Daniel José Formica Pestana

Epigenetic hotspots in cancer

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2020

Daniel José Formica Pestana

Epigenetic hotspots in cancer

Master in Oncobiology - Molecular Mechanisms of Cancer

This work was done under the supervision of:

Pedro Castelo-Branco, Ph.D

Ana Marreiros, Ph.D

UNIVERSIDADE DO ALGARVE

Departamento de Ciências Biomédicas e Medicina

2020

Epigenetic hotspots in cancer

Declaração de autoria do trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

"I declare that I am the author of this work, that is original and unpublished. Authors and works consulted are properly cited in the text and included in the list of references."

_____

(Daniel Pestana)

“If someone can prove me wrong and show me my mistake in any thought or action, I shall gladly change. I seek the truth, which never harmed anyone: the harm is to persist in one's own self-deception and ignorance.”

- Marcus Aurelius, Meditations

Agradecimentos

A realização desta jornada foi possível devido a um conjunto de importantes pessoas, às quais deixo o meu profundo agradecimento.

Ao Professor Pedro Castelo-Branco, pela sua confiança e por me ter continuamente incentivado a expandir horizontes e ultrapassar limites.

À Professora Ana Marreiros, por ter sido mais que uma orientadora. Por se ter tornado uma amiga sempre disponível que irradia felicidade e transmite sempre o lado mais positivo da vida.

À Professora Alexandra Binnie, pela sua simpatia e boa vontade em prestar apoio sempre que foi necessário.

À Professora Mónica Fernandes, pela constante disponibilidade para discutir e melhorar este trabalho.

À Cândida Cardoso, por ter sido uma verdadeira fonte de apoio incondicional. Por me ter sempre incentivado a progredir, e por dar um brilho especial à minha vida.

Ao Ricardo Pestana, por ser, e sempre ter sido, um amigo de confiança, independentemente da distância.

Por fim, o meu mais profundo agradecimento vai para as duas pessoas que tiveram, e têm, mais importância na minha vida.

Ao meu pai, que me transmitiu princípios, por depositar em mim a sua fé incondicional. Por ter sido sempre um amigo e um exemplo, mas acima de tudo, por me ter ensinado a ser o meu próprio exemplo.

À minha mãe, que me transmitiu virtudes, pelo seu constante apoio. Por me ter, desde que me lembro, incentivado a romper limites. Por ser uma verdadeira amiga e confidente.

Abstract

DNA methylation is one of the most studied epigenetic events. In normal cells, it assures the regulation of gene expression without changing the genetic code. However, alterations in DNA methylation are now widely recognized as a contributing factor in tumorigenesis.

The bulk of research done in cancer epigenetics focuses on one of two events: promoter hypermethylation and global hypomethylation. Advances in the understanding of how DNA methylation shapes the chromatin's organization and how the later affects gene expression have been made. Less is known about how DNA methylation affects genes not only locally but also at a distance.

We hypothesized that during tumorigenesis specific genomic regions are more susceptible to DNA methylation (epi-hotspots) and other are resistance to DNA methylation changes (epi-blackholes). We also hypothesized that these regions might persist in tumor cells by exerting some selective pressure in the primary tumor clones.

By performing a pan-cancer analysis comparing normal to stage-I primary stage-I primary tumor samples gathered from TCGA consortium, we observed that both epi-hotspots and epi-blackholes occurred in all of the analyzed cancer cohorts. Furthermore, generally, epi-hotspots were able to predict gene expression alterations during tumorigenesis, and epi-blackholes were predictors of maintenance of gene expression during tumor initiation, which was in accordance with our hypothesis.

We also found that several epi-hotspots and epi-blackholes are predictors of survival in stage-III tumor patients, which may provide potential study targets for candidate prognostic biomarkers.

In summary, this study provides new evidence that regional methylation patterns potentially might exert selective pressure in tumor initiation by influencing genome-wide gene expression, and that these traits might be used to develop novel diagnostic and prognostic candidate biomarkers.

Keywords: DNA methylation, Cancer, Gene Expression, Prognostic Biomarkers, DNA regions

Resumo

O cancro é um conjunto heterogéneo de várias doenças que são caracterizadas por uma taxa de crescimento e divisão celular anormais. Durante o processo tumorigénico, as células tumorais vão sucessivamente adquirindo alterações genéticas e epigenéticas, o que leva a uma continua seleção de subclones tumorais. Durante esta evolução tumoral, as células sofrem alterações a nível da metilação de DNA que, tal como as mutações, podem ser propagadas para as células-filha. Estes tipos de alterações contribuem não só para o início do processo tumorigénico, como também para o seu continuo desenvolvimento, sem alterarem a sequência de DNA.

Alterações a nível da metilação de DNA participam no processo tumorigénico influenciando diretamente a expressão génica, e afetando a conformação da cromatina que, por sua vez, está relacionada com a toda a expressão génica na célula. Para serem ativamente expressos, os genes têm de estar acessíveis a fatores regulatórios. Por outro lado, genes que têm a sua expressão silenciada tendem a estar compactados na cromatina, de forma a estarem inacessíveis às proteínas responsáveis pela sua transcrição. Alterações a nível da conformação da cromatina podem promover a tumorigénese pelo facto de mudarem a acessibilidade de certas regiões de DNA, assim alterando o padrão global de expressão génica da célula.

A maioria dos tumores apresenta um padrão de metilação de DNA global anormal. Uma vez que estes mesmos padrões têm um papel importante na modulação da acessibilidade da cromatina, que por sua vez tem impacto no fenótipo da célula, surgiu a pergunta biológica: " durante o processo de iniciação tumoral, serão certas regiões genómicas mais suscetíveis a alterações a nível de metilação de DNA?". E no caso da resposta a esta pergunta ser afirmativa, surge ainda a questão: "Será que estas regiões estão associadas à alteração de padrões de expressão génica nas células tumorais?".

No caso de existirem zonas genómicas de maior suscetibilidade a alterações de metilação de DNA, e estas estarem associadas a alterações a nível de expressão génica, surge ainda a hipótese que estas regiões poderão ter valor de prognóstico em pacientes com doença avançada.

Numa tentativa de respondermos a estas questões, realizámos uma análise a doze tipos de cancro (adenocarcinoma do colon, adenocarcinoma do pâncreas, carcinoma da mama, colangiocarcinoma, carcinoma do esófago, cancro da cabeça e pescoço, carcinoma de células

renais de células claras, carcinoma de células renais papilar, carcinoma hepatocelular, adenocarcinoma do pulmão, carcinoma do pulmão de células escamosas, e carcinoma da tireoide), onde comparámos dados de metilação entre amostras de tecido normal com amostras de tecido tumoral em estádio I. Por forma a se encontrarem regiões de maior suscetibilidade a alterações de metilação de DNA, aplicámos dois algoritmos de identificação de regiões diferencialmente metiladas e intercetámos os resultados. As regiões genómicas identificadas por ambos os métodos foram designadas epi-hotspots. De modo a aferir se os epi-hotspots estavam associados a alterações de expressão génica no processo de iniciação tumoral, efetuou-se ainda uma análise de regressão linear múltipla entre cada gene diferencialmente expresso em estádio I e cada epi-hotspot. Os genes diferencialmente expressos, cuja variação entre tecido normal e tumor estádio I podia ser explicada por epi-hotspots, foram sujeitos a um estudo de ontologia genética, por forma a se compreender se estes genes potencialmente epigeneticamente regulados enriqueciam algum processo celular.

Este processo foi também repetido por forma a se identificarem regiões de baixa suscetibilidade a alterações de metilação de DNA, que designámos epi-blackholes. De modo a testar se estas regiões estavam associadas a genes não-diferencialmente expressos, realizou-se ainda uma análise de regressão linear múltipla entre cada gene não-diferencialmente expresso em estádio I e cada epi-blackhole.

Estudou-se ainda o grau de semelhança entre os doze tipos de cancro aqui analisádos relativamente à presença de epi-hotspots e epi-blackholes por meio de uma análise de agrupamento hierárquico.

Finalmente, examinou-se o potencial de prognóstico de cada epi-hotspot e cada epi-blackhole em pacientes tumorais de estádio III, fazendo uso de uma análise baseada em regressão multivariada de Cox.

Os nossos resultados indicam que, apesar de existirem pequenas semelhanças, o número e localização de epi-hospots e epi-blackholes é característico de cada tipo de cancro, o que sugere que tanto a alteração como a manutenção dos padrões de metilação nestas regiões dependem da célula de origem.

Verificou-se ainda que os padrões de metilação em epi-hotspots estavam associados a padrões alterados de expressão génica, em amostras de tecido tumoral em estádio I. Este resultado suporta a hipótese de que alterações regionais de metilação de DNA podem conferir vantagem seletiva na evolução clonal do tumor por influenciarem a expressão génica. De uma

forma geral, os genes cuja variação em iniciação tumoral era explicada pela variação da metilação de epi-hotspots enriquecem processos celulares de forma distinta nos diferentes tipos de cancro analisados.

Por outro lado, padrões de metilação em epi-blackholes estavam associados à manutenção dos padrões de expressão génica, em amostras de tecido tumoral em estádio I, o que sugere que a conservação de padrões de metilação em certas regiões do DNA pode também ser relevante para a tumorigénese.

Observou-se também que em dois terços dos tipos de cancro analisados, a metilação das CpGs de pelo menos um epi-hotspot ou epi-blackhole foi capaz de dividir os pacientes oncológicos de estádio III em dois grupos com padrões de sobrevida distintos, independentemente da idade dos pacientes. Apesar de nem todas as regiões aqui identificadas terem demonstrado potencial de prognóstico, este estudo sugere que os padrões de metilação de DNA em epi-hotspots e epi-blackholes podem ser potênciais candidatos para biomarcadores de prognóstico em pacientes oncológicos de estádio III.

Em suma, este trabalho demonstra que, durante o processo de iniciação tumoral, há uma alteração do padrão de metilação de DNA de certas regiões genómicas (epi-hotspots). Por outro lado, parece também haver uma conservação do padrão de metilação de DNA de outras regiões (epi-blackholes). Além disso, parece existir uma associação entre a variação da expressão génica na iniciação tumoral e a metilação dos epi-hotspots. A manutenção do padrão de metilação dos epi-blackholes identificados parece estar associada com a ausência de variação de expressão de determinados genes. Este estudo revela ainda que epi-hotspots e epi-blackholes podem ter também exercer uma pressão seletiva no tumor, já que para além de estarem associados à expressão génica são ainda capazes de prever o prognóstico de pacientes oncológicos em estádio III.

Palavras-chave: Metilação de DNA, Cancro, Expressão Génica, Biomarcadores de Prognóstico, Regiões genómicas

Table of Contents

Index of Figures

Index of Tables

Table of Annexes

Annex XIX Epi-hotspot-related differentially expressed genes between Normal and Stage-I LIHC tumor samples.

Annex XX Epi-hotspot-related differentially expressed genes between Normal and Stage-I LUSC tumor samples.

Annex XXI Epi-hotspot-related differentially expressed genes between Normal and Stage-I THCA tumor samples.

Annex XXII Epi-hotspot-related differentially expressed genes between Normal and Stage-I KIRP tumor samples.

Annex XXIII Enriched gene sets between Normal and Stage-I COAD tumor samples.

Annex XXIV Enriched gene sets between Normal and Stage-I BRCA tumor samples.

Annex XXV Enriched gene sets between Normal and Stage-I CHOL tumor samples.

Annex XXVI Enriched gene sets between Normal and Stage-I ESCA tumor samples.

Annex XXVII Enriched gene sets between Normal and Stage-I HNSC tumor samples.

Annex XXVIII Enriched gene sets between Normal and Stage-I LIHC tumor samples.

Annex XXIX Enriched gene sets between Normal and Stage-I LUSC tumor samples.

Annex XXX Enriched gene sets between Normal and Stage-I THCA tumor samples.

Annex XXXI Identified epi-blackholes in the COAD dataset.

Annex XXXII Identified epi-blackholes in the BRCA dataset.

Annex XXXIII Identified epi-blackholes in the CHOL dataset.

Annex XXXIV Identified epi-blackholes in the ESCA dataset.

Annex XXXV Identified epi-blackholes in the HNSC dataset.

Annex XXXVI Identified epi-blackholes in the LIHC dataset.

Annex XXXVII Identified epi-blackholes in the LUSC dataset.

Annex XXXVIII Identified epi-blackholes in the THCA dataset.

Annex XXXIX Identified epi-blackholes in the KIRP dataset.

Annex XL Identified epi-blackholes in the KIRC dataset.

Annex XLI Identified epi-blackholes in the PAAD dataset.

Annex XLII Identified epi-blackholes in the LUAD dataset.

Annex XLIII Epi-blackhole-related non differentially expressed genes between Normal and Stage-I COAD tumor samples.

Annex XLIV Epi-blackhole-related non differentially expressed genes between Normal and Stage-I CHOL tumor samples.

Annex XLV Epi-blackhole-related non differentially expressed genes between Normal and Stage-I ESCA tumor samples.

Annex XLVI Epi-blackhole-related non differentially expressed genes between Normal and Stage-I HNSC tumor samples.

Annex XLVII Epi-blackhole-related non differentially expressed genes between Normal and Stage-I LUSC tumor samples.

Annex XLVIII Epi-blackhole-related non differentially expressed genes between Normal and Stage-I THCA tumor samples.

Annex XLIX Epi-blackhole-related non differentially expressed genes between Normal and Stage-I PAAD tumor samples.

Annex L Epi-blackhole-related non differentially expressed genes between Normal and Stage-I LUAD tumor samples.

Annex LI Epi-hotspots and epi-blackholes as candidate prognostic biomarkers in stage-III COAD patients.

Annex LII Epi-hotspot as a candidate prognostic biomarker in stage-III BRCA patients.

Annex LIII Epi-blackholes as candidate prognostic biomarkers in stage-III ESCA patients.

Annex LIV Epi-hotspots and epi-blackholes as candidate prognostic biomarkers in stage-III HNSC patients.

Annex LV Epi-hotspot as a candidate prognostic biomarker in stage-III LIHC patients.

Annex LVI Epi-hotspots as a candidate prognostic biomarker in stage-III LUSC patients.

Annex LVII Epi-blackholes as candidate prognostic biomarkers in stage-III KIRP patients.

Annex LVIII Epi-hotspots and epi-blackholes as candidate prognostic biomarkers in stage-III LUAD patients.

Annex LIX Genomic location of the common epi-hotspots in each epi-hotspot cluster.

Annex LX Genomic location of the common epi-blackholes in each epi-blackhole cluster.

Abbreviations

ADD ATRX-DNMT3-DNMT3L

ADP Adenosine diphosphate

AGO Argonaute

ALL Acute lymphoblastic leukemia

AML Acute myelogenous leukemia

c**AMP** Cyclic adenosine monophosphate

APC Adenomatous polyposis coli

ASC Adipose stem cells

ATP Adenosine triphosphate

ATRX X-linked helicase II

AUC Area under the ROC curve

BIK BCL2 Interacting Killer

BRCA Breast Invasive Carcinoma

CAF Cancer-associated fibroblasts

CBP CREB-Binding Protein

CCR-NOT Carbon Catabolite Repression—Negative On TATA-less

CG Guanine-Cytosine

CHOL Cholangiocarcinoma

CIMP CpG island methylator phenotype

CLL Chronic lymphoblastic leukemia

CML Chronic myelogenous leukemia

CNS Central nervous system

COAD Colon Adenocarcinoma

CREB cAMP-response element binding protein

CSC Cancer stem-cells

CVD Cardiovascular diseases

CXCL chemokine (C-X-C motif) ligand

DCP Decapping protein

DMAP DNA methyltransferase associated protein

DMR Differentially Methylated Region

DNA Deoxyribonucleic acid

DNMT DNA methyltransferases

ECM Extracellular Matrix

EHMT Euchromatic histone lysine methyltransferase

EMT Epithelial-mesenchymal transition

EMX Empty Spiracles Homeobox 2

ER Estrogen receptor

ESC Embryonic stem cells

ESCA Esophageal Carcinoma

EZH Enhancer of zeste homolog

FAS Fas Cell Surface Death Receptor

FDA Food and Drug Administration

FDR False discovery rate

FGF Fibroblast growth factor

GAGE Generally Applicable Gene-set Enrichment

GCT Germ cell tumors

GDC Genomic Data Commons

GNAT Gcn5-related N-acetyltransferases

GO Gene Ontology

GSA Gene set analysis

HAT Histone acetyltransferase

HDAC Histone deacetylase

HER Human epidermal growth factor receptor

HGF Hematopoietic growth factor

HNSC Head and Neck Squamous Cell Carcinoma

HR Hazard ratio

IARC International Agency for Research on Cancer

IDE Integrated development environment

IFN Interferon

IL Interleukin

IQR Interquartile range

KICH Kidney chromophobe

KIRC Renal Clear Cell Carcinoma

KIRP Renal Papillary Cell Carcinoma

LIAC Leading invasive anchor cell

LIHC Hepatocellular Carcinoma

LINE Long interspersed nuclear element 1 retrotransposon

LNCAP Lymph node carcinomas of the prostate

LUAD Lung Adenocarcinoma

LUSC Lung Squamous Cell Carcinoma

MCC Matthews correlation coefficient

MLH mutL homolog

MMP Metalloproteinases

MRE miRNA response elements

MSH MutS homolog 2

mTOR Mechanistic target of rapamycin

NCD Noncommunicable diseases

NCI National Cancer Institute

NGS Next generation sequencing

NHGRI National Human Genome Research Institute

NIH National Institutes of Health

NK Natural killer

NKT Natural killer T

NOE Number of Epi-hotspots

PAAD Pancreatic Adenocarcinoma

PDGF Platelet-derived growth factor

PR Progesterone receptor

PRC Polycomb repressive complex

PRMT Protein arginine N-methyltransferase

RFTS Replication Foci Targeting Sequence

RISC RNA-induced Silencing Complex

RNA Ribonucleic acid

ROC Receiver operating characteristic

ROS Reactive oxygen species

SAM S-adenosyl methionine

SCLC Small-cell lung carcinoma

SD Standard deviation

TBP TATA box-binding protein

TCGA The Cancer Genome Atlas Consortium

TERT Telomerase Reverse Transcriptase

TGF transforming growth factor

THCA Thyroid Carcinoma

TIC Tumor-initiating cell

TNF Tumor Necrosis Factor

UHRF Ubiquitin-like plant homeodomain and RING finger domain

UTR Untranslated region

VEGF Vascular endothelial growth factor

XRN Exoribonuclease

Chapter 1 Introduction

1.1 Cancer

Cancer is a group of over 100 highly heterogeneous complex diseases that can commonly be characterized by abnormal cell growth and division, disregarding normal cellular restraints [1,2]. This type of disease originates from abnormal cells that grow and proliferate indefinitely, giving rise to a neoplasm or tumor [1,2]. Although the terms cancer and tumor are frequently used interchangeably, a neoplasm can only be considered cancer if its cells exhibit malignancy, i.e., have the ability to invade surrounding and distant tissues [3]. Otherwise, the tumor is regarded as benign. In malignant neoplasms, the cancerous cells ultimately detach from the primary tumor and colonize other tissues in the body, forming secondary tumors known as metastases [3].

Cancers can be generally classified based on the cell type from which the primary neoplasm originates. The most common class of cancers arise from epithelial cells and are named carcinomas [4,5]. This set of cancers affect tissues that derive from all three embryonic germ layers, for example, epithelia from the lungs, liver, stomach, esophagus, gallbladder, and intestines, that derive from the endoderm (the most inner germ cell layer); ovaries that stem from the mesoderm (the middle germ cell layer); and skin that develops from the ectoderm (the outer germ cell layer) [5,6]. Generally, carcinomas can be further subdivided into two classifications that denote the biological function of the respective epithelial tissue. Mainly, epithelia can have one of two functions: 1) a protective function, like skin which protects the underlying cells from external agents, or 2) a secreting function, like endocrine glands, that secrete hormones into the bloodstream, or exocrine glands, that release secretions onto an epithelial sheet, through a duct. Cancers that stem from epithelial cells with a protective function are named squamous cell carcinomas, whereas the ones that arise from epithelial cells with a secreting function are named adenocarcinomas [5,6].

Cancers that do not originate from epithelial tissues are commonly referred to as nonepithelial cancers [7–12]. This class comprises three major subgroups: sarcomas, hematopoietic cancers, and neuroectodermal cancers.

Sarcomas originate from mesenchymal cells, such as fibroblasts, adipocytes, osteoblasts, myocytes, and even endothelial cell precursors [7,8]. Contrary to carcinomas, cells and tissues that give rise to sarcomas have a common origin in the embryonic mesoderm [7,8].

The second category of nonepithelial cancers is the hematopoietic malignancies group, frequently referred to as liquid tumors [9,10]. These types of malignancies are neoplastic tumors that affect various cell types from the blood and lymphatic systems, like lymphatic and leukocytic cells. Depending on which type of progenitors the affected cells derive from, the malignancy is classified as lymphoma, leukemia, or myeloma, being the first the most frequent hematological malignancy and the later the least frequent. Since hematopoiesis is a process involving many different cell types, these diseases can be further subclassified based on which type of cell is affected. For instance, leukemias can be clustered into acute lymphoblastic leukemia (ALL), acute myelogenous leukemia (AML), chronic lymphoblastic leukemia (CLL) chronic myelogenous leukemia (CML), and others. A particularity of hematopoietic malignancies is the common presence of chromosomal translocations, which is not as frequent in solid tumors [9,10].

The last nonepithelial cancer category is the neuroectodermal cancer subgroup [11,12]. This type of cancer arises from cells that comprise the central and peripheral nervous systems. As the group's designation suggests, the cells that originate these kinds of cancer stem from the embryonic ectoderm. Depending on the type of the nerve cell, different neuroectodermal cancers can arise, like glioblastomas, neuroblastomas, schwannomas, medulloblastomas, or gliomas[11,12]. Primary central nervous system (CNS) tumors are of particular importance since they remain among the most challenging cancers to treat, with a high mortality rate [4,11,12].

Although this classification system is able to roughly divide tumor types, there are some tumors that are hard to assign to a major classification. These include, for example, the small-cell lung carcinomas (SCLCs), melanomas, and teratomas [13–15]. SCLCs are constituted by cells that pose neurosecretory characteristics, and it is not clear whether these tumors stem from neuroectodermal cells that moved into the developing lung, or if they stem from endodermal cell populations of the lung that transdifferentiated, losing some of their epithelial attributes and acquiring neuroectodermal ones [15]. Melanomas are also a type of tumor challenging to fit into a major classification group [13]. They arise from melanocytes, which stem from the embryonic neural crest, and during development migrate to the skin and eye, having an origin close to that of neuroectodermal cells [13]. Another type of tumors hard to place into

one of the main classification clusters are the teratomas [14]. Teratomas are one of the most common forms of germ cell tumors (GCTs), which include not only teratomas, but also seminomas, choriocarcinomas, yolk sac tumors, embryonal cell carcinomas, and mixed GCTs [16]. GCTs arise from primordial germ cells, which become incorporated into the fetal gonads, being able to occur in the gonadal tissues themselves or in the path of primordial germ cell migration[16]. Interestingly, teratomas are histologically defined has having tissues derived from all three germ cell layers: ectoderm, mesoderm, and endoderm [14].

1.1.1 Cancer as a multistep microevolutionary process

In 1976, a reference article, published by Peter Nowell, proposed cancer development to be a stepwise Darwinian evolutionary process, characterized by the occurrence of mutations in somatic cells, and posterior subclonal selection [17] (Fig 1.1). Although cancer is mainly composed of cells that are a part of a larger organism, this perspective views cancer clones as unicellular quasi-species that reproduce asexually [18,19]. Cancer clones, and even normal cells, are a part of an ecosystem where its inner components collaborate in order to optimize the function of the whole organism. In contrast to typical ecosystems, in an organism, competition between individual cells is limited, being every somatic cell committed to eventually die, in favor of the organism's, or its progeny's survival. In fact, most multicellular organisms, especially long-lived animals like humans, have evolved to restrict clonal expansion of cells that contain certain renegade-traits such as continuous cellular self-renewal, extensive proliferation, and even unforeseen cell migration and invasion capabilities [18,19]. It is this restriction that causes cancer development to be a long process, during which the cells from

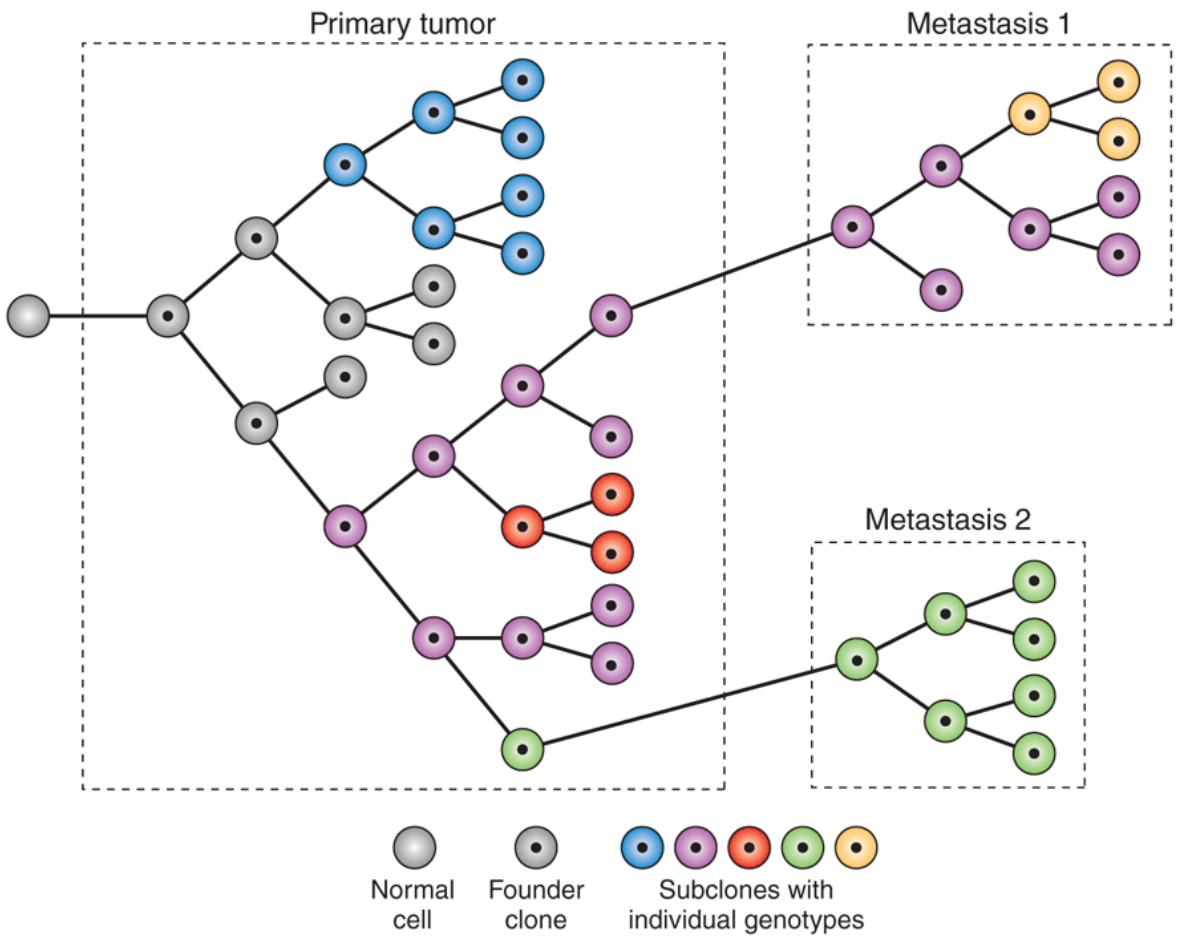

Figure 1.1 – Model for tumor clonal evolution. The entire population of tumor cells descends from a founder clone. The tumor cells acquire successive somatic mutations, which leads to a continuous selection of tumor subclones (here depicted in different colors). The primary tumor is not formed by a single clone, but rather several ones. Some of these clones can cease to exist, remain dormant, or expand. Additionally, metastases can stem from either minor primary tumor clones (metastasis 2) or major primary tumor clones (metastasis 1). Reprinted by permission from Springer Nature: Nature Biotechnology, Cancer sequencing unravels clonal evolution (19), Copyright (2012).

which the cancer arises randomly and sequentially acquire different mutations, some providing phenotypic bypasses to the constraints imposed by the micro-environment [18,20]. Some of the main micro-environmental restrictions include its structure, and limitation of resources, making tumoral natural selection a process based on cellular competition for space and resources. This tumoral clone microevolution is based on the acquisition of deoxyribonucleic acid (DNA) mutations and micro-environmental changes that alter the fitness impact of those DNA mutations.

Mutations can have different effects on the evolution of the clone that possesses it. These can be considered driver mutations that provide a selective advantage over the remaining

clones, passenger mutations which are selectively neutral, and deleterious mutations that are selectively negative. There is also a fourth type of mutation involved in the microevolutionary process of cancer, referred to as the mutator mutations. This type of lesion greatly increases the rate of other genetic alterations. Although not a mutation, another type of cellular alterations that influence tumor evolution are epigenetic changes, these alterations do not directly change the DNA sequence; however, impact gene expression. Interestingly, the rate of epigenetic alterations is several orders of magnitude higher than the ones of a genetic nature, possibly being a major determinant of clonal evolution [18,20].

It may be intuitive to think that a very fit tumor clone would largely expand and dominate the neoplasm. However, large clonal expansions after cellular transformation are very rare [18,21,22]. In fact, it seems that, originally, parallel clonal expansions occur, and only posteriorly subclones start to dominate in early cancer development [18,21,22].

The clonal fitness provided by genetic and epigenetic alterations is strongly dependent on the complex and dynamic cellular microenvironment [2,18,20].

1.1.2 Tumor microenvironment and its components

Although frequently, the microenvironment, or niche, where the tumor cells reside is often referred to as a cellular microenvironment, it is not comprised solely by cells [20]. The tumor microenvironment is constituted by cellular components, not all tumoral, such as fibroblasts, neuroendocrine cells, adipose cells, immune cells, myoepithelial cells, stromal cells, and endothelial cells, and non-cellular components like extracellular matrix (Fig. 1.2). Each microenvironmental component plays a different role in cancer development, and many of them acquire non-malignant phenotypic alterations associated with cancer [20].

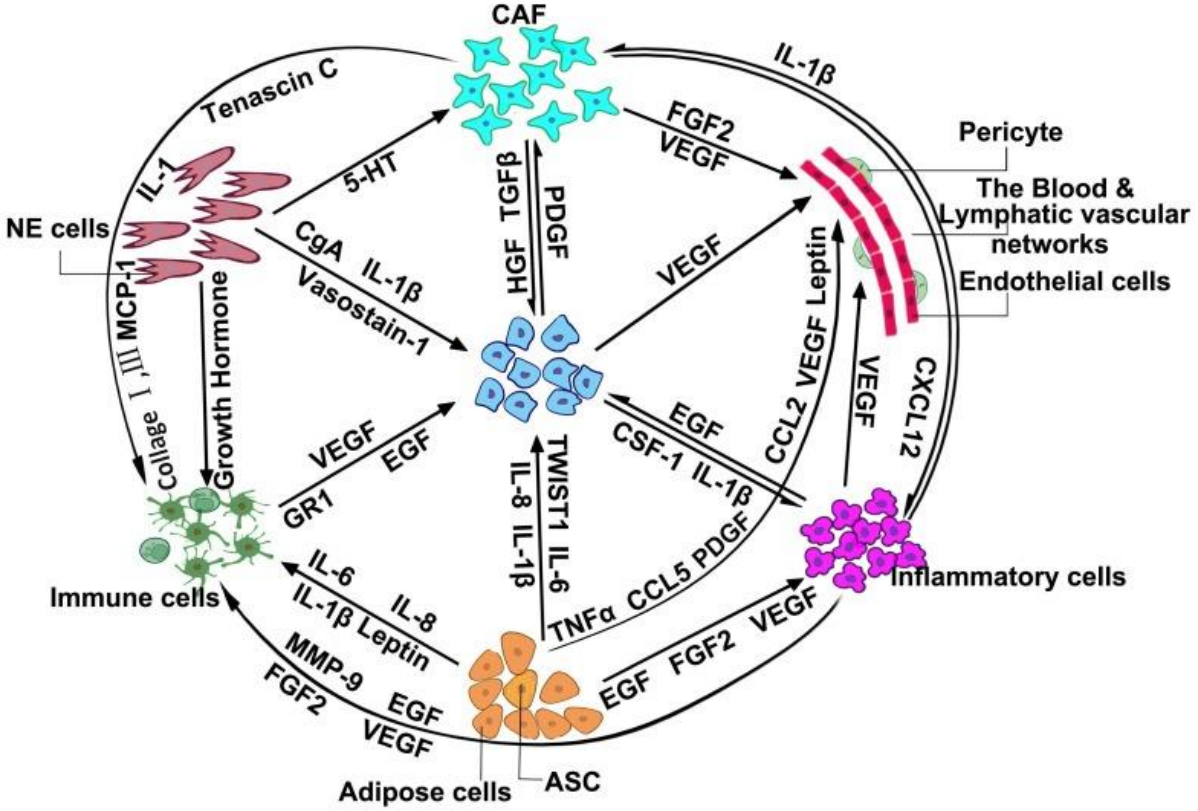

Figure 1.2 – Tumor cell's interaction with the microenvironment. The primary tumor does not solely comprise of tumor cells. It forms a communication network of different cell types that influence each other by emitting and receiving several mediators that impact the entire microenvironment. Reprinted under a Creative Commons Attribution-Noncommercial 4.0 International Public License, from Ivyspring International Publisher: Journal of Cancer, Role of tumor microenvironment in tumorigenesis (2017) [20].

1.1.2.1 Cancer-Associated Fibroblasts

Fibroblasts are a main component of the tumor microenvironment, and in this specific context, assimilate a myofibroblastic phenotype, often being referred to as cancer-associated fibroblasts (CAFs) [20,23]. The activation of fibroblasts to CAFs is parallel to the myofibroblastic activation during natural wound healing. The key distinction is that while myofibroblasts are transiently present in normal wound repair, CAFs remain permanently activated at the tumor site, much like in tissue fibrosis. There are several microenvironmental players that are able to induce fibroblastic activation. These include adhesion molecules contacting leukocytes, reactive oxygen species (ROS), micro Ribonucleic acid (miRNAs), cell-cell communication, and growth factors. After being fully activated, CAFs play a major role in cancer progression, mainly by contributing to inflammatory cell recruitment, stimulating angiogenesis, and remodeling the extracellular matrix (ECM). CAFs can also directly induce neoplastic cell

proliferation by secreting growth factors and immune-suppressive cytokines, and even with mesenchymal-epithelial cell interactions. Not only do CAFs contribute to tumor development, but they also support tumor invasion and metastasis. For example, proteins as chemokine (C-X-C motif) ligand 12 (CXCL12) and interleukin-22 (IL-22) are strongly overexpressed in CAFs and are thought to induce epithelial-mesenchymal transition (EMT) of certain types of cancer types, like gastric and prostate cancers. Other proteins ubiquitously expressed by myofibroblasts and CAFs, which may be an important contributing factor to tumor progression, are hematopoietic growth factor (HGF), transforming growth factor-beta 1 (TGF-β1), platelet-derived growth factor (PDGF), amongst others. CAFs also seem to have a role in tumor progression by promoting angiogenesis, through the secretion of pro-angiogenic factors like fibroblast growth factor 2 (FGF2), vascular endothelial growth factor (VEGF), and others [20,23].

1.1.2.2 Immune System cells

Another crucial cellular component of the tumor microenvironment are immune and inflammatory cells [20,24]. It is hypothesized that cells in a mammalian organism are permanently examined by the immune system [25,26]. This constant immune surveillance is able to recognize and eliminate most of the newly transformed tumor cells [25,26]. It is, then, as proposed by Hanahan and Weinberg, a necessity that tumoral cells manage to avoid detection by the immune system and subsequent immunological killing [2].

Although the normal function of the immune system would be to prevent tumor formation and control tumor outgrowth, it can also facilitate cellular transformation [26,27]. Research has shown that the immune system can both prevent and promote cancer, in a process commonly referred as 'Cancer Immunoediting'. This dynamic process is divided in three major phases: elimination, equilibrium, and escape. During elimination, the first immunoediting stage, is where immunosurveillance take place [26,27]. The immune system actively eliminates nascent tumors [26–29]. Initial tumor growth promotes inflammatory signals that induce the activation of the innate immune system, stimulating the recruitment of Natural killer T (NKT), Natural killer (NK), γδ T cells, macrophages and dendritic cells. Tumor infiltrating lymphocytes recognize structures on the transformed cells and produce Interferon (IFN)-γ, which activates antiproliferative and pro-apoptotic mechanisms. Even though the IFN-γ itself can induce some amount of tumor cell death, it also stimulates the release of the chemokines CXCL10, CXCL9, and CXCL11, by the tumor cells and normal cells present in the

microenvironment. These chemokines possess a strong anti-angiogenic effect, halting the formation of new intra-tumoral blood vessels, leading to even more tumor cell elimination. This inflammatory process is an escalating one, where the cytokines produced recruit more NK cells and macrophages to the tumor microenvironment. These two types of tumor-infiltrating immune cells produce IFN-γ and IL-12, transactivating each other, and thus contributing even more to tumoral elimination. The cellular debris produced by cellular elimination is ingested by local dendritic cells, that later migrate to draining lymph nodes and activate CD4+ T_H1 helper cells, which in turn promote the development of tumor specific CD8+ T cells. CD4+ and CD8+ T cells then return to the tumor site, killing the remaining immunogenic tumor cells. After all tumor cells with enough immunogenicity have been eliminated, the second stage of immunoediting begins: the equilibrium stage. In this phase, the immune system serves as a potent selective pressure to tumor cells, that although is not sufficient to eradicate the tumor, can restrain its development. This period, although long, is critical for tumor microevolution, where clones arise with new mutations that confer them resistance to the immune system. The final immunoediting stage is the escape process. The clones that survived the elimination process, and underwent the equilibrium process have a low immunogenicity, and thus are able to escape immunologic detection and elimination. In this way, the immune system serves as a selective pressure to tumor clonal evolution, and cancer development. In this context, tumor clones with survival advantage will eventually dominate the neoplasm [26–29].

1.1.2.3 Cells from the circulatory and lymphatic systems

The large network formed by the circulatory and lymphatic systems are of key importance to cells and tissues all around the organism, providing sustainable nutrition and oxygen to local cells and by removing carbon dioxide and other metabolic debris [20,30]. This is no different in a tumor context, as the neoplastic cells have the same vascular need than normal cells. However, there are some differences that distinguish the vasculature of a normal tissue than that of a tumoral context. One of these differences is that the angiogenic process, in a tumorigenic process, is not only almost always activated, but also remains continuously activated during the whole process, making the tumor environment highly pro-angiogenic. Even though the tumor site is characterized by an unceasing formation of new blood vessels, these are, in contrast to a normal context, usually leaky and inefficient.

Cells that constitute the primary tumor, that gradually increases in mass, will eventually find themselves in a low oxygen environment. To overcome this issue, tumor cells will have to not only adapt to a hypoxic setting, but also to recruit additional blood vessels, that increase the oxygen supply [20,30].

1.1.2.4 Adipocytes

Although adipose tissue is not usually associated with tumoral microenvironment, there is now evidence that establishes a link between them [20,31]. It has been shown, for example, that high quantities of adipocytes in a tissue, like in an obesity context, promotes tumor-site hypoxia, which leads to a tumor promoting proinflammatory state. Besides their hypoxia inducing abilities, adipocytes also secrete factors that may be involved in tumor initiation and/or progression. These include more than 50 cytokines, chemokines, and hormone-like factors.

An interesting type of adipose cells associated with tumor development and progression is the adipose stem cells (ASC), that have the ability to differentiate into several cell lineages. ASCs are powerful tumor promoters in several ways, being able to modulate other components of the tumoral microenvironment, inducing tumor-promoting inflammation, and even stimulate angiogenesis. One of the most interesting tumor-promoting properties ASCs possess is the ability to differentiate into cancer-associated cells. It has, for instance, been shown that ASCs, in some cancer models, can differentiate into CAFs that promote tumor proliferation [20,31].

1.1.2.5 Neuroendocrine cells

In normal situations, neuroendocrine cells have several regulating roles in different tissues, examples include cholecystokinin secreting cells, secretin-secreting S cells, gastric inhibitory polypeptide-secreting cells, motilin-secreting M cells and neurotensin-secreting N cells, all a type of neuroendocrine cells found in the small intestine named enteroendocrine cells [20,32]. Although neuroendocrine cells play an important role in the digestive track, these can also be found in glands or tissues like the hypothalamus, anterior pituitary gland, pineal gland, thyroid gland (calcitonin secreting cells), breast, thymus, and in the pancreas (islets of Langerhans) [20,32].

Neuroendocrine cells can also have a role in tumor formation and/or progression, mainly by having a substantial impact in the immune system [20]. The mechanisms that link neuroendocrine cells, the immune system and tumor formation are plentiful. For instance, substance P, a neuropeptide that acts as a neurotransmitter and neuromodulator of nociceptive afferents, is known to increase migratory activity of T lymphocytes by blocking its β1-integrin mediated adhesion, as well as inducing production of leukocytic cytokines. The catecholamine norepinephrine, also a neurotransmitter, can function as a production suppressor of the antitumoral cytotoxic T-lymphocytes, by inhibiting the synthesis of Tumor Necrosis Factor alpha (TNF-α) [20].

Interestingly, neuroendocrine cells are not only parallel players in tumor formation and growth but can also be the cells from which tumor originate [20,33]. In fact, there is a whole class of rare malignancies that stem from the hormone-producing cells of the neuroendocrine system, named neuroendocrine carcinomas [20,33]. For example, in 2004, R.J. Jin and colleagues, showed that androgen-dependent lymph node carcinomas of the prostate (LNCAP) were only able to develop in the presence of neuroendocrine tumors in castrating mice, thus attributing some specificity to these cells in tumor formation [33].

1.1.2.6 Extracellular Matrix (ECM)

As previously referred, the tumoral microenvironment is not comprised solely of cellular components [20]. The main non-cellular component being the extracellular matrix. Depending on the tissue, ECM consists of several components like collagens, laminins, fibronectins, proteoglycans, and hyaluronans, molded in a tissue-specific manner. It is the ECM that truly forms the tumoral microenvironment, containing all the growth factors, cytokines, and hormones secreted by the local tumor and non-tumor cells [20].

The ECM is present in any tissue and provides biochemical and structural support for the cells there located, and although it's a non-cellular component, is a physiologically active part of the tissue, making fundamental processes like cell-cell communication, cell adhesion, and cell proliferation, possible [20,34,35]. Being of such importance in these processes, and in maintaining the delicate tissue homeostasis, small changes in the ECM can have significant effects at the cellular level, being a key dynamic player not only in the regulation of healthy tissue homeostasis, but also in tumor progression.

During tumor growth, microenvironmental ECM continuously interacts with the tumor cells, and goes through active changes that are a crucial part of the tumor progression process. During this context, there is an increased secretion of fibronectin and collagen I, III, and IV. The augmented deposition of these matrix proteins interferes with biological processes like cell-cell adhesion, cell polarity, and growth factor signaling amplification. The exact role of collagen deposition in tumor progression is not yet completely understood [20,34,35]. Levental, and Karangiannis, with respective colleagues, showed, in 2009 and 2012, respectively, that collagen deposition and cross linking promotes tumor progression by increasing integrin signaling [36,37]. In contrast, it has also been shown by other groups, that depletion of fibrillar collagens I and III can also promote tumor progression, meaning that the effect of these matrix proteins in tumor development can be both beneficial and deleterious to the process [38,39].

One of the main events in tumor progression and cancer is the cellular migration, moving through surrounding tissues and penetrating the adjacent basement membrane [20,34,35]. The dense and highly cross-linked ECM in the microenvironment serves as a strong barrier to epithelial cell migration. One of the ways this physical barrier can be crossed is with the mechanical force generated by cellular proliferation. As cells proliferate continuously, the tumor increases in size and is increasingly spatially constrained by the basement membrane. This continual growth, despite the physical limits imposed by the environment's architecture, produces an ever-increasing mechanical stress across the membrane, eventually causing its rupture and allowing cells to escape the current location. There is also another way in which the tumor can breach the collagenous barrier, often denoted as anchor cell invasion. In this method some tumor cells, named leading invasive anchor cells (LIACs), play a key role by leading the membrane's traversing and opening way for the rest of the tumor cells to cross the basement membrane. Initially, LIACs extend an invadopodium – a protrusive, F-actin rich, subcellular "arm" – into the basement membrane, breaching the barrier. The membranal fissure, caused by the initial breaching, then widens, allowing the remaining tumor cells to cross the barrier.

It was long thought that the sole contributing factor to tumor cell membrane traversing were Metalloproteinases (MMPs), mainly due to the fact that there is an increased accumulation of these enzymes along the basement membrane. On the contrary, it has been shown that laminin and collagen IV are not fully degraded, rather mechanically pushed aside by the invadopodia, indicating that MMPs are more likely to play a role in the initial membrane breaching or by softening the matrix while LIACs allow the invasion. Although MMPs are not

a single player in tumor cell invasion, they play a major role in tumor progression and invasion, by contributing to the degradation of surrounding ECM barriers, and even by resealing growth factors that promote neo-angiogenesis. The ECM itself is a powerful proliferation ignitor, by directly contacting with the integrin family of cell surface receptors. The problem is that many ECM binding sites, crucial for cell survival and proliferation signals, are usually partially hidden by the ECM itself, often being referred to as "cryptic binding sites". It is the MMP's enzymatic activity that allows for the degradation of the surrounding collagen, which uncovers the binding sites and subsequently permits that the integrins in the cell membrane interact directly with the matrix.

The microenvironmental ECM also stores, embedded within collagen, various inactive tumor promoting growth factors that upon matrix degradation, mainly by MMPs, are released and activated. For example, during ECM degradation by MMPs, the active form of TGF-β is released to the tumor microenvironment, where it modulates several tumor-associated processes such as cell invasion, cell proliferation and immune response [20,34,35].

1.1.3 Cancer stem-cells and their niche

Many tissues in a grown human organism contain a specific population of adult stem-cells dedicated to continued self-renewal of the tissue where they reside [40]. In contrast to the other cells that form the tissues, adult stem-cells are long-lived and give rise to the short-lived, specialized cells that perform the tissue-specific functions [40].

It is hypothesized that tumor progression is, like many normal tissues, also driven by a dedicated subpopulation of stem cells, named cancer stem-cells (CSCs) [41,42]. This model is often known as the Cancer stem-cell theory, and it states that CSCs are the tumor cells with the main self-renewal proprieties, clonal tumor initiation capacity, and clonal long-term repopulation potential. The model also proposes that CSCs can transition between stem and non-stem cell states, in a reversible manner. Although these cells can evade cell death, and even metastasize, they can remain in a dormant state for long periods of time. The relevancy of the CSC theory is ever-increasing, since research has shown that CSCs are able to survive many current cancer therapeutics [41,42].

Similar to normal adult stem-cells, it is believed that CSCs also reside in specialized microenvironment, named niche [43]. Niches, in normal settings, are comprised of immune cells,

fibroblasts, endothelial cells, perivascular cells, and ECM [20]. In a tumoral context, the CSC niche is itself part of the tumor microenvironment [20,43].

1.1.3.1 Tumor development models and the role of CSC's and their niche

There are two main models that have been proposed to understand tumor progression and heterogeneity: the hierarchical and the stochastic models [42–45].

The hierarchical model proposes that CSC are the true malignant tumor-propagating cells, and that these are a biologically distinct subpopulation of cells within the total tumor cell population [42–45]. This model suggests that tumorigenesis only arises when a normal stem cell escapes regulation and becomes a stem-like-cell tumoral cell, a CSC. Since these cells have a very high self-renewal capacity, they are considered to be the unit of selection in a tumor. In contrast, all of the other tumoral non-CSCs lead to clonal exhaustion. The hierarchical model defends that the only possible way to eliminate clinical relapse, would be to completely eradicate CSC's, since these are the fuel for tumor growth [42–45].

It is important to recognize that CSCs, unlike normal stem cells, aren't truly multipotent, don't divide asymmetrically, and can only differentiate into one type of cell that can't generate an entire array of cellular lineages [42–45]. In fact, tumors, during their development and progression, tend to anatomically and functionally stray away from the original organ, possibly indicating that CSC's have deregulated self-renewal, proliferation and differentiation capabilities. This gap that separates CSC's from normal stem cells has led to the suggestion, from several proponents of the hierarchical model, that the term tumor-initiating cells (TICs) is more suited for the description of these tumorigenic cells. This recommendation has led to an indifferentiable use of both CSCs and TICs, often being implied that these are two terms to designate the same cell. However, the TIC refers to the cell-of-origin, the first abnormal tumoral cell that gives rise to the tumor, and not the subpopulation of cells within the tumor that sustain the tumoral growth and proliferation, i.e. the CSCs, which are not necessarily the cell-of-origin [42–45].

Although the hierarchical model defends that some cells are the fuel for the whole tumor, it does not exclude tumoral clonal evolution, asserting that one or several CSCs can generate different tumor clones, and that each clone evolves hierarchical with their own CSCs [42–45].

The second model that attempts to explain tumor progression and heterogeneity is named stochastic model [42–45]. This model proposes that every cell that constitutes a tumor is of equal likelihood to be the cell-of-origin and initiate the tumorigenic process. Here it is thought that cancer, being a hyperproliferative disease, evolves through the acquisition of genetic mutations in a sequential manner, that promotes subsequent clonal expansion. Contrarily to the hierarchical model, the stochastic model reasons that cellular transformation into tumor cells is largely explained by stochastically varying intrinsic factors, and only partially explained by the surrounding environment in which the cells reside [42–45].

There are tumor types that seem to fit this last model in an almost unflawed way, as is the case of some colorectal cancers, whose sequential progression was described by Vogelstein, in 1988 [46].

Although the hierarchical and stochastic models seem to be mutually exclusive and a theoretical dichotomy, the phenomenon of cellular phenotypic plasticity is able to merge both models into one [42–45]. There is, within the tumor, a subpopulation of tumoral cells that have the capacity to transit between a differentiated state and a stem-like state, i.e. cells phenotypically plastic. Depending on their genotype or on the environmental signals received by these cells, they can dedifferentiate and reenter into the CSC pool to repopulate the tumor [42–45].

This dedifferentiation ability can wither be inherited (hierarchical model) or acquired through random mutations (stochastic model) [42–45]. It has, in fact, been shown that certain genotypic changes can contribute to the cellular acquisition of phenotypic plasticity by tumor cells. For example, upregulation of NODAL, NOTCH, and WNT proteins, activation of human Telomerase Reverse Transcriptase (hTERT), and p53 inhibition, promotes phenotypic plasticity [42–45].

1.1.4 Cancer epidemiology

At the time of 2018, chronic diseases, also referred as noncommunicable diseases (NCDs) constitute the main cause of death worldwide [4]. Cancer is one of these diseases and is now considered the biggest hindrance to the increase in life expectancy in every country in the world [4].

The most impactful NCD, globally, are cardiovascular diseases (CVDs) [47,48]. In fact, on a yearly basis, more people die from CVDs than any other cause, representing the number one

cause of death worldwide [4,47,48]. However, it is expected that cancer will outrank CVDs and become the leading cause of death [4]. The reasons cancer is expected to surpass CVD's mortality rates are several but can be summed into two key points: 1) cancer incidence and mortality is rapidly rising in every country, and 2) deaths due to stroke and coronary heart disease are declining in many countries [4].

The steep rise in the number of cancer new cases and deaths across the globe reflect both aging and growth of the population, but also the prevalence of risk factors associated with socioeconomic development of populations [4]. For example, it is repeatedly observed that in new growing economies there is a shift in the pattern of cancer types that affect these populations from infection and poverty related cancers, such as cervix, stomach, and liver, to cancer types that frequently have a higher incidence in more developed populations, like Europe or North America [4].

1.1.4.1 Cancer incidence and mortality worldwide

It is estimated, by the International Agency for Research on Cancer (IARC), that there were, in 2018, 18.1 million new cancer cases and 9.6 million cancer deaths [4]. The most affected continent, regarding both incidence and mortality, is Asia, since, in 2018, half of all new cases (48.4%), and more than half (57.3%) of all deaths occurred in this continent. One of the main reasons that the cancer burden is higher in Asia is because this continent corresponds to 60% of the world population. The second continent most affected by cancer is Europe, in which 23.4% of the total new cancer cases and 20.3% of deaths occur. It is arguable that Europe may be the most affected continent by this disease, since it only represents 9% of the world population. Ranking third is the American continent which accounts for 21% and 14.4% of incidence and mortality globally [4].

Interestingly, the Asian and African continents, are the only ones in which the global shares of cancer mortality (57.3% for Asia, and 7.3% for Africa) are higher than the global shares of cancer incidence (48.4% for Asia, and 5.8% for Africa) [4].

According to 2018 data, the most frequent cancer type worldwide, for males and females combined, is lung cancer, accounting for 11.6% of the total new cases (~2.09 million new cases) [4]. Lung cancer is also the number one cause of cancer death, representing 18.4% of all cancer deaths (~1.76 million deaths). Closely following lung cancer is breast cancer, the

second most common cancer, which represents 11.6% (~2.09 million new cases) of all new cases and 6.6% of all cancer deaths (~ 0.63 million deaths). The third to tenth most common cancers are prostate (7.1% of all new cases and 3.8% of all deaths), colon (6.1% of all new cases and 5.8% of all deaths), nonmelanoma skin cancer (5.8% of all new cases and 0.7% of all deaths), stomach (5.7% of all new cases and 8.2% of all deaths), liver (4.7% of all new cases and 8.2% of all deaths), rectum (3.9% of all new cases and 3.2% of all deaths), esophagus (3.2% of all new cases and 5.3% of all deaths), and cervix uteri (3.2% of all new cases and 3.3% of all deaths) [4].

On the other side of the spectrum, the least frequent cancer types worldwide are vaginal (0.1% of all new cases and 0.1% of all deaths), mesothelioma (0.2% of all new cases and 0.3% of all deaths), penis (0.2% of all new cases and 0.2% of all deaths), Kaposi sarcoma (0.2% of all new cases and 0.2% of all deaths), vulva (0.2% of all new cases and 0.2% of all deaths), anus (0.3% of all new cases and 0.2% of all deaths), salivary glands (0.3% of all new cases and 0.2% of all deaths), testis (0.4% of all new cases and 0.1% of all deaths), Hodgkin lymphoma (0.4% of all new cases and 0.3% of all deaths), and hypopharynx (0.4% of all new cases and 0.4% of all deaths) [4].

In males, lung cancer is the one with the highest incidence and mortality (14.5% of all new cases and 22% of all deaths) [4]. In terms of incidence in males, lung cancer is followed by prostate, colorectal, and stomach cancers, which account for 13.5%, 10.9%, and 7.2% of all new male cancer cases, respectively. Regarding mortality in men, after lung cancer, the highest shares in cancer deaths are hold by liver, stomach, and colorectal cancers, which are responsible for 10.2%, 9.5%, and 9.0% of all male cancer deaths, respectively [4].

In females, the most frequently diagnosed cancer is breast cancer, being also the one that contributes the most to female cancer deaths (24.2% of all new cases and 15.0% of all deaths) [4]. The second to fourth most commonly diagnosed cancers in women are colorectal, lung, and cervix uteri cancers, which represent 9.5%, 8.4%, and 6.6% of all female new cancer cases. These three cancer types are also the ones that contribute mostly to women deaths by cancer, but in a different order, lung cancer ranking second (13.8% of all female deaths), and colorectal (9.5% of all female deaths) and cervix uteri cancers (7.5% of all female deaths) ranking third and fourth in women cancer mortality, respectively [4].

1.1.4.2 Cancer incidence and mortality in Europe

The European continent has a total population of approximately 744 million people, which represents around 9% of the total world population [4]. Even though Europe only represents less than a tenth of the global population, in 2018, 23.4% of all new cancer cases and 20.3% of all cancer deaths occurred in this continent. In fact, only in 2018, there were more than 4.2 million newly diagnosed cancer cases and more than 1.9 million cancer related deaths in Europe. In addition, the number of 5-year prevalent cases, in 2018, surpassed the 12.1 million [4].

In Europe, at the time of 2018, the five most frequently diagnosed cancer types were breast (12.4% of new cases), lung (11.1% of new cases), prostate (10.6% of new cases), colon (7.4% of new cases), and bladder (4.7% of new cases) cancers, while the least frequent were Kaposi sarcoma (0.06% of new cases), vagina (0.07% of new cases), nasopharynx (0.12% of new cases), penis (0.15% of new cases), and salivary glands (0.22% of new cases) cancers [4]. In terms of number of deaths, the most impactful cancers were lung (20% of deaths), colon (8.1% of deaths), breast (7.1% of deaths), pancreas (6.6% of deaths), and prostate (5.5% of deaths) cancers, while the least impactful were Kaposi sarcoma (0.02% of deaths), vagina (0.07% of deaths), testis (0.08% of deaths), penis (0.09% of deaths), nasopharynx (0.13% of deaths), and anus (0.19% of deaths) cancers. By sex, prostate is the most commonly diagnosed cancer in men (20.0% of new cases), followed by lung (13.9% of new cases), colorectal (12.1% of new cases), bladder (6.8% of new cases), and kidney (3.8% of new cases) cancers. In females, the top-ranking cancer in incidence is breast cancer (26.4% of new cases), followed by colorectal (10.6% of new cases), lung (8% of new cases), corpus uteri (6.1% of new cases), and skin melanoma (3.7% of new cases) cancers [4].

1.1.4.3 Cancer incidence and mortality in Portugal

Portugal is a small European country with a total population of around 10.3 million people which represents 1.4% of the European population and approximately 0.1% of the world population [4]. In this country, in the year of 2018, there were approximately 58.2 thousand new cancer cases and 29 thousand cancer related deaths. Furthermore, the number of 5-year prevalent cancer cases was 155.6 thousand [4].

In Portugal, at the time of 2018, the most commonly diagnosed cancers were breast (12.0% of new cases), prostate (11.4% of new cases), colon (9.7% of new cases), lung (9.1% of new cases), and rectum (7.6% of new cases) cancers, while the least commonly diagnosed were vagina (0.06% of new cases), mesothelioma (0.09% of new cases), Kaposi sarcoma (0.18% of new cases), salivary glands (0.18% of new cases), penis (0.19% of new cases), and nasopharynx (0.24% of new cases) cancers [4]. Mortality wise, the cancers that hold the largest number of death shares are lung (16.1% of deaths), colon (10.5% of deaths), stomach (7.9% of deaths), prostate (6.5% of deaths), and breast (6.0% of deaths) cancers, while the cancers with the lowest death count were Kaposi sarcoma (0.05% of deaths), vagina (0.06% of deaths), testis (0.06% of deaths), penis (0.14% of deaths), anus (0.16% of deaths), and Hodgkin lymphoma (0.18% of deaths) cancers. In males, the cancers with the highest number of new cases, in 2018, were prostate (20.4% of new cases), colorectal (18.8% of new cases), lung (12.3% of new cases), bladder (5.4% of new cases), and stomach (5.3% of new cases) cancers, while in females were breast (27.1% of new cases), colorectal (16.2% of new cases), thyroid (5.4% of new cases), lung (5.0% of new cases), and stomach (4.5% of new cases) [4].

Cancer is usually described as a genetic disease; however epigenetic abnormalities play profound roles in tumorigenesis [49]. Transformed cells consistently exhibit alterations in DNA methylation in a genome-wide level, aberrant chromatin structures, and altered regulatory element activities [49].

2.1 Epigenetics

2.1.1 Gene Expression

In humans, and other mammals, genetic information is embedded in DNA and is passed down to daughter cells somatically, and to the upcoming generation through the germline [50]. The decoding of the genetic information is carried out in a process of gene expression, in which the information stored in DNA is used to build functional molecules. The basic unit of this process is the gene, a section of DNA sequence that is able to give arise to a functional RNA molecule, a transcript, which in turn may have different functions, such as being translated into a polypeptide [50].

The first step in gene expression, where RNA is constructed from a DNA template, is named transcription and, in eukaryotes, is carried out by three key nuclear enzymes: 1) RNA

polymerase I (Pol I), that transcribes rRNA precursors, 2) RNA polymerase II (Pol II), that transcribes protein-coding genes to mRNAs, and 3) RNA polymerase III (Pol III), that transcribes small non-coding RNAs like tRNAs [51].

In eukaryotes, regulation of Pol II is of uttermost importance, since it is the foundation for cellular differentiation and identity [50–52]. Regulation of Poll II can happen in any of the three major stages of the transcription process: initiation, elongation, and termination. In the initiation phase, the Poll II assembles, at the DNA promoter region, with the transcription factors TFIIB, TFIID, TFIIE, TFIIF, and TFIIH (known as the general transcription factors), forming the pre-initiation complex. The general transcription factors are highly important to the transcription initiation, contributing to the binding of Pol II to the promoter, to initiate RNA synthesis, and to drive the Pol II to move forward. The assembly of the pre-initiation complex begins with the binding of a Pol II – TFIIF complex to a pre-assembled multipart in the promoter, composed of TFIIB and TATA box-binding protein (TBP). These constituents, bound to DNA, in the promoter, form the core initiation complex, which is conserved in the Pol I and Pol III transcription systems. The full pre-initiation complex is then completely formed when TFIIE and TFIIH bind to the core initiation complex. After assembly, the pre-initiation complex, in the presence of nucleoside triphosphates, promotes the opening of the double-stranded DNA in the promoter region, allowing for the DNA template strand to pass close to the Pol II active site, thus inducing the synthesis of the RNA chain, marking the beginning of the elongation phase. In some higher eukaryotes there is also an intermediate phase between initiation and elongation, where the polymerase lags in the proximal promoter region before the elongation starts. The RNA synthesis concludes in the termination phase, in which the transcript is released from the Poll II, that is also released from the DNA [50–52].

It is the reiteration of these stages, in a cyclical fashion, over a gene that defines its expression levels, of which regulation is of great importance to cellular function [50–52]. In fact, while all cells within an organism contain the same DNA sequence, it is the quantitative and qualitative regulation of gene expression that determines cell function and fate [50–52].

2.1.2 Chromatin dynamics and DNA organization

Although the negatively charged DNA is, in humans, linear, it is substantially compacted and organized into 3 dimensional structures known as chromosomes [53,54]. The double-chain DNA, in the cellular nucleus, is wound around proteins named histones, the main

proteinaceous component of chromosomes. The histone-DNA complex forms the basic structural chromosomal unit, the nucleosome, which consists of a histone octamer, with two of each histone monomers (H2A, H2B, H3, and H4), and 147 bp's of DNA. Most genomic DNA (around 80%) is organized into nucleosomes, and the remaining is contained in regions that connect neighboring nucleosomes, known as linker regions. The nucleosomes are wrapped in larger genomic structures of chromatin fibers and chromosomes [53,54].

This type of genomic compaction and organization permits selective accessibility of the transcription machinery, like transcription factors, to specific DNA regions, being compaction itself a key regulatory mechanism in gene expression [53,54]. Increased accessibility to different genomic regions renders specific effects, for example, enhancers to facilitate transcription, promoters to initiate transcription, open reading frames that are transcribed and translated into proteins, silencers to suppress transcription, or insulators that block interactions between promoters and enhancers. It is, then, the chromatin's organization dynamics, that allow cells with the same DNA sequence to have different functions and specializations [53,54].

The sum of mechanisms that regulate the chemical and/or structural alteration of the chromatin, collectively establishing different gene expression patterns in the same genome is known as epigenetics [53,54].

In higher eukaryotes, like humans, epigenetic modifications are several and include posttranslational modifications of histones, chromatin remodeling, DNA methylation, and noncoding RNA interactions [53,54]. The main similarity between these different epigenetic mechanisms is that none alters the primary DNA [53,54].

Epigenetic mechanisms of regulation are widely influenced by developmental and environmental stimuli, and although the DNA sequence per se is not altered, cells are able to transmit an "epigenetic memory" to daughter cells, passing their genetic information, but also their associated phenotype [53,54].

2.1.3 Histone Modifications

As previously discussed, gene expression is, in a large part, regulated by how the chromatin is organized in the cellular nucleus [55]. In fact, nucleosomes, by bending and cluttering DNA, greatly reduce its accessibility do transcription factors. The protein part of the nucleosome, the histones, are also subject to post-translational modifications, that influence

how the chromatin is compacted and its accessibility to nuclear enzymes. There are several types of histone post-translational modifications, such as methylation, acetylation, phosphorylation, ubiquitinoylation, sumoylation, deamination, ADP ribosylation, propionylation, and butyrylation [55].

Histones contain characteristic protrusions, that are projected away from the nucleosome, and thus more accessible, commonly known as histones N-terminal tails [55]. The modes of action by which post-translational histone modification affect the chromatin's structure and gene expression are various. These modifications can directly influence chromatin compaction, and thus transcription; for example, acetylation of lysine 16 of histone H4 (H4K16ac) is a type of modification that reduces chromatin compaction and increases transcription, while di- or tri- methylation of H4k20 promotes chromatin condensation. However, these types of epigenetic events do not only contribute to the direct chromatin remodeling, but can also have indirect modes of action, such as by recruiting effector proteins that activate signaling cascades, by causing obstruction to remodeling complexes, or by affecting the recruitment of transcription factors or chromatin remodelers [55].

One of the most dynamic epigenetic modifications of histones is acetylation [55]. This process is modulated by two families of enzymes, histone acetyltransferases (HATs), that promote the acetylation, and histone deacetylases (HDACs), that trigger the opposite effect. HATs catalyze the addition of an acetyl group, using Acetyl-Coenzyme A (CoA) as a cofactor, to the ε-amino group of a lysine side chain. Lysine amino acids residues, in the histone, have a positive charge, and by being acetylated, the charge is neutralized, potentially weakening histones' interaction with DNA, which is negatively charged. The family of HATs comprises of two main classes of enzymes, type-A HATs and type-B HATs. The class B HATs are mainly cytoplasmatic proteins, whose main role is to acetylate free cytoplasmatic histones, rather than acetylating the ones deposited in the chromatin. One of the key functions of type-B HATs is to acetylate histone H4 at the K5 and K12 residues, immediately after it is synthesized. These two marks allow for the deposition of the histone in the chromatin, and after this event has occurred successfully, the residues are deacetylated [55].

The type-A HAT class can be further subdivided into three type-A subgroups, namely 1) the Gcn5-related N-acetyltransferases (GNAT), 2) the MYST family (named after the founding members: MOZ, Ybf2/ Sas3, Sas2 and Tip60), and 3) the cyclic

adenosine monophosphate (cAMP)-response element binding protein (CREB)-Binding Protein (CBP)/p300 family [55].

On the opposing side of the spectrum of histone acetylation are the HDACs [55]. These are unspecific enzymes regarding substrate selection, being a single HDAC able to deacetylate several sites within histones. Similarly to HATs, HDACs are also subdivided in categories: class I (HDAC1, HDAC2, HDAC3, and HDAC8), class II (HDAC4, HDAC5, HDAC6, HDAC7, HDAC9, and HDAC10), class III (SIRT1, SIRT2,SIRT3, SIRT4, SIRT5, SIRT6, and SIRT7), and class IV (HDAC11). Although HDACs are subdivided into classes, this categorization is many times erroneously noted as a taxonomic classification, but truthfully all HDACs fit in the α and β protein classes [55].

Another type of histone post-translational modification is phosphorylation, and similarly to histone acetylation, this is a very dynamically regulated epigenetic process [55]. Contrarily to acetylation, however, phosphorylation mainly affects the residues serine, threonine, and tyrosine, on the histone tails, but also in their globular portion. In histones, as well as in other proteins, phosphorylation is mainly modulated by two types of enzymes: kinases, that catalyze the addition of a phosphate group, and phosphatases that have the reverse function. Histone kinases act by transferring a phosphate group from Adenosine triphosphate (ATP) to a certain amino-acid's side chain, thus creating a negative charge in that region of the histone, which in its turn influences the chromatin's organization [55].

The most studied histone post-translational modifications are the ones in these N-terminal tails, nonetheless these are not the only regions that can be modified [55].

2.1.4 Noncoding RNA

Epigenetic regulation of cellular protein levels is also carried out by untranslated RNA molecules termed as noncoding RNAs (ncRNAs) [56]. These RNA molecules aren't translated into proteins and are classified based on length, function, and cellular localization. Five major ncRNA classifications, with an epigenetic role, exist: microRNAs (miRNAs), small interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), small nucleolar RNAs (snRNAs), and long ncRNAs (lncRNAs) [56].

2.1.4.1 miRNA

MiRNAs are ncRNA molecules of small size, with an averaging length of 22 nucleotides [57]. miRNAs mediate gene silencing by directing proteins from the Argonaute (AGO) family to the target sites of mRNA molecules, which are typically in the 3' untranslated region (UTR)'s region. Nonetheless, miRNAs can also interact with other regions of the target mRNA, such as its coding sequence, 5'UTR region, or its promoter region [57].

These molecules are usually transcribed from a DNA template into primary miRNAs (pri-miRNAs), which are processed into precursor miRNAs (pre-miRNAs), and later into mature miRNA molecules [57]. These fully matured miRNAs are then able to regulate gene expression by interacting with target mRNA molecules. miRNA's expression patterns are tissue-specific, and its deregulation is associated with several diseases, including cancer. In fact, although these molecules are not translated, miRNAs can be considered tumor suppressors or oncogenes, in the later case being called oncomirs [57].

General classification of miRNAs is done by clustering these molecules into families, based on the similarity of their seed, which is a small 2-8 nucleotide sequence that is largely responsible for mRNA target recognition [57].

miRNA molecules mediate gene silencing by forming a complex known as minimal miRNA-induced silencing complex (miRISC), which is composed by a guide strand and an AGO protein [57]. The miRISC specifically interacts with its target mRNA by binding, through its seed sequence, to complementary sequences known as miRNA response elements (MREs). The miRNA - MRE complementary percentage dictates one of two possible fates for the target mRNA: 1) miRISC-mediated translational inhibition and decay, or 2) AGO2-dependent slicing of the target mRNA. A total miRNA – MRE base complementary promotes AGO2 endonuclease activity and, thus, cleavage of the target mRNA. However, in animal cells the vast majority miRNA – MRE contain at least one mismatch, thus preventing AGO2 endonuclease activation. In this case, AGO2 will act as a silencing mediator of RNA interference (RNAi), rather than actively cleaving mRNA, by recruiting the poly(A)-deadenylases PAN2/ and Carbon Catabolite Repression—Negative On TATA-less (CCR4-NOT), that will initiate and complete the poly(A)-deadenylation process, respectively. The target mRNA will then be decapped by the decapping protein 2 (DCP2) and degraded by exoribonuclease1 (XRN1) [57].

2.1.4.2 siRNA

siRNAs are small double-stranded RNA molecules with lengths between 20 – 25 bp. These molecules constitute an epigenetic player in gene silencing, acting through the biological mechanism of RNAi [58].

siRNAs are produced from longer dsRNA molecules, which are cleaved into shorter fragments, by the RNase III-like enzyme Dicer and, like miRNAs, form an effector complex with an Argonaut protein [58]. Although exogenous and/or artificial siRNA molecules are powerful and well-studied tools in biomedical research and potential therapeutic weapons in genetic diseases, mechanistic understanding of endogenous siRNAs in mammals is scarcer. One of the reasons for the lack of total systematic comprehension of the biosynthesis and biological role of endogenous siRNAs in mammalian cells, is that its existence causes some theoretical conflict with the fact that the occurrence of dsRNA in these cells is a hallmark of viral infection and prompts a powerful immune response via protein kinase R/ interferon [58]. Nonetheless, reports show that endo-siRNA expression in mouse oocytes, do not trigger an interferon response against dsRNA [58,59]. Some research seems to link endogenous siRNAs with retrotransposons, such as the Long interspersed nuclear element 1 retrotransposon (LINE-1). For instance, in breast cancer cells, LINE-1 transcripts are frequently enriched, whereas LINE-1 related endogenous siRNAs are usually depleted. Furthermore, overexpression of these siRNAs greatly silences LINE-1 expression by augmenting DNA methylation of its promoter [58,59].

2.1.4.3 PIWI-interacting RNAs

piRNAs are yet another class of ncRNAs of small size [60]. With lengths that range from 21 to 35 nucleotides, these animal-specific molecules have various biological roles such as gene expression regulation, viral combat, and silencing of transposable elements [60].

Unlike miRNAs and siRNAs, that stem from double-stranded RNA precursors, piRNAs originate, without the action of Dicer, from long ssRNA precursors, that are transcribed from genomic regions named piRNA clusters [60]. One of the main functions of piRNAs is to protect the germline genome against transposon mobilization, however, it is not understood how piRNAs are able to differentiate self-transcripts from non-self-transcripts [60].

piRNAs, similarly to miRNAs and siRNAs, interact with a protein from the Argonaute family [60]. However, this family of proteins can be divided into the AGO and PIWI clades, and whereas the siRNAs and miRNAs interact with an AGO effector protein, the piRNAs interact with a PIWI clade. One of the differences between these two groups of proteins, is that while AGO proteins are ubiquitously expressed, PIWI-clade proteins are usually restricted to gonadal cells, which explain some of the piRNA main functions [60].

2.1.4.4 small nucleolar RNAs

snRNAs are ncRNAs that mainly accumulate in nucleoli and their lengths can vary between 60 to 300 nucleotides [61]. This class of RNA molecules is responsible for post-transcriptional and maturation of other cellular RNA molecules, such as ribosomal RNAs. Most snRNAs are encoded in introns, and their genesis usually involves co-transcription with the gene where they reside, splicing, debranching of the intron lariat, and subsequent exonucleolytic digestion [61].

The main function of snRNAs is in the maturation of rRNAs [61]. For example, snRNA molecules promote modifications within conserved and functional regions of rRNAs, such as triggering the 2'-O-methylation of the fifth nucleotide by the methylase fibrillarin, prompting the conversion of uridines to pseudo-uridines by the dyskerin protein, and even by participating in pre-rRNA cleavage [61].

2.1.4.5 long ncRNAs (lncRNAs).

lncRNAs are, unlike the previously discussed ncRNAs, longer non-coding transcripts with lengths that range from 200 to 100.000 nucleotides [62]. These molecules are similar to mRNA transcripts, but don't have stable open reading frames, and their expression levels are not only tissue specific, but also appear to be generally lower than protein-coding transcripts [62].

LncRNAs are important cis and trans-acting modulators of protein-coding gene's expression [62]. These molecules can recruit chromatin-remodeling enzymes, like histone methylases, acetylases, and deacetylases, to specific chromatin loci, mediating the chromatin state and thus activating or repressing local genes. LncRNAs can also interact with other RNA-

binding factors to form RNA-protein complexes, that can promote transcription by recruiting key proteins to gene promoters or repress transcription by binding to existing gene repressors. These molecules have also been shown to be intricated in the repression of apoptotic genes, like B-cell lymphoma 2 (BCL2) Interacting Killer (BIK) and Fas Cell Surface Death Receptor (FAS) [62].

2.1.5 DNA methylation

Cytosines, one of the four bases that can be found in DNA, not only are a part of the genetic code, but also contain epigenetic information through chemical modification of its pyrimidine ring, a process named DNA methylation [63–68]. In this process, a methyl (CH3) group is enzymatically and covalently added to the fifth position of cytosines, giving rise to 5-methylcytosines (5-mC). DNA methylation is one of the most studied epigenetic events in mammals, but is also considerably conserved amongst other animals, plants, and fungi [63–68].

Although cytosine methylation is an event that is temporally and spatially regulated, is also a frequent one [63–68]. In fact, in humans, methylation occurs in approximately 60 – 80% of the 28 million CpG dinucleotides present in the human genome. These CpG dinucleotides are not evenly distributed across the genome, but rather condensed in regions named CpG islands. These regions have, normally, at least 200bp and is comprised mostly of CpG dinucleotides (more than 50% of its nucleotides) .Even though any cytosine in the genome can be methylated, this process is mostly constrained to palindromic CpG dinucleotides, being cytosine methylation in a non-CpG context (CpH, H = A, T, C) rare in most mammals. However, methylation is not biologically exclusive to CpG dinucleotides, indeed non-CpG methylation is a frequent event in plants, oocytes, pluripotent embryonic stem cells (ESCs), and mature neurons [63–68].

DNA methylation is not a static process or pattern, but rather a dynamic one [63–68]. A specific methylation pattern in a specific genomic region, or methylation mark, can be de novo synthesized, maintained or removed. The dynamics of DNA methylation is regulated by a meticulous balance between DNA methyltransferases (DNMTs), that methylate cytosines, and DNA demethylases, that remove the methyl group from 5mC [63–68].

DNA methylation and demethylation are two very distinct processes, with distinct players [63–68]. DNA methylation is catalyzed by three key enzymes, with methyltransferase

activity: DNMT1, DNMT3A, and DNMT3B. DNMT1 major role is in the maintenance of methylation patterns following, for example, DNA replication or genomic damage repair, being mostly active on DNA that has one of its two strands already methylated, also known as hemi-methylated DNA. De novo methylation is mostly carried out by the other two methylation enzymes DNMT3A and DNMT3B, which are not able to discriminate methylated and hemi-methylated DNA substrates [63–68].

Global DNA methylation can only be maintained with the action of DNMT1; thus, this enzyme is constitutively active in human cells [63–68]. The amine-terminus of the DNMT1 protein includes various regulatory domains, such as 1) the Replication Foci Targeting Sequence (RFTS), which is involved in DNMT1 dimerization and possibly implicated in hemi-methylated DNA recognition, 2) the DNA methyltransferase associated protein 1 (DMAP1)-binding domain, that interacts with Histone Deacetylase 2 (HDAC2), mediating transcription co-repression, 3) the Bromo homology domain, possibly important for protein-protein interactions, and 4) the cysteine rich CXXC domain, that enables the DNMT1's interaction with the unmethylated DNA. On the other hand, while the critical regulatory domains are located in the protein's N-terminus, it's catalytic domain rests in the carboxy-terminus. In fact, it seems that it is the interaction between the protein's regulatory domains, in the N-terminus, and the catalytic domain, in the C-terminus, that allows the allosteric activation of the DNA methyltransferase. DNMT1 also has an obligate partner, the ubiquitin-like plant homeodomain and RING finger domain 1 (UHRF1), a protein that preferentially recognizes hemi-methylated CpG sites [63–68].

Unlike DNMT1, the de novo methyltransferases DNMT3A and DNMT3B are usually downregulated in adult somatic cells, nevertheless they are highly expressed in undifferentiated embryonic stem cells [63–68]. To catalyze DNA methylation, DNMT3A and DNMT3B interact with a histone protein, usually with the unmodified lysine 4 residue of histone 3 (H3K4me0), through the Pro-Trp-Trp-Pro domain. Another important domain present in the de novo methyltransferases is the X-linked helicase II (ATRX)-DNMT3-DNMT3L (ADD) domain, a zinc finger binding domain that continuously inhibits the catalytic methyltransferase domain until the enzyme binds to the histone [63–68].

De novo DNA methylation is modulated by a third member of the DNMT3 family, the DNMT3-like protein (DNMT3L) [63–68]. This modulator for DNMT3A and DNMT3B activity is not catalytically active, in fact it does not have a functional catalytic domain and, unlike the

other DNMTs, doesn't interact with S-adenosyl methionine (SAM), the methyl group donor for DNA methylation. DNMT3L is able to physically interact with the catalytic domains of DNMT3A and DNMT3B, promoting their affinity to SAM, and thus stimulating their DNA methyltransferase activity [63–68].

Even though DNA methylation is a crucial cellular process, the loss of 5mC, or DNA demethylation, is of equal importance [63–68]. Genome wide DNA demethylation is of great importance in particular settings, such as creating and maintaining a pluripotent state in early embryos, or for eliminating parental imprints in developing primordial germ cells [63–68].

The process of DNA demethylation can happen in a passive manner or an active one [63–68]. In every cell cycle, after DNA replication, the action of DNMT1 leads to the maintenance of the methylation patterns, by the symmetrically methylation the nascent DNA strand. However, in the absence of functional maintenance machinery (DNMT1 and/or UHRF1), loss of 5mC occurs and, after successive cycles of DNA replication, there is a passive decrease in methylated cytosines, a phenomenon known as passive dilution of 5mC [63–68].

Like DNA mutations, changes in DNA methylation can be propagated to daughter cells [69]. However, the latter is much more vulnerable to environmental stimuli than the former. In fact, it is the dynamic nature of the methylation and demethylation processes that allow for cells with identical genetic sequences to display different phenotypes. It has been increasingly shown that, during the tumor initiation, development, and metastasis processes, mutations are not the only factor for clonal evolution. Events of aberrant DNA methylation, sometimes referred as epi-drivers, might also provide selective advantage to a tumor clone, contributing to the tumorigenic process and metastatic cascade (Fig. 1.3) [69].

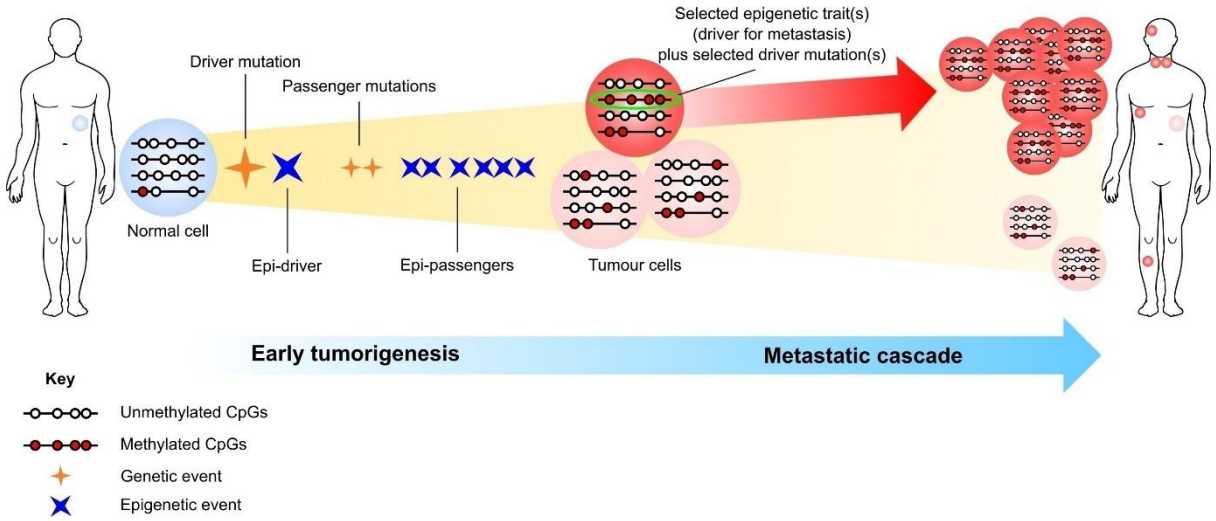

Figure 1.3 – Epigenetic alterations during tumor initiation, tumor development, and metastasis and clonal selection of epigenetic traits. Epi-drivers and driver mutations are key events that occur during tumor initiation. During tumor development several epigenetic changes can provide selective advantage to the tumor clone, being kept in the population and thus promoting tumor development. The same process might impact the metastatic cascade. Reprinted under a Creative Commons Attribution-Noncommercial 4.0 International Public License, from Elsevier: Seminars in Cancer Biology, Epigenetic drivers of tumorigenesis and cancer metastasis (2018) [69].

2.1.6 Relationship between DNA methylation and histone modifications

DNA methylation and histone modification are closely linked [70]. In fact, it seems that both processes seem to reciprocally influence each other, where histone modifications can direct DNA methylation patterns, and DNA methylation may provide a template to certain histone modification after DNA replication. In fact, several histone modifications have been suggested to be directly implicated in DNA methylation. For example, trimethylation of histone H3 lysine 9 (H3K9), histone H3 lysine 27 (H3K27), and histone H4 lysine 20 (H4K20) seem to be a prerequisite for local DNA methylation [70]. Enhancer of zeste homolog 2 (EZH2), one of the main constituents of the polycomb repressive complex 2 (PRC2) and a histone methyltransferase, is known to directly interact with DNMTs [71]. Knockdown of EZH2 not only reduces H3K7 methylation but also DNA methylation at specific EZH2 target genes. Furthermore, overexpression of EZH2 has been shown to increase DNA methylation at CpG sites [71]. Another histone methylase, named G9A, responsible for the catalysis of mono- and di-methylation of H3K9 and H3K27, is also involved in DNA methylation [72,73]. In 2007, Ikegami K. and colleagues, demonstrated that G9A knockout in murine cells led to site specific reduction of DNA methylation [73]. The G9A histone methylase can form an heteromeric complex with the euchromatic histone lysine methyltransferase 1 (EHMT1) and not only

promote methylation of H3K9, but also DNA methylation in site [72,73]. Together, these two epigenetic phenomena cause a local reduction in gene transcription. Although these two proteins are histone methylases, it has been shown that their knockout in Embryonic Stem Cells (ESCs) leads to DNA promoter hypomethylation of their target genes [72,73].

Protein arginine N-methyltransferase 5 (PRMT5), another histone methylase, catalyzes the di-methylation of histone H4 arginine 3 (H4R3) [74]. This epigenetic histone mark is a binding target for DNMT3A, which in turn methylate adjacent CpG dinucleotides [74].

2.1.7 Disruption of chromatin homeostasis and tumorigenesis

Chromatin's conformation is of major importance to cellular functions, since it is what allows for proper global gene expression [75]. Active genes and elements must be accessible to regulatory factors, whereas inactive genes are compacted inside inaccessible structures that prevent their transcription [75]. The disruption of this epigenetic homeostasis by genetic, environmental, and metabolic stimuli can promote tumor initiation and/or accelerate tumor development, by making certain chromatin regions abnormally compacted and consequently restrictive, or aberrantly unfolded, and thus permissive [76,77].

DNA methylation plays a profound role in the aberrant alteration of chromatin permissiveness during tumor initiation [77]. In normal cells, cytosines in guanine-cytosine (CG) enriched genomic regions are usually unmethylated, and cytosines in regions with a low CpG count are frequently highly methylated. However, in several cancers this methylation model is not observed. This common aberrant epigenetic phenomenon is termed as CpG island methylator phenotype (CIMP), where CpG islands become hypermethylated and CpG-poor loci become hypomethylated. This occurs in a wide range of cancer types, and is considered to be a restrictive epigenetic event, since CpG-rich loci hypermethylation has been shown to silence tumor suppressor genes like p16, and DNA mismatch repair genes like the mutL homolog 1 (MLH1) and the MutS homolog 2 (MSH2) [77].

DNA methylation can not only be intricated in epigenetic restriction events, but also induce a permissive chromatin state [77]. This permissive status allows cells to switch transcriptional states, genetic pathways, or developmental programs, many of which can be pro-oncogenic. The propagation of a selectively advantageous plastic chromatin state that is

propagated to daughter cells through mitosis, will cause the establishment of a new tumoral clone with increased fitness, and thus a step forward in the tumor development process [77].

Most cancer types follow an aberrant global methylation pattern, and since these patterns are a fundamental mechanism that modulates the chromatin's permissiveness, which in turn triggers different cellular pathways, a central biological question remains to be answered: "Are there genomic regions that are more susceptible to alterations is DNA methylation during tumorigenesis?". If this question yields an affirmative answer, then a secondary question arises: "Are these regions related to altered patterns of gene expression and consequently altered cellular patterns?".

Chapter 2 Aims

We hypothesize that upon tumor initiation certain genomic regions are differentially methylated and others are resistant to changes. We termed the first Epi-Hotspots and the later Epi-Blackholes. We also hypothesize that these genomic regions might be somehow related to alterations in gene expression during oncogenesis. Since gene expression and DNA methylation patterns are cell-type specific, we also expect that these genomic oncogenic events are also primarily tumor specific.

In addition, we also theorize that if these events are not present uniformly in every patient of each cohort, there might be some that might act as potential prognostic biomarkers able to predict patient survival in a more advanced phase of the disease.

To test our hypothesis, we aim to perform a pan-cancer analysis to:

- Search for genomic regions that are more susceptible to alterations in DNA methylation (Epi-hotspots) during tumor initiation;
- Search for genomic regions that are immutable regarding DNA methylation (Epi-Blackholes) in the normal – tumoral transition;
- Identify alterations in gene expression during tumor initiation and check which of them could be explained by either Epi-hotspots and/or blackholes.
- Perform a pan-cancer cross-examination of the previous alterations to understand common and differentiating patterns between different tumor types.
- Understand which biological processes are possibly being modulated by Epi-Hotspots or Epi-Blackholes.
- Examine if the identified Epi-Blackholes and epi-hotspots can predict survival of stage-III cancer patients.

Chapter 3 Methodology

3.1 Analytic Tools

All statistical analysis in this study was performed using R, which is a language and environment that can be used for statistical computing and graphics.

The R language choice for this analysis was due to several criteria, such as 1) is an open-source language, 2) provides the ability to perform several statistical and graphical techniques, 3) its capabilities can be easily extended through packages, 4) is widely used in the scientific community, which allows for increased study replicability and easy employment of methods described in other studies.

The R code developed herein was manipulated via a software named RStudio, which is an integrated development environment (IDE) specifically designed to operate R code. This open-source manipulation software was selected due to the several user-friendly features it possesses, such as a workspace browser, data viewer, ability of managing multiple working directories, having integrated R documentation, interactive debugger, among others.

A large part of this analysis was completed using several R packages, that were developed and published by other users. These are collection of R functions and compiled code, that allow for an easier (and pre-optimized) implementation of statistical tests, mathematical operations, or other computational tasks.

3.2 Data Source

In order to assess DNA methylation alterations in tumor initiation, we analyzed data from several cancer types originally generated by The Cancer Genome Atlas Consortium (TCGA). The TCGA was a program that started in 2006 by the National Human Genome Research Institute (NHGRI) and the National Cancer Institute (NCI) [78]. During more than 12 years, this cancer genomics program was able to examine 33 cancer types and characterize over 20,000 cancer and matched normal samples, generating proteomic, genomic, transcriptomic, and epigenomic pan-cancer data that totals more than 2.5 petabytes of publicly available information (equivalent to more than 2,500,000,000 megabytes of data) [78].

Being our goal to analyze DNA methylation variation during tumor initiation, we selected all the cohorts in the TCGA program that had publicly available DNA methylation and gene expression data for both normal and stage-I tumoral samples. Our study started from level-3 DNA methylation data generated from Infinium HumanMethyation450 bead array, and level-3 gene expression data generated from Illumina RNA-sequencing (RNA-Seq).

The selected datasets had already been through level-3 processing using the latest Human Genome Assembly hg38 and were downloaded from the National Institutes of Health's (NIH) Genomic Data Commons (GDC) Data portal (https://portal.gdc.cancer.gov/). The TCGA data is archived in the database of Genotypes and Phenotypes (dpGaP) under the accession number phs000178. In our analysis we used the v0.p8 version of the dataset, publicly released on August 17, 2018.

The TCGA cohorts used in the present study, respective identification code, disease, and primary site are listed below (Table 3.1).

Table 3.1 - Summary description of the analyzed datasets.

TCGA identification code	Disease	Primary site
TCGA-COAD	Colon Adenocarcinoma	Colon
TCGA-PAAD	Pancreatic Adenocarcinoma	Pancreas
TCGA-BRCA	Breast Invasive Carcinoma	Breast
TCGA-CHOL	Cholangiocarcinoma	Bile ducts (liver)
TCGA-ESCA	Esophageal Carcinoma	Esophagus
TCGA-HNSC	Head and Neck Squamous Cell Carcinoma	Oral cavities and pharynges
TCGA-KIRC	Renal Clear Cell Carcinoma	Kidney
TCGA-KIRP	Renal Papillary Cell Carcinoma	Kidney
TCGA-LIHC	Hepatocellular Carcinoma	Liver
TCGA-LUAD	Lung Adenocarcinoma	Bronchus and lung
TCGA-LUSC	Lung Squamous Cell Carcinoma	Bronchus and lung
TCGA-THCA	Thyroid Carcinoma	Thyroid gland

3.2.1 Infinium HumanMethyation450 bead array

As previously described, we opted to analyze DNA methylation data spawned by Infinium HumanMethyation450 bead array [79]. This type of assay has the ability to assess the methylation status of over 450 thousand CpGs sited across the genome, covering 96% of CpG islands in the human genome. This array utilizes two distinct probe types: 135501 Infinium-I probes and 350076 Infinium-II probes. An Infinium-I CpG target site's methylation status is assessed by a 50bp probe that detects a "methylated" (M) intensity, and by another equal sized probe that detects an "unmethylated" (U) intensity. On the other hand, Infinium-II CpG sites are targeted by only one probe that distinguishes "M" and "U" intensities employing a green dye, and a red dye [79]. The methylation level of a single CpG site is then represented by a β-Value, which can be computed with the following formula [80]:

$$\beta = \frac{M}{M + U}$$

3.2.2 Illumina RNA-sequencing (RNA-Seq)

In RNA-Seq complementary DNAs (cDNAs) are directly sequenced using high-throughput next generation sequencing (NGS) [81]. After this step, the sequencing reads are mapped to the reference genome for gene expression analysis [81].

In this process, a population of RNA is first converted into a library of cDNA fragments [81]. Each cDNA fragment has adaptors in one or in both extremities. Each of these fragments are then sequenced to produce short sequences reads, which are then aligned with the reference genome (or transcriptome), generating a base-resolution expression profile for each gene [81].

3.2.3 Data levels

Genomic data can be obtained with different degrees of treatment and depending on the type of treatment that was already applied to the data, this can be categorized in what are called Data Levels [82].

Typically, there are 4 data levels:

- Level 1 – This type of data is usually untreated and un-normalized. It is commonly referred as raw data.

- Level 2 – This represents data that was already normalized. It can be viewed as an intermediate degree of data processing.
- Level 3 – Data with this level of treatment is generally normalized, aggregated, and, sometimes, segmented.
- Level 4 – This is the combination of the previous treatments, but in an integrative multi-cohort analysis, such as a pan-cancer analysis [82].

The TCGA data that was obtained for the present study, had already level-3 treatment.

The DNA methylation datasets were imported as a combined genomic matrix of β-Values, in which the observations were the samples and the variables corresponded to the CpG sites. All β-Values that corresponded to the same sample, but from different aliquots where averaged.

The gene expression data was also aggregated into a genomic matrix, in which observations corresponded to samples and variables corresponded to genes. Data that matched the same sample, but different aliquots was also averaged. The data was also normalized according to the following function:

$$y = \log_2(x + 1)$$

3.3 Data preparation

Before initiating our analysis, all datasets were subjected to a pre-analytic processing, which involved variable and observation selection, outlier removal, missing data treatment, and age and gender calibration.

3.3.1 Variable and observation selection

Since the main goal of our research lies in tumor initiation, only samples that corresponded to either normal tissue or primary stage-I tumor tissue were analyzed.

RNA-Seq technology is able to generate reads for all kinds of transcripts, but in the current analysis our aim is to link methylation regions with genes. For this reason, all transcripts that did not correspond to a protein-coding gene were removed from the gene expression genomic matrixes.

3.3.2 Missing Data

Missing values in a sample can compromise the reliability of the results and can also produce bias [83]. Several procedures exist to deal with missing values, such as imputing missing values - which involves substituting the absent values by a numeric value derived from statistical analysis - or using only samples without missing values [83].

In our study, we removed all variables and observations in which at least half the data was missing.

3.3.3 Outlier removal

Outliers are extreme values that fall outside the common pattern of distribution of values [83]. These are data points that lie far away from the majority of the other data points and can originate from several factors including data entry errors, measurement errors, or abnormal values that are not common in a population. By creating bias, outliers can greatly affect statistical analysis, leading to over- or under- estimated statistical estimates. Various methods for identifying outliers exist, some of which use the distribution's mean and standard deviation (SD) to select them. However, since the mean and SD themselves are outlier sensitive, we identified aberrantly extreme values using the box plot method. In this method, any data point that is located outside 1.5 times the interquartile range (IQR), which is the distance between the third and first quartiles, either above the 75^{th} percentile or below the 25^{th} percentile is deemed an outlier [83].

Outlier detection and deletion was performed for every variable of every dataset.

3.3.4 Age and gender calibration

Due to the small size of the available cohorts, comparisons between normal and stage-I tumoral populations was not done via matched pairs. Gender and age differences between groups can lead to a biased analysis and since we did not find any good existing protocol to address this problem, we developed a method for inter-group homogenization of these variables described below (Fig. 3.1). To assess gender and age differences between groups, the nonparametric Wilcoxon rank-sum test, and the chi-squared test were used, respectively. Inter-group homogenization was achieved by systematically identifying the observation (or

individual) that caused the most inter-group variability with respect to age and gender. A significance threshold (α) of 0.05 was chosen for both the Wilcoxon and chi-squared tests. The next stage of the pipeline was to identify the age at which there was the greatest difference in terms of frequency of individuals and subsequently, which group had the highest number of individuals of this age. Since we also wanted to decrease inter-group heterogeneity for gender, the decision to remove a male or a female was based on gender frequency. If, at any given cycle, more than one individual was fit for exclusion, the removal was done randomly. After exclusion of the selected observation, the Wilcoxon and chi-squared tests were repeated, and the whole process was repeated until statistical significance was above the chosen threshold.

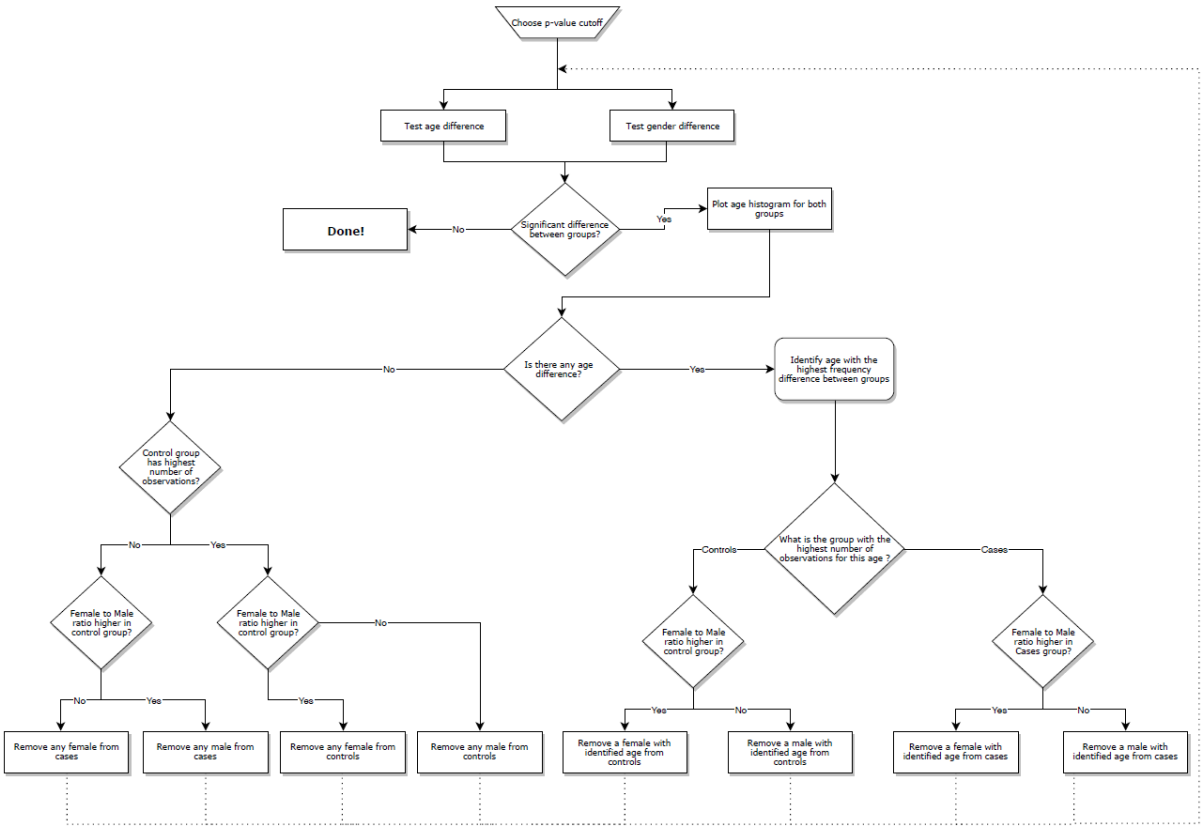

Figure 3.1 - Decision tree for age and gender calibration between cohorts. The algorithm here developed to decrease inter-group heterogeneity for age and sex, between cases and controls, systematically identifies and removes the individual that causes the largest difference regarding sex and age. The algorithm continuously iterates until the difference is below a user-defined threshold.

3.4 Epi-Hotspot Identification

In order to identify genomic regions that are more susceptible to methylation, i.e. Epi-Hotspots, we searched for Differentially Methylated Regions (DMRs) between normal and stage-I tumor samples using two different DMR-seeking algorithms, and by intersecting the results.

In September 2018, Mallik and colleagues performed a comprehensive analysis of four popular DMR finding methods: Bumphunter, Comb-p, DMRcate, and ProbeLasso [84]. The researchers generated forty simulated DNA methylation datasets and evaluated several performance indicators of the algorithms, such as sensitivity (predicted positives / actual positives), precision (actual positives / predicted positives), area under precision-recall curve (AuPR) (a comparison of sensitivity and precision, representing the overall discriminatory ability of a method to assess whether a region is associated with disease), Matthews correlation coefficient (MCC) (the correlation between the observed and the predicted binary classification), F1 score (F1) (measures accuracy taking into account precision and sensitivity), and type I error rate [84].

The Comb-p algorithm was quickly eliminated as an option because no computational tools existeded that allowed for its practical implementation in R. Of the remaining three algorithms, DMRcate had the best performance but Bumphunter was the one that identified the greatest number of DMRs, in a real dataset [84]. For these reasons, we elected to use both the DMRcate and Bumphunter algorithms and then intersect the results.

3.4.1 The Bumphunter Algorithm

In the Bumphunter methodology, the DNA methylation genomic matrix is converted from β-Values to M-Values, that is $M = \log(\frac{\beta Value}{1-\beta Value})$ [85]. A linear regression model is then applied by group in order to model differential methylation between cases (normal samples) and controls (stage-I primary tumor samples), at each CpG site. All consecutive CpG sites with a t-statistic that exceeds a certain threshold are clustered together into candidate DMRs, referred as “bumps”. Null distributions are generated for the candidate regions to estimate the statistical significance of the candidate DMRs separated by a minimum distance defined by the user (in this case 300 bp) [85].

To implement this algorithm in R we utilized the *Bumphunter()* function from the Bumphunter Bioconductor package. The parameter settings are described in Annex I.

3.4.2 The DMRcate Algorithm

The DMRcate method, like the Bumphunter method, starts with the logit transformation of β-Values to M-Values [86]. The algorithm continues by fitting a linear model at every CpG site, where the M-Value is regarded as the outcome variable and the group status (normal vs. stage-I tumor) is considered as the independent variable. For every CpG site, the statistic $Y = t^2$ is calculated, where the "t" corresponds to the t-statistics from the linear model. The method proceeds with the application of kernel smoothing, using a Gaussian smoother with bandwith λ (in this study, λ = 1000), scaled by a scaling factor C (in this study, C = 2). At each CpG site, the p-value is computed using the Sattarerhwaite method, and then corrected using the Benjamini and Hochberg method. Significant CpG sites that are within λ nucleotides from each other are collapsed into DMR regions. A p-value for each DMR is finally calculated using Stouffer's method [86]. Here, a DMR was considered to be statistically significant if the final p-value was lower than a fixed α of 0.05.

The DMRcate method was implemented in this study using the *dmrcate()* function from the Bioconductor package *DMRcate*. The specific function arguments and respective parameters are available in Annex I.

3.4.3 Identification of Epi-Hotspots

Having applied both DMRcate and Bumphunter algorithms to the datasets of interest we then proceeded to identify Epi-Hotspots, which we defined as differentially-methylated regions that were identified by both algorithms and in which the median point of one region overlapped with the other region. Only the segment identified by both methods was considered an Epi-Hotspot (Fig. 3.2).

To understand which genomic regions were enriched by the identified epi-hotspots, we used the annotation data provided for the Illumina 450k array to analyze the genomic locations of the epi-hotspots[79].

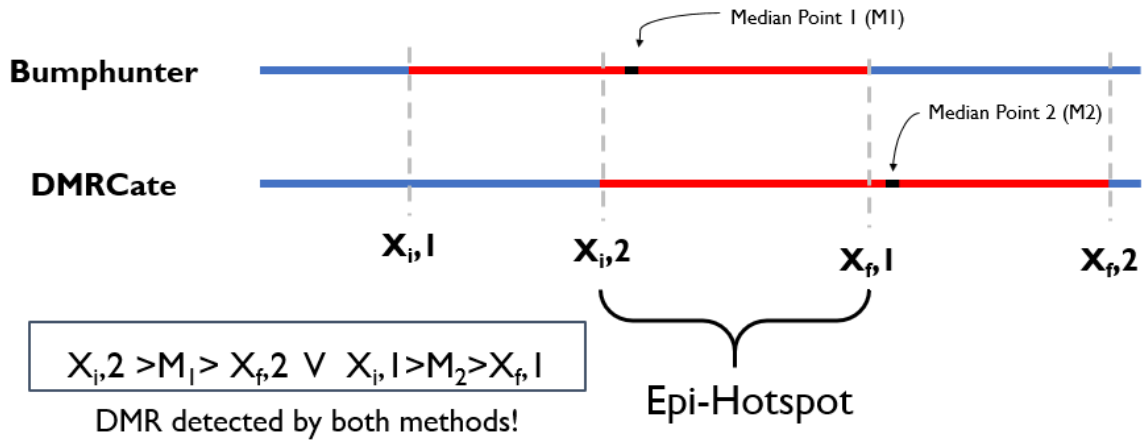

Figure 3.2 - Graphical illustration of the epi-hotspot identification method. Epi-hotspots were identified by intersecting the output of Bumphunter and DMRCate algorithms. If either the median point of the region detected by Bumphunter (M1) or the median point of the region detected by DMRCate (M2) was in the region detected by the other method, we considered that both algorithms identified the region. The intersection segment was considered an epi-hotspot.

3.4.4 Hierarchical Clustering between cohorts

Having identified and mapped the regions that were differentially methylated between normal and stage-I primary tumor samples (Epi-Hotspots) across multiple cancer types, we next aimed to understand which cancers were more similar and which ones were more divergent from each other, with respect to these regions.

We, therefore, performed a hierarchical clustering analysis, using the Epi-Hotspot overlapping percentage as a similarity unit. Dissimilarity (Diss) was determined as Diss = 1 – Similarity.

Hierarchical clustering is a type of clustering algorithm that partitions several objects into a tree of nodes, in which each node corresponds to a cluster [87]. Each of these nodes can have zero or more child nodes placed below the parent node in the tree. Usually, hierarchical clustering trees are built and read in a vertical downward fashion. This type of clustering algorithm can be performed employing several methods. In this analysis, the *complete linkage* method was used, where the distance between two clusters represents their maximum distance.

This method represents a conservative clustering approach, which is why it was chosen in this study [87].

3.5 Epi-Hotspot's relation with gene expression alterations

Having identified the Epi-Hotspots in the analyzed cancer types, our next goal was to understand if these regions are able to predict (or explain) gene expression alterations during the normal to stage-I tumoral transition. The first step was to identify which protein-coding genes had statistically significant differential expression in stage-I primary tumor samples relative to normal samples.

3.5.1 Searching for differentially expressed genes

To identify genes which were differentially expressed between normal and stage-I primary tumor samples, we employed several statistical inference techniques. Statistical inference is a strategy that makes use of data from a sample to describe the distribution of data in the population. In this analysis our null hypothesis (H_0) was that the mean difference in gene expression between normal and stage-I primary tumor samples would be zero. Our alternative hypothesis (H_1), was that the mean difference in gene expression between normal and stage-I primary tumor samples would be significantly different from zero.

In hypothesis testing, a sample is used to determine which of the hypotheses is rejected, and which is accepted. In all of our statistical inferences, we rejected the null hypothesis at a 5% alpha level.

In this process several statistical tests were used:

i. Shapiro-Wilks test – To assess if the given sample was drawn from a normal distribution [88].
ii. Levene's test – To test if two samples possess equal variances [89].
iii. Two-sample t-test – A parametric test employed to understand if the mean expression value of normally distributed genes is statistically different between the normal sample group and the stage-I primary tumor sample group [90].

iv. Wilcoxon-Mann-Whitney test – A non-parametric test used to test if the median value of a gene, drawn from an unknown or not-normal distribution, is statistically different between the normal sample group and the stage-I primary tumor sample group [91].

v. False discovery rate (FDR) multiple test correction – A method correcting the test-generated p-values to diminish the probability of a type-I error [92].

3.5.1.1 Shapiro-Wilks test: Assessing normality

Many statistical techniques, including parametric tests, are based on the assumption that the given data is taken from a population that follows a normal distribution [88]. Importantly, when this assumption is incorrect, it is impossible to draw accurate conclusions about the population [88].

When dealing with large samples however, comprising more than thirty or forty individuals, assumption of normality is a lesser issue [88]. In fact, with large enough sample sizes, it is possible to use parametric methods, even if the data is not normally distributed. Nonetheless, in our analysis we always tested for normality and our subsequent analysis proceeded in either a parametric or non-parametric manner, depending on the distribution of the data [88].

There are several methods to test normality, however, the Shapiro-Wilk test provides better power than other options [88]. Power is the most important measure of a test for normality, since it dictates the test's ability to detect whether a sample comes from a non-normal distribution. Furthermore, Shapiro-Wilk test has been cited as the best choice for testing normality, and for this reason it was the one employed in this study [88].

The Shapiro-Wilk test determines the correlation between the data and their corresponding normal scores [88]. A significant test statistic (an alpha of 5% was used in this study) leads to rejection of the hypothesis that the data is normally distributed.

To perform the Shapiro-Wilk test, the *shapiro.test()* R function from the "stats" package was used.

3.5.1.2 Levene's test: Comparing sample variance

Levene's test was employed to understand if the variance of a given gene in the normal (non-tumor) group was equal to the variance of the same gene in the stage-I primary tumor group.

This test was implemented in genes that were found to have a normal distribution, and its purpose was to examine if the genes with normal distribution were to be submitted to a two-sample unpaired Student's t-test (if the normal and tumoral groups have equal variances), or to two-sample t-test with the Welch approximation to the degrees of freedom (if the normal and tumoral samples had unequal variances) [89].

In Levene's test, H_0 states that the variances between the two groups are equal, whereas H_1 states the variances are not equal [89]. In the present study, H_0 was rejected if p-value was lower than a fixed alpha of 5%.

To perform the Levene's test, the *leveneTest()* R function from the "car" package was used.

3.5.1.3 Two-sample t-test

A t-test is a statistical test that compares the means between two groups [90]. It belongs to a branch of statistical inference methods deemed as parametric. In parametric techniques, the probability distribution of probability variables is defined, and inferences about the distribution's parameters are made [90].

In general, two types of t-tests exist: 1) the independent t-test, used when the two groups are independent of each other, and 2) the paired t-test, used when the two groups are dependent. In our analysis, an independent t-test was applied [90].

The calculation of the t statistic is based in the assumption that the samples are drawn from a population that displays a normal distribution and have an equal variance [90]. In cases in which our variables displayed unequal variances between normal and stage-I primary tumor groups (determined by Levene's test), the t-statistic was calculated differently, using the degrees of freedom calculated by the Welch Satterthwaite equation.

To perform the Two-sample t-test test, the *t.test()* R function from the "stats" package was used.

3.5.1.4 Wilcoxon-Mann-Whitney test

In the situations where the probability distribution cannot be defined, nonparametric methods are employed [91]. Nonparametric methods are part of a second branch of statistical inference that require little to no assumptions about the data to be examined [91].

The nonparametric alternative to the parametric unpaired t-test is the Wilcoxon rank sum test (also referred as the Mann-Whitney test) [91]. Unlike its parametric homologue, this test does not assume a normal distribution, however it does assume that the groups to be compared are independent. The Wilcoxon-Mann-Whitney test is a technique to examine the difference between the medians of two independent populations [91]. In the present study, the null hypothesis (H_0) was that the median expression value of a given gene was the same in the normal sample group and the stage-I primary tumor group. We considered a given gene to be differentially expressed between normal and stage-I primary tumor samples when this hypothesis was rejected, with a fixed alpha value of 5%.

To perform the Wilcoxon-Mann-Whitney test, the *wilcox.test()* R function from the "stats" package was used.

3.5.1.5 FDR multiple test correction

In the statistical inference process described above, all of the previous tests quantify the probability of a given result (for example, that a gene is differentially expressed in tumoral tissue relative to normal tissue) being significant as a result of mere random chance, given that the null hypothesis is true [92]. In this study we consider that there is enough evidence to reject H_0, when a p-value is lower the 0.05, which would signify a more extreme event, being H_0 true. However, a multiplicity problem, common to many genomic studies, arises since, by pure chance, at 5% level of significance it is expected that one in every twenty genes will appear to be significant [92].

Multiple hypothesis testing automatically inflates the probability of committing a type-I error [92]. At a 5% alpha level, the probability of committing a type-I error ($\alpha_{\text{single-error}}$), while testing a single gene, is $\alpha_{\text{single-error}} = 0.05$, on the other hand, the probability of not committing a type-I error is $1 - \alpha_{\text{single-error}} = 0.95$ (Fig. 3.3) [92]. However, in this part of our analysis we simultaneously tested approximately 20.000 genes. With this amount of multiplicity, the

probability of not committing a type-I error is $(1- \alpha)^{20.000} \approx 2.97x10^{446}$, thus the probability of committing a type-I error in our analysis is $\alpha_{multiple\text{-}error} = 1-(1-\alpha)^{20.000} \approx 1$.

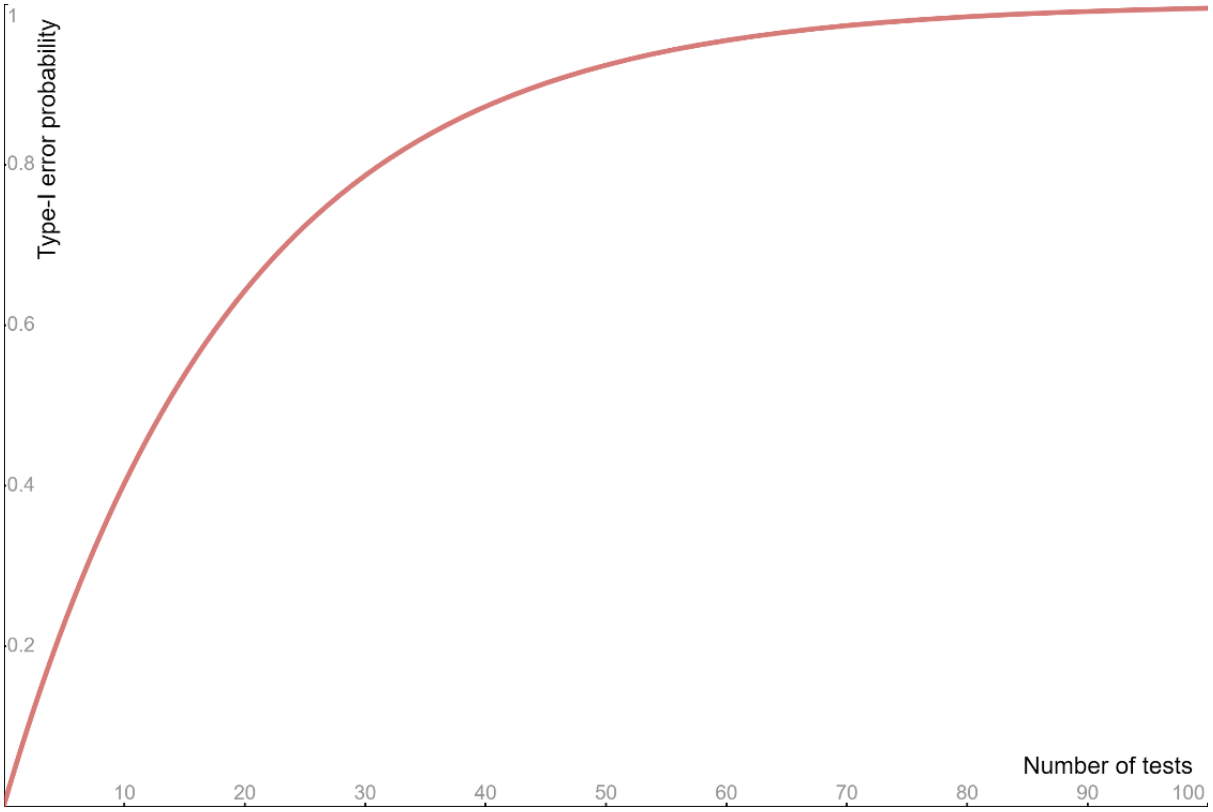

Figure 3.3 – Type-I error probability in function of the number of independent tests, at a 5% significance level.

In order to avoid an exponential false positive rate inflation, we applied a False Discovery Rate (FDR) adjustment method (known as The Benjamini and Hochberg adjustment approach). FDR is the expected proportion of false positives among all positives where the null hypothesis was rejected [92].

$$FDR = Expected\left(\frac{False\ Positives}{False\ Positives + True\ Positives}\right)$$

In this methodology, all p-values are ranked in an ascending order, and multiplied by m/k, where m is the number of p-values to be adjusted and k corresponds to the position where that same p-value is located in the sorted vector [92].

To perform the FDR p-value adjustment methodology, the *p.adjust()* R function from the "stats" package was used.

3.5.2 Identifying genes to differentiate normal and stage-I tumor groups

Having identified the differentially expressed genes, between normal tissue and stage-I primary tumor tissue, in all cancer datasets, our next aim was to understand if the tumorigenic alterations in gene expression could be explained, or predicted, by the previously identified Epi-Hotspots.

Our methodology involved applying a multiple regression technique to each differentially expressed gene and each Epi-Hotspot. The goal was to understand which of the Epi-Hotspot regions might explain gene expression alterations during tumor initiation.

To minimize the number of false positives in this analysis we opted only to analyze genes whose expression variability was primarily associated with the normal to stage-I tumor transition. Many genes labeled as differentially expressed between normal and stage-I tumor tissue actually have high levels of expression variability in each of the groups. This could instigate serious type-I errors in our analysis, because even if an Epi-Hotspot could explain a gene's variability, it would not be clear if this corresponded to the normal to stage-I tumor transition.

We therefore divided the differentially-expressed genes into two groups: the first group of genes were those whose variability was primarily associated with the transition from normal to stage-I tumor, while the second group of genes were those whose expression values were not able to clearly differentiate stage-I tumor tissue from normal tissue. Only genes whose variability was located primarily in the tumorigenic transition were subjected to the Epi-Hotspot – Gene Prediction analysis.

To identify which of the differentially expressed genes were good differentiators of normal and stage-I tumor tissue, we performed Receiver operating characteristic (ROC) curve analysis for each gene. Those genes with an area under the ROC curve (AUC) greater than 0.8 were used in the subsequent analysis.

3.5.2.1 ROC curves and AUC

The ROC curve is a plot in which the y-axis corresponds to the sensitivity of a given test, and the x-axis represents the false positive rate (or 1 – sensitivity) of the same test [93]. ROC curve analysis is a common and effective method to evaluate the quality and performance of a

given differentiator test, such as a diagnostic test. In the present study, the ROC curve methodology was used to understand which differentially expressed genes were good differentiators of normal tissue and stage-I tumor tissue. The ROC curve functions on the concept of a "separator" scale, where the values for the "cases" and "controls" form a pair of overlapping distributions. If the two distributions are completely separated, then the test is perfectly discriminant. Conversely, overlapping distributions signify that the test is not able to discriminate cases from controls [93].

Visually, increasingly discriminant tests are progressively closer to the upper left-hand corner of the "ROC space" [93]. A non-discriminant test results in a diagonal ROC curve and implies a test with a performance equal to chance (a test that will randomly yield a positive or negative result, in a manner that is unrelated to underlying disease status) [93].

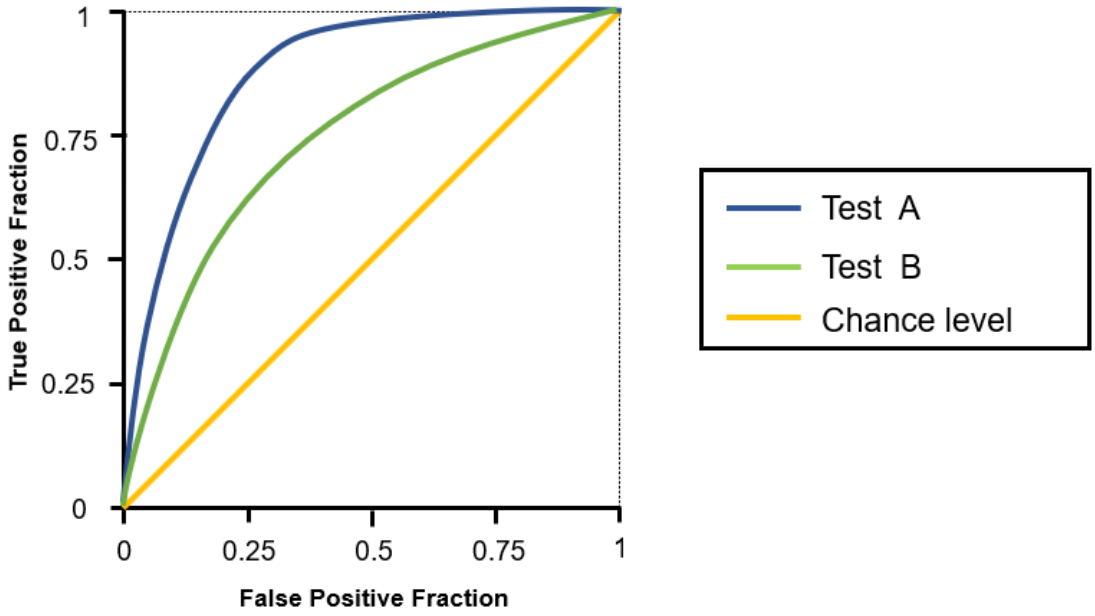

Figure 3.4 - Illustration of a roc curve for two mock tests and chance level. The variables specificity and 1-sensitivity are represented in the y and x axis, respectively. Test-A has a greater AUC than Test-B, and both are greater than chance level, where the AUC is 0.5.

The ROC curve can be summarized by the area under the curve (AUC), which denotes the entire area under the ROC curve [93]. The AUC is commonly utilized as unidimensional measure of combined sensitivity and specificity of a given test. The AUC value can be understood as the probability that a randomly chosen case observation will be identified as a "positive" relative to a randomly chosen control. The AUC value can also be interpreted as the average sensitivity for all possible values of specificity. This index ranges from 0 to 1, where

the maximum (AUC = 1) indicates a test is able to perfectly discriminate two groups, and the minimum (AUC = 0) implies that the test will incorrectly classify all observations. An AUC of 0 is extremely unlikely, since it would mean that a test would be incorrect at all times, which in a way also makes it correct at all times. The diagonal ROC curve, that indicates randomness sorting, corresponds to an AUC of 0.5 [93].

In the present study, only differentially expressed genes showing an AUC equal or greater than 0.8 for differentiating normal and stage-I tumor samples were subjected to further analysis.

To calculate the AUC index, the *roc()* R function from the "pROC" package was used.

3.5.3 Multiple Linear Regression Analysis

After identifying which genes were good differentiators (AUC ≥ 0.8), of normal and stage-I tumor samples we sought to understand which alterations in gene expression could be explained by aberrantly methylated regions, i.e. Epi-Hotspots.

To do this we performed a multiple linear regression analysis between each Epi-Hotspot and each differentially expressed gene.

Simple linear regression is based on the concept that one predictor variable X is used to model a response variable Y [94–96]. However, in our case, one Epi-Hotspot is a collection of multiple CpG sites, each bearing its own beta-value. Thus, one Epi-Hotspot consists of multiple factors on which a single response variable (the gene expression value) depends [94–96]. Our analysis therefore modeled how the variability of one gene, in the normal to stage-I tumor transition, depended linearly on the methylation variability of the CpG sites located in each Epi-Hotspot.

The multiple linear regression model, where gene expression is deemed as a response variable, and the beta-values of each CpG site within an Epi-Hotspot are predictor variables, is written as follows [94–96]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_k X_k + \varepsilon_k.$$

Where Y represents the response variable, here the expression of a given gene, which depends on k predictor values X_1, X_2, X_3, …, X_k, where k corresponds to the number of CpG sites in an Epi-Hotspot, and X is the corresponding beta-value of each CpG site. ε is the residual

term of the model. Note that in the equation above, β_0, β_1, β_2, β_3, …, β_k, do not represent beta-values, but rather the regression coefficients in the model [94–96].

Fitting a multiple linear regression is, almost always, possible [94–96]. However, it is of utmost importance to understand if the model is well fitted. It is also good practice to determine whether all the predictor variables in the model are necessary, since it is better to have a model with fewer predictors and the same explanatory power. In this analysis, however, we aimed to understand whether the entire Epi-Hotspot was a good predictor of gene expression behavior. For this reason we did not reduce the number of predictor variables. Furthermore, in this study, such reduction would almost always lead to models with single explanatory variables, which would hinder our goal of analyzing regions [94–96].

For each multiple linear regression model, we also performed significance testing of the given regression, employing the F-test [94–96]. This test determines whether any of the CpG predictor variables in the model have some relation with the response (in this case gene expression variation). The null hypothesis for the F-test states that all the predictor coefficients in the model are equal to zero, whereas the alternative hypothesis states that at least one coefficient is not zero. If the value of the F-test statistic is large enough, it can be concluded that the model is well-fitted, since at least one predictor in the model is relevant to the response [94–96]. In this study, we considered a model to be statistically significant if the F-test yielded a p-value lower than 0.05.

Another way to understand whether a given Epi-Hotspot is a good predictor of gene expression variation is using R^2. In simple linear regression, R^2 represents the square of the correlation coefficient between the single predictor variable and the response variable [94–96]. In multiple linear regression, R^2 represents the proportion of variation in the response that can be explained through regression of all the predictors, i.e. how well can the variation in the expression of a given gene be explained through the variation of all of the CpG sites in a given Epi-Hotspot [94–96]. Here, we considered that a gene's expression variability was well explained by an Epi-hotspot if R^2 (which varies between 0 and 1) was greater than 0.7.

To perform the multiple linear regression and the respective F-test we applied the *lm()* and *pf()* R functions, both available in the "stats" package.

3.6 Literature search

After identifying which genes were 1) differentially expressed between normal and stage-I primary tumor tissue, 2) good discriminators of both sample types, and 3) related to Epi-Hotspot regions, we sought to review if and how many times these genes were cited in the literature, and whether they were cited in a cancer-related way.

We used the OncoScore tool, which is available as an open-source tool and an R package [97]. This tool is able to query a given gene in the biomedical PubMed literature, and count the number of times such gene was mentioned in the database [97]. Additionally, we performed our own PubMed queries to determine the number of citations of each gene in the cancer literature and in the specific's disease literature.

3.7 Gene Ontology: Functional analysis

Having identified individual genes whose expression during the normal to stage-I tumor transition correlated with changes in DNA methylation, we next sought to understand which gene functions might be involved. We therefore performed a Gene Ontology analysis using the previously identified Epi-Hotspot related genes. The goal was to gather further insights about which cellular mechanisms could be putatively altered by Epi-Hotspots.

Although many strategies exist to perform functional genomic analysis, we chose to perform a gene ontology analysis, since it provides the most information regarding gene function [98]. The gene ontology resource is more than 20 years old and is constantly evolving and gathering new scientific information as it is released [99]. This methodology is based on a knowledge database (or "knowledgebase") organized with formal ontology, where classes of gene functions are defined (here referred as terms) with specific relations to each other [98,99].

In the gene ontology database, more than 45,000 terms, or classes of gene function, exist, all linked by approximately 134,000 relations [99]. Ontology terms are generally classified in three major groups: 1) cellular components, related to where the product of a given gene is located and in which cellular structures it acts, 2) molecular function, describing what activity the gene product performs at the molecular level, and 3) biological process, representing the cellular program or process in which the gene product participates [99].

This analysis was performed using a type of Gene set analysis (GSA) named Generally Applicable Gene-set Enrichment (GAGE) [98]. GAGE has been described in the literature as yielding more reliable results in comparison to other GSA methods [98].

GSA is a very common way to analyze gene expression data using the knowledge of cellular pathways or gene sets [98]. It is important to understand that, to its core, GSA utilizes full genomic information, and it is designed to detect pathway or gene set enrichment using all available genes [98]. However, in our study, this tool was used with the intent of understanding which ontology terms were enriched by our differentially expressed and Epi-Hotspot related genes, which was why only these genes were used in the present analysis.

The GAGE method employs a two-sample t-test to determine whether a given gene set (in this case defined by a gene ontology term) is significantly differentially expressed relative to the full background set of genes [98]. In this study, the background consisted only of differentially expressed genes related to epi-hotspots.

To perform this methodology, we employed the *gage()* R function, from the "gage" package, and considered a gene set (or term) as statistically significant if the resultant p-value was lower than 0.05.

3.8 Epi-Blackholes

In addition to understanding which genomic regions were consistently altered in the normal to stage-I tumor transition (Epi-Hotspots), we also aimed to identify regions that retained the same methylation statuses during this oncogenic transition.

In this analysis, we employed the same strategy was we used in Epi-Hotspot identification, but in an inverted way. However, instead of applying both DMRcate and BumpHunter algorithms, and intersecting the results, we only utilized the DMRcate methodology. The first clustering step used in the BumpHunter algorithm made this algorithm unreliable in this inverted search strategy, as it, would always favor regions with only one CpG site. In order to identify these "Epi-Blackholes" we identified regions that were unlikely to be altered, using a reversed alpha value ($1 - \alpha$) of 0.95%. If the resultant p-value was higher than 0.95, we considered that there was enough evidence to sustain the algorithm's null hypothesis, rather than the alternative hypothesis. In a sense, we overturned the statistical inference process

previously used, and tried to not accept the alternative hypothesis instead of rejecting the null hypothesis.

3.9 Putative genetic mechanism of Epi-Blackholes

In the same way that Epi-Hotspots may be related to differential gene expression, Epi-Blackholes could be, putatively, related to genes that do not change expression during tumor initiation.

To perform our analysis we hypothesized that if the basal variation of a Epi-Blackhole was a predictor of a related gene's basal variation, then one of the variables might prevent change of the other variable or, alternatively, that both variables might be affected in the same way by a third party.

To assess if the identified Epi-Blackholes might be determinants of gene expression, we performed a multiple linear regression analysis, using the same parameters described above. The difference here being that the predictor variables were the CpG sites located inside each Epi-Blackhole, and the response variable was the expression value of each non-differentially expressed gene.

The resultant non-differentially expressed and Epi-Blackhole related genes were subsequently subjected to functional analysis and count of literature citations as described above.

3.10 Epi-Hotspots and Blackholes as potential prognostic biomarkers in stage-III patients

We hypothesized that the identified Epi-hotspots and blackholes, which are potentially relevant for tumor initiation, might also provide insight into patient survival in a more advanced phase of disease. We therefore proceeded to study the impact of each previously identified Epi-Blackhole and epi-hotspot on survival of stage-III cancer patients.

The stage-III patient cohorts were collected from the same previously mentioned TCGA consortium and prepared applying the same tools and techniques described above. For each Epi-Blackhole and each hotspot we applied a multivariate Cox proportional-hazards model

analysis, to assess if the group of CpG sites of each given region could, as a whole, be associated with survival of the stage-III patients.

3.10.1 Multivariate Cox proportional-hazards model analysis

The Cox model is the most widely used technique for analysis of survival time in medical and biomedical research, in a multivariate way [100,101]. This approach is used to study the association between survival time and one (bivariate) or more (multivariate) predictor variables [100,101].

The choice for this method, in the present study, was due to the need of investigating the impact of the methylation status of a group of CpG sites, as a group, in patient survival.

The Cox model simply depicts the interaction between the event incidence (i.e. death), and a set of covariates (or predictor variables) [100,101]. This model can be expressed by what is called as the hazard function, h(t), which is fundamentally the probability that a given individual experiences an event (death) in a certain time point (t). This can be mathematically described as follow [100,101]:

$$h(t) = h_0(t) \times \exp\{B_1X_1 + B_2X_2 + \cdots + B_pX_p\}$$

where the hazard function h(t) is determined by a group of p predictor variables. The predictor variables do not contribute equally to the hazard function, instead each predictor variable has a specific impact size (or weight), and this is measured by each respective coefficient (B_1, B_2, …, B_p). The baseline hazard, h_0, corresponds to the value of the hazard when all covariates are equal to zero [100,101].

It is important to note that although certain covariates can have a much greater impact on the hazard function than others, the model is not fitted to only the most impactful and/or univariately significant predictor variables [100,101]. This is because a certain covariate can have a small overall prediction ability but can be an important contributor to the prediction power of a group of covariates [100,101].

As previously described, each covariate has its own weight on survival prediction, which can be quantified by the covariate's coefficient [100,101]. Each of these coefficients can be exponentiated to obtain hazard ratios (HR), which are in turn used to assess the impact of each

covariate on survival [100,101]. HR's are indicators of a covariate's impact on survival that can be intuitively interpreted, since there are only three distinct HR ranges [100,101]:

1. An HR greater that one (or a covariate's coefficient greater than zero) reveals that as the value of its respective covariate increases, so does the event hazard, i.e. death. These predictor variables are positively associated with decreased survival times and are commonly referred as poor (or bad) prognostic factors.
2. An HR lower than one (or a covariate's coefficient lower than zero) indicates that its respective covariate is negatively associated with the event probability, and thus positively associated with survival length. These types of predictor variables are classified as good prognostic factors.
3. Finally, an HR equal to one (or a covariate's coefficient equal than zero) implies that the covariate is not associated with survival [100,101].

To perform the multivariate Cox proportional-hazards model analysis we employed two publicly available R packages, the "survival" package for computation of the survival models, and the "survminer" package for subsequent graphical visualization.

After fitting each Cox model to each set of CpG sites, we evaluated the statistical significance of each model. To do this, we performed three different statistical tests, 1) the Wald test, or Z-test, 2) the likelihood ratio test, and 3) the score test [100,101]. These three tests all evaluate the null hypothesis that all of the analyzed covariate's coefficients are equal to zero [100,101]. We rejected the null hypothesis when each respective p-value was lower than 0.05. The model for a set of CpG sites was considered a good fit and statistically significant if all of three null hypotheses were rejected.

It is important to note that, as the name implies, the Cox proportional-hazards model makes the fundamental assumption that the hazards of the groups of individuals are proportional [100,101].

Due to its own nature, the cox model only makes sense under the assumption that the hazards are proportional [100,101]. In other words, the event hazard in a group must be, over time, a constant multiple of the event hazard of another group. If this assumption is ignored no real survival information can be drawn from the model [100,101].

To test if each model violated this assumption, we performed the Schoenfeld residuals test. This method tests whether the model's residuals are independent of time, i.e. if they

emerge randomly [102]. The Schoenfeld residuals are simply the difference between the covariate's value of a failed individual and the expected model value [102]. To perform the Schoenfeld residuals test we employed the *coxph()* R function, from the "survival" package, and considered that the proportional-hazards assumption was not violated when the Schoenfeld residuals test was neither globally statistically significant, with an alpha fixed at 5%, nor statistically significant for each individual covariate.

3.10.2 Partitioning of patients based on risk

Having identified which epi-hotspots and epi-blackholes were able to predict survival at stage-III, we aimed to understand which of these regions could split stage-III patients into two distinct risk groups: high hazard vs low hazard. We considered that these regions would be even better predictors of survival and would provide a more practical indicator of patient risk.

In each of the previously generated models the patients were divided into two groups: 1) high hazard and 2) low hazard. This was done by calculating a hazard score for each patient as follows:

$$hazard\ score = \exp\left\{B_1X_1 + B_2X_2 + \cdots + B_pX_p\right\}$$

where B1, B2, …, Bp are the previously computed coefficients of each respective X CpG site.

3.10.2.1 Maximally Selected Rank Statistics

To identify the hazard score threshold for which the patients are divided, a Maximally Selected Rank Statistics (Maxstat) method was applied [103]. This methodology computes a standardized two-sample linear rank statistic between the two groups of observations generated for every possible cutpoint. The maximum of the standardized statistics is considered to be the value that provides the best possible separation between both groups [103].

This method was performed using the *surv_cutpoint()* R function, from the "survminer" package.

3.10.2.2 Survival analysis between risk groups

After identifying the optimal hazard score cutpoint to segregate patients into different risk groups, we sought to confirm whether the high hazard group had a lower survival prognosis that the low hazard group. To do so, we generated Kaplan Meier estimators and compared the curves from both groups. To assess if the survival distributions from both groups were significantly different, we performed either a 1) Log rank test, if the survival curves were non-crossing, or 2) a two-stage test, if the survival curves crossed at least one time. In both tests an alpha of 5% was used to reject the null hypothesis.

3.10.2.2.1 Log-Rank test

The log rank test is one of the most widely used methods to compare survival between groups by testing the null hypothesis that there is no between-group difference in the probability of death at any point in time [104,105].

Essentially, this methodology compares the total number of observed deaths in both groups with the total number of expected deaths, were the null hypothesis to be true [104]. The log-rank test was performed using the *surv_pvalue*() R function in the "survminer" package.

It is important to note that the log-rank test functions optimally under the assumption of proportional hazard rates, and it has been demonstrated that this test loses power when the survival curves of the compared groups cross [104,105]. For this reason, when comparing groups in which survival curves crossed, we utilized a two-stage test. This choice was based on a study published in 2015 by Li and colleagues that extensively evaluated several methodologies to compare survival in crossed survival curves [106]. The two-stage test maintained a robust power and a low type-I error rate in various simulations [106]. In both log-rank and two-stage tests a fixed alpha of 5% was used.

3.10.2.2.2 Two-stage test

As the name implies, the two-stage test consists of two distinct phases [107]. In the first phase, a conventional log-rank test is performed, and if this test yields a statistically significant output it is possible to conclude that survival in the two groups is significantly different. If not, the second phase of the process is performed. In this phase, a certain weight is chosen for a

weighted log-rank test, which change signs before and after a potential crossing point. The weight is defined in such a way that the test statistics of the two stages are independent, allowing for a redefinition of the resultant p-value [107]. The two-stage test was performed using the *twostage*() R function, available in the "TSHRC" package.

To confirm that the differences in survival were not due to differences in patient age, we also performed a Wilcoxon-Mann-Whitney test to assess whether the median age between the high and low risk groups was statistically significantly different, at an alpha of 5%.

Of all the regions that could successfully differentiate two stage-III patient risk groups, we only considered potential prognostic biomarker regions as the ones that generated clusters with a high-to-low hazard ratio between 0.6 and 1.67 (1/0.6).

Chapter 4 Results

4.1 Data Preparation

It is a well-known fact that tumoral cells display aberrant DNA methylation patterns in comparison to non-tumoral cells [77]. The goal of the present study was to understand whether specific genomic regions have DNA methylation patterns that are consistently altered in certain tumors and whether these epigenetic modifications are related to altered patterns of gene expression.

The first step in our analysis involved the cleaning of the datasets to prepare them for further analysis. This process was composed of three main steps: 1) detecting and removing outlier data-points, 2) handling missing data, and 3) calibrating age and gender.

All of the steps mentioned above involved either variable or observation removal, as explained in the methodology section. For this reason, data cleaning, although making the data more suitable for analysis, also reduced the data available for analysis.

The number of variables and observations available for analysis after data cleaning are displayed in table 4.1.

Table 4.1 - Sample sizes and variables for individual cohorts.

	Gene Expression datasets			**Methylation datasets**			
Cohort	Number of Genes	Number of Normal patients	Number of stage-I Tumor patients	Number of CpG probes	Number of Normal patients	Number of tumor patients	
						Stage-I	Stage-III
COAD	19657	41	78	396048	38	47	87
PAAD	19657	4	21	396005	10	21	4
BRCA	19657	113	182	394738	96	127	199
CHOL	19657	9	19	394084	9	19	1
ESCA	19657	11	16	396055	16	18	56
HNSC	19657	44	25	396053	46	27	82
KIRC	19657	72	268	396063	157	157	73
KIRP	19657	32	172	396055	45	168	51
LIHC	19657	50	171	396054	50	175	86
LUAD	19657	59	284	395858	32	257	73
LUSC	19657	49	244	396060	42	172	56
THCA	19657	58	281	395795	49	171	113

4.2 Epi-hotspot identification

After the data was prepared for analysis, we next searched within each of the twelve datasets for genomic regions that were differentially methylated between normal and stage-I tumoral samples. This analysis was achieved by intersecting the outputs of two DMR searching methods: the DMRCate and Bumphunter algorithms. The complete lists of epi-hotspots for each dataset are available in Annex II through Annex XIII.

The number of epi-hotspots differed between cohorts, ranging from 2 epi-hotspots in esophageal cancer to 208 in liver cancer (Table 4.2). The size of the identified regions also varied greatly within each cohort. In fact, the standard deviation to the mean size of an epi-hotspot is considerably large in every cohort. For example, in the colon adenocarcinoma cohort (COAD), the shortest epi-hotspot was 5 bp long, while the longest was almost 2.9 kb. The number of CpG sites in each region also varied greatly. The smallest number of analyzed CpG sites in a region was usually three since this was one of our epi-hotspot selection criteria (described in the methods section). As shown in Table 4.2, the epi-hotspot regions also comprised a very small portion of the entire methylome, from as little as 0.006% to 0.37%.

Table 4.2 - Summary description of the identified Epi-hotspots. For each analyzed disease (first column), the number of identified hotspots, the average epi-hotspot size (Avg Size) and standard deviation, the sizes of the longest (Max Size) and shortest (Min Size) identified epi-hotspots, the average number CpG sites present in an epi-hotspot with standard deviation (Avg CpG number), the minimum and maximum CpG counts in a region (Min Cgs, Max Cgs), and the percentage of visible methylome corresponding to epi-hotspots (Methylome %).

Cohort	Epi-hotspot number	Avg Size (bp)	Max Size (bp)	Min Size (bp)	Avg cg number	Min Cgs	Max Cgs	Methylome %
COAD	75	744.4± 496.9	2898	5	11.6 ± 8.1	3	51	0.22
PAAD	3	677± 609.9	1381	308	11.3 ± 4.7	6	15	0.009
BRCA	76	380.7± 302	1193	6	6.8 ± 5.7	3	32	0.132
CHOL	70	639.1± 364.6	1748	97	9.6 ± 3.9	5	26	0.171
ESCA	2	752.5± 181.7	881	624	12 ± 2.8	10	14	0.006
HNSC	203	409.4± 292	1573	15	6.9 ± 4.7	3	34	0.354
KIRC	8	568.6± 268.4	1073	206	9.9 ± 3.9	6	18	0.02
KIRP	29	392.7± 343.4	1624	37	6.7 ± 4	3	18	0.049
LIHC	208	465.6± 449.3	3258	7	7.2 ± 6.1	3	54	0.377
LUAD	117	526.3± 460.1	3090	22	8.8 ± 7.3	3	52	0.259
LUSC	97	544.9± 452.1	3238	61	8.3 ± 6.8	3	53	0.203
THCA	7	356± 391.2	1142	36	8.7 ± 9.1	3	29	0.015

The resultant data may suggest that the number of epi-hotspots is greater in cohorts with larger sample size. However, this premise was tested, and we concluded that the number of identified epi-hotspots is not correlated, in a statistically significant manner, with 1) the number of normal patients in the dataset (Pearson's test r = -0.08, p-value = 0.8), 2) the number of tumoral patients in the dataset (Pearson's test r = 0.14, p-value = 0.7), and 3) the number of total patients in the sample (Pearson's test r = 0.08, p-value = 0.8). Not only do the p-values in the correlation tests show that these are undoubtedly not statistically significant, but the correlation's coefficients are extremely close to zero, suggesting that the variables are not dependent on each other.

Even though there is not a statistically significant correlation between sample size and the number of identified epi-hotspots, certain cohorts with very small sample size, such as PAAD and ESCA, may not provide enough statistical power to identify the true number of epi-hotspot in such diseases. On the other hand, cohorts such as KIRC and THCA have sample sizes that provide enough statistical power, but a very low epi-hotspot count (eight and seven, respectively). This may suggest that these diseases (or cell types) may in fact have a lower number of clustered methylation alterations in the normal to stage-I tumoral transition.

Next, we analyzed the genomic locations of the epi-hotspots using the annotation data provided for the Illumina 450k array [79]. We observed that the vast majority of Epi-hotspots (>60% in all datasets) occurred within promoter regions (Fig. 4.1).

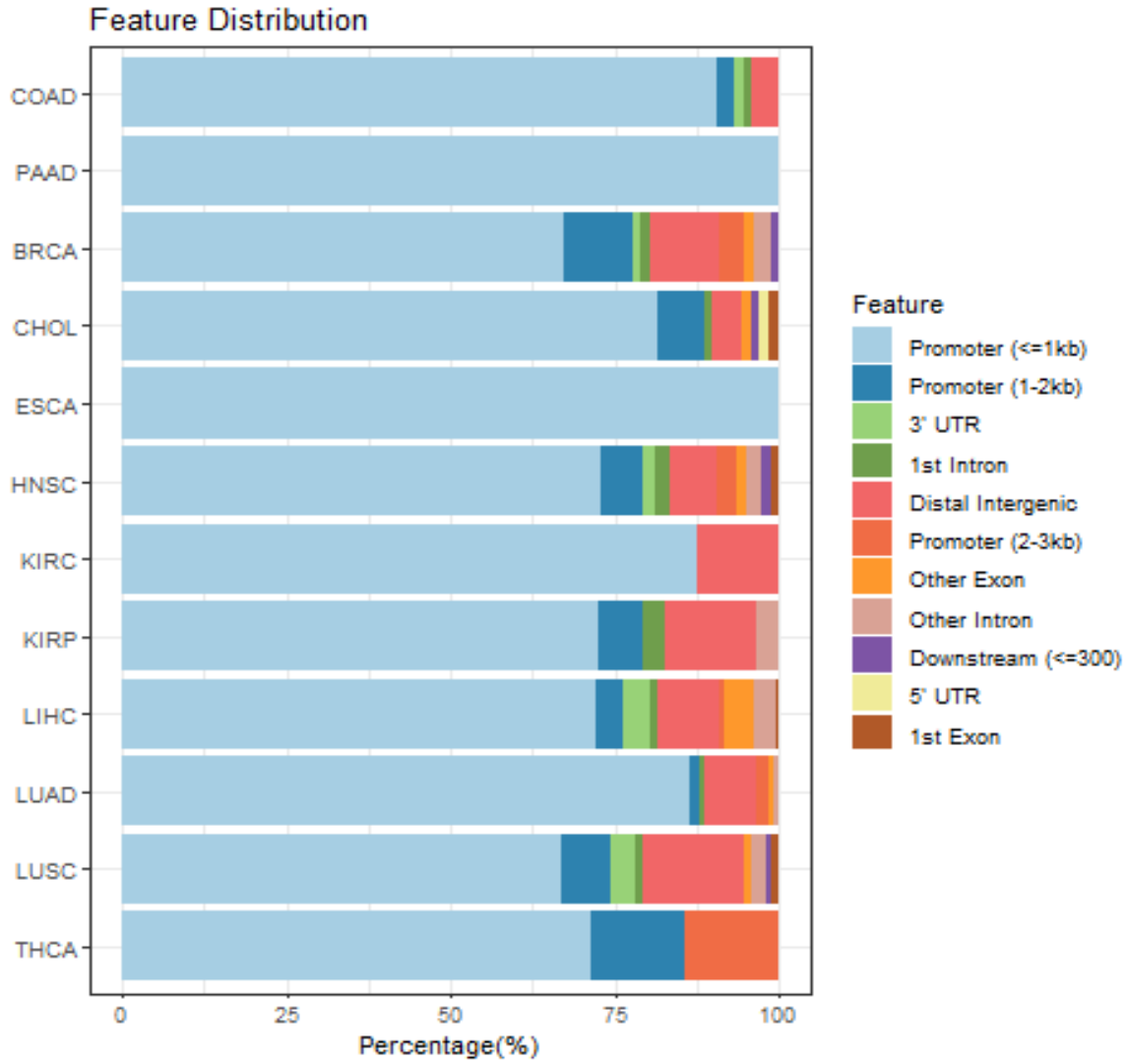

Figure 4.1 - Distribution of epi-hotspots across the genome. For each dataset (horizontal bars), the percentage of epi-hotspots that are located in each genomic region is represented in the x axis. In every cohort most of epi-hotspots are located in promoter regions.

While the fact that epi-hotspots occur mainly in promoter regions may have biological significance for gene expression, it is important to note that CpG sites exist in higher numbers within promoters, which could explain the higher number of epi-hotspots in these regions.

We also analyzed the distribution of Epi-hotspots relative to the genome. We noted that Epi-hotspot distribution varies across datasets. Nonetheless, certain patterns emerged, such as the presence of epi-hotspots on chromosome 6 in every cohort except ESCA (data shown in the sections 4.7.1 through 4.7.12).

4.3 Epi-hotspots and altered gene expression patterns

One of our goals was to understand if Epi-hotspots could be, in some way, associated with altered patterns of gene expression during the normal to stage-I tumor transition. We first sought to find out which genes were differentially expressed between normal and stage-I tumor patients in the twelve examined diseases. Using gene expression data accompanying each dataset, we first determined which genes showed differential expression. This was achieved by performing the series of statistical hypothesis tests and methodologies described in the Methods section.

Since our goal was to understand if Epi-Hotspots could predict gene expression changes during tumor initiation, we restricted our analysis to genes that presented an AUC higher than 0.8 for distinguishing normal and stage-I tumor samples. This restriction does not signify that genes with a lower AUC could not be associated with epi-hotspots. However, their inclusion introduced significant noise in the output data. The differentially expressed genes were then subjected to multiple linear regression analysis to test whether the Epi-hotspots predicted gene expression alterations in tumor initiation. The number of differentially expressed genes that were associated with epi-hotspots varied across cohorts. In some datasets, like CHOL, COAD, and HNSC, most of the detected differentially expressed genes were associated with epi-hotspot variation. In contrast, in the BRCA, ESCA, KIRP, LUSC, and THCA datasets, most of the differentially expressed genes were not associated with epi-hotspots. The number of differentially expressed genes that were associated with epi-hotspots relative to the number of differentially expressed genes that were not associated with epi-hotspots is represented in Figure 4.2. A PubMed query was also executed for each individual gene, in each cohort, to understand if it was already cited in the literature. The complete list of differentially expressed genes that were associated with epi-hotspot variation is available, for each cohort, in the Annexes XIV through XXII.

Epi-hotspots that showed correlation with gene expression changes were subsequently subjected to gene ontology analysis to identify groups of genes and pathways showing possible epigenetic regulation. The results from the gene ontology analysis and literature search are individually described for each dataset in the section 4.7.1 through 4.7.12. The complete list of enriched gene sets from this analysis are available in Annexes XXIII through XXX.

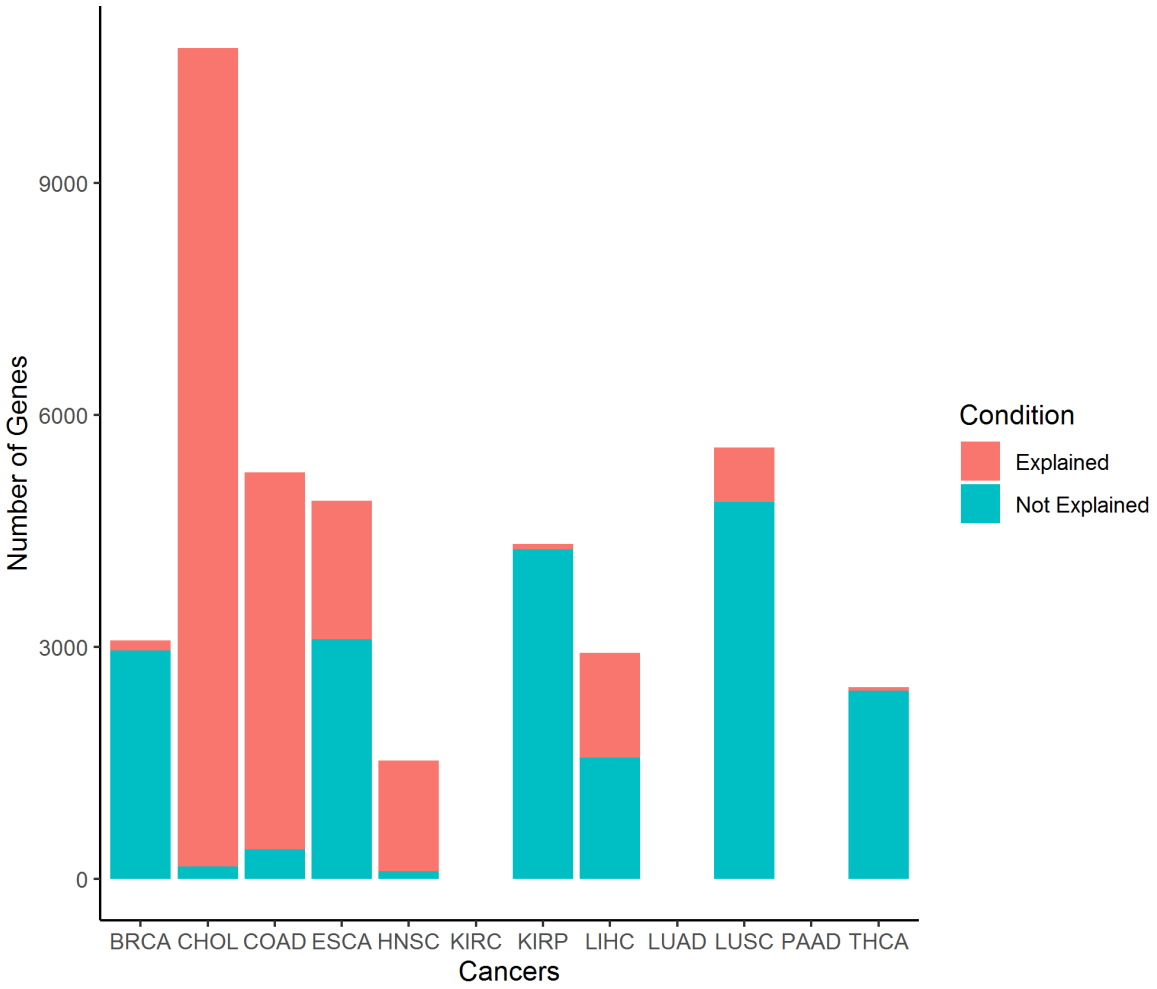

Figure 4.2 - Number of differentially expressed genes between normal and stage-I tumor tissue, in each cohort. The number genes whose variation can be explained by epi-hotspots shown in red and the number of genes whose variability cannot be explained shown in blue.

4.4 Epi-Blackhole identification

After identification and analysis of Epi-hotspots in all twelve cohorts, we next searched for genomic regions that showed minimal DNA methylation change between normal and stage-I tumors. To identify these regions, referred to here as Epi-Blackholes, we applied the same methodology used to identify Epi-hotspot regions. However, as described in the methods section, we only employed the DMRcate algorithm. We used an alpha greater than 95% as evidence to sustain the algorithm's null hypothesis that the DNA methylation pattern was not significantly changed between normal tissue and stage-I tumors.

Similar to the Epi-hotspot analysis, the number of identified Epi-Blackholes was different in each cohort, ranging from 3 Epi-Blackholes in the renal clear cell cancer (KIRC) dataset to 2455 epi-blackholes in cholangiocarcinoma (CHOL) (Table 4.3). The size of these

regions also varied greatly between the analyzed cohorts. In general, however, they were far larger than the epi-hotspot regions (Table 4.3). The exception was the renal clear cell cancer cohort, in which the average size of an Epi-Blackhole was smaller than the epi-hotspots. The average number of CpGs per Epi-Blackhole varied between cohorts, although this variation was not large, ranging from 3.7 in the KIRC cohort to 10 in the ESCA cohort (Table 4.3). As shown in the Table 4.3, Epi-Blackhole regions also represented a very small portion of the visible methylome. Although this was true across all cohorts, the esophageal and bile duct cancers stood out, having 6.1% and 5% of their visible methylomes within Epi-Blackholes. A complete list of the epi-blackholes identified in each dataset is available in the Annexes XXXI through XLII.

Table 4.3 - Summary description of the identified Epi-Blackholes. For each analyzed dataset (first column), the number of identified Epi-Blackholes, the average Epi-Blackhole size (+/- standard deviation), the sizes of the longest and shortest identified Epi-Blackholes, the average number of CpG sites present in the Epi-Blackholes (+/- standard deviation), the minimum and maximum CpG counts in a blackhole, and the percentage of the visible methylome that corresponds to identified Epi-Blackholes in each dataset.

Cohort	Epi-blackhole Number	Avg Size (bp)	Max Size (bp)	Min Size (bp)	Avg cg number	Min Cgs	Max Cgs	Methylome %
COAD	274	1007.1 ± 699.8	5430	12	8.4 ± 4.3	3	27	0.58
PAAD	898	991.1 ± 839.2	8017	19	6.7 ± 4.3	3	37	1.52
LUAD	134	850.6 ± 627.2	3411	44	6.3 ± 3.4	3	18	0.21
KIRC	3	173 ± 87.7	251	78	3.7 ± 1.2	3	5	0
BRCA	9	471.8 ± 448	1363	60	4.2 ± 2.3	3	10	0.01
CHOL	2455	1186.6± 1027.5	12915	18	8 ± 6.8	3	133	5
ESCA	2420	1283.5 ± 928.2	12287	15	10 ± 6.1	3	133	6.1
HNSC	55	636.7 ± 520.1	2983	34	4.8 ± 2.9	3	19	0.07
KIRP	17	400.7 ± 298	1121	72	3.9 ± 1.6	3	9	0.02
LIHC	11	547.5 ± 595.3	2061	82	4 ± 2.4	3	11	0.01
LUSC	18	999.9 ± 953.4	4283	42	9.6 ± 4.6	3	19	0.04
THCA	126	731.7 ± 665.7	4438	27	4.6 ± 2.2	3	16	0.15

To determine whether the number of Epi-Blackholes was dependent on sample size we performed the same statistical analysis described for Epi-hotpots (see methods section). Although the number of identified Epi-Blackholes was not correlated with the number of normal patients in each dataset (Pearson's test r = -0.5, p-value = 0.08), it was significantly negatively correlated with 1) the number of tumor samples (Pearson's test r = -0.6, p-value = 0.02) and 2) the total number of samples (Pearson's test r = -0.7, p-value = 0.009). This suggests that as the number of tumor samples decreases, the number of Epi-Blackholes increases. However, this correlation may be strongly influenced by some cohorts containing a reduced number of samples. Cohorts with low statistical power, such as the PAAD, ESCA, and CHOL datasets, may produce a higher number of false positive Epi-Blackholes, since the chance of accepting a given region as an epi-hotspot is lower. In fact, when we test whether the number of Epi-Blackholes is correlated with sample size in only the large cohorts we conclude that the number of identified regions is not correlated, in a statistically significant manner with 1) the number of normal patients in the dataset (Pearson's test r = -0.45, p-value = 0.23), 2) the number of tumor samples in the dataset (Pearson's test r = -0.27, p-value = 0.49), or 3) the number of total patients in the dataset (Pearson's test r = -0.45, p-value = 0.22) in these larger samples. This suggests that cohorts above a certain size are less susceptible to this bias.

We next analyzed the location of the Epi-Blackholes. We observed that, in most cohorts, the majority of these regions occurred within promoter regions (Fig. 4.3). However, in contrast to epi-hotspots, it was noticeable that Epi-Blackholes overlap at a higher frequency with other genomic regions such as exons and introns.

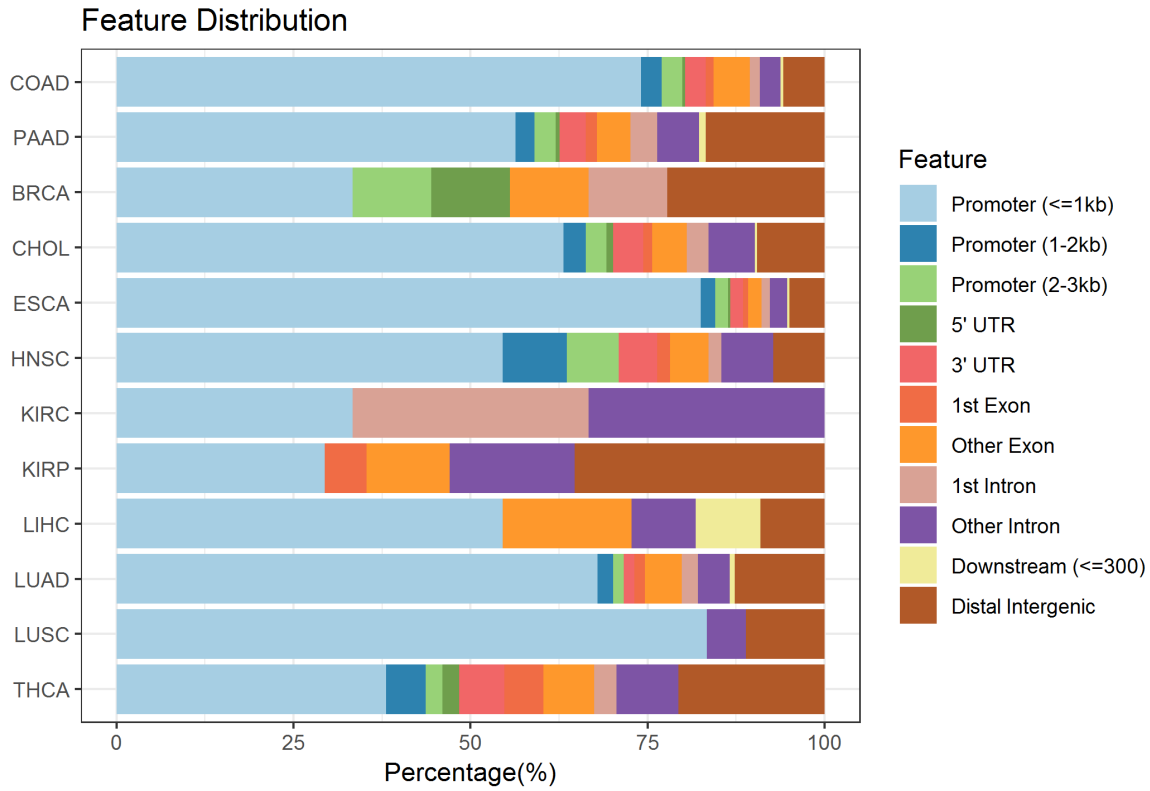

Figure 4.3 - Distribution of epi-blackholes across the genome. For each dataset (horizontal bars), the percentage of epi-hotspots that are located in each genomic region is represented in the x axis. In every cohort most of epi-hotspots are located in promoter regions.

4.5 Epi-Blackholes and altered gene expression patterns

Just as Epi-hotspots were hypothesized to have predictive capabilities regarding gene expression alterations in the normal to stage-I tumor transition, Epi-Blackholes might be correlated with genes that do not change expression during tumor initiation. We hypothesized that if Epi-Blackhole was a predictor of a non-differentially expressed gene's variation, then either one of the variables preventing change in the other, or possibly both variables were being affected by a third unknown factor.

To test whether Epi-Blackholes predict variation of non-differentially expressed genes between normal samples and stage-I tumors, we again utilized multiple linear regression to compare each Epi-Blackhole and each non-differentially expressed gene.

In most cancer cohorts we found that Epi-Blackholes were not strong predictors of variation in the expression of non-differentially expressed genes (Fig. 4.4). Nonetheless, in some cancer types, such as CHOL, ESCA, and PAAD, there was a substantially higher

proportion of genes whose variability could be explained by Epi-Blackholes. It is, however, important to remember that these cancer cohorts are the ones with the lowest sample sizes and the highest number of Epi-Blackholes, which may be a source of bias in this particular analysis. The lists of non-differentially expressed genes that were associated with epi-blackholes are available, for each dataset, in the Annexes XLIII through Annex L.

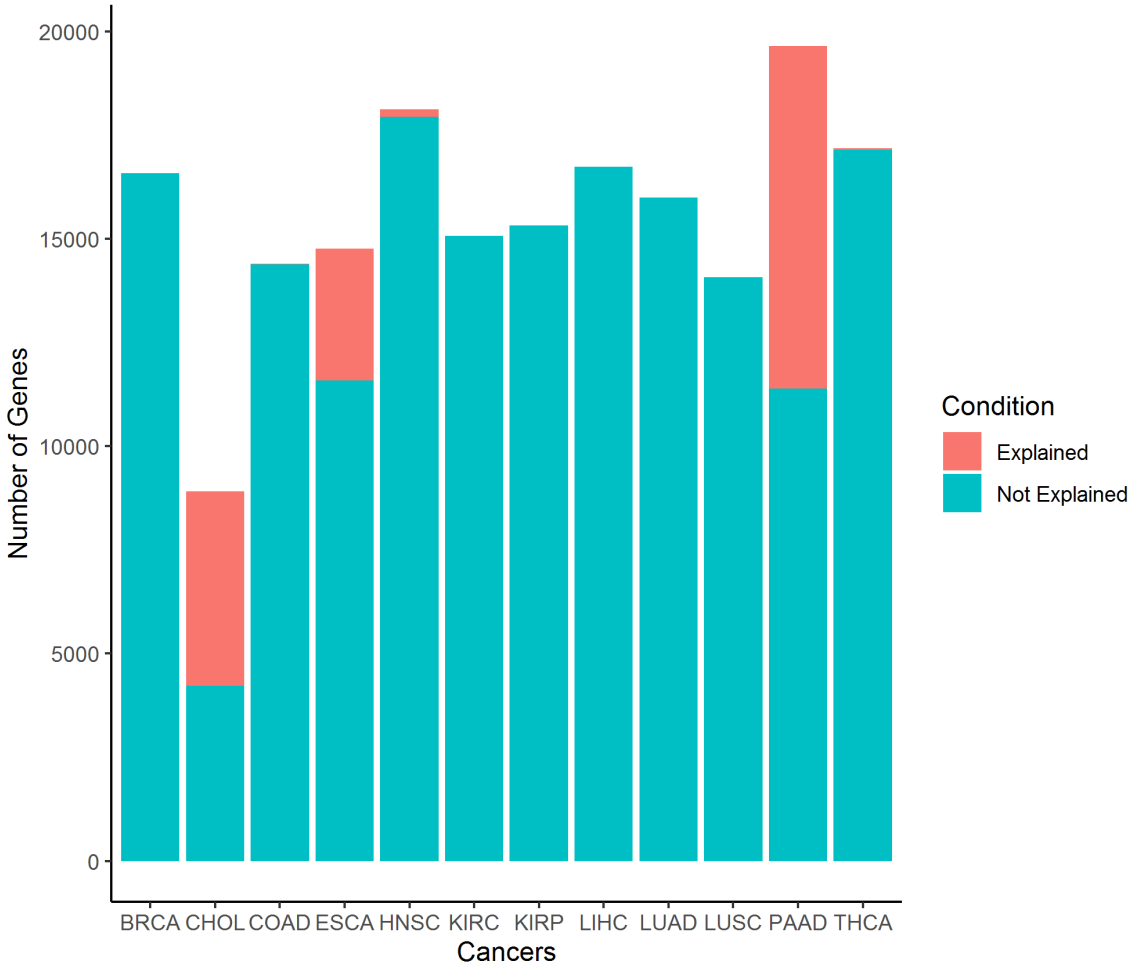

Figure 4.4 - Number of differentially expressed genes between normal and stage-I tumor tissue, in each cohort. With the number genes whose variation can be explained by epi-blackholes (in red) and the number of genes whose variability cannot be explained (in blue).

4.6 Epi-Hotspots and Epi-Blackholes as prognostic biomarkers

Having identified the genomic regions with altered DNA methylation patterns during tumor initiation (Epi-hotspots), as well as the regions that preserve their DNA methylation

status (Epi-Blackholes), we next aimed to understand whether heterogeneity in these regions might correlate with patient survival in a more advanced stage of the disease.

We therefore analyzed survival and DNA methylation data from stage-III patients in the same twelve TCGA cohorts.

For each identified Epi-Hotspot and Epi-Blackhole we individually applied a multivariate Cox proportional-hazards model, using each region's CpG values as predictor variables. All computed models were tested for proportional hazards using the Schoenfeld residuals test, and global significance was tested using the Wald test, the likelihood ratio test, and the score test

For the models that met the previous criteria, a hazard score was calculated as described in the methods section.

The hazard score was then used to discriminate patients into high vs low hazard groups, using the Maxstat method. These two groups were submitted to further survival analysis through the generation of Kaplan Meier estimators and assessment of differential survival by performing either log rank or Two-Stage tests, depending if the curves crossed

In every survival comparison we tested, using a Wilcoxon-Mann-Whitney test, if age was significantly different in both groups, and only considered the genomic regions that were able to discriminate patients into two groups with distinct survival distributions with statistically non-significant age differences. Furthermore, all comparisons for which the proportion of patients in one group to the second group was greater than 60% were discarded.

As shown in Table 4.4, in 8 of the 12 cohorts we found at least one region that was able to predict survival in stage-III patients. These regions included both Epi-hotspots (in 6 cohorts) and Epi-Blackholes (in 5 cohorts). The complete data regarding the epi-hotspots and epi-blackholes that were able to predict prognosis in stage-III cancer patients is available, for each dataset, in the Annexes LI through LVIII.

Table 4.4 - Number of epi-hotspots and epi-blackholes with prognostic potential in stage-III cancer patients. For each analyzed disease (first column), the number of identified epi-hotspots and epi-blackholes that are able to successfully discriminate stage-III patients into two groups with distinct survival distributions are represented.

Cohort	Prognosis Epi-blackholes	Prognosis Epi-Hotspots
BRCA	0	1
COAD	8	1
ESCA	19	0
HNSC	3	8
KIRP	2	0
LIHC	0	1
LUAD	2	6
LUSC	0	1
PAAD	0	0
KIRC	0	0
CHOL	0	0
THCA	0	0

4.7. Summary of results for each studied dataset

4.7.1 Summary of results from Colon Adenocarcinoma

The COAD cohort represents patients with colon cancer. In the COAD cohort we identified 75 epi-hotspots and 274 Epi-Blackholes (Annexes II and XXXI). The location of the identified regions is graphically represented in the Figure 4.5.

We also identified 5253 genes that were differentially expressed between the normal and stage-I tumor groups, of which 4873 (~93%) can be explained by epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XIV). Of these differentially expressed, epi-hotspot correlated genes, 217 have never been cited in the literature, 297 have previously been cited in the non-cancer literature, and 1705 have been cited in the cancer literature, but never in colon adenocarcinoma. By submitting these differentially expressed, epi-hotspot associated genes to a GAGE analysis, we observed that 121 Gene Ontology (GO) terms were up-regulated in the stage-I tumoral group and 457 were down-regulated (Table 4.5, Annex XXIII). By analyzing the five most enriched Go terms for each ontology class, we observed an

upregulation of gene sets related to cell polarity and a downregulation of gene sets related to ion transport (Fig. 4.6).

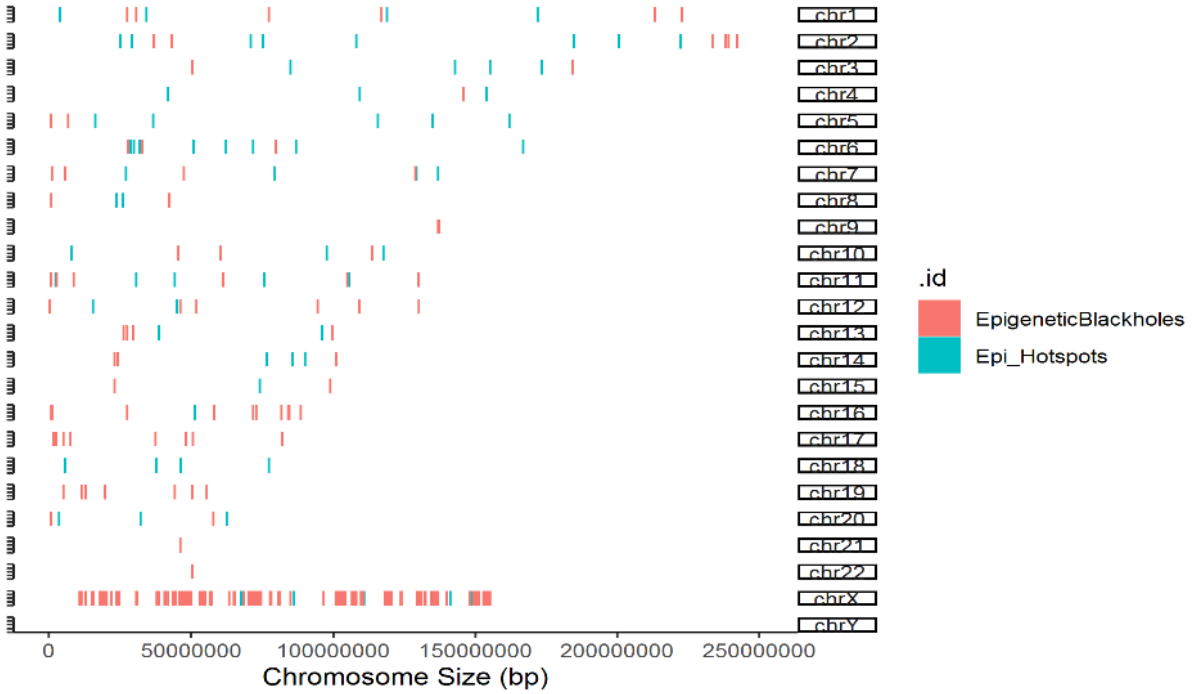

Figure 4.5 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in colon adenocarcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

Table 4.5 – Number of enriched GO terms in stage-I COAD samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	58	331
Cellular Component	28	57
Molecular Function	35	69

The remaining 14404 genes were not differentially expressed between the normal and stage-I tumor groups, of which 16 (~0.11%) are correlated with Blackhole variability (multiple linear regression r-squared ≥ 0.7) (Annex XLIII). Of these non-differentially expressed,

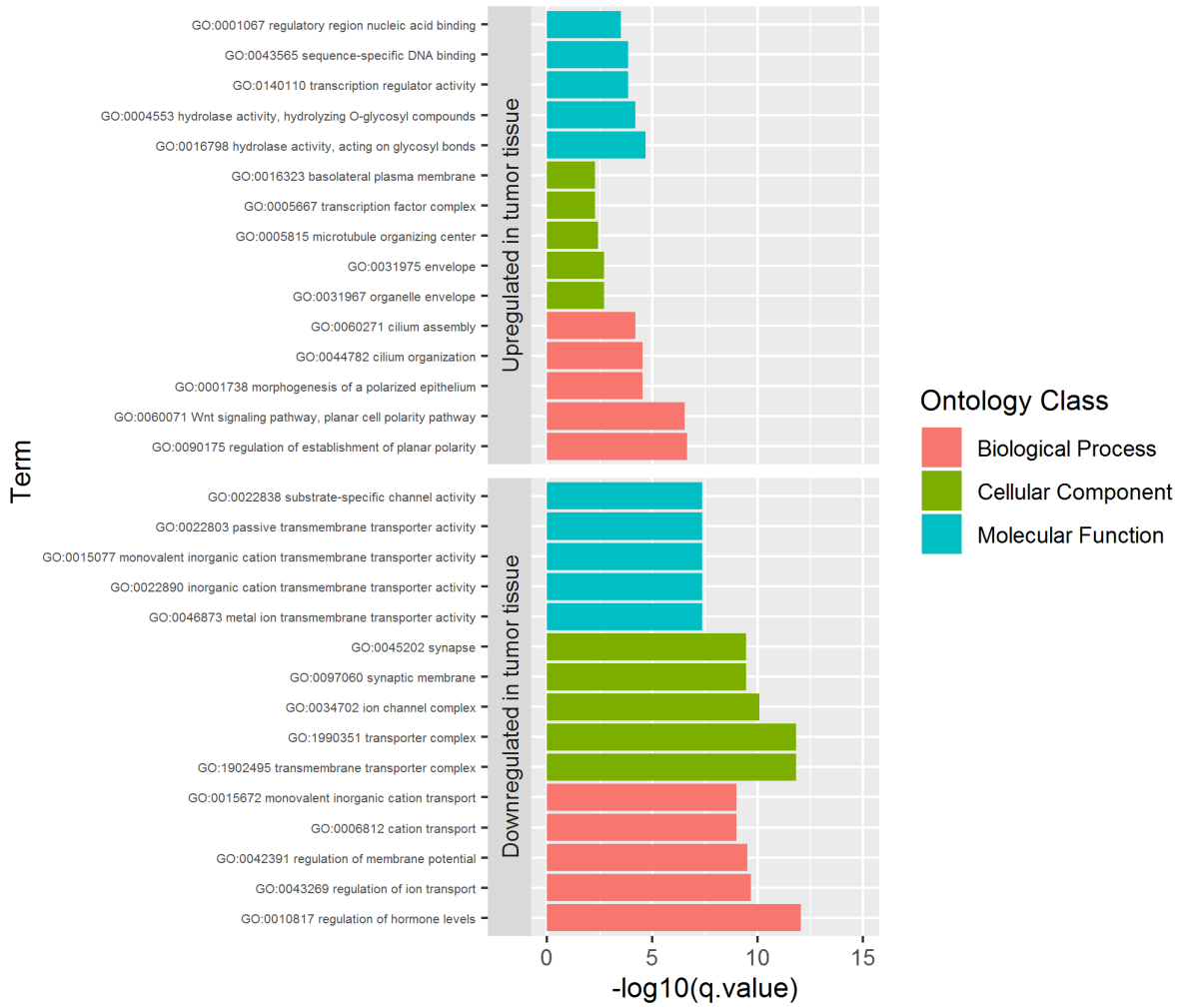

Figure 4.6 – Top five most significantly enriched GO terms in stage-I COAD samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, molecular functions, and cellular components. The upper portion of the chart depicts GO terms that were upregulated in stage-I COAD samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I COAD samples.

blackhole-associated genes, 2 were never cited, 0 were cited in the non-cancer literature, and 10 were cited in cancer, but never in colon adenocarcinoma.

When testing for the ability of epigenetic hotspots and blackholes to predict survival in stage-III colon adenocarcinoma patients, we found that eight Epi-Blackholes and one epi-hotspot were putative prognostic biomarkers (Annex LI). Figures 4.7 and 4.8 below represent one Epi-Blackhole and one epi-hotspot that can differentiate stage-III colon adenocarcinoma patients into two groups with distinct survival distributions. The survival curves for all other predictive regions are in Annex X.

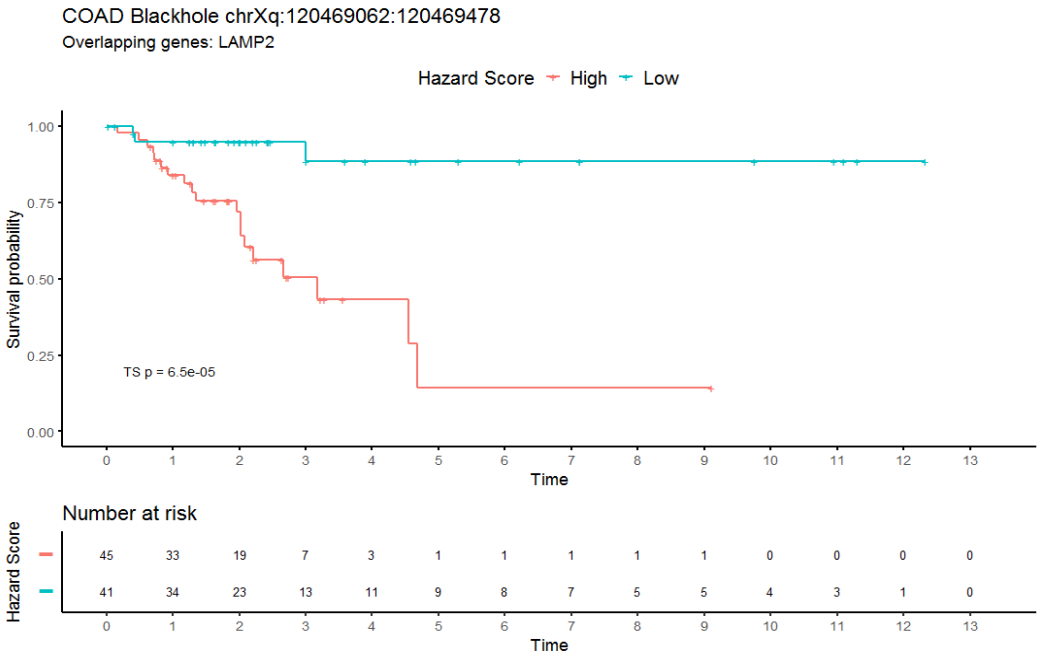

Figure 4.7 - Kaplan-Meier estimator of survival in stage-III COAD for two groups with different epi-blackhole methylation levels. Two groups of stage-III COAD patients with different hazard scores show significantly different survival (p = 0.000065). Statistical significance tested by Two-Stage. Time is represented in

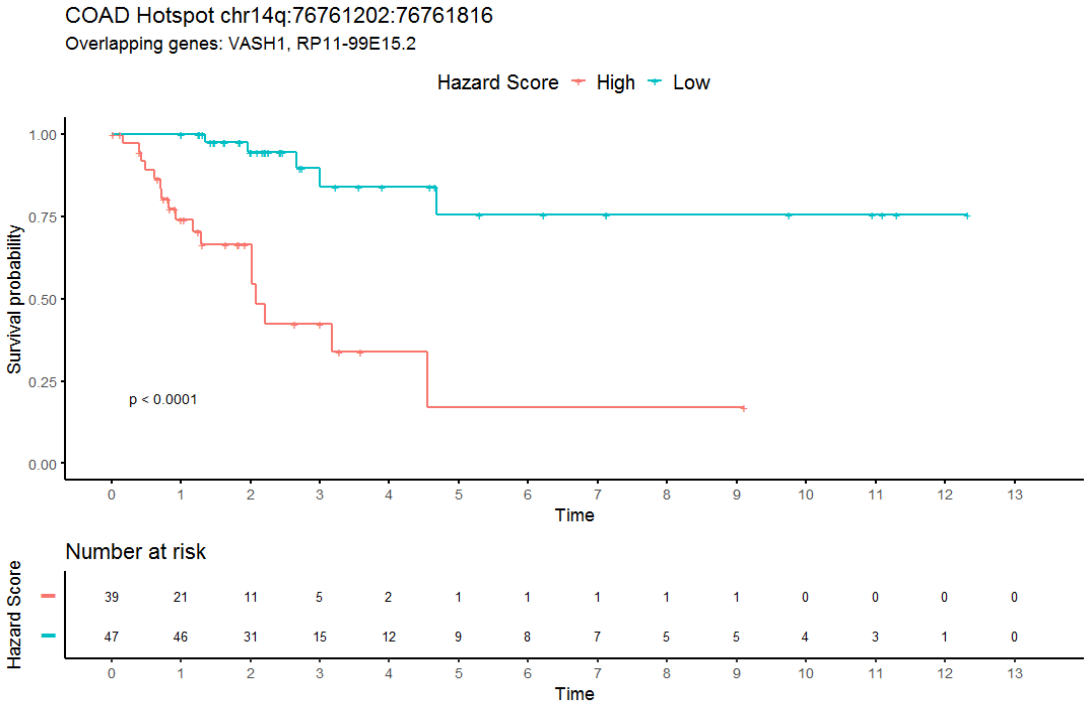

Figure 4.8 - Kaplan-Meier estimator of survival in stage-III COAD for two groups with different epi-hotspot methylation levels. Two groups of stage-III COAD patients with different hazard scores show significantly different survival (p < 0.0001). Statistical significance tested by Two-Stage. Time is represented in years.

4.7.2 Summary of results from Breast Invasive Carcinoma

In the breast invasive carcinoma (BRCA) cohort we identified 76 epi-hotspots and 9 Epi-Blackholes (Annexes III and XXXII). The location of the identified regions is graphically represented in Figure 4.9.

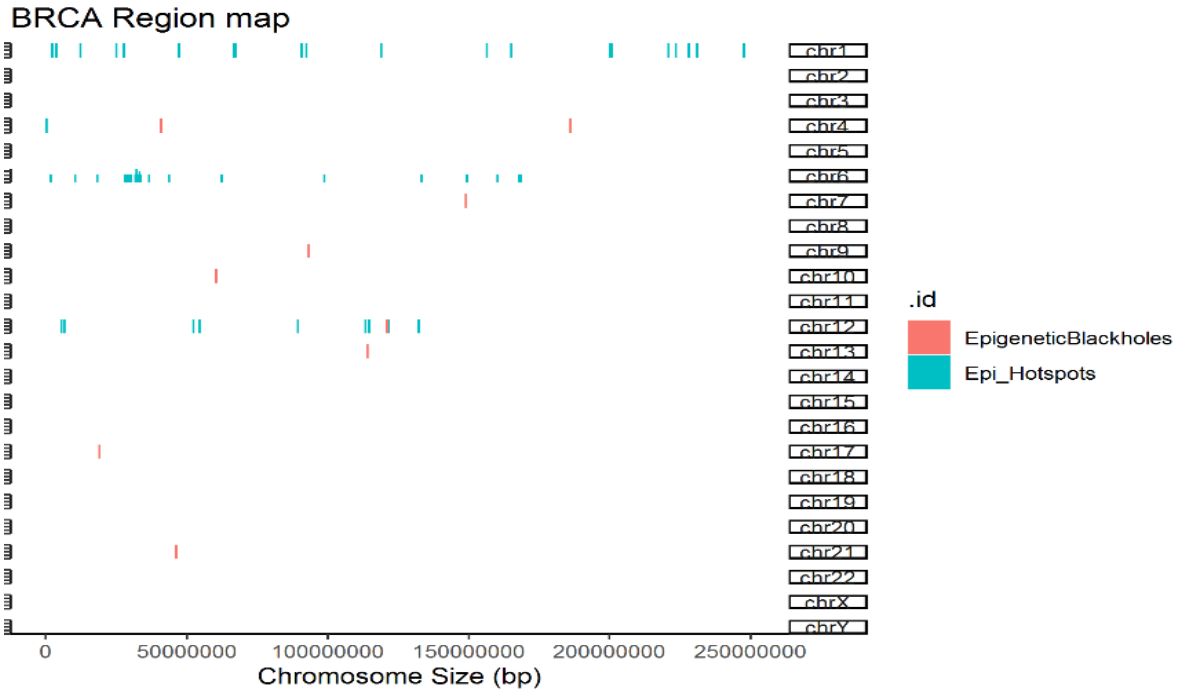

Figure 4.9 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in breast invasive carcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the BRCA cohort we found 3079 genes that were differentially expressed between the normal and stage-I tumor groups, of which 125 (~4%) were correlated with (multiple linear regression r-squared $\geq$ 0.7) (Annex XV). Of these differentially expressed, epi-hotspot associated genes, 5 were never cited in the literature, 8 were cited in the non-cancer literature, and 25 were cited in cancer, but never in breast cancer. By submitting these differentially expressed genes and potentially epigenetically regulated genes to a GAGE analysis, we observed that 24 GO terms were up-regulated in the stage-I tumoral group and 6 were down-regulated (Table 4.6, Annex XXIV). By analyzing the five most enriched Go terms for each ontology class, we observed an upregulation of gene sets related to non-membrane-bound cellular components and a downregulation of GO terms related to the cellular membrane (Fig. 4.10).

Table 4.6 - Number of enriched GO terms in stage-I BRCA samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	24	0
Cellular Component	6	6
Molecular Function	0	0

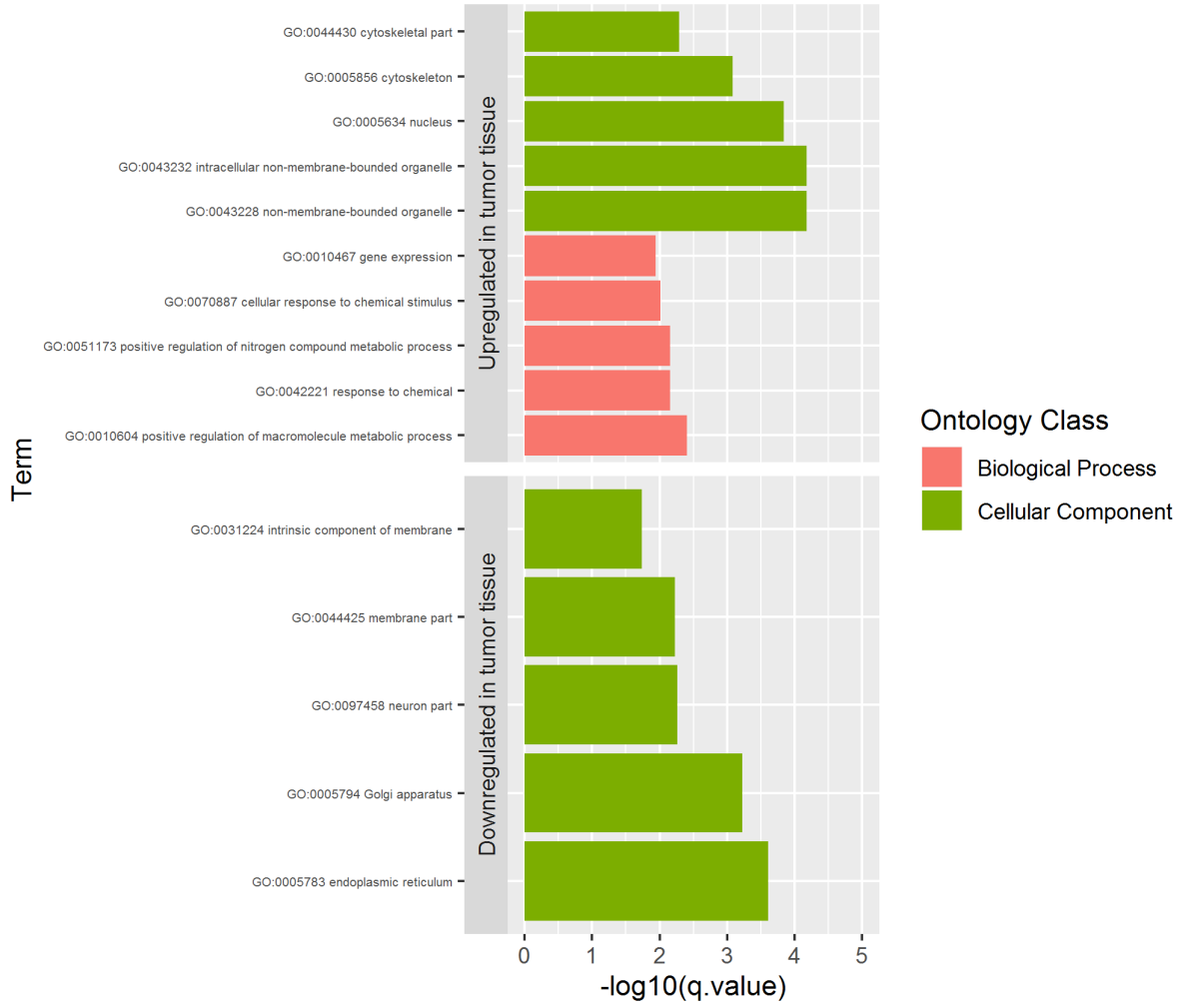

Figure 4.10 - Top five most significantly enriched GO terms in stage-I BRCA samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, and cellular components. No molecular function was significantly enriched. The upper portion of the chart depicts GO terms that were upregulated in stage-I BRCA samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I BRCA samples.

The remaining 16578 genes were not differentially expressed between the normal and stage-I tumor groups, of these none could be explained by Blackhole variability.

When testing for the ability of epigenetic hotspots and blackholes to predict survival in stage-III breast invasive carcinoma patients, we found that one epi-hotspot, was a putative good prognostic biomarker. The survival curve for this epi-hotspot is represented in Figure 4.11. Additional data regarding this region as a prognostic biomarker is available in Annex LII.

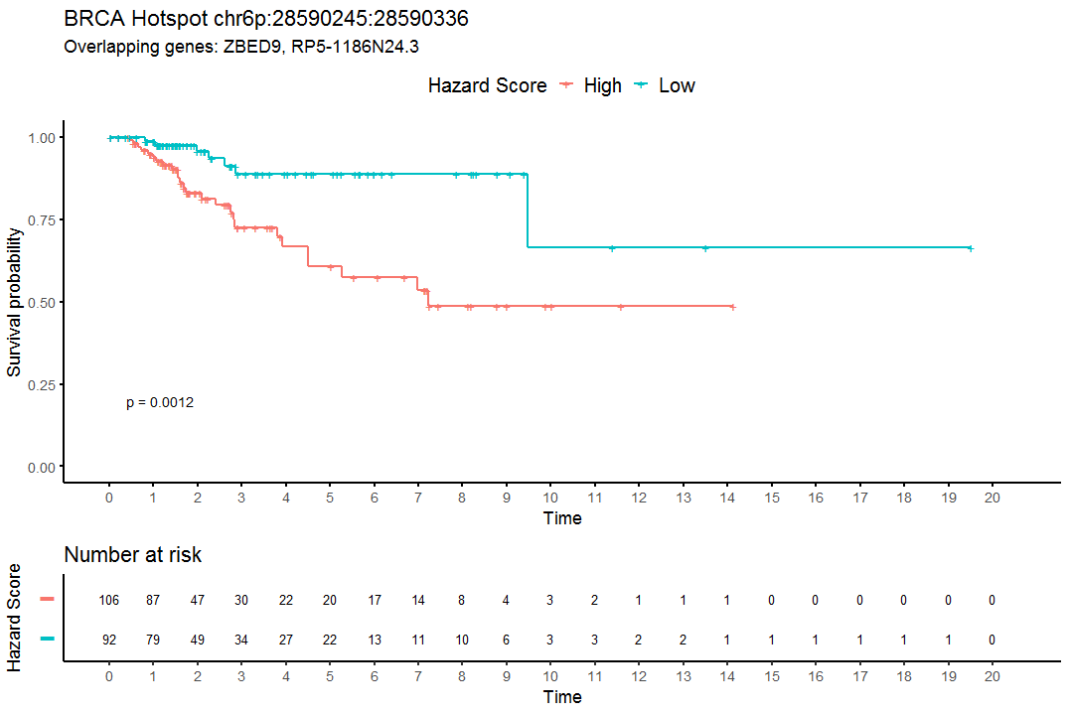

Figure 4.11 - Kaplan-Meier estimator of survival in stage-III BRCA for two groups with different epi-hotspot methylation levels. Two groups of stage-III BRCA patients with different hazard scores show significantly different survival (p = 0.0012). Statistical significance tested by log-rank. Time is represented in years.

4.7.3 Summary of results from Cholangiocarcinoma

In the cholangiocarcinoma (CHOL) cohort we identified 70 epi-hotspots and 2455 Epi-Blackholes (Annexes IV and XXXIII). The location of the identified regions is graphically represented in Figure 4.12.

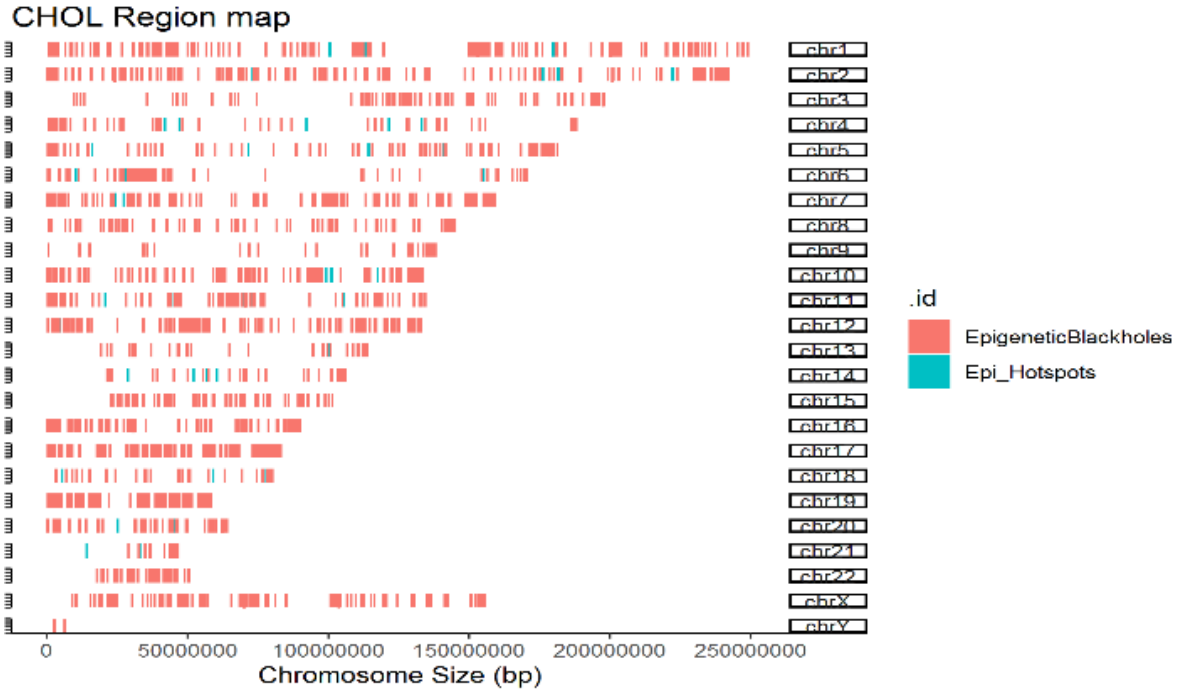

Figure 4.12 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in cholangiocarcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the CHOL cohort we found 10748 differentially expressed genes between the normal and stage-I tumor groups, of which 10591 (~99%) are correlated with epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XVI). Of these differentially expressed, epi-hotspot associated genes, 539 were never cited, 751 were cited in the non-cancer literature, and 8079 were cited in cancer literature, but never in the bile duct cancer literature. When these differentially expressed and epi-hotspot-associated genes were subjected to GAGE analysis, we observed that 321 GO terms were up-regulated in the stage-I tumoral group and 572 were down-regulated (Table 4.7, Annex XXV). By analyzing the five most enriched GO terms for each ontology class, we observed an upregulation of gene sets related to cell division, and a downregulation of gene sets related to ion transport (Fig. 4.13).

Table 4.7 - Number of enriched GO terms in stage-I CHOL samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	244	437
Cellular Component	45	36
Molecular Function	32	99

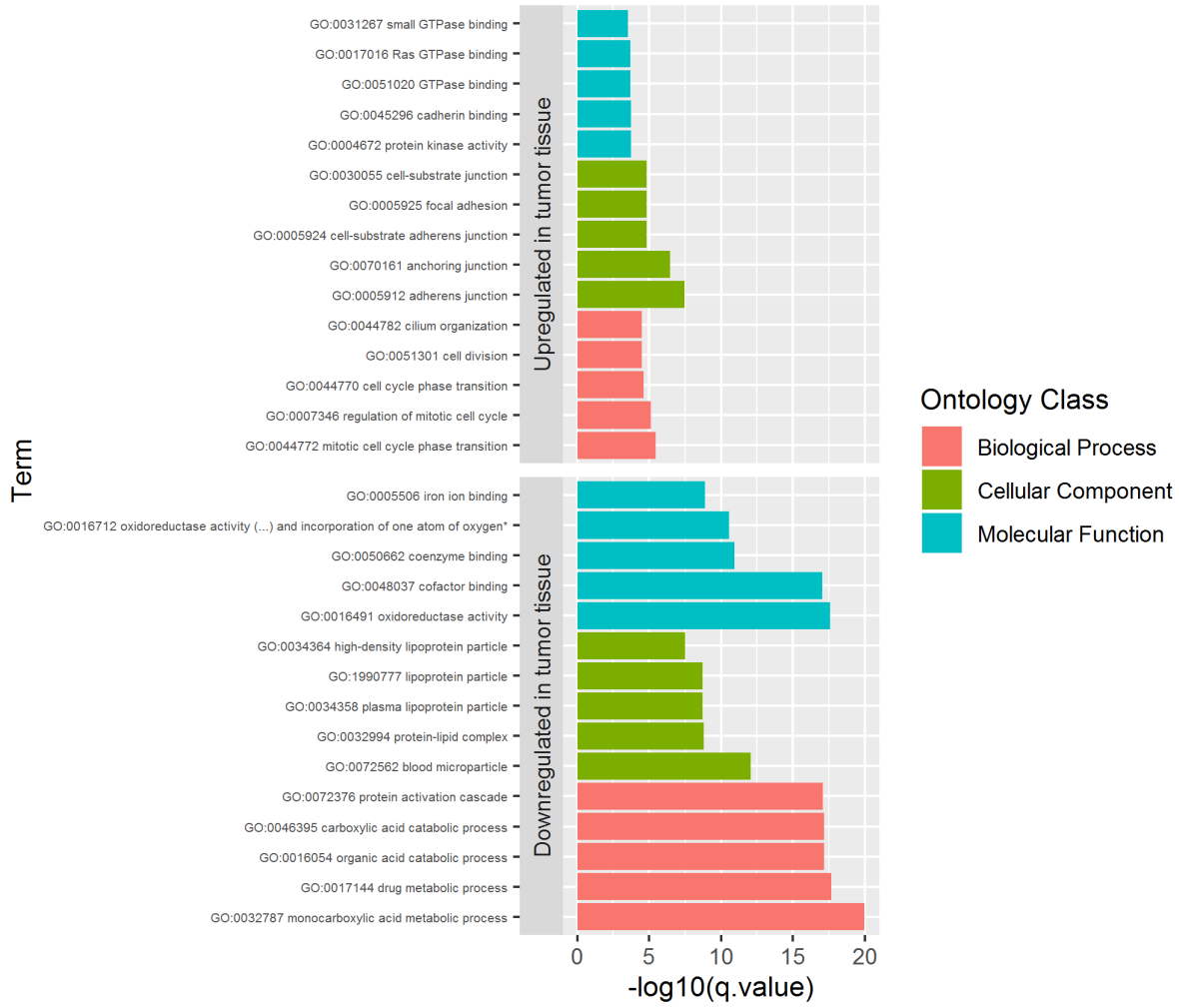

Figure 4.13 - Top five most significantly enriched GO terms in stage-I CHOL samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, molecular functions, and cellular components. The upper portion of the chart depicts GO terms that were upregulated in stage-I CHOL samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I CHOL samples.

The remaining 8909 genes were not differentially expressed between the normal and stage-I tumor groups, of which 4693 (~52.68%) can be explained by epi-blackhole variability (multiple linear regression r-squared ≥ 0.7) (Annex XLIV). Of these non-differentially

expressed, epi-blackhole-associated genes, 531 were never cited, 400 were cited in the non-cancer literature, and 3309 were cited in the cancer literature, but never in the bile duct cancer literature.

Due to the fact that only one stage-III patient had available methylation data in the TCGA dataset, testing whether the identified epi-hotspots and epi-blackholes were able to predict survival in stage-III cholangiocarcinoma patients was not possible.

4.7.4 Summary of results from Esophageal Carcinoma

In the esophageal carcinoma (ESCA) cohort we identified 2 epi-hotspots and 2420 Epi-Blackholes (Annexes V and XXXIV). The location of the identified regions is graphically represented in Figure 4.14.

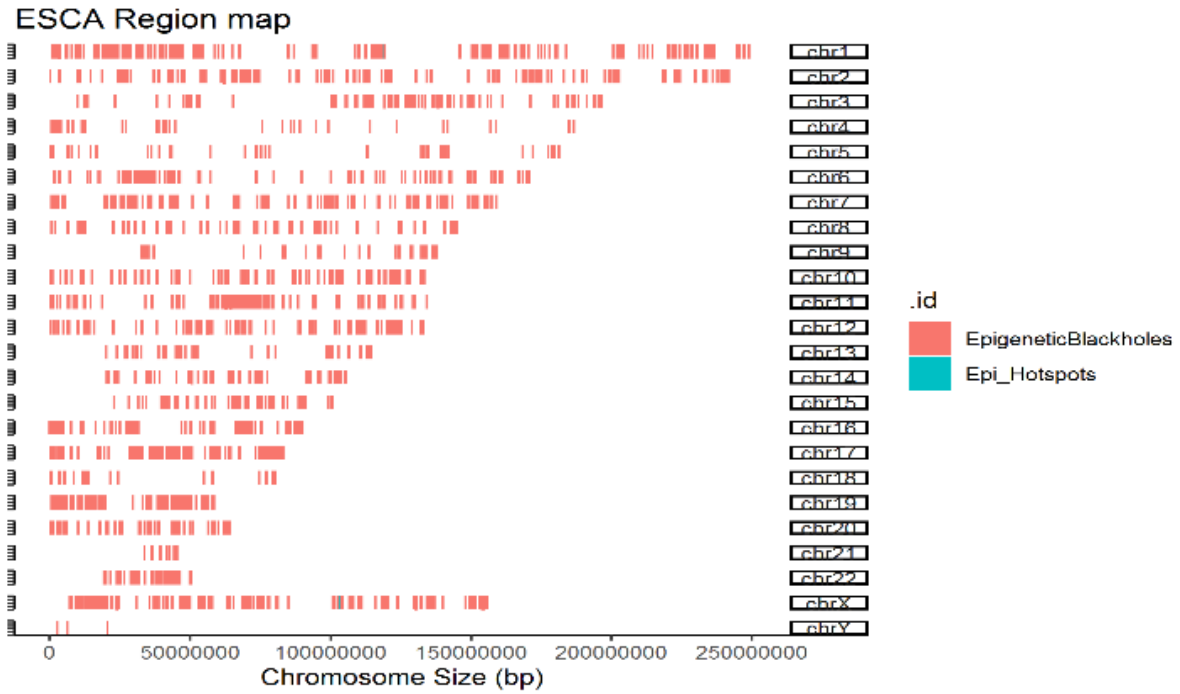

Figure 4.14 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in esophageal cancer. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the ESCA cohort we found 4890 differentially expressed genes between the normal and stage-I tumor groups, of which 1792 (~37%) are associated with variability (multiple linear regression r-squared ≥ 0.7) (Annex XVII**).** Of these differentially expressed epi-hotspot-associated genes, 67 were never cited, 106 were cited in the non-cancer literature, and 1215

were cited in cancer-literature, but never in the esophageal cancer literature. When submitting these differentially expressed and epi-hotspot-associated genes to a GAGE analysis, we observed that 119 GO terms were up-regulated in the stage-I tumoral group and 2 were down-regulated (Table 4.8, Annex XXVI). By analyzing the five most enriched Go terms for each ontology class, we observed an upregulation of gene sets related to cell division and a downregulation of only two gene sets related to ribosomes (Fig. 4.15).

Table 4.8 - Number of enriched GO terms in stage-I ESCA samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	88	0
Cellular Component	24	2
Molecular Function	7	0

The remaining 14767 genes were not differentially expressed between the normal and stage-I tumor groups, of which 3183 (~21.55%) can be explained by epi-blackhole variability (multiple linear regression r-squared $\geq$ 0.7) (Annex XLV). Of these non-differentially expressed, epi-blackhole-associated genes, 215 were never cited, 251 where cited in the non-cancer literature, and 2030 were cited in the cancer literature, but never in the esophageal cancer literature.

When testing whether the identified epi-hotspots and epi-blackholes were able to predict survival in stage-III esophageal cancer patients, we found that nineteen epi-blackholes, were putative prognostic biomarkers. Data regarding these prognosis-predicting regions is available in Annex LIII.

Figure 4.16 represents one epi-blackhole that can differentiate stage-III esophageal carcinoma patients into two groups with distinct survival distributions.

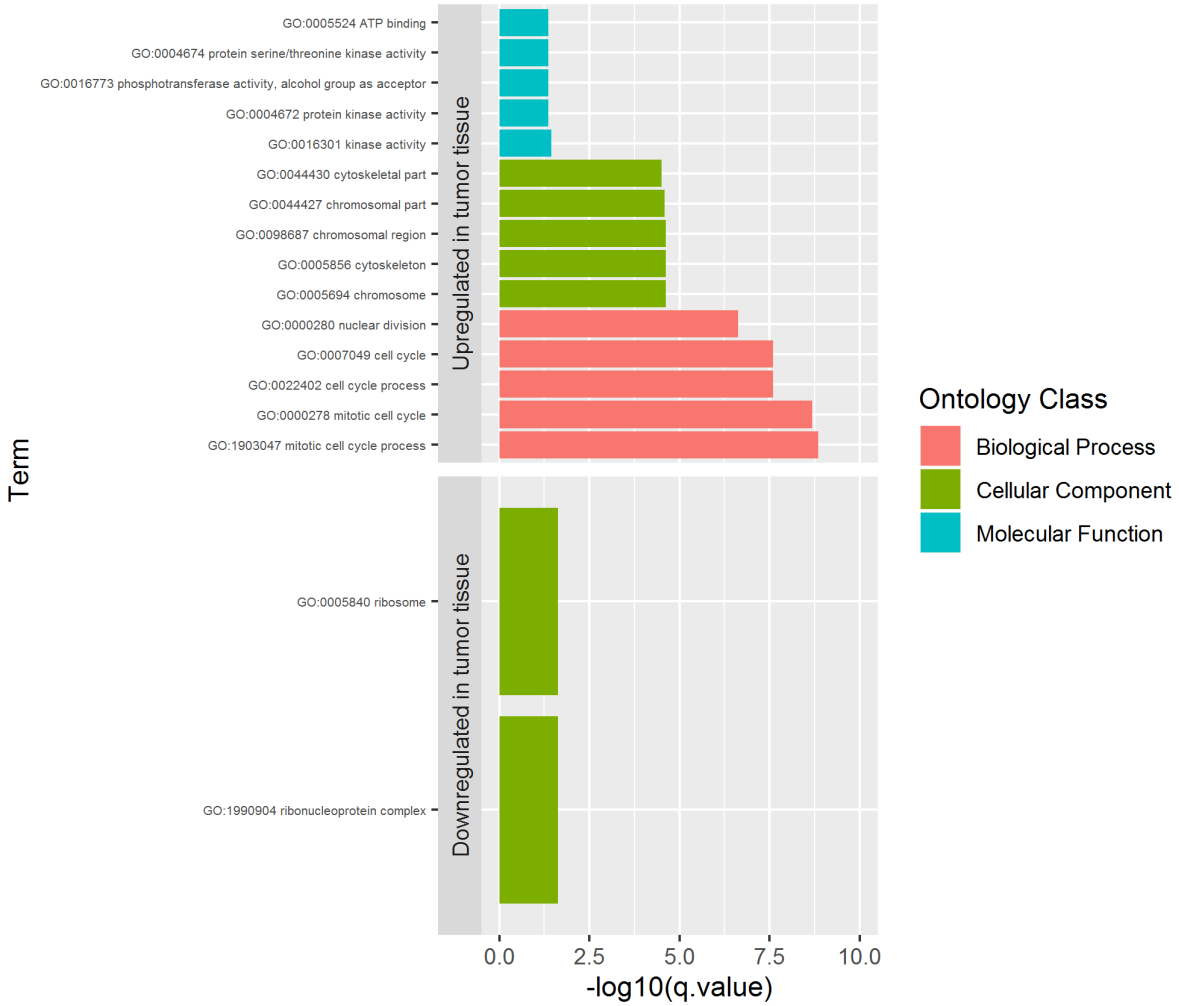

Figure 4.15 - Top five most significantly enriched GO terms in stage-I ESCA samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, molecular functions, and cellular components. The upper portion of the chart depicts GO terms that were upregulated in stage-I ESCA samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I ESCA samples.

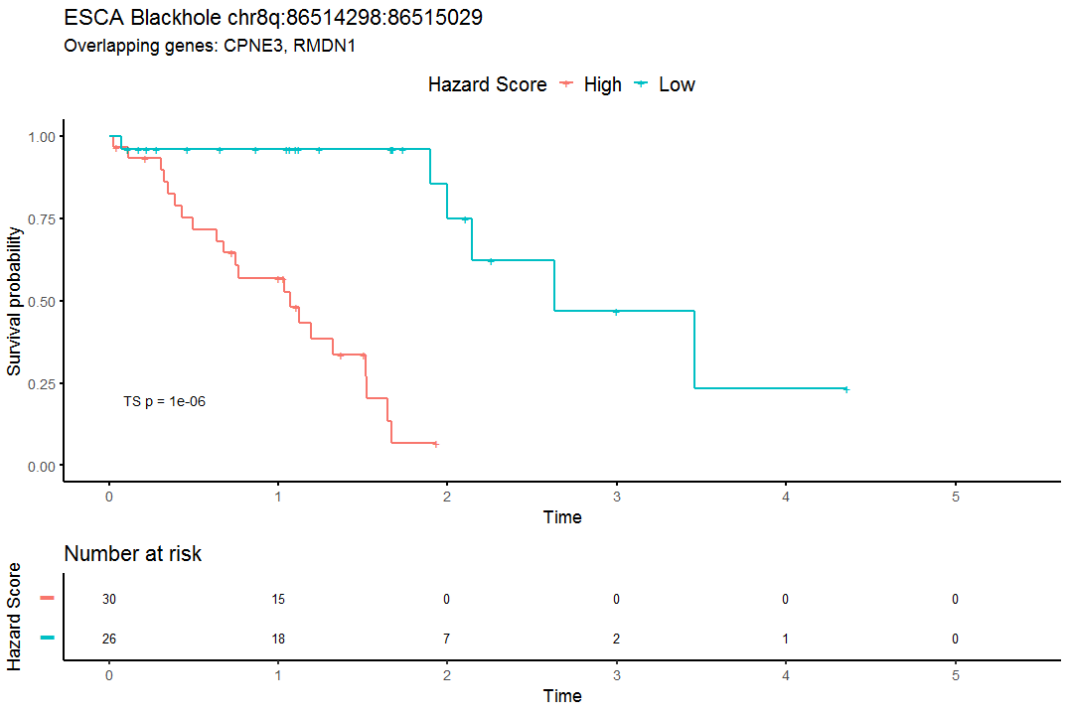

Figure 4.16 - Kaplan-Meier estimator of survival in stage-III ESCA for two groups with different epi-blackhole methylation levels. Two groups of stage-III ESCA patients with different hazard scores show significantly different survival (p = 0.000001). Statistical significance tested by Two-Stage. Time is represented in years.

4.7.5 Summary of results from Head and Neck Squamous Cell Carcinoma

In the head and neck squamous cell carcinoma (HNSC) cohort we identified 203 epi-hotspots and 55 Epi-Blackholes (Annexes VI and XXXV). The location of the identified regions is graphically represented in Figure 4.17.

In the HNSC cohort we found 1528 genes that were differentially expressed between the normal and stage-I tumor groups, of which 1431 (~94%) were associated with epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XVIII). Of these differentially expressed, epi-hotspot-associated genes, 56 were never cited, 81 were cited in the non-cancer literature, and 573 where cited in the cancer-literature, but never in the head and neck cancer literature. By submitting these differentially expressed and potentially epigenetically regulated genes to a GAGE analysis, we observed that 358 GO terms were up-regulated in the stage-I tumoral group and 115 were down-regulated (Table 4.9, Annex XXVII). By analyzing the five most enriched Go terms for each ontology class, we observed an upregulation of gene sets related to cell adhesion and extracellular organization and a downregulation of gene sets related to several metabolic processes (Fig. 4.18).

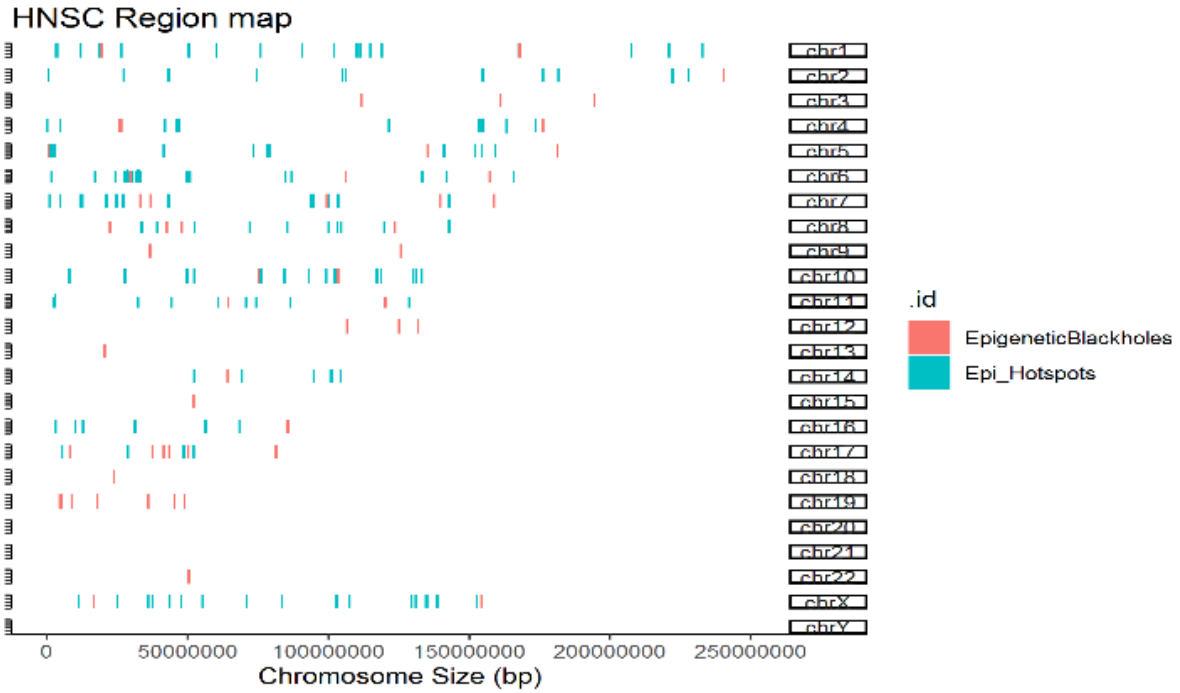

Figure 4.17 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in head and neck cancer. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

Table 4.9 - Number of enriched GO terms in stage-I HNSC samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	254	84
Cellular Component	51	15
Molecular Function	53	16

The remaining 18129 genes were not differentially expressed between the normal and stage-I tumor groups, of which 181 (~1%) could be explained by epi-blackhole variability (multiple linear regression r-squared $\geq$ 0.7) (Annex XLVI). Of these non-differentially expressed, epi-blackhole-associated genes, 12 were never cited, 15 were cited in the non-cancer literature, and 102 where cited in the cancer literature, but never in the head and neck cancer literature.

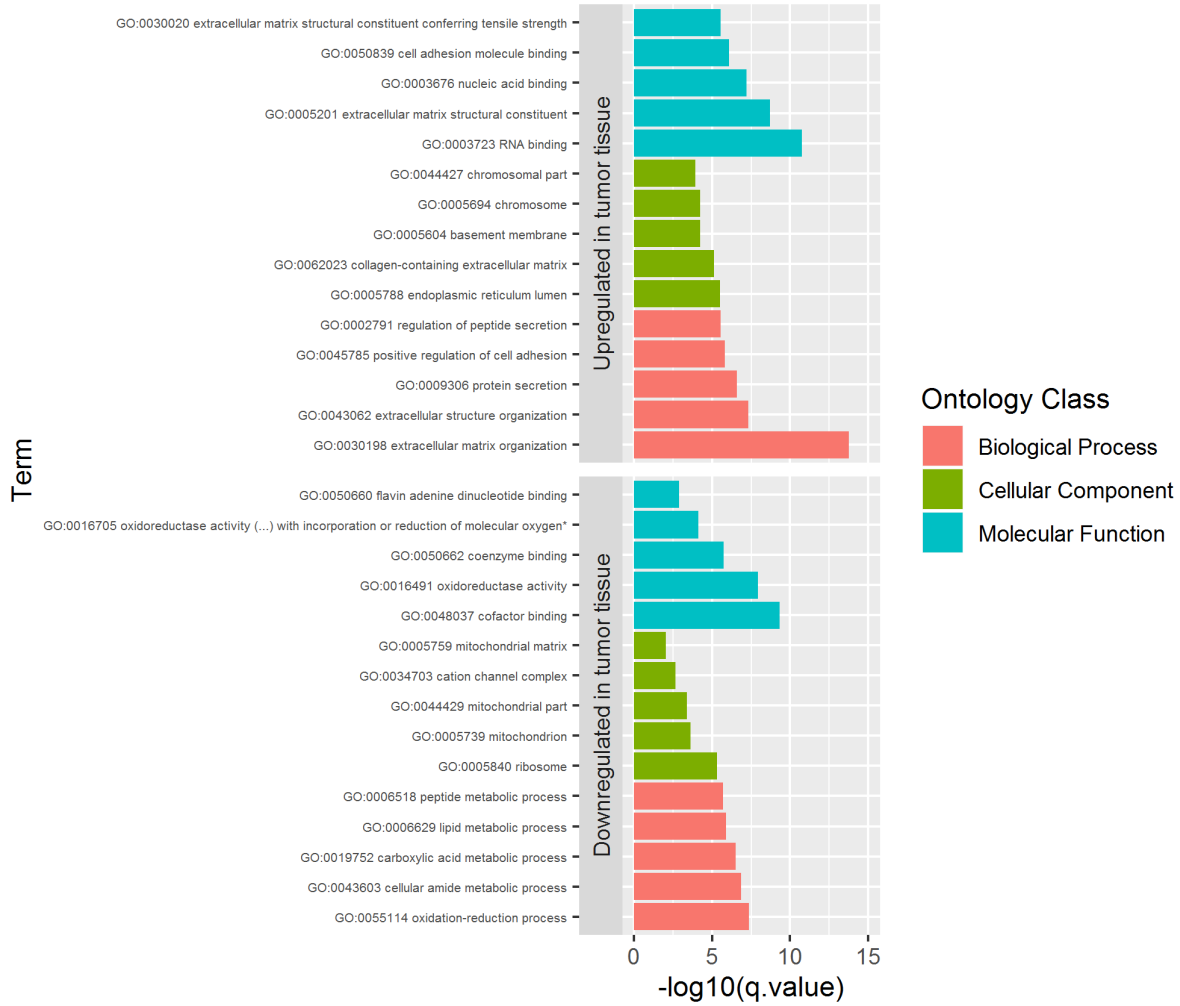

Figure 4.18 - Top five most significantly enriched GO terms in stage-I HNSC samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, molecular functions, and cellular components. The upper portion of the chart depicts GO terms that were upregulated in stage-I HNSC samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I HNSC samples.

When testing whether the identified epi-hotspots and epi-blackholes were able to predict survival in stage-III head and neck squamous cell carcinoma patients, we found that three Epi-Blackholes and eight epi-hotspots were putative prognostic biomarkers.

Figures 4.19 and 4.20 represent one epi-hotspot and one epi-blackhole that can differentiate stage-III head and neck squamous cell carcinoma patients into two groups with distinct survival distributions. Data for all other predictive regions is available in Annex LIV.

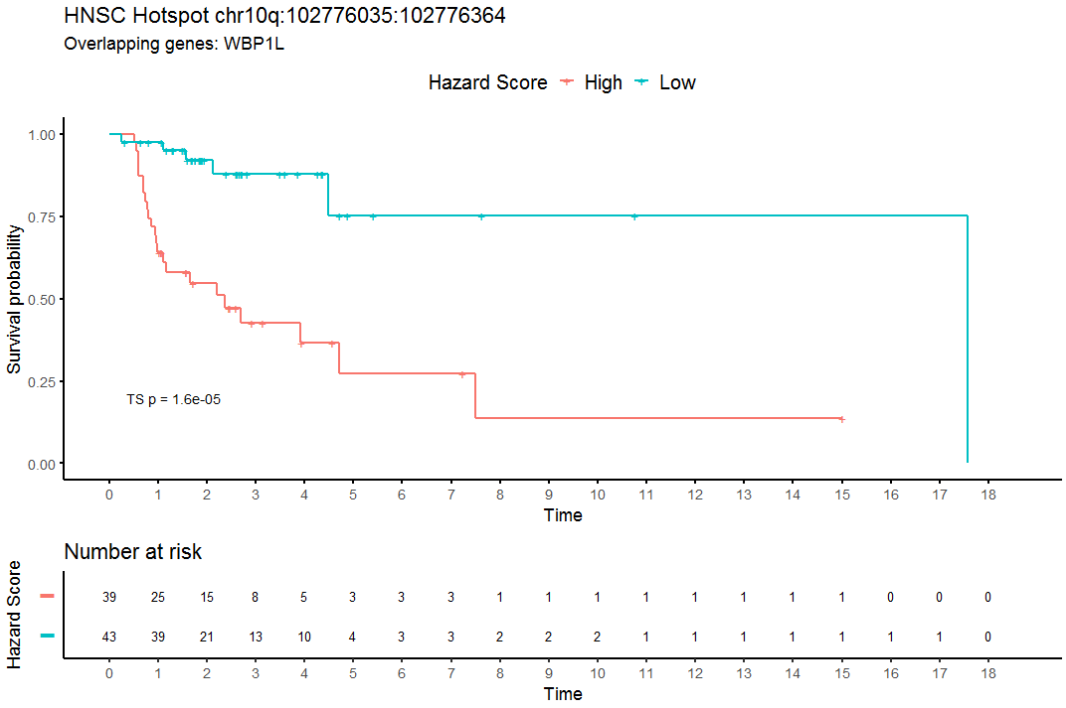

Figure 4.19 - Kaplan-Meier estimator of survival in stage-III HNSC for two groups with different epi-hotspot methylation levels. Two groups of stage-III HNSC patients with different hazard scores show significantly different survival (p = 0.000016). Statistical significance tested by Two-Stage. Time is represented in years.

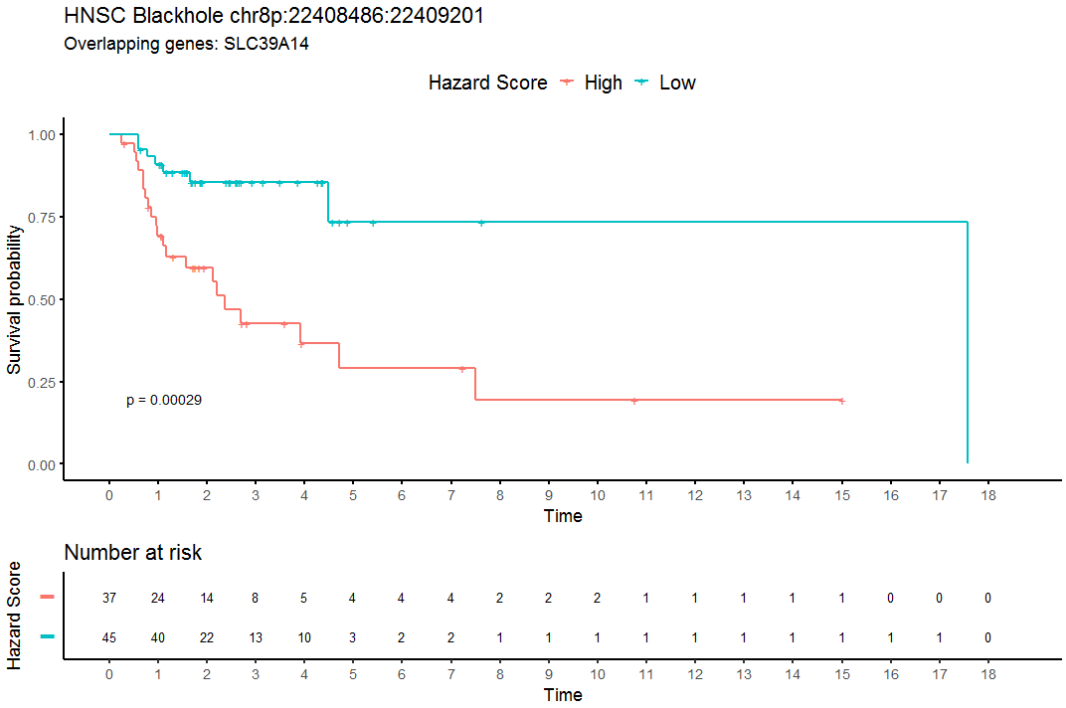

Figure 4.20 - Kaplan-Meier estimator of survival in stage-III HNSC for two groups with different epi-blackhole methylation levels. Two groups of stage-III HNSC patients with different hazard scores show significantly different survival (p = 0.00029). Statistical significance tested by log-rank. Time is represented in years.

4.7.6 Summary of results from Hepatocellular Carcinoma

In the Hepatocellular Carcinoma (LIHC) cohort we identified 208 epi-hotspots and 11 Epi-Blackholes (Annexes VII and XXXVI). The location of the identified regions is graphically represented in Figure 4.21.

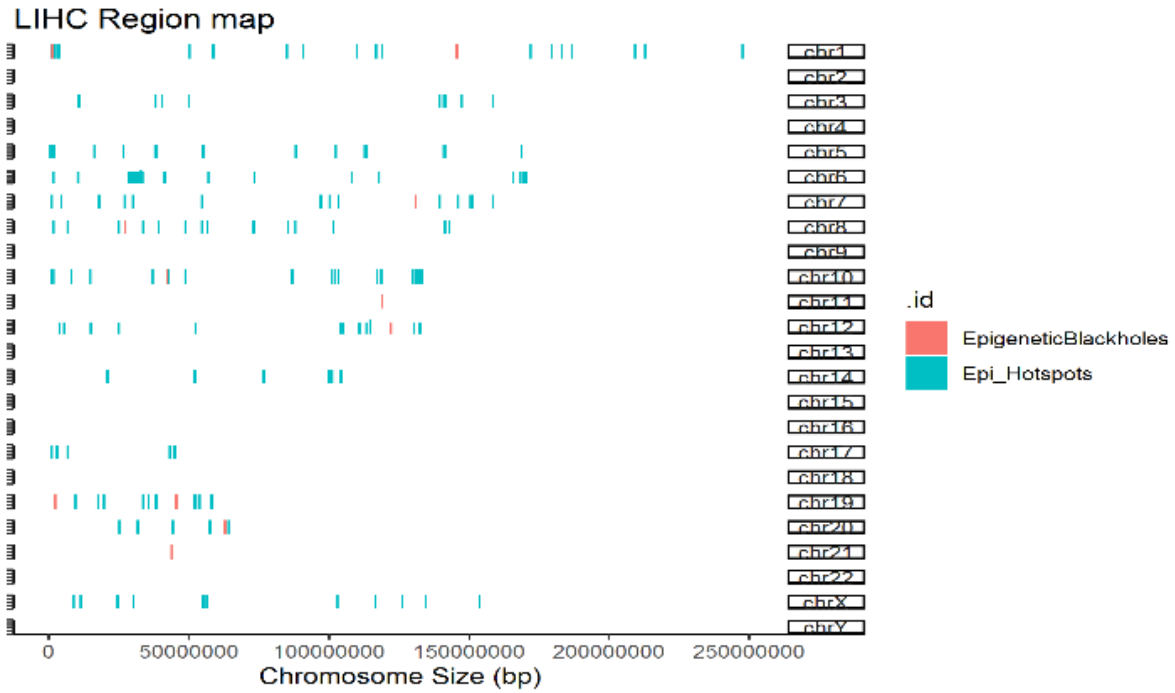

Figure 4.21 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in hepatocellular carcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the LIHC cohort we found 2921 genes that were differentially expressed between the normal and stage-I tumor groups, of which 1355 (~46%) were associated with epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XIX). Of these differentially expressed, epi-hotspot-associated genes, 49 were never cited, 59 were cited in the non-cancer literature, and 460 were cited in cancer literature, but never in the liver cancer literature. By submitting these differentially expressed and potentially epigenetically regulated genes to a GAGE analysis, we observed that 375 GO terms were up-regulated in the stage-I tumoral group and 397 were down-regulated (Table 4.10, Annex XXVIII). By analyzing the five most enriched Go terms for each ontology class, we observed an upregulation of gene sets related to cell division and a downregulation of gene sets related to cellular response to stimuli (Fig. 4.22).

Table 4.10 - Number of enriched GO terms in stage-I LIHC samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	261	318
Cellular Component	62	38
Molecular Function	52	41

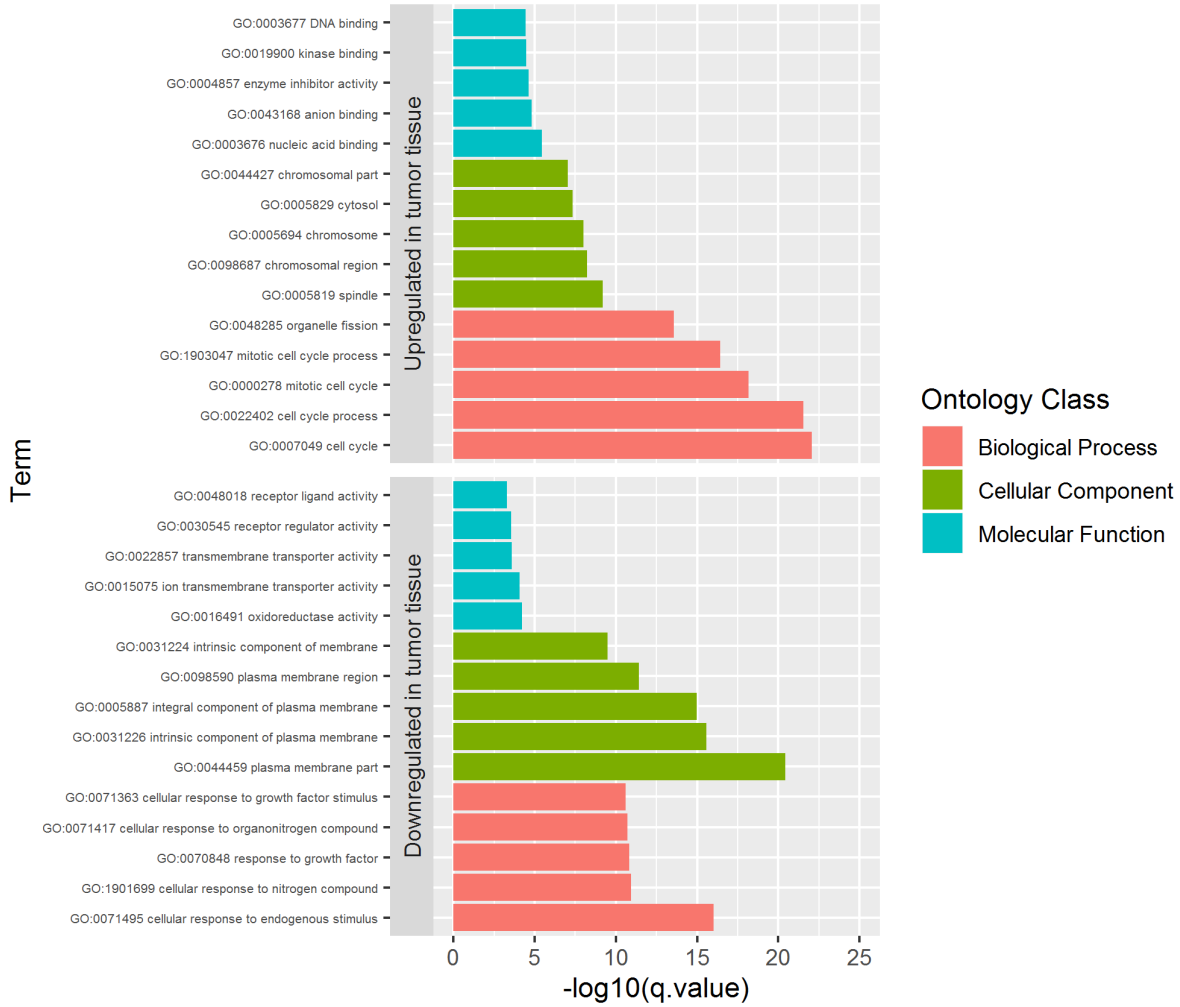

Figure 4.22 - Top five most significantly enriched GO terms in stage-I LIHC samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, molecular functions, and cellular components. The upper portion of the chart depicts GO terms that were upregulated in stage-I LIHC samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I LIHC samples.

The remaining 16736 genes were not differentially expressed between the normal and stage-I tumor groups, of which none could be explained by epi-blackhole variability (multiple linear regression r-squared ≥ 0.7).

When testing whether the identified epi-hotspots and epi-blackholes were able to predict survival in stage-III hepatocellular carcinoma patients, we found that one epi-hotspot was a putative prognostic biomarker. Figure 4.23 below represents one epi-hotspot that can differentiate stage-III hepatocellular carcinoma patients into two groups with distinct survival distributions. Additional data regarding this epi-hotspot as a prognosis predicting region is in Annex LV.

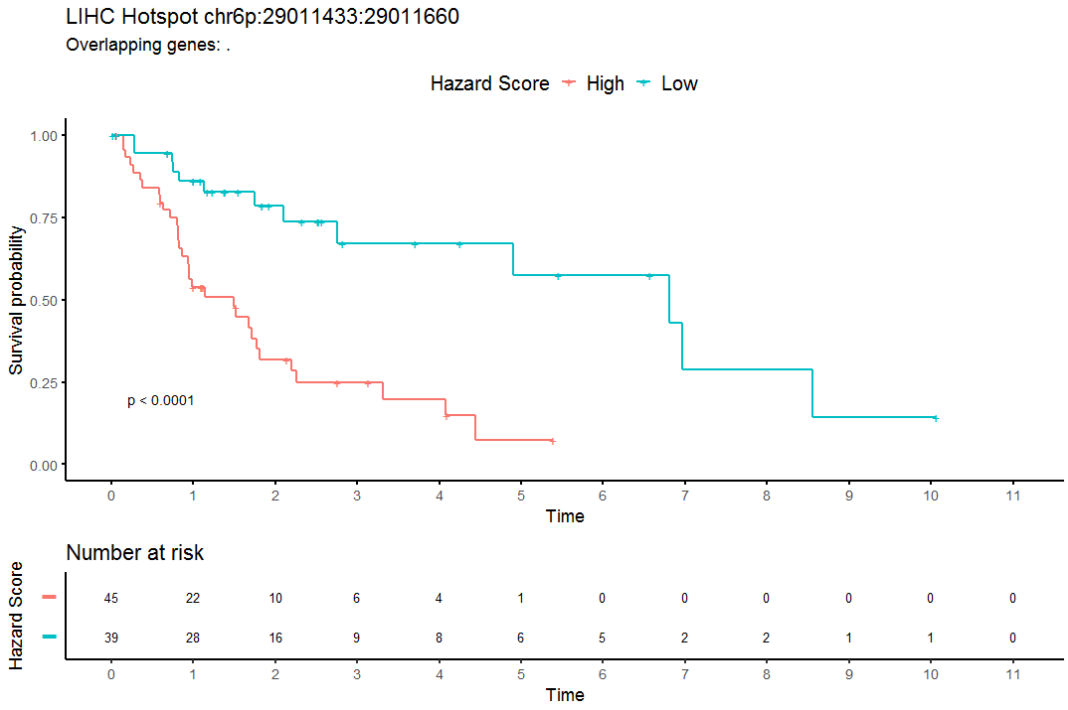

Figure 4.23 - Kaplan-Meier estimator of survival in stage-III LIHC for two groups with different epi-hotspot methylation levels. Two groups of stage-III LIHC patients with different hazard scores show significantly different survival ($p < 0.0001$). Statistical significance tested by Two-Stage. Time is represented in years.

4.7.7 Summary of results from Lung Squamous Cell Carcinoma

In the Lung Squamous Cell Carcinoma (LUSC) cohort we identified 97 epi-hotspots and 18 epi-blackholes (Annexes VIII and XXXVII). The location of the identified regions is graphically represented in the Figure 4.24.

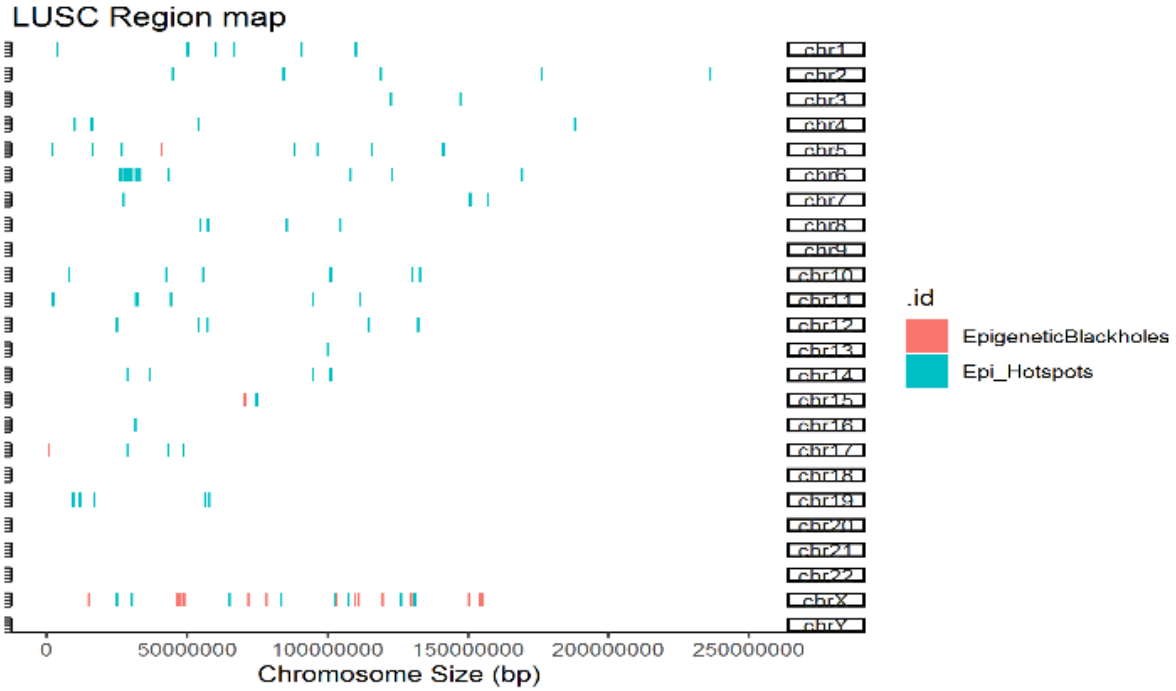

Figure 4.24 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in lung squamous cell carcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the LUSC cohort we found 5577 genes that were differentially expressed between the normal and stage-I tumor groups, of which 706 (~13%) were associated with epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XX). Of these differentially expressed, epi-hotspot explained genes, 24 were never cited, 27 were cited in the non-cancer literature, and 448 were cited in the cancer literature, but never in the squamous cell lung cancer literature. By submitting these differentially expressed and potentially epigenetically regulated genes to a GAGE analysis, we observed that 91 GO terms were up-regulated in the stage-I tumoral group and 268 were down-regulated (Table 4.11, Annex XXIX). By analyzing the five most enriched Go terms for each ontology class, we observed an upregulation of gene sets related to chromosome and organelle organization and a downregulation of gene sets related to immune response (Fig. 4.25). The remaining 14080 genes that were not differentially expressed between the normal and stage-I tumor groups, 22 (~0.16%) could be explained by epi-

blackhole variability (multiple linear regression r-squared ≥ 0.7) (Annex XLVII). Of these non-differentially expressed, epi-blackhole-associated genes, 7 were never cited, 4 were cited in the non-cancer literature, and 11 were cited in cancer literature, but never in squamous cell lung cancer literature.

Table 4.11 - Number of enriched GO terms in stage-I LUSC samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	38	234
Cellular Component	18	25
Molecular Function	35	9

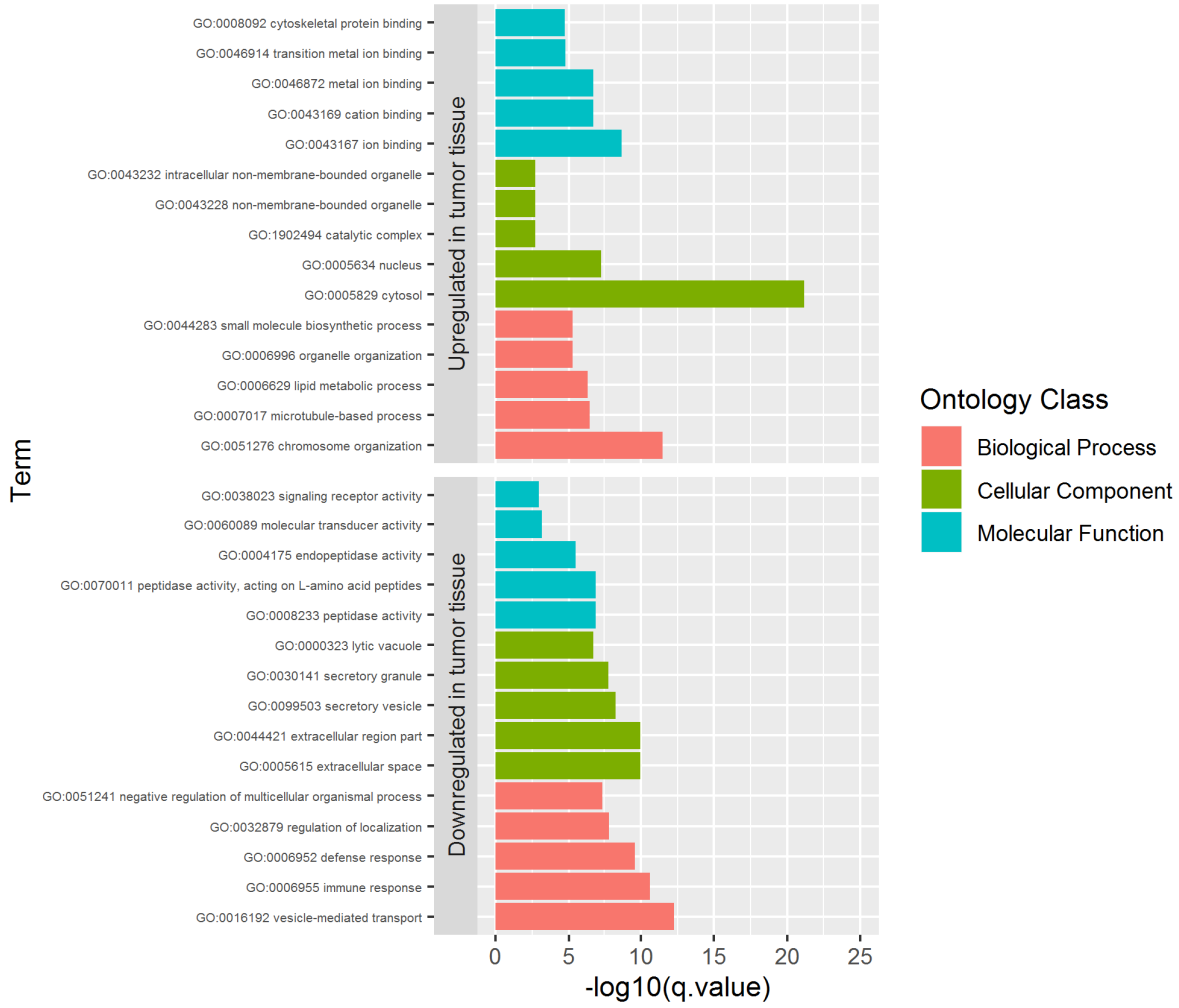

Figure 4.25 - Top five most significantly enriched GO terms in stage-I LUSC samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, molecular functions, and cellular components. The upper portion of the chart depicts GO terms that were upregulated in stage-I LUSC samples, whereas the bottom portion of the chart represent GO terms that were downregulated in stage-I LUSC samples.

When testing for the ability of epi-hotspots and epi-blackholes to predict survival in stage-III Lung Squamous Cell Carcinoma patients, we found that one epi-hotspot was a putative prognostic biomarker. The survival curve for this region is represented below, in Figure 4.26. Additional data regarding this epi-hotspot as a prognosis predicting region is in Annex LVI.

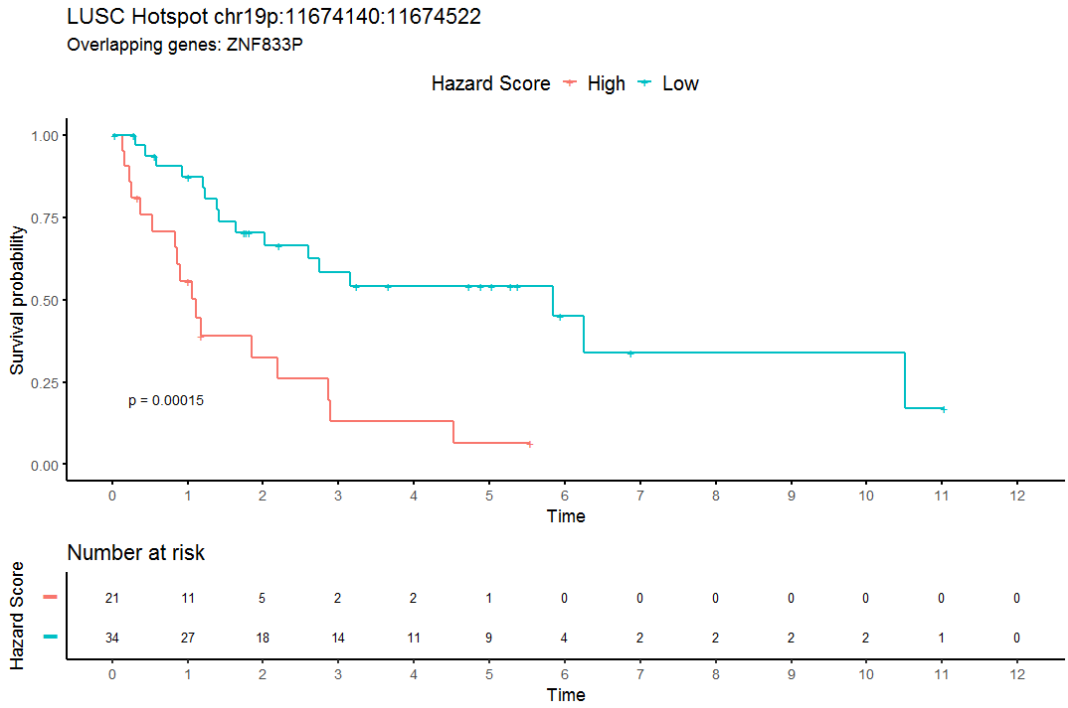

Figure 4.26 - Kaplan-Meier estimator of survival in stage-III LUSC for two groups with different epi-hotspot methylation levels. Two groups of stage-III LUSC patients with different hazard scores show significantly different survival (p = 0.00015). Statistical significance tested by Two-Stage. Time is represented in years.

4.7.8 Summary of results from Thyroid Carcinoma

In the thyroid carcinoma (THCA) cohort we identified 7 epi-hotspots and 127 Epi-Blackholes (Annexes IX and XXXVIII). The location of the identified regions is graphically represented in Figure 4.27.

In the THCA cohort we found 2476 genes that were differentially expressed between the normal and stage-I tumor groups, of which 44 (~2%) were associated with epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XXI). Of these differentially

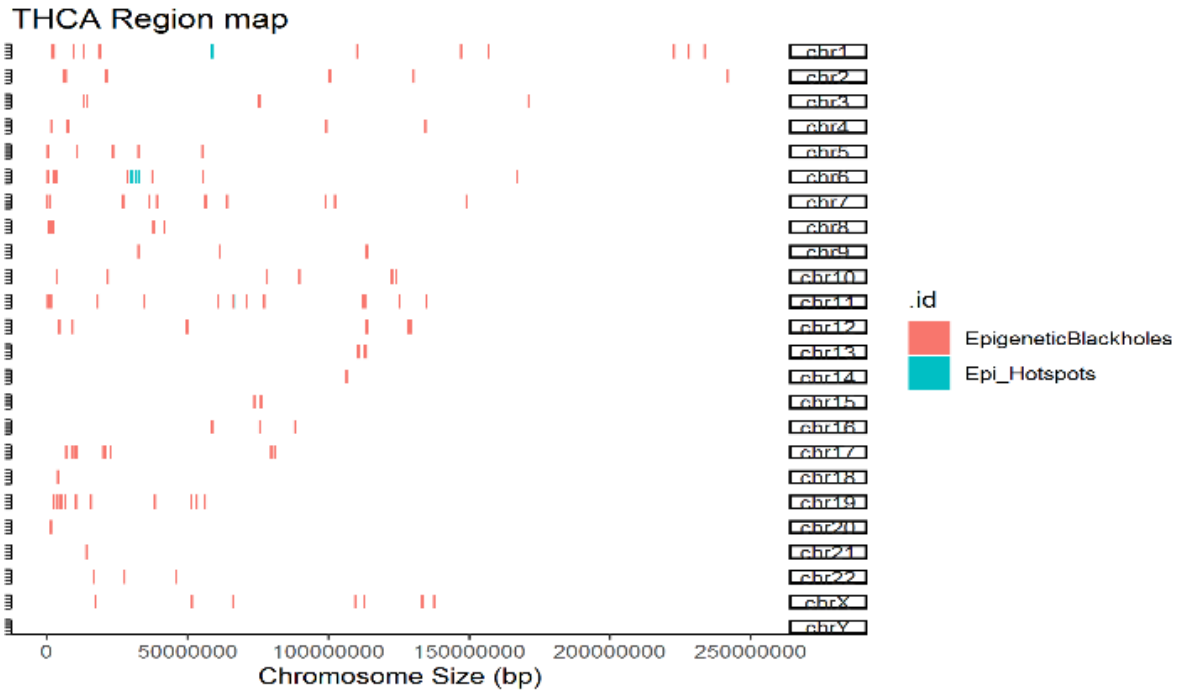

Figure 4.27 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in thyroid carcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

expressed, epi-hotspot-associated genes, 0 were never cited, 1 was cited in the non-cancer literature, and 17 where cited in the cancer literature, but never in the thyroid cancer literature. By submitting these differentially expressed and potentially epigenetically regulated genes to a GAGE analysis, we observed that 16 GO terms were down-regulated in the stage-I tumoral group (Table 4.12, Annex XXX). By analyzing the five most enriched Go terms for each ontology class, we observed a downregulation of gene sets related to cellular response to stimuli (Fig. 4.28).

Table 4.12 - Number of enriched GO terms in stage-I THCA samples

GO Class	Upregulated in tumor tissue	Downregulated in tumor tissue
Biological Process	0	15
Cellular Component	0	0
Molecular Function	0	1

The remaining 17181 genes were not differentially expressed between the normal and stage-I tumor groups, of which 39 (~0.23%) could be explained by epi-blackhole variability (multiple linear regression r-squared ≥ 0.7) (Annex XLVIII). Of these non-differentially

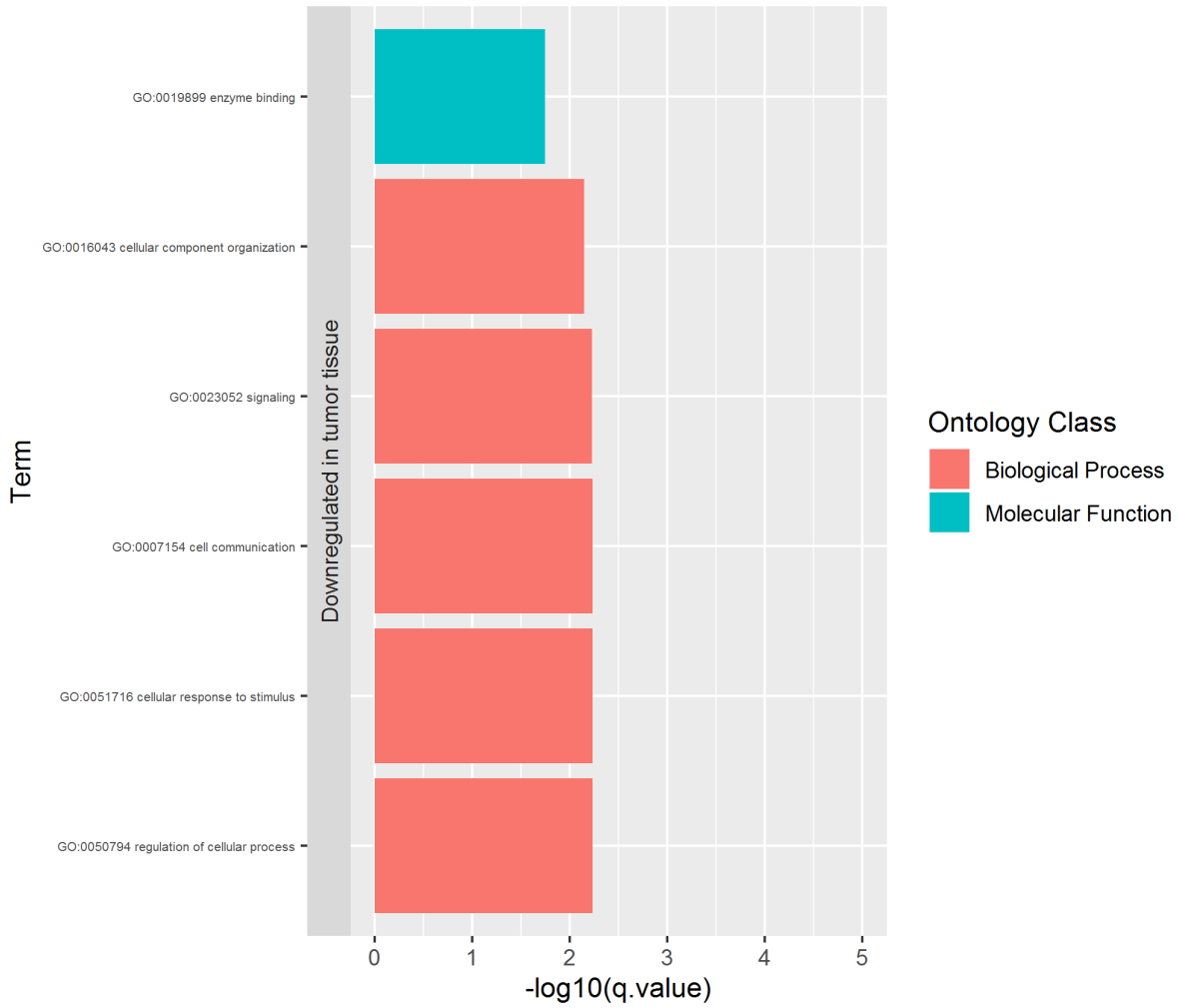

Figure 4.28 - Top five most significantly enriched GO terms in stage-I THCA samples. These terms result from a GAGE analysis and represent the top five most significantly enriched biological processes, and molecular functions. No cellular components were significantly enriched. The chart depicts GO terms that were downregulated in stage-I THCA samples. No downregulated processes were detected.

expressed, epi-blackhole explained genes, all were already cited in the cancer-literature, and 31 were cited in the cancer literature, but never in the thyroid cancer literature.

When testing for the ability of epi-hotspots and epi-blackholes to predict survival in stage-III thyroid carcinoma patients, we found no putative prognostic biomarkers.

4.7.9 Summary of results from Kidney Renal Papillary Cell Carcinoma

In the Kidney Renal Papillary Cell Carcinoma (KIRP) cohort we identified 29 epi-hotspots and 17 epi-blackholes (Annexes X and XXXIX). The location of the identified regions is graphically represented in the figure **(**Fig. 4.29).

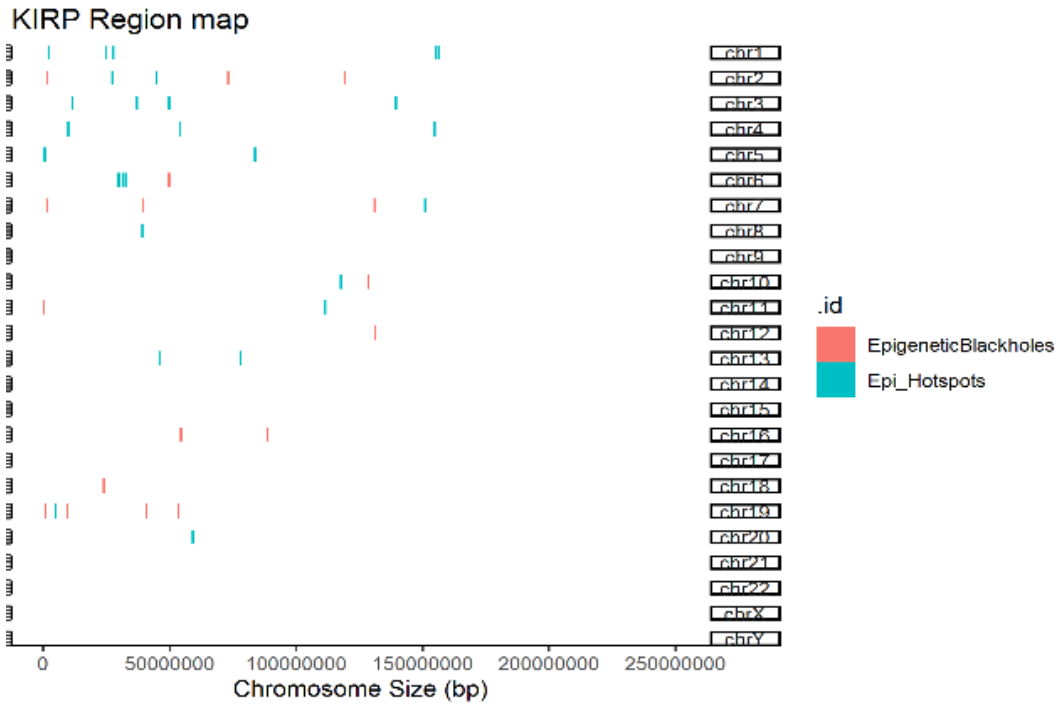

Figure 4.29 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in kidney renal papillary cell carcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the KIRP cohort we found 4326 genes that were differentially expressed between the normal and stage-I tumor groups, of which 66 (~2%) were associated with epi-hotspot variability (multiple linear regression r-squared ≥ 0.7) (Annex XXII). Of these differentially expressed epi-hotspot-associated genes, 3 were never cited, 3 were cited in the non-cancer literature, and 35 where cited in the cancer literature, but never in the Kidney Renal Papillary Cell Carcinoma literature.

The genes that were differentially expressed and whose variation during normal to stage-I tumor tissue could be explained by epi-hotspot variability were submitted to a GAGE analysis. However, no GO term was significantly enriched.

The remaining 15331 genes were not differentially expressed between the normal and stage-I tumor groups, none of which was associated with epi-blackhole variability (multiple linear regression r-squared ≥ 0.7).

When testing for the ability of epi-hotspots and epi-blackholes to predict survival in stage-III Kidney Renal Papillary Cell Carcinoma patients, we found that two epi-blackholes were putative prognostic biomarkers. Figures 4.30 and 4.31 below represent the two epi-

blackholes that could differentiate stage-III kidney Renal Papillary Cell Carcinoma patients into two groups with distinct survival distributions. The complete data regarding epi-blackholes as predictive prognostic regions in KIRP is available in Annex LVII.

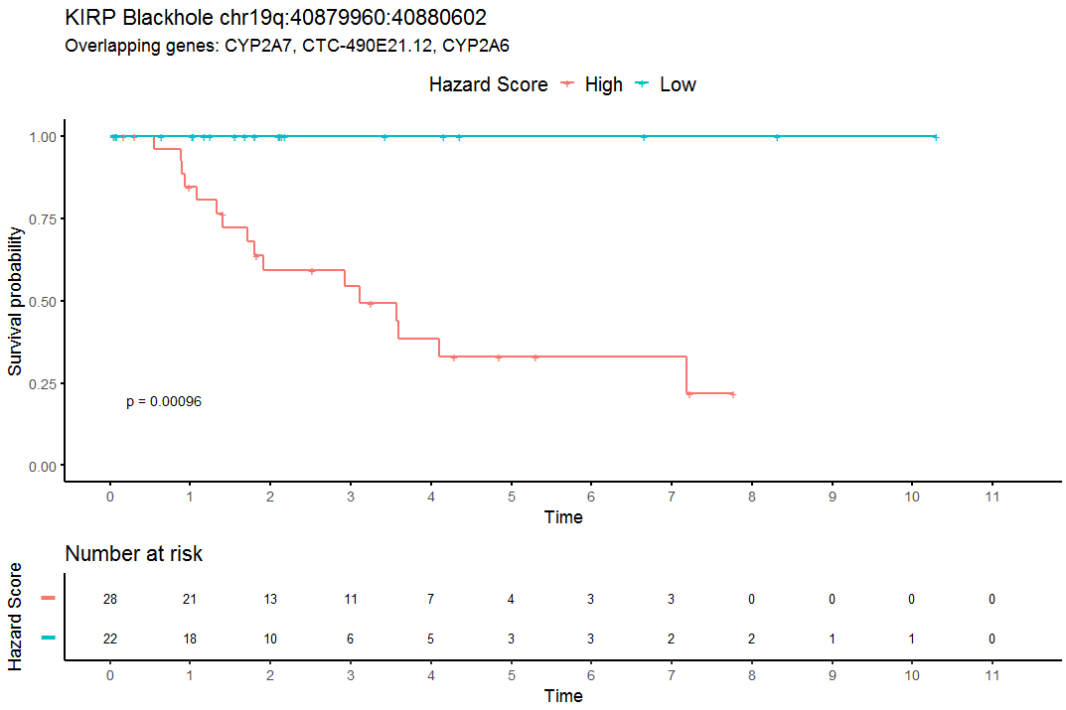

Figure 4.30 - Kaplan-Meier estimator of survival in stage-III KIRP for two groups with different epi-blackhole methylation levels. Two groups of stage-III KIRP patients with different hazard scores show significantly different survival (p = 0.00096). Statistical significance tested by log-rank. Time is represented in years.

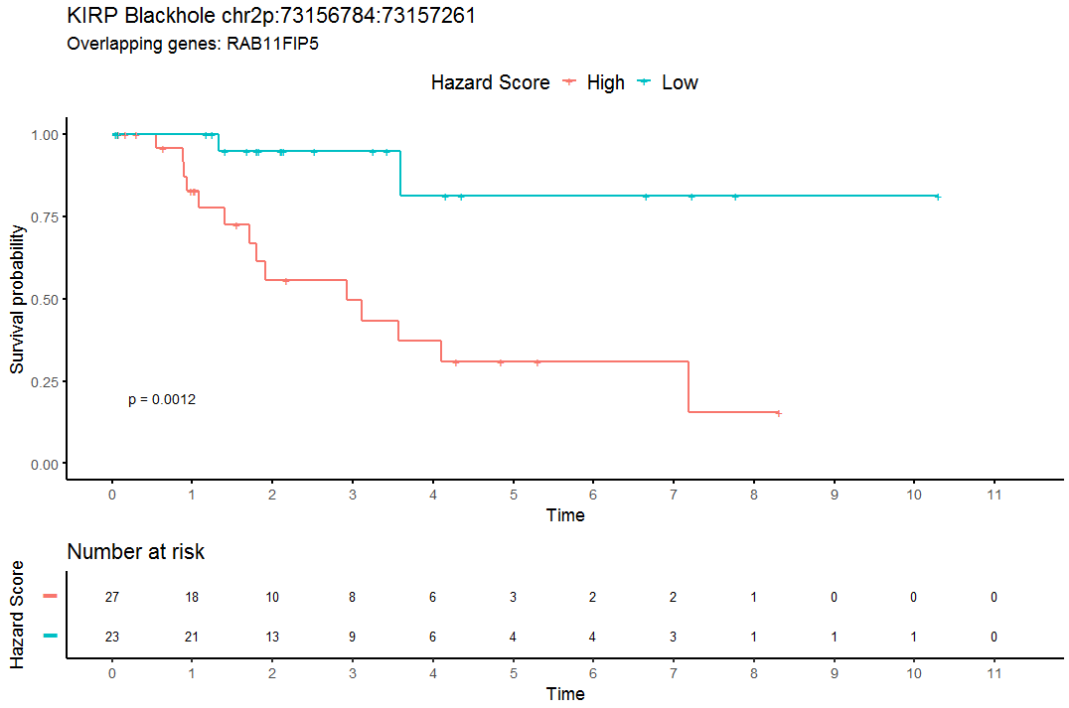

Figure 4.31 -Kaplan-Meier estimator of survival in stage-III KIRP for two groups with different epi-hotspot methylation levels. Two groups of stage-III KIRP patients with different hazard scores show significantly different survival (p = 0.0012). Statistical significance tested by log-rank. Time is represented in years.

4.7.10 Summary of results from Kidney Renal Clear Cell Carcinoma

In the kidney renal clear cell carcinoma (KIRC) cohort we identified 8 epi-hotspots and 3 epi-blackholes (Annexes XI and XL). The location of the identified regions is graphically represented in Figure 4.32.

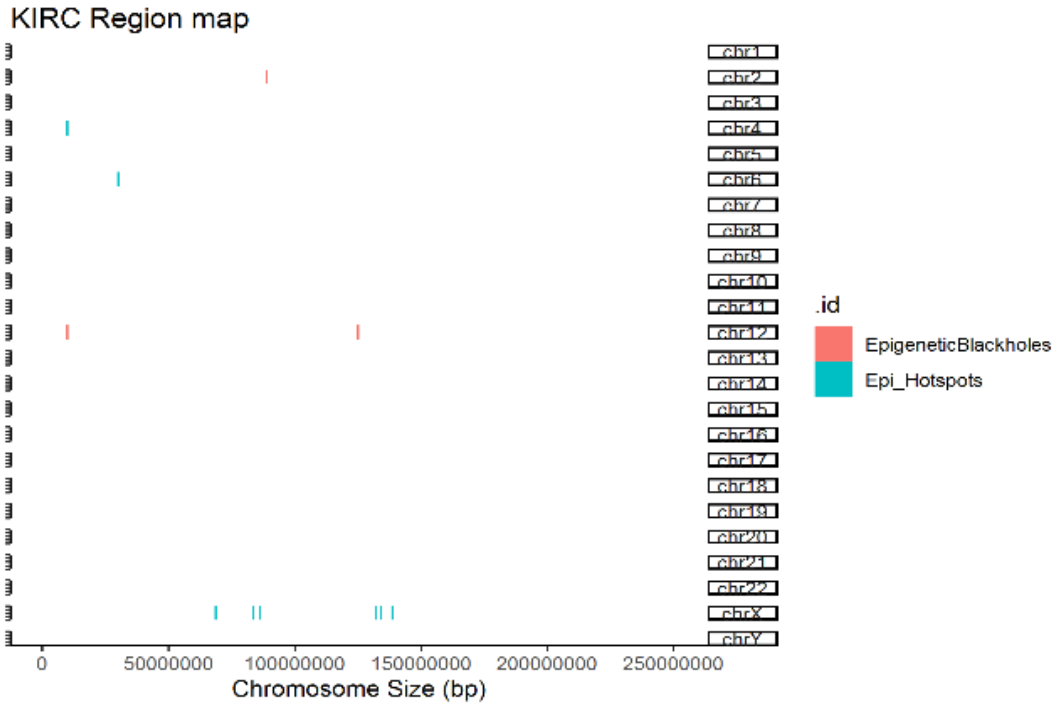

Figure 4.32 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in Kidney Renal Clear Cell Carcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the KIRC cohort we found 4588 genes that were differentially expressed between the normal and stage-I tumor groups, of which none could be explained by epi-hotspot variability (multiple linear regression r-squared $\geq$ 0.7). Of the remaining 15069 non-differentially expressed genes, none was associated with epi-blackhole variability. In addition, no region was able to predict survival in stage-III kidney clear cell carcinoma.

4.7.11 Summary of results from Pancreatic Adenocarcinoma

In the pancreatic adenocarcinoma (PAAD) cohort we identified 3 epi-hotspots and 898 epi-blackholes (Annexes XII and XLI). The location of the identified regions is graphically represented in Figure 4.33.

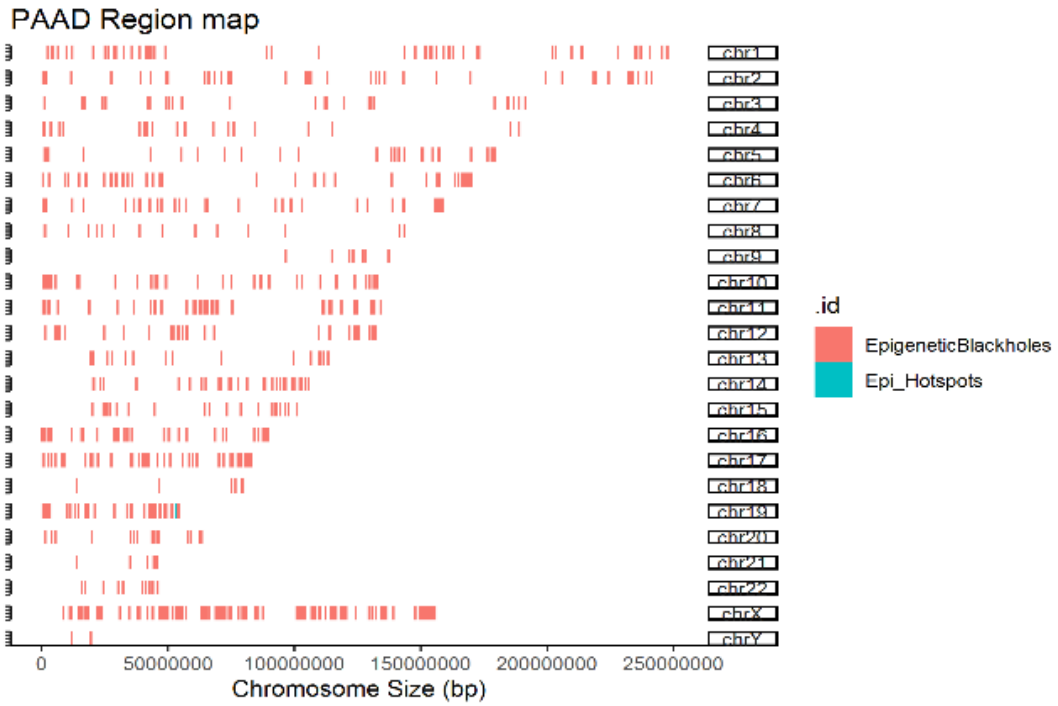

Figure 4.33 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in pancreatic adenocarcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the PAAD cohort we found no differentially expressed gene between the normal and stage-I tumor groups. It is important to note that such analytic outcome does not imply that there is no alteration in gene expression patterns during in this transition, but rather that such alterations were not detected by our analysis. In fact, it is possible that the very small number of normal samples (4 normal samples and 21 stage-I tumoral samples in the gene expression dataset) reduced the statistical power of our analysis in such a way that the detection of gene expression variations is highly unlikely.

The remaining 19657 genes were not differentially expressed between the normal and stage-I tumor groups, of which 8268 (~42.06%) were associated with epi-blackhole variability (multiple linear regression r-squared $\geq$ 0.7) (Annex XLIX). Of these non-differentially expressed epi-blackhole-associated genes, 216 were never cited, 480 were cited in the non-cancer literature, and 4472 where cited in the cancer literature, but never in the pancreatic cancer literature.

Furthermore, none of the epi-hotspots or epi-blackholes identified in this cohort was able to discriminate survival in stage-III pancreatic adenocarcinoma patients.

4.7.12 Summary of results from Lung Adenocarcinoma

In the lung adenocarcinoma (LUAD) cohort we identified 117 epi-hotspots and 134 epi-blackholes (Annexes XIII and XLII). The location of the identified regions is graphically represented in Figure 4.34.

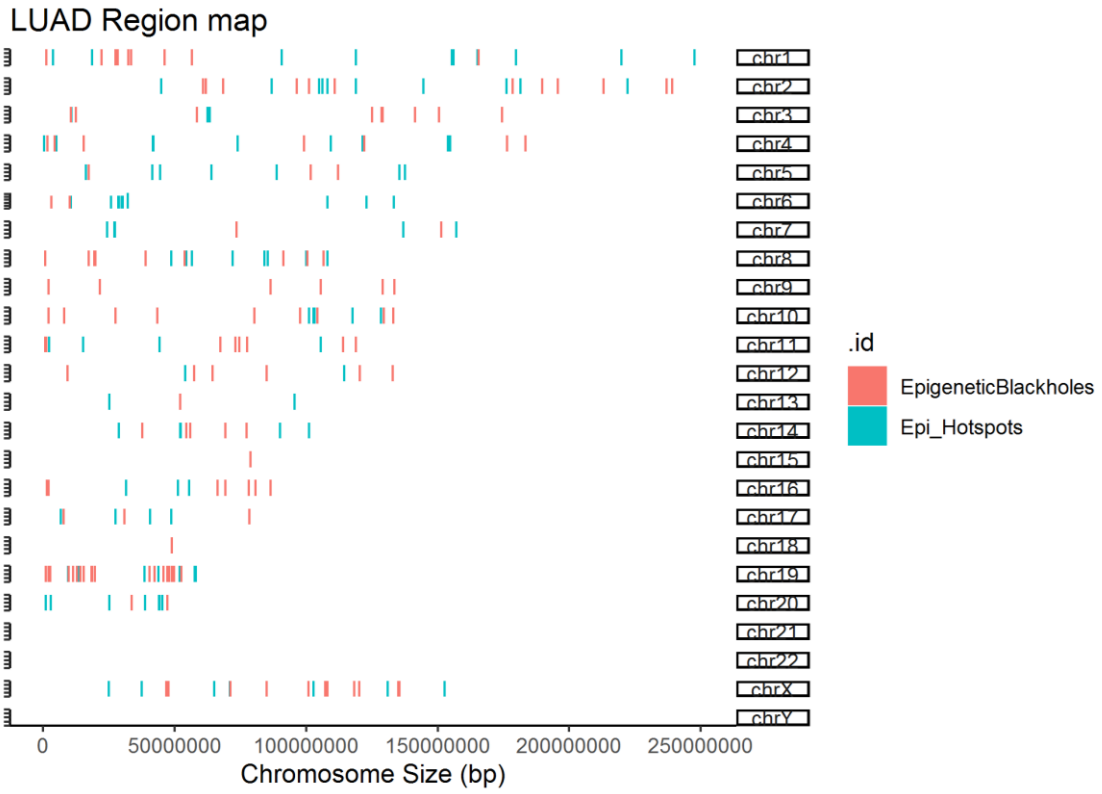

Figure 4.34 - Graphical representation of the location of the identified Epi-Hotspots (in blue) and Epi-Blackholes (in red) in lung adenocarcinoma. The chromosomes are represented in the Y-axis, and the genomic location (in bp) is represented in the X-axis.

In the LUAD cohort we found 3656 genes that were differentially expressed between the normal and stage-I tumor groups. However, none of these genes was associated with epi-hotspot variability.

Of the remaining 16001 non-differentially expressed genes, only 10 were associated with epi-blackhole variability (Annex L). These 10 genes were already cited in the cancer-literature, but never in the lung adenocarcinoma literature.

When testing for the ability of epi-hotspots and epi-blackholes to predict survival in stage-III lung adenocarcinoma patients, we found that six epi-hotspots and two epi-blackholes were putative prognostic biomarkers. Figures 4.35 and 4.36 below represent one epi-hotspot and one epi-blackhole that can differentiate stage-III lung adenocarcinoma patients into two

groups with distinct survival distributions. The complete data for all other predictive regions is in Annex LVIII.

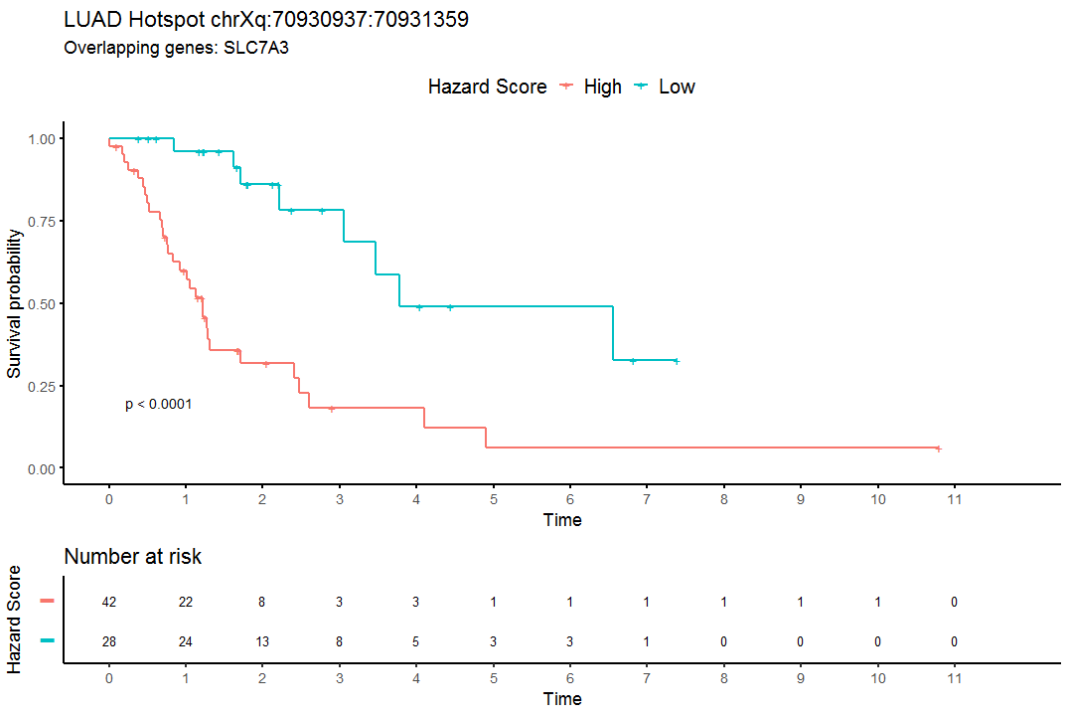

Figure 4.35 - Kaplan-Meier estimator of survival in stage-III KIRP for two groups with different epi-hotspot methylation levels. Two groups of stage-III KIRP patients with different hazard scores show significantly different survival ($p < 0.001$). Statistical significance tested by log-rank. Time is represented in years.

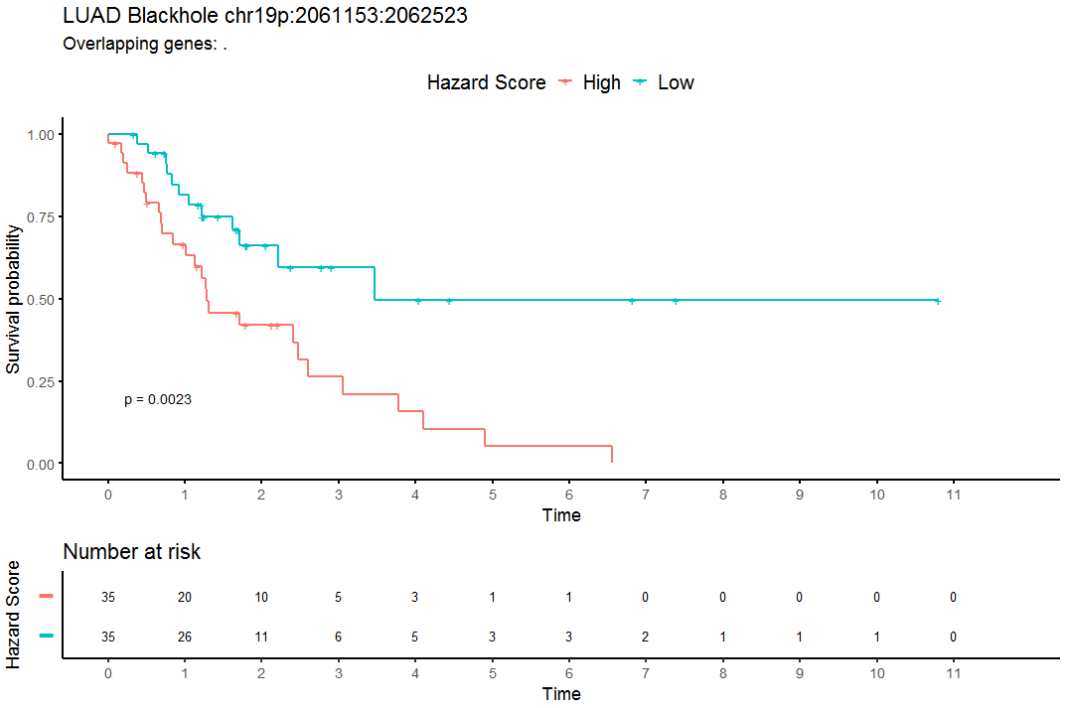

Figure 4.36 - Kaplan-Meier estimator of survival in stage-III LUAD for two groups with different epi-blackhole methylation levels. Two groups of stage-III LUAD patients with different hazard scores show significantly different survival ($p = 0.0023$). Statistical significance tested by log-rank. Time is represented in years.

4.8 Similarity between cancer types regarding Epi-hotspots

We were able to identify Epi-Hotspots in all of our cancer cohorts. We therefore wanted to determine the similarities between cancer types in terms of epi-hotspots. Using the Epi-Hotspot overlapping percentage as a similarity unit we performed hierarchical clustering analysis to determine the relative similarities between cancer types (Fig. 4.37).

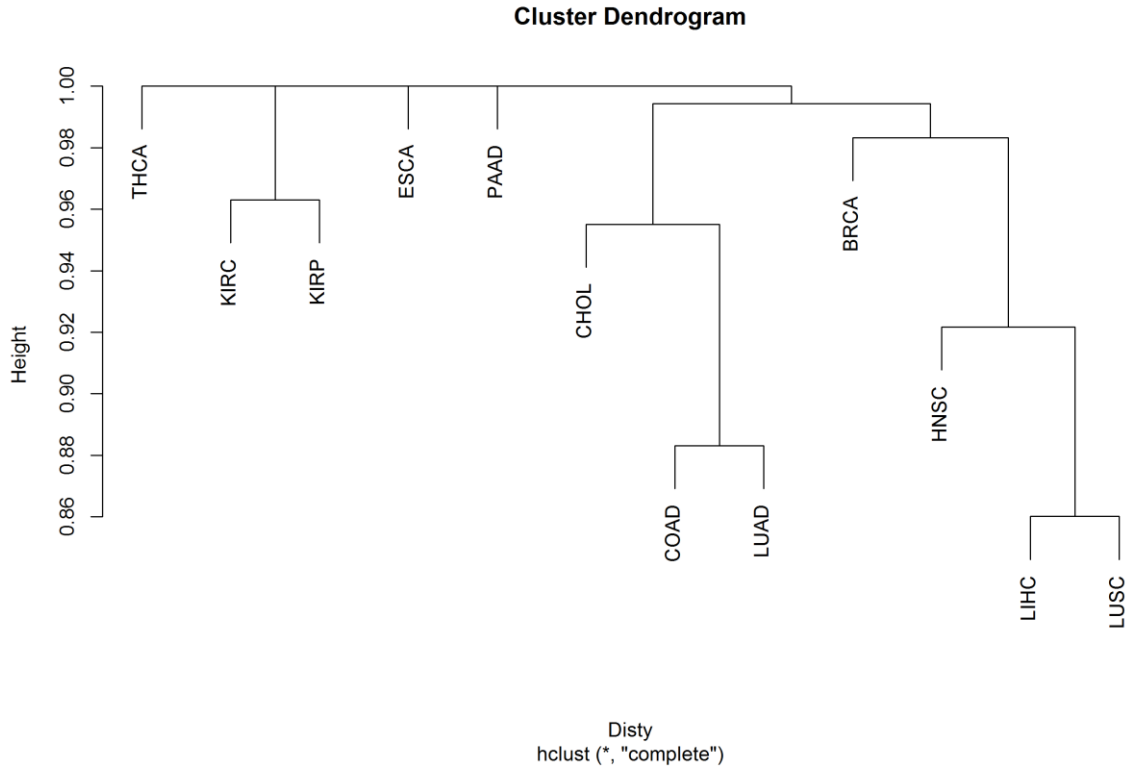

Figure 4.37 – Dendrogram representation obtained from hierarchical clustering based on epi-hotspots. Using the overlapping percentage of epi-hotspots between each pair of cohorts as similarity unit, a complete-method hierarchical clustering was performed. In the dendrogram, the height represents dissimilarity in a scale from 0 – 1.

In the dendrogram shown in Figure 4.37, there are five major clusters: 1) THCA, 2) ESCA, 3) PAAD, 4) KIRC and KIRP, and 5) CHOL, COAD, LUAD, BRCA, HNSC, LIHC, LUSC. The first three cancer types (each a separate cluster) were expected to be independent from each other and the remaining cohorts, since in these cohorts a very small number of epi-hotspots were found. Furthermore, the epi-hotspots found in these cohorts are different from all other cancer types It's important to note that it is not possible to conclude that these three cancer types are different from all other types, since the only reason they are independently

clustered is the very low epi-hotspot count in each of them. The fourth cluster is composed by the two cancer types that originate from the kidney. Even though this cluster is significantly distant from the remaining, these two cancer types have a similarity under 4%. The fifth cluster can be subdivided into two child clusters, one composed of CHOL, COAD, and LUAD, and another composed of BRCA, HNSC, LIHC, and LUSC. The two most similar cancer types are LIHC and LUSC, which share approximately 14% overlapping epi-hotspot regions, followed by COAD and LUAD, which share around 12% overlapping epi-hotspot regions (Fig 4.37). A description of the number of common epi-blackholes in each cluster is presented in Table 4.13. The genomic location of the common epi-hotspots in each epi-hotspot cluster is available in Annex LIX.

Table 4.13 – Summary of each epi-hotspot cluster. The table shows for each epi-hotspot cluster (first column), the cancers that are in the cluster (second column), the number of common epi-hotspots in the cluster (third column), and the genes that overlap with those epi-hotspots (fourth column).

Cluster	Cancers	NOE	Genes that overlap
H_1	KIRC, KIRP	2	SLC2A9, DRD5
H_2	CHOL, COAD, LUAD, BRCA, HNSC, LIHC, LUSC	0	-
H_2.1	CHOL, COAD, LUAD	6	PAX3, CCDC140, LINC00682, MARCH11, EMX2OS, EMX2, GRIA4
H_2.1.1	COAD, LUAD	19	WRAP73, TBX15, SLC5A7, PAX3, CCDC140, LINC00682, COL25A1, MARCH11, HLA-G, HCG4P8, HOXA3, RP1-170O19.22, HOXA-AS3, AC009264.1, CHRM2, GATA3-AS1, RP11-379F12.4, RP11-379F12.3, GATA3, EMX2OS, EMX2, GRIA4, SALL1, BHLHB9
H_2.2	BRCA, HNSC, LIHC, LUSC	2	WRAP73, COL11A2
H_2.2.1	HNSC, LIHC, LUSC	9	WRAP73, TNXB, COL11A2, EVX1-AS, CA3, GATA3, BHLHB9, BEX1
H_2.2.1.1	LIHC, LUSC	31	WRAP73, ELAVL4, ZIC1, MARCH11, CTD-2533K21.3, HLA-G, HCG4P8, TNXB, TNXB, COL11A2, NR2E1, EVX1-AS, GIMAP8, ASB10, RP1, CA3, GATA3-AS1, RP11-379F12.4, RP11-379F12.3, GATA3, CCNYL2, TLX1, CFAP46, BCAT1, RP11-662I13.3, MIR411, ZNF559-ZNF177, ZNF177, ZNF814, CTD-2583A14.9, BHLHB9, BEX1

NOE: Number of Epi-hotspots

4.9 Similarity between cancer types regarding epi-blackholes

Epi-Blackholes were identified in all twelve cancer cohorts analyzed. We next sought to determine whether there were groups of cancer types with higher similarity between them, based on the locations of Epi-Blackholes. A hierarchical clustering analysis was performed, using the Epi-Blackhole overlapping percentage as the similarity unit (Fig. 4.38).

The resultant clustering of cancer types is significantly different from the one generated

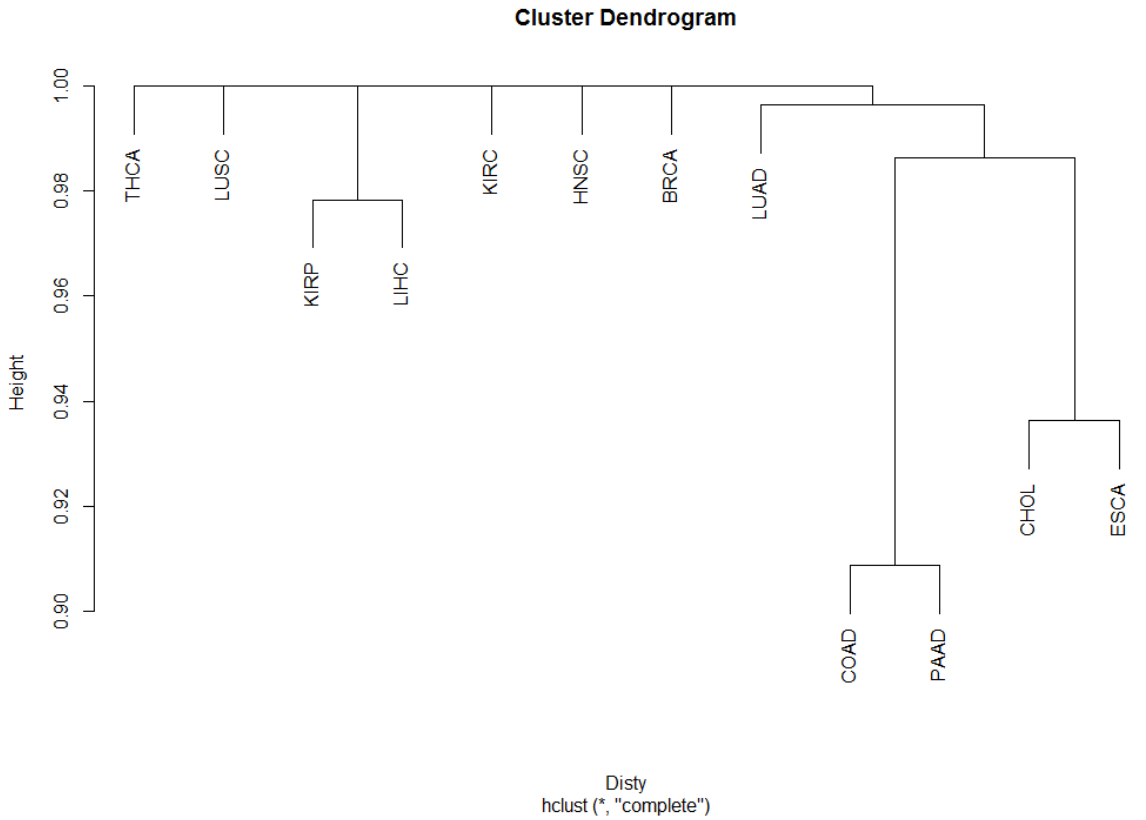

Figure 4.38 - Dendrogram representation obtained from hierarchical clustering based on epi-blackholes. Using the overlapping percentage of epi-blackholes between each pair of cohorts as similarity unit, a complete-method hierarchical clustering was performed. In the dendrogram, the height represents dissimilarity in a scale from 0 – 1.

from epi-hotspot analysis. From the resulting clustering dendrogram, seven major clusters can be discerned: 1) THCA, 2) LUSC, 3) KIRC, 4) HNSC, 5) BRCA, 6) KIRP and LIHC, and 7) LUAD, COAD, PAAD, CHOL, and ESCA. The first five clusters consist of only one cancer type, having no similarity between them or with any other cancer type. It is important to note that cancer types with low Epi-Blackhole count will tend to be dissimilar from other cancer types, whereas cancer types with a higher number of Epi-Blackholes will tend to be clustered together, since it is more likely that overlapping regions exist. The sixth cluster consists of a

KIRP and a LIHC, with around 2% similarity between them. The seventh cluster comprises five cancer types and can be subdivided in two subgroups that share less than 1% similarity, one composed of LUAD, and another composed of COAD, PAAD, CHOL, and ESCA. The two most similar cancer types are COAD and PAAD, which share approximately 91% of overlapping Epi-Blackhole regions, followed by the sub-cluster composed of CHOL and ESCA, which share around 94% similarity (Fig. 4.38). A description of the number of common epi-blackholes in each cluster is presented in Table 4.14. Due to the high number of genes that overlap with epi-blackholes, such information is presented in Annex LX.

Table 4.14 - Summary of each epi-hotspot cluster. The table shows for each epi-blackhole cluster (first column), the cancers that are in the cluster (second column), and the number of common epi-blackholes in the cluster (third column).

Cluster	Cancers	Number of common epi-blackholes
B_1	KIRP, LIHC	1
B_2	LUAD, COAD, PAAD, CHOL, ESCA	3
B_2.1	COAD, PAAD, CHOL, ESCA	20
B_2.1.1	COAD, PAAD	89
B_2.1.2	CHOL, ESCA	81

Chapter 5 Discussion

The chromatin's accessibility, in each region, is part of what allows cells to shift transcriptional states, activate and deactivate genetic pathways, and even commit to developmental programs. The permissiveness, or accessibility, state of the chromatin is, in part, controlled by DNA methylation, which displays aberrant patterns in several diseases such as cancer. In fact, it is well known that during tumor initiation, the tumoral cells follow abnormal methylation patterns, which can lead to alterations in the network of cellular pathways.

In this study, we pondered if certain genomic regions were preferentially targeted for abnormal DNA methylation during the normal to tumor transition and if these regions could be somehow related to altered patterns of gene expression.

In fact, we theorized that, during the normal to tumoral transition, certain genomic regions would be preferably targeted for alterations in DNA methylation, and others would remain unchanged. Our rationale was that certain modifications in the chromatin's accessibility would provide a selective advantage to the clone, while other modifications would constitute a selective disadvantage. As such, we hypothesized that during tumor initiation, specific genomic regions are differentially methylated, and others are kept unchanged. We decided to designate the former as Epi-Hotspots and the latter as Epi-Blackholes.

The first portion of this study was dedicated to testing this hypothesis. To do this, we made use of existing DNA methylation data from TCGA, and since cancer is a group of diseases, we analyzed all the twelve cohorts that had available data on DNA methylation and gene expression. The Epi-hotspots were identified by intersecting the outputs of two DMR searching methods named DMRCate and Bumphunter.

5.1 Epi-hotspots

We found that the number of epi-hotspots was different for every cohort. In fact, some cohorts revealed a low Epi-hotspot count, such as esophageal and pancreatic cancer cohorts, with 2 and 3 identified epi-hotspots, respectively, and other cohorts revealed a higher Epi-Hotspot count, such as liver and head and neck cancer cohorts, with 208 and 203 epi-hotspots, respectively. The reason for this discrepancy is hard to explain since all cohorts have not only different total sample sizes but also a different number of normal and tumor samples. As

described in the previous section, our analysis showed that there was not any correlation between the number of identified epi-hotspots and the number of normal, tumor, and total samples in the cohort. This is something that would, in fact, be expected, if there were not any analytic errors since an increase in statistical power would render an increase in the detection of real epi-hotspots, rather than an increase in the total number of detections. Nonetheless, this absence of correlation is to be taken carefully, since cohorts with a very low sample size may lack the statistical power to identify all epi-hotspots. In our epi-hotspot identification methodology, we generally opted for an alpha level of 5% in the statistical tests, and although this is a commonly accepted threshold, it is important to understand that in low power settings (such as in small sample size), it may be insufficient to detect a relevant amount of epi-hotspots.

Our data also demonstrated that the size of epi-hotspots greatly varied not only between cohorts but also within each analyzed cohort. In every cohort, we noted that the length of the epi-hotspots, as well as the number of CpG sites it contained, was variable. This was interesting to note since it demonstrated that these regions are not fixed in size, and can either be very short epi-hotspots (< 10 bp) with less than five CpG sites, or very long epi-hotspots (> 3 kbp) with dozens of CpG sites.

We also observed that most of the detected epi-hotspots, in every analyzed cohort, were located in promoter regions. It is long established that anomalous DNA methylation in promoters is characteristic of cancer initiation and development [108,109]. This epigenetic reprogramming was thought to follow a pattern of global hypomethylation, to activate oncogenes, and promoter hypomethylation specifically to repress tumor suppressor genes [109]. However, it is now known that this epigenetic reprograming in promoters is not contributing to cancer development by directly repressing the respective genes [108]. Our finding might, of course, reflect this epigenetic reprogramming hallmark in cancer, but there are other possible reasons that epi-hotspots are mostly concentrated in promoter regions: 1) the detection of epi-hotspots is, in a way, related to the presence of various CpG sites in a given region, so it is not surprising that high-density CpG regions, such as promoters, which are rich in CpG islands, are where most epi-hotspots are detected; 2) the visible methylome in all cancer cohorts here analyzed is also mostly concentrated in promoter regions, as such it is expected the most of the detected epi-hotspots would fall in promoter regions; 3) finally, it should not be excluded the possibility that this event is mechanistically important, and is part of a biological process that we do not yet understand.

It is also interesting to note that these regions that are consistently methylated in an aberrant manner in stage-I tumor samples represent a very small portion of the entire methylome. In fact, the cohort that revealed the largest total area of epi-hotspots was LIHC, where 0.377% of the whole methylome consisted of epi-hotspots. On the other hand, the cohort with the smallest total area of epi-hotspots was ESCA, with only 0.006% of the methylome consisting of epi-hotspots. This, of course, only represents a portion of the aberrantly methylated methylome. There might be several reasons for such small numbers, one of them being that a normal cell transitioning to a tumor cell must target specific regions of the DNA not only to increase proliferation but also to keep viability. It is also possible that these specific regions are amplified via clonal selection. In fact, one could theorize that if such regions are consistently altered in stage-I tumor cells, then it is possible that they exert some kind of advantageous selection pressure.

5.2 Epi-hotspots are related to aberrant gene expression during tumor initiation

We figured that one-way epi-hotspots could exert its selective pressure was by influencing gene expression, not only where they spawn but also in other parts of the genome. As previously explained, we first identified all genes that were differentially expressed between normal and stage-I tumor patients in all analyzed cohorts. After identifying these genes, we proceeded to filter out all of those that did not have a normal-tumor AUC lower than 0.8. It is important to discuss that the genes that were excluded in this step were not considered to be non-differentially expressed. However, we rationalized that these genes would have a high variability across samples, and, on the other hand, genes with high AUC values would have a higher gap in expression values between normal and tumor samples. This was an important step to assure that the predictive power of epi-hotspots was represented in the normal to tumor transition, rather than in intra-tumor variability.

In this study, we performed multiple linear regression models to understand if a given epi-hotspot, as a whole, was able to predict expression alteration of a given gene in tumor initiation. It is highly relevant to mention that this methodology was not developed to demonstrate a causal relationship between methylation of an epi-hotspot and expression levels of any gene, as we would not be able to determine causation with this type of analysis. Instead, we aimed to comprehend if there was any predictability in these variables.

It was interesting to observe that the number of differentially expressed genes between normal and stage-I tumor samples, with an AUC higher than 0.8, varied across all diseases. Nonetheless, in all cohorts in which differentially expressed genes were detected, there were always epi-hotspots that could predict the gene expression variation in tumor initiation. It is impossible, with the present study, to confirm the hypothesis that epi-hotspots are influencing gene expression in tumor initiation. However, our data suggest that there is some relationship between the former and the latter, which provides some evidence, although not causational, that DNA methylation in specific regions might impact gene expression during tumor initiation. This is particularly interesting due to the fact that our results are extremely narrowed, not only in the detection of the epi-hotspots but also in upholding genes to be considered differentially expressed. Furthermore, the epi-hotspot – gene relationships were also tested with a considerable degree of restriction by using high R^2 values and testing for model significance. Considering that our datasets do not have relatively large sample sizes, and by using such degree of restriction in the detection of epi-hotspot – gene relationships, it is arguable that in reality, there are many more relationships, with different degrees of association.

5.3 Epi-blackholes

Having identified genomic regions that are consistently aberrantly methylated in stage-I tumor samples, we also sought to understand if there were regions that were consistently unaltered regarding DNA methylation during tumor initiation. These regions, which were herein labeled as epi-blackholes, were identified by applying the same principles previously described to detect epi-hotspots but, instead, using an alpha level for the DMRcate algorithm only, higher than 95%. The goal of this methodology was to detect the regions that are most likely not to be altered during tumor initiation. It is important to explain that by using this method, we were trying to find enough evidence to sustain the algorithm's null hypothesis, rather than rejecting it. As such, the detected regions should be regarded not as regions that do not change its methylation patterns, but as regions that are highly unlikely to be epi-hotspots.

Our data suggested that, like epi-hotspots, epi-blackholes seem to vary in size and in number, across the analyzed diseases. These regions also portray a small portion of the visible methylome. However, it is interesting to observe that in the esophageal and bile duct cancers, epi-blackholes represent 6.1% and 5% of the visible methylome. In fact, there is a substantial difference in epi-hotspot area from epi-blackhole area in diseases such as bile duct and esophageal cancers. In a low statistical power cohort such as ESCA, such discrepancy might

be expected, as a low count of detectable epi-hotspots would cause us to detect a large number of regions that are highly unlikely to be epi-hotspots. This is also true in the CHOL cohort, although it is much more evident in the ESCA cohort.

Generally, like epi-hotspots, epi-blackholes seem to be mostly located in promoter regions, but to a smaller degree. It is possible to observe that epi-blackholes also populate other regions of the genome, such as exons or introns. We would argue that this is expected in two ways: 1) most of the visible methylome and CpG-dense regions are found in promoters, as such it is more likely to find epi-blackholes in these areas; and 2) most epi-hotspots are already mainly occupying promoters, so it is likely to find epi-blackholes in areas where epi-hotspots do not frequently spawn.

5.4 Epi-blackholes might be related to the maintenance of gene expression in tumor initiation

We also sought to understand if epi-blackholes were able to explain the variation of non-differentially expressed genes in the normal to stage-I tumor transition. Again, we were not able to look for causation; rather, we aimed to find relationships between the variation of epi-blackholes and the variation of non-differentially expressed genes during tumor initiation. Fundamentally, our data showed that epi-blackholes were not strong predictors of non-differentially expressed gene's expression variation. This may be due to the strong restrictions for true-positive detection applied. Interestingly, in some of the analyzed diseases, such as CHOL, ESCA, and PAAD, we found a considerable amount of non-differentially expressed genes whose normal-tumor variability could be predicted by epi-blackholes. These are cohorts with low statistical power and, as such, a high number of detectable epi-blackholes might increase the probability of finding a relationship. Nevertheless, considering that we only selected statistically significant models with a high R^2, we cannot exclude the possibility that these regions may play a role in tumor initiation or cell viability.

5.5 Epi-hotspots and epi-blackholes predict survival of stage-III tumor patients

We theorized that if both epi-hotspots and epi-blackholes have some role in exerting selective pressure in tumor initiation and/or development, different methylation patterns in these regions could differently impact the survival of patients. Therefore, we questioned if at least some of the identified regions could be predictors of prognosis, in a later phase of the

disease. To perform this analysis, we studied the identified epi-hotspots and epi-blackholes in stage-III patients of the same previously evaluated diseases. Initially, we aimed to study the impact of the regions in the survival of patients as the disease progressed, but this was not possible due to the lack of necessary data from several phases of the tumor. Nevertheless, we had sufficient data to analyze prognosis in stage-III patients, which is a tumor stage where patient survival is sufficiently diverse to be studied.

As explained in the previous sections, we performed multivariate Cox proportional-hazards models to assess if each of the identified epi-hotspots and epi-blackholes were able to predict survival of stage-III patients. Since this type of study deals with methylation patterns that can have small fluctuations, we sought to increase clinical relevance by creating a hazard score, which would divide patients into one of two groups: 1) a low-hazard, favorable prognosis group, or 2) a high-hazard, poor prognosis group. Although the latter step might reduce the resolution of the results, we reasoned that when dealing with β-values, that are not exact representations of true methylation status, in a set of low statistical power, this over-simplification would increase clinical relevance, and result interpretation.

As we hypothesized, in two-thirds of our cohorts, we observed that at least one region was not only able to predict survival in stage-III patients, but also divide them into two groups with distinct prognosis. This was an interesting result because although a very small amount of regions were able to predict survival, it is likely that, in reality, the number of epi-hotspots and/or epi-blackholes that predict prognosis is much larger. To make sure that this phase of the study would not render weak or false conclusions, we sought to reduce type-I errors as much as possible. This was done at the cost of an increase of false negatives, but, in fact, our primary goal was not to find all regions that predict survival, but if these regions were, in some way, related to patient prognosis. The course to profoundly decrease type-I error in this stage was meticulous and included several steps, such as 1) only analyzing models that fulfilled the assumption of proportional hazards, 2) only considering a model to be a positive result if it was simultaneously significant for three different statistical tests, 3) maintaining only the regions that were able to distinguish patients into two groups with a statistically significant difference in survival, a statistically non-significant different in patient age, and with a proportion of patient number between groups not greater than 60%. With such a level of restriction, although we are not able to understand the full mechanistic picture, our data suggest that epi-hotspots and epi-blackholes can be predictors of survival in stage-III tumor patients. Yet, we are not

able to conclude causality, since the aggressiveness of the disease, and even differences in treatment might cause different patterns of methylation in these regions.

5.6 Colon adenocarcinoma

Colon and rectal cancers represent a considerable burden of death in the world, representing approximately 9% of all cancer-related deaths worldwide [4]. As such, the search for clinically relevant diagnosis and prognosis biomarkers is of great importance.

Diagnostic tools using DNA methylation biomarkers are already employed in clinical practice, and these include methylation of the SEPT9 gene (commercialized as Epi-proColon), and methylation of the VIM gene (commercialized as ColoSure test) [110]. Although these two are the most studied, they present some caveats, such as the low sensitivity in identifying adenomas, in the former, and the lack of Food and Drug Administration (FDA) approval in the latter [110].

There are several candidate biomarkers that are in study for colorectal diagnosis using both blood and stool-based samples, which reveals the ever increasing need to create such tools [110]. In this study we identified 75 regions that are aberrantly altered in the normal-tumor transition, and although detected in tumor samples, can potentially be candidate diagnostic biomarkers of colon adenocarcinoma [110].

The case for prognostic biomarkers in colorectal cancer is quite different from diagnostic biomarkers, as there are not any that have entered clinical guidelines [110]. Nonetheless, the need for such tools is ever increasing, and there are several ones in study as candidate prognostic biomarkers in colorectal cancer. Here we found a total of nine regions (eight epi-blackholes and one epi-hotspot) that were able to not only predict survival in stage-III colon tumor patients, but also divide them into two groups with distinct survival distributions. These regions could potentially be candidate prognosis biomarkers for colon adenocarcinoma. Furthermore, none of the identified regions overlaps with genes already used as candidate prognostic or diagnostic biomarkers in colorectal cancer.

It is important to note that the analysis herein performed uses data from colon adenocarcinoma samples alone, and the biomarkers are generally used to identify colorectal cancer. As such, it is not certain that the same results would be obtained if using samples from both colon and rectum tissues. This is especially relevant in the identification of DNA methylation biomarkers in this disease, since different parts of the digestive tube are differently

exposed to carcinogens that can cause epimutations. Hence, this type of analyses might be biased if the utilized data is not representative of the whole colorectum.

It was interesting to find that of 5253 differentially expressed genes between normal and stage-I colon tumor samples, 93% were explained by epi-hotspot variability. This type of outcome might suggest that, in fact, the aberrant patterns of gene expression in stage-I colon adenocarcinoma are, somehow, connected to aberrant methylation in epi-hotspot regions. Furthermore, by performing a GAGE analysis to the differentially expressed genes that were related to epi-hotspot regions, we found that several GO terms were enriched comparing to the baseline of aberrantly expressed genes, 121 being up-regulated in stage-I tumor, and 457 down-regulated. Finding such amount of enriched GO terms might also suggest that epi-hotspots are related to altered cellular processes in colon tumorigenesis, even though our data is insufficient to make such conclusion. Some of these enriched GO terms are expected in colon tumorigenesis. For example, we detected a strong up-regulation of Wnt signaling pathway. In colon adenocarcinoma, mutation of the tumor suppressor gene Adenomatous polyposis coli (APC) is usually an initiating event in the majority of both sporadic and familial colon and rectal cancers [111]. APC inactivation tends to drive hyperactivation of the WNT signaling pathway, which is thought to be a major tumorigenic event in nearly all colorectal cancers. The WNT signaling pathways is involved in many cellular processes in both normal and tumorigenic settings, such as cell proliferation, cell migration, asymmetric cell division, amongst many others [111]. It is interesting to observe that several of the top enriched upregulated GO terms are related do cell polarity, and several downregulated GO terms are related to ion transport, as the latter usually impacts the former [112]. This is not surprising, since dysfunction of several ion channels are related to initiation and development of tumor of the gastrointestinal system [112]. Since our study does not demonstrate causality, it would be relevant, in the future, to understand if aberrant methylation in epi-hotspot regions are associated to ion transporter dysfunction, and consequently altered cell polarity.

5.7 Breast Invasive Carcinoma

Breast cancer is the second most frequently diagnosed cancer for both sexes combined, and the most diagnosed in females, where it stands as the leading cause of cancer-related death [4]. In a disease with such burden, diagnostic and prognostic tools are critical. The most frequently used method for breast cancer screening worldwide is the mammography [113]. This type of screening is shown to decrease breast cancer mortality, especially in women aged

between 50 and 69 years old. It is important to mention that the harms of overdiagnosis in breast cancer are still debated in the scientific community [113,114]. Nonetheless, the benefits of early detection in breast cancer far surpass the possible harms, as prognosis in this diseases largely depends on early detection [113–115].

In the present analysis we detected 76 regions that are abnormally methylated in stage-I breast tumor samples, consisting of potential candidate diagnostic biomarkers for breast tumor.

The primary challenge in breast cancer, immediately after diagnosis, is the assessment of patient prognosis and predicting treatment response [116]. The former can be evaluated by clinicopathological factors, such as tumor staging, estrogen receptor (ER) status, progesterone receptor (PR) status, and Human epidermal growth factor receptor 2 (HER2)/neu status, but also by molecular biomarker tools such the Oncotype DX or MammaPrint. The later two are amongst the most extensively validated prognostic biomarkers in breast cancer and are increasingly being employed in clinical practice. The Oncotype DX is a multi-analytic tool that evaluates expression levels of 21 genes, of which 16 are cancer-linked and 5 are controls. The relative expression of the former compared to the controls allows for the calculation of a recurrence score, which is a continuous variable that segregates breast cancer patients into low, intermediate, or high-risk groups. The MammaPrint, on the other hand, is considered to be extremely useful not only in predicting disease recurrence, but also in aiding in the decision making of treatment choice. This tool makes measures, through microarray, the expression levels of 70 genes, which allows for the division of breast cancer patients into low-risk and high-risk groups [116].

It is very interesting to observe that the most validated prognostic biomarker tools in breast cancer are multi-analytic, rather than single-variable centered. In this study we identified one epi-hotspot region, which consisted of three different CpG sites, that was able to divide patients into two groups with distinct prognosis.

We also identified 3079 differentially expressed genes between normal and stage-I breast tumor samples, and only 4% of which could be explained by epi-hotspot variability. In contrast, none of the non-differentially expressed genes could be explained by epi-blackhole variability. The low number of genes that could be explained by epi-hotspot variability lead to a GAGE analysis with only 36 enriched GO terms, 30 of which where up-regulated. At first, these results might seem surprising since statistical power in this cohort is not a major issue.

However, in fact, it is relevant to state that "breast tumor" is a term that refers to many different diseases. For example, a luminal breast tumor's pathological mechanism is entirely different from a triple negative's one. In our analysis, we analyzed cohorts that had a mix of different breast cancer subtypes, which affects the true identification of aberrantly methylated regions, all type of correlations, and, most importantly, in the assessment of prognosis. Although further investigation needs to be made in the different breast tumor subtypes, it was interesting to find an epi-hotspot that could predict survival in stage-III breast tumor patients. Nonetheless, solid conclusions are not possible at the time since it is unknown if the region is really distinguishing survival of two clusters of patients with different diseases.

5.8 Cholangiocarcinoma

Bile duct cancer, also known as cholangiocarcinoma, is a low incidence tumor, representing approximately 3% of all tumor in the gastrointestinal system [117]. However, the number of cholangiocarcinoma diagnosis are increasing, in the entire world [117,118].

This malignancy presents several challenges, the first being that it is an asymptomatic disease during most of its course, which makes early detection extremely difficult [118,119]. In fact, the median survival of cholangiocarcinoma patients is less than two years, and the five-year survival rates are around 20%. Another challenge in this disease is the fact that the only potentially curative treatment is surgery, but only in early stages of this disease. The treatment in later-stages is limited to cisplatin and gemcitabine, which seem to have only mediocre effects. It is clear that the biggest problem in this illness is early detection, and therefore the identification of diagnostic biomarkers is increasingly more urgent [117,118].

It is already been described that aberrant DNA methylation might constitute a possible way to detect early cholangiocarcinoma [118]. In this study we found 70 aberrantly methylated regions in the normal to stage-I tumor transition, which can be potential candidate multi-analytic biomarkers in this disease. Interestingly, we also found 10748 differentially expressed genes, of which virtually all (except only 1%) could be explained by epi-hotspot variability. This corroborates, in a way, with the existing literature, since it has been showed that aberrant DNA methylation, possibly caused by chronic inflammation, might be a key event that triggers an accelerated proliferation of biliary epithelial cells [118]. Although intriguing, the results in this cohort need to be taken extremely cautiously, since statistical power is considerably low. This also applies to the high amount of detected epi-blackholes, since the low number of samples in

both normal and tumoral groups greatly increases the chance of finding a region that is likely not an epi-hotspot.

5.9 Esophageal Carcinoma

Esophageal cancer is the seventh most diagnosed cancer worldwide, with 572 thousand new cases every year [4]. Although incidence of this malignancy is declining, esophageal cancer is still responsible for approximately 509 deaths a year. In fact, it is estimated that, in 2018, 1 in every 20 cancer-related deaths were due to esophageal cancer [4]. It is thought that a key contributing factor (although not the only one) from which esophageal cancer develops is gastroesophageal reflux disease and/or Barrett's esophagus [120]. Nonetheless, the mechanisms are not entirely clear, as only one in every 20 patients with Barrett's esophagus develops esophageal cancer. Furthermore, is also known that individuals with chronic gastroesophageal reflux are at risk for developing Barrett's esophagus [120]. As such, screening methods that would complement endoscopic surveillance to identify early stage esophageal cancer would be valuable.

In this analysis we identified 2 epi-hotspots and 2420 epi-blackholes. Statistical power in this cohort is extremely low, so it is important to interpret results cautiously. Nonetheless, it is curious to observe that the two identified epi-hotspots can explain the variation of 1792 differentially expressed genes between normal and stage-I tumor samples.

We also found that epi-blackholes could predict variation of 3183 non-differentially expressed genes, and some could also predict survival in stage-III esophageal cancer. However, due to low sample-size in this cohort, the number of detected epi-blackholes might be inflated which, by means of probability, could lead to positive results merely by chance. Therefore, if we take the results from this cohort as part of a greater compilation of results, there is some evidence that suggests that epi-hotspots and epi-blackholes might be regulators of (or maybe regulated by) gene expression during tumor initiation. Moreover, although not conclusive, this portion of the study may provide some substantiation that methylation of certain genomic regions may be fair predictors of prognosis in later stages of disease.

5.10 Head and neck squamous cell carcinoma

Head and neck squamous cell carcinoma is a term used to designate a group of malignancies that originate from the oropharynx, hypopharynx, lip, nasopharynx, larynx, or

oral cavity [121]. This group of diseases represents the sixth most common malignancy worldwide, as more than 500 thousand new cases are diagnosed each year. This disease is extremely heterogeneous and the main therapeutic options usually include surgery and/or radiotherapy [121,122]. With this type of treatment there is great benefit to early detection, which why diagnostic biomarkers are also useful. Here we identified 203 aberrantly methylated regions in stage-I tumor, which could potentially be used as candidate diagnostic biomarkers. We also found 1528 differentially expressed genes, of which almost all (94%) could be explained by epi-hotspots. This clearly suggests that, during tumor initiation, altered methylation in certain regions is related to altered patterns of gene expression.

When submitting the epi-hotspot explained differentially expressed genes to a gage analysis we observed an up-regulation of several GO terms associated with extracellular matrix remodeling. It is not surprising to see such up-regulation, as this is an event that has already been described as a key event in tumor initiation, not only in the head and neck carcinomas, but also in others tumor types [123]. However, it is interesting to observe that this event is somehow possibly related to DNA methylation in epi-hotspots. Additionally, we also identified 181 non-differentially expressed genes that could be explained by epi-blackhole variability. The data from this cohort also seems to suggest that epi-hotspot and epi-blackhole regions might be involved in alteration and maintenance of gene expression patterns during tumor initiation.

In this analysis we also found that eight epi-hotspots and three epi-blackholes were good predictors of survival in stage-III patients. Although such results are highly relevant, it is important to remind that HNSC is a group of several tumors, and a candidate prognostic biomarker would be more trustworthy if it took into account the specific ailment, rather a the generic classification of the disease.

5.11 Hepatocellular carcinoma

Liver cancer is the sixth most frequently diagnosed cancer, with approximately 841 thousand new cases and 782 thousand deaths in 2018, worldwide, the majority of which consist of hepatocellular carcinomas (accounting for 75% - 85% of all cases) [4].

The high lethality of this disease creates an increasing demand for reliable diagnostic and prognostic biomarkers. Interestingly, hepatocytes constantly adapt to environmental conditions, which are ever changing in the liver [124]. Several factors such as viruses, xenobiotics,

changes in metabolic processes, and others, have an effect in the hepatic methylome, which in turn could contribute to tumorigenesis [124].

The high volatility in DNA methylation in hepatocytes is clearly an opportunity to identify regions that could differentiate hepatocellular carcinoma samples from normal liver samples. In fact, we identified 208 abnormally methylated genomic regions and only 11 epi-blackholes. This difference in the number of detected epi-hotspots and epi-blackholes was expected, since hepatocytes have very fluctuating DNA methylation patterns [124].

We also found 2921 differentially expressed genes in stage-I LIHC tumor samples, of which about half (1355) could be explained by epi-hotspot variability. This also provides evidence that epi-hotspots might be influencing (or being influenced by) gene expression alterations in tumor initiation.

No non-differentially expressed gene were found to be explained by epi-blackholes, which could be explained by the volatility in hepatocyte's epigenome.

Although no epi-blackhole was a good predictor of survival, we found one epi-hotspot that was able to differentiate stage-III LIHC patients into two groups with distinct survival distributions. Even though this is a region that could potentially be used as a candidate prognostic biomarker, it is important to stress that this analysis deals with a group of LIHC samples from different etiologies. It is well known that patient prognosis is different in different types of LIHC [125]. So, in the development of a good prognostic biomarker, this would have to be accounted.

5.12 Lung Squamous Cell Carcinoma

Lung cancer leads the ranking in both cancer incidence and mortality, worldwide, representing 11.6% of all cancer diagnoses and 18.4% of all cancer-related deaths [4]. LUSC used to be the most common type of lung cancer, but in the last few decades, with the decrease in smoking behaviors (which constitute the main risk factor for this type of cancer), incidence in this disease has been gradually falling [126].

Even though LUSC diagnoses have been decreasing, it is still an incredibly deadly form of lung cancer, and diagnostic biomarkers able to screen for early-stage disease would be clinically advantageous.

In this study we found 97 epi-hotspots and 18 epi-blackholes. Furthermore, we identified 5577 differentially expressed, 706 of which could be explained by epi-hotspot variability, and 22 non-differentially expressed genes that could be explained by epi-blackhole variability. A higher number of epi-hotspots than epi-blackholes was expected in the LUSC cohort, since this is a disease highly associated with smoking, which it is known to cause a large mutational and epi-mutational load [126,127]. This might also be the reason for detecting 328 altered GO terms related to epi-hotspots.

Additionally, we found that one epi-hotspot was a good predictor of prognosis in stage-III LUSC patients, which, in such a deadly ailment, could potentially constitute a helpful candidate prognostic biomarker.

5.13 Thyroid Carcinoma

Thyroid cancer is the ninth most frequently diagnosed cancer, worldwide, with an incidence of 567 thousand cases per year [4]. Although somewhat frequent, THCA's mortality is much lower than the previously discussed diseases, being estimated that mortality rates in THCA are approximately 0.4 to 0.5 in both men and women [4].

Although THCA is a disease with a fair prognosis for most patients, about 10% of all cases may develop into more undifferentiated tumors, and 2% may progress to anaplastic carcinoma [128]. Although this progression does not happen in most THCA patients, mortality highly increases with the decrease of tumor differentiation [128]. Additionally, the etiology in this disease is not completely understood, and the only clear and well documented risk factor is ionizing radiation [4,128]. Therefore, the study of early events in DNA methylation might be of use not only for early-stage THCA detection, but also to gain further insights about the tumorigenic process of this disease.

In this study we identified 7 epi-hotspot regions and 127 epi-blackholes. Furthermore, we detected 2476 differentially expressed genes, 44 of which could be explained by epi-hotspot variability. The higher number of epi-blackholes compared to epi-hotspots was expected, since this is a slow-developing tumor, that is not as highly exposed to epi-mutagens as other types of tumors.

By submitting the previously identified 44 genes to a GAGE analysis, we only found 16 enriched GO terms, all downregulated. This was not entirely surprising, due to the low number of genes in the analysis's input. Additionally, we also detected 39 non-differentially

expressed genes that could be explained by epi-blackhole variability, which provides additional evidence that DNA methylation in these regions is linked to gene expression during tumor initiation.

Unlike in several other cohorts, we did not find any region that was able to predict prognosis in stage-III THCA patients successfully. We believe that this is due to the fact that THCA is a slow-progressing tumor with a very good overall prognosis, the exception being the infrequent anaplastic carcinomas [4,128]. This does not mean that THCA patients cannot be stratified into risk categories, but the nature of this disease renders most cohorts, by means of probabilities, highly homogeneous. Such cohorts provide the researcher great benefits when searching for relevant characteristics (like epi-hotspots) that can represent the target population, but also great difficulties when seeking to partition the population into clusters of individuals that are similar to each other, but distinct from other clusters.

Histopathological analysis of THCA samples already provides a fair predictor of prognosis, yet, it would be useful to specifically analyze poor-prognosis tumors, such as undifferentiated and anaplastic thyroid tumors, and identify biomarkers that could subdivide these high-risk patients into different prognostic categories.

5.14 Papillary renal cell carcinoma

Kidney cancer represents the fourteenth most diagnosed cancer, worldwide, with an estimated incidence of 403 thousand cases per year [4]. Regarding mortality, it occupies the sixteenth place in the ranking tables, causing approximately 175 thousand deaths per year [4].

There are several types of malignancies that originate in the kidney, the most common type being the renal cell carcinomas, which represent at least 85% of cases [129,130]. Most renal cell carcinomas can be subclassified into kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), and kidney chromophobe (KICH). There are also other renal cell subclassifications, but these are very rare, representing less than 1% of all renal cell cases [129,130].

KIRP represents about 15% - 20% of all renal cell carcinomas, being the second most frequent histological subtype [131]. In our analysis we found 29 epi-hotspots and 17 epi-blackholes. However, only 66 of the 4326 (~2%) differentially expressed genes could be explained by epi-hotspots, and none of the remaining 15331 non-differentially expressed genes could be explained by epi-blackholes. Additionally, no enriched GO term was detected when

submitting the 66 genes to a GAGE analysis. This may be because two distinct types of KIRP exist: type I and type II. The two types of KIRP are not only histologically different, but also present distinct genetic patterns, with characteristic gene mutations and chromosomal abnormalities[131]. Furthermore, even the prognosis is unequal in both tumor types, being good in type I, and poor in type II. Our cohort consisted in a balanced mix of type I, type II, and unspecified tumor types, which could explain the fact that we did not find many relationships between gene expression and epi-hotspots and epi-blackholes.

We also found two epi-blackholes that could predict survival in stage-III KIRP patients, but it is relevant to mention that we did not examined if these regions where accidentally segregating the patients based on tumor type. If that is the case, these two epi-blackholes might not be great prognostic biomarker, as both tumor types already have distinct prognostic patterns.

5.15 Kidney Renal Clear Cell Carcinoma

KIRC is the most frequent type of renal cell carcinoma, and accounts for 70-80% of all kidney originating neoplasms [129,132].

In this disease we found only 8 epi-hotspots and 3 epi-blackholes, which could suggest that DNA methylation in this disease is highly heterogenous. We theorize that this heterogeneity is not random, and that there are several clusters of KIRC patients, with distinct methylation patterns. To identify the true epi-hotspots and epi-blackholes in this malignancy, one would have first to identify, if they exist, the true sub-groups of KIRC patients. Nonetheless, it would be clinically relevant if any of the identified regions could be used as a potential candidate diagnostic biomarker, since the 5-year survival rate of this disease is approximately 92% if detected early, but only 67% if the tumor has spread locally, and 12% if it has already metastasized [132].

Interestingly, although we detected 4588 differentially expressed genes, none could be explained by epi-hotspot variability. Likewise, none of the remaining non-differentially expressed genes could be expressed by epi-blackholes. One explanation for such result might be that, unlike other tumor, KIRC is an extremely heterogenous disease, regarding not only gene mutations, but also chromosomal abnormalities [132]. Therefore, although the sample size of this cohort was sufficient to study it as one uniform disease, it may not be enough to detect associations in a highly heterogenous malignancy.

Additionally, we did not find any region to be a good prognostic predictor, but this might be explained by the low number of identified epi-hotspots and epi-blackholes. With the rigorous filters and restrictions here applied, it would be unlikely to yield a positive result, with such a low number of input regions.

5.16 Pancreatic Adenocarcinoma

Pancreatic cancer is the twelfth most frequently diagnosed cancer, in the world, with around 459 thousand cases per year [4]. This is an extremely deadly condition, with a number of fatalities that almost matches the incidence (432 thousand deaths in 2018) [4].

The poor prognosis in this disease is largely caused by late diagnosis, as most patients with this malignancy stay asymptomatic throughout most of the course of the tumor development and progression [133]. Reliable early-detection screening tools would be useful, since incidence of pancreatic adenocarcinoma seems to be rising every year, especially in developed countries [4,133].

In this study we identified 3 epi-hotspots and 898 epi-blackholes. Such results were most likely due to the low sample-size in this cohort, which reduced statistical power to such extent, that most regions were just extremely unlikely to be epi-hotspots, giving rise to a high number of epi-blackholes.

This was the only cohort in which no differentially expressed genes were found. It would be, certainly, preposterous to conclude that genetic patterns are maintained during tumor initiation, and it is highly probable that such result is due to an inadequate alpha level in this low statistical power setting.

We also observed that 8268 non-differentially expressed genes could be explained by epi-blackhole variability. However, this result is to be taken carefully, because since all genes were considered non-differentially expressed, and the number of epi-hotspots is very high, it is likely that, even with rigorous methodological restrictions, most of the detected links are simply due to chance. Nonetheless, true-positive relations surely exist but, with this analysis, we cannot detect which ones they are.

Additionally, due to the very low number of stage-III PAAD patients, the survival analysis did not generate any output.

5.17 Lung adenocarcinoma

As previously described, lung cancer is not only the most diagnosed cancer type, but also the deadliest, worldwide [4]. With the decreased smoking behaviors, LUAD became the most prevalent form of non-small cell lung cancer, a position previously occupied by LUSC [127,134,135].

In this cohort we found 117 epi-hotspot regions and 3656 differentially expressed genes. However, no relation was found between the former and the later. This was someway surprising, since aberrant DNA methylation is a documented early-event hallmark in lung cancer [135]. It is also been documented that these events trigger dysregulation of several oncogenes and tumor suppressor gene. For this reason, we were expecting to see at least some relation between epi-hotspots and differentially expressed genes. We did, however, found that 10 non-differentially expressed genes could be explained by epi-blackholes, a result that is closer to what is already documented in the scientific literature [135].

In this cohort, we also found six epi-hotspots and two epi-blackhole regions that were able to discriminate stage-III LUSC patients into two groups with distinct survival distributions, all of which could possibly be used as a potential candidate prognostic biomarker in this deadly disease.

5.18 Similarity in Epi-hotspots

One of the goals we ought to achieve was to understand if there were epi-hotspot regions that were common in more than one cohort, and if the analyzed malignancies could be assembled into clusters. This was achieved by performing a complete-type hierarchical clustering analysis using the epi-hotspot region's overlapping percentage as a similarity unit.

We identified five major clusters, three of which consist of a singular cohort. The three single-element clusters were THCA, ESCA, and PAAD, which were expected to be assembled in this way due to the very low number of identified regions. The results might seem to suggest that, regarding DNA methylation in epi-hotspots, these are diseases entirely different from each other and from all others. However, the detected difference is not a true difference, since the low number of input regions diminishes the probability of finding an overlap between them.

Apart from the three clusters mentioned above, it is interesting to observe that the overlapping percentage of epi-hotspots is very small, even in the most similar cohorts. This is in accordance with previous unpublished work from our group that showed that changes in

DNA methylation patterns, during tumor initiation, seemed to be specific to the cell-of-origin, rather than common to all malignancies.

Interestingly, the two renal cell carcinomas, KIRP and KIRC, were clustered together, sharing two epi-hotspot regions, one of which overlaps with two genes: SLC2A9, and DRD5. Curiously, both of these genes have been recently shown to be tumor growth suppressors [136,137]. In 2017, Leng and colleagues, showed that activating DRD5 (dopamine receptor D5) in tumoral cells inhibited the mechanistic target of rapamycin (mTOR) pathway by inducing reactive oxygen species (ROS) production, thus stimulating autophagy and, consequently, autophagic cell death [136]. Recently, in 2019, Han X. *et all*, showed that overexpression of the uric acid transporter SLC2A9, in LIHC cells, inhibited the expression of caspase 3, thus inducing apoptosis [137]. Curiously, none of these two genes was differentially expressed in KIRC or KIRP cohorts. Furthermore, contrarily to the discovery, by Han X. *et all*, that SLC2A9 is downregulated in LIHC cells, we did not find any alteration regarding expression levels in the stage-I LIHC cohort.

Although it is interesting that both renal cell carcinomas were clustered together, it is important to note that the similarity percentage regarding epi-hotspots is only 4%. Therefore, it is not entirely possible conclude that these two malignancies constitute a real epi-hotspot cluster.

The second multi-cohort cluster clearly consists of two major child clusters, the first being comprised of CHOL, COAD, and LUAD, with six common epi-hotspot regions. These epi-hotspots overlapped with five genes, none of which was found to be differentially expressed in all cohorts. The Empty Spiracles Homeobox 2 (EMX2) and GRIA4 genes were found to be differentially expressed, but only in colon adenocarcinoma.

It was surprising to detect differential expression of EMX2 only in colon cancer, because it has been shown that this gene is usually downregulated in lung cancer [138]. Nonetheless, it has also been shown that colorectal tumor tissues have decreased expression levels of EMX2, which our results also support [139] . Furthermore, it has also been shown that EMX2 expression levels are associated with worse prognosis in colorectal cancer [139]. Although we did not assess gene expression as a survival predictor, the epi-hotspots that overlapped with EMX2 were not good predictors of prognosis in stage-III colon adenocarcinoma. Additionally, the only epi-hotspot that was able to distinguish stage-III COAD patients into two groups with different prognosis was not associated with the downregulation of EMX2 in stage-I.

The second child cluster from the major multi-cohort cluster entails BRCA, HNSC, LIHC, and LUSC. We observed that two epi-hotspot regions, that overlapped with the genes WRAP73 and COL11A2, were common between the four cancers. The WRAP73 gene was overexpressed, in a highly statistically significant manner, in LUAD, LIHC, CHOL, and ESCA. This gene's product is a constitutive part of the centrosome, and is essential to correct mitotic spindle assembly [140]. In theory, it is not surprising that tumoral cells have higher expression levels of WRAP73, and it is possible that, in fact, all our analyzed cohorts exhibit this pattern, although not detectable in high statistical restrictions.

The COL11A2 gene was only differentially expressed in CHOL, but due to the low statistical power and the very high number of detected differentially expressed genes in this cohort, it is hard to draw trustworthy conclusions.

5.19 Similarity in Epi-blackholes

We also aimed to understand how the analyzed diseases related to each other concerning epi-blackhole location. Therefore, we repeated the previous hierarchical clustering analysis, but this time using as a similarity unit the overlapping percentage of epi-blackholes.

By doing so, we observed five single-cohort and two multi-cohort clusters. The THCA, LUSC, KIRC, HNSC, and BRCA cohorts were, in fact, expected to be clustered individually, as these were cohorts with a low number of detected epi-blackholes.

The first multi-cohort cluster consisted of KIRP and LIHC, with only one common epi-blackhole region that not overlapped with any protein-coding gene. Although these two diseases were clustered together, the very low similarity does not allow for the assumption that KIRP and LIHC are alike diseases concerning epi-blackholes.

The second multi-cohort cluster comprises LUAD, COAD, PAAD, CHOL, and ESCA, which are the cohorts with the highest number of detected epi-blackholes. It is important to note that drawing conclusions from this multi-cohort cluster might be misleading. With such a high number of epi-blackhole regions in these cohorts, the amount of detected overlaps radically increases. In fact, it is much likely that these diseases were clustered together, not for a true biological reason, but simply because of the very high probability of overlap by chance.

5.20 Limitations

It is relevant to mention that, like all studies, this analysis has several limitations and caveats that, if not taken seriously, can lead to incorrect assumptions of the tumorigenic process. In this subsection, we try to summarize some of the most critical limitations of our research:

- In this analysis we use both normal and tumoral samples in an independent manner, a limitation imposed by the absence of data from normal-matched samples. Although it is possible to understand, by the means of statistical inference, the changes that happen during tumorigenesis, we cannot delineate the true alterations that happen in a normal cell that is acquiring tumorigenic traits.
- It is difficult, if not impossible, to truly assess if normal samples here utilized are indeed normal. Usually, these samples are obtained during surgical removal of a tumor, being possible that these cells already have acquired some pre-tumoral characteristics. This is especially relevant in epigenetic analyses since changes in the cell's epigenome can happen before tumorigenesis.
- It is well known that individual characteristics like gender, age, or weight, as well has environmental factors and individual behaviors such as diet, smoking habits, exercise, sleeping patterns, and even chronic stress, affect the epigenome. This adds a layer of complexity when analyzing DNA methylation patterns of tumor-patients as a group, particularly in independent-group comparisons, like the one herein performed, because it is not possible to truly understand if a specific difference is due to the variable of interest (normal versus tumoral) or due to an individual characteristic that is randomly different in one of the groups.
- One of the most evident limitation in this study is the small size of the cohorts here analyzed. Low sample-size inherently causes a low power during hypothesis testing. It is critical that statistical power is increased, so that the probability of making a type-II error is minimal. It is clear that, in this analysis, many true epi-hotspots and epi-blackholes were not detected, primarily due to low-power settings, but also due to our attempt to contain type-I errors.
- Type-I error minimization and alpha-level setting is another caveat in our analysis. In fact, in the twelve studied cohorts, alpha-level was generally set to

5%, in an attempt to reduce type-I errors. However, it is relevant to note that alpha-level reduction naturally increases the probability of type-II errors. In a low-power setting, a low alpha level additionally increases the probability of type-II errors. For example, in the ESCA cohort, it is clear that a very high number of epi-hotspot regions are undetected, which is even more evident in the artificially high number of epi-blackholes in this cohort. In fact, alpha-level should have been set dynamically, depending on statistical power. This would be even more critical in this study, since we considered that type-II errors were the most serious.

- It is well known that cancer is not a single disease, but a set of many different diseases. In this study we analyzed cohorts of samples from tumor-patients that were, fundamentally, grouped by the organ from which the primary tumor arose. Although this is the only a practical way of researching using publicly available data, this simplification is highly fallacious. It is well known that each of the diseases herein analyzed are, in fact, a very heterogeneous mix of tumor subtypes that are completely different from each other, regarding the molecular mechanisms, pathophysiology, therapeutics, and even prognosis. Indeed, we are detecting general results that would, in theory, be common to at least most of the tumor subtypes. However, there is a strong possibility that one or more cohorts are enriched with certain tumor subtypes, thus generating biased results.
- Epi-blackhole regions were here detected by the probability of not being epi-hotspots. This give us a rough view of the regions that are most likely to be kept unchanged during tumorigenesis, regarding DNA methylation. Yet just because a region is unlikely to be an epi-hotspot, it is not possible to know for sure that it is an epi-blackhole.
- The epi-hotspot and epi-blackhole regions that were identified in this study are really candidate regions. The lack of normal matched samples does not allow for the detection of individual changes of the methylome, which is why it was not possible to calculate methylation thresholds that could be applied individually.
- When identifying altered patterns of DNA methylation and gene expression, we considered that these were changes that occurred in most patients, but in reality, it is likely that patients can be clustered into groups with different genetic and epigenetic patterns.

- When studying how DNA methylation is linked to gene expression, during tumorigenesis, we greatly under-detect these relationships, mainly because, in most cases, we do not have enough evidence to support it. This leads to an erroneous count in the number of genomic regions that could be potentially linked to gene expression.
- One of our goals was to understand which cellular pathways could be influenced by DNA methylation of epi-hotspots during tumor initiation, which was investigated trough a GAGE analysis. This was performed by detecting an enrichment of the group of genes that were related with epi-hotspots, thus although we gained insights of what GO-terms were the most enriched, relatively to the background of genes, we do not know if DNA methylation in epi-hotspots is really influencing these processes. Furthermore, even regarding the unknown true positives, it is not possible to understand if it is a true biological association or just a coincidence caused by the tumorigenic process itself.
- When analyzing the relations between DNA methylation in certain genomic regions and patterns of gene expression, during tumorigenesis, there is an intrinsic bias caused by the number of tested regions. It is, by means of probabilities, much more likely to find relations in a cohort that has more variables to be tested. For example, in cohorts like ESCA, where more than two thousand epi-blackholes were found, there is an increased chance of detecting a relation between epi-blackholes and gene expression.
- Since genomic data is constantly updated, gene names and symbols change over time. Although we strove to avoid mistakes, there is always some risk that an error was made when refereeing to or describing a gene.
- In some survival analyses, we detected a low number of potential prognosis biomarkers. This is not because such biomarkers do not exist, but because the characteristics of the sample make them undetectable. For example, in diseases with an extremely poor or fair prognosis, it is rather difficult to generate groups of patients with distinct survival distributions.
- Although this work stems from the theorization that, during tumorigenesis, DNA methylation in specific genomic regions would impact gene expression, the generated results do not imply causality between the studied events. So, even though this analysis presents evidence that there is, in fact, a relation between

the variable, it is not possible to conclude that one is influencing the other, or if both are being impacted by other factors.

- Reliable prognostic biomarkers are in increasing demand in cancer since these allow for better clinical decisions. While we have identified several potential candidates for further study, we did not search for predictors of therapeutic response. This type of biomarkers is even more relevant than prognosis biomarkers since they allow for a sustained decision of therapeutics. By using both prognostic biomarkers and predictor of therapeutic response, the risk/reward ratio of applying given therapy is dramatically decreased.
- Probably one of the biggest caveats of this work is the limited in-depth detailed analysis of the generated results. This limitation was mainly caused by the vast amount of generated data. A thorough examination of each individual region here identified, as well as their particular relation with each gene, remains to be done.
- In this work, several diagnostic and prognostic potential candidate biomarkers were identified. A good screening and diagnostic tool is one that can be utilized non-invasively, such as by using blood, serum, feces, or urine. The biomarkers here identified were detected using normal and tumor samples that were obtained directly from the affected tissue. While such biomarkers are still possible candidates, we do not know if these are detectable in a non-invasive way.
- Lastly, to draw more solid conclusions, this study needs validation of both the methodology and the results.

Chapter 6 Conclusion

It is well known that DNA methylation is a key epigenetic mechanism for the modulation of cellular processes that determine cell function and fate. Since tumor cells consistently have aberrant DNA methylation patterns, we hypothesized that certain genomic regions would be susceptible to alterations in DNA methylation while others would be resistant to it. Our data provides evidence to support this hypothesis, suggesting that such regions are systematically detectable in various tumor types.

We also hypothesized that one of the ways that such regions could exert selective pressure in the tumor was by influencing gene expression. Even though it was not possible to determine causality, our work clearly shows that, during tumor initiation, DNA methylation patterns in both epi-hotspots and epi-blackholes are related to gene expression, which is in accordance with our hypothesis.

One of our consequent biological questions was if epi-hotspots and epi-blackholes occurred in similar regions in every tumor type, for which we demonstrated that although some similarity exists between tumor types, the genomic location of such regions is mainly specific to the tissue-of-origin.

The fact that epi-hotspot and epi-blackhole regions are detectable, related to tumorigenic gene expression patterns, and are specific for each tumor type, lead us to hypothesize that, if these regions do exert selective pressure in the tumor clones, then different methylation patterns would lead to tumors with different degrees of aggressiveness, which would be translated into different patient prognoses. Again, our results support this hypothesis, demonstrating that DNA methylation in several of these regions could clearly distinguish patients with poor prognosis from patients with favorable prognosis.

In summary, our work provides evidence that, during tumorigenesis, there is a tissue-specific regional susceptibility to altered patterns of DNA methylation and, in parallel, regional conservation of normal patterns. Moreover, such regions seem to be associated with the alteration and maintenance of tumorigenic gene expression patterns and can be used to predict stage-III patient prognosis.

Bibliography

1. Hassanpour, S. H. & Dehghani, M. Review of cancer from perspective of molecular. *J. Cancer Res. Pract.* **4**, 127–129 (2017).

2. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).

3. Clark, W. Tumour progression and the nature of cancer. *Br. J. Cancer* **64**, 631–644 (1991).

4. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).

5. Wang, C.-C., Jamal, L. & Janes, K. A. Normal morphogenesis of epithelial tissues and progression of epithelial tumors. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **4**, 51–78 (2012).

6. Macara, I. G., Guyer, R., Richardson, G., Huo, Y. & Ahmed, S. M. Epithelial Homeostasis. *Curr. Biol.* **24**, R815–R825 (2014).

7. Dancsok, A. R., Asleh-Aburaya, K. & Nielsen, T. O. Advances in sarcoma diagnostics and treatment. *Oncotarget* **8**, (2017).

8. Hoang, N. T., Acevedo, L. A., Mann, M. J. & Tolani, B. A review of soft-tissue sarcomas: translation of biological advances into treatment measures. *Cancer Manag. Res.* **Volume 10**, 1089–1114 (2018).

9. Taylor, J., Xiao, W. & Abdel-Wahab, O. Diagnosis and classification of hematologic malignancies on the basis of genetics. *Blood* **130**, 410–423 (2017).

10. Allart-Vorelli, P., Porro, B., Baguet, F., Michel, A. & Cousson-Gélie, F. Haematological cancer and quality of life: a systematic literature review. *Blood Cancer J.* **5**, e305–e305 (2015).

11. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).

12. Montano, N. *et al.* Tumors of the peripheral nervous system: analysis of prognostic factors in a series with long-term follow-up and review of the literature. *J. Neurosurg.* **125**, 363–371 (2016).

13. Leonardi, G. *et al.* Cutaneous melanoma: From pathogenesis to therapy (Review). *Int. J. Oncol.* (2018) doi:10.3892/ijo.2018.4287.

14. Peterson, C. M., Buckley, C., Holley, S. & Menias, C. O. Teratomas: A Multimodality Review. *Curr. Probl. Diagn. Radiol.* **41**, 210–219 (2012).

15. Gazdar, A. F., Bunn, P. A. & Minna, J. D. Small-cell lung cancer: what we know, what we need to know and the path forward. *Nat. Rev. Cancer* **17**, 725–737 (2017).

16. Oosterhuis, J. W. & Looijenga, L. H. J. Human germ cell tumours from a developmental perspective. *Nat. Rev. Cancer* **19**, 522–537 (2019).

17. Nowell, P. The clonal evolution of tumor cell populations. *Science (80-.).* **194**, 23–28 (1976).

18. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).

19. Caldas, C. Cancer sequencing unravels clonal evolution. *Nat. Biotechnol.* **30**, 408–410 (2012).

20. Wang, M. *et al.* Role of tumor microenvironment in tumorigenesis. *J. Cancer* **8**, 761–773 (2017).

21. Anderson, K. *et al.* Genetic variegation of clonal architecture and propagating cells in leukaemia. *Nature* **469**, 356–361 (2011).

22. Maley, C. C. *et al.* Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat. Genet.* **38**, 468–473 (2006).

23. Sahai, E. *et al.* A framework for advancing our understanding of cancer-associated fibroblasts. *Nat. Rev. Cancer* **20**, 174–186 (2020).

24. Gonzalez, H., Hagerling, C. & Werb, Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. *Genes Dev.* **32**, 1267–1284 (2018).

25. Burnet, M. Cancer--A Biological Approach: I. The Processes Of Control. II. The Significance of Somatic Mutation. *BMJ* **1**, 779–786 (1957).

26. Kim, R., Emi, M. & Tanabe, K. Cancer immunoediting from immune surveillance to immune escape. *Immunology* **121**, 1–14 (2007).

27. Fridman, W. H. From Cancer Immune Surveillance to Cancer Immunoediting: Birth of Modern Immuno-Oncology. *J. Immunol.* **201**, 825–826 (2018).

28. Mittal, D., Gubin, M. M., Schreiber, R. D. & Smyth, M. J. New insights into cancer immunoediting and its three component phases—elimination, equilibrium and escape. *Curr. Opin. Immunol.* **27**, 16–25 (2014).

29. O'Donnell, J. S., Teng, M. W. L. & Smyth, M. J. Cancer immunoediting and resistance to T cell-based immunotherapy. *Nat. Rev. Clin. Oncol.* **16**, 151–167 (2019).

30. De Bock, K., Cauwenberghs, S. & Carmeliet, P. Vessel abnormalization: another hallmark of cancer?Molecular mechanisms and therapeutic implications. *Curr. Opin. Genet. Dev.* **21**, 73–79 (2011).

31. Scioli, M. G. *et al.* Adipose-Derived Stem Cells in Cancer Progression: New Perspectives and Opportunities. *Int. J. Mol. Sci.* **20**, 3296 (2019).

32. Gunawardene, A. R., Corfe, B. M. & Staton, C. A. Classification and functions of enteroendocrine cells of the lower gastrointestinal tract. *Int. J. Exp. Pathol.* **92**, 219–231 (2011).

33. Jin, R. J. *et al.* NE-10 Neuroendocrine Cancer Promotes the LNCaP Xenograft Growth in Castrated Mice. *Cancer Res.* **64**, 5489–5495 (2004).

34. Walker, C., Mojares, E. & del Río Hernández, A. Role of Extracellular Matrix in Development and Cancer Progression. *Int. J. Mol. Sci.* **19**, 3028 (2018).

35. Eble, J. A. & Niland, S. The extracellular matrix in tumor progression and metastasis. *Clin. Exp. Metastasis* **36**, 171–198 (2019).

36. Levental, K. R. *et al.* Matrix Crosslinking Forces Tumor Progression by Enhancing Integrin Signaling. *Cell* **139**, 891–906 (2009).

37. Karagiannis, G. S. *et al.* Cancer-Associated Fibroblasts Drive the Progression of Metastasis through both Paracrine and Mechanical Pressure on Cancer Tissue. *Mol. Cancer Res.* **10**, 1403–1418 (2012).

38. Özdemir, B. C. *et al.* Depletion of Carcinoma-Associated Fibroblasts and Fibrosis

Induces Immunosuppression and Accelerates Pancreas Cancer with Reduced Survival. *Cancer Cell* **25**, 719–734 (2014).

39. Arnold, S. A. *et al.* Lack of host SPARC enhances vascular function and tumor spread in an orthotopic murine model of pancreatic carcinoma. *Dis. Model. Mech.* **3**, 57–72 (2010).

40. Zakrzewski, W., Dobrzyński, M., Szymonowicz, M. & Rybak, Z. Stem cells: past, present, and future. *Stem Cell Res. Ther.* **10**, 68 (2019).

41. Kuşoğlu, A. & Biray Avcı, Ç. Cancer stem cells: A brief review of the current status. *Gene* **681**, 80–85 (2019).

42. Najafi, M., Farhood, B. & Mortezaee, K. Cancer stem cells (CSCs) in cancer progression and therapy. *J. Cell. Physiol.* **234**, 8381–8395 (2019).

43. Plaks, V., Kong, N. & Werb, Z. The Cancer Stem Cell Niche: How Essential Is the Niche in Regulating Stemness of Tumor Cells? *Cell Stem Cell* **16**, 225–238 (2015).

44. Wang, W. *et al.* Dynamics between Cancer Cell Subpopulations Reveals a Model Coordinating with Both Hierarchical and Stochastic Concepts. *PLoS One* **9**, e84654 (2014).

45. Rich, J. N. Cancer stem cells. *Medicine (Baltimore).* **95**, S2–S7 (2016).

46. Vogelstein, B. *et al.* Genetic Alterations during Colorectal-Tumor Development. *N. Engl. J. Med.* **319**, 525–532 (1988).

47. Roth, G. A. *et al.* Global, Regional, and National Burden of Cardiovascular Diseases for 10 Causes, 1990 to 2015. *J. Am. Coll. Cardiol.* **70**, 1–25 (2017).

48. Roth, G. A. *et al.* Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1736–1788 (2018).

49. Sharma, S., Kelly, T. K. & Jones, P. A. Epigenetics in cancer. *Carcinogenesis* **31**, 27–36 (2010).

50. Venters, B. J. & Pugh, B. F. How eukaryotic genes are transcribed. *Crit. Rev. Biochem. Mol. Biol.* **44**, 117–141 (2009).

51. van Steensel, B. & Furlong, E. E. M. The role of transcription in shaping the spatial

organization of the genome. *Nat. Rev. Mol. Cell Biol.* (2019) doi:10.1038/s41580-019-0114-6.

52. Wade, J. T. & Struhl, K. The transition from transcriptional initiation to elongation. *Curr. Opin. Genet. Dev.* **18**, 130–136 (2008).

53. Chen, Z., Li, S., Subramaniam, S., Shyy, J. Y.-J. & Chien, S. Epigenetic Regulation: A New Frontier for Biomedical Engineers. *Annu. Rev. Biomed. Eng.* **19**, 195–219 (2017).

54. Mariño-Ramírez, L., Kann, M. G., Shoemaker, B. A. & Landsman, D. Histone structure and nucleosome stability. *Expert Rev. Proteomics* **2**, 719–729 (2005).

55. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21**, 381–395 (2011).

56. Wei, J.-W., Huang, K., Yang, C. & Kang, C.-S. Non-coding RNAs as regulators in epigenetics. *Oncol. Rep.* **37**, 3–9 (2017).

57. O'Brien, J., Hayder, H., Zayed, Y. & Peng, C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. *Front. Endocrinol. (Lausanne).* **9**, (2018).

58. Piatek, M. J. & Werner, A. Endogenous siRNAs: regulators of internal affairs. *Biochem. Soc. Trans.* **42**, 1174–1179 (2014).

59. Chen, L., Dahlstrom, J. E., Lee, S.-H. & Rangasamy, D. Naturally occurring endo-siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. *Epigenetics* **7**, 758–771 (2012).

60. Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D. & Zamore, P. D. PIWI-interacting RNAs: small RNAs with big functions. *Nat. Rev. Genet.* **20**, 89–108 (2019).

61. Liang, J. *et al.* Small Nucleolar RNAs: Insight Into Their Function in Cancer. *Front. Oncol.* **9**, (2019).

62. Bhat, S. A. *et al.* Long non-coding RNAs: Mechanism of action and functional utility. *Non-coding RNA Res.* **1**, 43–50 (2016).

63. Mazzio, E. A. & Soliman, K. F. A. Basic concepts of epigenetics. *Epigenetics* **7**, 119–130 (2012).

64. Kim, J. K., Samaranayake, M. & Pradhan, S. Epigenetic mechanisms in mammals. *Cell. Mol. Life Sci.* **66**, 596–612 (2009).

65. Jin, Z. & Liu, Y. DNA methylation in human diseases. *Genes Dis.* **5**, 1–8 (2018).

66. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).

67. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).

68. Hernando-Herraez, I., Garcia-Perez, R., Sharp, A. J. & Marques-Bonet, T. DNA Methylation: Insights into Human Evolution. *PLOS Genet.* **11**, e1005661 (2015).

69. Chatterjee, A., Rodger, E. J. & Eccles, M. R. Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin. Cancer Biol.* **51**, 149–159 (2018).

70. Jin, B., Li, Y. & Robertson, K. D. DNA Methylation: Superior or Subordinate in the Epigenetic Hierarchy? *Genes Cancer* **2**, 607–617 (2011).

71. Viré, E. *et al.* The Polycomb group protein EZH2 directly controls DNA methylation. *Nature* **439**, 871–874 (2006).

72. Tachibana, M., Matsumura, Y., Fukuda, M., Kimura, H. & Shinkai, Y. G9a/GLP complexes independently mediate H3K9 and DNA methylation to silence transcription. *EMBO J.* **27**, 2681–2690 (2008).

73. Ikegami, K. *et al.* Genome-wide and locus-specific DNA hypomethylation in G9a deficient mouse embryonic stem cells. *Genes to Cells* **12**, 1–11 (2007).

74. Zhao, Q. *et al.* PRMT5-mediated methylation of histone H4R3 recruits DNMT3A, coupling histone and DNA methylation in gene silencing. *Nat. Struct. Mol. Biol.* **16**, 304–311 (2009).

75. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* **20**, 207–220 (2019).

76. Morgan, M. A. & Shilatifard, A. Chromatin signatures of cancer. *Genes Dev.* **29**, 238–249 (2015).

77. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science (80-.).* **357**, eaal2380 (2017).

78. Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell* **173**, 283–285 (2018).

79. Dedeurwaerder, S. *et al.* A comprehensive overview of Infinium HumanMethylation450 data processing. *Brief. Bioinform.* **15**, 929–941 (2014).

80. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).

81. Hrdlickova, R., Toloue, M. & Tian, B. RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA* **8**, e1364 (2017).

82. Silva, T. C. *et al.* TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research* **5**, 1542 (2016).

83. Kwak, S. K. & Kim, J. H. Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.* **70**, 407 (2017).

84. Mallik, S. *et al.* An evaluation of supervised methods for identifying differentially methylated regions in Illumina methylation arrays. *Brief. Bioinform.* **20**, 2224–2235 (2019).

85. Jaffe, A. E. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *Int. J. Epidemiol.* **41**, 200–209 (2012).

86. Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* **8**, 6 (2015).

87. Pirim, H., Ekşioğlu, B., Perkins, A. D. & Yüceer, Ç. Clustering of high throughput gene expression data. *Comput. Oper. Res.* **39**, 3046–3061 (2012).

88. Ghasemi, A. & Zahediasl, S. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *Int. J. Endocrinol. Metab.* **10**, 486–489 (2012).

89. Dumitrascu, B., Darnell, G., Ayroles, J. & Engelhardt, B. E. Statistical tests for detecting variance effects in quantitative trait studies. *Bioinformatics* **35**, 200–210 (2019).

90. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68**, 540 (2015).

91. Nahm, F. S. Nonparametric statistical tests for the continuous data: the basic concept and the practical use. *Korean J. Anesthesiol.* **69**, 8 (2016).

92. Stephens, M. False discovery rates: a new deal. *Biostatistics* kxw041 (2016) doi:10.1093/biostatistics/kxw041.

93. Obuchowski, N. A. & Bullen, J. A. Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Phys. Med. Biol.* **63**, 07TR01 (2018).

94. Neely, J. G. *et al.* Practical Guide to Understanding Multivariable Analyses. *Otolaryngol. Neck Surg.* **148**, 185–190 (2013).

95. Neely, J. G. *et al.* Practical Guide to Understanding Multivariable Analyses, Part B. *Otolaryngol. Neck Surg.* **148**, 359–365 (2013).

96. Slinker, B. K. & Glantz, S. A. Multiple linear regression is a useful alternative to traditional analyses of variance. *Am. J. Physiol. Integr. Comp. Physiol.* **255**, R353–R367 (1988).

97. Piazza, R. *et al.* OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. *Sci. Rep.* **7**, 46290 (2017).

98. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161 (2009).

99. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).

100. Lee, S. & Lim, H. Review of statistical methods for survival analysis using genomic data. *Genomics Inform.* **17**, e41 (2019).

101. Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B* **34**, 187–202 (1972).

102. Xue, X. *et al.* Testing the proportional hazards assumption in case-cohort analysis. *BMC Med. Res. Methodol.* **13**, 88 (2013).

103. Lausen, B. & Schumacher, M. Maximally Selected Rank Statistics. *Biometrics* **48**, 73 (1992).

104. Lucijanic, M., Skelin, M. & Lucijanic, T. Survival analysis, more than meets the eye. *Biochem. Medica* 14–18 (2017) doi:10.11613/BM.2017.002.

105. Zhang, D. & Quan, H. Power and sample size calculation for log-rank test with a time lag in treatment effect. *Stat. Med.* **28**, 864–879 (2009).

106. Li, H., Han, D., Hou, Y., Chen, H. & Chen, Z. Statistical Inference Methods for Two Crossing Survival Curves: A Comparison of Methods. *PLoS One* **10**, e0116774 (2015).

107. Qiu, P. & Sheng, J. A two-stage procedure for comparing hazard rate functions. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **70**, 071103032514002-??? (2007).

108. Moarii, M., Boeva, V., Vert, J.-P. & Reyal, F. Changes in correlation between promoter methylation and gene expression in cancer. *BMC Genomics* **16**, 873 (2015).

109. Esteller, M. CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future. *Oncogene* **21**, 5427–5440 (2002).

110. Jung, G., Hernández-Illán, E., Moreira, L., Balaguer, F. & Goel, A. Epigenetics of colorectal cancer: biomarker and therapeutic potential. *Nat. Rev. Gastroenterol. Hepatol.* **17**, 111–130 (2020).

111. Schatoff, E. M., Leach, B. I. & Dow, L. E. WNT Signaling and Colorectal Cancer. *Curr. Colorectal Cancer Rep.* **13**, 101–110 (2017).

112. Anderson, K. J., Cormier, R. T. & Scott, P. M. Role of ion channels in gastrointestinal cancer. *World J. Gastroenterol.* **25**, 5732–5772 (2019).

113. The benefits and harms of breast cancer screening: an independent review. *Lancet* **380**, 1778–1786 (2012).

114. Guan, Z., Yu, H., Cuk, K., Zhang, Y. & Brenner, H. Whole-Blood DNA Methylation Markers in Early Detection of Breast Cancer: A Systematic Literature Review. *Cancer Epidemiol. Biomarkers Prev.* **28**, 496–505 (2019).

115. Tang, Q., Cheng, J., Cao, X., Surowy, H. & Burwinkel, B. Blood-based DNA methylation as biomarker for breast cancer: a systematic review. *Clin. Epigenetics* **8**, 115 (2016).

116. Nicolini, A., Ferrari, P. & Duffy, M. J. Prognostic and predictive biomarkers in breast cancer: Past, present and future. *Semin. Cancer Biol.* **52**, 56–73 (2018).

117. Khan, S. A., Tavolari, S. & Brandi, G. Cholangiocarcinoma: Epidemiology and risk factors. *Liver Int.* **39**, 19–31 (2019).

118. Cners, R., Shomer, R. W. & Rodrigues, A. Aberrant DNA Methylation as a Biomarker and a Therapeutic Target of Cholangiocarcinoma. *Int. J. Mol. Sci.* **18**, 1111 (2017).

119. Blechacz, B., Komuta, M., Roskams, T. & Gores, G. J. Clinical diagnosis and staging of cholangiocarcinoma. *Nat. Rev. Gastroenterol. Hepatol.* **8**, 512–522 (2011).

120. Kailasam, A., Mittal, S. K. & Agrawal, D. K. Epigenetics in the Pathogenesis of Esophageal Adenocarcinoma. *Clin. Transl. Sci.* **8**, 394–402 (2015).

121. Lo Nigro, C., Denaro, N., Merlotti, A. & Merlano, M. Head and neck cancer: improving outcomes with a multidisciplinary approach. *Cancer Manag. Res.* **Volume 9**, 363–371 (2017).

122. Alsahafi, E. *et al.* Clinical update on head and neck cancer: molecular biology and ongoing challenges. *Cell Death Dis.* **10**, 540 (2019).

123. Plzák, J. *et al.* The Head and Neck Squamous Cell Carcinoma Microenvironment as a Potential Target for Cancer Therapy. *Cancers (Basel).* **11**, 440 (2019).

124. Toh, T. B., Lim, J. J. & Chow, E. K.-H. Epigenetics of hepatocellular carcinoma. *Clin. Transl. Med.* **8**, 13 (2019).

125. Kao, W.-Y. *et al.* Prognosis of Early-Stage Hepatocellular Carcinoma. *Medicine (Baltimore).* **94**, e1929 (2015).

126. Gandara, D. R., Hammerman, P. S., Sos, M. L., Lara, P. N. & Hirsch, F. R. Squamous Cell Lung Cancer: From Tumor Genomics to Cancer Therapeutics. *Clin. Cancer Res.* **21**, 2236–2243 (2015).

127. Langevin, S. M., Kratzke, R. A. & Kelsey, K. T. Epigenetics of lung cancer. *Transl. Res.* **165**, 74–90 (2015).

128. Shah, J. P. Thyroid carcinoma: epidemiology, histology, and diagnosis. *Clin. Adv. Hematol. Oncol.* **13**, 3–6 (2015).

129. Nabi, S., Kessler, E. R., Bernard, B., Flaig, T. W. & Lam, E. T. Renal cell carcinoma: a review of biology and pathophysiology. *F1000Research* **7**, 307 (2018).

130. Hsieh, J. J. *et al.* Renal cell carcinoma. *Nat. Rev. Dis. Prim.* **3**, 17009 (2017).

131. Rhoades Smith, K. E. & Bilen, M. A. A Review of Papillary Renal Cell Carcinoma and MET Inhibitors. *Kidney Cancer* **3**, 151–161 (2019).

132. Sanchez, D. J. & Simon, M. C. Genetic and metabolic hallmarks of clear cell renal cell carcinoma. *Biochim. Biophys. Acta - Rev. Cancer* **1870**, 23–31 (2018).

133. McGuigan, A. *et al.* Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *World J. Gastroenterol.* **24**, 4846–4861 (2018).

134. Ansari, J., Shackelford, R. E. & El-Osta, H. Epigenetics in non-small cell lung cancer: from basics to therapeutics. *Transl. Lung Cancer Res.* **5**, 155–171 (2016).

135. Shi, Y.-X., Sheng, D.-Q., Cheng, L. & Song, X.-Y. Current Landscape of Epigenetics in Lung Cancer: Focus on the Mechanism and Application. *J. Oncol.* **2019**, 1–11 (2019).

136. Leng, Z. G. *et al.* Activation of DRD5 (dopamine receptor D5) inhibits tumor growth by autophagic cell death. *Autophagy* **13**, 1404–1419 (2017).

137. Han, X., Yang, J., Li, D. & Guo, Z. Overexpression of Uric Acid Transporter SLC2A9 Inhibits Proliferation of Hepatocellular Carcinoma Cells. *Oncol. Res. Featur. Preclin. Clin. Cancer Ther.* **27**, 533–540 (2019).

138. Okamoto, J. *et al.* EMX2 is epigenetically silenced and suppresses growth in human lung cancer. *Oncogene* **29**, 5969–5975 (2010).

139. Aykut, B. *et al.* EMX2 gene expression predicts liver metastasis and survival in colorectal cancer. *BMC Cancer* **17**, 555 (2017).

140. Hori, A. *et al.* The conserved Wdr8-hMsd1/SSX2IP complex localises to the centrosome and ensures proper spindle length and orientation. *Biochem. Biophys. Res. Commun.* **468**, 39–45 (2015).