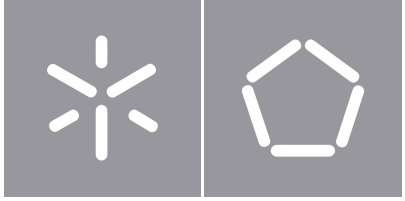
University of Minho
School of Engineering

Catarina Gomes Ferreira

MitoProfiles: Cancer mitochondrial profiles in high metabolic rate organs

october 2023



University of Minho
School of Engineering

Catarina Gomes Ferreira

MitoProfiles: Cancer mitochondrial profiles in high metabolic rate organs

Masters Dissertation
Master's in Bioinformatics

Dissertation supervised by
Professor Miguel Francisco Almeida Pereira Rocha, PhD
Guilhermina Isabel dos Santos Duarte, PhD

october 2023

Copyright and Terms of Use for Third Party Work

License granted to users of this work:

Acknowledgements

Writing a dissertation was a transformative process that enabled my personal and professional growth in this field. It helped me increase my sense of critical thinking, leading me to question the ”how” and ”why” before making important decisions. It also sensitized me to details that would have consumed a significant amount of time for correction. Therefore, I extend my sincerest gratitude to the following individuals:

First and foremost, I express my deepest appreciation to my incredible supervisor, Dr. Isabel Duarte. Throughout the journey of writing this dissertation, she provided unwavering guidance and continually challenged me to expand my knowledge. Her patience and adept problem-solving skills were invaluable, aiding me in overcoming various obstacles. Words cannot adequately convey the extent of my gratitude for her role in shaping a better version of myself.

I also wish to thank the esteemed teacher Miguel Rocha, whose expertise in the fields of Bioinformatics and Systems Biology shed light on alternative methods for analyzing results, enriching the quality of my work.

I am grateful to Ramiro Magno for his valuable suggestions that significantly enhanced the efficiency of my efforts.

My heartfelt thanks go to my cherished partner, who provided unwavering support throughout the dissertation-writing process and consistently encouraged me to strive for excellence.

I am deeply appreciative of the unflagging support from my friends and family, who stood by me and continued to believe in my capabilities.

Last but not least, the computational resources used for this thesis work were provided by the FCT Advanced Computing Project number 2022.46814.CPCA.A0. This support was instrumental in the successful completion of this research.

Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, october 2023

Catarina Gomes Ferreira

Abstract

Metabolic reprogramming is recognized as a critical hallmark of cancer, influencing cancer initiation and progression. Emerging evidence suggests that the metabolism of non-cancer cells within the tumor microenvironment plays a pivotal role in modulating tumor development, underscoring the importance of metabolic variables for better understanding cancer.

The main goal of this study is to identify genes exhibiting differential expression in cancer, with a specific emphasis on distinguishing between organs with high metabolic rates (brain, liver, and kidneys) and organs with low metabolic rates (bladder, colon, and skin), particularly focusing on genes encoding mitochondrial proteins.

For this, we used two databases containing RNA-seq samples from normal and cancer tissues, obtained from the Genotype-Tissue Expression (**GTEx**) and The Cancer Genome Atlas (**TCGA**) projects, respectively. General Linear Models (**GLMs**) were applied for differential expression analysis, and hierarchical clustering e soft fuzzy clustering to identify distinct gene expression profiles.

Our research showed that many of the differentially expressed mitochondrial genes, such as **ACSM1** and **ACSM5**, and **PRODH**, represent potential adaptations of cancer cells to metabolic and microenvironmental stress. Additionally, **FDX2**, a crucial player in iron-sulfur protein biogenesis, and **ACSM2B**, responsible for catalyzing the activation of free fatty acids (**FFAs**) to CoA, showed substantial expression differences, highlighting the importance of these two pathways for the oncogenic process. The most substantial genetic expression differences were observed between normal and cancer tissues, rather than between high and low metabolic rate organs, suggesting that the signal from the metabolic rate could be masked by the pronounced changes that cancer induces in cells.

Despite the unequal sample sizes and the usage of two different data sources, our findings provide valuable insights into the complex interplay between metabolism and gene expression in cancer.

Keywords Cancer, Metabolic rate, Mitochondrial proteins, Differential gene expression, Clustering

Resumo

A reprogramação metabólica é reconhecida como um *hallmark* do cancro, influenciando a sua iniciação e progressão. Estudos recentes mostram que o metabolismo das células não cancerígenas desempenha um papel crucial no microambiente tumoral e na modulação do seu desenvolvimento; demonstrando a importância do metabolismo neste processo. Neste estudo identificaram-se genes mitocondriais que exibem expressão diferencial em cancro, com particular ênfase na distinção entre órgãos com elevada taxa metabólicas (cérebro, fígado e rins) e órgãos com baixa taxa metabólica (bexiga, cólon e pele).

Para tal, foram utilizados dados de RNA-seq provenientes de duas bases de dados: Genotype-Tissue Expression (**GTEx**) e The Cancer Genome Atlas (**TCGA**), contendo amostras de tecidos normais e cancerígenos, respetivamente. Os genes diferencialmente expressos foram obtidos através de uma análise de expressão diferencial usando General Linear Models (**GLMs**), e os perfis de expressão foram obtidos por hierarchical clustering e soft fuzzy clustering.

Os resultados demonstraram que muitos dos genes mitocondriais diferencialmente expressos, tais como **ACSM1** e **ACSM5**, e **PRODH**, poderão representar potenciais adaptações das células cancerígenas ao stress metabólico e microambiental. Adicionalmente, a **FDX2**, uma proteína crucial para a biogénese de proteínas ferro-enxofre, e a **ACSM2B**, responsável pela ativação de ácidos gordos livres (**FFAs**) transformando-os em CoA, mostraram diferenças significativas de expressão, demonstrando a importância destes dois processos na carcinogénese. As diferenças de expressão entre tecidos normais e cancerígenos mostraram ser mais acentuadas do que entre órgãos com taxas metabólicas alternativas, sugerindo que a magnitude do sinal gerado pelas diferenças moleculares produzidas pelo tipo de taxa metabólica poderá não ser suficiente para se sobrepor à magnitude do sinal provocado pelo cancro.

Apesar do tamanho diferente das amostras, e da utilização de duas bases de dados diferentes, estes resultados contribuem para elucidar a complexa relação entre metabolismo e expressão genética em cancro.

Palavras-chave Cancro, Taxa metabólica, Proteínas mitocondriais, Expressão genética diferencial, Clustering

Contents

List of Figures

List of Tables

Acronyms

4E-BPs Eukaryotic translation initiation factor 4E (eIF4E)-binding proteins.

ACO1 Aconitase 1.

ACOT2 Acyl-CoA thioesterase 2.

ACSM1 Acyl-Coenzyme A Synthetase 1.

ACSM2B Acyl-CoA Synthetase Medium Chain Family Member 2B.

ACSM5 Acyl-Coenzyme A Synthetase 5.

ADGRF2 Adhesion G protein-coupled receptor F2.

AIFM3 Apoptosis inducing factor mitochondria associated 3.

AKR7A2 Aldo-Keto reductase family 7 member A2.

ATP Adenosine triphosphate.

B-raf v-raf murine sarcoma viral oncogene homolog B1.

Bak BCL-2-antagonist/killer.

Bax Bcl-2 Associated X-protein.

Bcl-2 B-cell lymphoma 2.

Bcl-xL B-cell lymphoma-extra large.

BCL2 BCL2 apoptosis regulator.

BCL2L10 BCL2 like 10.

Bim Bcl-2 Interacting Mediator of cell death.

BLOC1S1 Biogenesis of lysosomal organelles complex 1 subunit 1.

BNIP3 BCL2/adenovirus E1B 19 kDa protein-interacting protein 3.

BNIP3L/NIX BCL2/adenovirus E1B 19 kDa protein-interacting protein 3-like.

c-Myc Cellular Myelocytomatosis.

CHN2-AS1 CHN2 antisense RNA 1.

CIA Cytosolic iron-sulfur assembly machinery.

CLT Central limit theorem.

CMPK2 Cytidine/Uridine monophosphate kinase 2.

CPM Counts per Million.

CTCFL CCCTC-binding factor like.

DEGs Differentially Expressed Genes.

DNA Deoxyribonucleic Acid.

Drp1 Dynamin-related protein-1.

EM Expectation-Maximization.

EMT Epithelial-to-mesenchymal Transition.

ERK Extracellular signal-regulated kinase.

ETC Electron Transport Chain.

FC Fold Change.

FCD Fold Change Detection.

FDR False Discovery Rate.

FDX2 Ferrodoxin 2.

FDXR Ferredoxin reductase.

FFAs Free Fatty Acids.

GLMs General Linear Models.

GLOBOCAN Global Cancer Observatory.

GLRX5 Glutaredoxin 5.

GLUT1 Glucose Transporter Protein Type 1.

GMM Gaussian Mixture Models.

GPAT2 Glycerol-3-phosphate acyltransferase 2.

GTEx Genotype-Tissue Expression.

HDHD3 Haloacid dehalogenase like hydrolase domain containing 3.

HDHD5 Haloacid dehalogenase like hydrolase domain containing 5.

HDI Human Development Index.

HIF-1a Hypoxia-inducible factor 1-alpha.

IDE ntegrated development environment.

IFI27 Interferon alpha inducible protein 27.

Igf1/2 Insulin-like growth factor 1 and 2.

IMM Inner Mitochondrial Membrane.

ISC Iron-sulfur cluster machinery.

ISCU Iron-Sulfur Cluster Assembly Enzyme.

K-Ras Kirsten rat sarcoma virus.

KRT26 Keratin 26.

KRT85 Keratin 85.

LHFPL3-AS1 LHFPL3 Antisense RNA 1.

LINC00462 Long intergenic non-protein coding RNA 462.

LINC01833 Long intergenic non-protein coding RNA 1833.

LINC02247 Long intergenic non-protein coding RNA 2247.

logCPM Log Counts per Million.

logFC Log2 Fold Change.

LR Likelihood Ratio.

MCUR1 Mitochondrial calcium uniporter regulator 1.

MEK Mitogen-activated protein kinase kinase.

Mfn1 Mitofusin-1.

Mfn2 Mitofusin-2.

MOMP mitochondrial outer membrane permeabilization.

MPST Mercaptopyruvate sulfurtransferase.

MRPL24 Mitochondrial ribosomal protein L24.

MRPL43 Mitochondrial ribosomal protein L43.

MT-ATP6 Mitochondrially encoded ATP synthase membrane subunit 6.

MT-ATP8 Mitochondrially encoded ATP synthase membrane subunit 8.

MT-CO1 Mitochondrially encoded cytochrome c oxidase I.

MT-CO2 Mitochondrially encoded cytochrome c oxidase II.

MT-CO3 Mitochondrially encoded cytochrome c oxidase III.

MT-CYB Mitochondrially encoded cytochrome b.

MT-ND1 Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 1.

MT-ND2 Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 2.

MT-ND3 Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 3.

MT-ND4 Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 4.

MT-ND4L Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 4L.

MT-ND5 Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 5.

MT-ND6 Mitochondrially encoded NADH:ubiquinone oxidoreductase core subunit 6.

MT-TM Mitochondrially encoded tRNA-Met (AUA/G).

MT-TV Mitochondrially encoded tRNA-Val (GUN).

mtDNA Mitochondrial DNA.

mTOR Mammalian Target of Rapamycin.

MUC21 Mucin 21.

NADH Nicotinamide adenine dinucleotide.

NADPH Nicotinamide Adenine Dinucleotide Phosphate.

NGS Next-Generation Sequencing.

NRF2 nuclear factor (erythroid-derived 2)-like 2.

OMM Outer Mitochondria Membrane.

Opa1 Optic atrophy 1.

PAM Partitioning Around Medoids.

PCA Principal Component Analysis.

PGC-1a Proliferator-activated Receptor Gamma Coactivator-1 alpha.

PINK1 Phosphatase and Tensin Homolog deleted Induced Kinase 1.

PRODH Proline Dehydrogenase 1.

Puma p53 upregulated modulator of apoptosis.

RB Retinoblastoma Protein.

REE Resting Energy Expenditure.

RNA Ribonucleic Acid.

ROS Reactive Oxygen Species.

RPUSD3 RNA pseudouridine synthase D3.

SLC25A44 Solute carrier family 25 member 44.

SNPs Single Nucleotide Polymorphisms.

SOD2 superoxide dismutase 2.

Src Steroid Receptor Coactivator.

SVD Singular Value Decomposition.

TCA Tricarboxylic Acid.

TCGA The Cancer Genome Atlas.

TMM Trimmed Mean of M values.

TP53 Tumor Protein 53.

TTC21B-AS1 TTC21B antisense RNA 1.

WFDC5 WAP four-disulfide core domain 5.

YY1 Yin Yang 1.

ZIC5 Zic family member 5.

Chapter 1

Introduction

1.1 Motivation

Cancer is still one of the deadliest diseases in the world today. While there have been great advancements in treatments, the mortality rate for many types of cancer remains high. To develop therapies to treat this disease it is important to study its unique characteristics, which have been explored and described in detail in the last decades [Hanahan and Weinberg, 2000]; [Hanahan and Weinberg, 2011]; [Hanahan, 2022].

In recent years, a greater understanding of the role of metabolic reprogramming in cancer development has been gained. It has become clear that mitochondria play a central role in this process, as they are responsible for energy metabolism, and are involved in other cellular functions relevant to cancer progression, such as apoptosis, reactive oxygen species management, and cellular-stress signaling [Vyas et al., 2016]. Therefore, mitochondrial dysfunction and abnormal metabolism can lead to uncontrolled cell growth and proliferation.

However, we still have a very limited understanding of the different mitochondrial metabolic states present in cancers from different organs. Since each healthy organ has a different baseline metabolic-rate [Elia, 1992], by comparing the expression profiles of mitochondrial proteins across cancers from different organs (high versus low metabolic-rate organs), we can gain insights into the mitochondrial metabolic state in those cancers.

Overall, this analysis can potentially provide a better understanding of the molecular mechanisms driving differences in cancer metabolism, energy production, and organ dysfunction in cancer.

1.2 Main goal of this work

Knowing that high metabolism organs such as the liver, brain, or kidneys provide a tumor microenvironment that contains mitochondria already primed for the production of high amounts of energy, the main goal of this work is to find mitochondrial expression profiles (mitoprofiles) using data obtained from both normal tissues and cancer tissues and compare them to verify whether there are unique patterns of expression characteristic of organs with high metabolic-rate.

1.3 Thesis outline

Chapter II, 'State of the art', provides an overview of the subject areas addressed in this work. It begins with cancer and mitochondrion overview, with a brief description of the interactions between them and the importance that this can have for future discoveries about the disease. The chapter ends with a brief explanation of what kind of methodologies will be used to achieve the goal.

Chapter III, 'Materials and methods', contains a detailed description of the workflow for this research. The chapter starts with the collection of the transcriptomic data from the **TCGA** and the **GTEx** databases and their preparation for the statistical analyses. Subsequently, a differential gene expression analysis was performed between the normal and cancer samples for each organ, followed by a clustering analysis to build profiles for the differentially expressed genes.

Chapter IV, 'Results and Discussion', covers the findings from this study and discusses the major results.

Chapter V, 'Conclusions and Future work', summarizes the work highlighting what was found in this thesis as well as providing perspectives on future work to better understand the role of the mitochondria in cancer.

Chapter 2

State of the art

2.1 Cancer

Cancer is a complex disease characterized by genetic and epigenetic changes that result in the uncontrolled growth and division of cells with the potential to spread to other parts of the body. These changes, such as mutations in **DNA**, changes in **DNA** methylation or histone modification patterns, and alterations in non-coding **DNA** expression, can occur in oncogenes (genes that promote cell division) or tumor suppressor genes (genes that regulate cell division and prevent uncontrolled growth), as weel as **DNA** repair mechanisms, or cellular signaling pathways [Holland, 1996], impacting cell type regulation, cell division, apoptosis, angiogenesis, and metabolism.

Cancer development is influenced by a complex interplay of genetic and environmental factors. A few cancers are inherited and run in families, while others are caused by mutations that occur during a person's lifetime. Environmental factors, such as exposure to radiation, chemicals, or certain viruses, can also increase the risk of developing cancer [Anand et al., 2008].

The sequence of genetic and cellular events underlying cancer onset and development typically involves several key steps (Figure 1). Firstly there is the **initiation**: in this step, normal cells undergo genetic changes that result in the acquisition of malignant traits. These changes can be caused by mutations in **DNA**, changes in **DNA** methylation patterns, or alterations in the expression of non-coding **RNA**s; then the **promotion** step occurs when the cells start to divide and grow uncontrollably, leading to the formation of a tumor mass; followed by the **progression**, where cells acquire additional genetic changes that allow them to invade nearby tissues and spread to other parts of the body. The final step is the **metastasis**: a process in which cancer cells spread from the primary tumor to other parts of the body, forming secondary tumors [Hamidi and Ivaska, 2018].

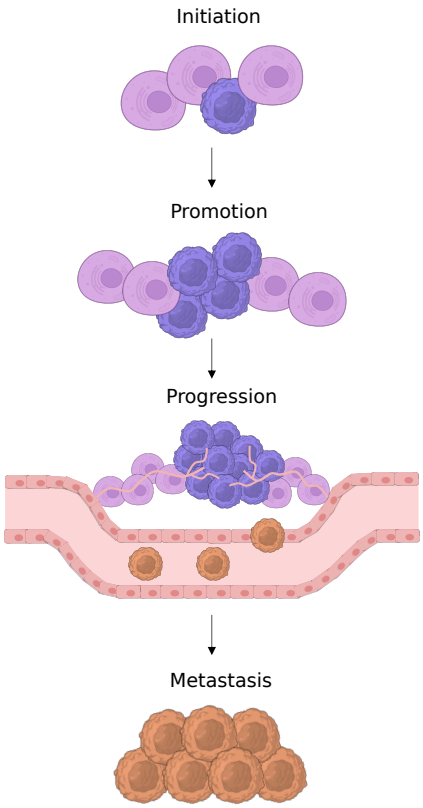


Figure 1: Four key steps in cancer development. Adapted from: [Vyas et al., 2016]

2.1.1 Prevalence and mortality of cancer worldwide

Looking at the most recent statistics made by **GLOBOCAN** in 2020 we can observe cancer incidence and mortality, with a focus on geographic variability across 20 world regions. They estimated there were 19.3 million new cancer cases and 9.96 million cancer deaths in that same year worldwide.

Looking at both genders combined, breast cancer is the most commonly diagnosed cancer (11.7% of the total cases) and lung cancer the leading cause of cancer death (18% of the total cancer deaths). The next most incident cancers are lung cancer (11.4%), colorectal cancer (10%) and prostate cancer (7.3%), and the most lethal are colorectal cancer (9.4%), liver cancer (8.3%) and stomach cancer (7.7%), (Figure 2 A). However, if we look at the statistics for each gender, we can see that for males lung cancer is the most frequent and is quickly followed by prostate and colorectal cancer. The most deadly cancer in males is lung, followed by liver and colorectal cancer (Figure 2 B). Among females, breast cancer is the most commonly diagnosed cancer, followed by colorectal and lung cancer, these being also the leading mortality cancers in females (Figure 2 C).

Cancer incidence and mortality are rapidly growing worldwide. The reasons can be complex but in general, they reflect both aging and growth of the population, as well as changes in the prevalence and

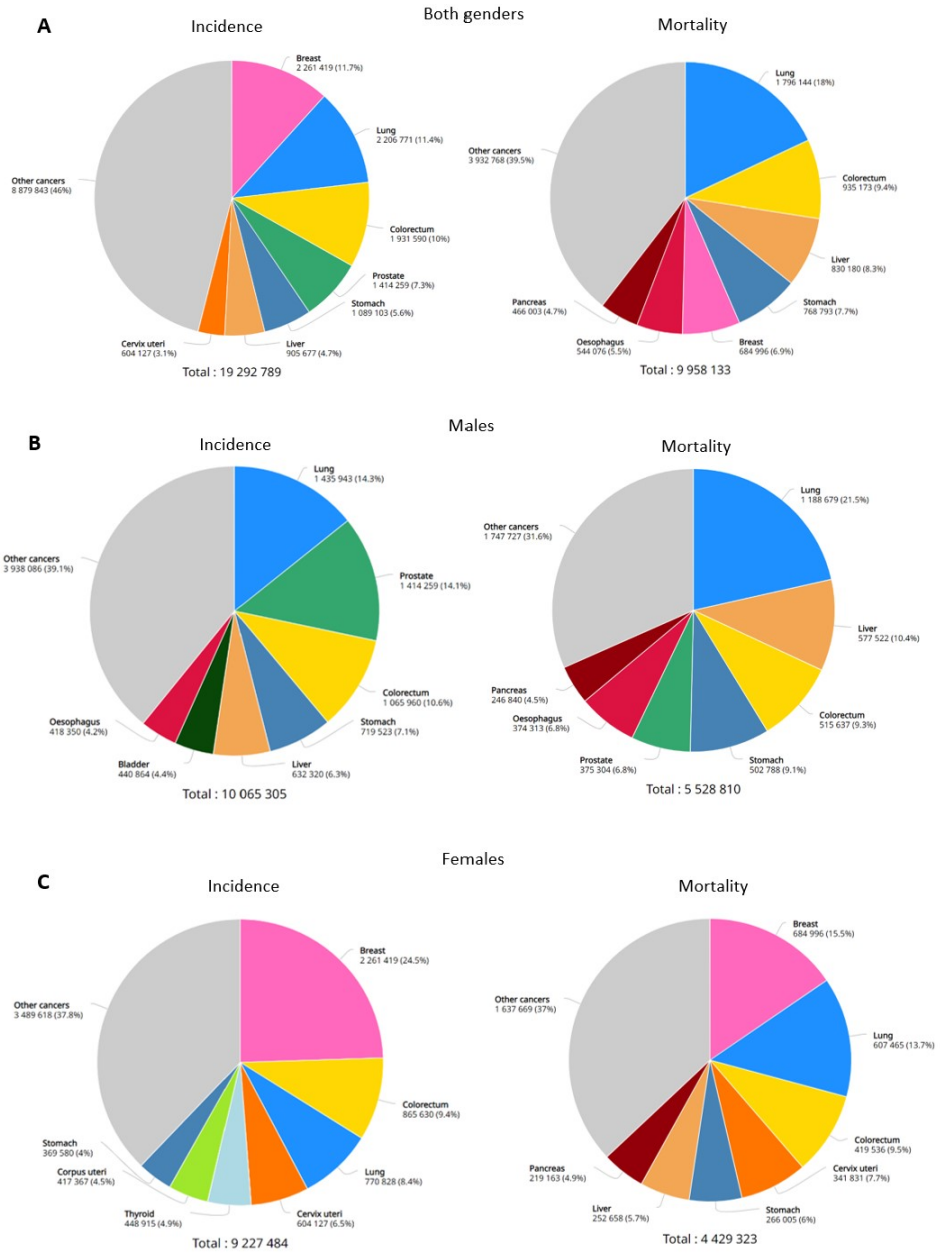


Figure 2: Cancer incidence and mortality for the 10 Most Common Cancers in 2020 for (A) Both Genders, (B) Males, and (C) Females. For each gender, the area of the pie chart reflects the proportion of the total number of cases (incidence) or deaths. Retrieved from: **GLOBOCAN** 2020.

distribution of the main risk factors for cancer, several being associated with socioeconomic development [Omran, 1998]; [Gersten and Wilmoth, 2002]. The influence of these facts can be seen by comparing the maps in Figures 3 and 4 to see the extent to which cancer's position as a cause of premature death reflects national social and economic development levels. As we can see the more developed countries (Figure 3 - shades of blue) have higher cancer related mortality rates (Figure 4 - darker red) than the developing countries, despite the disparity in disease and death official records.

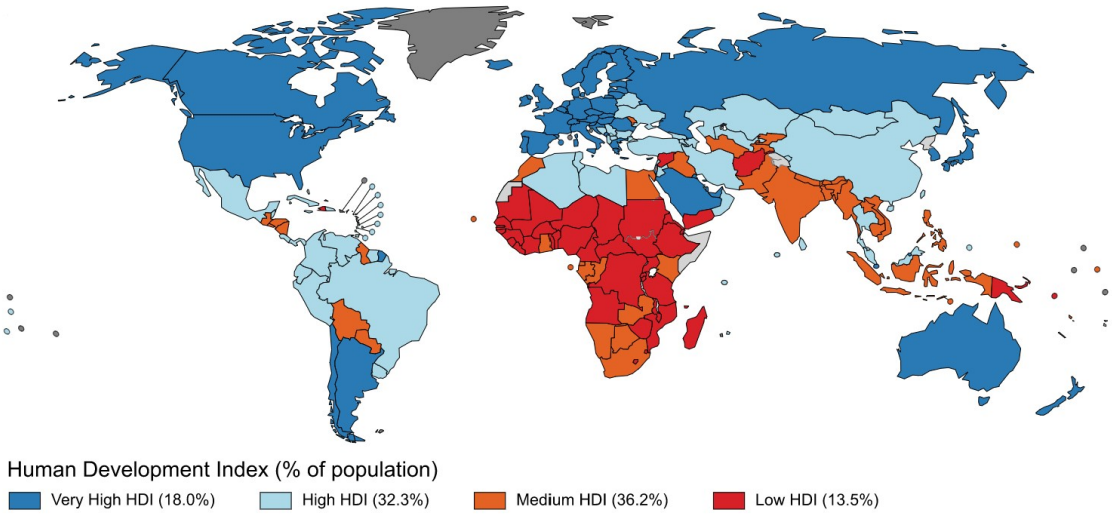


Figure 3: 4-Tier Human Development Index (**HDI**). Each country has been classified into 1 of 4 human development indices: Very high (dark blue), High (light blue), Medium (orange), and Low (red). Retrieved from: United Nations Procurement Division/United Nations Development Program.

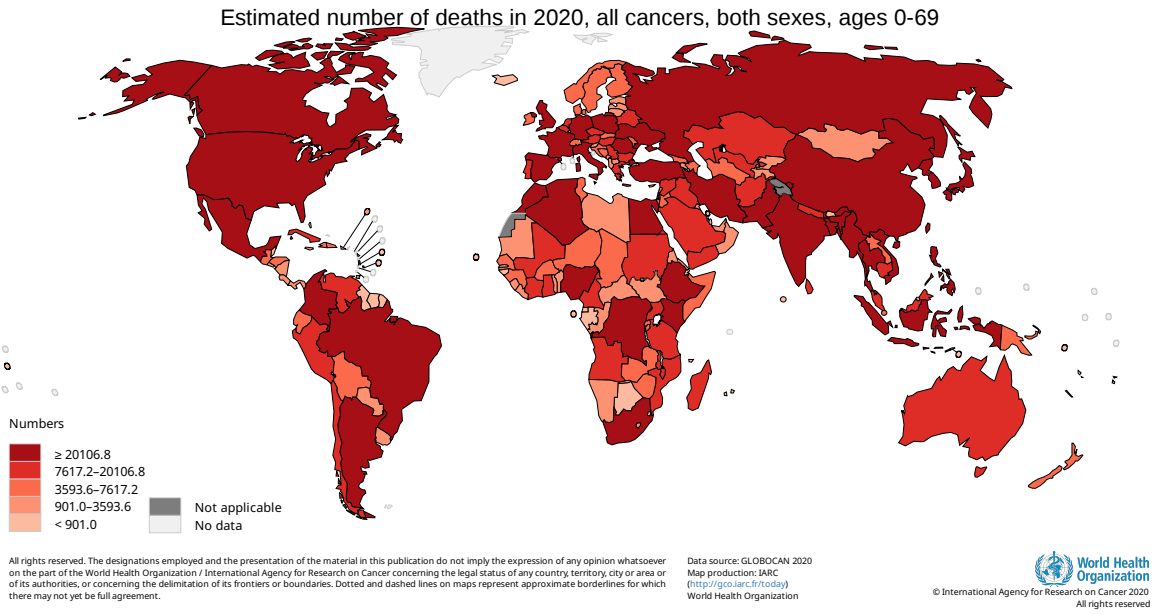


Figure 4: Estimated number of deaths in 2020 per country. Representation of the National Ranking of Cancer as a cause of death for all cancers, in both sexes, in ages below 70 years in 2020. The colors represent the incidence of the disease in different countries. Retrieved from: World Health Organization.

2.2 Understanding cancer development

To better understand cancer we must learn about its molecular and cellular characteristics. These are called hallmarks and they can be defined as distinctive and complementary capabilities that allow the tumor to grow and enable its metastatic dissemination. There are six hallmarks of cancer that help us understand the biology of cancer: sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis (Figure 5 - light blue text boxes) [Hanahan and Weinberg, 2000].

These hallmarks can be acquired in different tumor types via distinct mechanisms and at different times during the course of tumorigenesis. Their acquisition is possible via two enabling characteristics, being the most prominent the development of genomic instability in cancer cells. This generates random mutations including chromosomal rearrangements, and among these are the rare genetic changes that can lead to the initiation of the hallmark capabilities. A second enabling characteristic involves the inflammatory state of premalignant and malignant lesions driven by immune system cells, some of which serve to promote tumor progression (Figure 5 - light green text boxes) [Hanahan and Weinberg, 2011].

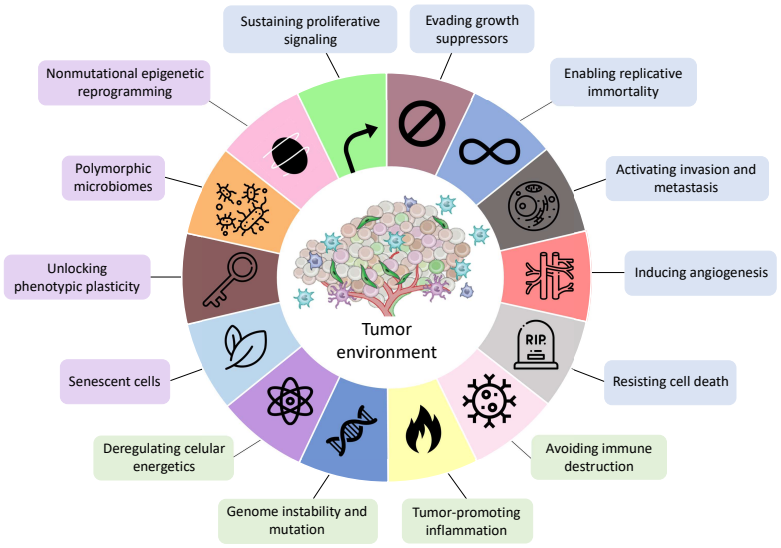


Figure 5: Current hallmarks of cancer. These hallmarks have emerged progressively as more research accumulated about cancer cells. The hallmarks with light blue text box represent the first hallmarks to be described. Those in light green represent the next group of hallmarks described. Finally, the ones with a light purple color show the hallmarks that represent the most recent discoveries about cancer cell biology. Adapted from: [Hanahan, 2022]

Hallmark 1: Sustaining proliferative signaling

This hallmark is one of the most important traits of cancer cells being related to the ability to sustain chronic proliferation.

Usually, normal tissues control with extreme caution the production and release of growth-promoting signals that instruct the cell to start its growth-and-division cycle, ensuring not only the homeostasis of cell number, but also the maintenance of normal tissue architecture and function. However, cancer cells, by disrupting these signals, become capable of deciding their own fate. These signals are effected by growth factors that bind to cell-surface receptors that typically contain intracellular tyrosine kinase domains. With this connection, the cell starts to send signals using signaling pathways that are responsible for the regulation of the cell cycle and its growth (increasing the cell size). This growth-promoting signals can also influence other biological properties of the cell, such as its energy metabolism and ability to survive [Hanahan and Weinberg, 2000]; [Lemmon and Schlessinger, 2010]; [Witsch et al., 2010]. As such, proliferative signaling is considered an important hallmark due to its contribution for increasing the cell's energy leading to the acquisition of more capabilities.

Hallmark 2: Evading growth suppressors

Cancer cells must circumvent the mechanisms that healthy cells have to negatively regulate their proliferation. These programs are dependent on the actions of dozens of tumor suppressor genes. The two prototypical tumor suppressors encode the retinoblastoma protein **(RB)**, which decides if the cell should or not proceed with its growth-and-division cycle [Burkhart and Sage, 2008], and tumor protein 53 **(TP53)**, which regulates cell division by keeping cells from proliferating too fast or in an uncontrolled way. Both of them operate as central control nodes within two key complementary cellular regulatory circuits where the decisions are made about the proliferation of a cell, or if something goes wrong and normal proliferation is not possible, they proceed with the activation of senescence and apoptotic programs [Ghebranious and Donehower, 1998]. With this, cancer cells are able to continue their growth without being stopped by normal mechanisms that would lead to their death.

Hallmark 3: Resisting cell death

The concept that cells are programmed to die by apoptosis serves as a natural barrier to cancer development [Adams and Cory, 2007].

The apoptotic machinery is composed of both upstream regulators and downstream effector components [Adams and Cory, 2007]. The regulators are divided into two major circuits: one receives and processes extracellular death-inducing signals, and the other senses and integrates a variety of signals of intracellular origin. Each of these circuits culminates in the activation of a protease that is normally inactivated (caspases 8 and 9), initiating a cascade of proteolysis involving effector caspases responsible for the apoptosis, when the cell is progressively disassembled and then consumed, either by its neighbors, by phagocytic cells or both.

Despite all of these mechanisms, there are several abnormality sensors that were identified and play key roles in tumor development [Adams and Cory, 2007]. The most common is the loss of **TP53** tumor suppressor function, which eliminates this critical damage sensor from the apoptosis-inducing course. Alternatively, a similar result can be achieved by increasing the expression of antiapoptotic regulators **(Bcl-2, Bcl-xL)** or of survival signals **(Igf1/2)**, by downregulating proapoptotic factors **(Bax, Bim, Puma)** [Bose et al., 2015]. The multiple ways of apoptosis-avoiding mechanisms presumably reflect the diversity of apoptosis-inducing signals that cancer cell populations encounter during their evolution to the malignant state. Therefore, this characteristic is crucial for the cancer cells to circumvent pathways that would lead to its destruction.

Hallmark 4: Enabling replicative immortality

The telomeres, which are composed of multiple tandem hexanucleotide repeats, shorten progressively in non-immortalized cells and eventually lose their ability to protect the ends of chromosomal **DNA** from end-to-end fusions. Such fusions generate unstable dicentric chromosomes that threaten cell viability. Accordingly, the length of telomeric **DNA** in a cell indicates how many generations a cell has until its telomeres are largely eroded and lose their protective functions, triggering entrance into crisis [Blasco, 2005].

Telomerase is the specialized **DNA** polymerase that adds telomere repeat segments to the ends of telomeric **DNA**. The presence of this polymerase is very rare in normal cells, but can be found expressed at significant levels in the vast majority (practically 90%) of spontaneously immortalized cells, including human cancer cells [Hanahan and Weinberg, 2011]. With the extension of the telomeric **DNA**, telomerase is able to counter the progressive telomere erosion that would otherwise occur in its absence. This phenomenon gives these cells some resistance to the induction of both senescence and apoptosis, effectively immortalizing them. As such, this ability allows the cell to always replicate since the size of the telomeres will never decrease and so it will never reach the critical point that would activate apoptosis.

Hallmark 5: Inducing angiogenesis

Just like normal tissues, tumors require sustenance such as nutrients and oxygen to survive, as well as the ability to remove metabolic waste and carbon dioxide. The tumor-associated neovasculature, created by the process of angiogenesis, addresses these needs.

Angiogenesis is the development of vasculature and involves the generation of new endothelial cells as well as their assembly into tubes (vasculogenesis) in addition to their sprouting (angiogenesis) from already existing blood vessels. This phenomenon is induced very early during the development of invasive cancers in humans [Raica et al., 2009]. Historically, angiogenesis was seen as an important step only when rapidly growing macroscopic tumors had formed. However, more recent data is starting to suggest that angiogenesis also contributes to the microscopic premalignant phase of cancer progression, further impling its status as an integral hallmark of cancer. This allow cancer cells to obtain enough nutrients to sustain themselves, making this ability crucial for their survival [Hanahan and Weinberg, 2011].

Hallmark 6: Activating invasion and metastasis

The invasion and metastasis is a multistep process that has been schematized as a sequence of discrete steps, often called the invasion-metastasis cascade [Talmadge and Fidler, 2010]. This depicts a succession

of cell-biologic changes that start with a local invasion. After that, cancer cells enter nearby blood and lymph vessels which transport them through the lymphatic system, which results in the escape of cancer cells from the lumina into the parenchyma of distant tissues (extravasation). Finally, when they escape the vessels they create small clusters of cancer cells (micrometastases), and start to develop micrometastatic lesions into macroscopic tumors (colonization). This characteristic is important for cancer cells because it allows them to spread to other areas to harness new nutrients continuing their growth, and hindering their treatment. Therefore this ability represents the last great frontier for exploratory cancer research, and imposes a greater challenge, that if overcame, will give new insights about tissue invasiveness and metastasis for the development of effective therapeutic strategies [Hanahan and Weinberg, 2000].

With growing knowledge about cancer cells, four new hallmarks have emerged: genome instability and mutation, tumor-promoting inflammation, reprogramming energy metabolism, and evading immune destruction (Figure 5 - light green text boxes) [Hanahan and Weinberg, 2011].

Hallmark 7: Genome Instability and Mutation

The acquisition of the previous six hallmarks is dependent on a sucession of alterations in the genomes of neoplastic cells. In other words, specific mutant genotypes confer a selective advantage on subclones of cells, allowing them to outgrow and eventually dominate the local tissue environment. Knowing this, multistep tumor progression can be seen as a succession of clonal expansions, where each of them is triggered by the possibility of enabling a mutant genotype [Berdasco and Esteller, 2010]; [Esteller, 2007]; [Jones and Baylin, 2007].

The ability to detect and resolve alterations that can occur in **DNA** made by genome maintenance systems ensures that the possibility of having a spontaneous mutation is usually very low during each cell generation. To orchestrate tumorigenesis, cancer cells need to acquire a roster of mutant genes and they often achieve this by increasing the rates of mutations [Negrini et al., 2010]; [Salk et al., 2010]. This mutability is either obtained by increasing the sensitivity to specific genes or through a breakdown in one or several components of the genomic maintenance machinery, or in some cases both. In addition, this mechanism can be accelerated by compromising the surveillance systems that usually are responsible for the monitoring of the genomic integrity and as a consequence, force the genetically damaged cells into either senescence or apoptosis [Jackson and Bartek, 2009]; [Sigal and Rotter, 2000]. **TP53** has a very important role in these steps, leading to its being called the ''guardian of the genome'' [Lane, 1992]. Therefore, by having the ability to affect such genes, cancer cells are able to acquire capabilities that will

be fundamental for other hallmarks.

Hallmark 8: Tumor-Promoting Inflammation

In the year 2000 there were already some clues about the tumor-associated inflammatory response having the paradoxical effect of enhancing tumorigenesis and progression, helping to acquire hallmark capabilities. Since then, research about the intersections between inflammation and cancer pathogenesis has evolved, providing abundant and compelling demonstrations of the functionality of tumor-promoting effects that immune cells have on neoplastic progression [DeNardo et al., 2010]; [Grivennikov et al., 2010]; [Qian and Pollard, 2010]; [Colotta et al., 2009].

Inflammation can contribute to the appearance of multiple hallmark capabilities due to its ability to supply the tumor microenvironment with bioactive molecules. These include growth factors that help sustaining the proliferative signaling, survival factors that help limit cell death, proangiogenic factors which are extracellular matrix-modifying enzymes that facilitate angiogenesis, invasion, and metastasis, and also inductive signals that initiate the activation of epithelial-to-mesenchymal transition **(EMT)** and other hallmark-facilitating programs [DeNardo et al., 2010]; [Grivennikov et al., 2010]; [Qian and Pollard, 2010]; [Karnoub and Weinberg, 2007].

Another important fact is that inflammation is, in some cases, evident in very early stages of neoplastic progression and is capable of fostering the development of incipient neoplasias into full-blown cancers [Qian and Pollard, 2010]; [De Visser et al., 2006]. Additionally, inflammatory cells are able to release chemicals, such as reactive oxygen species, that are actively mutagenic for nearby cancer cells, speeding their genetic evolution toward states of malignancy [Grivennikov et al., 2010]. As such, inflammation can be considered an important hallmark due to its contributions to the acquisition of core hallmark capabilities.

Hallmark 9: Reprogramming Energy Metabolism

Uncontrolled cell proliferation involves both deregulation of cell proliferation and energy adjustments related to the metabolism in order to fuel the cell's growth and division. In aerobic conditions, normal cells process glucose by first creating pyruvate via glycolysis in the cytosol which is then imported to the mitochondria to produce energy (**ATP**) via the oxidative phosphorylation pathway leaving carbon dioxide. Under anaerobic conditions, glycolysis is favored and only very low quantities of pyruvate and water are dispatched to the oxygen-consuming mitochondria cycle.

Otto Warburg was the first to observe anomalous characteristics of cancer cell energy metabolism [Warburg, 1930]; [Warburg, 1956a]; [Warburg, 1956b]. He noticed that even when in an environment with

oxygen, cancer cells are able to reprogram their glucose metabolism, and as a consequence their energy production. With this ability they are capable of limiting their energy metabolism to glycolysis, leading to a state that has been termed ''aerobic glycolysis''. At first glance, this reprogramming of energy metabolism seems to be a little counterintuitive, since with this, cancer cells must compensate for the lower **ATP** production from glycolysis, when compared to the mitochondrial oxidative phosphorylation. However, to compensate for this, cells upregulate glucose transporters, notably **GLUT1**, which considerably increases glucose import into the cytoplasm of the cell [Jones and Thompson, 2009]; [DeBerardinis et al., 2008]; [Hsu and Sabatini, 2008].

In an attempt to explain this phenomenon, an hypothesis made by Potter VR [1958], and refined by Vander Heiden et al. [2009], indicates that the increased glycolysis allows the spread of glycolytic intermediates into different biosynthetic pathways, including those responsible for generating nucleosides and amino acids. This will facilitate the biosynthesis of the macromolecules and organelles, components needed to assemble new cells. Additionally, the Warburg-like metabolism seems to be present in various rapidly dividing embryonic tissues, suggesting once again that this alteration supports the large-scale biosynthetic programs needed for active cell proliferation.

Interestingly, it was found that some tumors have two subpopulations of cancer cells that differ in their energy-generating pathways. One of the subpopulations consists of glucose-dependent cells that produce lactate (''Warburg-effect'') , and the second subpopulation preferentially import and utilize the lactate produced by neighbor cells as their main energy source, utilizing a portion of the citric acid cycle to achieve that purpose [Kennedy and Dewhirst, 2010]; [Feron, 2009]; [Semenza et al., 2008]. These subpopulations function symbiotically: the cancer cells who were deprived of oxygen depend exclusively on glucose for fuel and secrete lactate as waste. This molecule is then imported and preferentially used as fuel by better-oxygenated cells [Kennedy and Dewhirst, 2010]; [Feron, 2009]; [Semenza et al., 2008]. As such, energy metabolism reprogramming can be considered an important hallmark due to the ability to make cancer cells adapt their energy production to their local environment.

Hallmark 10: Evading Immune Destruction

Another unresolved issue that surrounds tumor formation has to do with the role that the immune system has in the eradication or resistance to the progression of late-stage tumors and micrometastases. The immune surveillance theory proposes that both cells and tissues are being continuously monitored by an ever-alert immune system that is responsible for recognizing and eliminating a huge majority of incipient cancer cells and thus tumors. According to this line of thought, when a solid tumor appears they either

had to avoid being detected by the numerous strategies of the immune system or have been able to limit the extent of immunological killing.

This flawed immunological monitoring of tumors appeared to be validated since there was a noticeable increase of certain types of cancers appearing in immunocompromised individuals [Vajdic and Van Leeuwen, 2009]. However, most of these cancers are virus-induced, suggesting that the control for this type of cancers is usually dependent on the reduction of the viral burden in infected individuals, by eliminating virus-infected cells. These observations, and the support of evidence from clinical epidemiology and genetically engineered mice, suggest that the immune system operates as a barrier to tumor formation and progression, at least in some forms of non-virus induced cancer. This fact makes this specific hallmark important due to its ability to shutdown one of our most important lines of defense [Teng et al., 2008].

Recently in 2022, four new hallmarks were proposed: unlocking phenotypic plasticity, nonmutational epigenetic reprogramming, senescent cells and polymorphic microbiomes (Figure 5 - light purple), that are still not commonly referred but that are based on strong scientific findings [Hanahan, 2022].

2.3 Resting metabolic-rates of major organs

The resting energy expenditure **(REE)** reflects the baseline energy needs of the body to maintain its essential functions while at rest. Different organs and tissues in the body have varying energy requirements and metabolic functions, leading to differences in their baseline metabolic rates. These differences are usually represented by individual organ-specific values, known as "ki values".

The measurement of ki values in each individual organ is challenging, with the most widely accepted values being determined in 1992 by Marinos Elia [Elia, 1992]. This method involved the catheterization of arterial and venous ends of a circulatory region *in vivo*, making it a highly invasive and error-prone approach.

In contrast, **REE** can be estimated through a simpler, non-invasive gasometry approach, which measures the amount of oxygen consumed and carbon dioxide released. This method provides a measure of **REE** in kilocalories per kilogram per day, and it forms the foundation for understanding an individual's total daily energy expenditure.

It is important to note that the summation of all ki values in the body must equal the resting energy expenditure, which is the total amount of energy spent by the body at rest. In the anatomogram (Figure 6)

are represented the ki values for specific organs measured by Marinos Elia [Elia, 1992]. Due to their different values they can be classified in high and low metabolic-rate organs.

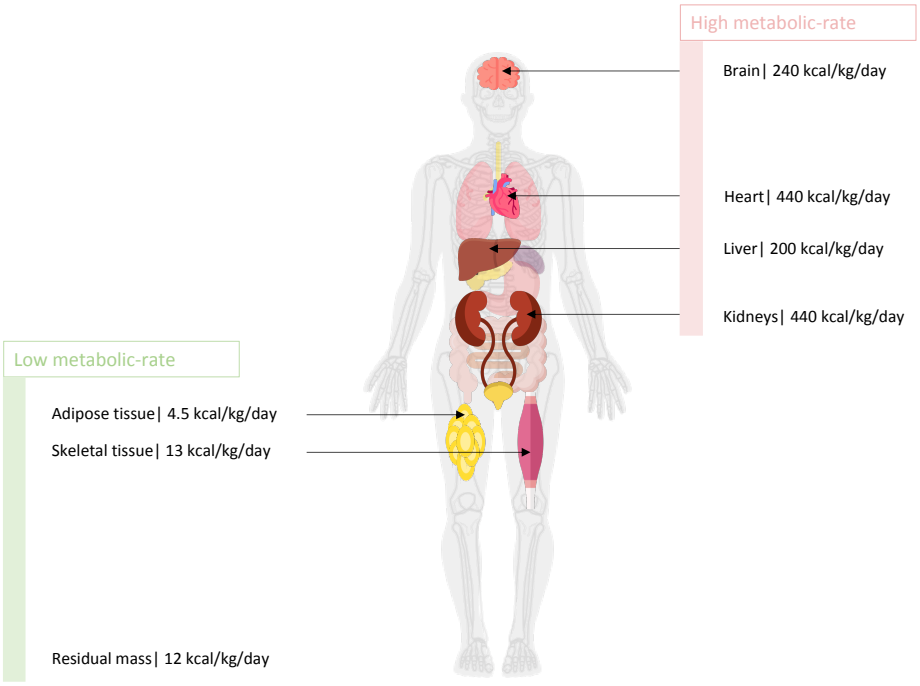


Figure 6: Metabolic rates per organ, as determined by Marinos Elia [Elia, 1992]. High metabolic-rate organs coloured in pink, and the low metabolic-rate organs are coloured in green.

2.4 The mitochondrion

2.4.1 Historical perspective

The discovery of mitochondria dates back to the 1890s, were it was described cytologically by Richard Altmann [Altmann, 1890]. The name mitochondrion was introduced in 1898 by Carl Benda [Benda, 1898], and originates from the Greek "mitos" (thread) and "chondros" (granule), referring to the appearance of these structures during spermatogenesis [Ernster and Schatz, 1981]. In 1913, the biochemist Otto Warburg linked cellular respiration to grana derived from guinea pig liver extracts, which functioned to enhance the activity of iron-containing enzymes [Ernster and Schatz, 1981]. In the following decades, many scientists started to study and discover the machinery that drives mitochondrial respiration, including tricarboxylic acid (**TCA**) cycle and fatty acid b-oxidation enzymes in the mitochondrial matrix that generate electron donors in order to fuel respiration and electron transport chain (**ETC**) complexes, and **ATP** syn-

thase in the inner mitochondrial membrane (**IMM**) that carry out oxidative phosphorylation [Ernster and Schatz, 1981].

2.4.2 Functions

Mitochondria are small, membrane-bound organelles found in eukaryotic cells with a complex internal structure. They have a smooth outer membrane that encloses the entire organelle, while the inner membrane is folded into projections called cristae, which increases the surface area for cellular respiration [Giacomello et al., 2020].

They are often referred to as the "powerhouses of the cell" because they generate the majority of a cell's energy supply through a process called cellular respiration, which takes place through a series of chemical reactions that transfer electrons from the fuel molecules, usually glucose, to oxygen, generating a proton gradient that drives the production of **ATP**, which is used as a source of chemical energy [Sweeney and Williamson, 2006].

Therefore, the oxidative phosphorylation uses oxygen as the terminal electron acceptor: the electrons from **NADH** are transported to oxygen by the proton-pumping electron transport chain, and the backflow of the pumped protons results in **ATP** formation by the mitochondrial **ATP** synthase. Such typical mitochondria occur in mammals, plants and various groups of unicellular eukaryotes, all of which are dependent on oxygen and thrive exclusively in oxic environments [Voet et al., 2016]. In humans, these organelles harbor a circular genome encoding their own **RNA**s and 13 proteins, including many of the essential subunits present in the protein complexes of the proton-pumping electron transport chain.

The ''Warburg effect'', refers to the fermentation of glucose to lactate in the presence of oxygen as opposed to the complete oxidation of glucose to fuel mitochondrial respiration, and this observation brought attention to the role of mitochondria in tumorigenesis [Warburg, 1956b]. Nowadays, we understand that although damaged mitochondria drive the Warburg effect in some cases, many cancer cells that display Warburg metabolism possess intact mitochondrial respiration, and some cancer subtypes are even dependent on mitochondrial respiration. Decades of studies on mitochondrial respiration in cancer have set the framework for a new frontier focused on additional functions of mitochondria in cancer, making possible the identification of pleiotropic roles in tumorigenesis.

Mitochondria are involved in a variety of other cellular processes such as (Figure 7): **regulation of cellular metabolism** by controlling the balance between glucose oxidation and lipid synthesis; **calcium storage**, which is important for several cellular processes such as muscle contraction and neurotransmitter release. Mitochondria also play a key role in **apoptosis** in response to cellular damage or stress; in

maintenance of cellular redox state, which is important for proper cellular function and survival; in **signalling**, therefore influencing cellular processes such as gene expression, differentiation, and growth [Voet et al., 2016]; and the **biogenesis of iron sulfur-clusters** which play important roles in pathways ranging from metabolism to DNA repair [Rouault, 2012].

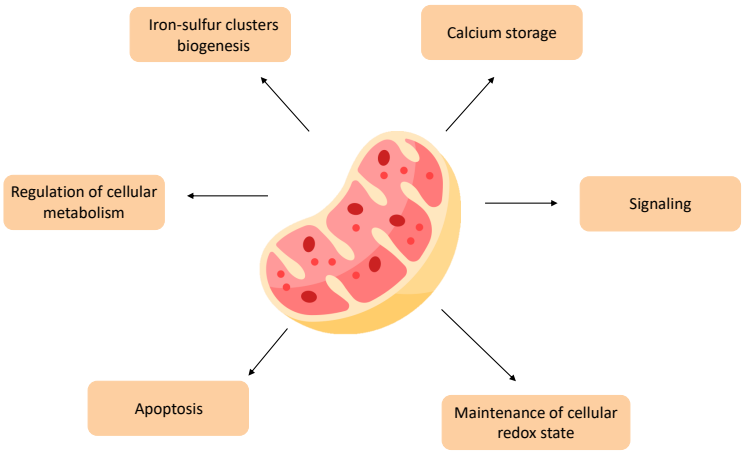


Figure 7: Mitochondrial central functions.

These functions make them important cellular stress sensors and allow for cellular adaptation to the environment. Mitochondria also have considerable flexibility for tumor cell growth and survival in harsh environments, such as during nutrient depletion, hypoxia, and cancer treatments, making them key players in tumorigenesis. There is no specific role for mitochondria in cancer development. Instead, mitochondrial functions in cancer cells vary depending on genetic, environmental, and tissue differences between tumors. This highlights the fact that the biology of mitochondria in cancer is fundamental to our understanding of cancer biology since many classical cancer hallmarks result in altered mitochondrial function [Vyas et al., 2016].

2.4.3 Mitochondrial biogenesis and Turnover

Mitochondrial mass is influenced by two opposing pathways, biogenesis and turnover, and these have been shown to be both positive and negative regulators of tumorigenesis. The role of mitochondrial biogenesis in cancer is regulated by many factors, including metabolic state, tumor heterogeneity, tissue type, microenvironment, and tumor stage *(vide infra)*. Additionally, mitophagy, which is the selective autophagic pathway for mitochondrial turnover, is able to maintain a healthy mitochondrial population. The regulation of these mitochondrial pathways is central to key oncogenic signaling pathways [Vyas et al., 2016].

2.4.4 Transcriptional and Signaling networks regulating biogenesis

The coordination of mitochondrial and nuclear-localized genes that are responsible for encoding mitochondrial proteins regulates mitochondrial biogenesis. The transcriptional coactivator peroxisome proliferator-activated receptor gamma coactivator-1 alpha **(PGC-1a)** is able to interact with multiple transcription factors making it a central regulator of mitochondrial biogenesis [Tan et al., 2016]. Different levels of this coactivator often reveal how much is the tumor reliant on mitochondrial mass, the bigger the expression of **PGC-1a** the more dependent the tumor is on mitochondrial respiration [Tan et al., 2016]. This phenomenon however is only valid for some cancer types, because the overexpression of **PGC-1a** on those cases can induce apoptosis [Tan et al., 2016]. Therefore, it is important to identify factors that contribute to the dichotomous effect of **PGC-1a** on tumor viability, as this has the potential to identify specific susceptibilities for cancer subtypes.

The transcription factor of the Myelocytomatosis family **c-Myc** is a key activator of mitochondrial biogenesis in cancer and it regulates cell cycle, growth, metabolism, and apoptosis. Over 400 mitochondrial genes have been identified as **c-Myc** targets, and initial studies demonstrated that the gain or loss of Myc resulted in the increase or reduction of mitochondrial mass [Li et al., 2005]. In normal physiology, **c-Myc** is able to couple mitochondrial biogenesis with cell-cycle progression. However, due to oncogenic **c-Myc**, the mitochondrial biogenesis increases as well as the biosynthetic and respiratory capacity of the cell by upregulating mitochondrial metabolism to support rapid proliferation which, complementing **c-Myc**'s effects, stimulate cell-cycle progression and glycolytic metabolism to coordinate rapid cell growth (Figure 8).

Another molecule responsible for mitochondrial biogenesis is the mammalian target of rapamycin **(mTOR)** signaling pathway. This target is critical for cellular growth and energy homeostasis and is misregulated in many diseases including cancer. **mTOR** regulates mitochondrial biogenesis both transcriptionally via **PGC-1a**/Yin Yang 1 **(YY1)** activation, which results in mitochondrial gene expression, and translationally via repression of inhibitory 4E-binding proteins **(4E-BPs)** that downregulate the translation of nuclear-encoded mitochondrial proteins [Morita et al., 2015] (Figure 8).

The transcriptional networks responsible for the regulation of mitochondrial biogenesis impact therapeutic outcomes by providing cancer cells with metabolic flexibility making them able to adapt to targeted treatment and tumor microenvironment. Cancer cells are capable of adapting their mitochondrial function according to a specific stress situation. An example is the upregulation of **c-Myc** and glycolytic gene expression that confer resistance to metformin, which is a complex I inhibitor in pancreatic cancer cells that actively utilize mitochondrial respiration due to **PGC-1a** expression [Sancho et al., 2015].

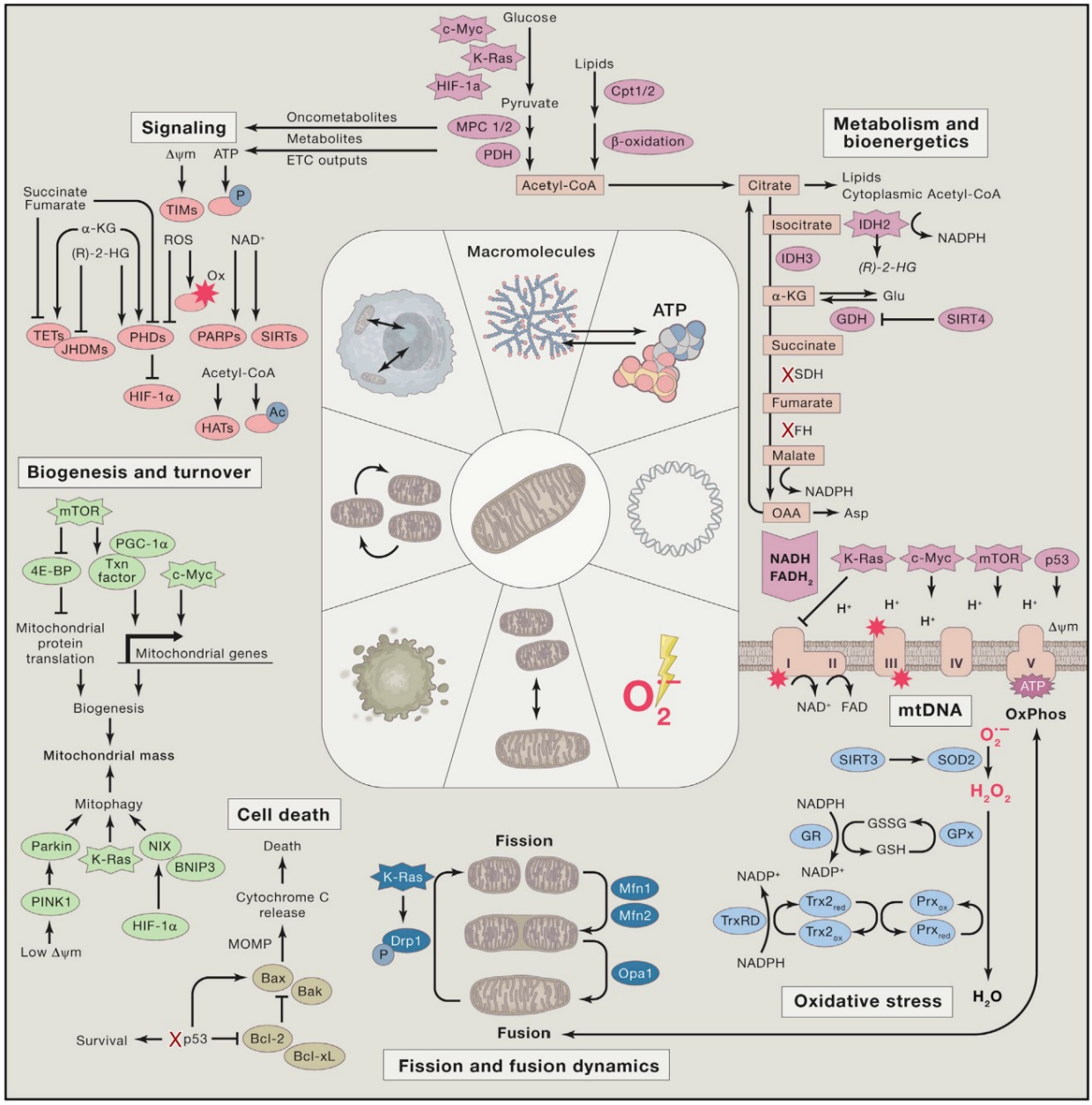


Figure 8: The role of the mitochondrion in the metabolism and bioenergetics, oxidative stress regulation, fission and fusion dynamics, cell death, biogenesis and turnover, and signaling in tumorigenesis. Retrieved from: [Vyas et al., 2016]

On a similar note, **c-Myc** dependent mitochondrial biogenesis is normally counteracted by the **HIF-1a** signaling pathway, yet this balance is altered during the oncogenic **c-Myc**-driven transformation [Dang et al., 2008]. With this information, an important consideration to have in cancer therapeutics will be addressing different routes of bioenergetic plasticity provided by the mitochondrion.

2.4.5 Mitophagy

The elimination of damaged mitochondria via mitophagy is critical for cellular fitness since dysfunctional mitochondria can impair **ETC** function and increase oxidative stress. An important trigger to initialize mitophagy is via the PTEN-induced putative kinase 1 **(PINK1/ Parkin** pathway. This pathway is activated upon mitochondrial membrane depolarization, which is a signal of mitochondrial dysfunction that can result from multiple causes such as lack of reducing equivalents, hypoxia, and impaired electron transport. Another pathway for mitophagy induction is through the **HIF-1a** target genes **Bcl-2** and adenovirus E1B 19 kDa-interacting protein 3 **(BNIP3)** and **BNIP3**-like **(BNIP3L/NIX)**, responsible for the inhibition of mitochondrial respiration during hypoxic conditions that could result in excessive **ROS**.

Similar to autophagy, which is shown to be both pro- and anti-tumorigenic based on context, the function of mitophagy depends on tumor stage [Mancias and Kimmelman, 2016]. Mitophagy-deficient Parkin null mice develop spontaneous hepatic tumors, and this loss increases tumorigenesis in multiple cancer models [Matsuda et al., 2015]. In addition, **BNIP3L/NIX** are identified as tumor suppressors in multiple cancer models [Chourasia et al., 2015]. With this in mind, in certain stages of tumorigenesis, decreased mitophagy may allow some dysfunctional mitochondria to persist, generating increased tumor-promoting **ROS** or other tumorigenic mitochondrial signals. In contrast, established tumors may also require mitophagy for stress adaptation and survival. Supporting this concept, **BNIP3** is normally induced in patient glioblastoma samples as a response to hypoxia caused by anti-angiogenic therapy and combinatorial angiogenesis. Additionally, oncogenic **K-Ras**-driven transformation upregulates mitophagy, making the accumulation of dysfunctional mitochondria switch the adenomas' tumor state from carcinoma to benign oncocytoma [Guo et al., 2013].

2.4.6 Fission and Fusion dynamics

The balance between fission and fusion is what dictates the morphology of mitochondria which are very dynamic organelles. With morphologies ranging from nearly transparent spherical organelles with hardly any cristae, to highly dense networks of cristae inside a big interconnected structure, bounded inside the same double membrane [Mishra and Chan, 2016]; [Kasahara and Scorrano, 2014]. To perform the

critical step in mitochondrial membrane fission is used the dynamin-related protein-1 **(Drp1)** recruitment to mitochondria and the interaction with its outer mitochondria membrane (**OMM**) receptors, allowing the membrane constriction fueled by GTPase activity. The regulation of the mitochondrial translocation activity made by **Drp1** initializes with phosphorylation mediated by multiple kinases that respond to distinct cell-cycle and stress conditions [Mishra and Chan, 2016].

Mitochondrial fusion is mediated by mitofusins, **Mfn1** and **Mfn2**, along side the optic atrophy-1 **(Opa1)** protein.

Various studies have shown that an imbalance of fission and fusion activities plays a role in cancer, where increased fission activities play a role and/or decreased fusion, resulting in a fragmented mitochondrial network [Senft and Ze'ev, 2016]. It is important to note that during these studies the ability to restore fused mitochondrial networks through either **Drp1** knockdown/ inhibition or **Mfn2** overexpression, was able to damage cancer cell growth, suggesting that the remodeling of the mitochondrial network is important for tumorigenesis. In contrast, the increase of **Drp1** expression is associated with a migratory phenotype in multiple cancer types, further highlighting the role of mitochondrial dynamics in metastasis [Senft and Ze'ev, 2016].

2.4.7 Cell death

One important characteristic of cancer cells is their ability to evade cell death, a phenomenon tightly linked to mitochondria. The pro-apoptotic **Bcl-2** family members **Bax** and **Bak** are taken to the outer mitochondrial membrane **(OMM)** and oligomerize in order to mediate mitochondrial outer membrane permeabilization (**MOMP**), which results in pore formation and cytochrome c release from mitochondria into the cytosol to allow the activation of caspases, enzymes needed by the apoptotic program. In healthy cells, anti-apoptotic family members like **Bcl-2** and **Bcl-xL** bind and inhibit the pro-apoptotic family members **Bax**/**Bak**. Tumors found a way to avoid apoptosis by downregulating pro-apoptotic **Bcl-2** and/or upregulating anti-apoptotic **Bcl-2** genes [Lopez and Tait, 2015]. The balance between pro- and anti-apoptotic proteins affects the susceptibility of a cancer cell to apoptotic stimuli and this behavior may help predict how a tumor will respond to chemotherapy [Sarosiek et al., 2013].

The shape of a mitochondrion can also dictate apoptotic susceptibility, as **Drp1** loss delays cytochrome c release and apoptotic induction [Martinou and Youle, 2011]. To avoid this problem, a GTPase-independent function of **Drp1** in membrane remodeling and hemifusion results in **Bax** oligomerization mediation of mitochondrial outer membrane permeabilization, which indicates that **Drp1** can promote apoptosis independent of fission [Martinou and Youle, 2011]. This can be further proved since the inhibi-

tion of **Drp1** was able to recover the sensitivity to apoptotic stimuli by restoring a balanced mitochondrial network [Renault et al., 2015]. Additionally, **Mfn1** is a target of the **MEK**/**ERK** signaling pathway and phosphorylated **Mfn1** inhibits mitochondria fusion and interacts with **Bak** to stimulate its oligomerization and subsequent mediation of mitochondrial outer membrane permeabilization [Pyakurel et al., 2015]. Therefore, while fission and fusion do not necessarily regulate apoptosis, a balance between these phenomena can generate different mitochondrial shapes that support the interactions with pro-apoptotic **Bcl-2** proteins.

2.4.8 Oxidative stress

Reactive oxygen species **(ROS)** can appear in the form of hydrogen peroxide, superoxide, and hydroxyl free radicals that are produced during physiological metabolic reactions. Mitochondria are one of the biggest contributors to cellular **ROS** and to compensate for it they have multiple antioxidant pathways to neutralize these molecules, including superoxide dismutase **(SOD2)**, glutathione, thioredoxin, and peroxiredoxins. The observation that the **ROS** levels are high in cancer cells allowed the formulation of a very simple hypothesis where **ROS** inhibition could be a successful therapeutic strategy. However, with advances in the study of cancer cells, a more complex scenario is being pictured where **ROS** stimulates signaling and proliferation, and the concomitant upregulation of antioxidant pathways prevent ROS-mediated cytotoxicity and may even enhance tumor survival [Shadel and Horvath, 2015]; [Sullivan and Chandel, 2014].

Multiple physiological reactions, such as electron transport by the **ETC** and NAD(P)H oxidases result in **ROS** production, and these are often aggravated during tumorigenesis by oncogenic signaling, **ETC** mutations, and hypoxic microenvironments. High levels of **ROS** contribute to the oxidation of macromolecules, such as lipids, proteins, and **DNA**, which can lead to genomic instability promoting mutations. However, many tumors show slightly higher levels of **ROS** that can help to regulate cell signaling via cysteine oxidation [Sullivan and Chandel, 2014]. Additionally, ROS-mediated regulation of oncogenic signaling also affects oxidation of cysteines in metastasis in **Src** which are able to then increase their oncogenic ability, promoting their migration and metastasis in multiple tumor types. These phenotypes were successfully blocked by **ROS** scavengers such as hydrogen peroxide and nitric oxide [Porporato et al., 2014].

In order to respond to a higher **ROS** level, many tumor cells try to upregulate protective antioxidant pathways. Examples such as the oncogenic **K-Ras**, **B-raf**, and **c-Myc** actively inhibit **ROS** through the regulation of the nuclear factor (erythroid-derived 2)-like 2 **(NRF2)**. This is a transcriptional regulator of the antioxidant response, which helps to promote tumorigenesis [DeNicola et al., 2011]. Similarly, a study made in melanoma found that migrating tumor cells had higher levels of **NADPH** than primary tumor

sites. This might be associated with combating the increased **ROS** caused by the stress associated with the metastatic process [Piskounova et al., 2015]. Thus, successful tumors maintain **ROS** levels within a specific window that helps them stimulate proliferation without causing cytotoxicity. The balance of **ROS** production and antioxidant expression is critical for the maintenance of the tumor-promoting **ROS** levels.

2.5 Analysis methods

2.5.1 Principal Component Analysis (PCA)

Principal Component Analysis (**PCA**) is a mathematical technique used to transform a set of potentially correlated variables into a smaller set of independent variables, referred to as principal components [Richardson, 2009]. **PCA** reduces the dimensionality of large datasets through vector space transformations and mathematical projections, making it easier to interpret the data by identifying trends, patterns, and outliers [Holmes and Huber, 2018].

The **PCA** algorithm is as follows: first it is essential to standardize numeric values, ensuring meaningful comparisons. This typically involves centering the data (subtracting the mean) and scaling it (dividing by the standard deviation). However, in some cases, maintaining different scales might be necessary if variable importance varies. The aim of these transformations is usually variance stabilization to replicate measurements' variances consistent across a variable's dynamic range. In contrast, standardization aims to make different variables' scales comparable [Holmes and Huber, 2018].

Then a covariance matrix is created which is a squared matrix, with dimensions equal to the number of features in the dataset. Each entry in the matrix captures the covariance between two separate dimensions, revealing how they vary together. The main diagonal contains variances, representing how each dimension varies with itself, while off-diagonal entries show the covariances between pairs of dimensions. The covariance matrix is symmetrical about the main diagonal, and its values help identify the relationships and variability between different dimensions in the data, a crucial step in **PCA** for dimensionality reduction and feature analysis [Smith, 2002].

With the matrix created we then can find the eigenvectors and eigenvalues. Eigenvectors are special vectors that play a crucial role in transforming and understanding data. They arise from the interaction between a transformation matrix and vectors. Eigenvectors represent directions in the original feature space that remain unchanged when transformed by the matrix. When multiplied by the matrix, they are only scaled, but their direction remains the same. Eigenvectors are significant in **PCA** as they form a new coordinate system for the data, allowing for dimensionality reduction. They are typically normalized to have

a length of one and are orthogonal, meaning they are at right angles to each other, making them a useful basis for expressing data in terms of these directions rather than the original axes. Finding eigenvectors can be challenging for larger matrices, but specialized math libraries are available for this purpose [Smith, 2002].

As for the eigenvalues these are closely linked to eigenvectors and always appear as pairs. An eigenvalue represents the scaling factor by which an eigenvector is stretched or compressed when multiplied by a square matrix. For example, if the scaling factor is 4, then 4 is the associated eigenvalue for that specific eigenvector. Regardless of how the eigenvector is scaled before matrix multiplication, the result will be a multiple of the original eigenvector, with the eigenvalue determining the factor. These pairs of eigenvalues and eigenvectors provide essential information about how data is transformed and oriented in the new coordinate system [Smith, 2002].

Finally, in a **PCA** analysis, there is a selection of components using eigenvectors and eigenvalues from the covariance matrix which allows dimensionality reduction. The eigenvalues associated with the eigenvectors provide a measure of their significance. The eigenvector with the highest eigenvalue is considered the principal component, representing the most prominent relationship among data dimensions. Typically, after finding the eigenvectors, they are ordered by eigenvalue in descending order, indicating their importance. You can choose to ignore components with lower eigenvalues, sacrificing some information but reducing dimensionality [Smith, 2002].

For our work, in order to combine gene expression values from two distinct datasets, namely, **TCGA** and **GTEx**, it is crucial to ensure comparability between the data. Inconsistencies could impact the conclusions about gene interactions. Therefore, by applying **PCA** to our data and visualizing the first two components (that together explain over 85% of the variance), we can verify if the gene expression values derived from two alternative databases are comparable.

PCA in R

There are several methods in R for conducting this analysis, such as the PCAtools package available in BioConductor [Gentleman et al., 2004], the FactoMineR [Lê et al., 2008] package, or using the base R functions prcomp() and princomp(). The main difference between these functions is their underlying approach: princomp() uses spectral decomposition, while prcomp() and PCA() from FactoMineR use singular value decomposition (**SVD**). (Further details regarding these two alternative **PCA** approaches are beyond the scope of this work, but a good overview is presented here: [Jolliffe and Cadima, 2016]). In this work, we used the prcomp() function from base R.

2.5.2 Differential Gene Expression

Gene expression technologies play a pivotal role in molecular biology research, enabling the assessment of transcriptional activity in various tissues or cell populations. These assessments help to identify alterations in gene expression linked to specific treatment conditions or phenotypes of interest. Gene expression studies can take the form of randomized experiments, involving perturbations like gene knock-outs or induced stressors, providing valuable insights into both normal cellular processes and disease mechanisms. Alternatively, they can be observational studies, comparing different phenotypes, such as diseased versus healthy tissues or cells from distinct populations. This approach is frequently employed in cancer research and the study of cell development [Hurd and Nelson, 2009]. There are several methods for measuring gene expression, with microarrays and RNA-sequencing being two of the most commonly used.

Microarray technology gained prominence after genome sequencing projects because it relies on prior knowledge of the query genome. Although valuable for gene expression analysis, it has certain limitations. One significant constraint is its reliance on prior knowledge of the genome or genomic features, which poses challenges when dealing with incomplete, incorrect, or outdated genome annotations [Hurd and Nelson, 2009].

In contrast, **RNA** sequencing (RNA-Seq) leverages high-throughput sequencing techniques to offer a comprehensive view of a cell's transcriptome. RNA-Seq surpasses previous microarray-based methods by providing deeper coverage and higher resolution for understanding the dynamic nature of gene expression. Recent advances in RNA-Seq workflows, encompassing sample preparation, sequencing platforms, and bioinformatics data analysis, have enabled thorough profiling of the transcriptome, offering insights into various physiological and pathological conditions [Kukurba and Montgomery, 2015].

However, the reliability of RNA-Seq data hinges on the accurate mapping of sequencing reads to reference genomes or efficient *de novo* assembly. This process can be computationally intensive, requiring substantial computing resources to handle the vast volume of small reads within a reasonable timeframe. Moreover, the relatively higher error rate associated with next-generation sequencing (**NGS**) data necessitates consideration of non-perfect matches during read mapping. This becomes particularly relevant when identifying allele-specific expressions in RNA-Seq data, especially in the context of detecting single nucleotide polymorphisms (**SNPs**) [Marguerat and Bähler, 2010].

Differential expression analysis of **RNA** sequencing experiments, relies on linear models to quantify the magnitude and direction of changes in gene expression. However, this process can be challenging due to two critical steps: setting up an appropriate model using design matrices and defining the desired comparisons through contrast matrices. The complexity arises because there is currently no comprehensive

catalog available for design and contrast matrices.

The design matrix serves a dual purpose: it shapes the model by defining the relationship between genes and explanatory variables, and it stores the values of these variables for each sample; on the other hand, the contrast matrix complements the design matrix by enabling the calculation of specific values of interest based on estimated parameters. The design matrix typically features columns associated with parameters and rows associated with individual samples. If the estimated parameters themselves are not the focus of interest, a contrast matrix can be employed to compute meaningful contrasts between these parameters. Multiple contrasts can be combined, with each column in the contrast matrix representing a distinct contrast and corresponding rows associated with columns in the related design matrix [Law et al., 2020].

In differential expression analysis, it is crucial to consider two types of explanatory variables: covariates and study variables. These variables can be numerical values representing quantitative measurements associated with the experiment. Examples include an individual's age, weight, or other molecular/cellular phenotypes. Categorical variables, on the other hand, are classifiers linked to the experiment's samples. These factors can be biological (e.g., disease status, genotype, metabolic rate, cell type) or technical (e.g., experiment time, sample batch, sequencing lane). Levels within a category represent unique values; for instance, the genotype factor might have two levels: "wildtype" and "mutant" [Law et al., 2020]. In the case of this work, our variables consist of factors with different levels. We have the 'cancer status' factor with two levels, 'cancer' and 'non cancer', and the 'metabolic rate' factor with two levels, 'high metabolic rate' and 'low metabolic rate'.

Another consideration relevant to the linear models used for differential gene expression, is whether to include or not an intercept term in the models. For example, with the model.matrix($\sim$ cancer status) model, where the formula contains only the 'cancer status' variable, the intercept term in the first column of the design matrix represents the y-intercept, while the second column represents the slope of the regression line. If the model forces the intercept to be zero, for example, model.matrix($\sim$ 0 + cancer status), the intercept term is removed from the design matrix, meaning that the regression line must intercept the y-axis at 0.

When comparing models with and without an intercept term for numeric variables, it is unsurprising that the model with an intercept term generally provides a better fit to the data. This is because it is less restrictive, allowing the y-intercept to be at any point, making it more flexible due to the extra parameter. However, for factor variables (i.e. categorical variables), models with and without an intercept term are equivalent, differing only in parameterization. In an intercept term model, it is determined by summing

both parameter estimates, while in a non-intercept term model, it is estimated directly as the second parameter. Consequently, the choice of design matrix is equivalent for categorical variables [Law et al., 2020].

In addition to considering intercept terms, it is essential to account for interactions between factors. An additive effect exists when the combined treatment effect equals the sum of the two individual effects (C - A - B = 0 or $\delta = 0$). In contrast, an interaction effect occurs when the combined treatment effect differs from the sum of the individual effects ($\delta \neq 0$). An interaction is considered synergistic if the combined effect is greater than the sum of the individual effects ($\delta > 0$) and repressive if it is smaller ($\delta < 0$) [Law et al., 2020].

In order to interpret the output from a differential expression analysis, it is crucial to grasp key concepts like p-values, fold change, and false discovery rate (**FDR**). The p-value quantifies the probability of observing the test statistic's given value or a more extreme one under the null hypothesis. Traditionally, a p-value cutoff of 0.05 is used to reject the null hypothesis [Ferreira and Patino, 2015]. Fold change detection (**FCD**) is a concept where the output depends on relative changes in input. Identical relative changes in the input result in identical output dynamics [Adler et al., 2014]. False discovery rate (**FDR**) procedures provide a more sensitive analysis compared to conventional family-wise error control, such as p-values. They also help control the rate of false positives [Genovese et al., 2002]. These concepts also help us prioritize genes for our study since they can help us find those with values more likely to be biologically meaningful.

For this study, there are two relevant alternative models that should be tested, namely the additive effect between metabolic rate and cancer status, with and without assuming an interaction between them:

1. Without assuming interactions between variables:

```
model.matrix(~ 0 + metabolic rate + cancer status)
```

2. Assuming a non-additive interaction between variables:

```
model.matrix(~ 0 + metabolic rate + cancer status +
+ cancer status:metabolic rate)
```

Considering that it is unlikely that there is no biological interaction between the cancer status of a cell and its metabolic rate, the second model (with an interaction term, and therefore more complete) was the one chosen for analysis and discussion.

Differential expression in R

For differential expression analysis of RNA-Seq data, several statistical methods and packages are available, such as DESeq [Anders and Huber, 2010] and edgeR [Robinson et al., 2010] (Table 1). However, it is crucial to interpret the results carefully as each method makes specific assumptions that might not always hold in the context of the data. Understanding the model parameters and their constraints is vital for drawing meaningful biological conclusions [Bullard et al., 2010].

Table 1: Information about the methods of normalization used for each package that can perform differential expression analysis.

Normalization method	Description	Accounted factors	Recommendations for use
DESeq2's median of ratios	Counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene.	Sequencing depth and RNA composition.	Gene count comparisons between samples and for DE analysis; NOT for within sample comparisons.
EdgeR's trimmed mean of M values (TMM)	Uses a weighted trimmed mean of the log expression ratios between samples.	Sequencing depth, RNA composition, and gene length.	Gene count comparisons between and within samples, and for DE analysis.

2.5.3 Clustering

Clustering is an unsupervised modeling technique that helps to analyze complex multivariate data by grouping similar data points into categories, often simplifying decision-making. Therefore, cluster validation is crucial, especially when no prior domain knowledge supports the existence of clusters [Holmes and Huber, 2018].

For instance, in cancer biology, clustering has provided valuable insights. Tumors previously considered identical based on location and histopathology have been categorized into distinct clusters based on molecular signatures, such as gene expression data [Marguerat and Bähler, 2010]. These clusterings

could lead to the identification of new, more relevant disease types, often associated with different patient outcomes.

To perform clustering, we first need to define what we mean by 'similarity.' Once relevant features are selected, we must decide how to combine the differences between these features into a single numerical value. One widely-used method for clustering is the **PAM** (Partitioning Around Medoids) or k-medoids method, outlined as follows [Kaufman and Rousseeuw, 2009]:

Begin with a matrix of 'p' features measured on 'n' observations. Randomly select 'k' distinct cluster centers from the 'n' observations as 'seeds.' Assign each remaining observation to the nearest cluster center. For each cluster, choose a new center (medoid) from the observations within that cluster, minimizing the sum of distances to cluster members. Repeat steps 3 and 4 until the clusters stabilize.

It's important to note that different initial seeds in Step 2 can yield varying results. A variant of this method, called k-means, replaces medoids with the arithmetic means (centers of gravity) of clusters. In **PAM**, cluster centers are observations, but this is not always the case with k-means.

These 'k-methods' are commonly used for clustering, especially when dealing with clusters of similar size and convex (blob-shaped) structures. However, they may fail when clusters differ significantly in size or exhibit non-spherical or non-elliptical shapes. To determine the number of clusters, one can perform sub-sampling and clustering repeatedly, identifying tight clusters that frequently group together [Holmes and Huber, 2018].

Various methods can calculate similarity, including the Euclidean distance, which measures the straight-line distance between two points in multidimensional space [Holmes and Huber, 2018].

$$d(A,B)=\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2}$$

The Manhattan distance sums the absolute differences in all coordinates.

$$d(A,B)=|a_1 - b_1| + |a_2 - b_2| + ... + |a_n - b_n|$$

When performing clustering, two key considerations are the choice of distance measure and the determination of the number of clusters, denoted as 'k.'

Clustering can be approached in different ways; for instance, density-based clustering is suitable when data consists of a few markers and numerous cells. This method identifies regions of high density separated by sparser regions, making it adaptable to non-convex clusters.

Hierarchical clustering

Hierarchical clustering is another commonly used variant. It follows a bottom-up approach, assembling similar observations and subclasses iteratively. This hierarchical structure, with its roots traced back to Aristotle's 'ladder of nature,' has practical applications across various fields. To implement hierarchical clustering, we need more than just the distances between individual objects. We must also decide how to compute the distance between a newly formed cluster and all other points or existing clusters. This choice influences the type of hierarchical clustering produced [Holmes and Huber, 2018].

Minimal Jump (Single Linkage): This method calculates the distance between clusters as the smallest distance between any two points in the two clusters. It tends to create clusters that resemble strings of contiguous points.

$$d_{\mathsf{SL}}(C_1, C_2) = \min_{x \in C_1, y \in C_2} \mathsf{dist}(x, y)$$

Maximum Jump (Complete Linkage): Here, the distance between clusters is defined as the largest distance between any two objects in the two clusters.

$$d_{\mathsf{CL}}(C_1, C_1) = \max_{x \in C_1, y \in C_1} \mathsf{dist}(x, y)$$

Group Average: This method computes the distance between clusters as an average of distances between all pairs of objects in the two clusters.

$$d_{\mathsf{GA}}(C_1, C_1) = \frac{1}{|C_1| \cdot |C_2|} \sum_{x \in C_1, y \in C_2} \mathsf{dist}(x, y)$$

These choices shape the hierarchical clustering trees. One notable advantage of hierarchical clustering is that it provides a graphical representation of grouping strength, reflected in the length of inner edges in the tree.

When clusters are expected to be of similar size, using group average, which minimizes within-class variance, is typically the preferred strategy [Holmes and Huber, 2018].

Fuzzy clustering

Fuzzy clustering is a soft clustering technique and represents a more nuanced approach compared to traditional hard clustering methods. In traditional clustering, each data point is unequivocally assigned to a single cluster, resulting in 'hard' boundaries between clusters. However, fuzzy clustering introduces a level of ambiguity by allowing data points to partially belong to multiple clusters. This ambiguity is quantified using membership coefficients, which range from 0 to 1 [Kaufman and Rousseeuw, 2009].

The key advantage of fuzzy clustering is its ability to provide detailed insights into the data's underlying structure. It can express that one data point predominantly belongs to a particular cluster while another data point may have a nearly equal association with multiple clusters.

However, this richness of information can also be overwhelming, especially as the number of data points and clusters increases. Still, the concept of fuzziness is appealing because it reflects the uncertainties often inherent in real-world data [Kaufman and Rousseeuw, 2009].

In this project, focusing on gene expression, clustering serves the purpose of grouping genes with similar expression patterns. This approach provides valuable insights into the genes' biological functions and the potential mechanistic relationships between them.

Clustering in R

Several packages and tools are available for performing clustering in this context. Notable options include 'fastcluster' [Müllner, 2013] which provides an efficient implementation of hierarchical clustering algorithms, 'cluster' [Maechler, 2018] which offers various partitioning methods like **PAM**, 'mclust' [Scrucca et al., 2016] a tool for model-based clustering using Gaussian Mixture Models (**GMM**) with the Expectation-Maximization (**EM**) algorithm, which iteratively refines **GMM**s providing a probabilistic framework for clustering and density estimation, and 'dbscan' [Hahsler et al., 2019] which efficiently implements the DBSCAN clustering algorithm.

Chapter 3

Methodology

Aiming to search for unique gene expression patterns characteristic of high metabolic-rate organs in cancer, an automated analysis pipeline was developed to ensure efficient, and reproducible data analysis. Therefore, for this study, we used the R programming language (version 4.2.2), which is freely available and particularly suited for statistical analysis and graphical data visualization. Additionally, this language is supported by an active development community, which greatly extends its functionality (Figure 9).

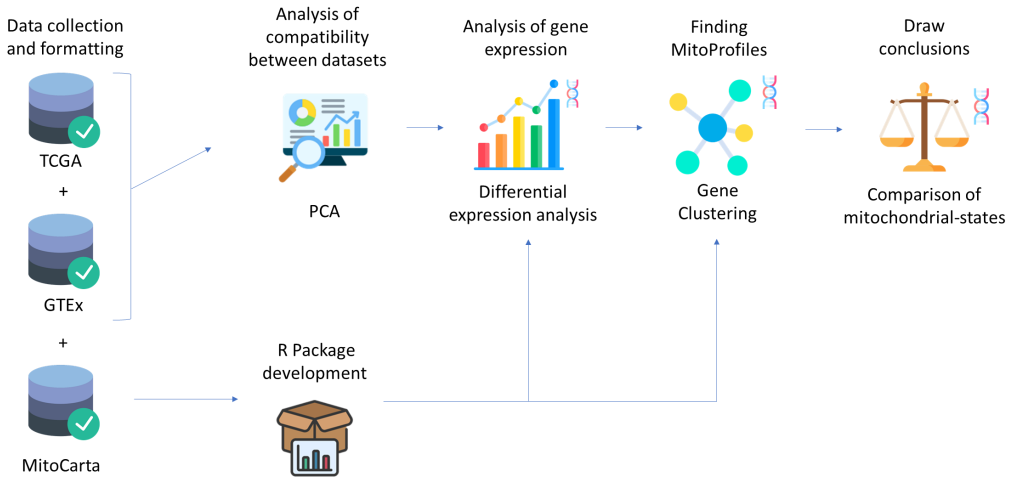


Figure 9: Project analysis workflow.

To analyze, explore, and understand the data, the free open-source integrated development environment (**IDE**) RStudio® (version 2023.03.0) was used. All data analyses were undertaken using custom R scripts, implementing additional functions to develop a pipeline suited to answer the research questions. As such, the R analysis pipeline presented in Figure 10 showcases the programming work developed throughout the thesis. All analyses were conducted in a Linux environment (Ubuntu distribution version 22.04.2), running on a virtual machine capable of storing big data and facilitating faster analyses.

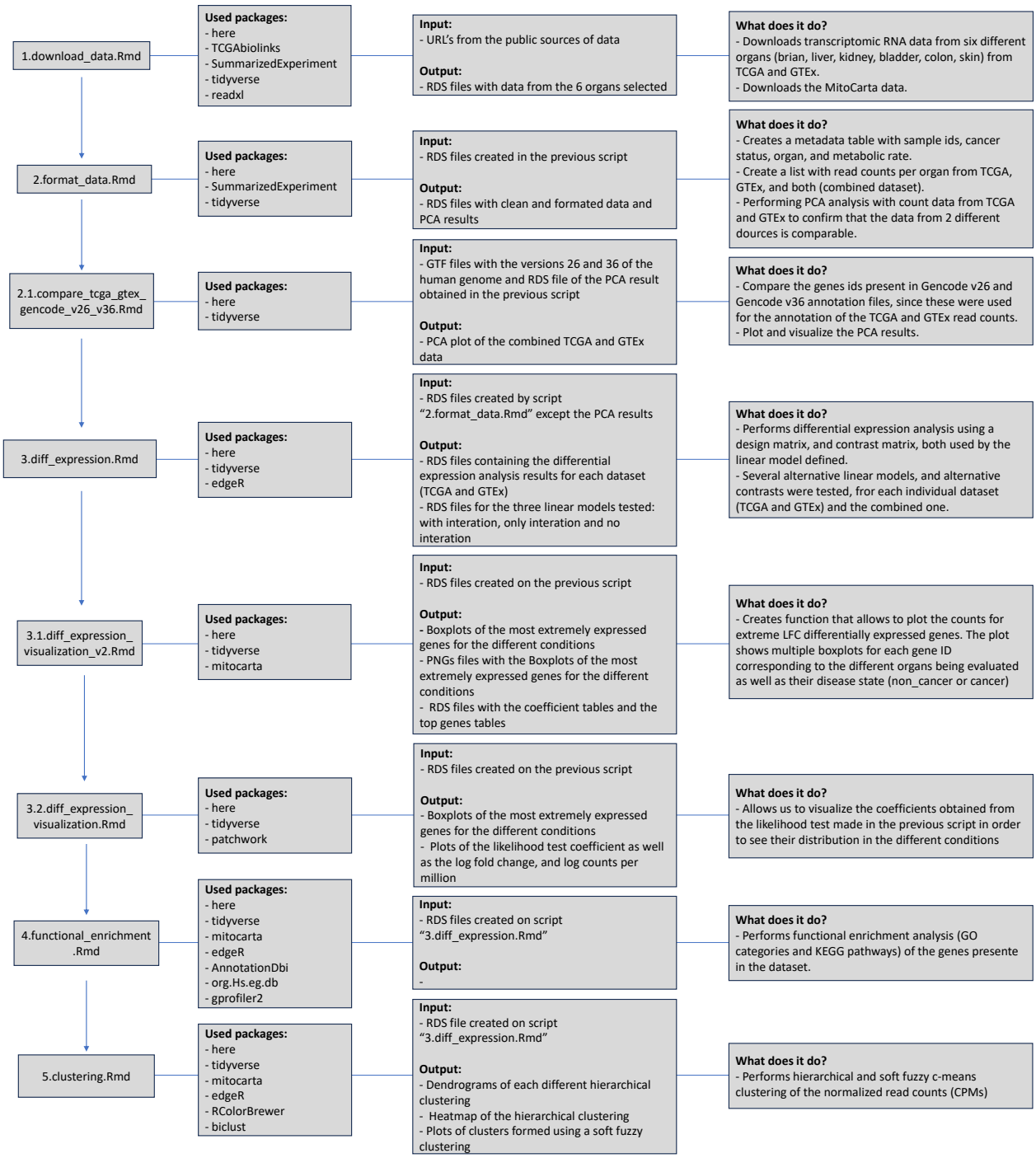


Figure 10: Structure of the data analysis workflow.

The data analysis scripts were version-controlled using Git and deposited in the GitHub repository (https://github.com/MitoProfiles/mitoprofiles) to comply with reproducible and open research best practices.

3.1 Data collection

Data collection is a crucial step for successful analysis. To test our hypothesis, we gathered data from three distinct databases: cancer patient data from The Cancer Genome Atlas (**TCGA**) [tcg], data from healthy patients in the Genotype-Tissue Expression (**GTEx**) [gte] database, and data related to human mitochondrial genes using MitoCarta [Mit]. For the first two databases, transcriptomics RNA-seq data were obtained for six organs, categorized into two groups: high-metabolism organs: brain, liver, and kidneys; and low-metabolism organs: bladder, colon, and skin. The annotated R markdown script "1.download_data.Rmd" provides the R code used for this step.

All data used in this project are open and publicly accessible, eliminating the need for additional access permissions and enabling their immediate usage.

To enhance data usability while maintaining a clear working environment, we established directories to systematically organize data in a meaningful file structure as shown in Annex A.1. The data collection phase focused on querying and downloading relevant datasets for this study.

3.1.1 TCGA data

To query and download data from the **TCGA** database, we used the TCGAbiolinks package [Colaprico et al., 2016] to retrieve sample population data from each organ into our local system. The database was accessed on 15 May 2023.

Three distinct functions were used, namely: (i) the GDCquery function, responsible for querying the database to retrieve public data; (ii) GDCdownload, enabling the bulk transfer of data from multiple organs simultaneously; and (iii) the GDCprepare function was employed to read the data and construct a SummarizedExperiment object. This object comprises one or more assays, with each assay represented by a matrix-like numeric object. Rows often represent genomic ranges of interest, while columns signify individual samples. The acquired data were subsequently saved in the form of an RDS file.

3.1.2 GTEx data

Regarding the **GTEx** database, data acquisition took place via programmatic download of the RNA-seq data from **GTEx** Analysis version 8 on the database's dedicated website [gte], accessed on 15 May 2023.

Gene read counts from the tissues of interest were downloaded from the following URLs (Table 2).

Table 2: Files downloaded from the GTEx portal.

Metabolic Type	Organ	URL
High	Brain	gene_reads_2017-06-05_v8_brain_cortex.gct.gz
	Liver	gene_reads_2017-06-05_v8_liver.gct.gz
	Kidney	gene_reads_2017-06-05_v8_kidney_medulla.gct.gz
	Kidney	gene_reads_2017-06-05_v8_kidney_cortex.gct.gz
Low	Bladder	gene_reads_2017-06-05_v8_bladder.gct.gz
	Colon	gene_reads_2017-06-05_v8_colon_sigmoid.gct.gz
	Colon	gene_reads_2017-06-05_v8_colon_transverse.gct.gz
	Skin	gene_reads_2017-06-05_v8_skin_sun_exposed_lower_leg.gct.gz
	Skin	gene_reads_2017-06-05_v8_skin_not_sun_exposed_suprapubic.gct.gz

The data was stored as a list object for clarity and organization. The file names were programmatically shortened and modified to incorporate user-friendly labels, including the organ's name. The final object was saved in RDS format.

3.1.3 MitoCarta data

Finally, for MitoCarta, the procedure closely resembled that of **GTEx**, where genomic data of the human genome were obtained from the database's public website [Mit]. Data download was performed on May 15 2023. Subsequently, the three data sheets in the Excel file were individually saved in RDS format.

3.2 Mitocarta R package

In order to enhance interaction with the data derived from MitoCarta, an RDatapackage was developed (Figure 11). The MitoCarta R package is publicly available at the following GitHub address:

https://github.com/MitoProfiles/MitoCarta.

The three MitoCarta data tables were cleaned and wrangled. The first dataframe, labeled "A_Human _MitoCarta3" contains information pertaining to genes that code for proteins predicted to be targeted to the human mitochondrion. The second dataframe contains information concerning all genes present in the human genome"B_Human_All_Genes". Lastly, the third dataframe, denoted as "C_MitoPathways" lists all the currently known metabolic pathways active in the mitochondrion.

The data was validated and wrangled into an R-friendly format using the R script "mitocarta.R" con-

tained in the package. Special care was taken to eliminate leading and trailing empty columns present in the Excel tables, which were causing serious errors when loading the data into memory. Additional errors were found in the MitoPathways table, specifically in the "Genes" and the "MitoPathway" columns: in the "Genes" column there was no fixed field separator between the individual gene names (spaces between commas had to be removed); and in the "MitoPathway" column, trailing underscores and commas had to be removed between pathway names to allow the identification of similar pathway names.

After resolving data structure issues, we generated documentation for the package using Roxygen. This documentation explains the data's source, details the information in each dataframe, and specifies the content in each column (Figure 11).

The MitoCarta R package can be installed using the following R code:

```
install.packages("remotes")
remotes::install\_github("MitoProfiles/MitoCarta")
```

The help page of the package can be accessed by running: `?mitocarta::mitocarta\_data` after installing it.

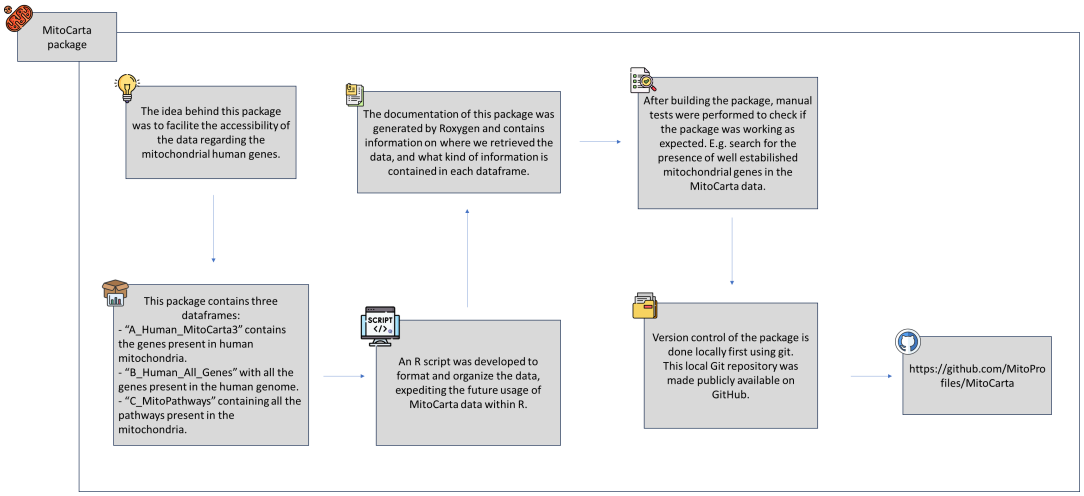


Figure 11: Description of the mitocarta package creation.

3.3 Data formatting and Compatibility between databases

After data download, the R scripts "2.format_data.Rmd" and "2.1.compare_tcga_gtex_gencode_v26_v36.Rmd" were developed for data preprocessing, cleaning, and assessment of the compatibility of gene expression data from **TCGA** and **GTEx** datasets. This first script defines a function (create_metadata) to

generate a metadata table containing sample IDs, cancer status, organ, and metabolic rate for each dataset downloaded. Next, the script processes **TCGA** and **GTEx** gene expression data separately, organizing it by organ and generating uniform datasets merged with the corresponding metadata. Finally, a Principal Component Analysis (**PCA**) is performed on the combined **TCGA** and **GTEx** data.

The **PCA** step is pivotal to verify if the data from different studies, and from the two different databases is comparable, and therefore suitable for drawing meaningful conclusions about differential expression. This involves determining if the RNA-seq protocols (both from the lab and the bioinformatics processing) are similar, particularly related to: (i) the strandedness of the sequencing library; (ii) the reference genome used; and (iii) the genome annotation version. If all three are similar, then, we can test if the overall gene counts do not show marked study biases.

The following information was retrieved for TCGA and for GTEx (Table 3):

Table 3: Comparison between data in TCGA and GTEx.

	TCGA	**GTEx** V8
RNA-seq library protocol	Illumina, non-stranded,	Illumina, non-stranded
Reference genome	GRCh38 reference genome	GRCh38 reference genome
Gene annotation version	GENCODE **v36**	GENCODE **v26**

Given the difference in genome annotation (in bold), we proceeded with comparing the two annotation files to decide if there were substantial differences between the genes present in both versions. We found that only 52 new gene identifiers were present in GENCODE v36 (not present in v26), and all common ones (5779 unique stable Ensembl identifiers) shared the exact same reference genome coordinates, rendering them fully comparable. These results can be inspected in the script file "2.1.compare_tcga_gtex_gencode_v26_v36.Rmd".

3.3.1 PCA

To visualize the variability in gene expression between both datasets, to discard possible study biases, a **PCA** was performed, using the built in R function `prcomp()`, centered and not scaled. It showed a good level of overlap between the two datasets, providing confidence to proceed with the planned differential expression analyses.

3.4 Differential Gene Expression

To conduct a differential gene expression analysis, three R packages were considered: limma [Ritchie et al., 2015], DESeq2 [Anders and Huber, 2010], and edgeR [Robinson et al., 2010]. EdgeR was the chosen package primarily because of its comprehensive normalization method - Trimmed Mean of M values (**TMM**), which uses a weighted trimmed mean of the log expression ratios between samples. This normalization accounts for differences in sequencing depth, **RNA** composition, and gene length, allowing for gene count comparisons between and within samples, as well as differential expression analysis [Robinson and Oshlack, 2010] (Table 1).

This step was performed using the script "3.diff_expression.Rmd", and the results were visualized with "3.1.diff_expression_visualization_v2.Rmd". Briefly, edgeR uses a negative binomial model to detect differentially expressed genes, while also modeling the biological and technical variability, and accounting for library size differences. Its approach involves estimating the dispersion parameters (which quantify the degree of variation), and applying a statistical test, such as the likelihood ratio test, to identify genes whose expression significantly differs between conditions [Love et al., 2014].

Three alternative linear models were considered, namely:

(i) The additive model:

```
~ 0 + metabolic rate + cancer status;
```

(ii) The interaction model:

```
~ 0 + metabolic rate:cancer status;
```

(iii) The additive with interaction model:

```
~ 0 + metabolic rate + cancer status + metabolic rate:cancer status.
```

After a brief comparison between models, the third (and most complete) model was chosen for analysis. The rationale behind this choice is the fact that this model is the one that best describes the prevailing understanding of cancer biology, specifically the alteration in cellular metabolism within cancer cells.

Since the initial question for this study is related to the profiles of gene expression in cancers from high metabolic rate organs, the analyzed contrast tested if the linear model coefficient "High metabolic rate in Cancer" is equal to zero. In other words, the null hypothesis for the hypothesis tests is that there are no expression changes in high metabolic-rate organs in cancer when compared to the low metabolic-rate organs in cancer, using as a reference the healthy (non-cancer) status. For further details, please see [Law et al., 2020]. Therefore, the significant differential expression (significant p-values) is related to the

metabolic type of the organs in cancer.

Data visualization was achieved via volcano plots (for all genes, and filtered only for mitochondrial ones); and boxplots for the top differentially expressed genes (also from the full set of genes, and only for the set of mitochondrial genes). All these visualizations are reproducible and openly available for detailed inspection at "3.1.diff_expression_visualization_v2.Rmd". An overview of the differential expression analysis workflow is presented in Figure 12.

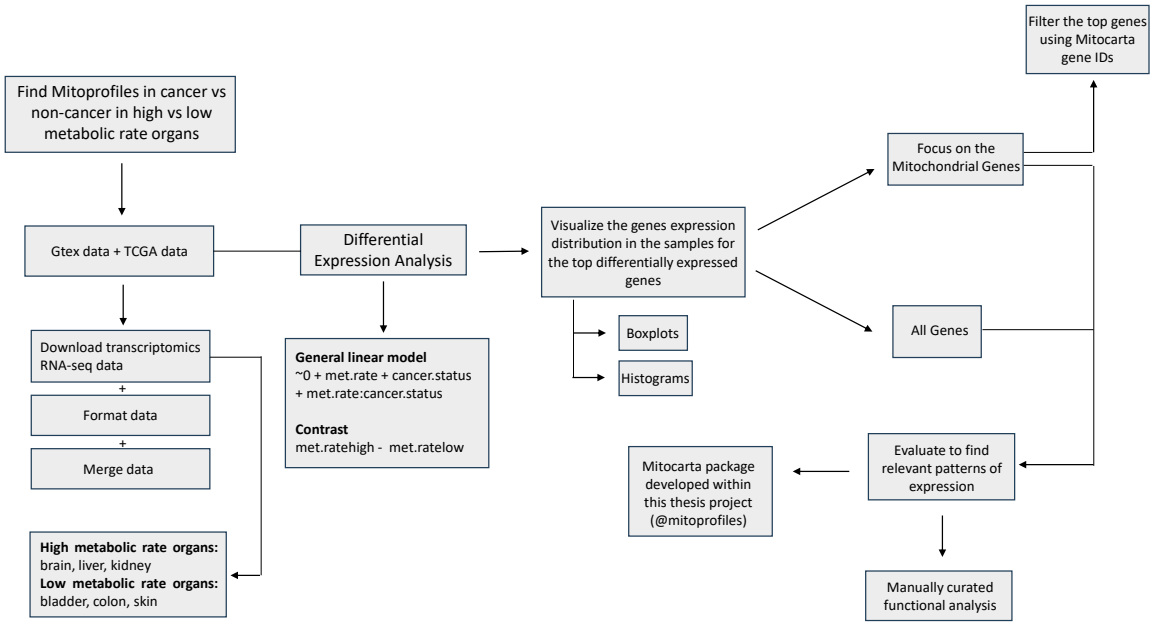


Figure 12: Detailed workflow for this project - Differential Expression.

3.5 Functional enrichment and Functional interactions

The top 500 differentially expressed genes were subject to a functional enrichment analysis using the gprofiler2 package [Reimand et al., 2007]. Only terms from the CORUM database (comprehensive resource of mammalian protein complexes) [Tsitsiridis et al., 2023] showed significant enrichment (figure shown in Annex B.2). This analysis was terminated since there was no significant functional enrichment for any of the other ontologies tested (GO Molecular Function, GO Cellular Component, GO Biological Process, KEGG, Reactome, TRANSFAC, miRTarBase, Human Protein Atlas, Human Phenotype, and WikiPathways). The R code used for this analysis can be inspected at "4.functional_enrichment.Rmd".

3.5.1 STRING Functional interaction networks

Despite the absence of statistically significant functional enrichment, we aimed to visualize potential functional relationships among the top differentially expressed genes. To achieve this, we used the STRING online platform [str] with default settings. The process involved inputting the gene IDs for each **DEGs** set into STRING's multiple protein search function. The gene IDs are converted to protein IDs, and the resulting network shows the known and predicted functional interactions between the proteins. The images were manually downloaded in PNG format.

3.6 Clustering

To find groups of genes with similar expression behavior across the different organs in cancer and non-cancer, we selected the 91 differentially expressed genes encoding mitochondrial proteins and applied two alternative clustering methodologies: hierarchical clustering, and fuzzy (soft) clustering. For both techniques, normalized data (**CPM** counts per million) was used (instead of the raw counts matrix), and the Manhattan distance method was used to calculate the distance matrix. The R code used for the clustering analysis can be fully inspected in "5.clustering.Rmd". The clustering analysis workflow is presented in Figure 13.

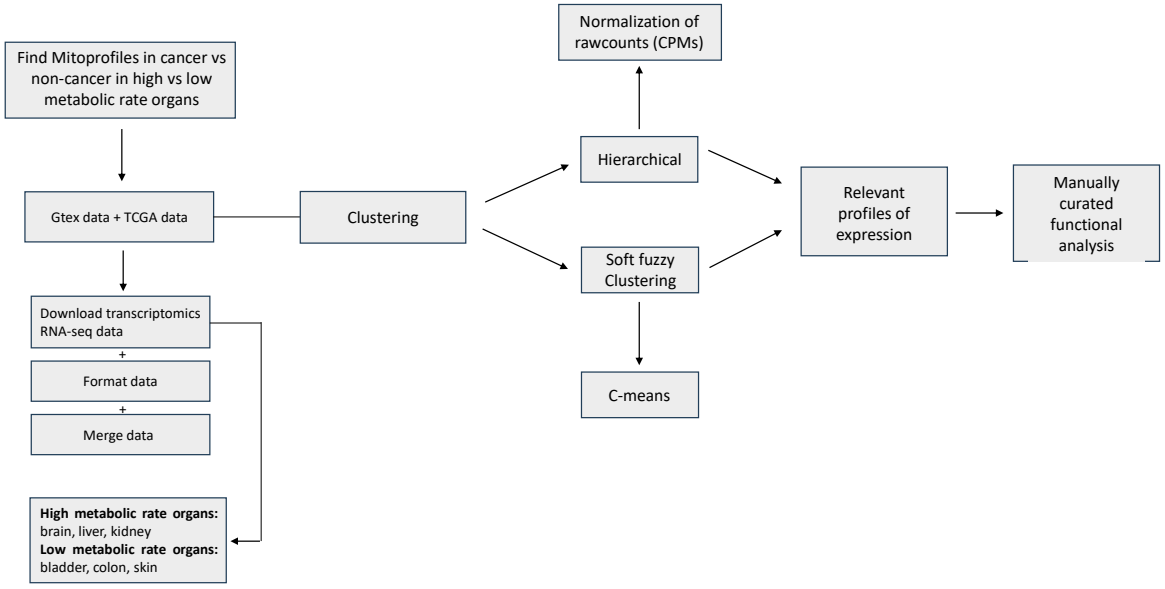


Figure 13: Detailed workflow for this project - Clustering.

Hierarchical clustering organizes data points into a dendrogram (a tree-like structure), based on their similarities, allowing the easy visualization of clusters. The choice of linkage method (e.g., complete, single, average) can greatly influence the clustering outcome, leading us to apply all three methods to our data, using the `hclust()` function from base R. The results from the group average linkage method were chosen for discussion given the overlap with the results from the gene expression data shown in the heatmap drawn. The alternative dendrograms are presented in Annex B.3.

Fuzzy clustering is a soft clustering technique that extends hard clustering methods allowing the same data point to belong to more than one cluster. Therefore, fuzzy clustering assigns degrees of membership, expressed as probabilities, to each data point across multiple clusters, allowing the same gene to belong to more than one cluster of gene expression. This approach is implemented in the function `cmeans()` from the R package e1071 that we used for this study. We requested 20 clusters, with a fuzzification parameter m = 1.25 (m determines the degree of fuzziness in the membership assignments of data points to clusters - smaller m values make the memberships more distinct and data points are more sharply assigned to a single cluster). The different clusters represent the alternative mitochondrial gene expression profiles in cancer from high and low metabolic rate organs.

3.7 English language editing

In compliance with international standards for transparency regarding the usage of AI technology in scientific studies, this thesis made use of the Large Language Model (LLM) ChatGPT 3.5, developed by OpenAI, exclusively for grammar and spelling correction, ensuring that the text maintained high standards of clarity. While ChatGPT 3.5 was used for linguistic enhancements, it did not influence or modify in any way the scientific content of the research data and findings reported. It was not used in any other part of the work developed.

Chapter 4

Results and Discussion

4.1 PCA analysis to compare GTEx and TCGA data

To compare the different datasets used from **TCGA** and **GTEx**, a Principal Component Analysis (**PCA**) was performed (Figure 14). **PCA** reduces the dimensions of a multivariate data table into a condensed set of variables known as summary indices, facilitating the observation of trends, shifts, clusters, and outliers. This analysis provides an overview of the relationships amongst observations and variables, as well as inter-variable relationships.

The overall **PCA** analysis shows a significant overlap between the samples from both databases (**TCGA** and **GTEx**), revealing a small variance between the datasets, making them broadly comparable (Figure 14 A). The few visible outliers (less than 25 samples) are all provenient from healthy samples (**GTEx**) but from different organs, namely from the colon, kidney, liver, and brain. To expand on this observation, and ensure that there are no significant asymmetries regarding the variability of the expression values from samples from the same organ coming from the two databases, the data were colored according to the organ (Figure 14 B).

Here, it is shown that regardless of the dataset (**GTEx** or **TCGA**), samples from the same organ exhibit suitably similar behavior, as indicated by the prominent clusters of each color, with few samples presenting outlier behavior. These outliers, however, were not removed given their limited number and their association with the **GTEx** dataset, which has fewer samples for analysis.

These results provide confidence that the data is indeed comparable, and therefore amenable for the next analysis steps required for this study.

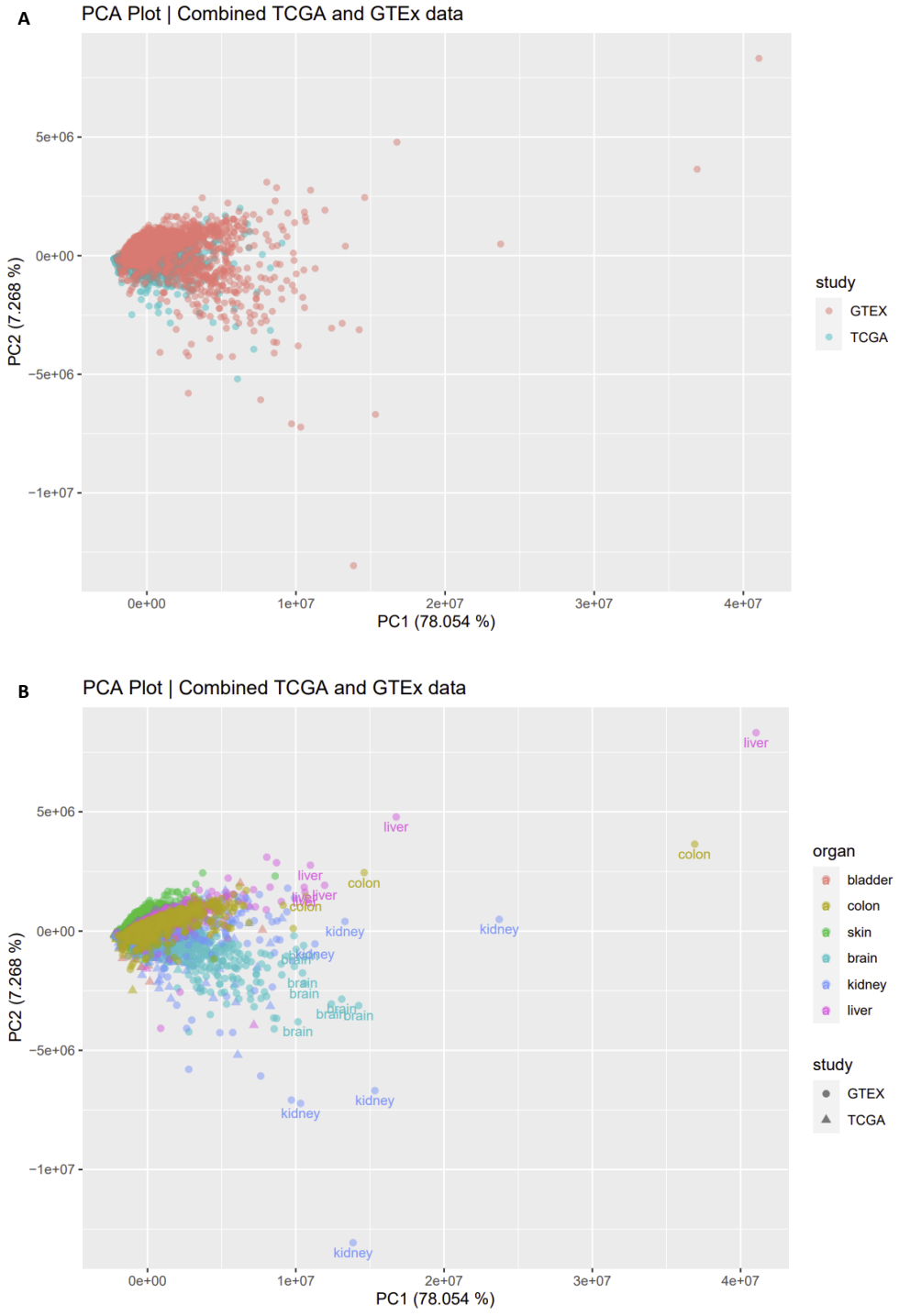


Figure 14: **PCA** analysis. **A)** This visualization shows the variability between the **GTEx** dataset (red) and the **TCGA** dataset (green). **B)** This visualization shows the variability between organs. Samples with PC1 > 1e+07 are identified with an organ label.

4.2 Differential Expression Analysis

To discover the genes that might be associated with cancer in organs with alternative metabolic types (high or low metabolic rate), a differential gene expression analysis was performed. Genes exhibiting differential expression between conditions can provide valuable insights into the biological processes active in the two conditions under study (high versus low metabolic rate in cancer).

In total, 79806 unique transcripts (from now on referred to as genes) were fitted using the additive with interaction model:

```
~ 0 + metabolic rate + cancer status + metabolic rate:cancer status
```

The evaluated contrast was "High metabolic rate in Cancer", meaning that the significant differential expression (significant p-values) pertains to changes observed in cancer from high versus low metabolic rate organs when compared to the reference status (i.e. healthy non-cancer samples).

This analysis outputs five calculated metrics: Log2 Fold Change (**logFC**), Log Counts per Million (**logCPM**), Likelihood Ratio (**LR**), p-value, and False Discovery Rate (**FDR**). A total of 4142 genes displayed significant fold change (p-value < 0.01, and **FDR** < 0.01).

4.2.1 Differential Expression visualization: Volcano plots

To visualize the global patterns of differential expression, two volcano plots were generated. These are scatterplots that display statistical significance (-log10 p-value) versus the magnitude of change (log2 Fold Change). They allow for quick visual identification of genes with both significant statistical changes and substantial fold changes.

Figure 15 A includes all genes, and Figure 15 B displays only the genes encoding mitochondrial proteins (i.e. proteins that are predicted to be targeted to the mitochondrion according to MitoCarta3.0 [Rath et al., 2021]). Upregulated genes (**FC** > 1) are identified by the purple color, downregulated genes (**FC** < -1) are depicted in green, and genes that do not exhibit significant Fold Change (p-value < 0.01) are represented in gray.

Figure 15 A, shows numerous genes with a very low p-value, indicating highly significant differential expression, some even present on the boundary of the x-axis corresponding to p-values of zero (i.e. -log10(0) = Infinite).

Given the substantial number of statistically significant genes, we include the gene symbol on genes with a log2 fold change that markedly deviated from the majority, i.e. log2 FC > 7 and log2 FC < -7. Only 13 genes pass this filtering criteria: four downregulated (**MT-TM**, **LINC01833**, **CHN2-AS1**, and **LHFPL3-AS1**) and nine upregulated (ENSG00000236740, **ADGRF2**, **LINC00462**, ENSG00000267774, **WFDC5**, **LINC02247**, ENSG00000273664, **KRT85**, and **TTC21B-AS1**).

Regarding the mitochondrial plot (Figure 15 B), we observe that the majority of genes also exhibit highly significant differential expression, however with more variable fold changes, leading to the choice of 15 genes with log2 fold change higher than 1, and lower than -1: 11 downregulated (**FDX2**, **RPUSD3**, **AIFM3**, **MT-CO2**, **CMPK2**, **MT-ATP6**, **MT-ND4L**, **MT-ND4**, **MT-ATP8**, **HDHD3**, and **GPAT2**), and four upregulated (**ACSM1**, **PRODH**, **BCL2L10**, and **ACSM5**).

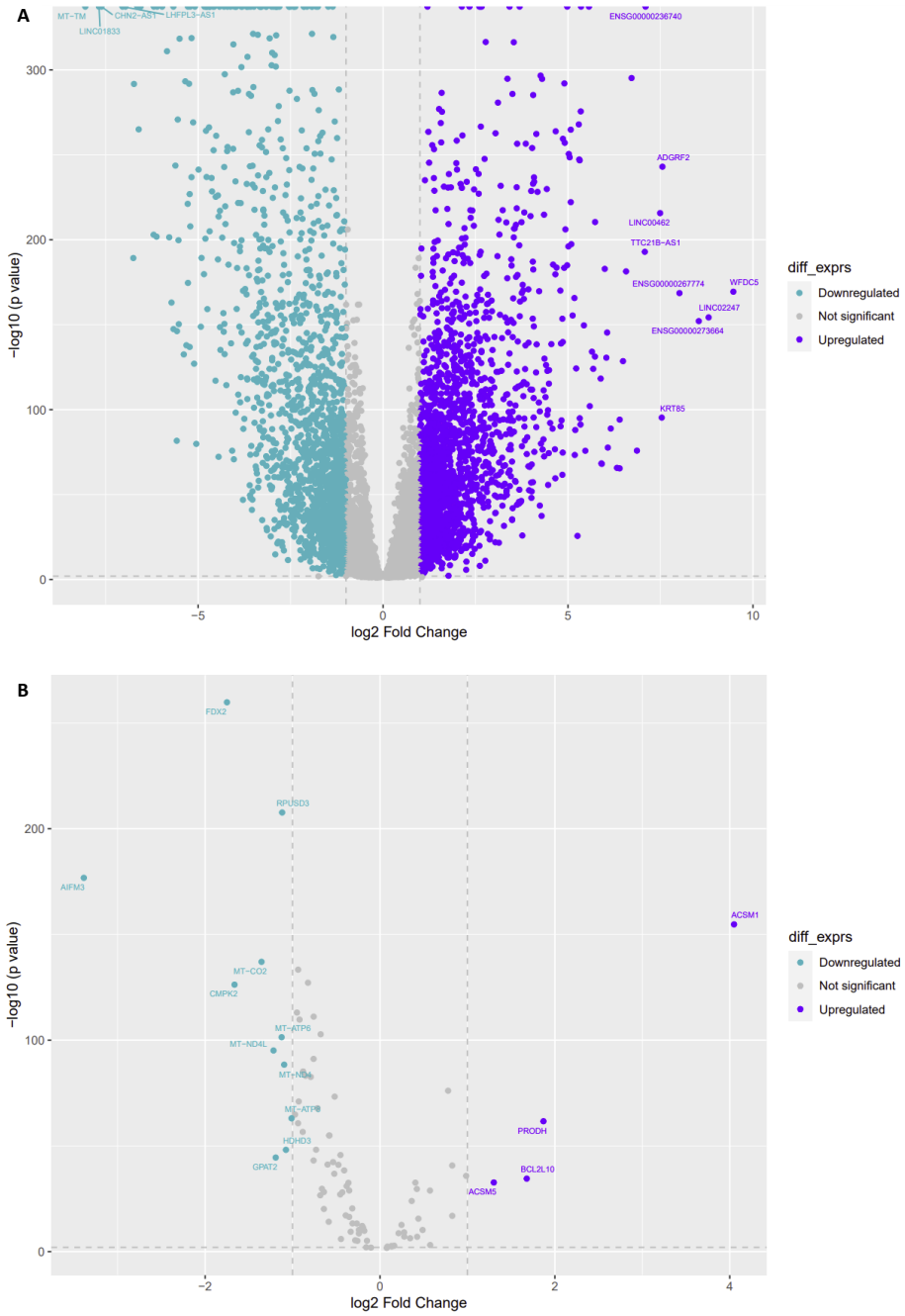


Figure 15: Differential expression visualization using volcano plots. **A)** Results for the global gene set analysis. Genes with a log fold change < -7 and > 7 are individually labeled. **B)** Results for the mitochondrial gene set analysis. Genes with a log fold change < -1 and > 1 are individually labeled. Downregulated genes are highlighted in green, upregulated genes in purple, and in gray are the genes with smaller log Fold Change.

4.2.2 Top Differentially Expressed genes

A selection of the top 24 differentially expressed genes (12 up and 12 downregulated genes) was filtered and further analyzed. These results are displayed in two tables: Table 4 containing the top 24 genes from the full set of **DEGs**, and Table 5 showing the top 24 differentially expressed mitochondrial genes.

Table 4: Top 24 differentially expressed genes. The first 12 genes are downregulated (sorted from lowest to highest log fold change). The next 12 genes are upregulated (sorted from highest to lowest log fold change).

gene_id	gene_symbol	logFC	logCPM	PValue	FDR	Functional description	Molecular activity
ENSG00000210112	MT-TM	-8.05	0.904	0	0	Mt tRNA	mitochondrially encoded tRNA-Met
ENSG00000259439	LINC01833	-7.68	2.025	0	0	long intergenic non-protein coding RNA 1833	LncRNA
ENSG00000235669	CHN2-AS1	-7.61	0.298	0	0	CHN2 antisense RNA 1	LncRNA
ENSG00000226869	LHFPL3-AS1	-7.04	5.504	0	0	LHFPL3 antisense RNA 1	LncRNA
ENSG00000256817	TPT1P12	-6.96	-0.454	0	0	TPT1 pseudogene 12	Processed pseudogene
ENSG00000254547	ANKRD33BP6	-6.75	0.613	2.19E-191	6.07E-190	ANKRD33B pseudogene 6	Unprocessed pseudogene
ENSG00000204544	MUC21	-6.73	2.966	2.20E-294	2.05E-292	Mucin-21	negative regulation of cell adhesion
ENSG00000210077	MT-TV	-6.64	0.085	0	0	Mt tRNA	mitochondrially encoded tRNA-Val
ENSG00000283178	ENSG00000283178	-6.61	-0.210	1.45E-267	1.07E-265	chromosome X open reading frame 49	Unprocessed pseudogene
ENSG00000124092	CTCFL	-6.59	1.827	0	0	Transcriptional repressor CTCFL	C2H2 zinc finger transcription factor
ENSG00000139800	ZIC5	-6.56	2.903	0	0	Zinc finger protein ZIC 5	C2H2 zinc finger transcription factor
ENSG00000227300	KRT16P2	-6.20	2.415	3.47E-205	1.14E-203	keratin 16 pseudogene 2	Transcribed unprocessed pseudogene
ENSG00000175121	WFDC5	9.47	7.499	1.73E-171	3.91E-170	WAP four-disulfide core domain protein 5	protease inhibitor
ENSG00000223729	LINC02247	8.80	1.096	2.88E-156	5.20E-155	long intergenic non-protein coding RNA 2247	LncRNA
ENSG00000273664	ENSG00000273664	8.54	0.529	4.78E-154	8.30E-153	novel transcript	LncRNA
ENSG00000267774	ENSG00000267774	8.01	1.656	1.16E-170	2.58E-169	novel transcript, antisense to CCBE1	LncRNA
ENSG00000164393	ADGRF2	7.55	3.904	1.92E-245	1.03E-243	adhesion G protein-coupled receptor F2	Transcribed unitary pseudogene
ENSG00000135443	KRT85	7.54	7.623	6.33E-97	5.09E-96	Keratin, type II cuticular Hb5	structural constituent of skin epidermis
ENSG00000233610	LINC00462	7.49	3.275	5.44E-218	2.07E-216	long intergenic non-protein coding RNA 462	LncRNA
ENSG00000236740	ENSG00000236740	7.10	4.022	0	0	muscular LMNA interacting protein	transcriptional regulator of the myogenic program
ENSG00000224490	TTC21B-AS1	7.08	3.543	4.50E-195	1.30E-193	TTC21B antisense RNA 1	LncRNA
ENSG00000186393	KRT26	6.86	2.362	2.36E-77	1.41E-76	Keratin, type I cytoskeletal 26	intermediate filament
ENSG00000246740	PLA2G4E-AS1	6.72	3.608	6.72E-298	6.64E-296	PLA2G4E antisense RNA 1	LncRNA
ENSG00000251491	OR7E28P	6.57	0.975	1.69E-183	4.27E-182	olfactory receptor family 7 subfamily E member 28 pseudogene	Transcribed unprocessed pseudogene

Table 5: Top 24 mitochondrial differentially expressed genes. The first 12 genes are downregulated (sorted from lowest to highest log fold change). The next 12 genes are upregulated (sorted from highest to lowest log fold change)

gene_id	gene_symbol	logFC	logCPM	PValue	FDR	Functional description	Molecular activity
ENSG00000183773	AIFM3	-3.39	6.103	1.63E-177	3.86E-176	Apoptosis-inducing factor 3	oxidoreductase
ENSG00000267673	FDX2	-1.75	4.481	1.54E-260	9.95E-259	Ferredoxin-2, mitochondrial	oxidoreductase
ENSG00000134326	CMPK2	-1.66	5.485	5.47E-127	6.70E-126	UMP-CMP kinase 2, mitochondrial	nucleotide kinase
ENSG00000198712	MT-CO2	-1.35	15.816	8.61E-138	1.21E-136	Cytochrome c oxidase subunit 2	oxidoreductase
ENSG00000212907	MT-ND4L	-1.22	13.082	7.55E-96	5.96E-95	NADH-ubiquinone oxidoreductase chain 4L	oxidoreductase
ENSG00000186281	GPAT2	-1.19	3.790	3.23E-45	1.12E-44	Glycerol-3-phosphate acyltransferase 2, mitochondrial	acyltransferase
ENSG00000198899	MT-ATP6	-1.12	15.326	3.96E-102	3.46E-101	ATP synthase subunit a	ATP synthase
ENSG00000156990	RPUSD3	-1.12	7.289	2.05E-208	7.00E-207	Mitochondrial mRNA pseudouridine synthase RPUSD3	RNA processing factor
ENSG00000198886	MT-ND4	-1.10	16.707	3.84E-89	2.72E-88	NADH-ubiquinone oxidoreductase chain 4	oxidoreductase
ENSG00000119431	HDHD3	-1.08	7.874	7.05E-49	2.61E-48	Haloacid dehalogenase-like hydrolase domain-containing protein 3	---
ENSG00000228253	MT-ATP8	-1.01	12.275	9.05E-64	4.32E-63	ATP synthase protein 8	ATP synthase
ENSG00000198727	MT-CYB	-0.97	15.746	1.51E-65	7.39E-65	Cytochrome b	---
ENSG00000166743	ACSM1	4.05	5.119	1.69E-155	3.02E-154	Acyl-coenzyme A synthetase ACSM1, mitochondrial	ligase
ENSG00000100033	PRODH	1.87	6.602	2.19E-62	1.02E-61	Proline dehydrogenase 1, mitochondrial	oxidase
ENSG00000137875	BCL2L10	1.68	3.972	3.00E-35	8.46E-35	Bcl-2-like protein 10;BCL2L10;PTN002477644;orthologs	---
ENSG00000183549	ACSM5	1.30	7.199	2.00E-33	5.48E-33	Acyl-coenzyme A synthetase ACSM5, mitochondrial	ligase
ENSG00000076555	ACACB	0.99	8.720	1.48E-36	4.30E-36	Acetyl-CoA carboxylase 2	---
ENSG00000198754	OXCT2	0.83	1.672	1.21E-17	2.34E-17	Succinyl-CoA:3-ketoacid coenzyme A transferase 2, mitochondrial	transferase
ENSG00000176171	BNIP3	0.83	8.911	1.84E-41	5.97E-41	BCL2_adenovirus E1B 19 kDa protein-interacting protein 3	---
ENSG00000164983	TMEM65	0.78	7.358	8.03E-77	4.74E-76	Transmembrane protein 65	---
ENSG00000066813	ACSM2B	0.57	8.859	7.65E-04	9.67E-04	Acyl-coenzyme A synthetase ACSM2B, mitochondrial	ligase
ENSG00000089163	SIRT4	0.57	3.279	1.32E-29	3.33E-29	NAD-dependent protein lipoamidase sirtuin-4, mitochondrial	---
ENSG00000134278	SPIRE1	0.49	7.357	5.83E-11	9.47E-11	Protein spire homolog 1	actin or actin-binding cytoskeletal protein
ENSG00000156026	MCU	0.44	7.360	2.63E-16	4.93E-16	Calcium uniporter protein, mitochondrial	---

4.2.3 Differential Expression visualization: Boxplots

Boxplots were also generated for each of these 48 top genes: 24 from the full gene set (Figure 16), and 24 from the mitochondrial subset (Figure 17). These plots allow the visualization of the patterns of expression of each of these top genes in the different conditions under study.

The distribution of the gene counts is grouped per metabolic type and cancer status. Thus, samples from **GTEx** (healthy tissue) are shown with striped boxes, and samples from **TCGA** (cancer tissue) are shown with dotted boxes. For each of these groups, low metabolic rate organs (bladder, colon, and skin) are represented in red, and high metabolic rate organs (brain, kidney, and liver) are represented in green.

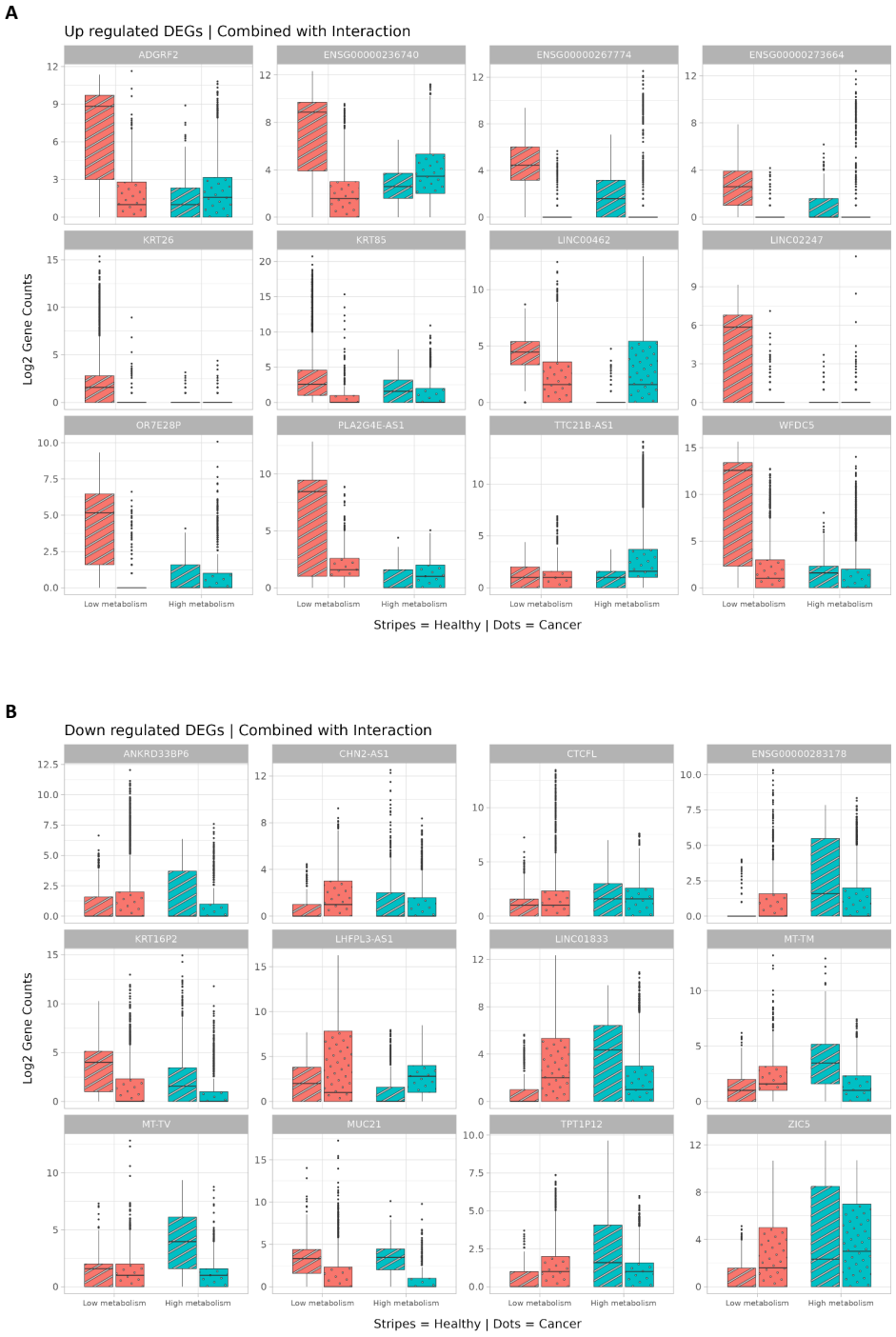


Figure 16: Boxplots of the 24 top differentially expressed genes from the **global gene set**. (**A**) Upregulated genes. (**B**) Downregulated genes. Each plot represents the expression of one gene, where the color distinguishes low metabolism (in red) from high metabolism (in green), and the pattern distinguishes between healthy (stripes) and cancer (dots).

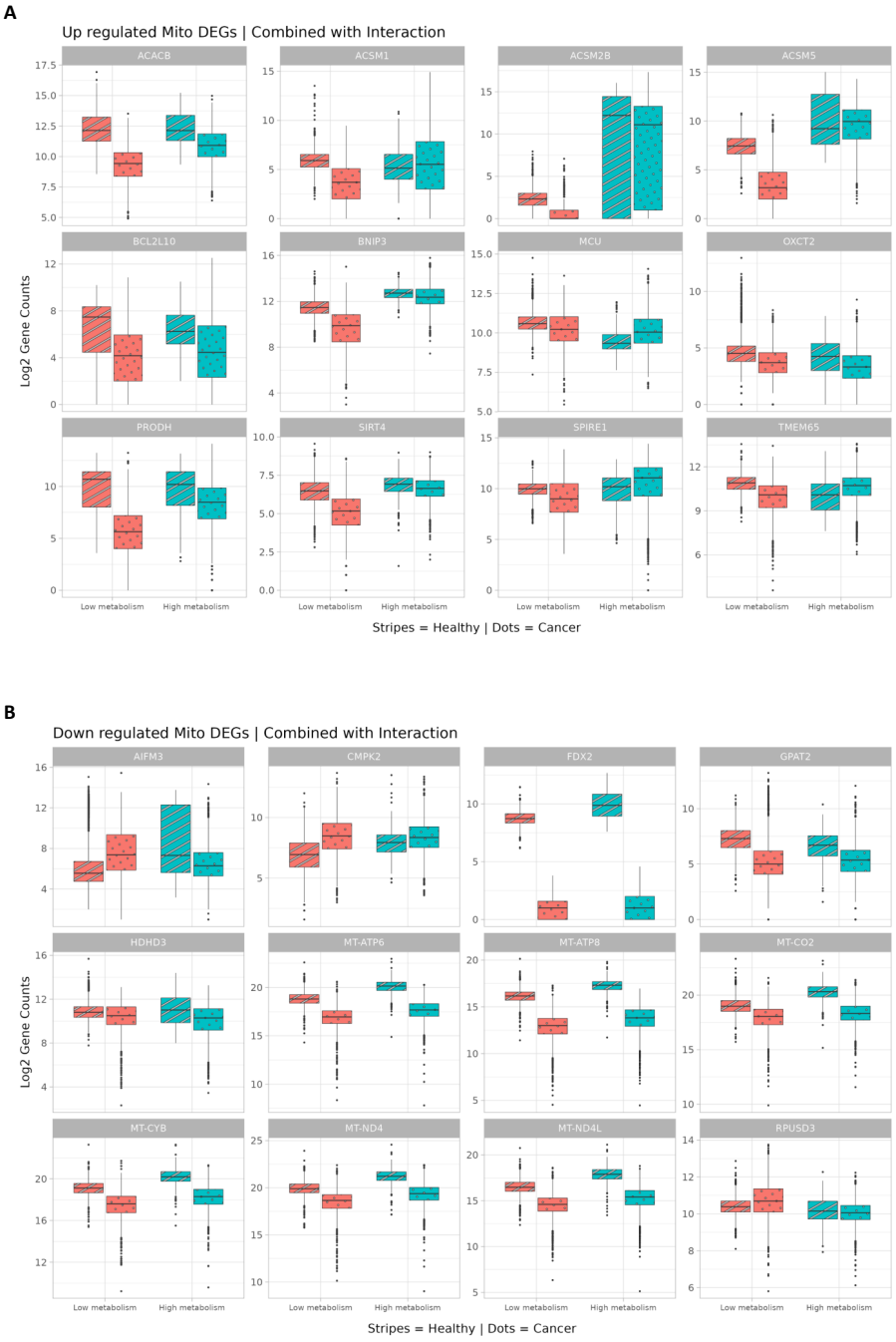


Figure 17: Boxplots of the 24 top differentially expressed genes from the **mitochondrial gene set**. (**A**) Upregulated genes. (**B**) Downregulated genes. Each plot represents the expression of one gene, where the color distinguishes low metabolism (in red) from high metabolism (in green), and the pattern distinguishes between healthy (stripes) and cancer (dots).

4.2.4 Differential Expression Discussion

The distributions observed for the differentially expressed genes reveal that the expression differences are more pronounced when comparing values between healthy and cancer samples than when comparing alternative metabolic rates. This suggests that the cancer status of a cell predominantly accounts for the majority of the signal present in gene expression changes. In other words, gene expression changes are primarily attributed to whether a cell is healthy or cancerous, regardless of its organ metabolic rates (Figures 16 and 17).

This behavior is most notorious in the mitochondrial differentially expressed genes (**DEGs**), where the asymmetry between striped and dotted boxes is particularly pronounced. The gene **FDX2** [ens, g] exhibits the most extreme pattern, making it a prime candidate for future experimental validation in the laboratory.

Global Differential Expression

The list of top differentially expressed genes includes many genes for which there is no functional annotation, specifically nine long non-coding **RNA**s and six pseudogenes without recognized regulatory elements. Since pseudogenes are traditionally considered non-functional **DNA** segments, their functionality remains uncharacterized. However, the development of high-throughput sequencing technology has facilitated the discovery that these gene remnants can regulate gene expression at different levels and play important roles in tumorigenesis [Hu et al., 2018]. For example, the pseudogene **LHFPL3-AS1** [ens, j] has recently been associated with the upregulation of **BCL2**, a protein that suppresses apoptosis [Zhang et al., 2020], a characteristic hallmark of cancer. Therefore, instead of removing these genes from subsequent analysis, we decided to report these findings, since these novel genomic features could be valuable candidates for future collaborations for experimental validation *in vitro*.

The remaining genes are involved in a diverse range of functions, including skin development (keratinization involving **KRT26** [ens, h], and **KRT85** [ens, i]), negative regulation of cell adhesion (**MUC21** [ens, m]), transcription factors (**CTCFL** [ens, f], and **ZIC5** [ens, p]), protease inhibition (**WFDC5** [ens, o]), and also mitochondrial Aminoacyl-tRNA biosynthesis (**MT-TM** [ens, k], and **MT-TV** [ens, l]). This diversity is most likely due to the fact that the samples used are from a diverse set of organs which will have some functions performed by organ-specific proteins.

From the top downregulated gene set, two genes, Mucin 21 (**MUC21**) [ens, m] and the long non-coding RNA **LHFPL3-AS1** [ens, j], are particularly interesting. Mucin 21 is a molecule previously demonstrated to inhibit cell-cell and cell-matrix adhesion, particularly relevant in lung carcinomas [Yoshimoto et al., 2019]. Mucins are heavily glycosylated proteins secreted by various cell types and play diverse roles in cancer progression, including cell adhesion, epithelial-mesenchymal transition, cell signaling, and influencing the tumor microenvironment [Liu et al., 2022]. While cancer cells with reduced cell adhesion are associated with cancer spreading through vessels or alveolar spaces in lung [Yoshimoto et al., 2019], the protective role of this gene in the intestinal tract suggests that its decreased expression might be a general feature of cancer, irrespective of its primary tissue location.

Although **LHFPL3-AS1** is annotated as a long-non-coding RNA (as mentioned before), it has been linked to the regulation of the **BCL2** a well known mediator of the suppression of cell apoptosis [Zhang et al., 2020]. Interestingly, **BCL2L10** (ENSG00000137875) [ens, e] which is a member of the **BCL2** protein located in the mitochondrion [Zhang et al., 2001] is upregulated in our analysis, therefore validating our results, since we observe a downregulation of the mediator of the apoptosis suppressor **LHFPL3-AS1** and an upregulation of the apoptosis suppressor **BCL2L10**.

Overall, this finding show that cell apoptosis is a recurring theme in our dataset, in line with the well recognized fact that cancer cells tend to silence genes that promote cell death, and overexpress genes that confer growth advantage [Hanahan and Weinberg, 2011].

Mitochondrial Differential Expression

In the analysis of differential expression among mitochondrial genes, the first notable observation is that six genes are encoded within mitochondrial DNA (**MT-CO2**, **MT-ND4L**, **MT-ATP6**, **MT-ND4**, **MT-ATP8**, and **MT-CYB**). Five of these genes are clustered closely together in the volcano plot (Figure 15 B), and a detailed examination of the mitochondrial boxplots (Figure 17) reveals that they all belong to the down-regulated set (which searched for metabolic rate differences), and all exhibit a similar pattern of higher expression in healthy tissue compared to cancer (i.e. they are also downregulated in cancer). This finding aligns with the mitochondrial dysfunction often associated with cancer development and progression. This dysfunction arises, not only due to the accumulation of **mtDNA** mutations caused mostly by the accumulation of **ROS** as a stress response in cancer [can], but can also be due to the dysregulation of the mitochondrial retrograde signaling pathways, mediated by **ROS**, and calcium, that can lead to the downregulation of **mtDNA** gene expression [Wallace, 2012].

Another notable observation concerning the genes encoded by **mtDNA** is that their distribution is narrower than that observed for nuclear-encoded mitochondrial proteins. This phenomenon is likely attributable to two factors: (i) **mtDNA** is circular and thus similar to a bacterial operon, leading to an almost stoichiometric expression of its genes [Pearce et al., 2017]; and (ii) the number of **mtDNA** molecules in each cell is very large, leading to a narrow gene expression distribution due to the central limit theorem (**CLT**). The **CLT** states that the sum of many independent and identically distributed random variables approaches a normal distribution. In the context of **mtDNA**, this means that the expression fluctuations from individual mitochondria are averaged out, and the resulting value approaches the mean expression of the entire population [Wolff et al., 2017].

Since mitochondria are known as the powerhouses of the cell, it might seem surprising that genes contributing to energy production are underexpressed in cancer cells. However, cancer cells can suppress programmed cell death [Hanahan and Weinberg, 2011], and reach a state in their proliferation cycle where they grow in microenvironments with increased **ROS** species levels, potentially compromising cell function [Hanahan and Weinberg, 2011]. As a result, the downregulation of mitochondrial genes related to energy production might represent an adaptation to an environment less conducive to aerobic respiration. In this context, it becomes less beneficial to have numerous active mitochondrial complexes if they are not being

efficiently used, especially when substrates are needed for other essential functions. This hypothesis gains support from the observed increase in the expression of genes, such as **ACSM1**, **ACSM5**, and **PRODH** that help maintain mitochondrial stability, and ensure **ATP** production even under extreme conditions [Bender et al., 2005].

ACSM1 [ens, a] and **ACSM5** [ens, c] encode acyl-CoA synthetases which are catalyzers of the activation of fatty acids by CoA to produce acyl-CoA, the first step in fatty acid metabolism [Vessey et al., 1999], in order to meet rapidly the demand for high quantities of energy. The increase seen in the expression of these genes can boost intracellular acetyl-CoA levels, which are subsequently used in the **TCA** cycle to produce **ATP** for processes that help the survival of cancer cells under metabolic stress conditions [Tang et al., 2018].

PRODH [ens, n] codes for a proline dehydrogenase which appears essential for maintaining normal mitochondrial function, **ATP** levels, and redox balance, especially in hypoxic conditions [Bender et al., 2005] that arise when fast growing solid tumors become large and poorly vascularized.

Conversely, **AIFM3**, the Apoptosis Inducing Factor Mitochondria Associated 3 gene [ens, d] is encoded in the nucleus and acts as an inducer of cellular apoptosis in the mitochondrion, which is in line with its downregulation in cancer.

Close examination of the expression profiles in Figure 17, shows one particular gene that stands out: **FDX2** which presents a clear downregulation in cancer samples compared to the control samples.

The functional annotation for **FDX2**, as deposited in five widely-used genomics databases, shows that this protein is associated with mitochondrial myopathy, but does not present a strong link with cancer: Ensembl [ens, g]; UniProt [uni]; NCBI [nih]; GeneCards [gen]; and ProteinAtlas [pro].

Nonetheless, given the known functions of **FDX2**, it is tempting to speculate about its potential role in cancer. **FDX2** is a mitochondrial protein responsible for electron transfer from **NADPH** via **FDXR**, playing a crucial role in iron-sulfur protein biogenesis (also termed Fe/S cluster biogenesis) (Figure 18) [Shi et al., 2012].

Iron-sulfur clusters, which are essential inorganic protein cofactors, have a wide-ranging impact on cellular processes, including **DNA** repair, cell cycle regulation, metabolism, and oxidative respiration [Petronek et al., 2021]; [Shi et al., 2012]; [Shi et al., 2021]; [Braymer and Lill, 2017]. The biogenesis of these clusters is a complex, multistep process, involving catalyzed protein-protein interactions and conformational changes [Maio and Rouault, 2015]; [Rouault, 2012]; [Maio and Rouault, 2015]. These steps include: (i) Cluster synthesis: Fe/S clusters are synthesized on a scaffold protein, which is typically an IscU-like protein in bacteria or an **ISCU** protein in eukaryotes; (ii) Cluster transfer to target proteins: Fe/S

clusters are transferred from the scaffold protein to target proteins by a set of specialized chaperones and transfer proteins; and (iii) Cluster insertion into target proteins: Fe/S clusters are inserted into target proteins by a set of specialized Fe/S cluster insertion proteins (Figure 18) [Maio and Rouault, 2015]; [Rouault, 2012].

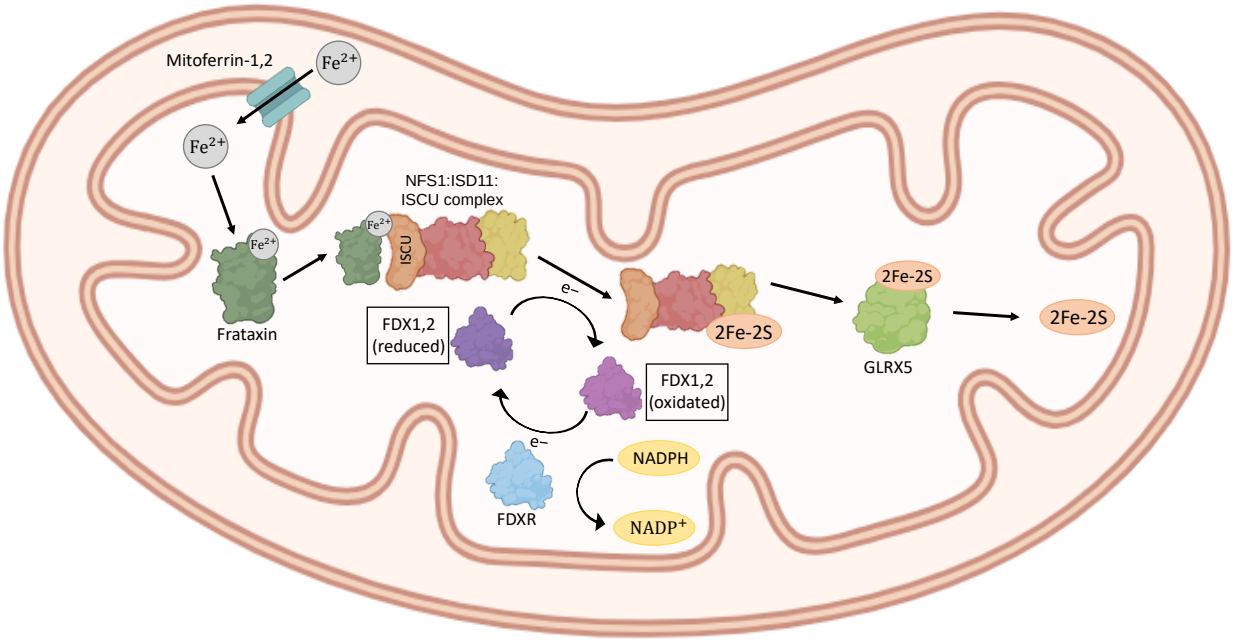


Figure 18: Iron-sulfur cluster biogenesis. Ferrous iron is transported across the inner mitochondrial membrane into the mitochondrial matrix by Mitoferrin-1 and Mitoferrin-2. Then frataxin binds ferrous iron in the mitochondrial matrix. The cysteine desulfurase NFS1, in a subcomplex with ISD11, provides the sulfur by converting cysteine into alanine and forming a persulfide which is used for cluster formation on ISCU. Interaction between NFS1 and ISD11 is necessary for desulfurase activity. Frataxin binds to the complex NFS1:ISD11:ISCU and is proposed to function as an iron donor to ISCU or as an allosteric switch that activates sulfur transfer and Fe-S cluster assembly. Cluster formation also involves the electron transfer chain ferredoxin reductase and ferredoxins 1 and 2. ISCU initially forms clusters containing 2 iron atoms and 2 sulfur atoms ([2Fe-2S] clusters). They are released by the monothiol glutaredoxin GLRX5 and used for assembly of [2Fe-2S] proteins.

In eukaryotes, iron-sulfur cluster biogenesis takes place in both the cytosol/nucleus and mitochondria. The cytosolic pathway involves the cytosolic iron-sulfur assembly (**CIA**) machinery, while the mitochondrial pathway involves the iron-sulfur cluster (**ISC**) machinery. The **CIA** machinery is responsible for the biogenesis of [2Fe-2S] and [4Fe-4S] clusters in the cytosol and nucleus, while the **ISC** machinery is responsible

for the biogenesis of [2Fe-2S], [4Fe-4S], and [2Fe-3S] clusters in mitochondria [Shi et al., 2021]; [Rouault, 2012]; [Braymer and Lill, 2017].

Disruption of iron-sulfur cluster biogenesis due to **FDX2** silencing can lead to oxidative stress and **DNA** damage, both recognized hallmarks of cancer [Hanahan, 2022]. Furthermore, impaired Fe/S cluster biogenesis may result in mitochondrial dysfunction which is common in cancer cells and is associated with altered metabolism, increased reactive oxygen species (**ROS**) production, and resistance to apoptosis [Petronek et al., 2021].

On the other hand, **FDX2** silencing may sensitize cancer cells to chemotherapy. For example, silencing of **GLRX5**, which is involved in iron-sulfur cluster biogenesis, can reverse cisplatin resistance and enhance the induction of ferroptosis in head and neck cancer cells [Petronek et al., 2021]. Similarly, **FDX2** silencing inhibited colony formation and reduced the activity of **ACO1** in colorectal cancer cells [Lin et al., 2022]. Therefore, it is tempting to speculate that **FDX2** silencing may enhance the efficacy of chemotherapy and improve cancer treatment outcomes. However, further research is needed to determine the specific role of **FDX2** in cancer and the potential therapeutic implications of the observed **FDX2** downregulation in cancer cells.

4.2.5 Functional relationships between DEGs

As previously mentioned in the Methods section, the functional enrichment analysis (conducted in R using ggprofiler2) did not yield statistically significant results. The fact that many of the top **DEGs** do not code for proteins (long non-coding RNAs and pseudogenes) explains the lack of functional results for the global gene set. This is also visible in the STRING network visualization of the top **DEGs** in Figure 19: no functional relationships between the global set of genes (no edges connecting the protein nodes).

The mitochondrial **DEGs** set is enriched in the cellular compartment mitochondrion (in GO CC ontology) as expected since this set was selected to be mitochondrial. The STRING functional network (Figure 20) shows strong links between the mitochondrially encoded proteins, and a few predicted (weaker evidence) links to nuclear-encoded proteins: **BCL2L10** and **BNIP3** (two **Bcl-2** family proteins involved in the balance between cell survival and apoptosis), **CMPK2** and the electron transport chain (homologous genes from other organisms are neighbors and coexpressed), and **AIFM3**, **FDX2**, and **PRODH** (discussed above).

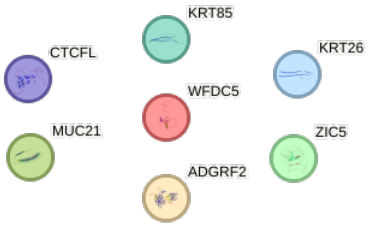


Figure 19: STRING functional interaction network for differentially expressed genes from the **global gene set**.

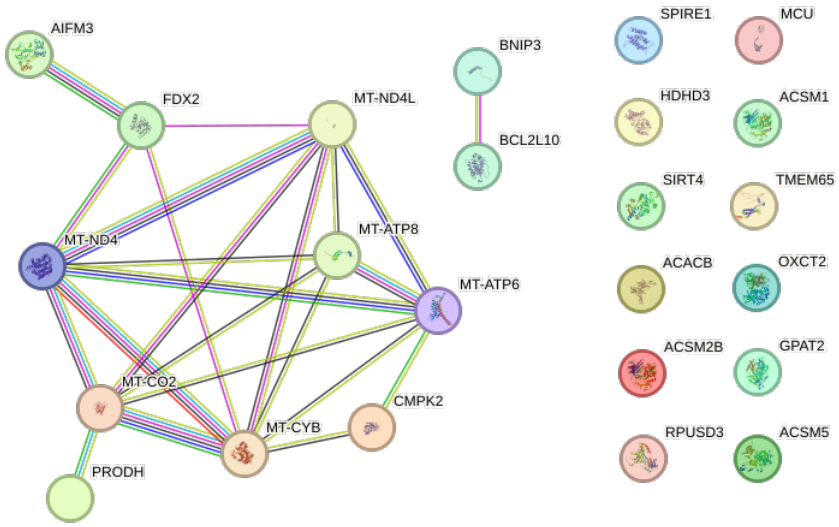


Figure 20: STRING functional interaction network for differentially expressed genes from the **mitochondrial gene set**.

4.3 Clustering

4.3.1 Hierarchical clustering

The primary objective of this study is to identify the gene expression profiles of mitochondrial genes in cancers originating from organs with alternative metabolic rates. To achieve this, two clustering analyses were conducted using the differentially expressed genes that encode mitochondrial proteins (102 genes out of 1136 that are present in MitoCarta3.0). The R code for this analysis is in the script "5.clustering.Rmd", and can be inspected at the GitHub repository.

Hierarchical clustering was performed first to visualize clusters of genes with similar expressions (Fig-

ure 21), followed by soft clustering using the c-means algorithm (Figure 22) with a fuzzification parameter, allowing genes to belong to more than one cluster [Ferraro and Giordani, 2020].

For hierarchical clustering, three alternative ways of computing the distance between a newly formed cluster and all other points or existing clusters were tested, namely, single linkage, complete linkage, and average linkage. Briefly, the single linkage calculates the distance between clusters as the smallest distance between any two points in the two clusters; in complete linkage the distance between clusters is defined as the largest distance between any two objects in the two clusters; and the average linkage computes the distance between clusters as an average of distances between all pairs of objects in the two clusters [Holmes and Huber, 2018]. This last one was chosen for analysis and discussion, because its results generated nearly the same clusters, but with longer inner branches (representing the strength of the clustering) (Annex B.3).

To accompany the hierarchical clustering results, a heatmap was generated, showing gene expression levels through a color gradient - intense colors representing higher expression and lighter shades indicating lower expression (Figure 21). The rows have been ordered according to the lateral dendrogram obtained from the hierarchical clustering. The columns (representing the individual samples) have been ordered by organ and cancer status prior to the clustering, as shown by the top color bar legend (NC = non-cancer, C = Cancer).

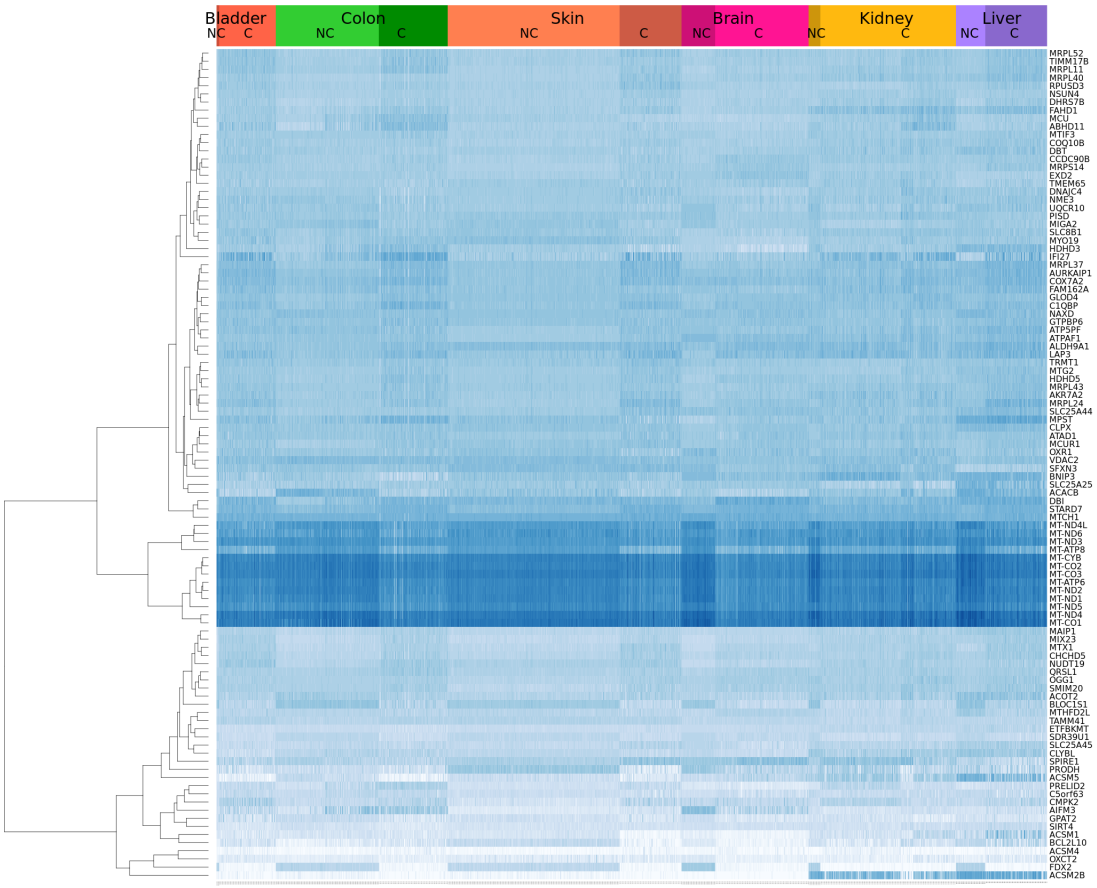


Figure 21: Heatmap expression of mitochondrial proteins grouped by hierarchical clustering. The intensity of the blue color indicates the level of gene expression (darker blue indicates higher expression). The dendrogram on the left shows the hierarchical clustering of the genes. The color code at the top indicates the organs (bladder in coral; colon in green; skin in orange; brain in pink; kidney in yellow; liver in purple) labeled as non cancer (NC lighter shade) or cancer (C darker shade).

Hierarchical clustering discussion

The hierarchical clustering analysis reveals a visible group of 13 genes that are more expressed (darker color) than the remaining clusters. This cluster groups the 13 mitochondrially encoded genes, namely, **MT-ND4L**, **MT-ND6**, **MT-ND3**, and **MT-ATP8** in one cluster, which then groups with the rest of the mitogenes (**MT-CYB**, **MT-CO2**, **MT-CO3**, **MT-ATP6**, **MT-ND2**, **MT-ND1**, **MT-ND5**, **MT-ND4**, and **MT-CO1**). The fact that they are encoded in mitochondrial **DNA** offers an immediate explanation for this clustering, owing to their shared biological function and coordinated regulation inside mitochondria, as discussed above. Interestingly, the exact same clusters are shown in the other two hierarchical clustering methods tested (Annex B.3).

As expected, **FDX2** clusters poorly with other genes and shows a very distinctive pattern of expression highly expressed in healthy patients across various organs, but lowly expressed in cancer samples, in line with the observed results shown in the boxplots discussed above (Figure 17).

The hierarchical clustering results also show clusters of genes, such as **ACSM2B** (the last gene on the heatmap) [ens, b], that are highly expressed in some organs regardless of the samples' health status. This gene encodes an acyl-CoA synthetase that catalyzes the activation of free fatty acids (**FFAs**) to CoA esters through a two-step thioesterification reaction [Vessey et al., 1999]. The absence of differences between healthy and cancer samples prompts us to look at the type of organ where the gene is most expressed. We can observe that it is primarily expressed in the kidneys and liver, both of which are high-metabolism organs that play an important role in lipid metabolism. The liver regulates the synthesis, degradation, and storage of lipids. Acyl-CoA synthetases in the liver are essential for activating fatty acids, which can then be utilized for energy production (via beta-oxidation) or incorporated into lipids for storage or export as lipoproteins [Vessey et al., 1999].

While the kidneys are primarily known for their role in filtration and excretion, they also participate in lipid metabolism. Proximal tubules utilize a variety of substrates for energy metabolism, but fatty acids are preferred, especially during conditions of fasting or increased energy demand [Bobulescu, 2010]. Acyl-CoA synthetases in the kidneys facilitate the activation of fatty acids for this purpose.

4.3.2 Fuzzy clustering: Mitochondrial expression profiles

The distinct clusters of mitochondrial proteins found by the fuzzy clustering algorithm are shown in Figure 22 and the genes listed in Table 6. These clusters can be interpreted as the individual mitochondrial expression profiles sought by this study.

Most profiles cluster several genes, however, a few clusters comprise only a single gene, showing their unique pattern of expression across the different sample groups (Table 6). These clusters (1, 3, 4, 5, 11, and 16) highlight many of the genes previously discussed, such as **PRODH**, **FDX2**, and **ACSM2B**.

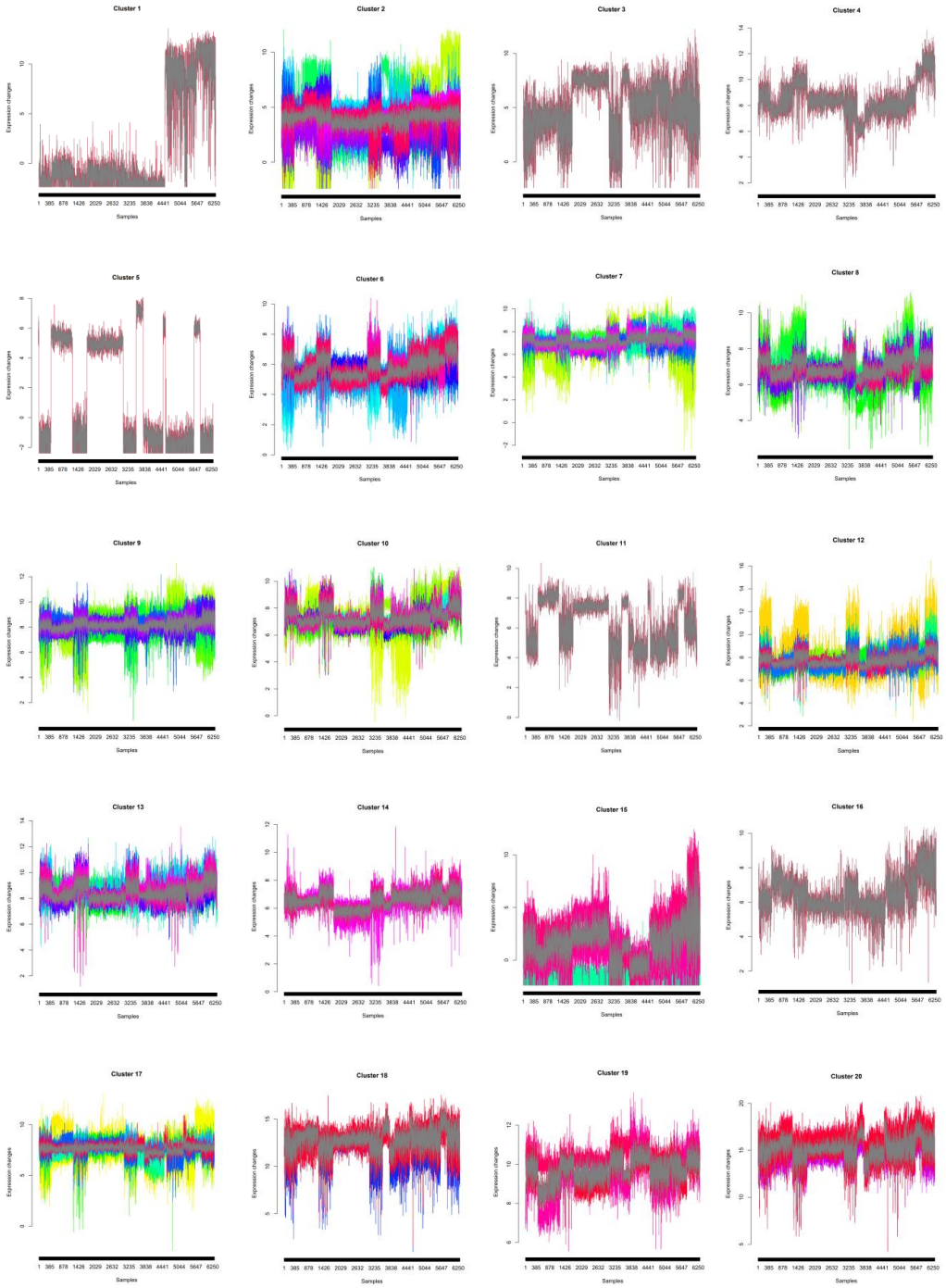


Figure 22: Fuzzy clustering results for the differentially expressed mitochondrial genes. Clusters that exhibit a single grey color overlaid with red indicate the presence of only one gene in that cluster, while clusters with multiple colors show more than one gene present. The algorithm was parameterized requesting for 20 clusters.

Table 6: Mitochondrial genes present in each of the 20 clusters presented in Figure 22.

Cluster	Genes
1	ACSM2B
2	SIRT4, SDR39U1, CMPK2, ETFBKMT, C5orf63, ACSM5, AIFM3, GPAT2, PRELID2
3	PRODH
4	MPST
5	FDX2
6	CLYBL, CHCHD5, TAMM41, MIX23, SLC25A45, MAIP1, MTHFD2L, MTX1
7	EXD2, MRPS14, MTIF3, SPIRE1, CCDC90B, DBT, TMEM65
8	ABHD11, DHRS7B, COQ10B, NSUN4, MCU, NUDT19
9	SFXN3, ATPAF1, ATAD1, OXR1, CLPX, GLOD4, BNIP3, GTPBP6, NAXD
10	MTG2, HDHD3, TIMM17B, RPUSD3, MRPL52, MRPL11, FAHD1, UQCR10, MRPL40
11	BLOC1S1
12	MCUR1, AKR7A2, MRPL43, HDHD5, MRPL24, SLC25A44, IFI27
13	LAP3, C1QBP, COX7A2, FAM162A, MRPL37, ALDH9A1, ATP5PF, VDAC2, AURKAIP1
14	OGG1, QRSL1, SMIM20
15	BCL2L10, ACSM1, OXCT2, ACSM4
16	ACOT2
17	ACACB, SLC8B1, NME3, TRMT1, DNAJC4, SLC25A25, MIGA2, PISD, MYO19
18	MT-ND6, MT-ND3, MT-ND4L, MT-ATP8
19	STARD7, MTCH1, DBI
20	MT-CO2, MT-CYB, MT-ND2, MT-ND5, MT-CO1, MT-ND4, MT-ND1, MT-ATP6, MT-CO3

Others, such as **MPST** (Mercaptopyruvate sulfurtransferase), **BLOC1S1** (Biogenesis Of Lysosomal Organelles Complex 1 Subunit 1), and **ACOT2** (Acyl-CoA thioesterase 2) appear here as potential markers for alternative regulation modes and/or unique expression programs, and therefore interesting candidates to shed light on potential new active mechanisms in cancer.

As expected, the 13 mitochondrially encoded genes form two individual clusters (cluster 18 and cluster 20), corresponding to the main darker cluster shown in the heatmap (Figure 21). These clusters support the hierarchical clustering results (Figure 21), as genes encoded by the mitochondria displayed a consistently high level of expression across all samples, distinguishing them from the remaining genes.

Finally, cluster 12 shows a particularly distinctive pattern of expression higher in cancer tissues than in

healthy controls (seen by the alternating up and down pattern). This cluster groups seven genes (**MCUR1**, **AKR7A2**, **MRPL43**, **HDHD5**, **MRPL24**, **SLC25A44**, and **IFI27**), all are enzymes functionally unrelated (confirmed in STRING), except for **MRPL43** and **MRPL24** which are both subunits of the mitochondrial ribosome. Since these proteins are encoded in the nucleus, this co-expression pattern might suggest a coordinated regulation in the nucleus to ensure proper mitochondrial function in response to cellular and environmental changes occurring in cancer. However, further research is required to clarify the potential mechanisms underlying the different reported expression profiles.

Chapter 5

Conclusions and Future work

This thesis presents research aimed at identifying differentially expressed genes in cancer, specifically focused on the genes that encode mitochondrial proteins, between high metabolic rate organs and low metabolic rate organs (Figure 23). The research involved the collection and analysis of two independent transcriptomics datasets comprising samples of normal tissue (**GTEx**) and cancer tissue (**TCGA**). Principal Component Analysis (**PCA**) was employed, revealing a significant overlap between samples from both **TCGA** and **GTEx** datasets, indicating the suitability of these datasets for further investigation.

Subsequently, a differential expression analysis was conducted using general linear models, identifying 8651 differentially expressed genes. These genes were inspected using volcano plots and the distribution of the top genes was visualized with boxplots. Among these genes, notable differences in expression were observed for mitochondrial protein-encoding genes like **ACSM1**, **ACSM5**, and **PRODH**.

Furthermore, the boxplots provided insights into the expression patterns of each gene across different organs and health status (non cancer and cancer). Interestingly, the main differences were observed between normal and cancer tissues rather than between high and low metabolic rate organs.

The final step of this study applied hierarchical and fuzzy clustering methods to the gene expression matrix to find expression profiles. These analyses highlighted distinct clusters of genes encoded in mitochondrial **DNA**, emphasizing their role in energy production. Additionally, unique expression patterns were identified for relevant genes, namely **FDX2** and **ACSM2B**, making them good candidates for future experimental validation (see Figure 23).

This study offers an overview of candidate genes exhibiting differential expression between normal and cancer tissues across six distinct organs. Specifically, the bladder, colon and skin represent low metabolic rate organs, while the brain, liver and kidney belong to the high metabolic rate category. The selection of these organs was based on the availability of comprehensive data within the **TCGA** dataset, offering the potential to identify common expression profiles.

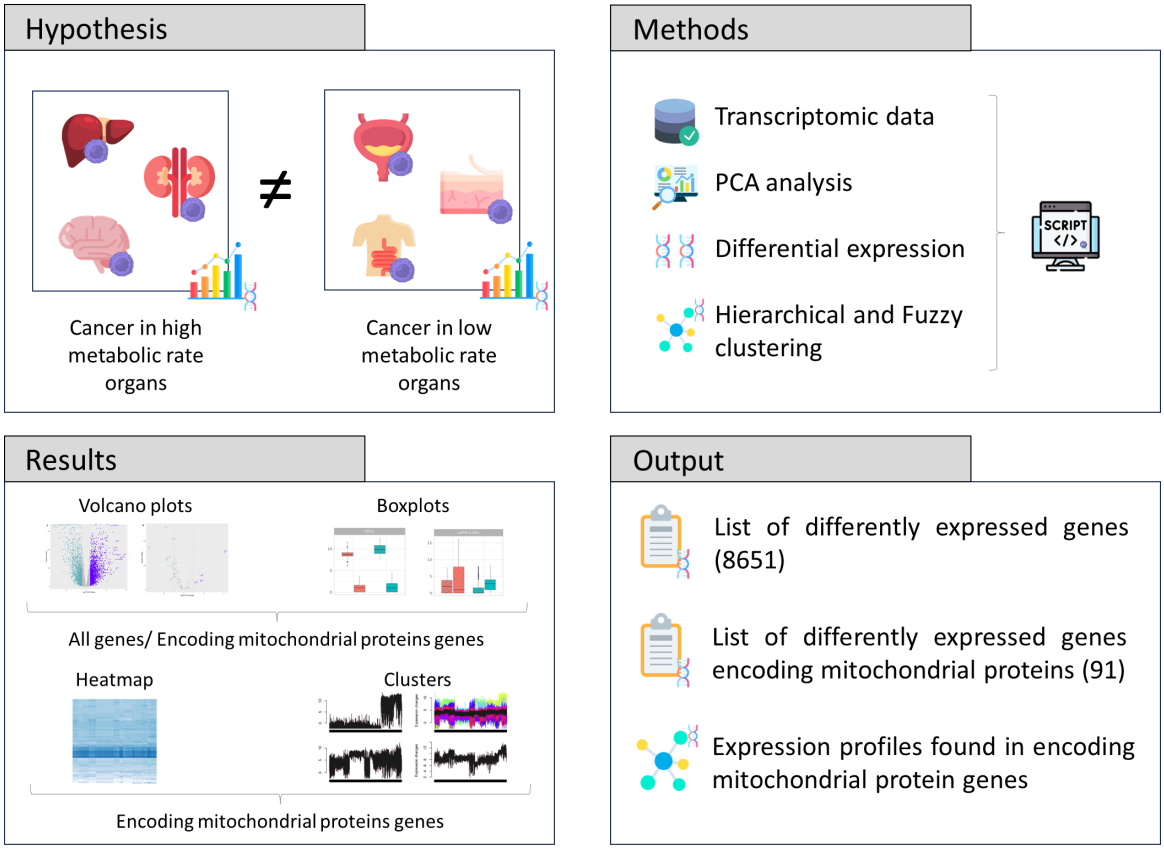


Figure 23: Overview of the thesis outline. This study aimed to identify differentially expressed genes associated with cancer, specifically focusing on the comparison between high metabolic rate and low metabolic rate organs, particularly those encoding mitochondrial proteins. To accomplish this, we used two RNA-seq datasets derived from both normal (**GTEx**) and cancer tissues (**TCGA**). A total of 8651 genes were analysed to find differential expression. Subsequently, hierarchical and fuzzy clustering algorithms were applied to a subset of 91 genes that encode mitochondrial proteins. These analyses showed 20 distinct clusters, ultimately presenting a valuable list of candidate genes for future validation studies.

Transcriptomics data analysis was a pivotal component of this study. RNA-seq data was used due to its numerous advantages when compared to other alternative methods like microarrays. These advantages encompass comprehensive transcript coverage, heightened sensitivity, the capacity to detect allele-specific differential expression, and the identification of novel transcripts [Bradford et al., 2010].

To conduct the differential expression analysis, the edgeR package was selected. This analytical approach enabled the identification of differentially expressed genes across distinct cancer tissues, subsequently allowing for their grouping into clusters if their expression profiles resembled those of other genes. Following this analysis, we pinpointed 8651 genes with significant changes in expression that were later visualized through volcano plots. Among these genes, thirteen emerged as particularly noteworthy, encom-

passing both upregulated and downregulated ones, thereby offering valuable insights into the biological processes associated with cancer status.

Within the downregulated genes featured **MUC21** and **LHFPL3-AS1**, both recognized for their roles in inhibiting cell adhesion and apoptosis, aligning with well established cancer hallmarks.

Regarding the upregulated genes encoding mitochondrial proteins, **ACSM1**, **ACSM5**, and **PRODH** suggested a potential adaptation of cancer cells to metabolic stress, ensuring energy production under adverse conditions. Conversely, downregulated mitochondrial genes like **MT-ATP6** indicate a potential reduction in aerobic respiration in cancer cells, possibly associated with their adaptation to harsher conditions.

In order to visualize the expression profiles of genes encoding mitochondrial proteins, an hierarchical and fuzzy clustering analysis was performed. These analyses unveiled clusters of genes with coordinated expression patterns. Particularly noteworthy were the genes encoded in mitochondrial **DNA**, namely **MT-ND6**, **MT-ND3**, **MT-ND4L**, **MT-ATP8**, **MT-CO2**, **MT-CYB**, **MT-ND2**, **MT-ND5**, **MT-CO1**, **MT-ND4**, **MT-ND1**, **MT-ATP6**, and **MT-CO3**, which formed distinct clusters exhibiting high expression levels. This finding strongly validates this analysis, since the fact that they are encoded in mitochondrial **DNA** leads to a shared regulation inside the organelle, leading to their co-regulation appearing in a different cluster from all other nuclear encoded genes. Additionally, individual gene expression profiles, exemplified by **FDX2** and **ACSM2B**, displayed unique patterns that warrant further in-depth investigation.

Overall, this study has illuminated the intricate interplay among gene expression, cancer status, and metabolic rates across diverse organs. The identification of specific genes and clusters exhibiting significant expression changes lays the groundwork for further research and experimental validation. Comprehending these molecular expression profiles holds the potential to yield valuable insights into cancer biology, metabolic regulation, and prospective therapeutic targets. This, in turn, can advance our understanding and pave the way for personalized medicine approaches.

However, it is important to notice that our initial hypothesis was not conclusively shown since when visualizing the gene expression through boxplots for each dataset and organ, we observed that the primary differences in gene expression stemmed from the comparison between normal and cancer samples, irrespective of the metabolic rates of the organs under analysis. This suggests that the signal associated with metabolic rate was relatively weak (if it is there at all) when compared to the dominant influence of the cancer status. Additional limitations, including unequal sample sizes for each organ, and the usage of two different data sources, must be considered when interpreting these results. However, despite these limitations, our findings provide valuable insights into the complex interplay between metabolism and gene

expression in cancer, providing some additional clues to the molecular mechanisms underlying cancer onset and development.

In the future it would be interesting to reformat this analysis including only cancer samples, enabling a more concentrated examination of the metabolic rate signal. Alternatively, we could explore a strategy focused on filtering the expression for the genes involved in cellular metabolism, irrespective of their association with mitochondria. Additionally, expanding our analysis to encompass a broader range of organs and subsequently categorizing them into high or low metabolic rate groups could be explored as another avenue to achieve our research objectives.

Bibliography

MitoCarta3.0. https://personal.broadinstitute.org/scalvo/MitoCarta3.0/Human.MitoCarta3.0.xls.

Mitochondrial DNA | Cancer Genetics Web. http://www.cancerindex.org/geneweb/gmtdna.htm.

ACSM1 (ENSG00000166743) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000166743;r=16:20623235-20698890, a. [Accessed 2-10-2023].

ACSM2B (ENSG00000066813) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000066813;r=16:20536226-20576427, b. [Accessed 02-10-2023].

ACSM5 (ENSG00000183549) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000183549;r=16:20409534-20441336, c. [Accessed 02-10-2023].

AIFM3 (ENSG00000183773) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000183773;r=22:20965108-20981360, d. [Accessed 02-10-2023].

BCL2L10 (ENSG00000137875) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000137875;r=15:52109263-52112775, e. [Accessed 02-10-2023].

CTCFL (ENSG00000124092) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000124092;r=20:57495966-57525652, f. [Accessed 2-10-2023].

FDX2 (ENSG00000267673) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000267673;r=19:10309916-10316015, g. [Accessed 02-10-2023].

KRT26 (ENSG00000186393) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000186393;r=17:40766238-40772201;t=ENST00000335552, h. [Accessed 2-10-2023].

KRT85 (ENSG00000135443) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000135443;r=12:52360006-52367481, i. [Accessed 2-10-2023].

LHFPL3-AS1 (ENSG00000226869) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000226869;r=7:104738597-104804107, j. [Accessed 02-10-2023].

MT-TM (ENSG00000210112) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000210112;r=MT:4402-4469;t=ENST00000387377, k. [Accessed 02-10-2023].

MT-TV (ENSG00000210077) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000210077;r=MT:1602-1670;t=ENST00000387342, l. [Accessed 2-10-2023].

MUC21 (ENSG00000204544) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000204544;r=6:30983718-30989903, m. [Accessed 02-10-2023].

PRODH (ENSG00000100033) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000100033;r=22:18912777-18936553, n. [Accessed 02-10-2023].

WFDC5 (ENSG00000175121) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000175121;r=20:45109452-45115174, o. [Accessed 2-10-2023].

ZIC5 (ENSG00000139800) | Ensembl. https://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000139800;r=13:99962964-99971767;t=ENST00000267294, p. [Accessed 2-10-2023].

FDX2 ferredoxin 2 | Gene Cards. https://www.genecards.org/cgi-bin/carddisp.pl?gene=FDX2. [Accessed 18-10-2023].

GTEx Portal. https://gtexportal.org/home/datasets.

FDX2 ferredoxin 2 | NCBI. https://www.ncbi.nlm.nih.gov/gene/112812. [Accessed 18-10-2023].

FDX2 protein expression summary | The Human Protein Atlas. https://www.proteinatlas.org/ENSG00000267673-FDX2. [Accessed 18-10-2023].

STRING: functional protein association networks. https://string-db.org/.

TCGA. https://portal.gdc.cancer.gov/.

UniProt. https://www.uniprot.org/uniprotkb/Q6P4F2/entry.

Jerry M Adams and Suzanne Cory. The bcl-2 apoptotic switch in cancer development and therapy. *Oncogene*, 26(9):1324–1337, 2007.

Miri Adler, Avi Mayo, and Uri Alon. Logarithmic and power law input-output relations in sensory systems with fold-change detection. *PLoS computational biology*, 10(8):e1003781, 2014.

Richard Altmann. *Die Elementarorganismen und ihre Beziehungen zu den Zellen*. Leipzig : Veit comp., 1890.

Preetha Anand, Ajaikumar B Kunnumakara, Chitra Sundaram, Kuzhuvelil B Harikumar, Sheeja T Tharakan, Oiki S Lai, Bokyung Sung, and Bharat B Aggarwal. Cancer is a preventable disease that requires major lifestyle changes. *Pharmaceutical research*, 25:2097–2116, 2008.

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Nature Precedings*, pages 1–1, 2010.

Carl Benda. Ueber die spermatogenese der vertebraten und höherer evertebraten, ii. theil: Die histiogenese der spermien. *Arch. Anat. Physiol*, 73:393–398, 1898.

Hans-Ulrich Bender, Shlomo Almashanu, Gary Steel, Chien-An Hu, Wei-Wen Lin, Alecia Willis, Ann Pulver, and David Valle. Functional consequences of prodh missense mutations. *The American Journal of Human Genetics*, 76(3):409–420, 2005.

María Berdasco and Manel Esteller. Aberrant epigenetic landscape in cancer: how cellular identity goes awry. *Developmental cell*, 19(5):698–711, 2010.

Maria A Blasco. Telomeres and human disease: ageing, cancer and beyond. *Nature Reviews Genetics*, 6 (8):611–622, 2005.

Ion Alexandru Bobulescu. Renal lipid metabolism and lipotoxicity. *Current opinion in nephrology and hypertension*, 19(4):393, 2010.

Sayantan Bose, Abir Kumar Panda, Shravanti Mukherjee, and Gaurisankar Sa. Curcumin and tumor immune-editing: resurrecting the immune system. *Cell division*, 10(1):1–13, 2015.

James R Bradford, Yvonne Hey, Tim Yates, Yaoyong Li, Stuart D Pepper, and Crispin J Miller. A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, 11(1):1–12, 2010.

Joseph J Braymer and Roland Lill. Iron–sulfur cluster biogenesis and trafficking in mitochondria. *Journal of Biological Chemistry*, 292(31):12754–12763, 2017.

James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC bioinformatics*, 11(1):1–13, 2010.

Deborah L Burkhart and Julien Sage. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nature Reviews Cancer*, 8(9):671–682, 2008.

Aparajita H Chourasia, Michelle L Boland, and Kay F Macleod. Mitophagy and cancer. *Cancer & metabolism*, 3(1):1–11, 2015.

Antonio Colaprico, Tiago C Silva, Catharina Olsen, Luciano Garofano, Claudia Cava, Davide Garolini, Thais S Sabedot, Tathiane M Malta, Stefano M Pagnotta, Isabella Castiglioni, et al. Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data. *Nucleic acids research*, 44(8):e71–e71, 2016.

Francesco Colotta, Paola Allavena, Antonio Sica, Cecilia Garlanda, and Alberto Mantovani. Cancer-related inflammation, the seventh hallmark of cancer: links to genetic instability. *Carcinogenesis*, 30(7):1073–1081, 2009.

Chi V Dang, Jung-whan Kim, Ping Gao, and Jason Yustein. The interplay between myc and hif in cancer. *Nature Reviews Cancer*, 8(1):51–56, 2008.

Karin E De Visser, Alexandra Eichten, and Lisa M Coussens. Paradoxical roles of the immune system during cancer development. *Nature reviews cancer*, 6(1):24–37, 2006.

Ralph J DeBerardinis, Julian J Lum, Georgia Hatzivassiliou, and Craig B Thompson. The biology of cancer: metabolic reprogramming fuels cell growth and proliferation. *Cell metabolism*, 7(1):11–20, 2008.

David G DeNardo, Pauline Andreu, and Lisa M Coussens. Interactions between lymphocytes and myeloid cells regulate pro-versus anti-tumor immunity. *Cancer and Metastasis Reviews*, 29(2):309–316, 2010.

Gina M DeNicola, Florian A Karreth, Timothy J Humpton, Aarthi Gopinathan, Cong Wei, Kristopher Frese, Dipti Mangal, Kenneth H Yu, Charles J Yeo, Eric S Calhoun, et al. Oncogene-induced nrf2 transcription promotes ros detoxification and tumorigenesis. *Nature*, 475(7354):106–109, 2011.

Marinos Elia. Organ and tissue contribution to metabolic rate. *Energy Metabolism. Tissue Determinants and Cellular Corrolaries*, pages 61–77, 1992.

Lars Ernster and Gottfried Schatz. Mitochondria: a historical review. *The Journal of cell biology*, 91(3): 227s–255s, 1981.

Manel Esteller. Cancer epigenomics: Dna methylomes and histone-modification maps. *Nature reviews genetics*, 8(4):286–298, 2007.

Olivier Feron. Pyruvate into lactate and back: from the warburg effect to symbiotic energy fuel exchange in cancer cells. *Radiotherapy and oncology*, 92(3):329–333, 2009.

Maria Brigida Ferraro and Paolo Giordani. Soft clustering. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(1):e1480, 2020.

Juliana Carvalho Ferreira and Cecilia Maria Patino. What does the p value really mean? *Jornal Brasileiro de Pneumologia*, 41(5):485, 2015.

Christopher R Genovese, Nicole A Lazar, and Thomas Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878, 2002.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1–16, 2004.

Omer Gersten and John R Wilmoth. The cancer transition in japan since 1951. *Demographic Research*, 7:271–306, 2002.

Nader Ghebranious and Lawrence A Donehower. Mouse models in tumor suppression. *Oncogene*, 17 (25):3385–3400, 1998.

Marta Giacomello, Aswin Pyakurel, Christina Glytsou, and Luca Scorrano. The cell biology of mitochondrial membrane dynamics. *Nature reviews Molecular cell biology*, 21(4):204–224, 2020.

Sergei I Grivennikov, Florian R Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140 (6):883–899, 2010.

Jessie Yanxiang Guo, Gizem Karsli-Uzunbas, Robin Mathew, Seena C Aisner, Jurre J Kamphorst, Anne M Strohecker, Guanghua Chen, Sandy Price, Wenyun Lu, Xin Teng, et al. Autophagy suppresses progression of k-ras-induced lung tumors to oncocytomas and maintains lipid homeostasis. *Genes & development*, 27(13):1447–1461, 2013.

Michael Hahsler, Matthew Piekenbrock, and Derek Doran. dbscan: Fast density-based clustering with r. *Journal of Statistical Software*, 91:1–30, 2019.

Hellyeh Hamidi and Johanna Ivaska. Every step of the way: integrins in cancer progression and metastasis. *Nature Reviews Cancer*, 18(9):533–548, 2018.

Douglas Hanahan. Hallmarks of cancer: new dimensions. *Cancer discovery*, 12(1):31–46, 2022.

Douglas Hanahan and Robert A Weinberg. The hallmarks of cancer. *cell*, 100(1):57–70, 2000.

Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

John H Holland. *Hidden order: How adaptation builds complexity*. Addison Wesley Longman Publishing Co., Inc., 1996.

Susan Huber Holmes and Wolfgang Huber. *Modern statistics for modern biology*. Cambridge University Press, 2018.

Peggy P Hsu and David M Sabatini. Cancer cell metabolism: Warburg and beyond. *Cell*, 134(5):703–707, 2008.

Xinling Hu, Liu Yang, and Yin-Yuan Mo. Role of pseudogenes in tumorigenesis. *Cancers*, 10(8):256, 2018.

Paul J Hurd and Christopher J Nelson. Advantages of next-generation sequencing versus the microarray in epigenetic research. *Briefings in Functional Genomics and Proteomics*, 8(3):174–183, 2009.

Stephen P Jackson and Jiri Bartek. The dna-damage response in human biology and disease. *Nature*, 461(7267):1071–1078, 2009.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374 (2065):20150202, 2016.

Peter A Jones and Stephen B Baylin. The epigenomics of cancer. *Cell*, 128(4):683–692, 2007.

Russell G Jones and Craig B Thompson. Tumor suppressors and cell metabolism: a recipe for cancer growth. *Genes & development*, 23(5):537–548, 2009.

Antoine E Karnoub and Robert A Weinberg. Chemokine networks and breast cancer metastasis. *Breast disease*, 26(1):75–85, 2007.

Atsuko Kasahara and Luca Scorrano. Mitochondria: from cell death executioners to regulators of cell differentiation. *Trends in cell biology*, 24(12):761–770, 2014.

Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

Kelly M Kennedy and Mark W Dewhirst. Tumor metabolism of lactate: the influence and therapeutic potential for mct and cd147 regulation. *Future Oncology*, 6(1):127–148, 2010.

Kimberly R Kukurba and Stephen B Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015.

David P Lane. p53, guardian of the genome. *Nature*, 358(6381):15–16, 1992.

Charity W Law, Kathleen Zeglinski, Xueyi Dong, Monther Alhamdoosh, Gordon K Smyth, and Matthew E Ritchie. A guide to creating design matrices for gene expression experiments. *F1000Research*, 9, 2020.

Sébastien Lê, Julie Josse, and François Husson. Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25:1–18, 2008.

Mark A Lemmon and Joseph Schlessinger. Cell signaling by receptor tyrosine kinases. *Cell*, 141(7): 1117–1134, 2010.

Feng Li, Yunyue Wang, Karen I Zeller, James J Potter, Diane R Wonsey, Kathryn A O'Donnell, Jung-whan Kim, Jason T Yustein, Linda A Lee, and Chi V Dang. Myc stimulates nuclearly encoded mitochondrial genes and mitochondrial biogenesis. *Molecular and cellular biology*, 25(14):6225–6234, 2005.

Jin-Fei Lin, Pei-Shan Hu, Yi-Yu Wang, Yue-Tao Tan, Kai Yu, Kun Liao, Qi-Nian Wu, Ting Li, Qi Meng, Jun-Zhong Lin, et al. Phosphorylated nfs1 weakens oxaliplatin-based chemosensitivity of colorectal cancer by preventing panoptosis. *Signal transduction and targeted therapy*, 7(1):54, 2022.

Xueping Liu, Yajun Xiao, Xia Xiong, and Xiaoyi Qi. Muc21 controls melanoma progression via regulating slitrk5 and hedgehog signaling pathway. *Cell Biology International*, 46(9):1458–1467, 2022.

Jonathan Lopez and SWG Tait. Mitochondrial apoptosis: killing cancer using the enemy within. *British journal of cancer*, 112(6):957–962, 2015.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

Martin Maechler. Cluster: cluster analysis basics and extensions. *R package version 2.0. 7–1*, 2018.

Nunziata Maio and Tracey A Rouault. Iron–sulfur cluster biogenesis in mammalian cells: new insights into the molecular mechanisms of cluster delivery. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1853(6):1493–1512, 2015.

Joseph D Mancias and Alec C Kimmelman. Mechanisms of selective autophagy in normal physiology and cancer. *Journal of molecular biology*, 428(9):1659–1680, 2016.

Samuel Marguerat and Jürg Bähler. Rna-seq: from technology to biology. *Cellular and molecular life sciences*, 67:569–579, 2010.

Jean-Claude Martinou and Richard J Youle. Mitochondria in apoptosis: Bcl-2 family members and mitochondrial dynamics. *Developmental cell*, 21(1):92–101, 2011.

Satoru Matsuda, Atsuko Nakanishi, Akari Minami, Yoko Wada, and Yasuko Kitagishi. Functions and characteristics of pink1 and parkin in cancer. *Front Biosci (Landmark Ed)*, 20:491–501, 2015.

Prashant Mishra and David C Chan. Metabolic regulation of mitochondrial dynamics. *Journal of Cell Biology*, 212(4):379–387, 2016.

Masahiro Morita, Simon-Pierre Gravel, Laura Hulea, Ola Larsson, Michael Pollak, Julie St-Pierre, and Ivan Topisirovic. mtor coordinates protein synthesis, mitochondrial activity and proliferation. *Cell cycle*, 14 (4):473–480, 2015.

Daniel Müllner. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. *Journal of Statistical Software*, 53:1–18, 2013.

Simona Negrini, Vassilis G Gorgoulis, and Thanos D Halazonetis. Genomic instability—an evolving hallmark of cancer. *Nature reviews Molecular cell biology*, 11(3):220–228, 2010.

Abdel R Omran. The epidemiologic transition theory revisited thirty years later. *World health statistics quarterly*, 53(2, 3, 4):99–119, 1998.

Sarah F Pearce, Pedro Rebelo-Guiomar, Aaron R D'Souza, Christopher A Powell, Lindsey Van Haute, and Michal Minczuk. Regulation of mammalian mitochondrial gene expression: recent advances. *Trends in biochemical sciences*, 42(8):625–639, 2017.

Michael S Petronek, Douglas R Spitz, and Bryan G Allen. Iron–sulfur cluster biogenesis as a critical target in cancer. *Antioxidants*, 10(9):1458, 2021.

Elena Piskounova, Michalis Agathocleous, Malea M Murphy, Zeping Hu, Sara E Huddlestun, Zhiyu Zhao, A Marilyn Leitch, Timothy M Johnson, Ralph J DeBerardinis, and Sean J Morrison. Oxidative stress inhibits distant metastasis by human melanoma cells. *Nature*, 527(7577):186–191, 2015.

Paolo E Porporato, Valéry L Payen, Jhudit Pérez-Escuredo, Christophe J De Saedeleer, Pierre Danhier, Tamara Copetti, Suveera Dhup, Morgane Tardy, Thibaut Vazeille, Caroline Bouzin, et al. A mitochondrial switch promotes tumor metastasis. *Cell reports*, 8(3):754–766, 2014.

Aswin Pyakurel, Claudia Savoia, Daniel Hess, and Luca Scorrano. Extracellular regulated kinase phosphorylates mitofusin 1 to control mitochondrial morphology and apoptosis. *Molecular cell*, 58(2):244–254, 2015.

Bin-Zhi Qian and Jeffrey W Pollard. Macrophage diversity enhances tumor progression and metastasis. *Cell*, 141(1):39–51, 2010.

Marius Raica, Anca Maria Cimpean, and Domenico Ribatti. Angiogenesis in pre-malignant conditions. *European journal of cancer*, 45(11):1924–1934, 2009.

Sneha Rath, Rohit Sharma, Rahul Gupta, Tslil Ast, Connie Chan, Timothy J Durham, Russell P Goodman, Zenon Grabarek, Mary E Haas, Wendy HW Hung, et al. Mitocarta3. 0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic acids research*, 49 (D1):D1541–D1547, 2021.

Jüri Reimand, Meelis Kull, Hedi Peterson, Jaanus Hansen, and Jaak Vilo. g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl_2): W193–W200, 2007.

Thibaud T Renault, Konstantinos V Floros, Rana Elkholi, Kelly-Ann Corrigan, Yulia Kushnareva, Shira Y Wieder, Claudia Lindtner, Madhavika N Serasinghe, James J Asciolla, Christoph Buettner, et al. Mitochondrial shape governs bax-induced membrane permeabilization and apoptosis. *Molecular cell*, 57 (1):69–82, 2015.

Mark Richardson. Principal component analysis. *URL: http://people. maths. ox. ac. uk/richardson-m/SignalProcPCA. pdf (last access: 3.5. 2013). Aleš Hladnik Dr., Ass. Prof., Chair of Information and Graphic Arts Technology, Faculty of Natural Sciences and Engineering, University of Ljubljana, Slovenia ales. hladnik@ ntf. uni-lj. si*, 6(16):4, 2009.

Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):1–9, 2010.

Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, 26(1):139–140, 2010.

Tracey A Rouault. Biogenesis of iron-sulfur clusters in mammalian cells: new insights and relevance to human disease. *Disease models & mechanisms*, 5(2):155–164, 2012.

Jesse J Salk, Edward J Fox, and Lawrence A Loeb. Mutational heterogeneity in human cancers: origin and consequences. *Annual review of pathology*, 5:51, 2010.

Patricia Sancho, Emma Burgos-Ramos, Alejandra Tavera, Tony Bou Kheir, Petra Jagust, Matthieu Schoenhals, David Barneda, Katherine Sellers, Ramon Campos-Olivas, Osvaldo Grana, et al. Myc/pgc-1α balance determines the metabolic phenotype and plasticity of pancreatic cancer stem cells. *Cell metabolism*, 22(4):590–605, 2015.

Kristopher A Sarosiek, Xiaoke Chi, John A Bachman, Joshua J Sims, Joan Montero, Luv Patel, Annabelle Flanagan, David W Andrews, Peter Sorger, and Anthony Letai. Bid preferentially activates bak while bim preferentially activates bax, affecting chemotherapy response. *Molecular cell*, 51(6):751–765, 2013.

Luca Scrucca, Michael Fop, T Brendan Murphy, and Adrian E Raftery. mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, 8(1):289, 2016.

Gregg L Semenza et al. Tumor metabolism: cancer cells give and take lactate. *The Journal of clinical investigation*, 118(12):3835–3837, 2008.

Daniela Senft and A Ronai Ze'ev. Regulators of mitochondrial dynamics in cancer. *Current opinion in cell biology*, 39:43–52, 2016.

Gerald S Shadel and Tamas L Horvath. Mitochondrial ros signaling in organismal homeostasis. *Cell*, 163 (3):560–569, 2015.

Ruifeng Shi, Wenya Hou, Zhao-Qi Wang, and Xingzhi Xu. Biogenesis of iron–sulfur clusters and their role in dna metabolism. *Frontiers in Cell and Developmental Biology*, 9:735678, 2021.

Yanbo Shi, Manik Ghosh, Gennadiy Kovtunovych, Daniel R Crooks, and Tracey A Rouault. Both human ferredoxins 1 and 2 and ferredoxin reductase are important for iron-sulfur cluster biogenesis. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1823(2):484–492, 2012.

Alex Sigal and Varda Rotter. Oncogenic mutations of the p53 tumor suppressor: the demons of the guardian of the genome. *Cancer research*, 60(24):6788–6793, 2000.

Lindsay I Smith. A tutorial on principal components analysis. 2002.

Lucas B Sullivan and Navdeep S Chandel. Mitochondrial reactive oxygen species and cancer. *Cancer & metabolism*, 2:1–12, 2014.

Diane Sweeney and Brad Williamson. *Biology: Exploring Life: Laboratory Manual*. Pearson Education, Incorporated, 2006.

James E Talmadge and Isaiah J Fidler. Aacr centennial series: the biology of cancer metastasis: historical perspective. *Cancer research*, 70(14):5649–5669, 2010.

Zheqiong Tan, Xiangjian Luo, Lanbo Xiao, Min Tang, Ann M Bode, Zigang Dong, and Ya Cao. The role of pgc1α in cancer metabolism and its therapeutic implicationspgc1α in cancer metabolism and its therapeutic implications. *Molecular cancer therapeutics*, 15(5):774–782, 2016.

Yue Tang, Jing Zhou, Shing Chuan Hooi, Yue-Ming Jiang, and Guo-Dong Lu. Fatty acid activation in carcinogenesis and cancer development: Essential roles of long-chain acyl-coa synthetases. *Oncology letters*, 16(2):1390–1396, 2018.

Michele WL Teng, Jeremy B Swann, Catherine M Koebel, Robert D Schreiber, and Mark J Smyth. Immune-mediated dormancy: an equilibrium with cancer. *Journal of Leucocyte Biology*, 84(4):988–993, 2008.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. Corum: the comprehensive resource of mammalian protein complexes–2022. *Nucleic acids research*, 51(D1):D539–D545, 2023.

Claire M Vajdic and Marina T Van Leeuwen. Cancer incidence and risk factors after solid organ transplantation. *International journal of cancer*, 125(8):1747–1754, 2009.

Matthew G Vander Heiden, Lewis C Cantley, and Craig B Thompson. Understanding the warburg effect: the metabolic requirements of cell proliferation. *science*, 324(5930):1029–1033, 2009.

Donald A Vessey, Michael Kelley, and Robert S Warren. Characterization of the coa ligases of human liver mitochondria catalyzing the activation of short-and medium-chain fatty acids and xenobiotic carboxylic acids. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1428(2-3):455–462, 1999.

Donald Voet, Judith G Voet, and Charlotte W Pratt. *Fundamentals of biochemistry: life at the molecular level*. John Wiley & Sons, 2016.

POTTER VR. The biochemical approach to the cancer problem. In *Federation proceedings*, volume 17, pages 691–697, 1958.

Sejal Vyas, Elma Zaganjor, and Marcia C Haigis. Mitochondria and cancer. *Cell*, 166(3):555–566, 2016.

Douglas C Wallace. Mitochondria and cancer. *Nature Reviews Cancer*, 12(10):685–698, 2012.

Otto Warburg. On the origin of cancer cells. *Science*, 123(3191):309–314, 1956a.

Otto Warburg. On respiratory impairment in cancer cells. *Science*, 124(3215):269–270, 1956b.

Otto Heinrich Warburg. *The metabolism of tumours: investigations from the Kaiser Wilhelm Institute for Biology, Berlin-Dahlem*. Constable & Company Limited, 1930.

Esther Witsch, Michael Sela, and Yosef Yarden. Roles for growth factors in cancer progression. *Physiology*, 2010.

Jonci Nikolai Wolff, Neil J Gemmell, Daniel M Tompkins, and Damian K Dowling. Introduction of a male-harming mitochondrial haplotype via 'trojan females' achieves population suppression in fruit flies. *Elife*, 6:e23551, 2017.

Taichiro Yoshimoto, Daisuke Matsubara, Manabu Soda, Toshihide Ueno, Yusuke Amano, Atsushi Kihara, Takashi Sakatani, Tomoyuki Nakano, Tomoki Shibano, Shunsuke Endo, et al. Mucin 21 is a key molecule involved in the incohesive growth pattern in lung adenocarcinoma. *Cancer Science*, 110(9):3006–3011, 2019.

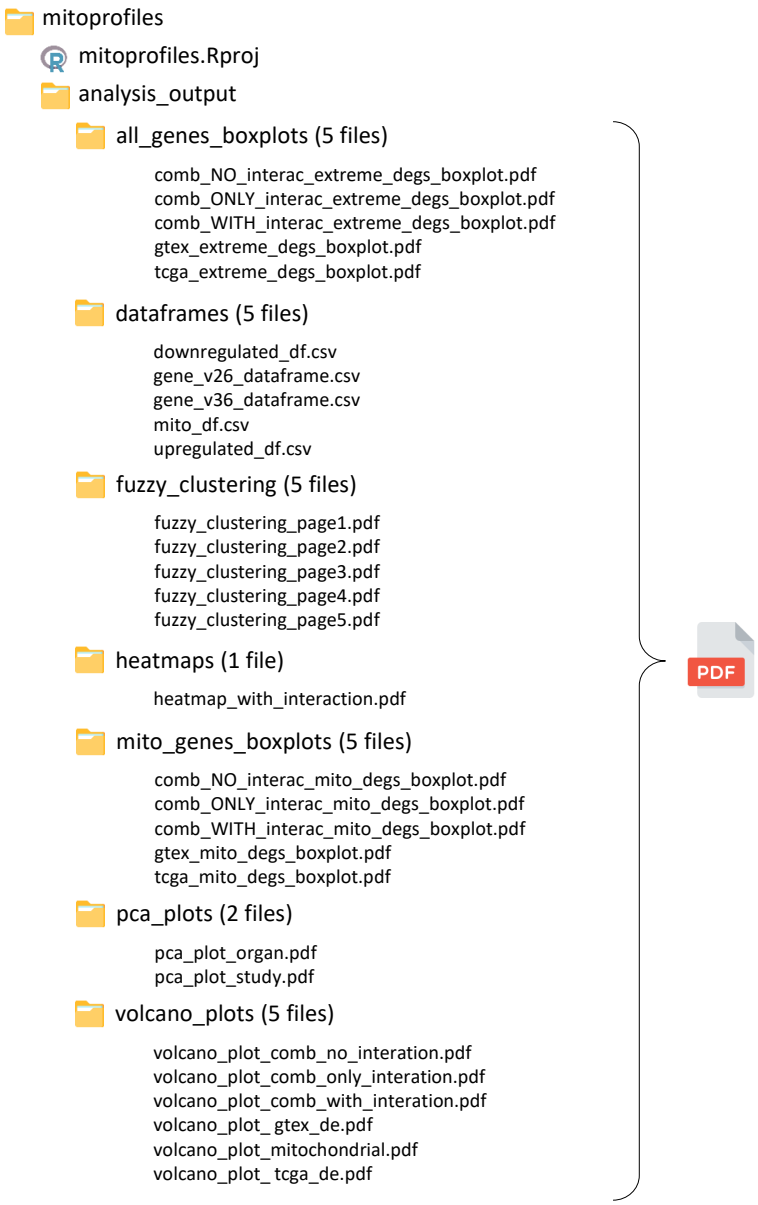
Hong Zhang, Wolfgang Holzgreve, and Christian De Geyter. Bcl2-l-10, a novel anti-apoptotic member of the bcl-2 family, blocks apoptosis in the mitochondria death pathway but not in the death receptor pathway. *Human molecular genetics*, 10(21):2329–2339, 2001.

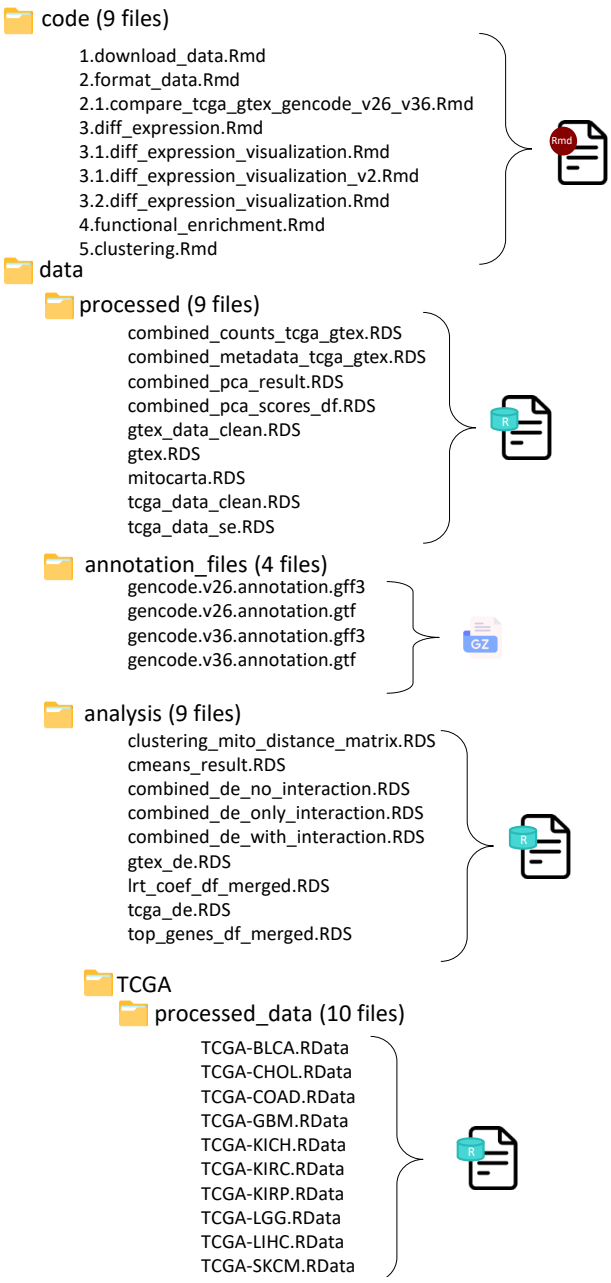
Song Zhang, Haitao Wan, and Xiaobo Zhang. Lncrna lhfpl3-as1 contributes to tumorigenesis of melanoma stem cells via the mir-181a-5p/bcl2 pathway. *Cell death & disease*, 11(11):950, 2020.

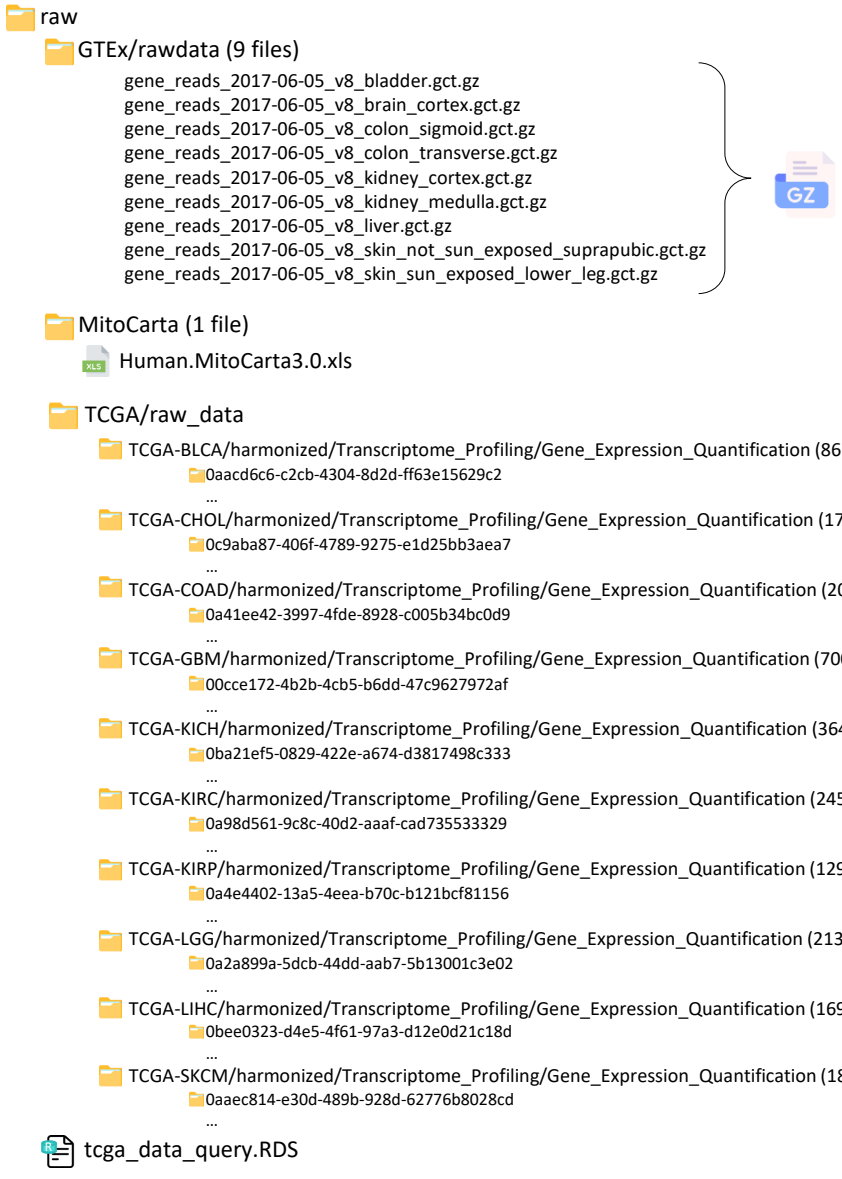
Part I
Appendices

Appendix A
Support work

A.1 Data analysis directory structure




code (9 files)
1.download_data.Rmd
2.format_data.Rmd
2.1.compare_tcga_gtex_gencode_v26_v36.Rmd
3.diff_expression.Rmd
3.1.diff_expression_visualization.Rmd
3.1.diff_expression_visualization_v2.Rmd
3.2.diff_expression_visualization.Rmd
4.functional_enrichment.Rmd
5.clustering.Rmd
Rmd
data
processed (9 files)
combined_counts_tcga_gtex.RDS
combined_metadata_tcga_gtex.RDS
combined_pca_result.RDS
combined_pca_scores_df.RDS
gtex_data_clean.RDS
gtex.RDS
mitocarta.RDS
tcga_data_clean.RDS
tcga_data_se.RDS
annotation_files (4 files)
gencode.v26.annotation.gff3
gencode.v26.annotation.gtf
gencode.v36.annotation.gff3
gencode.v36.annotation.gtf
GZ
analysis (9 files)
clustering_mito_distance_matrix.RDS
cmeans_result.RDS
combined_de_no_interaction.RDS
combined_de_only_interaction.RDS
combined_de_with_interaction.RDS
gtex_de.RDS
lrt_coef_df_merged.RDS
tcga_de.RDS
top_genes_df_merged.RDS
TCGA
processed_data (10 files)
TCGA-BLCA.RData
TCGA-CHOL.RData
TCGA-COAD.RData
TCGA-GBM.RData
TCGA-KICH.RData
TCGA-KIRC.RData
TCGA-KIRP.RData
TCGA-LGG.RData
TCGA-LIHC.RData
TCGA-SKCM.RData


raw
GTEx/rawdata (9 files)
gene_reads_2017-06-05_v8_bladder.gct.gz
gene_reads_2017-06-05_v8_brain_cortex.gct.gz
gene_reads_2017-06-05_v8_colon_sigmoid.gct.gz
gene_reads_2017-06-05_v8_colon_transverse.gct.gz
gene_reads_2017-06-05_v8_kidney_cortex.gct.gz
gene_reads_2017-06-05_v8_kidney_medulla.gct.gz
gene_reads_2017-06-05_v8_liver.gct.gz
gene_reads_2017-06-05_v8_skin_not_sun_exposed_suprapubic.gct.gz
gene_reads_2017-06-05_v8_skin_sun_exposed_lower_leg.gct.gz
GZ
MitoCarta (1 file)
Human.MitoCarta3.0.xls
TCGA/raw_data
TCGA-BLCA/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (863 files)
0aacd6c6-c2cb-4304-8d2d-ff63e15629c2
...
TCGA-CHOL/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (176 files)
0c9aba87-406f-4789-9275-e1d25bb3aea7
...
TCGA-COAD/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (2096 files)
0a41ee42-3997-4fde-8928-c005b34bc0d9
...
TCGA-GBM/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (700 files)
00cce172-4b2b-4cb5-b6dd-47c9627972af
...
TCGA-KICH/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (364 files)
0ba21ef5-0829-422e-a674-d3817498c333
...
TCGA-KIRC/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (2456 files)
0a98d561-9c8c-40d2-aaaf-cad735533329
...
TCGA-KIRP/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (1292 files)
0a4e4402-13a5-4eea-b70c-b121bcf81156
...
TCGA-LGG/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (2136 files)
0a2a899a-5dcb-44dd-aab7-5b13001c3e02
...
TCGA-LIHC/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (1696 files)
0bee0323-d4e5-4f61-97a3-d12e0d21c18d
...
TCGA-SKCM/harmonized/Transcriptome_Profiling/Gene_Expression_Quantification (1892 files)
0aaec814-e30d-489b-928d-62776b8028cd
...
tcga_data_query.RDS

Appendix B
Details of results

B.1 Differential Expression Analysis

B.1.1 Upregulated genes (all genes set)

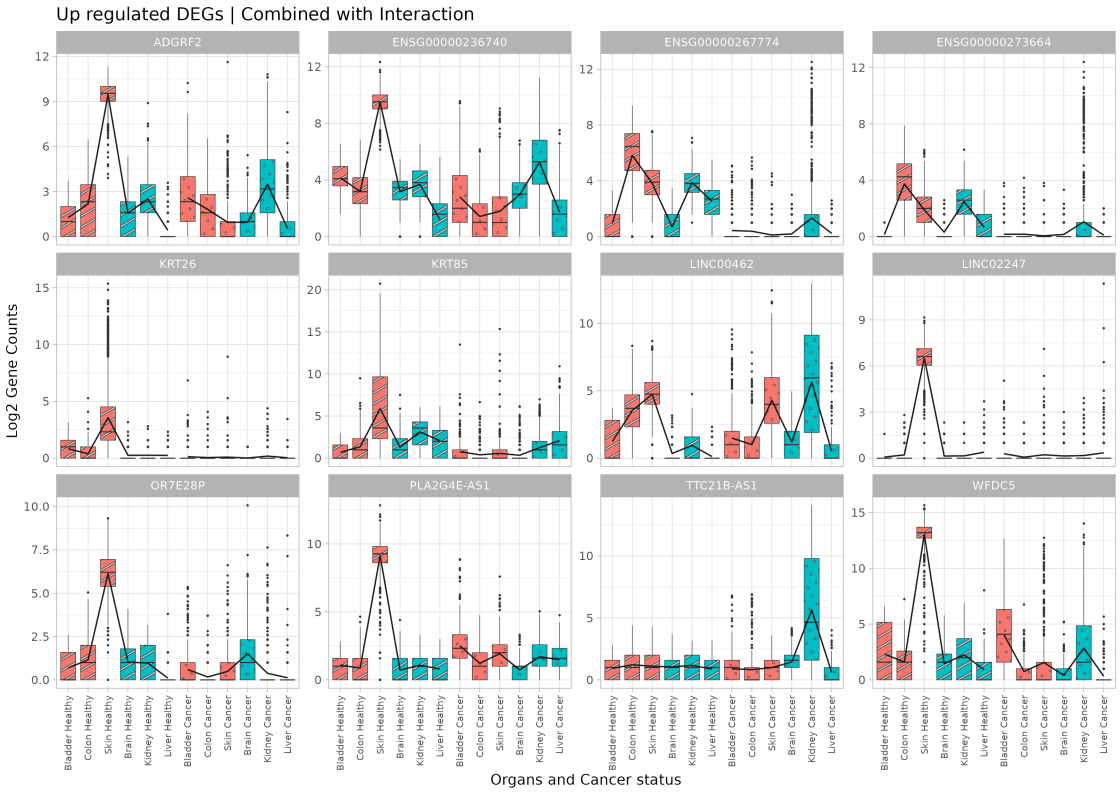


Figure 24: Boxplots of the most upregulated genes. Each plot represents the expression of a gene for the six organs studied (bladder, colon, skin, brain, kidney, and liver) in the two datasets used (**GTEx** and **TCGA**). The first six box plots correspond to **GTEx** data, with low metabolic rate organs marked in red and high metabolic rate organs in green. The following six plots are for **TCGA** data, and the arrangement of organs is the same.

B.1.2 Downregulated genes (all genes set)

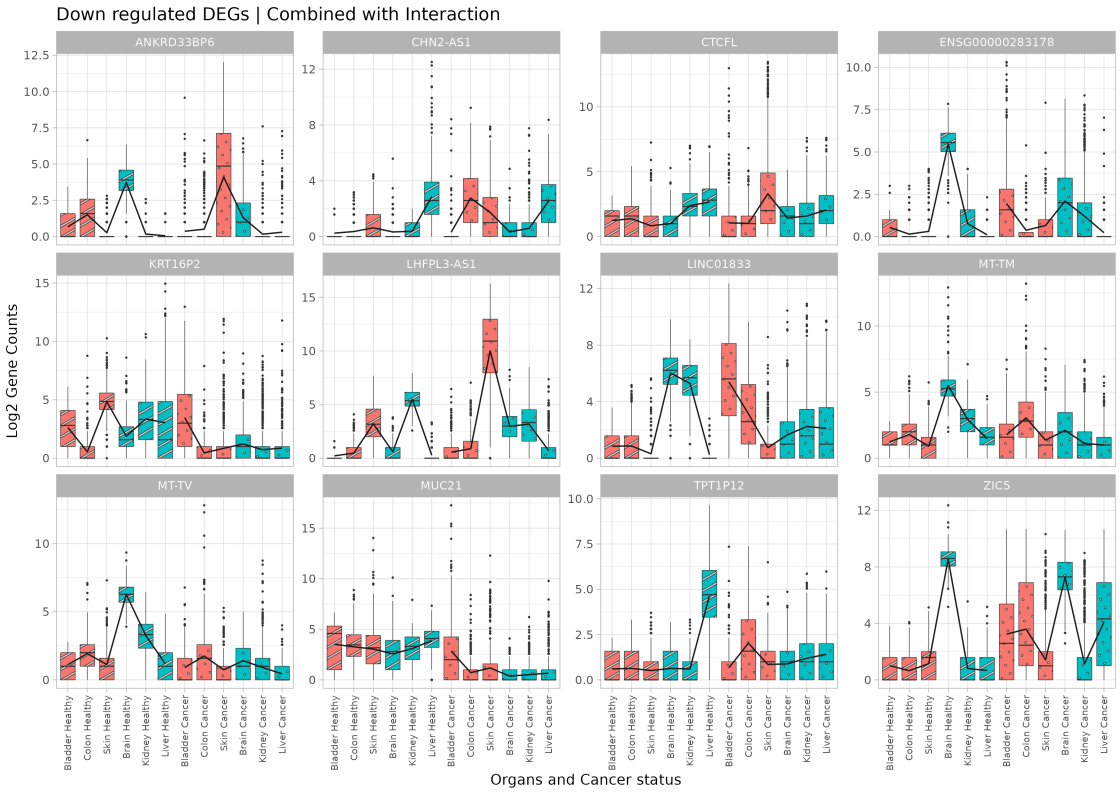


Figure 25: Boxplots of the most downregulated genes. Each plot represents the expression of a gene for the six organs studied (bladder, colon, skin, brain, kidney, and liver) in the two datasets used (**GTEx** and **TCGA**). The first six box plots correspond to **GTEx** data, with low metabolic rate organs marked in red and high metabolic rate organs in green. The following six plots are for **TCGA** data, and the arrangement of organs is the same.

B.1.3 Mitochondrial genes upregulated

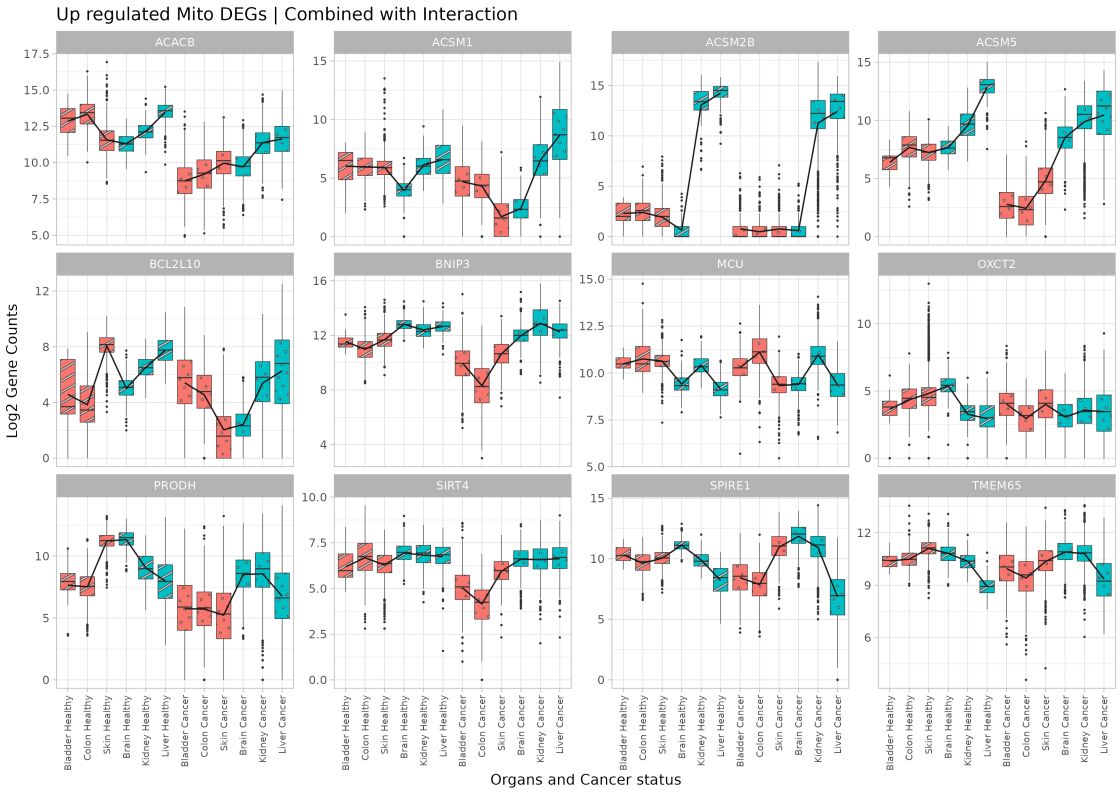


Figure 26: Boxplots of the most upregulated genes encoding mitochondrial proteins. Each plot represents the expression of a gene for the six organs studied (bladder, colon, skin, brain, kidney, and liver) in the two datasets used (**GTEx** and **TCGA**). The first six box plots correspond to **GTEx** data, with low metabolic rate organs marked in red and high metabolic rate organs in green. The following six plots are for **TCGA** data, and the arrangement of organs is the same.

B.1.4 Mitochondrial genes downregulated

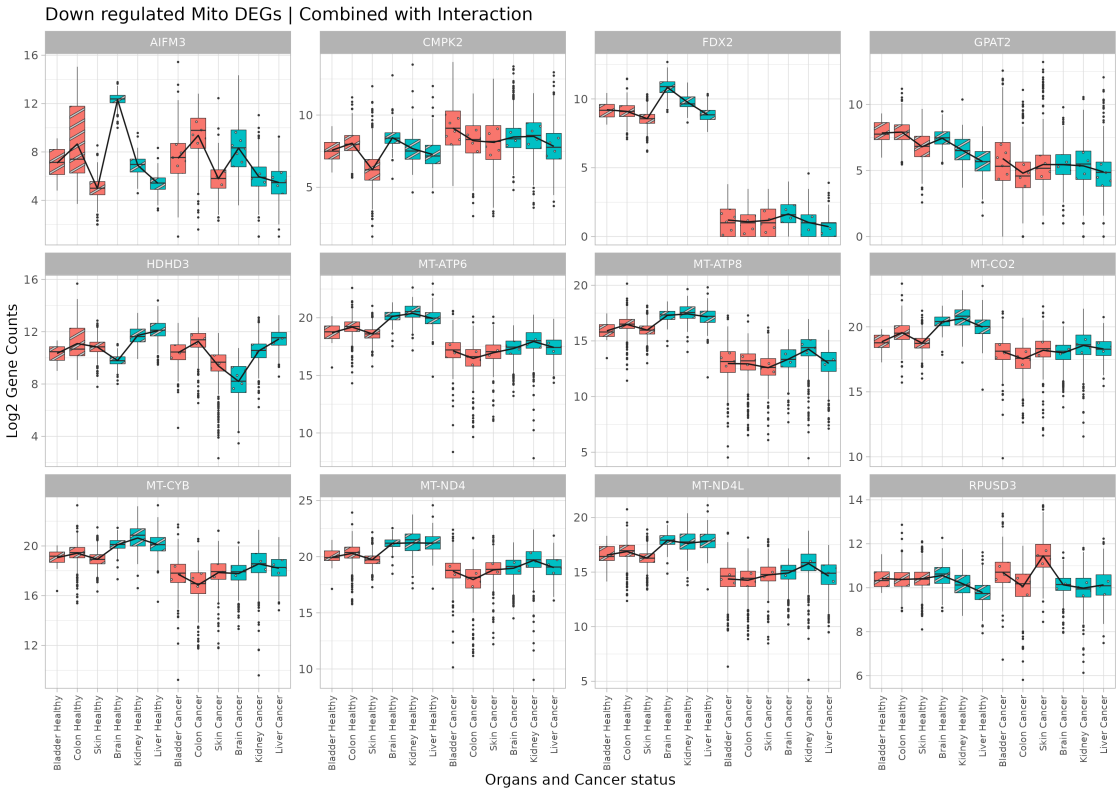


Figure 27: Boxplots of the most downregulated genes encoding mitochondrial proteins. Each plot represents the expression of a gene for the six organs studied (bladder, colon, skin, brain, kidney, and liver) in the two datasets used (**GTEx** and **TCGA**). The first six box plots correspond to **GTEx** data, with low metabolic rate organs marked in red and high metabolic rate organs in green. The following six plots are for **TCGA** data, and the arrangement of organs is the same.

B.2 Functional Enrichment

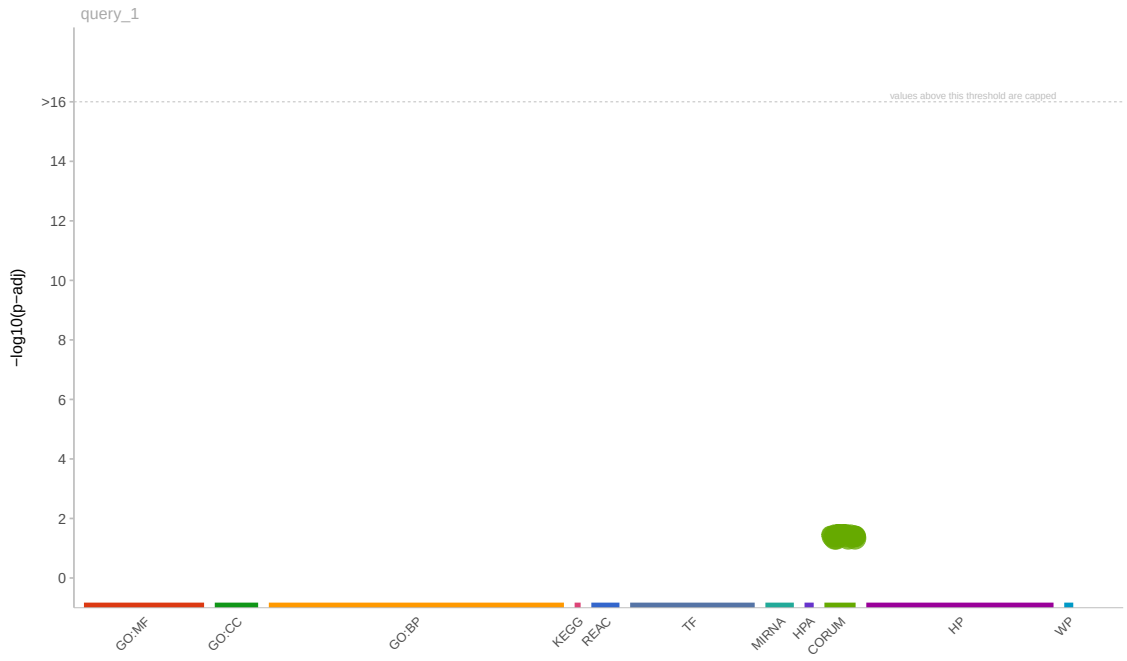


Figure 28: Functional enrichment analysis performed in this thesis.

B.3 Clustering

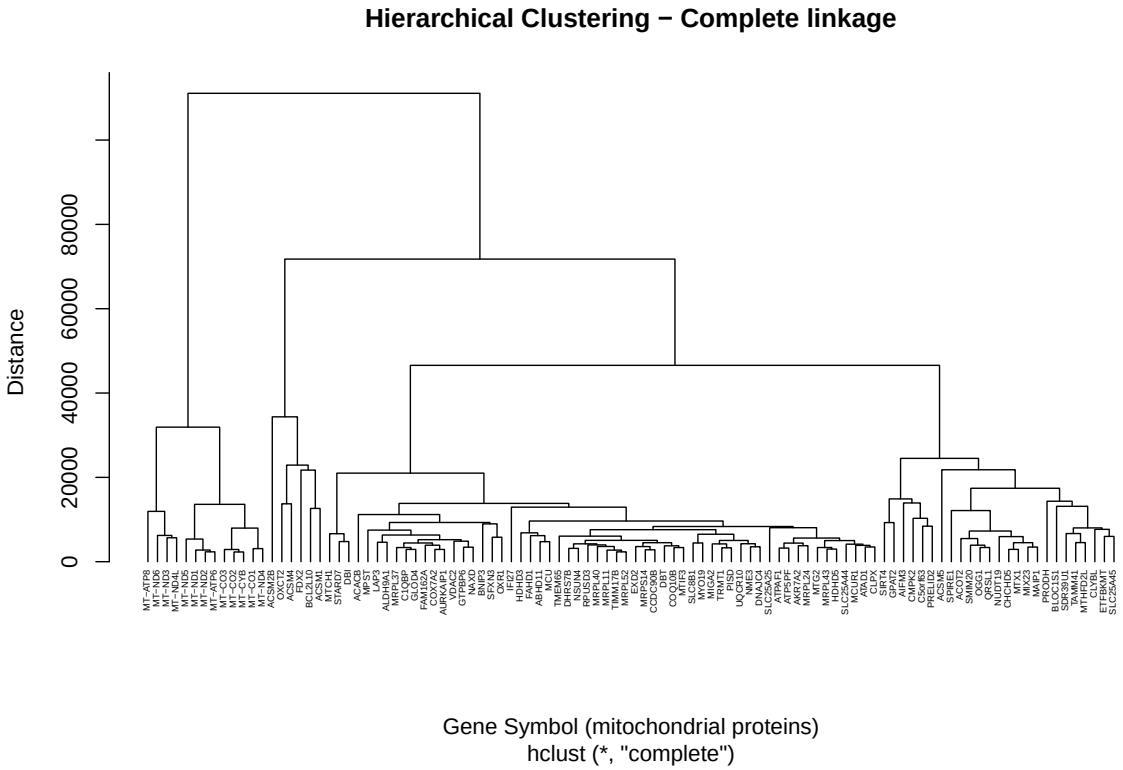


Figure 29: Dendrogram for the complete link.

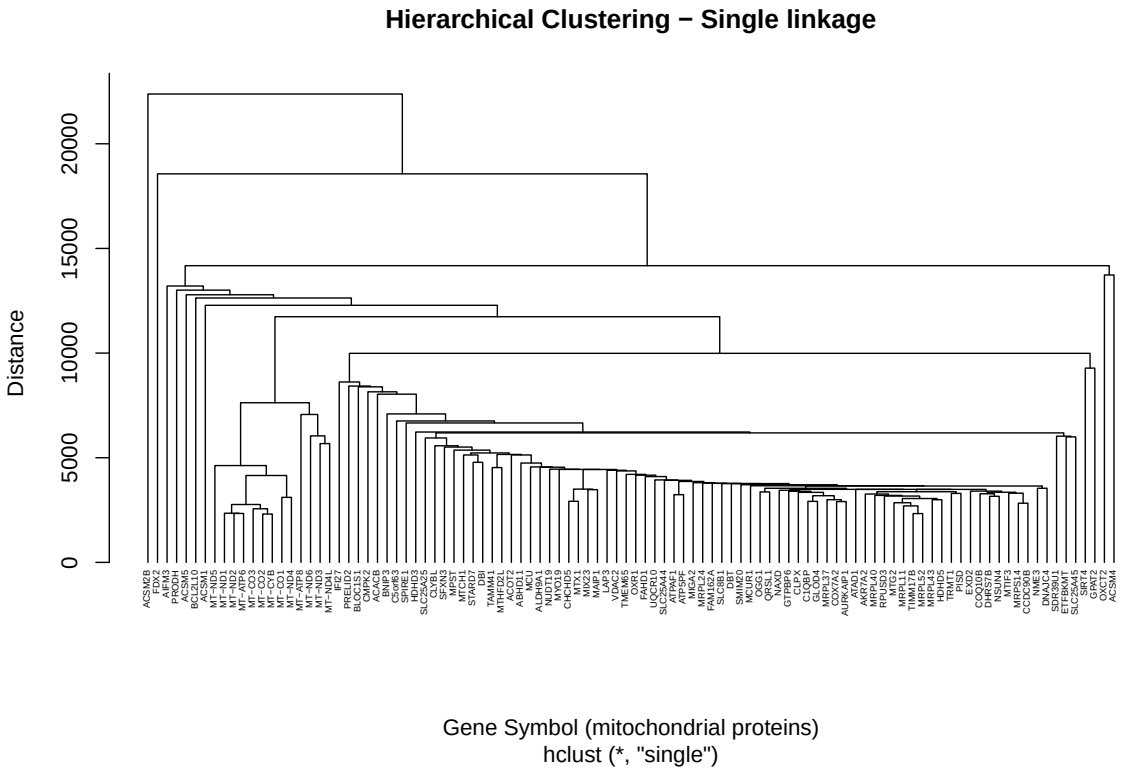


Figure 30: Dendrogram for the single link.

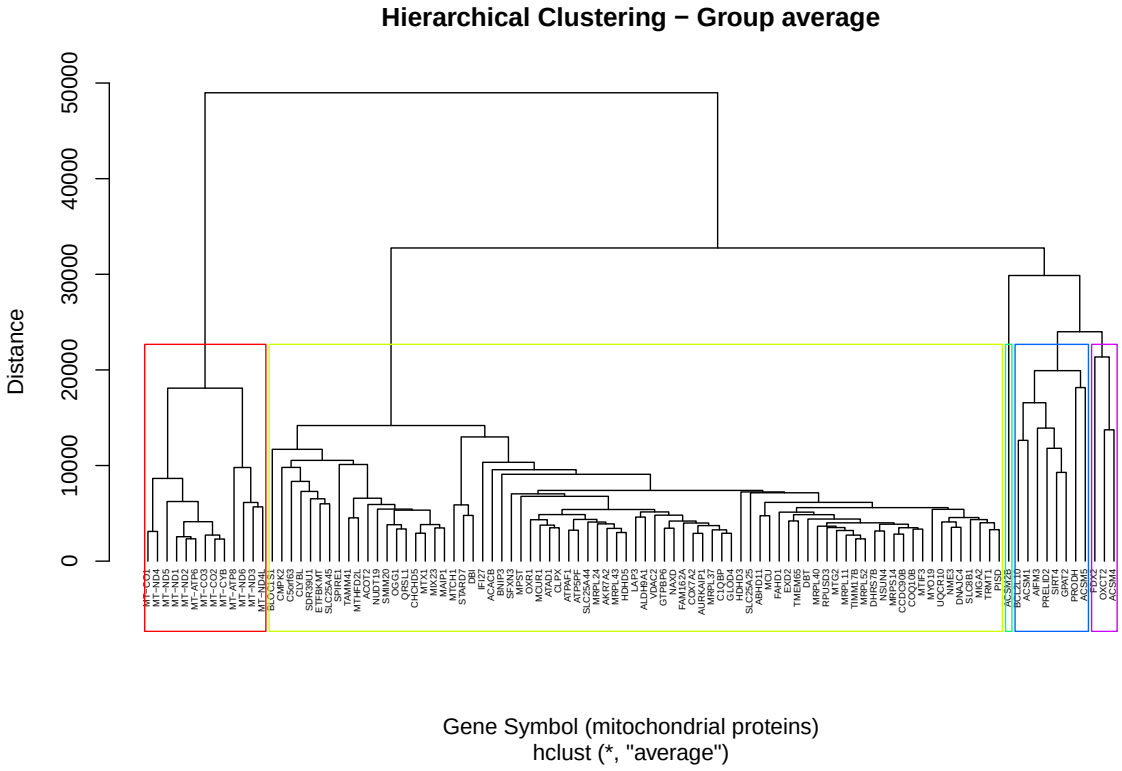


Figure 31: Dendrogram for the group average.

Appendix C
Software tools

C.1 R Packages

C.1.1 CRAN repository

here - enable easy file referencing in project-oriented workflows. The here package creates paths relative to the top-level directory.

tidyverse - helps to transform and better present data. It assists with data import, tidying, manipulation, and data visualization.

readxl – allows Reading data from Excel files and into R.

patchwork - combine separate ggplots into the same graphic in a simple and straightforward manner.

gprofiler2 - functional enrichment analysis and visualization, gene/protein/SNP identifier conversion and mapping orthologous genes across species.

RColorBrewer - provides color schemes for R graphics.

C.1.2 Bioconductor repository

TCGAbiolinks - the aim of TCGAbiolinks is : i) facilitate the GDC open-access data retrieval, ii) prepare the data using the appropriate pre-processing strategies, iii) provide the means to carry out different standard analyses.

SummarizedExperiment – provides the SummarizedExperiment class and associated methods allowing the setting and retrieval of data from SummarizedExperiment objects. These contain one or more assays, each represented by a matrix-like object of numeric or other mode, where the rows typically represent genomic ranges of interest and the columns represent samples.

edgeR - performs differential expression analysis of RNA-seq expression profiles with biological replication. Implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests.

AnnotationDbi - Implements a user-friendly interface for querying SQLite-based annotation data pack-

ages.

org.Hs.eg.db - Genome wide annotation for Human genomic features.

C.1.3 GitHub R packages

mitocarta - R data packege created within this project to provide access in R to the MitoCarta data, formated and clean.