

UNIVERSIDADE ESTADUAL DE MONTES CLAROS

Carlos Rafael Lima Monção

Análise de bioinformática das neoplasias de glândulas salivares

Montes Claros  
2019

Carlos Rafael Lima Monção

Análise de bioinformática das neoplasias de glândulas salivares

Dissertação apresentada ao Programa de Pós-graduação em Ciências da Saúde da Universidade Estadual de Montes Claros - Unimontes, como parte das exigências para a obtenção do título de Mestre em Ciências da Saúde.

Área de Concentração: Mecanismos e aspectos clínicos das doenças.

Orientador: Prof. Dr. André Luiz Sena Guimarães.

Coorientador: Prof. Dr. Marcos Flávio Silveira

Vasconcelos D'Angelo.

Montes Claros  
2019

M737a Monção, Carlos Rafael Lima.  
Análise de bioinformática das neoplasias de glândulas salivares [manuscrito] /  
Carlos Rafael Lima Monção. – 2019.  
73 f. : il.

Bibliografia: 61-64.  
Dissertação (mestrado) - Universidade Estadual de Montes Claros -  
UNIMONTES, Programa de Pós-Graduação em Ciências da Saúde/PPGCS, 2019.

Orientador: Prof. Dr. André Luiz Sena Guimarães.  
Coorientador: Prof. Dr. Marcos Flávio Silveira Vasconcelos D'Angelo.

1. Bioinform. 2. Glândulas salivares - Tumores. 3. Biologia computacional. 4.  
Genes líderes. 5. Biomarcadores do câncer. I. Guimarães, André Luiz Sena. II.  
D'Angelo, Marcos Flávio Silveira Vasconcelos. III. Universidade Estadual de  
Montes Claros. IV. Título.

Catálogo: Biblioteca Central Professor Antônio Jorge.

## UNIVERSIDADE ESTADUAL DE MONTES CLAROS-UNIMONTES

Reitor: Professor Antônio Alvimar Souza

Vice-reitora: Professora Ilva Ruas de Abreu

Pró-reitor de Pesquisa: Professor José Reinaldo Mendes Ruas

Coordenadoria de Acompanhamento de Projetos: Virgílio Mesquita Gomes

Coordenadoria de Iniciação Científica: Sônia Ribeiro Arruda

Coordenadoria de Inovação Tecnológica: Sara Gonçalves Antunes

Pró-reitor de Pós-graduação: Professor André Luiz Sena Guimarães

Coordenadoria de Pós-graduação Lato-sensu: Professor Marcos Flávio S. V. D'Angelo.

Coordenadoria de Pós-graduação Stricto-sensu: Professor Marcelo Perim Baldo

### PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE

Coordenador: Prof. Dr. Alfredo Maurício Batista de Paula

Subcoordenadora: Prof<sup>ª</sup>. Dr<sup>ª</sup> Marise Fagundes Silveira



UNIVERSIDADE ESTADUAL DE MONTES CLAROS  
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA SAÚDE



MESTRANDO(A): CARLOS RAFAEL LIMA MONÇÃO

TÍTULO DO TRABALHO: "Análise de bioinformática das neoplasias de glândulas salivares".

ÁREA DE CONCENTRAÇÃO: Mecanismos e aspectos clínicos das doenças

LINHA DE PESQUISA: Etiopatogenia e Fisiopatologia das Doenças

**BANCA (TITULARES)**

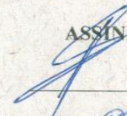
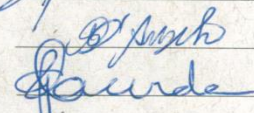
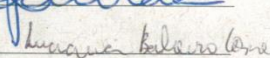
PROF. DR. ANDRÉ LUIZ SENA GUIMARÃES - ORIENTADOR/PRESIDENTE

PROF. DR. MARCOS FLÁVIO SILVEIRA VASCONCELOS D'ANGELO

PROF. DR. GUILHERME ARAÚJO LACERDA

PROF.ª DR.ª LUCIANA BALIEIRO COSME

ASSINATURAS

**BANCA (SUPLENTES)**

PROF.ª DR.ª LUCYANA CONCEIÇÃO FARIAS

PROF. DR. ALFREDO MAURÍCIO BATISTA DE PAULA

ASSINATURAS

\_\_\_\_\_  
\_\_\_\_\_

APROVADO

REPROVADO

Hospital Universitário Clemente Farias – HUCF

<http://www.unimontes.br> / [ppgcs@unimontes.br](mailto:ppgcs@unimontes.br)

Telefone: (0xx38) 3224-8372 / Fax: (0xx38) 3224-8372

Av. Cula Mangabeira, 562, Santo Expedito, Montes Claros – MG, Brasil – Cep: 39401-001

Dedico este trabalho

A Deus que sempre me guiou com sua luz e com sua eterna bondade pelos caminhos que trilhei. Dando-me forças e conforto quando mais precisei.

A minha esposa Caroline por sempre me apoiar e me compreender nessa gratificante, mas árdua caminhada.

Ao meu filho ou minha filha, que mesmo sem ainda ter nascido, já tem todo o meu amor.

Aos meus pais Carlos Monção e Maria Otaviana que me deram grandes exemplos de dedicação, honestidade e amor.

Aos meus familiares e amigos que compreenderam minha ausência.

## AGRADECIMENTOS

A Deus por me conceder a dádiva da vida e por sempre me dar forças para seguir a diante.

A minha esposa pelo estímulo e pelo apoio para me tornar uma pessoa melhor.

Aos meus pais, irmãos, tios, sobrinhos e cunhados pela ajuda e compreensão.

Ao meu orientador Prof. Dr. André Luiz Sena Guimarães pela oportunidade e por todo conhecimento transmitido. Agradeço ainda pelo exemplo de pessoa e profissional que é.

Aos colegas mestrandos e doutorandos por todos os ensinamentos e pelas experiências compartilhadas.

## RESUMO

É importante obter conhecimentos sobre os processos biológicos subjacentes ao desenvolvimento dos tumores de glândulas salivares com o objetivo de identificar potenciais biomarcadores ou novos alvos terapêuticos moleculares. Esse estudo foi realizado para investigar os genes líderes dos tumores de glândulas salivares. Inicialmente, os genes relacionados a todos os tumores de glândulas salivares foram identificados utilizando as bases de dados do Pubmed, OMIM e Genecards. Então, a base de dados do STRING foi utilizada para obter redes de interação entre as proteínas específicas de cada tipo tumoral e em seguida os valores de weighted number of links (WNL) e de total interaction score (TIS) foram calculados. Foi encontrado um total de 3070 genes relacionados com os 35 tipos de tumores de glândula salivar. Posteriormente foram identificados os genes líderes através dos algoritmos de cluster K-means e abordagem imuno/neural. Esse último método de clusterização de dados permitiu a utilização de uma estratégia inovadora que pode complementar ou substituir métodos mais tradicionais como o K-means. Foram identificados 764 genes líderes e detectados os genes com maiores valores de WNL, ou seja com as vias biológicas mais específicas, e com menor TIS, o que significa menor participação em processos biológicos essenciais. Do total de genes líderes, somente 92 estavam exclusivamente associados aos tumores benignos e 189 estavam exclusivamente associados aos malignos. Com esses dois últimos grupos de genes foi realizado o mesmo método de bioinformática descrito, com a criação de duas novas redes de interações proteicas no STRING, a obtenção de WNL e TIS seguida dos algoritmos K-means e imuno/neural. Os genes *UBE2K*, *CUL1*, *HGS*, *UBE2D1*, *SDHB*, *UBQLN1*, *UBE2D2* e *RAD23A* foram identificados como genes líderes exclusivamente de tumores benignos. Já os genes *INS*, *HRAS*, *SRC* e *JUN* foram os genes líderes exclusivamente dos tumores malignos. A partir desses dados, mais duas redes de interação proteica foram obtidas e foram identificados, então, o gene *RPS15* nos tumores benignos e o *PIK3CA* nos malignos. Esses dois genes poderiam ser possíveis alvos terapêuticos ou biomarcadores. Todas essas informações estão apresentadas detalhadamente em tabelas nos APÊNDICES A, B e C. Para validar os dados e hipóteses geradas por esse método de bioinformática, foram realizados experimentos com isolamento de RNA e real time PCR em amostras de tumores de glândulas salivares que evidenciaram maior expressão gênica do *RPS15* em tumores benignos comparativamente aos malignos.

Palavras-chave: Bioinformática. Neoplasia de glândula salivar. Biologia computacional. Genes líderes. Biomarcadores.



## ABSTRACT

It is important obtain knowledge of the biological processes underlying the developing of tumors of salivary glands with the objective of identify potentials biomarkers or new therapeutic molecular targets. This study was made to investigate the genes leaders of tumors of salivary glands. At first the genes related to all type of tumors of the salivary glands were identified using the databases Pubmed, OMIM and Genecards. Then, the database STRING was used to construct networks of protein-protein interactions to each specific tumor and the values of weighted number of links (WNL) and total interaction score (TIS) were obtained. In this first phase of this study, it was found a total of 3070 genes related with the 35 types of salivary gland tumors. The leaders genes were identified using the algorithm of K-means and Immune/Neural approach. The latter method of data clustering allowed the use of another innovated method which could add or replace more traditional methods like the K-means. Using these two algorithms, 764 genes were identified as leaders genes. And the identification of the genes with the higher values of WNL, in the other words, the genes with more specific biological ways in the tumors in study, and the genes with lower values of TIS, it means that they have a little importance in essential biological process. From the total of leader genes, only 92 leader genes were associated exclusively with benign tumors and 189 others leaders genes were associated exclusively with malignant tumors. Using the last two groups of genes, it was achieved two new networks of protein-protein interactions with the same bioinformatics method, including the calculation of values of WNL and TIS and K-means and Immune/Neural algorithms. The genes *UBE2K*, *CUL1*, *HGS*, *UBE2D1*, *SDHB*, *UBQLN1*, *UBE2D2* and *RAD23A* were identified as leaders genes exclusively of benign tumors. And the genes *INS*, *HRAS*, *SRC* and *JUN* were established as leader genes exclusively of malignant tumors. Again, from these data, another search for leader gene was proceeded. Then, it was identified the gene *RPS15* associated with benign tumors, and the gene *PIK3CA* in the malignant tumors. These two genes could be possible therapeutic molecular targets or biomarkers. All this data be available in the APPENDIXES A, B and C. With the objective of validation of the hypothesis created using that bioinformatic method, it was achieved experiments with RNA isolation and real time PCR in samples of salivary glands neoplasm that demonstrated higher level of genetic expression of *RPS15* in benign tumors in comparison with malignant tumors.

Keywords: Bioinformatics. Salivary Gland Neoplasm. Computational biology. Leader genes. Biomarkers.

## LISTA DE ILUSTRAÇÕES

Figura 1 – Anatomia das glândulas salivares .....	14
Figura 2 – Unidade funcional das glândulas salivares.....	15
Figura 3 – Rede de interações proteína-proteína do STRING.....	24
Figura 4 – Rede biológica de genes visualizada com o Cytoscape.....	25
Figura 5 – Gráfico do BiNGO com análise ontológica .....	26
Figura 6 – Algoritmo K-means.....	28
Fluxograma 1 – Abordagem imuno/neural.....	29
Fluxograma 2 – Metodologia de pesquisa.....	34

## LISTA DE TABELAS

Tabela 1 Classificação histológica dos tumores de glândulas salivares de acordo com a Organização Mundial de Saúde (OMS).....	17
Tabela 2 Genes líderes de cada subtipo histológico dos tumores de glândulas salivares.....	65
Tabela 3 Genes com maiores WNL e menores TIS.....	71
Tabela 4 Genes exclusivamente associados a tumores malignos e a tumores benignos.....	72

## LISTA DE ABREVIATURAS E SIGLAS

BiNGO	Biological Networks Gene Ontology tool
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
HER2	Human Epidermal Growth Factor Receptor 2
HIV	Human Immunodeficiency Virus
HPV	Human Papilloma Virus
MeSH	Medical Subject Headings
NCBI	National Center for Biotechnology Information
NTRK	Neurotrophic Tyrosine Kinase Receptor
OMIM	Online Mendelian Inheritance in Man
OMS	Organização Mundial de Saúde
PAAF	Punção aspirativa por agulha fina
PG	Parotid gland
qRT-PCR	Quantitative reverse transcription polymerase chain
RNA	Ribonucleic acid
RPS15	Ribosomal Protein S15
SLG	Sublingual gland
SMG	Submandibular gland
SPSS	Statistical Package for the Social Sciences
STRING	Search Tool for the Retrieval of Interacting Genes
TIS	Total interaction score
WHO	World Health Organization
WNL	Weighted number of links

## SUMÁRIO

1 INTRODUÇÃO .....	12
2 OBJETIVOS .....	13
2.1 Objetivo Geral .....	13
2.2 Objetivos Específicos .....	13
3 REVISÃO DE LITERATURA .....	14
3.1 Tumores de glândulas salivares.....	14
3.2 Análise de bioinformática.....	20
3.3 Genes líderes e redes de interação biológicas.....	21
3.4 Métodos de clusterização.....	27
3.5 Pesquisa experimental.....	28
4 METODOLOGIA .....	30
5 PRODUTOS .....	31
5.1 Artigo 1: Immune/Neural approach to characterize salivary gland neoplasms....	32
6 CONSIDERAÇÕES FINAIS.....	60
REFERÊNCIAS .....	61
APÊNDICES .....	65
ANEXOS .....	73

## 1 INTRODUÇÃO

O conhecimento dos processos biológicos responsáveis pelo desenvolvimento de neoplasias de glândulas salivares é essencial para identificar potenciais alvos terapêuticos moleculares, marcadores prognósticos e biomarcadores. Esses tumores possuem a particularidade de serem heterogêneos e raros, por isso existem limitações práticas que dificultam seu estudo, ao contrário de outras neoplasias mais comuns como de mama, cólon e pulmão (1). Outro aspecto a ser considerado é a melhora da custo-efetividade das pesquisas realizadas, uma vez que já existe uma extensa base de dados disponíveis na internet para pesquisas em biologia molecular de todos os tipos de tumores (2). Diante dessa extensa quantidade de dados e informações, ao uso da bioinformática com suas diferentes técnicas e algoritmos é essencial para a interpretação desses dados ao gerar novas hipóteses e teorias. A utilização de algoritmos validados de bioinformática é uma ferramenta relevante para pesquisas nessa área das ciências biológicas e biomédicas. O presente estudo visa estudar e analisar os genes líderes dos tumores de glândulas salivares através de métodos de bioinformática validados com ranqueamento dos genes baseado em dados matemáticos e estatísticos como o Weighted number of links (WNL) e o Total interaction score (TIS) acrescidos de um método de clusterização de dados tradicional denominado K-means e outro método inovador, chamado de abordagem imuno/neural (3). Posteriormente, para validar os resultados obtidos e as hipóteses geradas pela bioinformática, foram realizadas pesquisas experimentais apropriadas com o qRT-PCR do termo em inglês quantitative reverse transcription polymerase chain reaction real time.

## 2 OBJETIVOS

### 2.1 Objetivo geral

- Identificar os genes mais associados (genes líderes) às neoplasias de glândulas salivares que possam futuramente ser auxiliares no diagnóstico, no prognóstico e no tratamento.

### 2.2 Objetivos específicos

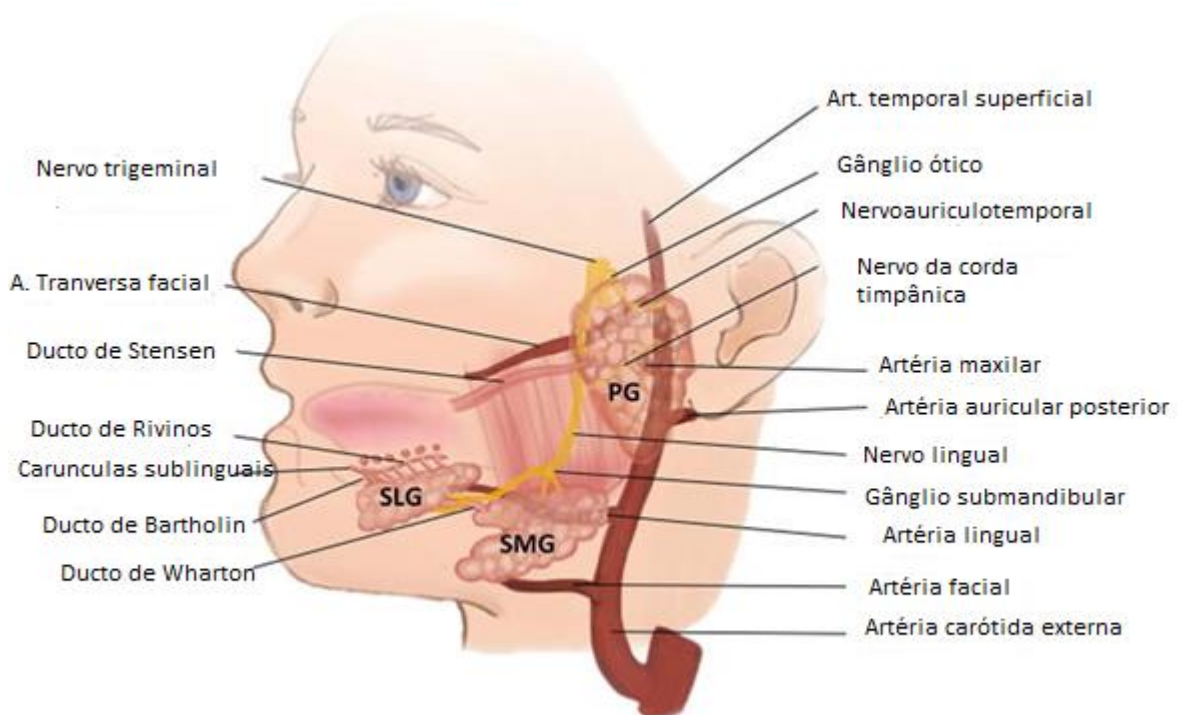
- Verificar quais os genes líderes de cada tipo histológico das neoplasias de glândulas salivares podem ser identificados pela bioinformática.
- Analisar quais os genes líderes poderiam ser potenciais biomarcadores ou utilizados para terapia de alvo molecular.

### 3 REVISÃO DE LITERATURA

#### 3.1 Tumores de glândulas salivares

As glândulas salivares são órgãos exócrinos responsáveis pela produção de saliva, líquido corporal que possui funções protetoras, digestivas e lubrificantes. São constituídas por três pares de glândulas maiores: parótidas, submandibulares e sublinguais além das glândulas menores. Essas últimas são numerosas e se distribuem pela mucosa e submucosa da cavidade oral e pelo trato aerodigestivo superior (4). Na figura 1 está representada a anatomia das glândulas salivares, com a parótida descrita como parotid gland (PG), glândulas sublinguais como sublingual gland (SLG) e glândulas submandibulares como submandibular gland (SMG).

Figura 1 – Anatomia das glândulas salivares

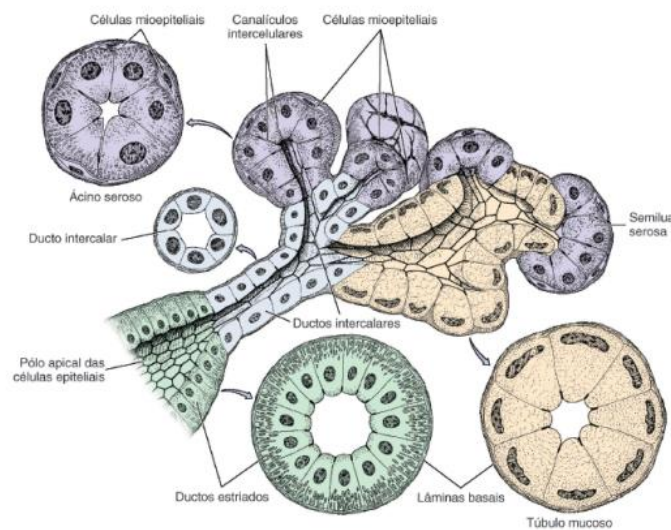


Fonte: Adaptado de Holmberg KV. Anatomy, biogenesis and regeneration of salivary glands. Monographs in oral science. 2014



Como é mostrado na Figura 2, a estrutura histológica de todas as glândulas salivares é constituída por um sistema de ductos ramificados e terminações secretoras que se abrem na mucosa oral e na orofaringe. Sua unidade funcional é formada pelos ácinos serosos, mucosos ou mistos, pelos ductos e pelas células mioepiteliais. As células acinares são envoltas por miofibroblastos, células endoteliais, fibras nervosas, por fibras colágenas e por células imunológicas como linfócitos (4, 5).

Figura 2 – Unidade funcional das glândulas salivares



Fonte: JUNQUEIRA, L. C.; CARNEIRO, J.; ABRAHAMSOHN, P. Histologia básica: texto e atlas. 13. ed. Rio de Janeiro: Guanabara Koogan, 2017.

As glândulas salivares dão origem a mais de 30 tipos de tumores malignos e benignos. Essa grande variedade se deve a heterogeneidade entre os diferentes tipos de tumores. São histologicamente distintos e constituem um grupo raro de neoplasias que podem variar consideravelmente no seu sítio anatômico, no seu comportamento biológico e em aspectos biomoleculares (5, 6).

A glândula parótida é o sítio mais frequente dos tumores de glândulas salivares e constituem aproximadamente 80 a 85% desses tumores (5, 7). Em torno de 75% das lesões parotídeas são benignas e 25% malignas (8). De modo menos frequente, os tumores de glândulas salivares podem surgir nas glândulas submandibular, sublingual e nas glândulas salivares menores. Estes tumores, ao contrário dos originados na parótida, são malignos em 40 a 45% no sítio das

glândulas submandibulares, em 70 a 90% nas glândulas sublinguais e de 50 a 75% nas glândulas salivares menores. Portanto, a probabilidade de um tumor ser maligno é inversamente proporcional ao tamanho da glândula salivar em que se originou (5).

A classificação da Organização Mundial de Saúde (OMS) dos tumores de glândulas salivares foi atualizada em 2017. Conforme essa nova classificação, existem 25 tipos de tumores malignos e 10 tipos de tumores benignos, ambos de origem epitelial (Tabela 1). Essa grande variedade histológica, a existência de tumores híbridos e a possibilidade de tumores benignos progredirem para malignidade, torna muitas vezes o diagnóstico anatomopatológico difícil de ser realizado mesmo para os patologistas mais experientes (7).

Tabela 1: Classificação histológica dos tumores de glândulas salivares de acordo com a Organização Mundial de Saúde (OMS). 2017

Classificação histológica da OMS dos tumores de glândulas salivares OMS

---

Tumores epiteliais malignos	Tumores epiteliais benignos
Adenocarcinoma de células basais	Adenoma canalicular
Adenocarcinoma mucinoso	Adenoma de células basais
Adenocarcinoma não especificado	Adenoma pleomórfico
Adenocarcinoma polimórfico de baixo grau	Adenoma sebáceo
Adenoma pleomórfico metastático	Cistadenoma
Carcinoma adenoide cístico	Linfadenoma
Carcinoma celular escamoso	Mioepitelioma
Carcinoma de células acinares	Oncocitoma
Carcinoma de células claras não especificado	Papiloma ductal
Carcinoma de grandes células	Tumor de Warthin
Carcinoma de pequenas células	
Carcinoma ductal salivar	
Carcinoma epitelial-mioepitelial	
Carcinoma ex adenoma pleomórfico	
Carcinoma linfoepitelial	
Carcinoma mioepitelial	
Carcinoma mucoepidermóide	
Carcinoma oncocítico	
Carcinoma sebáceo	
Carcinoma secretório análogo de mama	
Carcinosarcoma	
Cistoadenocarcinoma	
Cistoadenocarcinoma cribiforme	
Linfadenocarcinoma sebáceo	
Sialoblastoma	

---

Fonte: El-Naggar AK, 2017

A atual classificação da OMS ainda não considera nenhum marcador molecular específico para cada um dos 35 subtipos histológicos. Espera-se que o reconhecimento dessas proteínas e genes possam auxiliar o patologista a realizar o diagnóstico, o estadiamento e até definir o prognóstico de cada tipo de tumor de glândula salivar (6).

O tumor benigno mais comum é o adenoma pleomórfico que é responsável por metade de todos os tumores de glândulas salivares. Já o carcinoma mucoepidermóide e o carcinoma adenoide cístico são responsáveis pela metade de todas as neoplasias malignas dos tumores de glândulas salivares (9, 10).

Tumores de glândulas salivares são incomuns e representam 6 a 8% de todos os tumores de cabeça e pescoço (5, 7). Embora não exista nenhum fator de risco predominante para o seu desenvolvimento, há vários fatores que são possivelmente envolvidos como a radiação (11), cosméticos, compostos de níquel e produtos utilizados na fabricação de determinados tipos de borracha (5, 12, 13). Já os tumores de Warthin têm uma forte associação com o tabagismo, ao contrário dos outros tipos histológicos em que essa associação não está claramente estabelecida (7, 14, 15). Infecções virais como Epstein Barr (7), vírus da imunodeficiência humana (HIV) (16, 17) e papilomavírus humano (HPV) (18) podem estar associadas ao aumento no risco de neoplasias de glândulas salivares.

A manifestação clínica tipicamente se apresenta como uma massa indolor ou inchaço da parótida, glândulas submandibulares ou sublinguais. Outro modo de apresentação é a presença de massas ou ulcerações em palato, lábios ou mucosa bucal. Os sítios de metástases mais comuns são pulmão, seguido de ossos e fígado (18). A tomografia computadorizada e a ressonância magnética podem ser necessárias para avaliar a extensão dos tumores de glândulas salivares assim como uma possível invasão óssea, do espaço perifaríngeo e o acometimento linfonodal (19, 20). Apesar de não ser útil para o planejamento cirúrgico o ultrassom pode facilitar a punção aspirativa por agulha fina (PAAF). O diagnóstico citológico deve sempre ser obtido para evitar cirurgias agressivas para tumores benignos ou linfomas e pode ser realizada pela PAAF ou biópsia por agulha guiada por ultrassom (21, 22).

A base do tratamento é a ressecção cirúrgica. Pacientes com tumores benignos ou malignos de baixo grau são tipicamente tratados somente com cirurgia (23). Aqueles com tumores de alto grau e margens da ressecção cirúrgica positiva são tratados com cirurgia e radioterapia adjuvante (24). Os tumores irresssecáveis são tratados com radioterapia isolada ou radioterapia associada à quimioterapia. O tratamento paliativo nos pacientes com doença metastática é feito com

quimioterapia e terapia de alvo molecular. Não há ensaios clínicos adequados para definir a melhor abordagem terapêutica para neoplasias de glândulas salivares metastáticas principalmente por causa da raridade desses tumores (1). A terapia de alvo molecular se tornou possível com a melhor compreensão das modificações moleculares que ocorrem nos tumores malignos de glândulas salivares, permitindo a identificação de inúmeros potenciais alvos terapêuticos. Drogas como os inibidores de tirosina quinase: imatinib (1) e dasatinib (25), os inibidores do receptor do fator de crescimento epidérmico: gefitinib (26) e cetuximab (27), o inibidor seletivo de *NTRK*: larotrectinib (28) e o anti-*HER2*: trastuzumab são os principais exemplos da terapia de alvo molecular utilizados para o tratamento das neoplasias de glândulas salivares (29).

Nos últimos anos, houve um aumento considerável nas pesquisas em biologia molecular nos tumores de glândulas salivares que são realizadas na tentativa de definir novos biomarcadores que auxiliem na classificação e no diagnóstico histopatológico dos trinta e cinco diferentes tipos de tumores que exibem diferentes propriedades moleculares. O contínuo desenvolvimento de novas técnicas biomoleculares vem contribuindo para o conhecimento da biologia e da gênese desses tumores, embora esse processo seja muito mais lento quando comparados com as outras neoplasias como as de cólon, pulmão ou mama. Isso se deve a relativa raridade dos tumores de glândulas salivares e do longo tempo necessário para obter amostras teciduais (6).

Os mesmos tipos estudos também procuram marcadores moleculares para a avaliação de prognóstico e para desenvolvimento de novos tratamentos. Um dos exemplos disso é o estudo que mostram altos níveis de expressão nos genes *EGFR/HER-2* no carcinoma mucoepidermóide e no carcinoma adenoide cístico (30). Outras pesquisas evidenciaram que no carcinoma mioepitelial e no carcinoma adenoide cístico, existe uma maior expressão do proto-oncogene tirosina quinase *C-Kit*, do fator de transcrição NF- $\kappa$ B e das proteínas Ras e Wnt/ $\beta$  cateninas (31). No adenoma pleomórfico, a proteína reguladora da actina (*ENAH*), a proteína do tumor de Wills 1 (*WT1*), aquaporina 5 (*AQP5*), gene 1 do adenoma pleomórfico (*PLAG1*), catenina Beta 1 (*CTNNB1*) e beta defensina 1 (*DEFB1*) apresentaram aumento ou diminuição da expressão gênica. Já no carcinoma mucoepidermóide foi identificado uma translocação entre os genes CREB coativador de transcrição regulada (*CRTC1*), anteriormente identificado pelo símbolo *MECT1* e o gene mastermind like transcriptional coactivator 2 (*MAML2*) (6).

Outros estudos ainda evidenciam que a maior expressão do fator indutor de hipóxia 1 (*HIF-1*) dependente de fator vascular de crescimento endotelial A (*VEGFA*) está associado com a uma tendência de produzir metástases no carcinoma de células acinares e teoricamente poderiam ser um alvo terapêutico (32). Portanto o desenvolvimento de diferentes e novas técnicas de biologia

molecular torna possível o maior conhecimento da gênese e da biologia dos tumores de glândulas salivares principalmente ao estudar a regulação genética de diferentes genes e proteínas (6).

### 3.2 Análise de bioinformática

Bioinformática pode ser definida como uma abordagem interdisciplinar que utiliza métodos estatístico-matemáticos e ciências da computação para visualizar e analisar dados em ciências biológicas e biomédicas com a finalidade de descobrir novos padrões e gerar novas teorias, hipóteses e modelos. Disciplinas como bioquímica, biologia celular, biologia do desenvolvimento, genética, genômica e fisiologia contribuem para uma enormidade de informações geradas pelos estudos do funcionamento dos genes em um organismo, tornando necessário o domínio de programação de softwares, o desenvolvimento e o manejo de extensos bancos de dados. Por isso pesquisadores como biólogos, engenheiros elétricos, engenheiros mecânicos, cientistas da computação e matemáticos usualmente conduzem estudos em bioinformática para facilitar a análise e a interpretação de dados provenientes de experimentos biológicos e estudos observacionais (33, 34). Em outros termos, a bioinformática é um campo da biologia que desenvolve técnicas para coleta e manipulação de dados biológicos para fazer previsões biológicas ou possíveis novas descobertas. Pertencem a esse campo todos os métodos e teorias computacionais para solucionar problemas biológicos (35).

A bioinformática surgiu como uma importante disciplina logo após o desenvolvimento de modernas tecnologias de sequenciamento de DNA. E o término e a publicação do projeto genoma humano foi o marco de uma nova era nas pesquisas genéticas e isso levou a um aumento significativo dos estudos e das publicações em bioinformática (36, 37). Nas últimas décadas, após o desenvolvimento de tecnologias de alto rendimento como os microarrays, houve um aumento exponencial no volume de informações e dados gerados em pesquisas científicas de genética e biologia molecular. A interpretação desse grande volume de informações tornou se cada vez mais difícil. Por isso diferentes técnicas de bioinformática foram desenvolvidas (38).

### 3.3 Genes líderes e redes de interação biológicas

As associações potenciais entre determinadas variações gênicas com certas doenças ou a susceptibilidade a elas tornaram-se um novo e atrativo campo de pesquisa. Muitos genes são associados com uma única patologia de transmissão mendeliana. Entretanto, essa teoria não explica nem é capaz de associar determinadas sequências gênicas com doenças complexas devido a três fatores: o primeiro, doenças complexas são associadas a um grande número de genes, cada um contribuindo com um pequeno aumento do risco de desenvolvimento da patologia. O segundo fator é que patologias complexas têm apresentações clínicas variadas, o que dificulta ou inviabiliza o reconhecimento de um fenótipo específico e, conseqüentemente, a escolha da melhor amostra da população a ser estudada. Por último, os mecanismos dessas patologias envolvem muitos processos biológicos diferentes. A identificação de genes específicos para determinados processos patológicos sempre foi desafiador. Por isso foram criadas bases de dados que registram as variações genéticas e padrões de variações genéticas mais comuns para enfermidades específicas. Essas bases de dados se tornam recursos para facilitar estudos sobre associações genéticas (39).

Em qualquer sistema biológico as moléculas interagem entre si, formando complexos moleculares e vias bioquímicas que, por sua vez, geram funções biológicas (40). Sabe-se que as interações que ocorrem entre diferentes proteínas são o cerne de importantes e complexos processos celulares e seu estudo sistemático torna possível uma compreensão mais completa dos sistemas moleculares biológicos e do funcionamento da maquinaria celular (41, 42). Para melhor compreendermos a função biológica de cada proteína é importante conhecer sua interação com outras proteínas, assim conhecendo melhor a organização funcional de um proteoma (43, 44). Pode-se usar o conhecimento produzido nas interações proteína-proteína para entender os mecanismos das doenças e para construir mapas de interações mais específicos em determinadas enfermidades como já foi realizado, por exemplo, na coreia de Huntington (45).

Um dos métodos de bioinformática mais tradicionais é a pesquisa de genes líderes que consiste na análise sistemática de genes envolvidos em determinados processos patológicos, seguida de sua classificação por ordem de importância de acordo com o número de interações estabelecidas experimentalmente. Esse ranqueamento é baseado no cálculo das interações entre os genes o que é denominado *Weighted number of links (WNL)* que, por sua vez, é calculado a partir dos valores das interações proteicas obtidos no *STRING* para cada gene e a soma dos escores dessa associação multiplicado por 1000. Os genes que possuem maiores interações e, portanto, são

ranqueados nas primeiras posições são definidos como genes líderes porque, presumivelmente, possuem um importante papel na patologia ou processo biológico estudado. Outro escore importante é o total interaction score (TIS) que é calculado usando todas as interações gênicas de um gene específico com todo genoma da espécie em estudo. Os genes que apresentam os maiores valores da razão entre WNL/TIS possuem um escore de especificidade mais alto e por isso possuem influência mais específica para determinada via biológica ou conjuntos de genes e proteínas em análise. Os valores desses dois escores de interações (WNL e TIS) são encontrados a partir de bancos de dados disponíveis na internet como Search Tool for the Retrieval of Interacting Genes (STRING) (46-48). A identificação dos genes líderes está diretamente relacionada com o controle da expressão gênica e pode ser aplicada praticamente em qualquer investigação de regulação genética e de biologia molecular. Esse método sugere os genes mais relevantes no processo celular que está sendo pesquisado, baseado nos dados experimentais disponíveis. O que é útil para a interpretação da tecnologia de microarray e para auxiliar na elaboração de novas pesquisas como, por exemplo, experimentos clínicos (49, 50). O método de bioinformática com a pesquisa de genes líderes utilizado no presente estudo foi desenvolvido com procedimentos e recursos descritos abaixo.

O Medical Subject Headings (MeSH) da U.S. National Library of Medicine (NLM) é um vocabulário controlado utilizado para permitir o uso de uma terminologia comum e uma linguagem única para indexar, catalogar e pesquisar artigos e outros produtos científicos na base de dados da Medline e PubMed nas áreas biomédicas e de ciências da saúde (51).

A base de dados PubMed possui um conjunto de mais de 29 milhões de citações indexadas da literatura biomédica da Medline (medicina, enfermagem, farmácia, odontologia, veterinária e outras ciências da saúde), além de jornais em ciências naturais e livros online armazenados e organizados para acesso e pesquisa. Está em funcionamento desde 1996 (52).

Genecards® é um outro tipo de banco de dados que integra as informações obtidas de outros 150 bancos de dados que armazenam informações de genes humanos, as funções das proteínas que esses genes codificam e as doenças e enfermidades relacionadas. Fornece informações importantes como o nome oficial do gene, seus sinônimos, nomes previamente utilizados, os genes homólogos, a localização no cromossoma, o nome da proteína codificada com descrição da sua função celular, o padrão de expressão gênica, as semelhanças com outras proteínas e as doenças possivelmente relacionadas (53).



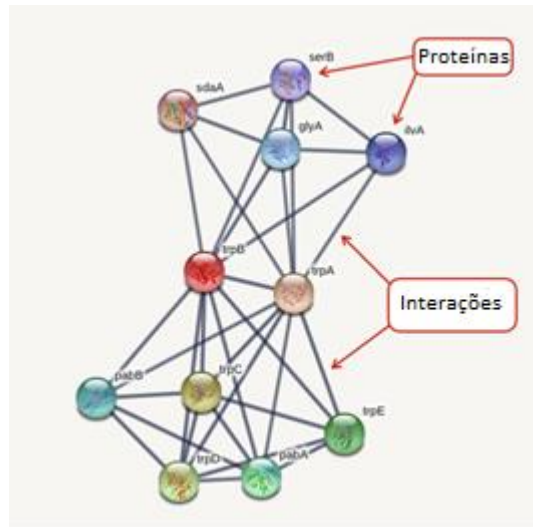
Online Mendelian Inheritance in Man® (OMIM) é um compêndio de genes humanos e fenótipos genéticos disponíveis gratuitamente na internet com informações de mais de 15.000 genes. Tem foco na relação entre genótipo e fenótipo e está disponível na base de dados do National Center for Biotechnology Information (NCBI). É atualizado diariamente e está disponível no site <http://www.ncbi.nlm.nih.gov/omim/>. O OMIM é amplamente utilizado por médicos, estudantes e pesquisadores e facilita as pesquisas e estudos sobre a complexa relação entre os genes e as doenças (54, 55).

O HUGO Gene Nomenclature Committee foi elaborado para padronizar a nomenclatura em genética humana e para assegurar que os nomes aprovados e aceitos estejam em alinhamento com as necessidades da comunidade científica mundial. Para cada gene é aprovada uma abreviação denominada como símbolo do gene seguida de um nome descritivo. Cada símbolo é único para evitar equívocos nas publicações científicas (56).

Search Tool for the Retrieval of Interacting Genes (STRING) é uma base de dados disponível na internet que armazena e integra a maior parte das informações disponíveis sobre as interações entre proteínas, incluindo as diretas (físicas) como as indiretas (funcionais). O conhecimento das interações proteína-proteína são oriundas de estudo experimentais com técnicas bioquímicas, biofísicas e genéticas, e também de interações proteicas previstas através de métodos computacionais, assim como o estudo de vias bioquímicas comuns entre proteínas, de técnicas de extração automática de resumos do Medline (text-mining) e de co-expressão gênica (41). Assim que o conjunto de genes ou proteínas em estudo é inserido no STRING, é criado um mapa de interação proteína-proteína, como é mostrado na Figura 3, e ainda é possível analisar a confiabilidade das interações e reajustar os níveis de corte de confiabilidade que serão aceitos. Também é possível extrair dados e análises estatísticas do mapa de interações obtido como, por exemplo, obter os valores de Weighted number of links (WNL).

A Figura 3 mostra um exemplo de uma rede de interações proteína-proteína do STRING que sumariza as interações existentes entre determinado grupo de proteínas. As esferas representam as proteínas, e as linhas as associações preditas entre elas. A cor das linhas que interligam as proteínas representam qual tipo de informação e evidência científica foi utilizada para definir as interações. Por exemplo, linhas amarelas são derivadas de interações obtidas em técnicas de mineração de dados de resumos dos artigos e as linhas púrpuras representam interações obtidas por através de estudos experimentais.

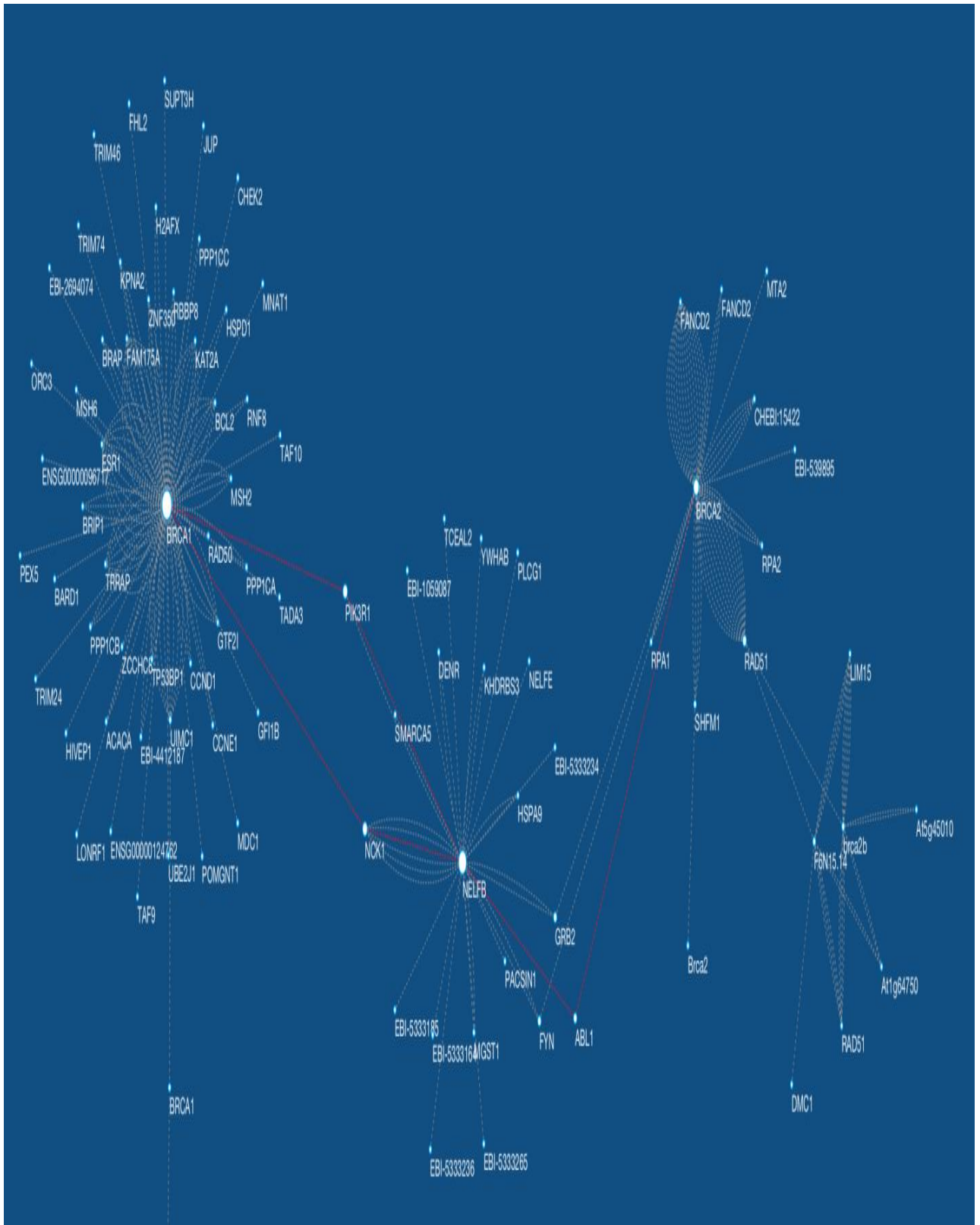
Figura 3 – Rede de interações proteína proteína do STRING



Fonte: Adaptado de <https://string-db.org/cgi/help.pl?sessionId=u50CKfYmJVzA>

As redes de interação biológicas criadas por softwares são essenciais para compreensão dos sistemas biológicos e das vias biológicas. O Cytoscape é uma plataforma de software de código aberto que integra redes de interações biomoleculares em um mapa de conceitos unificado, permitindo a visualização de uma representação gráfica usualmente gerada para tentar um melhor entendimento de sistemas biológicos como determinados processos celulares ou mesmo enfermidades. Ele explora visualmente redes biomédicas compostas principalmente por genes e proteínas. Além de permitir a integração, a análise e a visualização de redes complexas geradas em pesquisas biológicas ou em outras áreas do conhecimento, o Cytoscape também é capaz de gerar textos explicativos como perfis de expressão gênica e anotação genômica assim como dados numéricos como o coeficiente topológico e o power law (40, 57). Uma das ferramentas mais importantes desse software é a criação de uma interface visual interativa e versátil composta de nós, arestas e componentes conectados em uma rede de duas dimensões como representado na figura 4 (58). Os nós, também chamados de vértices, habitualmente representam proteínas, genes ou metabólitos. Enquanto as arestas representam as inter-relações como interações físicas ou regulação de expressão gênica (59).

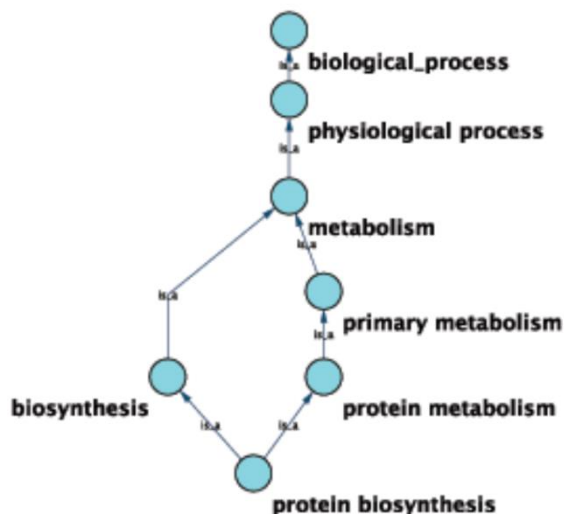
Figura 4 – Rede biológica de genes visualizada com o Cytoscape



Fonte: [https://cytoscape.org/what\\_is\\_cytoscape.html](https://cytoscape.org/what_is_cytoscape.html)

Uma das características mais interessantes e úteis do Cytoscape é a disponibilidade de Apps que podem realizar análises adicionais aos protocolos básicos (40). Anteriormente denominados como Plugins (até a versão 2.x), os Apps ou aplicativos são recursos adicionais disponíveis gratuitamente. Foram desenvolvidos para realizar diferentes funções e estender o Cytoscape de várias maneiras como, por exemplo, realizar outro método de analisar redes e perfis moleculares ou importar dados de algum banco de dados on-line. Podem ser desenvolvidos por qualquer usuário já que este é um programa de código aberto. Um dos aplicativos do Cytoscape mais utilizados é o Biological Networks Gene Ontology tool (BiNGO) que realiza uma análise ontológica caracterizando quais princípios e funções celulares que fazem parte de um determinado conjunto de genes ou proteínas como é demonstrado na Figura 5 (38).

Figura 5 – Gráfico do BiNGO com análise ontológica

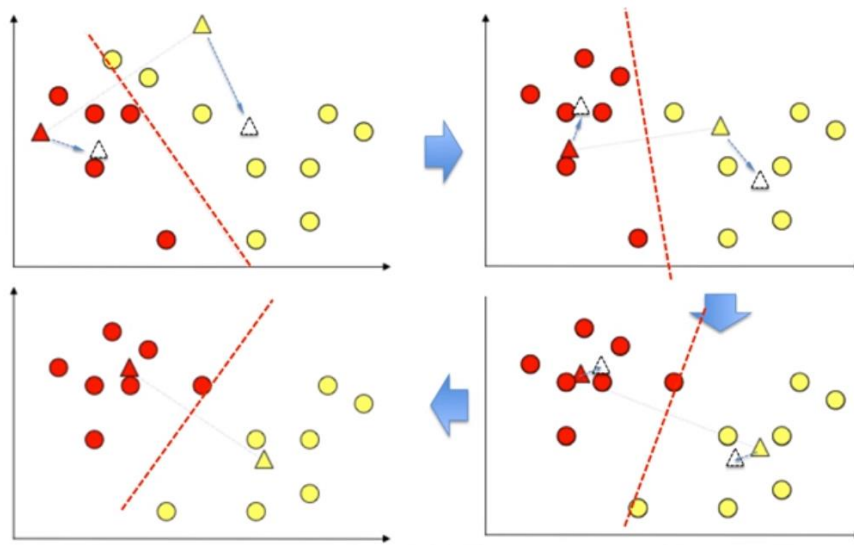


Fonte: <https://www.psb.ugent.be/cbd/papers/BiNGO/Tutorial.html>

### 3.4 Métodos de clusterização

Métodos ou análises de clusterização tem o objetivo de reunir em subgrupos homogêneos um determinado conjunto de dados ou de observações. Sendo que os fatores agrupados em um cluster possuem mais similaridades que os agrupados em outro. Portanto esses algoritmos buscam agrupar diferentes observações ou fatores em diferentes agrupamentos. Tais análises são utilizadas em diversos tipos de aplicações como em processamento de documentos ou mesmo em biologia molecular e na genética moderna. Exemplos mais concretos da aplicação desses algoritmos de clusterização são o agrupamento de emails em spam e não spam e as observações e dados obtidos em estudo de genética como nos microarrays (60). O algoritmo de K-means é um dos métodos mais populares de clusterização e é aplicável para qualquer dado que seja quantitativo e que se possa mensurar uma distância euclidiana entre os fatores que possuem determinados valores plotados em um gráfico ao longo de duas variáveis distintas, formando uma figura bidimensional. Assim como mostrado na figura 6, aqueles dados que tem suas mensurações em duas variáveis mais próximas do outro determinado dado são aglomerados em um subgrupo. A distância euclidiana entre eles é calculada usando geometria básica. Ao se desenhar uma linha reta entre duas observações, e então, traçar se uma bissetriz perpendicular cruzando um ângulo de  $90^\circ$  no ponto médio da linha que une os dois pontos em um gráfico, é possível definir quais observações, dados ou fatores estão mais próximos entre si, dessa maneira são formados diferentes aglomerados ou agrupamentos. Os pontos ou as observações iniciais a partir dos quais a distância euclidiana é calculada são definidos aleatoriamente pelo próprio algoritmo. Nesse estudo, as variáveis são os valores de WNL e do TIS, dividindo os genes em agrupamentos distintos. No K-means o número desses subgrupos homogêneos, ou seja o valor do K, são fixos e são definidos pelo pesquisador previamente ao emprego do algoritmo. O próprio pesquisador tem que decidir qual é o número exato de agrupamentos ou aglomerados antes de iniciar o processo de clusterização (61). O problema de estimar o número de subgrupos para um determinado conjunto de dados é complexo e difícil, sendo considerada uma importante limitação do uso do K-means (2). Existem vários métodos propostos para escolher o número de subgrupos ou clusters K sendo um dos mais usados o “gap statistic”(62).

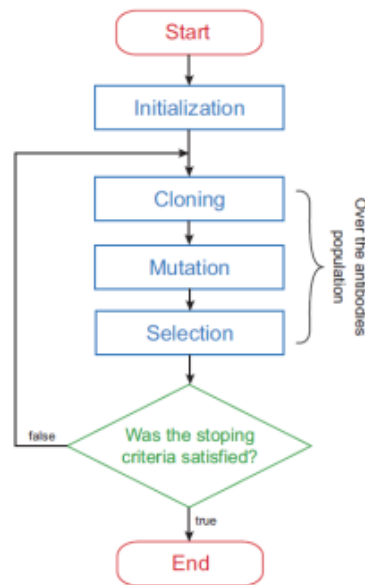
Figura 6 – Algoritmo K-means



Fonte: <https://webcourse.cs.technion.ac.il>

O outro método de clusterização utilizado foi criado recentemente e é baseado no algoritmo da abordagem imuno/neural que é inspirado no funcionamento do sistema imune associado com a rede neural de Kohonen (63). O sistema imune tem a capacidade de responder contra qualquer substância estranha ao organismo, o que é denominado com antígeno, usando elementos chamados anticorpos. A abordagem imuno/neural, um algoritmo de clusterização, considera a afinidade dos anticorpos formados pelo sistema imune, e escolhe aquele que melhor responde ao anticorpo através de um processo chamado de mutação ou maturação. No caso desse estudo a maior afinidade do anticorpo é aquela que melhor detecta a distância euclidiana dos genes em suas variáveis de WNL e TIS em um gráfico bidimensional. O processo se inicia com a criação de um anticorpo em uma posição central nesse gráfico e, após isso, cada anticorpo passa por três fases: clonagem, mutação e seleção, conforme é ilustrado no fluxograma abaixo (64).

## Fluxograma 1 – Abordagem imuno/neural



Fonte: D'Angelo et al, 2016

Na abordagem imuno/neural, durante o processo de clonagem cada anticorpo é clonado duas vezes e um dos clones é submetido a uma mutação aleatória na sua posição no gráfico bidimensional. Essa etapa é importante para garantir a grande variedade de anticorpos. Essa mutação é aprimorada pela rede neural de Kohonen com uma redução linear da distância entre cada antígeno e seu anticorpo correspondente. Já a etapa de seleção acontece quando dois anticorpos próximos em sua posição reconhecem o mesmo antígeno (gene com o valor de WNL e TIS) e então um novo anticorpo é criado em uma posição equidistante entre os dois anticorpos inicialmente citados. Os critérios de parada para as inúmeras repetições dos processos de clonagem, mutação e seleção são a convergência do algoritmo e a limitação de um número máximo de interação pelo integer K (64).

Uma limitação que deve ser considerada nesse estudo é uma metodologia de pesquisa eminentemente de base teórica usando diferentes bases de dados e métodos estatístico-matemáticos para criar novas hipóteses ou teorias. Por outro lado, uma grande vantagem, é a melhora da relação custo-benefício, uma vez que torna a análise dos dados obtidos menos complexa e possibilita a utilização de métodos experimentais mais específicos, racionais e custo efetivos (2, 65).

### 3.5 Pesquisa experimental

Com base nas hipóteses geradas pela bioinformática que evidenciou quais são os genes líderes associados aos tumores de glândulas salivares e buscando o método experimental mais adequado para validação dos dados, foi proposto a utilização da técnica de isolamento de RNA. Esse método tem a vantagem de avaliar o transcriptoma, ou seja, quais genes estão sendo expressados nesse contexto biológico. O estudo e análise isolada do DNA, isto é, do genoma só pode explorar os genes e não sua expressão e sua tradução em cadeias polipeptídicas (as proteínas ) que são responsáveis pela maioria das funções biológicas na maquinaria celular. O desenvolvimento do qRT-PCR quantitativo em tempo real (do termo em *inglês quantitative reverse transcription polymerase chain reaction real time*) permitiu a análise da transcrição gênica. O que representou um grande avanço nas pesquisas em genética e biologia molecular. O papel central do RNA mensageiro na transmissão da informação genética torna seu estudo essencial para aumento do conhecimento no estudo do câncer (66).



## 4 METODOLOGIA

A primeira etapa deste estudo foi a fase de revisão bibliográfica após a definição dos termos de busca pelo Medical Subject Headings (MeSH). Cada tipo histológico dos tumores de glândulas salivares foi pesquisado na base de dados do Pubmed e em seguida foram selecionados os artigos sobre a expressão gênica com identificação de genes relacionados aos tumores de glândulas salivares. Posteriormente, com os mesmo termos de busca e operadores booleanos, foram realizadas pesquisas em duas outras bases de dados, o Genecards® e o OMIM® que também mostraram os genes relacionados a cada tipo histológico de todos os tumores de glândulas salivares. Para certificar que não haveria problemas com a nomenclatura e com nas abreviações dos genes humanos pesquisados, foram realizadas pesquisas no HUGO Gene Nomenclature Committee.

Após a identificação do conjunto de genes mais relacionados a cada tipo histológico, esses genes foram expandidos na ferramenta Search Tool for the Retrieval of Interacting Genes (STRING) versão 10.0, na qual foi feita uma expansão dos genes identificados, sendo criada uma rede de interação gênica e protéica baseada somente em observações experimentais e com nível de confiança máximo (0,900). Desta rede pode se extrair diversos dados como o descritivo dos genes e os valores numéricos das interações entre proteínas como o WNL e o TIS, que, por sua vez, foram inseridos no programa Cytoscape versão 3.2.0. Este programa criou outra representação gráfica das vias de interação proteicas além de fazer análises destas redes, criando outros dados como o coeficiente topológico e o power law.

A análise ontológica que é a representação das vias moleculares foi obtida pelo aplicativo BinGO dentro do próprio Cytoscape. Além disso, a partir dos valores das interações proteicas obtidos no STRING foram calculados, para cada gene, a soma do escores de associação e este valor ajustado com a multiplicação por 1000, obtendo se assim um valor único denominado weighted number of links (WNL) que corresponde ao valor total de ligação dentro de uma determinada via biológica.

Outro dado obtido é o Total Interaction Score (TIS) que corresponde ao número de ligação de determinado gene ou proteína com todas as proteínas humanas. O TIS é calculado através de uma base de dados com mais de 8.500.000 interações gênicas, sendo necessário o uso do programa de gerenciamento de dados com alta capacidade de armazenamento como o Microsoft Access 2010.

Usando os valores de weighted number of links (WNL) e do Total Interaction Score (TIS) de cada gene relacionado a um tipo específico de tumor de glândula salivar, os genes foram agrupados utilizando no SPSS (versão 18.0, IBM, NewYork, NY, USA) o algoritmo de clusterização K-means seguido da análise estatística dos resultados pelo ANOVA e Tukey-Kramer post hoc test, considerando o valor de  $P < 0,001$ . Os genes de maior escore foram identificados como genes líderes.

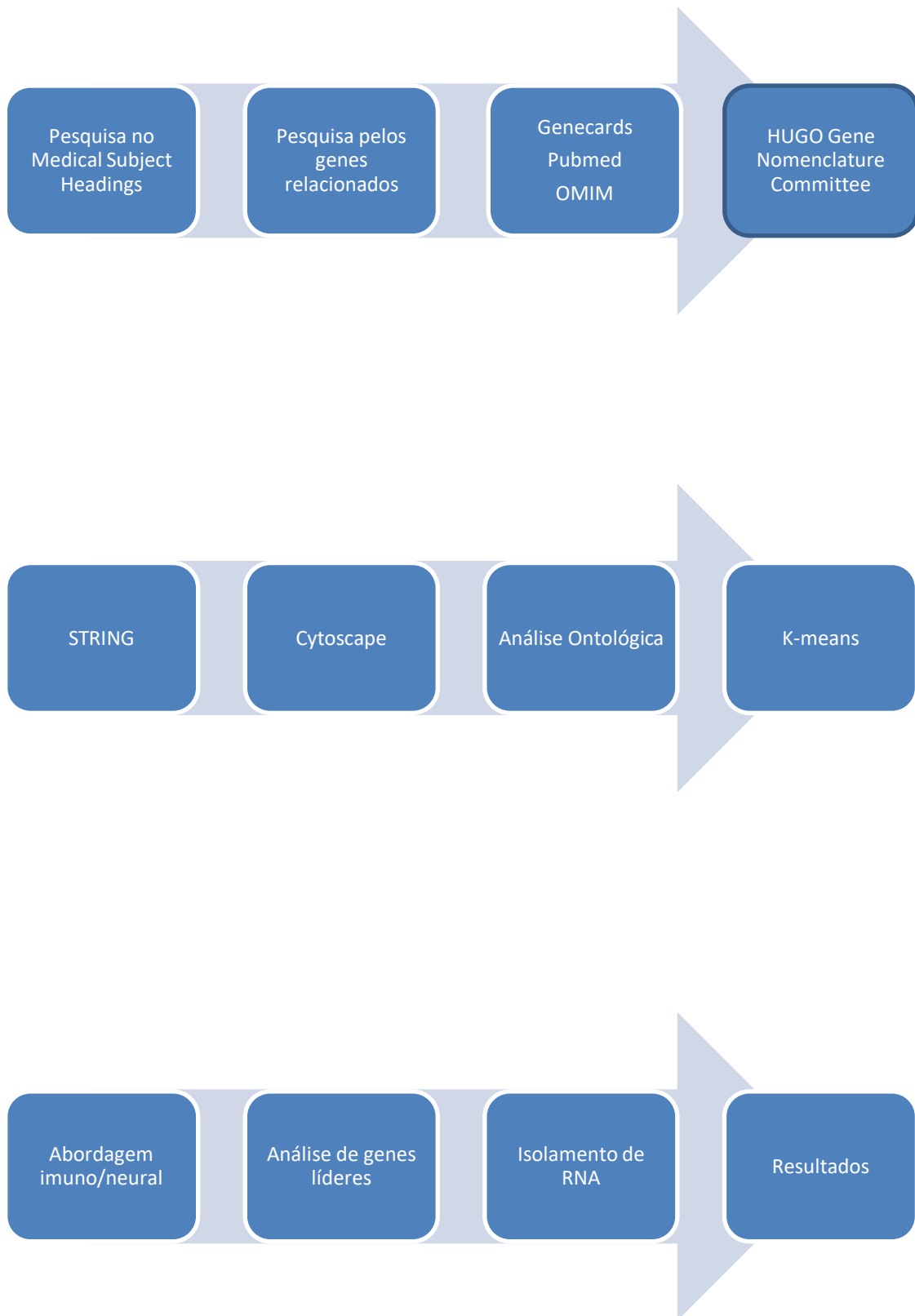
Outro algoritmo de clusterização usado neste estudo foi o chamado de abordagem imuno/neural que é baseado e inspirado no sistema imune que considera a afinidade dos anticorpos pelos antígenos estabelecida pela distância euclidiana entre os pontos que, por sua vez, são determinados pelos WNL e pelo TIS em um gráfico bidimensional.

Após a identificação dos genes líderes de cada tipo histológico dos tumores de glândulas salivares, eles foram divididos nos genes líderes associados exclusivamente aos tumores benignos, os exclusivamente associados aos malignos e os associados a ambos. Para cada um dos dois primeiros grupos, foi realizada uma nova expansão pelo STRING, seguida da obtenção dos valores de WNL e TIS e uso do K-means e do imuno/neural para identificação de novos genes líderes que, por sua vez, foram submetidos a nova expansão pelo STRING com a identificação de um gene com maior WNL e menor TIS dos tumores benignos e dos malignos. Para avaliação da transcrição gênica desses dois últimos genes identificados foi realizada uma pesquisa experimental.

Para validar os resultados em bioinformática obtidos pelos métodos citados acima, o isolamento de RNA foi realizado com o método quantitative reverse transcription polymerase chain reaction real time (qRT-PCR). Usando o RNA mensageiro de 13 amostras de tumores de glândulas salivares, o RNA foi isolado com o reagente Trizol (Thermo Fisher Scientific, Waltham, MA, USA) e posteriormente tratado com a enzima DNase I- Amplification Grade (referência de catálogo 18068015, Invitrogen, Carlsbad, CA, USA) e and 1.5 µg de RNA foi submetido a transcrição reversa usando o SuperScript® First-Strand Synthesis System para qRT-PCR (referência de catálogo 11904018, Invitrogen, Carlsbad, CA, USA). Cada reação de SYBR green com um volume total de 20 µl com 10 µl de cDNA e 8.5 µl de H<sub>2</sub>O. As condições de ciclagem térmica foram 95 °C por dez minutos, seguidas de 40 ciclos de 95 °C por 15 segundos e de 60 °C por um minuto. Primers para *RPS15* foram o forwad 5' - TGGAGGCTCGAATTATTGCCTTG - 3' e o reverse 5' - TAGCGTTTGTGGGCTTTGTC - 3'. A beta actina foi utilizada para

normalizar a expressão de *RPS15*. O PCR quantitativo foi realizado no StepOne Real-Time PCR System (Life Technologies, Carlsbad, CA, USA).

## Fluxograma 2 – Metodologia de pesquisa



## 5 PRODUTOS

5.1 Produto 1: *Immune/Neural approach to characterize salivary gland neoplasm (SGN)* formatado segundo as normas para publicação do periódico Applied Soft Computing, enviado em fevereiro de 2019.

Immune/Neural approach to characterize salivary gland  
neoplasms (SGN)

Carlos Rafael Lima Monção<sup>a</sup>, Eloa Mangabeira Santos<sup>a</sup>, Thiago Silva Prates<sup>a</sup>,  
Alfredo Maurício Batista de Paula<sup>b</sup>, Claudio Marcelo Cardoso<sup>a,d</sup>, Lucyana  
Conceição Farias<sup>b</sup>, Sérgio Henrique Sousa Santos<sup>e</sup>, Marcos Flávio Silveira  
Vasconcelos D'Angelo<sup>c,\*</sup>, André Luiz Sena Guimarães<sup>b,\*</sup>

<sup>a</sup>Graduate Program in Health Sciences,  
State University of Montes Claros, Montes Claros, Brazil

<sup>b</sup>Department of Dentistry,  
State University of Montes Claros, Montes Claros, Brazil

<sup>c</sup>Department of Computer Science,  
State University of Montes Claros, Montes Claros, Brazil

<sup>d</sup>Dilson Godinho Hospital, Montes Claros, Brazil

<sup>e</sup>Institute of Agricultural Sciences,  
Federal University of Minas Gerais, Montes Claros, Brazil

\*Corresponding author

Email addresses: marcos.dangelo@unimontes.br (Marcos Flávio Silveira Vasconcelos  
D'Angelo ), andreluizguimaraes@gmail.com (André Luiz Sena Guimarães )

## Abstract

Considering the controversial literature regarding the relationship between hypoxia and SGNs, the purpose of the current study was to use immune-inspired algorithm ClonALG whose performance is increased by using the Kohonen neural network training algorithm (Immune/Neural approach), to characterize the nature of salivary gland neoplasms (SGNs). The leader gene approach in order to identify biomarkers for SGNs. Extensive data were obtained for each of the 35 types of neoplasms. The gene leaders for each type of SGN were identified in a table and then divided according to the two different methods: K-means clustering and immune/Neural approach. Genes related to SGNs were identified using PubMed, OMIM and Genecards databases. A bioinformatics algorithm was then applied, and the STRING database was employed to build networks of protein-protein interactions for each nature of an SGNs. The weighted number of links (WNL) and total interactions score (TIS) values were then obtained. Finally, the genes were clustered, and the gene leaders were identified using the K-means clustering method and the Immune/Neural approach.

**Keywords:** Bioinformatics, Salivary gland neoplasm, Immune/Neural Approach

## 1. Introduction

Salivary gland neoplasms (SGNs) are a group of heterogeneous and rare neoplasms [19]. They vary considerably in their anatomic site of origin, histology and biological behavior. The World Health Organization (WHO) classifies SGNs according to their histopathological features [3] [37]. The great variety of named neoplasms is due to the heterogeneity between different types of neoplasms [3] [37]. It is possible that the presence of hybrid neoplasms and progression to malignancy are associated with this morphological variability [3] [37]. These characteristics can make histopathological diagnosis difficult [3] [12]. There is no one predominant factor known to be implicated in the development of cancer of the salivary gland, although there are many factors that may play a role [35], such as radiation, cosmetics, nickel compounds and tobacco [21] [34] [41]. A biopsy of an SGN is required to make the histopathological diagnosis and to conclude whether it is benign or malignant [19]. However, diagnosis can be challenging, even for an expert pathologist [19]. Moreover, complementary molecular tests are often required to diagnose SGNs [36]. Surgical resection is the ideal treatment for SGNs. However, radiation therapy is also used in high-grade neoplasms, those with a positive surgical margin, and for the inoperable disease. Currently, chemotherapy and molecular-targeted therapies are reserved for the treatment of recurrent and metastatic disease [25]. Some studies have focussed on the expression of biological targets in salivary gland neoplasms, including c-kit, EGFR, HER2, and AR, among others, all of which could be potential biomarkers, prognostic markers or molecular targets [1] [4] [7] [14] [16] [20] [22] [24] [30] [39] [40]. However, only a few studies have explored large-scale screening techniques. Bioinformatics approaches have provided meaningful results in the analysis of a wide variety of diseases [17] [18] [10] [11] [32]. The use of bioinformatics is helpful for genomic and proteomic analyses. Moreover, bioinformatics could contribute to the discovery of new cancer biomarkers. The leader gene approach has been used and improved for specific conditions [17] [18] [10] [11] [32]. In the current study, we incorporated an immune/neural classification [8] with the leader gene approach in order to characterize the nature of SGNs. A method based on hybrid immune/neural formulation is used to increase the classification performance in comparison with K-means. This immune/neural classification method is based on the association of ClonALG immune systems [9] with the Kohonen neural network [23]. In addition, to combine these two approaches, the stopping criteria and selection mechanism had to be changed to accommodate a non fixed number of antibodies. For this methodology, no mathematical or statistical models are required, decreasing the problem of implementation complexity of the algorithm.



## 2. Materials and Methods

### 2.1. Bioinformatics and Systems Biology Analysis

A bioinformatics approach was performed as described previously [17] [18] [10] [11] [32]. Briefly, we firstly performed searches using Medical Subject Headings (MeSH) terms, comprising all types of SGNs. The second step was to identify genes involved in SGNs in the on the subject was held search in the MEDLINE / PubMed (<http://www.ncbi.nlm.nih.gov/>), Online Mendelian Inheritance in Man (OMIM) and Genecards ([www.genecards.org](http://www.genecards.org)) databases, only considering human genes [15] [31]. We then checked the names and abbreviations of each gene with the Human Genome Organization (HUGO) to ensure consensus with gene names used by the scientific community. After this, the list of potential genes involved in SGNs was expanded using the web-available Search Tool for the Retrieval of Interacting Genes (STRING) software (version 10.0), which is a database that lists protein-protein interactions [42]. Only interactions with the highest degrees of confidence (above 0.900, range 0-0.99) were considered. Bioinformatics is an approach based on a systematic search to identify genes involved in the specific pathological process, which are then classified in order of importance according to the number of interactions determined experimentally. This rank was created based on a calculation of the interactions between genes, called the "weighted number of links" (WNL), which represents the gene interactions in a closed network. These interactions were identified using the STRING database, which allows scores to be obtained for each interaction. Maps are subsequently built among the identified genes to create networks [42]. Every gene identified was adjusted by multiplying to 1000 [17] [29] [11]. We also calculated the total interaction score (TIS), which comprises all of the gene interactions in the STRING database. A topological analysis was performed with Cytoscape, which is a software platform commonly used to visualize molecular interaction networks [5] [38]. The Ontological analysis, performed using the Biological Networks Gene Ontology (BiNGO) tool according to a previously described method [6] [26].

### 2.2. Clustering Procedures

Using the values of WNL and TIS for each gene related to the various types of SGN, the following clustering procedure was carried out using SPSS (version 18.0; IBM, New York, NY, USA). We firstly used K-means clustering, then validated the results with ANOVA and Tukey-Kramer post hoc test, with a p-value of <0.001 considered statistically significant [9] [17]. The genes with the highest rank of associations were identified as leader genes. The other clustering

procedure is based on the immune-inspired algorithm ClonALG [9], aided by the Kohonen neural network [23] training algorithm in order to increase its performance, as presented in [8]. The ClonALG approach explicitly takes into account the antibodies affinity where only the fittest are selected to proliferate through a process called maturation (or mutation). The affinity criterion is determined by the Euclidean distance.

As in almost all the evolution-based algorithms, this approach uses the concept of populations and relations between its individuals to proliferate the fittest. In this specific case, we have a population of antigens and a population of antibodies who will recognize the antigens. Both populations can be seen as matrices (Equation 1), with each line corresponding to an individual with dimension  $m$ .

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \in \mathbb{R}^{k \times m} \quad \text{and} \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \in \mathbb{R}^{n \times m}, \quad (1)$$

Where  $X$  is the antigen's population and  $B$  the antibodies' population;  $x_i \in \mathbb{R}^m$  and  $b_i \in \mathbb{R}^m$ . The general idea of an evolution-based algorithm is to repeat the mechanisms of cloning, mutation and selection until a stopping criterion is satisfied.

### 2.2.1. Stopping criteria

It was used two stopping criteria, the first one is a simple limit to the number of generations (an iteration composed by cloning, mutation and selection). The second one analyses the amount of created and deleted antibodies on each generation; if the amount of deleted antibodies is equal or bigger than the amount of created antibodies, the algorithm can be stopped.

### 2.2.2. Cloning

Each antibody is cloned twice and each one of these clones suffers a random mutation using a uniform distribution between -1 and 1 on its space position as shown in Equation 2, guaranteeing diversity in the antibody population.

$$B = \{B + \lambda \mathcal{U}_{n \times m}(-1, 1)\} \cup \{B + \lambda \mathcal{U}_{n \times m}(-1, 1)\} \cup B \quad (2)$$

where  $U_{n \times m}(-1, 1)$  represents a matrix with  $n \times m$  uniformly distributed random variables between -1 and 1 and  $\lambda$  is a constant that will be chosen in regard to the data, it could be  $\max(B)$  or even  $\text{mean}(B)$ ; the designer will choose whatever fits best.

### 2.2.3. Mutation

The mutation began at the cloning mechanism and will be improved by a Kohonen neural network where each antigen linearly reduces the distance between him ( $x_j$ ) and the closest antibody ( $\hat{b}_{x_j}$  given by Equation 3). This algorithm changes the antibody's position as if it was weights of a neuron as shown in Equation 4.

$$\hat{b}_{x_j} = \underset{b_i}{\operatorname{argmin}} \{ \|x_j - b_i\| : b_i \in B \} \quad (3)$$

$$\hat{b}_{x_j} = \hat{b}_{x_j} + \alpha_k (x_j - \hat{b}_{x_j}) \quad (4)$$

$$\alpha_k = \alpha_{k-1} \left( 1 - \frac{k}{K_m} \right) \quad (5)$$

Notice that Equations 3–5 iterate over  $k$  with  $0 < k \leq K_m$  for all  $x_j \in X$ . In order to make this linear reduction work, we must define  $\alpha_0$  and  $K_m$ ; in this paper, we defined  $\alpha_0 = 0.8$  and  $K_m = 2$ .

### 2.2.4. Selection

The selection uses two information to prune antibodies. The first one is the type of antigens mostly recognized by an antibody, given by Equation 6; the second one uses the proximity between two antibodies, where the threshold for considering an antibody close to another is defined in equation 7.

$$r(b_i) = \operatorname{Mo} \left( \{t(x_j) : x_j \in X \text{ and } \hat{b}_{x_j} = b_i\} \right) \quad (6)$$

where  $t(x_j)$  represents the class of antigen  $x_j$  and  $\operatorname{Mo}$  is the mode of the set.

$$c = 0.25 \left( \frac{2}{3n(3n-1)} \sum_{i=0}^{3n-1} \sum_{j=i+1}^{3n} \|b_i - b_j\| \right) \quad (7)$$

In other words, the threshold is 25% of the average distance between every two antibodies. We start by removing antibodies who did not recognize any class of antigen (Equation 8).

$$B = \{\hat{b}_{x_j} : x_j \in X\} \cap B \quad (8)$$

Then, all combinations of two elements out of the antibodies set are analyzed; when the set of antigens recognized by two of them are of the same type (Equation 6) and they are close with each other (Equation 7), a new antibody is created between them and those two must be deleted (Equation 9).

$$B_1 = \left\{ \frac{b_i + b_j}{2} : (b_i, b_j) \in \binom{B}{2} \text{ and } r(b_i) = r(b_j) \text{ and } \|b_i - b_j\| < c \right\}$$

$$B_2 = \left\{ b_i; b_j : (b_i, b_j) \in \binom{B}{2} \text{ and } r(b_i) = r(b_j) \text{ and } \|b_i - b_j\| < c \right\}$$

$$B = (B \setminus B_2) \cup B_1 \quad (9)$$

From Equation 9, it is possible to see that this method does not guarantee the final population of antibodies to be of size  $n$ .

### 2.3. Identification of targets

After clustering genes related to benign and malignant neoplasms, these genes were placed in a string to identify the central gene interactions. After expansion of the targets, genes were designated as described above in the bioinformatic and systems biology analysis section.

## 2.4. Validation

Ethical approval for this study was obtained from the Institutional Review Board. The inclusion criteria were the diagnose of SGN according to WHO classification of head and neck tumors [13]. Thirteen patients with benign and malign SGNs were included in the current study from surgery services from 2014 to 2016. The samples were divided into two groups according to the WHO classification and SGNs nature. Group 1 (Benign, n = 6) was comprised of patients with pleomorphic adenoma. For group 1 the exclusion criteria were the presence of malignant neoplasia in medical history. Group 2 (Malignant, n= 7) was comprised of patients with mucoepidermoid carcinoma (n = 4) and myoepithelial carcinoma (n = 3).

## 2.5. RNA isolation and real-time PCR

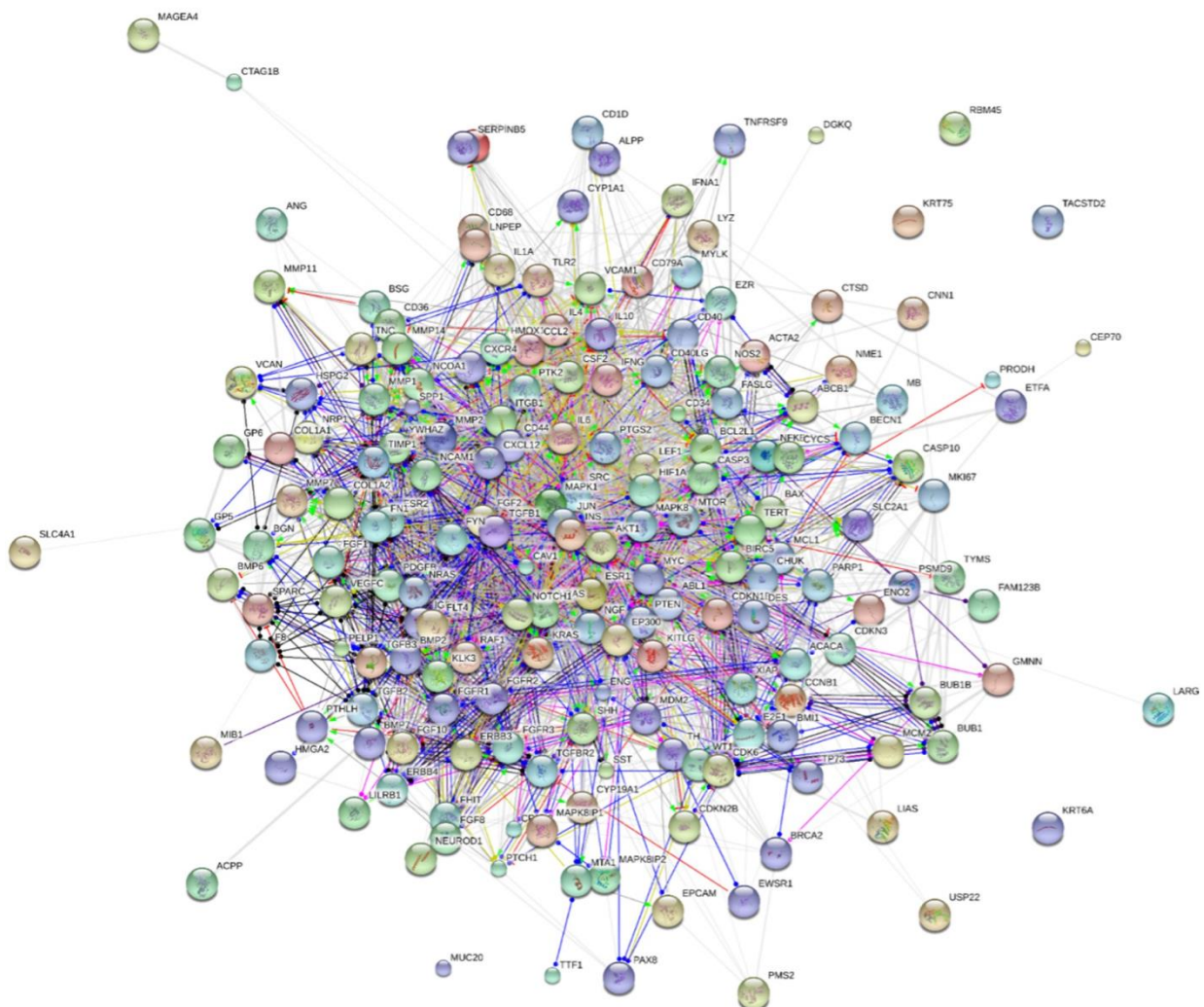
The mRNA from 13 samples was used for qRT-PCR. Isolation of RNA was performed using Trizol (Thermo Fisher Scientific, Waltham, MA, USA), according to a previously published protocol [17]. Total RNA was treated with amplification grade DNase I (catalog number 18068015, Invitrogen, Carlsbad, CA, USA), and 1.5 µg of RNA was reverse-transcribed using the SuperScript® First-Strand Synthesis System for qRT-PCR (catalog number 11904018, Invitrogen, Carlsbad, CA, USA). Each reaction for SYBR green-based qRT-PCR (total volume 20 µl) contained 10 µl of SYBR Green master mix, 0.25 µM of both forward and reverse primers, 1 µl of cDNA (66 ng/reaction) and 8.5 µl H<sub>2</sub>O. A non-template control (NTC) was included for each assay. The thermal cycling conditions were 95°C for 10 min, followed by 40 cycles of 95°C for 15 sec and 60°C for 1 min. The specific primers/probes (Life Technologies, Carlsbad, CA, USA) have been described previously [17]. Specifically, the primers for RPS15 were forward 5' - TGGAGGCTCGAATTATTGCCTTG - 3' and reverse 5' - TAGCGTTTGTGGGCTTTGTC - 3'. Beta-actin was used to normalize RPS15 expression and was amplified using the following primers: 5'- TGCCGACAGGATGCAGAAG-3' and 5'- CTCAGGAGGAGCAATGATCTTGA- 3'. Quantitative PCR was performed on a StepOne Real-Time PCR System (Life Technologies, Carlsbad, CA, USA). The C<sub>q</sub> values of endogenous controls were subtracted from the C<sub>t</sub> values of the respective targets to calculate the C<sub>q</sub>. The C<sub>q</sub> values from each experimental group were averaged and converted to log base 2 using the equation  $2^{-Cq}$  to compare expression among different samples.

### 3. Results

#### 3.1. Network characteristics

Our database was derived from bioinformatics analysis of 35 different histological classes of SGNs. There is extensive data about each type of neoplasia (Table in Supplementary Material). All networks presented significant representation based on the values for correlation and R2. The ontological analysis identified the common biological paths among specific groups of genes, and certain diseases showed mechanisms related to the three primary neoplasms of salivary glands. The network of malignant and benign neoplasms showed completely different patterns (Figure 1A and Figure 1B). Interestingly, the leader genes of malignant and benign lesions also showed different patterns of interaction (Figure 1C and Figure 1D).

A







D

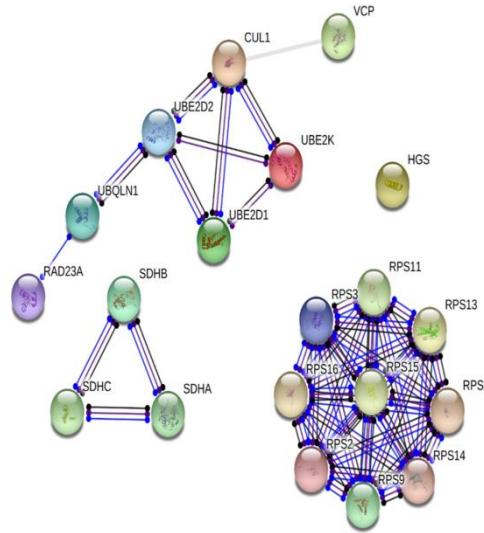


Figure 1: The network of malignant (A) and benign (B) salivary gland tumours, including all identified genes. Network with only the leader genes for malignant (C) and benign (D) tumours.

### 3.2. Targets

To identify genes related to benign SGNs, we combined K-means with ClonALG approaches. Genes exclusively related to malignant or benign SGNs were selected to obtain two networks for comparison. As suggested previously [28], in order to determine therapeutic targets, genes with a stronger relationship between WNL and TIS were considered as target genes. A network was built for malignant and benign neoplasms (Figure 2). The leader genes of each of the 35 SGNs are presented according to which of the two methods (K-means and ClonALG) they were obtained (Table in Supplementary Material). In 62.85% (22/35) of lesions, ClonALG identified fewer genes when compared to the K-means method. The genes with a higher WNL and a lower TIS, associated with malignant neoplasms, included INS, HRAS, SRC, and JUN (Table 1).

Genes leaders of genes associated exclusively with malignant tumors	Genes leaders of genes associated exclusively with benign tumors
INS HRAS, SRC, JUN	UBE2K, CUL1, HGS, VCP, UBE2D1, SDHB, UBQLN1, UBE2D2, RPS3, RAD23A

Table 1: Genes associated with malignancies.

On the other hand, benign neoplasms were related to UBE2K, CUL1, HGS, VCP, UBE2D1, SDHB, UBQLN1, UBE2D2, RPS3, and RAD23A. Two new networks were produced based on the STRING expansion. In this new expansion, RPS15 was the primary target for benign neoplasms, while PIK3CA was the target for malignant neoplasms (Figure 2).



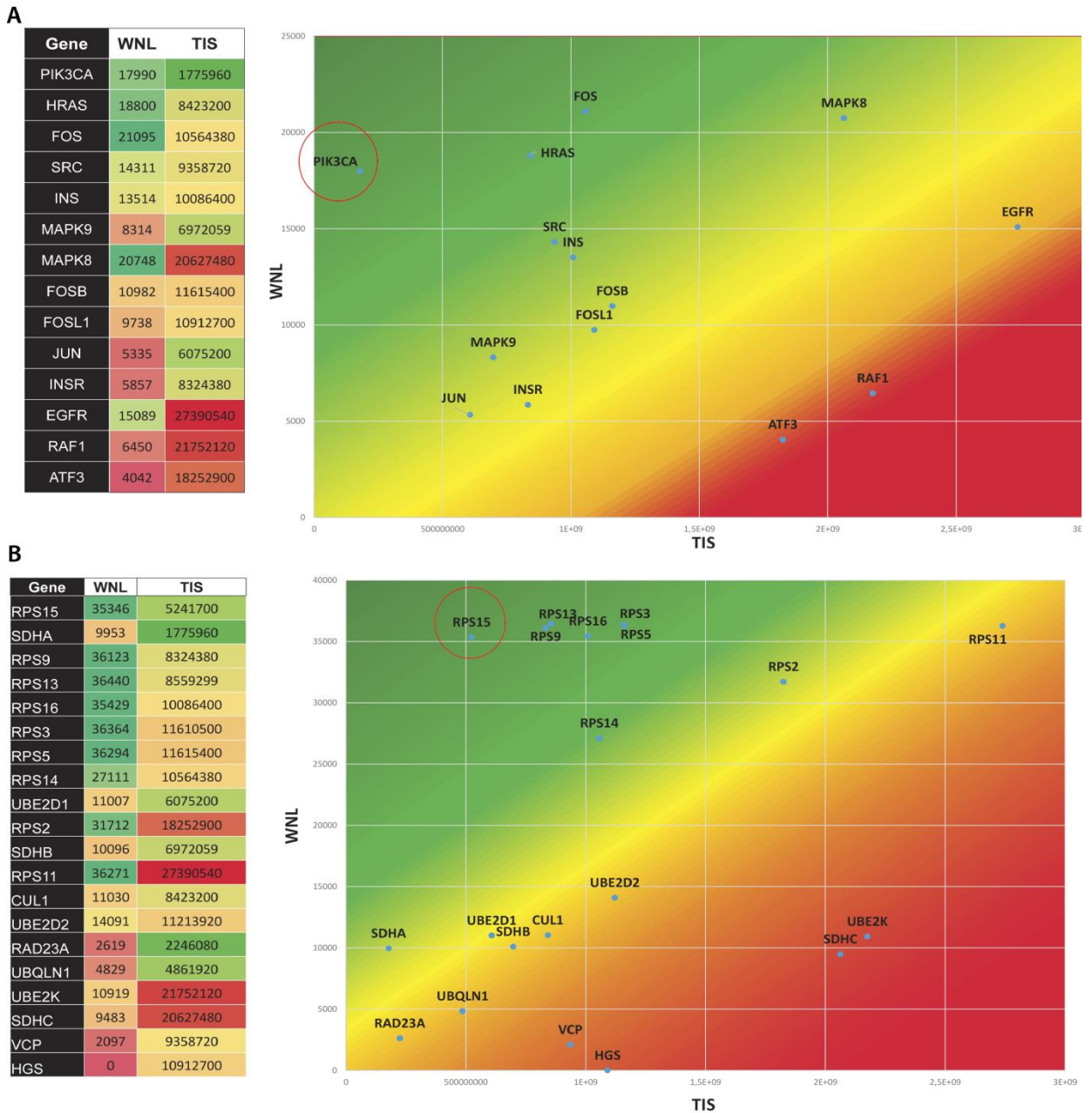


Figure 2: The relationship between a weighted number of links (WNL) and total interaction score (TIS) of leader genes associated with malignant (A) and benign (B) neoplasms.

Ontological analyses demonstrated that malignant SGNs were related to positive regulation of DNA replication (Figure 3A). On the other hand, benign SGNs were found to be related to translational elongation (Figure 3B). The strength of the benign ontology was higher than that for malignant neoplasms (Figure 3).

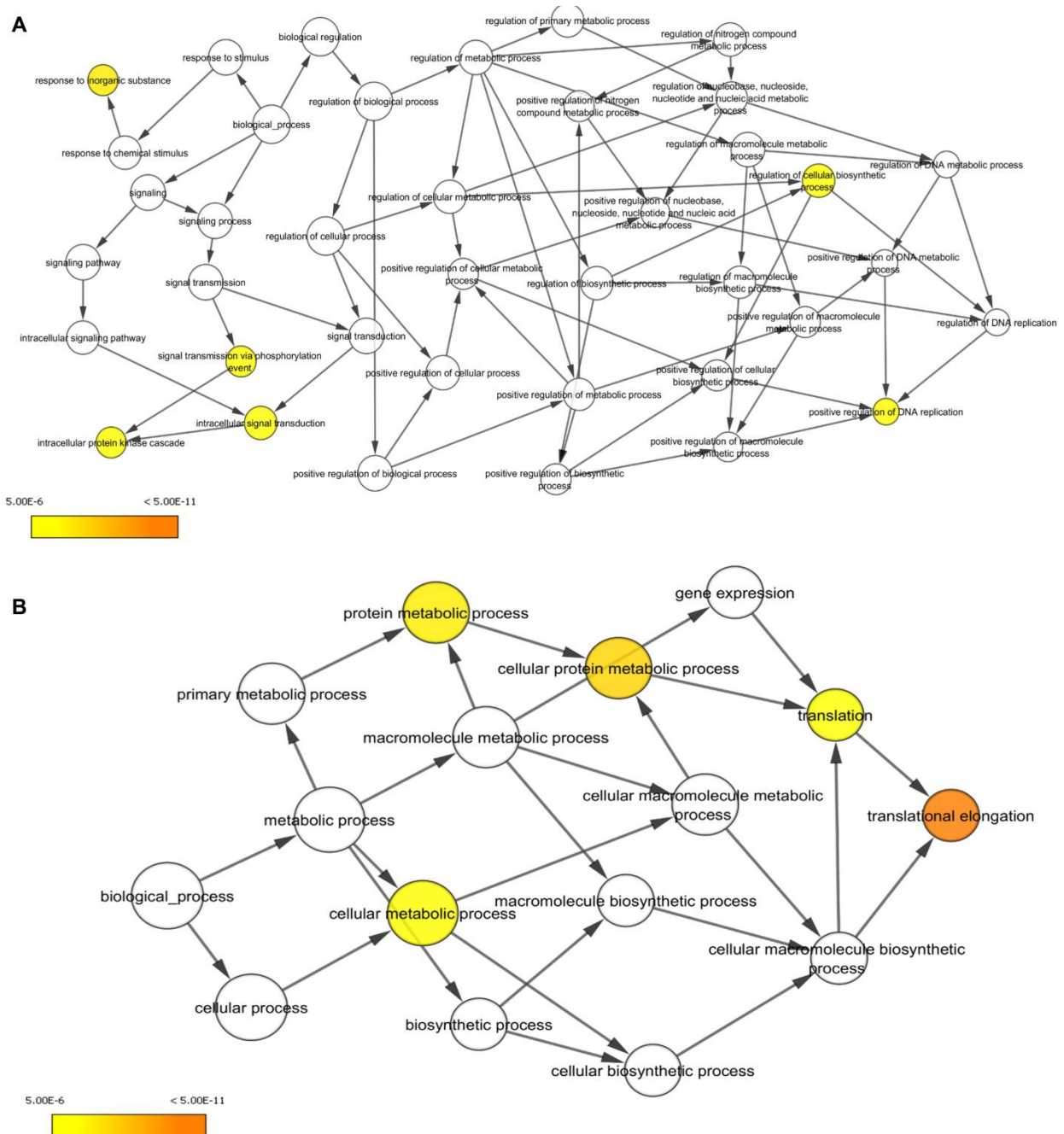


Figure 3: Ontological analyses for malignant (A) and benign (B) neoplasms, showing that the translational elongation process is more robust for differentiating between malignant and benign salivary gland neoplasms.

To validate the results of the theoretical study, mRNA of RSP15 was evaluated in malignant and benign SGNs. Malignant SGNs presented a reduction in RSP15 expression, confirming the results (Figure 4).

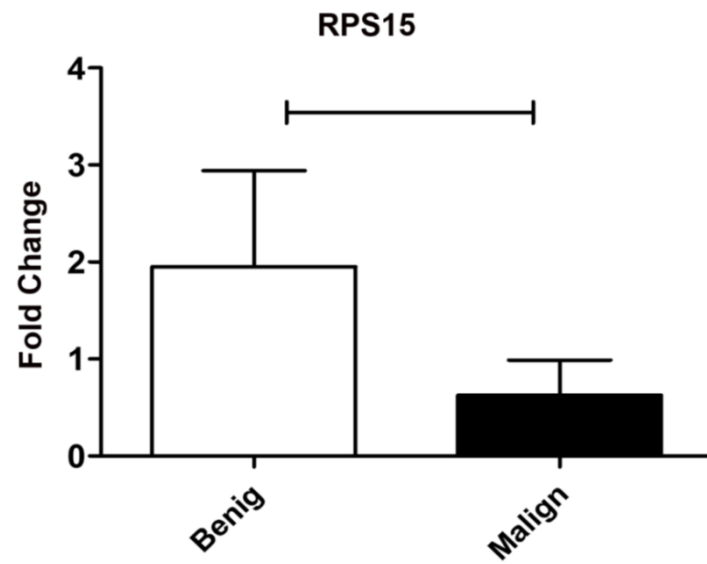


Figure 4: Gene expression analysis for RPS15. Levels of RPS15 mRNA differed depending on the nature of salivary gland neoplasms. Expression of RPS15 was reduced in malignant neoplasms.

#### 4. Discussion

There are many data in the literature about certain histological types of SGNs; however, some subtypes are very rare [19] [35]. As a consequence, it is difficult to obtain tissue samples for these SGNs, making it challenging to identify therapeutical targets. As an alternative, bioinformatics approaches may help to identify genes with a significant association with this neoplasia, which could represent prognostic markers or potential therapeutic targets [11]. The objective of the current study was to identify specific therapeutic targets using two different bioinformatics methods. Another critical aspect, which should be considered, is the fact that most neoplasms of the salivary glands are rare, and because of this, there were not enough tissue samples to proceed with experimental studies.

Classic malignant epithelial neoplasia biomarkers, such as epidermal growth factor receptor (EGFR), epidermal growth factor receptor 2 (HER2) and androgen receptor (AR), have been observed in malignant SGNs [4] [33] [39]. EGFR has been reported to be present in the majority of SGNs, mainly in malignant histotypes [33]. HER2 is usually expressed in neoplasia derived from the secretory duct [39]. C-kit is present mainly in neoplasms originating from the intercalated duct in both benign and malignant histologies [4]. The AR, as well as estrogen and progesterone receptors, are rarely present in SGNs [4] [16]. Another gene that is also cited in several studies is tumor protein 53 (TP53) [27]. The present study identified a significant number of genes related to each a neoplasia nature, which made it possible to make comparisons between groups according to whether they were benign or malignant neoplasia, their histological origin, and invasiveness, among other factors.

The K-means approach is a more traditional and well-studied procedure for clustering genes. It has some limitations which should be considered, for example, the definition of the number of clusters [2]. Because of this, the results should be carefully interpreted. There is evidence that clustering, precisely the number of clusters that are defined, is one of the most critical processes that need improvement [2] [32]. Because of this, our study used an alternative approach in addition to K-means.

The main limitation of the current study is the theoretical basis of the methodology, which is considered a disadvantage of this data mining technique. On the other hand, there is a significant cost benefit for this type of analysis [10] [11]. To obtain data for ClonALG, which is inspired by the biological immune system, all genes are known to be involved in SGNs were collected with

their respective values for WNL and TIS. This is an alternative method to find possible relationships and associations of the particular gene with different histological types.

Other interesting findings were obtained from our database. After reviewing the NCBI, Genecardsr and OMIM databases, followed by expansion using STRING software, we identified a total of 764 genes related to all 35 histological types of SGN. Of these, 24 are malignant, and 11 are benign. The distribution of these genes between malignant and benign neoplasms is remarkable. Some 92 genes are related only to benign lesions, while 189 were associated exclusively with malignant lesions. The remaining genes were linked to both types of neoplasms. This could be relevant information for future studies to identify genes and proteins that are more common in benign or malignant neoplasms. All of this data is available in the Supplementary Material.

Additionally, using the K-means method of clustering genes, 258 of the 764 genes were considered to be gene leaders. It is important to note that the distribution of gene leaders showed considerable variation, and also varied among the 35 types of SNGs. Therefore, each histological type should be analyzed with caution, looking for peculiarities. For example, of all of the leader genes of malignant neoplasms, four were found to be the most common. These include the genes CCND1, ERBB2, PGR, and TP63. On the other hand, five central leader genes were detected in benign neoplasms, which were KRT14, KRT19, KRT7, KRT8, and S100B.

Cyclin D1 (CCND1) is a regulatory constituent of the cyclin D1-CDK4 (DC) complex, which phosphorylates and inhibits members of the retinoblastoma (RB) protein family including RB1, as well as regulating the cell-cycle during the G(1)/S transition. Amplification, overexpression, and mutations in this gene usually cause alterations in cell cycle progression, and consequently, may be involved in tumorigenesis. CCND1 is present in seven neoplasms, including not otherwise specified adenocarcinomas, ductal papillomas, malignant sebaceous neoplasia, mammary analog secretory carcinomas, metastasizing pleomorphic adenomas, mucoepidermoid carcinomas, and myoepithelial carcinomas. Interestingly, the findings of this study showed that CCND1 had the highest WNL and lowest TIS in 16 of the 35 neoplasms (see Table in Supplementary Material).

Erb-B2 receptor tyrosine kinase 2 (ERBB2), which controls peripheral microtubule stabilization and outgrowth, was identified as a gene leader in seven different types of neoplasia. Interestingly, however, this gene showed the highest WNL and lowest TIS in 12 different SGNs. The progesterone receptor (PGR) is involved in the activation of c-SRC/MAPK signaling in

hormone stimulation. It was found to be a gene leader in six malignant neoplasms. It is fascinating that none of the 35 neoplasms showed PGR to have the highest WNL and lowest TIS. Tumour protein P63 (TP63) is responsible for initiating apoptosis in response to cellular insults and activated oncogenes. Seven malignant and two benign neoplasia histotypes showed the TP63 as a gene leader. TP63 had the highest WNL and lowest TIS in only three neoplasia (epithelial-myoepithelial carcinoma, mammary analog secretory carcinoma, and oncocytoma), but it was not a gene leader. The genes KRT14, KRT19, KRT7 and KRT8, all members of the keratin family, and S100B, a calcium-binding protein, were present most of the time in benign neoplasms.

Genes with higher WNL and relatively lower TIS are probably more acceptable as therapeutic targets, as genes with a higher TIS act in a wide variety of biological process [11] [32] (Table 1). It is widely accepted that the TIS is associated with a broad range of interactions in the STRING database. On the other hand, the WNL is associated with a distinct network interaction [11]. In conclusion, there is the difference in targets of malignant and benign SGNs. The ontological analyses demonstrated that malignant SGNs are related to positive regulation of DNA replication. On the other hand, benign SGNs are related to translational elongation. The investigation suggested by immune-based algorithm ClonALG and confirmed in patients' sample demonstrated that RPS15 levels were reduced in malignant SGNs.

## 5. Conclusions

In this paper an approach to characterize salivary gland neoplasms is presented. The proposed approach to characterize salivary gland neoplasms consists in two main modules: (i) a bioinformatics algorithm was then applied, and the STRING database was employed to build networks of protein-protein interactions for each nature of an SGNs and, (ii) a immune/neural approach which indicates leader genes. This method for data classification is based on the association of ClonALG immune systems with the Kohonen neural network. To show the advantages of this approach, experiments over K-means was carried out. The advantage of the proposed approach over K-means in controlled scenarios makes it a good candidate for real world testing.

## Acknowledgment

This study was supported by grants from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG). Dr. Guimarães, Dr. D'Angelo, Dr. Santos and Dr. de Paula are research fellows of the CNPq.



### Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Ethical approval for this study was obtained from the Institutional Review Board (52767316600005146).

### Informed consent

Informed consent was obtained from all individual participants included in the study.

## References

- [1] Dimitrios Andreadis, Apostolos Epivatianos, Athanasios Pouloupoulos, Alexandros Nomikos, Georgios Papazoglou, Demetrios Antoniadis, and Calypso Barbatis. Detection of c-kit (cd117) molecule in benign and malignant salivary gland tumours. *Oral Oncology*, 42(1):56 – 64, 2006.
- [2] Eric Bair. Semi-supervised clustering methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(5):349–361, 2013.
- [3] L. Barnes, J. W. Eveson, P. Reichart, and D. Sidransky. *Pathology and Genetics of Head and Neck Tumors*. Lyon: IARC Press, 2005.
- [4] Vishal S. Chandan, David Wilbur, William C. Faquin, and Kamal K. Khurana. Is c-kit (cd117) immunolocalization in cell block preparations useful in the differentiation of adenoid cystic carcinoma from pleomorphic adenoma? *Cancer Cytopathology*, 102(4):207–209, 2004.
- [5] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, Christopher Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A.R. Pico, A. Vailaya, P.L. Wang, A. Adler, B.R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G.J. Warner, T. Ideker, and G.D. Bader. Integration of biological networks and gene expression data using cytoscape. *Nature Protocols*, 2(10): 2366–2382, 2007.
- [6] Ugo Covani, Simone Marconcini, Luca Giacomelli, Victor Sivozhelevov, Antonio Barone, and Claudio Nicolini. Bioinformatic prediction of leader genes in human periodontitis. *Journal of Periodontology*, 79(10):1974–1983, 2008.
- [7] R. Dahse, O. Driemel, S. Schwarz, J. Dahse, K. Kromeyer-Hauschild, A. Berndt, and H. Kosmehl. Epidermal growth factor receptor kinase domain mutations are rare in salivary gland carcinomas. *British journal of cancer*, 100(4):623–625, 2009.
- [8] M. F. S. V. D’Angelo, R. M. Palhares, M. C. O. Camargos Filho, R. D. Maia, J. B. Mendes, and P. Ya. Ekel. A new fault classification approach applied to tennessee eastman benchmark process. *Applied Soft Computing*, 49:676–686, 2016.
- [9] L.N. de Castro and F.J.V. Zuben. Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation*, 6(3):239–251, 2002.
- [10] Fabiano de Oliveira Poswar, Lucyana C. Farias, Carlos Alberto de Carvalho Fraga, Wilson Bambirra, Manoel Brito-Jr, Manoel D. Sousa-Neto, S. Henrique Souza Santos, Alfredo M. B. de Paula, Marcos F. Silveira Vasconcelos D’Angelo, and A. L. S. Guimarães. Bioinformatics, interaction network analysis, and neural networks to characterize gene expression of radicular cyst and periapical granuloma. *Journal of Endodontics*, 41(6): 877 – 883, 2015.
- [11] Fabiano de Oliveira Poswar, L. Ives Santos, Lucyana C. Farias, Talita Antunes Guimarães, S. Henrique Souza Santos, Kimberly Marie Jones, Alfredo M. Batista de

- Paula, Reinaldo Martinez Palhares, Marcos F. Silveira Vasconcelos D'Angelo, and A. Luiz Sena Guimarães. An adaptation of particle swarm clustering applied in basal cell carcinoma, squamous cell carcinoma of the skin and actinic keratosis. *Meta Gene*, 12:72 – 77, 2017.
- [12] A. K. El-naggar. What is new in the world health organization 2017 histopathology classification? *Current Treatment Options in Oncology*, 18(7):1–4, 2017.
- [13] Adel K El-Naggar. What is new in the world health organization 2017 histopathology classification? *Current treatment options in oncology*, 18 (7):43, July 2017.
- [14] T Ettl, S Schwarz, N Kleinsasser, A Hartmann, T E Reichert, and O Driemel. Overexpression of egfr and absence of c-kit expression correlate with poor prognosis in salivary gland carcinomas. *Histopathology*, 53 (5):567–577, 2008.
- [15] Simon Fishilevich, Shahar Zimmerman, Asher Kohn, Tsippi Iny Stein, Tsviya Olender, Eugene Kolker, Marilyn Safran, and Doron Lancet. Genic insights from integrated human proteomics in genecards. Database, page baw030, 2016.
- [16] Afina S. Glas, Harry Hollema, Raoul E. Nap, and John T. Plukker. Expression of estrogen receptor, progesterone receptor, and insulin-like growth factor receptor-1 and of mib-1 in patients with recurrent pleomorphic adenoma of the parotid gland. *Cancer*, 94(8):2211–2216, 2002.
- [17] T. A. Guimaraes, L. C. Farias, C. A. Fraga, J. D. Feltenberger, G. A. Melo, R. D. Coletta, S. H. S. Santos, A. M. B. de Paula, and A. L. S. Guimaraes. Evaluation of the antineoplastic activity of gallic acid in oral squamous cell carcinoma under hypoxic conditions. *Anticancer Drugs*, 27(5):407–416, 2016.
- [18] T. A. Guimaraes, L. C. Farias, E. S. Santos, C. A. C. Fraga, L. A. Orsini, L. F. Teles, J. D. Feltenberger, S. F. de Jesus, M. G. de Souza, S. H. S. Santos, R. S. Gomez, A.M. B. de Paula, and A. L. S. Guimaraes. Metformin increases pdh and suppresses hif-1a under hypoxic conditions and induces cell death in oral squamous cell carcinoma. *Oncotarget*, 7(34):55057–55068, 2016.
- [19] Marco Guzzo, Laura D. Locati, Franz J. Prott, Gemma Gatta, Mark McGurk, and Lisa Licitra. Major and minor salivary gland tumors. *Critical Reviews in Oncology/Hematology*, 74(2):134 – 148, 2010.
- [20] V. A. Holst, C. E. Marshall, C. A. Moskaluk, and H. F. Jr. Frierson. Kit protein expression and analysis of c-kit gene mutation in adenoid cystic carcinoma. *Modern pathology: an official journal of the United States and Canadian Academy of Pathology*, 12:956–960, 1999.
- [21] P. L. Horn-Ross, B. M. Ljung, and M. Morrow. Environmental factors and the risk of salivary gland cancer. *Epidemiology*, 8:414 – 419, 1997.
- [22] Yung-Ming Jeng, Chiao-Ying Lin, and Hey-Chi Hsu. Expression of the c-kit protein is associated with certain subtypes of salivary gland carcinoma. *Cancer Letters*, 154(1):107 – 111, 2000.

- [23] T. Kohonen. Self-organizing maps. Springer, 2001. 408
- [24] L. D. Locati, P. Bossi, and S. E. A. Lo Vullo. Prognostic role of epidermal growth factor receptors 1 & 2, androgen receptor and kit in 111 salivary gland cancer. In 7th international conference on head & neck cancer, 2008.
- [25] Laura D. Locati, Federica Perrone, Marco Losa, Micol Mela, Paola Casieri, Marta Orsenigo, Barbara Cortelazzi, Tiziana Negri, Elena Tamborini, Pasquale Quattrone, Paolo Bossi, Gaetana Rinaldi, Cristiana Bergamini, Rosa G. Calderone, Cecilia Liberatoscioli, and Lisa Licitra. Treatment relevant target immunophenotyping of 139 salivary gland carcinomas (sgcs). *Oral Oncology*, 45(11):986 – 990, 2009.
- [26] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, 2005.
- [27] Toshitaka Nagao, Isamu Sugano, Yasuo Ishida, Akira Asoh, Shigeru Munakata, Kazuto Yamazaki, Akiyoshi Konno, Keiichi Iwaya, Tohru Shimizu, Hiromi Serizawa, and Yoshiro Ebihara. Hybrid carcinomas of the salivary glands: Report of nine cases with a clinicopathologic, immunohistochemical, and p53 gene alteration analysis. *Modern Pathology*, 15(7):724–733, July 2002.
- [28] Niyaz A Naykoo, Dil-Afroze, Roohi Rasool, Sonallah Shah, A.G Ahangar, Imtiyaz A Bhat, Iqbal Qasim, Mushtaq A Siddiqi, and Zafar A Shah. Single nucleotide polymorphisms, haplotype association and tumour expression of the vascular endothelial growth factor (vegf) gene with lung carcinoma. *Gene*, 608:95 – 102, 2017.
- [29] B. Orlando, N. Bragazzi, and C. Nicolini. Bioinformatics and systems biology analysis of genes network involved in olp (oral lichen planus) pathogenesis. *Archives of Oral Biology*, 58(6):664 – 673, 2013.
- [30] Michael F. Press, Malcolm C. Pike, Gene Hung, Jian Yuan Zhou, Yanling Ma, Jay George, Jeanne Dietz-Band, William James, Dennis J. Slamon, John G. Batsakis, and Adel K. El-Naggar. Amplification and overexpression of her-2/neu in carcinomas of the salivary gland: Correlation with poor prognosis. *Cancer Research*, 54(21):5675–5682, 1994.
- [31] Michael Rebhan, Vered Chalifa-Caspi, Jaime Prilusky, and Doron Lancet. Genecards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 13(4):163, 1997.
- [32] E. M. S. Santos, H. O. Santos, S. N. Fonseca, T. A. Guimarães, C. A. C. Fraga, K. M. Jones, A. M. B. de Paula, S. H. S. Santos, A. L. S. Guimarães, and L. C. Farias. Obesity-related genes and oral cancer: A bioinformatics approach and systematic review. *Journal of Applied Bioinformatics & Computational Biology*, 5(1), 2016.
- [33] Seta A. Sarkis, Bashar H. Abdullah, Ban A. Abdul Majeed, and Nazar G. Talabani. Immunohistochemical expression of epidermal growth factor receptor (egfr) in oral squamous cell carcinoma in relation to proliferation, apoptosis, angiogenesis and lymphangiogenesis. *Head & Neck Oncology*, 2 (1):13, Jun 2010.

- [34] Arthur B. Schneider, Jay Lubin, Elaine Ron, Cyril Abrahams, Marilyn Stovall, Ankush Goel, Eileen Shore-Freedman, and Theresa C. Gierlowski. Salivary gland tumors after childhood radiation treatment for benign conditions of the head and neck: Dose-response relationships. *Radiation Research*, 149(6):625 – 630, 1998.
- [35] Raja R. Seethala. Salivary gland tumors: Current concepts and controversies. *Surgical Pathology Clinics*, 10(1):155 – 176, 2017.
- [36] Raja R. Seethala and Christopher C. Griffith. Molecular pathology: Predictive, prognostic, and diagnostic markers in salivary gland tumors. *Surgical Pathology Clinics*, 9(3):339 – 352, 2016.
- [37] Raja R. Seethala and Göran Stenman. Update from the 4th edition of the world health organization classification of head and neck tumours: Tumors of the salivary gland. *Head and Neck Pathology*, 11(1):55–67, 2017.
- [38] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504, 2003. doi:10.1101/gr.1239303.
- [39] A Skalova, I Starek, T Vanecek, V Kucerova, L Plank, P Szepe, S Di Palma, and I Leivo. Expression of her-2/neu gene and protein in salivary duct carcinomas of parotid gland as revealed by fluorescence in-situ hybridization and immunohistochemistry. *Histopathology*, 42(4):348–356, 2003.
- [40] Kristine Bjorndal Sorensen, Christian Godballe, Karin de Stricker, and Annelise Krogdahl. Parotid carcinoma: expression of kit protein and epidermal growth factor receptor. *Journal of Oral Pathology & Medicine*, 35(5):286–291, 2006.
- [41] G.Marie Swanson and Patricia Brissette Burns. Cancers of the salivary gland: Workplace risks among women and men. *Annals of Epidemiology*, 7(6):369 – 374, 1997.
- [42] Christian von Mering, Lars J. Jensen, Berend Snel, Sean D. Hooper, Markus Krupp, Mathilde Foglierini, Nelly Jouffre, Martijn A. Huynen, and Peer Bork. String: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research*, 33 (suppl1):D433–D437, 2005.

## 6 CONSIDERAÇÕES FINAIS

O presente estudo apresentou uma nova metodologia para agrupamento/clusterização de dados em pesquisas genéticas e de biologia molecular e o compara com o método tradicional do K-means. Os resultados dos genes líderes encontrados em ambos os métodos podem ser utilizado em estudos experimentais posteriores para identificação dos genes que potencialmente poderiam ser utilizados como biomarcadores ou possíveis agentes de interesse em terapia de alvo molecular.

Na interpretação dos resultados deve ser considerado que a principal limitação deste tipo de estudo de mineração de dados usando a bioinformática é a base puramente teórica da sua metodologia, embora esta mesma característica trás grandes vantagens na relação custo benefício. Ou seja, ao mesmo tempo em que os métodos de bioinformática criam novas hipóteses e teorias, também permitem que as pesquisas experimentais que validam essas hipóteses sejam elaboradas de um modo mais adequado e custo-efetivo.

Os resultados desse estudo demonstram que o gene *RPS15*, codificador da proteína ribossomal S15, é um possível biomarcador ou marcador prognóstico. Portanto novas pesquisas são necessárias para o melhor conhecimento da associação desse gene com os processos biológicos responsáveis pelo desenvolvimento de neoplasias de glândulas salivares.

## REFERÊNCIAS

1. Laurie SA, Ho AL, Fury MG, Sherman E, Pfister DG. Systemic therapy in the management of metastatic or locally recurrent adenoid cystic carcinoma of the salivary glands: a systematic review. *The Lancet Oncology*. 2011;12(8):815-24.
2. Poswar FdO, Santos LI, Farias LC, Guimarães TA, Santos SHS, Jones KM, et al. An adaptation of particle swarm clustering applied in basal cell carcinoma, squamous cell carcinoma of the skin and actinic keratosis. *Meta Gene*. 2017;12:72-7.
3. D'Angelo MFSV, Palhares RM, Camargos Filho MCO, Maia RD, Mendes JB, Ekel PY. A new fault classification approach applied to Tennessee Eastman benchmark process. *Applied Soft Computing*. 2016;49:676-86.
4. Holmberg KV, Hoffman MP. Anatomy, biogenesis and regeneration of salivary glands. *Monographs in oral science*. 2014;24:1-13.
5. Guzzo M, Locati LD, Prott FJ, Gatta G, McGurk M, Licitra L. Major and minor salivary gland tumors. *Critical reviews in oncology/hematology*. 2010;74(2):134-48.
6. Ochal-Choińska AJ, Osuch-Wójcikiewicz E. Particular aspects in the cytogenetics and molecular biology of salivary gland tumours - current review of reports. *Contemporary oncology (Poznan, Poland)*. 2016;20(4):281-6.
7. El-Naggat AK. What Is New in the World Health Organization 2017 Histopathology Classification? Current treatment options in oncology. 2017;18(7):43.
8. Spiro RH. Salivary neoplasms: overview of a 35-year experience with 2,807 patients. *Head & neck surgery*. 1986;8(3):177-84.
9. Xiao CC, Zhan KY, White-Gilbertson SJ, Day TA. Predictors of Nodal Metastasis in Parotid Malignancies: A National Cancer Data Base Study of 22,653 Patients. *Otolaryngology--head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery*. 2016;154(1):121-30.
10. Dillon PM, Chakraborty S, Moskaluk CA, Joshi PJ, Thomas CY. Adenoid cystic carcinoma: A review of recent advances, molecular targets, and clinical trials. *Head & neck*. 2016;38(4):620-7.
11. Saku T, Hayashi Y, Takahara O, Matsuura H, Tokunaga M, Tokunaga M, et al. Salivary gland tumors among atomic bomb survivors, 1950-1987. *Cancer*. 1997;79(8):1465-75.
12. Horn-Ross PL, Ljung BM, Morrow M. Environmental factors and the risk of salivary gland cancer. *Epidemiology (Cambridge, Mass)*. 1997;8(4):414-9.
13. Swanson GM, Burns PB. Cancers of the salivary gland: workplace risks among women and men. *Annals of epidemiology*. 1997;7(6):369-74.
14. Pinkston JA, Cole P. Cigarette smoking and Warthin's tumor. *American journal of epidemiology*. 1996;144(2):183-7.
15. de Ru JA, Plantinga RF, Majoor MH, van Benthem PP, Slootweg PJ, Peeters PH, et al. Warthin's tumour and smoking. *B-ent*. 2005;1(2):63-6.
16. Serraino D, Boschini A, Carrieri P, Pradier C, Dorrucchi M, Dal Maso L, et al. Cancer risk among men with, or at risk of, HIV infection in southern Europe. *AIDS (London, England)*. 2000;14(5):553-9.
17. Shebl FM, Bhatia K, Engels EA. Salivary gland and nasopharyngeal cancers in individuals with acquired immunodeficiency syndrome in United States. *International journal of cancer*. 2010;126(10):2503-8.
18. Licitra L, Grandi C, Prott FJ, Schornagel JH, Bruzzi P, Molinari R. Major and minor salivary glands tumours. *Critical reviews in oncology/hematology*. 2003;45(2):215-25.
19. Lee YY, Wong KT, King AD, Ahuja AT. Imaging of salivary gland tumours. *European journal of radiology*. 2008;66(3):419-36.
20. Hisatomi M, Asaumi J, Yanagi Y, Unetsubo T, Maki Y, Murakami J, et al. Diagnostic value of dynamic contrast-enhanced MRI in the salivary gland tumors. *Oral oncology*. 2007;43(9):940-7.

21. Bialek EJ, Jakubowski W, Karpinska G. Role of ultrasonography in diagnosis and differentiation of pleomorphic adenomas: work in progress. *Archives of otolaryngology--head & neck surgery*. 2003;129(9):929-33.
22. Feld R, Nazarian LN, Needleman L, Lev-Toaff AS, Segal SR, Rao VM, et al. Clinical impact of sonographically guided biopsy of salivary gland masses and surrounding lymph nodes. *Ear, nose, & throat journal*. 1999;78(12):905, 8-12.
23. O'Brien CJ. Current management of benign parotid tumors--the role of limited superficial parotidectomy. *Head & neck*. 2003;25(11):946-52.
24. Guntinas-Lichius O, Klussmann JP, Schroeder U, Quante G, Jungehuelsing M, Stennert E. Primary parotid malignoma surgery in patients with normal preoperative facial nerve function: outcome and long-term postoperative facial nerve function. *The Laryngoscope*. 2004;114(5):949-56.
25. Wong SJ, Karrison T, Hayes DN, Kies MS, Cullen KJ, Tanvetyanon T, et al. Phase II trial of dasatinib for recurrent or metastatic c-KIT expressing adenoid cystic carcinoma and for nonadenoid cystic malignant salivary tumors. *Annals of oncology : official journal of the European Society for Medical Oncology*. 2016;27(2):318-23.
26. Laurie SA, Licitra L. Systemic therapy in the palliative management of advanced salivary gland cancers. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*. 2006;24(17):2673-8.
27. Locati LD, Bossi P, Perrone F, Potepan P, Crippa F, Mariani L, et al. Cetuximab in recurrent and/or metastatic salivary gland carcinomas: A phase II study. *Oral oncology*. 2009;45(7):574-8.
28. Drilon A, Laetsch TW, Kummar S, DuBois SG, Lassen UN, Demetri GD, et al. Efficacy of Larotrectinib in TRK Fusion-Positive Cancers in Adults and Children. *The New England journal of medicine*. 2018;378(8):731-9.
29. Haddad R, Colevas AD, Krane JF, Cooper D, Glisson B, Amrein PC, et al. Herceptin in patients with advanced or metastatic salivary gland carcinomas. A phase II study. *Oral oncology*. 2003;39(7):724-7.
30. Gibbons MD, Manne U, Carroll WR, Peters GE, Weiss HL, Grizzle WE. Molecular differences in mucoepidermoid carcinoma and adenoid cystic carcinoma of the major salivary glands. *The Laryngoscope*. 2001;111(8):1373-8.
31. Jeng YM, Lin CY, Hsu HC. Expression of the c-kit protein is associated with certain subtypes of salivary gland carcinoma. *Cancer letters*. 2000;154(1):107-11.
32. Kondo S, Mukudai Y, Soga D, Nishida T, Takigawa M, Shirota T. Differential expression of vascular endothelial growth factor in high- and low-metastasis cell lines of salivary gland adenoid cystic carcinoma. *Anticancer research*. 2014;34(2):671-7.
33. Yon Rhee S. *Bioinformatics. Current Limitations and Insights for the Future*2005. 569-70 p.
34. Moore JH. *Bioinformatics. Journal of Cellular Physiology*. 2007;213(2):365-9.
35. Jones NC, Pevzner PA, Pevzner P. *An introduction to bioinformatics algorithms: MIT press*; 2004.
36. Boguski MS. *Bioinformatics. Current opinion in genetics & development*. 1994;4(3):383-8.
37. Finishing the euchromatic sequence of the human genome. *Nature*. 2004;431(7011):931-45.
38. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics (Oxford, England)*. 2005;21(16):3448-9.
39. Hirschhorn JN. Genetic approaches to studying common diseases and complex traits. *Pediatric research*. 2005;57(5 Pt 2):74r-7r.
40. Su G, Morris JH, Demchak B, Bader GD. Biological network exploration with Cytoscape 3. *Current protocols in bioinformatics*. 2014;47:8.13.1-8..24.



41. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research*. 2015;43(Database issue):D447-52.
42. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. *Nucleic acids research*. 2009;37(Database issue):D412-6.
43. Ouzounis CA, Coulson RM, Enright AJ, Kunin V, Pereira-Leal JB. Classification schemes for protein structure and function. *Nature reviews Genetics*. 2003;4(7):508-19.
44. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*. 2005;122(6):957-68.
45. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, et al. A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease. *Molecular cell*. 2004;15(6):853-65.
46. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*. 2005;33(Database issue):D433-7.
47. Sivozhelezov V, Giacomelli L, Tripathi S, Nicolini C. Gene expression in the cell cycle of human T lymphocytes: I. Predicted gene and protein networks. *J Cell Biochem*. 2006;97(5):1137-50.
48. Santos EMS, Santos HO, Dos Santos Dias I, Santos SH, Batista de Paula AM, Feltenberger JD, et al. Bioinformatics Analysis Reveals Genes Involved in the Pathogenesis of Ameloblastoma and Keratocystic Odontogenic Tumor. *International journal of molecular and cellular medicine*. 2016;5(4):199-219.
49. Nicolini C. Nanogenomics for medicine. *Nanomedicine (London, England)*. 2006;1(2):147-52.
50. Nicolini C, Spera R, Stura E, Fiordoro S, Giacomelli L. Gene expression in the cell cycle of human T-lymphocytes: II. Experimental determination by DNASER technology. *J Cell Biochem*. 2006;97(5):1151-9.
51. Lipscomb CE. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*. 2000;88(3):265-6.
52. Shashikiran ND. MEDLINE, pubmed, and pubmed central ((R)): Analogous or dissimilar. *Journal of the Indian Society of Pedodontics and Preventive Dentistry*. 2016;34(3):197-8.
53. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current protocols in bioinformatics*. 2016;54:1.30.1-1..3.
54. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*. 2002;30(1):52-5.
55. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic acids research*. 2009;37(Database issue):D793-D6.
56. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). *Human genetics*. 2001;109(6):678-80.
57. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13(11):2498-504.
58. Battista GD, Eades P, Tamassia R, Tollis IG. *Graph drawing: algorithms for the visualization of graphs*: Prentice Hall PTR; 1998.
59. Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. *Nature biotechnology*. 2009;27(10):921-4.
60. Bair E. Semi-supervised clustering methods 2013. 349-61 p.
61. Wu FX. Genetic weighted k-means algorithm for clustering large-scale gene expression data. *BMC bioinformatics*. 2008;9 Suppl 6(Suppl 6):S12.

62. Tibshirani R, Walther G, Hastie T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic 2001. 411-23 p.
63. Timmis J, Hone A, Stibor T, Clark E. Theoretical advances in artificial immune systems. *Theoretical Computer Science*. 2008;403(1):11-32.
64. D'Angelo MF, Palhares RM, Camargos Filho MC, Maia RD, Mendes JB, Ekel PY. A new fault classification approach applied to Tennessee Eastman benchmark process. *Applied Soft Computing*. 2016;49:676-86.
65. Poswar Fde O, Farias LC, Fraga CA, Bambilra W, Jr., Brito-Junior M, Sousa-Neto MD, et al. Bioinformatics, interaction network analysis, and neural networks to characterize gene expression of radicular cyst and periapical granuloma. *Journal of endodontics*. 2015;41(6):877-83.
66. Bastard JP, Chambert S, Ceppia F, Coude M, Grapez E, Loric S, et al. [RNA isolation and purification methods]. *Annales de biologie clinique*. 2002;60(5):513-23.

## APÊNDICES

## APÊNDICE A – Genes líderes de cada subtipo histológico dos tumores de glândulas salivares

Classificação do tumor	Genes líderes com K means	Genes líderes com a abordagem imuno/neural
Adenocarcinoma de células basais	<i>ABCB1 BAX BIRC5 BMP2 CASP3 CASP8 CAV1 CD40LG CD44 CDKN1B CXCL12 ENO2 ERBB4 ESR2 FGFR3 FN1 HIF1A IFNG IL10 IL1B IL2 IL4 ITGB1 LEF1 MDM2 MLH1 MSH2 NFKBIA NOS2 PARP1 PGR PTCH1 PTGS2 RB1 SMAD4 TERT TP63 VCAM1</i>	<i>ABCB1 BAX BIRC5 CASP8 CAV1 CXCL12 ENO2 ERBB4 ESR2 HIF1A IL10 IL1B IL2 IL4 ITGB1 LEF1 MDM2 MLH1 MSH2 NFKBIA NOS2 PARP1 PGR PTCH1 PTGS2 RB1 SMAD4 TERT VCAM1</i>
Adenocarcinoma mucinoso	<i>AKT1 HRAS TP53</i>	<i>HRAS</i>
Adenocarcinoma, N.O.S	<i>ACACA ACTC1 AR BCL2 CCND1 CDH1 FGF2 KIT MET MYC PCNA PTEN STAT3 TGFB1 TNF VEGFA</i>	<i>ACACA ACTC1 AR BCL2 CCND1 CDH1 FGF2 KIT MET MYC PCNA PTEN STAT3 TGFB1 TNF VEGFA</i>
Adenocarcinoma polimórfico de baixo grau	<i>TP53</i>	<i>TP53</i>
Adenoma pleomórfico metastático	<i>BAP1 CCND1 CDH1 CDKN1A CDKN2A ERBB2 KRT14 MUC1 PBRM1 PGR PLAG1 RB1 SETD2 TP63 VIM</i>	<i>CCND1 CDH1 CDKN1A ERBB2 RB1</i>
Carcinoma Adenóide Cístico	<i>BCL2 CTNNB1 EGFR MTOR NOTCH1 PCNA PTEN</i>	<i>BCL2 CTNNB1 EGFR MTOR NOTCH1 PCNA PTEN</i>

Obs.: Os genes líderes em vermelho não foram identificados pela abordagem imuno/neural.

Continua

Classificação do tumor	Genes líderes com K means	Genes líderes com a abordagem imuno/neural
Carcinoma de células escamosas	<p><i>ATM BCL2L1 BIRC5 BMP2 BRCA1 CASP3 CCNB1 CD4 CDKN1B CXCL12 ERBB2 FGF10 FGF2 FGFR1 FGFR2 GSTP1 HLA-DRB1 ICAM1 IFNA1 IL1A IL4 MAPK8 MMP1 MMP14 MYC PGR PIK3CG PTHLH PTK2 RARB RASSF1 S100B SHH SPP1 SRC STAT1 VEGFA</i></p>	<p><i>ATM BCL2L1 BIRC5 BMP2 FGF10 FGF2 FGFR1 GSTP1 HLA-DRB1 IFNA1 MAPK8 MMP1 MMP14 MYC PGR PTHLH RARB SHH SRC</i></p>
Carcinoma de células acinares	<p><i>INS JUN SRC</i></p>	<p><i>INS JUN SRC</i></p>
Carcinoma de células claras, N.O.S	<p><i>ABCB1 BAX BIRC5 CASP3 CASP8 CAV1 CD36 CD4 CD40 CD40LG CD44 CD79A CDKN1B COL1A1 CTSD CXCL12 CYP19A1 ENO2 ERBB3 ESR2 FGFR3 FLT4 FN1 HIF1A IFNG IL10 IL1B IL2 IL4 MDM2 MLH1 MME MMP2 MSH2 NCAM1 NOS2 PGR PTCH1 PTGS2 PTK2 RAF1 SMAD4 SST TERT TP63 WT1 TLR2 VCAM1 WT1</i></p>	<p><i>BAX BIRC5 CASP3 CASP8 CAV1 CD40 CD44 CD79A CDKN1B COL1A1 CTSD CXCL12 CYP19A1 ERBB3 ERBB3 ESR2 FLT4 FN1 HIF1A IFNG IL10 IL1B IL2 MMP2 MSH2 NCAM1 NOS2 PGR PTCH1 PTGS2 PTK2 RAF1 SMAD4 SST TERT TP63 WT1</i></p>
Carcinoma de grande células	<p><i>ABCB1 BAX BRCA2 CASP8 CAV1 CCL2 CD40 CD40LG CD44 CD79A CDKN2B CDKN3 CSF2 CTSD CXCR4 CYCS CYP19A1 ENO2 ERBB4 ESR2 FG8 FN1 IL10 IL1B IL4 ITGB1 MCL1 MDM2 MLH1 MME MMP1 MSH2 NCAM1 NME1 NOS2 PARP1 PGR PTCH1 PTGS2 PTK2 TERT TGFB2 TP63 TYMS VCAM1</i></p>	<p><i>BAX CAV1 CCL2 CD40 CD44 CDKN2B CSF2 CTSD CXCR4 CYCS CYP19A1 ESR2 IL10 IL1B IL4 ITGB1 MCL1 MDM2 MMP1 NME1 PARP1 PGR PTGS2 PTK2 TERT TGFB2 TP63 TYMS VCAM1</i></p>

Obs.: Os genes líderes em vermelho não foram identificados pela abordagem imuno/neural.

Continua

Classificação do tumor	Genes líderes com K means	Genes líderes com a abordagem imuno/neural
Carcinoma de pequenas células	<p><i>ABCB1 BAX BRCA2 CASP8 CAV1 CCL2 CD34 CD36 CD40 CD40LG CD44 CD79A CDKN2B CDKN3 CHUK CSF2 CTSD CXCR4 CYCS CYP19A1 ENO2 ERBB4 ESR2 FLT4 FN1 HSPG2 IL10 IL1B IL4 IL1B IL4 ITGB1 MDM2 MLH1 MME MMP1 MSH2 NCAM1 NME1 NOS2 PARG PGR PTCH1 PTGS2 PTK2 TERT TGFB2 TP63 VCAM1</i></p>	<p><i>ABCB1 BAX BRCA2 CASP8 CAV1 CCL2 CD36 CD40 CD40LG CD44 CD79A CDKN2B CDKN3 CHUK CSF2 CTSD CXCR4 CYP19A1 ENO2 ERBB4 ESR2 FLT4 FN1 HSPG2 IL10 IL1B IL4 ITGB1 MDM2 MLH1 MME MMP1 NCAM1 NME1 NOS2 PGR PTCH1 PTGS2 TERT TGFB2 TP63 VCAM1</i></p>
Carcinoma ex adenoma pleomórfico	<p><i>ABCB1 BECN1 BMI1 BMP6 BMP7 CD34 CD44 COL1A1 COL1A2 CYP1A1 E2F1 ENG ENO2 EWSR1 EZR FASLG FGF1 HBEGF HMOX1 IGF2 KITLG LYZ MCM2 MME NCOA1 NGF PDGFB PLAG1 PSMD9 PTGS2 S100B SLC2A1 SPARC SPP1 TERT TGFB2 TGFB3 TH TIMP1 TNC TP63 TP73 VCAN VIM XIAP</i></p>	<p><i>BECN1 BMP7 CD34 CD44 COL1A1 COL1A2 CYP1A1 E2F1 ENG EWSR1 EZR FGF1 HBEGF HMOX1 IGF2 KITLG MCM2 MME NCOA1 NGF PTGS2 SLC2A1 SPP1 TERT TGFB3 TH TIMP1 TNC TP73 VCAN VIM</i></p>
Carcinoma mioepitelial	<p><i>ABL1 ACTA1 ACTA2 ACTC1 CCND1 CDK2 CDK4 CDK6 CTNNB1 EP300 ERBB2 ESR1 FYN HDAC1 EGFR EP300 ERBB2 ESR1 FYN GSK3B HDAC1 KIT KRAS MTOR STAT3 YWHAZ</i></p>	<p><i>ACTA2 CDK2 CDK4 CDK6 CTNNB1 EP300 ERBB2 ESR1 FYN HDAC1 MTOR YWHAZ</i></p>
Carcinoma mucoepidermóide	<p><i>BUB1 CCND1 CDH1 CDKN2A ERBB2 FGF2 FGFR1 GATA3 KIT STAT3 TGFB1 VEGFA</i></p>	<p><i>CCND1 ERBB2 FGF2 FGFR1 GATA3 KIT STAT3 TGFB1</i></p>
Carcinoma oncocítico	<i>UBC</i>	-

Continua

Classificação do tumor	Genes líderes com K means	Genes líderes com a abordagem imuno/neural
Carcinoma sebáceo maligno	<i>AR CCND1 CDH1 CDKN1A CDKN2A</i> <i>CTNNB1 EGFR ERBB2 ESR1 FGFR2</i> <i>IL6 KIT KRAS PIK3CA PPARG PTEN</i> <i>TNF</i>	<i>CDKN1A CDKN2A CTNNB1 EGFR</i> <i>ERBB2 IL6 KIT PIK3CA PTEN TNF</i>
Carcinoma secretório-mamário análogo	<i>BCL2 CCND1 CDH1 EGFR ERBB2</i> <i>ESR1 PTEN</i>	<i>BCL2 ESR1 PTEN</i>
Carcinosarcoma	<i>ACPP ALPP ANG CD1D CD68</i> <i>CEACAM3 CEACAM5 CEP70 CHGA</i> <i>CLDN3 CLDN4 CTAG1B DES DGKQ</i> <i>E2F1 EPCAM F8 FAM123B GP5</i> <i>GP6 HMGA2 KRT13 KRT14 KRT18</i> <i>KRT19 KRT20 KRT7 KRT75 KRT8</i> <i>LARGE LIAS LNPEP MAGEA4 MAPK8IP1</i> <i>MAPK8IP1 MAPK8IP2 MB MKI67</i> <i>MMP7 MUC1 MUC16 MUC20 MYLK</i> <i>NEUROD1 NRP1 PAX8 PMS2 PTHLH</i> <i>S100A1 S100B SERPINA1 SLC4A1</i> <i>SPARC SYP TACSTD2 TNFRSF9 TTF1</i> <i>VEGFC</i>	<i>ACPP ALPP ANG CD1D CD68</i> <i>CEACAM3 CEACAM5 CEP70 CHGA</i> <i>CLDN3 CLDN4 CTAG1B DES DGKQ</i> <i>EPCAM F8 FAM123B GP5 GP6</i> <i>KRT14 KRT19 KRT20 KRT75 LARGE</i> <i>LIAS LNPEP MAGEA4 MAPK8IP1</i> <i>MAPK8IP2 MMP7 MUC1 MUC16</i> <i>MUC20 NEUROD1 NRP1 PAX8 PMS2</i> <i>PTHLH S100A1 SERPINA1 SPARC</i> <i>TACSTD2 TNFRSF9 TTF1</i>
Cistoadenocarcinoma	<i>UBC</i>	<i>UBC</i>
Sialoblastoma	<i>ACTA2 BCL2 CTNNB1 EGFR ESR1</i> <i>KRAS MAPK1 MYC NRAS PCNA PTEN</i>	<i>ACTA2 BCL2 EGFR ESR1 KRAS</i> <i>MAPK1 MYC PTEN</i>
Adenoma canalicular	<i>ABCC1 ABCG2 AHR GDNF KRT13</i> <i>KRT19 KRT5 KRT7 KRT72 KRT8 MME</i> <i>NFE2L2 RECK S100B SLC4A1 SLC4A3</i> <i>SOX10 SPN TP63 VIM</i>	<i>ABCC1 ABCG2 KRT19 KRT5 KRT8</i> <i>MME NFE2L2 RECK S100B SLC4A3</i> <i>SOX10 SPN TP63 VIM</i>

Continua

Classificação do tumor	Genes líderes com K means	Genes líderes com a abordagem imuno/neural
Adenoma de células basais	<i>BRAF CALCA CASP8 CD9 CHGA CYLD EGF FAS FASLG FGF10 GDNF GFAP GH1 HGF IL2RA KRT13 KRT14 KRT18 KRT19 KRT20 KRT5 KRT7 KRT72 KRT8 MGMT MKI67 MUC1 NES NKX2-1 PRL S100B S100P SLC4A1 SLC4A3 SYP TG TGFA TP63 VIM</i>	<i>BRAF CALCA CASP8 CD9 CHGA CYLD EGF FAS FASLG FGF10 GDNF GFAP HGF IL2RA KRT13 KRT14 KRT18 KRT19 KRT20 KRT5 KRT7 KRT72 KRT8 MGMT MKI67 MUC1 NES NKX2-1 PRL S100B S100P SLC4A1 SLC4A3 SYP TG TGFA TP63 VIM</i>
Adenoma pleomórfico	<i>ACTA1 BCL2 CDK2 CDK4 CTNNB1 EGFR GSK3B PCNA</i>	<i>BCL2 CDK2 CDK4 CTNNB1 EGFR GSK3B</i>
Adenoma sebáceo	<i>KRT14 MLH1 MSH2</i>	<i>MLH1</i>
Cistoadenoma	<i>ACPP AFP ALCAM AZGP1 BRAF CA9 CD2 CEACAM3 CEACAM5 CHGA CLDN3 CLDN4 COL6A3 CRTC1 CRTC3 CYP21A2 GAST HBEGF HGF KRT14 KRT15 KRT18 KRT19 KRT20 KRT7 KRT8 MAML2 MFGE8 MKI67 MT3 MUC1 MUC2 MUC4 MUC5AC MUC5B MUC6 NES RASSF1 REG4 S100A1 SERPINA3 SMUG1 SYP TGFA THY1 TIMP2</i>	<i>ACPP AFP ALCAM AZGP1 BRAF CA9 CD2 CEACAM3 CEACAM5 CHGA CLDN3 CLDN4 CRTC1 CRTC3 CYP21A2 GAST HGF KRT14 KRT15 KRT19 KRT20 KRT7 MAML2 MFGE8MT3 MUC1 MUC2 MUC4 MUC5ACMUC5B MUC6 RASSF1 S100A1 SERPINA3 SMUG1 SYP THY1 TIMP2</i>
Linfadenoma	<i>UBC</i>	<i>UBC</i>
Mioepitelioma	<i>UBC</i>	<i>UBC</i>

Obs.: Os genes líderes em vermelho não foram identificados pela abordagem imuno/neural.

Continua

Classificação do tumor	Genes líderes com K means	Genes líderes com a abordagem imuno/neural
Oncocitoma	<i>CTNNB1</i>	<i>CTNNB1</i>
Papiloma ductal	<i>ACTC1 AKT2 AKT3 AR ATM BRCA1</i> <i>CCND1 CDH1 CDKN1A CDKN2A</i> <i>ERBB2 GATA3 HSPD1 KDR KIT MET</i> <i>PIK3CA PIK3CG PPARG PRKG1 RB1</i> <i>RET STAT1 STAT3 STK11 TNF VEGFA</i>	<i>AKT3 AR ATM BRCA1 CCND1 CDH1</i> <i>CDKN1A ERBB2 HSPD1 KDR KIT MET</i> <i>PIK3CA PIK3CG PPARG PRKG1 RB1</i> <i>RET STAT1 STAT3 VEGFA</i>
Warthin tumour	<i>BCL2 CDH1 EGFR ICAM1 PCNA RET</i> <i>SDHB SMAD4 SST</i>	<i>BCL2 EGFR PCNA RET SDHB</i>

---

Conclusão



## APÊNDICE B – Genes com maiores WNL e menores TIS


Classificação histológica	Genes maiores WNL e menores TIS
Carcinoma de grandes células	<i>BCL2 CCND1 CTNNB1 MAPK8 MYC PIK3CA STAT3</i>
Cistoadenocarcinoma cribiforme de baixo grau	<i>ARID1A KRT14 KRT5</i>
Linfoadenomas	<i>EGFR GRB2 RPS27A</i>
Tumor maligno sebáceo	<i>AR CCND1 ERBB2 ESR1 PIK3CA</i>
Carcinoma secretário mamário análogo	<i>EGF KRT18 KRT8 MUC1 S100B TP63</i>
Adenoma pleomórfico metastático	<i>CCND1 CDH1 CDKN1A CDKN2A ERBB2 RB1</i>
Adenocarcinoma mucinoso	<i>CTNNB1 EGFR GRB2 PIK3R1</i>
Carcinoma mucoepidermóide	<i>CCND1 CDH1 CTNNB1 ERBB2 FGF2 STAT3 VEGFA</i>
Carcinoma mioepitelial	<i>E2F1 EGFR GRB2 RB1</i>
Mioepitelioma	<i>EGFR PCNA RPS27A</i>
Carcinoma oncocítico	<i>CTNNB1 EGFR GRB2 PIK3R1</i>
Oncocitoma	<i>KRT18 KRT8 TP63</i>
Adenoma pleomórfico	<i>MCM3 MCM4 MCM5 MCM7 PCNA</i>
Adenocarcinoma polimórfico de baixo grau	<i>CCND1 CDH1 ERBB2</i>
Carcinoma ductal salivar	<i>AR CCND1 CDH1 ERBB2 MMP9 PIK3CA</i> <i>SMAD4 VEGFA</i>

## APÊNDICE C – Genes exclusivamente associados a tumores malignos ou benignos

Genes exclusivamente associados a tumores de glândulas salivares malignos	Genes exclusivamente associados a tumores de glândulas salivares benignos
<p> <i>ABCB1 ABL1 ACACA ACPP ACTA2 AKT1 ALPP ANG BAX BCL2L1 BECN1 BGN            BIRC5 BMI1 BMP2 BMP6 BMP7 BRCA2 BSG BUB1 BUB1B CASP10 CASP3            CAV1 CCL2 CCNB1 CD1D CD34 CD36 CD4 CD40 CD40LG CD44 CD68 CD79A            CDK6 CDKN1B CDKN2B CDKN3 CEP70 CHUK CNN1 COL1A1 COL1A2 CREB3            CSF2 CTAG1B CTSD CXCL12 CXCR4 CYCS CYP19A1 CYP1A1 DES DGKQ E2F1            NAH ENG ENO2 EP300 EPCAM ERBB3 ERBB4 ESR1 ESR2 ETFA EWSR1 EZR F8            FAM123B FASLG FGF1 FGF10 FGF2 FGF8 FGFR1 FGFR2 FGFR3 FHIT FLT4 FN1            FYN GMNN GP5 GP6 GSTP1 HDAC1 HIF1A HLA-DRB1 HMGA2 HMOX1 HRAS            HSPG2 IFNA1 IFNG IGF2 IL10 IL1A IL1BIL2 IL4 IL6 INS ITGB1 JUN KITLG KLK3            KRAS KRT6A KRT75 LARGE LEF1 LIAS LNPEP LYZ MAGEA4 MAPK1 MAPK8            MAPK8IP1 MAPK8IP2 MB MCL1 MCM2 MDM2 MIB1 MKI67 MMP1 MMP11            MMP14 MMP2 MMP7 MTA1 MTOR MUC20 MYC MYLK NCAM1 NCOA1            NEUROD1 NFKBIA NGF NME1 NOS2 NOTCH1 NRAS NRP1 PARP1 PAX8            PDGFB PELP1 PMS2 PRODH PSMD9 PTCH1 PTEN PTGS2 PTHLH PTK2 RAF1            RARBSERPINA1 SERPINB5 SHH SLC2A1 SLC4A1 SPARC SPP1 SRC SST TACSTD2            TERT TGFB1 TGFB2 TGFB3 TGFB2 TH TIMP1 TLR2 TNC TNFRSF9 TP73 TTF1            TYMS USP22 VCAM1 VCAN VEGFC WT1 XIAP YWHAZ</i> </p>	<p> <i>ABCC1 ABCG2 ACPP AFP AGER AHR AKT2 AKT3 ALCAM AZGP1 BAP1 CA9            CALCA CD2 CD9 CDC34 COL6A3 CRTC1 CRTC3 CUL1 CYLD CYP21A2 FAS            FASLG FGF10 GAST GDNF GFAP GH1 HGF HGS HLA-DQB1 HMGCR            HSPD1 IL2RA KDR KRT15 KRT3 KRT72 MAGOH MAML2 MFGE8 MGMT            MKI67 MT3 MUC4 MUC5AC NES NFE2L2 NKX2-1 PBRM1 POLH PRKG1            PRL PSMC2 PSMD14 PSMD4 RABGEF1 RAD23A RAD23B RECK REG4 RET            RPS3 SDHB SERPINA3 SETD2 SLC4A1 SLC4A3 SMUG1 SOX10 SPN SST            STK11 TG THY1 TIMP2 TSG101 UBA1 UBE2D1 UBE2D2 UBE2D3 UBE2K            UBE2N UBQLN1 UCHL3 USP10 USP14 USP32 VCP VIM WIBG</i> </p>

## ANEXOS

ANEXO A – Aprovação da pesquisa pelo comitê de ética local da Universidade Estadual de Montes Claros / Brasil (número de protocolo: CAAE: 52767316.6.0000.5146).

<b>- DADOS DO PROJETO DE PESQUISA</b>	
<p>Título Público: Expressão de HIF1<math>\alpha</math> e miR-210 e proteômica em neoplasias de glândulas salivares e suas associações com critérios prognósticos. Pesquisador Responsável: André Luiz Sena Guimarães Contato Público: André Luiz Sena Guimarães Condições de saúde ou problemas estudados: Descritores CID - Gerais: Descritores CID - Específicos: Descritores CID - da Intervenção: Data de Aprovação Ética do CEP/CONEP: 20/02/2016</p>	
<b>- DADOS DA INSTITUIÇÃO PROPONENTE</b>	
<p>Nome da Instituição: Universidade Estadual de Montes Claros - UNIMONTES Cidade: MONTES CLAROS</p>	
<b>- DADOS DO COMITÊ DE ÉTICA EM PESQUISA</b>	
<p>Comitê de Ética Responsável: 5146 - Universidade Estadual de Montes Claros - UNIMONTES Endereço: Av. Dr Rui Braga s/n-Camp Univers Profº Darcy Rib Telefone: (38)3229-8180 E-mail: smelocosta@gmail.com</p>	