Machine Linear*, SVMR = *Support Vector Machine Radial*, BT = *Bagged Trees*, GBM = *Gradient Boosting Machine*, XGBOOST = *eXtreme Gradient Boosting*, NIR = *No Information Rate* (taxa de não-informação), Acu = Acurácia, AUC = *Area Under the Curve* (área sob a curva ROC).

Tabela 3. Predições com os modelos criados com o GTr superamostragem.

	kNN	**NB**	**DT**	**SVML**	**SVMR**	**BT**	**RF**	**GBM**	**XGBOOST**
VP	55	67	63	64	70	67	75	69	69
VN	43	34	33	44	37	28	29	32	31
FP	30	18	22	21	15	18	10	16	16
FN	14	23	24	13	20	29	28	25	26
Acurácia	0.6901	0.7113	0.6761	0.7606	0.7535	0.6690	0.7324	0.7113	0.7042
NIR	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986
Valor de *p* [Acu > NIR]	1.52E-02	3.47E-03	3.49E-02	3.56E-05	7.60E-05	5.07E-02	5.97E-04	3.47E-03	5.85E-03
Sensitividade	0.6471	0.7882	0.7412	0.7529	0.8235	0.7882	0.8824	0.8118	0.8118
Especificidade	0.7544	0.5965	0.5789	0.7719	0.6491	0.4912	0.5088	0.5614	0.5439
AUC	0.7743	0.7445	-	0.7909	0.7889	-	0.7831	0.7926	0.7773

kNN = *k-Nearest Neigbors*, NB = *Naive Bayes*, DT = *Decision Trees*, SVML = *Support Vector Machine Linear*, SVMR = *Support Vector Machine Radial*, BT = *Bagged Trees*, GBM = *Gradient Boosting Machine*, XGBOOST = *eXtreme Gradient Boosting*, NIR = *No Information Rate* (taxa de não-informação), Acu = Acurácia, AUC = *Area Under the Curve* (área sob a curva ROC).

Tabela 4. Predições com os modelos criados com o GTr subamostragem.

	kNN	**NB**	**DT**	**SVML**	**SVMR**	**BT**	**RF**	**GBM**	**XGBOOST**
VP	57	63	59	60	66	62	66	66	60
VN	45	38	32	42	40	37	41	39	35
FP	28	22	26	25	19	23	19	19	25
FN	12	19	25	15	17	20	16	18	22
Acurácia	0.7183	0.7113	0.6408	0.7183	0.7465	0.6972	0.7535	0.7394	0.669
NIR	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986
Valor de *p* [Acu > NIR]	2.00E-03	3.47E-03	1.73E-01	2.00E-03	1.56E-04	9.56E-03	7.60E-05	3.11E-04	5.08E-02
Sensitividade	0.6706	0.7412	0.6941	0.7059	0.7765	0.7294	0.7765	0.7765	0.7059
Especificidade	0.7893	0.6667	0.5614	0.7368	0.7018	0.6491	0.7193	0.6842	0.6140
AUC	0.8009	0.7721	-	0.7856	0.7804	0.7388	0.7801	0.7812	-

kNN = *k-Nearest Neigbors*, NB = *Naive Bayes*, DT = *Decision Trees*, SVML = *Support Vector Machine Linear*, SVMR = *Support Vector Machine Radial*, BT = *Bagged Trees*, GBM = *Gradient Boosting Machine*, XGBOOST = *eXtreme Gradient Boosting*, NIR = *No Information Rate* (taxa de não-informação), Acu = Acurácia, AUC = *Area Under the Curve* (área sob a curva ROC).

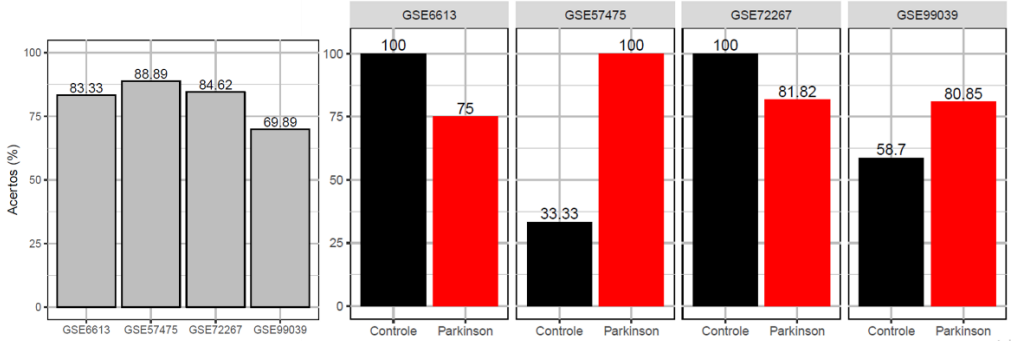
Tabela 5. Predições com os modelos criados com o GTr combinados.

	kNN	**NB**	**DT**	**SVML**	**SVMR**	**BT**	**RF**	**GBM**	**XGBOOST**
VP	60	63	64	62	67	67	71	65	31
VN	44	40	19	37	39	35	37	36	36
FP	25	22	21	23	18	18	14	20	24
FN	13	17	38	20	18	22	20	21	21
Acurácia	0.7324	0.7254	0.5845	0.6972	0,7035	0,6638	0,7015	0.7113	0.6831
NIR	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986	0.5986
Valor de *p* [Acu > NIR]	5.97E-04	1.11E-03	6.67E-01	9.56E-03	1.56E-04	2.00E-03	3.56E-05	3.47E-03	2.33E-02
Sensitividade	0.7059	0.7412	0.7529	0.7294	0.7212	0.7176	0.7529	0.7647	0.7176
Especificidade	0.7719	0.7018	0.3333	0.6491	0.6842	0.6140	0.6491	0.6316	0.6316
AUC	0.8032	0.7754	-	0.7567	0,7311	0,6713	0,7124	0.7373	0.7381

kNN = *k-Nearest Neigbors*, NB = *Naive Bayes*, DT = *Decision Trees*, SVML = *Support Vector Machine Linear*, SVMR = *Support Vector Machine Radial*, BT = *Bagged Trees*, GBM = *Gradient Boosting Machine*, XGBOOST = *eXtreme Gradient Boosting*, NIR = *No Information Rate* (taxa de não-informação), Acu = Acurácia, AUC = *Area Under the Curve* (área sob a curva ROC).

Figura 21. Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr.

A – Maior sensibilidade, *Support Vector Machine Linear*



B – Maior especificidade, *k-Nearest Neighbors*

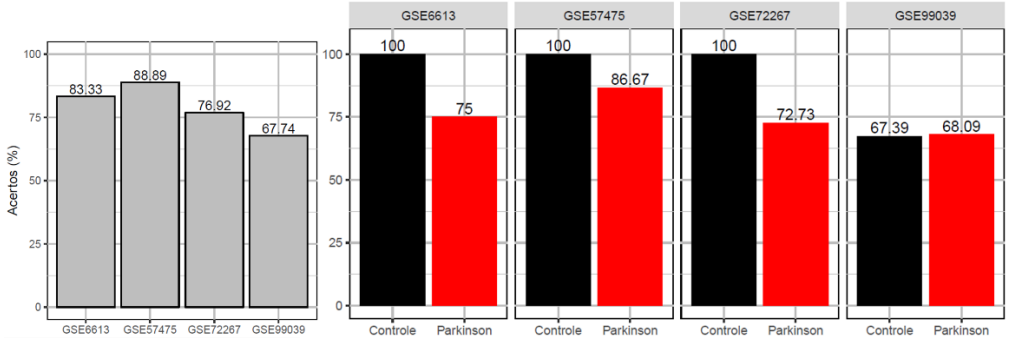
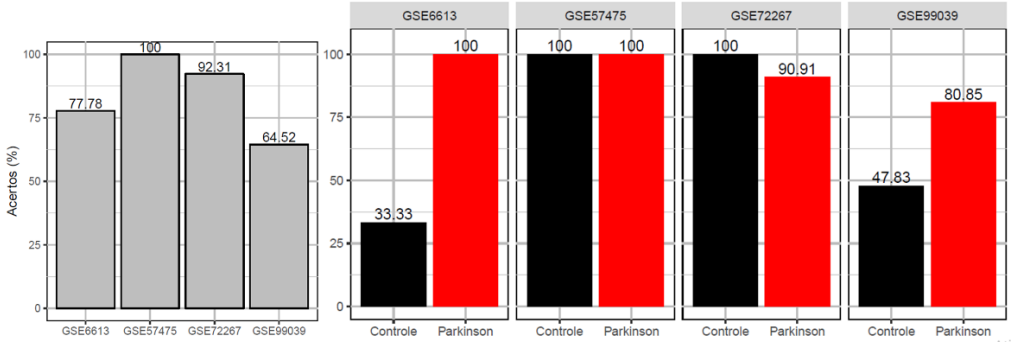


Figura 22. Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr superamostragem.

A – Maior sensibilidade, *Random Forest*



B – Maior especificidade, *Support Vector Machine Linear*

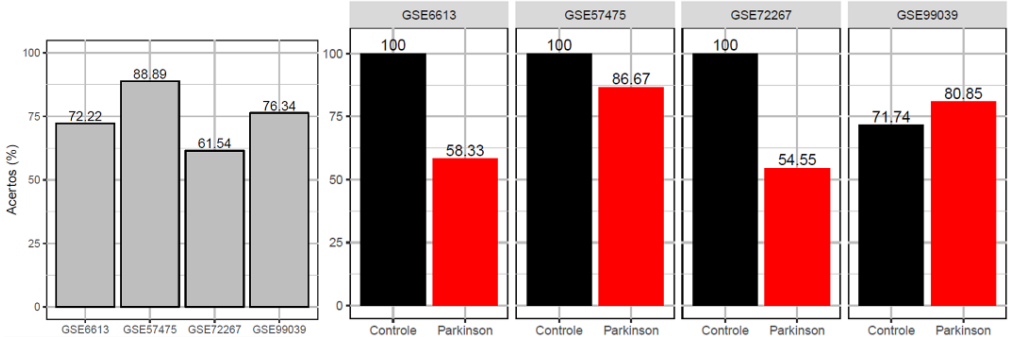
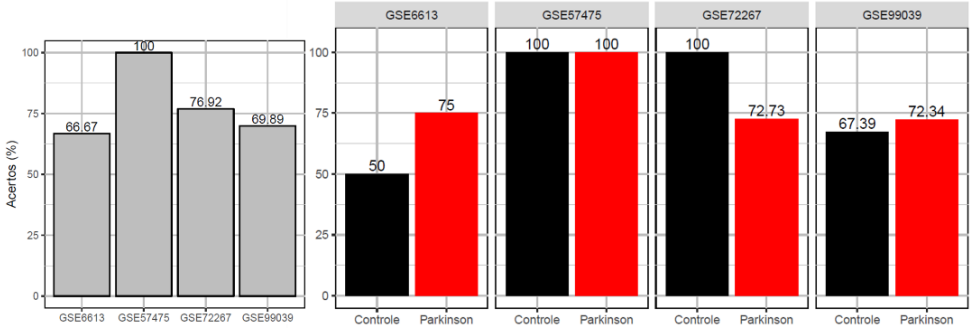


Figura 23. Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr subamostragem.

A – Maior sensibilidade, *Gradient Boosting Machine*



B – Maior especificidade, *k-Nearest Neighbors*

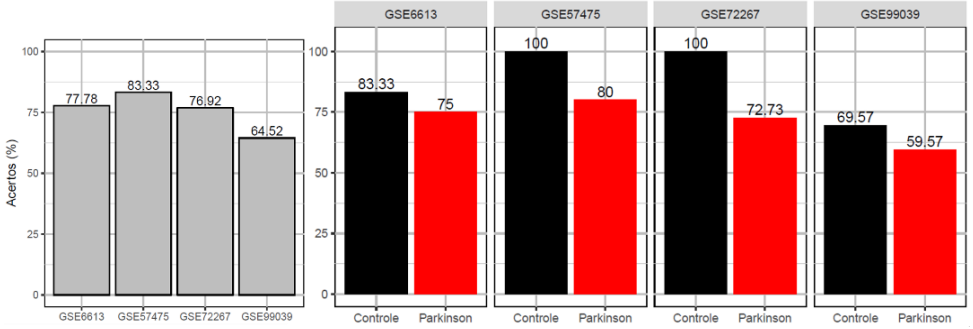
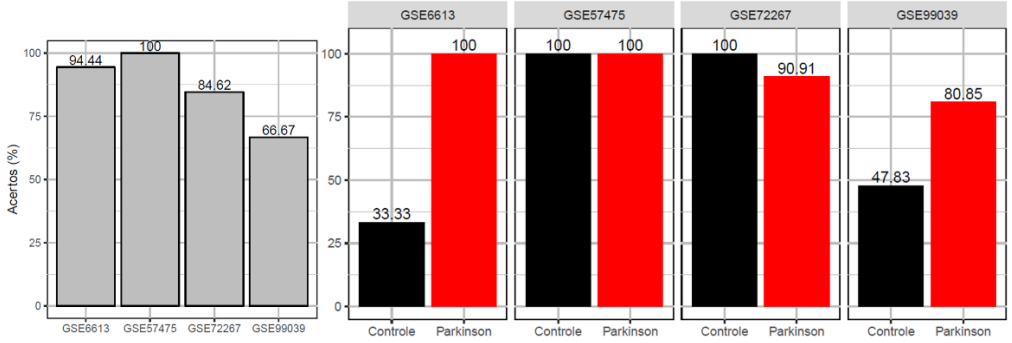


Figura 24. Porcentagens de acerto de algoritmos com a maior sensibilidade e especificidade utilizando o modelo criado com o GTr combinados.

A – Maior sensibilidade, *Gradient Boosting Machine*



B – Maior especificidade, *k-Nearest Neighbors*

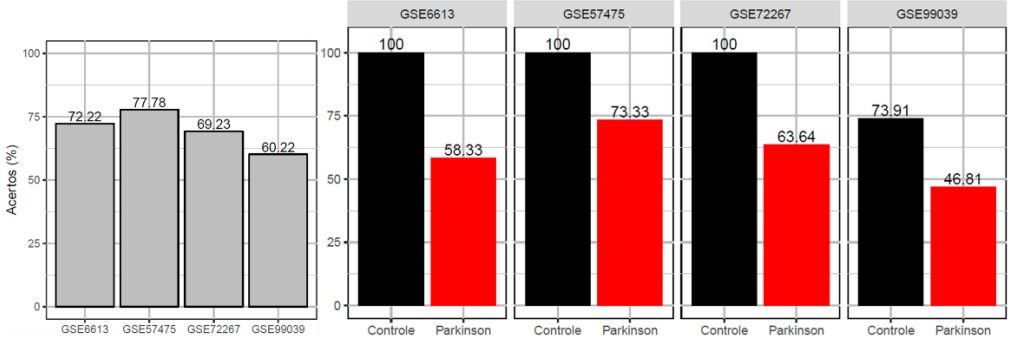
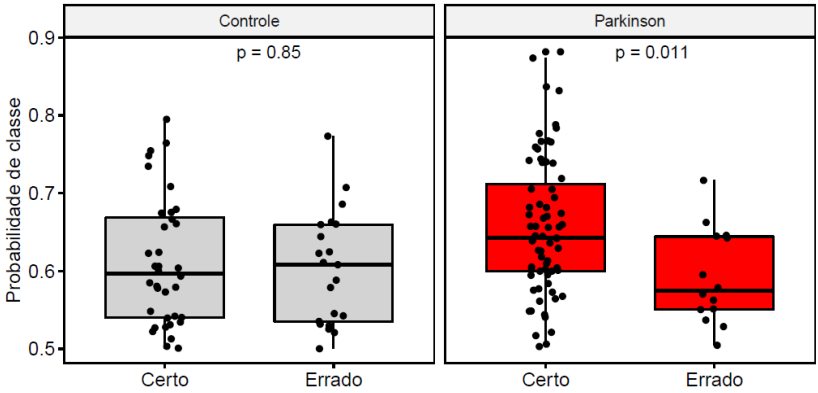


Figura 25. Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr.

A – Maior sensibilidade, *Support Vector Machine Linear*



B – Maior especificidade, *k-Nearest Neighbors*

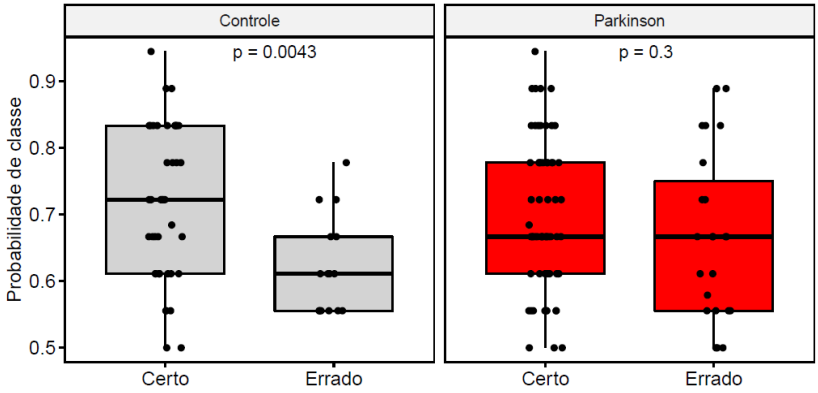
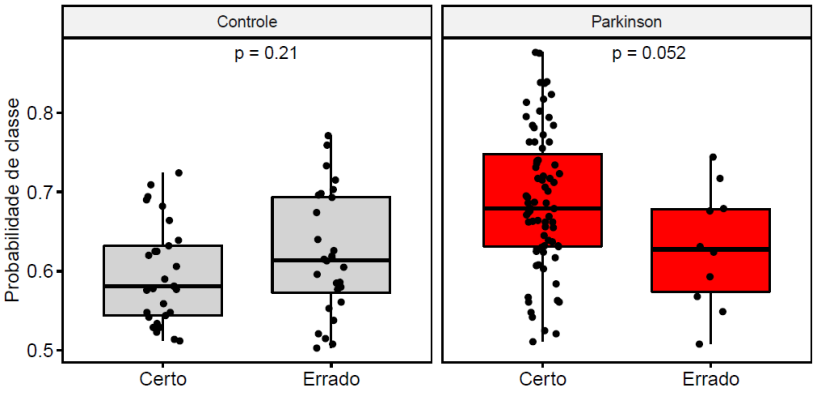


Figura 26. Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr superamostragem.

A – Maior sensibilidade, *Random Forest*



B – Maior especificidade, *Support Vector Machine Linear*

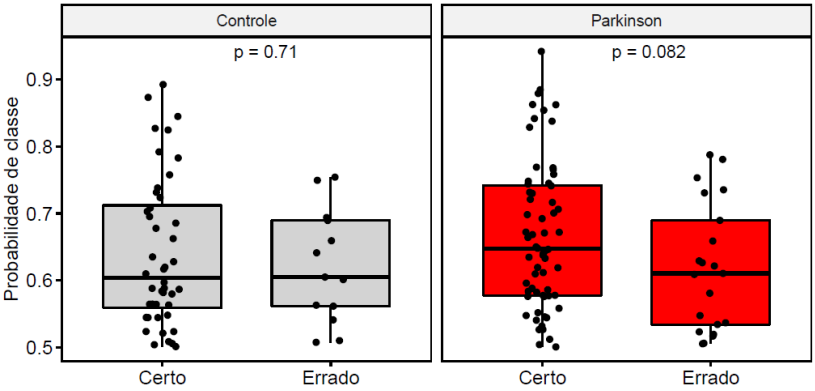
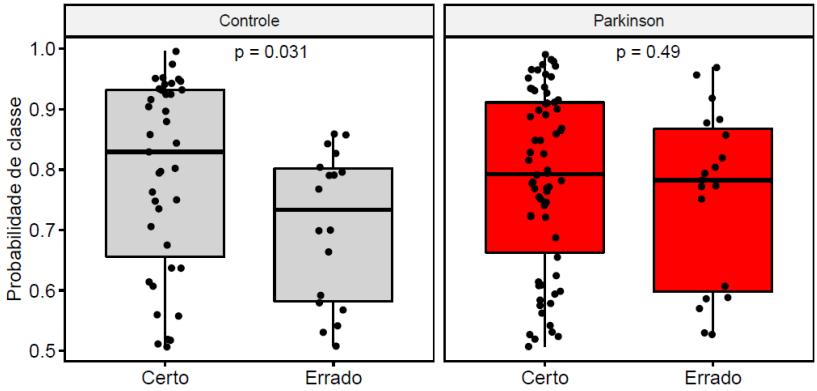


Figura 27. Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr subamostragem.

A – Maior sensibilidade, *Gradient Boosting Machine*



B – Maior especificidade, *k-Nearest Neighbors*

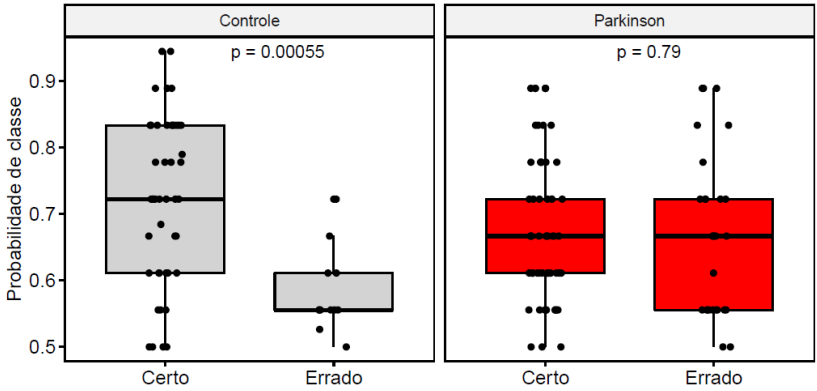
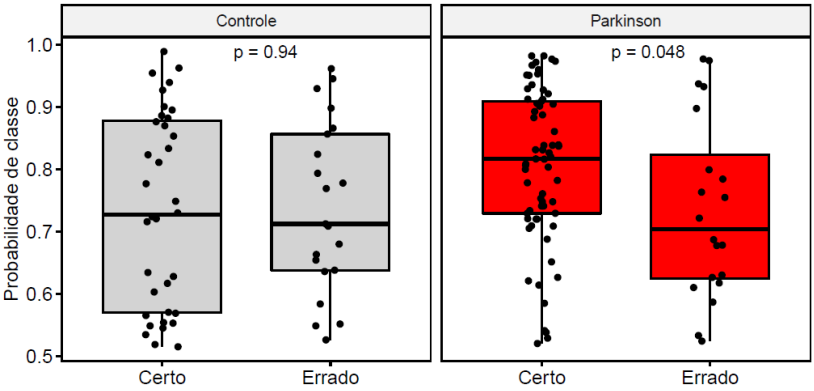


Figura 28. Probabilidades de classes nos modelos com maiores sensibilidade e especificidade elaborados com GTr combinados.

A – Maior sensibilidade, *Gradient Boosting Machine*



B – Maior especificidade, *k-Nearest Neighbors*

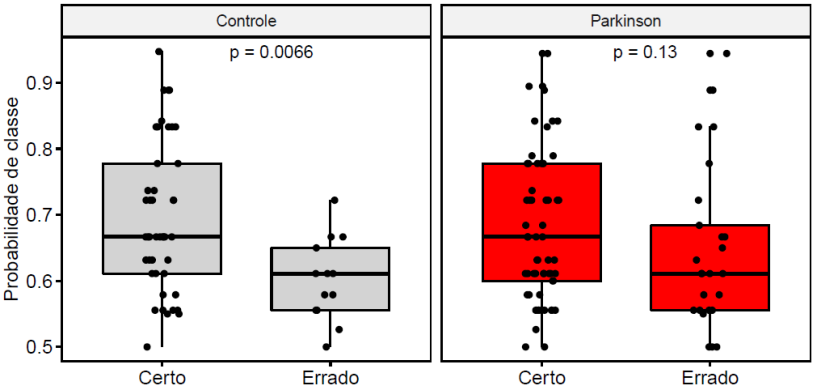
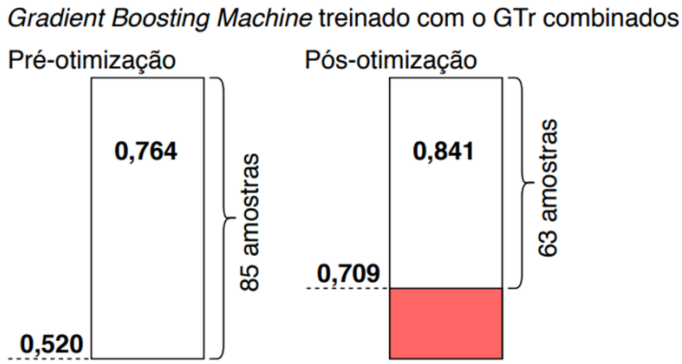


Figura 29. Modelos escolhidos com os limites de maior valor de probabilidade de classe.

A



B

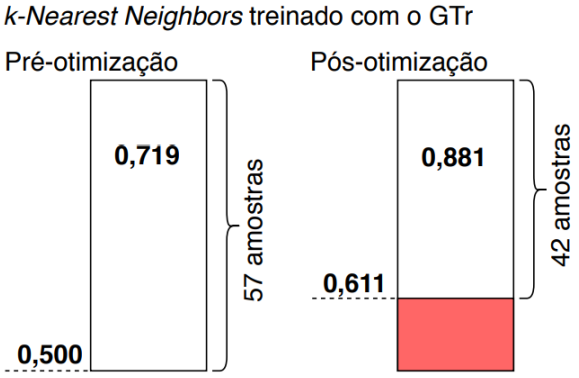
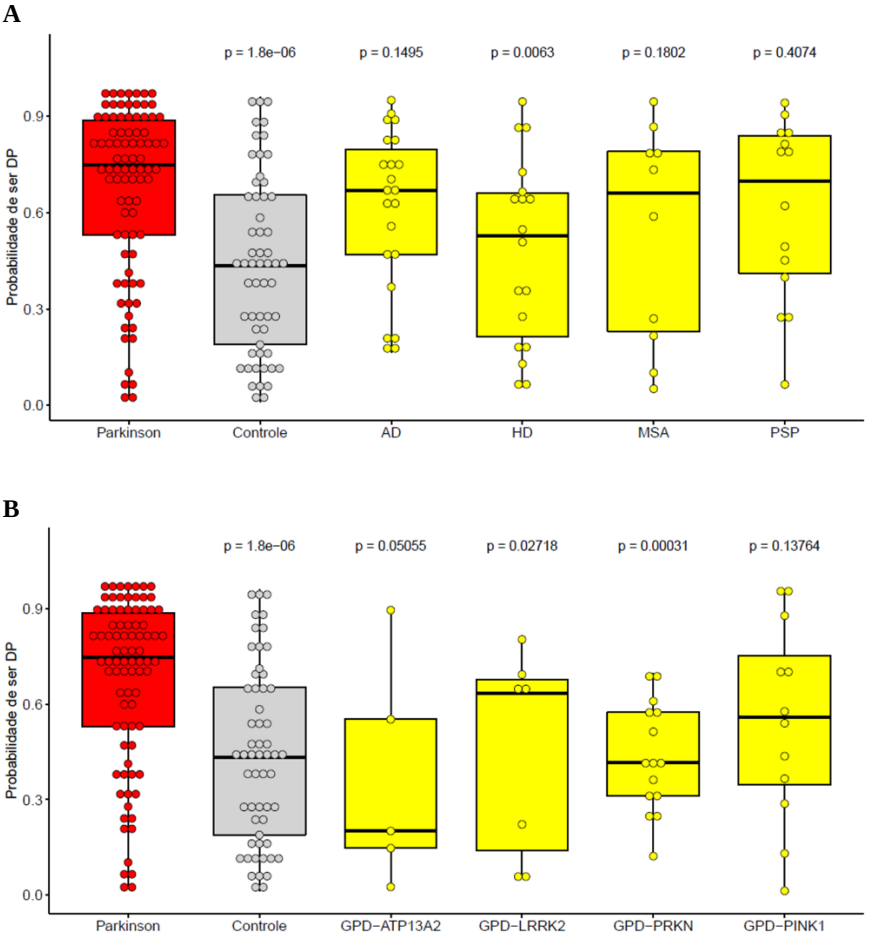


Figura 30. Probabilidades de predição na classe *Parkinson* de amostras de outras categorias.



AD: Doença de Alzheimer, HD: Doença de Huntington, MSA: Atrofia multissistêmica, PSP: Paralisia supranuclear progressiva, GPD-ATP13A2: DP genética com mutações no gene *ATP13A2*, GPD-LRRK2: DP genética com mutações no gene *LRRK2*, GPD-PRKN: DP genética com mutações no gene *PRKN* e GPD-PINK1: DP genética com mutações no gene *PINK1*.

5.11 COMPETÊNCIA PREDITIVA DO MODELO DE CLASSIFICAÇÃO BASEADO EM IMAGENS E ENRIQUECIDO POR COLINEARIDADE

Para elaboração de um modelo de classificação capaz de valorizar as colinearidades de expressão de genes foram utilizadas redes de correlação como estruturas para imputação de valores de expressão, como um mapa de paisagem, e utilizando classificações de imagens geradas.

Para elaboração da rede de correlação gênica os valores de correlação de Pearson entre cada par de genes de todos os conjuntos foram extraídos (**Apêndices AA a AH**) e os valores de *p* de significância estatística foram combinados utilizando o método de Fisher, um teste estatístico de χ^2 (**Apêndices AI e AJ**).

Os valores de χ^2 foram reescalonados para um intervalo de 0 a 1 e foram montadas redes com esses valores como ponderamento (pesos) entre os conectores. As maiores significâncias estatísticas combinadas resultavam em conectores com uma força de atração maior entre os nós.

As redes elaboradas com os 500 conectores com maior força de atração dos GSP e dos GSN estão apresentadas nas **Figuras 31 e 32** e a descrição dos nós e conectores estão nos **Apêndices AK e AL**. A rede com os GSP contém 69 nós (genes, círculos em vermelho) e dos 500 conectores 423 (84,6%) são de pares de genes co-expressos descritos na literatura (conectores, linhas, em vermelho). A rede com os GSN contém 65 nós (genes, círculos em azul) e dos 500 conectores 248 (49,6%) são de pares de genes co-expressos descritos na literatura (conectores, linhas, em azul).

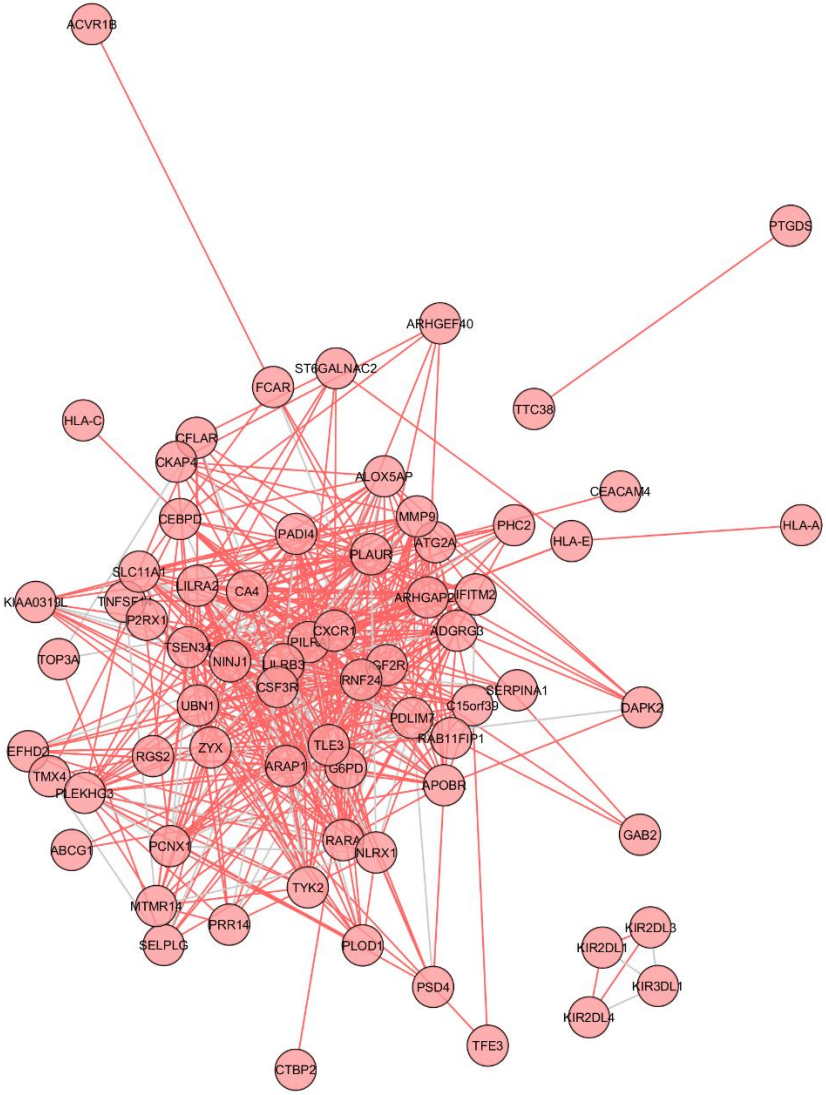
As redes foram agrupadas em um mesmo ambiente/espaço no programa Cytoscape (**Figura 33**) de forma a ocuparem o menor comprimento em ambos os eixos, condensando-as em um quadrado para formular as imagens.

Utilizando essa estrutura foram criadas imagens topográficas de cada amostra, utilizando o programa ViaComplex. Na **Figura 34** estão representados exemplos de imagens criadas pelo programa. Em A e B estão representadas uma amostra de DP idiopática e de um indivíduo do grupo controle, respectivamente. Percebe-se as diferenças entre as cores da parte superior e inferior de ambas as imagens. Na amostra de DP idiopática (A), os genes da parte de cima, correspondentes aos GSP, estão em um vermelho vivo, apontando a expressão elevada desses genes

nesses indivíduos, conforme era esperado. Entretando, os genes da parte de baixo, correspondentes aos GSN, estão entre os tons de azul e verde, apontando a expressão reduzida desses genes nesses indivíduos. Na amostra do grupo controle (B) não há uma inversão, propriamente, mas sim uma atenuação entre esses dois extremos marcados nas amostras de DP. Entre as 711 amostras foram reparadas variações entre esses padrões, próprios não somente da variação amostral, também do procedimento de reescalonamento, onde os valores tiveram que estar entre 0 e 1 no conjunto de amostras.

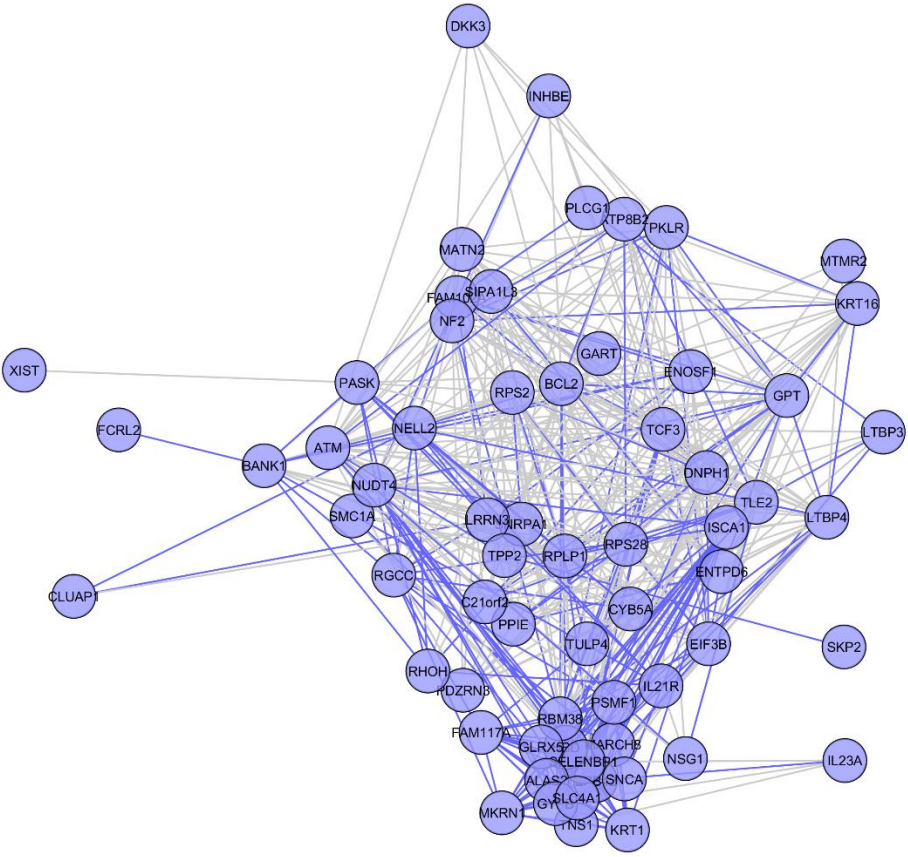
Estas imagens foram utilizadas como dados para elaboração de um modelo de classificação aplicando o algoritmo de *deep learning MXNET*. O GTr continha 80% das amostras e o GTe continha os 20% restantes. A predição resultou em uma acurácia de 0,8414, uma AUC de 0,8628, uma sensibilidade de 0,8142 e uma especificidade de 0,8800. A NIR continuava sendo 0,5986. Dessa forma, esse algoritmo, no sentido da sequência de passos do procedimento como um todo, apresentou bons resultados sem uma afinação com hiperparâmetros ou utilização de ferramenta de correção do desbalanço amostral. Entretanto ainda há muitos pontos de arbitrariedade no procedimento.

Figura 31. Estrutura da rede de GSP



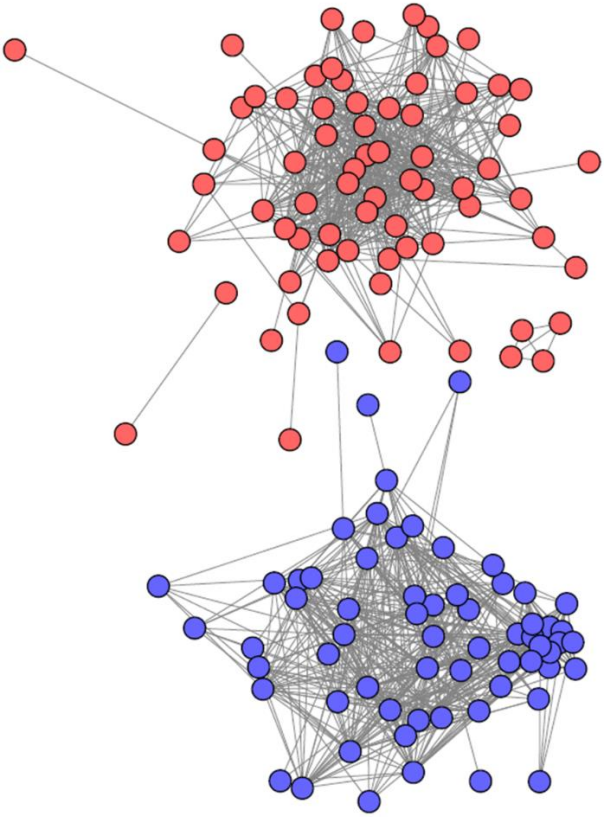
Nós (círculos) vermelhos: GSP. Extensão de conectores (linhas): correlação gênica. Conectores vermelhos: co-expressões identificadas no GeneMANIA.

Figura 32. Estrutura da rede de GSN



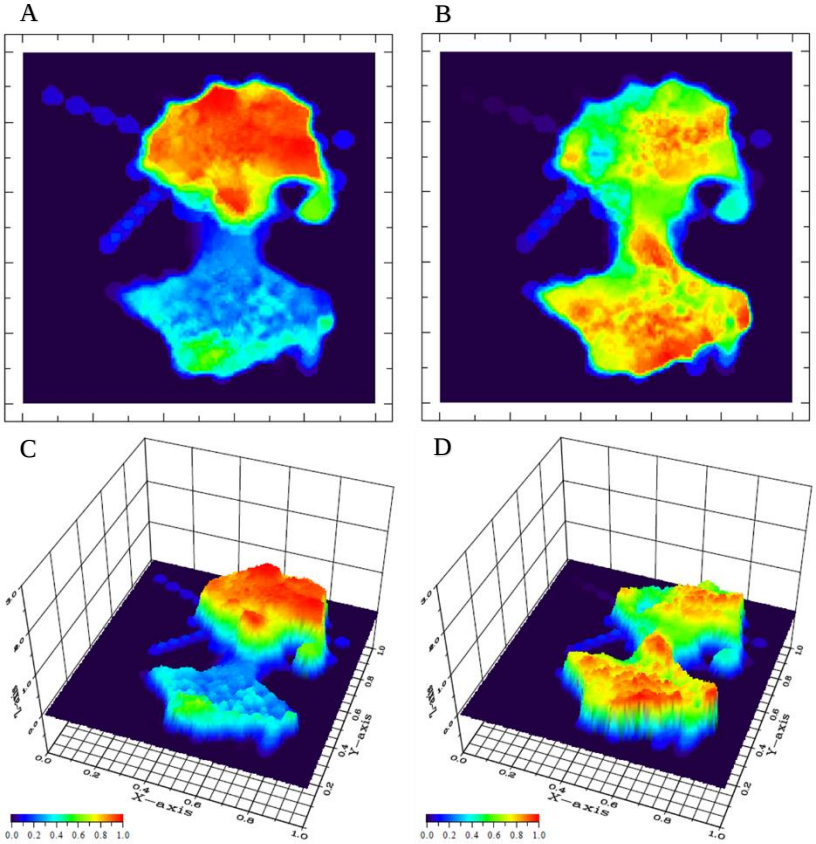
Nós (círculos) azuis: GSN. Extensão de conectores (linhas): correlação gênica. Conectores azuis: co-expressões identificadas no GeneMANIA.

Figura 33. Estruturas combinadas das redes de GSP e GSN.



Nós (círculos) vermelhos: GSP. Nós azuis: GSN.

Figura 34. Exemplos de topologias de redes geradas com o ViaComplex.



A e **C**: Exemplo de rede de amostras de pacientes com Parkinson idiopático, **B** e **D**: Exemplo de rede de amostras de pacientes do grupo controle. **A** e **B**: Gráfico 3D, **C** e **D**: Superfície 3D

6 DISCUSSÃO

No presente trabalho foi realizada a maior meta-análise de dados de microarranjos de DNA do sangue periférico de pacientes com a DP idiopática e indivíduos do grupo controle utilizando os tamanhos de efeito como recursos robustos de comparação de dados entre estudos. Os diferentes estudos analisados incluíam amostras de pacientes com a DP idiopática que utilizavam tratamento farmacológico e pacientes recém diagnosticados, ainda sem tratamento, pacientes com uma ampla variedade de idades e estágios da doença, montando um trabalho de meta-análise composto para enfrentar as diferenças encontradas em dados da literatura.

A primeira intenção da pesquisa científica é a estimação da magnitude e da direção do real, do existente. Os pesquisadores podem fazer isso pela produção de estimados de tamanhos de efeito (ELLIS, 2010). É essencial que não se busque apenas a interpretação da significância estatística, por mostrar somente a improbabilidade de achados e não a sua significância prática (SULLIVAN; FEINN, 2012). Para este estudo, buscou-se a modificação periférica processada em pacientes com uma doença neurodegenerativas, ou seja, distante do tecido afetado cerebral.

A mensuração de tamanhos de efeito, como realizada aqui, é primordial para interpretação de resultados de um estudo. A Associação Americana de Psicologia (APA ou *American Psychological Association* em inglês) identifica a "falha em reportar os tamanhos de efeito" como um dos sete erros comuns de manuscritos científicos (APA, 2001). Da mesma forma, a Associação Americana de Pesquisa Educacional (AERA ou *American Educational Research Association* em inglês) em seus padrões de publicação já recomendam que as análises estatísticas sejam acompanhadas de tamanhos de efeito e uma interpretação qualitativa do efeito (AERA, 2006).

É comum achar estudos de transcriptoma ou de associação genômica ampla (GWAS ou *Genome-wide association study*, em inglês) para as mesmas condições em que cada estudo contém um tamanho amostral pequeno com poder estatístico baixo. É necessária a combinação de estudos como este para aumentar a sensitividade e validar as conclusões (RAMASAMY et al., 2008). No caso da DP idiopática foram obtidos 4 estudos independentes que somavam 388 amostras de pacientes

com a doença e 323 amostras de indivíduos sadios no total. Deste total, 61% das amostras são provenientes do conjunto de dados GSE99039, depositado em 2017 (CALLIGARIS et al., 2015; LOCASCIO et al., 2015; SCHERZER et al., 2007; SHAMIR et al., 2017).

Foram utilizadas três fontes na obtenção de dados de microarranjo. Entre os conjuntos de dados utilizados, dois vieram do repositório público de dados Array Express e dois de artigos científicos. Apesar de não ter sido possível obter dados utilizando as palavras-chave no GEO, o maior repositório de dados atual, todos os dados achados estavam presentes nessa plataforma. Esse fato aponta para a necessidade de palavras-chave representativas e da busca em múltiplas fontes na elaboração de meta-análises de genômica funcional.

Neste estudo foram utilizados trabalhos com conjuntos de dados a nível de pacientes de forma as eliminar os vieses de algoritmos particulares utilizados em cada estudo original. Arquivos de imagem de *slides* de microarranjo não são normalmente depositadas em repositórios públicos e nem tecnologicamente uniformes para serem usadas em meta-análises (RAMASAMY et al., 2008). Listas gênicas publicadas em artigos dependem de cálculos de pré-processamento, métodos de análise, limiares de significância e mecanismos de anotação gênica específicos e não são recomendadas para meta-análises (TSENG; GHOSH; FEINGOLD, 2012). As matrizes de expressão de dados a nível de pacientes (sem pré-processamento) possuem os valores de expressão de cada gene para cada uma das amostras e são o *input* ideal para a meta-análise por possibilitarem os cálculos de tamanhos de efeito.

Os estudos analisados continham metadados muito distintos entre si. As informações sobre os dados eram principalmente relacionadas ao grupo de amostras e não aos pacientes. Os conjuntos GSE6613 e GSE72267 continham somente os metadados de fenótipo de amostras (*Parkinson* ou *Controle*), o GSE57475 ainda continha metadados de idade e sexo de amostras, enquanto o GSE99039 continha metadados de estagiamento da doença, informação sobre a idade atual da coleta do sangue e a idade do primeiro diagnóstico da DP para maior parte dos dados. A publicação de referências mais detalhadas para amostras auxiliaria a estratificação de pacientes com a DP para pesquisas mais direcionadas resultando em prognósticos, diagnósticos mais seguros e desenvolvimento de fármacos como já realizado para muitos tipos câncer (JASKOWIAK; COSTA; CAMPELLO, 2018). Heterogeneidades devido

a variáveis demográficas, clinicas e tecnológicas geralmente ocorrem entre estudos, porém a falha em reportar e refletir essas variáveis em modelamento estatístico e na meta-análise pode suceder em menores poderes estatístivos e maior número de falsos positivos (TSENG; GHOSH; FEINGOLD, 2012). Em meta-análises de microarranjos as variáveis sistemáticas devem ser consideradas e incorporadas sempre.

A DP idiopática é geralmente diagnosticada a partir de 55 - 60 anos e sua prevalência cresce a partir de 70 - 75 anos (POEWE et al., 2017). As médias de idade representadas neste estudo variam de 62 a 69,4 anos com desvio padrão de 51 a 77,8 anos. Quanto a severidade da doença, as médias representadas variam de 1,8 a 2,3 com desvio padrão de 0,9 a 3 na escala Hoehn-Yahr (HOEHN; YAHR, 1967). Nesta escala, a DP é classificada em 5 estágios tendo como base a sintomatologia motora. O tempo gasto em cada estágio é mutável em pacientes e o salto entre os estágios não é incomum (ZHAO et al., 2010). Os estágios abrangidos aqui variam de 1 a 3. O estágio 1 é a fase "inicial" da DP com sintomas que incluem a presença de tremores ou agitação em um membro (superior ou inferior) e atinge apenas um lado (unilateral – direito ou esquerdo), geralmente são observadas alterações na postura, perda de equilíbrio e de expressão facial (apatia facial). O estágio 2, por sua vez, os sintomas dos pacientes são bilaterais e afetam membros superiores e inferiores, o paciente encontra dificuldades para caminhar ou manter o equilíbrio e as tarefas físicas normais se tornam mais desafiadoras. No estágio 3 os sintomas são bastante agravados e incluem a incapacidade de andar em linha reta ou de ficar em pé, os movimentos físicos apresentam desaceleração observável e podem ocorrer episódios de congelamento da marcha (HOEHN; YAHR, 1967). Dessa forma, o presente estudo contém amostras de pacientes com Parkinson com fenótipos muito distintos e os achados devem mostrar robustez para muitos cenários.

No pré-processamento dos dados não foram observados arranjos de má qualidade, provavelmente por uma filtragem prévia do(s) autor(es) sobre os dados. Quanto ao processo de normalização de dados, não há um consenso na literatura sobre os algoritmos capazes de efetuar medidas de expressão comparáveis entre plataformas para meta-análise (RAMASAMY et al., 2008; TSENG; GHOSH; FEINGOLD, 2012). Entretando, foi concluído que algoritmos baseados no modelo (no caso na tecnologia ou na plataforma do microarranjo) são os que melhor executam seus empregos (RAMASAMY et al., 2008) e, sendo assim, foram

utilizados os algoritmos padrão-ouro para microarranjos de cDNA como o GSE57475, da plataforma Illumina e para *chips* de oligonucleotídeos como o GSE6613, GSE72267 e GSE99039, da plataforma Affymetrix. Uma etapa comum na análise de microarranjos é a eliminação não-específica de sondas, fazendo a análise em um filtrado de sondas, entretanto também não existe uma conclusão sobre essa etapa opcional ser benéfica na perspectiva de meta-análise (RAMASAMY et al., 2008). Com base nisso, todas as sondas foram mantidas e para a seleção, no caso de muitas sondas para um gene anotado, foi optado pelo menor valor de *p* no teste *t* de expressão diferencial (RHODES et al., 2002), identificando o valor que tem menor probabilidade de ocorrência ao acaso. A maioria dos pacotes de realização de meta-análises em R não possuem manuais práticos ou reuniam funções de difícil aplicação, principalmente para análises que utilizam as informações individuais de pacientes. Dessa forma, metodologias claras e objetivas para a análise, pacotes convenientes para seus aplicamentos (seja em R ou ambiente semelhante de programação) e o desenvolvimento de programas com interfaces gráficas favoráveis ao usuário permitirão os pesquisadores testar, comparar e, por fim, elaborar metodologias que auxiliarão os cientistas em suas conclusões (TSENG; GHOSH; FEINGOLD, 2012).

Neste trabalho foram realizadas meta-análises de valores de expressão de 17.712 genes obtidos em estudos independentes de microarranjo de DNA utilizando a combinação de tamanhos de efeito. Estima-se que no genoma humano contenha de 19.000 a 20.000 genes codificantes a proteínas (EZKURDIA et al., 2014) e, dessa forma, a meta-análise conteve aproximadamente (por contar alguns genes não codificantes a proteínas) de 88,5 a 93,2% transcritos destes do genoma humano analisados. Para meta-análise, os genes deveriam estar presentes em pelo menos 2 conjuntos de dados. Sendo assim, os genes representados em menos amostras foram os 5309 genes da intersecção do GSE57475 e do GSE99039, totalizando 580 amostras.

Há mais de 70 variedades de tamanhos de efeito (KIRK, 2003). A forma escolhida para representação e interpretação do efeito neste estudo foi o coeficiente de correlação produto-momento de Pearson (*r*). Os valores de correlação de Pearson, quando calculados a partir de um método de comparação de grupos como o Hedges' *g*, são referentes a uma correlação entre os valores de expressão de um gene e um fenótipo, aqui definido como a DP idiopática. Para o GSE6613, GSE57475, GSE72267

e GSE99039 respectivamente, utilizando o limiar descrito (ROSENTHAL; ROSNOW, 1984), 0,5%, 0%, 1,6% e 0% dos genes apresentaram tamanhos de efeito "grande", 10,7%, 0,4%, 13% e 0,1% dos genes apresentaram tamanhos de efeito "médio" e 48,0%, 30,5%, 44,3% e 22,7% dos genes apresentaram tamanhos de efeito "pequenos". Esses dados demonstram as diferenças de expressão entre os genes em cada estudo e esse fato deve se relacionar as variedades populacionais entre esses estudos, entendendo que alguns deles tinham objetivos distintos, como Calligaris e colaboradores (2015) (responsáveis pelo trabalho GSE72267) que visavam estudar o sangue de pacientes não medicados, diagnosticados para a DP pouco antes da coleta de sangue, enquanto os outros utilizavam amostras de pacientes medicados e com a DP mais avançada (LOCASCIO et al., 2015; SCHERZER et al., 2007; SHAMIR et al., 2017). Novamente a falta de dados, ou metadados destes trabalhos dificulta (ou impossibilita) as conclusões de conjuntos individuais.

Dentre os 17.712 genes da meta-análise, 0,01% (2) apresentaram tamanhos de efeito "médio" (*PTGDS* e *ADGRG3* com *r* igual a 0,273 e 0,256, respectivamente) e 13,2% (2343) apresentaram tamanhos de efeito "pequeno". A disparidade entre os valores de porcentagens da meta-análise e de cada análise individual reflete a dificuldade de identificar genes expressos diferencialmente em um grupo com amostras tão dispares.

O *PTGDS* (*Prostaglandin D2 Synthase* ou Prostaglandina D2 sintetase, em inglês) codifica a proteína prostaglandina D2 sintetase responsável pela catalisação da prostaglandina H2 (PGH2) em prostaglandina D2 (PGD2). A PGD2 atua como um neuromodulador e fator trófico no sistema nervoso central e é envolvida com a contração e o relaxamento da musculatura lisa além da inibição da agregação plaquetária. Estudos manipulando camundongos transgênicos para superexpressão de *PTGDS* sugerem que este gene possa estar envolvido na regulação do sono não-REM (*Rapid Eye Movement* ou Movimento ocular rápido, em inglês) (PINZAR et al., 2000). Essa proteína é expressa preferencialmente no fluido cerebroespinal, humor vítreo e urina. O estímulo de células da glia é uma característica geralmente encontrada em pacientes com a DP, tornando a variação na expressão e em isoformas de *PTGDS* um grande interesse para assinaturas moleculares uma vez que reflete patofisiologias em meninges e na glia, fontes comuns de PGD2 (HARRINGTON et al., 2006; MANDEL et al., 2010). A redução de

variantes 3 e 4 do gene em fluido cerebroespinal de pacientes com a DP identificada por Harrington e colaboradores (2006) é correlacionada a alterações retratadas em outros trabalhos, sugerindo a adequação do gene e de suas isoformas como candidatos diagnósticos para DP. De fato, o gene está presente na assinatura conseguida neste trabalho. O *ADGRG3* (*Adhesion G Protein-Coupled Receptor G3* ou Receptor G3 acoplado a proteína G de adesão, em inglês) codifica a proteína homônima relacionada as vias de interação molecular de sistema imune inato. Ainda são escassas as informações na literatura quanto a função da proteína, entretanto estudos *in vitro* apontam a regulação da migração de células do endotélio linfático através de GTPases RhoA pequenas como a CDC42.

Dentre os 17.712 genes da meta-análise, 2661 (15%) genes analisados apresentaram diferenças na distribuição de médias insinuando a existência de dois ou mais grupos populacionais de pacientes com a DP idiopática. Isso pode ser explicado por características distintas dos estudos, conforme discutido, porém não foram analisadas as causas. O reconhecimento de subtipos de amostras de pacientes pode favorecer a otimização de assinatura gênica, adequando-a aos grupos.

O passo mais comum em estudos de análise de microarranjos é a análise de expressão diferencial, onde se selecionam os genes cujas expressões são significativamente diferentes entre as categorias. Neste trabalho foram realizadas as reanálises de expressão diferencial de cada estudo utilizando estruturas comuns de análise. No total, os 4 conjuntos de dados apresentaram 578 GDE utilizando os limiares de p-valor $\leq 0{,}05$ para o teste-t e valor absoluto positivo de *fold change* $\geq 1{,}2$ e apenas 20 desses (3,4%) estavam presentes em dois ou mais conjuntos. Esses resultados auxiliam a explicar a inconsistência verificada em estudos de microarranjo da DP e a falta de marcadores utilizados na clínica para prognósticos, diagnóstico, foco de tratamentos e outros.

Os limiares (*thresholds*) utilizados do ponto de vista de análises de expressão diferencial são baixos, indicando diferenças de aproximadamente 20% (*fold change* absoluto $\geq 1{,}2$) entre as médias de expressão de pacientes com a DP idiopática e indivíduos sadios. Entretanto, esses valores foram anteriormente utilizados em outros trabalhos que aplicaram análise de expressão diferencial ou meta-análise em sangue de pacientes com a DP idiopática (SANTIAGO; LITTLEFIELD; POTASHKIN, 2016; SANTIAGO; POTASHKIN, 2017; SCHERZER et al., 2007). O estabelecimento de limiares baixos

para apuração dos GDE está pertinente as características da própria análise, que busca estes genes em um tecido distante da área onde ocorre a neurodegeneração e em um tecido de bastante variação transcricional, como é o sangue periférico, variando juntamente ao ciclo circadiano (BOIVIN et al., 2003), ingesta alimentar (LEONARDSON et al., 2009) e outros, principalmente em células do sistema imunológico (ECKER et al., 2018; MARTINEZ-JIMENEZ et al., 2017), como os leucócitos, identificados na análise funcional. Além do mais, um *fold change* absoluto ≥ 2, como aplicado em diversos trabalhos, resultaria em zero GDE em GSE6613 e GSE57475, dois no GSE72267 (*BTNL3* e *HLA-DQA1*) e um no GSE99039 (*XIST*). Da mesma forma, não foram aplicados cortes de testes múltiplos pelo mesmo motivo, apesar do conhecimento das repercussões sobre os aumentos de erros do tipo I.

Dos GDE que estavam presentes em mais de um conjunto de dados, três estavam presentes nos GSP (*ADGRG3*, *MMP9* e *PTGDS*), sete nos GSN (*EPB42*, *GYPB*, *HBD*, *LRRN3*, *MARCH8*, *SNCA* e *XIST*) e quatro na assinatura gênica criada (*LRRN3*, *MMP9*, *PTGDS* e *XIST*).

As amostras de cada estudo foram organizadas em um agrupamento hierárquico com base nos valores de expressão dos GSP e dos GSN, para observar se esses genes conseguiam as separar em dois grupos quanto ao fenótipo das amostras. Os valores de acurácia devem ser maiores do que os valores de taxa de não-informação salientando a capacidade de grupos produzidos de selecionar (por separação) as amostras no grupo "certo" de forma acurada. Tal metodologia não é um algoritmo de classificação com aprendizados, foi uma forma utilizada de idealizar a separação natural nos estudos com esses genes. O único conjunto que retratou acurácia maior que a taxa de não-informação foi o GSE99039 que apresentou uma classificação 1% maior que o acaso. Todavia, tal diferença não é relevante para aplicação clínica.

As análises funcionais de ontologias enriquecidas de genes selecionados foram realizadas utilizando a função goana do pacote *limma* e o banco de dados DAVID. As significâncias detectadas em genes designados nessa meta-análise indicam relevância de termos enriquecidos na DP, de forma geral. Os 10 termos enriquecidos utilizando o método goana/limma com maior significância estatística dos GSP foram todos relacionados a processos do sistema imune, principalmente da imunidade mediada pelos leucócitos, sua ativação e sua degranulação. Os leucócitos são células do sistema imune envolvidas com o processo inflamatório. A

inflamação é uma reação frente a uma infecção ou uma lesão tecidual. Sabe-se que a inflamação crônica é uma das muitas características de doenças neurodegenerativas (GLASS et al., 2010), incluindo a DP e que a ativação da micróglia no mesencéfalo, região cerebral onde se localiza a SN, é correlacionada a manifestação da sintomatologia motora na DP (MCGEER et al., 1988; OUCHI et al., 2005). Experimentos *in vitro* mostraram que a proteína α-sinucleína mal-conformada ativa a micróglia que, por sua vez, secreta citocinas como o TNF-α (SU et al., 2008) e a IL-β (KLEGERIS et al., 2008) e a produção de espécies reativas de oxigênio que danificam os neurônios dopaminérgicos (ZHANG et al., 2005). O TNF- α pode induzir a síntese a IL-β, uma interleucina pró-inflamatória (SU et al., 2008). Outros termos, vindos do método de busca no banco de dados DAVID distinguiram enriquecimento de vias de sinalização por interferon-α e interferon-γ. O interferon-α é produzido por leucócitos geralmente em resposta as infecções virais e sua produção é estimulada por TNF- α e IL-1, como a IL-β (SEN, 2001). O interferon-γ é produzido por leucócitos geralmente para ativação de macrófagos, potencialização de interferon e estimulação da expressão do complexo maior de histocompatibilidade, regulando as respostas inflamatórias, sua produção é estimulada por IL-12 E IL-18 (TEIJARO, 2016). Esses achados, somados a achados semelhantes em outros artigos de análise de expressão utilizando microarranjos de DNA e microarranjos de miRNA, sugerem uma relação complexa e dinâmica entre a neuroinflamação em processos de neurodegeneração e a ativação do sistema imune periférico.

Termos semelhantes ao sistema imune foram encontrados em outras pesquisas de avaliação transcritptomica da DP. Os termos relacionados ao sistema imune, como sistema imune inato e adaptativo e inflamação se apresentavam como enriquecidos em amostras de sangue periférico de pacientes (ALIEVA et al., 2014; CHIKINA et al., 2015), termos relacionados a inflamação, neuroinflamação e sinalizações do extravasamento de leucócitos se apresentavam como enriquecidos em células nucleadas sanguíneas de pacientes com a DP idiopática e com a DP genética com mutações no gene LRRK2 (MUTEZ et al., 2011; SOREQ et al., 2008, 2012)

Os 10 termos enriquecidos utilizando o método goana/limma com maior significância estatística dos GSN foram menos associados entre si, que os de GSP, porém muito associados aos 2 termos enriquecidos

utilizando o método de busca no banco de dados DAVID. O termo "poliubiquitinação de proteínas" e os termos pertinentes a organização do citoesqueleto estavam presentes em ambos os modos de busca. A ubiquitina é uma proteína que desempenha uma função na regulação de proteínas. Ela age como uma etiqueta que marca as proteínas indesejadas (como as mal-conformadas, como a proteína α-sinucleína na DP) para que elas sejam degradadas pelo complexo proteico proteossoma. Os processos de ubiquitinação e de poliubiquitinação, ambos enriquecidos por GSN, fazem parte da via da ubiquitina-proteassoma. Com poucas exceções, apenas proteínas poliubiquitinadas conseguem ter acesso ao núcleo proteolítico do proteassoma. Sabe-se que a homeostase intracelular da proteína α-sinucleína é mantida pelas ações do sistema ubiquitina-proteassoma e do sistema de autofagia lisossômica. A inibição ou prejuízo de qualquer um desses sistemas leva ao aumento de níveis de α-sinucleína e a sua possível acumulação (XILOURI et al., 2009). O envelhecimento, o maior fator de risco para o desenvolvimento da DP idiopática está correlacionado as reduções de ambos os sistemas, o que é plausível com as observações de aumento de níveis de α-sinucleína em neurônios dopaminérgicos da SN no decorrer normal do envelhecimento (CHU; KORDOWER, 2007; KAUSHIK; CUERVO, 2015). Adicionalmente, os oligômeros de α-sinucleína presentes na DP podem inibir o sistema ubiquitina-proteossoma (EMMANOUILIDOU; STEFANIS; VEKRELLIS, 2010). A presença de genes envolvidos na poliubiquitinação de proteínas (*BCL2*, *MARCH8*, *MKRN1*, *PSMF1*, *SKP2* e *TPP2*) entre os GSN e o enriquecimento de termos utilizando os dois métodos de análise funcional apontam para o reflexo no sangue periférico de um processo presente no sistema nervoso central complexo entre a proteína α-sinucleína e o sistema ubiquitina-proteossoma.

Termos semelhantes a ubiquitinação de proteínas foram encontrados em outras pesquisas de avaliação transcriptomica da DP. As vias de interação molecular ubiquitina-proteossoma se apresentaram como enriquecidas em amostras de sangue periférico de pacientes (SCHERZER et al., 2007) e os termos de ubiquitinação de proteínas e proteólise entreposta por ubiquitina foram identificados nas células nucleadas sanguíneas quando analisados os microRNA (CHATTERJEE et al., 2014; MARTINS et al., 2011). Termos semelhantes a organização do citoesqueleto, como "moléculas de adesão" e "organização da matriz extracelular", foram identificados no sangue periférico de pacientes com

a DP idiopática e com a DP genética com mutações no gene LRRK2 (INFANTE et al., 2016).

Dois diferentes mecanismos foram utilizados na seleção de preditores, resultando em uma assinatura com 58 genes. Tal apuração foi positiva do ponto de vista da aplicação clínica, onde uma assinatura com menos genes resulta em um teste barateado, e do ponto de vista matemático, de onde se buscam modelos que possuam menor complexidade e maior estabilidade, dois motivos para exclusão de preditores colineares (KUHN; JOHNSON, 2013), motivo da utilização da primeira operação. O segundo algoritmo utilizado para seleção de preditores, o algoritmo de eliminação de preditores recursivo (GUYON et al., 2002) busca o grupo de preditores de maior relevância na capacidade de predição do fenótipo testando todas as possíveis combinações de preditores em todos os tamanhos de grupo. É um algoritmo eficiente e com ampla aplicação. Dessa forma, foram escolhidos 58 genes presentes em todos os conjuntos de dados para elaboração da assinatura gênica de classificação diagnóstica da DP idiopática.

Funcionalmente, os GSP da assinatura apresentaram enriquecimento de termos relacionados a resposta imune (5 genes) e a regulação do processo apoptótico (4 genes), enquanto os GSN não apresentaram enriquecimento de termos de ontologias.

Utilizando a assinatura formada foram criados diversos modelos utilizando várias opções de hiperparâmetros em 9 algoritmos de classificação. Ao final foram gerados 9 modelos (referentes a 9 diferentes algoritmos) para cada GTr, ou seja, um total de 36 modelos representando os resultados superiores de algoritmos utilizando um devido GTr. Somente os modelos de *Decision Trees* de todos os GTr, o *Bagged Trees* do GTr superamostragem e o *eXtreme Gradient Boosting* do GTr subamostragem não conseguiram significância estatística da relação entre a acurácia e a taxa de não-informação. Todos esses foram modelos criados sobre algoritmos de árvores de decisão e é conhecida a baixa acurácia desses na formação de modelos com variáveis numéricas, geralmente expressando alta instabilidade (LANTZ, 2015). Uma vez que todos os dados preditores utilizados eram variáveis numéricas retratando as expressões de genes da assinatura, esperava-se que esses algoritmos mostrassem dificuldades de classificação.

Entre os modelos criados para cada GTr foram selecionados os de maiores valores de sensibilidade e especificidade. A sensibilidade indica a capacidade do modelo de classificar acertadamente as amostras de pacientes com a DP idiopática, enquanto a especificidade indica para as amostras de indivíduos sadios. Para o caso estudado, resultados falsos-positivos, ou seja, diagnosticar pacientes como a DP, quando são sadios, é menos grave que resultados contrários, de falsos-negativos, por atrapalharem um tratamento de pronto início e por deslocarem a descoberta de fármacos, de marcadores de diagnóstico, prognóstico, estratificação e outros. Dessa forma os valores de sensibilidade foram os de maior relevância.

O modelo que obteve o maior valor de sensibilidade foi o *Random Forest* treinado com o GTr superamostragem com 0,882, enquanto o modelo com o maior valor de especificidade foi o *k-Nearest Neighbors* treinado com o GTr com 0,835.

Utilizando os 8 modelos criados, 4 com maiores valores de sensibilidade e 4 de sensitividade, um de cada GTr, foram calculadas as probabilidades de classe de amostras do GTe. Dessa forma pôde-se considerar as diferenças de médias de classificações corretas e incorreta. Os dois modelos com maiores valores de sensibilidade que tiveram diferença significativa foram o *Support Vector Machine* do GTr e o *Gradient Boosting Machine* do GTr combinados. Os três modelos com maiores valores de especificidade que tiveram diferença significativa foram os *k-Nearest Neighbors* do GTr, GTr subamostragem e GTr combinados. Percebe-se a eficiência de algoritmos baseados em instâncias, como o *k-Nearest Neighbors* e o *Support Vector Machine*, esses algoritmos não empreendem generalizações explicitas e sim comparam as novas amostras com amostras observadas no treino. No caso do *k-Nearest Neighbors* a análise é realizada com base nas disposições de pontos em um espaço bidimensional, e no caso do *Support Vector Machine* a análise é realizada pela relação/posição da amostra e do suporte formado, em um espaço bi- ou tridimensional. Devido as diferenças de expressão entre as condições serem muito sutis, é provável que estes algoritmos atuem concedendo maiores valores de probabilidade de classes pelas classificações serem realizadas baseadas nas posições das amostras e das relações entre amostras e não somente nos valores.

Dos modelos que tiveram diferenças significativas entre as predições constatou-se que há mais amostras preditas incorretas com

menores valores de probabilidade de classe e, dessa forma, uma demarcação limite seria capaz de melhorar os resultados. O intervalo do limite criado foi com a eliminação de 25% de amostras de menor probabilidade. Todas as sensibilidades e especificidades foram otimizadas com tal tratamento, apresentando melhoras de 1,7% a 7,7% de sensibilidade e 1,3% a 14,5% de especificidade. Foram escolhidos os modelos que tiveram os maiores valores de probabilidade de classe acima do corte estabelecido por serem valores de maior improbabilidade de obtenção arbitrária. Dessa forma, foram selecionados para classificação de pacientes com a DP idiopática o modelo *Gradient Boosting Machine* treinado com o GTr combinados, que apresentou 84,12% de acerto nas amostras com probabilidade de classe > 0,7095 (ou 70,95%) para DP idiopática e para classificação de indivíduos sadios o modelo *k-Nearest Neighbors* treinado com o GTr, que apresentou 88,09% de acerto nas amostras com probabilidade de classe > 0,6666 (ou 66,66%). Apesar de apenas 105 amostras (75% do GTe) serem classificadas, as predições mostraram 85,71% de acerto.

Para testar caso a assinatura e o modelo *Gradient Boosting Machine* treinado com o GTr combinados são exclusivos a pacientes com a DP idiopática, a assinatura foi aproveitada para distribuição das 107 amostras de outras condições presentes no GSE6613 e GSE99039. Tais amostras eram de sangue periférico de pacientes com a doença de Alzheimer, a doença de Huntington, atrofia multissistêmica, paralisia supranuclear progressiva e 4 tipos da DP genética. A assinatura não foi capaz de distinguir robustamente as amostras de pacientes com a doença de Alzheimer, atrofia multissistêmica, paralisia supranuclear progressiva e DP genética com mutações no gene *PINK1* das amostras com a DP idiopática. Entretanto, nenhum destes apresentou média maior do que o grupo com pacientes com Parkinson idiopático, mantendo o modelo e a assinatura aptos a diagnosticar a DP idiopática. A assinatura teve capacidade de distinguir robustamente as amostras de pacientes com a doença de Huntington, DP genética com mutações nos genes *LRRK2* e (*PRKN*) e expressou tendência na DP genética com mutações no gene *ATP13A2* (p = 0,0505).

A grande problemática dessas análises é o pequeno tamanho amostral dos grupos com outras categorias. As análises foram realizadas com base nas diferenças entre médias (testes t), porém diversas amostras em grupos que não apresentaram diferenças estatísticas estão abaixo do

limiar de corte estabelecido para este modelo (0,7095) e, sendo assim, não seriam classificadas como DP idiopática. Sendo assim, são necessárias mais amostras para se ver uma distribuição mais "real" da população, como as de DP idiopática, mais uniformemente distribuídas. Entretanto, a seleção de preditores e a construção de modelos foi feita sobre essas amostras de DP idiopática e é esperado essa homogeneidade na classificação da classe.

Todas as outras categorias quadram a doenças neurodegenerativas. Excetuando-se a doença de Alzheimer, as outras condições abrangem aspectos motores muito similares a DP como anormalidades na marcha, instabilidade postural, rigidez e contração muscular (distonia). Esses aspectos geralmente são associados às disfunções na via nigro-estriatal causando neurodegeneração no estriado. Do ponto de vista fenotípico elas apresentam semelhança. Do ponto de vista neuropatológico, todas essas condições também exibem semelhanças, por serem doenças geradas por acumulação proteica e causarem morte neuronal. Todas são caracterizadas por disfunção mitocondrial, estresse oxidativo, apoptose, alteração da autofagia e neuroinflamação. Com base nisso é possível entender a dificuldade de classificação destes grupos e distinção da DP idiopática. Sendo que essas doenças compartilham características nas neuropatologias nas quais os genes da assinatura fazem parte, como modulação do sistema imune e apoptose, torna-se ainda mais essencial a aquisição de mais amostras para análises e ajustes dos modelos.

Ainda que hajam modelos com valores de sensibilidade e especificidade maiores do que os selecionados, como o *Random Forest* treinado com o GTr superamostragem que apresentou sensibilidade de 0,882 pré- e 0,920 pós-otimização e o *k-Nearest Neighbors* treinado com o GTr subamostragem que apresentou especificidade de 0,789 pré- e 0,904 pós-otimização, julgou-se melhor selecionar os modelos pela linha de corte. Como tais modelos ainda têm que ser testados em amostras de estudos independentes e em mais amostras de outras categorias, acredita-se que um limiar mais alto permitiria, a longo prazo, resultados melhores de predição por uma maior improbabilidade de se obter valores maiores de probabilidades de modo arbitrário.

Para testar um modelo que não perdesse informação na presença de preditores colineares, foi elaborada uma metodologia de classificação de utilização de redes ponderadas pelos 500 maiores valores de correlação entre os pares de genes selecionados como base para os valores de

expressão referidos a cores, gerando assim um *heatmap* de redes. Foram criadas imagens topográficas de cada amostra, utilizando o programa ViaComplex e estas imagens foram utilizadas como dados para elaboração de um modelo de classificação aplicando o algoritmo de *deep learning MXNET*. A predição resultou em uma AUC de 0,8628, uma sensibilidade de 0,8142 e uma especificidade de 0,8800. Apesar dos ótimos resultados para um modelo que não foi ajustado com diversas opções de hiperparâmetros ou sem as reparações do desbalanço amostral, conforme feito neste trabalho, há muitos pontos de arbitrariedade no procedimento.

Os dados foram reescalonados em um intervalo de 0 a 1 e, sendo assim, os valores não quadram aos valores "reais" obtidos do microarranjo, e sim uma forma de ranqueamento entre os valores de expressão de todas as amostras. Essa característica do procedimento se torna menos considerável em conjuntos de muitas amostras. Outros pontos de arbitrariedade foram os ponderamentos de redes. Foram combinados os valores de *p* pelo método de Fisher e os valores de χ^2 obtidos deste cálculo foram utilizados como ponderação de conectores. A combinação de valores de *p* muito baixos geram valores de χ^2 altos e a ideia utilizada foi de que os pares de genes cujos valores de *p* em vários conjuntos fossem baixos, teriam dinâmica trasncricional mais correlacionada que genes com valores de *p* altos. Esse fato foi confirmado por redes de co-expressão descritas obtidas do GeneMANIA. Na rede de GSP 84,6% dos 500 conectores com maior valor de χ^2 foram corroborados pelos dados de co-expressão gênica. Entretanto, na rede de GSN, apenas 49,6% dos 500 conectores foram corroborados, constituindo um fator de arbitrariedade.

Deve-se determinar como as redes serão ponderadas, quais os genes farão parte das redes e quantos conectores serão utilizados. Uma possibilidade clara seria utilizar as redes do GeneMANIA, utilizando unicamente genes co-expressos. É possível perceber a maior homogeneidade de transcricões na rede de GSP nas imagens obtidas com o ViaComplex, corroborando o valor correlação/co-expressão desta. É possível que uma eliminação dos conectores que não condizem a genes co-expressos descritos otimize as capacidades de classificação desses procedimentos. Essa metodologia pode auxiliar na classificação de condições em que as variações de expressão sejam muito sutis, como é o

caso do sangue periférico da DP, por poder valorizar genes agrupados na rede, intensificando a cor resultante.

De forma geral, este trabalho ratifica a relevância da meta-análise em estudos transcriptômicos para configuração de diferenças no perfil de expressão com uma maior segurança do que em estudos individuais e da elaboração de uma estrutura inclusiva e abrangente para realização de meta-análises de transcriptomas. Reforça-se a necessidade de mais estudos na área da DP utilizando as técnicas de alto rendimento, entretanto com uma maior descrição dos metadados das amostras. Por fim, destaca-se a relevância do trabalho em equipe de diversas áreas do conhecimento, onde os cientistas e profissionais da área da saúde devem atuar conjuntamente a matemáticos e cientistas da computação para elaboração de artifícios cada vez mais eficazes no processo da interpretação biológica.

A seleção dos preditores e dos modelos elaborando uma assinatura gênica e uma forma de utilização foram eficientes, entretanto não necessariamente tal assinatura é apontada como finalizada. É possível otimizar a assinatura com uma literatura enquadrando mais informações sobre os dados e com mais dados sobre outras doenças, podendo assim estratificar e categorizar as amostras. Dessa forma, a bioinformática pode auxiliar mais na pesquisa e, consequentemente, na qualidade de vida de pacientes da doença.

Apesar da variação amostral quanto aos tratamentos farmacológicos, idades e severidade da doença, esse trabalho pôde identificar uma assinatura e modelos confiáveis na classificação da DP idiopática em amostras de sangue e fornecer base bioinformática para suas otimizações.

7 PRINCIPAIS RESULTADOS

– Construiu-se um banco de dados contendo os valores de tamanho de efeito de quatro trabalhos independentes e da meta-análise de 17.712 genes calculados das diferenças entre pacientes com a DP idiopática e indivíduos sadios, em Hedges'g corrigido e em coeficientes de correlação produto-momento de Pearson.

– Foram identificados 200 genes que apresentam as principais diferenças de expressão entre amostras sanguíneas de pacientes com a DP idiopática e indivíduos sadios.

– Os genes selecionados estão relacionados ao sistema imunológico mediado por leucócitos, a poliubiquitinação de proteínas e a organização do citoesqueleto.

– Os genes selecionados não são capazes de separar as amostras de DP idiopática e indivíduos sadios quando avaliados individualmente por agrupamentos hierárquicos.

– Há limitada sobreposição de GDE colhidos de análises de expressão diferencial em estudos individuais.

– Elaborou-se uma assinatura de 58 genes para distinção dos grupos avaliados.

– Os genes da assinatura estão relacionados a resposta imune e ao processo apoptótico.

– Descreveu-se 30 modelos capazes de classificar os grupos avaliados.

– Selecionou-se e otimizou-se dois modelos criados de classificação, um favorecendo a sensibilidade e outro a especificidade que juntamente obtiveram 86% de acertos.

– O modelo de classificação para sensibilidade é capaz de diferenciar também os grupos de amostras da doença de Huntington, da DP genética por mutações nos genes *LRRK2* e *PRKN*.

– Os modelos selecionados foram obtidos de algoritmos *Gradient Boosting Machine* e *k-Nearest Neighbors*, apesar de alguns algoritmos apresentaram maiores valores absolutos de sensibilidade e especificidade, atingindo-se até 92% de sensibilidade e 90% de especificidade.

– Elaborou-se uma sequência de operações para classificações de imagens que busca valorizar dados colineares.

– O modelo de classificação baseada em imagens apresentou 84% de acertos.

8 CONCLUSÃO E PERSPECTIVAS

8.1 CONCLUSÃO

– Apesar da variação amostral quanto as características clínicas e demográficas dos pacientes, esse trabalho pôde determinar uma assinatura e diferentes modelos confiáveis na classificação da DP idiopática se baseando em perfis de expressão gênica em amostras de sangue periférico analisados por algoritmos de aprendizado de máquina.

8.2 PERSPECTIVAS

– Confirmar a variação de expressão de genes da assinatura em outras metodologias, como a RT-qPCR (reações de transcriptase reversa seguida por reação em cadeia da polimerase quantitativa).

– Para os modelos de classificação de imagens, determinar como as redes serão ponderadas, quais os genes farão parte das redes e quantos conectores serão utilizados.

– Testar sua eficácia de predição utilizando outros algoritmos de *deep learning*, hiperparâmetros e correção de desbalanço amostral.

– Testar cada modelo selecionado em amostras independentes de diversas categorias e dos grupos treinados.

REFERÊNCIAS

AERA. American Educational Research Association. In: Standards for reporting on empirical social science research in AERA publications, 2006.

AHLSKOG, J. E.; MUENTER, M. D. Frequency of levodopa–related dyskinesias and motor fluctuations as estimated from the cumulative literature. **Movement disorders: official journal of the Movement Disorder Society**, v. 16, n. 3, p. 448–458, 2001.

ALIEVA, A. K. et al. Involvement of endocytosis and alternative splicing in the formation of the pathological process in the early stages of Parkinson's disease. **BioMed Research International**, v. 2014, n. 1, p. 1–6, 2014.

ALLISON, D. B. et al. **DNA microarrays and related genomics techniques: Design, analysis, and interpretation of experiments**. Chapman & Hall/CRC, 2006.

ALTMAN, N. S. An introduction to kernel and nearest–neighbor nonparametric regression. **The American Statistician**, v. 46, n. 3, p. 175–185, 1992.

APA. American Psychological Association. In: Publication Manual of the American Psychological Association, 2001

ASHBURNER, M. et al. Gene Ontology: Tool for the unification of biology. **Nature Genetics**, v. 25, n. 1, p. 25–29, 2000.

BAI, Z. et al. Distinctive RNA expression profiles in blood associated with Alzheimer disease after accounting for white matter hyperintensities. **Alzheimer Disease and Associated Disorders**, v. 28, n. 3, p. 226–233, 2014.

BALDERESCHI, M. et al. Parkinson's disease and parkinsonism in a longitudinal study: Two–fold higher incidence in men. ILSA Working Group. Italian Longitudinal Study on Aging. **Neurology**, v. 55, n. 9, p. 1358–1363, 2000.

BARBOSA, M. T. et al. Parkinsonism and Parkinson's disease in the elderly: A community–based survey in Brazil (the Bambuí study). **Movement Disorders**, v. 21, n. 6, p. 800–808, 2006.

BEISSBARTH, T. et al. Processing and quality control of DNA array hybridization data. **Bioinformatics (Oxford, England)**, v. 16, n. 11, p. 1014–1022, 2000.

BIER, F. F. et al. DNA microarrays. **Advances in Biochemical Engineering/Biotechnology**, v. 109, n.1, p. 433–453, 2008.

BJORKLUND, G. et al. Metals and Parkinson's disease: Mechanisms and biochemical processes. **Current Medicinal Chemistry**, v. 25, n. 19, p. 2198–2214, 2018.

BOIVIN, D. B. et al. Circadian clock genes oscillate in human peripheral blood mononuclear cells. **Blood**, v. 102, n. 12, p. 4143–4145, 2003.

BRAZMA, A. et al. Minimum information about a microarray experiment (MIAME) – Toward standards for microarray data. **Nature Genetics**, v. 29, n. 4, p. 365–371, 2001.

BRAZMA, A. et al. ArrayExpress – A public repository for microarray gene expression data at the EBINucleic Acids Research. **Nucleic Acids Research**, v.31, n.1, p. 68–71, 2003.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123–140, 1996.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BRUNDIN, P. et al. Research in motion: The enigma of Parkinson's disease pathology spread. **Nature Reviews Neuroscience**, v. 9, n. 10, p. 741–745, 2008.

CALLIGARIS, R. et al. Blood transcriptomics of drug–naïve sporadic Parkinson's disease patients. **BMC Genomics**, v. 16, n. 1, p. 876–890, 2015.

CAPALDI, A. P. Analysis of gene function using DNA microarrays. In: **Methods in Enzymology**, v. 470, n.1, p. 3–17, 2010.

CARLSSON, A.; LINDQVIST, M.; MAGNUSSON, T. 3,4–Dihydroxyphenylalanine and 5–hydroxytryptophan as reserpine antagonists. **Nature**, v. 180, n. 4596, p. 1200–1200, 1957.

CASTRO, M. A. A. et al. ViaComplex: Software for landscape analysis of

gene expression networks in genomic context. **Bioinformatics**, v. 25, n. 11, p. 1468–1469, 2009.

CHARCOT, J. M. Oeuvres complètes (Tome 1). Leçons sur les maladies du système nerveux (Em francês). Bureaux du Progrès Médical, 1872.

CHATTERJEE, P. et al. Studying the system–level involvement of MicroRNAs in parkinson's disease. **PLoS ONE**, v. 9, n. 4, p. 1–15, 2014.

CHEN, T. et al. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems, 2015.

CHEN, T.; GUESTRIN, C. XGBoost. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16**. ACM Press, 2016

CHEN, Y. J. et al. Normalization methods for analysis of microarray gene–expression data. **Journal of Biopharmaceutical Statistics**, v. 13, n. 1, p. 57–74, 2003.

CHEN, Z. et al. Epigenetic regulation: A new frontier for biomedical engineers. **Annual Review of Biomedical Engineering**, v. 19, n. 1, p. 195–219, 2017.

CHIKINA, M. D. et al. Low–variance RNAs identify Parkinson's disease molecular signature in blood. **Movement Disorders**, v. 30, n. 6, p. 813–821, 2015.

CHILLAG–TALMOR, O. et al. Use of a refined drug tracer algorithm to estimate prevalence and incidence of Parkinson's disease in a large israeli population. **Journal of Parkinson's disease**, v. 1, n. 1, p. 35–47, 2011.

CHOI, J. K. et al. Combining multiple microarray studies and modeling interstudy variation. **Bioinformatics**, v. 19, n. 1, p. 84–90, 2003.

CHU, Y.; KORDOWER, J. H. Age–associated increases of α–synuclein in monkeys and humans are associated with nigrostriatal dopamine depletion: Is this the target for Parkinson's disease? **Neurobiology of Disease**, v. 25, n. 1, p. 134–149, 2007.

CICCHETTI, F.; DROUIN–OUELLET, J.; GROSS, R. E. Environmental toxins and Parkinson's disease: What have we learned from pesticide–induced animal models? **Trends in Pharmacological Sciences**, v. 30, n. 9,

p. 475–483, 2009.

COHEN, J. The statistical power of abnormal–social psychological research: A review. **The Journal of Abnormal and Social Psychology**, v. 65, n. 3, p. 145–153, 1962.

COHEN, J. The cost of dichotomization. **Applied Psychological Measurement**, v. 7, n. 3, p. 249–253, 1983.

CONNOLLY, B. S.; LANG, A. E. Pharmacological treatment of Parkinson disease. **JAMA**, v. 311, n. 16, p. 1670–1683, 2014.

CORTES, C.; VAPNIK, V. Support–vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995.

COVER, T. M.; HART, P. E. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

DAVIS, S.; MELTZER, P. S. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. **Bioinformatics**, v. 23, n. 14, p. 1846–1847, 2007.

DE BARROS, R. H.; JUNIOR, E. DE P. G. Por uma história do velho ou do envelhecimento no Brasil. **CES Revista**, v. 27, n. 1, p. 75–92, 2013.

DENNIS, G. et al. DAVID: Database for annotation, visualization, and integrated discovery. **Genome biology**, v. 4, n. 5, p. 1–11, 2003.

DO, J. H.; CHOI, D.–K. Normalization of microarray data: Single–labeled and dual–labeled arrays. **Molecules and Cells**, v. 22, n. 3, p. 254–261, 2006.

DYKSTRA–AIELLO, C. et al. Altered expression of long noncoding RNAs in blood after ischemic stroke and proximity to putative stroke risk loci. **Stroke**, v. 47, n. 12, p. 2896–2903, 2016.

ECKER, S. et al. Epigenetic and transcriptional variability shape phenotypic plasticity. **BioEssays**, v. 40, n. 2, p. 1–11, 2018.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. **Nucleic acids research**, v. 30, n. 1, p. 207–210, 2002.

ELLIS, P. D. **The essential guide to effect sizes**. Cambridge University

Press, 2010.

EMMANOUILIDOU, E.; STEFANIS, L.; VEKRELLIS, K. Cell–produced α–synuclein oligomers are targeted to, and impair, the 26S proteasome. **Neurobiology of Aging**, v. 31, n. 6, p. 953–968, 2010.

EZKURDIA, I. et al. Multiple evidence strands suggest that there may be as few as 19.000 human protein–coding genes. **Human Molecular Genetics**, v. 23, n. 22, p. 5866–5878, 2014.

FENG, Y.; JANKOVIC, J.; WU, Y.–C. Epigenetic mechanisms in Parkinson's disease. **Journal of the Neurological Sciences**, v. 349, n. 1, p. 3–9, 2015.

FENG, Y.; WANG, X. Systematic analysis of microarray datasets to identify Parkinson's disease–associated pathways and genes. **Molecular Medicine Reports**, v. 15, n. 3, p. 1252–1262, 2017.

FERESHTEHNEJAD, S.–M. et al. New clinical subtypes of Parkinson disease and their longitudinal progression. **JAMA Neurology**, v. 72, n. 8, p. 863–873, 2015.

FISHER, R. Statistical methods for research workers. Oliver and Boyd, 1932

FLINT, A. The skin in Parkinson's disease. **Primary care**, v. 4, n. 3, p. 475–480, 1977.

FREUDENBERG, J. M. Comparison of background correction and normalization procedures for high–density oligonucleotide microarrays. Universitat Leipzig, 2005.

GAUTIER, L. et al. affy – Analysis of Affymetrix GeneChip data at the probe level. **Bioinformatics**, v. 20, n. 3, p. 307–315, 2004.

GIROUX, M. L. Parkinson disease: Managing a complex, progressive disease at all stages. **Cleveland Clinic journal of medicine**, v. 74, n. 5, p. 313–322, 2007.

GLASS, C. K. et al. Mechanisms underlying inflammation in neurodegeneration. **Cell**, v. 140, n. 6, p. 918–934, 2010.

GLASS, G. V; MCGAW, B.; SMITH, M. L. **Meta–analysis in social research**. Sage Publications, 1981.

GOLUB, T. R. et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. **Science**, v. 286, n. 5439, p. 531–537, 1999.

GORELL, J. M. et al. Multiple risk factors for Parkinson's disease. **Journal of the neurological sciences**, v. 217, n. 2, p. 169–174, 2004.

GUI, X. et al. Improved statistical tests for differential gene expression by shrinking variance components estimates. **Biostatistics**, v. 6, n. 1, p. 59–75, 2005.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Machine Learning**, v. 46, n. 1, p. 389–422, 2002.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. **Journal of Machine Learning Research**, v. 3, n. 3, p. 1157–1182, 2003.

HARELL JR., F. E. et al. Hmisc: Harrell Miscellaneous. R package version 4.1–1, 2018.

HARRINGTON, M. G. et al. Prostaglandin D synthase isoforms from cerebrospinal fluid vary with brain pathology. **Disease Markers**, v. 22, n. 1, p. 73–81, 2006.

HARTMANN, A. Postmortem studies in Parkinson's disease. **Dialogues in Clinical Neuroscience**, v. 6, n. 3, p. 281–293, 2004.

HEDGES, L. V. Distribution theory for Glass's estimator of effect size and related estimators. **Journal of Educational Statistics**, v. 6, n. 2, p. 107–128, 1981.

HEDGES, L. V.; OLKIN, I. **Statistical methods for meta–analysis**. Academic Press, 1985.

HOEHN, M. M.; YAHR, M. D. Parkinsonism: Onset, progression, and mortality. **Neurology**, v. 17, n. 5, p. 427–442, 1967.

HUBER, W. et al. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. **Bioinformatics**, v. 18, n. 1. p. 96–104, 2002.

HUGHES, A. J. et al. A clinicopathologic study of 100 cases of Parkinson's disease. **Archives of neurology**, v. 50, n. 2, p. 140–148, 1993.

IBGE. Instituto Brasileiro de Geografia e Estatística (IBGE). In: Indicadores sociais, Censo Demográfico, 2010.

IBGE. Instituto Brasileiro de Geografia e Estatística (IBGE). In: Projeção da população do Brasil e de Unidades da Federação, 2018.

IBGE. Instituto Brasileiro de Geografia e Estatística (IBGE). In: Síntese de indicadores sociais, 2012.

INFANTE, J. et al. Comparative blood transcriptome analysis in idiopathic and LRRK2 G2019S – Associated Parkinson's disease. **Neurobiology of Aging**, v. 38, n. 1, p. 1–5, 2016.

IRIZARRY, R. A. et al. Summaries of Affymetrix GeneChip probe level data. **Nucleic acids research**, v. 31, n. 4, p. 1–8, 2003.

IRIZARRY, R. A. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. **Biostatistics**, v. 4, n. 2, p. 249–264, 2003.

JAMES, G. et al. **An introduction to statistical learning**. Springer New York, 2013.

JANKOVIC, J. Parkinson's disease: Clinical features and diagnosis. **Journal of Neurology, Neurosurgery & Psychiatry**, v. 79, n. 4, p. 368–376, 2008.

JASINSKA–MYGA, B. et al. Depression in Parkinson's disease. **The Canadian journal of neurological sciences**, v. 37, n. 1, p. 61–66, 2010.

JASKOWIAK, P. A.; COSTA, I. G.; CAMPELLO, R. J. G. B. Clustering of RNA–Seq samples: Comparison study on cancer data. **Methods**, v. 132, n. 1, p. 42–49, 2018.

KASSAHUN, Y. et al. Surgical robotics beyond enhanced dexterity instrumentation: A survey of machine learning techniques and their role in intelligent and autonomous surgical actions. **International Journal of Computer Assisted Radiology and Surgery**, v. 11, n. 4, p. 553–568, 2016

KAUFFMANN, A.; GENTLEMAN, R.; HUBER, W. arrayQualityMetrics – A Bioconductor package for quality assessment of microarray data.

Bioinformatics, v. 25, n. 3, p. 415–416, 2009.

KAUSHIK, S.; CUERVO, A. M. Proteostasis and aging. **Nature Medicine**, v. 21, n. 12, p. 1406–1415, 2015.

KIRK, R. E. The importance of effect magnitude. In: DAVIS, S. F. **Handbook of research methods in experimental psychology**. Blackwell Pub, 2003. p. 83–105.

KLEGERIS, A. et al. α–Synuclein activates stress signaling protein kinases in THP–1 cells and microglia. **Neurobiology of Aging**, v. 29, n. 5, p. 739–752, 2008.

KLEIN, C.; WESTENBERGER, A. Genetics of Parkinson's disease. **Cold Spring Harbor Perspectives in Medicine**, v. 2, n. 1, p. 1–15, 2012.

KOUROU, K. et al. Machine learning applications in cancer prognosis and prediction. **Computational and Structural Biotechnology Journal**, v. 13, n. 1, p. 8–17, 2015.

KUHN, M. et al. caret: Classification and Regression Training. R package version 6.0–80, 2018

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. Springer New York, 2013.

LANTZ, B. **Machine learning with R**. Packt, 2015.

LEE MOSLEY, R. et al. Neuroinflammation, oxidative stress, and the pathogenesis of Parkinson's disease. **Clinical Neuroscience Research**, v. 6, n. 5, p. 261–281, 2006.

LEE, P. C. et al. Traumatic brain injury, paraquat exposure, and their relationship to Parkinson disease. **Neurology**, v. 79, n. 20, p. 2061–2066, 2012.

LEONARDSON, A. S. et al. The effect of food intake on gene expression in human peripheral blood. **Human Molecular Genetics**, v. 19, n. 1, p. 159–169, 2009.

LESAGE, S.; BRICE, A. Parkinson's disease: From monogenic forms to genetic susceptibility factors. **Human Molecular Genetics**, v.18, n. 1, p 48–59, 2009.

LEWY, F. Zur pathologischen Anatomie der Paralysis agitans (Em alemão). **Dtsch. Z. Nervenheilk**, n. 50, v.1, p. 50–55, 1913.

LI, S.; BECICH, M. J.; GILBERTSON, J. Microarray data mining using gene ontology. **Studies in health technology and informatics**, v. 107, n. 2, p. 778–782, 2004.

LILL, C. M. Genetics of Parkinson's disease. **Molecular and Cellular Probes**, v. 30, n. 6, p. 386–396, 2016.

LIPPMANN, R. An introduction to computing with neural nets. **IEEE ASSP Magazine**, v. 4, n. 2, p. 4–22, 1987.

LIPSHUTZ, R. J. et al. High density synthetic oligonucleotide arrays. **Nature Genetics**, v. 21, n. 1, p. 20–24, 1999.

LOANE, C.; POLITIS, M. Positron emission tomography neuroimaging in Parkinson's disease. **American Journal of Translational Research**, v. 3, n. 4, p. 323–341, 2011.

LOCASCIO, J. J. et al. Association between α–synuclein blood transcripts and early, neuroimaging–supported Parkinson's disease. **Brain**, v. 138, n. 9, p. 2659–2671, 2015.

LUNARDON, N.; MENARDI, G.; TORELLI, N. ROSE: A Package for binary imbalanced learning. **The R Journal**, v. 6, n. 1, p. 79–89, 2014.

MAGLOTT, D. et al. Entrez Gene: Gene–centered information at NCBI. **Nucleic Acids Research**, v. 39, n. 1, p. 52–57, 2011.

MANDEL, S. A. et al. Biomarkers for prediction and targeted prevention of Alzheimer's and Parkinson's diseases: Evaluation of drug clinical efficacy. **EPMA Journal**, v. 1, n. 2, p. 273–292, 2010.

MARSH, L. Depression and Parkinson's disease: Current knowledge. **Current Neurology and Neuroscience Reports**, v. 13, n. 12, p. 1–17, 2013.

MARSIT, C. J. Influence of environmental exposure on human epigenetic regulation. **Journal of Experimental Biology**, v. 218, n. 1, p. 71–79, 2015.

MARTINEZ–JIMENEZ, C. P. et al. Aging increases cell–to–cell transcriptional variability upon immune stimulation. **Science**, v. 355, n. 6332,

p. 1433–1436, 2017.

MARTINS, M. et al. Convergence of mirna expression profiling, α–synuclein interacton and GWAS in Parkinson's disease. **PLoS ONE**, v. 6, n. 10, p. 1–11, 2011.

MASON, L. et al. Boosting algorithms as gradient descent. **In Advances in Neural Information Processing Systems 12**, v. 1, n. 1, p. 512–518, 2000.

MCGEER, P. L. et al. Reactive microglia are positive for HLA–DR in the substantia nigra of Parkinson's and Alzheimer's disease brains. **Neurology**, v. 38, n. 8, p. 1285–1291, 1988.

MCGEER, P. L.; MCGEER, E. G. **Inflammation and the degenerative diseases of aging**. Annals of the New York Academy of Sciences, v. 1035, n. 1, p. 104–116, 2004

MENDENHALL, WILLIAM M., SINCICH, TERRY L., S. BOUDREAU, N. **Statistics for engineering and the sciences**. Pearson Prentice–Hall, 2016.

MILLER, R. M.; FEDEROFF, H. J. Microarrays in Parkinson's disease: A systematic approach. **NeuroRX**, v. 3, n. 3, p. 319–326, 2006.

MULLAINATHAN, S.; SPIESS, J. Machine learning: An applied econometric approach. **Journal of Economic Perspectives**, v. 31, n. 2, p. 87–106, 2017.

MURTY, M. N.; DEVI, V. S. **Pattern recognition: An algorithmic approach (undergraduate topics in computer science)**. Springer, 2012.

MUTEZ, E. et al. Transcriptional profile of Parkinson blood mononuclear cells with LRRK2 mutation. **Neurobiology of Aging**, v. 32, n. 10, p. 1839–1848, 2011.

NAJAFABADI, M. M. et al. Deep learning applications and challenges in big data analytics. **Journal of Big Data**, v. 2, n. 1, p. 1–21, 2015.

NALLS, M. A. et al. Large–scale meta–analysis of genome–wide association data identifies six new risk loci for Parkinson's disease. **Nature Genetics**, v. 46, n. 9, p. 989–993, 2014.

NIELSEN, D. Tree boosting with XGBoost: Why does XGBoost win "every" machine learning competition? **NTNU Tech Report**, v. 1, n. 12, p. 1–110,

2016.

O'CONNELL, M. Differential expression, class discovery and class prediction using S–PLUS and S+ArrayAnalyzer. **ACM SIGKDD Explorations Newsletter**, v. 5, n. 2, p. 38–47, 2003.

ONU. Organização das Nações Unidas (ONU). In: Países dos BRICS terão 940 milhões de idosos até 2050, 2017.

OUCHI, Y. et al. Microglial activation and dopamine terminal loss in early Parkinson's disease. **Annals of Neurology**, v. 57, n. 2, p. 168–175, 2005.

PAN, W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. **Bioinformatics**, v. 18, n. 4, p. 546–554, 2002.

PARK, A.; STACY, M. Disease–modifying drugs in Parkinson's disease. **Drugs**, v. 75, n. 18, p. 2065–2071, 2015.

PARKINSON, J. An essay on the shaking palsy. Sherwood, Neely and Jones, 1817.

PATTERSON, J.; GIBSON, A. **Deep learning : A practitioner's approach**. O'Reilly Media, 2017.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science**, v. 2, n. 11, p. 559–572, 1901.

PINZAR, E. et al. Prostaglandin D synthase gene is involved in the regulation of non–rapid eye movement sleep. **Proceedings of the National Academy of Sciences**, v. 97, n. 9, p. 4903–4907, 2000.

POEWE, W. et al. Parkinson disease. **Nature Reviews Disease Primers**, v. 3, n. 1, p. 1–21, 2017.

POSTUMA, R. B.; BERG, D. Advances in markers of prodromal Parkinson disease. **Nature Reviews Neurology**, v. 12, n. 11, p. 622–634, 2016.

PRINGSHEIM, T. et al. The prevalence of Parkinson's disease: A systematic review and meta–analysis. **Movement Disorders**, v. 29, n. 13, p. 1583–1590, 2014.

PRZEDBORSKI, S. The two–century journey of Parkinson disease research. **Nature Reviews Neuroscience**, v. 18, n. 4, p. 251–259, 2017.

RAMASAMY, A. et al. Key issues in conducting a meta–analysis of gene expression microarray datasets. **PLoS Medicine**, v. 5, n. 9, p. 1320–1332, 2008.

RHODES, D. R. et al. Meta–analysis of microarrays: Interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. **Cancer Research**, v. 62, n. 15, p. 4427–4433, 2002.

RHODES, D. R. et al. Large–scale meta–analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. **Proceedings of the National Academy of Sciences**, v. 101, n. 25, p. 9309–9314, 2004.

RILEY, B. E. et al. Systems–based analyses of brain regions functionally impacted in Parkinson's disease reveals underlying causal mechanisms. **PLoS ONE**, v. 9, n. 8, p. 1–14, 2014.

RITCHIE, M. E. et al. A comparison of background correction methods for two–colour microarrays. **Bioinformatics**, v. 23, n. 20, p. 2700–2707, 2007.

RITCHIE, M. E. et al. limma powers differential expression analyses for RNA–sequencing and microarray studies. **Nucleic Acids Research**, v. 43, n. 7, p. 1–13, 2015.

ROSENTHAL, R.; ROSNOW, R. L. Applying Hamlet's question to the ethical conduct of research: A conceptual addendum. **American Psychologist**, v. 39, n. 5, p. 561–563, 1984.

SAMPAIO, T. B. et al. The relevance of intranasal route in Parkinson's disease: From physiopathological alterations to administration of neurotoxins. **Clinical Pharmacology and Translational Medicine**, v. 1, n. 2, p. 20–37, 2017.

SÁNCHEZ, A.; VILLA, M. C. R. DE. A tutorial review of microarray data analysis. **Bioinformatics**, v. 1, n. 1, p. 1–55, 2008.

SANTIAGO, J. A.; LITTLEFIELD, A. M.; POTASHKIN, J. A. Integrative transcriptomic meta–analysis of Parkinson's disease and depression identifies NAMPT as a potential blood biomarker for de novo Parkinson's

disease. **Scientific Reports**, v. 6, n. 1, p. 1–10, 2016.

SANTIAGO, J. A.; POTASHKIN, J. A. Blood transcriptomic meta–analysis identifies dysregulation of hemoglobin and iron metabolism in Parkinson' disease. **Frontiers in Aging Neuroscience**, v. 9, n. 1, p. 1–8, 2017.

SAVICA, R. et al. Incidence and pathology of synucleinopathies and tauopathies related to parkinsonism. **JAMA Neurology**, v. 70, n. 7, p. 859–866, 2013.

SCHAPIRE, R. E. The boosting approach to machine learning: An overview. **MSRI Workshop on Nonlinear Estimation and Classification**. v. 171, n. 1, p. 149–171, 2003.

SCHERZER, C. R. et al. Molecular markers of early Parkinson's disease based on gene expression in blood. **Proceedings of the National Academy of Sciences**, v. 104, n. 3, p. 955–960, 2007.

SEN, G. C. Viruses and interferons. **Annual Review of Microbiology**, v. 55, n. 1, p. 255–281, 2001.

SHAMIR, R. et al. Analysis of blood–based gene expression in idiopathic Parkinson disease. **Neurology**, v. 89, n. 16, p. 1676–1683, 2017.

SHARMA, S. et al. Biomarkers in Parkinson's disease (recent update). **Neurochemistry International**, v. 63, n. 3, p. 201–229, 2013.

SHI, W.; OSHLACK, A.; SMYTH, G. K. Optimizing the noise versus bias trade–off for Illumina whole genome expression BeadChips. **Nucleic Acids Research**, v. 38, n. 22, p. 1–11, 2010.

SIMON, R. Analysis of DNA microarray expression data. **Best Practice & Research Clinical Haematology**, v. 22, n. 2, p. 271–282, 2009.

SOLDNER, F. et al. Parkinson–associated risk variant in distal enhancer of α–synuclein modulates target gene expression. **Nature**, v. 533, n. 7601, p. 95–99, 2016.

SOREQ, L. et al. Advanced microarray analysis highlights modified neuro–immune signaling in nucleated blood cells from Parkinson's disease patients. **Journal of Neuroimmunology**, v. 201, n. 1, p. 227–236, 2008.

SOREQ, L. et al. Exon arrays reveal alternative splicing aberrations in

Parkinson's disease leukocytes. **Neurodegenerative Diseases**, v. 10, n. 1, p. 203–206, 2012.

STAFFORD, P. **Methods in microarray normalization**. CRC Press, 2008.

STAYTE, S.; VISSEL, B. Corrigendum: Advances in non–dopaminergic pharmacological treatments of Parkinson's disease. **Frontiers in Neuroscience**, v. 8, n. 1, p. 1–2, 2014.

STOLL, L. et al. Differential regulation of ionotropic glutamate receptors. **Biophysical Journal**, v. 92, n. 4, p. 1343–1349, 2007.

STROBL, C.; MALLEY, J.; TUTZ, G. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. **Psychological Methods**, v. 14, n. 4, p. 323–348, 2009.

SU, X. et al. Synuclein activates microglia in a model of Parkinson's disease. **Neurobiology of Aging**, v. 29, n. 11, p. 1690–1701, 2008.

SUBRAMANIAM, S. R.; CHESSELET, M.–F. Mitochondrial dysfunction and oxidative stress in Parkinson's disease. **Progress in Neurobiology**, v. 106, n. 1, p. 17–32, 2013.

SUBRAMANIAN, A. et al. Gene set enrichment analysis: A knowledge–based approach for interpreting genome–wide expression profiles. **Proceedings of the National Academy of Sciences**, v. 102, n. 43, p. 15545–15550, 2005.

SULLIVAN, G. M.; FEINN, R. Using effect size – Or why the p value is not enough. **Journal of Graduate Medical Education**, v. 4, n. 3, p. 279–282, 2012.

SWEENEY, T. E. et al. Methods to increase reproducibility in differential gene expression via meta–analysis. **Nucleic Acids Research**, v. 45, n. 1, p. 1–14, 2017.

TEIJARO, J. R. Type I interferons in viral control and immune regulation. **Current Opinion in Virology**, v. 16, n. 1, p. 31–40, 2016.

TONG, T.; WANG, Y. Optimal shrinkage estimation of variances with applications to microarray data analysis. **Journal of the American Statistical Association**, v. 102, n. 477, p. 113–122, 2007.

TORGERSON, W. S. **Theory and methods of scaling**. R.E. Krieger Pub. Co, 1958.

TRÉTIAKOFF, C. Contribution a l'etude de l'anatomie pathologique du locus niger de soemmering avec quelques déductions relatives à la pathogénie des troubles du tonus musculaire et de la maladie de Parkinson (em frânces). Université de Paris, 1919.

TSENG, G. C.; GHOSH, D.; FEINGOLD, E. Comprehensive literature review and statistical considerations for microarray meta–analysis. **Nucleic Acids Research**, v. 40, n. 9, p. 3785–3799, 2012.

TUSHER, V. G.; TIBSHIRANI, R.; CHU, G. Significance analysis of microarrays applied to the ionizing radiation response. **Proceedings of the National Academy of Sciences**, v. 98, n. 9, p. 5116–5121, 2001.

TWELVES, D.; PERKINS, K. S. M.; COUNSELL, C. Systematic review of incidence studies of Parkinson's disease. **Movement Disorders**, v. 18, n. 1, p. 19–31, 2003.

UNWIN, N.; ALBERTI, K. G. M. M. Chronic non–communicable diseases. **Annals of Tropical Medicine & Parasitology**, v. 100, n. 5–6, p. 455–464, 2006.

VAN'T VEER, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. **Nature**, v. 415, n. 6871, p. 530–536, 2002.

VAN DEN EEDEN, S. K. et al. Incidence of Parkinson's disease: Variation by age, gender, and race/ethnicity. **American journal of epidemiology**, v. 157, n. 11, p. 1015–1022, 2003.

VEKRELLIS, K. et al. Pathological roles of α–synuclein in neurological disorders. **The Lancet Neurology**, v. 10, n. 11, p. 1015–1025, 2011.

VREELING, F. W. et al. Primitive reflexes in Parkinson's disease. **Journal of Neurology, Neurosurgery & Psychiatry**, v. 56, n. 12, p. 1323–1326, 1993.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, n. 301, p. 236–244, 1963.

WEBB, G. I.; BOUGHTON, J. R.; WANG, Z. Not so Naive Bayes: Aggregating one–dependence estimators. **Machine Learning**, v. 58, n. 1, p. 5–24, 2005.

WEI, T.; SIMKO, V. R package "corrplot": Visualization of a Correlation Matrix (Version 0.84), 2017.

WEINSHEIMER, S. M. et al. Gene expression profiling of blood in brain arteriovenous malformation patients. **Translational Stroke Research**, v. 2, n. 4, p. 575–587, 2011.

WHITWORTH, G. B. **An introduction to microarray data analysis and visualization**. Elsevier Inc., 2010.

WU, H. et al. MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. In: **The Analysis of Gene Expression Data**. Springer, New York , p. 313–341, 2003.

XILOURI, M. et al. Abberant α–synuclein confers toxicity to neurons in part through inhibition of chaperone–mediated autophagy. **PLoS ONE**, v. 4, n. 5, p. 1–15, 2009.

XILOURI, M.; BREKK, O. R.; STEFANIS, L. Alpha–synuclein and protein degradation systems: A reciprocal relationship. **Molecular Neurobiology**, v. 47, n. 2, p. 537–551, 2013.

YANG, Y. H. et al. Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. In: **Selected Works of Terry Speed**. Oxford University Press, p. 591–600, 2012.

YOUNG, M. D. et al. Gene ontology analysis for RNA–seq: Accounting for selection bias. **Genome Biology**, v. 11, n. 2, p. 1–12, 2010.

ZHANG, W. et al. Aggregated α–synuclein activates microglia: A process leading to disease progression in Parkinson's disease. **The FASEB Journal**, v. 19, n. 6, p. 533–542, 2005.

ZHAO, Y. J. et al. Progression of Parkinson's disease as evaluated by Hoehn and Yahr stage transition times. **Movement Disorders**, v. 25, n. 6, p. 710–716, 2010.

APÊNDICES

Os Apêndices se encontram na página: bit.ly/2I7xDxJ. Caso haja necessidade, contatar via endereço eletrônico: userfalchetti@gmail.com.