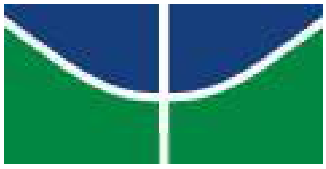


Universidade de Brasília

Instituto de Ciências Exatas - IE

Departamento de Estatística - EST

Análise de Expressões Gênicas em Síndrome de Sjögren

Bárbara Jacqueline Cruvinel Matos

16/0002796

Trabalho de Conclusão de Curso

Orientadora: Prof. Dra. Joanlise Marco de Leon Andrade

Brasília

2020

Bárbara Jacqueline Cruvinel Matos

Análise de Expressões Gênicas em Síndrome de Sjögren

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários à obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dra. Joanlise Marco de Leon Andrade

Universidade de Brasília – UnB

Instituto de Matemática

Departamento de Estatística

Brasília

14 de Dezembro de 2020

Bárbara Jacqueline Cruvinel Matos
Análise de Expressões Gênicas em Síndrome de Sjögren/ Bárbara Jacqueline Cruvinel Matos. – Brasília, 14 de Dezembro de 2020-
81 p. : il. (algumas cores.) ; 30 cm.

Orientador: Prof. Dra. Joanlise Marco de Leon Andrade

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Instituto de Matemática
Departamento de Estatística, 14 de Dezembro de 2020.

1. Análise de Expressões Gênicas. 2. Síndrome de Sjögren. I. Orientador. II. Universidade de Brasília. III. Faculdade de Ciências Exatas. IV. Departamento de Estatística.

CDU 02:141:005.7

Bárbara Jacqueline Cruvinel Matos

Análise de Expressões Gênicas em Síndrome de Sjögren

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários à obtenção do título de Bacharel em Estatística.

Trabalho aprovado. Brasília, 14 de Dezembro de 2020:

Prof. Dra. Joanlise Marco de Leon Andrade

Prof. Dr. George Freitas von Borries

Prof. Dra. Thais Carvalho Valadares Rodrigues

Brasília
14 de Dezembro de 2020

Agradecimentos

Agradeço a Deus, primeiramente. Sem Ele nada em minha vida seria possível. Agradeço também à minha família, em especial a minha mãe que sempre me apoiou em tudo que eu precisava durante a minha vida e sempre esteve disposta a me ajudar em tudo o que foi preciso.

Eternos agradecimentos à Universidade de Brasília que abriu tantas portas e me ensinou tanto, em especial a minha professora e orientadora, Dra. Joanlise Marco de Leon Andrade, pelas diretrizes, disponibilidade, ensinamentos, compreensão, incentivo, paciência, além de seu grande desprendimento em ajudar-me sempre, viabilizar este projeto e acreditar em minha capacidade. Dr. George Freitas Von Borries, agradeço pelo apoio, ajudas, incentivos e inúmeras oportunidades de conhecimento.

Resumo

Neste estudo foram analisados dados de expressão gênica de um grupo de casos de Síndrome de Sjögren Primária e um grupo controle, de indivíduos saudáveis. O principal objetivo foi a realização de análises de perfil de expressão gênica, com base na identificação de genes diferencialmente expressos entre os casos e controles. Três abordagens foram utilizadas para esse fim: testes de comparação de médias, testes não paraméticos e o algorítmo PPCLUST (Particionamento com uma robusta medida de distância para dados com altas dimensões e um tamanho amostral baixo). Análises de agrupamento por técnicas não supervisionadas dos transcritos selecionados foram realizadas para a visualização dos perfis de expressão. Destaca-se a superexpressão de genes ligados ou induzidos por interferons. Também foram identificados sistemas biológicos por Anotação Genômica (*Annotation*), a maior parte dos quais também associados a interferons. Os resultados obtidos são consistentes com os descritos em outros estudos de Síndrome de Sjögren.

Palavras-chaves: Análise de Perfil de Expressão Gênica. Síndrome de Sjögren. Interferons. Análise de Agrupamento. Análise de Cluster. Correção por testes múltiplos. FDR. PPCLUST.

Abstract

In this study, gene expression data from Primary Sjögren's Syndrome cases and a group of healthy controls were analyzed. The main objective was to perform gene expression profiling on the identified differentially expressed genes between cases and controls. Three different approaches were employed for this purpose: parametric tests, non-parametric tests and PPCLUST algorithm (Partition clustering of high dimensional low sample size data based on p-values). Unsupervised Cluster Analyses were performed for profiling visualization. Several interferon genes and interferon inducible genes were identified. Important biological pathways were identified by Annotation, most of which were related to interferon pathways. Results were consistent with other Sjögren's Syndrome genetic studies.

Key-words: Gene expression profiling. Sjögren's Syndrome. Interferon genes. Cluster analysis. Multiple Testing Correction. FDR. PPCLUST.

Lista de ilustrações

Figura 1 – Resumo esquemático dos mecanismos genéticos e epigenéticos associados com a suscetibilidade à SSp 21
Figura 2 – Resumo esquemático do processo de obtenção das intensidades de expressão gênica 27
Figura 3 – Processo da síntese proteica (ou expressão gênica) 28
Figura 4 – Ilustração de agrupamento hierárquico aglomerativo e divisivo. 36
Figura 5 – Exemplo de *Heatmap* de perfil de expressão gênica. 37
Figura 6 – Distribuição de expressões gênicas de casos e controles em escala original. 41
Figura 7 – Distribuição de expressões gênicas de casos e controles em escala log2. . 42
Figura 8 – Distribuição das expressões médias de casos e controles em escala original. 42
Figura 9 – Distribuição das expressões médias de casos e controles em escala log2. 43
Figura 10 – Distribuição de FC por médias e de FC por medianas. 43
Figura 11 – Distribuição dos valores utilizados para o PPCLUST. 44
Figura 12 – Distribuição de P-valores do Testes de normalidade Shapiro-Wilk. . . . 44
Figura 13 – Número de grupos definidos pelo algorítimo PPCLUST por P-valor. . . 45
Figura 14 – Estrutura dos grupos obtidos pelo algorítmo PPCLUST. 45
Figura 15 – Diagrama de Venn das 3 listas de transcritos. 47
Figura 16 – *Heatmap* das expressões gênicas em escala log2 para a Lista L1 48
Figura 17 – *Heatmap* das expressões gênicas em escala log2 para a Lista L2 48
Figura 18 – *Heatmap* das expressões gênicas em escala log2 para a Lista L3 49
Figura 19 – *Heatmap* das expressões gênicas em escala log2 para a Lista L4 49
Figura 20 – *Heatmap* das expressões gênicas em escala log2 para a Lista L5 50
Figura 21 – Distribuição do número de genes da L1 detectados nos sistemas biológicos mais significantes 52
Figura 22 – Distribuição do número de genes da L2 detectados nos sistemas biológicos mais significantes 53
Figura 23 – Distribuição do número de genes da L3 detectados nos sistemas biológicos mais significantes 53
Figura 24 – Distribuição do número de genes da L4 detectados nos sistemas biológicos mais significantes 54
Figura 25 – Distribuição do número de genes da L5 detectados nos sistemas biológicos mais significantes 54
Figura 26 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista L1 56
Figura 27 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista L2 56

Figura 28 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista L3 57
Figura 29 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista L4 57
Figura 30 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista L5 58
Figura 31 – Diagrama das ligações entre os sistemas biológicos identificados na Lista L1 59
Figura 32 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L2. 59
Figura 33 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L3. 60
Figura 34 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L4. 60
Figura 35 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L5. 61
Figura 36 – Gráfico de redes para os genes e seus sistemas biológicos relacionados identificados na Lista L1 62
Figura 37 – Cnetplot - WILCOX. 62
Figura 38 – Cnetplot - PPCLUST. 63
Figura 39 – Cnetplot - interseção dos transcritos entre PPCLUST, Teste T e Wilcoxon. 63
Figura 40 – Cnetplot - União dos transcritos entre PPCLUST, Teste T e Wilcoxon. 64
Figura 41 – Ilustração da ação de interferons. 66

Lista de tabelas

Tabela 1 – Tabela do método FDR 34
Tabela 2 – Número de transcritos selecionados para cada etapa das abordagens de identificação de TDEs. 46
Tabela 3 – Genes mais sub/super-expressos na **L5** 51
Tabela 4 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 1 76
Tabela 5 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 2 77
Tabela 6 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 3 78

Tabela 7 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 4 79
Tabela 8 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 5 80

Lista de abreviaturas e siglas

AECG	Grupo de Consenso Americano-Europeu, do inglês *American-European Consensus Group*;
ANA	Anticorpos Antinucleares;
ANTI-RO	Anticorpos contra o antígeno Ro (proteína citoplasmática);
CBC	Hemograma Completo, do inglês *Complete blood count*;
Células NK	Células Exterminadoras Naturais , do inglês *Natural Killer Cell*;
DHEA-S	Sulfato de Deidroepiandrosterona, do inglês *Dehidroepiadrosterona Sulfate*;
DHEA	Deidroepiandrosterona, do inglês *Dehydroepiandrosterone*;
DNA	Ácido Desoxirribonucleico, do inglês *Deoxyribonucleic acid*;
EAGC ou GWAs	Estudos de Associação de Genoma Completo, do inglês *Genome-wide association studies*;
FDR	Descobertas Taxa de Falsas, do inglês *False Discovery Rate*;
HLA	Sistema Antígeno Leucocitário Humano, do inglês *Human leukocyte antigen*;
IFN	Interferon;
IRF5	Fator Regulador do Interferão 5, do inglês *Interferon regulatory factor 5*;
meQTL	Loci Controladores de Traços Quantitativos da Metilação, do inglês *Methylation Quantitative Trait Loci*;
miRNA	Micro Ácido Ribonucleico, do inglês *Micro Ribonucleic acid*;
NF-κB	Fator Nuclear κB, do inglês *Factor Nuclear κB*;
NOD2	Domínio de Oligomerização Nucleotídica 2, do inglês *Nucleotide-binding Oligomerization Domain Containing 2*;
OMRF	Fundação de Pesquisa Médica de Oklahoma, do inglês *Oklahoma Medical Research Foundation*;

OR Razão de Chances, do inglês *Odds Ratio*;

PPCLUST Cluster de partição de dados de tamanho amostral alto e dimensional baixo com base em p-valores, do inglês *Partition clustering of high dimensional low sample size data based on p-values.*

RNA Ácido Ribonucleico, do inglês *Ribonucleic acid*;

RNAm ou mRNA Ácido Ribonucleico Mensageiro, do inglês *Messenger Ribonucleic Acid*;

SNP Polimorfismo de Nucleotídeo Simples, do inglês *Single Nucleotide Polymorphism*;

STAT4 Transdutor de Sinal e Ativador da Transcrição 4, do inglês *Signal transducer and activator of transcription 4*;

TDEs Transcritos diferencialmente expressos;

TLR Receptores do tipo Toll, do inglês *Toll-like receptors*;

TNF Fatores de Necrose Tumoral, do inglês *Tumor Necrosis Factor*;

SS Síndrome de Sjögren.

Sumário

Lista de ilustrações 11

Lista de tabelas 12

1 INTRODUÇÃO 19

2 METODOLOGIA 25

2.1 Conjunto de dados 25

2.2 Estimação de expressões gênicas 27

2.2.1 Procedimentos de controle de qualidade 29

2.2.2 Correção de *Background*, Normalização e Sumarização 29

2.3 Análises de expressões gênicas 30

2.3.1 Análise exploratória 30

2.3.2 Métodos de filtragem 30

2.3.3 Filtragem por magnitude de efeito 35

2.3.4 Análises de agrupamento 35

2.3.5 Anotação genômica 38

3 RESULTADOS 41

3.1 Análise exploratória 41

3.2 Análise de perfil de expressão gênica 45

3.3 Identificação de sistemas biológicos e funções genéticas 51

4 DISCUSSÃO E CONCLUSÕES 65

REFERÊNCIAS 67

REFERÊNCIAS 69

APÊNDICES 71

APÊNDICE A – DESCRIÇÃO DOS SISTEMAS BIOLÓGICOS IDENTIFICADOS EM ANÁLISES DE ANOTAÇÃO . . 73

APÊNDICE B – LISTA DE TRANSCRITOS ESTATISTICAMENTE SIGNIFICANTES 76

APÊNDICE C – INFORMAÇÕES RELACIONADAS AO SOFTWARE R E AO BANCO DE DADOS. 81

1 Introdução

A Síndrome de Sjögren (SS) é uma doença autoimune sistêmica caracterizada pela infiltração de linfócitos em glândulas exócrinas, causando processos inflamatórios e eventual diminuição em suas funções normais (Ramos-Casals e outros, 2012). As glândulas lacrimais e salivares são as mais comumente afetadas, o que resulta em sintomas e sinais de secura intensa na boca (xerostomia) e nos olhos (ceratoconjuntivite seca) devido à diminuição na produção de saliva e lágrima. Indivíduos portadores de SS podem apresentar outras complicações graves incluindo fadiga profunda, dor crônica e a infiltração em outras glândulas exócrinas (como o pâncreas, glândulas sudoríparas, glândulas mucosas dos tratos respiratório, gastrointestinal e uro-genital) e apresentar ainda neuropatias e risco aumentado de linfomas (Tapinos e outros, 1999).

A secura pode afetar profundamente a qualidade de vida dos portadores de SS, interferindo em funções diárias básicas, como comer, falar e dormir, dado que a redução do volume de saliva e a subsequente perda das propriedades antibacterianas da saliva podem acelerar infecções, como a cárie dentária e a doença periodontal. Glândulas mucosas do trato respiratório superior e inferior também podem ser afetadas, levando a um ressecamento do nariz, garganta e traqueia, resultando muitas vezes em tosse seca crônica (Kassan e Moutsopoulos, 2004).

A SS pode ser classificada em dois tipos: Síndrome de Sjögren Primária (SSp), que ocorre na ausência de outro distúrbio autoimune subjacente, e em Síndrome de Sjögren Secundária (SSs), que está associada a outros distúrbios autoimunes, principalmente lúpus eritematoso sistêmico (LES), artrite reumatoide e esclerose sistêmica (Pasoto, Martins e Bonfa, 2019).

Estima-se que a prevalência mundial de Síndrome de Sjögren esteja em torno de 0,5% a 1,0% da população adulta, ocorrendo em mulheres em 90% dos casos (Carsons e Patel, 2019), em sua maioria caucasianas (Patel e Shahane, 2014) em idades entre os 40 e 60 anos (Qin e outros, 2015). Estudos sugerem que o risco aumentado nessa faixa etária esteja relacionado à ocorrência da menopausa, que causa uma alteração na razão andrógeno-estrogênio e a diminuição de ambos os hormônios (Brandt e outros, 2015). Há evidências de que baixos níveis de andrógenos estão associados à SS. Estudos reportaram concentrações do andrógeno deidroepiandrosterona (DHEA) 40 a 50% mais baixas em pacientes com SS do que em controles pareados por idade e sexo (Valtysdóttir, Wide e Hällgren, 2001). Além disso, foi descoberto que, em glândulas salivares de pacientes com SS, o DHEA foi efetivamente convertido em testosterona e não em di-hidrotestosterona (o que seria o processo normal), provavelmente devido à expressão diminuída das principais

enzimas esteroidogênicas e sua localização subcelular anormal (Nikolov e Illei, 2009).

Fatores ambientais, como infecções virais prévias (a vírus Epstein-Barr, citomegalovírus, vírus herpes humano, vírus da hepatite C, dentre outros) ou bacterianas (a Helicobacter pylori) já foram apontados como potenciais desencadeadores da resposta imune ao tecido glandular, devido à sua frequente concomitância em pacientes com SS (Felberg e Dantas, 2006). Além disso, segundo pesquisa realizada (Theander e outros, 2015), autoanticorpos, que são anticorpos produzidos pelo sistema imune que atuam contra uma ou mais proteínas do próprio indivíduo, podem estar presentes 18 a 20 anos antes do diagnóstico de SS primária, principalmente em famílias com vários casos de doenças autoimunes, em que o perfil de autoanticorpos, juntamente com a avaliação do risco genético, permite a identificação de indivíduos suscetíveis em um estado pré-doença (Theander e outros, 2015).

Vários estudos de genes candidatos e, mais recentemente, estudos de associação de genoma completo (EAGC, do inglês *Genome-wide association studies* ou GWAs) reportaram associações de SSp com muitos fatores genéticos, destacando-se os maiores efeitos com genes *HLA* e efeitos moderados com genes *IRF5* e *STAT4*, consistentes em várias etnias (Imgenberg-Kreuz e outros, 2019), como ilustrado na Figura 1. Os genes incluídos nesta Figura, foram identificados com significância de estudos de associação de genoma completo (GWAS), após correção para testes múltiplos que envolvem centenas de milhares a milhões de testes realizados para todos os marcadores separadamente.

Além de associações com os genes *HLA*, *IRF5* e *STAT4*, estudos de genes candidatos para a SSp (Imgenberg-Kreuz e outros, 2019) identificaram outros *loci* associados a maiores riscos para o desenvolvimento da síndrome (Quadro 1 que contém alguns deles assim como a função de cada um).

Também foram observadas em estudos de perfil de expressão gênica associações de SSp com muitos desses genes candidatos, como *IFN, EPSTI1, STAT1* e *IFI44L*, que são relacionados sobretudo à atividade da doença e à presença de autoanticorpos (Yao e outros, 2019). Tais estudos utilizaram plataformas chamadas de chips de *microarray* (ou microarranjos em português), que permitem a estimação (por quantificação de RNA mensageiro) de expressão gênica de dezenas de milhares de transcritos (trecho do DNA já transcrito em uma molécula de RNA) simultaneamente. Representam uma das abordagens mais amplamente utilizadas para se determinar diferenças globais de transcriptoma (medição em larga escala de transcritos) entre casos e controles saudáveis. Estudos de perfil de expressão gênica permitiram a identificação das denominadas "assinaturas" de expressão da doença. Uma assinatura representa um grupo de genes co-expressos (e diferentemente expressos entre casos e controles) que pertencem a sistemas biológicos específicos. Os genes ou os sistemas identificados podem servir como biomarcadores para diagnóstico, classificação e previsão de resposta a medicamentos (Li e outros,2013).

Nesse contexto, o presente trabalho teve como objetivo a realização de uma análise de perfil de expressões gênicas com base na identificação de transcritos diferencialmente expressos (TDEs) entre casos com SSp e controles saudáveis.

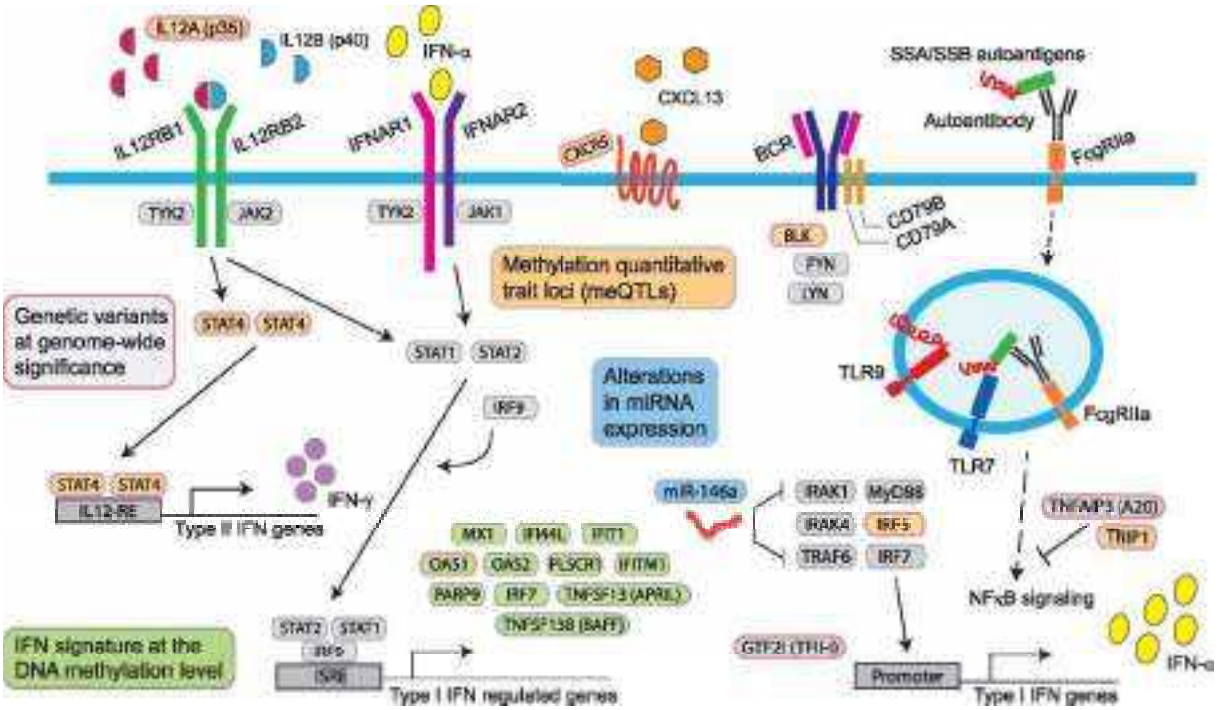


Figura 1 – Resumo esquemático dos mecanismos genéticos e epigenéticos associados com a suscetibilidade à SSp. Caixas com bordas em rosa representam genes localizados fora da região HLA com SNPs associados à suscetibilidade de SSp com significância de estudos de associação de genoma completo (GWAS) após correção para testes múltiplos e incluem *BLK*, *CXCR5*, *GTF2I*, *IL12A*, *IRF5*, *OAS1*, *STAT4*, *TNFAIP3* e *TNIP1*. Caixas verdes representam genes induzidos por interferon (IFN), como *MX1* , *IFI44L*, *OAS1*, *OAS2*, *TNFSF13B* e *IRF7*, com locais CpG hipometilados em SSp. Caixas laranjas incluem meQTLs (loci controladores de traços quantitativos da metilação), referentes à associação entre um variante genético (SNP) e o nível de metilação em um local CpG próximo como *BLK*, *CXCR5*, *IL12A*, *IRF5-TNPO3*, *STAT4* e *TNIP1*. Caixas azuis representam alterações na expressão do miRNA, incluindo o *miRNA-146*, foram identificadas em SSp (Imgenberg-Kreuz e outros, 2019).

Quadro 1: Listagem de genes que conferem maior risco para a SS identificados em estudos de genes candidatos (p-valores corrigidos por testes múltiplos inferiores a 5%).

Gene	SNP	OR (95% CI)	Função
IRF5-TNPO3	rs2004640	1.93 (1.15,3.42)	Tipo I de IFN e sinalização **TLR**, produção de citocina produção (por exemplo, IL-6, IL-12),apoptose.
	rs10488631	1.57 (1.23.1.99)	
	CGGGGindel	2.00 (1.50.2.70)	
		1.49 (1.24, 1.79)	
	rs13246321	1.70 (1.35,2.13)	
STAT4	rs7574865	1.46 (1.09, 1.97)	Tipo I e II de IFN, sinalização **NOD2**.
		1.40 (1.21, 1.62)	
	rs7582694	1.41 (1.14,1.73)	
		1.40 (1.15.1.70)	
BLK- FAM167A	rs12549796	1.37 (1.15,1.63)	Sinalização e desenvolvimento de células B. **timopoiese**, secreção e sintese de insulina e sinalização de **NF-kB**.
EBF1	rs3843489	1.68 (1.29.2.18)	Potenciador da atividade transcricional durante o desenvolvimento de células B.
TNSF4/OX40L	rs1234315	1.34 (1.14,1.64)	Co-estinulação, proliferação de células T e produção de citocinas.
BAFF/INFSF 13B	5 haplotype	2.60 (1.70,4.10)	Maturação, proliferação e sobrevivência de células B. além da sobrevivência de células epiteliais.
	5 haplotype (TATT,GTTC)	-	
NCR3/NKp30	rs11575837	0.48	Regulação das comunicações **dendríticas** e celulares **NK** e regulação das citocinas Thl (IL-12 e IFN-y).
	rs2736191	0.56	
PTPN22	rs2476601	2.42 (1.24,4.75)	Ativação de células T.
TNIP1	rs3792783	1.33 (1.16,1.52)	Sinalização de **NF-kB** e EGF/ERK e apoptose induzida por **TNF**.
	rs7708392	1.21 (1.08,1.36)	
TNFAIP3	rs2230926	3.26(1.31,8.12)	Sinalização de **NF-kB** (repressor), apoptose induzida por **TNF**, sinalização de **TLR**4, produção de citocinas (por exemplo, IL-1B).
BAFF-R	His159Tyr	4.13 (1.19.14.3)	Sinalização de NF-kB2.
LTA/LTB/INF	rs1800629	2.00 (1.61,2.49)	Organogênese linfoide, manutenção do tecido linfoide terciário.
	rs909253	1.59 (1.34,1.89)	
MECP2	rs17435	1.33 (1.12,1.59)	Silenciamento transcricional mediado por metilação.

Adaptado e traduzido de: (Imgenberg-Kreuz, et al.,2019)

OR: *odds ratio.*

TLR: toll-like receptors, receptores do tipo Toll, que são uma família de proteínas transmembrânicas de tipo I que formam uma parte do sistema imunológico inato.

NOD2: (em inglês: nucleotide-binding oligomerization domain containing 2) é uma proteína que desempenha um importante papel no sistema imune.

Timopoiese: processo em que dentro do timo os timócitos entram em um processo de seleção e maturação.

NF-κB: (factor nuclear κB) é um complexo proteico que desempenha funções como fator de transcrição e um papel fundamental na regulação da resposta imunitária à infecção.

Células dendríticas: nomeadas por suas formas de sondagem, "tipo árvore" ou dendríticas, são responsáveis pelo início das respostas imunes adaptativas e, portanto, funcionam como as "sentinelas" do sistema imunológico.

Células NK: células exterminadoras naturais (do inglês Natural Killer Cell) são um tipo de linfócitos citotóxicos necessários para o funcionamento do sistema imunitário inato.

TNF: (do inglês: Tumor necrosis factor), são os fatores de necrose tumoral e este gene codifica uma citocina pró-inflamatória importante, produzida principalmente por macrófagos, que tem sido relacionada com diversas doenças, como as autoimunes e as degenerativas. Estudos indicam que polimorfismos do TNF podem influenciar seu nível de expressão e atividade e, portanto, poderiam influenciar a susceptibilidade a tumores (Lopes e outros, 2014).

2 Metodologia

2.1 Conjunto de dados

Neste trabalho foram analisados dados de expressões gênicas de 222 indivíduos de descendência europeia, sendo 190 diagnosticados com SSp (grupo de casos) e 32 indivíduos saudáveis (grupo de controle) de um estudo de caso-controle (Lessard e outros, 2013) com 15.063 transctitos e aprovado pelos devidos comitês de ética[1] e consentimento informado (documento assinado) pelos participantes.

Os dados são públicos e fazem parte do repositório funcional de dados genômicos `Gene Expression Omnibus (GEO)` (Edgar, Domrachev e Lash, 2002).

Os pacientes com SS foram avaliados e recrutados por médicos especialistas na Universidade de Minnesota ou na *Oklahoma Medical Research Foundation (OMRF)*. O critério adotado para a classificação dos casos foi o da American-European Consensus Group (AECG) de 2002 (Vitali e outros, 2002), apresentado no Quadro 2.

Segundo o critério AECG, para ser diagnosticado com Síndrome de Sjögren Primária um indivíduo deve apresentar:

- Quaisquer 4 dos 6 critérios, incluindo obrigatoriamente o item IV (Histopatologia) ou o VI (Autoanticorpos) ou
- Quaisquer 3 dos 4 itens de critérios objetivos (III, IV, V, VI).

Tendo como critérios de exclusão o tratamento anterior de radiação de cabeça e pescoço, infecção por hepatite C, AIDS, linfoma preexistente, sarcoidose, enxerto por doença do hospedeiro ou o uso atual de drogas anticolinérgicas.

[1] Oklahoma Medical Research Foundation Review Boards e University of Minnesota Institutional Review Board.

Quadro 2: Critério adotado para a classificação dos casos da (AECG) de 2002.

I. Sintomas oculares - resposta positiva a pelo menos uma das seguintes perguntas:	**1.** Tem problemas oculares diários e persistentes, relacionados a quadro de olho seco há mais de 3 meses? **2.** Tem sensação de areia ou queimação ocular? **3.** Usa colírios lubrificantes mais de 3 vezes ao dia?
II. Sintomas orais - resposta positiva a pelo menos uma das seguintes perguntas:	**1.** Tem uma sensação diária de boca seca há mais de 3 meses? **2.** Tem inchaço recorrente ou persistente das glândulas salivares, na idade adulta? **3.** Sente necessidade frequente de ingerir líquidos para ajudar na deglutição de alimentos sólidos?
III. Sinais oculares - evidência objetiva de envolvimento ocular definido como resultado positivo para, pelo menos, um dos dois testes a seguir:	**1.** Teste de Schirmer I, realizado sem anestesia (<5 mm em 5 minutos). **2.** Pontuação da Rosa Bengala ou outra pontuação de corante ocular (>4 dna escala de van Bijsterveld's).
IV. Histopatologia: achados histopatológicos nas glândulas salivares menores (obtidas através da mucosa de aparência normal) sialoadenite linfocítica focal:	Avaliada por um histopatologista especialista, com um escore de foco >1, definido como um número de focos linfocitários adjacentes aos ácinos mucosos de aparência normal com pelo menos 50 linfócitos por 4 mm2 de tecido da glândula salivar.
V. Comprometimento da glândula salivar: evidência objetiva, definida por um resultado positivo para, pelo menos, um dos seguintes testes de diagnóstico:	**1.** Fluxo salivar total não estimulado (<1,5 ml em 15 minutos). **2.** Sialografia parotídea mostrando a presença de sialectasias difusas (padrão pontuado, cavitário ou destrutivo), sem evidência de obstrução nos ductos principais. **3.** Cintilografia salivar mostrando captação tardia, concentração reduzida e/ou excreção retardada do traçador.
VI. Autoanticorpos: presença de pelo menos um dos seguintes autoanticorpos séricos:	Anticorpos contra os antígenos Ro (SSA) ou La (SSB), ou ambos.

Os dados de expressões gênicas utilizados são resultantes de dois experimentos de *microarray* separados. O RNA total foi obtido pela coleta de sangue periférico em tubos PAXGene (BD Company) e extraídos seguindo protocolos do fabricante (Qiagen). Concentrações de RNA foram determinadas utilizando-se um espectômetro NanoDrop (Thermo Scientific). A qualidade de RNA foi avaliada pelo Agilent 2100 Bioanalyzer baseada na razão ribossômica 28S/18S e no número de integridade de RNA. cDNA de dupla hélice foi sintetizado usando um promotor T7 e cRNA biotin-labeled foi transcrito utilizando o sistema de amplificação TotalPrep de RNA da Illumina (Ambion). Amostras foram hibridizadas ao chip Human WG-6 v3.0 BeadChip microarrays (da empresa Illu-

mina), que contém 48.803 sequências de mRNA distintas. Os *microarrays* foram lavados para remoção dos "alvos" excedentes com alto rigor e marcados com estreptavidina-Cy3. Os dados de expressão gênica com base na intensidade fluorescente foram obtidos pela leitura dos chips no scanner BeadStation 500 da Illumina ou o iScan (Lessard e outros, 2013), conforme ilustrado na Figura 2.

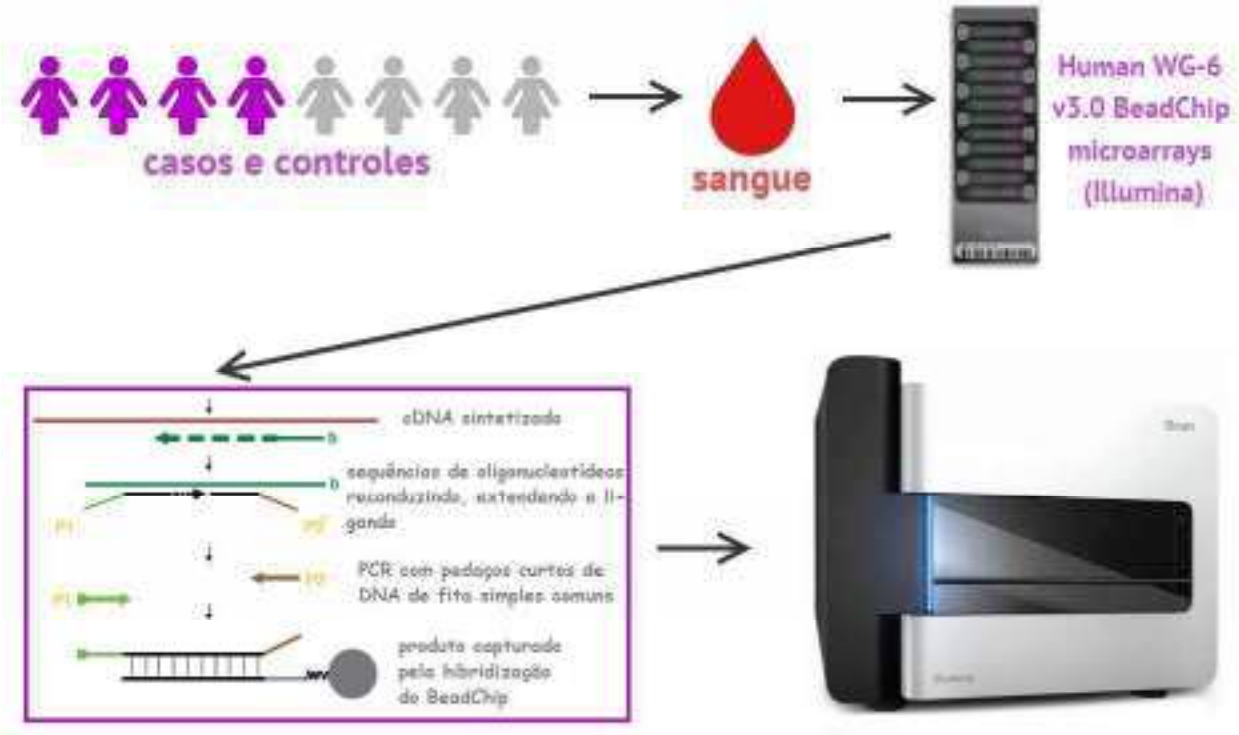


Figura 2 – Resumo esquemático do processo de obtenção das intensidades de expressão gênica. <https://www.illumina.com/>

2.2 Estimação de expressões gênicas

A expressão gênica é o processo pelo qual a informação codificada em uma sequência de DNA é utilizada para a formação de um produto genético funcional. Quando o produto é uma molécula de proteína, o processo é denominado síntese de proteína ou síntese protéica e ocorre quando RNA mensageiro (RNAm) é produzido (processo chamado de transcrição) e traduzido em proteína (processo chamado de tradução). Sequências de DNA não-codificante produzem RNA funcional, que inclui elementos regulatórios de atividade gênica, como por exemplo RNA transportador (tRNAs), RNA ribossômico (rRNAs), microRNAs (miRNAs), RNAs não-codificante longo (lncRNAs). A expressão gênica é um processo que regula rigidamente se proteínas devem ser produzidas ou não e em que quantidades, permitindo que a célula responda a estímulos diversos (Griffiths e outros, 2005).

Desafios associados à medição das expressão de proteínas, como a imprecisão na quantificação de algumas moléculas de proteína, diferentes alternativas de transcrição e variações no nível da proteína, maiores que no nível do RNA, estimularam a utilização da quantidade de mRNA produzido (transcrito) para se estimar a expressão gênica (Arivaradarajan e Misra, 2019). Com isso, para as análises de transcriptoma[2] utiliza-se a quantificação dos mRNAs em um dado organismo ou tecido. O perfil de expressão gênica em diferentes condições ambientais, estados patológicos, fisiológicos ou de desenvolvimento e finalmente a caracterização de polimorfismos associados aos genes transcritos (SNPs[3] e formas alternativas de *splicing*) (Bravim, 2013).

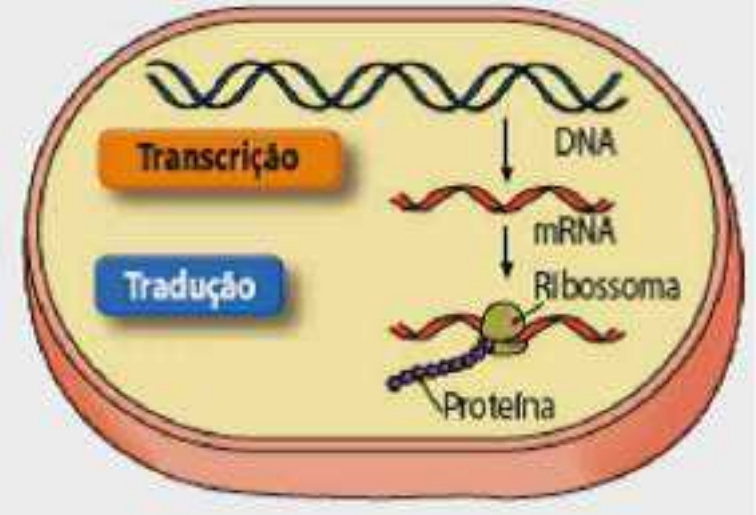


Figura 3 – Processo da síntese proteica (ou expressão gênica)
<https://i.pinimg.com/originals/96/6d/07/966d07e9195385b3c50d294e23ef44ff.png>

Com o objetivo de estimar os transcritos, utilizaram o chip de alto rendimento `Human WG-6 v3.0 BeadChip` (Illumina, 2008) da empresa `Illumina` para a análise das amostras hibridizadas que permite a medição de expressão do genoma inteiro e estudos integrados de genotipagem. Sua cobertura inclui mais de 48.000 transcritos escolhidos com base em dois diferentes bancos de dados que agrupam diversos genes que estão ligados a diferentes funções biológicas, são eles: *Reference Sequence (RefSeq)* e *UniGene*.

A análise de perfil de expressão gênica envolve uma sequência de etapas que podem ser divididas em **análises no nível de probe**, realizadas previamente por outros pesquisadores (Lessard e outros, 2013), e em **análises no nível de expressão gênica**, foco do presente trabalho.

[2] Transcriptoma é o conjunto completo (ou em larga escala) de transcritos (mRNAs, rRNAs, entre outros) de um dado organismo, órgão, tecido ou linhagem celular.
[3] Variações na sequência de DNA dos humanos.

As análises no nível de probe [4] que foram realizadas (Lessard e outros, 2013) incluíram em sequência: procedimentos de controle de qualidade, correção de *background*, normalização e sumarização, descritas a seguir.

2.2.1 Procedimentos de controle de qualidade

O conjunto de dados utilizado envolve dados de dois experimentos de *microarray* separados e por isso procedimentos de controle de qualidade foram realizados separadamente para cada subconjunto.

Com o intuito de identificar e eliminar dados de baixa qualidade, foram excluídos os transcritos expressos em menos de 10% dos indivíduos (limiar de detecção: P-valor $<$ 0,05) por subconjunto de dados. Foram ainda excluídos os transcritos com percentuais de valores faltantes estatisticamente diferentes entre os dois subconjuntos de dados (P-valor $<$ 0,001 pelo teste exato de Fisher (Agresti, 2019). Os probes relativos à genes não anotados ou com resultados inconsistentes entre os dois bancos de dados também foram excluídos.

2.2.2 Correção de *Background*, Normalização e Sumarização

O procedimento de correção de *background* tem por objetivo estimar o verdadeiro nível de hibridização do cDNA [5], removendo ruídos (sinais de intensidade não específicas ou *non-specific binding*) do sinal de intensidade total, detectado por leitura em scanner (Figura 2).

A etapa seguinte, de normalização, é essencial para garantir que as expressões gênicas das diferentes observações (dados de microarray para cada indivíduo) sejam comparáveis, removendo potenciais erros sistemáticos de medição associados aos processos de hibridização e leitura (por scanner) de imagens, problemas físicos nos chips, efeito de lote de reagentes, condições laboratoriais entre outras fontes de variação. A normalização quantílica tem por objetivo fazer duas distribuições idênticas em propriedades estatísticas, pra isso, transforma os dados impondo a mesma distribuição empírica da média obtida pela média de cada quantil nas amostras. Além disso, os dados são transformados para a escala logarítma de base 2.

Métodos de sumarização resumem os dados à nível de *probe* em uma única medida (por transcrito) que representa o nível estimado de expressão para cada transcrito. O método da mediana polida é comumente utilizado para esse fim dado que ele protege contra sondas outlier, extraindo efeitos de linhas e colunas através da subtração recursiva

[4] Fragmento de RNA usado em genética para pesquisar um determinado gene ou outra sequência de DNA

[5] cDNA ou DNA complementar, é o DNA sintetizado a partir de uma molécula de RNA mensageiro

das medianas de linhas e de colunas.

2.3 Análises de expressões gênicas

A avaliação dos dados de expressão gênica para os transcritos incluídos nesta etapa (após as exclusões de controle de qualidade) foi iniciada por uma análise descritiva dos dados. Em seguida, testes estatísticos e algorítmos multivariados foram utilizados para a identificação de TDEs entre grupos do estudo. Análises de agrupamento foram utilizadas para a visualização de perfis de expressão diferenciados entre os grupos de interesse. Além disso, procedimentos de anotação genômica foram realizados para identificar informações de genes (dos transcritos diferencialmente expressos que se ligam a algum gene) relacionados a sistemas biológicos associadas a Sindrome de Sjögren.

2.3.1 Análise exploratória

Gráficos e tabelas das distribuições (na escala original e em escala log2) foram elaborados para avaliar distribuições das expressões gênicas e identificar possíveis irregularidades e padrões nas observações. Testes de Shapiro-Wilk (Shapiro e Wilk, 1965) foram utilizados para avaliar a normalidade (hipótese nula) das distribuições de expressão gênica de todos os transcritos para casos e controles. Este teste foi escolhido por possuir maior poder para detectar desvios de normalidade por assimetria, padrão esperado para alguns transcritos.

2.3.2 Métodos de filtragem

Alguns métodos podem ser utilizados para a identificação dos chamados transcritos diferencialmente expressos (TDEs), cujas diferenças entre níveis de expressão são estatísticamente significantes. Comumente, são identificados por testes de comparação de médias ou de medianas, com correção por testes múltiplos. Modelos de regressão são utilizados quando covariáveis clínicas ou moleculares estão disponíveis.

Neste trabalho foram utilizadas três abordagens para a identificação de TDEs entre casos e controles para cada transcrito, listadas abaixo e descritas em sequência:

1. Testes T com correção de Welsh (para variâncias desiguais) para comparação de médias, seguidos por correção por testes múltiplos;

2. Testes não paramétricos de Wilcoxon-Mann-Whitney para comparação de medianas, seguidos por correção por testes múltiplos;

3. O Algoritmo PPCLUST (Partition clustering of high dimensional low sample size data based on p-values).

Sendo a primeira delas, o Teste T, a mais utilizada na literatura mas que depende da suposição de normalidade dos transcritos, que muitas vezes não é atendida, além da dependência de uma correção por testes múltiplos. A segunda, Wilcoxon-Mann-Whitney, um teste não paramétrico com os mesmos objetivos do Teste T, mas sem a necessidade de uma suposição de normalidade, mantendo-se ainda a necessidade de correção por testes múltiplos e, o PPCLUST, que mesmo sendo um algorítmo de agrupamento tem sua finalidade adaptada e é uma técnica invariante a transformações monótonas que não depende de correção por testes múltiplos.

1. Teste T - Welch de comparação de médias

Com o objetivos de identificar transcritos diferencialmente expressos, o Teste T foi utilizado. Seguindo para isso as seguintes hipóteses para cada um dos k transcritos (Morettin e Bussab, 2017):

$H_0 : \mu_{\text{(k-ésimo caso)}} = \mu_{\text{(k-ésimo controle)}}$

$H_A : \mu_{\text{(k-ésimo caso)}} \neq \mu_{\text{(k-ésimo controle)}}$

A estatística do teste é dada por:

$$T = \frac{\bar{X}_{casos} - \bar{X}_{controles}}{\sqrt{S^2_{casos}/n + S^2_{controles}/m}} \sim t(v), \tag{2.1}$$

em que $\bar{X}_{casos}$ e $\bar{X}_{controles}$ representam médias amostrais; S^2_{casos} e $S^2_{controle}$, variâncias amostrais e n e m, número de observações para os grupos casos e controles, respectivamente.

O número de graus de liberdade é aproximado pela expressão de Welch-Satterthwaite:

$$v = \frac{(A+B)^2}{(A^2/(n-1) + B^2/(m-1))}, \tag{2.2}$$

em que $A = S^2_{casos}/n$ e $B = S^2_{controles}/m$.

2. Teste de Wilcoxon-Mann-Whitney

Tendo em vista que alguns dos transcritos podem apresentar distribuições acentuadamente assimétricas, utilizou-se o teste não paramétrico de Wicoxon-Mann-whitney para a identificação de TDEs, cujas hipóteses são:

$H_0 : \theta_{casos} = \theta_{controles}$;

$H_A : \theta_{casos} \neq \theta_{controles}$,

em que θ_i representa a mediana de i. A estatística do teste é dada por:

$$W = S_n - \frac{1}{2}n(n+1), \tag{2.3}$$

em que S_i é somas dos postos dos elementos de i, $S_n = \frac{1}{2}(m+n)(m+n+1) - S_m$ e n e m o tamanho amostral dos casos e dos controles, respectivamente.

O p-valor do teste é calculado da seguinte maneira:

$$p - valor = \begin{cases} 2P(W > W_{obs} - 1) & \text{, se } W_{obs} > \frac{mn}{2}; \\ 2P(W > W_{obs}) & \text{, se } W_{obs} \leq \frac{mn}{2}, \end{cases} \tag{2.4}$$

onde segue-se uma distribuição normal com média $\frac{n(n+m+1)}{2}$ e variância $\frac{nm(n+m+1)}{12}$.

3. PPCLUST

O algorítmo *Partition clustering of high dimensional low sample size data based on p-values* (PPCLUST) (von Borries e Wang, 2009) também foi utilizado para a identificação de TDEs. Trata-se de um algoritmo de agrupamento de particionamento com uma robusta medida de distância que pode agrupar variáveis de um banco de dados com altas dimensões e um tamanho de amostra baixo (HDLSS - *High dimensional low sample size data based on p-values*). É um algorítmo não hierárquico baseado em modelagem de dados por mistura de distribuições não especificadas (abordagem não paramétrica), usando P-valores como medida de similaridade e regra automática de parada.

O PPCLUST tem como base a metodologia de ANOVA não-paramétrica e tem como funcionalidade agrupar variáveis com a mesma distribuição de dados, resultando ao final do procedimento em grupos com uma variância alta entre si e pequena dentro do grupo, sem que para isso seja preciso a pré-especificação do número de grupos a serem formados.

A primeira etapa envolve o cálculo das diferenças entre expressões dos casos e a mediana dos controles.

O algoritmo é invariante a transformações monótonas mas posto que esses dados foram disponibilizados na escala log2, essa escala foi mantida para o cálculo dessas diferenças, como na expressão:

$$Y_{ij} = log_2(X_{ij}^{casos}) - log_2(\theta^{controles}), \tag{2.5}$$

em que X_{ij}^{casos} denota o valor da i-ésima variável (transcrito) para o j-ésimo caso e $\theta^{controles}$ representa a mediana de expressão para o grupo controle. $\{Y_{ij}, 1 \leq j \leq n_i\}$ são consideradas observações independentes de alguma distribuição desconhecida $F_i(x)$, $i : 1, 2, ...n$. Quer-se testar a seguinte hipótese nula:

$$H_0 : F_1(x) = ... = F_2(x) = ... = F_n(x). \tag{2.6}$$

A Estatística do Teste se dá pela razão entre o quadrado médio do tratamento (MST_R) e o, do erro (MSE_R):

$$F_R = \frac{MST_R}{MSE_R} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\bar{R}_{i.} - \tilde{R}_{..})^2}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_1}S_{R,i}^2}, \tag{2.7}$$

em que R_{ij} representa o rank da observação Y_{ij} no conjunto de todas as observações $n_1 + n_2 + ... + n_a$ e R_{ij} é uniformemente distribuído entre 1 e $\sum_{i=1}^{a} n_i$; $\bar{R}_{i.}$ representa a classificação média das observações para o i-ésimo transcrito; $\tilde{R}_{..}$ é definido pela média geral não ponderada de classificações de todos os transcritos e $S_{R,i}^2$ é a variância da amostra calculada usando o rank das observações do i-ésimo transcrito.

Além disso, em relação a distribuição assintótica segue que:

Teorema 1 *((Wang e Akritas, 2004) Teste de ausência de efeito de grupo)*

Seja $H_0 : F_1(x) = ... = F_a(x)$ *atendido, com* $F_i(x)$ *arbitrário. Se* $n_i \geq 2$ *fixo, assumindo que as observações são independentes, os seguintes limites existem:*

$$v_2^2 = \lim_{a\to\infty} \frac{1}{a}\sum_{i=1}^{a}\frac{1}{n_i}\sigma_i^2 > 0; \tag{2.8}$$

e

$$\tau_2 = \lim_{a\to\infty} \frac{1}{a}\sum_{i=1}^{a}\frac{2\sigma_i^4}{n_i(n_i-1)}. \tag{2.9}$$

Então, se $a \to \infty$ tem-se que:

$$\sqrt{a}(F_R - 1) \xrightarrow{d} N(0, \tau_2/v_2^4). \tag{2.10}$$

Nesse algorítmo, para obter uma medida de similaridade com objetivo de agrupar os dados, utiliza-se o P-valor obtido com a estatística $\sqrt{a(F_R - 1)}$ e, então, compara-se o P-valor com o nível de significância α pré-especificado. O valor desse nível α pré-especificado geralmente é testado com diversos valores com o objetivo de verificar se os agrupamentos são consistentes e/ou quando vai ocorrer uma estabilização no número de grupos.

Ademais, um nível de significância usado que leva a clusters muito pequenos indica que o nível α não é pequeno o suficiente e os resultados de cluster obtidos não são confiáveis.

Além disso, quando esse nível α estável é elevado as variáveis testadas são dadas como distribuições similares e, quando pequeno, há evidências contra H_0 de que pelo menos uma variável tem distribuição diferente das demais.

Para informações sobre a implementação do PPCLUST no R ver: Lins (2018)

Correção por Testes Múltiplos

Estudos de *microarray* comumente envolvem a realização de dezenas de milhares de testes estatísticos (um para cada transcrito) para a identificação de TDEs. Se cada teste é realizado a um nível de significância α, a probabilidade de se rejeitar incorretamente pelo menos umas das hipóteses nulas será muito maior que o nível α. Métodos de correção por testes múltiplos são utilizados para controlar a probabilidade da ocorrência de erros do tipo I. O método de Benjamini e Hochberg (1995), abordagem menos conservadora para a correção de múltiplas, foi utilizado para ajustar os p-valores obtidos nas abordagens 1. e 2. Foram calculadas taxas de descobertas falsas, do inglês *False Discovery Rates* (FDRs) para controlar a proporção de falsos positivos entre o conjunto de hipóteses rejeitadas nesses conjuntos de testes.

A Tabela 1 apresenta a distribuição de testes múltiplos por resultado (significante ou não) e situação real (H_0 verdadeira ou não) (Garcia, 2019).

Tabela 1 – Tabela do método FDR - False Discovery Rate.

	Significante	**Não significante**	**Total**
H_0 **verdadeiro**	F	m_0 - F	m_0
H_A **verdadeiro**	T	m_1 - T	m_1
Total	S	m - S	m

A proporção de resultados incorretamente declarados como significantes entre todos os resultados significantes é dada por:

$$\frac{n^{\circ}\text{falsos positivos}}{n^{\circ}\text{testes significativos}} = \frac{F}{S}.$$

O FDR(T) é calculado através da seguinte esperança:

$$FDR(T) = E\left[\frac{F(T)}{S(T)}\right].$$

Aqui, T representa o limiar usado para considerar os P-valores como significativos ($0 < T \leq 1$).

Sendo esse cálculo de falsos positivos feito com os P-valores obtidos abaixo do α especificado, utilizando para isso uma distribuição binomial(r,α), onde r representa a quantidade de P-valores rejeitados pelo α especificado antes da correção. Ajustando assim a distribuição real dos P-valores dos dados.

2.3.3 Filtragem por magnitude de efeito

Geralmente, seleciona-se para as análises subsequentes, apenas os TDEs com maiores diferenças entre grupos, que corresponderiam aos maiores efeitos (positivos ou negativos). Para tanto, pode-se utilizar uma métrica denominada *fold-change*, definida como a razão de médias (ou de medianas) de dois grupos. Neste trabalho, foram considerados TDEs os transcritos com *fold-change* (FC), definido pela razão entre as médias (ou medianas) de casos e controles, superior a 1,5 ou inferior a 0,67.

Para os transcritos selecionados pela abordagem 1. (por testes t ajustados por testes múltiplos) FC foi calculado por:

$$FC_i = \frac{\bar{X}_i^{casos}}{\bar{X}_i^{controles}}, \tag{2.11}$$

em que $\bar{X}_i^l$ repesenta a média para o i-ésimo transcrito de t (casos ou controles).

Já para as abordagens 2. (por testes de Wilcoxon-Mann-witney) e 3. (pelo algorítmo PPLUST) bem como para a interseção e união das abordagens 1. a 3., FC foi calculado por:

$$FC_i = \frac{\theta_i^{casos}}{\theta_i^{controles}}, \tag{2.12}$$

em que θ_i^l repesenta a mediana para o i-ésimo transcrito de t (casos ou controles).

2.3.4 Análises de agrupamento

Análises de Agrupamento foram realizadas objetivando a definição de padrões de expressão gênica entre casos e controles com base nos transcritos selecionados, separando eles em grupos, baseando-se nas características que possuem. A idéia básica consiste em colocar em um mesmo grupo o que é similar de acordo com algum critério pré-determinado (Linden, 2009).

Além disso, as Análises de agrupamento são exemplos de técnicas de aprendizado não supervisionado em que as categorias de classificação (os agrupamentos) são desconhecidas, mas observa-se recursos relacionados às categorias não observadas (Agresti, 2019). Portanto, essas análises visam encontrar algorítmos ótimos para agrupar grupos que são semelhantes de acordo com algum critério, não existindo grupos pré-definidos nem quantos grupos se formarão.

Além disso, podem ser classificadas como de particionamento e métodos hierárquicos. Um método de particionamento constrói k *clusters*, ou seja, classifica os dados em k grupos, que juntos atendem aos requisitos de uma partição (cada grupo deve conter pelo menos um objeto e cada objeto deve pertencer a exatamente um grupo) (Kaufman e

Rousseeuw, 2019), um exemplo de um algoritmo desse tipo é o própio PPCLUST. Já os algoritmos hierárquicos não constroem uma única partição com *clusters* k, mas lidam com todos os valores de k na mesma execução, ou seja, a partição com $k = 1$ (todos os objetos estão juntos no mesmo *cluster*) faz parte dos resultados e também a situação com $k = n$ (cada objeto forma um *cluster* separado com apenas um elemento). Intermediariamente, todos os valores de $k = 2, 3,...,n - 1$ são abordados em um tipo de transição gradual, com a diferença que entre $k = r$ e $k = r + 1$ os agrupamentos de r se dividem para obter os agrupamentos de $r + 1$ (ou, de maneira diferente), dois dos *clusters* $r + 1$ se combinam para produzir o *cluster* r (Kaufman e Rousseeuw, 2019). Neste trabalho foi utilizado um algorítmo hierárquico, que pode ser subdividido de duas maneiras, descritas a seguir e ilustradas na Figura 4:

- Aglomerativos: Cada item de dados é considerado como um grupo individual, e grupos são recursivamente fundidos de acordo com algum critério de proximidade até produzir um bom agrupamento no final do processo (Miranda, 2012).
- Divisivos: Inicialmente, o conjunto de todos os dados é considerado como sendo um único grupo e, em seguida ele é recursivamente dividido de acordo com algum critério de proximidade para produzir um bom agrupamento final (Miranda, 2012).

Técnicas Aglomerativas e Divisivas constroem sua hierarquia em direção oposta e podem produzir resultados distintos (Kaufman e Rousseeuw, 2019).

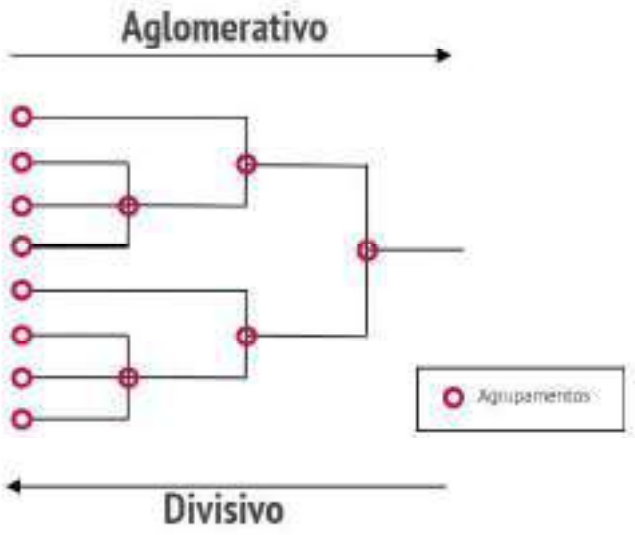


Figura 4 – Ilustração de agrupamento hierárquico aglomerativo e divisivo.

Neste trabalho foi utilizado um algorítmo Hierárquico Divisivo. A medida de similaridade utilizada foi a distância euclidiana, que além de ser a mais utilizada em dados

em escala LOG, também reduz a variabilidade desses dados por se tratar de uma medida quadrática.

Distância Euclidiana:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}, \tag{2.13}$$

p : número de variáveis;

x : i-esimo transctiro para o agrupamento dos transcritos ou o j-ésimo caso ou controle para o agrupamento de casos ou controles;

y : i-esimo transctiro para o agrupamento dos transcritos ou o j-ésimo caso ou controle para o agrupamento de casos ou controles, difetente de x.

O agrupamento hierárquico é normalmente apresentado em um dendrograma e para um visualização ainda melhor desses resultados, em genética principalmente, é comumente usado a apresentação desses resultados em *heatmaps* com dendogramas (na parte superior e ao lado) geralmente, como na Figura 5, retratando o processo de mesclagem dos *clusters* em função da métrica de proximidade. A associação dessas duas ferramentas gráficas possibilita melhor visualização das magnitudes dos efeitos. Além disso, uma separação por grupos, também é possivel, como é o caso da separação entre casos e controles.

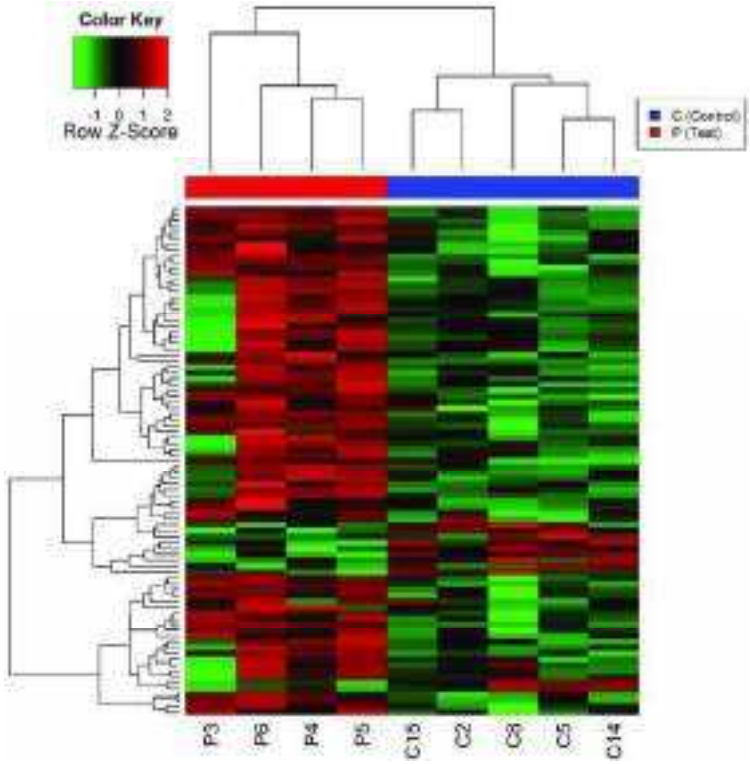


Figura 5 – Exemplo de *Heatmap* de perfil de expressão gênica.

2.3.5 Anotação genômica

A anotação genômica, do inglês *genomic annotation*, pode ser definida como o processo de descoberta de componentes importantes do genoma, principalmente genes e seus produtos, adicionando a eles análises e interpretações necessárias para se extrair sua importância biológica e colocando-os no contexto dos processos biológicos, pressupondo para isso, uma forte correlação entre coexpressão e a relação funcional (Stein, 2001).

Tipicamente, essa análise envolvem o denominado *pathway analysis* (análise de caminhos), que tem como objetivo identificar proteínas vinculadas a transcritos que são relacionadas dentro de uma função biológica ou que ligam uma função biológica a outra. Esse procedimento permite encontrar processos celulares distintos, doenças ou vias de sinalização estatisticamente associadas à seleção de genes diferencialmente expressos entre as duas amostras (Aguiar e Severino, 2010).

A análise de caminhos precisa de uma base de conhecimento com redes de coleta e interação de caminhos. O conteúdo, a estrutura e a funcionalidade das coleções de caminhos geralmente variam em diferentes bases de dados. Sendo aqui utilizada a coleção de caminhos da base `Reactome` (reactome.org, 2020).

A base `Reactome` é um recurso com curadoria on-line para dados de caminhos humanos, com base no qual pode-se inferir reações equivalentes em várias espécies não humanas. É utilizado também como recurso de aprendizado além de ser uma ferramenta computacional para auxiliar na interpretação de microarranjos e conjuntos de dados semelhantes em larga escala (Vastrik e outros, 2007) (reactome.org, 2020).

A base de dados `Reactome` tem como principal característica que as identificações de cada gene (colunas IDs, UniProt, símbolos genéticos ou IDs ChEBI) são mapeados para a análise de caminhos, análises de super-representação e da análise de conectividade entre moléculas.

A super-representação diz respeito ao cálculo de uma lista de proteínas, obtidas através de algumas das identificações supracitadas que contém mais anotações em cada caminho da base `Reactome` do que seria esperado. A análise de super-representação envolve um teste estatístico que utiliza a distribuição hipergeométrica e determina se certas vias da base `Reactome` estão super-representadas (enriquecidas) nos dados enviados.

Por meio desse cálculo da distribuição hipergométrica, um P-valor é obtido e é cálculado um ajuste para comparações múltiplas por Dunn-Sidak (Abdi, 2007) [6] que se assemelha bastante com o método por FDR, visando controlar o Erro do Tipo I.

[6] Ajuste derivado do método de Bonferroni (Bland e Altman, 1995)

Esse ajuste se dá seguindo a seguinte equação:

$$p' = 1 - (1 - p)^r, \tag{2.14}$$

em que p' representa o P-valor ajustado por Dunn-Sidak, p representa o P-valor obtido pela distribuição hipergeométrica e r é igual ao número de hipóteses (Biocomputing, 2007).[7]. Além disso, a medida *gene ratio* (razão de genes ou razão gênica) definida pela porcentagem do total de genes diferencialmente expressos é também utilizada na análise de anotação.

[7] Dunn-Sidak é derivado da correção de Bonferroni, este que possui $p' = \frac{p}{k}$

3 Resultados

Procedimentos no nível de *probe*, conduzidos por (Lessard e outros, 2013), envolveram a eliminação de dados de baixa qualidade com a filtragem dos transcritos que foram expressos em menos de 10% dos indivíduos (limiar de detecção: P <0,05) e probes com taxas diferenciais ausentes entre casos e controles (P<0,001 pelo teste exato de Fisher). Em seguida, os dados foram transformados para uma escala logarítmica de base 2 e foi realizada uma Normalização Quantílica. Tais procedimentos levaram à seleção de 15.063 (dos 48.803 transcritos incluídos no chip) para os 190 casos de pSS e 32, do grupo controle utilizados nas análises no nível de expressão gênica, cujos resultados são apresentados a seguir.

3.1 Análise exploratória

Os boxplots das Figuras 6 e 7 apresentam os dados de expressão gênica em escala original e em escala log2 por grupo. Observa-se acentuada assimetria na escala original em ambos os grupos, o que evidencia a necessidade da transformação em escala logarítima. Como tais gráficos incluem expressões dos 15.063 transcritos selecionados não é possível se observar diferenças nas distribuições dos grupos de casos e controles.

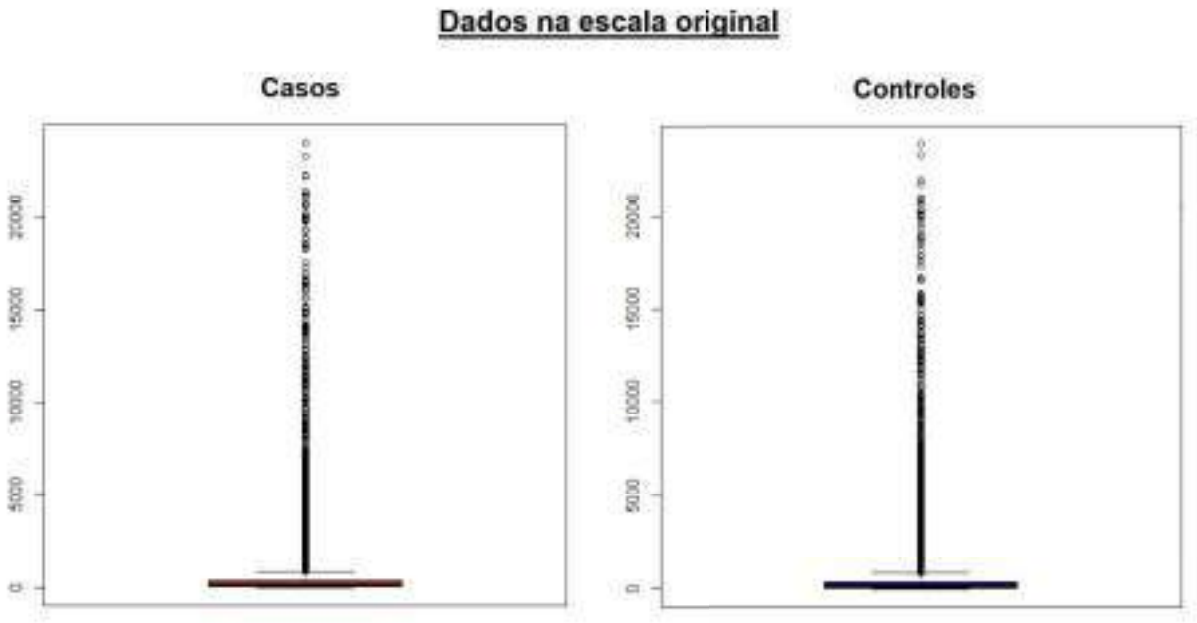


Figura 6 – Distribuição de expressões gênicas de casos e controles em escala original.

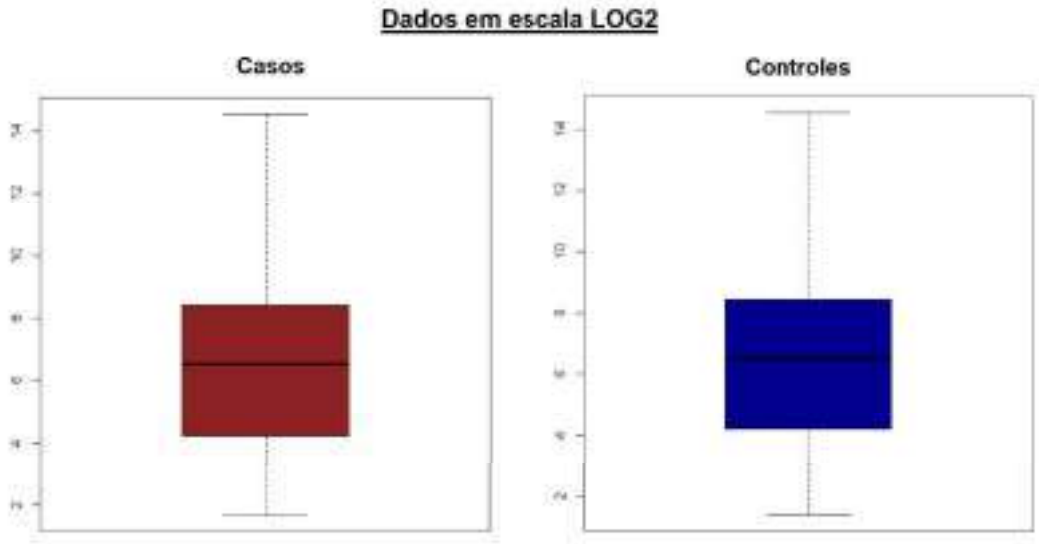


Figura 7 – Distribuição de expressões gênicas de casos e controles em escala log2.

A distribuição das médias dos transcritos (Figuras 8 e 9) destaca que apenas uma pequena parcela dos transcritos possui expressões médias elevadas em ambos os grupos.

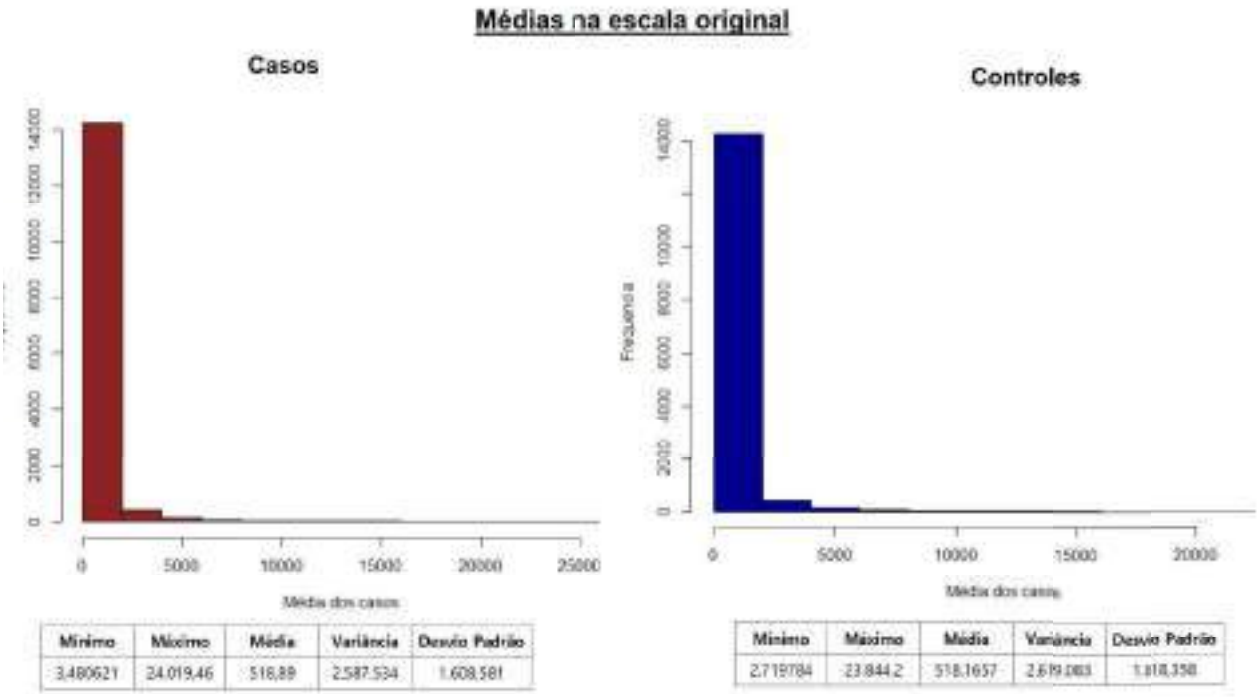


Figura 8 – Distribuição das expressões médias (de todos os transcritos) de casos e controles em escala original.

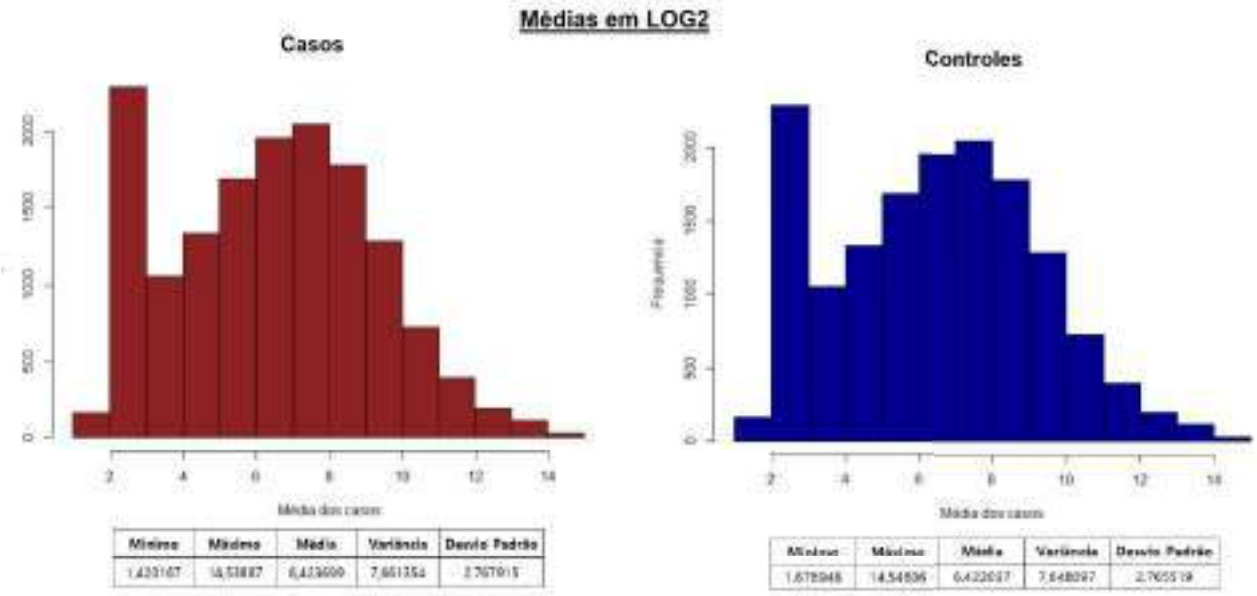


Figura 9 – Distribuição das expressões médias (por transcrito) de casos e controles em escala log2.

A Figura 10 apresenta as distribuições de FC dos 15.063 transcritos, definidas pela razão de médias e de medianas (de casos por controles), respectivamente. Para ambas as medidas, observa-se que as razões FC se concentram em valores muito próximos de 1 para a vasta maioria dos transcritos.

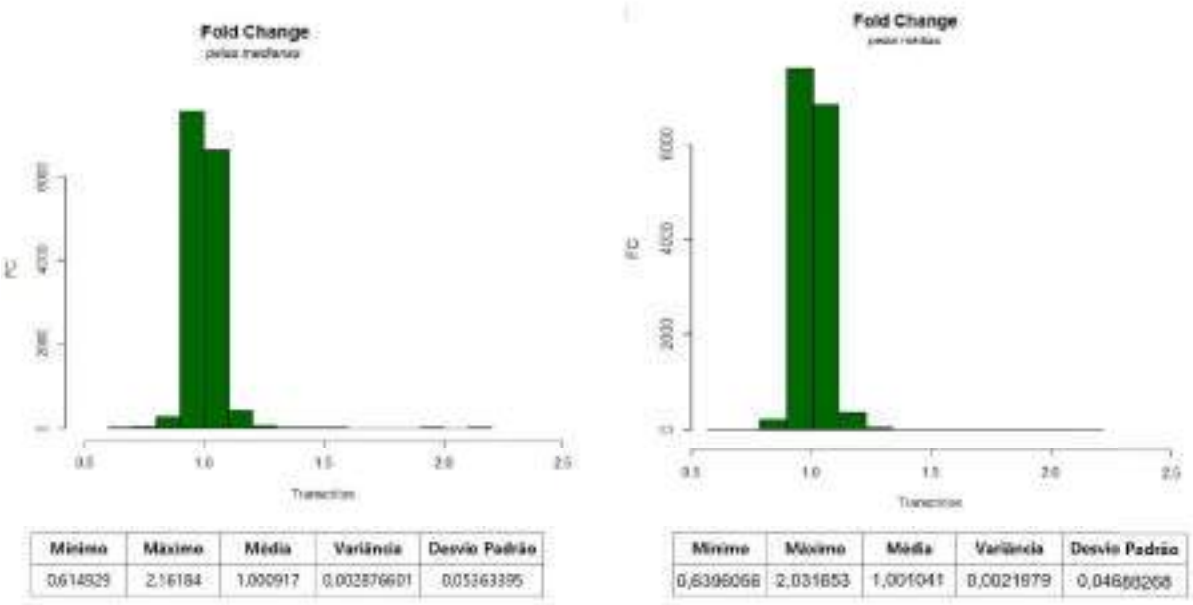
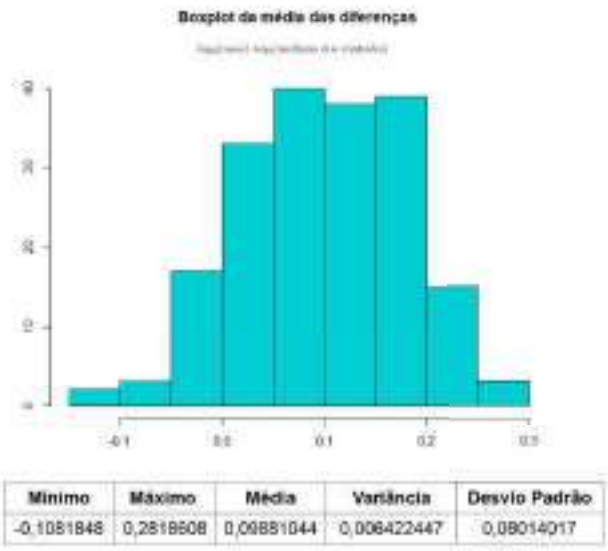


Figura 10 – Distribuição de FC por médias e de FC por medianas.

Para a implementação do algorítmo PPCLUST, para cada transcrito, foram utilizadas as expressões gênicas dos casos subtraídas pela mediana das expressões do grupo controle em escala log2 (escala que os dados foram disponibilizados), chegando então a: $(log_2(X^{casos}{}_i) - log_2(\theta_{controles}))$. A Figura 11 apresenta a distribuição das médias desses valores.



Mínimo	Máximo	Média	Variância	Desvio Padrão
-0,1081848	0,2818508	0,09881044	0,006422447	0,08014017

Figura 11 – Distribuição dos valores utilizados para o PPCLUST.

A Figura 12 apresenta a distribuição de p-valores de testes de Normalidade (Shapiro-Wilks) para os 15.063 transcritos em log2 . Considerando-se um nível α= 0,05, observou-se que 8.469 (aproximadamente 44%) dos transcritos apresentaram evidência contra a hipótese nula de normalidade. Tais resultados apontam para a necessidade da utilização das abordagens por Wilcoxon-Mann-Whitney e PPCLUST na identificação de TDEs.

Figura 12 – Distribuição de P-valores do Testes de normalidade Shapiro-Wilk.

3.2 Análise de perfil de expressão gênica

Após as análises exploratórias preliminares iniciaram-se os procedimentos para a identificação de TDEs sob as três diferentes abordagens (Tabela 2).

Para a escolha do nível α adequado no algorítmo PPCLUST, de modo que o número de grupos definidos fosse estável, foram testados níveis α de $10e^{-1}, 10e^{-4}, 10e^{-8}$ e $10e^{-12}$, que resultaram, respectivamente, em 13,13,14 e 13 grupos (Figura 13).

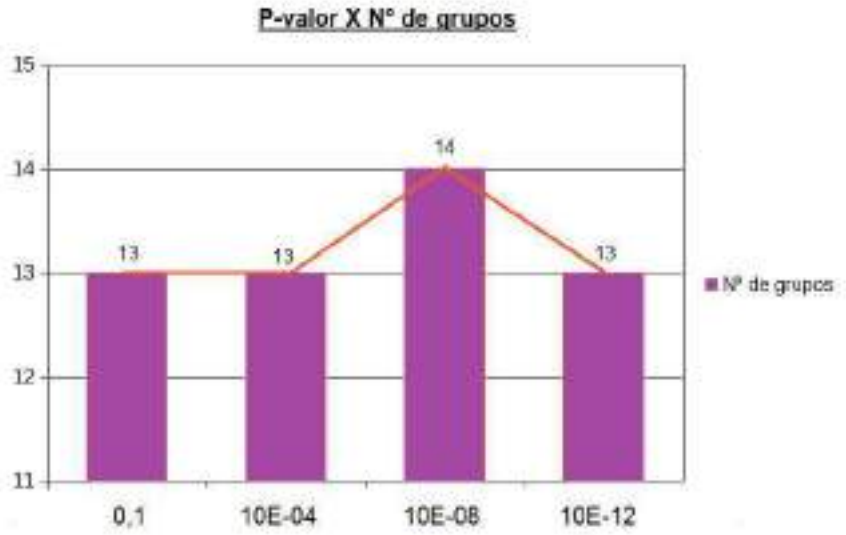


Figura 13 – Número de grupos definidos pelo algorítimo PPCLUST por P-valor.

Atingiu-se uma certa estabilidade em 13 grupos com o menor dos níveis α's testados de $10e^{-1}$. O agrupamento resultante é apresentado na Figura 14. Destaca-se que o grupo 0 foi criado para indicar transcritos que não puderam ser incluídos em nenhum dos grupos criados (são sobras do processo de agrupamento).

Grupo	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Quantidade transcritos	35	4	92	647	17	1865	2616	3586	2964	1976	891	277	89	4

Trasncritos que não se encaixam em nenhum grupo

Grupos extremos

Figura 14 – Estrutura dos grupos obtidos pelo algorítmo PPCLUST.

Após o processo de agrupamento, transcritos dos grupos extremos 1 a 4 e 10 a 11 foram considerados TDEs, totalizando 2.021 transcritos.

A separação em grupos extremos e não extremos foi arbitrária e baseada na visualização da distribuição dos transcritos que apresentassem menor quantidade nos grupos considerados extremos que nos demais grupos, sugerindo que esses seriam TDEs.

Para efeitos de consistência com as outras abordagens, uma filtragem adicional por magnitude de efeito for efetuada, excluindo-se transcritos com FC de medianas entre $(1.5)^{-1}$ e 1.5.

Os números de TDEs identificados por cada uma das abordagens, incluindo seus respectivos cortes por magnitude e correções por testes múltiplos são apresentados na Tabela 2. Destaca-se que, cerca de 50 % dos transcritos apresentou significância genômica (p-valores corrigidos pela taxa de FDR $< 0{,}05$) e apenas 10 % apresentou magnitudes de efeitos maiores, definidos por FC $< (1.5)^{-1}$ ou FC > 1.5.

Tabela 2 – Número de transcritos selecionados para cada etapa das abordagens de identificação de TDEs.

Etapa/Abordagem	Teste t	Wilcoxon-Mann-Whitney	PPCLUST
Após a realização do teste	4.214	4.340	2.021
Com correção pela taxa FDR	2.133	2.165	-
Com corte por magnitude (*fold change*)	232	257	295

O diagrama de Venn (Figura 15) abaixo apresenta o as relações entre os conjuntos de TDEs, identificados pelas 3 abordagens utilizadas. Observa-se que 318 transcritos foram identificados por qualquer das 3 técnicas (União dos conjuntos). Destaca-se que 50% destes (158 transcritos) foram detectados pelas 3 diferentes abordagens, enquanto que 35% foram detectados por apenas uma. Apesar da abordagem baseada no teste t detectar um menor número de TDEs que a abordagem baseada no teste de Wilcoxon-Mann-Witney, apenas 3 transcritos foram identificados unicamente pelo último. Cinquenta e seis transcritos foram detectados unicamente pela abordagem PPCLUST e 53, unicamente pela abordagem baseada no teste t.

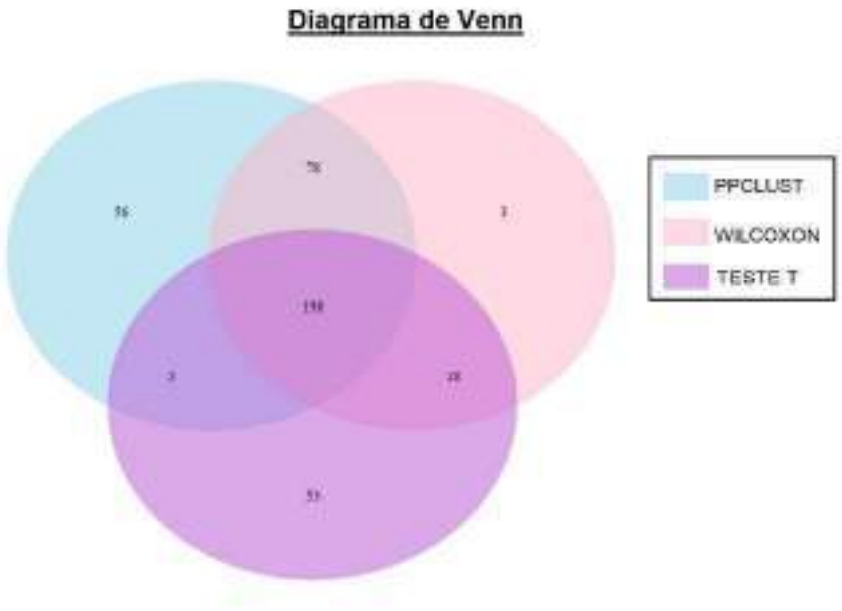


Figura 15 – Diagrama de Venn das 3 listas de transcritos (resultantes de testes paramétricos, não paramétricos e PPCLUST).

Em sequência, são apresentados conjuntos de gráficos para cada uma das cinco listas de TDEs, descritas a seguir (completar):

- Lista **L1**: TDEs identificados por teste t, corrigidos pela taxa de FDR e com $FC_{\text{medias}} > 1,5$ ou $FC_{\text{medias}} < 1,5^{(-1)}$;
- Lista **L2**: TDEs identificados por teste Wilcoxon-Mann-Witney, corrigidos pela taxa de FDR e com $FC_{\text{medianas}} > 1,5$ ou $FC_{\text{medianas}} < 1,5^{(-1)}$;
- Lista **L3**: TDEs identificados por algorítmo PPCLUST e com corte pelo $FC_{\text{medianas}} > 1,5$ ou $FC_{\text{medianas}} < 1,5^{(-1)}$;
- Lista **L4**: Interseção das listas L1, L2 e L3;
- Lista **L5**: União das listas L1, L2 e L3.

As Figuras 16 a 20 apresentam os perfis de expressão gênica para os TDEs selecionados para as listas **L1** a **L5**. Tais figuras evidenciam nitidamente a existência de grupos de transcritos super-expressos (em vermelho) e sub-expressos (em verde) no grupo de casos, comparado ao grupo de controles. Pode-se observar uma concentração maior de transcritos super-expressos entre os detectados como TDEs pelas abordagens 1 e 2 e pela intersecção (**L4**). A abordagem que envolve o algorítmo PPCLUST detectou mais trasncritos sub-expressos que as outras abordagens.

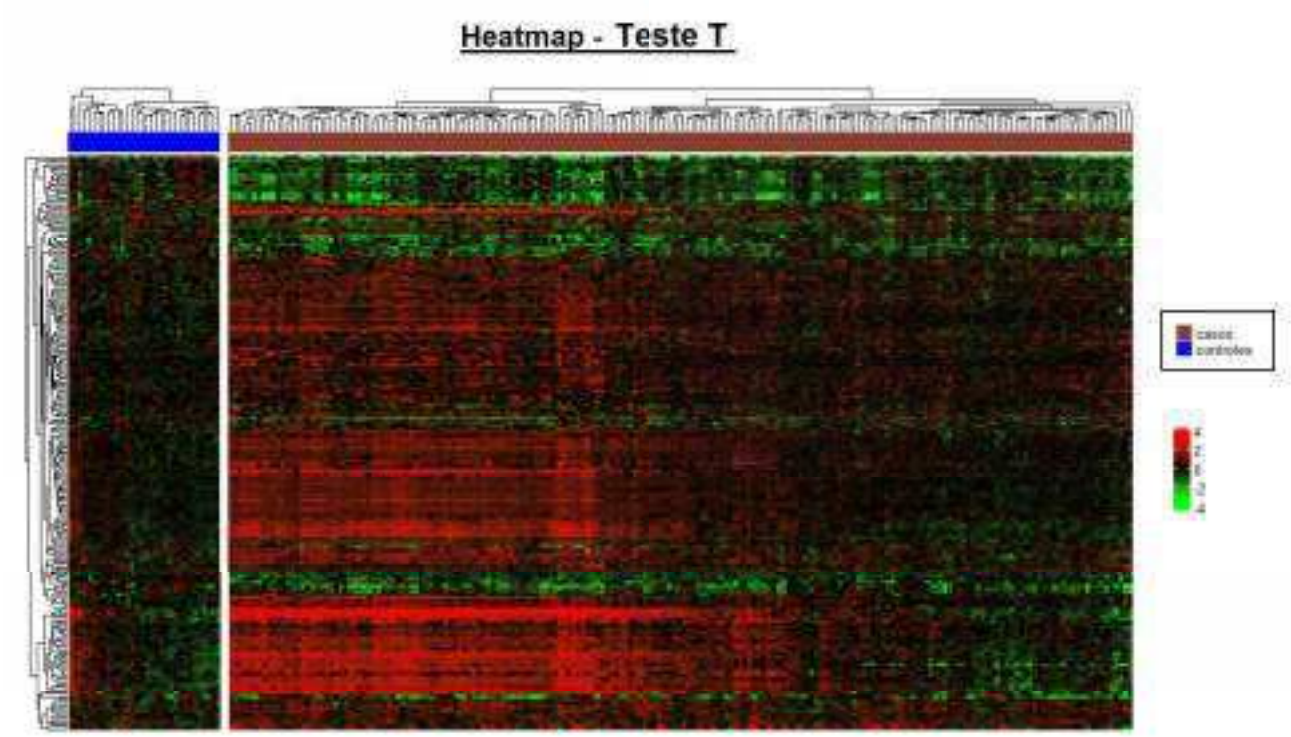


Figura 16 – *Heatmap* das expressões gênicas em escala log2 para a Lista L1 (testes t corrigidos por FDR e FCs por médias)

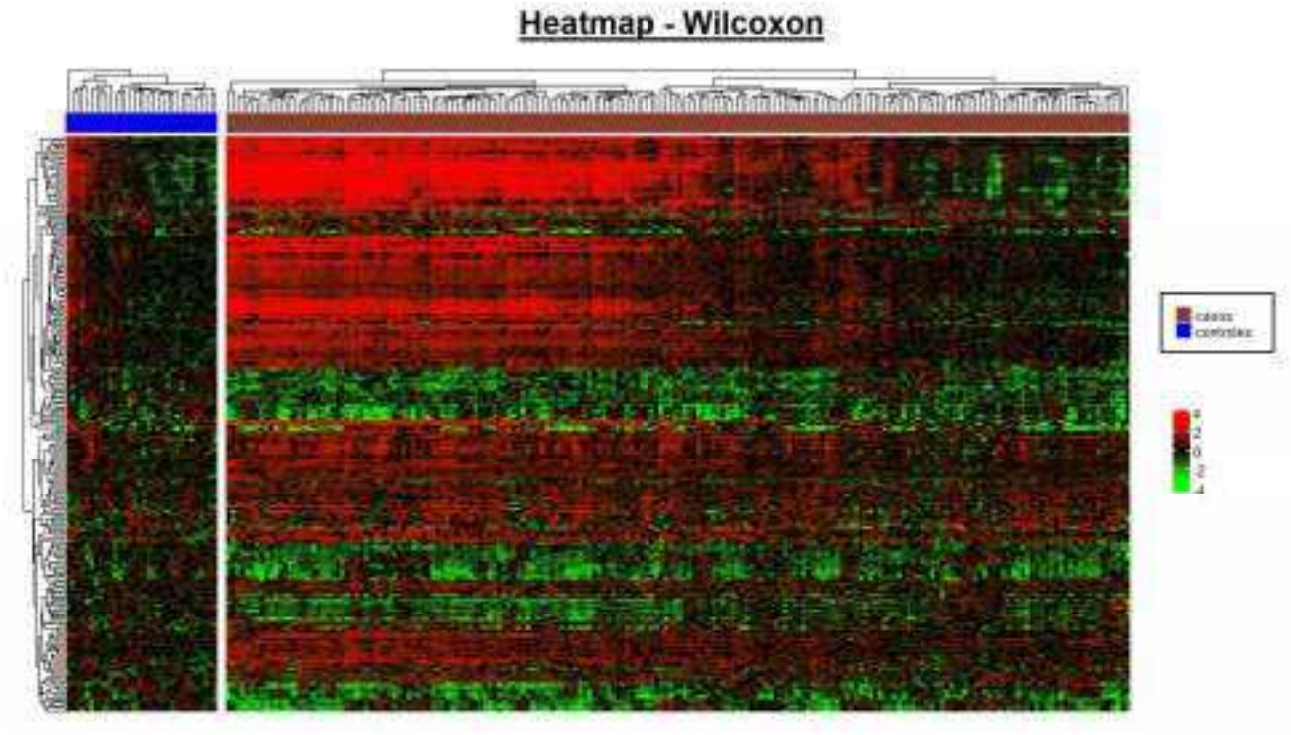


Figura 17 – *Heatmap* das expressões gênicas em escala log2 para a Lista L2 (testes Wilcoxon-Mann-Whitney corrigidos por FDR e FCs por medianas)

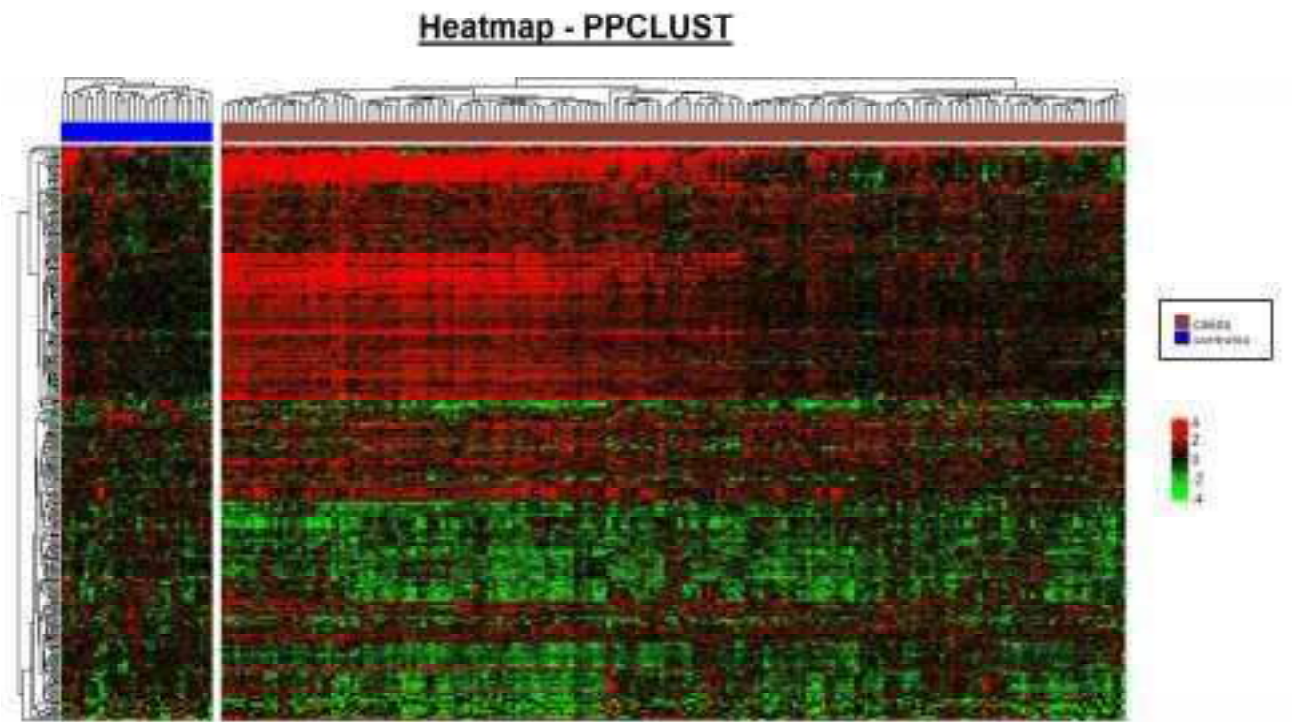


Figura 18 – *Heatmap* das expressões gênicas em escala log2 para a Lista L3 (algorítmo PPCLUST e FCs por medianas)

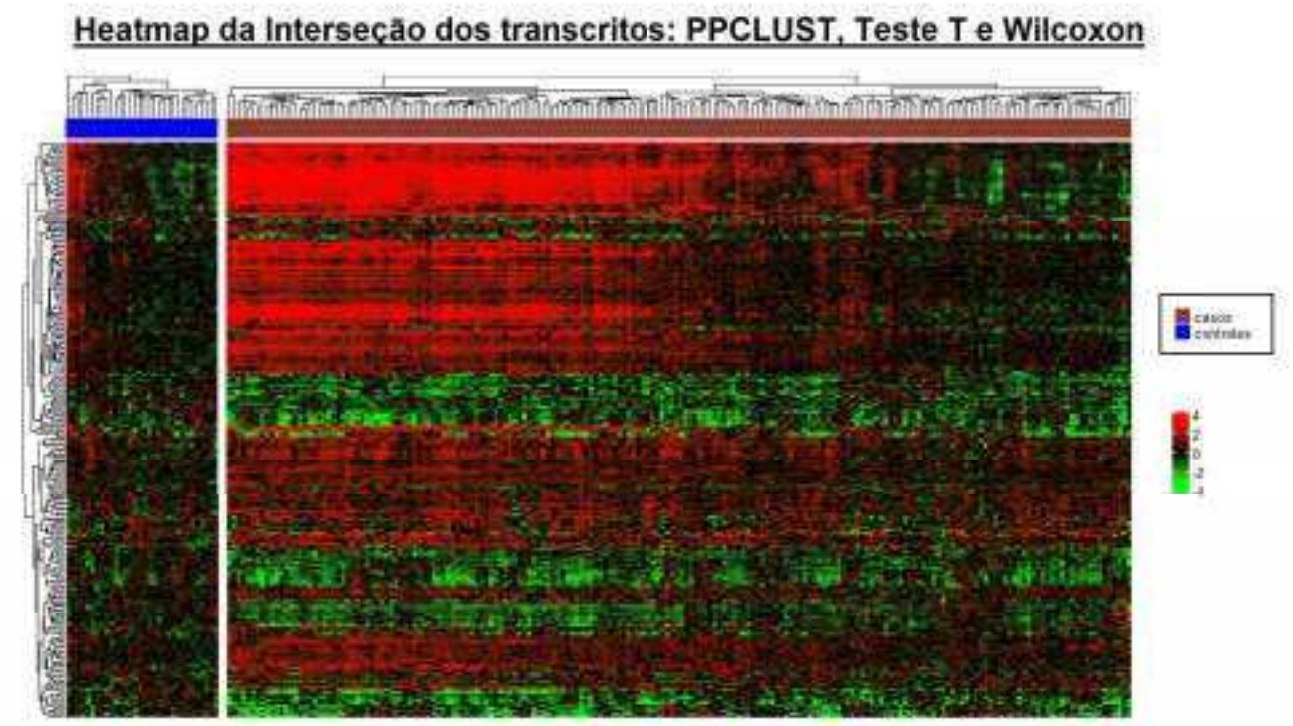


Figura 19 – *Heatmap* das expressões gênicas em escala log2 para a Lista L4 (Intersecção das abordagens)

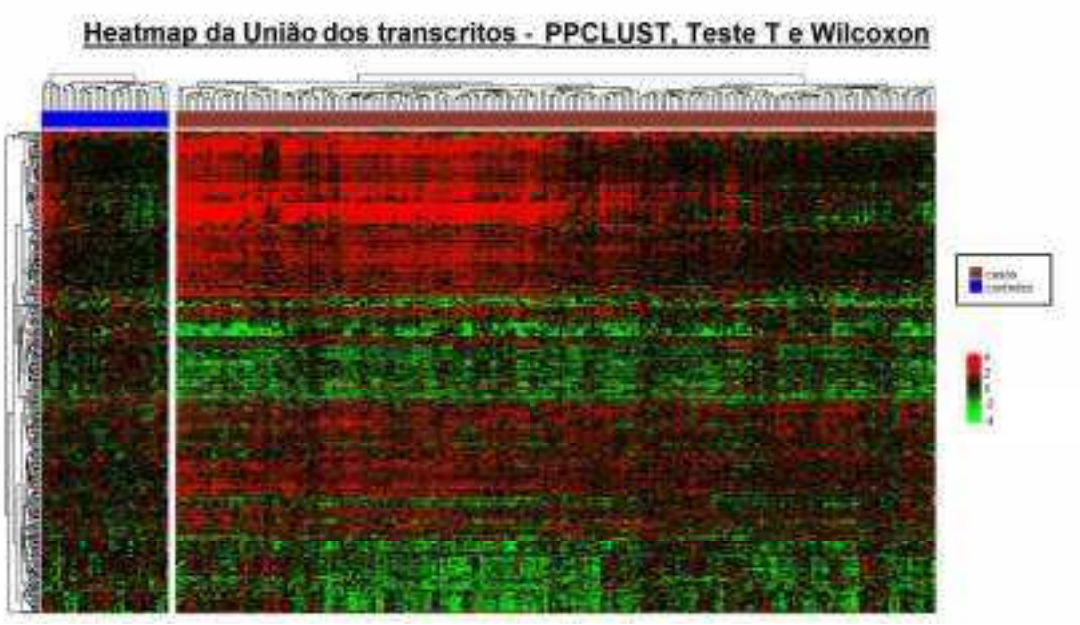


Figura 20 – *Heatmap* das expressões gênicas em escala log2 para a Lista L5 (União das abordagens)

A Tabela 3 apresenta os 10 genes para maiores FCs entre a união das 3 abordagens (**L5**). Destaca-se entre estes, genes de interteron (*IFI44L, IFI27, IFI44, IFIT1*) ou induzidos por diferentes tipos de interferons bem como outros genes, relacionados a função de imunidade e defesa contra vírus e bactérias.

Tabela 3 – Genes mais sub/super-expressos na lista de união dos TDEs detectados pelas 3 abordagens (**L5**).

GENE	**FUNÇÃO**	**SUB/ SUPER-EXPRESSO NOS CASOS**
CMPK2	Ligada a resposta celular ao lipopolissacarídeo. O lipopolissacarídeo é um dos componentes principais da membrana exterior da bactéria gram-negativa .	Super-expresso
RSAD2	É induzível por intelferon tipo I e tipo II e desempenha um papel importante no estado antiviral celular. Pode inibir uma ampla gama de vírus de DNA e RNA, incluindo citomegalovírus humano, vírus da hepatite C, vírus do Nilo Ocidental, vírus da dengue, vírus sindbis, vírus influenza A, vírus sendai, vírus estomatite vesicular e vírus da imunodeficiêcia humana (HIV-1).	Super-expresso
IFI44L	Um gene comum de ser encontrado em pacientes com doenças autoimunes que também é estimulado por interferon e é ligado a resposta de defesa imune e aresposta de defesa à vírus.	Super-expresso
IFI27	Pode desempenhar um papel na resposta vascular à lesão e é um novo modulador de respostas imunes inatas que regulam os receptores nucleares anti-inflamatórios.	Super-expresso
LAMP3	A expressão elevada de LAMP3 está associada à transição epitelial-mesenquimal (as células epiteliais perdem sua polaridade celular e adesão celular e ganham propriedades migratórias) e potencial metastático no câncer de esôfago.	Super-expresso
OTOF	Mutações nesse gene são uma causa de surdez recessiva não sindrômica neurossensorial.	Super-expresso
IFI44	É um paralogue do IFI44L que é estimulado por intelferon e ligado a respostas imunes, resposta às bactérias e a vírus.	Super-expresso
TUBB2A	Principal constituinte dos microtúbulos, que são estruturas proteicas que fazem parte do citoesqueleto nas células eucarióticas. Além de estar ligado ao processo do ócio celular mitótico e a migração de neurônios.	Super-expresso
EPSTI1	Gene de resposta ao interferon que foi originalmente identificado como um gene induzido por fibroblastos estremais no câncer de mama, mas novos estudos têm mostrado que ele está relacionado também a suscetibilidade à doença inflamatória crônica, lúpus eritematoso sistemático.	Super-expresso
IFIT1	Estimulado por interferon e ligado a resposta de defesa imune e a resposta de defesa à vírus.	Super-expresso

(Omim,2020) (Uniprot, 2020) (Genecards, 2020)

3.3 Identificação de sistemas biológicos e funções genéticas

Os resultados das análises de anotação são apresentados em conjuntos de 5 gráficos (para as listas **L1** a **L5**).

Gráficos de barras da razão gênica:

Os gráficos de barras do número de genes para cada uma das listas (Figuras 21 a 25) apresentam os sistemas mais significantes identificados com base no número de genes de cada lista (**L1** a **L5**). Utilizam gradação de cor de acordo com o P-valor ajustado que advém da hipergeométrica do teste que é realizado.

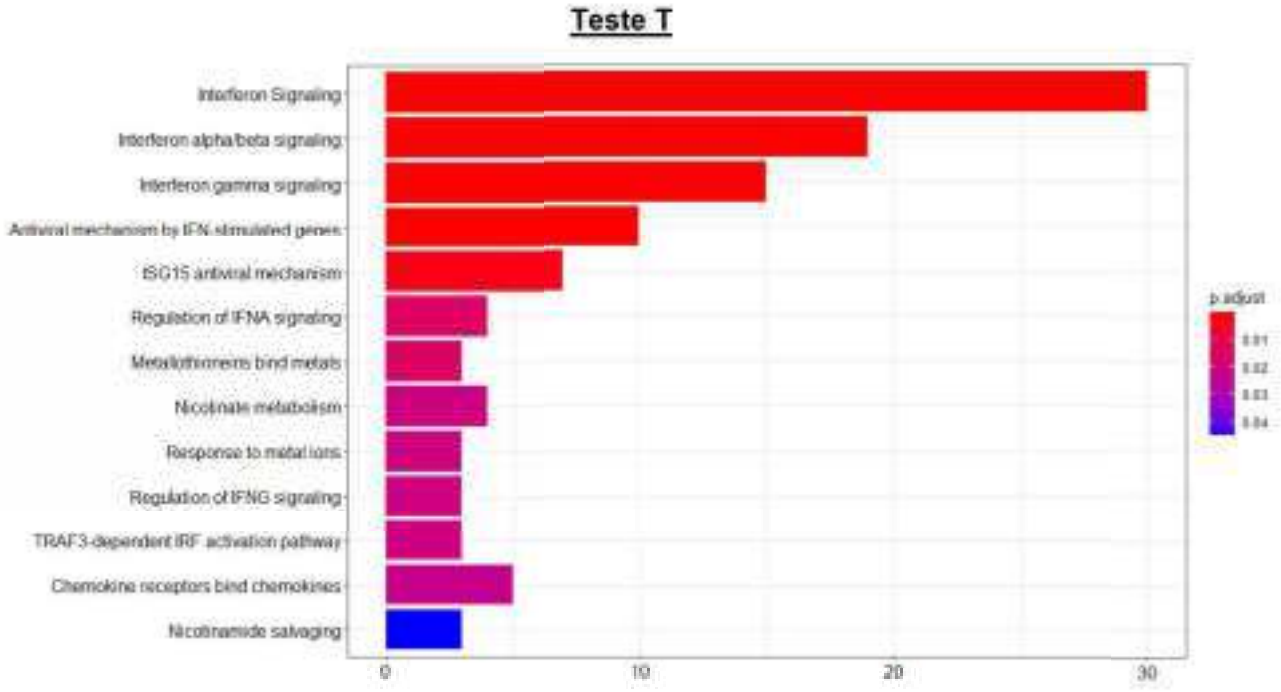


Figura 21 – Distribuição do número de genes da Lista **L1** detectados nos sistemas biológicos mais significantes.

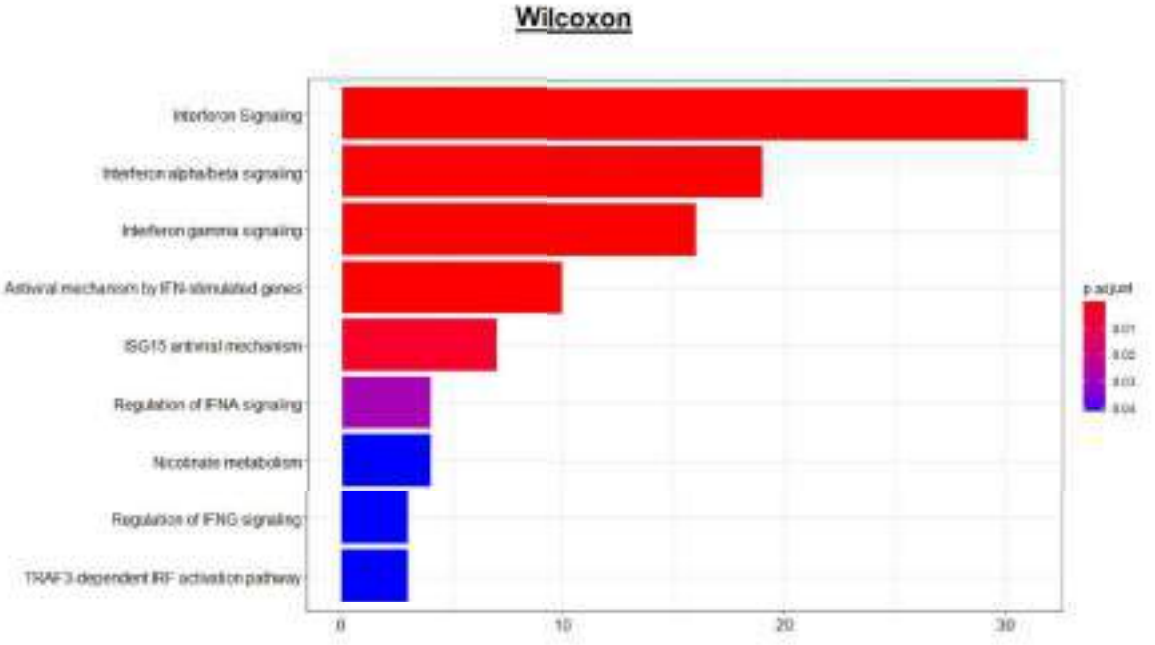


Figura 22 – Distribuição do número de genes da Lista **L2** detectados nos sistemas biológicos mais significantes.

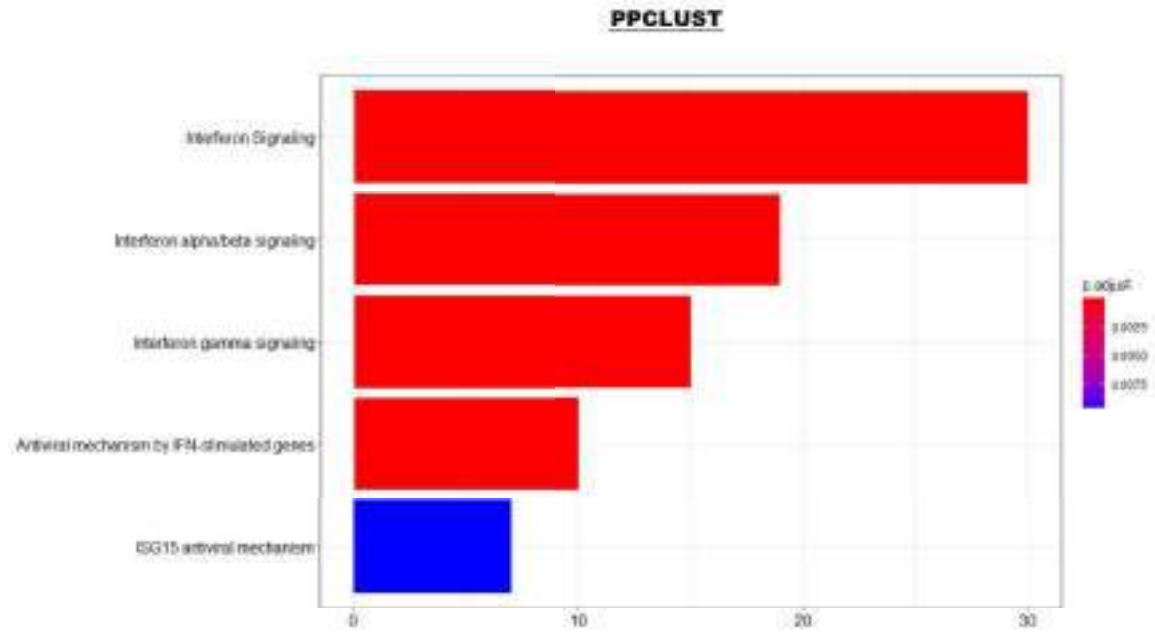


Figura 23 – Distribuição do número de genes da Lista **L3** detectados nos sistemas biológicos mais significantes.

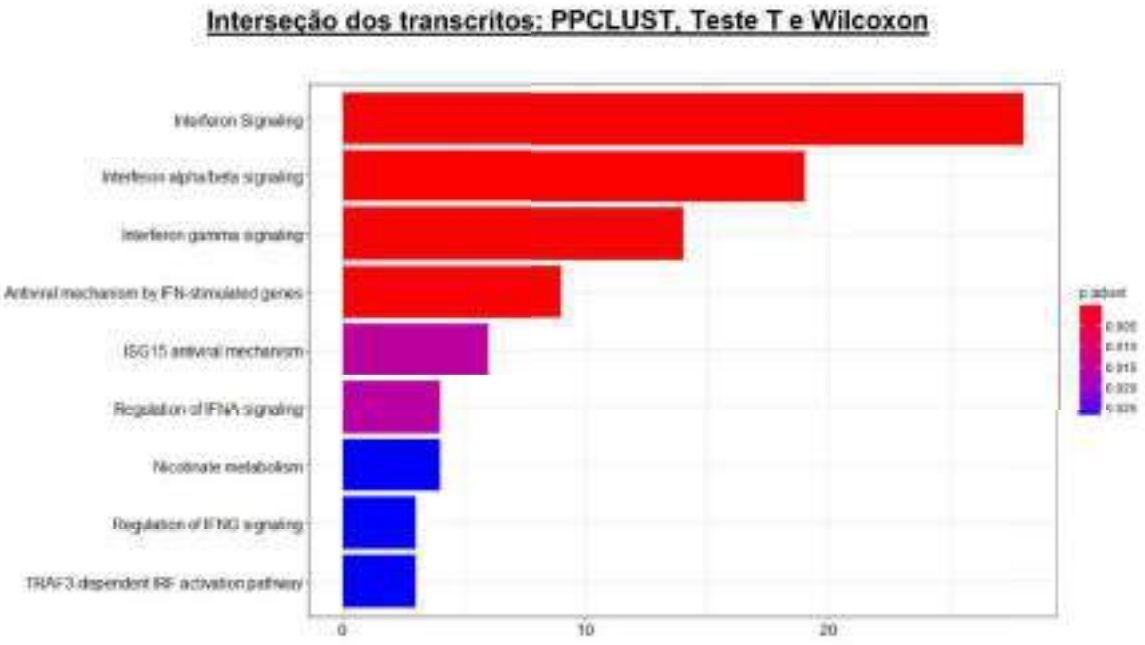


Figura 24 – Distribuição do número de genes da Lista **L4** detectados nos sistemas biológicos mais significantes.

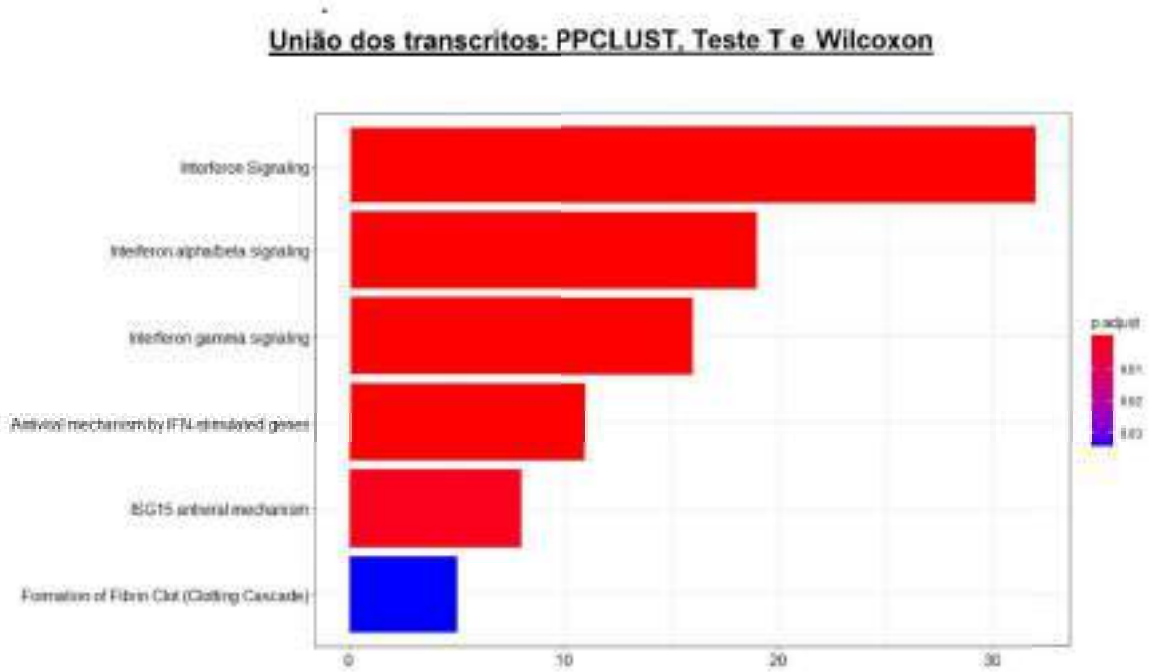


Figura 25 – Distribuição do número de genes da Lista **L5** detectados nos sistemas biológicos mais significantes.

Observa-se que a abordagem baseada no Teste T (**L1**) foi a que identificou o maior número de sistemas biológicos e a lista baseada no PPCLUST (**L3**) a que menos identificou, porém esta última com P-valores mais baixos que nas demais listas, indicando então uma maior consistência para esses sistemas nela identificados.

Todos os gráficos apresentam os sistemas de sinalização de interferon, de interferon alfa/beta e de interferon gama bem como o sistema de mecanismo antiviral por genes induzidos por interferon.

Além disso, nota-se que os sistemas identificados pela lista por Wilcoxon-Mann-Whitney (**L2**) e pela Interseção (**L4**) identificaram exatamente os mesmos sistemas, se diferenciando apenas em seu P-valor ajustado.

Ademais, partes dos sistemas identificados pela Lista **L1** foram visualizadas em todas as demais listas, sendo a lista da União, a Lista **L5**, a única em que foi identificado um sistema que não estava na **L1**, a formação de coágulos.

A descrição dos sistemas biológicos identificados pelos procedimentos das listas L1 a L5 estão na Tabela do Apêndice A.

Gráficos de pontos da razão gênica:

Os resultados fornecidos nas Figuras 21 a 25 podem ser alternativamente apresentados como nas Figuras 26 a 30, que apresentam os sistemas biológicos detectados pela medida razão gênica (*gene ratio*[1]). A gradação de cores indica a magnitude do P-valor e o tamanho do círculo, a quantidade de genes identificados como parte de cada um desses sistemas biológicos.

[1] Porcentagem do total de genes diferencialmente expressos.

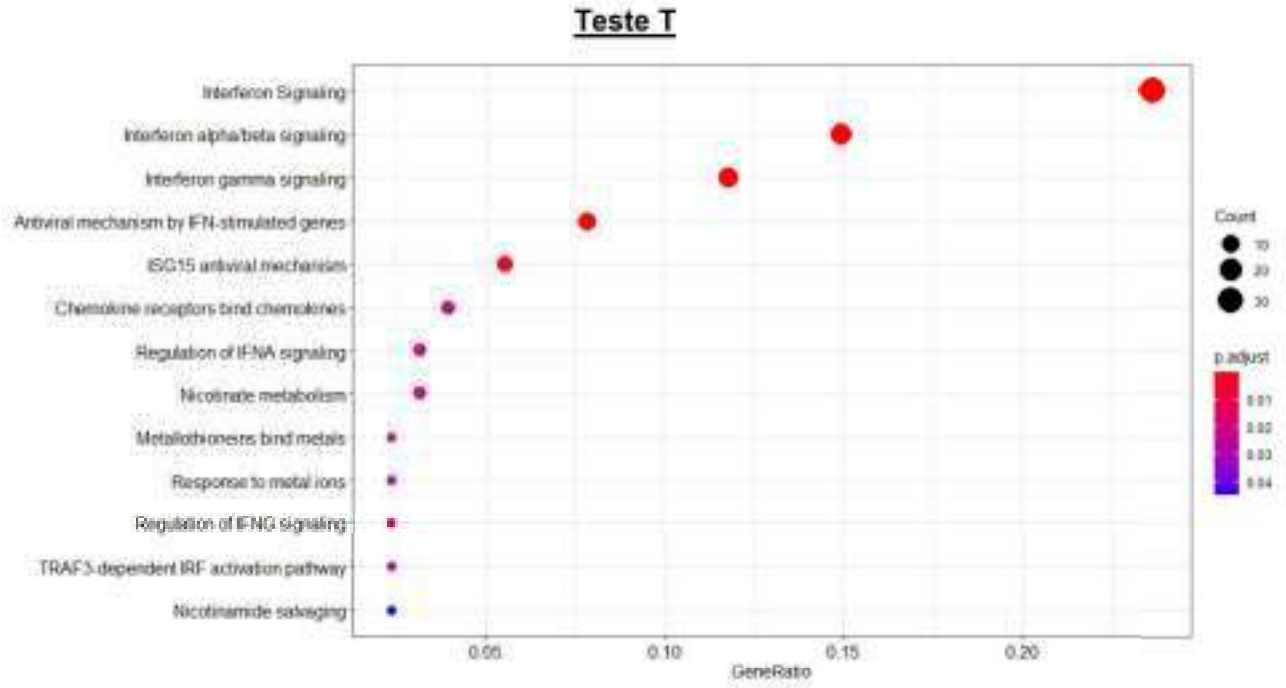


Figura 26 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista **L1**.

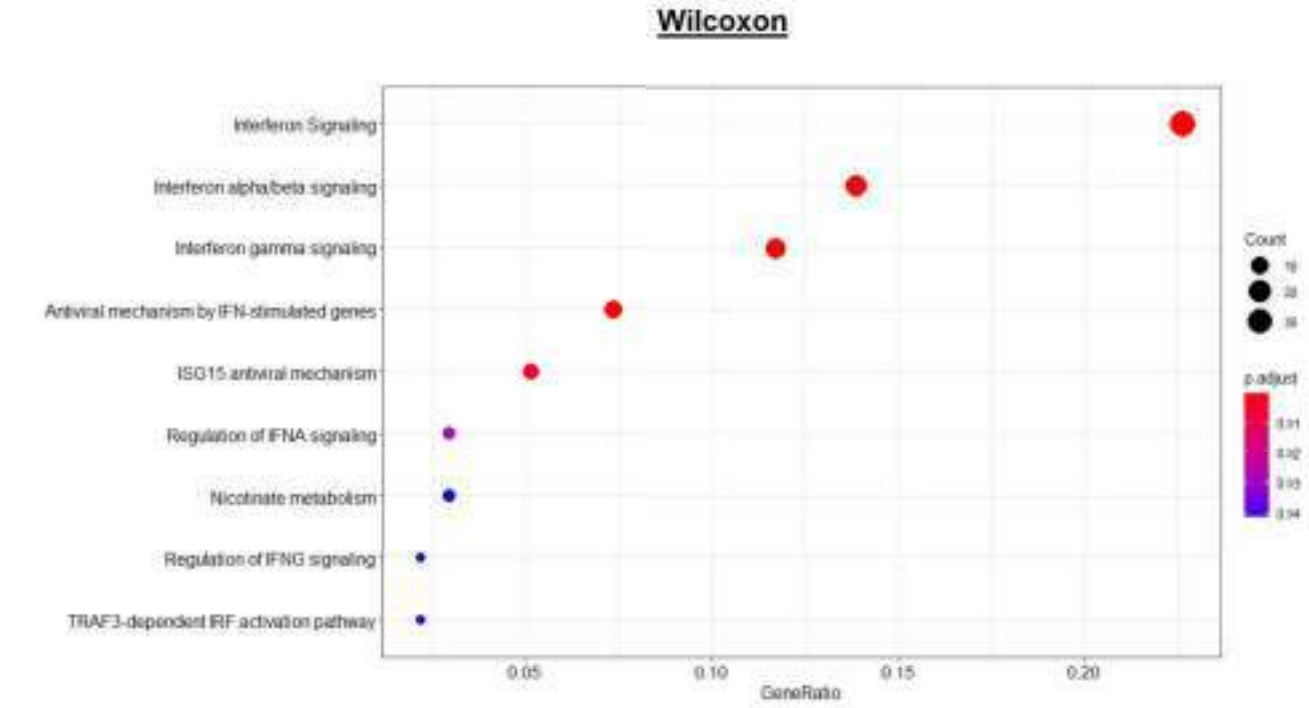


Figura 27 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista **L2**.

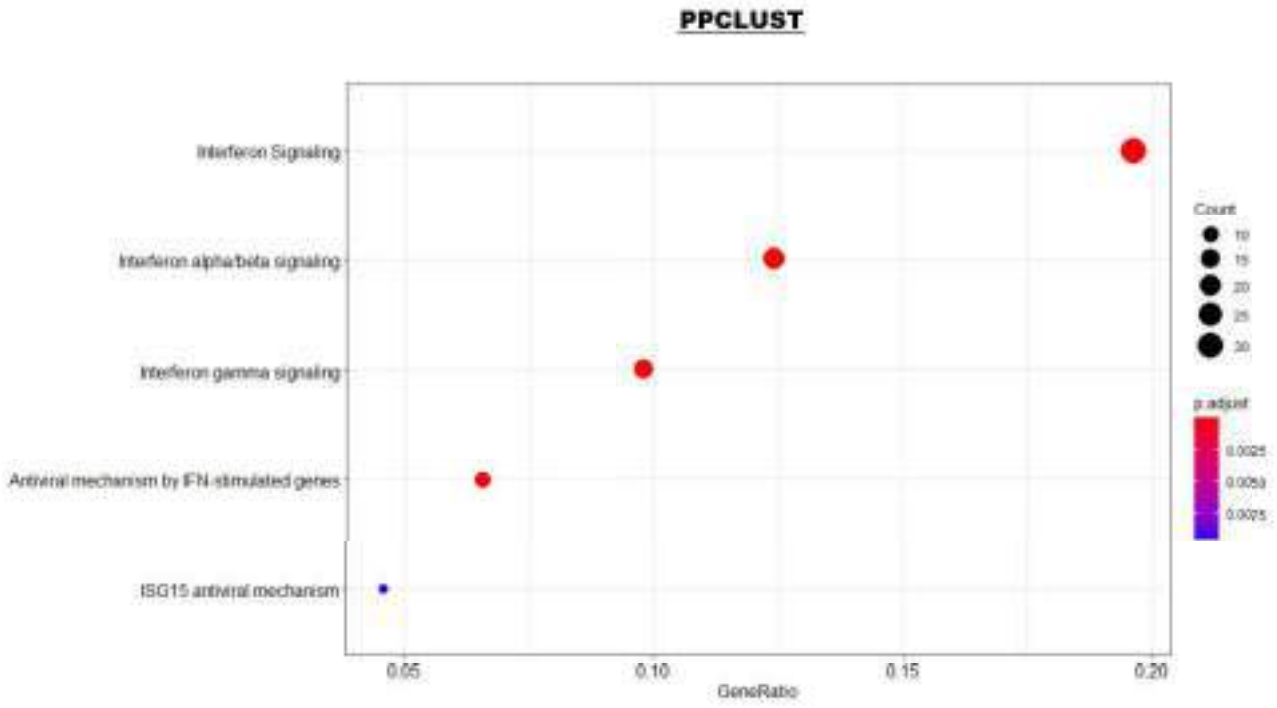


Figura 28 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista **L3**.

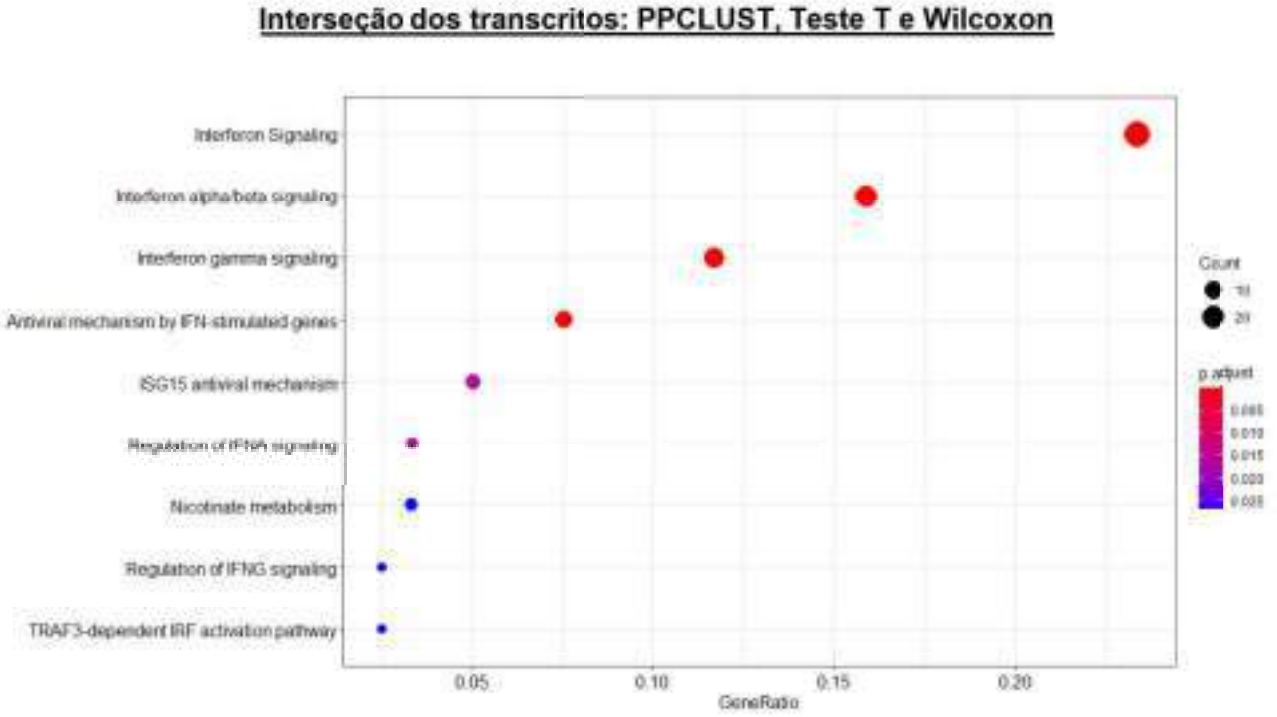


Figura 29 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista **L4**.

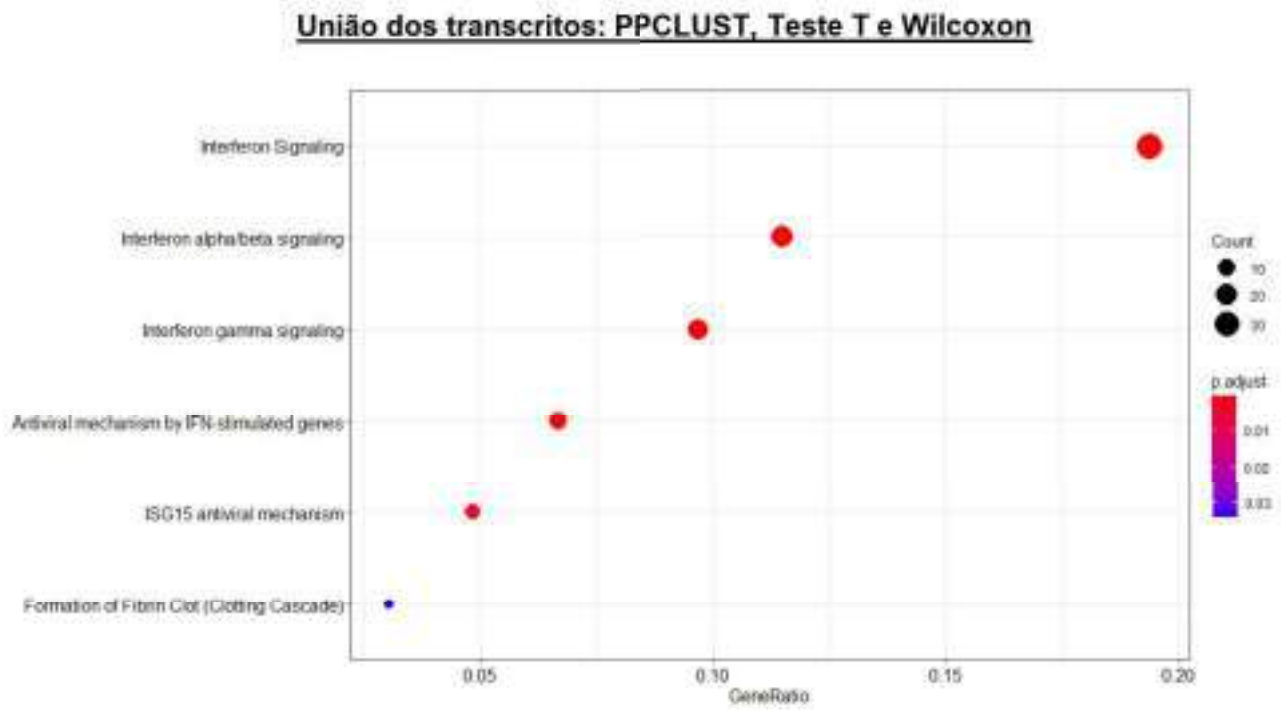


Figura 30 – Distribuição da razão gênica dos sistemas biológicos mais significantes identificados na Lista **L5**.

Diagramas de conexão de sistemas (*emapplots*):

São mapas de anotação que organizam os sistemas biológicos em uma rede com bordas conectando o conjunto desses sistemas quando relacionados. As Figuras 31 a 35 apresentam ligações entre sistemas biológicos mais significantes identificados nas 5 listas de TDEs. A gradação de cores indica a magnitude do P-valor e o tamanho do círculo, a quantidade de genes identificados como parte de cada um desses sistemas biológicos.

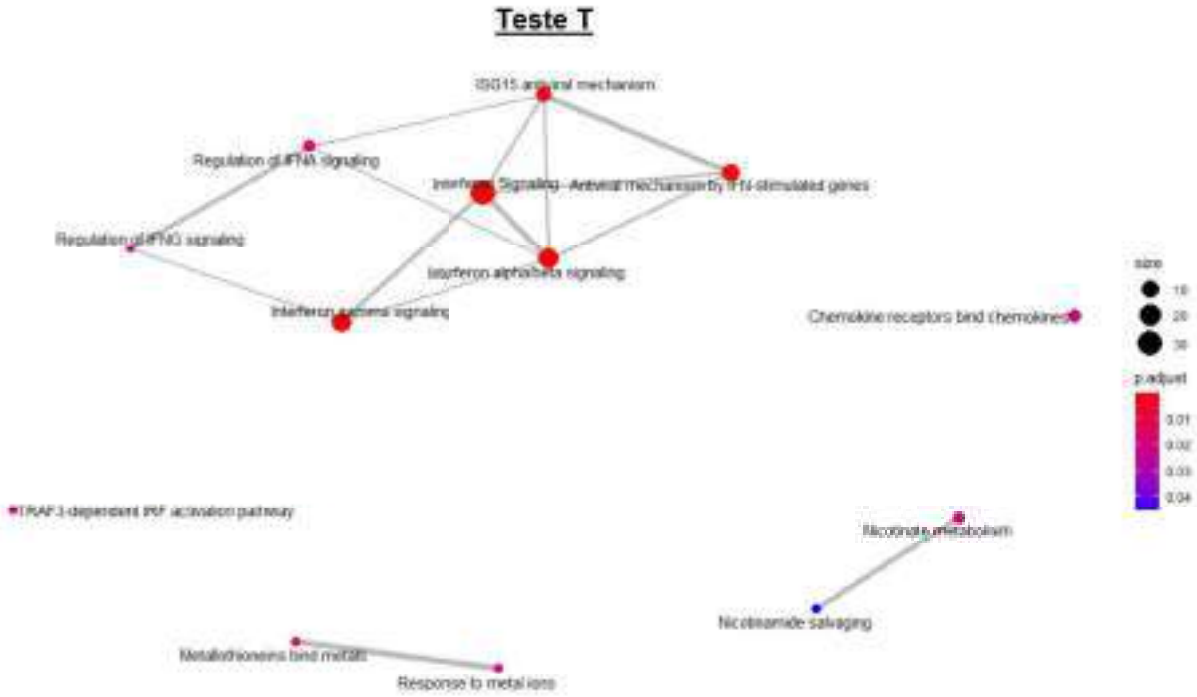


Figura 31 – Diagrama das ligações entre os sistemas biológicos identificados na Lista **L1**.

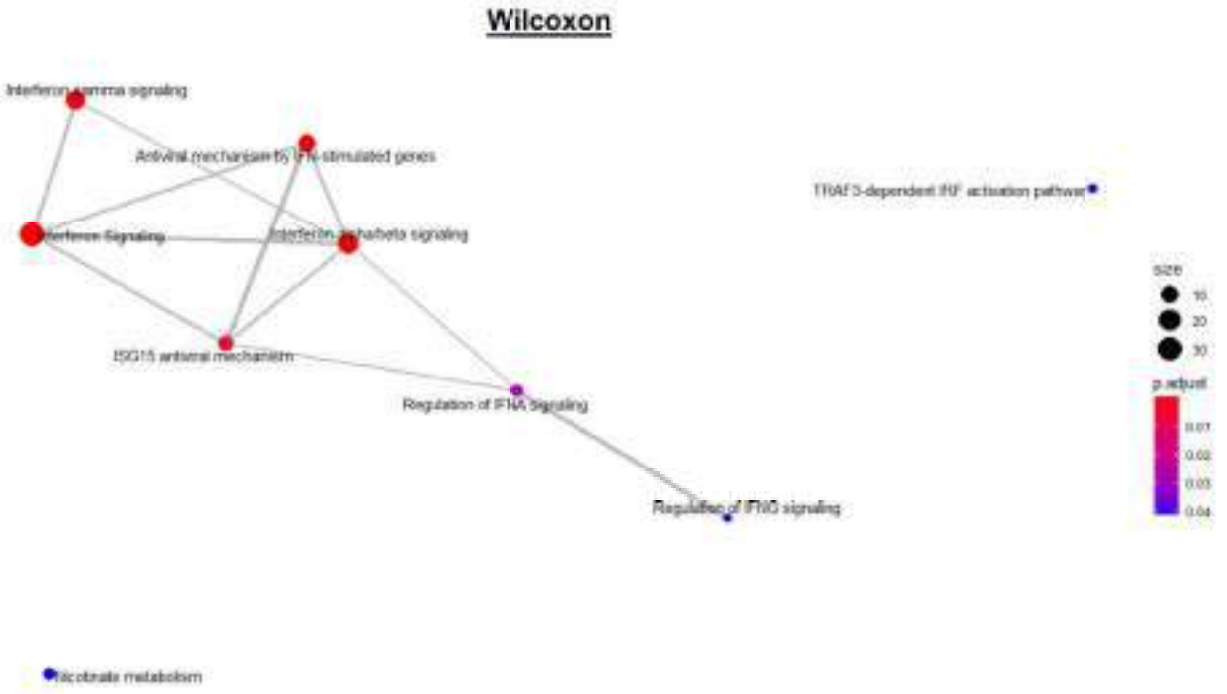


Figura 32 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L2.

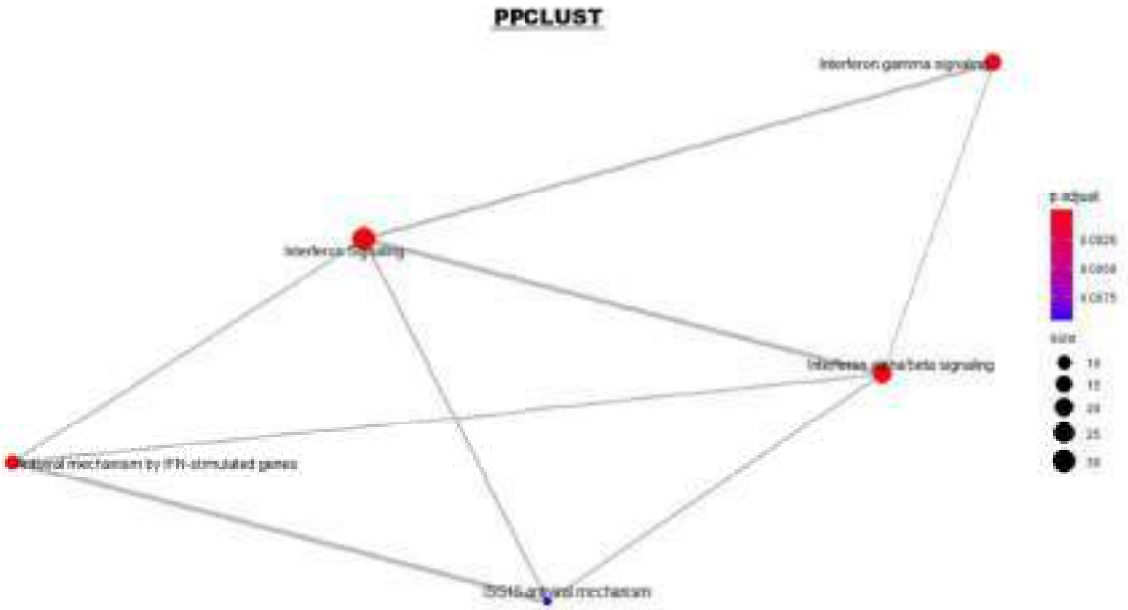


Figura 33 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L3.

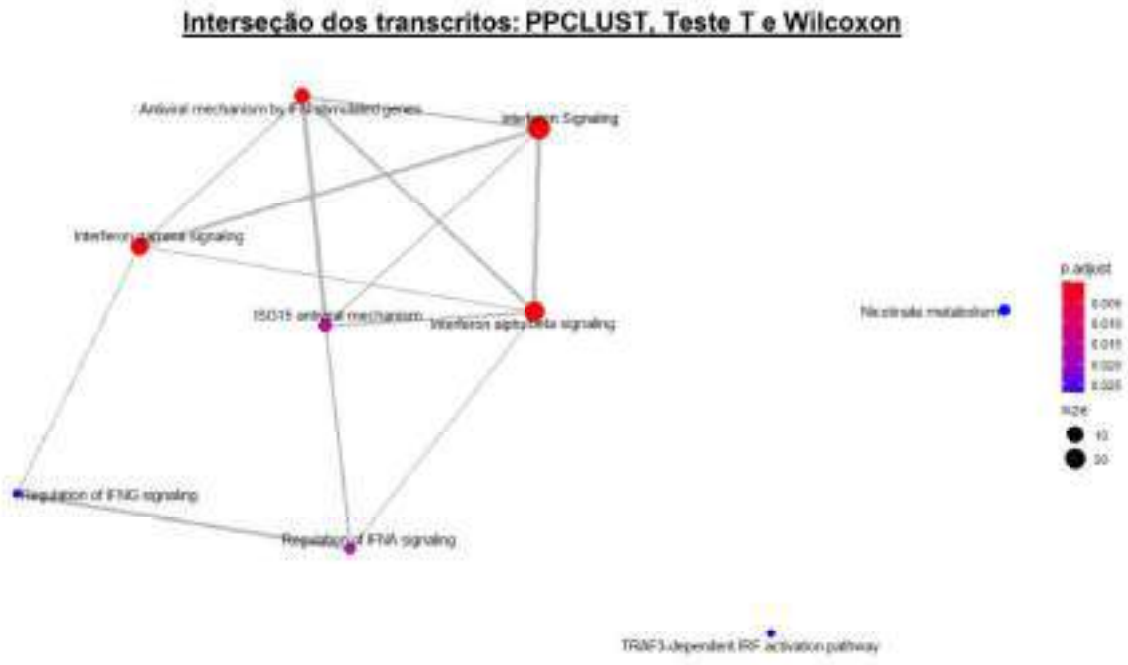


Figura 34 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L4.

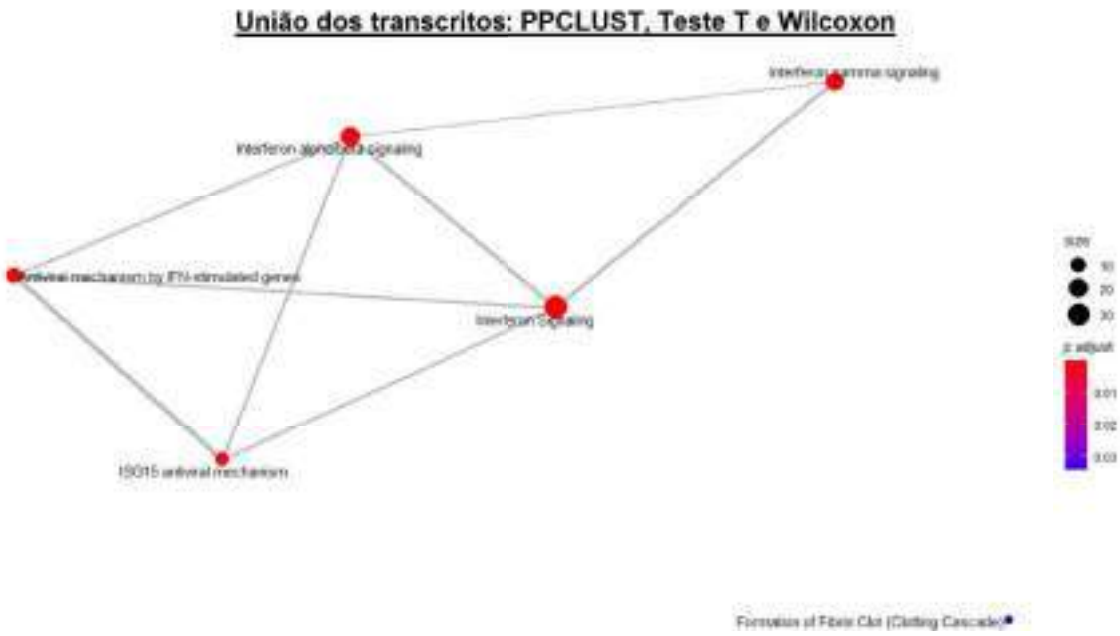


Figura 35 – Diagrama das ligações entre os sistemas biológicos identificados para a Lista L5.

Observa-se que os sistemas detectados para cada uma das listas estão bastante interligados, sendo o diagrama referente à Lista **L1** o que apresentou menos conexões, provavelmente porque apresentou mais sistemas. Sendo o referente a Lista **L3** o único que apresentou todos esses sitemas interligados.

Gráficos de rede, ou *cnetplots:*

Nas Figuras 36 a 40 são apresentadas ligações entre sistemas e seus respectivos genes dentre os que pertencem a uma das listas de transcritos.

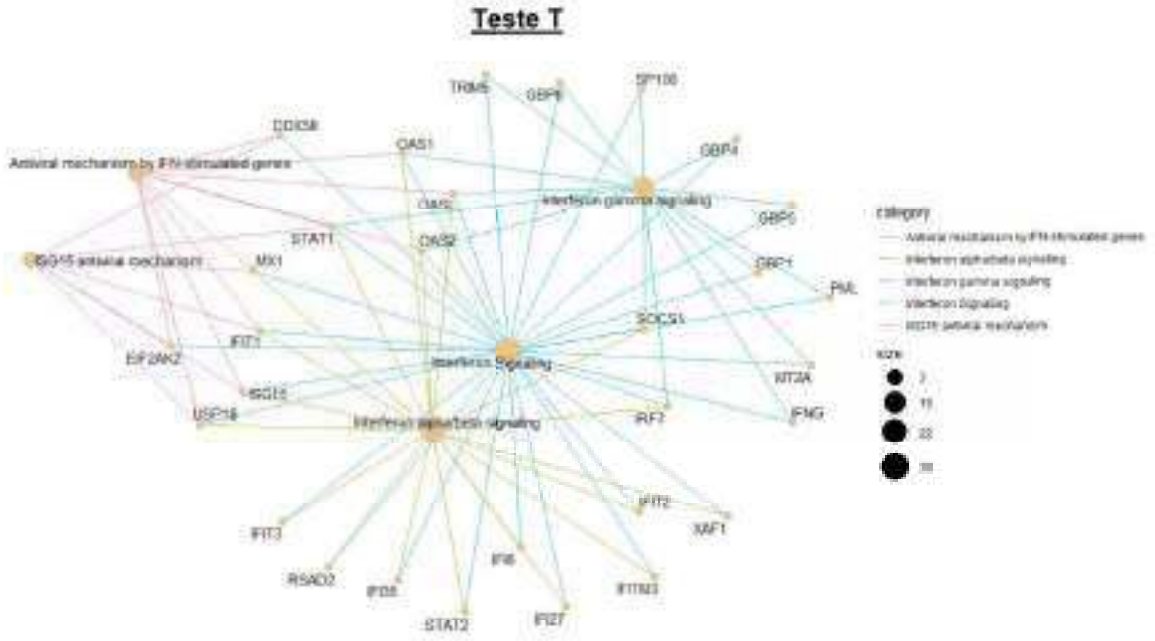


Figura 36 – Gráfico de redes para os genes e seus sistemas biológicos relacionados identificados na Lista **L1**.

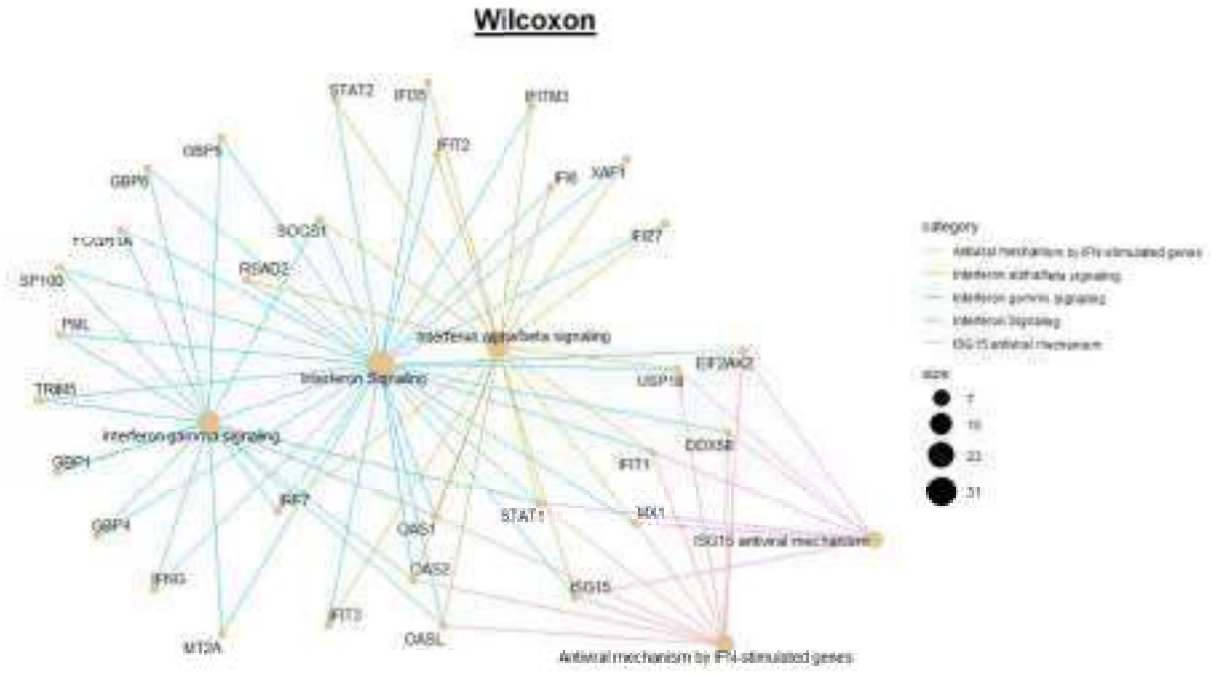


Figura 37 – Gráfico de redes para os genes e seus sistemas biológicos relacionados identificados para a Lista L2.

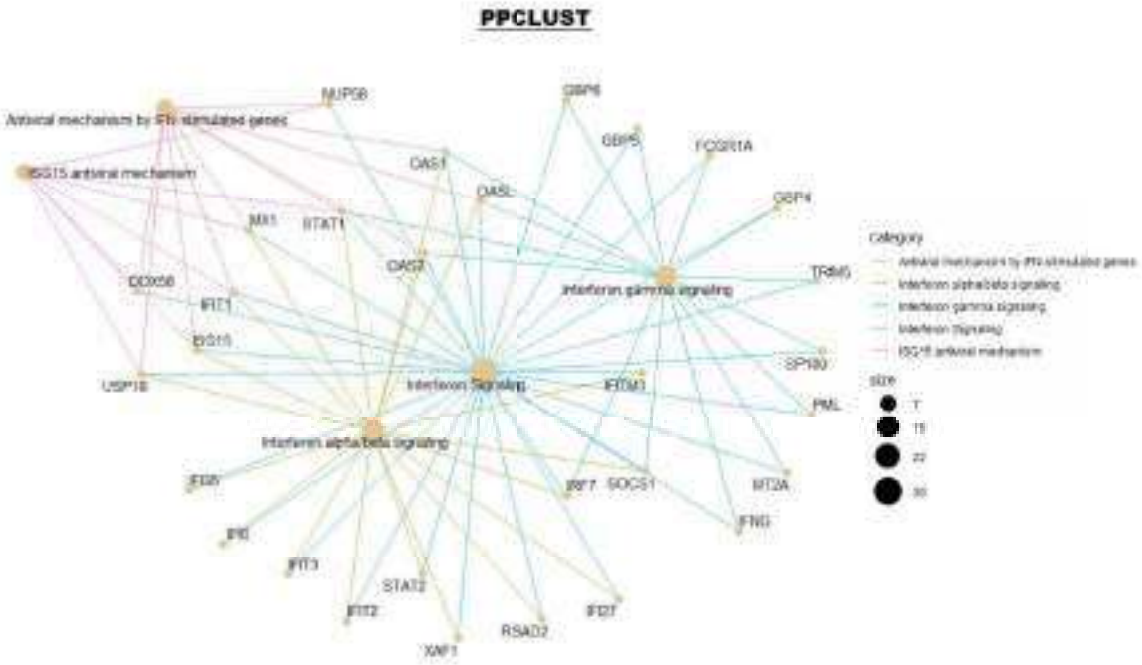


Figura 38 – Gráfico de redes para os genes e seus sistemas biológicos relacionados identificados para a Lista L3.

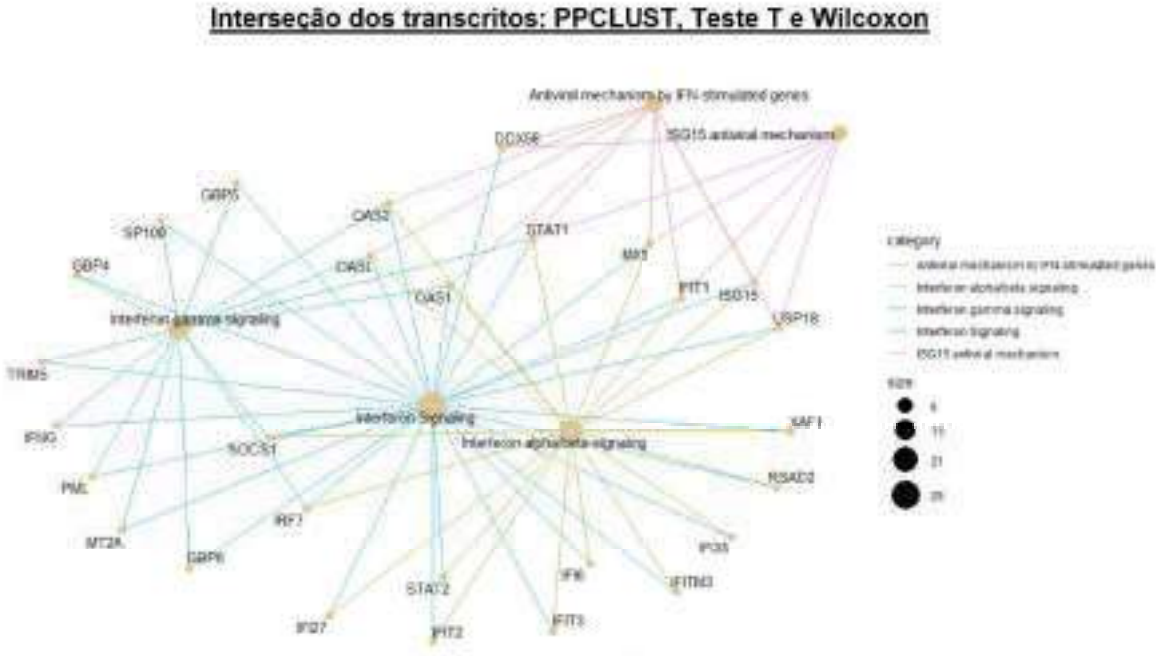


Figura 39 – Gráfico de redes para os genes e seus sistemas biológicos relacionados identificados para a Lista L4.

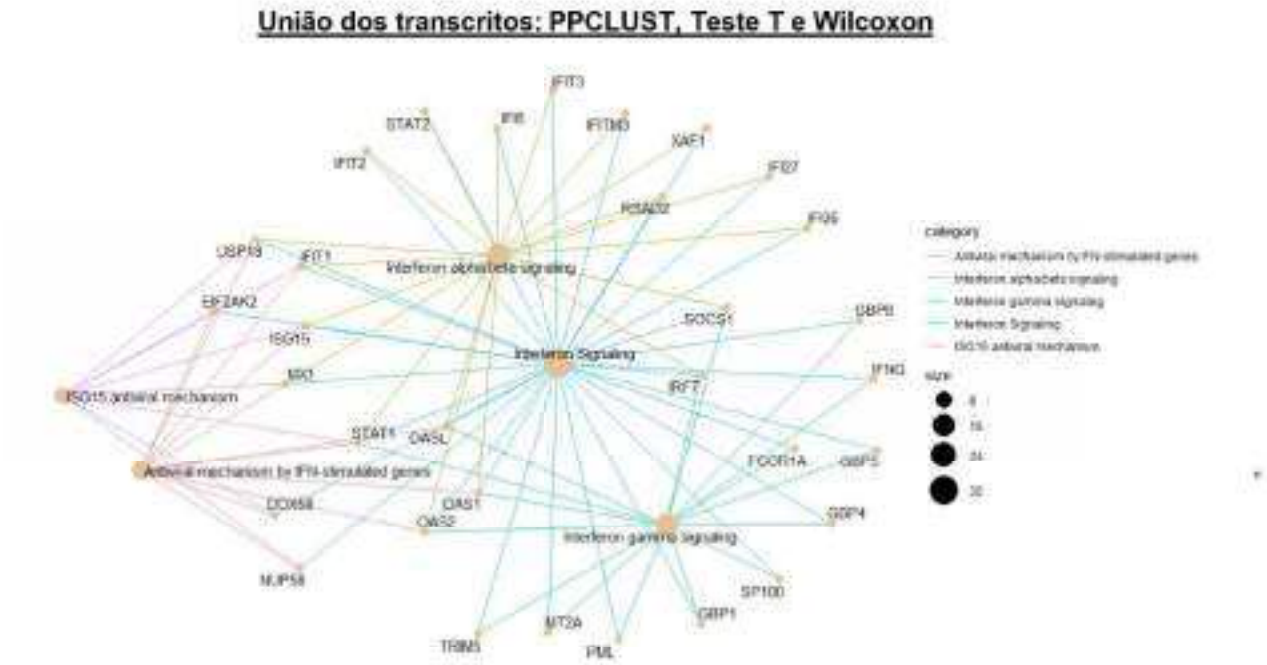


Figura 40 – Gráfico de redes para os genes e seus sistemas biológicos relacionados identificados para a Lista L5.

Observa-se que muitos genes pertencem a mais de um sistema biológico, sendo possível observar a complexidade nas relações entre genes e sistemas biológicos em todas as 5 listas.

4 Discussão e Conclusões

O presente trabalho envolveu uma análise de perfil de expressão gênica, comparando-se dados de casos de pSS e de controles saudáveis. Três abordagens distintas foram utilizadas para a identificação de transcritos diferencialmente expressos (TDEs), quais sejam, 1. Por p-valores de testes t, corrigidos por FDR $< 0{,}05$ e *Fold Change* (FC) de médias; 2. Por p-valores de testes Wilcoxon-Mann-Whitney, corrigidos por FDR e FC de medianas; 3. Por resultados do algorítmo PPCLUST e FC de medianas. Além das listas de TDEs obtidas por tais abordagens (**L1**, **L2** e **L3**, respectivamente), utilizou-se a lista composta pela intersecção (**L4**) e outra pela união (**L5**) de **L1**, **L2** e **L3**.

Observou-se que aproximadamente 50% dos TDEs estavam presentes em todas as técnicas. Em todas as listas de TDEs, foram identificados genes associados a SS também observados em outros estudos (em (Imgenberg-Kreuz e outros, 2019) e (Yao e outros, 2019)). Cinco dos 10 genes identificados como diferencialmente expressos estavam também presentes na lista dos 20 mais associados no estudo de YAO e outros (2019) e os 10 genes dados como mais diferencialmente expressos (Tabela 3) são também observados na Figura 1 (Imgenberg-Kreuz e outros, 2019), resultante de análises de estudos de genes candidatos.

Destaca-se entre os resultados obtidos, um número considerável de genes de interferons ou induzidos por interferons super-expressos. Interferons são citocinas que desempenham um papel central nas respostas imunes iniciais, com funções antivirais e antitumorais (Reactome, 2020) (Figura 41).

Entre os genes sub-expressos, foi identificado um gene relacionado a coagulação. Indivíduos com SS podem apresentar problemas de coagulação (Omim,2020) (Uniprot, 2020) (Genecards, 2020).

Com respeito às análise de anotação, houve bastante concordância entre os sistemas identificados nas Listas **L1** a **L5**, sendo **L1** a que identificou mais sistemas, seguida da **L2** e então **L3**, com apenas 5 sistemas, confirmados pelos resultados obtidos pelas outras duas técnicas. Os sistemas biológicos identificados já foram reportados em outros estudos genéticos de SS e algumas outras doenças autoimunes (Imgenberg-Kreuz e outros, 2019).

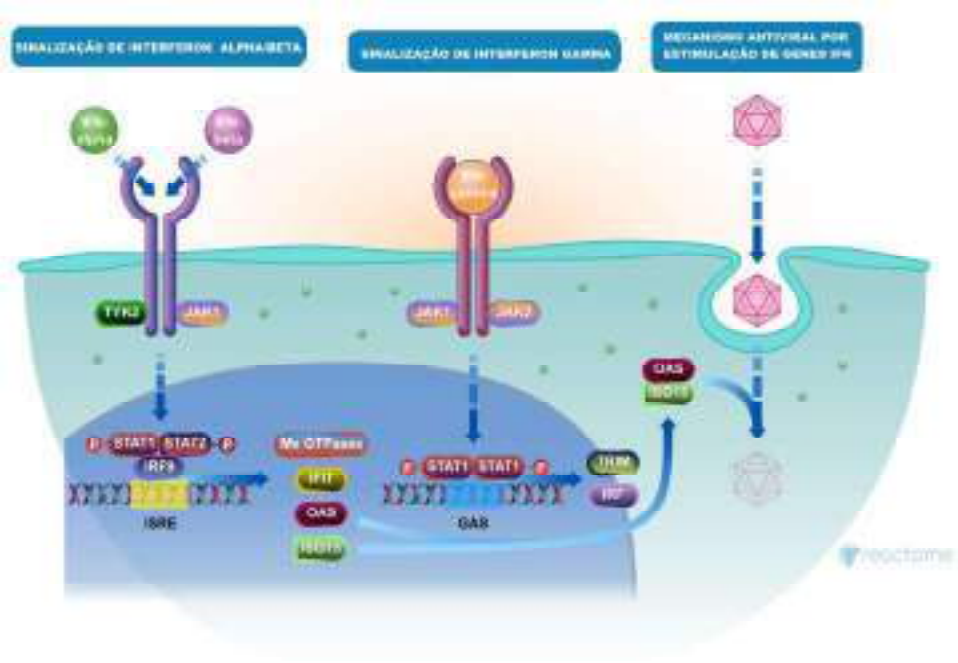


Figura 41 – Ilustração da ação de interferons. (Reactome, 2020)

Apesar de haver bastante concordância entre as listas (**L1**, **L2** e **L3**) e os sistemas biológicos identificados pelas 3 abordagens, tal comparação evidencia que a escolha do método de filtragem pode afetar substancialmente os resultados. Portanto, uma cuidadosa análise exploratória dos dados se faz necessária. Além disso, apesar de não ter sido tratado neste trabalho, procedimentos de controle de qualidade são essenciais e diferenças entre técnicas de normalização devem ser observadas.

Outras técnicas poderiam ter sido utilizadas buscando encontrar TDEs, como é o caso de uma Análise de Discriminante em que um gene teórico, super-expresso em todas as amostras do grupo 1 e sub-expresso em todas as amostras do grupo 2, seleciona genes que possuam a correlação de Pearson mais alta com um discriminador ideal.

Outra opção seria o método paramétrico de Bayes Empíricos para TDEs (`EBarrays`) que tem como base o cálculo de probabilidades a posteriori de padrões de expressão diferencial em várias condições.

O presente estudo permitiu uma análise global das expressões gênicas em SS. Mais estudos são necessários para um melhor entendimento dos processos moleculares que levam ao desenvolvimento e progressão da síndrome e a identificação de biomarcadores moleculares. Em particular, há o interesse em se identificar subtipos de pacientes e avaliar o perfil de expressão gênica por complicações clínicas, severidade da síndrome e resposta à medicamentos.

Referências

ABDI, H. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, Sage Thousand Oaks, CA, v. 3, p. 103–107, 2007. Citado na página 38.

AGRESTI, A. *An introduction to categorical data analysis third edition*. [S.l.]: Wiley New York, 2019. Citado 2 vezes nas páginas 29 e 35.

AGUIAR, P. M. d. C.; SEVERINO, P. Biomarkers in parkinson disease: global gene expression analysis in peripheral blood from patients with and without mutations in park2 and park8. *Einstein (Sao Paulo)*, SciELO Brasil, v. 8, n. 3, p. 291–297, 2010. Citado na página 38.

ARIVARADARAJAN, P.; MISRA, G. *Omics approaches, technologies and applications: integrative approaches for understanding OMICS data*. [S.l.]: Springer, 2019. Citado na página 28.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, Wiley Online Library, v. 57, n. 1, p. 289–300, 1995. Citado na página 34.

BIOCOMPUTING, P. S. on. *Pacific Symposium on Biocomputing 2007: Maui, Hawaii, 3-7 January 2007*. [S.l.]: World Scientific Publishing Company, 2007. Citado na página 39.

BIOCONDUTOR. *bioconductor.org*. 2020. <https://bioconductor.org/>. [Online; accessed 22-June-2020]. Citado na página 81.

BLAND, J. M.; ALTMAN, D. G. Multiple significance tests: the bonferroni method. *Bmj*, British Medical Journal Publishing Group, v. 310, n. 6973, p. 170, 1995. Citado na página 38.

BRANDT, J. E. et al. Sex differences in sjögren's syndrome: a comprehensive review of immune mechanisms. *Biology of sex differences*, Springer, v. 6, n. 1, p. 19, 2015. Citado na página 19.

BRAVIM, F. Análise de expressão gênica. In: . [S.l.]: Universidade Federal do Espírito Santo, Laboratório de Biotecnologia Aplicado ao Agronegócio, Brasil, 2013. Citado na página 28.

CARSONS, S. E.; PATEL, B. C. Sjogren syndrome. In: *StatPearls [Internet]*. [S.l.]: StatPearls Publishing, 2019. Citado na página 19.

EDGAR, R.; DOMRACHEV, M.; LASH, A. E. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic acids research*, Oxford University Press, v. 30, n. 1, p. 207–210, 2002. Citado na página 25.

FELBERG, S.; DANTAS, P. E. C. Diagnóstico e tratamento da síndrome de sjögren. *Arquivos brasileiros de oftalmologia*, Conselho Brasileiro de Oftalmologia, 2006. Citado na página 20.

GARCIA, A. A. F. Biometria de marcadores genéticos - tópico 3: Mapas genéticos i segregação mendeliana. In: . [S.l.]: Departamento de Genética - ESALQ/USP, 2019. Citado na página 34.

GENECARDS. *genecards.org*. 2020. <https://genecards.org/>. [Online; accessed 28-June-2020]. Citado 2 vezes nas páginas 51 e 65.

GRIFFITHS, A. J. et al. *An introduction to genetic analysis*. [S.l.]: Macmillan, 2005. Citado na página 27.

ILLUMINA, I. *HumanHT-12 v3 Expression BeadChip*. 2008. <https://www.illumina.com/Documents/products/datasheets/datasheet_humanht_12.pdf>. [Online; accessed 18-July-2020]. Citado na página 28.

IMGENBERG-KREUZ, J. et al. Genetics and epigenetics in primary sjögren's syndrome. *Rheumatology*, 2019. Citado 4 vezes nas páginas 20, 21, 22 e 65.

KASSAN, S. S.; MOUTSOPOULOS, H. M. Clinical manifestations and early diagnosis of sjögren syndrome. *Archives of internal medicine*, American Medical Association, v. 164, n. 12, p. 1275–1284, 2004. Citado na página 19.

KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 2009. v. 344. Citado 2 vezes nas páginas 35 e 36.

LESSARD, C. J. et al. Variants at multiple loci implicated in both innate and adaptive immune responses are associated with sjögren's syndrome. *Nature genetics*, Nature Publishing Group, v. 45, n. 11, p. 1284–1292, 2013. Citado 4 vezes nas páginas 25, 27, 28 e 41.

LI, H. et al. Interferons in sjögren's syndrome: genes, mechanisms, and effects. *Frontiers in immunology*, Frontiers, v. 4, p. 290, 2013. Citado na página 21.

LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, n, v. 4, n. 4, p. 18–36, 2009. Citado na página 35.

LINS, R. d. S. Implementação computacional de algoritmos para agrupamento de dados hdlss e hdllss. 2018. Citado na página 33.

LOPES, M. P. et al. Estrutura e variabilidade do promotor do gene do fator de necrose tumoral humano (tnf). Universidade Federal de Goiás, 2014. Citado na página 23.

MIOZZI, L. et al. Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *PLoS One*, Public Library of Science, v. 3, n. 6, 2008. Citado na página 38.

MIRANDA, P. A. V. de. Métodos de agrupamento (clustering) - aula 18. In: . [S.l.]: Universidade de São Paulo (USP), 2012. Citado na página 36.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Saraiva Educação SA, 2017. Citado na página 31.

NIKOLOV, N. P.; ILLEI, G. G. Pathogenesis of sjögren's syndrome. *Current opinion in rheumatology*, NIH Public Access, v. 21, n. 5, p. 465, 2009. Citado na página 20.

OMIM. *omim.org*. 2020. <https://omim.org/>. [Online; accessed 28-June-2020]. Citado 2 vezes nas páginas 51 e 65.

PASOTO, S. G.; MARTINS, V. A. de O.; BONFA, E. Sjögren's syndrome and systemic lupus erythematosus: links and risks. *Open access rheumatology: research and reviews*, Dove Press, v. 11, p. 33, 2019. Citado na página 19.

PATEL, R.; SHAHANE, A. The epidemiology of sjögren's syndrome. *Clinical epidemiology*, Dove Press, v. 6, p. 247, 2014. Citado na página 19.

QIN, B. et al. Epidemiology of primary sjögren's syndrome: a systematic review and meta-analysis. *Annals of the rheumatic diseases*, BMJ Publishing Group Ltd, v. 74, n. 11, p. 1983–1989, 2015. Citado na página 19.

RAMOS-CASALS, M. et al. Primary sjögren syndrome. British Medical Journal Publishing Group, v. 344, p. e3821, 2012. Citado na página 19.

REACTOME. *reactome.org*. 2020. <https://reactome.org/>. [Online; accessed 28-June-2020]. Citado 8 vezes nas páginas 65, 66, 73, 76, 77, 78, 79 e 80.

REACTOME.ORG. *reactome.org*. 2020. <https://reactome.org/>. [Online; accessed 22-June-2020]. Citado na página 38.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, v. 52, n. 3/4, p. 591–611, 1965. Citado na página 30.

STEIN, L. Genome annotation: from sequence to biology. *Nature reviews genetics*, Nature Publishing Group, v. 2, n. 7, p. 493–503, 2001. Citado na página 37.

TAPINOS, N. I. et al. *Sjögren's Syndrome*. Boston, MA: Springer US, 1999. 127–134 p. Citado na página 19.

THEANDER, E. et al. Prediction of sjögren's syndrome years before diagnosis and identification of patients with early onset and severe disease course by autoantibody profiling. *Arthritis & rheumatology*, Wiley Online Library, v. 67, n. 9, p. 2427–2436, 2015. Citado na página 20.

UNIPROT. *uniprot.org*. 2020. <https://uniprot.org/>. [Online; accessed 28-June-2020]. Citado 2 vezes nas páginas 51 e 65.

VALTYSDÓTTIR, S. T.; WIDE, L.; HÄLLGREN, R. Low serum dehydroepiandrosterone sulfate in women with primary sjögren's syndrome as an isolated sign of impaired hpa axis function. *The Journal of rheumatology*, The Journal of Rheumatology, v. 28, n. 6, p. 1259–1265, 2001. Citado na página 19.

VASTRIK, I. et al. Reactome: a knowledge base of biologic pathways and processes. *Genome biology*, BioMed Central, v. 8, n. 3, p. 1–13, 2007. Citado na página 38.

VITALI, C. et al. Classification criteria for sjögren's syndrome: a revised version of the european criteria proposed by the american-european consensus group. *Annals of the rheumatic diseases*, BMJ Publishing Group Ltd, v. 61, n. 6, p. 554–558, 2002. Citado na página 25.

VON BORRIES, G.; WANG, H. Partition clustering of high dimensional low sample size data based on p-values. *Computational statistics & data analysis*, Elsevier, v. 53, n. 12, p. 3987–3998, 2009. Citado na página 32.

WANG, H.; AKRITAS, M. G. Rank tests for anova with large number of factor levels. *Journal of Nonparametric Statistics*, Taylor & Francis, v. 16, n. 3-4, p. 563–589, 2004. Citado na página 33.

YAO, Q. et al. Identifying key genes and functionally enriched pathways in sjögren's syndrome by weighted gene co-expression network analysis. *Frontiers in genetics*, Frontiers, v. 10, p. 1142, 2019. Citado 2 vezes nas páginas 20 e 65.

Apêndices

APÊNDICE A – Descrição dos Sistemas Biológicos Identificados em Análises de Anotação

Descrição dos Sistemas Biológicos Identificados em Análises de Anotação - Parte 1. (Reactome, 2020)

- ***Interferon Sinaling*** **| Sinal de Interferon**: Citocinas que desempenham um papel central no início de respostas imunes, especialmente efeitos antivirais e antitumorais;
- ***Interferon alpha/beta sinaling*** **| Sinal de Interferon alfa/beta**: Tipo de interferon, citocinas, que desempenham um papel central no início de respostas imunes, especialmente efeitos antivirais e antitumorais;
- ***Interferon gamma sinaling*** **| Sinal de Interferon gama**: Tipo de interferon, citocinas, que desempenham um papel central no início de respostas imunes, especialmente efeitos antivirais e antitumorais;
- ***Antiviral mechanism by IFN-stimulated genes*** **| Mecanismo antiviral por genes estimulados por Interferons**: Os interferons ativam a sinalização que leva à indução transcricional de centenas de genes estimulados por interferon. As proteínas codificadas por inteferon incluem efetores diretos que inibem a infecção viral através de diversos mecanismos, bem como fatores que promovem respostas imunes adaptativas. As proteínas geradas pelas vias interferon desempenham papéis-chave na indução de respostas imunes inatas e adaptativas;
- ***ISG15 antiviral mechanism*** **| Mecanismo antiviral ISG15**: É fortemente induzido após a exposição a interferons tipo I, vírus, lipopolissacarídeo bacteriano e outros estresses. Uma vez liberado, o ISG15 maduro se conjuga com uma série de proteínas alvo, um processo denominado ISGylation;
- ***Regulation of IFNA signaling*** **| Regulação da sinalização IFNA**: Várias proteínas e mecanismos envolvidos no controle da extensão da estimulação do ligante da sinalização de interferons alpha e beta;
- ***Regulation of IFNG signaling*** **| Regulação da sinalização IFNG**: Várias proteínas e mecanismos envolvidos no controle da extensão da estimulação do ligante da sinalização de interferons alpha e beta;

- ***Nicotinamide salvaging*** **| Recuperação de nicotinamida**: Modulador da ação de três enzimas envolvidas na recuperação da nicotinamida;
- ***Nicotinate metabolism*** **| Metabolismo de nicotinato**: Modulador da ação de três enzimas envolvidas na recuperação da nicotinamida;
- ***Nicotinate metabolism*** **| Metabolismo de nicotinato**: São cofatores importantes em várias centenas de reações redox (oxidação-redução);
- ***Metallothioneins bind metals*** **| Metalotioneínas ligam metais**: Embora as funções das metalotioneínas não tenham sido totalmente elucidadas, elas parecem participar na desintoxicação de metais pesados, no armazenamento e transporte de zinco e na bioquímica redox;
- ***Response to metal ions*** **| Resposta a íons metálicos**: Embora metais como zinco, cobre e ferro sejam necessários como cofatores para enzimas celulares, eles também podem catalisar substituições de metais prejudiciais ou reações redox inespecíficas, se não forem retirados;
- ***Chemokine receptors bind chemokines*** **| Receptores de quimiocinas ligam quimiocinas**: As quimiocinas podem ser divididas em dois grupos funcionais, inflamatório e homeostático, mas a discriminação não é estrita e algumas sobreposições são encontradas. As quimiocinas inflamatórias são produzidas em condições inflamatórias por infiltração e células residentes em resposta a mediadores pró-inflamatórios (IL-1 e TNF-α), produtos bacterianos (LPS) e agentes infecciosos (vírus);
- ***Formation of Fibrin Clot*** **| Formação de coágulos de fibrina**: Ligados a cicatrização, formação de coagulos, para a cicatrização.

APÊNDICE B – Lista de transcritos estatisticamente significantes

Tabela 4 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 1. (Reactome, 2020)

ID	LOG FC	GENE SYMBOL	GENE TITLE
"ILMN_1751393"	0,977557056	"ZNF684"	"zinc finger protein 684"
"ILMN_1714402"	1,42041923	"ZCCHC2"	"zinc finger CCHC-type containing 2"
"ILMN_2402617"	-0,685994565	"ZBTB16"	"zinc finger and BTB domain containing 16"
"ILMN_1675566"	-0,659722728	"XKR8"	"XK related 8"
"ILMN_1742618"	1,482770738	"XAF1"	"XIAP associated factor 1"
"ILMN_1671260"	-0,728893603	"WLS"	"wntless Wnt ligand secretion mediator"
"ILMN_2399709"	-0,625034436	"WLS"	"wntless Wnt ligand secretion mediator"
"ILMN_1807379"	0,629798563	"WHSC1L1"	"Wolf-Hirschhorn syndrome candidate 1-like 1"
"ILMN_1752755"	-0,657468830	"VWF"	"von Willebrand factor"
"ILMN_1749200"	0,6027658	"USP18"	"ubiquitin specific peptidase 18"
"ILMN_2096012"	0,612713121	"UHMK1"	"U2AF homology motif (UHM) kinase 1"
"ILMN_1808040"	0,726100131	"TYMS"	"thymidylate synthetase"
"ILMN_2038775"	2,149127251	"TUBB2A"	"tubulin beta 2A class IIa"
"ILMN_2044813"	1,642306271	"TUBB2A"	"tubulin beta 2A class IIa"
"ILMN_1802731"	0,71873169	"TTYH3"	"tweety family member 3"
"ILMN_1710083"	-1,244351085	"TTC25"	"tetratricopeptide repeat domain 25"
"ILMN_1715007"	1,049445315	"TTC21A"	"tetratricopeptide repeat domain 21A"
"ILMN_1813436"	[illegible]	"TRIM69"	"tripartite motif containing 69"
"ILMN_1737599"	0,729915027	"TRIM5"	"tripartite motif containing 5"
"ILMN_1792359"	0,680910027	"TRIM5"	"tripartite motif containing 5"
"ILMN_1690783"	-1,770952695	"TREML1"	"triggering receptor expressed on myeloid cells like 1"
"ILMN_1686097"	1,073570049	"TOP2A"	"topoisomerase (DNA) II alpha"
"ILMN_1796508"	-0,748920809	"TOP1MT"	"topoisomerase (DNA) I, mitochondrial"
"ILMN_1699695"	-0,685182567	"TNFRSF21"	"TNF receptor superfamily member 21"
"ILMN_1759075"	-0,59680125	"TNFRSF13B"	"TNF receptor superfamily member 13B"
"ILMN_1785732"	[illegible]	"TNFAIP6"	"TNF alpha induced protein 6"
"ILMN_1770087"	-0,690035015	"TMPRSS9"	"transmembrane protease, serine 9"
"ILMN_2160577"	-0,647642663	"TMIGD2"	"transmembrane and immunoglobulin domain containing 2"
"ILMN_1706119"	0,620572664	"TMEM144"	"transmembrane protein 144"
"ILMN_2249082"	0,63336521	"TIPRL"	"TOR signaling pathway regulator"
"ILMN_1765307"	0,726002986	"TIMM10"	"translocase of inner mitochondrial membrane 10 homolog (yeast)"
"ILMN_2358969"	-0,772323619	"TIAM2"	"T-cell lymphoma invasion and metastasis 2"
"ILMN_1661603"	-0,640364189	"THRA"	"thyroid hormone receptor, alpha"
"ILMN_1759707"	-0,680793294	"THBD"	"thrombomodulin"
"ILMN_1728847"	-0,737391654	"TGM3"	"transglutaminase 3"
"ILMN_1788841"	0,647988877	"TCL1A"	"T-cell leukemia/lymphoma 1A"
"ILMN_1759250"	0,673779511	"TAP2"	"transporter 2, ATP binding cassette subfamily B member"
"ILMN_2413650"	0,607431877	"STIL"	"SCL/TAL1 interrupting locus"
"ILMN_1690921"	0,763880468	"STAT2"	"signal transducer and activator of transcription 2"
"ILMN_1690105"	0,772420607	"STAT1"	"signal transducer and activator of transcription 1"
"ILMN_1691364"	0,663930069	"STAT1"	"signal transducer and activator of transcription 1"
"ILMN_1777325"	0,661675490	"STAT1"	"signal transducer and activator of transcription 1"
"ILMN_2204545"	-0,650662423	"ST3GAL4"	"ST3 beta-galactoside alpha-2,3-sialyltransferase 4"
"ILMN_1708085"	-0,70380597	"SPRYD3"	"SPRY domain containing 3"
"ILMN_1651354"	[illegible]	"SPP1"	"secreted phosphoprotein 1"
"ILMN_2374449"	-0,860243155	"SPP1"	"secreted phosphoprotein 1"
"ILMN_1683678"	1,51600043	"SPATS2L"	"spermatogenesis associated serine rich 2 like"
"ILMN_1703263"	0,735450133	"SP140"	"SP140 nuclear body protein"
"ILMN_2246882"	0,706515583	"SP140"	"SP140 nuclear body protein"
"ILMN_2326596"	0,605396667	"SP100"	"SP100 nuclear antigen"
"ILMN_1774733"	0,73038091	"SOCS1"	"suppressor of cytokine signaling 1"
"ILMN_1775357"	0,615327981	"SNX29"	"sorting nexin 29"
"ILMN_1714445"	-1,082730541	"SLC6A8"	"solute carrier family 6 member 8"
"ILMN_1709695"	0,646542025	"SLC38A11"	"solute carrier family 38 member 11"
"ILMN_2384216"	0,614096078	"SLC26A8"	"solute carrier family 26 member 8"
"ILMN_1782938"	-0,660797625	"SLC16A10"	"solute carrier family 16 member 10"
"ILMN_2211039"	0,647132688	"SLC15A2"	"solute carrier family 15 member 2"
"ILMN_1725690"	0,692513564	"SIGLEC1"	"sialic acid binding Ig like lectin 1"
"ILMN_1776077"	0,589041533	"SF1"	"splicing factor 1"
"ILMN_1670305"	1,66873139	"SERPING1"	"serpin family G member 1"
[illegible]	[illegible]	"SERPINE2"	"serpin family E member 2"
"ILMN_1679606"	0,668966749	"SCGB3A1"	"secretoglobin family 3A member 1"
"ILMN_1805064"	0,719577883	"SCARNA9"	"small Cajal body-specific RNA 9"
"ILMN_1703688"	-1,032905616	"SAMD14"	"sterile alpha motif domain containing 14"
"ILMN_2173975"	1,101950046	"RTP4"	"receptor transporter protein 4"
"ILMN_1659981"	0,586150515	"RSPH9"	"radial spoke head 9 homolog"
"ILMN_1657871"	3,070865705	"RSAD2"	"radical S-adenosyl methionine domain containing 2"
"ILMN_1775178"	-0,738609211	"RPL3L"	"ribosomal protein L3 like"
"ILMN_1702955"	-0,650134968	"ROPN1B"	"rhophilin associated tail protein 1B"
"ILMN_2396207"	-0,774499431	"RFX2"	"regulatory factor X2"
"ILMN_1697490"	-0,835971767	"REEP6"	"receptor accessory protein 6"
"ILMN_2317751"	0,830574751	"REC8"	"REC8 meiotic recombination protein"
"ILMN_1808238"	-0,821094905	"RBPMS2"	"RNA binding protein with multiple splicing 2"
"ILMN_1740426"	-0,622131092	"RASD1"	"ras related dexamethasone induced 1"
"ILMN_1733045"	-0,621416812	"RAB36"	"RAB36, member RAS oncogene family"
"ILMN_2069224"	-0,760530635	"PVALB"	"parvalbumin"
[illegible]	[illegible]	"PTPRO"	"protein tyrosine phosphatase, receptor type O"
"ILMN_1810062"	0,641577905	"PTPRK"	"protein tyrosine phosphatase, receptor type K"

Tabela 5 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 2. (Reactome, 2020)

ID	LOG FC	GENE SYMBOL	GENE TITLE
"LMN_1695640"	0,627894165	"PTPN22"	"protein tyrosine phosphatase, non-receptor type 22"
"LMN_1736233"	0,588218022	"PSMB9"	"proteasome subunit beta 9"
"LMN_1810628"	1,296908943	"PRUNE2"	"prune homolog 2"
"LMN_1660017"	-0,609519843	"PRR5-ARHGAP8"	"PRR5-ARHGAP8 readthrough"
"LMN_1810608"	0,813404884	"PNPT1"	"polyribonucleotide nucleotidyltransferase 1"
"LMN_1726019"	0,634302235	"PML"	"promyelocytic leukemia"
"LMN_1745242"	0,96368421	"PLSCR1"	"phospholipid scramblase 1"
"LMN_1650857"	-0,581758804	"PLAU"	"plasminogen activator, urokinase"
"LMN_1704537"	-0,868271138	"PHGDH"	"phosphoglycerate dehydrogenase"
"LMN_2179823"	0,862194371	"PHC3"	"polyhomeotic homolog 3"
"LMN_1704870"	-0,908305549	"PGLYRP1"	"peptidoglycan recognition protein 1"
"LMN_1745522"	-0,882132907	"PF4V1"	"platelet factor 4 variant 1"
"LMN_1708580"	-0,847094455	"PDZK1IP1"	"PDZK1 interacting protein 1"
"LMN_1694982"	0,503515805	"PDK4"	"pyruvate dehydrogenase kinase 4"
"LMN_1815657"	1,082169247	"PDGFRB"	"platelet derived growth factor receptor beta"
"LMN_1731224"	0,807280265	"PARP9"	"poly(ADP-ribose) polymerase family member 9"
"LMN_2053527"	0,876819457	"PARP9"	"poly(ADP-ribose) polymerase family member 9"
"LMN_1691781"	1,014183113	"PARP14"	"poly(ADP-ribose) polymerase family member 14"
"LMN_1718558"	0,967677074	"PARP12"	"poly(ADP-ribose) polymerase family member 12"
"LMN_1721411"	0,73489482	"PARP10"	"poly(ADP-ribose) polymerase family member 10"
"LMN_1657373"	-0,693697125	"P3H2"	"prolyl 3-hydroxylase 2"
"LMN_1669847"	2,293122274	"OTOF"	"otoferlin"
"LMN_1732687"	0,814667796	"OR2AG1"	"olfactory receptor family 2 subfamily AG member 1 (gene/pseudogene)"
"LMN_1674811"	1,42249429	"OASL"	"2'-5'-oligoadenylate synthetase like"
"LMN_1661721"	1,181973523	"OASL"	"2'-5'-oligoadenylate synthetase like"
"LMN_1736729"	1,180695805	"OAS2"	"2'-5'-oligoadenylate synthetase 2"
"LMN_1654247"	0,955271848	"OAS1"	"2'-5'-oligoadenylate synthetase 1"
"LMN_1675048"	1,203441803	"OAS1"	"2'-5'-oligoadenylate synthetase 1"
"LMN_2410826"	1,303529656	"OAS1"	"2'-5'-oligoadenylate synthetase 1"
"LMN_1806446"	0,635703141	"NTNG2"	"netrin G2"
"LMN_1772627"	-0,812712746	"NSG1"	"neuron specific gene family member 1"
"LMN_2411236"	-0,803002708	"NRCAM"	"neuronal cell adhesion molecule"
"LMN_2049766"	0,858472389	"NFE2L3"	"nuclear factor, erythroid 2 like 3"
"LMN_1660437"	-0,616141450	"NECAB2"	"N-terminal EF-hand calcium binding protein 2"
"LMN_1727618"	0,813125136	"NDUFAF6"	"NADH ubiquinone oxidoreductase complex assembly factor 6"
"LMN_1769589"	-0,593241173	"NBL1"	"neuroblastoma 1, DAN family BMP antagonist"
"LMN_1716733"	-1,528175893	"MYOM2"	"myomesin 2"
"LMN_1716062"	-0,888940427	"MYL4"	"myosin light chain 4"
"LMN_1705257"	-0,597345444	"MYBPH"	"myosin binding protein H"
"LMN_1703915"	-0,896485226	"MXI1"	"MAX interactor 1, dimerization protein"
"LMN_1662358"	1,640073431	"MX1"	"MX dynamin like GTPase 1"
"LMN_1696664"	0,781161756	"MT2A"	"metallothionein 2A"
"LMN_1691156"	0,631734978	"MT1A"	"metallothionein 1A"
"LMN_2370836"	0,950522802	"MS4A4A"	"membrane spanning 4-domains A4A"
"LMN_1776939"	0,807224863	"MS4A1"	"membrane spanning 4-domains A1"
"LMN_1775016"	0,625536208	"MPZL2"	"myelin protein zero like 2"
"LMN_1717046"	0,589090902	"MOB3B"	"MOB kinase activator 3B"
"LMN_2396016"	-1,589756634	"MMP28"	"matrix metallopeptidase 28"
"LMN_1803673"	-0,862274548	"MISP3"	"MISP family member 3"
"LMN_1738588"	-0,840271481	"MGLL"	"monoglyceride lipase"
"LMN_2136589"	0,645269876	"MERTK"	"MER proto-oncogene, tyrosine kinase"
"LMN_1669321"	-0,997204607	"MATK"	"megakaryocyte-associated tyrosine kinase"
"LMN_2319000"	-0,891609262	"MATK"	"megakaryocyte-associated tyrosine kinase"
"LMN_1662336"	0,765084357	"MASTL"	"microtubule associated serine/threonine kinase like"
"LMN_1695404"	1,500611262	"LY9E"	"lymphocyte antigen 9 complex, locus E"
"LMN_2229242"	0,613622149	"LSM3"	"LSM3 homolog, U6 small nuclear RNA and mRNA degradation associated"
"LMN_1773856"	-1,105251381	"LRRN3"	"leucine rich repeat neuronal 3"
"LMN_2048591"	-0,921944621	"LRRN3"	"leucine rich repeat neuronal 3"
"LMN_1696004"	0,841754228	"LRRK1"	"leucine rich repeat kinase 1"
"LMN_2349018"	-1,261858321	"LRRC26"	"leucine rich repeat containing 26"
"LMN_1677196"	0,659291146	"LRCH3"	"leucine rich repeats and calponin homology domain containing 3"
"LMN_1665736"	-0,968886894	"LOC729451"	"uncharacterized LOC729451"
"LMN_1715760"	0,749247756	"LGALS9"	"galectin 9"
"LMN_1654890"	0,769329508	"LENG8"	"leukocyte receptor cluster member 8"
"LMN_1683792"	1,054036743	"LAP3"	"leucine aminopeptidase 3"
"LMN_1713581"	-0,829866134	"LAMP5"	"lysosomal associated membrane protein family member 5"
"LMN_2170814"	2,311554393	"LAMP3"	"lysosomal associated membrane protein 3"
"LMN_2386790"	-0,910938148	"KLRC3"	"killer cell lectin like receptor C3"
"LMN_1746123"	0,899101032	"KLHL14"	"kelch like family member 14"
"LMN_1673521"	-0,880030966	"KISS1R"	"KISS1 receptor"
"LMN_1771462"	-0,892017340	"KIAA1324"	"KIAA1324"
"LMN_1738481"	-0,597834393	"KEL"	"Kell blood group, metallo-endopeptidase"
"LMN_1796224"	0,607896028	"KDM5C"	"lysine demethylase 5C"
"LMN_1733811"	0,947561479	"JUP"	"junction plakoglobin"
"LMN_1741970"	0,881844541	"JUP"	"junction plakoglobin"
"LMN_2165993"	-0,827626834	"ITLN1"	"intelectin 1"
"LMN_1796755"	-0,592218613	"ITGB5"	"integrin subunit beta 5"
"LMN_2054019"	1,801347541	"ISG15"	"ISG15 ubiquitin-like modifier"

Tabela 6 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 3. (Reactome, 2020)

ID	LOG FC	GENE SYMBOL	GENE TITLE
"ILMN_1811468"	-0,8131923324	"IRX3"	"iroquois homeobox 3"
"ILMN_2349861"	0,710597537	"IRF7"	"interferon regulatory factor 7"
"ILMN_2340643"	-0,69407111	"INSC"	"inscuteable homolog (Drosophila)"
"ILMN_1697710"	0,614631828	"IL37"	"interleukin 37"
"ILMN_2207291"	1,054802508	"IFNG"	"interferon gamma"
"ILMN_1805750"	1,03215301	"IFITM3"	"interferon induced transmembrane protein 3"
"ILMN_1664543"	1,737876725	"IFIT3"	"interferon induced protein with tetratricopeptide repeats 3"
"ILMN_1701788"	1,552718774	"IFIT3"	"interferon induced protein with tetratricopeptide repeats 3"
"ILMN_2239754"	1,108690128	"IFIT3"	"interferon induced protein with tetratricopeptide repeats 3"
"ILMN_1738428"	1,024439699	"IFIT2"	"interferon induced protein with tetratricopeptide repeats 2"
"ILMN_1707695"	1,972678302	"IFIT1"	"interferon induced protein with tetratricopeptide repeats 1"
"ILMN_1781373"	0,940745037	"IFIH1"	"interferon induced with helicase C domain 1"
"ILMN_2347736"	1,729058907	"IFI6"	"interferon alpha inducible protein 6"
"ILMN_1723812"	2,836717413	"IFI44L"	"interferon induced protein 44 like"
"ILMN_1760062"	2,188853084	"IFI44"	"interferon induced protein 44"
"ILMN_1745374"	0,6841456471	"IFI35"	"interferon induced protein 35"
"ILMN_2058782"	2,834608357	"IFI27"	"interferon alpha inducible protein 27"
"ILMN_1686310"	0,8428286381	"IDO1"	"indoleamine 2,3-dioxygenase 1"
"ILMN_1671696"	-0,965306475	"HOMER2"	"homer scaffolding protein 2"
"ILMN_1705884"	0,684548429	"HNMT"	"histamine N-methyltransferase"
"ILMN_2066060"	0,8305736264	"HLA-DRB6"	"major histocompatibility complex, class II, DR beta 6 (pseudogene)"
"ILMN_2198239"	-0,94809176	"HGD"	"homogentisate 1,2-dioxygenase"
"ILMN_1652466"	1,476085795	"HES4"	"hes family bHLH transcription factor 4"
"ILMN_1787509"	0,787926187	"HELZ2"	"helicase with zinc finger 2"
"ILMN_2318568"	-1,194693209	"HCFC1R1"	"host cell factor C1 regulator 1"
"ILMN_2121408"	0,664686642	"HBEGF"	"heparin binding EGF like growth factor"
"ILMN_1723139"	1,037679171	"GPD2"	"glycerol-3-phosphate dehydrogenase 2"
"ILMN_1805221"	-0,633841206	"GOLGA7"	"golgin A7"
"ILMN_1723049"	-0,999186211	"GJC2"	"gap junction protein gamma 2"
"ILMN_1708932"	0,613526921	"GCH1"	"GTP cyclohydrolase 1"
"ILMN_1756953"	1,538310949	"GBP6"	"guanylate binding protein family member 6"
"ILMN_2114568"	1,066697476	"GBP5"	"guanylate binding protein 5"
"ILMN_1771385"	0,758321776	"GBP4"	"guanylate binding protein 4"
"ILMN_1701114"	1,396827094	"GBP1"	"guanylate binding protein 1"
"ILMN_2148785"	1,182911945	"GBP1"	"guanylate binding protein 1"
"ILMN_1698725"	0,831434156	"FRMD3"	"FERM domain containing 3"
"ILMN_1655525"	-0,840661787	"FLCN"	"folliculin"
"ILMN_1814952"	-0,788884214	"FLCN"	"folliculin"
"ILMN_1762531"	-0,717990803	"FGF9"	"fibroblast growth factor 9"
"ILMN_1727982"	-0,883486644	"FEZ1"	"fasciculation and elongation protein zeta 1"
"ILMN_1779071"	-0,65839678	"FEZ1"	"fasciculation and elongation protein zeta 1"
"ILMN_1654389"	0,8508886054	"FCGR1A"	"Fc fragment of IgG receptor Ia"
"ILMN_2302757"	-0,959266278	"FCGBP"	"Fc fragment of IgG binding protein"
"ILMN_1717063"	-0,640348463	"FBXO9"	"F-box protein 9"
"ILMN_1701455"	0,712288529	"FBXO6"	"F-box protein 6"
"ILMN_2380919"	-0,89407602	"FBLN2"	"fibulin 2"
"ILMN_1701082"	0,782688512	"FAM72D"	"family with sequence similarity 72 member D"
"ILMN_1797301"	-1,083752628	"FAM212B"	"family with sequence similarity 212 member B"
"ILMN_2401253"	0,602542867	"FAM13A"	"family with sequence similarity 13 member A"
"ILMN_1708105"	0,594883945	"EZH2"	"enhancer of zeste 2 polycomb repressive complex 2 subunit"
"ILMN_1708747"	1,030497658	"EXOG"	"endonuclease/exonuclease (5'-3'), endonuclease G-like"
"ILMN_1769092"	-0,598656142	"EVA1B"	"eva-1 homolog B"
"ILMN_1700671"	1,355596923	"ETV7"	"ETS variant 7"
"ILMN_2388547"	2,146902559	"EPSTI1"	"epithelial stromal interaction 1 (breast)"
"ILMN_2323427"	-0,635481464	"EPB41"	"erythrocyte membrane protein band 4.1"
"ILMN_1708502"	1,235838107	"EIF2AK2"	"eukaryotic translation initiation factor 2 alpha kinase 2"
"ILMN_1767322"	-0,588078442	"EDAR"	"ectodysplasin A receptor"
"ILMN_1730284"	-0,945551939	"DSC1"	"desmocollin 1"
"ILMN_2402640"	-0,767310987	"DSC1"	"desmocollin 1"
"ILMN_2047112"	-0,604385498	"DPCD"	"deleted in primary ciliary dyskinesia homolog (mouse)"
"ILMN_1678422"	0,924053541	"DHX58"	"DEXH-box helicase 58"
"ILMN_1795181"	1,354480158	"DDX60"	"DEXD/H-box helicase 60"
"ILMN_1797001"	0,85279248	"DDX58"	"DEXD/H-box helicase 58"
"ILMN_1819304"	-0,710833785	"DDX5"	"DEAD-box helicase 5"
"ILMN_1801065"	0,60658642	"DACT1"	"dishevelled binding antagonist of beta catenin 1"
"ILMN_1708303"	1,339231854	"CYP4F22"	"cytochrome P450 family 4 subfamily F member 22"
"ILMN_1705403"	-0,693477791	"CYP2S1"	"cytochrome P450 family 2 subfamily S member 1"
"ILMN_1651757"	0,722625373	"CXorf21"	"chromosome X open reading frame 21"
"ILMN_1674640"	-0,751509724	"CXCR6"	"C-X-C motif chemokine receptor 6"
"ILMN_1752562"	-0,716741967	"CXCL5"	"C-X-C motif chemokine ligand 5"
"ILMN_1791759"	1,772269071	"CXCL10"	"C-X-C motif chemokine ligand 10"
"ILMN_1812995"	0,66981082	"CTSL"	"cathepsin L"
"ILMN_2374036"	0,604339896	"CTSL"	"cathepsin L"
"ILMN_1737992"	-0,592405052	"CRAT"	"carnitine O-acetyltransferase"
"ILMN_1804724"	0,795898783	"CR2"	"complement component 3d receptor 2"
"ILMN_1766551"	-0,691475405	"CPA3"	"carboxypeptidase A3"
"ILMN_2184949"	0,657809139	"COX7B"	"cytochrome c oxidase subunit 7B"
"ILMN_1783909"	-1,052385664	"COL6A2"	"collagen type VI alpha 2 chain"

Tabela 7 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 4. (Reactome, 2020)

ID	LOG FC	GENE SYMBOL	GENE TITLE
"LMN_1711514"	-0,900845829	"COCH"	"cochlin"
"LMN_1763821"	3,123557895	"CMPK2"	"cytidine/uridine monophosphate kinase 2"
"LMN_2059274"	-0,845408694	"CLEC9A"	"C-type lectin domain family 9 member A"
"LMN_1662259"	-0,8007975	"CLEC4C"	"C-type lectin domain family 4 member C"
"LMN_1756926"	0,756265626	"CKS2"	"CDC28 protein kinase regulatory subunit 2"
"LMN_2072296"	0,587892761	"CKS2"	"CDC28 protein kinase regulatory subunit 2"
"LMN_1732534"	0,859730044	"CHMP5"	"charged multivesicular body protein 5"
"LMN_2270100"	0,770493447	"CEP85L"	"centrosomal protein 85 like"
"LMN_1710815"	0,75460514	"CEACAM1"	"carcinoembryonic antigen related cell adhesion molecule 1"
"LMN_2371724"	0,878864781	"CEACAM1"	"carcinoembryonic antigen related cell adhesion molecule 1"
"LMN_1663390"	0,7906861 72	"CDC20"	"cell division cycle 20"
"LMN_1651316"	0,865871514	"CD69"	"CD69 molecule"
"LMN_2233763"	0,791955954	"CD36"	"CD36 molecule"
"LMN_1701914"	0,591561057	"CD274"	"CD274 molecule"
"LMN_1726589"	-1,868045392	"CD248"	"CD248 molecule"
"LMN_1678833"	0,592756103	"CCR1"	"C-C motif chemokine receptor 1"
"LMN_1766125"	0,753668939	"CCNA2"	"cyclin A2"
"LMN_2157089"	0,612736046	"CCNA1"	"cyclin A1"
"LMN_1772964"	0,885724277	"CCL8"	"C-C motif chemokine ligand 8"
"LMN_1764038"	0,886797514	"CCL23"	"C-C motif chemokine ligand 23"
"LMN_1720048"	1,0966 68678	"CCL2"	"C-C motif chemokine ligand 2"
"LMN_1707979"	1,226984948	"CARD17"	"caspase recruitment domain family member 17"
"LMN_1737089"	-0,880769362	"CAPN5"	"calpain 5"
"LMN_1676973"	-0,750903789	"CABP5"	"calcium binding protein 5"
"LMN_1668917"	-0,587152851	"C5AR2"	"complement component 5a receptor 2"
"LMN_1810752"	-0,69210541	"C4BPA"	"complement component 4 binding protein alpha"
"LMN_2224486"	0,840252327	"C3orf14"	"chromosome 3 open reading frame 14"
"LMN_1710746"	0,616423176	"C2"	"complement component 2"
"LMN_1662578"	0,697859403	"C1GALT1"	"core 1 synthase, glycoprotein-N-acetylgalactosamine 3-beta-galactosyltransferase 1"
"LMN_1690241"	1,6051053	"BATF2"	"basic leucine zipper ATF-like transcription factor 2"
"LMN_1734655"	-0,585477781	"ATG9B"	"autophagy related 9B"
"LMN_2374865"	1,394758446	"ATF3"	"activating transcription factor 3"
"LMN_1815184"	0,864788709	"ASPM"	"abnormal spindle microtubule assembly"
"LMN_1696414"	-0,704974268	"ASF1B"	"anti-silencing function 1B histone chaperone"
"LMN_2188878"	0,606467828	"ARMT1"	"acidic residue methyltransferase 1"
"LMN_1664853"	0,657337512	"ARHGAP25"	"Rho GTPase activating protein 25"
"LMN_1660182"	0,773601001	"APOBEC3A"	"apolipoprotein B mRNA editing enzyme catalytic subunit 3A"
"LMN_1708648"	1,006218955	"ANKRD22"	"ankyrin repeat domain 22"
"LMN_2132539"	1,030068408	"ANKRD22"	"ankyrin repeat domain 22"
"LMN_1783443"	0,587794200	"ALOX15"	"arachidonate 15-lipoxygenase"
"LMN_2387712"	-0,813561555	"AK5"	"adenylate kinase 5"
"LMN_1681301"	0,625281321	"AIM2"	"absent in melanoma 2"
"LMN_1770454"	0,912073843	"AGRN"	"agrin"
"LMN_1707925"	-0,653624337	"ABHD12B"	"abhydrolase domain containing 12B"
"LMN_1652417"	0,684394564	""	""
"LMN_1654639"	0,76439301	""	""
"LMN_1654692"	0,610946546	""	""
"LMN_1657932"	-1,077861336	""	""
"LMN_1658579"	0,60943002	""	""
"LMN_1667444"	-1,287874744	""	""
"LMN_1676709"	-0,581046845	""	""
"LMN_1680757"	-1,115620056	""	""
"LMN_1680772"	-0,749120346	""	""
"LMN_1681726"	-0,866155826	""	""
"LMN_1689172"	-0,751942346	""	""
"LMN_1690365"	0,778591736	""	""
"LMN_1693836"	-0,79789856	""	""
"LMN_1694127"	-1,423217589	""	""
"LMN_1695485"	0,843233476	""	""
"LMN_1708004"	0,919373435	""	""
"LMN_1710644"	0,628541854	""	""
"LMN_1723433"	-0,615618915	""	""
"LMN_1730451"	0,645856005	""	""
"LMN_1735364"	0,620213353	""	""
"LMN_1746817"	0,669679185	""	""
"LMN_1751289"	-0,653698883	""	""
"LMN_1756952"	0,624307768	""	""
"LMN_1768551"	0,76327292	""	""
"LMN_1770765"	-0,77428903	""	""
"LMN_1773887"	-0,617404937	""	""
"LMN_1782467"	1,7067001	""	""
"LMN_1798101"	0,645663858	""	""
"LMN_1804356"	-1,061582287	""	""
"LMN_1808769"	0,679412761	""	""
"LMN_1822671"	-2,31230011	""	""
"LMN_1834898"	-0,625708477	""	""
"LMN_1860803"	0,636125151	""	""
"LMN_1862108"	-0,631667878	""	""

Tabela 8 – APÊNDICE: Lista de transcritos estatisticamente significantes - PARTE 5. (Reactome, 2020)

ID	LOG FC	GENE SYMBOL	GENE TITLE
"ILMN_1870885"	-1,300109527		
"ILMN_1887011"	-0,714887909		
"ILMN_1889958"	0,707857487		
"ILMN_1890691"	0,614521733		
"ILMN_2296880"	0,596810841		
"ILMN_2377829"	0,712354649		

APÊNDICE C – Informações relacionadas ao software R e ao banco de dados.

- Fonte dos dados: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51092>
  - Dados brutos: `GSE51092RAW.tar`
  - Dados não normalizados: `GSE51092nonnormalized.txt.gz`
  - Utilizou-se a opção: `"Analyze with GEO2R"`
- *Heatmap*: pacote *ComplexHeatmap* do projeto *Bioconductor*[1].
- Anotação genômica: *"ReactomePA"* do projeto *Biocondutor*.

[1] Bioconductor é um projeto de software com código aberto para bioinformática, que fornece ferramentas para a análise e compreensão de dados genômicos de alto rendimento utilizando linguagem de programação estatística R (BIOCONDUTOR, 2020) com o intuito de promover a compreensão dos ensaios biológicos atuais e emergentes de alto rendimento.