André Miguel Romeira Mestre

Whole-genome analysis of DNA methylation across cancer types reveals specific patterns in early stages

Universidade do Algarve

Departamento de Ciências Biomédicas e Medicina

2018

André Miguel Romeira Mestre

Whole-genome analysis of DNA methylation across cancer types reveals specific patterns in early stages

Master in Oncobiology – Molecular Mechanisms of Cancer

This work was done under the supervision of:

Ana Marreiros, PhD (Supervisor)

Pedro Castelo-Branco, PhD (Co-Supervisor)

Universidade do Algarve

Departamento de Ciências Biomédicas e Medicina

2018

Whole-genome analysis of DNA methylation across cancer types reveals specific patterns in early stages

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam na 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“*I declare that I am the author of this work that is original and unpublished.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.*”

____

(André Miguel Romeira Mestre)

“Viver não é necessário. Necessário é criar.”

Fernando Pessoa

Agradecimentos

A finalização desta dissertação marca o fim de mais um grau académico, um percurso de aprendizagem e de desenvolvimento pessoal em que em cada dia resultou num pouco mais de conhecimento e compreensão da vida científica. Um bem-haja a todas as pessoas que permitiram que este objetivo pessoal se concretizasse, especialmente aquelas que demostraram amizade e companheirismo.

Em especial, quero aqui deixar um sentido obrigado à Professora Doutora Ana Marreiros pela disponibilidade, apoio na concretização de toda a dissertação e amizade que me permitiu chegar até aqui. Muito obrigado.

Agradeço ao Professor Doutor Pedro Castelo-Branco por acreditar em mim, por tudo o que me ensinou e por todo o apoio que me deu que me permitiu chegar até aqui. Um sincero muito obrigado.

Aos meus colegas Sara Ramalhete e André Fonseca, um muito obrigado por todo o apoio, motivação e amizade que demonstraram ao longo deste último ano.

A todos os restantes membros desta equipa, em especial a Joana Apolónio, deixo aqui um sincero agradecimento pela disponibilidade, amizade e todos os momentos de convívio.

Agradeço a todos os meus amigos pelas conversas, apoio e compreensão ao longo deste ano que passou cheio de momento bons e outros menos bons que foram atenuados pela vossa presença.

O meu último, e mais importante, agradecimento é para toda a minha família, em particular pais e irmão, pelo carinho e confiança, pelo constante sacrifício que fizeram para que eu pudesse chegar até aqui, e por fazerem de mim a pessoa que sou hoje.

Abstract

Dynamic variations in DNA methylation are known to play an important role in cancer development through modulation of gene expression. Here, were developed a mathematical structured model to identify patterns of differentially methylated genes (cDMGs), across different cancers types that can act as epigenetic diagnostic biomarkers.

A Working Pipeline (WP), designed in R language, was applied to 8 cancer cohorts from The Cancer Genome Atlas (TCGA) aiming to analyze DNA methylation and gene expression alterations occurring during normal to stage I carcinogenic transition.

WP has a principal component which was divided in four steps: 0. Clinical characterization of patients; 1. Identification of cDMGs; 2. Identification of genetic/epigenetic patterns across different cancer type; and 3. Identification of diagnostic predictors. Additionally, the WP had a second component containing two more complementary steps: 4. Identification of CpG probes that better predict gene expression and 5. HJ-Biplot approach to visualize genes or CpG probes and its association with sample distribution. Appling the principal component of the WP to TCGA cohorts, we identified 117 cDMGs in breast cancer, 307 in colorectal cancer, 99 in head and neck cancer, 156 in kidney clear cell cancer, 106 in kidney papillary cancer, 349 in liver cancer, 180 in lung cancer and 25 in thyroid cancer. Analysis of patterns across these cancers revealed that the majority of cDMGs are cancer-specific. Moreover, we found cDMGs to be good predictors of diagnosis. When considering specific biomarkers for each cancer, only 19, 153, 27, 93, 53, 72, 38 and 14 genes were found to be good diagnostic biomarkers in breast, colorectal, head and neck, kidney$_R$, kidney$_P$, liver, lung and thyroid cancers, respectively.

Therefore, we developed a novel working pipeline that allowed data sets analyses available worldwide. Validation of this mathematical model evidences that normal-tumor transition is not a conserved process event across different cancers type, but specific to the cell of origin.

Keywords: cancer, DNA methylation, gene expression, diagnosis biomarker and computational analysis.

Resumo

O cancro é descrito como um grupo de doenças altamente complexas caracterizadas pelo crescimento anormal e descontrolado de células com a capacidade de invadir outros tecidos. A vasta maioria das células presentes no organismo adulto apresentam o genoma completo, altamente regulado, de forma, a manter os padrões de atividade específica para cada tecido. Assim, os mecanismos que regulam esta atividade são importantes objetos de estudo no desenvolvimento de cancro, nomeadamente, a metilação do DNA.

A metilação do DNA é um dos mecanismos epigenéticos mais estudados que ocorre pela adição de um grupo metil à sequência de DNA, modificando a função dos genes e influenciando a expressão genética.

O cancro é maior causa de morbilidade e mortalidade no mundo, contando com 18.1 milhões de novos casos e 9.6 milhões de mortes. Salienta-se, que os cancros do pulmão, mama e colorretal apresentam a maior taxa de incidência.

A presente dissertação teve como principais objetivos 1) criar um procedimento de trabalho, 2) identificar genes diferencialmente metilados associados a cancro (cDMGs), 3) identificar padrões de expressão/metilação entre diferentes tipos cancros e 4) identificar preditores de diagnóstico.

Metodologicamente, foi criado um procedimento de trabalho que teve aplicação na análise do genoma completo das coortes do *The Cancer Genome Atlas* (TCGA). A análise enunciada utilizou dados de expressão genética (*Illumina Hiseq*) e metilação de DNA (*Illumina HumanMethylation 450K array*) para 8 coortes dos seguintes tipos de cancro: cancro da mama, cancro colorretal, cancro da cabeça e pescoço, cancro das células renais (cancro do rim$_R$), cancro papilar do rim (cancro do rim$_P$), cancro do fígado, cancro do pulmão e cancro da tiroide. Neste projeto, foram comparados dois grupos, tecido sólido adjacente e tumor primário em estadio I com 84 e 126 em cancro da mama, 21 e 54 em cancro colorretal, 20 e 27 em cancro da cabeça e pescoço, 24 e 155 em cancro do rim$_R$, 23 e 167 em cancro do rim$_P$, 41 e 171 em cancro do fígado, 21 e 245 em cancro do pulmão e 50 e 284 em cancro da tiroide, respetivamente. Os dados mencionados foram analisados através de linguagem de programação em R.

Considerando os objetivos propostos, verificou-se que o primeiro objetivo é a chave para os restantes. O procedimento de trabalho foi estruturado com base em duas componentes distintas. A componente principal apresentou 4 fases: Fase 0 – Caracterização dos cohorts; Fase

1 – Identificar genes diferencialmente metilados associados a cancro; Fase 2 – Identificar padrões genéticos/epigenéticos entre diferentes tipos de cancro e Fase 3 – Identificar preditores de diagnóstico. Entretanto, a componente complementar apresentou 2 fases: Fase 4 – Identificar sítios de metilação com maior impacto na expressão e Fase 5 – Representação multivariada utilizando HJ-Biplot para visualizar genes ou sítios de metilação e a sua associação com a distribuição das amostras.

Dentro da componente principal, a Fase 0 foi considerada opcional e teve como intuito caracterizar os pacientes da coorte utilizando as variáveis clínicas disponíveis para tal. As fases seguintes estiveram dependentes da existência de dados de expressão genética (*Illumina HiSeq*) e metilação de DNA (*Illumina HumanMethylation 450K array*), assim como, pacientes que apresentem ambas as amostras. Deste modo, ambas as bases de dados foram importadas no início da Fase 1, os genes e sítios de metilação foram sujeitos a um pré-processamento, seguido de um processo de testes inferenciais distribuídos por níveis. Após seleção de genes com diferenças significativas de expressão e sítios de metilação com diferenças significativas de metilação estabeleceu-se os pontos de corte (valor absoluto $\log_2$(Foldchange)>1.5 e valor absoluto $\Delta\beta$>0.2). Assim, foram selecionados apenas genes e CpG com diferenças muito significativas com interesse de estudo. Posteriormente, o teste de correlação de Pearson avaliou a relação entre ambos e identificou os genes diferencialmente metilados associados a cancro. A Fase 2 procurou identificar padrões através da interseção das várias coortes. Por fim, a Fase 3 identificou os bons preditores de diagnóstico. De forma a complementar a análise, a Fase 4 utilizou os modelos lineares de regressão múltipla para identificar a metilação de sítios de metilação com maior impacto na expressão de gene. Entretanto, a Fase 5 procurou de forma multivariada identificar comportamentos de gene ou sítios de metilação com maior influência na distinção entre grupos e na distribuição das amostras.

Através do procedimento de trabalho estabelecido, foram identificados nas coortes mama, colorretal, cabeça e pescoço, rim$_R$, rim$_P$, fígado, pulmão e tiroide, diferenças na expressão de 117, 307, 99, 156, 106, 349, 180 e 25 genes (valor absoluto de $\log_2$(Foldchange) > 1,5 e p-value ajustado (FDR)<0.05) e diferencialmente metiladas 368, 924, 292, 299, 224, 1453, 601 e 40 sítios de metilação (valor absoluto de $\Delta\beta$>0,2 e p-value ajustado (FDR)<0.05), respetivamente, designados de cDMGs. Seguidamente, foi realizada uma análise de processo biológico que revelou a existência de enriquecimento de funções ligadas ao desenvolvimento e sistema nervoso. Entretanto, foi realizada uma análise anotação com objetivo de verificar quais

dos cDMGs nunca foram reportados em cancro. Esta análise sugere que nas coortes acima mencionadas 18, 36, 13, 18, 15, 48, 20 e 3 genes, respetivamente, nunca foram mencionados com cancro. Por outro lado, 62, 150, 28, 27, 20, 94, 100 e 6 genes, respetivamente, já foram mencionados no cancro específico. Entretanto, os restantes já foram mencionados em cancro, mas não no cancro específico.

De seguida, a intersecção dos genes ou sítios de metilação entre coortes mostrou que a maioria eram específicos para o tipo tumoral e apenas uma pequena quantidade deles tinham presença em mais de uma coorte. Assim, para as coortes da mama, colorretal, cabeça e pescoço, rim_{R}, rim_{P}, fígado, pulmão e tiroide, são específicos para a coorte 55, 202, 49, 100, 70, 240, 97 e 18 genes, respetivamente, e 261, 782, 223, 244, 189, 1339, 449 e 35 sítios de metilação, respetivamente. Seguidamente, foi realizada uma análise de vias de sinalização utilizando a base de dados Reactome que mostrou a cascata RAF/MAP quinase (*p-value=8.01e-05*) está muito presente em cancro colorretal, assim como, as interações L1CAM (p-value=0.004208). Adicionalmente, a ativação do recetor GABA A (*p-value=0.026896*) está enriquecido em cancro da cabeça e pescoço, os recetores péptido-ligando (*p-value=0.006942*) e a metilação de DNA (*p-value=0.024459*) em cancro do pulmão. Finalmente, os nossos resultados sugerem que o desenvolvimento de cancro em estadios precoces apresenta características intrínsecas ao tecido de origem.

Por último, a análise de bons preditores de diagnóstico teve como objetivo identificar biomarcadores com capacidade de discriminar tecido normal e tumoral em estadios precoces. Os nossos resultados mostraram que nas coortes previamente mencionados existiram 45, 238, 57, 142, 88, 126, 88 e 18 genes, respetivamente, juntamente com 340, 835, 286, 299, 200, 1129, 595 e 38 sítios de metilação, respetivamente. Destes, 44, 153, 68, 173, 111, 261, 128 e 24, respetivamente pertenceram aos padrões específicos encontrados.

Concluindo, nós criamos um procedimento de trabalho capaz de analisar bases de dados de todo o mundo. Como vimos, este estudo mostrou que o procedimento permitiu identificar diferenças de metilação significativas em estadios precoces. Estas alterações na sua grande maioria são específicas da transição normal-tumoral evidenciando que este evento não é conservado entre tipos de cancro, sugerindo que cada tecido apresenta características únicas do tipo de célula de origem.

Palavras-chave: cancro, metilação de DNA, expressão genética, biomarcador de diagnóstico e análise computacional.

INDEX OF CONTENTS

INDEX OF FIGURES

INDEX OF TABLES

INDEX OF ANNEXES

LIST OF ABBREVIATIONS

5-mC – 5-methylcytosine

BRCA – Breast Invasive Carcinoma

C – Cytosine

COAD – Colon Adenocarcinoma

CpG – Cytosine-Guanine pair

CRC – Colorectal Cancer

DNA – Deoxyribonucleic acid

DNMT – DNA methyltransferase

DMG – Differential methylated genes

FDR – False Discover Rate

G – Guanine

GO – Gene Ontology

GDC – Genomic Data Commons

HAT – Histone acetyltransferase

HDAC – Histone deacetylase

HDM – Histone demethylase

HMT – Histone methyltransferase

HNSC – Head and Neck Squamous Cell Carcinoma

LIHC – Liver Hepatocellular Carcinoma

LUAD – Lung Adenocarcinoma

MBD – Methyl-binding proteins

MLR – Multiple Linear Regressions

MIRNA – MicroRNA

NAT – Normal adjacent tissue

NCRNA – Non-coding RNA

KIDNEY$_R$ – Kidney Renal Clear Cell Carcinoma

KIDNEY$_P$ – Kidney Renal Papillary Cell Carcinoma

KIRC – Kidney Renal Clear Cell Carcinoma

KIRP – Kidney Renal Papillary Cell Carcinoma

TCGA – The Cancer Genome Atlas

TF – Transcription factor

TFBS – Transcription factors biding sites

THCA – Thyroid Carcinoma

READ – Rectum Adenocarcinoma

RNA – Ribonucleic acid

RNAi – Interfering ribonucleic acid

SIRNA – Small Interfering RNA

WP – Working pipeline

1 CHAPTER 1 – INTRODUCTION

1.1. Cancer

Organisms present highly differentiated and specialized tissues that allow them to perform their functions autonomously. Most cells of an adult organism have the complete genome, meaning that they present more information than is necessary for its functioning. Proliferative capacity is an intrinsic feature to cells, allowing the tissue to maintain its functions and characteristics. This constant maintenance involves the repair of damage as well as cell replacement when possible (Weinberg 2014). This key rule of regulation is the crucial feature for cancer development.

When normal and functional tissues lose their original characteristics for which they have been programmed, they become dangerous to the body. The altered cells have now access to genome information which they would not normally have. Hence, changes in the genome that are promoted by this cellular and functional instability allow the cells to acquire new abnormal phenotypes. Tissues with abnormal cells compromise their function and proliferate wildly forming an abnormal cell mass, and then the tumor (Weinberg 2014).

Cancer can be considered a set of highly complex diseases characterized by abnormal and uncontrolled cell growth that can proliferate and invade other tissues (Hanahan et al. 2000). It is characterized by extensive amounts of genetic mutations and chromosomal abnormalities. Recently, aberrant epigenetic modifications have been highlighted in cancer and, together with genetic alterations, they have been useful for understanding the complexity observed in neoplasms. Therefore, the cancer epigenome has contributed greatly to the understanding of the complexity and diversity of different types of cancer. Nevertheless, the characterization of the epigenetic events during the tumorigenesis remains unclear.

1.1.1. Epidemiology

Cancer can be considered the disease of the century and it is a major cause of morbidity and mortality worldwide, accounting for about 18.1 million new cases and 9.6 million cancer-related deaths, in 2018 (Bray et al. 2018). Develop countries with more resources a higher incidence of cancer. However, since they provide better healthcare services in terms of screening, diagnosis and treatment, the mortality rates are lower in the developed countries than in less developed ones (Bray et al. 2018). Among top seven of cancers with the highest incidence worldwide are lung (11.6%), breast (11.6%), colorectal (10.2%), prostate (7.1%), stomach (5.7%), liver (4.7%) and Oesophagus (3.2%), represented in ***Figure 1.1A***. Then, the top seven of most deadly cancers worldwide are lung (18.4%), colorectal (9.2%), liver (8.2%), stomach (8.2%), breast (6.6%), oesophagus (5.3%) and pancreas (4.5%), represented in ***Figure 1.1B***. Worldwide, lung cancer is the cancer with the highest incidence and mortality (Bray et al. 2018).

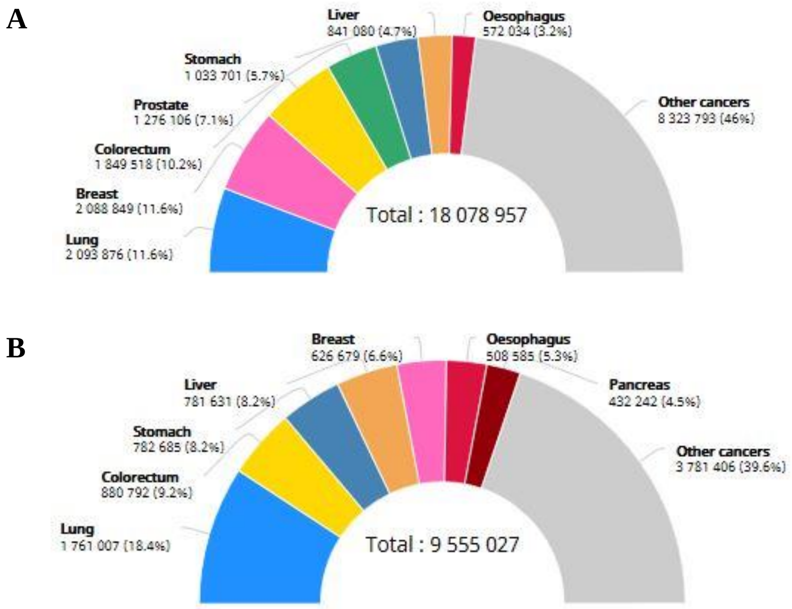


Figure 1.1 – Worldwide cancer estimated incidence and mortality rates for both sexes of 2018. Pie charts illustrated the different cancer estimated incidence **(A)** and mortality **(B)** rates in the worldwide population for the year 2018. From GLOBOCAN 2018 (IARC).

Although, Europe contains 9% of the world's population, it has 25% of the global cancer rate. Thus, the constant updating of European statistics becomes a strong ally in cancer planning. Moreover, in Europe, it is estimated 3.91 million new cases and 1.93 million deaths due to cancer in 2018 (Ferlay et al. 2018).

In 2018, in Portugal, from the total of 10 291 198 people, 58 199 people were diagnosed and 28 960 died with cancer (Anon n.d.). When we look at the incidence (***Figure 1.2A***), the seven most common cancers in both sexes are colorectal (17.6%), prostate (11.4%), breast (12%), lung (9.1%), stomach (5%), bladder (4%) and non-Hodgkin lymphomas (3.6%). When we look at mortality (***Figure 1.2B***), the seven most deadly cancers are colorectal (14.7%), lung (16.1%), stomach (7.9%), breast (6%) and prostate (6.5%), pancreas (6.5%) and liver (4.7%). In Portugal, colorectal cancer is the cancer with the highest incidence and mortality (Anon n.d.).

The exponential increase of cancer cases is due to population growth and aging, and the number of cases is estimated to increase up to more than 20 million per year in 2030 (Stewart, BWKP and Wild 2014).

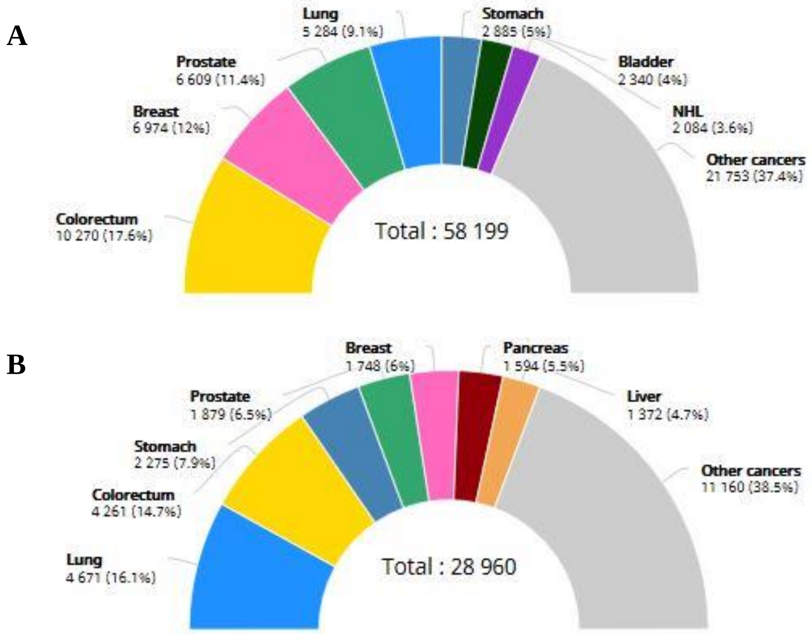


Figure 1.2 – Portugal cancer estimated incidence and mortality rates for both sexes of 2018. Pie charts illustrated the different cancer estimated incidence **(A)** and mortality **(B)** rates in the Portuguese population for the year 2018. From GLOBOCAN 2018 (IARC).

1.1.2. Hallmarks of cancer

Tumorigenesis results from the transformation of normal cells into transformed cells which have lost their original characteristics (Hanahan et al. 2000). This transformation is a multistep process characterized by five main histological states: hypertrophy, hyperplasia, metaplasia, dysplasia and neoplasia. First, hypertrophy is characterized essentially by the increase of cell size without increasing in number. Second, hyperplasia is characterized by an increased number of cells. Third, metaplasia contemplates cells from different lineages due to be a transition phase. Fourth, dysplasia reveals changes in the function and shape of cells due to tissue disorganization. Finally, neoplasia is characterized by uncontrolled growth and loss of function associated to invasion of adjacent tissues and metastasis. The cell mass formed is designated as tumor, however not all tumors reach a stage of neoplasia, termed cancer (Weinberg 2014).

In 2000, to characterize this transformation, Douglas Hanahan and Robert Weinberg proposed six key capabilities that cells acquire during the multistep process of tumor development which are necessary for the development of cancer (***Figure 1.3A***). These hallmarks include (Hanahan et al. 2000): sustaining proliferative signaling; evading growth suppressors; activating invasion and metastasis; enabling replicative immortality; inducing angiogenesis; and resisting cell death.

Although these hallmarks are attributed to cancer, there are *five characteristics* (all except invasion and metastasis) common to benign tumors (Lazebnik 2010). Importantly, we must consider that they were established aiming to create lines of investigation to combat the mechanisms underlying the ability of cancers to kill, as such these characteristics are called hallmarks of cancer and not hallmarks of tumor. The benign and malignant tumors are not highly related, since both types can be:

- Developed in the same organ;
- Arises from the same cellular type;
- Presented the same size;
- And have the same external influence and occur spontaneously.

Therefore, this arises the need to know the mechanisms that are behind tumor malignancy (invasion and metastasis- Lazebnik 2010).

In 2011, Hanahan and Weinberg re-evaluated the hallmarks to improve the characterization of cancer (Hanahan & Weinberg 2011). Hence, there is an extension of two new emerging hallmarks - deregulating cellular energetics and avoiding immune destruction - and two characteristics that enable the acquisition of all the previous hallmark capabilities: tumor-promoting inflammation and genome instability and mutation (Hanahan & Weinberg 2011; ***Figure1.3B***).

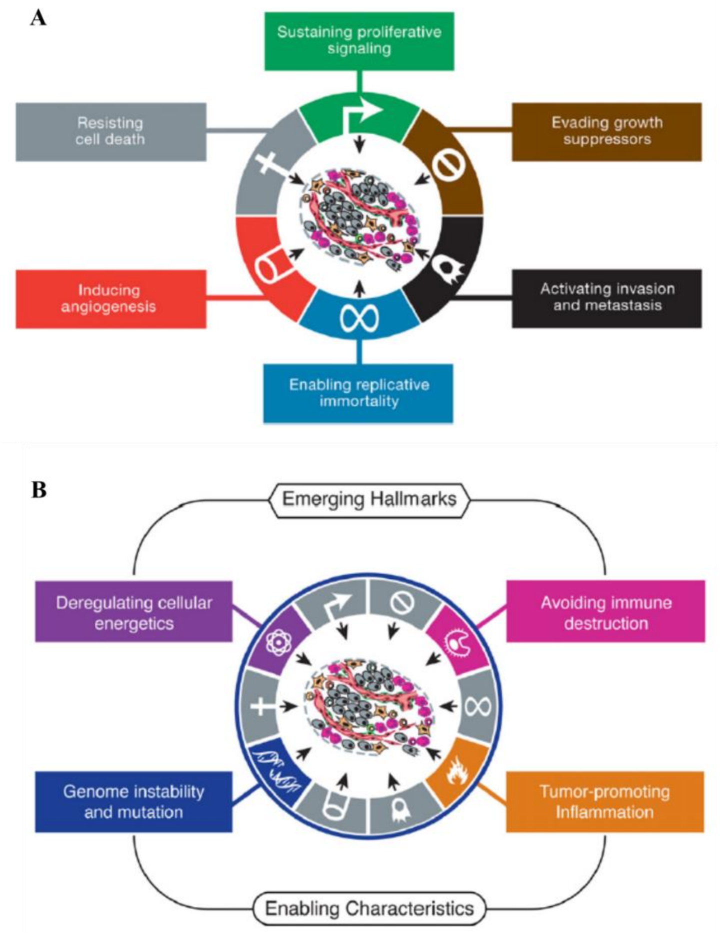

A
Sustaining proliferative signaling
Resisting cell death
Evading growth suppressors
Inducing angiogenesis
Activating invasion and metastasis
Enabling replicative immortality
B
Emerging Hallmarks
Deregulating cellular energetics
Avoiding immune destruction
Genome instability and mutation
Tumor-promoting Inflammation
Enabling Characteristics

Figure 1.3 – Hallmarks of cancer. (A) Illustration of the six first proposed hallmarks of cancer. **(B)** Recently proposed emerging hallmarks and enabling characteristics that contribute for tumorigenesis. From Hanahan et al, 2011.

1.1.3. Modification of signaling pathways

Over the last decades, the evolution of techniques has allowed to more effectively characterize the biology of cancer by identifying specific molecular patterns in solid tumors of various types (Ferté et al. 2010). These altered molecular pathways create the environment conducive to tumorigenesis. With the identification of participants of these circuits, there are new biomarkers that appears as potential biomarkers for clinical application (Ferté et al. 2010).

Although there are many potential new biomarkers, only a small amount is currently used. The idea that a single biomarker is the best way of diagnosis has been surpassed by the existence of a group of biomarkers which are most effective in the diagnosis (Ferté et al. 2010).

Transmission of signals depends on the molecular circuits. In fact, a molecule responds intensely to specific chemicals in its microenvironment: it can adjust its metabolism or alter gene expression patterns. Responses to physiological stimuli are mainly coordinated by chemical signals. Steps of this process which transform the message into a normal physiological response is are termed signal transmission (Berg et al. 2008). All these circuits present the main stages:

1st Liberation of the first messenger - external stimulus from the external environment (Berg et al. 2008);

2nd Reception of the first messenger - since most molecules do not enter the cell. Membrane proteins work as receptors that bind to the signaling molecules and transfer the information to the inside of the cell, according to the stimulatory molecule. Membrane receptors involve the entire cell, presenting multiple extra- and / or intracellular domains. The functional mechanism of this signal reception is promoted by a shift in the conformation of receptor domains upon binding of the stimulator. Importantly, an extracellular binding site recognizes specifically the signaling molecule (Berg et al. 2008);

3rd Delivery of the message inside the cell by the second messenger. Within cells, other small molecules are important for retransmitting information from receptor-ligand complexes. These molecules vary in concentration in response to environmental signals (Berg et al. 2008). Some of the second important messengers are cyclic AMP, cyclic GMP, calcium ion, inositol 1,4,5-triphosphate (IP_3) and diacylglycerol (DAG). The use of this second messenger molecule has several implications. First, the signal can be amplified significantly. Only a small number of receptors can be active by direct binding of the signaling molecules, but each receptor molecule can activate many second messengers. Secondly, the second messengers are commonly free to diffuse through the cell and influence various processes. Thirdly, the use of second messengers common to many routes creates both opportunities and potential threats (Berg et al. 2008);

4th Activation of effectors that alter the physiological response. The result of the signal in the signaling pathway is to activate or inhibit pumps, enzymes, and gene transcription factors that directly control metabolic pathways, the activation of gene expression, and processes such as nerve transmission (Berg et al. 2008);

5th Conclusion of the signal - after obtaining the physiological response to a signal, the signaling processes end, otherwise the cell loses its ability to respond to new signals (Berg et al. 2008).

Molecular circuits are strong allies in the development of cancer through changes in all phases of signal transmission. Despite external inhibition stimulus, the cell alters the levels of molecules that potentiate proliferation allowing that the cell remains alive as well as the interaction with the receptors. Furthermore, effectors which promote proliferation are often augmented or constitutively active whereas effectors responsible by growth inhibition are blocked and functionless. However, the proper response does not reach its destination, or it is not terminated by remaining active and leading to the abnormality. Thus, signal transduction pathways assume the most critical roles in cancer development and progression, from external stimulus to physiological response, leading to alterations in gene expression.

1.1.4. Tumors arise from many specialized cell types

Human tumors are mostly of epithelial origin, being named carcinomas (Weinberg 2014). The epithelium is a stratified structure and each lamina is constituted by cells whose function is to protect the organs against external aggressions. These structures line the walls of the cavities and channels and are separated from the conjunctive tissue by the basement membrane. This structure separates two types of tissue providing structural support (Weinberg 2014). Carcinomas are from various types and locations since gastrointestinal tract epithelium, to the skin, mammary gland, pancreas, lung, liver, ovary, uterus, prostate, gallbladder and bladder. On the other hand, there are also sarcomas, hematopoietic malignancies and neuroectodermal tumors (Weinberg 2014).

Carcinomas have different embryonic origins and can be classified according to the germ layer where were originated. Endoderm carcinomas are the most common and are constituted by the epithelia of the lung, liver, gallbladder, pancreas, oesophagus or intestines. Carcinomas of the mesoderm are constituted, for example, by the epithelium of the ovary or kidney. Ectodermal carcinoma arises mainly from the skin (Weinberg 2014).

In terms of function, carcinomas can be classified as squamous carcinomas, when they occur in tissues which have protective functions, or adenocarcinomas when they affect secretory tissues. Some organs such as the lung may have both (Weinberg 2014).

Importantly, the progression of the tumor allows to classify it into benign or malignant. The tumor may be aggressive or have a slow and harmless development (Weinberg 2014).

1.2. Mechanisms of gene expression regulation

Proteins expressed by the cell are codified in coding genes. However, deoxyribonucleic acid (DNA) is not the direct template for protein synthesis (Berg et al. 2008). A DNA sequence is copied by a class of RNA molecules called messenger RNA (mRNA). This process is composed for two steps, transcription and translation whether the gene is coding, or only by transcription, when the gene is non-coding and has regulatory functions. The most common flow of genetic information in normal cells is: DNA, RNA and Protein (Berg et al. 2008).

The genome can be considered a large file which is regulated in a meticulous way to minimize the damage caused by mutations. Most genes are present in identical amounts in all cells, i.e., one copy per haploid cell and two copies per diploid cell. Importantly, the level of gene expression, indicated by the number of mRNA copies, can vary widely, ranging from no expression to hundreds of copies. Additionally, the expression levels of the same gene may still vary and tends to account for the cell response to microenvironmental stimuli (Berg et al. 2008).

Many genes presented in eukaryotic cells are considered as housekeeping genes, since they are constitutively expressed at low levels in all cells, being essentially responsible for encoding metabolism enzymes or cellular components (Hartl 2014). The expression levels of the remaining genes differs according to the cell type or stage of the cell cycle, being regulated by the control of transcription (Hartl 2014). There are different mechanisms involved in the transcription regulation, and certainly several of them are still not known. From these, mechanisms linked to epigenetics have been extremely relevant in this area.

1.2.1. Epigenetic modifications

Epigenetic modifications are reversible and do not change the DNA sequence. Among these are histone post-translational modifications, DNA methylation and non-coding RNAs, especially miRNAs (Dawson & Kouzarides 2012). The epigenome is crucial for regulating the physiology and pathology characteristic of each cell type. The specific pattern of gene expression and cell phenotypes are controlled epigenetically by marking histones through chemical changes and DNA through methylation, as well as through mechanisms such as incorporation of histone variants, transcription of non-coding RNAs, editing of RNA and three-dimensional chromatin remodelling (Vogel & Lassmann 2014).

First fundamental epigenetic event in our body is called genomic imprinting and occurs in early stages during the embryonic development. This process affects dozens of mammalian genes and results in the expression of these genes from only one of the two parental chromosomes. Inactivation of one of the gene alleles is regulated by epigenetic instructions, established in the parental germ cells (Reik & Walter 2001). However, the development of the cellular progenitors and later the cellular differentiation and specialization of the tissues presents a unique epigenome that allows each cell type to present different functions and characteristics (Chang & Bruneau 2012; Cedar & Bergman 2011; Bharathy N., Ling B.M.T. 2013).

Contrary to the genome, the epigenome is dynamic and can be modified, and therefore some epigenetic risk markers have the potential to be reversed (***Figure 1.4***). External factors such as drugs, diet or environmental exposure can cause epigenetic changes (Bishop & Ferguson 2015). Fundamental understanding of epigenetic events in cell regulation opens a new window for altering the transcriptional state of cells, leading to changes in tumorigenesis.

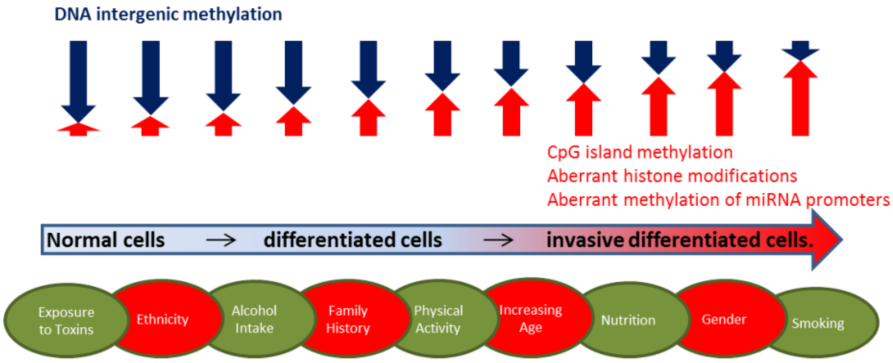

DNA intergenic methylation
CpG island methylation
Aberrant histone modifications
Aberrant methylation of miRNA promoters
Normal cells → differentiated cells → invasive differentiated cells.
Exposure to Toxins
Ethnicity
Alcohol Intake
Family History
Physical Activity
Increasing Age
Nutrition
Gender
Smoking

Figure 1.4 - Epigenetic modifications that promote risk and / or progression of cancer and some factors. The green represents factors that can be modified. In red are represented factors that cannot be modified, they are intrinsic to the individual. From: Bishop KS. et all, 2015.

1.2.1.1 Histone modifications

Histone modifications have a key role in the regulation of chromatin structure through dynamic patterns that can make the chromatin more or less condensed. These post-translational modifications regulate chromatin by influencing the folding, positioning and organization of DNA, altering processes such as gene expression. The complexity of this mechanism lies not only in the pattern of histone modification, but above all in the three-dimensionality of the structures and their dynamics (Bannister & Kouzarides 2011). Many of these modifications are illustrated in ***Figure 1.5*** and include histone methylation, acetylation, phosphorylation and ubiquitylation (Bannister & Kouzarides 2011; Dawson & Kouzarides 2012).

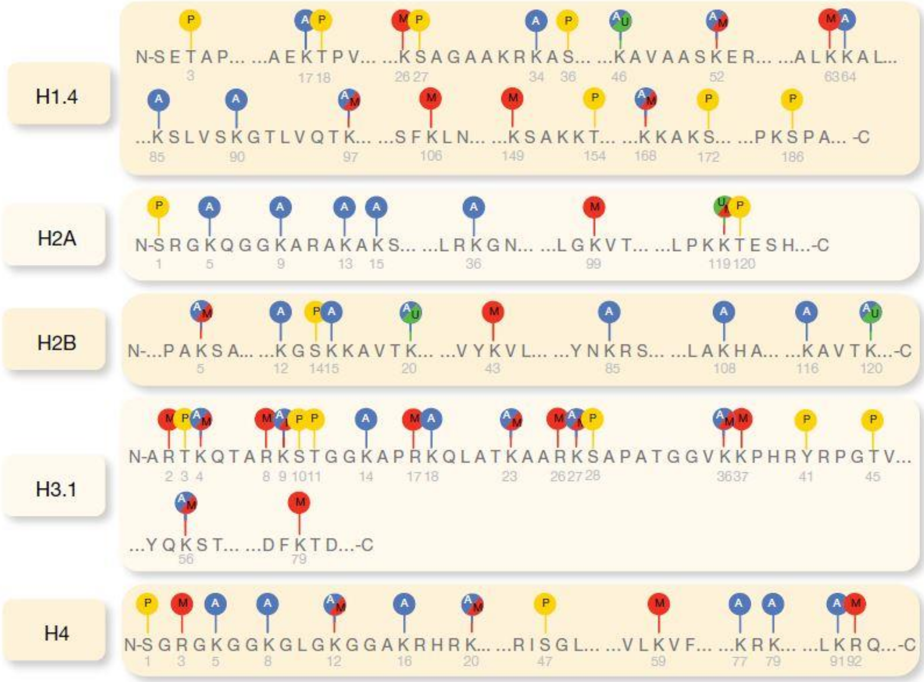

H1.4
N-SETAP... ...AEKTPV... ...KSAGAAKRKAS... ...KAVAASKER... ...ALKKAL...
3 17 18 26 27 34 36 46 52 63 64
...KSLVSKGTLVQTK... ...SFKLN... ...KSAKKT... ...KKAKS... ...PKSPA... -C
85 90 97 106 149 154 168 172 186
H2A
N-SRGKQGGKARAKAKS... ...LRKGN... ...LGKVT... ...LPKKTESH...-C
1 5 9 13 15 36 99 119 120
H2B
N-...PAKSA... ...KGSKKAVTK... ...VYKVL... ...YNKRS... ...LAKHA... ...KAVTK...-C
5 12 14 15 20 43 85 108 116 120
H3.1
N-ARTKQTARKSTGGKAPRKQLATKAARKSAPATGGVKKPHRYRPGTV...
2 3 4 8 9 10 11 14 17 18 23 26 27 28 36 37 41 45
...YQKST... ...DFKTD...-C
56 79
H4
N-SGRGKGGKGLGKGGAKRHRK... ...RISGL... ...VLKVF... ...KRK... ...LKRQ...-C
1 3 5 8 12 16 20 47 59 77 79 91 92
A
M
P
U

Figure 1.5 – Histone modifications. The histone code is defined by the post-translational changes that occur in their tails. The most common changes are illustrated in the figure: acetylation (blue), methylation (red), phosphorylation (yellow) and ubiquitination (green). The number below represents the position of the corresponding amino acid. From: Portela A. et all, 2010.

Alterations in histones can happen through changes directly in the dynamics of the chromatin or involving the binding of the effector molecules (Bannister & Kouzarides 2011; Dawson & Kouzarides 2012). In the first mechanism acetylation and phosphorylation, for example, alter the histone charge due to their negative charge, weakening condensation and promoting the accessibility of DNA to transcription factors. In the second mechanism, some histone modifications do not cause severe changes, however, factors that are associated with it can induce these changes, such as factors associated to chromatin due to their different domains (Bannister & Kouzarides 2011; Dawson & Kouzarides 2012).

Tumorigenesis presents aberrant epigenetic landscapes with post-translational modifications that activate oncogenes or inactivate tumor suppressor genes. The most frequent described alterations are both acetylation and methylation (Dawson & Kouzarides 2012). Acetylation is associated with genetic transcription and promotes chromatin opening due to its negative charge, which weakens the condensation due to neutralization of the positive histone tail. The addition of the acetyl group is carried out by histone acetyltransferases (HATs), and its removal is catalyzed by histone deacetylases (HDACs) (Yoo & Jones 2006). On the other hand, histone methylation acts as either the activator or repressor of gene transcription. The addition of a methyl group is performed by histone methyltransferase enzyme (HMTs) and removed by histone demethylases (HDMs) (Kooistra & Helin 2012).

1.2.1.2 DNA methylation

DNA methylation is one of the epigenetic mechanisms that occurs by adding a methyl (CH_3) group to the DNA sequence, modifying the gene function and, consequently, affecting gene expression. The most well-known and characterized methylation process is the covalent methylation of the carbon 5 of the cytosine pyrimidine ring in cytosine-guanine (CG) pair, resulting in 5-methylcytosine (5-mC).

Methylation is controlled in cells at different levels and the enzymatic reaction is performed by a family of enzymes called DNA methyltransferases (DNMTs). In mammals, DNMT1 is the more active enzyme, being responsible for restoring DNA post-replication methylation sites, called DNA methylation maintenance. DNMT2 exhibits reduced activity in this area with the remaining DNA methyltransferases in vitro (Hermann et al. 2003), but their deletion in embryonic stem cell studies has shown no effect on overall methylation, suggesting that the enzyme does not have a function important in the maintenance and regulation of DNA methylation patterns (Okano et al. 1998). DNMT3A and DNMT3B are responsible for the methylation of new sites through a process called "*de novo*" methylation (Laird 2003; Baubec et al. 2015).

In the *de novo* process, the recruitment of DNMTs to the DNA sequences target is unclear (Klose & Bird 2006). However, there are three possibilities: DNMTs recognize the specific site through motifs; recruitment of DNMTs may occur due to protein-protein contact with transcription repressors or other intervening agents; or the interference-mediated RNA (RNAi) system may target this process for specific targets (Hervouet et al. 2009; Ge et al. 2004; Klose & Bird 2006; Baubec et al. 2015).

The most common and widespread methylation sequence recognition motif is “5'-CpG-3'”. CpG dinucleotides are unevenly distributed throughout the human genome. Moreover, there are defective and highly enriched segments of these dinucleotides, called CpG islands (Laird 2003). About 50% of genes contain CpG islands in the region of the promoter, which is usually hypomethylated and associated to the activation of gene expression. In contrast, the literature frequently reported that hypermethylation in the same region inhibits gene expression. This process can occur through two pathways of inhibition. Firstly, through the recruitment of transcription inhibitors, such as methyl-binding proteins (MBDs), which form part of a large complex including histone deacetylases (HDACs), through co-repressor molecules to silence transcription and modify the surrounding chromatin, promoting a link between DNA methylation and chromatin remodeling and modification. Secondly, by directly blocking the binding site of methylation-sensitive transcription factors (TFs), such as *MYC* (Yoo & Jones 2006). Epigenetic dogma focuses on the hypermethylation associated with gene repression, while hypomethylation has been associated to gene activation. Importantly, studies have shown that hypermethylation may also be associated with the activation of gene expression (Castelo-Branco et al. 2013; Chao et al. 2015). Furthermore, genes that follow this pattern plays a key role in cancer (Bert et al. 2013).

Alterations in enzymes responsible for maintaining the epigenetic homeostasis may promote the deregulation of gene expression and, consequently, lead to tumorigenesis.

1.2.1.3 MicroRNAs (miRNAs)

Small non-coding RNAs (20-30 nucleotides) are associated to a family of proteins called *Argonaute family proteins* (AGO) and can be micro RNA (miRNA), siRNA and PIWI-interacting RNA (piRNA). miRNAs are the most abundant ncRNAs with about 22 nucleotides and act as regulators of protein-coding (Ha & Kim 2014).

Genes which codify to miRNA are transcribed by RNA polymerase II (Pol II) from intronic zones encoding protein genes or from dedicated miRNA gene loci (Lin & Gregory

2015). Primary miRNAs (primiRNAs) after being capped, spliced and polyadenylated are cleaved by the Microprocessor composed by DROSHA, RNase III and its cofactor; critical region of the DiGeorge syndrome (DGCR8); and double stranded RNA (dsRNA) binding protein (Lin & Gregory 2015). However, the pre-miRNAs formed are sent to the cytoplasm and processed by DICER which is helped by TRBP (Lin & Gregory 2015). Subsequently, TRBP assists the binding of DICER to AGO proteins to form the silencing complex (miRISC) which recruits a mature miRNA and promotes the binding to its complementary mRNA leading to post-translational gene silencing or DNA degradation (Lin & Gregory 2015). miRNA is constituted by a recognition domain termed "miRNA seed" which has a length of 6 nucleotides, located from position 2 to 7 of the 5'end. Importantly, the majority of the human protein coding genes have at least one binding site showing the relevance of miRNAs in the regulation of gene expression (Ha & Kim 2014).

Although miRNAs have distinct functions (oncogenic and tumor suppressor), studies have shown that their expression in a tumor is decreased compared to the normal tissue. It was also found that DROSHA and DICER are decreased in some types of cancer (Lin & Gregory 2015).

1.3. Epigenetic biomarkers have potential diagnostic in cancer

Tumorigenesis is associated with successive changes that can be used for diagnosis or prognosis. The great challenge is to catalog all these changes and to find common and distinct patterns that allow to characterize and classify neoplasms correctly. Although each cancer is unique, there are common features that may have potential for clinical application and make diagnosis increasingly early, reducing progression and mortality. Importantly, epigenetic modifications have the potential to be reversible and become good drug targets.

Studies for characterization of DNA methylation that occurs in different types of cancer have emerged in recent years. Specifically, hypermethylation analysis of gene promoters in serum has been shown to be a useful tool for cancer diagnosis (Fujiwara et al. 2005). Epigenetic changes are associated with pathological conditions such as neurological diseases, autoimmune diseases and cancer. Global methylation patterns change with tumorigenesis, causing hypomethylation of CpG probes located in the gene body and hypermethylation of CpG probes located in the gene promoter (De Carvalho et al. 2012). Aberrant DNA methylation of CpG islands is a shared feature of various neoplasms and it is frequently associated with repression of tumor suppressor genes (Fujiwara et al. 2005). Epigenetic modifications constitute an innovative cancer biomarker due to factors such as stability, frequency, reversibility and accessibility in body fluids. Studies have shown the potential of these markers and some of them have already been commercialized (Costa-Pinheiro et al. 2015; Nikolaidis et al. 2012; Kneip et al. 2011; Warren et al. 2011).

Genome-wide analysis of gene expression and DNA methylation using multiple cohorts is useful to identify common or specific patterns between them. Some studies have recently emerged using this more comprehensive approach (Wang et al. 2016; Zhang et al. 2015; Wei et al. 2016; Aran et al. 2017; Moon & Nakai 2018).

2 CHAPTER 2 – AIMS

Databases available online allow to perform analysis based on the use of computational tools to explore a large scale of hypotheses, aiming to find new patterns and eventually new biomarkers to characterize the various types of cancer, for example. The Cancer Genome Atlas (TCGA) provides gene expression and DNA methylation data from samples collected from multiples cohorts, which were obtained through the same techniques. Discovery of DNA methylation patterns associated with changes in gene expression was the major focus for the present study. Therefore, in this dissertation project, we intend to:

1. **Create a working pipeline for genome-wide analysis;**
2. **Identify cancer differentially methylated genes (cDMGs) in early stages**;
3. **Identify genetic/epigenetic patterns across different cancer types**;
4. **Identify epigenetic diagnostic biomarkers**.

3 CHAPTER 3 – MATERIALS AND METHODS

3.1. Bioinformatics principal resources

3.1.1. TCGA – The Cancer Genome Atlas

The Cancer Genome Atlas is a project that resulted from a collaboration between the *National Cancer Institute* (NCI) and the *National Human Genome Research Institute* (NHGRI). This publicly funded project has helped to generate comprehensive, multi-dimensional maps of the key genomic changes for 33 cancer cohorts (Tomczak et al. 2015). This large-scale project as brought together large amounts of genetic and epigenetic information that can be used by the cancer research community to improve prevention, diagnosis and treatment. This database is available to the community (https://cancergenome.nih.gov/) and presents a package in *R* software (*TCGAbiolinks*) that facilitates the selection and download of the study cohort information. This factor was decisive for the selection of the data we use.

3.1.2. Programming R language

R is a programming language within statistical and graphical computing. This open source software provides a wide range of statistical tools and graphics techniques that are characterized by being easily extensible (https://www.r-project.org/). On the other hand, Bioconductor is an open source open source software project that provides tools for the analysis and processing of high-throughput genomic data based on the program *R* language (http://www.bioconductor.org/). Available *Bioconductor* tool used in this study is the *TCGAbiolinks* core package (Anon n.d.). One of these very important tools is the *TCGAbiolinks* core package in this study.

3.1.3. TCGAbiolinks Package

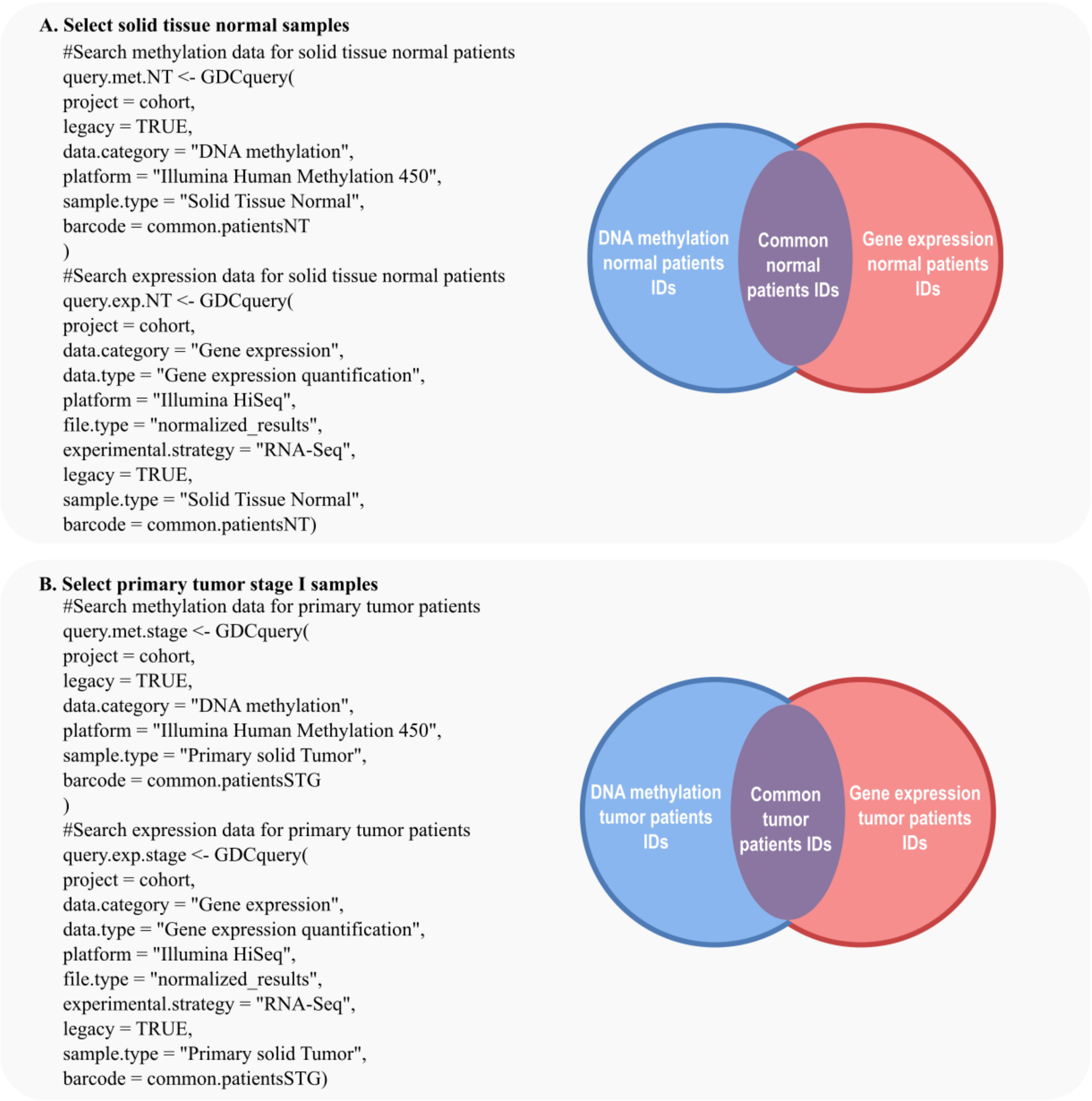
TCGAbiolinks version 2.7.3 is an *R/Bioconductor* package that allows performing a bioinformatic analysis in a sequenced way (Colaprico et al. 2016). The main functions that we used to prepare the database are: *GDCquery*, *GDCdonwload* and *GDCprepare*. *GDCquery* summarizes the information for the GDC sample, taking into account all additional information, such as data category, data type, platform and barcode; *GDCdownload* is a data transfer tool selected by *GDCquery*; *GDCprepare* imports the downloaded data and prepares it for *R* project (Colaprico et al. 2018).

3.2. Data extraction and prepare

3.2.1. Selection of samples

Samples were selected separately for each TCGA cohort with the help of the *GDCquery* function of package *TCGAbiolinks*, mentioned above. We performed an intersection of the patient's ID from both databases, DNA methylation and gene expression, and selected IDs of patients matched. Importantly, the result of two databases was select using *GDCquery* functions with different criterions (***Figure 3.1***). Firstly, for DNA methylation, unmodified data (*legacy = True*) of the *Illumina Human Methylation 450k* array platform (*platform = Illumina Human Methylation 450*) was selected for the matched patients above mentioned (barcode) from the specific cohort (*project = cohort*). Secondly, for gene expression, unmodified data (*legacy = True*) of the *Illumina HiSeq* platform (*platform = Illumina Hiseq*) was selected for the matched patients above mentioned (*barcode*) from the specific cohort (project = cohort) and only normalized results files (*file.type = normalized_results*) of the genetic expression quantification (*data.type = Gene Expression Quantification*) were selected. The samples were grouped according to their typology in normal solid tissue and primary solid tumor (***Figure 3.1A***, ***Figure 3.1B***). Since we considered early stages, the samples corresponding to stage I were collected from the total patient samples using the indicative clinical variable, pathological stage. Additionally, we characterized each type of cancer based directly on the online directory (https://xenabrowser.net/) and the IDs previously selected.

In summary, sample type differentiates the solid tissue normal and primary solid tumor for both databases resulting in four *GDCquery* that define the bases of DNA methylation and gene expression for different types of samples. However, of the 33 cohorts present in the TCGA, 9 cohorts were selected considering the match between patients and a minimum number of 20 samples per group. Of the selected cohorts, colon and rectal cancer were treated together given the similarity between them.



A. Select solid tissue normal samples

```
#Search methylation data for solid tissue normal patients
query.met.NT <- GDCquery(
project = cohort,
legacy = TRUE,
data.category = "DNA methylation",
platform = "Illumina Human Methylation 450",
sample.type = "Solid Tissue Normal",
barcode = common.patientsNT
)
#Search expression data for solid tissue normal patients
query.exp.NT <- GDCquery(
project = cohort,
data.category = "Gene expression",
data.type = "Gene expression quantification",
platform = "Illumina HiSeq",
file.type = "normalized_results",
experimental.strategy = "RNA-Seq",
legacy = TRUE,
sample.type = "Solid Tissue Normal",
barcode = common.patientsNT)
```

DNA methylation normal patients IDs
Common normal patients IDs
Gene expression normal patients IDs

B. Select primary tumor stage I samples

```
#Search methylation data for primary tumor patients
query.met.stage <- GDCquery(
project = cohort,
legacy = TRUE,
data.category = "DNA methylation",
platform = "Illumina Human Methylation 450",
sample.type = "Primary solid Tumor",
barcode = common.patientsSTG
)
#Search expression data for primary tumor patients
query.exp.stage <- GDCquery(
project = cohort,
data.category = "Gene expression",
data.type = "Gene expression quantification",
platform = "Illumina HiSeq",
file.type = "normalized_results",
experimental.strategy = "RNA-Seq",
legacy = TRUE,
sample.type = "Primary solid Tumor",
barcode = common.patientsSTG)
```

DNA methylation tumor patients IDs
Common tumor patients IDs
Gene expression tumor patients IDs

Figure 3.1 - Samples selection. (A) Solid tissue normal samples selection. **(B)** Primary solid tumor samples selection. DNA methylation samples are selected based on the following criteria: project that defines which cohort; legacy provides access to an unmodified copy of data; data.category is defined as DNA methylation; platform is given by Illumina Human Methylation 450; sample.type is set to Solid Tissue Normal; and the barcode defines the IDs for patients with both samples. However, gene expression samples are selected based on the criteria: project that defines the cohort; data.category is define like Gene Expression; legacy provides access to an unmodified copy of data; data.type is defined as Gene Expression quantification; platform is given by Illumina HiSeq; file.type selects normalized_results data; and experimental.strategy defines the purpose of the RNA-Seq study. Sample.type and barcode aggregate samples of patients by typology.

3.3. Characterization of cohorts

Additionally, were exported clinical variables for each type of cancer using an online repository (https://xenabrowser.net/) from IDs previously selected by *TCGAbiolinks* package to perform a descriptive analysis that characterized the patients. Some variables such as age, gender, race, pathologic T, pathologic M and pathologic N are present in all cohorts.

3.3.1. Breast invasive carcinoma (TCGA-BRCA)

Breast invasive carcinoma was selected from the repository with 84 normal solid tissue patients and 126 primary tumor patients in stage I, both with samples for DNA methylation and gene expression.

In ***Table 3.1*** are represented the characteristics of patients for this cohort: the patients are all female with mean age and standard deviation of 58±15 for normal and 60±13 for tumor. The predominant race in normal patients is white with approximately 93%, while in white tumor with approximately 75% of whites and 21% of blacks or African Americans.

Considering the histological type, we verified the high incidence of ductal infiltration carcinoma. Associated with this factor is the presence of estrogen, progesterone and HER_2 receptors. Many patients are positive for estrogen receptors with approximately 70% normal and 75% for tumor, progesterone receptor positive with approximately 60% for normal and 67% for tumor, and finally, approximately 56% of the tumors are negative for HER_2 (***Table 3.1***).

Clinically, the patients were mostly not exposed to neoadjuvant therapy and about half of both groups were subjected to radiation. Tumor patients are classified mainly in T1, N0 and M0 (***Table 3.1***).

Table 3.1 – Characteristics of the patients from TCGA breast invasive carcinoma cohort.

TCGA-BRCA			
Characteristics		Solid Tissue Normal (n = 84)	Primary Tumor (n = 126)
Age	Mean ± SD[1]	58 ± 15	60 ± 13
Gender	Female	84 (100%)	126 (100%)
Race	Asian	1 (01.19%)	3 (02.38%)
	Black or African American	4 (07.41%)	26 (20.63%)
	White	78 (92.86%)	95 (75.40%)
	Not reported	1 (01.19%)	2 (01.59%)
Pathologic Stage	Stage I	11 (13.10%)	126 (100%)
	Stage II	51 (60.71%)	0
	Stage III	20 (23.81%)	0
	Stage IV	1 (01.19%)	0
	Not Reported	1 (01.19%)	0
Histological Type	Infiltrating Ductal Carcinoma	68 (80.95%)	89 (70.63%)
	Infiltrating Lobular Carcinoma	4 (04.76%)	20 (15.87%)
	Medullary Carcinoma	2 (02.38%)	1 (00.79%)
	Metaplastic Carcinoma	0	1 (00.79%)
	Mixed Histology	9 (10.71%)	4 (03.17%)
	Mucinous Carcinoma	0	4 (03.17%)
	Other	1 (01.19%)	6 (04.76%)
	Not Reported	0	1 (00.79%)
History of Neoadjuvant Treatment	No	84 (100%)	125 (99.21%)
	Yes	0	1 (00.79%)
Radiation Therapy	No	26 (30.95%)	56 (44.44%)
	Yes	35 (41.67%)	60 (47.62%)
	Not Reported	23 (27.38%)	10 (07.94%)
Estrogen Receptor Status	No	13 (15.48%)	26 (20.63%)
	Yes	59 (70.24%)	95 (75.40%)
	Not Reported	12 (14.29%)	5 (03.97%)
Progesterone Receptor Status	No	21 (25.00%)	37 (29.37%)
	Yes	50 (59.52%)	84 (66.67%)
	Not Reported	13 (15.48%)	5 (03.97%)
HER2 Receptor Status	No	0	70 (55.56%)
	Yes	0	5 (03.97%)
	Not Reported	0	51 (40.48%)
Pathologic T	T1	18 (21.43%)	126 (100%)
	T2	52 (61.90%)	0
	T3	9 (10.71%)	0
	T4	5 (05.95%)	0
Pathologic N	N0	31 (36.90%)	120 (95.24%)
	N1	37 (44.05%)	4 (03.17%)
	N2	9 (10.71%)	0
	N3	4 (04.76%)	0
	NX	3 (03.57%)	2 (01.59%)
Pathologic M	M0	78 (92.86%)	107 (84.92%)
	M1	1 (01.19%)	0
	MX	5 (05.95%)	19 (15.08%)

1. Standard Deviation

3.3.2. Colorectal adenocarcinoma (TCGA-COADREAD)

Colorectal adenocarcinoma was selected from the repository with 21 normal solid tissue patients and 54 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.2*** are represented the characteristics of patients for this cohort: in normal patients, 22% are men and 17% are women, the mean age and standard deviation is 68±13. In tumor patients, 59% are men and 41% women, the mean age and standard deviation is 66±13. Both groups are mostly made up of white individuals.

Considering the histological type, we verified that 90% of the normal ones have samples of colon adenocarcinoma, whereas in the tumors 70% of colon adenocarcinoma and 19% of rectum adenocarcinoma. The most common anatomical subdivision in cancer development is cecum with 35% and colon sigmoid with 19% (***Table 3.2***).

Clinically, patients were mostly not exposed to neoadjuvant therapy or radiation. 57% of colorectal patients have a family history of polyps and are classified mainly in T2, N0 and M0 (***Table 3.2***).

Table 3.2 – Characteristics of the patients from TCGA colorectal adenocarcinoma cohort.

TCGA-COADREAD			
Characteristics		Solid Tissue Normal (n = 21)	Primary Tumor (n = 54)
Age	Mean ± SD[1]	68 ± 13	66 ± 13
Gender	Female	9 (16.67%)	22 (40.74%)
	Male	12 (22.22%)	32 (59.26%)
Race	Asian	0	0
	Black or African American	3 (14.29%)	8 (14.81%)
	White	10 (47.62%)	41 (75.93%)
	Not reported	8 (38.10%)	5 (09.26%)
Pathologic Stage	Stage I	2 (09.52%)	54 (100%)
	Stage II	11 (52.38%)	0
	Stage III	3 (14.29%)	0
	Stage IV	5 (23.81%)	0
Histological Type	Colon Adenocarcinoma	19 (90.48%)	38 (70.37%)
	Colon Mucinous Adenocarcinoma	0	5 (09.26%)
	Rectal Adenocarcinoma	2 (09.52%)	10 (18.52%)
	Not reported	0	1 (01.85%)
Anatomic Neoplasm Subdivision	Ascending Colon	2 (09.52%)	6 (11.11%)
	Cecum	4 (19.05%)	19 (35.19%)
	Descending Colon	1 (04.76%)	2 (03.70%)
	Hepatic Flexure	2 (09.52%)	2 (03.70%)
	Rectosigmoid Junction	0	6 (11.11%)
	Rectum	2 (09.52%)	4 (07.41%)
	Sigmoid Colon	8 (38.10%)	10 (18.52%)
	Splenic Flexure	0	1 (01.85%)
	Transverse Colon	0	3 (05.56%)
	Not reported	2 (09.52%)	1 (01.85%)
Lymphatic Invasion	No	11 (52.38%)	42 (77.78%)
	Yes	7 (33.33%)	6 (11.11%)
	Not reported	3 (14.29%)	6 (11.11%)
History of Colon Polyps	No	6 (28.57%)	31 (57.41%)
	Yes	8 (38.10%)	13 (24.07%)
	Not reported	7 (33.33%)	10 (18.52%)
History of Neoadjuvant Treatment	No	21 (100%)	54 (100%)
	Yes	0	0
Radiation Therapy	No	16 (76.19%)	46 (85.19%)
	Yes	1 (04.76%)	0
	Not reported	4 (19.05%)	8 (14.81%)
Pathologic T	T1	0	8 (14.81%)
	T2	2 (09.52%)	45 (83.33%)
	T3	17 (80.95%)	0
	T4	2 (09.52%)	0
	Tis	0	1 (01.85%)
Pathologic N	N0	14 (66.67%)	54 (100%)
	N1	4 (19.05%)	0
	N2	3 (14.29%)	0
Pathologic M	M0	11 (52.38%)	45 (83.33%)
	M1	5 (23.81%)	0
	MX	4 (19.05%)	9 (16.67%)
	Not reported	1 (04.76%)	0

1. Standard Deviation

3.3.3. Head and neck squamous cell carcinoma (TCGA-HNSC)

Head and neck squamous cell carcinoma was selected from the repository with 20 normal solid tissue patients and 27 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.3*** are represented the characteristics of patients for this cohort: in normal patients, 15% are men and 5% are women, the mean age and standard deviation is 64±12. In tumor patients, 52% are men and 48% women, the mean age and standard deviation is 62±16. Both groups are mostly made up of white individuals.

Considering the histological type, we verified that 100% of the normal ones have samples of head and neck squamous cell carcinoma in the tumors 96%. The most common anatomical subdivision in cancer development is oral tongue with 56% (***Table 3.3***).

Clinically, patients were mostly not exposed to radiation and are classified mainly in T1, N0 and M0 (***Table 3.3***).

Table 3.3 – Characteristics of the patients from TCGA head and neck squamous cell carcinoma cohort.

TCGA-HNSC			
Characteristics		Solid Tissue Normal (n = 20)	Primary Tumor (n = 27)
Age	Mean ± SD[1]	64 ± 12	62 ± 16
Gender	Female	5 (%)	13 (48.15%)
	Male	15 (%)	14 (51.85%)
Race	Asian	0	1 (03.70%)
	Black or African American	0	2 (07.41%)
	White	20 (100%)	23 (85.19%)
	Not reported	0	1 (03.70%)
Pathologic Stage	Stage I	0	27 (100%)
	Stage II	6 (30.00%)	0
	Stage III	4 (20.00%)	0
	Stage IV	10 (50.00%)	0
Anatomic Neoplasm Subdivision	Alveolar Ridge	0	1 (03.70%)
	Base of tongue	1 (05.00%)	0
	Floor of mouth	2 (10.00%)	1 (03.70%)
	Hard Palate	0	1 (03.70%)
	Larynx	5 (25.00%)	2 (07.41%)
	Lip	0	2 (07.41%)
	Oral Cavity	3 (15.00%)	1 (03.70%)
	Oral Tongue	9 (45.00%)	15 (55.56%)
	Oropharynx	0	1 (03.70%)
	Tonsil	0	3 (11.11%)
Histological Type	Head & Neck Squamous Cell Carcinoma	20 (100%)	26 (96.30%)
	Head & Neck Squamous Cell Carcinoma Basaloid Type	0	1 (03.70%)
Lymphovascular Invasion	No	6 (30.00%)	16 (59.26%)
	Yes	4 (20.00%)	2 (07.41%)
	Not Reported	10 (50.00%)	9 (33.33%)
Radiation Therapy	No	5 (25.00%)	18 (66.67%)
	Yes	6 (30.00%)	8 (29.63%)
	Not Reported	9 (45.00%)	1 (03.70%)
Pathologic T	T1	0	27 (100%)
	T2	7 (35.00%)	0
	T3	6 (30.00%)	0
	T4	7 (35.00%)	0
Pathologic N	N0	7 (35.00%)	24 (88.89%)
	N1	2 (10.00%)	0
	N2	6 (30.00%)	0
	NX	5 (25.00%)	3 (11.11%)
Pathologic M	M0	0	12 (44.44%)
	MX	0	3 (11.11%)
	Not Reported	20 (100%)	12 (44.44%)

1. Standard Deviation

3.3.4. Kidney renal clear cell carcinoma (TCGA-KIRC)

Kidney renal clear cell carcinoma was selected from the repository with 24 normal solid tissue patients and 155 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.4*** are represented the characteristics of patients for this cohort: in normal patients, 75% are men and 25% are women, the mean age and standard deviation is 67±13. In tumor patients, 57% are men and 43% women, the mean age and standard deviation is 60±13. The predominant race in normal patients is white with approximately 83%, while in white tumor with approximately 77% of whites and 23% of blacks or African Americans.

Considering the histological type, we verified that 100% of patients have samples of kidney clear cell renal carcinoma (***Table 3.4***).

Clinically, tumor patients were mostly not exposed to radiation and are classified mainly in T1, NX and M0 (***Table 3.4***).

Table 3.4 – Characteristics of the patients from TCGA kidney renal clear cell carcinoma cohort.

TCGA-KIRC

Characteristics		Solid Tissue Normal (n = 24)	Primary Tumor (n = 155)
Age	Mean ± SD[1]	67 ± 13	60 ± 13
Gender	Female	6 (25.00%)	67 (43.23%)
	Male	18 (75.00%)	88 (56.78%)
Race	Asian	0	0
	Black or African American	1 (04.17%)	35 (22.58%)
	White	20 (83.33%)	120 (77.42%)
	Not reported	3 (12.50%)	0
Pathologic Stage	Stage I	1 (04.17%)	155 (100%)
	Stage II	8 (33.33%)	0
	Stage III	7 (29.17%)	0
	Stage IV	8 (33.33%)	0
Histological Type	Kidney Clear Cell Rel Carcinoma	24 (100%)	155 (100%)
Radiation Therapy	No	2 (08.33%)	71 (45.81%)
	Not Reported	22 (91.67%)	84 (54.19%)
Pathologic T	T1	3 (12.50%)	155 (100%)
	T2	10 (41.67%)	0
	T3	10 (41.67%)	0
	T4	1 (04.17%)	0
Pathologic N	N0	10 (41.67%)	59 (38.06%)
	NX	14 (58.33%)	96 (61.94%)
Pathologic M	M0	17 (70.83%)	131 (84.52%)
	M1	7 (29.17%)	0
	MX	0	22 (14.19%)
	Not Reported	0	2 (01.29%)

1. Standard Deviatio

3.3.5. Kidney renal papillary cell carcinoma (TCGA-KIRP)

Kidney renal papillary cell carcinoma was selected from the repository with 23 normal solid tissue patients and 167 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.5*** are represented the characteristics of patients for this cohort: in normal patients, 65% are men and 35% are women, the mean age and standard deviation is 63±14. In tumor patients, 75% are men and 25% women, the mean age and standard deviation is 62±12. The predominant race in normal patients is white with approximately 74%, while in white tumor with approximately 70% of whites and 22% of blacks or African Americans.

Considering the histological type, we verified that 100% of patients have samples of kidney papillary cell renal carcinoma (***Table 3.5***).

Clinically, patients were mostly not exposed to neoadjuvant therapy or radiation. Tumor patients are classified mainly in T2, NX and MX (***Table 3.5***).

Table 3.5 – Characteristics of the patients from TCGA kidney papillary clear cell carcinoma cohort.

TCGA-KIRP

Characteristics		Solid Tissue Normal (n = 23)	Primary Tumor (n = 167)
Age	Mean ± SD[1]	63 ± 14	62 ± 12
Gender	Female	8 (34.78%)	42 (25.15%)
	Male	15 (65.22%)	125 (74.85%)
Race	American Indian or Alaska Native	0	2 (01.20%)
	Asian	0	3 (01.80%)
	Black or African American	3 (13.04%)	36 (21.56%)
	White	17 (73.91%)	117 (70.06%)
	Not reported	3 (13.04%)	9 (05.39%)
Pathologic Stage	Stage I	10 (43.48%)	167 (100%)
	Stage II	1 (04.35%)	0
	Stage III	9 (39.13%)	0
	Stage IV	3 (13.04%)	0
Histological Type	Kidney Papillary Rel Cell Carcinoma	23 (100%)	167 (100%)
History of Neoadjuvant Treatment	No	23 (100%)	167 (100%)
Radiation Therapy	No	0	125 (74.85%)
	Not Reported	23 (100%)	42 (25.15%)
Pathologic T	T1	10 (43.48%)	167 (100%)
	T2	1 (04.35%)	0
	T3	11 (47.83%)	0
	T4	1 (04.35%)	0
Pathologic N	N0	8 (34.78%)	26 (15.57%)
	N1	5 (21.74%)	0
	NX	10 (43.48%)	141 (84.43%)
Pathologic M	M0	16 (69.57%)	52 (31.14%)
	M1	2 (08.70%)	0
	MX	5 (21.74%)	112 (67.07%)
	Not Reported	0	3 (01.80%)

1. Standard Deviation

3.3.6. Liver hepatocelular carcinoma (TCGA-LIHC)

Liver hepatocellular carcinoma was selected from the repository with 41 normal solid tissue patients and 171 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.6*** are represented the characteristics of patients for this cohort: in normal patients, 56% are men and 44% are women, the mean age and standard deviation is 60±16.
In tumor patients, 71% are men and 29% women, the mean age and standard deviation is 61±12.

The predominant race in normal patients is white with approximately 63%, while in tumor with approximately 46% of whites and 46% of Asian.

Considering the histological type, we verified that 100% of the normal ones have samples of hepatocellular carcinoma and in the tumors 98% (***Table 3.6***).

Clinically, patients were mostly not exposed to neoadjuvant therapy or radiation. Tumor patients are classified mainly in T1, N0 and M0 (***Table 3.6***).

Table 3.6 – Characteristics of the patients from TCGA liver hepatocellular carcinoma cohort.

TCGA-LIHC

Characteristics		Solid Tissue Normal (n = 41)	Primary Tumor (n = 171)
Age	Mean ± SD[1]	60 ± 16	61 ± 12
Gender	Female	18 (43.90%)	50 (29.24%)
	Male	23 (56.10%)	121 (70.76%)
Race	Asian	5 (12.20%)	79 (46.20%)
	Black or African American	7 (17.07%)	8 (04.68%)
	White	26 (63.41%)	78 (45.61%)
	Not reported	3 (07.32%)	6 (03.51%)
Pathologic Stage	Stage I	17 (41.46%)	171 (100%)
	Stage II	7 (17.07%)	0
	Stage III	7 (17.07%)	0
	Stage IV	1 (02.44%)	0
	Not Reported	9 (21.95%)	0
Histological Type	Fibrolamellar Carcinoma	0	2 (01.17%)
	Hepatocellular Carcinoma	41 (100%)	167 (97.66%)
	Hepatocholangial Carcinoma (Mixed)	0	2 (01.17%)
History of Neoadjuvant Treatment	No	41 (100%)	169 (98.83%)
	Yes	0	2 (01.17%)
Radiation Therapy	No	32 (78.05%)	157 (91.81%)
	Yes	2 (04.88%)	2 (01.17%)
	Not Reported	7 (17.07%)	12 (07.02%)
Pathologic T	T1	19 (46.34%)	170 (99.42%)
	T2	10 (24.39%)	0
	T3	9 (21.95%)	0
	T4	3 (07.32%)	0
	Not Reported	0	1 (00.58%)
Pathologic N	N0	25 (60.98%)	127 (74.27%)
	N1	1 (02.44%)	0
	NX	14 (34.15%)	44 (25.73%)
	Not Reported	1 (02.44%)	0
Pathologic M	M0	27 (65.85%)	124 (72.52%)
	M1	1 (02.44%)	0
	MX	13 (31.71%)	47 (27.49%)

1. Standard Deviation

3.3.7. Lung adenocarcinoma (TCGA-LUAD)

Lung adenocarcinoma was selected from the repository with 21 normal solid tissue patients and 245 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.7*** are represented the characteristics of patients for this cohort: in normal patients, 67% are men and 33% are women, the mean age and standard deviation is 64±12. In tumor patients, 41% are men and 59% women, the mean age and standard deviation is 66±10. The predominant race in normal patients is white with approximately 86%, while in tumor with approximately 79% of whites and 10% of black or African Americans.

Considering the histological type, we verified that 76% of the normal ones have samples of “Not Otherwise Specified” and in the tumors 62%. The most common anatomical subdivision in cancer development is R-upper (upper right lung) with 39% (***Table 3.7***).

Clinically, patients were mostly not exposed to neoadjuvant therapy or radiation. When we look at the smoker's history indicator for tumor patients, 12% are "Current reformed smoker for < or = 15 years" and 13% "Current reformed smoker for > 15 years", 64% of which are unreported. Tumor patients are classified mainly in T1/T2, N0 and M0 (***Table 3.7***).

Table 3.7 – Characteristics of the patients from TCGA lung adenocarcinoma cohort.

TCGA-LUAD			
Characteristics		Solid Tissue Normal (n = 21)	Primary Tumor (n = 245)
Age	Mean ± Standard Deviation	64 ± 12	66 ± 10
Gender	Female	7 (33.33%)	145 (59.18%)
	Male	14 (66.67%)	100 (40.82%)
Race	Asian	0	4 (01.63%)
	Black or African American	3 (14.29%)	25 (10.20%)
	White	18 (85.71%)	193 (78.78%)
	Not reported	0	23 (09.39%)
Pathologic Stage	Stage I	12 (57.14%)	245 (100%)
	Stage II	4 (19.05%)	0
	Stage III	4 (19.05%)	0
	Stage IV	1 (04.76%)	0
Anatomic Neoplasm Subdivision	Bronchial	0	1 (00.41%)
	L-Lower	3 (14.29%)	36 (14.69%)
	L-Upper	5 (23.81%)	61 (24.90%)
	Other	0	1 (00.41%)
	R-Lower	1 (04.76%)	39 (15.92%)
	R-Middle	0	9 (03.67%)
	R-Upper	10 (47.62%)	96 (39.18%)
	Not Reported	2 (09.52%)	2 (00.82%)
Histological Type	Lung Acir Adenocarcinoma	0	12 (04.90%)
	Lung Adenocarcinoma Mixed Subtype	2 (09.52%)	44 (17.96%)
	Lung Adenocarcinoma- Not Otherwise Specified (NOS)	16 (76.19%)	152 (62.04%)
	Lung Bronchioloalveolar Carcinoma Mucinous	0	5 (02.04%)
	Lung Bronchioloalveolar Carcinoma Nonmucinous	0	13 (05.31%)
	Lung Micropapillary Adenocarcinoma	0	1 (00.41%)
	Lung Mucinous Adenocarcinoma	2 (09.52%)	1 (00.41%)
	Lung Papillary Adenocarcinoma	0	11 (04.49%)
	Lung Signet Ring Adenocarcinoma	0	1 (00.41%)
	Lung Solid Pattern Predomint Adenocarcinoma	0	4 (01.63%)
	Mucinous (Colloid) Carcinoma	1 (04.76%)	1 (00.41%)
History of Neoadjuvant Treatment	No	20 (95.24%)	245 (100%)
	Yes	1 (04.76%)	0
Radiation Therapy	No	15 (71.43%)	212 (86.53%)
	Yes	4 (19.05%)	16 (06.53%)
	Not Reported	2 (09.52%)	17 (06.94%)
Pathologic T	T1	8 (38.10%)	120 (48.98%)
	T2	11 (52.38%)	125 (51.02%)
	T3	1 (04.76%)	0
	T4	1 (04.76%)	0
Pathologic N	N0	11 (52.38%)	239 (97.55%)
	N1	4 (19.05%)	0
	N2	4 (19.05%)	0
	NX	2 (09.52%)	6 (02.45%)
Pathologic M	M0	18 (85.71%)	158 (64.49%)
	M1	1 (04.76%)	0
	MX	1 (04.76%)	85 (34.69%)
	Not Reported	1 (04.76%)	2

3.3.8. Thyroid carcinoma (TCGA-THCA)

Thyroid carcinoma was selected from the repository with 50 normal solid tissue patients and 284 primary tumor patients in stage I, both with samples for methylation and expression.

In ***Table 3.8*** are represented the characteristics of patients for this cohort: in normal patients, 28% are men and 72% are women, the mean age and standard deviation is 46±17. In tumor patients, 24% are men and 76% women, the mean age and standard deviation is 38±13. The predominant race in normal patients is white with approximately 76%, while in tumor with approximately 65% of whites and 12% of Asian.

Considering the histological type, we verified that 84% of the normal ones have samples of thyroid papillary carcinoma - classical/usual and in the tumors 75% (T***able 3.8***).

Clinically, patients were mostly exposed to radiation, but not neoadjuvant therapy. Tumor patients are classified mainly in T1, N0/N1 and M0/MX (***Table 3.8***).

Table 3.8 – Characteristics of the patients from TCGA thyroid carcinoma cohort.

TCGA-THCA			
Characteristics		Solid Tissue Normal (n = 50)	Primary Tumor (n = 284)
Age	Mean ± SD[1]	46 ± 17	38 ± 13
ºGender	Female	36 (72.00%)	216 (76.06%)
	Male	14 (28.00%)	68 (23.94%)
Race	American Indian or Alaska Native	0	1 (00.35%)
	Asian	3 (06.00%)	35 (12.32%)
	Black or African American	5 (10.00%)	11 (03.87%)
	White	35 (76.00%)	185 (65.14%)
	Not reported	7 (14.00%)	52 (18.31%)
Pathologic Stage	Stage I	30 (60.00%)	284 (100%)
	Stage II	5 (10.00%)	0
	Stage III	12 (24.00%)	0
	Stage IV	3 (06.00%)	0
Histological Type	Thyroid Papillary Carcinoma - Classical/usual	42 (84.00%)	212 (74.65%)
	Thyroid Papillary Carcinoma - Follicular	5 (10.00%)	56 (19.72%)
	Thyroid Papillary Carcinoma - Tall Cell	3 (06.00%)	10 (03.52%)
	Other, specify	0	6 (02.11%)
History of Neoadjuvant Treatment	No	49 (98.00%)	283 (99.65%)
	Yes	1 (02.00%)	1 (00.35%)
Radiation Therapy	No	14 (28.00%)	123 (43.31%)
	Yes	34 (68.00%)	155 (54.58%)
	Not Reported	2 (04.00%)	6 (02.11%)
Pathologic T	T1	9 (18.00%)	121 (42.61%)
	T2	17 (34.00%)	96 (33.80%)
	T3	21 (42.00%)	64 (22.54%)
	T4	3 (06.00%)	1 (00.35%)
	TX	0	2 (00.70%)
Pathologic N	N0	25 (50.00%)	139 (48.94%)
	N1	20 (40.00%)	114 (40.14%)
	NX	5 (10.00%)	31 (10.92%)
Pathologic M	M0	33 (66.00%)	33 (66.00%)
	M1	2 (04.00%)	2 (04.00%)
	MX	15 (30.00%)	15 (30.00%)

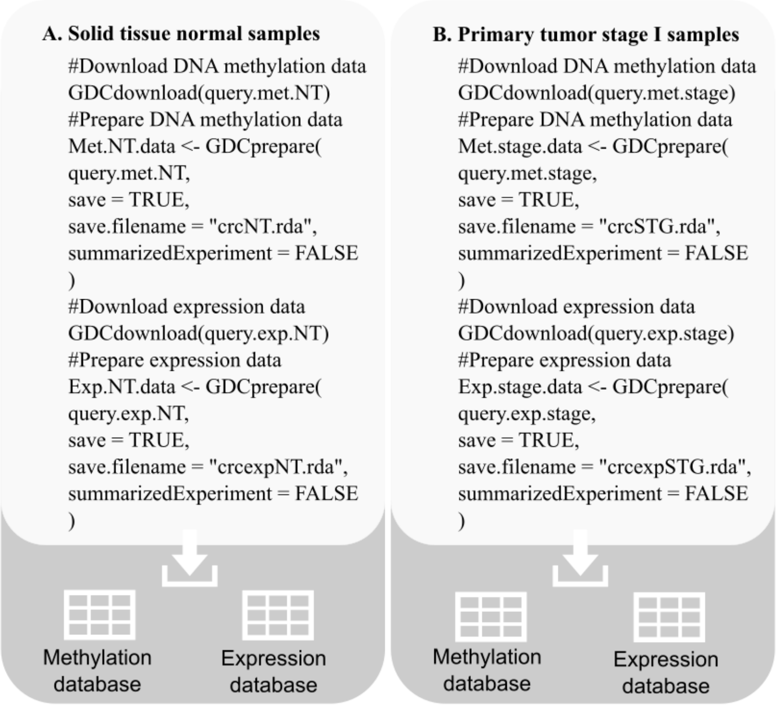
[1] Standard Deviation

3.4. Export and prepare databases

Databases were exported using the *GDCdownload* function and prepared for use in *R*, with the *GDCprepare* function of *TCGAbiolinks* package (Colaprico et al. 2016). Databases of DNA methylation and gene expression for normal solid tissue (**Figure 3.2A**) and primary solid tumor (***Figure 3.2B***) were downloaded, considering the previous function. Files were imported into R and organized by data frames available for use in *R* software.

Databases were also pre-processed excluding: CpG sites non-named, genes not matched to methylation and missing data gene.

This procedure was fundamental considering the work guidelines.



A. Solid tissue normal samples

```
#Download DNA methylation data
GDCdownload(query.met.NT)
#Prepare DNA methylation data
Met.NT.data <- GDCprepare(
query.met.NT,
save = TRUE,
save.filename = "crcNT.rda",
summarizedExperiment = FALSE
)
#Download expression data
GDCdownload(query.exp.NT)
#Prepare expression data
Exp.NT.data <- GDCprepare(
query.exp.NT,
save = TRUE,
save.filename = "crcexpNT.rda",
summarizedExperiment = FALSE
)
```

B. Primary tumor stage I samples

```
#Download DNA methylation data
GDCdownload(query.met.stage)
#Prepare DNA methylation data
Met.stage.data <- GDCprepare(
query.met.stage,
save = TRUE,
save.filename = "crcSTG.rda",
summarizedExperiment = FALSE
)
#Download expression data
GDCdownload(query.exp.stage)
#Prepare expression data
Exp.stage.data <- GDCprepare(
query.exp.stage,
save = TRUE,
save.filename = "crcexpSTG.rda",
summarizedExperiment = FALSE
)
```

Figure 3.2 - Export and prepare solid tissue normal samples. (A) Download and prepare DNA methylation and gene expression for solid tissue normal samples. **(B)** Download and prepare DNA methylation and gene expression for primary tumor samples. Samples were downloaded in separate folders and imported for use in R.

3.4.1. Gene expression Database

Gene expression (data level 3) was obtained from the *Illumina HiSeq 2000* sequencing platform of TCGA genome characterization center, in the University of North Carolina (Illumina 2010; Anon n.d.). This platform tool presents a high accuracy associated with an unprecedented output data demonstrating a breakthrough in the user experience (Illumina 2010). Its ability allows it to process 200 samples of gene expression in a single run by Next Generation Sequence (Illumina 2010). Dataset shows the gene-level transcription estimates in transformed RSEM normalized count (Li & Dewey 2011). RSEM provides accurate quantification of transcription for species with no sequenced genome. This user-friendly and precise software tool is useful for quantifying the abutment of RNA-Seq data transcripts (Li & Dewey 2011). Data were normalized with log(x+1) to linearize the relationship between expression and methylation (Silva et al. 2016) and presented normalized expression values for 20502 genes.

3.4.2. DNA methylation Database

Methylation quantification was obtained from the *Illumina Infinium HumanMethylation450 platform* by Johns Hopkins - USC Epigenome Center (Cost 2012). This high-throughput, low cost tool features extensive genome coverage, including more than 450,000 methylation sites per sample at single nucleotide resolution. Technical capacity is quite powerful, ensuring more than 98% of technical replicas with a simple working protocol (Cost 2012).

The workflow is simple without requiring PCR and need a sample amount as low as 500ng. The *HumanMethylation 450 Beadchip* applies two different chemical assays: *Infinium I* and *Infinium II,* to improve the coverage of the analysis (***Figure 3.3***). However, the development of Infinium II allows the use of degenerate oligonucleotide probes for a single bead type, making CpG sites with three nucleotides of separation have no interference in the methylation analysis. Finally, when we look to sensitivity of tool is able to detect the Δβ value of 0.2 with a lower than 1% false positive rate (Cost 2012).

Each probe has a methylation value (*β-value*) ranging from 0 (hypomethylated) to 1 (hypermethylated), that show the intensity ratio of the methylated (M) bead type according to the combined, methylated and unmethylated (U), locus intensity ($\beta = M/(M+U)$), recorded by GenomeStudio software (Illumina Inc 2010; Siegmund n.d.). The DNA methylation data files included information of signal intensities (raw and normalized), detection confidence, and calculated beta values for methylated (M) and unmethylated (U) probes (Tomczak et al. 2015). Data obtained presented values for 364643 probes.

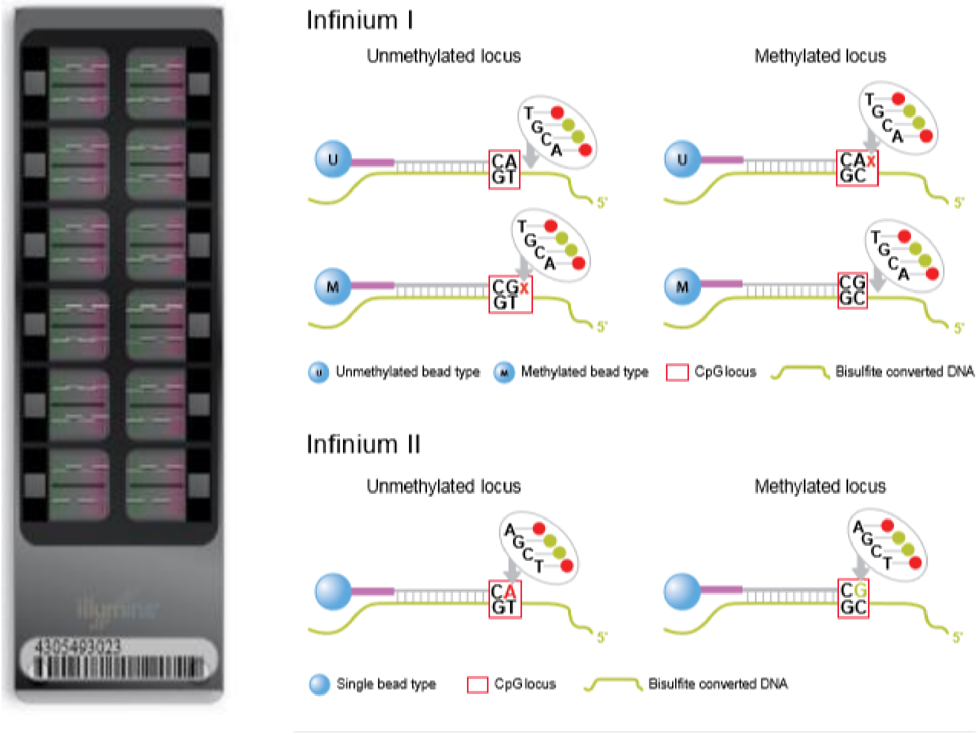


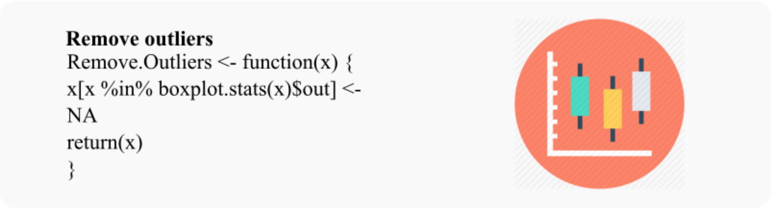
Figure 3.3 – Infinium HumanMethylation 450. In the left side, Infinium HumanMethylation 450 Beadchip with extensive coverage for more than 450 000 methylation sites. In the right side, Infinium I and Infinium II assays chemistry technologies to improve the coverage of the analysis.

3.4.3. Removal of outliers

Outliers assume much greater or less discrepant values in relation to most of the observations made (Cousineau 2011). These observations may not reflect the reality of sampling and may lead to data distortion reflecting changes in the mean value and variance of data that have an impact on results. The variance reflects how far in general the values are from the expected value, which value reflects where the data of a distribution is concentrated. Importantly, statistical inference processes are based on many of these dispersion measures.

The impact of these measures becomes greater when the sample is small or when the statistics are less robust (Cousineau 2011). Thus, it is imperative to carry out the analysis of the sample to reduce the impact of these values in the interpretation of the obtained results.

Procedures for outlier's analysis are diverse. They can be classified in univariate or multivariate fields (Aguinis et al. 2013). An outlier's analysis was performed considering one of the univariate methods analysis, the boxplot method. For gene expression and DNA methylation were performed boxplot representation (***Figure 3.4***). Inferior and superior limit were Q1-1.5 Inter Quartile Range (IQR) and Q3+1.5 IQR, respectively (Aguinis et al. 2013). and the points that lie outside this representation are replaced by *NA* (not applicable) and considered an outlier. This procedure was performed for both bases, DNA methylation and gene expression, by the function present in ***Figure 3.4***.



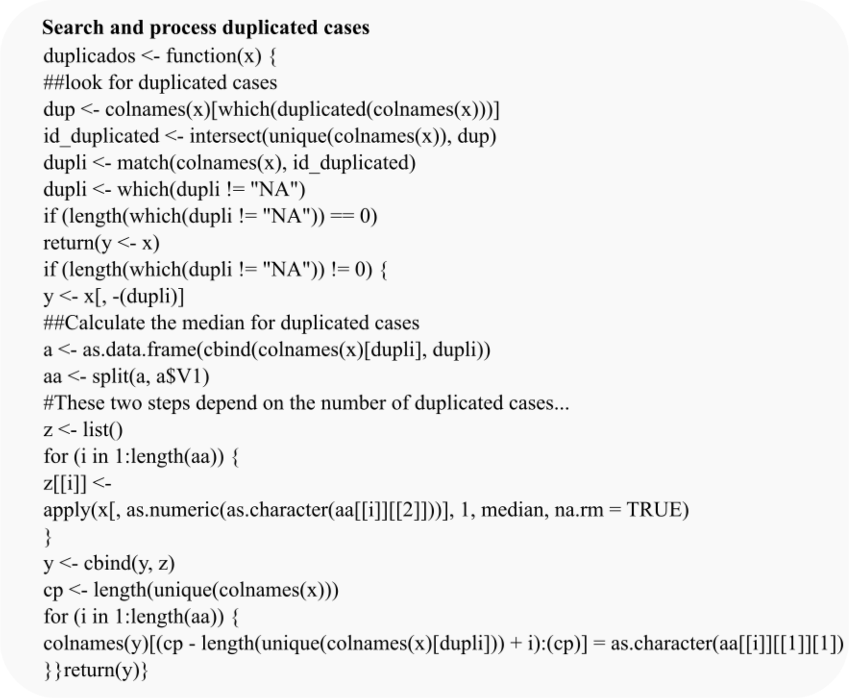
Remove outliers

```
Remove.Outliers <- function(x) {
x[x %in% boxplot.stats(x)$out] <-
NA
return(x)
}
```

Figure 3.4 - Remove outliers function. This function created in R considers the outliers based on the boxplot method, replacing these discrepant values by NA, that is, missing value.

3.4.4. Duplicated cases

In databases dimension analysis, after extraction for *R*, was considered the fact that each patient had two samples. Slight discrepancies were taking account since, in some cohorts, existed more than one sample for primary tumor. In those situations, through ID patient, was considered the median value and created a single data frame resulting from duplicate cases (***Figure 3.5***). Median is a good central tendency measure since extreme values have already been removed (Manikandan 2011).



Search and process duplicated cases

```
duplicados <- function(x) {
##look for duplicated cases
dup <- colnames(x)[which(duplicated(colnames(x)))]
id_duplicated <- intersect(unique(colnames(x)), dup)
dupli <- match(colnames(x), id_duplicated)
dupli <- which(dupli != "NA")
if (length(which(dupli != "NA")) == 0)
return(y <- x)
if (length(which(dupli != "NA")) != 0) {
y <- x[, -(dupli)]
##Calculate the median for duplicated cases
a <- as.data.frame(cbind(colnames(x)[dupli], dupli))
aa <- split(a, a$V1)
#These two steps depend on the number of duplicated cases...
z <- list()
for (i in 1:length(aa)) {
z[[i]] <-
apply(x[, as.numeric(as.character(aa[[i]][[2]]))], 1, median, na.rm = TRUE)
}
y <- cbind(y, z)
cp <- length(unique(colnames(x)))
for (i in 1:length(aa)) {
colnames(y)[(cp - length(unique(colnames(x)[dupli])) + i):(cp)] = as.character(aa[[i]][[1]][1])
}}return(y)}
```

Figure 3.5 - Duplicate cases function. This function identifies duplicate samples IDs and collects those data frames by performing median for be transformed into single data frame.

3.5. Statistical analysis

Descriptive statistics were performed to characterize the cancer cohorts. Inferential statistical analysis was executed to capture populational significant results regarding gene expression and DNA methylation, discriminated by solid tissue normal and stage I primary tumor. Before any comparative analyses (normal vs tumor), was performed a normality test to verify normal distribution. Results were considered statistically significant when *p-value< 0.05*, assuming the False Discovery Rate (FDR) correction < 0.05.

3.5.1. Shapiro–Wilk normality test

Shapiro-Wilk test was used to test the normal distribution. This procedure is important since the interpretation and statistical inference procedures must to be sustained. Shapiro–Wilk normality test represents the most powerful test in order to attest distribution behaviour considering sample size effect (Razali & Wah 2011). Parametric or non-parametric procedures were conducted based on the Shapiro-Wilk test results. A *p-value<0,05* lidded to the rejection of the null hypothesis, a sample came from a normally distributed population (Shapiro & Wilk 1965), and therefore we opted for a non-parametric tests version.

This test was performed using *shapiro.test* function available in *stats* R package.

3.5.2. Levene test

Groups variance is a very important assumption in statistical procedures, especially when we are testing, equal means by groups. Therefore, we performed Levene's test for homogeneity of variance across two samples (Levene 1960). The Levene's test results determined if we opted for Welch test (unequal variances t-test), which is more reliable in this context, or for the Student´s t-test (equal variances t-test).

This test was performed using *leveneTest* function available in *car* R package.

3.5.3. Unpaired two-sample statistical tests

Wilcoxon Rank Sum test (Wilcoxon 1945) or Mann Whitney U test, a non-parametric alternative to Student´s t-test was used to compare two unpaired samples. This test takes into account the ranking differences between the two samples, and tests if the distributions of both populations are equal, or if they were independently selected from populations with same distribution.

Student´s t-test is a parametric test used for testing equal populational means, in the presence of independent samples normally distributed, and assuming similar variance of groups. For unequal variances of groups we performed the Welch test (Welch 1947).

These tests were performed using *wilcox.test* and *t.test* function available in *stats* R package.

3.5.4. Correction of multiple testing

Multiple comparisons increase the probability of taking false positives. An approach to controlling the false discovery rate (FDR) was proposed by Benjamini & Hochberg. This approach takes in account the order of *p-values*, considering the minimum accumulative to verify the criteria $p_{(i)} \leq \left(\frac{i}{m}\right)q, i = 1,..., m$ (Benjamini & Hochberg 1995). This condition provides a new set of *p-values,* which compared to the significance associated with the test and verified the previous criteria can be actualized.

This correction was performed using *p.adjust* function available in *stats* R package.

3.5.5. Pearson correlation test

Pearson´s ρ test was performed aiming to find significant correlations between methylation in gene expression patterns, within primary tumor samples. Correlations between gene expression and DNA methylation of the respective CpG site was tested, where correlation coefficient, ranges from -1 to 1.

This test was performed using *cor.test* function available in *stats* R package.

3.5.6. Receiver operating characteristic (ROC curves)

The receiver operating characteristic (ROC) curve is a graphical representation of the pairs sensitivity, true positive fraction (TPR), and 1-specificity, false positive fraction (FPR). All cutoffs result from the coordinates that represents a compromise between sensitivity and specificity. This measure the quality of a diagnostic biomarker, representing the probability of discriminate stage I primary tumor from normal samples (Fawcett 2006; Robin et al. 2011). Moreover, the area under the curve (AUC) is used as an accuracy index (Fawcett 2006).

AUC is used as a quality measure of the curve and it is calculated through the trapezoid rule (Robin et al. 2011). We defined an AUC ≥ 0.8, a sensibility ≥ 0.6 and a specificity ≥ 0.6 to obtain the potential new biomarkers for further clinical applications (Greiner et al. 2000).
ROC curves are represented in ***Figure 3.6A*** for all genes differentially expressed which presents CpG probes differentially methylated in lung cancer and the ***Figure 3.6B*** represents ROC curves considering the mentioned cutoffs criteria.

This analysis was performed using *roc* and *ggroc* functions available in *pROC* R package (Robin et al. 2018).

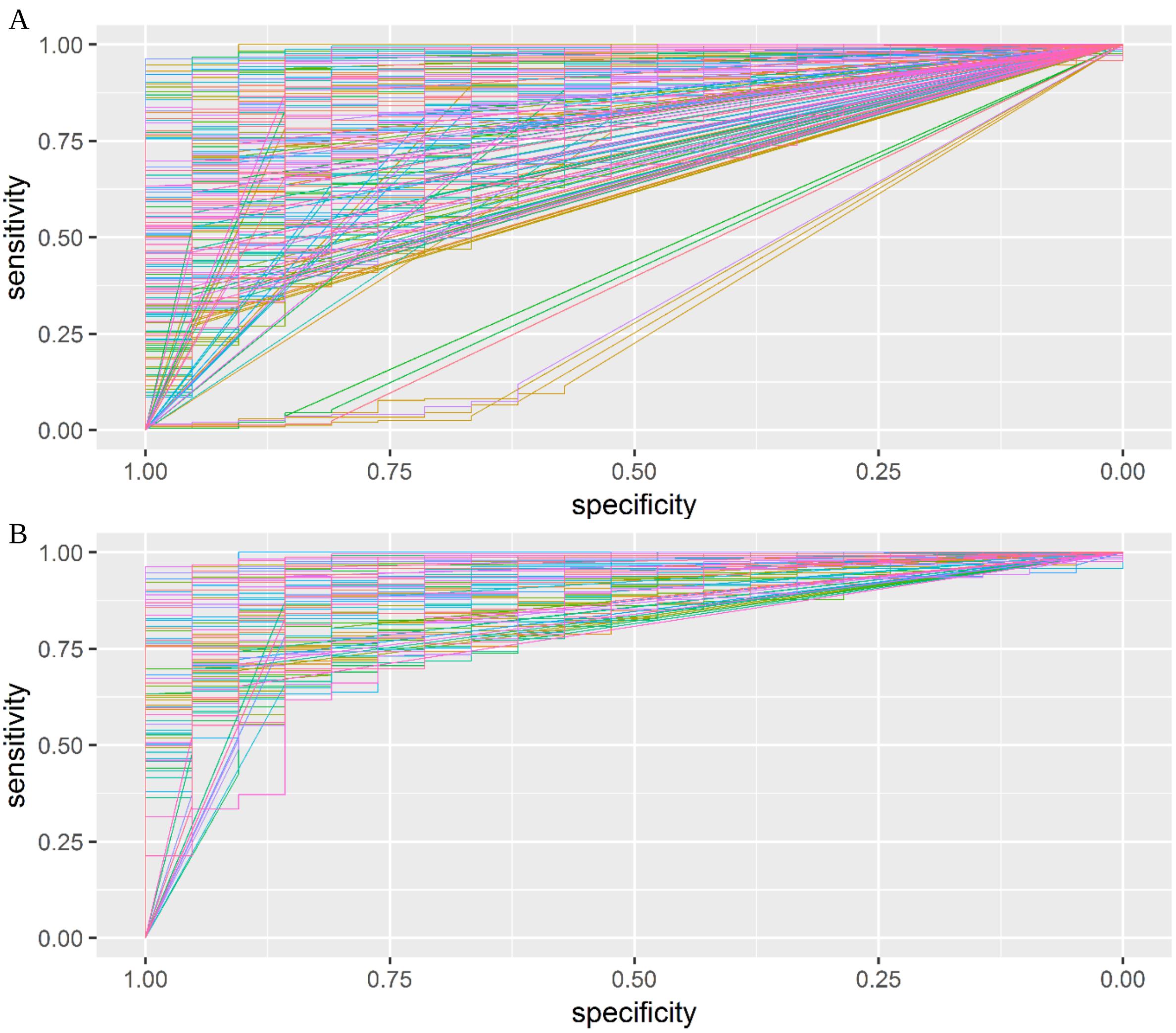


Figure 3.6 – ROC curve analysis of cDMGs of lung cancer. (A) All ROC curves were represented at various colors and **(B)** the selection of area under the curve (AUC) equal or greater than 0.8. Sensitivity represent the probability of true positives and 1-specificity represent the probability of 1-P (false positives), true negatives.

3.5.7. Multiple Linear Regression Model

Multiple linear regression (MLR) was used to analyse the relationships between a dependent variable (gene expression) and multiple independent variables (DNA methylation). The MLR was useful to verify which CpG probes are statistically significant on the variation of gene expression. The MLR model equation is given by,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + e \qquad (1)$$

where $\beta_i, i = 1,..., k$ are the weight that indicates the relative contribution for each unit measure of X_i in dependent variable variation. $X_i, i = 1,..., k$ are the original independent variables (Hair et al. 1999).

This analysis was performed using the *lm* function available in *stats* R package

3.6. Cut-offs selected for DNA methylation and gene expression

Expression databases for solid normal and stage I primary tumor samples were obtained through *TCGAbiolinks* package. Then, for each gene, the ratio between the mean of normal and tumor was measured and obtained foldchange value.

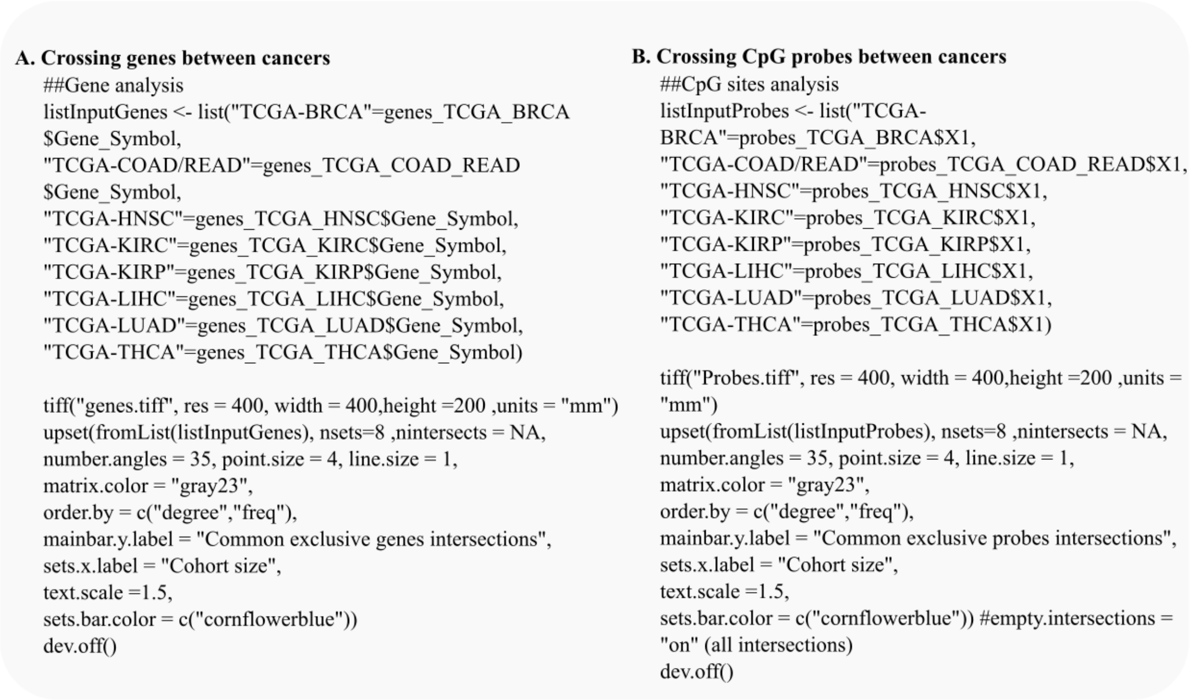
Importantly, to improve the visualization of these differences, the foldchange was transformed by to apply the $\log_2$(Foldchange). Moreover, only genes with $\log_2$(Foldchange)>1.5 (up-regulated) and $\log_2$(Foldchange)<-1.5 (down-regulated) were considered as differentially expressed. This cut-off implies a |Foldchange|>2.82.

Regarding to the DNA methylation data, for each probe, the difference between the mean of normal and tumor beta-value (Δβ) was calculated. Here, CpG sites were considered as differentially methylated when Δβ>0.2 (hypermethylated) or Δβ<-0.2 (hypomethylated).

3.7. Patterns across cancers

To identify common and specific, differentially expressed genes, as well as CpG sites differentially methylated were performed intersections across all cancer cohorts (***Figure 3.7A-B***).

This was performed automatically through the *UpSetR* package in R.



A. Crossing genes between cancers

```
##Gene analysis
listInputGenes <- list("TCGA-BRCA"=genes_TCGA_BRCA
$Gene_Symbol,
"TCGA-COAD/READ"=genes_TCGA_COAD_READ
$Gene_Symbol,
"TCGA-HNSC"=genes_TCGA_HNSC$Gene_Symbol,
"TCGA-KIRC"=genes_TCGA_KIRC$Gene_Symbol,
"TCGA-KIRP"=genes_TCGA_KIRP$Gene_Symbol,
"TCGA-LIHC"=genes_TCGA_LIHC$Gene_Symbol,
"TCGA-LUAD"=genes_TCGA_LUAD$Gene_Symbol,
"TCGA-THCA"=genes_TCGA_THCA$Gene_Symbol)

tiff("genes.tiff", res = 400, width = 400,height =200 ,units = "mm")
upset(fromList(listInputGenes), nsets=8 ,nintersects = NA,
number.angles = 35, point.size = 4, line.size = 1,
matrix.color = "gray23",
order.by = c("degree","freq"),
mainbar.y.label = "Common exclusive genes intersections",
sets.x.label = "Cohort size",
text.scale =1.5,
sets.bar.color = c("cornflowerblue"))
dev.off()
```

B. Crossing CpG probes between cancers

```
##CpG sites analysis
listInputProbes <- list("TCGA-
BRCA"=probes_TCGA_BRCA$X1,
"TCGA-COAD/READ"=probes_TCGA_COAD_READ$X1,
"TCGA-HNSC"=probes_TCGA_HNSC$X1,
"TCGA-KIRC"=probes_TCGA_KIRC$X1,
"TCGA-KIRP"=probes_TCGA_KIRP$X1,
"TCGA-LIHC"=probes_TCGA_LIHC$X1,
"TCGA-LUAD"=probes_TCGA_LUAD$X1,
"TCGA-THCA"=probes_TCGA_THCA$X1)

tiff("Probes.tiff", res = 400, width = 400,height =200 ,units =
"mm")
upset(fromList(listInputProbes), nsets=8 ,nintersects = NA,
number.angles = 35, point.size = 4, line.size = 1,
matrix.color = "gray23",
order.by = c("degree","freq"),
mainbar.y.label = "Common exclusive probes intersections",
sets.x.label = "Cohort size",
text.scale =1.5,
sets.bar.color = c("cornflowerblue")) #empty.intersections =
"on" (all intersections)
dev.off()
```

Figure 3.7 - Patterns across cancers analysis. (A) Gene analysis of patterns across cancers types in early stages. **(B)** CpG sites analysis of patterns across types of cancer in early stages.

3.8. HJ-biplot

HJ-Biplot (Galindo Villardón 1985) is a multivariate approach that allows to visualize the patient's distribution according to the variables that more contribute to this distribution, considering to their norm. Since this is a data reduction technique, the patient coordinates were used to apply a hierarchical cluster analysis, considering the square Euclidean distance and using the Ward method (Hair et al. 1999). Therefore, both normal and cancer samples were distributed by 3 groups for all cohorts. Were strategically 3 clusters to observed possible undefined samples.

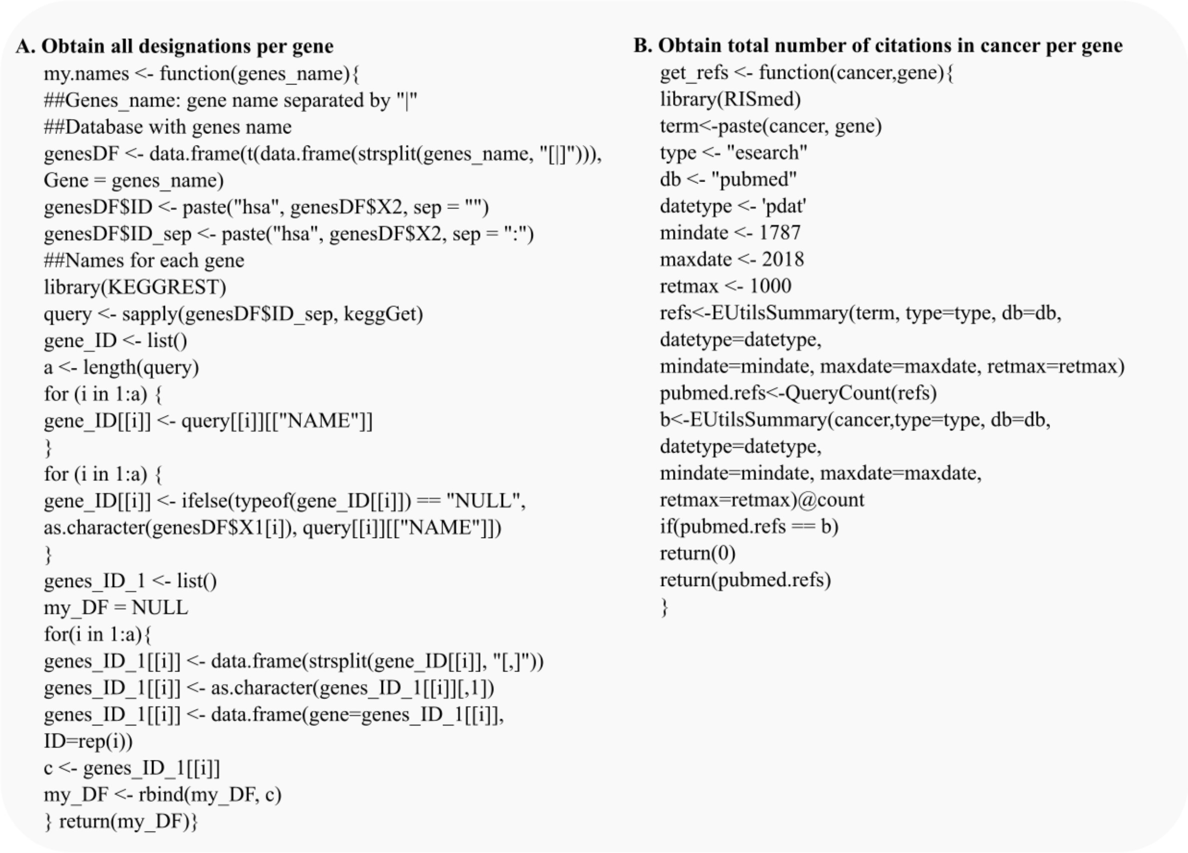
Only variables with contributions, of factor to the element, over than 0.7 were considered. Since this technique does not deal with missing data, these cases were replaced by the median value of that variable.

HJ-Biplot and hierarchical clusters were performed using *HJ.Biplot* and *AddCluster2Biplot* functions provided by the *MultBiplotR* R package (Vicente-villardon 2015).

3.9. Oncosearch algorithm

RISmed package provides a text-mining tool that was used to query PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) for references and selected keywords.

Firstly, *my.names* function, adapted from a function provided by *KEGGREST* package which returns all designations for each gene (***Figure 3.8A***). Then *get_refs* function, created for searching citations in PubMed according to the association between gene to cancer, or gene to specific cancer, was used (***Figure 3.8B***). Differentially methylated genes were analyzed by sets and the top 10 genes were manually verified on the site. Moreover, in order to verify the veracity of results previously obtained a random set of genes were also manually tested.



A. Obtain all designations per gene

```
my.names <- function(genes_name){
##Genes_name: gene name separated by "|"
##Database with genes name
genesDF <- data.frame(t(data.frame(strsplit(genes_name, "[|]"))),
Gene = genes_name)
genesDF$ID <- paste("hsa", genesDF$X2, sep = "")
genesDF$ID_sep <- paste("hsa", genesDF$X2, sep = ":")
##Names for each gene
library(KEGGREST)
query <- sapply(genesDF$ID_sep, keggGet)
gene_ID <- list()
a <- length(query)
for (i in 1:a) {
gene_ID[[i]] <- query[[i]][["NAME"]]
}
for (i in 1:a) {
gene_ID[[i]] <- ifelse(typeof(gene_ID[[i]]) == "NULL",
as.character(genesDF$X1[i]), query[[i]][["NAME"]])
}
genes_ID_1 <- list()
my_DF = NULL
for(i in 1:a){
genes_ID_1[[i]] <- data.frame(strsplit(gene_ID[[i]], "[,]"))
genes_ID_1[[i]] <- as.character(genes_ID_1[[i]][,1])
genes_ID_1[[i]] <- data.frame(gene=genes_ID_1[[i]],
ID=rep(i))
c <- genes_ID_1[[i]]
my_DF <- rbind(my_DF, c)
} return(my_DF)}
```

B. Obtain total number of citations in cancer per gene

```
get_refs <- function(cancer,gene){
library(RISmed)
term<-paste(cancer, gene)
type <- "esearch"
db <- "pubmed"
datetype <- 'pdat'
mindate <- 1787
maxdate <- 2018
retmax <- 1000
refs<-EUtilsSummary(term, type=type, db=db,
datetype=datetype,
mindate=mindate, maxdate=maxdate, retmax=retmax)
pubmed.refs<-QueryCount(refs)
b<-EUtilsSummary(cancer,type=type, db=db,
datetype=datetype,
mindate=mindate, maxdate=maxdate,
retmax=retmax)@count
if(pubmed.refs == b)
return(0)
return(pubmed.refs)
}
```

Figure 3.8 - Oncosearch algorithm functions. (A) *my.names* function is a set of instructions to provide all gene designations per gene, retrieves all entries from the KEGG database **(B)** *get_refs* function provide a search in PubMed citations association gene to cancer or gene to specific cancer, from the year 1787 until 2018.

4 CHAPTER 4 – RESULTS

4.1. Working pipeline

4.1.1. Introduction

Creating a working algorithm that can be applied to large data cohort is a challenge. Considering the main goal that compares two general groups, to find significant differences in sets of variables, can originate potentials predictors of diagnostic and might be useful for disease treatment planning with faster and more reliable outputs. Here, we used DNA methylation and gene expression as examples of input data to validate our pipeline.

4.1.2. Construction of the working pipeline

This working pipeline was based on statistical procedures, from data extraction to final outputs, with optional and complementary components (***Figure 4.1***). The pipeline is structured into 6 well defined steps:

Phase 0: Optional step – Characterization of cohorts;
Phase 1: Differential step – Identify DMG;
Phase 2: Patterns step – Intersections across cohorts;
Phase 3: Predictors step – Identify predictors of diagnosis;
Phase 4: Linear models step – CpG probes with more impact in gene expression;
Phase 5: Representation step – Multivariate approach.

The optional step (***Phase 0***) is based on the cohort's characterization using clinical data available. Patients should be descriptively characterized with common clinical variables between cohorts (e.g. gender, age, histological type, therapies and classification scales) or other important variables for a specific cohort (e.g. estrogen receptor status, progesterone receptors status and HER_2 receptor status specific for breast cancer).

Differential step (***Phase 1***) pretends to identify differentially methylated genes that resulted from significant differences between groups. For each cohort (***i***) were exported from repository two databases, DNA methylation and gene expression, with a dichotomy variable that defined groups (***Input C***). Just patients with both data were included.

In the gene expression database, all samples were considered and genes without designation were excluded since they were not relevant in this study. In DNA methylation database all samples were considered with respective CpG sites and corresponded gene. Only CpG sites with designation and corresponding gene were established for differential analysis. Additionally, all outliers were removed according boxplot method. The overall result of is the ***Output 1***.

Cohort dimension were a control criterion for decide if the algorithm proceeded (minimum of 20 samples per group) meaning the cohort were included or excluded and followed by other.

After the initial data pre-processing, the inferential tests were split according to the conclusions of each level of decisions tests. Shapiro-Wilk test is the first decision level to verify the normality distribution of genes or CpG probes and in case of null hypothesis rejection, *p-value<0.05*, we followed a parametric approach in next level. In this case, was performed the t-tests according the assumptions of similar variances given (Students t-test) or unequal variances (Welch test). In particular, Levene test, were the equal population variances are tested, was the intermediate procedure before performing the t-test. Thus, the *p-value* results of t-tests decides if the two selected groups were substantially different, meaning there were statistically significant in differences of means between groups. The general layout of this sequence of tests provided us the genes differentially expressed and CpG probes differentially methylated. Also, an important control of *p-values* was, in each test, performed through a multiple comparison procedure (*FDR<0.05*) aimed to control the probability of committing any type I error in families of comparisons under simultaneous consideration. This short sequence of tests represented the core of the working pipeline.

Based on the previous runs tests we were able to establish the magnitude of the differentially expressed genes and differentially methylated CpG probes through two cut-offs, $|\log_2(\text{foldchange})| > 1.5$ for genes and $|\Delta\beta| > 0.2$ for CpG probes (***Output 2***).

This step was concluded with correlation analysis (Pearson correlation test) between gene expression and DNA methylation considered only for tumor samples. Statistical significant correlations were saved when *p-value<0.05*. This final procedure completes Phase 1 where were saved all considered differentially methylated genes (***Output 3***). Phase 1 algorithm has as many iterations as the initial n of cohorts, meaning the number of outputs is the same of the number of cohorts.

Patterns step (***Phase 2***) was based at intersections across cohorts that pretended find common genetic/epigenetic patterns with other cohorts or specifics per cohort. This step was only performed in the presence of more than one cohort. For ***Output 3*** were performed an intersection between genes or CpG probes, and we obtained in ***Output 4***. Then, the patterns across cohorts were analyzed and selected those with interest in the study context.

Predictors step (***Phase 3***) pretended identify epigenetic diagnostic biomarkers through the ***Output 3***. For each cohort, was performed a ROC curve analysis to identify genes and CpG probes with the power to predict diagnosis. For this, were establish three cut-offs: AUC > 0.8, sensibility>0.6 and specificity>0.6. Those potential Genes and CpG probes revealed good influence discrimination between groups (***Output 5***). This output may play an important clinical role.

Linear models step (***Phase 4***) for identify CpG probes with more impact in gene expression, that is, CpG probes which are the best linear predictors in gene expression. For each cohort, were performed an MLR model analysis using ***Output 3***, considering gene expression as the dependent variable and respective CpG probes as independent variables. CpG probes with *p-value<0.05* were considerate statistically significant, and genes without significant CpG probes were excluded (***Output 6***). This step was important to check the which CpG probes are really important in gene expression variation.

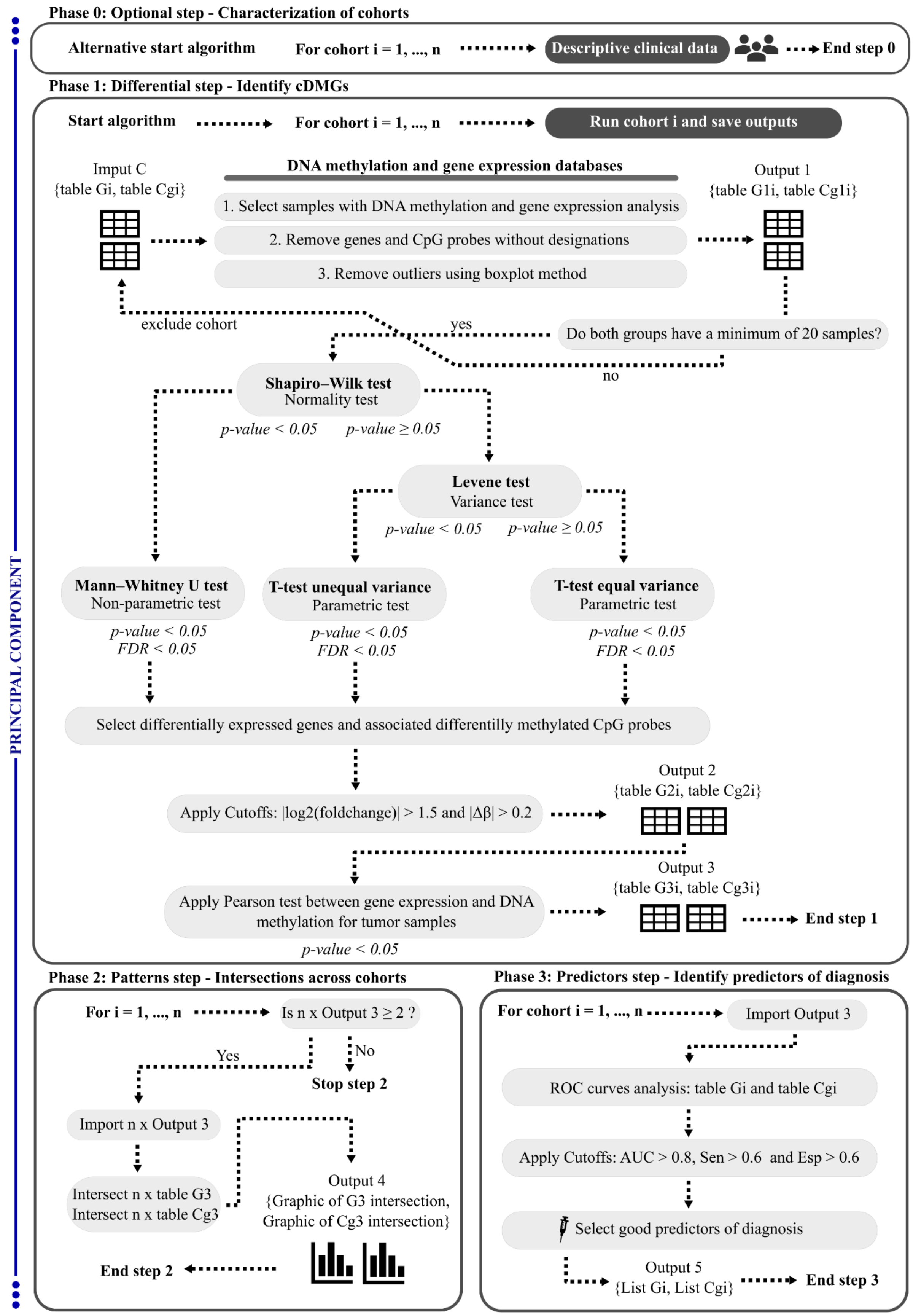
Representation step (***Phase 5***) based on the multivariate technique, the HJ-Biplot. For each cohort, using ***Output 3***. Gene expression or DNA methylation coordinates from HJ-Biplot were plotted in a principal plane (more accumulative variance, plan 1-2). Since the distribution of patients is influenced by gene expression or DNA methylation, the HJ-Biplot representation allowed better understanding of the impact of these variables in clustering behavior of samples. Additionally, to remove noise and select better explicability from genes and DNA methylation, were considered variables with contributions≥0.7. Hierarchical clustering of samples was performed based on HJ-Biplot coordinates, considering the ward method and the Euclidean distance. Results were presented in HJ-Biplot representation (***Output 7***).

Once the work pipeline was structured, we test the cohorts in *The Cancer Genome Atlas* (TCGA) database. 33 TCGA cohorts were available online and submitted to ***Phase 0*** and ***Phase 1***. Only 8 cohorts with solid tissue normal and primary tumor stage I samples (see in methods) results from previous phases, and there were:

Cohort 1 – Breast Invasive Carcinoma (TCGA-BRCA);
Cohort 2 – Colorectal Adenocarcinoma (TCGA-COADREAD);
Cohort 3 – Head and Neck Squamous Cell Carcinoma (TCGA-HNSC);
Cohort 4 – Kidney Renal Clear Cell Carcinoma (TCGA-KIRC);
Cohort 5 – Kidney Renal Papillary Cell Carcinoma (TCGA-KIRP);
Cohort 6 – Liver Hepatocellular Carcinoma (TCGA-LIHC);
Cohort 7 – Lung Adenocarcinoma (TCGA-LUAD);
Cohort 8 – Thyroid Carcinoma (TCGA-THCA)

This working pipeline section answers the first objective of this study and supports the others. The ***Phase 1*** answers the second aim: "Identify cancer differentially methylated genes (cDMGs) in early stages", ***Phase 2*** answers the third aim: "Identify patterns across cancers" and ***Phase 3*** answers the fourth and last aim: "Identify epigenetic biomarkers that predict diagnosis". Any other Phases were additional or complementary analysis that might be of interest to the researchers.

4.1.3. Working pipeline



PRINCIPAL COMPONENT
Phase 0: Optional step - Characterization of cohorts
Alternative start algorithm
For cohort i = 1, ..., n
Descriptive clinical data
End step 0
Phase 1: Differential step - Identify cDMGs
Start algorithm
For cohort i = 1, ..., n
Run cohort i and save outputs
Imput C
{table Gi, table Cgi}
DNA methylation and gene expression databases
1. Select samples with DNA methylation and gene expression analysis
2. Remove genes and CpG probes without designations
3. Remove outliers using boxplot method
Output 1
{table G1i, table Cg1i}
exclude cohort
yes
Do both groups have a minimum of 20 samples?
no
Shapiro–Wilk test
Normality test
p-value < 0.05
p-value ≥ 0.05
Levene test
Variance test
p-value < 0.05
p-value ≥ 0.05
Mann–Whitney U test
Non-parametric test
p-value < 0.05
FDR < 0.05
T-test unequal variance
Parametric test
p-value < 0.05
FDR < 0.05
T-test equal variance
Parametric test
p-value < 0.05
FDR < 0.05
Select differentially expressed genes and associated differentilly methylated CpG probes
Apply Cutoffs: |log2(foldchange)| > 1.5 and |Δβ| > 0.2
Output 2
{table G2i, table Cg2i}
Output 3
{table G3i, table Cg3i}
Apply Pearson test between gene expression and DNA methylation for tumor samples
p-value < 0.05
End step 1
Phase 2: Patterns step - Intersections across cohorts
For i = 1, ..., n
Is n x Output 3 ≥ 2 ?
Yes
No
Stop step 2
Import n x Output 3
Intersect n x table G3
Intersect n x table Cg3
Output 4
{Graphic of G3 intersection,
Graphic of Cg3 intersection}
End step 2
Phase 3: Predictors step - Identify predictors of diagnosis
For cohort i = 1, ..., n
Import Output 3
ROC curves analysis: table Gi and table Cgi
Apply Cutoffs: AUC > 0.8, Sen > 0.6 and Esp > 0.6
Select good predictors of diagnosis
Output 5
{List Gi, List Cgi}
End step 3

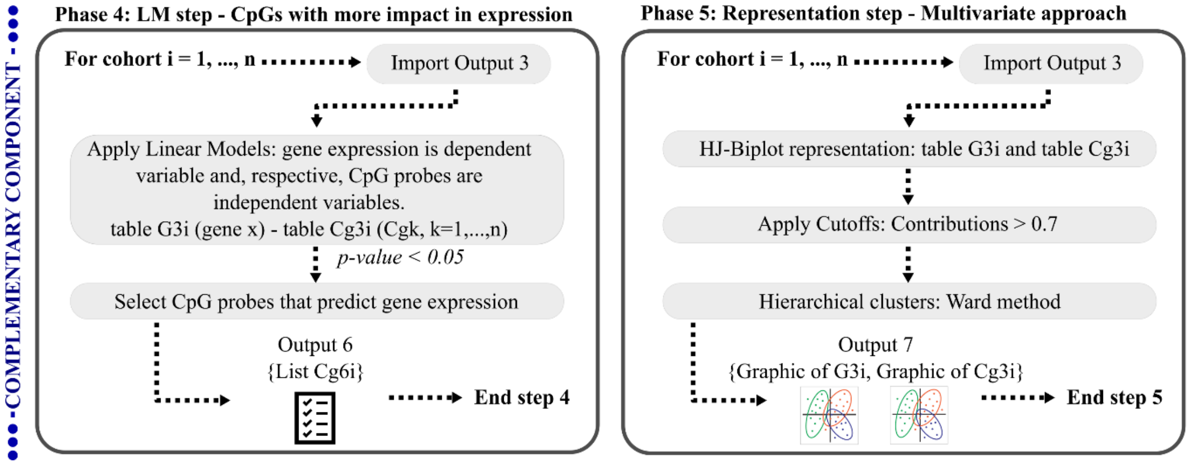

•••-COMPLEMENTARY COMPONENT -•••
Phase 4: LM step - CpGs with more impact in expression
For cohort i = 1, ..., n
Import Output 3
Apply Linear Models: gene expression is dependent variable and, respective, CpG probes are independent variables.
table G3i (gene x) - table Cg3i (Cgk, k=1,...,n)
p-value < 0.05
Select CpG probes that predict gene expression
Output 6
{List Cg6i}
End step 4
Phase 5: Representation step - Multivariate approach
For cohort i = 1, ..., n
Import Output 3
HJ-Biplot representation: table G3i and table Cg3i
Apply Cutoffs: Contributions > 0.7
Hierarchical clusters: Ward method
Output 7
{Graphic of G3i, Graphic of Cg3i}
End step 5

Figure 4.1 - Working pipeline. In **Principal component**, exists four Phases: *0. optional step*, *1. differential step*, *2. patterns step* and *3. predictors step*. Optional step (Phase 0) is based on the cohort's characterization using clinical data available. Differential step (Phase 1) pretends identify differentially methylated genes that presents significant differences between groups using gene expression and DNA methylation databases. After data preprocess, the normality distribution of genes and CpG probes were verified to be chosen in parametric or nonparametric way by Shapiro-Wilk test decision. Nonparametric way uses Mann-Whitney U test to check differences between group distributions. Before parametric way, similar variances were checked trough Levene test, and t-test were performed for equal or unequal variances (Welch test). These two ways were complemented with false positives correction procedure were FDR < 0.05. Then, were established two cut-offs, |log2(Foldchange)|>1.5 and |Δβ|>0.2. Pearson coefficient was used to verify correlations between genes and CpG probes in tumor group and select differentially methylated genes. This Phase 1 iterates per cohort and save the respective outputs. Two steps followed, using the Output 3, were designated patterns-step (Phase 2) and predictors-step (Phase 3). Phase 2 is based on intersection of genes or CpG across cohorts resulting in common patterns or specifics in each one. Predictors-Phase identify genes and CpG probes as good predictors of diagnosis based in ROC curve analysis, considering AUC > 0.8, sensibility >0.6 and specificity >0.6. In **Complementary component**, exist two Phases: *4. Linear models step* and *5. Representation step*. Linear models step (Phase 4) pretend to obtain CpG probes that has good predictors of gene expression trough MLR (*p-value* < *0.05*). Representation step (Phase 5) is based on HJ-Biplot representation of gene expression or DNA methylation were samples are distributed in a maximum retain variance scenario (plan). Samples were grouped trough hierarchical clustering, using Ward method and Euclidian distance.

4.2. Identification of cancer differentially methylated genes (cDMGs) – Phase I

4.2.1. Introduction

Initially, we performed ***Phase 1*** of the WP to identify cancer differentially methylated genes. Early stages analysis presupposes the existence of stage I primary tumor and solid tissue normal samples. All Genes and CpG sites probes for whole-genome were considered. A selecting procedure were computed for each type of cancer, and saved genes differentially expressed associated with differentially methylated CpG probes were saved.

4.2.2. Identification of cDMGs per cohort

We extracted expression for 20531 genes and methylation of 485577 CpG probes through pre-processing, the initial procedure of Phase 1. Missing gene expression values and nonidentity gene were excluded (***Figure 4.2A***). 20502 genes and 364643 CpG probes became available for analysis (***Figure 4.2B***). All 8 cohorts were submitted into ***Phase I*** to evaluate the differences between normal and tumor groups. Results reveals gene sets and respective CpG probes across for each type of cancer (***Figure 4.2C, Appendix 1***). Colorectal cancer has 307 cDMGs associated with 924 CpG probes, breast cancer has 117 with 368, head and neck cancer have 99 with 292, kidney$_R$ cancer has 156 with 299, kidney$_P$ has 106 with 224, liver cancer has 349 with 1453, lung cancer has 180 with 601 and thyroid cancer 25 with 40, respectively. Colorectal and Liver cancers were the cancers with more genes and CpG probes and thyroid cancer with fewer alterations.

We checked how many CpG probes regulated negatively and positively the up-regulated and down-regulated genes (***Figure 4.2D***). As a result, down-regulated genes were associated with hypermethylated and hypomethylated CpG probes for all cancers: head and neck cancer present 148 and 70, lung cancer present 74 and 61, liver cancer present 31 and 691, colorectal present 451 and 245, thyroid cancer present 2 and 3, breast cancer present 74 and 107, kidney$_R$ present 157 and 69 and kidney$_P$ present 112 and 39, respectively. Also, up-regulated genes were associated with hypermethylated and hypomethylated CpG probes: head and neck cancer present 65 and 9, lung cancer present 429 and 37, liver cancer present 113 and 618, colorectal present 146 and 82, thyroid cancer present 1 and 34, breast cancer present 123 and 64, kidney$_R$ present 15 and 58 and kidney$_P$ present 50 and 23, respectively.

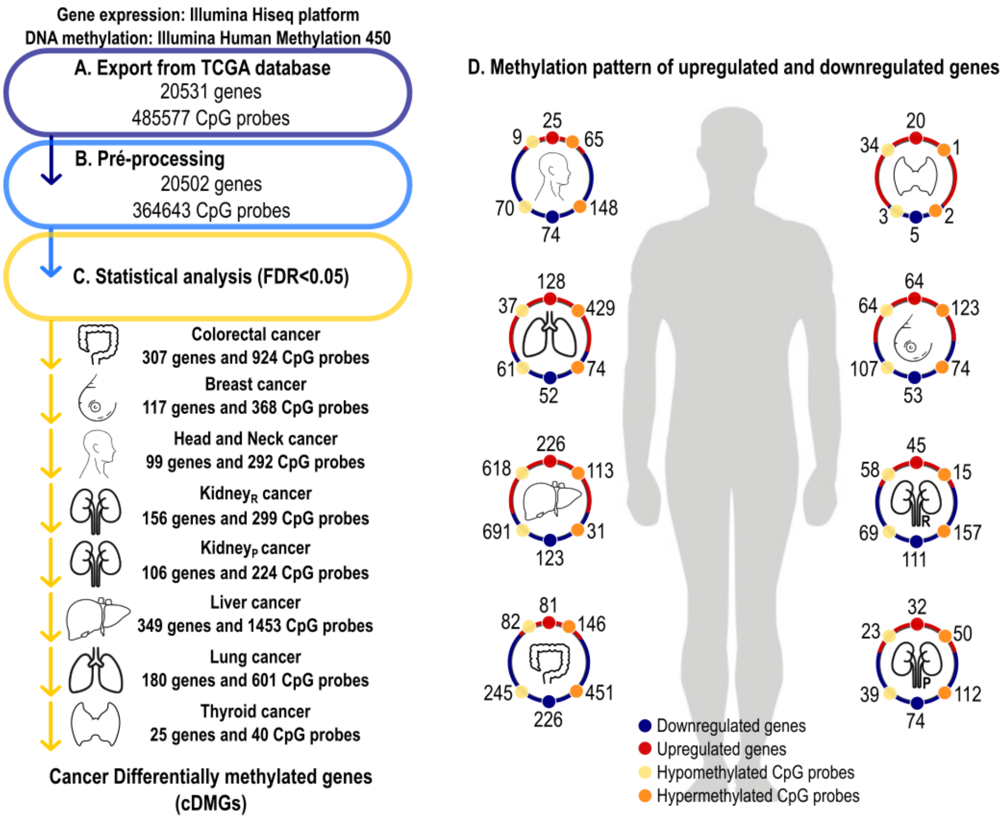

Gene expression: Illumina Hiseq platform
DNA methylation: Illumina Human Methylation 450
A. Export from TCGA database
20531 genes
485577 CpG probes
B. Pré-processing
20502 genes
364643 CpG probes
C. Statistical analysis (FDR<0.05)
Colorectal cancer
307 genes and 924 CpG probes
Breast cancer
117 genes and 368 CpG probes
Head and Neck cancer
99 genes and 292 CpG probes
Kidney$_R$ cancer
156 genes and 299 CpG probes
Kidney$_P$ cancer
106 genes and 224 CpG probes
Liver cancer
349 genes and 1453 CpG probes
Lung cancer
180 genes and 601 CpG probes
Thyroid cancer
25 genes and 40 CpG probes
Cancer Differentially methylated genes
(cDMGs)
D. Methylation pattern of upregulated and downregulated genes
25
9
65
70
148
74
20
34
1
3
2
5
128
37
429
61
74
52
64
64
123
107
74
53
226
618
113
691
31
123
45
58
15
69
157
111
R
81
82
146
245
451
226
32
23
50
39
112
74
P
● Downregulated genes
● Upregulated genes
● Hypomethylated CpG probes
● Hypermethylated CpG probes

Figure 4.2 – Cancer differentially methylated genes (cDMGs). (A) Number of genes and CpG probes exported from The Cancer Genome Atlas (TCGA) database. Gene expression was measured using *Illumina Hiseq* platform and DNA methylation by *Illumina HumanMethylation 450*. **(B)** Number of genes and CpG probes resulting from pre-processing of data. Missing values, non-identified genes and CpG probes, and CpG probes without matching gene was excluded. **(C)** Inferential statistical analysis applied for all cohorts with FDR < 0.05. **(D)** Pattern of upregulated and downregulated genes for all cohorts, and frequencies of hypermethylated and hypomethylated considering gene pattern type.

This Phase allowed ranked, trough up-regulated and down-regulated classification, a top five list of genes for each cancer cohort (***Table 4.1***). Starting with cancers that have more up-regulated genes which are: breast, liver, lung and thyroid, and those that had more down-regulated genes: colorectal, head and neck, kidney$_R$ and kidney$_P$ cancers. CpG probes were divided by the up or down regulated gene and classified in hyper- methylated and hypo-methylated (***Figure 4.2D***). Despite thyroid cancer in spite of presents few alterations, the majority of upregulated genes are associated to hypomethylated probes. Most cancers present more CpG probes hypermethylated in early stages, except for lung and thyroid cancers.

Table 4.1 – Top five genes for mostly up-regulated or down-regulated cohorts.

Cohort	Up-regulated genes	Log$_2$(foldchange)	Cohort	Down-regulated genes	Log$_2$(foldchange)
Breast cancer	*TLX1NB*	5.246988909	**Colorectal cancer**	*CACNG5*	-5.052460275
	METTL11B	4.624172361		*GABRG1*	-4.677518094
	APOBEC1	3.713445213		*HTR3B*	-4.59866826
	EFNA2	3.68424939		*CLVS2*	-4.216960803
	SP8	3.609379941		*MCHR2*	-4.005016352
Liver cancer	*CTAG2*	5.999762549	**Head and neck cancer**	*NKAIN3*	-5.189842923
	NAA11	5.824472442		*PRAMEF12*	-4.87102028
	REG1B	5.720864063		*ACCSL*	-4.465862478
	COX7B2	5.533620437		*PROKR2*	-3.999350407
	CDH9	5.501310695		*CSN2*	-3.872931671
Lung cancer	*TFAP2D*	6.177427934	**Kidney$_R$ cancer**	*DEFB132*	-4.666355345
	PITX2	5.551557746		*ROS1*	-3.616624535
	SP8	5.35736266		*VGLL1*	-3.550205623
	HOTAIR	5.346610236		*OLFM3*	-3.429478559
	FOXI3	5.256633941		*FXYD4*	-3.323236545
Thyroid cancer	*PLA2G2E*	7.060822742	**Kidney$_P$ cancer**	*CGA*	-5.499764391
	MS4A15	4.458326797		*CPNE6*	-4.838612363
	CSF2	3.55921961		*C16orf11*	-4.007789014
	AWAT2	3.513327584		*AMELY*	-3.976773857
	RNASE11	3.444127322		*BSND*	-3.896258476

4.2.3. CpG probes localization

Modified CpG probes for each cohort were distributed considering their location in 5 different sites: 5'UTR, TSS1500, TSS200, 1st Exon, Body and 3'UTR. In ***Figure 4.3*** are represented frequency and proportion of CpG probes per site, considering only CpG probes with a single localization (***Appendix 3***). Mostly CpG probes are located in gene body, except in thyroid cancer, that there was a large proportion of CpG sites located on transcription start sites.

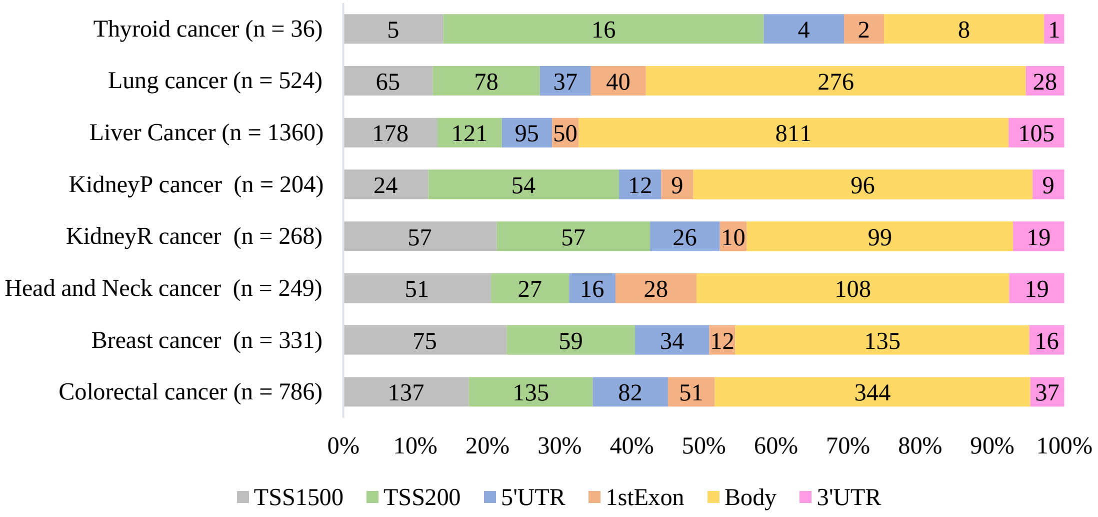


Figure 4.3 – CpG sites localization for each type of cancer. CpG probes can be in 5 main sites: TSS1500, TSS200, 5'UTR, 1stExon, Body, 3'UTR. TSS1500 - 1500bp upstream to Transcription Start Site (grey); TSS200 – 200bp upstream to Transcription Start Site (green); 5'UTR - Five prime untranslated region (blue); 1stExon – First exon of gene (orange); Body – Gene body (yellow); 3'UTR - Three prime untranslated region (pink). Only probes with single localization were represented.

4.2.4. Gene-annotation enrichment analysis

To evaluate the Gene Ontology (GO) of cDMGs by each type of cancer an enrichment analysis was performed using DAVID (https://david.ncifcrf.gov/), version 6.8, available *online*. Use default fields except that we defined as *p-value<0.05*. Rich-factor is the ratio of the number of cDMGs mapped to this GO term with total annotated in this term (Wang et al. 2016).

Results reveals the existence of some very significant gene terms in development, synaptic transmission, signaling and transport (***Figure 4.4, Appendix 4***). Breast cancer reveals very enrichment genes in the dorsal/ventral formation pattern (*p-value=0.0005* and rich-factor=0.125). Colorectal cancer reveals very enrichment genes in the chemical synaptic transmission (*p-value=1.309E-10*). Head and neck cancer is highly enriched in ionotropic glutamate receptor signaling (*p-value=0.0002* and rich-factor=0.167). Kidney$_R$ cancer is highly enriched in genes associated with excretion (*p-value = 0.002* and rich-factor=0.108). Kidney$_P$ cancer is more enriched in genes associated with kidney development (*p-value=0.009* and rich-factor=0.047). Liver cancer is enriched in visual perception (*p-value=0.0001* and rich-factor=0.137) and neurotransmitter secretion (*p-value=0.0002* and rich-factor=0.137). Insight, that thyroid cancer due to its number of genes did not present results.

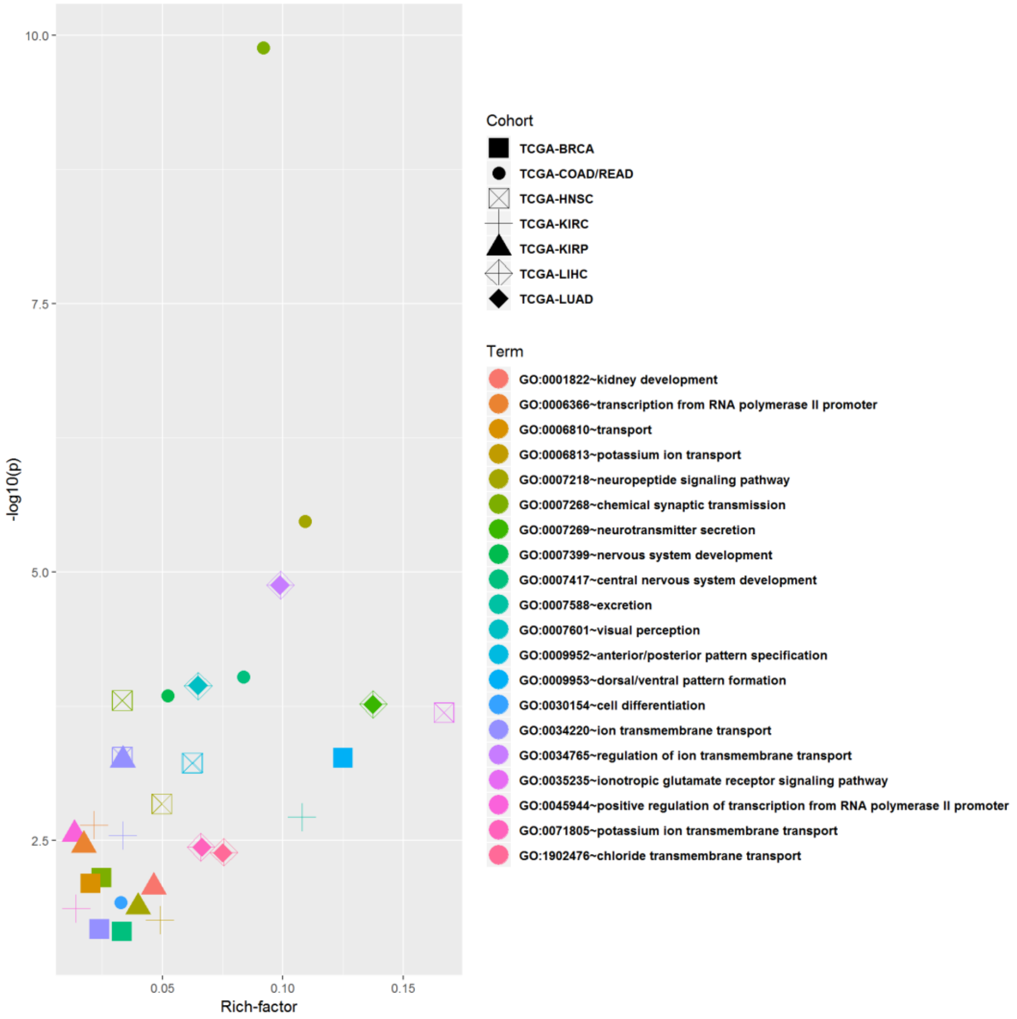


Figure 4.4 – Cancer-associated differentially methylated genes (cDMGs) identified in 7 cohorts. Scatterplot for statistics of biological enrichment. For all gene lists, we listed the top five enriched biological processes. Rich-factor is the ratio of the number of cDMGs mapped to this GO term with total annotated in this term, From: Wang et al. 2016. The higher rich factor means is the more significant enrichment. The higher -log10(p) also means the more significant enrichment, where p is the p-value for GO term. TCGA-BRCA – Breast Invasive Carcinoma; TCGA-COADREAD – Colorectal Adenocarcinoma; TCGA-HNSC – Head and Neck Squamous Cell Carcinoma; TCGA-KIRC – Kidney Renal Clear Cell Carcinoma; TCGA-KIRP – Kidney Renal Papillary Cell Carcinoma; TCGA-LIHC – Liver Hepatocellular Carcinoma. TCGA-THCA – Thyroid Carcinoma due to its reduced number of genes does not present results.

4.2.5. Analysis of annotation in the literature

Differential methylated genes in early stages were evaluated for their annotation in the literature with the algorithm called "*Oncosearch*" (see in the methods). The presence of cDMGs associated with the respective cancers gives consistency to the analysis and, however, not reported annotation genes might suggest clinical potential in diagnosis.

Associations between genes with cancer for each cohort gave us genes that are mostly reported in cancer type (***Appendix 5***). We verified that only a small set gene of was linked in cancer. Similar proportions in ***Figure 4.5***, tells us that approximately 13% weren´t reported in cancer. We also can see some reported genes in cancer in general: *COL9A1*, *CD5L*, *SLC4A1*, *CALCA*, *PROC*, *TNFSF14*, *SLC6A2*, *TMEFF2*, *CSF2* and *SPINK5*, related to body metabolism and homeostasis. Within genes never mentioned in cancer some of them are the following: *KCNJ9*, *GLB1L3*, *SLC27A6*, *SLC38A8*, *FLJ12825*, *HIST1H4E*, *FRMPD4*, *DCDC2B*, *DMRTA2*, *NXF5*, associated with distinct function in human body.

In cancer reported genes, approximately 87%, we had two scenarios (***Figure 4.5***): firstly, the cancers like thyroid, liver, kidney, and head and neck that had more genes not referenced in specific cancers, and, secondly, the cancers like: lung, breast and colorectal that had more referenced genes in specific cancers.

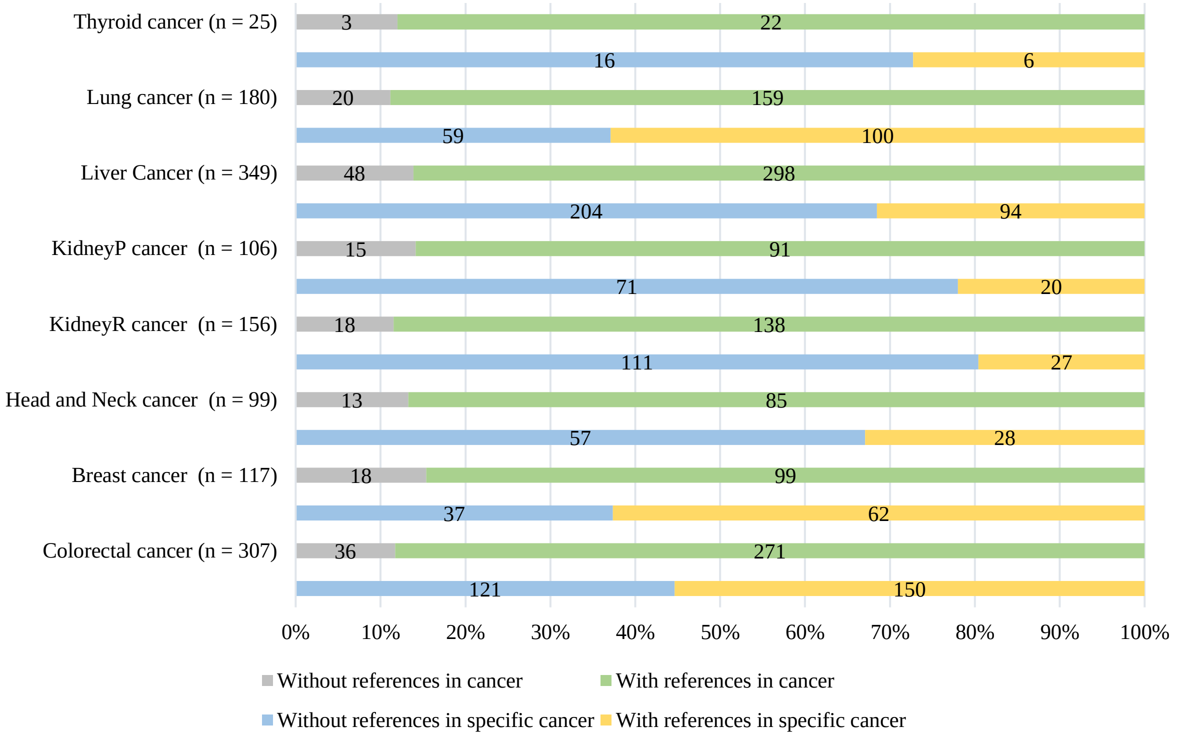


Figure 4.5 – Annotation analysis for cDMGs in literature using Oncosearch Algorithm. Association between cDMGs and cancer or specific cancer were performed by *Oncosearch* algorithm. *Oncosearch* algorithm use PubMed repository to search query citations. Grey represents genes without cancer references and green with references in cancer. Genes associated with cancer but not reported in specific cohorts were represented in blue. In yellow were genes that associated with specific cohort.

4.3. Specific methylation patterns across different cancers – Phase 2

4.3.1. Introduction

Genetic and epigenetic patterns were used to explore behavior among different types of cancer, aimed to find common or unique patterns according to the ***Phase 2*** of the WP. Completed this analysis, we focused on common biological patterns that allowed us to characterize basic tumorigenesis, or specific pattern to create an interest pattern suported on new potential diagnosis biomarker.

4.3.2. Crossing different cancers

Identification of genetic and epigenetic patterns occurred by crossing the sets of genes and CpG probes of each type of cancer. The intersections trough all cohorts results in combinations of common or unique patterns that might characterize the tumor development. The results reveal the existence of a set of common genes and several specific genes for each type of cancer (***Figure 4.6A***). Those specific genes were distributed by cohort: 18 in thyroid cancer, 49 in head and neck, 70 in kidney$_P$, 55 in breast cancer, 100 in kidney$_R$, 97 in lung cancer, 202 in colorectal and 240 in liver cancer (***Appendix 2***). Despite the high complexity of patterns, we saw that most behavior is characterized by particular genes. However, in adenocarcinomas, colorectal and lung cancers, presented 21 common genes. Kidney$_R$ and Kidney$_P$ cancers presented 13 common genes.

CpG probes have the same pattern results as the genes (***Figure 4.6B***). Those specific CpG probes were distributed by cohort: thyroid cancer with 35, head and neck cancer with 223, kidney$_P$ cancer with 189, breast cancer with 261, kidney$_R$ cancer with 244, lung cancer with 449, colorectal cancer with 782 and liver cancer with 1339.

In summary, the results reveal a specific pattern of cDMGs for each early stage cancer that showed to be unique in cancer development.

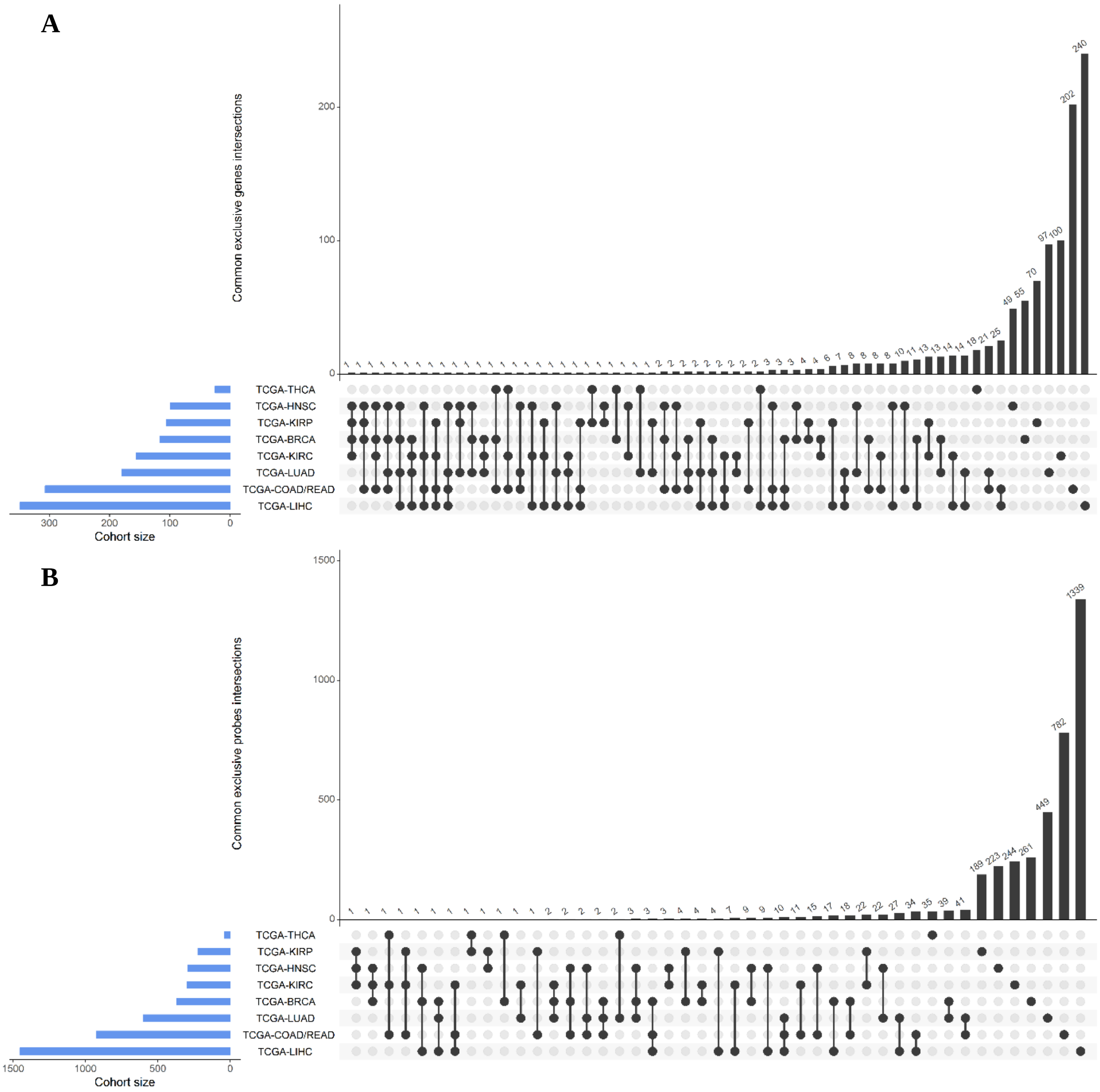


Figure 4.6 – Crossing different cancers. (A) represent the cDMGs patterns across cancers. **(B)** represented the CpG probes patterns across cancers. Blue represents the size cohort. The intersection of cDMGs is represented in graph. The single points represented the genes or CpG probes that are unique in each cancer. TCGA-BRCA – Breast Invasive Carcinoma; TCGA-COAD/READ – Colorectal Adenocarcinoma; TCGA-HNSC – Head and Neck Squamous Cell Carcinoma; TCGA-THCA – Thyroid Carcinoma; TCGA-KIRC – Kidney Renal Clear Cell Carcinoma; TCGA-KIRP – Kidney Renal Papillary Cell Carcinoma; TCGA-LIHC – Liver Hepatocellular Carcinoma.

4.3.3. Pathways characterization of the specific cDMGs

After identifying specific gene patterns for each type of cancer, we performed an enrichment analysis of pathways using the *Reactome Pathways Database*, helped with *DAVID* tool previously used. The default criteria were used and *p-value<0.05*.

Pathways results analysis reveal some enrichment pathways, the most significant being represented in ***Table 4.2***. It was important described which cancers presented significant enrichment pathways (***Appendix 6***).

For early stages, specific changes in colorectal cancer were strongly associated with RAF/MAP kinase cascade (*p-value = 8.01E-05*) through the genes: *FGF19*, *FGF8*, *GRIN2A*, *IL5RA*, *FGF20*, *NEFL*, *FGF3* and *GFRA3*; PI3K Cascade (*p-value = 0.004096*) through the genes: *FGF19*, *FGF8*, *FGF20* and *FGF3*; and L1CAM interactions (*p-value = 0.004208*) through the genes: *CNTN2*, *CNTN1* and *NCAN*.

Head and neck cancer is associated with GABA A receptor activation (*p-value = 0.026896*) and Neurotransmitter receptors and postsynaptic signal transmission (*p-value = 0.028936*) through the genes: *GABRG3* and *GABRB1*; Negative regulation of TCF-dependent signaling by WNT ligand antagonists (*p-value = 0.030972*) through the genes: *SOST* and *WIF1*; and G alpha (q) signaling events (*p-value = 0.045894*) through the genes: *NPFFR2, PROKR2* and *TRH*.

Lung cancer, the most significant pathways were Peptide ligand-binding receptors (*p-value = 0.006942*) through the genes: *SSTR4*, *EDN3*, *NPBWR1* and *OPRD1*; and RNA Polymerase I Promoter Opening (*p-value = 0.023072*), DNA methylation (*p-value = 0.024459*), Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3 (*p-value = 0.025879*) all through the genes: *HIST1H2BI, HIST1H3B* and *HIST1H4E*.

Kidney$_P$ cancer is associated with Ion homeostasis (*p-value = 0.015509*) through the genes: *FXYD3*, *CAMK2A* and *CASQ2*.

Liver cancer is associated with: Collagen biosynthesis and modifying enzymes (*p-value = 0.031157*) through the genes: *COL6A6*, *COL25A1*, *COL2A1* and *COL24A1*; Glucagon-type ligand receptors (*p-value = 0.037531*) through the genes: *GLP2R*, *GLP1R* and *GHRHR*; Adheres junctions interactions (*p-value = 0.042236*) through the genes: *CDH8*, *CDH9* and *CDH10*; and Signaling by PDGF (*p-value = 0.042236*) through the genes: *COL6A6*, *COL2A1* and *THBS4*. Importantly, breast, kidney$_R$ and thyroid cancers had no significant results.

Table 4.2 – Pathways enrichment analysis for specific genes for each type of cancer.

Cohort	Term	Count	*p-value*	Genes
Colorectal cancer	R-HSA-5673001: RAF/MAP kinase cascade	8	8.01E-05	FGF19, FGF8, GRIN2A, IL5RA, FGF20, NEFL, FGF3, GFRA3
	R-HSA-109704: PI3K Cascade	4	0.004096	FGF19, FGF8, FGF20, FGF3
	R-HSA-373760: L1CAM interactions	3	0.004208	CNTN2, CNTN1, NCAN
Head and Neck cancer	R-HSA-977441: GABA A receptor activation	2	0.026896	GABRG3, GABRB1
	R-HSA-112314: Neurotransmitter receptors and postsynaptic signal transmission	2	0.028936	GABRG3, GABRB1
	R-HSA-3772470: Negative regulation of TCF-dependent signaling by WNT ligand antagonists	2	0.030972	SOST, WIF1
	R-HSA-416476: G alpha (q) signalling events	3	0.045894	NPFFR2, PROKR2, TRH
Lung cancer	R-HSA-375276: Peptide ligand-binding receptors	4	0.006942	SSTR4, EDN3, NPBWR1, OPRD1
	R-HSA-73728: RNA Polymerase I Promoter Opening	3	0.023072	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-5334118: DNA methylation	3	0.024459	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-5625886: Activated PKN1 stimulates transcription of AR (androgen receptor) regulated genes KLK2 and KLK3	3	0.025879	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-427359: SIRT1 negatively regulates rRNA expression	3	0.026601	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-212300: PRC2 methylates histones and DNA	3	0.030335	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-2299718: Condensation of Prophase Chromosomes	3	0.031105	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-912446: Meiotic recombination	3	0.041806	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-201722: Formation of the beta-catenin:TCF transactivating complex	3	0.042679	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-73777: RNA Polymerase I Chain Elongation	3	0.044447	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-5250924: B-WICH complex positively regulates rRNA expression	3	0.045341	HIST1H2BI, HIST1H3B, HIST1H4E
	R-HSA-3214815: HDACs deacetylate histones	3	0.048063	HIST1H2BI, HIST1H3B, HIST1H4E
Kidney$_P$ cancer	R-HSA-5578775: Ion homeostasis	3	0.015509	FXYD3, CAMK2A, CASQ2
Liver cancer	R-HSA-1650814: Collagen biosynthesis and modifying enzymes	4	0.031157	COL6A6, COL25A1, COL2A1, COL24A1
	R-HSA-420092: Glucagon-type ligand receptors	3	0.037531	GLP2R, GLP1R, GHRHR
	R-HSA-418990: Adherens junctions interactions	3	0.042236	CDH8, CDH9, CDH10
	R-HSA-186797: Signaling by PDGF	3	0.042236	COL6A6, COL2A1, THBS4

4.4. Identification of biomarkers with clinical application – Phase 3

4.4.1. Introduction

Diagnostic biomarkers are essential for identify early stages in cancer. Find new biomarkers that can discriminate cancer outcome with high levels of sensitivity and specificity, become a truly powerful tool for clinicians in order to predict or adjust better diagnosis. ***Phase 3*** of the working pipeline is to performer those analyses.

4.4.2. cDMGs predict diagnostic of patients with cancer in early stages

ROC curve is the most common analysis for study evaluation in order to identify cDMGs as good predictors of cancer diagnostic. We intend to identify biomarkers that have the capacity to discriminate normal from tumor tissue for each type of cancer, and eventually distinguish cancers (***Appendix 7***).

Breast cancer has 45 genes and 340 CpG probes as good diagnosis predictors, but only 165 CpG probes were corresponded with the 45 genes. However, looking at the specific patterns, just 19 genes with 44 CpG sites revealed as potential diagnostic predictors for breast cancer.

Colorectal cancer has 238 genes and 835 CpG probes as good diagnosis predictors, but only 673 CpG probes corresponded with the 238 genes. However, looking at the specific patterns, just 153 genes with 461 CpG sites revealed as potential diagnostic predictors for colorectal cancer.

Head and neck cancer has 57 genes and 286 CpG probes as good diagnosis predictors, but only 156 CpG probes corresponded with the 57 genes. However, looking at the specific patterns, just 27 genes with 68 CpG sites revealed as potential diagnostic predictors for head and neck cancer.

Kidney$_R$ cancer has 142 genes and 299 CpG probes as good diagnosis predictors, but only 271 CpG probes corresponded with the 142 genes. However, looking at the specific patterns, just 93 genes with 173 CpG sites revealed as potential diagnostic predictors for kidney$_R$ cancer.

Kidney$_P$ cancer has 88 genes and 200 CpG probes as good diagnosis predictors, but only 181 CpG probes corresponded with the 88 genes. However, looking at the specific patterns, just 53 genes with 111 CpG sites revealed as potential diagnostic predictors for kidney$_P$ cancer.

Liver cancer has 126 genes and 1129 CpG probes as good diagnosis predictors, but only 619 CpG probes corresponded with the 126 genes. However, looking at the specific patterns, just 72 genes with 261 CpG sites revealed as potential diagnostic predictors for liver cancer.

Lung cancer has 88 genes and 595 CpG probes as good diagnosis predictors, but only 280 CpG probes corresponded with the 88 genes. However, looking at the specific patterns, just 38 genes with 128 CpG sites were revealed as potential diagnostic predictors for lung cancer.

Thyroid cancer has 18 genes and 38 CpG probes as good diagnosis predictors, but only 29 CpG probes corresponded with the 18 genes. However, looking at the specific patterns, just 14 genes with 24 CpG sites revealed as potential diagnostic predictors for thyroid cancer.

In summary, our results revealed that not all genes and CpG probes are good predictors of diagnostic, but a Specific pattern analysis revealed some sets of predictors in diagnostic for specific tumor types.

4.5. Complementary component

4.5.1. Introduction

A complementary component of WP intended to analyze the relations between genes and CpG probes wondering which methylation of CpG probes is effectively important in gene expression, using multiple linear regression analysis (***Phase 4***). Also, a multivariate approach, HJ-Biplot, where samples and genes (or CpG probes) are simultaneously presented in graphic layout, were useful for identifying cluster behaviors considering the most important genes or CpG probes in samples distribution (***Phase 5***).

4.5.2. CpG sites with more importance per gene – Phase 4

The followed analysis aimed to identify, which CpG probes were statically significant in gene variations that might be in origin of normal-tumor transition.

Results showed that, within cDMGs with more than one differentially methylated CpG probe, there existed methylation positions which play a more significant role in order to explain gene expression (***Appendix 8***). Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 71 CpG probes (57%).

Breast cancer has 125 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17%

TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to the *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 71 CpG probes (57%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 71 CpG probes corresponds to 63 genes that: 28 has CpG sites with single localization belong to the specific pattern for breast cancer. The top 20 most significant CpG probes are represented in ***Table 4.3***.

Table 4.3 – Top 20 of most significant CpG probes selected for breast cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg01747222	CMTM5	TSS200	0.262727	1.28E-33
cg15615793	CRHR2	TSS1500	0.24681	1.40E-15
cg05426601	CPA1	TSS1500	0.200353	2.86E-15
cg13696490	LOC201651	TSS1500	0.261073	3.65E-14
cg00635343	LOC642597	Body	-0.2126	3.97E-13
cg07504127	CLDN25	TSS1500	-0.23736	6.51E-12
cg10316270	RBM46	TSS200	-0.21748	1.17E-11
cg03543319	PROKR1	Body	-0.23093	1.70E-11
cg16310003	HPD	TSS1500	0.249793	8.51E-11
cg20742415	METTL11B	TSS1500	-0.26241	1.99E-10
cg17085688	GNGT1	Body	0.208609	2.47E-10
cg26925231	SGCZ	Body	0.273407	4.65E-10
cg01479664	TEPP	Body	-0.23518	5.58E-10
cg11473616	CYP1A2	TSS200	-0.22304	1.80E-09
cg01454519	CST5	1stExon	-0.21741	3.32E-09
cg01595325	HS3ST4	Body	-0.22664	8.13E-09
cg04561937	GAB4	3'UTR	-0.20994	8.34E-08
cg21977377	NKX2-2	Body	0.26497	1.10E-07
cg10322419	LOC284661	TSS1500	-0.26566	1.26E-07
cg07197831	DNAJC5G	3'UTR	-0.24136	2.90E-07

Colorectal cancer has 302 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 144 CpG probes (48%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 144 CpG probes corresponds to 121 genes that: 80 has CpG sites with single localization belong to the specific pattern for colorectal cancer. The top 20 most significant CpG probes are represented in ***Table 4.4***.

Table 4.4 – Top 20 of most significant CpG probes selected for colorectal cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg14075424	NANOS3	1stExon	-0.20987	3.13E-24
cg01618102	BAI3	Body	-0.49695	4.79E-23
cg20950932	GRIA3	TSS1500	0.315462	1.18E-16
cg03609960	ANKS1B	Body	0.387767	2.10E-16
cg00618450	SEZ6L	Body	0.391342	5.22E-16
cg16415058	SORCS1	1stExon	0.278692	1.04E-15
cg10735632	C2orf40	TSS1500	0.284777	1.44E-15
cg26165108	ENPP6	Body	-0.36758	3.13E-15
cg04306063	CRHBP	Body	0.258118	1.32E-14
cg20752831	RIMS4	3'UTR	-0.20803	1.50E-14
cg12232463	LONRF2	Body	-0.39412	3.65E-14
cg08104310	ASTN1	3'UTR	-0.31812	1.17E-13
cg01162507	GP2	3'UTR	-0.27092	2.64E-13
cg16080876	NEFL	1stExon	0.30605	3.03E-13
cg13206017	SST	TSS200	0.330157	9.84E-13
cg04678336	SGCG	Body	-0.25889	1.76E-12
cg01201932	CMA1	TSS1500	-0.35601	9.19E-12
cg02739437	SPOCK3	TSS1500	-0.28411	1.79E-11
cg20944283	CADM3	3'UTR	-0.32134	2.61E-11
cg09442828	ADRB3	1stExon	0.24909	3.29E-11

Head and neck cancer has 71 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 32 CpG probes (45%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 32 CpG probes corresponds to 31 genes that: 15 has CpG sites with single localization belong to the specific pattern for head and neck cancer. The 15 most significant CpG probes are represented in ***Table 4.5***.

Table 4.5 – Top 15 of most significant CpG probes selected for head and neck cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg16001323	ADH1B	3'UTR	-0.22058	2.85E-10
cg03222834	SHISA9	Body	-0.22171	3.95E-10
cg16638385	SOST	Body	0.364031	3.87E-08
cg02598319	C6	Body	-0.274	1.51E-07
cg11027140	GPR144	TSS1500	0.229374	1.52E-07
cg21171320	TRPM3	Body	-0.32901	2.66E-07
cg13672800	C20orf141	TSS1500	-0.25299	1.20E-06
cg18950108	DPCR1	Body	-0.24163	1.83E-06
cg06418867	C10orf90	Body	-0.2098	6.77E-06
cg24566400	RBP4	TSS1500	0.223605	1.41E-05
cg19112977	MGC16121	Body	0.317332	2.21E-05
cg20646280	KCTD8	1stExon	0.2793	2.31E-05
cg11213520	LHX5	Body	0.354345	5.92E-05
cg02227188	HOXC9	Body	0.366655	0.000153
cg24157814	DCT	Body	-0.20058	0.000312

Kidney$_R$ cancer has 202 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 146 CpG probes (72%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 146 CpG probes corresponds to 119 genes that: 87 has CpG sites with single localization belong to the specific pattern for kidney$_R$ cancer. The top 20 most significant CpG probes are represented in ***Table 4.6***.

Table 4.6 – Top 20 of most significant CpG probes selected for Kidney$_R$ cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg20610181	CA9	1stExon	-0.3922	4.58E-59
cg04511534	GGT6	Body	0.39297	4.89E-52
cg18390495	DEFB132	Body	-0.26447	3.76E-44
cg03762549	TMEM61	Body	0.239261	2.73E-42
cg13540480	C14orf50	TSS1500	0.283084	7.32E-41
cg07795964	NAT8L	3'UTR	0.253435	3.73E-35
cg08842032	EPN3	5'UTR	0.355429	4.41E-31
cg14895298	BIRC7	TSS200	-0.24693	8.66E-31
cg00373436	TAGLN3	Body	-0.3479	2.32E-29
cg21504505	C4orf6	3'UTR	-0.24695	5.47E-28
cg14021961	CLCNKA	5'UTR	0.223245	1.37E-24
cg10958362	C5orf38	Body	-0.21964	9.74E-23
cg13929970	FGFBP1	5'UTR	0.236362	7.41E-22
cg08568550	C11orf16	TSS200	0.23572	4.98E-21
cg10896586	GPR110	5'UTR	0.263813	5.03E-21
cg03211864	BTBD16	Body	-0.23755	7.16E-21
cg16804165	CKMT1A	Body	-0.26004	1.92E-20
cg10362335	TNFSF14	5'UTR	-0.20133	5.52E-20
cg25649889	MYO3B	Body	0.31116	9.35E-20
cg18565355	ESRP1	Body	0.280886	2.97E-19

Kidney$_P$ cancer has 119 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 91 CpG probes (76%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 91 CpG probes corresponds to 76 genes that: 55 has CpG sites with single localization belong to the specific pattern for kidney$_P$ cancer. The top 20 most significant are represented in ***Table 4.7***.

Table 4.7 – Top 20 of most significant CpG probes selected for Kidney$_P$ cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg04389897	TFAP2A	3'UTR	-0.23457	1.22E-42
cg10938046	C6orf223	Body	0.328454	3.56E-39
cg00061039	F11	TSS200	0.253093	2.26E-29
cg09075137	SCEL	3'UTR	0.289281	1.93E-28
cg16146033	SLC22A8	Body	-0.23207	2.69E-28
cg16806210	TTC36	1stExon	0.232749	1.52E-26
cg07493760	SLC4A9	TSS200	0.249099	1.40E-24
cg00144673	MCCD1	Body	0.303626	9.66E-24
cg03180302	TTR	1stExon	0.260254	5.62E-22
cg20688289	MUC12	TSS200	0.292753	3.36E-21
cg26433444	PI16	3'UTR	0.202948	5.93E-21
cg00498289	GCKR	1stExon	-0.28471	1.34E-19
cg11769400	PEBP4	Body	-0.30956	1.69E-18
cg19765377	MAT1A	TSS1500	0.205119	9.55E-18
cg14635269	LMX1B	Body	-0.26171	5.76E-17
cg08622198	CHRM3	5'UTR	-0.39448	3.32E-16
cg10974219	MYOCD	3'UTR	-0.34818	3.25E-15
cg16510654	C1orf64	TSS200	0.219073	8.19E-15
cg02704949	FXYD3	5'UTR	0.28631	9.15E-15
cg22324567	EBF2	Body	-0.20462	1.37E-13

Liver cancer has 425 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 185 CpG probes (44%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 185 CpG probes corresponds to 159 genes that: 111 has CpG sites with single localization belong to the specific pattern for breast cancer. The top 20 most significant CpG probes are represented in ***Table 4.8***.

Table 4.8 – Top 20 of most significant CpG probes selected for liver cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg10479063	PZP	1stExon	0.217808	2.77E-57
cg19358195	RPS6KA6	TSS1500	0.299307	7.74E-28
cg15452017	COX7B2	5'UTR	-0.2069	2.80E-25
cg02215603	HHIP	Body	-0.32207	3.73E-24
cg00012148	TINAG	TSS1500	-0.26392	8.24E-22
cg25368212	SSX1	TSS1500	-0.20673	3.17E-20
cg13510648	VCX3A	5'UTR	-0.42249	7.64E-19
cg20683151	TM4SF20	1stExon	-0.2109	1.04E-17
cg07041214	OR56A3	1stExon	-0.36219	2.29E-17
cg00974523	PRDM7	3'UTR	-0.26688	6.14E-17
cg09771429	LDLRAD1	TSS200	-0.26845	4.48E-16
cg25451456	OR56A3	TSS1500	-0.28525	5.79E-16
cg11357940	ZNF716	Body	-0.34037	1.17E-15
cg22467052	CFTR	Body	-0.2979	3.34E-15
cg17616453	C21orf62	5'UTR	-0.2414	3.21E-14
cg21860285	CPA6	TSS1500	-0.2323	2.16E-13
cg05626117	CLEC4G	TSS1500	-0.27335	9.50E-13
cg22165105	TINAG	3'UTR	-0.25925	1.11E-12
cg06563300	SLC17A8	TSS200	0.209888	1.13E-12
cg22799510	PROK2	3'UTR	-0.25783	1.32E-12

Lung cancer has 210 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, since the gene expression and methylation are “deep variables” (have many decimals), which resulted in 120 CpG probes (57%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 120 CpG probes corresponds to 86 genes that: 44 has CpG sites with single localization belong to the specific pattern for lung cancer. The top 20 most significant CpG probes are represented in ***Table 4.9***.

Table 4.9 – Top 20 of most significant CpG probes selected for lung cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg01392518	T	TSS1500	0.218394	6.42E-15
cg18768582	HBG1	Body	-0.22737	1.61E-14
cg12559170	HBG2	Body	-0.23612	9.96E-14
cg24748769	OTX2	Body	0.280923	8.20E-12
cg16856286	HOXC13	1stExon	0.270299	1.12E-11
cg11781718	HIST1H4E	TSS1500	0.237443	1.54E-11
cg06463958	T	TSS1500	0.288948	3.18E-11
cg07854132	OVCH1	TSS200	-0.29594	5.78E-11
cg19924352	FAM83A	1stExon	-0.28377	8.79E-11
cg23507945	IL22RA2	Body	-0.27795	1.06E-10
cg13791254	FOXE1	1stExon	0.297016	1.55E-10
cg01708273	HOXD11	3'UTR	0.363357	2.70E-10
cg16413687	ALX1	TSS1500	0.208811	3.67E-10
cg00633740	EDN3	Body	-0.20982	1.94E-09
cg26336935	KRT16	TSS200	-0.21271	4.47E-09
cg02650767	OR2B11	1stExon	-0.26845	1.09E-08
cg16464328	SLC4A1	TSS1500	-0.20833	1.20E-08
cg18451814	OTX2	TSS1500	0.327475	2.22E-08
cg19134945	PITX2	TSS200	0.236844	4.25E-08
cg06404175	OXT	Body	0.251856	4.37E-08

Thyroid cancer has 31 significant CpG probes that are distributed throughout gene considering only the probes with a single localization as followed: 27% TSS1500, 17% TSS200, 8% 5UTR, 2% 1st exon, 39% body and 4% 3'UTR. Assuming more demanding criteria related to *p-values* we considered, in this context, *p-values<0.0005*, which resulted in 27 CpG probes (87%). It should be noted that the majority of methylation of CpG probes were within an accurate range showing a strong relationship with the respective gene expression. The selected the 27 CpG probes corresponds to 23 genes that: 17 has CpG sites with single localization belong to the specific pattern for thyroid cancer. The 17 most significant are represented in ***Table 4.10***.

Table 4.10 – Top 17 of most significant CpG probes selected for thyroid cancer based on the MLR analysis.

cg	Gene	Localization	Delta-beta	p-value
cg23620049	LIPH	TSS200	-0.3957	2.02E-80
cg08328750	KRT15	TSS200	-0.30161	5.28E-68
cg12403889	C1orf187	5'UTR	-0.29036	7.35E-44
cg03255783	ESPN	Body	-0.20309	1.19E-38
cg01802532	NMU	Body	-0.22039	6.36E-37
cg04473405	KRT85	TSS1500	-0.30272	1.06E-28
cg25959149	BANF2	TSS200	0.248938	7.48E-24
cg22717825	PLA2G2E	TSS200	-0.36306	2.77E-16
cg03448202	MYBPH	TSS200	-0.33929	2.95E-16
cg15442792	MUC21	TSS200	-0.35681	5.90E-14
cg20695587	TMPRSS11F	TSS200	-0.26451	1.48E-11
cg25388882	C1orf180	Body	-0.34318	1.69E-07
cg19856444	SLC39A12	5'UTR	-0.26777	1.54E-05
cg02196805	CSF2	1stExon	-0.30557	5.18E-05
cg18959422	MYBPH	TSS1500	-0.22881	5.21E-05
cg18122696	SYT8	TSS1500	-0.3032	0.000195
cg23904115	AWAT2	TSS1500	-0.28298	0.000263

This significant CpG probes were majority located in genes that were specific for each type of cancer. This fact reveals that these specific patterns suggest high relevance in the initial distinction of the tumorigenesis, since the transcription start sites, 1500 and 200, were even more important after these results, making them preferred methylation sites soon after the gene body.

4.5.3. cDMGs using multivariate approach – Phase 5

Multivariate HJ-biplot technique was applied to all cohorts for the gene expression and DNA methylation. This procedure aimed to evaluate samples distribution, through the most important genes or CpG probes, in order to observe if effectively the solid tissue normal and stage I primary tumor samples were delimited in the same distribution space (***Appendix 9***). This multivariate data reduction approach also helped to corroborate the designed pipeline.

In breast cancer, HJ-biplot for gene expression were represented in ***Figure 4.7***. Results show that *KCNJ16*, *HPSE2*, *LRRC3B*, *DPP6* and *CPA1* were genes with high importance in the distribution of samples, when we consider their vector norm. We also verified that *HPSE2-LRRC3B* and *DPP6-CPA1* were strongly correlated. Colors of HJ-Biplot coordinates represent the output of hierarchical clustering procedure, which was discriminated in three groups characterized by: C1 with 80 normal samples, C2 with 4 normal and 29 tumor samples and C3 with 97 tumor samples. Additionally, results showed that C1 group is well differentiated and positioned in the same direction of genes and in opposition to C3. This fact suggests that higher levels of gene expression in normal samples have lower gene expression levels in tumor samples, considering the genes: *KCNJ16*, *HPSE2*, *LRRC3B*, *DPP6* and *CPA1*. Indeed, the $\log_2$(Foldchange) of that set of genes (-1.56, -1.79, -1.98, -1.71 and -2.41, respectively) confirmed this tendency of down-regulation of gene expression. However, there is an intermediate cluster, C2, where samples are also closer to the other clusters, meaning that those samples might be confused. This HJ-Biplot representation retains 85.7% of variance in the plan 1-2 (***Figure 4.7***). Importantly, in the HJ-Biplot representation, we recognized one gene (*CPA1*) which have been previously described in a specific pattern for breast cancer. Interestingly, that gene presents the lower $\log_2$(Foldchange).

Moreover, DNA methylation HJ-Biplot (***Annex 1***) is similar to the gene expression HJ-Biplot. Clusters are distributed by: C1 with 84 normal and 14 tumor samples, C2 with 75 tumor samples and C3 with 37 tumor samples. C2 cluster is an intermediate group doubtful in terms of effective classification of normal or tumor. C1 is essentially explained by the high variability in 14 CpG probes which are in the same samples direction. However, tumor samples were influenced by other 5 CpG probes. Interestingly, a more detailed analysis revealed that 19 CpG probes belong to only two of the referred genes. *DPP6* presents 16 CpG probes (5 hypermethylated and 11 hypomethylated) according to the previously verified in the graph. At last, the 4 CpG probes remaining (4 hypomethylated) are located in the *KCNJ16* gene. This suggests that *DPP6* gene is very important in breast cancer, although it is not specific. DNA methylation HJ-Biplot retained 88% of variance in the plan 1-2.

In breast cancer, in spite of the working pipeline performed, there is yet an intermediate group (for gene expression and methylation) in terms of centroids distance related to the normal exclusive (C1) group and the tumor exclusive samples, cluster C3. This fact reveals that remains a kind of a “doubt group” to take into account, since it does not distance itself sufficiently from the other clusters and, eventually, it can signify misrepresented or classified samples.

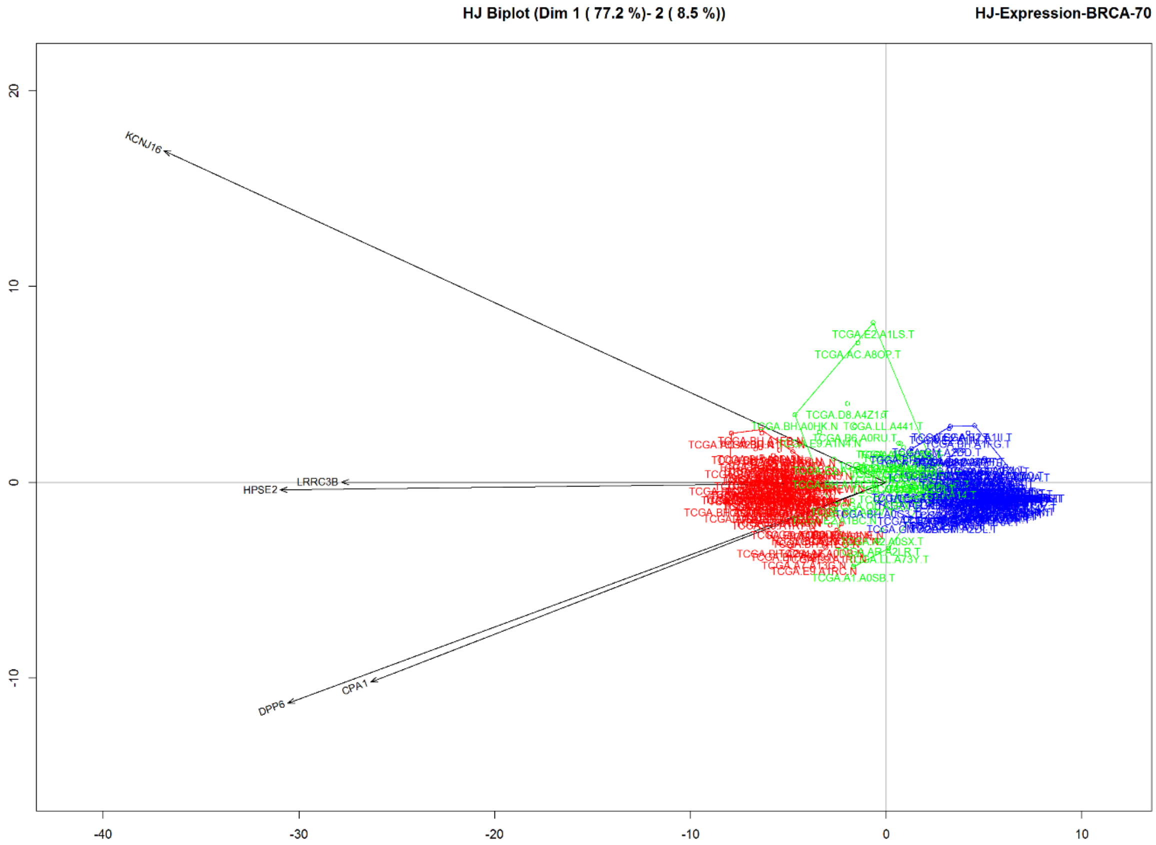


Figure 4.7 – HJ-biplot for gene expression in breast cancer. This representation results from a variable selection with more than 70% of contribution, retaining 85.7% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 80 normal, C2 with 4 normal and 29 tumor samples and C3 with 97 tumor samples.

In colorectal cancer, HJ-Biplot for gene expression is represented in ***Figure 4.8***. Results identified a set of 84 genes with impact in the distribution of samples. All of them are positively correlated. Cluster analysis reveals three well defined groups: C1 with 21 normal samples, C2 with 18 tumor samples and C3 with 36 tumor samples. Samples in C1 were distributed through influence of genes. Indeed, these genes are all down-regulated in tumor samples. *TMEFF2* gene presents the biggest difference (log$_2$(Foldchange) = -3.81). Other clusters are further away from C1 and without samples intersections. Gene expression HJ-Biplot retains 86.4% of the variance in the plan 1-2. Looking at specific gene pattern for each type of cancer, from third aim, we verified that 61 genes are specific for colorectal cancer, being *TMEFF2* one example of them.

DNA methylation HJ-Biplot had a similar behavior to gene distribution (***Annex 2***). From clustering, results in three clusters, C1 with 21 normal and 2 tumor samples, C2 with 25 tumor samples and C3 with 27 tumor samples. C1 was essentially explained by 14 CpG probes and C3 by 102 CpG probes, in opposite side. Interestingly, the same behavior was verified through the Δβ. Specifically, *TMEFF2* presents 5 hypermethylated CpG probes, suggesting a high control of gene expression. DNA methylation HJ-Biplot retained 85.3% of the variance in the plan 1-2.

In colorectal cancer, were found three well defined clusters. One of them corresponds to normal samples and the other two are tumor groups. Moreover, the group constituted by normal samples are sufficiently distant from tumor groups, suggesting a strong possibility that they might have subgroups in stage I tumor samples.

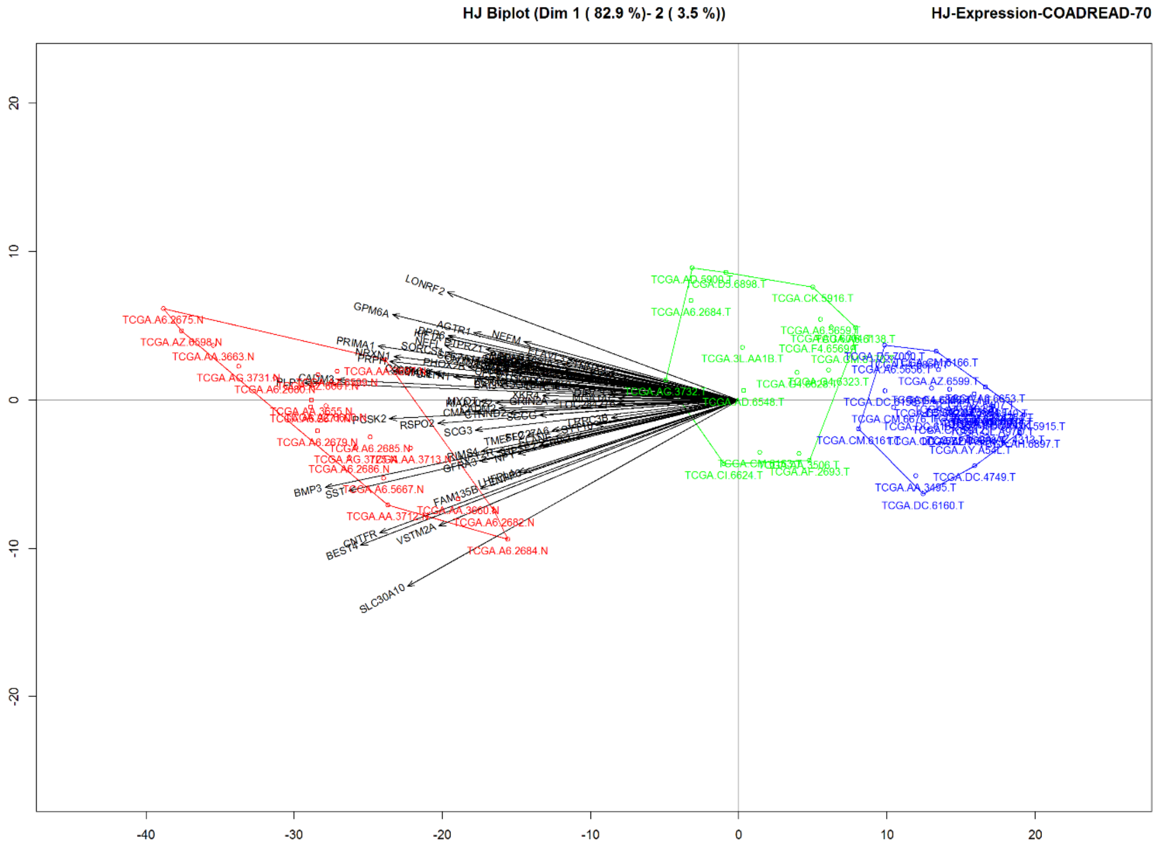


Figure 4.8 – HJ-biplot for gene expression in colorectal cancer. This representation results from a variable selection with more than 70% of contribution, retaining 86.4% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 21 normal samples, C2 with 18 tumor samples and C3 with 36 tumor samples.

In head and neck cancer, HJ-biplot for gene expression is represented in the ***Figure 4.9***. Results showed that there was a set of 7 genes with impact in the distribution of samples, and the hierarchical clusters analysis was distributed: C1 with 16 normal samples, C2 with 3 normal and 8 tumor samples and C3 with 1 normal and 19 tumor samples. Additionally, the group constituted only by normal samples (C1) was well differentiated and positioned in the same orientation of genes. Indeed, all genes were down-regulated in tumor samples according to the working pipeline. Furthermore, exists an intermediate group (C2) that translates doubt in terms of effective classification of samples. Gene expression HJ-Biplot retained 87.1% of the variance in the plan 1-2. Looking at the specific pattern of genes for each type of cancer from third aim, we verified that 5 genes are specific for head and neck cancer. *FOXI2* and *GRIK3* genes are examples of this with a $\log_2$(Foldchange) = -1.97 and -1.61, respectively.

However, the behavior of DNA methylation HJ-biplot is similar to the distribution in gene expression (***Annex 3***). From clustering, results three clusters (C1 with 19 normal and 1 tumor samples, C2 with 6 tumor samples and C3 with 1 normal and 20 tumor samples), C3 essentially explained by the variability power of the following 7 CpG probes. Interestingly, all CpG probes are hypermethylated and were located in *FOXI2* (5 CpG probes) and *GRIK3* (2 CpG probes). Importantly, C1 is very cohesive and defined. Additionally, it should also be noted that there is an intermediate cluster (C2) great defined with more variability that each other. DNA methylation analysis in head and neck cancer gives us a quantification of information from the 1-2 plans in the order of 95.8%.

In Head and neck cancer, three bordered clusters were found. One of them corresponds to the normal samples and two tumor groups sufficiently distant between them. However, since there are blue samples closer to C1 is doubtful to suggest that the C2 is well defined.

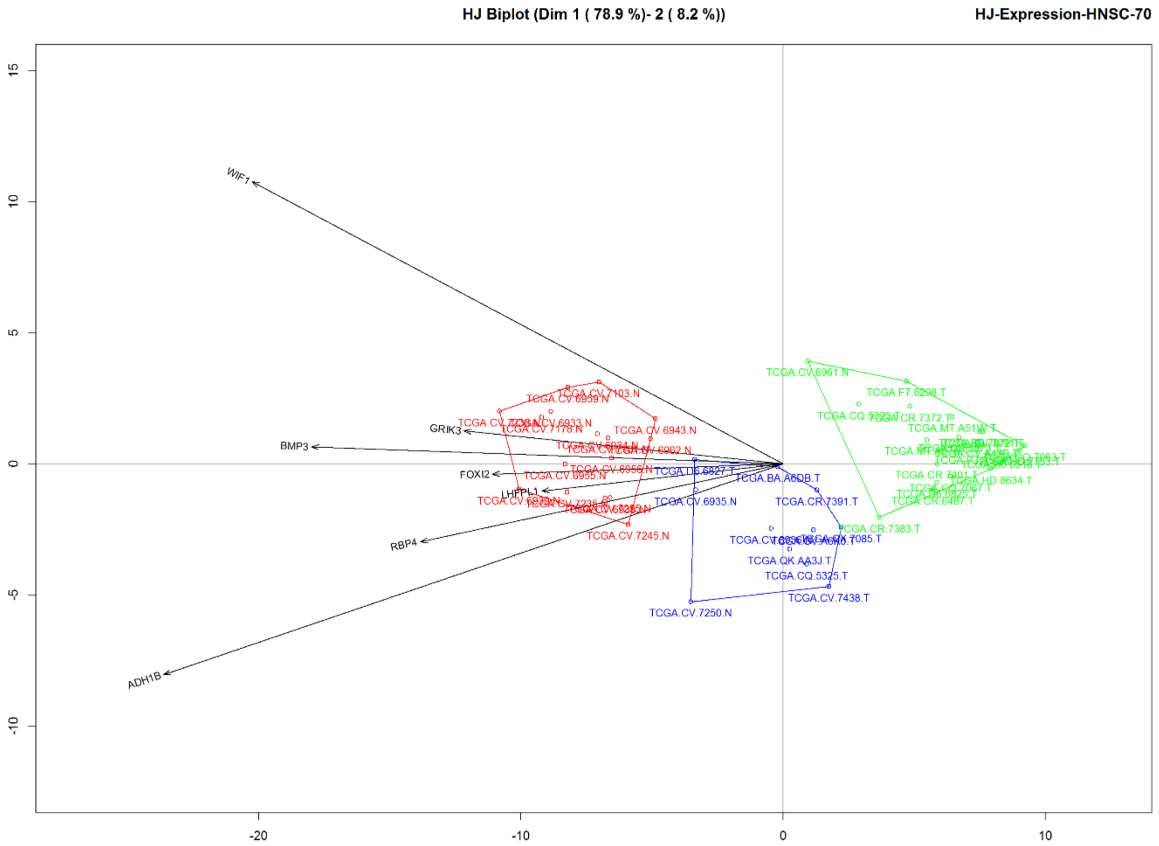


Figure 4.9 – HJ-biplot for gene expression in head and neck cancer. This representation results from a variable selection with more than 70% of contribution, retaining 87.1% of variance. Hierarchical clusters are represented in red (C1), blue (C2) and green (C3). C1 with 16 normal, C2 with 3 normal and 8 tumor samples and C3 with 1 normal and 19 tumor samples.

In kidney$_R$ cancer, HJ-biplot gene expression is represented in ***Figure 4.10***. Results showed that there are 2 sets of important genes in samples distribution: 44 genes with the principal impact on normal samples (left side) and 3 genes in the distribution of tumor samples (left right). Hierarchical clusters analysis is distributed: C1 with 24 normal samples, C2 with 41 tumor samples and C3 with 144 tumor samples. Indeed, this behavior agreed with working pipeline results (44 genes down-regulated and 3 genes up-regulated). Although well bordered clusters, C2 has samples closer to C1 and C3. Gene expression HJ-Biplot retained 85.9% of the variance in the plan 1-2. When we looked at the specific pattern of genes for each type of cancer from second objective, we verified that 27 genes have a specific gene for kidney$_R$ cancer. *RANBP3L* gene, a downregulated gene ($\log_2$(Foldchange) = -1.83), is an example of them.

DNA methylation HJ-biplot was different in to samples distribution (***Annex 4***). Clustering results showed: C1 with 21 normal and 12 tumor samples, C2 with 3 normal and 74 tumor samples and C3 with 69 tumor samples. C2 and C3 are merged and directly correlated with 25 hypermethylated CpG probes according to the results obtained by the working pipeline. DNA methylation HJ-Biplot retains 88.8% of the variance in the plan 1-2.

In kidney$_R$ cancer, seems that gene expression has more influence in differentiating samples, when compared to DNA methylation. In fact, clusters obtained by DNA methylation were undefined if we consider the intersections of samples.

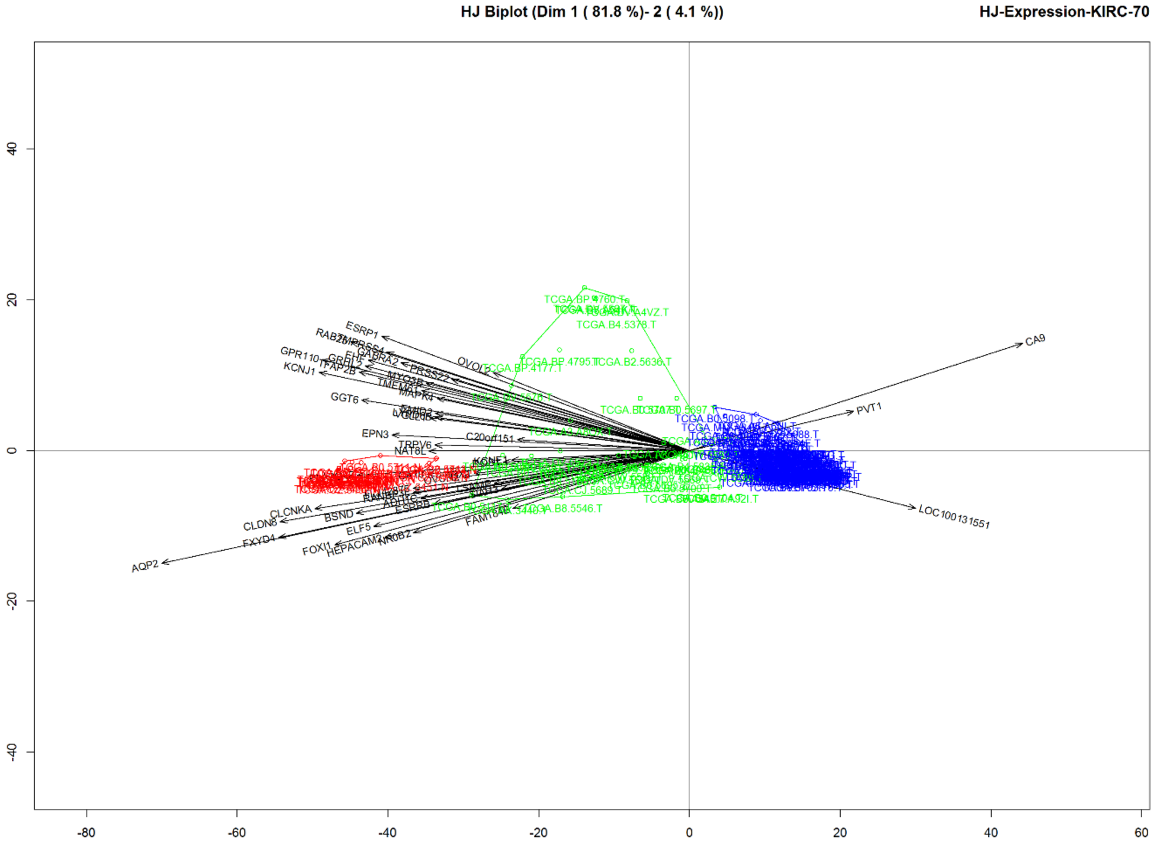


Figure 4.10 – HJ-biplot for gene expression in kidney$_R$ cancer. This representation results from a variable selection with more than 70% of contribution, retaining 85.9% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 24 normal, C2 with 41 tumor samples and C3 with 144 tumor samples.

In Kidney$_P$ cancer, HJ-biplot for gene expression is represented in ***Figure 4.11***. Results showed that there are 19 genes with impact in the samples distribution. Hierarchical clusters analysis is distributed: C1 with 23 normal and 4 tumor samples, C2 with 105 tumor samples and C3 with 58 tumor samples and verified that C2 and C3 overlapping. Genes are distributed according to normal samples and confirmed by working pipeline. Indeed, all of them are down-regulated in tumor. Gene expression HJ-Biplot retained 87.5% of the variance in the plan 1-2. Looking at the specific pattern of genes for each type of cancer from second objective, we verified that 10 genes are specific for kidney$_P$ cancer. *SLC4A9* is an example of a specific down-regulated gene for each cancer ($\log_2$(Foldchange) = -1.83).

DNA methylation HJ-biplot is similar to the gene expression distribution (***Annex 5***). Clustering results are distributed: C1 with 23 normal and 11 tumor samples, C2 with 63 tumor samples and C3 with 93 tumor samples, also considering that C2 and C3 are overlapped and explained by 7 CpG probes. Confirmed by the working pipeline as hypermethylated CpG probe. DNA methylation HJ-Biplot retained 89.9% of the variance in the plan 1-2.

In kidney$_P$ cancer, it was observed that in gene expression are not necessarily three clusters, since two of them are almost total overlapped. Same patterns were seen in the DNA methylation HJ-Biplot (***Annex 5***). This fact strongly suggests that we have truly two groups of samples. However, once the distribution of C1 samples was 32% for tumor and 68% for normal samples, since with substantial distance between C1 and C2-C3, the C1 must have further attention to answer why those tumor samples were in there. In summary, in this type of cancer is not clear that we can properly differentiate tumor from normal samples.

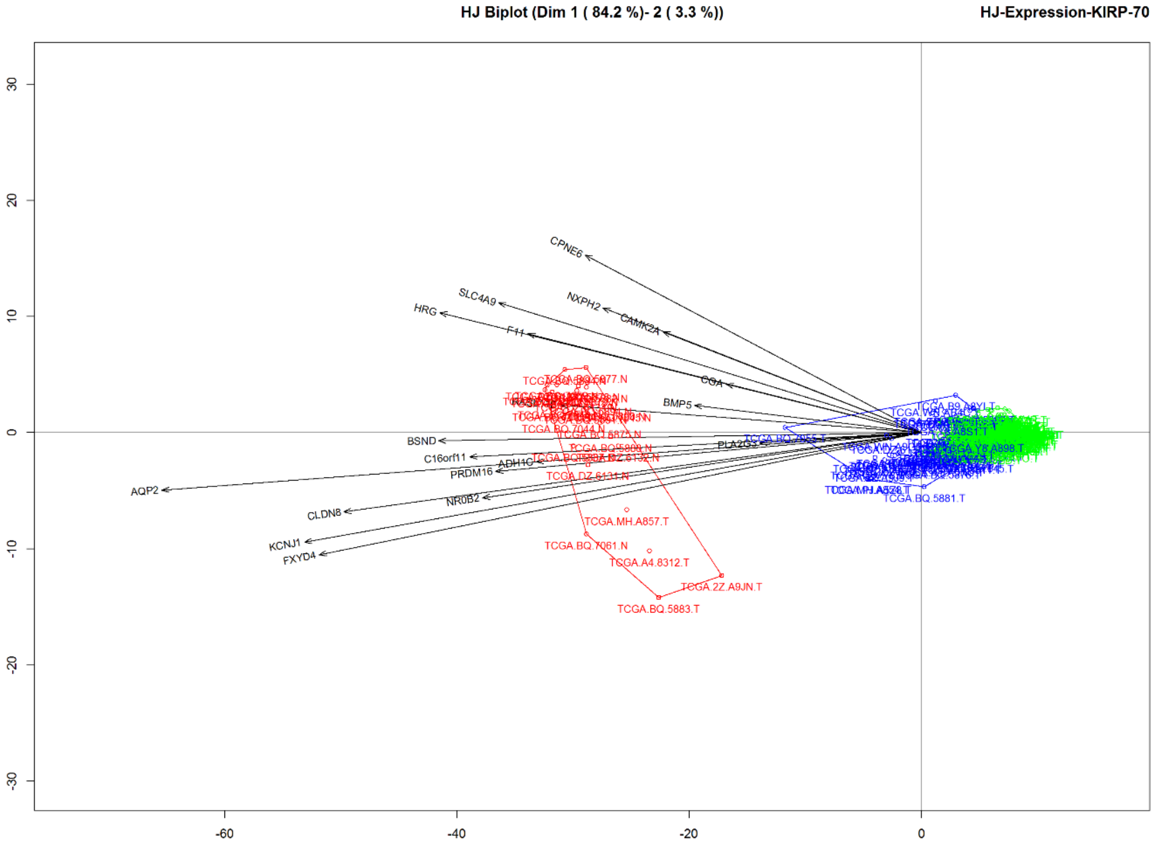


Figure 4.11 – HJ-biplot for gene expression in kidney$_P$ cancer. This representation results from a variable selection with more than 70% of contribution, retaining 87.5% of variance. Hierarchical clusters are represented in red (C1), blue (C2) and green (C3). C1 with 23 normal and 4 tumor samples, C2 with 105 tumor samples and C3 with 58 tumor samples.

In liver cancer, HJ-biplot for gene expression is represented in ***Figure 4.12***. Results showed that 13 genes have an impact in samples distribution. Interestingly, all genes are down-regulated with conferring of previous results of working pipeline. Hierarchical clusters are distributed: C1 with 40 normal and 6 tumor samples, C2 with 1 normal and 20 tumor samples and C3 with 145 tumor samples were C2 - C3 overlapping. These genes follow the same behavior of the previous cohort. All genes are down-regulated and confirmed the HJ-Biplot tendency. Looking at the specific gene pattern for each type of cancer from second objective, we verified that 11 genes are for liver cancer.

DNA methylation HJ-biplot was similar to the gene expression distribution (***Annex 6***). Clustering results are distributed: C1 with 41 normal and 70 tumor samples, C2 with 49 tumor samples and C3 with 52 tumor samples. CpG probes are in the same direction as C1, contrary to the other cohorts. Tumor patients of this cluster suggests a high proximity with normal samples. All CpG probes are Hypo-methylated in tumor confirmed by working pipeline. DNA methylation HJ-Biplot retained 81.9% of the variance in the plan 1-2.

In liver cancer, three clusters are defined, but the C1 presents a mix of samples suggesting a high proximity between normal samples and subgroup of tumor samples. Other clusters are well defined. Importantly, liver cancer presents hypomethylated CpG probes and down-regulated respective genes with more importance to differentiate groups.

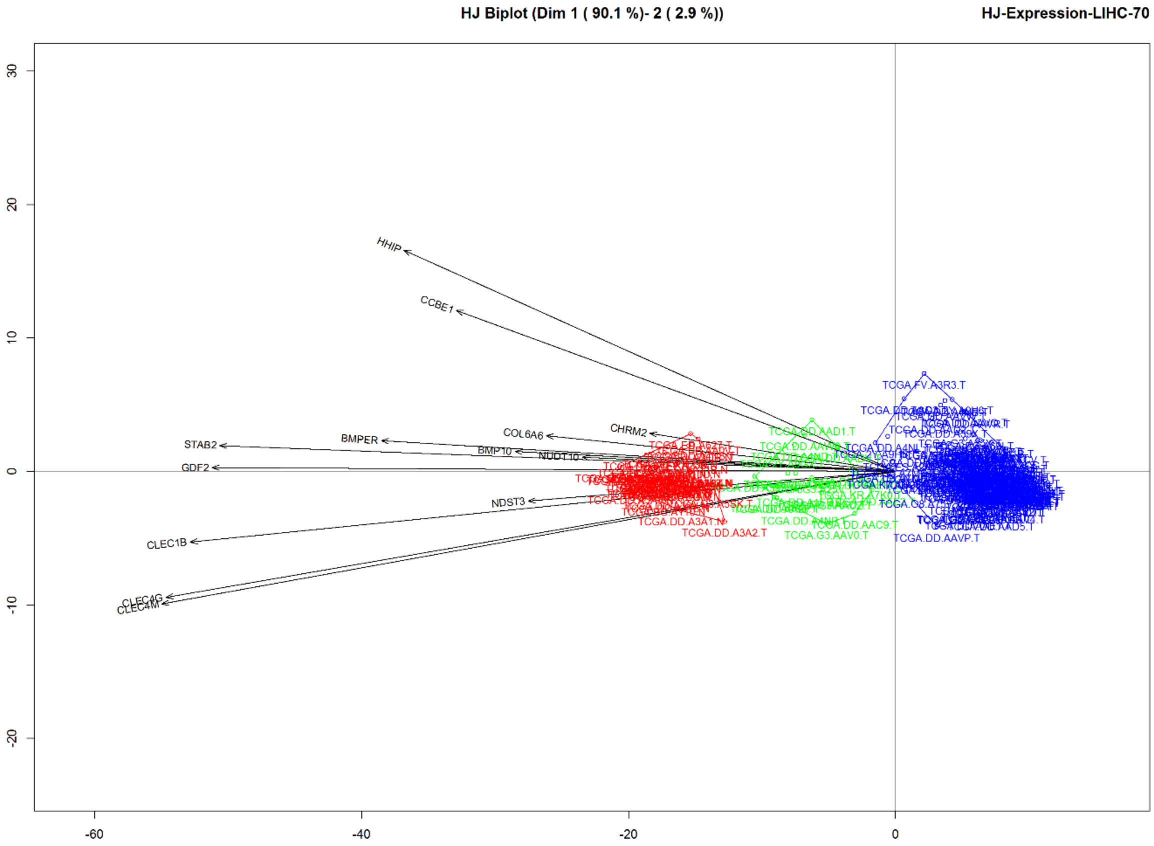


Figure 4.12 – HJ-biplot for gene expression in liver cancer. This representation results from a variable selection with more than 70% of contribution, retaining 93% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 40 normal and 6 tumor samples, C2 with 1 normal and 20 tumor samples and C3 with 145 tumor samples.

In lung cancer, HJ-biplot for gene expression is represented in ***Figure 4.13***. Results identified 7 genes with impact in samples distribution. Hierarchical clusters are distributed: C1 with 21 normal and 19 tumor samples, C2 with 149 tumor samples and C3 with 77 tumor samples. Indeed, all genes are down-regulated according to working pipeline analysis. Results showed that C1 was constituted by approximately 50% of tumor samples, and other clusters are overlapped. This fact means that it was difficult to differentiate clusters of samples in this cancer type. Gene expression HJ-Biplot retained 79.5% of the variance in the plan 1-2.

Looking at the specific pattern of genes for each type of cancer from second objective, we verified that only 2 genes (*AGBL1* and *OVCH1*) are specific for lung cancer, but do not have selected CpG probes.

DNA methylation HJ-biplot is similar to gene expression distribution (***Annex 7***). Clustering results are distributed: C1 with 21 normal and 24 tumor samples, C2 with 135 tumor samples and C3 with 86 tumor samples. Same behavior of hypermethylation was verified. DNA methylation HJ-Biplot retained 81.5% of the variance in the plan 1-2.

In liver cancer, C1 presents a merge of normal and tumor samples suggesting a high proximity. C2-C3 are merged forming only a cluster. In contrast, DNA methylation approximates samples of C1, expanding samples from C2 and C3. This fact suggests that methylation has more explicability in normal samples.

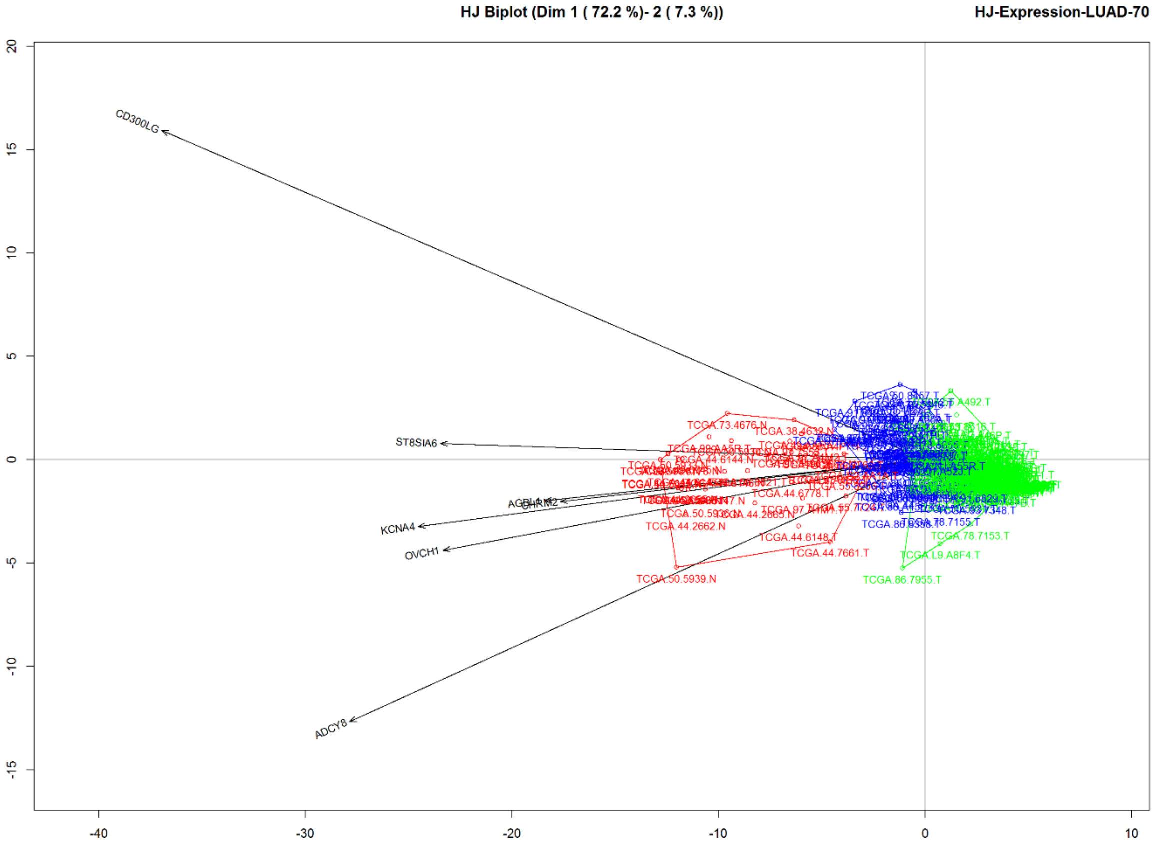


Figure 4.13 – HJ-biplot for gene expression in lung cancer. This representation results from a variable selection with more than 70% of contribution, retaining 79.5% of variance. Hierarchical clusters are represented in red (C1), blue (C2) and green (C3). C1 with 21 normal and 19 tumor samples, C2 with 149 tumor samples and C3 with 77 tumor samples.

In thyroid cancer, HJ-biplot for gene expression is represented in ***Figure 4.14***. Results showed that there are 6 genes which have an impact in samples distribution. Hierarchical clusters are distributed: C1 with 49 normal and 59 tumor samples, C2 with 1 normal and 97 tumor samples and C3 with 128 tumor samples. Results showed one group with tumor samples (C3) and two merged groups (C1-C2) with both tumor and normal samples. Genes are in the same direction of tumor groups revealing be up-regulated according to the results obtained by the working pipeline. Gene expression HJ-Biplot retained 88.1% of the variance in the plan 1-2. Looking at patterns of specific genes for each type of cancer from second objective, we verified that only 6 genes have a specific gene for thyroid cancer. *MUC21* is an example of them, presetting $\log_2$(Foldchange) = 2.26.

DNA methylation HJ-biplot was similar to the gene expression distribution. (***Annex 8***). Clustering results are distributed: C1 with 50 normal and 76 tumor samples, C2 with 77 tumor samples and C3 with 131 tumor samples. C3 is directly correlated with 25 hypomethylated CpG probes according to the results of the working pipeline. DNA methylation HJ-Biplot retained 88.1% of the variance in the plan 1-2.

In thyroid cancer, samples present high proximity. C1 is composed by a merge of normal and tumor samples, revealing few differences. C2 is an intermediary group that merged with C3 and C1. Contrary, DNA methylation presents more differences between groups distributed them more defined.

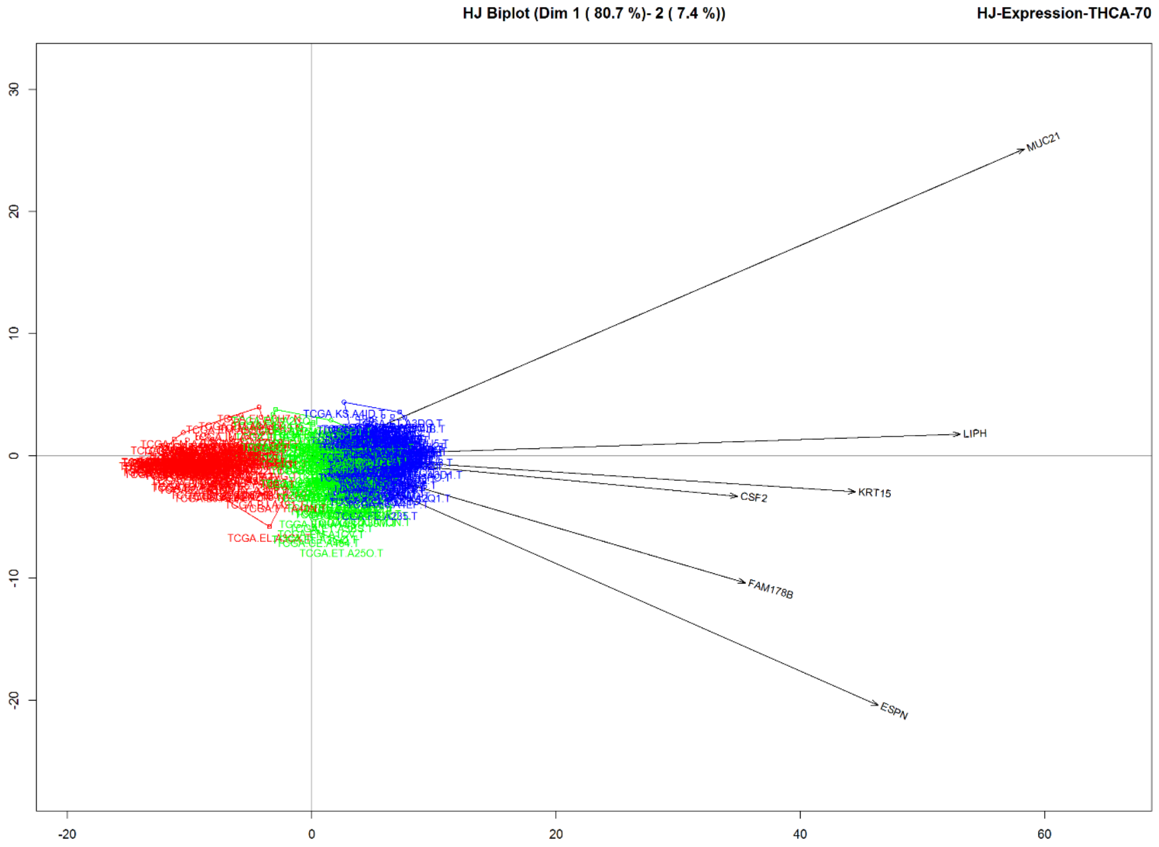


Figure 4.14 – HJ-biplot for gene expression in thyroid cancer. This representation results from a variable selection with more than 70% of contribution, retaining 88.1% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 49 normal and 59 tumor samples, C2 with 1 normal and 97 tumor samples and C3 with 128 tumor samples.

5 CHAPTER 5 – DISCUSSION

In this study, we developed a mathematical structured model to explore big data resources through the design of a Working Pipeline (WP).

An exponential increase in data repositories is being observed. TCGA is a good example of public platforms and was the resource of gene expression and DNA methylation of this study. We analyzed 20502 genes and 364643 CpG probes as our initial variables in order to compare normal tissue and stage I primary tumor groups.

Here, we developed a tool that pretends to identify specific patters across cancers. For that, we stratified our model into 6 different Phases distributed into two components: 4 phases in principal component and 2 phases in complementary component.

Firstly, to characterize descriptively the data information is essential to be aware of the main studied populations features (***Figure 4.1: Phase 0***).

Then, to organize the data is essential to prepare inputs in order to be used in the next phase. Performing tests to compare two general groups is crucial to sustain the effective difference in those subsets and apply strategic cut points is necessary to improve general accuracy (***Figure 4.1: Phase 1***).

After, the development of strategies to analyze intersections of variables is also necessary to identify patterns to characterize cancer populations (***Figure 4.1: Phase 2***).

Furthermore, identifying good predictors in order to discriminate groups is a fundamental procedure in cancer diagnosis (***Figure 4.1: Phase 3***).

Finally, it is important to perform other deep analysis such as: select variables that have a truly impact in other variables, such as which CpG probes were significant in the dynamics of gene expression (***Figure 4.1: Phase 4***); and to perform graphic representations to visualize latent correlations, between gene expression and DNA methylation, that only multivariate approaches can capture (***Figure 4.1: Phase 5***).

Next, we tested our developed model in 8 TCGA cohorts.

In this WP, the pre-processing data and the variables selection were performed in the ***Phase 1***, saved in ***Output 1*** and matched with the same criteria as founded in Zhang et al. 2015 study. We also included an outlier procedure (Boxplot method) to remove discrepant data, since we had substantial differences in sample sizes.

Moreover, in ***Phase 1*** we performed a t-test to differentiate the mean of gene expression levels in normal and tumor samples. Also, the Pearson test was used to identify significant correlations between patterns of gene expression and differentially methylated CpG probes and saved in ***Output 3***. These statistical procedures were also performed in the previously mentioned study (Zhang et al. 2015).

In order to cover all the tests in accordance to statistical assumptions, we also performed, in ***Phase 1***, the non-parametric Wilcoxon test (Mann-Whitney U test) for unpaired samples to compare methylation and gene expression mean levels (normal vs tumor), depending on the Shapiro-Wilk normality test results. The methylation cutoff was $|\Delta\beta| > 0.2$ and the considered gene expression cutoff was |Foldchange| > 2.8. This was also done in another study (Wei et al. 2016) although the gene expression cutoff was different (|Foldchange| > 2) (***Output 2***).

In ***Phase 2***, we computed an intersection function that combined set of genes or CpG probes across all cancer cohorts. This phase reported to ***Output 4*** the frequency of all intersections of genes.

In ***Phase 3***, we performed ROC curve analysis to identify potential new biomarkers for our cancer cohorts (***Output 5***). Also, this diagnostic tool was used previously to identify potential new biomarkers in rectal cancer (Wei et al. 2016).

Next, in ***Phase 4***, linear regression models were performed to capture the significant CpG probes in the variation of gene expression. We predefined the top 20 of more significant CpG probes in gene expression (***Output 6***).

Finally, in ***Phase 5***, we used the HJ-Biplot representation to assess sample distributions according to the multivariate behavior of genes and CpG probes. Aberrant methylation changes can start very early in tumor development promoting several signaling pathways abnormalities, such as genetic and epigenetic instability (Baylin et al. 2001). Therefore, identifying cDMGs for each cancer type is of most relevance.

Aberrant methylation changes can start very early in tumor development, mediate a several important signaling pathways abnormalities in cancer, such as genetics instability. Identifying cDMGs is very appellative reason for the main hypothesize if all types of cancer have the same dynamic changes.

Since tumorigenesis involves a lot of aberrant alterations (Hanahan et al. 2000) it was expected that a large sets of genes and CpG probes were altered between normal and stage I tissue from different tumors. For each tumor type, a coverage of 20531 genes and 48557 CpG probes is provided at TCGA data sets. (***Figure 4.2A***). Upon applying our ***Phase 1*** step we verified that a small set of genes (25 in thyroid cancer and 349 in liver cancer) and CpG probes (40 in thyroid cancer and 1453 in liver cancer) characterized cancer initiation (***Figure 4.2C***).

Also, our data showed that liver and colorectal cancer presents the most changes in gene expression and DNA methylation (***Figure 4.2C***). Both organs are directly linked to metabolism and they are constant contact with many insults. Previous studies revealed that lesions caused by inflammation, mechanical and chemical agents promote a chronic immune response that potentiates cell proliferation and regeneration (Mariani et al. 2014). Also, high risk of cancer is associated to life habits, such as: tobacco, diet, obesity, alcohol, among others (Simon 2016). In contrast, thyroid cancer showed to be the cancer with less epigenetic alterations, suggesting a slower development. Interestingly this tumor type is often indolent and is considered a curable cancer when diagnosed at early stages (Mazzaferri & Kloos 2001).

We also verified that there are cohorts defined by specific gene expression profiles. In deed breast, liver, lung and thyroid cancers present more upregulated genes suggesting a positive transcription regulation. In contrast, colorectal, head and neck, kidney$_R$ and kidney$_P$ cancers presents more downregulated genes revealing a negative transcription regulation. It is interesting to notice that the downregulated cancer cohorts were more associated to hypermethylated CpG probes where the upregulated cohorts were associated to hypomethylated CpG probes. This fact suggests that overall hypermethylation is associated to downregulation and hypomethylation to upregulation (Victoria Valinluck Lao and William M. Grady 2011). However, some tumor types including liver and breast cancer showed hypermethylation to be associated to upregulation. Previous studies have already reported this trend (Castelo-Branco et al. 2013; Bert et al. 2013).

This study identified, for each cohort, multiple CpG sites differentially methylated which are correlated to alterations in gene expression (***Figure 4.3***). We verified that most altered CpG sites were located in transcription start sites (TSS1500, TSS200) and in the gene body in agreement with other previous studies (Kumar Mishra & Guda 2017).

Interestingly, CpG presenting an hypomethylated status were mostly located in the gene body where hypermethylated CpGs were mostly at transcription starts site. This is consistent with the dogma where of cancer epigenetics where it is observed global hypomethylation and specific hypermethylation.

Biological process enrichment analysis was based on a gene ontology platform (***Figure 4.4***). Our results showed that events associated to the nervous system and development were enriched in all cohorts. Interestingly, chemical synaptic transmission was also presented in the majority of the analyzed cohorts.

Next, we performed literature searches and observed that approximately 87% of cancer differentially methylated genes that came out in our analysis were previously reported in cancer. (***Figure 4.5***). Strikingly, more than 10% of the genes identified by our model have not yet been reported in cancer and can therefore be potential cancer biomarkers.

Analysis of patterns across the different cancer cohorts revealed that the majority of the cDMGs are tissue-specific for early stages of the disease (***Output 4)***, suggesting that regulation of these key genes depends on the cell where it originates (***Figure 4.6***). In fact, genome-wide DNA methylation profiling study identified differentially methylated regions in 17 human somatic tissues which are also tissue-specific (Lokk et al. 2016).. However, more genome-wide multicenter studies are necessary to validate this hypothesis.

Next, we did pathways enrichment analysis in a *Reactome Pathway Database*. Regarding colorectal cancer, RAF/MAP kinase cascade that is involved in the regulation of cell proliferation, differentiation, migration and apoptosis was enriched in our study. (Slattery et al. 2012). In head and neck cancer, the GABA A receptor activation is an enriched pathway that plays a role in the vertebrate central nervous system (Simon et al. 2004). Importantly, in the present study, this pathway related to the nervous system was found to be enriched. In lung cancer *OPRD1*($\log_2$(foldchange)=4.29) is an example of a gene that is upregulated and part of the peptide ligand-biding receptors pathway. Interestingly, this gene was found to be overexpressed in lung cancer but not in normal lung (Cohen et al. 2016). In kidney$_P$ cancer the ion homeostasis pathway was found to be enriched and previous studies reported this mechanisms to be downregulated in kidney tumor cells (Boer et al. 2001). At last, adherent junction interactions is an enriched pathway in liver cancer.

We then searched for biomarkers with diagnostic potential and identified a subset of cDMGs for each cohort (***Output 5***). Good potential new biomarkers were discriminated to select the tissue-specific pattern of cDMGs. As examples we saw that in breast cancer out of the 19 specific cDMGs which were considered as good diagnostic predictors, 5 have never been mentioned in cancer, such as *METTL11B* (AUC = 0.83) with cg20742415 (AUC = 0.83). *METTL11B* or *NRMT2* (N-terminal RCC1 methyltransferase) is a methyltransferase primarily monomethylase to specifically methylate free α-amino group of proteins (Petkowski et al. 2013). In colorectal cancer, from a set of 153 specific genes, 64 have not yet been reported in cancer, such as *RIC3* (AUC=0.97) which presents the cg04886703 (AUC=0.99) as also a good diagnostic biomarker. Interestingly, this gene was reported to promote the expression of the nicotinic acetylcholine receptor alpha7 subunit (Halevi et al. 2003). In the complementary phase (Phase 4), we attested the significant methylation of CpG probes that contributes to the dynamic performance of gene expression. Regarding to the top of the most significant CpG probes, we found that, in all cancer cohorts, the region of the transcription start sites is more enriched, followed by gene body. In fact, cg20742415 (*p-value=1.99e-10*) in breast cancer was located in TSS1500 and cg03827337 (*p-value=0.028*) in head and neck. However, cg18390495 (*p-value=3.76e-44*) in kidney$_R$ cancer, as well as, cg20685897 (*p-value=7.67e-05*) in liver cancer were located in the gene body. Interestingly, gene body methylation has been reported as a potential therapeutic target for modulation of transcription levels (Yang et al 2018).

Finally, in order to validate the developed WP, we used HJ-Biplot for all cancer cohorts analyzed and observed that when genes or CpG probes were projected in the direction of normal samples, they were downregulated or hypomethylated, respectively. Indeed, these genes and CpG probes explained the distribution of normal samples. In opposition, when genes or CpG probes were projected in the direction of tumor samples, they were upregulated or hypermethylated, respectively. Additionally, since this analysis was performed based on normal samples these results were in agreement with the principal component of the WP that can be used in other scientific contexts.

5.1. Limitations of study

This present study presents some limitations, such as:

1. Number of samples varied among the different cohorts, which may contribute to a decrease robustness of our analyzes;
2. Normal patient samples were not obtained from normal patients, but rather extracted from tumor adjacent tissue of patients with disease. As these surrounding tissues may already experience some alterations this can affect the viability of our results.
3. These results were not validated by another data sets.

6 CHAPTER 6 – CONCLUSION

We develop a novel working pipeline that permits analyzing big data sets available worldwide. For that we used a mathematical structured model that imports, exports, cleans and computes statistical power techniques for big data repositories.

In order to validate our model, we use TCGA data on multiple tumor types.

Remarkably, our findings evidence that the transition between normal tissue into a carcinogenic stage is not a conserved event that occurs in different tumor types but specific to the cell of origin.

Indeed, specific patterns of gene expression and differentially methylated genes allowed to find new biomarkers with high capacity to discriminate normal and tumor samples in the initial stages of cancer.

BIBLIOGRAPHY

Aguinis, H., Gottfredson, R.K. & Joo, H., 2013. Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. *Organizational Research Methods*, 16(2), pp.270–301.

Anon, Bioconductor - About. Available at: https://www.bioconductor.org/about/ [Accessed September 11, 2018a].

Anon, CANCER TODAY. *IARC*. Available at: http://gco.iarc.fr/today/home [Accessed September 18, 2018b].

Anon, Platform Design - TCGA. Available at: https://cancergenome.nih.gov/abouttcga/aboutdata/platformdesign [Accessed August 7, 2018c].

Aran, D. et al., 2017. Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nature Communications*, 8(1), pp.1–13. Available at: http://dx.doi.org/10.1038/s41467-017-01027-z.

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell Research*, 21(3), pp.381–395. Available at: http://dx.doi.org/10.1038/cr.2011.22.

Baubec, T. et al., 2015. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. *Nature*, 520(7546), pp.243–247.

Baylin, S.B. et al., 2001. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human molecular genetics*, 10(7), pp.687–692.

Benjamini, Y. & Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Source Journal of the Royal Statistical Society. Series B (Methodological) Journal of the Royal Statistical Society. Series B J. R. Statist. Soc. B*, 57(1), pp.289–300. Available at: http://www.jstor.org/stable/2346101%5Cnhttp://about.jstor.org/terms.

Berg, J.M., Tymoczko, J.L. & Stryer, C., 2008. *Bioquímica* 6th ed., NOVA GUANABARA.

Bert, S.A. et al., 2013. Regional Activation of the Cancer Genome by Long-Range Epigenetic Remodeling. *Cancer Cell*, 23(1), pp.9–22. Available at: http://dx.doi.org/10.1016/j.ccr.2012.11.006.

Bharathy N., Ling B.M.T., T.R., 2013. *Epigenetic Regulation of Skeletal Muscle Development and Differentiation. In: Kundu T. (eds) Epigenetics: Development and Disease, Subcellular Biochemistry, vol 61.*, Springer, Dordrecht.

Bishop, K.S. & Ferguson, L.R., 2015. The interaction between epigenetics, nutrition and the development of cancer. *Nutrients*, 7(2), pp.922–947.

Boer, J.M. et al., 2001. Identification and Classification of Differentially Expressed Genes in Renal Cell Carcinoma by Expression Profiling on a Global Human 31,500-Element cDNA Array. *Genome Research*, 11(11), pp.1861–1870. Available at: http://genome.cshlp.org/lookup/doi/10.1101/gr.184501.

Bray, F. et al., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Journal for Clinicians*.

De Carvalho, D. et al., 2012. DNA Methylation Screening Identifies Driver Epigenetic Events of Cancer Cell Survival. *Cancer Cell*, 21(5), pp.655–667. Available at: http://dx.doi.org/10.1016/j.ccr.2012.03.045.

Castelo-Branco, P. et al., 2013. Methylation of the TERT promoter and risk stratification of childhood brain tumours: An integrative genomic and molecular study. *The Lancet Oncology*, 14(6), pp.534–542.

Cedar, H. & Bergman, Y., 2011. Epigenetics of haematopoietic cell development. *Nature Reviews Immunology*, 11(7), pp.478–488. Available at: http://dx.doi.org/10.1038/nri2991.

Chang, C.-P. & Bruneau, B.G., 2012. Epigenetics and Cardiovascular Development. *Annual Review of Physiology*, 74(1), pp.41–68. Available at: http://www.annualreviews.org/doi/10.1146/annurev-physiol-020911-153242.

Chao, W.R. et al., 2015. Unusual c-KIT (+) squamous cell carcinoma of the uterine cervix showing paradoxical hypermethylation of the c-KIT proto-oncogene. *European Journal of Obstetrics Gynecology and Reproductive Biology*, 184, pp.130–131. Available at: http://dx.doi.org/10.1016/j.ejogrb.2014.11.034.

Cohen, A.S. et al., 2016. Delta-Opioid Receptor (δOR) Targeted Near-Infrared Fluorescent Agent for Imaging of Lung Cancer: Synthesis and Evaluation In Vitro and In Vivo. *Bioconjugate Chemistry*, 27(2), pp.427–438. Available at: http://pubs.acs.org/doi/10.1021/acs.bioconjchem.5b00516.

Colaprico, A. et al., 2016. TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, 44(8), p.e71.

Colaprico, A.A. et al., 2018. *Package ' TCGAbiolinks ,'*

Cost, L., 2012. HumanMethylation450 BeadChip. *Illumina*, pp.1–4. Available at: papers2://publication/uuid/5ED8F284-E418-41CB-B648-D9DE0684ED66.

Costa-Pinheiro, P. et al., 2015. Diagnostic and prognostic epigenetic biomarkers in cancer. *Epigenomics*, 7(6).

Cousineau, D., 2011. Outliers detection and treatment : a review . , 3(1), pp.58–67.

Dawson, M.A. & Kouzarides, T., 2012. Cancer epigenetics: From mechanism to therapy. *Cell*, 150(1), pp.12–27. Available at: http://dx.doi.org/10.1016/j.cell.2012.06.013.

Fawcett, T., 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), pp.861–874.

Ferlay, J. et al., 2018. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*. Available at: https://doi.org/10.1016/j.ejca.2018.07.005.

Ferté, C., André, F. & Soria, J.C., 2010. Molecular circuits of solid tumors: Prognostic and predictive tools for bedside use. *Nature Reviews Clinical Oncology*, 7(7), pp.367–380. Available at: http://dx.doi.org/10.1038/nrclinonc.2010.84.

Fujiwara, K. et al., 2005. Identification of Epigenetic Aberrant Promoter Methylation in Serum DNA Is Useful for Early Detection of Lung Cancer Identification of Epigenetic Aberrant Promoter Methylation in Serum DNA Is Useful for Early Detection of Lung Cancer. , 11, pp.1219–1225.

Galindo Villardón, M.P., 1985. Una alternativa de Reprresentation Simultánea: HJ-Biplot. *Questíio*, 10(1), pp.13–23.

Ge, Y.Z. et al., 2004. Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. *Journal of Biological Chemistry*, 279(24), pp.25447–25454.

Greiner, M., Pfeiffer, D. & Smith, R.D., 2000. Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine*, 45(1–2), pp.23–41.

Ha, M. & Kim, V.N., 2014. Regulation of microRNA biogenesis. *Nature Reviews Molecular Cell Biology*, 15(8), pp.509–524. Available at: http://dx.doi.org/10.1038/nrm3838.

Hair, J.F. et al., 1999. *Análise Multivariada* 5th ed. A. Otero, ed., Madrid: Pearson educación, S. A.

Halevi, S. et al., 2003. Conservation within the RIC-3 Gene Family. *Journal of Biological Chemistry*, 278(36), pp.34411–34417. Available at: http://www.jbc.org/lookup/doi/10.1074/jbc.M300170200.

Hanahan, D., Weinberg, R. a & Francisco, S., 2000. The Hallmarks of Cancer Review University of California at San Francisco. , 100, pp.57–70.

Hanahan, D. & Weinberg, R.A., 2011. Hallmarks of cancer: The next generation. *Cell*, 144(5), pp.646–674. Available at: http://dx.doi.org/10.1016/j.cell.2011.02.013.

Hartl, D.L., 2014. *Essential Genetics: A Genomics Prespective* 6th ed., Burlington: Jones e Bartlett Publishers Inc.

Hermann, A., Schmitt, S. & Jeltsch, A., 2003. The human Dnmt2 has residual DNA-(Cytosine-C5) methyltransferase activity. *Journal of Biological Chemistry*, 278(34), pp.31717–31721.

Hervouet, E., Vallette, F.M. & Cartron, P.F., 2009. Dnmt3/transcription factor interactions as crucial players in targeted DNA methylation. *Epigenetics*, 4(7), pp.487–499.

Illumina, 2010. HiSeq^{TM} 2000 Sequencing System. *Specification Sheet: Illumina® Sequencing*, pp.1–4. Available at: https://www.illumina.com/documents/products/datasheets/datasheet_hiseq2000.pdf.

Illumina Inc, 2010. GenomeStudio Methylation Module v1.8 User Guide (11319130B). *Illumina documentation*, (November).

Klose, R.J. & Bird, A.P., 2006. Genomic DNA methylation: The mark and its mediators. *Trends in Biochemical Sciences*, 31(2), pp.89–97.

Kneip, C. et al., 2011. SHOX2 DNA methylation is a biomarker for the diagnosis of lung cancer in plasma. *Journal of Thoracic Oncology*, 6(10), pp.1632–1638. Available at: http://dx.doi.org/10.1097/JTO.0b013e318220ef9a.

Kooistra, S.M. & Helin, K., 2012. Molecular mechanisms and potential functions of histone demethylases. *Nature Reviews Molecular Cell Biology*, 13(5), pp.297–311. Available at: http://www.nature.com/doifinder/10.1038/nrm3327.

Kumar Mishra, N. & Guda, C., 2017. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. *Oncotarget*, 8(17), pp.28990–29012. Available at: http://www.oncotarget.com/fulltext/15993.

Laird, P.W., 2003. The power and the promise of DNA methylation markers. *Nature Reviews Cancer*, 3(4), pp.253–266.

Lazebnik, Y., 2010. What are the hallmarks of cancer? *Nature Reviews Cancer*, 10(4), pp.232–233. Available at: http://dx.doi.org/10.1038/nrc2827.

Levene, H., 1960. *Contributions to Probability and Statistics*, Redwood City: Stanford University Press.

Li, B. & Dewey, C.N., 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12.

Lin, S. & Gregory, R.I., 2015. MicroRNA biogenesis pathways in cancer. *Nature Reviews Cancer*, 15(6), pp.321–333. Available at: http://dx.doi.org/10.1038/nrc3932.

Lokk, K. et al., 2016. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biology*, 17(1).

Manikandan, S., 2011. Measures of central tendency: Median and mode. *Journal of Pharmacology and Pharmacotherapeutics*, 2(3), p.214. Available at: http://www.jpharmacol.com/text.asp?2011/2/3/214/83300.

Mariani, F., Sena, P. & Roncucci, L., 2014. Inflammatory pathways in the early steps of colorectal cancer development. *World Journal of Gastroenterology*, 20(29), pp.9716–9731.

Mazzaferri, E.L. & Kloos, R.T., 2001. Current Approaches to Primary Therapy for Papillary and Follicular Thyroid Cancer. *The Journal of Clinical Endocrinology & Metabolism*, 86(4), pp.1447–1463. Available at: https://academic.oup.com/jcem/article-lookup/doi/10.1210/jcem.86.4.7407.

Moon, M. & Nakai, K., 2018. Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers. *Journal of Bioinformatics and Computational Biology*, 16(2, SI).

Nikolaidis, G. et al., 2012. DNA methylation biomarkers offer improved diagnostic efficiency in lung cancer. *Cancer Research*, 72(22), pp.5692–5701.

Okano, M., Xie, S. & Li, E., 1998. Dnmt2 is not required for de novo and maintenance methylation of viral DNA in embryonic stem cells. *Nucleic Acids Research*, 26(11), pp.2536–2540.

Petkowski, J.J. et al., 2013. NRMT2 is an N-terminal monomethylase that primes for its homolog NRMT1. *Biochem J.*, 456(3), pp.453–462.

Razali, N.M. & Wah, Y.B., 2011. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1), pp.21–33.

Reik, W. & Walter, J., 2001. Genomic imprinting: Parental influence on the genome. *Nature Reviews Genetics*, 2(1), pp.21–32.

Robin, X. et al., 2018. *Package ' pROC ,'*

Robin, X. et al., 2011. pROC: an open source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12(Swets 1973), p.77.

Shapiro, S.S. & Wilk, M.B., 1965. An Analysis of Variance Test for Normailty (Complete Samples). *Biometrika*, 52(3–4), pp.591–611.

Siegmund, K., Methylation Array Data Analysis Tips. Available at: https://emea.illumina.com/techniques/microarrays/methylation-arrays/methylation-array-data-analysis-tips.html [Accessed August 6, 2018].

Silva, T.C. et al., 2016. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research*, 5(0), p.1542. Available at: https://f1000research.com/articles/5-1542/v2.

Simon, J. et al., 2004. Analysis of the Set of GABAA Receptor Genes in the Human Genome. *Journal of Biological Chemistry*, 279(40), pp.41422–41435. Available at: http://www.jbc.org/cgi/doi/10.1074/jbc.M401354200.

Simon, K., 2016. Colorectal cancer development and advances in screening. *Clinical Interventions in Aging*, 11, pp.967–976.

Slattery, M.L., Lundgreen, A. & Wolff, R.K., 2012. MAP kinase genes and colon and rectal cancer. *Carcinogenesis*, 33(12), pp.2398–2408.

Stewart, BWKP and Wild, C.P. and others, 2014. *World cancer report 2014* B. Stewart & C. Wild, eds., Lyon.

Tomczak, K., Czerwińska, P. & Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia*, 1A, pp.A68–A77.

Vicente-villardon, J.L., 2015. *Package ' MultBiplotR ,'*

Victoria Valinluck Lao and William M. Grady, 2011. Epigenetic and Colorectal Cancer. *Nat rev Gastroenterol Hepatol*, 8(12), pp.686–700.

Vogel, T. & Lassmann, S., 2014. Epigenetics: Development, dynamics and disease. *Cell and Tissue Research*, 356(3), pp.451–455.

Wang, Y. et al., 2016. The identification of age-associated cancer markers by an integrative analysis of dynamic DNA methylation changes. *Scientific Reports*, 6(March). Available at: http://dx.doi.org/10.1038/srep22722.

Warren, J.D. et al., 2011. Septin 9 methylated DNA is a sensitive and specific blood test for colorectal cancer. *BMC Medicine*, 9.

Wei, J. et al., 2016. Integrated analysis of genome-wide DNA methylation and gene expression profiles identifies potential novel biomarkers of rectal cancer. *Oncotarget*, 7(38), pp.62547–62558. Available at: http://www.oncotarget.com/fulltext/11534.

Weinberg, R.A., 2014. *The Biology of Cancer_2nd edition* 2nd ed. Garland Science, ed.,

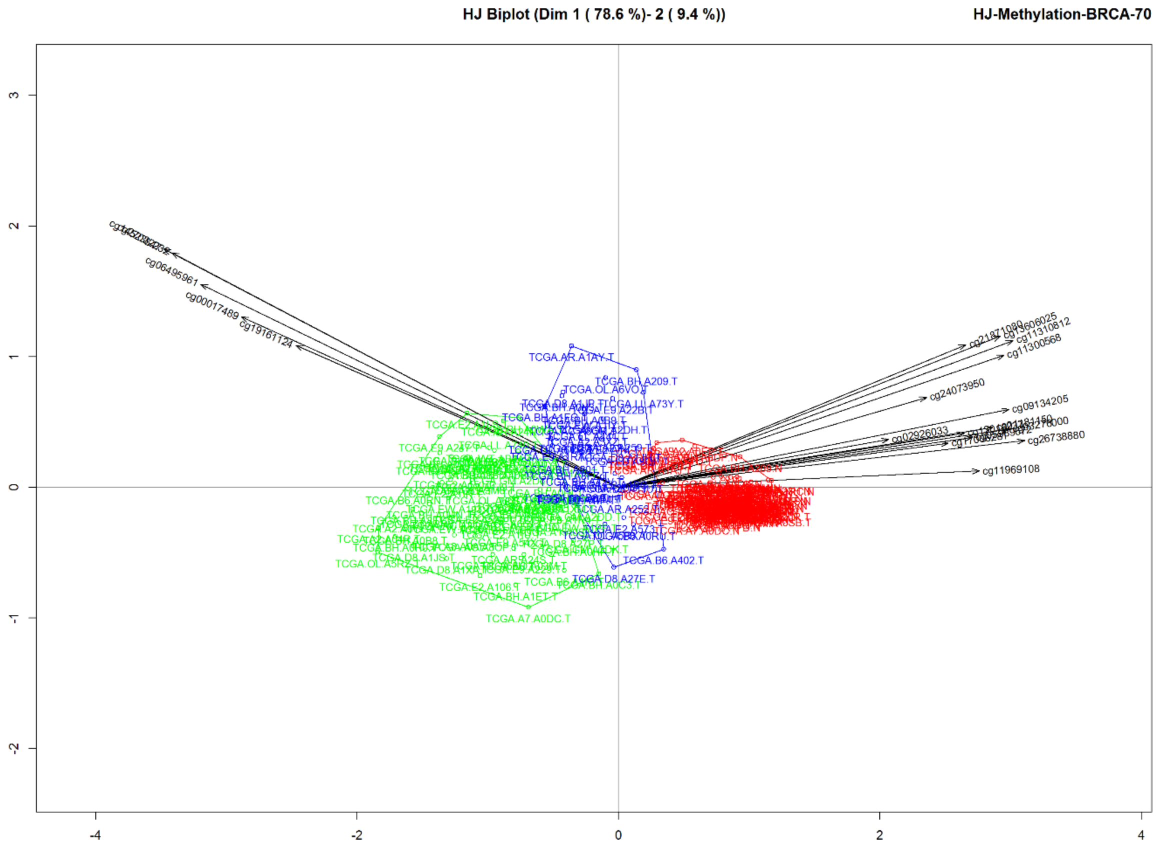
Welch, B.L., 1947. The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, 34(1/2), p.28. Available at: http://www.jstor.org/stable/2332510?origin=crossref.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), pp.80–83.

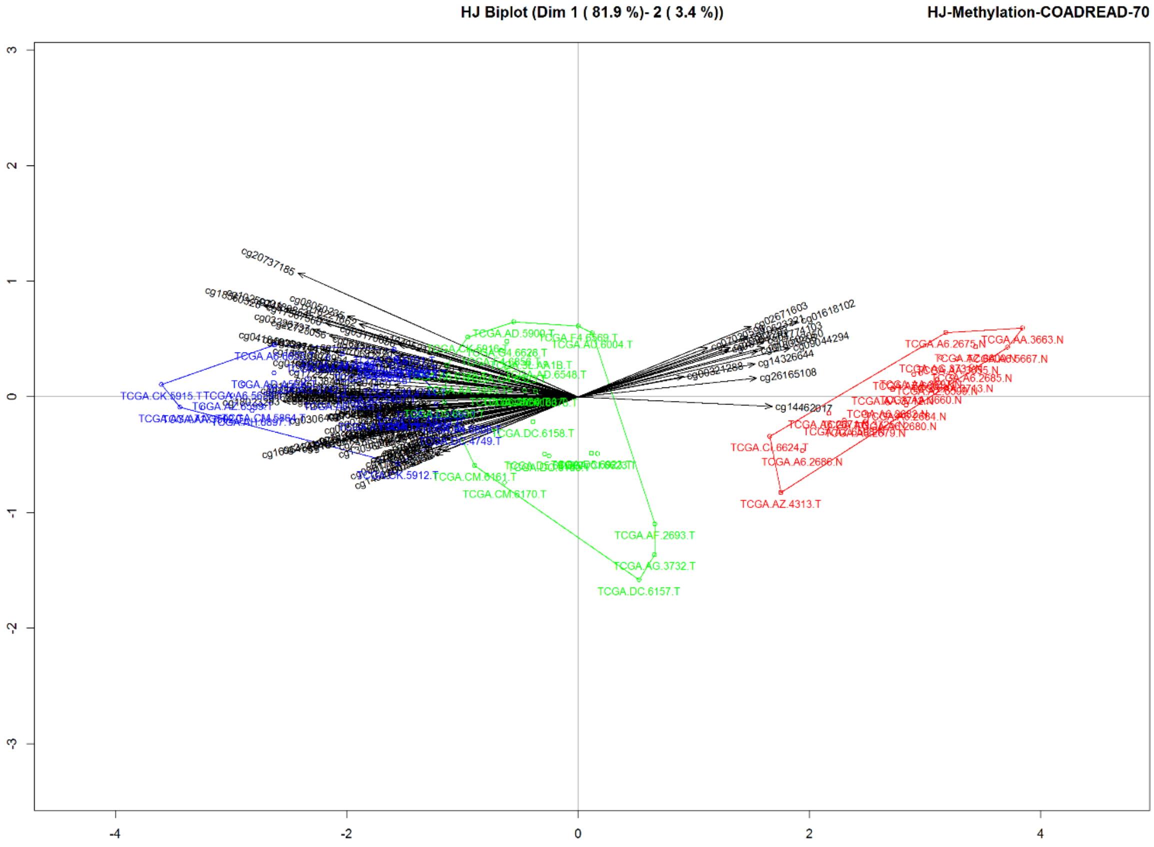
Yoo, C.B. & Jones, P.A., 2006. Epigenetic therapy of cancer: Past, present and future. *Nature Reviews Drug Discovery*, 5(1), pp.37–50.

Zhang, C. et al., 2015. The identification of specific methylation patterns across different cancers. *PLoS ONE*, 10(3), pp.1–16.

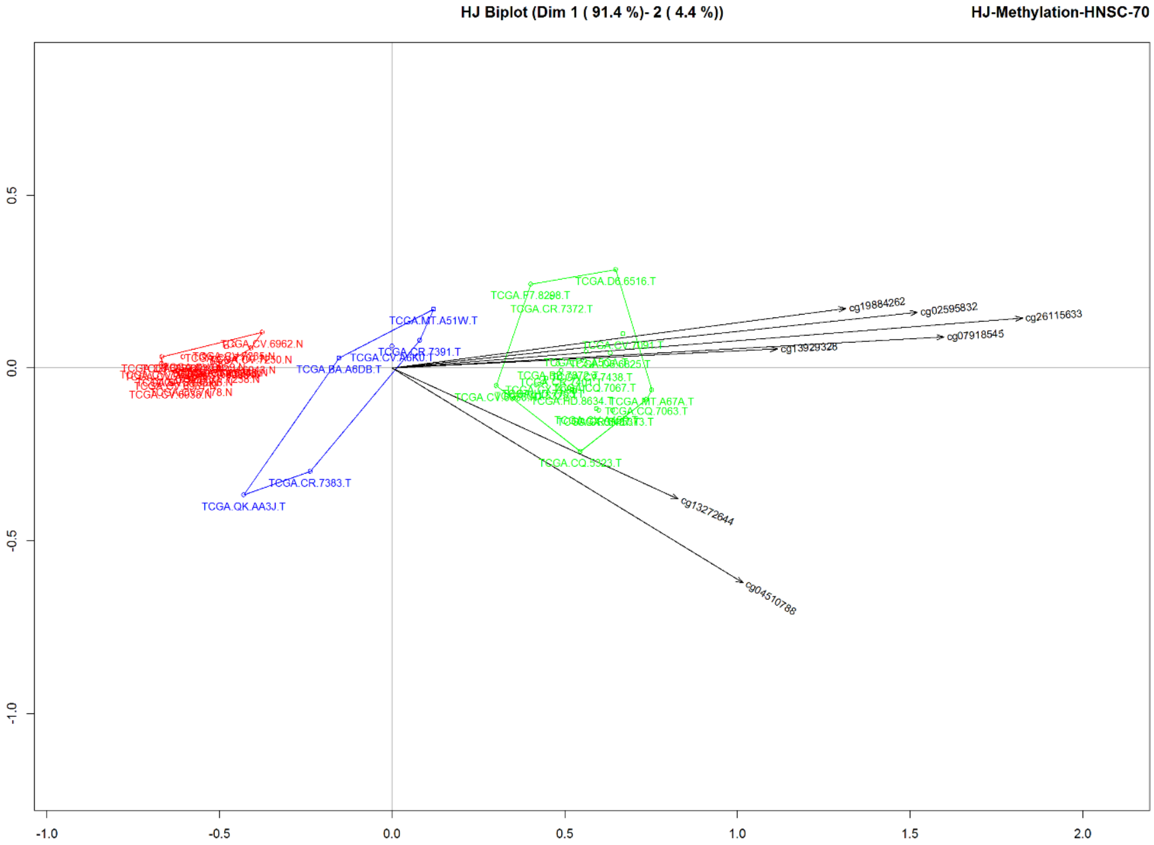
ANNEXES



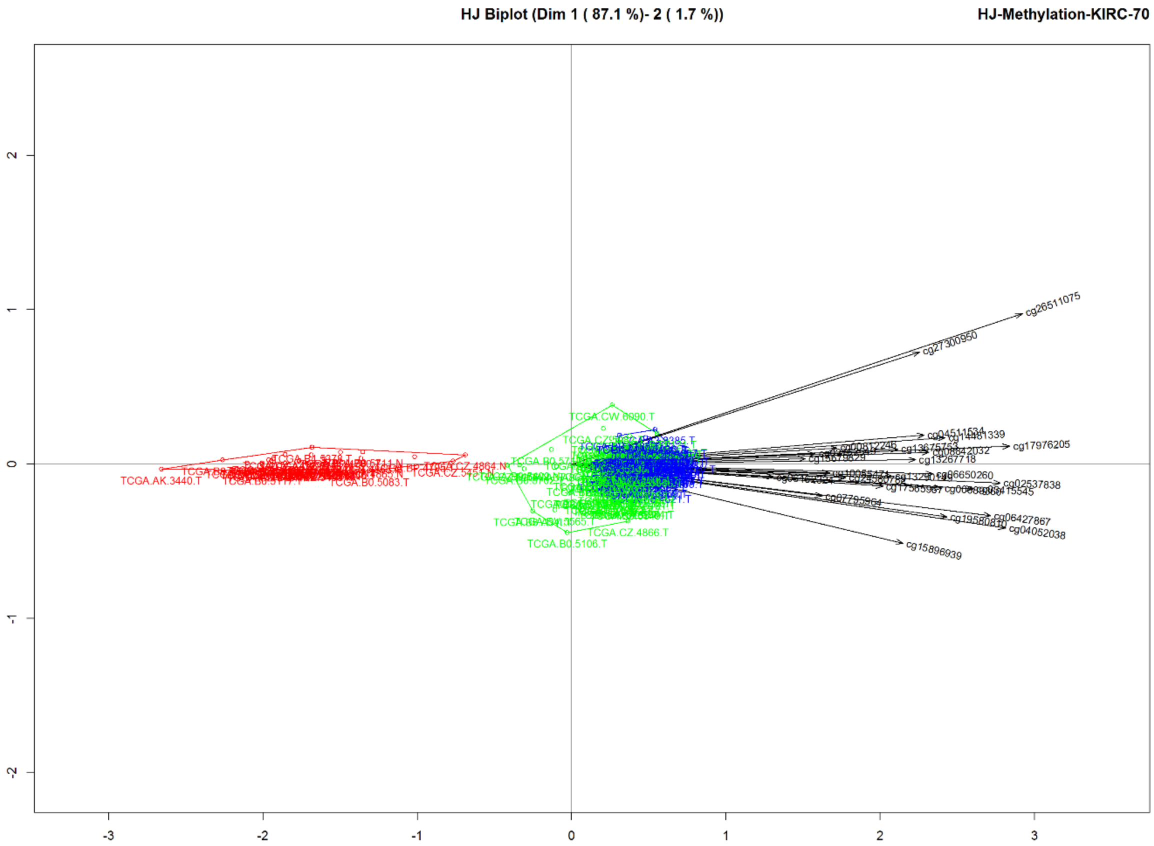
Annex 1 - HJ-biplot for DNA methylation in breast cancer. This representation results from a variable selection with more than 70% of contribution, retaining 88% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 84 normal and 14 tumor samples, C2 with 75 tumor samples and C3 with 37 tumor samples.



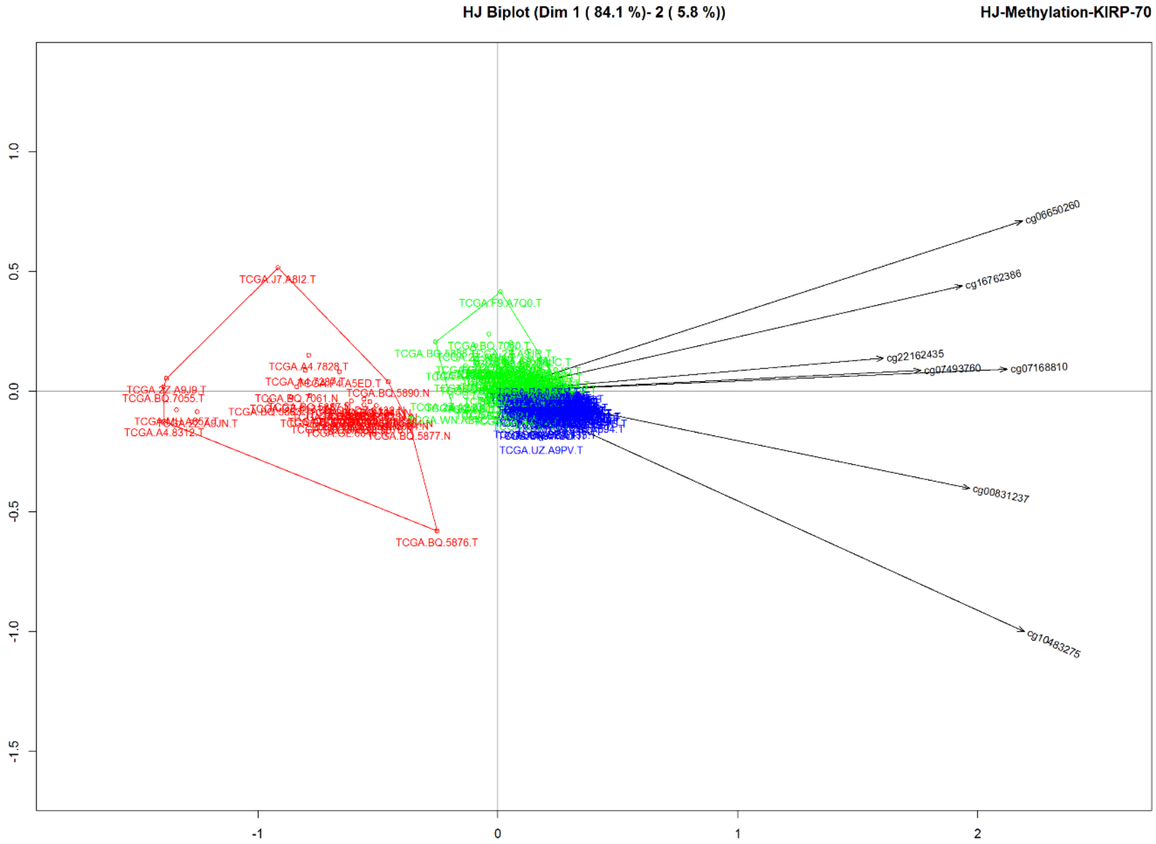
Annex 2 – HJ-biplot for DNA methylation in colorectal cancer. This representation results from a variable selection with more than 70% of contribution, retaining 85.3% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 21 normal and 2 tumor samples, C2 with 25 tumor samples and C3 with 27 tumor samples.



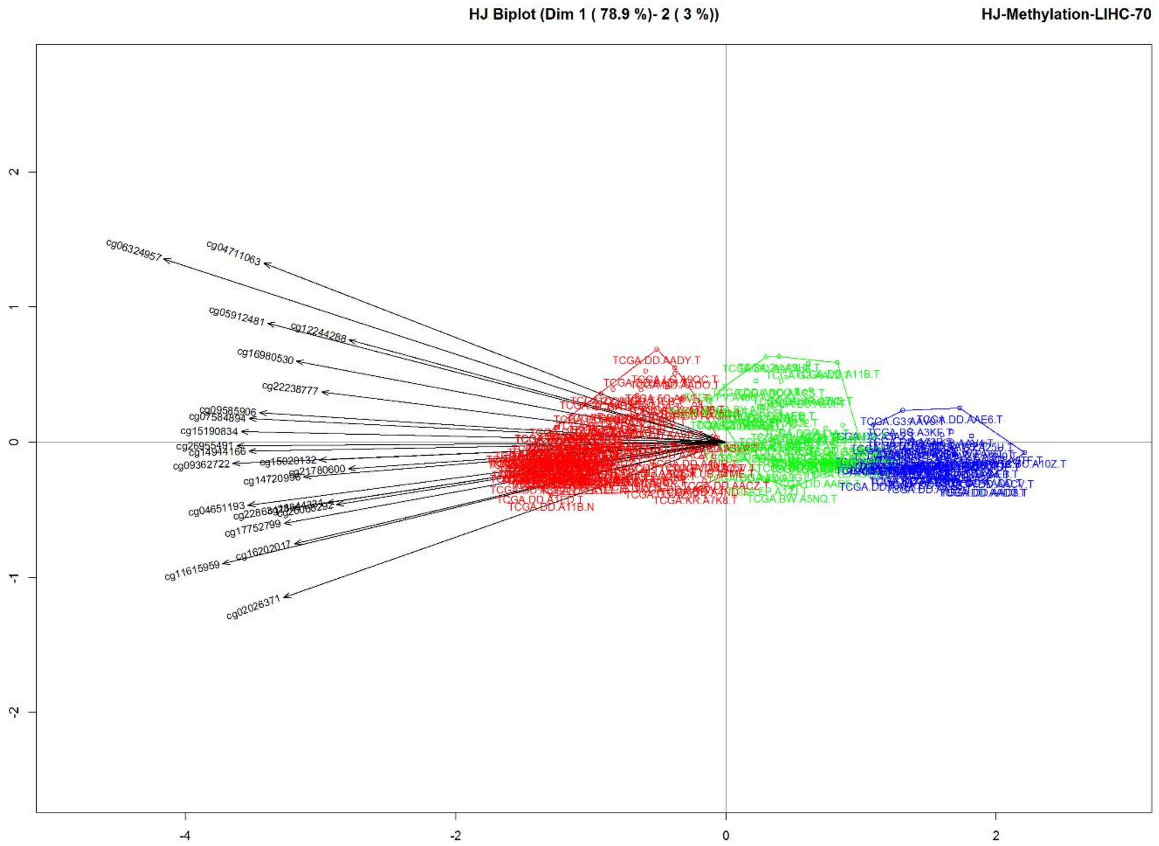
Annex 3 – HJ-biplot for DNA methylation in head and neck cancer. This representation results from a variable selection with more than 70% of contribution, retaining 95.8% of variance. Hierarchical clusters are represented in red (C1), blue (C2) and green (C3). C1 with 19 normal and 1 tumor samples, C2 with 6 tumor samples and C3 with 1 normal and 20 tumor samples.



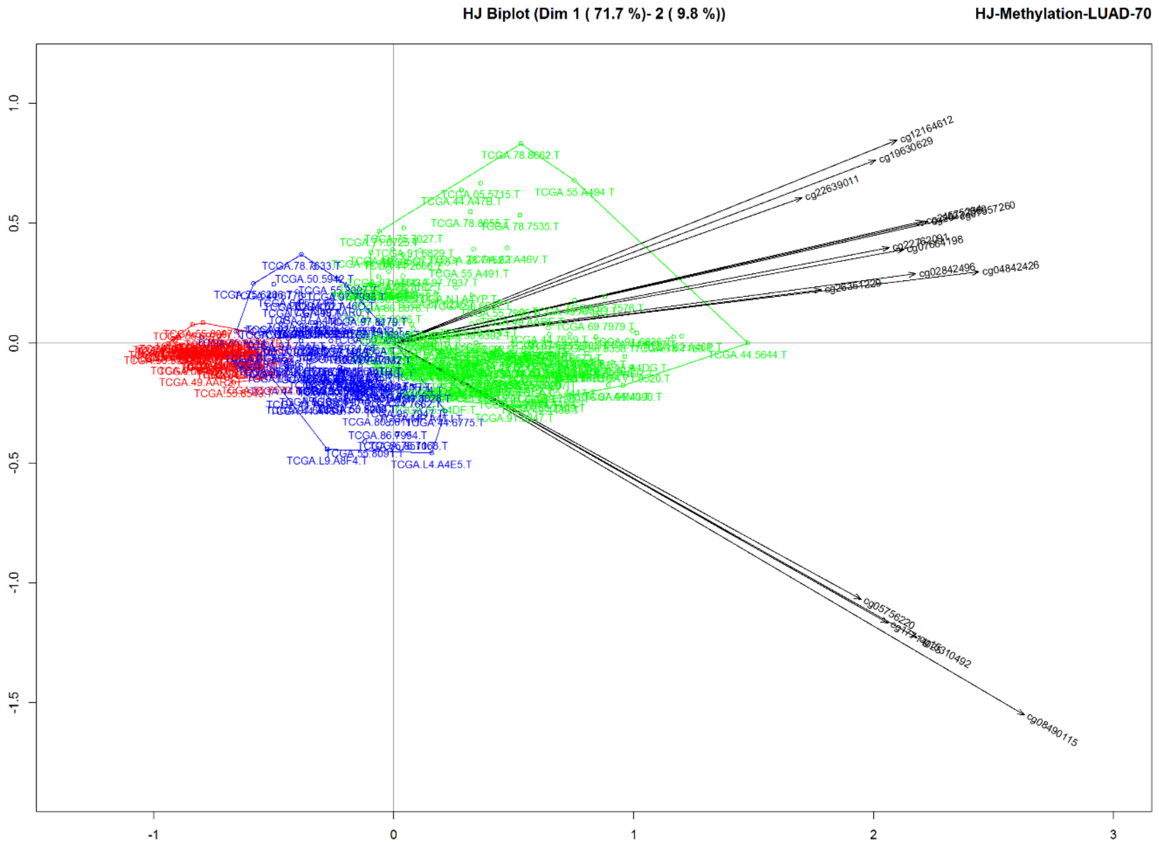
Annex 4 – HJ-biplot for DNA methylation in kidney$_R$ cancer. This representation results from a variable selection with more than 70% of contribution, retaining 88.8% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 21 normal and 12 tumor samples, C2 with 3 normal and 74 tumor samples and C3 with 69 tumor samples.



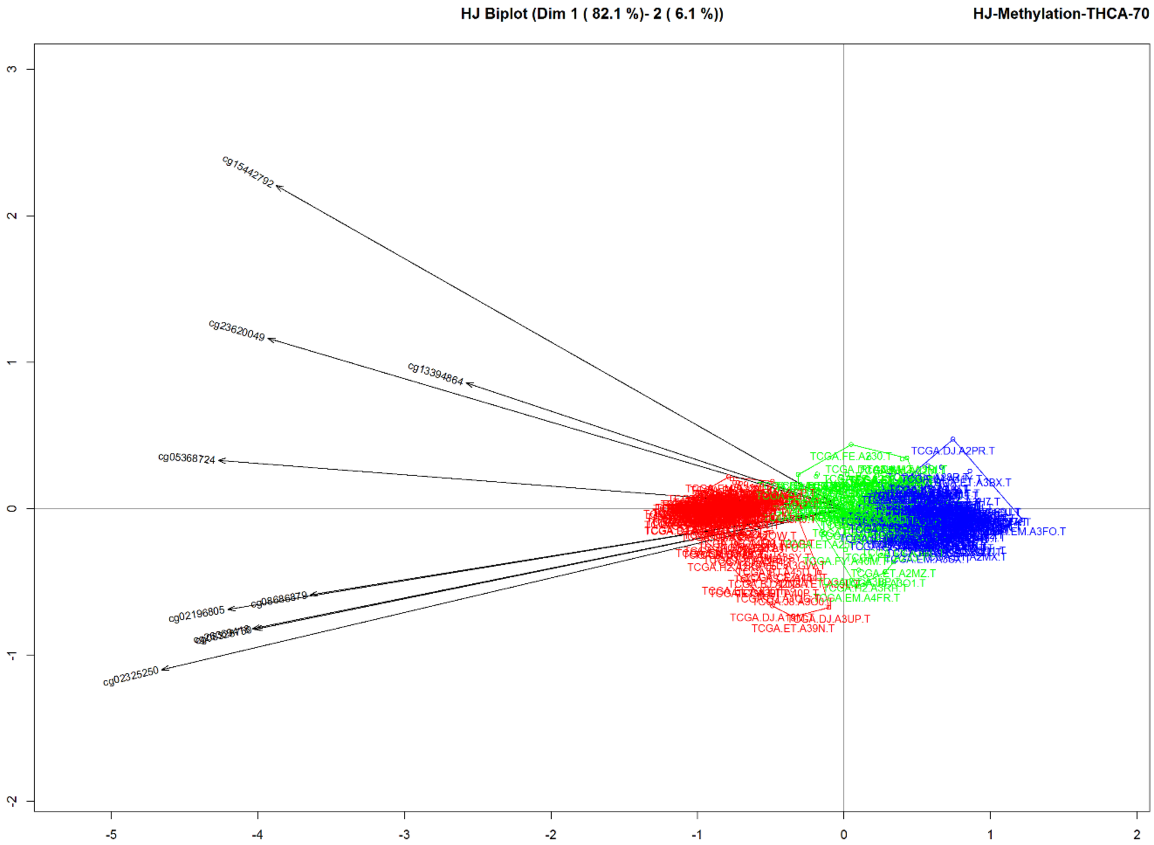
Annex 5 – HJ-biplot for DNA methylation in kidney$_P$ cancer. This representation results from a variable selection with more than 70% of contribution, retaining 89.9% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 23 normal and 11 tumor samples, C2 with 63 tumor samples and C3 with 93 tumor samples.



Annex 6 – HJ-biplot for DNA methylation in liver cancer. This representation results from a variable selection with more than 70% of contribution, retaining 81.9% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 41 normal and 70 tumor samples, C2 with 49 tumor samples and C3 with 52 tumor samples.



Annex 7 – HJ-biplot for DNA methylation in lung cancer. This representation results from a variable selection with more than 70% of contribution, retaining 81.5% of variance. Hierarchical clusters are represented in red (C1), blue (C2) and green (C3). C1 with 21 normal and 24 tumor samples, C2 with 135 tumor samples and C3 with 86 tumor samples.



Annex 8 – HJ-biplot for DNA methylation in thyroid cancer. This representation results from a variable selection with more than 70% of contribution, retaining 88.2% of variance. Hierarchical clusters are represented in red (C1), green (C2) and blue (C3). C1 with 50 normal and 76 tumor samples, C2 with 77 tumor samples and C3 with 131 tumor samples.