Joana Margarida Morgado Menoita

Perfil Genómico do Carcinoma da Cabeça e Pescoço: Existem Preditores para as Diferentes Taxas de Sobrevivência?

Dissertação de Mestrado em Engenharia Biomédica apresentada à Universidade de Coimbra

Setembro 2016

· U C ·

Universidade de Coimbra

• U C •
FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Joana Margarida Morgado Menoita

Perfil Genómico do Carcinoma da Cabeça e Pescoço: Existem Preditores para as Diferentes Taxas de Sobrevivência?

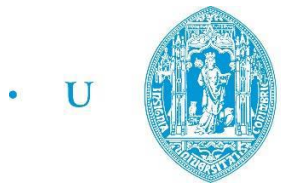
Dissertação apresentada à Universidade de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Biomédica

Orientador(es):
Professora Doutora Maria Joana Barbosa de Melo (Professora Auxiliar da Faculdade de Medicina da Universidade de Coimbra)
Professor Doutor Francisco José Santiago Fernandes Amado Caramelo (Professor Auxiliar da Faculdade de Medicina da Universidade de Coimbra)

Coimbra, 2016

Este trabalho foi desenvolvido em colaboração com:

Departamento de Física da Faculdade de Ciências e Tecnologia da Universidade de Coimbra



FCTUC FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE DE COIMBRA

Laboratório de Citogenética e Genómica da Faculdade de Medicina da Universidade de Coimbra

FMUC FACULDADE DE MEDICINA UNIVERSIDADE DE COIMBRA

Laboratório de Bioestatística e Informática Médica da Faculdade de Medicina da Universidade de Coimbra

FMUC FACULDADE DE MEDICINA UNIVERSIDADE DE COIMBRA

AGRADECIMENTOS

À Professora Doutora Maria Joana Barbosa de Melo pela orientação, disponibilidade e todo o otimismo incutido na concretização desta dissertação.

Ao Professor Doutor Francisco José Santiago Fernandes Amado Caramelo por todo o tempo disponibilizado, toda a ajuda incansável, todas as dúvidas esclarecidas e acima de tudo pela paciência.

À Ilda Ribeiro por toda a confiança depositada em mim, ter sempre acreditado que conseguiria, acudir-me em todas as aflições e problemas, ter sempre uma solução para tudo e, acima de tudo, pela amizade.

Aos meus "amigos bioestatísticos", André Santos e João Roque, porque sem eles este projeto não seria o mesmo e por todas as vezes que mandei *e-mails* e mensagens a pedir ajuda.

A toda a equipa do Laboratório de Citogenética e Genómica da Faculdade de Medicina da Universidade de Coimbra dirigido pela Professora Doutora Isabel Marques Carreira por me ter recebido tão bem.

Às "estagiárias" Camila Oliveira, Luísa Esteves, Mariana Tomás e Nicole Pedro por todo o companheirismo, compreensão, ajuda e amizade nesta nossa luta. E à Mariana Val e Joana Rodrigues por me ouvirem e apoiarem.

Aos meus amigos da Coimbra, Guarda e Vila Real que são uma parte crucial na minha vida e contribuíram de forma imprescindível para conseguir atingir os meus objetivos. E aos meus colegas de curso que me acompanharam nestes anos universitários.

E por fim e o mais importante, à minha família OBRIGADA. Aos meus pais, Manuel e Joaquina, por nunca duvidarem de mim e por todo o carinho, esforço, ajuda e paciência para comigo. Às minhas irmãs, Sara e Rita, que foram, são e serão o meu pilar. Ao meu avô José. E à restante família, tios e primos.

“Everything is theoretically impossible, until it is done”

Robert A. Heinlein

RESUMO

O Carcinoma Epidermoide da Cabeça e Pescoço (CECP) é uma neoplasia que surge nas regiões do trato aerodigestivo superior, incluindo a cavidade oral, orofaringe, laringe, hipofaringe e língua e é o sexto tumor mais comum em todo o mundo. O desenvolvimento deste carcinoma resulta da acumulação gradual de alterações e instabilidade genética e epigenética que levam à carcinogénese e progressão tumoral. Têm sido efetuados alguns estudos com vista a identificar alterações genómicas possíveis de serem utilizadas na classificação, diagnóstico e prognóstico, incluindo a previsão da eficácia do tratamento. Apesar dos avanços efetuados nas tecnologias de diagnóstico e modalidades de tratamento, estes tumores continuam a ser diagnosticados em estádios avançados sem significativas melhorias nas taxas de sobrevivência. O objetivo deste trabalho é testar a seguinte hipótese: será que as diferentes taxas de sobrevivência dos doentes de CECP estão associadas à presença de tumores com diferentes assinaturas genómicas? Recorrendo a resultados provenientes da técnica de Hibridização Genómica Comparativa em array (aCGH) referentes à variação do número de cópias (CNV) de material genético, conjuntamente com a informação clínico-patológica dos doentes com diagnóstico de CECP, foram analisados 104 doentes com recurso às plataformas de *software* Matlab R2014a, R3.1.2 e SPSS. Desta forma, como resultados deste projeto foi verificada a associação das variáveis dependentes estádio do tumor, localização do tumor e presença de metástases relativamente aos dados de aCGH e foram identificadas as regiões mínimas comuns alteradas mais frequentes nos cromossomas. Foi também possível avaliar a eficácia dos dois classificadores, Random Forest e SVM, para classificar dados e foram testadas as variáveis dependentes, onde apenas a condição metástases/recidiva mostrou mais significância em classificar novos dados. Similarmente foi quantificada a importância das variáveis para a classificação em relação às regiões obtidas pelo processo de redução de variáveis, e foram obtidas 11 regiões cromossómicas clinicamente relevantes, sendo que amplificações das regiões 11q13.5-q14.1, 22q11.22-q11.23, 3p14.3-p14.2 e 11p14.1-p13 e deleções das regiões 22q11.22-q11.23, 6q16.1-q16.3, 17q21.31-q21.32, 3p14.3-p14.2 e 3q26.31-q26.33 apresentaram pior prognóstico e um tempo de sobrevida menor; e amplificações das regiões 17q21.31-q21.32 e 1q21.1-q21.2 apresentaram melhor prognóstico e uma sobrevivência maior. Por fim, identificaram-se potenciais biomarcadores/preditores para o CECP: *APPL1*, contido na região 3p14.3-p14.2 quando amplificada, tem um maior

risco de metástases; e os genes *BCR* e *SMARCB1*, contidos na região 22q11.22-q11.23 quer quando deletada ou amplificada, apresentam um pior prognóstico. Em conclusão, respondendo à questão fundamental deste projeto, os doentes de CECP dependendo das suas assinaturas genómicas têm diferentes taxas de sobrevivência e, consequentemente, melhor ou pior prognóstico. A descoberta de biomarcadores com aplicabilidade na prática clínica auxiliará na antecipação do aparecimento e/ou progressão da doença.

Palavras-chave: carcinoma epidermoide da cabeça e pescoço; alterações genéticas; metástases/recidiva; biomarcadores/preditores.

ABSTRACT

Head and Neck Squamous Cell Carcinoma (HNSCC) is a neoplasia that appears on the upper aero-digestive tract, including the oral cavity, oral pharynx, larynx, hypopharynx and tongue and is the 6th most common tumour in the world. The development of this carcinoma results of the gradual build-up of changes and genetic and epigenetics instability that lead to carcinogenesis and tumoral progression. Some studies have been made in order to identify genomice changes that can be used in the classification, diagnosis and prognosis, including predicting the effectiveness of the treatment. Regarding to of the advances on the diagnosis technologies and types of treatment, these tumours are still diagnosed in advanced stages without an improve survival rates. The main goal of this study is to answer the following question: Are the different survival rates of patients with HNSCC associated to the presence of tumours with different genomic signatures? Using results from array-Comparative Genomic Hybridization (aCGH) technic, that refer to Copy Number Variation (CNV) of genetic material, jointly with the clinic/pathologic information of the patients with the HNSCC, we have analysed 104 patients using the software platforms Matlab R2014a, R3.1.2 and SPSS. This way, as a result of this project it was verified the association of variables dependent on the tumour stage, tumour localization and the presence of metastasis connected with the aCGH data and were identified minimum common regions more frequently changed in the chromosomes. It was also possible to check the efficiency of the two classifiers, Random Forest and SVM to classify data and the dependent variables were tested, where only the condition metastasis/relapse showed the significance of new data. The same way, were quantified the importance of the variables for the classification in relation to the regions obtained by the reduction of variables, and were obtained 11 regions chromosomal clinically relevant, being the amplification of regions 11q13.5-q14.1, 22q11.22-q11.23, 3p14.3-p14.2 and 11p14.1-p13 and deletion of regions 22q11.22-q11.23, 6q16.1-q16.3, 17q21.31-q21.32, 3p14.3-p14.2 and 3q26.31-q26.33 related with a worse prognosis and a lower survival rate; and the amplification of regions 17q21.31-q21.32 and 1q21.1-q21.2 related with a better prognosis. Finally, potential biomarkers/predictors for HNSCC were identified: *APPL1*, contained in the region 3p14.3-p14.2 when amplified, has a greater risk of metastasis; and the genes *BCR* and *SMARCB1,* contained in the region 22q11.22-q11.23 whether when deleted or amplified, represent a worse diagnosis. Concluding, answering the fundamental

question of this project, the patients of HNSCC depending on their genomic signatures have different survival rates and, therefore, better or worse prognosis. The discovery of biomarkers that can be applied in the clinical practice will help to foresee the onset and/or progress of the disease.

Key-words: head and neck squamous cell carcinoma; genetic alterations; metastasis/relapse; biomarkers/predictors.

LISTA DE ABREVIATURAS

5-FU	5-fluorouracilo
aCGH	Hibridização Genómica Comparativa em Array
APPL1	Adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper containing 1
ATM	ATM serine/threonine kinase
ATR	ATR serine/threonine kinase
AUC	Area Under Curve
BCR	Breakpoint Cluster Region
BLC6	B-cell Lymphoma 6
BRCA1	Breast cancer 1
C	Citosina
CCND1	Cyclin D1
CDC42	Cell Division Cycle 42
CDKN2A	Cyclin-dependent kinase inhibitor 2A
CDKN2B	Cyclin-dependent kinase inhibitor 2B
CDKN2C	Cyclin-dependent kinase inhibitor 2C
CDKN2D	Cyclin-dependent kinase inhibitor 2D
CE	Carcinoma Epidermoide
CECP	Cancro Epidermoide da Cabeça e Pescoço
cfDNA	DNA livre em circulação
CH_3	Grupo metilo
CHEK1	Checkpoint kinase 1
CNV	Variação do Número de Cópias
CRK	V-crk avian sarcoma vírus CT10 oncogene homolog
CSMD1	CUB and Sushi multiple domains 1
DCC	DCC netrin 1 receptor
DCUN1D1	DCN1, defective in cullin neddylation1, domain containing 1
DDR	Resposta ao dano do DNA
DNMT	DNA metiltransferase
EBER	Pequeno RNA codificado pelo Vírus de Epstein-Barr
EBV	Vírus de Epstein-Barr

EGFR	Recetor do Fator de Crescimento Epidérmico
FDA	Administração de Alimentos e Medicamentos
FHIT	Fragile histidine triad
FISH	Hibridização *in-situ* de Fluorescência
FRA11F	Fragile site, aphidicolin type, common, fra (11)(q14.2)
G	Guanina
GALR1	Galanin receptor 1
GATA4	GATA binding protein 4
H2AFX	H2A histone family, member X
HPV	Vírus do Papiloma Humano
HPV16	Vírus do Papiloma Humano do tipo 16
HR	Hazard Ratio
HSV	Vírus Herpes Simplex
IC	Intervalo de confiança
IL12A	Interleukin 12A
ING1	Inhibitor of growth family, member 1
LOH	Perda de heterozigotia
LOI	Perda de *Imprinting*
LRP12	Low density lipoprotein receptor-related protein 12
miRNA	microRNA
MME	Membrane metallo-endopeptidase
MMH	Maximum Marginal Hyperplane
MRE11A	MRE11 homolog A, double strand break repair nuclease
MTAP	Methylthioadenosine phosphorylase
MTUS1	Microtubule associated tumour supressor 1
MYC	V-myc avian myelocytomatosis viral oncogene homolog
OCT	Tomografia de Coerência Ótica
PARD6G	Par-6 family cell polarity regulator gamma
PCA	*Principal componente analysis*
PIK3CA	Phosphatidylinositol-4-5-biphosphate 3-kinase, catalytic subunit alpha
PTK2	Protein kinase 2
PTPRD	Protein tyrosine phosphatase, receptor type, D

RAC1	Ras-Related C3 Botulinum Toxin Substrate 1 (Rho Family, Small GTP Binding Protein Rac1)
RARβ	Retinoic acid receptor beta
Rb	Proteína associada ao Retinoblastoma
RB1	Retinoblastoma
ROC	Receiver Operating Characteristic
SCCA1	Squamous Cell Carcinoma Antigen 1
SERPINB13	Serpin peptidase inhibitor, clade B (ovalbumin) member 13
SMARCB1	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1
SOX2	Sex determining region Y-box 2
SPSS	Statistical Package for the Social Sciences
SVM	Support Vector Machine
Tblue	Azul de Toluidina
TCGA	The Cancer Genome Atlas
TNM	Tumor, Nódulo, Metástases
TP53	Tumour protein p53
TP63	Tumour protein p63
TUSC3	Tumour supressor candidate 3
WHO	World Health Organization
WHSC1L1	Wolf-Hirschhorn syndrome candidate 1-like 1
WNT1	Wingless-type MMTV integration site family, member 1

LISTA DE FIGURAS

Figura 1 - Diversidade morfológica e anatómica da cabeça e pescoço. Adaptado de: *Stadler et al., 2008*.[11] 3

Figura 2 - Representação esquemática da prevalência (linha superior), incidência (linha média) e mortalidade (linha inferior) em todo o mundo (A) e na Europa (B) relativamente ao cancro na cavidade oral e língua. Adaptado de: *GLOBOCAN 2012*.[20] 4

Figura 3 - Caraterísticas da carcinogénese da cabeça e pescoço. (A) O processo normal da morfogénese escamosa na mucosa adulta. (B) Esquema das caraterísticas do CECP. A ordem exata da aquisição de alterações distintas não é clara. Além disso, vários genes podem contribuir para mais do que uma caraterística. Adaptado de: Rothenberg and Ellisen, 2012.(32) 8

Figura 4 - Desregulação do ciclo celular pelo HPV. A consequência molecular da expressão de E6 e E7 é a entrada do ciclo celular e inibição da apoptose mediada por p53, o que permite que o vírus se replique. Adaptado de: *Leemans et al.*, 20011.[9] 15

Figura 5 - Modelo da progressão genética da carcinogénese da cabeça e pescoço. O CECP progride através de um processo de várias etapas a partir de caraterísticas histológicas normais à hiperplasia, displasia leve, displasia moderada, displasia severa, carcinoma *in situ*, carcinoma invasivo, e metástases. Vários genes estão envolvidos, principalmente na progressão das metástases e nos estádios iniciais da progressão do tumor. Adaptado de: *Haddad and Shin, 2008* e *Pai and Westra. 2009*.[16, 29]. 16

Figura 6 - Vias alteradas na patogénese do CECP identificadas em estudos de sequenciação de todo o exoma. Vermelho: supressores de tumores putativos e estabelecidos; verde: oncogenes; castanho: outros genes/proteínas relevantes; azul: proteínas virais. Adaptado de: *Rothenberg and Ellisen, 2012*.[32] 17

Figura 7 - Metilação do DNA e (des)regulação do genoma - representação esquemática das principais alterações que ocorrem nas células cancerígenas. (A) (Retângulo verde) Ilhas CpG estão frequentemente associadas a promotores de genes e são resistentes à metilação do DNA. (B) (Retângulo vermelho) Nas células cancerígenas, as ilhas CpG são propensas à hipermetilação do DNA, a qual resulta no silenciamento do gene aberrante.

Legenda: Círculo branco, CpG não metilado; círculo preto, CpG metilado. Adaptado de: *Stirzaker et al., 2014*.[79] 23

Figura 8 - Processo "classifica, constrói, testa" de Classificação. 26

Figura 9 - Gráfico referente à variável importance plot de forma a reduzir o número de variáveis. Adaptado de: *Metagenomics. Statistics*.[89]. 29

Figura 10 - Representação esquemática do hiperplano pretendido no classificador SVM. 30

Figura 11 - Representação esquemática da distinção entre precisão e exatidão. **A.** Precisão sem exatidão. **B.** Exatidão com precisão. 32

Figura 12 - Esquema representativo da precisão e sensibilidade. As bolas representam todos elementos pertencentes à amostra, ou seja, as pessoas; as bolas com preenchimento cinzento correspondem às pessoas com doença ou condição clínica; as bolas com delimitação cinzenta correspondem às pessoas saudáveis. 33

Figura 13 - Representação do aCGH. Esta técnica consiste em quatro etapas: A – marcação do DNA genómico com diferentes fluorocromos, Cy3 e Cy5; B – hibridação na plataforma de *microarray*; C – leitura dos sinais pelo *scanner*; D – processamento dos dados pelo *software* e análise dos resultados. Adaptado de *Sharkey et al, 200*5[105] e imagem cedida pelo Laboratório de Citogenética e Genómica – Faculdade de Medicina da Universidade de Coimbra 39

Figura 14 - Representação esquemática das etapas do trabalho. 42

Figura 16 - Fluxograma do algoritmo do classificador Random Forest 47

Figura 17 - Fluxograma do algoritmo do classificador Random Forest e do classificador SVM. 48

Figura 18 - Fluxograma do algoritmo para o Gráfico Radar. 49

Figura 19 – Fluxograma do algoritmo para o ideograma cromossómico. 50

Figura 20 - Representação da AUC para Random Forest e SVM em relação ao estádio… 58

Figura 21 - Representação da AUC para Random Forest e SVM em relação à localização. 59

Figura 22 - Representação da AUC para Random Forest e SVM em relação à presença de metástases. 61

Figura 23 - Gráfico da curva de sobrevivência da região 11q13.5-q14.1 quando amplificada. 64

Figura 24 - Gráfico da curva de sobrevivência da região 22q11.22-q11.23 quando amplificada. 65

Figura 25 - Gráfico da curva de sobrevivência da região 17q21.31-q21.32 quando amplificada. 66

Figura 26 - Gráfico da curva de sobrevivência da região 22q11.22-q11.23 quando deletada. 67

Figura 27 - Gráfico da curva de sobrevivência da região 6q16.1-q16.3 quando deletada. 68

Figura 28 - Gráfico da curva de sobrevivência da região 17q21.31-q21.32 quando deletada. 69

Figura 29 - Gráfico da curva de sobrevivência da região 1q21.1-q21.2 quando amplificada. 70

Figura 30 - Gráfico da curva de sobrevivência da região 3p14.3-p14.2 quando amplificada. 71

Figura 31 - Gráfico da curva de sobrevivência da região 3p14.3-p14.2 quando deletada. 72

Figura 32 - Gráfico da curva de sobrevivência da região 3q26.31-q26.33 quando deletada. 73

Figura 33 - Gráfico da curva de sobrevivência da região 11p14.1-p13 quando amplificada. 74

Figura 34 - Gráfico radial das regiões relevantes relativamente ao CECP. 76

Figura 35 - Ideograma cromossómico de todas as CNVs presentes em todos os autossomas em pelo menos 10% dos doentes. A verde do lado direito do cromossoma estão esquematizados os ganhos do número de cópias, e a vermelho do lado esquerdo do cromossoma estão visíveis as perdas do número de cópias. 79

Figura 36 - Ideograma cromossómico das CNVs presentes nas regiões cromossómicas mais relevantes para o CECP e em pelo menos 10% dos doentes. A verde do lado direito

do cromossoma estão esquematizados os ganhos do número de cópias, e a vermelho do lado esquerdo do cromossoma estão visíveis as perdas do número de cópias. 80

LISTA DE TABELAS

Tabela 1 - Estadiamento TNM para o cancro oral. Adaptado de: *Neville et al., 2002*.[33]..... 9

Tabela 2 - Estádios do CE tendo como base na classificação TNM. Adaptado de: *Trotta et al., 2011*.[34] 10

Tabela 3 – Regiões cromossómicas e genes mais comuns envolvidos no CECP, quer em ganho e perda, e sua respetiva função.....18

Tabela 4 - As alterações epigenéticas mais comuns. Adaptado de: *Mascolo et al., 2012*.[76] 23

Tabela 5 - Esquematização da matriz de confusão.....34

Tabela 6 - Caraterísticas clínicas dos doentes com carcinoma da cabeça e pescoço.....51

Tabela 7 - Regiões mais relevantes com interesse clínico para o CECP em função do estádio. A frequência representa o número de vezes que nas 1000 execuções do código a variável foi considerada entre as 16 mais importantes.....52

Tabela 8 - Regiões mais relevantes com interesse clínico para o CECP em função da localização. A frequência representa o número de vezes que nas 1000 execuções do código a variável foi considerada entre as 15 mais importantes.....53

Tabela 9 - Regiões mais relevantes com interesse clínico para o CECP em função da presença de metástases. A frequência representa o número de vezes que nas 1000 execuções do código a variável foi considerada entre as 16 mais importantes.....54

Tabela 10 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest em relação ao estádio.....55

Tabela 11 - Valores médios dos parâmetros e respetivo desvio padrão obtidos por Random Forest em relação à localização.....55

Tabela 12 - Valores médios dos parâmetros e respetivo desvio padrão obtidos por Random Forest em relação à presença de metástases.....56

Tabela 13 - Valores médios e respetivo desvio padrão dos parâmetros do conjunto de treino, *trainSet*, do modelo em função do estádio.....57

Tabela 14 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest e SVM em relação ao estádio no conjunto de teste.....57

Tabela 15 - Valores médios e respetivo desvio padrão dos parâmetros do conjunto de treino, *trainSet*, do modelo em função da localização..... 59

Tabela 16 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest e SVM em relação à localização no conjunto de teste. 59

Tabela 17 - Valores médios e respetivo desvio padrão dos parâmetros do conjunto de treino, *trainSet*, do modelo em função da presença de metástases. 60

Tabela 18 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest e SVM em relação à presença de metástases no conjunto de teste. 60

Tabela 19 - Regiões mais relevantes com interesse clínico para o CECP determinadas pela regressão logística..... 63

Tabela 20 - Tempo de sobrevida médio para a região 11q13.5-q14.1 quando amplificada. 64

Tabela 21 - Tempo de sobrevida médio para a região 22q11.22-q11.23 quando amplificada. 65

Tabela 22 - Tempo de sobrevida médio para a região 17q21.31-q21.32 quando amplificada. 66

Tabela 23 - Tempo de sobrevida médio para a região 22q11.22-q11.23 quando deletada. 67

Tabela 24 - Tempo de sobrevida médio para a região 6q16.1-q16.3 quando deletada. .. 67

Tabela 25 - Tempo de sobrevida médio para a região 17q21.31-q21.32 quando deletada. 68

Tabela 26 - Tempo de sobrevida médio para a região 1q21.1-q21.2 quando amplificada. 69

Tabela 27 - Tempo de sobrevida médio para a região 3p14.3-q14.2 quando amplificada. 70

Tabela 28 - Tempo de sobrevida médio para a região 3p14.3-q14.2 quando deletada. .. 71

Tabela 29 - Tempo de sobrevida médio para a região 3q26.31-q26.33 quando deletada. 72

Tabela 30 - Tempo de sobrevida médio para a região 11p14.1-q13 quando amplificada. 73

Tabela 31 – Influência das regiões mais relevantes relativamente ao prognóstico do CECP. 74

ÍNDICE

AGRADECIMENTOS vii
RESUMO xi
ABSTRACT xiii
LISTA DE ABREVIATURAS xv
LISTA DE FIGURAS xix
LISTA DE TABELAS xxiii
ÍNDICE xxvii
1. INTRODUÇÃO 1
1.1. Cancro 1
1.1.1. Caraterísticas do Cancro 1
1.2. Cancro da Cabeça e Pescoço 2
1.2.1. Epidemiologia 3
1.2.2. Fatores de Risco 5
1.2.3. Histologia e Progressão do CECP 6
1.2.4. Estadiamento 8
1.2.5. Deteção e Diagnóstico 10
1.2.6. Terapia para CECP 12
1.2.7. Carcinogénese 13
1.2.7.1. Carcinogénese HPV 14
1.2.7.2. Carcinogénese não-HPV 15
1.2.8. Alterações nas Vias de Sinalização 16
1.2.9. Alterações Citogenéticas e Genes Alterados no CECP 17
1.2.10. Alterações Epigenéticas 22
1.3. Ferramentas Bioinformáticas na análise do CECP 24
1.3.1. Aprendizagem Indutiva 24
1.3.1.1. Aprendizagem Supervisionada 25
1.3.1.2. Aprendizagem Não supervisionada 25
1.3.2. Classificação 25
1.3.2.1. Classificadores 27
1.3.2.1.1. Random Forest 27
1.3.2.1.2. SVM 29
1.3.3. Regressão Logística 30

1.3.4. Análise de dados ... 31
1.3.4.1. Precisão, exatidão, sensibilidade e especificidade ... 31
1.3.4.2. Matriz de confusão ... 33
1.3.4.3. Curvas ROC ... 34
1.3.4.4. Valor p ... 35
1.3.4.5. Análise de Sobrevivência ... 36
1.4. Técnicas Laboratoriais ... 37
1.4.1. Hibridização Genómica Comparativa com Array (aCGH) ... 37
2. FUNDAMENTAÇÃO E OBJETIVOS ... 41
3. MATERIAIS E MÉTODOS ... 43
3.1. População em estudo ... 43
3.2. Análise por aCGH ... 43
3.3. Análise estatística ... 44
3.3.1. Importância das variáveis ... 44
3.3.2. Rotinas de Classificação ... 46
3.3.3. Visualização dos dados ... 48
4. RESULTADOS E DISCUSSÃO DOS RESULTADOS ... 51
4.1. População em estudo ... 51
4.2. Resultados obtidos por *Importance Plot*: redução de variáveis ... 51
4.3. Resultados obtidos por Random Forest ... 54
4.4. Resultados obtidos na comparação Random Forest e SVM ... 56
4.5. Resultados obtidos por Regressão Logística: determinação de variáveis relevantes para CECP ... 61
4.6. Análise de sobrevivência ... 63
4.7. Interpretação do gráfico radar ... 76
4.8. Interpretação do ideograma cromossómico ... 78
5. CONCLUSÕES ... 81
6. PERSPETIVAS FUTURAS ... 83
7. REFERÊNCIAS BIBLIOGRÁFICAS ... 85

1. INTRODUÇÃO

1.1. Cancro

O cancro é uma doença caraterizada por instabilidade genómica, que envolve alterações dinâmicas no genoma, surgindo como resultado de alterações que ocorreram na sequência de DNA das células cancerígenas.[1-3] As células cancerígenas têm alterações de regulação da proliferação celular normal e consequente da homeostasia.[2, 4] Existem mais de 100 tipos de cancro distintos, e dentro do mesmo órgão existem subtipos de tumores específicos.[2]

Ao longo dos anos, no que respeita à instabilidade genómica do cancro, os investigadores mostraram que mutações em alguns genes estão diretamente relacionadas com o cancro, podendo os genes ser divididos em duas classes: oncogenes - que são derivados a partir de proto-oncogenes que sofrem mutações com ganho dominante da função - e genes supressores de tumor - em que mutações levam a uma perda da função.[2, 3]

O cancro continua a ser uma das principais causas de morbidade e mortalidade em todo o mundo.[5] A sua ocorrência está a aumentar por causa do crescimento e envelhecimento da população, bem como devido ao aumento da prevalência de fatores de risco estabelecidos, como tabagismo, excesso de peso, sedentarismo e mudança de padrões reprodutivos associados à urbanização e desenvolvimento económico.[6]

Com base em estimativas da GLOBOCAN, em 2012, ocorreram cerca de 14,1 milhões de novos casos de cancro, 8,2 milhões de mortes e 32,6 milhões de pessoas vivem com cancro durante 5 anos após o diagnóstico em todo o mundo.[6] Prevê-se que até 2020, o número de novos casos de cancro no mundo aumente para mais de 15 milhões e as mortes para 12 milhões.[5] Segundo a *World Health Organization* (WHO) estima-se que, em 2035, quase 15 milhões de pessoas por ano morram de cancro.[7]

Sendo o cancro uma doença do genoma, carateriza-se por um conjunto de alterações genómicas que podem ditar o comportamento e resposta ao tratamento clínico.[3, 4]

1.1.1. Caraterísticas do Cancro

As seis caraterísticas biológicas essenciais na fisiologia celular adquiridas durante o desenvolvimento de tumores humanos surgem de um vasto catálogo dos genótipos de células cancerígenas. Estas caraterísticas, que constituem um princípio organizador da complexa doença neoplásica, são: autossuficiência em sinais de crescimento;

insensibilidade aos sinais inibidores de crescimento (anti-crescimento); evasão à morte celular programada (apoptose); potencial replicativo ilimitado; angiogénese sustentada; e invasão de tecidos e metástases.[2, 8] Cada uma destas alterações fisiológicas representa a rutura de um mecanismo de defesa anticancerígena programado em células e tecidos. [2]

Subjacente a estas caraterísticas estão a instabilidade do genoma, que gera a diversidade genética que acelera a sua aquisição, e a inflamação, que promove muitas das caraterísticas já mencionadas. O progresso conceptual na última década acrescentou duas caraterísticas emergentes ao potencial generalizado a esta lista, sendo elas a reprogramação do metabolismo energético e a evasão da destruição imunológica. Além das células cancerígenas, os tumores exibem uma outra dimensão de complexidade: eles contêm um reportório de células aparentemente normais recrutadas que contribuem para a aquisição de traços caraterísticos, criando o "microambiente do tumor". O reconhecimento de uma aplicabilidade generalizada destes conceitos irá afetar cada vez mais o desenvolvimento de novos meios para o tratamento do cancro.[8]

1.2. Cancro da Cabeça e Pescoço

O Carcinoma Epidermoide da Cabeça e Pescoço (CECP) é o sexto tipo de cancro mais incidente no mundo.[9]

O CECP é uma neoplasia complexa que surge nas regiões do trato aerodigestivo superior, incluindo a cavidade oral, orofaringe, laringe, hipofaringe e língua. Os tumores a partir destes locais têm apresentações e resultados clínicos distintos, e são tratados com estratégias diferentes.[9, 10] Existe então uma matriz extremamente diversificada da anatomia e morfologias tumorais, com pelo menos dez sublocais anatómicos da cabeça e pescoço, tornando-se assim um desafio a definição com precisão da extensão da doença, conforme ilustrado na Figura 1.[11]

O CECP resulta da acumulação de alterações genéticas e epigenéticas numa variedade de vias celulares, e especificamente em oncogenes, genes supressores de tumor, e/ou genes de estabilidade do DNA.[12, 13] Contudo, o estudo do desenvolvimento e progressão do tumor é dificultado pela sua complexidade biológica sendo o CECP conhecido por ser muito heterogéneo em relação à localização inicial, a nível molecular e histológico, história natural e prognóstico do tumor primário tanto a nível molecular como histológico.[13, 14]

Apesar dos progressos tecnológicos e clínicos terem avançado muito, as taxas de sobrevivência a cinco anos, de aproximadamente 50%, têm apenas moderadamente melhorado durante os últimos 20 anos.[15]

Os sintomas deste tipo de cancro variam, dependendo do local de origem, e podem incluir uma dor de garganta, disfagia, odinofagia e rouquidão.[16]

Como já referido, os tumores desenvolvem-se geralmente dentro de campos pré-neoplásicos de células geneticamente alteradas e, dessa forma, a persistência desses campos após o tratamento representa um grande desafio, pois pode levar a recorrências locais e aparecimento de segundos tumores primários que são responsáveis por uma grande proporção de mortes, especificamente no CECP.[9]

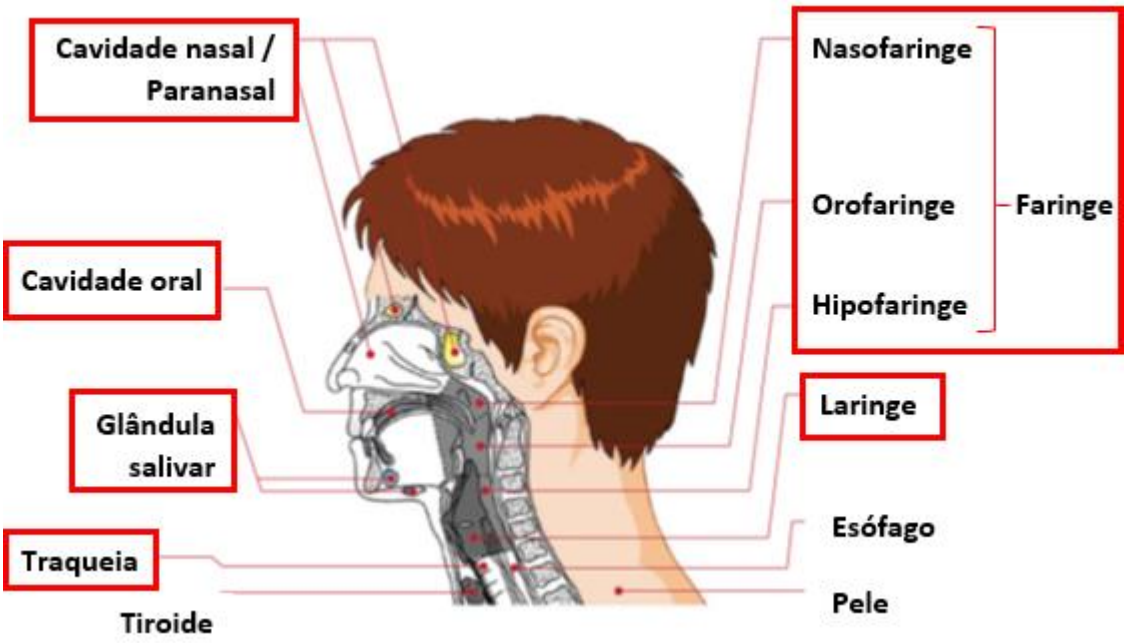


Figura 1 - Diversidade morfológica e anatómica da cabeça e pescoço. Adaptado de: *Stadler et al., 2008*.[11]

1.2.1. Epidemiologia

O Cancro da Cabeça e Pescoço é o sexto tipo mais comum de cancro, o que representa cerca de 6% de todos os casos, e em particular, o CECP é também o sexto tumor mais maligno em todo o mundo. Estimam-se 650000 novos casos de cancro por ano e 50% destes casos morrem devido à doença, ou seja 350000 mortes em todo o mundo a cada ano.[10, 17, 18]

A maioria, mais de 90%, dos casos do cancro oral, faringe e laringe (coletivamente referidos como Cancros da Cabeça e Pescoço) deriva de uma origem epitelial escamosa.[14, 19]

A Figura 2 representa a prevalência, incidência e mortalidade do cancro da cavidade oral e língua em todo o mundo e na Europa.[20]

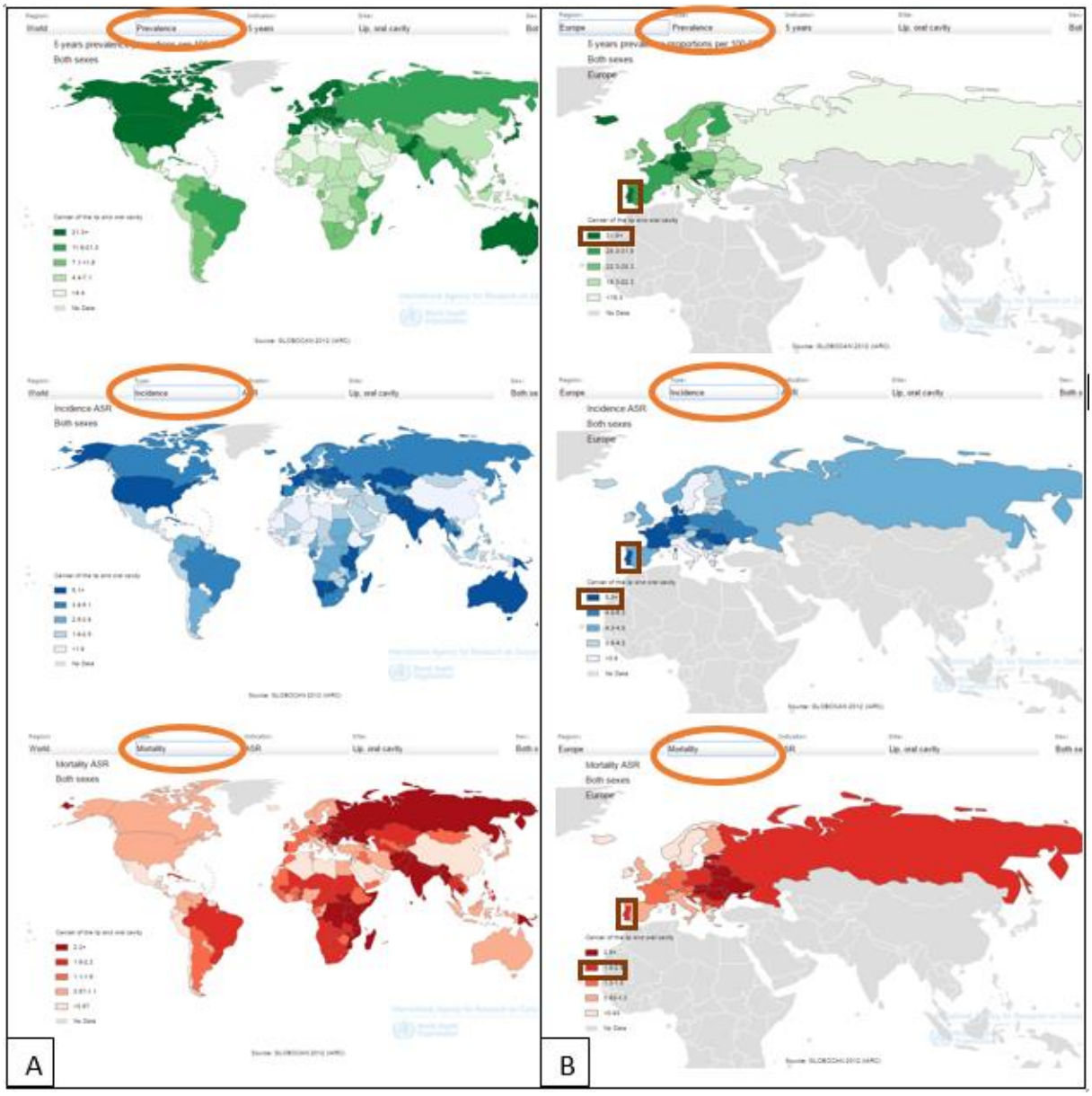


Figura 2 - Representação esquemática da prevalência (linha superior), incidência (linha média) e mortalidade (linha inferior) em todo o mundo (A) e na Europa (B) relativamente ao cancro na cavidade oral e língua. Adaptado de: *GLOBOCAN 2012*.[20]

1.2.2. Fatores de Risco

O CECP apresenta diversos fatores de risco, alguns dos quais já são bem conhecidos, enquanto outros ainda não estão bem esclarecidos face ao seu papel relativamente a este carcinoma. Existem fatores exógenos, endógenos, e/ou químicos e físicos.

Os fatores de risco mais importantes até agora identificados são o tabagismo e o consumo de álcool, o qual parece ter um efeito sinergético quando combinado com o tabaco.[9, 10, 17, 18] Diversos estudos demonstraram que o tabaco e álcool aumentam o risco de CECP de uma forma dose-resposta.[21] O tabaco e o álcool são conhecidos por induzir mutações e alterações cromossómicas, afetando genes relevantes para o controlo do ciclo celular.[14] No entanto, os mecanismos exatos pelos quais as substâncias cancerígenas do tabaco e do álcool induzem a transformação e progressão maligna de células epiteliais na cabeça e pescoço não são completamente compreendidos. E o facto de que a maioria dos consumidores de tabaco e álcool não desenvolverem CECP nas suas vidas e que aproximadamente 20% dos doentes com CECP, particularmente do sexo feminino e com idade mais jovem, não tenham nenhuma evidência clara de exposição aos fatores de risco comuns reforça a complexidade da tumorigénese do CECP e o importante papel das interações gene-ambiente no processo tumorigénico. Contudo, *Mao et al.* provaram que o risco de desenvolvimento do cancro oral encontra-se aumentado de três a nove vezes nas pessoas que fumam ou bebem e até cem vezes em pessoas que fumam ou bebem muito relativamente às que não fumam nem bebem.[10]

Um subgrupo dos CECPs, particularmente aqueles da orofaringe, é causado por infeção do Vírus do Papiloma Humano (HPV), maioritariamente o Vírus do Papiloma Humano do tipo 16 (HPV16) e em menor grau o tipo 18.[9, 17, 22, 23] Dessa forma, a associação entre HPV e CECP tem implicações potencialmente importantes para a prevenção, tratamento e prognóstico.[17, 19]

Outro vírus, o Vírus de Epstein-Barr (EBV), pode também ser um agente causador do Carcinoma Nasofaríngeo e a expressão de genes do EBV pode estar estritamente relacionada com a patogénese deste carcinoma.[22, 24] Pequenos RNAs codificados pelo Vírus Epstein-Barr (EBERs) induzem a transformação inicial das células epiteliais, contribuindo assim para a oncogénese do Carcinoma Nasofaríngeo.[24]

Um outro vírus, o Herpes Simplex (HSV), foi também demonstrado ser um agente promotor do cancro oral.[25]

Estes vírus são, desta forma, capazes de sequestrar o aparelho celular do hospedeiro e modificar o DNA e as estruturas cromossómicas e induzir alterações proliferativas nas células.[25]

Embora o CECP surja esporadicamente, a herança familiar tem sido também observada.[17] Além disso, certas doenças hereditárias, como a anemia de Fanconi que é uma doença autossómica recessiva caraterizada por anemia aplástica e alterações congénitas e com instabilidade genómica, apresentam uma suscetibilidade genética que parece predispor para CECP.[9, 18, 22, 26] O risco para CECP aumenta em indivíduos com síndromes de suscetibilidade ao cancro, como o cancro colo-retal não polipose hereditário, síndrome Li-Fraumeni, anemia de Fanconi, como descrito anteriormente, e ataxia telangiectasia.[17]

Outros fatores de risco para o CECP incluem a exposição a agentes cancerígenos exógenos como a nutrição, má higiene oral, infeções, contaminantes ambientais tais como fumos de tinta, subprodutos de plástico, e fumos de gasolina, doença do refluxo gastro-esofágico, fatores dietéticos e uso de marijuana. A sífilis, os fatores oro-dentais, e a candidíase crónica são mais fatores de risco do surgimento do CECP.[10, 18, 25]

Alguns cientistas relataram menor risco do CECP com maior ingestão de frutas e legumes.[17, 25]

1.2.3. Histologia e Progressão do CECP

A definição do Carcinoma Epidermoide (CE), segundo a *WHO Classification*, é "uma neoplasia epitelial invasiva com diferentes graus de diferenciação escamosa e uma propensão para metástases em nódulos linfáticos precoces e extensas, ocorrendo predominantemente em adultos consumidores de álcool e tabaco na quinta e sexta década de vida".[27]

Já relativamente à histopatologia do CE esta é definida segundo a *WHO Classification* como: "diferenciação escamosa, muitas vezes vista como queratinização com formação "pérola" variável, e crescimento invasivo. A invasão é manifestada pela rutura da membrana basal, e extensão no tecido subjacente, muitas vezes acompanhada por reação do estroma. Invasão angio-linfática e perineural são sinais adicionais de malignidade".[27]

Por sua vez, os tumores são tradicionalmente classificados em carcinoma bem, moderadamente ou pouco diferenciados.[27]

O CE bem diferenciado assemelha-se a epitélio escamoso estritamente normal. Por sua vez, o CE moderadamente diferenciado contém pleomorfismo nuclear distinto, atividade mitótica incluindo mitoses anormais e há geralmente menos queratinização neste tipo de carcinoma. Por fim, no CE pouco diferenciado predominam as células imaturas, com numerosas mitoses, típicas e atípicas, e a queratinização é mínima.[27]

O cancro oral é uma neoplasia epitelial que geralmente começa como uma proliferação celular anormal clonal de células estaminais alteradas perto da membrana basal, expandindo para cima e lateralmente, substituindo assim o epitélio normal. O processo neoplásico inicia-se com o epitélio normal progredindo da hiperplasia à displasia, ao carcinoma *in situ* e ao carcinoma invasivo, tal como ilustrado na Figura 3.[28] A cavidade oral é revestida por um epitélio escamoso estratificado que varia ligeiramente com respeito à espessura e à queratinização como uma função de exposição a forças de mastigação. A sua interface com a lâmina própria adjacente é abrupta e delineada por uma membrana basal contígua – uma estrutura que regula a diferenciação e a migração das células epiteliais e serve como uma barreira para a invasão do estroma durante a tumorigénese.[29]

Concomitante, o CE histologicamente é caraterizado por uma displasia escamosa que se refere a alterações neoplásicas do epitélio da superfície antes da invasão dos tecidos conjuntivos subepiteliais. Essas mudanças incluem a organização celular anormal, atividade mitótica aumentada e alargamento nuclear com pleomorfismo. Estas alterações são tipicamente classificadas numa escala de 1 a 3 com base na gravidade da atipia. Embora a terminologia varie, atipia limitada a um terço inferior do epitélio é geralmente referida como displasia leve, atipia limitada aos dois terços inferiores como displasia moderada, e atipia que envolve toda a espessura do epitélio como displasia grave/carcinoma *in situ*.[29-31]

Várias lesões orais são de particular relevância para o cancro oral, sendo elas a leucoplasia oral, líquen plano oral e eritroplasia oral. A leucoplasia oral é um diagnóstico clínico que descreve manchas brancas ou placas que não podem ser atribuídas a outra doença. Por sua vez, o líquen plano oral deverá ser uma doença autoimune. E por fim, a eritroplasia oral é considerada uma doença rara. Estas três lesões, leucoplasia oral, líquen plano oral e eritroplasia oral, podem mostrar diferentes graus de anormalidades histológicas, da displasia leve ao carcinoma *in situ*.[31]

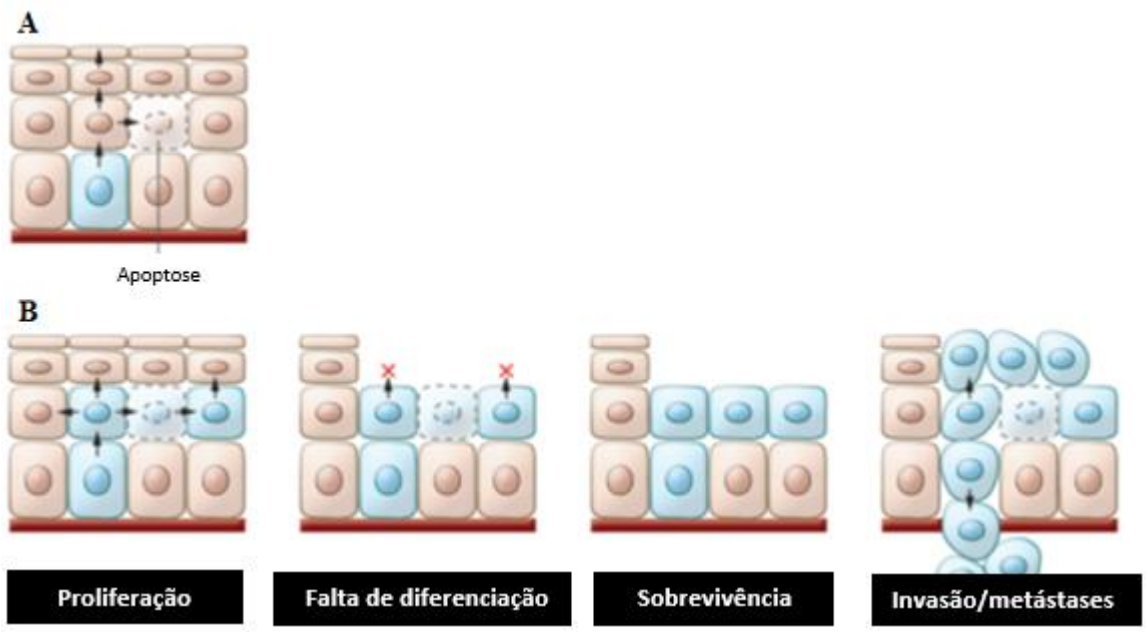


Figura 3 - Caraterísticas da carcinogénese da cabeça e pescoço. (A) O processo normal da morfogénese escamosa na mucosa adulta. (B) Esquema das caraterísticas do CECP. A ordem exata da aquisição de alterações distintas não é clara. Além disso, vários genes podem contribuir para mais do que uma caraterística. Adaptado de: Rothenberg and Ellisen, 2012.(32)

1.2.4. Estadiamento

O estadiamento do cancro oral é importante para estabelecer o tratamento adequado e determinar o prognóstico. O CE da cavidade oral e orofaringe e carcinoma não-epidermoide das glândulas salivares menores são convencionalmente estadiados usando o sistema de classificação "tumor, nódulo, metástases" TNM, o qual é baseado na avaliação clínica e patológica do tamanho do tumor e comprometimento dos nódulos linfáticos. O T representa o tamanho do tumor primário, N indica o estado dos nódulos linfáticos regionais, e M indica a presença ou ausência de metástases distantes.[31, 33, 34]

A sobrevivência dos doentes com cancro oral e orofaringe está fortemente relacionada com o estádio da doença no momento do diagnóstico.[33] No entanto, o estadiamento tradicional é muitas vezes insuficiente e nem sempre fornece informações precisas de prognóstico.[31]

Na Tabela 1 está exposto o estadiamento TNM para o cancro oral.

Tabela 1 - Estadiamento TNM para o cancro oral. Adaptado de: *Neville et al., 2002*.[33]

Tumor Primário (T)	
Tx	Tumor primário não pode ser avaliado
T0	Não há evidência do tumor primário
Tis	Carcinoma *in situ*
T1	Tumor com 2 cm ou menos na sua maior dimensão
T2	Tumor com mais de 2 cm mas não mais que 4 cm na sua maior dimensão
T3	Tumor com mais de 4 cm na sua maior dimensão
T4	Tumor invade estruturas adjacentes
Envolvimento ganglionar (N)	
Nx	Nódulos linfáticos regionais não podem ser avaliados
N0	Ausência de metástases dos nódulos linfáticos regionais
N1	Metástases num único nódulo linfático homolateral, com 3 cm ou menos na sua maior dimensão
N2	Metástases num único nódulo linfático homolateral, com mais de 3 cm, mas não mais que 6 cm na sua maior dimensão; ou em múltiplos nódulos linfáticos homolaterais, nenhum deles com mais de 6 cm na sua maior dimensão; ou nos nódulos linfáticos bilaterais ou contra-laterais, nenhum deles com mais de 6 cm na sua maior dimensão
N2a	Metástases num único nódulo linfático homolateral, com mais de 3 cm, mas não mais que 6 cm na sua maior dimensão
N2b	Metástases em múltiplos nódulos homolaterais, nenhum deles com mais de 6 cm na sua maior dimensão
N2c	Metástases em nódulos linfáticos bilaterais ou contra-laterais, nenhum deles com mais de 6 cm na sua maior dimensão
N3	Metástases num nódulo linfático com mais de 6 cm na sua maior dimensão
Metástases à distância (M)	
Mx	Metástase à distância não pode ser avaliada
M0	Nenhuma metástase à distância
M1	Metástases à distância

Atualmente há 32 possíveis combinações das categorias T, N e M, que são agregadas em sete estádios do CE: 0, I, II, III, IVA, IVB e IVC. (Tabela 2) O agrupamento das categorias

TNM em estádios representa ambos os dados de sobrevivência homogéneos e variações importantes nas caraterísticas da doença que podem afetar as opções de tratamento.[34]

Tabela 2 - Estádios do CE tendo como base na classificação TNM. Adaptado de: *Trotta et al., 2011*.[34]

Estádio do cancro	Categoria T	Categoria N	Categoria M
0	Tis	N0	M0
I	T1	N0	M0
II	T2	N0	M0
III	T1, T2 T3	N1 N0, N1	M0 M0
IVA	T1, T2, T3 T4a	N2 N0, N1, N2	M0 M0
IVB	Qualquer T4b	N3 Qualquer	M0 M0
IVC	Qualquer	Qualquer	M1

1.2.5. Deteção e Diagnóstico

A cavidade oral é uma região facilmente acessível que pode ser examinada visualmente com pouco desconforto, contudo a maioria dos cancros orais, mesmo em países desenvolvidos, são geralmente detetados em estádios tardios. Além disso, a maioria dos cancros orais, em particular o CE oral, é diagnosticado quando os sinais e sintomas já estão presentes.[35, 36] Não surpreendentemente, o custo para o tratamento de um doente com CE oral no estádio III ou estádio IV é muito maior do que o tratamento para um doente com CE oral no estádio I ou II.[35-38]

O rastreio, em medicina, é uma estratégia utilizada na população para detetar uma doença em indivíduos sem quaisquer sinais ou sintomas da doença ou examinação de um grupo de indivíduos assintomáticos para detetar aqueles com uma elevada probabilidade de ter ou vir a desenvolver uma dada doença.[38] O Exame Oral Convencional é o método mais comum para o rastreio e é útil na descoberta de algumas lesões orais, contudo, este método não tem a capacidade para identificar corretamente todas as lesões potencialmente malignas orais, nem detetar com precisão a pequena proporção de lesões biologicamente relevantes que são propensas a progredir para cancro. E é também, um teste subjetivo, fortemente dependente da experiência e

habilidade dos clínicos.[38] Este exame físico consiste na deteção visual de nódulos, alterações da mucosa (como alterações de cor ou textura e úlceras), inchaços e adenopatia da linfa inexplicável.[37]

Um outro teste de rastreio, com base na análise PCR em tempo real de SCCA1 (Squamous Cell Carcinoma Antigen 1) que é independente do examinador, foi desenvolvido apresentando uma sensibilidade e especificidade de aproximadamente 70%. No entanto, estudos complementares ainda se encontram em curso de forma a validar a sua utilização na prática clínica.[39]

Atualmente, o método padrão de diagnóstico para o CE é o exame clínico com endoscopia de luz branca, seguido de biópsia por agulha invasiva e análise histopatológica. As limitações clínicas deste método de diagnóstico, como por exemplo técnicas microscópicas inadequadas bem como análises histopatológicas propensas a favorecer um determinado diagnóstico, podem negligenciar o diagnóstico precoce da doença.[40, 41]

Para além deste existem outros métodos de diagnóstico, por exemplo: os diagnósticos óticos que são caracterizados pelo uso de diferentes comprimentos de onda de luz para examinar o tecido suspeito de uma forma não invasiva *in vivo*. A profundidade da penetração da luz no tecido é dependente do seu comprimento de onda. Desta forma, as biópsias óticas podem ser produzidas e fornecer uma avaliação rápida da arquitetura de tecido e detetar patologias da submucosa.[40] Desta forma são usadas para diagnóstico as modalidades óticas *ex vivo*, tais como espectroscopia de Raman, a Tomografia de Coerência Ótica (OCT), o diagnóstico de fluorescência, a endomicroscopia confocal a laser, a imagem de banda estreita e a microscopia confocal de reflectância.[40] As modalidades de diagnóstico ótico atualmente utilizadas sugerem a possibilidade de um custo eficaz do diagnóstico bem como a avaliação em tempo real o que poderá facilitar a deteção precoce, reduzir os custos de saúde e, concomitantemente melhorar a sobrevivência do doente e a sua qualidade de vida.[40] Estas técnicas, tais como autofluorescência e espectroscopia trimodal (espectroscopia de fluorescência, espectroscopia de dispersão elástica e espectroscopia de Raman), permitem assim reduzir o tempo de espera e prestar assistência na seleção do local da biópsia.[38, 40]

Outro método de deteção precoce é o Azul de Toluidina (Tblue), um corante metacromático que cora seletivamente os componentes ácidos dos tecidos do DNA e RNA e que se baseia na teoria de que o teor do componente celular acídico é maior no

tecido displásico do que no tecido não-displásico, utilizado assim na deteção de pré-cancro e cancro.[37, 38]

A citologia oral é uma outra técnica de diagnóstico utilizada para análise histopatológica, a qual inclui citomorfometria, citometria de DNA e análise imunocitoquímica. O uso da citologia oral na deteção de lesões displásicas é essencial, mas tem sido limitada, até agora, por resultados variáveis de falso-positivos e falso-negativos.[38]

Os métodos julgados mais promissores de deteção do cancro oral fazem uso de biomarcadores presentes nos fluidos biológicos, tais como saliva e plasma.[37]

Um biomarcador ideal deve ter algumas caraterísticas, tais como especificidade para a doença, presença obrigatória em todos os doentes afetados, reversibilidade após o tratamento adequado, e deteção precoce, isto é, antes dos doentes desenvolverem manifestações clínicas evidentes da doença. Além disso, os biomarcadores ideais devem refletir não só a gravidade da doença, mas também fornecer informações sobre a história cumulativa da doença.[41]

A evidência sugere que os biomarcadores salivares são uma opção de diagnóstico viável. No entanto, é necessário determinar qual dos diversos biomarcadores relatados exibe reprodutibilidade aceitável do teste de diagnóstico.[41]

Por fim, as análises sanguíneas são uma técnica minimamente invasiva que tem tido um ganho no valor clínico no diagnóstico do cancro. No plasma de vários doentes com cancro, a presença de DNA livre em circulação (cfDNA) a partir de células tumorais tem sido sugerido como tendo valor de diagnóstico. Já se descobriu que o DNA contém alterações genéticas e epigenéticas que estão relacionadas com a iniciação, progressão e resistência do cancro, tais como perda de heterozigotia (LOH) – do Inglês *Loss of Heterozigosity* - e mutações em genes supressores de tumores/oncogenes.[42]

Existem assim diversas técnicas que podem ser usadas na deteção e diagnóstico do CECP, contudo, a melhor técnica de diagnóstico continua a ser aquela que se tenha experiência e treino suficiente e mais à-vontade na sua realização.[43]

1.2.6. Terapia para CECP

As opções de tratamento atuais incluem radioterapia com ou sem quimioterapia, cirurgia com ou sem radioterapia ou quimioterapia, e mais recentemente radioterapia com agentes moleculares alvo, tal como Cetuximab. A escolha do tratamento dependerá de fatores clínicos, tais como o local, estádio, e tamanho do tumor. Além do sistema de

estadiamento, o fator clínico mais importante para prever o *outcome* é o tamanho do tumor. No entanto, há uma discrepância notável entre os doentes na probabilidade de recorrência e sobrevivência mesmo após se considerar o estádio e tamanho do tumor. Dada a instabilidade genómica inerente dos cancros, incluindo os da cabeça e pescoço, é provável que a grande disparidade possa ser explicada por fatores biológicos que diferem marcadamente entre tumores.[44, 45]

O CECP em fase inicial geralmente é curável com radioterapia definitiva/cirurgia. A cirurgia também pode ser o tratamento inicial para certos doentes com doença localmente avançada. Para doentes com caraterísticas pós-cirúrgicas adversas ou doença localmente avançada irressecável, a quimio-radioterapia é o padrão atual.[46, 47]

Por sua vez, para a maioria dos doentes que inicialmente são diagnosticados com doença localmente avançada, ou seja, estádio III/IV, muitas vezes é necessário uma abordagem multidisciplinar incorporando a quimioterapia, radioterapia, e cirurgia. Os doentes com CECP recidente/metastático têm um mau prognóstico e algumas outras opções de tratamento eficazes são necessárias, até mesmo a quimioterapia paliativa.[46]

O tratamento para o CECP requer um delicado equilíbrio entre a erradicação completa do cancro e a preservação da forma anatómica e função dos órgãos desta região sensível. Preservação de órgãos e qualidade de vida são fatores críticos para o planeamento do tratamento, uma vez que os órgãos afetados por este tipo de cancro são extremamente importantes para a vida do doente.[48]

1.2.7. Carcinogénese

A carcinogénese oral é o resultado de uma acumulação progressiva de alterações genéticas, que inclui três etapas principais: iniciação, promoção e progressão. Para este processo de múltiplas etapas ter sucesso, diversos processos celulares e desequilíbrios devem ocorrer. A exposição a fatores cancerígenos pode conduzir à expressão anormal de genes supressores de tumor e/ou proto-oncogenes, os quais, por sua vez, ativam as vias que conduzem à transformação maligna das células. Muitas vezes, essa expressão anormal pode incluir uma mutação esporádica, deleção, LOH, expressão aumentada, ou modificação epigenética.[11, 28]

Em neoplasias, a proliferação celular é excessiva e autónoma, não coordenada com a divisão celular a decorrer nos tecidos normais, provocada pelos danos do DNA devido à

perda dos *checkpoints* do ciclo celular. Estas células neoplásicas em última análise entram nos vasos linfáticos e metastizam para os nódulos linfáticos regionais.[28]

1.2.7.1. Carcinogénese HPV

Nas últimas décadas surgiu a associação entre o HPV e um subconjunto do cancro da cabeça e pescoço, como já referido. O rápido aumento dos tumores da orofaringe relacionados com HPV tem sido predominantemente relatado em países economicamente desenvolvidos e a infeção por HPV foi encontrada fortemente associada com uma boa resposta terapêutica e sobrevivência.[14]

O HPV é um grupo heterogéneo de pequenos vírus de DNA causando uma variedade de lesões epiteliais proliferativas em locais específicos do corpo.[14] O HPV pertence à família Papillomaviridae, vírus de DNA de cadeia dupla circular de 8 Kb que infeta as células basais da mucosa epitelial.[49]

A carcinogénese induzida pelo HPV pensa-se que está presente em cerca de 20-25% dos casos de CECP e os doentes tendem a ser mais jovens, sem histórico de tabaco ou consumo de álcool.[14] O perfil clínico e biológico destes doentes é distinto do de outros doentes com carcinoma da orofaringe, com início mais precoce, nódulos cervicais císticos, e histopatologia do carcinoma basaloide.[49]

O ciclo de vida do HPV está relacionado com a fase de diferenciação das células epiteliais que estão infetadas. A infeção começa nas células em proliferação indiferenciadas da camada epitelial basal, que conduz à expressão dos genes precoces virais, em particular os oncogenes E6 e E7 (Figura 4). Estes atingem as vias do antigénio do p53 e proteína associada ao Retinoblastoma (Rb), respetivamente. A proteína E6 liga-se à p53 e tem como alvo a degradação da proteína, que leva à inibição da apoptose mediada por p53. A proteína E7 liga-se à Rb e inativa-a.[9, 49]

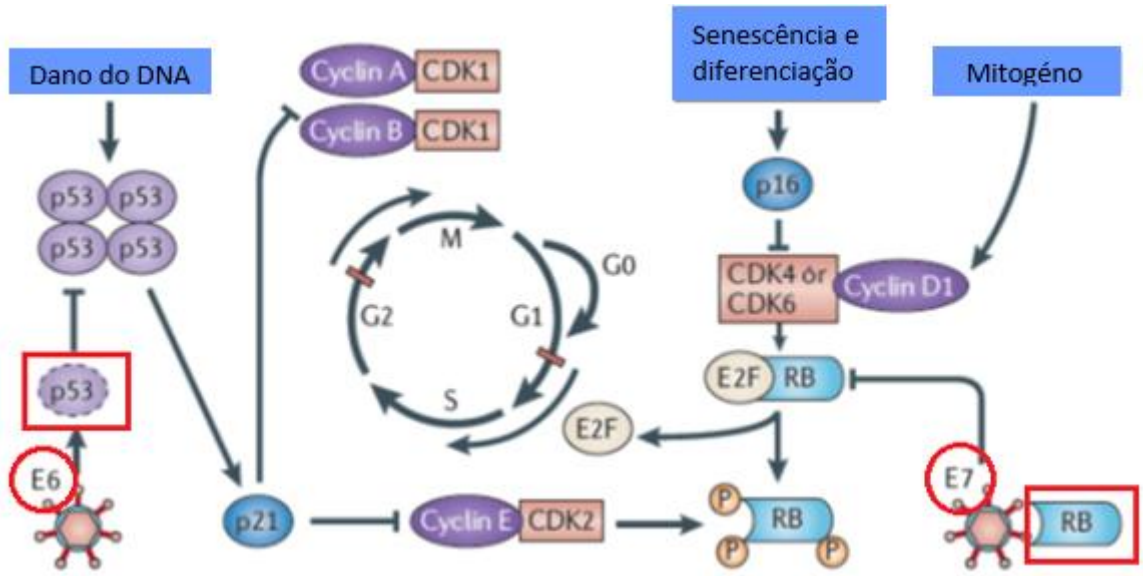


Figura 4 - Desregulação do ciclo celular pelo HPV. A consequência molecular da expressão de E6 e E7 é a entrada do ciclo celular e inibição da apoptose mediada por p53, o que permite que o vírus se replique. Adaptado de: *Leemans et al.*, 20011.[9]

1.2.7.2. Carcinogénese não-HPV

Uma grande quantidade de acontecimentos genéticos que conduzem à inativação de genes supressores de tumores e/ou à ativação dos proto-oncogenes, regulam o desenvolvimento do CECP.[10]

A maior parte do conhecimento sobre a carcinogénese do CECP foi adquirida a partir de estudos do cancro oral, principalmente porque o cancro oral é o CECP mais comumente diagnosticado. Além disso, lesões potencialmente malignas orais são a patologia mais frequentemente diagnosticada.[9, 50] Leucoplasia, lesão branca, e eritroplasia, lesão vermelha, são clinicamente lesões detetadas que podem histologicamente representar hiperplasia, displasia, e mesmo carcinoma *in situ* ou invasivo.[10]

O CECP é uma doença complexa caraterizada por heterogeneidade clínica, patológica, fenotípica e biológica.[16] Fiel ao modelo atual de desenvolvimento de neoplasias, o início e progressão do Cancro da Cabeça e Pescoço é um processo complexo de várias etapas que implicam a aquisição progressiva das alterações genéticas e epigenéticas e desregulação nas vias de sinalização associadas ao cancro (Figura 5).[29, 32, 50]

Carcinomas que se desenvolvem em grandes campos pré-neoplásicos geralmente estendem-se até às margens cirúrgicas da resseção do tumor, levando a recidivas e segundos tumores primários. Além de invasões locais, o CECP também é caraterizado por difusão dos nódulos cervicais, levando à conclusão de que é uma doença associada com progressão metastática e invasão. A progressão metastática e invasão de células

tumorais parece estar estritamente relacionada com as interações entre as células e o microambiente que as rodeia, que inclui a adesão celular, rearranjos do citoesqueleto, migração e degradação celular da membrana basal, intravasão, sobrevivência nos vasos sanguíneos, extravasão, crescimento das células tumorais no local distante e promoção da angiogénese.[50]

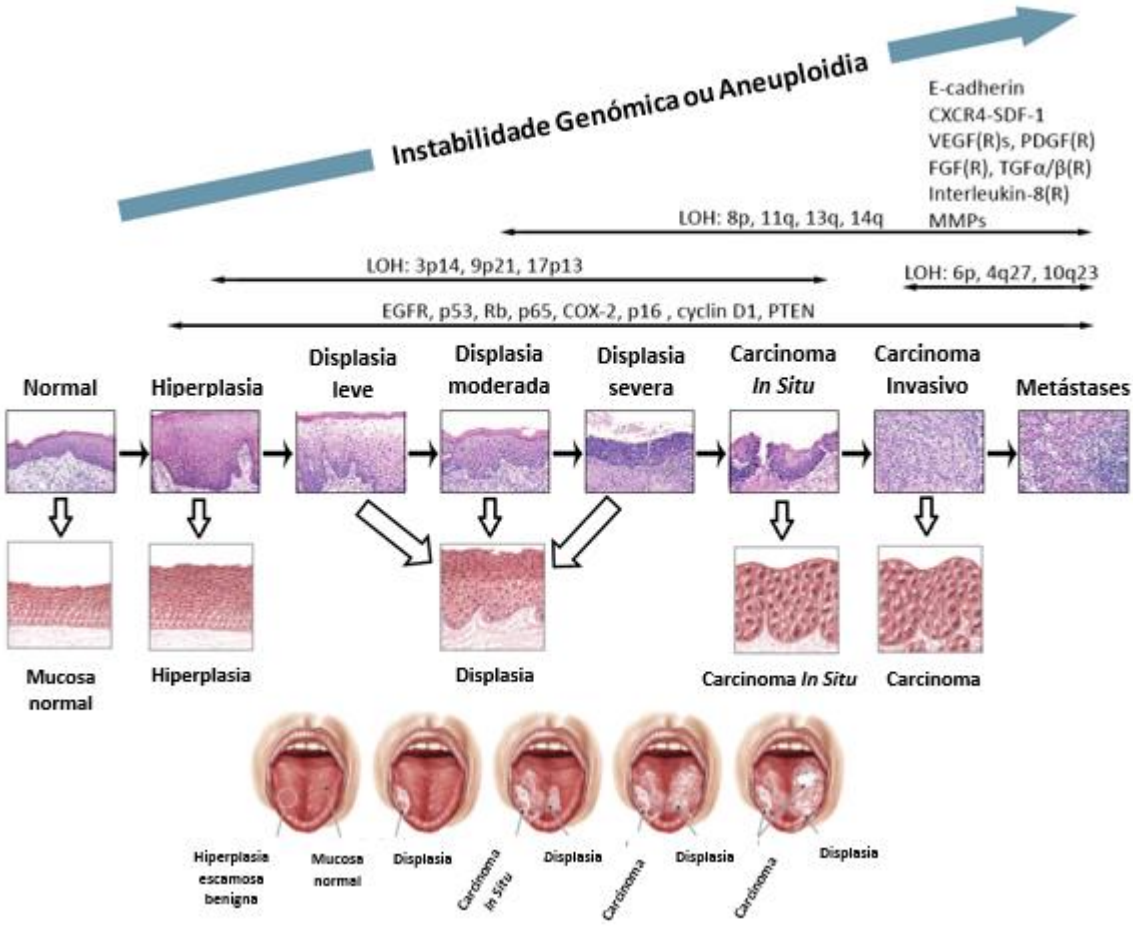


Figura 5 - Modelo da progressão genética da carcinogénese da cabeça e pescoço. O CECP progride através de um processo de várias etapas a partir de caraterísticas histológicas normais à hiperplasia, displasia leve, displasia moderada, displasia severa, carcinoma *in situ*, carcinoma invasivo, e metástases. Vários genes estão envolvidos, principalmente na progressão das metástases e nos estádios iniciais da progressão do tumor. Adaptado de: *Haddad and Shin, 2008* e *Pai and Westra. 2009*.[16, 29].

1.2.8. Alterações nas Vias de Sinalização

As alterações genéticas no CECP estão frequentemente associadas a diferentes vias envolvidas em eventos reguladores fundamentais. As vias supressoras de tumor, incluindo p53 e Rb, NOTCH, PI3KCA (Phosphatidylinositol-4-5-bisphosphate 3-kinase, catalytic subunit alpha) e EGFR, e TGF-β/SMAD têm um papel fundamental na

patogénese da doença. Estes genes e outros genes do CECP desempenham papéis importantes em, pelo menos, quatro vias funcionais principais: proliferação celular (vias p53 e Rb), diferenciação terminal (via NOTCH), sobrevivência celular (vias PI3KCA e EGFR), e adesão e invasão (via TGF-β/SMAD), com muitos dos genes englobados em mais do que uma única via.[50-54] (Figura 6)

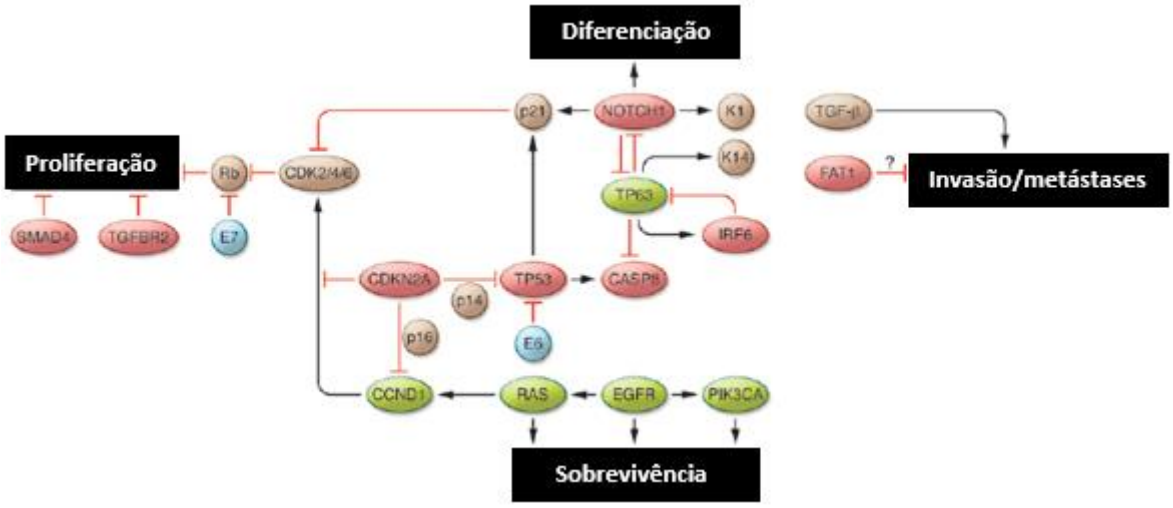


Figura 6 - Vias alteradas na patogénese do CECP identificadas em estudos de sequenciação de todo o exoma. Vermelho: supressores de tumores putativos e estabelecidos; verde: oncogenes; castanho: outros genes/proteínas relevantes; azul: proteínas virais. Adaptado de: *Rothenberg and Ellisen, 2012*.[32]

1.2.9. Alterações Citogenéticas e Genes Alterados no CECP

Alterações citogenéticas demonstraram ser marcadores genéticos, biomarcadores, úteis para o diagnóstico, prognóstico e deteção precoce de doenças malignas, controlo de recidivas e, em última instância, locais de genes específicos em que ocorreram interrupções cromossómicas e que permitem delinear terapias individualizadas.[55-59]

À semelhança de outros tumores, o CECP desenvolve-se como resultado da desregulação de múltiplos genes relacionados com o cancro, ou seja, oncogenes, genes supressores de tumor, e integridade do genoma ou genes de resposta ao dano do DNA (DDR).[59]

As alterações citogenéticas mais comuns no CECP são 3q, 5p, 6p, 7p, 8q, 9q, 11q, 16p, 17p, 17q, 19q, 20q e 11q13 que apresentaram ganhos do número de cópias mais frequentemente, e 2q, 3p, 4q, 5q, 8p, 9p, 11q, 13q, 18q e 21q que apresentaram perdas do número de cópias mais frequentemente.[56, 58-60]

Observou-se um padrão consistente da distribuição de alterações genéticas em termos de perdas e ganhos para alguns cromossomas, em particular para os cromossomas 3, 8 e 11.[57]

Outros dados sugerem também que a perda de 9p e a amplificação de 11q13 podem ser de melhor prognóstico no CECP.[60]

A análise citogenética de amostras do CECP revelou *breakpoints* cromossómicos consistentes, rearranjos cromossómicos estruturais e numéricos.[58] Os genes mais comumente alterados no CECP são divididos em dois grupos principais: oncogenes e genes supressores do tumor.[13, 32] (Tabela 3)

Tabela 3 – Regiões cromossómicas e genes mais comuns envolvidos no CECP, quer em ganho e perda, e sua respetiva função.

Região cromossómica	Gene	Papel na carcinogénese	Tipo de alteração	Ref.
3q	*TP63* (Tumour protein p63)	Sobreexpressão de genes relacionados com cancro	Ganho	(56) (59) (61)
	DCUN1D1 (DCN1, defective in cullin neddylation1, domain containing 1)	Sobreexpressão de genes relacionados com cancro	Ganho	
	SOX2 (Sex determining region Y-box 2)	Sobreexpressão de genes relacionados com cancro	Ganho	
	CCND1 (Cyclin D)	Sobreexpressão de genes relacionados com cancro	Ganho	
	PI3KCA	Sobreexpressão de genes relacionados com cancro Estado avançado no CECP	Ganho	
	ATR (ATR serine/threonine kinase)	Danos do DNA	Ganho	
	MME (Membrane metallo-endopeptidase)	Carcinogénese oral	Ganho	
	BLC6 (B-cell Lymphoma 6)	Carcinogénese oral	Ganho	
	Hs.570518	Carcinogénese oral	Ganho	
	IL12A (Interleukin 12A)	Carcinogénese oral	Ganho	

Região cromossómica	Gene	Papel na carcinogénese	Tipo de alteração	Ref.
3p	*FHIT* (Fragile histidine triad)	Danos do DNA Instabilidade genética Desenvolvimento e progressão do CECP Gene supressor de tumor	Perda	(56) (58) (62)
	RARβ (Retinoic acid receptor beta)	Presente na lesão maligna Carcinogénese	Perda	
7p	*EGFR*	Pior prognóstico Aumento da recidiva local Carcinogénese do CECP	Ganho	(58) (59) (60) (63)
8q	*MYC* (V-myc avian myelocytomatosis viral oncogene homolog)	Pior prognóstico Baixa taxa de sobrevivência Preditor da progressão de displasia a hiperplasia no CECP Tumores localizados no pavimento da boca	Ganho	(56) (57) (59) (64)
	PTK2 (Protein kinase 2)	Capacidade de invasão	Ganho	
	LRP12 (Low density lipoprotein receptor-related protein 12)		Ganho	
	WNT1 (Wingless-type MMTV integration site family, member 1)	Tumores localizados no pavimento bucal	Ganho	
8p	*WHSC1L1* (Wolf-Hirschhorn syndrome candidate 1-like 1)		Ganho	(56) (57) (58) (59)
	CSMD1 (CUB and Sushi multiple domains 1)	Cancros epiteliais Prognóstico pobre Sobrevivência global baixa Curto intervalo sem doença	Perda	
	GATA4 (GATA binding protein 4)	Gene supressor de tumor	Perda	
	MTUS1 (Microtubule associated tumour supressor 1)	Gene supressor de tumor Estado avançado no CE Oral da língua	Perda	
	TUSC3 (Tumour supressor candidate 3)		Perda	

Região cromossómica	Gene	Papel na carcinogénese	Tipo de alteração	Ref.
9q	*NOTCH1*	Gene supressor de tumor ou oncogene	Ganho	(53)
9p	*PTPRD* (protein tyrosine phosphatase, receptor type, D)	Crescimento tumoral	Perda	(31) (56) (59) (65)
	CDKN2B (Cyclin-dependent kinase inhibitor 2B)	Genes supressores de tumor Bloqueia a progressão do ciclo celular Desregulação da proliferação celular	Perda	
	CDKN2A (Cyclin-dependent kinase inhibitor 2A)	Genes supressores de tumor Bloqueia a progressão do ciclo celular Desregulação da proliferação celular	Perda	
	CDKN2C (Cyclin-dependent kinase inhibitor 2C)	Genes supressores de tumor Bloqueia a progressão do ciclo celular Desregulação da proliferação celular	Perda	
	CDKN2D (Cyclin-dependent kinase inhibitor 2D)	Genes supressores de tumor Bloqueia a progressão do ciclo celular Desregulação da proliferação celular	Perda	
	MTAP (Methylthioadenosine phosphorylase)		Perda	
11q	*CCND1*	Prognóstico pobre Recidiva Envolvimento de nódulos linfáticos Reduz a sobrevivência global	Ganho	(56) (58) (59) (65)
	MRE11A (MRE11 homolog A, double strand break repair nuclease)	Genes DDR Biomarcadores para prognóstico	Perda	
	H2AFX (H2A histone family, member X)	Genes DDR Biomarcadores para prognóstico	Perda	
	CHEK1 (Checkpoint kinase 1)	Genes DDR Biomarcadores para prognóstico	Perda	

Região cromossómica	Gene	Papel na carcinogénese	Tipo de alteração	Ref.
11q	*ATM* (ATM serine/threonine kinase)	Genes DDR Biomarcadores para prognóstico Desenvolvimento do CECP Resistência à radioterapia	Perda	(56) (58) (59) (65)
	CASP1 (Caspase 1)	Genes DDR Biomarcadores	Perda	
11q13	*FRA11F* (Fragile site, aphidicolin type, common, fra (11)(q14.2))		Ganho	(17) (58) (59) (65)
	CCND1	Ambiente tumoral agressivo	Ganho	(66)
13q	*RB1* (Retinoblastoma)	Desenvolvimento de tumores Supressão da sua função normal Gene supressor de tumor	Perda	(31) (59) (66)
	ING1 (Inhibitor of growth family, member 1)	Desenvolvimento de tumores Supressão da sua função normal Gene supressor de tumor	Perda	(67) (68)
17p	*TP53* (Tumour protein p53)	"O guardião do genoma" Gene supressor de tumor Pobre prognóstico Diminui taxas de sobrevivência Aumenta risco de recidiva loco-regional	Perda	(13) (17) (56) (63) (69) (70) (71) (72)
	BRCA1 (Breast cancer 1)		Perda	
	CRK (V-crk avian sarcoma vírus CT10 oncogene homolog)		Perda	
18q	*GALR1* (Galanin receptor 1)		Perda	(56) (58) (59) (65) (70)
	PARD6G (Par-6 family cell polarity regulator gamma)	Interfere no ciclo celular	Perda	
	DCC (DCC netrin 1 receptor)	Sobrevivência diminuída	Perda	
	SERPINB13 (Serpin peptidase inhibitor, clade B (ovalbumin) member 13)	Desenvolvimento do CECP	Perda	

1.2.10. Alterações Epigenéticas

Alterações epigenéticas podem ser herdadas e reversíveis, afetando a conformação espacial do DNA e a sua atividade de transcrição, não afetando a estabilidade e integridade do DNA, mas conduzindo a alterações na estrutura da cromatina. Este processo conduz a alterações no fenótipo, sem alterar a sequência de bases do DNA. A regulação epigenética ocorre naturalmente, nomeadamente no desenvolvimento do organismo.[73-78]

Estudos recentes têm mostrado que a regulação epigenética desempenha um papel importante na biologia do cancro, mas também em infeções virais e anomalias do desenvolvimento, por exemplo.[74, 75, 77-79]

Existem vários mecanismos epigenéticos conhecidos (Tabela 4), que incluem a metilação do DNA, modificação de histonas e modificações pós-transcricionais que podem influenciar a expressão génica.[73, 74, 76-79]

A metilação do DNA é um importante regulador da transcrição do gene, (Figura 7) sendo um processo biológico que consiste na adição de um grupo metilo (CH_3) no carbono 5 das bases azotadas citosinas (C). Os alvos principais deste processo são as ilhas de CpG, regiões ricas em dinucleótidos "CG" principalmente presentes na região promotora de determinados genes. A presença de metilação aumentada na região promotora dos genes está associada com a supressão da expressão do gene.[73-75, 77-80]

Este processo envolve uma reação enzimática catalisada por uma família de enzimas denominadas DNA metiltransferases (DNMTs). Três subtipos de DNMT foram identificados: DNMT1, DNMT3A e DNMT3B. Enquanto DNMT1 é responsável por manter a metilação padrão nas células filhas durante a mitose ou meiose, DNMT3A e DNMT3B estão envolvidos no processo de metilação *de novo*.[74, 77, 78]

Uma vez que, existem mais de 28 milhões de locais CpG no genoma humano, será necessário avaliar o estado de metilação de cada um desses locais para compreender inteiramente o papel da metilação do DNA na saúde e na doença.[79]

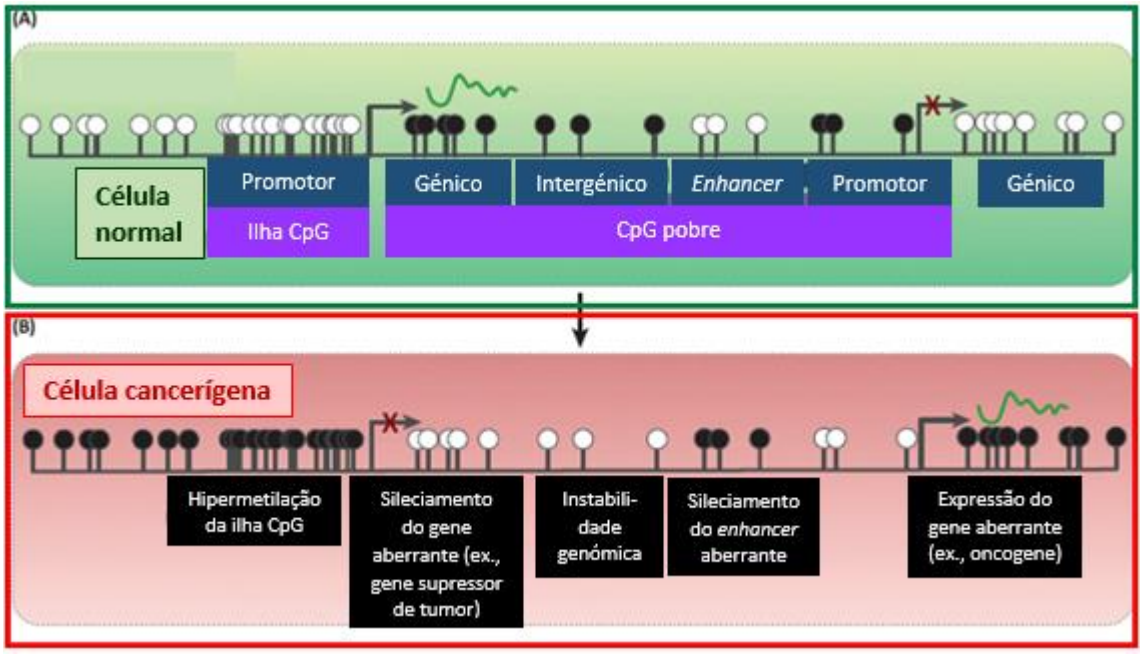


Figura 7 - Metilação do DNA e (des)regulação do genoma - representação esquemática das principais alterações que ocorrem nas células cancerígenas. (A) (Retângulo verde) Ilhas CpG estão frequentemente associadas a promotores de genes e são resistentes à metilação do DNA. (B) (Retângulo vermelho) Nas células cancerígenas, as ilhas CpG são propensas à hipermetilação do DNA, a qual resulta no silenciamento do gene aberrante. Legenda: Círculo branco, CpG não metilado; círculo preto, CpG metilado. Adaptado de: *Stirzaker et al., 2014*.[79]

Tabela 4 - As alterações epigenéticas mais comuns. Adaptado de: *Mascolo et al., 2012*.[76]

Alteração epigenética	**Mecanismo putativo**	**Consequência biológica**
Hipometilação do DNA	Ativação dos oncogenes celulares	Aumento da proliferação, vantagem de crescimento
	Ativação de elemento transponível	Instabilidade genómica, ruído de transcrição
Hipermetilação do DNA	Hipermetilação *de novo* das ilhas CpG dentro dos promotores de genes que levam ao silenciamento dos genes supressores de tumores e associados ao cancro	Instabilidade genómica e cromossómica, aumento da proliferação, vantagem do crescimento
Perda de *Imprinting* (LOI – do Inglês *Loss of Imprinting*)	Reativação de alelos silenciosos, expressão bialélica de genes *imprinted*	Expansão da população de célula precursora
Inativação do cromossoma X	Mecanismos são desconhecidos, mas parece estar relacionado com a idade	Dosagem do gene alterado, vantagem do crescimento

Alteração epigenética	Mecanismo putativo	Consequência biológica
Acetilação de histonas	Ganho da função	Ativação dos genes promotores de tumores
	Perda da função	Defeitos na reparação do DNA e *checkpoints*/pontos de verificação
Desacetilação de histonas	Silenciamento dos genes supressores de tumores	Instabilidade genómica, aumento da proliferação
Metilação de histonas	Perda de padrões hereditários da expressão génica ("memória celular")	Instabilidade genómica, vantagem do crescimento
Amplificação de microRNAs (miRNAs) no cancro	Funcionam como oncogenes	Transformação neoplásica
Deleção de miRNAs no cancro	Funcionam como supressores de tumores	Transformação neoplásica

1.3. Ferramentas Bioinformáticas na análise do CECP

Os sistemas de *Data Mining* podem ser classificados de muitas formas distintas dependendo dos algoritmos que cada aproximação usa, uma vez que certos algoritmos de *Data Mining* são mais apropriados para determinados tipos de problemas do que outros. Contudo, atualmente, ainda não existem padrões bem definidos e generalizadamente aceites.[81-83] Desta forma, numa dada área procuram-se sistemas de *Data Mining* que cubram as suas necessidades e, assim, o primeiro requisito que se coloca é que os resultados obtidos de um estudo de *Data Mining* sejam compreensíveis, o segundo requisito é o de que os sistemas tenham bons desempenhos, de modo a que os modelos possam ser construídos em tempo adequado, e por último, os estudos de *Data Mining* forneçam resultados precisos.[81, 82, 84]

1.3.1. Aprendizagem Indutiva

A indução pode ser encarada como a construção de conhecimento a partir dos dados. A aprendizagem indutiva é o método de construção de um modelo em que a base de dados é analisada na pesquisa de tendências e padrões. Ou seja, objetos com características semelhantes são agrupados em classes podendo ser formuladas regras que permitam prever a classe dos objetos que venham a ser analisados futuramente. Há que ter em atenção que geralmente a base de dados é dinâmica, logo o modelo deve ter a capacidade de se adaptar, isto é, deve ter a capacidade de aprender e integrar.[82-85]

A aprendizagem indutiva é constituída por duas abordagens diferentes, ou seja, pode ser realizada através da aprendizagem supervisionada e aprendizagem não supervisionada.
A indução significa, assim, a extração de padrões. E a qualidade do modelo produzido por métodos de aprendizagem indutiva é tal que pode ser utilizado para prever resultados de situações futuras.[82-85]

1.3.1.1. Aprendizagem Supervisionada

A aprendizagem supervisionada é feita a partir de exemplos. O analista ajuda o sistema a criar o modelo, através da definição das classes e dos exemplos em cada classe. O sistema tem que determinar a descrição para cada classe, isto é, o conjunto de propriedades comuns nos exemplos que lhe são fornecidos. Posteriormente à descrição determinada, é possível formular uma regra de classificação que pode ser utilizada para prever a classe de um objeto que não tenha sido considerado aquando da aprendizagem.[81, 86]
Assim, os problemas que envolvem aprendizagem supervisionada são geralmente conhecidos como problemas de Classificação.

1.3.1.2. Aprendizagem Não supervisionada

A aprendizagem não supervisionada é efetuada com base em observação e descoberta. Na aprendizagem não supervisionada não são definidas classes à priori, pelo que o sistema de *Data Mining* necessita de observar os exemplos e reconhecer os padrões por si próprio, ou seja, tem de descobrir sozinho relações, padrões, regularidades ou categorias nos dados que lhe vão sendo apresentados. Assim resulta um conjunto de descrições de classes, uma para cada classe descoberta na base de dados.[81, 86]
Neste caso, os problemas associados à aprendizagem não supervisionada são geralmente referidos como problemas envolvendo Clusterização.

1.3.2. Classificação

Data Mining significa extração de dados, a qual faz uso de ferramentas informáticas baseadas em modelos matemáticos e estatísticos, que recorrem às séries registadas nas bases de dados com o intuito de identificar eventuais correlações e padrões consistentes e/ou relacionamentos sistemáticos entre as variáveis. Posteriormente, aplica os padrões detetados a novos dados.[85, 87, 88]

O processo de *data mining* consiste em três etapas, sendo elas: a exploração; a construção do modelo ou definição do padrão a identificar; e a validação/verificação.[85, 87, 88]

Como já referido acima, a classificação é definida como o método pelo qual se obtém: regras que permitem a distinção de objetos provenientes de classes distintas. Por sua vez, a clusterização define-se como uma técnica que agrupa os dados em conjuntos por semelhança de propriedades dos seus elementos.[85, 87, 88] Relativamente à classificação, este processo baseia-se no processo "classifica, constrói, testa" (Figura 8).

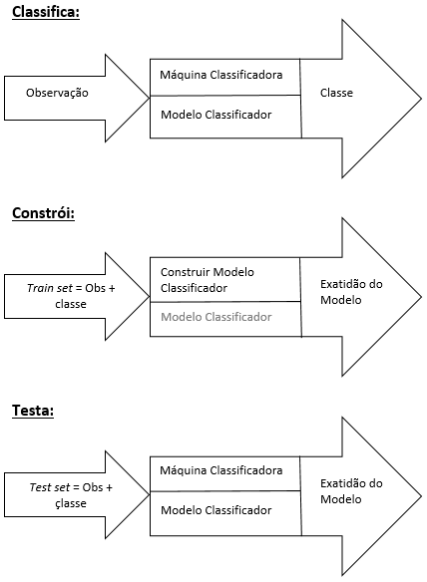


Figura 8 - Processo "classifica, constrói, testa" de Classificação.

A classificação é então um processo de dois passos, sendo eles[85, 87, 88]:

- Construção do modelo, onde se faz a descrição de um conjunto de classes pré-determinadas, sendo que cada observação pertence a uma classe pré-definida, determinado pelo atributo de classe; o conjunto de treino é o conjunto de observações usadas para treino; e o modelo estatístico de classificação é representado como regras de decisão, árvores de decisão, ou fórmulas matemáticas que possibilitam a atribuição de uma classe a uma dada observação;
- Utilização do modelo para classificação futura ou objetos desconhecidos, estima-se a exatidão do modelo, onde a classificação dos elementos do conjunto de teste é comparada com a sua classe conhecida e a exatidão é a percentagem de elementos de teste que são classificadas corretamente pelo modelo. O conjunto de teste tem de ser independente do conjunto de treino; e se a exatidão for

aceitável, utiliza-se o modelo para classificar novas observações cuja classe é desconhecida.

1.3.2.1. Classificadores

1.3.2.1.1. Random Forest

A técnica Random Forest examina um grande conjunto de árvores de decisão. Essa análise é feita através da primeira geração de uma amostra aleatória dos dados originais com *bootstrapping* (substituição), e utilizando um número definido pelo utilizador de variáveis selecionadas aleatoriamente de todas as variáveis para determinar a divisão em cada nó. Vários subconjuntos de árvores são construídos, e o papel de cada variável em cada decisão é tido em conta.[89, 90]

O classificador Random Forest apresenta várias características cruciais fazendo-o ser um classificador muito usado na programação, pois[90]:

- é um dos melhores na exatidão entre os algoritmos atuais;
- é executado de forma eficiente em grandes bases de dados;
- pode lidar com milhares de variáveis de entrada, sem exclusão de variáveis;
- pode estimar a importância das variáveis na classificação;
- gera uma estimativa imparcial interna do erro de generalização com a construção do classificador;
- é um método razoavelmente eficaz para estimar os dados em falta e manter a exatidão quando uma grande proporção dos dados estão em falta;
- tem métodos para equilibrar o erro em conjuntos de dados desequilibrados da população da classe;
- pode guardar resultados gerados para uso futuro em outros dados;
- são computados protótipos que dão informações sobre a relação entre as variáveis e a classificação;
- calcula proximidades entre pares de casos que podem ser usados na clusterização, *outliers* da localização, ou (por dimensionamento) permitir uma visualização interessante dos dados;
- permite a extensão das capacidades anteriores para dados não classificados, levando à clusterização, visualizações de dados e deteção de *outliers*;
- oferece um método para detetar interações de variáveis.

Numa outra perspetiva, em concordância com este classificador é usada uma técnica, o *importance plot*, que permite a redução de variáveis de forma fidedigna.

O *importance plot* é um método que indica a importância de cada variável para o modelo de classificação. Para cada variável usada na classificação é obtido um score que indica a sua importância na classificação dos dados. Geralmente este resultado é traduzido num gráfico, de onde deriva o nome. O gráfico mostra cada variável sobre o eixo y, e a sua importância no eixo x (Figura 9). As variáveis são ordenadas de maior para menor importância. Assim, as variáveis mais importantes estão no topo e uma estimativa da sua importância é dada pela posição do ponto sobre o eixo x. E é dessa forma que se deve usar as variáveis mais importantes, como determinado a partir da variável *importance plot* e *principal componente analysis* (PCA), por exemplo.[89, 91]

Normalmente, deve-se olhar para um grande intervalo entre as variáveis para decidir quantas variáveis importantes escolher. Esta é uma ferramenta importante para reduzir o número de variáveis para outras técnicas de análise de dados, como neste caso o Random Forest, mas também é necessário ter cuidado para não ficar com poucas variáveis (o que não irá distinguir os dados) ou muitas variáveis (o que irá explicar muito pormenorizadamente as diferenças e também devido à ocorrência de *overfitting*).[89, 91]

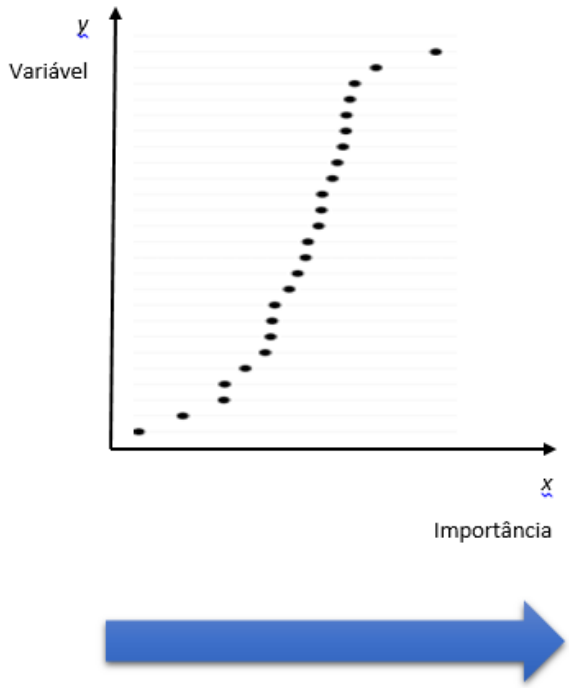


Figura 9 - Gráfico referente à variável importance plot de forma a reduzir o número de variáveis. Adaptado de: *Metagenomics. Statistics.*[89].

1.3.2.1.2. SVM

O *support vector machine*, SVM, é um classificador que transforma os dados num espaço de alta dimensão através de mapeamento não linear. Nesse espaço, o SVM procura o melhor hiperplano de separação.[92]

Isto é, sejam os dados $(x_1, y_1), \ \ldots, (x_{|D|}, y_{|D|})$, e sendo x_i o conjunto de treino associado com a classe y_i, existem infinitas superfícies, designadas de hiperplanos, que separam as duas classes. Contudo desejamos encontrar o melhor hiperplano, isto é, a linha que minimiza o erro de classificação sobre os dados novos.[92]

Desta forma, o SVM tem o intuito de procurar o hiperplano com a margem maior, isto é, *maximum marginal hyperplane* (MMH) (Figura 10).

Assim, o hiperplano de separação é definido como

$$w \bullet x + b = 0$$

onde $w = \{w_1, w_2, \ \ldots, w_n\}$ é um vetor de pesos e b é um escalar.[92]

Em $2D$ pode ser reescrito como:

$$w_0 + w_1 x_1 + w_2 x_2 = 0.$$

E os pontos acima da linha têm:

$$w_0 + w_1 x_1 + w_2 x_2 > 0,$$

e os pontos abaixo da linha têm:

$$w_0 + w_1 x_1 + w_2 x_2 < 0.$$

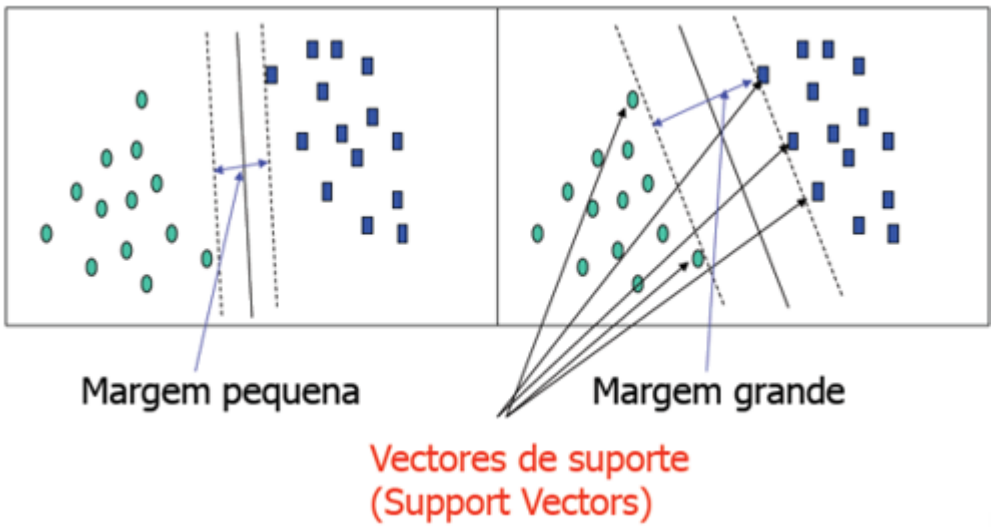


Figura 10 - Representação esquemática do hiperplano pretendido no classificador SVM.

1.3.3. Regressão Logística

A regressão logística é um classificador bem conhecido com propriedades bem determinadas, constituindo por isso um dos classificadores mais usados.

A análise estatística de dados necessita de bastantes ferramentas estatísticas, e quando se pretende modelar relações entre variáveis, uma das mais importantes são os modelos de regressão. O intuito principal dos modelos de regressão é explorar a relação entre uma ou mais variáveis independentes e uma variável dependente.[93]

Um dos casos específicos dos modelos lineares generalizados são os modelos onde a variável dependente apresenta apenas duas hipóteses lógicas ou que de alguma forma foi dicotomizada assumindo valores 0 ou 1, e dessa forma, o modelo de regressão logística é o modelo mais usado e familiar.[93, 94]

A regressão logística é uma técnica estatística que tem como objetivo modelar, tendo como base um conjunto de observações, a relação "logística" entre uma variável dependente dicotómica e uma série de variáveis independentes (continuas, discretas) e/ou categóricas.[93, 94]

Na regressão logística, a variável dependente é dicotómica, atribuindo-se geralmente o valor 1 ao acontecimento de interesse (ou seja, sucesso) e 0 ao acontecimento complementar (isto é, insucesso).[95]

Qualquer que seja a regressão, a quantidade crucial é o valor médio da variável dependente dado o valor da variável independente. Este valor médio é designado como

valor médio condicional e é expresso como $E[Y/X]$ onde Y representa a variável dependente e X a variável independente. A quantidade $E[Y/X]$ é definida como "valor esperado de Y dado $X = x$".(95)

No modelo de regressão linear admitimos que o valor médio condicional pode ser expresso como uma equação linear em x:

$$E[Y/X = x] = \beta_0 + \beta_1 x$$

e repare-se que $E[Y/X = x]$ pode tomar valores entre $-\infty$ e $+\infty$.(95)

No modelo de regressão logística a variável dependente toma dois valores distintos, admitindo-se que:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}},$$

em que $\pi(x) = E[Y/X = x]$.

Posteriormente, uma transformação fundamental no estudo de modelos de regressão logística é a transformação *logit* que tem por objetivo linearizar o modelo. Essa transformação é:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \Leftrightarrow$$

$$\Leftrightarrow g(x) = ln\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}}{1 - \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}}\right) = ln\left(\frac{\frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}}{\frac{1}{1 + e^{\beta_0+\beta_1 x}}}\right) \Leftrightarrow$$

$$\Leftrightarrow g(x) = \ln\left(e^{\beta_0+\beta_1 x}\right) = \beta_0 + \beta_1 x.$$

O modelo com esta transformação possui as seguintes propriedades do modelo de regressão linear: a função *logit* é linear nos parâmetros, pode ser contínua e os seus valores variam em $\mathbb{R}$. E esta transformação é chamada *Transformação logit* de $\pi(x)$ e a razão $\frac{\pi(x)}{1-\pi(x)}$ é designada *Odds*.(95)

1.3.4. Análise de dados

1.3.4.1. Precisão, exatidão, sensibilidade e especificidade

A precisão e a interpretabilidade, isto é, a compreensão e informação dadas pelo modelo são formas de avaliar os métodos de classificação.

A precisão (ou rigor) traduz quão bem determinado foi um resultado, sem o relacionar com o verdadeiro valor da grandeza. Em geral, um método é considerado preciso se uma grande quantidade de análises são repetidas na mesma amostra e facultam resultados

semelhantes. Neste caso, a variação aleatória é considerada pequena e o método mostra-se confiável, uma vez que os resultados podem ser reproduzidos (Figura 11). Ou seja, a precisão é boa (ou o resultado rigoroso), quando os erros acidentais são pequenos.[85, 88]
Por sua vez, exatidão (ou fidelidade) avalia quão perto o resultado está do verdadeiro valor. Assim, um procedimento é considerado exato quando o valor do resultado se aproxima do valor absoluto verdadeiro para a amostra em questão. Esses resultados são comparados com valores conhecidos. A exatidão é grande (ou o resultado fiel) se os erros sistemáticos são pequenos (Figura 11).
Por outro lado, sensibilidade é a capacidade de um teste identificar corretamente as pessoas que têm a doença ou condição clínica. Por exemplo, considerando um exame com 90% de sensibilidade significa que se 100 pessoas possuem uma determinada doença, o teste deverá identifica-la corretamente em 90. As outras 10 pessoas não irão mostrar o resultado esperado para o teste. Para estes 10% restantes a constatação dita normal será falsa, designada de resultado falso-negativo. Em suma, quanto mais sensível for um teste, menos resultados falso-negativos serão produzidos. Matematicamente, sensibilidade é a percentagem de objetos classificados corretamente como *X* de todos os *X* (Figura 12).[85, 88]
Por último, especificidade é a capacidade de um teste excluir corretamente pessoas que não têm uma doença ou condição clínica. Baseando no exemplo da sensibilidade, tendo um exame com 90% de especificidade significa que se 100 pessoas saudáveis são testadas, apenas em 90 será encontrado o resultado normal. As restantes terão resultado positivo para o teste. Para essas 10 pessoas a constatação considerada anormal será falsa, designada de resultado falso-positivo.

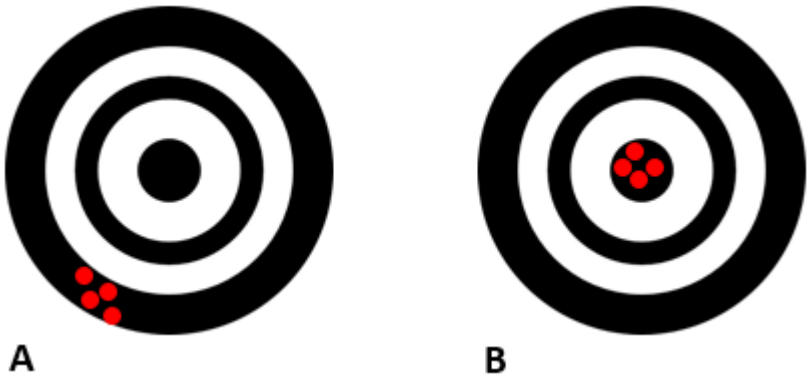


Figura 11 - Representação esquemática da distinção entre precisão e exatidão. **A.** Precisão sem exatidão. **B.** Exatidão com precisão.

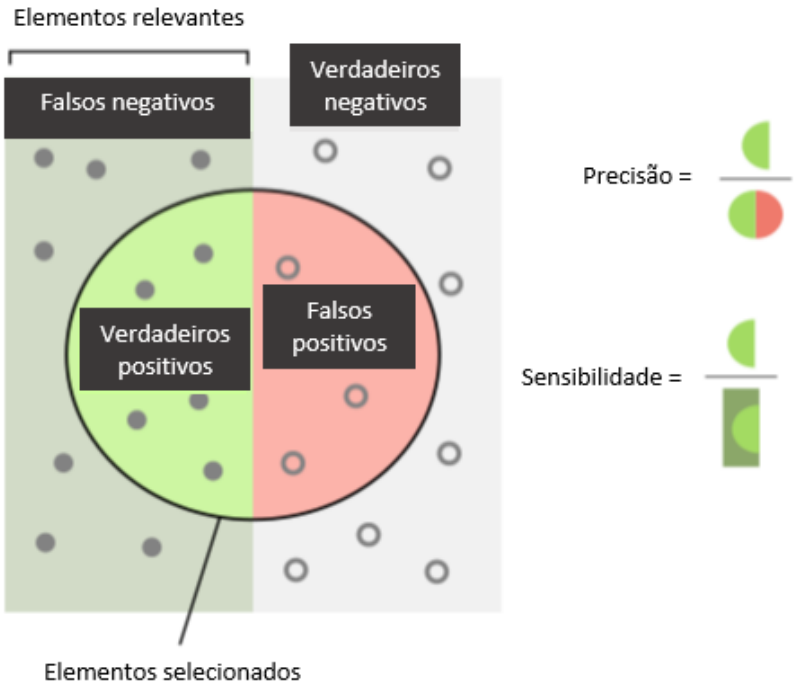


Figura 12 - Esquema representativo da precisão e sensibilidade. As bolas representam todos elementos pertencentes à amostra, ou seja, as pessoas; as bolas com preenchimento cinzento correspondem às pessoas com doença ou condição clínica; as bolas com delimitação cinzenta correspondem às pessoas saudáveis.

1.3.4.2. Matriz de confusão

A matriz de confusão dos resultados associados ao modelo, oferece uma medida efetiva do modelo de classificação, ao mostrar o número de classificações corretas *versus* as classificações preditas para cada classe, sobre um conjunto de exemplos.

Como se pode verificar na Tabela 5, o número de verdadeiros positivos para cada classe localiza-se na diagonal principal da matriz, os elementos restantes da matriz para os casos em que a coluna é diferente da linha representam erros na classificação, e a matriz de confusão de um classificador ideal possui todos esses elementos iguais a zero uma vez que nesse caso o classificador não comete erros.[85, 88]

Tabela 5 - Esquematização da matriz de confusão.

		Classificados como…				
		Classe 1	Classe 2	…	Classe n	
Atual…	Classe 1	Verdeiro Positivo	(C1, C2)	…	(C1, Cn)	Totais de C1
	Classe 2	(C2, C1)	Verdeiro Positivo	…	(C2, Cn)	Totais de C2
	…	…	…	Verdeiro Positivo	…	
	Classe n	(Cn, C1)	(Cn, C2)	…	Verdeiro Positivo	
	Totais de classificação					Totais de Cn

Então, a determinação da exatidão segunda a matriz de confusão é[85, 88]:

$$Exatidão = \frac{soma\ das\ diagonais}{soma\ de\ todas\ as\ observações},$$

Já a exatidão da classe x é[85, 88]:

$$Exatidão\ da\ classe\ x = \frac{Verdadeiro\ Positivo\ da\ coluna}{soma\ da\ coluna} = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo\ + Falsos\ Positivos},$$

E a sensibilidade da classe x é[85, 88]:

$$Sensibilidade\ da\ classe\ x = \frac{Verdadeiro\ Positivo\ da\ linha}{soma\ da\ linha} = \frac{Verdadeiro\ Positivo}{Verdadeiro\ Positivo\ + Falsos\ Negativos}.$$

1.3.4.3. Curvas ROC

As curvas ROC (*Receiver operating characteristic*) são bastante importantes para a classificação uma vez que permitem avaliar o modelo de forma independente englobando medidas como a sensibilidade, a especificidade e a exatidão.

Optando como exemplo ter doença ou não ter doença, a sensibilidade, como já referido anteriormente, é a capacidade que um teste tem de discriminar, de entre os suspeitos de uma patologia, aqueles efetivamente doentes. Ou como definido por Galen e Gambino (1975), a sensibilidade é "a positividade na doença". Por sua vez, a especificidade, também já mencionado, é a capacidade que o mesmo teste tem de ser negativo, em face

de uma amostra de indivíduos que não têm a doença em questão e, como definido por Galen e Gambino (1975), é "a negatividade da doença".[96, 97] A sensibilidade e a especificidade são duas características difíceis de conciliar, isto é, é extremamente complicado aumentar a sensibilidade e a especificidade de um teste em simultâneo.[96]
As curvas ROC são uma forma de expor a relação, geralmente antagónica, entre a sensibilidade e a especificidade de um teste diagnóstico quantitativo.[96]
As curvas ROC são construídas com base na sensibilidade e especificidade de um sistema que classifique os dados em duas classes mutuamente exclusivas para todos os pontos de corte possíveis na variável de decisão.[98]
As curvas ROC descrevem a capacidade discriminativa de um teste diagnóstico para um determinado número de valores e assim permite colocar em evidência os valores para os quais existe maior otimização da sensibilidade em função da especificidade.[96]
Por outro lado, as curvas ROC permitem quantificar a exatidão de um teste diagnóstico, já que, esta é proporcional à área sob a curva ROC.[96]
A área abaixo da curva ROC - *area under curve* (AUC) - é a medida de avaliação mais popular na análise ROC representando o equilíbrio entre sensibilidade e especificidade, e este valor varia entre 0,5 e 1, sendo que quanto mais próximo de 1 melhor é o poder de discriminar entre as duas observações. No contexto da análise da expressão diferencial a AUC pode ser interpretada como uma medida que avalia a distância entre as distribuições correspondentes aos dois grupos em análise.[98]

1.3.4.4. Valor p

Outro parâmetro muito importante também aquando da classificação é o valor de p.
Sabe-se que quando se utilizam amostras pequenas num estudo, os resultados podem ser influenciados por fatores aleatórios. Assim, é fundamental quando se comparam dois grupos saber se as diferenças encontradas nestes se devem ou não ao acaso e uma forma plausível de se fazer é conseguir fórmulas matemáticas, cujo resultado forneça uma ideia da probabilidade dos valores resultantes serem devido ao acaso. Essa probabilidade é então traduzida pelo valor de p.[99]
O valor de p é uma estimativa da probabilidade de se conseguir resultados iguais ou mais extremos, ou seja, com maiores diferenças entre os grupos, partindo da hipótese que não existem diferenças entre os grupos em análise relativamente à variável em estudo. Esta hipótese é definida em termos estatísticos por hipótese nula, isto é, de

forma simplista é a probabilidade de obter esse resultado se na realidade não existem discrepâncias entre os grupos em questão. Na verdade, o valor de p informa a probabilidade que uma associação seja um falso-positivo decorrente do acaso.[99]

Como o valor de p traduz uma probabilidade, este valor varia entre 0 e 1. Sendo que quanto mais próximo de 0, maior a probabilidade da diferença encontrada entre os grupos não ser causada por fatores aleatórios. No entanto, inicialmente é fundamental saber a partir de que valor de p se deve desprezar erros aleatórios, eliminando a hipótese nula. O valor de p abaixo do qual se rejeita erros aleatórios é designado por nível de significância.[99]

Unanimemente, considera-se que a probabilidade dos resultados advirem de erros aleatórios é razoavelmente baixa para ser menosprezada quando o valor de p é inferior a 0,05. Por conseguinte, os resultados são considerados como "estatisticamente significativos". Por sua vez, quando o valor de p é superior a 0,05 os resultados são ponderados como "estatisticamente não significativos". Contudo, não é correto avaliar o valor de p consoante esteja abaixo ou acima de 0,05, mas deve-se indicar sempre o valor exato do valor de p.[99]

Quando um resultado é considerado "estatisticamente significativo" não significa que o resultado seja correto, este apenas indica que a probabilidade da diferença encontrada entre os grupos ser devida a fatores aleatórios é mínima e deste modo pode ser desprezada. Uma vez que, qualquer resultado pode ser também afetado por erros sistemáticos, os quais não são avaliados por valor de p.[99]

Regra geral, o valor de p é tanto menor quanto maiores as diferenças entre os grupos e quanto maior o tamanho da amostra. E dessa forma, resultados estatisticamente significativos podem não ter ainda assim relevância clínica, visto que depende de uma avaliação subjetiva por parte de múltiplos outros fatores.[99]

1.3.4.5. Análise de Sobrevivência

A análise de sobrevivência tem como variável dependente o tempo até à ocorrência de determinado acontecimento. Ou seja, a análise de sobrevivência equipara a rapidez com que os intervenientes desenvolvem determinado acontecimento ao fim de determinado período de tempo. Na análise de sobrevivência, o acontecimento final pode ser a morte, ou qualquer outro acontecimento, tal como recidiva, progressão da doença, efeito

lateral, mudança de estado. Assim, o termo " tempo de sobrevida" na análise de sobrevivência refere-se ao tempo até ao acontecimento em questão.[100]

A vantagem mais relevante da análise de sobrevivência é a de utilizar a informação de todos os intervenientes até ao momento em que desenvolvem o acontecimento ou são censurados - os intervenientes não desenvolveram o acontecimento até ao fim da observação no estudo. Deste modo, na análise de respostas binárias, ocorrer ou não um acontecimento, em estudos longitudinais que se descrevem por tempo de seguimento diferente entre os intervenientes e perdas de *follow-up*, a análise de sobrevivência é a técnica eleita. A análise de sobrevivência também avalia o ritmo a que os acontecimentos decorrem ao longo do tempo nos diferentes grupos em estudo.[100]

A análise dos dados na análise de sobrevivência pode ser realizada por dois métodos: Actuarial e Kaplan-Meier. Especificamente, o método de Kaplan-Meier ou Curva de Sobrevivência é o mais usado atualmente uma vez que é usada a data exata do acontecimento e por consequente os resultados são mais precisos. O método de Kaplan-Meier consta em dividir o tempo de seguimento em intervalos, onde os limites dizem respeito ao tempo de seguimento em que houve acontecimentos.[100]

O Hazard Ratio (HR) é a medida de associação para comparar grupos utilizada na análise de sobrevivência. HR é a probabilidade de algum interveniente que não teve o acontecimento até determinado momento, tê-lo nesse mesmo instante, ou seja, compara a incidência instantânea com que os acontecimentos ocorrem nos diferentes grupos. Outra forma de comparar é através da sobrevida mediana – tempo ao fim do qual 50% dos intervenientes atingem o evento de interesse.[100]

1.4. Técnicas Laboratoriais

1.4.1. Hibridização Genómica Comparativa com Array (aCGH)

A técnica aCGH, ou em Inglês *Array Comparative Genomic Hybridization*, também é designada por cariótipo molecular.

O aCGH foi criado para o *screening* de elevada resolução do genoma humano, com o intuito de identificar desequilíbrios no número de cópias do DNA.[101] Desta forma, o aCGH pode avaliar todo o genoma ou ser direcionado para regiões específicas.

Nesta técnica, o DNA genómico marcado, quer seja refente à amostra em estudo ou à amostra controlo, tem iguais quantidades que são co-hibridizadas num *array* contendo sondas de DNA. Estas sondas de DNA estão num suporte físico, sendo este uma lâmina

de vidro, e podem ser fragmentos de DNA genómico humano clonados a partir de bactérias (BAC – *Bacterial Artificial Chromosome*) ou de PAC (P1 – *derived Artificial Chromosomes*), com um tamanho de cerca de 75-200 kb, ou ainda oligonucleotídeos sintéticos, 25-85 mers.[102] O espaçamento e comprimento das sondas de DNA determinam a resolução genómica das diferentes plataformas de aCGH. Ou seja, nas plataformas de aCGH para todo o genoma os alvos estão espaçados com uma cobertura de cerca de um clone por megabase a um clone por 100 Kb[103], já nos *arrays* comerciais de oligonucleótidos variam de uma sonda por 6 kb a uma sonda por 70 Kb.

O protocolo e a plataforma selecionados influenciam a quantidade de DNA utilizado e de regiões estudadas.

A análise do número de cópias com elevada resolução através do uso de *arrays* de oligonucleótidos foi descrita inicialmente por *Lucito et al.* em 2013[104] com o objetivo de detetar tanto amplificações como deleções.

Relativamente ao procedimento esta técnica consiste na marcação do DNA genómico do caso em estudo e da amostra controlo com flurocromos diferentes, nomeadamente *Cyanine 5* (Cy5) e *Cyanine 3* (Cy3). As sondas estão fixadas em lâminas de vidro, às quais é co-hibridizado o DNA genómico do caso em estudo e do controlo, antecipadamente marcado, sendo as lâminas posteriormente lidas num *scanner* para avaliar a intensidade da fluorescência, sendo quantificada a partir de *software* específico para a análise do número de cópias. Para cada locus, o rácio resultante da intensidade de fluorescência entre a amostra em estudo e de controlo é proporcional ao rácio do número de cópias do DNA em estudo comparativamente ao controlo. Desta forma, um rácio Cy5:Cy3 alterado é indicador de perda ou ganho no DNA em estudo (Figura 13).

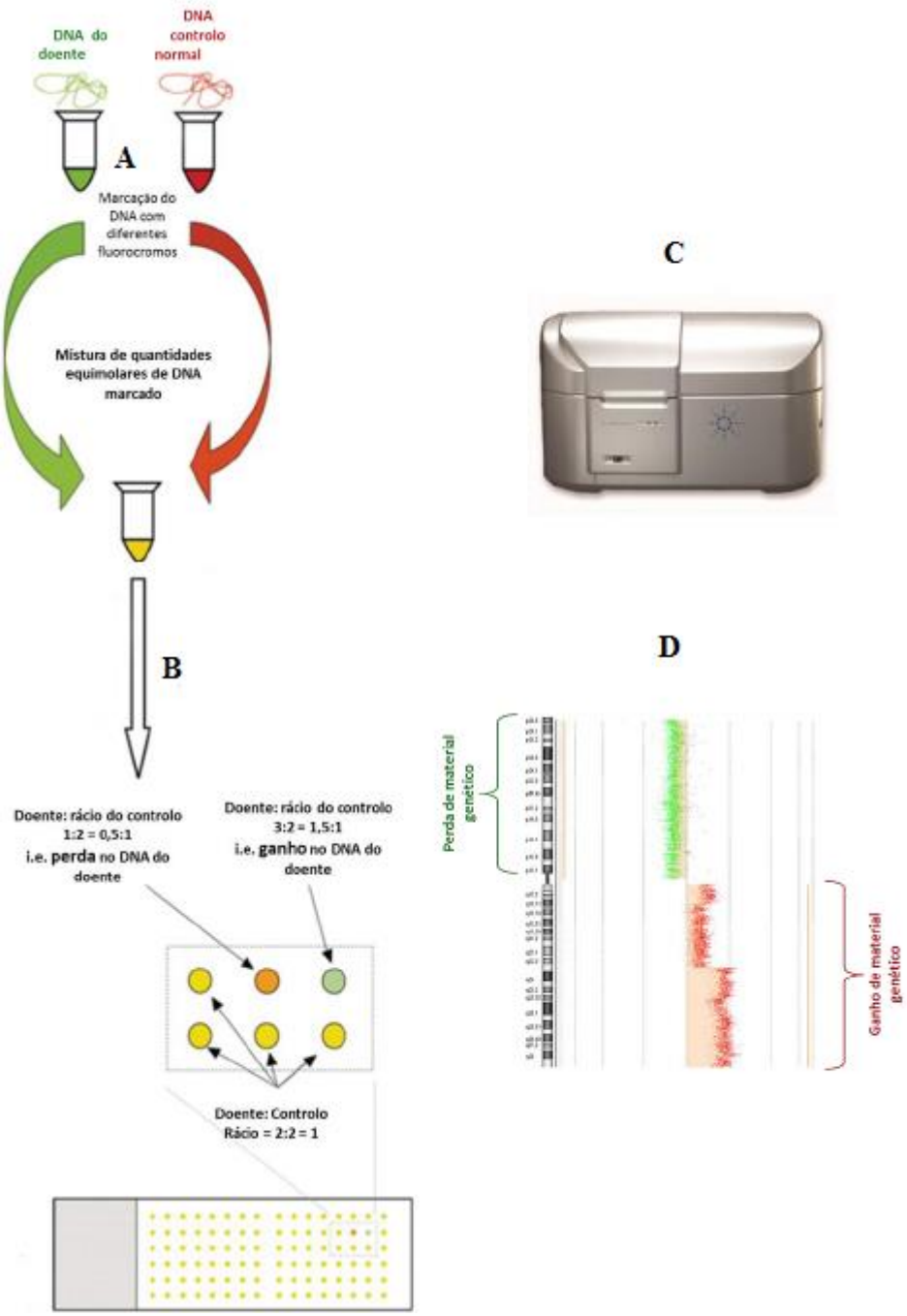


Figura 13 - Representação do aCGH. Esta técnica consiste em quatro etapas: A – marcação do DNA genómico com diferentes fluorocromos, Cy3 e Cy5; B – hibridação na plataforma de *microarray*; C – leitura dos sinais pelo *scanner*; D – processamento dos dados pelo *software* e análise dos resultados. Adaptado de *Sharkey et al, 200*5[105] e imagem cedida pelo Laboratório de Citogenética e Genómica – Faculdade de Medicina da Universidade de Coimbra.

O desenvolvimento de tumores tem como base processos combinados de instabilidade genética e seleção, emergindo na expansão clonal de células que acumularam o conjunto de alterações genéticas mais vantajosas. Então, os tumores contêm uma história genética do seu desenvolvimento, embora essa história possa ser difícil de interpretar.

Algumas alterações que são cruciais no desenvolvimento inicial do tumor podem ser perdidas ou omitidas por eventos subsequentes. Por sua vez outras alterações podem ser neutras ou mesmo prejudiciais para o tumor, mas identificadas, visto que estão presentes nas células que desenvolvem uma alteração suficientemente pró-tumorigénica, ou porque são um produto de um evento que produz uma alteração fundamental.

O aCGH tem a capacidade de analisar DNA de uma grande variedade de amostras com elevada resolução e desse modo providencia um excelente ponto de partida para o estudo em cancro. Esta tecnologia tem potencial para ser usada na classificação tumoral e para predizer a progressão e prognóstico tumoral. No entanto, a sua aplicação para o prognóstico é relativamente reduzida tendo como base a incapacidade de deteção de alterações equilibradas.[106]

Primeiramente, o aCGH foi desenvolvido como uma ferramenta de investigação de desequilíbrios genómicos no cancro, contudo, atualmente tem-se tornado essencial como uma ferramenta de rotina no diagnóstico, e gradualmente está a substituir as metodologias de citogenética clássica em vários laboratórios de genética.[107] Porém, uma vez que esta técnica não permite visualizar alterações equilibradas nem variações de ploidia, o mais correto será conciliar os diferentes métodos disponíveis no laboratório de forma a responder com máximo rigor às questões científicas colocadas.

2. FUNDAMENTAÇÃO E OBJETIVOS

Este estudo experimental tem como intuito principal responder à seguinte questão: será que as diferentes taxas de sobrevivência dos doentes de CECP estão associadas à presença de tumores com diferentes assinaturas genómicas? Ou seja, pretende-se verificar a existência de pelo menos dois grupos de doentes com CECP que respondem diferentemente ao tratamento, apresentando diferentes taxas de recidivas locais/regionais e, por conseguinte associar um perfil genómico específico a cada um desses grupos.

Assim, foi propósito deste estudo:

- Recorrer a resultados provenientes da técnica aCGH referentes à variação do número de cópias de material genético, conjuntamente com a informação clínico-patológica dos doentes com diagnóstico de CECP, de forma a procurar associações entre estes dados;
- Identificar as regiões mínimas comuns alteradas, mais frequentes nos cromossomas, de forma a determinar a relevância das diferentes regiões alteradas dos cromossomas no CECP;
- Classificar os dados genómicos relativamente a várias condições, tais como estadiamento, localização, e principalmente a nível metastático/recidiva;
- Determinar o risco de recidivas/metástases em dois grupos distintos segundo diversas variáveis;
- Associar genes de interesse com funções essenciais no CECP;
- Procurar biomarcadores que possam prever o melhor/pior prognóstico do CECP;

A Figura 14 esquematiza o trabalho desenvolvido neste estudo.

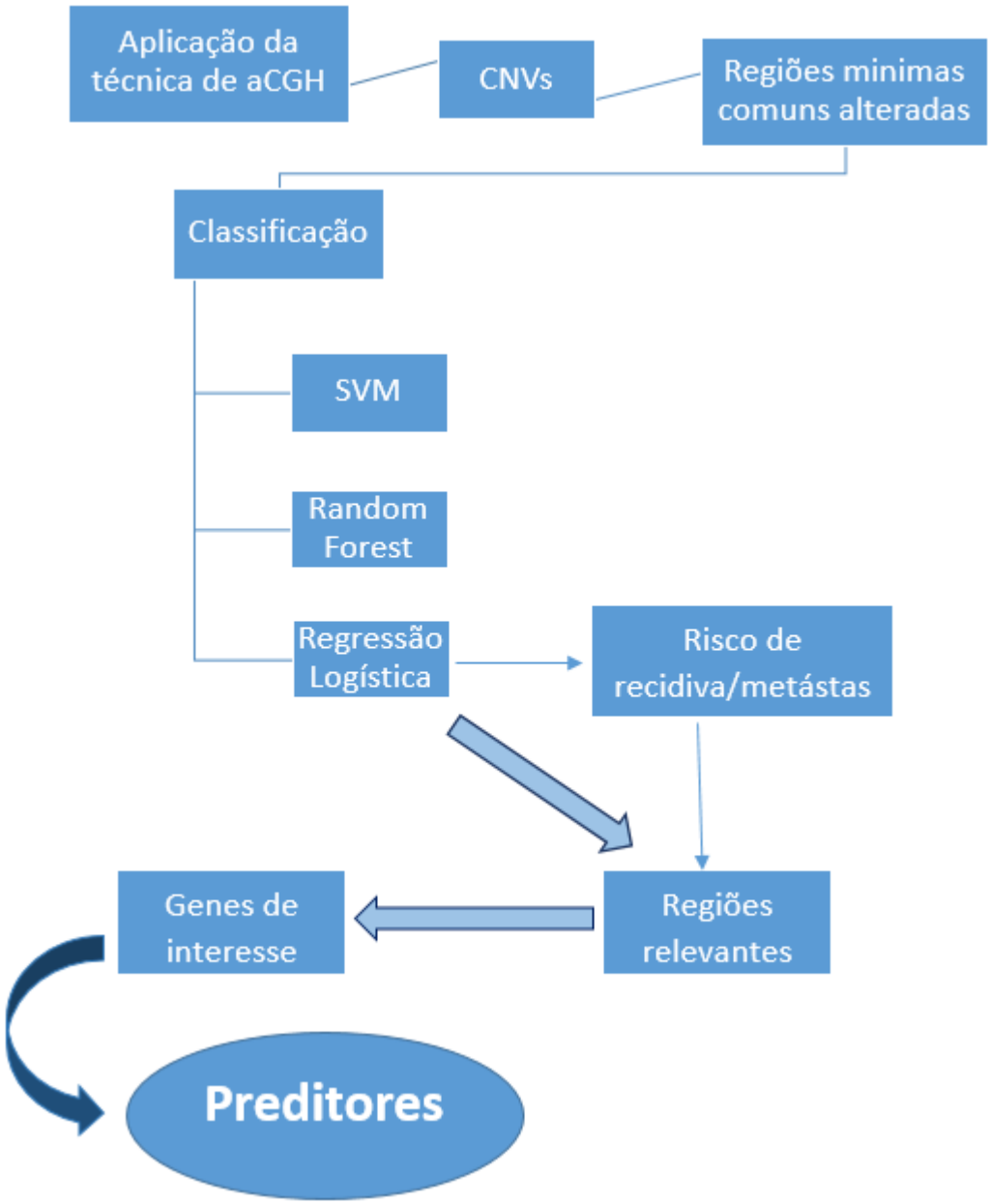


Figura 14 - Representação esquemática das etapas do trabalho.

3. MATERIAIS E MÉTODOS

3.1. População em estudo

De Outubro de 2010 a Agosto de 2015 foram recolhidas biópsias a doentes com diagnóstico de cancro da cabeça e pescoço, provenientes das consultas de Cirurgia Maxilofacial e/ou Estomatologia, dos Hospitais da Universidade de Coimbra, EPE. Este estudo foi aprovado pela Comissão de Ética em Investigação da Faculdade de Medicina da Universidade de Coimbra. Todos os doentes e indivíduos controlo deram o seu consentimento informado e escrito para participar no estudo de acordo com a Declaração de Helsínquia.

Posteriormente foram analisados 104 doentes usando a técnica de aCGH de forma a permitir o estudo relativamente à variação do número de cópias (CNV – do Inglês *Copy Number Variation*) de segmentos genómicos. Assim, ficou-se com a informação genómica de todos os doentes participantes do estudo.

As amostras de controlo usadas foram amostras de tecido gengival de indivíduos saudáveis.

Por fim, juntamente com o processo clínico, foram recolhidos os seguintes dados de todos os doentes: idade à data do diagnóstico, sexo, localização do tumor, estádio usando a classificação TNM, os hábitos tabágicos e o consumo de álcool, infeção por HPV, tratamento, o compromisso das margens, a invasão perineural, diferenciação histológica, presença de recidivas/metástases, bem como a data do diagnóstico anatomo-patológico confirmativo da recidiva/metástase, de forma a estabelecer o tempo de sobrevida e o intervalo livre de doença (este definido como o período entre a data do tratamento inicial e a data do diagnóstico da recidiva, caso exista). O período de follow up dos doentes foi até Fevereiro de 2016.

3.2. Análise por aCGH

A técnica de aCGH foi executada utilizando um *Agilent Human Genome microarray* 4x180K (Agilent Technologies Inc, Santa Clara, USA), constituído por sondas de oligonucleótidos com 60-mer. Assim, este *microarray* contém 180000 sondas de oligonucleótidos que abrangem sequências codificantes e não codificantes de todo o genoma.

Para esta configuração de *microarray* foi utilizada uma quantidade de DNA genómico de 1,1 µg para um volume final de 26 µL e, neste caso concreto sempre que necessário perfez-se o volume com água *Nuclease-free* (Promega Corporation, Wisconsin, USA). Esta tecnologia foi feita de acordo com as instruções fornecidas pela Agilent. Sumariamente, as amostras a analisar foram marcadas com o fluorocromo Cy5-dUTP e as amostras de controlo com Cy3-dUTP (Agilent Technologies Inc, Santa Clara, USA). O passo seguinte foi a realização da purificação do DNA genómico marcado usando filtros individuais Amicon 30KDa (Millipore, Massachusetts, USA) e TE (pH 8,0) (Promega Corporation, Wisconsin, USA), alcançando-se um volume final de 21 µL de DNA genómico marcado após a purificação. De seguida, efetuou-se a determinação do grau de marcação e da atividade específica, recorrendo a um espectrofotómetro (NanoDrop 1000, Thermo Scientific, Wilmington, USA). A mistura de hibridização foi adicionada a cada tubo que contém o DNA da amostra controlo marcado com Cy3 e o respetivo DNA da amostra em estudo marcado com Cy5. A hibridização ocorreu a 65 ºC durante 24 h, num forno de hibridização (Agilent Technologies Inc, Santa Clara, USA). Passado este tempo, realizaram-se as lavagens com *Agilent Oligo aCGH Wash Buffer 1* e 2 (Agilent Technologies Inc, Santa Clara, USA), durante 1 minuto à temperatura ambiente, seguido de 5 minutos a 37 ºC, respetivamente. Posteriormente fez-se o *scanner* das lâminas com o *scanner* de *microarray* G2565ca (Agilent Technologies Inc, Santa Clara, USA) e, as imagens foram processadas com o *software Feature Extraction v10.7* (Agilent Technologies Inc, Santa Clara, USA). Depois do processamento e controlo de qualidade os resultados foram analisados com o *software Agilent Genomic Workbench v6.5*.

3.3. Análise estatística

3.3.1. Importância das variáveis

Para determinação das variáveis que mais contribuem para a classificação recorreu-se ao *Importance Plot* inerente ao package da plataforma estatística R que contém o classificador Random Forest. Este procedimento teve por objetivo, como já dito, reduzir o número de variáveis tendo em conta a relevância das mesmas.

Para fazermos este processo recorremos ao R3.1.2, visto que é um ambiente de *software* livre para computação estatística e gráficos. Deste modo, criou-se um algoritmo em R (Figura 15).

Este código começa por importar a base de dados que contém toda a informação a nível genómico, obtida através de aCGH, de todos os doentes e correspondentes a todas as regiões dos autossomas. Posteriormente, os dados foram divididos em dois conjuntos, *testSet* (conjunto de teste) e *trainSet* (conjunto de treino), garantindo-se a proporção das variáveis no que concerne à variável dependente. Por exemplo, relativamente a ter ou não metástases, 1 ou 0 respetivamente na base de dados, há 40 doentes que têm metástases e 64 doentes que não têm metástases; já a variável dicotómica estádio (estádio I + II *vs*. estádio III + IV), apresenta uma divisão de doentes 45/ 59. Por fim, a localização pode ser 1 (tumor localizado na língua) ou 2 (tumor localizado em outras zonas), e 44 doentes correspondem a 1 e 60 doentes a 2.

A fase seguinte, do código desenvolvido, comporta a realização do *Importance Plot* a partir do método Random Forest com o objetivo de redução de variáveis. Para este efeito, fazem-se 1000 execuções do algoritmo acumulando a frequência com que cada variável (que corresponde a uma região genética) foi caracterizada como relevante. Uma vez que o número de variáveis é bastante superior ao número de sujeitos disponíveis (658 variáveis e 104 sujeitos), este processo de escolha aleatória de variáveis em cada execução do algoritmo permite manter um ratio adequado entre variáveis e sujeitos e obter uma medida de importância de cada variável.

No final, e com base nas variáveis independentes mais importantes (16 para o estádio e metástases, 15 para a localização), calculou-se a exatidão, sensibilidade e especificidade do classificador Random Forest no conjunto de teste.

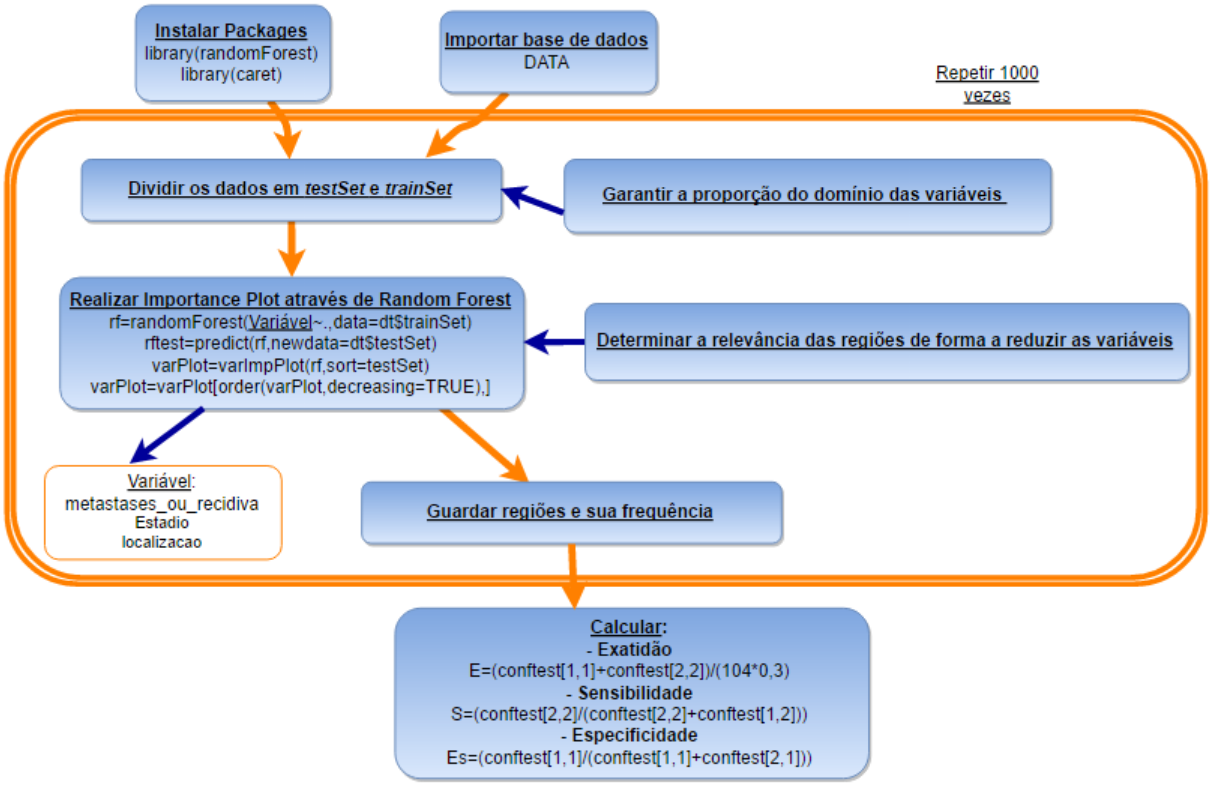


Figura 15 - Fluxograma do algoritmo para obter a variável *Importance Plot*.

3.3.2. Rotinas de Classificação

De modo a proceder à classificação dos dados usando o Random Forest criou-se uma outra rotina na mesma plataforma de *software* (R3.1.2) (Figura 16).

O código recorre naturalmente à preparação das packages necessárias à função do Random Forest, nomeadamente 'pROC', 'caret', 'MASS' e 'randomForest'. Após a importação da base de dados obtida pelos dados de aCGH, a qual foi apelidada de 'DATA', procedeu-se à mesma divisão aleatória dos dados num conjunto de treino do classificador e num conjunto de teste garantindo, uma vez mais, a proporção respeitante à variável dependente. Tendo em conta as variáveis independentes mais importantes determinadas anteriormente e para cada uma das variáveis dependentes (metástases, estádio e localização) fez-se o treino do classificador no grupo de treino aplicando-se o mesmo ao grupo de teste. Os resultados obtidos no conjunto de teste foram então usados para determinação da exatidão, da sensibilidade e da especificidade. O procedimento foi repetido aleatoriamente 1000 vezes e tendo-se calculado a média e desvio padrão destes parâmetros.

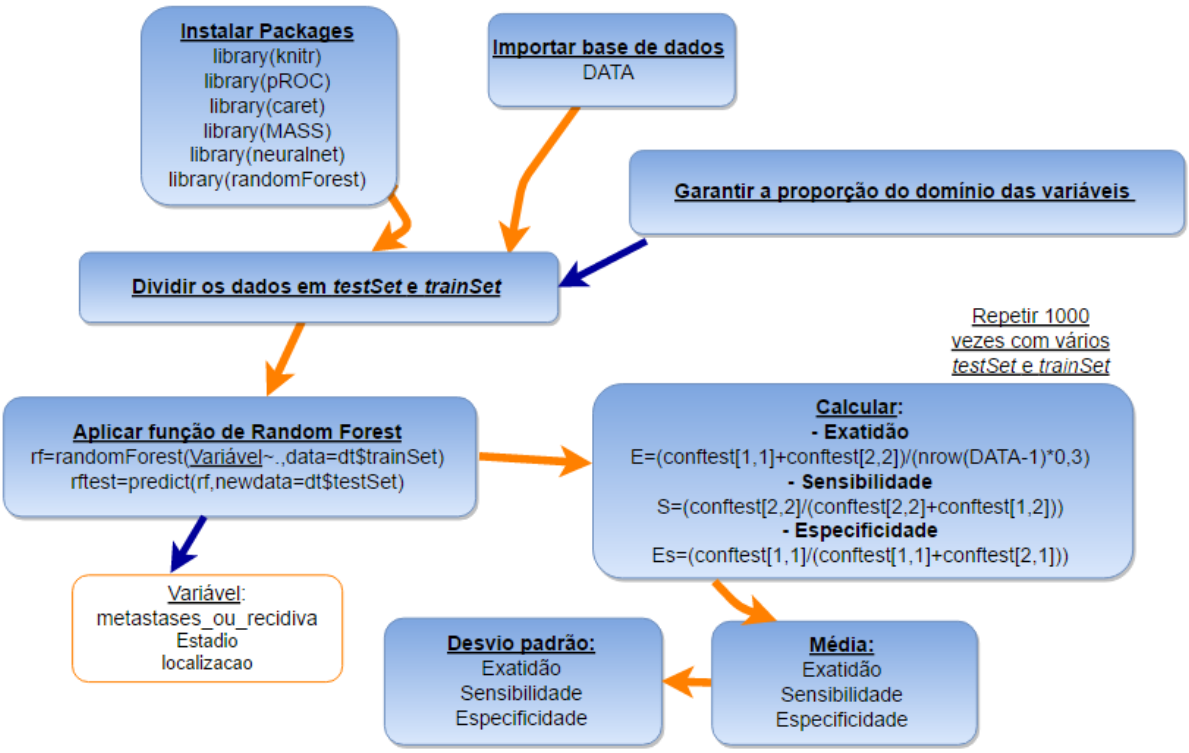


Figura 16 - Fluxograma do algoritmo do classificador Random Forest.

Posteriormente, usou-se também o SVM com o intuito de classificar os dados. E compararam-se os dois classificadores de forma a concluir qual a melhor opção a ser adotada para os dados em questão recorrendo a um algoritmo também criado na mesma plataforma tal como ilustrado na Figura 17. Nesta rotina calculou-se também a AUC com o intuito da comparação ser mais quantitativa relativamente aos dois classificadores.

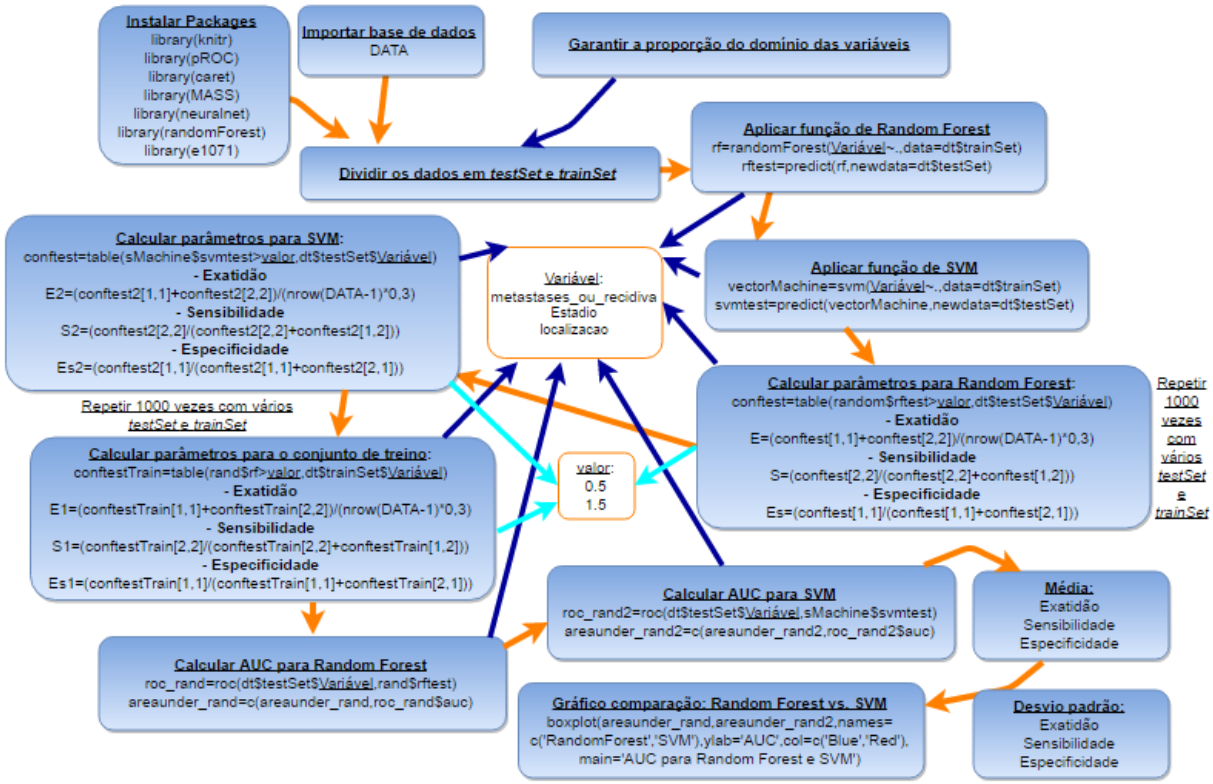


Figura 17 - Fluxograma do algoritmo do classificador Random Forest e do classificador SVM.

Por fim, a análise de dados recorrendo à regressão logística foi realizada no programa *Statistical Package for the Social Sciences* (SPSS) de IBM. Esta técnica estatística devolveu valores relativos à significância, ao coeficiente β e ao intervalo de confiança (IC).

Posteriormente, foram obtidas as curvas de sobrevivência relacionadas com as regiões relevantes determinadas pela regressão logística e em função da presença de metástases.

3.3.3. Visualização dos dados

A visualização dos resultados foi feita por recurso a gráficos comuns, no entanto, também se implementou um algoritmo em MATLAB R2014a para visualização de alguns dados genéticos relativos a informação de perda e ganho.

O algoritmo desenvolvido cria um gráfico radar que permite avaliar rapidamente a informação contida numa base de dados de genética, neste caso específico, a base de dados obtida pelos resultados de aCGH (Figura 18).

O intuito do gráfico radar era condensar diversa informação na mesma visualização, daí a escolha de uma cor de fundo adequada à parte da informação que se pretendia representar.

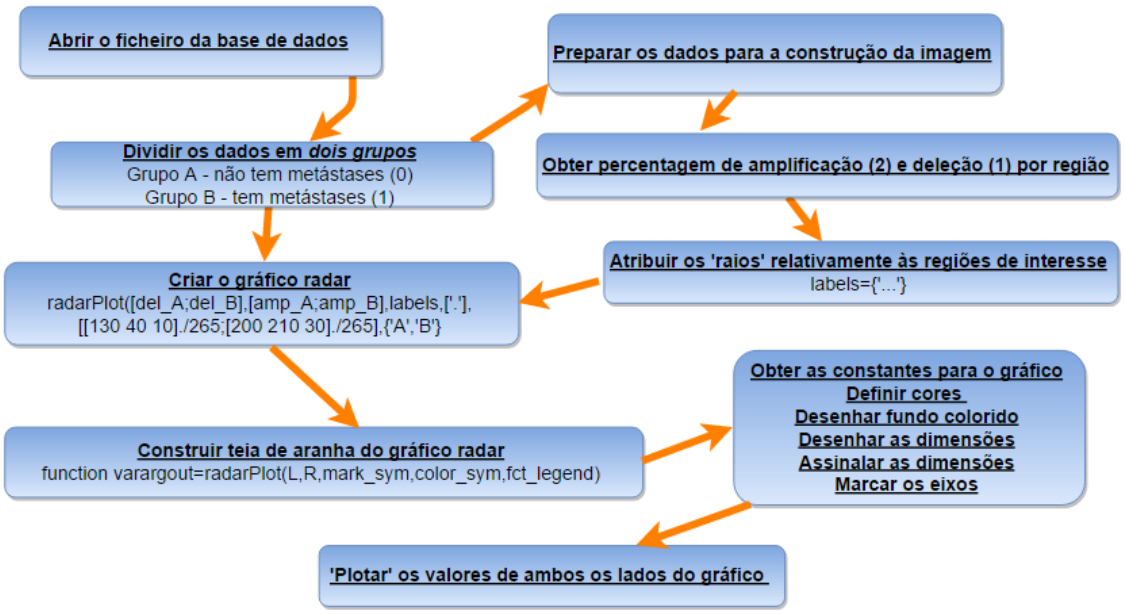


Figura 18 - Fluxograma do algoritmo para o Gráfico Radar.

Uma outra forma de visualizar os resultados foi através de um ideograma cromossómico, o qual foi criado através de um algoritmo também implementado em Matlab R2014a. O algoritmo representa assim todos os cromossomas com as respetivas variações do número de cópias (do Inglês, *copy number variation*), CNVs, nos autossomas. Os cromossomas sexuais não foram tidos em conta nesta análise, não apresentando assim qualquer CNV no ideograma.

Para implementar este algoritmo foi necessário importar uma base de dados contendo as informações dos cromossomas disponível em ftp://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/cytoBandIdeo.txt.gz.

A Figura 19 representa o fluxograma correspondente ao algoritmo.

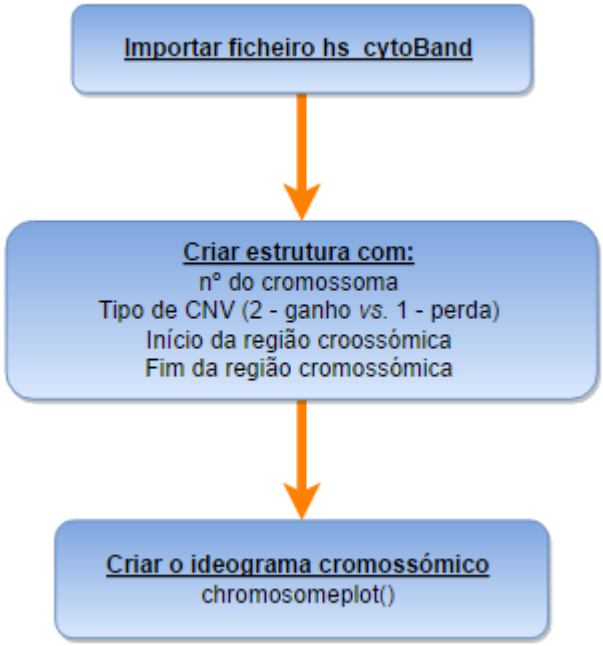


Figura 19 – Fluxograma do algoritmo para o ideograma cromossómico.

4. RESULTADOS E DISCUSSÃO DOS RESULTADOS

4.1. População em estudo

As caraterísticas clínicas dos 104 doentes incluídos no estudo encontram-se discriminadas na Tabela 6.

Tabela 6 - Caraterísticas clínicas dos doentes com carcinoma da cabeça e pescoço.

	Número de doentes	
Sexo	♂	88
	♀	16
Idade	$<$ 60 anos	52
	$\geq$ 60 anos	52
Localização tumoral	Língua	44
	Pavimento da boca	28
	Trigono-retromolar	8
	Mucosa	6
	Rebordo-gengival	3
	Amígdala	2
	Rebordo-alveolar	3
	Palato	4
	Hipofaringe	2
	Laringe	2
	Supraglote	1
	Epiglote	1
Estádio	I	18
	II	27
Estádio	III	20
	IV	39
Metástases	Sim	40
	Não	64
Tabaco	Sim	76
	Não	28
Álcool	Sim	70
	Não	31
	Sem informação	3
Tratamento	Quimioterapia	26
	Radioterapia	70
	Cirurgia	89

4.2. Resultados obtidos por *Importance Plot*: redução de variáveis

Os dados obtidos do aCGH a nível genómico retrataram 658 regiões correspondentes aos autossomas, onde as regiões foram distribuídas por cromossoma começando a

região 1 no início do braço curto do cromossoma 1 e terminando no cromossoma 22 na extremidade do braço longo. A correspondência entre as regiões e as bandas dos cromossomas foi obtida a partir do *Genome Browser - University of California Santa Cruz*[108] recorrendo ao *Genomes Human GRCh37/hg19*.

De forma a reduzir as variáveis, visto que o modelo não aceita mais variáveis que casos (104 doentes), usou-se o método associado ao *Importance Plot* descrito anteriormente.

Como já referido na seção Materiais e Métodos a redução das variáveis é feita em função da presença de metástases, do estádio em que se encontra o tumor do doente ou da localização do tumor.

Relativamente à redução das variáveis em função do estádio obtiveram-se 21 regiões, sendo estas as que apresentaram uma frequência superior 40%. Contudo, do ponto de vista clínico apenas 16 destas regiões apresentaram interesse para o CECP, tendo como base GeneCards®. As regiões mais relevantes em relação ao estádio estão representadas na Tabela 7.

Tabela 7 - Regiões mais relevantes com interesse clínico para o CECP em função do estádio. A frequência representa o número de vezes que nas 1000 execuções do código a variável foi considerada entre as 16 mais importantes.

Região	Frequência
3q28-q29	884
7q34-q35	627
6p25.3-p25.2	627
2p23.1-p22.3	597
7q31.33-q32.3	556
22q11.21	540
1q23.2-q23.3	514
1q42.2-q43	504
8p11.23-p11.1	502
17q21.31-q21.32	495
5q35.2-q35.3	494
3p14.3-p14.2	484
1p36.21-p36.13	484
10q11.22-q11.23	462
2p22.3-p22.2	435
8p23.3-p23.1	434

Por sua vez, a restrição das variáveis em função da localização a partir do *Importance Plot* reduziu as regiões a 19, sendo que apenas 15 apresentaram interesse clínico para o CECP, tendo como base GeneCards®. Neste processo também se fez a restrição a partir

da frequência superior a 40%. As regiões mais relevantes em relação à localização estão representadas na Tabela 8.

Tabela 8 - Regiões mais relevantes com interesse clínico para o CECP em função da localização. A frequência representa o número de vezes que nas 1000 execuções do código a variável foi considerada entre as 15 mais importantes.

Região	Frequência
6p22.1-p21.32	996
3q26.33-q28	959
8p23.3-p23.1	863
1q23.2-q23.3	820
5q13.2-q13.3	689
17q21.31-q21.32	665
1p36.33-p36.32	512
22q11.21	491
8p12-p11.23	479
20q11.1-q11.22	473
2q37.3	470
20q13.31-q13.33	457
4p16.3	452
11p15.5-p15.4	435
12p13.33-p13.32	429

Por último, a eliminação de variáveis em função da presença de metástases limitou as regiões a 21, contudo apenas 16 demonstraram interesse clínico para o CECP, tendo como base GeneCards®. Neste processo também se fez a restrição a partir da frequência superior a 40%. As regiões mais relevantes em relação à presença de metástases estão representadas na Tabela 9.

Tabela 9 - Regiões mais relevantes com interesse clínico para o CECP em função da presença de metástases. A frequência representa o número de vezes que nas 1000 execuções do código a variável foi considerada entre as 16 mais importantes.

Região	Frequência
8q12.1-q13.1	665
6p25.3-p25.2	652
15q26.3	651
5p15.33-p15.32	649
22q11.22-q11.23	601
6q16.1	566
3q26.31-q26.33	554
11q13.5-q14.1	498
5q21.3-q22.2	492
11p14.1-p13	471
6q16.1-q16.3	450
4p14-p13	431
1q31.3-q32.1	430
3p14.3-p14.2	423
17q21.31-q21.32	418
1q21.1-q21.2	402

4.3. Resultados obtidos por Random Forest

Nesta fase o que se obteve de Random Forest foi a média e o respetivo desvio padrão dos parâmetros exatidão, sensibilidade e especificidade tendo em conta a variável de referência, ou seja, estádio, localização ou presença de metástases.

Este modelo em função do estádio mostra uma exatidão média de cerca de 64%, mostrando que em 100 pessoas apenas 64 apresentam um resultado próximo do valor verdadeiro do estádio. Também mostra uma especificidade média aproximada de 49%, o que mostra que este modelo tem uma capacidade baixa de excluir corretamente pessoas de um determinado estádio, somente cerca de metade das pessoas. E apresenta uma sensibilidade média de cerca de 73%, onde se pode verificar que este método tem cerca de $\frac{3}{4}$ de probabilidade de identificar corretamente o estádio das pessoas, concomitantemente $\frac{1}{4}$ de produzir resultados falso-negativos.

A Tabela 10 esquematiza os valores médios e respetivo desvio padrão dos parâmetros relativamente ao estádio.

Tabela 10 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest em relação ao estádio.

Parâmetros	Média	Desvio padrão
Exatidão	0,64	0,07
Especificidade	0,49	0,13
Sensibilidade	0,73	0,11

Por sua vez, este processo em função da localização apresenta uma exatidão média aproximadamente de 71%, ou seja, em 100 pessoas 71 apresentam uma localização próxima da localização verdadeira. É também interpretada a especificidade média de cerca de 54%, significando que em 100 pessoas 54 são excluídas corretamente devido a não ter determinada localização que era suposto, neste caso a aptidão do modelo é reduzida. E a sensibilidade média é de 72%, isto é, em 100 pessoas 72 são identificadas corretamente relativamente à localização e 28 pessoas são avaliadas localmente como falso-negativas.

Na Tabela 11 são visíveis os valores médios e respetivo desvio padrão dos parâmetros em função da localização.

Tabela 11 - Valores médios dos parâmetros e respetivo desvio padrão obtidos por Random Forest em relação à localização.

Parâmetros	Média	Desvio padrão
Exatidão	0,71	0,07
Especificidade	0,54	0,13
Sensibilidade	0,81	0,10

Por fim, relativamente à presença de metástases o Random forest obteve os seguintes parâmetros: exatidão média sensivelmente de 71%, o que quer dizer que em 100 pessoas 71 são consideradas bem classificadas se têm ou não metástases; especificidade média de 84%, ou seja, o modelo tem uma aptidão para excluir corretamente 84 pessoas de 100 devido à presença ou não de metástases; e sensibilidade média próxima de 51%, isto é, 51 pessoas de 100 são classificadas corretamente em relação à presença de metástases.

Como se pode constatar na Tabela 12 estão esquematizados os valores médios e respetivo desvio padrão dos parâmetros.

Tabela 12 - Valores médios dos parâmetros e respetivo desvio padrão obtidos por Random Forest em relação à presença de metástases.

Parâmetros	Média	Desvio padrão
Exatidão	0,73	0,07
Especificidade	0,84	0,08
Sensibilidade	0,51	0,14

4.4. Resultados obtidos na comparação Random Forest e SVM

Nesta etapa, usaram-se dois classificadores, independentemente da variável dependente, de forma a poder comparar a eficácia de cada um para classificar corretamente cada uma das situações testadas.

Após correr o algoritmo desta secção obteve-se a média dos parâmetros exatidão, especificidade e sensibilidade dos classificadores Random Forest e SVM, e do grupo de treino, *trainSet*, e o respetivo desvio padrão para cada valor. Também se obteve a AUC para cada um destes dois classificadores.

Relativamente à variável estádio, quando os dados são tratados aleatoriamente, ou seja, sem recorrer a nenhum classificador, a probabilidade de acertar corretamente na classificação dos dados é cerca de 57%, obtido por

$$\frac{59}{59+45} = 0.57,$$

sendo que 59 corresponde ao número de casos no estádio $III + IV$ e 45 o número de casos no estádio $I + II$.

Por sua vez, quando se utiliza um classificador o conjunto de treino tem exatidão média de 60%, especificidade média próxima de 46% e sensibilidade média cerca de 72%, tal como ilustrado na Tabela 13.

Em contrapartida, os parâmetros obtidos pelo Random Forest em função do estádio, como se pode constatar na Tabela 14, são: exatidão 64%, especificidade 49% e sensibilidade 73%; já os parâmetros obtidos pelo SVM são: exatidão 69%, especificidade 55% e sensibilidade 77%. A conclusão que se pode tirar destes pontos é: a exatidão dos classificadores é melhor que a exatidão correspondente ao modelo aleatório, mas esse aumento não é muito significativo, e uma das causas desse evento é a exatidão do próprio modelo não ser também muito boa, e pode-se também visualizar que a exatidão do SVM é ligeiramente melhor que a do Random Forest; a especificidade também melhora quando se recorre aos classificadores e o mesmo ocorre relativamente à

sensibilidade. Sumariando, tendo em conta o estádio, esta variável não é conclusiva para classificar novos dados que surjam.

Tabela 13 - Valores médios e respetivo desvio padrão dos parâmetros do conjunto de treino, *trainSet*, do modelo em função do estádio.

	Parâmetros	**Média**	**Desvio padrão**
TrainSet	Exatidão	0,60	0,05
	Especificidade	0,46	0,08
	Sensibilidade	0,72	0,05

Tabela 14 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest e SVM em relação ao estádio no conjunto de teste.

	Random Forest		**SVM**	
Parâmetros	**Média**	**Desvio padrão**	**Média**	**Desvio padrão**
Exatidão	0,64	0,07	0,69	0,07
Especificidade	0,49	0,13	0,55	0,12
Sensibilidade	0,73	0,10	0,77	0,10

Falando agora da AUC, o valor de AUC é inferior no Random Forest, 0,68 com um desvio padrão de 0,08, em relação ao SVM, 0,74 com um desvio padrão de 0,08. Ou seja, visto que a AUC varia entre 0,5 e 1, e que quanto mais próximo de 1 melhor é o poder de discriminar entre duas observações, SVM apresenta um melhor poder de discriminação, como se pode visualizar na Figura 20.

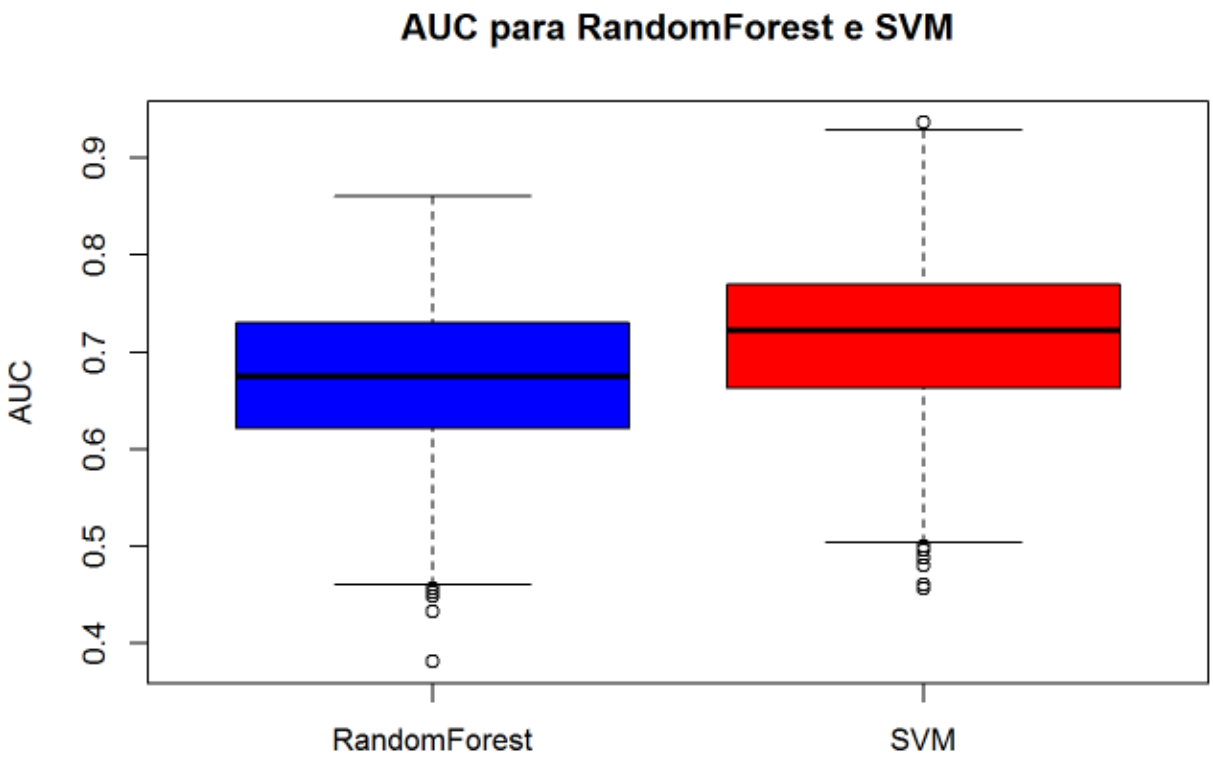


Figura 20 - Representação da AUC para Random Forest e SVM em relação ao estádio.

Tendo agora em consideração a variável localização, tratando os dados de forma aleatória, isto é, sem qualquer técnica, a probabilidade de acertar corretamente na classificação dos dados é cerca de 57%, calculada a partir de

$$\frac{60}{60+44} = 0.57,$$

sendo que 60 corresponde a outras localizações e 44 à localização língua.

Porém, quando se utiliza um classificador o conjunto de treino tem exatidão média de 69%, especificidade média aproximadamente 54% e sensibilidade média cerca de 80%, tal como ilustrado na Tabela 15.

Por sua vez, os parâmetros obtidos pelo Random Forest em função da localização, como se pode certificar na Tabela 16, são os seguintes: exatidão 71%, especificidade 54% e sensibilidade 81%; já os parâmetros obtidos pelo SVM: exatidão 69%, especificidade 52% e sensibilidade 80%. Assim, o que se pode constatar é: a exatidão de Random Forest e SVM é superior à exatidão sem recorrer a classificadores, mas a exatidão de SVM é ligeiramente inferior a Random Forest; a especificidade e a sensibilidade são semelhantes em todos os eventos desta fase. Finalizando, tendo em conta a localização, esta variável não é conclusiva para classificar novos dados que surjam.

Tabela 15 - Valores médios e respetivo desvio padrão dos parâmetros do conjunto de treino, *trainSet*, do modelo em função da localização.

	Parâmetros	**Média**	**Desvio padrão**
TrainSet	**Exatidão**	0,69	0,04
	Especificidade	0,54	0,04
	Sensibilidade	0,80	0,06

Tabela 16 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest e SVM em relação à localização no conjunto de teste.

	Random Forest		**SVM**	
Parâmetros	**Média**	**Desvio padrão**	**Média**	**Desvio padrão**
Exatidão	0,71	0,07	0,69	0,07
Especificidade	0,54	0,13	0,52	0,13
Sensibilidade	0,81	0,10	0,80	0,10

Referindo agora a AUC, o valor de AUC é ligeiramente superior em SVM, 0,78 com um desvio padrão de 0,07, do que em Random Forest, 0,76 com um desvio padrão de 0,07. Ou seja, SVM aproxima-se mais de 1, logo tem melhor poder de discriminação entre duas observações que Random Forest, tal como ilustrado na Figura 21.

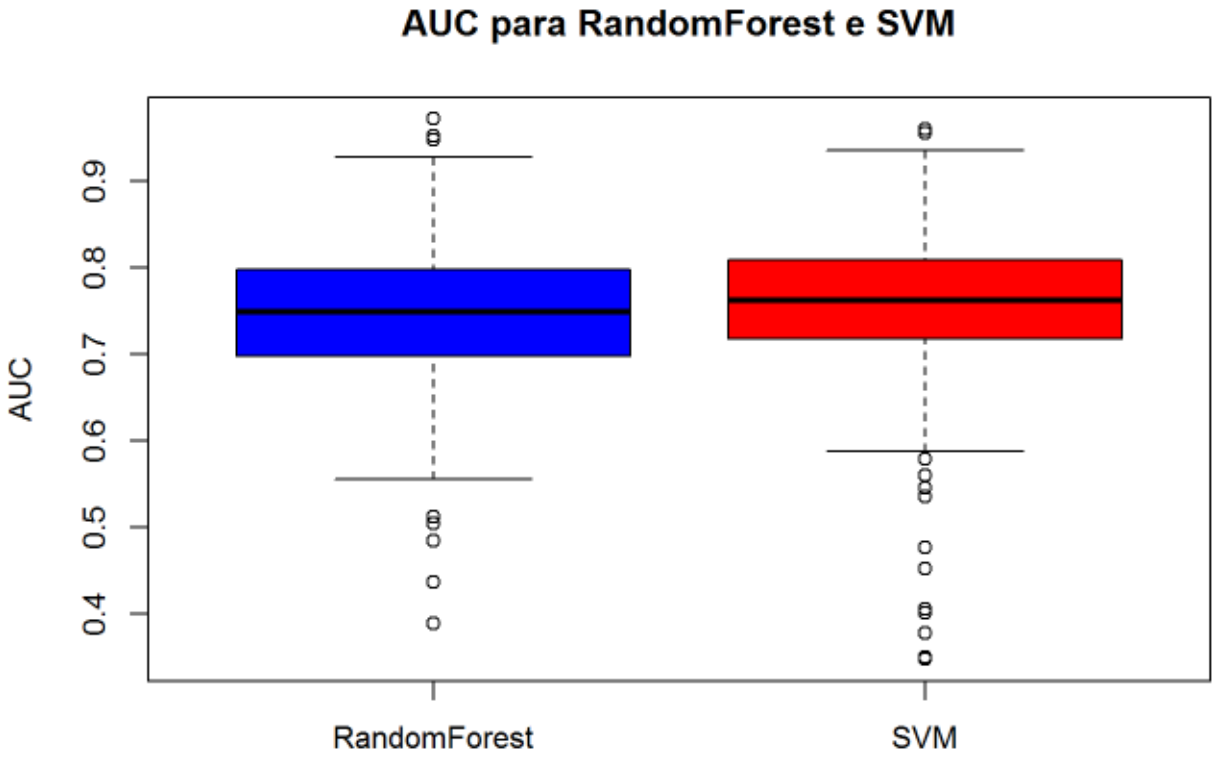


Figura 21 - Representação da AUC para Random Forest e SVM em relação à localização.

Por fim considerando a variável presença de metástases, e tratando os dados sem recorrer a qualquer técnica, a probabilidade de acertar corretamente na classificação dos dados é cerca de 62%, calculada a partir de

$$\frac{64}{64+40} = 0.62,$$

sendo que 40 corresponde a metástases e 64 a não metástases.

Ao invés, quando se usa um classificador o conjunto de treino tem exatidão média de 69%, especificidade média aproximadamente 83% e sensibilidade média cerca de 48%, como é visível na Tabela 17.

Contudo, os parâmetros obtidos pelo Random Forest em função da presença de metástases, tal como mencionado na Tabela 18, são os seguintes: exatidão 73%, especificidade 84% e sensibilidade 50%; por sua vez, os parâmetros obtidos pelo SVM: exatidão 73%, especificidade 85% e sensibilidade 47%. Por conseguinte, a exatidão de Random Forest e SVM são semelhantes e superiores à exatidão sem recorrer a classificadores; a especificidade e a sensibilidade são similares em todos os eventos desta fase. Em suma, tendo em conta a presença de metástases, esta variável é a que mostra mais significância em classificar novos dados que surjam.

Tabela 17 - Valores médios e respetivo desvio padrão dos parâmetros do conjunto de treino, *trainSet*, do modelo em função da presença de metástases.

	Parâmetros	**Média**	**Desvio Padrão**
TrainSet	Exatidão	0,69	0,04
	Especificidade	0,83	0,04
	Sensibilidade	0,48	0,07

Tabela 18 - Valores médios e respetivo desvio padrão dos parâmetros obtidos por Random Forest e SVM em relação à presença de metástases no conjunto de teste.

	Random Forest		**SVM**	
Parâmetros	**Média**	**Desvio Padrão**	**Média**	**Desvio padrão**
Exatidão	0,73	0,07	0,73	0,07
Especificidade	0,84	0,08	0,85	0,09
Sensibilidade	0,50	0,14	0,47	0,15

Aludindo agora a AUC, o valor de AUC é ligeiramente superior em SVM, 0,76 com um desvio padrão de 0,07 do que em Random Forest, 0,74 com um desvio padrão de 0,08. E assim, SVM aproxima-se mais de 1, logo tem melhor poder de discriminação entre duas observações que Random Forest, tal como se pode ver na Figura 22.

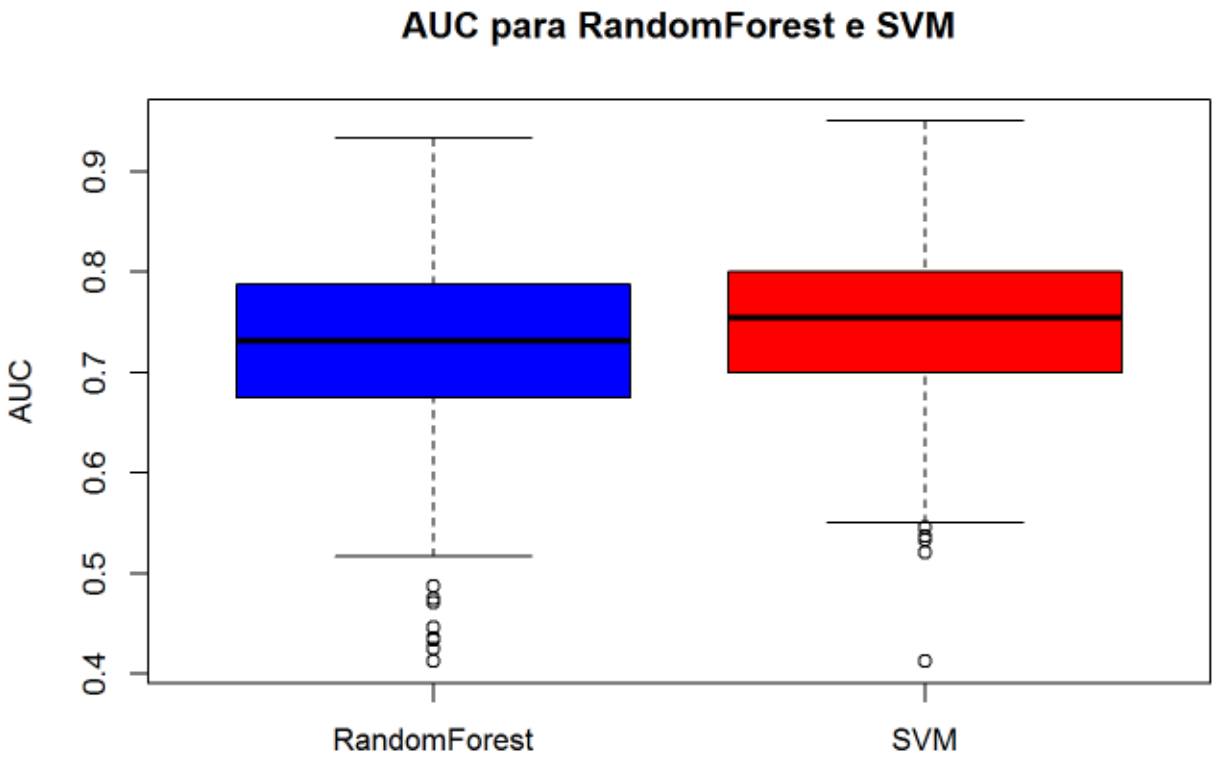


Figura 22 - Representação da AUC para Random Forest e SVM em relação à presença de metástases.

4.5. Resultados obtidos por Regressão Logística: determinação de variáveis relevantes para CECP

Nesta secção, a partir do *software* SPSS, foi obtido os valores de coeficiente β e seu IC e também o valor de significância (Sig).

A partir das regiões obtidas pelo processo de eliminação de variáveis ajustou-se um modelo logístico com o objetivo de obter uma quantificação da importância das variáveis para a classificação. Contudo, discriminou-se estas regiões em duas partes: amplificação, designado como 'A', e deleção, definida como 'D'; de forma a saber quando é mais relevante a região, se amplificada ou deletada. Posteriormente, após correr os dados no SPSS, a forma de analisá-los foi: ordenar pela significância, eliminar as regiões cujos valores do limite inferior do IC têm valores nulos, filtrar as regiões mais relevantes e discutir os valores de β.

Na Tabela 19 estão esquematizadas as regiões mais relevantes determinadas por este modelo e estão ordenadas de forma decrescente relativamente ao limite inferior de 95% do IC.

Nesta tabela, a coluna "Região" refere se aquela banda está amplificada ou deletada, como referido anteriormente.

Por sua vez a coluna "β" refere os valores do coeficiente obtidos pela regressão logística. Neste caso, o valor positivo deste coeficiente é "favorável" à recidiva e o valor negativo é "favorável" à ausência de metástases; quanto mais elevado em valor absoluto mais influência tem essa região.

Relativamente à "Sig", este valor refere-se ao valor de p, o qual informa a probabilidade que um conjunto de dados seja um falso-positivo decorrente do acaso. Contudo, apenas uma região tem um valor de p menor que 0,05 pelo que todos os elementos vão ser tratados como descritivos.

Por fim, o IC é o intervalo estimado no qual a média de um parâmetro de uma amostra tem uma dada probabilidade de ocorrer. Comumente descreve-se como o intervalo onde há 95% de probabilidade da média verdadeira da população inteira ocorrer, sendo este intervalo designado 95% IC.

Nesse caso, em conclusão, as regiões 11q13.5-q14.1, 22q11.22-q11.23, 3q26.31-q26.33, 3p14.3-p14.2, 6p25.3-p25.2, 15q26.3 e 11p14.1-p13 quando amplificadas apresentam um pior prognóstico, visto que apresentam valores β que influenciam positivamente a presença de metástases; as regiões 5p15.33-p15.32, 22q11.22-q11.23, 6q16.1-q16.3, 17q21.31-q21.32, 3p14.3-p14.2 e 3q26.31-q26.33 quando deletadas apresentam também pior prognóstico pelo mesmo motivo que as regiões acima referidas; as regiões 17q21.31-q21.32, 1q21.1-q21.2, 4p14-p13, 1q31.3-q32.1, 5p15.33-p15.32 e 6q16.1-q16.3 quando amplificadas apresentam melhor prognóstico, uma vez que os valores negativos de β influenciam negativamente a presença de metástases; e as regiões 1q21.1-q21.2, 6p25.3-p25.2 e 1q31.3-q32.1 quando deletadas têm melhor prognóstico.

Tabela 19 - Regiões mais relevantes com interesse clínico para o CECP determinadas pela regressão logística.

	Região	β	Sig.	95% IC Exp(β)	
				Lower	Upper
11q13.5-q14.1	A	4,322	0,031	1,497	3793,73
5p15.33-p15.32	D	2,503	0,059	0,905	165,088
22q11.22-q11.23	A	2,384	0,072	0,808	145,696
1q21.1-q21.2	D	−2,527	0,073	0,005	1,26
17q21.31-q21.32	A	−1,626	0,115	0,026	1,484
22q11.22-q11.23	D	2,046	0,128	0,554	108,121
6q16.1-q16.3	D	2,805	0,131	0,432	631,437
17q21.31-q21.32	D	1,387	0,176	0,537	29,823
1q21.1-q21.2	A	−1,45	0,238	0,021	2,611
3q26.31-q26.33	A	1,39	0,252	0,372	43,342
3p14.3-p14.2	A	1,769	0,351	0,143	241,248
6p25.3-p25.2	A	0,963	0,378	0,308	22,243
4p14-p13	A	−1,202	0,434	0,015	6,101
1q31.3-q32.1	A	−0,909	0,578	0,016	9,875
3p14.3-p14.2	D	0,483	0,627	0,23	11,421
3q26.31-q26.33	D	1,938	0,684	0,001	79172,2
5p15.33-p15.32	A	−0,383	0,687	0,105	4,41
15q26.3	A	0,463	0,759	0,082	30,854
6p25.3-p25.2	D	−0,27	0,817	0,077	7,545
1q31.3-q32.1	D	−0,487	0,818	0,01	38,677
6q16.1-q16.3	A	−0,486	0,82	0,009	40,377
11p14.1-p13	A	0,337	0,852	0,041	47,546

4.6. Análise de sobrevivência

De forma a determinar o tempo de sobrevida média recorreu-se a uma análise de sobrevivência usando gráficos de Kaplan-Meyer os quais foram realizados tendo em conta as regiões em análise.

Para a presente análise apenas 11 das 22 regiões inicialmente escolhidas foram consideradas clinicamente relevantes.

Essas 11 regiões são: 11q13.5-q14.1 amplificada, 22q11.22-q11.23 amplificada, 3p14.3-p14.2 amplificada, 11p14.1-p13 amplificada, 17q21.31-q21.32 amplificada, 1q21.1-q21.2 amplificada, 22q11.22-q11.23 deletada, 6q16.1-q16.3 deletada, 17q21.31-q21.32 deletada, 3p14.3-p14.2 deletada e 3q26.31-q26.33 deletada.

De seguida, apresenta-se discriminada e individualmente cada região em função dos parâmetros referidos inicialmente.

Cada região amplificada está codificada na base de dados com os valores 0 e 1, significando o valor 1 que a região apresenta amplificação e o valor 0 significa que a região não tem alteração ou está deletada, mas neste caso surge como não alterada porque não é diferenciada. O mesmo esquema é usado para as regiões deletadas: o valor 1 significa que a região está deletada e o valor 0 que a região não tem alteração ou tem amplificação.

A região 11q13.5-q14.1 quando amplificada apresenta um tempo de sobrevida médio de 33,7 meses que é mais baixo do que quando a região não apresenta alteração (64,8 meses), e tem um valor de p de 0,375 o que mostra que este resultado é descritivo (Tabela 20). A curva de sobrevivência mostra a mesma informação: quando a região não está alterada tem uma sobrevivência maior e quando amplificada o tempo de falecimento em meses é muito menor (Figura 23). Assim, esta região apresenta um pior prognóstico em concordância com a regressão logística.

Tabela 20 - Tempo de sobrevida médio para a região 11q13.5-q14.1 quando amplificada.

11q13.5-q14.1 A	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	64,807	5,427	54,171	75,444	0,375
1	33,708	8,655	16,745	50,671	

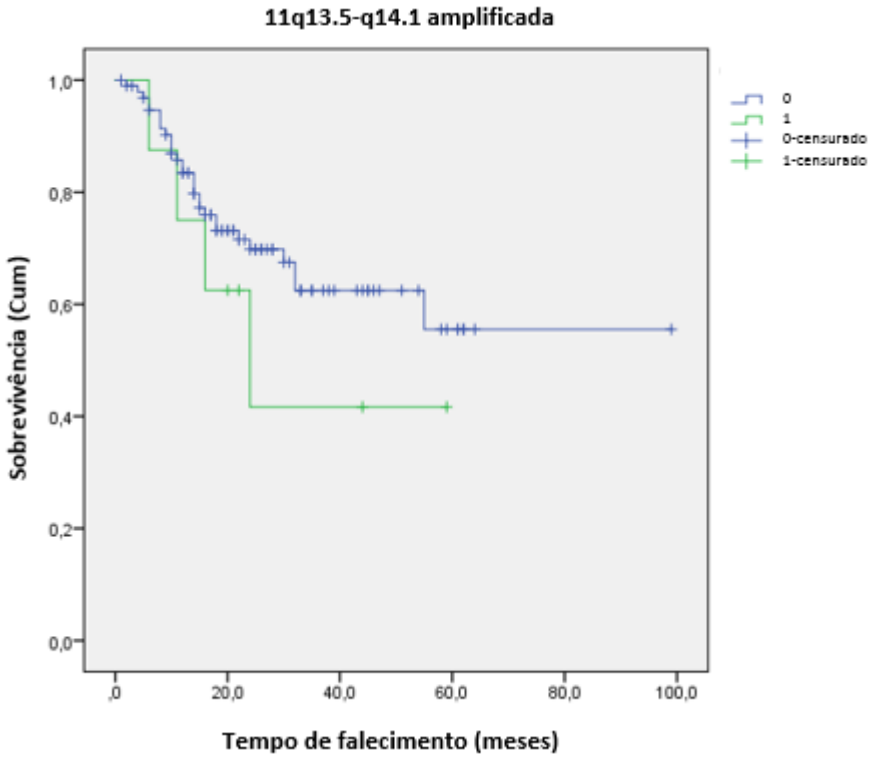


Figura 23 - Gráfico da curva de sobrevivência da região 11q13.5-q14.1 quando amplificada.

A região 22q11.22-q11.23 quando amplificada exibe também um tempo de sobrevida médio de 37,4 meses que é mais baixo do que quando a região não apresenta alteração (64,8 meses), e tem um valor de p de 0,450 o que mostra que este resultado é descritivo (Tabela 21). Também a curva de sobrevivência mostra que quando a região não está alterada tem uma sobrevivência maior mas não muito acentuada e quando amplificada o tempo de falecimento em meses é relativamente menor, tal como ilustrado na Figura 24. Logo, esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 21 - Tempo de sobrevida médio para a região 22q11.22-q11.23 quando amplificada.

22q11.22-q11.23 A	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	63,942	5,855	52,467	75,418	0,450
1	37,417	8,102	21,536	53,297	

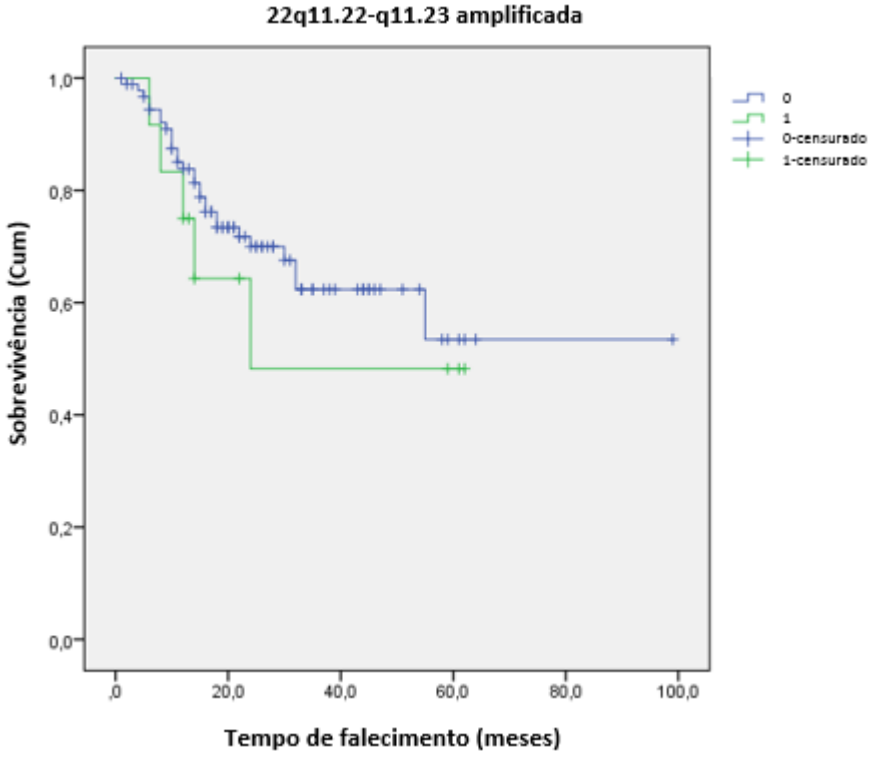


Figura 24 - Gráfico da curva de sobrevivência da região 22q11.22-q11.23 quando amplificada.

A região 17q21.31-q21.32 quando amplificada exibe um tempo de sobrevida 76,4 meses que é mais alto do que quando a região não apresenta alteração (38,5 meses), e tem um valor de p de 0,016 o que mostra que este resultado é estatisticamente significativo (Tabela 22). E a curva de sobrevivência indica que quando a região não está alterada

tem uma sobrevivência menor e quando amplificada o tempo de falecimento em meses é relativamente maior, tal como ilustrado na Figura 25. Portanto, esta região expõe um melhor prognóstico em concordância com a regressão logística.

Tabela 22 - Tempo de sobrevida médio para a região 17q21.31-q21.32 quando amplificada.

17q21.31-q21.32 A	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	38,524	3,445	31,772	45,276	0,016
1	76,351	7,545	61,563	91,139	

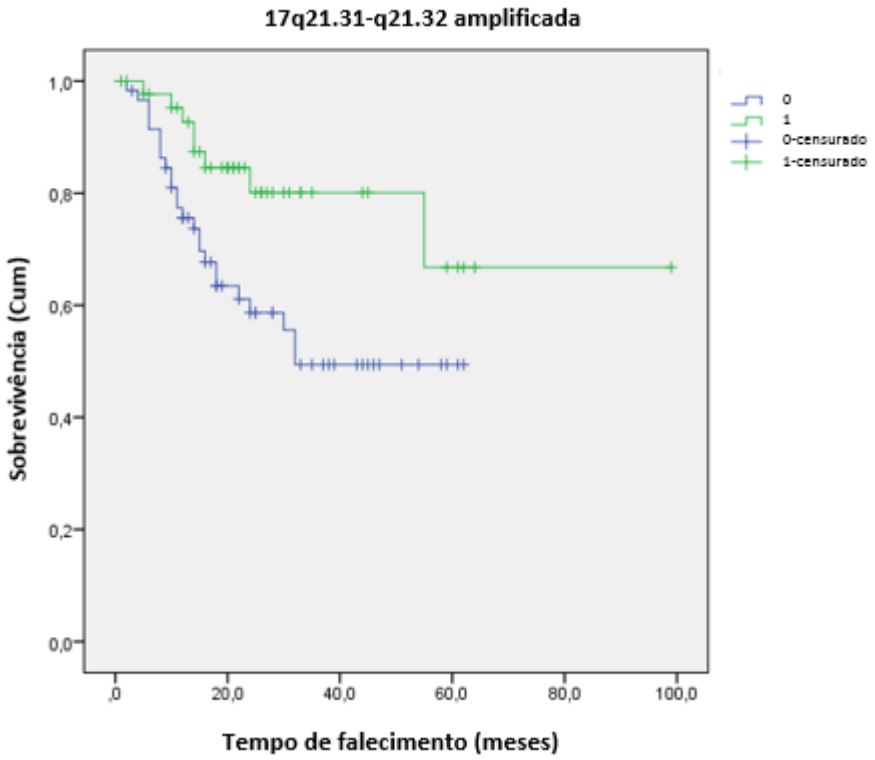


Figura 25 - Gráfico da curva de sobrevivência da região 17q21.31-q21.32 quando amplificada.

A região 22q11.22-q11.23 quando deletada ostenta um tempo de sobrevida 40,667 meses que é mais baixo do que quando a região não apresenta alteração (66,146 meses) mas esta diferença não é muito acentuada, e tem um valor de p de 0,504 o que mostra que este resultado é descritivo (Tabela 23). A curva de sobrevivência mostra que inicialmente quando a região não está alterada tem uma sobrevivência menor mas depois ocorre o inverso e quando deletada o tempo de falecimento em meses é relativamente menor, tal como figurado na Figura 26. Prontamente, esta região expõe um pior prognóstico em concordância com a regressão logística

Tabela 23 - Tempo de sobrevida médio para a região 22q11.22-q11.23 quando deletada.

22q11.22-q11.23 D	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	66,146	5,275	55,807	76,484	0,504
1	40,667	7,116	26,718	54,615	

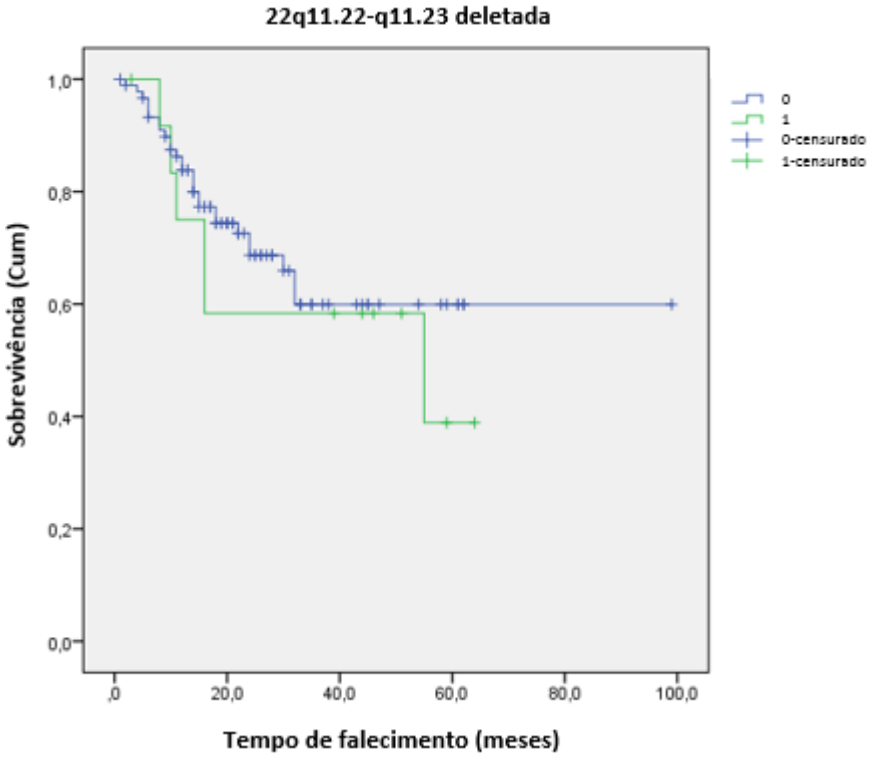


Figura 26 - Gráfico da curva de sobrevivência da região 22q11.22-q11.23 quando deletada.

A região 6q16.1-q16.3 quando deletada aparenta um tempo de sobrevida 27 meses que é mais baixo do que quando a região não apresenta alteração (65,9 meses), e tem um valor de p de 0,027 o que mostra que este resultado é estatisticamente sgnificativo (Tabela 24). A curva de sobrevivência exibe que inicialmente quando a região não está alterada tem uma sobrevivência menor mas depois ocorre o oposto de forma pronunciada e quando deletada a região o tempo de falecimento em meses é relativamente menor, tal como figurado na Figura 27. Dessa forma, esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 24 - Tempo de sobrevida médio para a região 6q16.1-q16.3 quando deletada.

6q16.1-q16.3 D	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	65,921	5,415	55,308	76,535	0,027
1	27,000	8,157	11,013	42,987	

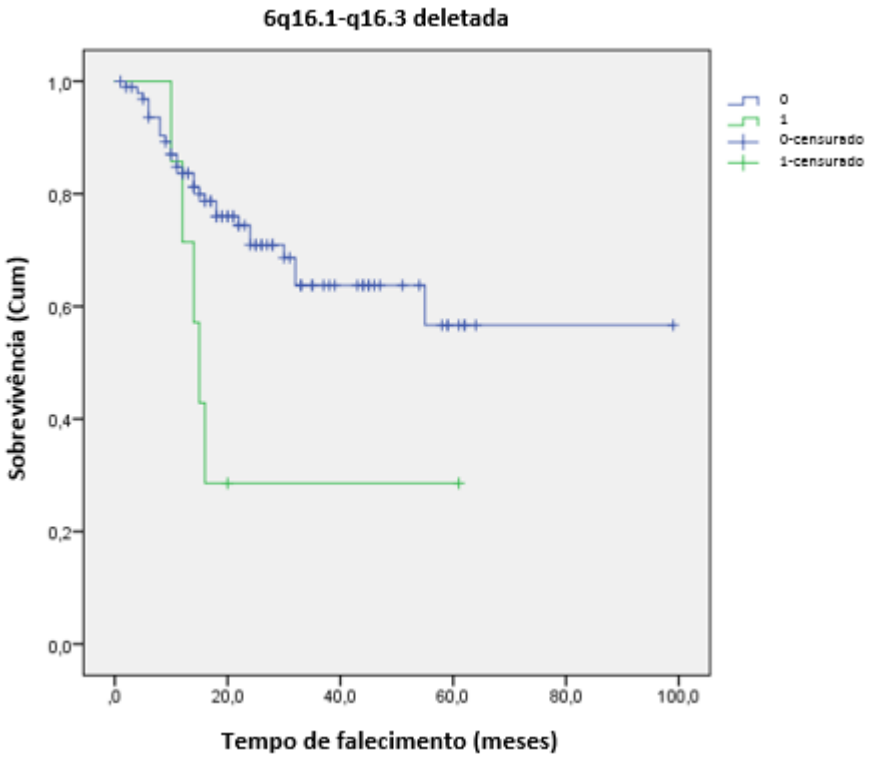


Figura 27 - Gráfico da curva de sobrevivência da região 6q16.1-q16.3 quando deletada.

A região 17q21.31-q21.32 quando deletada exibe um tempo de sobrevida 35,8 meses que é mais baixo do que quando a região não apresenta alteração (68 meses), e tem um valor de p de 0,071 o que mostra que este resultado é descritivo (Tabela 25). A curva de sobrevivência indica que quando a região está alterada tem uma sobrevivência menor e quando deletada o tempo de falecimento em meses é relativamente menor, tal como esquematizado na Figura 28. Portanto, esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 25 - Tempo de sobrevida médio para a região 17q21.31-q21.32 quando deletada.

17q21.31-q21.32 D	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	68,006	6,026	56,196	79,817	0,071
1	35,830	5,144	25,749	45,911	

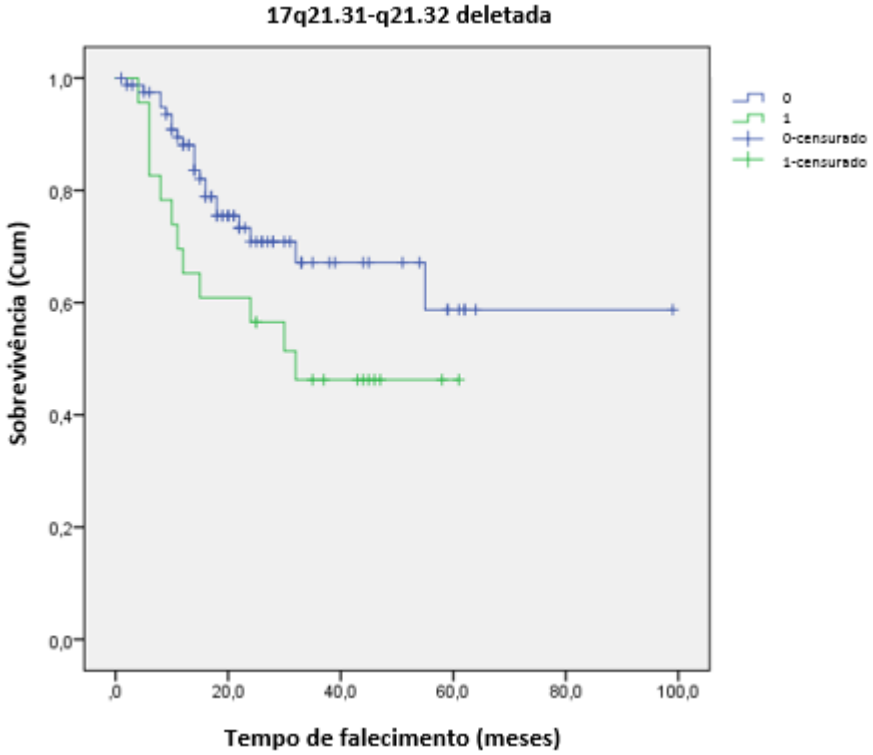


Figura 28 - Gráfico da curva de sobrevivência da região 17q21.31-q21.32 quando deletada.

A região 1q21.1-q21.2 quando amplificada exibe um tempo de sobrevida 74,4 meses que é maior do que quando a região não apresenta alteração (42,8 meses), e tem um valor de p de 0,322 o que mostra que este resultado é descritivo (Tabela 26). E a curva de sobrevivência sugere que quando a região não está alterada tem uma sobrevivência menor e quando amplificada o tempo de falecimento em meses é relativamente maior, tal como ilustrado na Figura 29. Logo, esta região mostra um melhor prognóstico em concordância com a regressão logística.

Tabela 26 - Tempo de sobrevida médio para a região 1q21.1-q21.2 quando amplificada.

1q21.1-q21.2 A	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	42,797	3,174	36,575	49,018	0,322
1	74,422	8,712	57,346	91,497	

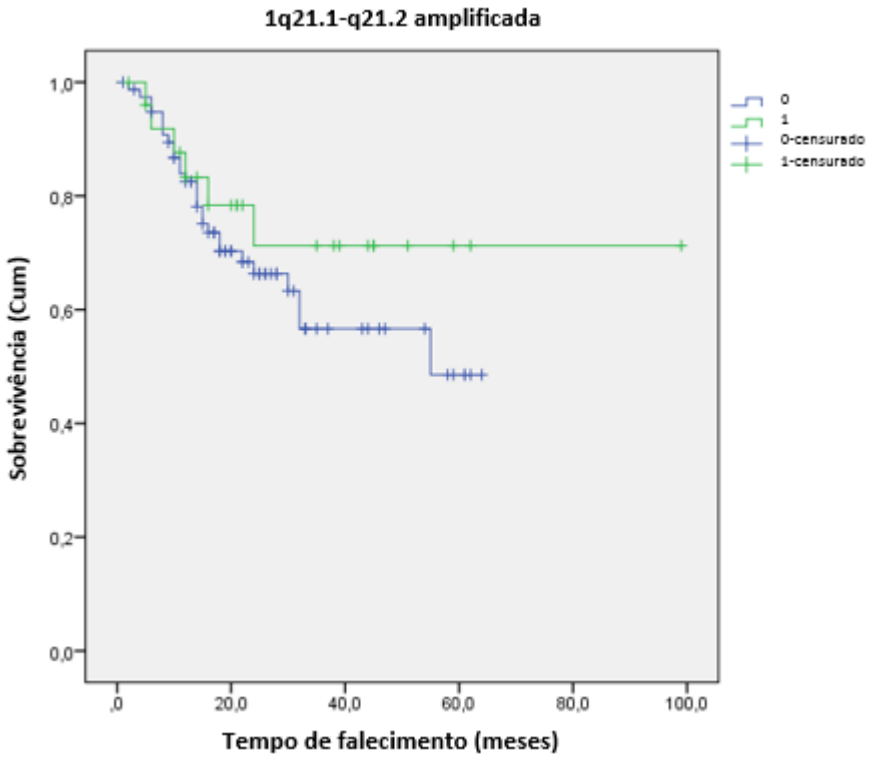


Figura 29 - Gráfico da curva de sobrevivência da região 1q21.1-q21.2 quando amplificada.

A região 3p14.3-p14.2 quando amplificada mostra um tempo de sobrevida 37,7 meses que é menor do que quando a região não apresenta alteração (63,8 meses), e tem um valor de p de 0,769 o que mostra que este resultado é descritivo (Tabela 27). A curva de sobrevivência mostra "períodos" diferentes, ou seja, inicialmente quando a região está alterada tem uma sobrevivência maior, depois ocorre o oposto, de seguida volta a ser maior a sobrevivência quando a região está alterada e por fim volta a acontecer o inverso. E quando a região está amplificada o tempo de falecimento em meses é relativamente menor, tal como esquematizado na Figura 30. Portanto, em suma esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 27 - Tempo de sobrevida médio para a região 3p14.3-q14.2 quando amplificada.

3p14.3-p14.2 A	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	63,758	5,362	53,248	74,268	0,769
1	37,733	11,230	15,723	59,744	

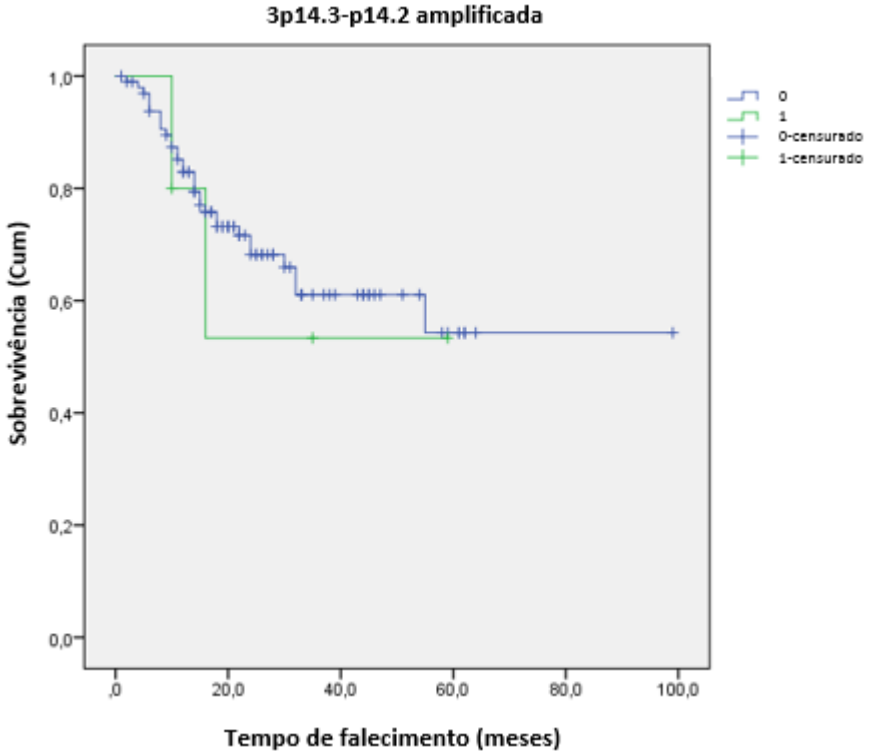


Figura 30 - Gráfico da curva de sobrevivência da região 3p14.3-p14.2 quando amplificada.

A região 3p14.3-p14.2 quando deletada expõe um tempo de sobrevida 41,1 meses que é menor do que quando a região não apresenta alteração (64,1 meses), e tem um valor de p de 0,627 o que mostra que este resultado é descritivo (Tabela 28). A curva de sobrevivência indica "fases" diferentes, ou seja, inicialmente quando a região está alterada tem uma sobrevivência maior, depois ocorre o oposto, e no final, volta a ter uma sobrevivência maior quando alterada a região. E quando a região está deletada o tempo de falecimento em meses é relativamente menor, tal como esquematizado na Figura 31. Pois, somando esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 28 - Tempo de sobrevida médio para a região 3p14.3-q14.2 quando deletada.

3p14.3-p14.2 D	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	64,123	5,881	52,596	75,650	0,627
1	41,119	5,541	30,259	51,978	

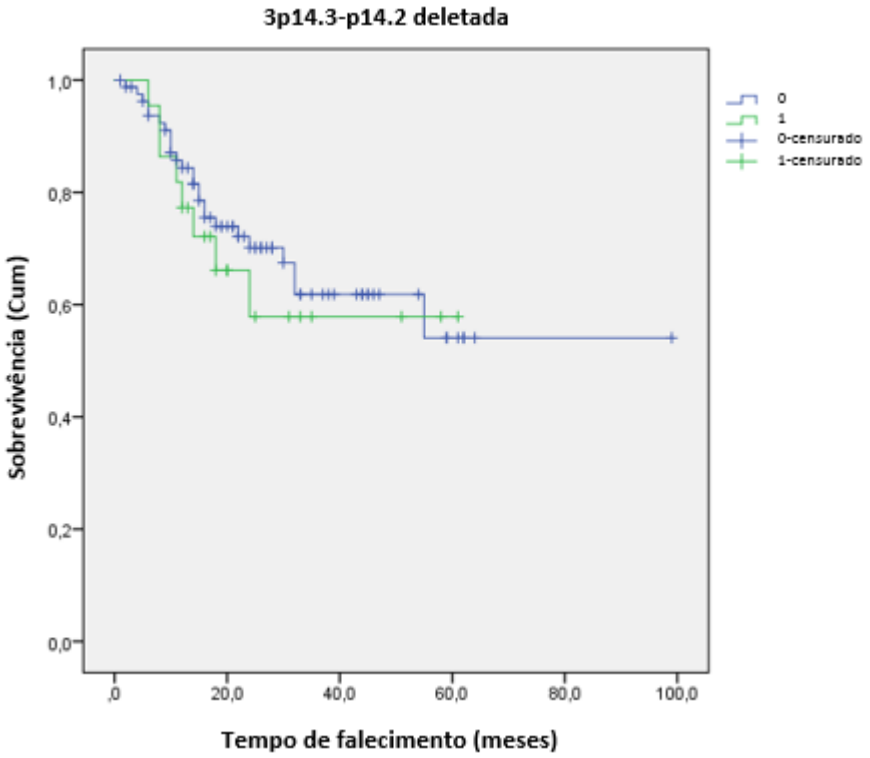


Figura 31 - Gráfico da curva de sobrevivência da região 3p14.3-p14.2 quando deletada.

A região 3q26.31-q26.33 quando deletada exibe um tempo de sobrevida em média 10,8 meses que é menor do que quando a região não apresenta alteração (66 meses), e tem um valor de p de 0 o que mostra que este resultado é estatisticamente significativo (Tabela 29). A curva de sobrevivência assinala inicialmente que quando a região está alterada tem uma sobrevivência maior, e de seguida ocorre o oposto de forma muito acentuada. E quando a região está deletada o tempo de falecimento em meses é muito menor, tal como esquematizado na Figura 32. Logo, esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 29 - Tempo de sobrevida médio para a região 3q26.31-q26.33 quando deletada.

3q26.31-q26.33 D	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	66,018	5,217	55,792	76,244	0
1	10,800	2,143	6,600	15,000	

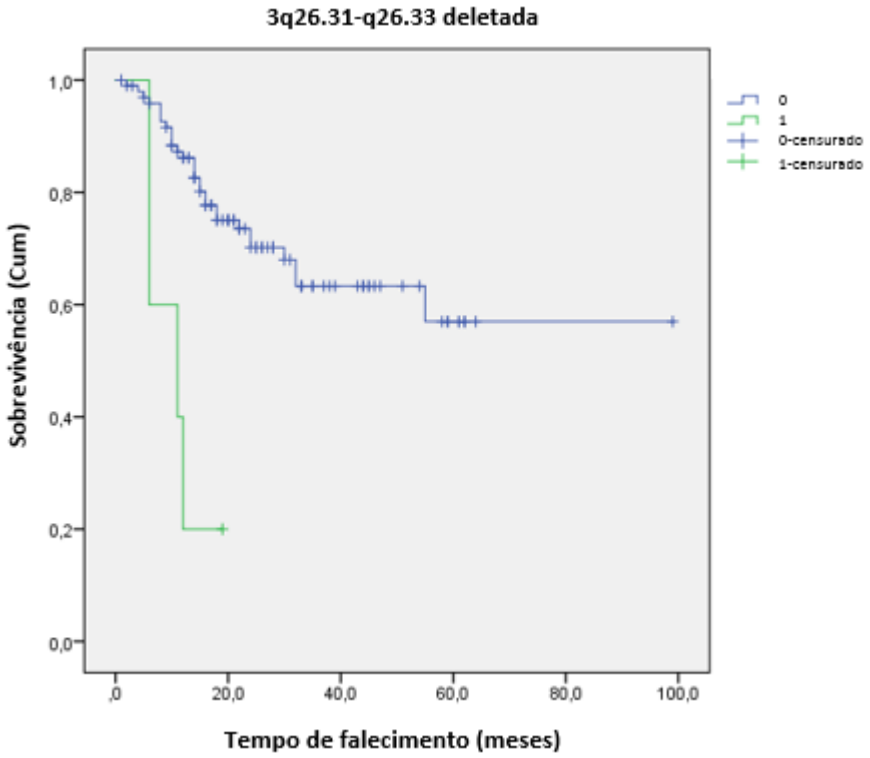


Figura 32 - Gráfico da curva de sobrevivência da região 3q26.31-q26.33 quando deletada.

A região 11p14.1-p13 quando amplificada expõe um tempo de sobrevida 35,6 meses que é menor do que quando a região não apresenta alteração (64,8 meses), e tem um valor de p de 0,355 o que mostra que este resultado é descritivo (Tabela 30). Relativamente à curva de sobrevivência, esta indica que inicialmente quando a região está alterada tem uma sobrevivência maior, e de seguida ocorre o oposto. E quando a região está amplificada o tempo de falecimento em meses é relativamente menor, tal como esquematizado na Figura 33. Assim, esta região expõe um pior prognóstico em concordância com a regressão logística.

Tabela 30 - Tempo de sobrevida médio para a região 11p14.1-q13 quando amplificada.

11p14.1-p13 A	Média (meses)	$\delta_{Média}$	95% IC		Valor de p
			Lower	Upper	
0	64,753	5,524	53,925	75,580	0,355
1	35,600	7,405	21,086	50,114	

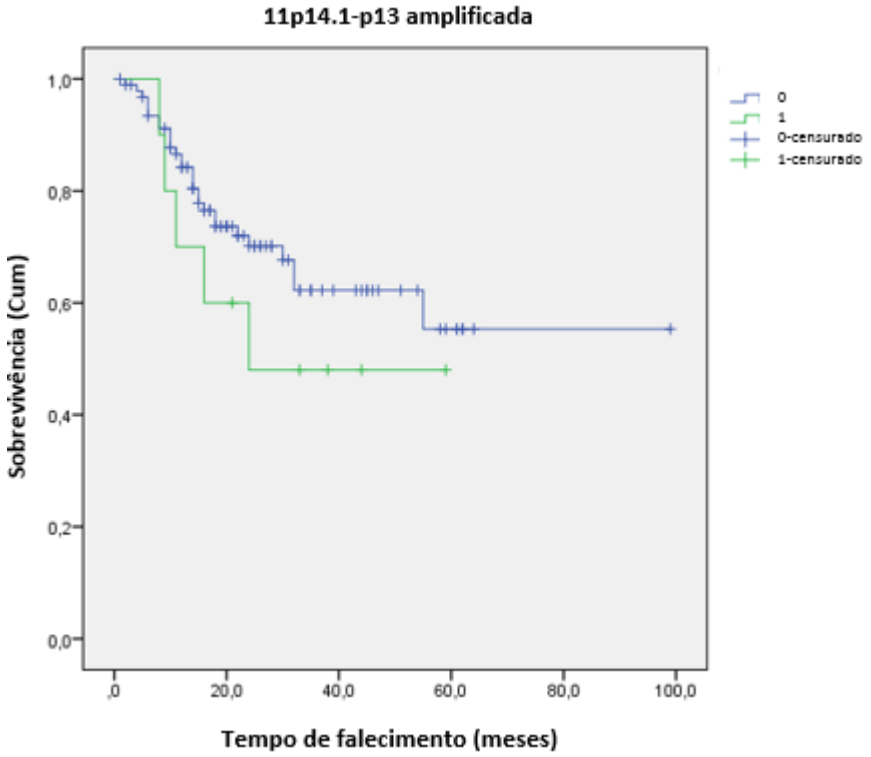


Figura 33 - Gráfico da curva de sobrevivência da região 11p14.1-p13 quando amplificada.

Sumariamente, na Tabela 31 está esquematizada as regiões mais relevantes para o CECP, qual o seu "estado", ou seja, amplificada, deletada ou sem alterações, e qual a sua influência a nível de prognóstico para este tipo de cancro.

Tabela 31 – Influência das regiões mais relevantes relativamente ao prognóstico do CECP.

	Pior Prognóstico	**Melhor Prognóstico**
Amplificada	11q13.5-q14.1 22q11.22-q11.23 3p14.3-p14.2 11p14.1-p13	17q21.31-q21.32 1q21.1-q21.2
Deletada	22q11.22-q11.23 6q16.1-q16.3 17q21.31-q21.32 3p14.3-p14.2 3q26.31-q26.33	

Segundo estudos anteriores sabe-se que as alterações citogenéticas mais comuns no CECP são 3q, 5p, 6p, 7p, 8q, 9q, 11q, 16p, 17p, 17q, 19q, 20q e 11q13 que apresentaram ganhos do número de cópias mais frequentemente, e 2q, 3p, 4q, 5q, 8p, 9p, 11q, 13q, 18q e 21q que apresentaram perdas do número de cópias mais frequentemente, como já referido no ponto de Alterações Citogenéticas.[56, 58-60] E sabe-se também que outros

dados sugerem que a perda de 9p e a amplificação de 11q13 podem ser indicadoras de melhor prognóstico no CECP.[60]

E desse modo este trabalho faz alusão a regiões que têm um papel crucial no prognóstico do CECP, tal como referido na Tabela 31.

Posteriormente num trabalho paralelo realizado pela mesma instituição deste projeto, e recorrendo a uma base de dados disponível *online* em *The Cancer Genome Atlas* (TCGA)[109] comparou-se os dados obtidos por aCGH, que depois foram reduzidos a 11 regiões referidas neste ponto, e com os dados fornecidos por TCGA e alcançou-se duas regiões que apresentavam o mesmo gene nas duas base de dados e que coincidiam na sua função e no seu papel a nível do prognóstico.

Conclui-se assim que a região 3p14.3-p14.2 quando amplificada tem um maior risco de metástases/recidiva do que quando deletada, uma vez que o valor de β obtido na Tabela 19 quando amplificada é 1,769 e quando deletada é 0,483, mostrando assim que esta região amplificada tem maior risco de metástase/recidiva, e por consequente pior prognóstico. E assim, em concordância com a base de dados de TCGA conclui-se que esta região contém o gene *APPL1* (Adaptor protein, phosphotyrosine interaction, PH domain and leucine zipper containing 1), o qual se sabe, segundo GeneCards®, que a sua função é regular a proliferação celular em resposta a sinais extracelulares a parir de um compartimento endossomal precoce.[110] Logo, este gene é um potencial biomarcador/preditor para o CECP.

Também se conclui que a região 22q11.22-q11.23, quer quando está deletada como quando está amplificada, apresenta pior prognóstico, em consonância com os dados obtidos por TCGA. E esta região contém dois genes, *BCR* (Breakpoint Cluster Region), cuja função conhecida é ser uma proteína ativadora de GTPase para *RAC1* (Ras-Related C3 Botulinum Toxin Substrate 1 (Rho Family, Small GTP Binding Protein Rac1)) e *CDC42* (Cell Division Cycle 42), promover o intercâmbio de GTP ligado a *RAC1* ou *CDC42* através de GTP, ativando-os, e exibir atividade de serina/treonina quinase; e *SMARCB1* (SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily b, member 1), cuja função conhecida é ser um componente central do complexo BAF (hSWI/SNF), tendo este complexo remodelado de cromatina dependente de ATP um papel importante na proliferação e diferenciação celular em atividades antivirais celulares e inibe a formação de tumores.[110] Assim, estes dois genes são também potenciais biomarcadores/preditores do CECP.

4.7. Interpretação do gráfico radar

Nesta seção, fazendo uso do algoritmo criado em MATLAB, como referido no capítulo Materiais e Métodos, originou-se o gráfico radar.

Fez-se a identificação de dois grupos, A e B, distintos que têm riscos diferentes, sendo que o grupo B apresenta metástases em *follow up*, em oposição ao grupo A. Desse modo, o grupo B está associado a maior risco de metástases e/ou recidivas e o grupo A associado a menor risco de aparecimento de metástases e/ou recidivas.

É também evidente que cada um dos grupos apresenta um perfil genómico diferente.

Referente ao gráfico radial, este foi dividido em duas áreas semi-circulares, uma representada a vermelho que corresponde a perda e outra representada a azul que condiz com ganho, como evidenciado na Figura 34.

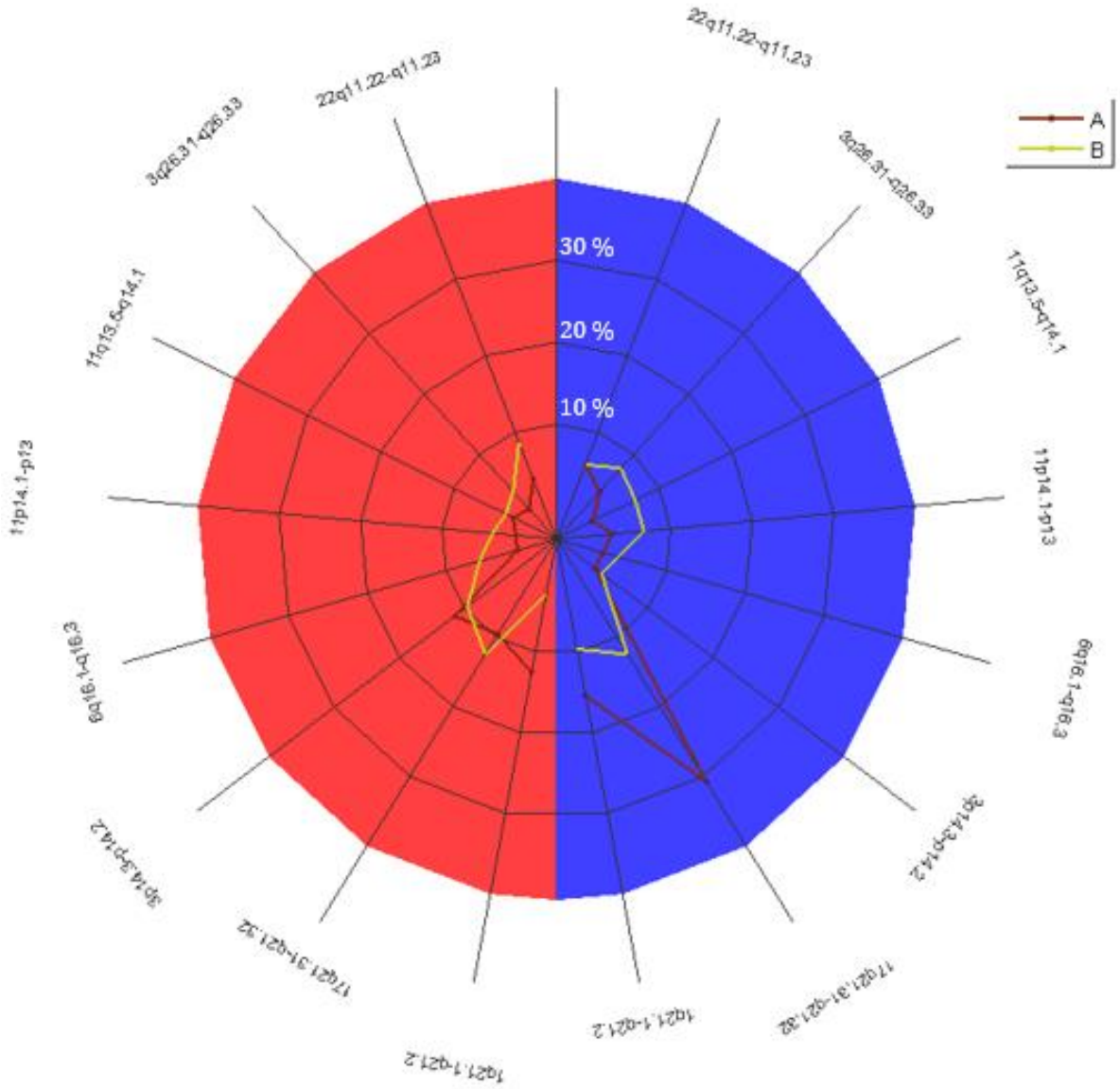


Figura 34 - Gráfico radial das regiões relevantes relativamente ao CECP.

As regiões cromossómicas encontram-se representadas radialmente e em espelho relativamente às áreas semi-circulares.

A região 22q11.22-q11.23 quando está deletada apresenta um maior risco de metástases e/ou recidiva no grupo B, aproximadamente 10%, que no grupo A. Por sua vez, esta região quando está amplificada apresenta igual risco de metástases e/ou recidiva nos dois grupos.

Quanto à região 3q26.31-q26.33, esta ostenta um maior risco de metástases e/ou recidiva no grupo B quer estando ela amplificada ou deletada. Contudo, quando esta região apresenta ganho o grupo B tem um maior risco.

Já a região 11q13.5-q14.1 mostra que quando está amplificada o grupo B tem um maior risco de metástases e/ou recidiva, em oposição a quando a região está deletada, onde o risco é semelhante nos dois grupos e mais baixo que o risco do grupo B da região amplificada.

De seguida a região 11p14.1-p13 mostra um maior risco de metástases e/ou recidiva no grupo B em ambas as situações de ganho e perda, contudo quando ostenta ganho esta região tem um maior risco neste grupo.

Em relação à região 6q16.1-q16.3, o grupo A apresenta risco de metástases e/ou recidiva semelhante quer quando está amplificada ou deletada a região. Por sua vez, o grupo B mostra um maior risco de metástases e/ou recidiva quando apresenta perda.

Quanto à região 3p14.3-p14.2, tanto o grupo A como o grupo B apresenta risco de metástases e/ou recidiva relativamente similar, mas quando a região mostra perda o risco é comparativamente maior do que quando a região mostra ganho.

Por sua vez a região 17q21.31-q21.32 quando apresenta perda relativamente ao grupo B mostra maior risco de metástases e/ou recidiva que o grupo A, sendo esta diferença pouco acentuada. Em contrapartida, a região 17q21.31-q21.32 quando apresenta ganho mostra uma discrepância bastante saliente, a qual mostra maior risco de metástases e/ou recidiva no grupo A, cerca de 30%, que no grupo B. Este resultado foi surpreendente, uma vez que neste caso o grupo que não contem metástases tem maior propensa a desenvolver metástases. E juntando o facto que esta região quando amplificada mostra melhor prognóstico, como percetível na Tabela 22 e Figura 25, pode-se afirmar que quando amplificada a região e apresentando metástases, o risco de metástases e/ou recidiva é mais baixo, indicando melhor prognóstico.

E por fim, a região 1q21.1-1q21.2, apresentando ganho ou perda mostra um risco de metástases e/ou recidiva maior no grupo A que no grupo B, e tal como descrito na região anterior, e segundo a Tabela 26 e Figura 29, esta região quando amplificada tem melhor prognóstico, logo segundo a Figura 34 o grupo B, que tem metástases, tem menor risco de desenvolvimento de mais metástases e/ou recidiva

4.8. Interpretação do ideograma cromossómico

Neste ponto, recorrendo ao algoritmo desenvolvido em Matlab criou-se o ideograma cromossómico, como mencionado no capítulo Materiais e Métodos.

Inicialmente, tendo em conta todas as regiões cromossómicas dos autossomas eliminaram-se as regiões que apresentavam CNVs em menos de 10% dos doentes, como está representado na Figura 35.

Posteriormente, após a redução e seleção das variáveis mais relevantes para o CECP criou-se um novo ideograma cromossómico. Na Figura 36 estão representadas apenas as CNVs contidas nas regiões cromossómicas mais relevantes e que estão presentes em pelo menos 10% dos doentes.

Como se pode constatar pela Figura 36 apenas os cromossomas 1, 3, 17 e 22 apresentaram regiões cromossómicas relevantes no CECP indicando perda e ganho nas regiões 1q21.1-q21.2, 17q21.31-q21.32 e 22q11.22-q11.23, perda na região 3p14.3-p14.2 e ganho na região 3q26.31-q26.33.

Figura 35 - Ideograma cromossómico de todas as CNVs presentes em todos os autossomas em pelo menos 10% dos doentes. A verde do lado direito do cromossoma estão esquematizados os ganhos do número de cópias, e a vermelho do lado esquerdo do cromossoma estão visíveis as perdas do número de cópias.

Figura 36 - Ideograma cromossómico das CNVs presentes nas regiões cromossómicas mais relevantes para o CECP e em pelo menos 10% dos doentes. A verde do lado direito do cromossoma estão esquematizados os ganhos do número de cópias, e a vermelho do lado esquerdo do cromossoma estão visíveis as perdas do número de cópias.

5. CONCLUSÕES

Tendo em conta os propósitos iniciais e com a realização deste trabalho foi possível:

- Estudar a associação entre o conjunto de dados procedentes da técnica aCGH referentes à variação do número de cópias de material genético e a informação clínico-patológica dos doentes com diagnóstico de CECP relativamente a três variáveis dependentes: estádio do tumor, localização do tumor e presença de metástases.
- Identificar as regiões mínimas comuns alteradas mais frequentes nos cromossomas. Após a restrição de variáveis (658 regiões correspondentes aos autossomas), uma vez que se tinham mais variáveis que número de casos, relativamente à variável estádio reduziram-se as variáveis a 16 regiões cromossómicas; à variável localização reduziu-se a 15 regiões cromossómicas; e à variável metástases reduziu-se a 16 regiões cromossómicas.
- Avaliar dois classificadores, Random Forest e SVM, com o intuito de comparar a eficácia de cada um para classificar corretamente cada uma das observações testadas. E para ambos os classificadores foram testadas as variáveis estádio, localização e presença de metástases quanto à sua capacidade para classificar novos casos, e desse modo, concluiu-se que a condição metástases/recidiva foi a que mostrou mais significância em categorizar novos dados.
- Quantificar a importância das variáveis para a classificação em relação às regiões obtidas pelo processo de eliminação de variáveis e responder à questão: Quando é mais relevante a região: amplificada ou deletada? E recorrendo à regressão logística concluiu-se que as regiões 11q13.5-q14.1, 22q11.22-q11.23, 3q26.31-q26.33, 3p14.3-p14.2, 6p25.3-p25.2, 15q26.3 e 11p14.1-p13 quando amplificadas e as regiões 5p15.33-p15.32, 22q11.22-q11.23, 6q16.1-q16.3, 17q21.31-q21.32, 3p14.3-p14.2 e 3q26.31-q26.33 quando deletadas apresentam um pior prognóstico; e as regiões 17q21.31-q21.32, 1q21.1-q21.2, 4p14-p13, 1q31.3-q32.1, 5p15.33-p15.32 e 6q16.1-q16.3 quando amplificadas e as regiões 1q21.1-q21.2, 6p25.3-p25.2 e 1q31.3-q32.1 quando deletadas têm melhor prognóstico.
- Analisar a taxa de sobrevivência em doentes com CECP. Apenas 11 das regiões cromossómicas inicialmente escolhidas foram consideradas clinicamente relevantes. E deste modo verificou-se que as regiões 11q13.5-q14.1 amplificada,

22q11.22-q11.23 amplificada, 3p14.3-p14.2 amplificada, 11p14.1-p13 amplificada, 22q11.22-q11.23 deletada, 6q16.1-q16.3 deletada, 17q21.31-q21.32 deletada, 3p14.3-p14.2 deletada e 3q26.31-q26.33 deletada apresentam pior prognóstico e um tempo de sobrevida menor; e as regiões 17q21.31-q21.32 amplificada e 1q21.1-q21.2 amplificada apresentam melhor prognóstico e uma sobrevivência maior.

- Descobrir potenciais biomarcadores/preditores para o CECP. Em comparação com um estudo realizado em paralelo usando uma base de dados de TCGA duas regiões mostraram concordância nos resultados e concluiu-se que a região 3p14.3-p14.2 quando amplificada tem um maior risco de metástases e contém o gene *APPL1* que segundo a sua função já conhecida é um bom potencial biomarcador/preditor para o CECP; a região 22q11.22-q11.23, quando deletada ou amplificada, apresenta um pior prognóstico e contém os genes *BCR* e *SMARCB1* que são bons candidatos a biomarcadores/preditores para o CECP.

E assim, à questão: será que as diferentes taxas de sobrevivência dos doentes de CECP estão associadas à presença de tumores com diferentes assinaturas genómicas?, a resposta será Sim.

6. PERSPETIVAS FUTURAS

Embora se tenha identificado e quantificado a importância das regiões mais relevantes para o CECP e, posteriormente, determinado potenciais biomarcadores/preditores para esta patologia, futuramente é necessário abranger mais processos de análise de dados e confirmar laboratorialmente. Dessa forma, dever-se-á:

- Melhorar a classificação. Uma vez que os dois classificadores analisados apresentam uma exatidão próxima apenas de 70%, seria vantajoso, por exemplo, combinar classificadores usando a técnica *Boosting*, a qual tenta construir um classificador forte a partir de um conjunto de classificadores fracos;
- Usar uma base de dados maior de forma a estudar os preditores já encontrados, visto que desta forma os resultados seriam mais fidedignos;
- Validar em laboratório os protenciais biomarcadores/preditores (*APPL1*, *BCR* e *SMARCB1*) para o CECP, nomeadamente através da tecnologia de *Touch* FISH (Hibridização *in-situ* de Fluorescência).

7. REFERÊNCIAS BIBLIOGRÁFICAS

1. Stratton MR, Campbell PJ, Futreal PA. ***The cancer genome***. Nature. 2009;458(7239):719-24.

2. Hanahan D, Weinberg RA. ***The hallmarks of cancer***. Cell. 2000;100(1):57-70.

3. Garraway LA, Lander ES. ***Lessons from the cancer genome***. Cell. 2013;153(1):17-37.

4. Macconaill LE, Garraway LA. ***Clinical implications of the cancer genome***. J Clin Oncol. 2010;28(35):5219-28.

5. Kanavos P. ***The rising burden of cancer in the developing world***. Ann Oncol. 2006;17 Suppl 8:viii15-viii23.

6. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. ***Global cancer statistics, 2012***. CA Cancer J Clin. 2015;65(2):87-108.

7. WHO. ***Global Initiative for Cancer Registry Development*** [29-03-2016]. Available from: http://gicr.iarc.fr/en/The-Problem.

8. Hanahan D, Weinberg RA. ***Hallmarks of cancer: the next generation***. Cell. 2011;144(5):646-74.

9. Leemans CR, Braakhuis BJ, Brakenhoff RH. ***The molecular biology of head and neck cancer***. Nat Rev Cancer. 2011;11(1):9-22.

10. Mao L, Hong WK, Papadimitrakopoulou VA. ***Focus on head and neck cancer***. Cancer Cell. 2004;5(4):311-6.

11. Stadler ME, Patel MR, Couch ME, Hayes DN. ***Molecular biology of head and neck cancer: risks and pathways***. Hematol Oncol Clin North Am. 2008;22(6):1099-124, vii.

12. Smeets SJ, Braakhuis BJ, Abbas S, Snijders PJ, Ylstra B, van de Wiel MA, et al. ***Genome-wide DNA copy number alterations in head and neck squamous cell carcinomas with or without oncogene-expressing human papillomavirus***. Oncogene. 2006;25(17):2558-64.

13. Tan M, Myers JN, Agrawal N. ***Oral cavity and oropharyngeal squamous cell carcinoma genomics***. Otolaryngol Clin North Am. 2013;46(4):545-66.

14. Tornesello ML, Perri F, Buonaguro L, Ionna F, Buonaguro FM, Caponigro F. ***HPV-related oropharyngeal cancers: from pathogenesis to new therapeutic approaches***. Cancer Lett. 2014;351(2):198-205.

15. Llewellyn CD, Johnson NW, Warnakulasuriya KA. ***Risk factors for squamous cell carcinoma of the oral cavity in young people--a comprehensive literature review***. Oral Oncol. 2001;37(5):401-18.

16. Haddad RI, Shin DM. ***Recent advances in head and neck cancer***. N Engl J Med. 2008;359(11):1143-54.

17. Argiris A, Karamouzis MV, Raben D, Ferris RL. ***Head and neck cancer***. Lancet. 2008;371(9625):1695-709.

18. Wittekindt C, Wagner S, Mayer CS, Klussmann JP. ***Basics of tumor development and importance of human papilloma virus (HPV) for head and neck cancer***. GMS Curr Top Otorhinolaryngol Head Neck Surg. 2012;11:Doc09.

19. Michaud DS, Langevin SM, Eliot M, Nelson HH, Pawlita M, McClean MD, et al. ***High-risk HPV types and head and neck cancer***. Int J Cancer. 2014;135(7):1653-61.

20. GLOBOCAN. ***GLOBOCAN 2012*** [29-03-2016]. Available from: http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx e http://globocan.iarc.fr/Pages/Map.aspx.

21. Ragin CC, Modugno F, Gollin SM. ***The epidemiology and risk factors of head and neck cancer: a focus on human papillomavirus***. J Dent Res. 2007;86(2):104-14.

22. Polanska H, Raudenska M, Gumulec J, Sztalmachova M, Adam V, Kizek R, et al. ***Clinical significance of head and neck squamous cell cancer biomarkers***. Oral Oncol. 2014;50(3):168-77.

23. Gillison ML, Koch WM, Capone RB, Spafford M, Westra WH, Wu L, et al. ***Evidence for a causal association between human papillomavirus and a subset of head and neck cancers***. J Natl Cancer Inst. 2000;92(9):709-20.

24. Yoshizaki T, Endo K, Ren Q, Wakisaka N, Murono S, Kondo S, et al. ***Oncogenic role of Epstein-Barr virus-encoded small RNAs (EBERs) in nasopharyngeal carcinoma***. Auris Nasus Larynx. 2007;34(1):73-8.

25. Ram H, Sarkar J, Kumar H, Konwar R, Bhatt ML, Mohammad S. ***Oral cancer: risk factors and molecular pathogenesis***. J Maxillofac Oral Surg. 2011;10(2):132-7.

26. Butturini A, Gale RP, Verlander PC, Adler-Brecher B, Gillio AP, Auerbach AD. ***Hematologic abnormalities in Fanconi anemia: an International Fanconi Anemia Registry study***. Blood. 1994;84(5):1650-5.

27. Barnes L, Eveson J, Reichart P, Sidransky DEWHO. ***Classification of Tumours. Pathology and Genetics of Head and Neck Tumours.*** IARC Press. 2005.

28. Nagpal JK, Das BR. ***Oral cancer: reviewing the present understanding of its molecular mechanism and exploring the future directions for its effective management***. Oral Oncol. 2003;39(3):213-21.

29. Pai SI, Westra WH. ***Molecular pathology of head and neck cancer: implications for diagnosis, prognosis, and treatment***. Annu Rev Pathol. 2009;4:49-70.

30. Thompson L. ***World Health Organization classification of tumours: pathology and genetics of head and neck tumours***. Ear Nose Throat J. 2006;85(2):74.

31. Tsantoulis PK, Kastrinakis NG, Tourvas AD, Laskaris G, Gorgoulis VG. ***Advances in the biology of oral cancer***. Oral Oncol. 2007;43(6):523-34.

32. Rothenberg SM, Ellisen LW. ***The molecular pathogenesis of head and neck squamous cell carcinoma***. J Clin Invest. 2012;122(6):1951-7.

33. Neville BW, Day TA. ***Oral cancer and precancerous lesions***. CA Cancer J Clin. 2002;52(4):195-215.

34. Trotta BM, Pease CS, Rasamny JJ, Raghavan P, Mukherjee S. ***Oral cavity and oropharyngeal squamous cell cancer: key imaging findings for staging and treatment planning***. Radiographics. 2011;31(2):339-54.

35. Hassona Y, Scully C, Shahin A, Maayta W, Sawair F. ***Factors Influencing Early Detection of Oral Cancer by Primary Health-Care Professionals***. J Cancer Educ. 2016;31(2):285-91.

36. van der Waal I, de Bree R, Brakenhoff R, Coebergh JW. ***Early diagnosis in primary oral cancer: is it possible?*** Med Oral Patol Oral Cir Bucal. 2011;16(3):e300-5.

37. Kao SY, Lim E. ***An overview of detection and screening of oral cancer in Taiwan***. Chin J Dent Res. 2015;18(1):7-12.

38. Awan K. ***Oral Cancer: Early Detection is Crucial***. J Int Oral Health. 2014;6(5):i-ii.

39. Kugimoto T, Morita K, Omura K. ***Development of oral cancer screening test by detection of squamous cell carcinoma among exfoliated oral mucosal cells***. Oral Oncol. 2012;48(9):794-8.

40. Davies K, Connolly JM, Dockery P, Wheatley AM, Olivo M, Keogh I. ***Point of care optical diagnostic technologies for the detection of oral and oropharyngeal squamous cell carcinoma***. Surgeon. 2015;13(6):321-9.

41. Guerra EN, Acevedo AC, Leite AF, Gozal D, Chardin H, De Luca Canto G. ***Diagnostic capability of salivary biomarkers in the assessment of head and neck cancer: A systematic review and meta-analysis***. Oral Oncol. 2015;51(9):805-18.

42. Lee YH, Wong DT. ***Saliva: an emerging biofluid for early detection of diseases***. Am J Dent. 2009;22(4):241-8.

43. Carreras-Torras C, Gay-Escoda C. ***Techniques for early diagnosis of oral squamous cell carcinoma: Systematic review***. Med Oral Patol Oral Cir Bucal. 2015;20(3):e305-15.

44. Akervall J, Nandalur S, Zhang J, Qian CN, Goldstein N, Gyllerup P, et al. ***A novel panel of biomarkers predicts radioresistance in patients with squamous cell carcinoma of the head and neck***. Eur J Cancer. 2014;50(3):570-81.

45. Begg AC. ***Predicting recurrence after radiotherapy in head and neck cancer***. Semin Radiat Oncol. 2012;22(2):108-18.

46. Langer CJ. ***Targeted therapy in head and neck cancer: state of the art 2007 and review of clinical applications***. Cancer. 2008;112(12):2635-45.

47. Ko C, Citrin D. ***Radiotherapy for the management of locally advanced squamous cell carcinoma of the head and neck***. Oral Dis. 2009;15(2):121-32.

48. Prince A, Aguirre-Ghizo J, Genden E, Posner M, Sikora A. ***Head and neck squamous cell carcinoma: new translational therapies***. Mt Sinai J Med. 2010;77(6):684-99.

49. Dufour X, Beby-Defaux A, Agius G, Lacau St Guily J. ***HPV and head and neck cancer***. Eur Ann Otorhinolaryngol Head Neck Dis. 2012;129(1):26-31.

50. Koontongkaew S. ***The tumor microenvironment contribution to development, growth, invasion and metastasis of head and neck squamous cell carcinomas***. J Cancer. 2013;4(1):66-83.

51. Bussink J, van der Kogel AJ, Kaanders JH. ***Activation of the PI3-K/AKT pathway and implications for radioresistance mechanisms in head and neck cancer***. Lancet Oncol. 2008;9(3):288-96.

52. Coehn J, Chen Z, Lu S, Yang X, Arun P, Ehsanian R, et al. ***Attenuated Transforming Growth Factor β Signaling Promotes Nuclear Factor-κB Activation in Head and Neck Cancer***. Cancer Cell. 2009;69(8):3415-24.

53. Sun W, Gaykalova DA, Ochs MF, Mambo E, Arnaoutakis D, Liu Y, et al. ***Activation of the NOTCH pathway in head and neck cancer***. Cancer Res. 2014;74(4):1091-104.

54. Partridge M, Costea DE, Huang X. ***The changing face of p53 in head and neck cancer***. Int J Oral Maxillofac Surg. 2007;36(12):1123-38.

55. Bockmuhl U, Petersen I. ***DNA ploidy and chromosomal alterations in head and neck squamous cell carcinoma***. Virchows Arch. 2002;441(6):541-50.

56. Ribeiro IP, Marques F, Caramelo F, Pereira J, Patricio M, Prazeres H, et al. ***Genetic gains and losses in oral squamous cell carcinoma: impact on clinical management***. Cell Oncol (Dordr). 2014;37(1):29-39.

57. Ribeiro IP, Marques F, Caramelo F, Ferrao J, Prazeres H, Juliao MJ, et al. ***Genetic imbalances detected by multiplex ligation-dependent probe amplification in a cohort of patients with oral squamous cell carcinoma-the first step towards clinical personalized medicine***. Tumour Biol. 2014;35(5):4687-95.

58. Gollin SM. ***Chromosomal alterations in squamous cell carcinomas of the head and neck: window to the biology of disease***. Head Neck. 2001;23(3):238-53.

59. Gollin SM. ***Cytogenetic alterations and their molecular genetic correlates in head and neck squamous cell carcinoma: a next generation window to the biology of disease***. Genes Chromosomes Cancer. 2014;53(12):972-90.

60. Martin CL, Reshmi SC, Ried T, Gottberg W, Wilson JW, Reddy JK, et al. ***Chromosomal imbalances in oral squamous cell carcinoma: examination of 31 cell lines and review of the literature***. Oral Oncol. 2008;44(4):369-82.

61. Kozaki K, Imoto I, Pimkhaokham A, Hasegawa S, Tsuda H, Omura K, et al. ***PIK3CA mutation is an oncogenic aberration at advanced stages of oral squamous cell carcinoma***. Cancer Sci. 2006;97(12):1351-8.

62. Lotan R, Xu XC, Lippman SM, Ro JY, Lee JS, Lee JJ, et al. ***Suppression of retinoic acid receptor-beta in premalignant oral lesions and its up-regulation by isotretinoin***. N Engl J Med. 1995;332(21):1405-10.

63. Guo T, Califano JA. ***Molecular biology and immunology of head and neck cancer***. Surg Oncol Clin N Am. 2015;24(3):397-407.

64. Jin Y, Jin C, Wennerberg J, Hoglund M, Mertens F. ***Cytogenetic and fluorescence in situ hybridization characterization of chromosome 8 rearrangements in head and neck squamous cell carcinomas***. Cancer Genet Cytogenet. 2001;130(2):111-7.

65. Soni S, Kaur J, Kumar A, Chakravarti N, Mathur M, Bahadur S, et al. ***Alterations of rb pathway components are frequent events in patients with oral epithelial dysplasia and predict clinical outcome in patients with squamous cell carcinoma***. Oncology. 2005;68(4-6):314-25.

66. Szyfter K, Wierzbicka M, Hunt JL, Rinaldo A, Rodrigo JP, Takes RP, et al. ***Frequent chromosomal aberrations and candidate genes in head and neck squamous cell carcinoma***. Eur Arch Otorhinolaryngol. 2016;273(3):537-45.

67. Koontongkaew S, Chareonkitkajorn L, Chanvitan A, Leelakriangsak M, Amornphimoltham P. ***Alterations of p53, pRb, cyclin D(1) and cdk4 in human oral and pharyngeal squamous cell carcinomas***. Oral Oncol. 2000;36(4):334-9.

68. Pande P, Mathur M, Shukla NK, Ralhan R. ***pRb and p16 protein alterations in human oral tumorigenesis***. Oral Oncol. 1998;34(5):396-403.

69. Lane DP. ***Cancer. p53, guardian of the genome***. Nature. 1992;358(6381):15-6.

70. Scully C, Field JK, Tanzawa H. ***Genetic aberrations in oral or head and neck squamous cell carcinoma 2: chromosomal aberrations***. Oral Oncol. 2000;36(4):311-27.

71. Blons H, Laurent-Puig P. ***TP53 and head and neck neoplasms***. Hum Mutat. 2003;21(3):252-7.

72. Loyo M, Li RJ, Bettegowda C, Pickering CR, Frederick MJ, Myers JN, et al. ***Lessons learned from next-generation sequencing in head and neck cancer***. Head Neck. 2013;35(3):454-63.

73. Taby R, Issa JP. ***Cancer epigenetics***. CA Cancer J Clin. 2010;60(6):376-92.

74. Arantes LM, de Carvalho AC, Melendez ME, Carvalho AL, Goloni-Bertollo EM. ***Methylation as a biomarker for head and neck cancer***. Oral Oncol. 2014;50(6):587-92.

75. Das PM, Singal R. ***DNA methylation and cancer***. J Clin Oncol. 2004;22(22):4632-42.

76. Mascolo M, Siano M, Ilardi G, Russo D, Merolla F, De Rosa G, et al. ***Epigenetic disregulation in oral cancer***. Int J Mol Sci. 2012;13(2):2331-53.

77. Li YF, Hsiao YH, Lai YH, Chen YC, Chen YJ, Chou JL, et al. ***DNA methylation profiles and biomarkers of oral squamous cell carcinoma***. Epigenetics. 2015;10(3):229-36.

78. Kaur J, Demokan S, Tripathi SC, Macha MA, Begum S, Califano JA, et al. ***Promoter hypermethylation in Indian primary oral squamous cell carcinoma***. Int J Cancer. 2010;127(10):2367-73.

79. Jones PA, Baylin SB. ***The epigenomics of cancer***. Cell. 2007;128(4):683-92.

80. Stirzaker C, Taberlay PC, Statham AL, Clark SJ. ***Mining cancer methylomes: prospects and challenges***. Trends Genet. 2014;30(2):75-84.

81. PORTO FDEDUD. ***Tipos de Data Mining*** [07-03-2016]. Available from: http://paginas.fe.up.pt/~mgi99021/it/tipos.htm.

82. Cohen W. ***Fast Effective Rule Induction***. Machine Learning: Proceedings of the Twelfth International Conference. 1995(07-03-2016).

83. Tan P, Steinbach M, Kumar V. Introduction to Data Mining. First Edition ed: Addison-Wesley Longman Publishing Co; 2005.

84. Weiss S, Indurkhya N. Predicitve Data Mining: A Practical Guide: Morgan Kaufmann Publishers, Inc.; 1997.

85. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition ed: Springer; 2009.

86. COIMBRA IDSERDUD. ***Aprendizagem Supervisionada e Aprendizagem não Supervisionada*** [07-03-2016]. Available from: http://home.isr.uc.pt/~paulo/PROJ/NN95/node31.html.

87. Duda R, Hart P, Stork D. Pattern Classification. Second Edition ed: John Wiley & Sons; 2000.

88. Lim T, Loh W, Shih Y. ***A Comparison of Predicition Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms*** Machine Learning. 2000(40):203-29.

89. Statistics. M. ***Random Forests*** [07-03-2016]. Available from: https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.html.

90. Breiman L, Cutler A. ***Random Forests*** [07-03-2016]. Available from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

91. R-bloggers. ***'Variable Importance Plot' and Variable Selection*** [07-03-2016]. Available from: http://www.r-bloggers.com/variable-importance-plot-and-variable-selection/.

92. Burges C. ***A Tutorial on Support Vector Machines for Pattern Recognition***. Data Mining and Knowledge Discovery. 1998(2):121-67.

93. R-bloggers. ***How to perform a Logistic Regression in R*** [07-03-2016]. Available from: http://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/.

94. DEPARTMENT OF STATISTICS CUITCONY. ***Generalized Linear Models*** [29-06-2016]. Available from: http://www.stat.columbia.edu/~martin/W2024/R11.pdf.

95. Cabral CIS. Aplicação do Modelo de Regressão Logística num Estudo de Mercado: Universidade de Lisboa; 2013.

96. MEDSTATWEB. ***Avaliação de Testes Diagnósticos*** [07-03-02016]. Available from: http://stat2.med.up.pt/cursop/print_script.php3?capitulo=tdiagnosticos&numero=4&titulo=Avaliacao%20de%20testes%20de%20dianostico.

97. Galen R, Gambino S. ***Beyond normality: the predictive value and efficiency of medical diagnoses***. J Wiley & Sons. 1975:10-40.

98. Silva-Fortes C, Turkman M, Sousa L. ***Curvas ROC degeneradas na análise de genes diferencialmente expressos***. Conferência Estatística e Qualidade na Saúde, EQS. 2011.

99. Botelho F, Silva C, Cruz F. ***Epidemiologia explicada - O valor de prova (p)***. Acta Urológica. 2008;25(3):55-7.

100. Botelho F, Silva C, Cruz F. ***Epidemiologia explicada - Análise de Sobrevivência***. Acta Urológica. 2009;26(4):33-8.

101. Knuutila S, Bjorkqvist AM, Autio K, Tarkkanen M, Wolf M, Monni O, et al. ***DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies***. Am J Pathol. 1998;152(5):1107-23.

102. Shinawi M, Cheung SW. ***The array CGH and its clinical applications***. Drug Discov Today. 2008;13(17-18):760-70.

103. Veltman JA, de Vries BB. ***Diagnostic genome profiling: unbiased whole genome or targeted analysis?*** J Mol Diagn. 2006;8(5):534-7; discussion 7-9.

104. Lucito R, Healy J, Alexander J, Reiner A, Esposito D, Chi M, et al. ***Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation***. Genome Res. 2003;13(10):2291-305.

105. Sharkey FH, Maher E, FitzPatrick DR. ***Chromosome analysis: what and when to request***. Arch Dis Child. 2005;90(12):1264-9.

106. Blaveri E, Simko JP, Korkola JE, Brewer JL, Baehner F, Mehta K, et al. ***Bladder cancer outcome and subtype classification by gene expression***. Clin Cancer Res. 2005;11(11):4044-55.

107. Shaffer LG, Kashork CD, Saleki R, Rorem E, Sundin K, Ballif BC, et al. ***Targeted genomic microarray analysis for identification of chromosome abnormalities in 1500 consecutive clinical cases***. J Pediatr. 2006;149(1):98-102.

108. California Uo. ***UCSC - Genome Browser*** [02-03-2016]. Available from: https://genome.ucsc.edu/.

109. TCGA. ***The Cancer Genome Atlas*** [10-07-2016]. Available from: http://cancergenome.nih.gov/.

110. GeneCards®. ***GeneCards®: The Human Gene Database*** [21-04-2016]. Available from: http://www.genecards.org/.