UNIVERSITAT POLITÈCNICA DE VALÈNCIA

ESCOLA TÈCNICA SUPERIOR D'ENGINYERIA AGRONÒMICA I DEL MEDI NATURAL

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Identification of blood-based tumour specific markers for early detection of lung, oesophageal and head and neck cancer

---

TRABAJO FIN DE GRADO EN BIOTECNOLOGÍA

Curso 2019 – 2020

Alumno: Jorge Mota Pino

Tutora: Esther Giraldo Reboloso

Cotutora: Rosa Farràs Rivera

Valencia, 6 de julio de 2020

TITLE: Identification of blood-based tumour specific markers for early detection of lung, oesophageal and head and neck cancer.

SUMMARY

Lung, oesophageal and head and neck cancer are the first, seventh and sixth most frequently diagnosed cancers worldwide, respectively. In most cases, diagnosis of these cancers occurs at advanced stages, when surgical resection is no longer possible. The sharp decrease in the 5-year survival rate observed when comparing non-metastatic cases with widespread tumours highlights that early detection is a key factor in reducing the number of cancer-related deaths. In recent years, the numerous advantages of liquid biopsy compared to tissue biopsy (the former being minimally invasive and low risk) have generated a great level of interest regarding the use of this technique to diagnose cancer. As a result, the development of multi-analyte blood tests has made it possible to identify driver mutations in circulating tumour DNA (ctDNA) or specific surface markers in circulating tumour cells (CTCs), among others. According to this, the main objective of the present study is the development and subsequent bibliographic validation of a multigene panel comprised of genes that are overexpressed in tumour cells of four specific cancer types (lung adenocarcinoma, lung squamous cell carcinoma, head and neck squamous cell carcinoma and oesophageal carcinoma) in order to develop a test that allows early identification of CTCs in blood samples obtained from cancer patients, thereby facilitating an early diagnostic method specific for these cancers.

To generate the aforementioned multigene panel, expression data obtained from different cancer genetic databases (among which GEPIA stands out) was analysed. The main criteria for gene selection were high expression of the candidate gene in specific cancer cells and minimum expression not only in normal cells, but also in tumour cells from different cancers. Then, an extensive bibliographic search was conducted to gather relevant information regarding the suitability of the candidate genes. Last of all, primers for amplification of the selected genes were designed, with a short-term goal to validate them in cancer cell lines.

The proposed multigene panel was divided in 5 different subgroups. The first contains 12 genes whose combined expression in CTCs could allow for detection of any of the four cancer types assessed in this study. Following the same approach, the aim of the rest of subgroups is to allow identification of just one specific cancer type. In the case of lung cancer, 5 genes were chosen for lung adenocarcinoma and 4 genes for lung squamous cell carcinoma detection. For head and neck squamous cell carcinoma diagnosis, the combination of 6 different genes was found to be very promising, whilst only 4 genes were identified as promising candidates for diagnosis of oesophageal carcinoma.

The results presented in this study confirm that a combination of different markers is essential in order to obtain a successful multiparameter blood test that allows early detection of a specific cancer type. Further studies confirming gene expression data in solid tumour samples and cancer patients' CTCs would be required so that the suitability of the developed multigene panel for early cancer diagnosis is experimentally corroborated.

KEYWORDS: cancer; liquid biopsy; circulating tumour cells (CTCs); non-invasive biopsy; detection test.

Author: Jorge Mota Pino.

Location and date: Valencia, July 2020.

Academic tutor: Dra. Esther Giraldo Reboloso.

Cotutor: Dra. Rosa Farràs Rivera.

TÍTULO: Identificación de marcadores específicos de tumor en sangre para la detección precoz del cáncer de pulmón, esófago y cabeza y cuello.

RESUMEN

Los cánceres de pulmón, esófago y cabeza y cuello se encuentran entre aquellos que presentan una mayor incidencia mundial, siendo el cáncer de pulmón el más frecuentemente diagnosticado. En la mayoría de casos, la detección de estos tipos de cáncer ocurre en etapas avanzadas, cuando ya no es posible extirpar el tumor mediante cirugía. Al comparar tumores no diseminados con metastásicos, el gran descenso que se observa en la tasa de supervivencia a 5 años evidencia la importancia de la detección precoz para la reducción del número de muertes causadas por cáncer. En los últimos años, las numerosas ventajas que ofrece la biopsia líquida en comparación con la biopsia tisular (es mínimamente invasiva, conlleva un bajo riesgo, etc.) han despertado un gran interés por esta técnica en el diagnóstico de pacientes con cáncer. En esta línea se han desarrollado análisis de sangre multianalito que permiten determinar mutaciones *driver* en el ADN tumoral circulante (ctDNA) o marcadores de superficie específicos en células tumorales circulantes (CTCs). El objetivo principal del presente estudio es la elaboración y posterior validación bibliográfica de un panel de genes que se encuentren sobreexpresados en células tumorales de cuatro tipos de cáncer específico (adenocarcinoma de pulmón, cáncer de pulmón de célula escamosa, cáncer de cabeza y cuello de célula escamosa y cáncer de esófago), con el fin de desarrollar un test que permita la identificación precoz de CTCs en muestras de sangre de pacientes con cáncer, consiguiendo así un método de diagnóstico temprano y específico del tipo de cáncer en cuestión.

Para la obtención del panel de genes se analizaron datos de expresión provenientes de distintas bases de datos genéticas, entre las que destaca *GEPIA*, aplicando como criterio prioritario para la selección de cada gen una alta expresión en células del cáncer de interés y una mínima expresión tanto en células de tejido sano como en células tumorales de un cáncer distinto al de interés. Después, se llevó a cabo una amplia búsqueda bibliográfica para recopilar información que avalara la idoneidad de los genes candidatos. Por último, se llevó a cabo el diseño de *primers* para la amplificación de los genes seleccionados, en vistas a realizar una validación futura de los mismos en diferentes líneas celulares tumorales.

El panel de genes propuesto se dividió en 5 subgrupos. El primero contiene 12 genes y se basa en que la expresión combinada de los mismos en CTCs permita la detección de cualquiera de los cuatro tipos de cáncer evaluados. Siguiendo este mismo planteamiento, el objetivo del resto de subgrupos es la identificación temprana de un solo tipo específico de cáncer. Así, se seleccionaron 5 genes para la detección del adenocarcinoma de pulmón y 4 para la del cáncer de pulmón de célula escamosa. Por otro lado, para diagnosticar el cáncer de cabeza y cuello de célula escamosa la combinación de 6 genes demostró ser muy prometedora, mientras que solo 4 genes fueron seleccionados para el diagnóstico del cáncer de esófago.

Los resultados presentados en este estudio confirman que la combinación de diferentes marcadores es esencial a la hora de obtener un test que permita la detección precoz de un tipo específico de cáncer a partir de muestras de sangre. Sin embargo, con el fin de corroborar de manera experimental la idoneidad del panel de genes propuesto sería necesario realizar futuros estudios que confirmen los datos de expresión de los genes seleccionados en muestras de tumor sólido y CTCs procedentes de pacientes.

PALABRAS CLAVE: cáncer; biopsia líquida; células tumorales circulantes (CTCs); biopsia no invasiva; test de detección.

Autor: Jorge Mota Pino.
Localidad y fecha: Valencia, julio 2020.
Tutora académica: Dra. Esther Giraldo Reboloso.
Cotutora: Dra. Rosa Farràs Rivera.

INDEX

FIGURE INDEX

TABLE INDEX

ABBREVIATIONS

ADC: Adenocarcinoma.

ALK: Anaplastic Lymphoma Kinase.

APC/C: Anaphase Promoting Complex/Cyclosome.

bp: base pairs.

CK: Cytokeratin.

cfDNA: circulating free DNA.

CRC: Colorectal Cancer.

CT: Computed Tomography.

CTC: Circulating Tumour Cell.

ctDNA: circulating tumour DNA.

CTM: Circulating Tumour Microemboli.

DAPI: 4',6-diamidino-2-phenylindole.

DEP-FFF: Dielectrophoretic Field-Flow Fractionation.

DNase: Deoxyribonuclease.

EBI: European Bioinformatics Institute.

EC: Endometrial Cancer.

ECM: Extracellular Matrix.

EBV: Epstein-Barr Virus.

EGFR: Epidermal Growth Factor Receptor.

EMT: Epithelial-Mesenchymal Transition.

EMBL: European Molecular Biology Laboratory.

EpCAM: Epithelial Cell Adhesion Molecule.

FC: Fold Change.

FDA: Food and Drug Administration.

GEPIA: Gene Expression Profiling Interactive Analysis.

GTEx: Genotype-Tissue Expression.

HNC: Head and Neck Cancer.

HNSqCC: Head and Neck Squamous Cell Carcinoma.

HPA: Human Protein Atlas.

HPV: Human Papillomavirus.

IS: Immune System.

LADC: Lung Adenocarcinoma.

LB: Liquid Biopsy.

LC: Lung Cancer.

LCC: Large Cell Carcinoma.

lncRNA: long non-coding RNA.

LSqCC: Lung Squamous Cell Carcinoma.

log_2FC: log_2 Fold Change.

MMP: Matrix Metalloproteinase.

MRI: Magnetic Resonance Imaging.

NCBI: National Centre for Biotechnology Information.

NCI: National Cancer Institute.

NHS: National Health Security.

NSCLC: Non-Small Cell Lung Carcinoma.

NTS: Neurotensin.

OADC: Oesophageal Adenocarcinoma.

OC: Oesophageal Cancer.

OSqCC: Oesophageal Squamous Cell Carcinoma.

OVC: Ovarian Cancer.

PB: Peripheral Blood.

PBMC: Peripheral Blood Mononuclear Cell.

PC: Pancreatic Cancer.

PET: Positron Emission Tomography.

PPV: Positive Predictive Value.

PRC1: Polycomb Repressive Complex 1.

RT-PCR: Reverse Transcription Polymerase Chain Reaction.

RT-qPCR: Reverse Transcription Quantitative Polymerase Chain Reaction.

SqCC: Squamous Cell Carcinoma.

SCLC: Small Cell Lung Carcinoma.

SEOM: Sociedad Española de Oncología Médica (Spanish Society of Medical Oncology).

TEP: Tumour Educated Platelet.

Tm: Melting Temperature.

tpm: transcripts per million.

VEGF: Vascular Endothelial Growth Factor.

WBC: White Blood Cell.

WHO: World Health Organization.

1. INTRODUCTION

1.1. Concept of cancer

Cancer is defined as a collection of related diseases in which the body's cells divide without control and spread into surrounding tissues. Cancer cells may also spread to other parts of the body through the blood and lymph systems. Since cancer can start almost anywhere in the human body, there are several types of cancer: carcinoma, which begins in skin or tissues that line or cover internal organs; sarcoma, which starts in bone, muscle, fat, blood vessels or other connective tissue; leukaemia, which begins in blood-forming tissue (mainly in the bone marrow) and causes abnormal blood cells to enter the blood; lymphoma, which begins in cells of the immune system; and central nervous system cancers, which start in the brain and spinal cord (National Cancer Institute; NCI, 2015).

Cancer incidence and mortality are rapidly growing worldwide. The reasons are complex but reflect aging, growth of the population and changes in the prevalence and distribution of the main risk factors for cancer, many of which are associated with socioeconomic development (Bray *et al*., 2018). According to the World Health Organization (2020), cancer is already the first or second leading cause of premature death in 134 of 183 countries, being responsible for one in six deaths globally. In 2018, there were an estimated 9.6 million deaths and 18.1 million new cases of cancer worldwide (World Health Organization; WHO, 2020).

1.1.1. Molecular biology of cancer

The process by which normal cells are progressively transformed into tumour cells is known to require a sequential acquisition of mutations that arise as a consequence of DNA damage. This damage can be the result of endogenous processes (e.g. errors in DNA replication), or interactions with exogenous mutagens (e.g. UV radiation, chemical carcinogens, ionising radiation, etc.). Even though cells have developed means to repair such damage, the fact that some of these mutations affect genes responsible for genome integrity maintenance has facilitated the acquisition of permanent changes in the genome. These mutations can be acquired, being the most frequent cause of cancer, or inherited, although this last case is less frequent (Bertram, 2000).

This way, tumorigenesis is a multistep process involving a succession of genetic alterations, each conferring one or another type of growth advantage, that drives the progressive transformation of normal cells into highly malignant derivatives. Hanahan & Weinberg suggested in 2000 that there existed six essential alterations in cell physiology dictating malignant growth: self-sufficiency in growth signals, insensitivity to anti-growth signals, evasion of apoptosis, limitless replicative potential, sustained angiogenesis and tissue invasion and metastasis. These novel capabilities acquired in the course of tumour development, known as hallmarks of cancer, represent the successful breaching of the main anticancer defence mechanisms hardwired into cells (Hanahan & Weinberg, 2000).

Even though nowadays the aforementioned six hallmarks of cancer continue to provide a solid foundation for the comprehension of the molecular biology of cancer, Hanahan & Weinberg reviewed these essential alterations in 2011 (Hanahan & Weinberg, 2011) on the basis of the remarkable progress in cancer research subsequent to their first publication, and added four new hallmarks to the list: genome instability and mutation, cellular energetics deregulation, evasion of the immune system (IS) and tumour-promoting inflammation (*Figure 1)*.

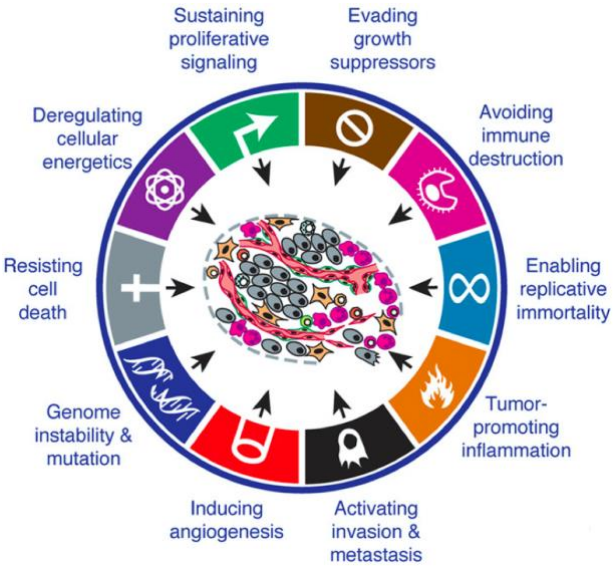


Figure 1. *The ten hallmarks of cancer described by Hanahan & Weinberg.* Adapted from Hanahan & Weinberg, 2011.

It is important to note that the biology of tumours cannot be completely understood by simply paying attention to the traits of cancer cells, but instead must take into account the contributions of the tumour microenvironment. In this respect, tumours constitute complex tissues comprising several distinct cell types that participate in heterotypic interactions with one another. For instance, normal cells forming the tumour-associated stroma play an important role as active participants in tumorigenesis and even contribute to the expression and development of certain hallmark capabilities (Hanahan & Weinberg, 2011).

1.2. Lung cancer

1.2.1. Incidence and risk factors

Lung cancer (LC) was the most frequently diagnosed type of cancer worldwide in 2018 (11.6% of all cases). It was also responsible for 18.4% of all cancer deaths that year, therefore accounting as the deadliest cancer type worldwide (*Figure 2*) (WHO, 2020).

In Spain, researchers estimate that out of 277,000 newly diagnosed cancer cases in 2020, almost 30,000 will be LC, therefore becoming the fourth most frequently diagnosed cancer type in the country, only after colorectal, prostate and breast cancer. Following the worldwide trend, LC was also the leading cause of death from cancer in Spain during 2018. It is important to note that, whilst the number of LC cases and deaths in men has seen an important reduction in the last decade due to a decrease in the smoking habit, the contrary situation is now observed in women, as the increase in female smokers in the 1970s is starting to become evident today. In fact, LC has changed from being the country's fourth most frequently diagnosed cancer type in women in 2015 to being third in 2019. It also accounted as the second cause of death by cancer in women in 2018 (only after breast cancer), having displaced colon cancer to third position (SEOM, 2020).

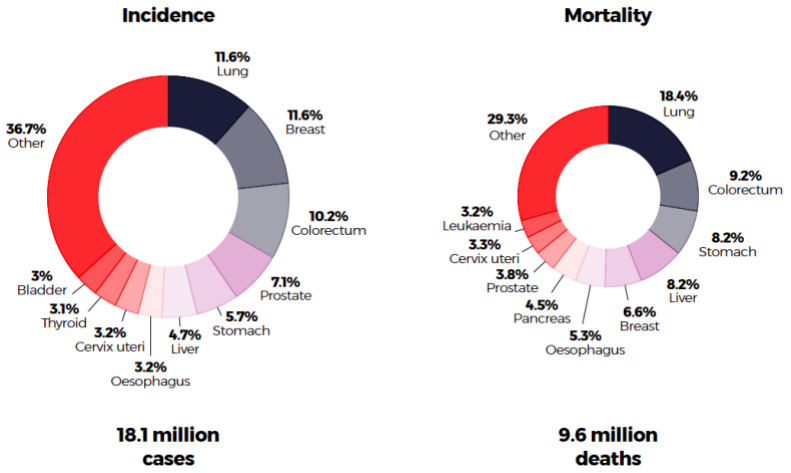


Figure 2. *Distribution of cases and deaths caused by the leading 10 cancer types for both sexes in 2018.* Retrieved from WHO, 2020.

Regarding risk factors, tobacco consumption is by far the main etiological factor in lung carcinogenesis (de Groot *et al.*, 2018; Malhotra *et al.*, 2016). In fact, it is estimated that approximately 90% of LC deaths in men and 80% of LC deaths in women are caused by tobacco smoking (Ridge *et al.*, 2013). In addition, there are other important risk factors, such as occupational exposure to lung carcinogens, being the most frequent asbestos, radon, silica and heavy metals (Malhotra *et al.*, 2016). In France, the estimated proportion of LC cases attributable to occupational agents in men is 12.5% (Boffeta *et al.*, 2010), and in the UK this number rises to 14.5% for both sexes (Rushton *et al.*, 2012). Air pollution, ionising radiation, alcohol consumption and dietary habits are also important etiological factors that explain many cases of LC in non-smokers. Genetic risk factors may also play an important role since several studies have demonstrated that the presence of specific single nucleotide polymorphisms increases the susceptibility to get certain types of LC (Malhotra *et al.*, 2016).

1.2.2. Diagnosis

LC diagnosis is primarily based on symptoms, and the detection normally happens when curative interventions are no longer possible (Jantus-Lewintre *et al.*, 2012). The most common LC symptoms are dyspnoea, cough, haemoptysis and systemic symptoms, such as weight loss (Latimer & Mott, 2015). However, signs and symptoms are notoriously variable and highly dependent on the tumour type and the extent of metastases (Collins, 2007).

In an initial evaluation, non-invasive techniques are used: history and physical examination, complete blood count and imaging techniques such as chest radiography or contrast-enhanced computed tomography (Collins, 2007; Latimer & Mott, 2015). These are followed by a diagnostic evaluation, which comprises three simultaneous steps: tissue diagnosis, staging and functional evaluation. The choice of procedure for tissue sample acquisition depends on the type, location and size of the tumour, but the general rule is to use the least invasive method (Latimer & Mott, 2015). Using all the collected information, staging is then performed, generally following the 7th edition of the TNM system, which is based on the primary tumour size in the long axis (T), the degree of spread to regional lymph nodes (N) and the presence of metastases beyond regional lymph nodes (M). Finally, functional evaluation of the patient is required in order to determine the most appropriate treatment (Latimer & Mott, 2015; Mirsadraee *et al.*, 2012).

1.2.3. Pathology and classification

Lung cancers are traditionally divided into two broad histologic classes, which grow and spread differently: small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC), with the former accounting for approximately 15% of cases and the latter for the remaining 85% (Lemjabbar-Alaoui *et al*., 2015; Zheng, 2016). The advent of molecular profiling and targeted therapy has led to further subclassification of NSCLC into adenocarcinoma (ADC), squamous cell carcinoma (SqCC) and large cell carcinoma (LCC) (Rodríguez-Canales *et al.,* 2016). The prevalence, subtype and anatomic location of the most important LC types are shown in *Table 1*.

Table 1. *Histologic classification of the most prevalent lung cancer types.* SCLC: Small Cell Lung Carcinoma. Adapted from Lemjabbar-Alaoui *et al*., 2015.

Lung cancer type		Prevalence	Origin and development	Subtypes
Non-small cell lung carcinomas	Adenocarcinomas (ADCs)	40 %	Arise in peripheral bronchi and advance by producing lobar atelectasis and pneumonitis.	Bronchioalveolar, acinar, papillary, solid with mucus formation, mixed.
	Squamous cell carcinomas (SqCCs)	25 – 30 %	Arise in the main bronchi and advance to the carina.	-
	Large cell lung carcinomas (LCLCs)	10 %	Lack the classic glandular or squamous morphology, are more proximal in location and tend to invade the mediastinum in early stages.	Large cell neuroendocrine, lymphoepithelial-like, basaloid, large cell with rhabdoid phenotype.
Small cell lung carcinomas (SCLCs)		10 – 15 %	Derive from hormonal cells. Tend to be central mediastinal tumours, are extremely aggressive and disseminate rapidly into submucosal lymphatic vessels and regional lymph nodes.	Pure SCLC, combined SCLC.

In the last 20 years, important progresses have been made in the understanding of the molecular alterations underlying NSCLCs. Both ADCs and SqCCs are characterised by a high average number of somatic mutations per megabase in comparison with several other tumour types. Identification of these mutations has allowed for implementation of precision oncology clinical trials aimed at matching patients with specific targeted therapies based on identification of genomic driver events, which have resulted in an improvement of clinical outcomes (Skoulidis & Heymach, 2019). In general, the more frequently mutated genes in NSCLC with known potential function of driver genes are: *EGFR* (10−30% of cases), *KRAS* (15−30%), *FGFR1* (20%), *PIK3CA* (2−5%), *ERBB2* (*HER2*) (2−5%), *BRAF* (1−3%), *ALK* (3%), *ROS1* (1%), *MAP2K1/MEK1* (1%), *RET* (1%), *NRAS* (1%) and *AKT1* (< 1%) (Testa *et al.*, 2018). In *Figure 3*, an "oncogenic pie chart" exhibits the prevalence of individual genomic alterations in early-stage lung ADC (Skoulidis & Heymach, 2019).

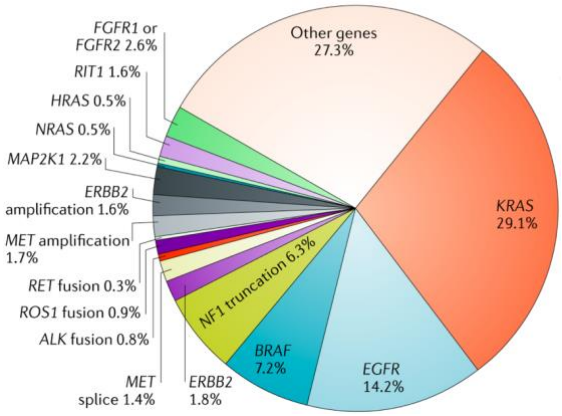


Figure 3. *Prevalence of individual genomic alterations in early-stage lung adenocarcinoma.* Retrieved from Skoulidis & Heymach, 2019.

1.2.4. Treatment and prognosis

Treatment and prognosis are closely tied to the tumour type and stage (Collins *et al.*, 2007). The recommended treatment for patients with NSCLC in stages I-II is surgery, with a 5-year survival of 77-92% for stage IA, 68% for stage IB, 60% for stage IIA and 53% for stage IIB. Advanced NSCLC (stages IIIA-B) in patients not amenable to surgical resection is treated with a multi-modality approach that comprises thoracic radiotherapy combined with concurrent delivery of doublet chemotherapy using either carboplatin or cisplatin and a second drug. The 5-year survival drops in these cases to 15-20% for stage IIIA and 3-7% for stage IIIB (Hirsch *et al.*, 2017). In patients with stage IV NSCLC, the treatment of choice depends on many factors, including histology, comorbidity and molecular genetic features of the cancer. The standard treatment options include radiation therapy, combination chemotherapy, laser therapy or targeted therapies with epidermal growth factor receptor (EGFR), anaplastic lymphoma kinase (ALK), vascular endothelial growth factor (VEGF), receptor tyrosine kinase C-ROS oncogene 1 (ROS1) or proto-oncogene B-Raf (BRAF) inhibitors, depending on the case (Lemjabbar-Alaoui *et al*., 2015).

Regarding SCLCs, in most cases they are treated non-surgically, being chemotherapy in combination with radiotherapy the mainstay of treatment for this type of LC. Even though SCLC is initially more responsive to these therapies than all other LC types, it usually behaves very aggressively and is widely disseminated at the time of diagnosis, which makes it very difficult to treat. This fact explains that the overall survival at 5 years of all population of SCLC patients is 5-10% (Lemjabbar-Alaoui *et al*., 2015)

1.3. Oesophageal and head and neck cancer

1.3.1. Pathology and classification

Oesophageal cancer (OC) typically involves malignancy that arises from the epithelium or surface lining of the oesophagus. The two most common histological types of OC include oesophageal squamous cell carcinoma (OSqCC), which arises from cells that line the upper part of the oesophagus; and oesophageal adenocarcinoma (OADC), which arises from glandular cells localised at the junction between the oesophagus and the stomach (Mao *et al.*, 2011; Napier *et al.*, 2014). Less than 2% of all oesophageal cancers are categorised as sarcomas or small cell carcinomas. Globally, OSqCC is the most common type of OC and accounts for the vast majority of cases (Napier *et al.*, 2014).

Head and neck cancer (HNC) is not a specific entity, but a broad category of diverse tumour types arising from different anatomic structures. The vast majority (over 90%) begin in the squamous cells that line the mucosal surfaces inside the head and the neck. This way, HNC includes those cancers that start in the oral cavity, pharynx, larynx, paranasal sinuses, nasal cavity and salivary glands (*Figure 4)*, being the latter location relatively uncommon (NCI, 2017). The microscopic appearance in head and neck squamous cell carcinoma (HNSqCC) may vary, but the prototypic HNSqCC is moderately differentiated. There are 3 subtypes of HNSqCC: the basaloid variant, the spindle-cell variant and the papillary variant (Pai & Westra, 2009).

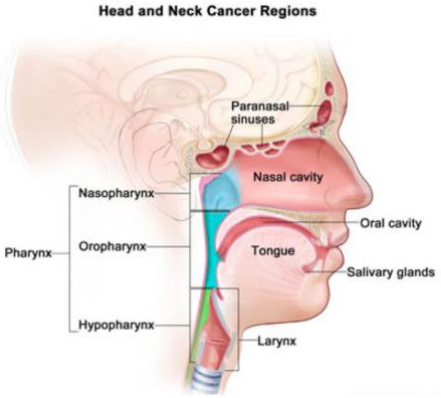


Figure 4. *Head and neck cancer regions.* Retrieved from NCI, 2017.

1.3.2. Incidence and risk factors

OC is the seventh most common cancer and accounts for 5.3% of deaths caused by cancer worldwide (*Figure 2*), with approximately 570,000 new cases and 500,000 deaths in 2018 (WHO, 2020). In Spain the situation is quite different, since OC is not among the top 15 most frequently diagnosed cancers. It is estimated that in 2020 only 0.86% of all cancers diagnosed in the country will correspond to this type. Similarly, in 2018 OC only accounted for 1.61% of all cancer deaths in Spain. It is important to note that over 80% of OC new cases and deaths occur in men (SEOM, 2020).

Even though nowadays OSqCC continues to be responsible for the majority of OC cases worldwide, it is important to note that in the past decades a marked histologic shift has taken place, especially in developed countries. For instance, before 1990 OADC accounted for less than 15% of all OC cases in the US, while nowadays this subtype accounts for more than 60%. This shift has led to the current situation, where generally the majority of OSqCC cases occur in developing countries whereas OADC is mainly diagnosed in developed countries (Mao *et al.*, 2011; Umar & Fleischer, 2008).

Regarding OSqCC incidence, there is a vast geographic variation globally, with up to 10-fold differences in areas located less than a few hundred kilometres apart. In Western countries the overall incidence of OSqCC is currently quite low, with Asian countries being responsible for the majority of cases worldwide. This way, the populations at higher risk of OSqCC are in north and central China, north-eastern Iran and countries between the two, comprising the so-called "Asian belt" of OC (Umar & Fleischer, 2008).

These differences between territories in the incidence and type of OC can be linked to exposure to different risk factors (Napier *et al.*, 2014). The risk factors associated with OC are well known, and even though OSqCC and OADC share some of these factors, there are unique elements that contribute to the development of each histologic type (Mao *et al.*, 2011; Umar & Fleischer, 2008). Risk factors associated with OSqCC include gender and race, smoking, alcohol consumption, dietary components and genetic aspects (*Supplementary Table 1, Appendix I)*. In the case of OADC, it is associated with gender and race, gastroesophageal reflux disease,

Barrett's oesophagus (a metaplastic transformation of normal stratified squamous oesophageal epithelium), obesity, tobacco, nutritional deficit, drugs and genetic aspects (Domper Arnal *et al.*, 2015).

Before talking about incidence and risk factors of HNC, some considerations need to be made. Because cancers of the head and neck usually appear subdivided in cancer incidence databases depending on their origin (larynx cancer, pharynx cancer, etc.), HNC does rarely appear among the most frequently diagnosed cancers. However, if all the cancer types that HNC comprises are considered altogether, in 2018 there were approximately 900,000 new HNC cases and almost half a million deaths caused by HNC globally, therefore becoming the sixth most common cancer worldwide (WHO, 2020). In 2018, HNC accounted for 3.25% of all cancer deaths in Spain (3,671 out of 112,714). The country seems to follow the global trend, since HNC is also the sixth most frequently diagnosed cancer in Spain. Similar to OC, more than 75% of HNC new cases and deaths occur in men (SEOM, 2020).

The development of HNSqCC is associated with exposure to carcinogens, diet, oral hygiene, family history or infectious agents. Of these, alcohol and tobacco use (including smokeless tobacco, also known as "chewing tobacco", and passive smoking) are the dominant risk factors, being the direct cause of at least 75% of all HNCs (NCI, 2017; Pai & Westra, 2009). The nitrosamines and polycyclic hydrocarbons present in tobacco smoke seem to be responsible for this increased risk. Interestingly, even though heavy alcohol consumption alone is recognised as an independent risk factor for HNSqCCs, it is more relevant for its ability to magnify the effects of tobacco smoke synergistically. This ability of alcohol more likely resides in its nature as a chemical solvent, which enhances and prolongs mucosal exposure to tobacco smoke carcinogens. Although tobacco and alcohol account for the vast majority of HNSqCCs originated in the oral cavity, larynx and hypopharynx, their role in oropharyngeal cancer is less consequential. Instead, human papillomavirus (HPV), particularly type 16, is known to be the causative agent of up to 70% of cancers originated in the oropharynx. In this sense, the role of HPV-16 in oropharyngeal cancer has become more evident in the last decades where the number of tobacco smokers has substantially decreased in many areas of the world (Pai & Westra, 2009). HPV carcinogenesis occurs due to the action of viral proteins E6 and E7. E6 is capable of degrading the tumour suppressor p53, which leads to uncontrolled proliferation and genomic instability, among other effects. Similarly, E7 is capable of degrading the tumour suppressor retinoblastoma, causing a dysregulation of the cell cycle often associated with the first steps leading to tumorigenesis (Alfouzan, 2019). Other risk factors for HNC include betel quid, consumption of certain preserved or salted foods during childhood, poor oral hygiene, occupational exposure to dust, asbestos or synthetic fibres, exposure to radiation and Epstein-Barr virus (EBV) infection (NCI, 2017).

1.3.3. Diagnosis

Because there are no specific symptoms of early OC, most oesophageal cancers are diagnosed after development of dysphagia (reduction of oesophagus lumen by 50%), when tumours are locally advanced. This way, only one in eight oesophageal cancers are identified at an early stage (T1), either by incidental finding during a gastroscopy performed for other reasons or by means of a Barrett's oesophagus surveillance programme. Apart from dysphagia, other typical symptoms of OC include vomiting, weight loss, dysphonia and gastrointestinal bleeding. The two main tests used to diagnose OC are gastroscopies and X-rays after barium swallow, being the former the gold standard for OC diagnosis. These gastroscopies may include a tissue biopsy in suspect areas to confirm the diagnosis of the endoscopist. OSqCC is more likely to be found in the upper and middle part of the oesophagus, whereas OADC is usually detected in the lower part. Further tests, such as computed tomography (CT) scan, endoscopic ultrasound scan,

positron emission tomography (PET) scan or laparoscopy may be carried out for staging of OC (Meves *et al.*, 2015; NHS, 2019).

In the case of HNC, diagnosis at early stages is also difficult since patients present with vague symptoms and minimal physical findings. The clinical presentation and the time at which symptoms can be detected will depend on the primary site involved (*Table 2*). In this sense, cancers of the glottis and the oral cavity are usually diagnosed at an early stage, whilst patients with cancer of the hypopharynx or the supraglottis present symptoms quite late in the course of the disease (Marur & Forastiere, 2008).

Table 2. *Main clinical presentations of head and neck squamous cell carcinoma.* Retrieved from Marur & Forastiere, 2008.

Subsite	Clinical presentation
Oral cavity	Sores, ulcers, pain.
Oropharynx	Sore throat, chronic dysphagia otalgia, odynophagia.
Hypopharynx	Soreness, otalgia, dysphagia, hoarseness.
Larynx	Persistent hoarseness, shortness of breath.
Supraglottis	Neck mass.
Nasopharynx	Otitis media unresponsive to antibiotics, nasal obstruction, epitaxis.

If mucosal abnormalities or lumps are detected on a physical examination, a nasopharyngolaryngoscopy is generally performed, followed by a tissue biopsy (usually fine needle aspiration) for cytologic examination of tumour cells. Once a diagnosis has been established, the extent of the disease needs to be determined for accurate staging. Imaging techniques are normally used for this purpose, such as CT, magnetic resonance imaging (MRI) or PET. In fact, PET has become a useful diagnostic technique for both initial staging and restaging of HNC, as it can be used not only to localise and stage an unknown primary tumour, but also to identify persistent disease after treatment (Marur & Forastiere, 2008).

1.3.4. Treatment and prognosis

The management of OC is challenging not only in terms of detecting the tumour at an early stage, but also because of the overall poor prognosis of the disease. In this sense, accurate staging information is crucial for establishing the appropriate treatment choice (Meves *et al.*, 2015). Surgical resection can be a definitive treatment for Tis (carcinoma *in situ*), T1 and some T2 OCs. Oesophageal mucosal resection is normally the treatment of choice if OC is diagnosed very early on, as it is a much less aggressive surgery compared to oesophagectomy, which comprises the removal of a section of the oesophagus. However, the latter is normally the only choice for T2 OC. In the case of tumours in stages III and IV, neoadjuvant treatment (usually chemotherapy or chemoradiotherapy) is used to render the tumour resectable by surgical excision (Napier *et al.*, 2014).

Regarding mortality and prognosis, OC is a serious malignancy due to its aggressiveness and poor survival and its incidence is expected to increase over the next 10 years. Because only 12.5% of oesophageal tumours are found at a stage where they are endoscopically resectable, this type of cancer has the sixth worst prognosis worldwide (Meves *et al.*, 2015; Napier *et al.*, 2014). The modest progress made in the past decades regarding treatment has translated into an increase in the overall 5-year survival rate from 5% in the 1960s to 20% nowadays. Not surprisingly, this rate changes depending on the stage at which the tumour is diagnosed, ranging

from 47% in early stages to 5% in cases where metastasis has already occurred (Huang & Yu, 2018).

HNC treatment requires a multidisciplinary approach with a team including a medical oncologist, a head and neck surgeon and a radiation oncologist. Determination of the stage and resectability of the tumour is essential to establish the most effective treatment. HNCs in early stages (T1 or T2 with no nodal involvement) are usually treated with surgery or radiation, depending on their location. For its part, intermediate-stage tumours (poor-prognosis T2 or exophytic T3) are treated with a combined-modality approach (normally radiotherapy followed by surgery or chemoradiotherapy), which is likely to provide better results. In patients with advanced tumours (T3 and T4 primary tumours with N2 or N3 lymphadenopathy), complete surgical excision and pre or post-operative radiation is used. In cases where the tumour is unresectable or preservation of the organ is desired, concurrent chemoradiation is the treatment of choice (Marur & Forastiere, 2008). It is important to mention that some novel strategies have become quite important in the last decades regarding HNC treatment. Since more than 90% of HNSqCCs overexpress EGFR, targeted therapy strategies have been developed to block EGFR function. In this sense, the use of a monoclonal antibody (cetuximab) directed against the extracellular receptor domain of EGFR (thus preventing agonist binding and dimerization of the receptor), in conjunction with chemotherapy or radiation therapy has shown promising results (Pai & Westra, 2009).

When considering all different types of HNC altogether, the overall 5-year survival rate according to the American Cancer Society (2020) is around 60%. However, there are significant differences depending on the origin of the tumour. In this sense, cancers of the salivary glands show the best overall 5-year survival rate (71%), whilst hypopharyngeal cancers show the worst prognosis, with an overall 5-year survival rate of only 33%. The importance of early diagnosis becomes evident once again since the 5-year survival rate improves by approximately 20% in tumours diagnosed at an early stage (*Table 3*).

Table 3. *5-year survival rate of the different types of head and neck cancer depending on the stage at which the cancer is diagnosed.* Adapted from American Cancer Society, 2020.

	5-year survival rate			
Subsite	**Overall**	**Local, no spread**	**Spread to surrounding tissues or lymph nodes**	**Spread to distant parts of the body**
Larynx	60%	77%	45%	33%
Hypopharynx	32%	59%	33%	21%
Nasal cavity & paranasal sinus	58%	84%	51%	42%
Nasopharynx	61%	82%	73%	48%
Oral cavity & oropharynx	65%	84%	66%	39%
Salivary glands	71%	94%	65%	35%

1.4. Liquid biopsy in oncology

Historically, cancers have been diagnosed, categorised and subclassified by means of histologic analyses of tumour tissue. In recent years, advances in the field of molecular oncology have expanded the array of tests available for accurate cancer diagnosis at the genomic level, essential for patient management and treatment decision. The era of precision cancer medicine,

boosted by the development of the so-called "-omics" technologies, has heightened the need for high-quality diagnostic material. Given the associated costs and risks of the assessment of cancer mutational profiles through fragments of tumour obtained by interventional biopsies, the discovery of circulating tumour DNA (ctDNA) and circulating tumour cells (CTCs) in blood, among others, have presented an attractive opportunity for minimally invasive and low risk liquid biopsy genomic diagnostics (Cescon *et al*., 2020; Poulet *et al.*, 2019).

1.4.1. Current status. Advantages and limitations.

The numerous advantages that liquid biopsy (LB) has to offer have generated a great level of interest regarding the use of this technique in the past few years. A quick search in Pubmed (https://pubmed.ncbi.nlm.nih.gov) using the search string (liquid AND biopsy) reveals a huge increase in the number of published articles in the last decade. This rise becomes more evident from 2014 on, since the number of search results from 2014 to 2019 has seen a fivefold increase in comparison with prior years.

The main advantage of LB when compared to tissue biopsy is the avoidance of surgical interventions which, apart from posing great risks for the patient, limit largely the possibility of biopsy sampling and are subjected to accessibility of tumour tissue (Poulet *et al.*, 2019). Furthermore, the fact that tissue biopsies are normally obtained from the primary tumour presents two problems. Firstly, because tumours tend to be heterogeneous there is a great likelihood that valuable information for treatment is missed (the most aggressive subclones may remain undetected). In contrast, LB reflects the broad range of malignancy's properties, including possible metastases that may not yet have been detected. Secondly, monitoring of tumour progression and evolution with time may require a serial obtention of samples, which is not easily feasible when using the invasive procedures required for tissue biopsy. However, repeated samples can be taken whenever necessary in the case of LB due to the minor invasiveness of the technique, that allows for systematic and real-time monitoring of the tumour's molecular alterations (Mader & Pantel, 2017; Poulet *et al*., 2019).

Despite the fact that the concept of LB has been introduced with the potential to revolutionise the management of cancer patients eliminating invasive interventions, there are some hurdles in the way before its full deployment (Costa & Schmitt, 2019). In this sense, the possibility to perform a histological analysis for staging of the tumour is limited to the obtention of CTCs in LB samples, whereas this problem does not exist when performing a conventional tissue biopsy. Furthermore, the low level of tumour-derived products in body fluids leads to a high risk of false negative results (Poulet *et al.*, 2019). A full comparison of the advantages and limitations of LB and conventional tissue biopsy is shown in *Supplementary Table 2 (Appendix I).*

1.4.2. Applications of liquid biopsy.

Even though the idea of LB was initially related with CTCs, this technique is now extended to cell-free circulating nucleic acids (DNA, mRNA, long non-coding RNAs, microRNAs, etc.), "tumour-educated platelets" (TEPs) or vesicles (mainly exosomes), among others (*Figure 5)*. Furthermore, their applications are not restricted to oncology, as this technique is also applicable to cardiovascular diseases, prenatal diagnosis or atherosclerosis (Mader & Pantel, 2017). However, it must be noted that most of the articles published nowadays in relation to LB are somehow cancer-related because of the quite interesting applications of this technique in the field, some of which have already been discussed: the possibility of early detection of the disease, real-time monitoring of responses and resistance to therapy, tumour heterogeneity analysis, etc. (Calabuig-Fariñas *et al*., 2016).

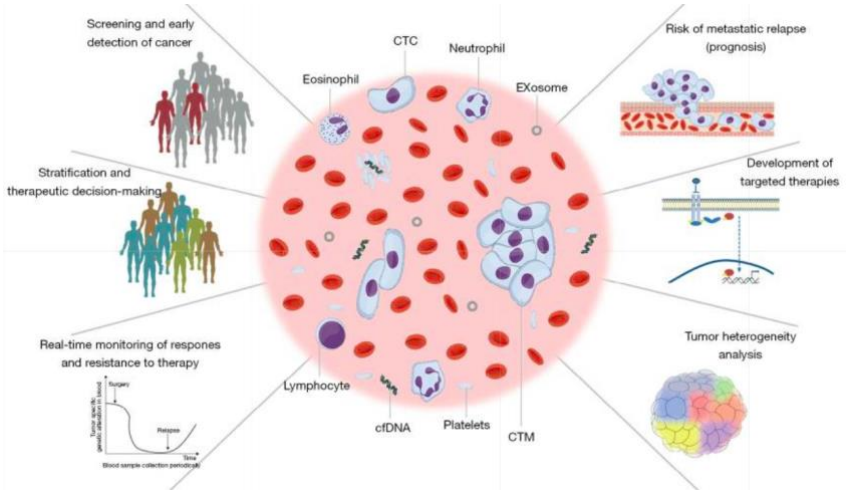


Figure 5. *Main applications of liquid biopsy in oncology.* CTC: Circulating Tumour Cell; cfDNA: circulating free DNA; CTM: Circulating Tumour Microemboli. Retrieved from Calabuig-Fariñas *et al.*, 2016.

1.4.3. Circulating tumour cells (CTCs). Isolation and detection.

First described in 1869 by the Australian pathologist Thomas Ashworth as a curiosity in the blood of a man with metastatic cancer, CTCs have now assumed, together with ctDNA, immense importance in liquid biopsy (Calabuig-Fariñas *et al.*, 2016; Katz *et al.*, 2020; Mader & Pantel, 2017). CTCs are cancer cells that detach from a primary tumour or metastatic lesion and spread via blood or lymphatic vessels to other parts of the body with a reported half-life of 1 to 2.4 hours. Their abundance is quite low (less than 10 cells/ml of blood), even in cases presenting with metastasis, and varies between tumour types (Poulet *et al.*, 2019). CTCs have been found in blood both as single cells and as cell clusters of 2 to 50 cells, known as circulating tumour microemboli (CTM). Since they are postulated to contain subpopulations of cells that can potentially initiate distant metastases, several models have appeared in order to describe the dissemination process to colonise distant organs (Calabuig-Fariñas *et al.*, 2016). In this sense, a deep analysis of these metastasis-initiating cells could shed light on the molecular processes of the metastatic cascade. Gaining plasticity and motility seems essential for intravasation and survival in the bloodstream, which usually involves epithelial-mesenchymal transition (EMT). This seems to be the reason why both single CTCs and CTMs show enrichment of mesenchymal markers, which would indicate increased plasticity and has been shown to be related with more aggressive behaviour, thus supporting their role in the initiation of metastatic outgrowth. (Calabuig-Fariñas *et al.*, 2016; Mader & Pantel, 2017). The presence of CTMs has been reported to be a negative prognostic factor in several types of cancer, such as LC, probably because tumour cells are protected by being enveloped in various cells derived from the tumour microenvironment, such as lymphocytes, neutrophils, platelets or macrophages (Calabuig-Fariñas *et al.*, 2016; Katz *et al.*, 2020).

Since the discovery of CTCs, many technologies have been developed for their isolation and detection in the peripheral blood (PB) of patients. Even though this task remains challenging due to the fact that CTCs are present in very low concentrations in the bloodstream (1 ml of PB contains 1-10 CTCs against a background of 10^6 – 10^7 nucleated blood cells and 10^9 red blood cells), currently there exist numerous commercial devices and methods that maximise tumour cell yield, even in the earliest stages of the disease (Calabuig-Fariñas *et al.*, 2016; Katz *et al.*, 2020). Due to the extreme sensitivity required to detect CTCs, most of the existing technologies

consist on a two-step process comprising cell enrichment (to increase CTCs concentration) and a subsequent detection step (Sharma *et al.*, 2018). In this sense, CTC detection methods can be broadly classified as: (i) label-dependent, based on positive enrichment usually involving cell surface markers such as epithelial cell adhesion molecule (EpCAM) or cytokeratin (CK); or (ii) label-independent, based on negative selection usually involving size, density, charge, elasticity or other biophysical properties (Calabuig-Fariñas *et al.*, 2016).

Most label-independent methods comprise assays based on cell size, such as Metacell® filtration device, isolation by size of epithelial tumour cells (ISET®, Rarecells Company, France) or ScreenCell®. Some use combined physical properties of the cells, such as dielectrophoretic field-flow fractionation (DEP-FFF), which employs separation by size and polarizability. As a whole, size-based isolation methods provide high throughput but they have limited applicability due to the heterogeneity in size of CTCs (Sharma *et al.*, 2018).

The wide variety of functional assays available in the market that allow detection of only viable CTCs overcomes the limitations of physical heterogeneity. These include from analysing CTC invasiveness via collagen adhesion matrix protein (Cam assay, Vita-Assay™), to indirectly detecting CTCs via a telomerase-specific adenovirus that replicates in cancer cells and marks them with green fluorescent protein (GFP) in a method known as TelomeScan. However, the high number of false-positive results makes some of these functional assays unfeasible (Sharma *et al.*, 2018).

The most common methods for CTC isolation undoubtedly involve immunobead assays and microdevices, using either positive selection for direct CTC isolation or negative selection to remove blood cells. CellSearch® (Veridex, Raritan, NJ, USA) is currently the only device approved by the Food and Drug Administration (FDA) for CTC detection in breast, prostate and colorectal cancer. It is based on an initial enrichment of EpCAM positive cells using immunobeads, followed by immunofluorescent staining using epithelial markers (CK8, CK18 and CK19), a leukocyte marker (CD45) and 4',6-diamidino-2-phenylindole (DAPI) for nuclear staining. The major drawback of technologies relying on EpCAM positive selection is they are not capable of detecting neither cancer stem cells that have not yet started epithelial differentiation nor CTCs that have undergone EMT (Calabuig-Fariñas *et al.*, 2016; Sharman *et al.*, 2018).

In recent years, the development of microfluidics has motivated the design of microdevices that utilise various antibodies together with separation by size, such as CTC iChip, which integrates size-based enrichment with either EpCAM positive enrichment or CD45 negative depletion (Karabacak *et al.*, 2014).

1.4.4. Multiparameter analyses

A clear long-term goal of LB is to increase resolution, enabling not only identification of minimal residual disease, but also early detection of cancer. For some applications, using only one LB analyte may be sufficient. For early detection of nasopharyngeal carcinoma in asymptomatic individuals, analysis of EBV DNA in plasma is enough due to the fact that these cancer cells contain more than 500 copies of the EBV target sequence, making its detection in peripheral blood quite straightforward. However, achieving early cancer detection with one single analyte will likely remain an exception (Heitzer *et al.*, 2019).

The first attempts at multiparameter analyses focused on ctDNA and protein biomarkers. In this sense, Cohen *et al.* (2017) presented a test capable of identifying the majority of patients with resectable pancreatic cancer (PC) using a combination of protein biomarkers (CA19-9, TIMP1 and LRG1) with ultra-sensitive *KRAS* mutation detection, which showed significantly higher detection rates than ctDNA testing alone. The same group presented another strategy for early cancer detection, named CancerSEEK, which aims for early detection of eight common cancer types (lung, ovarian, liver, stomach, oesophageal, pancreatic, breast and

colorectal) using combined protein and genetic biomarkers. Due to the fact that driver gene mutations are not usually tissue-specific, protein markers were most informative regarding localisation of the tumour. The sensitivities ranged from 69% to 98% in the detection of ovarian, liver, pancreatic, stomach and oesophageal cancer, and the specificity was greater than 99% (Cohen *et al.*, 2018). On the basis of this data, Wong *et al.* (2019) have proposed another approach, Cancer A1DE, which utilises a different modelling paradigm in data analysis to outperform CancerSEEK. One of the most interesting features of CancerA1DE is that it can double CancerSEEK sensitivity in tumours detected at stage I from 38% to 77%, maintaining the 99% specificity level. Despite this improvement, a key concern with this test continues to be its positive predictive value (PPV). In this sense, even though CancerSEEK could achieve a 99% sensitivity and 99% specificity, because the prevalence of the eight cancers in healthy individuals of more than 64 years of age is approximately 1%, the resulting PPV would be only 50%, meaning 50% of test positives would actually be false positives (Heitzer *et al.*, 2019).

Following CancerSEEK, UroSEEK and PapSEEK multi-analyte assays were also developed. In the case of UroSEEK, urothelial cells shed into urine are assayed using a ten-gene multiplex assay, a TERT singleplex assay and an aneuploidy assay with a sensitivity of 75% for urothelial cancer. For its part, PapSEEK is a multiplex-PCR-based test that detects alterations in Pap or Tao brush samples, enabling detection of not only endometrial cancer, but also a substantial fraction of ovarian cancers, for which the sensitivity could be increased including ctDNA assays (Heitzer *et al.*, 2019).

These examples are proof that multi-analyte tests have the power to greatly improve LB analyses, especially when orthogonal analytes are combined to improve signal. This multiparameter approach could open the door to early detection of lethal cancer types (i.e. pancreatic cancer) as described in *Supplementary Figure 1 (Appendix II)*, which would notoriously improve their prognosis (Heitzer *et al.*, 2019).

1.4.5. Future perspectives for liquid biopsy

The large potential of LB in oncology research and clinics is just starting to be explored efficiently. In particular, CTCs and ctDNA have gained remarkable attention as biomarkers, even though there are still technical challenges that remain to be solved. In this sense, cancer detection using CTCs and ctDNA is not equally feasible in all tumour entities, and there exists an important inter-patient variability among patients with the same cancer type. For instance, in patients with NSCLC, the amount of ctDNA and CTCs is much lower to that expected for such an aggressive cancer type (Mader & Pantel, 2017).

However, the potential applications of LB will surely be reinforced in the coming years because of the large number of clinical studies going on nowadays. Furthermore, the improvements made over the past few years in precise and highly sensitive technologies will undoubtedly yield numerous new applications once they are applied to LB (Poulet *et al.*, 2019).

The major problem that LB encounters nowadays is that the majority of assays lack evidence of clinical utility. This implies their use is confined purely to research purposes within clinical studies. To achieve efficient clinical usage, assay developers need to perform an important work of standardization of both pre-analytical and analytical procedures for all LB components (Poulet *et al.*, 2019). Furthermore, most existing assays focus on a single analyte despite resolution and the range of suitable applications could be vastly extended by adopting multiparametric assays. In order to integrate the large amounts of data obtained from multiple analytes, more powerful statistical tools that make use of high-dimensional machine learning approaches must be developed. In any case, it will not be until clinical validity and utility are demonstrated that LB will reach its full potential and have the expected impact on the clinical management of cancer patients (Heitzer *et al.*, 2019).

2. OBJECTIVES

In recent years, the great potential of multi-analyte blood tests in the field of medical oncology has become evident. The main objective of the present study is the development and subsequent bibliographic validation of a multigene panel comprised of genes that are overexpressed in tumour cells of four specific cancer types (lung adenocarcinoma, lung squamous cell carcinoma, head and neck squamous cell carcinoma and oesophageal carcinoma) in order to develop a test that allows early identification of CTCs in blood samples, thereby facilitating an early diagnostic method specific for these cancers.

In order to achieve this purpose, the specific aims of this study are:

1. To identify overexpressed genes specifically in lung adenocarcinoma, lung squamous cell carcinoma, oesophageal carcinoma and head and neck squamous cell carcinoma.

2. To select candidate genes with potential prognostic value attending to the information available in genetic and bibliographic databases.

3. To design primers for the candidate genes, which will be validated first in tumour cell lines and then in CTCs, and will allow for quantification of gene expression in CTCs using quantitative reverse transcriptase polymerase chain reaction (RT-qPCR).

4. To integrate the results and evaluate the clinical utility of the proposed multiparameter blood test in cancer patients.

3. MATERIALS AND METHODS

3.1. Multigene panel design

The proposed multigene panel was divided in 5 different subgroups:

1. Genes that are overexpressed in lung adenocarcinoma (LADC) cells in comparison with normal lung cells, showing a fold change (FC) greater than 2. At the same time, these genes must show low or no expression in both tumoral and normal cells from other tissues different than lung.
2. Genes that are overexpressed in lung squamous cell carcinoma (LSqCC) cells in comparison with normal lung cells, showing a FC > 2. At the same time, these genes must show low or no expression in both tumoral and normal cells from other tissues different than lung.
3. Genes that are overexpressed in HNSqCC cells in comparison with normal head and neck squamous cells, showing a FC > 2. At the same time, these genes must show low or no expression in both tumoral and normal cells from other tissues.
4. Genes that are overexpressed in OC cells in comparison with normal oesophageal cells, generally showing a FC > 2. At the same time, these genes must show low or no expression in both tumoral and normal cells from other tissues.
5. Genes that are overexpressed in LADC, LSqCC, HNSqCC and OC cells in comparison with normal lung cells, head and neck squamous cells and oesophageal cells, respectively. At the same time, these genes must show low or no expression in both tumoral and normal cells from other tissues.

3.2. Analysis of differential gene expression data

The aforementioned multigene panel was mainly developed using the interactive web application GEPIA, which stands for Gene Expression Profiling Interactive Analysis (http://gepia.cancer-pku.cn). GEPIA is a tool that allows for fast and customizable data analysis of 9,736 tumours and 8,587 normal samples obtained from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov) and Genotype-Tissue Expression (GTEx) database (https://www.gtexportal.org/home), using a standard processing pipeline for RNA sequencing data. Among the several functionalities that this time-saving web-based application offers, differential expression analysis, profiling plotting, survival analysis, correlation analysis and similar gene detection stand out (Tang *et al.*, 2017).

In order to find differentially expressed genes (DEGs), the differential expression analysis tool that GEPIA offers was used. Since the purpose was finding overexpressed genes, the $\log_2$ fold change ($\log_2$FC) was set to > 1.5 and an adjusted p-value (q-value) cutoff of 0.05 (q-value < 0.05) was established as threshold. ANOVA was chosen as differential method and the obtained data was downloaded for 4 different datasets: LADC, LSqCC, HNSqCC and OC.

The five subgroups of the multigene panel described above were obtained in a similar way. The only major difference was that for identification of the fifth subgroup, a comparison between datasets needed to be made. For that purpose, LADC, LSqCC, HNSqCC and OC datasets were listed and the genes common to all datasets were identified. Apart from this additional step, the following applies for all 5 subgroups comprising the multigene panel.

Each candidate gene was then studied individually using the information provided by GEPIA. This way, the histogram describing the gene expression profile across all tumour samples and paired normal tissues (*Figure 6*) proved to be of great use to visually and quickly identify whether the candidate gene is specific for the cancer type (or types) of interest or, on the

contrary, it is a marker for other cancers. Because the aim of this project is to develop a sensitive and specific blood-based test, genes showing an expression higher than 15 transcripts per million (tpm) in white blood cells (WBCs) were discarded. In a similar way, experimental data regarding expression in peripheral blood mononuclear cells (PBMCs) obtained in collaboration with Dr. Esplugues' team at CIPF will be considered, if available.

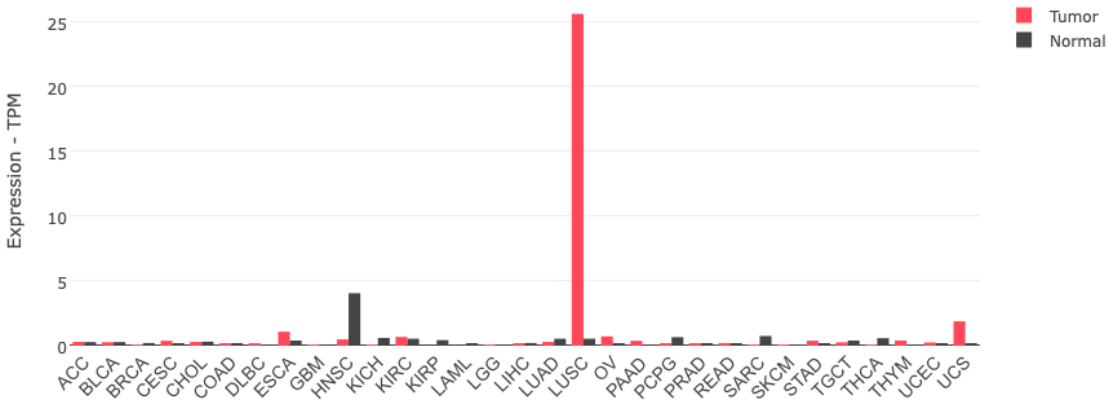


Figure 6. *Neurotensin (NTS) gene expression profile across all tumour samples and paired normal tissues*. This gene expression profile allows for rapid and quickly identification of *NTS* as candidate gene for the LSqCC subgroup of the multigene panel, since it shows high expression in LSqCC cells and low or no expression in all other cancer types and in all normal tissues. Retrieved from GEPIA, 2017.

3.3. Genetic databases analysis

After discarding genes expressed in other cancer types and normal tissues using GEPIA, more information about the candidate genes was obtained using different databases. The National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov) gene database was used in order to obtain general information about the function of the candidate gene. This information was mainly used to discard genes expressed in cells of the immune system, which are therefore not specific of CTCs. Next, the Human Protein Atlas (HPA) database (https://www.proteinatlas.org) was used in order to obtain more information about the subcellular localization of the candidate genes' products. In this sense, proteins located to the cell surface were preferred, since this would allow for future improvements of the multiparameter test by using antibodies against specific antigens in the surface of CTCs. Gene expression data in different tissues and cell lines is also available in the HPA database. Furthermore, the "cancer types summary" and "expression" sections from cBioPortal for Cancer Genomics (https://www.cbioportal.org) were also used to evaluate gene expression across different cancer types. Last of all, Expression Atlas from the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI) (https://www.ebi.ac.uk/gxa) also proved to be useful. More specifically, the Cancer Cell Line Encyclopedia, comprising RNA sequencing information from 934 human cancer cell lines, was of great use because it allowed for selection of all tumour cell lines related with a specific cancer type at once. This way, it was possible to download expression data of the candidate genes in all tumour cell lines related to a specific cancer type at the same time. An average expression value (in tpm) was calculated for each candidate gene and cancer type. Moreover, cell lines exhibiting the most frequent driver mutations for each cancer type were selected in order to assess how expression of candidate genes changes depending on different genetic alterations. For that cell line selection, ATCC (https://www.atcc.org) and DepMap (https://depmap.org) databases were used.

3.4. Bibliographic search strategy

A literature search using the main international electronic databases was conducted between 11th January 2020 and 24th June 2020. The aim was to perform an extensive search, using different synonyms when possible in order to improve the results.

The keywords used for this bibliographic search were: "lung cancer", "non-small cell lung carcinoma", "lung adenocarcinoma", "lung squamous cell carcinoma", "head and neck cancer", "head and neck squamous cell carcinoma", "oesophageal carcinoma", "CTCs", "liquid biopsy", "serum" and the selected genes' official symbols and/or names, which can be found in *Supplementary Table 4 (Appendix I)*.

3 bibliographic databases containing life sciences and biomedical information were used: PubMed (https://pubmed.ncbi.nlm.nih.gov), Scopus (https://www.scopus.com/home.uri) and Google Scholar (https://scholar.google.com). The same search strings were used in all databases, in the same order and for all selected genes. In order to illustrate the approach followed, the gene *NTS* will be used as example. The following applies for all selected genes.

The first search string used was (NTS AND CTCs), and no filter was applied. This search did normally yield no results. The second search string used was (NTS AND liquid AND biopsy), with no filters again, generally yielding less than 5 or no results at all. Then, (NTS AND serum) was used, with no filter. Finally, the last search string used was (NTS AND lung AND squamous AND cell AND carcinoma) and a date filter would usually be set so that only studies published in the last 10 years would be shown. In some cases, when very little information was available, this filter was not used. Although the bibliographic search did normally finish at this point, if the gene at issue was not very well known and little information was available, a last more general search string would be used, (NTS AND cancer). Another strategy followed in these cases was to replace the original gene symbol for a gene alias or for the name of the gene product. In addition, reverse search also proved to be quite useful for the purpose of finding literature in such cases.

Even though more than 10 references were analysed when researching the majority of genes, the general rule was to display information of up to 3 different studies per gene and cancer type, which proved to be enough to summarise the most important findings for the purpose of this study. In cases in which several studies providing similar information were found, the most relevant publications were used.

3.5. Primer design

For primer design, the Primer-BLAST tool that NCBI offers was used (https://www.ncbi.nlm.nih.gov/tools/primer-blast), since it finds specific primers for a given PCR template allowing the user to set certain parameters. This tool is based on Primer3 (https://primer3.org), the most widely used programme for PCR primers design. The settings used were:

1. PCR product size: minimum of 70 and maximum of 150 base pairs (bp). Primers were selected taking into account an optimum amplicon size of 100 bp.
2. Primer melting temperatures (Tm): minimum of 57 °C and maximum of 63 °C. Maximum Tm difference of 3 °C.
3. The option "primer must span an exon-exon junction" was selected. This is to make sure no DNA present in the sample is amplified in the qPCR step, thus obviating the need to use deoxyribonuclease (DNase).
4. "Refseq mRNA" was selected as database, which contains all mRNA data from NCBI's reference sequence collection.
5. Guanine-cytosine content was set to be lower than 65% and low self-complementarity was also taken into account in primer selection.

Once primers were selected, their specificity was checked using the same tool (Primer-BLAST) to avoid products on unintended templates.

3.6. General criteria for gene selection

The criteria used to obtain the genes included in the multigene panel are listed in order of importance below:

1. High expression in cancer cells in comparison with non-cancerous cells ($log_2FC > 1.5$) in GEPIA.
2. High expression in cancer type/s of interest and low or no expression in normal tissues or other cancer types.
3. Very low (< 15 tpm) or no expression in WBCs (genetic databases).
4. If experimental data is available, very low or no expression in PBMCs.
5. Candidate gene not related to the IS.
6. Minimum expression of 50 tpm in GEPIA. However, if the gene appears to be a very promising candidate, expressions ranging 25-50 tpm in tumour tissues may be accepted.
7. Bibliographic evidence that the gene may be a good candidate.
8. Availability of primers providing amplicons of the right size (100 bp) and spanning an exon-exon junction.
9. High expression in specific cancer cell lines.
10. Cell surface localisation of the gene product.

It is important to note that it was almost impossible to find a candidate gene that meets all the cited criteria. Therefore, the aim of this section is to establish a general guideline that has indeed proved to be very useful in the selection process. In any case, it should be clear that each candidate gene was analysed individually and incorporated to the multigene panel after integrating information from different databases and taking into account several different parameters. To set an example, some genes might not fulfil the first or second most important criterion but may have been selected either way because the information found in the literature strongly suggests they are very promising candidates.

Interestingly, this study has been designed to also allow expression analysis of specific genes for detection of other cancer types than the assessed here (data not shown).

4. RESULTS AND DISCUSSION

4.1. LADC, LSqCC, HNSqCC and OC common gene selection

There were 45 genes in common in the 4 datasets obtained from GEPIA containing overexpressed genes ($log_2FC > 1.5$) in LADC, LSqCC, HNSqCC and OC. After analysing each gene individually, 33 out of the 45 genes were discarded following the criteria established in section 3.6. Not showing specificity for the cancer types of interest was one of the most frequent reasons why candidate genes were discarded, along with the fact that some showed high expression in WBCs (> 15 tpm) or were expressed in cells of the IS, as 6 of the 45 candidate genes coded for immunoglobulins. The list of discarded candidate genes along with specific details as to why each of them was discarded can be found in *Supplementary Table 3 (Appendix I)*.

The final 12 genes selected for this subgroup of the multigene panel are, in alphabetical order: *COL10A1, CST1, CTHRC1, CXCL9, CXCL13, EPCAM, KRT17, LAMB3, MMP1, MMP11, MMP12* and *UBE2C*. *Supplementary Table 4 (Appendix I)* displays genes' official full names, along with forward and reverse primers for their amplification, designed as described in section 3.5. A comparison between expression of the listed genes in each of the assessed tumour tissues (LADC, LSqCC, HNSqCC, OC) and their corresponding normal tissues is shown below, in *Figure 7*.

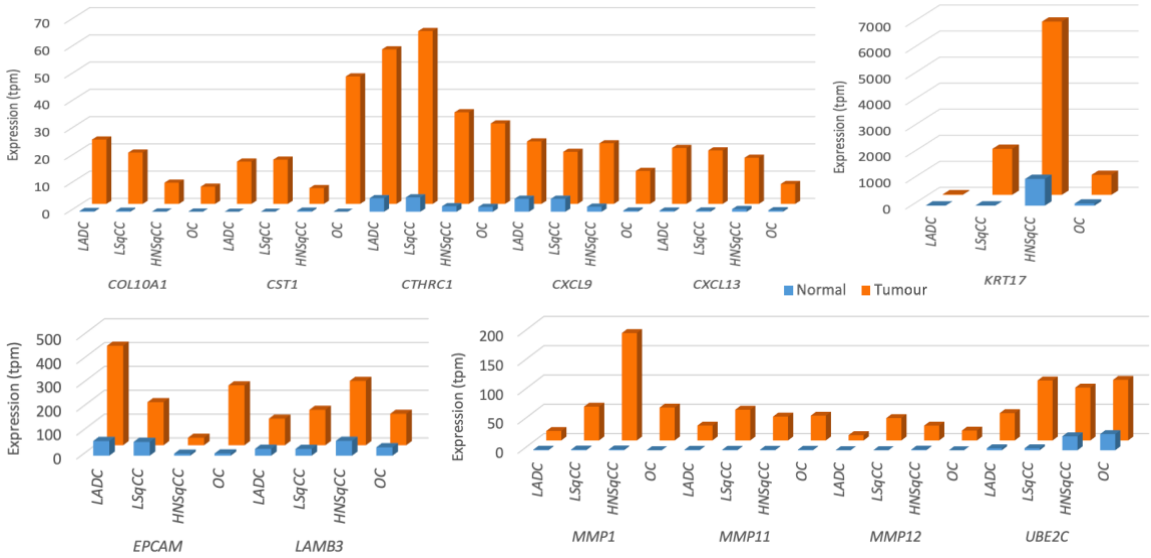


Figure 7. *Expression profile of the selected genes in LADC, LSqCC, HNSqCC and OC tumour samples (orange bars) and paired normal tissues (blue bars).* Data retrieved from GEPIA (http://gepia.cancer-pku.cn/index.html). LADC: Lung Adenocarcinoma; LSqCC: Lung Squamous Cell Carcinoma; HNSqCC: Head and Neck Squamous Cell Carcinoma; OC: Oesophageal Carcinoma.

The suitability of the selected genes was corroborated by a bibliographic search, the results of which have been summarised in *Table 4*. In general, all genes assessed in this section fulfil at least seven out of the ten established general criteria for gene selection. Undoubtedly, the requirement that proved to be most difficult to meet in this subgroup of the multigene panel was specificity. In this sense, it was quite difficult to find a gene involved in tumorigenesis that is overexpressed only in the four cancer types of interest. However, this fact may not pose a problem at all, since the purpose of this study is to come up with a multigene panel that can be used in a multiparameter blood test. Therefore, it is expected that the combination of parameters solves the potential problems derived from the lack of specificity that certain candidate genes present with. As can be observed in *Table 4*, the majority of studies found are related with expression of the gene in tumour samples and paired normal tissues. Moreover, some of them try to establish a relationship between gene expression and prognosis.

***Table 4**. Overview of results obtained from the literature search of the 4 studied cancer types' common gene selection.* The nature of the bibliographic source is indicated with an asterisk (*) in the case of studies based on bioinformatics analysis of genetic databases. Experimental studies based on analysis of protein or mRNA levels in cell lines, tissues or blood plasma are not marked. DEG: Differentially Expressed Gene; FC: Fold Change; LADC: Lung Adenocarcinoma; LC: Lung Cancer; LSqCC: Lung Squamous Cell Carcinoma; HNC: Head and Neck Cancer; HNSqCC: Head and Neck Squamous Cell Carcinoma; nd: no data; NSCLC: Non-Small Cell Lung Carcinoma; OADC: Oesophageal Adenocarcinoma; OC: Oesophageal Carcinoma; OSqCC: Oesophageal Squamous Cell Carcinoma; qPCR: quantitative Polymerase Chain Reaction; REF: References; RT-PCR: Reverse Transcription Polymerase Chain Reaction; SqCC: Squamous Cell Carcinoma.

Candidate gene	NSCLC		HNSqCC		OC	
	Gene information	**REF**	**Gene information**	**REF**	**Gene information**	**REF**
COL10A1	- Higher plasma levels in NSCLC patients compared to healthy smokers. - Valuable diagnostic tool to distinguish LC patients from smokers.	Andriani *et al.*, 2018	- Identified as DEG and hub gene involved in HNC development. - Potential biomarker for HNC.	Chen *et al.*, 2019*	- Overexpressed in OSqCC. - Might be a potential diagnostic and prognostic biomarker.	Li *et al.*, 2019*
	Among most significant DEGs in LADC tumour tissue compared to normal tissue.	Wu *et al.*, 2015*	Overexpressed in laryngeal SqCC.	Lapa *et al.*, 2019		Li *et al.*, 2019*
CST1	- Overexpressed in LADC. - Important role in LADC development.	Chen *et al.*, 2017*	Upregulation in HNSqCC in tobacco smokers.	Shaikh *et al.*, 2019	Considered as tumour marker for gastrointestinal tract cancer.	Chen *et al.*, 2013
	Hypomethylation of its promoter region leads to high expression in NSCLC patients.	Liu & Yao, 2019*				
CTHRC1	- Proteomic analysis shows overexpression in NSCLC. - Pro-metastatic gene contributing to invasion through *MMP7* and *MMP9* upregulation.	He *et al.*, 2018	- Significantly overexpressed at the mRNA level in oral SqCC. - Related to poor prognosis.	Lee *et al.*, 2015	- Identified as one of the most upregulated genes in OSqCC. - Oncogenic driver in progression of OSqCC. - May serve as potential biomarker.	Wang *et al.*, 2017*
					- Significantly overexpressed in OC early tumour stages (transcriptome analysis).	Warnecke-Eberz *et al.*, 2016
CXCL9	*CXCL9* is upregulated in early-stage NSCLC.	Metodieva *et al.*, 2011	- Expression was significantly higher in oral SqCC compared to normal epithelium. - Serum CXCL9 levels were also significantly higher in oral SqCC patients.	Chang *et al.*, 2013	Satisfactory OADC prognostic factor.	Babar *et al.*, 2019
	- High levels of CXCL9 in serum of NSCLC patients. - CXCL9 is one of the most sensitive and specific biomarkers in early-stage NSCLC.	Spaks *et al.*, 2015				

Candidate gene	NSCLC		HNSqCC		OC	
	Gene information	**REF**	**Gene information**	**REF**	**Gene information**	**REF**
CXCL13	High serum CXCL13 levels in NSCLC patients.	Singh *et al.*, 2014	Microarray analysis shows high levels of gene expression.	Sambandam *et al.*, 2013	*nd*	*nd*
	Overexpressed in 90% of NSCLC cases.	Wang *et al.*, 2015				
EPCAM	High expression in NSCLC cells (flow cytometry, RT-PCR).	Kim *et al.*, 2009	- EpCAM expression absent in healthy oral mucosa. - High expression in 85% oral SqCC cases evaluated.	Sen & Carnelio, 2016	Expression significantly higher in OC in comparison with normal tissue (RT-qPCR, ELISA, immunohistochemistry).	Kimura *et al.*, 2007
KRT17	Both RNA and protein levels were upregulated in LADC tissues compared to normal lung tissue (qPCR, immunohistochemical staining).	Liu *et al.*, 2018	Invariably and permanently induced in oral SqCC (immunohistochemistry and cDNA microarray analysis using two oral SqCC cell lines).	Khanom *et al.*, 2016	Validation of *KRT17* overexpression in OSqCC using new technologies: expression microdissection (protein analysis) and RNAscope (mRNA expression).	Du *et al.*, 2013
	Significantly higher expression in NSCLC tissue in comparison with normal lung tissues, assessed by immunohistochemistry.	Wang *et al.*, 2019			- Upregulated in OSqCC (RNA sequencing). - May serve as prognostic biomarker in OSqCC.	Liu *et al.*, 2020
LAMB3	- Gene array and bioinformatics analyses demonstrate *LAMB3* implication in LC. - Gene knockdown implies suppressed cell invasion and metastasis.	Wang *et al.*, 2013	Immunohistochemistry analyses show upregulation in both HNSqCC cell lines and patient tissues.	Liu *et al.*, 2019	- Expression levels found to be higher in malignant OSqCC tissues than in corresponding normal tissues. - May predict prognosis.	Kita *et al.*, 2009
MMP1	Protein levels high in plasma from LC patients in comparison with healthy controls (ELISA).	Li *et al.*, 2010	DEG and hub gene in HNC development.	Chen *et al.* 2019	Included in gene cluster used as "diagnostic signature" for detection of early OC due to its high expression levels.	Warnecke-Eberz *et al.*, 2016
MMP11	DEG studies show FC > 4 when comparing expression in LADC and LSqCC tissue with corresponding normal tissues.	Gobin *et al.*, 2019*	DEG studies show FC > 4 when comparing expression in HNSqCC tissue with normal tissue.	Gobin *et al.*, 2019*	DEG studies show FC > 4 when comparing expression in OC tissue with normal tissue.	Gobin *et al.*, 2019*
MMP12					Included in gene cluster used as "diagnostic signature" for detection of early OC due to its high expression levels.	Warnecke-Eberz *et al.*, 2016

Candidate gene	NSCLC		HNSqCC		OC	
	Gene information	**REF**	**Gene information**	**REF**	**Gene information**	**REF**
UBE2C	- mRNA and protein levels were significantly upregulated in NSCLC tissues in comparison with normal lung tissue. - May also be an indicator of poor survival.	Kadara *et al.*, 2009	- Overexpression in HNSqCC cells versus normal oral keratinocytes. - Involved in tumorigenesis through important pathways: proteasome, cell cycle, ubiquitin proteolysis.	Jin *et al.*, 2020*	- RT-qPCR reveals overexpression in 73% OSqCC tumour samples in comparison with normal tissues. - Immunohistochemistry shows overexpression of the gene in all OSqCC samples assessed.	Palumbo *et al.*, 2016

If the candidate genes are divided depending on the function of their products, *COL10A1*, *CTHRC1* and *LAMB3* fall into the category of genes coding for proteins related to the extracellular matrix (ECM). The function of the ECM is to provide a mechanical and biochemical support to the surrounding cells (Andriani *et al.*, 2018). Since it is actively involved in cell proliferation and migration, many ECM related genes are overexpressed in several cancer types from very early stages, therefore being the perfect candidates for the purpose at issue.

Whereas *COL10A1* encodes the alpha chain of type X collagen, expressed mainly by hypertrophic chondrocytes during endochondral ossification, *CTHRC1* encodes collagen triple helix repeat containing 1. The overexpression in fibroblasts of these types of collagen is associated with increased cell migration, motility and invasion (Warnecke-Eberz *et al.*, 2016). It must be noted that *CTHRC1* has been strongly associated with progression of OC by MAPK/MEK/ERK/FRA-1 pathway activation. Regarding *LAMB3*, it encodes the β3 subunit of a protein of the basement membrane, laminin 5. This protein is involved in epithelial cell migration, regeneration and repair processes, and although it was initially related with gastric carcinogenesis, its involvement in LADC progression has also been reported (Liu *et al.*, 2019).

As for the limitations of these candidate genes, *COL10A1*, though more specific, shows expression lower than 25 tpm in HNSqCC and OSqCC. Furthermore, expression data in tumour cell lines (*Supplementary Table 5, Appendix I*) is very low in all cancer types, with values close to 0 tpm, so its detection may be complicated. *LAMB3*, however, shows high expression (> 100 tpm) in both cancerous tissue samples and tumour cell lines. *CTHRC1* exhibits moderate expression (> 50 tpm) in NSCLC but low in HNSqCC and OC (≈ 30 tpm) tissue samples and very low in HNSqCC and OC cell lines (≈ 10 tpm). In any case, because of the continuous remodelling taking place in the ECM of the tumour microenvironment, these genes' products are likely to be released into blood, constituting potential protein circulating biomarkers (Andriani *et al.*, 2018).

Also related to the ECM, *KRT17* codes for keratin 17, a type I intermediate filament mainly present in epithelial basal cells (Wang *et al.*, 2019). However, under normal circumstances *KRT17* is not expressed in the epidermis of normal skin. Stress conditions such as skin scratching are required for the gene to be expressed. In cancer, *KRT17* plays an important role in the occurrence and development of different tumours (Liu *et al.*, 2020). Overexpression of *KRT17* upregulates β-catenin activity and levels of Wnt target genes (cyclin D1, c-Myc, etc.), enhancing proliferation and invasiveness of lung cancer cells. Besides, it promotes ECM development by upregulation of MMP9, Vimentin and Snail expression and downregulation of E-cadherin (Wang *et al.*, 2019). In the case of oral SqCC, *KRT17* has been shown to promote cell proliferation and migration by stimulating the Akt/mTOR pathway. This way, Khanom *et al.* (2016) highlight that this type of keratin would act as a pathogenic protein that facilitates tumour growth via stimulation of different signalling pathways. Regarding expression data, *KRT17* shows very high expression in HNSqCC (> 2500 tpm), OC (> 750 tpm) and LSqCC (> 350 tpm) tissue samples and cancer cell lines. In the case of LADC, expression was found to be quite low in tissue samples (≈ 20 tpm) but cell lines show better values (≈ 170 tpm). Taking into

account all this information and the fact that it is a very specific gene for the cancer types at issue, *KRT17* constitutes one of the most promising candidate genes of this study.

Another group that can be distinguished in this gene selection is the one comprised by enzymes. In this sense, *CST1* codes for cystatin-SN, which belongs to the type 2 cystatin superfamily. Type 2 cystatin proteins are cysteine proteinase inhibitors found mainly in human fluids and secretions, where they appear to have protective functions. Cystatin-SN is involved in inflammation, cell cycle, cellular senescence, tumorigenesis and metastasis. Its involvement in several signalling pathways has been reported, such as Wnt, GSK3, Akt or IL-6 (Liu & Yao, 2019). Although the expression data in tissue samples and tumour cell lines is not very promising, with values close to 0 tpm in HNSqCC, the literature found (*Table 4*) and its apparent specificity, together with the fact that experimental expression data in healthy individuals' PBMCs shows a value very close to 0 (0.03 tpm) have encouraged its incorporation into this subgroup of the multigene panel.

MMP1, MMP11 and *MMP12* are also included in the category of genes coding for enzymes. More specifically, they encode different types of matrix metalloproteinases (MMPs), which have been studied for more than 40 years and have always been related with the degradation of the ECM. Moreover, MMPs have been found to play several roles at the cellular level in pathways such as immunity, apoptosis, angiogenesis and cellular migration (Chen *et al.*, 2019; Gobin *et al.*, 2019). In cancer, it has been demonstrated that this family of proteolytic enzymes promotes invasion and metastasis, mainly due to their ability to degrade components of the ECM. Traditional classification of MMPs is based on the first identified target of degradation (Gobin *et al.*, 2019). In this sense, *MMP1* codes for a collagenase involved in initial invasion and metastasis. It shows a high expression both in tumour tissue and cell lines of the cancer types of interest. Moreover, protein levels have been found to be high in LC patients (Li *et al.*, 2010). This way, and despite its overexpression in other cancer types (colorectal and pancreatic, mainly), it could be a very useful asset in combination with other markers.

In the case of *MMP11*, it encodes a stromelysin produced by peritumoral stromal fibroblasts that has been related with the regulation of early tumour invasion, implantation and expansion. It may also be implicated in evasion of apoptosis of early cancer cells. Of the 3 MMPs assessed in this section, it is by far the least specific one, as according to Gobin *et al.* (2019), it happens to be ubiquitously upregulated across most cancer types. However, expression data of the assessed tumour cell lines indicates otherwise (with values below 3 tpm). In any case, it was selected mainly because of the great FC obtained when comparing expression in tumour versus normal tissues. This is a consequence of the very low expression of the gene in almost all healthy tissue types, with values very close to 0 tpm. Moreover, experimental studies with healthy individuals' PBMCs have determined expression of *MMP11* in these cells to be of 0 tpm, therefore reducing the possibility of false positives in a multiparameter blood test. In any case, if its low specificity does not allow to detect the cancer types of interest, there is evidence it could be useful either way as a general cancer marker.

Regarding *MMP12*, this gene codes for a metalloelastase that does not show very high expression neither in tumour tissue (< 40 tpm for all cancer types assessed) nor in tumour cell lines (< 3 tpm). However, it appears to be the most specific selected MMP for the cancer types at issue (Gobin *et al.*, 2019). Moreover, *MMP12* shows very low expression in normal tissues and the information found in the literature suggests it could be a good candidate. For all these reasons and despite its apparent low expression, it has been incorporated in the hopes that it provides the specificity that *MMP1* and *MMP11* lack.

The last member of this group is *UBE2C*, which codes for ubiquitin conjugating enzyme E2. This protein is involved in several molecular functions, such as the degradation of mitotic cyclins involved in cell cycle regulation, or the suppression of premature DNA replication through degradation of cyclin A by the anaphase promoting complex/cyclosome, APC/C (Jin *et al.*, 2020; Palumbo *et al.*, 2016). Its expression in tumour cell lines and cancer types of interest

is moderate (> 100 tpm, except in LADC tissue). Moreover, in healthy individuals' PBMCs its expression is quite low (0.66 tpm), so there should not be any problems in this respect. The major potential limitation in this case is the fact that *UBE2C* also appears to be overexpressed in ovary, breast, thyroid and uterine carcinomas (Palumbo *et al.*, 2016).

2 genes coding for CXC chemokine ligands, *CXCL9* and *CXCL13* have also been included in this subgroup of the multigene panel. Chemokines are small proteins that regulate the migration of cells towards a chemokine gradient detected by G-protein-coupled chemokine receptors. They are essential for homeostasis of the immune and stem cell systems. In cancer, they are involved in neoplastic transformation of cells, tumour cell growth and organ-specific metastasis (Singh *et al.*, 2015). Both *CXCL9* and *CXCL13* show similar levels of expression in the studied tumour tissues (generally around 20 tpm) and cell lines (around 0.2 tpm). However, *CXCL9* appears to be a bit more specific and serum proteins have been found in NSCLC and oral SqCC patients, as described in *Table 4.* In the case of CXC chemokine ligand-13, usually induced under inflammatory conditions (Sambandam *et al.*, 2013), it is present at high levels in serum of NSCLC patients (Singh *et al.*, 2014). However, there is no literature about its role in OC, and according to Vachani *et al.* (2007) it could be used to differentiate LSqCC from HNSqCC, so its suitability for this common gene selection would have to be demonstrated experimentally.

Last of all, *EPCAM*, coding for a type I transmembrane glycoprotein involved in intercellular adhesion (Kim *et al.*, 2009), was the last gene to be incorporated. Its expression appears to be quite high in tumour cell lines (> 200 tpm) and in the cancer types' tissues at issue (> 180 tpm except in HNSqCC). Moreover, expression in healthy individuals' PBMCs was close to 0 (0.02 tpm). Undoubtedly, *EPCAM* is in most cases expressed in CTCs, since it is used by the only clinical system approved by the FDA to detect CTCs, CellSearch® (Veridex, Raritan, NJ, USA). However, this lack of specificity is its major drawback for the purpose at issue, along with the fact that some CTCs have undergone EMT and no longer express *EPCAM* (Sharman *et al.*, 2018).

4.2. Lung adenocarcinoma (LADC) gene selection

After studying each of the genes found in the LADC dataset ($log_2FC > 1.5$) individually in GEPIA, 15 genes were initially chosen. Of these, 10 were discarded as even though they were quite specific, the majority of genes initially selected presented with very low expressions (< 10 tpm in most cases) in LADC tissue. Moreover, there was no literature ratifying their suitability for this subgroup of the multigene panel. This way, the final five selected genes are *AGR2*, *CEACAM6*, *SMIM22*, *UBD* and *WFDC2*. Genes' official full names, along with forward and reverse primers designed for their amplification can be found in *Supplementary Table 4 (Appendix I).* A comparison between their expression profile in LADC tissue and normal lung tissue is shown in *Figure 8*.

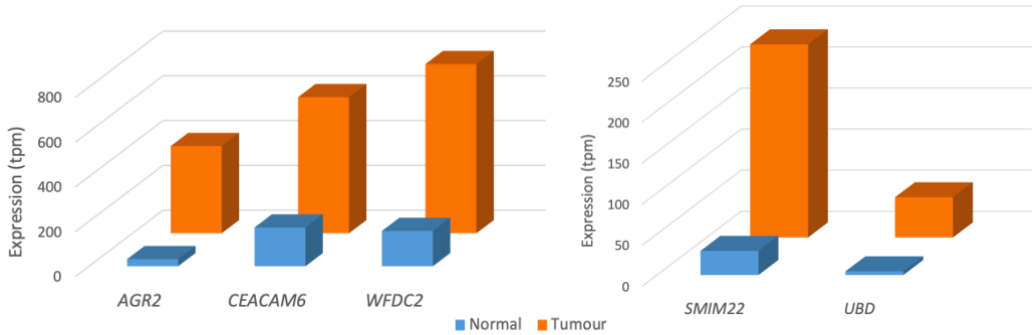


Figure 8. *Expression profile of the selected genes in LADC tumour samples (orange bars) and paired normal tissues (blue bars)*. Data retrieved from GEPIA (http://gepia.cancer-pku.cn/index.html).

From the very first analysis in GEPIA, it became evident that it was going to be difficult to find cancer-related genes that are overexpressed only in a very specific cancer type. Because of this, the strategy followed was to select a group of genes whose combined expression may allow to differentiate one cancer from another. To set an example, *CEACAM6* shows high expression in LADC, colorectal cancer (CRC) and pancreatic cancer (PC), but no expression in ovarian cancer (OVC) or endometrial cancer (EC). *WFDC2*, however, shows high expression in LADC, OVC and EC, but no expression in CRC or PC. This way, cases of CRC or PC could be ruled out because of high *WFDC2* expression. Similarly, cases of OVC and EC could be discarded because of high *CEACAM6* expression. This strategy was also used in the following sections. *Table 5* summarises the results of the literature search concerning these LADC genes and offers information about their expression in healthy individuals' PBMCs and other cancer types where they appear to be overexpressed.

***Table 5**. Overview of the literature search results regarding LADC gene selection along with information about expression in other cancer types and healthy individuals' PBMCs.* All studies refer to experimental analyses of protein or mRNA levels in cell lines, tissues or blood plasma. LADC: Lung Adenocarcinoma; LSqCC: Lung Squamous Cell Carcinoma; nd: no data; NSCLC: Non-Small Cell Lung Carcinoma; qPCR: quantitative Polymerase Chain Reaction; REF: References; tpm: transcripts per million.

Candidate gene	Gene information	REF	Other cancer types	PBMCs' expression
AGR2	Serum AGR2 protein levels are high in patients with stage I LADC, so it is a very promising early prognostic biomarker.	Chung *et al.*, 2011	Breast, colorectal, pancreatic, prostate, stomach.	0 tpm
	Found to be overexpressed in 94% of LADC patients (assessed by liquid chromatography tandem mass spectrometry and immunohistochemistry).	Chung *et al.*, 2012		
	May be of clinical value in differentiating LADC from LSqCC, since it happens to be strongly expressed in LADC and shows weak or no expression in LSqCC cases.	Pizzi *et al.*, 2012		
CEACAM6	Can be used to distinguish LADC from LSqCC patients.	Relli *et al.*, 2018	Colorectal, pancreatic, stomach.	0 tpm
	Serum levels frequently upregulated in LADC patients (assessed by enzyme-linked immunosorbent assay)	Singer *et al.*, 2010		
SMIM22	*nd*	*nd*	Bladder, breast, cervical, colorectal, ovarian, pancreatic, stomach, uterine.	0 tpm
UBD/FAT10	Elevated levels in quick chemoresistant NSCLC tissues (qPCR). Promotes LADC cell proliferation, migration and invasion. Its knockdown reduces drug resistance in NSCLC cells.	Xue *et al.*, 2016	Cholangiocarcinoma, diffuse large B-cell lymphoma, liver, pancreatic, stomach.	0.03 tpm
WFDC2/HE4	Serum HE4 levels higher in NSCLC in comparison with tuberculosis patients and healthy individuals (detected by electrochemiluminescence method).	Wang *et al.*, 2019	Kidney, ovarian, endometrial.	*nd*
	Serum HE4 is one of the biomarkers with the highest sensitivity (43.8%) and specificity (95%) for early diagnose of NSCLC when compared to other frequently used biomarkers (detected with Roche Elecsys assays).	Zeng *et al.*, 2016		

Apart from the literature search, expression data in different LADC cell lines was also considered to assess the suitability of the selected genes. Cell lines with different driver mutations were selected (*Supplementary Table 6, Appendix I*) to discuss variations in candidate genes expression depending on certain genetic alterations: A549 (mutated *KRAS* and *CDKN2A*), NCI-H1395 (mutations in *BRAF*), NCI-H1975 (mutated *EGFR*, *CDKN2A*, *PIK3CA* and *TP53*) and NCI-H2228 (*EML4-ALK* fusion). Information about selected genes' expression in these LADC cell lines can be found in *Supplementary Table 7 (Appendix I)*.

The most promising candidate genes are *AGR2* and *CEACAM6*. Anterior gradient homolog 2 (AGR2) is a human ortholog of a frog protein involved in embryo development. In humans, it is a chaperon involved in protein maturation in the endoplasmic reticulum as a member of the protein disulphide isomerase family. In cancer, it promotes cell growth, migration and transformation (Chung *et al.*, 2012; Pizzi *et al.*, 2012). *AGR2* is a very suitable candidate since it would presumably allow for discrimination between LADC and LSqCC (Pizzi *et al.*, 2012) and it has been detected in serum of early-stage LADC patients (Chung *et al.*, 2011). Its expression is high in A549 (156 tpm), NCI-H2228 (167 tpm) and NCI-H1395 (2670 tpm), but low in NCI-H1975 (2 tpm), which presents with mutations in *TP53*. Since AGR2 itself is a p53 inhibitor (Chung *et al.*, 2012), it is possible that cells with mutated *TP53* do not have the need to express *AGR2* to inhibit p53, which would explain why NCI-H1975 cells present with low *AGR2* levels.

Regarding *CEACAM6*, this gene codes for carcinoembryonic antigen-related cell adhesion molecule 6, which mediates homotypic and heterotypic interactions between cells through integrin receptors. It is involved in cell adhesion, proliferation, migration and invasion. Similar to *AGR2*, it exhibits high gene expression and protein serum levels in LADC cases (Singer *et al.*, 2010), but low in LSqCC, so it could be used as a distinctive marker between these two cancer types. Interestingly, even though expression is relatively high in NCI-H1395 (393 tpm) and moderate in NCI-H2228 (97 tpm), A549 and NCI-H1975 show low expression values (< 12 tpm).

Even though no literature was found linking small integral membrane protein 22 (encoded by *SMIM22*) with LADC diagnosis, this microprotein was selected because of its possible key role in cancer progression, since its knockdown is related with decreased proliferation in many breast cancer cell lines. This could be explained because this small open reading frame-encoded protein seems to be related with cell cycle control, cell mobility and organization of the actin cytoskeleton. Even though its expression in most LADC cell lines is low (< 25 tpm), the latest publications reporting the use of these microproteins for prostate and breast cancer diagnosis (Polycarpou-Schwarz *et al.*, 2018) have encouraged its incorporation.

Contrary to *SMIM22, UBD* and *WFDC2* appear to be more solid candidates. *UBD*, or *FAT10*, encodes ubiquitin D, whose physiological function has been largely unknown. However, their interaction partners have recently been identified, and include Mad2, p53, p62 and huntingtin. Some of these interactions have confirmed its involvement in cancer progression. In fact, there is evidence of its implication in nuclear kappa B signalling pathway in LADC cases (Xue *et al.*, 2016). Despite its low expression in LADC cell lines, it could be an important asset due to its relatively high specificity.

Last of all, whey-acidic-protein 4-disulfide core domain 2 (*WFDC2*) or human epididymis 4 (*HE4*) encodes a novel biomarker highly expressed in ovarian cancer but showing very low expression in normal tissues. As it was previously explained, it has been incorporated in the hopes that it can complement other biomarkers, since there is strong evidence that it could be used for early diagnosis of LC (Wang *et al.*, 2019). Even though it presents with high expression in tissue samples and cell lines exhibiting *EML4-ALK* fusion, its suitability will have to be checked experimentally due to the low expression values found in other LADC cell lines (*Supplementary Table 7, Appendix I*).

4.3. Lung squamous cell carcinoma (LSqCC) gene selection

Out of the 17 genes initially chosen for this subgroup of the multigene panel, 13 were discarded mainly due to very low expression (< 15 tpm) or lack of specificity. The 4 final selected genes were *HAS3*, *NTS*, *SOX2* and *TP63*, whose official full names together with primers for their amplification are displayed in *Supplementary Table 4 (Appendix I).* The expression profile of the genes in LSqCC tissue in comparison with normal lung tissue is shown in *Figure 9*.

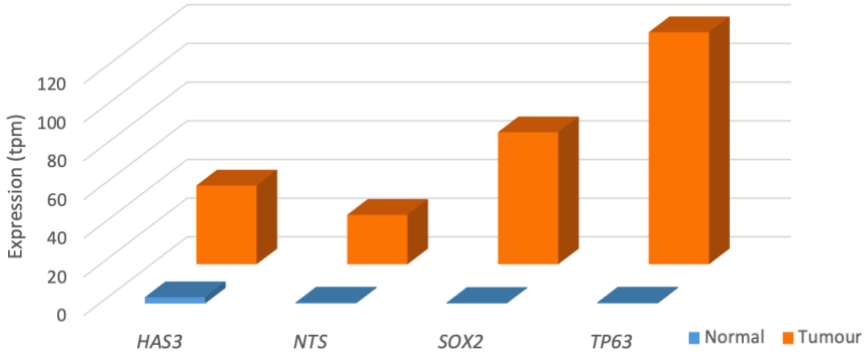


Figure 9. *Expression profile of the selected genes in LSqCC tumour samples (orange bars) and paired normal tissues (blue bars)*. Data retrieved from GEPIA (http://gepia.cancer-pku.cn/index.html).

The major problem found in this selection was that most genes showing high expression in LSqCC were also overexpressed in LADC and HNSqCC. Similar to section 4.2, the strategy was to use a combination of genes that could rule out other cancer types different than LSqCC. *Table 6* shows the findings of the bibliographic search and provides information about other cancer types where selected genes are overexpressed.

In this section, the selected LSqCC cell lines (*Supplementary Table 6, Appendix I*) were NCI-H2170, showing mutations in *TP53* and *CDKN2A*; SW 900, with mutated *KRAS*, *CDKN2A* and *TP53*; and HCC-95, showing high *PIK3CA* copy number. Information about candidate genes' expression in these cell lines can be found in *Supplementary Table 8 (Appendix I).*

Table 6*. Overview of the literature search results regarding LSqCC gene selection and expression in other cancer types.* LSqCC: Lung Squamous Cell Carcinoma; nd: no data; NSCLC: Non-Small Cell Lung Carcinoma; qPCR: quantitative Polymerase Chain Reaction; REF: References; SqCC: Squamous Cell Carcinoma.

Gene	Gene information	REF	Other cancer types
HAS3	*nd*	*nd*	Bladder, breast, cervical, oesophagus, head and neck.
NTS	Expression in more than half of the assessed NSCLC stage I tissues and no expression in corresponding normal tissues.	Dupouy *et al.*, 2011	-
SOX2	Frequency of amplification in LSqCC: 20 – 60% of cases. Might be a general marker for SqCC differentiation.	Karachaliou *et al.*, 2013	Brain, breast, lung adenocarcinoma, lung small cell lung carcinoma, prostate.
	Quantification of serum *SOX2* DNA by qPCR may be a novel accessory diagnostic tool for lung cancer detection.	Wu *et al.*, 2013	
TP63	Amplification in 88% of early LSqCCs. High expression in most LSqCC patients but low in other lung cancer types' patients.	Massion *et al.*, 2004	Bladder, cervical, head and neck, oesophageal.
	Shows high expression in LSqCC but low in lung adenocarcinoma, which could provide a novel differential diagnosis strategy.	Peng *et al.*, 2019	

The first selected gene was *HAS3*, which codes for one of the three hyaluronan synthase enzymes involved in the production of hyaluronic acid. Even though it was difficult to find bibliographic evidence linking *HAS3* with LSqCC diagnosis, it has been included because increased synthesis of hyaluronan has been closely related with tumorigenesis in lung and breast cancer, among others (Chow *et al.*, 2010). Furthermore, on the basis of expression data, the gene appears to be quite specific and more expressed in SqCCs, which would help distinguish LSqCC from LADC. The fact that its use for LC diagnosis has not yet been tested has also encouraged its incorporation. However, LSqCC cell lines' data (*Supplementary Table 8, Appendix I*) shows its expression might be limited to cells showing *PIK3CA* amplifications, so its suitability will have to be checked experimentally.

Regarding *NTS*, this gene codes for neurotensin, a 13 amino acid peptide normally released by N cells of the gastrointestinal tract that predominantly exerts hormonal and neurocrine regulation on the digestive process. Neurotensin's action is mediated by two G protein coupled receptors, NTSR1 and NTSR2. NTSR1 appears to be abnormally expressed during early stages of cell transformation due to Wnt/β-catenin pathway deregulation. Several oncogenic effects have been reported concerning this NTS/NTSR1 complex, mostly related to tumour growth (Dupouy *et al.*, 2011). The gene has been mainly selected because, in addition to this information, its expression appears to be very specific to LSqCC, as shown in *Figure 6*. The only limitation of *NTS* is its relatively low expression in tissue samples (26 tpm) and cell lines (15 tpm on average), which could be a major problem when it comes to its detection in CTCs.

Last of all, copy number alterations occur in several cancer types, including LC. Although many of these are common to both LADC and LSqCC, gains in 3q26 and 8p12 chromosome areas seem to be more common in squamous histology. Genes in 3q26 include *PIK3CA*, *SOX2, TP63* and *TERC* (Karachaliou *et al.*, 2013). *SOX2* and *TP63* were selected due to their relative specificity and the information found in the literature, as they appear to be very promising, particularly when combined with other markers (*Table 6*). *SOX2* belongs to the SOX family of transcription factors. It is involved in the regulation of embryonic development and determination of cell fate, as it downregulates genes responsible for differentiation. In cancer, it has also been linked to Wnt/β-catenin pathway deregulation. Even though *SOX2* overexpression has been related with all types of LC (Karachaliou *et al.*, 2013), the analysed expression data exhibits a much more prominent FC in LSqCC tissues. Furthermore, expression data in LC cell lines shows the same trend. In this respect, a remarkable rise in expression is seen in HCC-95, consistent with the fact that this cell line shows *PIK3CA* amplification. The same applies to *TP63* expression, which shows a considerable increase in HCC-95 compared to average expression in all assessed cell lines (*Supplementary Table 8, Appendix I*). However, the suitability of this gene, coding for a member of the p53 family of transcription factors, will have to be checked experimentally as cell lines presenting with other driver mutations, such as SW 900, show a very low expression (3 tpm).

4.4. Head and neck squamous cell carcinoma (HNSqCC) gene selection

In the case of HNSqCC, a higher number of genes (31) were initially selected, although only 7 would finally be incorporated to the multigene panel: *CDH3*, *KRT16*, *LAMC2*, *MMP10*, *PI3*, *PTHLH* and *SLC2A1* (official full names and primers for their amplification displayed in *Supplementary Table 4, Appendix I*). In this case, it was easier to find specific genes for HNSqCC in the initial selection, but some of them were discarded because of their low expression (< 10 tpm) or because they were expressed in normal tissues as well. The expression profile of the selected genes in HNSqCC tissue in comparison with normal tissue is shown in *Figure 10*.

Similar to the previous section, the major problem found in this gene selection was that most genes also showed high expression in LSqCC, which is consistent with the fact that both LSqCC and HNSqCC are tumours of squamous histology, and therefore share some metabolic alterations. The combination of genes used aimed at solving this concern.

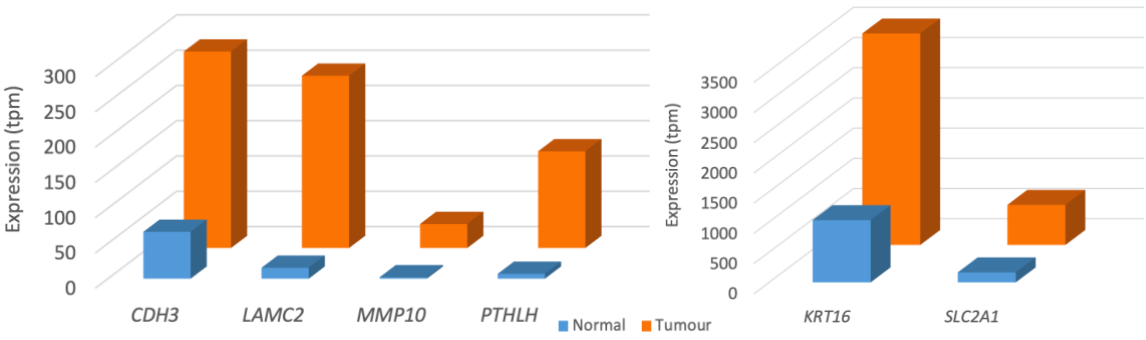


Figure 10. *Expression profile of the selected genes in HNSqCC tumour samples (orange bars) and paired normal tissues (blue bars)*. Data retrieved from GEPIA (http://gepia.cancer-pku.cn/index.html).

In *Supplementary Table 9 (Appendix I)*, information about expression of the mentioned genes in HNSqCC cell lines is shown. In this case, the selected cell lines (*Supplementary Table 6, Appendix I*) were: FaDu, which was established from a hypopharyngeal SqCC with mutated *CDKN2A*, *SMAD4* and *TP53*; HSC-2, established from an oral cavity SqCC, with mutated *CASP8*, *CDKN2A, PIK3CA, TP53* and *TP63*; and HSC-3, established from a tongue SqCC, with mutated *CASP8, CDKN2A, NOTCH1, SMAD4* and *TP53*. The findings of the literature search are shown in *Table 7*, along with other cancer types where they appear to be overexpressed.

Regarding *CDH3,* this gene codes for P-cadherin, a classical cadherin of the cadherin superfamily. It is a calcium-dependent cell-cell adhesion transmembrane protein important for maintaining cellular localization and tissue integrity. Strongly linked to E-cadherin, its role in tumorigenesis has been reported in different cancer types (Psyrri *et al.*, 2014). Apart from the evidence found in the literature (*Table 7*), *CDH3* has been included in this subgroup of the multigene panel for its constant and high expression across tumour tissues (278 tpm) and cell lines (values between 115 and 154 tpm) corresponding to HNSqCCs of different origin, as described above. Undoubtedly, its major limitation resides in its low specificity, which hopefully will be counterbalanced by the remaining selected markers.

A member of the keratin gene family, *KRT16*, has also been included due to its specificity. Coding for an intermediate filament protein, *KRT16* has been linked to tumorigenesis and metastasis, since its product regulates ECM molecules and integrins. Furthermore, *KRT16* silencing RNAs have been tested in oral SqCC cells, resulting in enhanced cytotoxicity and tumour killing effects (Huang *et al.*, 2019). Interestingly, its co-expression with *KRT14* led to assess the adequacy of this other keratin for this subgroup of the multigene panel. However, *KRT14* would be finally discarded due to high expression in several normal tissues. In fact, one of the limitations of *KRT16* for the purpose at issue is actually its high expression in normal head and neck tissues, together with relatively low expression in some of the most representative HNC cell lines. Nevertheless, both average cell lines' expression (*Supplementary Table 9, Appendix I*) and the data found in GEPIA (*Figure 10*) offers the hope that it might still be a promising candidate.

In section 4.1, *LAMB3*, encoding β3 chain of laminin 5, was included in the common genes' selection. In this case, it is the gene coding for laminin 5 γ2 chain, *LAMC2*, the one being incorporated. As it was already mentioned, laminins play important roles in keeping tissue architecture and regulating cell growth, migration and differentiation (Kuratomi *et al.*, 2008). HNSqCC cell lines exhibit quite high *LAMC2* expression, with values even higher than 1000 tpm in HSC-3 *(Supplementary Table 9, Appendix I)*. Despite this gene has been found to also be overexpressed in many other cancer types (*Table 7*), the fact that when comparing expression data in different cancerous and corresponding normal tissues the highest FC is by far observed in HNSqCC has encouraged its incorporation.

***Table 7**. Overview of the literature search results regarding HNSqCC gene selection and expression in other cancer types.* HNSqCC: Head and Neck Squamous Cell Carcinoma; REF: References; RT-qPCR: Reverse Transcription quantitative Polymerase Chain Reaction; SqCC: Squamous Cell Carcinoma.

Gene	Gene information	REF	Other cancer types
CDH3	High P-cadherin (*CDH3*'s product) expression in 84% of oral SqCCs tissues and cell lines analysed by immunohistochemistry.	Lo Muzio *et al.*, 2004	Breast, cervical, colorectal, lung, oesophagus, pancreatic, testicular, uterine.
	Immunohistochemistry analysis reveals strong membranous P-cadherin staining in laryngeal SqCC.	Psyrri *et al.*, 2014	
KRT16	Upregulation in invasive oral SqCC lines and tumour tissues (microarray analysis).	Huang *et al.*, 2019	Cervical, lung SqCC.
	Mass spectrometry analysis reveals high expression in laryngeal SqCC in comparison with non-cancerous samples.	Zha *et al.*, 2015	
LAMC2	Co-expression network analysis using integrative transcriptome datasets relates its overexpression with initiation and prognosis of oral SqCC. Might serve as potential target for early diagnosis.	Kisoda *et al.*, 2020	Cervical, colorectal, lung, oesophagus, pancreatic, thyroid.
	Serum protein levels (measured by immunoassay) can be used to monitor HNSqCC patients' progression.	Kuratomi *et al.*, 2008	
	Bioinformatic analyses show significant correlation with HNSqCC.	Zhao *et al.*, 2019	
MMP10	Immunohistochemistry reveals high expression in HNSqCCs. Significant correlation with invasiveness and metastasis.	Deraz *et al.*, 2011	Breast, lung, oesophagus, pancreatic.
	Immunohistochemistry shows high expression in HNSqCCs in comparison with basal cell carcinomas.	Kadeh *et al.*, 2016	
PTHLH	Expression data of different databases indicates overexpression in primary HNSqCCs in comparison with normal tissues.	Chang *et al.*, 2017	Lung SqCC.
	Real-time PCR shows upregulation in 89% of oral SqCC tissues assessed. This was confirmed using Western blot.	Lv *et al.*, 2014	
SLC2A1	Immunohistochemistry reveals higher expression in HNSqCC tissues in comparison with adjacent normal tissues.	Lin *et al.*, 2018	Bladder, cervical, colorectal, lung SqCC, oesophageal, ovarian, pancreatic.
	RT-qPCR shows expression in 84% of laryngeal cancer samples analysed. Significant differences are found when comparing expression data in tumour and adjacent normal laryngeal tissues.	Starska *et al.*, 2015	

Similar to *MMP11* (described in section 4.1), *MMP10* encodes a stromelysin. In this case, this metalloproteinase has been associated with invasion, migration, growth, apoptosis evasion and production of angiogenic and metastatic factors (Gobin *et al.*, 2019). *MMP10*'s incorporation was initially considered together with *MMP3* and *MMP13*, but these last two genes would finally be discarded because its expression data in HNSqCC was not considered to be high enough. According to Gobin *et al.* (2019), *MMP10* is almost universally upregulated across all cancer types (FC > 2). Expression data retrieved from GEPIA confirms this lack of specificity. However, this does not necessarily mean that the gene shows a very high and detectable expression across all cancer types. For instance, FC > 4 is obtained when analysing OC's expression data, but this is just because no expression is found in normal tissues (0 tpm) and a low expression of 4 tpm is found in OC tissues. This fact, together with the evidence provided by Kadeh *et al.* (2016) suggesting *MMP10* could be used to distinguish different HNC types, has led to its incorporation.

Regarding *PTHLH*, it is one of the most promising candidates. This gene encodes an autocrine/paracrine ligand that binds a family B G-protein coupled receptor and regulates cell proliferation and differentiation through activation of the cAMP/PKA or IP3/PKC signalling cascades (Chang *et al.*, 2017). Even though some HNSqCC cell lines (e.g. FaDu) show very low

expression of the gene (< 5 tpm), average data and both HSC-2 and HSC-3 show expressions higher than 70 tpm. However, the main reason why it was chosen was its apparent specificity, as it presents with quite high expression only in HNSqCC and LSqCC tissues.

Finally, *SLC2A1*, coding for one of the 14 members of the GLUT family (GLUT-1) of glucose transporter proteins, has also been incorporated. Its overexpression in certain tumours has been related with hypoxic conditions and the need to meet the high energy supply of cancerous cells on the basis of the Warburg effect (Lin *et al.*, 2018). Its expression is moderately high in the HNSqCC cell lines assessed, with FaDu (of hypopharyngeal origin) showing the highest values (331 tpm). Its major limitation is, as might be expected, its lack of specificity, since high *SLC2A1* expression has been reported in a wide range of carcinomas. Nevertheless, the fact that expression data in HNSqCC tissue doubles the values found for cervical, oesophageal or pancreatic cancer has endorsed its addition to this subgroup of the multigene panel.

4.5. Oesophageal carcinoma (OC) gene selection

For this last subgroup, 24 genes where initially selected, of which only 4 would end up being included in the multigene panel: *CYP2S1*, *MALAT1*, *PHC3* and *PLEC* (official full names and primers for their amplification displayed in *Supplementary Table 4, Appendix I*). *Figure 11* shows expression data in oesophageal normal and tumour tissues. The reason why most genes were discarded was that, despite being quite specific, they coded for small nucleolar RNAs, with very low expressions (< 3 tpm). In this sense, OC gene selection was by far the most difficult, since there were very few genes specific for this cancer type.

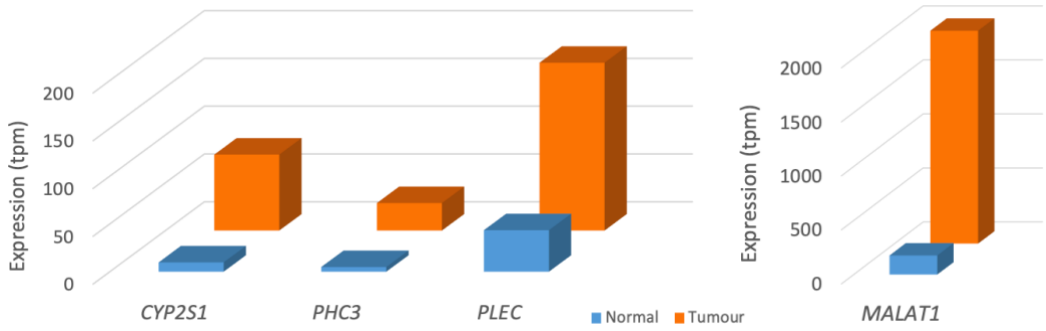


Figure 11. *Expression profile of the selected genes in OC tumour samples (orange bars) and paired normal tissues (blue bars)*. Data retrieved from GEPIA (http://gepia.cancer-pku.cn/index.html).

The results of the literature search concerning these genes can be found in *Table 8*, along with other cancer types where the selected genes appear to be overexpressed. 3 cell lines presenting OC's most frequent mutations were selected: OE19, established from an OADC, with mutated *SMAD2* and *TP53*; TE-1, from an OSqCC, with mutated *ERBB2, KRAS, SMAD4* and *TP53*; and KYSE-30, established from an OSqCC as well, with mutated *CDKN2A* and *TP53*. OC cell lines' selection is also summarised in *Supplementary Table 6 (Appendix I*) and data about candidate genes' expression in these OC cell lines can be found in *Supplementary Table 10 (Appendix I)*.

Perhaps the least well-known gene in this selection is *PHC3*, that encodes a member of the Polycomb repressive complex 1 (PRC1), which catalyses histone H2A ubiquitination. This complex includes several proteins responsible for epigenetic regulation. PRC1 has also been implicated in LC tumorigenesis, as it has been related with one of the most important risk factors in both LC and OC: tobacco smoke. In this sense, the carcinogens found in tobacco smoke are able to activate PRC1 genes, thereby increasing tumorigenicity of cancer cells (Crea *et al.*, 2013). Despite the literature search makes *PHC3* look quite promising, the obvious limitation of this candidate is its relatively low expression in OC tissues (30 tpm) and cell lines (10-15 tpm).

***Table 8**. Overview of the literature search results regarding OC gene selection and expression in other cancer types.* EMT: Epithelial-Mesenchymal Transition; OC: Oesophageal Carcinoma; OSqCC: Oesophageal Squamous Cell Carcinoma; REF: References; RT-qPCR: Reverse Transcription quantitative Polymerase Chain Reaction; SqCC: Squamous Cell Carcinoma.

Gene	Gene information	REF	Other cancer types
CYP2S1	High expression in tumour cells in hypoxic state compared to normal tissue. May play a role in metabolism and activation of carcinogens. Its expression is induced by polycyclic aromatic hydrocarbons.	Szaefer *et al.*, 2013	Colorectal, head and neck, lung SqCC, stomach, testicular.
MALAT1	Upregulation in OC tissues compared to para-tumour tissues (RT-qPCR). Its suppression inhibits tumour growth and EMT in mice.	Li *et al.*, 2020	Leukaemia, lung, ovarian, stomach.
	RT-qPCR results show remarkably increased expression in OSqCC cells compared to normal oesophageal cells. Its suppression attenuates stemness and decreases migration of OSqCC cells	Yao *et al.*, 2019	
PHC3	Gene amplification in epithelial neoplasms, which was found to be correlated with mRNA overexpression. May emerge as a novel oncogene, prognostic marker or target for epigenetic therapy.	Crea *et al.*, 2013	Leukaemia, lung, stomach.
PLEC	Bioinformatic analysis shows 2-fold upregulation in SqCC tissue. Immunohistochemistry shows overexpression in 84% of OSqCC cases. Expression was mainly cytoplasmic and membranous.	Pawar *et al.*, 2011	Head and neck, pancreatic, stomach, among others.

Regarding *CYP2S1*, this gene encodes a member of the cytochrome P450 superfamily of enzymes. These proteins, localised to the endoplasmic reticulum, are monooxygenases that catalyse different reactions related to drug metabolism and synthesis of lipids. Similar to *PHC3*, this gene has been selected due to its involvement in environmental carcinogens' metabolism and activation (Szaefer et al., 2013), since tobacco and alcohol consumption are widely known risk factors for OC. Regarding expression data, a great FC is observed in tissue samples (*Figure 11*), and cell lines' gene expression is not discouraging either (≈ 50 tpm on average). Interestingly, *CYP2S1*'s expression appears to be much higher in OADC cell lines when compared to OSqCC, so it may even be useful to differentiate between the two histologies. Its major limitation, however, is the lack of bibliographic evidence supporting its suitability, along with the fact that it happens to be overexpressed in other cancer types (*Table 8*).

In the case of *MALAT-1*, this gene does not code for a protein, but produces a precursor transcript from which a long non-coding RNA (lncRNA) is derived. Aberrant expression of lncRNAs has been reported to be involved in tumorigenesis through different mechanisms. Although *MALAT1* has been shown to have opposite effects depending on the cancer type (Yao *et al.*, 2019), recent studies show its expression is abnormally high and related to poor prognosis in OC (Li *et al.*, 2020). This information, corroborated by the expression data obtained from GEPIA (*Figure 11*) and OC cell lines, has encouraged its incorporation. However, its suitability must be tested experimentally, since although the data found in GEPIA makes *MALAT1* look very specific for OC, the information found in the literature reveals it has successfully been used as a prognostic biomarker for other cancer types, including lung, liver, or renal cancer (Li *et al.*, 2020).

Finally, *PLEC* codes for plectin 1, a member of the plakin family involved in crosslinking different cytoskeletal proteins for successful maintenance of cell architecture. Its cleavage by caspase-8 during the early stages of apoptosis has been reported. In this respect, the fact that most apoptotic pathways are dysregulated in cancer cells probably leads to plectin 1 accumulation (Pawar *et al.*, 2011). *PLEC* shows high expression in both OC tissues (176 tpm) and cell lines, particularly in those established from OSqCCs (260 – 309 tpm). Nevertheless, due to the fact that expression data in normal tissues displays values higher than 50 tpm in lung, bladder or uterus, its suitability for this subgroup of the multigene panel will have to be checked experimentally.

5. CONCLUSIONS

1. The results presented in this study confirm the unlikelihood of finding an overexpressed gene in a specific cancer type that shows low or no expression in all other cancer types and normal tissues. This way, the only option to obtain a successful multi-analyte blood test that allows early identification of a specific cancer type is using a combination of several different markers.

2. According to the evidence found in different genetic and bibliographic databases, *COL10A1, CST1, CTHRC1, CXCL9, CXCL13, EPCAM, KRT17, LAMB3, MMP1, MMP11, MMP12* and *UBE2C* are overexpressed in LADC, LSqCC, HNSqCC and OC cells (log_2FC > 1.5), and a multi-analyte blood test based on their combined expression in CTCs could allow for early detection of the mentioned cancer types.

3. Early-stage NSCLC's most frequent histologic types could be detected and distinguished using a multiparameter blood test on the basis of the combined expression in CTCs of different genes: *AGR2*, *CEACAM6*, *SMIM22*, *UBD* and *WFDC2* in the case of LADC; and *HAS3*, *NTS*, *SOX2* and *TP63* in the case of LSqCC.

4. In a similar way, the combined expression in CTCs of *CDH3*, *KRT16*, *LAMC2*, *MMP10*, *PI3*, *PTHLH* and *SLC2A1* in the case of HNSqCC, and of *CYP2S1*, *MALAT1*, *PHC3* and *PLEC* in the case of OC could be used for early diagnosis of these cancer types thanks to the development of the aforementioned multiparameter blood test.

5. Despite being overexpressed, the controversy generated by studies showing opposite conclusions regarding the suitability of some of the selected genes, together with the fact that some show lower expression than 100 tpm in cancer cells have made it clear that the adequacy of this gene selection for the early diagnostic cancer test at issue must be experimentally validated in tumour cell lines and CTCs derived from patients.

6. REFERENCES

ANDRIANI, F., LANDONI, E., MENSAH, M., FACCHINETTI, F., MICELI, R., TAGLIABUE, E., GIUSSANI, M., CALLARI, M., DE CECCO, L., COLOMBO, M. P., ROZ, L., PASTORINO, U., & SOZZI, G. (2018). Diagnostic role of circulating extracellular matrix-related proteins in non-small cell lung cancer. *BMC Cancer, 18*(1), 899.

ALFOUZAN, A. F. (2019). Head and neck cancer pathology: Old world versus new world disease. *Nigerian Journal of Clinical Practice, 22*(1), 1.

AMERICAN CANCER SOCIETY (2020). *Cancer Facts & Figures 2020*. Viewed on May 18th, 2020. Retrieved from: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2020/cancer-facts-and-figures-2020.pdf

BABAR, L., KOSOVEC, J. E., JAHANGIRI, V., CHOWDHURY, N., ZHENG, P., OMSTEAD, A. N., SALVITTI, M. S., SMITH, M. A., GOEL, A., KELLY, R. J., JOBE, B. A., & ZAIDI, A. H. (2019). Prognostic immune markers for recurrence and survival in locally advanced esophageal adenocarcinoma. *Oncotarget, 10*(44), 4546-4555.

BERTRAM, J. S. (2000). The molecular biology of cancer. *Molecular Aspects of Medicine, 21*(6), 167-223.

BRAY, F., FERLAY, J., SOERJOMATARAM, I., SIEGEL, R. L., TORRE, L. A., & JEMAL, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, *68*(6), 394-424.

BOFFETTA, P., AUTIER, P., BONIOL, M., BOYLE, P., HILL, C., AURENGO, A., MASSE, R., THÉ, G. DE, VALLERON, A.-J., MONIER, R., & TUBIANA, M. (2010). An estimate of cancers attributable to occupational exposures in France. *Journal of Occupational and Environmental Medicine, 52*(4), 399-406.

CALABUIG-FARIÑAS, S., JANTUS-LEWINTRE, E., HERREROS-POMARES, A., & CAMPS, C. (2016). Circulating tumor cells versus circulating tumor DNA in lung cancer—Which one will win? *Translational Lung Cancer Research, 5*(5), 466-482.

CESCON, D. W., BRATMAN, S. V., CHAN, S. M., & SIU, L. L. (2020). Circulating tumor DNA and liquid biopsy in oncology. *Nature Cancer, 1*(3), 276-290.

CHANG, K.-P., WU, C.-C., FANG, K.-H., TSAI, C.-Y., CHANG, Y.-L., LIU, S.-C., & KAO, H.-K. (2013). Serum levels of chemokine (C-X-C motif) ligand 9 (CXCL9) are associated with tumor progression and treatment outcome in patients with oral cavity squamous cell carcinoma. *Oral Oncology, 49*(8), 802-807.

CHANG, W.-M., LIN, Y.-F., SU, C.-Y., PENG, H.-Y., CHANG, Y.-C., HSIAO, J.-R., CHEN, C.-L., CHANG, J.-Y., SHIEH, Y.-S., HSIAO, M., & SHIAH, S.-G. (2017). Parathyroid hormone-like hormone is a poor prognosis marker of head and neck cancer and promotes cell growth via RUNX2 regulation. *Scientific Reports, 7*(1), 41131.

CHEN, Y.-F., MA, G., CAO, X., LUO, R.-Z., HE, L.-R., HE, J.-H., HUANG, Z.-L., ZENG, M.-S., & WEN, Z.-S. (2013). Overexpression of Cystatin SN positively affects survival of patients with surgically resected esophageal squamous cell carcinoma. *BMC Surgery, 13*(1), 15.

CHEN, B., GAO, S., JI, C., & SONG, G. (2017). Integrated analysis reveals candidate genes and transcription factors in lung adenocarcinoma. *Molecular Medicine Reports, 16*(6), 8371-8379.

CHEN, F., ZHENG, A., LI, F., WEN, S., CHEN, S., & TAO, Z. (2019). Screening and identification of potential target genes in head and neck cancer using bioinformatics analysis. *Oncology Letters 18*(3), 2955-2966.

CHOW, G., TAULER, J., & MULSHINE, J. L. (2010). Cytokines and growth factors stimulate hyaluronan production: Role of hyaluronan in epithelial to mesenchymal-like transition in non-small cell lung cancer. *Journal of Biomedicine & Biotechnology, 2010*, 485468.

CHUNG, K., NISHIYAMA, N., YAMANO, S., KOMATSU, H., HANADA, S., WEI, M., WANIBUCHI, H., SUEHIRO, S., & KAKEHASHI, A. (2011). Serum AGR2 as an early diagnostic and postoperative prognostic biomarker of human lung adenocarcinoma. *Cancer Biomarkers: Section A of Disease Markers, 10*(2), 101-107.

CHUNG, K., NISHIYAMA, N., WANIBUCHI, H., YAMANO, S., HANADA, S., WEI, M., SUEHIRO, S., & KAKEHASHI, A. (2012). AGR2 as a potential biomarker of human lung adenocarcinoma. *Osaka City Medical Journal, 58*(1), 13-24.

COHEN, J. D., JAVED, A. A., THOBURN, C., WONG, F., TIE, J., GIBBS, P., SCHMIDT, C. M., YIP-SCHNEIDER, M. T., ALLEN, P. J., SCHATTNER, M., BRAND, R. E., SINGHI, A. D., PETERSEN, G. M., HONG, S.-M., KIM, S. C., FALCONI, M., DOGLIONI, C., WEISS, M. J., AHUJA, N., … LENNON, A. M. (2017). Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers. *Proceedings of the National Academy of Sciences, 114*(38), 10202-10207.

COHEN, J. D., LI, L., WANG, Y., THOBURN, C., AFSARI, B., DANILOVA, L., DOUVILLE, C., JAVED, A. A., WONG, F., MATTOX, A., HRUBAN, R. H., WOLFGANG, C. L., GOGGINS, M. G., DAL MOLIN, M., WANG, T.-L., RODEN, R., KLEIN, A. P., PTAK, J., DOBBYN, L., … PAPADOPOULOS, N. (2018). Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science, 359*(6378), 926-930.

COLLINS, L. G., HAINES, C., PERKEL, R., & ENCK, R. E. (2007). Lung cancer: Diagnosis and management. *American Family Physician, 75*(1), 56-63.

COSTA, J. L., & SCHMITT, F. C. (2019). Liquid biopsy: A new tool in oncology. *Acta Cytologica, 63*(6), 448-448.

CREA, F., SUN, L., PIKOR, L., FRUMENTO, P., LAM, W. L., & HELGASON, C. D. (2013). Mutational analysis of Polycomb genes in solid tumours identifies PHC3 amplification as a possible cancer-driving genetic alteration. *British Journal of Cancer, 109*(6), 1699-1702.

DE GROOT, P. M., WU, C. C., CARTER, B. W., & MUNDEN, R. F. (2018). The epidemiology of lung cancer. *Translational lung cancer research, 7*(3), 220–233.

DERAZ, E. M., KUDO, Y., YOSHIDA, M., OBAYASHI, M., TSUNEMATSU, T., TANI, H., SIRIWARDENA, S. B. S. M., KEIKHAEE, M. R., KIEKHAEE, M. R., QI, G., IIZUKA, S., OGAWA, I., CAMPISI, G., LO MUZIO, L., ABIKO, Y., KIKUCHI, A., & TAKATA, T. (2011). MMP-10/stromelysin-2 promotes invasion of head and neck cancer. *PloS One, 6*(10), e25438.

DOMPER ARNAL, M. J., FERRÁNDEZ ARENAS, Á., & LANAS ARBELOA, Á. (2015). Esophageal cancer: Risk factors, screening and endoscopic treatment in Western and Eastern countries. *World journal of gastroenterology*, *21*(26), 7933–7943.

DU, Q., YAN, W., BURTON, V. H., HEWITT, S. M., WANG, L., HU, N., TAYLOR, P. R., ARMANI, M. D., MUKHERJEE, S., EMMERT-BUCK, M. R., & TANGREA, M. A. (2013). Validation of esophageal squamous cell carcinoma candidate genes from high-throughput transcriptomic studies. *American Journal of Cancer Research, 3*(4), 402-410.

DUPOUY, S., MOURRA, N., DOAN, V. K., GOMPEL, A., ALIFANO, M., & FORGEZ, P. (2011). The potential use of the neurotensin high affinity receptor 1 as a biomarker for cancer progression and as a component of personalized medicine in selective cancers. *Biochimie, 93*(9), 1369-1378.

GEPIA (2017). *NTS*. Viewed on May 25th, 2020. Retrieved from: http://gepia.cancer-pku.cn/detail.php?gene=nts

GOBIN, E., BAGWELL, K., WAGNER, J., MYSONA, D., SANDIRASEGARANE, S., SMITH, N., BAI, S., SHARMA, A., SCHLEIFER, R., & SHE, J.-X. (2019). A pan-cancer perspective of matrix metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. *BMC Cancer, 19*(1), 581.

HANAHAN, D., & WEINBERG, R. A. (2000). The hallmarks of cancer. *Cell, 100*(1), 57-70.

HANAHAN, D., & WEINBERG, R. A. (2011). Hallmarks of cancer: The next generation. *Cell, 144*(5), 646-674.

HE, W., ZHANG, H., WANG, Y., ZHOU, Y., LUO, Y., CUI, Y., JIANG, N., JIANG, W., WANG, H., XU, D., LI, S., WANG, Z., CHEN, Y., SUN, Y., ZHANG, Y., TSENG, H.-R., ZOU, X., WANG, L., & KE, Z. (2018). CTHRC1 induces non-small cell lung cancer (NSCLC) invasion through upregulating MMP-7/MMP-9. *BMC Cancer, 18*(1), 400.

HEITZER, E., HAQUE, I. S., ROBERTS, C. E. S., & SPEICHER, M. R. (2019). Current and future perspectives of liquid biopsies in genomics-driven oncology. *Nature Reviews Genetics, 20*(2), 71-88.

HIRSCH, F. R., SCAGLIOTTI, G. V., MULSHINE, J. L., KWON, R., CURRAN, W. J., WU, Y.-L., & PAZ-ARES, L. (2017). Lung cancer: Current therapies and new targeted treatments. *The Lancet, 389*(10066), 299-311.

HUANG, F.-L., & YU, S.-J. (2018). Esophageal cancer: Risk factors, genetic association, and treatment. *Asian Journal of Surgery, 41*(3), 210-215.

HUANG, W.-C., JANG, T.-H., TUNG, S.-L., YEN, T.-C., CHAN, S.-H., & WANG, L.-H. (2019). A novel miR-365-3p/EHF/keratin 16 axis promotes oral squamous cell carcinoma metastasis, cancer stemness and drug resistance via enhancing β5-integrin/c-met signaling pathway. *Journal of Experimental & Clinical Cancer Research: CR, 38*(1), 89.

JANTUS-LEWINTRE, E., USÓ, M., SANMARTÍN, E., & CAMPS, C. (2012). Update on biomarkers for the detection of lung cancer. *Lung Cancer: Targets and Therapy, 3*, 21–29.

KADARA, H., LACROIX, L., BEHRENS, C., SOLIS, L., GU, X., LEE, J. J., TAHARA, E., LOTAN, D., HONG, W. K., WISTUBA, I. I., & LOTAN, R. (2009). Identification of gene signatures and molecular markers for human lung cancer prognosis using an in vitro lung carcinogenesis system. *Cancer Prevention Research (Philadelphia, Pa.), 2*(8), 702-711.

KADEH, H., SARAVANI, S., HEYDARI, F., & SHAHRAKI, S. (2016). Differential immunohistochemical expression of matrix metalloproteinase-10 (MMP-10) in non-melanoma skin cancers of the head and neck. *Pathology, Research and Practice, 212*(10), 867-871.

KARABACAK, N. M., SPUHLER, P. S., FACHIN, F., LIM, E. J., PAI, V., OZKUMUR, E., MARTEL, J. M., KOJIC, N., SMITH, K., CHEN, P., YANG, J., HWANG, H., MORGAN, B., TRAUTWEIN, J., BARBER, T. A., STOTT, S. L., MAHESWARAN, S., KAPUR, R., HABER, D. A., & TONER, M. (2014). Microfluidic, marker-free isolation of circulating tumor cells from blood samples. *Nature protocols, 9*(3), 694-710.

KARACHALIOU, N., ROSELL, R., & VITERI, S. (2013). The role of SOX2 in small cell lung cancer, lung adenocarcinoma and squamous cell carcinoma of the lung. *Translational Lung Cancer Research, 2*(3), 172-179.

KATZ, R. L., ZAIDI, T. M., & NI, X. (2020). Liquid biopsy: Recent advances in the detection of circulating tumor cells and their clinical applications. In: M. M. Bui & L. Pantanowitz (Eds.), *Monographs in Clinical Cytology vol. 25*. Karger. Basilia: 43-66.

KHANOM, R., NGUYEN, C. T. K., KAYAMORI, K., ZHAO, X., MORITA, K., MIKI, Y., KATSUBE, K.-I., YAMAGUCHI, A., & SAKAMOTO, K. (2016). Keratin 17 is induced in oral cancer and facilitates tumor growth. *PloS One, 11*(8), e0161163.

KIM, Y., KIM, H. S., CUI, Z. Y., LEE, H.-S., AHN, J. S., PARK, C. K., PARK, K., & AHN, M.-J. (2009). Clinicopathological implications of EpCAM expression in adenocarcinoma of the lung. *Anticancer Research, 29*(5), 1817-1822.

KIMURA, H., KATO, H., FARIED, A., SOHDA, M., NAKAJIMA, M., FUKAI, Y., MIYAZAKI, T., MASUDA, N., FUKUCHI, M., & KUWANO, H. (2007). Prognostic significance of EpCAM expression in human esophageal cancer. *International Journal of Oncology, 30*(1), 171-179.

KISODA, S., SHAO, W., FUJIWARA, N., MOURI, Y., TSUNEMATSU, T., JIN, S., ARAKAKI, R., ISHIMARU, N., & KUDO, Y. (2020). Prognostic value of partial EMT-related genes in head and neck squamous cell carcinoma by a bioinformatic analysis*. Oral Diseases,* odi.13351.

KITA, Y., MIMORI, K., TANAKA, F., MATSUMOTO, T., HARAGUCHI, N., ISHIKAWA, K., MATSUZAKI, S., FUKUYOSHI, Y., INOUE, H., NATSUGOE, S., AIKOU, T., & MORI, M. (2009). Clinical significance of LAMB3 and COL7A1 mRNA in esophageal squamous cell carcinoma. *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 35*(1), 52-58.

KURATOMI, Y., SATO, S., MONJI, M., SHIMAZU, R., TANAKA, G., YOKOGAWA, K., INOUE, A., INOKUCHI, A., & KATAYAMA, M. (2008). Serum concentrations of laminin γ2 fragments in patients with head and neck squamous cell carcinoma. *Head & Neck, 30*(8), 1058-1063.

LAPA, R. M. L., BARROS-FILHO, M. C., MARCHI, F. A., DOMINGUES, M. A. C., DE CARVALHO, G. B., DRIGO, S. A., KOWALSKI, L. P., & ROGATTO, S. R. (2019). Integrated miRNA and mRNA expression analysis uncovers drug targets in laryngeal squamous cell carcinoma patients. *Oral Oncology, 93*, 76-84.

LATIMER, K. M., & MOTT, T. F. (2015). Lung cancer: Diagnosis, treatment principles, and screening. *American Family Physician, 91*(4), 250-256.

LEE, C. E., VINCENT-CHONG, V. K., RAMANATHAN, A., KALLARAKKAL, T. G., KAREN-NG, L. P., GHANI, W. M. N., RAHMAN, Z. A. A., ISMAIL, S. M., ABRAHAM, M. T., TAY, K. K., MUSTAFA, W. M. W., CHEONG, S. C., & ZAIN, R. B. (2015). Collagen triple helix repeat containing-1 (CTHRC1) expression in oral squamous cell carcinoma (OSqCC): Prognostic value and clinico-pathological implications. *International Journal of Medical Sciences, 12*(12), 937-945.

LEMJABBAR-ALAOUI, H., HASSAN, O., YANG, Y.-W., & BUCHANAN, P. (2015). Lung cancer: Biology and treatment options. *Biochimica et biophysica acta, 1856*(2), 189-210.

LI, J., WANG, X., ZHENG, K., LIU, Y., LI, J., WANG, S., LIU, K., SONG, X., LI, N., XIE, S., & WANG, S. (2019). The clinical significance of collagen family gene expression in esophageal squamous cell carcinoma. *PeerJ*, 7, e7705.

LI, M., XIAO, T., ZHANG, Y., FENG, L., LIN, D., LIU, Y., MAO, Y., GUO, S., HAN, N., DI, X., ZHANG, K., CHENG, S., & GAO, Y. (2010). Prognostic significance of matrix metalloproteinase-1 levels in peripheral plasma and tumour tissues of lung cancer patients. *Lung Cancer, 69*(3), 341-347.

LI, Q., DAI, Z., XIA, C., JIN, L., & CHEN, X. (2020). Suppression of long non-coding RNA MALAT1 inhibits survival and metastasis of esophagus cancer cells by sponging miR-1-3p/CORO1C/TPM3 axis. *Molecular and Cellular Biochemistry, 470*(1-2), 165-174.

LI, Y., WANG, X., SHI, L., XU, J., & SUN, B. (2019). Predictions for high COL1A1 and COL10A1 expression resulting in a poor prognosis in esophageal squamous cell carcinoma by bioinformatics analyses. *Translational Cancer Research, 9*(1), 85-94.

LIN, W., YIN, C.-Y., YU, Q., ZHOU, S.-H., CHAI, L., FAN, J., & WANG, W.-D. (2018). Expression of glucose transporter-1, hypoxia inducible factor-1α and beclin-1 in head and neck cancer and their implication. *International Journal of Clinical and Experimental Pathology, 11*(7), 3708-3717.

LIU, J., LIU, L., CAO, L., & WEN, Q. (2018). Keratin 17 promotes lung adenocarcinoma progression by enhancing cell proliferation and invasion. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research, 24*, 4782-4790.

LIU, L., JUNG, S.-N., OH, C., LEE, K., WON, H.-R., CHANG, J. W., KIM, J. M., & KOO, B. S. (2019). LAMB3 is associated with disease progression and cisplatin cytotoxic sensitivity in head and neck squamous cell carcinoma. *European Journal of Surgical Oncology: The Journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology, 45*(3), 359-365.

LIU, Y., & YAO, J. (2019). Research progress of cystatin SN in cancer. *OncoTargets and therapy, 12*, 3411-3419.

LIU, Z., YU, S., YE, S., SHEN, Z., GAO, L., HAN, Z., ZHANG, P., LUO, F., CHEN, S., & KANG, M. (2020). Keratin 17 activates AKT signalling and induces epithelial-mesenchymal transition in oesophageal squamous cell carcinoma. *Journal of Proteomics, 211*, 103557.

LO MUZIO, L., PANNONE, G., MIGNOGNA, M. D., STAIBANO, S., MARIGGIÒ, M. A., RUBINI, C., PROCACCINI, M., DOLCI, M., BUFO, P., DE ROSA, G., & PIATTELLI, A. (2004). P-cadherin expression predicts clinical outcome in oral squamous cell carcinomas. *Histology and Histopathology, 19*(4), 1089-1099.

LV, Z., WU, X., CAO, W., SHEN, Z., WANG, L., XIE, F., ZHANG, J., JI, T., YAN, M., & CHEN, W. (2014). Parathyroid hormone-related protein serves as a prognostic indicator in oral squamous cell carcinoma. *Journal of Experimental & Clinical Cancer Research: CR, 33*, 100.

MADER, S., & PANTEL, K. (2017). Liquid biopsy: Current status and future perspectives. *Oncology Research and Treatment, 40*(7-8), 404-408.

MALHOTRA, J., MALVEZZI, M., NEGRI, E., LA VECCHIA, C., & BOFFETTA, P. (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, *48*(3), 889-902.

MAO, W.-M., ZHENG, W.-H., & LING, Z.-Q. (2011). Epidemiologic risk factors for esophageal cancer development. *Asian Pacific Journal of Cancer Prevention: APJCP*, *12*(10), 2461-2466.

MARUR, S., & FORASTIERE, A. A. (2008). Head and neck cancer: Changing epidemiology, diagnosis, and treatment. *Mayo Clinic Proceedings, 83*(4), 489-501.

MASSION, P. P., TAFLAN, P. M., RAHMAN, S. M. J., YILDIZ, P., SHYR, Y., CARBONE, D. P., & GONZALEZ, A. L. (2004). Role of p63 amplification and overexpression in lung cancer development. *Chest, 125*(5), 102S.

METODIEVA, S. N., NIKOLOVA, D. N., CHERNEVA, R. V., DIMOVA, I. I., PETROV, D. B., & TONCHEVA, D. I. (2011). Expression analysis of angiogenesis-related genes in Bulgarian patients with early-stage non-small cell lung cancer. *Tumori, 97*(1), 86-94.

MEVES, V., BEHRENS, A., & POHL, J. (2015). Diagnostics and Early Diagnosis of Esophageal Cancer. *Viszeralmedizin*, *31*(5), 315–318.

MIRSADRAEE, S., OSWAL, D., ALIZADEH, Y., CAULO, A., & VAN BEEK, E., JR (2012). The 7th lung cancer TNM classification and staging system: Review of the changes and implications. *World journal of radiology, 4*(4), 128–134.

NATIONAL CANCER INSTITUTE. (2015). *What is cancer?* Viewed on May 1st, 2020. Retrieved from: https://www.cancer.gov/about-cancer/understanding/what-is-cancer

NATIONAL CANCER INSTITUTE. (2017). *Head and Neck cancers.* Viewed on May 10th, 2020. Retrieved from: https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet

NATIONAL HEALTH SECURITY (2019). *Oesophageal cancer.* Viewed on May 16th, 2020. Retrieved from: https://www.nhs.uk/conditions/oesophageal-cancer/

NAPIER, K. J., SCHEERER, M., & MISRA, S. (2014). Esophageal cancer: A Review of epidemiology, pathogenesis, staging workup and treatment modalities. *World journal of gastrointestinal oncology*, *6*(5), 112–120.

PAI, S. I., & WESTRA, W. H. (2009). Molecular pathology of head and neck cancer: implications for diagnosis, prognosis, and treatment. *Annual review of pathology, 4*, 49–70.

PALUMBO, A., DA COSTA, N. M., DE MARTINO, M., SEPE, R., PELLECCHIA, S., DE SOUSA, V. P. L., NICOLAU NETO, P., KRUEL, C. D., BERGMAN, A., NASCIUTTI, L. E., FUSCO, A., & PINTO, L. F. R. (2016). UBE2C is overexpressed in ESCC tissues and its abrogation attenuates the malignant phenotype of ESCC cell lines. *Oncotarget, 7*(40), 65876-65887.

PAWAR, H., KASHYAP, M. K., SAHASRABUDDHE, N. A., RENUSE, S., HARSHA, H. C., KUMAR, P., SHARMA, J., KANDASAMY, K., MARIMUTHU, A., NAIR, B., RAJAGOPALAN, S., MAHARUDRAIAH, J., PREMALATHA, C. S., KUMAR, K. V. V., VIJAYAKUMAR, M., CHAERKADY, R., PRASAD, T. S. K., KUMAR, R. V., KUMAR, R. V., & PANDEY, A. (2011). Quantitative tissue proteomics of esophageal squamous cell carcinoma for novel biomarker discovery. *Cancer Biology & Therapy, 12*(6), 510-522.

PENG, S., LI, X., LIU, Q., ZHANG, Y., ZOU, L., GONG, X., WANG, M., & MA, X. (2019). Identification of differentially expressed genes between lung adenocarcinoma and squamous cell carcinoma using transcriber signature analysis. *Journal of Southern Medical University, 39*(6), 641-649.

PIZZI, M., FASSAN, M., BALISTRERI, M., GALLIGIONI, A., REA, F., & RUGGE, M. (2012). Anterior gradient 2 overexpression in lung adenocarcinoma. *Applied Immunohistochemistry & Molecular Morphology: AIMM, 20*(1), 31-36.

POLYCARPOU-SCHWARZ, M., GROß, M., MESTDAGH, P., SCHOTT, J., GRUND, S. E., HILDENBRAND, C., ROM, J., AULMANN, S., SINN, H.-P., VANDESOMPELE, J., & DIEDERICHS, S. (2018). The cancer-associated microprotein CASIMO1 controls cell proliferation and interacts with squalene epoxidase modulating lipid droplet formation. *Oncogene, 37*(34), 4750-4768.

POULET, G., MASSIAS, J., & TALY, V. (2019). Liquid biopsy: General concepts. *Acta Cytologica, 63*(6), 449-455.

PSYRRI, A., KOTOULA, V., FOUNTZILAS, E., ALEXOPOULOU, Z., BOBOS, M., TELEVANTOU, D., KARAYANNOPOULOU, G., KRIKELIS, D., MARKOU, K., KARASMANIS, I., ANGOURIDAKIS, N., KALOGERAS, K. T., NIKOLAOU, A., & FOUNTZILAS, G. (2014). Prognostic significance of the Wnt pathway in squamous cell laryngeal cancer. *Oral Oncology, 50*(4), 298-305.

RELLI, V., TREROTOLA, M., GUERRA, E., & ALBERTI, S. (2018). Distinct lung cancer subtypes associate to distinct drivers of tumor progression. *Oncotarget, 9*(85), 35528-35540.

RIDGE, C. A., MCERLEAN, A. M., & GINSBERG, M. S. (2013). Epidemiology of lung cancer. *Seminars in interventional radiology*, *30*(2), 93–98.

RODRIGUEZ-CANALES, J., PARRA-CUENTAS, E., & WISTUBA, I. I. (2016). Diagnosis and molecular classification of lung cancer. In: K. L. Reckamp (Eds.), *Lung Cancer. Cancer Treatment and Research vol. 170*. Springer. Cham: 25-46.

RUSHTON, L., HUTCHINGS, S. J., FORTUNATO, L., YOUNG, C., EVANS, G. S., BROWN, T., BEVAN, R., SLACK, R., HOLMES, P., BAGGA, S., CHERRIE, J. W., & VAN TONGEREN, M. (2012). Occupational cancer burden in Great Britain. *British Journal of Cancer, 107*(S1), S3-S7.

SAMBANDAM, Y., SUNDARAM, K., LIU, A., KIRKWOOD, K. L., RIES, W. L., & REDDY, S. V. (2013). CXCL13 activation of c-myc induce rank ligand expression in stromal/preosteoblast cells in the oral squamous cell carcinoma tumor-bone microenvironment. *Oncogene, 32*(1), 97-105.

SEN, S., & CARNELIO, S. (2016). Expression of epithelial cell adhesion molecule (EpCAM) in oral squamous cell carcinoma. *Histopathology, 68*(6), 897-904.

SHAIKH, I., ANSARI, A., AYACHIT, G., GANDHI, M., SHARMA, P., BHAIRAPPANAVAR, S., JOSHI, C. G., & DAS, J. (2019). Differential gene expression analysis of HNSqCC tumors deciphered tobacco dependent and independent molecular signatures. *Oncotarget, 10*(58), 6168-6183.

SHARMA, S., ZHUANG, R., LONG, M., PAVLOVIC, M., KANG, Y., ILYAS, A., & ASGHAR, W. (2018). Circulating tumor cell isolation, culture, and downstream molecular analysis. *Biotechnology advances, 36*(4), 1063-1078.

SINGER, B. B., SCHEFFRAHN, I., KAMMERER, R., SUTTORP, N., ERGUN, S., & SLEVOGT, H. (2010). Deregulation of the CEACAM expression pattern causes undifferentiated cell growth in human lung adenocarcinoma cells. *PloS One, 5*(1), e8747.

SINGH, R., GUPTA, P., KLOECKER, G. H., SINGH, S., & LILLARD, J. W. (2014). Expression and clinical significance of CXCR5/CXCL13 in human non-small cell lung carcinoma. *International Journal of Oncology, 45*(6), 2232-2240.

SKOULIDIS, F., & HEYMACH, J. V. (2019). Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nature Reviews Cancer, 19*(9), 495-509.

SPAKS, A., JAUNALKSNE, I., SPAKA, I., CHUDASAMA, D., PIRTNIEKS, A., & KRIEVINS, D. (2015). Diagnostic value of circulating cxc chemokines in non-small cell lung cancer. *Anticancer Research, 35*(12), 6979-6983.

STARSKA, K., FORMA, E., JÓŹWIAK, P., BRYŚ, M., LEWY-TRENDA, I., BRZEZIŃSKA-BŁASZCZYK, E., & KRZEŚLAK, A. (2015). Gene and protein expression of glucose transporter 1 and glucose transporter 3 in human laryngeal cancer—The relationship with regulatory hypoxia-inducible factor-1α expression, tumor invasiveness, and patient prognosis. *Tumor Biology, 36*(4), 2309-2321.

SZAEFER, H., CICHOCKI, M., & MAJCHRZAK-CELIŃSKA, A. (2013). New cytochrome P450 isoforms as cancer biomarkers and targets for chemopreventive and chemotherapeutic agents. *Postepy Higieny I Medycyny Doswiadczalnej, 67*, 709-718.

TANG, Z., LI, C., KANG, B., GAO, G., LI, C., & ZHANG, Z. (2017). GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research, 45*(W1), W98-W102.

TESTA, U., CASTELLI, G., & PELOSI, E. (2018). Lung cancers: Molecular characterization, clonal heterogeneity and evolution, and cancer stem cells. *Cancers, 10*(8).

UMAR, S. B., & FLEISCHER, D. E. (2008). Esophageal cancer: Epidemiology, pathogenesis and prevention. *Nature Clinical Practice. Gastroenterology & Hepatology, 5*(9), 517-526.

VACHANI, A., NEBOZHYN, M., SINGHAL, S., ALILA, L., WAKEAM, E., MUSCHEL, R., POWELL, C. A., GAFFNEY, P., SINGH, B., BROSE, M. S., LITZKY, L. A., KUCHARCZUK, J., KAISER, L. R., MARRON, J. S., SHOWE, M. K., ALBELDA, S. M., & SHOWE, L. C. (2007). A 10-gene classifier for distinguishing head and neck squamous cell carcinoma and lung squamous cell carcinoma. *Clinical Cancer Research, 13*(10), 2905-2915.

WANG, C., LI, Z., SHAO, F., YANG, X., FENG, X., SHI, S., GAO, Y., & HE, J. (2017). High expression of Collagen Triple Helix Repeat Containing 1 (CTHRC1) facilitates progression of oesophageal squamous cell carcinoma through MAPK/MEK/ERK/FRA-1 activation. *Journal of Experimental & Clinical Cancer Research, 36*(1), 84.

WANG, G.-Z., CHENG, X., ZHOU, B., WEN, Z.-S., HUANG, Y.-C., CHEN, H.-B., LI, G.-F., HUANG, Z.-L., ZHOU, Y.-C., FENG, L., WEI, M.-M., QU, L.-W., CAO, Y., & ZHOU, G.-B. (2015). The chemokine CXCL13 in lung cancers associated with environmental polycyclic aromatic hydrocarbons pollution. *ELife, 4*.

WANG, X.-M., LI, J., YAN, M.-X., LIU, L., JIA, D.-S., GENG, Q., LIN, H.-C., HE, X.-H., & YAO, M. (2013). Integrative analyses identify osteopontin, LAMB3 and ITGB1 as critical pro-metastatic genes for lung cancer. *PloS One, 8*(2), e55714.

WANG, Y., WANG, Z., DING, Y., SUN, F., & DING, X. (2019). The application value of serum HE4 in the diagnosis of lung cancer. *Asian Pacific Journal of Cancer Prevention: APJCP, 20*(8), 2405-2407.

WANG, Z., YANG, M.-Q., LEI, L., FEI, L.-R., ZHENG, Y.-W., HUANG, W.-J., LI, Z.-H., LIU, C.-C., & XU, H.-T. (2019). Overexpression of KRT17 promotes proliferation and invasion of non-small cell lung cancer and indicates poor prognosis. *Cancer Management and Research, 11*, 7485-7497.

WARNECKE-EBERZ, U., METZGER, R., HÖLSCHER, A. H., DREBBER, U., & BOLLSCHWEILER, E. (2016). Diagnostic marker signature for esophageal cancer from transcriptome analysis. *Tumor Biology, 37*(5), 6349-6358.

WONG, K.-C., CHEN, J., ZHANG, J., LIN, J., YAN, S., ZHANG, S., LI, X., LIANG, C., PENG, C., LIN, Q., KWONG, S., & YU, J. (2019). Early cancer detection from multianalyte blood test results. *iScience, 15*, 332-341.

WORLD HEALTH ORGANIZATION. (2020). *WHO report on cancer: setting priorities, investing wisely and providing care for all*. Viewed on May 1st, 2020. Retrieved from: https://apps.who.int/iris/handle/10665/330745

WU, X., ZHANG, W., HU, Y., & YI, X. (2015). Bioinformatics approach reveals systematic mechanism underlying lung adenocarcinoma. *Tumori Journal, 101*(3), 281-286.

WU, Y., DU, X., XUE, C., LI, D., ZHENG, Q., LI, X., & CHEN, H. (2013). Quantification of serum SOX2 DNA with FQ-PCR potentially provides a diagnostic biomarker for lung cancer. Medical Oncology, 30(4), 737.

XUE, F., ZHU, L., MENG, Q.-W., WANG, L., CHEN, X.-S., ZHAO, Y.-B., XING, Y., WANG, X.-Y., & CAI, L. (2016). FAT10 is associated with the malignancy and drug resistance of non-small-cell lung cancer. *OncoTargets and Therapy, 9*, 4397-4409.

YAO, Q., YANG, J., LIU, T., ZHANG, J., & ZHENG, Y. (2019). Long non-coding RNA MALAT1 promotes the stemness of esophageal squamous cell carcinoma by enhancing YAP transcriptional activity. *FEBS Open Bio, 9*(8), 1392-1402.

ZENG, Q., LIU, M., ZHOU, N., LIU, L., & SONG, X. (2016). Serum human epididymis protein 4 (HE4) may be a better tumor marker in early lung cancer. *International Journal of Clinical Chemistry, 455*, 102-106.

ZHA, C., JIANG, X. H., & PENG, S. F. (2015). iTRAQ-based quantitative proteomic analysis on S100 calcium binding protein A2 in metastasis of laryngeal cancer. *PloS One, 10*(4), e0122322.

ZHAO, L., CHI, W., CAO, H., CUI, W., MENG, W., GUO, W., & WANG, B. (2019). Screening and clinical significance of tumor markers in head and neck squamous cell carcinoma through bioinformatics analysis. *Molecular Medicine Reports, 19*(1), 143-154.

ZHENG, M. (2016). Classification and Pathology of Lung Cancer. *Surgical Oncology Clinics of North America, 25*(3), 447–468.

7. APPENDICES

7.1. Appendix I. Supplementary tables.

Supplementary Table 1*. Risk factors associated with oesophageal squamous cell carcinoma and adenocarcinoma.* GERD: Gastroesophageal Reflux Disease; +: associated risk; -: no risk associated. Retrieved from Domper Arnal *et al.*, 2015.

Risk factor	Oesophageal squamous cell carcinoma	Oesophageal adenocarcinoma
Geography	South-eastern Africa, Asia, Iran, South America.	Western Europe, North America, Australia.
Race	Black > White	White > Black
Gender	Male > Female	Male > Female
Alcohol	++++	-
Tobacco	++++	++
Obesity	-	+++
GERD	-	++++
Diet low in fruits and vegetables	++	+
Socioeconomic conditions	++	-
Genetic aspects	++	+

Supplementary Table 2. *Comparison of the advantages and limitations of conventional tissue biopsy and liquid biopsy.* CTCs: Circulating Tumour Cells. Adapted from Poulet *et al.*, 2019.

Conventional tissue biopsy	Liquid biopsy
Gold standard.	Clinical interest under investigation.
Accessible to histological analysis and staging.	The possibility of a histological analysis is limited to obtention of CTCs.
Sometimes unavailable.	Easy to obtain & faster turn-around time. Low level of tumour-derived products in body fluids, increasing risk of false negative results.
Invasive procedure, discomfort for patients.	Minimally invasive.
Potential high yield of DNA but risk of DNA degradation/cross-link. DNA quantity highly variable with sampling method.	Quantity and quality of DNA strongly dependent on pre-analytical and analytical processes.
Localised analysis, no characterization of intra- or inter-tumour heterogeneity (metastasis), especially in advanced stages.	Allows, if enough DNA is available, to highlight both intra- and inter-tumour heterogeneity.
Not applicable to serial monitoring. No possibility of dynamic follow-up of cancer molecular modifications.	Applicable to serial monitoring. Dynamic follow-up of tumour evolution.

***Supplementary Table 3**. Main reasons why candidate genes of the initial common gene selection were discarded.* This initial selection comprises overexpressed genes in tumour tissue of all 4 cancer types assessed (LADC, LSqCC, HNSqCC and OC) with a $log_2FC > 1.5$. In the case of low expression in cancer types of interest, < 25 transcripts per million (tpm) was set as threshold. For high expression in normal tissues a > 50 tpm cutoff was set and in the case of expression in white blood cells (WBCs), > 15 tpm was used as threshold. Primer availability was not checked (-) for genes failing to meet two or more of the other selection criteria. The criteria used to establish primer availability are listed in section 3.5. IS: immune system; WBCs: white blood cells.

Discarded gene	Low expression in cancer types of interest	Not specific for cancer types of interest	High expression in normal tissues	Expression in WBCs	IS gene	Primers availability
ALG1L	X	X				-
ANLN	X	X		X		-
AURKA	X	X		X		-
CA9	X		X			-
CDKN2A		X				X
CEP55	X	X		X		-
COL1A1		X	X			-
CXCL10	X	X				-
DSG2		X		X		-
DTL	X	X		X		-
FOXM1		X		X		-
IGF2BP3	X			X		-
IGFBP3		X	X			-
IGHG1					X	-
IGHG2		X			X	-
IGHG3		X			X	-
IGHG4		X			X	-
IGHV1-69-2					X	-
IGHV4-34		X			X	-
ITGB4		X	X			-
MARCKSL1		X	X	X		-
MCM2	X	X		X		-
MMP7	X		X			-
MMP9		X	X			-
MYBL2		X		X		-
PLK1	X	X		X		-
SPP1		X	X			-
SULF1		X	X			-
TOP2A		X		X		-
TPX2	X	X		X		-
TRIP13	X	X		X		-
UBE2T		X		X		-

***Supplementary Table 4**. Candidate genes' official symbol, name and forward and reverse primers for their amplification*. Primer design details are explained in section 3.5. Official gene full names retrieved from HUGO (https://www.genenames.org).

Official gene symbol	Official gene full name	Forward primer	Reverse primer
AGR2	Anterior gradient 2, protein disulphide isomerase family member	AAGGCAGGTGGGTGAGGAAATC	TGGGTCGAGAGTCCTTTGTGT
CDH3	Cadherin 3	GGGAGCCTGTGTGTGTCTAC	GTCTCTCAGGATGCGGTAGC
CEACAM6	Carcinoembryonic antigen cell adhesion molecule 6	ACTCAGCGTCAAAAGGAACG	GACGGTAATTGGCCTTTGAG
COL10A1	Collagen type X alpha 1 chain	AAAGGCCCACTACCCAACAC	GTGGACCAGGAGTACCTTGC
CST1	Cystatin-SN	CCCGGGTGGCATCTATAACG	GGTCTGTTGCCTGGCTCTTA
CTHRC1	Collagen triple helix repeat containing 1	GATCCCCAAGGGGAAGCAAA	GGCCCTTGTAAGCACATTCC
CXCL9	CXC motif chemokine ligand 9	GTGCAAGGAACCCCAGTAGT	GGTGGATAGTCCCTTGGTTGG
CXCL13	CXC motif chemokine ligand 13	CAGCCTCTCTCCAGTCCAAG	ATTCCCACGGGGCAAGATTT
CYP2S1	Cytochrome P450 family 2 subfamily S member 1	GGCTATACCCTCCTGCTCCT	CTCCCGATTCAGCTCCTCAC
EPCAM	Epithelial cell adhesion molecule	TACAAGCTGGCCGTAAACTG	GCCAGCTTTGAGCAAATGAC
HAS3	Hyaluronan synthase 3	ATCCCCAAGTAGGGGGAGTC	CAGCCAAAGTAGGACTGGCA
KRT16	Keratin 16	ACGAGCAGATGGCAGAGAAAAA	GCTGCTCTGTACCAGTTCGC
KRT17	Keratin 17	AATCCTGCTGGATGTGAAGACG	GTACTGAGTCAGGTGGGCATC
LAMB3	Laminin subunit beta 3	CTTCTACAACAACCGGCCCT	CAAACACAGCGGGGTCAAAG
LAMC2	Laminin subunit gamma 2	GGAGCTGGAGTTTGACACGA	CAGCGTTCTTGGCTCTGGTA
MALAT1	Metastasis associated lung adenocarcinoma transcript 1	CTGGGGCTCAGTTGCGTAAT	CTCACAAAACCCCCGGAACT
MMP1	Matrix metallopeptidase 1	AGAGCAGATGTGGACCATGC	TTGTCCCGATGATCTCCCCT
MMP10	Matrix metallopeptidase 10	AGTTTGGCTCATGCCTACCC	CAGGGAGTGGCCAAGTTCAT
MMP11	Matrix metallopeptidase 11	AAGAGGTTCGTGCTTTCTGG	ATCGCTCCATACCTTTAGGG
MMP12	Matrix metallopeptidase 12	TTTGGTGGTTTTTGCCCGTG	TCGAAATGTGCATCCCCTCC
NTS	Neurotensin	GCAGGGCTTTTCAACACTGG	TCATACAGCTGCCGTTTCAGA
PHC3	Polyhomeotic homolog 3	GCTGCTGTTCAGGCAAGTTT	GAAGCCTGGGAACGGCTTAT

Official gene symbol	Official gene full name	Forward primer	Reverse primer
PLEC	Plectin	ACCAAGTGGGTCAACAAGCA	CCAGCAGGGAGATGAGGTTG
PTHLH	Parathyroid hormone like hormone	GGAGACTGGTTCAGCAGTGG	CCCTTGTCATGGAGGAGCTG
SLC2A1	Solute carrier family 2 member 1	TGGCATCAACGCTGTCTTCT	AGCCAATGGTGGCATACACA
SMIM22	Small integral membrane protein 22	CCCCAGGAAGGAAAGACCCA	CAGACGGGGACTGGAAGACA
SOX2	SRY-box transcription factor 2	AGGATAAGTACACGCTGCCC	TAACTGTCCATGCGCTGGTT
TP63	Tumour protein p63	CTGCCCTGACCCTTACATCC	TGGGACATGGTGGATCGGTA
UBD	Ubiquitin D	AGATGGCTCCCAATGCTTC	TCACGCTGTCATATGGGTTG
UBE2C	Ubiquitin conjugating enzyme E2C	TTCCTGTCTCTCTGCCAACG	CTCCTGCTGTAGCCTTTTGC
WFDC2	WAP four-disulfide core domain 2	CCCTAGTCTCAGGCACAGGA	CTGTCCGAGACGCACTCTTG

Supplementary Table 5*. Common selection genes' average expression data in LADC, LSqCC, HNSqCC and OC cell lines.* The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). LADC: Lung Adenocarcinoma; LSqCC: Lung Squamous Cell Carcinoma; HNSqCC: Head and Neck Squamous Cell Carcinoma; OC: Oesophageal Carcinoma; tpm: transcripts per million.

Gene	LADC expression (tpm)	LADC cell lines	LSqCC expression (tpm)	LSqCC cell lines	HNSqCC expression (tpm)	HNSqCC cell lines	OC expression (tpm)	OC cell lines
COL10A1	0.51	51	0.30	20	0.17	10	0.27	21
CST1	27.40	43	39.60	13	0.26	8	0.39	12
CTHRC1	44.83	57	25.50	22	9.16	13	10.18	24
CXCL9	0.33	4	0.10	5	0.23	4	0.10	2
CXCL13	0.18	19	0.13	4	0.15	4	0.31	10
EPCAM	389.28	57	332.88	22	239.00	13	296.68	25
KRT17	172.05	57	340.58	22	2543.38	13	1241.20	25
LAMB3	220.80	57	193.28	22	405.92	13	286.04	25
MMP1	107.84	57	39.10	22	316.54	13	24.52	25
MMP11	2.23	57	1.85	22	0.96	13	2.86	25
MMP12	0.24	19	0.21	8	2.23	12	1.50	13
UBE2C	278.23	57	302.18	22	180.54	13	243.64	25

***Supplementary Table 6**. Selected cell lines' genetic alterations and cancer type of origin.* Data retrieved from ATCC (https://www.atcc.org) and DepMap (https://depmap.org). LADC: Lung Adenocarcinoma; LSqCC: Lung Squamous Cell Carcinoma; OADC: Oesophageal Adenocarcinoma; OSqCC: Oesophageal Squamous Cell Carcinoma; SqCC: Squamous Cell Carcinoma.

Cell line	Cancer type	Genetic alterations
A549	LADC	Mutated *CDKN2A* and *KRAS*.
NCI-H1395	LADC	Mutated *BRAF*.
NCI-H1975	LADC	Mutated *CDKN2A, EGFR, PIK3CA* and *TP53*.
NCI-H2228	LADC	*EML4-ALK* fusion.
NCI-H2170	LSqCC	Mutated *CDKN2A* and *TP53*.
HCC-95	LSqCC	*PIK3CA* amplification.
SW 900	LSqCC	Mutated *CDKN2A, KRAS* and *TP53*.
FaDu	Hypopharyngeal SqCC	Mutated *CDKN2A, SMAD4* and *TP53*.
HSC-2	Oral cavity SqCC	Mutated *CASP8, CDKN2A, PIK3CA, TP53* and *TP63*.
HSC-3	Tongue SqCC	Mutated *CASP8, CDKN2A, NOTCH1, TP53* and *SMAD4*.
OE19	OADC	Mutated *SMAD2* and *TP53*.
TE-1	OSqCC	Mutated *ERBB2, KRAS, SMAD4* and *TP53*.
KYSE-30	OSqCC	Mutated *CDKN2A* and *TP53*.

***Supplementary Table 7**. LADC candidate genes' expression data in selected LADC cell lines.* An average expression value was obtained using expression data of the genes of interest in several LADC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). LADC: Lung Adenocarcinoma; tpm: transcripts per million.

Cell line	A549 expression (tpm)	NCI-H1395 expression (tpm)	NCI-H1975 expression (tpm)	NCI-H2228 expression (tpm)	Average expression in tpm (all LADC cell lines)	LADC cell lines
AGR2	156.0	2670.0	2.0	167.0	430.4	57
CEACAM6	11.0	393.0	0.5	97.0	575.1	57
SMIM22	0.1	73.0	2.0	17.0	24.2	53
UBD	0.3	-	-	0.3	8.9	41
WFDC2	0.3	1.0	2.0	83	45.4	55

***Supplementary Table 8**. LSqCC candidate genes' expression data in selected LSqCC cell lines.* An average expression value was obtained using expression data of the genes of interest in several LSqCC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). LSqCC: Lung Squamous Cell Carcinoma; tpm: transcripts per million.

Cell line	NCI-H2170 expression (tpm)	HCC-95 expression (tpm)	SW 900 expression (tpm)	Average expression in tpm (all LSqCC cell lines)	LSqCC cell lines
HAS3	0.5	55.0	4.0	55.5	22
NTS	0.2	7.0	0.3	14.8	18
SOX2	50.0	229.0	25.0	136.5	19
TP63	-	483.0	3.0	61.2	18

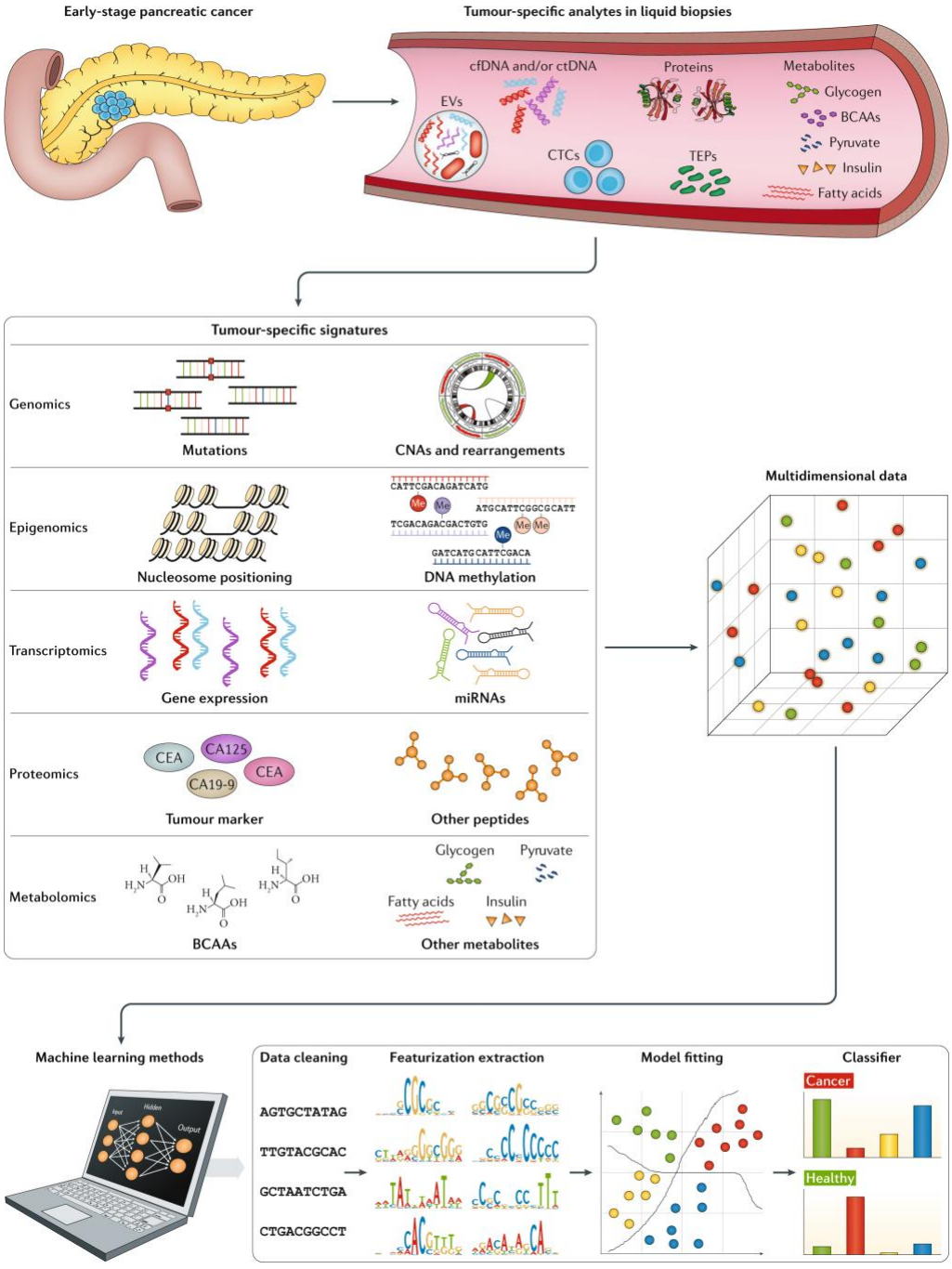
***Supplementary Table 9**. HNSqCC candidate genes' expression data in selected HNSqCC cell lines.* An average expression value was obtained using expression data of the genes of interest in several HNSqCC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). HNSqCC: Head and Neck Squamous Cell Carcinoma; tpm: transcripts per million.

Cell line	FaDu expression (tpm)	HSC-2 expression (tpm)	HSC-3 expression (tpm)	Average expression in tpm (all HNSqCC cell lines)	HNSqCC cell lines
CDH3	148.0	115.0	154.0	267.6	13
KRT16	37.0	8.0	2.0	270.2	13
LAMC2	52.0	118.0	1245.0	513.5	13
MMP10	17.0	13.0	24.0	74.9	13
PI3	37.0	100.0	2.0	403.3	13
PTHLH	2.0	74.0	241.0	61.2	13
SLC2A1	331.0	145.0	186.0	196.5	13

***Supplementary Table 10**. OC candidate genes' expression data in selected OC cell lines.* An average expression value was obtained using expression data of the genes of interest in several OC cell lines. The number of cell lines for which candidate genes' expression data was available is also indicated. Data retrieved from the Cancer Cell Line Encyclopedia, EMBL-EBI (https://www.ebi.ac.uk/gxa). OC: Oesophageal Carcinoma; tpm: transcripts per million.

Cell line	OE19 expression (tpm)	TE-1 expression (tpm)	KYSE-30 expression (tpm)	Average expression in tpm (all OC cell lines)	OC cell lines
CYP2S1	201.0	13.0	36.0	45.8	25
MALAT1	166.0	113.0	590.0	242.7	25
PHC3	8.0	14.0	11.0	15.2	25
PLEC	113.0	309.0	260.0	132.3	25

7.2. Appendix II. Supplementary figures.



Supplementary Figure 1. *Combination strategies for early detection of cancer from liquid biopsy samples.* Various tumour specific circulating analytes yield different information about the genome (mutations, copy number alterations, etc.), the epigenome, the proteome, the transcriptome or the metabolome. This data is to be combined in innovative ways and used for machine learning purposes. The machine learning workflow comprises the four steps shown in the figure and allows for distinction between tumour and normal states. BCAAs: Branched-Chain Amino Acids; cfDNA: circulating free DNA; CNAs: Copy Number Alterations; CTCs: Circulating Tumour Cells; ctDNA: circulating tumour DNA; EVs: Extracellular Vesicles; TEPs: Tumour-Educated Platelets. Retrieved from Heitzer *et al.*, 2019.