Detecting Novel Subtypes of Cancer Using Bayesian Unsupervised Clustering

Sergio Llaneza Lago

Norwich Medical School
University of East Anglia

A thesis submitted to Norwich Medical School at the University of East Anglia in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

August 2023

Word Count: 56993

Abstract

Although there have been many advances in screening programs and treatments in recent years that have reduced the mortality rate of cancer, it remains the second leading cause of death worldwide, accounting for almost 10 million deaths worldwide in 2020. Identifying and characterising subtypes based on molecular classifications can help identify the aggressiveness of the disease so that the best treatment pathway can be identified, and new treatment options developed. This has been exemplified in breast cancer. Latent Process Decomposition (LPD) is a soft clustering technique that has been successfully applied to expression data to discover subtypes, including a poor prognosis subtype called DESNT. The benefit of LPD is that it better models the heterogenous structure of tumours.

The aim of this thesis is to apply LPD on transcriptome data from The Cancer Genome Atlas to detect and characterise subtypes of numerous cancer types and create a resource of the results. This was achieved through the development of Automata, an R package used to automate this methodology.

In total I have identified 168 cancer subtypes spanning across 28 cancer types. Moreover, I have characterised the features of each subtype, generating a unique encyclopaedic compendium of molecular subtypes of cancer that provides an in-depth source of information for the research community. I have successfully validated my findings by comparing them with known subtypes from breast carcinoma, prostate adenocarcinoma, colorectal adenocarcinoma and lung cancer. Additionally, I have discovered common features that characterise subtypes across cancer types. Finally, I have identified 26 subtypes which have a significant association with outcome including some that were not picked up by traditional clustering methods.

The results presented in this thesis are the foundation for the long-term impact of a more personalised approach to cancer patient care.

Acknowledgements

I feel grateful to Professor Daniel Brewer for his astounding guidance and continuous support during the past four years. I am also thankful to Professor Colin Cooper for his insights and supervision of my research. Furthermore, I am indebted to Professor Vincent Moulton for sharing his expertise and support for this project.

Big thanks to the Cancer Genetics team of UEA. Their support and company in the past years helped me pull through and finish this project.

Furthermore, I would like to dedicate this research to my loving family. Without their love and constant support from the very first years of moving to the UK, I would not stand where I am now. I wish to dedicate this work to the loving memory of my father. His support and affection will always be with me.

I would also like to thank my friends back in Spain. Sergio, Fer and Sandra have kept me company throughout these years, despite the distance between us. But also, my friends in Norwich. Claudia, Tasos, Ryan, Monica and many others that I was fortunate to meet and share memories with.

To my dear Agapimú. Thank you for your support in the past two years. I would not be able to do this without you. Thank you for being with me in my worst and my best moments, for the never-ending phone calls, and for your patience.

Finally, I would like to thank the Big C. Without their generous funding, this research would not have come to fruition.

Table of Contents

List of Figures

List of Tables

Acronyms

Acronym	Description
ρ	Spearman's rank correlation coefficient
ADT	Androgen Deprivation Therapy
AJCC	American Joint Committee on Cancer
ANOVA	Analysis of Variance
BCR	Biospecimen Core Resource
BH	Benjamin-Hochberg adjustment
CGC	Cancer Gene Census
CMS	Consensus Molecular Subgroups
COSMIC	Catalogue of Somatic Mutations in Cancer
CRPC	Castration Resistant Prostate Cancer
DCC	Data Coordinating Centre
DEG	Differentially expressed gene
DMG	Differentially methylated gene
DNA	Deoxyribonucleic Acid
DRE	Digital Rectal Exam
ER	Estrogen Receptor
GCC	Genome Characterisation Centre
GO	Gene Ontology
GRC	Genome Reference Consortium
GSC	Genome Sequencing Centre
GSEA	Gene set enrichment analysis
HBOC	Hereditary Breast and Ovarian Cancer
HPC	High Powering Research Clustering
IHC	Inmunohistochemistry
IQR	Interquartile range

ITH	Intra-tumour heterogeneity
KEGG	Kyoto Encyclopedia of Genes and Genomes
LPD	Latent Process Decomposition
MRI	Magnetic Resonance Imaging
PI	Proximal-Inflammatory
PP	Proximal-Proliferative
PR	Progesterone receptor
PSA	Proste-Specific Antigen
RNA	Ribonucleic Acid
RR	Relative Risk
TCGA	The Cancer Genome Atlas
TNM	Tumour - Nodes - Metastases
TRU	Terminal Respiratory Unit
TSS	Tissue Source Site
VST	Variance Stabilising Transformation
WXS	Whole Exome Sequencing
r	Pearson's correlation coefficient
pre-mRNA	pre-messenger Ribonucleic Acid

Chapter 1

Introduction

1.1 Summary

In this chapter, I provide an overview of the key biological and medical concepts and methodologies relevant to identifying cancer subtypes. I explain the biological mechanisms behind the cancer disease and briefly describe the clinical aspects and current subtype perspectives of breast cancer, prostate cancer, colorectal cancer, lung adenocarcinoma, and lung squamous cell carcinoma. The database utilised as the data source for this thesis is also disclosed. Finally, the applications of machine learning are briefly reviewed, with an emphasis on the method used in this thesis.

1.2 Cancer

Cancer is a collection of related genetic diseases characterised by the uncontrolled growth and limitless division of abnormal human cells[1]. These cells replace their normal functions for others, such as the capacity to spread into surrounding healthy tissues and generate tumours. Cancer is a global problem. The WHO (2019)[2] estimates that it is the second leading cause of death in the world (Fig. 1.1). In 2018, 17 million cases and 9.6 million deaths were reported worldwide[3]. A report published by DEMOS in 2020[4] estimates an annual cost of £7.6bn to the UK economy due to this illness.

Cancer originates from modifications in the genome that lead to alterations in gene expression and regulation processes. Genome alterations can occur naturally during cellular division, can be inherited from the parents, and can be derived from exposure to environmental and lifestyle factors – ultraviolet radiation, tobacco smoke, and infectious agents[6,7]. Most of these alterations (termed mutations) are repaired by DNA repair mechanisms, but occasionally they persist, and even more rarely, they confer an environmental advantage to the cell through gained characteristics (see section 1.2.1)[8]. Malignant cells gain more and more mutations, increasing the genetic variability of the tumour in the long term[9]. When a cell acquires enough advantages, it gains the capacity to proliferate and invade tissues[8]. This leads to cancer cells spreading around the body (metastasis) and eventually death.

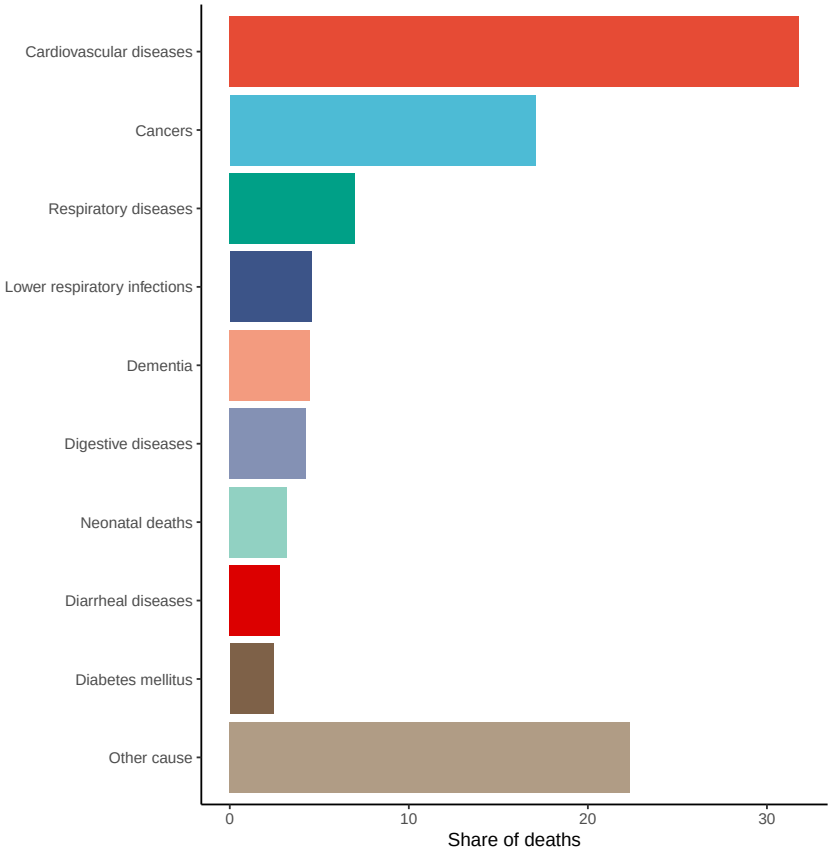

Figure 1.1: Share of deaths by cause worldwide in 2017. Cancers are the second leading cause of death after cardiovascular diseases. Adapted from Ritchie (2018)[5].

1.2.1 Hallmarks of cancer

Hanahan and Weinberg (2000)[10] described six hallmarks commonly present in cancer cells: self-sufficiency to growth signals, insensitivity to anti-growth signals, evasion of the cell programmed death, limitless replicative potential, continuous generation of blood vessels, and tissue invasion and metastasis. An updated review in 2011 described four more hallmarks: avoidance of the immune system, deregulated metabolism, promotion of the inflammatory response, and genome instability[11].

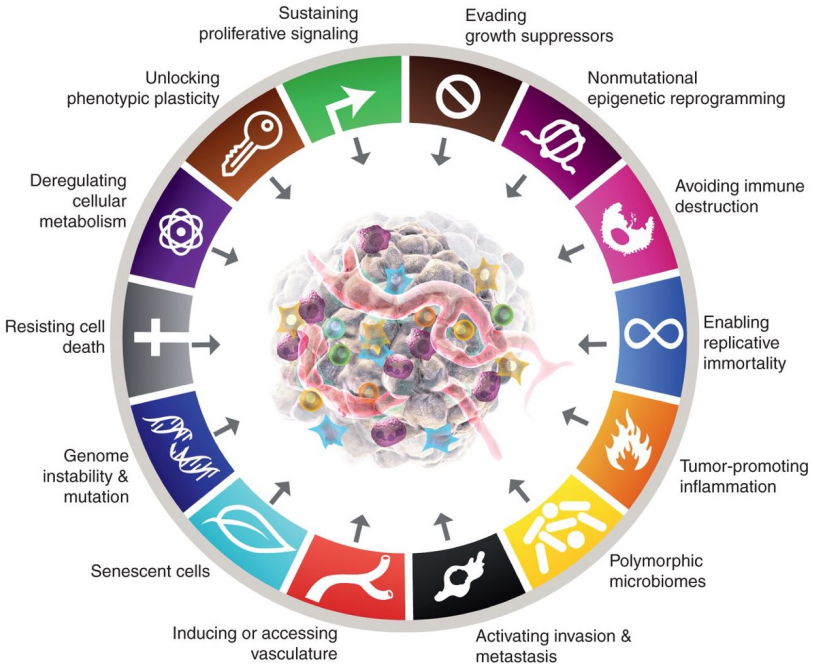

Figure 1.2: Schematic representation of the hallmarks of cancer. Obtained from Hanahan (2022)[12].

In other words, cancer cells produce their own growth signals and inhibit any type of external regulation mechanisms so they can develop and divide independently of the control exerted by the human body. Additionally, they modify their antigen presentation to either target cells related to the immune system or stay dormant for long periods so they can avoid an immune response[13]. Malignant cells lose the ability to program their own death (apoptosis) when damaged, infected, or no longer required. Also, they bypass the limited number of times normal cells can divide, which causes an accumulation of alterations, leading to high genomic instability.

Due to their unlimited growth, tumours require a large amount of oxygen and nutrients that are provided by forming new blood vessels (angiogenesis). Tumours enhance this process by promoting an inflammatory response and inducing the metabolic production of angiogenesis stimulators such as lactate[14]. In the later stages of the disease, tumours invade surrounding tissues and break free, travelling through the body via lymph or blood vessels to expand into other organs.

Recently, Hanahan (2022)[12] defined four new hallmarks, termed enabling characteristics, that represent the molecular and cellular mechanisms by which the previous ten hallmarks

are acquired: unlocking phenotypic plasticity, non-mutational epigenetic reprogramming, polymorphic microbiomes, and senescent cells (Fig 1.2)[12]. Unlocking phenotypic plasticity refers to the phenomenon in which embryonic cells gradually lose their ability to produce different phenotypes during embryo development. This supposes an obstacle to the uncontrolled growth and spread of cancer; hence, malignant cells can reverse this process. Growing evidence supports the non-mutational epigenetic reprogramming hallmark that suggests that carcinogenic genome modifications are not only caused by genome instability and mutations but also by epigenetic regulated mechanisms. The third condition, polymorphic microbiomes, refers to the symbiotic associations between microorganisms and the internal organs and how they can positively and negatively affect cancer development. The last condition alludes to the role of the senescent cells (cells that cannot divide again) in the tumour environment. The senescence process has long been viewed as a protective mechanism against cancer since it prevents illimitable replication. However, increasing evidence suggests the opposite[15,16].

1.3 Genes and gene expression

To completely understand the causes and consequences of tumours, we first need a solid grasp of the molecular mechanisms involved in gene expression. Genetic information is stored in deoxyribonucleic acid (DNA), a helicoidal double-stranded macromolecule located in the nucleus of cells[17]. DNA consists of a sequence of nucleotides, an organic molecule composed of a nitrogen base and a phosphate group. Depending on the nitrogenous base, nucleotides are classified into guanine, adenine, cytosine and thymine (G, A, C, and T, respectively). The DNA is organised into structures known as chromosomes which in turn are formed into functional subunits, in which the main type are genes. According to the central dogma of the molecular biology, the DNA is transcribed into ribonucleic acids (RNAs) and then translated into proteins, the functional units of the cell (Fig 1.3).

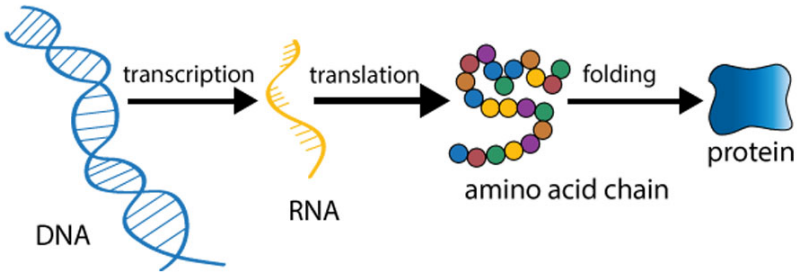

Figure 1.3: Gene expression process. DNA is transcribed into RNA, which in turn, is translated into proteins.

In human cells, the transcription process begins with the binding of transcription factor proteins to the *promotor* regions of the DNA located at the beginning of each gene[18]. This causes an opening in the helicoidal structure of the DNA that gives access to the polymerase RNA, responsible for generating a copy of the nucleotide chain known as pre-messenger RNA (pre-mRNA). This copy is an immature version of mRNA and contains both regions with protein-building information (exons) and non-protein-coding regions (introns)[18]. To reach the mature state, pre-mRNA undergoes a splicing process in which a selection of

exons joins together, and introns are removed[19]. Splicing is an important mechanism in cell development since the same pre-mRNA can result in multiple different mRNAs, increasing the protein diversity[20].

The resultant mRNA is exported from the nucleus of the cell to the ribosomes, located in the cytoplasm, to start the process of translation. The ribosomes read the chain of nucleotides in groups of three (codon) and synthesise a specific amino acid according to the genetic code[21]. Each newly-formed amino acid binds to the one previously synthesised, generating a chain. Once the mRNA has been read, the chain is released from the ribosomes and goes through a series of post-translational modifications to become a protein. The function of the proteins encompasses many roles such as structural component, immune defence, transport, signal transmission, storage, and catalyst of biochemical reactions[22].

The mRNA is only one of the multiple types of RNA involved in the process[21]. Non-protein-coding RNAs are also present in the cell and are classified according to their roles or size[23]: ribosomal RNA forms the ribosomes, transfer RNA carries the amino acids during translation, microRNA inhibits the translation of specific mRNA, small nuclear RNA regulates the removal of introns, and small interfering RNA degrades pre-mRNA.

1.4 Genetic and epigenetic alterations in cancer

Cancer is a disease of heritable changes in cells. Alterations in the genome cause abnormalities in the gene expression and, therefore, in the final protein produced by the cells. The most common abnormalities are genetic alterations –point mutations, indels, and numerical and structural alterations of chromosomes- and epigenetic alterations, which do not change the DNA sequence[24].

1.4.1 Small-scale mutations

Small-scale mutations affect one or a few nucleotides in the genome and are classified into single nucleotide variants (SNVs; a nucleotide is switched to another) or indels (a nucleotide or group of nucleotides are inserted or deleted).

The repercussions of small-scale gene mutation depend on how they affect the final protein[25]. Mutations in non-coding regions may affect transcription and prevent pre-mRNA generation or alter the splicing step. Mutations in coding regions can lead to *synonymous* mutations (the same amino acid is produced, occasionally with aberrant function) or *missense* mutations (the alteration causes a change to the amino acid sequence). Missense mutations that result in a change that stops the translation and generates a premature shortened protein are termed *nonsense* mutations[26]. Another possibility is that one or multiple nucleotides are inserted or deleted, which causes a change in the translation reading frame (*frameshift*) since all the triplets of nucleotides beyond that point are pushed or pulled along the sequence. However, frameshifts of three nucleotides are less likely to have an effect since they do not change the remaining protein sequence.

The most commonly mutated genes in cancer are the gene *TP53* that is a tumour-suppressor gene; the gene *BRAF* which is involved in cell division; the gene *JAK2* which participates in growth factor signalling; and the gene *KRAS* which is associated with several cancer types such as lung adenocarcinoma or colorectal carcinoma[27–31].

1.4.2 Chromosomal abnormalities

Chromosomes can be altered in the number of copies or in the structure. Structural changes refer to when a part is missing, there are extra parts, a part is switched, or a part is inverted[32]. Depending on whether the genetic material is equally exchanged between two chromosomes, structural changes can be reciprocal or nonreciprocal. These types of aberrations range in size and can affect from a small number of genes to all the genes on an entire chromosome[33]. Structural changes can lead to gene fusions when two independent genes are placed together (Fig 1.4)[34]. Fusion genes can have a higher expression level than normal genes, produce aberrant proteins, and are recurrent in prostate adenocarcinoma, lung cancer, and leukaemia[34,35]. Cell division errors can result in whole-genome duplication, contributing to genome instability and vulnerability to chromosomic abnormalities[36].

Chromosomal abnormalities occur at a higher rate in cancerous cells than in healthy cells[37]. The most common ones in cancer are gene amplifications –an increase of a specific part of a chromosome, thus, an overexpression of the affected genes[38]. Another common type is deletions, which result in the loss of genetic material.

Genes commonly affected by chromosomal abnormalities in cancer are the gene *EGFR* and *CDK12*, which are involved in cell division, cell proliferation, gene expression regulation, splicing, and DNA reparation[27,40,41].

1.4.3 Epigenetic alterations and DNA methylation

Epigenetic alterations are heritable changes that regulate the expression of genes without modifying the DNA sequence. In healthy cells, these alterations are controlled processes that work as mechanisms to regulate gene expression. However, when they malfunction and work improperly, they can contribute to the rise of various diseases, including cancer[42]. There are three main types of epigenetic alterations: DNA methylation, histone modifications, and RNA silencing. Histone modifications regulate the physical compression of the DNA chain, which determines whether the transcription machinery can access the genetic material. RNA silencing refers to the degrading effect of non-coding RNAs on mRNA transcripts.

Here we focus on DNA methylation, a mechanism in which a methyl group (CH3) is added to a gene of the DNA sequence, inhibiting its expression[43]. Cancer cells display an altered methylation profile that is passed to subsequent generations and is characterised by the presence of hypo and hypermethylated regions. Hypomethylation increases the possibility of genomic mutations by activating mechanisms that change the positions of some regions of the DNA and has been reported to be linked to early tumour development in breast cancer and tumour progression in ovarian cancer[44–46]. Hypermethylation inhibits the expression of tumour suppressor and DNA repair genes, which increases the number of mutations in the cell and leads to the acquirement of carcinogenic mutations and the hallmarks of cancer[42]. Examples of common hypermethylated tumour-supressor genes in breast cancer are *BRCA1* (DNA repair function), *PR* (hormone signalling pathways), and *ATM* (cell cycle regulation)[47].

1.4.4 Role of mutations in cancer development

Due to the continuous growth, cancer cells accumulate mutations at a higher rate than healthy cells[24]. Not all mutations change the function of a cell and provide an environmental

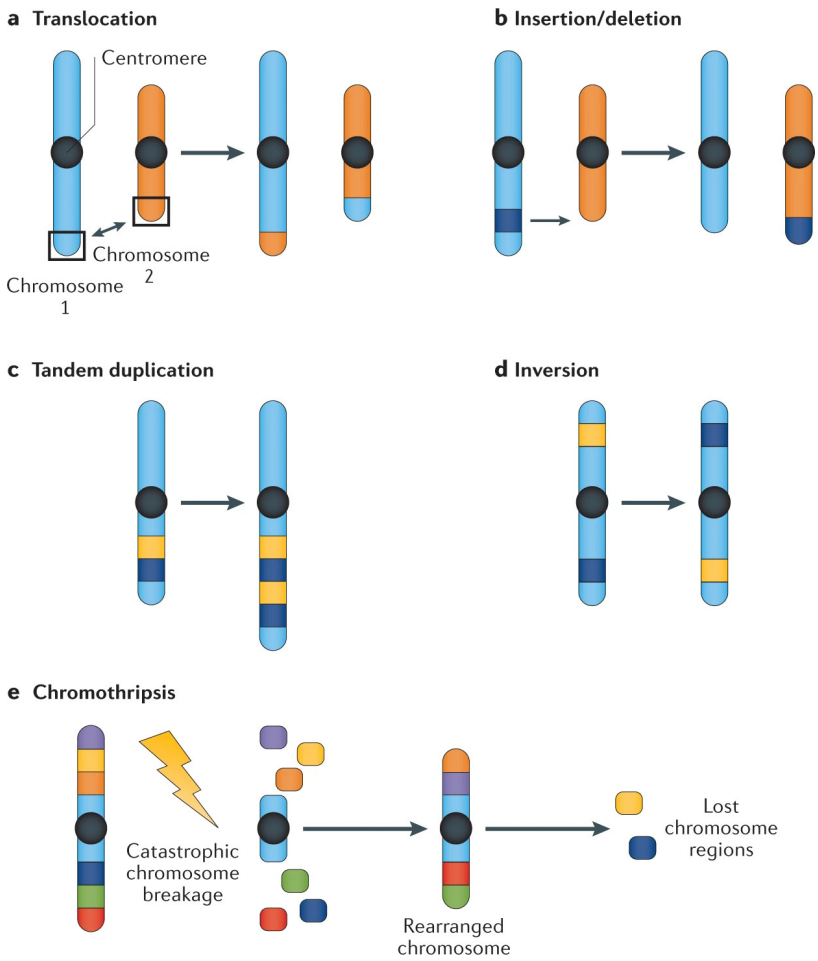

Figure 1.4: The different mechanisms that can cause a gene fusion. (a) Translocation, in which two chromosomes interchange a small portion. (b) Insertion/deletion is when a part of one chromosome breaks and is inserted into another chromosome. (c) Tandem duplication is when a chromosome section is duplicated and placed adjacent to the original. (d) Inversion is when a portion of the chromosome reinserts in the opposite direction. (e) Chromothripsis is when a chromosome is broken into multiple segments that are rearranged, usually with the consequent loss of some fragments. Obtained from Pederzoli et al. (2020)[39].

advantage (passenger mutations), and determining which ones do (driver mutations) is essential to understand cancer evolution and design new targeted therapies[48]. Weinberg (1996)[49] stated that two classes of genes when mutated contribute to carcinogenesis: proto-oncogenes and tumour suppressor genes. Proto-oncogenes participate in the development and division of the cell and are susceptible to mutations that stimulate their expression, such as chromosomal amplifications, gene fusions or whole genome duplications. On the contrary, tumour suppressor genes slow cell division, repair DNA mistakes and induce death cell. When both copies of a tumour suppressor gene are inhibited by disruptive point mutations, deletions, or hypermethylation, they cannot stop the carcinogenesis.

1.5 Types and subtypes of cancer

Although cancer is often referred as one single condition, it actually consists of more than 100 different diseases with shared features[50]. These diseases are classified according to the type of cell and tissue where the tumour originates from, leading to a plethora of types with their own traits and evolution. Thus, each cancer type possesses a unique behaviour, symptomatology, diagnosis, and response to treatment that leads to a wide range of outcomes.

The most common cancer types in England are breast carcinoma (~46,000 cases) followed by prostate adenocarcinoma (~40,500 cases), lung cancer (~37,500 cases), and colorectal adenocarcinoma (~34,000 cases) (Fig. 1.5)[51].

A single type of cancer can be further subdivided into subtypes shaped by distinct molecular pathways with their own composition and appearance, intrinsic molecular features, treatment pathways, aggressiveness, and prognosis[53]. Correct classification of subtypes has proved critical for optimising patient pathways, developing new treatments and moving the clinical practice towards treatment individualisation[53,54]. Traditionally, classification approaches were based on the architectural features and appearance of the tumour under the microscope (histology) and their location and spread (imaging)[55,56]. Advances in genome-wide analysis techniques now allow us to use molecular data when determining subtypes in cancer[57]. This includes data at the genomic, transcriptomic, epigenomic, and proteomic levels. The combination of imaging, histology, and molecular approaches created a vast amount of data that allows new subtypes to be identified and existing subtypes to be refined.

Since the scope of this thesis is about identifying cancer subtypes, the following subsections focus on the most important cancer types, in terms of incidence, with a brief description of their molecular subtypes. These cancer types will be examined in detail in this thesis.

1.5.1 Breast cancer

The breast

The breast is the area of the body that covers the chest. It is made of fat and specialised tissue and its main function, in the case of women, is to provide lactation to babies.

Breast cancer is the most common cancer type in the UK, with a diagnosis rate of 1 in 8 women[58]. Recovery is achievable when the tumour is detected in the early stages, and 8 out of 10 women survive for 10 years or more[58–62]. However, when the tumour is detected

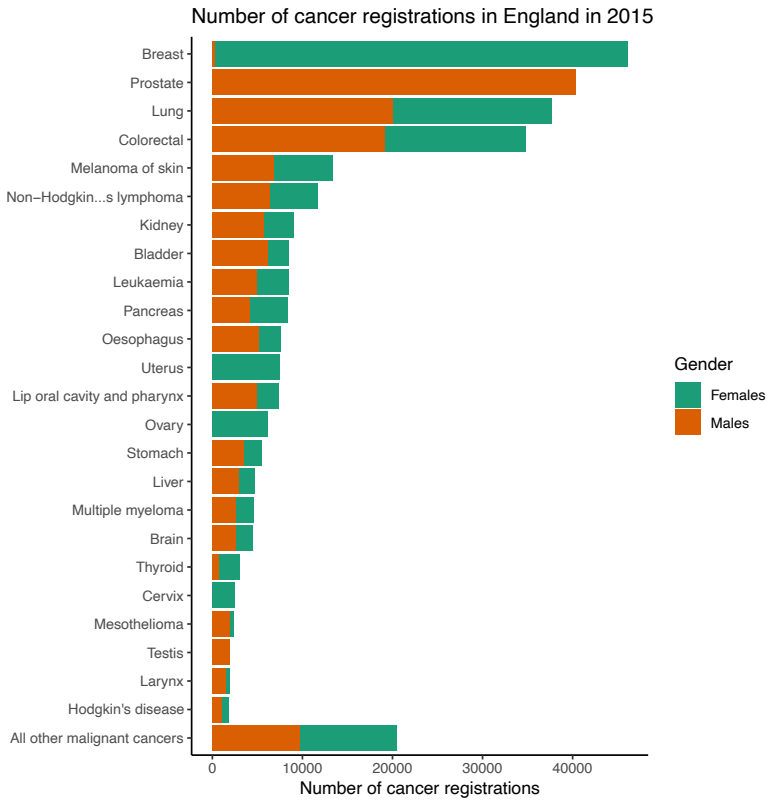

Figure 1.5: Number of cancer registrations in England during 2015 for different cancer types. Breast cancer is the most common type among the whole population and females. Prostate cancer is the second most common type in the whole population and the first in men. Lung and colorectal cancer are the third and fourth most common cancer types. Adapted from ONS (2016)[52].

in later stages, the likelihood of metastasis to bone, liver, lung and brain raises to up to 6 in 10 women[61,62]. The main symptoms can include a change in the size or shape of the breast and nipples, a discharge from the nipples, and/or the appearance of a lump in the armpit[58].

Risk factors

A risk factor is any attribute or characteristic of an individual that increases the likelihood of developing a disease. Age is the main risk factor for breast cancer. In 2016, 99.3% of associated deaths in America were in women over the age of 40[59]. Familiar history also plays a critical role in this cancer: women with a familial history of cancer from first-degree relatives have from 1.75 (one relative) to 2.75-fold higher risk (two or more relatives)[59]. The inheritance risk is mainly related to the mutation of the breast cancer susceptibility genes *BRCA1* or *BRCA2*[63]. These are tumour suppressor genes, and their inheritance is dominant, meaning that when one parent has the mutation, the probability of passing it on is 50%. Breast cancer patients with this type of mutation develop hereditary breast cancer and ovarian syndrome (HBOC), which is associated with early-onset breast cancer and an increased risk of ovarian, pancreatic, stomach, laryngeal, fallopian tube, and prostate cancer[63]. HBOC represents 5-7% of breast cancer cases and increases the risk of developing the disease to 50-80%, while for ovarian cancer, it increases to 30-50%[63].

Other risk factors include early menarche, late menopause, and late age at first pregnancy[59,64]. Each year delay in menopause adds a 3% increase in breast cancer risk, while in the case of menarche, there is a 5% increase[59]. First births in women over the age of 35 increase the hazard ratio to 1.54[64].

Hormone levels play an important role in risk assessment too, especially estrogen. According to Banks (2003)[65], common sources of exogenous estrogen, such as the use of oral contraceptives and hormone replacement therapy, raises the relative risk (RR) to 1.66. Moreover, alcohol consumption elevates levels of estrogen-related hormones in the blood, with a 7.1% increment of the RR for every 10 grams of alcohol consumed per day[66].

Detection and diagnosis

Breast cancer prevention begins with early detection. There are two major screening methods: mammography and Magnetic Resonance Imaging (MRI). Mammography is a reliable and time-wise efficient method that uses low-energy X-rays to generate high-resolution images of the breast[59]. MRI, on the other hand, is a sensitive scan using magnetic fields[67]. MRI has poorer specificity but is not affected by breast density and can detect hidden primary breast cancer, residual tumours, and metastasis[68].

In the UK, the screening schedule consists of a mammography every three years for women between 50 and 70 years old or those with a family history of cancer. In high-risk populations, such as the ones with a family history of cancer, if the mammography turns normal is followed by an MRI as validation[69]. When an abnormality is detected in either of the scans, a biopsy (a sampling from an organ) is performed to confirm the suspicion of cancer and the cells are graded according to their resemblance to healthy cells[69]. Other factors considered include tumour localisation, detection of tubule formations (malignant cells inside the space of a tubular shape), the presence of cells with abnormal nuclear appearance, and the quantity of cells duplicating[70].

Across different cancer types, there is a common staging framework to represent the state of the tumour based on five levels (0, I, II, III, and IV), with possible sublevels (A, B, C) named American Joint Committee on Cancer (AJCC) or numeric staging. The criterion for each level varies across cancer types, but it usually consists of a combination of the grade of differentiation and the anatomical extent of the tumour, which is measured using the tumour-nodes-metastases (TNM) system. The tumour (T) stage is defined by how far the cancer has spread in and around the starting organ; the nodes (N) show if the cancer has spread to the lymph nodes, and the metastases (M) indicate whether the cancer has spread to other parts of the body (Table 1.1). The stages of breast cancer are described in Table 1.2.

Table 1.1: TNM classification for breast cancer. Adapted from Cancer.Net (2020)[71].

Stage	Definition
Primary tumour (T)	
TX	Primary tumour cannot be assessed
T1	Tumour smaller than 20 mm.
T2	Tumour larger than 20 mm but smaller than 50 mm.
T3	Tumour larger than 50 mm.
T4	Tumour is inflammatory or has grown into either the chest wall, into the skin, or between both.
Regional lymph nodes (N)	
NX	Not looked or unclear scans
N0	No cancer in the lymph nodes or cancer smaller than 0.2 mm.
N1	Cancer spread to 1 to 3 lymph nodes.
N2	Cancer spread to 4 to 9 lymph nodes.
N3	Cancer spread to more than 10 lymph nodes.
Distant metastases (M)	
MX	Not looked or unclear scans
M0	Not evidence
M1	Spread

Subtypes of breast cancer

Breast carcinoma has a well-defined molecular subclassification that is effectively used to treat patients[72]. Traditionally, immunohistochemistry markers –estrogen receptor (ER), progesterone receptor (PR) and HER2 receptor- in combination with the tumour appearance and size were used for patient stratification[73]. Sørlie et al. (2001)[74] described a molec-

Table 1.2: AJCC staging for breast cancer. Adapted from Cancer.Net (2020)[71].

AJCC Stage	TNM stage	Description
Stage IA	T1, N0, M0	Small, invasive and not spread tumour
Stage IB	T0, N0, M0 | T1, N0, M0	Spread cancer with small size (less than 20 mm)
Stage IIA	T0, N1, M0 | T1, N1, M0 | T2, N0, M0	Small to medium size tumour (20 to 50 mm) spread to 1 to 3 lymph nodes.
Stage IIB	T2, N1, M0 | T3, N0, M1	Not spread tumour of medium size or spread tumor of small size
Stage IIIA	Any T, N2, M0	Cancer spread to 4 to 9 lymph nodes.
Stage IIIB	T4, N0, M0 | T4, N1, M4 | T4, N2, M4	Inflammatory breast cancer or tumour has spread to the chest wall
Stage IIIC	Any T, N3, M0	Cancer spread to 10 or more lymph nodes
Stage IV	Any T, any N, M1	Tumour has spread to other organs.

ular framework in which tumours were classified into five distinct groups using a hierarchical clustering technique on microarray data that measured gene expression levels for all genes in the genome (Fig. 1.6)[72]. Afterwards, these five molecular groups were mapped to an immunohistochemistry (IHC) profile giving result to the Luminal A, Luminal B, HER-2 overexpressed, Basal-like (also referred to as triple-negative), and Normal-like subtypes (Table 1.3)[75–77]. Luminal A leads in frequency and possesses a good prognosis and a lower relapse rate than the other four subtypes; its treatment consists of hormonal therapy[78]. Normal-like has very similar molecular features, protein expression, and treatment to Luminal A, but it has a slightly worse prognosis[79]. Luminal B, although often grouped with Luminal A, is more aggressive and can require chemotherapy[78,80]. Similarly, Basal-like and HER-2 overexpressed subtypes display a more aggressive behaviour with a higher recurrence rate and probability of metastasis, worst prognosis, and poorer outcome after hormonal therapy[78,80,81].

This classification framework of the five breast subtypes is currently utilised in the clinical practise and research. However, instead of analysing the whole genome, only the gene expression profile of 50 genes (termed as PAM50) is studied[82].

Treatment

Treatment for breast cancer generally depends on the disease stage. For stages I and II, the common treatment pathway is to undergo tumour resection surgery, which can be a mastectomy (breast removal) or a breast conservation surgery[83]. Higher-stage patients

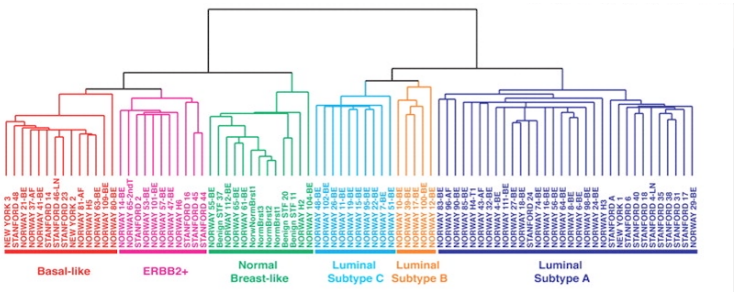

Figure 1.6: Hierarchical clustering performed by Sørlie et al. (2001)[74] classifying breast carcinoma into several molecular subtypes. Adapted from Sørlie et al. (2001)[74].

Table 1.3: Breast cancer subtypes. For each one, the immunohistochemistry (IHC) profile, the grade associated with the cancer, the associated clinical outcome, and the prevalence of the cancer are given. Adapted from Dai et al. (2015)[72].

Subtype	IHC status	Grade	Outcome	Prevalence
Luminal A	[ER+|PR+] HER2-KI67-	1|2	Good	23.70%
Luminal B	[ER+|PR+] HER2-KI67+ | [ER+|PR+] HER2+KI67+	2|3	Intermediate | Poor	38.8% | 14%
HER2 overexpressed	[ER-PR-] HER2+	2|3	Poor	11.20%
Basal	[ER-PR-] HER2-, basal marker+	3	Poor	12.30%
Normal-like	[ER+|PR+] HER2-KI67-	1|2|3	Intermediate	7.80%

need systemic therapy, such as chemotherapy or hormone therapy, which is proven to make 80% of cases suitable for surgery[83]. Cases considered incurable –usually in stage IV- receive systemic therapy alone as a palliative[83]. However, the treatment choice can also vary depending on the dominant subtype of the patient; for example, HER-2 overexpressed tumours have a unique hormone therapy option named Herceptin that targets the cells overexpressing the HER2 protein and improves the survival probability of the patients[84]. Such treatment option is not possible in the basal-like subtype due to its characteristic lack of hormone receptors.

1.5.2 Prostate cancer

The prostate

The prostate is a walnut-sized gland that surrounds the urethra and is situated in front of the rectum, between the bladder and the penis. This organ is only present in men, and its function is to secrete a fluid to the urethra that enhances and protects the sperm[85].

Prostate cancer is the most common cancer in men (26% of all male cancer diagnoses in the UK), and, although it has relatively low mortality, is the second leading cause of death by cancer in men (1 of every 41 men)[86–88].

It is predominantly a disease of older men, with more incidence in people of black African-Caribbean family origin[89,90]. Presenting symptoms are commonly related to difficulties in urination and the presence of blood in urine and semen[91].

Risk factors

The main nonmodifiable factor is age, with more than a third of new cases being men aged 75 years or over and an average incidence rising steeply at the age of 40-44 years to peak at the age of 70–74 years[92,93]. It has been reported that one of every three men older than 50 years has evidence of prostate cancer, but in 80% of cases, the tumours are too small and not aggressive enough to be clinically significant[94]. Other risk factors include ethnicity (increased in African origin, decreased in East Asian origin) and a positive family history of prostate cancer[92].

Modifiable factors such as obesity (relative risk of 1.05 per 5 kg/m2 increment of BMI; RR of 1.03 per 10 cm increment of the waist to hip ratio), smoking (RR=1.22), and the consumption of alcohol (RR=1.25) also have a positive association with prostate cancer[92]. By contrast, physical activity (RR=0.81), the use of anti-inflammatory drugs (RR=0.9), and the consumption of tomatoes (RR=0.81-0.89) and cruciferous vegetables (RR=0.6) reduce the risk of the disease[92].

Detection and diagnosis

The possibility of prostate cancer is investigated if the patient asks for it or when relevant symptoms are present such as frequent urination, problems emptying the bladder or the presence of blood in the urine. The investigation consists of a combination of checking the levels of prostate-specific antigen (PSA) in blood, and a physical examination of the prostate by inserting a gloved finger through the rectum, a process known as digital rectal examination (DRE), to detect abnormal sizes or lumps[95].

Despite their widespread use, the reliability of the PSA and DRE has been called into question. PSA levels can be altered by physical and sexual activity, prostate-related diseases and race; while small tumours can remain undetected by DRE[94,96–100]. It is estimated that PSA, when a cut-off of 3.0ng/mL is used, has a sensitivity of 32% to detect any prostate cancer and 68% to detect high-grade cancers (with a specificity of 85%)[101,102]. Similarly, DRE possesses an overall sensitivity and specificity of 53.2% and 83.6%, respectively[103].

Due to this reason, alternatives have been researched: Paul et al. (2005)[104] proposed the use of the early prostate cancer antigen, and Varambally et al. (2008)[105] suggested the use of Golgi membrane protein 1 for clinically localised prostate cancer. However, PSA, in combination with DRE, remains the clinical standard for triggering further tests[106]. When the results of these tests indicate the possibility of cancer, an MRI is performed to detect and localise the cancer.

However, the gold standard measure to diagnose prostate cancer is to perform a biopsy that is examined by a histopathologist who will assign a score known as Gleason grade[95]. The Gleason grade is evaluated according to differentiation level and glandular patterns presented on the biopsy (Table 1.4)[107].

Table 1.4: Gleason grade criteria according to the appearance of the cells in prostate cancer. Adapted from Humphrey (2004)[107].

Pattern	Tumour Cell Arrangements
1	Single, rounded, closely packed but separated glands
2	Single, rounded, loosely packed but separated glands with variation in size and shape.
3A	Single, widely separated glands of variable shape and size with elongated, angular and twisted forms.
3B	Single, small, widely separated glands of variable shape with elongated, angular and twisted forms.
3C	Glands are pierced with holes (cribriform shape) or with finger-like shape (papillary shape), but no presence of necrosis.
4A	Glands are fused together creating masses, cords or chains.
4B	Glands are fused together with a clear cytoplasm and arranged in masses, cords or chains.
5A	Papillary or cribriform masses with central necrosis.
5B	Masses and sheets of carcinoma differentiated cells, with presence of a few tiny glands.

Staging in prostate adenocarcinoma in the UK is based on the TNM system. However, the use of the AJCC staging is also very common and, for this cancer, is based on a combination of TNM (Table 1.5), Gleason score, DRE, and PSA levels[109–111]. Stage 0 is used when abnormal cells are present but are not extended. Stage I defines cancer that cannot be felt, occupies only one-half of one side of the prostate or less, presents cells with a healthy look, and with low PSA levels. Stage II represents tumours that can or cannot be felt, occupies the inside of the prostate or less, presents cells with moderate or poor

Table 1.5: TNM classification for prostate cancer. Adapted from Prostate Cancer UK (2019)[108].

Stage	Definition
Primary tumour (T)	
TX	Primary tumour cannot be assessed
T1	Cancer only appreciable through a biopsy
T2	Cancer felt during DRE or seen on scans. Still contained on the prostate
T3	Cancer felt during DRE and breaking through the outer layer of the prostate
T4	Cancer spread into nearby organs
Regional lymph nodes (N)	
NX	Not looked or unclear scans
N0	No cancer in the lymph nodes
N1	Cancer in the lymph nodes
Distant metastases (M)	
MX	Not looked or unclear scans
M0	Not spread
M1	Spread

differentiation, and with medium PSA levels. Stage III is characterised by tumours growing outside of the prostate, with poor differentiation and high PSA levels. Stage IV depicts cancers spread beyond the prostate (Table 1.6).

Table 1.6: AJCC staging for prostate cancer. Adapted from American Cancer Society (2021)[111].

AJCC Stage	TNM stage	Gleason score	PSA level
Stage I	T1, N0, M0 | T2, N0, M0	3 + 3 or less	Less than 10
Stage IIA	T1, N0, M0 | T2, N0, M0	3 + 3 or less	At least 10 but less than 20
Stage IIB	T1, N0, M0 | T2, N0, M1	3 + 4	Less than 20
Stage IIC	T1, N0, M0 | T2, N0, M1	4 + 3 | 4 + 4	Less than 20
Stage IIIA	T1, N0, M0 | T2, N0, M3	4 + 4 or less	At least 20
Stage IIIB	T3, N0, M0 | T4, N0, M4	4 + 4 or less	Any
Stage IIIC	Any T, N0, M0	5 + 4 | 5 + 5	Any
Stage IVA	Any T, N1, M1	Any	Any
Stage IVB	Any T, any N, M1	Any	Any

Treatment

Treatment for prostate adenocarcinoma varies greatly depending on the stage of the patient. Stage I cancers are considered low-risk since the tumour grows slowly and may never cause symptoms. Therefore, the recommended treatment is active surveillance, which consists of watching a patient's condition without providing any treatment unless there are changes in the patient's condition. However, the patient can opt for brachytherapy (local radiotherapy) or radiation. Stage II is treated with radiation or brachytherapy in combination with hormone therapy (androgen deprivation therapy, ADT) or alternatively with radical prostatectomy (removal of the prostate gland and surrounding tissue)[112]. Stage IIIA, IIB and IIIC are identical to the previous one, with the difference that during the prostatectomy, if the cancer has spread to the lymphs, the surgery is combined with hormone therapy or, less frequently, radiotherapy. After surgery, radiotherapy and hormone treatment are applied to avoid cancer recurrence. From 20 to 40% of the patients undergoing a prostatectomy suffer a biochemical recurrence, in which their PSA levels exponentially increase, indicating that the cancer has come back[113]. Stage IVA is treated with radiotherapy, with or without brachytherapy, along with hormone therapy and, occasionally, chemotherapy for older populations. For young men, stage IVA treatment consists of a prostatectomy with radiotherapy and hormone treatments afterwards. In stage IVB, the cancer has spread to distant organs and cannot be cured. The main treatments are hormone therapy or surgery for palliative effects, chemotherapy to slow down the disease, and clinical trials to find a new treatment pathway.

Subtypes of prostate cancer

Prostate cancer is very heterogeneous compared to other cancers due to its highly variable clinical course[114]. Many attempts have been made to detect clinically relevant subtypes able to classify patients into a low-risk/high-risk cohort, but no standard classification framework based on molecular features is used in the clinic[114–116]. Despite this, there are several examples of prostate tumours stratified based on mRNA expression signatures and patterns of somatic copy number alterations that provide a better insight into the natural history of the disease[116,117]. Molecular alterations occurring early in the timeline define primary subclasses that accumulate specific additional mutations and drive localised prostate cancer to metastasis[116]. However, after the patient is treated, the molecular lineage landscape shifts and is defined instead by the existing resistance mechanisms[116]. According to this molecular landscape, three major stages of prostate cancer are defined: clinically localised, metastatic ADT-sensitive, and castration-resistant prostate cancer (CRPC). The three stages represent a timeline of events in which clinically localised defines a non-aggressive prostate cancer that can be treated by surgery or ADT therapy. Metastatic ADT-sensitive represents a recurring early tumour that is still sensitive to ADT hormonal therapy. Afterwards, the tumour generates resistance to the treatment and enters the CRPC stage.

The TCGA (2015)[115] proposed to define the clinically localised prostate cancer stage by specific genomic alterations that are often mutually exclusive and can potentially categorise patients into seven different subclasses according to the presence or absence of early gene fusions (in genes *ERG*, *ETV1/4* and *FLI1*) and point mutations (in genes *SPOP*, *FOXA1* and *IDH1*). However, no molecular classification framework is currently being used in clinical practice.

Recently, Luca (2018)[118] proposed that a single prostate cancer sample contains multiple contributing lineages. They opted for a novel unsupervised machine learning methodology – Latent Process Decomposition (LPD)- that fitted their hypothesis better[119]. This approach successfully identified a poor prognosis subtype named DESNT and created a framework to classify patients according to their risk cohort[120].

1.5.3 Lung cancer

The lungs

The lungs are two spongy organs located on either side of the chest. They are made up of sections called lobes; the right has three while the left has two, leaving room for the heart. Both lungs are connected to the windpipe (or trachea) through the bronchi, and their function is to absorb the oxygen from breathing and transfer it into the bloodstream.

Lung cancer is the second most common cancer in men and women[121]. Every year 47,000 new cases of this pathology arise in the UK, and it was estimated to be responsible for 20% of cancer deaths in the European Union in 2016, with a ten-year survival of 5%[122,123]. In addition, survival rates for this disease have plummeted to 2% since 2020 due to delays in diagnosis caused by the COVID-19 pandemic[124].

The associated symptoms of this cancer develop in later stages and often include a persistent cough, breathlessness, coughing up blood, tiredness, weight loss, and pain when breathing or coughing[122].

Risk factors

Age plays a critical role in lung cancer risk. Population older than 75 years represent 40% of lung cancer cases. However, the major cause of this disease is tobacco smoking (risk increment of 20 to 50-fold compared to non-smokers)[122,125].

Other risk factors are diet, alcohol consumption, related diseases, and exposure to chemicals[125–132]. A diet rich in vegetables and fruits decreases the risk of the disease, while consumption of coffee and alcohol is reported to increase it[126,127,133–135]. Respiratory diseases –asthma and tuberculosis, among others- also increase the risk of developing lung cancer (RR of 1.8 and 1.5-2.0, respectively)[129,130]. Lastly, exposure to chemical agents and air pollution are associated with lung cancer and represent over 14% of cases in the UK[125,131,132].

Classification

Lung cancer is classified into three main types depending on which cell they originate from: non-small cell lung cancer, small cell lung cancer, and other types[136]. Small lung cancer represents 15-20% of the cases and is characterised by quick growth and spread[136,137]. Due to this rapid spread rate and their tendency to relapse, radio or chemotherapy are considered the best treatment option[138,139].

Depending on the affected cell, non-small lung cancers are divided into adenocarcinomas, squamous cell carcinoma, and large cell carcinoma[137]. In this work, we focus on adenocarcinomas and squamous cell carcinoma. Both types share similarities in treatment and prognosis but develop in populations with different features: adenocarcinomas develop in mucus-secretor cells and are the most common cancer type in non-smokers, especially in women and young people, although it has a higher incidence in current or former smokers[137]. Squamous cell carcinomas develop on squamous cells inside the airways, and it is often caused by a history of smoking[137]. Early stages of both types are treated with surgery, while more advanced stages require adjuvant chemotherapy[140].

Detection and diagnosis

Lung cancer symptoms manifest in later stages of the disease and can be confused with other problems such as infections or smoking consequences – this hampers the early detection of this disease[141]. Screening is recommended for former or current smokers between 50 to 80 years old and consists of an annual tomography (scan similar to X-rays that provides images of different sections of the human body) to find lesions. When there is a suspicion of cancer, chest x-rays are used for validation[142,143].

A biopsy is the gold standard for diagnosis and is performed either via bronchoscopy (a lighted tube through the throat), mediastinoscopy (incision at the base of the neck to take tissue from lymph nodes), or computerised tomography-guided needle biopsy[143]. The staging framework is applied to all non-small cell cancer lungs independently of further subclassification (Table 1.7).

Treatment

Surgery is the optimal treatment option to remove the tumour when it hasn't spread far[145]. If surgery is not possible, the NICE UK guidelines recommend radiotherapy or chemother-

Table 1.7: AJCC and TNM staging for non-small cell lung cancer. Adapted from American Cancer Society (2019)[144].

AJCC Stage	TNM stage	Description
Stage IA	T1, N0, M0	Small, not invasive tumour less than 3 cm of size.
Stage IB	T2, N0, M0	Small, not invasive tumour between 3 to 4 cm of size.
Stage IIA	T2, N0, M1	Not spread tumour between 4 to 5 cm of size.
Stage IIB	T1, N1, M0 | T2, N1, M0 | T3, N0, M0	Spread tumour less than 5 cm of size or not spread tumour bigger than 5 cm.
Stage IIIA	Any T, N1, M0 | Any T, N2, M0	Tumour of 3-5 cm of size that is spread to the mediastinum or the chest wall.
Stage IIIB	Any T, N2, M0 | Any T, N3, M0	Tumour of 5-7 cm of size that is spread to the mediastinum or the chest wall.
Stage IIIC	Any T, N3, M0	Tumour larger than 7 cm of size that is spread to the mediastinum or the chest wall.
Stage IVA	Any T, any N, M1	Tumour has spread to the other lung.
Stage IVB	Any T, any N, M1	Tumour has spread to distant organs.

apy if the cancer is too widespread[145]. For non-small cell lung cancer, it is common to use targeted therapies to slow down the progression of the tumour[146]. This treatment works by targeting specific changes in cancer cells or by removing environmental advantages for the cancerous cells.

Subtypes of lung cancer

Due to the scope of this thesis, we will only describe the subtypes of non-small cell lung cancers, specifically the ones from adenocarcinoma and squamous lung carcinoma.

Beer et al. (2002)[147] pioneered subtyping by identifying a set of 50 genes that could be used to classify patients into high and low risk according to their gene expression profiles. This led to identifying three clusters of tumours using hierarchical clustering, which were characterised and termed by The Cancer Genome Atlas as the terminal respiratory unit (TRU), the proximal-inflammatory (PI), and the proximal-proliferative (PP) subtypes[148]. TRU was characterised by a good prognosis and an accumulation of mutations in the gene *EGFR* and kinase fusion expression (an abnormal fusion of two genes, one of which is a kinase protein-coding that regulates biological activity). The PI subtype was associated with the mutation of the genes *NF1* and *TP53*. PP, on the other hand, showed enrichment of mutations of *KRAS* and inactivation of the *STK11* gene by chromosomal deletion, reduced gene expression, and deleterious mutation. However, this classification framework is still being validated and is not used in clinical practice.

For squamous cell carcinoma, unsupervised discovery approaches successfully detected subtypes with significant clinical divergencies but failed in establishing the number or nature of these subtypes[149–151]. In 2010, Wilkerson et al.[152] defined and validated four subtypes: primitive, classical, secretory, and basal. The Cancer Genome Atlas (2012)[153] reiterated the existence of these four subtypes and defined them as follows: The classical subtype was characterised by alterations in *KEAP1*, *NFE2L2* and *PTEN*, overall hypermethylation, and chromosomal instability; The primitive subtype was associated with alterations in *RB1* and *PTEN*; the basal subtype was defined by alterations in *NF1*; and the secretory subtype was characterised in later research with elevated immune cell response and slow growth[154,155]. However, this classification framework is still being validated and is not used in clinical practice.

1.5.4 Colorectal cancer

The large intestine

The large intestine is the last part of the gastrointestinal tract in humans. It is formed by the colon, the rectum, and the anal canal (Fig. 1.7). As part of the digestive system, its function is to absorb water and salts from non-digested food and to dispose of the remains via defecation.

Colorectal cancer, also called bowel cancer, is the 4th most common cancer in the UK; it accounted for 11% of all new cancer cases from 2016 to 2018 and has a ten-year survival rate of 52%[156]. The most common symptoms are changes in bowel habits such as defecating more often, presence of blood in the faeces without any other reason, abdominal pain, and a feeling of discomfort after eating[157].

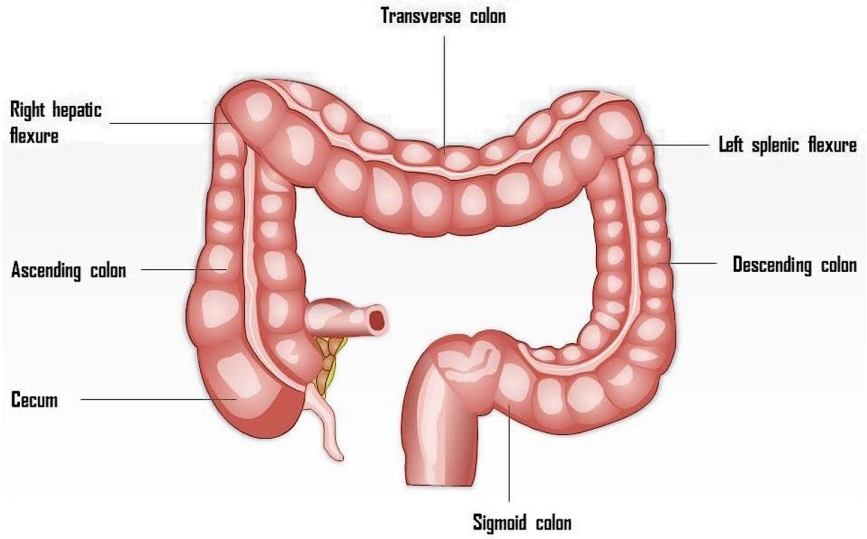

Figure 1.7: Anatomy of the large intestine. Adapted from Slide (2022)[158].

Risk factors

The likelihood of developing colorectal cancer increases with age, specifically after reaching 40 years old[159]. More than 90% of cases occur in people aged 50 or older[159]. A previous history of polyps, adenomas (benign tumours in the bowel mucose), ulcerative colitis, and Chron's disease are positively associated with increased risk[160–162].

Among the modifiable factors, diets high in fat are considered a major risk for colorectal cancer, in addition to lack of physical activity and excess body weight[160,161]. Tobacco and alcohol consumption are also positively associated with colorectal cancer risk[128,163].

Detection and diagnosis

Screening is offered every two years to men and women aged from 60 to 74[164]. The test consists of a faecal immunochemical test which detects small amounts of blood in the faeces. When the result is abnormal, the procedure is followed by a colonoscopy in which a long tube with a camera at the tip is inserted through the rectum to study its appearance[164]. A biopsy confirms the diagnosis after colonoscopy raises suspicions of a possible tumour.

Colorectal cancer staging is based solely on the TNM staging (Table 1.8). In this type of cancer, tumour growth is measured as the depth from the inner of the intestine to the outer (Fig.1.8). T0 represents tumours in the surface layer, T1 stands for tumours in the submucosa, T2 tumours are located in the muscle layer, T3 refers to the subserosa, and finally, T4 means that the tumour has grown through all layers of the bowel[165].

Treatment

Treatment varies depending on whether the cancer is localised on the colon or on the rectum. Surgery is the most common option for both in the first three stages and is supported by the use of chemotherapy as an adjuvant treatment[167]. Additionally, rectal cancers apply

Table 1.8: AJCC and TNM staging for colorectal cancer. Adapted from The American Cancer Society (2020)[166].

AJCC Stage	TNM stage	Description
Stage IA	T1, N0, M0	Small, not invasive tumour less than 3 cm of size.
Stage IB	T2, N0, M0	Small, not invasive tumour between 3 to 4 cm of size.
Stage IIA	T2, N0, M1	Not spread tumour between 4 to 5 cm of size.
Stage IIB	T1, N1, M0 | T2, N1, M0 | T3, N0, M0	Spread tumour less than 5 cm of size or not spread tumour bigger than 5 cm.
Stage IIIA	Any T, N1, M0 | Any T, N2, M0	Tumour of 3-5 cm of size that is spread to the mediastinum or the chest wall.
Stage IIIB	Any T, N2, M0 | Any T, N3, M0	Tumour of 5-7 cm of size that is spread to the mediastinum or the chest wall.
Stage IIIC	Any T, N3, M0	Tumour larger than 7 cm of size that is spread to the mediastinum or the chest wall.
Stage IVA	Any T, any N, M1	Tumour has spread to the other lung.
Stage IVB	Any T, any N, M1	Tumour has spread to distant organs.

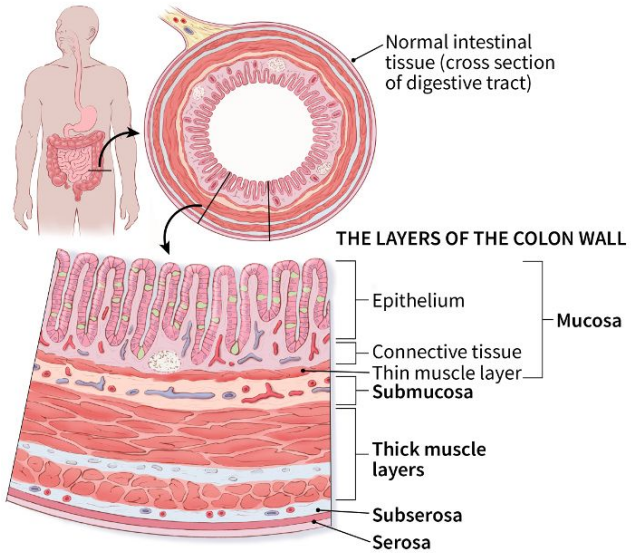

Figure 1.8: Diagram of the layers of the large intestine. Adapted from The American Cancer Society (2020)[166].

radiotherapy or chemotherapy before surgery to reduce the probability of relapse. Surgery can be applied to remove a small part of the bowel, to remove the totality of it, to redirect the bowel to an opening on the abdomen (colostomy), or to remove a bowel obstruction[167]. Stage IV disease requires a combination of surgery and either chemotherapy or radiotherapy, or sometimes both[167].

Subtypes of colorectal cancer

The first studies on subtype identification in colorectal cancer divided it into colon and rectal cancer. However, The Cancer Genome Atlas (2012)[168] concluded that both types are "nearly indistinguishable" on a molecular level. Guinney (2015)[169] proposed a framework for colorectal subtypes based on a consensus from six different analyses, defining four subtypes named Consensus Molecular Subgroups (CMS): CMS1 (microsatellite instability immune), CMS2 (canonical), CMS3 (metabolic) and CMS4 (mesenchymal). CMS1 was characterised by microsatellite instability (parts of the DNA with repeated fragments that varies across humans) and upregulation of immune genes, and it was often detected in female patients with lesions on the right side of the intestine and higher tumour grade. CMS2 was associated with the canonical pathway of carcinogenesis (mutations in *APC*, *p53*, and *RAS*) and with overexpression of epidermal growth factor receptors; in contrast to the previous subtype, CMS2 was more common on the left side and had the best overall survival[170]. CMS3 was defined as metabolic dysregulation with high glutaminolysis and lipogenesis. Finally, CMS4 is characterised by an overexpression of the tissue growth factor pathway, an epithelial-mesenchymal transition, and resistance to chemotherapy, which contributed to being the subtype with the worst overall survival. These findings were corroborated by Ellis (2020)[171] using the clustering methodology of LPD that was also used in prostate cancer. They identified four subtypes with features like the CMS subtypes, especially a low prognosis group referred to as Pericol.

1.6 The Cancer Genome Atlas database

The Cancer Genome Atlas database is a project created in 2005 by the National Cancer Institute of the USA with the purpose of creating an atlas of cancer genomic profiles[172]. The main objective of this proposal was to catalogue and discover major cancer-causing genome alterations in 33 different human tumours through genome sequencing and high-throughput genome analysis techniques[173]. The selection criteria for the studied cancers were a poor prognosis, a current impact on public health, and the availability of high-quality samples[174].

1.6.1 Structure and data organisation

According to Tomczack et al. (2015)[173], the TCGA is organised into multiple cooperative centres that collect and process samples, apply high-throughput sequencing technologies or perform data analyses. The Tissue Source Sites (TSSs) collect blood and tissue from a selection of patients and deliver them to the Biospecimen Core Resource (BCR). The BCR verifies the quality and quantity of the samples and submits: clinical and meta-data to the Data Coordinating Center (DCC) and biological samples (analytes) to the Genome Characterisation Centres (GCCs) and to the Genome Sequencing Centers (GSCs). The GCCs and GSCs perform genetic characterisation, high-throughput sequencing, and

alignment mapping and deposit the resulting data into the DCC. From the DCC, the data is shared with the scientific community through free-access databases.

The TCGA database is structured into projects, each of which specialises in a distinct cancer class, although one class can include multiple disease types with different primary locations, such as adenocarcinomas and neoplasms. Projects are formed of cases that represent a unique patient, and in turn, a single case contains multiple data categories: sequencing reads, transcriptome profiling, simple nucleotide variations, copy number variations, DNA methylation, clinical data, and biospecimen data. All types of data in the TCGA are subclassified in four levels that range from raw and non-normalised to interpreted and summarised data[173]. Only the two most processed levels are publicly available due to confidentiality issues.

Cases can have data from multiple sample sources, normally both healthy tissue (benign) and tumour tissue. To easily identify samples, the TCGA created a barcode composed of a series of identifiers which illustrate the TSS of origin, the participant number, the sample and vial number, the portion and analyte identifiers, the plate sequence, and the centre responsible for studying the properties of the sample (process known as characterisation) (Fig.1.9)[175].

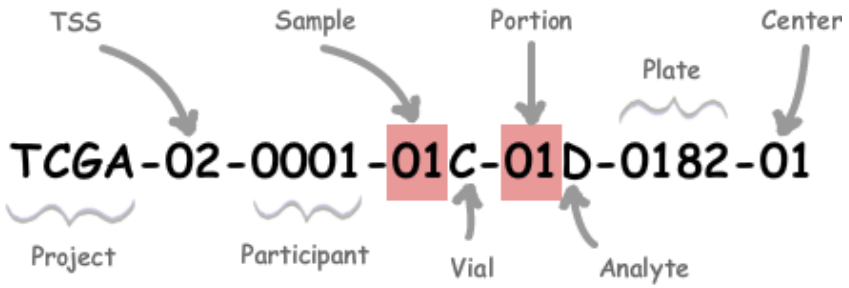

Figure 1.9: TCGA identifying barcode. Obtained from The Cancer Genome Atlas (n.d)[175].

TCGA samples were originally aligned to the Genome Reference Consortium (GRC) version h37, but, since then, the scientific community has evolved drastically due to technological advances and decreased costs[176]. In 2016, the TCGA updated their data to align to GRCh38 and applied a common process to all the samples independently of their cancer type (or project)[177]. Thus, this new version was denominated as *harmonised* TCGA. To this date, the TCGA contains 33 cancer types across 11315 cases, with each cancer type having an average of 235 (interquartile range (IQR) = 357). A complete list of the available cancer types and the number of harmonised cases can be found in appendix A.1. Due to the costs of re-analysing all the samples, microarray data was not uplifted and is only available in the previous version of the TCGA, known as *legacy* TCGA, in the GRCh37 format.

1.6.2 Available data types

The TCGA contains a plethora of data for each sample. The most common measurement experiments applied to the samples are transcriptome analysis, methylation level analysis, somatic mutations and copy number changes analysis. In this section, I describe the data

used in this thesis.

Transcriptome data

The transcriptome is a collective name for all the RNA molecules present in a group of cells at the moment that the sample is taken. TCGA possesses two platforms for obtaining this data: microarray (Agilent G4502A) and RNA-seq (Illumina Sequencing System).

Microarray technology benefits from the DNA hybridisation phenomenon, which describes that complementary nucleotides from different DNA strands can bind to each other: cytosine is complementary to guanine, while adenine is complementary to thymine[178]. The process by which microarrays measure expression is broadly as follows. mRNA is extracted from the samples and converted to DNA by reverse transcription and labelled with a fluorescent dye[179]. The dyed sequences are then applied to a surface where there are spots consisting of multiple copies of DNA (probes) representing a particular gene or region. The applied DNA bind the matching molecules, and fluorescence intensity is measured. Therefore, the higher the fluorescence in a spot, the higher the amount of mRNA containing the sequence present in the probe in that location[179].

In RNA-seq technology, the mRNA from a cell or group of cells is sequenced. The higher the number of copies of mRNA present for a gene, the higher its gene expression is (Fig.1.10)[180]. It also allows the study of splicing outcomes, mutations in the DNA sequence, and post-transcriptional modifications. The main differences between microarrays and RNA-seq are that RNA-seq does not require transcript-specific probes, can be used to detect SNVs or indels, and is more sensitive to weakly expressed genes[181]. However, RNA-seq is more expensive and its results are more complex and difficult to analyse as they require high-power computing[182,183].

DNA Methylation data

The methylation files available in the TCGA were obtained from two platforms: Infinium Human Methylation 27k and Infinium Human Methylation 450k, which differ in the number of probes they cover[176]. Both methodologies follow the sample hybridisation principle than microarrays, but with three fundamental modifications: (i) DNA is used instead of reverse-transcripted mRNA; (ii) the DNA goes through a bisulfite treatment that converts unmethylated cytosines to another nucleotide known as uracil; and (iii) the chip contains two copies of each sequence, one with cytosine and one with uracil[185]. Once the hybridisation is complete, the probes are stained with a specific immunohistochemical assay, so the colour differs according to the methylation status, and the intensity of each colour is scanned[185].

Somatic copy number variations and mutations data

Copy number variations were processed using single nucleotide polymorphism array (Affymetrix Genome-Wide Human SNP6.0), a method that is similar to arrays but with the purpose of detecting modifications or variations in the sequence[176].

Somatic mutations were obtained by whole-exome sequencing (WXS), a process to sequence only protein-coding regions of DNA, by using an Agilent SureSelect Human All Exon kit[176]. Afterwards, TCGA analysed each sample with four different variant detection algorithms:

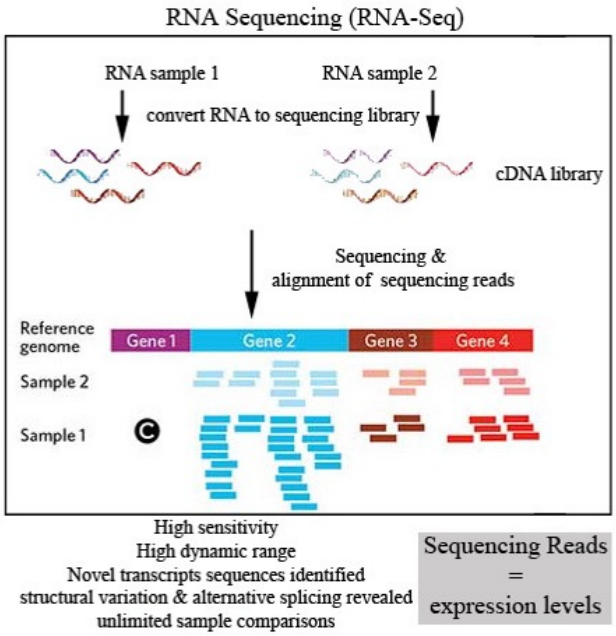

Figure 1.10: Schematic representation of RNA sequencing. In this example, two samples are sequenced at the same time so their gene expression can be compared. Adapted from Otogenetics (2022)[184].

VarScan2, MuTect, Muse, and SomaticSniper[186–189]. Each algorithm has its own strengths and shortcomings.

Clinical data

The clinical data collects all information related to patient diagnosis, demographics, exposures, laboratory tests, and family relationships[190]. Although all cancer types have a fixed set of parameters in common, each type has parameters that are relevant only for that specific cancer[191]. An example of this would be the Gleason score for prostate cancer or the presence of progesterone receptors in breast cancer.

1.7 Clustering and machine learning

Common approaches to detect molecular subtypes require the use of machine learning and clustering techniques. Machine learning is a branch of artificial intelligence that uses data to perform tasks without explicitly programmed instructions. An example is the recommendation systems employed by many streaming services that learn the preferences of the users to recommend new content[192,193].

The corpus of machine learning is divided into two types: supervised and unsupervised[194]. Supervised is when the algorithm is *trained* on *a priori* knowledge and only can be applied to a data pool similar to the one used for training. Unsupervised machine learning lacks *a priori* information and works by finding structure and relationships in the provided dataset.

Clustering analysis is an unsupervised approach that groups samples (or data points) with similar characteristics. More than a hundred clustering techniques exist, and each one

assesses the similarity between data points using a different set of rules[195]. The most widely used ones are based on the notion that the *distance* between two data points in the data space is equivalent to their similarity degree. Distance between data points could be calculated as the spatial distance of one data point from any group to another data point from another group (complete linkage), the minimum spatial distance between two data points from different groups (single linkage), or the mean distance between all the samples from one cluster to the mean from other (average linkage)[196]. Likewise, distance definition can be understood as a metric-based distance such as Euclidian approaches or as a similarity-based distance such as in correlation approaches. Depending on how the data points are classified into groups, clustering can be considered as hard or soft[197].

A common issue with clustering analysis is the selection of the optimal number of clusters. Since the algorithm has no prior information, the number of clusters must be determined statistically to ensure that it is the ideal one for each dataset. A common approach is the elbow method, which is based on the idea of calculating the average internal dispersion of the data points from each cluster and gradually adding more clusters until the dispersion no longer variates and stabilises, meaning that additional clusters are not needed (Fig. 1.11)[198]. Another approach is the silhouette method, which measures how well each data point lies within its cluster and compares the average values while using different numbers of clusters[198].

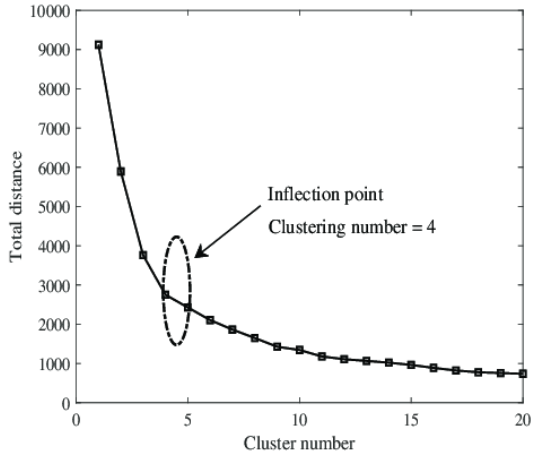

Figure 1.11: Example of applying the elbow method. The selected number of clusters would be 4 since the dispersion stabilises despite adding more clusters.

1.7.1 Hard clustering

Hard clustering is a type of clustering in which each sample is assigned to a single cluster. Some examples of this approach are hierarchical clustering, k-means clustering, density-based spatial clustering of applications with noise clustering, and mean shift clustering[199].

Hierarchical clustering considers each individual data point (or sample) as its own group and merges it with the closest one[200]. The merging process is repeated until all the groups are merged into one single big cluster, and the output is represented as a tree (dendrogram) that

indicates the distance between each group (Fig.1.12). This type of hierarchical clustering is named *agglomerative* clustering, whereas divisive clustering is the opposite process, in which the starting point is one large cluster that is gradually fragmented into smaller clusters. Hierarchical clustering is known for being the method applied to breast cancer to define the PAM50 subtypes framework mentioned in section 1.5.1[201].

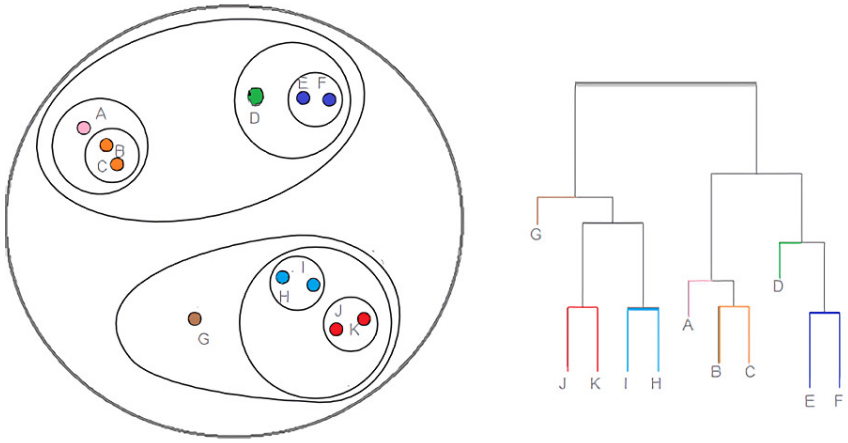

Figure 1.12: Hierarchical clustering process and corresponding dendrogram. Data points are accumulatively clustered together depending on their distance from each other. Obtained from Glen (2016)[202].

K-means clustering works by grouping the data points into a k number of clusters. It does this by generating K random centroids and measuring the distance of each data point to each centroid. The algorithm then groups the data point with the nearest centroid and repeats until all data points are allocated. Afterwards, the centroids are regenerated, but instead of being assigned random values, they are given the sum of all the points assigned to them in the previous iteration divided by the number of points in the group[203]. The algorithm repeats this procedure until the value of the centroids when generated remains constant, meaning that the centroid has centred itself in the middle of its cluster (Fig. 1.13)[203].

1.7.2 Soft clustering

Soft clustering is a type of clustering in which each sample can belong to multiple clusters simultaneously with different degrees of membership. Examples of soft clustering algorithms include fuzzy c-means, gaussian mixture models, and the latent process decomposition (LPD) algorithm that was applied to discover a poor prognosis subtype in prostate cancer and to study the subtypes of colorectal cancer as mentioned in section 1.5.2 and section 1.5.4 respectively.

LPD is a hierarchical Bayesian technique based on latent Dirichlet allocation and is described in detail by Rogers et al. (2005)[205,206]. In terms of cancer and gene expression, LPD defines a cluster as a biological process that leads to a particular expression pattern. Therefore, the expression profile of a sample can be explained as the combination of different proportions of subtype representative expression patterns or processes. As proved by Luca et al. 2008[118], this approach fits better with the high heterogeneity present in cancer that complicates a clear-cut between subtypes.

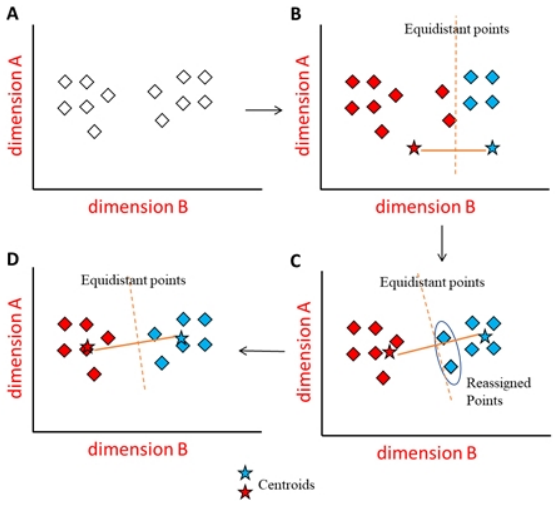

Figure 1.13: Schematic representation of the K-means clustering process. Two random centroids are generated in B, and the data points are grouped according to their closest centroid. In C, the two centroids are regenerated, and the process is repeated, resulting in the reassignment of two data points. In D, the outcome of the analysis is shown. Obtained from Muzhingi[204].

In the following sections, I will describe the algorithm behind LPD in more detail.

Fundaments of LPD

For each given sample a from a dataset D, LPD assumes the existence of a particular distribution of processes θ that contribute to the observed expression profile of the sample. Assuming that the number, K, of processes is known in advance, the distribution θ is defined as a K-dimensional vector whose elements, θ_k, are mixture components. These values reflect the contribution of each process to the sample, i.e. the probability of each process being involved in the gene expression profile of the sample. The distribution θ is assumed to be sampled from a dataset-specific Dirichlet distribution $Dir(\alpha)$ that represents the variance of the mixture components across all the samples that make up the dataset D (Fig. 1.14).

Then, we can define that each gene g contained in the sample a, possess an expression level, e_{ga}, that is sampled from the normal distribution corresponding to each process k, with a mean μ_k and variance σ_k.

Parameter estimation

LPD is a Bayesian model and as such, it has two sets of parameters: observed data D, and hidden or unknown parameters H. Hidden parameters need to be estimated, and in the case of LPD, they are $H = \{\alpha, \mu, \sigma, \gamma\}$, where μ denotes the set of parameters μ_{gk} and σ denotes the set σ_{gk} . When LPD is applied to a dataset, it calculates which parameters H give the optimal result, i.e. the values that maximise the posterior probability $p(H|D)$ which indicates the probability of parameters given the data. The maximum $p(H|D)$ is usually referred to as the *maximum a posteriori* (MAP) probability.

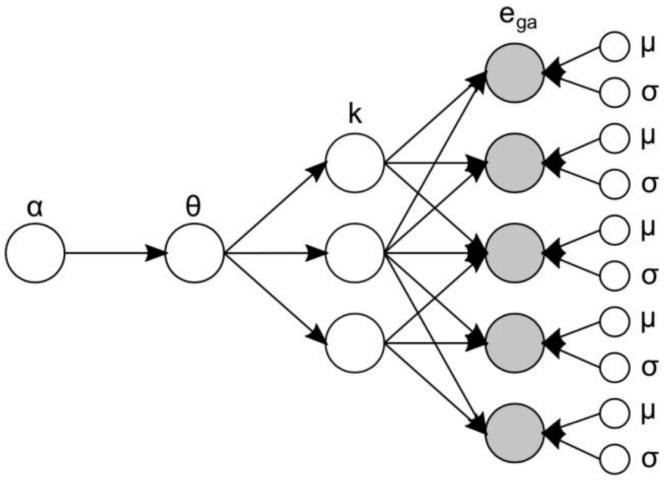

Figure 1.14: Schematic representation of the LPD technique. Each circle represents a variable, and the arrows represent the dependencies between the variables. White circles are assigned to hidden variables, while black circles are observed variables. Obtained from Bogdan-Alexandru (2017)[207].

To estimate MAP, Bayes' rule can be applied, which states that

$$p\left(H|D\right) = \frac{p\left(D|H\right)p\left(H\right)}{p\left(D\right)}, \tag{1.1}$$

where *p(H)* represents the prior, which describes any prior knowledge or belief about the data before observing it, and *p(D/H)* represents the likelihood of the data given the parameters. Since *p(D)* stays constant when trying to find the best *H*, it can be deprecated, therefore

$$p(H|D) \propto p(D|H)p(H). \tag{1.2}$$

If there is no prior knowledge or belief about the data, it is considered a uniform prior as *P(H)* remains constant across *H*. In this case, MAP would be equivalent to the maximum *p(D/H)*, commonly known as maximum likelihood estimation (MLE).

Accordingly, depending on the nature of the prior, there are two ways to find the best hidden parameter values. If the prior is uniform, the best way is by finding MLE, however, one common problem is that MLE tends to over-fit. This is an error in which the model fits too closely to the data and so does not work well for any new data. One possible solution to this problem, known as the MLE solution, is to perform cross-validation, a procedure in which a part of the samples is used to train the model, and the other part is used to test that the model works well. On the other hand, if the prior is not uniform the best way to estimate the right values is by finding the MAP; this approach is known as the MAP solution. Both MLE and MAP solutions, are fully described in Roger et al. (2005)[205].

The MLE solution consists of finding the log-likelihood, defined as *log p(D/H)*, instead of the likelihood. This is because, in the practice, the results from both methodologies are equivalent and finding the log-likelihood is simpler. Therefore, the likelihood

$$p(D|H) \rightarrow p(D|\mu,\sigma,\alpha) = \prod_{a=1}^{A} \int_{\theta} p(a|\mu,\sigma,\alpha)p(\theta|\alpha)d\theta, \tag{1.3}$$

would be equivalent to the log-likelihood

$$\log p(a|\mu,\sigma,\alpha) = log \int_{\theta} \left\{ \prod_{g=1}^{G} \sum_{k=1}^{K} N\left(e_{ak}|k,\mu_{gk},\sigma_{gk}\right) \right\} p(\theta|\alpha)d\theta. \tag{1.4}$$

This approach is a simpler version of LPD but is vulnerable to over and underfitting the data. This means that if the given number of processes is superior or inferior to the actual number of processes inherent in the data, the model can fail to find a good representation for each process.

The MAP solution, on the other hand, is based on the idea that non-uniform priors reflect beliefs about the data in the form of parameters. For example, in a dataset in which the expression level of each gene has been normalised across samples to a normal distribution with mean 0 and variance 1, there could be the belief that the expression across all the genes g is not expected to be significative different in a given process k. This could be represented as the parameter μ_{gk} which would be sampled from a normal distribution $N(0, \sigma_{\mu})$, so

$$p(\mu_{gk}) \propto N(0,\sigma_{\mu}). \tag{1.5}$$

In the same way, it can be assumed that the variance parameter σ^2_{gk} will tend to be close to 1, and it can be designed in a way that it will never be 0, so

$$p(\sigma^2_{gk}) \propto \exp\left\{-\frac{s}{\sigma^2_{gk}}\right\}. \tag{1.6}$$

The MAP solution, in comparison with the MLE, is more complex as it introduces additional parameters that, if they are properly chosen, can protect the model from overfitting but not from underfitting. Among these parameters, only the parameter s has shown a great impact on the results. The parameter s is prior for the variances σ_{gk}, and it would be referred to as *sigma* throughout this thesis.

LPD algorithm

The MAP solution has a better performance for classification than the MLE solution. However, it needs to be provided with the number of processes and the sigma value. In order to estimate these parameters, the MLE solution is applied for each possible combination of number of processes and sigma so that the log-likelihood is interpreted as an indicator of fitness. To ensure the robustness of these values, 100 cross-validations are applied to each possible combination.

The number of processes is usually between 2 and 15, while the sigma value frequently variates between 0.001 and 1.5. During the process of testing each sigma with each number of processes, the log-likelihood increases with every process until one maximum point where it remains stable, commonly known as plateau. This phenomenon occurs due to the overfitting prevention provided by choosing the right sigma, and, therefore, the sigma value with the highest log-likelihood at that point would be the optimal one. The optimal number of processes would be the one before reaching the plateau as it will be the highest one without reaching overfitting.

1.8 Survival analysis of the clinical data

Survival analyses are used in clinical analysis to compare the survival prognostic between two or more groups according to different factors. In cancer research, this methodology is broadly applied to identify cancer subtypes with lower prognoses in comparison to the others.

The fundamental object of study in this analysis is the time to an event of interest (survival time) during the follow-up period of a patient. For cancer research, an event of interest comprehends the remission and relapse of the tumour or the death of the patient. The particularity of this type of analysis is that most events tend to occur early, which requires the adoption of a statistical test that can fit such distribution[208]. However, many patients do not undergo an event during their follow-up period and their true time to an event remains unknown[208]. This is known as censoring and can also occur when the patient is lost during the study period or when the patient experiences a different event that makes the follow-up impossible.

Three common approaches in survival analyses are Kaplan-Meier (KM) survival estimator, log-rank test, and Cox regression analysis. KM estimator is based on the idea that events occur in independent intervals of time. Thus the probabilities of surviving from one interval to the next are the cumulative product of the survival probabilities from the previous intervals[209]. Log-rank test compares the survival between two or more groups and is often used along the KM estimator. Specifically, log-rank calculates for each group at each event time the number of expected events since the previous event if there were no difference between groups[209]. Cox analysis is employed in multivariate models to describe the relationship between the event incidence (defined as the instantaneous event probability at a given time) and a set of covariates[210].

1.9 Thesis overview, aims and objectives

1.9.1 Hypothesis

The application of the Latent Process Decomposition algorithm to analyze transcriptome data from The Cancer Genome Atlas will lead to the identification and multi-omic characterization of distinct molecular subtypes across several cancer types. The identified subtypes will exhibit unique gene expression patterns, differential methylation, genetic alterations, and prognosis. The study of these subtypes will contribute to the understanding of tumour progression and heterogeneity, assist in the transition to personalized cancer therapy, and aid in developing potential biomarkers and therapeutic targets effective across different cancer types. The results obtained from this research will serve as a foundation for future validation and exploration of additional datasets, potentially extending the understanding of cancer subtypes and their clinical significance. Additionally, developing pipelines to automate the process of detecting and characterizing these subtypes will provide a robust and automated approach for detecting and studying molecular subtypes, contributing to a comprehensive resource for cancer subtype analysis.

1.9.2 Aims

To utilize the Latent Process Decomposition (LPD) algorithm to analyze RNA-seq and microarray transcriptome data sourced from The Cancer Genome Atlas in order to identify and characterize subtypes across several cancer types, contributing to a comprehensive resource of the results. In addition to develop a pipeline that automates the process of identifying and characterizing subtypes.

1.9.3 Objectives

- To apply LPD to all the TCGA expression datasets with appropriate size ($n > 100$) to detect subtypes using a novel approach to determine the optimal number of subtypes.
- To develop an R pipeline to automate the methodology.
- To characterise the stratifications detected by LPD by studying differentially expressed genes, clinical associations, and associations with genetic alterations.
- To validate this approach by comparison with other subtypes for four common cancers.
- To perform a pancancer review of the subtype characteristics which are in common across all studied datasets.
- To identify novel subtypes with differential prognosis.
- To generate interactive reports of the results for each cancer dataset to be used as a rich resource for scientists to generate hypotheses associated with molecular subtypes.

1.9.4 Chapter overview

- Chapter 2: Detailed descriptions of all methods applied in this thesis are given, serving as a reference for analytical and statistical methods.
- Chapter 3: Introduction the R package 'Automata' developed to automate the methodology used in this thesis.
- Chapter 4: Pancancer analysis across all the subtypes detected in this study independently of their cancer type.
- Chapter 5: Validation of the Latent Process Decomposition algorithm by testing it in breast carcinoma, prostate adenocarcinoma, colorectal adenocarcinoma and lung cancer.
- Chapter 6: Identification of subtypes with significant associations with survival probability. The ones in skin cutaneous melanoma and in bladder cancer are explored in detail.
- Chapter 7: The results from all previous chapters are considered as a whole, discussing the strengths and weaknesses of the analyses within this thesis. Several future directions for this research area are also discussed.

Additional supplementary material is available as a separate file.

Chapter 2

Methods

2.1 Statistical tests, models, and transformations

All statistical analysis was performed in R 3.6.2, and unless otherwise specified used default parameters and two-tailed tests of significance, with $P < 0.05$ accepted as the threshold for "significance".

2.1.1 Data transformations: variance-stabilising transformation (VST) and log2

Data transformation is the application of a mathematical function to all the datapoints that form a dataset with the purpose of making them suitable to be used as input for specific statistical tests[211].

During the processing and analysis of the transcriptome, the mRNA is fragmented with different abundance and can lead to an overrepresentation of specific genes that distorts any downstream statistical analysis. Two common data transformation approaches to counter this are variance-stabilising transformation and log-2 transformation (Fig 2.1). VST minimises the impact of the most abundant fragments and therefore reduces the dependency of the standard deviation on the fragment abundance. Specifically, VST finds a simple function f to apply to values x in a dataset to create new values $y = f(x)$ such that the variability of the values y is not related to their mean value. Log-2 transformation instead solves this same problem by magnifying the impact of the low abundant fragments by applying a binary logarithm.

2.1.2 Correlation tests: Pearson's and Spearman's coefficients

In statistics, a correlation is any relationship between two (or more) variables[213]. Correlation coefficients measure this relationship and range from a fully negative relationship (-1) to a fully positive relationship (+1), while 0 corresponds to no relationship[213]. Popular coefficients in genome analysis are Pearson's correlation coefficient (denoted as r) and Spearman's rank correlation coefficient (denoted as ρ). Pearson's coefficient measures the strength of a linear association between the variables by drawing a line of best fit through the data and calculating how far each data point is from the line[214]. Alternatively, Spearman's coefficient is a non-parametric test that measures the fitness of the data points to a

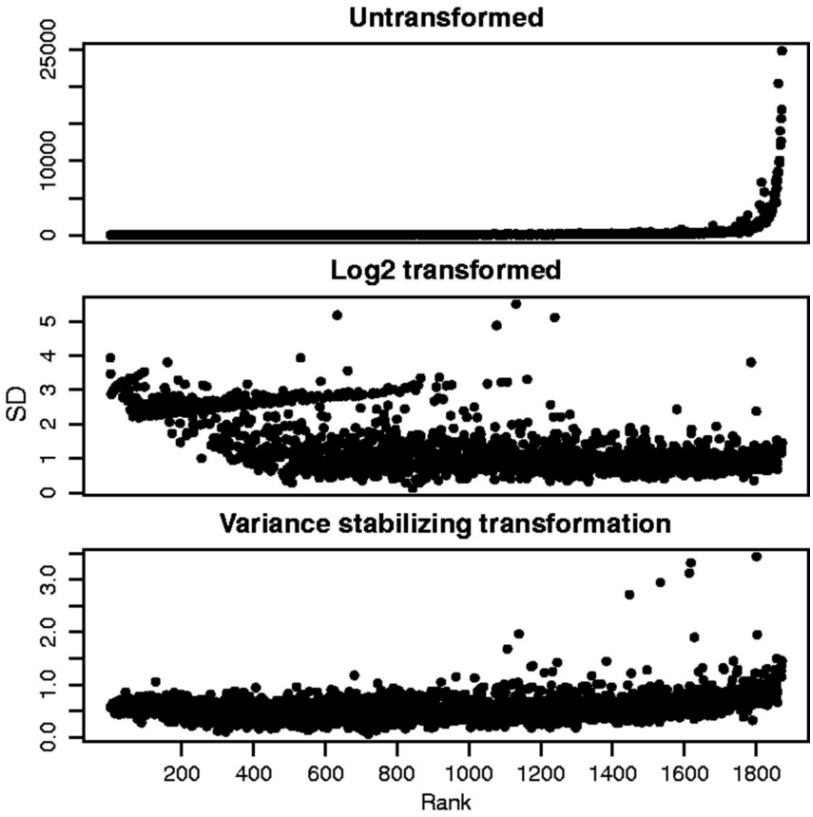

Figure 2.1: Visual representation of VST and log2 transformation on the data points compared to untransformed data. Log2 transformation magnifies the low abundant reads, while VST minimises the impact of the high abundant reads. Obtained from Klein (2015)[212].

monotonic function (a function that increases or decreases over its entire range)[214]. Because of their differences, Pearson's is the preferred choice for raw data values, while Spearman's is better suited for rank-ordered values.

2.1.3 Student's t-test

Student's t-test is a hypothesis testing technique for comparing the mean of two paired samples drawn from a normally distributed populations with an unknown standard deviation[215]. When a sample is drawn from a normally distributed population, the sample is also normally distributed as long as the sample size is large (more than 30). The t distribution assumes that the likelihood of extreme values is greater in smaller sample sizes, and therefore the distribution curve becomes flatter and broader (Fig 2.2)[215]. With this in mind, the advantage of the t-test over other statistical test is that can be applied to any sample size, including very small (less than 10).

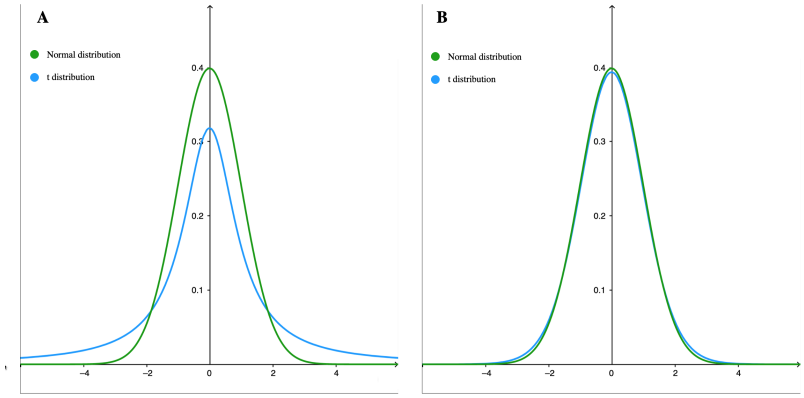

Figure 2.2: Example of the normal and t distribution for a sample size of (A) 2 and (B) 20. With a large sample size, the t distribution becomes closer to the normal distribution . Edited from Raystuckey1 (n.d.)[216].

2.1.4 Wilcoxon signed-rank test

The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test for comparing the mean of two populations using two unpaired samples. It is considered the non-parametric alternative to the Student's t-test, which means that can be applied to datapoints not normally distributed. Essentially, it calculates the difference between two sets of independent samples and analyses these differences to establish if they are statistically significantly different from one another[217].

2.1.5 ANOVA test

Analysis of variance (ANOVA) tests if there are statistically significant differences between three or more independent groups. In addition to the sample size and the mean values per group, the test analyses the variance between the groups across the samples drawn[218].

2.1.6 Chi-squared test

A chi-squared test determines whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table[219]. It is calculated as

$$\mathcal{X}_c^2 = \sum \frac{(Oi - E_i)^2}{E_i} \tag{2.1}$$

where c denotes degrees of freedom, O denotes observed values, and E denotes expected values.

2.1.7 Post-hoc analysis and Tukey test

When the results of a statistical test are significant, a post hoc analysis is used to determine where the differences came from. An example is the Tukey test, which is the post-hoc for ANOVA and performs pair-wase comparisons to detect which groups are different[220].

2.1.8 P-value adjustment for multiple testing

When performing multiple statistical analyses, p-values below the threshold of significance may occur randomly. P-value adjustments are one way to avoid these false positive errors. In transcriptome studies, Benjamin-Hochberg (BH) is a common approach because it works best with large datasets and does not produce false negative results. BH orders and ranks the p-values of all analyses from lowest to highest and then calculates the critical value as

$$(i/m) * Q \tag{2.2}$$

where i denotes the rank of the p-value, m defines the total number of tests, and Q represents the chosen false discovery rate. The rank of the largest p-value that is less than the critical value is selected, and all the ranks below it are considered true significant.

2.1.9 Kernel density estimation

Kernel density estimation is a non-parametric method for estimating the probability density function of a random variable[221]. This method is similar to histograms, with the difference that instead of being a discrete representation divided into bins, they are displayed in a smooth curve (Fig. 2.3)

2.1.10 Jaccard Similarity Index

Jaccard similarity index compares the member of two sets, determining which members are shared and which are distinct[223]. According to this, it measures the similarity between the two sets on a scale of 0% to 100%. It is calculated as

$$J(X,Y) = (X \cap Y) / (X \cup Y), \tag{2.3}$$

where X and Y represent each set.

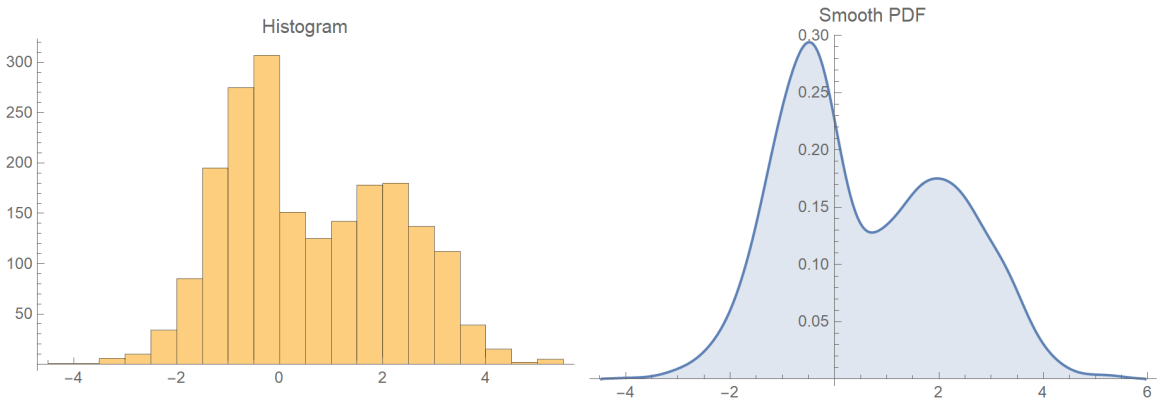

Figure 2.3: Representation of the density of a population through a histogram and a kernel-density estimation function. Obtained from Kamperis (2020)[222].

2.1.11 Survival analysis: Kaplan-Meier estimator, log-rank test, and Cox regression analysis

Survival analyses are used on clinical analysis to compare the survival prognostic between two or more groups according to different factors. Three common approaches in survival analyses are Kaplan-Meier estimation, log-rank test, and Cox regression analysis. These three methods and their application to cancer studies are discussed in section 1.8.

2.1.12 Latent Process Decomposition

Latent process decomposition is a hierarchical Bayesian technique based on latent Dirichlet allocation that classifies samples into soft clusters. This technique is analysed in depth in section 1.7.2.

2.1.13 Limma

Limma is an R package that provides a framework for analysis gene expression experiments based on microarrays[224]. Limma assumes that in a matrix of expression values, each row represents a gene and each column corresponds to a RNA sample, and it fits a linear model to each row of data.

2.2 Programming resources and tools

2.2.1 R programming language and libraries

Developed by Robert Gentleman and Ross Ihaka, R is an open-source programming language and environment specialised in statistical analysis[225]. The language also provides a wide range of graphical tools for creating high-quality plots and has access to thousands of packages that extends its functionalities. Libraries are denoted as 'packages' in R and contain code, data, and documentation in a standardised fashion. In recent years, R has emerged as one of the most popular programming languages for machine learning, biomedical research, bioinformatics, data mining, and financial mathematics[226,227]. The version of R utilised for this thesis was 3.6.2.

2.2.2 Rmarkdown format

The Markdown format is a lightweight language that employs simple syntax and facilitates the transformation of human-readable text files into HTML or PDF publications[228]. Through the `Rmarkdown` package, R is able to integrate embedded code snippets into the Markdown language to create dynamic documents and facilitate reproducible research[229]. The version of Rmarkdown utilised for this thesis was 2.1.

2.2.3 Flexdashboard framework

The `Flexdashboard` package provides a framework for Rmarkdown to create reproducible web-based dashboards[230]. Dashboard layouts are highly customisable and automatically adjust for optimal display on differently sized screens[231]. The package also provides functionalities to the web-documents such as value boxes, gauges, text annotations, and interactive JavaScript-based data visualisations and tables[231]. The version of Flexdashboard utilised for this thesis was 0.5.1.1.

2.2.4 High Performance Computing

High Performance Computing (HPC) refers to the practice of aggregating computing power to process data and perform complex calculations at high speeds. The University of East Anglia possesses an HPC server that runs the Centos 7 Linux operative system and consists of more than 7000 CPU cores, over 6 TB of RAM memory and further than 100 TB of storage[232]. The server employs the open-source SLURM[233] job scheduler, which allows users to submit scripts (termed as jobs) with an allocated RAM memory and core power to be run in the server.

Unless otherwise indicated, all statistical analysis and machine learning applications described in this thesis were performed on the HPC of the University of East Anglia using a maximum of 96 GB of RAM and 24 cores from an Intel Xeon E5-2620 v4 2.1Ghz node.

2.2.5 Gene set enrichment analysis

Gene set enrichment analysis (GSEA) is a method for identifying gene classes that are over-represented in a large gene set and that may be associated with diseases. The method works by considering a measure of association between genes and the phenotype of interest (e.g., fold change for differential expression) and ranking the genes according to that measure of association. A test is then performed for each annotation category to determine whether the ranks of genes in that category are evenly distributed across the rank list, or appear more towards the top or bottom of the list. When Tian et al. (2005)[234] proposed this methodology, they applied Student's t-test for two group experiments and ANOVA for multi-group experiments as a test. The p-values from the GSEA are obtained by permutating the tests for each category.

2.3 Databases

2.3.1 Catalogue of Somatic Mutations in Cancer

The Catalogue of Somatic Mutations in Cancer (COSMIC) is one of the largest and most comprehensive curated datasets of the impact of somatic mutations in human cancer[235]. The

data is derived from the scientific literature after an in-depth curation process combined with data imported from the major cancer data portals such as the TCGA[236]. COSMIC describes over 5,900,000 coding mutations across more than 1,300,000 samples[236].

Based on COSMIC, genes with strong evidence that they are functionally affected by driver mutations are collated into the Cancer Gene Census (CGC). The CGC classifies genes into two tears based on the strength of the evidence supporting their role in carcinogenesis[235]. The first tier requires two publications from two independent groups describing the role of the gene on cancer development, while the second tier requires extensive bibliographic evidence[236].

In addition, COSMIC contains a set of mutational signatures, which are particular combinations of mutations that occur due to specific processes such as exposure to ultraviolet light or mismatches during DNA replication.

2.3.2 Kyoto Encyclopaedia of Genes and Genomes database

The Kyoto Encyclopaedia of Genes and Genomes (KEGG) was born to link genomic information to a network of interacting molecules in the cell representing a higher-order biological function[237,238]. Currently, KEGG consists of fifteen curated datasets divided into four categories: SYSTEMS for high-level systemic functions, GENOMIC for molecular-level functions, CHEMICAL for metabolites and products of biochemical reactions, and HEALTH for drug effects[239].

In the work described in this thesis, we focus on the pathways database contained in SYSTEMS. This data is commonly applied in transcriptome analysis to identify the biological processes associated with genes of interest. It consists of manually drawn reference pathway maps combined with organism-specific pathway maps[239].

2.3.3 Gene Ontology consortium

The Gene Ontology (GO) Consortium provides a systematic, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism[240]. The consortium is divided into three levels: molecular function, biological process, and cellular component[241]. Molecular function describes activities that occur at molecular level such as catalysis or transport. Biological processes are the biological objectives to which a gene or gene product contributes. Cellular component contains the location in which molecular functions and biological processes occur.

2.3.4 The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas database contains a catalogue of major-causing genome alterations in 33 different cancer types through genome sequencing and high-throughput genome analysis techniques. The structure and types of data in the TCGA are described in depth in section 1.6.

Chapter 3

The Automata package

3.1 Background

Previous research using the LPD algorithm has predominantly focused on analysing a single cancer type. However, in this work, my objective was to gain results from multiple cancer types, which required the sequential iteration of the methodology for each suitable cancer project dataset available in the TCGA database. To accomplish this, I have developed pipelines to automate and streamline the execution of this methodology and integrated them into an R package. The decision to create an R package was motivated by its easiness of being installed in R by any user regardless of coding experience, its modular use that allows a flexible utilisation, and to facilitate the improvement and editing of the code in future updates and bug fixes.

This package was named Automata, an abbreviation for *Automatic TCGA download and processing*. I have divided Automata into six major steps to facilitate the understanding of the package's workflow (Fig. 3.1). The first step is data download, which involves acquiring all relevant data associated with a specific TCGA cancer project. In the second step, data preprocessing, the obtained data is cleaned, normalised, and prepared for subsequent analysis. The third step, LPD application, encompasses estimating the optimal parameters for LPD and applying the model to the prepared data through multiple runs to account for randomness. Following that, in the data postprocessing step, the best run from LPD is selected, and each sample is assigned to a potential subtype. The fifth step, differential analysis, involves characterising each LPD-identified subtype by comparing their clinical and molecular associations. Finally, the results are presented in an interactive HTML report generated in the sixth step, which provides a comprehensive summary of the findings for each subtype. With this structured approach, Automata enables streamlined processing and analysis of TCGA data, facilitating the investigation and interpretation of cancer subtypes across multiple cancer types.

3.2 Automata workflow

In this section, I provide a detailed description of the six major steps covered by the Automata workflow. Each of these steps consists on several programming functions that serve different purposes (Table 3.1).

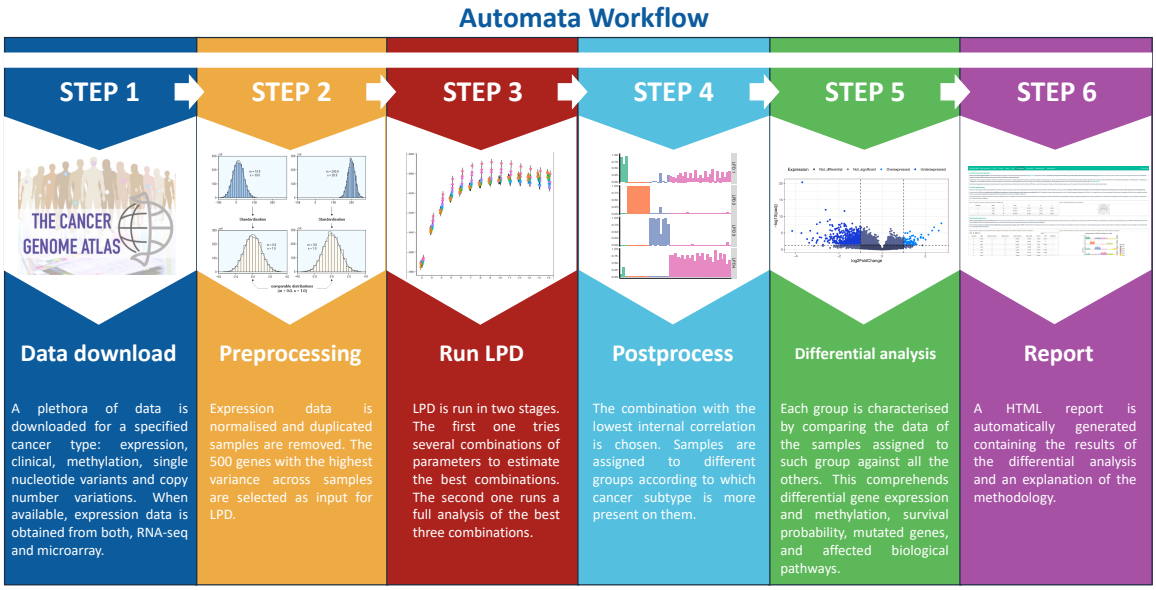

Figure 3.1: Automata workflow. The workflow of the package is divided into six major steps that encompass the entire process from downloading cancer data from the TCGA to reporting the findings of the analysis.

Table 3.1: Functions included in the Automata R package and their description. The functions are divided into six major steps representing the workflow of the package.

Function	Description
Data download	
createTCGAfolder	Creates the folder TCGA_results to store the data generated by Automata.
downloadClinical	Downloads gene expression data from the TCGA for a given cancer type.
downloadCNV	Downloads CNV data from the TCGA for a given cancer type.
downloadExpression	Downloads clinical data from the TCGA for a given cancer type.
downloadMeth	Downloads methylation data from the TCGA for a given cancer type.
downloadSNV	Downloads SNV data from the TCGA for a given cancer type.
getBarcodes	Queries the TCGA database to obtain the barcode of the suitable samples for a given cancer type.
getQueryFilePaths	Retrieves the filepaths to which the TCGA data is downloaded to aid in locating in future analyses.
Preprocessing	
cleanMatrix	Removes duplicated samples.
normaliseExpression	Normalises expression values.
topGenes	Pick the 500 genes with the highest variance.

Table 3.1: Functions included in the Automata R package and their description. The functions are divided into six major steps representing the workflow of the package. *(continued)*

Function	Description
LPD application	
estimateParameters	Selects the best combinations of parameters.
generateJobFolder	Creates folders to run LPD on.
runLPD_A_ADA	Executes the first stage of LPD.
runLPD_B_ADA	Executes the second stage of LPD.
Postprocess	
assignGammas	Assigns samples into LPD groups according to their gamma values.
calculateCorrelation	Performs Pearson's correlation analysis in the gamma values from the three representative runs.
compareGamma	Compares the gammas values of all the runs within the same combination to calculate which one is closer to the medoid.
postProcess	Applies all the functions of the postprocess step.
sortGamma	Sortes the processes from runs within the same combination to match the reference run.
Differential analysis	
batchEffect	Analyses the presence of batch effect in the data.
clinicalAnalysis	Performs an analysis of the clinical data downloaded from the TCGA.
cnvAnalysis	Analyse the CNV data from the TCGA after creating genomic coordinates.
createIntersection	Creates a Venn diagram comparing the results from the differential analysis.
diffExpGenes	Performs a differential expression analysis.
extraClinical	Performs additional clinical analysis for prostate cancer data such as Gleason score, PSA levels etc.
generateAberrations	Creates genomic ranges from the CNV data downloaded from the TCGA and finds overlaps with reference genomic ranges from bioMart.
methylationAnalysis	Performs a differential analysis of the methylation data from the TCGA.
runDendrogram	Performs hierarchical clustering in the samples of the TCGA based on their gene expression profile.
runPathwayAnalysis	Performs enrichment analysis of a subset of genes in the KEGG and GO database.
snvAnalysis	Analyse the SNV data from the TCGA
somaticSignaturesAnalysis	Analyses the associations of the COSMIC Somatic Signatures with the LPD groups.
Report	
createChildsRMD	Creates the files for the sections of the TCGA report file.
createMainRMD	Creates the main framework for the TCGA report file.

Table 3.1: Functions included in the Automata R package and their description. The functions are divided into six major steps representing the workflow of the package. *(continued)*

Function	Description
generateOverviewRMD	Creates the overview sections for the TCGA report file.
generatePathways	Compiles the paths of every individual file that is required to generate the TCGA report file.
generateReport	Generates the HTML file containing the TCGA report file.
generateSubChilds	Creates the files for the subsections of the TCGA report file.

3.2.1 Data download

To ensure the availability of diverse cancer types, a large sample size, and the utilisation of multiple data platforms, all data for this research was obtained from The Cancer Genome Atlas (TCGA). An additional advantage of using the TCGA is that, due to their harmonisation process (see chapter 1.6), all the different cancer types go through the same analysis and batch effect removal, enabling biologically meaningful comparisons across them. This standardisation allows for robust and reliable comparisons of molecular features and clinical characteristics across various cancer types, enhancing the validity and universality of the research findings. The studies conducted by Luca et al. (2018)[118] and Ellis et al. (2021)[171] to identify subtypes of prostate and colorectal cancer, respectively, were conducted using LPD in the TCGA data. As such, their studies can be used as valuable references and offer further validation of the findings in this research.

One common challenge with unsupervised algorithms like LPD is the requirement for a sufficient number of samples to ensure reliable results without introducing statistical distortions. To address this concern, I made the decision to implement a cut-off value of 100 samples as a minimum requirement for TCGA cancer datasets to be eligible for analysis. This cut-off value was chosen to ensure an adequate sample size that represents the inherent biological heterogeneity of cancer disease, and mitigate the risk of statistical biases. In total, 28 cancer datasets passed this threshold out of the 33 available (a complete list can be found in appendix A.

To obtain the TCGA data, I utilized the R package `TCGAbiolinks` to query and download various levels of information in a formatted manner. This included transcriptome counts (as RNA-seq HTseq counts), transcriptome intensity (from microarray data), clinical features, methylation data (with priority given to 450K over 27K), SNV data (MAF files from VarScan, Mutect2, MuSe, and Somatic Sniper platforms), and copy number segment data. It is noteworthy that transcriptome microarray data was only available for 11 of the 28 cancer types in the legacy TCGA dataset. Only samples with available expression data, either in RNA-seq or microarray format, were selected for inclusion in this study. This selection criterion is because LPD, the algorithm employed in this research, was specifically developed for the sole analysis of expression data. To ensure high-quality reads of the SNV data, I generated a consensus MAF file by selecting only the SNPs present in all four platforms and indels present in both Mutect2 and VarScan.

3.2.2 Data preprocessing

The data preprocessing stage consists of several steps to prepare the downloaded data for LPD analysis, including data cleaning, normalisation, and gene selection.

In the TCGA dataset, multiple samples from the same patient are often available, derived from different tissues or separate laboratory analyses. To ensure an appropriate representation of patients and to avoid excessively complex results, a maximum of one normal tissue sample and one tumour sample were selected from each patient. In terms of sample selection, solid tumour samples were prioritized over other sample types due to their higher availability across the TCGA and overall higher quality. Samples derived from fresh-frozen vials were given priority over paraffin-fixed (FFPE) samples. This decision was made based on the superior quality of RNA obtained from fresh-frozen samples. Additionally, in the TCGA, re-runs of samples due to technical errors are assigned a higher barcode value than their initial runs[242]. Therefore, samples with higher portion and/or plate numbers were chosen to ensure consistency (Fig. 1.9). By applying these selection criteria, the data preprocessing step aimed to optimize the quality and representativeness of the samples used in the subsequent LPD analysis.

Transcriptome data was normalised through two different methodologies depending if it was derived from RNA-seq or microarray. The RNA-seq counts were normalised across samples using the R package `DESeq2` and the application of VST (see chapter 2.1.1). This normalisation method accounts for variations in sequencing depth and gene-specific biases, ensuring that the expression data are comparable across samples and suitable for subsequent analysis. In the case of microarray data from the legacy TCGA dataset, the publicly available expression data had already undergone normalisation through log2 transformation. However, this transformation can result in negative values, which are not compatible with the LPD algorithm design. To address this issue, a practical solution was implemented. The minimum intensity value for each gene across all samples was identified, and this value was added to the data for that specific gene. By adding this minimum value, the negative values are adjusted, making them compatible with the LPD algorithm.

Due to the resource-intensive nature of the LPD algorithm and technical limitations, it was not feasible to apply the algorithm to the entire gene expression dataset. The execution time of the LPD algorithm is known to increase exponentially as the data size expands. To overcome this limitation, a proposed solution was to select a subset of genes for analysis. Specifically, the 500 genes with the highest expression variance among samples were chosen. This selection criterion was based on the findings of Rogers et al. (2005)[205], who suggested that this minimum sample size adequately represents the transcriptome heterogeneity of the patient population for LPD analysis.

3.2.3 Applying LPD

The application of the LPD algorithm is divided into two distinct stages: parameter estimation and model application (Fig. 3.3). Both stages were specifically designed to be executed on the High-Performance Computing (HPC) infrastructure at the University of East Anglia. The decision to utilize the HPC environment was driven by several factors, including the substantial amount of RAM and disk space required for processing the data, as well as the extended duration of the computational job.

To estimate the parameters for the LPD algorithm, a comprehensive evaluation was conducted by testing 90 combinations of two hyperparameters: sigma values and the number of processes. The sigma values were varied between a range of -0.0001 to -1.5, while the number of processes ranged from 2 to 15. Each combination was repeated five times with 1000 iterations per repetition, and the average log-likelihood was used as a measure of fitness for each combination. The selection of the optimal combination of parameters was guided by several considerations. Firstly, as explained in 1.7.2, the fitness of the model tends to increase with the number of processes until overfitting occurs. Therefore, the goal was to identify the number of processes just before reaching the point of overfitting (Fig. 3.2). Secondly, within that optimal number of processes, the sigma value with the highest log-likelihood was selected as a reference. Given the objective of this research to identify distinct subtypes with clinically significant differences, it was recognized that defining a large number of processes for the same cancer type may result in overlapping molecular features across subtypes or clinical outcomes that lack meaningful relevance in treatment development. To address this, a pragmatic approach was taken by selecting the combination with the least number of processes within the range of the standard deviation of the reference sigma value. Additionally, adjacent combinations in terms of the number of processes were also taken into consideration as a precautionary measure. In previous applications of LPD, this approach was done manually by visually inspecting the log-likelihood graph (Fig. 3.2), however this was streamlined in the Automata developing process to improve the consistency of the approach[171,243,244].

The second stage of the LPD application involved running the LPD model multiple times on each of the selected combinations. Specifically, the model was run 100 times for each combination, with each run consisting of 1000 iterations. To account for the randomness inherent in the unsupervised algorithms, each run was initialized with a different random seed, which is a selected number used to initiate pseudo-number generation. From each run, a matrix was generated, capturing the degree of membership of each sample to each detected process. This information was represented as "gamma values" ranging from 0 to 1, indicating the strength of association between each sample and each detected process. This approach ensures the robustness and reliability of the LPD results, minimizing the impact of randomness and providing a comprehensive assessment of the sample-process associations.

3.2.4 Postprocessing the LPD outcome

The postprocessing step involves comparing the output of each run to determine the best result among the three selected combinations (Fig. 3.3). These encompass three steps: (1) comparing the runs within the same combination to identify the most representative run for that specific combination, (2) comparing the representative runs from each combination to determine the overall best outcome, and (3) assign each sample to a group according to the process more prevalent on them.

Given the probabilistic nature of LPD, the starting point of the analysis for each run is randomly determined. As a result, the identification and labelling of processes can vary between different runs. For example, a subset of runs may label a particular process as process 1, while another subset of runs may label the same process as process 4. A randomly selected run was chosen as reference to address this variability and facilitate comparison between runs. The gamma values of this reference run were then compared to the gamma

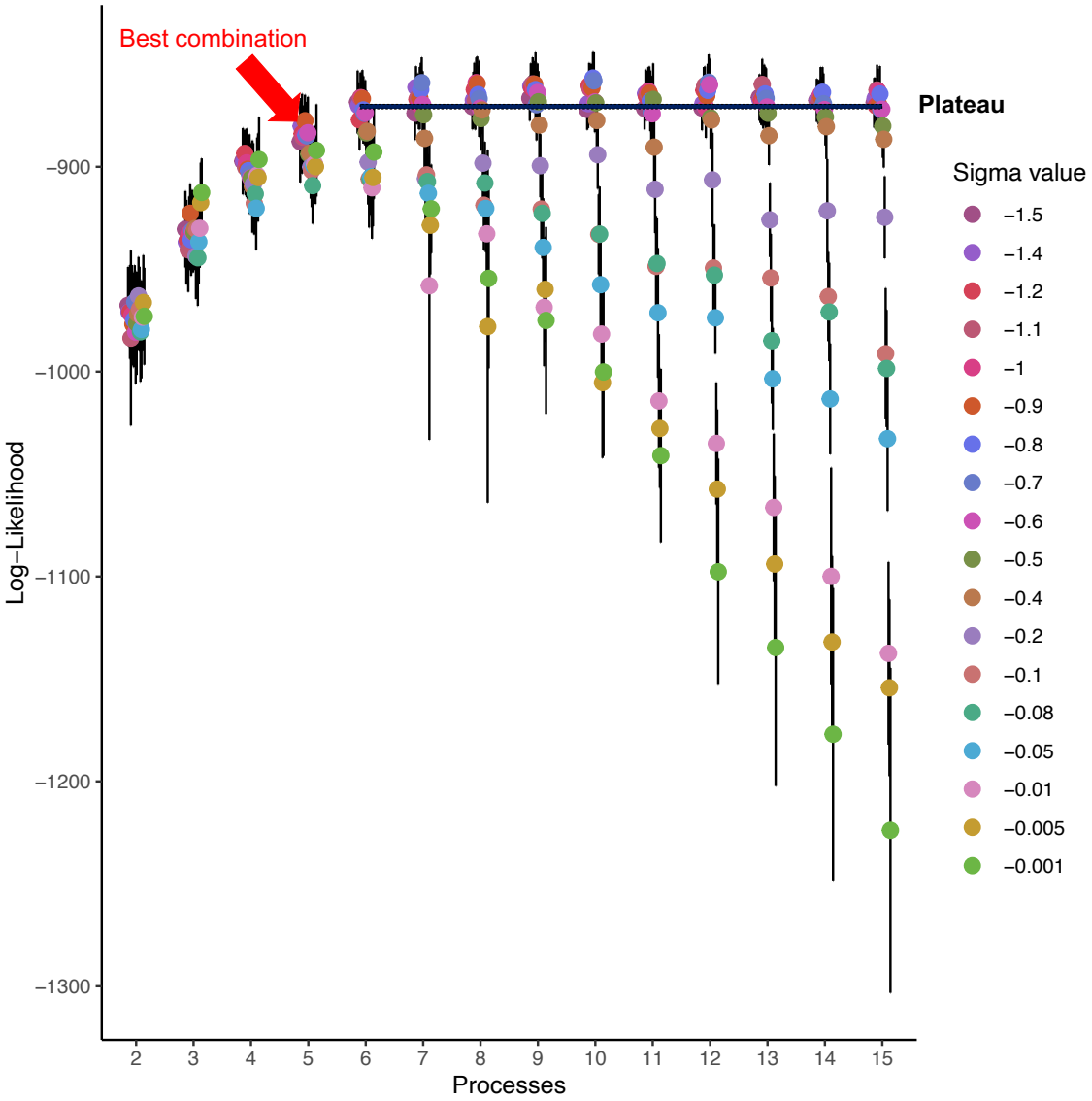

Figure 3.2: Example of log-likelihood estimation to identify the best combination of processes and sigmas. Each curve represents a specific sigma value and is colour-coded for clarity. Most curves exhibit a similar pattern. Initially, the log-likelihood increases as the number of processes increases, indicating a better fit to the data. However, as the number of processes increases, the curves reach a consensus plateau. This plateau signifies the point of overfitting, where additional processes do not significantly improve the performance of the model. To determine the best combination, the sigma value with the highest log-likelihood is picked as reference. From this sigma value, the combination with the least number of processes within the range of the standard deviation of the sigma value is selected. By picking this combination, we achieve a balance between capturing the complexity of the data without introducing unnecessary complexity. Typically, this combination is located just before the plateau.

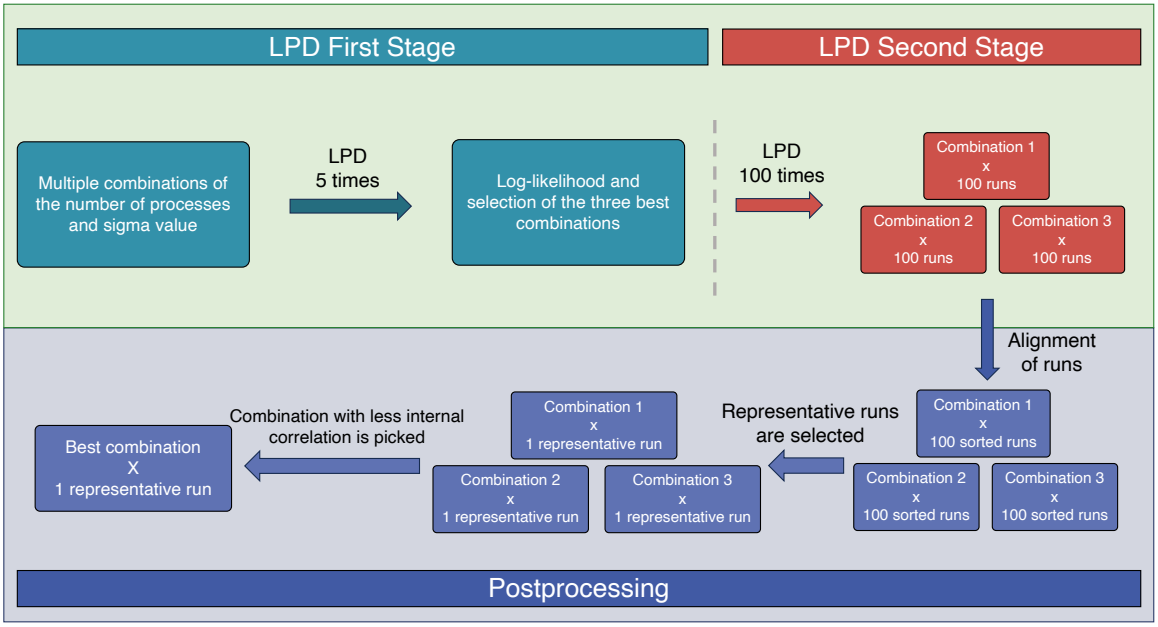

Figure 3.3: Schematic representation of the LPD application (green square) and the postprocessing step (purple square) of Automata. LPD application is divided into two stages: the first step (depicted in teal), in which the best parameters for the dataset are estimated, and the second step (shown in red), in which LPD is executed for the three best combinations of parameters. In the postprocessing step, the output from each of the combinations is compared, and the best combination is picked.

values of all other runs within the same combination. This comparison was performed using Spearman's rank correlation coefficient to assess the similarity between the gamma values of different runs. By aligning the processes based on their correlation with the reference run, a consistent comparison was achieved. The processes were then labelled according to the order observed in the reference run.

To identify the most representative run within each combination, I calculated the medoid gamma values and selected the run that was closest to the medoid. This approach ensured that the selected run captured the central tendency of the gamma values for each process within the same combination. These runs were considered as "representants" of their respective combination of parameters.

To compare the three representative runs obtained from the different combinations of parameters, I calculated the internal correlation for each run using Pearson's correlation coefficient. This correlation analysis was performed within the gamma values of the processes detected by each representative run. This analysis aimed to assess the degree of similarity or overlap between the processes identified by each run. Based on the results of the correlation analysis, I discarded the two representative runs with the highest correlation. The rationale behind this decision was to prioritize identifying uniquely distinct processes and minimize redundancy in the analysis.

In the final step, each sample of the same cancer type was assigned to a specific group, referred to as "LPD groups," based on the prevalence of a particular process within that sample. This assignment was determined by examining the gamma values from the output of the run with the lowest internal correlation. The LPD groups were labelled with the corresponding number of the most abundant process within each group.

3.2.5 Differential analysis

The differential analysis step compares the molecular landscape and clinical outcomes of the samples assigned to a specific LPD group with those of all other samples within the same cancer type. Specifically, this analysis aims to identify differences in gene expression patterns, methylation profiles, mutations, and copy number variations, as well as clinical characteristics, including patient survival. By comparing these variables, I aim to characterise the LPD process that is more prevalent in each LPD group, which serves as a quantifiable representation of the subtypes present in the samples. Therefore, the term "subtype" will be used with the semantic meaning of a process that holds biological significance.

Batch effect proportions

The batch effect refers to a non-biological variation in experimental data that can impact the outcome of an analysis. In this research, the batch effect was defined as factors that the LPD algorithm may mistakenly identify as molecular processes but are actually caused by variations in sample handling or experimental procedures. Two potential sources of batch effect in the TCGA data were identified: the tissue source site (TSS) centre and the preservation technique used for the samples. Samples processed at different TSS centres may exhibit discrepancies due to variations in machinery calibration or experimental protocols, leading to distinct molecular features. Additionally, differences in RNA preservation quality between frozen and FFPE samples could introduce variability that LPD may interpret as

separate sample populations. By identifying and accounting for these batch effects, the accuracy and reliability of the LPD analysis can be enhanced, ensuring that the identified molecular processes are truly reflective of the underlying biology.

To assess the potential presence of a batch effect in the assignment of samples to LPD groups, a chi-square test was conducted. This statistical test was used to detect significant disproportions of the tissue source site (TSS) centres or the type of sample preservation (frozen vs FFPE) across the LPD groups.

Diferentially expressed genes (DEGs)

The differential expression analysis was performed in two separate ways depending on if the expression data was obtained through RNA-seq or microarray. In the RNA-seq count samples, the R package `DESeq2` was utilized, whilst the R package `limma` was used in the microarray samples. Genes were considered significantly differentially expressed if they exhibited an absolute log2 fold change greater than 1 when comparing across LPD groups.

Methylation level analysis

A matrix was generated to capture the methylation level per CpG site in the genome and the associated genes using the R package `limma` in the downloaded methylation array data. To ensure data quality and eliminate redundancy, duplicated probes were filtered out. Mean log2 fold change was calculated for probes covering the same genes, and a threshold of significance of absolute log2 fold change of 1.5 was applied.

Differential analysis of single nucleotide variants (SNVs)

The R package `maftools` was employed to analyze the MAF consensus file, focusing on various aspects such as variant classification ratio, variant types, number of variants per sample, and single base mutation types. To assess the impact of single nucleotide variants (SNVs) on gene-level alterations within the LPD groups, a log2 ratio was computed to determine whether a gene displayed a higher (termed as overmutated) or lower (termed as undermutated) frequency of SNVs in one group compared to the other groups.

Analysis of the copy number variations (CNVs)

Segment mean values per probe were obtained from the TCGA, and a threshold of significance of an absolute segment value greater than 2 was applied. Positive values were interpreted as amplifications, while negative values were interpreted as deletions. Through the R package `GenomicRanges`, probes were mapped to their respective genes based on genomic coordinates. The proportion of base pairs affected by CNVs was calculated for each sample and compared across LPD groups through a Wilcox test. Additionally, a log2 ratio was computed to evaluate whether a gene exhibited higher (termed as over-affected) or lower (termed as under-affected) frequency of CNVs in the samples belonging to a group when compared to other groups.

Euclidean hierarchical clustering

To compare the results of LPD with a traditional clustering approach, a Euclidean hierarchical clustering with complete linkage was conducted. The input for the clustering model

consisted of the same set of expression data from the 500 genes used in LPD. To facilitate direct comparisons, the number of clusters chosen for hierarchical clustering was the same as the number of LPD groups defined by LPD. This approach allowed for a straightforward evaluation of the similarities and differences between both approaches.

Analysis of COSMIC mutational signatures

An analysis of COSMIC signatures was conducted for all the LPD groups using the R package `MutationalPatterns`. The reference genome used for this analysis was the `BSgenome Homo sapiens UCSC hg38`. A heatmap was generated to visualize the contribution of each COSMIC signature to the average mutational profile of the LPD groups. This heatmap provided insights into the specific mutational patterns and underlying mutational processes associated with each LPD group. A description of each COSMIC mutational signature can be found in Appendix B.

Biological pathways and processes enrichment analysis

Enrichment analysis was performed to gain further insights into the biological significance of the differentially expressed genes, methylated genes, genes affected by SNVs, and genes impacted by CNVs. This analysis was conducted using the R packages `clusterProfiler` and `msigdb`, in combination with the KEGG database for biological pathways (p-value cut-off: 0.5; q-value cut-off: 0.01) and the GO database for biological processes (p-value cut-off: 0.5; q-value cut-off: 0.01; ontology: biological processes).

Additional evidence of a functional effect for differentially expressed genes

To gain a deeper understanding of the functional impact of DEGs within each LPD group, it is important to explore the connections between DEGs and other molecular alterations. To achieve this, I examined the overlap between DEGs, DMGs, genes affected by SNV, and genes impacted by CNVs through a Venn diagram. Identifying the common genes among these differentially altered gene sets can reveal potential mechanisms that contribute to the observed differential expression.

Clinical analysis

Clinical data obtained from the TCGA was utilized to investigate the association between LPD groups and various clinical factors. A proportion analysis was conducted to examine the enrichment or depletion of cancer stages, gender, race, primary pathology, and Gleason score (specific to prostate cancer) within each LPD group.

Survival analysis was performed to explore the relationship between LPD groups and patient survival. This analysis used the R packages `survival` and `survminer` to generate Kaplan-Meier estimation curves and conduct log-rank tests. The survival probability was assessed based on the number of days until death or the last follow-up as the time parameter, and the vital status of the patient as the event parameter.

In the case of prostate cancer, the time parameter was modified to biochemical recurrence when available. Additionally, a Wilcox test was conducted to compare the proportions of PSA across LPD groups in prostate cancer cases. Furthermore, a Chi-square analysis was

performed to evaluate the proportion of high and low Gleason score samples within each LPD group.

3.2.6 Report

The final step of the Automata workflow involves generating an HTML report that provides a comprehensive summary of the research findings (Fig. 3.4). The report includes significant outcomes from the differential analysis step, a detailed description of the methodology followed, and graphical representations and tables to present the results. The report is designed as a dashboard, organized into different tabs to showcase each detected LPD group separately.

Rmarkdown was employed in conjunction with the R package `Flexdashboard` to create this report. The interactive nature of the report allows users to explore the data more effectively. Graphics can be displayed with interactive features such as zooming in and filtering results. Similarly, tables in the report offer search and sorting functionalities, allowing users to locate specific entries or rearrange them based on their preferences (Fig. 3.5).

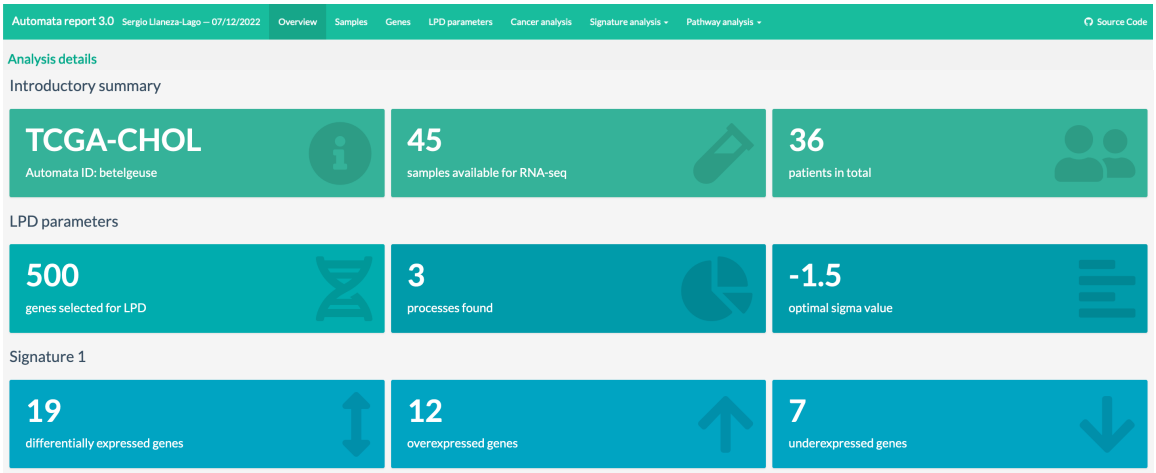

Figure 3.4: Screenshot of the Automata Report showcasing the Overview page, providing an overview of the example dataset for cholangiocarcinoma (TCGA-CHOL). It displays the number of samples and patients available in the dataset, the count of identified processes and the number of genes found to be differentially expressed within one of the processes.

3.3 Data availability

After submitting this dissertation, the generated reports for each of the 28 cancer types will be publicly available on a dedicated repository hosted on GitHub.

3.4 Conclusion

In this chapter, I have introduced the R package Automata, showcasing its significant role in automating the methodology employed in this dissertation. The availability of Automata as a freely accessible R package holds great value for the research community, as it streamlines

Figure 3.5: Screenshot of the Automata Report featuring the presentation of results. It includes a detailed explanation of the differential expression analysis, an interactive table presenting a list of the differentially expressed genes, and an interactive volcano plot displaying the distribution of such genes in terms of fold change.

the processing of TCGA data and facilitates the generation of interactive reports with user-friendly descriptions of the methodology.

By providing an automated solution, Automata enhances the reproducibility of research in cancer genomics. Researchers can utilise the functionalities of the package to efficiently process and analyse multi-omics data from TCGA, promoting transparency and ensuring that the results can be replicated and validated by others in the scientific community.

Furthermore, the user-friendly nature of the interactive reports generated by Automata enables non-technical readers to comprehend the methodology and findings of the research easily. This accessibility encourages broader engagement and understanding of the research outcomes beyond the academic and scientific community.

In summary, the availability of Automata as an R package, its automation of TCGA data processing, and the generation of interactive reports with non-technical descriptions all contribute to its significance as a valuable resource for cancer multi-omics research. It serves as a tool to promote reproducibility and facilitate the dissemination of research findings to a wider audience.

3.5 Summary

In this chapter, I presented the R package Automata, designed to streamline the analysis of multi-omics cancer data from the TCGA. The package's workflow is organized into six major steps: data download, data preprocessing, LPD application, postprocessing, differential analysis, and report generation. Automata facilitates the analysis of multiple cancer types, and its modular approach ensures ease of use, flexibility, and reproducibility, making it accessible to researchers regardless of coding experience.

A novel advancement presented in this chapter is the automated determination of the optimal number of processes for the LPD algorithm. Traditionally, this selection was done manually by visually inspecting a log-likelihood graph. The development of an automated approach streamlines the analysis and improves its consistency and accuracy.

In conclusion, Automata represents a significant advancement in cancer genomics research, providing a user-friendly tool to efficiently process and analyze multi-omics data from TCGA and to facilitate the investigation and interpretation of cancer subtypes across diverse cancer types.

Chapter 4

Pancancer analysis of subtypes detected by Latent Process Decomposition

4.1 Introduction

Previous research using the Latent Process Decomposition (LPD) algorithm has focused on its application to a single cancer type but lacks the overall and shared molecular characteristics across cancers[118,171]. The meta-analysis of multiple cancer types for shared genomic features is frequently referred to as "pancancer" and is key to accessing a better understanding of cancer biology, as illustrated by previous research: Ma et al. (2021)[245] described that mutations in the *TP53* gene drive high-grade serous ovarian, serous endometrial and basal breast carcinomas; likewise, Weinstein et al. (2013) found that the gene *ERBB2-HER2* is amplified in subsets of glioblastoma, gastric, serous endometrial, bladder and lung cancer. In some cases, the same genetic aberration can have different effects depending on which organ occurs, e.g. the NOTCH gene family is activated by mutations in leukaemias but stays inactivated in squamous cell cancers[246–248]. Not only gene mutations can be analysed in pancancer studies but also molecular features such as the intratumoral genetic heterogeneity (ITH), meaning by this the coexistence of genetically distinct subpopulations in a tumour[249]. ITH may be a critical factor in the resistance to targeted cancer therapy and is considered a likely indicator of the potential of the tumour for evolutionary adaptation[249,250].

Pancancer analyses may focus on specific cancer groups to analyse traits that distinguish them from other groups[251]. One possible classification criterion is the histology of the cancer type. TCGA contains carcinoma, sarcoma, melanoma, leukaemia, thymoma, and mixed histologies (Table 4.1). Carcinomas arise in the epithelial tissue covering the lining of the organs, passageways, and skin[252]. They can be further subclassified into adenocarcinomas when they originate from mucous membranes or squamous cell carcinomas when they arise in the squamous cells of the epithelium. Sarcomas are cancers that originate in connective and supportive tissue such as bone, cartilage, muscles, tendons, and fat, although some rare types can develop in the brain (gliosarcomas)[252,253]. Melanoma refers to skin cancers that develop in the melanocytes, a group of cells responsible for skin pigmentation, and are associated with exposure to ultraviolet rays[254,255]. Leukaemia includes tumours that

affect the bone marrow, where blood cells are made, which is why it is commonly referred to as blood cancer[252]. Thymomas are a particular type of carcinoma that originates in the epithelial cells of the thymus and histologically resemble non-cancerous cells[256]. The mixed type is employed when a cancer type can be classified into more than one type simultaneously; an example would be an adenosquamous carcinoma[252].

Table 4.1: Primary histological type and subtype for each cancer project downloaded from TCGA.

Cancer type	Project code	Primary histological type	Histologial subtype
Bladder Urothelial Carcinoma	TCGA-BLCA	Carcinoma	Mixed
Breast Invasive Carcinoma	TCGA-BRCA	Carcinoma	Mixed
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	TCGA-CESC	Carcinoma	Adenosquamous
Cholangiocarcinoma	TCGA-CHOL	Carcinoma	Mixed
Colon Adenocarcinoma	TCGA-COAD	Carcinoma	Adenocarcinoma
Esophageal Carcinoma	TCGA-ESCA	Carcinoma	Mixed
Glioblastoma Multiforme	TCGA-GBM	Sarcoma	Glioma
Head and Neck Squamous Cell Carcinoma	TCGA-HNSC	Carcinoma	Squamous
Kidney Chromophobe	TCGA-KICH	Carcinoma	Mixed
Kidney Renal Clear Cell Carcinoma	TCGA-KIRC	Carcinoma	Mixed
Kidney Renal Papillary Cell Carcinoma	TCGA-KIRP	Carcinoma	Mixed
Acute Myeloid Leukemia	TCGA-LAML	Leukemia	Myelogenus
Brain Lower Grade Glioma	TCGA-LGG	Sarcoma	Glioma
Liver Hepatocellular Carcinoma	TCGA-LIHC	Carcinoma	Carcinoma
Lung Adenocarcinoma	TCGA-LUAD	Carcinoma	Adenocarcinoma
Lung Squamous Cell Carcinoma	TCGA-LUSC	Carcinoma	Squamous
Ovarian Serous Cystadenocarcinoma	TCGA-OV	Carcinoma	Adenocarcinoma
Pancreatic Adenocarcinoma	TCGA-PAAD	Carcinoma	Adenocarcinoma
Pheochromocytoma and Paraganglioma	TCGA-PCPG	Sarcoma	Glioma
Prostate Adenocarcinoma	TCGA-PRAD	Carcinoma	Adenocarcinoma
Rectum Adenocarcinoma	TCGA-READ	Carcinoma	Adenocarcinoma
Sarcoma	TCGA-SARC	Sarcoma	Sarcoma
Skin Cutaneous Melanoma	TCGA-SKCM	Melanoma	Melanoma
Stomach Adenocarcinoma	TCGA-STAD	Carcinoma	Adenocarcinoma
Testicular Germ Cell Tumors	TCGA-TGCT	Mixed	Mixed
Thyroid Carcinoma	TCGA-THCA	Carcinoma	Mixed
Thymoma	TCGA-THYM	Thymona	Neoplasm
Uterine Corpus Endometrial Carcinoma	TCGA-UCEC	Carcinoma	Mixed

A few illnesses are associated with inherited genetic mutations that can increase cancer risk by affecting tumour-suppressor genes. Multiple studies perform pancancer analysis of the cancer types related to these diseases[257]. Some of the most common are (Table 4.2): (i) Mutations in the BRCA1 and BRCA2 genes are typically associated with breast cancer (70% of those who carry this mutation) but also increase the risk of ovarian cancer (45% of carriers) and, to a lesser extent, of prostate cancer and pancreatic cancer[258]. (ii) Lynch syndrome is a hereditary predisposition to colorectal cancer (70% of carriers) in addition to uterine, ovarian, stomach, prostate, and bladder cancer[259]. This syndrome is associated with mutations in the *MLH1*, *MSH2*, *MSH6*, and *PMS2* genes[259]. (iii) Li-Fraumeni syndrome is caused by a mutation in the gene *TP53* (involved in cell division), and it increases the risk of developing breast cancer, bone cancer, myeloid leukaemia, soft tissue sarcoma, brain tumours, and adrenal gland cancer[260]. (iv) PTEN hamartoma tumour syndrome is linked to a mutation

in the gene *PTEN* and increases the risk of developing breast cancer, thyroid cancer, uterine corpus cancer, colorectal cancer, kidney cancer and skin melanoma[261,262]. (v) Familial adenomatous polyposis is caused by a mutation in the *APC* gene and is associated with 1% of all colorectal cancers; additionally, it increases the risk of developing stomach, pancreatic and liver cancer[263]. (vi) Mutations in the MUTYH gene can lead to a MUTYH-associated polyposis disease, which increases the risk of colorectal, bladder, breast, uterine corpus, and ovarian cancer[259]. (vii) Finally, Peutz-Jeghers syndrome is linked to mutations in the *STK11* gene that increase the risk of developing breast, colorectal, pancreatic, stomach, and ovarian cancer[259].

Table 4.2: Inherited mutations or syndromes and to which cancer types they are related.

TCGA project	BRCA cancer	Lynch	Li-Fraumeni	PTEN cancer	Adenomatous	MUTYH cancer	Peutz-Jeghers
TCGA-BLCA		X				X	
TCGA-BRCA	X		X	X		X	X
TCGA-COAD		X		X	X	X	X
TCGA-GBM			X				
TCGA-KIRC				X			
TCGA-KIRP				X			
TCGA-LAML			X				
TCGA-LGG			X				
TCGA-LIHC					X		
TCGA-OV	X	X				X	X
TCGA-PAAD	X				X		X
TCGA-PRAD	X	X					
TCGA-READ		X		X	X	X	X
TCGA-SARC			X				
TCGA-SKCM				X			
TCGA-STAD		X			X		X
TCGA-THCA				X			
TCGA-UCEC		X		X		X	

Despite the rising popularity of pancancer analyses, most are performed by comparing whole cancer types, thereby overlooking the shared features and similarities between subtypes of different cancers. Whole-cancer comparisons conceal the molecular processes that thrive the stratification of the tumour into subpopulations by favouring the detection of differential carcinogenic processes instead. Thus, in this chapter, I aim to perform a pancancer analysis across the subtypes of the 28 cancer types obtained through Automata to unravel the biological processes shared across cancer types that play a role in the stratification and ITH of the tumour. In addition, the gamma values generated by LPD on each cancer type will be examined to gain an insight into the genetic diversity and variance of the disease. I hypothesise that the ITH is related to the number of processes detected for each cancer type and that several genes are commonly differentially expressed across subtypes from distinct cancer types.

4.2 Methodology

The data used in this chapter was gathered and processed following the Automata workflow that is explained in detail in section 3.2. The specifics of the statistical tests, databases and computational resources are described in chapter 2.

4.2.1 Study of the gamma values

Descriptive statistics were calculated for the gamma values from all cancer datasets. The same operation was repeated separating the data into RNA-seq and microarray sets. For all the sets, the frequency of each process number was calculated and represented by kernel density estimation. A linear model was built to examine the relationships between the total number of processes and the number of available samples for each cancer type.

To study the ITH level of each cancer dataset, it was assumed that cancers with perfect ITH would have uniformly distributed gamma values. Following this principle, the mean difference $\bar{X}$ between these uniformly distributed values and the observed values O for each cancer type was calculated as

$$\bar{X} = \frac{\sum_{i=1}^{n} \left| \frac{1}{n} - O_i \right|}{n}, \tag{4.1}$$

where n represents the total number of processes for a cancer type. The first quartile of cancers with the smallest difference was ranked as "High", the second and third quartile as "Medium", and the fourth quartile as "Low". The ratio of datasets from each level of ITH assigned between RNA-seq and microarray sets was compared with a Chi-squared test. An ANOVA and Student's t-test were performed test to compare the distribution of each ITH level according to the number of samples and the number of processes.

Cancer types with both microarray and RNA-seq datasets were selected to check whether the gamma values in the microarray and the RNA-seq datasets were equivalent and behaved similarly. The pattern in sample assignment to LPD groups was compared through an alluvial plot to look for similar behaviours. Spearman's correlation coefficient was used to detect hidden resemblances between the distribution of the gamma values of each microarray LPD group and all RNA-seq LPD groups for the same cancer type.

Cancer types datasets were categorised according to histology and, in parallel, their link to hereditary disorders and mutations to check whether this would reveal new molecular features. Microarray datasets were excluded due to their lack of representation of histologies and cancer types associated with hereditary disorders and mutations. Kernel density estimation was performed on the gamma values for all RNA-seq datasets, a linear model was built to analyse the correlations between the total number of processes and the number of samples for each category, and the proportion of ITH levels was assessed via Chi-squared test.

4.2.2 Common differentially expressed genes across cancers

The differentially expressed genes were only examined in the RNA-seq data.

To find common differentially expressed genes (DEGs) for all cancer types, the DEGs obtained through Automata were filtered to pick only the significant ones (absolute log2 Fold Change >1; adjusted p-value < 0.05). The same criteria were followed for differentially

methylated genes (DMGs). In the case of genes differentially mutated by single nucleotide variants (SNVs) and impacted by copy number variations (CNVs), only those with an absolute log2 ratio over 1 were considered significant.

The most common DEGs across subtypes were selected and matched with the significant DMGs, significant genes differentially affected by SNVs, and significant genes differentially impacted by CNVs to gain an insight into the mechanisms affecting the normal expression profile. An enrichment analysis was conducted through KEGG (*p-value cutoff: 0.05, q-value cutoff: 0.01*) and GO (*p-value cutoff: 0.05, q-value cutoff: 0.01, ontology: BP, adjustment: BH, universe: all significant DEG*). This procedure was repeated for each category of cancer based on histology or association with inherited diseases (see table 4.1 and table 4.2).

Driver genes ($n = 199$) were imported from Bailey et al. (2018)[264] to learn more about the roles of the detected DEGs. Driver genes present as DEG in at least two molecular subtypes from distinct cancer types were selected and, similarly to earlier, they were matched with the significant DMGs, significant genes differentially affected by SNVs, and significant genes differentially impacted by CNVs.

Similarities between molecular subtypes from distinct cancer types were calculated through a Jaccard similarity index.

4.3 Results

4.3.1 Overview of the gamma values

The mode number of processes for all datasets was seven (37% of datasets). For microarray datasets, the median was five processes while for RNA-seq ones it was seven (Fig. 4.1). Both transcriptome platforms displayed a significant positive relationship between the number of processes and the number of samples (Fig. 4.2).

Each dataset was classified as low, median and high ITH (Table 4.3). All datasets classified as high ITH were RNA-seq datasets (Table 4.3). Datasets with high ITH contained gamma values around 0.1 more often than those datasets with medium and low levels, which instead tended towards zero (Table 4.3). However, there seemed to be a pattern where the less ITH, the more gamma values were distributed over the range from zero to one (Fig. 4.3). There was a significant association between a dataset's ITH level and the number of processes detected and the number of samples available: datasets with high ITH had higher numbers of processes and samples compared to the datasets with low ITH (Fig. 4.4). In RNA-seq, the datasets with high ITH had a significantly higher number of processes than the medium datasets (Fig. 4.4). When comparing the distribution of the level of ITH between the microarray and RNA-seq datasets, the Chi-squared test returned a significant difference ($p - value : 0.001, X^2 = 13.03$), which was attributed to the overrepresentation of datasets with low ITH in the microarray set.

When comparing the LPD assignment between microarray and RNA-seq datasets, the alluvial plot revealed a lack of perfect matches between the homologous LPD groups, except for those composed of normal tissue samples (Fig. 4.5). However, the correlation analysis showed strong correlations with matches for 90% or more of the LPD groups for TCGA-BRCA, TCGA-COAD, TCGA-KIRC, TCGA-LAML, TCGA-OV and TCGA-READ; average results with matches ranging from 50-75% for TCGA-GBM, TCGA-LGG and TCGA-

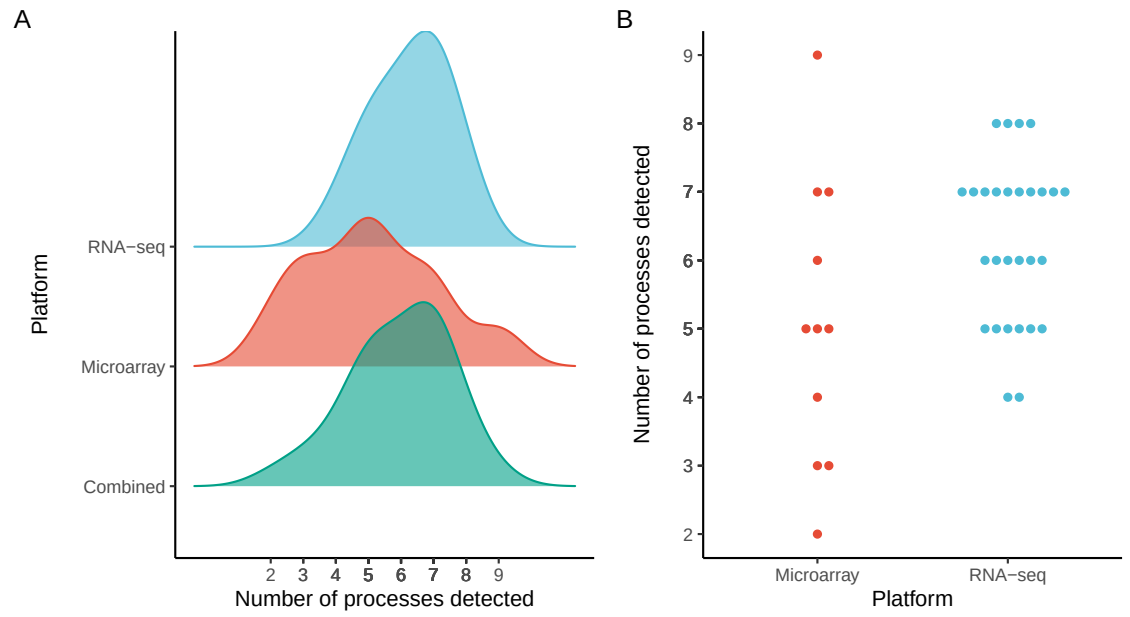

Figure 4.1: The distribution of the number of processes detected for TCGA cancer datasets obtained from RNA-seq and microarray is represented as (A) a density plot and (B) a beeswarm plot.

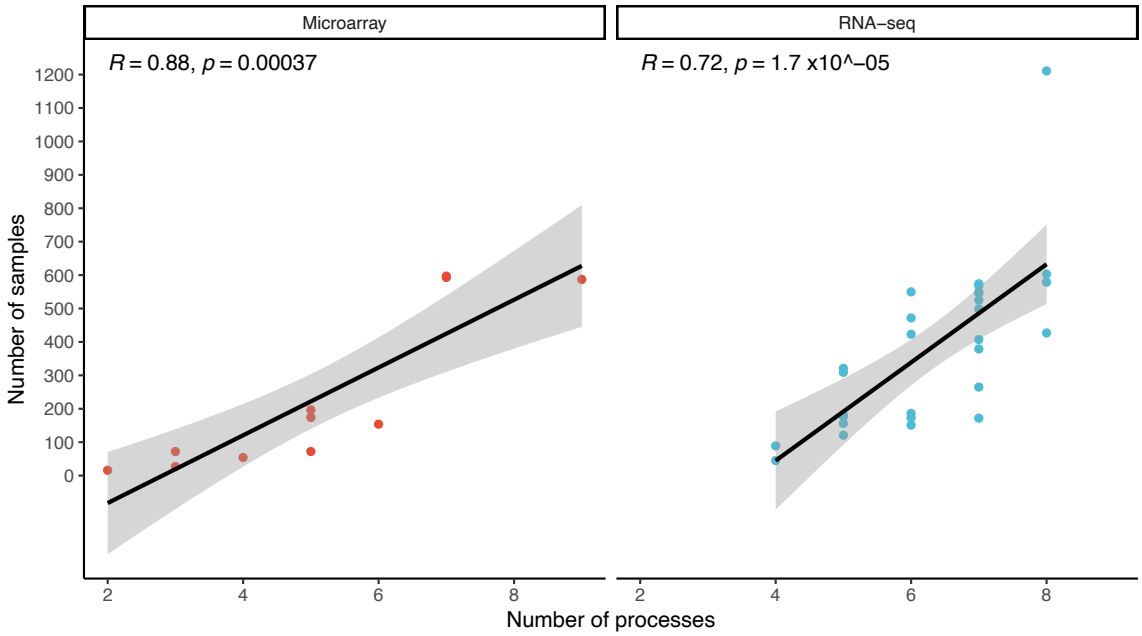

Figure 4.2: Scatter plot of the number of processes detected by LPD and the number of samples in both RNA-seq and microarray datasets. The trend line is calculated using a linear model and the grey shading is the confidence interval.

Table 4.3: Heterogeneity level for each of the TCGA datasets. Projects are allocated into high, medium, or low tier according to their mean difference. The classification process is repeated separately for projects within the same platform (RNA-seq and microarray) and across different platforms.

TCGA project	Mean difference	ITH level across the same platform	ITH level across both platforms
RNA-seq			
TCGA-SKCM	0.1034418	High	High
TCGA-OV	0.1070768	High	High
TCGA-UCEC	0.1079444	High	High
TCGA-STAD	0.1106305	High	High
TCGA-LUAD	0.1109764	High	High
TCGA-BRCA	0.1132777	High	High
TCGA-BLCA	0.1133887	High	High
TCGA-SARC	0.1188576	Medium	High
TCGA-KIRC	0.1193732	Medium	High
TCGA-PCPG	0.1199632	Medium	High
TCGA-PRAD	0.1229284	Medium	Medium
TCGA-COAD	0.1312153	Medium	Medium
TCGA-LIHC	0.1326119	Medium	Medium
TCGA-HNSC	0.1334104	Medium	Medium
TCGA-LUSC	0.1362890	Medium	Medium
TCGA-READ	0.1380267	Medium	Medium
TCGA-CESC	0.1476822	Medium	Medium
TCGA-THCA	0.1533392	Medium	Medium
TCGA-LGG	0.1618897	Medium	Medium
TCGA-KIRP	0.1656156	Medium	Medium
TCGA-PAAD	0.1685146	Medium	Medium
TCGA-GBM	0.1767068	Low	Medium
TCGA-LAML	0.1809022	Low	Medium
TCGA-ESCA	0.1868125	Low	Medium
TCGA-THYM	0.1875569	Low	Medium
TCGA-TGCT	0.2129326	Low	Low
TCGA-KICH	0.2618135	Low	Low
TCGA-CHOL	0.2835094	Low	Low
Microarray			
TCGA-LAML	0.1461098	High	Medium
TCGA-OV	0.1515230	High	Medium
TCGA-BRCA	0.1631411	High	Medium
TCGA-GBM	0.1791827	Medium	Medium
TCGA-COAD	0.2194057	Medium	Low
TCGA-LUSC	0.2438373	Medium	Low
TCGA-READ	0.2453690	Medium	Low
TCGA-KIRC	0.2484054	Medium	Low
TCGA-UCEC	0.2533182	Low	Low
TCGA-LGG	0.3191980	Low	Low
TCGA-KIRP	0.3690866	Low	Low

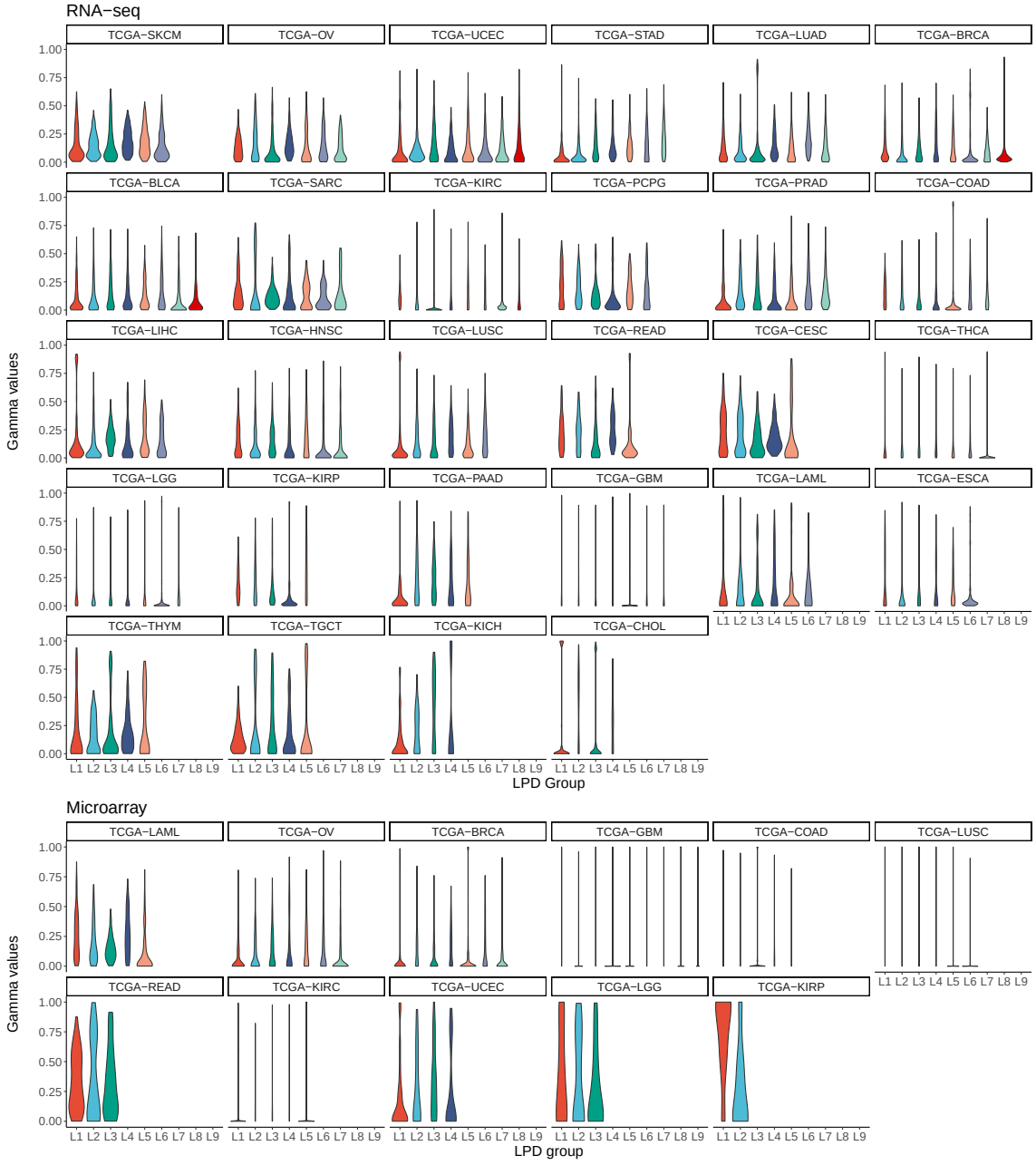

Figure 4.3: Violin plot showing the distribution of gamma values for all cancer datasets sorted by ITH level. Datasets are divided according to their platform of origin into RNA-seq and microarray.

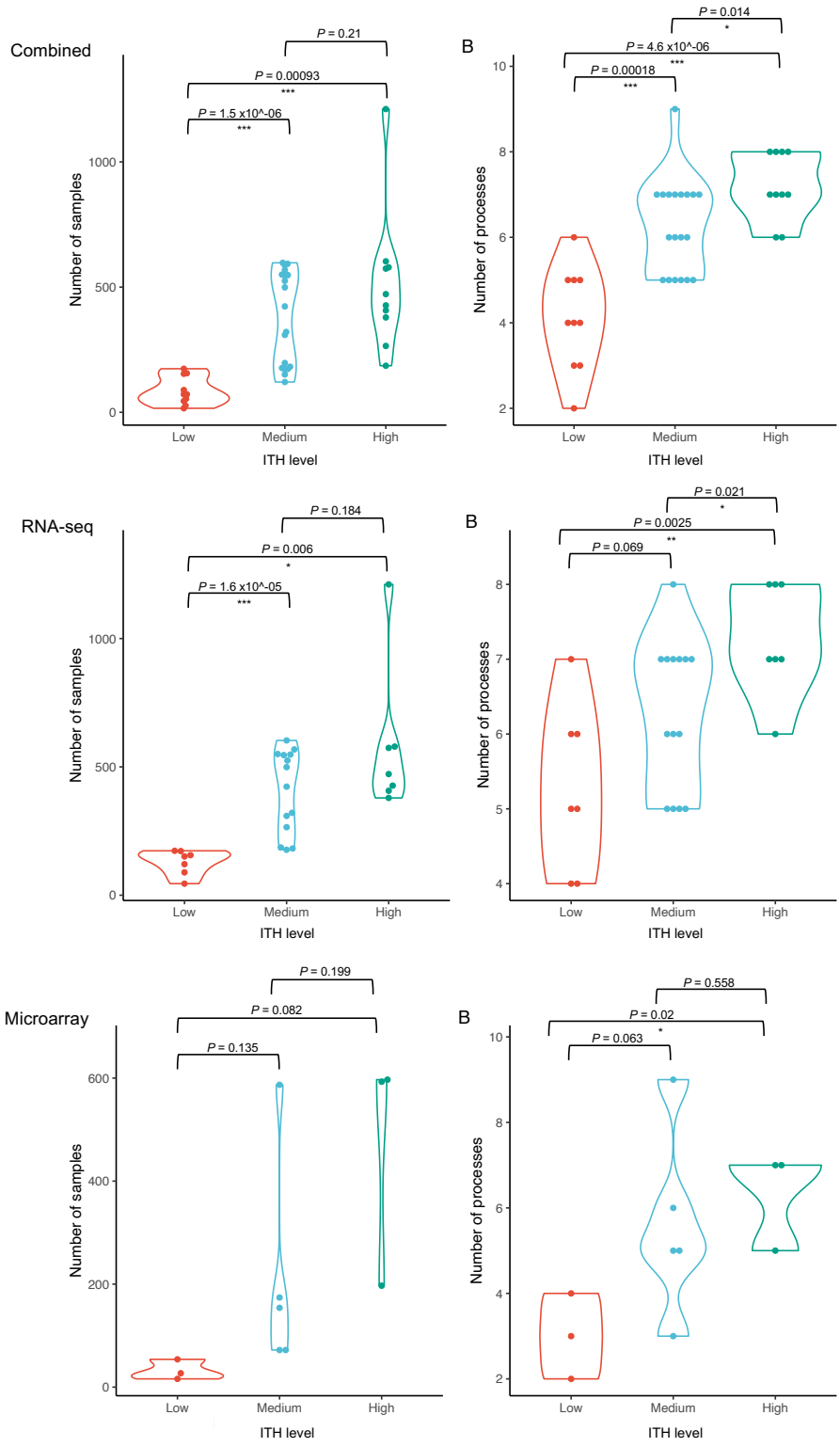

Figure 4.4: Violin plot illustrating the relationship between ITH and the number of processes and samples for each dataset. The datasets are presented both combined and divided based on their platform of origin (RNA-seq and microarray). The left column of the plot represents the correlation between the number of samples and the ITH level, while the right column displays the relationship between the number of processes and the ITH level.

UCEC; and no matches for TCGA-KIRP and TCGA-LUSC. All correlation matrix plots are depicted in the appendix C.

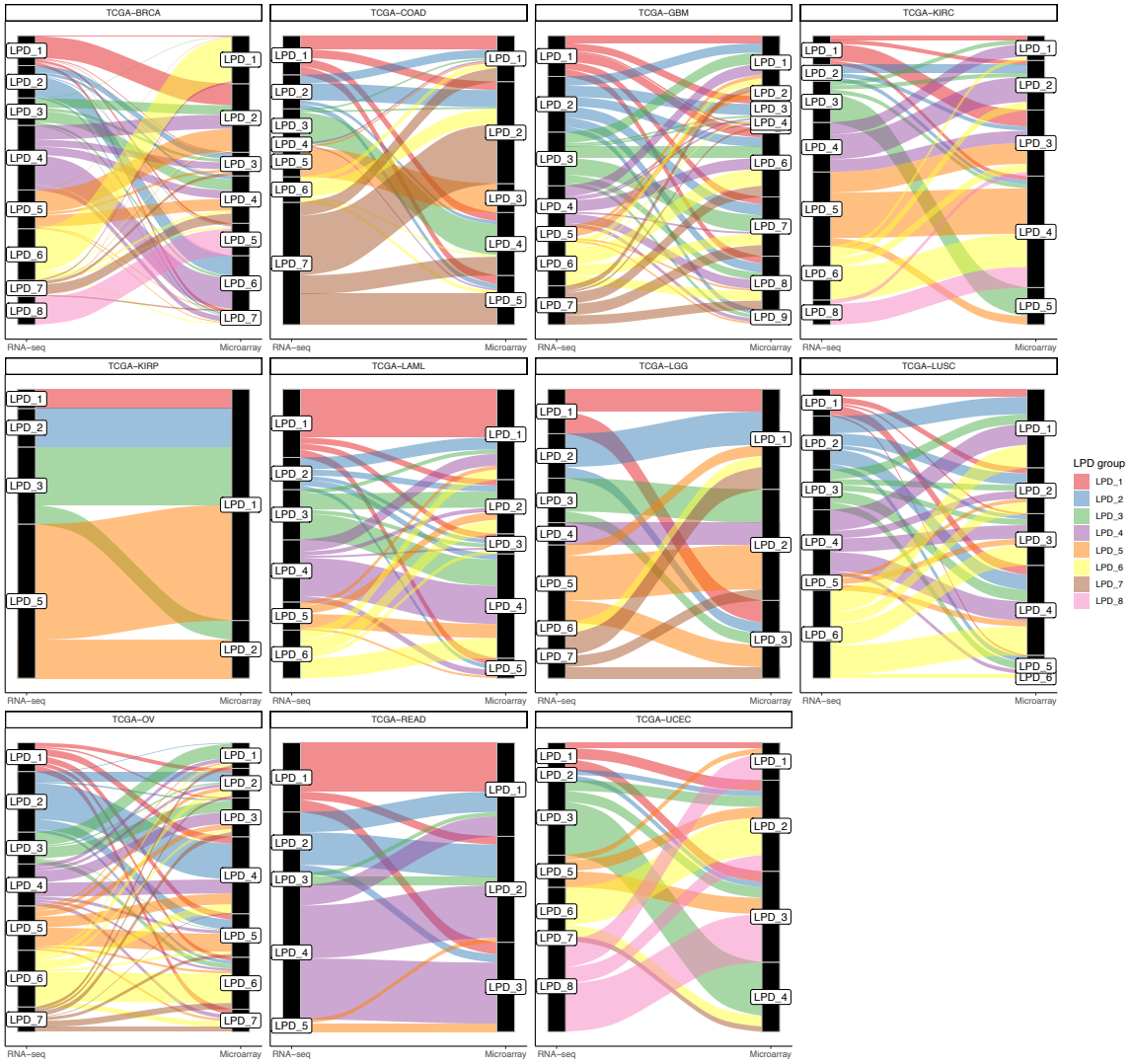

Figure 4.5: Alluvial plot comparing the assignment of matching samples in RNA-seq to microarray across the 11 datasets in common for both platforms. Each colour represents an LPD group detected in the RNA-seq analysis and how it is allocated in the microarray analysis.

When comparing cancer types based on their primary histological group (Table 4.1), I observed that certain types, including melanoma, leukaemia, thymoma, and mixed histologies, had only one cancer type available in the TCGA. Specifically, melanoma had only available skin cutaneous melanoma, leukaemia had only acute myeloid leukaemia, thymoma was the only thymoma-like cancer type, and the mixed histologies category was formed solely by testicular germ cell tumours. Due to the limited representation and diversity in these groups, I decided to exclude them from further analysis to ensure robust and meaningful comparisons. On the other hand, carcinomas had 20 different cancer types available, while sarcomas had four, making it possible to conduct a meaningful comparison between

these categories. Despite the variation in the number of cancer types, both carcinomas and sarcomas showed a similar distribution of gamma values, with seven processes being the most frequently occurring value. Additionally, when analyzing the distribution of gamma values according to specific diseases, no significant differences were observed, indicating a consistent pattern across various cancer types within each histological group.

4.3.2 Recurrent differentially expressed genes that define subtypes across cancers

Across the 28 cancer types, Automata identified 33,217 different significant DEGs formed by protein-coding genes, pseudogenes and antisense genes, as well as 33,283 DMGs, 19,174 mutated genes, and 2620 genes impacted by aberrations. The complete list of all the genes is available in Supplementary Material A. Of the total DEGs, 25,304 were simultaneously present in the subtypes of two to 23 cancer types. A total of 160 distinct DEGs met my criteria for being designated "the most common ones", which required them to be present in at least 19 cancer types. Then, I matched each case of common DEG and cancer type (3188 cases) with the DMGs, mutated and chromosomic impacted genes (Figure 4.6).

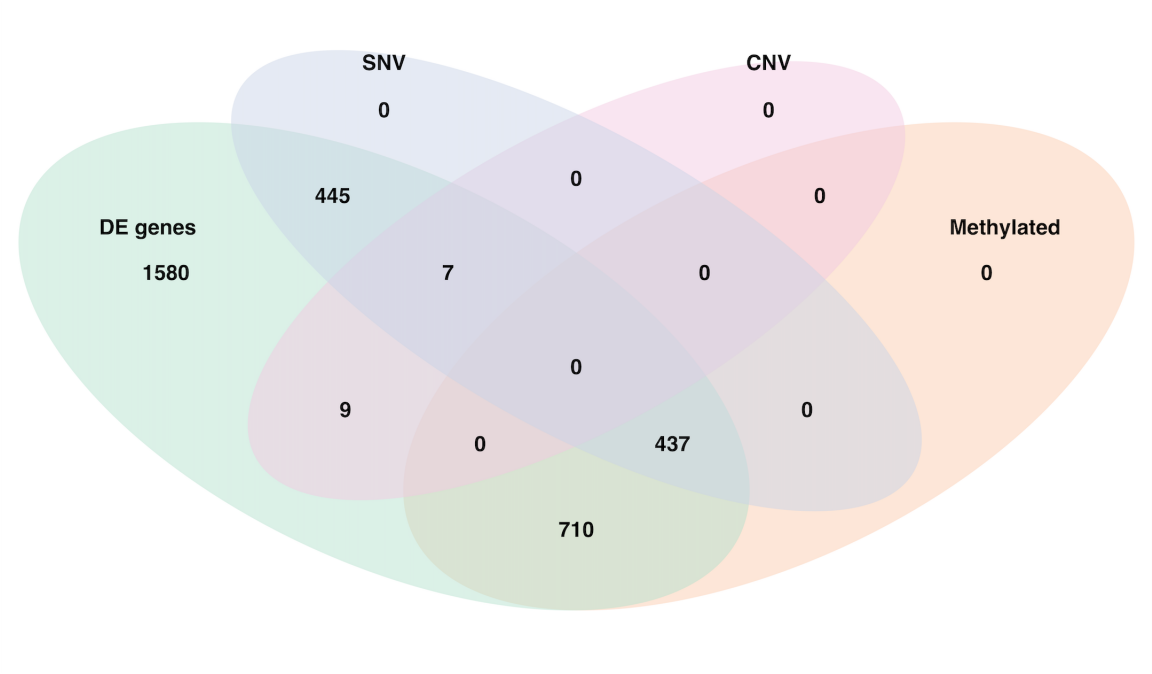

Figure 4.6: Venn diagram showing the matches of cases of most common DEGs with DMGs, significantly mutated genes (SNV) and genes affected by chromosomic aberrations (CNV).

The KEGG enrichment analysis of the 160 DEGs only returned three biological processes: PPAR signalling pathway, complement and coagulation cascades, and neuroactive ligand-receptor. GO enrichment analysis, on the other hand, revealed multiple immune-related pathways (Table 4.4). Individual enrichment studies for DMGs, altered genes, and chromosomal abnormalities yielded identical results with no discernible differences. The same phenomenon happened when cancer types were analysed based on their association with inheritable syndromes or their histology.

Table 4.4: Gene Ontology enrichment analysis outcome when comparing the most common DEGs against all DEGs. For each biological process, the gene ratio and the adjusted p-value is included. Only significant biological processes are shown.

Description	GeneRatio	p.adjust
humoral immune response	44/138	< 0.0001
immunoglobulin production	33/138	< 0.0001
production of molecular mediator of immune response	37/138	< 0.0001
protein activation cascade	35/138	< 0.0001
humoral immune response mediated by circulating immunoglobulin	31/138	< 0.0001
complement activation, classical pathway	30/138	< 0.0001
complement activation	32/138	< 0.0001
adaptive immune response	48/138	< 0.0001
immunoglobulin mediated immune response	31/138	< 0.0001
B cell mediated immunity	31/138	< 0.0001
phagocytosis	34/138	< 0.0001
immune response-activating signal transduction	39/138	< 0.0001
immune response-activating cell surface receptor signaling pathway	35/138	< 0.0001
activation of immune response	41/138	< 0.0001
immune response-regulating signaling pathway	39/138	< 0.0001
immune response-regulating cell surface receptor signaling pathway	35/138	< 0.0001
lymphocyte mediated immunity	32/138	< 0.0001
regulation of humoral immune response	24/138	< 0.0001
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	31/138	< 0.0001
endocytosis	40/138	< 0.0001
regulation of complement activation	22/138	< 0.0001
regulation of protein activation cascade	22/138	< 0.0001
immune response-regulating cell surface receptor signaling pathway involved in phagocytosis	20/138	< 0.0001
Fc-gamma receptor signaling pathway involved in phagocytosis	20/138	< 0.0001
Fc-epsilon receptor signaling pathway	20/138	< 0.0001
Fc receptor mediated stimulatory signaling pathway	20/138	< 0.0001
Fc-gamma receptor signaling pathway	20/138	< 0.0001
acute inflammatory response	25/138	< 0.0001

Table 4.4: Gene Ontology enrichment analysis outcome when comparing the most common DEGs against all DEGs. For each biological process, the gene ratio and the adjusted p-value is included. Only significant biological processes are shown. *(continued)*

Description	GeneRatio	p.adjust
regulation of acute inflammatory response	22/138	< 0.0001
regulation of protein processing	22/138	< 0.0001
regulation of protein maturation	22/138	< 0.0001
Fc receptor signaling pathway	20/138	< 0.0001
protein processing	26/138	< 0.0001
protein maturation	26/138	< 0.0001
regulation of immune effector process	28/138	< 0.0001
receptor-mediated endocytosis	24/138	< 0.0001
leukocyte migration	30/138	< 0.0001
response to bacterium	32/138	< 0.0001
regulation of inflammatory response	27/138	< 0.0001
defense response to bacterium	23/138	< 0.0001
phagocytosis, recognition	13/138	< 0.0001
B cell receptor signaling pathway	14/138	< 0.0001
defense response to other organism	23/138	< 0.0001
phagocytosis, engulfment	13/138	< 0.0001
plasma membrane invagination	13/138	< 0.0001
membrane invagination	13/138	< 0.0001
positive regulation of B cell activation	13/138	< 0.0001
regulation of proteolysis	26/138	< 0.0001
B cell activation	16/138	< 0.0001
regulation of B cell activation	13/138	< 0.0001
antigen receptor-mediated signaling pathway	14/138	< 0.0001
antimicrobial humoral response	11/138	< 0.0001
cell recognition	13/138	< 0.0001
positive regulation of leukocyte activation	16/138	< 0.0001
positive regulation of cell activation	16/138	0.0001
positive regulation of lymphocyte activation	14/138	0.0002
regulation of cell activation	18/138	0.0012

Table 4.4: Gene Ontology enrichment analysis outcome when comparing the most common DEGs against all DEGs. For each biological process, the gene ratio and the adjusted p-value is included. Only significant biological processes are shown. *(continued)*

Description	GeneRatio	p.adjust
antimicrobial humoral immune response mediated by antimicrobial peptide	7/138	0.0014
regulation of leukocyte activation	16/138	0.0049
regulation of lymphocyte activation	14/138	0.0091
lymphocyte activation	17/138	0.011

When I compared the DEGs to driver genes, I discovered a total of 75 genes matching (Table 4.5). I also noticed that 13 of them were DMGs, mutated and affected by chromosomic aberrations too: *AXIN2*, *CACNA1A*, *CCND1*, *CD798*, *CDKN2A*, *CYSLTR2*, *EGFR*, *EGR3*, *ERBB2*, *FGFR1*, *FOXA1*, *KRT222*, and *PGR*.

Table 4.5: Driver genes found to be differentially expressed across cancer types and the number of cancer types in which they were found.

Gene	In how many cancer types	Gene	In how many cancer types
ALB	19	GRIN2D	4
ALK	7	HGF	11
APOB	17	IL7R	5
AR	7	IRF6	5
AXIN2	4	KEL	6
CACNA1A	4	KIF1A	18
CARD11	2	KIT	9
CCND1	3	KLF5	4
CD70	9	KRT222	8
CD79B	8	MECOM	4
CDH1	2	MET	2
CDKN2A	3	MUC6	19
CNBD1	4	MYCN	11
COL5A1	6	PAX5	15
CREB3L3	11	PDGFRA	4
CYSLTR2	5	PGR	9
DACH1	6	PIK3CG	5

Table 4.5: Driver genes found to be differentially expressed across cancer types and the number of cancer types in which they were found. *(continued)*

Gene	In how many cancer types	Gene	In how many cancer types
DMD	4	PLCB4	5
EGFR	4	PTCH1	3
EGR3	7	PTPRC	5
ELF3	3	PTPRD	4
EPAS1	2	RET	10
EPHA3	10	RHOB	2
ERBB2	4	RNF43	2
ERBB3	3	RXRA	2
ERBB4	12	SETBP1	2
ESR1	5	SMARCA1	2
FAT1	2	SOX17	3
FGFR1	2	SOX9	2
FGFR2	2	SPTA1	8
FGFR3	5	TBX3	5
FLNA	2	U2AF1	3
FLT3	3	UNCX	6
FOXA1	6	WT1	14
FOXA2	12	ZBTB20	5
FOXQ1	2	ZCCHC12	11
GABRA6	3	ZNF750	7
GATA3	8	NA	NA

Lastly, the Jaccard similarity index returned significant matches only for colon and rectal cancer (36%) when considering whole cancer types. When I focused on the subtypes, I discovered that the majority of them were between subtypes of the same cancer type , but I also detected a match between rectal and colon cancer subtypes (Table 4.6).

4.4 Discussion

4.4.1 Study of the gamma values

Gamma values provide information about the biological processes that contribute to the molecular diversity of the tumour. As predicted, the positive association between the number of samples and the number of processes suggests that the bigger the sample set, the

Table 4.6: Jaccard Similarity Index across LPD groups. Only the top five matches are shown.

LPD Group A	LPD Group B	Percentage of similarity
Kidney clear cell cancer (LPD_8)	Kidney clear cell cancer (LPD_7)	0.3911231
Rectal cancer (LPD_7)	Colon cancer (LPD_7)	0.3725998
Rectal cancer (LPD_8)	Rectal cancer (LPD_2)	0.3290226
Colon cancer (LPD_7)	Colon cancer (LPD_6)	0.3201380
Breast carcinoma (LPD_4)	Breast carcinoma (LPD_6)	0.3094262

more processes are required to reflect the genetic variation throughout the set. This dependence on sample count might explain the observed differences between the RNA-seq and microarray datasets. Because I used a minimum sample size cutoff solely on RNA-seq and not on microarrays, several cancer types exhibited a considerable variation in sample size and hence in the number of processes. The same issue arises at the ITH level since it is derived from the gamma values: a larger number of processes have a higher possibility of their characteristics overlapping, and therefore their gamma values would be evenly distributed, resulting in high ITH. However, because most microarray datasets contain fewer samples than their counterparts, microarray datasets account for the vast bulk of low ITH datasets. Thus, this phenomenon may cast doubt on my results at the ITH level analysis.

Still, I found consistency with the study performed by Morris et al. (2016)[250], in which they also classified as high ITH the RNA-seq datasets for TCGA-SKCM, TCGA-LUAD, TCGA-BRCA, TCGA-BLCA and TCGA-KIRC; in addition to TCGA-PRAD, TCGA-HNSC and TCGA-LUSC which in my case I set as the upper half of the medium level. As a result, while my criteria for dataset categorisation into tiers appeared capable of partially reflecting ITH, I believe it would be essential to revise them in future research for more accurate findings. Only 11 of the 28 cancer types in the TCGA have microarray data, and only four datasets exceed the sample size criteria of 100 that was applied to all RNA-seq datasets. This resulted in various inconsistencies, including those stated above, and the inability to compare the output from counterparty pairings reliably. As a result, I decided to exclude all microarray datasets from future studies. This posed the question of whether removing the information produced from the microarray datasets would prevent me from making an appropriate interpretation of the subsequent analyses. Nevertheless, Gao et al. (2019)[176] demonstrated in their study that, despite slight discrepancies, the data from both platforms were highly concordant. Hence, I assumed it was safe to discard the microarray data.

4.4.2 Common differentially expressed genes across cancers

Pancancer analyses reveal genetic similarities between cancer types. In this work, I focused on identifying genes that are differentially expressed, methylated, mutated, or affected by chromosomal abnormalities for each subtype of a particular cancer type in order to characterise the molecular characteristics that differentiated them. However, by comparing the existence of those differential genes across cancer types, it is possible to gain a pancancer understanding of the common biological pathways that underlie cancer subtype stratification.

Three of the identified shared DEGs coexist in 23 cancer types: *CHGA*, *CPLX2*, and *ITLN1*. The *CHGA* gene product stimulates gastric acid secretion and is involved in the innate immune response[265]. It is considered a prognostic marker for breast, liver, urothelial, and pancreatic cancer, as well as a possible marker for prostate and colon cancer[266–268]. *CPLX2* product is involved in electrical signal transmission between neurons and is a possible predictive biomarker in neuroendocrine lung tumours[269,270]. *ITLN1* product is involved in ion binding, glucose control, and protein phosphorylation regulation[271]. It is linked to Type 2 diabetes, the innate immune system, and IL-9 signalling (responsible for immune cell development and activity)[271]. It has also been identified as DEG in cancers of the gastrointestinal tract, prostate, gynaecological system, breast, bladder, and renal system[272]. The differential expression of these three genes has been linked to cancer classification into subtypes with distinct prognoses, most notably in gastric and lung adenocarcinoma, glioblastoma and colon cancer[273–276]. According to the Human Protein Atlas database, the protein product of these genes is present in the majority of the TCGA cancer types. Therefore, I believe these genes play a critical role in cancer stratification and thus could be a potential therapeutic target that requires further investigation.

Matches between common DEGs and genes affected by chromosomic aberrations were relatively low compared to methylated or mutated genes. The fact that all matches occurred in cancers strongly dominated by copy number changes (breast, ovarian, endometrial, and lung cancer) suggests that the thresholds applied to copy number variations in Automata were too strict and should be reduced for future studies[277–279]. About the DEGs that remained unmatched, they are most likely associated with mechanisms that were not studied in this work and may require sequence analysis, such as microRNAs, transcription factors, and histone deacetylation (a mechanism that overcompresses the DNA in specific regions to prevent the transcription).

KEGG enrichment analysis only showed three enriched biological pathways: PPAR signalling pathway, complement and coagulation cascades, and neuroactive ligand-receptor. PPAR signalling regulates metabolic balance, sugar, lipid, and energy metabolism, as well as insulin sensitivity; complement and coagulation cascades participate in the immune system response; neuroactive ligand-receptors are formed mainly by neuroreceptor genes[280,281]. These three pathways play a role in carcinogenesis and have been linked to a plethora of cancer types, although it is unclear whether they operate as oncogenes, tumour suppressors, or both[281–288]. Different subsets of PPAR genes may have different effects on cancer development, whereas the complement cascade is known to promote tumour growth by inducing chronic inflammation[281,283]. I believe that these pathways function as double agents, promoting the development of certain cancer subtypes while inhibiting others, and hence directly contribute to the increase of ITH. However, further research is needed to understand the specific relevance of these mechanisms across the various subtypes.

On the other hand, GO enrichment analysis only returned an enrichment of genes related to the immune response. This is not surprising, given that immune response is a well-established criterion for classifying tumours into subtypes[289–291].

Regarding driver genes, 13 of the 75 driver DEGs were affected simultaneously by differential methylation, mutation, and chromosomal abnormalities. Bailey et al. (2018)[264] identified eleven of these genes as pancancer drivers (*AXIN2*, *CACNA1A*, *CCND1*, *CDKN2A*, *EGFR*, *EGR3*, *ERBB2*, *FGFR1*, *FOXA1*, *KRT222*, *PGR*), whereas *CD79B* and *CYSLTR2* are associated to lymphoma and uveal melanoma respectively. The fact that these genes

were impacted by all three genetic processes analysed indicates their importance in tumour growth and the surge of subtypes and emphasises their potential to be therapeutic targets.

The Jaccard similarity index revealed a statistically significant similarity between colon and rectum adenocarcinoma. This is consistent with the findings of The Cancer Genome Atlas (2012)[168], which declared that both cancer types "are nearly indistinguishable on a molecular level".

Finally, none of my findings from examining cancer types split by inheritable illness and histology differed from the results found before when analysing all cancer types. One probable rationale is that analysing such characteristics would require knowing which samples belonged to individuals suffering from the disease. Additionally, in terms of histologies, the TCGA lacks enough representation to conduct a meaningful study.

4.5 Conclusions

In this chapter, I have performed a pancancer analysis across the subtypes detected by LPD in 28 cancer types to unravel the biological processes shared across cancer types that contribute to tumour stratification and ITH. I have successfully identified a set of genes and biological pathways that were differentially expressed across several subtypes of different cancer types. Further research will be required to validate these findings and their potential as therapeutic targets that are effective across cancer types.

4.6 Summary

In this chapter, I conducted a comprehensive pancancer analysis to explore the molecular diversity and ITH of tumours across multiple cancer types. The first part of this chapter focused on analysing the gamma values, which represent the presence of subtypes in the tumour. Notably, the analysis revealed a positive correlation between the number of subtypes and sample size, underscoring the significance of larger datasets in accurately capturing molecular variations within these tumours.

The analysis of common differentially expressed genes across multiple cancer types revealed the presence of three genes -*CHGA*, *CPLX2*, and *ITLN1*- in the subtypes of 23 different cancer types, suggesting their potential as biomarkers for tumour stratification. Moreover, the pathway analysis of these differentially expressed genes highlighted three pathways that appear to play crucial roles in the progression or recession of specific subtypes: the PPAR signalling pathway, complement and coagulation cascades, and immune-related pathways. Additionally, identifying driver DEGs affected by differential methylation, mutations, and chromosomal abnormalities underscores their significance in tumour growth and the emergence of subtypes. Overall, this study deepened our understanding of ITH and subtype stratification, setting the stage for future investigations in this field.

Chapter 5

Validation of LPD and the study of breast, prostate, colorectal and lung carcinoma

5.1 Introduction

Large-scale genomics datasets are becoming increasingly prevalent as technology advances and costs fall, for example the TCGA project and the Gene Expression Omnibus[292]. These datasets are used as the raw material for discovering cancer subtypes, which considerably contributes to forwarding cancer therapy from a standard universal approach to personalised treatment practices. Several mathematical techniques have proven helpful in grouping patients into distinct subtypes with different survival patterns: hierarchical clustering, k-means clustering, and self-organising maps. For example, application of hierarchical clustering led to the discovery of five molecular breast cancer types (Basal, Luminal A, Luminal B, HER2-overexpressing and Normal-like, see section 1.5.1). However, in other cancer types these approaches have been less successful. Luca et al. (2018)[243] hypothesised that these sorts of analyses were limited due to their implicit assumption of sample assignment to one particular group or cluster, which is in stark contrast to the well-documented heterogeneous composition of most individual cancer samples. To address this drawback, Luca et al. (2018)[243] proposed using the LPD algorithm, which is more suited to the concept of a single sample containing more than one contributing lineage. Using this method, they successfully detected a poor prognosis subtype of prostate cancer denoted as DESNT and established a framework to predict the outcome of prostate cancer patients (see section 1.5.2)[120,243]. Later, Ellis (2021)[171] used LPD to identify four subtypes of colorectal cancer, one of which was a low-prognosis subtype that was named "Pericol" (see section 1.5.4).

In this chapter, I aim to analyse the results of LPD applied to TCGA in a selection of well-studied cancer types to (i) compare my output with previous subtype discovering approaches, (ii) validate the sample classification by the LPD step integrated into Automata, and (iii) gain a better insight into the cancer biology. I will analyse the four cancers with the highest mortality[293]: breast, prostate, colorectal, and lung cancer (adenocarcinoma and squamous cell carcinoma). For breast cancer, I will compare the LPD output with the five classic subtypes; in the case of prostate and colorectal cancer, I will check whether

Automata's LPD was able to detect DESNT and Pericol; finally, in lung cancer, I will test whether LPD can differentiate between lung adenocarcinoma and lung squamous cell carcinoma. Finally, I will compare the LPD output to the results of hierarchical clustering.

5.2 Methods

The results used in this chapter were gathered and processed from the output of Automata (see chapter 3) in the projects breast carcinoma (TCGA-BRCA), prostate adenocarcinoma (TCGA-PRAD), colon adenocarcinoma (TCGA-COAD), lung adenocarcinoma (TCGA-LUAD), and lung squamous cell carcinoma (TCGA-LUSC). The specifics of the statistical tests, databases and computational resources are described in chapter 2.

5.2.1 Exploring the LPD output

With a Chi-square test, Automata examined the presence of batch effects due to the Tissue Source of the samples, as well as whether the presence of healthy tissue samples was uniformly distributed across groups. Additionally, the mean gamma values of all samples allocated to the same group was calculated. Those that showed a mean gamma value larger than 0.5 were considered a *robust assignment*.

5.2.2 Clinicopathologic characteristics

Boruta[294] was used to select the clinical features analysed by Automata that were important in predicting the assignment of samples into LPD groups for each of the five cancer types. A Chi-squared test was performed to find if there were significant differences in the selected features across the LPD groups. Survival analyses were performed by Automata using Kaplan-Meier curves and log-rank tests to compare the prognosis of each LPD group. In prostate cancer, a Cox analysis was conducted to study associations between PSA values and each LPD group.

5.2.3 Identification of differentially expressed genes (DEGs)

Automata calculated the number of DEGs across each group for each cancer type and represented their differential expression as log_2 *fold change*. An enrichment analysis using the KEGG and GO databases was conducted to gain insights into the biological processes influenced by DEGs. The GO terms obtained were then studied to identify the ones involved in cancer hallmarks. The complete list of GO terms related to cancer hallmarks was obtained from Chen et al. (2021)[295]. Genes with positive fold change values were classified as *overexpressed* or *upregulated*, while those with negative values were labelled as *underexpressed* or *downregulated*. The ratio of overexpressed to underexpressed genes was calculated. Additionally, Cancer driver genes from Bailey et al. (2018)[264] (n = 299) were cross-referenced with the identified DEGs to explore potential associations between DEGs and known cancer driver genes.

5.2.4 Identification of differentially methylated genes, genes affected by mutations and genes affected by copy number changes

The workflow described in the previous section was repeated for genes that were differentially methylated, genes that were affected by single nucleotide variants, and genes that

were affected by copy number changes. Genes with positive values were labelled as *hypermethylated*, *overmutated* and *overimpacted* respectively, while their analogues for negative values were labelled as *hypomethylated*, *undermutated* and *underimpacted*.

In addition, in the case of the genes affected by single nucleotide variants, Automata classified the SNVs into Single-nucleotide polymorphisms (SNPs), insertions or deletions. The pipeline categorised each SNV according to its effect (frameshift, missense, stop...) based on the affected nucleotide. Furthermore, Automata also performed the identification of mutational signatures denoted by COSMIC for each LPD group. In the analysis of prostate cancer, the genes *SPOP*, *FOXA1*, and *IDH1* were assessed to determine if they were more frequently mutated (overmutated) or less frequently mutated (undermutated) in any of the LPD groups when compared to each other. Similarly, in lung adenocarcinoma, the genes *EGFR*, *NF1*, *TP53*, and *KRAS* were also analyzed for over and undermutation in the LPD groups.

In the analysis of copy number variations in the lung adenocarcinoma dataset, the *STK11* gene was specifically examined to verify and validate the findings in relation to previous results reported in the literature.

Automata also analysed the presence of genes that were co-occurring as DEG, DMG, affected by SNVs, and impacted by CNVs. This was genes that had additional evidence for functional importance. The DEGs were split into overexpressed and underexpressed. Only co-occurances with hypomethylated or amplified genes were judged relevant for overexpressed genes, whereas co-occurances with hypermethylated, deleted, and mutated genes were considered relevant for underexpressed genes. A Chi-squared test was used to compare the frequency of co-occurrences across LPD groups.

5.2.5 Comparison of the LPD output with Euclidian hierarchical clustering

The Automata pipeline performed hierarchical clustering of the samples based on Euclidean distance and complete linkage to compare its output with the LPD approach. For each cancer type, a dendrogram was generated to visualise and compare both assignments.

5.2.6 The PAM50 classification of the BRCA samples

The assignment of TCGA BRCA samples to the five PAM50 subtypes was obtained from Netanely et al. (2016)[296]. A Chi-squared test was used to compare the presence of progesterone, estrogen, and HER2 receptors across the five groups. Kaplan-Meier estimators along with a log-rank test were used to compare the prognosis of the groups. Hierarchical clustering of the samples based on Euclidean distance (complete linkage) was performed to compare LPD assignment with the PAM50 classification. The classification of the samples according to PAM50 was compared to the LPD output through an alluvial plot to study the overlaps between the two approaches.

5.2.7 Comparison of the LPD output with DESNT

The assignment of the TCGA-PRAD samples into the DESNT classification and the set of 45 genes associated with DESNT were obtained from Luca et al. (2018)[118]. An alluvial plot was performed to compare the assignment of Luca et al. with the LPD step integrated

in Automata. The set of 45 genes were compared to the DEGs, DMGs, genes differentially mutated by SNVs, and genes differentially impacted by CNVs identified by Automata in TCGA-PRAD.

5.2.8 Comparison of the LPD output with Pericol

The assignment of the TCGA-COAD samples into the Pericol classification was obtained from Ellis (2021)[171]. An alluvial plot was performed to compare the assignment of Ellis with the LPD step integrated in Automata. A Pearson correlation analysis between each of Ellis' identified subtypes and the LPD groups identified by Automata was performed.

5.2.9 LPD and Euclidean hierarchical clustering applied to the combined lung carcinoma dataset

The LUAD and LUSC dataset from TCGA were combined and LPD was applied on it following the same workflow described in section 3.2. An alluvial plot was used to visually compare the assignment into groups of the combined dataset by LPD with the cancer type they belong to. Similarly a Euclidean hierarchical clustering (complete linkage) was performed to compare classifications.

5.2.10 Comparison of the LPD output with previous subtyping frameworks in lung squamous cell carcinoma

The genes *KEAP1*, *NFE2L2*, *PTEN*, *RB1*, and *NF1* were studied regarding expression and methylation profile, SNV mutations, and CNV alterations. The primary objective of this comparative study was to draw parallels and contrasts with the study by the TCGA (2012)[153].

5.3 Results

5.3.1 Breast cancer

Exploring the LPD output for BRCA

Automata was used to analyse a total of 1211 samples from 1095 different patients. Eight LPD groups were found optimal, which were called named LPD_1 (n = 140, 11.56%), LPD_2 (n = 139, 11.47%), LPD_3 (n = 110, 9.08%), LPD_4 (n = 267, 22.04%), LPD_5 (n = 160, 13.21%), LPD_6 (n = 209, 17.25%), LPD_7 (n = 67, 5.53%), and LPD_8 (n = 119, 9.82%) (Fig 5.1). No tissue source site (TSS) was significantly associated with the sample distribution into LPD groups (P = 0.18; Chi-squared test). There was significant overrepresentation of healthy tissue samples in LPD_8 ($n_{healthy}$ = 98, 82.35%, $P = 1.49x10^{-178}$; Chi-squared test). When the mean gamma values for each group were calculated, LPD_6 and LPD_8 exhibited a robust assignment (Fig 5.2).

Clinicopathologic characteristics of the clusters in BRCA

Table 5.1 shows the clinicopathologic characteristics of the tumour samples. Patients in the LPD_2, LPD_4, and LPD_7 groups were older ($P = 1.11x10^{-09}$; Chi-squared test) than those in the other groups. LPD_6 had a larger proportion of patients of black or African American ethnicity than the other groups (P = 0.0004; Chi-squared test). Pathological

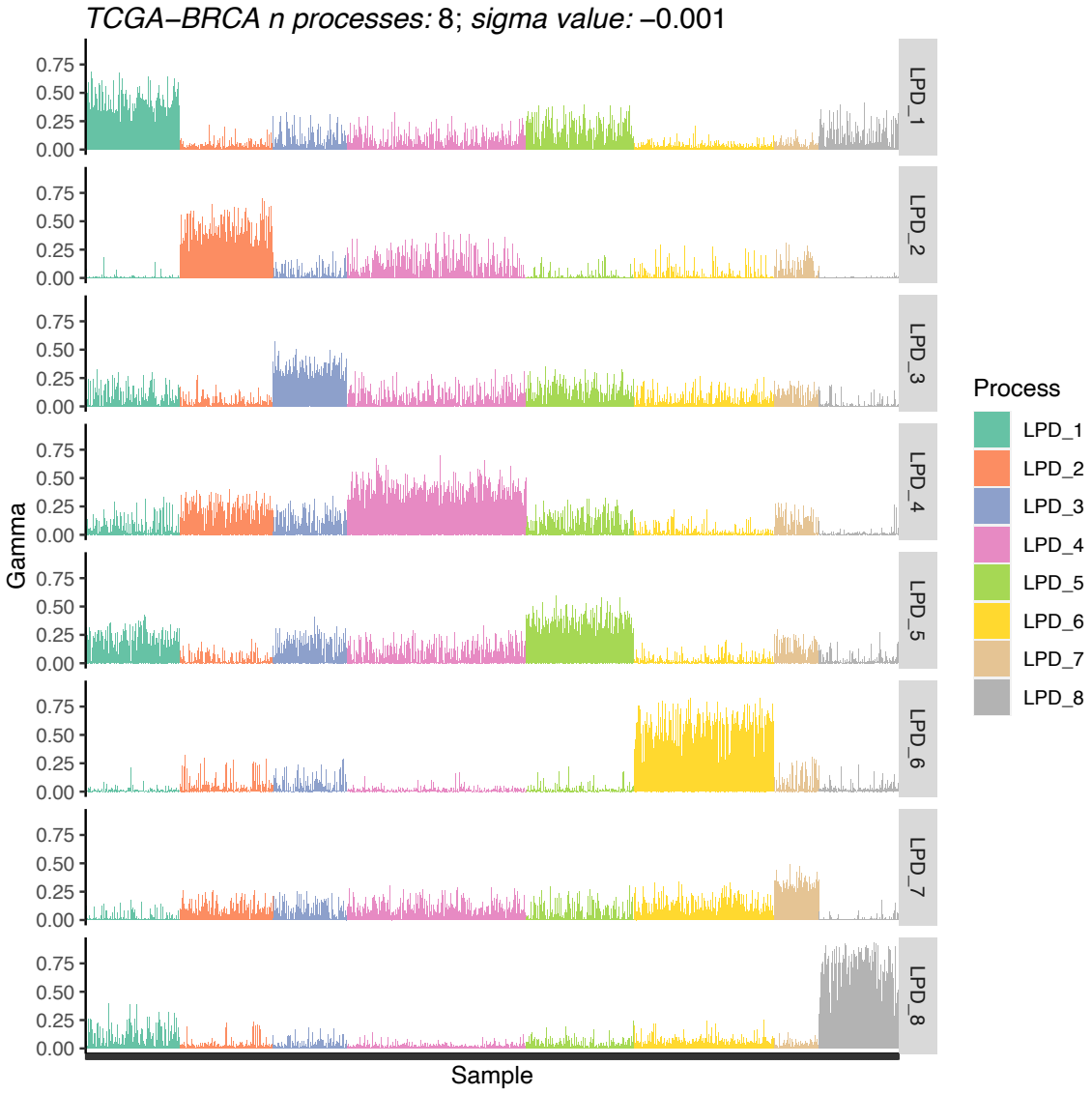

Figure 5.1: Gamma values of all samples for each detected LPD process in breast carcinoma. A total of 8 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_8. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD process is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

Figure 5.2: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in breast carcinoma. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

stages I and III were distributed similarly across all LPD groups; however, stage II was enriched in LPD_6 ($P = 0.0004$; Chi-squared test). LPD_6 had a larger percentage that were negative for estrogen ($P = 0.001$; Chi-squared test) and progesterone receptor ($P = 0.001$; Chi-squared test). Overall the groups had a significant association with event-free survival ($P = 0.04$; Log-rank test) was identified, with LPD_1 ($P = 0.019$; Log-rank test) and LPD_7 ($P = 0.015$; Log-rank test) showing the most significant association with prognosis (Fig 5.3).

Table 5.1: Clinicopathologic features of the detected subtypes for breast carcinoma. chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	P-value
Age (years; mean (sd))	58.9 (13.3)	58.2 (12.1)	64.9 (13.2)	58.7(13.2)	62.3 (13.9)	57.02 (12.5)	56.6 (12.1)	62.8 (14.8)	< 0.0001
Race									
Asian	61	4	9	7	10	11	10	10	
Black or african american	179	9	23	20	31	19	66	11	
White	755	121	88	73	193	119	122	39	
American indian or alaska native	1	0	0	0	0	0	1	0	0.0004
Pathological Stage									
Stage I	180	30	12	16	53	34	27	8	
Stage II	610	68	76	53	154	84	137	38	
Stage III	256	38	34	38	52	39	35	20	0.0009
Estrogen receptor									
Positive	788	134	114	96	241	143	22	38	
Negative	232	3	3	9	5	12	172	28	0.001
Progesterone receptor									
Positive	684	124	89	75	222	137	12	25	
Negative	333	11	28	31	23	19	180	41	0.001

Identification of differentially expressed genes in BRCA

The differential analysis of the gene expression for the eight LPD groups in breast cancer revealed 11,626 significant differentially expressed genes (Table 5.2; median across groups = 1097; IQR = 521). The top overexpressed and underexpressed genes, ranked by *log2 fold change*, are presented in Table 5.3. LPD_2 was the only group that exhibited more overexpressed genes than underexpressed genes. Among the DEGs, 13 were identified as cancer driver genes, including *ALB*, *APOB*, *CD79B*, *EGR3*, *EPHA3*, *ERBB2*, *FOXA2*, *KIF1A*, *KIT*, *MUC6*, *WT1* and *ZBTB20* (Fig. 5.4). Further analysis revealed the enrichment of 58 biological processes associated with these DEGs (Figure 5.5.A for KEGG and Figure 5.6.A for GO). Notably, all LPD groups in KEGG showed downregulation of neuroactive ligand-receptor interaction, and four LPD groups exhibited associations with the PPAR signalling pathway, consistent with the pancancer analysis results discussed in chapter 4. LPD_4 and LPD_6 displayed similar profiles with downregulation of systemic lupus erythematosus and upregulation of metabolism of xenobiotics, drug metabolism, and chemical carcinogenesis. In GO, the most common altered biological process was epidermis development, which was upregulated in LPD_4 and downregulated in LPD_2, LPD_3, LPD_5, LPD_7, and LPD_8. Two distinct patterns were observed in the analysis checking for associations to cancer hallmarks (Figure 5.7). LPD_2 displayed enrichment in biological processes related to unlimited replication, indicating a potential association with enhanced cell proliferation and growth. On the other hand, LPD_3 and LPD_8 showed an association between underexpressed genes and tumour inflammation caused by tumoural cells in healthy cells.

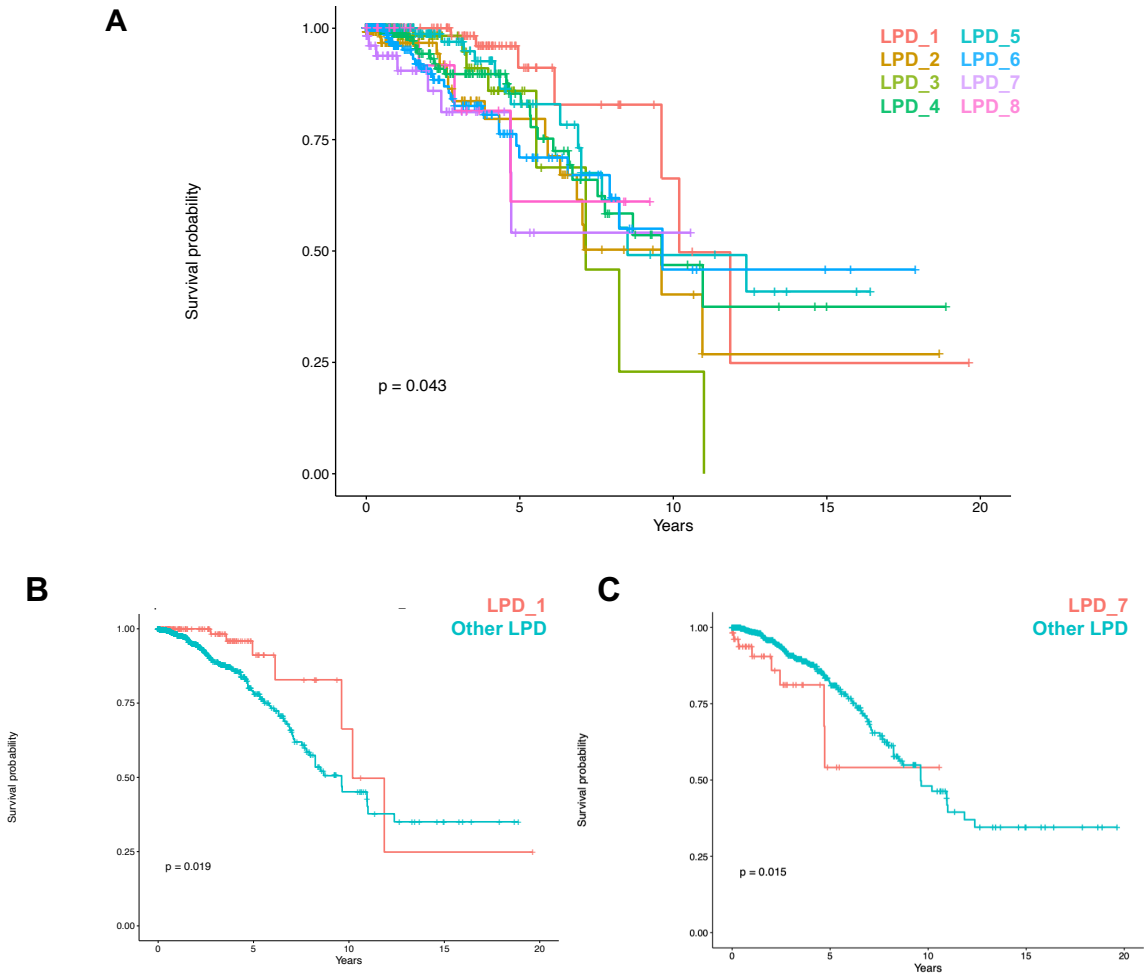

Figure 5.3: (A) Kaplan-Meier curves for all the LPD groups in breast carcinoma showing the survival probability over time of the patients allocated to each group. Log-rank test was conducted across the survival curves and the corresponding p-value is displayed. (B) Kaplan-Meier curve for LPD_1 (red) in comparison to the other LPD groups (blue). (C) Kaplan-Meier curve for LPD_7 (red) in comparison to the other LPD groups (blue).

Table 5.2: Gene counts for various categories in BRCA. These include the number of genes exhibiting significant differential expression and differential methylation, the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively), and the number of genes with significantly higher or lower frequency of CNV (referred to as overimpacted and underimpacted, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	LPD_8
n health tissue samples	1	10	0	0	0	4	0	98
DEGs								
Upregulated	235	2426	100	69	132	22	55	741
Downregulated	508	962	1033	787	1118	1039	946	1453
Total	743	3388	1133	856	1250	1061	1001	2194
Ratio	0.46	2.52	0.1	0.09	0.12	0.02	0.06	0.51
DMGs								
Hypermethylated	0	9	2	9	0	175	81	613
Hypomethylated	1	22	0	50	0	339	45	349
Total	1	31	2	59	0	514	126	962
Ratio	0	0.41	1	0.18	1	0.52	1.8	1.76
Mutated								
Overmutated	71	229	179	683	247	892	440	0
Undermutated	264	349	315	342	308	306	404	21
Total	335	578	494	1025	555	1198	844	21
Ratio	0.27	0.66	0.6	2	0.8	2.92	1.1	0
Affected by CNV								
Overimpacted	79	748	393	586	336	389	990	0
Underimpacted	92	43	18	76	45	85	2	8
Total	171	791	411	662	381	474	992	8
Ratio	0.86	17.4	21.83	7.71	7.47	4.58	495	0

Table 5.3: The five genes more overexpressed (log2FoldChange > 1) and underexpressed (log2FoldChange < -1) for each LPD group in breast carcinoma. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status
LPD_1		
SNORA74B	4.152843	Overexpressed
RNU4-2	4.071574	Overexpressed
RNU4-1	3.854122	Overexpressed
SNORA74A	3.707482	Overexpressed
RNY3	3.611527	Overexpressed
CSN2	-5.363128	Underexpressed
LALBA	-4.834109	Underexpressed
DCAF4L2	-4.274501	Underexpressed
CSN1S1	-4.040578	Underexpressed
MYL1	-4.030537	Underexpressed
LPD_2		
RN7SL3	4.484153	Overexpressed
RN7SKP227	4.355101	Overexpressed
RNU1-88P	3.842152	Overexpressed
ENSG00000253456	3.778723	Overexpressed
ENSG00000259001	3.764864	Overexpressed
CSN2	-8.144828	Underexpressed
LALBA	-5.369773	Underexpressed
CSN3	-5.195181	Underexpressed
ENSG00000231683	-4.391698	Underexpressed
SULT1C3	-4.216631	Underexpressed
LPD_3		
SMR3B	4.103980	Overexpressed
ENSG00000225840	2.294302	Overexpressed
ENSG00000224467	2.287689	Overexpressed
PRR27	2.263271	Overexpressed
DCAF8L1	2.217291	Overexpressed
CSN2	-5.356047	Underexpressed
CHGA	-4.774411	Underexpressed
SULT1C3	-4.409733	Underexpressed
MTND1P23	-4.153919	Underexpressed
SLCO1B3-SLCO1B7	-3.696974	Underexpressed
LPD_4		
ADH7	4.242276	Overexpressed
ENSG00000231683	4.152039	Overexpressed
ENSG00000261409	3.475462	Overexpressed
SFTPB	3.403842	Overexpressed
SFTPA1	2.998193	Overexpressed
MUC2	-4.869433	Underexpressed

CHGA	-4.415999	Underexpressed
CARTPT	-4.305338	Underexpressed
ACTBP12	-3.808859	Underexpressed
RNU1-11P	-3.576751	Underexpressed
LPD_5		
ENSG00000237527	3.598906	Overexpressed
FGA	3.120956	Overexpressed
APOC3	2.870472	Overexpressed
LINC00261	2.507229	Overexpressed
APOA2	2.388342	Overexpressed
PRR27	-4.853331	Underexpressed
MYL1	-4.638612	Underexpressed
CSN2	-4.235598	Underexpressed
FTHL17	-4.146464	Underexpressed
DCAF4L2	-4.137787	Underexpressed
LPD_6		
CHGB	2.941945	Overexpressed
MYOC	2.604452	Overexpressed
CYP2A6	2.362853	Overexpressed
NOBOX	2.248615	Overexpressed
LACRT	1.627733	Overexpressed
SMR3B	-4.627655	Underexpressed
MYL1	-4.374082	Underexpressed
MTND1P23	-4.264854	Underexpressed
KRT13	-3.823832	Underexpressed
SCARNA5	-3.708685	Underexpressed
LPD_7		
GKN2	1.842305	Overexpressed
TUBA3D	1.798026	Overexpressed
ANKRD18B	1.758612	Overexpressed
LINC01224	1.695252	Overexpressed
ENSG00000223023	1.641827	Overexpressed
CSN2	-8.005280	Underexpressed
LALBA	-7.904983	Underexpressed
SULT1C3	-6.081887	Underexpressed
LACRT	-5.748902	Underexpressed
CARTPT	-5.179050	Underexpressed
LPD_8		
MYPN	2.903648	Overexpressed
PRAMENP	2.762662	Overexpressed
GCG	2.755927	Overexpressed
IGLL1	2.705269	Overexpressed
RPS26P34	2.655842	Overexpressed
CARTPT	-7.229038	Underexpressed
LALBA	-7.050424	Underexpressed
CSN2	-6.772605	Underexpressed
KLHL1	-5.683357	Underexpressed

CHGA	-5.002839	Underexpressed

Identification of differentially methylated genes in BRCA

The eight LPD groups differed greatly in terms of the number of DMG (median across groups = 45; IQR = 221), with LPD_6 and LPD_8 accumulating 80% of the DMGs ($n = 1476$; Table 5.2). Only LPD_7 and LPD_8 showed more hyper than hypomethylation. Five driver genes were effected: *PMS2*, *APOB*, *FOXA2*, *SOX17* and *SPTA1* (Fig. 5.4). KEGG returned biological processes that were significantly enriched only for LPD_8 (Fig. 5.5.B), whereas GO detected processes in LPD_6 as well as LPD_8 (Fig. 5.6.B).

Identification of genes affected by single nucleotide variants in BRCA

A median of 566 genes per LPD group were enriched or depleted in single nucleotide variations (SNVs) in comparison to other groups (IQR = 435). LPD_8 exhibited a relatively low number of genes affected by mutations (less than 1% of the total), which was most likely due to the high proportion of healthy tissue samples in the group (82% of the samples) (Table 5.2). LPD_4, LPD_6 and LPD_8 were the only groups defined by a higher occurrence of mutations. A total of 146 of the genes differentially affected by SNVs were identified as driver genes (Fig. 5.4). KEGG enrichment analysis returned several processes, all of which were related to genes relatively depleted of SNVs (Fig. 5.5.C). Most LPD groups had two associated processes: the neurodegeneration pathway, which was linked to LPD_1, LPD_3, LPD_5, LPD_6 and LPD_7; and the human papillomavirus infection, which was linked to LPD_1, LPD_3, LPD_4, LPD_5, LPD_6 and LPD_7. The GO enrichment analysis, on the other hand, returned only one term, gland development, that was linked to under-mutated genes in LPD_8 (Fig. 5.5.C). No differences were observed when comparing the SNP type, variant type, and variant class frequency across LPD groups ($P > 0.05$; Chi-squared test; Fig. 5.8). The bulk of total detected SNVs were missense mutations caused by the point replacement of cytosine to thymine.

The proportion of COSMIC mutational signatures associated with SNVs in samples in each group can be seen in Fig. 5.9. LPD_1, LPD_3, LPD_5 and LPD_7 all had a similar pattern defined by a uniform contribution from signatures 1, 2, and 13, although the contribution of signature 1 for LPD_7 was weaker than for the rest. LPD_2 and LPD_8 showed a strong proportion in a single signature: signature five and signature two, respectively. LPD_4 was strongly related to signature one and, to a lesser extent, with signature ten. The pattern of LPD_8 mutational signatures was unclear.

Identification of genes impacted by copy number variations in BRCA

A median of 442 genes per LPD group were enriched or depleted in copy number variations (CNVs) compared to other LPD groups (IQR = 365). Similarly to the SNVs, the number of genes impacted by CNV for LPD_8 was relatively low (Table 5.2). Except for LPD_1, the ratio of overimpacted to underimpacted was relatively high in favour of an increase in copy number alterations. Eighteen genes affected by copy number variations were driver genes (Fig. 5.4). KEGG and GO enrichment analyses did not yield any significantly enriched biological pathways.

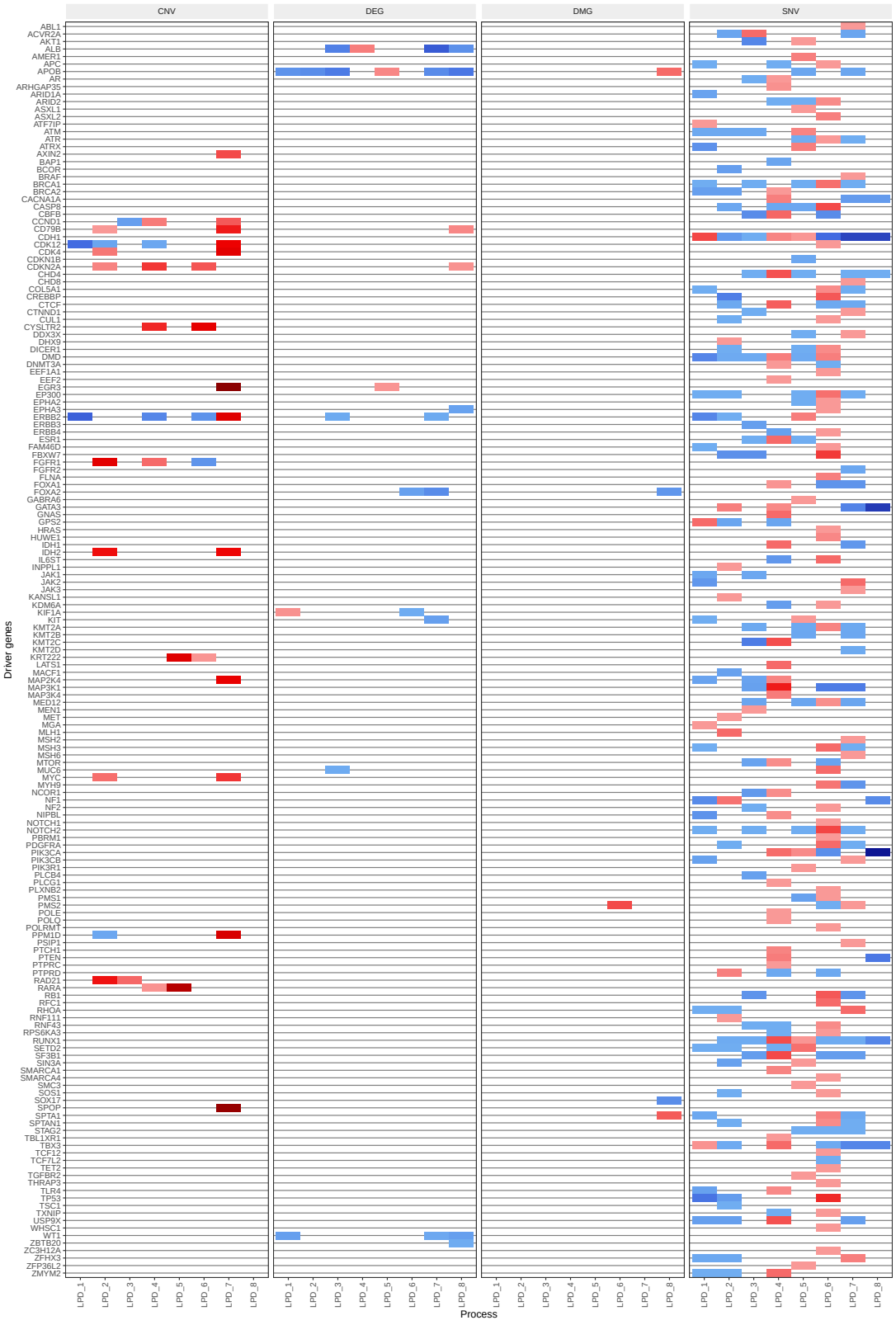

Figure 5.4: Heatmap showing the presence of driver genes across different categories in breast carcinoma, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV) and copy number variants (CNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV and CNV are represented in red, while genes with opposite characteristics are depicted in blue.

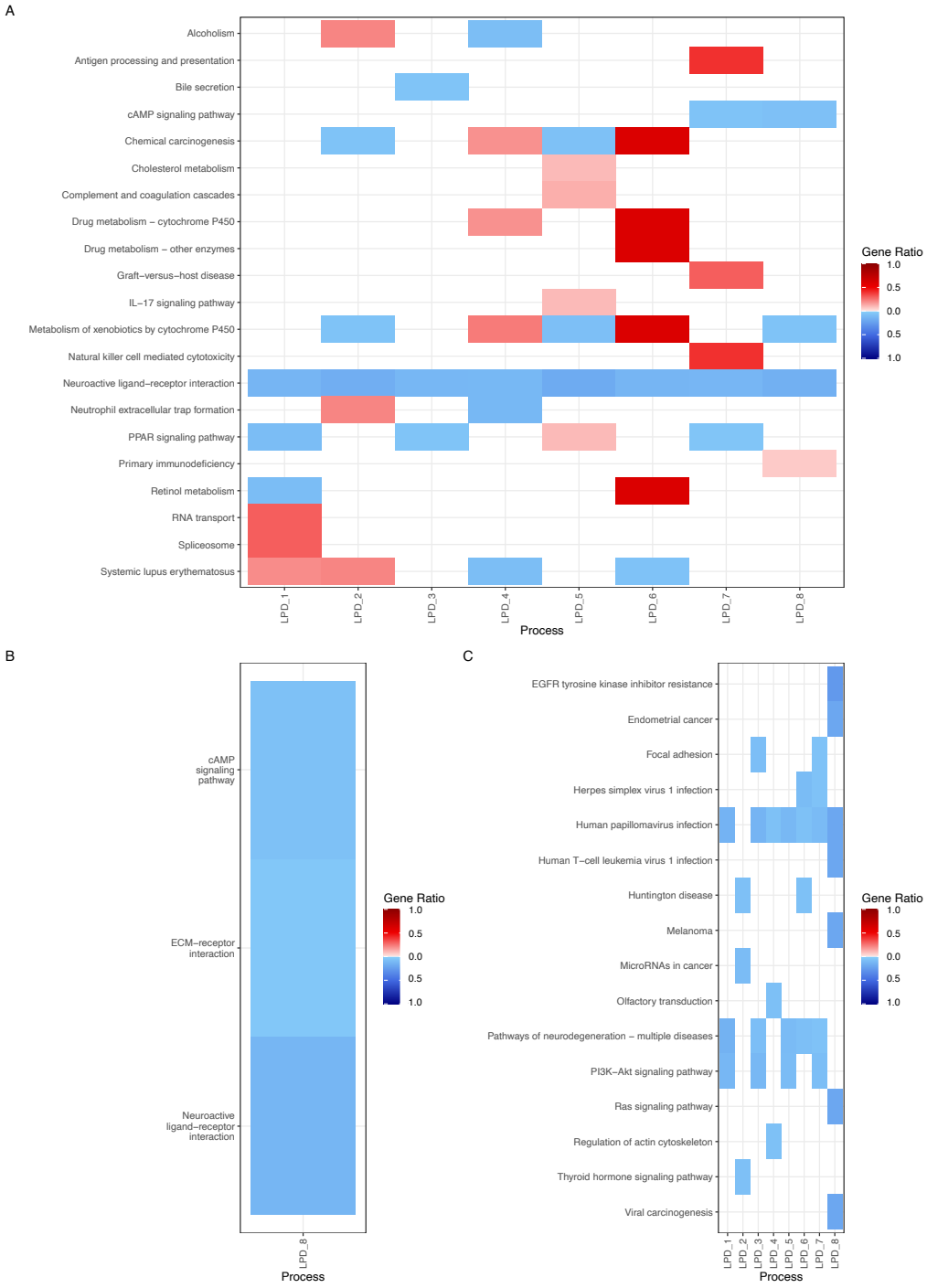

Figure 5.5: Biological pathways associated with different categories in BRCA determined using KEGG enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, biological pathways associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the pathways associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological pathways is available in Supplementary Material B.

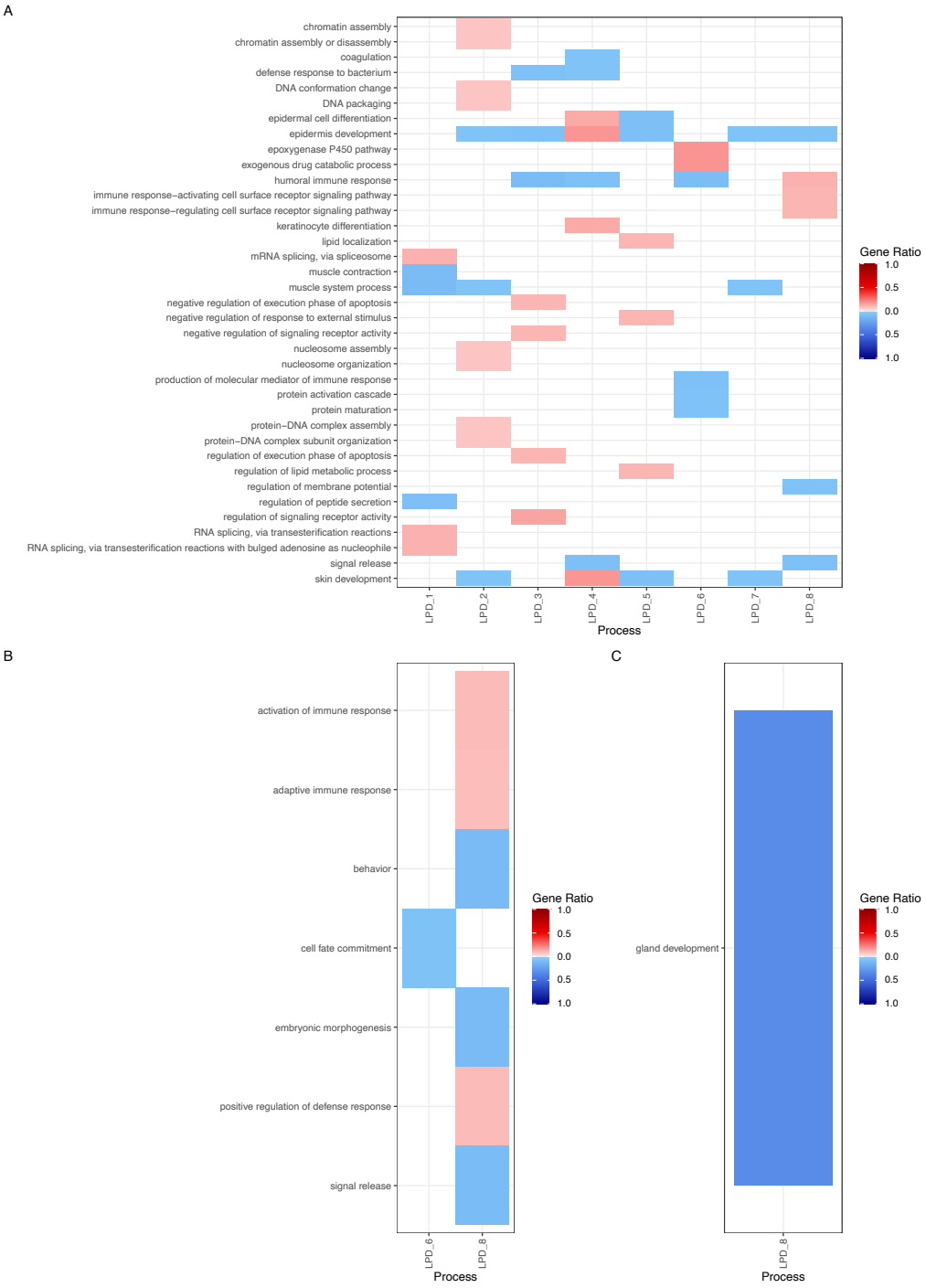

Figure 5.6: Biological processes associated with different categories in BRCA determined using GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant processes with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while processes linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the processes associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological processes is available in Supplementary Material B.

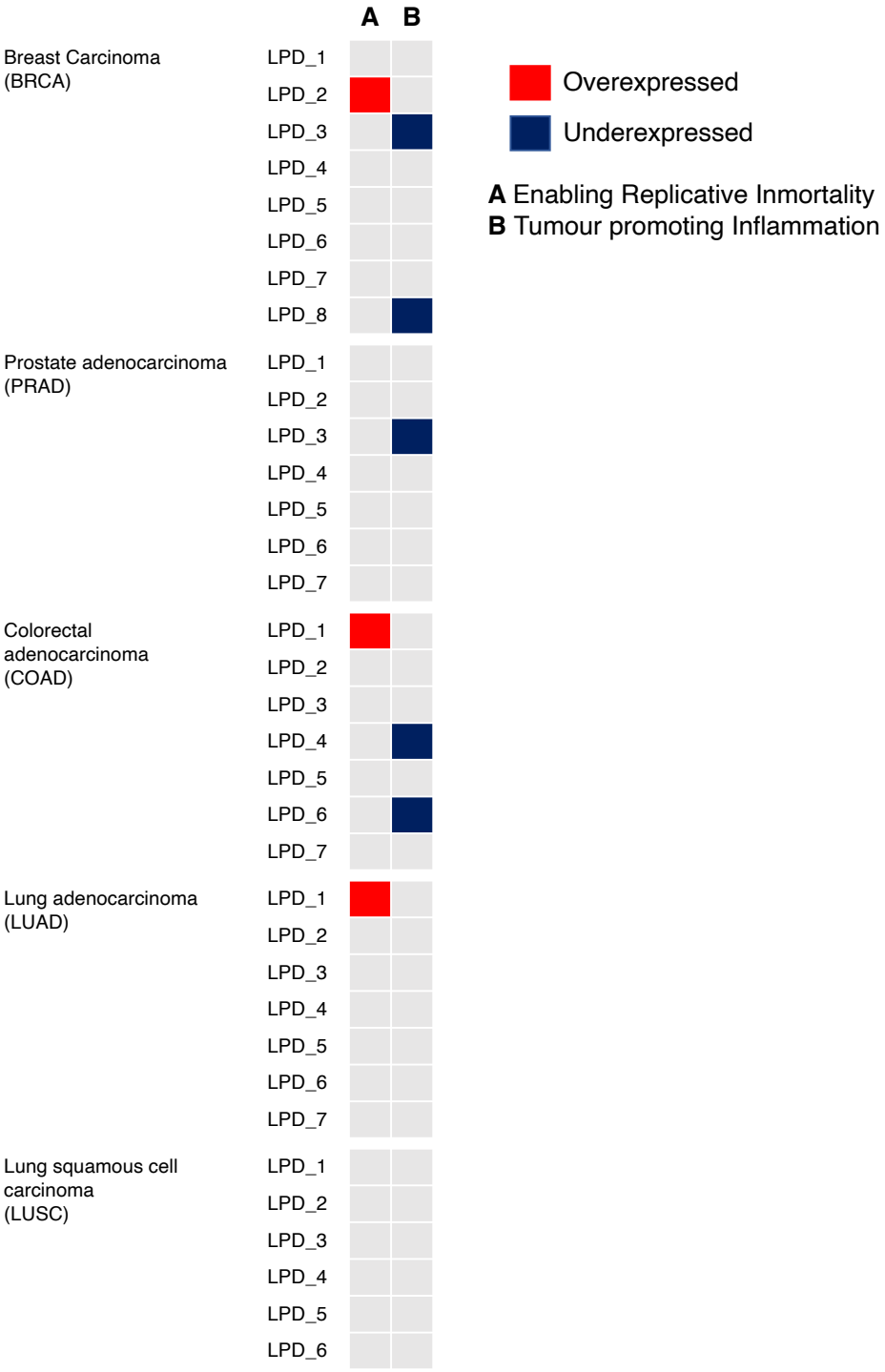

Figure 5.7: Hallmarks of cancer associated to the DEGs of each LPD group detected in BRCA, PRAD, COAD, LUAD, and LUSC. Two hallmarks were found differentially associated to the LPD groups across the five cancer types: enabling replicative inmortality (A), and tumour promoting inflammation (B). The hallmarks associated with overexpressed genes are depicted in red, while those associated with underexpressed genes are depicted in blue.

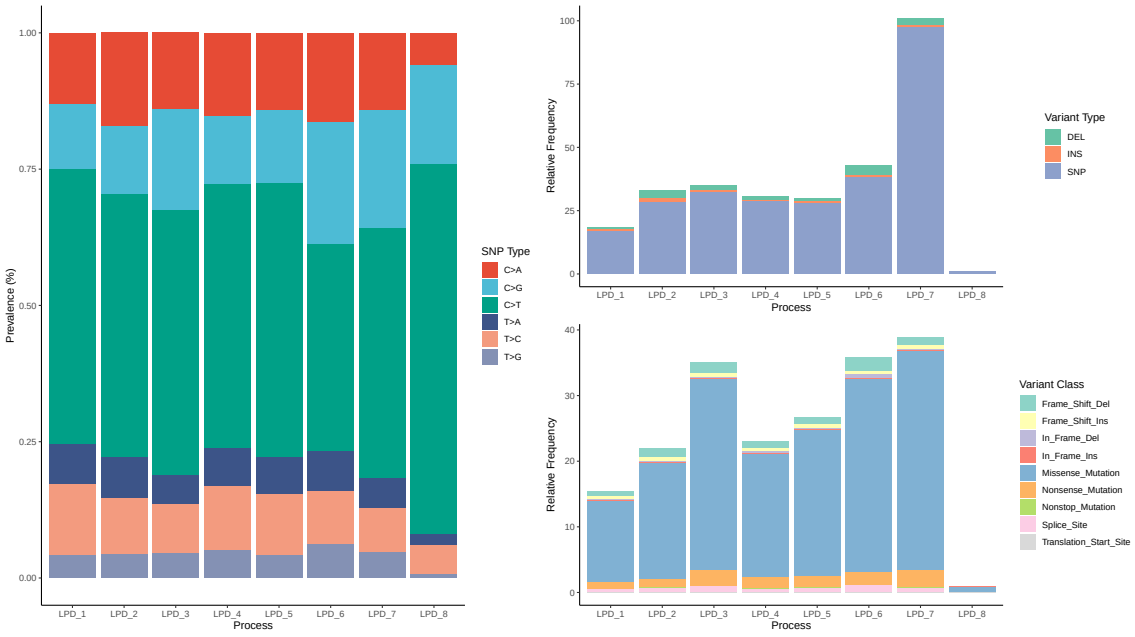

Figure 5.8: Detected single nucleotide variants (SNVs) within each LPD group for BRCA. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

Additional evidence of a functional effect of differentially expressed genes in BRCA

Identifying DEGs provides an overview of the biological pathways and molecular processes altered in each group, but it does not explain the factors that affected the gene expression in the first place. Matching DEGs to DMGs, genes influenced by SNVs, and genes impacted by CNVs can reveal some of the mechanisms underlying in the differential expression and indicate the enrichment of LPD groups in any of the three alterations.

Matches between overexpressed, hypomethylated and amplified genes are shown in figure 5.10, while matches between underexpressed, hypermethylated, deleted and mutated genes are shown in figure 5.11. No significant overlaps were observed in the overexpressed DEGs ($P = 0.47$; Chi-squared test). In the underexpressed genes, LPD_6 had a significant overlap with 7 hypermethylated, 5 deleted and 2 mutated genes ($P = 0.0008$; Chi-squared test). The complete list of matched genes is available in Supplementary Material B.

Comparison of the LPD output in BRCA with Euclidean hierarchical clustering

The distribution of the LPD groups according to Euclidean hierarchical clustering displayed a well-defined separation of LPD_6 and LPD_8 from the other groups (Fig. 5.12). Samples belonging to LPD_1 and LPD_3 seemed to be frequently clustered together, as well as LPD_2 and LPD_4. In general, the hierarchical clusters appeared to have no associations with the LPD groupings.

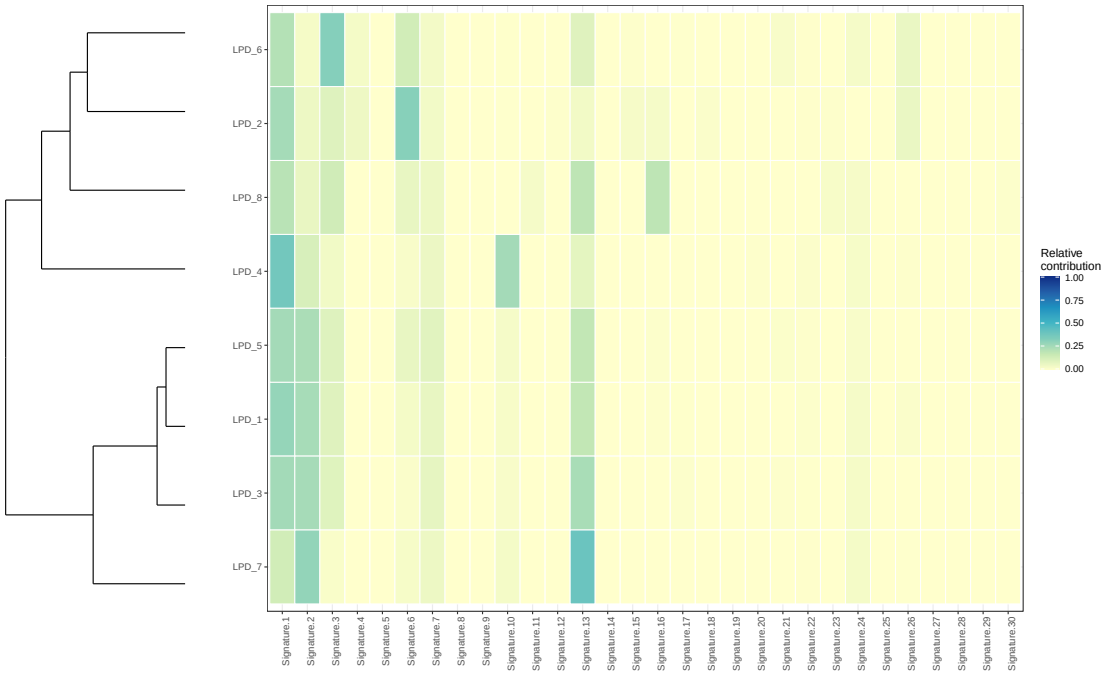

Figure 5.9: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in BRCA. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

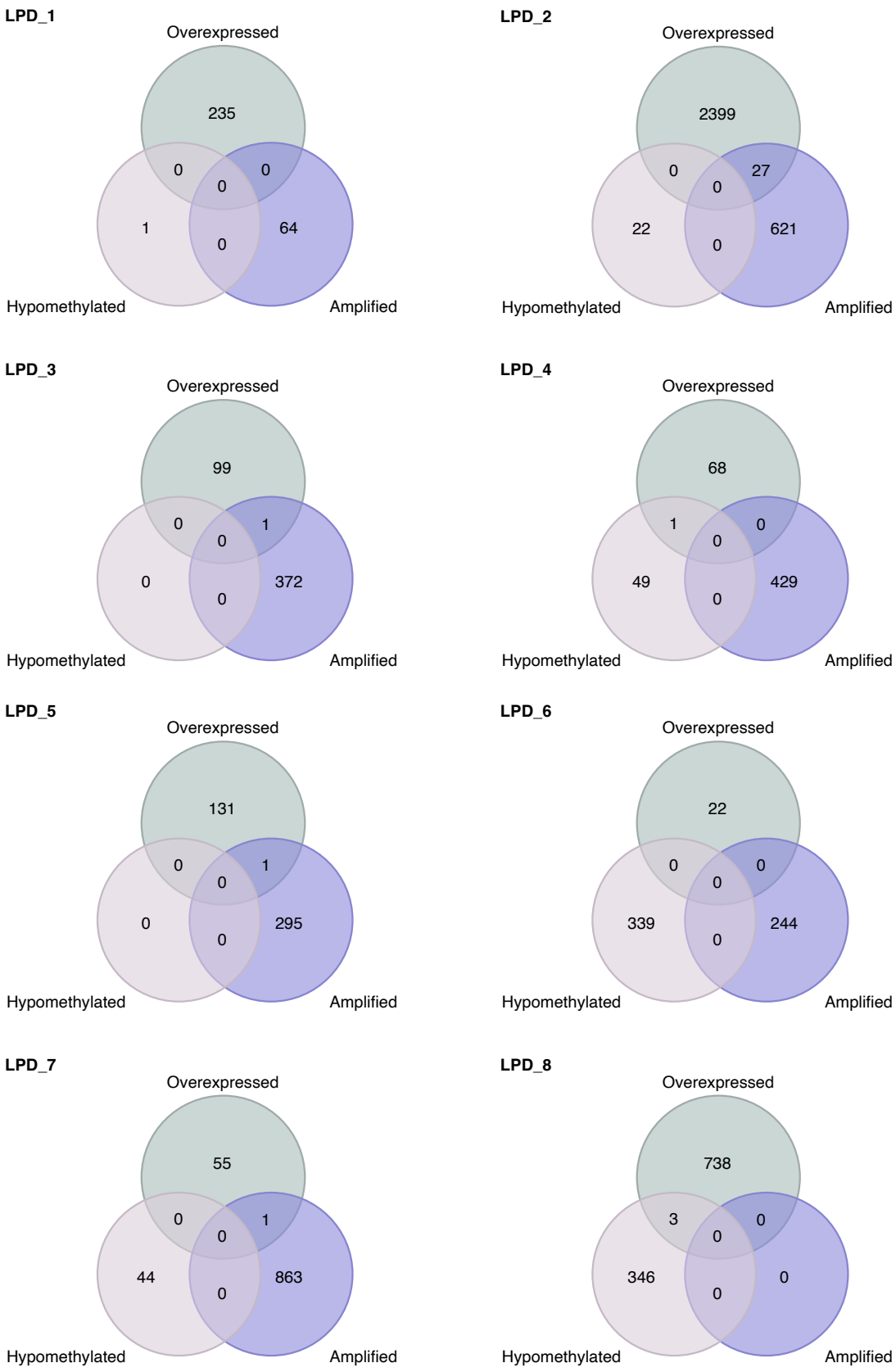

Figure 5.10: Venn diagram displaying the overlaps between three categories in genes in BRCA for each LPD group: overexpressed genes, hypomethylated genes, and amplified genes. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit overexpression, hypomethylation, and amplification simultaneously.

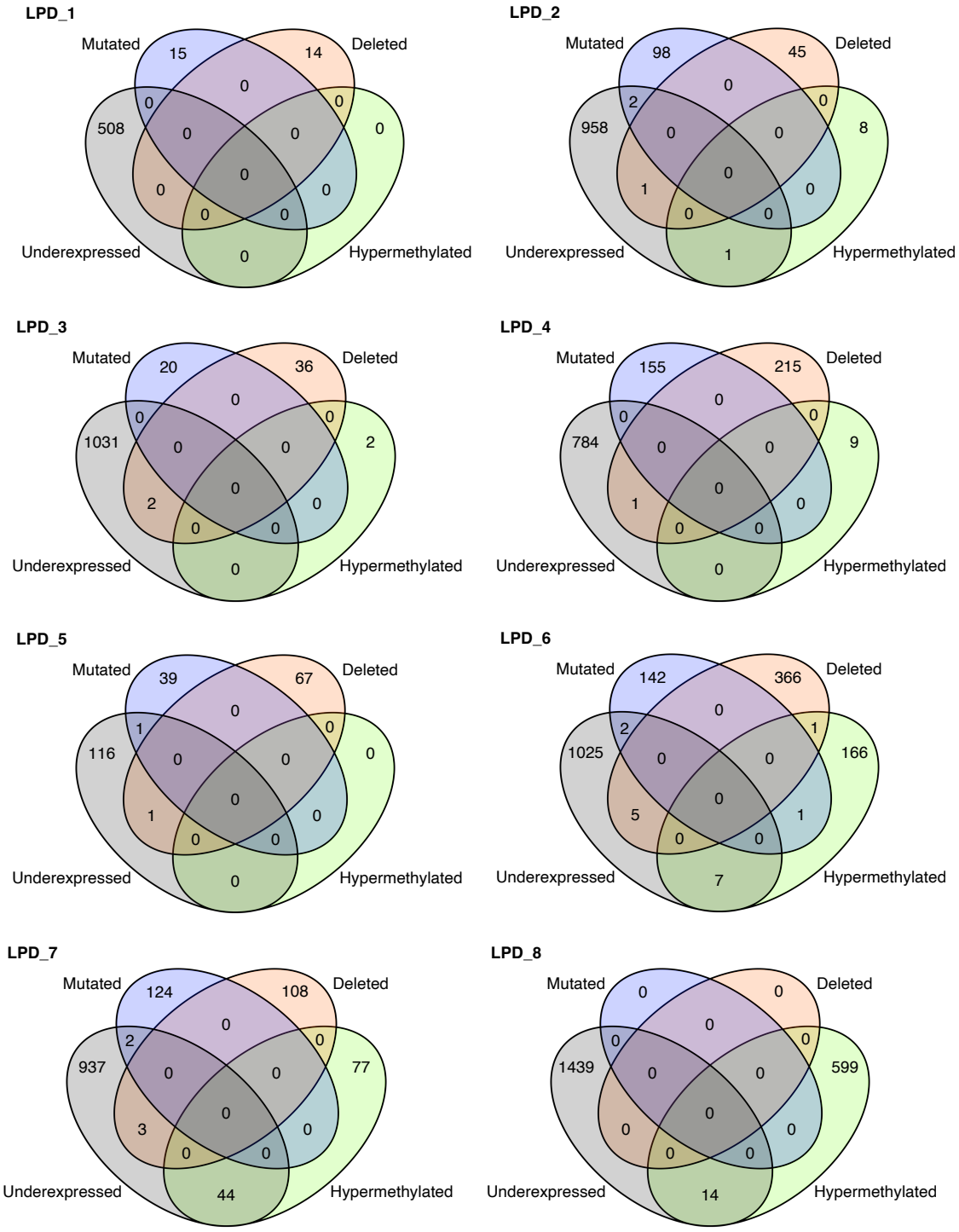

Figure 5.11: Venn diagram displaying the overlaps between four categories in genes in BRCA for each LPD group: underexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit underexpression, hypermethylation, mutations and deletions.

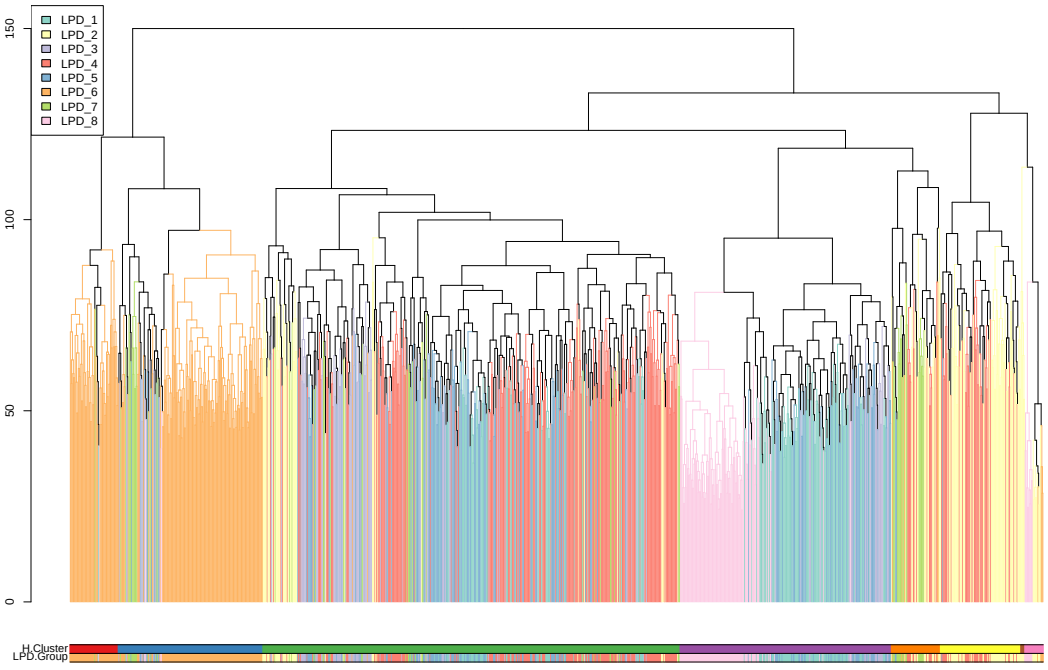

Figure 5.12: Dendrogram showing the sorting of the BRCA samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into eight groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

The PAM50 classification of the BRCA samples

The presence of estrogen, progesterone and HER2 receptors was not uniformly distributed across the TCGA samples classified according to their PAM50 subtype (Table 5.4; $P < 0.0001$; Chi-squared test). The three receptors were significantly absent in the Basal samples, whereas estrogen and progesterone were enriched in the Luminal A and B samples. The HER2 samples were the only ones with enrichment of HER2 receptors, but they lacked estrogen and progesterone receptors. The Kaplan-Meier to compare the survival probability of the five groups showed a significant difference in the curves of the five groups (Figure 5.13; $P < 0.001$; Log-rank test). Basal, together with Luminal A, had the best survival probability, whereas Normal and Luminal B exhibited the worse probability. The Euclidian hierarchical clustering of the PAM50 depicted a good separation for the Basal, the Normal and the HER2 samples., whereas Luminal A and B were slightly mixed (Figure 5.14).

Table 5.4: The count of samples expressing estrogen, progesterone, and HER2 receptors for each of the PAM50 groups. Chi-squared tests were conducted to assess any disparities in the proportion of receptor expression across the groups and P-values are provided. The total count and ratio for each receptor category within the PAM50 groups are also included.

	Total	Basal	Her2	Luminal A	Luminal B	Normal	P-value
n	1140	181	78	531	201	149	
Estrogen							
Yes	833	20	27	499	186	101	
No	247	155	45	11	3	33	
Total	1080	175	72	510	189	134	
Ratio	3.37	0.13	0.6	45.36	62	3.06	< 0.0001
Progesterone							
Yes	722	11	14	457	152	88	
No	355	162	60	50	37	46	
Total	1077	173	74	507	189	134	
Ratio	2.03	0.07	0.23	9.14	4.11	1.91	< 0.0001
HER2							
Yes	108	3	47	26	30	2	
No	643	127	16	345	137	18	
Total	751	130	63	371	167	20	
Ratio	0.17	0.02	2.94	0.08	0.22	0.11	< 0.0001

Comparison of the BRCA LPD output with the PAM50 classification

I compared the PAM50 classification of the five established molecular subtypes with the LPD groups for the BRCA samples revealed distinct patterns for each LPD group (Fig. 5.15). LPD_1 was largely conformed of Luminal A samples (90.29% of LPD_1 samples). LPD_2 and LPD_4 resulted in a varied percentage of Luminal A and Luminal B; LPD_2 possessed a more balanced distribution, with 45% of the samples allocated to Luminal A

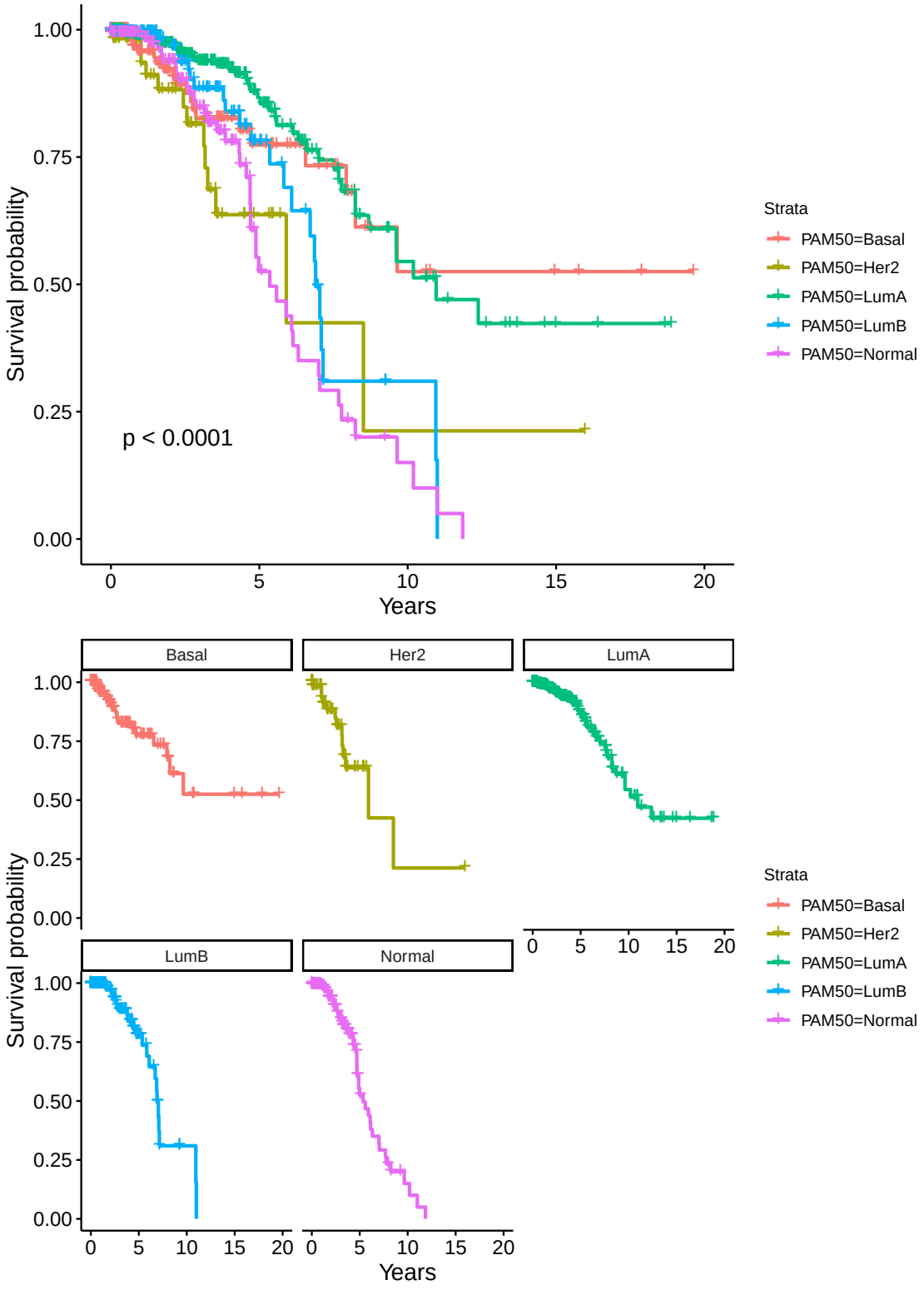

Figure 5.13: Kaplan-Meier curves showing the survival probability of the TCGA breast carcinoma samples classified according to the PAM50 system. Log-rank test was conducted in the combined graph.

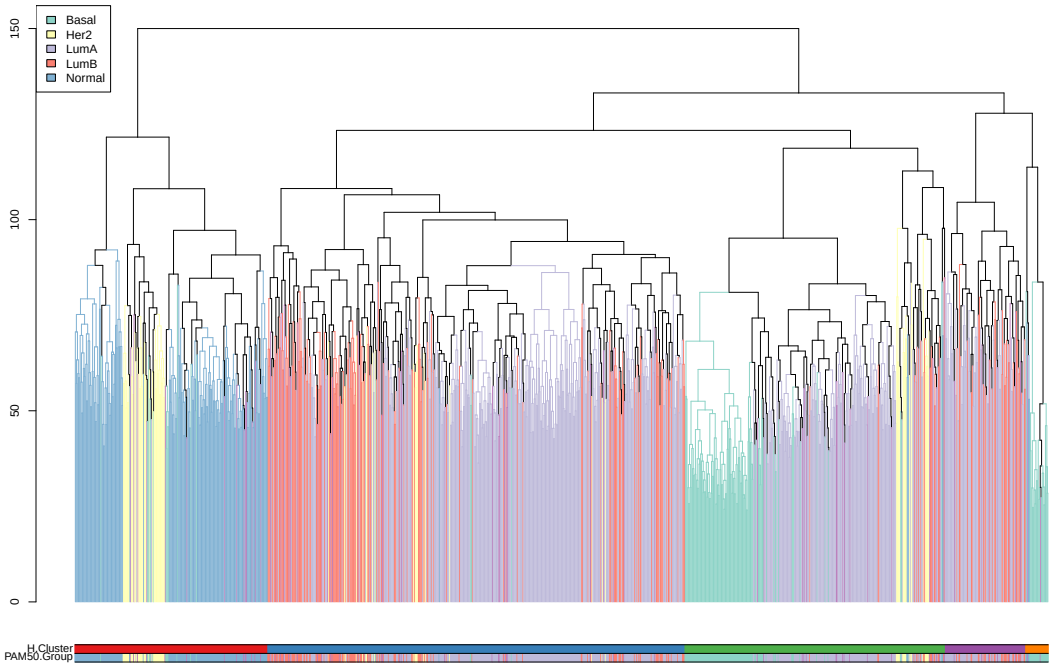

Figure 5.14: Dendrogram showing the sorting of the BRCA samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into five groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into the PAM50 groups (PAM50.Group). The dendrogram branches are colour-coded according to the corresponding PAM50 group.

and 43.3% assigned to Luminal B. LPD_4, on the other hand, had a higher amount of Luminal A, accounting for 70.07% of the samples, while Luminal B accounted for 29.13%. A similar pattern was seen in LPD_3 and LPD_5: LPD_3 is represented mainly by Luminal A (53.92%), followed by Luminal B (29.41%) and HER2 (14.7%); whereas LPD_5 is dominated by Luminal A (76.15%), followed by Luminal B (13.9%) and HER2 (7.94%). LPD_6 had the most distinct behaviour of any group, consisting virtually entirely of Basal samples (86.43%). LPD_7 was characterised as having a HER2 content of 51.56% and a Luminal B content of 35.93%. Finally, the bulk of Normal samples were assigned to LPD_8 accounting for 94.87% of this group.

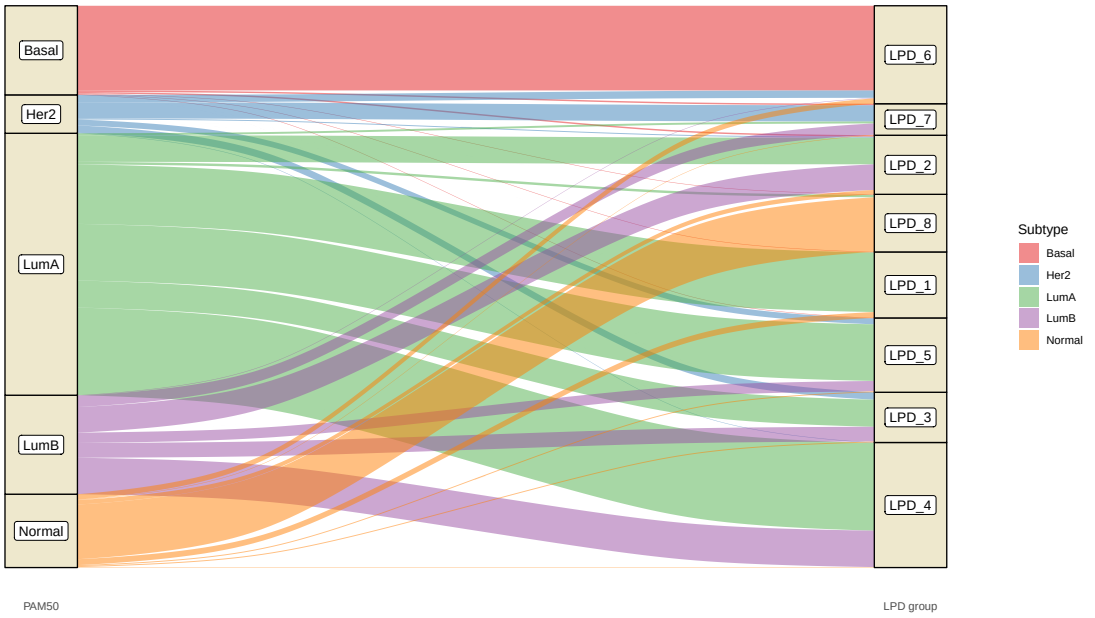

Figure 5.15: Alluvial plot representing the overlaps between the PAM50 classification and the LPD assignment for the BRCA samples. Each PAM50 group is assigned a distinct colour, enabling the visualization of how samples from each PAM50 group are allocated among the LPD groups.

5.3.2 Prostate cancer

Exploring the LPD output for PRAD

A total of 548 samples from 496 patients were analysed and seven LPD groups were found optimal, which were termed LPD_1 (n = 71, 12.95%), LPD_2 (n = 94, 17.15%), LPD_3 (n = 103, 18.79%), LPD_4 (n = 56, 10.21%), LPD_5 (n = 64, 11.67%), LPD_6 (n = 89, 16.24%), and LPD_7 (n = 71, 12.95%) (Fig 5.16). The Chi-squared test returned a significant overrepresentation of healthy tissue samples for LPD_5 ($n_{healthy}$ = 30, 48.88%, $P = 3.24\text{x}10^{-27}$; Chi-squared test), but the sample distribution into LPD groups was independent of TSS ($P = 0.13$; Chi-squared test). LPD_5 was the only group that showed a robust assignment (Fig 5.17). LPD_3, on the other hand, displayed shared assignment with LPD_7, as did LPD_5 with LPD_6.

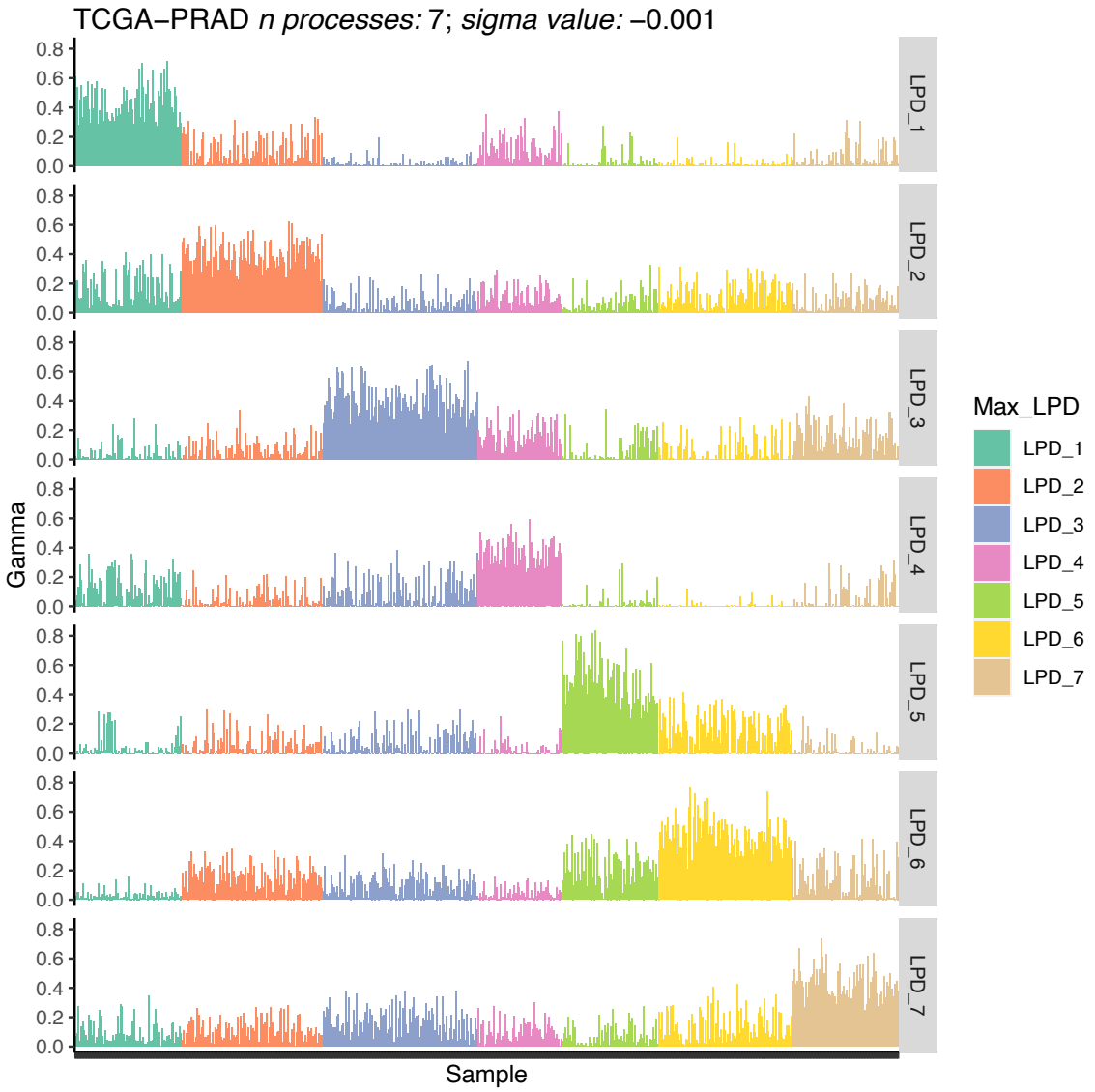

Figure 5.16: Gamma values of all samples for each detected LPD process in prostate adenocarcinoma. A total of 7 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_7. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD signature is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

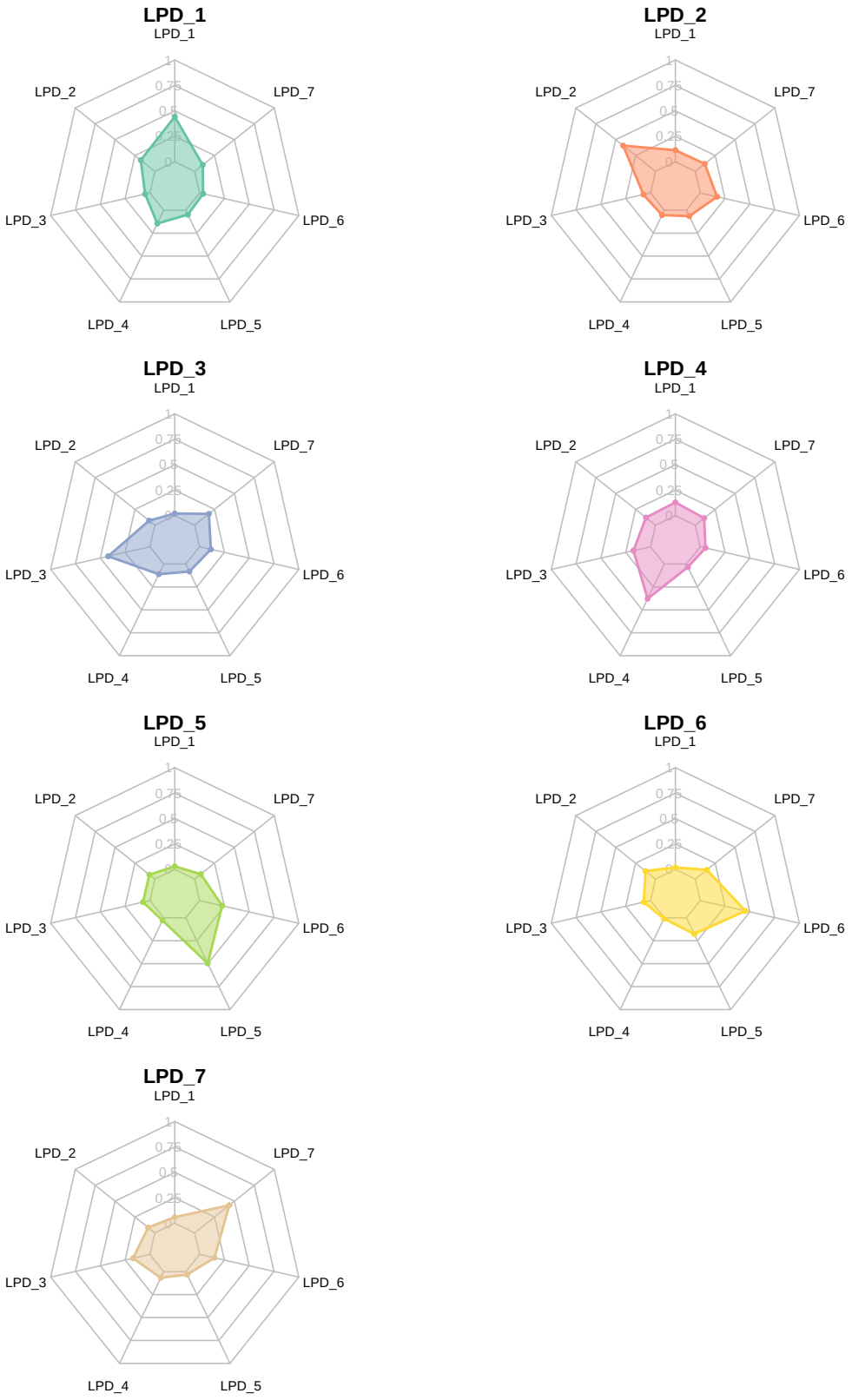

Figure 5.17: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in prostate adenocarcinoma. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

Clinicopathologic characteristics of the PRAD clusters

The clinical features of the tumour samples are available in Table 5.5. There were no significant association with age ($P = 0.0528$; Chi-squared test) or race ($P = 0.078$; Chi-squared test) across groups. Although the bulk of the samples were Gleason grade 7, LPD 1 and LPD 4 were enriched for grade 9 ($P = 0.0009$; Chi-squared test). There was a significant event-free survival in LPD_4 ($P = 0.016$; log-rank test), which had the worst prognosis across all groups (Fig 5.18). None of the LPD groups showed a significant association with PSA values (Table 5.6).

Table 5.5: Clinicopathologic features of the detected subtypes for prostate adenocarcinoma. chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	P-value
Age (years; mean (sd))	62.3 (6.85)	62.7 (7.45)	62.6 (6.82)	63.3 (6.69)	62.9 (5.99)	62.9 (6.44)	62.2 (7.14)	59.8 (7.46)	0.05
Race									
Asian	12	2	3	2	4	0	0	1	
Black or african american	63	11	12	6	4	9	14	7	
White	457	56	72	95	47	53	73	61	
American indian or alaska native	1	1	0	0	0	0	0	0	0.07
Gleason Score									
Grade 6	45	3	13	6	0	4	7	12	
Grade 7	246	16	57	56	11	15	48	43	
Grade 8	63	16	4	9	7	4	12	11	
Grade 9	137	26	20	29	36	10	11	5	
Grade 10	4	3	0	0	1	0	0	0	0.0009

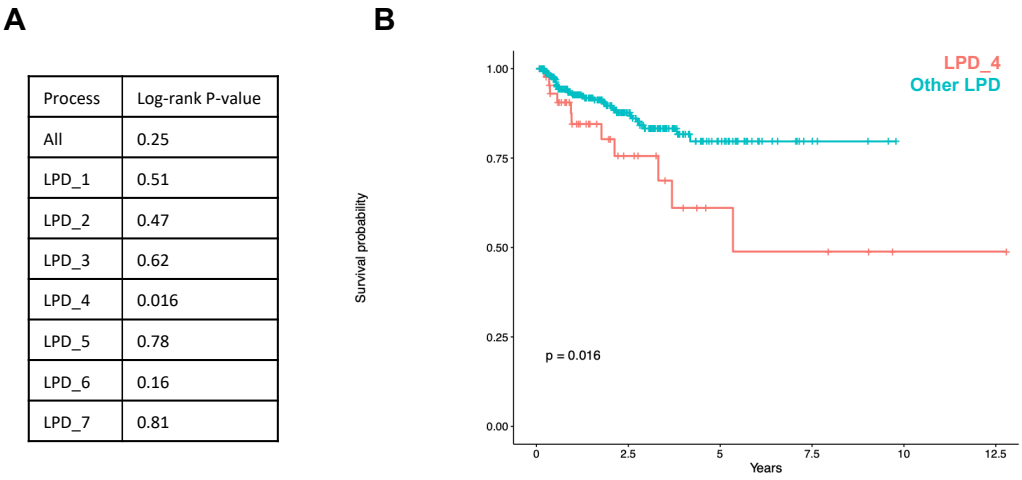

Figure 5.18: (A) Log-rank test outcome from assessing the survival curves for the samples of each LPD group when compared to the rest of the samples in prostate adenocarcinoma. (B) Kaplan-Meier curve illustrating the survival probability over time of the patients allocated in LPD_4 (displayed in red) in comparison to samples in other groups (displayed in blue). Log-rank test was conducted to compare both curves, and the p-value is provided.

Table 5.6: Cox analysis was conducted to examine the associations between PSA values and each LPD group. The coefficients in the analysis indicate whether the hazard risk increases (if positive) or decreases (if negative) with respect to PSA values. The standard error of the coefficient (SE) provides information about the precision of the estimate. The Wald statistical significance value (z) indicates the level of significance of the coefficient.

LPD Group	Coefficient	SE	z	p-value
LPD_1	-0.0011	0.0013	-0.9017	0.3672
LPD_2	-0.0012	0.0013	-0.9088	0.3635
LPD_3	-0.0012	0.0013	-0.9077	0.3640
LPD_4	-0.0012	0.0013	-0.8880	0.3745
LPD_5	-0.0012	0.0013	-0.9073	0.3642
LPD_6	-0.0012	0.0013	-0.8985	0.3689
LPD_7	-0.0012	0.0013	-0.9087	0.3635

Identification of differentially expressed genes in PRAD

A total of 5954 significant differentially expressed genes were identified throughout the eight LPD groups (Table 5.7; median across groups = 837; IQR = 531). According to their ratios, all groups were largely underexpressed. The top overexpressed and underexpressed genes ranked by *log2 fold change* are shown in Table 5.8. When it comes to driver genes, LPD_4 was the only group that did not include the driver gene *MUC6* among its DEGs, while LPD_6 was the only group that did not include *ALB* (Fig. 5.21). All the LPD groups showed distinct enrichment profiles both in KEGG (Fig. 5.19) and GO (Fig. 5.20). In KEGG, LPD_4 showed the largest size effect across all groups, affecting the overexpression of Cytokine-cytokine receptor interaction and JAK-STAT signalling pathway. In GO, LPD_4 was the group with the greatest number of overexpressed processes (9 out of 25). In the analysis checking for associations to cancer hallmarks (Figure 5.7), LPD_3 showed an association between underexpressed genes and tumour inflammation caused by tumoural cells.

Table 5.7: Gene counts for various categories in PRAD. These include the number of genes exhibiting significant differential expression and differential methylation, the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively), and the number of genes with significantly higher or lower frequency of CNV (referred to as overimpacted and underimpacted, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7
n health tissue samples	6	0	3	1	31	11	0
DEGs							
Upregulated	106	243	402	39	156	177	269
Downregulated	1145	739	922	350	523	315	568
Total	1251	982	1324	389	679	492	837
Ratio	0.09	0.33	0.44	0.11	0.3	0.56	0.47
DMGs							
Hypermethylated	6091	1497	335	2164	3629	1012	1123
Hypomethylated	2311	1514	144	2013	7824	3943	570
Total	8402	3011	479	4177	11453	4955	1693
Ratio	2.64	0.99	2.33	1.08	0.46	0.26	1.97
Mutated							
Overmutated	166	63	57	54	0	10	19
Undermutated	73	36	55	48	18	24	24
Total	239	99	112	102	18	34	43
Ratio	2.27	1.75	1.04	1.12	0	0.42	0.79

Table 5.8: The top five significantly differentially expressed genes that were overexpressed and underexpressed for each LPD group according to the *log2 fold change*. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status
LPD_1		
ENSG00000224099	3.615036	Overexpressed
MTNR1A	2.248674	Overexpressed
NETO1	1.856292	Overexpressed
S100A7A	1.825112	Overexpressed
MTND4P21	1.772190	Overexpressed
SEMG1	-7.171716	Underexpressed
SEMG2	-5.892613	Underexpressed
PAEP	-5.395901	Underexpressed
PATE1	-5.034581	Underexpressed
CRISP1	-4.583755	Underexpressed
LPD_2		
ENSG00000237250	2.669610	Overexpressed
KCNH5	2.332661	Overexpressed
TMEM196	2.284011	Overexpressed
SMR3B	2.205399	Overexpressed
CLC	2.105240	Overexpressed
APOA2	-4.298991	Underexpressed
FGA	-3.962494	Underexpressed
LIPF	-3.845629	Underexpressed
FGB	-3.845334	Underexpressed
ACTBP12	-3.737143	Underexpressed
LPD_3		
EDDM3B	6.279508	Overexpressed
EDDM3A	5.593542	Overexpressed
PATE4	5.146797	Overexpressed
PAEP	4.778401	Overexpressed
PATE1	4.362297	Overexpressed
DEFA5	-5.396080	Underexpressed
DEFA6	-4.494971	Underexpressed
APOA2	-3.955551	Underexpressed
COX7B2	-3.722701	Underexpressed
LINC00993	-3.543722	Underexpressed
LPD_4		
ENSG00000231421	3.011507	Overexpressed
RFTN1P1	2.032715	Overexpressed
SAA1	1.615941	Overexpressed
SAA2-SAA4	1.524770	Overexpressed
ENSG00000250658	1.495579	Overexpressed
XIRP2	-6.324237	Underexpressed

MYH7	-5.378590	Underexpressed
SMYD1	-5.292855	Underexpressed
APOA2	-5.152843	Underexpressed
MYL1	-5.030023	Underexpressed
LPD_5		
SCGB2A2	2.523154	Overexpressed
SCRT2	2.086026	Overexpressed
KRTAP20-2	1.970307	Overexpressed
ENSG00000224750	1.904636	Overexpressed
ENSG00000259788	1.865306	Overexpressed
MYH7	-7.353021	Underexpressed
MYL1	-7.247763	Underexpressed
MYL2	-7.047821	Underexpressed
SEMG1	-6.733322	Underexpressed
PATE1	-5.440793	Underexpressed
LPD_6		
ENSG00000257870	2.065291	Overexpressed
ENSG00000256101	1.982568	Overexpressed
NEUROD4	1.901945	Overexpressed
DBET	1.797888	Overexpressed
ANKRD20A9P	1.698838	Overexpressed
SEMG1	-6.401434	Underexpressed
SEMG2	-5.756053	Underexpressed
DEFA5	-5.220362	Underexpressed
PATE1	-5.110311	Underexpressed
SMR3B	-5.001748	Underexpressed
LPD_7		
CARD18	3.361582	Overexpressed
LINC01647	2.699024	Overexpressed
FAHD2P1	2.466740	Overexpressed
ENSG00000203987	2.370935	Overexpressed
ENSG00000242790	2.359671	Overexpressed
SEMG1	-7.156033	Underexpressed
SEMG2	-6.808584	Underexpressed
PATE1	-5.708379	Underexpressed
ACTBP12	-4.660729	Underexpressed
CRISP1	-4.254212	Underexpressed

Identification of differentially methylated genes in PRAD

Except for LPD_5, which gathered 33% of the DMGs, the number of DMGs were evenly distributed throughout the other groups (Table 5.7; median across groups = 4177; IQR = 4326). For LPD_1, LPD_3, LPD_4, and LPD_7 the majority of the DMGs were hypermethylated, whereas LPD_2 was slightly hypomethylated (ratio = 0.98) and LPD 5 was largely hypomethylated (ratio = 0.46). In terms of the overlap with known driver genes, LPD_1 and LPD_5 accumulated the majority of the differentially methylated driver

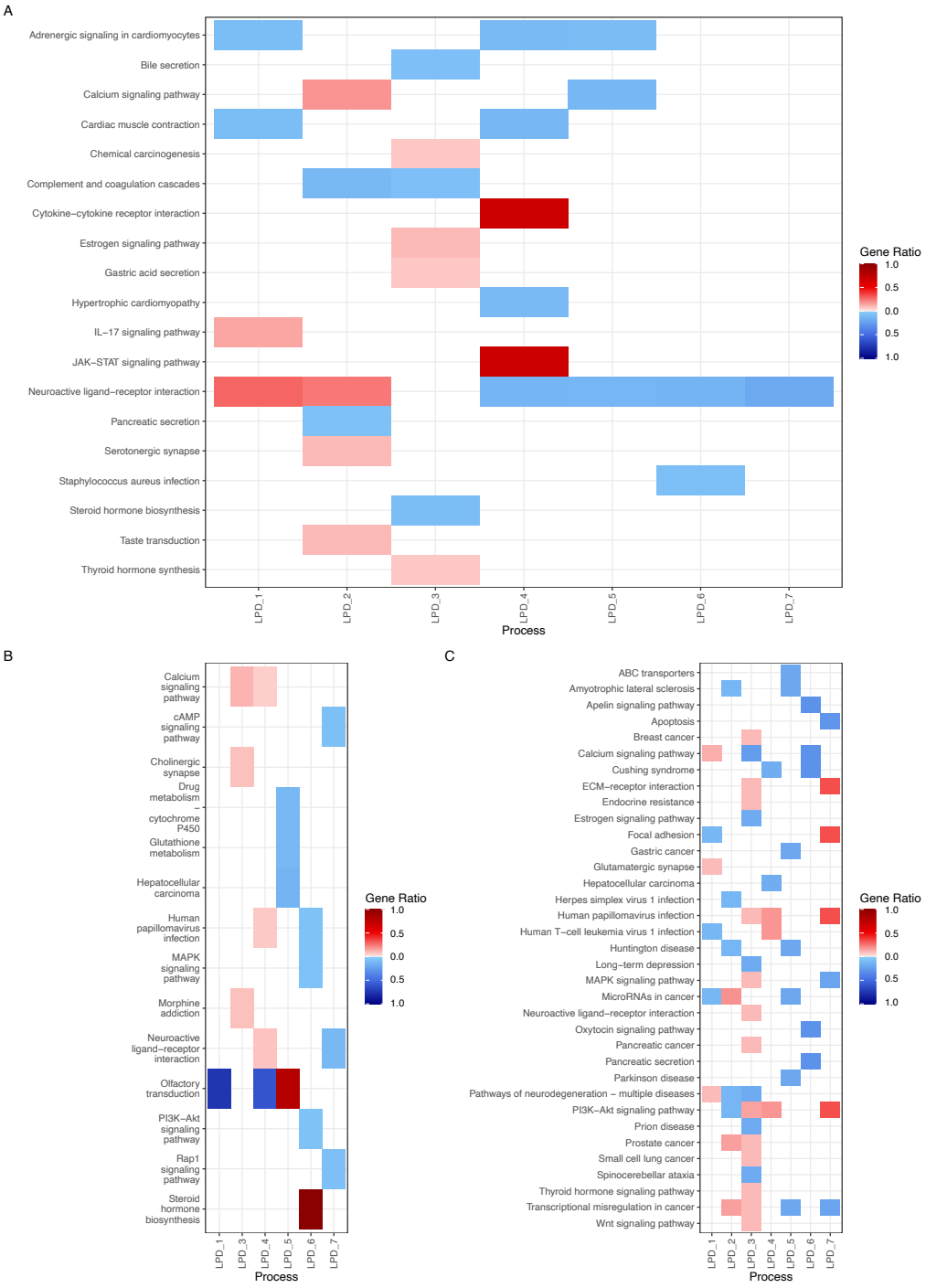

Figure 5.19: Biological pathways associated with different categories in prostate adenocarcinoma determined using KEGG enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, biological pathways associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the pathways associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological pathways is available in Supplementary Material B.

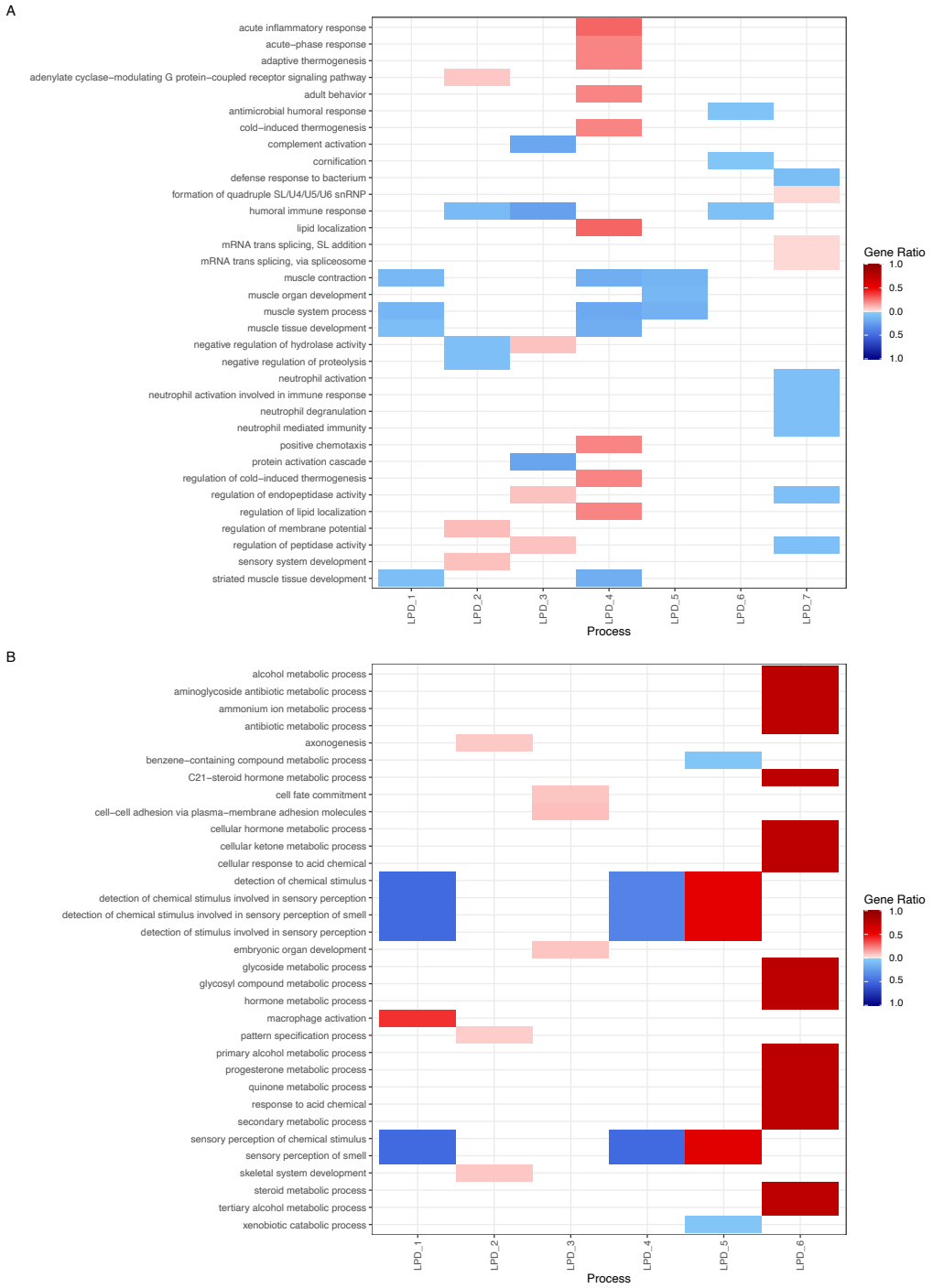

Figure 5.20: Biological processes associated with different categories in PRAD determined using GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant processes with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while processes linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, hypermethylated genes are depicted in red while hypomethylated are represented in blue. The complete list of associated biological processes is available in Supplementary Material B.

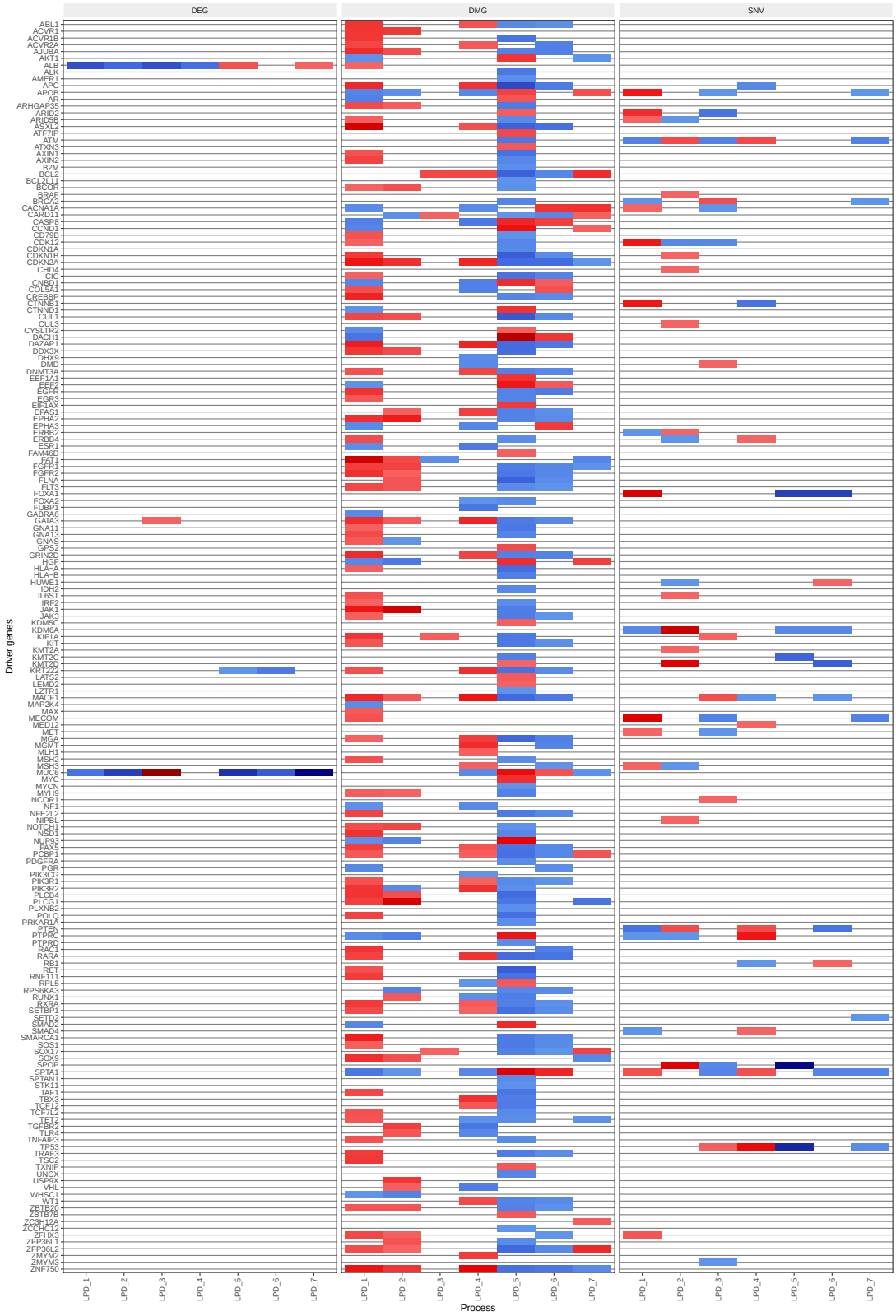

Figure 5.21: Heatmap showing the presence of driver genes across different categories in prostate adenocarcinoma, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV) and copy number variants (CNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV and CNV are represented in red, while genes with opposite characteristics are depicted in blue.

genes, but displayed opposing profiles: LPD_1 exhibited a hypermethylation for the same genes than LPD_5 exhibited a hypomethylation, and vice versa, when LPD_1 exhibited hypomethylation, LPD_5 exhibited hypermethylation (Fig. 5.21). In KEGG, both LPD_1 and LPD_4 revealed a large impact size for underexpression of olfactory transduction, which was countered by a large overexpression for the same pathway in LPD_5 (Fig. 5.19.B). This effect was also evident in the GO enrichment analysis findings, which revealed the same opposing impact directions in processes involved in the sensory detection and perception of chemical stimulus and smell (Fig. 5.20.B). LPD_6 was the group with the largest number of overexpressed pathways in GO (18 out of 34), all of which were related to metabolic processes.

Identification of genes affected by single nucleotide variants in PRAD

A median of 99 genes per LPD group (IQR = 68) were enriched or depleted of SNVs when compared to the rest of the samples. LPD_1 was the only group notably defined by a large overmutation (ratio = 2.27), while LPD_6 was the only group with a large undermutation (ratio = 0.41) (Table 5.7). Focusing in the driver genes, LPD_1, LPD_2, and LPD_4 had strong overmutations across several genes, whereas LPD_5 and LPD_6 had strong undermutations (Fig. 5.21). In KEGG enrichment analysis, LPD_7 showed the largest size effect that was involved in the overexpression of ECM-receptor interaction, focal adhesion, human papilloma virus infection and PI3K-Ak1 signalling pathway (Fig. 5.19.C). According to GO enrichment analyses, no biological processes were significantly enriched with subtype DMGs. No significant differences were observed between subtypes when comparing the SNP type, variant type, and variant class frequency across LPD groups ($P > 0.05$; Chi-squared test; Fig. 5.22). The majority of SNVs discovered were missense mutations induced by the point substitution of cytosine with thymine. LPD_2 displayed a significant enrichment in mutations in the *SPOP* gene (Fig. 5.23).

In the COSMIC mutational process profile of the LPD groups, all of them showed a strong contribution from signature 1 (Fig. 5.24), but there are differences. LPD_1 and LPD_3 showed a contribution from signature 15, and, along with LPD_4 and LPD_5, a partial contribution from signature 6. LPD_2 and LPD_6 had also a strong proportion of signature 3.

Additional evidence of a functional effect of differentially expressed genes in PRAD

Matches between overexpressed, and hypomethylated genes are shown in figure 5.25, while matches between underexpressed, hypermethylated, and mutated genes are shown in figure 5.26. In the overexpressed genes, LPD_2 had more matches with 20 hypomethylated genes (median of matches across groups = 5), while LPD_3 and LPD_7 had less with only one match ($P < 0.0001$; Chi-squared test). In the underexpressed genes, LPD_1 (137 hypermethylated, 2 mutated; median of matches across groups = 19) and LPD_5 (56 hypermethylated) were enriched in matches, while LPD_3 and LPD_6 were depleted with less than 7 matches each ($P < 0.0001$; Chi-squared test). The complete list of matched genes is available in Supplementary Material B.

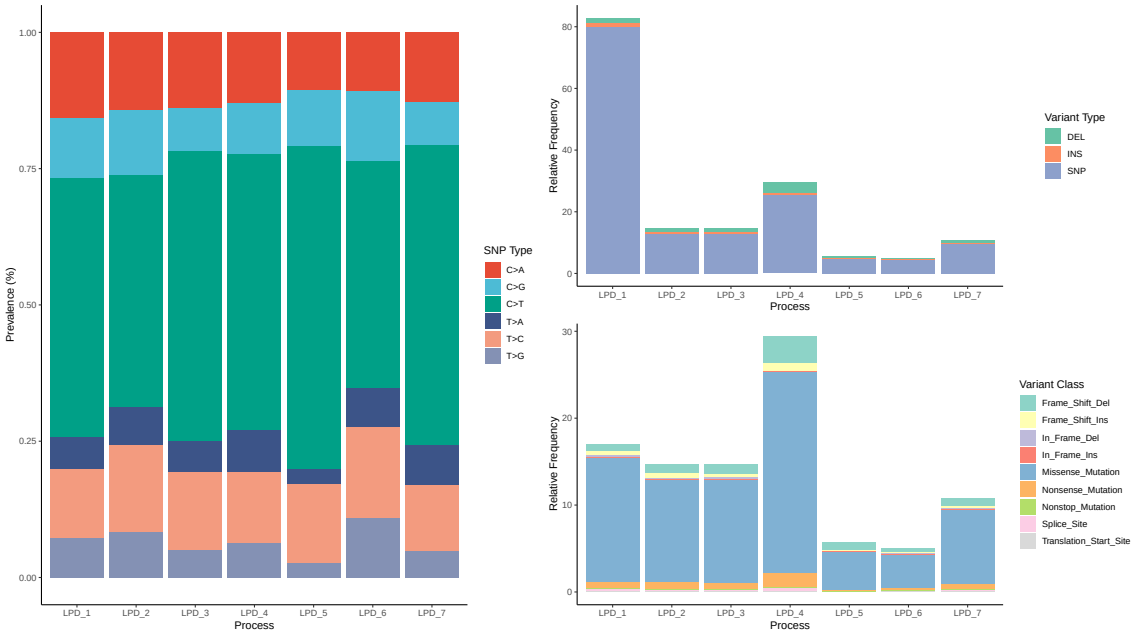

Figure 5.22: Detected single nucleotide variants (SNVs) within each LPD group for PRAD. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

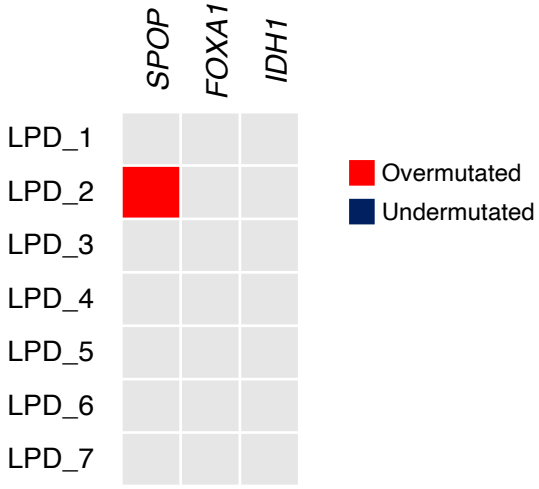

Figure 5.23: Heatmap displaying the mutational status of three genes (*SPOP*, *FOXA1*, *IDH1*) that play a key role in prostate cancer. Each column in the heatmap represents one of these genes, while each row represents a different LPD group in prostate cancer. The heatmap uses color-coding, with red indicating a significant overmutation (gene more frequently mutated) of a gene in a specific LPD group, and blue indicating a significant undermutation (gene less frequently mutated).

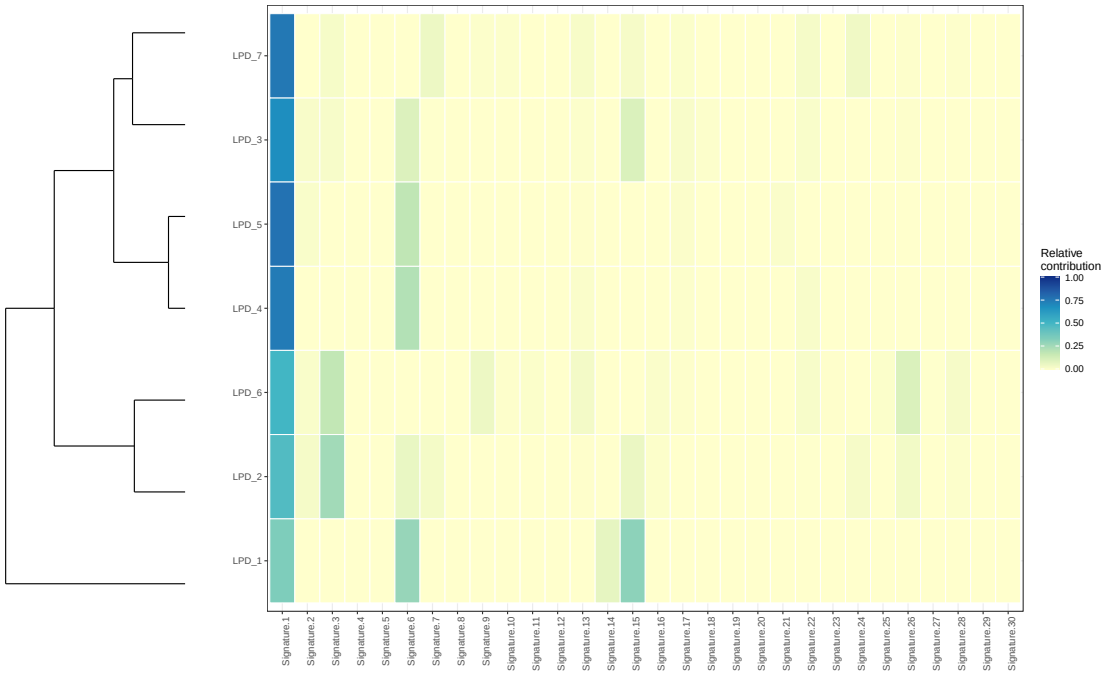

Figure 5.24: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in PRAD. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

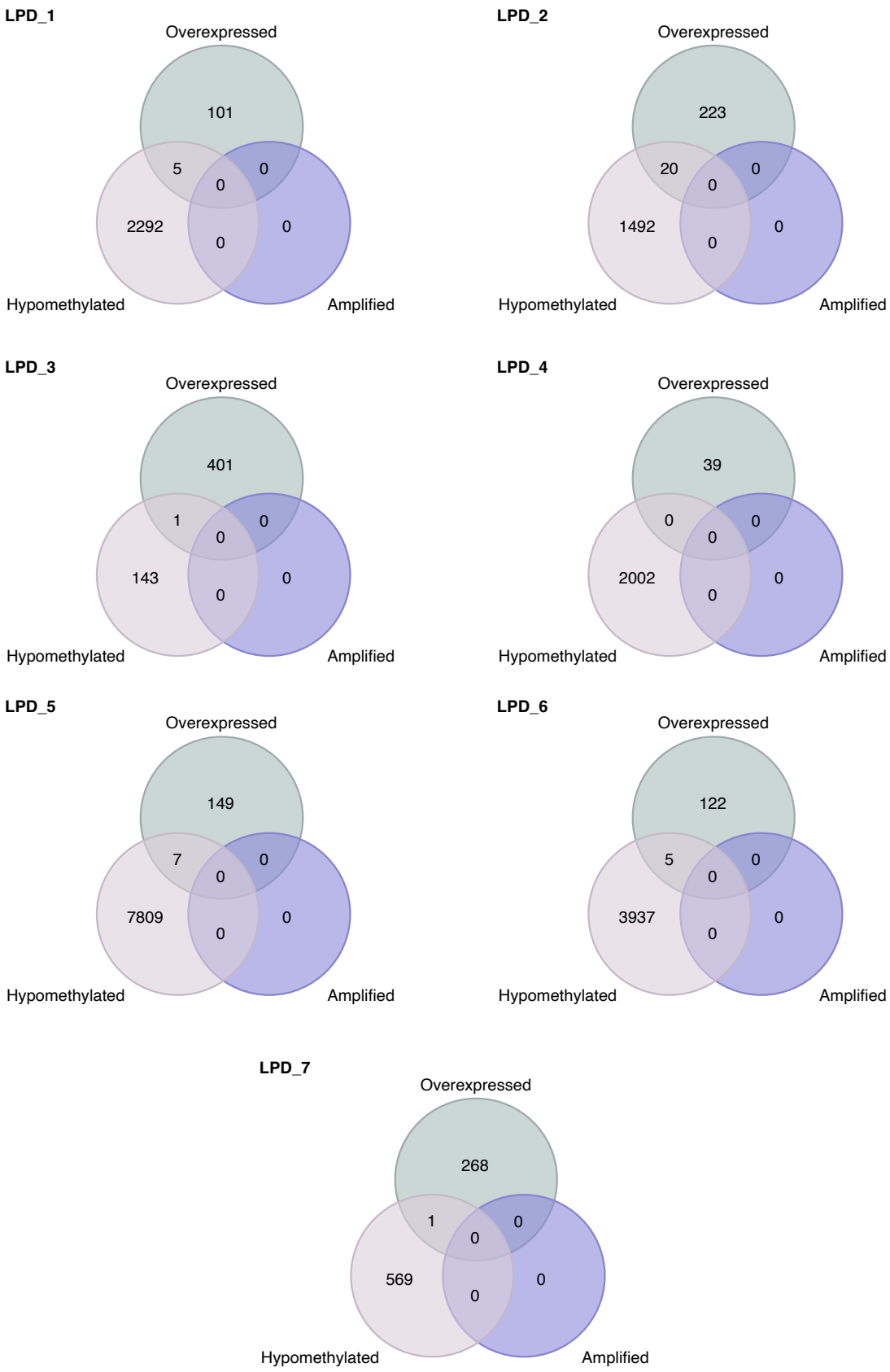

Figure 5.25: Venn diagram displaying the overlaps between three categories in genes in PRAD for each LPD group: overexpressed genes, hypomethylated genes, and amplified genes. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit overexpression, hypomethylation, and amplification simultaneously.

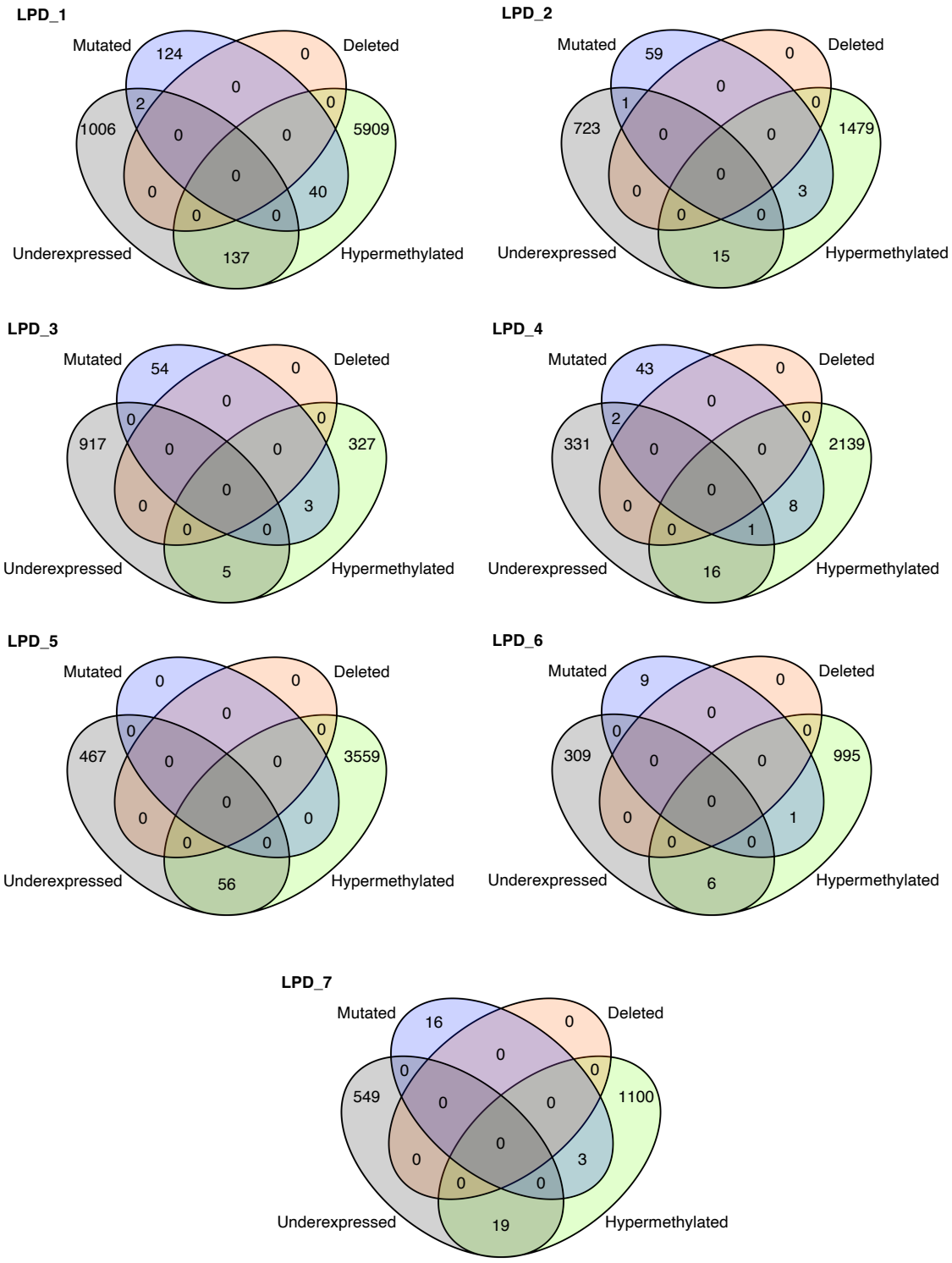

Figure 5.26: Venn diagram displaying the overlaps between four categories in genes in PRAD for each LPD group: uunderexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit underexpression, hypermethylation, mutations and deletions.

Comparison of the PRAD LPD output with Euclidian hierarchical clustering

Although samples from the same group tended to cluster together, there was no well-defined separation across the groups (Fig. 5.27). The green hierarchical clustering was entirely formed by a portion of the LPD_1 samples. Two other clusters, red and blue, appeared to be formed of a mix of LPD_2 and LPD_5. The remainder of the clusters had a heterogeneous composition.

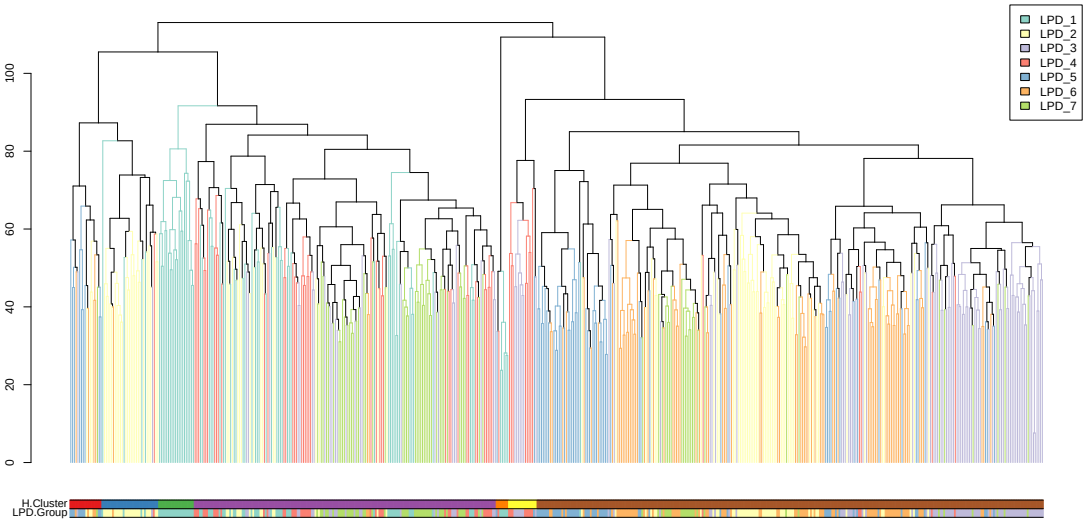

Figure 5.27: Dendrogram showing the sorting of the PRAD samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into seven groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

Comparison of the LPD output with DESNT

In their study, Luca et al. (2018)[118] used LPD to identify and characterise a low prognosis subtype of prostate cancer termed as DESNT (Figure 5.28). Aside from the poor prognosis, another characteristic of this subtype was the underexpression of a set of 45 genes, 16 of which were found hypermethylated in the TCGA. These 45 genes were consistent across various datasets and named as *core 45*.

When Luca's approach and my LPD results were compared, 56% of the samples assigned to LPD_1 and 77% of the samples assigned to LPD_4 were classified as DESNT (Fig. 5.29). LPD_2 (8%), LPD_3 (15%), and LPD_7 (22%), all had some DESNT samples in their groups. Additionally, the gamma values of LPD_4 exhibited a moderate positive

correlation (Pearson = 0.33) with the gamma value distribution of the DESNT samples, whereas LPD_2 exhibited a moderate negative correlation (Pearson = -0.34) (Fig. 5.30). The core 45 genes were only found as DEG in LPD_1 and LPD_3 (Table 5.9). However, LPD_1 had a match of 24 of them as DMG, LPD_4 showed a match of 13, LPD_5 had a match of 34, and LPD_6 had a match of 19. From those, I compared the set of 16 genes that Luca found hypermethylated in the TCGA. LPD_1, LPD_2, LPD_3, and LPD_4 showed them as hypermethylated, whereas LPD_5 and LPD_6 showed them as hypomethylated. No matches between the core 45 and genes affected by SNVs were observed.

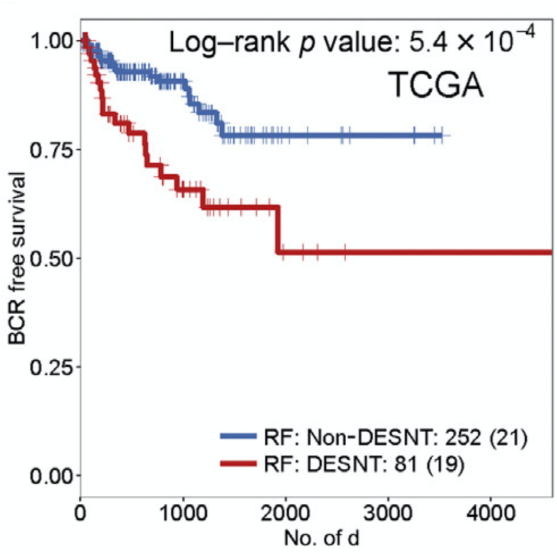

Figure 5.28: Kaplan-Meier curve showing the survival probability over time of the samples associated to Luca's DESNT subtype in comparison to the other samples in the TCGA. Obtained from Luca et al. (2018)[118].

5.3.3 Colorectal cancer

Exploring the LPD output for COAD

499 samples from 458 patients were collected and analysed. Seven LPD groups were found optimal by LPD: LPD_1 (n = 58, 11.62%), LPD_2 (n = 86, 17.23%), LPD_3 (n = 82, 16.43%), LPD_4 (n = 46, 9.21%), LPD_5 (n = 45, 9.01%), LPD_6 (n = 70, 14.02%), and LPD_7 (n = 112, 22.44%) (Fig 5.31). Normal tissue samples were significantly overrepresented in LPD_5 ($n_{healthy}$ = 41, 91.11%, $P = 3.1x10^{-96}$; Chi-squared test), but the sample distribution of LPD groups was independent of TSS ($P = 0.21$; Chi-squared test). LPD_5 and LPD_7 showed a robust assignment (mean gamma>0.5; Fig 5.32). LPD_1 displayed a shared assignment with LPD_3, as did LPD_2 with LPD_4.

Clinicopathologic characteristics of the COAD clusters

The clinical features of the tumour samples are available in Table 5.10. There were no significant differences in age ($P = 0.12$; Chi-squared test) or race ($P = 0.72$; Chi-squared test) across groups. LPD_3 was enriched for the colon mucious adenocarcinoma ($P = 0.0099$; Chi-squared test). No significant event free survival events were found ($P = 0.73$; Log-rank test).

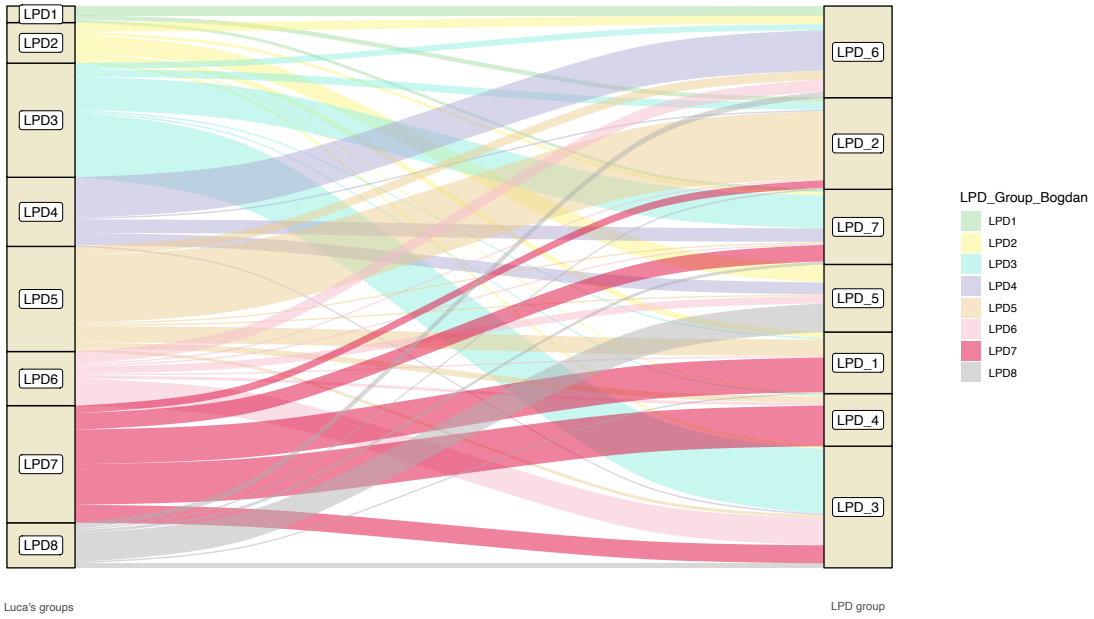

Figure 5.29: Alluvial plot showcasing a comparison between the sample assignments using Luca's approach (displayed on the left) and the LPD approach described in this thesis (displayed on the right). Luca's LPD_7 is the DESNT subtype.

Table 5.9: Number of genes shared between the LPD groups and the core 45 genes of DESNT, as well as the subset of 16 hypermethylated genes found in the TCGA by Luca et al. (2018)[118]. Within each LPD group, a comparison is made to identify the number of genes that are overexpressed, underexpressed, hypermethylated, and hypomethylated and are also present in Luca's core 45 gene set. Additionally, the hypermethylated and hypomethylated genes are compared to the set of 16 genes found to be hypermethylated in the TCGA by Luca et al. (2018)[118].

	Core 45				16 hypermethylated	
LPD Group	Overexpressed	Underexpressed	Hypermethylated	Hypomethylated	Hypermethylated	Hypomethylated
LPD_1	0	2	22	2	11	0
LPD_2	0	0	9	3	7	0
LPD_3	3	0	2	0	1	0
LPD_4	0	0	11	2	6	1
LPD_5	0	0	1	33	0	13
LPD_6	0	0	2	16	0	9
LPD_7	0	0	3	0	0	0

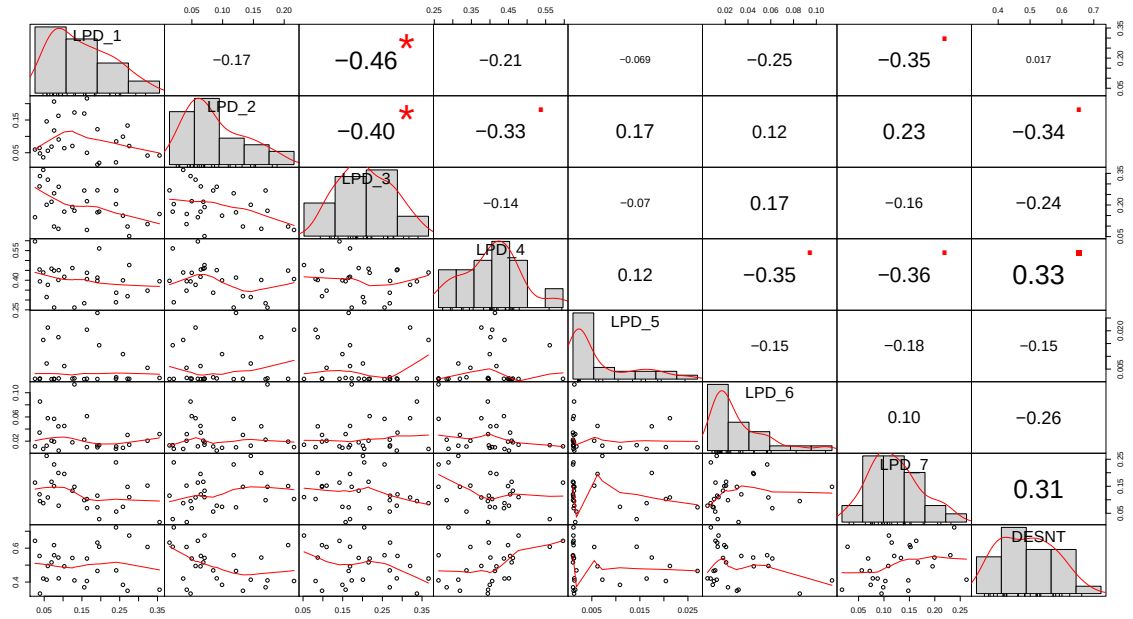

Figure 5.30: Correlation chart of the LPD groups found in prostate adenocarcinoma and the DESNT subtype. Correlations were calculated by comparing the gamma values of the samples assigned to each of the groups. The diagonal of the chart represents the distribution of the data points from each LPD group and DESNT through a barplot. The lower triangle displays a bivariate scatter plot of the data points from both groups with a fitted line indicating the trend. The upper triangle of the chart displays the correlation values and significance levels, with significant correlations highlighted in red (red square indicating p-value < 0.05, * indicating p-value < 0.01).

Table 5.10: Clinicopathologic features of the detected subtypes for colorectal adenocarcinoma. chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	P-value
Age (years; mean (sd))	68.6 (13.3)	69.9 (13.9)	67.9 (12.2)	69.5 (13.9)	66.8 (13.7)	70.5 (14.0)	64.8 (13.7)	70.1 (12.4)	0.124
Race									
Asian	11	0	3	4	1	0	2	1	
Black or african american	63	5	14	9	12	3	10	10	
White	230	26	44	47	31	20	38	24	
American indian or alaska native	2	0	0	1	0	1	0	0	0.72
Histology									
Colon adenocarcinoma	387	45	79	55	39	3	66	100	
Colon mucious adenocarcinoma	62	12	3	25	6	1	4	11	0.0099

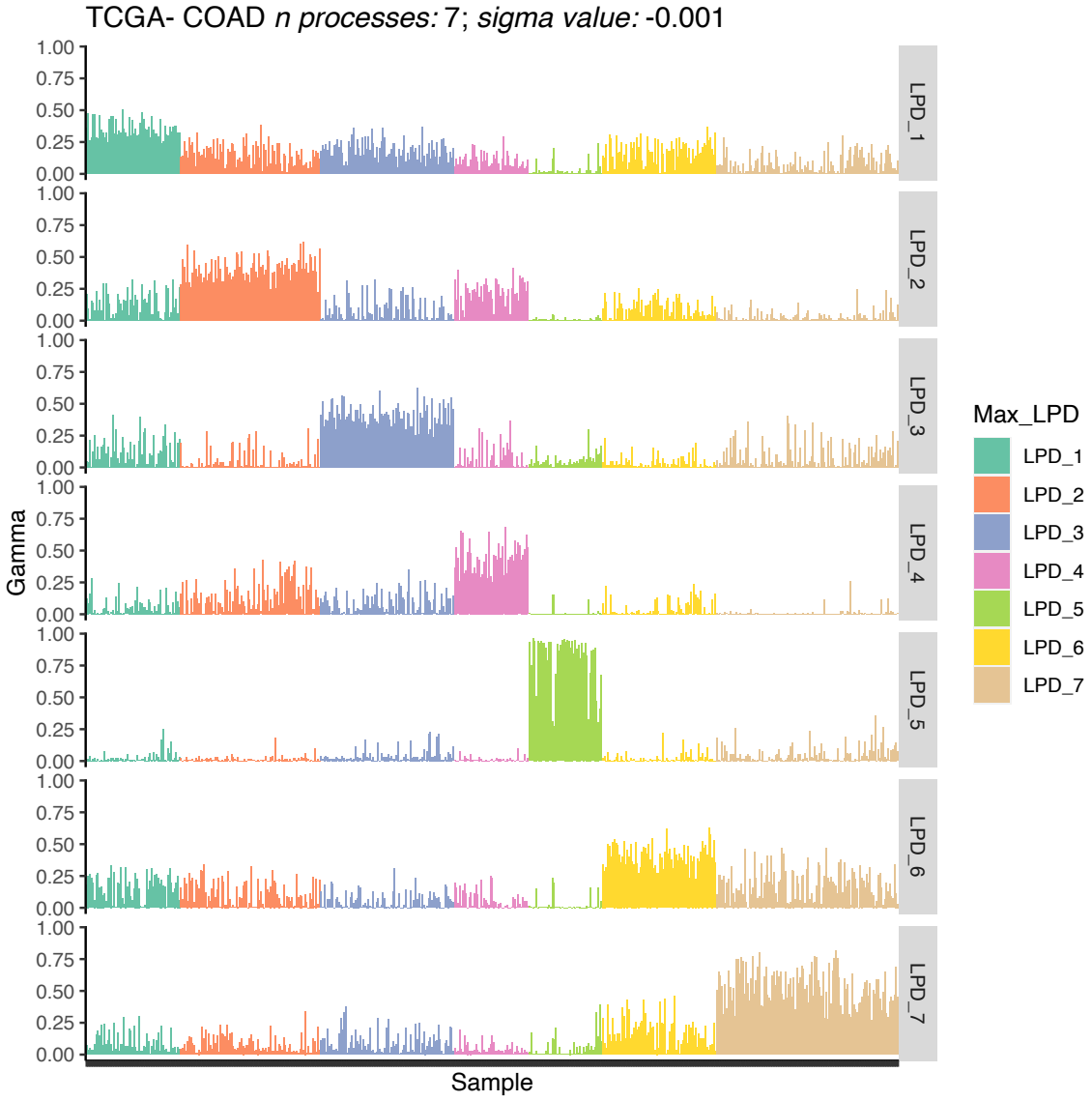

Figure 5.31: Gamma values of all samples for each detected LPD process in colon carcinoma. A total of 7 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_7. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD signature is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

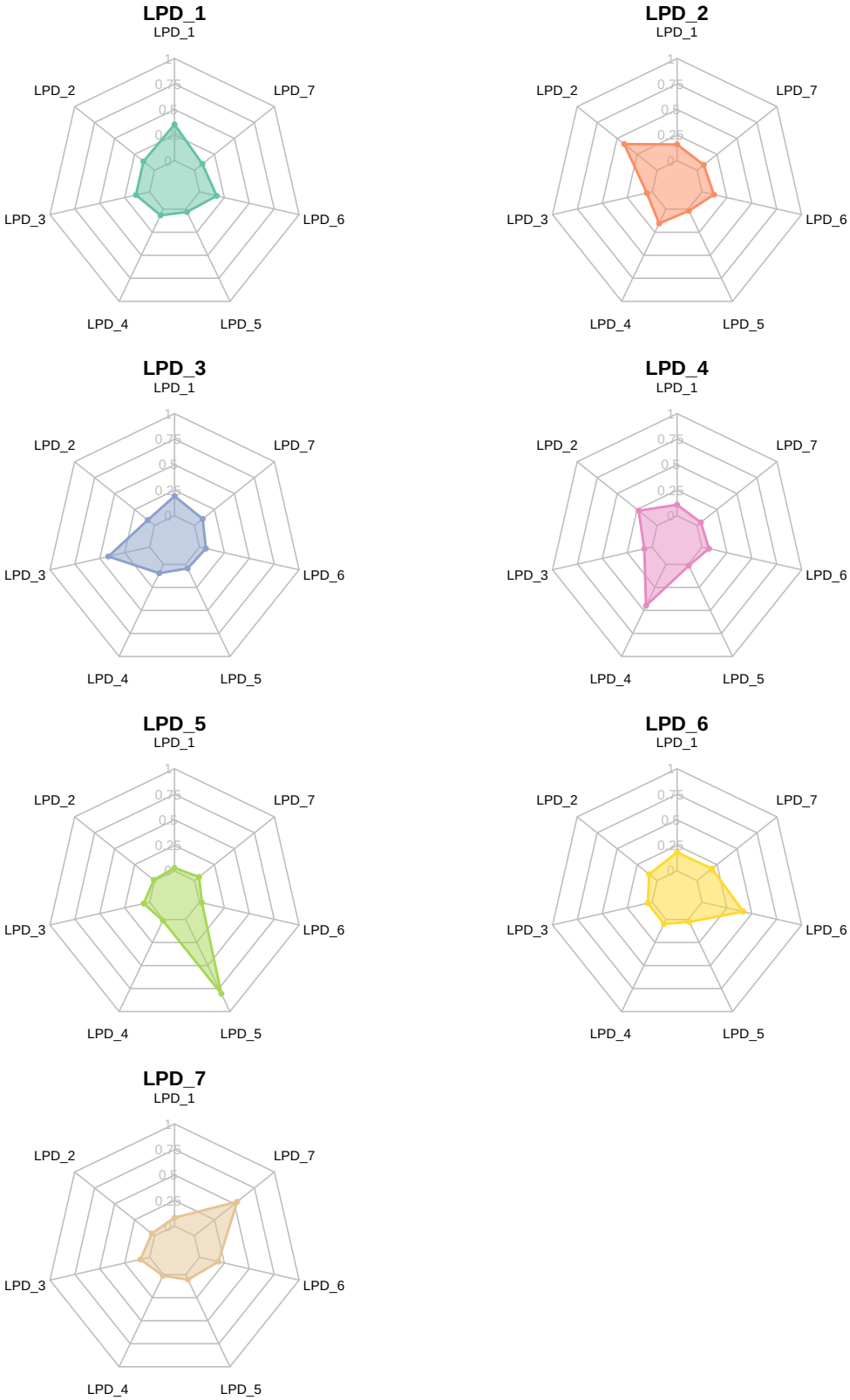

Figure 5.32: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in colon adenocarcinoma. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

Identification of differentially expressed genes in COAD

A total of 15560 significant differentially expressed genes were identified throughout the seven LPD groups (Table 5.11; median per group = 2503; IQR = 1333). The top overexpressed and underexpressed genes ranked by *log2 fold change* are shown in Table 5.12. According to their ratios, LPD_1 (ratio = 7.85), and LPD_2 (ratio = 4.14) both had a strong weighting to overexpressed DEGs. Of the known driver genes, LPD_1 showed a strong underexpression of the gene *ALB* compared to the other subtypes, LPD_7 had a strong underexpression of the gene *APC*, and LPD_5 had a strong overexpression of the gene *KIF1A* (Fig. 5.34). In KEGG, the neuroactive ligand-receptor interaction pathway had a significantly enrichment of underexpressed DEGs identified in all LPD groups except LPD_1 (Fig. 5.33). No significant pathways were detected by GO enrichment analysis (P>0.05; hypogeometric test). Two distinct patterns were observed in the analysis checking for associations to cancer hallmarks (Figure 5.7). LPD_1 displayed enrichment in biological processes related to unlimited replication, indicating a potential association with enhanced cell proliferation and growth. On the other hand, LPD_4 and LPD_6 showed an association between underexpressed genes and tumour inflammation caused by tumoural cells in healthy cells.

Table 5.11: Gene counts for various categories in COAD. These include the number of genes exhibiting significant differential expression and differential methylation, the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively), and the number of genes with significantly higher or lower frequency of CNV (referred to as overimpacted and underimpacted, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7
n health tissue samples	0	0	0	0	41	0	0
DEGs							
Upregulated	2561	892	982	724	283	33	45
Downregulated	326	215	899	1779	396	2734	3691
Total	2887	1107	1881	2503	679	2767	3736
Ratio	7.86	4.16	1.2	0.41	0.71	0.01	0.01
DMGs							
Hypermethylated	34	198	7142	1667	8553	75	138
Hypomethylated	5	1587	889	1809	4974	2257	766
Total	39	1785	8031	3476	13527	2332	904
Ratio	6.8	0.12	8.02	0.92	1.72	0.03	0.18
Mutated							
Overmutated	829	66	7156	28	0	25	846
Undermutated	3030	2133	545	1983	91	1611	2884
Total	3859	2199	7701	2011	91	1636	3730
Ratio	0.27	0.03	13.13	0.01	0	0.02	0.27

Table 5.12: The five genes more overexpressed (log2FoldChange > 1) and underexpressed (log2FoldChange < -1) for each LPD group in colorectal adenocarcinoma. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status
LPD_1		
RN7SL507P	5.032915	Overexpressed
ENSG00000202198	4.768499	Overexpressed
RNU1-88P	4.728191	Overexpressed
RN7SKP255	4.472533	Overexpressed
H1-5	4.465810	Overexpressed
CLDN18	-4.916779	Underexpressed
CSAG1	-4.652485	Underexpressed
REG1B	-4.253761	Underexpressed
MAGEA12	-3.979264	Underexpressed
HBZ	-3.940116	Underexpressed
LPD_2		
RNA5SP215	4.969380	Overexpressed
RNA5SP242	4.963592	Overexpressed
RNA5SP191	4.602151	Overexpressed
RNA5SP19	4.468207	Overexpressed
RNA5SP141	4.080148	Overexpressed
CT83	-3.526326	Underexpressed
CALB1	-3.511444	Underexpressed
MTND1P23	-3.084556	Underexpressed
CLDN18	-2.927589	Underexpressed
ENSG00000251577	-2.860505	Underexpressed
LPD_3		
ENSG00000259098	4.071201	Overexpressed
IGLJ6	3.010844	Overexpressed
IGKV3-25	2.884749	Overexpressed
ENSG00000244239	2.803467	Overexpressed
HBZ	2.760149	Overexpressed
MEP1AP4	-3.626357	Underexpressed
PNRC2P1	-3.306243	Underexpressed
MTND1P23	-3.221205	Underexpressed
CTAGE8	-3.212173	Underexpressed
TBC1D3L	-3.122095	Underexpressed
LPD_4		
ENSG00000227706	3.478345	Overexpressed
LINC01029	3.205211	Overexpressed
MT-TE	3.130519	Overexpressed
HBZ	3.030297	Overexpressed
KRTAP21-4P	2.778233	Overexpressed
OTOP2	-5.877087	Underexpressed

ADIPOQ	-5.197516	Underexpressed
IGHV3-7	-4.605075	Underexpressed
OTOP3	-4.241783	Underexpressed
CA1	-4.087425	Underexpressed
LPD_5		
KIF1A	2.710388	Overexpressed
SLC7A14	2.477655	Overexpressed
PCSK2	2.456239	Overexpressed
SPOCK3	2.427280	Overexpressed
HOXC13	2.423373	Overexpressed
RNA5SP141	-5.029740	Underexpressed
RNA5SP149	-3.980359	Underexpressed
IGF2	-3.905600	Underexpressed
FOXG1	-3.789494	Underexpressed
RNA5SP202	-3.783853	Underexpressed
LPD_6		
SULT1E1	2.689792	Overexpressed
LINC00607	2.524841	Overexpressed
GABRP	2.293974	Overexpressed
LINC00523	2.145855	Overexpressed
C8orf49	1.647288	Overexpressed
OTOP2	-5.331797	Underexpressed
PYY	-4.912999	Underexpressed
GUCA2B	-4.883646	Underexpressed
RNA5SP141	-4.632673	Underexpressed
HBZ	-4.345070	Underexpressed
LPD_7		
MAGEB2	1.800112	Overexpressed
ENSG00000215512	1.783016	Overexpressed
SPRR2F	1.631415	Overexpressed
TBC1D3D	1.540495	Overexpressed
TBC1D3L	1.506476	Overexpressed
RNA5SP141	-5.837125	Underexpressed
IGLJ3	-5.823278	Underexpressed
RNA5S9	-5.769115	Underexpressed
CNN2P4	-5.725669	Underexpressed
ENSG00000226532	-5.672581	Underexpressed

Identification of subtype characteristic differentially methylated genes in COAD

LPD_1 had a very low number of DMGs compared to the other groups, as it only accounted for the 1% of the total DMGs for COAD (Table 5.11; median across groups = 2332; IQR = 4409). LPD_1, LPD_3 and LPD_5 were characterised by the majority of DMGs being hypermethylated, while the other groups were strongly hypomethylated. Focusing in the driver genes, LPD_ 3 and LPD_5 accumulated most of the driver DMGs, while LPD_2,

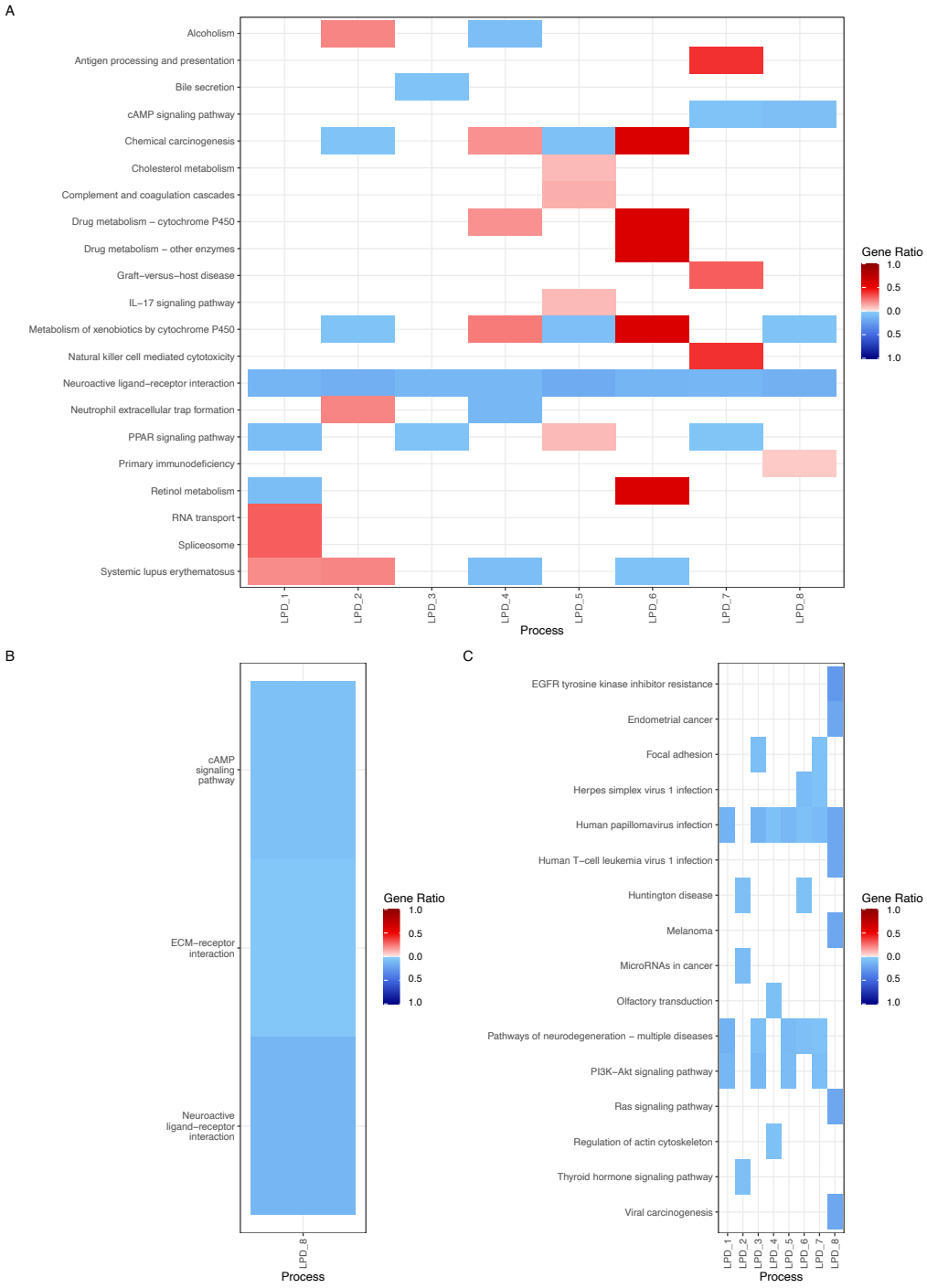

Figure 5.33: Biological pathways associated with different categories in colorectal adenocarcinoma determined using KEGG enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, biological pathways associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the pathways associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological pathways is available in Supplementary Material B.

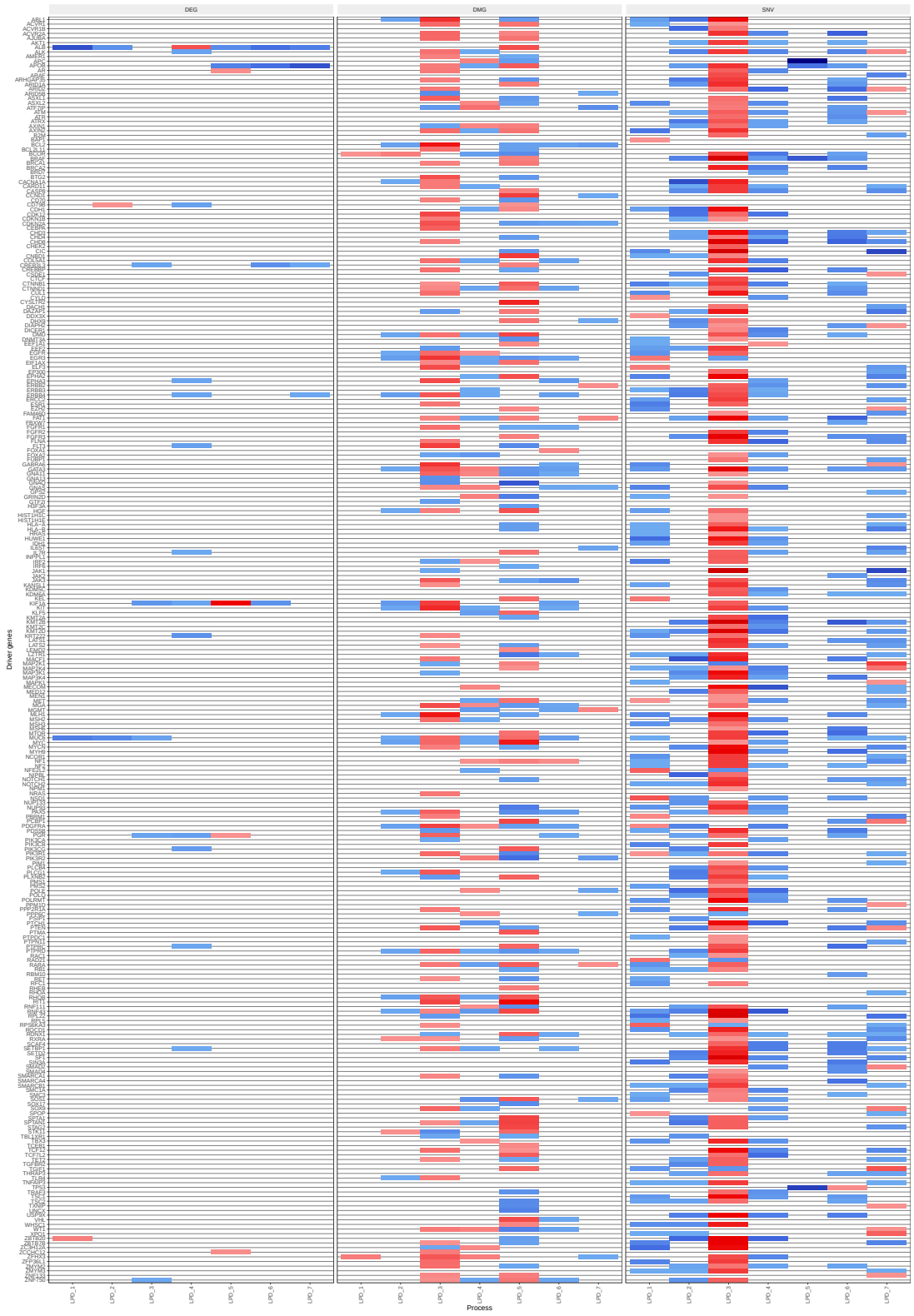

Figure 5.34: Heatmap showing the presence of driver genes across different categories in colorectal adenocarcinoma, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV) and copy number variants (CNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV and CNV are represented in red, while genes with opposite characteristics are depicted in blue.

LPD_4 and LPD_6 displayed similar profiles between them (Fig. 5.34). LPD_1 DMGs were not found to be significantly enriched in any KEGG pathway (Fig. 5.33.B). Neuroactive ligand-receptor interactions was the most commonly altered pathway, found in all subtype DMGs except for LPD_1 and LPD_7. None of the LPD groups showed a strong size effect for any pathway. The analysis in GO yielded similar results, with again LPD_1 DMGs lacking significant enriched pathways (Fig. 5.35.A).

Identification of genes affected by single nucleotide variants in COAD

A median of 2199 genes per LPD group were enriched or depleted in single nucleotide variations (SNVs) in comparison to other groups (IQR = 1971). LPD_3 was the sole overmutated group (ratio of enriched genes to depleted genes = 13.13), while LPD_2 (ratio = 0.03) and LPD_6 (ratio = 0.01) had the lowest ratios (Table 5.11). LPD_5 had a considerable smaller number of enriched/depleted SNVs in known driver genes than the other groups (Fig. 5.34). LPD_3 had the majority of the driver genes that were enriched in SNVs, whereas the remaining groups appeared to be defined by depleteion in driver genes. According to KEGG enrichment analysis, LPD_6 had the largest effect size across all groups, which included T cell differentiation, herpes virus infection, hepatitis C, and atherosclerosis that had a significant overrepresentation of genes enriched in SNVs compared to other subtypes (Fig. 5.33.C). The enrichment in GO, on the other hand, returned biological processes only for LPD_2, LPD_4, LPD_6, and LPD_7 and with none of them showing a strong size effect (Fig. 5.33.B). No differences were observed when comparing the SNP type, variant type, and variant class frequency across LPD groups (P>0.05; Chi-squared text; Fig. 5.36). The majority of SNVs discovered were missense mutations induced by the point substitution of cytosine with thymine.

In the COSMIC mutational signature profile of the LPD groups, all of them showed a strong contribution from signature 1 (Fig. 5.37). Additionally, except for LPD_5, all groups showed to an extent a contribution from signature 6. LPD_1, LPD_3, LPD_4 and LPD_7 had a contribution from signature 15, and LPD_1 along with LPD_7 had also a strong link to signature 10.

Additional evidence of a functional effect of differentially expressed genes in COAD

Matches between overexpressed, and hypomethylated are shown in Figure 5.38, while matches between underexpressed, hypermethylated, and mutated genes are shown in Figure 5.39. In the overexpressed genes, LPD_5 was enriched with 74 matches with hypomethylated genes, whereas LPD_1 and LPD_3 were depleted with 5 or less matches ($P < 0.0001$; Chi-squared test). In the underexpressed genes, LPD_3, LPD_4, and LPD_5 were enriched in matches, while LPD_6 and LPD_7 were depleted ($P < 0.0001$; Chi-squared test). The complete list of matched genes is available in Supplementary Material B.

Comparison of the COAD LPD output with Euclidian hierarchical clustering

LPD_5 showed a well-defined separation from the other groups (Fig. 5.40). Samples belonging to LPD_7, although scattered across different hierarchical clusters, tended to

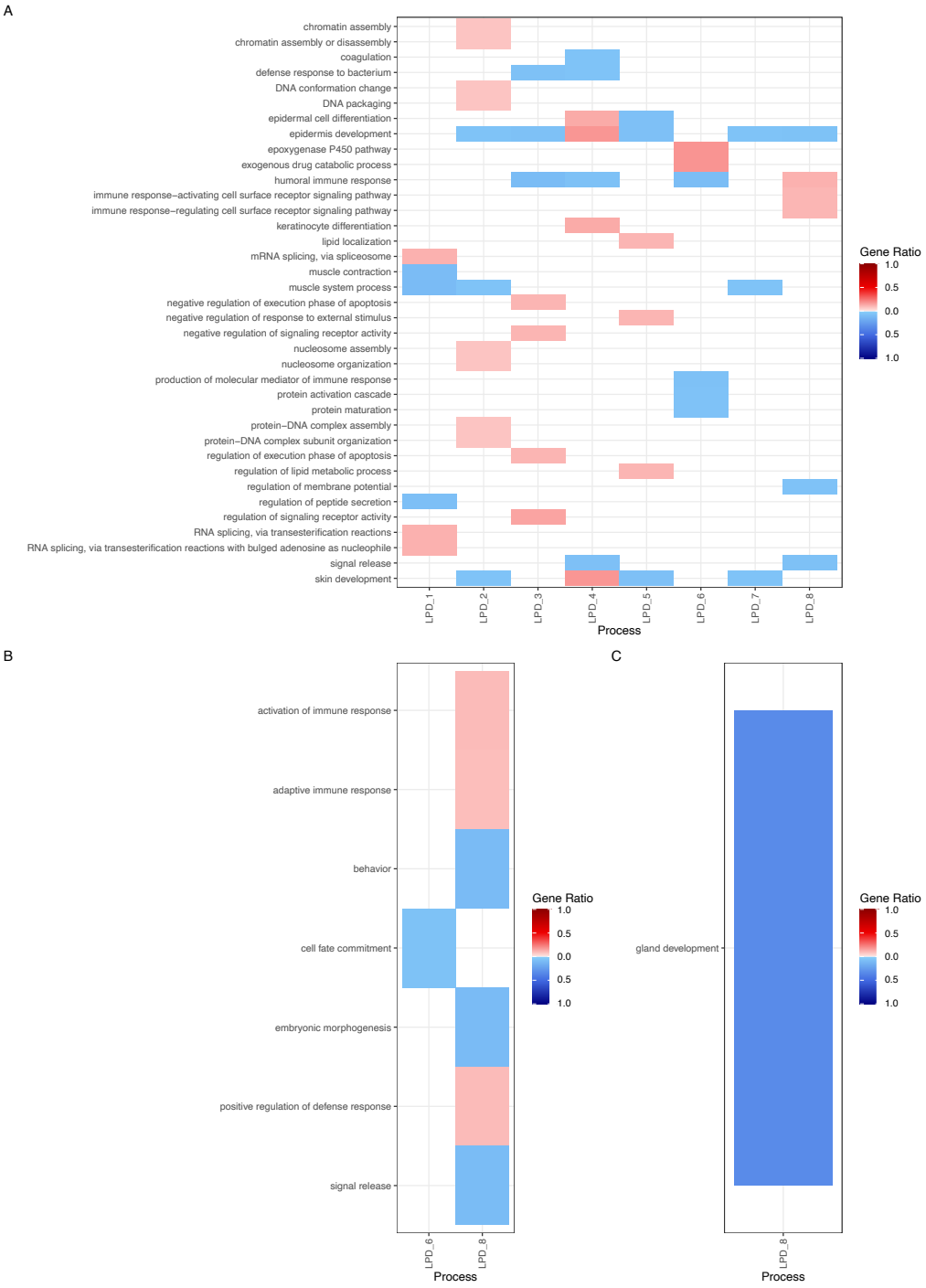

Figure 5.35: Biological processes associated with different categories in COAD determined using GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant processes with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while processes linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the processes associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological processes is available in Supplementary Material B.

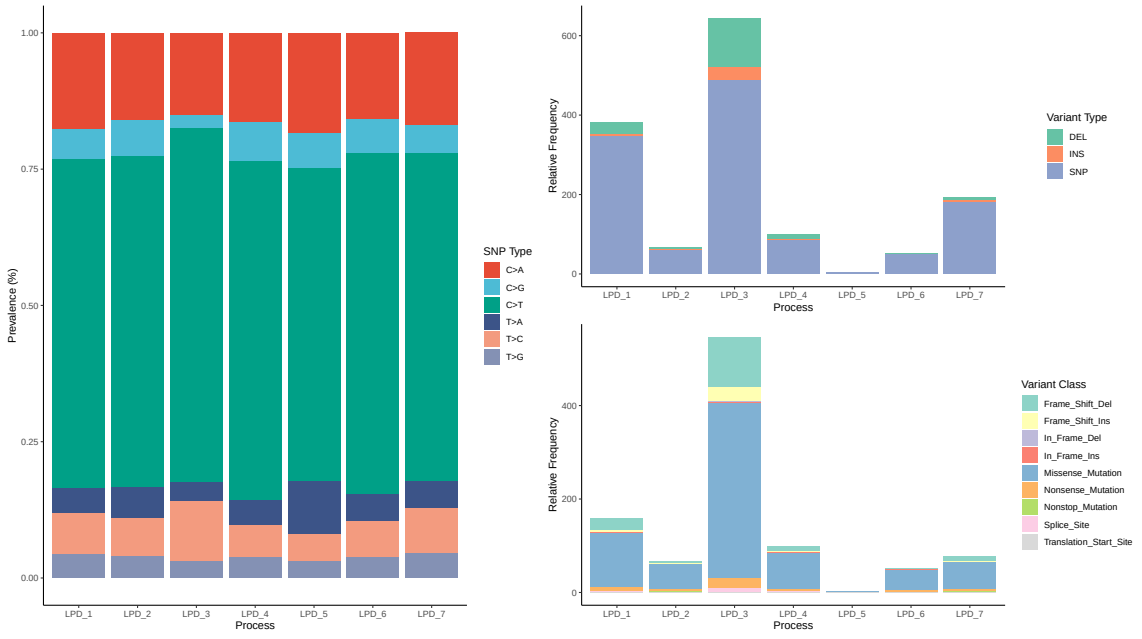

Figure 5.36: Detected single nucleotide variants (SNVs) within each LPD group for COAD. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

be grouped together. LPD_2 and LPD_4 appeared to be clustered together. Except for the blue hierarchical cluster, all the hierarchical clusters showed a mix of different LPD groups.

Comparison of the LPD output with Pericol

In their work, Ellis (2021)[171] used LPD in the TCGA-COAD dataset and detected six subtypes, labelled as C1, C2, C3, C4, C5, and C6. The subtype C3 showed a good correlation to the Pericol signature, which was a low-prognosis subtype of colorectal cancer detected by Ellis in other four datasets.

When the outcome of Ellis's LPD approach and my LPD results were compared, only the subtypes C1, C4, and C5 displayed a clear match with my LPD groups (Fig. 5.41). LPD_2 shared samples with 49% of the samples forming C4, while LPD_4 had 25% of them allocated. C5 shared 62% of its samples with LPD_3. LPD_7 was allocated to 67% of the samples in C1. These patterns were reaffirmed in the Pearson correlation analysis between Ellis' groups and my groups (Fig. 5.42): LPD_2 and LPD_4 showed a positive significant correlation with C5, and likewise, LPD_3 did with C5, and LPD_7 with C1. LPD_7 also showed a significant correlation with C2, but was discarded due to the small sample size of C2.

5.3.4 Lung cancer

The results for lung cancer are divided into LUAD and LUSC initially.

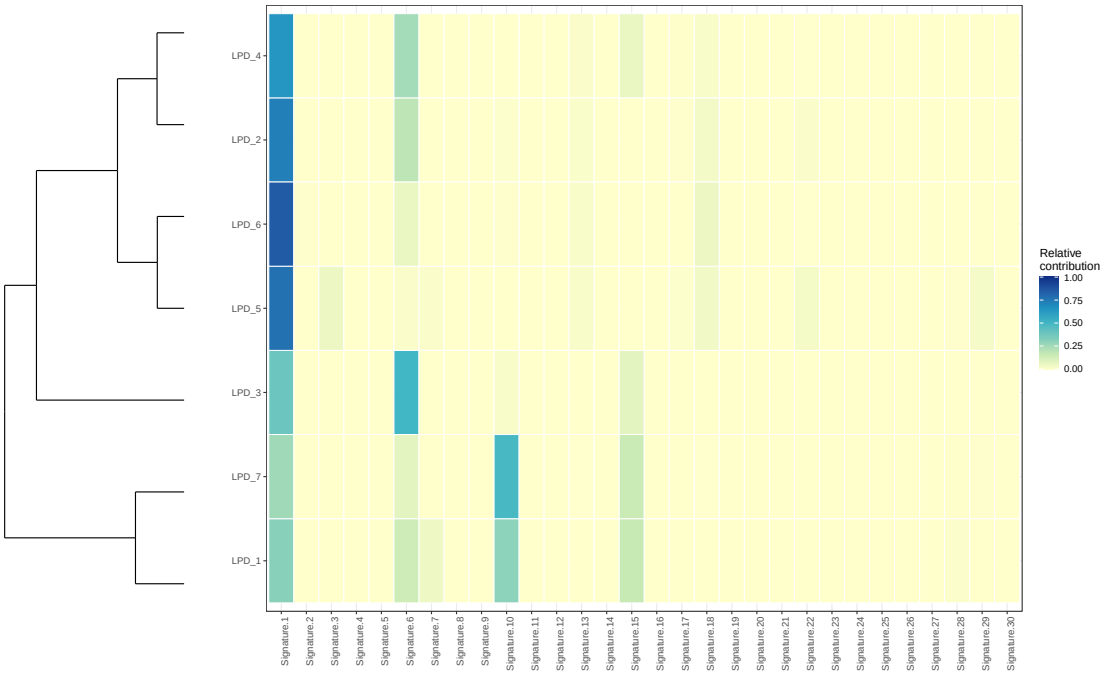

Figure 5.37: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in COAD. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

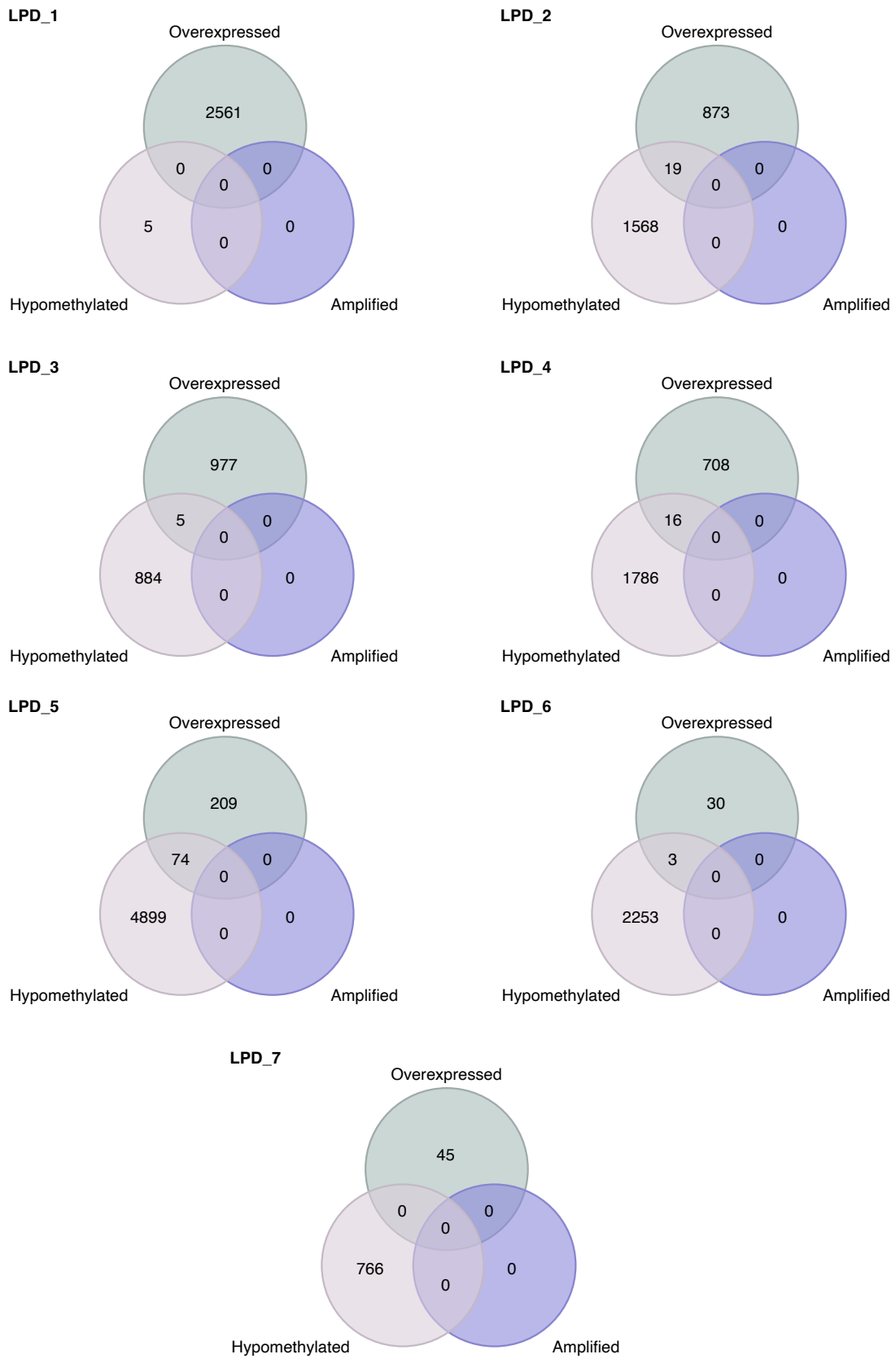

Figure 5.38: Venn diagram displaying the overlaps between three categories in genes in COAD for each LPD group: overexpressed genes, hypomethylated genes, and amplified genes. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit overexpression, hypomethylation, and amplification simultaneously.

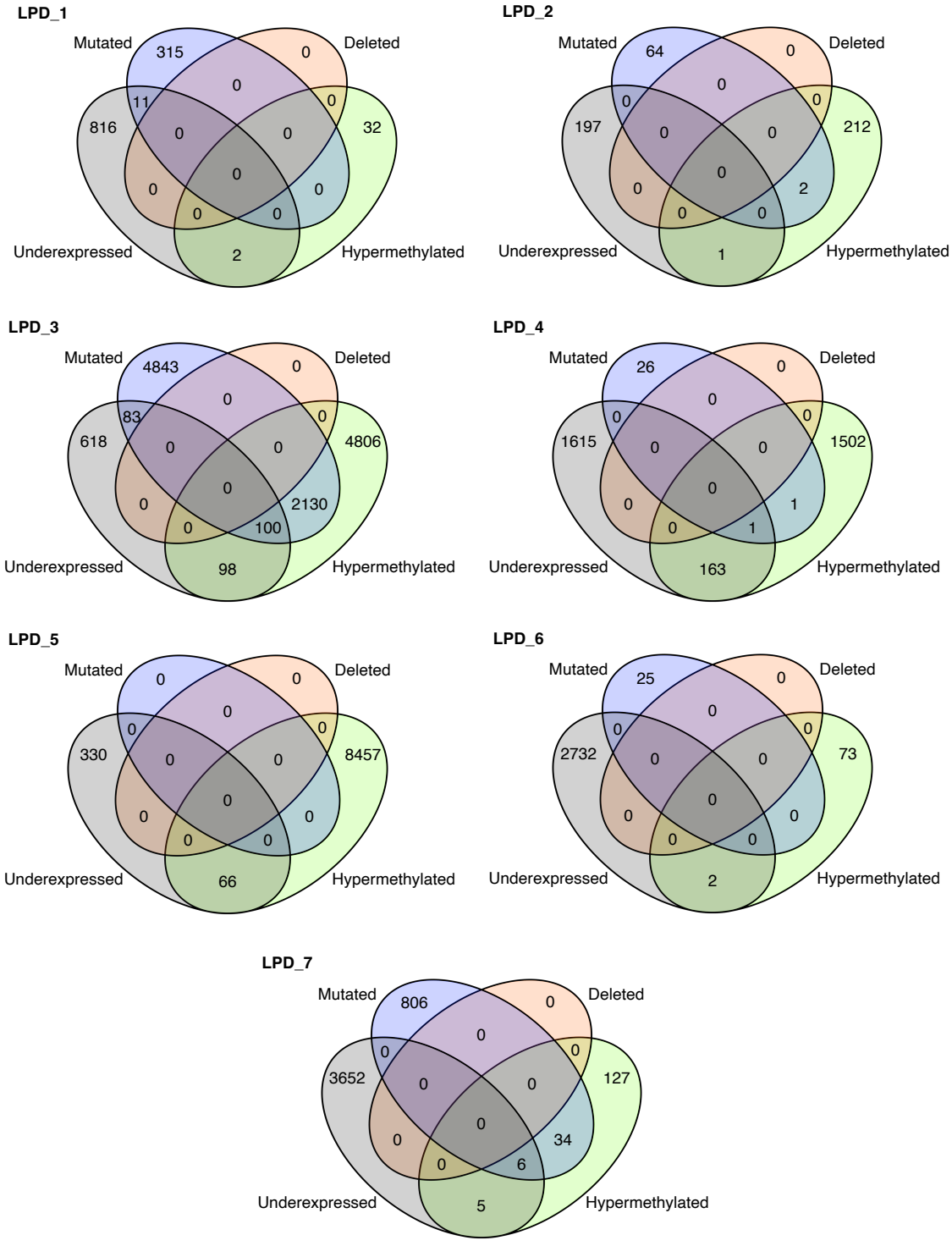

Figure 5.39: Venn diagram displaying the overlaps between four categories in genes in COAD for each LPD group: underexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit underexpression, hypermethylation, mutations and deletions.

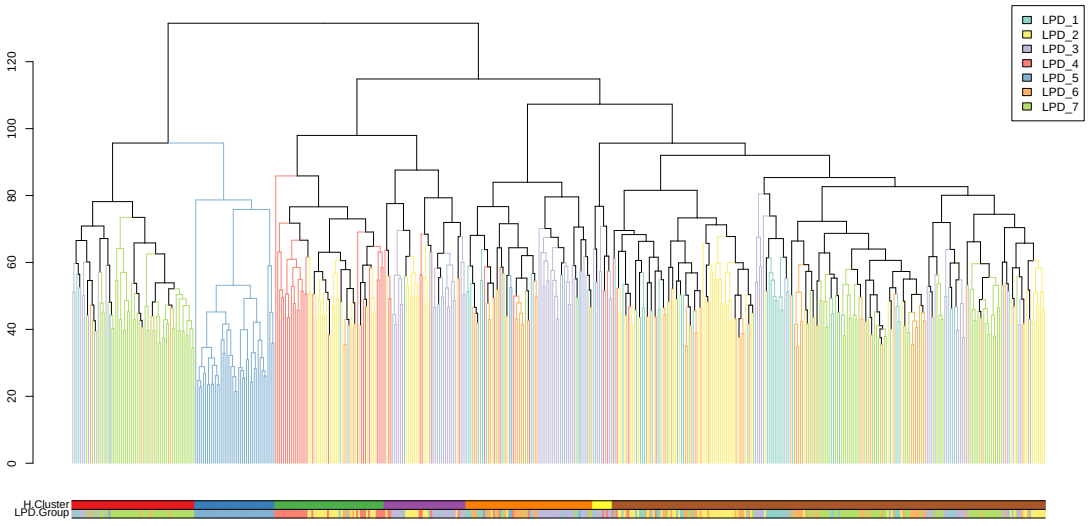

Figure 5.40: Dendrogram showing the sorting of the COAD samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into seven groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

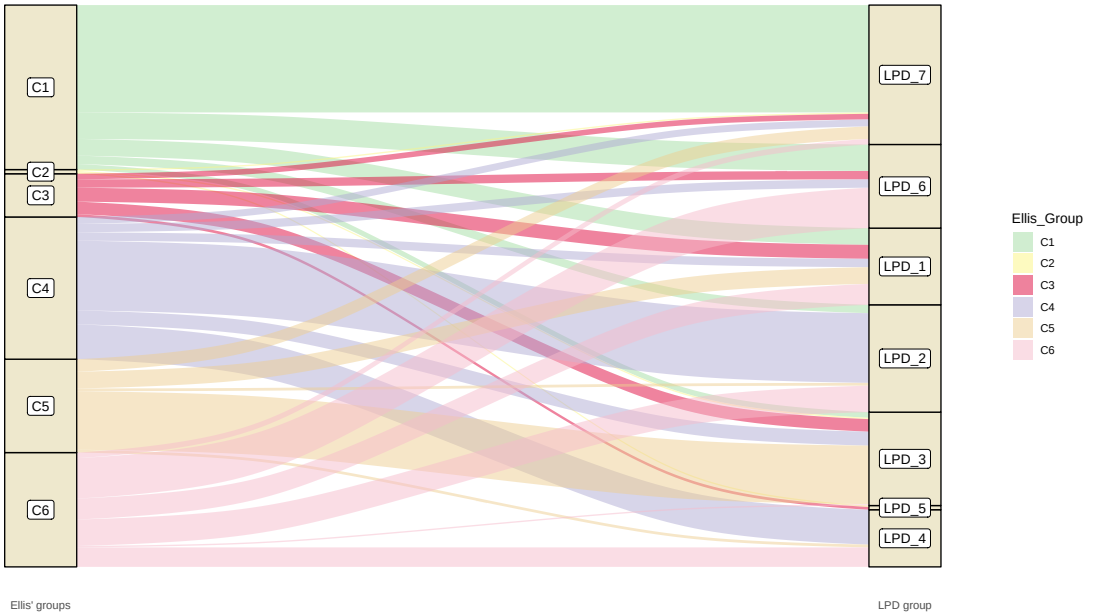

Figure 5.41: Alluvial plot showcasing a comparison between the sample assignments using Ellis' approach (displayed on the left) and the LPD approach described in this thesis (displayed on the right). Ellis' C3 is the Pericol subtype.

Exploring the LPD output for LUAD

A total 572 samples from 515 different patients were studied. Seven LPD groups were found optimal, which were termed as LPD_1 ($n = 84$, 14.63%), LPD_2 ($n = 60$, 10.45%), LPD_3 ($n = 70$, 12.19%), LPD_4 ($n = 89$, 15.5%), LPD_5 ($n = 70$, 12.19%), LPD_6 ($n = 129$, 22.47%), and LPD_7 ($n = 72$, 12.54%) (Fig 5.43). The sample distribution into LPD groups was independent of TSS ($P = 0.34$; Chi-quared test), although there was significant overrepresentation of normal tissue samples for LPD_3 ($n_{healthy} = 59$, 84.39%, $P = 2.54x10^{-100}$). When comparing the mean gamma values for each group, LPD_3 showed a robust assignment, while LPD_4 and LPD_6 displayed a shared distribution (Fig 5.44).

Clinicopathologic characteristics of the clusters in LUAD

Table 5.13 shows the clinicopathologic characteristics of the tumour samples in LUAD. No difference due to age ($P = 0.61$; Chi-squared test), race ($P = 0.41$; Chi-squared test) and pathological stage was detected ($P = 0.11$; Chi-squared test). LPD_6 had a higher proportion of female patients than the other groups ($P = 0.00049$; Chi-squared test). When the groups were compared, a significant event-free survival ($P = 0.0026$; log-rank test) was identified, with LPD_5 ($P = 0.0019$; log-rank test LPD_5 vs the rest) showing the poorest prognosis (Fig 5.45).

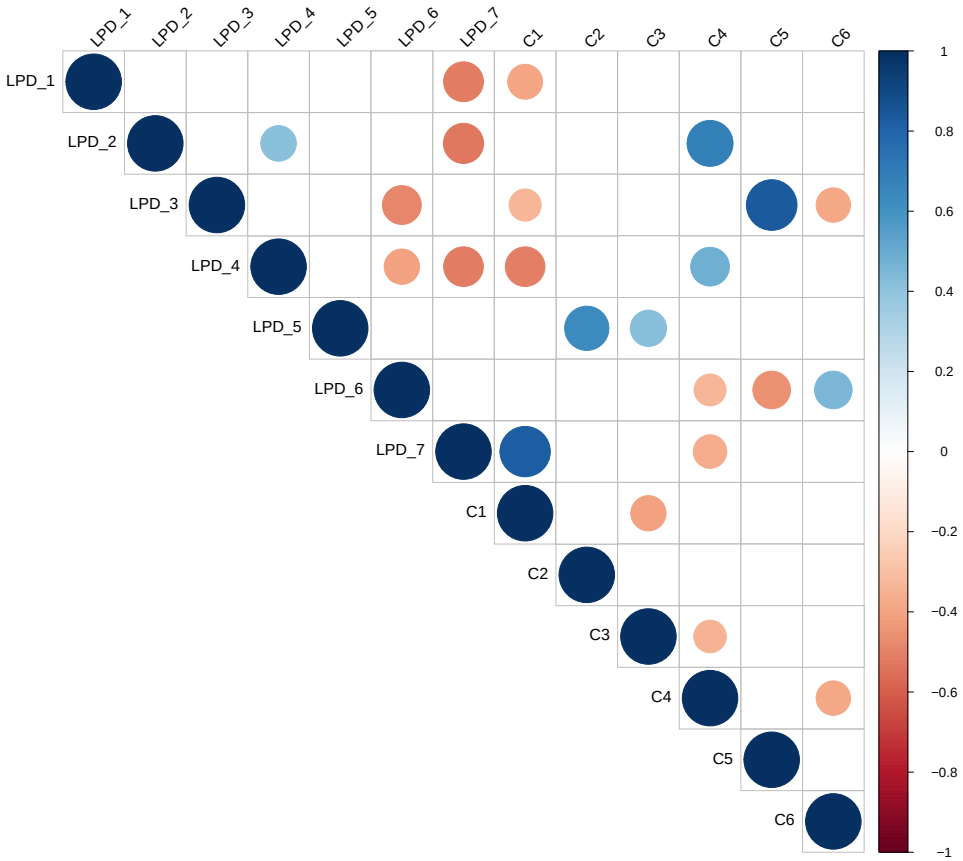

Figure 5.42: Pearson correlation matrix between Ellis' detected subtypes of COAD and the LPD groups. Subtype C3 is Pericol.

Table 5.13: Clinicopathologic features of the detected subtypes for lung adenocarcinoma. chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	P-value
Age (years; mean (sd))	66.7 (10.2)	67.2 (10.6)	64.6 (11.8)	66.3 (10.6)	68 (9.04)	66.5 (10.2)	67.2 (9.83)	66.1 (10.4)	0.61
Race									
Asian	7	0	1	1	3	2	0	0	
Black or african american	57	9	5	5	7	7	14	10	
White	443	60	48	63	67	52	102	51	
American indian or alaska native	1	0	0	0	0	0	0	1	0.41
Gender									
Female	310	34	22	42	53	24	100	35	
Male	264	50	38	28	36	46	29	37	0.0004
Pathological stage									
Stage I	274	52	29	6	50	29	76	32	
Stage II	122	19	15	0	19	19	29	21	
Stage III	84	8	9	2	13	18	19	15	
Stage IV	26	3	6	2	6	3	4	2	0.11

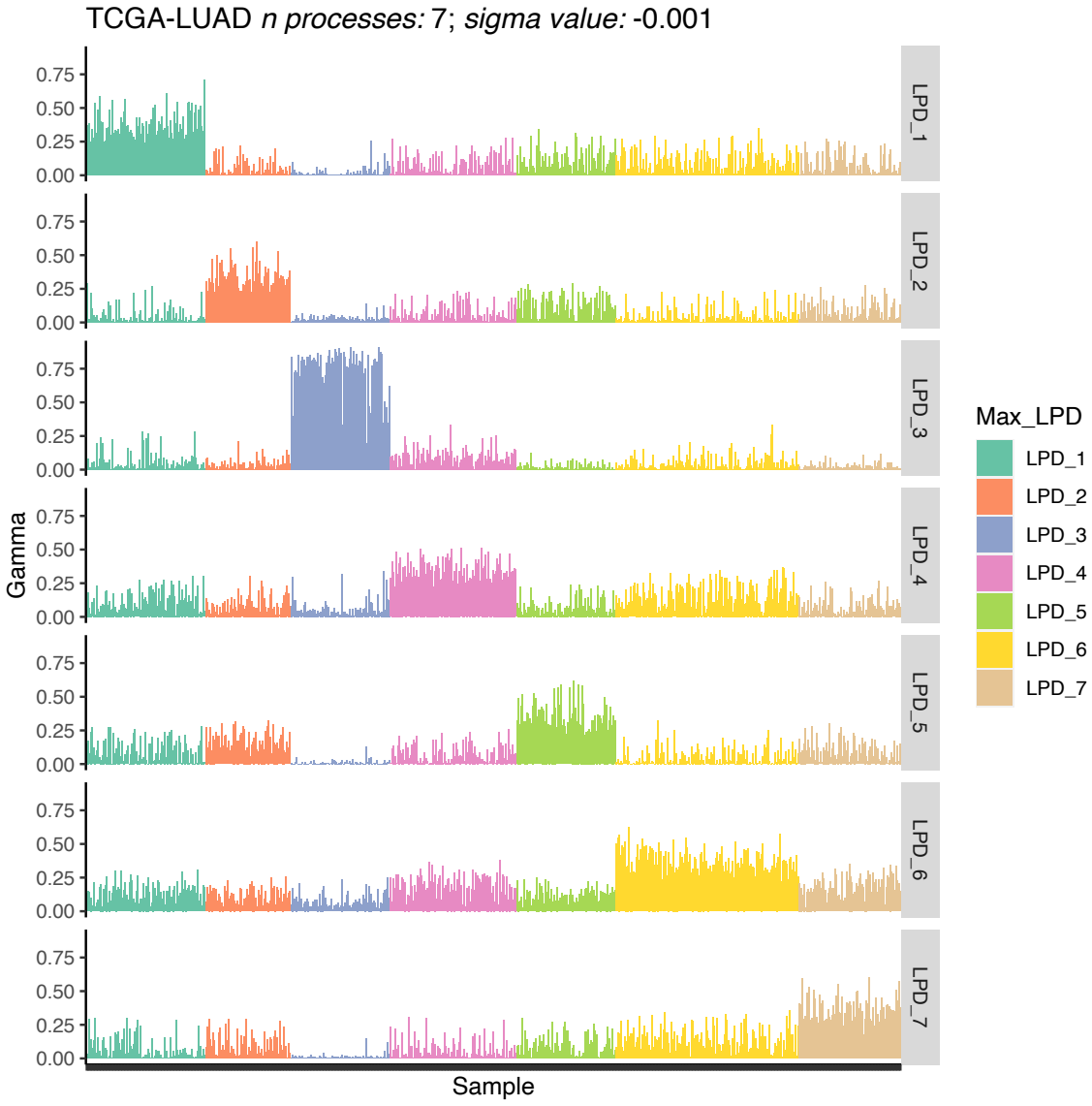

Figure 5.43: Gamma values of all samples for each detected LPD process in lung adenocarcinoma. A total of 7 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_7. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD signature is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

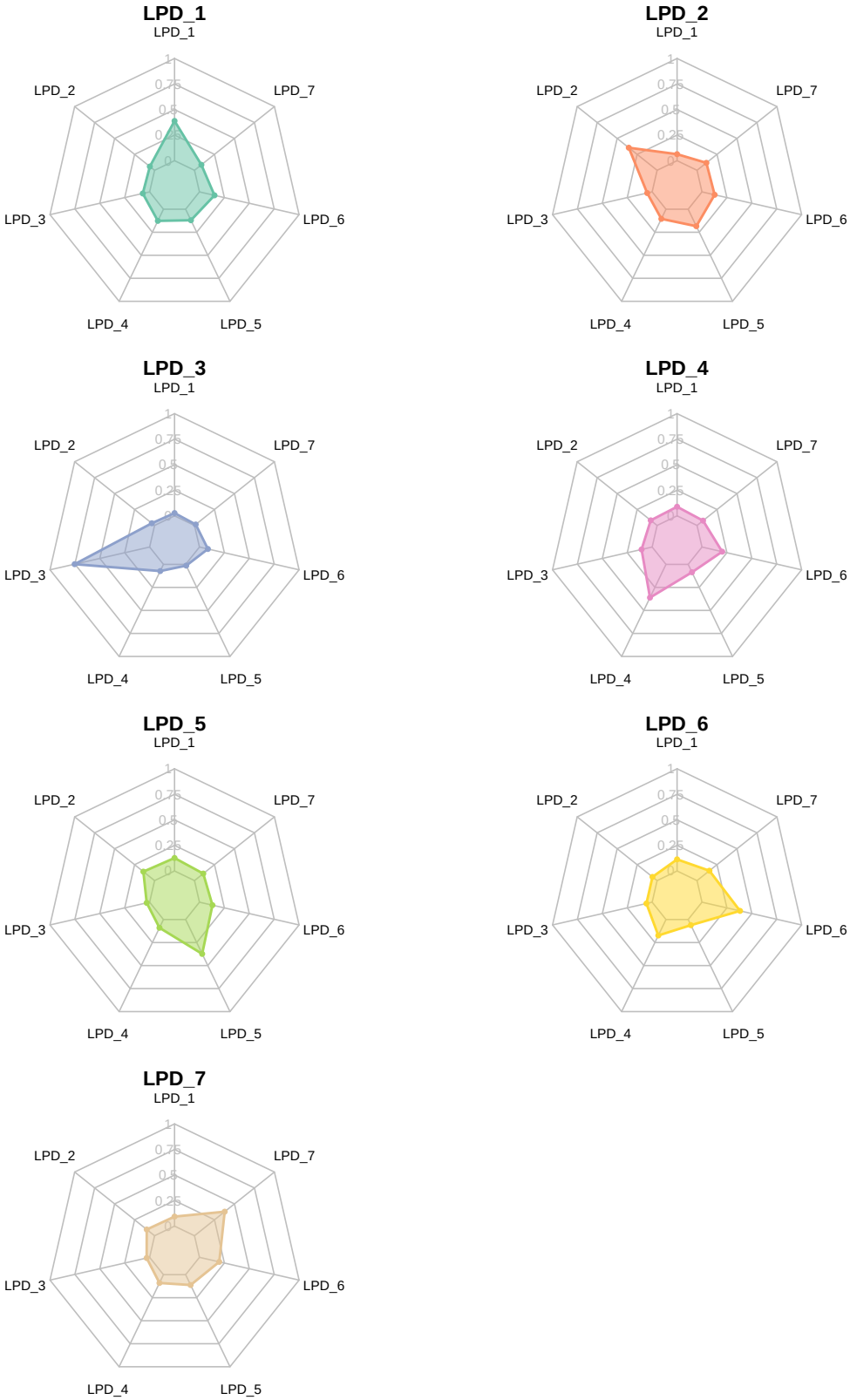

Figure 5.44: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in lung adenocarcinoma. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

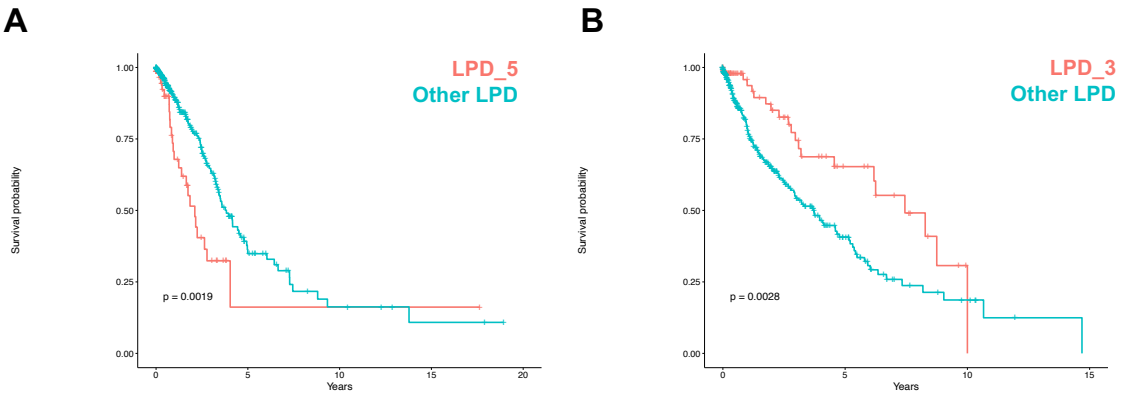

Figure 5.45: Kaplan-Meier estimator curve for (A) samples allocated in LPD_5 in LUAD (red) in comparison to the other samples (blue), and (B) samples allocated in LPD_3 in LUSC (red) in comparison to the other samples (blue). Log-rank test was conducted across the survival curves and the corresponding p-value is displayed.

Identification of differentially expressed genes in LUAD

Across the seven LPD groups detected in LUAD, a total of 8123 DEGs were identified (Table 5.14; median across groups = 765; IQR = 787). The top overexpressed and underexpressed genes ranked by *log2 fold change* are shown in Table 5.15. LPD_1 was the only group defined by a majority of overexpressed DEGs, whereas the other groups exhibited a large underexpression with ratios less than 0.5. Focusing on the known driver genes, the gene *ALB* was identified as DEG in all groups except LPD_6. This gene was strongly overexpressed in LPD_5, and strongly underexpressed in LPD_2, LPD_3, and LPD_4 (Fig. 5.48). KEGG enrichment analysis revealed that the neuroactive ligand-receptor interaction pathway was affected by DEGs in all the LPD groups (Fig. 5.46.A). GO, on the other hand, had no affected process that was shared by all groups (Fig. 5.47.A). LPD_1 displayed enrichment in biological processes related to unlimited replication (Figure 5.7).

Table 5.15: The five genes more overexpressed (log2FoldChange > 1) and underexpressed (log2FoldChange < -1) for each LPD group in lung adenocarcinoma. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status
LPD_1		
ENSG00000259001	7.554622	Overexpressed
RNU5B-1	6.959635	Overexpressed
SCARNA5	6.947593	Overexpressed
H4C6	6.904643	Overexpressed
RNY1	6.317706	Overexpressed
CALCA	-5.455774	Underexpressed
ASCL1	-5.159458	Underexpressed
OTX2	-4.175863	Underexpressed

LINC00676	-3.900353	Underexpressed
PAGE1	-3.861308	Underexpressed
LPD_2		
LINC01733	3.648242	Overexpressed
EZHIP	3.335493	Overexpressed
SCARNA10	2.606762	Overexpressed
PAGE5	2.464287	Overexpressed
LINC01287	2.215750	Overexpressed
REG4	-5.629096	Underexpressed
CALCA	-4.617435	Underexpressed
MUC17	-4.226491	Underexpressed
PRB4	-3.863112	Underexpressed
ALB	-3.736734	Underexpressed
LPD_3		
CYP11B1	6.249645	Overexpressed
HSD3B2	3.036680	Overexpressed
PANTR1	2.347195	Overexpressed
PSG1	2.140873	Overexpressed
VAX1	2.125592	Overexpressed
PAGE2	-5.353004	Underexpressed
TRIM48	-4.351509	Underexpressed
REG4	-4.332929	Underexpressed
TUBA3C	-4.322058	Underexpressed
ALB	-4.079147	Underexpressed
LPD_4		
RPTN	2.553057	Overexpressed
PRB1	2.110110	Overexpressed
DSG3	2.050485	Overexpressed
ENSG00000267706	1.805970	Overexpressed
GDPD2	1.673839	Overexpressed
ALB	-4.175284	Underexpressed
PASD1	-4.004412	Underexpressed
DEFA5	-3.926615	Underexpressed
ZIC1	-3.922928	Underexpressed
MAGEA4	-3.916776	Underexpressed
LPD_5		
SPAG11B	4.679144	Overexpressed
ALB	4.030163	Overexpressed
TNMD	2.775628	Overexpressed
CGA	2.746014	Overexpressed
LINC02582	2.580005	Overexpressed
REG4	-4.870428	Underexpressed
MAGEA10	-4.789649	Underexpressed
MAGEA4-AS1	-4.508396	Underexpressed
DLK1	-4.352166	Underexpressed
MARCHF11	-4.182055	Underexpressed

LPD_6		
LINC02672	2.315253	Overexpressed
WFDC5	2.157474	Overexpressed
ENSG00000261166	1.999574	Overexpressed
ENSG00000231317	1.851959	Overexpressed
ATOH1	1.845386	Overexpressed
SSX1	-4.227539	Underexpressed
PRB4	-4.050408	Underexpressed
G6PC	-3.590065	Underexpressed
PRB1	-3.559352	Underexpressed
H4C6	-3.431266	Underexpressed
LPD_7		
APOC3	2.980410	Overexpressed
CDH16	2.428384	Overexpressed
AACSP1	2.296290	Overexpressed
ANKRD20A19P	2.266112	Overexpressed
FAM166A	2.203936	Overexpressed
REG4	-5.200540	Underexpressed
MAGEA9B	-4.872916	Underexpressed
DSCR8	-4.512208	Underexpressed
DSCR4	-4.493593	Underexpressed
UPK1B	-4.413044	Underexpressed

Identification of differentially methylated genes in LUAD

Most of the LPD groups had less than a dozen of DMGs, except LPD_3 that accumulated 209, accounting for the 88% of the total (Table 5.14; median across groups = 4; IQR = 6). In terms of driver genes, LPD_3 was defined as a strong hypermethylation of *RUNX1*, *SPTA1*, and *USP9X* (Fig. 5.48). KEGG enrichment analysis detected four pathways significantly enriched by DMGs for LPD_3 (Fig. 5.46.B), while GO returned seven processes (Fig. 5.47.B).

Identification of genes affected by single nucleotide variants in LUAD

A median of 1534 genes per LPD group were enriched or depleted in SNVs in comparison to other groups (IQR = 447). LPD_3 exhibited a relatively low number of genes affected by mutations (less than the 1% of the total) (Table 5.14). LPD_2 was significantly overmutated in comparison to the other groups, while LPD_1 and LPD_4 exhibited a significant undermutation. About the driver genes, a plethora of them were detected as differentially affected across groups (Fig. 5.48). LPD_1, LPD_4 and LPD_7 had a majority of depleted for SNVs of the driver genes, whereas LPD_2 and LPD_7 appeared to have a predisposition for overmutation. LPD_3 showed strong depletion of mutations in *TP53*, and *EGFR* when compared with other groups. The KEGG enrichment analysis returned several pathways, but three of them were affected in all groups except LPD_3: herpes virus infection, neuroactive ligand-receptor interaction and olfactory transduction (Fig. 5.46.C). LPD_1 and LPD_3 showed similar profiles with overmutation of neurodegeneration, olfactory transduction and neuroactive ligand-receptor interaction, and undermutation of herpes

Table 5.14: Gene counts for various categories in LUAD. These include the number of genes exhibiting significant differential expression and differential methylation, the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively), and the number of genes with significantly higher or lower frequency of CNV (referred to as overimpacted and underimpacted, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7
n health tissue samples	0	0	59	0	0	0	0
DEGs							
Upregulated	1480	120	99	62	61	42	546
Downregulated	939	435	628	703	1267	636	1105
Total	2419	555	727	765	1328	678	1651
Ratio	1.58	0.3	0.16	0.09	0.05	0.08	0.49
DMGs							
Hypermethylated	0	0	53	2	2	12	3
Hypomethylated	3	2	156	1	2	0	2
Total	3	2	209	3	4	12	5
Ratio	0	0	0.33	2	1	1	1.5
Mutated							
Overmutated	320	1576	0	330	778	1028	650
Undermutated	961	616	32	944	876	767	884
Total	1281	2192	32	1274	1654	1795	1534
Ratio	0.33	2.56	0	0.35	0.89	1.34	0.74
Affected by CNV							
Overimpacted	182	244	56	80	135	92	153
Underimpacted	2	2	1	0	1	0	0
Total	184	246	57	80	136	92	153
Ratio	91	122	56	1	135	1	1

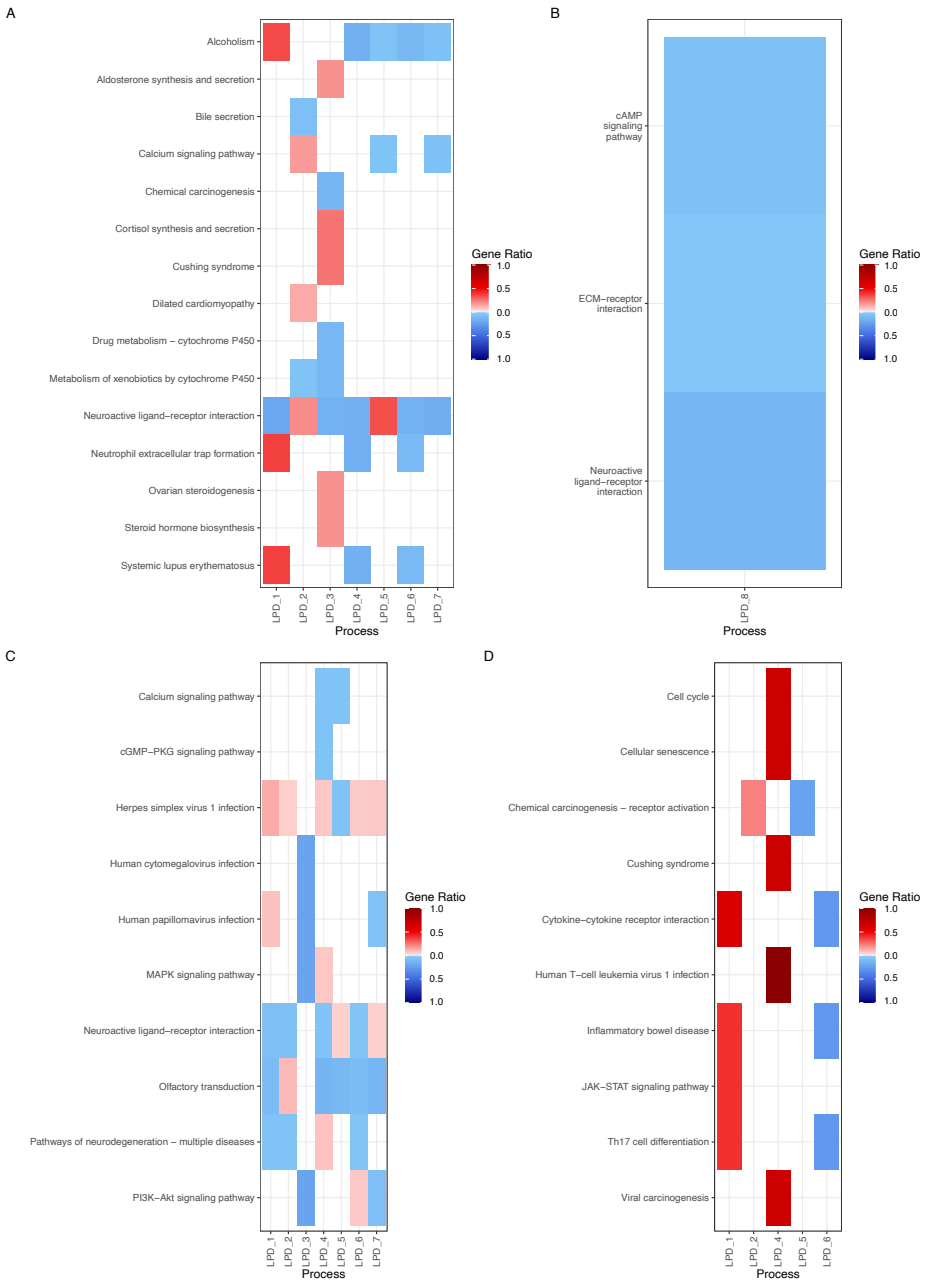

Figure 5.46: Biological pathways associated with different categories in lung adenocarcinoma determined using KEGG enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, biological pathways associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the pathways associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. (D) Copy number variations, the pathways associated to genes with significant higher frequency of being affected by copy number variations are represented in red, while the opposite is represented in blue. The complete list of associated biological pathways is available in Supplementary Material B.

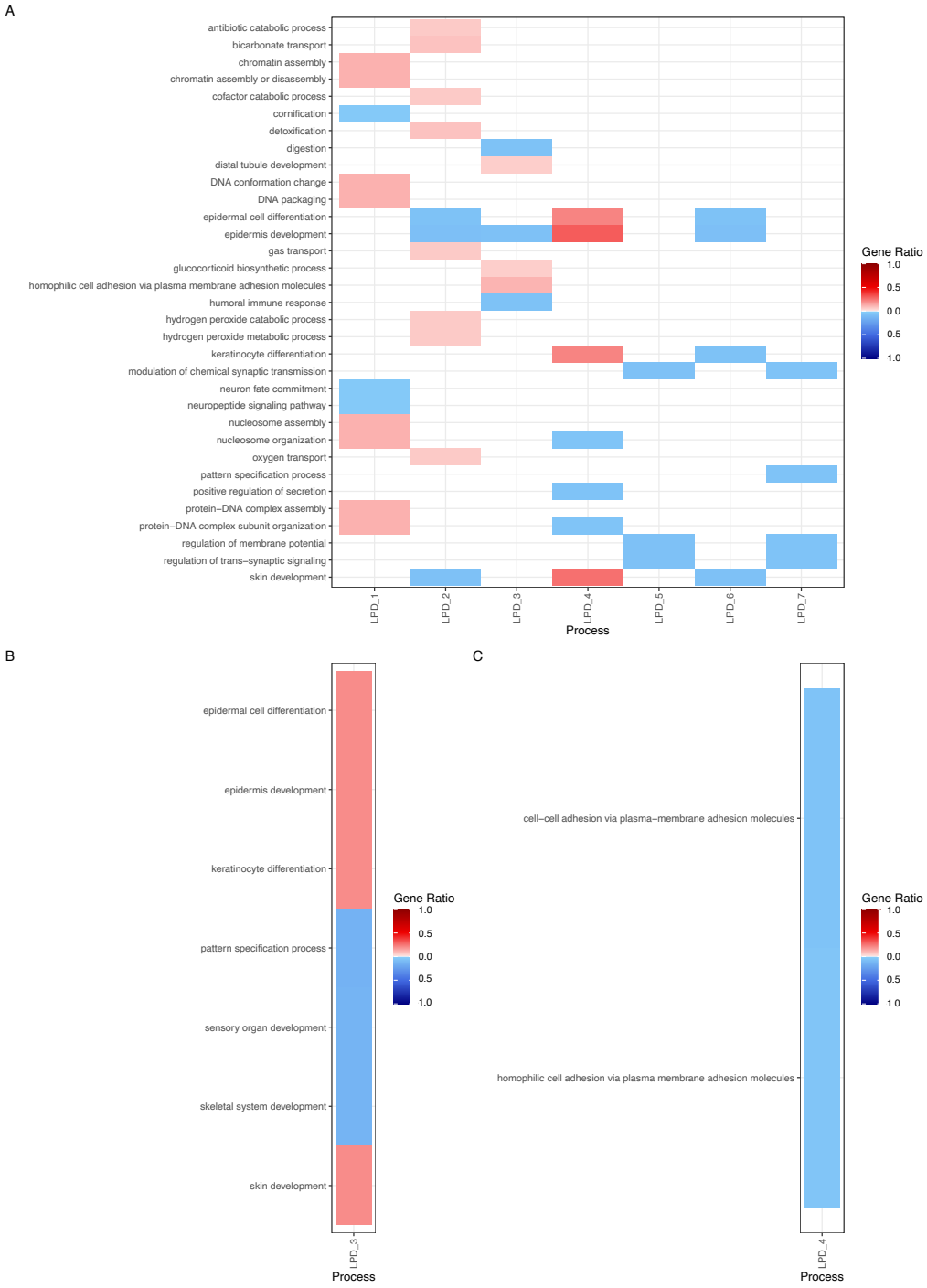

Figure 5.47: Biological processes associated with different categories in LUAD determined using GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant processes with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while processes linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the processes associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological processes is available in Supplementary Material B.

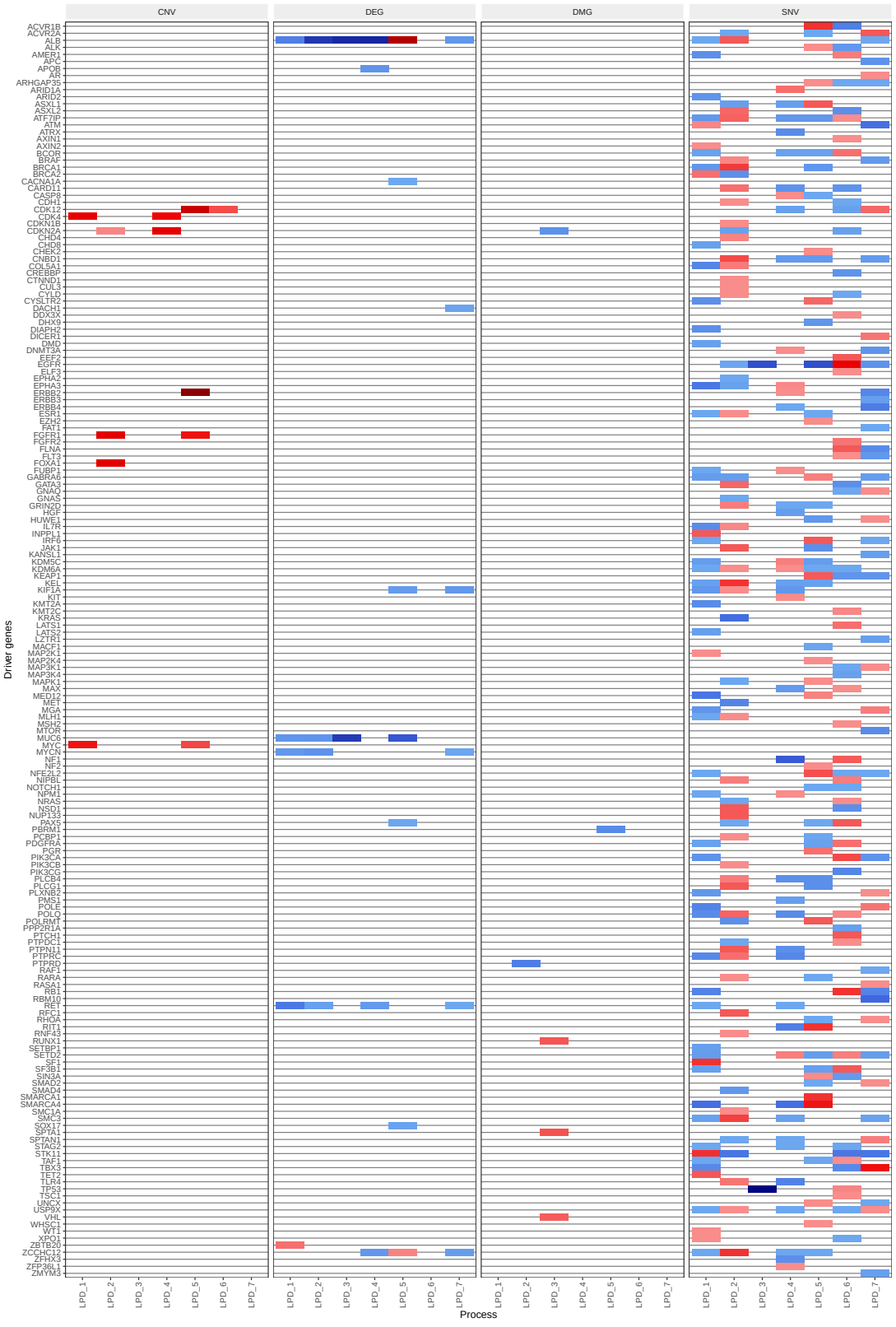

Figure 5.48: Heatmap showing the presence of driver genes across different categories in lung adenocarcinoma, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV) and copy number variants (CNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV and CNV are represented in red, while genes with opposite characteristics are depicted in blue.

infection. GO analysis only detected enriched processes in LPD_4 consisting in overmutations of cell adhesion processes (Fig. 5.47.C). When comparing the SNP type, variant type, and variant class frequency across LPD groups, no differences were observed (P>0.05; chi-squared text; Fig. 5.49) The bulk of detected SNVs were missense mutations caused by the point replacement of cytosine to adenine. LPD_6 displayed a significant enrichment in mutations in the *EGFR* gene, whilst no associations were found for the remaining LPD groups (Fig. 5.50.A).

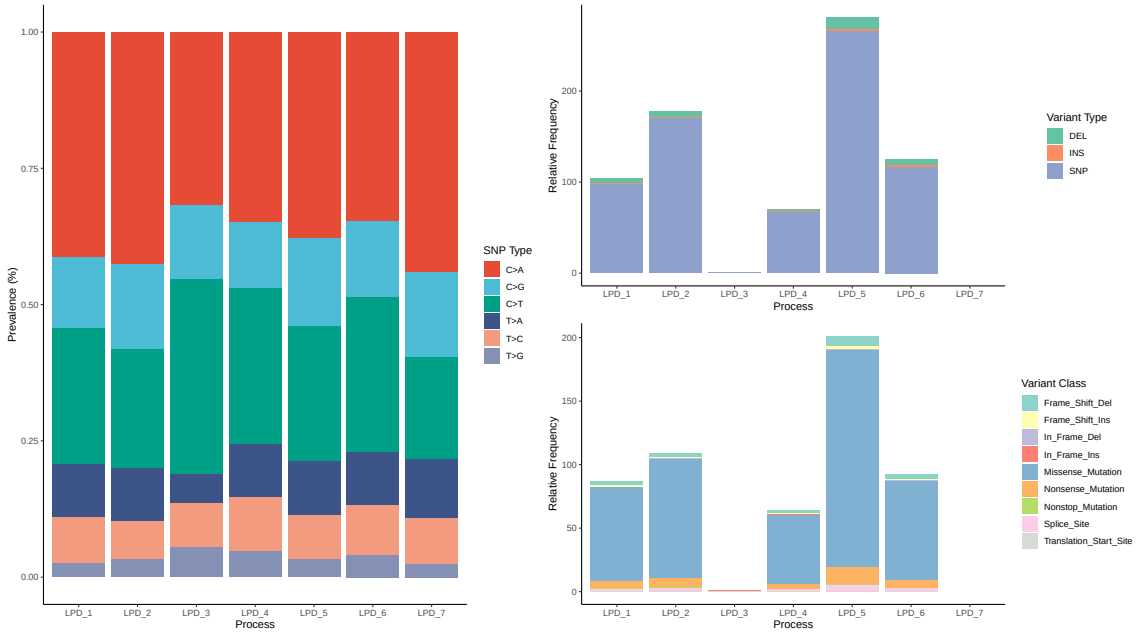

Figure 5.49: Detected single nucleotide variants (SNVs) within each LPD group for LUAD. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

The COSMIC profiles of the LPD groups revealed that, with the exception of LPD_3, all groups had a strong contribution from signature 4, whereas LPD_3 was the only group with a relationship to signature 1 (Fig. 5.51).

Identification of genes impacted by copy number variations in LUAD

A median of 56 genes per LPD group were enriched or depleted in CNVs in comparison to other groups (IQR = 106). The number of genes underimpacted by CNVs was very low across all groups, with LPD_4, LPD_5, and LPD_6 having none (Table 5.14). As a result, the ratio for all groups indicated an overimpact. This was also visible in the KEGG enrichment analysis, in which only LPD_5 and LPD_6 had pathways enriched by underimpacted genes (Fig. 5.46.D). GO enrichment analysis yielded no significant results. Seven driver genes were identified as CNV-impacted genes, with *ERBB2* and *CDK12* having the largest effect size in LPD_5 (Fig. 5.48). None of the LPD groups showed associations with the gene *STK11* (Fig. 5.50.B).

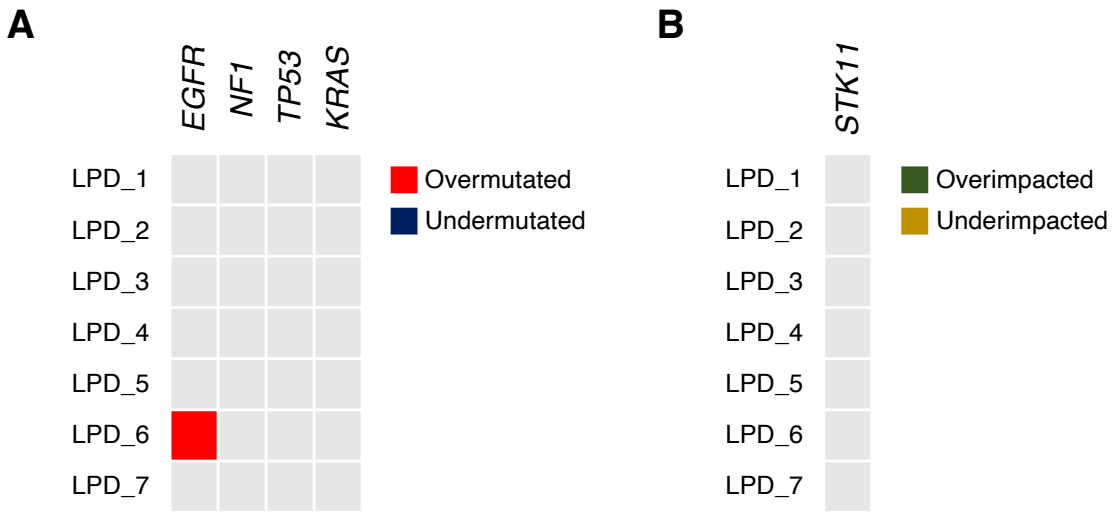

Figure 5.50: Heatmaps illustrating the mutational status of multiple genes in the lung adenocarcinoma (LUAD) dataset across different LPD groups. In the first heatmap (A), each column represents one of the four genes (*EGFR*, *NF1*, *TP53*, and *KRAS*), while each row corresponds to a distinct LPD group. The colour red indicates a significant overmutation of a gene in a specific LPD group, meaning that the gene is more frequently mutated in that group. Conversely, blue indicates a significant undermutation, indicating that the gene is less frequently mutated in the group. In the second heatmap (B), the status of the *STK11* gene in terms of copy number variations (CNV) is displayed. Each column represents the *STK11* gene, and each row represents an LPD group. Green indicates a significant overimpact, suggesting that the gene is more frequently affected by CNV in that group. Yellow indicates a significant underimpact, implying that the gene is less frequently affected by CNV in the group. However, none of the LPD groups showed a significant association in this particular case, indicating that the *STK11* gene did not exhibit distinct CNV patterns among the LPD groups in the LUAD dataset.

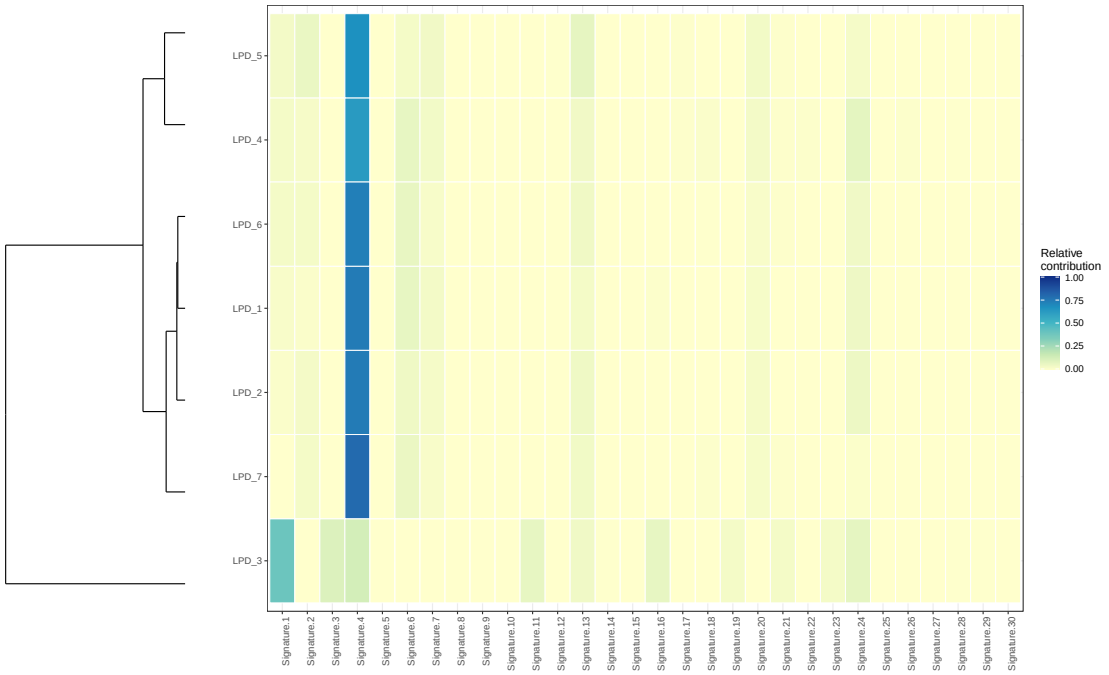

Figure 5.51: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in LUAD. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

Functional analysis of differentially expressed genes in LUAD

Matches between overexpressed, hypomethylated and amplified genes are shown in figure 5.52, while matches between underexpressed, hypermethylated, deleted and mutated genes are shown in figure 5.53. For the overexpressed genes, LPD_3 showed an enrichment of matches with 4 genes hypomethylated ($P = 0.004$; Chi-squared analysis). For underexpressed genes, LPD_2 was enriched with 20 matches, whereas LPD_1 was depleted ($P < 0.0001$; Chi-squared analysis). The complete list of matched genes is available in Supplementary Material B.

Comparison of the LPD output in LUAD with Eucledian hierarchical clustering

The distribution of the LPD groups were spread across the dendrogram, with the exception of LPD_3 that was clearly differentiated from the other groups (Fig. 5.54). A portion of LPD_2 samples was also separated from the others groups and accounting for the totality of samples assigned to the hierarchical brown cluster. Samples from LPD_4, LPD_6 and LPD_7 tended to be located in the same hierarchical groups.

Exploring the LPD output for LUSC

In the case of LUSC, 550 samples from 501 patients were examined. Six LPD groups were found optimal, named as LPD_1 ($n = 80$, 14.54%), LPD_2 ($n = 98$, 17.81%), LPD_3 ($n = 98$, 17.81%), LPD_4 ($n = 113$, 20.54%), LPD_5 ($n = 45$, 8.18%), LPD_6 ($n = 116$, 21.09%) (Fig 5.55). The sample distribution into LPD groups was independent of TSS ($P = 0.14$; Chi-squared test), although there was significant overrepresentation of healthy tissue samples for LPD_1 ($n_{healthy} = 49$, 70%, $P = 1.31x10^{-68}$; Chi-squared test). According to the mean gamma values, the assignments were relatively robust, especially in LPD_1 (Fig 5.56). LPD_2, LPD_4, and LPD_6 all exhibited a shared assignment.

Clinicopathologic characteristics of the clusters in LUSC

The clinicopathologic characteristic of LUSC are collected in Table 5.16. Patients classified as LPD_2 were older, while the ones in LPD_3 were younger ($P = 0.008$; Chi-squared test). LPD_3 had a higher proportion of male patients and LPD_6 of female ones ($P = 0.00049$; Chi-squared test). No difference according to race ($P = 0.1$; Chi-squared test) and pathological stage ($P = 0.02797$; Chi-squared test) was detected. When the survival probability was compared, LPD_3 showed a worse prognosis than the other groups ($P = 0.0028$; log-rank test) (Fig 5.45).

Identification of differentially expressed genes in LUSC

A total of 5285 DEGs were detected, with a median across groups of 729 and an IQR of 382. Except for LPD_2, the ratios from all groups indicated a big underexpression preference. The top overexpressed and underexpressed genes ranked by *log2 fold change* are shown in Table 5.18. Thirteen differentially expressed driver genes were detected across the six LPD groups, with LPD_2 exhibiting the strongest effect size for the overexpression of *ALB* and *APOB*(Fig. 5.59). As from the results from LUAD, the neuroactive ligand-receptor pathway was affected by DEGs in all the LPD groups with only LPD_2 showing it as enriched in overexpressed DEGs (Fig. 5.57.A). LPD_1 and LPD_4 showed the strongest

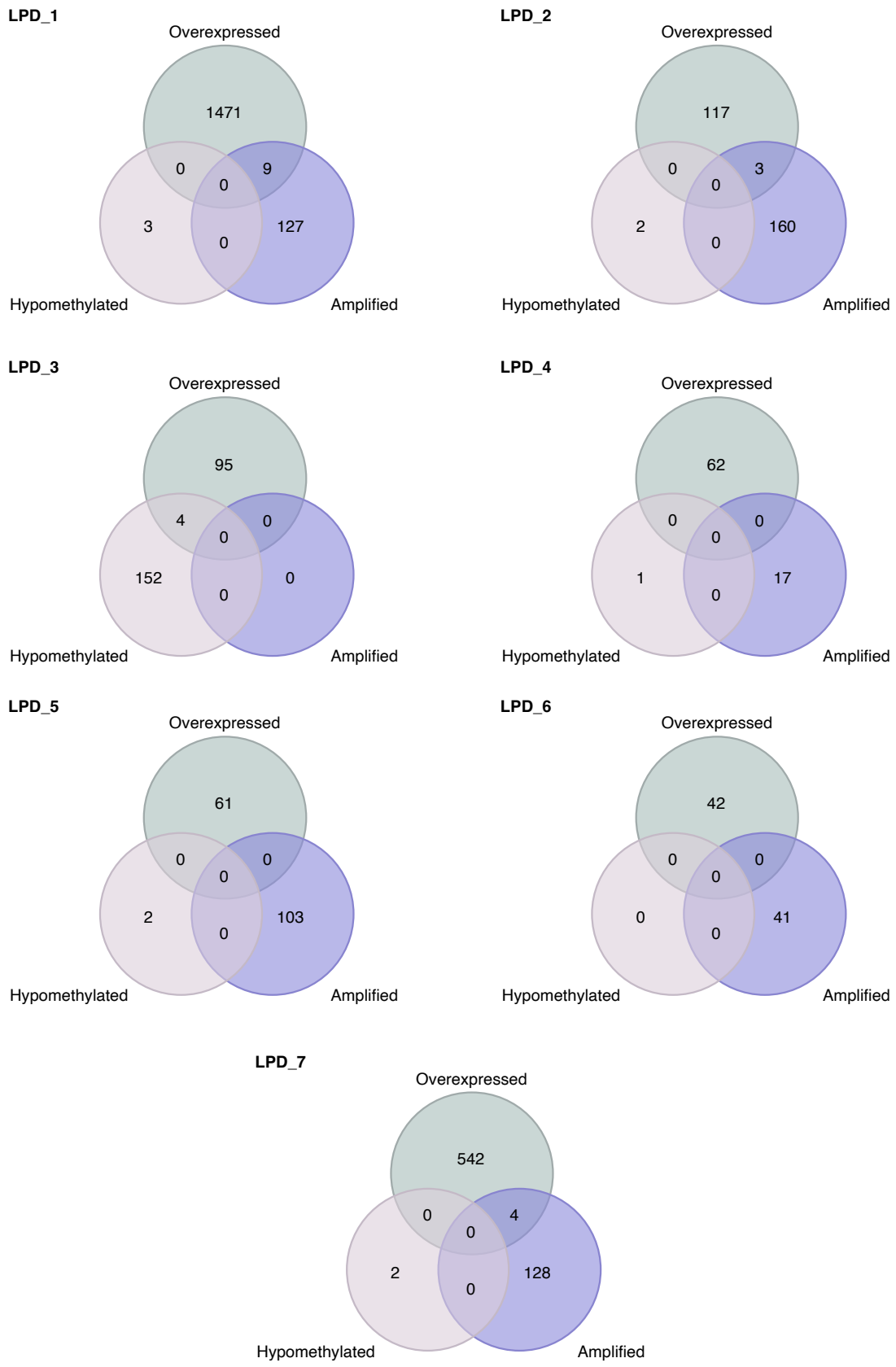

Figure 5.52: Venn diagram displaying the overlaps between three categories in genes in LUAD for each LPD group: overexpressed genes, hypomethylated genes, and amplified genes. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit overexpression, hypomethylation, and amplification simultaneously.

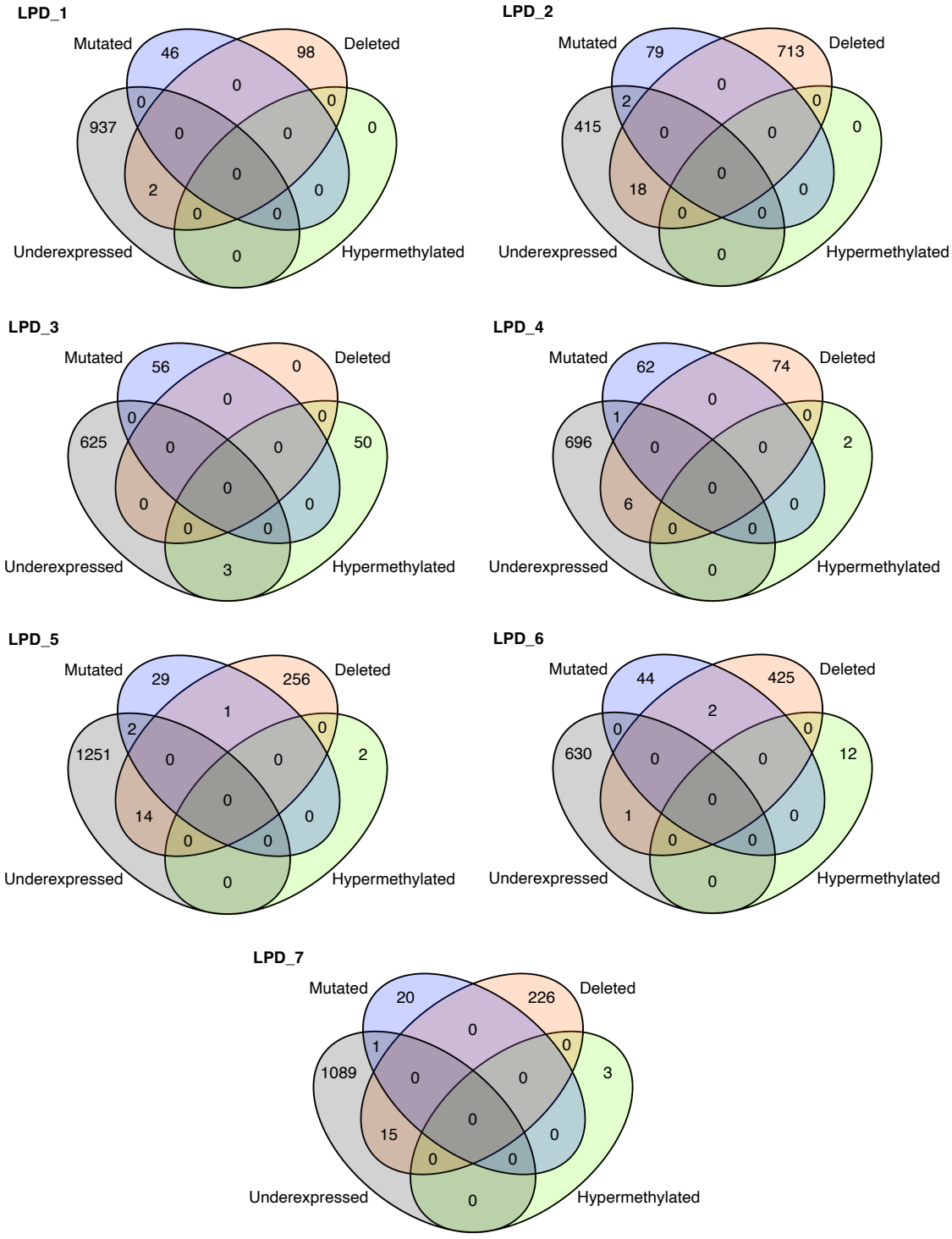

Figure 5.53: Venn diagram displaying the overlaps between four categories in genes in LUAD for each LPD group: underexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit underexpression, hypermethylation, mutations and deletions.

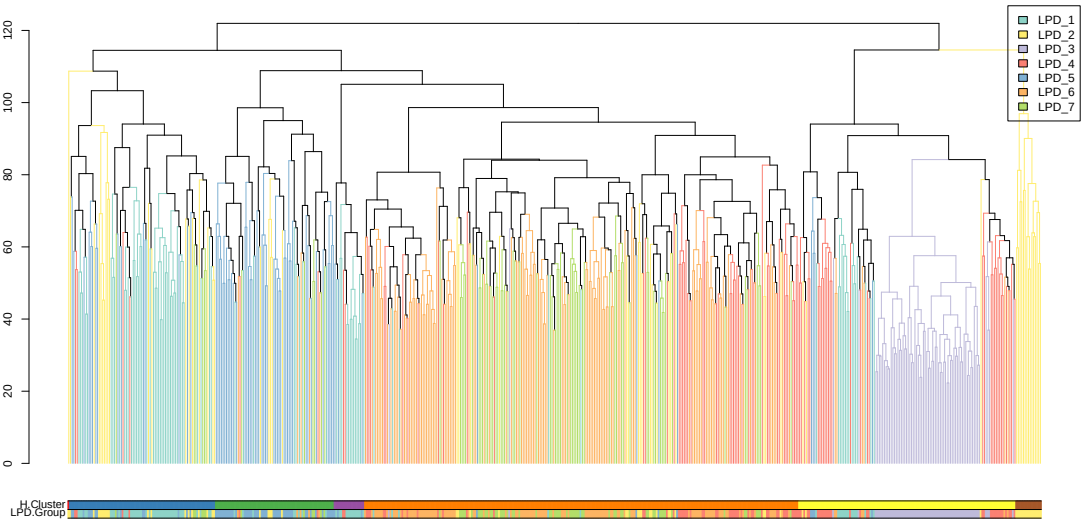

Figure 5.54: Dendrogram showing the sorting of the LUAD samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into seven groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

Table 5.16: Clinicopathologic features of the detected subtypes for lung squamous cell carcinoma. chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	P-value
Age (years; mean (sd))	68.8 (8.69)	68.9 (8.23)	71.1 (7.85)	66.5 (9.45)	69.4 (8.21)	67.1 (11.2)	68.8 (7.91)	0.008
Race								
Asian	9	1	4	1	2	0	1	
Black or african american	32	2	2	9	5	6	8	
White	391	68	68	66	75	34	80	0.1
Gender								
Female	144	26	23	14	23	12	46	
Male	406	54	75	84	90	33	70	0.0004
Pathological stage								
Stage I	244	14	55	34	56	19	66	
Stage II	162	11	23	47	39	15	27	
Stage III	84	4	17	16	15	11	21	
Stage IV	7	1	2	1	2	0	1	0.02

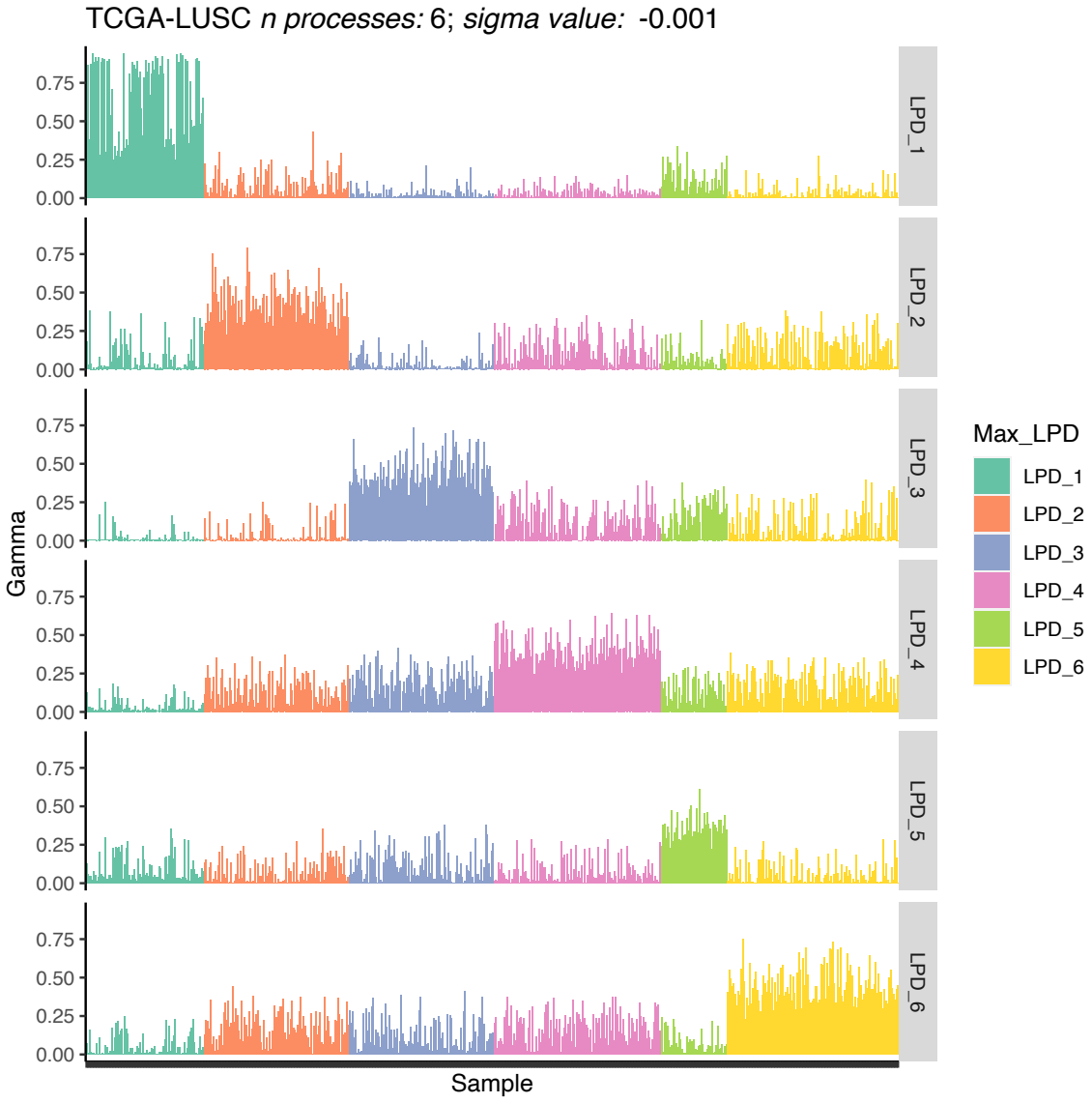

Figure 5.55: Gamma values of all samples for each detected LPD process in lung squamous cell carcinoma. A total of 6 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_6. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD signature is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

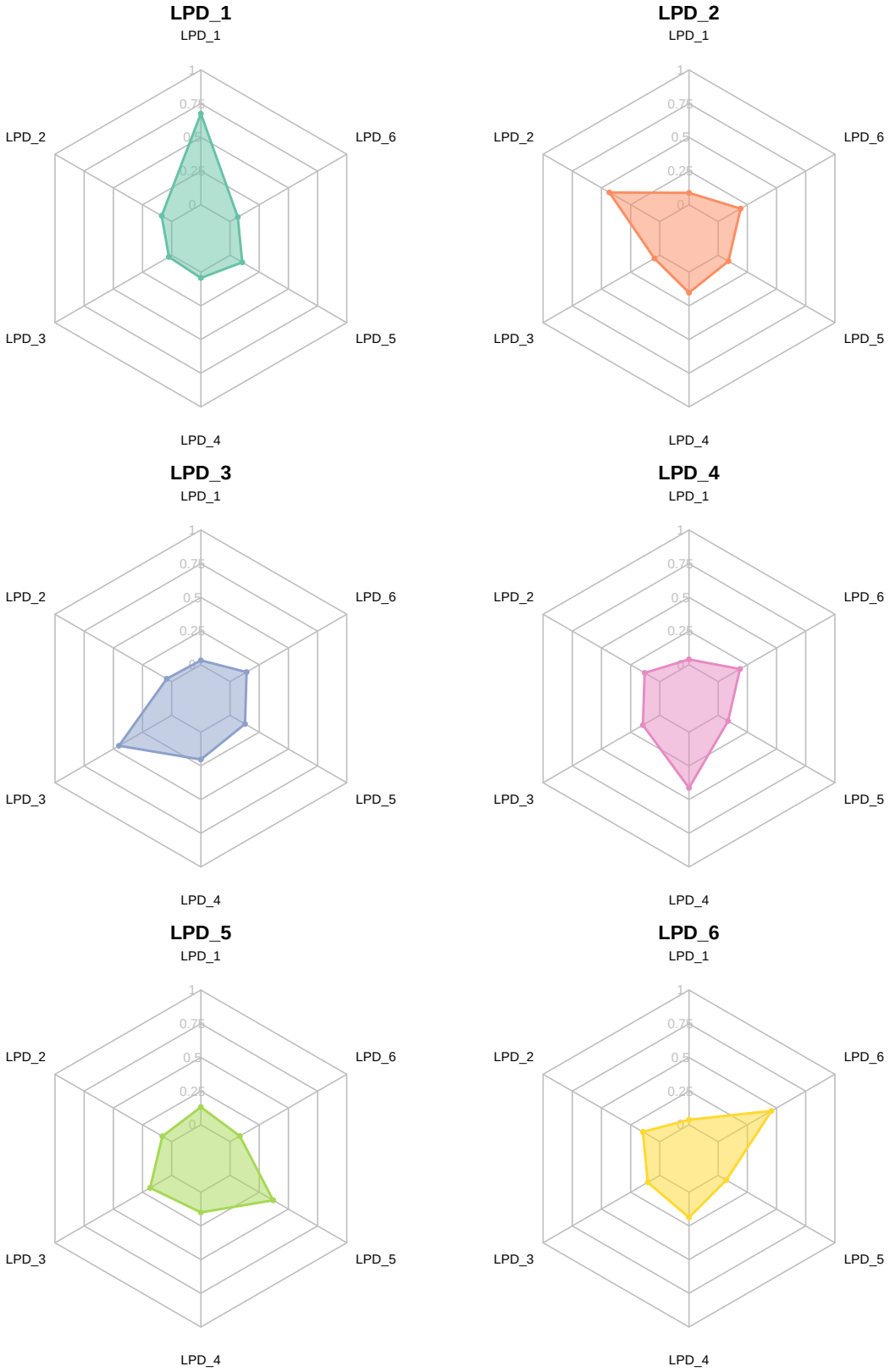

Figure 5.56: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in lung squamous cell carcinoma. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

size effect for the overexpression of *Staphylococcus aureus* infection and steroid biosynthesis respectively. In GO enrichment analysis, LPD_1 exhibited the strongest size effect too, which involved the overexpression of defense response to bacterium and humoral immune response (Fig. 5.58.A). None of the LPD groups showed associations to cancer hallmarks (Figure 5.7).

Table 5.17: Gene counts for various categories in LUSC. These include the number of genes exhibiting significant differential expression and differential methylation, the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively), and the number of genes with significantly higher or lower frequency of CNV (referred to as overimpacted and underimpacted, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7
n health tissue samples	0	0	59	0	0	0	0
DEGs							
Upregulated	1480	120	99	62	61	42	546
Downregulated	939	435	628	703	1267	636	1105
Total	2419	555	727	765	1328	678	1651
Ratio	1.58	0.3	0.16	0.09	0.05	0.08	0.49
DMGs							
Hypermethylated	0	0	53	2	2	12	3
Hypomethylated	3	2	156	1	2	0	2
Total	3	2	209	3	4	12	5
Ratio	0	0	0.33	2	1	1	1.5
Mutated							
Overmutated	320	1576	0	330	778	1028	650
Undermutated	961	616	32	944	876	767	884
Total	1281	2192	32	1274	1654	1795	1534
Ratio	0.33	2.56	0	0.35	0.89	1.34	0.74
Affected by CNV							
Overimpacted	182	244	56	80	135	92	153
Underimpacted	2	2	1	0	1	0	0
Total	184	246	57	80	136	92	153
Ratio	91	122	56	1	135	1	1

Table 5.18: The five genes more overexpressed (log2FoldChange > 1) and underexpressed (log2FoldChange < -1) for each LPD group in lung squamous cell carcinoma. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status
LPD_1		
ENSG00000259001	7.554622	Overexpressed

RNU5B-1	6.959635	Overexpressed
SCARNA5	6.947593	Overexpressed
H4C6	6.904643	Overexpressed
RNY1	6.317706	Overexpressed
CALCA	-5.455774	Underexpressed
ASCL1	-5.159458	Underexpressed
OTX2	-4.175863	Underexpressed
LINC00676	-3.900353	Underexpressed
PAGE1	-3.861308	Underexpressed
LPD_2		
LINC01733	3.648242	Overexpressed
EZHIP	3.335493	Overexpressed
SCARNA10	2.606762	Overexpressed
PAGE5	2.464287	Overexpressed
LINC01287	2.215750	Overexpressed
REG4	-5.629096	Underexpressed
CALCA	-4.617435	Underexpressed
MUC17	-4.226491	Underexpressed
PRB4	-3.863112	Underexpressed
ALB	-3.736734	Underexpressed
LPD_3		
CYP11B1	6.249645	Overexpressed
HSD3B2	3.036680	Overexpressed
PANTR1	2.347195	Overexpressed
PSG1	2.140873	Overexpressed
VAX1	2.125592	Overexpressed
PAGE2	-5.353004	Underexpressed
TRIM48	-4.351509	Underexpressed
REG4	-4.332929	Underexpressed
TUBA3C	-4.322058	Underexpressed
ALB	-4.079147	Underexpressed
LPD_4		
RPTN	2.553057	Overexpressed
PRB1	2.110110	Overexpressed
DSG3	2.050485	Overexpressed
ENSG00000267706	1.805970	Overexpressed
GDPD2	1.673839	Overexpressed
ALB	-4.175284	Underexpressed
PASD1	-4.004412	Underexpressed
DEFA5	-3.926615	Underexpressed
ZIC1	-3.922928	Underexpressed
MAGEA4	-3.916776	Underexpressed
LPD_5		
SPAG11B	4.679144	Overexpressed
ALB	4.030163	Overexpressed
TNMD	2.775628	Overexpressed
CGA	2.746014	Overexpressed

LINC02582	2.580005	Overexpressed
REG4	-4.870428	Underexpressed
MAGEA10	-4.789649	Underexpressed
MAGEA4-AS1	-4.508396	Underexpressed
DLK1	-4.352166	Underexpressed
MARCHF11	-4.182055	Underexpressed
LPD_6		
LINC02672	2.315253	Overexpressed
WFDC5	2.157474	Overexpressed
ENSG00000261166	1.999574	Overexpressed
ENSG00000231317	1.851959	Overexpressed
ATOH1	1.845386	Overexpressed
SSX1	-4.227539	Underexpressed
PRB4	-4.050408	Underexpressed
G6PC	-3.590065	Underexpressed
PRB1	-3.559352	Underexpressed
H4C6	-3.431266	Underexpressed
APOC3	2.980410	Overexpressed
CDH16	2.428384	Overexpressed
AACSP1	2.296290	Overexpressed
ANKRD20A19P	2.266112	Overexpressed
FAM166A	2.203936	Overexpressed
REG4	-5.200540	Underexpressed
MAGEA9B	-4.872916	Underexpressed
DSCR8	-4.512208	Underexpressed
DSCR4	-4.493593	Underexpressed
UPK1B	-4.413044	Underexpressed

Identification of differentially methylated genes in LUSC

All the groups showed a majority of hypermethylation DEGs with the exception of LPD_4 and LPD_5 (Table 5.17; median across groups = 124; IQR = 304). Twelve DM genes were detected as driver genes (Fig. 5.59). In KEGG analysis, LPD_3 showed the strongest size effect that involved the hypermethylation of arginine and proline metabolism and the protein digestion and absorption (Fig. 5.57.B). However, GO enrichment analysed showed the strongest enrichment in LPD_5 for the hypermethlation of cell-cell adhesion (Fig. 5.58.B).

Identification of genes affected by single nucleotide variants in LUSC

A median of 1842 genes across groups were affected by SNVs, with an IQR of 593. Two groups, LPD_1 and LPD_2, showed strong undermutated ratios (Table 5.17). Focusing on the driver genes, several of them were detected as genes affected by SNVs with LPD_1 exhibiting the highest size effect as the undermutation of *LUSC2*, *EPHA3*, *KMT2D*, *NFE2L2* and *PTCH1* (Fig. 5.59). In KEGG, the pathway olfactory transduction was differentially affected by SNVs in all the groups, as well as the PI3K-Akt signalling pathway (Fig. 5.57.C). With the exception of LPD_1, the pathway related to the herpes virus infection was also

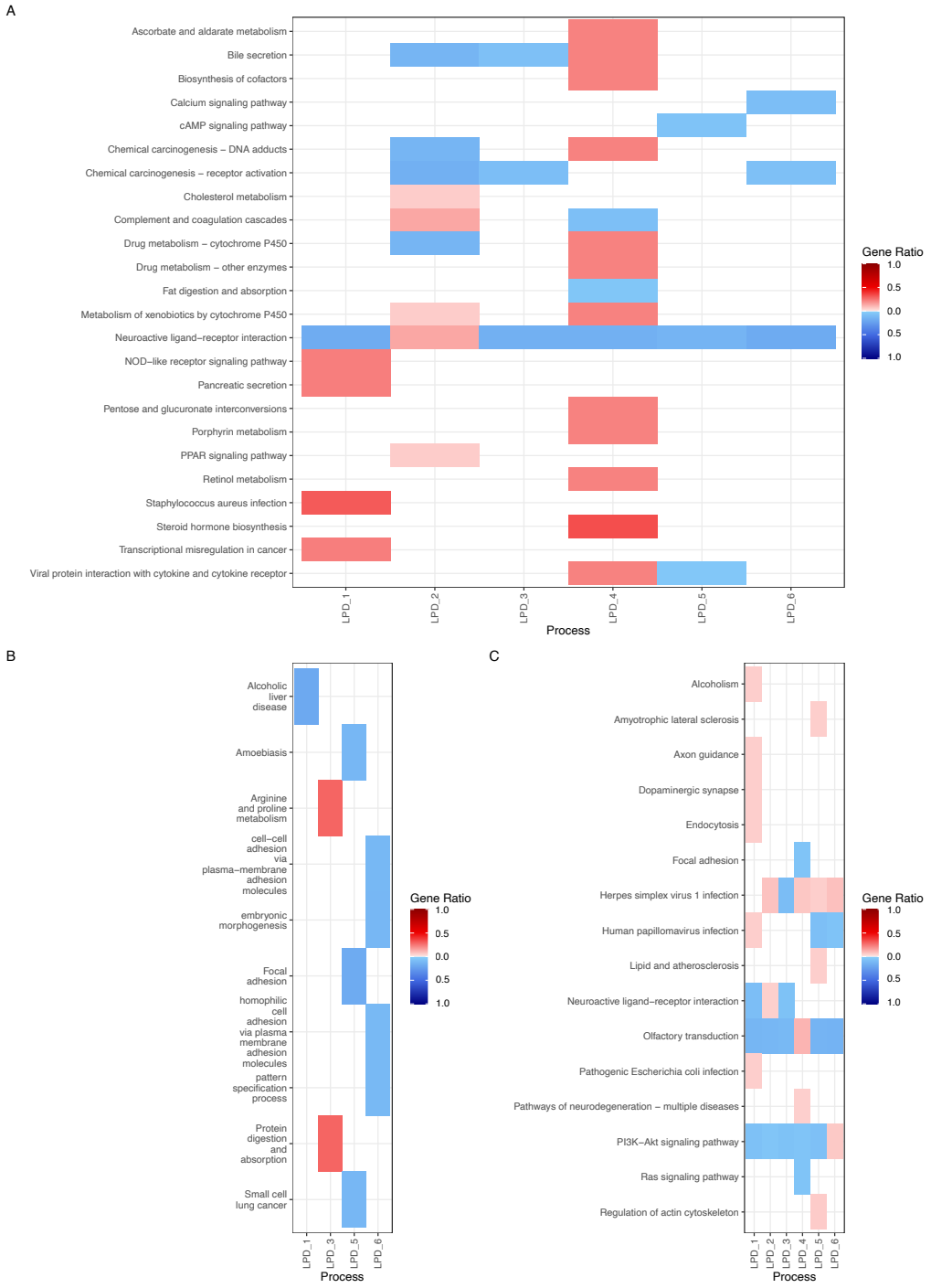

Figure 5.57: Biological pathways associated with different categories in lung squamous cell carcinoma determined using KEGG enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, biological pathways associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the pathways associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological pathways is available in Supplementary Material B.

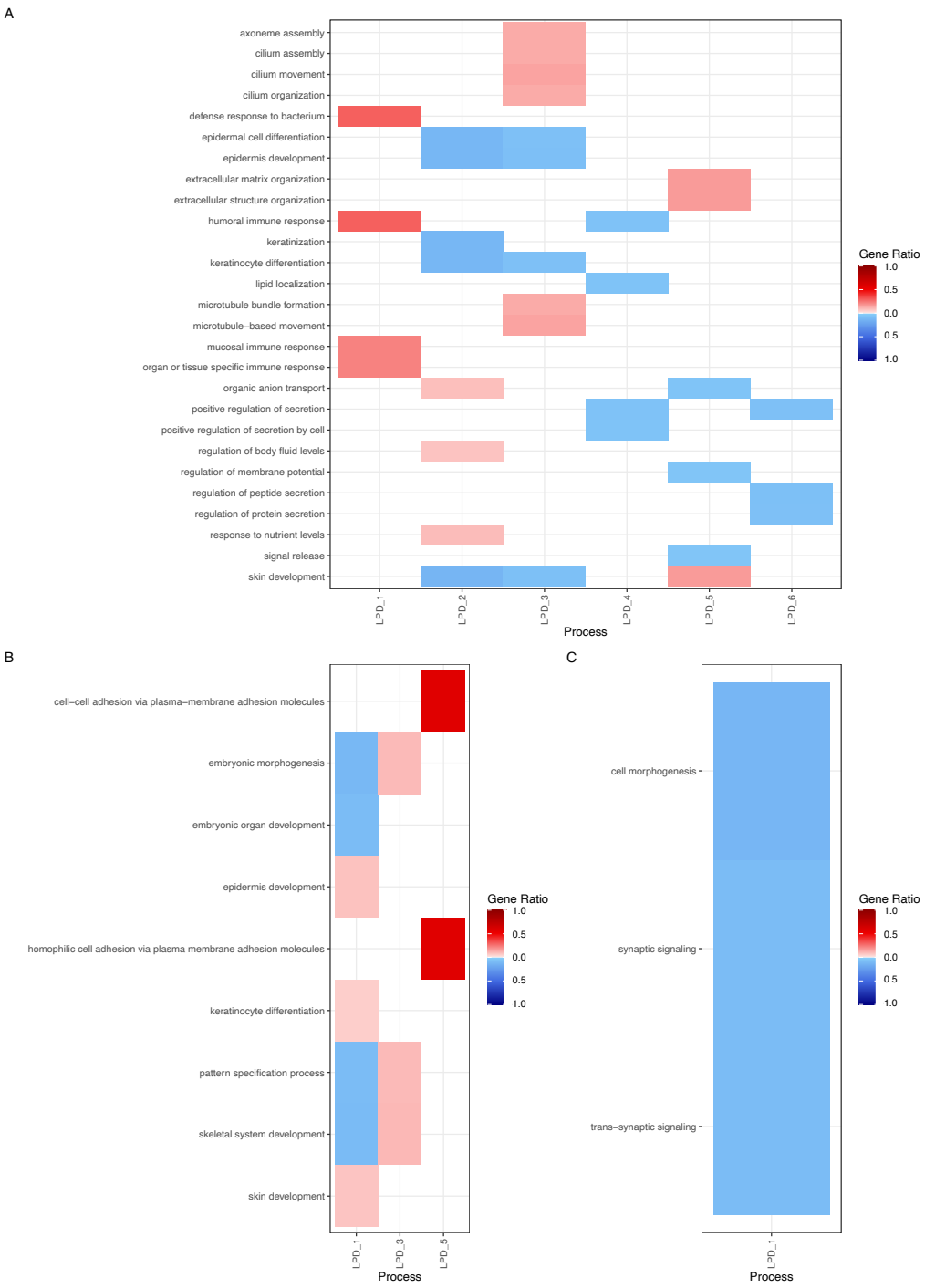

Figure 5.58: Biological processes associated with different categories in LUSC determined using GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant processes with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while processes linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, hypermethylated genes are depicted in red while hypomethylated are represented in blue. (C) Single nucleotide variants, the processes associated to genes with significant higher frequency of mutation are represented in red, while the opposite is represented in blue. The complete list of associated biological processes is available in Supplementary Material B.

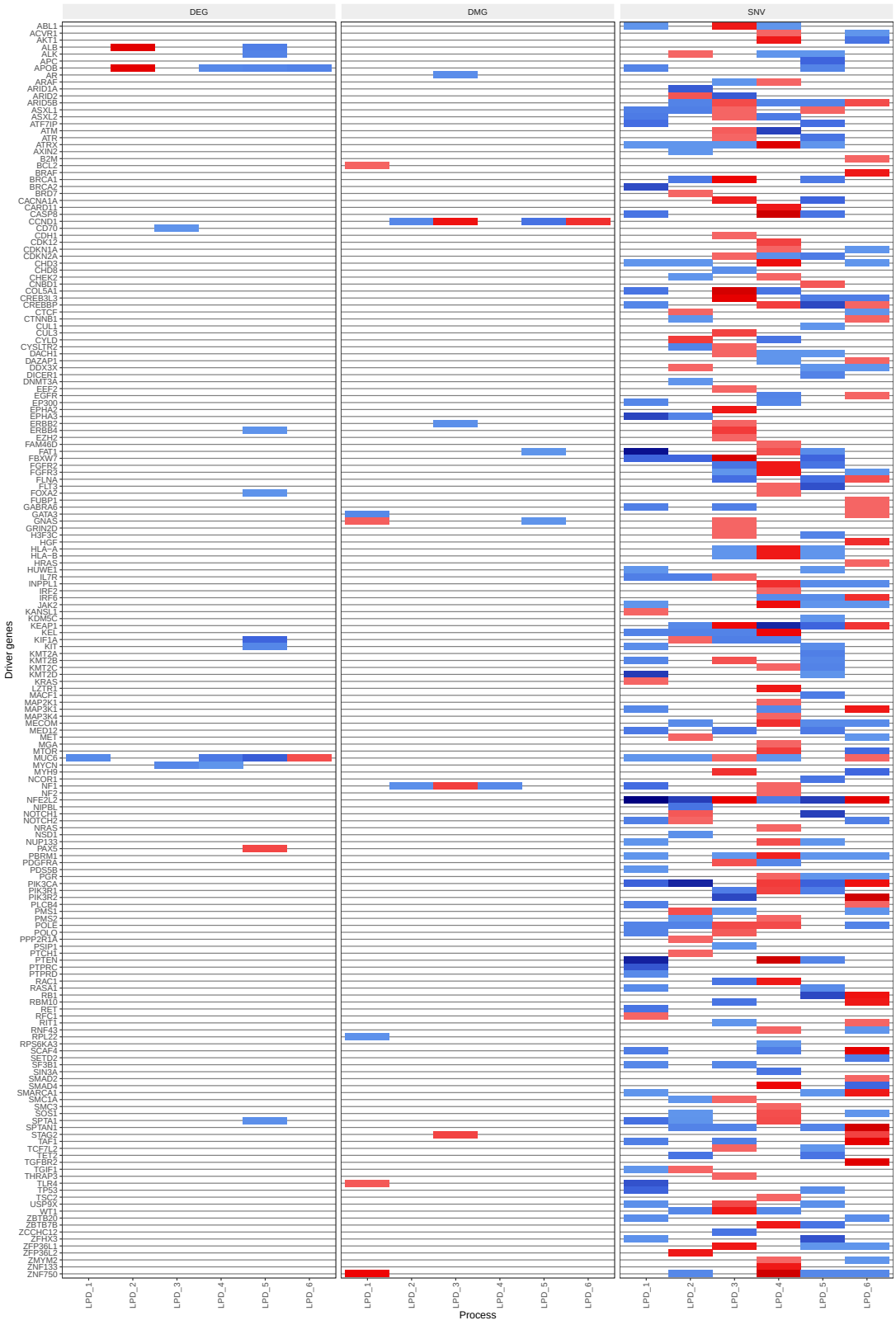

Figure 5.59: Heatmap showing the presence of driver genes across different categories in lung squamous cell carcinoma, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV) and copy number variants (CNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV and CNV are represented in red, while genes with opposite characteristics are depicted in blue.

differentially affected in all groups. GO enrichment analysis only yielded results for LPD_1 where undermutatied genes were enriched in three biological processes: cell morphogenesis, synaptic signalling and tras-synaptic signalling (Fig. 5.58.C). When comparing the SNP type, variant type, and variant class frequency across LPD groups, no significant differences were observed (P>0.05; Chi-squared test; Fig. 5.60). The bulk of detected SNVs were missense mutations caused by the point replacement of cytosine to adenine.

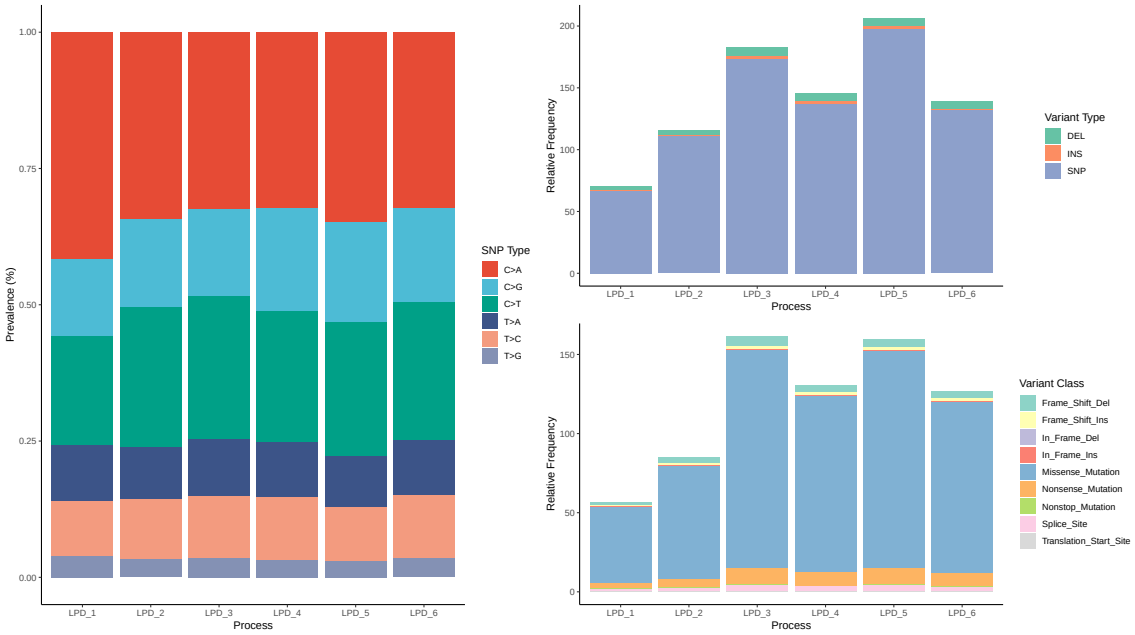

Figure 5.60: Detected single nucleotide variants (SNVs) within each LPD group for LUSC. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

When the COSMIC mutational signature profiles was examined, all the groups exhibited a significant contribution from signature 4, with only LPD_2 showing an additional strong relative contribution from signature 7 (Fig. 5.61).

Functional analysis of differentially expressed genes in LUSC

Matches between overexpressed, hypomethylated and amplified genes are shown in figure 5.62, while matches between underexpressed, hypermethylated, deleted and mutated genes are shown in figure 5.63. For overexpressed genes, no differences across groups were observed ($P = 0.17$; Chi-squared test). On the other hand, for underexpressed genes, LPD_3 and LPD_4 were enriched in matches, while LPD_5 was depleted ($P < 0.0001$; Chi-squared test). The complete list of matches genes is available in Supplementary Material B.

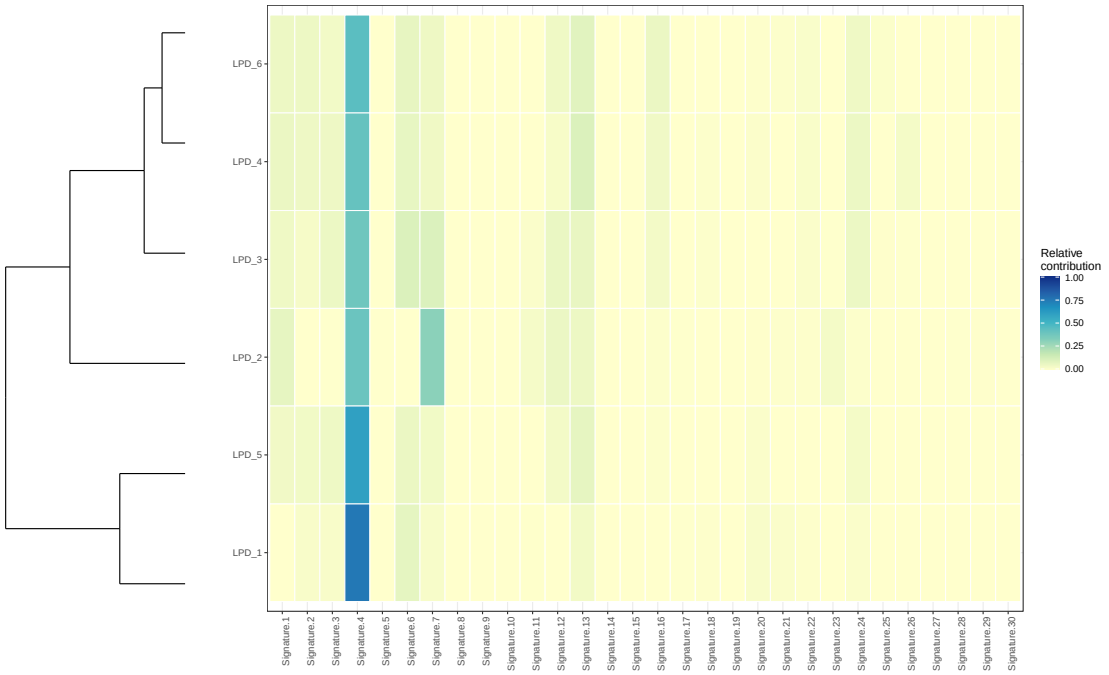

Figure 5.61: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in LUSC. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

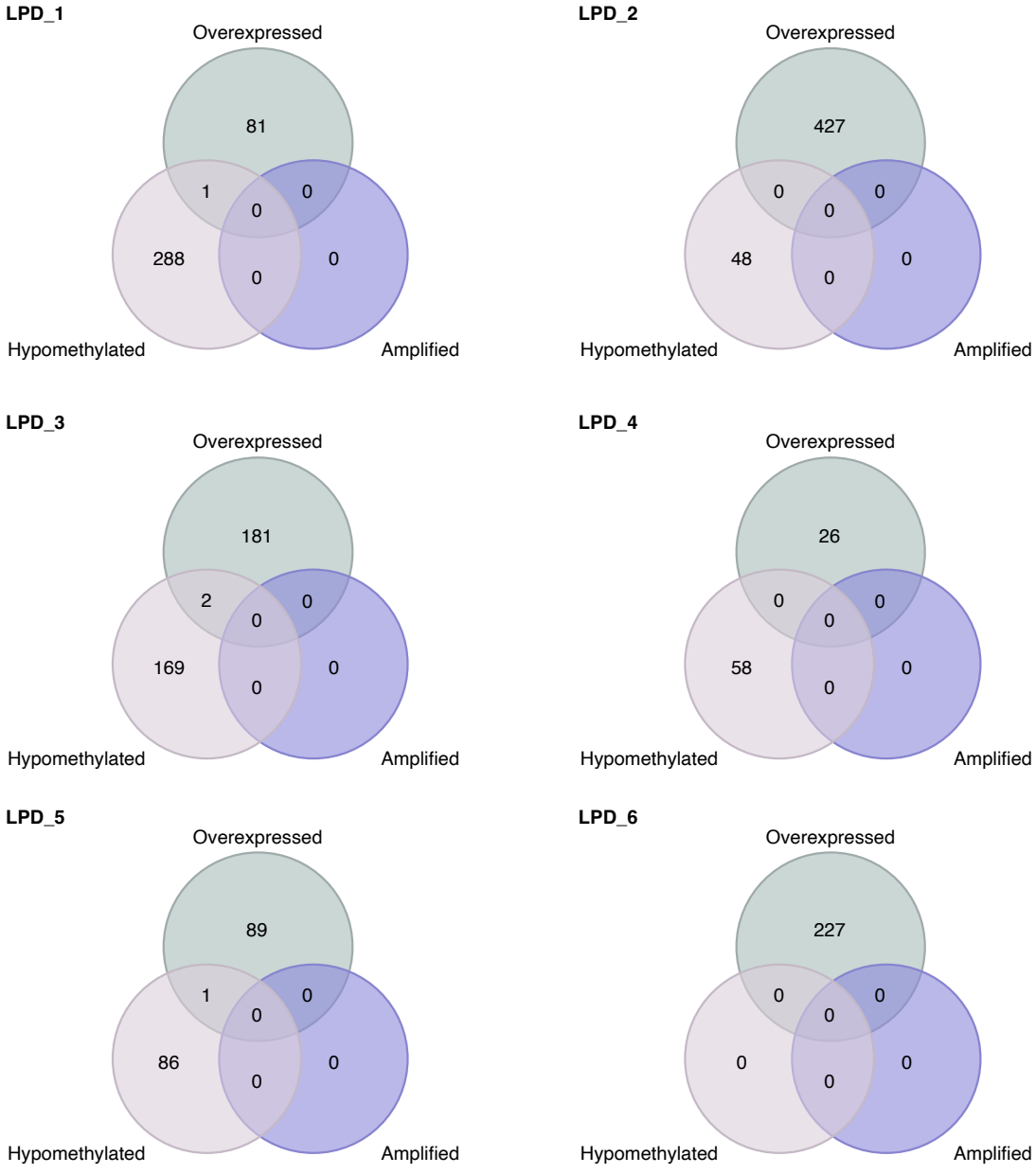

Figure 5.62: Venn diagram displaying the overlaps between three categories in genes in LUSC for each LPD group: overexpressed genes, hypomethylated genes, and amplified genes. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit overexpression, hypomethylation, and amplification simultaneously.

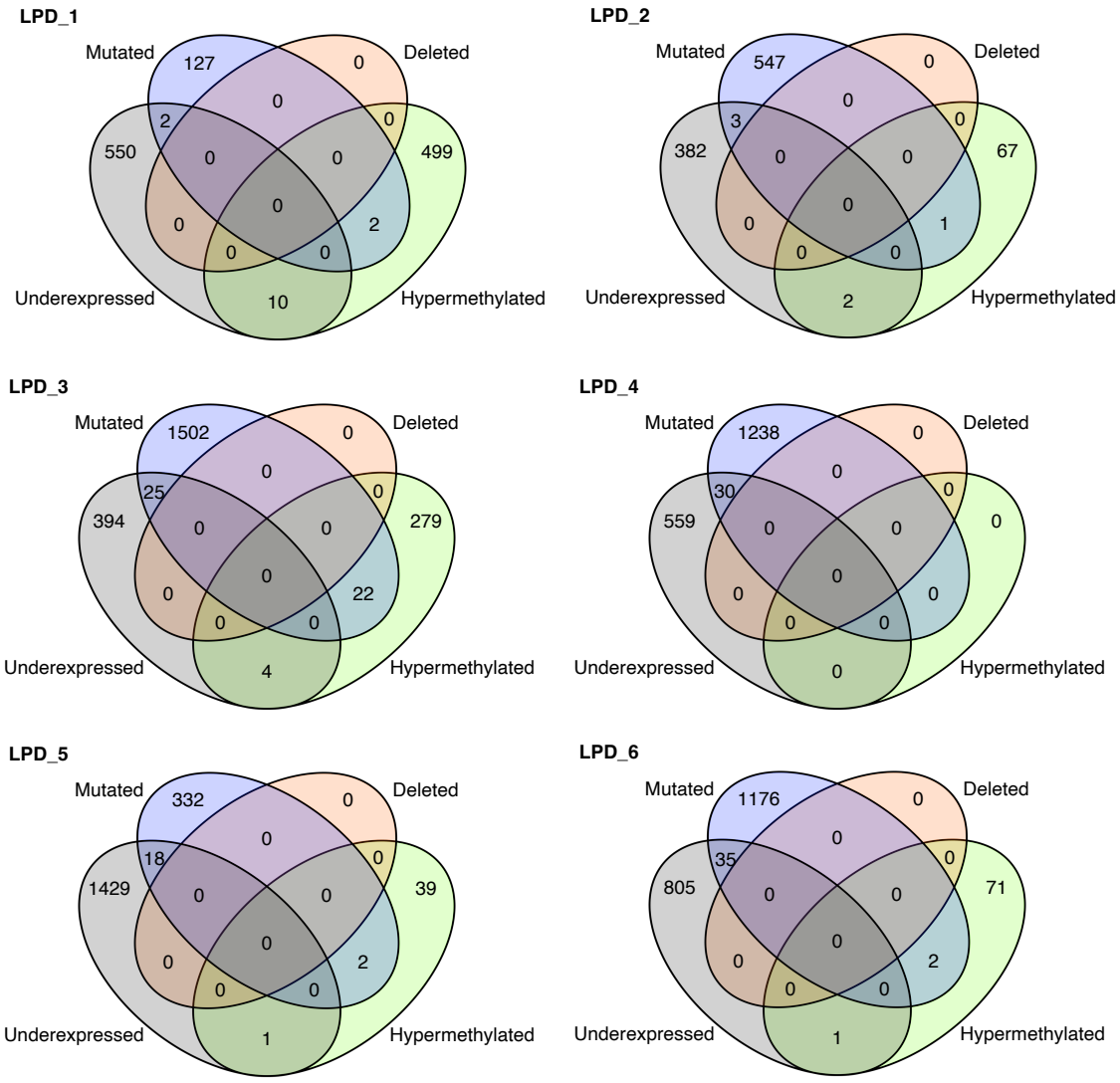

Figure 5.63: Venn diagram displaying the overlaps between four categories in genes in LUSC for each LPD group: underexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit underexpression, hypermethylation, mutations and deletions.

Comparison of the LPD output in LUSC with Euclidian hierarchical clustering

Although the samples were mixed, the samples assigned to LPD_1 and LPD_3 were separated from the others (Fig. 5.64). The red hierarchical cluster seemed to be virtually a perfect match with LPD_1, while LPD_3 was mixed with LPD_4 and LPD_6 to form the yellow hierarchical cluster. LPD_2 and LPD_4 tended to be clustered together across all the hierarchical groups.

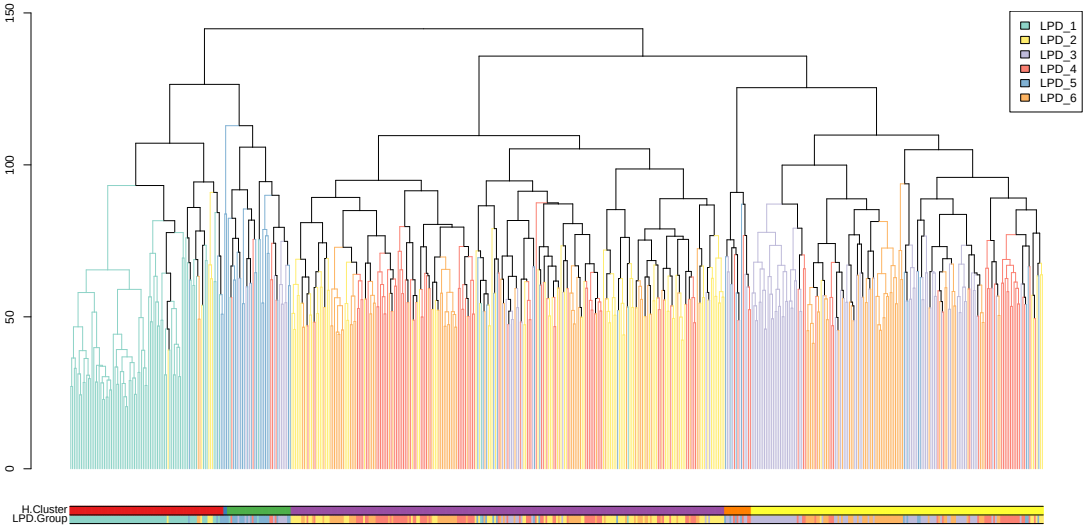

Figure 5.64: Dendrogram showing the sorting of the LUSC samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into seven groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

Comparison of the LPD output with previous subtyping frameworks in LUSC

The Cancer Genome Atlas has previously stated the existence of four subtypes in lung squamous cell carcinoma[153]. These subtypes were classified by diverse features, including alterations in the genes *KEAP1*, *NFE2L2*, *PTEN*, *RB1*, and *NF1*. In my analysis comparing the LPD groups, only LPD_1 showed an alteration, specifically an overmutation in the gene *RB1* (Fig. 5.65).

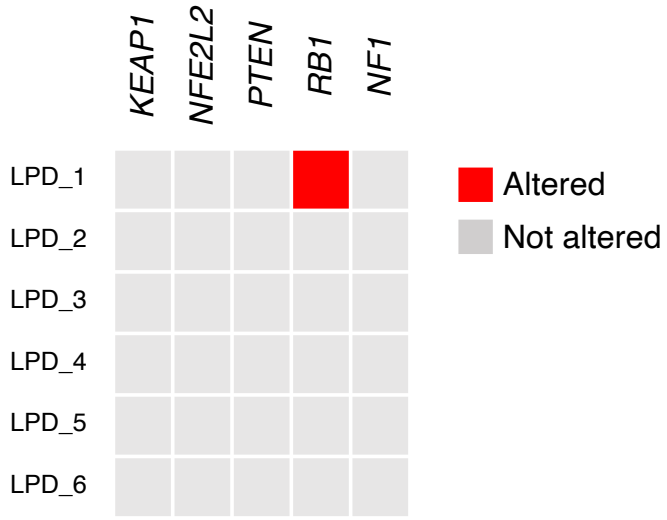

Figure 5.65: Heatmap displaying the presence of alterations in specific genes (*KEAP1*, *NFE2L2*, *PTEN*, *RB1*, and *NF1*) across each LPD group in lung squamous cell carcinoma (LUSC). Each row represents one of the LPD groups, while each column represents a gene to analyze. The genes were selected based on their previous link to classification in LUSC[153]. In the analysis, a gene is considered altered if it shows significant differences across LPD groups in terms of expression and methylation profiles, single nucleotide variant mutations, and copy number variations.

LPD applied to the combined lung carcinoma dataset

LPD was applied to the combined dataset of lungs carcinomas and eight groups were found to be the optimal number of groups (Fig. 5.66). Three of them (G2, G5, G8) were dominated by LUAD with over 93% of their samples assigned to this cancer type, although all of them had a few LUSC samples. Likewise, G3, G4, and G6 showed an allocation of over 93% of their samples to LUAD. G1 and G7 showed a more mixed profile with 55% and 64% of their samples assigned to LUAD, respectively. G1 was formed by 91 primary solid tumour samples, while G7 displayed a total of 105 normal tissue samples (75% of the total) and 33 primary solid tumour samples. The assigments are available in Supplementary Material B.

Euclidean hierarchical clustering applied to the combined lung carcinoma dataset

The Euclidean hierarchical clustering drew a clear separation between the samples belonging to LUAD and the samples belonging to LUSC, although slightly mixed (Fig. 5.67). Most of the hierarchical clusters were visibly dominated by one of the cancer types, except for the brown cluster and a small portion of the yellow cluster that seemed more mixed.

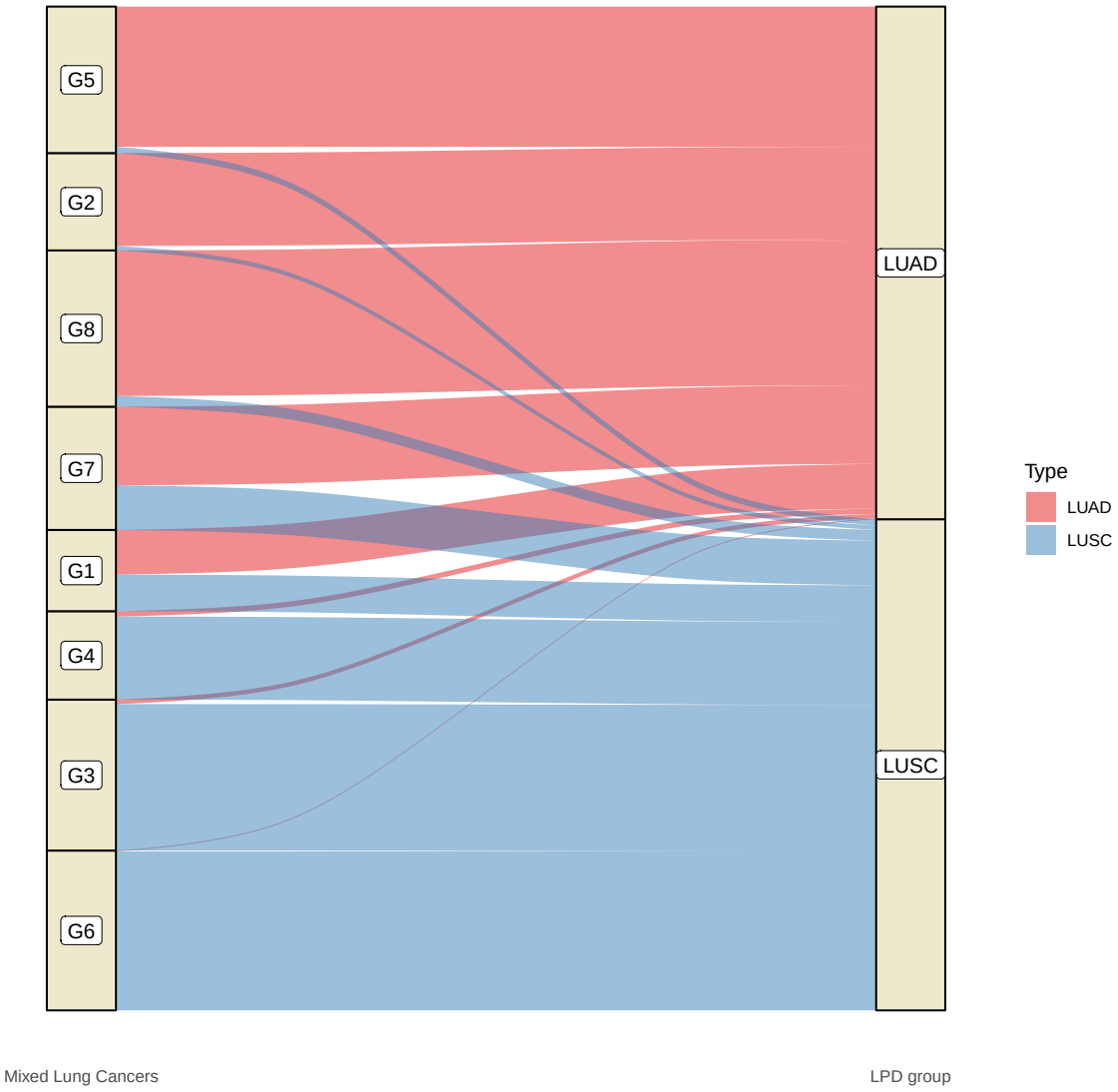

Figure 5.66: Alluvial plot illustrating the assignment of the combined lung dataset samples by LPD and their corresponding lung cancer types. Each cancer type is assigned a distinct colour and the plot represents the presence of samples from each cancer type in each groupd identified in LPD when analysing the combined lung datasets.

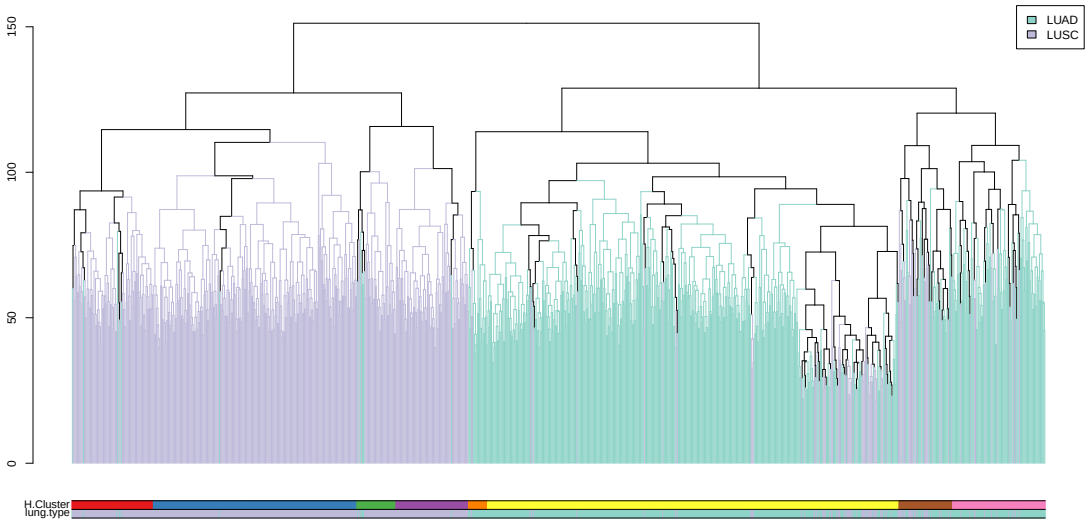

Figure 5.67: Dendrogram showing the sorting of the mixed lung samples using Euclidean hierarchical clustering. Samples with similar expression profiles are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into eight groups (H.Clusters) based on hierarchical clustering, while the second bar (lung.type) indicates the allocation of the samples into their corresponding lung cancer type. The dendrogram branches are colour-coded according to the corresponding lung cancer type.

5.4 Discussion

In this chapter, the LPD algorithm integrated into Automata was applied to four cancer types to validate its potential as an unsupervised cluster approach that works across several cancer types. In breast cancer, LPD was able to distinguish across the five histological types described in the PAM50 classification; in prostate cancer, the algorithm detected a low prognosis subtype with overlapping characteristics to DESNT; in lung cancer, LPD demonstrated its capacity to discriminate between cancer types; however, in colorectal cancer, it failed to detect a low prognosis subtype with shared characteristics with Pericol.

5.4.1 The validation of LPD

Breast cancer

LPD distinguishes three major patterns in breast carcinoma. In the study of breast carcinoma data, LPD identified eight subtypes, each with its traits. However, compared with the PAM50 classification, these subtypes seemed to be broadly categorized into Normal, Basal-like, and Mixed Luminal.

The first category, labeled as "Normal," corresponds to LPD_8 in which the samples were predominantly matched with the PAM50-Normal subtype (Fig. 5.15). This alignment is consistent with the composition of LPD_8, with 98 out of 119 samples representing normal non-malignant tissue (Table 5.1). The dominance of normal tissue samples in this group may account for the high gamma values observed (Fig. 5.2), as well as the substantial number of genes affected by methylation, SNV, and CNV (Table 5.2). Furthermore, this is consistent with the hallmark of cancer associated with promoting inflammation being underexpressed in LPD_8. The distinct tissue composition and molecular features of this group also contribute to its well-defined separation in the hierarchical clustering analysis (Fig. 5.12).

The Basal-like pattern is based on the PAM50-Basal subtype, which is characterised by the absence of progesterone, estrogen, and HER2 receptors. This pattern corresponds to the group LPD_6, which was significantly lower in the presence of the three receptors. The PAM50 classification performed by Netanely et al. (2016)[296] revealed that over 85% of the samples assigned to this group were identified as authentic Basal samples. Consistently, the samples in LPD_6 demonstrated a robust assignment, a distinct COSMIC profile, and a good separation from the other groups in hierarchical clustering.

The Mixed Luminal pattern includes all the remaining subtypes (LPD_1, LPD_2, LPD_3, LPD_4, LPD_5 and LPD_7) and is defined by the varied presence of Luminal A and B samples, as well as HER2 samples in certain groups. These six subtypes were also found to be mixed in the hierarchical clustering and there were no clinicopathologic distinctions between them. However, LPD_1 and LPD_7 demonstrated distinct prognostic outcomes compared to the other groups. In the COSMIC profiling, LPD_1, LPD_3, LPD_5, and LPD_7 showed a strong resemblance, while LPD_2 and LPD_4 seemed uniquely distinct from the other groups. An additional difference between LPD_2 and the rest of the groups was observed in the associations to hallmarks, as LPD_2 was the only group with overexpression of biological process related to replicative inmortality. These findings suggest that LPD has the potential to be able to find subsets of the luminal subtypes, however further work needs to be done to validate these results.

Luminal A, Luminal B and HER2 are mixed. Despite the clear differentiation of subtypes outlined by Sorlie et al. (2001)[297], the LPD groups appeared to be very mixed, with five LPD subtypes displaying varied amounts of Luminal A and B, in addition to HER2. This suggests that the LPD algorithm may not be as efficient for breast cancer as hierarchical clustering. Nonetheless, my results were consistent with those of Netanely et al. (2016)[296]. They analysed RNA-seq data for TCGA-BRCA through unsupervised clustering and reported that the separation between Luminal A and Luminal B is a continuum. According to the proportion of Luminal A or B, they could classify the samples into processes with superior predictive values than the PAM50 classification. Similarly, Daemen et al. (2018)[298] speculated that HER2 overexpression, rather than being a subtype marker, may be an event during the carcinogenesis that occurs in all cancer subtypes regardless of their PAM50 classification. Thus, in terms of differential molecular landscape, the classification performed by LPD may be more accurate than the PAM50 categories as it can distinguish Normal and Basal samples and reflect with exactitude the molecular compositions of the Luminal samples.

LPD results are consistent with previous applications in breast carcinoma. The results obtained from applying the LPD step integrated into Automata to breast carcinoma were consistent with the previous findings of Carrivick et al. (2006)[244]. Their research analysed several breast carcinoma datasets (50-200 sample size). They discovered four subtypes, one indolent and three of wide-ranging aggressiveness, with one matching the Basal subtype and another the Luminal subtype. Similarly, my data revealed the presence of a Basal group and a Normal group but differed in the number of Luminal subtypes. This suggests that the identified Mixed Luminal groups may be subdivisions of the Luminal subtype outlined in Carrivick's study. A possible explanation of this difference is that Carrivick's LPD approach was built by executing 50 iterations to select the optimal number of processes in combination with Kaplan-Meier estimators, which could bias the results towards the detection of clinical stratifications over the detection of gene expression signatures. Furthermore, Carrivick remarked that larger sample sizes might result in the finding of additional subtypes.

Overall, the application of LPD to breast carcinoma data provides a more refined and comprehensive characterization of subtypes, raising questions about the reliability of the current PAM50-based classification system. Previous studies have acknowledged limitations in the PAM50 classification, as it may not fully capture the clinical response and complex underlying biology of the molecular pathways driving these subtypes[299]. The findings from LPD offer promising insights into the molecular heterogeneity of breast cancer and have the potential to revolutionize our understanding of the disease.

Prostate cancer

LPD detected DESNT in the TCGA. In their work, Luca et al. (2018)[118] applied LPD to five datasets including the TCGA. Except for the TCGA, they detected the DESNT signature in all of them. As a workaround, Luca developed a random forest model based on the gene expression profiles of the 45 core underexpressed genes linked with DESNT and used it to identify the signature in the TCGA. In my study, I successfully identified a poor prognosis group (LPD_4) that has the same prognosis as DESNT and a significant correlation when comparing their gamma values distribution. Furthermore, 77% of the samples of LPD_4 were allocated to DESNT in Luca et al..

I also found that 56% of the samples assigned to LPD_1 were also classified as DESNT, despite the fact that this group has no differential prognosis. LPD_1 shared some molecular traits with DESNT, such as having a match of 11 genes out of 16 with the hypermethylated genes defined by Luca as hypermethylated in the TCGA, and the overlap of 22 DMGs with the core 45. Additionally, LPD_1 has shown through the pathways and driver genes analysis an opposing molecular profile to normal samples, suggesting that LPD_1 is a potential aggressive subtype. I believe that LPD_1 and LPD_4 are subdivisions of the DESNT group detected by Luca, but further research would be needed to confirm this.

However, upon comparing the LPD output with the classification frameworks proposed by the TCGA in 2015[115] for prostate cancer (see section 1.5.2), no matches were found. The TCGA paper classified prostate cancer into groups based on various criteria, one of which included a high number of point mutations in three key genes: *SPOP*, *FOXA1*, and *IDH1*. Although LPD_2 displayed an overmutation of *SPOP* (Fig. 5.23), the other features of the samples in this LPD group did not align with the characteristics described by the TCGA, such as significant hypermethylation (Table 5.7). This indicates that LPD-identified subgroups may not fully correspond to the established classifications, highlighting the need for further investigation and validation of LPD results.

Colorrectal cancer

Pericol was not detected by LPD in the TCGA. Ellis (2021)[171] characterised a low prognosis subtype named Pericol by building an LPD model with samples from four datasets retrieved from the Gene Expression Omnibus database. However, when LPD was applied to the TCGA, six processes were found with no differences in prognosis across them. Ellis instead conducted a correlation analysis between the TCGA processes and Pericol and observed a significant positive association for process C3. In my study, the LPD model built using only TCGA-COAD data failed to identify low prognosis groups, correlations with the process C3 (Fig. 5.42), and similarities in sample assignment with C3 (Fig. 5.41). These findings suggests that the LPD model built in TCGA-COAD is incapable of detecting Pericol. More research and analysis of the TCGA sample pool will be needed to determine the cause of this phenomenon.

Automata LPD found three subtypes with similar characteristics with Ellis' results. Aside to Pericol, Ellis (2021)[171] detected three more processes that were conserved across several datasets and labelled as LPD A, LPD B, and LPD C. These three processes correlated to the processes C2, C4, and C5, respectively.

LPD A (C2) was defined by an overrepresentation of normal samples and enrichment in pathological stage I. The presence of normal samples is also present in my group LPD_5. However, both of these groups failed to show a significant correlation or similarities in their sample assignment (Fig. 5.42; Fig. 5.41).

LPD B (C4) was characterised as a subtype formed by distal samples, with microsatellite instability, overexpression of mismatch repair pathways, chromosomal instability, and overmutation of the gene *TP53*. This subtype matched with the process LPD_2 and LPD_4. Both of these groups showed a shared gamma value distribution (Fig 5.32), were frequently clustered together in the hierarchical clustering (Fig. 5.40), and were the only groups associated with COSMIC mutational signature 6 (Fig. 5.37). However, none of these groups

shared the overexpression of mismatch repair pathways and overmutation of the gene *TP53* with LPD B. The remaining characteristics of LPD B were not included in the scope of this thesis and therefore their comparison could not be performed.

LPD C (C5) was described as a subtype formed by proximal samples, enriched in pathological stage II, with high microsatellite instability, global genome hypermethylation, and overmutations of the gene *BRAF*. This subtype matched with LPD_3, which was characterised as enriched in colon mucious adenocarcinoma samples, a hypermethylation of several driver genes, an enrichment of DEG genes due to hypermethylation and mutations, and a overmutation of the gene *BRAF* among others (Fig. 5.34).

Finally, although not conserved across several datasets, the process C1 was matched to LPD_7. This LPD group was defined as a robust assigned group, with good separation from other groups in the hierarchical clustering, with a moderate underexpression (Ratio = 0.71) and hypermethylation (Ratio = 1.72), and a total undermutation (Ratio = 0).

Although I have successfully found matches for LPD_2, LPD_3, LPD_4, and LPD_7 with processes previously described by Ellis, the matches were only partial. This reiterates my hypothesis that the LPD model is unable to fully grasp the complexity of the colorectal cancer when using the TCGA data.

Lung carcinomas

LPD was able to differentiate between both lung carcinomas. Lung adenocarcinoma and lung squamous cell carcinoma are two of the most common types of lung cancer and they are often treated similarly and classified together despite their different molecular landscape[300,301]. To validate the potential of LPD as a tool for analysing different cancer types, I combined both lung datasets and applied LPD to see whether the algorithm could discriminate between them. I tried this same experiment before successfully with Euclidean hierarchical clustering, therefore I expected a similar if not better outcome from LPD. Except for two processes, G1 and G7, my results demonstrated that LPD was capable of separating samples from both cancer types. The reason that LPD was unable to categorise correctly G7 was attributed to the fact that 75% of the samples in this group were healthy tissue samples with virtually no molecular differences between them. On the other hand, a possible explanation for G1 is that is a group composed of samples with cancer signatures common to both types of cancer. As discussed in chapter 4, several distinct cancer types share common mutations and driving processes that stratify them into tumours. As a result, my findings demonstrate that LPD can distinguish between various cancer types.

No matches with previous subtyping frameworks. In the LUAD dataset, the TCGA (2014)[148] defined three subtypes: terminal respiratory unit (TRU), proximal-inflammatory (PI), and proximal-proliferative (PP) subtypes. These subtypes were primarily characterized by specific mutations in genes such as *EGFR*, *NF1*, *TP53*, and *KRAS*, as well as chromosomal deletion of *STK11*. When analyzing the LPD groups derived from the LUAD dataset, the only significant association detected was between LPD_6 and the gene *EGFR*, which is a distinctive trait of the TRU subtype (Fig. 5.50). However, LPD_6 did not exhibit any of the other traits that defined the TRU subtype, such as a good prognosis. This indicates that LPD was able to identify a specific genetic trait associated with the TRU subtype in the LPD_6 group but did not fully capture all the characteristics that define

the TRU subtype as described in the TCGA classification.

In the LUSC dataset, the TCGA (2012)[153] identified four distinct subtypes based on specific genetic and molecular characteristics, including alterations in the genes *KEAP1*, *NFE2L2*, *PTEN*, *RB1*, and *NF1*, hypermethylations, and overexpression of immune system-associated genes. When comparing these TCGA-defined subtypes with the LPD groups, no significant associations were found, except for LPD_1, which showed an overmutation in the *RB1* gene (Fig. 5.65). However, this single association is not sufficient to conclude that LPD_1 fully match one of the TCGA subtypes. Further research would be required to conclude whether the results from LPD align witht the ones from the TCGA.

5.4.2 Subtypes with differential prognosis

Breast cancer displayed two differential prognosis subtypes. Among the eight LPD groups identified in breast carcinoma, LPD_1 and LPD_7 showed significant differences in prognosis compared to the other groups. LPD_1 exhibited a poor prognosis, while LPD_7 displayed a good prognosis (Fig. 5.3). Both of these groups were classified within the Mixed Luminal category, making the good prognosis of LPD_7 consistent. However, the poor prognosis of LPD_1 was unexpected. Nevertheless, the survival analysis conducted using the PAM50 classification (Fig. 5.13) revealed that the PAM50 categories did not behave as expected in terms of prognosis. An additional detected anomaly was that, although not significant, LPD_6 showed overall good prognosis when compared to the other LPD groups. The LPD_6 group was associated with the Basal subtype, which is known to have the worst prognosis of the five PAM50 subtypes due to its quick growth and lack of receptors that difficult treatment[302]. Nonetheless, the analysis of survival probability of the BRCA samples categorized according to their PAM50 subtype also showed that Basal had the best prognosis, with a similar curve to LPD_6 (Fig. 5.13). This indicates that the PAM50 classification derived from the TCGA-BRCA dataset may not function as expected and could suggests an issue with the clinical or molecular data of the project rather than with the LPD algorithm. Still, LPD was able to distinguish the Basal, Normal, and Luminal samples, confirming the potential of LPD as a viable methodology for the analysis of cancer.

LPD_1 was characterised as a robust group with overexpression of splicing mechanisms and underexpression of muscle system processes. The genes most differentially expressed in this group were the underexpression of *CSN2*, *LALBA*, and *DCAF4L2*, and the overexpression of *SNORA74B*, *RNU4-2*. *CSN2* and *LALBA* are genes associated with milk production and are believed to be involved in the development of the mammary gland[303]. *DCAF4L2* is a gene that participates in the protein-protein interplay and promotes hepatocellular carcinoma, colorectal cancer invasion, and metastasis[304,305]. *SNORA74B* expression is positively associated with the development of gallbladder cancer[306]. *RNU4-2* is a small nuclear RNA that regulates the splicing process. In addition, this group was associated with 37 driver genes, defined by the overmutation of tumour suppressor genes such as *ATF7IP*, *CDH1*, *GPS2*, *MGA*, and *TBX3*.

LPD_7 was characterised as a group of older patients and the underexpression of epidermis development and muscle system process. LPD_7 shares *CSN2* and *LALBA* with LPD_1 as the most differentially expressed genes, along with *SULTIC3*, *LACRT*, and *CARTPT*. *SULTIC3* is a sulfotransferase gene, whereas *LACR* promotes ductal cell proliferation, and

its expression is associated with human breast cancer[307]. Lastly, *CARTPT* is involved in the appetite, energy balance and maintenance of body weight. In addition, this group was associated with 61 driver genes, including the overmutation of tumour suppressor genes (*CHD8*, *CTNND1*, *DDX3X*, *MSH6*, *PMS2*, *PSIP1*, and *ZFHX3*) and the amplification of seven oncogenes (*CCND1*, *CD79B*, *CDK4*, *ERBB2*, *IDH2*, *MYC* and *SPOP*). Moreover, my analysis found additional evidence of a functional effect of nine underexpressed genes in this group: *PLIN1*, *ALDH1L1*, *OLIG1*, and *ALDH1L1-AS2* that were underexpressed and hypermethylated; *UGT2B28* and *RPS26P38* that were underexpressed and mutated; and *MYOM1* and *GRIA1* that were underexpressed and deleted. *PLIN1* is a significantly underexpressed gene in breast cancer and it is considered an independent predictor of survival in estrogen receptor positive cancers[308]. *ALDH1L1* is usually hypermethylated in cancers including lung adenocarcinomas and hepatocellular carcinoma[309].

The characteristics of these groups did not seem to align with any molecular features of breast cancer I could find in the literature, suggesting that these subtypes would require further research and validation to fully understand their biology and the significance of their differential expressed genes in cancer development.

Prostate cancer and DESNT. LPD_4 in prostate cancer exhibited several similarities to the DESNT subtype as discussed in the above section. The genes most differentially expressed in this group were the underexpression of *XIRP2*, *MYH7*, *SMYD1*, *APOA2*, and *MYL1*. Four of these genes, *XIRP2*, *MYH7*, *SMYD1*, and *MYL1*, are involved in the organisation, regulation, and development of the cardiac muscle tissue. *APOA2*, on the other hand, is associated with the regulation of lipids and its expression level could be a potential biomarker of prostate cancer, along with hepatocellular carcinoma, gastric cancer, myeloma, and pancreatic cancer[310].

5.5 Conclusions

In this chapter, I demonstrated the potential of the LPD algorithm for the analysis of cancer transcriptome by testing it in different scenarios and comparing it with previous results. Moreover, I have analysed and characterised the detected cancer subtypes in breast carcinoma, prostate adenocarcinoma, colorectal cancer, lung adenocarcinoma, and lung squamous cell carcinoma. Furthermore, I have proved the capacity of LPD to detect differential prognosis subtypes, which suggests that LPD could be a valuable tool to comprehend the heterogeneous response to treatment present across many cancer types.

5.6 Summary

This chapter presents the outcome of applying the LPD algorithm integrated into Automata in breast, prostate, colorectal, and lung cancer. In breast cancer, LPD identified eight subtypes, broadly categorized into Normal, Basal-like, and Mixed Luminal patterns. LPD_1 and LPD_7 showed distinct prognostic outcomes, with LPD_1 displaying a poor prognosis and LPD_7 a good prognosis. In prostate cancer, LPD detected a poor prognosis group (LPD_4) with similarities to the DESNT subtype. In colorectal cancer, LPD failed to detect the Pericol subtype in the TCGA dataset, requiring further research for validation. In lung cancer, LPD successfully differentiated between lung adenocarcinoma and lung squa-

mous cell carcinoma, with a few exceptions. These findings demonstrated the potential of LPD to identify cancer subtypes with differential prognosis and to provide insights into the molecular landscape of various cancer types.

Chapter 6

Identifying and characterising subtypes with a significant association with outcome

6.1 Introduction

Although there have been many advances in screening programs and treatments in recent years that have reduced the mortality rate of cancer, it still remains the second leading cause of death worldwide, accounting for almost 10 million deaths worldwide in 2020[311]. One possible explanation for this is the heterogeneity found across samples from the same cancer that is driven by unknown processes. In many cancer types, clinicopathological features, immunohistochemistry, and gene mutations are employed as prognostic and therapeutic indicators. However, the current known molecular markers fail to represent the complexity of the tumour environment and still result in variable outcome of therapy response. Molecular classifications of cancers into subtypes based on gene expression surged as a solution for this issue and helped to identify good and poor prognosis subtypes that could potentially be translated to clinical practice and treatment selection. Examples are the classification of breast carcinomas (see section 1.5.1) and the DESNT signature for prostate adenocarcinoma (see section 1.5.2).

In this chapter I aim to (i) identify molecular subtypes with differential prognosis detected by LPD, and (ii) characterise them to uncover potential novel therapeutic targets and potential biomarkers to improve patient classification. I will focus on the characterisation of the differential prognosis subtypes in skin cutaneous melanoma (SKCM) and bladder urothelial carcinoma (BLCA). SKCM is a disease that affects the pigment-making cells in the skin and causes more than 10,000 deaths each year, yet it has a high five-year survival rate (98%) when detected early[312]. This malignancy is commonly associated with overexposure to ultraviolet radiations from the sun or indoor tanning[313]. Its symptoms include the apparition of a new skin mole or the change in the appearance of an existing mole, such as increasing size, changing shape or colour, or being itchy. The most prevalent type is distinguished by a superficial spread over the outermost layer of the skin before infiltrating deeper layers[313]. BLCA affects the bladder and the urothelial tract, organs responsible for storing and eliminating urine from the body. It is the eighth most common cancer in

England, with over 500,000 cases and about 230,000 deaths globally each year (five-year survival rate of 90%)[314,315]. The symptoms of this malignancy are the presence of blood in the urine, a burning sensation during urination and a high frequency of urination[315].

6.2 Methods

The data used in this chapter was gathered and processed following the Automata workflow that is explained in detail in section 3.2. The specifics of the statistical tests, databases and computational resources are described in chapter 2.

6.2.1 Survival analyses

The Kaplan-Meier curves and log-rank tests performed by Automata for each of the LPD groups. From the subtypes with significant log-rank test (p-value < 0.05), the subtypes were classified into *good* or *poor* prognosis according to their survival probability.

A Cox regression model was built using the gamma value output from LPD to see if there was a significant association with outcome and the size of the effect (hazard ratio).

6.2.2 Exploring the LPD output

For the majority of this chapter I focused on the results from the projects corresponding to bladder carcinoma (TCGA-BLCA) and skin cutaneous melanoma (TCGA-SKCM).

With a Chi-square test, Automata examined the presence of batch effects due to the Tissue Source of the samples, as well as whether the presence of benign tissue samples was uniformly distributed across groups. Additionally, the mean gamma values of all samples allocated to the same group was calculated. Those that showed a mean gamma value larger than 0.5 were considered a *robust assignment*.

6.2.3 Determing Important Clinicopathologic characteristics

Boruta[294] was used to select the clinical features that were important in predicting the assignment of samples into LPD groups. A Chi-squared test was performed to find if there were significant differences in the selected features across the LPD groups.

6.2.4 Identification of differentially expressed genes (DEGs)

Automata calculated the number of DEGs across each group for each cancer type and represented their differential expression as log_2 *fold change*. Additionally, the biological processes associated with DEGs were studied using an enrichment analysis in the KEGG and GO databases. The GO terms obtained were then studied to identify the ones involved in cancer hallmarks. The complete list of GO terms related to cancer hallmarks was obtained from Chen et al. (2021)[295]. Genes found higher in the subtype compared to all other samples were labelled as *overexpressed* or *upregulated*, while those that were lower were labelled as *underexpressed* or *downregulated*. The ratio of the number of overexpressed genes divided by underexpressed genes was calculated.

A list of known cancer driver genes from Bailey et al. (2018)[264] (n = 299) was cross-referenced against the identified DEGs.

6.2.5 Identification of differentially methylated genes and genes enriched or depleted with mutations

The workflow described in the previous section was repeated for genes that were differentially methylated and genes that were affected by single nucleotide variants. Genes with that had higher beta values in the subset were labelled as *hypermethylated* while those with lower values werelabelled as *hypomethylated*. Genes that had an enrichment of mutations (SNVs) in a subset compared to all other samples are labelled *overmutated* and those significantly depleted, *undermutated*.

Automata classifies SNVs into Single-nucleotide polymorphisms (SNPs), insertions or deletions and further subdivides them into the predicted effect (frameshift, missense, stop etc.). Differences in the proportions of each type are detected using a Chi-squared test. Furthermore, Automata also determines the proportion of each mutation in an LPD groups that is associated with each COSMIC mutational signature.

Automata also provides a list of genes that were co-occurring as DEG, DMG, and affected by SNVs for the subtypes with differential prognosis. This was genes that had additional evidence for functional importance. The DEGs were split into overexpressed and underexpressed. Only co-occurances with hypomethylated genes were judged relevant for overexpressed genes, whereas co-occurances with hypermethylated, and mutated genes were considered relevant for underexpressed genes. A Chi-squared test was used to compare the frequency of co-occurrences across LPD groups.

6.2.6 Comparison of the LPD output with Euclidian hierarchical clustering

Hierarchical clustering of the samples based on Euclidean distance and complete linkage to compare its output with the LPD approach. For each cancer type, a dendrogram was generated to visualise and compare assignments.

6.3 Results

6.3.1 Differential prognosis subtypes in the TCGA

A total of 26 subtypes detected by LPD across the 28 analysed cancer types showed a significant association with prognosis (Table 6.1; $P < 0.05$; Log-rank test). These 26 subtypes were distributed across 17 cancer types, with ten showing a good prognosis and 16 a poor prognosis compared to the other subtypes in that cancer type (Fig. 6.1). Subtypes with differential prognosis for breast cancer, prostate cancer, lung adenocarcinoma and lung squamous cell carcinoma are examined in chapter 5.

6.3.2 Characterisation of differential prognosis subtypes in SKCM

Two LPD groups returned a differential prognosis in SKCM: LPD_1 was significantly associated with a poor prognosis ($P = 0.0063$; Log-rank test; Fig. 6.1), and LPD_3 with a good prognosis ($P = 0.0330$; Log-rank test; Fig. 6.1).

The cox regression analysis showed a significant association between LPD_1 gamma values and time to death (Hazard ratio = 8.83; $P < 0.001$). Likewise LPD_3 had a significant

Table 6.1: Log-rank test outcome to assess the differential prognosis for each subtype detected across the TCGA datasets. The prognosis status is provided, indicating whether a subtype is associated with a better or worse prognosis compared to other subtypes within the same cancer type. P-values from the log-rank test are also reported. Only the LPD groups with significantly differential prognosis are displayed, highlighting those subtypes that exhibit statistically significant variations in survival outcomes.

Cancer type	Differential prognosis process	Prognosis status	P-value
TCGA-BLCA	LPD_7	Good	0.0470
TCGA-BRCA	LPD_1	Poor	0.0190
TCGA-BRCA	LPD_7	Good	0.0150
TCGA-GBM	LPD_2	Poor	0.0170
TCGA-GBM	LPD_6	Poor	0.0470
TCGA-GBM	LPD_7	Poor	0.0290
TCGA-HNSC	LPD_4	Poor	0.0063
TCGA-HNSC	LPD_7	Good	0.0350
TCGA-KIRC	LPD_4	Poor	0.0001
TCGA-KIRC	LPD_6	Poor	0.0330
TCGA-KIRC	LPD_8	Good	0.0170
TCGA-KIRP	LPD_3	Poor	0.0460
TCGA-LAML	LPD_1	Poor	0.0190
TCGA-LAML	LPD_5	Good	0.0035
TCGA-LGG	LPD_3	Good	0.0360
TCGA-LGG	LPD_4	Poor	0.0001
TCGA-LUAD	LPD_5	Good	0.0019
TCGA-LUSC	LPD_3	Poor	0.0028
TCGA-PAAD	LPD_1	Good	0.0020
TCGA-PRAD	LPD_4	Poor	0.0160
TCGA-SARC	LPD_1	Poor	0.0014
TCGA-SKCM	LPD_1	Poor	0.0063
TCGA-SKCM	LPD_3	Good	0.0330
TCGA-THCA	LPD_7	Poor	0.0005
TCGA-THYM	LPD_1	Good	0.0320
TCGA-UCEC	LPD_3	Poor	0.0150

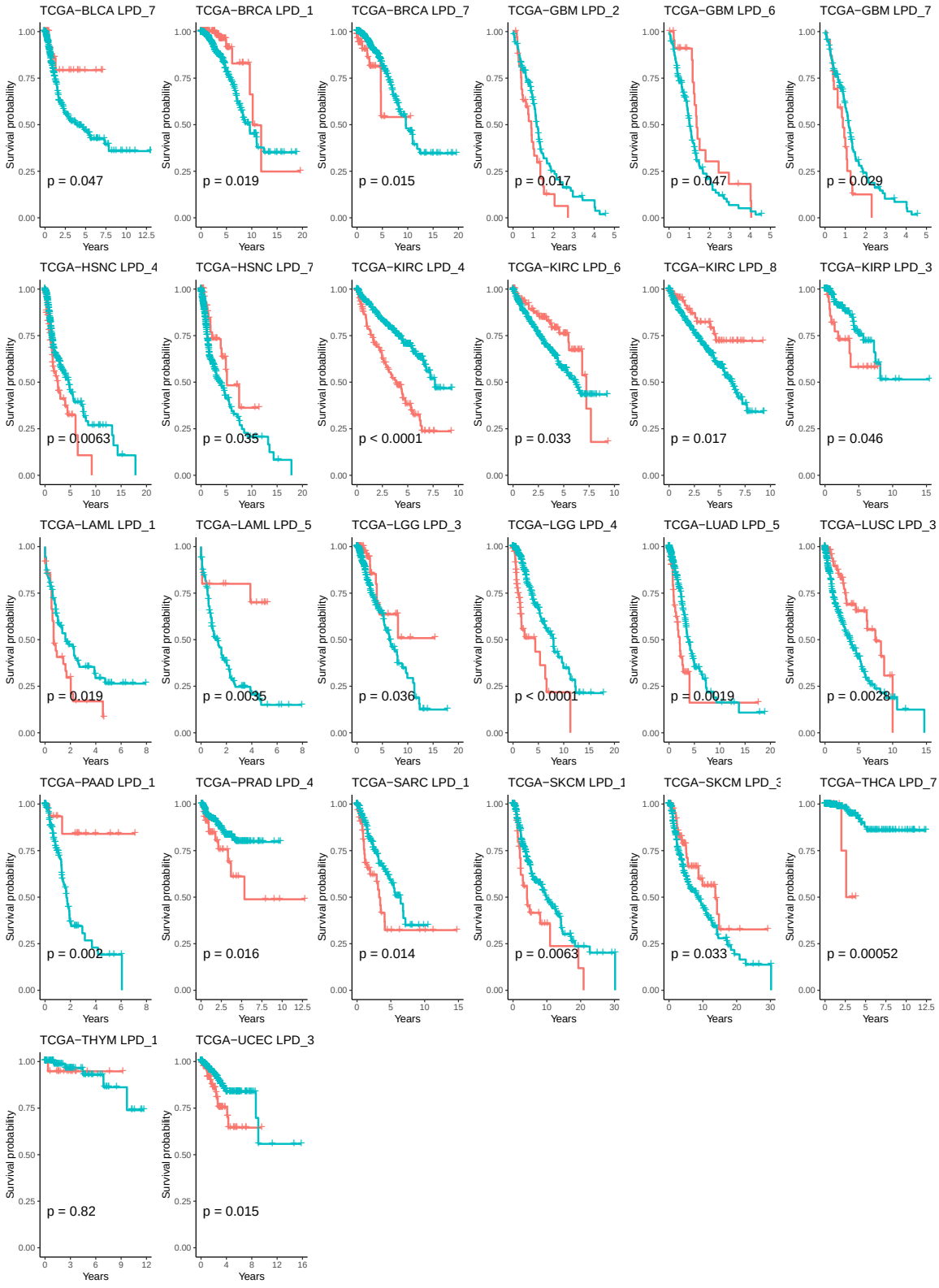

Figure 6.1: Kaplan-Meier estimator curves and log-rank tests to examine all the differential prognosis subtypes identified across the TCGA datasets. These curves compare the survival probability of each differential prognosis subtype (represented in red) against all other subtypes for the same cancer type (represented in blue). The log-rank test assesses the statistical significance of the differences in survival outcomes between the specific differential prognosis subtype and the remaining subtypes within the same cancer type.

association (Hazard ratio = 0.20; $P < 0.001$).

Exploring the LPD output for SKCM

The SKCM dataset consisted of 472 samples from 468 patients. Automata found six subtypes to be optimal, termed as LPD_1 ($n = 83$, 17.58%), LPD_2 ($n = 60$, 12.71%), LPD_3 ($n = 87$, 18.43%), LPD_4 ($n = 77$, 16.31%), LPD_5 ($n = 94$, 19.91%), and LPD_6 ($n = 71$, 15.04%)(Fig 6.2). Sample distribution into LPD groups was independent of the tissue source site (TSS) where they were processed ($P = 0.34$; Chi-squared test). There was no overrepresentation of benign tissue in any of the groups. When the mean gammas for each group were calculated, LPD_1 showed a robust assignment (mean gamma > 0.5) with minor overlap with LPD_2, whereas LPD_3 likewise showed a robust assignment with minor overlap with LPD_4 (Fig. 6.3).

Clinicopathologic characteristics of the differential prognosis subtypes in SKCM

Table 6.2 shows the clinicopathologic characteristics of the tumour samples. Patients assigned to LPD_1 and LPD_5 were significantly older, while the ones in LPD_6 were significantly younger ($P = 0.0005$; Chi-squared test). LPD_3 had no significant association with age. LPD_1 had a smaller proportion of patients that underwent radiation therapy ($P = 0.03$; Chi-squared test). LPD_3 had a bigger proportion of female patients than the other groups ($P = 0.009$; Chi-squared test). No differences were detected for the pathological stage.

Table 6.2: Clinicopathologic features of the detected subtypes for SKCM. Chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	P-value
Age (years; mean (sd))	59.5 (15.9)	62.6 (15.2)	61.6 (16.4)	57.2 (15.5)	58.8 (14.7)	63 (15.2)	53.1 (17.1)	0.0005
Race								
Asian	12	3	2	2	2	2	1	
White	449	78	55	82	74	90	70	
Black or african american	1	0	0	1	0	0	0	0.96
Gender								
Female	179	39	19	45	24	32	20	
Male	293	44	41	42	53	62	51	0.009
Pathological Stage								
Stage I	2	2	0	0	0	0	0	
Stage II	66	29	8	3	8	16	2	
Stage III	27	10	3	4	6	2	2	
Stage IV	3	2	0	0	0	1	0	0.5
Radiation therapy								
Yes	76	4	13	18	16	12	13	
No	350	64	42	63	56	69	56	0.03

Identification of differentially expressed genes in the differential prognosis subtypes in SKCM

Across the six subtypes, 7450 significant differentially expressed genes were identified with LPD_1 accounting for 780 and LPD_3 for 1983 (Table 6.3; median across groups = 1161;

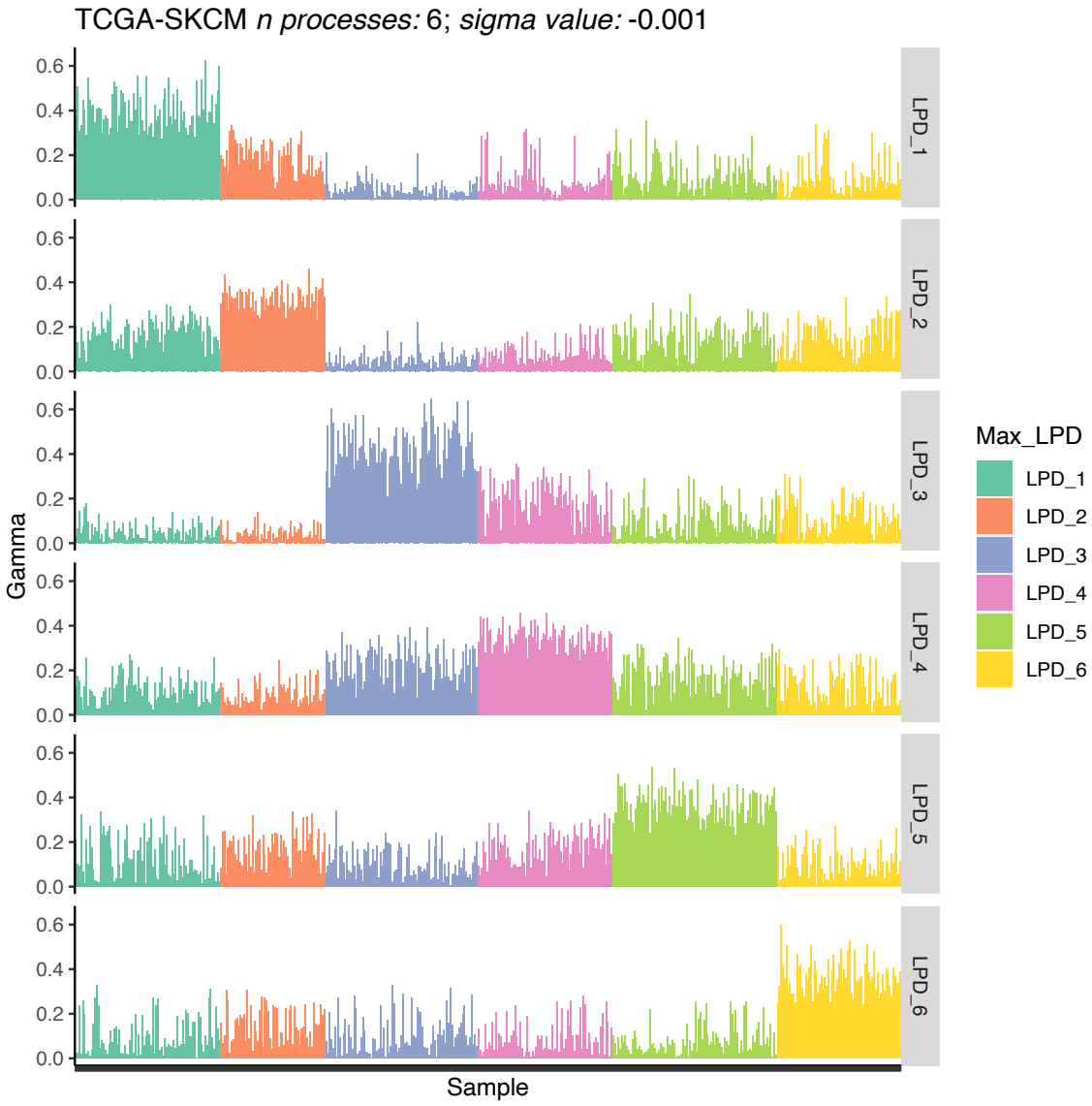

Figure 6.2: Gamma values of all samples for each detected LPD process in SKCM. A total of 6 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_6. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD signature is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

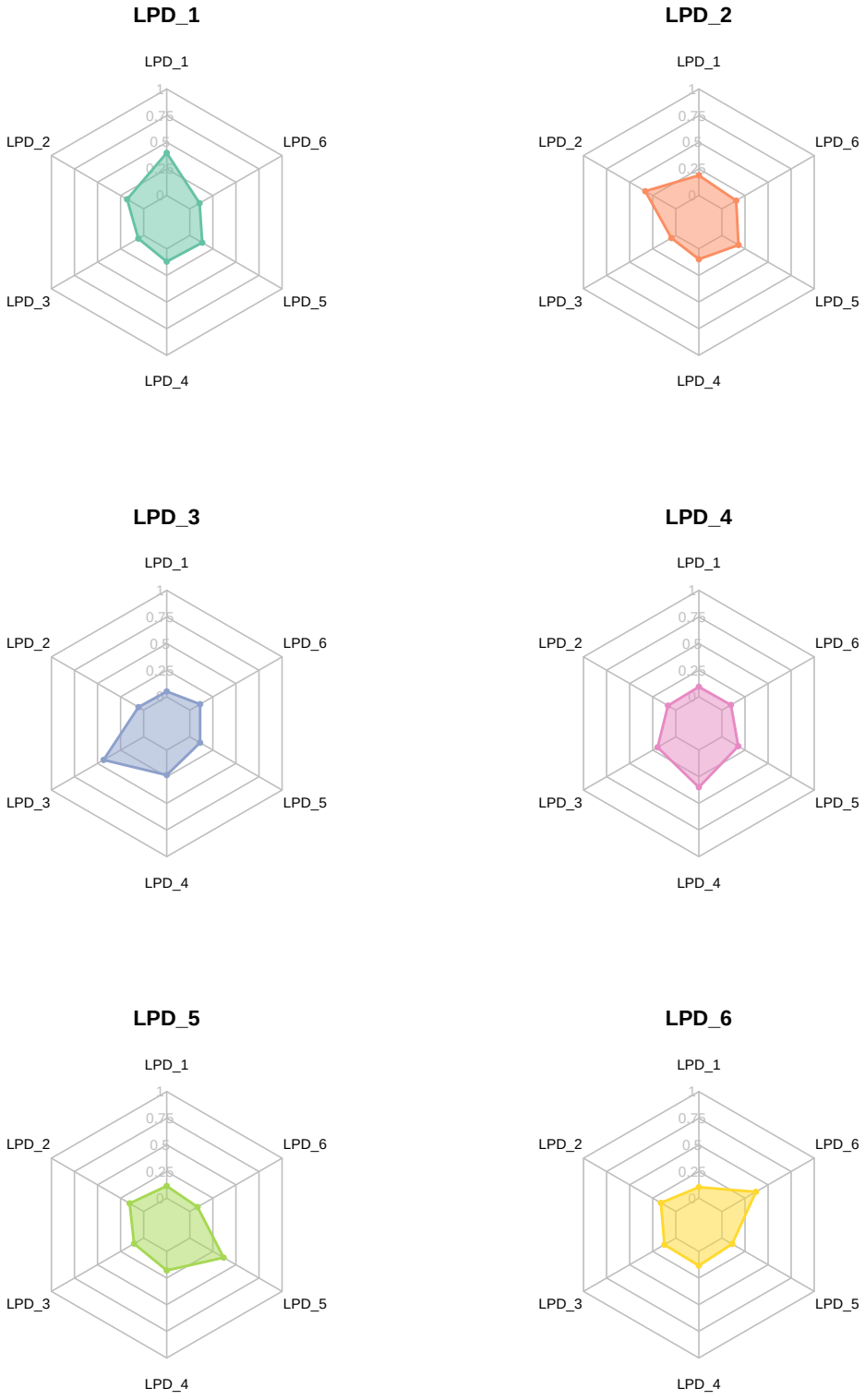

Figure 6.3: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in SKCM. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

IQR = 684). The most overexpressed and underexpressed genes in both groups are available in Table 6.4. LPD_1 was strongly weighted to underexpressed genes (Ratio = 0.13), whereas LPD_3 had a moderate weighting to underexpressed (Ratio = 0.64). The list of biological processes associated with the DEGs of both of this groups is shown in Figure 6.4.A for KEGG and Figure 6.5.A for GO. In KEGG, the groups shared opposing profiles affecting two pathways: LPD_1 showed an enrichment in underexpressed genes in *Staphylococcus aureus* infection and estrogen pathway, whereas these pathways consisted of overexpressed genes in LPD_3. This same phenomenon was also observed in GO enrichment analysis, in which LPD_1 showed enrichment of under expressed DEGs in epidermal cell differentiation, epidermis and skin development, while the same terms were found for LPD_3 but for over expressed DEGs. LPD_3 exhibited two associations with cancer hallmarks. Firstly, it exhibited an enrichment of biological processes related to invasion and metastasis. Secondly, LPD_3 displayed an association between underexpressed genes and tumour inflammation caused by tumoral cells in healthy cells. Out of the known driver genes *FGFR2*, *FGFR3*, *FOXQ1*, and *ZNF750* were found to be DEGs (over expressed in LPD_3 and under expressed in LPD_1).

Table 6.3: Gene counts for various categories in SKCM. These include the number of genes exhibiting significant differential expression and differential methylation, and the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6
n health tissue samples	0	0	1	0	0	0
DEGs						
Upregulated	93	69	779	164	75	164
Downregulated	687	1267	1204	1416	709	823
Total	780	1336	1983	1580	784	987
Ratio	0.135371179	0.054459353	0.647009967	0.115819209	0.105782793	0.19927096
DMGs						
Hypermethylated	928	5618	7687	118	1399	385
Hypomethylated	3163	3113	3996	55	517	5264
Total	4091	8731	11683	173	1916	5649
Ratio	0.293392349	1.80469001	1.923673674	2.145454545	2.705996132	0.073138298
Mutated						
Overmutated	622	630	254	1115	4212	806
Undermutated	2083	1995	2428	1537	683	1799
Total	2705	2625	2682	2652	4895	2605
Ratio	0.298607777	0.315789474	0.10461285	0.725439167	6.166910688	0.448026681

Table 6.4: The five genes more overexpressed (log2FoldChange > 1) and underexpressed (log2FoldChange < -1) for each LPD group associated with significantly differential prognosis in SKCM. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status

LPD_1		
LINC00347	2.839688	Overexpressed
NAP1L4P2	2.386833	Overexpressed
ENSG00000234713	2.213128	Overexpressed
CA6	2.137206	Overexpressed
PRB3	2.121801	Overexpressed
FLG2	-6.173211	Underexpressed
LCE3D	-5.793414	Underexpressed
WFDC12	-5.584058	Underexpressed
KRT71	-5.077534	Underexpressed
C1orf68	-5.064496	Underexpressed
LPD_3		
KRT71	5.912358	Overexpressed
KRT25	5.367328	Overexpressed
LOR	4.946126	Overexpressed
AADACL3	4.929217	Overexpressed
LCE3C	4.730980	Overexpressed
OTOR	-6.388640	Underexpressed
HTN3	-5.117165	Underexpressed
SMR3B	-4.715372	Underexpressed
SFTPA2	-4.564585	Underexpressed
SFTPB	-4.459785	Underexpressed

Identification of differentially methylated genes in the differential prognosis subtypes in SKCM

A median of 4870 significant DMGs associated with a subtype, with an IQR of 5500. LPD_1 had a predominance of genes that were hypomethylated (Ratio = 0.29) whereas LPD_3 showed a moderate weighting towards hypermethylation (Ratio = 1.92) (Table 6.3). In the enrichment analysis of KEGG and GO, LPD_1 and LPD_3 showed distinct profiles, with LPD_3 showing the strongest effect size (measured as gene ratio) for the hypermethylation of olfactory transduction. The driver genes affected by DMG showed again an opposing profile for several genes, with LPD_1 enriched in hypomethylation and LPD_3 enriched in hypermethylation (Fig. 6.6).

Identification of genes affected by single nucleotide genes in the differential prognosis subtypes in SKCM

A median of 2667 genes were enriched or depleted by SNVs, with an IQR of 68. In both, LPD_1 and LPD_3 there was a strong weighting towards depleted genes (Ratio = 0.29 and 0.1 respectively). No pathways were detected as affected by SNV in both KEGG and GO enrichment analysis. The driver genes profile was distinct although both of them were dominated by undermutations (Fig. 6.6). No differences were observed across groups when comparing the variant type, variant class, and SNP type (Fig. 6.7). All LPD groups showed a similiar COSMIC signature profile (Fig. 6.8).

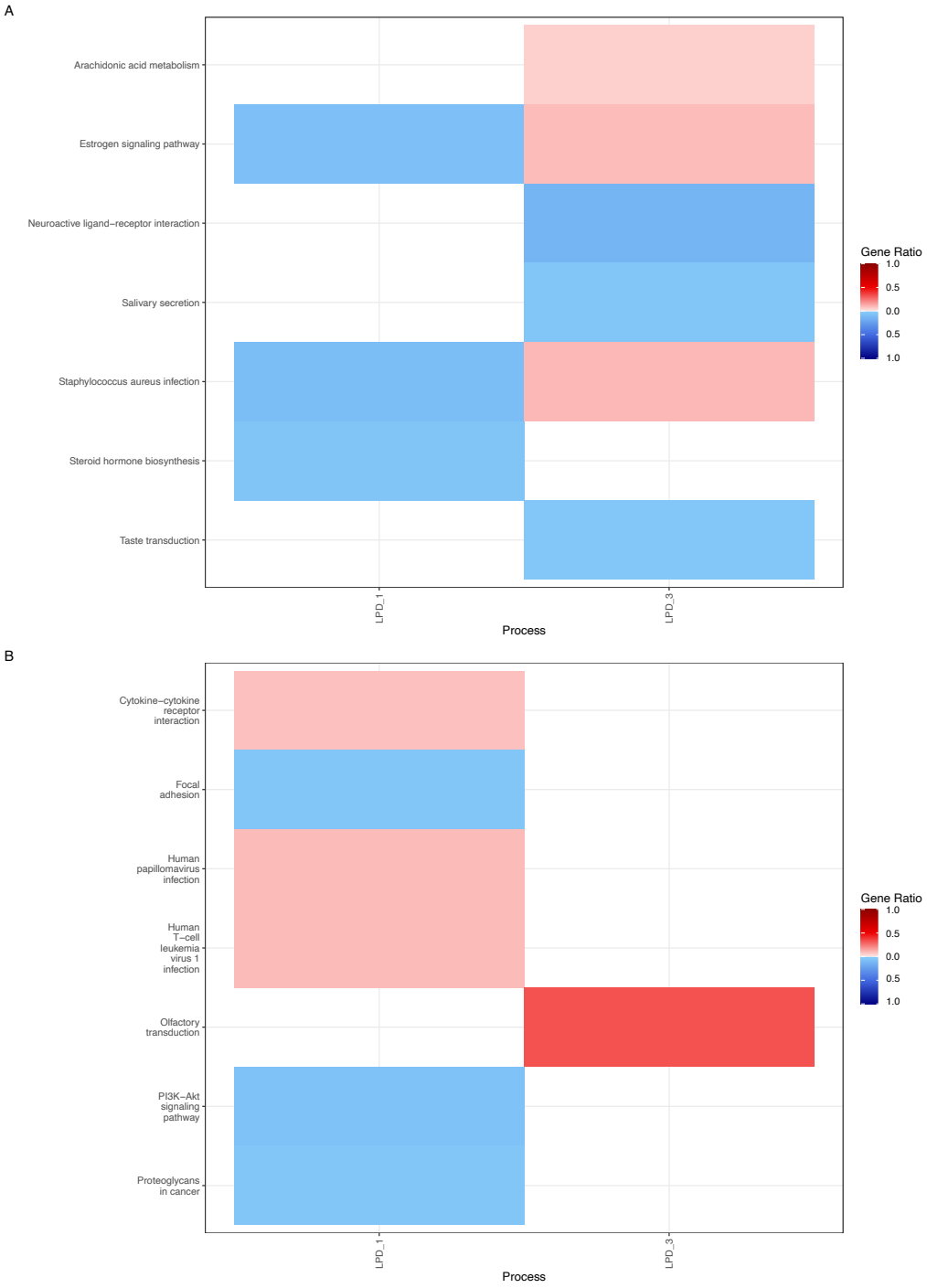

Figure 6.4: Biological pathways associated with different categories in SKCM determined using KEGG enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, biological pathways associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. Only LPD groups associated to significantly differential prognosis are displayed. The complete list of associated biological pathways is available in Supplementary Material B.

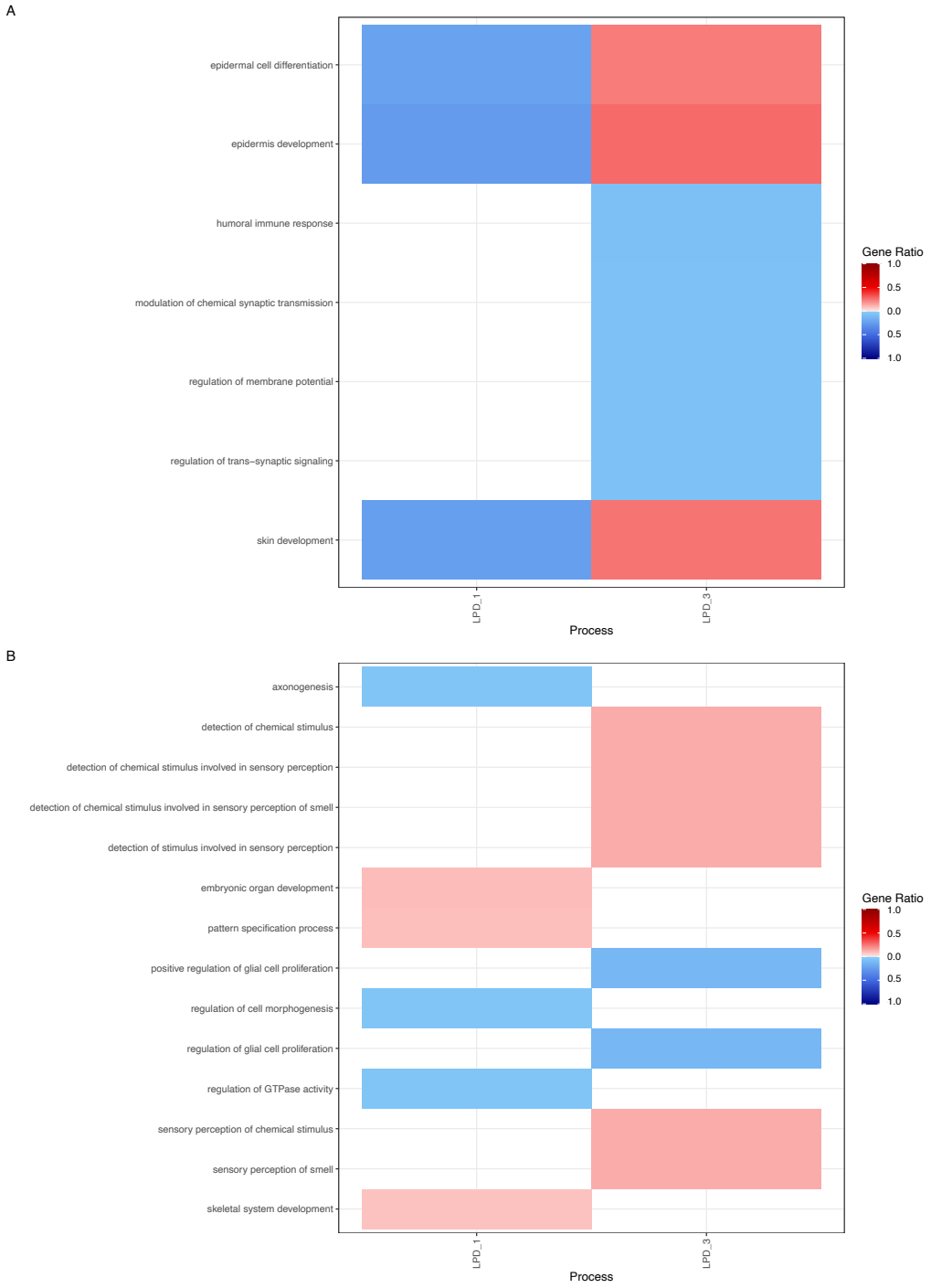

Figure 6.5: Biological processes associated with different categories in SKCM determined using GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant processes with the highest gene ratio are displayed. (A) Differentially expressed genes, processes associated with overexpressed genes are highlighted in red, while processes linked to underexpressed genes are shown in blue. (B) Differentially methylated genes, hypermethylated genes are depicted in red while hypomethylated are represented in blue. Only LPD groups associated to significantly differential prognosis are displayed. The complete list of associated biological processes is available in Supplementary Material B.

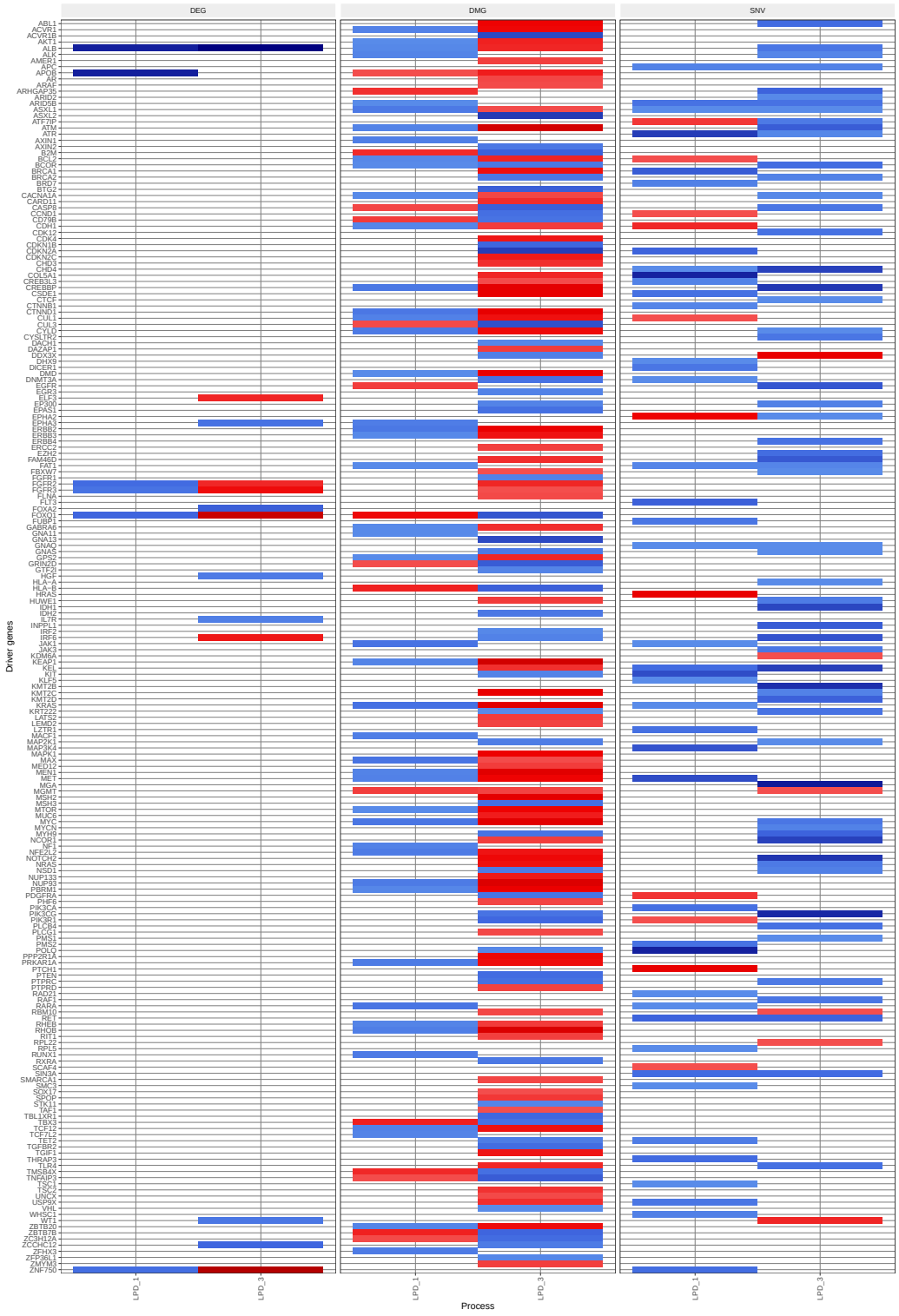

Figure 6.6: Heatmap showing the presence of driver genes across different categories in SKCM, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV are represented in red, while genes with opposite characteristics are depicted in blue. Only LPD groups associated to significantly differential prognosis are displayed.

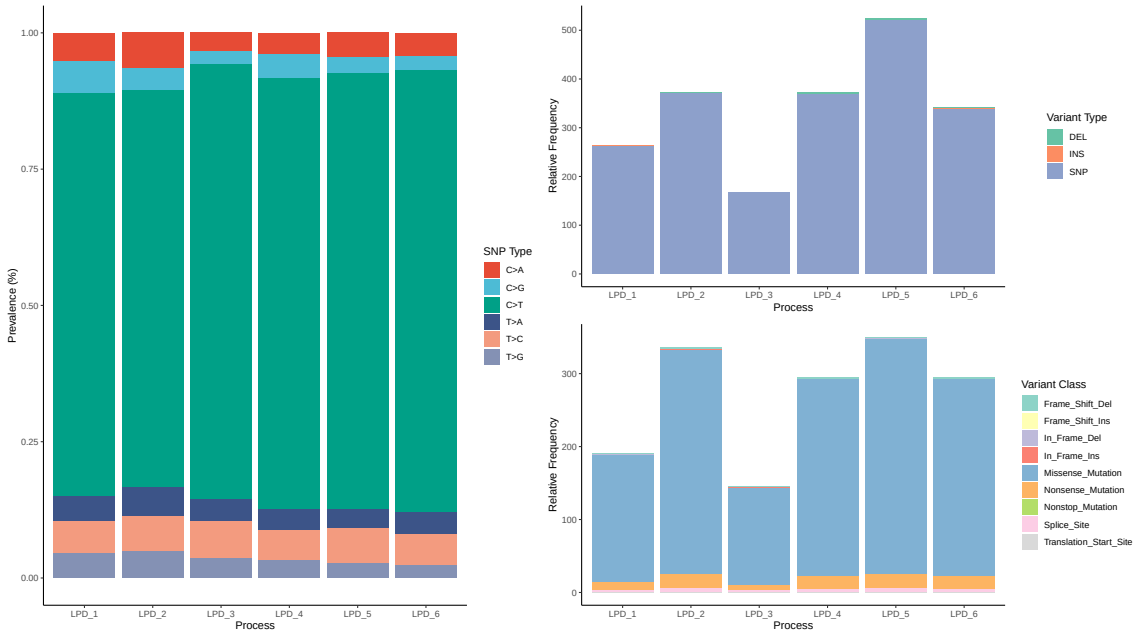

Figure 6.7: Detected single nucleotide variants (SNVs) within each LPD group for SKCM. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

Functional analysis of the differential prognosis subtypes in SKCM

The matches between differentially expressed genes, differentially methylated genes, and genes affected by SNVs variations are shown in Figure 6.9 for overexpressed DEGs, and in Figure 6.10 for underexpressed DEGs. In LPD_1 6 genes were hypomethylated and overexpressed and 34 genes were two of being underexpressed, hypermethylated or enriched in mutations. In LPD_3 90 genes were hypomethylated and overexpressed and two genes were underexpressed, hypermethylated and enriched in mutations (240 had at least two). LPD_3 was enriched in matches for overexpressed genes in comparison to LPD_1, while no differences were observed between both groups for underexpressed genes. The complete list of matched genes is available in Supplementary Material B.

Comparison of the LPD output in SKCM with Euclidean hierarchical clustering

The Euclidean hierarchical clustering returned a dendrogram with very mixed samples with no clear distinctions between the LPD groups (Fig. 6.11). The blue cluster was mostly composed of LPD_1 samples, whereas the remaining clusters showed assignments to at least three LPD groups. So LPD_3 was a novel subtype associated with a good prognosis that was not picked up by traditional clustering techniques.

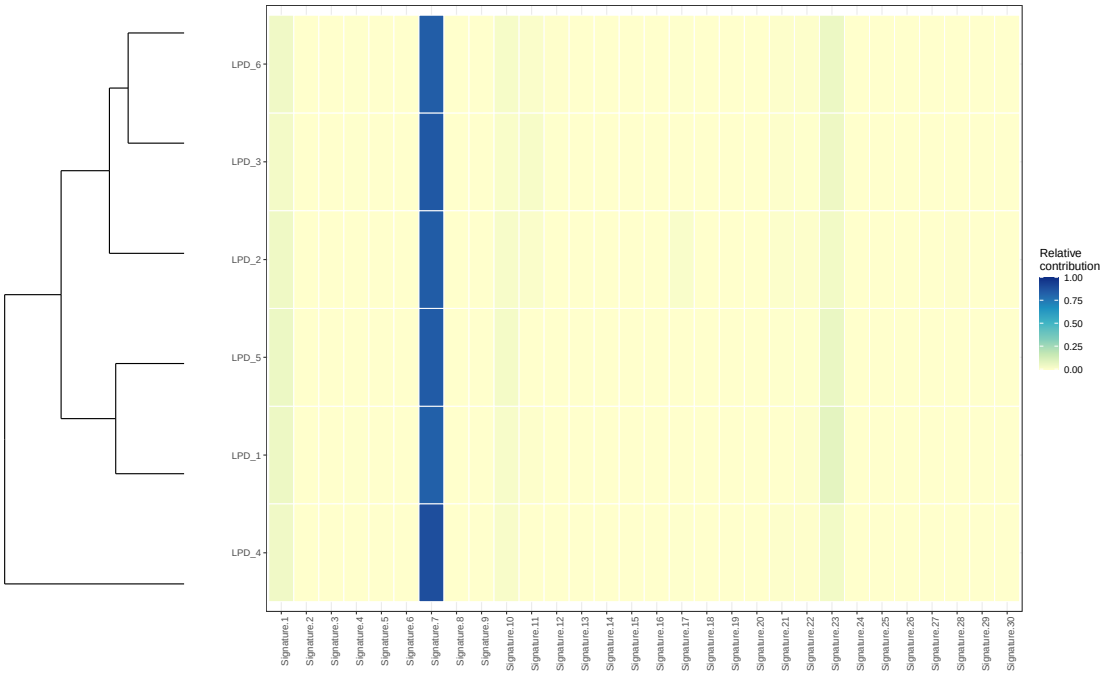

Figure 6.8: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in SKCM. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

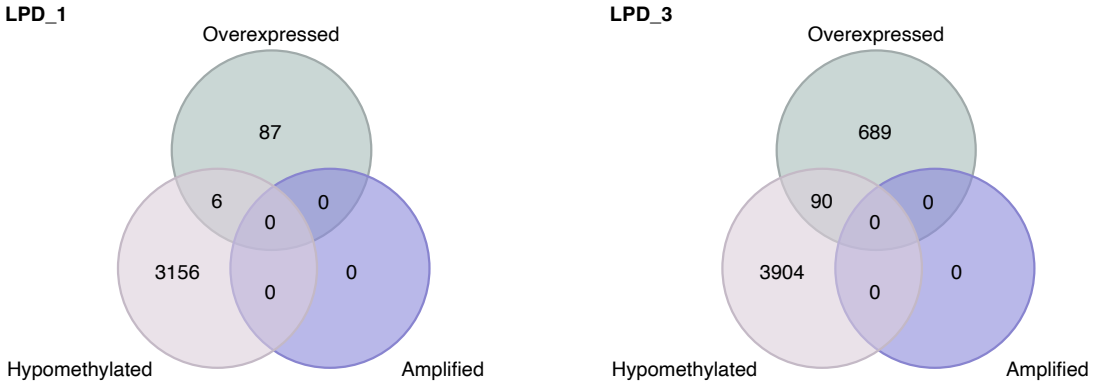

Figure 6.9: Venn diagram displaying the overlaps between three categories in genes in SKCM for each LPD group associated to significantly differential prognosis: overexpressed genes, hypomethylated genes, and amplified genes. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit overexpression, hypomethylation, and amplification simultaneously.

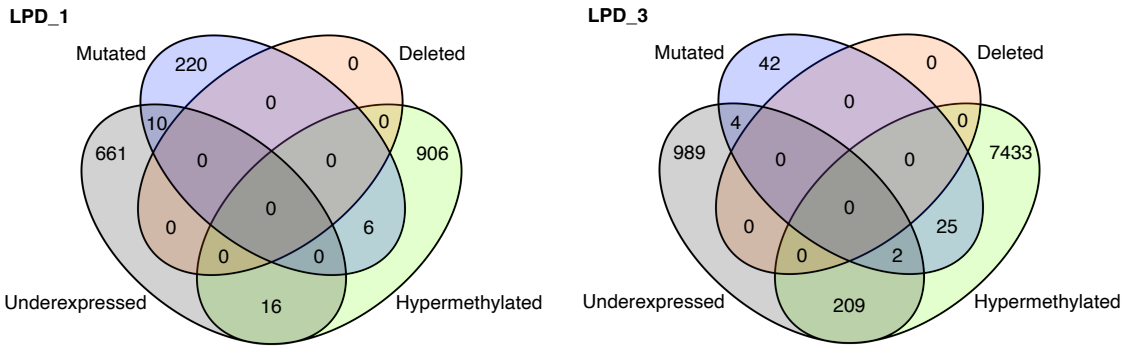

Figure 6.10: Venn diagram displaying the overlaps between four categories in genes in SKCM for each LPD group associated to significantly differential prognosis: underexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit underexpression, hypermethylation, mutations and deletions.

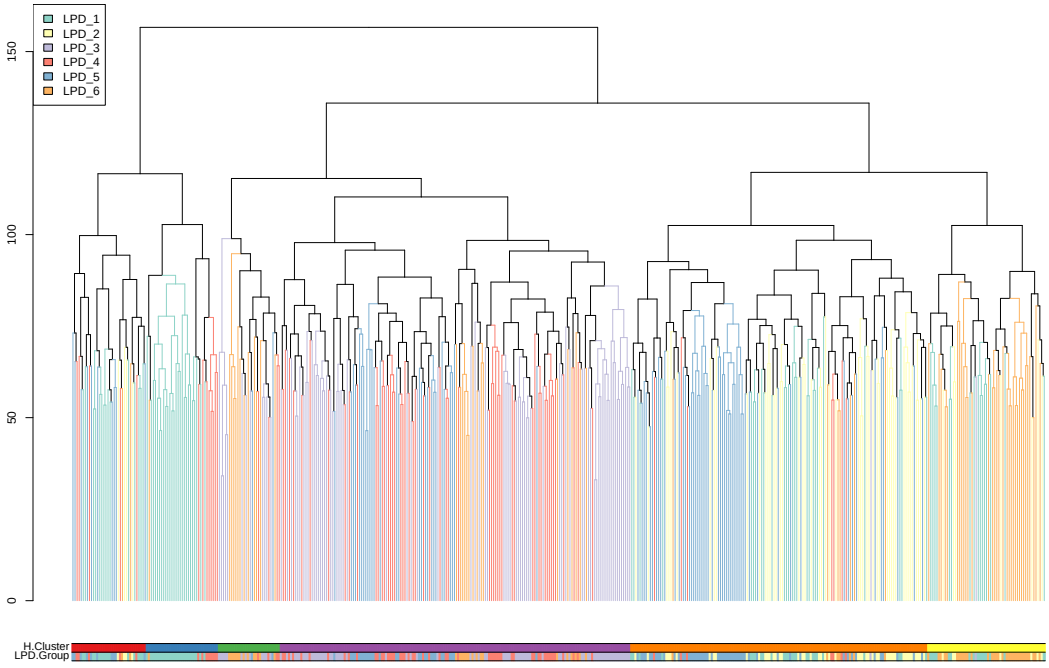

Figure 6.11: Dendrogram showing the sorting of the SKCM samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into six groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

6.3.3 Characterisation of differential prognosis subtypes in BLCA

BLCA only had one group with differential prognosis, LPD_7, which showed a statistically significant association with a good prognosis ($P = 0.047$; Log-rank test; Fig. 6.1) in comparison to the other subtypes. The cox regression analysis did not show a significant association between LPD_7 gamma values and time to death (Hazard ratio = 0.25; $P = 0.09$).

Exploring the LPD output for BLCA

The BLCA dataset consisted of 427 samples from 408 patients. LPD split the samples into eight LPD groups, termed as LPD_1 ($n = 45$, 10.5%), LPD_2 ($n = 50$, 11.7%), LPD_3 ($n = 78$, 18.3%), LPD_4 ($n = 60$, 14.1%), LPD_5 ($n = 44$, 10.3%), LPD_6 ($n = 75$, 17.6%), LPD_7 ($n = 36$, 8.43%), and LPD_8 ($n = 39$, 9.13%) (Fig 6.12). Sample distribution into LPD groups was independent of TSS ($P = 0.31$). LPD_6 was enriched in benign tissue samples ($P = 0.0001$; Chi-squared test). When the mean gammas for each group were calculated, LPD_7 showed a shared assignment with LPD_2 (Fig. 6.13).

Clinicopathologic characteristics of the differential prognosis subtypes in BLCA

Table 6.5 shows the clinicopathologic characteristics of the tumour samples. LPD_7 was significantly enriched in non-papillary samples ($P = 0.02$; Chi-squared test) but no other clinical variables. Patients assigned to LPD_5 and LPD_6 were significantly older, while the ones in LPD_2 were significantly younger ($P < 0.0001$; Chi-squared test). LPD_2 showed the most distinct clinical profile with enrichment of Asian patients ($P = 0.0004$; Chi-squared test), stage II samples ($P = 0.0009$; Chi-squared test), low histological grade ($P = 0.0004$; Chi-squared test), and Papillary samples ($P = 0.002$; Chi-squared test). Chi-squared returned a significant enrichment for gender ($P = 0.009$), however, the post-hoc tests of individual subtypes did not return anything significant.

Table 6.5: Clinicopathologic features of the detected subtypes for BLCA. Chi-squared test was conducted for each category across LPD groups. The resulting p-values are displayed.

Variable	All	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	LPD_8	P-value
Age (years; mean (sd))	69.6 (10.8)	69.1 (11.3)	63.1 (11)	71.7 (10.5)	70 (11.1)	71.4 (8.45)	72 (10.1)	66.1 (12.1)	70.5 (9.02)	< 0.0001
Race										
Asian	44	1	19	10	2	2	2	7	1	
Black or african american	24	4	2	5	4	4	4	1	0	
White	341	36	26	62	49	38	68	25	37	0.0004
Gender										
Female	116	13	9	18	23	5	29	6	13	
Male	311	32	41	60	37	39	46	30	26	0.009
Pathological Stage										
Stage I	2	0	1	0	0	0	0	1	0	
Stage II	130	15	31	21	17	9	14	17	6	
Stage III	140	20	11	24	19	18	24	11	13	
Stage IV	134	10	3	32	23	17	27	7	15	0.0009
Radiation therapy										
High	403	45	34	75	59	43	75	33	39	
Low	21	0	16	1	1	0	0	3	0	0.0004
Non-Papillary	285	37	13	50	47	32	59	15	32	
Papillary	137	8	37	28	11	11	15	21	6	0.002

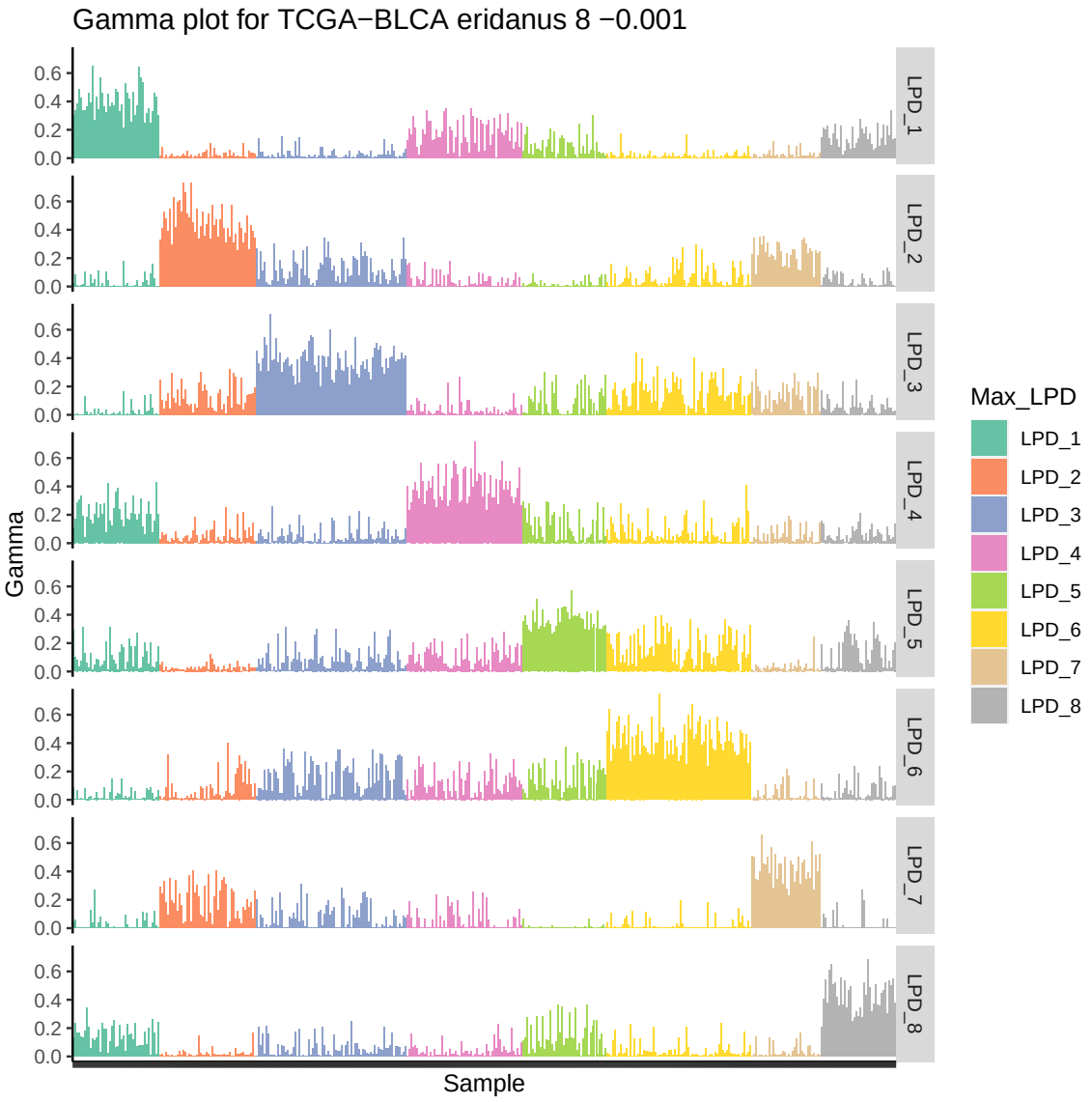

Figure 6.12: Gamma values of all samples for each detected LPD process in BLCA. A total of 8 processes were detected, each one represented in a horizontal lane numbered from LPD_1 to LPD_8. The gamma value of each sample to each process is represented as a barplot along the lanes. The colour of the sample indicates the which LPD signature is more dominant in the sample, and therefore to which LPD group the sample is assigned to.

Figure 6.13: Radar plot illustrating the mean gamma values for the samples allocated to each LPD group in BLCA. Each LPD group is represented by a separate axis on the radar plot, and the mean gamma value for each group is depicted as a data point along the respective axis.

Identification of differentially expressed genes in the differential prognosis subtypes in BLCA

Across the eight LPD groups, 10566 significant differentially expressed genes were identified (Table 6.6; median across groups = 1227; IQR = 674) with LPD_7 accounting for 1312. The most overexpressed and underexpressed genes in LPD_7 are available in Table 6.7. LPD_7 had a heavy weighting towards underexpressed DEGs (Ratio = 0.19). Eight driver genes were found as DEG in LPD_7, from which seven were underexpressed with *ALB* showing the highest *log2 fold change* (Fig. 6.15). KEGG enrichment analysis showed an overexpression of cytokine-cytokine receptor interaction and renin-angiotensin system, and an underexpression of chemical carcinogenesis, retinol metabolism, metabolism of xenobiotics and nueroactive ligand-receptor interaction (Figure 6.14.A). GO enrichment reiterated the overexpression renin-angiotensis processes and showed an unexpression of skin and epidermis development (Figure 6.14.B). No associations to cancer hallmarks were detected.

Table 6.6: Gene counts for various categories in BLCA. These include the number of genes exhibiting significant differential expression and differential methylation, and the count of genes showing a significantly higher or lower frequency of SNV mutations (referred to as overmutated and undermutated, respectively). The ratio of genes between these different gene statuses and the total gene count in each category are calculated. The number (n) of healthy samples within each LPD group is also provided.

Gene count	LPD_1	LPD_2	LPD_3	LPD_4	LPD_5	LPD_6	LPD_7	LPD_8
n health tissue samples	0	3	0	1	0	10	0	5
DEGs								
Upregulated	1271	362	52	273	58	149	211	87
Downregulated	1054	2055	996	868	390	379	1101	1260
Total	2325	2417	1048	1141	448	528	1312	1347
Ratio	1.205882353	0.176155718	0.052208835	0.314516129	0.148717949	0.393139842	0.19164396	0.069047619
DMGs								
Hypermethylated	3626	2539	1199	852	4973	461	1743	8417
Hypomethylated	2109	10446	4074	1174	481	203	11668	1307
Total	5735	12985	5273	2026	5454	664	13411	9724
Ratio	1.719298246	0.243059544	0.294305351	0.72572402	10.33887734	2.270935961	0.149382928	6.439938791
Mutated								
Overmutated	311	123	1602	673	257	558	416	140
Undermutated	449	391	329	441	419	462	503	406
Total	760	514	1931	1114	676	1020	919	546
Ratio	0.692650334	0.314578005	4.869300912	1.526077098	0.613365155	1.207792208	0.827037773	0.344827586

Table 6.7: The five genes more overexpressed (log2FoldChange > 1) and underexpressed (log2FoldChange < -1) for each LPD group associated with significantly differential prognosis in BLCA. The complete list of genes is available in Supplementary Material B.

Gene	log2FoldChange	Status
LPD_7		
TRH	6.237348	Overexpressed
CGB5	5.806779	Overexpressed
FGF4	5.767181	Overexpressed
NOTUM	4.571882	Overexpressed

CGB3	4.120817	Overexpressed
FABP1	-6.587514	Underexpressed
MAGEA1	-5.594863	Underexpressed
CRNN	-5.191205	Underexpressed
GC	-5.157953	Underexpressed
APOA2	-5.112752	Underexpressed

Identification of differentially methylated genes in the differential prognosis subtypes in BLCA

A median of 5595 genes were found DMG across all groups, with an IQR of 6078 (Table 6.6). LPD_7 had 13411 DMGs and showed a strong weighting to DMGs being hypomethylated (Ratio = 0.14). 147 DMGs were also driver genes for LPD_7, with 122 of them being hypomethylated (Fig. 6.15). No pathways were detected as enriched by KEGG. GO analysis revealed three hypermethylated biological processes related to embrionic development and four hypomethylated processes associated to cell morphogenesis, neuron development and cation regulation (Figure 6.14.C).

Identification of genes affected by single nucleotide genes in the differential prognosis subtypes in BLCA

A median of 840 genes were enriched or depleted in SNVs in a subtype compared to the rest, with an IQR of 400. LPD_7 had roughly equal enriched and depleted genes (Ratio = 0.82). 37 driver genes were detected across the genes affected by SNVs, from those 32 being undermutated (Fig. 6.15). No pathways or biological processes were detected as affected by SNV in both KEGG and GO enrichment analysis. No differences were observed across groups when comparing the variant type, variant class, and SNP type (Fig. 6.16). In the COSMIC profiling, LPD_7 was the only group to show a contribution from signature 10 and its proportion of signature 13 was lower than the one in the other groups (Fig. 6.17).

Functional analysis of differentially expressed genes in BLCA

The matches between DEGs, DMGs and genes affected by SNVs are gathered in Figure 6.18. 74 matches were detected between overexpressed genes and hypomethylated, while a total of 45 matches were found for underexpressed, mutated and hypermethylated genes.

Comparison of the LPD output in BLCA with Euclidean hierarchical clustering

The Euclidean hierarchical clustering returned a very heterogeneous dendrogram with no clear distinction between the LPD groups and the hierarchical clusters (Fig. 6.19).

6.4 Discussion

In this chapter, I have successfully identified 26 subtypes that have a significant association with time to survival probability across 17 different cancer types from the TCGA. From those, I have characterised two subtypes in skin cutaneous melanoma and one in bladder cancer.

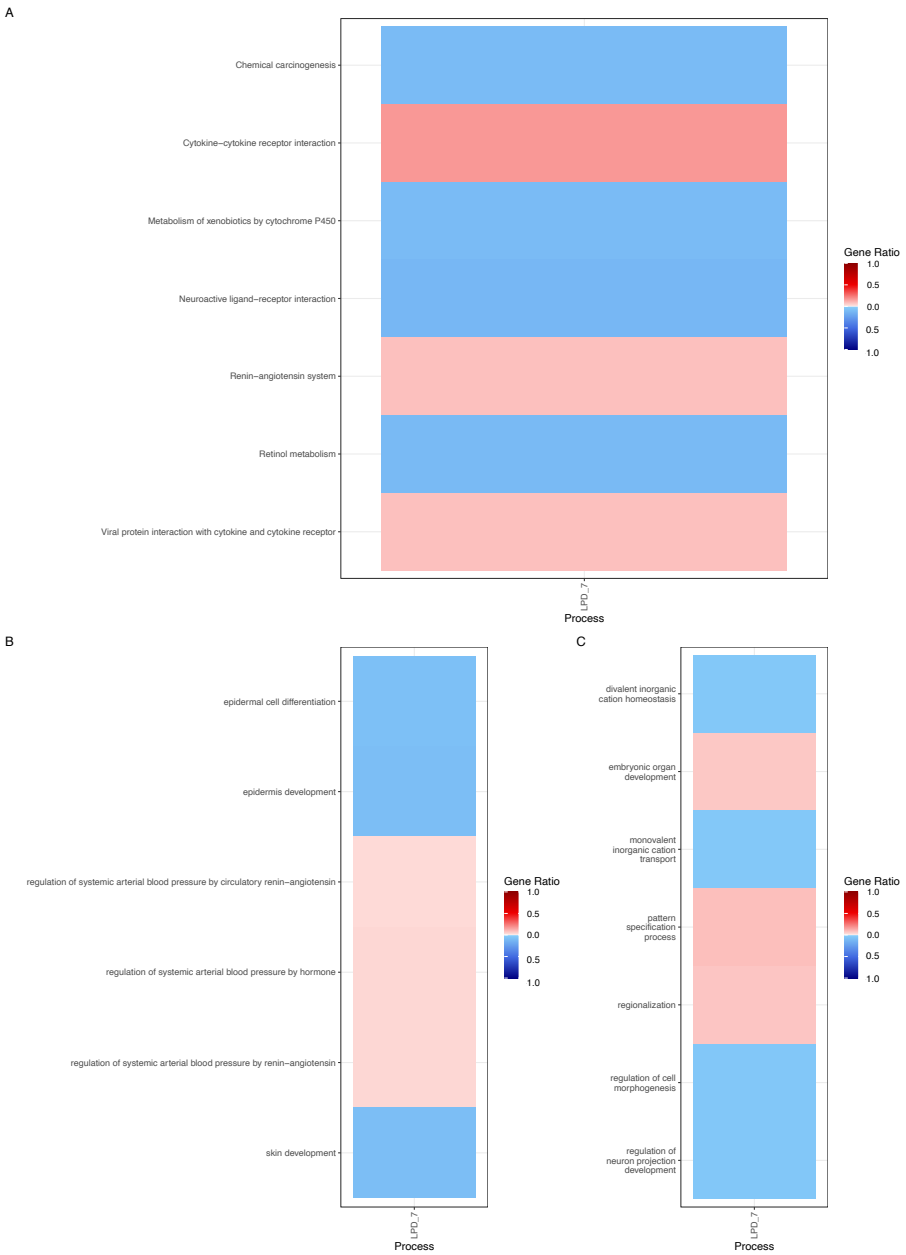

Figure 6.14: Biological pathways and processes associated with different categories in BLCA determined using KEGG and GO enrichment analysis. The gene ratio quantifies how many genes are associated to the processes. Only significant pathways and processes with the highest gene ratio are displayed. (A) Enrichment of differentially expressed genes according to KEGG, pathways associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (B) Enrichment of differentially expressed genes according to GO, pathways associated with overexpressed genes are highlighted in red, while pathways linked to underexpressed genes are shown in blue. (C) Enrichment of differentially methylated genes according to GO, biological processes associated to hypermethylated genes are depicted in red while hypomethylated are represented in blue. Only LPD groups associated to significantly differential prognosis are displayed. The complete list of associated biological pathways and processes is available in Supplementary Material B.

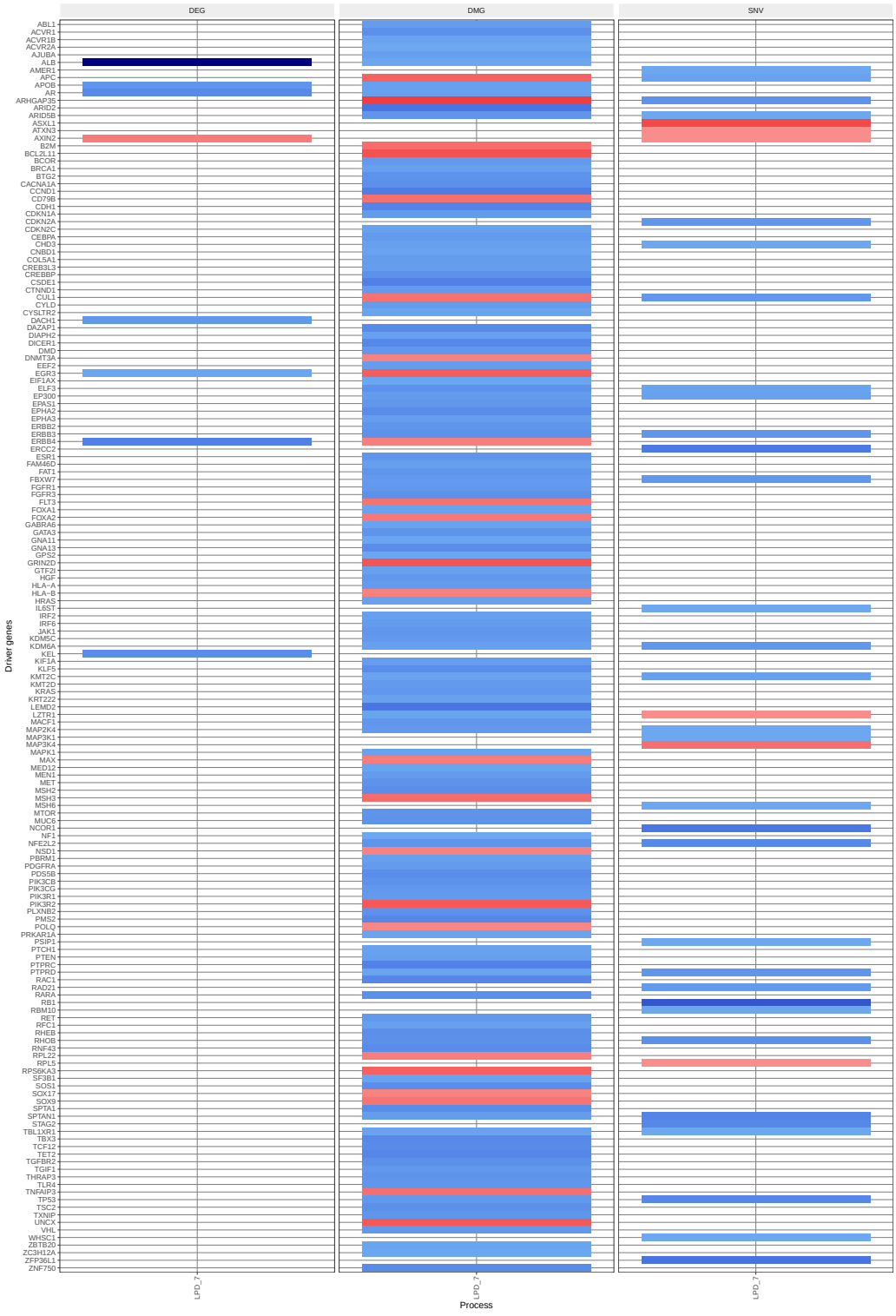

Figure 6.15: Heatmap showing the presence of driver genes across different categories in BLCA, including differentially expressed genes (DEG), differentially methylated genes (DMG), and genes affected by single nucleotide variants (SNV). Significantly overexpressed genes, hypermethylated genes, and genes more frequently altered by SNV are represented in red, while genes with opposite characteristics are depicted in blue. Only LPD groups associated to significantly differential prognosis are displayed.

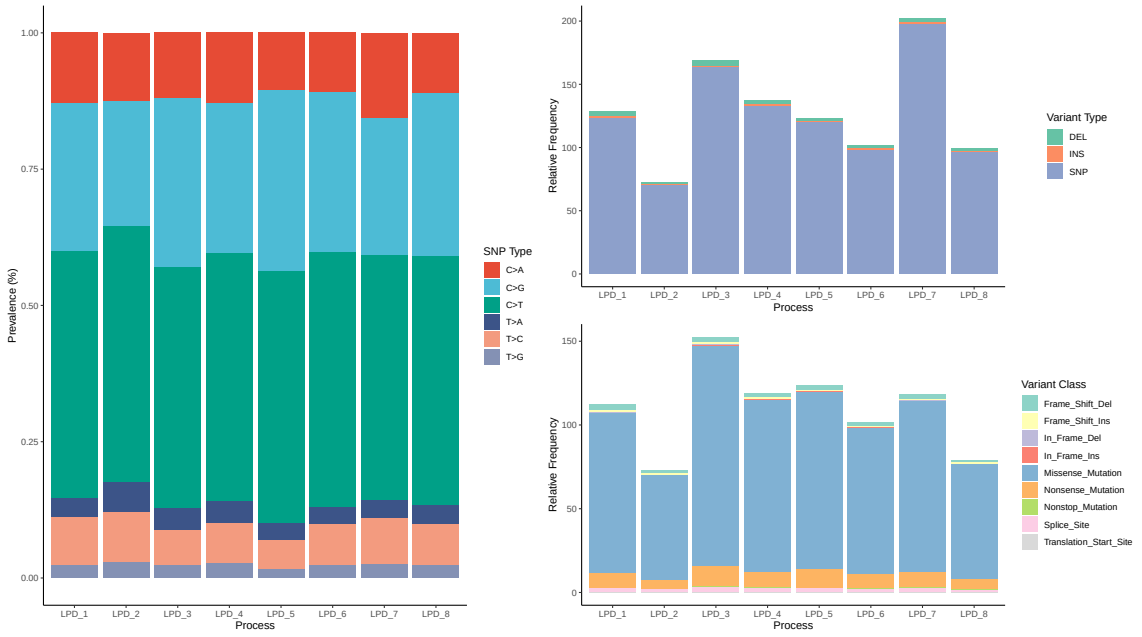

Figure 6.16: Detected single nucleotide variants (SNVs) within each LPD group for BLCA. It consists of three separate barplots showcasing the proportion of each category within each group, including SNP type, variant type (DEL for deletion, INS for insertion, SNP for point mutation), and variant class.

In skin cutaneous melanoma, LPD_1 and LPD_3 showed opposing expression and survival probability profiles, with LPD_1 having a good prognosis while LPD_3 was identified as a poor prognosis subtype. LPD_3 was a novel subtype associated with a good prognosis that was not picked up by traditional clustering techniques. LPD_1 was defined as a subtype with low number of patients undergoing radiation therapy, with underexpression of *Staphylococcus aureus* infection, estrogen pathway, and epidermis development, and hypomethylation of embryonic and skeletal development, and hypermethylation of cell morphogenesis regulation. The significantly differentially expressed genes with the largest change for this subtype are the underexpressed genes *FLG2*, *LCE3D*, *WFDC12*, *KRT71*, *C1orf68*. *FLG2* is involved in the maintenance of the epithelial homeostasis[316]. *LCE3D* is involved in the keratinization and nervous system development, and its depletion is associated to psoriasis risk[317]. *WFDC12* codes a protease inhibitor protein. *KRT71* is involved in the root of the hair follicles in the epidermis. *C1orf68* encodes a skin-specific protein and was reported as an upregulated gene in melanoma development[318]. In addition, the driver gene *APOB* was detected as underexpressed and hypermethylated.

LPD_3 was defined as a subtype with enrichment of female patients with overexpression of *Staphylococcus aureus* infection, estrogen pathway, arachidonic acid metabolism, epidermis development, and processes associated to invasion and metastasis; underexpression of taste transduction and salivary secretion, regulation of synapsis, humoral immune response, and hypermethylation of olfactory decection of stimulus; and hypomethylation of glial cell proliferation. The differentially expressed genes for this subtype that show the largest changes are the underexpression of *OTOR* and *HTN3*, and the overexpression of *KRT71*, *KRT25*

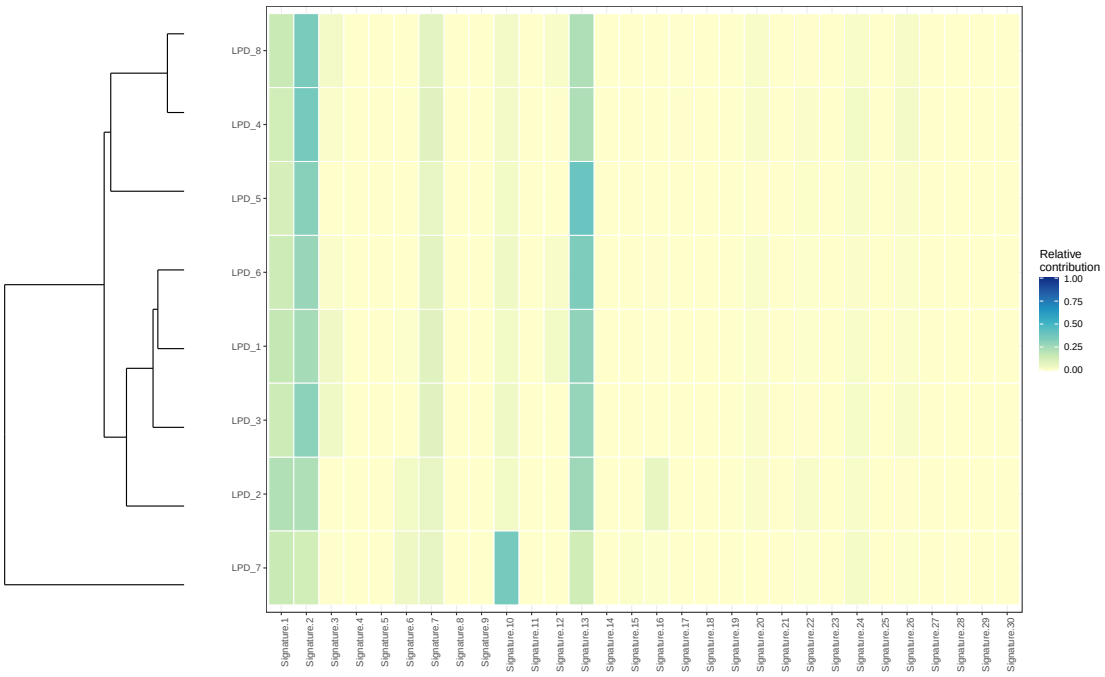

Figure 6.17: Heatmap representing the relative contribution of the COSMIC signatures to each LPD group found in BLCA. The LPD groups are sorted using hierarchical clustering, therefore groups with similar profiles are placed side by side. A summary of each of the COSMIC signatures can be found in Appendix B.

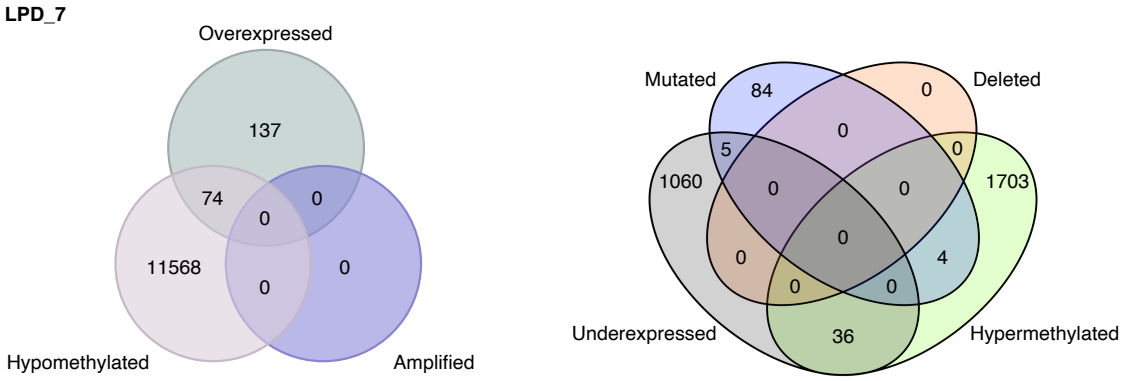

Figure 6.18: Venn diagram displaying the overlaps across multiple categories in genes in BLCA for LPD_7. In the left, overexpressed genes, hypomethylated genes, and amplified genes. In the right,underexpressed genes, hypermethylated genes, mutated genes by SNVs, and genes deleted by CNV. The diagram illustrates the shared genes among these categories, highlighting the common genes that exhibit simultaneously either underexpression, hypermethylation, mutations and deletions, or overexpression, hypomethylation, and amplification.

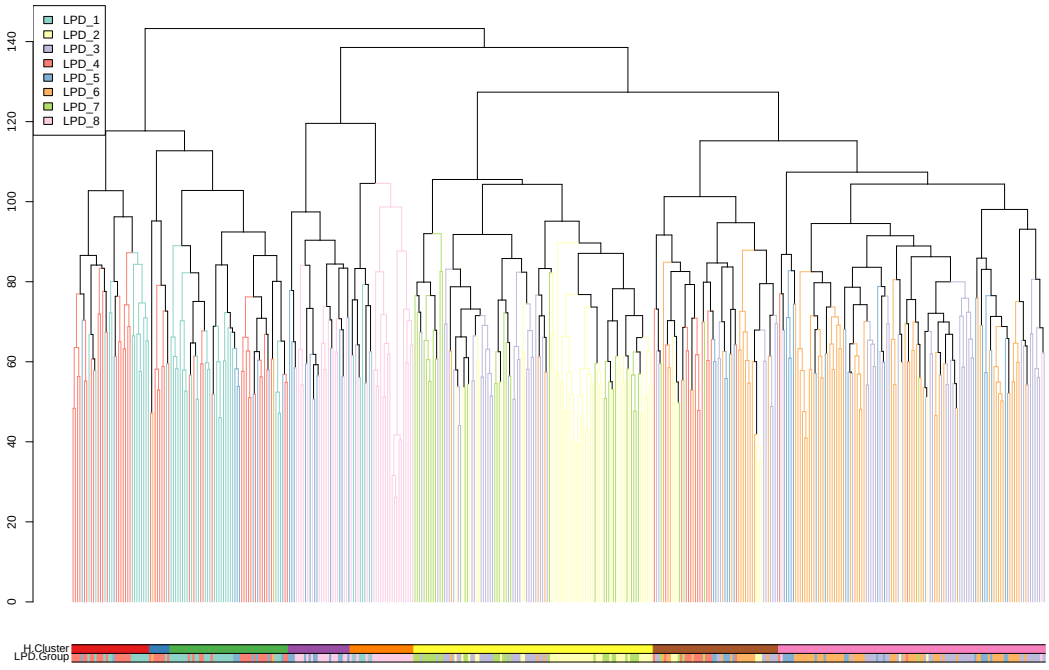

Figure 6.19: Dendrogram showing the sorting of the BLCA samples using Euclidean hierarchical clustering. Samples with similar expression profile are arranged side by side. At the bottom, two bars provide additional information: the first bar represents the classification of samples into eight groups (H.Clusters) based on hierarchical clustering, while the second bar indicates the allocation of the samples into LPD groups (LPD.Group) determined by the LPD algorithm. The dendrogram branches are colour-coded according to the corresponding LPD group.

and *LOR*. *OTOR* is a gene that participates in the inhibition of melanoma[319]. *HTN3* is involved in the antimicrobial resistance of the immune system, and, along with the gene *MSANTD3*, is reported as a potential biomarker to distinguish salivary gland tumours[320]. Additionally, this gene was detected as differentially expressed in tumoral skin tissue when comparing to healthy tissue in a study performed by Li et al. (2021)[321]. *KRT25* and *KRT71* are responsible for the structural integrity of the hair follicles and are reportedly underexpressed in melanoma[322,323]. *LOR* encodes the protein loricrin which is a major component of epidermal cells[324]. Additionally, LPD_1 exhibited the overexpression and hypomethylation of the driver genes *FOXQ1* and *IRF6*. Likewise, the driver gene *ALB* was underexpressed and hypermethylated, while the driver gene *WT1* was underexpressed and enriched in mutations.

In bladder cancer, LPD_7 was characterised as a good prognosis subtype enriched in non-papillary samples, with overexpression of the region-angiotensis system, underexpression of chemical carcinogenesis and retinol metabolism, hypermethylation of embryonic development processes and hypomethylation of neuron development. The differentially expressed genes for this subtype with the largest changes are the underexpression of *FABP1* and *CRNN*, and the overexpression of *TRH*, *CGB5*, and *FGF4*. *FABP1* participates in the transport of fatty acids and its protein product is linked to the carcinogenesis of the bladder tumours, along with prostate and kidney tumours, but their exact role remains unclear[325,326]. *CRNN* plays a role in the epidermis differentiation and was studied as a possible urine biomarker for the diagnosis of bladder carcinoma[327]. *TRH* is involved in the regulation of the secretion of the thyroids and is reported as differentially expressed in the carcinogenesis of bladder cancer[328]. *CGB5* participates in the secretion of hormones for the maintenance of pregnancy and its expression has been reported as linked to bladder cancer, in addition to esophageal carcinoma, head and neck squamous cell carcinoma, lung squamous cell carcinoma, pancreatic adenocarcinoma, and rectum adenocarcinoma[329]. *FGF4* is responsible of embryonic development, cell growth and tissue repair; the dysregulation of the expression of this gene is associated to carcinogenesis of several cancer types including bladder cancer[330]. It is encouraging that there is evidence for an association with bladder cancer for a large number of these genes. LPD_7 was not picked up by hierarchical clustering suggesting that LPD can provide additional biologically useful information.

The identification and characterisation of subtypes with differential prognosis may serve as a source of prospective targets for diagnostics and development of new treatments or developing new biomarkers to improve the selection of treatment. Further research and validation will be required to build a new classification framework that can be applied in the clinical practice.

6.5 Summary

In this chapter, I employed the Automata workflow to identify and characterize 26 subtypes with differential prognoses. I specifically focused on the ones detected in skin cutaneous melanoma and bladder urothelial carcinoma. Two subtypes, LPD_1 and LPD_3, were identified for skin cutaneous melanoma. LPD_1 exhibited a good prognosis and was associated with specific genetic and epigenetic alterations, while LPD_3 represented a novel and previously unrecognized subtype with a favourable prognosis that traditional clustering methods failed to identify. Similarly, in bladder urothelial carcinoma, LPD_7 emerged as

a good prognosis subtype with distinct molecular features that were not captured by hierarchical clustering. Evidence of an association with bladder cancer with a large number of detected differentially expressed genes was provided.

The significant contributions of this research are the discovery of previously unknown molecular subtypes across 17 cancer types exhibiting differential prognoses. Among these subtypes, ten demonstrated a good prognosis, while 16 showed a poor prognosis. A comprehensive analysis of these subtypes will shed light on their genetic and epigenetic alterations, offering insights into potential therapeutic targets and biomarkers. However, further research and validation are essential to establish a new classification framework with clinical applicability.

Chapter 7

Conclusions and future work

7.1 Summary of findings

7.1.1 The Automata package

In this thesis, I have conducted the analysis of 28 cancer types from The Cancer Genome Atlas to detect and characterise subtypes and created a resource of results. In this chapter I will summarise my findings, put those results into a broader context and indicate some potential future directions to carry on this work.

In chapter 3, I introduce Automata, an R package developed to automate the detection and characterisation of molecular subtypes using Latent Process Decomposition in the cancer genome atlas (TCGA) dataset. The functions included in Automata are summarised in six major steps: downloading each of the 28 datasets, preprocessing the data to make it statistically fit, applying the Latent Process Decomposition (LPD) algorithm to find stratifications on the tumour dataset, postprocessing the samples to assign them to groups for downstream analysis, differential analysis between the groups to characterise them, and generating an interactive report of the results.

In an important advance for this type of analysis, I developed an approach to automatically determine the optimal number of processes for the LPD algorithm. This had previously been selected by manually observing a likelihood-ratio graph.

Characterisation of the detected subgroups included the identification of genes that were differentially expressed, differentially methylated, enriched or depleted in single nucleotide variants, insertions and deletions, and genes affected by copy number variations. A comparison of the association with survival outcomes across groups was also generated along with whether a group had an over-representation of benign samples. The COSMIC mutational signature profiles of each group were produced. A comparison of the LPD subgroups was made with Euclidean hierarchical clustering. Finally, the possibility of a batch effect associated with the data generation centre was examined.

The interactive report of results consists of a lightweight HTML file containing a summary of the significant outcomes from the differential analysis, a description of the followed methodology, and the differential analysis results via graphical representation or table. The report is presented as a website and is divided into different tabs to show each of the de-

tected processes separately. Additionally, pertinent graphics can be displayed interactively, allowing users to zoom in and filter results. A similar option is available for the tables, giving the user the possibility to search for specific entries or sort the table at will.

The free access of Automata as an R package, as well as its ability to automate the processing of TCGA data and to generate interactive reports with non-technical descriptions of the methodology, highlights the importance of this package as a source of cancer multi-omics data for the research community and an essential asset in aid of the reproducibility in research.

7.1.2 Pancancer analysis of subtypes detected by LPD across the TCGA

In chapter 4, I described the pancancer analysis carried out by comparing the outcome of Automata pipeline in each of the analysed 28 cancer types. The application of latent process decomposition to such a large and diverse dataset has never been performed before. I examined three main areas.

In the first area I focused on the factors involved in selecting the optimal number of processes. I found a significant positive relationship between the number of processes and the sample size of the datasets that was also present at different ITH levels of the cancer type. I concluded that a larger number of processes is required to explain the genetic variability present in highly heterogeneous cancer types.

Secondly, I identified shared molecular traits that distinguish subtypes across different cancer types. I successfully found a set of differentially expressed genes that met this condition, with 75 of them being driver genes. Three genes were found to be as differentially expressed in 23 cancer types: *CHGA*, *CPLX2*, and *ITLN1*. These three genes have been previously linked to cancer classification in several cancer types[273–276]. Enrichment of three biological pathways were characteristic of a subtype detected across subtypes: PPAR signalling pathway, complement and coagulation cascades, and neuroactive ligand-receptor. The role of these pathways in cancer remains uncertain as it is unclear whether they operate as tumour promoters, tumour suppressors, or both[281–288]. I concluded that these pathways might act as double agents, promoting the development of particular cancer subtypes while inhibiting others, and hence directly contribute to the increase of tumour heterogeneity. The identification of these altered genes and pathways indicates their importance in tumour growth to particular subtypes and emphasises their potential to be therapeutic targets that could work in specific subtypes of different cancer types.

Thirdly, I identified the subtype characteristic enrichment of specific molecular pathways or biological processes that were common across the cancer types grouped by their histology or association to hereditary diseases. No satisfactory results were obtained from this analysis. I attribute this to the lack of balanced representation of cancer histologies in the TCGA and the lack of complete clinical information on which cancer samples were caused by hereditary factors.

I identified a number of shortcomings in this chapter. The most flagrant one was the poor performance of the microarray samples in comparison to their RNA-seq counterparts, which may be due to the much smaller sample sizes in the microarray projects. The microarray data was not examined further in this thesis. A second shortcoming was the unusually

low number of identified genes significantly enriched or depleted in copy number variations. This may be due to the strict thresholds applied in the Automata pipeline or the algorithm used by the TCGA.

7.1.3 Validation of LPD and the study of breast, prostate, colorectal and lung carcinoma

In chapter 5, I described the validation of the LPD as an unsupervised cluster approach that works across several cancer types by applying it to the four cancer types that effect the most people in the UK: breast carcinoma, prostate adenocarcinoma, colorectal adenocarcinoma and lung cancer.

In breast carcinoma, LPD successfully distinguished the majority of the five histological types used in clinical classification[297]. Specifically, I found a process corresponding to the Normal subtype, a process corresponding to the Basal subtype, and then five processes defined by a mix of Luminal A and Luminal B in different proportions, along with HER2 samples. I concluded that these five processes were subdivisions of the Luminal category, which was consistent with similar studies such as Netanely et al. (2016)[296] and Daemen et al. (2018)[298]. Moreover, this result suggested that LPD may be able to sub-divide the established categories as it can distinguish Normal and Basal samples and reflect with exactitude the molecular compositions of the Luminal samples. This has the potential to improve clinical interventions for these additional types.

In prostate adenocarcinoma, a process with similar characteristics to a previously described poor prognosis subtype labelled as DESNT was found[118]. This process was characterised by a poor prognosis and a significant correlation in its gamma values to the ones from DESNT. Additionally, 77% of the samples assigned to this group were classified as DESNT in previous research[118].

In colorectal cancer, a similar approach was followed to detect a poor prognosis subtype labelled as Pericol[171], although in this case, a matching group was not found. However, this was consistent with the results outlined in the Pericol study, in which they had perform a correlation analysis in TCGA data with other databases in order to detect Pericol. Due to this, I believe that the LPD model built in the colorectal data of the TCGA is incapable of detecting Pericol.

In lung cancer, the datasets from two lung cancer types (lung adenocarcinoma and lung squamous cell carcinoma) were mixed to test the capacity of LPD to distinguish between them. Except for samples taken from normal tissue, the algorithm accurately discriminated between both cancer types. This was deemed acceptable since the molecular differences between benign tissue from both cancer types were considered minimal.

7.1.4 Identifying and characterising subtypes with a significant association with outcome

In chapter 6, I identified subtypes that had a significant association with time to death when compared to other samples in analysed cancer types: 26 subtypes across 17 different cancer types were detected, with ten showing a good prognosis and 16 showing a poor prognosis. In skin cutaneous melanoma I examined two of these subtypes in detail with many characteristics of these groups having been previously associated with skin cancers. I

also examined one good prognosis subtype in bladder cancer. There was good evidence for an association with bladder cancer for a large number of the differentially expressed genes for this subtype. Although not all the subtypes were explored, this showcased the enormous resource of molecular subtypes provided by the Automata pipeline.

7.2 Results in a broader context

7.2.1 LPD as a tool for the identification of cancer subtypes

Despite the advances in early detection and treatments, cancer remains as the second leading cause of death worldwide[311]. An explanation for this is that even within a single cancer type, a high degree of genetic heterogeneity is present across its samples, hindering the diagnosis and the development of new treatments. Driven by technological advances and decreased costs a plethora of 'omic datasets now exist that can be used to characterise individual samples and group cancers into subtypes. For transcriptome datasets, traditional unsupervised clustering approaches such as hierarchical clustering and k-means have been used to determine subtypes. There has been a notable success in breast cancer where five subgroups have been determined: Normal-like, Luminal A, Luminal B, Triple negative, and HER2[297]. Each subtype has distinct characteristics, such as the age of onset or the prognosis of the disease, and a different clinical treatment pathway. This framework is used routinely in clinical practice. However, for other cancer types there has been less success.

The inherent shortcoming of the traditional approaches is the implicit assumption of sample assignment to a particular cluster or group. Such analyses are in complete contrast to the complex composition of many cancers, with individual cancer samples containing multiple distinct biological processes simultaneously present and jointly contributing to their expression profile. Soft clustering methods, however, allow the assignment of samples into multiple groups with different levels of membership, which represents better the underlying biology in the tumour. Latent process decomposition (LPD) is a Bayesian unsupervised clustering technique that takes heterogeneity into account and represents the expression profile of a cancer as a combination of underlying latent processes. Each latent process is considered as an underlying functional state and a given sample can be represented over a number of these underlying functional states.

The potential of LPD for high heterogeneous cancers in comparison to hierarchical clustering and k-means has been evidenced by Luca (2017)[207]. In his work, LPD was able to detect a poor prognosis subtype in prostate cancer labelled as DESNT across several datasets. However, hierarchical clustering failed to detect any differential prognosis groups, while k-means did not consistently identified the DESNT group across datasets. In this work the capacity of LPD to detect additional biologically useful structure across a wide range cancer types when compared to hierarchical clustering has been shown. Except for breast carcinoma, the hierarchical clustering failed to represent the complexity of the tumour and to separate the samples according to their prognosis. I have also shown that LPD can propose novel clinically relevant subtypes i.e. subtypes that indicate a worse or better prognosis and detect differential prognosis has been demonstrated in this thesis. At this stage these novel subtypes are proposals as they have only been detected in one dataset, but the potential for an impact to improve patient care is clear.

7.2.2 The Automata package

For the development of this thesis, an R package named as Automata was developed to automate the application of LPD across datasets from The Cancer Genome Atlas (TCGA) database and the characterisation of the detected processes. Due to the nature of packages in R, the source code from Automata can be easily modified and adapted to work with further datasets, include new analysis, or change any functionality at will.

In addition to allow the greater utility of our results, the package generates an interactive report of the results for each analysed dataset. The work has resulted in a rich publicly available resource that can be mined by the scientific community to answer specific questions and generate hypotheses for future work. This work is the first time that subtypes and their characterisation has been made available for the TCGA dataset.

The report relies heavily in the use of visual elements to represent the data and is structured ain an interactive dashboard. This facilitates browsing across the report, makes exploring the data more intuitive, and improves the user-experience. This increases the accessibility to scientists that may be wary of dealing with the raw results. The use of dashboards is becoming increasingly popular and a necessary step to make the results of large and complex results to the wider scientific community. Good examples are the Integrative OncoGenomics web portal[331], the ICGC data portal[332], and the cBioPortal[333].

7.2.3 The importance of pancancer studies

Until recently, cancer genomic project initiatives such as the TCGA have focused on the analysis of specific tumour types. This has been very successful and has helped to identify novel driver genes, new biomarkers and define molecular subtypes specific for those cancers[334]. However, it is known that tumours from different organs cancer can share many characteristics whereas tumours from the same organ can be different. For example, Ma et al. (2021)[245] described that mutations in the *TP53* gene drive high-grade serous ovarian, serous endometrial and basal breast carcinomas; This highlights the need of developing a holistic framework across tumours independent of their type, since findings in one can be applied to another. Furthermore, integrated data interpretation will aid in determining how the effect of gene alterations differ across tumours and their therapeutical implications.

The TCGA has performed pancancer analysis across 12 cancers: glioblastoma multiforme, lymphoblastic acute myeloid leukemia, head and neck squamous carcinoma, lung adenocarcinoma, lung squamous carcinoma, breast carcinoma, kidney renal clear-cell carcinoma, ovarian carcinoma, bladder carcinoma, colon adenocarcinoma, uterine cervical and endometrial carcinoma and rectal adenocarcinoma[334]. In their studies they reported the discovery of new driver mutations due to the increment in statistical power caused by the addition of sample size and a better understanding of the differences between driver and passenger mutations[335–337]. Weinstein et al. (2013)[334] also found that the gene *ERBB2* is amplified in subsets of glioblastoma, gastric, serous endometrial, bladder and lung cancer. Additionally, they proposed that mixing cancer types aids in the identification of carcinogenetic processes not related to specific tissue types[338].

Most pancancer studies are performed by comparing whole cancer types, thereby overlooking the shared features and similarities between subtypes of different cancers. Whole-

cancer comparisons conceal the molecular processes that thrive the stratification of the tumour into subpopulations by favouring the detection of differential carcinogenic processes instead.

In this research, I performed a pancancer analysis across the subtypes detected by LPD in 28 cancer types to unravel the biological processes shared across cancer types that contribute to tumour stratification and heterogeneity. I have successfully identified a set of genes and biological pathways that were characteristic of a subtype that were recurrently identified across subtypes from different cancer types. Although the data I collected was only partially investigated, it remains as one-of-a-kind data compilation that can be used in future research to discover new potential biomarkers or therapeutic targets that are effective across cancer types.

7.2.4 The expansion of personalised medicine

Emerging, high-throughput, data-intensive biomedical assays, such as DNA sequencing, proteomics, or imaging protocols, has put into evidence the heterogeneous disease course and response to treatment across individuals. This has caused a gradual paradigm shift from traditional medicine towards a personalised medicine approach in which the treatment of the patients is tailored to their unique molecular and genetic profile[339]. The use of genomic profiling in research is already well-established and it is becoming increasingly so in clinical practice. This is exemplified by the development of initiatives such as the 100,000 genomes project by Genomics England[340] and the use of next generation sequencing in the diagnosis of hereditary diseases[341].

Two examples of how personalised medicine is currently being used in clinical practice are mutation-specific therapies and personalising disease prevention[342]. Mutation-specific therapies consist of identifying the genetic profile of the patient and then prescribing a drug that is effective for that profile. An example of this is the drug ivacaftor that is used to treat patients with cystic fibrosis but is only effective for patients with a specific mutation (G551D mutation) in the gene *CFTR*[343]. Another example is, in colorectal cancer, patients with a mutation in the *PIK3CA* gene who took post-operative aspirin showed an improvement in overall survival in comparison to those who did not[344]. Personalised disease prevention entails developing personalised disease prevention strategies based on the genomic profile of the patient. For example, screening for BRCA mutations in families with a history of breast cancer and offering risk-reducing surgery to remove breast tissue in those with mutations[345].

Cancer subtypes play a critical role in the transition of cancer treatment to personalised medicine. For example, individuals with breast cancer dominated by the HER2 subtype have a unique drug, Herceptin, that improves their overall survival[346]. In this study, I have identified and characterised 168 subtypes across 28 cancer types. Of these 26 had a significant association with outcome in comparison to other subtypes in that cancer type: with ten exhibiting a good prognosis and 16 exhibiting a poor prognosis. The data gathered during this work constitutes a valuable source of data to identify subtypes that could potentially be used in clinical practice, understand the biological mechanisms behind the stratification of tumours into subtypes, and to support the transition to individualised cancer therapy. An example to illustrate the potential of the data analysed in this research is the DESNT subtype[118]. This subtype was identified in prostate cancer using the LPD algorithm that

was also used in this work and was shown to be robustly a poor prognosis subtype. In further research, Luca et al. (2020)[120] reported that the presence of the DESNT subtype, even in small quantities, was an indicator of poor outcome. This has the potential to improve the decision of what is the best treatment for a patient. DESNT is now the basis for a large research programme in the Cancer Genetics team at the Norwich Medical School. Most importantly, a diagnostics lab has been set up to translate DESNT into a prognostic test that can be used at the time of diagnosis of prostate cancer in clinical practice and is in the process of being accredited[347].

7.3 Novel Findings and Publishable Contributions

The research presented in this thesis has yielded several novel findings and contributions to the field of cancer genomics, which hold the potential for publication in reputable academic journals. The following key findings stand out as noteworthy:

7.3.1 The Automata Package: Automation of Cancer Subtype Analysis

The development of the "Automata" R package represents a significant contribution to the field of cancer genomics. This package allows for the automated detection and multi-omic characterisation of molecular subtypes using the Latent Process Decomposition algorithm applied to data from The Cancer Genome Atlas dataset. The functions encompass six major steps: data downloading, data preprocessing, LPD algorithm application, data postprocessing, differential analysis, and generating interactive reports with non-technical methodology descriptions. The capability of the package to characterise the detected subtypes through various analyses, including differential gene expression, methylation patterns, single nucleotide variants, insertions and deletions, copy number variations, mutational signature profiles, and clinical associations, provides a holistic perspective of cancer subtypes, enhancing the understanding of their molecular features.

Automata introduces a user-friendly and accessible tool for cancer researchers, eliminating the need for laborious manual analysis of TCGA data. The package's open-source nature ensures it can be easily modified and adapted to work with other cancer genomics databases and datasets. As such, it represents a valuable resource for the research community, contributing to improved data accessibility and reproducibility and aiding in collaborative efforts in cancer genomics research. Moreover, the Automata package generates interactive reports in the form of lightweight HTML files, making the results easily accessible and understandable to researchers, clinicians, and non-technical users. The report provides a summary of significant outcomes and a description of the methodology, along with graphical representation and tables that can be interactively filtered and explored.

Additionally, a notable advancement in this research is the automated determination of the optimal number of processes for the LPD algorithm. Traditionally, this selection was done manually by visual inspection of a log-likelihood graph. The development of an automated approach streamlines the analysis and improves its consistency and accuracy.

7.3.2 Pan-Cancer Analysis of Subtypes: Uncovering Shared Molecular Traits

The pancancer analysis conducted in Chapter 4 comparing the outcome of the Automata pipeline in 28 different cancer types is a comprehensive and systematic approach that has not been previously performed. This analysis has provided invaluable insights into shared molecular traits that distinguish subtypes across diverse cancer types. By successfully identifying differentially expressed genes and associated biological pathways and processes across multiple cancer types, this research has uncovered potentially relevant biomarkers that could have implications for clinical practice.

7.3.3 Validation of LPD and Study of Major Cancer Types

The validation of the LPD algorithm integrated into the Automata pipeline and its application to four major cancer types in Chapter 5 have provided valuable insights into their molecular heterogeneity and the potential for subtype-specific treatment strategies. Notably, in breast carcinoma, LPD successfully identified distinct subdivisions within the Luminal subtype and offered a more precise and accurate characterization of the basal and normal-like subtypes. Similarly, in prostate adenocarcinoma, the detection of a poor prognosis subtype with similarities to the DESNT subtype further validates the efficacy of the Automata's LPD approach and highlights its clinical relevance. Furthermore, exploring colorectal and lung cancer has revealed new insights into poor prognosis subtypes. Although limitations were encountered, this study demonstrates the potential of LPD to contribute to personalized cancer therapy. These findings lay the foundation for future investigations and clinical translation of the identified subtypes in the studied cancer types.

7.3.4 Identifying Subtypes with Clinical Relevance

Chapter 6 describes the identification of 26 distinct subtypes across 17 cancer types that exhibited a significant differential prognosis in terms of survival probability. Among these subtypes, ten showed a good prognosis, and 16 showed a poor prognosis. This novel finding has considerable implications for precision medicine and personalized treatment strategies. These subtypes provide valuable prognostic information, potentially guiding treatment decisions and improving patient outcomes. Moreover, they present promising targets for further investigation and potential translation to clinical applications.

7.3.5 Resource for the Scientific Community

The comprehensive analysis and generation of an extensive resource of molecular cancer subtypes in this research provide valuable data for the scientific community. This resource can be used to generate hypotheses, identify potential biomarkers, and support further research in cancer genomics.

7.4 Limitations

During the course of this research, three main limitations were identified, impacting the interpretation and generalization of the findings.

7.4.1 Exclusively reliant on TCGA data for subtype analysis

A major limitation of this study is the exclusive use of data from the TCGA database for gathering, processing, and analysis. While the TCGA is a valuable resource, using it solely raises concerns about the external validity and generalization of the results to other datasets and patient populations. The lack of independent validation datasets makes it difficult to ascertain whether the identified subtypes and their characteristics are truly representative of the molecular landscape of the analyzed cancers or if they are specific to the TCGA datasets. Consequently, all findings presented in this thesis should be considered hypotheses requiring further validation and investigation.

Additionally, it is important to acknowledge the potential bias in the TCGA dataset, as it predominantly comprises samples from Caucasian populations. The limited representation of ethnic diversity within the dataset may result in subtype definitions specific to certain ethnic backgrounds, hindering the generalization of the findings to individuals from other racial and ethnic groups. To address this limitation, future studies should aim to include more diverse samples, actively involving participants from ethnic minority backgrounds to ensure more comprehensive and generalizable results. Three main limitations were identified while conducting this research: the analysis focused solely on data from the TCGA, the need for an in-depth analysis of each cancer type by experts in the respective fields, and time limitations for a thorough examination of the results.

Furthermore, the TCGA lacks representation of certain rare and less-studied cancer types. Consequently, the findings and subtypes identified in this study may not apply to these underrepresented cancer types, limiting this research's overall generalization and clinical utility.

7.4.2 Grouping samples in LPD analysis

Although LPD is a soft clustering technique, in the postprocessing step of Automata, the samples are allocated into LPD groups, and subsequent differential analysis is performed by comparing these groups. One of the main reasons for adopting the grouping strategy instead of analyzing the effect of each subtype based on their presence percentage in the sample is the practical complexity arising from the vast number of cancer types under investigation. Individually examining each subtype would make the task practically infeasible within the scope of this study. The wide variability in the proportions of subtypes across individual samples would make it challenging to derive meaningful insights for clinical use. By grouping the subtypes into more manageable and cohesive clusters, it becomes more feasible to effectively characterize and compare the subtypes across different samples.

However, it is essential to acknowledge that the grouping approach may also have limitations. Although it streamlines the analysis process, it may oversimplify the representation of sample heterogeneity, potentially overlooking finer distinctions between samples within the same group.

7.4.3 Adequate sample size requirement for LPD analysis

Due to the inherent nature of LPD, which requires a sufficient number of samples for robust and biologically meaningful results, some cancer types with less than 100 samples were not selected for this study. Consequently, the insights gained from this research might not be

fully generalizable to all cancer types, especially those with smaller sample sizes.

7.4.4 Insufficient disease-specific expert analysis

While this research has included a comprehensive characterization of the identified subtypes, it is essential to recognize the need for in-depth analysis by disease-specific experts. Expert input and expertise are crucial for fully understanding the complexity and implications of the identified subtypes for each specific cancer type and enhancing their accuracy and clinical relevance.

7.4.5 Time constraints on thorough examination

The vast amount of data generated during this research difficult the in-depth examination of all the analyzed cancer types due to time limitations. The thorough analysis and comparison of each cancer type to existing literature were not feasible within the scope of this thesis. As a consequence, some cancer types may not have received the level of scrutiny needed to fully understand the biological mechanisms behind the obtained results.

7.5 Abundance of inmune system processes and impact on the interpretation of results

Throughout this research, the characterization of subtypes revealed a notable abundance of immune system-related processes associated with differentially expressed, methylated, mutated or copy-number affected genes. Several factors could contribute to this observation:

Firstly, cancer is a disease characterized by complex interactions between tumour cells and the immune system, as reflected in the hallmarks of cancer (see 1.2.1). This dynamic interplay activates and modulates numerous immune-related processes, influencing gene regulation and expression levels. As the LPD algorithm was applied to expression data, the influence of the immune system on clustering outcomes became apparent. Moreover, since the immune system plays a critical role in driving heterogeneity within cancer samples, it can potentially lead to the stratification of cancer into distinct subtypes, which is effectively captured by the LPD algorithm.

Secondly, immune-related processes may also be attributed to tumour infiltration, that is to say, the presence of immune cells within the tumour environment. Consequently, some subtypes detected in this study may not exclusively reflect the tumour biology but instead represent the intricate tumour-immune system interactions. Further analysis will be required to quantify the impact of tumour infiltration on the identified subtypes.

In conclusion, the abundant presence of immune system-related processes in this research underscores the complex role the immune system plays in cancer development and progression. Moreover, it highlights the potential influence of tumour infiltration on tumour heterogeneity and subtype discovery. Understanding these intricate interactions will contribute to advancing the comprehension of cancer biology and may open new pathways for targeted therapeutic approaches.

7.6 Adaptability to other databases

The approach presented in this research has been primarily developed and demonstrated using data from the TCGA. However, it is essential to emphasize its potential for adaptation to other databases and datasets.

While the TCGA is one of the largest and most comprehensive resources for cancer genomics data, there are other public repositories and databases that contain multiple levels of data from various cancer types, such as the Expression Omnibus database. Researchers can easily modify the Automata R package to accommodate different data formats, experimental designs, and preprocessing steps specific to alternative databases. The accessible and open-source nature of the package converts it into a flexible tool that can be tailored to fit various datasets and analyses. It is also important to notice that the LPD algorithm is not restricted to the TCGA database alone, and it can be applied to any data that can be represented in a numeric matrix format.

By making the Automata package adaptable to other databases, this research becomes relevant to a broader audience of cancer genomics researchers working with different datasets. It encourages other scientists to explore the utility of LPD and the Automata package in their studies, potentially leading to more comprehensive and comparative analyses across diverse datasets and advancing the understanding of the biological mechanisms that drive cancer subtypes and their implications in oncology research.

7.7 Future work

Here I present some possible directions of research that could be pursued in continuation of the work presented in this thesis.

7.7.1 Validation of the results in other datasets and further research

In this work, five cancer types were analysed in-depth: breast carcinoma, prostate adenocarcinoma, colorectal cancer, lung adenocarcinoma, and lung squamous cell carcinoma. In addition to this, 176 subtypes were detected across cancer types. However, these results were not validated in other datasets. Since the TCGA is one of the largest hubs of data from distinct cancer types, it would be difficult to find another database that contains transcriptome data for all the analysed cancer types in this study. A possible alternative would be the use of specific datasets of each cancer type.

A plethora of datasets are available in the Gene Expression Omnibus[292] and in the Pan-cancer Analysis of Whole Genomes[348] data portals from different cancer types, however their sample size could be too low in comparison to the TCGA, therefore, several datasets may need to be combined. Some example of other datasets are the dataset of 3,273 breast cancer samples by Brueffer et al. (2018)[349], the PanProstate Cancer Group with over 2000 samples of prostate adenocarcinoma[350], the dataset of 593 samples of colorectal cancer samples by Lin et al. (2018)[351], the dataset of 1,118 lung adenocarcinoma samples by Lim et al. (2018)[352], and the dataset of 80 lung squamous cell carcinoma samples by Setpahty et al. (2021)[353].

Furthermore, to understand the results, the biological mechanisms behind them, and the potential clinical implications, the assistance of scientific experts specific for each of the

analysed cancer types would be required.

7.7.2 Gamma values as a continuous variable

For each subtype identified LPD provides a score between 0 and 1 that identifies what proportion of the expression from that sample cn be explained by that process (gamma value). In this study, the samples were then categorised into groups based on the most prevalent process. This allowed the characteristics of each group to be determined. An alternative approach to identifying characteristics would be to treat gamma values as a continuous variable and study how the molecular features of the samples variate depending on the fraction of each LPD signature present on them. Although making the statistical tests of association more complex, this may take advantage of the heterogeneity of the scores better. This would be especially interesting to look at when testing the clinical associations. Using this approach I could determine the optimal combination of process gammas to predict outcome.

The idea of using gamma values as a continuous metric has already been studied with successful results by Luca et al. (2020)[120]. In their work, they reported that the presence of the gamma values corresponding to the DESNT subtype, even in small quantities, was an indicator of poor outcome.

7.7.3 Applying LPD on methylation data

Applying the LPD algorithm in the methylation data of the TCGA to determine subtypes driven by changes in epigenetic factors would provide a new dimension to my previous findings. A comparison between the processes discovered in transcriptome data with the ones from epigenetics could provide a better understanding of the etiology molecular landscape of cancer. Moreover, this approach may provide more accurate results for those cancer types known to be dominated by malignant methylation processes such as lung cancer[354].

7.8 Conclusion

Driven by technological advances, the treatment of cancer is gradually shifting towards a personalised medicine approach in which the identification of subtypes is critical to fully comprehend the genomic profile of each patient and provide accurate prognosis and treatment. In this thesis, I have developed a methodology that can be easily reproduced and adapted to new data and analyses to identify 168 cancer subtypes with varied prognosis spanning across 28 cancer types. Moreover, I have characterised the features of each subtype, generating an encyclopaedic compendium of molecular subtypes of cancer that provides an inestimable source of information for the research community and for future studies.

Appendix A

TCGA available data

Table A.1: List of available cancer data in the TCGA database. For each cancer type it is shown the TCGA project ID and the number of available cases.

Cancer type	TCGA project ID	Number of cases available
Acute Myeloid Leukemia	TCGA-LAML	200
Adrenocortical Carcinoma	TCGA-ACC	92
Bladder Urothelial Carcinoma	TCGA-BLCA	412
Brain Lower Grade Glioma	TCGA-LGG	516
Breast Invasive Carcinoma	TCGA-BRCA	1098
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	TCGA-CESC	307
Cholangiocarcinoma	TCGA-CHOL	51
Colon Adenocarcinoma	TCGA-COAD	461
Esophageal Carcinoma	TCGA-ESCA	185
Glioblastoma Multiforme	TCGA-GBM	617
Head and Neck Squamous Cell Carcinoma	TCGA-HNSC	528
Kidney Chromophobe	TCGA-KICH	113
Kidney Renal Clear Cell Carcinoma	TCGA-KIRC	537
Kidney Renal Papillary Cell Carcinoma	TCGA-KIRP	291
Liver Hepatocellular Carcinoma	TCGA-LIHC	377
Lung Adenocarcinoma	TCGA-LUAD	585
Lung Squamous Cell Carcinoma	TCGA-LUSC	504
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	TCGA-DLBC	58
Mesothelioma	TCGA-MESO	87
Ovarian Serous Cystadenocarcinoma	TCGA-OV	608
Pancreatic Adenocarcinoma	TCGA-PAAD	185
Pheochromocytoma and Paraganglioma	TCGA-PCPG	179
Prostate Adenocarcinoma	TCGA-PRAD	500
Rectum Adenocarcinoma	TCGA-READ	172
Sarcoma	TCGA-SARC	261
Skin Cutaneous Melanoma	TCGA-SKCM	470
Stomach Adenocarcinoma	TCGA-STAD	443
Testicular Germ Cell Tumors	TCGA-TGCT	150
Thymoma	TCGA-THYM	124
Thyroid Carcinoma	TCGA-THCA	507
Uterine Carcinosarcoma	TCGA-UCS	57
Uterine Corpus Endometrial Carcinoma	TCGA-UCEC	560
Uveal Melanoma	TCGA-UM	80

Appendix B

COSMIC Mutational Signatures in Human Cancer

Table B.1: List of COSMIC mutational signatures in human cancer. For each one, it is described the cancer types in which they are more predominant, the proposed aetiology, and additional mutational features. Adapted from Tate et al. (2019)[236].

Signature	Cancer Types	Proposed Aetiology	Additional Mutational Features
Signature 1	All	Endogenous mutational process initiated by spontaneous deamination of 5-methylcytosine	Associated with small numbers of small insertions and deletions in most tissue types
Signature 2	22 cancer types, commonly in cervical and bladder cancers	Activity of the AID/APOBEC family of cytidine deaminases	Transcriptional strand bias observed in exons; associated with Signature 13
Signature 3	Breast, ovarian, and pancreatic cancers	Failure of DNA double-strand break-repair by homologous recombination	Strongly associated with elevated numbers of large insertions and deletions with overlapping microhomology at breakpoint junctions
Signature 4	Head and neck cancer, liver cancer, lung adenocarcinoma, lung squamous carcinoma, small cell lung carcinoma, and esophageal cancer	Smoking	Exhibits transcriptional strand bias for C>A mutations and associated with CC>AA dinucleotide substitutions
Signature 5	All	Unknown	Exhibits transcriptional strand bias for T>C substitutions at ApTpN context
Signature 6	17 cancer types, most common in colorectal and uterine cancers	Defective DNA mismatch repair	Associated with high numbers of small insertions and deletions at mono/polynucleotide repeats
Signature 7	Skin cancers and cancers of the lip	Ultraviolet light exposure	Associated with large numbers of CC>TT dinucleotide mutations at dipyrimidines; exhibits transcriptional strand-bias

Table B.1: List of COSMIC mutational signatures in human cancer. For each one, it is described the cancer types in which they are more predominant, the proposed aetiology, and additional mutational features. Adapted from Tate et al. (2019)[236]. *(continued)*

Signature	Cancer Types	Proposed Aetiology	Additional Mutational Features
Signature 8	Breast cancer and medulloblastoma	Unknown	Exhibits weak strand bias for C>A substitutions; associated with double nucleotide substitutions, notably CC>AA
Signature 9	Chronic lymphocytic leukaemias and malignant B-cell lymphomas	Somatic hypermutation by AID	N/A
Signature 10	6 cancer types, notably colorectal and uterine cancer	Altered activity of the error-prone polymerase POLE	Exhibits strand bias for C>A mutations at TpCpT context and T>G mutations at TpTpT context
Signature 11	Melanoma and glioblastoma	Resembles mutational pattern of alkylating agents	Exhibits strong transcriptional strand-bias for C>T substitutions
Signature 12	Liver cancer	Unknown	Exhibits strong transcriptional strand-bias for T>C substitutions
Signature 13	22 cancer types, commonest in cervical and bladder cancers	Attributed to activity of the AID/APOBEC family of cytidine deaminases converting cytosine to uracil	Transcriptional strand bias observed in exons; associated with Signature 2
Signature 14	4 uterine cancers and a single adult low-grade glioma sample	Unknown	N/A
Signature 15	Several stomach cancers and a single small cell lung carcinoma	Defective DNA mismatch repair	Associated with high numbers of small insertions and deletions at mono/polynucleotide repeats

Table B.1: List of COSMIC mutational signatures in human cancer. For each one, it is described the cancer types in which they are more predominant, the proposed aetiology, and additional mutational features. Adapted from Tate et al. (2019)[236]. *(continued)*

Signature	Cancer Types	Proposed Aetiology	Additional Mutational Features
Signature 16	Liver cancer	Unknown	Exhibits extremely strong transcriptional strand bias for T>C mutations at ApTpN context, occurring almost exclusively on the transcribed strand
Signature 17	Esophagus cancer, breast cancer, liver cancer, lung adenocarcinoma, B-cell lymphoma, stomach cancer, and melanoma	Unknown	N/A
Signature 18	Neuroblastoma, and also observed in breast and stomach carcinomas	Unknown	N/A
Signature 19	Pilocytic astrocytoma	Unknown	N/A
Signature 20	Stomach and breast cancers	Defective DNA mismatch repair	Associated with high numbers of small insertions and deletions at mono/polynucleotide repeats
Signature 21	Stomach cancer	Unknown	N/A
Signature 22	Urothelial (renal pelvis) carcinoma and liver cancers	Exposures to aristolochic acid	Exhibits very strong transcriptional strand bias for T>A mutations
Signature 23	Found only in a single liver cancer sample	Unknown	Exhibits very strong transcriptional strand bias for C>T mutations

Table B.1: List of COSMIC mutational signatures in human cancer. For each one, it is described the cancer types in which they are more predominant, the proposed aetiology, and additional mutational features. Adapted from Tate et al. (2019)[236]. *(continued)*

Signature	Cancer Types	Proposed Aetiology	Additional Mutational Features
Signature 24	Subset of liver cancers	Exposures to aflatoxin	Exhibits very strong transcriptional strand bias for C>A mutations
Signature 25	Hodgkin lymphomas	Unknown	Exhibits transcriptional strand bias for T>A mutations
Signature 26	Breast cancer, cervical cancer, stomach cancer, and uterine carcinoma	Defective DNA mismatch repair	Associated with high numbers of small insertions and deletions at mono/polynucleotide repeats
Signature 27	Subset of kidney clear cell carcinomas	Unknown	Exhibits very strong transcriptional strand bias for T>A mutations; associated with high numbers of small insertions and deletions at mono/polynucleotide repeats
Signature 28	Subset of stomach cancers	Unknown	N/A
Signature 29	Gingivo-buccal oral squamous cell carcinoma	Tobacco chewing habit	Exhibits transcriptional strand bias for C>A mutations; associated with CC>AA dinucleotide substitutions
Signature 30	Breast cancers	Unknown	N/A

Appendix C

Correlations between RNA-seq and Microarray LPD groups

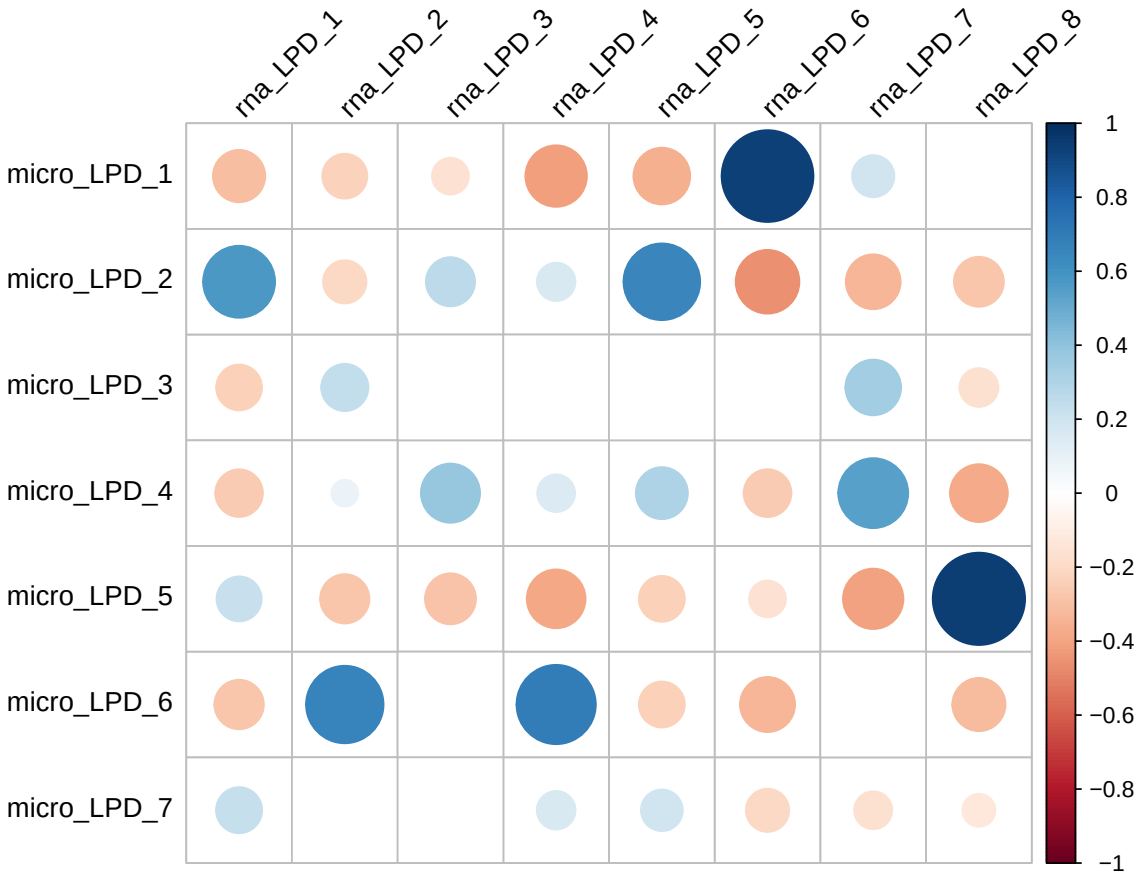

Figure C.1: Correlation matrix for TCGA-BRCA between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

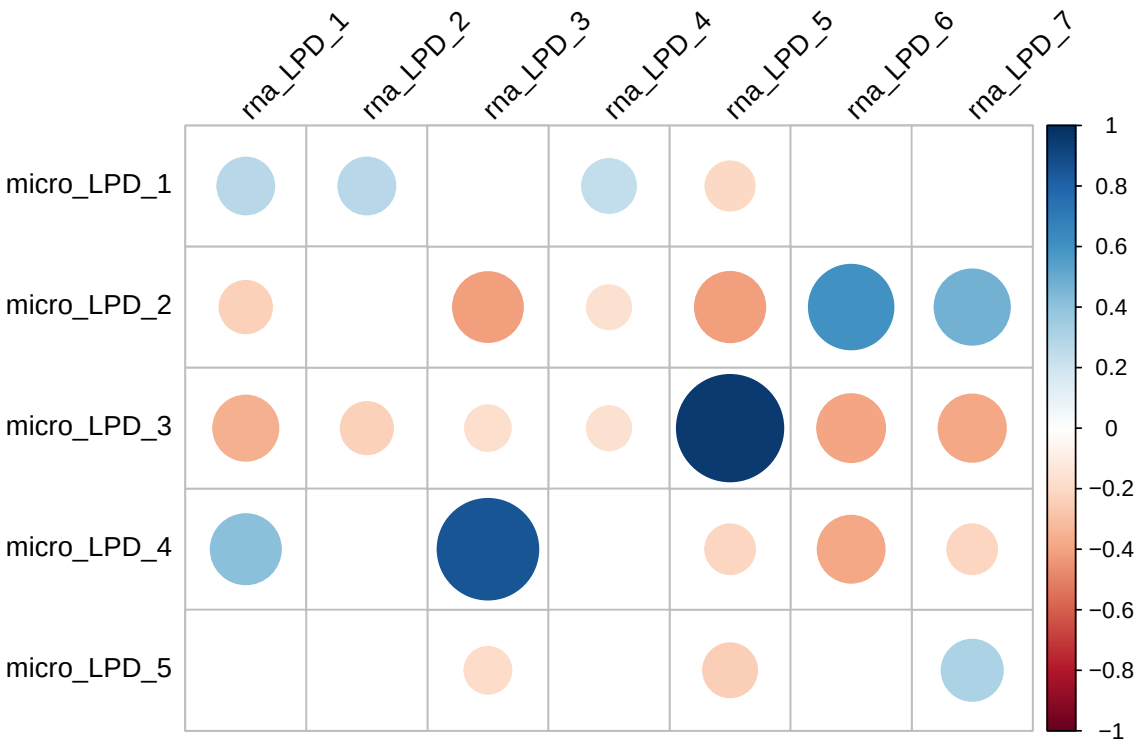

Figure C.2: Correlation matrix for TCGA-COAD between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

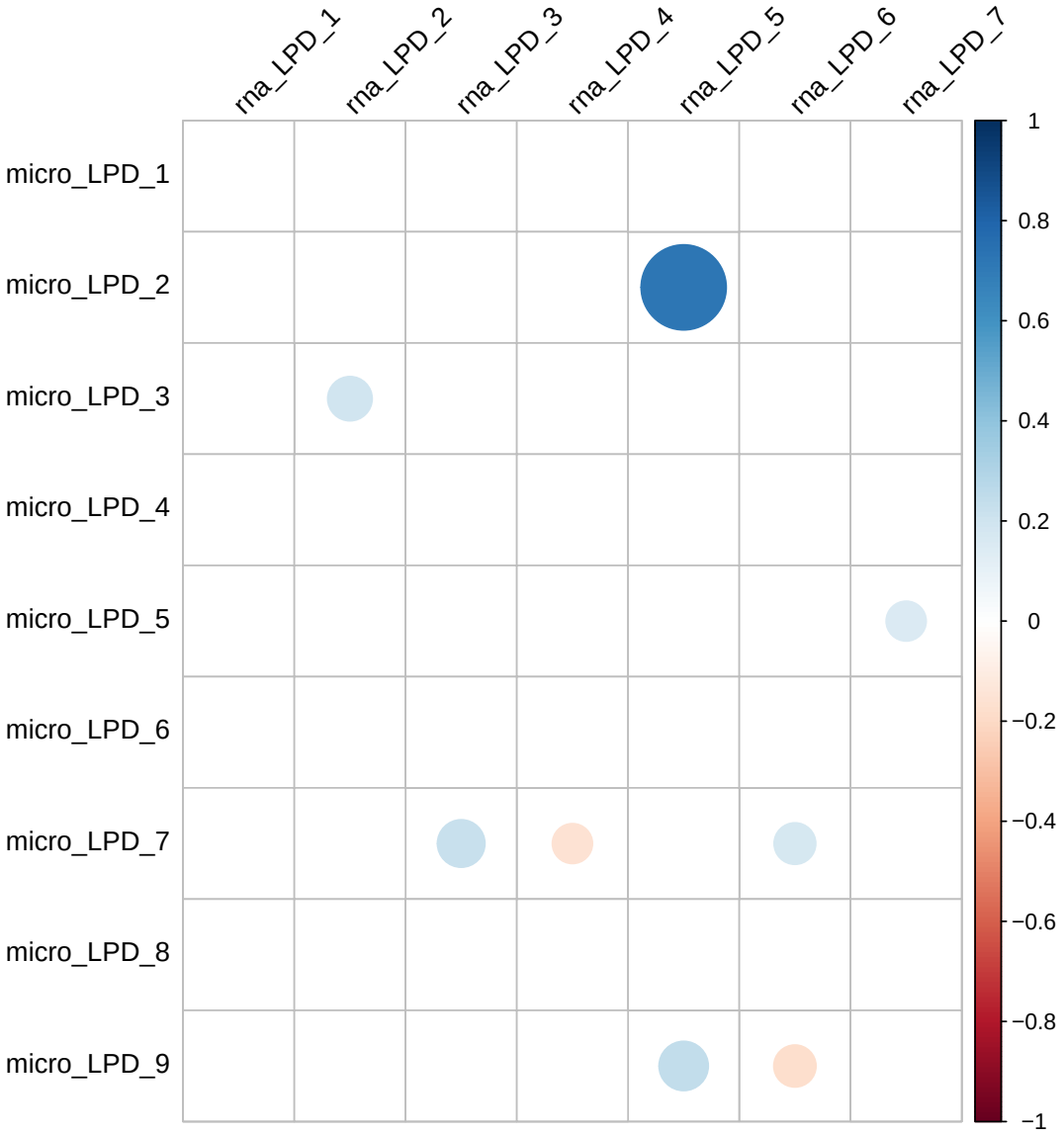

Figure C.3: Correlation matrix for TCGA-GBM between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

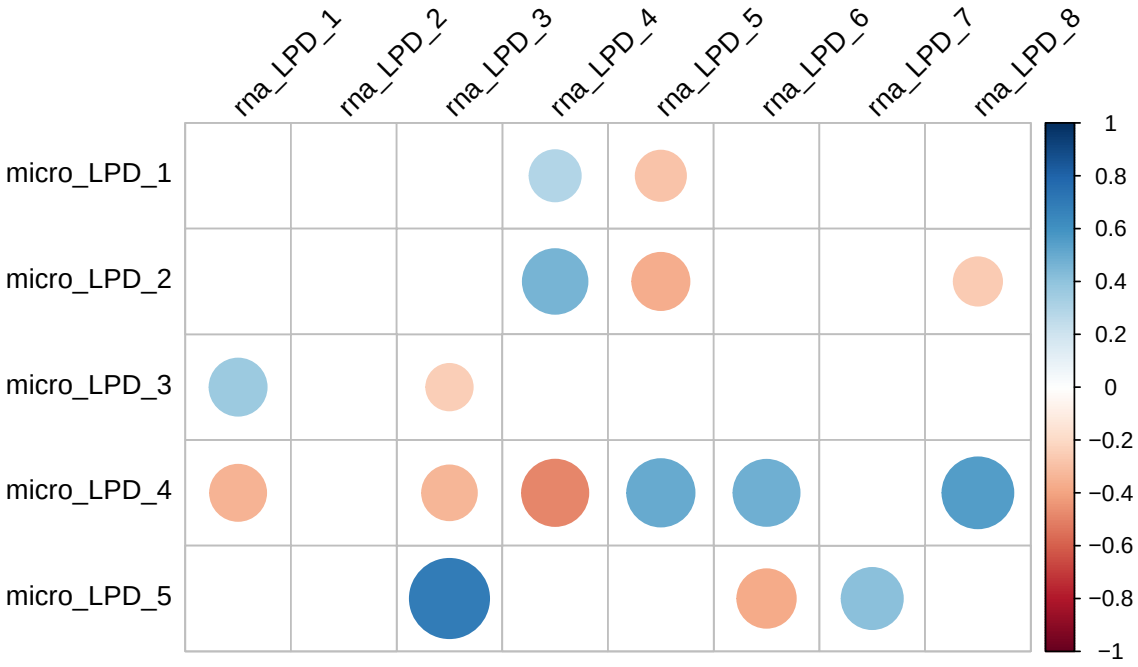

Figure C.4: Correlation matrix for TCGA-KIRC between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

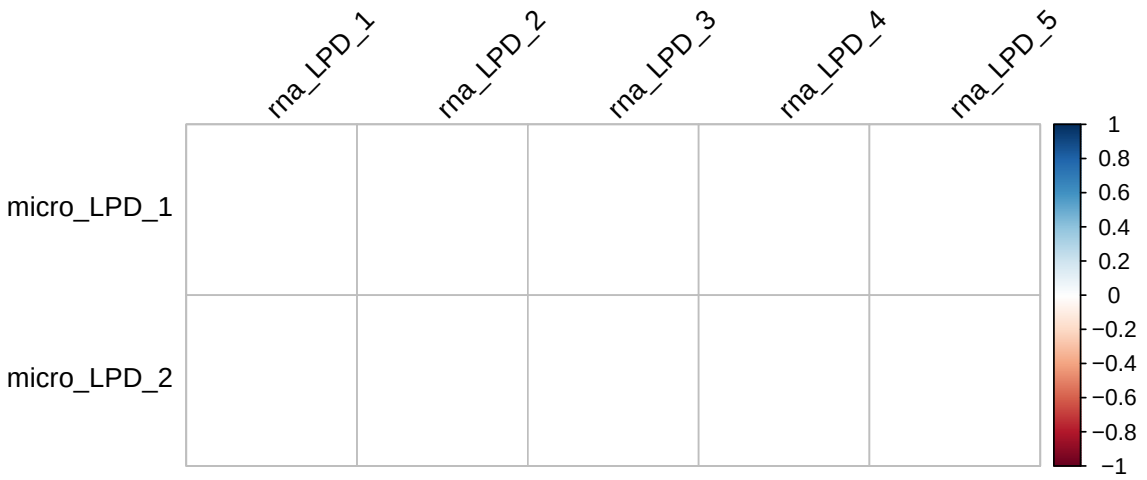

Figure C.5: Correlation matrix for TCGA-KIRP between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

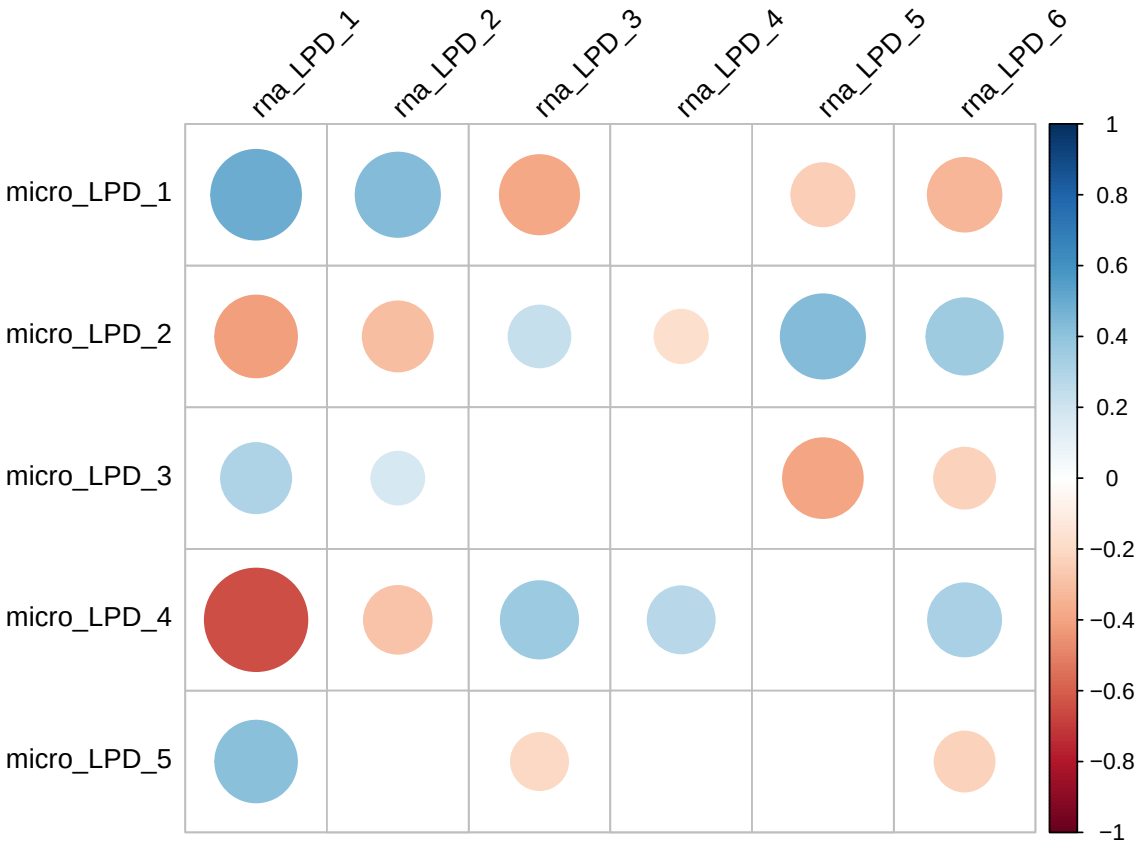

Figure C.6: Correlation matrix for TCGA-LAML between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

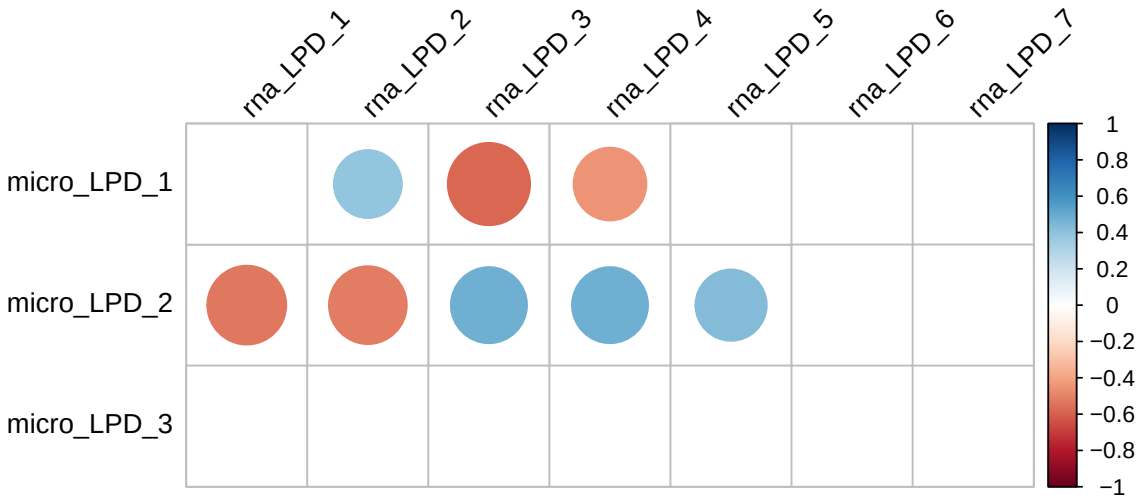

Figure C.7: Correlation matrix for TCGA-LGG between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

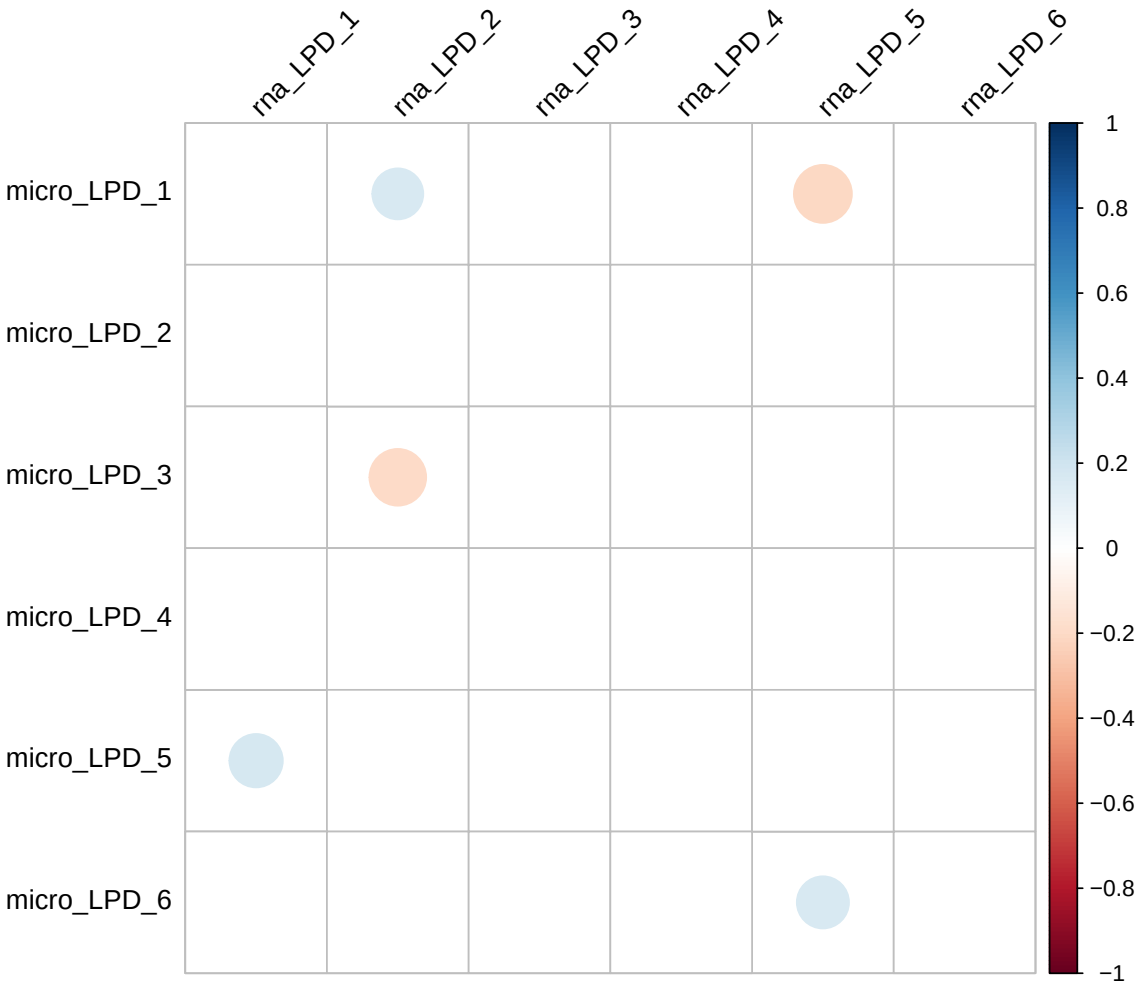

Figure C.8: Correlation matrix for TCGA-LUSC between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

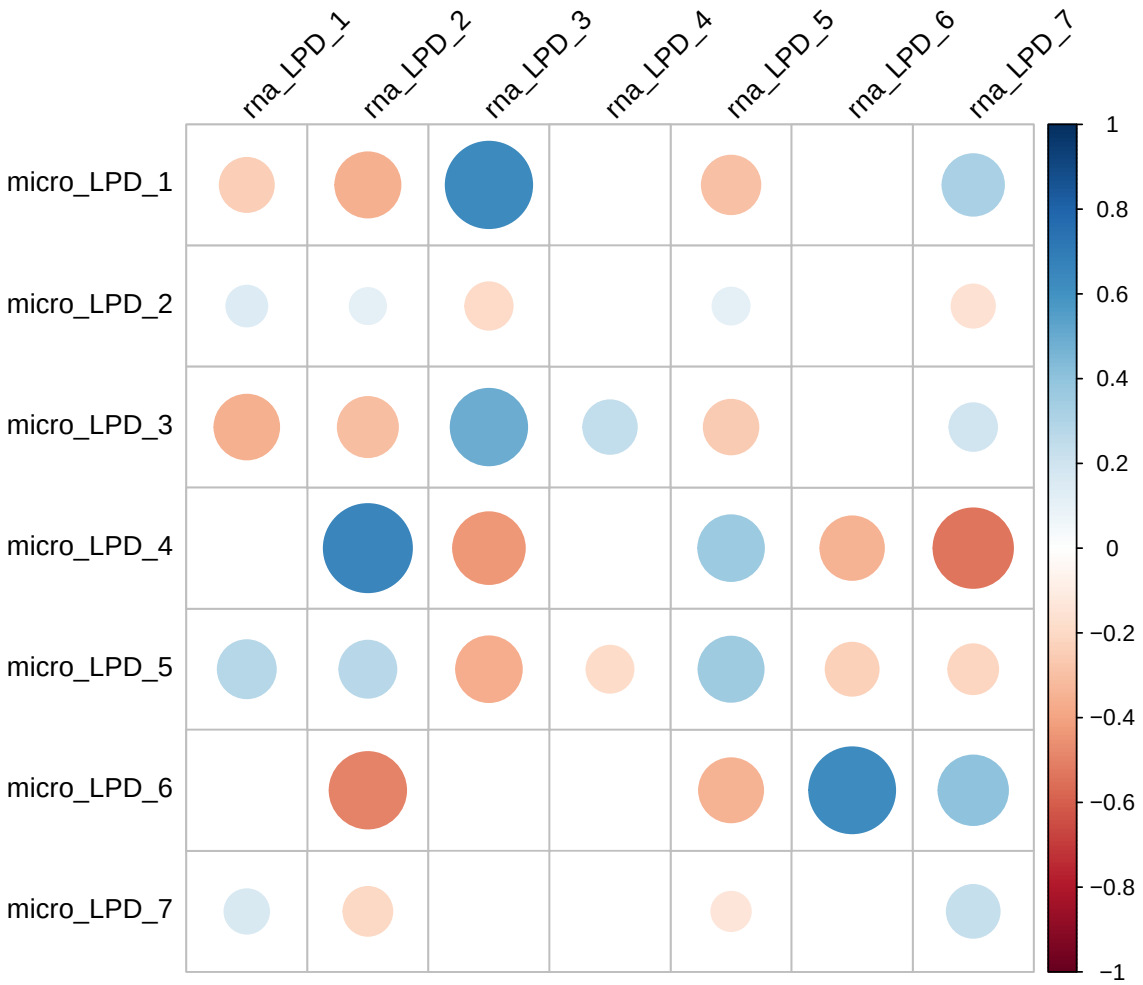

Figure C.9: Correlation matrix for TCGA-OV between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

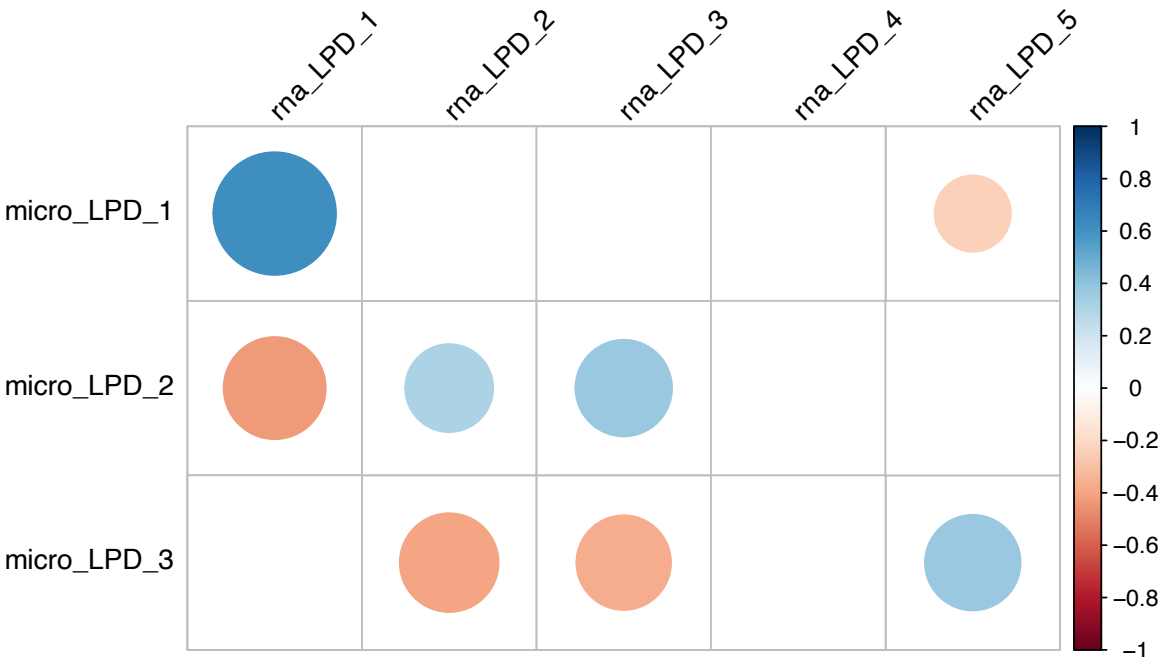

Figure C.10: Correlation matrix for TCGA-READ between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

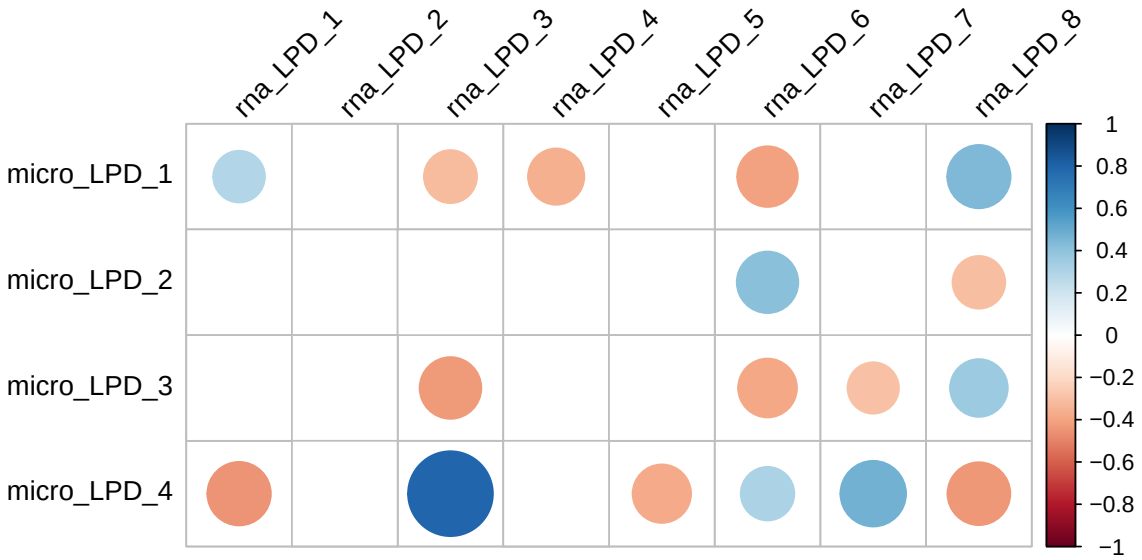

Figure C.11: Correlation matrix for TCGA-UCEC between the LPD groups assigned using the microarray platform (micro) and the LPD groups assigned when using the RNA-seq platform (rna). The color-coded matrix displays positive or negative correlations, with the intensity of the color indicating the strength of the correlation. For visual aid, only correlations above a threshold of 0.33 or below -0.33 are represented. Additionally, the size of the circle associated with each correlation is proportional to the strength of the correlation.

References

1. Vella F. *The cell. A molecular approach.* Vol 26. ASM Press; 1998:98-99. doi:10.1016/s0307-4412(98)00065-x
2. World Health Organization. Global health estimates: Leading causes of death. 2020:1-2. `https://www.who.int/data/gho/data/themes/mortality-and-global-health-estimates/ghe-leading-causes-of-death`. Accessed August 9, 2021.
3. UK CR. Worldwide cancer mortality statistics | Cancer Research UK. 2018:Oral cancer mortality trends over time. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer#heading-One%0Ahttp://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer%0Ahttps://www.cancerresearchuk.org/health-professional/cancer-stat`. Accessed January 31, 2019.
4. Hilhorst S, Lockey A, DEMOS. *A 'Ripple Effect' Analysis of Cancer's Wider Impact.*; 2020:26-38.
5. Ritchie H. What do people die from? - Our World in Data. 2018. `https://ourworldindata.org/what-does-the-world-die-from`. Accessed December 2, 2021.
6. Wei Dai YY. Genomic Instability and Cancer. *Journal of Carcinogenesis & Mutagenesis*. 2014;05(02). doi:10.4172/2157-2518.1000165
7. National Cancer Institute. Cancer-Causing Substances in the Environment - National Cancer Institute. 2015. `http://www.cancer.gov/about-cancer/causes-prevention/risk/substances`. Accessed April 8, 2019.
8. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature*. 2009;458(7239):719-724. doi:10.1038/nature07943
9. McFarland CD, Yaglom JA, Wojtkowiak JW, et al. The damaging effect of passenger mutations on cancer progression. *Cancer Research*. 2017;77(18):4763-4772. doi:10.1158/0008-5472.CAN-15-3283-T
10. Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000;100(1):57-70. doi:10.1016/S0092-8674(00)81683-9
11. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. *Cell*. 2011;144(5):646-674. doi:10.1016/j.cell.2011.02.013
12. Hanahan D. Hallmarks of Cancer: New Dimensions. 2022;12:31-46. doi:10.1158/2159-8290.CD-21-1059

13. Vinay DS, Ryan EP, Pawelec G, et al. Immune evasion in cancer: Mechanistic basis and therapeutic strategies. 2015;35:S185-S198. doi:10.1016/j.semcancer.2015.03.004

14. Pinheiro C, Garcia EA, Morais-Santos F, et al. Reprogramming energy metabolism and inducing angiogenesis: Co-expression of monocarboxylate transporters with VEGF family members in cervical adenocarcinomas. *BMC Cancer*. 2015;15(1):1-11. doi:10.1186/s12885-015-1842-4

15. Kowald A, Passos JF, Kirkwood TBL. On the evolution of cellular senescence. *Aging Cell*. 2020;19(12). doi:10.1111/acel.13270

16. Wang B, Kohli J, Demaria M. Senescent Cells in Cancer Therapy: Friends or Foes? *Trends in cancer*. 2020;6(10):838-857. doi:10.1016/J.TRECAN.2020.05.004

17. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. The Structure and Function of DNA. 2002. https://www.ncbi.nlm.nih.gov/books/NBK26821/.

18. Hampton O. Molecular Biology Review. 2008. `https://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html`. Accessed April 30, 2019.

19. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences of the United States of America*. 1977;74(8):3171-3175. doi:10.1073/pnas.74.8.3171

20. Chen L, Tovar-Corona JM, Urrutia AO. Alternative Splicing: A Potential Source of Functional Innovation in the Eukaryotic Genome. *International Journal of Evolutionary Biology*. 2012;2012:1-10. doi:10.1155/2012/596274

21. Berg JM(JeremyM, Tymoczko JL, Stryer Lubert, Stryer Lubert. *Biochemistry*. W.H. Freeman; 2002. https://www.ncbi.nlm.nih.gov/books/NBK21154/.

22. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Protein Function. 2002. https://www.ncbi.nlm.nih.gov/books/NBK26911/.

23. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in Endocrinology*. 2018;9(AUG):402. doi:10.3389/FENDO.2018.00402/BIBTEX

24. Herceg Z, Hainaut P. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Molecular Oncology*. 2007;1(1):26-41. doi:10.1016/j.molonc.2007.01.004

25. Katsonis P, Koire A, Wilson SJ, et al. Single nucleotide variations: Biological impact and theoretical interpretation. *Protein Science*. 2014;23(12):1650-1666. doi:10.1002/pro.2552

26. Zhang Z, Miteva MA, Wang L, Alexov E. Analyzing effects of naturally occurring missense mutations. 2012;2012:15. doi:10.1155/2012/805827

27. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*. 2017;45(D1):D777-D783. doi:10.1093/nar/gkw1121

28. Gaedcke J, Grade M, Jung K, et al. KRAS and BRAF mutations in patients with rectal cancer treated with preoperative chemoradiotherapy. *Radiotherapy and Oncology*. 2010;94(1):76-81. doi:10.1016/j.radonc.2009.10.001

29. Bercovich D, Ganmore I, Scott LM, et al. Mutations of JAK2 in acute lymphoblastic leukaemias associated with Down's syndrome. *The Lancet*. 2008;372(9648):1484-1492. doi:10.1016/S0140-6736(08)61341-0

30. Auner V, Kriegshäuser G, Tong D, et al. KRAS mutation analysis in ovarian samples using a high sensitivity biochip assay. *BMC Cancer*. 2009;9. doi:10.1186/1471-2407-9-111

31. American Society of Clinical Oncology. The Genetics of Cancer | Cancer.Net. 2018. `https://www.cancer.net/navigating-cancer-care/cancer-basics/genetics/genetics-cancer`. Accessed September 19, 2022.

32. Thompson SL, Compton DA. Chromosomes and cancer cells. 2011;19:433-444. doi:10.1007/s10577-010-9179-y

33. Kapur RP, Siebert JR. Chromosomal abnormalities. *Stocker and Dehner's Pediatric Pathology: Third Edition*. July 2012. https://www.ncbi.nlm.nih.gov/books/NBK115545/.

34. Panagopoulos I, Heim S. Interstitial deletions generating fusion genes. 2021;18:167-196. doi:10.21873/CGP.20251

35. Edwards PAW. Fusion genes and chromosome translocations in the common epithelial cancers. 2010;220:244-254. doi:10.1002/path.2632

36. Quinton RJ, DiDomizio A, Vittoria MA, et al. Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. *Nature*. 2021;590(7846):492-497. doi:10.1038/s41586-020-03133-3

37. Mitelman F, Johansson B, Mertens F. The impact of translocations and gene fusions on cancer causation. 2007;7:233-245. doi:10.1038/nrc2091

38. Albertson DG. Gene amplification in cancer. 2006;22:447-455. doi:10.1016/j.tig.2006.06.007

39. Pederzoli F, Bandini M, Marandino L, et al. Targetable gene fusions and aberrations in genitourinary oncology. 2020;17:613-625. doi:10.1038/s41585-020-00379-4

40. Liang S, Hu L, Wu Z, et al. CDK12: A Potent Target and Biomarker for Human Cancer Therapy. 2020;9. doi:10.3390/cells9061483

41. Sigismund S, Avanzato D, Lanzetti L. Emerging functions of the EGFR in cancer. 2018;12:3-20. doi:10.1002/1878-0261.12155

42. Kulis M, Esteller M. DNA Methylation and Cancer. In: *Advances in Genetics*. Vol 70.; 2010:27-56. doi:10.1016/B978-0-12-380866-0.60002-2

43. Jin B, Li Y, Robertson KD. DNA methylation: Superior or subordinate in the epigenetic hierarchy? *Genes and Cancer*. 2011;2(6):607-617. doi:10.1177/1947601910393957

44. McMahon KW, Karunasena E, Ahuja N. The Roles of DNA Methylation in the Stages of Cancer. 2017;23:257-261. doi:10.1097/PPO.0000000000000279

45. Costa FF, Paixão VA, Cavalher FP, et al. SATR-1 hypomethylation is a common and early event in breast cancer. *Cancer Genetics and Cytogenetics*. 2006;165(2):135-143. doi:10.1016/j.cancergencyto.2005.07.023

46. Widschwendter M, Jiang G, Woods C, et al. DNA hypomethylation and ovarian cancer biology. *Cancer Research*. 2004;64(13):4472-4480. doi:10.1158/0008-5472.CAN-04-0238

47. Xiang TX, Yuan Y, Li LL, et al. Aberrant promoter CpG methylation and its translational applications in breast cancer. 2013;32:12-20. doi:10.5732/cjc.011.10344

48. Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine. 2014;6:5. doi:10.1186/gm524

49. R.A. Weinberg. How cancer arises. *Scientific American*. 1996;(September):62-70. https://pdfs.semanticscholar.org/3b1c/7940f6ef101124cf478f5f8de2113ee1c5fb.pdf.

50. National Cancer Institutes: Seer Training Modules. What Is Cancer? | SEER Training. `https://training.seer.cancer.gov/disease/cancer/`. Accessed December 14, 2021.

51. Office for National Statistics. Office for National Statistics: Cancer Registration Statistics, England. 2014. `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/firstrelease2014`. Accessed January 31, 2019.

52. Office for National Statistics. Office for National Statistics: Cancer Registration Statistics, England. 2014. `https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/cancerregistrationstatisticsengland/firstrelease2014`. Accessed April 30, 2021.

53. Rossing M, Pedersen CB, Tvedskov T, et al. Clinical implications of intrinsic molecular subtypes of breast cancer for sentinel node status. *Scientific Reports*. 2021;11(1):2259. doi:10.1038/s41598-021-81538-4

54. Collisson EA, Bailey P, Chang DK, Biankin AV. Molecular subtypes of pancreatic cancer. 2019;16:207-220. doi:10.1038/s41575-019-0109-y

55. Malhotra GK, Zhao X, Band H, Band V. Histological, molecular and functional subtypes of breast cancers. 2010;10:955-960. doi:10.4161/cbt.10.10.13879

56. Rathore S, Akbari H, Rozycki M, et al. Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. *Scientific Reports*. 2018;8(1):5087. doi:10.1038/s41598-018-22739-2

57. Tran D, Nguyen H, Le U, Bebis G, Luu HN, Nguyen T. A novel method for cancer subtyping and risk prediction using consensus factor analysis. *Frontiers in Oncology*. 2020;10:1052. doi:10.3389/fonc.2020.01052

58. Breast cancer in women - NHS. 2019. `https://www.nhs.uk/conditions/breast-cancer/`. Accessed August 11, 2021.

59. Sun YS, Zhao Z, Yang ZN, et al. Risk factors and preventions of breast cancer. 2017;13:1387-1397. doi:10.7150/ijbs.21635

60. Mayer M. Metastatic Breast Cancer: Symptoms, Treatment, and More. 2021. `https://www.breastcancer.org/symptoms/types/recur_metast`. Accessed September 8, 2021.

61. Redig AJ, Mcallister SS. Breast cancer as a systemic disease: A view of metastasis. 2013;274:113-126. doi:10.1111/joim.12084

62. Cancer Research UK. Bowel cancer survival statistics | Cancer Research UK. 2014. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/survival`. Accessed January 3, 2022.

63. Roy R, Chun J, Powell SN. BRCA1 and BRCA2: Different roles in a common pathway of genome protection. 2012;12:68-78. doi:10.1038/nrc3181

64. Horn J, Åsvold BO, Opdahl S, Tretli S, Vatten LJ. Reproductive factors and the risk of breast cancer in old age: A Norwegian cohort study. *Breast Cancer Research and Treatment*. 2013;139(1):237-243. doi:10.1007/s10549-013-2531-0

65. Banks E, Beral V, Bull D, et al. Breast cancer and hormone-replacement therapy in the Million Women Study. *Lancet*. 2003;362(9382):419-427. doi:10.1016/S0140-6736(03)14065-2

66. Jung S, Wang M, Anderson K, et al. Alcohol consumption and breast cancer risk by estrogen receptor status: In a pooled analysis of 20 studies. *International Journal of Epidemiology*. 2016;45(3):916-928. doi:10.1093/ije/dyv156

67. NHS. MRI scan - NHS. 2018. `https://www.nhs.uk/conditions/mri-scan/`. Accessed September 8, 2021.

68. Morrow M, Waters J, Morris E. MRI for breast cancer screening, diagnosis, and treatment. 2011;378:1804-1811. doi:10.1016/S0140-6736(11)61350-0

69. West Midlands Expert Advisory Group Breast Cancer. Clinical Guidelines for the Management of for Breast Cancer. *Clinical Guidelines*. 2015:12. https://www.england.nhs.uk/mids-east/wp-content/uploads/sites/7/2018/02/guidelines-for-the-management-of-breast-cancer-v1.pdf.

70. Dalle JR, Leow WK, Racoceanu D, Tutac AE, Putti TC. Automatic breast cancer grading of histopathological images. In: *Proceedings of the 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS'08 - "Personalized Healthcare Through Technology"*. IEEE Computer Society; 2008:3052-3055. doi:10.1109/iembs.2008.4649847

71. Cancer.Net. Breast Cancer: Stages | Cancer.Net. 2020. `https://www.cancer.net/cancer-types/breast-cancer/stages`. Accessed October 18, 2021.

72. Dai X, Li T, Bai Z, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. 2015;5:2929-2943. /pmc/articles/PMC4656721/ /pmc/articles/PMC4656721/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4656721/.

73. Vallejos C, Gómez H, Cruz W, et al. Breast cancer classification according to immunohistochemistry markers: Subtypes and association with clinicopathologic variables in a peruvian hospital database. *Clinical Breast Cancer*. 2010;10(4):294-300. doi:10.3816/CBC.2010.n.038

74. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(19):10869-10874. doi:10.1073/pnas.191367098

75. Fragomeni SM, Sciallis A, Jeruss JS. Molecular Subtypes and Local-Regional Control of Breast Cancer. 2018;27:95-120. doi:10.1016/j.soc.2017.08.005

76. Haibe-Kains B, Desmedt C, Loi S, et al. A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*. 2012;104(4):311-325. doi:10.1093/jnci/djr545

77. Rouzier R, Perou CM, Symmans WF, et al. Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical Cancer Research*. 2005;11(16):5678-5685. doi:10.1158/1078-0432.CCR-04-2421

78. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. 2014;5:412-424. doi:10.5306/wjco.v5.i3.412

79. Almarzooq R, Alrayes A, Alaradi H, Abdulla H. Molecular subtypes of breast cancer. *Bahrain Medical Bulletin*. 2018;40(4):222-225. https://www.breastcancer.org/symptoms/types/molecular-subtypes.

80. Parker J, Prat A, Cheang M, Lenburg M, Paik S, Perou C. Breast Cancer Molecular Subtypes Predict Response to Anthracycline/Taxane-Based Chemotherapy. In: *Cancer Research*. Vol 69. American Association for Cancer Research (AACR); 2009:2019-2019. doi:10.1158/0008-5472.sabcs-09-2019

81. Creighton CJ. The molecular profile of luminal B breast cancer. 2012;6:289-297. doi:10.2147/BTT.S29923

82. Pu M, Messer K, Davies SR, et al. Research-based PAM50 signature and long-term breast cancer survival. *Breast Cancer Research and Treatment*. 2020;179(1):197. doi:10.1007/S10549-019-05446-Y

83. Moo TA, Sanford R, Dang C, Morrow M. Overview of Breast Cancer Therapy. 2018;13:339-354. doi:10.1016/j.cpet.2018.02.006

84. Stebbing J, Copson E, O'Reilly S. Herceptin (trastuzamab) in advanced breast cancer. 2000;26:287-290. doi:10.1053/ctrv.2000.0182

85. Hoffman M. Prostate Gland (Human Anatomy): Prostate Picture, Definition, Function, Conditions, Tests, and Treatments. 2014. `https://www.webmd.com/urinary-incontinence-oab/picture-of-the-prostate#1`. Accessed April 3, 2019.

86. Ebili HO, Olayiwola AO, Olopade OI. Molecular subtypes of breast cancer. *Personalized Management of Breast Cancer*. 2014:21-33. doi:10.2217/EBO.13.374

87. The American Cancer Society. Key Statistics for Prostate Cancer. 2022. https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html.

88. NICE. Introduction | Prostate cancer: diagnosis and management | Guidance | NICE. 2014. https://www.nice.org.uk/guidance/cg175/chapter/Introduction.

89. Rawla P. Epidemiology of Prostate Cancer. *World Journal of Oncology*. 2019;10(2):63. doi:10.14740/WJON1191

90. Panigrahi GK, Praharaj PP, Kittaka H, et al. Exosome proteomic analyses identify inflammatory phenotype and novel biomarkers in African American prostate cancer patients. *Cancer medicine*. 2019;8(3):1110-1123. doi:10.1002/CAM4.1885

91. NHS. Prostate cancer - Symptoms - NHS. 2018. `https://www.nhs.uk/conditions/prostate-cancer/symptoms/`. Accessed April 8, 2019.

92. Leitzmann MF, Rohrmann S. Risk factors for the onset of prostatic cancer: Age, location, and behavioral correlates. *Clinical Epidemiology*. 2012;4(1):1-11. doi:10.2147/CLEP.S16747

93. Cancer Research UK. Cancer Research UK Cancer incidence statistics. 2018. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer/incidence#heading-One`. Accessed April 26, 2019.

94. Dall'era MA, Cooperberg MR, Chan JM, et al. Active surveillance for early-stage prostate cancer: Review of the current literature. *Cancer*. 2008;112(8):1650-1659. doi:10.1002/cncr.23373

95. NHS. Prostate cancer - Diagnosis - NHS. 2018. `https://www.nhs.uk/conditions/prostate-cancer/diagnosis/`. Accessed April 4, 2019.

96. Oremek GM, Seiffert UB. Physical activity releases prostate-specific antigen (PSA) from the prostate gland into blood and increases serum PSA concentrations. *Clinical Chemistry*. 1996;42(5):691-695. doi:10.1093/clinchem/42.5.691

97. Demir K, Tarhan F, Orçun A, Aslan H, Türk A. Effects of ejaculation on serum prostate-specific antigen levels. *Turk Uroloji Dergisi*. 2014;40(1):40-45. doi:10.5152/tud.2014.03704

98. NHS. Why should I avoid sexual activity before a PSA test? - NHS. 2017. `https://www.nhs.uk/common-health-questions/mens-health/why-should-i-avoid-sexual-activity-before-a-psa-test/`. Accessed April 30, 2019.

99. Azab S, Osama A, Rafaat M. Does normalizing PSA after successful treatment of chronic prostatitis with high PSA value exclude prostatic biopsy? *Translational Andrology and Urology*. 2012;1(3):148-152. doi:10.3978/j.issn.2223-4683.2012.07.02

100. Hosain GMM, Sanderson M, Du XL, Chan W, Strom SS. Racial/ethnic differences in predictors of PSA screening in a tri-ethnic population. *Central European Journal of Public Health*. 2011;19(1):30-34. doi:10.21101/cejph.a3622

101. Adhyam M, Gupta AK. A Review on the Clinical Utility of PSA in Cancer Prostate. 2012;3:120-129. doi:10.1007/s13193-012-0142-6

102. Meigs JB, Barry MJ, Oesterling JE, Jacobsen SJ. Interpreting results of prostate-specific antigen testing for early detection of prostate cancer. *Journal of General Internal Medicine*. 1996;11(9):505-512. doi:10.1007/BF02599596

103. Mistry K, Cable G. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. *Journal of the American Board of Family Practice*. 2003;16(2):95-101. doi:10.3122/jabfm.16.2.95

104. Paul B, Dhir R, Landsittel D, Hitchens MR, Getzenberg RH. Detection of prostate cancer with a blood-based assay for early prostate cancer antigen. *Cancer Research*. 2005;65(10):4097-4100. doi:10.1158/0008-5472.CAN-04-4523

105. Varambally S, Laxman B, Mehra R, et al. Golgi protein GOLM1 is a tissue and urine biomarker of prostate cancer. *Neoplasia*. 2008;10(11):1285-1294. doi:10.1593/neo.08922

106. Obort AS, Ajadi MB, Akinloye O. Prostate-specific antigen: any successor in sight? *Reviews in urology*. 2013;15(3):97-107. http://www.ncbi.nlm.nih.gov/pubmed/24223021%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi

107. Humphrey PA. Gleason grading and prognostic factors in carcinoma of the prostate. 2004;17:292-306. doi:10.1038/modpathol.3800054

108. Prostate Cancer UK. Scans to see if your cancer has spread | Prostate Cancer UK. 2019. `https://prostatecanceruk.org/prostate-information/prostate-tests/scans-to-see-if-your-cancer-has-spread`. Accessed April 4, 2019.

109. National Cancer Institute. Cancer Staging - National Cancer Institute. 2015. https://www.cancer.gov/about-cancer/diagnosis-staging/staging.

110. Cancer.net. Prostate Cancer: Stages and Grades | Cancer.Net. 2020. `https://www.cancer.net/cancer-types/prostate-cancer/stages-and-grades`. Accessed October 18, 2021.

111. American Cancer Society. Prostate Cancer Stages. 2021.

112. Radical Prostatectomy | Johns Hopkins Medicine. 2021. `https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/radical-prostatectomy`. Accessed August 10, 2021.

113. Tourinho-Barbosa RR, Srougi V, Nunes-Silva I, et al. Biochemical recurrence after radical prostatectomy: What does it mean? 2018;44:14-21. doi:10.1590/S1677-5538.IBJU.2016.0656

114. Kaffenberger SD, Barbieri CE. Molecular subtyping of prostate cancer. 2016;26:213-218. doi:10.1097/MOU.0000000000000285

115. Abeshouse A, Ahn J, Akbani R, et al. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*. 2015;163(4):1011-1025. doi:10.1016/j.cell.2015.10.025

116. Arora K, Barbieri CE. Molecular Subtypes of Prostate Cancer. 2018;20:1-9. doi:10.1007/s11912-018-0707-9

117. Ross-Adams H, Lamb A, Dunning M, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine*. 2015;2(9):1133-1144. doi:10.1016/j.ebiom.2015.07.017

118. Luca BA, Brewer DS, Edwards DR, et al. DESNT: A Poor Prognosis Category of Human Prostate Cancer. *European Urology Focus*. 2018;4(6):842-850. doi:10.1016/j.euf.2017.01.016

119. Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2005;2(2):143-156. doi:10.1109/TCBB.2005.29

120. Luca BA, Moulton V, Ellis C, et al. A novel stratification framework for predicting outcome in patients with prostate cancer. *British Journal of Cancer*. 2020;122(10):1467-1476. doi:10.1038/s41416-020-0799-5

121. American Cancer Society. Lung Cancer Statistics. 2021. `https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html`. Accessed September 13, 2021.

122. NHS. Lung cancer - NHS. 2019. `https://www.nhs.uk/conditions/lung-cancer/`. Accessed September 13, 2021.

123. Malvezzi M, Carioli G, Bertuccio P, et al. European cancer mortality predictions for the year 2016 with focus on leukaemias. *Annals of Oncology*. 2016;27(4):725-731. doi:10.1093/annonc/mdw022

124. Press E. September 2021. https://www.infosalus.com/salud-investigacion/noticia-supervivencia-cancer-pulmon-descendido-retrasos-diagnosticos-causados-pandemia-20210913110027.html.

125. Malhotra J, Malvezzi M, Negri E, La Vecchia C, Boffetta P. Risk factors for lung cancer worldwide. *European Respiratory Journal*. 2016;48(3):889-902. doi:10.1183/13993003.00359-2016

126. Tram KL, Gallicchio L, Lindsley K, et al. Cruciferous vegetable consumption and lung cancer risk: A systematic review. *Cancer Epidemiology Biomarkers and Prevention*. 2009;18(1):184-195. doi:10.1158/1055-9965.EPI-08-0710

127. Guertin KA, Freedman ND, Loftfield E, Graubard BI, Caporaso NE, Sinha R. Coffee consumption and incidence of lung cancer in the NIH-AARP Diet and Health Study. 2016;45:929-939. doi:10.1093/ije/dyv104

128. Fedirko V, Tramacere I, Bagnardi V, et al. Alcohol drinking and colorectal cancer risk: An overall and dose-Response meta-analysis of published studies. 2011;22:1958-1972. doi:10.1093/annonc/mdq653

129. Santillan AA, Camargo CA, Colditz GA. A meta-analysis of asthma and risk of lung cancer (United States). 2003;14:327-334. doi:10.1023/A:1023982402137

130. Zheng W, Blot WJ, Liao ML, et al. Lung cancer and prior tuberculosis infection in Shanghai. *British Journal of Cancer*. 1987;56(4):501-504. doi:10.1038/bjc.1987.233

131. Lissowska J, Bardin-Mikolajczak A, Fletcher T, et al. Lung cancer and indoor pollution from heating and cooking with solid fuels: The IARC international multicentre case-control study in Eastern/Central Europe and the United Kingdom. *American Journal of Epidemiology*. 2005;162(4):326-333. doi:10.1093/aje/kwi204

132. Rushton L, Hutchings SJ, Fortunato L, et al. Occupational cancer burden in Great Britain. *British Journal of Cancer*. 2012;107(Suppl 1):S3-S7. doi:10.1038/bjc.2012.112

133. Bagnardi V, Rota M, Botteri E, et al. Alcohol consumption and lung cancer risk in never smokers: A meta-analysis. *Annals of Oncology*. 2011;22(12):2631-2639. doi:10.1093/annonc/mdr027

134. Djoussé L, Dorgan JF, Zhang Y, et al. Alcohol consumption and risk of lung cancer: The Framingham study. *Journal of the National Cancer Institute*. 2002;94(24):1877-1882. doi:10.1093/jnci/94.24.1877

135. Fehringer G, Brenner DR, Zhang ZF, et al. Alcohol and lung cancer risk among never smokers: A pooled analysis from the international lung cancer consortium and the SYNERGY study. *International Journal of Cancer*. 2017;140(9):1976-1984. doi:10.1002/ijc.30618

136. Zheng M. Classification and Pathology of Lung Cancer. 2016;25:447-468. doi:10.1016/j.soc.2016.02.003

137. Society TAC. About Lung Cancer. 2019. `https://www.cancer.org/cancer/lung-cancer/about/what-is.html`. Accessed September 13, 2021.

138. Alvarado-Luna G, Morales-Espinosa D. Treatment for small cell lung cancer, where are we now?-A review. 2016;5:26-38. doi:10.3978/j.issn.2218-6751.2016.01.13

139. Cooper S, Spiro SG. Small cell lung cancer: Treatment review. 2006;11:241-248. doi:10.1111/j.1440-1843.2006.00850.x

140. Zarogoulidis K, Zarogoulidis P, Darwiche K, et al. Treatment of non-small cell lung cancer (NSCLC). 2013;5:S389. doi:10.3978/j.issn.2072-1439.2013.07.10

141. Ozdemir O, Russell RL, Berlin AA. A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans. *IEEE Transactions on Medical Imaging*. 2020;39(5):1419-1429. doi:10.1109/TMI.2019.2947595

142. American Cancer Society. *Lung Cancer Early Detection, Diagnosis, and Staging.*; 2019.

143. Detterbeck FC, Lewis SZ, Diekemper R, Addrizzo-Harris D, Alberts WM. Executive Summary: Diagnosis and Management of Lung Cancer, 3rd ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. *Chest*. 2013;143(5):7S-37S. doi:10.1378/chest.12-2377

144. American Cancer Society. Non-Small Cell Lung Cancer Stages. 2019. `https://www.cancer.org/cancer/lung-cancer/detection-diagnosis-staging/staging-nsclc.html`. Accessed October 19, 2021.

145. NHS. Lung cancer - NHS. 2019. `https://www.nhs.uk/conditions/lung-cancer/`. Accessed October 8, 2021.

146. Dempke WCM, Suto T, Reck M. Targeted therapies for non-small cell lung cancer. 2010;67:257-274. doi:10.1016/j.lungcan.2009.10.012

147. Beer DG, Kardia SLR, Huang CC, et al. Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine*. 2002;8(8):816-824. doi:10.1038/nm733

148. Collisson EA, Campbell JD, Brooks AN, et al. Comprehensive molecular profiling of lung adenocarcinoma: The cancer genome atlas research network. *Nature*. 2014;511(7511):543-550. doi:10.1038/nature13385

149. Larsen JE, Pavey SJ, Passmore LH, et al. Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis*. 2007;28(3):760-766. doi:10.1093/carcin/bgl207

150. Inamura K, Fujiwara T, Hoshida Y, et al. Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene*. 2005;24(47):7105-7113. doi:10.1038/sj.onc.1208858

151. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research*. 2006;66(15):7466-7472. doi:10.1158/0008-5472.CAN-06-1191

152. Wilkerson MD, Yin X, Hoadley KA, et al. Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clinical Cancer Research*. 2010;16(19):4864-4875. doi:10.1158/1078-0432.CCR-10-0199

153. Hammerman PS, Voet D, Lawrence MS, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489(7417):519-525. doi:10.1038/nature11404

154. Faruki H, Mayhew GM, Serody JS, Hayes DN, Perou CM, Lai-Goldman M. Lung Adenocarcinoma and Squamous Cell Carcinoma Gene Expression Subtypes Demonstrate Significant Differences in Tumor Immune Landscape. *Journal of Thoracic Oncology*. 2017;12(6):943-953. doi:10.1016/j.jtho.2017.03.010

155. Eldridge L. Squamous Cell Carcinoma of the Lungs: Symptoms and Treatment. 2019. `https://www.verywellhealth.com/squamous-cell-carcinoma-of-the-lungs-in-depth-2249367`. Accessed October 11, 2021.

156. Cancer Research UK. Bowel cancer statistics | Cancer Research UK. 2017. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero.

157. NHS. Bowel cancer - Symptoms - NHS. 2019. `https://www.nhs.uk/conditions/bowel-cancer/symptoms/`. Accessed October 12, 2021.

158. Slide Team. 0514 Anatomy Of Large Intestine Medical Images For Powerpoint | PPT Images Gallery | PowerPoint Slide Show | PowerPoint Presentation Templates. 2021. `https://www.slideteam.net/0514-anatomy-of-large-intestine-medical-images-for-powerpoint.html`. Accessed October 19, 2021.

159. Haggar FA, Boushey RP. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. *Clinics in Colon and Rectal Surgery*. 2009;22(4):191-197. doi:10.1055/s-0029-1242458

160. Janout V, Kollárová H. Epidemiology of colorectal cancer. *Biomedical papers of the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia*. 2001;145(1):5-10. doi:10.5507/bp.2001.001

161. De Jong AE, Morreau H, Nagengast FM, et al. Prevalence of adenomas among young individuals at average risk for colorectal cancer. *American Journal of Gastroenterology*. 2005;100(1):139-143. doi:10.1111/j.1572-0241.2005.41000.x

162. Stidham RW, Higgins PDR. Colorectal Cancer in Inflammatory Bowel Disease. *Clinics in Colon and Rectal Surgery*. 2018;31(3):168-178. doi:10.1055/s-0037-1602237

163. Limsui D, Vierkant RA, Tillmans LS, et al. Cigarette smoking and colorectal cancer risk by molecularly defined subtypes. *Journal of the National Cancer Institute*. 2010;102(14):1012-1022. doi:10.1093/jnci/djq201

164. GOV.UK. Bowel cancer screening: programme overview - GOV.UK. 2019. https://www.gov.uk/guidance/bowel-cancer-screening-programme-overview#target-population https://www.gov.uk/guidance/bowel-cancer-screening-programme-overview. Accessed October 5, 2021.

165. Cancer.Net. Colorectal Cancer: Stages | Cancer.Net. *American Society of Clinical Oncology (ASCO)*. 2020:1-5. https://www.cancer.net/cancer-types/colorectal-cancer/stages.

166. The American Cancer Society. Colorectal Cancer Stages | Rectal Cancer Staging | Colon Cancer Staging. 2020. `https://www.cancer.org/cancer/colon-rectal-cancer/detection-diagnosis-staging/staged.html`. Accessed October 19, 2021.

167. Cancer Research UK. Bowel cancer | Cancer Research UK. 2018. https://www.cancerresearchuk.org/about-cancer/bowel-cancer/treatment https://www.cancerresearchuk.org/about-cancer/bowel-cancer. Accessed October 5, 2021.

168. Muzny DM, Bainbridge MN, Chang K, et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337. doi:10.1038/nature11252

169. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*. 2015;21(11):1350-1356. doi:10.1038/nm.3967

170. Stintzing S, Wirapati P, Lenz HJ, et al. Consensus molecular subgroups (CMS) of colorectal cancer (CRC) and first-line efficacy of FOLFIRI plus cetuximab or bevacizumab in the FIRE3 (AIO KRK-0306) trial. *Annals of Oncology*. 2019;30(11):1796-1803. doi:10.1093/annonc/mdz387

171. Ellis C. Using latent process decomposition to classify prostate and colorectal cancers. August 2021.

172. National Institutes of Health. The Cancer Genome Atlas, Program Overview. 2018. `https://cancergenome.nih.gov/abouttcga/overview`. Accessed April 10, 2019.

173. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. 2015;1A:A68-A77. doi:10.5114/wo.2014.47136

174. NIH. TCGA Cancers selected for study. `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/studied-cancers`. Accessed April 10, 2019.

175. The Cancer Genome Atlas. TCGA Barcode - GDC Docs. 2021. `https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/`. Accessed October 25, 2021.

176. Gao GF, Parker JS, Reynolds SM, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Systems*. 2019;9(1):24-34.e10. doi:10.1016/j.cels.2019.06.006

177. The Cancer Genome Atlas. Harmonized Data - GDC Docs. `https://docs.gdc.cancer.gov/Encyclopedia/pages/Harmonized_Data/`. Accessed October 25, 2021.

178. De Bellis G, Battaglia C. Introduction to DNA microarray technologies. *Seminars in Organic Synthesis*. 2006;15:59-65. doi:10.1002/0471670278.ch1

179. Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *American Journal of Obstetrics and Gynecology*. 2006;195(2):373-388. doi:10.1016/j.ajog.2006.07.001

180. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. 2009;10:57-63. doi:10.1038/nrg2484

181. Illumina. RNA-Seq Technology | Comparison vs Microarrays. 2019. https://emea.illumina.com/science/technology/next-generation-sequencing/microarray-rna-seq-comparison.html https://www.illumina.com/science/technology/next-generation-sequencing/microarray-rna-seq-comparison.html. Accessed April 27, 2019.

182. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. Zhang S-D, ed. *PLoS ONE*. 2014;9(1):e78644. doi:10.1371/journal.pone.0078644

183. Martin SAM, Dehler CE, Król E. Transcriptomic responses in the fish intestine. 2016;64:103-117. doi:10.1016/j.dci.2016.03.014

184. Otogenetics. RNA Sequencing VS Microarray - Otogenetics. 2022. `https://www.otogenetics.com/rna-sequencing-vs-microarray/`. Accessed January 26, 2022.

185. Illumina. Methylation Arrays | Analyze methylation sites across the genome. 2021. `https://www.illumina.com/techniques/microarrays/methylation-arrays.html`. Accessed October 29, 2021.

186. Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. 2012;22(3):568-576. doi:10.1101/gr.129684.111

187. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*. 2013;31(3):213-219. doi:10.1038/nbt.2514

188. Fan Y, Xi L, Hughes DST, et al. MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome biology*. 2016;17(1):178. doi:10.1186/s13059-016-1029-6

189. Larson DE, Harris CC, Chen K, et al. Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. *Bioinformatics*. 2012;28(3):311-317. doi:10.1093/bioinformatics/btr665

190. National Cancer Institute. Clinical Data - GDC Docs. `https://docs.gdc.cancer.gov/Encyclopedia/pages/Clinical_Data/`. Accessed November 1, 2021.

191. National Cancer Institute Genomic Data Commons Data Model Working Group. Selecting Common Cross-Study Clinical Data Elements | NCI Genomic Data Commons. 2017. https://gdc.cancer.gov/node/777/ https://gdc.cancer.gov/documentation/selecting-common-cross-study-clinical-data-elements. Accessed November 1, 2021.

192. Yu A. How Netflix Uses AI and Machine Learning. 2019;24:2019. https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe https://becominghuman.ai/how-netflix-uses-ai-and-machine-learning-a087614630fe%3E.

193. Ciocco S. How Does Spotify Know You So Well? | by Sophia Ciocca | Medium. 2017. `https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe`. Accessed April 19, 2019.

194. Sathya R, Abraham A. Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. *International Journal of Advanced Research in Artificial Intelligence*. 2013;2(2). doi:10.14569/ijarai.2013.020206

195. Sauravkaushik K. Clustering | Types Of Clustering | Clustering Applications. 2020. `https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/`. Accessed September 20, 2022.

196. Moseley B, Wang JR. Approximation bounds for hierarchical clustering: Average linkage, bisecting K-means, and Local Search. In: *Advances in Neural Information Processing Systems*. Vol 2017-Decem.; 2017:3095-3104. https://papers.nips.cc/paper/6902-approximation-bounds-for-hierarchical-clustering-average-linkage-bisecting-k-means-and-local-search.

197. Bora DJ, Gupta DrAK. A Comparative study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. *International Journal of Computer Trends and Technology*. 2014;10(2):108-113. doi:10.14445/22312803/ijctt-v10p119

198. Kassambara A. Determining The Optimal Number Of Clusters: 3 Must Know Methods - Datanovia. 2020. `https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/#elbow-method`. Accessed September 20, 2022.

199. Vardhan A, Sarmah P, Das A. A Comprehensive Analysis of the Most Common Hard Clustering Algorithms. *Lecture Notes in Networks and Systems*. 2020;98:48-58. doi:10.1007/978-3-030-33846-6_6/COVER

200. Bock T. What is Hierarchical Clustering? | Displayr.com. 2020. `https://www.displayr.com/what-is-hierarchical-clustering/`. Accessed April 16, 2019.

201. Lin IH, Chen DT, Chang YF, et al. Hierarchical clustering of breast cancer methylomes revealed differentially methylated and expressed breast cancer genes. El-Maarri O, ed. *PLoS ONE*. 2015;10(2):e0118453. doi:10.1371/journal.pone.0118453

202. Glen S. Hierarchical Clustering / Dendrogram: Simple Definition, Examples - Statistics How To. 2016. `https://www.statisticshowto.com/hierarchical-clustering/`. Accessed February 11, 2022.

203. Al-Masri A. How Does k-Means Clustering in Machine Learning Work? | by Anas Al-Masri | Towards Data Science. 2019. `https://towardsdatascience.com/how-does-k-means-clustering-in-machine-learning-work-fdaaaf5acfa0`. Accessed September 20, 2022.

204. Muzhingi I. K-Means clustering made simple | Oxford Protein Informatics Group. 2020. `https://www.blopig.com/blog/2020/07/k-means-clustering-made-simple/`. Accessed September 25, 2022.

205. Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray data sets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. 2005;2(2):143-156. doi:10.1109/TCBB.2005.29

206. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2003;3:993-1022. doi:10.1162/jmlr.2003.3.4-5.993

207. Bogdan-Alexandru L. Identification of biomarkers for the management of human prostate cancer. 2017.

208. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival Analysis Part I: Basic concepts and first analyses. 2003;89:232-238. doi:10.1038/sj.bjc.6601118

209. Kishore J, Goel M, Khanna P. Understanding survival analysis: Kaplan-Meier estimate. *International Journal of Ayurveda Research.* 2010;1(4):274. doi:10.4103/0974-7788.76794

210. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival Analysis Part II: Multivariate data analysis- An introduction to concepts and methods. 2003;89:431-436. doi:10.1038/sj.bjc.6601119

211. McDonald JH. Data Transformations. In: *Biological Statistics.* LibreTexts; 2022. https://stats.libretexts.org/Bookshelves/Applied_Statistics/Book%3A_Biological_Statistics_(Mc

212. Klein FA, Pakozdi T, Anders S, Ghavi-Helm Y, Furlong EEM, Huber W. FourCSeq: Analysis of 4C sequencing data. *Bioinformatics.* 2015;31(19):3085-3091. doi:10.1093/bioinformatics/btv335

213. Journey TBM. 11. Correlation and regression. `https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression`. Accessed May 15, 2022.

214. Schober P, Schwarte LA. Correlation coefficients: Appropriate use and interpretation. *Anesthesia and Analgesia.* 2018;126(5):1763-1768. doi:10.1213/ANE.0000000000002864

215. Hazra A, Gogtay N. Biostatistics series module 3: Comparing groups: Numerical variables. *Indian Journal of Dermatology.* 2016;61(3):251-260. doi:10.4103/0019-5154.182416

216. Raystuckey1. Normal Distribution vs. t-distribution – GeoGebra. `https://www.geogebra.org/m/xp7A3A53`. Accessed June 22, 2022.

217. Kim H-Y. Statistical notes for clinical researchers: Nonparametric statistical methods: 2. Nonparametric methods for comparing three or more groups and repeated measures. *Restorative Dentistry & Endodontics.* 2014;39(4):329. doi:10.5395/rde.2014.39.4.329

218. Sthle L, Wold S. Analysis of variance (ANOVA). 1989;6:259-272. doi:10.1016/0169-7439(89)80095-4

219. McHugh ML. The Chi-square test of independence. *Biochemia Medica.* 2012;23(2):143-149. doi:10.11613/BM.2013.018

220. Statistics How to. Tukey Test / Tukey Procedure / Honest Significant Difference - Statistics How To. 2019. `https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/post-hoc/tukey-test-honest-significant-difference/`. Accessed September 30, 2022.

221. DeepAI. Kernel Density Estimation Definition | DeepAI. 2020. `https://deepai.org/machine-learning-glossary-and-terms/kernel-density-estimation`. Accessed September 22, 2022.

222. Kamperis S. A gentle introduction to kernel density estimation | Let's talk about science! 2020. `https://ekamperi.github.io/math/2020/12/08/kernel-density-estimation.html`. Accessed September 22, 2022.

223. Ivchenko GI, Honov SA. On the jaccard similarity test. *Journal of Mathematical Sciences 1998 88:6*. 1998;88(6):789-794. doi:10.1007/BF02365362

224. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. 2015;43(7):e47-e47. doi:10.1093/NAR/GKV007

225. The R Foundation. R: What is R? 2022. `https://www.r-project.org/about.html`. Accessed May 9, 2022.

226. Hackenberger BK. R software: Unfriendly but probably the best. *Croatian Medical Journal*. 2020;61(1):66-68. doi:10.3325/cmj.2020.61.66

227. Kumar Dhanda S, Kumar R, Giorgi FM, Ceraolo C, Mercatelli D. The R Language: An Engine for Bioinformatics and Data Science. *Life 2022, Vol 12, Page 648*. 2022;12(5):648. doi:10.3390/LIFE12050648

228. Bartlein PJ. Markdown and Html. 2019. `https://pjbartlein.github.io/REarthSysSci/markdown.html`. Accessed May 9, 2022.

229. Grolemund G. Introduction to R Markdown. 2014. doi:10.1201/9780429341823-11

230. Sievert C, Iannone R, Allaire J BB. flexdashboard: R markdown format for flexible dashboards. 2022;1. `https://pkgs.rstudio.com/flexdashboard/`. Accessed May 9, 2022.

231. Haymond S. Create laboratory business intelligence dashboards for free using R: A tutorial using the flexdashboard package. *Journal of Mass Spectrometry and Advances in the Clinical Lab*. 2022;23:39-43. doi:10.1016/j.jmsacl.2021.12.002

232. University of East Anglia. High Performance Computing - History of HPC at UEA - Groups and Centres. 2020. `https://www.uea.ac.uk/groups-and-centres/research-and-specialist-computing/high-performance-computing/history-of-hpc-at-uea`. Accessed April 20, 2022.

233. Slurm. Slurm Workload Manager - Overview. 2022. `https://slurm.schedmd.com/overview.html`. Accessed September 30, 2022.

234. Tian L, Greenberg SA, Kong SW, Altschuler J, Kohane IS, Park PJ. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(38):13544-13549. doi:10.1073/pnas.0506577102

235. Forbes SA, Beare D, Boutselakis H, et al. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Research*. 2017;45(D1):D777-D783. doi:10.1093/nar/gkw1121

236. Tate JG, Bamford S, Jubb HC, et al. COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015

237. Kanehisa M, Sato Y, Kawashima M. KEGG mapping tools for uncovering hidden features in biological data. *Protein Science*. 2022;31(1):47-53. doi:10.1002/pro.4172

238. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. 2000;28:27-30. doi:10.1093/nar/28.1.27

239. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*. 2017;45(D1):D353-D361. doi:10.1093/nar/gkw1092

240. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: Tool for the unification of biology. 2000;25:25-29. doi:10.1038/75556

241. Gene Ontology. Gene Ontology overview. 2022. `http://geneontology.org/docs/ontology-documentation/`. Accessed September 22, 2022.

242. Heiman D. FAQ - Broad TCGA GDAC - Confluence. 2018. `https://broadinstitute.atlassian.net/wiki/spaces/GDAC/pages/844334036/FAQ`. Accessed March 20, 2022.

243. Luca B-A, Brewer DS, Edwards DR, et al. DESNT: A Poor Prognosis Category of Human Prostate Cancer. *European Urology Focus*. 2018;4(6):842-850. doi:10.1016/j.euf.2017.01.016

244. Carrivick L, Rogers S, Clark J, Campbell C, Girolami M, Cooper C. Identification of prognostic signatures in breast cancer microarray data using Bayesian techniques. *Journal of the Royal Society Interface*. 2006;3(8):367-381. doi:10.1098/rsif.2005.0093

245. Ma WF, Boudreau HE, Leto TL. Pan-cancer analysis shows tp53 mutations modulate the association of nox4 with genetic programs of cancer progression and clinical outcome. *Antioxidants*. 2021;10(2):1-17. doi:10.3390/antiox10020235

246. Tosello V, Ferrando AA. The NOTCH signaling pathway: Role in the pathogenesis of T-cell acute lymphoblastic leukemia and implication for therapy. 2013;4:199-210. doi:10.1177/2040620712471368

247. Fukusumi T, Califano JA. The NOTCH Pathway in Head and Neck Squamous Cell Carcinoma. 2018;97:645-653. doi:10.1177/0022034518760297

248. Aster JC, Pear WS, Blacklow SC. The Varied Roles of Notch in Cancer. 2017;12:245-275. doi:10.1146/annurev-pathol-052016-100127

249. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: A looking glass for cancer? 2012;12:323-334. doi:10.1038/nrc3261

250. Morris LGT, Riaz N, Desrichard A, et al. Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*. 2016;7(9):10051-10063. doi:10.18632/oncotarget.7067

251. Yang S, Michalski ST, Holle J, et al. Unexpected germline mutations in a pan-cancer analysis including sarcoma, renal, and other cancers. *Journal of Clinical Oncology*. 2017;35(15_suppl):1584-1584. doi:10.1200/jco.2017.35.15_suppl.1584

252. National Cancer Institute. Cancer Classification | SEER Training. 2020. https://training.seer.cancer.gov/disease/categories/classification.html https://training.seer.cancer.gov/disease/categories/classification.html%0Ahttps://training.seer.can Accessed May 25, 2022.

253. Pardo J, Murcia M, García F, Alvarado A. Gliosarcoma: A rare primary CNS tumor. Presentation of two cases. *Reports of Practical Oncology and Radiotherapy*. 2010;15(4):98-102. doi:10.1016/j.rpor.2010.05.003

254. Sun X, Zhang N, Yin C, Zhu B, Li X. Ultraviolet Radiation and Melanomagenesis: From Mechanism to Immunotherapy. 2020;10:951. doi:10.3389/fonc.2020.00951

255. Mayo Clinic. Melanoma - Symptoms and causes - Mayo Clinic. 2018. https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884.

256. Robinson S, Hossein A. Thymomas. *Asian Cardiovascular and Thoracic Annals*. 1996;4(4):206-207. doi:10.1177/021849239600400404

257. Latham A, Srinivasan P, Kemel Y, et al. Microsatellite instability is associated with the presence of Lynch syndrome pan-cancer. *Journal of Clinical Oncology*. 2019;37(4):286-295. doi:10.1200/JCO.18.00283

258. Petrucelli N, Daly MB, Pal T. *BRCA1- and BRCA2-Associated Hereditary Breast and Ovarian Cancer*. University of Washington, Seattle; 1993. https://www.ncbi.nlm.nih.gov/books/NBK1247/ http://www.ncbi.nlm.nih.gov/pubmed/20301425.

259. Jang E, Chung DC. Hereditary colon cancer: Lynch syndrome. 2010;4:151-160. doi:10.5009/gnl.2010.4.2.151

260. Sorrell AD, Espenschied CR, Culver JO, Weitzel JN. Tumor protein p53 (TP53) testing and li-fraumeni syndrome: Current status of clinical applications and future directions. 2013;17:31-47. doi:10.1007/s40291-013-0020-0

261. Plamper M, Gohlke B, Woelfle J. PTEN hamartoma tumor syndrome in childhood and adolescence—a comprehensive review and presentation of the German pediatric guideline. *Molecular and Cellular Pediatrics*. 2022;9(1). doi:10.1186/s40348-022-00135-1

262. Pilarski R. PTEN hamartoma tumor syndrome: A clinical overview. 2019;11:844. doi:10.3390/cancers11060844

263. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and Familial Colon Cancer. *Gastroenterology*. 2010;138(6):2044-2058. doi:10.1053/j.gastro.2010.01.054

264. Bailey MH, Tokheim C, Porta-Pardo E, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*. 2018;173(2):371. doi:10.1016/J.CELL.2018.02.060

265. GeneCards. CHGA Gene - GeneCards | CMGA Protein | CMGA Antibody. 2022. `https://www.genecards.org/cgi-bin/carddisp.pl?gene=CHGA`. Accessed August 7, 2022.

266. The Human Protein Atlas. Tissue expression of CHGA - Summary - The Human Protein Atlas. 2022. `https://www.proteinatlas.org/ENSG00000100604-CHGA/tissue`. Accessed August 7, 2022.

267. Guo Z, Wang Y, Xiang S, Wang S, Chan FL. Chromogranin A is a predictor of prognosis in patients with prostate cancer: A systematic review and meta-analysis. 2019;11:2747-2758. doi:10.2147/CMAR.S190678

268. Zhang X, Zhang H, Shen B, Sun XF. Chromogranin-a expression as a novel biomarker for early diagnosis of colon cancer patients. *International Journal of Molecular Sciences*. 2019;20(12). doi:10.3390/ijms20122919

269. GeneCards. CPLX2 Gene - GeneCards | CPLX2 Protein | CPLX2 Antibody. 2022. `https://www.genecards.org/cgi-bin/carddisp.pl?gene=CPLX2&keywords=CPLX2#aliases_descriptions`. Accessed August 7, 2022.

270. Komatsu H, Kakehashi A, Nishiyama N, et al. Complexin-2 (CPLX2) as a potential prognostic biomarker in human lung high grade neuroendocrine tumors. *Cancer Biomarkers*. 2013;13(3):171-180. doi:10.3233/CBM-130336

271. GeneCards. ITLN1 Gene - GeneCards | ITLN1 Protein | ITLN1 Antibody. 2022. `https://www.genecards.org/cgi-bin/carddisp.pl?gene=ITLN1&keywords=ITLN1`. Accessed August 7, 2022.

272. Paval DR, Di Virgilio TG, Skipworth RJE, Gallagher IJ. The Emerging Role of Intelectin-1 in Cancer. *Frontiers in Oncology*. 2022;12:105. doi:10.3389/FONC.2022.767859/BIBTEX

273. Speck O, Tang W, Morgan DR, et al. Three molecular subtypes of gastric adenocarcinoma have distinct histochemical features reflecting Epstein-Barr virus infection status and neuroendocrine differentiation. *Applied Immunohistochemistry and Molecular Morphology*. 2015;23(9):633-645. doi:10.1097/PAI.0000000000000122

274. Fujiwara T, Hiramatsu M, Isagawa T, et al. ASCL1-coexpression profiling but not single gene expression profiling defines lung adenocarcinomas of neuroendocrine nature with poor prognosis. *Lung Cancer*. 2012;75(1):119-125. doi:10.1016/j.lungcan.2011.05.028

275. Feng P, Li Z, Li Y, Zhang Y, Miao X. Characterization of Different Subtypes of Immune Cell Infiltration in Glioblastoma to Aid Immunotherapy. *Frontiers in Immunology*. 2022;13. doi:10.3389/fimmu.2022.799509

276. Liu J, Jiang C, Xu C, et al. Identification and development of a novel invasion-related gene signature for prognosis prediction in colon adenocarcinoma. *Cancer Cell International*. 2021;21(1):1-20. doi:10.1186/s12935-021-01795-1

277. Macintyre G, Goranova TE, De Silva D, et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*. 2018;50(9):1262-1270. doi:10.1038/s41588-018-0179-8

278. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics*. 2013;45(10):1127-1133. doi:10.1038/ng.2762

279. O'Hara AJ, Le Gallo M, Rudd ML, Bell DW. High-resolution copy number analysis of clear cell endometrial carcinoma. *Cancer Genetics*. 2020;240(301):5-14. doi:10.1016/j.cancergen.2019.10.005

280. Creative Diagnostics. PPAR Signaling Pathway. 2019:1-9. https://www.creative-diagnostics.com/ppar-signaling-pathway.htm.

281. Pio R, Corrales L, Lambris JD. The role of complement in tumor growth. In: *Advances in Experimental Medicine and Biology*. Vol 772. NIH Public Access; 2014:229-262. doi:10.1007/978-1-4614-5915-6_11

282. Chen YZ, Xue JY, Chen CM, et al. PPAR signaling pathway may be an important predictor of breast cancer response to neoadjuvant chemotherapy. *Cancer Chemotherapy and Pharmacology*. 2012;70(5):637-644. doi:10.1007/s00280-012-1949-0

283. Fanale D, Amodeo V, Caruso S. The Interplay between Metabolism, PPAR Signaling Pathway, and Cancer. 2017;2017. doi:10.1155/2017/1830626

284. Lin P, Huang Z. Correlation Analysis Connects Cancer Subtypes. *PLoS ONE*. 2013;8(7):e69747. doi:10.1371/journal.pone.0069747

285. Wang R, Wei J, Li Z, Tian Y, Du C. Bioinformatical analysis of gene expression signatures of different glioma subtypes. *Oncology Letters*. 2018;15(3):2807-2814. doi:10.3892/ol.2017.7660

286. Wei P, Tang H, Li D. Insights into Pancreatic Cancer Etiology from Pathway Analysis of Genome-Wide Association Study Data. *PLoS ONE*. 2012;7(10):e46887. doi:10.1371/journal.pone.0046887

287. Li H, Liu J wei, Liu S, Yuan Y, Sun L ping. Bioinformatics-Based Identification of Methylated-Differentially Expressed Genes and Related Pathways in Gastric Cancer. *Digestive Diseases and Sciences*. 2017;62(11):3029-3039. doi:10.1007/s10620-017-4740-6

288. Ji X, Bossé Y, Landi MT, et al. Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nature Communications*. 2018;9(1):1-15. doi:10.1038/s41467-018-05074-y

289. Kim TS, Da Silva E, Coit DG, Tang LH. Intratumoral Immune Response to Gastric Cancer Varies by Molecular and Histologic Subtype. *American Journal of Surgical Pathology*. 2019;43(6):851-860. doi:10.1097/PAS.0000000000001253

290. Zhang H, Chen Y. Identification of glioblastoma immune subtypes and immune landscape based on a large cohort. *Hereditas*. 2021;158(1):1-15. doi:10.1186/s41065-021-00193-x

291. Liu Q, Nie R, Li M, et al. Identification of subtypes correlated with tumor immunity and immunotherapy in cutaneous melanoma. *Computational and Structural Biotechnology Journal*. 2021;19:4472-4485. doi:10.1016/j.csbj.2021.08.005

292. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. 2002;30(1):207-210. doi:10.1093/nar/30.1.207

293. Cancer Research UK. Cancer mortality for common cancers | Cancer Research UK. 2022. `https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared#heading-Zero`. Accessed September 16, 2022.

294. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. *Journal of Statistical Software*. 2010;36(11):1-13. doi:10.18637/JSS.V036.I11

295. Chen Y, Verbeek FonsJ, Wolstencroft K. Establishing a consensus for the hallmarks of cancer based on gene ontology and pathway annotations. *BMC Bioinformatics*. 2021;22(1):178. doi:10.1186/s12859-021-04105-8

296. Netanely D, Avraham A, Ben-Baruch A, Evron E, Shamir R. Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. *Breast Cancer Research*. 2016;18(1):1-16. doi:10.1186/s13058-016-0724-2

297. Sørlie T, Perou CM, Tibshirani R, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*. 2001;98(19):10869-10874. doi:10.1073/PNAS.191367098/SUPPL_FILE/INDEX.HTML

298. Daemen A, Manning G. HER2 is not a cancer subtype but rather a pan-cancer event and is highly enriched in AR-driven breast tumors. *Breast Cancer Research.* 2018;20(1):1-16. doi:10.1186/s13058-018-0933-y

299. Dawson S-J, Rueda OM, Aparicio S, Caldas C. A new genome-driven integrated classification of breast cancer and its implications. *The EMBO Journal.* 2013;32(5):617-628. doi:10.1038/emboj.2013.19

300. Chen JW, Dhahbi J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Scientific Reports 2021 11:1.* 2021;11(1):1-15. doi:10.1038/s41598-021-92725-8

301. Faruki H, Mayhew GM, Serody JS, Hayes DN, Perou CM, Lai-Goldman M. Lung Adenocarcinoma and Squamous Cell Carcinoma Gene Expression Subtypes Demonstrate Significant Differences in Tumor Immune Landscape. *Journal of Thoracic Oncology.* 2017;12(6):943-953. doi:10.1016/j.jtho.2017.03.010

302. Rakha EA, Reis-Filho JS, Ellis IO. Basal-like breast cancer: a critical review. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* 2008;26(15):2568-2581. doi:10.1200/JCO.2007.13.1748

303. Wellberg E, Metz RP, Parker C, Porter WW. The bHLH/PAS transcription factor singleminded 2s promotes mammary gland lactogenic differentiation. *Development.* 2010;137(6):945-952. doi:10.1242/DEV.041657/-/DC1

304. Wang H, Chen Y, Han J, et al. DCAF4L2 promotes colorectal cancer invasion and metastasis via mediating degradation of NFκb negative regulator PPM1B. *American Journal of Translational Research.* 2016;8(2):405. /pmc/articles/PMC4846892/ /pmc/articles/PMC4846892/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4846892/.

305. Zhang X, Xu S, Hu C, et al. LncRNA ST8SIA6-AS1 promotes hepatocellular carcinoma progression by regulating MAGEA3 and DCAF4L2 expression. *Biochemical and Biophysical Research Communications.* 2020;533(4):1039-1047. doi:10.1016/j.bbrc.2020.09.115

306. Qin Y, Meng L, Fu Y, et al. SNORA74B gene silencing inhibits gallbladder cancer cells by inducing PHLPP and suppressing Akt/mTOR signaling. *Oncotarget.* 2017;8(12):19980-19996. doi:10.18632/ONCOTARGET.15301

307. Weigelt B, Bosma AJ, Van 't Veer LJ. Expression of a novel lacrimal gland gene lacritin in human breast tissues. *Journal of cancer research and clinical oncology.* 2003;129(12):735-736. doi:10.1007/S00432-003-0514-Y

308. Zhou C, Wang M, Zhou L, et al. Prognostic significance of PLIN1 expression in human breast cancer. *Oncotarget.* 2016;7(34):54488. doi:10.18632/ONCOTARGET.10239

309. Oleinik NV, Krupenko NI, Krupenko SA. Epigenetic Silencing of ALDH1L1, a Metabolic Regulator of Cellular Proliferation, in Cancers. *Genes & Cancer.* 2011;2(2):130. doi:10.1177/1947601911405841

310. Ren L, Yi J, Li W, et al. Apolipoproteins and cancer. *Cancer Medicine.* 2019;8(16):7032. doi:10.1002/CAM4.2587

311. World Health Organisation. Cancer Key Facts. 2022. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed September 13, 2022.

312. Hillman Cancer Centre. Cutaneous Melanoma | UPMC Hillman Cancer Center. 2019. https://hillman.upmc.com/cancer-care/melanoma-skin/types/cutaneous-melanoma. Accessed September 13, 2022.

313. Melanoma Research Alliance. Cutaneous Melanoma: What You Need to Know - Melanoma Research Alliance. 2020. https://www.curemelanoma.org/about-melanoma/types/cutaneous-melanoma. Accessed September 13, 2022.

314. Safiri S, Kolahi AA, Naghavi M, et al. Global, regional and national burden of bladder cancer and its attributable risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease study 2019. *BMJ Global Health.* 2021;6(11):e004128. doi:10.1136/BMJGH-2020-004128

315. American Cancer Society. Survival Rates for Bladder Cancer. 2022. https://www.cancer.org/cancer/bladder-cancer/detection-diagnosis-staging/survival-rates.html. Accessed September 13, 2022.

316. The Human Protein Atlas. FLG2 protein expression summary - The Human Protein Atlas. 2022. https://www.proteinatlas.org/ENSG00000143520-FLG2. Accessed September 30, 2022.

317. Bergboer JGM, Tjabringa GS, Kamsteeg M, et al. Psoriasis Risk Genes of the Late Cornified Envelope-3 Group Are Distinctly Expressed Compared with Genes of Other LCE Groups. *The American Journal of Pathology.* 2011;178(4):1470. doi:10.1016/J.AJPATH.2010.12.017

318. Han Y, Li X, Yan J, et al. Bioinformatic Analysis Identifies Potential Key Genes in the Pathogenesis of Melanoma. *Frontiers in Oncology.* 2020;10:581985. doi:10.3389/FONC.2020.581985

319. GeneCards. OTOR Gene - GeneCards | OTOR Protein | OTOR Antibody. 2022. https://www.genecards.org/cgi-bin/carddisp.pl?gene=OTOR. Accessed September 28, 2022.

320. Andreasen S, Varma S, Barasch N, et al. The HTN3-MSANTD3 Fusion Gene Defines a Subset of Acinic Cell Carcinoma of the Salivary Gland. *The American journal of surgical pathology.* 2019;43(4):489. doi:10.1097/PAS.0000000000001200

321. Li Y, Qi J, Yang J. RTP4 is a novel prognosis-related hub gene in cutaneous melanoma. *Hereditas.* 2021;158(1):1-11. doi:10.1186/S41065-021-00183-Z/FIGURES/6

322. GeneCards. KRT71 Gene - GeneCards | K2C71 Protein | K2C71 Antibody. 2022. https://www.genecards.org/cgi-bin/carddisp.pl?gene=KRT71. Accessed September 28, 2022.

323. Raskin L, Fullen DR, Giordano TJ, et al. Transcriptome Profiling Identifies HMGA2 as a Biomarker of Melanoma Progression and Prognosis. *The Journal of investigative dermatology.* 2013;133(11):2585. doi:10.1038/JID.2013.197

324. GeneCards. LORICRIN Gene - GeneCards | LORI Protein | LORI Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=LORICRIN. Accessed September 28, 2022.

325. Tölle A, Suhail S, Jung M, Jung K, Stephan C. Fatty acid binding proteins (FABPs) in prostate, bladder and kidney cancer cell lines and the use of IL-FABP as survival predictor in patients with renal cell carcinoma. *BMC Cancer*. 2011;11(1):1-10. doi:10.1186/1471-2407-11-302/FIGURES/5

326. Boiteux G, Lascombe I, Roche E, et al. A-FABP, a candidate progression marker of human transitional cell carcinoma of the bladder, is differentially regulated by PPAR in urothelial cancer cells. *International journal of cancer*. 2009;124(8):1820-1828. doi:10.1002/IJC.24112

327. Song Y, Jin D, Ou N, et al. Gene Expression Profiles Identified Novel Urine Biomarkers for Diagnosis and Prognosis of High-Grade Bladder Urothelial Carcinoma. *Frontiers in Oncology*. 2020;10:394. doi:10.3389/FONC.2020.00394/FULL

328. Bondaruk J, Jaksik R, Wang Z, et al. The Origin of Bladder Cancer from Mucosal Field Effects. *bioRxiv*. May 2021:2021.05.12.443785. doi:10.1101/2021.05.12.443785

329. Białas P, Śliwa A, Szczerba A, Jankowska A. The Study of the Expression of CGB1 and CGB2 in Human Cancer Tissues. *Genes*. 2020;11(9):1-11. doi:10.3390/GENES11091082

330. Lim EC, Lim SW, Tan KJ, et al. In-Silico Analysis of Deleterious SNPs of FGF4 Gene and Their Impacts on Protein Structure, Function and Bladder Cancer Prognosis. *Life*. 2022;12(7). doi:10.3390/LIFE12071018

331. Integrative OncoGenomics. IntOGen - Cancer Mutations Browser. 2022. `https://www.intogen.org/search`. Accessed September 29, 2022.

332. ICGC Data portal. Advanced Search | ICGC Data Portal. 2022. `https://dcc.icgc.org/search/m?filters=%7B%22mutation%22:%7B%22study%22:%7B%22is%22:%5B%22PCAWG%22%5D%7D%7D,%22donor%22:%7B%22studies%22:%7B%22is%22:%5B%22PCAWG%22%5D%7D%7D%7D&donors=%7B%22from%22:1%7D&mutations=%7B%22from%22:1%7D`. Accessed September 29, 2022.

333. CBioPortal. cBioPortal for Cancer Genomics. 2022. `https://www.cbioportal.org/`. Accessed September 29, 2022.

334. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics 2013 45:10*. 2013;45(10):1113-1120. doi:10.1038/ng.2764

335. Tamborero D, Gonzalez-Perez A, Perez-Llamas C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Scientific Reports 2013 3:1*. 2013;3(1):1-10. doi:10.1038/srep02650

336. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nature Methods 2013 10:11*. 2013;10(11):1108-1115. doi:10.1038/nmeth.2651

337. Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. *Nature Genetics 2013 45:10*. 2013;45(10):1127-1133. doi:10.1038/ng.2762

338. Li J, Lu Y, Akbani R, et al. TCPA: a resource for cancer functional proteomics data. *Nature Methods 2013 10:11*. 2013;10(11):1046-1047. doi:10.1038/nmeth.2650

339. Agyeman AA, Ofori-Asenso R. Perspective: Does personalized medicine hold the future for medicine? *Journal of Pharmacy & Bioallied Sciences*. 2015;7(3):239. doi:10.4103/0975-7406.160040

340. Genomics England. 100,000 Genomes Project | Genomics England. 2022. `https://www.genomicsengland.co.uk/initiatives/100000-genomes-project`. Accessed September 30, 2022.

341. Bansal V, Boucher C. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? *iScience*. 2019;18:37. doi:10.1016/J.ISCI.2019.06.035

342. Goetz LH, Schork NJ. Personalized Medicine: Motivation, Challenges and Progress. *Fertility and sterility*. 2018;109(6):952. doi:10.1016/J.FERTNSTERT.2018.05.006

343. Davis PB, Yasothan U, Kirkpatrick P. Ivacaftor. *Nature reviews Drug discovery*. 2012;11(5):349-350. doi:10.1038/NRD3723

344. Liao X, Lochhead P, Nishihara R, et al. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. *The New England journal of medicine*. 2012;367(17):1596-1606. doi:10.1056/NEJMOA1207756

345. Li X, You R, Wang X, et al. Effectiveness of Prophylactic Surgeries in BRCA1 or BRCA2 Mutation Carriers: A Meta-analysis and Systematic Review. *Clinical cancer research : an official journal of the American Association for Cancer Research*. 2016;22(15):3971-3981. doi:10.1158/1078-0432.CCR-15-1465

346. Baselga J. Clinical trials of Herceptin® (trastuzumab). *European Journal of Cancer*. 2001;37(SUPPL. 1):18-24. doi:10.1016/s0959-8049(00)00404-4

347. University of East Anglia. Prostate Cancer Research - Giving To UEA - About. 2020. `https://www.uea.ac.uk/about/giving-to-uea/our-causes/prostate-cancer-research`. Accessed September 30, 2022.

348. Hutter C, Zenklusen JC. The Cancer Genome Atlas: Creating Lasting Value beyond Its Data. *Cell*. 2018;173(2):283-285. doi:10.1016/J.CELL.2018.03.042

349. Brueffer C, Vallon-Christersson J, Grabau D, et al. Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative. *JCO precision oncology*. 2018;2(2):1-18. doi:10.1200/PO.17.00135

350. Pan Prostate Cancer Group. Pan Prostate Cancer Group - Home. `https://panprostate.org/about.html`. Accessed September 30, 2022.

351. Lin HH, Wei NC, Chou TY, et al. Building personalized treatment plans for early-stage colorectal cancer patients. *Oncotarget*. 2017;8(8):13805-13817. doi:10.18632/ONCOTARGET.14638

352. Lim SB, Tan SJ, Lim W-T, et al. A merged lung cancer transcriptome dataset for clinical predictive modeling. *Scientific Data 2018 5:1*. 2018;5(1):1-8. doi:10.1038/sdata.2018.136

353. Satpathy S, Krug K, Jean Beltran PM, et al. A proteogenomic portrait of lung squamous cell carcinoma. *Cell*. 2021;184(16):4348-4371.e40. doi:10.1016/J.CELL.2021.07.016

354. Das PM, Singal R. DNA methylation and cancer. *Journal of Clinical Oncology*. 2004;22(22):4632-4642. doi:10.1200/JCO.2004.07.151