· U C ·

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Luísa Maria Pais Esteves

Genomic Profiling of Head and Neck Carcinoma for the determination of different predictors of survival rates - *In silico* analysis

Dissertation submitted to the Faculty of Sciences and Technology in partial fulfilment of the requirements for the Master's degree in Biomedical Engineering

Mentors:

Prof. Dr.ª Maria Joana Lima Barbosa de Melo

Prof. Dr. Francisco José Santiago Fernandes Amado Caramelo

Coimbra, 2016

Este trabalho foi desenvolvido em colaboração com:

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

· U C ·

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Laboratório de Citogenética e Genómica da Universidade de Coimbra

· U C ·

FMUC FACULDADE DE MEDICINA
UNIVERSIDADE DE COIMBRA

Laboratório de Bioestatística e Informática Médica

· U C ·

FMUC FACULDADE DE MEDICINA
UNIVERSIDADE DE COIMBRA

Agradecimentos

Gostaria de agradecer à **Professora Doutora Joana Barbosa Melo** e ao P**rofessor Doutor Francisco Amado Caramelo** pelo apoio imprescindível e valiosas críticas construtivas que direcionaram esta dissertação.

À **Ilda Ribeiro** gostaria de agradecer a constante disponibilidade, paciência e atenção. Por mais ocupada, sempre dedicou um momento do seu tempo às minhas questões e uma palavra amiga.

A toda a equipa do Laboratório de Citogenética e Genómica da Faculdade de Medicina da Universidade de Coimbra por me receberem num ambiente de trabalho simpático e acolhedor, ao qual é sempre bom regressar.

Ao **André Santos**, gostaria de agradecer a disponibilidade e ajuda indispensáveis durante a realização deste trabalho.

Às minhas queridas "miguitas" do Lab, **Camila**, **Joana**, **Mariana** e **Nicole** quero agradecer a amizade, toda a paciência e maravilhosas pausas para o café. Apoiaram-me nos momentos mais complicados e ajudaram-me a sorrir quando tudo parecia mal. Sem vocês não teria sido a mesma coisa, foi um ano repleto de momentos felizes.

Aos meus amigos **Beatriz**, **Catarina**, **Nuno** e **João** pelas gargalhadas, horas incontáveis de conversa, paciência e por estarem comigo desde o início. Obrigada por me acolherem tantas vezes no Gabi e por me fazerem ver as coisas pelo lado positivo.

À **Ju** quero agradecer a confiança e a amizade. Mesmo estando longe, os teus e-mails alegraram os meus dias. Obrigada por tudo.

À **Necas** quero agradecer a ajuda e amizade. Obrigada pelos passeios e pelas conversas bem-dispostas que tanto me ajudaram a descontrair.

Ao **Hugo,** cuja dedicação e paciência não conhecem igual. Foste o meu maior apoio, não me deixaste entrar em pânico e tornaste tudo melhor. Obrigada por me confortares nos momentos menos bons e por estares sempre presente nos melhores. Sem ti, este trabalho não teria sido possível.

Aos meus **pais** por todo o apoio, carinho e encorajamento que me dedicaram não só ao longo deste trabalho, mas também ao longo da vida. Obrigada por estarem sempre presentes e por se orgulharem das minhas pequenas conquistas.

Abstract

Head and Neck Cancer (HNC) refers to a group of biologically similar malignancies arising in the upper aerodigestive tract. Approximately 95% of these cancers are Squamous Cell Carcinomas (SCC). Head and Neck Squamous Cell Carcinomas (HNSCC) exhibit extremely malignant phenotypes, frequently with surrounding tissue invasion and distant metastasis, having a 50% five-year survival rate. The detection of these tumors in early stages and the identification of characteristics associated with the prediction of clinical progression, remain challenging matters in clinical practice. The aim of this study was to characterize the genomic profile of HNSCC using copy number data from The Cancer Genome Atlas (TCGA) containing detailed information on 528 patients, in order to contribute to the development of clinical solutions, ultimately seeking to reduce the number of deaths caused by this cancer.

This study identified several genomic alterations consistent with the changes described in literature as being associated to HNSCC. Chromosomes 3, 5, 8, 9 and 11 were the ones that registered copy number alterations in a higher number of patients. In particular, the most frequently amplified regions were located at 8q24.21, 3q22.23, 5p15.33 and 11q13.3 and the most deleted regions were located at 3p21.2, 9p21.3, 8p22.3 and 11q23.2.

Using the clustering algorithm *k-means*, two distinct groups of patients were obtained per chromosome, however no underlying mechanism for the cluster assignments was uncovered.

Certain genes with the possibility of being biomarkers for prognosis were identified. The deletion *APPL1* was found to be statistically significant for the risk of death of HNSCC patients, with 50% of patients that did not present deletion of the *APPL1* loci surviving 1046 days more than those who did, conferring *APPL1* the possibility of application in a clinical context. Along with the deletion of *APPL1*, the deletion of *FER* as well as the non-amplification of *BCR* seem to be biomarkers for worse prognosis in HNSCC patients.

Keywords: Head and Neck Cancer, Head and Neck Squamous Cell Carcinoma, Copy number alterations, The Cancer Genome Atlas, Clustering, Biomarker

Resumo

O Cancro da Cabeça e Pescoço (CCP) refere-se a um grupo de doenças biologicamente semelhantes que aparecem no trato aerodigestivo superior. Aproximadamente 95% destes cancros são Carcinomas de Células Escamosas (CCE). Carcinomas Epidermoides da Cabeça e Pescoço (CECP) apresentam fenótipos extremamente malignos, muitas vezes com invasão de tecidos circundantes e metástases à distância, com uma taxa de sobrevivência a cinco anos de 50%. A deteção destes tumores em estádios iniciais e a identificação de características associadas com a previsão da evolução clínica, permanecem questões difíceis na prática clínica. Assim, o objetivo deste estudo foi a caracterização do perfil genómico de CECP usando dados relativos ao número de cópias do The Cancer Genome Atlas (TCGA) contendo informações detalhadas sobre 528 pacientes, a fim de contribuir para o desenvolvimento de soluções clínicas, em última análise, procurando reduzir o número de mortes causadas por esse tipo de cancro.

Este estudo identificou um conjunto de alterações genómicas consistentes com as modificações descritas na literatura como estando associadas a CECP. Os cromossomas 3, 5, 8, 9 e 11 foram os que registaram alterações no número de cópias num maior número de pacientes. Em particular, as regiões mais frequentemente amplificadas localizam-se em 8q24.21, 3q22.23, 5p15.33 e 11q13.3 e as regiões mais deletadas localizam-se em 3p21.2, 9p21.3, 8p22.3 e 11q23.2.

Usando o algoritmo de *clustering*, *k-means*, dois grupos distintos de pacientes foram obtidos por cromossoma, no entanto nenhum mecanismo subjacente para as atribuições de cluster foi descoberto.

Foram identificados alguns genes com a possibilidade de serem biomarcadores para o prognóstico. Descobriu-se que a deleção do gene *APPL1* foi estatisticamente significante para o risco de morte de pacientes com CECP, com 50% dos pacientes que não apresentaram perda do loci de *APPL1* sobrevivendo 1046 dias mais do que aqueles que a apresentavam, conferindo ao gene *APPL1* a possibilidade de aplicação em contexto clínico . Juntamente com a deleção do *APPL1*, a deleção do *FER*, bem como a ausência de amplificação do *BCR* parecem ser biomarcadores para o pior prognóstico em pacientes com CECP.

Palavras-chave: Carcinoma da Cabeça e Pescoço, Carcinoma Epidermóide da Cabeça e Pescoço, Alteração do Número de Cópias, The Cancer Genome Atlas, *Clustering*, Biomarcador

List of Acronyms and Abbreviations

5-FU	5 fluorouracil
AKT	Protein kinase B
APPL1	Adaptor Protein, Phosphotyrosine Interacting With PH Domain And Leucine Zipper 1
BCR	Breakpoint Cluster Region
CNV	Copy Number Variation
CT	Computed Tomography
E2F	E2 Factor
EGFR	Epidermal Growth Factor
FER	(Fps/Fes Related) Tyrosine Kinase
HNC	Head and Neck Cancer
HNSCC	Head and Neck Squamous Cell Carcinoma
HPV	Human Papillomavirus
IHC	Immunohistochemical
ISH	In situ Hybridization
JAK/STAT	Janus kinase/signal transducer and activator of transcription
LOH	Loss of Heterozygosity
NEK7	NIMA Related Kinase 7
NF- κB	Nuclear factor kappa B
OPSCC	Oropharyngeal Squamous Cell Carcinoma
p16INK4a	Cyclin dependent kinase inhibitor 4A
p53	Cellular tumor antigen p53
PCR	Polymerase Chain Reaction
PET	Positron Emission Tomography
PI3K	Phosphatidylinositol 4, 5 bisphosphate 3 kinase
PI3KCA	Phosphatidylinositol 4, 5 bisphosphate 3 kinase, catalytic subunit alpha
pRB	Retinoblastoma Protein
RB	Retinoblastoma
SCC	Squamous Cell Carcinoma
SCCA	Squamous Cell Carcinoma Antigen

SMARCB1	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily B, Member 1
SNP	Single Nucleotide Polymorphisms
TCGA	The Cancer Genome Atlas
TGF-β	Transforming growth factor β
TNM	Tumor Node Metastasis
TORS	Transoral Robotic Surgery
TP53	Tumor protein 53

List of Figures

List of Tables

Contents

1. Goals

1. Goals

One of the goals of this work is the identification of, at least, two groups of patients of HNSCC from The Cancer Genome Atlas (TCGA) with different disease outcomes, presenting with different loco-regional relapse rates and, consequently be able to associate a genomic profile to each of these groups. Another important goal is the identification of biomarkers/predictors with a possible clinical application, helping with the anticipation of disease progression.

Identification of the most frequently altered chromosomic regions and the minimum common regions for the cohort was also an objective of this work.

It is important to add that this work is incorporated within a project from Laboratório de Citogénica e Genómica da Universidade de Coimbra. Therefore, one of the main goals of this project was to use data from TCGA to validate results from this laboratory.

2. Introduction

2. Introduction

2.1. Cancer Overview

Cancer is one of the leading causes of morbidity and mortality worldwide, in part because of the population's growth and aging and the prevalence of risk behaviors like alcohol and tobacco consumption, inactivity, bad eating habits and the changing of reproductive patterns. According to GLOBOCAN, in 2012 there were 14.1 million new cancer cases and 8.2 million cancer-related deaths [1] .

All the cells that constitute the human body are direct descendants of the fertilized egg that originates it and, consequently, they all carry a copy of its diploid genome. Nevertheless, a cancer cell genome possesses a set of acquired alterations that were not present in the fertilized egg, termed somatic mutations, different from germline mutations inherited from progenitors and passed on to the next generation [2].

Regarding the genomic instability of cancer, researchers settled two classes of genes that, when mutated, were directly linked to cancer development: the oncogenes (derived from proto oncogenes that suffer mutations with dominant gain of function) and the tumor suppressors genes (in which mutations lead to a recessive loss of function).[3] The genetic alteration mechanisms involve several genomic alterations such as nucleotide substitution or deletion, copy number alterations in chromosomes and DNA rearrangements.[4]

Thus, in humans, tumorigenesis is a multistep process arising from genetic aberrations that culminates in the transformation of normal cells into malignant ones.[5]

Despite the differences between cancer types they all share the same essential modifications in cell physiology that lead to tumor formation, as suggested by Hanahan and Weinberg : self sufficiency in growth signals, insensivity to growth inhibitory signals, evasion of apoptosis, infinite replicative potential, sustained angiogenesis, tissue invasion and metastasis [5].

2.2. Head and Neck Squamous Cell Carcinoma

Head and Neck Cancer (HNC) refers to a group of biologically similar cancers that originate in a variety of subsites: oral cavity (including the lip, oral tongue, maxilla, floor of the mouth, buccal mucosa, gingiva, retromolar trigone and hard palate), nasal cavity/paranasal sinuses, pharynx (nasopharynx, oropharynx and hypopharynx), larynx, thyroid, trachea and salivary gland (Figure 1). About 95% of HNC are squamous cell carcinomas (SCC), arising in the mucosal lining of the upper aerodigestive tract [6-8].

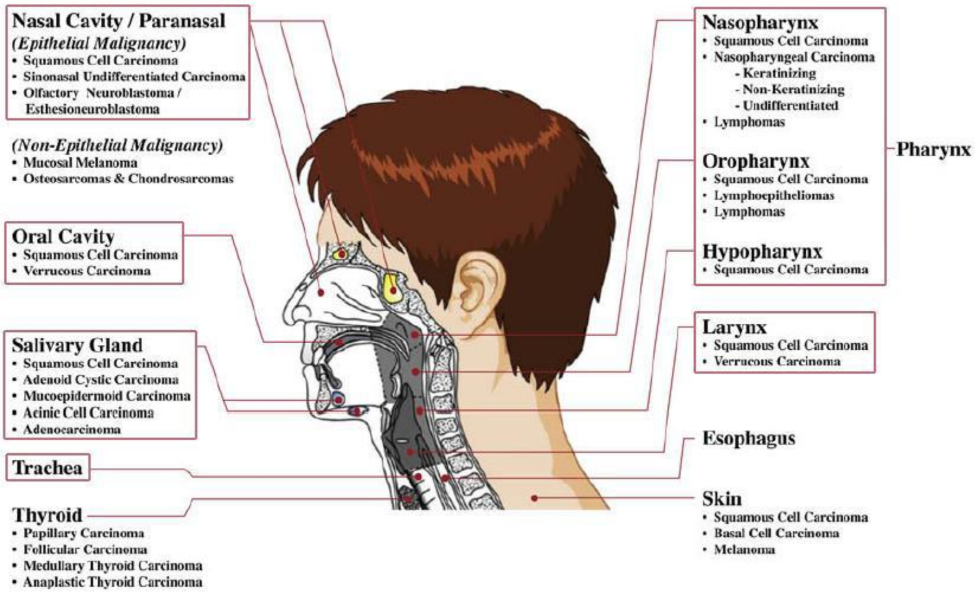
Nasal Cavity / Paranasal
(Epithelial Malignancy)
• Squamous Cell Carcinoma
• Sinonasal Undifferentiated Carcinoma
• Olfactory Neuroblastoma / Esthesioneuroblastoma
(Non-Epithelial Malignancy)
• Mucosal Melanoma
• Osteosarcomas & Chondrosarcomas
Oral Cavity
• Squamous Cell Carcinoma
• Verrucous Carcinoma
Salivary Gland
• Squamous Cell Carcinoma
• Adenoid Cystic Carcinoma
• Mucoepidermoid Carcinoma
• Acinic Cell Carcinoma
• Adenocarcinoma
Trachea
Thyroid
• Papillary Carcinoma
• Follicular Carcinoma
• Medullary Thyroid Carcinoma
• Anaplastic Thyroid Carcinoma
Nasopharynx
• Squamous Cell Carcinoma
• Nasopharyngeal Carcinoma
- Keratinizing
- Non-Keratinizing
- Undifferentiated
• Lymphomas
Oropharynx
• Squamous Cell Carcinoma
• Lymphoepitheliomas
• Lymphomas
Hypopharynx
• Squamous Cell Carcinoma
Pharynx
Larynx
• Squamous Cell Carcinoma
• Verrucous Carcinoma
Esophagus
Skin
• Squamous Cell Carcinoma
• Basal Cell Carcinoma
• Melanoma

Figure 1- Diversity of head and neck cancer and histopathologic diagnosis that present at the various subsites in the head and neck. HNSCC regions are marked. Adapted from: Stadler et al. (2008) [8]

Head and Neck Squamous Cell Carcinoma (HNSCC) patients exhibit extremely malignant phenotypes, most of which present invasion of surrounding tissue and distant metastasis, even at early stages and a 50% five-year survival rate [6, 7]. In recent years, numerous efforts have been directed towards the development of new strategies for detection, diagnosis and treatment of HNSCC as well as the improvement of already existing ones.[6]

It is generally assumed that, during the past two decades there has been a significant improvement in life quality of HNSCC patients, mainly due to the use of advanced surgical and radiotherapeutic techniques and organ preservation protocols. However, no increase in five year survival rates has been registered in recent decades, mainly because of the frequent development of metastasis, loco regional recurrences and second primary tumors [7].

There is also a limited knowledge about the molecular pathways that lead to HNSCC carcinogenesis. The fact that HNSCC is a genetically and biologically heterogeneous disease has hindered the many efforts to precisely prognosticate, treat and identify the cancer genes that are behind its origin. [7]

2.2.1. Epidemiology

Approximately 560,000 new cases of HNSCC and 300,000 deaths are reported annually worldwide. In 2012, 300,400 new cases and 145,400 deaths from oral cavity cancer and 86,700 new cases and 50,800 deaths from nasopharyngeal carcinoma alone, were reported [1, 6].

Across the last decade, the primary site distribution of HNSCC has shifted to a steady increase of oropharyngeal squamous cell carcinoma (OPSCC) and a decline in larynx and hypopharynx cancer. This tendency has been observed in parallel with a decline in cigarette smoking and the identification of human papillomavirus (HPV) as a risk factor for OPSCC development. [9]

For most HPV-negative HNSCCs, the age at diagnosis is over 60 years and for HPV-positive HNSCC's it is under 60 years of age. [7]

The highest rates of oral cavity cancer are found in Melanesia, South Central Asia, Central and Eastern Europe and the lowest are found in Western Africa and Eastern Asia.[1] Portugal is also a part of the highest incidence group, according to Globocan 2012 (Figure 2). [10]

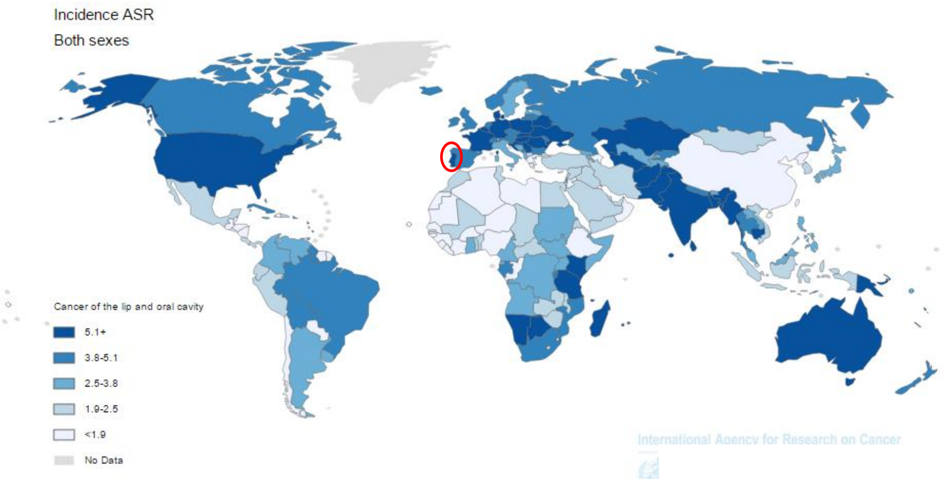
Incidence ASR
Both sexes
Cancer of the lip and oral cavity
5.1+
3.8-5.1
2.5-3.8
1.9-2.5
<1.9
No Data
International Agency for Research on Cancer

Figure 2 - Schematic representation of the worldwide incidence o *s with Portugal highlighted in red. Adapted from GLOBOCAN 2012 [10]*

2.2.2. Etiology

Several risk factors related to HNSCC carcinogenesis have been described, including tobacco use, alcohol consumption, human papillomavirus (HPV) and other viral infections, syphilis, oro dental factors, dietary factors and chronic candidiasis. [7, 11]

<u>Tobacco Use</u>

Tobacco smoking or chewing are lifestyle behaviors that have been linked to oral cancer development, accounting for 25% of these cancers worldwide. [12]

Tobacco smoking is a very prominent HNSCC risk factor, correlated with the intensity and duration of exposure. This risk markedly increases when patients have been smoking for more than 20 years and the daily smoked cigarettes exceed 20 units. [12]

Interruption of this habit reduces the risk of developing HNSCC but it does not necessarily eliminate it. Consequently, risk among former smokers is steadily lower than among current smokers, with a trend of decreasing risk with the number of years since quitting. Passive smoking also seems to increase the risk of cancer development, even for people who have never smoked actively. [12, 13]

Tobacco smoke contains some carcinogenic substances such as nitrosamines and polycyclic hydrocarbons which have genotoxic effects. For example, *TP53* (tumor protein 53) mutations are more frequent in HNSCC patients that smoke. [13] *TP53* works mainly as a transcription factor, being known for its tumor suppressor action. It has two major roles: cell cycle arrest and initiation of apoptosis after genotoxic stress. Loss of *TP53* function through mutations of the *TP53* gene, leads to cellular transformation. [14]

While heavy smokers are at higher risk of developing HNSCC, light smokers do not seem to exhibit higher risk than non-smokers, in the absence of coexisting risk factors. [12]

<u>Alcohol Consumption</u>

Although heavy alcohol drinking is recognized as a risk factor for HNSCC, it is most relevant when in combination with tobacco smoking, because of its ability to synergistically magnify the effects of tobacco smoke. [12]

Alcohol in itself is not a carcinogen, however, acetaldehyde (one of its metabolites) forms DNA adducts that affect DNA synthesis and repair. Moreover, alcohol’s nature as a chemical solvent may increase and extend the exposure of the mucosa to carcinogens from tobacco smoke. [13]

Other Lifestyle Factors

It has been suggested by multiple sources that nutrition plays a crucial role in the development of HNC, particularly in the case of oral cancer. The ingestion of vegetables and fruits, rich in antioxidants and anti-carcinogenic substances like vitamins A, C and E, carotenoids, fibres, phytosterols, folates and flavonoids, may help counterbalance the effects of other carcinogens like tobacco and alcohol. [12]

Also regarding oral cancer, poor oral hygiene associated to dental sepsis seems to play a part in the tumorigenesis process. It is also believed that the carcinogenic effect of tobacco may be potentiated by lack of oral hygiene. [15]

Human Papillomavirus (HPV)

HPVs are a group of small heterogeneous DNA viruses that cause a variety of proliferative epithelial lesions, in some specific body sites. More than 50 HPV genotypes have been found to infect the human mucosa, of which type 16 in particular has been associated with HNC. About 90% of HPV-related HNC are linked to the presence of HPV16 while other HPV genotypes have a prevalence below 5%. [16]

HPV16 is also a proven etiological agent in up to 70% of oropharyngeal cancers. This is becoming particularly relevant in a younger nonsmoking nondrinking demographic, mainly due to the shifting of sexual patterns in these populations. In fact, risk factors like tobacco and alcohol do not seem to appreciatively contribute to the HPV mediated carcinogenesis of the oropharynx. [13]

HPV – positive HNSCC are specific clinical entities in regard to treatment response and survival outcome. [16] HPV mediates carcinogenesis mainly through the products of its E6 and E7 viral oncogenes, that interact an inactivate p53 and retinoblastoma (RB) respectively. [7] E6 oncoprotein targets p53 tumor suppressor for proteasomal degradation in a ubiquitin dependent way, compromising p53 –induced cell cycle arrest and apoptosis, while E7 oncoprotein can bind to and inhibit the retinoblastoma tumor suppressor (pRb).This leads to accumulation of free E2F(E2 Factor) in the cell, inducing the expression of S phase genes, causing an increase of cyclin dependent kinase inhibitor p16 ($p16^{INK4a}$) and aberrant cell proliferation (***Figure 3***). [11]

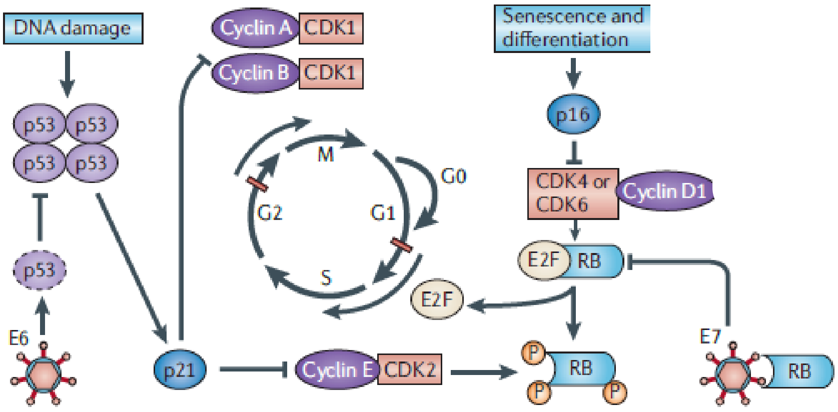

Figure 3 - HPV-mediated cell cycle deregulation. Adapted from Leemans et al. [7]

HPV induced p16 overexpressing HNSCC have a generally better prognosis when compared to HPV negative HNSCC. [11]

2.2.3. Histology and Progression

Undoubtedly, oral squamous cell carcinomas are the most studied HNCs in term of their pathogenesis. Oral precursor lesions are the most frequently diagnosed out of premalignant lesions in all HNSCC cancers. [7]

The oral cavity is lined by a stratified squamous epithelium, whose interface with the underlying lamina propria is outlined by a basement membrane that regulates differentiation and migration of epithelial cells and serves as a barrier to invasion during tumorigenesis. [13]

Early squamous cell carcinoma often presents as a white lesion (leukoplakia), red patch (erythroplakia) or a mixed red and white lesion (erythroleukoplakia). Later, superficial ulceration of the mucosal surface may be developed. These lesions increase the probability of a tumor development. [17]

Early squamous cell carcinoma progresses from a benign squamous hyperplasia through the stages of squamous dysplasia to invasive squamous cell carcinoma (***Figure 4***). [13]

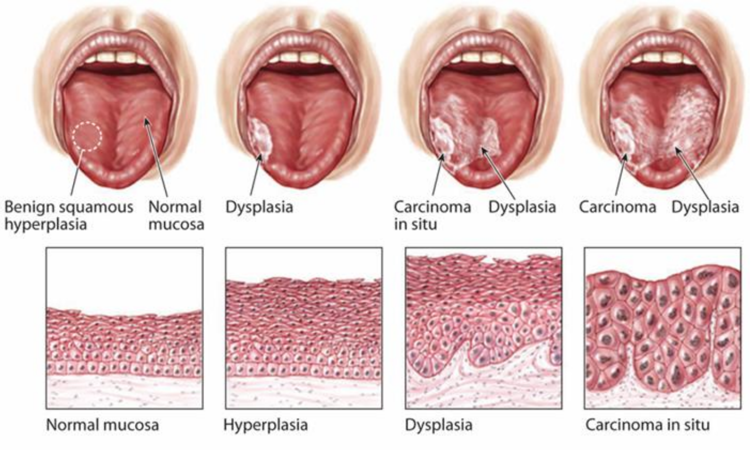

Figure 4 - Histological progression of oral cancer, a type of HNSCC. Adapted from Pai et al. [13]

Squamous dysplasia, the neoplasic alterations in the surface epithelium, includes abnormal cellular organization, an increase in mitotic activity and nuclear enlargement.

These alterations are usually graded on an atypia scale: atypia limited to the lower third of the epithelium is termed mild dysplasia, atypia limited to the lower two thirds of the epithelium is classified as moderate dysplasia and atypia that involves the full thickness of the epithelium is considered severe dysplasia/carcinoma *in situ*. [13]

If allowed to progress the carcinoma *in situ* breaks through the basement membrane and infiltrates the underlying connective tissue. As it grows, in more advanced stages, through lymphatic spaces and perineural invasion, the tumor overruns skeletal muscle, bones and skin. [13]

2.2.4. Changes in Signaling Pathways

HNSCC cancer genes play major roles in at least four important functional pathways: cellular proliferation, squamous epithelial differentiation, cell survival, and invasion/metastasis. [18]

Limitless replicative potential: p53 and RB pathways

To overcome senescence and obtain limitless replicative potential, cancer cells exhibit cell cycle alterations. In HNSCC, crucial genes involved in the regulation of the cell cycle that are targeted by mutations or by HPV oncogenes are those encoding proteins in the p53 and RB pathways. [7]

TP53 is a tumor suppressor gene that is commonly mutated in HNSCC, being one of the earliest identified genetic alterations in this type of cancer and occurring in about half of all cases. The inactivation of *p53* stimulates cellular proliferation and also originates abnormal responses to DNA damage. Data studies point to the downregulation of the p53 pathway in 60-80% of all HNSCC cases.[7, 18]

The retinoblastoma protein (pRb) plays a pivotal role in the negative control of the cell cycle and in tumor progression. It has been shown that pRb is responsible for a major G1 checkpoint, blocking S phase entry and cell growth. In HNSCC, the inactivation of this protein promotes cell cycle progression and together with the alterations in the p53 pathway, these changes lead to cellular immortalization. [18]

Terminal differentiation – NOTCH pathway

NOTCH signaling has been linked to multiple biological functions, such as regulation of self-renewal capacity, cell cycle exit and cell survival. Some NOTCH family mutations have been detected in HNSCC and several of those encode inactivating mutations, suggesting a tumor suppressor function. The NOTCH pathway is involved in squamous epithelium terminal differentiation promotion: the *NOTCH* gene is inhibited in the basal epithelial cells by the transcription factor p63 that becomes downregulated during terminal differentiation coincident with NOTCH1 upregulation. Reactivation of p63 expression was observed in the dysplasic stage of HNSCC and overexpression and/or genomic amplification of *TP63* (tumor protein 63), that induces p63 activation, was observed in the majority of invasive HNSCCs. [18]

Cell Survival: EGFR and PIK3CA Pathways

The PI3K (phosphatidylinositol 4, 5 bisphosphate 3 kinase) signaling pathway is often activated in HNSCC. PI3Ks are a family of enzymes that play roles in numerous cellular processes such as apoptosis, proliferation, cell cycle progression, cytoskeletal stability, motility, and metabolism. Many of these functions are related with the activation of protein kinase B (AKT) by PI3Ks. The class Ia PI3Ks, most frequently associated with cancer, are heterodimers coupled to receptor tyrosine kinases such as EGFR or adaptor molecules that may become active after receptor phosphorylation. [7, 8, 18]

Epidermal growth factor receptor (EGFR) signaling has been strongly implicated in carcinogenesis, tumor progression, and response to therapy in HNSCC. [8] The EGFR is a transmembrane receptor tyrosine kinase of the ErbB (epidermal growth factor receptor) family. After activation by ligand binding, EGFR forms a dimer and activates downstream pathways such as PI3K (phosphoinositide 3 kinase), AKT, JAK/STAT (Janus kinase/signal transducer and activator of transcription) and Ras. These pathways are involved in proliferation, evasion of apoptosis, invasion, angiogenesis and metastasis.[19] Overexpression of EGFR in HNSCC is associated with aggressive and treatment resistant tumors with poor prognosis. [20]

Adhesion and invasion signaling: TGF β/SMAD

Another important alteration associated with HNSCC is the inactivation of the transforming growth factor β (TGF β) pathway. TGF β1 signals through the TGF β receptors and these transduce the signal by phosphorylating SMAD proteins. Downregulation of TGF β receptors is often found in HNSCC tumors. This might be related to the recurrent loss of chromosome 18q in HNSCC patients, which contains the *SMAD* and *TGF β* receptor genes. The inactivation of this pathway components is associated with tumor initiation and TGF β1 is linked to metastasis development, in the absence of functional TGF β receptor. [7, 18]

Recently, it was reported that abrogation of the TGF β pathway was linked to the activation of nuclear factor κB (NF-κB), a transcription factor involved in cell survival. [7]

2.2.5. Detection and Methods of Diagnosis

Head and neck carcinoma is frequently preventable and those that are diagnosed early have a good prognosis, often being curable. Patients, however, normally are diagnosed with advanced carcinomas which are incurable or require aggressive therapy.[21] Thus it is essential for HNSCC to be diagnosed at an early stage in order to improve the clinical outcome of the patients.

Early oral cancers and precursor lesions are often very subtle and asymptomatic. Invasive oral squamous cell carcinoma is often preceded by clinically identifiable premalignant changes of the mucosa: leukoplakia and erythroplakia, followed by superficial ulceration. Later stage symptoms include bleeding, loosening of teeth, dysphagia and growth of a neck mass. [17] Enlargement of cervical lymph nodes is common in certain sites, like the tongue and the nasopharynx. [21]

Generally, since the oral cavity is easily accessible, the most common method for oral cancer screening is oral examination, which is a subjective test dependent on the clinician's experience and skill. Therefore, currently the main method for identification and diagnosis of malignant disorders is the biopsy of the suspicious tissue, followed by histopathological analyses. This method also encompasses some limitations since it is invasive, expensive and its results suffer from variability related to the observer. [22]

Some imaging techniques like nasopharyngolaryngoscopy, computed tomography (CT), and magnetic resonance are routinely used to identify the extent of the disease and to help with staging HNSCC. Positron emission tomography (PET) has been useful in the detection of small primary tumors and nodal disease undetectable through palpation examination.[9]

The available methods for early diagnosis also include brush biopsy, toluidine blue staining, auto fluorescence and spectroscopy. [23]

HPV positive OPSCC's constitute a different biological and molecular entity when compared to HPV negative OPSCC's. Immunohistochemical (IHC) analysis of the tumor site for $p16^{INK4A}$ is being used as the preferred initial test to identify high risk HPV infection. There are also methods of direct viral DNA and RNA identification: polymerase chain reaction (PCR) an in situ hybridization (ISH). Even though these methods are not yet completely established as regular clinical procedures, they can be applied to formalin fixed paraffin embedded biopsy samples. [9]

The most promising method for HNSCC early detection and diagnosis is the use of molecular biomarkers. A biomarker is defined as a biochemical, molecular or genetic parameter that can be objectively measured and evaluated in order to access the presence and progress of disease. In the past, biomarkers were primarily used as prognostic tools for HNSCC patients. Recently, biomarkers have been used to address multiple disease related aspects, such as early detection and diagnosis, staging, therapy planning and follow up surveillance. [13]

Although considered to be at risk of progression to malignant HNSCC, premalignant lesions of the upper aero digestive tract can be confused with nonneoplasic reactive processes at a histopathological level. Some biomarkers can be useful in these situations. Amongst those, LOH (loss of heterozygosity) at determined chromosomal loci seem to be the most promising: studies have shown that LOH at 3p and 9p are successfully able to establish that distinction. [13]

The detection of genomic, transcriptomic, proteomic and metabolomic biomarkers in bio fluids such as saliva and serum has been regarded as a very promising procedure, especially given its noninvasive nature. A recent systematic review by Guerra et al (2016) suggested that a combination of serum biomarkers resulted in diagnostic values with higher sensitivity and specificity than when the biomarkers were tested independently. The combined biomarkers with higher diagnostic capability were *EGFR* + *CCND1* and *SCCA* + *EGFR* + *CCND1*. [23] The same group has also shown similar results for saliva derived biomarkers. Salivary biomarkers appear to detect early stages of HNSCC better than the advanced ones. A set of salivary single biomarkers (interleukin 8, choline, pipecolinic acid, l phenylalanine, and S carboxymethyl l cysteine) as well as in combination, demonstrated excellent diagnostic test accuracy. [24]

Despite the advances in diagnostic techniques, the existing detection methods of HNSCC in early stages remain insufficient, therefore there is a global need for new methodologies, both less invasive and more accurate.

2.2.6. Staging

Cancer staging is essential for establishing proper treatment and determining prognosis. The anatomical location of HNSCC is important for their clinical classification, as the head and neck region comprises a variety of anatomic sites. More importantly, these tumors have diverse clinical behaviors and outcomes. For example, tumors in the hypopharynx have a higher probability of metastasizing compared to tumors in the oral cavity or larynx. [25, 26] The most commonly diagnosed HNSCCs are those located in the oral cavity and in the oropharynx. Survival of patients with these types of cancer is strongly related to the stage of disease at diagnosis. [17]

Tumors of the oral cavity and oropharynx are staged anatomically according to the TNM system, where T stands for the size of the primary tumor, N represents the status of lymph node spread and M indicates the presence or absence of distant metastases. [17, 27] However, patients in the same stage of the disease may show different responses to the same treatment and different clinical outcomes. [26] The TNM system used is shown in Table 1.

Table I - *TNM Classification of HNSCC of the oral cavity and oropharynx. Adapted from Trotta et al. [27]*

Primary Tumor of Oral Cavity (T)	
Tx	Primary tumor cannot be assessed
T0	No evidence of primary tumor is seen
Tis	Primary tumor is carcinoma *in situ*
T1	Primary tumor has a maximal diameter of 2 cm or less
T2	Primary tumor has a maximal diameter of more than 2 cm but no more than 4 cm
T3	Primary tumor has a maximal diameter of more than 4 cm
T4	
Lip	Primary tumor involves cortical bone, inferior alveolar nerve, floor of the mouth, skin
Oral cavity	Primary tumor involves cortical bone, intrinsic or extrinsic muscles of the tongue, maxillary sinus, skin
T4b	Primary tumor involves lateral pterygoid muscle, pterygoid plates, lateral nasopharynx, skull base, carotid artery
Primary Tumor of Oropharynx (T)	
Tx	Primary tumor cannot be assessed
T0	No evidence of primary tumor is seen
T1	Primary tumor has a maximal diameter of less than 2 cm
T2	Primary tumor has a maximal diameter of 2–4 cm
T3	Primary tumor has a maximal diameter of more than 4 cm
T4a	Primary tumor involves the larynx, intrinsic or extrinsic muscles of the tongue, medial pterygoid, hard palate, mandible
T4b	Primary tumor involves lateral pterygoid muscle, pterygoid plates, lateral nasopharynx, skull base, carotid artery
Regional Metastasis (N)	
Nx	Regional lymph nodes cannot be assessed
N0	No regional lymph node metastasis is evident
N1	Ipsilateral single enlarged node with a maximal diameter of less than 3 cm
N2a	Ipsilateral single enlarged node with a maximal diameter of 3–6 cm
N2b	Ipsilateral multiple enlarged nodes with a maximal diameter of less than 6 cm
N2c	Bilateral or contralateral enlarged nodes with a maximal diameter of less than 6 cm

Table I *(Continuation)* ***-*** *TNM Classification of HNSCC of the oral cavity and oropharynx. Adapted from Trotta et al. [27]*

N3	Enlarged node with a maximal diameter of more than 6 cm
Distant Metastasis (M)	
M0	No distant metastasis is evident
M1	Distant metastasis is evident

T, N and M categories may be combined in 32 different ways, which result into seven different stages of HNSCC: 0, I, II, III, IVA, IVB and IVC, represented in Table 2. [27]

Table II - *Oral Cavity and Oropharyngeal SCC Staging based on TNM Classification. Adapted from Trotta et al. [27]*

Stage	T Category	N Category	M Category
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	Any T4b	N3 Any	M0 M0
IVC	Any	Any	M1

2.2.7. Therapy

Several therapeutic approaches are used in the management of HNSCC. The most used of those are surgery, radiotherapy, chemotherapy, or a combination of two or more of these modalities, depending on TNM stage and primary site. [9, 23]

Early stage disease (stage I and II) is usually treated with surgery or radiation alone. Most patients with locally advanced disease (stage III and IVA/B) are treated with platinum based chemoradiation with or without chemotherapy as a sequential therapy. Metastatic disease is treated with combination chemotherapy for patients with good performance status and single agent chemotherapy for patients with poor performance status.[9]

All of these treatments have some level of toxicity, possibly leading to late organ dysfunction. [9]

Surgery

Curative surgery for HNSCC is used for resectable tumors in which clear margins can be achieved and function is preserved. Classic open surgery or minimally invasive surgery like transoral robotic surgery (TORS) or laser surgery can be employed, depending on the particular characteristics and anatomical location of the tumor. [9] Primary site, location, size, proximity to bone, and depth of infiltration are factors that influence a particular surgical approach.[28]

Radiotherapy

Radiotherapy consists in the use of high energy radiation from x rays, gamma rays, neutrons, protons, and other sources to destroy cancer cells and reduce tumor size. [29] This practice is very important for the management of early stage and locally advanced (HNSCC) either alone or combined with surgery and/or chemotherapy. [30] Radiotherapy is either used as an alternative to surgery in small primary tumors, to reduce tumor size prior to surgery in operable large tumors, after surgery to remove tumor cells that may be left at the tumor site or as palliative treatment in incurable carcinomas. [31]

Chemotherapy

Chemotherapy is a treatment that stops the growth of cancer cells by the administration of anti-cancer drugs. In HNSCC, the most used drugs for this purpose are cisplatin, 5 fluorouracil (5 FU), carboplatin and bleomycin, generally used in combinations of two. Chemotherapy has proven effective for HNSCC, especially when used in conjunction with radiation therapy or surgery. [9, 31] More recently, a three drug combination of taxane added to cisplatin and 5 fluorouracil has become the standard regimen for chemotherapy in HNSCC. A study reported a significant improvement in progression free survival and overall survival for the taxane triple drug regimen. [9]

Target Therapy

A novel understanding of the molecular genetic behind HNSCC has made possible the development of new therapies targeting specific cell membrane growth factors or downstream signaling pathway mutations. In HNSCC, *EGFR* is a particularly interesting target, since its overexpression in HNSCC is generally linked to high loco regional recurrence and low survival rates. EGFR acts as a central transducer of multiple signaling pathways that are involved in tumor cell growth, invasion and angiogenesis.[9, 13]

Targeted therapy has been especially directed to blockade EGFR function. The most generally applied strategy is the use of monoclonal antibodies, namely Cetuximab, directed against the extracellular receptor domain of EGFR blocking ligand binding and, consequently, preventing ligand dimerization and activation and triggering antitumoral immune responses. [13]

Cetuximab has been successfully used in combination with radiotherapy and chemotherapy, improving locoregional control and overall survival in in patients with locoregionally advanced HNSCC.[13]

2.2.8. Genomic and Cytogenetic Alterations

The initiation and development of HNC is a multistep process involving the progressive acquisition of genetic and epigenetic aberrations. Early studies of head and neck tumorigenesis showed that genetic impairment often precedes microscopic modifications and the number of acquired genetic alterations increases from squamous hyperplasia through dysplasia to invasive carcinoma, roughly following a sequential order. It was also shown that chromosomal loss and gain favor genetic pathways that regulate cell growth, motility and stromal interactions. [13]

Concerning HNSCC, cytogenetic analysis of solid tumors can be difficult to attain due to various factors, such as low mitotic index and small specimen size. HNSCC tumors remain difficult to culture, with, reportedly only 30% of these growing in culture and yielding analyzable metaphase spreads. [32]

HNSCC develops, like other tumors, as a result of dysregulation of oncogenes, tumor suppressor genes and DNA damage response genes. The karyotypes of HNSCC typically are complex, near triploid, containing multiple clonal numerical and structural chromosome aberrations. [32] The most common structural alterations present in HNSCC include deletions, translocations and isochromosomes followed by less common alterations like duplications, insertions, inversions, ring chromosomes, endoreduplication and dicentric chromosomes. [20]

Chromosomal rearrangements are one of the most easily recognizable features of cancer and the mechanism behind the most common ones is genomic instability influenced by accumulating DNA damage, defective DNA damage repair and replication stress. Critical components of carcinogenesis are copy number variations (CNV) – deletions and duplications, and allelic loss or loss of heterozygosity (LOH) – the loss of one allele of a gene for which the other allele is already inactivated. [7, 33]

CNVs are regions of genetic structural variation observed between two or more genomes larger than 1 kilobase (kb) in size that can involve gains or losses of genomic DNA. These regions can be microscopic or submicroscopic and, therefore, are not easily visible by standard G banding karyotyping. [34]

CNVs affect a greater fraction of the genome than single nucleotide polymorphisms (SNP). High resolution SNP arrays have allowed CNV identification. The characterization of germline CNVs has helped understand the susceptibility to various diseases and somatic CNVs are being used in the identification of genome regions involved in pathogenic phenotypes, such as in cancer. Although some cancer cases are the result of genetic predisposition that increase an individual’s risk, cancer is a somatic genetic disease. [32, 34]

The most frequent cytogenetic alterations in HNSCC are gains at 3q, 5p, 7p, 8q, 11q, and 20q and losses at 3p, 4q, 5q, 8p, 9p, 11q, 13q, 18q, and 21q. [20, 35]

The correlation between copy number alteration of multiple genes, tumor progression and clinical outcome suggest that these genes and/or the proteins they encode and their interactions in the pathways in which they take part may be targets for early detection methods or therapeutic intervention.

Region 3q

Gains in the long arm of chromosome 3 have been reported as some of the most frequent chromosomal alterations in HNSCC, being present in 72% of HNSCC.[20] Gains of 3q25.29 are associated with lower survival rates.[36] Gains in this region are related with overexpression of cancer related genes such as *TP63* (tumor protein p63), *CCNL1* (cyclin L1), and *PIK3CA* (phosphatidylinositol 4, 5 bisphosphate 3 kinase, catalytic subunit alpha). [20]

PIK3CA is an oncogene mapped at 3q26.3. The overexpression of this gene leads to uncontrolled cell growth, cell invasion, drug resistance and metastasis. This oncogene has been associated with HNSCC, since its amplification has been detected in precursor oral dysplasia, reportedly being altered in 6% to 29% of HNSCC. *PIK3CA* copy number amplification has been connected to cancer relapse and poor prognosis in patients without lymph node metastasis. [20, 37]

CCNL1, mapped at 3q25.32 is thought to be involved in HNSCC progression, as it was shown to be amplified in 26% of cases and overexpressed in 57% of tumors, by Muller and colleagues. [38] *CCNL1* (isoform α) is thought to be a regulator of the G0 to G1 cell cycle transition. *CCNL1* amplification has been linked with lymph node metastasis and with an advanced clinical stage, being also associated with shorter overall survival. [20, 38]

TP63 is a homolog of *TP53* which shows copy number gain and overexpression in HNSCC, and is linked to poor survival rates in OSCC patients. [20]

The copy number gain and overexpression of various genes on distal 3q is associated with tumor development, poor prognosis and aggressive clinical course in HNSCC. [20]

Region 3p

Loss of the short arm of chromosome 3 can be detected in 56% to 78% of oral dysplasias and in more than 90% of OSCC, making it one of the earliest and most frequent changes in HNSCC. TCGA described loss of one or more segments of 3p in 71% of its HNSCC. 3p loss is either mediated by isochromosome formation or chromosome breakage, most frequently at 3p14. [20] The most common losses in 3p include *FHIT* (fragile histidine triad) and *RARB* (retinoic acid receptor, beta) tumor suppressor genes. [39]

3p14 is the site of *FHIT* gene and the most common chromosomal fragile site, FRA3B. It has been suggested that inactivation of *FHIT* is important for progression of HNSCC and that loss of expression of Fhit protein causes DNA damage and genome instability. [20, 39] Regarding *RARB*, it was verified that this gene was not expressed in 60% of potentially malignant oral lesions. The loss of this gene could enhance carcinogenesis through loss of response to retinoids. [39]

Region 7p

HNSCC cell lines demonstrate gain of 7p region, specifically in the copy number of 7p12 22. *EGFR* is coded at 7p12 and is amplified in 10% of TCGA HNSCC and overexpressed in about 90% of HNSCC. *EGFR* overexpression is caused by gene copy number increase, gene amplification, increased mRNA synthesis, decreased downregulation or expression of EGFRvIII, an active truncated form of the protein present in almost 50% of HNSCC. *EGFR* ligands are also overexpressed in HNSCC. *EGFR* plays a critical role in HNSCC growth, invasion, metastasis and angiogenesis. [19, 20]

Region 8q

One of the most frequent copy number alterations in HNSCC, present in 74% of tumors, is gain of 8q involving bands 8q23.1-q24.22. 8q gain is an early change present in oral dysplasia, and is frequently due to isochromosome formation. [20]

MYC (v myc avian myelocytomatosis viral oncogene homolog) is an oncogene located at 8q, that is overexpressed in HNSCC because of gene amplification or copy number gain and that has been associated with poor survival rates. [39]

PTK2 (protein tyrosine kinase 2) is mapped at 8q24 and is overexpressed in HNSCC, being associated with the invasive potential of the tumor. [39]

LRP12 (low density lipoprotein receptor related protein 12) and *WNT1* (wingless type MMTV integration site family, member 1) are also overexpressed in HNSCC. [39]

Region 8p

Loss of the short arm of chromosome 8 is detected in 58% of HNSCC, with nearly half of all HNSCC showing allelic loss of 8p23.2. This band maps *CSMD1* (CUB and Sushi multiple domains 1) which expression is abnormal in several HNSCC as a consequence of deletion, epigenetic silencing or aberrant splicing. Loss of 8p23 in HNSCC is an established predictor of poor prognosis, shortened disease free interval and low survival rates. [20]

Region 9p

One of the most frequent genetic changes in HNSCC is losses in band 9p21. 9p loss occurs via isochromosome formation for 9q as well as 9p multigene deletions. Genes in this region include *PTPRD* at 9p23 24 and *CDKN2A*, *CDKN2B*, and *MTAP* at 9p21.3. [20]

PTPRD is a receptor protein phosphatase that is very important in cellular signaling and inhibition of tumor cell growth and that is lost in about 50% of TCGA HNSCC. *PTPRD* deletions or mutations can drive tumor growth by hyper activation of its substrate, STAT3, an important transcription factor in HNSCC. [20]

CKN2A encodes p16$^{\mathrm{INK4a}}$ protein, which is important to cell cycle regulation due to its interaction with Rb (retinoblastoma) protein. [40] Loss of *CDKN2A* (p16) gene was reported in 59% of TCGA HNSCC. Deletions, somatic mutations of *CDKN2A* and promoter hyper methylation result in *CDKN2A* inactivation in about 80% of HNSCC.[20]

The reported genetic alterations can lead to uncontrolled cell proliferation, by loss of cell cycle checkpoint control leading to tumorigenesis. [20]

Region 11q

11q13 amplification is an early change in HNSCC, playing an important part in the transition from moderate to severe dysplasia. The core of the 11q13 amplicon contains 13-14 genes where all but three or four are overexpressed in HNSCC tumors. *CCDN1* oncogene is considered the most important oncogenic driver of this amplicon. *CCDN1* plays an important role in promoting G1/S cell cycle transition and the overexpression of this gene leads to a faster transition from G1 to S. The overexpression of *CCDN1* protein is associated with disease recurrence, lymph node involvement and reduced overall survival. [20]

Along with 11q13 amplification OSCC cancer cell lines have also been shown by Jin et al. (2002) and later, Martin et al. (2008) to present distal 11q loss. [41, 42] Amplification of 11q13 with distal 11q loss was found to occur more frequently in tumors of the tongue, retromolar trigone and buccal mucosa. Various groups reported a correlation between 11q13 amplification/distal 11q loss (11q22-qter) and decreased patient survival, which further validates the use of 11q13 amplification/distal 11q loss as a biomarker for patient prognosis.[20]

Region 13q

A large number of HNSCC cases present loss of 13q, especially 13q12.11 and 13q14.2 bands. *RB1* is a gene encoded in region 13q that plays a crucial role in cell cycle arrest and control and its loss is associated with development of tumors.[20]

ING1 (inhibitor of growth family, member 1) is mapped at 13q.34 and mutations on this gene lead to uncontrolled cell growth that may be associated with tumor development. [20]

Region 18q

Loss of 18q is a common alteration in HNSCC, especially loss of 18q23 band. This occurrence is related to advanced tumor stages and poor prognosis. *GALR1* (galanin receptor 1) and *PARD6G* (par 6 family cell polarity regulator gamma) are both affected by this loss of genetic material. *GALR1* is mapped at 18q23 and is frequently lost in HNSCC as a result of promoter methylation. *GALR1* is a G protein coupled receptor that is important in the inhibition of cell proliferation [20].

PARD6G deletion affects interphase and spindle microtubule organization and it also leads to defects in centrosome organization and function. [20]

2.3. Databases for Cancer Genomics Study: Overview and Relevance

Accumulating evidence has stated that cancer is a disease of the genome. With the development of high-throughput sequencing technologies, previous one-by-one studies to explore the molecular mechanisms of cancer have been left behind. Recently, cancer research has become more dependent on data sharing and the systematic study of the cancer genome (Figure 5). Data are available from various platforms for the complete genome sequences of different cancer types, allowing for a wider accessibility to a global view on cancer.

Many web-based cancer genomics databases are in operation nowadays, several of those supplying, along with the data, web-tools and resources. Although very valuable to the progression of knowledge on cancer, these platforms are not without their liabilities: they depend greatly on the collaboration of others and the complexity and overwhelming quantity of information makes it challenging for computational methodologies to be applied successfully. [43]

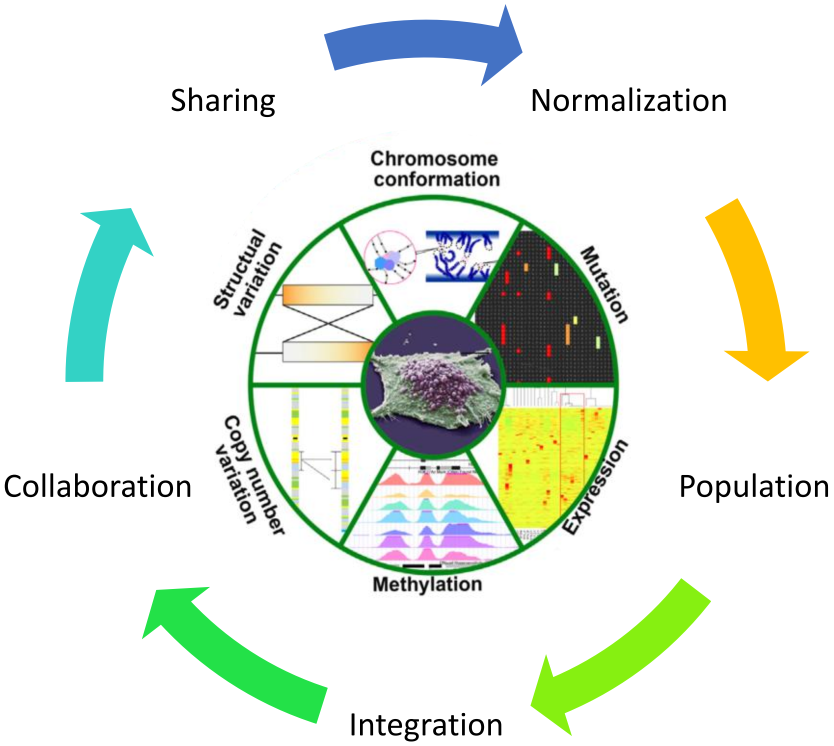

Figure 5 -The future of cancer research lies on making full use of the data coming from heterogeneous sources, including genomics, metabolomics and proteomics data and a vast collection of clinical information. It will depend largely on the effort to obtain representative data for the population, use advanced data mining algorithms and adequate collaboration and sharing of the information. Adapted from Yang et al. (2015) [43]

Various data portals in existence have led to the identification of recurrent point mutations, translocations and a great number of potential therapeutic targets in various different cancer subtypes. Many researchers have been seeking to translate this data into clinical applications, which is made possible with the help of emerging complex computational technologies.

2.3.1. The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA) is a United States collective project involving the National Cancer Institute (NCI) and the National Human Genome Research Institute

(NHGRI) with the major objective of “understanding the molecular basis of cancer through the application of genome analysis technologies, including large-scale genome sequencing” in order to “improve our ability to diagnose, treat, and prevent cancer”. [44] The complex genetic nature of cancer has made it a subject of continued interest in finding genetic pathways and chromosomic aberrations in its origin. This is enabled by the extensive mapping of different types of cancer, including HNSCC.[45, 46]

In order to generate molecular profiles for each tumor type, TCGA used different methodologies including whole genome and exon sequencing, SNP genotyping, CNV profiling using microarrays, DNA methylation profiling, genome-wide expression, functional proteomic analysis, and microRNA (miRNA) expression profiling through RNA sequencing. [45]

With regard to HNSCC, TCGA has identified trademarks of HPV-associated and tobacco-associated cancer sub-types, helping to improve the classification of this cancer type. In 2009, Parfenov and colleagues described how HPV integration affects the host’s genome, by amplification of oncogenes and disruption of tumor suppressors as well as leading to chromosomal rearrangements.[47] They also have determined that non-HPV HNSCCs were different entities: they possess different gene expression profile and DNA methylation patterns. [47] In 2015, Lawrence and associates established a comprehensive genomic characterization of HNSCC tumors from TCGA, showing that HPV-associated tumors were dominated mutations of the oncogene *PIK3CA*, loss of *TRAF3*, and amplification of *E2F1*, and that smoking-related HNSCCs demonstrated loss-of-function *TP53* mutations and *CDKN2A* inactivation with frequent amplification of 3q26/28 and 11q13/22. They have also identified a number of potential genetic therapeutic candidates, including *PIK3CA* and *TP53*. [48]

2.3.1.1. SNP Microarray - DNA Copy Number Analysis in HNSCC

The CNV data from TCGA for HNSCC tumors was obtained using Affymetrix Genome-Wide Human SNP Array 6.0 (Figure 6). [49]

The presence of single base polymorphisms described as SNPs – naturally occurring germline point mutations with a minor allele frequency of at least 1% in a given population - is the most common source for genetic variation in the human genome. [50] High-resolution SNP microarrays have been used for the detection of CNVs. [51]

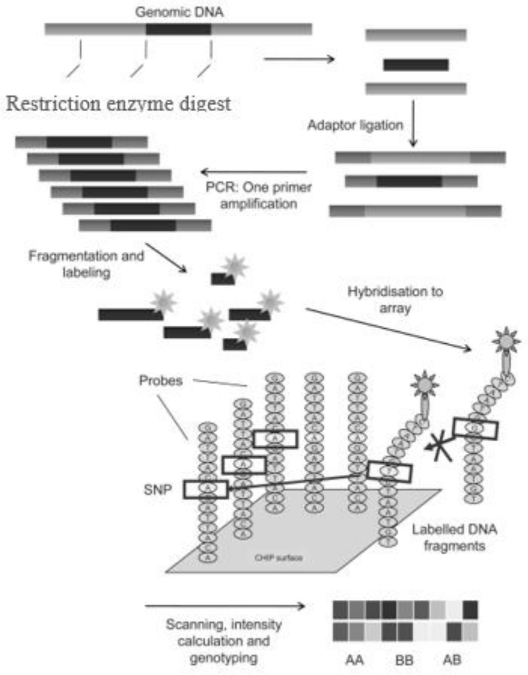
Genomic DNA
Restriction enzyme digest
Adaptor ligation
PCR: One primer amplification
Fragmentation and labeling
Hybridisation to array
Probes
SNP
CHIP surface
Labelled DNA fragments
Scanning, intensity calculation and genotyping
AA
BB
AB

Figure 6 *– Scheme of the principles of Affymetrix SNP array technology. Adapted from Nowak et al (2009) [50].*

The different alleles of SNPs can be examined by sequence-specific oligonucleotide microarrays, synthesized onto gene chips. Probes containing perfect matches and mismatches are combined for the examination of a single SNP. [50]

In this technique, genomic DNA is digested by restriction enzymes, forming fragments of different sizes. These fragments bind to adaptors, enabling a one-primer PCR, in order to form fragments of chosen size (200-1,100 bp) that will be labeled with a fluorochrome and subsequently hybridized to the microarray. There, DNA fragments containing a SNP bind specifically to their perfect match probes. The hybridized array is then scanned by a laser that detects fluorescence, which intensity depends on the binding degree. The resulting intensity data give information about DNA copy number and the determination of SNP alleles provides information about the genotype (Figure 6). [50]

2.3.1.2. SNP Microarray CNV Data Interpretation

The data obtained through the fluorescence analysis of the microarray probes are a set of values given in the form of base-2 logarithms. Each logarithmic value is associated to a specific region of the human genome and represents the binding fraction of patient DNA relatively to the control DNA. [52, 53] The proportions of each DNA sample can be estimated taking the inverse of the logarithm:

$$x = \log_2(ratio) \Leftrightarrow \\ ratio = 2^x \qquad (1.1)$$

For each probe, *x* is a logarithmic value in the *log2* scale, obtained from image processing whereas the *ratio* is the proportion of each DNA sample present in the probe. [52] For probes where *x* is less than zero there is a loss of genetic material. In these cases, less quantity of patient DNA sample was present in the probes when comparing to the control DNA sample (with a normal copy number of 2). A value of *x* greater than zero denotes a gain in genetic material, which means that a larger quantity of patient DNA was present. If *x* is zero, there are no genomic differences for that region.

This can be used to infer the copy number of any given genomic region containing the SNP or other marker sequences. [50] This is possible using equation (1.1), where *x* is the mean of the logarithmic values of the consecutive probes that reported the alteration. In the TCGA data set, this value is termed *Segment Mean.* The regions that register alterations are referred to as CNVs.

2.4. Algorithms and Statistical Methods

2.4.1. Machine Learning and Pattern Classification

Machine learning is a core subarea of artificial intelligence, a subject within computer science. The main goal of machine learning is the development of algorithms that learn autonomously and automatically, without the aid or intervention of humans. One of its most common applications is in data mining - the process of analyzing data from different perspectives, finding patterns and summarizing it into useful information. [54]

Of great importance in the application of machine learning is to obtain a prediction rule that is as accurate as possible and, in certain contexts (such as in medical diagnosis) a prediction rule that is effortlessly comprehensible by human experts. Machine learning also intersects broadly with statistics, mathematics, physics, theoretical computer science, and other.

In machine learning, pattern classification is the subfield dedicated to the study of methods to label data into distinct classes. This categorization can be made by distinctively labeling the data (supervised learning), dividing the data into classes (unsupervised learning), selection of the most significant features of the data (feature selection) or a combination of two or more of these tasks. [55]

Pattern classification tasks are typically divided into distinct blocks:

1. Data collection and representation;
2. Feature selection and/or feature reduction;
3. Clustering;
4. Classification.

Unsupervised Learning

In unsupervised learning, the machine is given inputs but does not receive a set of desired outputs. The goal of the machine in this instance is to build representations of the input, later used for decision making and prediction of future inputs, for example. It can be regarded as the finding of patterns within the data, extracted from unstructured noise. Two classic examples of unsupervised learning are clustering and dimensionality reduction. [55]

Clustering

This step endeavors to cluster the data and find representative data points (cluster centers, for example) or to remove superfluous data points. These techniques are applicable when the objective is the division of the data into natural groups. These clusters should reflect some mechanism that causes some instances to bear a stronger resemblance between each other than with the remaining instances. [56]

In practice, usually a criterion for joining instances into clusters and the preferred number of clusters is specified to be used by the clustering algorithm. This results in a problem of clustering algorithms: they can find clusters even if there are no clusters in the data. [55]

K-means clustering

One of the most common clustering algorithms is *k-means* clustering. This algorithm starts by picking K random points in the data set, defining them as centroids. Then each data point is assigned into a cluster number closest to each different centroid. The clustering thus obtained is based on the original randomized centroid, which is not exactly what is intended, so the centroids are updated using a mean of the data. This step is repeated until the centroids no longer move. [57] In this work, Euclidean distance was used for calculating the distance in *k-means*:

$$d_{euclid}\,(x,y) = \sqrt{\sum_{f=1}^{n}\left(x_f - y_f\right)^2} \quad (1.2)$$

One of the limitations of *k-means* clustering is that clusters must have hard boundaries, which means that a data point must only be part of a single cluster. Besides, k-means prefers spherical data. These limitations are evident in some cases, like in Figure 7, where the central data could either be placed at cluster 1 or cluster 2. [57]

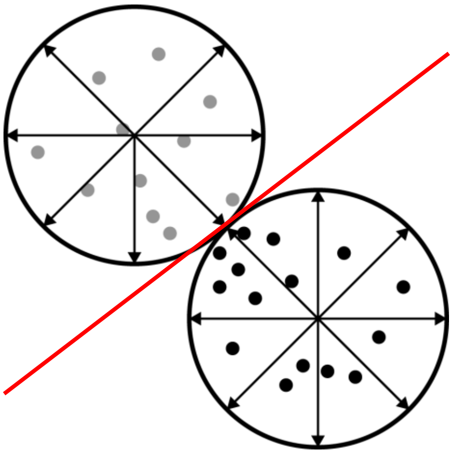

Figure 7 - k-means behaves in a circular manner. Adapted from Kirk (2014) [57]

2.4.2. Survival Analysis

Survival analysis focuses on data with three main characteristics: (1) the dependent (or response) variable is the time to the occurrence of a particular event, often death, (2) observations can be censored, meaning that the event of interest has not occurred yet or is not known to have occurred, and (3) there are predictors or explanatory variables that have an effect on the time to the occurrence of the event, that can be assessed or controlled. [58, 59]

For this work, the event is the death of the patient, and so the censored data are those where the outcome is unknown or survival. Time to the event is referred to as survival time.

The Survival Function

The survival function $S(t)$ is defined as the probability that the event has not taken place by duration, which is the probability of surviving at least until time t. [58]

Assuming that T is a continuous random variable with probability density function (p.d.f.) $f(t)$ its cumulative distribution function (c.d.f.) is $F(t) = P\{T < t\}$,resulting in the probability of the event having occurred by duration t. [59] The survival function $S(t)$ is then given by the complement of (t) :

$$S(t) = P\{T \geq t\} = 1 - F(t) = \int_t^\infty f(x)dx \quad (1.3) \quad [59]$$

The graphical representation of $S(t)$ against t is called a survival curve.

Kaplan-Meier Method

The Kaplan-Meier method can be used to estimate the survival curve from the observed survival times, without assuming any underlying probability distribution.[58] At any given time i, the survival probability is calculated by the formula given below:

$$p_i = \frac{r_i - d_i}{r_i} \quad (1.4) \quad [58]$$

Where r_i is the number of patients alive by the beginning of the period and d_i is the number of dead patients within the period. This method is based on the assumption that the probability of surviving k or more periods after entering the study (the cumulative proportion surviving) is given by the product of the k observed survival rates for each period:

$$S(k) = p_1 \times p_2 \times \ldots \times p_k \quad (1.5)\ [58]$$

The Kaplan–Meier method allows to estimate survival probabilities and to compare survival between groups, however it can only study the effect of one factor at the time, and consequently it cannot be used for multivariate analysis. For these purposes, a regression technique like the Cox proportional hazards model may be more of use. [60]

Cox's Proportional Hazards Model (Cox Regression)

Cox Regression enables the testing of differences between survival times of particular groups of patients while allowing explanatory variables to be considered. In this model, the dependent variable is hazard - the instantaneous probability of occurrence of the event, i.e. the risk of death of a patient at a given moment. It is assumed that the hazard does not follow a particular probability distribution and that the hazard ratio does not depend on time: the risk of a group of patients dying relatively to the other group does not vary from one moment to the other. [58]

The model is defined as follows:

$$\ln h(t) = \ln h_0(t) + b_1x_1 + \cdots + b_px_p \Leftrightarrow$$
$$\ln \frac{h(t)}{h_0(t)} = b_1x_1 + \cdots + b_px_p \quad (1.6)\ \ [58]$$

Here, $h(t)$ is the hazard at time t;$x_1, x_2, \ldots, x_p$ are the explanatory variables and $h_0(t)$ is the baseline hazard (hazard when all the explanatory variables are null). The coefficients $b_1, b_2, \ldots, b_p$are estimated from the data, using an optimization method. [58]

Because hazard measures the risk of death at instant t, it is easier to examine the cumulative hazard function $H(t)$, that can be obtain from the cumulative survival function S(t):

$$H(t) = -\ln S(t)\ (1.7) \quad [58]$$

3. Materials and Methods

3. Materials and Methods

3.1. Data collected from the TCGA Data Portal

Copy Number Data

Copy number data obtained by SNP array and patients' clinical data were downloaded from the TCGA Data Portal, available at https://tcga-data.nci.nih.gov/tcga/ , on the 23rd October, 2015.

Tumor samples were collected with appropriate informed consent from newly diagnosed HNSCC patients at the time of their surgical resection n.

The available copy number data was Level 3 data, meaning that it was not raw, having suffered some kind of processing before being made available. In this case, the accessible data was normalized copy number and purity/ploidy data, per sample. This included copy number information for tumor samples, normal solid tissue collected close to the tumor and blood samples from the patients, distinguishable by a Sample Type TCGA barcode identifier that assumes different values for different sample types: 01 for tumor samples, 10 for blood samples and 11 for solid normal tissue (Figure 8).

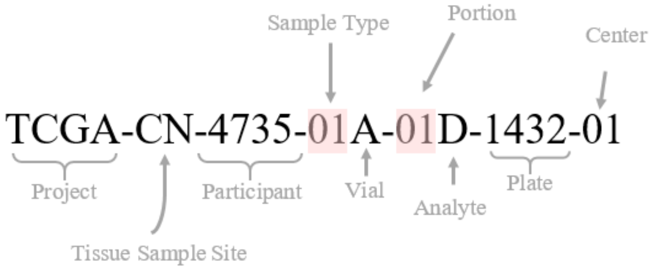

Figure 8 - Break down of a TCGA barcode from a HNSCC tumor sample into its components and translation into its metadata.

For each sample four files were included: two files where the germline CNVs had been removed and the other two including the patient's germline CNVs – in each case there is one file for each of two versions of the human genome references (Human Genome Version 18 and Human Genome Version 19). The selected files were the ones without germline CNVs that used Human Genome Version 19, since this is the current version in use. The copy number data was organized as exemplified in Table III, for each sample.

Table III *- Here are shown two lines from two different samples' copy number data sets as mere exemplification. Each line represents a CNV and each CNV is characterized by the name of the sample, the chromosome where the variation appears(Chr.), a start and an end reference, the number of probes which reported the alterations (Num_probes) and the mean of all log2(ratio) registered by all the probes involved in detecting a certain CNV(Segment_Mean).*

Sample	Chr.	Start	End	Num_probes	Segment_Mean
BALMS_p_TCGA b54and67_SNP_N _GenomeWideSN P_6_A01_730336	1	3218610	83929928	45758	-0,04
MIRES_p_TCGA _151_SNP_N_Gen omeWideSNP_6_ C06_831610	11	56242354	61930617	3143	0,2305

Clinical Data

Patient's clinical data encompassed information about a large array of features, including sample code, age at initial diagnosis, gender, ethnicity, country of origin, vital status at time of last follow-up, smoking history, tumor status (with tumor or tumor free following the tumor resection) at the time of enrollment, new tumor event status (metastization/relapse), follow-up times, days to death (when applicable), and HPV status among others. In addition, TNM staging components were also shown for both clinical and pathological staging and a compiled tumor stage using the standard AJCC staging criteria was also available. Almost all patients were subjected to treatment with curative intent (radiotherapy, chemotherapy or target therapy).

3.2. Software and Online Tools

In this work, the following software versions were used:

- MATLAB R2015a;
- IBM SPSS Statistics v.23 (Chicago, Illinois, USA).

The online tools used were:

- USCC Genome Browser;
- Ensembl Genome Browser 85.

3.3. Data selection

For the CNV data, a MATLAB (R2015a) routine was implemented to:

1. Organize data into tables discriminating between genetic material deletions and amplifications – a segment mean less than 0 denotes a deletion and one above 0 designates an amplification;
2. Select only tumor samples, based on the TCGA barcode associated to each sample- samples with a sample type code 01;
3. Select tumors located in specific histological sites (since the data from Laboratório de Citogenética e Genómica that were used as a comparison means was based on these locations) : oral tongue, base of tongue, floor of mouth, buccal mucosa, oral cavity, hard palate and alveolar ridge;
4. Reduce data volume, considering a minimum of 3 consecutive probes and a 0.1 segment mean, in modulus ;
5. Divide data into chromosome files;
6. Determine the minimum common regions with alteration for each chromosome;
7. Organize histograms using those alterations and set a threshold of region size and number of patients which have a certain altered region in common.

The limits of the most frequently altered regions (start and end) were introduced in the USCC Genome browser platform, in order to determine the chromosomic region where the alteration was located.

3.3.1. Division of data into chromosome files

The division of the data into chromosome files was done in order to increase the level of organization of the data and to try and condense the most information on the same document. These files were then used to ease the way for the subsequent tasks.

This division was achieved by taking each file from each patient and running it through a MATLAB R2015a routine, which fluxogram is represented in Figure 9.

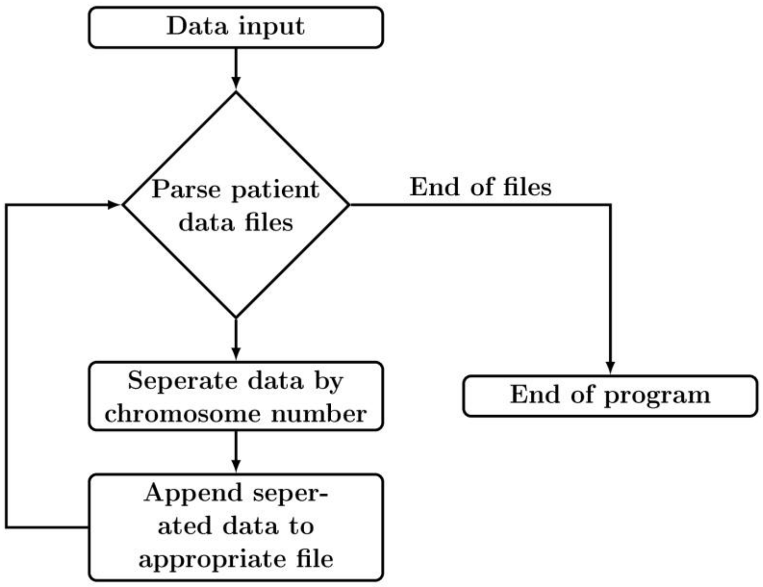
Data input
Parse patient data files
End of files
End of program
Seperate data by chromosome number
Append seper- ated data to appropriate file

Figure 9 – Fluxogram of the MATLAB R2015a routine implemented to divide the data into one file for each chromosome

3.3.2. Minimum common region determination

The determined minimum common regions constitute the minimum regions that are common between ranges of closed intervals established by each altered region from each patient, per chromosome. In other words, they represent the intersections of the altered regions present in the same interval. (Figure 9) Genetic material deletions and amplifications were discriminated.

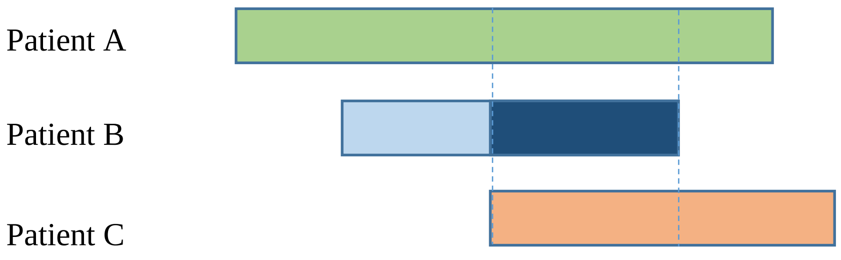

Figure 10 – Simplified schematic representation for the process of determination of the minimum common regions. The bars represent the portion of region that is alter for each of the patients (A, B and C) in the same given chromosomic band. The region in navy blue is the one that is common to all three patients –the minimum common region.

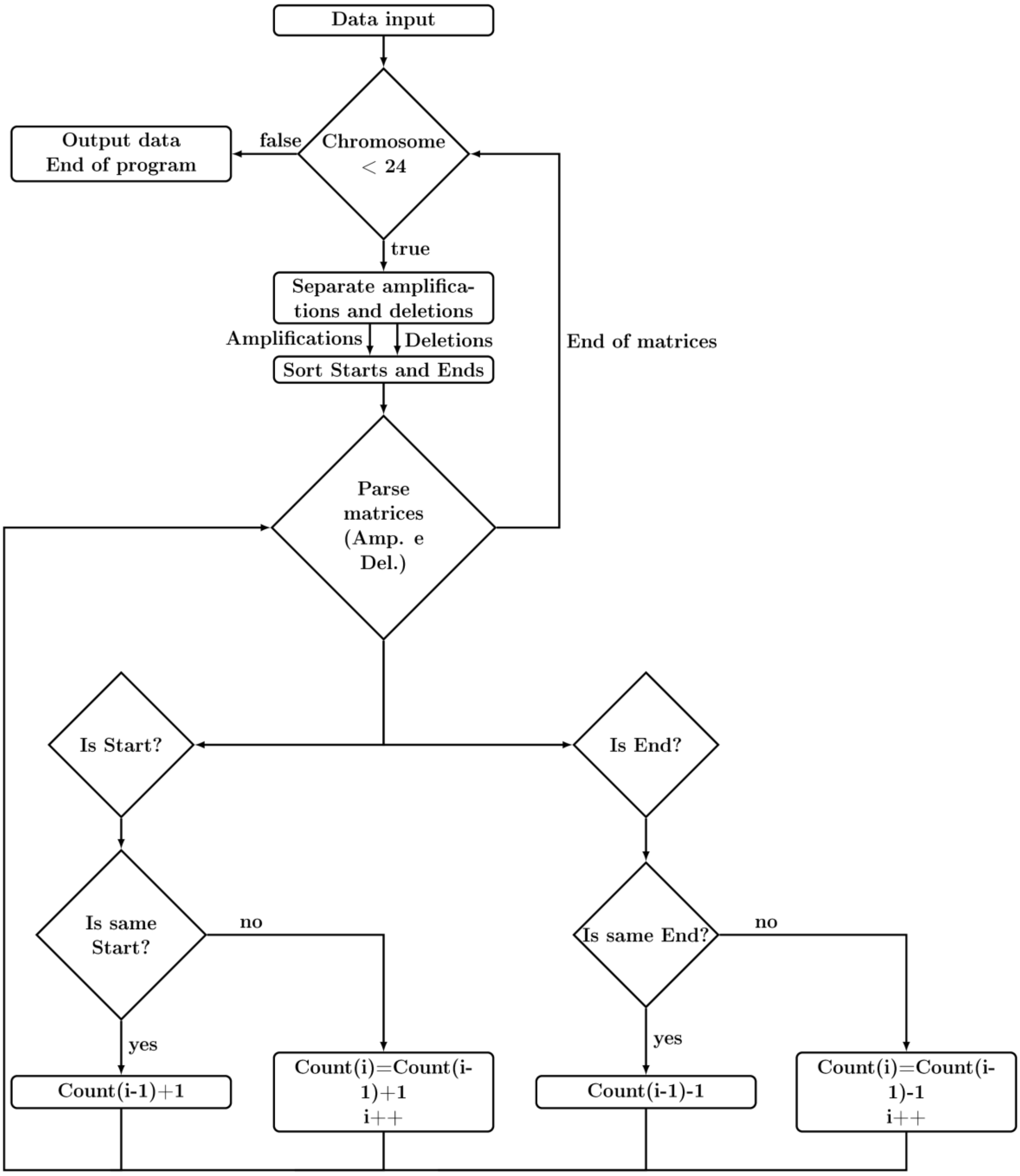

Figure 11 - Fluxogram for the MATLB R2015a routine implemented to find the minimum common regions for each chromosome

3.4. Feature Selection and Clustering

Again using MATLAB, the minimum common regions data, after some reduction, were organized into documents containing the alterations suffered by each one of the patients in the form of a dichotomic variable (0 for a normal region and 1 for an altered region) for each chromosome (Figure 12).

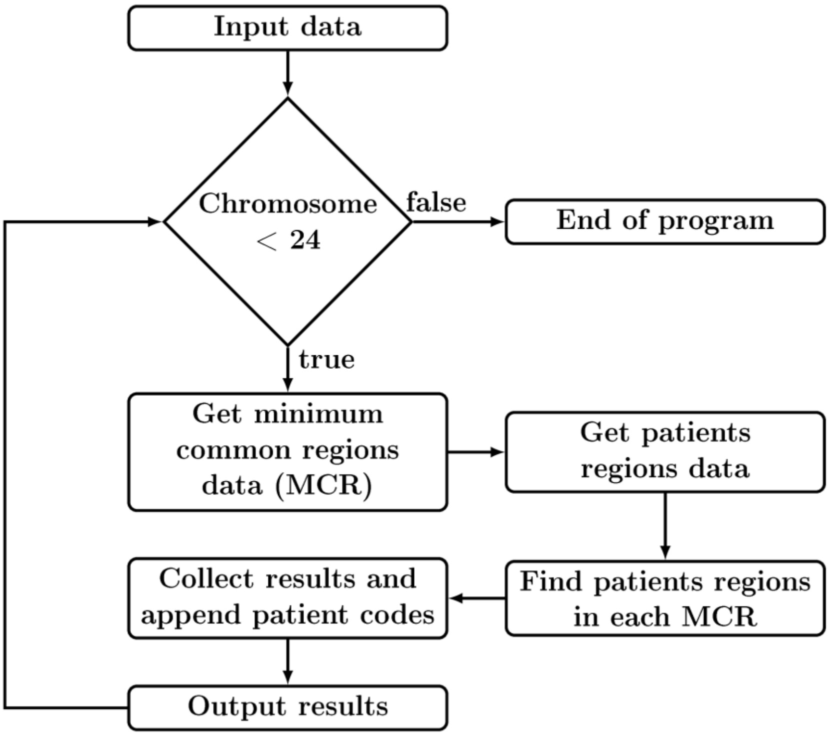

Figure 12 – Fluxogram for the MATLAB R2015a routine to construct the documents with the alterations present in each patient

These documents were then utilized with the clustering algorithm k-means in MATLAB. The k-means algorithm was chosen since it is the most common clustering method as well as being versatile and easy to use.

On the second iteration of this method, the genes for each chromosomic region were determined using the Ensembl Gene Browser. The resulting document contained all the genes for the minimum common regions, after reduction, of each chromosome as well as their respective description and biological functions. These genes were filtered using the key-words : cancer, head cancer, neck cancer, oral cancer, tumorigenesis, metastasis, angiogenesis, differentiation, proliferation, apoptosis, cell cycle, DNA repair gene, repair gene, damage repair, oncogene, tumor suppressor, tumor suppressor gene, chromatin remodeling and histone modification.

The selected genes were used to reduce the number of regions involved in the clustering algorithm, maintaining only the ones that codified at least one of the genes. The resulting regions were then used to repeat the clustering.

The cluster assignment was cross validated with some clinical features (metastasis/relapse information and clinical stage) to ascertain the relation between the clusters and their meaning at a phenotypic level, evaluated using a chi-squared test.

3.5. Survival Analysis

The genes selected at the previous step were compared with the genes considered more important by the bioinformatical analysis performed at Laboratório de Citogenética e Genómica of FMUC on their array CGH CNV data in a cohort of HNSCC patients, from HNSCC patients being followed at Hospitais da Universidade de Coimbra. The genes common to both data sets were then evaluated using survival analysis.

The survival analysis was performed considering the survival time (in days) for every patient and their vital status. When the subject was still alive or the outcome (alive or dead) was unknown, the days to the latest follow-up were used as the survival time and the patients were censured. The "days to last follow-up" variable was included in the clinical information provided by TCGA, meaning the number of days since the patients were diagnosed until they had their latest follow-up, after being included in the database.

First, using SPSS the Cox's Regression Model was applied to access the risk of death considering all the genes simultaneously. Then a Kaplan-Meier method was used to obtain the survival curves for each individual gene.

The survival analysis was conducted in a total of 312 patients, since 2 out of the original 314 didn't have any available information about the survival time.

3.6. Statistical Analysis

Statistical analysis was performed with the aid of the statistical analysis software package SPSS. To characterize the sample, descriptive measures of dispersion and central tendency (mean and standard deviation, quantitative variables) and absolute and relative frequencies (nominal variables) were used.

The results with *P* value smaller than 0.05 were considered statistically significant.

4. Results and Discussion

4. Results and Discussion

4.1. Cohort description

Clinical and copy number data was available from TCGA for a total of 528 patients. In this work, 314 of those patients were taken into account after some degree of selection: only the patients that had available tumor sample files and whose tumors were found in predetermined locations (Figure 8), were selected. All 314 patients exhibited HNSCC and all the tumor samples were from primary tumors. The majority of the cohort consists of tumors from the oral tongue (42%, n= 131), the oral cavity (23%, n= 73) and the floor of the mouth (20%, n= 63).

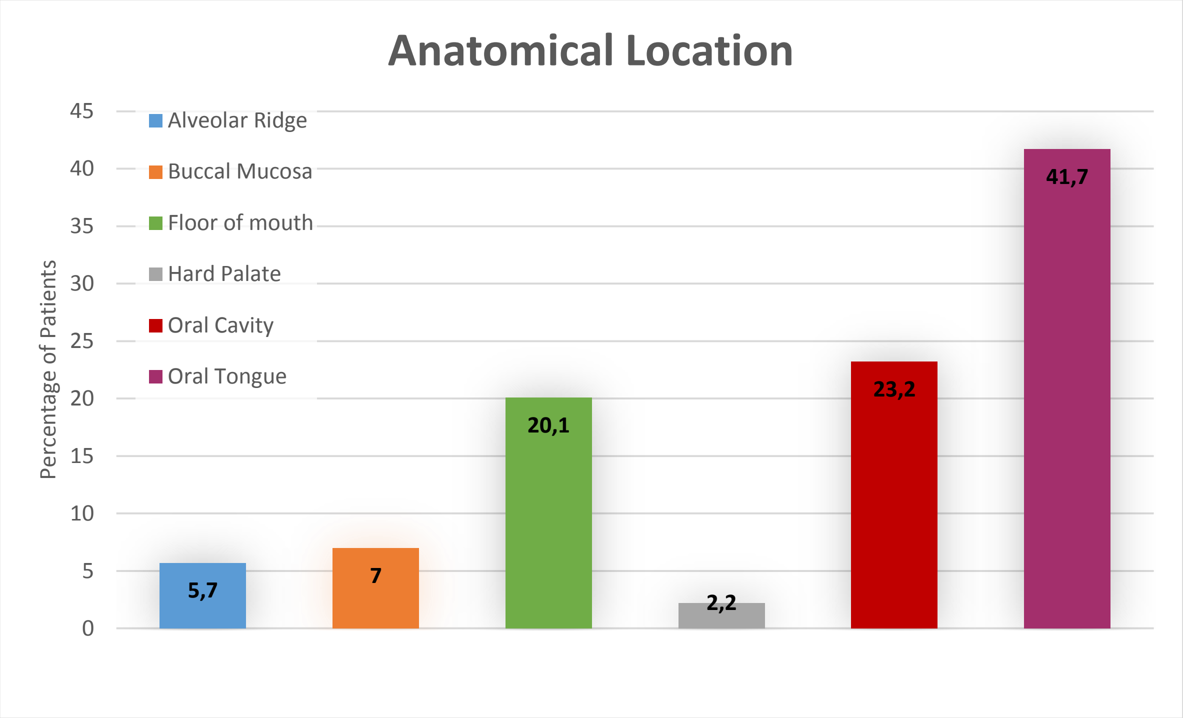

Figure 13 - Anatomical locations of the 314 tumor samples from the cohort, in percentage of patients

In this cohort, 66.56% (n = 209) of the patients were male and 33.44% (n = 105) were female. Figure 9 shows the distribution of age at initial diagnosis, for both sexes. The ages range from 19 to 90 years, with a mean value of 61.91±13.187 years of age at initial diagnosis, which is consistent with the reported majority of diagnosed HNSCCs. [61]

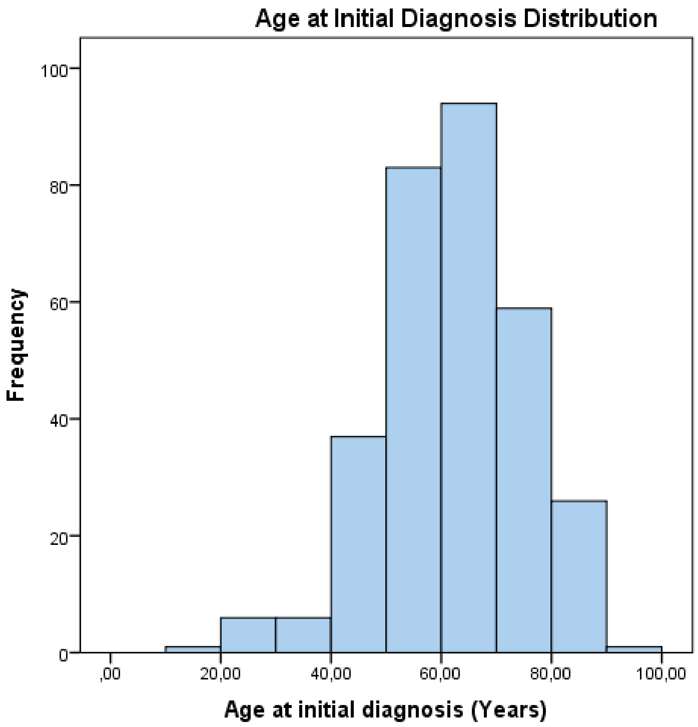

Figure 14 - Age at initial diagnosis distribution, for both sexes for the selected cohort

The samples were collected from an array of different countries as shown on Figure 10. The great majority of patients was from the United States of America (67%, n = 212).

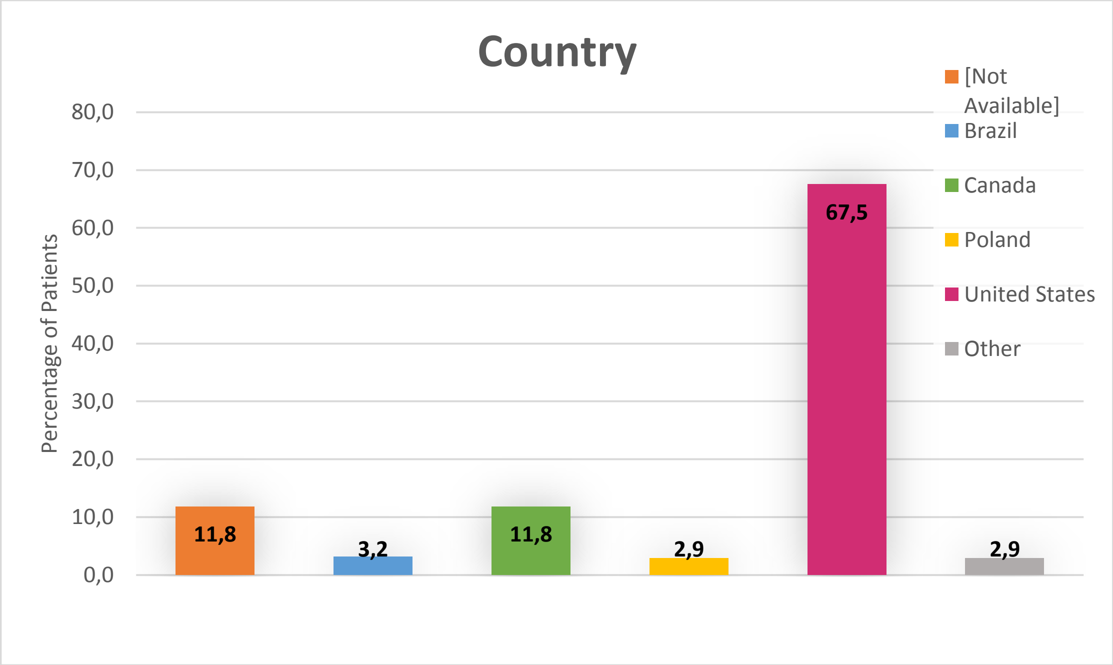

Figure 15 - Percentage of patients per country of origin present in the cohort

Table IV - Assessment of the greatest risk factors for HNSCC (smoking, alcohol drinking, presence of HPV) within the cohort

	Tobacco History	Alcohol History	HPV Status
Yes	215	203	32
No	90	104	281
Unknown	9	7	1

Concerning the most referred to risk factors for the development of HNSCC, the cohort exhibited a remarkable 68.5% (n=215) of patients with tobacco history, who were heavy smokers (mean pack years = 44. 94). Previous smokers comprised 53.6% (n= 121) of those and current smokers covered 43.72% (n=94). 64.7% of patients had an alcohol history. Furthermore, 162 of the patients with a tobacco smoking history also had an alcohol drinking history. Only 10.2% (n=32) of the cohort had HPV(+) status, which is to be expected since all the oropharyngeal tumors were excluded. (Table IV).

The clinical stages of the disease, upon diagnosis are dominated by Stage IV tumors (49.4%,n =155), as seen on Figure 11.

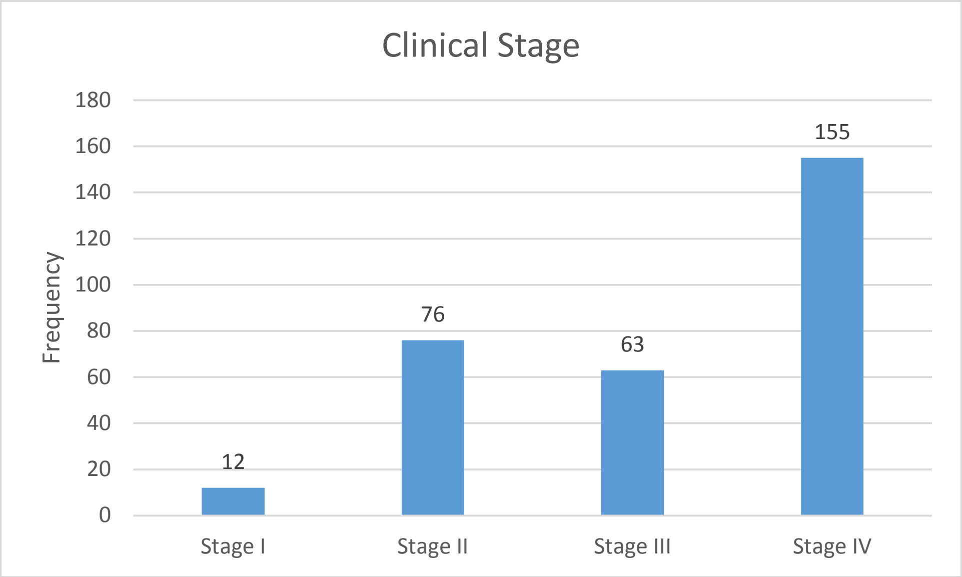

Figure 16 - Clinical stages based on TNM classification of the tumors in the cohort.

In this cohort, only 8.3% (n=26) of patients developed metastasis/relapse, however only 30.3% (n=95) of the patients had available information for the metastasis/relapse status. In average, patients died within 670.62 days (approximately 2 years) of being diagnosed with HNSCC.

4.2. Minimum common regions and most frequently altered regions

Genomic alterations were detected in all chromosomes, except chromosome Y, for which there were no data available. The size of the altered regions was variable, from patient to patient.

The determined minimum common regions were organized into tables and then represented into graphs of region start/end (in kbases) versus the number of patients carrying the alteration (absolute frequency). This representation establishes a visual template for the detected alterations along the chromosome.

As a way of simplifying the presentation of the results, only one of the twenty-two graphs - one for each chromosome - is shown (Figure 13).

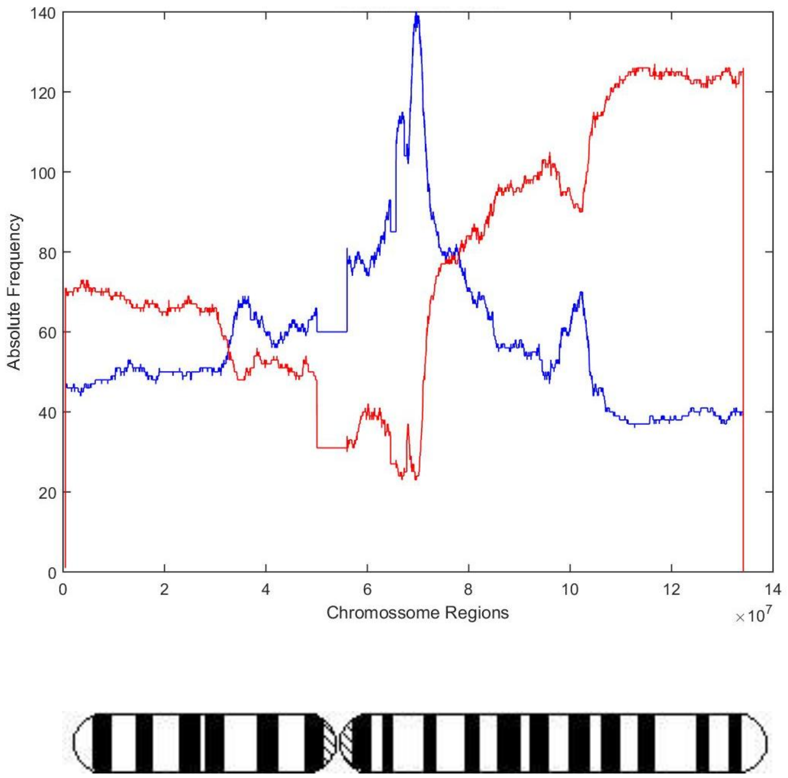

Figure 17- Representation of the minimum common regions and the frequency at which they are found in the cohort. This particular representation is for chromosome 11. Blue represents genetic material amplifications and red represents deleted chromosomic regions.

Chromosomes 3,5,8,9 and 11 registered the most frequently altered regions. The locations of the most recurrently altered regions are represented on Table V.

Table V – Chromosomic locations of the most frequently altered regions in the cohort, with type of alteration, percentage of tumors and absolute frequency

Type of Alteration	Chromosomic band	Percentage of patients	Absolute frequency
Amplification	8q24.21	79%	248
Amplification	3q22.23	67%	210
Amplification	5p15.33	44%	138
Amplification	11q13.3	44%	139
Deletion	3p21.2	69%	216
Deletion	8p22.3	66%	208
Deletion	9p21.3	62%	195
Deletion	11q23.2	40%	126

These results are in accordance with the reported alterations that are most frequently found in HNSCC tumors. [20, 35]

Concerning amplification of genomic material, gain of 8q is one of the most frequently reported copy number alterations in HNSCC, especially involving bands 8q23.1-q24.22.[20] Here, amplification of 8q24.21 band was the most frequently detected alteration. Since this is considered an early change already detectable in oral dysplasia, it is expected that it should be present in a large number of patients given that the distribution of clinical stage tends to later stage carcinomas (Figure 16).

Region 3q is also one of the most identified amplified regions. 3q22–q24 is a frequently amplified region reported by both *Bockmühl et al* (2000) and *Patmore et al (2002)*. [62, 63] Gain of 3q is correlated to poor prognosis and tumor development and is also an early marker associated with invasion and metastasis. [20, 63]

Region 5p is reported as one of the most frequently amplified chromosomic regions in HNSCC, especially 5p14-15 gains. 5p15 gain is associated with the transition from mild to moderate dysplasia.[20]

Amplifications of 11q13 are also very frequent in HNSCC and seem to be associated to a poor prognosis and poor survival, independently of the stage of the tumor.

[20, 63, 64] They are early alterations in HNSCC, having an important role in the transition from moderate to severe dysplasia. [20]

In what concerns the genetic material deletions, losses at the short arm of chromosome 3 are among the most common alterations in HNSCC. They are considered early changes in the progression of the disease. [20]

Genetic material loss of 8p is also a recurrent report for HNSCC. Losses at 8p21-22 are associated to poor prognosis and some studies have determined that loss at this loci is also associated with recurrence. [65]

Losses located at the short arm of chromosome 9 are also prevalent alterations of HNSCC, observed in pre-invasive and invasive lesions alike, suggesting that loss of 9p is an early event in HNSCC progression. The minimal area of loss identified in the same study was 9p21-22. [66]

Deletions located at the distal long arm of chromosome 11 (11q22-qter) are also a common staple for HNSCC cases and have been associated with loco-regional recurrence and poor survival. [67]

4.3. Region size and patient number threshold setting

After determining the minimum common regions for both the deleted and the amplified regions, there still were many regions to manage. The logical step was to try to reduce that volume of data.

First, histograms of region size versus number of regions per chromosome with that size, with various bin widths were established (100 kbases, 300 kbases, 500kbases) for each chromosome, for both the genetic material deletion and amplification (Figures 18-23).

Since this process was done for every single chromosome, only the resulting graphs for one representative chromosome are shown (chromosome 11).

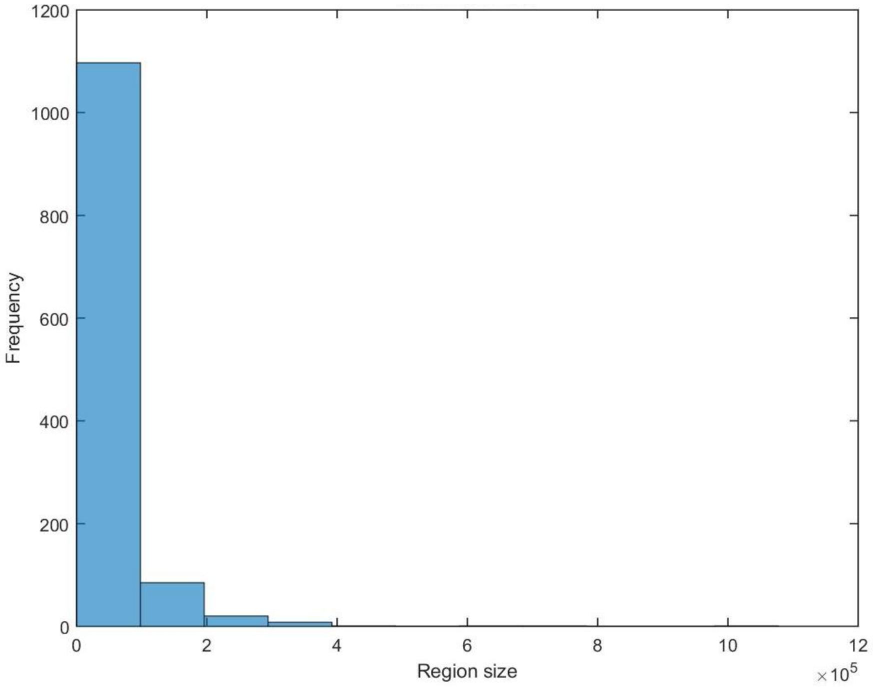

Figure 18- Region size histogram, with a bin width of 100 kbases for amplifications in chromosome 11

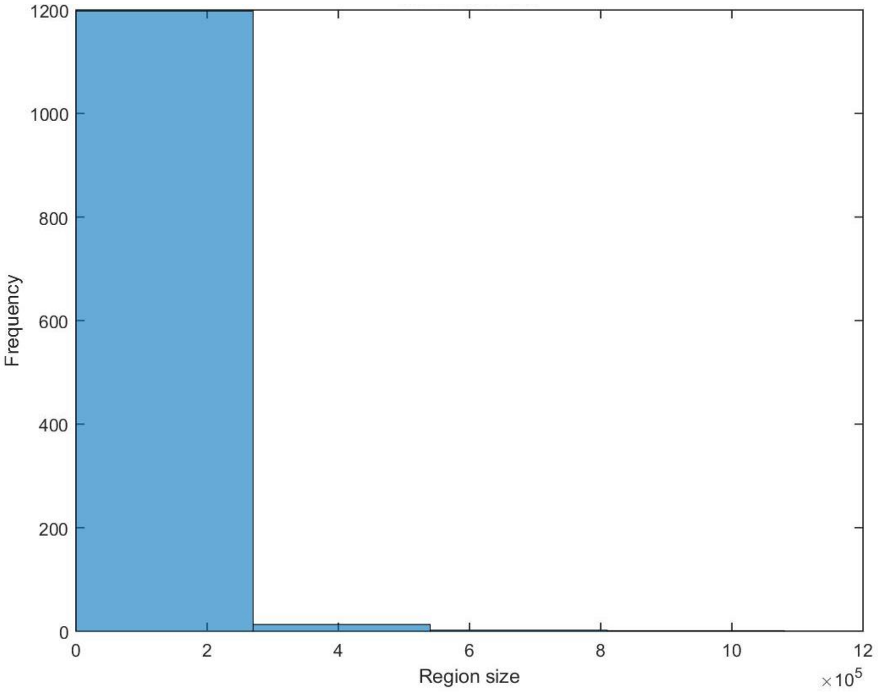

Figure 19 - Region size histogram, with a bin width of 300 kbases for amplifications in chromosome 11

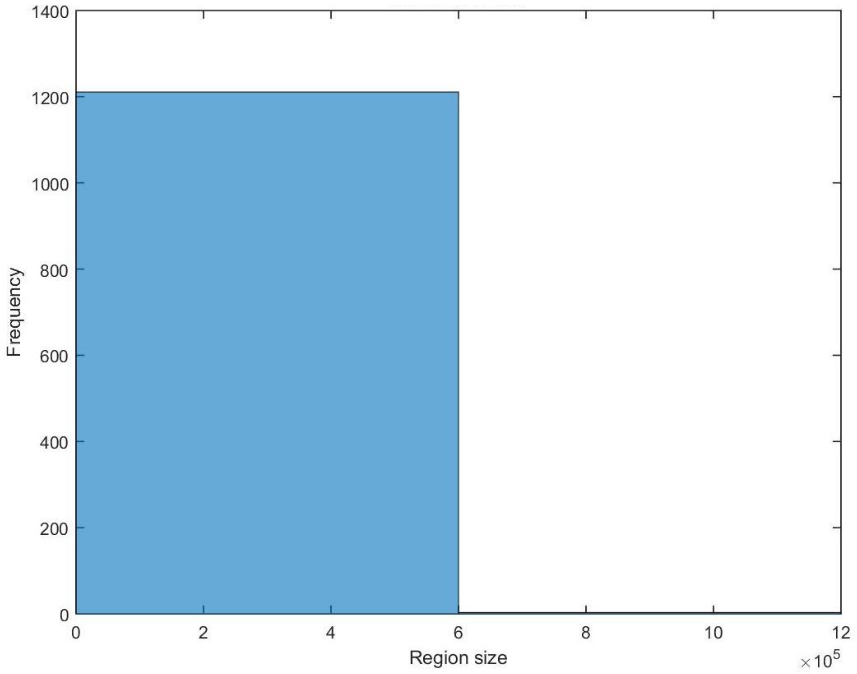

Figure 20 - Region size histogram, with a bin width of 500 kbases for amplifications in chromosome 11

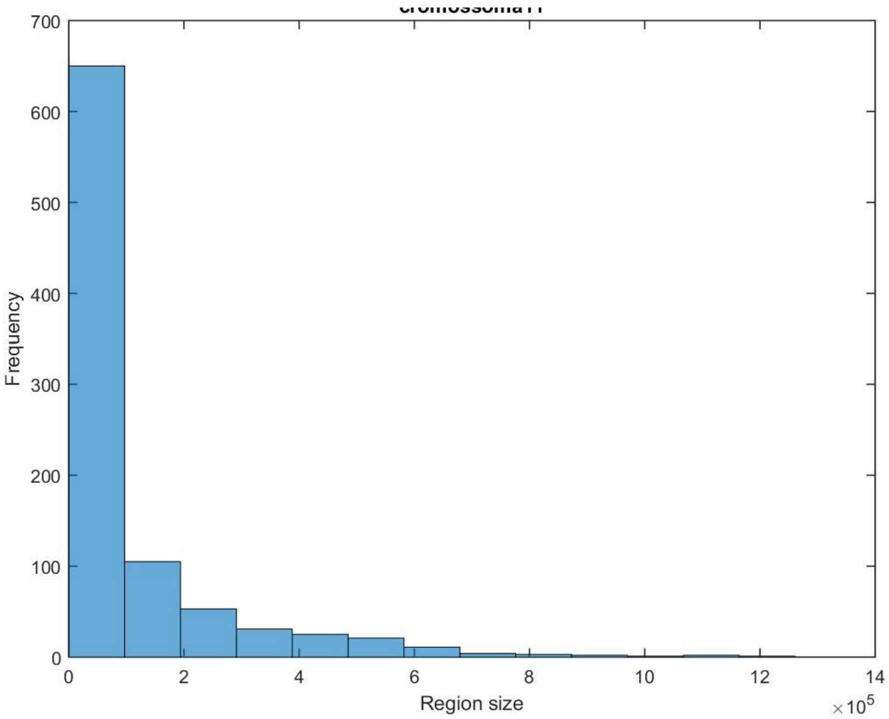

Figure 21- Region size histogram, with a bin width of 100 kbases for deletions in chromosome 11

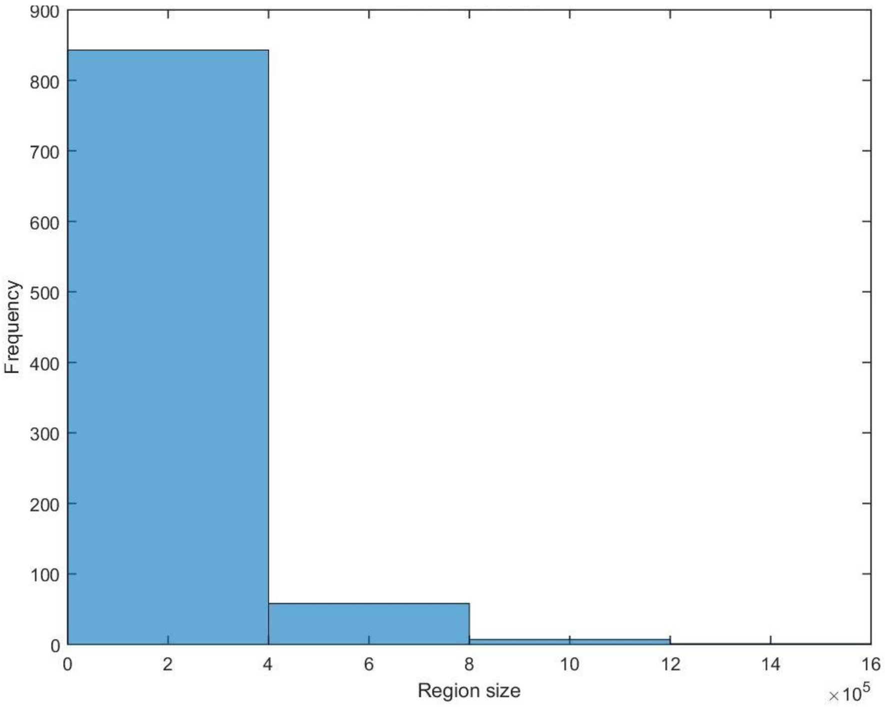

Figure 22 - Region size histogram, with a bin width of 300 kbases for deletions in chromosome 11

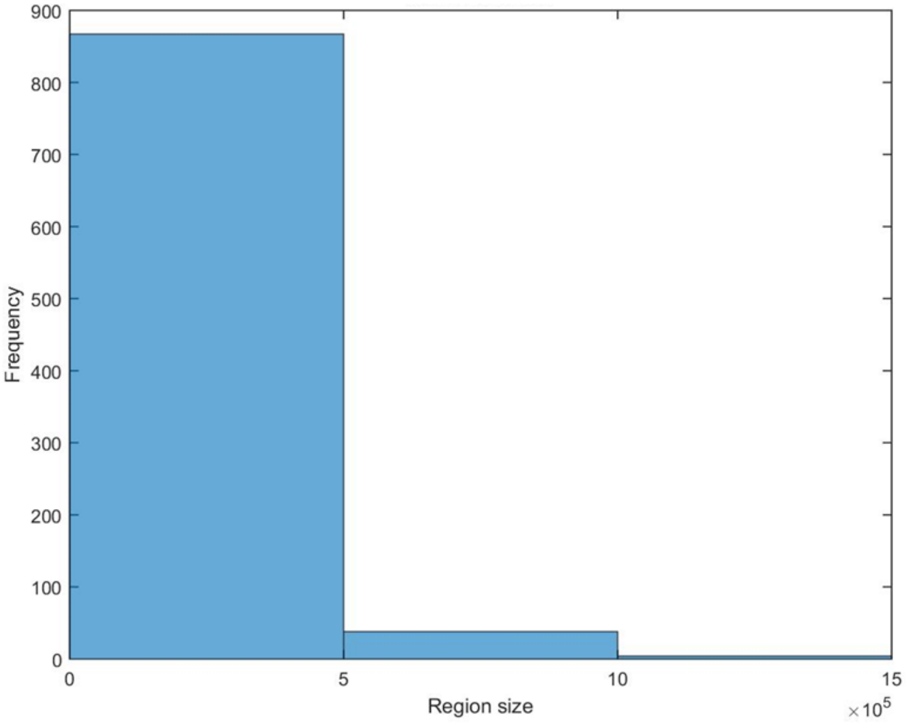

Figure 23 - Region size histogram, with a bin width of 300 kbases for deletions in chromosome 11

The next step in data volume reduction was the setting of a threshold in patient number based on the number of regions per quantile of patients, for each chromosome. Once again only the results for chromosome 11 are shown (Figures 24 and 25).

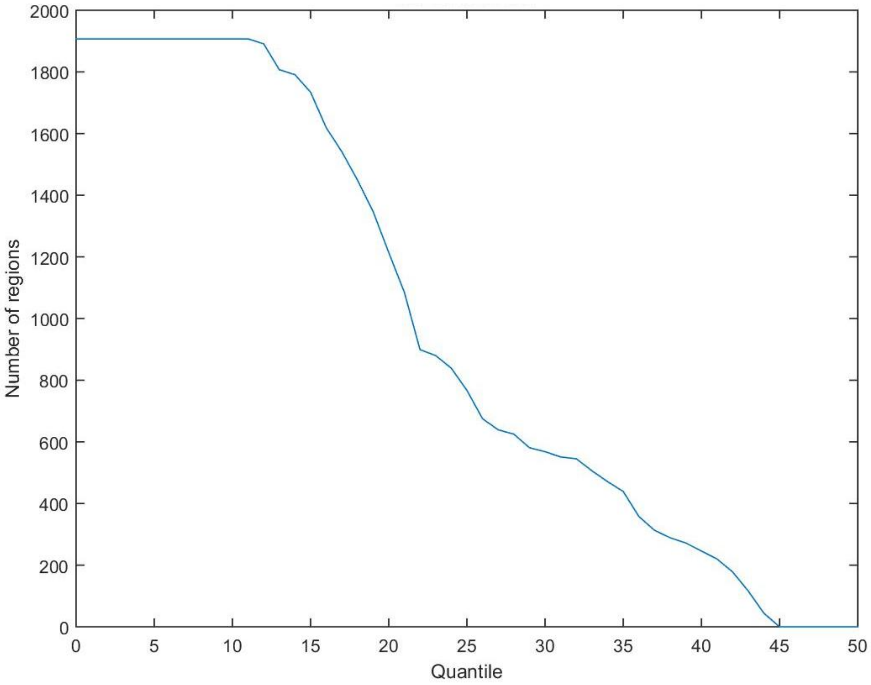

Figure 25 – Graphical representation of the number of regions present in each quantile of patients, for genetic material amplification in chromosome 11

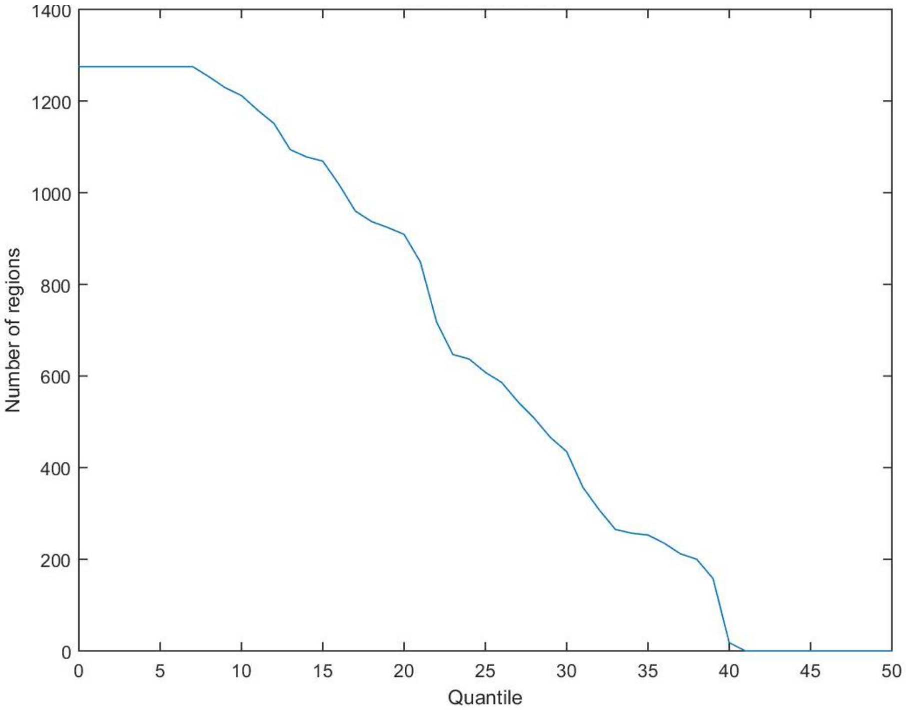

Figure 24 – Graphical representation of the number of regions present in each quantile of patients, for genetic material deletion in chromosome 11

After careful consideration of both criterions of selection and reduction, the decision was made to establish unanimous region size and patient number thresholds for every chromosome.

In terms of region size, the choice fell on the 300 kbases mark, meaning that it was decided to keep the regions that were 300 kbases or larger in size, since this was the option where the compromise between region size and number of regions was the most balanced for all of the chromosomes.

With regard to the number of patients, it was noticed that the graphical representations tended to have an inflexion point at around quantile 20. However, some chromosomes either at the deletions or amplifications did not meet that criterion, because of the reduced number of patients that presented those alterations. In those situations the criterion was still applied resulting in no regions being kept.

Accordingly, it was decided that the regions to keep would have to be present in at least 20% of the patients as well as being at least 300 kbases in size. As previously stated, the compromise was made to apply the same criteria for each chromosome. The resulting number of regions, for each chromosome, is shown in Table VI.

Table VI - Number of amplified and deleted regions kept, per chromosome, after reduction of the volume of data

	Number of regions	
Chromosome	**Amplification**	**Deletion**
1	99	8
2	3	34
3	100	74
4	0	183
5	39	138
6	4	0
7	99	39
8	96	30
9	103	15
10	0	68
11	11	98
12	30	0
13	4	106
14	87	0

Table VII (continuation) - Number of amplified and deleted regions kept, per chromosome, after reduction of the volume of data

	Number of regions	
Chromosome	**Amplification**	**Deletion**
15	0	61
16	33	0
17	0	18
18	9	39
19	0	13
20	48	0
21	0	34
22	12	0
X	57	47

4.4. Cluster Analysis

The cluster analysis was performed with the algorithm *k-means*, using MATLAB R2015a. The documents used to perform this analysis contained the regions in which every patient had alterations, with a 0 for a normal region and 1 for an altered region, for every chromosome.

The first application of this method resulted in a cluster assignment, of k =2 (two groups of patients) for every chromosome, based on the altered chromosomic regions. Most chromosome’s cluster assignments were very unbalanced, as made clear by the cluster size ratios presented in Table VII.

Table VIII – Number of patients by cluster and the ratio between cluster sizes, for each chromosome

Chromosome	Cluster 1	Cluster 2	Ratio
1	242	72	3,36
2	232	82	2,83
3	115	199	1,73
4	234	80	2,93
5	121	193	1,60
6	248	66	3,76
7	191	123	1,55

Table IX (continuation) – Number of patients by cluster and the ratio between cluster sizes, for each chromosome

Chromosome	Cluster 1	Cluster 2	Ratio
8	230	84	2,74
9	198	116	1,71
10	105	209	1,99
11	125	189	1,51
12	223	91	2,45
13	89	225	2,53
14	208	106	1,96
15	248	66	3,76
16	249	65	3,83
17	245	69	3,55
18	170	144	1,18
19	218	96	2,27
20	191	123	1,55
21	193	121	1,60
22	69	245	3,55
X	198	116	1,71

A visual depiction of the cluster assignments, by patient with the altered regions represented was made for every 23 chromosomes. However, here only that for chromosome 11 is present (Figure 26).

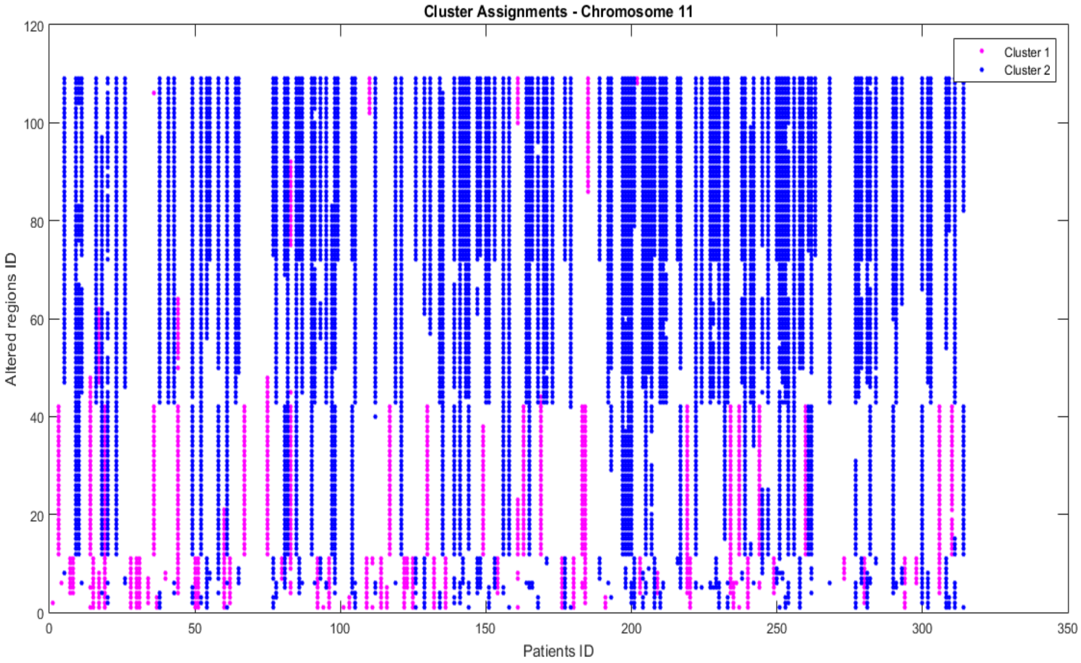

Figure 26 - Depiction of cluster assignment with representation of the altered regions by patient. As an example, the case of chromosome 11 is shown.

The representations for most chromosomes were crowded due mainly to the large number of regions considered for each chromosome, nonetheless an apparent change in the density in altered regions distribution can be seen allowing for the clustering of patients into distinct groups.

Bearing in mind that clusters should reflect a mechanism that causes some instances to strongly resemble each other in contrast with the remaining instances, the relationship between some phenotype features from the patients were tested.

The first of those to be tested was the tumor relapse/metastasis clinical feature. The results for chromosome 11 are shown in Figure 27.

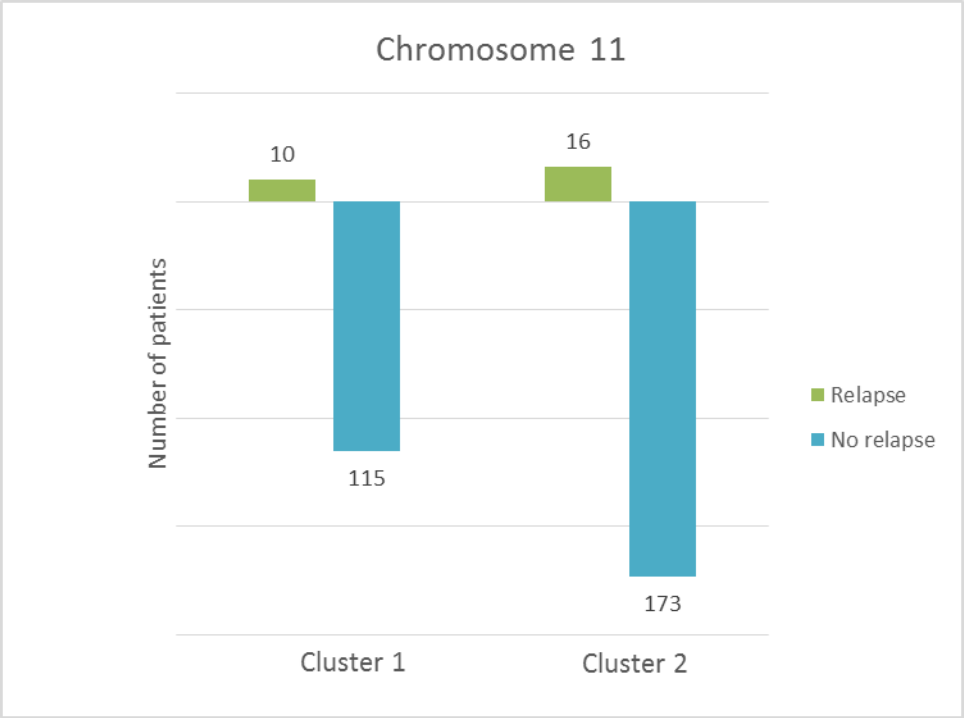

Figure 27 - Bar plot representation of the number of patients that suffered tumor relapse and those that did not, within the same cluster. As an example, the case of chromosome 11 is shown.

Here only the results for chromosome 11 are depicted, however the relationship between identified clusters and the phenotype feature tumor relapse/metastasis is not evident, for any of the analysed chromosomes. In order to determine the relation between the clusters and their meaning at a phenotypic level, these results were evaluated using a chi-squared test. For chromosome 11, the determined chi-square value was 2,000 with a significance of 0,157, leading to the conclusion that the tumor relapse/metastasis feature is not statically significant when trying to identify an underlying mechanism for cluster assignment. For the remaining chromosomes, the conclusions were identical.

In what concerns the relapse/metastasis clinical feature these results were somewhat expected, given the reduced number of patients from the cohort that had available information (n=95, 30.3%). A cluster analysis and further comparison with the relapse/metastasis clinical feature was performed considering only this group of patients, the results were, however, once again inconclusive.

In order to further explore the possibility of a clinical feature being the basis for cluster establishment, the clinical stage was tested. The results are shown in Figures 28-31 , for chromosome 11.

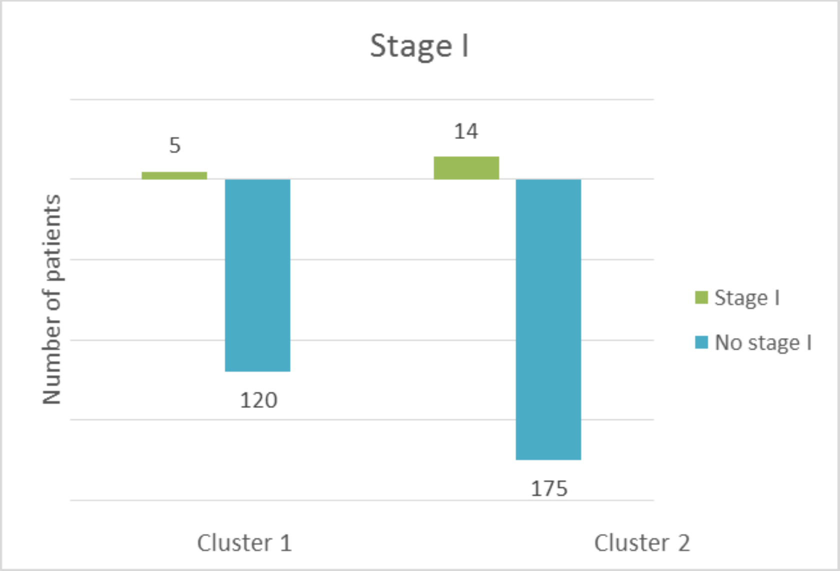

Figure 28 - Bar plot representation of the number of patients that were on clinical stage I and those that did not, within the same cluster. As an example, the case of chromosome 11 is shown.

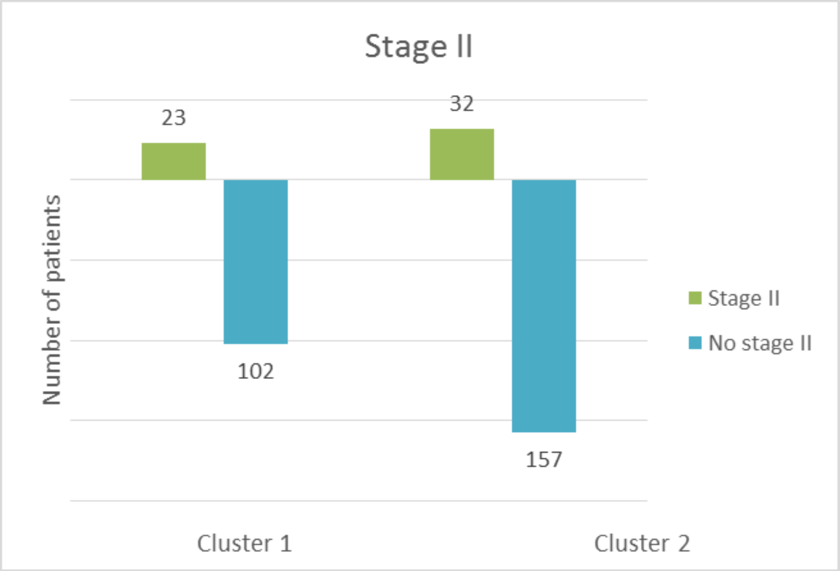

Figure 29 - Bar plot representation of the number of patients that were on clinical stage II and those that did not, within the same cluster. As an example, the case of chromosome 11 is shown.

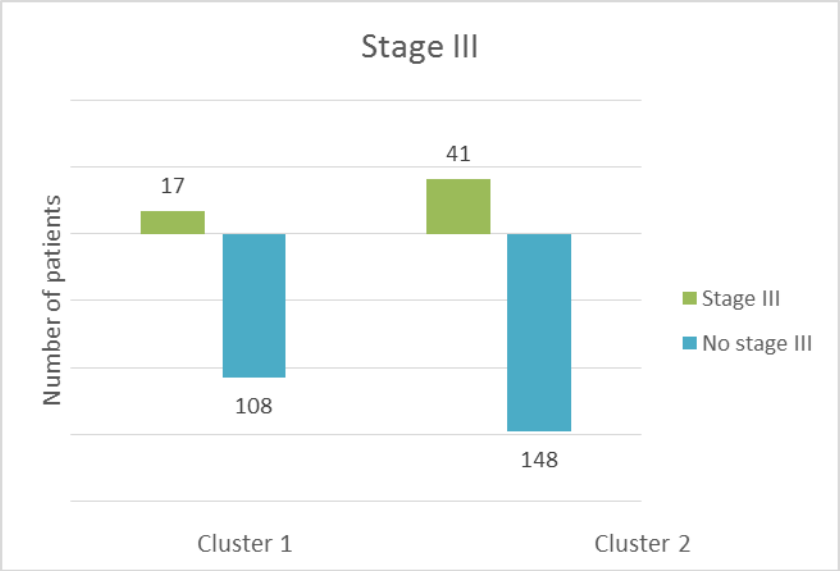

Figure 30 - Bar plot representation of the number of patients that were on clinical stage III and those that did not, within the same cluster. As an example, the case of chromosome 11 is shown.

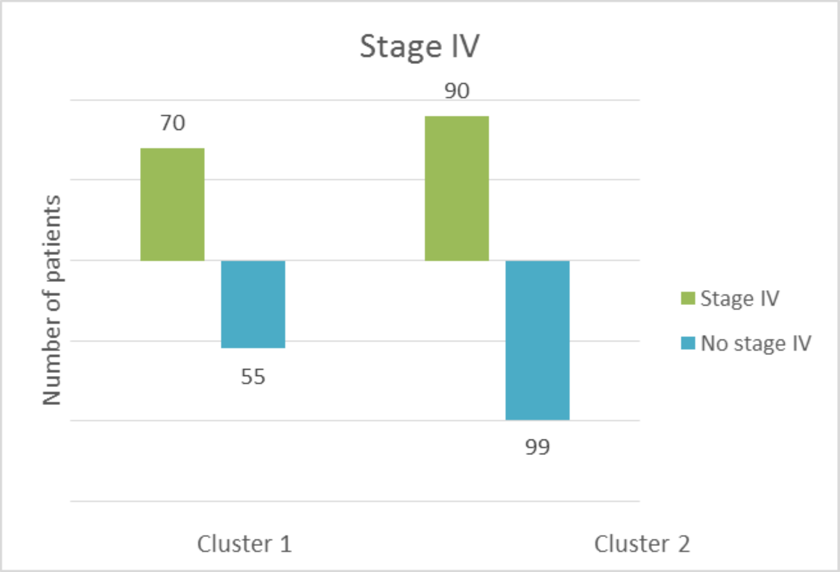

Figure 31 - Bar plot representation of the number of patients that were on clinical stage IV and those that did not, within the same cluster. As an example, the case of chromosome 11 is shown.

For every stage, a chi-square test was performed, the results for chromosome 11 are shown in Table VIII.

Table X – Chi-square test results for each of the Clinical Stage comparisons between clusters along with the significance

Clinical Stage	χ^2	Significance
Stage I	1,537	0,215
Stage II	0,112	0,737
Stage III	3,273	0,070
Stage IV	2,115	0,146

Significance < 0,05 – Statistical Significance

From the observation of the bar plots for clinical stage distribution among the clusters, no apparent distinction is evident between the two clusters for any of the stages. This assumption was further consolidated by the determined chi-square values along with their significance (higher than 0,05 for every instance).

All of the above mentioned tasks were performed for every chromosome, with the inferences for each chromosome being akin to the ones shown for chromosome 11.

One of the concerns when performing cluster analysis was the elevated number of regions used for some chromosomes. In order to reduce this number, the genes for each region were obtained and filtered by known function and cancer involvement. The resulting filtered genes were used to decrease the number of regions, considering every region that coded at least one gene. The previously described procedure was applied to the resulting regions, to no avail since the results were once more inconclusive.

In conclusion, the results obtained by the *k-means* clustering algorithm in conjunction with the testing of clinical features were not conclusive to a possible underlying mechanism of separation of the patients into clusters. One of the possible explanations for this outcome is the fact that the *k-means* algorithm separates the data into clusters even when there isn't a natural separation within the data to begin with and so, the obtained clusters could be the result of this problem. Other drawback could be the metric used by the algorithm to determine de distance between instances, which in this case was the Euclidean distance.

4.5. Survival Analysis – validation of results from the Laboratório de Citogenética e Genómica

After having reduced the number of genes from 4700 to 824 based on their biological function and cancer involvement, they were compared to a dataset of genes deemed the most relevant by bioinformatical analysis of patient tumor samples collected at the University of Coimbra Hospitals, within the scope of a study from the Laboratório de Citogenética e Genómica da Universidade de Coimbra. This dataset comprised a total of 70 genes that were cross-examined with the dataset of genes obtained from the most relevant regions determined by the treatment of the TCGA data. The genes common to both datasets are represented in Table XI.

Table XI – Genes common to both datasets, containing the gene designation, the chromosome that codes it, the start and the end of the region that codes it in the same chromosome, the alteration status in the dataset from this work and the number and percentage of patients where that region is altered.

Gene	Chromosome	Start	End	Status	Number of patients	Percentage of patients
APPL1	3	57227737	57273468	Deleted	224	71%
BCR	22	23179704	233180037	Amplified	69	22%
FER	5	108747822	10919196841	Deleted	118	37%
NEK7	1	198156963	198322420	Amplified	71	23%
SMARCB1	22	23786963	23834516	Amplified	69	22%

These five genes were then used to perform a survival analysis upon the dataset of patients from the TCGA project.

4.5.1. Cox's proportional hazards model (Cox Regression)

Using Cox's proportional hazards model it was possible to assess all of the five genes at once. The output is shown in Table XII.

Table XII – Application of Cox's regression to the data, using the presence of the gene alteration as explanatory variables.

Gene	Coefficient (b)	Standard error	P	e^b	95,0% CI for e^b
APPL1	0,491	0,243	0,044*	1,634	1,014 -2,632
BCR	-0,163	0,248	0,511	0,849	0,522 -1,382
FER	-0,201	0,219	0,360	0,818	0,533 -1,257
NEK7	-0,154	0,243	0,526	0,857	0,532 -1,380

**P-value* < 0,05 – Statistical Significance

The *P* values indicate that the presence or the absence of the alteration was only statistically significant in the case of the *APPL1* gene (*P* = 0,044), whereas for the remaining genes (except *SMARCB1* for which the algorithm did not even calculate any coefficients) the presence or the absence of the alteration was not significant for the survival.

The coefficient b is the logarithm of the hazard ratio for a patient that carries the alteration of the region that codes the gene compared to a patient that does not. In the case of *APPL1* this coefficient is 0.491 and the exponential of this value is 1.634, indicating that a patient that carries the deletion of the *APPL1* loci is in average 1.634 times more likely to die from HNSCC at any given time than a patient that does not carry the alteration. This indicates that the risk associated with the deletion of the *APPL1* gene loci in patients of HNSCC is much higher. In this case, the confidence interval for e^b does not contain 1, which indicates a difference between the risk associated with the two situations and the statistical significance of *APPL1's* loci deletion.

4.5.2. Kaplan-Meier Method

The Kaplan-Meier Method was applied in order to determine a more readily available measurement for survival that could easily be translated into a tangible quantity. Therefore, this method was primarily used to assess the median, in days, of the survival time for both groups of patients – the ones with alteration of the gene loci and those where the alteration is absent.

Table XIII - Median of survival time, in days, for the APPL1 gene loci deletion presence and absence in HNSCC patients

APPL1	**Median (days)**	
	Estimate (days)	**Std. Error (days)**
Deletion present	1671,000	388,761
Deletion absent	2717,000	1306,592
Overall	2002,000	361,869

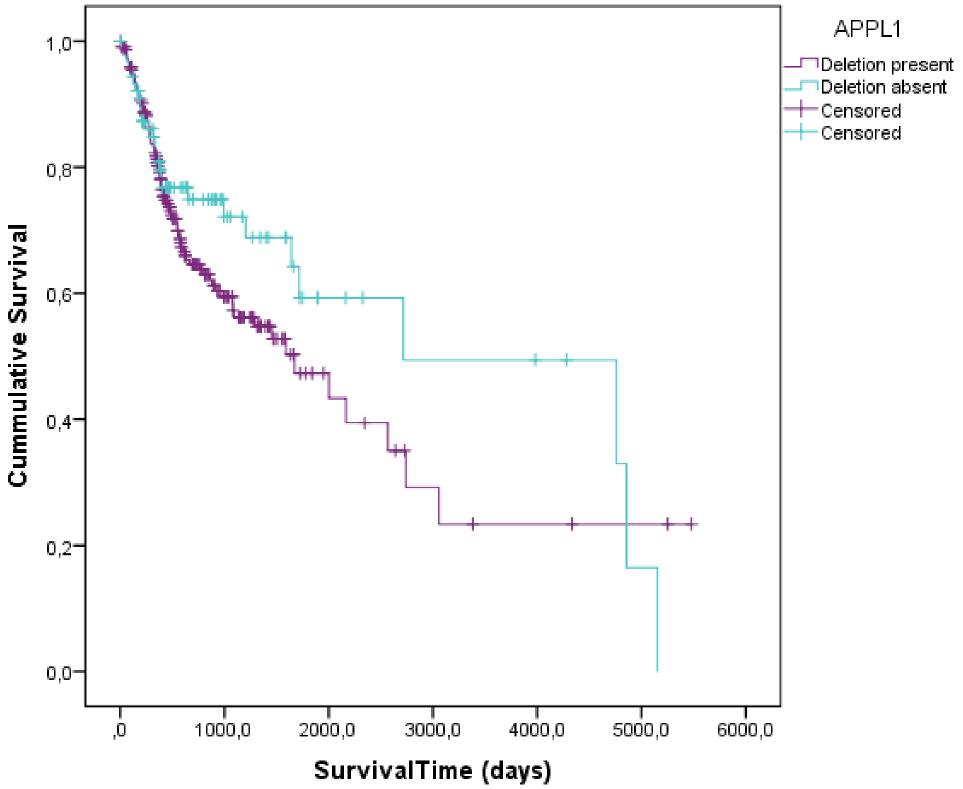

Figure 32- Cumulative survival versus survival in days for patients of HNSCC when considering the alteration status of the region that codes APPL1 gene

In the case of *APPL1*, the median survival estimate is 1046 days (approximately 2 years and 10 months) higher for the group of patients that does not carry the region deletion. The difference in probability of survival between the two groups of patients is also evident in the cumulative survival functions graph (Figure 26).

This may have some clinical implications: knowing not only that the patients that carry the alteration have a higher probability of survival and that 50% of the HNSCC patients that do not carry an *APPL1* loci deletion have a higher survival time by almost 3 years may have real importance when it comes to establishing a prognosis and defining a course of action for the patient's treatment and life quality improvement.

APPL1 is a protein coding gene which protein is involved in the regulation of cell cycle. The encoded protein binds other proteins, including RAB5A, AKT and PIK3CA, all of those being involved in cell cycle progression and cell replicative potential in HNSCC. [68] Although alteration on the expression of *APPL1* has been reported in several types of cancer, no literature referring to its connection with HNSCC was found.

Besides *APPL1*, the cases of *BCR*, *FER* and *NEK7* were also tested. However, from their cumulative survival curves there is no evident difference between the survivals in both groups of patients. Nevertheless, for *FER* and *BCR* the survival time median estimates may have some clinical importance as well. For *BCR*, the survival time median estimate was 546 days higher for patients that exhibited gain in the *BCR* coding region and for *FER* the survival time median estimate was 568 days higher for patients that did not have loss in the *FER* coding region.

5. Conclusions

5. Conclusions

As previously mentioned throughout the course of this work, HNSCC is a highly malignant disease, with invasion of surrounding tissue and distant metastization and a 50% five year survival rate. Although many efforts have been directed at discovering its genomic and metabolomic background, the pathways that lead to carcinogenesis in this type of cancer are widely unknown. It is known, however, that HNSCC is the product of genetic impairment and the accumulation of these damages lead to the progression of the disease: out of these, copy number alterations are of particular importance to this work, in the sense that tumor progression and clinical outcome may be affected by these modifications and the genes that those regions encode and may be targets for early detection and/or therapy.

With this work, it was possible to identify the most commonly altered chromosomic regions for a cohort of HNSCC patients obtained from TCGA data portal. From these, two groups of patients per chromosome were identified however no association with the phenotypes metastasis/relapse and cancer stage was established.

It was also possible to identify relevant genes present in the most frequently altered regions in the cohort. After comparison with the most frequently altered genes in a cohort of HNSCC patients from the Hospitais da Universidade de Coimbra, five potential target genes were identified: *APPL1, BCR, FER, NEK7* and *SMARCB1*.

Out of those, the deletion *APPL1* was found to be statistically significant for the risk of death of HNSCC patients ($P = 0.044$) who were, in average, 1.634 times more likely to die from HNSCC at any given time if they carried this alteration. Additionally, 50% of patients that did not present deletion of the *APPL1* loci survived 1046 days more than the other group, conferring *APPL1* the possibility of application in a real life context.

FER and *BCR's* survival time median estimates may have some clinical importance as well. The survival time median estimate for *BCR* was 546 days higher for patients that exhibited gain in the gene's coding region and the survival time median estimate for *FER* was 568 days higher for patients that did not have loss in the *FER* loci.

The deletion of *APPL1* and *FER* as well as the non-amplification of *BCR* seem to be biomarkers for worse prognosis in HNSCC patients.

In the clinical context, the knowledge that a patient may present better survival odds depending on their genetic alterations, may affect both the patient's prognosis and the applied type and duration of treatment. As such, the prediction of a worse prognosis may lead to a closer monitoring of the patients' disease progression, in order to provide a higher quality of life or eventually increase the patient's expected survival time.

6. Future Perspectives

6. Future Perspectives

Cancer is a disease of the genome and this work succeeded in finding some potential genetic biomarkers for survival, however it is not yet concluded and some additional procedures must be implemented.

It is important to evaluate the validity of these results at a biological level, using biopsies of patients of HNSCC to perform fluorescence in situ hybridization (FISH) confirming the relationship between presence or absence of alteration and survival, using probes directed at *APPL1, FER* and *BCR*, but mainly at *APPL1* since this was the only gene that presented statistically significant survival.

The clustering algorithm used in this work was *k-means* implemented used Euclidean distance as metric, however the results were far from promising. At a future attempt, a different metric could be used to calculate the distance using this algorithm. Another approach would be the use of another algorithm to perform the cluster analysis in the attempt to find a correlation between HNSCC patients’ phenotypic features and the cluster assignments.

7. Bibliography

Bibliography

1. Torre, L.A., et al., *Global cancer statistics, 2012.* CA Cancer J Clin, 2015. **65**(2): p. 87-108.
2. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-24.
3. Garraway, L.A. and E.S. Lander, *Lessons from the cancer genome.* Cell, 2013. **153**(1): p. 17-37.
4. Macconaill, L.E. and L.A. Garraway, *Clinical implications of the cancer genome.* J Clin Oncol, 2010. **28**(35): p. 5219-28.
5. Hanahan, D. and R. Weinberg, *The Hallmarks of Cancer* Cell 2000. **100**(1): p. 57-70.
6. Safdari, Y., et al., *Recent advances in head and neck squamous cell carcinoma--a review.* Clin Biochem, 2014. **47**(13-14): p. 1195-202.
7. Leemans, C.R., B.J. Braakhuis, and R.H. Brakenhoff, *The molecular biology of head and neck cancer.* Nat Rev Cancer, 2011. **11**(1): p. 9-22.
8. Stadler, M.E., et al., *Molecular biology of head and neck cancer: risks and pathways.* Hematol Oncol Clin North Am, 2008. **22**(6): p. 1099-124, vii.
9. Marur, S. and A.A. Forastiere, *Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment.* Mayo Clin Proc, 2016. **91**(3): p. 386-96.
10. *GLOBOCAN* 2012 [cited 2016 16-05]; Available from: http://globocan.iarc.fr/Pages/Map.aspx.
11. Bose, P., N.T. Brockton, and J.C. Dort, *Head and neck cancer: from anatomy to biology.* Int J Cancer, 2013. **133**(9): p. 2013-23.
12. Petti, S., *Lifestyle risk factors for oral cancer.* Oral Oncol, 2009. **45**(4-5): p. 340-50.
13. Pai, S.I. and W.H. Westra, *Molecular pathology of head and neck cancer: implications for diagnosis, prognosis, and treatment.* Annu Rev Pathol, 2009. **4**: p. 49-70.
14. Andrews, G.A., et al., *Mutation of p53 in head and neck squamous cell carcinoma correlates with Bcl-2 expression and increased susceptibility to cisplatin-induced apoptosis.* Head Neck, 2004. **26**(10): p. 870-7.

15. Ram, H., et al., *Oral cancer: risk factors and molecular pathogenesis.* J Maxillofac Oral Surg, 2011. **10**(2): p. 132-7.

16. Tornesello, M.L., et al., *HPV-related oropharyngeal cancers: from pathogenesis to new therapeutic approaches.* Cancer Lett, 2014. **351**(2): p. 198-205.

17. Neville, B.W. and T.A. Day, *Oral cancer and precancerous lesions.* CA Cancer J Clin, 2002. **52**(4): p. 195-215.

18. Rothenberg, S.M. and L.W. Ellisen, *The molecular pathogenesis of head and neck squamous cell carcinoma.* J Clin Invest, 2012. **122**(6): p. 1951-7.

19. Kalyankrishna, S. and J.R. Grandis, *Epidermal growth factor receptor biology in head and neck cancer.* J Clin Oncol, 2006. **24**(17): p. 2666-72.

20. Gollin, S.M., *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): p. 972-90.

21. Sanderson, R.J. and J.A. Ironside, *Squamous cell carcinomas of the head and neck.* BMJ, 2002. **325(**7368): p. 822-7.

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

23. Guerra, E.N., et al., *Diagnostic accuracy of serum biomarkers for head and neck cancer: A systematic review and meta-analysis.* Crit Rev Oncol Hematol, 2016. **101**: p. 93-118.

24. Guerra, E.N., et al., *Diagnostic capability of salivary biomarkers in the assessment of head and neck cancer: A systematic review and meta-analysis.* Oral Oncol, 2015. **51**(9): p. 805-18.

25. Patel, S.G. and J.P. Shah, *TNM staging of cancers of the head and neck: striving for uniformity among diversity.* CA Cancer J Clin, 2005. **55**(4): p. 242-58; quiz 261-2, 264.

26. Mao, L., W.K. Hong, and V.A. Papadimitrakopoulou, *Focus on head and neck cancer.* Cancer Cell, 2004. **5**(4): p. 311-6.

27. Trotta, B.M., et al., *Oral cavity and oropharyngeal squamous cell cancer: key imaging findings for staging and treatment planning.* Radiographics, 2011. **31**(2): p. 339-54.

28. Shah, J.P. and Z. Gil, *Current concepts in management of oral cancer--surgery.* Oral Oncol, 2009. **45**(4-5): p. 394-401.

29. *NCI Dictionary of Cancer Terms*. 5/4/2015 [cited 2016 3/5/2016]; Available from: http://www.cancer.gov/publications/dictionaries/cancer-terms?cdrid=44971.

30. Mazeron, R., et al., *Current concepts of management in radiotherapy for head and neck squamous-cell cancer.* Oral Oncol, 2009. **45**(4-5): p. 402-8.

31. Foulkes, M., *Oral cancer: risk factors, treatment and nursing care.* Nurs Stand, 2013. **28**(8): p. 49-57.

32. Gollin, S.M., *Chromosomal alterations in squamous cell carcinomas of the head and neck: window to the biology of disease.* Head Neck, 2001. **23**(3): p. 238-53.

33. Jenkins, G., et al., *Genome stability pathways in head and neck cancers.* Int J Genomics, 2013. **2013**: p. 464720.

34. Shlien, A. and D. Malkin, *Copy number variations and cancer.* Genome Med, 2009. **1**(6): p. 62.

35. Ashman, J.N., et al., *Prognostic value of genomic alterations in head and neck squamous cell carcinoma detected by comparative genomic hybridisation.* Br J Cancer, 2003. **89**(5): p. 864-9.

36. Sticht, C., et al., *Amplification of Cyclin L1 is associated with lymph node metastases in head and neck squamous cell carcinoma (HNSCC).* Br J Cancer, 2005. **92**(4): p. 770-4.

37. Murugan, A.K., A.K. Munirajan, and N. Tsuchida, *Genetic deregulation of the PIK3CA oncogene in oral cancer*. Cancer Lett, 2013. **338**(2): p. 193-203.

38. Muller, D., et al., *Cyclin L1 (CCNL1) gene alterations in human head and neck squamous cell carcinoma*. Br J Cancer, 2006. **94**(7): p. 1041-4.

39. Ribeiro, I.P., 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): p. 4687-95.

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

41. Jin, Y., et al., *Cyclin D1 amplification in chromosomal band 11q13 is associated with overrepresentation of 3q21-q29 in head and neck carcinomas.* Int J Cancer, 2002. **98**(3): p. 475-9.

42. Martin, C.L., et al., *Chromosomal imbalances in oral squamous cell carcinoma: examination of 31 cell lines and review of the literature.* Oral Oncol, 2008. **44**(4): p. 369-82.

43. Yang, Y., et al., *Databases and web tools for cancer genomics study*. Genomics Proteomics Bioinformatics, 2015. **13**(1): p. 46-50.

44. *Mission and Goal-TCGA* [cited 2016 July 7]; Available from: http://cancergenome.nih.gov/abouttcga/overview/missiongoal.

45. Chang, J.T., Y.M. Lee, and R.S. Huang, *The impact of the Cancer Genome Atlas on lung cancer*. Transl Res, 2015. **166**(6): p. 568-85.

46. *TCGA Data Portal - Available Cancer Types*. [cited 2016 July 20th]; Available from: https://tcga-data.nci.nih.gov/docs/publications/tcga/.

47. Parfenov, M., et al., *Characterization of HPV and host genome interactions in primary head and neck cancers*. Proc Natl Acad Sci U S A, 2014. **111**(43): p. 15544-9.

48. Cancer Genome Atlas Network, *Comprehensive genomic characterization of head and neck squamous cell carcinomas*. Nature, 2015. **517**(7536): p. 576-82.

49. TCGA. *Data Levels and Data Types*. [cited 2016 24th July]; Available from: https://tcga-data.nci.nih.gov/docs/publications/tcga/datatype.html.

50. Nowak, D., W.K. Hofmann, and H.P. Koeffler, *Genome-wide Mapping of Copy Number Variations Using SNP Arrays*. Transfus Med Hemother, 2009. **36**(4): p. 246-251.

51. Mao, X., B.D. Young, and Y.J. Lu, *The application of single nucleotide polymorphism microarrays in cancer research*. Curr Genomics, 2007. **8**(4): p. 219-28.

52. Bengtsson, H., P. Wirapati, and T.P. Speed, *A single-array preprocessing method for estimating full-resolution raw copy numbers from all Affymetrix genotyping arrays including GenomeWideSNP 5 & 6*. Bioinformatics, 2009. **25**(17): p. 2149-56.

53. Trevino, V., F. Falciani, and H.A. Barrera-Saldana, *DNA microarrays: a powerful genomic tool for biomedical and clinical research*. Mol Med, 2007. **13**(9-10): p. 527-41.

54. Schapire, R. *Theoretical Machine Learning*. 2008 [cited 2016 1st August]; Available from: http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf.

55. Olivier Bousquet, U.v.L., Gunnar Rätsch, *Advanced Lectures on Machine Learning*. 2004: Springer-Verlag.

56. Ian H. Witten, E.F., Mark A. Hall, *Data Mining*. 2011, Burlington, USA: Morgan Kaufman

57. Kirk, M., *Thoughtful Machine Learning*. 1st ed. 2014: O'Reilly Media, Inc.

58. Bewick, V., L. Cheek, and J. Ball, *Statistics review 12: survival analysis.* Crit Care, 2004. **8**(5): p. 389-94.

59. Rodríguez, G. *Lecture Notes on Generalized Linear Models* 2007 [cited 2016 2nd August]; Available from: http://data.princeton.edu/wws509/notes/.

60. Jager, K.J., et al., *The analysis of survival data: the Kaplan-Meier method.* Kidney Int, 2008. **74**(5): p. 560-5.

61. Vigneswaran, N. and M.D. Williams, *Epidemiologic trends in head and neck cancer and aids in diagnosis.* Oral Maxillofac Surg Clin North Am, 2014. **26**(2): p. 123-41.

62. Patmore, H.S., et al., *Can a genetic signature for metastatic head and neck squamous cell carcinoma be characterised by comparative genomic hybridisation?* Br J Cancer, 2004. **90**(10): p. 1976-82.

63. Bockmuhl, U., et al., *Chromosomal alterations during metastasis formation of head and neck squamous cell carcinoma.* Genes Chromosomes Cancer, 2002. **33**(1): p. 29-35.

64. Akervall, J.A., et al., *Amplification of cyclin D1 in squamous cell carcinoma of the head and neck and the prognostic value of chromosomal abnormalities and cyclin D1 overexpression.* Cancer, 1997. **79**(2): p. 380-9.

65. Chen, Y. and C. Chen, *DNA copy number variation and loss of heterozygosity in relation to recurrence of and survival from head and neck squamous cell carcinoma: a review.* Head Neck, 2008. **30**(10): p. 1361-83.

66. van der Riet, P., et al., *Frequent loss of chromosome 9p21-22 early in head and neck cancer progression.* Cancer Res, 1994. **54**(5): p. 1156-8.

67. Ambatipudi, S., et al., *Genomic profiling of advanced-stage oral cancers reveals chromosome 11q alterations as markers of poor clinical outcome.* PLoS One, 2011. **6**(2): p. e17250.

68. *APPL1 Gene.* [cited 2016 September 6]; Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=APPL1.