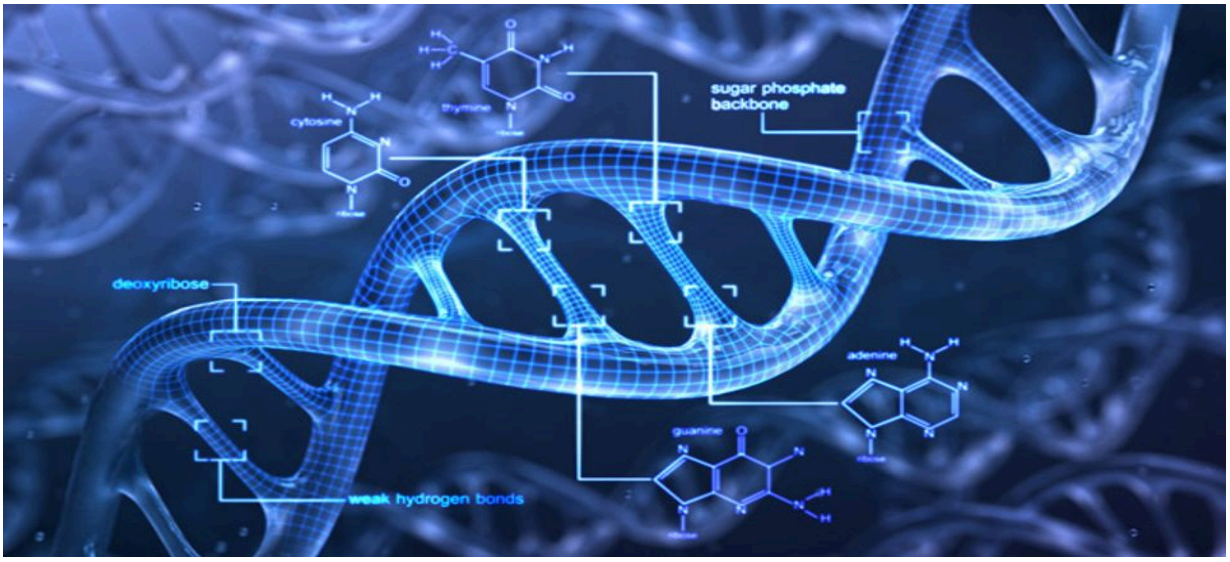
Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Exploratory Analysis of DNA Methylation Patterns in Colon Cancer Patients



The Role of DNA Methylation (Bose, 2022)

Supervisor's name: Dr. Nicholas P. Danks

MSc Business Analytics 2024

Submitted in partial fulfilment of the requirements of the examination for

MSc Business Analytics, Trinity College Dublin, July 2024.

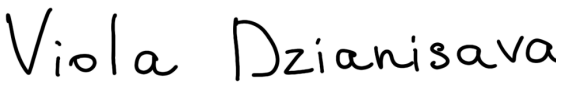
Author's Declaration

We have read the University's code of practice on plagiarism. We hereby certify this material, which we now submit for assessment on the programme of study leading to the award of MSc Business Analytics is entirely our own work and has not been taken from the work of others, save and to the extent that such work has been cited within the text of our work.

Student Name: Xijun Lin
Student ID Number: 23340637
Student Signature:

Date: 19th July 2024

Student Name: Viola Dzianisava
Student ID Number: 23359986
Student Signature:



Date: 19th July 2024

Student Name: Henry Lynam
Student ID Number: 23369189
Student Signature:

Date: 19th July 2024

Abstract

This dissertation presents a comprehensive analysis of DNA methylation patterns in colon cancer, focusing on their role in gene regulation and implications for disease progression and treatment. Utilising advanced computational tools, experimental methods and machine learning models, this study systematically investigates the alterations in the epigenetic landscape that lead to the dysregulation of gene expression, a hallmark of cancer pathology. Drawing on extensive genomic datasets from The Cancer Genome Atlas (TCGA), this research aims to delineate specific methylation profiles that correlate with tumour phenotype, progression, and patient prognosis.

The analytical approach includes a detailed examination of methylation data across various stages and subtypes of colon cancer, with the objective of identifying biomarkers that can predict disease outcomes and responses to therapy. By integrating methylation profiles with comprehensive clinical data, the project evaluates the prognostic significance of these epigenetic markers and their potential as targets for therapeutic intervention.

The findings of this dissertation aim to advance the understanding of epigenetic modifications in colon cancer and their practical applications in oncology. This study underscores the potential of epigenetic therapy as an integral component of personalised medicine, enhancing the efficiency and specificity of treatment regimens. By providing new insights into the molecular drivers of colon cancer, this research contributes to the development of more effective diagnostic tools and therapeutic strategies, ultimately aiming to improve clinical outcomes for patients afflicted by this debilitating disease.

Acknowledgements

The team would like to thank our professor Dr. Nicholas Danks, our dissertation supervisor, for his guidance over the course of this project and our MSc in Business Analytics programme. His technical insights enabled us to cover a broad range of analytical methodologies.

We would also like to thank Dr. Daithi Heffernan for pointing us in the direction of DNA Methylation Data analysis, for the theory guidance, and for always making the time for meetings when questions arose.

Finally, we sincerely appreciate the unconditional support and encouragement of our friends and families.

Table of content

List of Tables

List of Figures

Chapter 1: Introduction

Section 1.1 Overview

Sepsis is an extreme immune response brought on as a response to infection by a foreign agent. It is a primary cause of mortality/morbidity rates in surgical patients. It is characterised by an initial extreme case of inflammation, wherein the immune system attacks infected and surrounding areas, followed by a dampening of gene expression via epigenetic alterations to the genome that may or may not be recovered from over time. Sepsis is a complex condition influenced not only by internal biological factors such as Genetics, Epigenetics, Age, Health, but also based on clinical treatment and aftercare.

DNA Methylation is a major epigenetics mechanism in regulating immune response and maintaining the expression of genes, DNA Methylation is a major epigenetic mechanism in regulating immune response and maintaining the expression of genes, acting as a control for the level of gene expression. Non-standard methylation is implicated in many diseases, both as cause or as symptom. This is particularly observable in cancer patient data, and understanding how methylation patterns differ between subjects in different conditions could provide a deeper understanding of sepsis pathophysiology and reveal biomarkers to be used for early screening and new medical treatments.

In this dissertation, we aim to investigate the DNA Methylation patterns in patient data taken from the Cancer Genome Atlas (TCGA). By utilising various data processing, analytic, and visualisation techniques, we will explore the differences in how genes are methylated between normal and tumour tissue samples of colon cancer patients, particularly as it pertains to sepsis and sepsis-related concepts.

Section 1.2 Importance of studying DNA methylation and its role in cancer development

Epigenetics and Gene Regulation: Genetics is the study of the changes in the structure of particular genes and whether these gene sequences are altered. Alterations include: mutations, deletions, insertions and translocation. Whereas on the other hand, epigenetics studies

changes in the gene activity or function that are not associated with the change of the DNA sequence itself. DNA methylation is an epigenetic mechanism involving the transfer of a methyl group onto the C5 position of the cytosine to form 5-methylcytosine (Moore, Le and Fan, 2013).

DNA methylation, a critical epigenetic mechanism, plays a pivotal role in gene regulation by modulating the accessibility of DNA to transcriptional machinery. This process effectively controls which genes are activated ("switched on") or repressed ("switched off"), ensuring proper cellular function and development. The aberration in these methylation patterns is closely linked to the onset and progression of various cancers. As cells transform from a normal state to a cancerous one, specific changes in DNA methylation contribute to this transition by altering gene expression profiles, which can promote cellular proliferation and inhibit programmed cell death mechanisms (Jones and Baylin, 2007).

Epigenetics broadly encompasses the study of changes in gene expression that do not involve alterations in the underlying DNA sequence.The reversible nature of DNA methylation, with far-reaching effects on genome stability and cellular phenotypes, has been a key focus of research in this field. This analysis will delve into the associations noted within DNA methylation changes that may affect cancer development, highlighting its dual role of silencing tumour suppressor genes through hypermethylation and activating oncogenic pathways through hypomethylation (Feinberg, Ohlsson, & Henikoff, 2006). This thesis will analyse how DNA methylation could act as a key regulatory mechanism and serve as a potential therapeutic target. By understanding the specific methylation patterns associated with different types of cancer, including colon cancer, researchers can develop targeted epigenetic therapies aimed at restoring normal methylation patterns, thereby inhibiting cancer progression and potentially reversing malignant phenotypes (Baylin and Jones, 2011).

This section highlights the need to study DNA methylation in the broader context of cancer research and its potential impact on the development of novel therapeutic strategies.

Section 1.3 Background information

Section 1.3.1 Definition of cells and DNA

To understand DNA methylation, reference to the DNA and the cell needs to be made first.

Cells are a basic form of life. All the living organisms on earth consist of enormous quantities of cells, making them the building blocks working together to maintain life.

Initially, all cells in the body are the same, the entire genome is in every cell, but only certain parts of the genome are active in specific cells to perform different functions due to the fact that different genes are expressed in different cells (Costa and Johannes, 2020).

The cell contains DNA and RNA that expresses the genetic code. DNA holds all the information that is needed to build new cells and maintain their living. RNA is responsible for maintaining the expression of the information that is stored in DNA. To help with the cell division process cells use proteins that are made from amino acids (*What Is a Cell*, no date). The production of protein starts with transcription - it is a process where a portion of the cell's DNA becomes a template creation of an RNA molecule (*The Information in DNA Is Decoded by Transcription*, no date). And then continues with translation - it is a process where mRNA becomes a template for protein assembly (*The Information in DNA Determines Cellular Function via Translation*, no date). During cell's division, DNA is replicated and passed on to new cells, ensuring each cell has the same genome (*Gene Expression*, no date).

DNA is organised into cells the following way. Wrapped DNA around the proteins form histones, collection of which is called the nucleosome as can be seen on figure 1. Histones contain genetic code. All nucleosomes together form chromatin. The process of chromatin opening determines which genes will become active and inactive. Active genes (that are switched "on") determine what type the cell it becomes or how it functions - for example, will it be a liver, lung or the hair cell. Both DNA methylation and histone modification are involved in this process and establish patterns of gene repression during development (Cedar and Bergman, 2009). Histone modification shows how much the whole genome opens up or not.

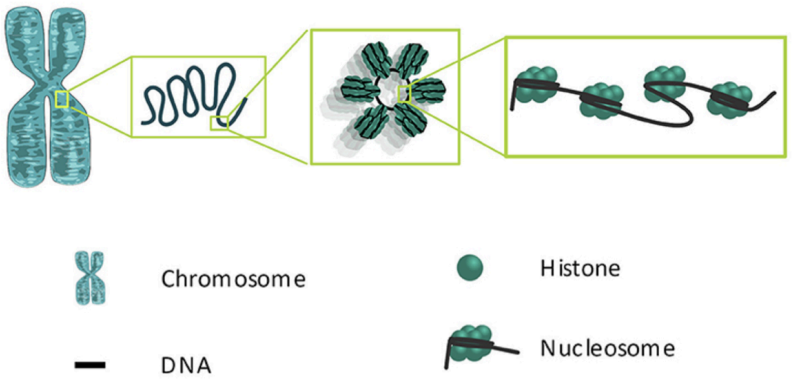


Figure 1: DNA explanation (Costa and Johannes, 2020)

Section 1.3.2 DNA methylation

DNA methylation regulates gene expression by recruiting proteins involved in gene repression (Moore et al., 2012). It includes controlling gene expression without changing the DNA sequence by appending a methyl group to a cytosine residue in a CpG dinucleotide at its 5' position (Lanata et al., 2018). These methyl groups can turn off genes, resulting in these genes stopped being used. Methyl groups are added to precise locations in the DNA, usually where a cytosine base is followed by a guanine base. When they are added to the DNA they can turn genes off, as a result the cell doesn't read these genes and it doesn't produce the protein that these genes normally produce. DNA methylation controls how easy it is for individual base pairs to get transcribed through RNA into protein. Which means that if the cell is methylated, the base pairs stick together or become nonfunctional and the gene cannot open up.

It is significant that DNA methylation works correctly to ensure that normal development of the cells occurs and they perform their normal functions. For example, the liver cell should make liver related proteins. When errors happen in the DNA methylation process, certain genes get turned off incorrectly or in the wrong timing. This can lead to the development of a malignant cell and ultimately into cancer, in which the cell's control system breaks which leads to abnormal growth (Moore et al., 2012). As said in the article "DNA Methylation and Its Basic Function", written by Moore, Le and Fan in 2013 - DNA methylation is also very

essential for silencing retroviral elements, regulating tissue-specific gene expression, genomic imprinting, and X chromosome inactivation.

Section 1.3.3 Immune system

Immune system plays a very important role in a person's well-being. It protects the body from harmful diseases that can attack from outside as well as inside. Immune system is always running in the background of our bodies and as long as it is working, we don't notice it, but when the body comes across a new disease it has never seen before, in some cases the person might become ill (Institute for Quality and Efficiency in Health Care, 2023). Immune system also plays a big role in recognising and destroying the cancer cells within the body.

How does the immune system recognize cancer cells?

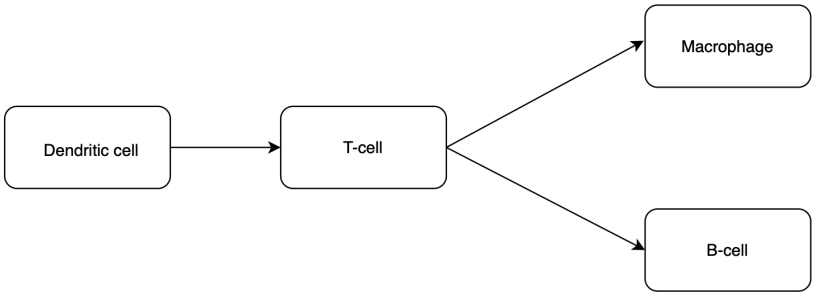


Figure 2: Immune system recognize cancer cells progress

Immune system response starts with dendritic cells. They are antigen cells that capture pathogens and receive signals from pathogens that influence the outcome of immune responses (Liu, 2016). They call for help from the T-cells that are pure immune cells, and are capable of controlling both cancer and infections within the body. Immune system searches through all the T-cells in the body and finds the right matching T-cell that can destroy the specific bacteria mentoring the body. T-cells and T-lymphocytes cells are immune cells that among other things identify cancer cells on a regular basis and try to remove them. This is why if there happens to be an association in the abnormalities in the T-cell functioning in the patients with colon cancer, it can be said they influence the cancer appearance.

Section 1.3.4 Metastasis

Metastasis marks the stage at which cancer cells spread from their original tumour site to other parts of the body. Understanding the underlying mechanisms that facilitate this process plays a key role in the development of effective treatments for cancer prognosis. Recent research advances suggest that DNA methylation plays a key role in regulating genes involved in metastasis.

Epigenetic Regulation of Metastatic Pathways: DNA methylation alters the expression of various genes associated with cell adhesion, migration, and invasion—the key processes involved in metastasis. Specifically, hypermethylation of tumour suppressor genes often leads to reduced expression of proteins that inhibit metastasis, thereby enabling cancer cells to detach, invade adjacent tissues, and eventually disseminate to distant organs (Casalino and Verde, 2020).

Impact on the tumour microenvironment: The tumour microenvironment, which includes the various cells surrounding the tumour, has a considerable impact on the metastatic potential of colon cancer cells. Changes in DNA methylation in cancer cells can secrete factors that alter the microenvironment and thus support invasion and metastasis. For example, methylation-induced silencing of certain genes in colon cancer cells leads to increased secretion of pro-metastatic cytokines and growth factors, thereby preparing for distant tumour colonisation (Guo et al., 2017).

Section 1.4 Objectives of this dissertation

The aim of this study is to provide an in-depth analysis of DNA methylation patterns in colon cancer patients and to explore their role in disease progression and treatment. Specific objectives include:

1. **Correlation of methylation values for solid normal and tumour tissue:** To determine whether there is a correlation between methylation values for genes, associated with colon cancer, to check how the average values for each gene differ for normal and tumour tissue. This objective is based on the hypothesis that there is a difference in gene expression between 2 categories of tissues.

2. **Assessing the prognostic significance of methylation patterns:** Integrating methylation data with clinical outcome data to assess the prognostic significance of these methylation markers. Explore the relationship between methylation levels and response to treatment through experimental methods, e.g. with the help of survival analysis and multivariate regression models.
3. **Investigating methylation pathways and modules:** Enriching genetic data with biological pathway data to establish methylation pathway network maps from which to identify clusters of genes associated by function or interaction, and comparing networks built under different conditions to identify differentially methylated genes and different gene module manifestation.

Section 1.5 Report overview

In the first section **"**Introduction" the purpose of the project and all the needed domain knowledge is described to allow better understanding of the biology sphere for non-medical readers. The topic of DNA methylation is very broad. To narrow it down literature review was conducted to search for the similar research done in the analysis of DNA methylation and description of technical approaches used. The findings were then noted in the "Literature review" section.

The third section **"**Methodology**"** covers the TCGA data extraction process and cleaning steps. Since different analyses were performed in parallel, there are also multiple cleaning steps described. In this section overview of two datasets is performed to show the key characteristics of the data.

The fourth section **"**Findings and Results" focuses on the analysis and results, presenting graphs and tables for an easier understanding. The analysis was split into 3 parts, first starting from a defined set of genes (26 and 17 genes respectively) to then expanding to the top 5000 genes by variance.

The last chapter "Conclusions" focuses on results summary, also done in three parts per corresponding analysis and states the report limitations and further areas to expand the research.

Chapter 2: Literature Review

Section 2.1 Overview of colon cancer

Section 2.1.1 Introduction

Colon cancer forms in the tissues of the colon. Most colon cancers are adenocarcinomas, which begin in cells that make and release mucus and other fluids (*NCI Dictionary of Cancer Terms*, no date). It is one of the most common types of cancers diagnosed in the United States (*Cancer facts and statistics*, no date). Global statistics from the World Health Organization (2022) also highlight its prevalence, emphasising the need for effective screening and early detection strategies.

Colon cancer typically starts as small, noncancerous (benign) clumps of cells called polyps that form on the inside of the colon or rectum. Over time, genetic mutations within these polyps can accumulate, leading to their transformation into cancer (Siegel *et al*., 2020). Inherited mutations in some genes are commonly associated with colon cancer, especially in inherited disorders such as Lynch syndrome and familial adenomatous polyposis (Lynch and De La Chapelle, 2003).

Section 2.1.2 Epigenetic modifications and risk factors

One of the critical factors in the development of colon cancer is DNA methylation, an epigenetic modification involving the addition of a methyl group to DNA. This process plays a significant role in the regulation of gene expression. Abnormal methylation patterns, such as hypermethylation of tumour suppressor gene promoters and hypomethylation of oncogenes, are frequently observed in colon cancer (Jones and Baylin, 2007).

The process of hypermethylation involves the silencing of promoter regions of tumour suppressor genes. By inactivating crucial genes, which are involved in cell cycle regulation, DNA repair, and apoptosis, hypermethylation further leads to cancer development (Feinberg *et al*., 2006).

The process of hypomethylation can activate oncogenes and lead to genomic instability. This is associated with increased expression of genes that lead to cell proliferation and cancer progression (Feinberg *et al*., 2006).

Risk factors for colon cancer include age, with a higher incidence in individuals over 50, hereditary genetic disorders, as well as dietary factors, smoking and lack of physical activity. WCRF International (2018) claims that diets high in red and processed meats and low in fibre are linked to an increased risk of colon cancer.

Section 2.1.3 Diagnosis, prognosis, and future research

Early detection of colon cancer significantly improves the prognosis of patients. Screening method colonoscopy allows for the identification and removal of precancerous polyps, effectively preventing the development of cancer. Other diagnostic tools include faecal occult blood tests, flexible sigmoidoscopy, and CT colonography. Therefore, the American Cancer Society (2022) recommends that individuals at average risk begin screening at age 45.

Besides, the prognosis for colon cancer varies widely based on the cancer stage at diagnosis. Early-stage cancers have a high survival rate, while advanced cancers with distant metastases have a considerably lower survival rate. Prognostic factors include tumour size, lymph node involvement, and the presence of metastases, as well as the patient's overall health and response to treatment (*NCCN guidelines for patients: Colon cancer*, 2022).

Current research is exploring the use of biomarkers for early diagnosis and the development of novel therapeutic agents that target specific molecular pathways involved in cancer development. Additionally, there is an increasing understanding of the role of the microbiome in colon cancer development and its significant potential as a therapeutic target (Smith *et al*., 2021).

In summary, colon cancer is a major health concern due to its high prevalence and significant impact on mortality rate. Understanding the epigenetic modifications, risk factors, and advancements in diagnosis and treatment is crucial for improving patient outcomes. Continued research into biomarkers and novel therapies holds promise for more effective prevention, as well as early detection, personalised treatment strategies, which ultimately reduce the burden of this disease.

Section 2.2 DNA methylation genes description

The explanation of DNA methylation genes heavily relies upon the article - Moore, Le and Fan (2013) and others referred.

Dnmt (DNA Methyltransferases) family genes that catalyse the addition of methyl groups onto DNA (Dnmt1, Dnmt3a, and Dnmt3b genes) and control DNA methylation are analysed in the methylation gene analysis. DNA methylation, catalysed by the DNMTs, plays an important role in maintaining genome stability. Abnormal levels of DNMTs and changes in DNA methylation patterns are strongly linked to various types of cancer, though the exact reasons for this connection are not fully understood (Jin and Robertson, 2012). Dnmt3a and Dnmt3b genes are referred to as de novo Dnmt because they have the ability to add a new methylation pattern to unmodified DNA. Results, shown in the paper, illustrate that Dnmt3b is required during early development and Dnmt3a is required for normal cellular differentiation. Dnmt1 copies the parental DNA strand's DNA methylation pattern onto the newly synthesised daughter strand during DNA replication. Dnmt1 attaches itself to the freshly synthesised DNA and methylates it in a way that exactly resembles the pre-DNA replication methylation pattern (Hermann *et al*., 2004). Dnmt1 can also repair DNA methylation, according to Mortusewicz *et al*. (2005) and is essential for both cell division and cellular differentiation.

DNA methylation is recognized by three separate families of proteins: the MBD proteins, the UHRF proteins, and the zinc-finger proteins (ZBTB4, and ZBTB38). The MBD family includes MeCP2, the first identified methyl-binding protein, with MBD1,MBD2, MBD3, MBD4 (Meehan *et al*, 1989; Lewis *et al*, 1992; Hendrich and Bird, 1998). They give instructions for creating proteins that control gene activity (expression) by modifying chromatin, which is the combination of DNA and protein that organises DNA into chromosomes (*MBD5 gene: MedlinePlus Genetics*, no date). MBD genes are important for normal neuronal development and functioning (Amir *et al*, 1999).

The UHRF proteins, including UHRF1 and UHRF2, are multidomain proteins that flip out and bind methylated cytosines via a SET- and RING-associated DNA-binding domain (Hashimoto *et al*, 2008, 2009). The primary function of UHRF proteins is not to bind to DNA and repress transcription, they first bind to Dnmt1 and then target it to hemimethylated DNA

in order to maintain DNA methylation, especially during DNA replication (Sharif *et al*, 2007; Bostick *et al*, 2007; Achour *et al*, 2008). Using their functional domains, they interact with other molecules and act as central points in regulatory networks for important biological processes, such as maintaining DNA methylation and repairing DNA damage (Unoki and Sasaki, 2022).

Zinc-finger domain proteins - Kaiso, ZBTB4, and ZBTB38 (Prokhortchouk *et al*, 2001; Filion *et al*, 2006) repress transcription in a DNA methylation-dependent manner (Prokhortchouk *et al*, 2001; Yoon *et al*, 2003; Filion *et al*, 2006; Lopes *et al*, 2008). Zinc finger proteins are common in eukaryotic genomes and have many functions: they help recognize DNA, package RNA, activate transcription, regulate cell death, fold and assemble proteins, and bind lipids (Laity, Lee and Wright, 2001).

Section 2.3 Pathways, weight gene analysis, hub genes, gene clusters

Section 2.3.1 Pathways

Biological pathways are the sequences of actions taken by intracellular molecules to perform functions, induce changes, or create molecules. These pathways include multiple types such as metabolic pathways, signal transduction pathways, and gene regulation pathways. Metabolic pathways involve chemical reactions within a cell, transforming molecules through a series of enzyme-mediated steps. Signal transduction pathways enable cells to respond to external signals through a cascade of molecular events that ultimately result in a cellular response. Gene regulation pathways control the expression of genes, determining when, where, and how much gene product is made.

The integrity of these pathways is essential for proper biological function, and disruptions can lead to or result from diseases. For example, alterations in signal transduction pathways are implicated in various cancers due to their role in cell growth and division control (Hanahan & Weinberg, 2011). Disruptions in metabolic pathways can lead to metabolic disorders such as diabetes (Ashcroft & Rorsman, 2012).

Section 2.3.2 Gene variance

In data analysis, variance refers to the degree of dispersion or variation in a set of values. In the context of this data, variance measures how much the expression or methylation levels of a particular gene vary across different samples. High variance in gene expression indicates significant differences in how a gene is expressed in different conditions or among different individuals, suggesting it may play a role in responding to environmental or biological changes.

Section 2.3.3 Gene modules

Gene modules are groups of interconnected genes identified via network analysis. By selecting the top 5,000 genes by variance, we can focus on genes that are likely biologically significant, especially when working with limited computational resources. These modules often correspond to specific biological pathways, as genes within the same module are likely to be co-regulated and functionally related (Langfelder & Horvath, 2008).

Comparing gene modules from tumour and normal datasets can reveal differences in gene connectivity. Visual representations of gene connectivity using heatmaps show regions of high connectivity (red), indicating strong interactions between genes. These patterns can help identify critical pathways and potential targets for therapeutic intervention (Zhang & Horvath, 2005).

Section 2.3.4 Weighted gene co-expression network analysis (WGCNA)

Weighted Gene Co-expression Network Analysis (WGCNA) is a data mining and analysis technique used to analyse high-dimensional datasets, particularly genomic data. WGCNA functions include clustering, data reduction, feature selection, and exploratory analysis. Its primary strength in bioinformatics is its ability to integrate genetic data with other biological markers based on weighted correlations, enabling the mapping of gene pathways as well as ancillary pathways influenced by non-genetic factors, such as epigenetic methylation data (Langfelder & Horvath, 2008).

A standard WGCNA can identify:

- **Gene Co-expression Modules:** Clusters of genes with highly correlated expression patterns. Genes in a module are often co-regulated or functionally related.
- **Hub Genes:** Highly connected genes within a module that regulate the module's combined functions. Changes in the expression of hub genes significantly impact the module's behavior.
- **Non-Genomic Trait Correlations:** Relationships between gene modules and external data, such as clinical data, revealing associations not evident in the genetic data alone.
- **Functional Annotation:** Using Gene Ontology to identify pathways and relationships between modules, including their functions and regulatory mechanisms.

Section 2.3.5 Methylation beta value-enriched WGCNA

By enriching a WGCNA with DNA methylation beta values instead of more standard pathways used , the analysis shifts to identifying co-methylated modules. These modules are clusters of co-methylated CpG sites, likely representing regions of shared methylation regulation or function. The Methylated WGCNA also measures connectivity within modules to identify hub CpG sites, which are key regulators of methylation patterns within a module (Horvath, 2011).

From these clusters and hubs, we can derive insights related to immune system interactions:

- **Immune Response Interactions:** Identifying modules associated with reactions to specific diseases can inform us about the epigenetic nature of immune responses, including natural immune responses and those triggered by injury or illness.
- **Trait Association:** Associating traits such as ageing or health conditions with specific methylation patterns can provide insights into the biological underpinnings of these traits.
- **Other Biological Functions:** Annotating modules, hubs, and networks with Gene Ontology helps explore the impact of methylation on pathways, revealing potential knock-on effects or pathway restructuring (Zhang et al., 2013).

By leveraging these methodologies, we can enhance our understanding of the complex interplay between genetics, epigenetics, and disease, ultimately leading to more targeted and effective therapeutic strategies.

Section 2.4 Previous research on DNA methylation in cancer

Section 2.4.1 Key findings

Research over the past few decades has profoundly expanded human understanding of DNA methylation and its critical role in gene regulation and cancer development. DNA methylation, a process involving the addition of a methyl group to the cytosine residues in CpG dinucleotides, influences gene expression by modifying the accessibility of DNA to transcription machinery. In 2007 Jones and Baylin found that this epigenetic alteration is essential for normal cellular function, and its disruption is frequently associated with cancer, leading to abnormal cell growth and differentiation.

More recently, the association between aberrant methylation patterns and cancer has been a focal point of numerous studies. One of the earliest and most consistent findings is the hypermethylation of promoter regions leading to the silencing of tumour suppressor genes, this gene silencing can affect crucial cellular processes, including apoptosis, DNA repair, and cell cycle regulation, thereby promoting uncontrolled cellular proliferation (Esteller, 2008). For instance, the hypermethylation of the MGMT gene promoter in glioblastoma patients correlates with a favourable response to alkylating agents (Hegi *et al*., 2005), demonstrating that methylation patterns can serve as biomarkers for predicting treatment outcomes.

Besides, the potential of DNA methylation patterns as biomarkers extends beyond treatment prediction to early cancer detection and prognosis. However, the variability in methylation patterns among individuals is a significant challenge. This inter-individual variability complicates the creation of universal diagnostic and therapeutic strategies, necessitating personalised approaches to cancer treatment (Eckhardt *et al*., 2006).

Moreover, technological advances have significantly enhanced our ability to detect and quantify methylation patterns. Next-generation sequencing (NGS) and bisulfite sequencing have revolutionised the field by providing comprehensive coverage and high-resolution maps of methylation at the genome-wide level. These technologies have facilitated a deeper understanding of the methylation landscape in various cancers, enabling the identification of key epigenetic changes associated with tumorigenesis (Bock *et al*., 2008), thus people can better observe and analyse their epigenetic changes.

Longitudinal studies have found that methylation patterns change over time, a phenomenon known as "epigenetic drift" (Milicic *et al*., 2023). This drift is influenced by ageing and may affect cancer risk, highlighting the importance of temporal dynamics in methylation studies. Understanding these changes is therefore important for developing effective cancer prevention strategies (Fraga *et al*., 2005).

Furthermore, environmental factors, including diet, stress, and exposure to toxins, also influence DNA methylation (Feil and Fraga, 2012). Studies of these influences have had varying degrees of impact on cancer prevention and treatment, illustrating the complex interplay between genetics and the environment, and therefore cancer research needs to integrate genetic and environmental factors.

Further, the application of machine learning techniques to large methylation datasets has further revolutionised cancer research. Machine learning models can identify patterns that predict cancer types and outcomes more accurately than traditional methods. These computational approaches have enabled researchers to interpret complex biological data, leading to more precise and predictive cancer diagnostics and prognostics (Babu *et al.*, 2008). Analysing large methylation datasets using different models facilitates researchers to derive more accurate potential information about methylation by looking at different model results.

In general, the large body of research on DNA methylation in cancer has highlighted its critical role in gene regulation, cancer development and progression. Advances in detection technologies have enabled more detailed and accurate mapping of methylation patterns, while data integration and computational modelling have led to a deeper understanding of the biological significance of methylation. Identifying specific methylation changes associated with different cancer types has helped develop biomarkers for early detection, prognosis and personalised treatment strategies that could save more patients in the future. Furthermore, continued research in this area is expected to shed further light on the complexity of cancer and improve patient prognosis through personalised and targeted interventions.

Section 2.4.2 Limitations

Although there have been many important advances in DNA methylation and its role in cancer in the last decade, there are still limitations that prevent the full potential of this research from being realised.

Firstly, the complexity of methylation patterns is a great challenge for researchers. Methylation is highly context-dependent and influenced by a variety of factors such as tissue type, developmental stage and environmental exposure (Feinberg *et al*., 2006). Due to this complexity, it is often difficult to distinguish between pathogenic methylation changes that lead to cancer and those that are simply a consequence of disease. Furthermore, distinguishing between benign and malignant changes in methylation patterns remains difficult. Not all methylation changes lead to functional consequences, and some may be part of the normal ageing process rather than indicators of cancer (Fraga *et al*., 2005). Thus, identifying specific methylation markers that are universally applicable to different cancers remains a difficult research problem.

Secondly, there is considerable individual variation in methylation patterns. Factors such as genetic background, age and lifestyle contribute to this variation, complicating the development of standardised diagnostic and therapeutic tools (Eckhardt *et al*., 2006). Personalised therapeutic approaches are therefore important, but these require extensive validation and are unlikely to be feasible in all clinical settings due to resource constraints.

Drug resistance is another major limitation. Although drugs targeting DNA methylation (e.g. DNA methyltransferase inhibitors) have shown promise, cancer cells can become resistant to these therapies (Stresemann and Lyko, 2008). Improving the efficacy of epigenetic therapies requires understanding the mechanisms behind this resistance and developing strategies to overcome it.

The integration of methylation data with other types of genomic and epigenomic data is still in its infancy. Although multi-omics approaches hold the promise of providing a comprehensive view of cancer biology, the complexity of integrating and interpreting these different types of data poses significant challenges (Feinberg *et al*., 2006). Therefore, researchers need powerful computational tools and frameworks to effectively integrate these data and extract meaningful insights.

Finally, there are many technical and methodological challenges to detecting and quantifying DNA methylation. Although next-generation sequencing (NGS) and bisulfite sequencing have greatly improved scientific research capabilities, these methods are still very expensive and resource-intensive. Furthermore, NGS and bisulfite sequencing require sophisticated bioinformatics tools and expertise to analyse the large amounts of data generated, which are

not readily available to all research or clinical laboratories (Bock *et al*., 2008) This means that it is difficult for researchers to gain access to data on people's DNA methylation without significant financial support.

Table 1: Overview of previous research on DNA methylation in cancer

Category	Key Concepts and Findings	Challenges and Limitations
Basic Mechanism	❖ DNA methylation involves adding a methyl group to cytosine in CpG dinucleotides. ❖ It influences gene expression by modifying DNA accessibility.	❖ Complexity of methylation patterns. ❖ Difficult to distinguish between pathogenic changes and consequences of disease.
Clinical Impact	❖ Methylation can silence tumour suppressor genes, affecting cell processes like apoptosis. ❖ Methylation patterns serve as biomarkers for cancer prognosis and treatment response.	❖ Variability in methylation among individuals complicates universal diagnostic and therapeutic strategies.
Technological Advances	❖ Advances in NGS and bisulfite sequencing provide detailed methylation maps. ❖ Machine learning enhances the precision of cancer diagnostics.	❖ High costs and resource-intensive nature of advanced sequencing technologies. ❖ Need for sophisticated bioinformatics tools.
Research Developments	❖ Studies have shown that methylation patterns change over time (epigenetic drift). ❖ Environmental factors like diet and stress influence DNA methylation.	❖ Difficulty in integrating methylation data with other genomic data. ❖ Challenges in interpreting complex multi-omics data.
Future Directions	❖ Continued research is enhancing understanding of methylation's role in cancer. ❖ Identification of specific methylation changes could lead to better biomarkers and treatments.	❖ Resistance to drugs targeting DNA methylation. ❖ Need for more effective integration of multi-omics approaches in research.

In conclusion, while DNA methylation research has greatly advanced understanding of cancer, there are limitations that must be addressed if its full potential is to be realised. These

include the complexity and variability of methylation patterns, technical challenges in detection and quantification, difficulty distinguishing benign from malignant changes, therapeutic resistance, and the need for better integration of multi-omics data. Addressing these challenges will require continued innovation and collaboration across multiple disciplines to discover more possibilities.

Section 2.5 Technique for balancing the data

Quasi - experimental methods and propensity score matching.

Propensity score matching allows to match patients from cancer and normal solid tissue datasets upon similar calculated propensity scores and thus balancing the datasets to only matched records which leads to accurate predictions (Wang, 2022). One-to-one matching is the most common approach, found in scientific literature and is used in the current analysis (Wilson, Giovannucci and Mucci, 2011).

Steps:

1. Calculate the propensity score for each gene
2. Decide how this score will be used
3. Run the logistic regression model

Firstly, random numbers cannot just be assigned to each patient's gene. To generate these scores a multivariable logistic regression model is run. Probability that the patient has higher methylation value for this gene conditional upon their patient’s data. The weights are assigned in such a way to the variables.

Table 2: Simplified explanation of propensity score matching

Id	Gene	Cancer	Propensity score
1	BCL2	Yes	0.6
2	BCL2	No	0.5
3	BCL2	Yes	0.5
4	BCL2	No	0.6
5	BCL2	Yes	0.7
6	BCL2	No	0.8

Calculated propensity score will be used for matching the records. Patients with the same propensity scores will be matched, for comparison normal tissue with tumour paerson, but with the same score. For an easier understanding, normal and tumour tissues were merged in the same table on figure one. In Python code matching is performed using 2 tables, one with normal tissues, the other one with tumour tissues. As can be seen from the table, patient 2 and patient 3 have the same propensity score of 0.5 for the gene BCL2. They will be analysed against each other. The same is done for patient 1 and patient 3, with their propensity score of 0.6. Patient 5 with propensity score 0.7 doesn't have a match, so he will be excluded from the analysis. The same will be done to patient 6, because his propensity score does not match other patients.

This is a simplified description of how matching occurs .Propensity scores values are also simplified. The actual python code uses the "psm.knn_matched" function (*PSMpy*, 2023). It uses k-nearest neighbours algorithm to find top closest neighbours for each patient's records (Wang, 2022). For 1:1 matching, the closest propensity score is picked as a match.

After the matching process is completed, a logistic regression model is run to predict the target variable "Tumour". The graph should show a major overlap of propensity scores for both groups in order for good matching results (an example of a good graph is shown on Table 2). If no matched records are found, propensity score matching cannot be used (Wang, 2022).

Chapter 3: Methodology

Section 3.1 Data sources and description

This section will provide a detailed description of the data sources used for the analysis of DNA methylation in colon cancer and processes involved in acquiring and preparing these datasets. It will also discuss the R scripts used to handle and analyse the data from the TCGA database.

Section 3.1.1 Data sources

The primary data source for this study is The Cancer Genome Atlas (TCGA), specifically the TCGA-COAD (Colon Adenocarcinoma) project. TCGA is a comprehensive and publicly accessible database that provides genomic, epigenomic, transcriptomic, and clinical data for various types of cancer. For this study, we focused on DNA methylation data, which is important for understanding the epigenetic modifications associated with colon cancer.

To acquire and prepare DNA methylation data from the TCGA-COAD project, we interacted directly with the Genome Data Commons (GDC) using the TCGAbiolinks software package in R (National Cancer Institute, no date). This package is essential for querying and accessing publicly available genomic datasets, including DNA methylation data collected through the Illumina Human Methylation 450 platform, which provides comprehensive coverage of more than 450,000 CpG sites (Colaprico *et al.*, 2015).

Section 3.1.2 Data acquisition and preparation

1. **Setting Up the R Environment**

The first step involves setting up the R environment by loading the necessary software packages and specifying the directory on the local computer where the data will be downloaded.

```
# Load required packages

library(TCGAbiolinks)
```

```
library(SummarizedExperiment)

library(dplyr)

library(IlluminaHumanMethylation450kanno.ilmn12.hg19)

# Specify download directory

download_dir <- "/~"

if (!dir.exists(download_dir)) {

  dir.create(download_dir, recursive = TRUE)

}
```

2. Querying and Downloading Data

Data queries specify key parameters based on the needs of the study, with the data for this thesis focusing on the data type of DNA methylation, the data format (Methylation Beta value) and the type of sample of interest, such as 'metastasis', 'primary tumour', 'recurrent tumour' and 'solid tissue normal'. In addition to this, there is information about the patient's clinic information.

```
# Query DNA methylation data for TCGA-COAD

query_TCGA_COAD <- GDCquery(

  project = "TCGA-COAD",

  data.category = "DNA Methylation",

  data.type = "Methylation Beta Value",

  workflow.type = "SeSAMe Methylation Beta Estimation",

  sample.type = c("Metastatic", "Primary tumour", "Recurrent tumour", "Solid Tissue Normal"),

  platform = "Illumina Human Methylation 450"
```

```
)

# Download data to the specified directory

GDCdownload(query = query_TCGA_COAD, directory = download_dir)

# Getting Clinical Data

clinical_coad <- GDCquery_clinic("TCGA-COAD")
```

3. **Annotating Methylation Data**

The IlluminaHumanMethylation450kanno.ilmn12.hg19 package was used to incorporate genome-annotated data, linking each CpG locus to a specific genomic position for subsequent biological interpretation and analysis.

```
# Load annotation data from the package

data(IlluminaHumanMethylation450kanno.ilmn12.hg19)

# Retrieve the detailed annotation data

annot <- getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)

# Convert the annotation data to a data frame

annot_df <- as.data.frame(annot)
```

4. **Merging and Cleaning Data**

The methylation data are merged with genomic information based on common identifiers, aligning the molecular data with the genomic context.

```
# Merge methylation data with annotation

merged_data <- merge(gene_methylation_data, annot_df, by.x = "row.names", by.y = "Name", all.x = TRUE)

# Select and rename necessary columns using base R

tcga_columns <- grep("^TCGA", colnames(merged_data), value = TRUE)
```

```
coad_data <- merged_data[, c("Row.names", "chr", "pos", "UCSC_RefGene_Name",
tcga_columns)]

colnames(coad_data)[1:4] <- c("Composite", "Chr", "Coordinate", "Gene")

# Function to remove duplicate gene names separated by semicolons

clean_gene_names <- function(gene) {

  if (!is.na(gene) && gene != "") {

    unique_genes <- unique(unlist(strsplit(gene, ";")))

    return(paste(unique_genes, collapse = ";"))

  }

  return(gene)

}

# Apply the function to clean up gene names

coad_data$Gene <- sapply(coad_data$Gene, clean_gene_names)

# Remove rows where all TCGA columns are NA or NULL

coad_data <- coad_data[rowSums(is.na(coad_data[tcga_columns]) |
coad_data[tcga_columns] == "") < length(tcga_columns), ]

# Display the final data

print(head(coad_data))
```

The final merged_data has 478 rows and 485,689 columns of data.

Section 3.2 The cancer genome atlas (TCGA)

The Cancer Genome Atlas (TCGA) is a collaborative effort of two organisations: the National Cancer Institute (NCI) and the National Human Genome Research Institute

(NHGRI). They are both part of the National Institutes of Health, U.S. Department of Health and Human Services (*DBGAP Study*, no date).

TCGA is an open data portal that has patient data collected from more than 11,000 patients. The data can be taken out and analysed by any person.

Initially, the goal of the project was to bring data together across thousands of patients with all cancer types into one place, available for further research. The data on the portal was also used by scientists to drive the progress of the medical field. Using the mutations, gene expression and methylation data TCGA team has been able to develop better sub-classifications of individual cancers and see that in some cases patients might respond differently to treatments. For example, the classification that TCGA team did indicated that there was a subgroup of low-grade gliomas that had a highly malignant potential, but they were treated as the regular low grade gliomas. Using that past knowledge, today the patients are being treated and followed up differently because they get classified in the correct group from the beginning.

The possibilities of TCGA portal are much bigger than only being able to classify patients into correct groups upon their disease, but also to get greater insights into the molecular mechanisms for the different classifications, which in future can lead to better treatments for those individual patients.
Some of the past TCGA discoveries led to new therapeutic targets and even new drugs that are influenced by genetic variation and molecular characterization and biomarkers. As well, it really helped with precision medicine nowadays.

Recent version of TCGA data portal provides the valuable cases for colon cancer patients that come from verified sources. As well, TCGA portal offers built in analysis tools such as “Gene expression clustering”, “Cohort comparison” and many others, that might be valuable for analysis in-parallel with other machine learning models.

Section 3.3 Data preprocessing

Section 3.3.1 Data types and sources

DNA Methylation Data: Quantitative methylation data from Illumina's Human Methylation 450 platform. Methylation Beta values represent the proportion of methylation at each site and are used to assess the genome-wide methylation status of COAD patients.

Gene Expression Data: RNA-Seq data provides insight into gene expression levels to correlate changes in methylation with changes in gene activity.

Clinical Data: Patient information including age, gender, cancer stage, treatment history and survival outcomes.

Section 3.3.2 Preprocess

Data cleaning was performed after data collection to eliminate any inconsistent or incomplete data entries. This step was essential to ensure the accuracy of the analysis, and the cleaning process included checking for and dealing with missing values and outliers.

Standardisation is another key step aimed at minimising technical variability that may mask true biological differences. The process is mainly performed using this limma software package, where we apply standardisation techniques to adjust methylation data to be comparable across samples.

Finally download all this data into csv files: one file contains the patient's **clinical data**, another contains **the gene methylation data for normal tissues** (normal solid tissues), and the third contains **the gene methylation data for cancerous tissues** (metastases, primary tumours and recurrent tumours). The analytical approach was partitioned into three main sections.

Section 3.3.3 Data cleaning for analysis 1 and dataset description

Section 3.3.3.1 Preparation phase

For the analysis methylation dataset was required to be prepared. Firstly, "Composite", “Coordinate" and "Chr" columns were dropped, because they were not needed. This work

was decided to be concentrated on the gene name and methylation values for each gene. It was observed that the dataset contained double names of the genes for some of the gene names, divided by the “;” sign. In these cases a separate row is made for each gene name defined in the name. New rows' values are duplicated from the original row.

Missing gene names were filled with “Unknown" names.

For the patient data only the following columns were included in the dataset for analysis: "submitter_id", “race", “gender", “ajcc_pathologic_stage", “tissue_or_organ_of_origin", “primary_diagnosis", “prior_malignancy”. These columns were chosen as they most closely describe the patients.

For each dataset unique gene names were calculated.

There were 23335 unique genes for the tumour dataset and 23950 unique genes for the normal dataset. Unique genes presented in both datasets (normal and tumour) was only 20448.

It was also found that gene names can be repeated multiple times, so the mean of all its values was calculated, resulting in the gene name being listed only once.

Within the dataset, there are 314 samples of tumour tissue (Metastatic, Primary Tumour, Recurrent Tumour) and 38 normal tissue samples (Solid Tissue Normal).

Both tumorous tissue and normal tissue were taken from Colon Adenocarcinoma patients.

Tumour tissue is removed directly from sites in which the cancer has proliferated.

Normal tissue is taken from areas of the body that have not yet been metastasized to by cancer.

Section 3.3.3.2 Analysis phase

Firstly, barplot was made upon patient data to make sure it is a representation of the US population, since TCGA data is taken from the US population. The dataset contains 46% of male representatives and 40% of female representatives with colon cancer as seen on figure 3.

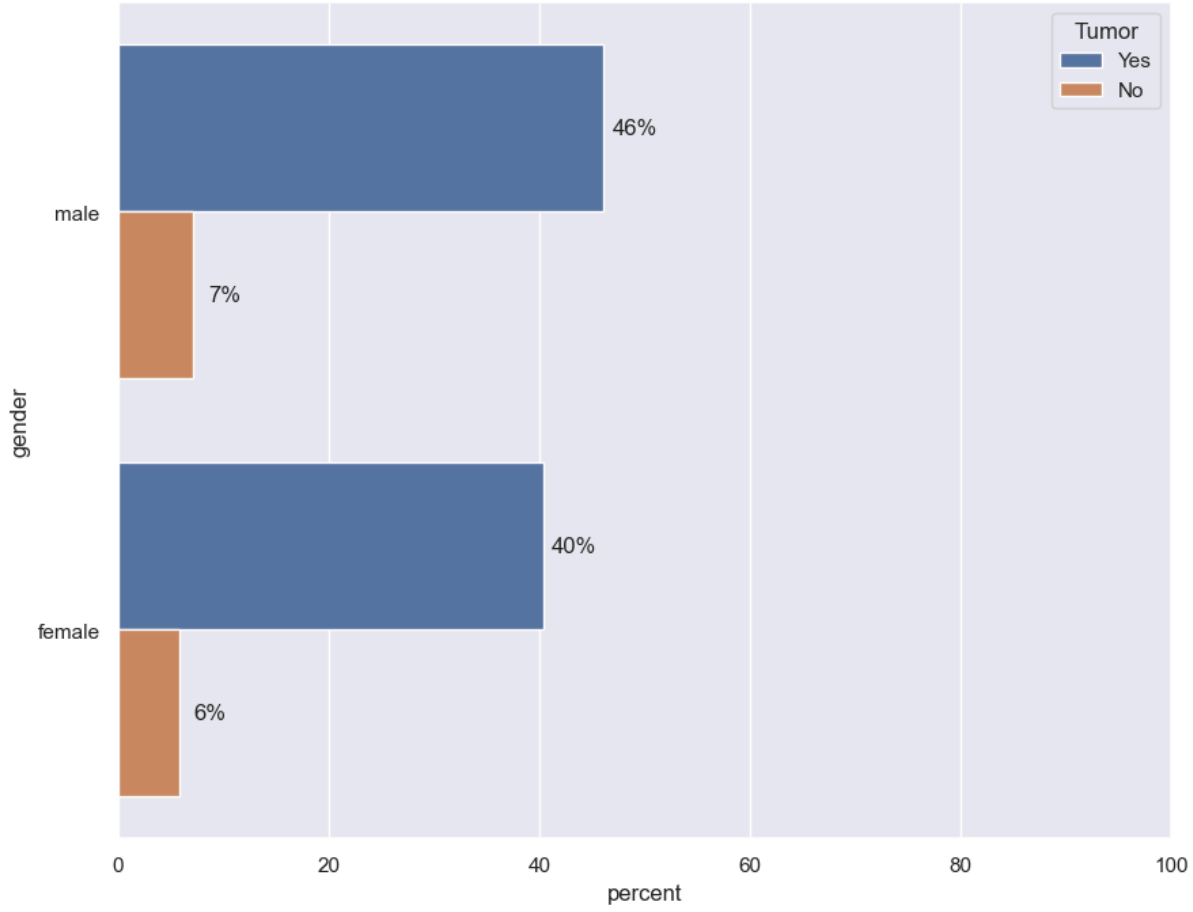


Figure 3: Dataset's overview of gender types

Then statistics for colon cancer in the US for the year 2024 was analysed. *Cancer facts and statistics* estimated that approximately 106,590 new cases of colon cancer were found in 2024 (with the following numbers in men - 54,210 and in women - 52,380) as shown on figure 2 (*How common is colorectal cancer?*, no date). Showing these values in percentage format clearly shows that analysed colon cancer datasets from TCGA are a right representation of the US population (Statista, 2024).

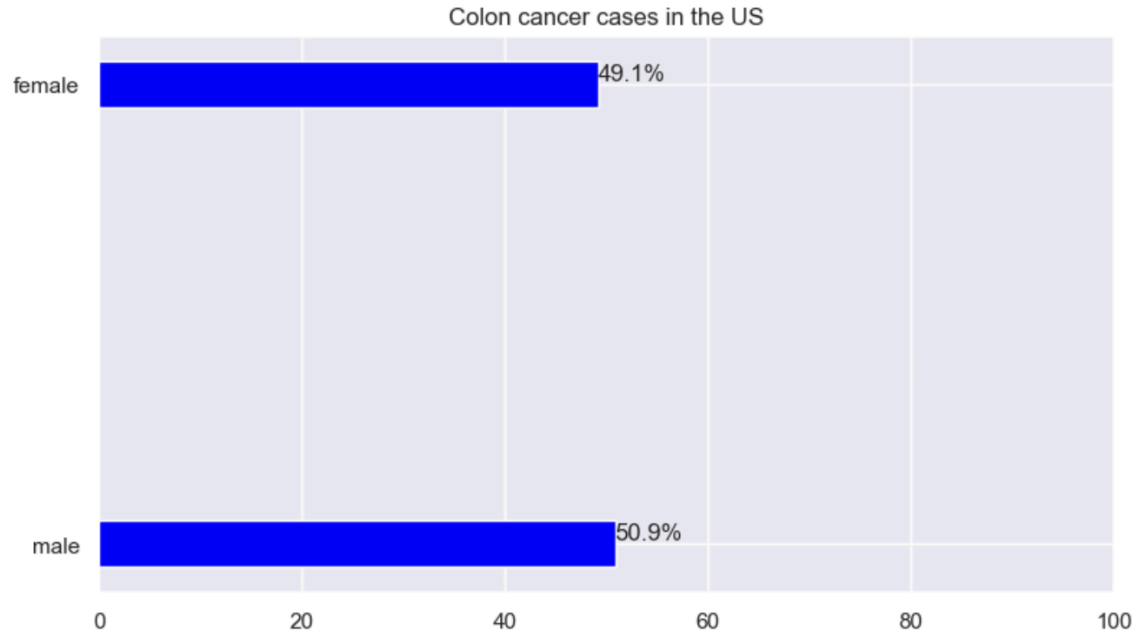


Figure 4: Colon cancer cases in US for male and female

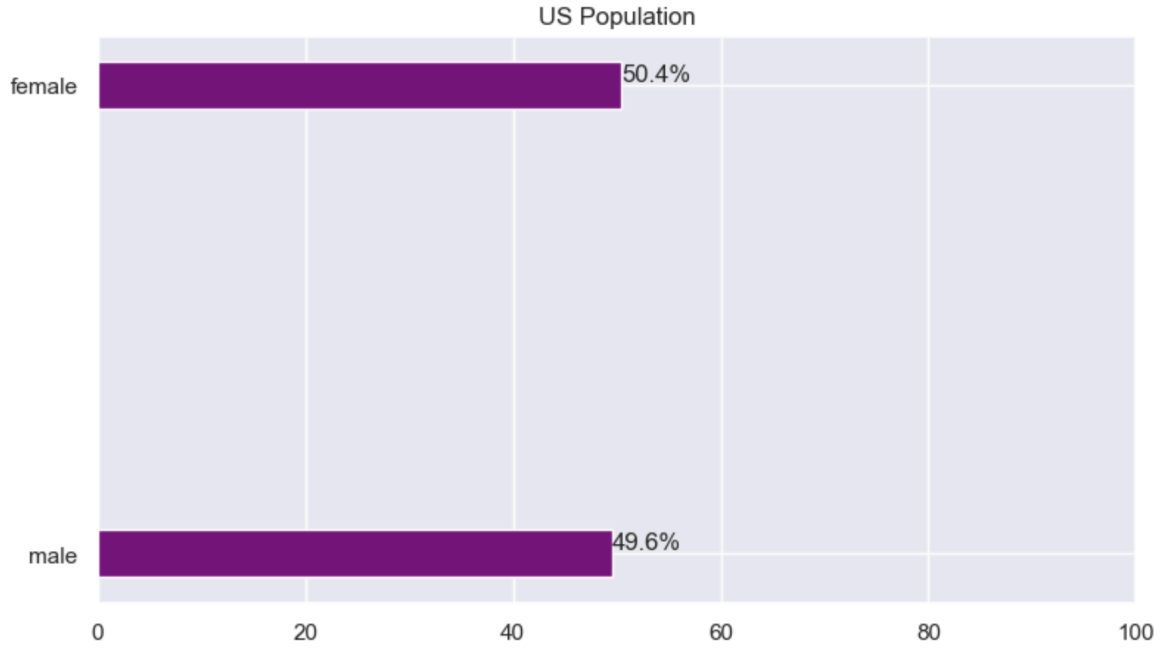


Figure 5: Comparison of colon cancer gender types with the whole US population

Patients were also split by ethnic groups and the most common type in the analysis is "white". It cannot be said that "white" people are more inclined to get colon cancer, it is just that there is more information about them in the current dataset with the leading 63% for

patients with cancer. Barplot graph illustrates this on figure 6. Ethnic groups were not fully analysed in the current work for the lack of data, further investigation should be carried on because there is a research already made that states that if taking the male category, there is an increase in chance (of 81.4%) of getting colon cancer in patients of African Americans ethnic group (Ollberding *et al*., 2011).

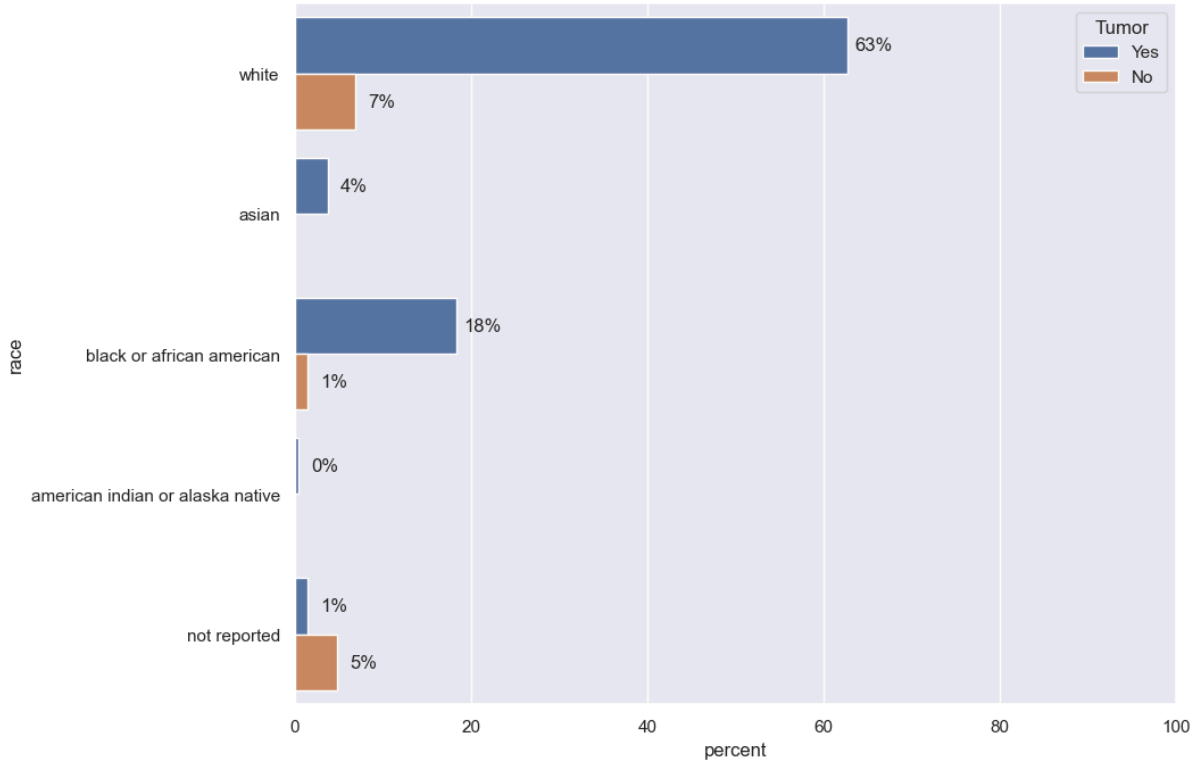


Figure 6: Dataset's overview of race types

Cancer stage was analysed to see the most common cancer types. Top 3 most common for this dataset stages are "Stage IIA", "Stage IIIB", "Stage I".

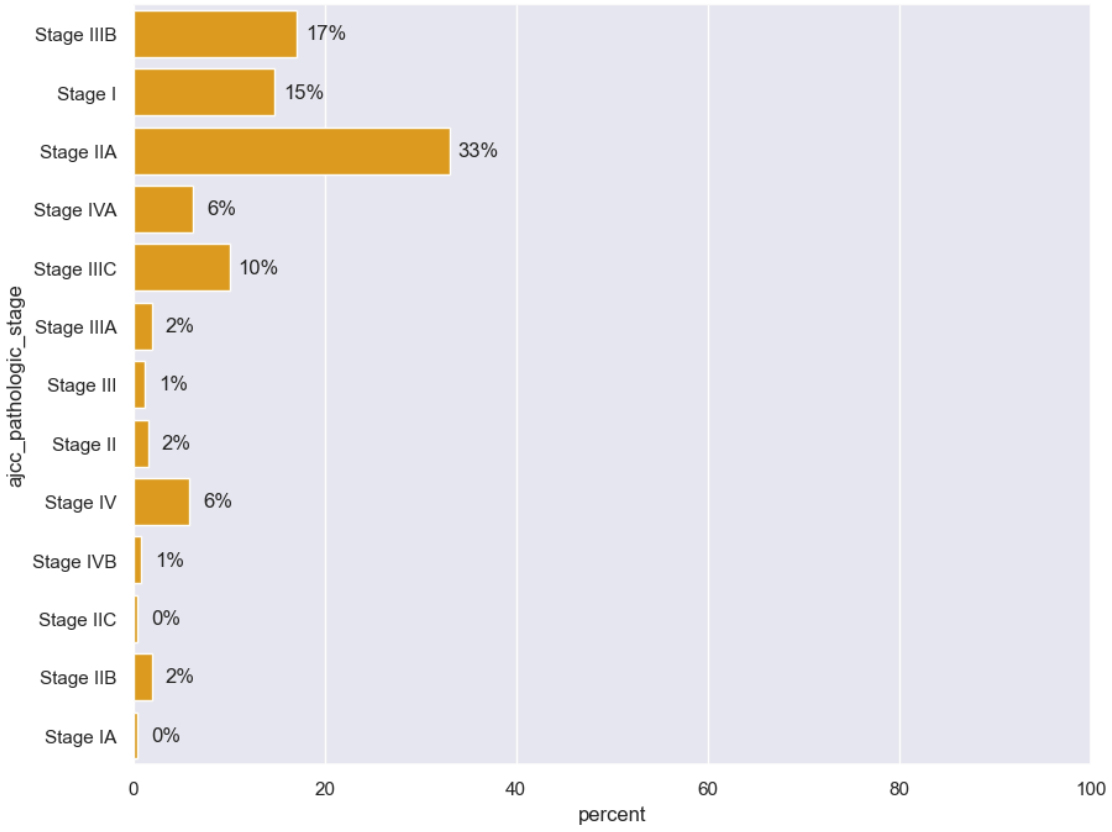


Figure 7: Dataset's overview of cancer stages

Section 3.3.4 Data cleaning for analysis 2 and dataset description

Section 3.3.4.1 Data cleaning

1. Handling Missing Values

One of the first steps in our data cleaning process involved handling missing values within the dataset. DNA methylation data, obtained through the Illumina Human Methylation 450 platform, often contain missing values due to various reasons such as technical errors or insufficient coverage. To address this, using the function to check for any rows in our data matrix that contained missing values. This step ensures that only the data with complete information is carried forward for analysis.

2. Normalisation and Transformation

Following the filtering of missing data, normalisation is applied to correct for potential batch effects and other systematic variations that are not related to biological differences. Using the

voom function from the limma package, that transformed the count data to log-counts per million (CPM), enabling appropriate modelling of the data. The voom function also accounts for the mean-variance relationship inherent in count data, making it suitable for linear modelling.

3. Validation of Group Variables

Another critical aspect of our data cleaning involved the validation of group variables used for comparisons in our differential expression analysis. We ensured that the condition variables, such as 'definition' which might refer to cancer stages or types, and 'gender' are present and correctly formatted in the dataset.

4. Annotation Merging

Finally, gene annotations were integrated with the methylation data to facilitate biological interpretation. This involved merging CpG site information with gene symbols to link methylation patterns to specific genomic locations.

Section 3.3.4.2 PCA of the data

Principal Component Analysis (PCA) is performed to reduce the dimensionality of a dataset and highlight the most important patterns of variation (Ringnér, 2008).PCA transforms the raw data into a new set of uncorrelated variables, called principal components, which capture the maximum variance in the data. The process involves normalising the data, calculating the covariance matrix and then extracting eigenvalues and eigenvectors to determine the principal components.

PCA results can be visualised by scatter plots of the first two principal components (PC1 and PC2), which typically capture the most significant variation in the data set. This report focuses on the analysis of three illogical metadata, namely **tissue type, gender and ethnicity**, for colouring to provide insight into how these factors affect overall methylation status.

Section 3.3.4.3 Volcano plot

Differential methylation analysis aims to identify CpG sites with significant differences in methylation levels between normal and cancerous tissues, which may be potential regulatory regions affecting changes in cancer-related gene expression.

In this analysis, the normal and tumour data were combined into one dataset and a design matrix was created to model the group variables (normal vs tumour). A linear model was then fitted to the data using the limma software package, which is particularly suited to high-dimensional data such as methylation profiles. Empirical Bayesian conditioning was applied to improve the stability and reliability of the results.

The analysis extracts the most differentially methylated CpG sites and adjusts for multiple comparisons using the Benjamini-Hochberg method to control for false discovery rates (Smyth, 2004). Results included gene names and were categorised according to adjusted p-values and fold changes, making the visualisation effect more prominent for significant changes (e.g. 'up-regulated', 'down-regulated', 'not significant'). Furthermore, the altered analyses set specific genes of interest, which were focussed on because of the relatively significant changes in these genes in previous studies (Moore, Le and Fan, 2012). Therefore, this analysis focused on these genes (DNMT1, DNMT3A, DNMT3B, DNMT3L, MBD1, MBD2, MBD3, MBD4, UHRF1, UHRF2, ZBTB)

Generated volcano plots visualise the results, showing the relationship between the magnitude of methylation change (log2 fold change) and statistical significance (-log10 adjusted p-value). These plots help to identify key gene sites with significant methylation differences.

Section 3.3.4.4 Survival analysis

This work began with the collection and preparation of clinical data for the TCGA-COAD project, which included basic details such as patient identifiers, days to death, and vital status. These data points allow for tracking the long-term prognosis of the patient and help to correlate these prognoses with epigenetic data. To prepare for survival analyses, vital status information was converted into a binary format to differentiate between patients who died and those who survived, as one of the key variables in the survival model.

This analysis performed a comprehensive survival analysis using well-established statistical methods. First, Kaplan-Meyer survival curves were generated to describe survival probabilities over time in a nonparametric manner. These curves were stratified according to different clinical and demographic factors, including cancer stage, gender, and ethnicity, to visually assess differences in survival outcomes among these groups.

The effects of methylation and other covariates on survival were then quantitatively assessed using Cox proportional hazards models. This approach allows for simultaneous adjustment for multiple factors and provides a robust framework for subsequent analyses to determine the specific impact of methylation patterns on survival outcomes.

Section 3.3.5 Data cleaning for analysis 3 and WGCNA analysis

This analysis used Weighted Gene Co-expression Network Analysis (WGCNA) to investigate gene co-expression networks in tumour and normal tissue samples.Gene modules derived from normal and tumour datasets using top 5000 genes by variance were compared. To do this, conditions for analysis experiments were designed: Tumour Data with Tumour Variance, Normal Data with Normal Variance, Tumour Data with Normal Variance, and Normal Data with Tumour Variance.

The top 5,000 genes by variance were chosen as variance is an indicator of biological significance, and it differentiates our past analysis based on gene selection.

Section 3.3.5.1 Data preparation

The datasets for tumour and normal samples were obtained from TCGA and loaded in R via RStudio. For each dataset, the CpG site names, gene names, and methylation values were extracted. The methylation values were cleaned to ensure all data were numeric, and any columns with more than 50% missing values were removed.

In each dataset, the top 5000 genes were selected based on variance. This selection helps to focus on the most variable and potentially significant genes. The datasets were then transposed for compatibility with WGCNA.

From the available normal tissue samples, a subset of 35 patients was randomly selected to ensure a balanced comparison with the 35 tumour samples. This random selection helps to reduce bias and allows for a more manageable analysis.

Section 3.3.5.2 Network construction and module detection

For each dataset, we checked for good samples and genes using the goodSamplesGenes function from the WGCNA package. The samples and genes that did not meet the quality criteria were excluded.

We then selected a set of soft-thresholding powers to construct the network topology. The optimal power was chosen based on the criterion of achieving a scale-free topology (a network distributed based on a power law), standard for biological networks.

The blockwiseModules function from WGCNA was used to perform module detection. This function partitions the dataset into blocks and constructs a network for each block. The parameters used were a maximum block size of 5000, unsigned Topological Overlap Matrix (TOM), minimum module size of 30, and a merge cut height of 0.25.

Section 3.3.5.3 Gene connectivity and hub genes

After constructing the networks, the adjacency matrix and TOM were calculated. Gene connectivity within modules was assessed using intramodular connectivity. The top hub genes for each module were identified as those with the highest connectivity.

Section 3.3.5.4 Visualization and heatmaps

Dendrograms were graphed to visualise the found modules. Heatmaps for each module were created and saved. The TOM heatmap for the entire dataset and the eigengene adjacency heatmap for the module eigengenes were also generated and saved.

For each experimental condition, the WGCNA results, including module labels, module colours, eigengenes, gene connectivity, and hub genes, were saved to files. The resulting dendrograms and heatmaps were stored in specific directories for further analysis.

Chapter 4: Findings and Results

Section 4.1 Overview of analysis 1

Section 4.1.1 Immune genes analysis

In order to address the question of why cancer patients have a higher rate of infection (*Why Are People with Cancer More Likely to Get Infections?*, no date) the following set of genes that control T-cells, which in turn play a role in cancer development and sepsis, were determined:

- "BCL2",
- "BTLA",
- "CD274",
- "CD44",
- "CTLA4",
- "FOXP3",
- "GATA3",
- "HAVCR2",
- "HMGB1",
- "IL13",
- "IL2",
- "IL2RB",
- "IL33",
- "ITK",
- "LAG3",
- "LAT",
- "MIF",
- "NFKB1",
- "PDCD1",
- "RORC",
- "SPP1",
- "STAT1",

- "STAT2",
- "TBX21",
- "TNFSF10",
- "WNT1".

One of the most important factors is that these genes should be present in both datasets in order to compare the values. If one gene is not present in one dataset, analysis is not possible, because there are no values for comparison. It was verified that all of the specified above genes are present in both datasets.

Then it was checked for the missing methylation values in the datasets for patients.

There were only 10 missing values in the dataset with patients with cancer in some of their genes and those values were excluded from the analysis because of the extremely small number compared to the total of defined values. There were none missing values in the dataset with normal tissue data.

Section 4.1.1.1 Unbalanced immune genes dataset

The next step was to build a correlation plot for the specified set of genes for normal tissue and tumour tissue. Correlation plot shows how closely two genes are related and do they move in the same or opposite directions.
On the first plot (Figure 8) and second plot (Figure 9), the greener the values show that when one methylation value of the gene rises, the other one rises as well. The more pink the values are means that when one methylation value goes up, the other one goes down.

To explain the values on the graph general guidelines for correlation values should be specified. References for these guidelines were taken from “Integrative analysis of DNA methylation and gene expression data in breast cancer” article (Akoglu, 2018) and Operations Analytics Trinity lectures slides.

Table 3: Interpretation of correlation values

Correlation values	Interpretation
0	No relationship
0.01 - 0.19	Little to no relationship
0.20 - 0.29	Weak relationship
0.30 - 0.39	Moderate relationship
0.40 - 0.69	Strong relationship
0.70 - 0.99	Very strong relationship
1	Perfect relationship

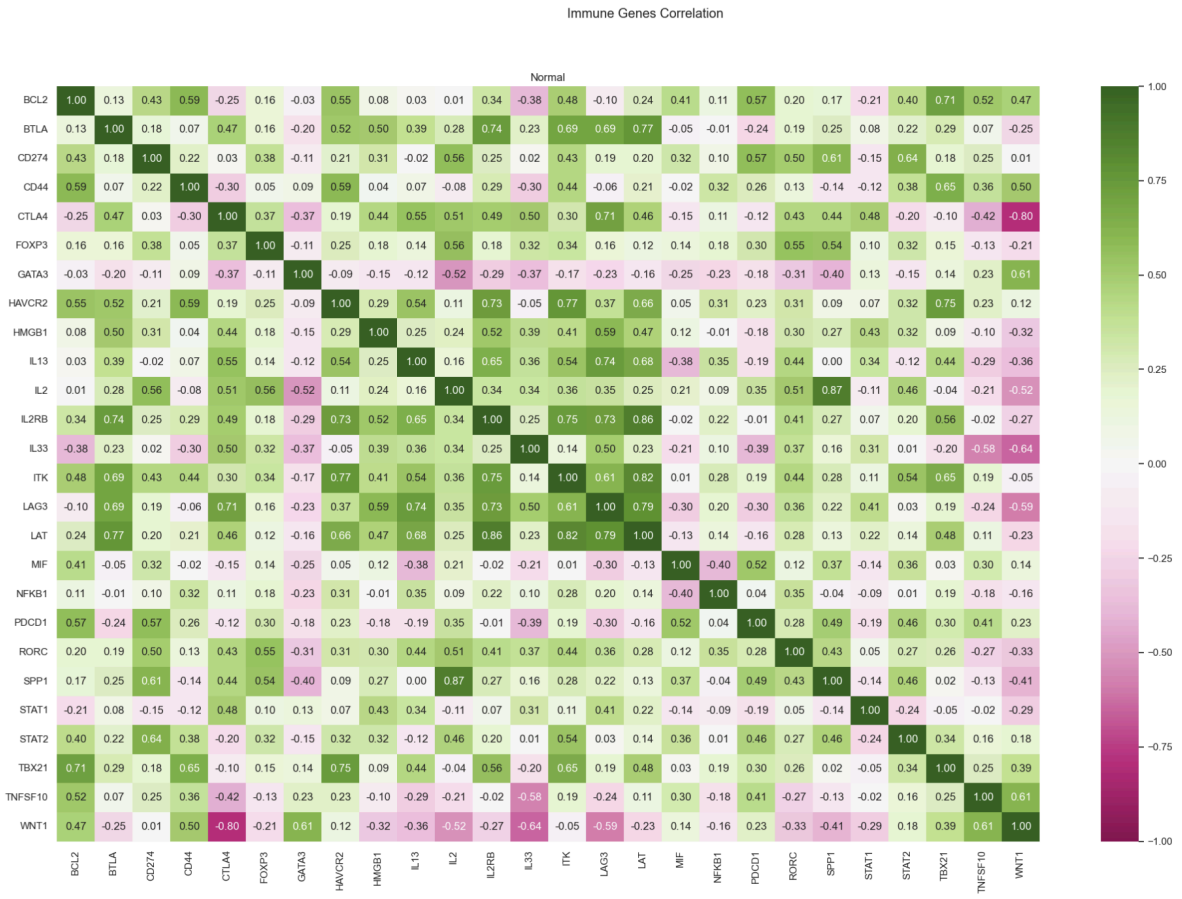


Figure 8: Unbalanced immune gene dataset correlation plot for normal tissue.

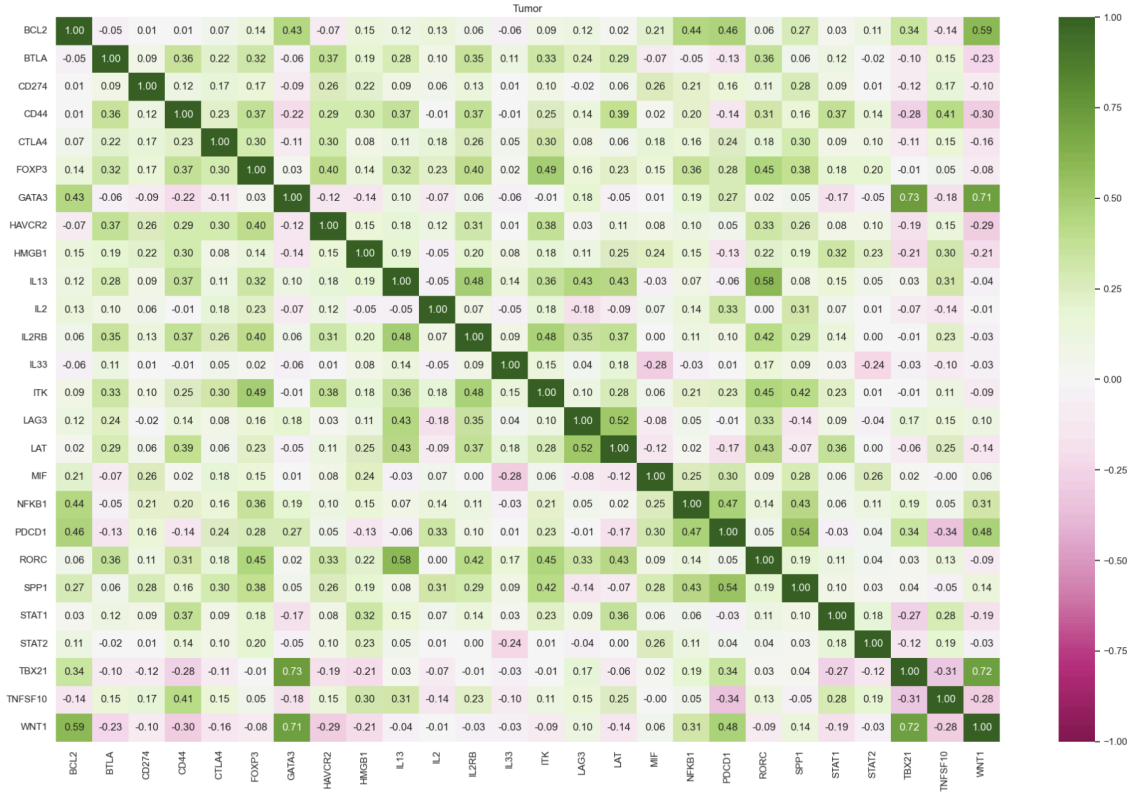


Figure 9: Unbalanced immune gene dataset correlation plot for tumour tissue

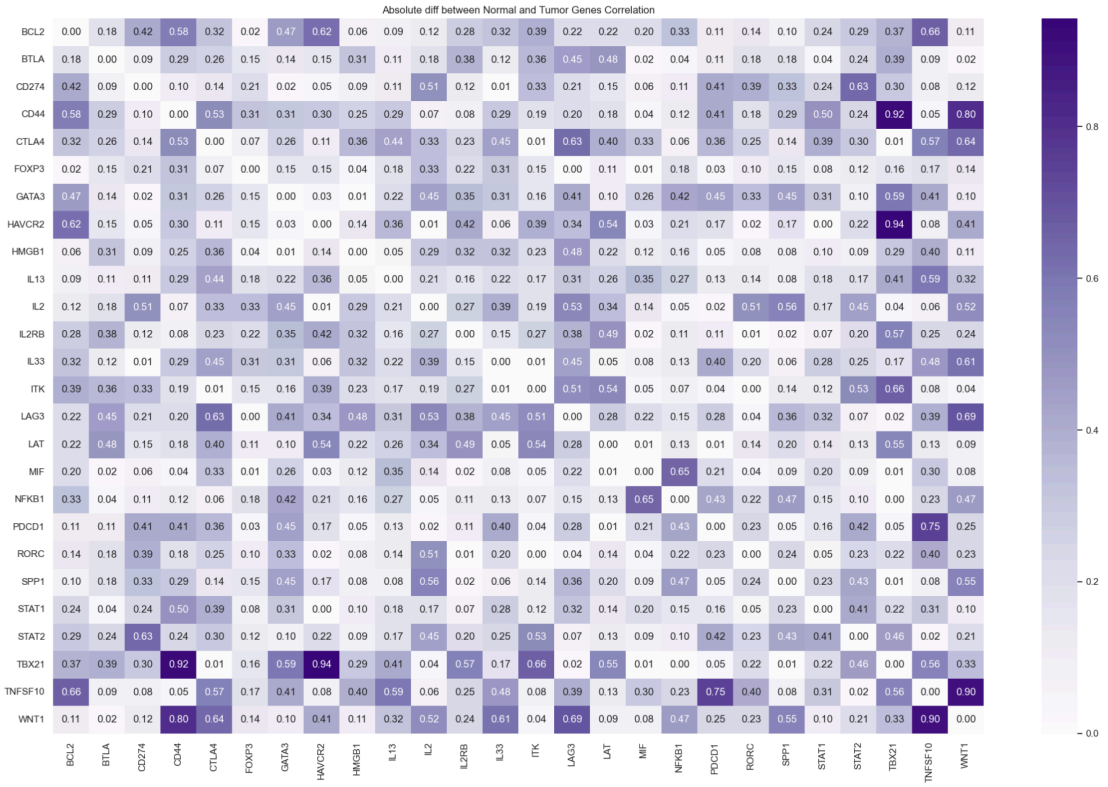


Figure 10: Unbalanced immune gene dataset correlation plot with absolute value

The overall tendency on the immune graphs show that for people with normal tissue methylation values are more correlated compared to the tissue with cancer where values are more stable with no significant difference in the value changes. The pale areas, where values are in a range between 0 and 0.19 show little or no correlation, which means those genes have no linear relationship between them. The absolute value graph (Figure 10) shows the difference between normal and tumour graphs. If the difference is big, which shows positive correlation, the result will be darker.

Since the graph is made on an unbalanced dataset, which needs more further investigations, top 3 correlated genes will be identified on the absolute value graph. Following are the pairs of genes that have the highest correlation: “TBX21” and “HAVCR2” with 0.94 correlation value, “TBX21” and “CD44” with the 0.92 correlation value, “TNFSF10” and “WNT1” with 0.90 correlation value.

To illustrate the relationships further, pairplot graphs for each gene relationship are made. On the diagonal line histograms with the value ranges are made with blue colours showing the methylation value range for tissue with cancer and orange colours showing the methylation

value ranges for normal tissue for each gene individually. For some genes, zoomed in on figure 13, significant differences in value ranges are seen. The rest of scatterplots show relationships between all the pairs of genes. On the X and Y axis are the full list of immune genes. Pairplot graphs present a great opportunity to see the spread of values for the chosen genes for normal and tumour tissue patients and check if there is a certain pattern. The pairplot graph was made using Python library "Seaborn", used for data visualisation (*seaborn.pairplot — seaborn 0.13.2 documentation*, no date).

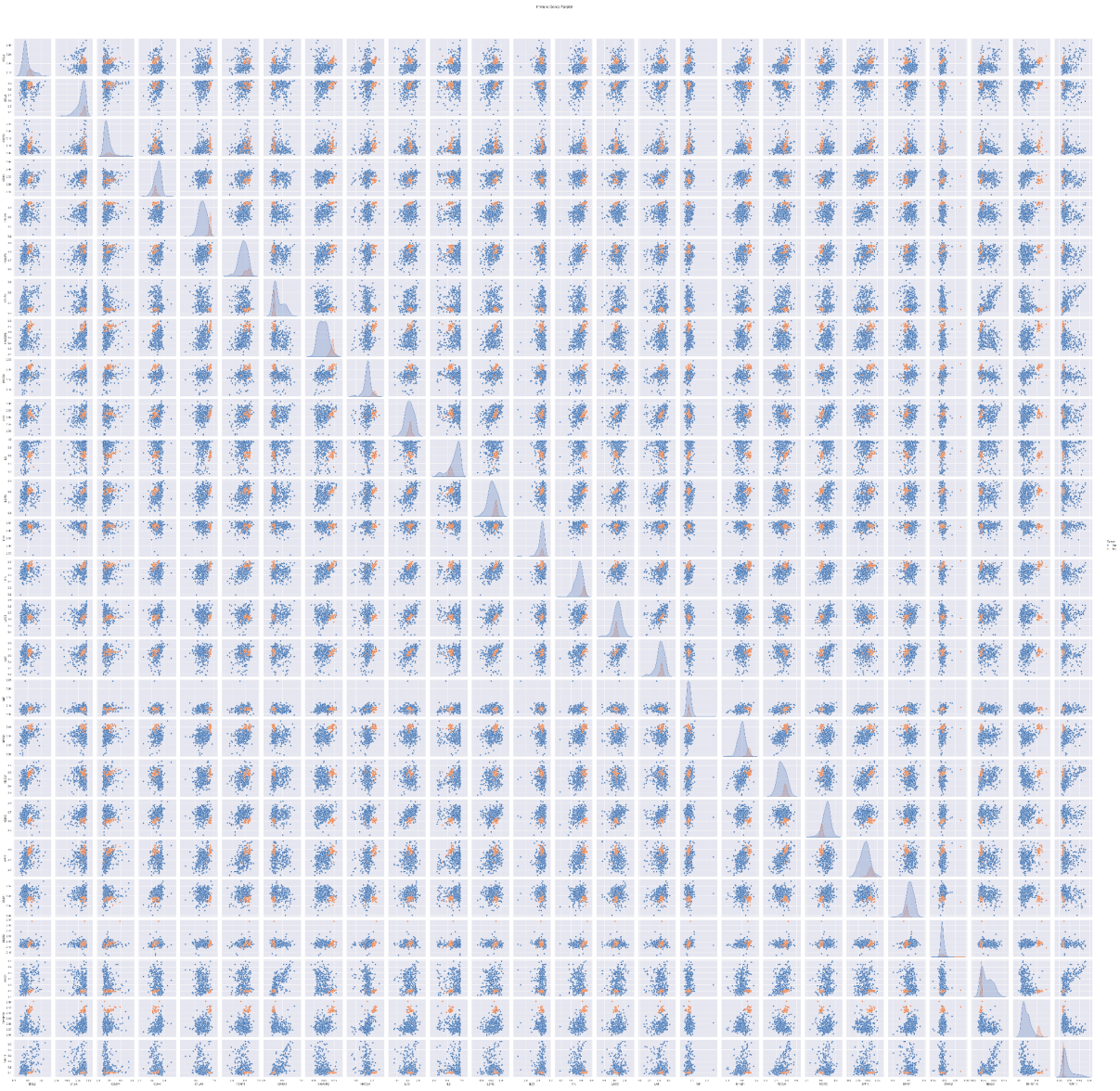


Figure 11: Immune genes pairplot graph

On the next figure 12 zoomed in, a pairplot graph is shown, listing the first 5 genes for a closer look on how it is made. Top genes are self written on the screenshot, because on the initial graph (figure above) those values are written at the bottom.

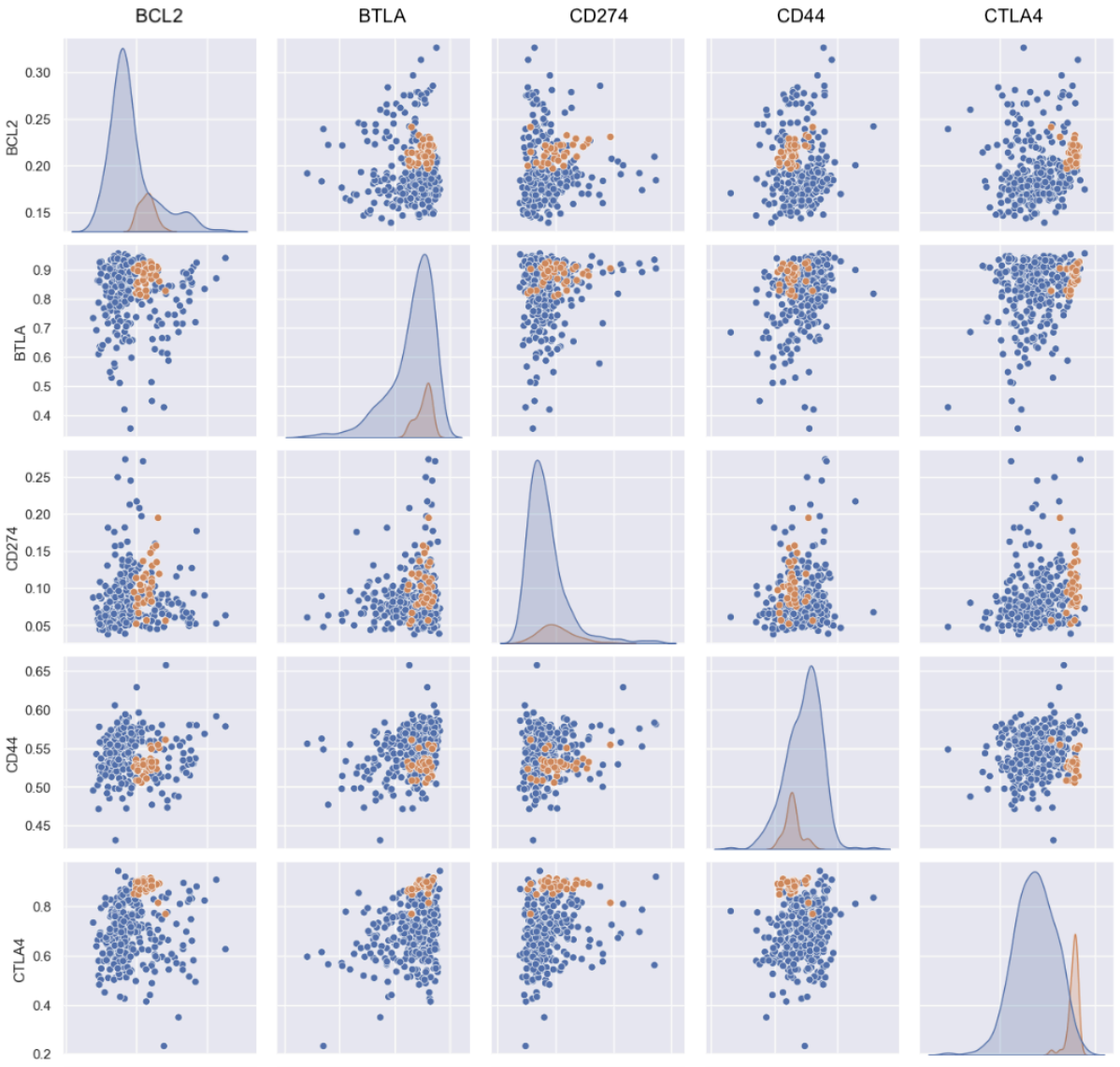


Figure 12: Immune genes pairplot graph zoomed in on the first 5 pairs of genes

From the graph on figure 11 it can be seen that the difference in means for most genes for cancer and non-cancer tissue are relatively similar that suggest that those genes are not significant in immune system response that appears after sepsis and might lead to cancer. Although some genes like “TNFSF10”, “HAVCR2”, “CTLA4” show significantly different mean value ranges for cancer and non-cancer tissue.

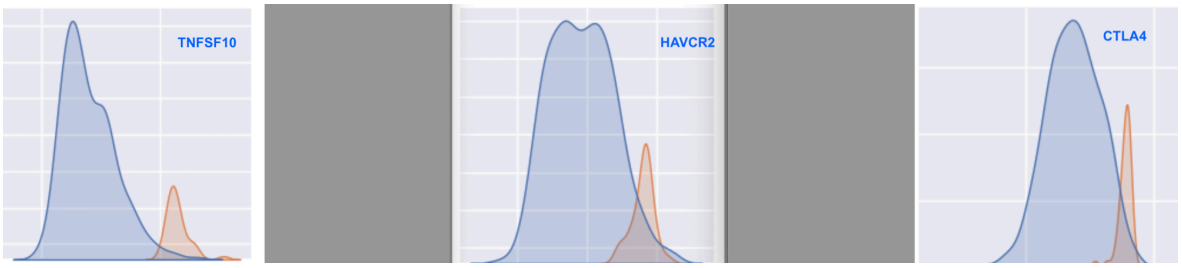


Figure 13: Genes with the biggest correlation difference

Figure 13 illustrates zoomed in focus of these genes, where blue area indicated tumour tissue and orange area normal tissue values.

Section 4.1.1.2 Balanced immune genes dataset

The previous results were identified on an unbalanced dataset. Thus to check if the pattern will remain the same on a balanced dataset, quasi-experimental method propensity score matching was used to balance the data in the datasets by downsampling the tumour dataset, because it has more values than the normal tissue patient's dataset. Downsampling is performed in such a way: individual genes are given a propensity score in each dataset. Then genes with matched propensity scores in each of the datasets are left. Then they are compared with each other. Detailed overview of how this algorithm works is described in the literature review section 2.5. Unbalanced dataset is shown on figure 14.

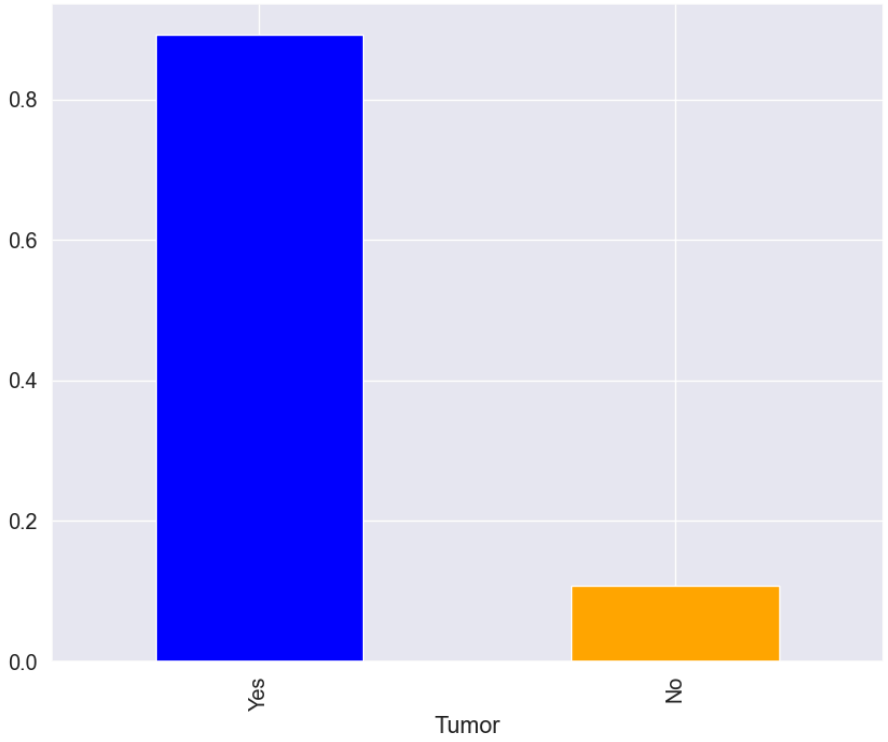


Figure 14: Patient ratio before the propensity score matching

Propensity score matching is run on each gene individually. For the limited space of this thesis only 4 graphs out of 26 will be shown on figures 15 - 18.

Figure 15: Propensity score matching for gene BCL2

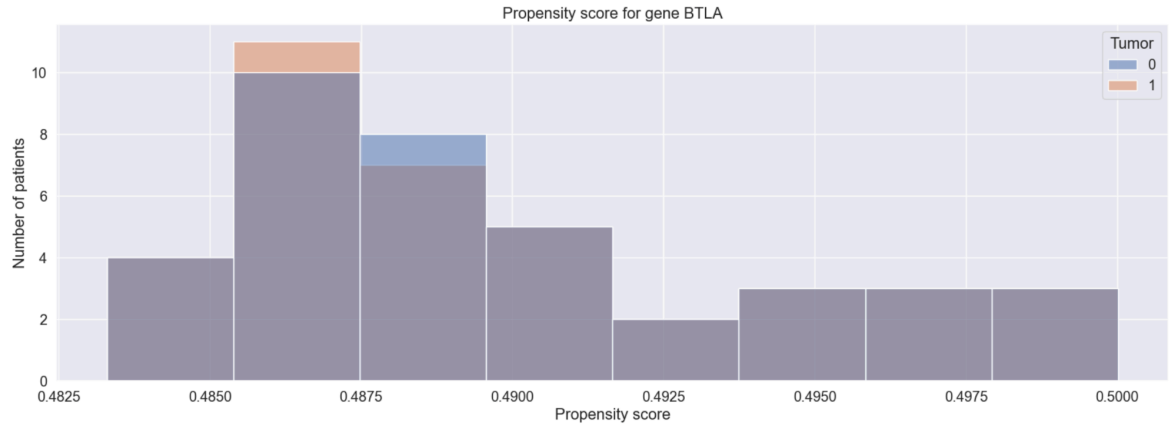


Figure 16: Propensity score matching for gene BTLA

Figure 17: Propensity score matching for gene CD44

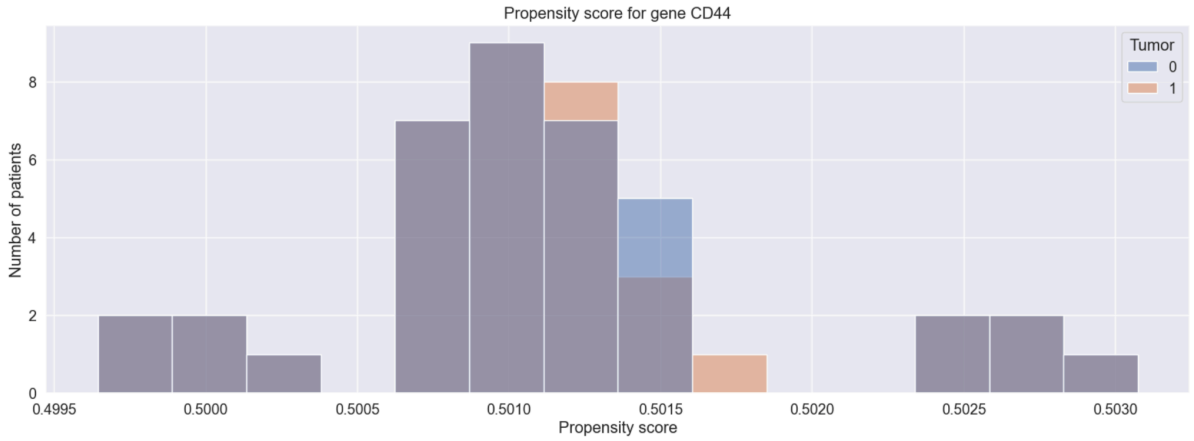


Figure 18: Propensity score matching for gene CD44

After running the propensity score algorithm, all genes had a major overlap of propensity score values. This means that there are enough matches of the data.

As a result, tumour patients' dataset became downsampled to match the records from the normal tissue patients dataset. Balanced dataset's overview is shown on figure 19.

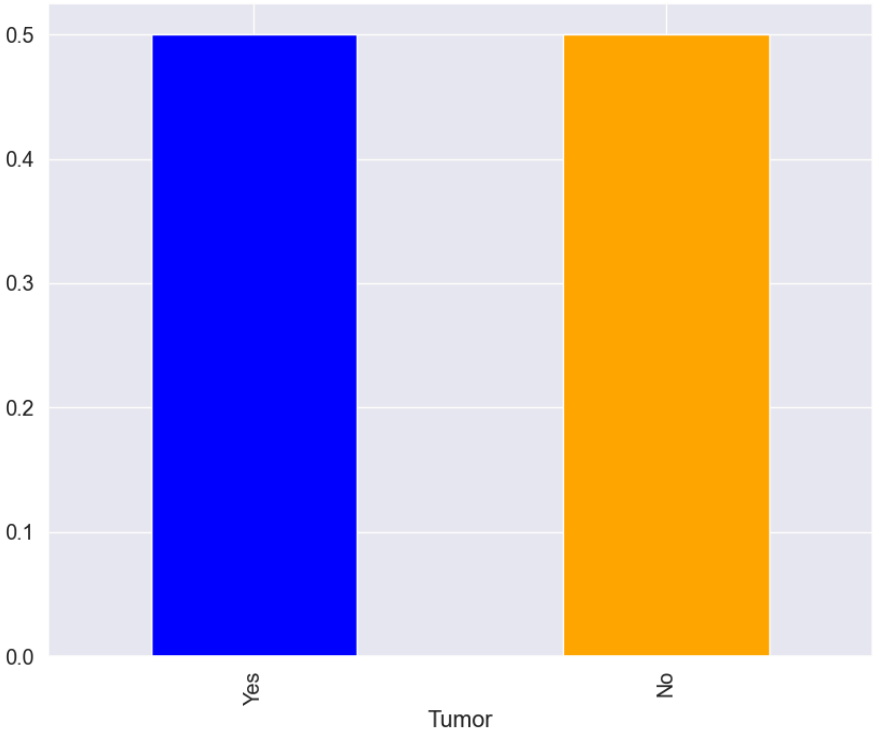


Figure 19: Normal and tumour patient's datasets with equal number of rows after propensity score matching

After the propensity score matching has been completed correlation and pairplot graphs are run again to see if results remain the same on the balanced datasets.

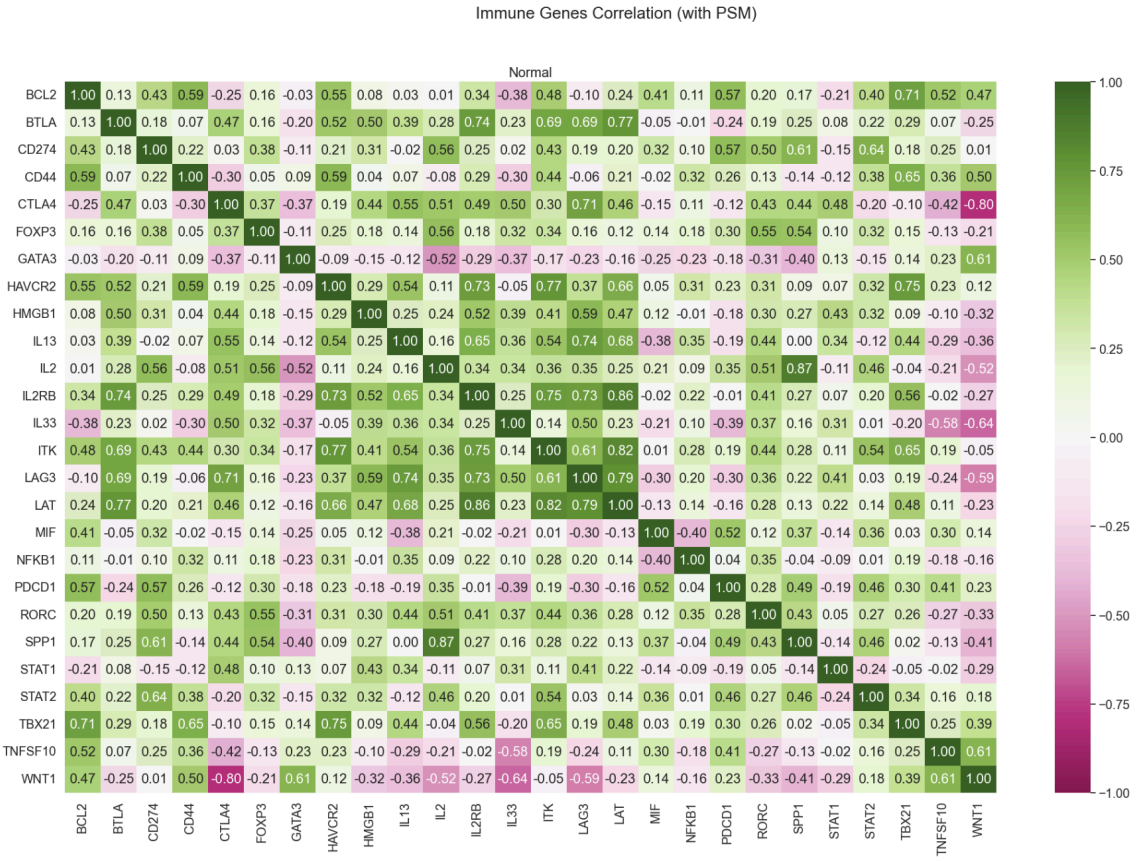


Figure 20: Balanced immune gene dataset correlation plot for normal tissue

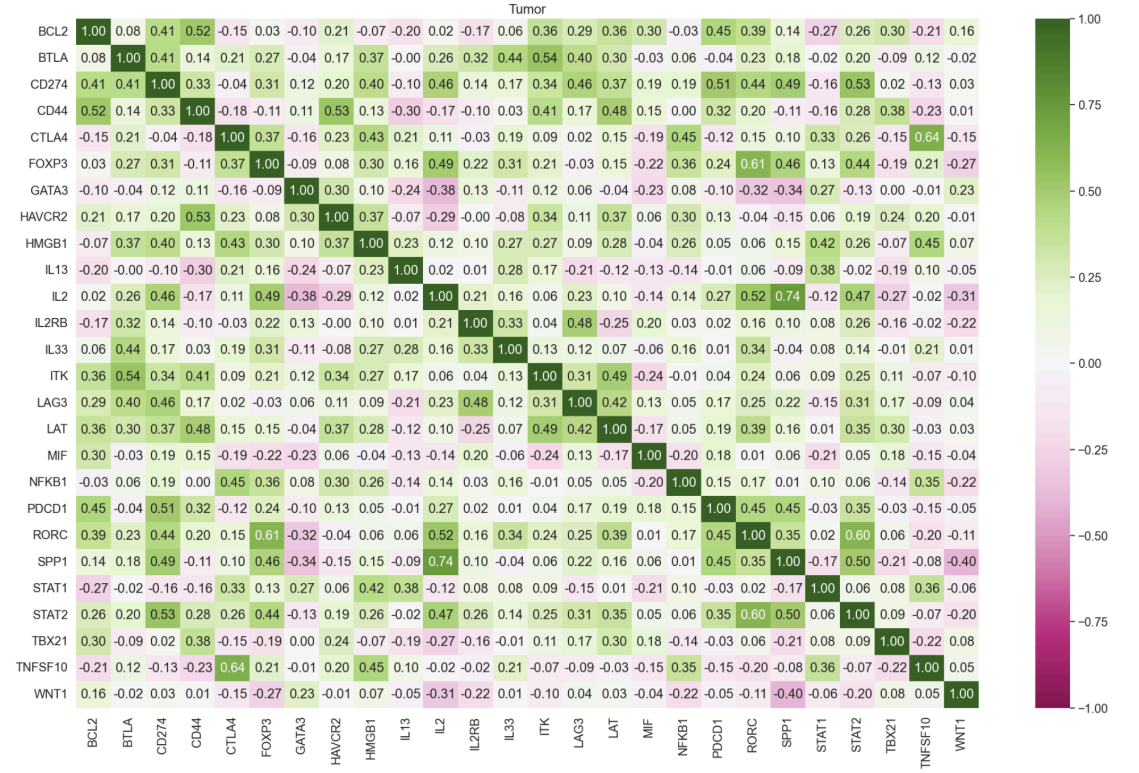


Figure 21: Balanced immune gene dataset correlation plot for tumour patients

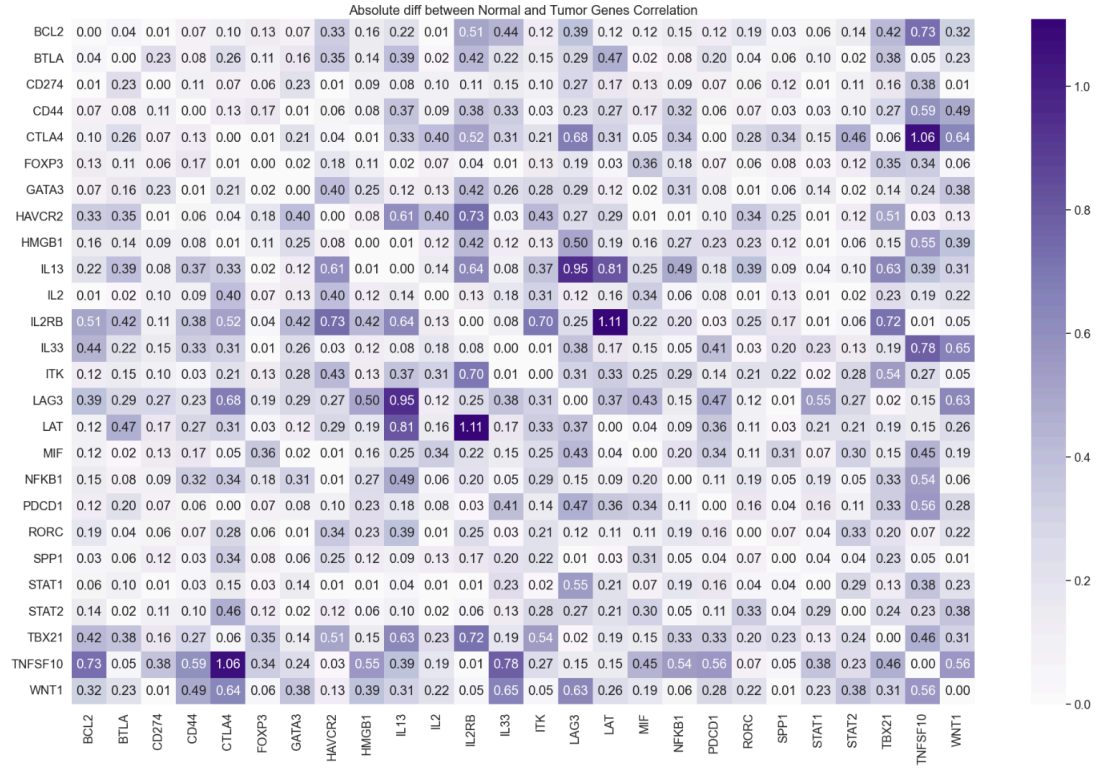


Figure 22: Balanced immune gene dataset correlation plot with absolute values

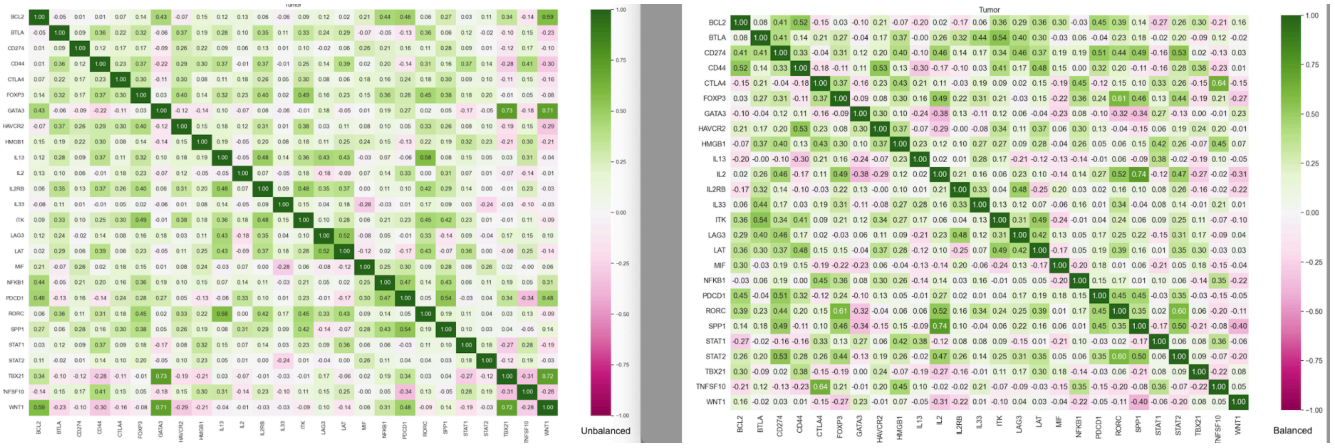


Figure 23: Unbalanced immune gene dataset (left) and balanced dataset (right)

It can be observed that the overall pattern for more highly correlated values for normal tissue still exists. Although for tumour tissue, slightly more correlated values appeared in the ranges of 0.50-0.80, shown on Table 3, which suggests a good, strong relationship for the pair of genes: IL2 + SPP1 with correlation 0.74, CTLA4 + TNFSF10 with correlation 0.64, FOXP3 + RORC with correlation 0.61.

For comparison, on unbalanced dataset the most correlated pairs of genes for tumour dataset were the following: GATA3 +TBX21 with correlation 0.73, GATA3 + WNT1 with correlation 0.71, IL13 + RORC with correlation 0.58.

When looking at the absolute difference between Normal and Tumour genes correlation graph for balanced data the most correlated pairs of genes are the following: IL2RB + LAT with correlation 1.11, TNFSF10 + CTLA4 with correlation 1.06, LAG3 + IL13 with correlation 0.95, LAT + IL13 with correlation 0.81. The absolute value difference graph (purple colour) is the most convenient way to search for the most correlated values, as it allows one to see the difference in values for normal and tumour graphs on one graph. It can be observed that for **both** unbalanced and balanced datasets gene **TNFSF10** showed highly correlated values.

Next pairplot graph is analysed for the balanced immune genes dataset, shown on Figure 24.

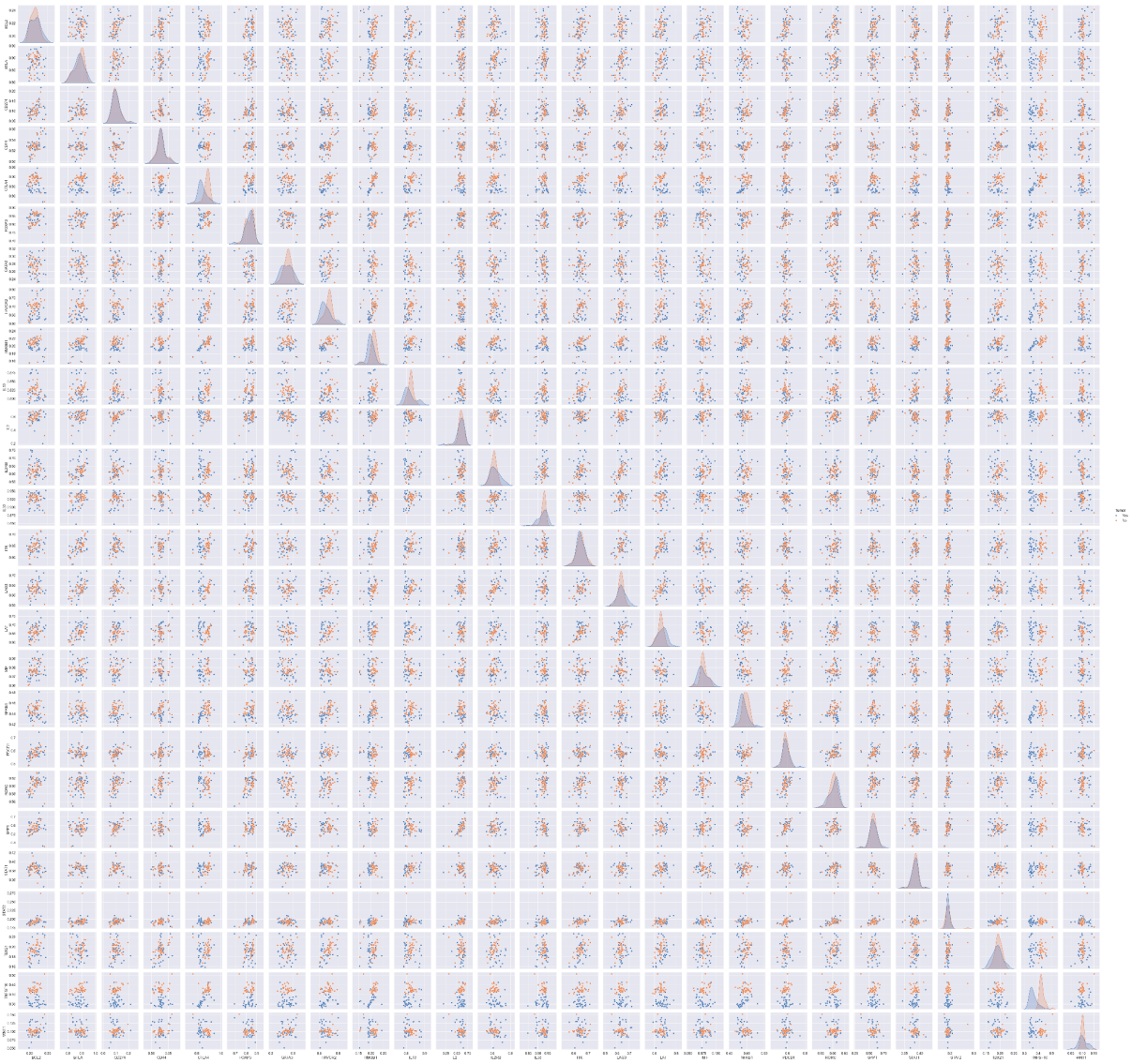


Figure 24: Pairplot graph for Immune genes balanced dataset

On figure 25 zoomed in histograms of genes are shown where average values are significantly different. It can be observed that gene TNFSF10 is also here, confirming that he should be deeply looked at. The rest of genes are: CTLA4, WNT1, IL13, LAT, HAVCR2, some of which had been pointed out in the correlation plot.

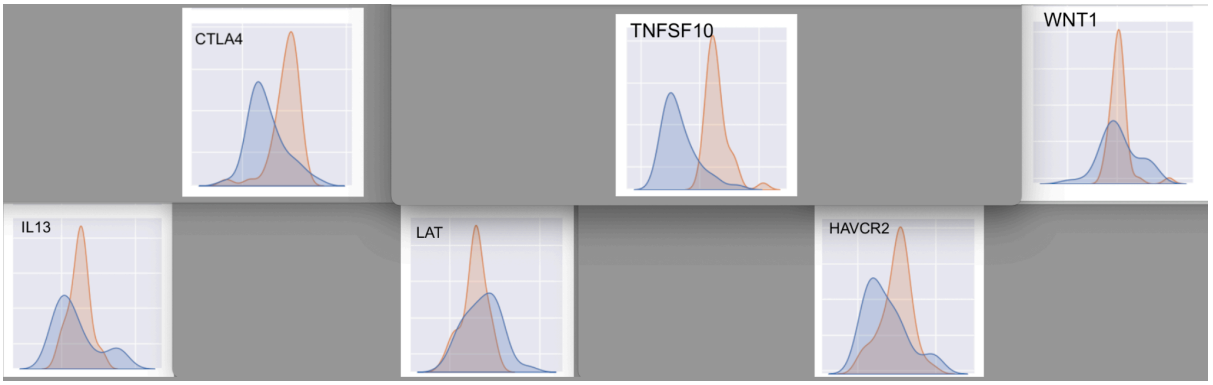


Figure 25: Zoomed in pairplot graph for Immune genes balanced dataset for certain genes

To confirm the significance of certain genes in the analysis, T-test should be done to form the hypothesis and calculate p-value for each gene. Hypothesis testing for immune gene sets is described in detail in the 4.1.4 section of this report.

To conclude the section for the immune gene analysis the following table 4 with significant genes is made to use later in the report.

Table 4: Results in genes significance in immune gene analysis

Immune gene analysis					
Unbalanced dataset			**Balanced dataset**		
Absolute difference correlation plot			**Absolute difference correlation plot**		
Gene name	**Gene name**	**Correlation value**	**Gene name**	**Gene name**	**Correlation value**
TBX21	HAVCR2	0.94	IL2RB	LAT	1.11
TBX21	CD44	0.92	TNFSF10	CTLA4	1.06
TNFSF10	WNT1	0.90	IL13	LAG3	0.95
			LAT	IL13	0.81
Pairplot histogram			**Pairplot histogram**		
TNFSF10			TNFSF10		
HAVCR2			CTLA4		
CTLA4			IL13		
			LAT		
			HAVCR2		
			WNT1		

Section 4.1.2 Methylation genes analysis

The next set of genes for analysis were the genes that affect methylation.

Below the full list of them is shown:

- 'DNMT1',
- 'DNMT3A',
- 'DNMT3B',
- 'DNMT3L',
- 'MBD1',
- 'MBD2',
- 'MBD3',
- 'MBD4',
- 'UHRF1',
- 'UHRF2',
- 'ZBTB4',
- 'ZBTB38',
- 'TET1',
- 'BEND3',
- 'TET2',
- 'TET3',
- 'NLRC5'

The same checks were carried out as for the immune genes dataset. Although in this one no missing values for normal and tumour datasets were found.

It was verified that all genes are present in both datasets.

Section 4.1.2.1 Unbalanced methylation genes dataset

Figure 26 correlation plot for the specified set of genes for normal and tumour tissue is shown.

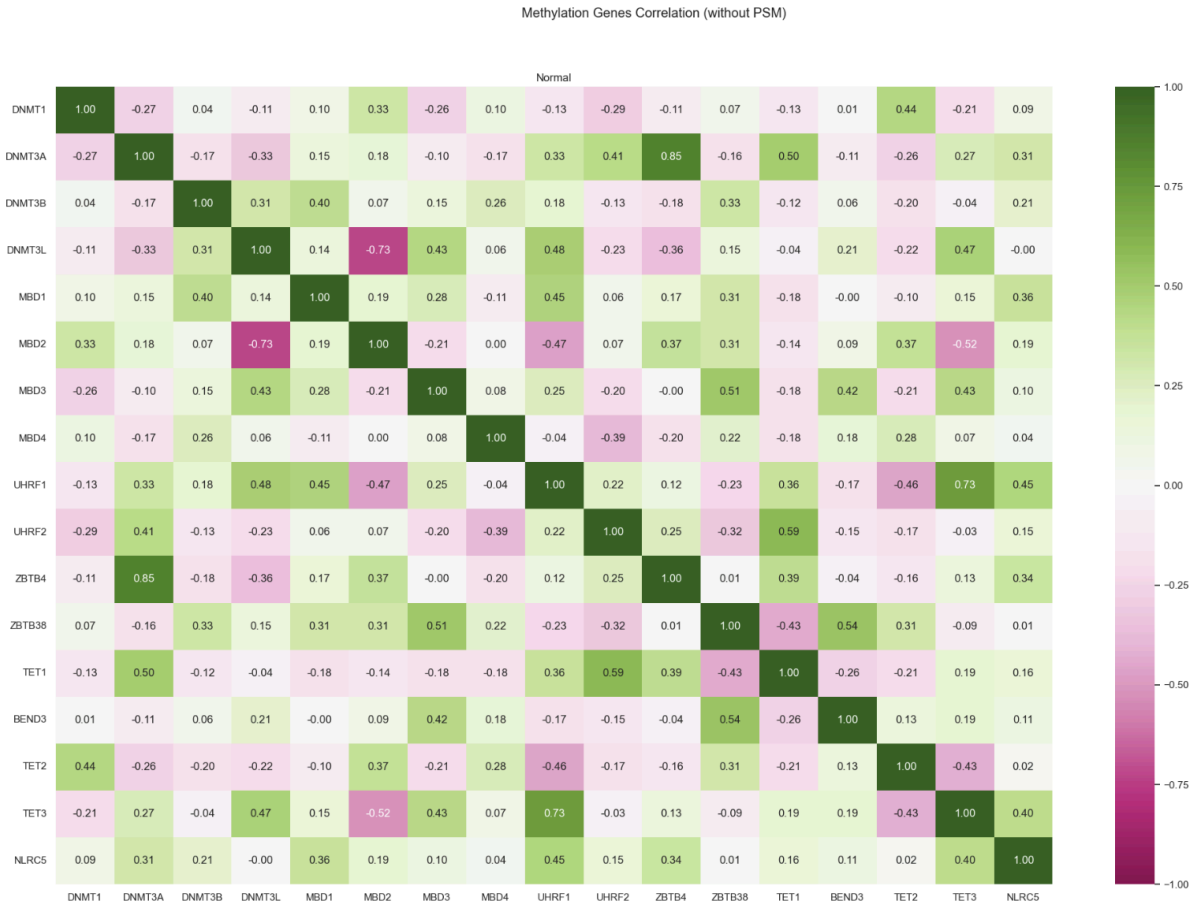


Figure 26: Unbalanced methylation gene dataset correlation plot for normal tissue

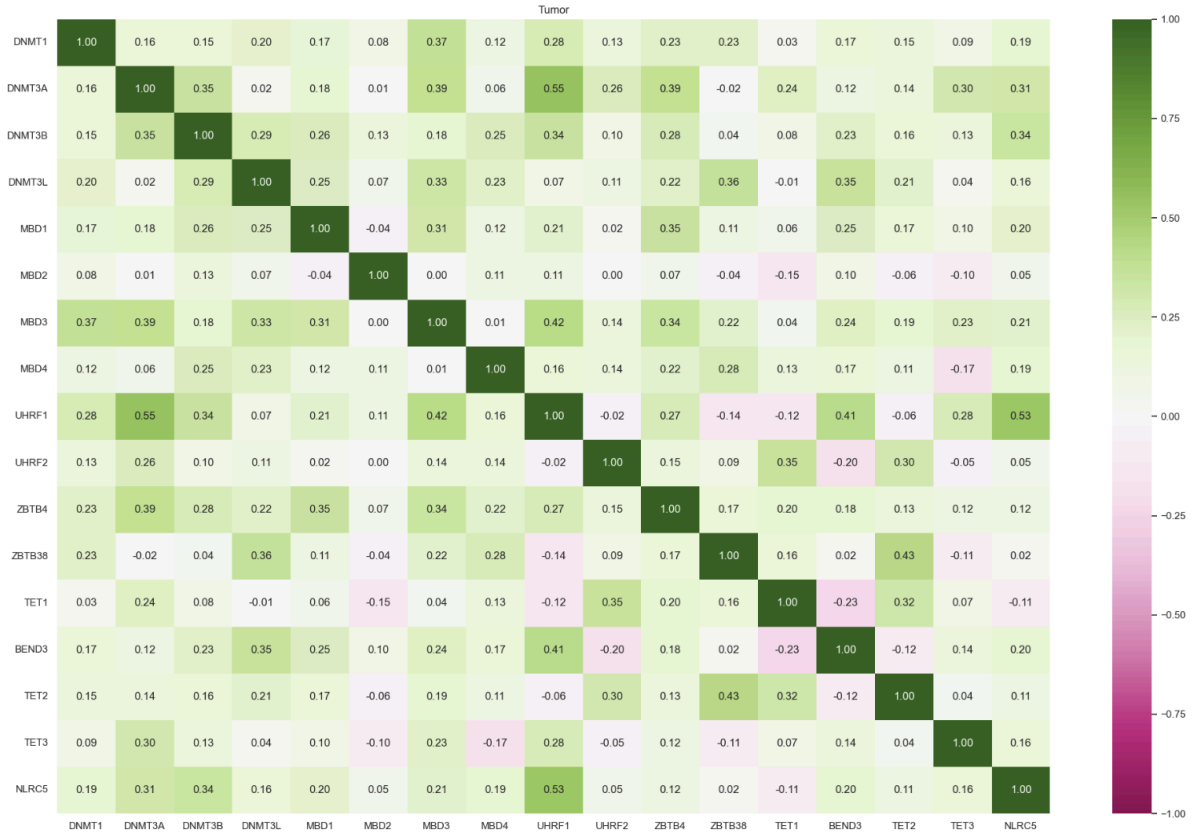


Figure 27: Unbalanced methylation gene dataset correlation plot for tumour tissue

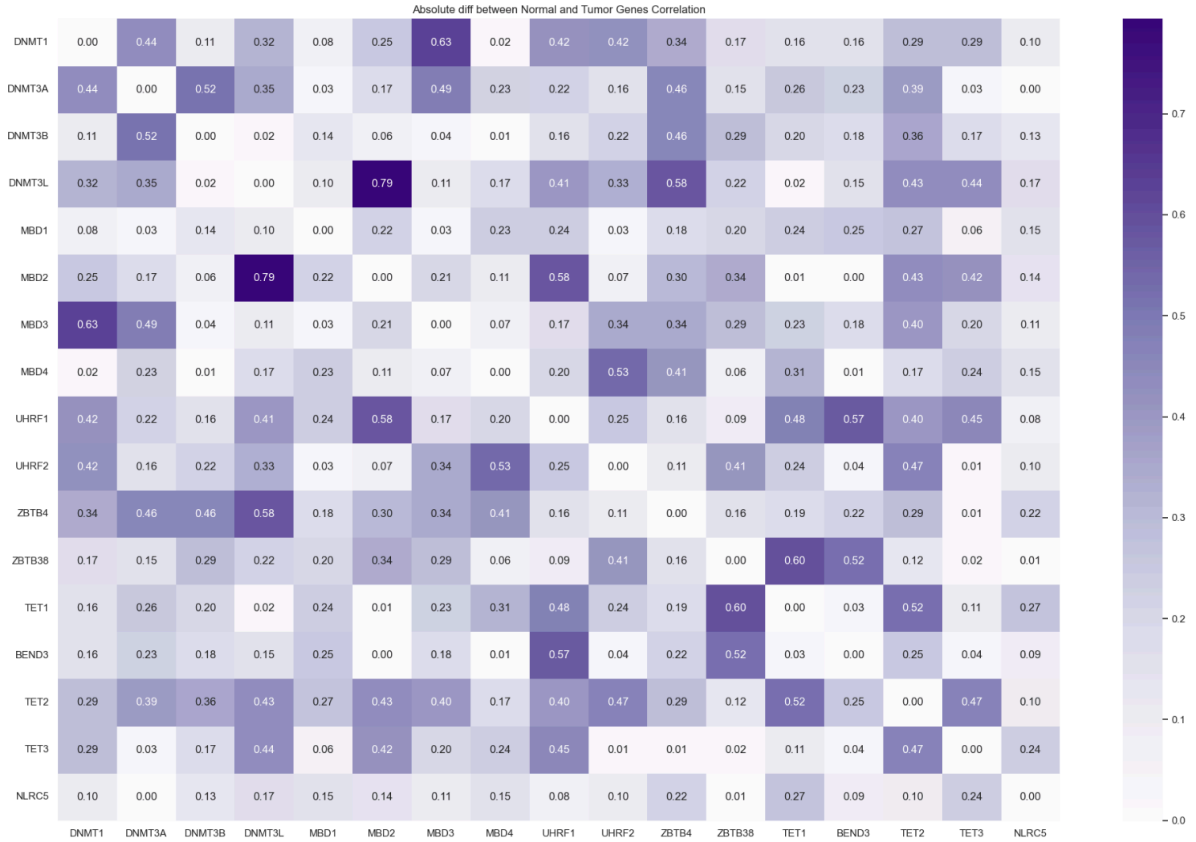


Figure 28: Unbalanced methylation gene dataset correlation plot with absolute value

Upon those graphs the following conclusions can be made.

Normal Tissue:

Very strong positive relationship, where one gene's methylation value rises, the other value rises as well.

DNMT3A + ZBTB38 with positive correlation 0.85

TET3 + UHRF1 with positive correlation 0.73

Very strong negative relationship, where one methylation value goes up, the other one goes down.

MBD2 + MNMT3L with negative correlation -0.73

Tumour Tissue:

Strong positive relationship, where one gene's methylation value rises, the other value rises as well.

NLRC5 + UHRF1 with positive correlation 0.53

DNMT3A + UHRF1 with positive correlation 0.53

Absolute values difference for Normal and Tumour unbalanced datasets:

Very strong and strong positive relationship, where one gene's methylation value rises, the other value rises as well.

DNMT3L + MBD2 with positive correlation 0.79

DNMT1 + MBD3 with positive correlation 0.63

ZBTB38 + TET1 with positive correlation 0.60

A Pairplot graph is also built and shown on figure 29.

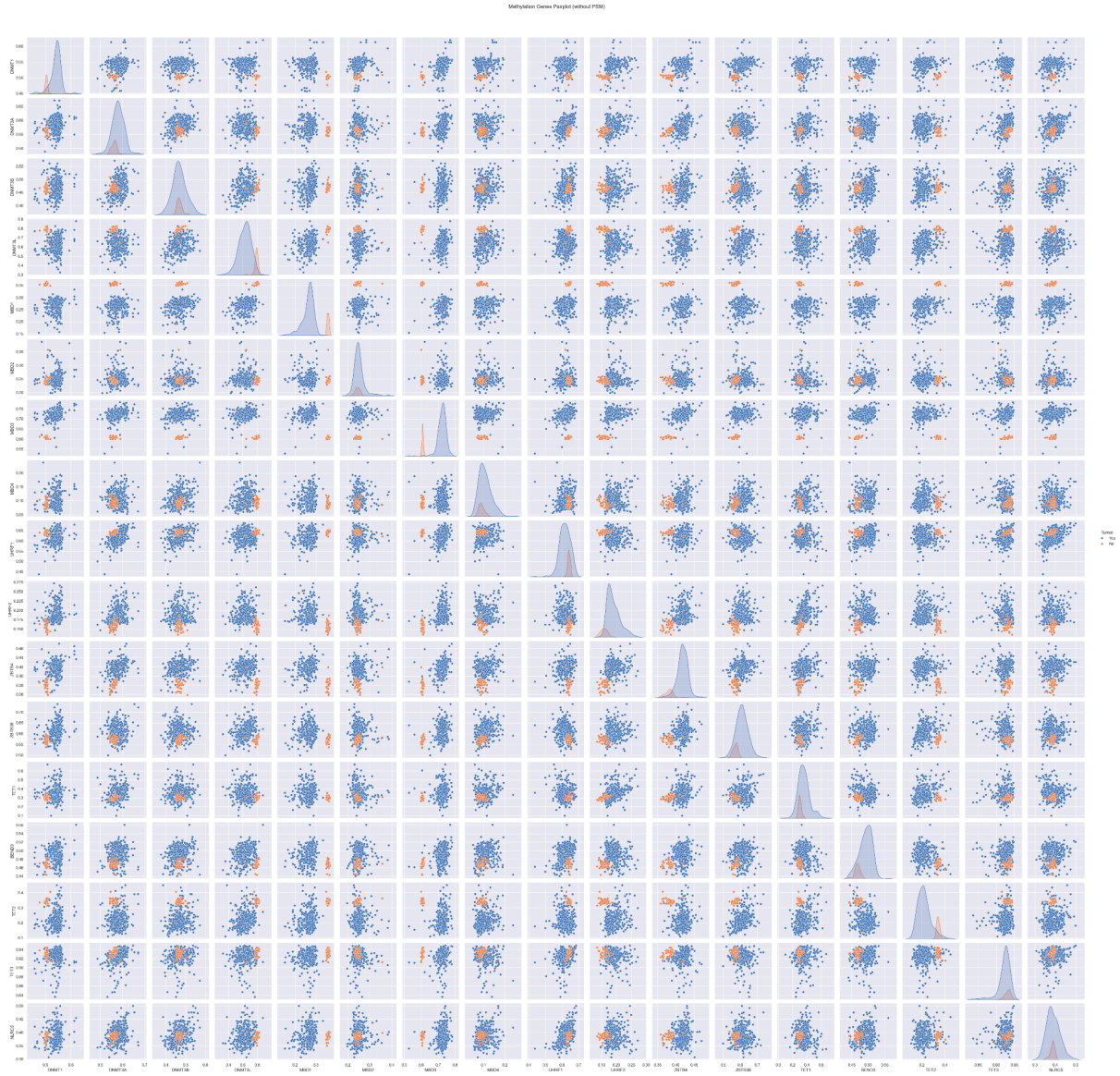


Figure 29: Pairplot graph for unbalanced methylation genes dataset

Big differences in average values for both datasets were found in the following genes - **DNMT1, DNMT3L, MBD1, MBD3, UHRF2, ZBTB4, TET2**. Zoomed in histograms are shown on figure 30.

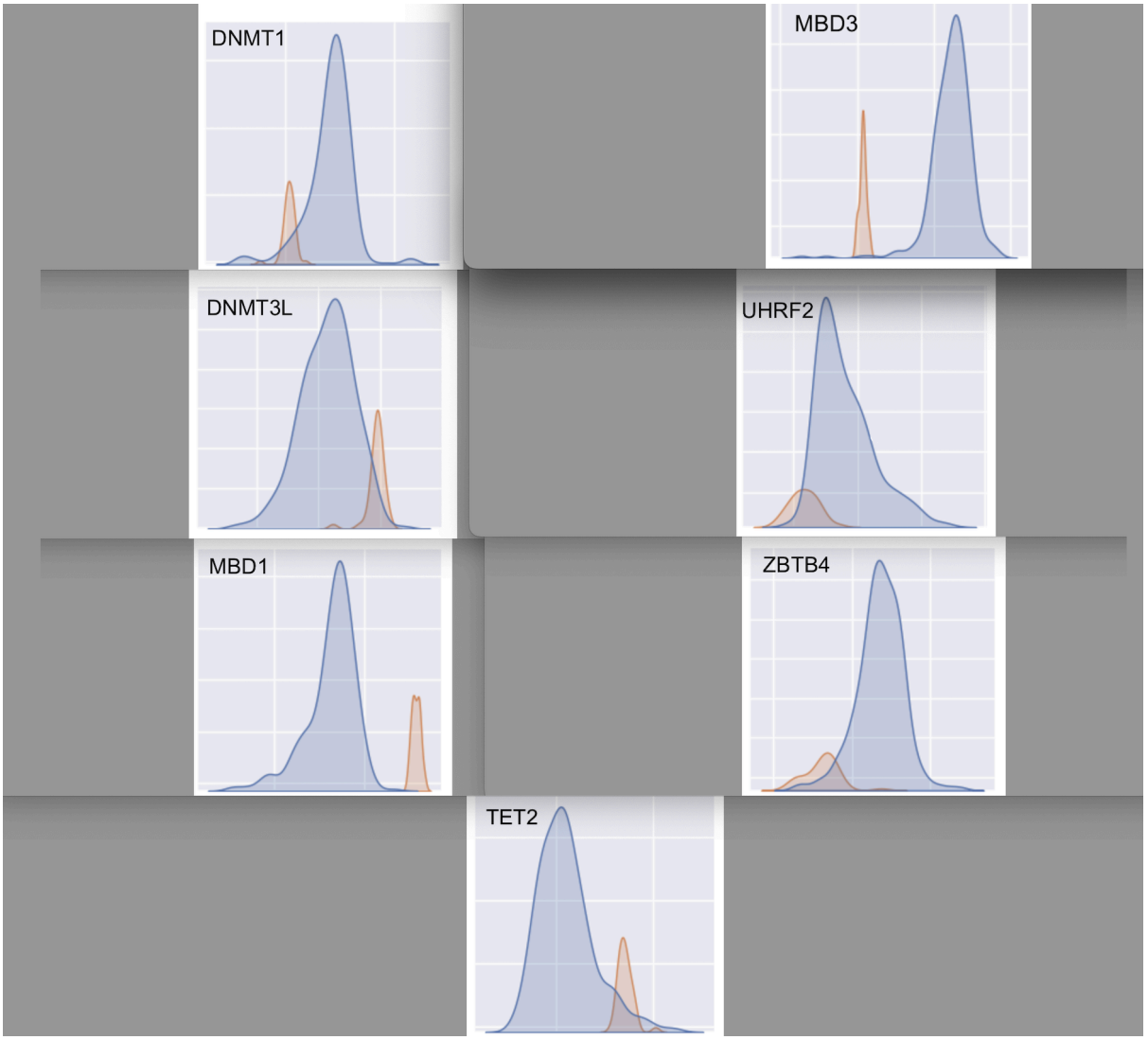


Figure 30: Zoomed in histograms of some genes from a pairplot graph for unbalanced methylation genes dataset

Section 4.1.2.2 Balanced methylation genes dataset

It was verified in the 4.1.4 section of this report, that balancing the datasets makes the analysis more accurate, thus a balanced dataset using propensity score matching is built for the methylation set of genes.

Dataset overview before the propensity score matching is shown on figure 31.

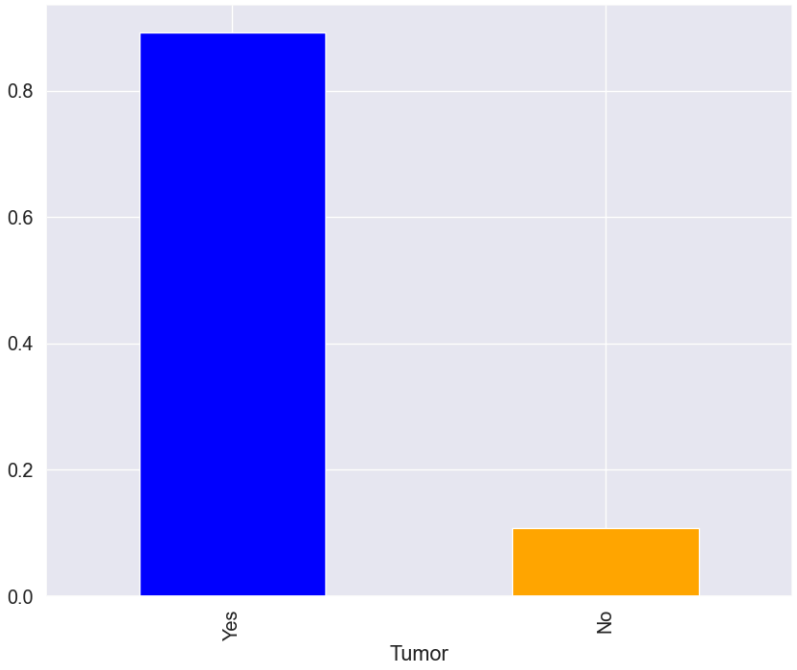


Figure 31: Unbalanced methylation genes dataset overview

Propensity score matching is run on each gene individually. For the limited space of this thesis only 3 graphs out of 17 will be shown on figures 32-34.

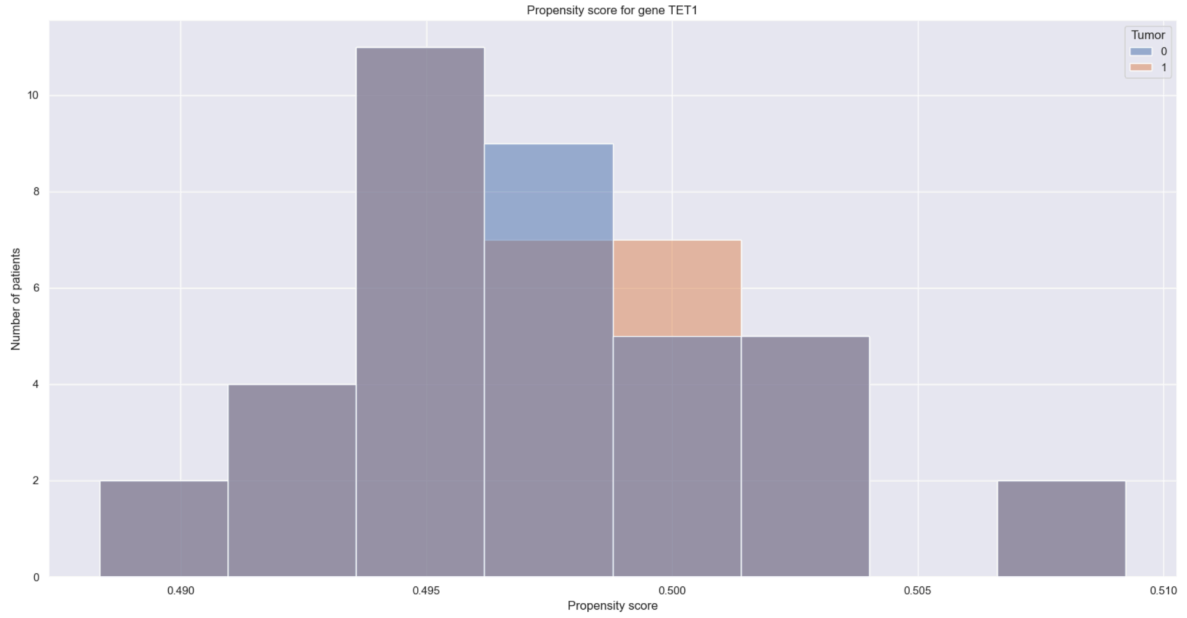


Figure 32: Propensity score matching for gene TET1

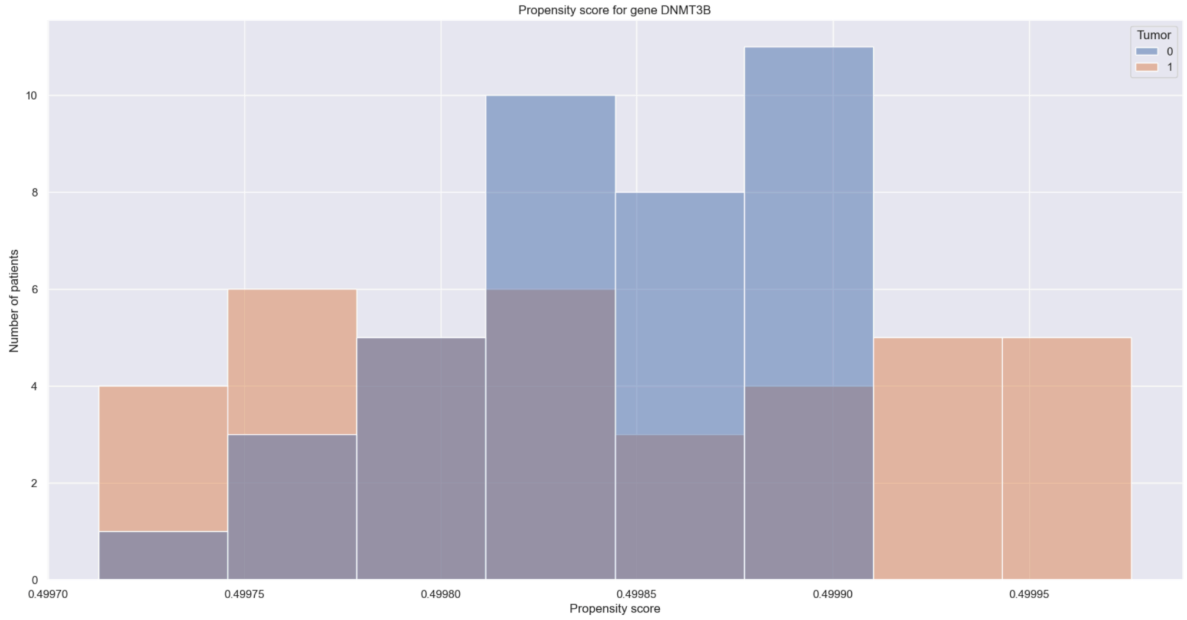


Figure 33: Propensity score matching for gene DNMT3B

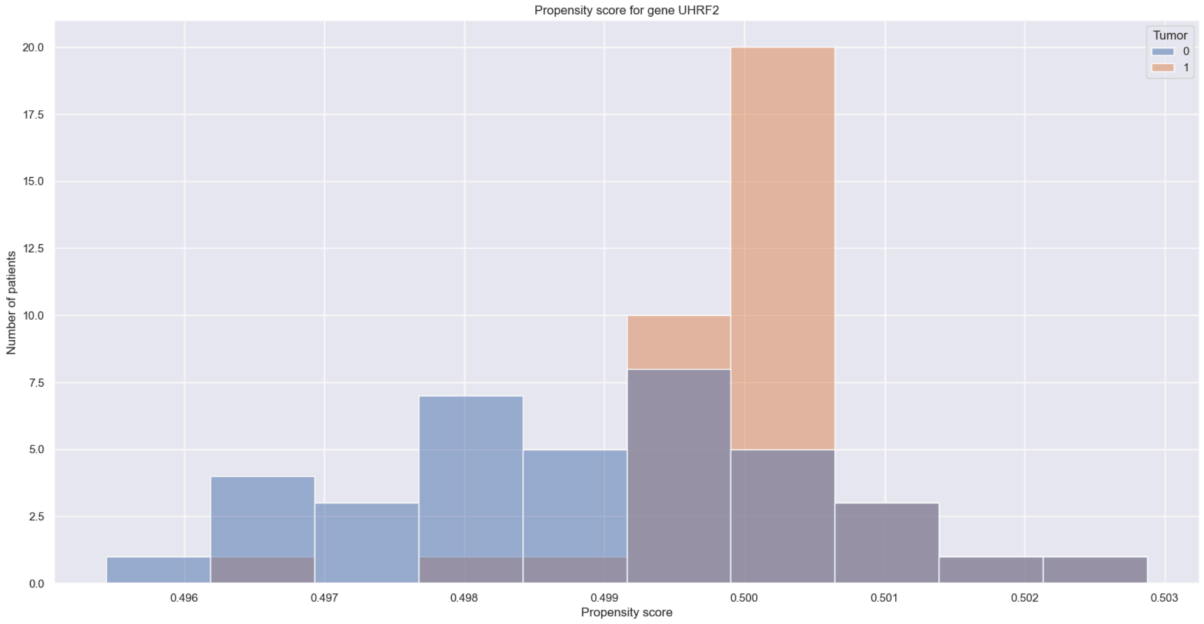


Figure 34: Propensity score matching for gene UHRF2

As it can be observed from the graphs all genes had a major overlap of propensity score values. This means that there are enough matches of the data and the propensity score algorithm is valid to use. After running the matching algorithm both datasets show equal size, as shown on figure 35.

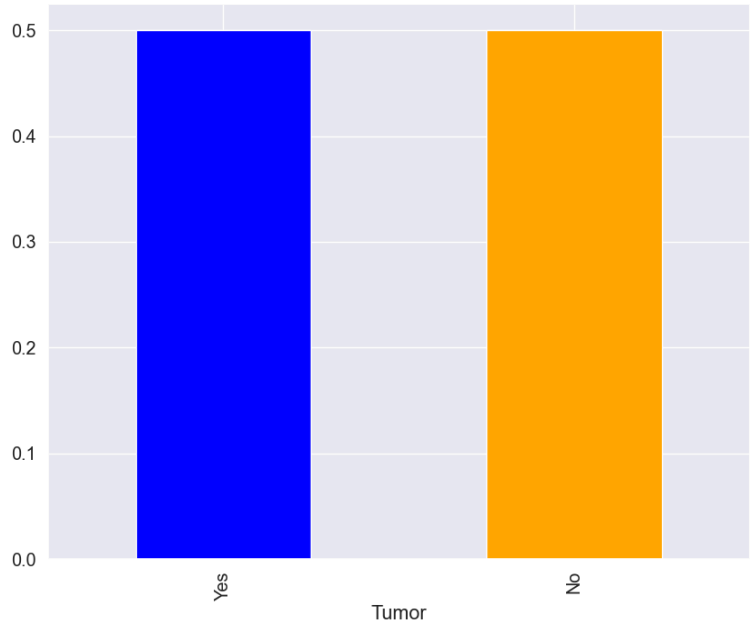


Figure 35: Balanced methylation genes dataset overview

Next correlation graph after propensity score matching is shown on figure 36.

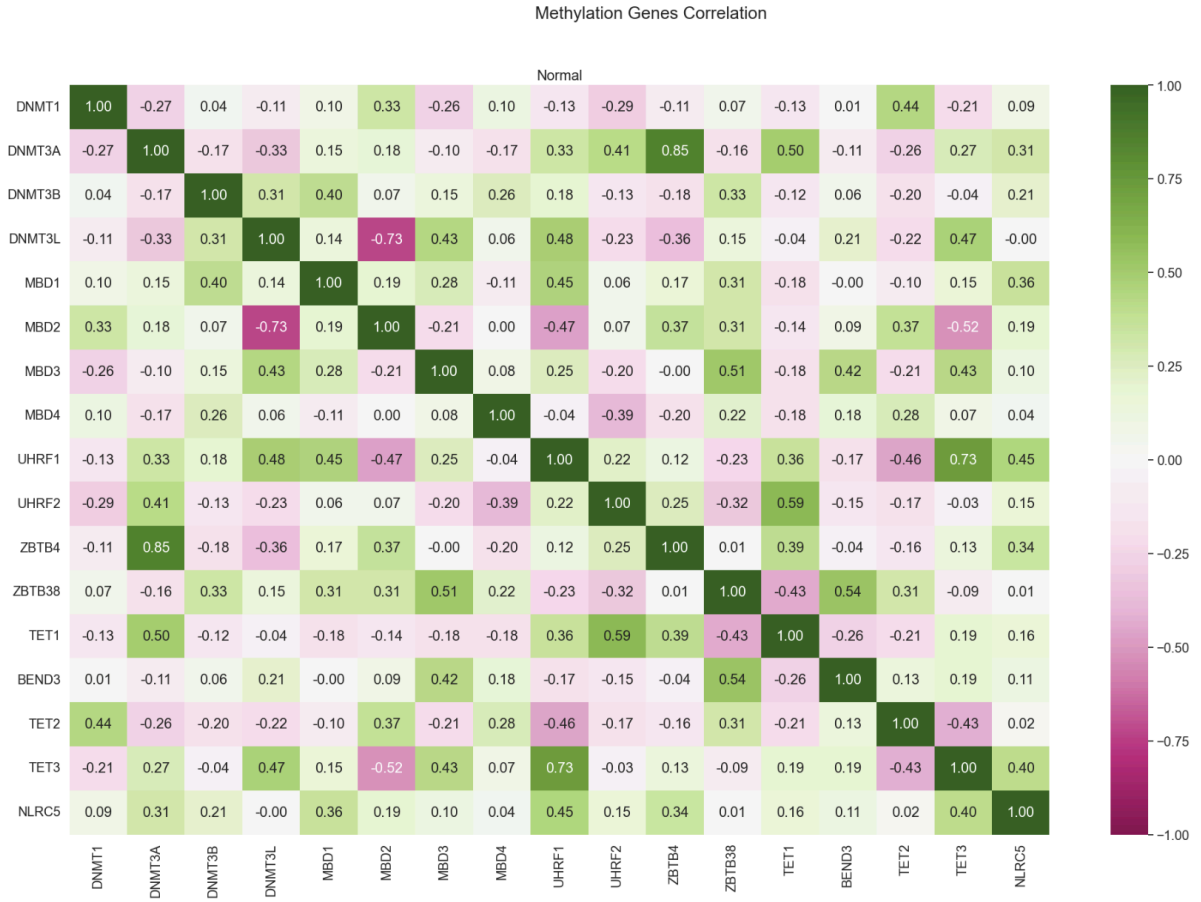


Figure 36: Balanced methylation gene dataset correlation plot for normal tissue


Tumor
DNMT1 1.00 -0.08 -0.11 -0.26 -0.34 0.19 0.26 0.08 0.02 0.15 0.29 -0.07 0.09 -0.06 -0.01 -0.10 0.16
DNMT3A -0.08 1.00 -0.05 0.13 0.05 -0.17 0.15 -0.17 0.50 0.18 0.30 -0.17 0.62 -0.20 -0.28 -0.06 0.02
DNMT3B -0.11 -0.05 1.00 -0.05 -0.09 0.05 0.04 0.08 0.04 -0.03 -0.00 0.05 -0.14 0.11 0.28 0.01 0.17
DNMT3L -0.26 0.13 -0.05 1.00 0.27 -0.04 -0.35 -0.23 0.23 -0.36 -0.44 0.03 0.04 0.03 0.03 0.12 0.06
MBD1 -0.34 0.05 -0.09 0.27 1.00 -0.25 -0.51 -0.22 0.08 -0.15 -0.49 0.00 0.03 0.03 0.08 0.01 0.00
MBD2 0.19 -0.17 0.05 -0.04 -0.25 1.00 0.17 0.03 -0.14 -0.08 0.01 0.24 -0.34 -0.00 0.27 -0.01 -0.22
MBD3 0.26 0.15 0.04 -0.35 -0.51 0.17 1.00 0.22 0.08 0.48 0.40 0.02 0.20 -0.33 -0.11 0.23 -0.17
MBD4 0.08 -0.17 0.08 -0.23 -0.22 0.03 0.22 1.00 -0.25 -0.01 0.06 0.17 -0.06 0.20 0.05 -0.00 -0.11
UHRF1 0.02 0.50 0.04 0.23 0.08 -0.14 0.08 -0.25 1.00 0.15 0.02 0.03 0.23 -0.22 -0.25 0.26 0.06
UHRF2 0.15 0.18 -0.03 -0.36 -0.15 -0.08 0.48 -0.01 0.15 1.00 0.41 -0.13 0.27 -0.32 -0.21 0.26 -0.04
ZBTB4 0.29 0.30 -0.00 -0.44 -0.49 0.01 0.40 0.06 0.02 0.41 1.00 0.08 0.15 -0.01 -0.14 -0.19 0.05
ZBTB38 -0.07 -0.17 0.05 0.03 0.00 0.24 0.02 0.17 0.03 -0.13 0.08 1.00 -0.48 0.51 0.46 0.15 -0.24
TET1 0.09 0.62 -0.14 0.04 0.03 -0.34 0.20 -0.06 0.23 0.27 0.15 -0.48 1.00 -0.34 -0.38 0.01 0.04
BEND3 -0.06 -0.20 0.11 0.03 0.03 -0.00 -0.33 0.20 -0.22 -0.32 -0.01 0.51 -0.34 1.00 0.34 -0.01 0.04
TET2 -0.01 -0.28 0.28 0.03 0.08 0.27 -0.11 0.05 -0.25 -0.21 -0.14 0.46 -0.38 0.34 1.00 -0.09 -0.29
TET3 -0.10 -0.06 0.01 0.12 0.01 -0.01 0.23 -0.00 0.26 0.26 -0.19 0.15 0.01 -0.01 -0.09 1.00 -0.20
NLRC5 0.16 0.02 0.17 0.06 0.00 -0.22 -0.17 -0.11 0.06 -0.04 0.05 -0.24 0.04 0.04 -0.29 -0.20 1.00
DNMT1 DNMT3A DNMT3B DNMT3L MBD1 MBD2 MBD3 MBD4 UHRF1 UHRF2 ZBTB4 ZBTB38 TET1 BEND3 TET2 TET3 NLRC5
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Figure 37: Balanced methylation gene dataset correlation plot for tumour tissue

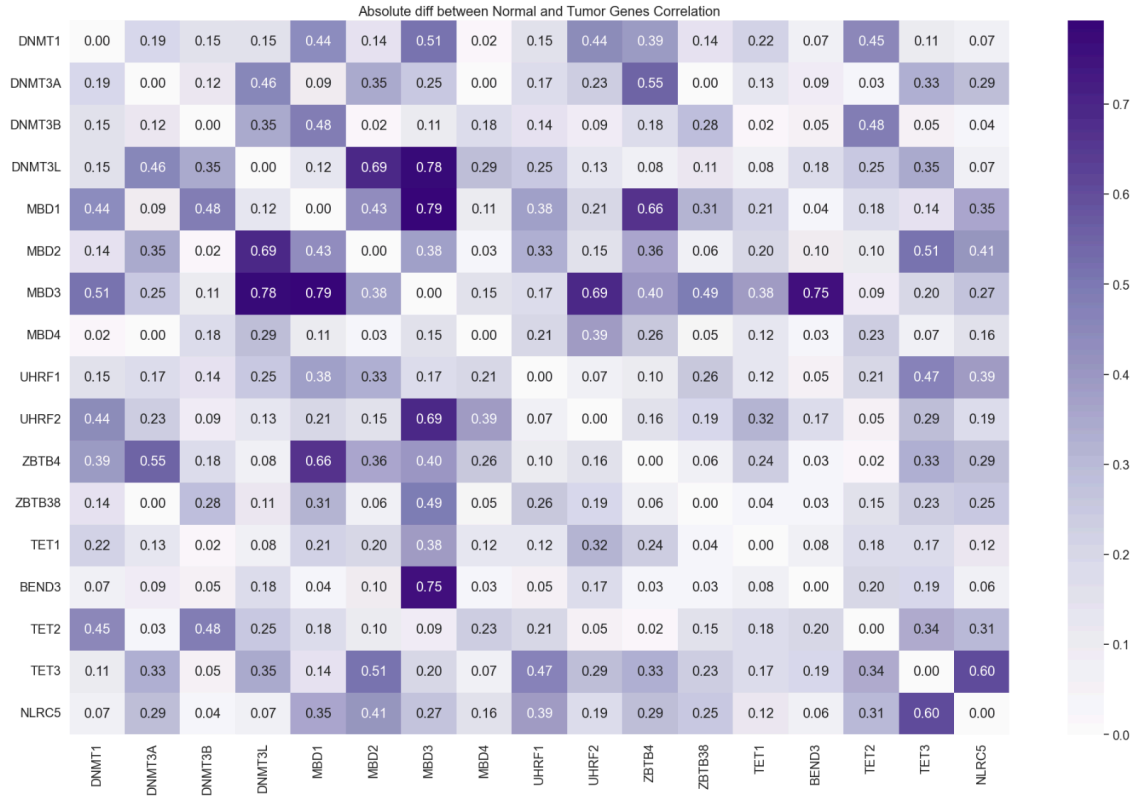


Figure 38: Balanced methylation gene dataset correlation plot with absolute values

From the correlation graphs a strong pattern among MBD3 gene (in the absolute value) can be seen and DNMT3A gene in the tumour dataset.

Strong and very strong relationships can be seen in the following set of genes.

Normal tissue dataset:

One gene's methylation value rises, the other value rises as well.

DNMT3A + ZBTB4 with positive correlation 0.85

TET3 + UHRF1 with positive correlation 0.73

UHRF2 + TET1 with positive correlation 0.59

When one methylation value goes up, the other one goes down

MBD2 + DNMT3L with negative correlation -0.73

MBD2 + TET3 with negative correlation -0.52

Tumour tissue dataset:

One gene's methylation value rises, the other value rises as well.

DNMT3A + TET1 with positive correlation 0.62

DNMT3A + UHRF1 with positive correlation 0.50

When one methylation value goes up, the other one goes down

MBD1 + MBD3 with negative correlation -0.51

ZBTB4 + MBD1 with negative correlation -0.49

Absolute values for both datasets:

One gene's methylation value rises, the other value rises as well.

MBD1 + MBD3 with positive correlation 0.79

DNMT3L + MBD3 with positive correlation 0.78

BEND3 + MBD3 with positive correlation 0.75

UHRF2 + MBD3 with positive correlation 0.69

Then the pairplot graph was made for a balanced methylation genes dataset.

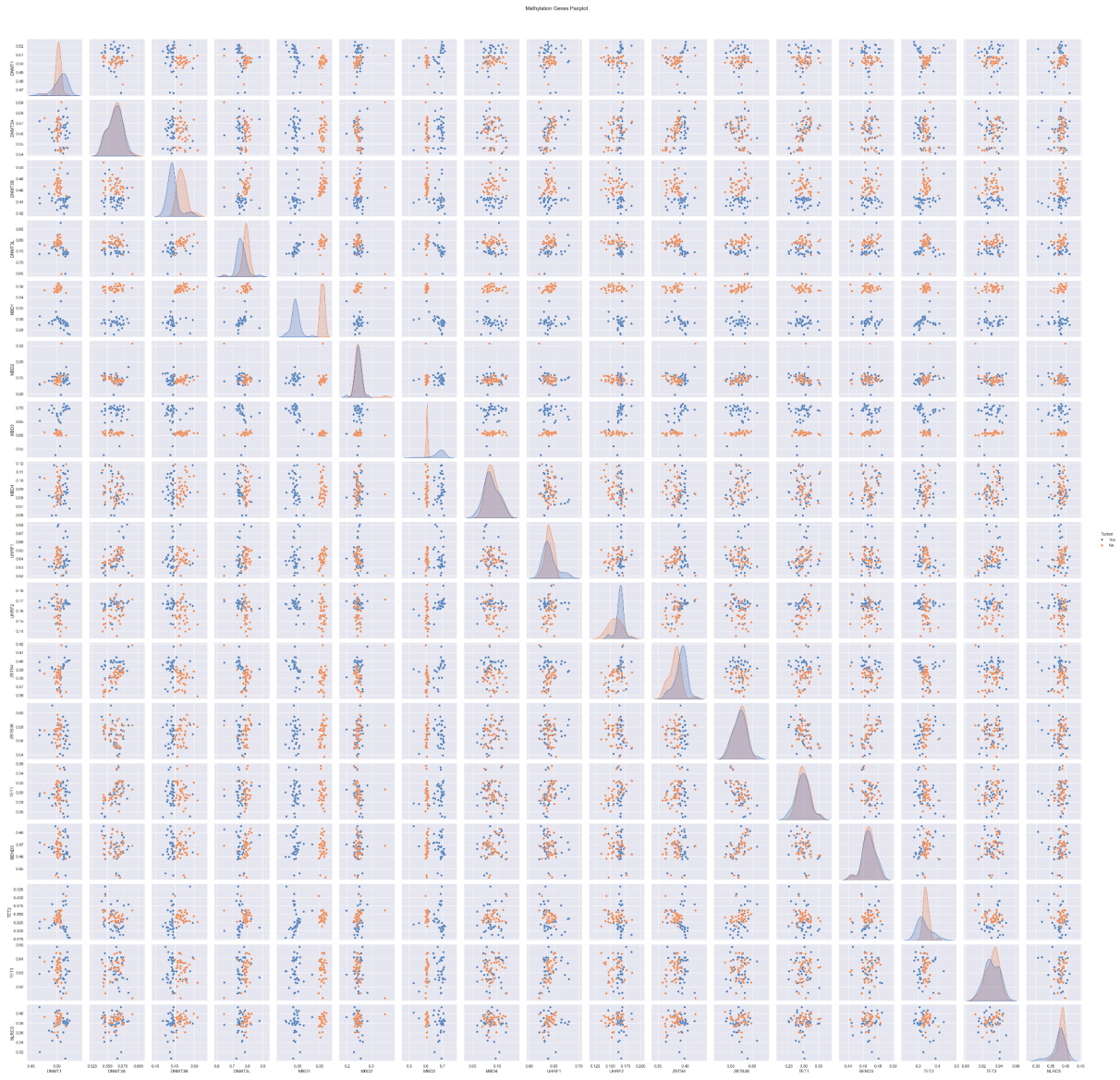


Figure 39: Pairplot graph for balanced methylation genes set

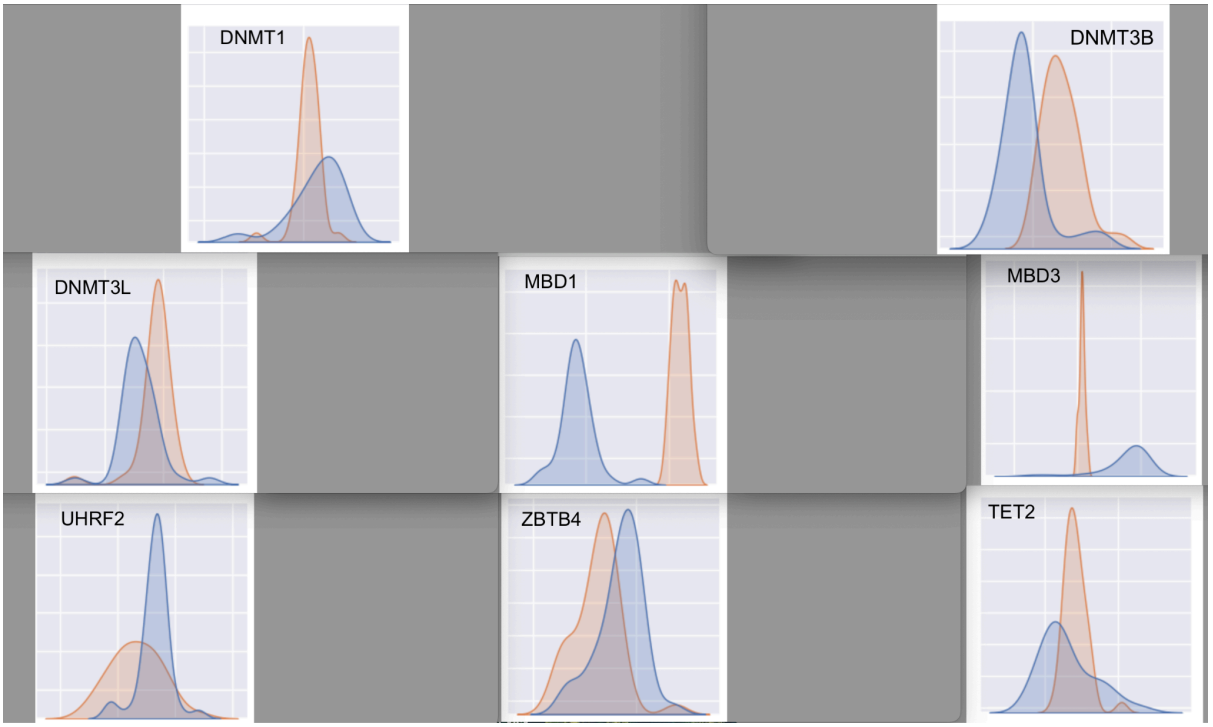


Figure 40: Zoomed in focus on the significant genes histograms from the pairplot graph for balanced methylation genes set

The biggest difference in average value ranges for normal and tumour patients were found in the following genes: **MBD1, MBD3, DNMT3B, DNMT3L, ZBTB4, UHRF2, DNMT1, TET2.**

To confirm the significance of these genes in the methylation gene analysis, T-test should be done to form the hypothesis and calculate p-value for each gene. Hypothesis testing for methylation gene sets is described in detail in the 4.1.3 section of this report.

To conclude the section for the immune gene analysis the following table 5 with significant genes is made to use later in the report.

Table 5: Methylation gene analysis

Methylation gene analysis					
Unbalanced dataset			Balanced dataset		
Absolute difference correlation plot			Absolute difference correlation plot		
Gene name	Gene name	Correlation value	Gene name	Gene name	Correlation value
DNMT3L	MBD2	0.79	MBD3	MBD1	0.79
DNMT1	MBD3	0.69	MBD3	DNMT3L	0.78
ZBTB38	TET1	0.60	MBD3	BEND3	0.75
			MBD3	UHRF2	0.69
Pairplot histogram			Pairplot histogram		
DNMT1			MBD1		
DNMT3L			MBD3		
MBD1			DNMT3B		
MBD3			DNMT3L		
UHRF2			ZBTB4		
ZBTB4			UHRF2		
TET2			DNMT1		
			TET2		

Section 4.1.3 Hypothesis testing using T-test for methylation set of genes

To answer the question if there are differences in genes between Normal and Tumour tissue, a T-test is used. It is a statistical method to see the differences in means between two groups (Hosseini, 2023).

T-test is calculated in Python using “scipy.stats.” library with the function “ttest_ind”. It calculates the T-test for means of 2 independent datasets, in the current example for normal and tumour tissue and shows the difference between them. This test is used for setting the null hypothesis indicating that 2 samples (independent) have the same average values by default (*ttest_ind — SciPy v1.14.0 Manual*, no date).

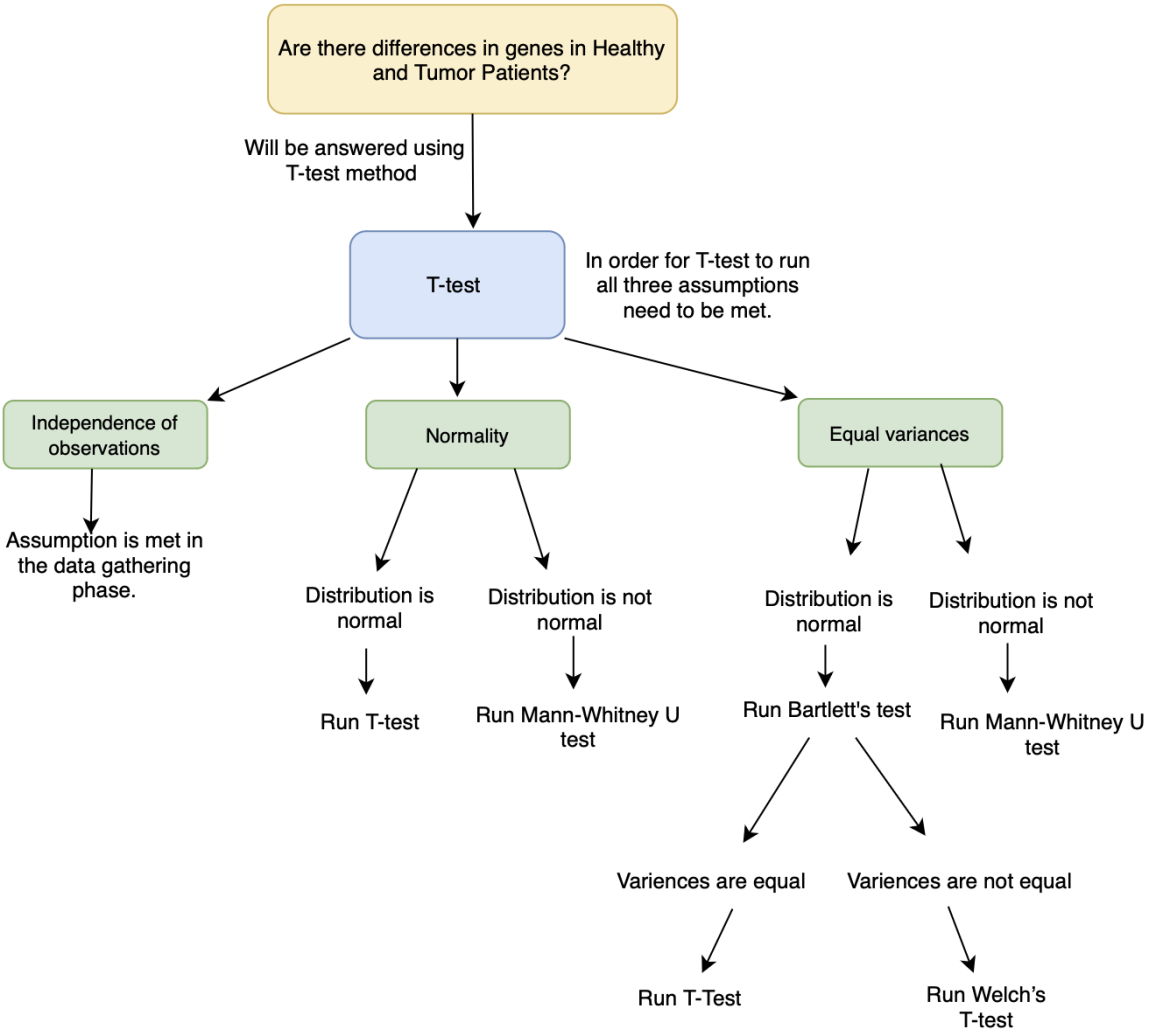


Figure 41: T-test flow chart

In order to perform a T-test all three criteria should be met for results to be considered valid:

- Independence of observations
- Normality
- Equal variances

1. **Independence of observations.**

Both datasets for normal and tumour tissue are completely independent. This was observed on the stage of downloading data from TCGA.

2.**Normality.**

In order to meet this criteria data should be in a form of normal distribution. To check normality, a normal test is run in Python and the “scipy.stats.normaltest” function is used. Normal test checks if the sample is different from the normal distribution. The null hypothesis in this function is defined as a sample following normal distribution. This is based on D’Agostino and Pearson’s test which combines skew and kurtosis to produce an omnibus test of normality (*normaltest — SciPy v1.14.0 Manual*, no date). Normal tests are run by each gene individually for normal and tumour tissue. Input data is the gene data for each gene individually for normal tissue and separately for tumour tissue. And in the output p-value is calculated for the corresponding gene and patient. Then distributions are tested for normality as illustrated on figure 42.

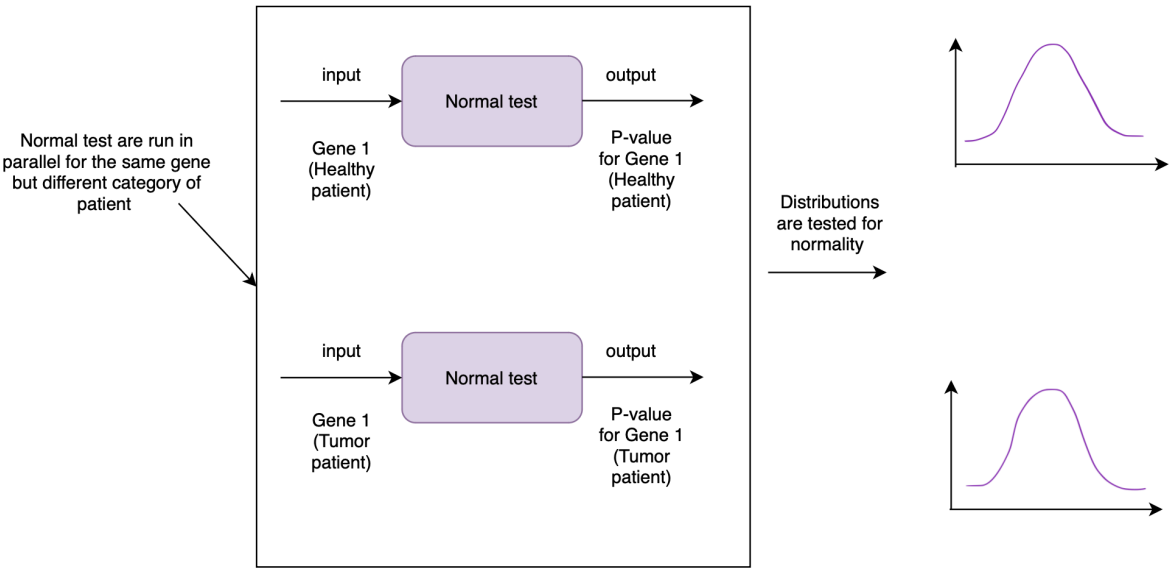


Figure 42: Normality test explanation, showing input and output values

Hypothesis for normality:

Null hypothesis (H0): The sample comes from a normal distribution.

Alternative hypothesis (H1): The sample does not come from a normal distribution.

Interpretation:

P-value:

- If the p-value is greater than 0.05, fail to reject the null hypothesis (H0).
- If the p-value is 0.05 or less, reject the null hypothesis (H0).

	Normal Patients statistic	Normal Patients p-value	Tumor Patients statistic	Tumor Patients p-value	Normal Patients Fail to reject H0	Tumor Patients Fail to reject H0	Both Fail to reject H0
DNMT1	32.338972	0.0001	16.278461	0.0032	False	False	False
DNMT3A	0.006032	0.9969	1.384035	0.4804	True	True	True
DNMT3B	11.245193	0.0119	24.504989	0.0003	False	False	False
DNMT3L	51.771143	0.0001	13.742832	0.0063	False	False	False
MBD1	1.669963	0.4080	17.344560	0.0033	True	False	False
MBD2	73.631914	0.0001	9.456222	0.0197	False	False	False
MBD3	0.087898	0.9558	32.988381	0.0001	True	False	False
MBD4	2.318973	0.2797	0.703757	0.6878	True	True	True
UHRF1	1.182754	0.5323	11.184782	0.0146	True	False	False
UHRF2	0.159108	0.9262	12.386189	0.0072	True	False	False
ZBTB4	6.981779	0.0393	4.748825	0.0914	False	True	False
ZBTB38	0.965725	0.5994	0.272192	0.8740	True	True	True
TET1	3.791320	0.1361	0.796015	0.6624	True	True	True
BEND3	6.989299	0.0406	2.172236	0.3051	False	True	False
TET2	24.505823	0.0003	6.174073	0.0562	False	True	False
TET3	3.007782	0.2050	1.664744	0.4069	True	True	True
NLRC5	10.328065	0.0133	17.375270	0.0017	False	False	False

Figure 43: Normality check on the methylation set of genes

On figure 43 it can be seen that DNMT3A, MBD4, ZBTB38, TET1, TET3 genes fail to reject the null hypothesis, which means their samples follow normal distribution. On those genes T-test will be run. On the rest of genes, which do not follow normal distribution, Mann-Whitney U test is run, with the corresponding Python function "scipy.stats.mannwhitneyu" (*normaltest — SciPy v1.14.0 Manual*, no date). This test is used for non-normal data and compares distributions of independent samples in comparison to Wilcoxon Signed-Rank Test that compares related samples (as an example: before and after the treatment) (*Mann Whitney U Test (Wilcoxon Rank Sum Test)*, no date).

3. Equal variances.

The variances of both datasets should be equal. To check the equality of variances Bartlett's test is used with the following Python function "scipy.stats.bartlett" (*bartlett — SciPy v1.14.0 Manual*, no date).

If the variances are not equal, Welch's t-test is used because it does not assume equal variances (Hosseini, 2023).

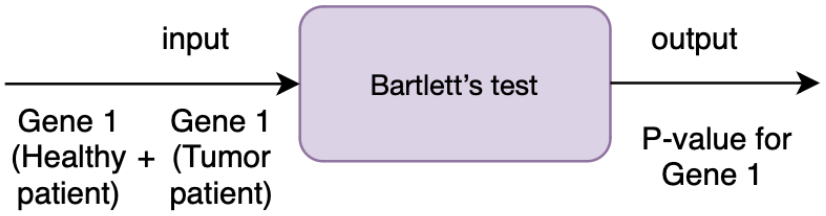


Figure 44: Bartlett's test explanation, showing input and output values

Hypothesis:

- Null hypothesis (H0): Sample variances are equal.
- Alternative hypothesis (H1): Sample variances are not equal.

Interpretation:

P-value:

- If p-value is greater than 0.05, fail to reject the null hypothesis (H0).
- If p-value is 0.05 or less, reject the null hypothesis (H0).

After running Bartlett's test, genes on figure 45 show those genes that have equal variances (failed to reject the null hypothesis). On those genes T-test will be run. In the current example no genes with unequal variance were found, so Welch's test is not performed.

	Test Name	statistic	p-value	Fail to reject H0
MBD4	Bartlett	1.140965	0.285448	True
TET1	Bartlett	0.56194	0.45348	True
TET3	Bartlett	0.472272	0.491944	True
ZBTB38	Bartlett	0.447688	0.503435	True
DNMT3A	Bartlett	0.002964	0.956581	True

Figure 45: Equal variances check on the methylation genes set

T-test

Tests for independence of observations, normality and equal variances were successfully run on 5 genes. Upon those genes T-test is performed, on the rest corresponding tests as described before are run. Results can be seen in the table on figure 46. The significance level is chosen to be 0.05. The test is 2-tail because it is easier to determine if there is any difference between the groups, examining both directions of data ranges (Hayes, 2024).

Hypothesis:

- Null hypothesis (H0): There is no difference in gene expression between normal and tumour tissues (no effect).
- Alternative hypothesis (H1): There is a difference in gene expression (effect present).

T-test interpretation:

- p-value:
    - If p-value is greater than 0.05, fail to reject the null hypothesis (H0). There is no statistically significant difference in gene expression between normal and tumour tissues.
    - If p-value is 0.05 or less, reject the null hypothesis (H0). There is a statistically significant difference in gene expression between normal and tumour tissues.
- t-statistic:
    - If t-statistic is less than 0, the mean of the normal group is less than the mean of the tumour group.
    - If t-statistic is greater than 0, the mean of the normal group is greater than the mean of the tumour group.

	gene	Test Name	p-value	normal genes size	tumor genes size	Fail to reject H0
4	MBD1	Mann-Whitney U rank test	2.901654e-22	38	38	False
6	MBD3	Mann-Whitney U rank test	8.093069e-13	38	38	False
2	DNMT3B	Mann-Whitney U rank test	6.767822e-12	38	38	False
3	DNMT3L	Mann-Whitney U rank test	1.235647e-07	38	38	False
10	ZBTB4	Mann-Whitney U rank test	2.611436e-06	38	38	False
9	UHRF2	Mann-Whitney U rank test	3.467971e-04	38	38	False
0	DNMT1	Mann-Whitney U rank test	4.522748e-03	38	38	False
14	TET2	Mann-Whitney U rank test	5.177072e-03	38	38	False
16	NLRC5	Mann-Whitney U rank test	8.558218e-02	38	38	True
8	UHRF1	Mann-Whitney U rank test	3.765288e-01	38	38	True
12	TET1	T-test	3.846154e-01	38	38	True
7	MBD4	T-test	5.064935e-01	38	38	True
5	MBD2	Mann-Whitney U rank test	5.318039e-01	38	38	True
15	TET3	T-test	8.061938e-01	38	38	True
13	BEND3	Mann-Whitney U rank test	8.162447e-01	38	38	True
1	DNMT3A	T-test	8.581419e-01	38	38	True
11	ZBTB38	T-test	9.370629e-01	38	38	True

Figure 46: T-test results on methylation set of genes

Table on figure 46 shows the results of the T-test. The first eight genes: **MBD1, MBD3, DNMT3B, DNMT3L, ZBTB4, UHRF2, DNMT1, TET2** fail to reject the null hypothesis, based on their p-values, which are less than 0.05. These genes have significant differences in values between normal and tumour tissue. These genes are responsible for the regulation of gene expression through DNA methylation and chromatin remodelling. In the rest of genes it can be concluded that there is no significant difference in their values between normal and tumour tissue.

Comparing the T-test results with the correlation and pairplot analysis for unbalanced and balanced datasets it can be said that the balanced dataset gave the same results as T-test has shown on the balanced dataset pairplot graph! Unbalanced pairplot graph missed just one gene - DNMT3B if looking at histograms with bare eyes. Which proves that both analysis (correlation, pairplot and T-test) should be performed together to double check the results and build more accurate conclusions. Balancing the datasets is also important, as shown in the analysis it prevents losing important data.

Section 4.1.4 Hypothesis testing using T-test for immune set of genes

Normality check on immune genes set.

Hypothesis for normality:

Null hypothesis (H0): The sample comes from a normal distribution.

Alternative hypothesis (H1): The sample does not come from a normal distribution.

Interpretation:

P-value:

- If the p-value is greater than 0.05, fail to reject the null hypothesis (H0).
- If the p-value is 0.05 or less, reject the null hypothesis (H0).

Results of normality check on the set of immune genes are shown on figure 47.

	Normal Patients statistic	Normal Patients p-value	Tumor Patients statistic	Tumor Patients p-value	Normal Patients Fail to reject H0	Tumor Patients Fail to reject H0	Both Fail to reject H0
BCL2	0.810348	0.6595	1.559754	0.4312	True	True	True
BTLA	4.543433	0.1004	0.587552	0.7497	True	True	True
CD274	7.666275	0.0315	16.312561	0.0025	False	False	False
CD44	4.413454	0.1065	3.735221	0.1479	True	True	True
CTLA4	36.752924	0.0001	5.444279	0.0702	False	True	False
FOXP3	2.856738	0.2175	15.640558	0.0040	True	False	False
GATA3	1.429477	0.4690	5.031705	0.0837	True	True	True
HAVCR2	2.720749	0.2342	6.417692	0.0488	True	False	False
HMGB1	29.738578	0.0002	5.591425	0.0669	False	True	False
IL13	0.401919	0.8084	6.921026	0.0405	True	False	False
IL2	29.018966	0.0001	34.869239	0.0001	False	False	False
IL2RB	0.754918	0.6758	2.123437	0.3209	True	True	True
IL33	27.402706	0.0002	10.831667	0.0119	False	False	False
ITK	1.546340	0.4448	1.566794	0.4414	True	True	True
LAG3	6.596405	0.0453	0.495135	0.7771	False	True	False
LAT	2.779877	0.2326	0.152866	0.9286	True	True	True
MIF	3.311074	0.1735	3.091682	0.1912	True	True	True
NFKB1	4.532116	0.0977	13.837464	0.0049	True	False	False
PDCD1	4.742481	0.0907	16.046820	0.0025	True	False	False
RORC	5.005421	0.0857	5.725998	0.0611	True	True	True
SPP1	14.413571	0.0037	5.418328	0.0729	False	True	False
STAT1	7.407867	0.0344	10.676276	0.0121	False	False	False
STAT2	72.987486	0.0001	3.764931	0.1448	False	True	False
TBX21	0.147296	0.9310	0.266837	0.8696	True	True	True
TNFSF10	29.655623	0.0002	15.806459	0.0035	False	False	False
WNT1	52.511365	0.0001	0.413931	0.8075	False	True	False

Figure 47: Normality check on the Immune set of genes

On figure 47 it can be seen that BCL2, BTLA, CD44, GATA3, IL2RB, ITK, LAT, MIF, RORC, TBX21 genes fail to reject the null hypothesis, which means their samples follow

normal distribution. On those genes T-test will be run. On the rest of genes, which do not follow normal distribution, Mann-Whitney U test is run.

Equal variances check on immune genes set.

As noted in the previous section, this check is responsible for checking that the variances of both datasets are equal.

Hypothesis:

- Null hypothesis (H0): Sample variances are equal.
- Alternative hypothesis (H1): Sample variances are not equal.

Interpretation:

p-value:

- If p-value is greater than 0.05, fail to reject the null hypothesis (H0).
- If p-value is 0.05 or less, reject the null hypothesis (H0).

The results of equal variance check are shown on figure 48.

	Test Name	statistic	p-value	Fail to reject H0
IL2RB	Bartlett	9.882008	0.001669	False
TBX21	Bartlett	6.302674	0.012056	False
LAT	Bartlett	5.464649	0.019405	False
GATA3	Bartlett	3.759564	0.052506	True
MIF	Bartlett	3.271039	0.070513	True
BCL2	Bartlett	2.209905	0.137127	True
RORC	Bartlett	1.647875	0.199248	True
ITK	Bartlett	0.277114	0.598599	True
BTLA	Bartlett	0.172715	0.67771	True
CD44	Bartlett	0.015806	0.899952	True

Figure 48: Equal variances check on immune genes set

After running Bartlett's test, genes on figure 48 show the genes: GATA3, MIF, BCL2, RORC, ITK, BTLA, CD44 that have equal variances (failed to reject the null hypothesis). On those

genes T-test will be run. The genes with unequal variance were also found in this case: IL2RB, TBX21, LAT - Welch's test will be performed upon them.

T-test check on immune genes set.

Tests for independence of observations, normality and equal variances were successfully run on 7 genes. Upon those genes T-test is performed, on the rest corresponding tests as described before are run. The significance level is chosen to be 0.05. The test is a 2-tailed test.

Hypothesis:

- Null hypothesis (H0): There is no difference in gene expression between normal and tumour tissue(no effect).
- Alternative hypothesis (H1): There is a difference in gene expression (effect present).

T-test interpretation:

- p-value:
    - If p-value is greater than 0.05, fail to reject the null hypothesis (H0). There is no statistically significant difference in gene expression between normal and tumour tissue.
    - If p-value is 0.05 or less, reject the null hypothesis (H0). There is a statistically significant difference in gene expression between normal and tumour tissue.
- t-statistic:
    - If t-statistic is less than 0, the mean of the normal group is less than the mean of the tumour group.
    - If t-statistic is greater than 0, the mean of the normal group is greater than the mean of the tumour group.

The results of the T-test are shown on figure 49.

	gene	Test Name	statistic	p-value	normal genes size	tumor genes size	Fail to reject H0
24	TNFSF10	Mann-Whitney U rank test	1359.000000	6.797180e-14	38	38	False
4	CTLA4	Mann-Whitney U rank test	1162.000000	1.942491e-06	38	38	False
8	HMGB1	Mann-Whitney U rank test	1096.000000	6.741593e-05	38	38	False
17	NFKB1	Mann-Whitney U rank test	1029.000000	1.224868e-03	38	38	False
7	HAVCR2	Mann-Whitney U rank test	998.000000	3.809360e-03	38	38	False
9	IL13	Mann-Whitney U rank test	940.000000	2.324614e-02	38	38	False
15	LAT	Welch's t-test	-2.290634	2.597403e-02	38	38	False
11	IL2RB	Welch's t-test	-2.029909	5.394605e-02	38	38	True
23	TBX21	Welch's t-test	1.366147	1.718282e-01	38	38	True
6	GATA3	T-test	0.803830	4.125874e-01	38	38	True
14	LAG3	Mann-Whitney U rank test	648.000000	4.473216e-01	38	38	True
25	WNT1	Mann-Whitney U rank test	794.000000	4.598067e-01	38	38	True
0	BCL2	T-test	-0.555387	5.774226e-01	38	38	True
16	MIF	T-test	0.467270	6.403596e-01	38	38	True
5	FOXP3	Mann-Whitney U rank test	766.000000	6.531467e-01	38	38	True
13	ITK	T-test	0.375418	7.052947e-01	38	38	True
22	STAT2	Mann-Whitney U rank test	689.000000	7.371004e-01	38	38	True
10	IL2	Mann-Whitney U rank test	754.000000	7.449070e-01	38	38	True
1	BTLA	T-test	0.296612	7.562438e-01	38	38	True
19	RORC	T-test	-0.217038	8.121878e-01	38	38	True
21	STAT1	Mann-Whitney U rank test	702.000000	8.403947e-01	38	38	True
18	PDCD1	Mann-Whitney U rank test	704.000000	8.565791e-01	38	38	True
12	IL33	Mann-Whitney U rank test	730.000000	9.382641e-01	38	38	True
2	CD274	Mann-Whitney U rank test	718.000000	9.711676e-01	38	38	True
20	SPP1	Mann-Whitney U rank test	719.000000	9.794033e-01	38	38	True
3	CD44	T-test	-0.019284	9.870130e-01	38	38	True

Figure 49: T-test results on immune set of genes

From the results on figure 49, it can be seen that the first seven genes: **TNFSF10, CTLA4, HMGB1, NFKB1, HAVCR2, IL13, LAT** fail to reject the null hypothesis, based on their p-values, which are less than 0.05. These genes, responsible for immune system functionality, including such processes as regulation, signalling, and inflammatory responses have significant differences in values between normal and tumour tissue.

To compare the t-test result with the table from the correlation and pairplot analysis of immune genes from section 4.1.1 from this report following conclusions can be made. Gene TNFSF10 and CTLA4 showed high correlation values in the correlation plot and difference in the average values on the pairplot in both unbalanced and balanced datasets. Genes LAT and IL13 stood out in correlation and pairplot for the balanced dataset. And gene HAVCR2 stood out in an unbalanced dataset in correlation plot and pairplot. To summarise these findings more genes that eventually showed up as significant in the T-test analysis were found when the datasets were balanced, which should be a good practice for further analysis for the current topic. 2 significant genes HMGB1 and NFKB1 stood out only after

performing the T-test. Which means for an investigation for significant genes and full coverage of the analysis, both correlation plot, pairplot and T-test should be performed, ideally on balanced datasets.

Full summary of T-test analysis is shown in Table 6.

Table 6: Full summary of T-test for significant genes

T-test results for significant genes	
Immune genes set	Methylation genes set
TNFSF10	MBD1
CTLA4	MBD3
HMGB1	DNMT3B
NFKB1	DNMT3L
HAVCR2	ZBTB4
IL13	UHRF2
LAT	DNMT1
	TET2

Section 4.2 Overview of analysis 2

Section 4.2.1 PCA findings

PCA plots based on definition (metastatic tumour, primary solid tumours, recurrent solid tumour, solid tissue normal) show a clear separation of the normal tissue samples (blue dots) from the cancer samples (black, red and green dots). The cancer samples overlap but there is a degree of separation between the different tumour types.

In contrast, the distinct clustering of normal tissues suggests that methylation profiles are effective in distinguishing between normal and cancerous tissues. The overlap between different tumour types suggests the presence of common epigenetic alterations in cancer and also indicates specific changes associated with tumour progression and metastasis.

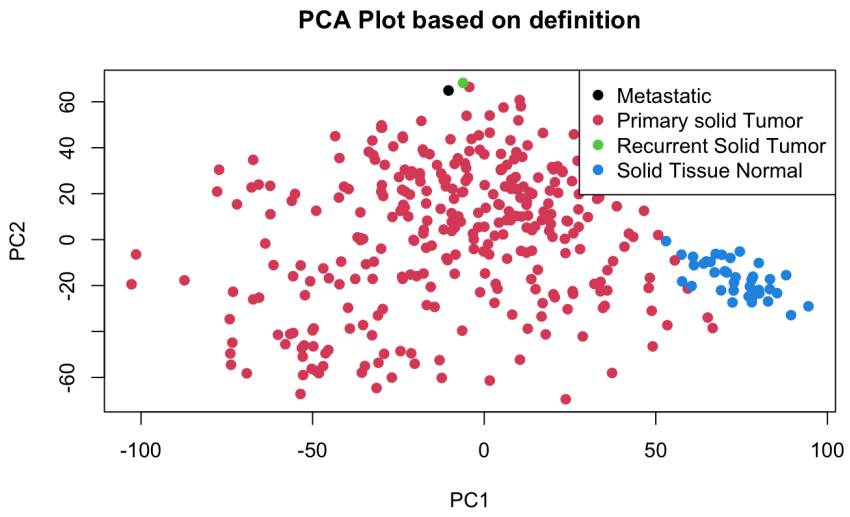


Figure 50: PCA plot based on definition

Secondly, the analysis of the largest contributing genes to the first and second principal components (PC1 and PC2) deepened understanding of the methylation profiles associated with the different tissue types analysed in the PCA plots. Among these genes, DNAJC6, NTNG1, KCNK12 and RSPO3 were significant contributors to PC1. These genes are known for their roles in neurodevelopment, ion channel regulation and the Wnt signalling pathway, which are critical in both normal physiology and cancer progression (Casalino and Verde, 2020; Johnson and Lal, 2016).

```
> print(top_genes_PC1_names)
          CpG_id                        gene
21827  cg26615127                      DNAJC6
25452  cg07155336                       NTNG1
55096  cg13913015                      KCNK12
65118  cg22830113 RP11-521016.1;RP11-521016.2
118793 cg09111223            GRID2;RP11-9B6.1
124385 cg12861945              FAM218A;TRIM61
180322 cg19365062                       RSPO3
209459 cg25167643                      PTPRZ1
261512 cg03052869                       GRID1
278392 cg15344220                         TUB
281502 cg11855526       MPPED2;RP5-1024C24.1
286137 cg25303599                 FADS1;FADS2
330563 cg26224671                     TNFSF11
381534 cg06728579          GNAO1;RP11-46107.1
436559 cg15424739                        COMP
```

Figure 51: Top gene of PC1

```
> print(top_genes_PC2_names)
          CpG_id                       gene
16479  cg17449954                       HEYL
36271  cg05719164                       LHX4
100611 cg10752315                       <NA>
108284 cg05006473                       <NA>
199796 cg09921682                       ZPBP
213176 cg01274524                 AC073055.2
256215 cg26718433                     CXCL12
257968 cg11914795                   C10orf107
306600 cg06951626                       ABCC9
309918 cg15336765      AQP5;RP11-469H8.6
311665 cg20643952 HOXC6;RP11-834C11.12
460963 cg07246713                      OPRL1
468629 cg21634064      RP3-412A9.16;SMTN
468631 cg09745440                       SMTN
468632 cg20934596                       SMTN
```

Figure 52: Top gene of PC2

Except for some genes that change a lot because of the cancer itself, the genes (Figure 51) **NTNG1**, involved in neural development, and **RSPO3**, a regulator of the Wnt signalling pathway, are of particular interest. These findings suggest that changes in the methylation status of these genes may be a marker of the transition from normal to tumour tissue status. Furthermore, RSPO3 is associated with the process of tumourigenesis and has the potential to be a biomarker of cancer progression (Guo *et al.*, 2017).

The genes **HEYL**, **LHX4** and **CXCL12** are more prominent in Figure 52. HEYL is part of the Notch signalling pathway and CXCL12 is a chemokine involved in the migration and invasion of cancer cells, which further emphasises the underlying epigenetic mechanisms

driving tumour progression and metastasis (Yuan *et al.*, 2019). Therefore, in the PCA plot, Metastatic and Recurrent Solid Tumours are positioned as off top in PC2.

The gender-based (male and female) PCA plot shows that there is no clear distinction between male samples (red dots) and female samples (black dots). The data points are interspersed with each other, indicating that gender has no significant effect on the overall methylation profile. For the methylation sites analysed, there was no significant difference by gender. This implies that other factors, such as tissue type or gene mutations, may have a more significant impact on methylation patterns.

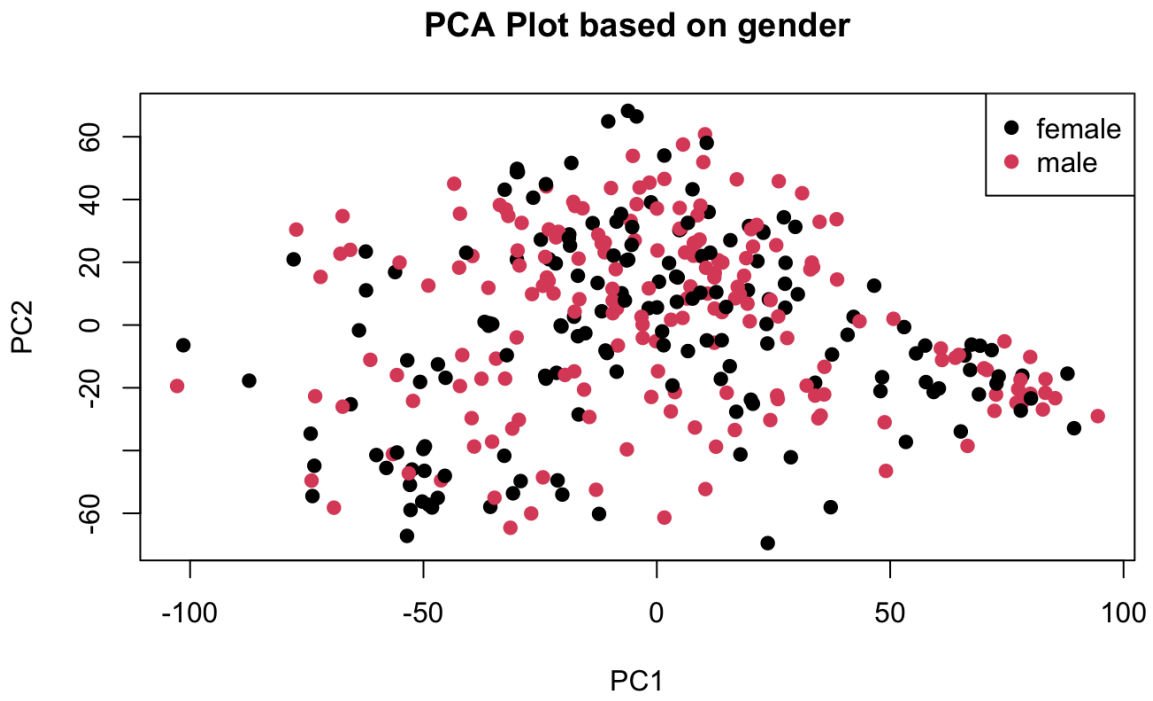


Figure 53: PCA plot based on gender

PCA plots based on race (Asian, Black or African American, White) show extensive overlap between racial groups. While there are subtle differences, there is no significant clustering between the groups. The extensive overlap suggests that racial differences, like gender differences, do not substantially affect the methylation profiles in this dataset. However, there will be small differences in racial differences that can be further explored in the future with more specific analyses or larger sample sizes.

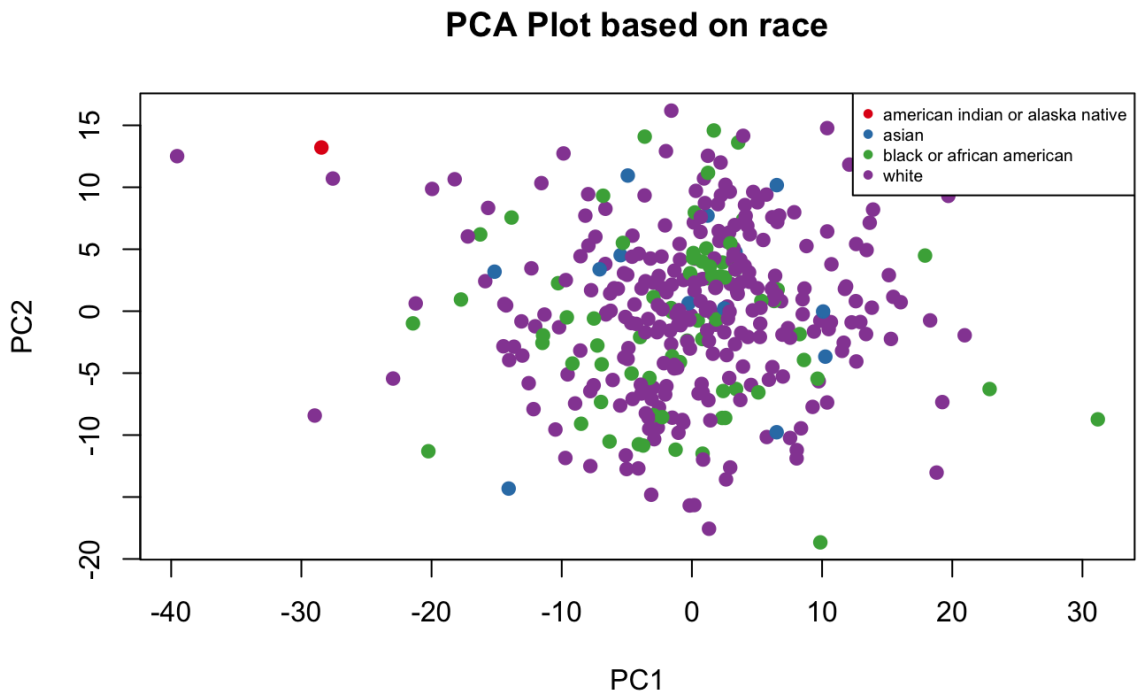


Figure 54: PCA plot based on race

Limitations of this study include the inability of PCA to definitively categorise samples based on gender or race. This suggests that in the case of colon adenocarcinoma, the methylation profiles may not be significantly different, or the ability to detect such differences may be lacking due to the dataset. Future studies could better understand subtle variations in methylation across populations by exploring larger, more diverse datasets. The distinction between sex and race in the PCA analysis of this paper was not sufficient to significantly differentiate methylation profiles, suggesting that the study may not be able to fully assess or identify subtle differences that may exist in methylation patterns influenced by gender or race. This limitation suggests that more refined methods or larger datasets will be needed in the future to reveal more distinct methylation patterns associated with different races.

Section 4.2.2 Volcano plot of differential methylation analysis

The differential methylation analysis between normal and cancerous tissues revealed significant epigenetic alterations in several key genes, notably **DNMT3A**, **DNMT3L**, and

UHRF1. These genes play crucial roles in DNA methylation and gene regulation, processes that are often disrupted in cancer.

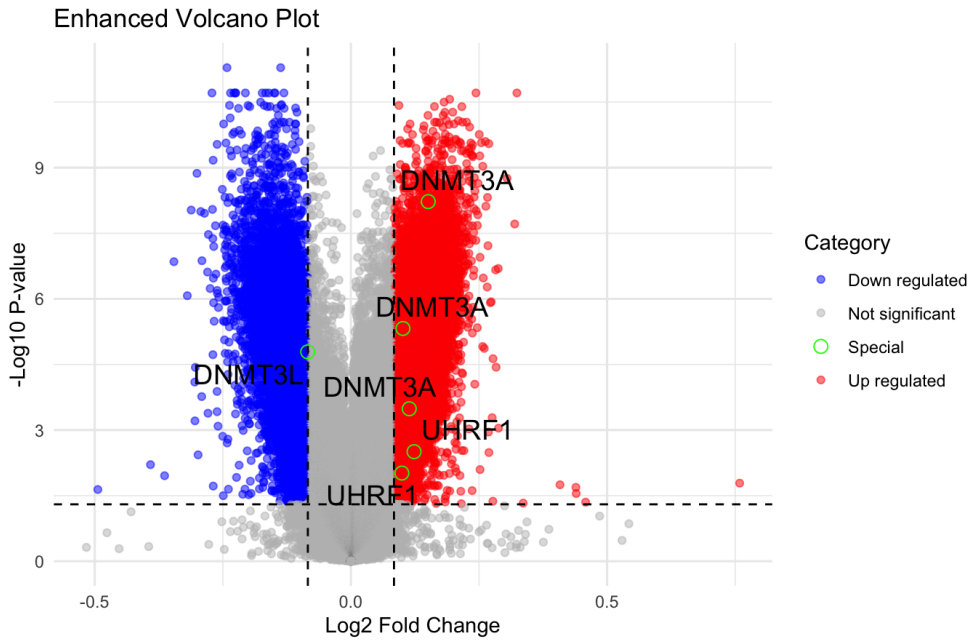


Figure 55: Volcano plot of key genes

1. **DNMT3A (DNA Methyltransferase 3 Alpha)**:
    - **Role**: DNMT3A is instrumental in DNA methylation, a process critical for gene regulation and cellular differentiation. It catalyses the transfer of methyl groups to DNA, impacting gene expression and maintaining genomic stability.
    - **Findings**: The analysis indicated that DNMT3A is significantly upregulated in cancer tissues compared to normal tissues. This upregulation is evident from the increased methylation levels, as highlighted in the volcano plot. The increased activity of DNMT3A in cancer tissues suggests its involvement in promoting tumourigenesis by altering the expression of genes that control cell proliferation and survival (Moore, Le and Fan, 2012).

2. **DNMT3L (DNA Methyltransferase 3 Like)**:
    - **Role**: Although DNMT3L lacks catalytic activity, it functions as a regulatory factor that enhances the activity of other DNA methyltransferases like

DNMT3A and DNMT3B. DNMT3L is crucial in establishing maternal genomic imprints and epigenetically regulating gene expression.

- **Findings**: The analysis found that DNMT3L is downregulated in cancer tissues, indicating reduced levels of methylation compared to normal tissues. This reduction in DNMT3L activity might lead to dysregulation of methylation patterns, contributing to abnormal gene expression and cancer progression (Moore, Le and Fan, 2012).

3. **UHRF1 (Ubiquitin-like with PHD and Ring Finger Domains 1)**:
    - **Role**: UHRF1 is essential for maintaining DNA methylation and is involved in the regulation of the cell cycle and apoptosis. It often shows overexpression in cancer cells and regulates tumour suppressor genes through methylation.
    - **Findings**: UHRF1 showed significant changes in its methylation status, being upregulated in cancer tissues. This upregulation suggests that UHRF1 may play a role in silencing tumour suppressor genes, thereby promoting the proliferation and survival of cancer cells (Moore, Le and Fan, 2012).

The visualisation through the volcano plot underscore the significant epigenetic changes occurring in colon cancer. Genes such as DNMT3A and UHRF1, with their altered methylation patterns, highlight the crucial role of epigenetic regulation in cancer development. The findings suggest that targeting these epigenetic modifications could provide novel therapeutic strategies for treating colon cancer, focusing on the reason for restoring normal methylation patterns and gene expression profiles. The different methylation patterns observed provide different avenues for the development of targeted epigenetic therapies. For example, DNMT inhibitors, which are already used to treat certain types of haematological malignancies, could be tailored to target specific methylation changes in colorectal cancer.

Section 4.2.3 Survival analysis findings

The survival analysis revealed distinct survival patterns across different groups:

- **By Cancer Stage**: In Figure 56-57, patients with Stage I colon cancer exhibited the highest survival probability, progressively decreasing with more advanced stages. This trend was statistically significant, underscoring the prognostic importance of cancer staging (Therneau, 2015).

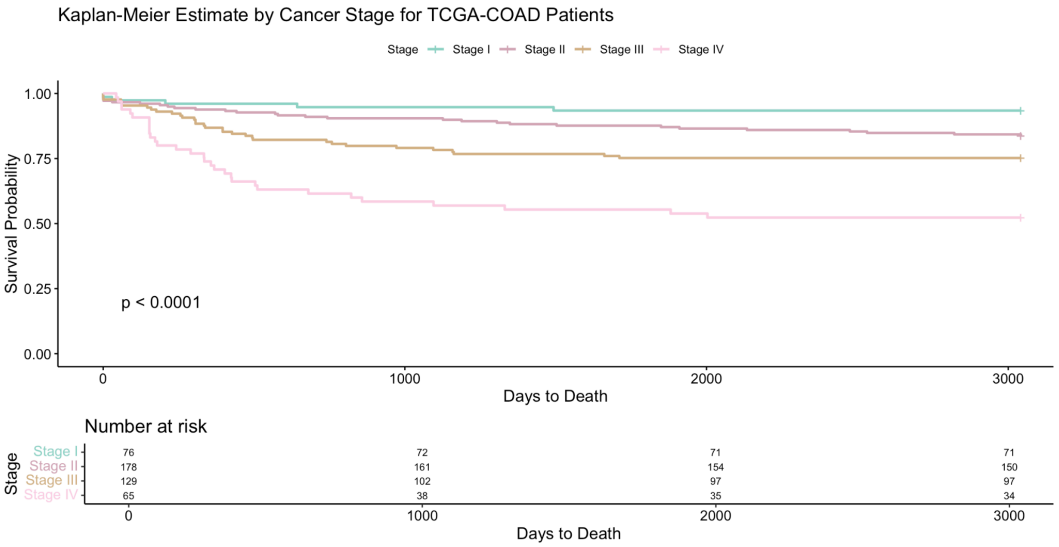


Figure 56: Kaplan-Meier estimate by cancer stage

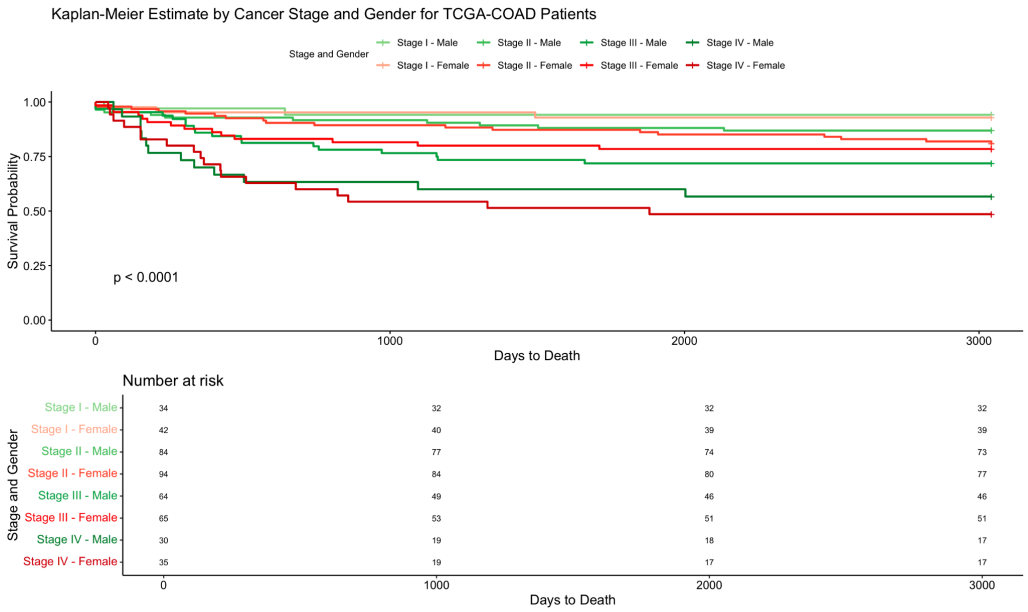


Figure 57: Kaplan-Meier estimate by cancer stage and gender

- **By Gender**: The survival curves for male and female tissueoverlapped significantly, suggesting that gender does not have a statistically significant impact on survival in colon cancer (Figure 58).

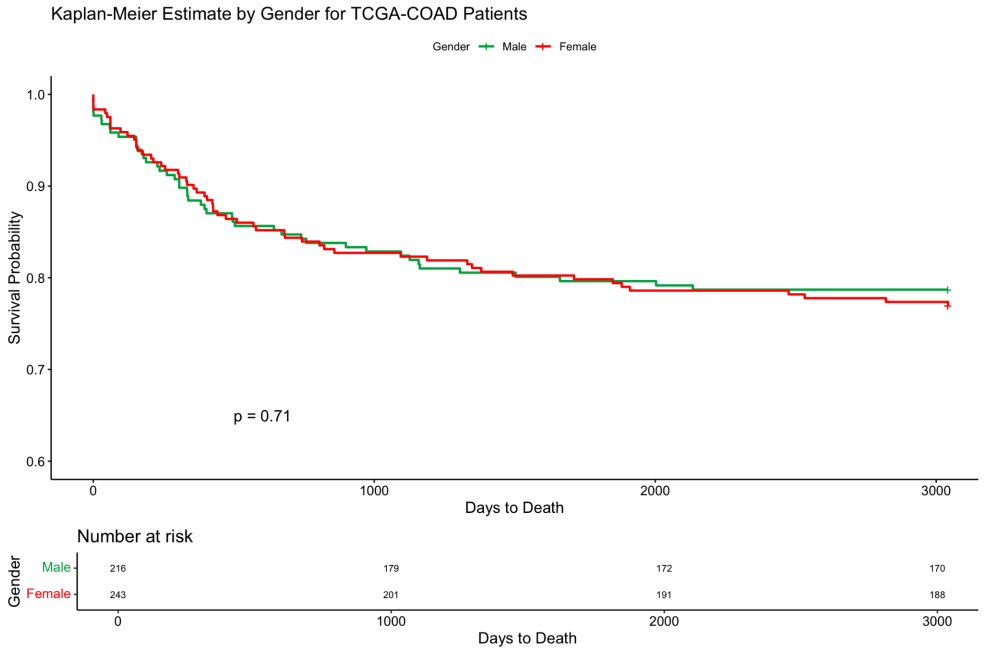


Figure 58: Kaplan-Meier estimate by gender

- **By Race**: Survival probabilities varied slightly among different racial groups, but these differences were not statistically significant, indicating that race alone is not a strong predictor of survival outcome in colon cancer.

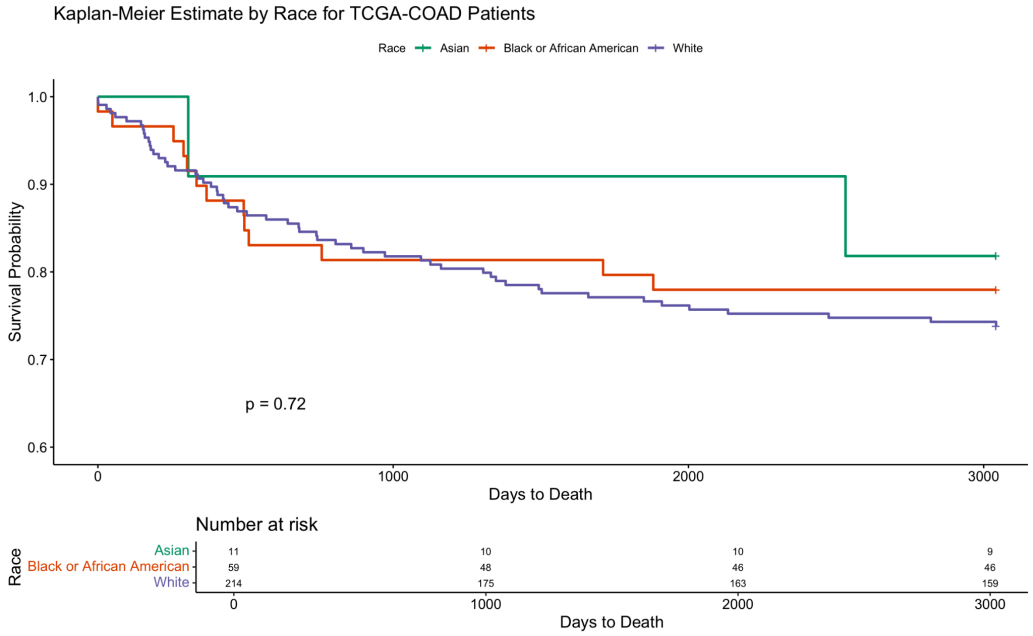


Figure 59: Kaplan-Meier estimate by race

Kaplan-Meier curves provide a clear visualisation of the different survival probabilities, especially when analysed by cancer stage. The separation of the curves indicates how the prognosis worsens for advanced cancers. However, when analysed by gender or ethnicity, the curves are very close together, suggesting that these factors alone have a negligible effect.

Survival analyses conducted in this study highlight the critical role of clinical stage in determining the prognosis of tissue with colorectal adenocarcinoma (COAD) Tissue with stage I colon cancer have the highest probability of survival, which decreases with more advanced stage. This apparent gradient in survival probability emphasises the aggressive nature of advanced cancers, which are often accompanied by more disease, metastasis and resistance to conventional therapies (Therneau, 2015).

Contrary to expectations, the analyses showed no statistically significant impact of gender and different ethnic groups on the survival rates of tumour patients, which were only slightly different, but these differences were not statistically significant. This suggests that biological and molecular factors may have a more significant impact on colon cancer prognosis than demographic variables (Smith *et al.*, 2018). This finding is consistent with other studies that have shown the limited prognostic value of demographic factors compared to genetic or molecular markers (Altman *et al.*, 2012).

As gender and ethnicity have a negligible impact on survival outcomes, this highlights the potential for further research into how underlying genetic, epigenetic and molecular traits interact with demographic factors to influence disease progression and survival. For example, combining genomic and epigenomic data with clinical outcomes could drive the development of predictive models that better account for individual differences in prognosis and treatment response (Hassan *et al.*, 2022). The tailored treatments will be more effective, especially in the early stages of developing cancer.

In conclusion, the most significant genes in PCA and volcano plot are:

Table 7: Summary of significant genes in analysis 2

Significant Genes in PCA and Volcano Plot		
DNA Methylation Regulators	**Signal Transduction and Cell Communication**	**Transcription Factors and Signalling Molecules**
DNMT3A	RSPO3	HEYL
DNMT3L	NTNG1	CXCL12
UHRF1		

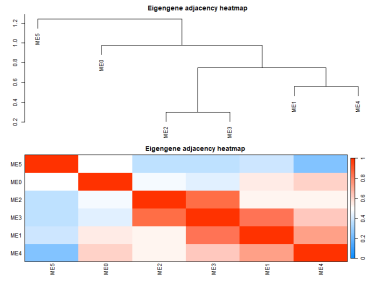
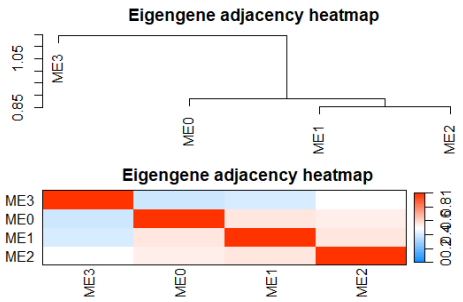
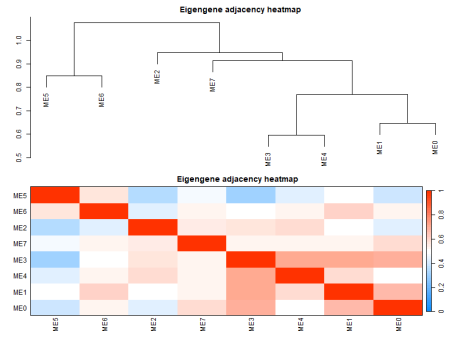
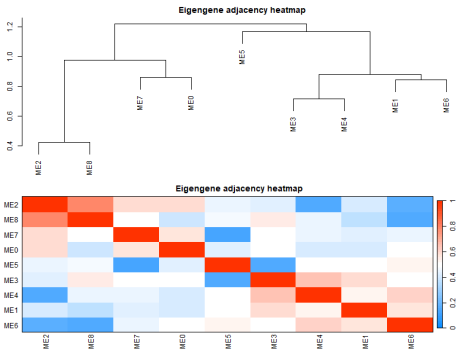
These genes collectively influence a wide range of cellular functions from epigenetic regulation to signal transduction and cellular communication. In colon cancer, dysregulation of these genes most likely contributes to all aspects of cancer biology, including tumour growth, resistance to cell death, metastasis and the tumour microenvironment. Understanding the pathways and mechanisms of these genes could provide therapeutic targets and diagnostic biomarkers for more effective treatment of colon cancer.

Section 4.3 Overview of analysis 3

Weighted gene co-expression network analysis - Morphology.

Section 4.3.1 Analysis of eigengene module heatmap

Table 8: Analysis of eigengene model heatmap

	Tumour Gene Selection	Normal Gene Selection
Tumour Data		
Normal Data		

Eigengenes:

An eigengene, calculated as the first principal component of a cluster, is a measurement of the degree of expression in a gene cluster. It generalises the behaviour of the modules into a single value.

Heatmap Explanation:

- Heatmaps graph the interactions between the eigengene values of our modules.
- Rows and columns map to different modules

- Colours represent the strength of eigengene relationship, red is a strong positive correlation, blue is a strong negative correlation, and white is little correlation.
- Dendrogram represents hierarchical clustering of gene modules, highly interactive modules are nested closer together.

Section 4.3.1.1 Explanation of results:

Top Left (Tumour Data with Tumour Gene Selection):

Represents the relationships between eigengenes of gene modules taken from tumour data, using the 5,000 top genes by variance in the tumour dataset.

This serves as the baseline for colon cancer tissue physiological processes and their module relationships.

Top Right (Tumour Data with Normal Gene Selection):

Represents the relationships between eigengenes of gene modules taken from tumour data, using the top 5,000 genes by variance in the normal dataset.

Differences from the top left heatmap highlight how tumour data organised differently when normal variability genes are considered.

Bottom Left (Normal Data with Tumour Gene Selection):

Represents the relationships between eigengenes of modules taken from normal data, using the top 5,000 genes by variance in the tumour dataset.

Differences from the bottom right heatmap highlight how normal data organised differently when tumour variability genes are considered.

Bottom Right (Normal Data with Normal Gene Selection):

This shows the relationships between eigengenes of modules derived from normal data, using the top 5,000 most variable genes in the normal dataset.

This serves as the baseline for normal physiological processes and their module relationships.

Section 4.3.1.2 Analysis of results

Tumour vs. Normal:

Comparing tumour data with normal data can reveal which modules are standard human biology and which are cancer-associated modules

Gene Selection Impact:

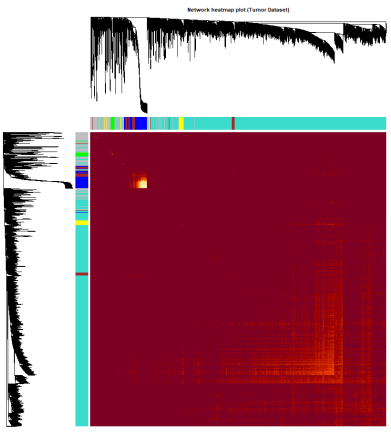
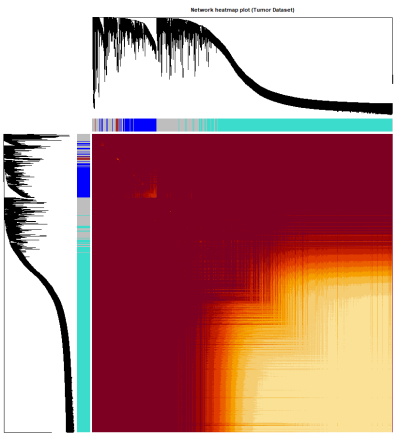
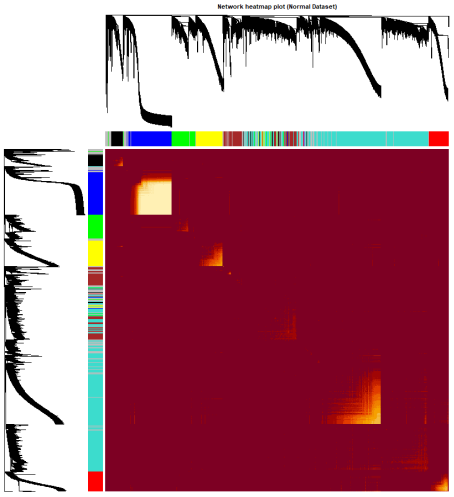
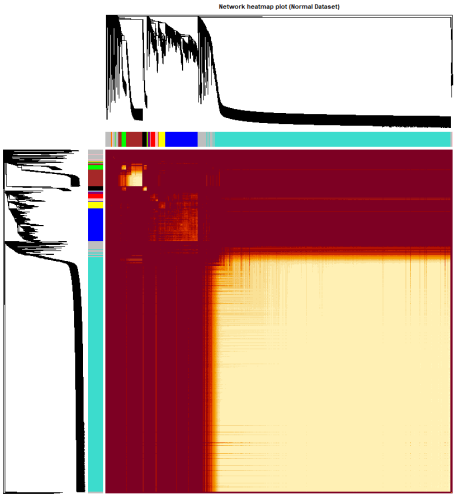
By mapping the top 5,000 genes by variance from each dataset onto each dataset via enrichment analysis, we can identify how both sets of top-variance genes interact in their networks in both a normal and tumour state. Modules derived from tumour-variable genes might highlight pathways critical in cancer, while those from normal-variable genes might highlight general physiological processes.

Correlations:

Strong Red correlations indicate the modules are co-regulated or functionally related. Differences in correlation patterns indicate the disrupted regulatory pathways and module function in cancer tissue

Section 4.3.2 Analysis of WGCNA module heatmap

Table 9: Analysis of WGCNA module heatmap

	Tumour **Gene Selection**	**Normal Solid Gene Selection**
Tumour Data		
Normal Solid Tissue Data		

Section 4.3.2.2 Explanation - WGCNA Module Heatmap

Top Left (Tumour Data with Methylation Gene Selection)

This heatmap shows the TOM for the tumour dataset using the top 5000 genes by variance in the tumour samples.

The blocks of similar colour indicate modules (clusters) of highly interconnected genes. A denser red colour suggests a higher degree of connectivity among the genes within that module.

Top Right (Tumour Data with Normal Gene Selection)

This heatmap shows the TOM for the tumour dataset using the top 5000 genes by variance in the normal samples.

Comparing this to the top left heatmap, different patterns of connectivity, reflecting how the gene selection based on normal sample variance influences the network structure in tumour samples. Differences indicate how gene connectivity patterns vary between gene sets.

Bottom Left (Normal Data with Methylation Gene Selection)

This heatmap shows the TOM for the normal dataset using the top 5000 genes by variance in the tumour samples.

This allows for comparison of connectivity patterns between normal and tumour datasets using the same gene set. Differences from the tumour dataset heatmaps can indicate how gene connectivity is disrupted or altered in cancer.

Bottom Right (Normal Data with Normal Gene Selection)

This heatmap shows the TOM for the normal dataset using the top 5000 genes by variance in the normal samples.

This represents the baseline connectivity in normal samples using genes most variable in normal conditions. Comparing this with the bottom left heatmap shows how gene selection criteria affect the network structure in normal samples.

Section 4.3.2.2 Differences and Indications:

- **Gene Pathway Trends**:
    - Differences in module density and structure (visualised as coloured blocks) can tell how modules are influenced by the status(tumour vs. normal) and the selection of genes (tumour variance vs. normal variance).

- **Disrupted Connectivity in Tumour Samples**:
    - Comparing the top left and bottom left heatmaps can reveal disrupted connectivity patterns in tumour samples. Modules present in normal samples but not in tumour samples (or vice versa) can indicate key pathways affected by the tumour.
- **Impact of Gene Selection**:
    - Comparing the top left and top right, as well as bottom left and bottom right, shows the impact of gene selection criteria on the network structure. This can highlight how genes are differently interconnected in certain conditions.

Section 4.3.3 Results for each WGCNA + Identified Gene Modules

Gene Modules: Below are the Heatmaps of the WGCNA and Eigengene clusters. Included are the individual gene modules identified in each network analysis, and represent functionally associated or highly connective genes.

Section 4.3.3.1. WGCNA tumour tissue methylation data via top 5000 genes by variance from tumour tissue methylation data

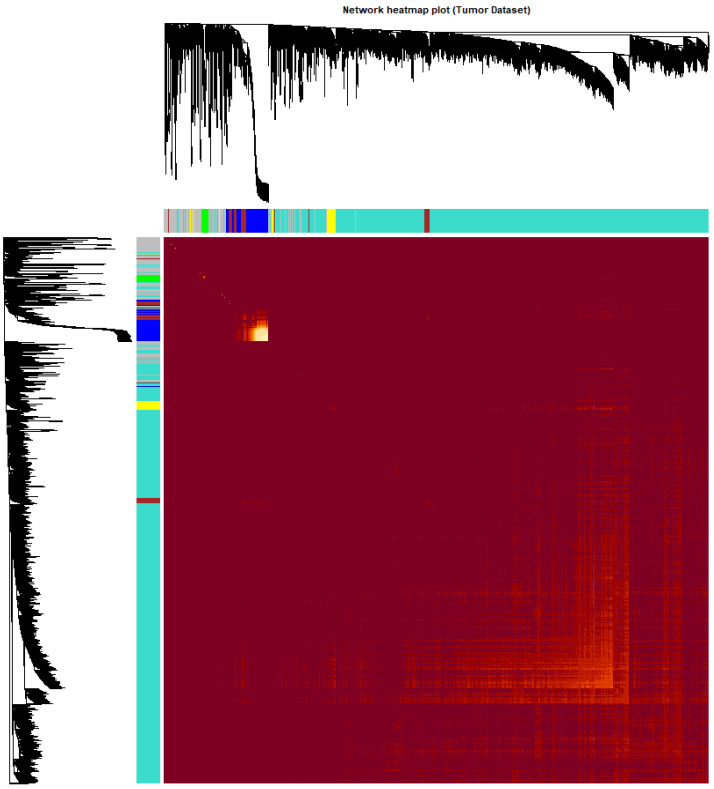


Figure 60: Network heatmap plot (Tumour dataset)

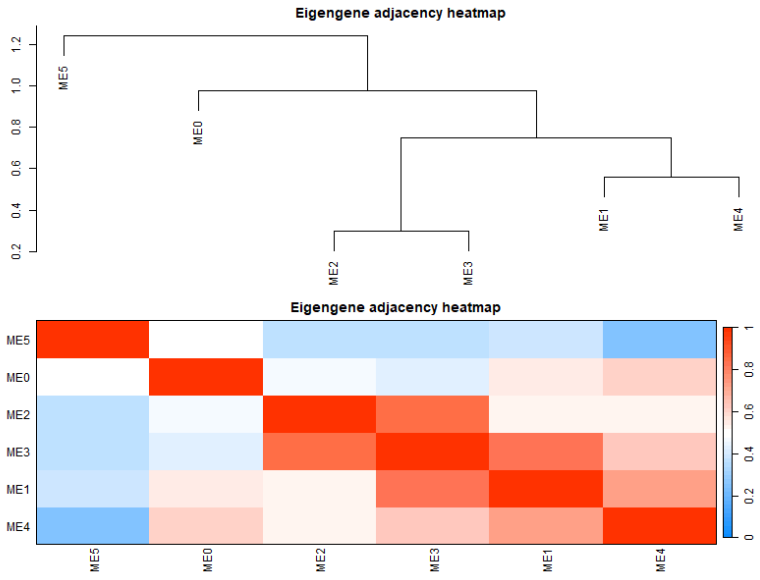
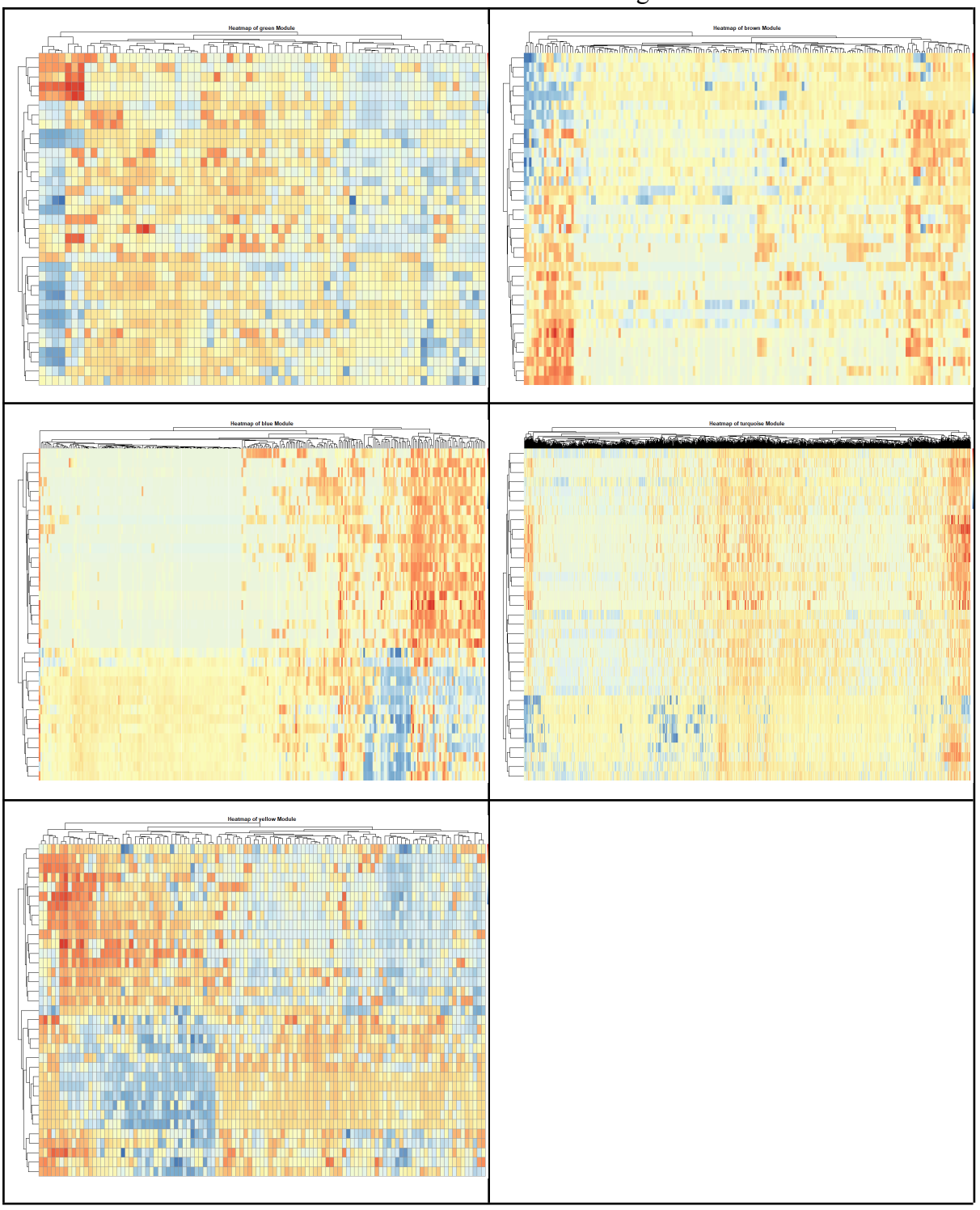


Figure 61: Eigengene adjacency heatmap (Tumour)

Table 10: Gene modules tumour + tumour gene selection



Section 4.3.3.2. WGCNA normal tissue methylation data via Top 5000 genes by variance from normal tissue methylation data

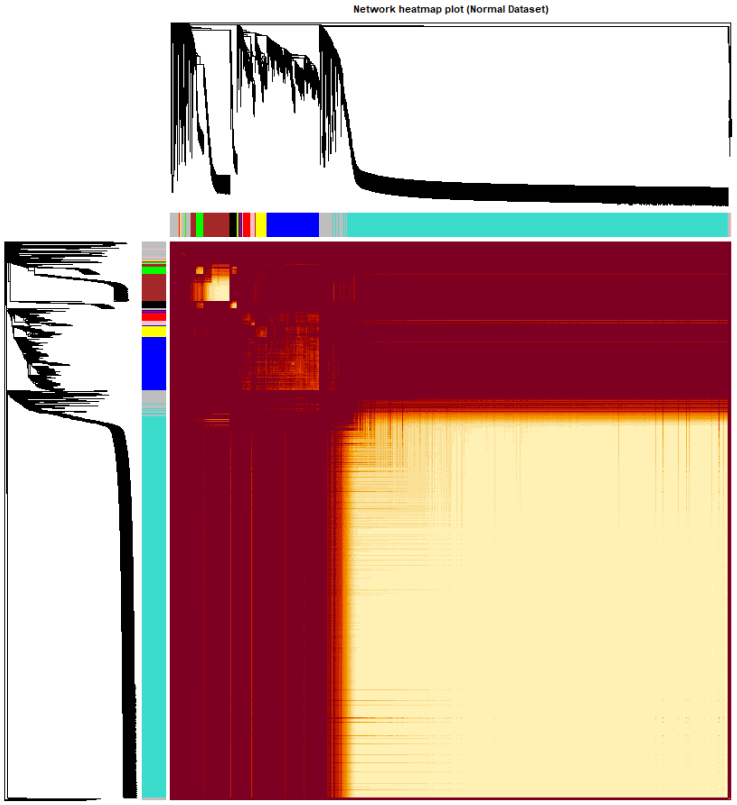


Figure 62: Network heatmap plot (Normal dataset)

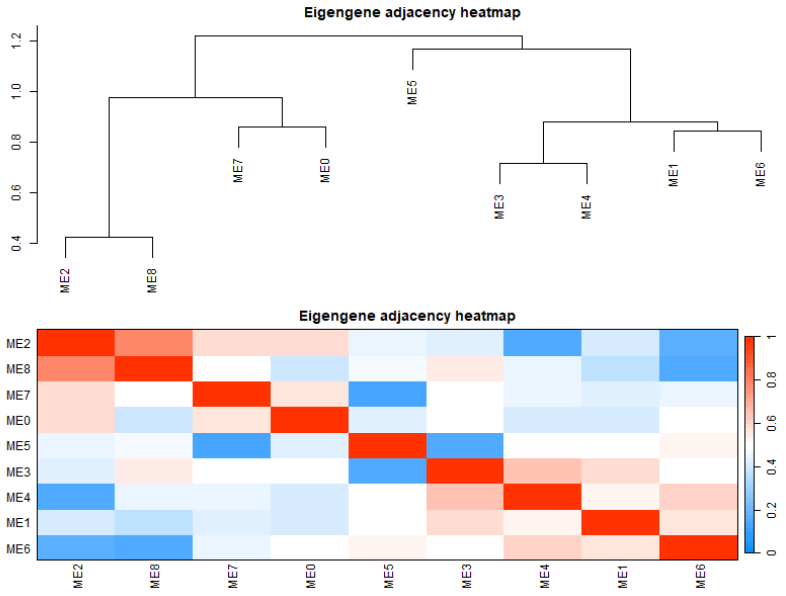
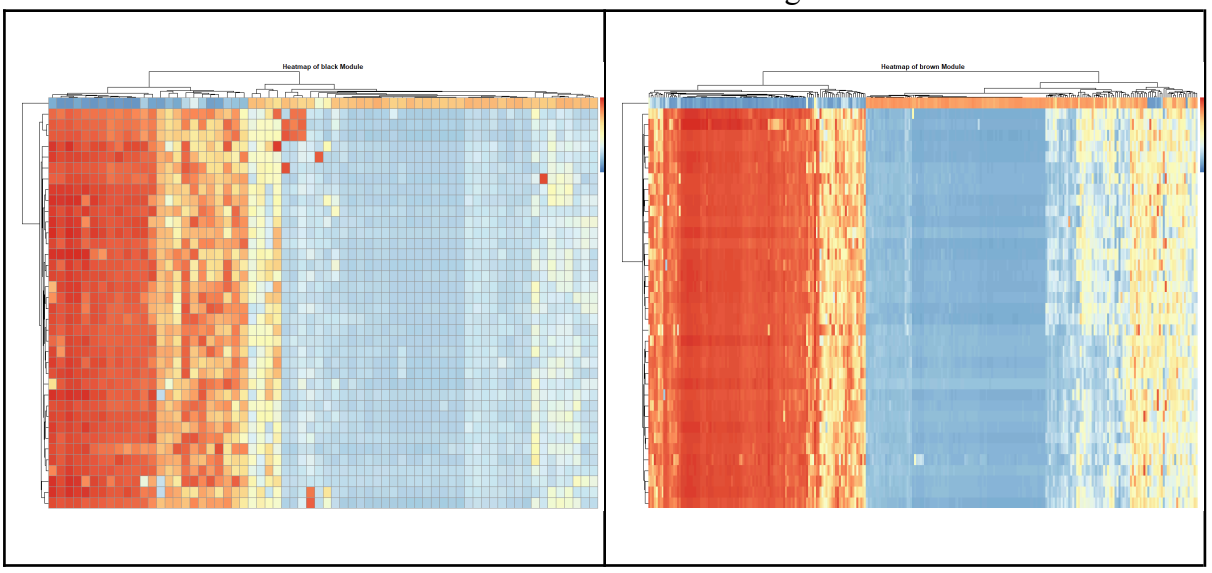


Figure 63: Eigengene adjacency heatmap (Normal)

Modules Identified:

Table 11: Gene modules normal + normal gene selection



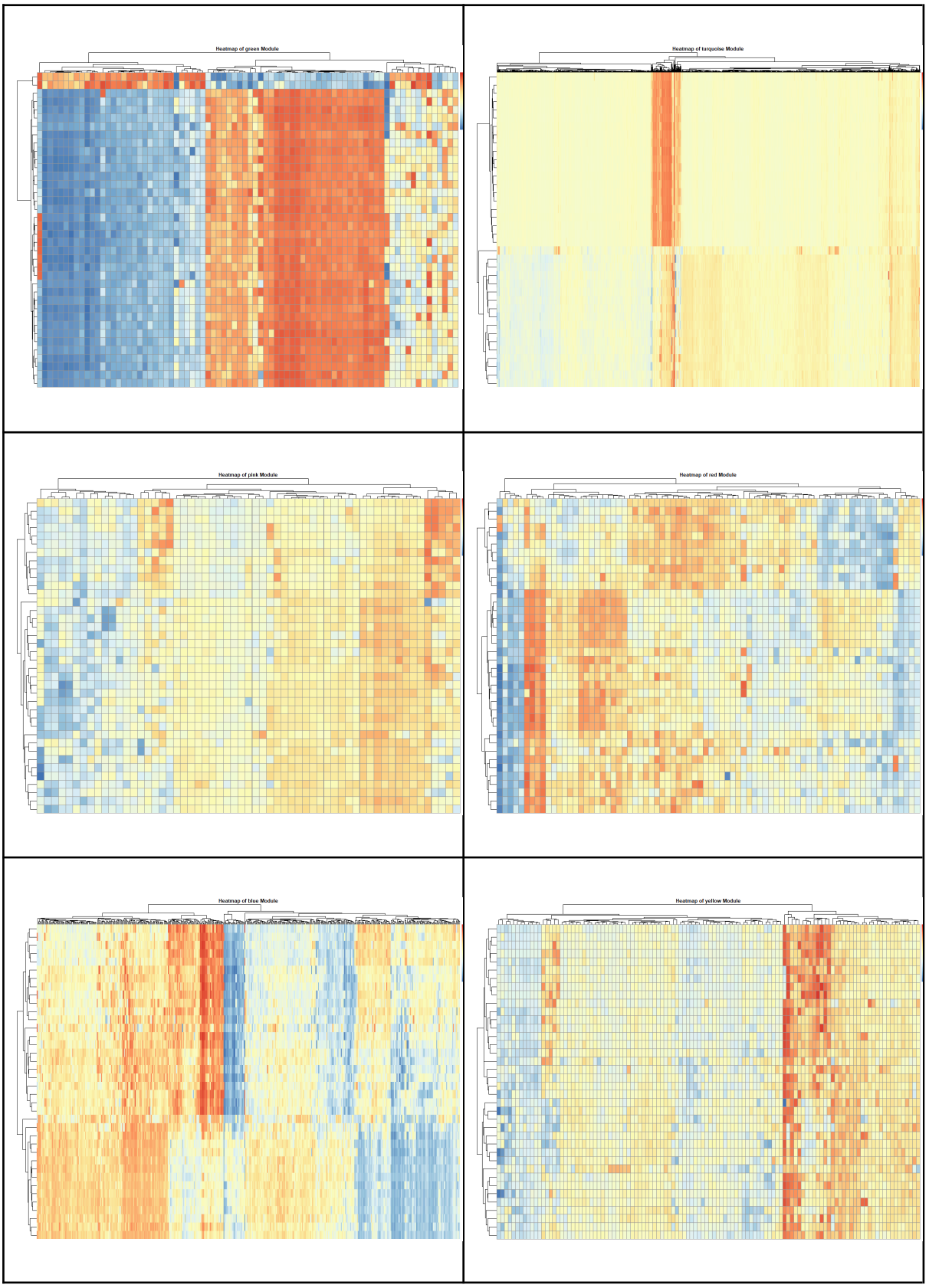

Heatmap of green Module
Heatmap of turquoise Module
Heatmap of pink Module
Heatmap of red Module
Heatmap of blue Module
Heatmap of yellow Module

Figure 64: Scale independence and mean connectivity

Section 4.3.3.3. WGCNA tumour tissue methylation data via Top 5000 genes by variance from normal tissue methylation data

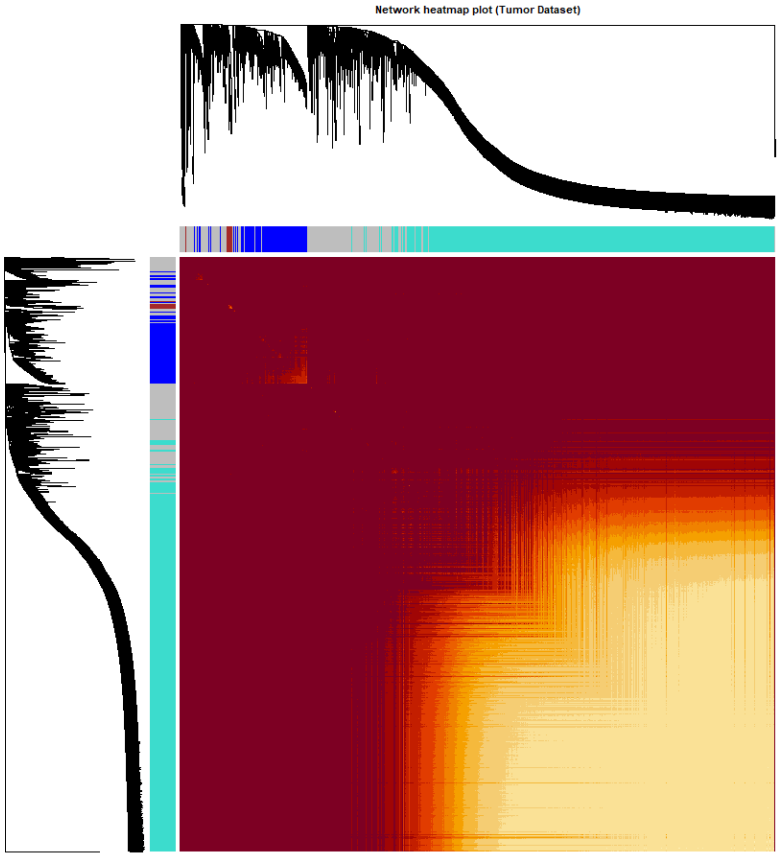


Figure 65: Network heatmap plot (Tumour)

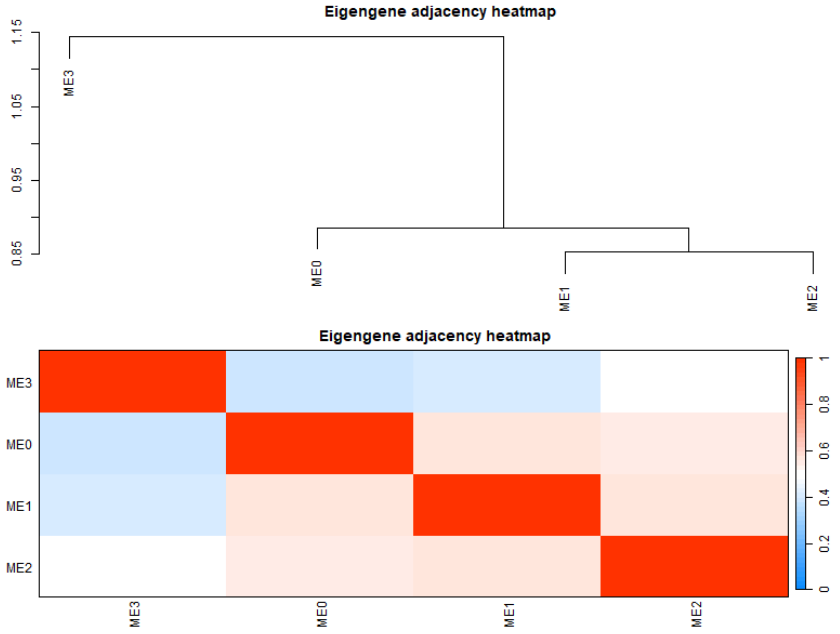
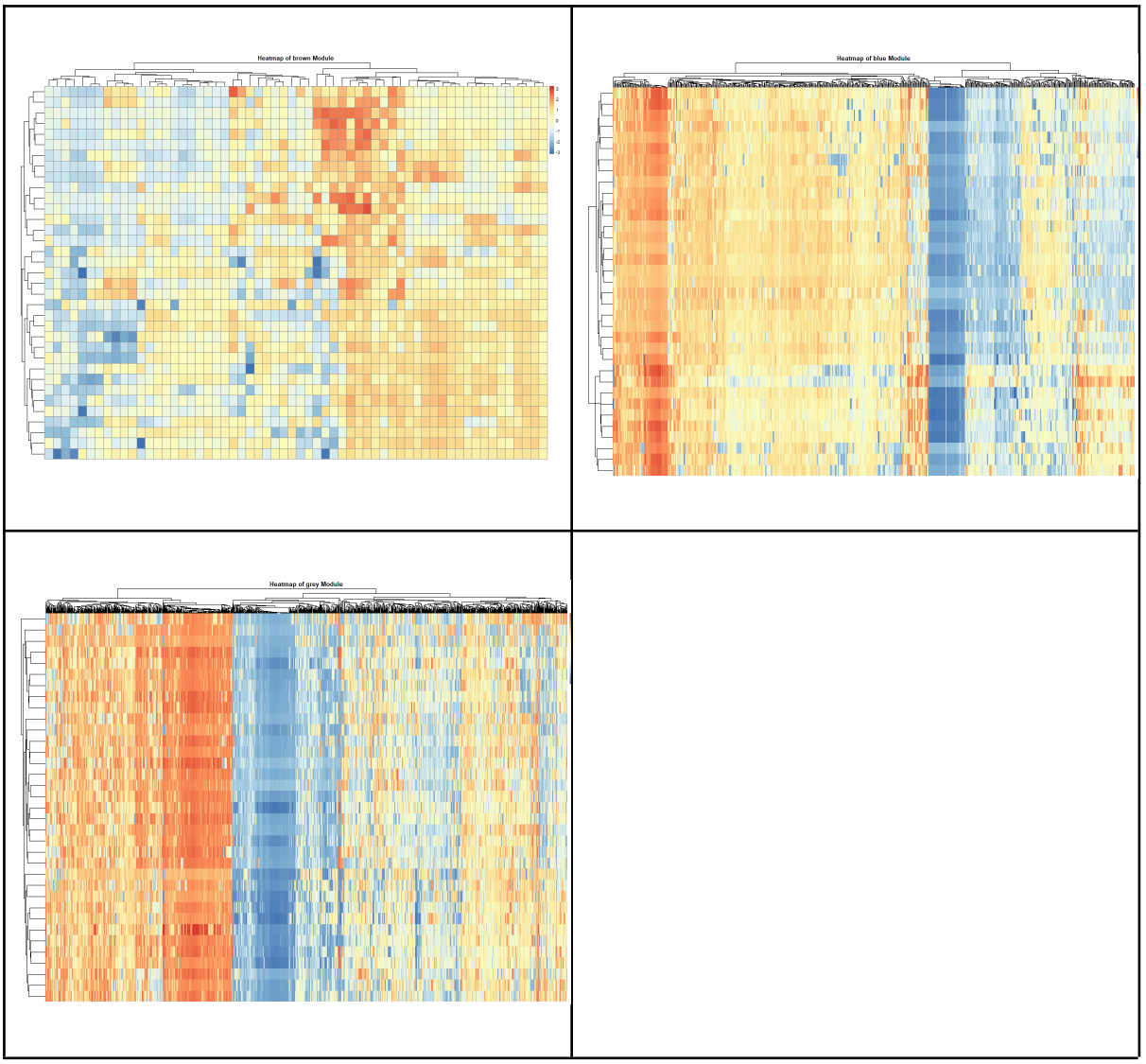


Figure 66: Eigengene adjacency heatmap (Tumour)

Modules Identified:

Table 12: Gene modules tumour + normal gene selection



Section 4.3.3.4. WGCNA normal solid tissue methylation data via top 5000 genes by variance from tumour patient methylation data

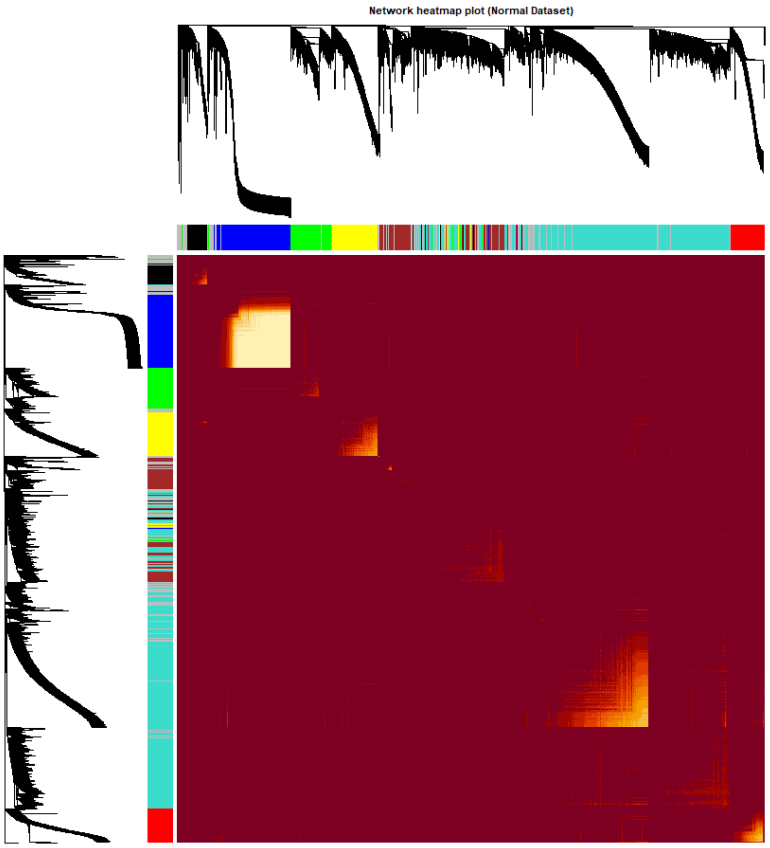


Figure 67: Network heatmap plot (Normal)

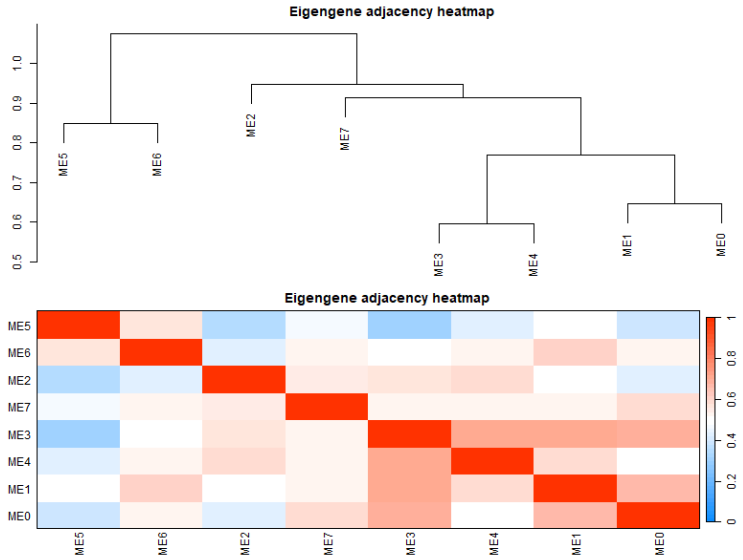
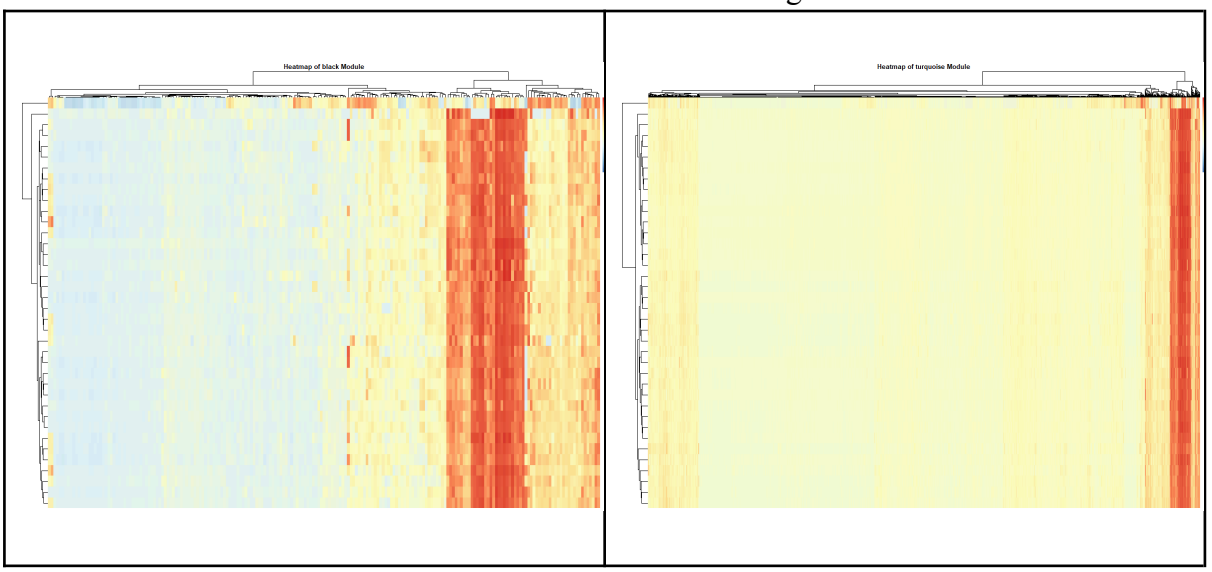
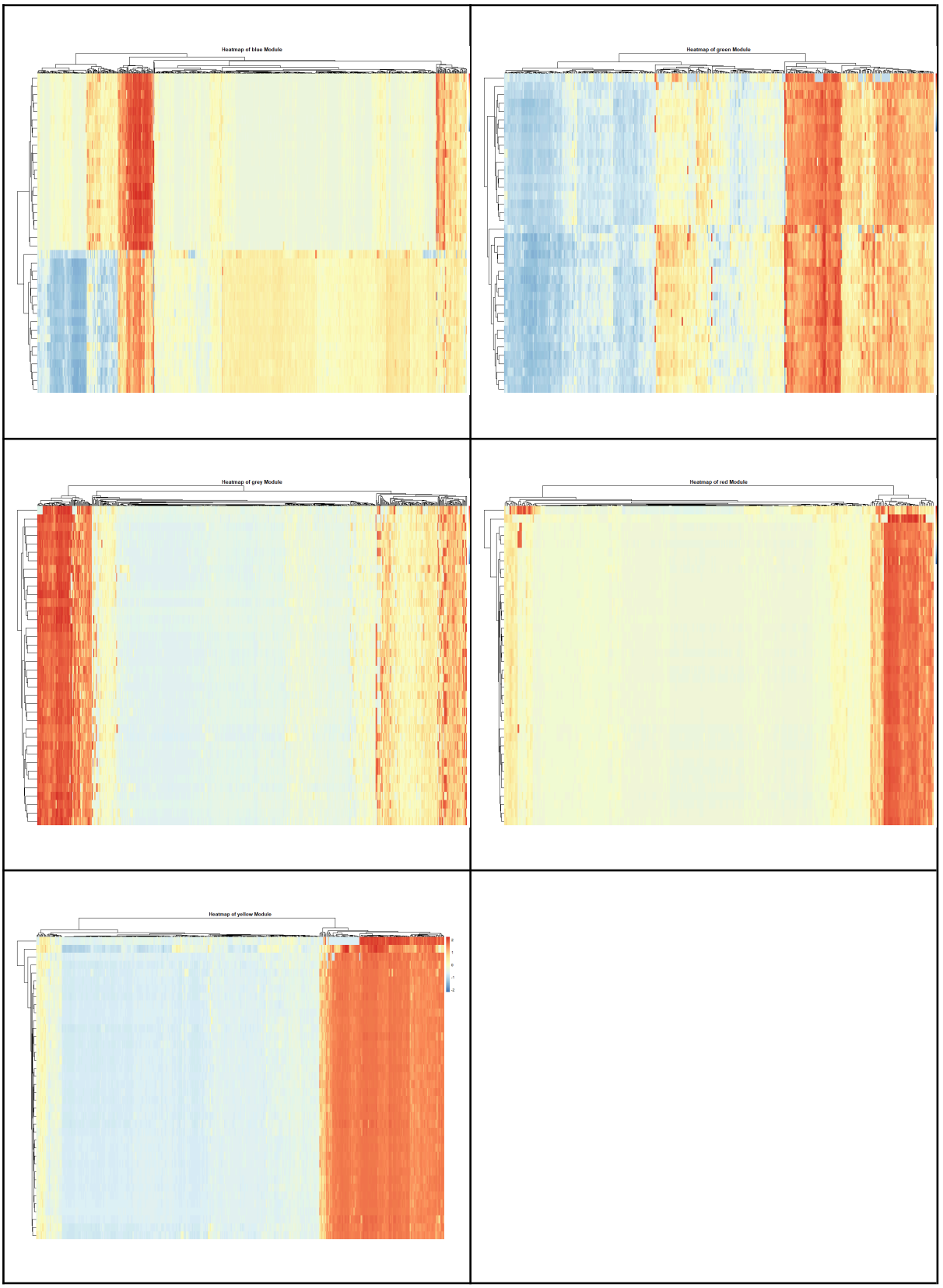


Figure 68: Eigengene adjacency heatmap (Normal)

Modules Identified:

Table 13: Gene modules normal + tumour gene selection




Heatmap of blue Module
Heatmap of green Module
Heatmap of grey Module
Heatmap of red Module
Heatmap of yellow Module

Chapter 5: Conclusion

Section 5.1 Conclusion for each section of analysis

In our analyses of DNA methylation using TCGA colon cancer, we utilised various bioinformatic methodologies and disciplines to capture and differentiate methylation and gene expression in both normal solid tissue and tumour tissues. We first utilised correlation and pairplots to explore the relationship amongst immune and methylation-related sets of genes, isolating differences in variance and stability. We then carried out a PCA (Principal Component Analysis) and survival analysis to identify genes core to regulation, as well as clinical and demographic correlations/associations. Finally, we employed Weighted Gene Co-expression Network Analyses (WGCNA) to enrich genes from normal and tumour tissue samples with pathway data and formulate functionally connected gene modules, providing insights into how pathways and functions are disturbed by cancer. By combining these analyses, we aimed broadly to understand the epigenetic modifications that occur in the body subject to a sepsis-associated illness like cancer, and in doing so have identified biomarkers with potential value for diagnosis and/or prognosis.

Table 14: Comparative analysis for key genes

DNA Methylation Regulators	**Immune System Related**	**Signal Transduction and Cell Communication**
DNMT3A	TNFSF10	RSPO3
DNMT3B	CTLA4	NTNG1
DNMT3L	HMGB1	**Transcription Factors and Signalling Molecules**
UHRF1	NFKB1	HEYL
UHRF2	HAVCR2	CXCL12
TET2	IL13	
MBD1	LAT	
MBD3		
ZBTB4		
DNMT1		

For a clearer presentation of results, conclusions were divided into 3 parts, as per corresponding analysis.

Section 5.1.1 Analysis 1 conclusion

The main focus of the analysis was to use correlation and pair plot to check the relationship of methylation and immune set of genes among themselves. Correlation plot showed whether the relationship between the genes was positive or negative and to what extent. Absolute value correlation plot allowed to see the biggest difference in genes relationships. Pairplot showed a scatterplot of the values for each gene individually and the histogram showed average range of values for each gene. Those analysis graphs were really important as they showed that for a normal solid tissue methylation values between the defined set of genes were very unstable, with some genes rising, the others going down, or both genes rising and going down simultaneously. In the article of Moore, Le and Fan (2012) it is proven that in a normal solid tissue methylation values usually change among themselves. Whereas on the other hand correlation plot showed that for both sets of genes for tumour tissue methylation values stayed relatively stable for all genes. Out of these plots the most significant genes were defined, whose values changed between normal solid tissue and tumour tissue significantly for the immune set of genes TNFSF10, CTLA4, HMGB1, NFKB1, HAVCR2, IL13, LAT and for methylation set - MBD1, MBD3, DNMT3B, DNMT3L, ZBTB4, UHRF2, DNMT1, TET2 , which were proved to be the same on the pairplot and using the t-test analysis. This research answers the question of “Does cancer have an association with the methylation value of the gene” and looks at the correlation of methylation patterns between normal solid tissue and tumour tissue.

Section 5.1.2 Analysis 2 conclusion

For **PCA analysis**, in PC1 specific genes such as NTNG1 and RSPO3 have significant contributions in PC1. According to pca analysis, the methylation values of these genes would be higher in normal human genes than in tissue with cancer. While HEYL and CXCL12 are more prominent in PC2. The higher the methylation values of these genes in PC2, the higher the degree of cancer metastasis and progression. However, the number of metastatic cases is too small to support this idea. These genes are essential for understanding the epigenetic regulation associated with tumour development, tumour type differentiation and metastatic processes.

Furthermore, **differential methylation analysis** (volcano plot) identified significant alterations in key genes such as DNMT3A, DNMT3L and UHRF1, which are essential for DNA methylation and gene regulation and are frequently disrupted in cancer. The up-regulation of DNMT3A and altered methylation status of UHRF1 in cancer tissues underscore their potential role in tumourigenesis.

Survival analyses showed that clinical stage had a significant impact on prognosis, with early-stage colon cancer having the highest survival rates. Interestingly, demographic factors such as gender and ethnicity had no significant impact on survival, as could demographic factors on methylation in a previous PCA analysis. Therefore, biologic and molecular markers will be able to provide a more accurate prognostic tool than demographic variables.

Section 5.1.3 Analysis 3 conclusion

The greater frequency of identifiable modules in the normal solid tissue data regardless of gene selection (9 modules for normal solid tissue Gene Selection, 8 for tumour tissue) might imply an overall higher combined function in normal solid tissue data overall.

The lower number of identifiable modules in Colon Cancer patient data (6 with tumour data gene selection, 4 with health gene selection) indicates lesser interconnected module function, and thus implies overall lower function.

The colour strength indicator of the adjacency matrix heatmap for normal solid tissue data with health gene selection implies specifically correlated modules, indicated by the contrasting highly negatively correlated modules. This is seen to a lesser extent in the tumour data and tumour gene selection.

WGCNA module heatmap conclusion

The similarities and differences seen in the heatmaps based on different conditions and criteria indicate biological pathways that are constant and those which are not. From this we can intuit what overall effect the condition has on genes in terms of module pathways.

The modules that we can only identify in tumour sample data likely stems from cancer-related biological processes, whereas modules identifiable in normal sample data that don’t appear in tumour tissuedata might be pathways affected by cancer-related processes.

Each of the four Heatmaps represents the Topological Overlap Matrix (TOM) for each network created.

The first two networks are the baseline. The Tumour Data + Tumour Variance Genes represents the Colon cancer tissue dataset networked via the top 5,000 genes by variance. It is highly methylated with certain modules or sub-modules unmethylated

The Normal Data+Normal Variance, is normal tissue data networked via the top 5,000 genes by variation in the normal solid tissue data. It is conversely is primarily weakly methylated, though strong overlap along the left and top sides indicated that this methylation is a constant between the two and not influenced by the patient's cancer specifically.

The remaining two heatmaps represent The Tumour tissue Data networked via the top 5,000 normal genes, and the normal solid tissue data networked via the top 5,000 genes. They share similarities with both their data source and their gene references from the alternative dataset. This indicates overlaps and underlaps of methylated areas in the gene networks.

It can be said that correlation analysis and the eigengene and WGCNA module heatmaps demonstrated a higher frequency of identifiable functional modules in normal data, indicating greater overall functionality, whereas cancer data showed fewer, less interconnected modules or less "stable" relationships between genes. This research provides valuable insights into the epigenetic changes associated with cancer and emphasises the potential of methylation patterns as biomarkers for diagnosis and prognosis.

Section 5.2 Discussion based on past literature

Similar research was done by Dobre et al. (2021) related to analysing colon adenocarcinoma patient outcomes using methylation-driven genes and Chen et al. (2023) research, where the main focus was more on gene mutations. Both researchers compared normal solid tissue and tumour tissue data as well, although in the first paper 9807 genes were analysed and in the second paper the sample included only 22 genes.

What is different from the current research, Chen et al. (2023) studied mutations in the KRAS and BRAF genes, which occur in about 10% and 40% of colorectal cancer cases respectively. Additionally these genes are proved to be linked to dysregulated DNA methylation and

miRNA expression. Mutations were not analysed in the current piece of work and is definitely an interesting area for further research.

Dobre *et al*. (2021) analysis shows very similar background information and dataset description to the current work, also unbalanced datasets for normal solid tissue and tumour tissue, consisting of 41 normal solid tissue samples and 476 tumour tissue samples. The analysis focused on building the model to predict patient outcomes, similar to the current research, which concentrates more on defining the different set of genes. Overall, these researches show that the analysed topic is very valuable to the study of methylation genes and has a lot of future potential for more broad analysis.

In a study by Casalino and Verde (2020) it was noted that changes in gene methylation status are key to differentiating between normal and cancerous tissues. Methylation values of these genes were found to be higher in normal tissues than in cancer tissue in PCA analysis, which is consistent with the emphasis on the role of methylation in the maintenance of normal cellular function and suppression of tumourigenesis (Johnson and Lal, 2016).

Furthermore, the findings on HEYL and CXCL12 in PC2 contributing to cancer metastasis and progression resonate with Yuan *et al* (2019), who highlighted the involvement of CXCL12 in the migration and invasion of cancer cells. However, due to the limited number of metastatic cases in the dataset used for the analysis, data from more metastatic cases are needed to guess more definitively the relationship between the methylation levels of these genes and the extent of cancer metastasis.

Differential methylation analysis revealed that DNMT3A, DNMT3L and UHRF1 were significantly altered in cancer tissue. Moore, Le and Fan (2012), who discussed the key role these genes play in DNA methylation and gene regulation, which are frequently disrupted in cancer. Besides, the upregulation of DNMT3A and the altered methylation status of UHRF1 highlight their presence and potential role in promoting tumourigenesis.

Survival analyses have shown that clinical stage has a significant impact on prognosis, with early-stage colon cancers having the highest survival rates.This is supported by Therneau (2015), who emphasised the importance of cancer stage on prognosis. Interestingly, survival analyses found that demographic factors such as gender and ethnicity did not have a significant impact on survival, suggesting that biological and molecular markers may be more

accurate in prognostic judgements. This result is consistent with the findings of Hassan *et al*. (2022), who noted that genetic and molecular markers have better predictive power than demographic variables in cancer prognosis.

Section 5.3 Clinical implications

These analyses provide insights into the implications of DNA Methylation in Colon Cancer tissue, and may enable further understanding of methylation differences in Sepsis tissue. By examining methylation patterns from three directions, we have identified differences between normal solid tissues and tumour tissues that might have possible diagnostic and prognostic applications.

Section 5.3.1 Identification of key genes

We identified several key genes with significant differences in methylation between normal solid tissue and tumour tissue. Notably, genes such as DNMT3A, DNMT3B, and UHRF1 showed altered methylation patterns in cancer tissues, highlighting their potential roles in tumourigenesis. These findings are consistent with the literature, suggesting that disruptions in these genes contribute to the development and progression of cancer. Clinically, these genes could serve as biomarkers for early detection and targeted therapy.

Section 5.3.2 Survival analysis

The survival analysis underscored the impact of clinical stage on prognosis, with early-stage colon cancer tissue exhibiting higher survival rates. This aligns with existing studies emphasising the importance of early detection and intervention. Interestingly, demographic factors such as gender and ethnicity were not significant predictors of survival, suggesting that molecular markers are more reliable indicators of patient outcomes. This finding supports the use of genetic and epigenetic markers in personalised medicine to improve prognostic accuracy.

Section 5.3.3 Gene modules and functional pathways

The eigengene and WGCNA module heatmaps demonstrated a higher frequency of identifiable functional modules in normal solid tumour data compared to cancer data. This indicates that cancer disrupts the stability and connectivity of gene networks, which could be linked to the loss of normal cellular functions and the emergence of cancer-related pathways. These altered pathways could aid in the development of targeted strategies targeting specific gene modules affected by cancer.

Section 5.3.4 Tumour-specific methylation patterns

The differential methylation analysis revealed tumour-specific patterns, with certain modules only identifiable in tumour samples. These tumour-specific modules likely represent cancer-related biological processes, providing insights into the mechanisms of tumour development and progression. Findings like these could influence patient-specific therapies aimed at these pathways, potentially improving treatment quality and outcomes.

The potential of DNA methylation patterns as biomarkers for diagnosis, prognosis, and targeted therapy in colorectal cancer are only beginning to be explored. By identifying key genes and disrupted pathways, this research contributes to the understanding of cancer biology and the nature of sepsis.

Section 5.4 Limitations of the study

1. **Sample size:**

One of the limitations was the small sample size of normal solid tissue in the analysis. There were 38 records in the normal tissue dataset and 314 records in the tumour tissue dataset.

To balance the data, propensity score matching was used to downsample the tumour dataset, ensuring a more balanced comparison between groups. However, this method led to leaving out some data, which may have impacted the final results. And since smaller datasets have less ability to detect significant differences in the data, a bigger amount of patient's data to analyse would help to make more discoveries about the genes relationships and methylation patterns (Analytics, 2024).

Besides, a major limitation is the relatively small number of metastatic cases in the dataset. The overall sample size and diversity may limit the ability to detect subtle differences in methylation profiles between different demographic groups (e.g., Asian, mixed race, etc.) or less common cancer subtypes.

2. **Methodological limitations:**

This section relies heavily on PCA to distinguish between different types of cancer tissue based on methylation profiles. However, PCA may not be effective in capturing non-linear relationships or interactions between genes, which are critical for understanding complex

diseases such as cancer. Furthermore, PCA relies on variance as a measure of significance, which may overlook biologically significant but less variable features.

3. **Computational limitations**

As we were working with such large datasets that interact in complex ways, owing to the complexity of human biology, computing power and time were major limitations on the scope of research we could undertake. Our equipment consisted of two 16GB RAM laptops and a 32GB of RAM laptop, sufficient for working with some less computationally intensive analysis techniques but not for the more complex calculations required for certain techniques.

To get around this, an Amazon Web Service Elastic 2 remote computing instance was set up. By loading the instance with 128GB of RAM and 16 cores, we expanded the scope of our analysis. Parallel Computing on 15 of the cores allowed simultaneous calculations for these intensive analytics techniques. Ultimately, though computationally powerful enough, the AWS EC2 free trial expired, incurring fees, and the instance was shut down before it could finish its processes.

Section 5.5 Future research directions

A lot of future research can be done based on the foundations that the current report focused on. Firstly, it would be very interesting to observe if the same patterns of gene relationships would stay the same for the other cancer types. By comparing these patterns, common epigenetic changes and unique characteristics specific to certain cancers, can be possibly identified thus leading to improvements in cancer treatments.

Given greater computational resources and a less strict time frame, a much higher-resolution network analysis of biological pathways is achievable. In isolating significantly bigger clusters, gene ontology could be employed to identify individual genes and annotate with tangible biological function information. Larger, more functional modules make a comparison of normal and cancerous tissue more indicative of the health of the patients.

Bibliography

Akoglu, H. (2018) User's guide to correlation coefficients, *Turkish Journal of Emergency Medicine/TüRkiye Acil Tıp Dergisi*, 18(3), pp. 91–93. Available at: https://doi.org/10.1016/j.tjem.2018.08.001. (Accessed: 16 July 2024)

Altman, D.G. *et al.* (2012) Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK): Explanation and Elaboration, *PLoS Medicine*, 9(5), p. e1001216. Available at: https://doi.org/10.1371/journal.pmed.1001216. (Accessed: 6 July 2024)

Analytics Emerging India (2021) Decoding data size: Pros and cons of working with small data sets, *Medium,* Available at: https://medium.com/@analyticsemergingindia/decoding-data-size-pros-and-cons-of-working-with-small-data-sets-bc1ea0792da6#:~:text=Unlike%20big%20data%2C%20which%20offers,significant%20impact%20on%20the%20results. (Accessed: 12 July 2024)

Ashcroft, F. M., & Rorsman, P. (2012). *Diabetes mellitus and the β cell: the last ten years.* Cell, 148(6), 1160-1171.

Babu, M. *et al.* (2008) Altered gene expression changes in Arabidopsis leaf tissues and protoplasts in response to Plum pox virus infection, *BMC Genomics*, 9(1). Available at: https://doi.org/10.1186/1471-2164-9-325. (Accessed: 13 July 2024)

bartlett — SciPy v1.14.0 Manual (no date). Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html.

Baylin, S.B. and Jones, P.A. (2011) A decade of exploring the cancer epigenome — biological and translational implications, *Nature Reviews. Cancer*, 11(10), pp. 726–734. Available at: https://doi.org/10.1038/nrc3130. (Accessed: 12 July 2024)

Bock, C. *et al.* (2006) 'CpG Island Methylation in Human Lymphocytes Is Highly Correlated with DNA Sequence, Repeats, and Predicted DNA Structure,' *PLOS Genetics*, 2(3), p. e26.Available at: https://doi.org/10.1371/journal.pgen.0020026. (Accessed: 16 June 2024)

Bose, P., PhD. (2022, September 15). *The role of DNA methylation in human disease*. Genomics Research From Technology Networks. Available at: https://www.technologynetworks.com/genomics/articles/the-role-of-dna-methylation-in-human-disease-355520.(Accessed: 18 July 2024)

Cancer facts and statistics (no date). Available at: https://www.cancer.org/research/cancer-facts-statistics.html. (Accessed: 5 July 2024)

Casalino, L. and Verde, P. (2020) Multifaceted Roles of DNA Methylation in Neoplastic Transformation, from Tumor Suppressors to EMT and Metastasis, *Genes*, 11(8), p. Available at: 922. https://doi.org/10.3390/genes11080922. (Accessed: 2 July 2024)

Cedar, H. and Bergman, Y. (2009) 'Linking DNA methylation and histone modification: patterns and paradigms,' *Nature Reviews. Genetics*, 10(5), pp. 295–304. Available at: https://doi.org/10.1038/nrg2540. (Accessed: 7 July 2024)

Chen, D. *et al.* (2023b) 'Development of a prognostic model for personalized prediction of colon adenocarcinoma (COAD) patient outcomes using methylation-driven genes,'

Journal of Applied Genetics/Journal of Applied Genetics, 64(4), pp. 713–721. Available at: https://doi.org/10.1007/s13353-023-00778-4. (Accessed: 11 July 2024)

Colaprico, A. *et al.* (2015) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data, *Nucleic Acids Research*, 44(8), p. e71. Available at: https://doi.org/10.1093/nar/gkv1507.(Accessed: 16 May 2024)

Costa, M.-C.D. and Johannes, F. (2020) Epigenetics: Switching genes on and off, *Frontiers for Young Minds*, 8. Available at: https://doi.org/10.3389/frym.2020.554136. (Accessed: 13 May 2024)

DBGAP study (no date). Available at: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v3.p3. (Accessed: 17 June 2024)

Dobre, M. *et al.* (2021b) Crosstalk between DNA methylation and gene mutations in colorectal cancer, *Frontiers in Oncology*, 11. Available at: https://doi.org/10.3389/fonc.2021.697409. (Accessed: 17 June 2024)

Eckhardt, F. *et al.* (2006) 'DNA methylation profiling of human chromosomes 6, 20 and 22,' *Nature Genetics*, 38(12), pp. 1378–1385. Available at: https://doi.org/10.1038/ng1909. (Accessed: 26 June 2024)

Esteller, M. (2008) 'Epigenetics in cancer,' *New England Journal of Medicine*, 358(11), pp. 1148–1159. Available at: https://doi.org/10.1056/nejmra072067. (Accessed: 2 July 2024)

Feil, R. and Fraga, M.F. (2012) 'Epigenetics and the environment: emerging patterns and implications,' *Nature Reviews. Genetics*, 13(2), pp. 97–109. Available at: https://doi.org/10.1038/nrg3142. (Accessed: 2 July 2024)

Feinberg, A.P., Ohlsson, R. and Henikoff, S. (2006) 'The epigenetic progenitor origin of human cancer,' *Nature Reviews. Genetics*, 7(1), pp. 21–33. Available at: https://doi.org/10.1038/nrg1748. (Accessed: 2 July 2024)

Fraga, M.F. *et al.* (2005) Epigenetic differences arise during the lifetime of monozygotic twins, *Proceedings of the National Academy of Sciences of the United States of America*, 102(30), pp. 10604–10609. Available at: https://doi.org/10.1073/pnas.0500398102. (Accessed: 2 July 2024)

GeeksforGeeks (2022) *Python seaborn.pairplot() method*. Available at: https://www.geeksforgeeks.org/python-seaborn-pairplot-method/. (Accessed: 5 July 2024)

Gene expression (no date). Available at: https://www.nature.com/scitable/topicpage/gene-expression-14121669/. (Accessed: 7 May 2024)

Guo, F. *et al.* (2017) Truncated apolipoprotein C-I induces apoptosis in neuroblastoma by activating caspases in the extrinsic and intrinsic pathways, *Oncology Reports*, 38(3), pp. 1797–1805. Available at: https://doi.org/10.3892/or.2017.5819. (Accessed: 2 July 2024)

Hanahan, D., & Weinberg, R. A. (2011). *Hallmarks of cancer: the next generation.* Cell, 144(5), 646-674.

Hassan, M. *et al.* (2022) Innovations in Genomics and big data analytics for personalized medicine and health Care: a review, *International Journal of Molecular Sciences*, 23(9), p. 4645. Available at: https://doi.org/10.3390/ijms23094645. (Accessed: 11 July 2024)

Hayes, A. (2024) *What is a Two-Tailed Test? Definition and example*. Available at: https://www.investopedia.com/terms/t/two-tailed-test.asp. (Accessed: 27 June 2024)

Hegi, M.E. *et al.* (2005) MGMTGene Silencing and Benefit from Temozolomide in Glioblastoma, New *England Journal of Medicine*, 352(10), pp. 997–1003. Available at: https://doi.org/10.1056/nejmoa043331. (Accessed: 19 June 2024)

Horvath, S. (2011). Weighted network analysis: applications in genomics and systems biology. *Springer Science & Business Media.*

Hosseini, S. (2023) *How to do a T-Test in Python*. Available at: https://builtin.com/data-science/t-test-python. (Accessed: 27 June 2024)

How common is colorectal cancer? (no date). Available at: https://www.cancer.org/cancer/types/colon-rectal-cancer/about/key-statistics.html. (Accessed: 1 July 2024)

Institute for Quality and Efficiency in Health Care (IQWiG) (2023) *In brief: How does the immune system work?* Available at: https://www.ncbi.nlm.nih.gov/books/NBK279364/. (Accessed: 9 July 2024)

Jin, B. and Robertson, K.D. (2012) 'DNA methyltransferases, DNA damage repair, and cancer,' in *Advances in experimental medicine and biology*, pp. 3–29. Available at: https://doi.org/10.1007/978-1-4419-9967-2_1. (Accessed: 12 June 2024)

Jones, P.A. and Baylin, S.B. (2007) 'The epigenomics of cancer,' *Cell*, 128(4), pp. 683–692. Available at: https://doi.org/10.1016/j.cell.2007.01.029. (Accessed: 11 July 2024)

Laity, J.H., Lee, B.M. and Wright, P.E. (2001) 'Zinc finger proteins: new insights into structural and functional diversity,' *Current Opinion in Structural Biology*, 11(1), pp. 39–46. Available at: https://doi.org/10.1016/s0959-440x(00)00167-6. (Accessed: 8 July 2024)

Lanata, C.M., Chung, S.A. and Criswell, L.A. (2018) 'DNA methylation 101: what is important to know about DNA methylation and its role in SLE risk and disease heterogeneity,' *Lupus Science & Medicine*, 5(1), p. e000285. Available at: https://doi.org/10.1136/lupus-2018-000285. (Accessed: 3 July 2024)

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics, 9*(1), 559.

Liu, K. (2016) 'Dendritic cells,' in *Elsevier eBooks*, pp. 741–749. Available at: https://doi.org/10.1016/b978-0-12-394447-4.30111-0. (Accessed: 24 June 2024)

Lynch, H.T. and De La Chapelle, A. (2003) 'Hereditary colorectal cancer,' *New England Journal of Medicine*, 348(10), pp. 919–932. Available at: https://doi.org/10.1056/nejmra012242. (Accessed: 1 July 2024)

Mann Whitney U Test (Wilcoxon Rank Sum Test) (no date). Available at: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html. (Accessed: 2 July 2024)

MBD5 gene: MedlinePlus Genetics (no date). Available at: https://medlineplus.gov/genetics/gene/mbd5/. (Accessed: 3 July 2024)

Milicic, L. *et al.* (2023) 'Utility of DNA methylation as a biomarker in aging and Alzheimer's disease,' *Journal of Alzheimer's Disease Reports*, 7(1), pp. 475–503. Available at: https://doi.org/10.3233/adr-220109. (Accessed: 3 June 2024)

Moore, L.D., Le, T. and Fan, G. (2012) 'DNA methylation and its basic function,' *Neuropsychopharmacology*, 38(1), pp. 23–38. Available at: https://doi.org/10.1038/npp.2012.112. (Accessed: 1 July 2024)

NCCN guidelines for patients: Colon cancer (2022). Available at: https://www.nccn.org/patientresources/patient-resources. (Accessed: 29 June 2024)

NCI Dictionary of Cancer Terms (no date). Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/colon-cancer#. (Accessed: 6 July 2024)

normaltest — SciPy v1.14.0 Manual (no date). Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html. (Accessed: 3 July 2024)

Ollberding, N.J. *et al.* (2011) 'Racial/ethnic differences in colorectal cancer risk: The multiethnic cohort study,' *International Journal of Cancer*, 129(8), pp. 1899–1906. Available at: https://doi.org/10.1002/ijc.25822. (Accessed: 1 July 2024)

PSMpy (2023).Available at: https://pypi.org/project/psmpy/. (Accessed: 3 July 2024)

Siegel, R.L., Miller, K.D. and Jemal, A. (2020) 'Cancer statistics, 2020,' *CA: A Cancer Journal for Clinicians*, 70(1), pp. 7–30. Available at: https://doi.org/10.3322/caac.21590. (Accessed: 28 May 2024)

Statista (2024) *U.S. population by sex 1980-2022*. Available at: https://www.statista.com/statistics/241495/us-population-by-sex/. (Accessed: 20 May 2024)

Stresemann, C. and Lyko, F. (2008) 'Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine,' *International Journal of Cancer*, 123(1), pp. 8–13. Available at: https://doi.org/10.1002/ijc.23607. (Accessed: 16 June 2024)

The Cancer Genome Atlas Program (TCGA) (no date). Available at: https://www.cancer.gov/ccg/research/genome-sequencing/tcga. (Accessed: 12 May 2024)

The information in DNA determines cellular function via translation (no date). Available at: https://www.nature.com/scitable/topicpage/the-information-in-dna-determines-cellular-function-6523228/. (Accessed: 3 July 2024)

The information in DNA is decoded by transcription (no date). Available at: https://www.nature.com/scitable/topicpage/the-information-in-dna-is-decoded-by-6524808/. (Accessed: 5 July 2024)

Therneau, T. (2024) *A package for survival analysis in R*. Available at: https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf. (Accessed: 20 May 2024)

ttest_ind — SciPy v1.14.0 Manual (no date). Available at: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html. (Accessed: 3 July 2024)

Unoki, M. and Sasaki, H. (2022) 'The UHRF protein family in epigenetics, development, and carcinogenesis,' *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, 98(8), pp. 401–415. Available at: https://doi.org/10.2183/pjab.98.021. (Accessed: 2 July 2024)

Wang, H. (2022) 'A Practical Guide to Quasi-Experimental Methods (PSM and DID),' *harrywang.me*, 1 June. Available at: https://harrywang.me/psm-did. (Accessed: 19 July 2024)

WCRF International (2022) *Diet, activity and cancer - WCRF International*. Available at: https://www.wcrf.org/dietandcancer. (Accessed: 14 July 2024)

What is a cell? (no date). Available at: https://www.nature.com/scitable/topicpage/what-is-a-cell-14023083/. (Accessed: 15 July 2024)

Why Are People with Cancer More Likely to Get Infections? (no date). Available at: https://www.cancer.org/cancer/managing-cancer/side-effects/infections/why-people-with-cancer-are-at-risk.html. (Accessed: 19 June 2024)

Wilson, K.M., Giovannucci, E. and Mucci, L.A. (2011) 'Response: Re: Coffee consumption and prostate Cancer risk and progression in the Health Professionals follow-up study,' *Journal of the National Cancer Institute*, 103(19), pp. 1481–1482. Available at: https://doi.org/10.1093/jnci/djr306. (Accessed: 1 July 2024)

World Health Organization: WHO (2022) *Cancer*. Available at: https://www.who.int/news-room/fact-sheets/detail/cancer. (Accessed: 1 June 2024)

Yuan, L. *et al.* (2019) 'Overexpression of LINC00037 represses cervical cancer progression by activating mTOR signaling pathway,' *Journal of Cellular Physiology*, 234(8), pp. 13353–13360. Available at: https://doi.org/10.1002/jcp.28012. (Accessed: 12 July 2024)

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology, 4*(1).

Zhang, B., Zhu, J., & Horvath, S. (2013). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology, 4*(1).

Appendix

Link for the code:

https://github.com/viola-dzianisava/Analysing-DNA-methylation-in-patients-with-colon-cancer/tree/main