THE UNIVERSITY *of* EDINBURGH

Natural therapeutics: Cross-kingdom microRNA and its potential to target the RNA landscape of glioblastoma

Amanda Vanessza Fentor

THE UNIVERSITY *of* EDINBURGH

Doctor of Philosophy with Integrated Study

Institute of Genetics and Cancer

University of Edinburgh

2025

Declaration

I hereby declare that this thesis, has been composed by me, this work is my own unless otherwise stated in the text. This work has not been submitted for any other degree or professional qualification.

Vanessza Fentor, June 2025

Additional declarations

The original aim of this work was to collaborate with MirNat on the improvement of the MirCompare algorithm, which was built to assess sequence complementarity between plant microRNAs and mammalian mRNAs. This collaboration would have built on their findings plant microRNAs being involved in cross-kingdom interactions. However, due to personnel changes 8 months after starting my PhD, I had to deviate from this plan, as the collaboration was no longer possible. Therefore, my supervisor and I switched to using the only available cell lines and tissues in the lab, glioblastoma, as a proof of concept in order to carry forward the idea of the potential of plant microRNA medicines in humans.

As a student on the Precision Medicine doctoral programme, I also had to take classes for credits and complete the associated coursework. Additionally, this work began during Covid lockdown in 2020; therefore, the project's initial progression was much slower than the pace of a normal PhD experience.

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF) (http://www.ecdf.ed.ac.uk/).

Abstract

GBM is the most common and lethal primary brain tumour, characterized by rapid growth and a high propensity for recurrence. Despite advancements in surgical, chemotherapeutic, and radiotherapeutic interventions, the median survival time for GBM patients remains low, often less than 15 months post-diagnosis. This underscores the urgent need for innovative research to unravel the molecular underpinnings of GBM and identify novel therapeutic targets. Glioblastoma Stem Cells (GSCs) play a pivotal role in enhancing the stem-like state of tumour cells, promoting pro-migratory and pro-invasive factors that fortify the tumour's immunosuppressive microenvironment. This immunosuppression facilitates tumour maintenance, progression, recurrence, and resistance to conventional therapies, posing substantial challenges for immunotherapy and novel drug development. GSCs are commonly used to model GBM due to their ability to reciprocate key features of the disease. However, in this study, which integrates both cell line and tissue transcript expression analyses, demonstrates that this approach might not fully capture the complexity of the disease. The discrepancies between cell line models and actual tissue samples highlight the need for more representative models in GBM research.

Cell adaptation to external and internal stressors is fundamentally governed by modifications in gene expression, which can be quantitatively assessed at the transcriptomic level. Hypoxic stress, a hallmark of Glioblastoma (GBM) tumorigenesis, has profound effects on brain cells, contributing significantly to the disease's aggressiveness and poor prognosis. In my PhD research, I employed comprehensive transcriptomic analyses to uncover both novel hypoxia-induced RNA transcripts in both GSCs and primary GBM tissues with particular focus on those implicated in tumour recurrence. This study integrated bulk RNAseq data and bioinformatics pipelines to map hypoxia-associated transcriptomic alterations in GSC models and recurrent GBM samples. Strikingly, tissue-specific analyses revealed distinct molecular signatures that were not detect in the cell lines: collagens associated with tumour recurrence, and HOX family alleles linked to tumour grade progression. These findings underscore the limitations of cell line models and emphasize the need for systems that better recapitulate the spatial and molecular heterogeneity of GBM. Utilizing a pathological approach, I conducted tissue microarray (TMA) staining to validate the presence and spatial distribution of these identified transcripts within tumour samples. These validated genes were subsequently subjected to medicinal microRNA target prediction using miTAR, a deep learning algorithm, to identify microRNAs that potentially regulate these novel driver transcripts. Given that microRNAs exert post-transcriptional control by downregulating their target mRNAs, the identification of such regulatory interactions is crucial for understanding the mechanisms driving GBM pathology.

Moreover, my research highlights the therapeutic potential of medicinal plant-derived microRNAs. MicroRNAs have the ability to remain stable, enabling them to traverse the harsh gut environment and exert cross-kingdom effects. This property positions them to potentially serve as natural adjuncts to patient diets in clinical settings, offering a complementary approach to traditional therapies. Specifically, medicinal plant microRNAs could be integrated as dietary supplements to enhance recovery and mitigate disease progression, providing a novel, non-invasive strategy to bolster patient health during treatment and recovery phases.

In conclusion, my findings underscore the intricate relationship between hypoxic stress, gene expression, and GBM progression, while also illuminating the promising role of medicinal plant microRNAs in therapeutic interventions. These microRNAs not only represent a natural, easily integrable supplement to patient diets but also offer potential for novel therapeutic avenues aimed at combating GBM and improving patient outcomes. Given the poor prognosis and limited treatment options for GBM, advancing our understanding of its molecular mechanisms and exploring innovative treatments is of utmost importance.

Lay Summary

Glioblastoma (GBM) is one of the most aggressive and deadly brain cancers, with most patients surviving less than 15 months after diagnosis, even with advanced treatments like surgery, chemotherapy, and radiation. A major challenge in fighting GBM lies in the tumour's ability to resist therapies and return after treatment. My research explores an unconventional and innovative approach: using small, naturally occurring molecules microRNAs (miRNAs) found in plants as a potential therapeutic agent to treat GBM. These miRNAs act like tiny genetic regulators (also found in humans and animals), capable of targeting specific genes and turning them "off". By harnessing this natural mechanism, we could potentially disrupt key processes that allow cancer cells to grow and survive.

To identify promising miRNAs, I analysed gene activity in GBM cell lines and tissues using advanced RNA sequencing techniques. Tissues were used in addition to cell lines as they contain immune cells and other factors that aren't modelled by cell lines alone. My analysis pinpointed several genes overactive in GBM, like NDRG1 and EGLN3 from hypoxic glioblastoma cell lines. But the tissue data told the more interesting story - uncovering recurrence-linked factors that cell lines missed entirely, particularly HOX gene family members (HOXC10 and HOXB3). These aren't just bystanders; they're active players in tumour growth and progression, making them prime targets for therapy.

I then used computational tools, including miTAR and MepmiRDB, to screen plant miRNAs that could theoretically bind to and silence these cancer-driving genes. My focus was on commercially available medicinal plants, such as kiwi, to ensure that any discoveries could feasibly developed into treatments. The computational process generated a vast dataset of potential miRNA-mRNA pairs. These results represent a key starting point for future researchers, offering thousands of candidate interactions that may regulate critical GBM genes. This extensive dataset provides the bases to begin experimental validation, enabling the scientific community to continue my project.

This research provides a foundation for a novel approach to GBM treatment. By leveraging plant miRNAs, we could develop therapies that are affordable, accessible, and based on natural biological processes. While experimental validation is still needed, this work opens the door to an entirely new way of tackling cancer using the power of plants.

Acknowledgements

I would like to thank my supervisors, for providing me with this PhD project opportunity and for helping me develop the skills to lead a project independently.

To my friends who have given their time and support, thank you! Especially, Sára, and Sarah – thank you so much for getting me through this journey. Every tear of sadness and joy helped immensely, and without your support I would have never gotten here.

To my family back home – Anya, Apa, Bazsi, Papika, Muci mama and Nagymama – thank you for giving me all the opportunities that got me here. Thank you for your support even when my requests and boundaries didn't make sense to you. Thank you for always giving me perspective and making sure I stayed on this path. I would never have gotten here without your sacrifices.

Lastly, but absolutely most importantly, I'd like to thank my husband, Tony! During this crazy time, you have stuck by me at every corner, so much so that you proposed, and we got married. This might just be the ultimate show of love and support. Your belief in me never wavered, not even when I doubted myself – which, honestly, was a lot. I proudly share this accomplishment with you, and I will forever be grateful for every effort you made to get me across the finish line – especially bribing me with doughnuts. Thank you for always being there for me, but especially during these past years when I didn't even know I needed you the most.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Appendix Figures

List of Appendix Tables

Abbreviations

AI	Artificial Intelligence
ALT	Alternative Lengthening of Telomeres
ANOVA	Analysis Of Variance
AOX1	Aldehyde Oxidase 1
ASRP	Arabidopsis Small RNA Project
ATACseq	Assay For Transposase-Accessible Chromatin Sequencing
BBB	Blood-Brain Barrier
ceRNA	Competing Endogenous RNA
ChIPseq	Chromatin Immunoprecipitation Sequencing
CLCBio	CLC Genomics Workbench
CNS	Central Nervous System
EMT	Epithelial-Mesenchymal Transition
GBM	Glioblastoma Multiforme
GSC	Glioblastoma Stem Cell
HIF	Hypoxia-Inducible Factors
KEGG	Kyoto Encyclopaedia Of Genes and Genomes
MAOB	Monoamine Oxidase B
miRNA	microRNA
ML	Machine Learning
mRNA	messengerRNA
PCR	Polymerase Chain Reaction
QC	Quality Control
qPCR	Quantitative Real Time PCR
RB	Retinoblastoma-Associated
RISC	RNA-Induced Silencing Complex
ROS	Reactive Oxygen Species
shRNA	Short Hairpin RNA
TCTCE	Tumour Cell to Tumour-Cell Environment
TE	Tumour Environment
TERTp	Telomerase Reverse Transcriptase Promoter
TME	Tumour Microenvironment
TOE	Tumour Organismal Environment
UPR	Unfolded Protein Response
WHO	World Health Organization

CHAPTER 1. Introduction

1.1. Precision Medicine

Various definitions exist for precision medicine, or sometimes referred to as personalised medicine, Precision Medicine UK defines it follows:

> "Precision medicine refines our understanding of disease prediction and risk, onset and progression in patients, informing better selection and development of evidence-based targeted therapies and associated diagnostics."
>
> "Disease treatment and other interventions are better targeted to take into account the patient's genomic and other biological characteristics, as well as health status, medications patients are already prescribed and environmental and lifestyle factors." *UK Programme Coordination Group for precision medicine (11)*

According to the University of Edinburgh Precision Medicine Doctoral Training Programme's website, it "identifies disease endotypes that are diagnostically, prognostically, or mechanistically meaningful, with the aim of improving patient stratification and informing development of novel therapies" (12). In other words, precision medicine aims to provide the right treatment/prevention technique to the right patient at the right time. It is a new focus area of medicine where the aim is to move away from a "one size fits all" method to a patient specific approach. This method aims to combine all aspects of science such as multi-omics (genomics, transcriptomics, proteomics), imaging, and clinical data to develop a tailored, individualised treatment plan.

As it is an emerging field, there's no established clinical practice and infrastructure, yet. In the UK, recently, the government has shown efforts to incorporate routine genome sequencing to aid patient diagnosis and treatment (13). As precision medicine approaches become more widely available and used, clinicians need to be provided the right support, framework and guidance as genomic data often raises ethical dilemmas, which require clear and understandable communication to patients.

Therefore, it's important for all parties – researchers, clinicians, government - to work together in developing this new approach for effective patient wellbeing. For heterogenous diseases, such as cancer, the 'one-size-fits-all' treatment method is less efficient in patients with rarer disease phenotypes or unique healthcare needs. An example of 'one-size-fits-all' would be the use of chemotherapy in a wide range of cancers that is not tailored to a biomarker or a genomic feature of the patient's specific cancer type (14). Precision medicine can leverage unique, more focused – multiomic approaches – to design an individual, patient specific treatment (15).

In terms of impact on clinicians, as well as having a bigger workload to combine and understand all available data, they would also have a better understanding of patient needs, which advises them to provide a more suitable treatment plan for patients. The abundance of additional data would also aid clinicians in the decision-making process. Perhaps the most notable study of the usage of "non-conventional" approaches in a clinical setting was when genotype information was used as a guideline to help determine the correct dose of warfarin (16). One could argue that "phenotypic"-guided (protein biomarker guided) cancer treatment is a type of precision medicine and;

therefore, precision medicine is not ‘new’. For example, breast cancer patients that are estrogen receptor positive can be stratified for tamoxifen treatment and cervical cancer can be detected using protein biomarker in cells from cervical smears (17). Nevertheless, genotype-guided treatment seems to be the most well-studied impact of precision medicine. As mentioned before, highly heterogeneous diseases, such as cancer, can benefit the most from additional data collection from patients under the precision medicine umbrella, as genomic profiling of tumours can advise targeted therapy for patients with breast and lung cancer (18).

To summarise, precision medicine is an emerging field of patient treatment approach where patient specificity, guided by ‘omics’ platforms, is the priority as opposed to a ‘one-size-fits-all’ approach. The impact can span across all aspects of patient care such as patient, clinician, and research. Possible benefits can include: (i) improved diagnosis, (ii) predict susceptibility to conventional treatment, (iii) pre-empt disease progression, (iv) tailored treatment, which could reduce the cost, time, and failure rate of treatment. Current examples of using genomics to personalize treatment is for example (i) in the use of epidermal growth factor receptor (EGFR) mutations to stratify the use of kinase inhibitors that are mutant kinase-specific ; (ii) the use of kinase inhibitors stratified by patients that have the BCR-ABL genetic fusion (19); and (iii) the emerging use of cancer vaccines that require mutated neoantigens be detected in patients’ cancer using genome sequencing, RNA sequencing and/or mass spectrometry (20). Rather than defaulting to conventional targets, I explored how medicinal plant microRNAs could be used vulnerabilities in glioblastoma, while examining other aspects of conventional research practices such as the use of cell lines. The coming chapters will interrogate whether this approach can reconcile two seemingly contradictory needs: the scalability of dietary interventions and the patient-specificity demanded by GBM’s notorious heterogeneity.

1.2. What is Cancer?

Cancer is an extensive group of diseases that can originate in nearly any organ or tissue of the body. It occurs when cells become abnormal by undergoing uncontrolled growth and failure of the immune system to rectify the situation, surpassing their usual boundaries to invade neighbouring body parts and/or spread to other organs (21, 22). This spread is known as metastasis and is a significant contributor to mortality. Alternative terms for cancer include neoplasm and malignant tumour (21). Cancer ranks as the second leading cause of death, according to the WHO, contributing to an estimated 9.6 million or 1 in 6 deaths in 2018 worldwide (21). The most common types of cancer in men are lung, prostate, colorectal, stomach and liver; while in women it’s breast, colorectal, lung, cervical and thyroid (21).

The average new cases were 438.7 per 100,000 men and women per year; the death rate was 149.4 per 100,000 per year in the US (data age adjusted and based on 2016-2020 cases and deaths) (5). Close to 2 million people were estimated to have gotten cancer in 2023 alone in the US with no sign of a decline (Figure 1.1) (5). Between 2013 and 2019 the estimated 5-year relative survival was 68.7%, a statistic that is hard to calculate as patients cause of death can be impacted by other factors. Unfortunately, cancer has been a steady player in medicine for centuries, with only limited number of treatments available such as surgery, chemotherapy, and radiotherapy – whose effects vary from patient to patient. As cancer stays on the radar,

governments, hospitals, and patients themselves spend tremendous amount of money on patient care.

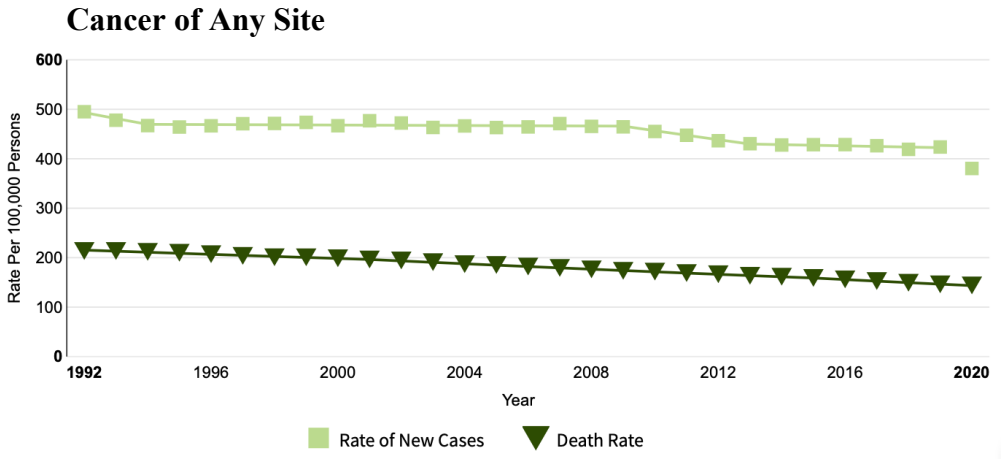


Figure 1.1: Cancer Statistics of Any Site

Rate of new cases and deaths in the Unites States between 1992 and 2020 (5)

A study conducted in 2016 examined patient costs for four cancer types (colorectal, breast, prostate, and lung) in England and found that in 2010 they cost 3% of the total cost of hospital care or £1.5 billion. Thinking about all other cancer types and expanding this knowledge to the rest of the world, billions and billions of pounds are spent each year on just hospital care due to cancer worldwide. In addition, increased delay in diagnosis and treatment, due to the COVID-19 pandemic, the occurrence of preventable cancer deaths has increased dramatically (23, 24).

Therefore, it is vital to conduct further research to understand the mechanisms of cancer, find new genetic markers and drug targets, and develop preventative procedures for early diagnosis. In this chapter, I will highlight key mechanisms, genetic markers, and currently available treatment procedures in the field of cancer research.

1.2.1. Hallmarks of cancer

Cancer cells exhibit faults in the regulatory mechanisms that dictate their rate of division as well as feedback loops that control these regulatory mechanisms (homeostasis). Normally, healthy cells grow and divide in a tightly controlled fashion. Growth only occurs when growth factors are activated, and division occurs only within the bounds of their environment that is supplied by sufficient blood flow. Otherwise, programmed cell death is triggered (apoptosis). This tightly controlled mechanism needs to be disrupted in order to turn a cell from normal into a cancerous one. Each controlling mechanism is supported by a number of proteins. A mechanism is considered disrupted if a critical protein malfunctions causing the signalling pathway to fail. Proteins can become non-functional or malfunctioning when their DNA, RNA or amino acid sequence is damaged through somatic mutations (these are non-inherited mutations but rather acquired later). This can occur in a series of steps, which were outlined by Hanahan and Weinberg as the ‘hallmarks of cancer’ (25). These initially included: (a) sustained proliferative signalling, (b) evading growth suppressors; (c) activating invasion & metastasis, (d) enabling replicative immortality, (e) inducing angiogenesis, and (f) resisting cell death; then later expanded with: (g) deregulated

cellular energetics, (h) avoiding immune destruction, (i) tumour-promoting inflammation, and (j) genome instability & mutation (Figure 1.2) (25, 26).

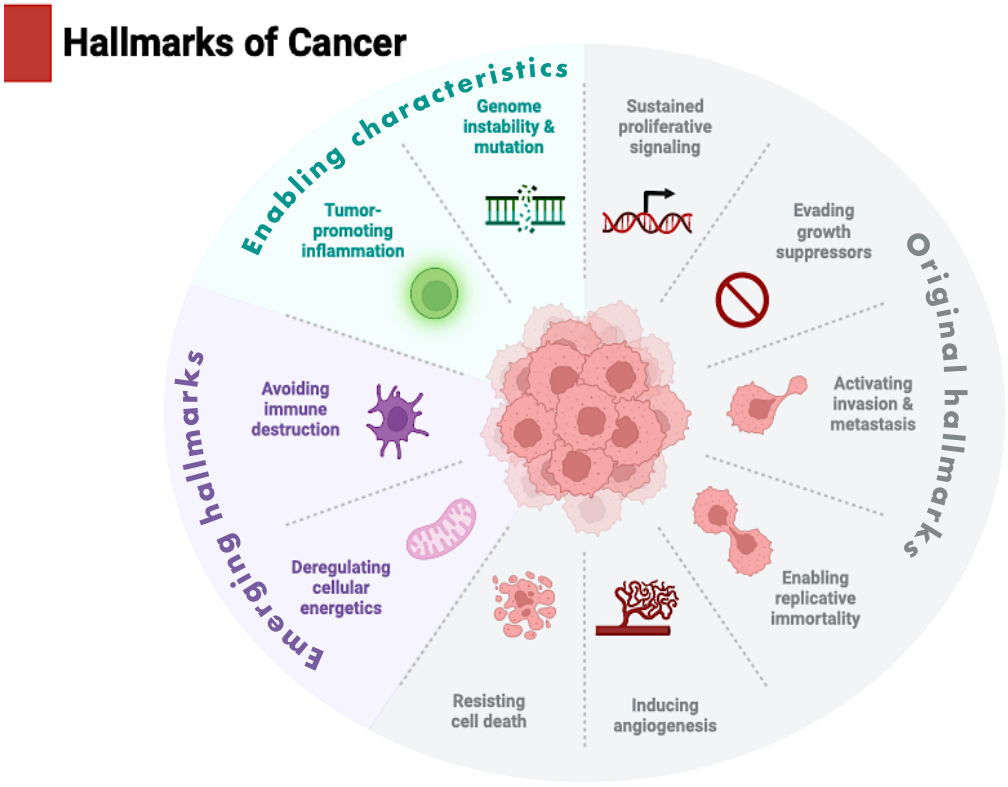


Figure 1.2: Hallmarks of cancer Image created with BioRender.com

Sustained Proliferative Signalling

It refers to a cells natural requirement for hormones and other molecules to stimulate growth and division; however, cancer cells acquire the ability to grow and divide without depending on such external signals (25, 26). Cancer cells can obtain the ability to continuously proliferate in a couple of ways. They can produce their own growth factors themselves, which then binds to their own surface, leading to self-stimulation of proliferation or autocrine proliferative stimulation (26). Alternatively, cancer cells can prompt neighbouring normal cells in their surroundings to produce growth factors, which in turn nourish the cancer cells. Additionally, cancer cells can alter the structure or the amount of growth factor receptors on their surfaces, resulting in an oversensitivity to even the smallest amount of growth signal (26). Furthermore, the constant activation of downstream signalling pathways can completely illuminate the need of growth factors, resulting in cancer cell independence from external growth factors.

Evading Growth Suppressors

To maintain the capability of constant growth, cancer cells must also acquire the capability to evade powerful mechanism that negatively regulate proliferation, mostly driven by tumour suppressor genes. The two most notable tumour suppressors have been identified as the retinoblastoma-associated (RB) and TP53 proteins; they play

central roles in two crucial cell signalling cascades responsible for cell division and activating processes like senescence or apoptosis (25, 26). RB protein integrates signals from both intracellular and extracellular signals to determine whether a cell should enter the growth-and-division-cycle (26). RB protein malfunction, therefore, causes a cell to lack this evaluation step and proliferate continuously. In contrast, TP53 responds to only internal signals, halting the cell cycle under unfavourable conditions or triggering apoptosis in severe cases of cell damage, though TP53 effects are highly dependent on cell type and context (26, 27). Although RB and TP53 have been identified as key regulators in tumorigenesis, their malfunction does not necessarily mean definitive tumour development. This is due to the large number of other regulative pathways with overlapping purpose that are present in cells to combat such malfunctions.

Activating Invasion & Metastasis

The process of invasion and metastasis was conceptualized as a series of distinct steps, termed the invasion-metastasis cascade (28, 29). This sequence starts with (i) local invasion, then (ii) intravasation by cancer cells into nearby blood and/or lymphatic vessels, (iii) transit through these systems, (iv) extravasation into distant tissues, (v) formation of micrometastases and (vi) eventual growth into macroscopic tumours, termed colonization. For example, mutant TP53 facilitates metastasis in epithelial carcinomas via the epithelial-mesenchymal transition (EMT), which is a key player in metastasis impacting various stages such as malignant conversion, extracellular matrix (ECM) degradation, primary tumour invasion, intravasation and extravasation. Key transcription factors (TFs) relating to EMT, such as basic helix-loop-helix TFs, zinc-finger E-box-binding (ZEB), and SNAIL, drive these cellular changes (30).

Enabling Replicative Immortality

The ability to replicate unlimitedly was established very early on in the cancer research field as one of the main features of a cancer cell in order to amass macroscopic tumours. In contrast, normal cells in the body are only allowed through the growth-and-division cycle a limited number of times. This transition has been termed ‘immortalization’, a trait that cancerous cells acquire to proliferate without any signs of senescence or crisis that would prompt for apoptosis. Telomeres, made up of multiple tandem hexanucleotide repeats, have been suggested as one of the key players in immortalization, where its shortening causes the loss of DNA sequence protection (25, 26, 31).

Inducing Angiogenesis

Angiogenesis the growth of blood vessels from existing vasculature (32). More so than normal cells, cancer cells require the rapid development of blood vessels due to their increased proliferation rate. For example, the VEGFA gene encodes ligands that are responsible for angiogenesis during embryonic and postnatal development (26). The VEGF gene family is regulated at multiple levels and can be regulated by both hypoxia and oncogene signalling (33).

Resisting Cell Death

Apoptosis or programmed cell death is the most radical mechanism for eliminating faulty cells. The apoptotic machinery is comprised of two major circuits: one receiving external death-inducing signals, and one responding to internal triggers. Both circuits

result in activating a normally latent protease (caspase 8 and 9, respectively), which are responsible to activate the downstream execution phase of apoptosis, where the cell is disassembled and consumed by its neighbours and phagocytic cells (25, 26). S is a hallmark of cancer cells. Recent studies suggest that apoptotic cells within high-grade tumours may contribute to tumour survival, challenging conventional views on the role of apoptosis in cancer therapy (34). These dual effects of apoptosis in cancer poses critical questions about the efficacy of inducing cell death as a therapeutic approach. Therapy-induced apoptosis may trigger responses leading to therapy failure or aggressive tumour evolution. Addressing the imbalance between cell birth and death in tumours is crucial for understanding how apoptotic cell death influences cancer progression or suppression, prompting a need for focused research on molecular mechanisms within the tumour microenvironment (TME).

Deregulated Cellular Energetics

Cancer cells exhibit altered energy metabolism characterized by a preference for glycolysis even in the presence of oxygen, a phenomenon termed "aerobic glycolysis". This metabolic switch is driven by mutations in oncogenes and tumour suppressors, facilitating the acquisition of hallmark capabilities such as continuous proliferation and evasion of cell death. Glycolysis also supports biosynthetic pathways necessary for cell growth and division. Some tumours contain subpopulations of cells with different energy-generating pathways, displaying symbiotic behaviour by exchanging lactate as fuel. This altered energy metabolism, orchestrated by proteins involved in cancer hallmarks, suggests it as an emerging hallmark of cancer (26).

Avoiding Immune Destruction

The role of the immune system in preventing or eliminating tumour formation remains a complex and unresolved issue. While immune surveillance theory suggests that the immune system can recognize and eliminate incipient cancer cells, the increased incidence of certain cancers in immunocompromised individuals supports this notion. Experimental studies in mice deficient in immune components show increased tumour susceptibility, indicating the role of both innate and adaptive immunity in tumour eradication. Clinical evidence also supports antitumoral immune responses in certain human cancers. However, cancer cells can evade immune destruction through various mechanisms, suggesting ‘immunoevasion’ as an emerging hallmark of cancer (26).

Tumour-Promoting Inflammation

Pathologists have observed that tumours often exhibit dense infiltration by immune cells, resembling inflammatory conditions in non-neoplastic tissues. Advanced markers reveal immune cell presence in virtually all neoplastic lesions, ranging from subtle infiltrations to noticeable inflammations. Initially viewed as attempts to eradicate tumours, immune responses have been found to paradoxically enhance tumorigenesis and progression. Research on inflammation and cancer has shown immune cells, especially from the innate immune system, promoting neoplastic progression by supplying bioactive molecules that sustain proliferation, limit cell death, and facilitate angiogenesis and metastasis. Inflammation, present even at early neoplastic stages, can foster cancer development and accelerate genetic evolution towards malignancy, making it an enabling characteristic in acquiring hallmark capabilities (26).

Genome Instability & Mutation

Cancer progression involves a succession of genetic alterations that confer selective advantages to subclones of neoplastic cells, leading to clonal expansions. These alterations can result from mutations or epigenetic changes, such as DNA methylation and histone modifications, affecting gene expression regulation. Dysfunctions in the DNA maintenance machinery, including those affecting DNA repair and surveillance systems, contribute to increased mutation rates in cancer cells, leading to genomic instability. Telomerase, previously known for its role in replicative potential, is now recognized as crucial for maintaining genome integrity. Advanced genomic analyses have revealed distinct mutation patterns across different tumour types, highlighting genome instability as a pervasive characteristic in cancer development. These defects in genome maintenance and repair mechanisms are selectively advantageous, accelerating tumour progression by facilitating the accumulation of favourable genotypes (26).

In conclusion, cancer is characterized by a multitude of dysregulated processes, termed the 'hallmarks of cancer', that enable the continuous growth and spread of malignant cells. These dysfunctions disrupt the finely tuned regulatory mechanisms governing cell division, apoptosis, and energy metabolism. From sustained proliferative signalling to evading immune destruction, each hallmark of cancer represents a critical step in tumorigenesis, driven by genetic mutations and alterations in cellular pathways. The emerging hallmarks, such as deregulated cellular energetics and immunoevasion, further underscore the complexity of cancer biology. Understanding these hallmarks not only sheds light on the molecular mechanisms driving cancer, but also holds promise for developing targeted therapies aimed at disrupting these aberrant processes and ultimately improving patient outcomes.

1.2.2. Cancer microenvironment

The concept of a tumour microenvironment (TME) was first introduced by Virchow in 1863, when he highlighted a relationship between inflammation and cancer, then in 1889 upon the formulation of Paget's "seed and soil" theory (35). Then, most notably, in 2011 Hannah and Weinberg recognised the emerging participation of TME in cancer development and progression in their 'hallmarks of cancer' proposal (26). TME comprises of many components such as non-malignant cells, vessels, nerves, intracellular components, and metabolites – all of which play critical roles in cancer development and progression. As the TME is a complex and dynamic space, there's a constant concern of its accurate characterisation. Therefore, the conventional understanding of TME has been broadened to include the tumour organismal environment (TOE), which also includes microenvironments distant from cancer lesions; but influencing their development. Additionally, TME has been divided into six layers, acknowledging the heterogeneity across different tumour locations: tumour cell to tumour-cell environment (TCTCE), niche, confined tumour environment (TE), proximal TE, peripheral TE, and TOE (36). TME formation is a complex interplay between oncogenic mutations, chronic inflammation, and wound-healing processes. It is heavily influenced by cancer cells to support their own survival, migration, and response to intrinsic and extrinsic factors. Specialized microenvironments within TME, such as the hypoxic niche and immune microenvironment, have emerged as crucial targets for cancer therapy, including immunotherapy and targeted drug therapy.

The hypoxic niche, characterized by low oxygen levels, influences various aspects of cancer biology, and is associated with poor prognosis in patients. It activates vascular

endothelial cells (upregulates VEGF transcription), promotes angiogenesis, and alters tumour metabolism, contributing to tumour progression and therapeutic resistance. In response to decreasing oxygen levels cells, rely on the increased activity of hypoxia-inducible factors (HIF) and HIF signalling to adapt to hypoxic conditions. The 15 top-ranked hypoxia associated genes have been named by Buffa et al. VEGFA, SLC2A1, PGAM1, ENO1, LDHA, TP11, P4HA1, MRPS1, CDKN3, ADM, NDRG1, TUBB6, ALDOA, MIF, and ACOT7, which are collectively considered as the hypoxia signature (Buffa signature) to assess the hypoxic state (37). The presence of hypoxia also results in increased heterogeneity between patients with the same tumour types, and increases the occurrence of somatic mutations in oncogenes and tumour suppressors such as TP53, PTEN and MYC (38, 39).

The immune microenvironment, made up of a diverse array of immune cells, plays a dual role in cancer progression. While it can exert antitumor effects, immune suppression within TME often enables cancer cells to evade the immune system and promote metastasis. Immune cells such as T cells, B cells, natural killer (NK) cells, tumour-associated macrophages (TAMs), myeloid-derived suppressor cells (MDSCs), mast cells, granulocytes, dendritic cells, tumour-associated neutrophils, cancer-associated fibroblasts, adipocytes, vascular endothelial cells, and pericytes play important roles in shaping the antitumour immune response and determining cancer outcomes.

Metastasis is responsible for 90% of cancer-related deaths in cases of solid tumours, and poses a challenge due to undetectable micrometastases and inadequate treatment responses (36). Advanced computation techniques, like deep-learning, has newly enabled metastasis analysis through identification and targeting (40). The TME is largely responsible for orchestrating the metastatic cascade, with immune cells playing vital roles. $CD4^+CD25^+$ $FOXP3^+$ regulatory T cells, $Ly6G^+$ neutrophils, MDSCs, and macrophages contribute to immunosuppressive pre-metastatic niches, while other immune cells, such as TH1 $CD4^+$ or $CD8^+$ T cells, $Ly6G^-$ neutrophils, and NK cells, exert antitumor effects (41). Brain metastases exhibit distinct immune landscapes compared to primary tumours, affecting treatment responses and clinical trial inclusion criteria. Understanding these differences enhances treatment strategies for advanced cancer patients, particularly those with brain metastases, emphasizing the need for further research and inclusive clinical trials (36).

Metabolic reprogramming (a hallmark of cancer), an alteration in metabolism or nutrient supply, includes increased glucose, lipid, glutamine, and amino acids, as well as lactate accumulation and ROS addiction. Cancer cells exhibit a preference for glycolysis and elevated lactate metabolism, even in normoxic conditions, instead of oxidate phosphorylation known as the Warburg effect or aerobic glycolysis (42). Lactate, once considered a metabolic by-product, now emerges as a key player in reprogramming cancer and stromal cells within the TME, promoting an immunosuppressive environment, angiogenesis, and most notably the survival of hypoxic cells. Glutamine metabolism also fuels cancer growth and stromal cell function through the production of energy, carbon, and nitrogen. Additionally, elevated ROS levels in cancer also impact tumorigenesis, tumour immunity, and TME reprogramming with cancer cells developing tolerance to ROS accumulation. The excessive proliferative rate results in long diffusion distances, where blood vessels are unable to keep up with the rapid expansion rate of the tumour leading to hypoxia and

upregulation of HIF expression. Under hypoxia, to enhance tumorigenesis, mitochondrial ROS are required for HIF stabilization (43). Dysregulation of ROS can also affect the regulation of other immune components in the TME such as MDSCs, TAMs, and T cells fuelling further tumorigenesis. Finally, lipid metabolism, including cholesterol and fatty acids, are components that promote cancer growth, recurrence and metastasis formation through post-translational modification of proteins, energy availability of cancer cells and generation of cancer cells membranes (44).

The acidic niche, result of dysregulated or reversed pH, promotes cancer cell survival, proliferation, migration, invasion, glycolysis and inhibits apoptosis. A cancer cell feature is to have slightly higher intracellular pH (~7.2 compared to 7.4) and lower extracellular pH (~7.4 compared to 6.7-7.1) (45). Acidic niche, hypoxic niche and the metabolic, particularly lactose, microenvironment are heavily intertwined because acidic niche is a product of either lactate metabolism or CO_2 hydration (36). Hypoxia is responsible, among other things, for tumour cell adaption to the acidic TME and lactate production. In addition, acidic niche generation is also driven by oncogene activation, such as Ras and Myc, and inactivation of tumour suppressors, like p53.

The hypoxic niche, prevalent throughout tumours and their surroundings, profoundly influences cancer cells and triggers a cascade of hypoxia-induced effects such as immune, lactate, ROS metabolism ME and acidic niche (36). A typical result of immune microenvironment change is hypoxia-induced VEGF expression, which promotes the ‘glycolytic switch’ and ‘metabolic symbiosis’ in cancer cells (46). It’s a phenomenon where oxidative cancer cells favour lactate over glucose utilization, mediated by MCT1, leading to glucose starvation and necrosis in hypoxic regions. It was further demonstrated that the hypoxia-lactate axis directly regulates gene expression through post-translational modification of histones, referred to as histone lactylation (47).

In summary, the concept of the tumour microenvironment (TME) has evolved significantly since its conception in the 19th century, with modern understanding highlighting its pivotal role in cancer development and progression. Comprising of diverse components like immune cells, vessels, nerves, and metabolites, the TME houses intricate interactions shaping cancer biology. Hypoxia, immune dysregulation, and metabolic reprogramming emerge as key drivers, influencing tumour behaviour and response to therapy. Specialized microenvironments within the TME, such as the hypoxic niche, acidic niche, and immune microenvironment, have emerged as critical therapeutic targets. Understanding these complexities is essential for developing effective treatments and improving outcomes for cancer patients, underscoring the importance of ongoing research and clinical innovation.

1.2.3. Glycolysis

To better understand the complexity of cancer cell development, many studies have found that glycolysis plays an important role in numerous cancer related processes such as tumorigenesis, metastasis, and chemoresistance, as well as resisting cell death (48). As part of the deregulated cellular energetics or metabolic reprogramming, lactate accumulation in the TME leads to reduced pH, part of the acidic niche, and immunosuppression in the TME. Briefly, glycolysis refers to the oxidative breakdown of glucose, forming lactate with a small amount of ATP in the absence of oxygen or under anaerobic conditions (Figure 1.3).

Normal, healthy cells produce energy by glucose conversion through aerobic oxidation, whereby glucose oxidizes to water and carbon dioxide. However, in the absence of oxygen glucose or glycogen is broken down to produce lactate and energy (Figure 1.3). This process is known as anaerobic respiration.

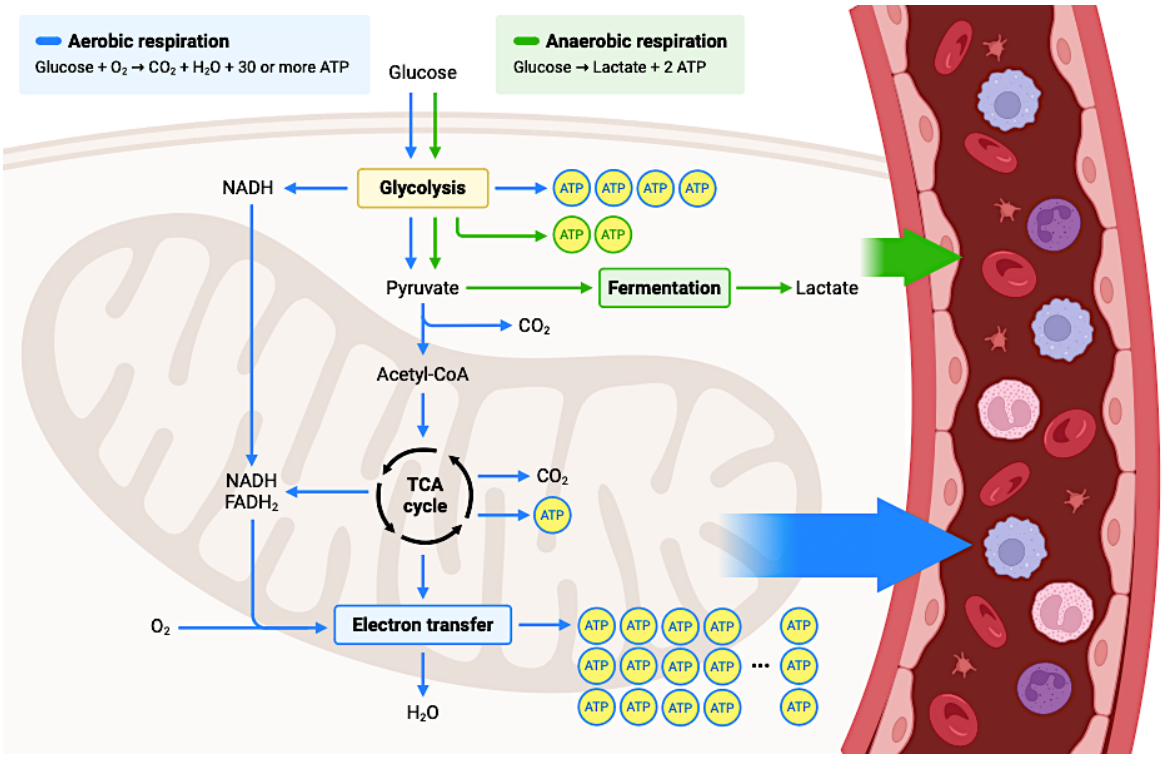


Figure 1.3: Comparison of aerobic and anaerobic respiration) in cells. Created with Biorender.com

Glycolysis exhibits three main characteristics. Firstly, it serves as the sole method to produce ATP under anaerobic condition or in cells lacking mitochondria, like erythrocytes and hyperthyroid cells, through pyruvate conversion to lactate (Figure 1.3). Secondly, under normal conditions (i.e. in the presence of oxygen), glycolysis produces pyruvate, which subsequently enters the tricarboxylic acid (TCA) cycle (also known as the citric acid or Krebs cycle) in the mitochondria leading to ATP synthesis. Thirdly, some components from the glycolytic and TCA cycle can participate in anabolic pathways, generating NADPH and intermediary molecules necessary for glycogen, lipids, nucleotides and protein syntheses (48, 49). In essence, glycolysis aims to provide intermediary molecules that activate biosynthetic pathways. When oxygen is available, mitochondria predominantly supply plenty of ATP to cells, but when there is an oxygen deficit, glycolysis becomes activated to generate ATP, ensuring cell survival (48, 49).

Aerobic glycolysis is further supported by the upregulation of various enzymes and transporters, including GLUT1, HK2, PFK1, PKM2, LDHA, and MCT1 (50). Inhibition or deficiency of these molecules has shown promise in inhibiting cancer cell proliferation, migration, and glycolysis. For instance, GLUT1 knockdown inhibits prostate and colorectal cancer growth (51). HK is an enzyme that converts glucose to glucose-6-phosphate, thereby limiting the glycolytic reaction very early on. HK2, an isoform of HK, was shown to suppress glycolysis and tumour growth in glioblastoma by separating from mitochondria due to elevated glucose levels, which phosphorylates

IκBα, causing its elimination, thereby transcriptionally enhancing PDL1, which promotes tumour cell immune evasion and brain tumorigenesis (48, 52). The absence of HK2 negatively regulates glycolysis and tumour cell growth in liver cancer, while in glioblastoma HK2 activity inhibition promotes cell death (53, 54). PFK1, is another rate-limiting enzyme in glycolysis, and its deficiency promotes cell death in inhibited cell migration in rectal cancer, proving to a potential new target for other cancer types (55). PKM2, a mediator of tumour metabolism, LDHA, glycolysis inhibitor, and MCT1, regulates tumour cell survival through proton-linked transmembrane lactate transport, are have been shown to be promising therapy targets in cancer (48).

In addition, there are many other therapeutic targets for cancer. For instance, mTOR, a major regulatory component of the Warburg effect, HIFα, a downstream participant in the mTOR pathway, and PKM2, a tumorigeneses promoter, are part of a complex signalling cascade that regulate tumour progression. Thus, all of them can be therapeutic targets (56).

In summary, TME-influenced glycolysis is closely intertwined with numerous participants of the immune system. Recent research highlights tumour glycolytic activity as a potential biomarker for predicting responses to immunotherapy. Targeting aerobic glycolysis has thus emerged as a key focus in cancer treatment development. Tumour cells heavily rely on glycolysis, involving many metabolic enzymes, suggesting the development of drugs that specifically target these enzymes to inhibit tumour growth while minimizing harm to normal tissues. Combining targeted glycolytic therapy with immunotherapy and chemotherapy holds promise for enhancing cancer treatment outcomes.

1.2.4. Genetic properties

Carcinogenesis is a highly complex process, involving various levels of regulation. Our understanding of cancer is deeply connected with our knowledge of genetics. Milestones over the past 200 years such as Mendelian genetic inheritance, discovery of chromosomes, and DNA structure in conjunction with the development of sequencing have allowed the identification of the first cancer gene in avian species 1976: SRC (57, 58, 59, 60, 61). Oncogenes, such as HRAS, were identified in humans soon after, in 1982 (62, 63). A second type of cancer gene, tumour suppressor, was also identified in the mid-1980, shining a light on the complexity of the genetic basis of tumorigenesis (64). By 2004, close to 300 oncogenes have been identified, that were altered by either point mutations, translocations or copy-number variations, affecting cancer hallmarks (65). By the 2010s, established next-generation sequencing methods have identified hundreds of mutated genes assumed to participate in tumorigenesis at one point during formation; however, it was also established that only a few of these mutational events affect driver genes, which are responsible for the beginning stages of malignization (66).

Tumours are highly heterogeneous both between and within tumour types. A compendium found that around the majority of driver genes (~360) are only present on one or two tumour types; however, a select few (~10) are capable of driving mutations in over 20 different malignancies, making them significant cancer-wide driver genes, like : KRAS, TP53, PTEN, and PIK3CA, to name a couple (66). More specifically, KRAS (v-Ki-ras2 Kirsten rat sarcoma 2 viral oncogene homolog) is a small GTPase that switches between inactive (GDP-bound) and active state (GTP-bound)

to aid cellular response to external stimuli through the PI3K and MEK/ERK pathways (67). Activating mutations cause a continuously active state which promotes cancer hallmarks like cell replicative immortality, angiogenesis and proliferative signalling (68, 69). TP53 (tumour protein 53) is probably the most widely known tumour related protein; it was named Molecule of the Year in 1993, it is a transcription factor that manages cell responses to a variety of stress inducing factors such as DNA damage, abnormal growth signalling, hypoxia, and various drugs or environmental factors like ultraviolet light (70, 71). TP53 malfunction induces cellular senescence and suppresses most of the other cancer hallmarks like cell death, and inflammatory response (69, 72). PTEN (phosphatase and tensin homolog gene) is a lipid phosphatase that also regulates the PI3K/AKT pathway, which controls cell and tumour fate in terms of cancer hallmarks such as elevated invasion & metastasis and increased genomic mutation rates (73).

On the other hand, some mutations are highly specific to particular cancer types; for instance, MYC126 and cyclin D3 (CCDN3) mutations are prevalent in Burkitt lymphomas, with 60% and 47% of cases affected, respectively (74, 75). In uveal melanoma, half of the cases exhibit activating mutations in specific hotspots of the guanine nucleotide-binding protein G_q subunit-α (GNAQ) gene, while nearly all others show mutations in one of two homologous hotspots of its paralogue, GNA11 (66).

Furthermore, the mutational characteristics of driver genes were suggested to be crucial for their function in tumorigenesis. For example, PTEN, a tumour suppressor, exhibits an abundance of both nonsense and missense mutations in glioblastomas (76). Nonsense mutations trigger nonsense-mediated decay, reducing functional PTEN protein production, while missense mutations hinder enzymatic activity or membrane recruitment, disrupting its role in regulating cell functions like the cell cycle, apoptosis and protein synthesis (66). In contrast, missense mutations in EGFR exhibit different tumorigenic effects across different cancer types. In glioblastomas, they cluster in the extracellular domains, promoting receptor activation independent of ligands. In lung adenocarcinomas, mutations occur in the tyrosine kinase domain, leading to increased receptor activity while decreasing ATP affinity. Overall, mutations tend to concentrate in specific domains of protein products across various genes and tumour types. The p53 DNA-binding domain stands out as particularly enriched for somatic mutations across 43 different cancer types, surpassing other protein domains. Although, this enrichment is primarily due to TP53 (66).

In addition to malfunctioning internal molecular pathways, environmental factors can also damage the DNA makeup of crucial genes leading to mutations. Environmental chemicals have numerous abilities like inducing gene mutations, acting as tumour promoters, and enhancing cell proliferation via mutated oncogenes. Furthermore, chemical exposure can exacerbate mutation causing endogenous pathways, such as reactive oxygen species and impaired DNA repair, leading to cancer development.

Despite the complex underlying mechanism of cancer development, numerous driver genes have been identified and successfully targeted that led to improved disease management and treatment. For instance, the development of tamoxifen, an estrogen receptor antagonist, and herceptin, which attracts immune cells to target cancer cells with HER2 gene overexpression, has resulted in decreased mortality in breast cancer patients (77, 78, 79). Although, the same improvement cannot be said about other cancer types like glioblastoma (Figure 1.4).

In conclusion, the intricate process of carcinogenesis underscores the profound interplay between genetics and cancer. From foundational discoveries like Mendelian inheritance to the identification of oncogenes and tumour suppressors, our understanding of cancer genetics has evolved significantly over the past two centuries. The discovery of driver genes like KRAS, TP53, and PTEN has shed light on the complexity of tumorigenesis, with certain mutations exhibiting widespread impact across multiple malignancies. Moreover, the specificity of mutations to particular cancer types highlights the heterogeneous nature of tumours. Understanding the mutational characteristics of driver genes has proven crucial in elucidating their role in tumorigenesis, offering insights into potential therapeutic targets. Despite the complex interplay of genetic and environmental factors in cancer development, targeted therapies have shown promise in improving patient outcomes in certain cancers, though challenges persist in others. Continued research and advancements in cancer genetics hold the key to further enhancing disease management and treatment strategies in the future.

1.3. Glioblastoma

In this section, glioblastoma multiforme will be introduced in more detail to highlight the necessity and impact of further research on patient wellbeing, disease diagnosis and treatment.

1.3.1. Clinical and genetic features

Glioblastoma is the most common and aggressive primary brain tumour in adults. Based on the most recent update of the World Health Organization (WHO) in 2021, it has the highest grade classification (grade 4) of brain tumours (80, 81). Glioblastoma can arise anywhere in the central nervous system (CNS); however, it most commonly forms in the frontal or temporal lobes of the brain (82). Its unique origin in the CNS distinguishes it from other, more common, secondary cancers that form as a result of metastasis from distant primary locations such as lung, breast or skin (81, 83). Glioblastoma belongs to a group of tumours that are characteristically heterogeneous called gliomas, which are believed to originate from glial cells or their precursors, and include astrocytomas and oligodendrogliomas (84). Astrocytomas are categorised into two grade groups: grade 1 when localised, and grade 2-4 when diffused, where the increasing grade level reflects an increasingly aggressive tumour phenotype (80). Glioblastoma, or sometimes referred to as grade 4 astrocytoma tumour, is a highly malignant tumour with high rates of cell division, vascular proliferation and, therefore, high level of necrosis in the tumour's central areas (80, 81, 85).

In 2015, the age-standardized incidence rate of glioblastoma in Europe was 5.02 per 100,000 people for glioblastoma. The same figure for Canada and the United States were 4.5 and 4.32, respectively (86). These numbers show a steadily increasing trend since 1995 (from 3.56 and 3.92, respectively), however, it isn't clear why – most likely better diagnostic and reporting practices. In the 2021 update of the WHO, the criteria to categorize astrocytomas and glioblastomas has changed; therefore, some of the following survival statistics will be based on the previous system from 2016 (80). For diffuse astrocytomas (grade 2), 45% of patients survive for more than 5 years, for anaplastic astrocytoma (grade 3), only more than 20% of patients survive for more than 5 years and for glioblastoma (grade 4), this figure drops to only 5% (87). In the last 10 years, roughly 2000 people die every year in England and Wales combined

(88). In the US, the rate of new cases and deaths have been stagnating since 1992, with no hope for change in sight (Figure 1.4).

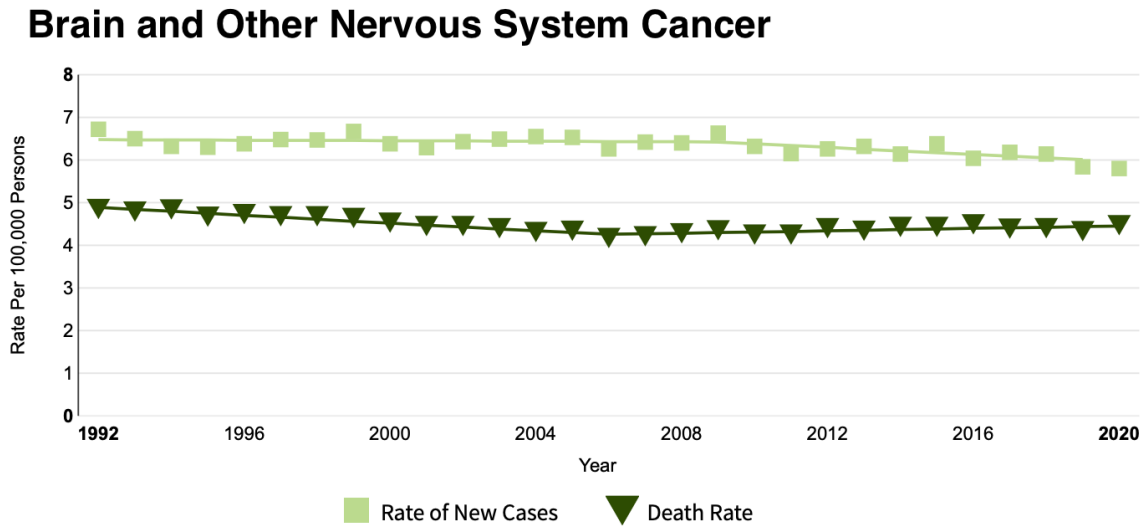


Figure 1.4: Brain and other Nervous System cancer statistics in US between 1992 and 2020 (4)

Brain tumours remain a challenging frontier in medicine, with limited effective treatment options, primarily relying on surgery (81). Despite comprising only 2% of all primary cancers, they contribute to 7% of cancer-related deaths in individuals under 70 years old, and survival rates have seen minimal improvement since 1975 (89). While slight increases in the 5-Year Relative Survival rate may be attributed to advancements in healthcare and surgical techniques, these improvements are not specific to addressing the complexities of combating brain tumours (90).

Types

There are over 100 different types of brain tumours, typically named after the cell they develop from. Clinicians consider several factors to classify brain tumours including (i) the type of cell they develop from, (ii) the tumour's location in the brain, (iii) genetic characteristics, (iv) whether it mostly develops in children or adults, and (v) the grade of the tumour. Some genetic characteristics have been identified that aid clinicians and scientist to diagnose and categorise tumours such as IDH mutations, 1p/19q codeletion, and TERT promoter mutation, etc. Understanding the different types of brain tumours can be very overwhelming. So, while introducing some tumour types, I'd like to focus on the development of glioblastoma specifically, as it was the tumour of interest in my project.

Glioma the most common type of cancerous (malignant) brain tumour in adults that develop from glial cells, which play a supporting role in the brain and spinal cord. There are three types of gliomas: astrocytomas (grades 1-4), oligodendrogliomas (grades 2-3) and glioblastomas (also called grade 4 astrocytoma) (Table 1.1). In addition, there are also four other types in the table below to highlight the versatility of brain tumours (Table 1.1).

Astrocytoma is the most frequently occurring type of glioma. It develops from a type of glial cell called astrocyte. Its main genetic biomarker is IDH-mutation, which is

followed by TP53 and ATRX mutations. There are three subtypes grade 2 (diffuse), grade 3 (anaplastic) and grade 4 (glioblastoma) (Table 1.1).

Oligodendroglioma is a rare type of glioma that develop from glial cells called oligodendrocytes. They can be categorised as low-grade or high-grade depending on the speed of tumour growth. They can be referred to as oligodendroglioma, IDH mutant and 1p/19q co-deleted (91, 92) (Table 1.1).

Glioblastoma is a grade 4 astrocytoma; however, its genetic biomarker deviates to IDH-wild-type (Table 1.1).

IDH
Isocitrate dehydrogenase is an enzyme with three isoforms (IDH1, IDH2 and IDH3), that participate is a number of major metabolic processes like the Krebs cycle, glutamine metabolism, lipogenesis and redox regulation (93). In a number of human malignancies such as gliomas and acute myeloid leukemia, the genes encoding IDHs are frequently mutated, which is believed to be an early event in gliomagenesis (92, 93). The mutant IDH genes alter downstream IDH enzymes causing them to repress DNA methylation leading to genome-wide DNA hypermethylation. These are more commonly found in low grade gliomas, denoted with “IDH-mutant”, and tend to have better prognosis than IDH-wild-type (92, 94).

1p/19q
The complete deletion of both the short arm of chromosome 1 (1p) and the long arm of chromosome 19 (19q) (hence the name: 1p/19q co-deletion) is a genetic signature of oligodendrogliomas, a subtype of gliomas. This molecular change causes an unbalanced whole-arm translocation between chromosomes 1 and 19, with the loss of the derivative t(1p;19q), that occurs early in the pathogenesis of oligodendrogliomas. Patients exhibiting this genetic signature tend to have better prognosis as these diffuse gliomas have a favourable response to alkylating chemotherapy (95).

TERT promoter mutation
Telomeres, which are additional non-coding DNA sequences found at the ends of chromosomes, serve the crucial function of safeguarding the coding DNA from shortening. Telomerase, a ribonucleoprotein complex, plays a vital role in preserving telomeres. Deregulation of telomere maintenance is a common occurrence in 90% of advanced cancers. Telomerase reverse transcriptase (TERT), a catalytic subunit of the telomerase complex, is frequently associated with poor patient prognosis across various tumour types, including breast cancer, sarcomas, and brain tumours.

Glioma Type	Location	Grade	Characteristics	Genetic Markers
Astrocytoma	Throughout the brain and spinal cord (Diffuse)	Grade 1	Well-differentiated, slow-growing, often cured by surgical resection.	• IDH-Mutant, • TP53 mutation, • CDKN2A/B homozygous deletion absent
		Grade 2	Well-differentiated, lacks anaplasia, low or no mitotic activity. Slower-growing, may progress to higher grades.	
		Grade 3	Focal/dispersed anaplasia, significant mitotic activity. Faster-growing, more aggressive.	
		Grade 4 (Also known as Glioblastoma)	Microvascular proliferation or necrosis or CDKN2A/B homozygous deletion or any combination. Fast-growing, highly aggressive.	• IDH-Mutant, • TP53 mutation, • CDKN2A/B homozygous deletion • CDK4 amplification • RB1 mutation/homozygous deletion • PIK3A/PIK3R1 mutation
Oligodendroglioma	Cerebral hemispheres, often in frontal lobes (Diffuse)	Grade 2	Arises from oligodendrocytes. Slow-growing, tends to have better prognosis	• IDH mutation, 1p/19q codeletion
Anaplastic Oligodendroglioma		Grade 3	Arises from oligodendrocytes. More aggressive, tends recur.	• IDH mutation, 1p/19q codeletion, • CDKN2A homozygous deletion
Glioblastoma (Glioblastoma multiforme or GBM)	Cerebral hemispheres, but can occur anywhere	Grade 4	Presents microvascular proliferation, necrosis, cellular atypia or specific genetic alterations. Highly aggressive, fast-growing, most common, and malignant.	• IDH-Wild-Type • H3-wild-type And one or more of the following: • TERT promoter mutation • EGFR gene amplification • +7/−10 chromosome copy-number alterations • PTEN mutation
Ependymoma	Ventricles of the brain and the spinal cord.	Grade 1	Well-differentiated, slow-growing.	RELA fusion (in some posterior fossa ependymomas)
		Grade 2	More cellular, moderate growth potential.	
		Grade 3	Anaplastic, faster-growing, higher recurrence risk.	
Mixed Gliomas	Various locations	Grades 2-3	Variable characteristics depending on the dominant component. May have features of both astrocytoma and oligodendroglioma.	• IDH mutation • 1p/19q codeletion (in oligodendroglioma component)
Pilocytic Astrocytoma	Mostly in the cerebellum also in optic nerve.	Grade 1	Well-circumscribed, often cystic with a mural nodule. Typically, slow-growing, often occurs in children.	BRAF mutation (commonly V600E)
Subependymoma	Near ventricles, particularly the fourth ventricle.	Grade 1	Slow-growing, often asymptomatic. Well-differentiated with low proliferative potential.	Variable, often diploid.

Table 1.1: Incomplete list of different types of Glioma tumours.

Symptoms

A considerable limiting factor for successful treatment is the time of diagnosis, which is largely due to non-specific or general symptoms such as headache, fatigue, and nausea. Around half of brain tumour diagnoses occur after emergency hospital visits, even though patients had previously sought help from their general practitioners (GPs) for symptoms (96). Only 2% of patients in England are diagnosed through the "suspected cancer" pathway, which grants general practitioners (GPs) direct access to magnetic resonance imaging (MRI) scans within two weeks (96, 97). The challenge lies in identifying early brain tumour symptoms in primary care, given their similarity to common benign conditions. While a successful national campaign for children (HeadSmart) has shortened diagnosis times, there's currently no equivalent strategy for adults (98). Headaches, the most common early symptom, result in only a small percentage of GP consultations leading to a brain tumour diagnosis. Additionally, headache characteristics vary based on tumour factors, and the likelihood of an underlying brain tumour significantly increases when additional neurological symptoms are present. Seizures are observed in approximately 20% of glioblastoma patients initially, with an additional 20% developing seizures later in the disease (99, 100). Although less common in primary care, new onset seizures in adulthood have the highest positive predictive value (PPV) among individual symptoms (1.6%), followed by motor weakness (1.5%) and confusion (1.4%) (101). Various other symptoms associated with brain tumours have a PPV of less than 1% (101). However, the combination of symptoms, particularly if they are progressive, significantly increases the likelihood of identifying an intracranial tumour on MRI. Additional symptoms reported to GPs 6 months before brain tumour diagnosis by a small percentage (< 5%) of patients are vertigo, anxiety, depression, sleep problems, poor concentration, dizziness, altered sensation, visual impairment.

Causes and risk factors

Due the presence of overwhelmingly general symptoms leading to late diagnosis, causes and risk factors of the disease have been difficult the identify. Therefore, most patients do not have identifiable risk factors. Nevertheless, limited resources draw the conclusion that a small number of primary brain tumours are due to genetic predisposition syndromes, but this only equates to less than 5%. The only uncontested cause of the disease is ionising radiation (102); however, no other environmental factors have been unquestionably linked to glioblastoma development. Interestingly, large cell phone usage studies have been conducted to determine the causative effect of radiofrequency electromagnetic fields (RF-EMF) on brain tumours. As a result, some evidence from these studies might suggest that long term exposure of RF-EMF (>10 years) could be a risk factor for developing glioma. Therefore, the WHO classified cell phone use and RF-EMF as “possibly carcinogenic to human beings” in 2011, then in 2015, upgraded to “probably carcinogenic” (103, 104). However, this has remained highly controversial, so much so that the WHO has launched an investigation into the validity of such papers, as there are papers sighting no effect (105, 106).

Prognosis and patient experience

In terms of prognosis, patients under 70 years old with glioblastoma exhibit median survival (without treatment) of approximately 3-4.5 months. An extended median survival of 8-10 months is possible via biopsy followed by chemotherapy (with or without radiotherapy). Furthermore, maximal treatment with debulking surgery followed by chemoradiotherapy improves median survival to approximately 15-16

months, with associated survival rates of 27-31% at 2 years and 7-10% at 5 years (81). Elderly patients receiving best supportive care alone have a median survival of less than 4 months. For patients over 65 years old, hypo-fractionated radiotherapy and chemotherapy following biopsy or resection yield better median survival of 7-9 months compared to radiation alone, although the addition of adjuvant chemotherapy does not improve quality of life. Patients with glioblastoma experience typical trajectories of physical, social, psychological, and existential distress, with existential distress being acute at specific stages: around diagnosis, after initial treatment, at disease progression, and at the end of life. Understanding these illness experiences allows for tailored support and communication throughout the disease course, with early provision of a palliative care approach significantly improving quality of life and death for patients and their carers (81).

1.3.2. Microenvironment

The glioma microenvironment is an intricate and highly dynamic landscape that plays a fundamental role in shaping tumour growth. Unlike many other solid tumours, gliomas arise within the central nervous system (CNS), a uniquely specialized and tightly regulated environment where interactions between tumour cells and their surroundings – including the extracellular matrix (ECM), vasculature, immune components and metabolic landscape – ultimately define disease trajectory. The microenvironment of gliomas is far from a passive bystander; rather, it actively participates in tumour evolution, driving key phenotypic changes in cancer cells and significantly influencing treatment outcomes (107). Understanding the molecular and cellular composition of the glioma microenvironment is therefore important to take into consideration when developing effective therapeutic strategies, particularly in overcoming resistance mechanisms that stem from its highly adaptive nature.

A major component of the glioma microenvironment is the brain extracellular matrix (ECM), which provides both structural integrity and biochemical signalling cues that regulate cellular behaviour. Unlike ECM in peripheral tissues, which is largely composed of fibrous proteins such as collagen and fibronectin, the brain ECM is primarily composed of hyaluronic acid, proteoglycans and glycoproteins, making it relatively soft and highly hydrated (108). Glioma cells interact with the ECM through integrins and other adhesion molecules (109), which facilitates invasion and promote the remodelling of ECM components to favour tumour expansion. ECM degradation and remodelling, driven by matrix metalloproteinases and other enzymes, further contribute to glioma infiltration into healthy brains tissue, a hallmark feature that makes complete surgical resection nearly impossible (110). Additionally, glioma cells can hijack native ECM signalling pathways to maintain a proliferative and invasive phenotype (110), underscoring the importance of ECM dynamics in glioma pathophysiology.

Another critical feature of the glioma microenvironment is its tumour vasculature, which exhibits abnormalities compared to normal brain blood vessels. Gliomas induce angiogenesis, the formation of new blood vessels, through upregulation of pro-angiogenic factors such as vascular endothelial growth factor (VEGF) (33). However, the newly formed vasculature is structurally and functionally defective, characterized by disorganised, highly permeable and inefficiently perfused vessels. This leads to irregular oxygen and nutrient distribution, promoting regions of hypoxia and increased interstitial pressure within the tumour. The dysfunctional vasculature not only facilitates

tumour progression but also has profound implications for therapy: poor perfusion limits drug delivery, while increased vessel permeability contributes to oedema and intracranial pressure, exacerbating disease symptoms.

The immune landscape within gliomas is equally complex. Despite the brain being considered an immune-privileged organ, gliomas actively engage with and manipulate the immune system to evade detection. One of the most prominent immune players in glioma progression is the tumour-associated macrophages (TAMs), which are largely composed of microglia (the brain's resident immune cells) and monocyte-derived macrophages infiltrating from the periphery (111). TAMs can adopt a pro-tumorigenic phenotype, supporting glioma progression by secreting cytokines, growth factors, and matrix-degrading enzymes that enhance invasion and angiogenesis. Notably, gliomas exploit the immune system's regulatory mechanisms to suppress anti-tumour immune responses, creating an immunosuppressive microenvironment that allows tumours to grow unchecked. For example, the upregulation of programmed cell death-ligand 1 (PD-L1) on glioma cells (112) and the secretion of immunosuppressive cytokines such as TGFβ and IL10 contribute to immune evasion (113), effectively dampening the activity of cytotoxic T cells.

Beyond general immune suppression, gliomas also foster a direct interplay with macrophages, which further contributes to tumour progression. TAMs are often found at the invasive front of gliomas, where they actively engage in remodelling the extracellular matrix and facilitating tumour cell migration. Moreover, macrophages are heavily recruited to necrotic and hypoxic regions within the tumour, responding to signals such as colony-stimulating factor 1 (CSF1) and HIF1a. In these, regions, macrophages secrete angiogenic factors that help sustain glioma growth despite a challenging microenvironment (113). Recent studies have also demonstrated that glioma-derived extracellular vesicles play an important role in altering macrophage behaviour, delivering microRNAs, lipids, and proteins that reprogram macrophages toward a tumour-supportive state (114).

One of the defining features of the glioma environment is hypoxia, which arises due to inadequate oxygen supply caused by the abnormal vasculature. While glioblastomas are highly vascularized, their blood vessels are poorly organized and inefficient, leading to regions of sever oxygen deprivation that drive tumour adaptation and progression (115, 116). Hypoxia is a powerful selective pressure that triggers the activation of hypoxia-inducible factors (HIFs), which regulate genes involved in angiogenesis, metabolic reprogramming, invasion and therapy resistance (117). This hypoxic stress forces tumour cells to switch to aerobic metabolism, known as the Warburg effect, increasing reliance on alternative nutrient sources to sustain proliferation and survival (56). It also promotes the maintenance of glioma stem-like cells (GSCs) – a subpopulation with heightened tumorigenic potential that contributes to recurrence and treatment resistance (110). A striking pathological hallmark of glioblastomas is the presence of necrotic foci surrounded by cellular pseudopalisades, where migrating tumour cells actively respond to oxygen gradients by upregulating invasive and pro-survival pathways (109). Oxygen concentration in the healthy human brain typically ranges around 4.6% O_2, but within gliomas, this can drop to as low as 1.7% O_2, creating a tumour core with extreme hypoxic stress (112). In contrast, under laboratory conditions, GBM cell lines are traditionally cultured at 20.(% O_2, an artifically hyperoxic condition that does not accurately reflect the physiological hypoxia of the

tumour microenvironment (110). The response to hypoxia is mediated by HOF stabilization, where HIFa subunits translocate to the nucleus and heterodimerize with HIFβ, initiatingthe transcription of genes associated with glycolysis, angiogenesis, and tumour invasion (110). This includes key regulators such as VEGF, GLUT1, BCL2, survivin, GFAP, and vimentin, all of which drive tumour cell survival and plasticity (117). Furthermore, HIF-driven pathways enhance the expression of pluripotency-associated transcription factors such as SOX2, OCT4, KLF4, MYC, and NANOG, facilitating the reprogramming of glioblastoma cells toward a more stem-like phenotype that reinforces tumour heterogeneity and therapy resistance (110, 111). Ultimately, the profound impact of hypoxia on glioblastoma biology makes it a crucial factor in tumour progression and patient prognosis, emphasizing the need for more physiologically relevant experimental models that mimic the hypoxic tumour microenvironment in vitro.

A key aspect of glioma adaptation to its microenvironment is the secretion of cytokines and growth factors, which act as essential mediators of cell-to-cell communication within the tumour niche (48). Cytokines such as IL6, IL8 and TNFα promote tumour cell survival, proliferation and migration, while growth factors like platelet-derived growth factors (PDGF) and epidermal factor (EGF) drive oncogenic signalling pathways (48, 113). The presence of these signalling molecules not only orchestrates interactions between tumour cells, immune cells and stromal components, but also contributes to chemotherapy and radiotherapy resistance, allowing gliomas to persist despite aggressive treatment.

Extracellular vesicles (EVs) represent another crucial, yet often overlooked, component of the glioma microenvironment (118). Glioma cells release exosomes and microvesicles carrying a diverse cargo of RNA, proteins, and lipids, which facilitate communication with neighbouring cells and aid in reshaping the tumour microenvironment. EVs can transfer oncogenic signals to nearby cells, enhance invasion and metastasis, and contribute to drug resistance by exporting cytotoxic agents out of tumour cells (118). Importantly, EV-mediated crosstalk extends beyond glioma cells themselves, influencing immune cells, endothelial cells, and even astrocytes to create a pro-tumorigenic state.

From a metabolic perspective, gliomas undergo profound metabolic reprogramming, shifting their energy production towards aerobic glycolysis, commonly known as the Warburg effect (42). Even in the presence of oxygen, glioma cells preferentially metabolize glucose into lactate, a process that enhances biosynthetic precursor availability for rapid cell proliferation. This altered metabolism not only fuels tumour growth but also alters the nutrient composition of the microenvironment, influencing surrounding cells and further driving tumour adaptation.

Finally, the glioma microenvironment is heavily influenced by the blood-brain barrier (BBB), a highly selective endothelial barrier that normally protects the CNS from harmful substances in circulation (119, 120). However, gliomas disrupt the integrity of the BBB, leading to increased permeability and the formation of a “blood-tumour barrier” (BTB). While this disruption facilitates the infiltration of tumour-promoting immune cells and macromolecules, it also presents a major therapeutic challenge: the altered permeability is still highly selective, preventing many chemotherapeutic agents from effectively reaching tumour cells. Consequently, the presence of a compromised but still partially functional BBB remains a major obstacle in glioma treatment,

highlighting the need for targeted drug delivery strategies that can efficiently penetrate the tumour microenvironment (120).

In summary, the glioma microenvironment is an extraordinarily complex and adaptive system that actively contributes to tumour progression, invasion, immune evasion, and therapy resistance. Its unique composition – including ECM alterations, abnormal vasculature, immune suppression, hypoxia, metabolic shifts, and BBB disruption – creates significant barriers to effective treatment. As the understanding of the glioma microenvironment deepens, it is becoming increasingly clear that successful therapeutic interventions must consider not only glioma cells themselves, but also their interactions with the surrounding niche.

1.3.3. Treatments

As the most common and aggressive malignant primary brain tumour in adults, it is crucial to establish effective treatment options. However, currently, these options consist of surgery, typically followed by chemotherapy or radiotherapy, but unfortunately, still result in patient survival of just over 1 year (121). The surgical step not only aims to obtain tissue for diagnosis, but also to improve neurological function, facilitate steroid weaning, prolong survival, and enhance quality of life. In cases where surgery is not feasible, such as in elderly patients or those with a poor performance status, a less invasive biopsy may be considered (81, 122). However, if the risk of biopsy is deemed too high or prognosis is very unfavourable, best supportive care or palliative radiotherapy may be recommended. Extent of resection has been linked to improved survival, but care must be taken to avoid injury to critical brain regions. Various surgical adjuncts, including neuroanatomical navigation systems and fluorescent dyes, aid in maximizing tumour removal while minimizing postoperative disability (81, 122). Tumour specimens are classified and graded according to the WHO classification for CNS tumours, with genetic distinctions such as IDH mutation status influencing prognosis. Treatment following resection or biopsy depends on factors such as age and performance status, with standard protocols involving radiation therapy and temozolomide chemotherapy. Response to temozolomide can be predicted by the MGMT promoter methylation status, with methylated tumours showing improved survival compared to unmethylated tumours (81).

Even with extensive surgical resection and intensive adjuvant therapy, nearly all GBM tumours experience local recurrence post-treatment. In addition, persistent challenges in treating GBM include incomplete tumour removal, significant genetic heterogeneity, the presence of the BBB, and an immunosuppressive microenvironment (120).

High Infiltration

GBM is characterized by its highly infiltrative nature, which poses significant challenges for achieving complete cellular resection. Even with maximal surgical efforts, microscopic tumour cells often remain, leading to tumour recurrence. Additionally, within GBM tumours, there are abundant hypoxic regions that create perivascular niches. These niches provide a conducive environment for glioma initiating cells (GICs) to thrive. GICs are self-renewing cells with stem-like properties that have been implicated in the initiation and progression of GBM. Moreover, these cells are notoriously resistant to conventional treatments such as radiotherapy and chemotherapy. As a result, the presence of GICs in hypoxic niches contributes to the development of more aggressive and treatment-resistant recurrent tumours (123).

Intertumor and Intratumor Heterogeneity
GBM exhibits significant heterogeneity both between different tumours (intertumoral heterogeneity) and within individual tumours (intratumor heterogeneity). This diversity in genetic and epigenetic characteristics complicates the development of targeted therapies. Previous efforts by The Cancer Genome Atlas (TCGA) categorized GBMs into four distinct molecular subtypes: mesenchymal, classical, proneural, and neural. Each subtype is associated with specific genetic alterations and clinical features. However, recent studies have revealed further complexity, showing that different molecular subtypes can coexist within the same tumour. This spatial and temporal variation underscores the challenges in devising effective treatments that target the diverse molecular landscapes of GBM.

Blood Brain Barrier (BBB)
The blood-brain barrier (BBB) is a specialized structure that tightly regulates the passage of molecules between the bloodstream and the brain parenchyma. While the BBB serves a crucial role in protecting the brain from harmful substances, it also presents a formidable obstacle to delivering therapeutic agents to GBM tumours. In healthy brain tissue, the BBB restricts the entry of most molecules, including chemotherapeutic drugs. However, in GBM, the BBB is often disrupted due to the presence of leaky blood vessels and altered expression of transporter proteins. Despite this disruption, the permeability of the BBB can vary within individual tumours, with some regions exhibiting enhanced permeability while others remain relatively intact. Furthermore, even if drugs manage to penetrate the tumour tissue, they may encounter resistance mechanisms, such as upregulation of efflux pumps by glioblastoma cells, which limit their efficacy.

Immunosuppressive Microenvironment
The microenvironment of GBM plays a crucial role in tumour progression and treatment response. Some GBMs are characterized by an immunosuppressive microenvironment, often referred to as "cold tumours." In these tumours, there is a paucity of tumour-infiltrating lymphocytes (TILs), particularly cytotoxic T cells, which are essential for mounting an effective anti-tumour immune response. Additionally, these tumours exhibit defects in antigen presentation and high levels of immunosuppressive cells, such as regulatory T cells and myeloid-derived suppressor cells. Consequently, GBMs with an immunosuppressive microenvironment are resistant to immunotherapy, including immune checkpoint inhibitors. In contrast, "hot tumours" are characterized by robust T cell infiltration and greater immunogenicity, making them more responsive to immunotherapy. Strategies aimed at converting "cold" tumours into "hot" ones hold promise for enhancing the efficacy of immunotherapy in GBM.

In light of these challenges, ongoing research efforts are focused on improving systemic therapies for GBM. These efforts encompass a range of approaches, including the development of novel pharmacologic agents, elucidation of resistance mechanisms, and exploration of innovative treatment strategies aimed at overcoming the complex biological features of GBM.

1.3.4. Genetic drivers of Glioblastoma

Although glioblastoma is a disease of heterogenous nature, previous studies have identified consistently occurring genetic alterations, also known as driver genes, in

TP53, PTEN, EGFR, PIK3CA, PIK3R1, NF1, and RB1 (Figure 1.5); as well as the order of mutations and patterns of tumour growth (41). These key genetic events occurring in IDH-wildtype and IDH-mutant GBM will be briefly summarized in the following paragraphs.

Genetic Alterations (based on cBioPortal data)

IDH-wt GBM

Mutations

Genes	Frequency (%)	Ref.
TERT	78 - 84	13,14
PTEN	21 - 39	3,10,13,14
TP53	7 - 31	3,10,13,14
EGFR	4 - 22	3,10,13,14
NF1	10 - 19	3,10,14
RB1	4 - 12	3,10,13,14
PIK3CA	9 - 11	3,13,14

Copy number variations

Genes	Frequency (%)	Ref.
CDKN2A	25 - 57	3,13,14
CDKN2B	25 - 56	3,13,14
EGFR	29 - 46	3,13,14
CDK4	12 - 15	3,13,14
PDGFRA	9 - 14	3,13,14

IDH-mut High-grade Glioma

Mutations

Genes	Frequency (%)	Ref.
TP53	56 - 75	3,9,13,14
ATRX	45 - 63	9,13,14
FUBP1	10 - 50	9,10,14
PIK3CA	10 - 19	3,9,10,13,14
CIC	7 - 50	3,9,10,13,14
NOTCH1	7 - 15	9,14
TERT	0.2 - 50	9,13,14

Copy number variations

Genes	Frequency (%)	Ref.
CDKN2A	9 - 43	3,14
CDKN2B	9 - 43	3,14
CDK4	5 - 36	3,14
MYCN	5 - 14	3,14
PDGFRA	5 - 7	3,14

Figure 1.5: Overview of genetic alterations in IDH-wildtype and IDH-mutant GBMs. Frequently occurring driver mutations and CNVs in IDH-wildtype and IDH-mutant high-grade gliomas (WHO grade 3 & 4) were listed as above. Image adapted from (10)

IDH-wildtype Glioblastoma

The above-mentioned genes (Figure 1.5) affect important signalling pathways such as growth factor, p53, and retinoblastoma. These alterations contribute to uncontrolled proliferation and are crucial for GBM development. Notably, IDH-wildtype GBMs, frequently exhibit mutations in the telomerase reverse transcriptase promoter (TERTp), promoting increased TERT expression and telomere activation. This mutation is suggested to an early event, potentially triggering other GBM driver mutations.

IDH-mutant Glioblastoma

Approximately 12% of GBMs carry IDH mutations, with 73%-83% of IDH1/2 mutations found in secondary GBMs, making them a key early event in gliomagenesis (124, 125). These mutations significantly alter metabolic pathways, epigenetic regulation, and reactive oxygen species (ROS) homeostasis, contributing to tumour progression (125, 126). Alongside IDH1 mutations, TP53 loss-of-function mutations frequently occur, often followed by ATRX alterations, which drive alternative lengthening of telomeres (ALT) as a mechanism of telomere maintenance in GBM (10). This highlights how core driver mutations influence tumour cell survival and adaptability. Additionally, hypoxia-induced ROS triggers hypoxia-inducible factor-1α (HIF-1α) signalling, further shaping the tumour microenvironment. Attempts to categorize GBMs based on genetic similarities have identified proneural, classical, and mesenchymal subtypes, suggesting that distinct cells-of-origin may dictate glioblastoma gene expression patterns and tumour behaviour (10, 127).

As described in this chapter, GBM is a highly aggressive tumour with poor prognosis and without affective treatment options beyond surgery, which does not guarantee non-recurrence. In a recent clinical update in the British Medical Journal, the authors posed the following questions for future research:

- Are there any biomarkers that could expedite glioblastoma diagnosis and monitor treatment response?
- What interventions can shorten the time to diagnosis, and do these affect survival outcomes?
- How can clinical trial design be adapted to examine the efficacy of future targeted drug therapies? (81)

In my research, my aim was to add scientific knowledge and value to the last point, by exploring alternative drug targets: microRNAs.

1.4. microRNAs

Disclaimer: microRNA was chosen as a target because I was supposed to go to Italy and work with a research group that studies plant microRNA-based drugs in mice. A group member there, had developed a microRNA-mRNA binding prediction algorithm, which I was supposed to collaborate to improve. Unfortunately, the internship fell through because of COVID and other staffing reasons. Since the idea was already in place, my supervisor and I decided to stick with miRNAs – as the funding was for this project - but switch to glioblastoma as the model organism because the Edinburgh group was already working on this disease and had samples available in-house.

Ribonucleic acids (RNAs) are essential molecules that play a wide variety of roles in cells. For a long time, they were thought to be just intermediaries, carrying genetic information from DNA to make proteins. But as research has progressed, it's become clear that RNA does so much more. Beyond its part in protein synthesis, RNA is involved in regulating genes, supporting cellular structures, and even catalyzing reactions. The idea that RNA could be more than a messenger started to emerge in the mid-20th century, and since then, scientists have uncovered its central role in nearly every biological process. Understanding RNA also means understanding what happens when it malfunctions – especially how that contributes to diseases like cancer. The most well-known types of RNA are messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA) (Figure 1.6) (128). mRNA was first identified in the early 1960s as the molecule that carries instructions from DNA to the ribosome, where proteins are built (129, 130). This process begins with DNA being transcribed into mRNA via steps such as: pre-mRNA transcription, addition of a protective 5' cap, splicing, polyA tail attachment to the 3' end. This mature mRNA then acts as the blueprint for protein synthesis (131).

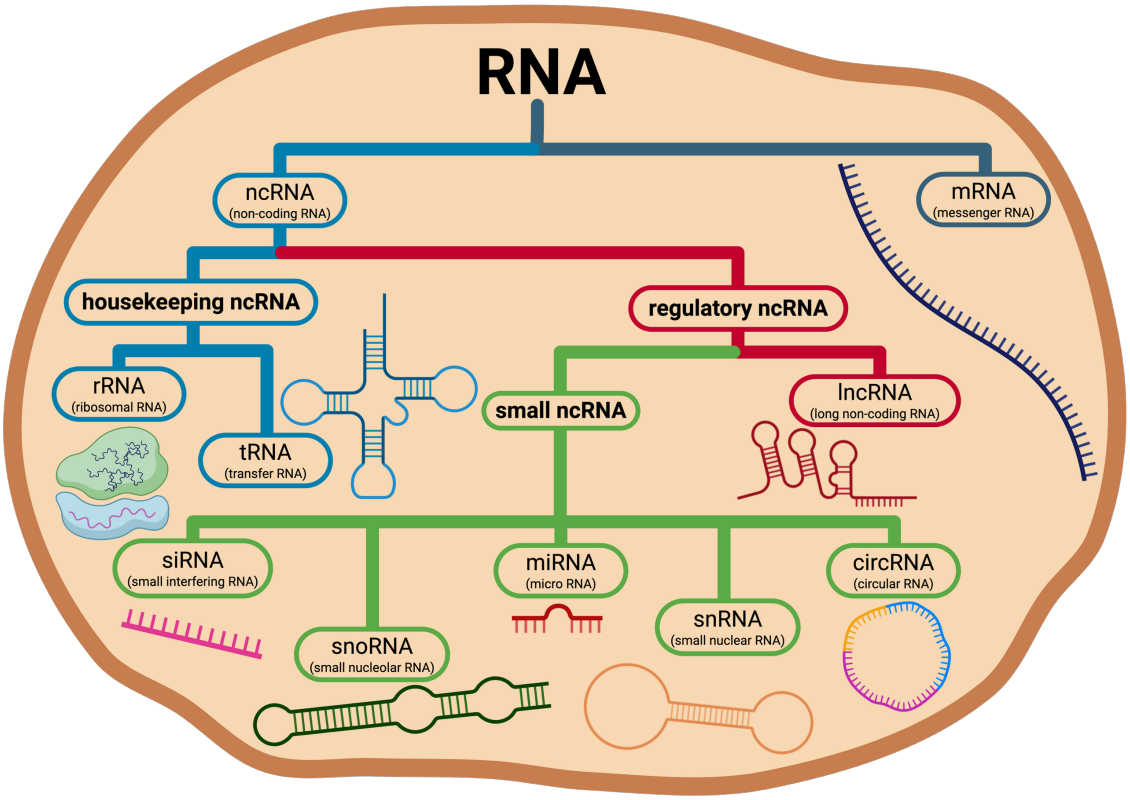


Figure 1.6: Types of RNA. RNAs are categorized by cellular role, structure, and origin. Messenger RNA (mRNA) is the transient informational molecule transcribed from DNA and translated to protein – coding RNA. Non-coding RNA (ncRNA) is a term encompassing a vast number of RNA molecules with specific roles. Housekeeping ncRNAs include: ribosomal RNA (rRNA) which constitutes the catalytic and structural core of the ribosome; transfer RNA (tRNA) adopts a cloverleaf structure, delivering specific amino acids to the ribosome via codon-anticodon pairing. Regulatory ncRNAs can be long ncRNAs (lncRNA), which is a heterogeneous class (>200 nt) involved in transcriptional and epigenetic control, or small ncRNA like small interfering (siRNA), small nucleolar (snoRNA), micro (miRNA), small nuclear (snRNA), and circular (circRNAs) – and many more. SiRNA typically derives from exogenous double-stranded RNA and mediates precise target cleavage via perfect complementarity. SnoRNA primarily guides nucleotide modifications in rRNA. miRNA guides post-transcriptional gene silencing by imperfectly binding target mRNAs. SnRNA directs pre-mRNA splicing within the spliceosome. CircRNA often functions as a miRNA sponge or protein scaffold.

TRNA and rRNA, on the other hand, work directly in the translation process. tRNAs act as adaptors, matching mRNA codons with the correct amino acids to build proteins. rRNAs are a core part of ribosomes, which are the cellular machines where proteins are made. Unlike mRNA, tRNA and rRNA are transcribed by RNA polymerase III and RNA polymerase I, respectively. They also go through their own maturation steps, including folding into precise structures and undergoing chemical modifications.

When these "classic" RNAs don’t work properly, it can lead to serious diseases, including cancer. For example, errors in mRNA splicing can create dysfunctional proteins that drive cancer progression. Problems with ribosomes, often caused by mutations in rRNA or disruption in its production, can boost the translation of oncogenes or lower the expression of tumour suppressors. Even tRNAs, when improperly processed, can disrupt the balance of protein production and contribute to cancer. In addition to these well-known RNAs, researchers have discovered a whole world of other RNA types, many of which don’t code for proteins but still have important jobs. These are non-coding RNAs (ncRNAs), and they can be split into two main

groups: housekeeping ncRNAs and regulatory ncRNAs (Figure 1.6). Small nuclear RNAs (snRNAs) are one type of housekeeping ncRNA that help splice pre-mRNA, removing introns and connecting exons to form mature mRNA. This is a crucial step in gene expression, and when it goes wrong, it can lead to abnormal proteins being produced. This has been linked to various cancers, as defective splicing can activate cancer-promoting genes or silence genes that normally prevent tumours. Another group, small nucleolar RNAs (snoRNAs), works behind the scenes to chemically modify other RNAs like rRNAs and tRNAs. These modifications help stabilize RNA structures and fine-tune their functions. Changes in snoRNA levels or activity have been connected to cancer too, as they can affect how well ribosomes work, shifting the balance of protein production in ways that promote tumour growth. Regulatory ncRNAs are where things get really interesting. These include microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs), which all play roles in controlling gene expression. miRNAs are short, about 20–22 nucleotides long, and they post-transcriptionally act as molecular switches, turning genes on or off by binding to mRNAs and preventing them from making proteins. lncRNAs, on the other hand, are much longer and can interact with DNA, RNA, or proteins to regulate genes in more complex ways. circRNAs form closed loops, which are more stable than other RNAs and have also been shown to act similarly to lncRNAs as "sponges" for miRNAs, soaking them up and limiting their activity (132).

These different RNA types don’t work in isolation—they interact in complex networks. For example, lncRNAs can guide or block the activity of miRNAs, while circRNAs can compete with mRNAs for miRNA binding, regulating how genes are expressed. This interconnected system allows cells to precisely control which proteins are made and when. However, it also means that a problem in one part of the network can ripple through the system, causing imbalances that contribute to diseases like cancer. Of all the regulatory ncRNAs, miRNAs are particularly fascinating because they are involved in so many processes majorly due to their small size and are often dysregulated in disease. This thesis focuses on miRNAs, starting with their discovery and biogenesis before diving into their roles in normal and pathological conditions. Specifically, I will explore their potential as drug targets, given their central role in controlling gene expression.

In recent years, miRNAs have emerged as key players in the regulation of gene expression. These small, non-coding RNAs fine-tune the activity of target genes by binding to messenger RNAs, typically at the 3′ untranslated region (3′ UTR), to suppress their translation or promote their degradation (133). The discovery of miRNAs in 1993 by the Ambros and Ruvkun groups profoundly shaped the ‘then’-understanding of molecular biology, revealing a previously unknown layer of gene regulation (134). Since then, countless miRNAs have been discovered in all plant and animal systems, even to this day (135, 136). Unlike protein-coding genes, miRNAs don’t code for proteins themselves; instead, their effects are mediated by their ability to regulate multiple mRNAs simultaneously, making them master regulators of cellular pathways. What makes miRNAs particularly interesting is their role in human health and disease. Each miRNA can regulate hundreds of target genes, and their activity is tightly controlled to maintain cellular homeostasis. When miRNA expression is dysregulated, it can lead to various diseases, including cancer (137, 138). In fact, miRNAs have been implicated in nearly every stage of tumour development, from initiation to metastasis. Moreover, miRNAs can be secreted into extracellular fluids,

where they have been recognized as possible biomarkers for various diseases (139, 140). Some miRNAs act as tumour suppressors, while others function as oncogenes. For instance, the miR-17-92 cluster is known to promote cell proliferation and survival, contributing to tumour growth in cancers such as lymphoma and lung cancer. Conversely, let-7, one of the first miRNAs discovered, acts as a tumour suppressor by targeting oncogenes like MYC and RAS.

RNA-based therapies represent a groundbreaking approach to treatment, harnessing the unique properties of RNA molecules to influence gene expression, regulate cellular processes, and even correct genetic abnormalities. Unlike traditional drugs that target proteins by binding to active sites to trigger functional changes, RNA therapies work directly at the genetic level, enabling the modulation of virtually any gene or pathway within the cell—even those that do not code for proteins but are still implicated in disease pathways, such as non-coding RNAs (141). This approach is particularly notable given that proteins constitute only about 1.5% of the human genome, and only an estimated 10-14% of those proteins have structures that make them “druggable” by conventional small molecules (136, 142). RNA-based therapies bypass this limitation by targeting the genetic information itself, providing the ability to intervene with high specificity across a much broader range of cellular processes. This high specificity not only enhances therapeutic efficacy, but also holds potential for reduced side effects compared to tradition therapies, a significant advantage in treating complex and chronic diseases (143, 144). The applications of RNA therapies span numerous disease categories, including genetic disorders, various types of cancer, and infectious diseases. The recent success of mRNA vaccines against COVID-19, which use mRNA encoding the SARS-CoV-2 spike protein, highlighted the therapeutic potential of RNA technologies on a global scale, inspiring further development across diverse therapeutic areas (144). Through advancements in RNA delivery methods and chemical modifications, RNA-based therapies are quickly becoming a versatile and potent class of drugs that could reshape the future of medicine (145).

Beyond vaccines, mRNA therapies are being explored for protein replacement in genetic diseases and cancer immunotherapy. Protein replacement therapy involves delivering mRNA that encodes missing or defective proteins in diseases like cystic fibrosis or certain metabolic disorders. In cancer, mRNA can be engineered to produce tumour-specific antigens that stimulate an immune response, training the body to recognise and attack cancer cells. These strategies showcase the versatility of mRNA in addressing complex disease mechanisms (145). The rapid advancements in lipid nanoparticle delivery systems have played a pivotal role in these applications, making it possible to efficiently deliver mRNA into target cells without triggering significant immune responses (144, 145).

1.4.1. MicroRNA Biogenesis

MiRNA biogenesis is a complex process that includes transcription, processing, nuclear export, and maturation in the cytoplasm. It relies on coordinated action of specialized proteins and complexes, each designed to recognize specific RNA structures and sequences. While the canonical pathway is the most well-studied method of miRNA production, there are also non-canonical pathways that provide alternative means of generating functional miRNAs under certain conditions. These alternative pathways demonstrate the flexibility of miRNA biogenesis in adapting to cellular environments and external signals. Nevertheless, the canonical pathway

remains the primary method for miRNA production and is essential for understanding how miRNAs mature, function and ultimately use them as drug targets.

The regulation of miRNA biogenesis is just as important as the process itself. Every step in the pathway is controlled by various cellular factors, which help ensure that mature miRNAs are expressed at the right time and place. The transcriptional regulation of miRNA genes involves transcription factors and epigenetic modifications like DNA methylation and histone acetylation. On the other hand, post-transcriptional regulation includes proteins that interact with primary and precursor miRNA molecules, influencing their processing or stability. For example, RNA-binding proteins such as Lin28 can prevent the processing of certain miRNAs, adding another layer of control. This dynamic regulation guarantees that miRNA expression is carefully adjusted to meet the needs of both the cell and the organism. Disruptions in miRNA biogenesis can significantly impact cellular function. Changes in the expression or activity of essential biogenesis enzymes, such as Drosha, Dicer, and Argonaute proteins, have been associated with a range of diseases. In cancer, for instance, the overexpression or downregulation of certain miRNAs—often referred to as oncomiRs and tumour-suppressor miRNAs, respectively—play a role in tumour development, metastasis, and resistance to treatment (114, 146). Likewise, abnormal miRNA biogenesis is linked to cardiovascular diseases, neurodegenerative disorders, and immune system dysfunctions (138). These insights emphasize the clinical importance of miRNA biogenesis and its potential as a target for therapy.

The evolutionary conservation of miRNAs and their biogenesis machinery highlight their essential role in the survival and adaptability of various organisms. Research conducted on model organisms like Drosophila melanogaster, Arabidopsis thaliana, and mice has provided valuable insights into how miRNAs are produced and function (147). These investigations have shown that the canonical biogenesis pathway is remarkably conserved across different species (148), underscoring its vital role in maintaining cellular and organismal balance. In the last twenty years, breakthroughs in high-throughput sequencing and computational biology have transformed miRNA research (135). These advancements have allowed scientists to identify thousands of miRNAs and their targets (149), revealing the extensive regulatory networks they influence. Additionally, structural studies of miRNA biogenesis proteins, including Drosha, DGCR8, and Dicer, have clarified the molecular mechanisms behind their function and specificity. These findings not only enhance our understanding of miRNA biology but also open up new avenues for innovative therapeutic approaches.

While significant progress has been made, many questions still linger regarding the complexities of miRNA biogenesis. For example, the mechanisms that regulate miRNA processing during stress or developmental stages are not completely understood. Additionally, the interaction between canonical and non-canonical pathways, along with their respective contributions to the miRNA pool in various cell types and contexts, continues to be a focus of research (141, 148). Tackling these questions will necessitate a blend of molecular, computational, and systems biology methods. In this subchapter, I will explore the biogenesis of miRNAs. I will firstly highlight the canonical pathway, emphasizing its molecular machinery and regulatory mechanisms, following with an introduction to non-canonical pathways as alternative methods for miRNA generation. The canonical pathway is the primary mechanism through which most miRNAs are processed and matured. This process involves a

series of key enzymes and complexes, such as Drosha, DGCR8, Dicer, Exportin-5, and Argonaute proteins. It showcases the precision and efficiency needed to produce functional miRNAs that can regulate gene expression with high specificity. In the next subchapter, I will provide a detailed analysis of the canonical pathway, outlining each step and its regulation.

1.4.2. Canonical pathway

The biogenesis of miRNAs starts in the nucleus with the transcription of miRNA genes. These genes can be found in various genomic locations, such as intergenic regions, introns of protein-coding genes, and exonic regions of non-coding genes (148, 150). Most miRNA genes are transcribed by RNA polymerase II, while RNA polymerase III is responsible for transcribing a smaller group of miRNAs. This transcription results in primary miRNAs (pri-miRNAs), which are long transcripts that are capped and polyadenylated (Figure 1.7) (148, 150). Typically, pri-miRNAs span several kilobases and feature one or more stem-loop structures that are crucial for further processing. These stem-loop regions include a double-stranded RNA stem of about 33–35 base pairs, a terminal loop, and single-stranded flanking sequences. The structural characteristics of pri-miRNAs play a vital role in their recognition by the nuclear processing machinery. Once synthesized, pri-miRNAs undergo their first maturation step in the nucleus, mediated by the Microprocessor complex (151). This complex consists of the ribonuclease III enzyme Drosha and its essential cofactor, DGCR8 (DiGeorge Syndrome Critical Region 8). DGCR8 recognizes and binds to the stem-loop structure of the pri-miRNA, while Drosha performs the enzymatic cleavage (148). Drosha cuts the pri-miRNA approximately 11 base pairs away from the junction between the stem-loop and the single-stranded RNA flanking sequences. This cleavage generates a precursor miRNA (pre-miRNA), a shorter, hairpin-shaped molecule approximately 70 nucleotides in length (148). The pre-miRNA possesses a characteristic 2-nucleotide overhang at its 3′ end, which is a hallmark feature recognized by downstream processing enzymes. The precision of Drosha's cleavage is important, as errors at this stage can disrupt subsequent processing steps and compromise the functionality of the mature miRNA.

After nuclear processing, pre-miRNA is transported to the cytoplasm for further maturation (Figure 1.7). This transport is carried out by Exportin-5, which is part of the karyopherin family of nuclear transport receptors. Exportin-5 identifies the double-stranded RNA structure of the pre-miRNA along with its 2-nucleotide 3′ overhang. The transport process requires energy and depends on the small GTPase Ran. When Ran-GTP is present, Exportin-5 binds to the pre-miRNA and helps it move through the nuclear pore complex (152). Once in the cytoplasm, Ran-GTP is converted to Ran-GDP, which releases the pre-miRNA for the next processing stage. The efficiency and specificity of nuclear export are vital for ensuring that pre-miRNA is correctly localized in the cytoplasm and protected from premature degradation.

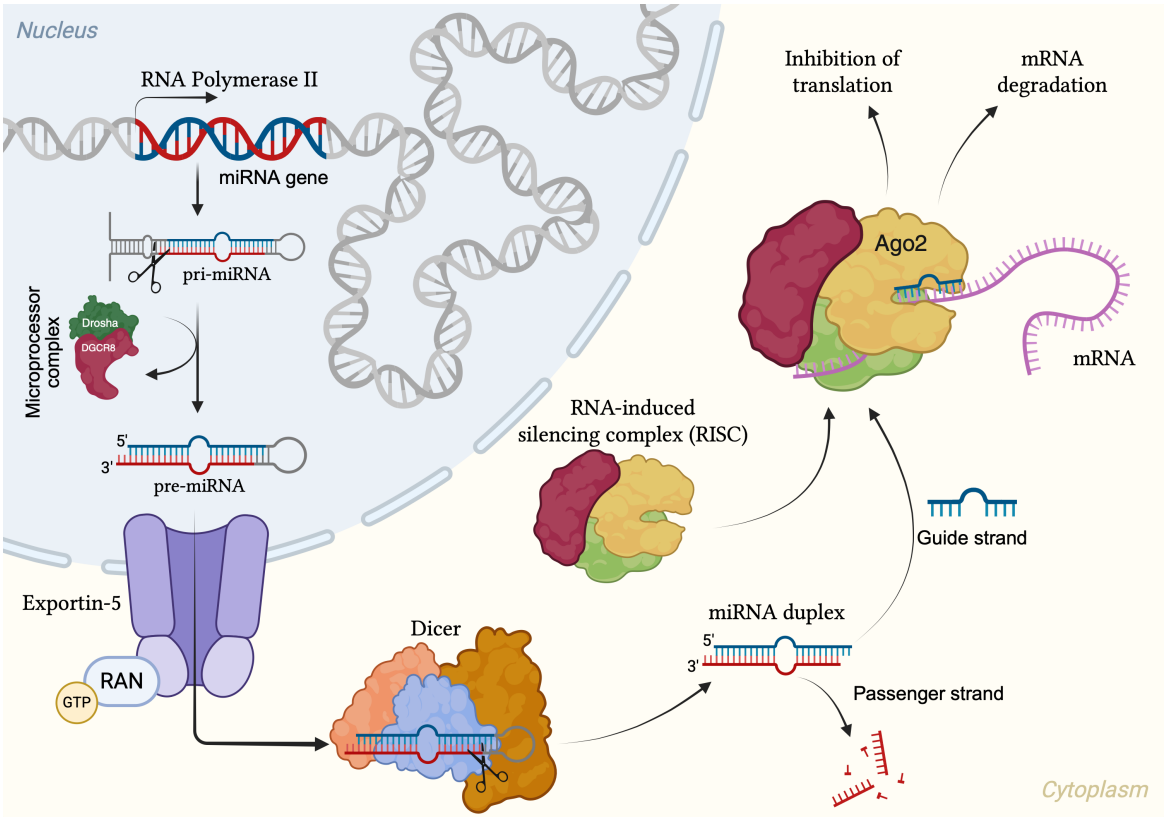


Figure 1.7: Canonical pathway of miRNA biogenesis. MicroRNA genes are transcribed by RNA polymerase II (or III) to generate primary miRNA transcripts (pri-miRNA) in the nucleus. The Microprocessor complex, comprising Drosha and DGCR8, cleaves the pri-miRNA to release a ~70 nucleotide precursor miRNA (pre-miRNA) hairpin. This pre-miRNA is exported to the cytoplasm by Exportin-5 and Ran-GTP. The RNase III enzyme Dicer, incomplex with TRBP, processes the pre-miRNA into a transient ~22 bp miRNA duplex. One strand (the guide strand) is selectively loaded into the Argonaute (AGO) protein within the RNA-induced silencing complex (RISC), while the passenger strand (miRNA) is typically degraded. The mature miRNA-RICS complex then binds to complementary sites in the 3' untranslated region (3'UTR) of target messenger RNAs (mRNAs), leading to translational repression or mRNA degradation.*

In the cytoplasm, the pre-miRNA undergoes its second and final cleavage, which is catalyzed by the ribonuclease III enzyme Dicer (Figure 1.7). Dicer is a highly conserved protein that has specialized domains for the precise processing of pre-miRNA. The PAZ domain of Dicer attaches to the 2-nucleotide 3′ overhang of the pre-miRNA, positioning it for cleavage by the RNase III domains. Dicer then cleaves the pre-miRNA within its stem region, resulting in a miRNA duplex that is about 20–24 nucleotides long. This duplex contains two strands: the guide strand, which will develop into the functional mature miRNA, and the passenger strand (miRNA*), which is usually degraded. The accuracy of Dicer's cleavage is essential for producing miRNAs of the correct length and sequence, as any deviations can hinder their ability to effectively bind to target mRNAs. After being processed by Dicer, the miRNA duplex is integrated into the RNA-induced silencing complex (RISC), which is a multi-protein complex that carries out the regulatory roles of miRNAs. The main component of RISC is an Argonaute (AGO) protein, usually AGO2 in mammals. During this loading phase, one strand of the miRNA duplex is chosen to be the guide strand, while the other, known as the passenger strand, is discarded. The selection of strands is mainly influenced by the thermodynamic stability at the ends of the duplex; the strand with the less stable 5′ end is more likely to be included in RISC. Once the guide strand is incorporated, it directs RISC to bind with complementary sequences in target mRNAs.

The mature RISC-miRNA complex plays a crucial role in gene silencing through two primary mechanisms: translational repression and mRNA degradation. In animals, miRNAs typically attach to partially complementary sequences found in the 3′ untranslated regions (UTRs) of target mRNAs. This interaction brings in additional proteins that either inhibit translation or encourage mRNA destabilization via processes like deadenylation and decapping. In certain instances, especially in plants, miRNAs can bind to their targets with nearly perfect complementarity, resulting in the direct cleavage of the mRNA by AGO2. The method of silencing is influenced by how closely the miRNA matches its target and the specific cellular environment.

The canonical pathway of miRNA biogenesis is regulated at multiple levels, ensuring that mature miRNAs are expressed accurately in terms of timing and location. At the transcriptional stage, the expression of miRNA genes is influenced by transcription factors and epigenetic changes, including DNA methylation and histone acetylation. Post-transcriptional regulation involves proteins that interact with pri- and pre-miRNAs, affecting their processing or stability. For instance, RNA-binding proteins like Lin28 can attach to specific pre-miRNAs and inhibit their processing by Dicer, which prevents the formation of mature miRNAs. Furthermore, the localization and compartmentalization of the miRNA processing machinery also play a role in regulating miRNA biogenesis.

The canonical miRNA biogenesis pathway plays a crucial role in maintaining cellular balance and managing intricate regulatory networks. When this pathway is disrupted, it can lead to significant issues, resulting in abnormal miRNA expression that is linked to various diseases. For example, the overproduction of oncogenic miRNAs (oncomiRs) or the reduced expression of tumour-suppressor miRNAs is commonly seen in cancer, which aids in tumour growth and spread. Likewise, problems in miRNA biogenesis are connected to cardiovascular diseases, neurodegenerative conditions, and immune system malfunctions. These insights highlight the necessity of comprehending the canonical pathway and its regulatory functions.

In conclusion, the canonical pathway serves as the primary mechanism for miRNA maturation, encompassing a series of well-coordinated steps that guarantee the creation of functional miRNAs. Each phase, from transcription and nuclear processing to cytoplasmic maturation and RISC assembly, is carefully regulated to yield mature miRNAs that can precisely control gene expression. Gaining insight into this pathway not only enhances our understanding of miRNA biology but also opens up potential therapeutic avenues for diseases linked to miRNA dysregulation.

1.4.2.1. Non-canonical pathway

While the canonical pathway is the main and well-understood mechanism for microRNA (miRNA) maturation, there are also alternative or non-canonical pathways that play a role in producing functional miRNAs. These non-canonical routes skip one or more steps of the canonical process and use different molecular machinery or processing strategies. Although these pathways are less frequent, they demonstrate the flexibility and adaptability of the cellular systems involved in RNA processing. Non-canonical pathways often emerge in specific cellular situations, such as under stress, in tissue-specific regulation, or during viral infections, highlighting the complexity of miRNA biology.

One significant type of non-canonical miRNA biogenesis is the Drosha-independent pathway, which eliminates the need for the Microprocessor complex that usually processes primary miRNAs (pri-miRNAs) into precursor miRNAs (pre-miRNAs). Among the most researched examples in this category are mirtrons. Mirtrons are short hairpin structures that originate from the introns of protein-coding genes. These introns are removed during mRNA splicing and subsequently debranched by the lariat debranching enzyme, forming a hairpin-shaped RNA that resembles canonical pre-miRNAs. Mirtrons are transported to the cytoplasm by Exportin-5 and then processed by Dicer to produce mature miRNAs. Based on their structure, mirtrons can be categorized into 5′-tailed, 3′-tailed, or canonical mirtrons. Even though they bypass Drosha, mirtrons function effectively as they are incorporated into the RNA-induced silencing complex (RISC) and regulate target mRNAs in a manner similar to canonical miRNAs. Another group of Drosha-independent miRNAs comes from transfer RNAs (tRNAs), referred to as tRNA-derived miRNAs (tRNA-miRs). These miRNAs are produced from precursor tRNAs (pre-tRNAs) that are cleaved by the tRNA splicing endonuclease (TSEN) complex during the maturation of tRNAs. The resulting fragments then fold into structures similar to pre-miRNAs, which are recognized by Exportin-5 and processed by Dicer. tRNA-miRs have been linked to stress response pathways and mechanisms that promote cell survival, especially in situations that inhibit the typical miRNA biogenesis. By utilizing pre-tRNA fragments for miRNA production, the cell maintains its ability to regulate gene expression even in challenging conditions.

The Dicer-independent category of non-canonical miRNA biogenesis, on the other hand, does not require Dicer-mediated cleavage in the cytoplasm. A key example of this is the Argonaute 2 (AGO2)-dependent pathway. In this process, pre-miRNAs that have been processed by Drosha in the nucleus are directly loaded into AGO2 in the cytoplasm, where AGO2 cleaves the pre-miRNA to generate a mature miRNA-like strand. This pathway is particularly important in cells that lack functional Dicer or in situations where Dicer activity is reduced. AGO2-dependent miRNA biogenesis is believed to be crucial for developmental processes and stress responses, serving as a backup mechanism for miRNA production. A distinct variation of the Dicer-independent pathway features simtrons—introns that are spliced out and, unlike mirtrons, do not form pre-miRNA-like hairpins. Instead, RNA molecules derived from simtrons are directly incorporated into AGO proteins, allowing for maturation without the involvement of Drosha or Dicer. This process exemplifies an even more efficient method of miRNA biogenesis and is believed to function under very specific conditions, highlighting the flexibility of miRNA processing systems.

Additional non-canonical pathways utilize unusual RNA substrates or processing enzymes. For example, small nucleolar RNAs (snoRNAs), which usually direct chemical modifications of ribosomal RNAs (rRNAs) and small nuclear RNAs (snRNAs), can be repurposed to serve as miRNA precursors. Some snoRNAs form hairpin structures that resemble pre-miRNAs, enabling them to be recognized and processed by Dicer. Likewise, endogenous short hairpin RNAs (shRNAs) are transcribed as single hairpin structures that skip Drosha cleavage and move directly to Dicer-mediated maturation. Although these occurrences are relatively rare, snoRNA-derived miRNAs and endogenous shRNAs enhance the variety of RNA substrates that can produce functional miRNAs, showcasing the adaptability of the cellular machinery.

Viruses also take advantage of non-canonical pathways to create miRNAs that influence host cell functions. For instance, herpesviruses produce miRNAs from short hairpin structures that skip the Drosha step and depend on Dicer for processing. Other viruses imitate the standard miRNA biogenesis by encoding RNA substrates that hijack the host cell's machinery. These viral miRNAs enable pathogens to control host gene expression, avoid immune detection, and boost their replication and survival. Investigating viral miRNAs not only uncovers new mechanisms of RNA processing but also sheds light on host-pathogen interactions. The regulation of non-canonical pathways is highly context-dependent, reflecting their specific roles in cellular and organismal biology. These pathways often become active when canonical miRNA biogenesis is disrupted, such as during stress, developmental changes, or in response to viral infections. By using alternative substrates and processing methods, non-canonical pathways ensure the production of miRNAs even when the standard machinery is not available. This flexibility is essential for maintaining cellular balance and adapting to environmental challenges.

Even though non-canonical miRNAs have specialized functions, they are similar in their roles to canonical miRNAs. After processing, they become part of the RISC complex and help regulate gene expression by pairing with target mRNAs. However, the targets for non-canonical miRNAs can vary because of differences in their sequences and structural characteristics, which may lead to unique regulatory networks. This variation adds to the complexity of gene regulation, enabling cells to adjust their responses to specific stimuli or conditions.

In summary, the non-canonical pathways of miRNA biogenesis offer alternative methods for creating functional miRNAs, skipping one or more steps of the traditional process. These pathways, which include mechanisms that do not rely on Drosha or Dicer, utilize a variety of RNA substrates and processing techniques. They demonstrate the flexibility of miRNA biogenesis and its capacity to adapt to the changing needs of the cell. Non-canonical miRNAs are essential for development, responses to stress, and various diseases, and their improper regulation has been linked to numerous health issues. Gaining insight into these pathways enhances our understanding of miRNA biology and opens up potential therapeutic avenues for conditions related to miRNA dysfunction.

1.4.3. Mechanisms of miRNA-mediated gene regulation

MiRNAs play a crucial role in regulating gene expression after transcription. They achieve this by binding to complementary or partially complementary sequences in target messenger RNAs (mRNAs). This binding usually takes place in the 3′ untranslated region (3′ UTR) of the mRNA, although it can also happen in the coding sequence or the 5′ UTR. The interaction between miRNAs and mRNAs can lead to either a reduction in translation or the degradation of the mRNA, depending on how well the sequences match and the specific cellular environment. Through these mechanisms, miRNAs help fine-tune protein expression, which is essential for regulating various physiological processes such as development, cell differentiation, proliferation, and apoptosis. The process of gene regulation by miRNA starts when a mature miRNA is incorporated into the RNA-induced silencing complex (RISC). The main component of RISC is an Argonaute (AGO) protein, which attaches to the guide strand of the miRNA. This guide strand acts as a template, helping the RISC complex locate complementary sequences in target mRNAs. Among the four Argonaute

proteins found in mammals (AGO1–4), AGO2 is particularly important due to its endonuclease activity, which allows it to directly cleave mRNA when there is near-perfect complementarity. However, in most animal cells, miRNAs typically bind to their targets with imperfect complementarity, resulting in translational repression and/or destabilization of the mRNA instead of direct cleavage.

One key way that miRNAs regulate gene expression is through translational repression. In this mechanism, the RISC complex, loaded with miRNA, attaches to the target mRNA, which prevents ribosomes and other components necessary for translation from being recruited. This repression can take place at various points during translation, such as during initiation and elongation. For instance, translation initiation is inhibited by blocking the formation of the eukaryotic initiation factor 4F (eIF4F) complex at the 5′ cap of the mRNA. On the other hand, miRNAs can also disrupt elongation by causing ribosomes to stall as they progress along the mRNA. Both of these mechanisms effectively decrease protein production without altering mRNA stability, allowing for a flexible regulatory approach that enables quick adjustments in protein levels in response to changes within the cell.

Another important way that miRNAs regulate gene expression is through the degradation of mRNA. When the RNA-induced silencing complex (RISC) binds to a target mRNA that matches it well enough, it brings in enzymes that trigger the destabilization of the mRNA via two main pathways: deadenylation and decapping. Deadenylation is the process of removing the poly(A) tail from the mRNA, which is carried out by deadenylase complexes like CCR4-NOT and PAN2-PAN3. This shortening of the poly(A) tail makes the mRNA more vulnerable to degradation by 3′-to-5′ exoribonucleases. In contrast, decapping involves taking away the 5′ cap structure of the mRNA, which is crucial for its protection. The decapping enzyme DCP2, often working with cofactors such as DCP1, removes this cap, leaving the mRNA open to degradation by 5′-to-3′ exonucleases like XRN1. Together, these processes of deadenylation and decapping significantly lower the stability of target mRNAs, leading to a decrease in their levels within the cell.

The decision between translational repression and mRNA degradation depends on how well the miRNA pairs with its target and the specific cellular context. In plants, where there is often perfect or nearly perfect complementarity, this typically results in the direct cleavage of mRNA by AGO2. In contrast, in animals, miRNA-target interactions are usually less than perfect, leading to a mix of translational repression and mRNA destabilization. Notably, recent research indicates that translational repression usually occurs before mRNA degradation, suggesting a stepwise regulatory process where miRNAs first inhibit translation and then facilitate mRNA decay.

In addition to the well-known mechanisms, miRNAs can also regulate gene expression through alternative pathways. For example, certain miRNAs have been found to stabilize mRNAs in specific conditions by binding to their 3′ UTRs. Although this mechanism is less common, it demonstrates the flexibility of miRNA-mediated regulation. Moreover, miRNAs can have an indirect effect on gene expression by influencing the expression or activity of transcription factors, signalling pathways, and other regulatory molecules. These indirect interactions can create feedback loops and crosstalk among various regulatory networks, enhancing the role of miRNAs in cellular functions.

The spatial and temporal dynamics of miRNA-mediated regulation play a crucial role in its effectiveness. miRNAs can target specific cellular compartments, like stress granules or processing bodies (P-bodies), which act as centers for mRNA silencing and degradation. These structures create a platform for the concentration and coordination of RISC complexes, enabling efficient regulation of target mRNAs. Furthermore, the expression levels and activity of miRNAs and their cofactors can be finely adjusted through post-transcriptional modifications, such as phosphorylation, ubiquitination, and methylation, which influence their stability and function. MiRNA-mediated gene regulation is crucial for maintaining cellular balance and adapting to changes in the environment. When miRNA expression or function is altered, it can disrupt these regulatory systems, resulting in diseases like cancer, neurodegenerative disorders, and immune issues. For example, the overproduction of oncogenic miRNAs (oncomiRs) can inhibit tumour suppressor genes, leading to increased cell growth and spread, while the absence of tumour-suppressive miRNAs can result in uncontrolled growth and survival of cancer cells. Likewise, abnormal miRNA activity has been linked to Alzheimer's disease, where it plays a role in the buildup of harmful proteins and nerve cell dysfunction.

In conclusion, miRNA-mediated gene regulation is a complex process that includes translational repression, mRNA degradation, and sometimes the stabilization of target transcripts. By fine-tuning protein expression, miRNAs serve as essential modulators of cellular processes, promoting adaptability and resilience in gene regulatory networks. Their capacity to integrate into various pathways and establish intricate feedback loops highlights their importance in both normal physiology and disease. Gaining a deeper understanding of the specific mechanisms behind miRNA-mediated regulation is vital for deciphering the complexities of gene expression and could lead to potential therapeutic strategies for conditions linked to miRNA dysregulation.

1.4.4. MicroRNA Target Recognition

MicroRNAs (miRNAs) play a crucial role in regulating gene expression by binding to specific sequences in target messenger RNAs (mRNAs). This process is facilitated by the mature miRNA strand that is part of the RNA-induced silencing complex (RISC). The recognition of targets mainly occurs through complementary base-pairing between the miRNA and its corresponding mRNA. However, this interaction is influenced by a variety of complex rules and factors that determine specificity, affinity, and the resulting regulatory effects. Gaining insight into the mechanisms of miRNA target recognition is vital for understanding their biological functions and clarifying their roles within intricate gene regulatory networks. The key factor in how miRNAs recognize their targets is the "seed region," which is a highly conserved sequence found at positions 2–8 from the 5′ end of the miRNA. This seed region is responsible for the initial base-pairing with the target mRNA, typically located in the 3′ untranslated region (3′ UTR). The degree of complementarity between the seed region and the target sequence is crucial for effective binding and regulatory function. There are different types of seed matches, including 6-mer, 7-mer, and 8-mer sites, with stronger interactions and regulatory capabilities as the match length increases. For instance, an 8-mer site features perfect pairing with the seed region and may include an additional match at position 1 or a supplementary base pair at the 3′ end, leading to the most significant repression. While the seed region lays the groundwork for target recognition, other factors influence the specificity and effectiveness of the interaction. In addition to the seed, miRNAs can create extra or compensatory base-pairing with

the target mRNA in the 3′ region of the miRNA. These interactions, commonly referred to as 3′ pairing, enhance the stability of the binding between the miRNA and the mRNA and can offset weaker seed matches. This adaptability enables miRNAs to regulate a wider array of targets, though with differing levels of efficiency. The context of the target site within the mRNA plays a significant role in how miRNAs bind and regulate gene expression. Various sequence and structural characteristics in the 3′ UTR affect how accessible and effective miRNA binding sites are. For example, sites found in areas with low secondary structure are generally more accessible to the RISC complex than those that are hidden within tightly folded RNA structures. Moreover, the closeness of the target site to the mRNA's poly(A) tail can enhance repression by promoting interactions with deadenylase complexes that lead to mRNA destabilization. Additionally, sites located in AU-rich regions are preferred, as these sequences usually form less stable secondary structures, which increases their accessibility.

Cooperativity among miRNAs plays a significant role in how they recognize their targets. Many mRNAs have several miRNA-binding sites, which enables different miRNAs to work together in regulating gene expression. When two or more miRNAs attach to nearby sites on the same mRNA, their combined influence can enhance gene silencing. This cooperative effect often relies on how the binding sites are arranged, with sites that are close together yielding the strongest synergy. On the other hand, competition between miRNAs and RNA-binding proteins (RBPs) for the same or adjacent binding sites can reduce miRNA effectiveness, adding another layer of complexity to the regulation of targets. While the 3′ UTR is the main area where miRNAs target, you can also find miRNA-binding sites in the coding sequence (CDS) and 5′ UTR of mRNAs. Although these sites are less frequent and generally not as effective as those in the 3′ UTR, they still play a role in miRNA-mediated regulation in certain situations. Targeting the coding sequence often involves non-canonical seed matches and depends on structural features that help RISC binding. On the other hand, 5′ UTR sites might be less accessible because of ribosome scanning during the start of translation, but they can still be important under specific cellular conditions or stress situations.

The recognition of miRNA targets is also shaped by post-transcriptional modifications occurring in both the miRNA and the mRNA. For instance, chemical changes like N6-methyladenosine (m6A) on the mRNA can either promote or hinder miRNA binding, depending on where they are located in relation to the target site. Likewise, modifications to the miRNA, such as 3′ adenylation or uridylation, can influence the stability and binding characteristics of the miRNA. These modifications contribute to a dynamic regulatory environment, allowing miRNA-mRNA interactions to be finely adjusted in response to cellular signals and environmental factors.

Recent findings indicate that miRNAs can also identify targets through non-canonical binding methods that differ from the usual seed region-based interactions. For instance, certain miRNAs exhibit extensive complementarity beyond the seed region, utilizing alternative areas of the miRNA to form stable pairings. These non-canonical interactions frequently take place in coding sequences or unusual regions of the mRNA, broadening the regulatory potential of miRNAs. Moreover, "pivot pairing," where a single mismatched nucleotide disrupts the standard seed match but is offset by pairing further downstream, has been noted in specific situations. These non-

canonical mechanisms underscore the adaptability of miRNA target recognition. The abundance and availability of miRNAs and their targets play a crucial role in how effectively these targets are recognized. miRNA expression levels differ across various tissues, developmental stages, and disease conditions, leading to unique regulatory profiles. Likewise, the levels and turnover rates of mRNA influence the pool of targets that miRNAs can bind to. The competition among mRNAs for the same miRNAs, known as the competing endogenous RNA (ceRNA) hypothesis, adds further complexity to this process. In this framework, non-coding RNAs like long non-coding RNAs (lncRNAs) and circular RNAs (circRNAs) function as miRNA sponges, capturing miRNAs and limiting their availability for other targets. This competitive interaction significantly affects the overall regulatory network that miRNAs mediate. The dynamics of how miRNAs recognize their targets are also influenced by cellular compartmentalization. miRNAs and their targets can be found in specific subcellular areas, such as processing bodies (P-bodies) or stress granules, where the silencing and degradation of mRNA are concentrated. Being localized in these structures boosts the efficiency of miRNA-mediated repression by creating microenvironments that are rich in regulatory components. Additionally, the localization of miRNAs to distinct cellular compartments, like synapses in neurons or the leading edge of migrating cells, allows for spatially restricted regulation of target mRNAs, enabling precise control over localized gene expression.

In conclusion, the process of miRNA target recognition is complex and involves various factors, including sequence complementarity, structural accessibility, cooperative interactions, and the dynamic nature of cellular environments. While the seed region is the main factor determining target specificity, many other elements affect the stability, efficiency, and results of miRNA-mRNA interactions. These intricate mechanisms allow miRNAs to precisely regulate a wide range of gene networks, playing essential roles in development, homeostasis, and disease. Ongoing research into the details of miRNA target recognition is likely to uncover new insights into the fundamental principles of gene regulation and may open up potential therapeutic options for disorders related to miRNAs.

1.4.5. Plant microRNAs

Plant microRNAs (miRNAs) play essential roles similar to those of mammalian miRNAs in regulating gene expression, but they also have several important differences in their structure, biogenesis, target recognition, and mechanisms of action. These variations highlight the evolutionary differences between plants and animals, as well as the distinct regulatory needs of plant biology. Gaining a deeper understanding of plant miRNAs is crucial for uncovering their specific functions in growth, development, stress responses, and adaptation to environmental changes.

The biogenesis of plant miRNAs shares some fundamental steps with the canonical pathway found in mammals, but it utilizes different enzymes, processing sites, and regulatory elements. In plants, miRNAs are transcribed by RNA polymerase II into primary miRNAs (pri-miRNAs), which feature a cap structure at the 5′ end and a poly(A) tail at the 3′ end. The processing of pri-miRNAs in plants takes place in the nucleus through a single-step cleavage facilitated by DICER-LIKE1 (DCL1), a plant-specific counterpart of the Dicer enzyme seen in animals. DCL1 operates within a nuclear complex that includes accessory proteins like HYPONASTIC LEAVES1 (HYL1) and SERRATE (SE). This complex identifies the stem-loop structure of pri-

miRNAs and accurately excises the mature miRNA duplex, eliminating the need for a separate Drosha-like enzyme, which is essential in the biogenesis of miRNAs in mammals.

Plant miRNAs have several structural differences compared to mammalian miRNAs. For instance, plant pri-miRNAs, which are the primary transcripts that mature miRNAs are derived from, tend to be longer and more diverse in structure than their mammalian counterparts. They can create complex secondary structures with multiple stem-loop regions. Typically, these pri-miRNAs encode a single miRNA within one hairpin, although some can produce multiple miRNAs. Additionally, plant pre-miRNAs, which result from the processing of pri-miRNAs, often vary in length from 60 to 200 nucleotides, while mammalian pre-miRNAs are usually around 70 nucleotides long. The mature miRNAs found in plants are highly conserved across different species and generally measure 20–22 nucleotides in length, much like mammalian miRNAs, but they frequently exhibit more perfect or near-perfect sequence complementarity with their targets.

In contrast to mammals, the process of miRNA biogenesis in plants does not depend on Exportin-5 for the nuclear export of pre-miRNAs. Instead, the mature miRNA/miRNA* duplex is directly transported to the cytoplasm by HASTY, which is a homolog of Exportin-5. Furthermore, before export, plant miRNA duplexes are frequently methylated at their 3′ ends by the RNA methyltransferase HUA ENHANCER1 (HEN1). This methylation serves to prevent uridylation and degradation of miRNAs, acting as a protective mechanism that is less common in mammalian systems. After export, the mature miRNA strand is integrated into the RNA-induced silencing complex (RISC), where ARGONAUTE1 (AGO1) functions as the main effector protein in plants. Plant AGO1 has slicer activity that enables the cleavage of target mRNA, which is a mechanism that differs from the translational repression typically seen in mammals. The guide strand is preferentially loaded into AGO1 based on the thermodynamic stability of the duplex ends, a selection process that is also observed in mammals.

The way plant miRNAs recognize their targets is significantly different from the mechanism in mammals, primarily due to the high level of sequence complementarity between plant miRNAs and their targets. Typically, plant miRNAs bind to their target mRNAs with nearly perfect or perfect complementarity, often within the coding sequence instead of the 3′ untranslated region (3′ UTR), which is where mammalian miRNAs usually target. This strong complementarity enables plant miRNAs to directly trigger mRNA cleavage through the endonucleolytic activity of AGO1, a process similar to RNA interference (RNAi) in mammals, though it is uncommon for mammalian miRNAs. This precise targeting mechanism leads to effective and specific regulation of target mRNAs in plants, resulting in minimal off-target effects compared to the broader and more varied targeting observed in mammalian systems. Another key aspect of plant miRNA function is the conservation of their binding sites. In plants, miRNA target sites are often highly conserved across different species, indicating evolutionary pressure to preserve these regulatory networks. On the other hand, mammalian miRNA target sites tend to be less conserved and more diverse, which allows for greater flexibility in gene regulation. The conservation of plant miRNA binding sites supports strong regulation of essential genes that are crucial for important biological processes such as development, hormone signalling, and

responses to stress. In addition to cleavage, plant miRNAs can also regulate their targets through translational repression, although this mechanism is not as common as in mammals. Recent research has shown that plant miRNAs can inhibit translation by disrupting ribosome assembly or elongation, often alongside mRNA cleavage. This dual role of plant miRNAs in both mRNA cleavage and translational repression underscores their adaptability in managing gene expression. Plants also have distinct regulatory networks that involve miRNAs and their targets. For instance, many plant miRNAs target transcription factors that influence developmental processes like leaf shape, flowering time, and root growth. By adjusting the expression of these transcription factors, miRNAs create complex feedback loops and help coordinate intricate developmental programs. Additionally, plant miRNAs are essential in responding to both abiotic and biotic stresses. They regulate genes that are part of pathways related to drought resistance, nutrient absorption, and defence against pathogens, allowing plants to adjust to changes in their environment.

An interesting feature of plant miRNA function is the ability of certain miRNAs to act as long-distance signalling molecules. These mobile miRNAs can be transported between cells and tissues via plasmodesmata or the phloem, enabling systemic regulation of gene expression. For example, miR399, which responds to phosphate starvation, travels from shoots to roots to help regulate phosphate homeostasis by targeting PHO2 mRNA. This long-distance signalling mechanism is specific to plants and highlights the unique roles of miRNAs in coordinating physiological responses throughout the organism. Post-transcriptional modifications and regulatory mechanisms also set plant miRNA function apart from that of mammals. For instance, plant miRNAs undergo differential processing and stability regulation in response to environmental or developmental signals. Proteins like DAWDLE (DDL) and TOUGH (TGH) play a role in influencing the stability and efficiency of miRNA biogenesis complexes, thereby modulating miRNA levels based on context. Furthermore, plant miRNAs can create regulatory circuits with small interfering RNAs (siRNAs) and other non-coding RNAs, which further enhances their regulatory capabilities.

In conclusion, plant miRNAs have unique structural, biogenetic, and functional traits that set them apart from mammalian miRNAs. Their longer and more varied precursor structures, dependence on DCL1 for processing in the nucleus, strong complementarity to their targets, and a tendency for mRNA cleavage showcase the distinctive aspects of plant miRNA pathways. These differences reflect the specific regulatory requirements of plants, allowing them to manage complex developmental processes, adapt to environmental stresses, and maintain homeostasis. Gaining a deeper understanding of how plant miRNAs work and their functions can provide important insights into plant biology and open up possibilities for enhancing crop resilience and productivity through miRNA-based approaches.

1.4.6. miRNA drugs in development

The detailed exploration of miRNA biology – from their precise biogenesis to their network-level regulation of gene expression – naturally leads to a compelling question: can we harness these molecules as medicines? As detailed in the preceding chapters, the dysregulation of specific miRNAs is a hallmark of numerous diseases, including cancer, where they can act as potent oncogenes or tumour suppressors. This central role, combined with the ability to theoretically modulate any gene pathway by targeting (a) miRNA(s), has fueled research into translating miRNA biology into clinical

therapeutics. This subchapter highlights the current landscape of miRNA-based drug development, highlighting key candidates, delivery strategies and the significant challenges that must be navigated to bring these interesting small molecules from the laboratory to the clinic.

The therapeutic strategies themselves build directly on the mechanisms described earlier. MiRNA mimics are synthetic double-stranded RNAs designed to mimic the endogenous mature miRNA. Once delivered into the cell, they are loaded into RISC to restore the lost function of a tumour-suppressor miRNA. Conversely, miRNA inhibitors (often called antagomiRs, or anti-miRs) are single-stranded, chemically modified antisense oligonucleotides. They work by binding tightly to the mature oncogenic miRNA, sequestering it and preventing it from interacting with its target mRNAs, thereby blocking its harmful activity. While the design is straightforward, the challenge lies in delivery and specificity. As highlighted in the discussion of miRNA target recognition, a single miRNA can influence hundreds of genes. Therefore, achieving targeted delivery to diseased tissue is critical to minimize off-target effects and potential toxicity.

The clinical landscape for miRNA drugs is still emerging, particularly for aggressive cancers like glioblastoma. Reflecting the preclinical focus of much research, there are currently no FDA-approved miRNA drugs on the market for any cancer type, and trials specifically for GBM are generally in early phases or have been terminated. One promising candidate for GBM was RGLS5579 (Table 1.2), an inhibitor targeting the pro-invasive miR-10b. It reached late preclinical development with plans for a Phase I clinical trial, though its current recruiting status is now unclear. The broader field, however, offers instructive examples. For blood cancer, Cobomarsen (MRG-106) (153), an inhibitor of miR-155, however promise in a completed Phase I trial for certain lymphomas, though its Phase II development was later terminated for business reasons, not due to safety or efficacy. In solid tumours, the TargomiRs (MesomiR1) (154) platform, which delivers a miR-16 mimic using engineered minicells, completed a Phase I trial in malignant pleural mesothelioma.

These efforts have also faced serious setbacks that highlight the challenges of this new drug class. The most notable case is MRX34, a liposomal nanoparticle-formulated miR-34a mimic (155). Despite preclinical data supporting its tumour-suppressor role in various solid tumours, its Phase I trial was terminated due to severe immune-related adverse events. This event highlighted the critical importance of refining delivery systems to improve tolerability. In response, next-generation candidates are employing advanced technologies. For example, TTX-MC138, a miR-10b inhibitor for metastatic solid tumours, is in Phase I/II trials utilizing a targeted, brain-penetrating nanoparticle delivery system (156). Likewise, INT-1B3, a miR-193a-3p mimic for advanced solid tumours, is undergoing Phase I evaluation (157). These efforts represent a concerted push to enhance tumour-specific delivery and pharmacokinetics.

Overcoming the BBB for glioblastoma therapy remains a paramount challenge, driving innovation in preclinical settings. Strategies such as convection-enhanced delivery, brain-penetrating nanoparticles coated with targeting ligands, and the use of focused ultrasound with microbubbles to temporarily disrupt the BBB (158) are being actively investigated to enable efficient drug accumulation in the tumour bed. Parallel to these therapeutic endeavours, the clinical utility of miRNAs as non-invasive biomarkers for

diagnosis, prognosis, and treatment monitoring in glioblastoma and other cancers has become a robust and distinct area of translational research, offering a more immediate path for miRNAs to impact patient care.

Drug name	Target miRNA	Mode of Action	Cancer Type(s)	Status	Clinical Trial Identifier(s)
RGLS5579	miR-10b	Inhibitor	glioblastoma	Plans for Phase I	
Cobomarsen (MRG-106)	miR-155	Inhibitor	Lymphomas / leukemias (eg. CTCL DLBCL)	Phase I completed; Phase II terminated (for business reasons)	NCT03837457 NCT02580552 NCT03713320
TargomiRs (MesomiR 1)	miR-16	Mimic	Malignant Pleural Mesothelioma, Non-Small Cell Lung Cancer	Phase I	NCT02369198
MRX34	miR-34a	Mimic	Various solid tumours (eg. lung, liver)	Phase I terminated due to immune-related toxicities	NCT01829971
TTX-MC138	miR-10b	Inhibitor	Advanced solid tumours	Phase I/II (active)	NCT06260773 NCT06260774
INT-1B3	miR-193a-3p	Mimic	Solid tumours	Phase I (terminated due to insufficient funding)	NCT04675996

Table 1.2: Clinical trials of miRNA therapeutics for various cancer types. NCT numbered trials are registered at ClinicalTrials.gov.

In conclusion, the development of miRNA-based therapeutics is a dynamic and evolving frontier, built directly upon the foundational science of miRNA biogenesis and function. While the path has proven more treacherous than initially hoped, with clinical progress tempered by delivery and safety challenges, the strategic lessons learned from early trials are catalysing the design of more sophisticated, target agents. The ongoing preclinical and clinical work – spanning mimics, inhibitors, and novel delivery approaches – continues to test the hypothesis that modulating these master regulatory networks can yield potent and specific anti-cancer effects. Success will likely depend on achieving the delicate balance between leveraging the broad regulatory capacity of miRNA and constraining their activity to the precise pathological context, thereby turning their biological complexity from a translational hurdle into a therapeutic asset.

This chapter has described miRNAs as important regulators of gene expression and their profound implications in disease and therapy. Beginning with the fundamental roles of RNA, the evolution of RNA was traced from viewing RNA merely as a messenger to recognizing its diverse regulatory functions, culminating in the discovery and characterization of non-coding RNAs, particularly miRNAs. The detailed examination of miRNA biogenesis revealed the precision of the canonical pathway – from Pol II transcription and Drosha/DGCR8 processing in the nucleus to Exportin-5-mediated export and Dicer/Ago2 maturation in the cytoplasm – alongside the flexibility afforded by non-canonical pathways like mirtrons. The mechanisms by which mature miRNAs, loaded into RISC, post-transcriptionally regulate gene expression were explored through target mRNA destabilization and translational repression, a process governed by the intricate rules of seed-based target recognition and influenced by cellular context. Further highlighting the versatility of these molecules, a comparison

with plant miRNAs underscored how conserved core principles are adapted to meet specific organismal needs, particularly through near-perfect complementarity leading to direct mRNA cleavage. However, the key takeaway from this chapter should be that the very properties which make miRNAs powerful endogenous regulators – their ability to fine-tune vast gene networks – also presents unique challenges and opportunities for therapeutic intervention, setting the stage for the miRNA drugs in current clinical trials to harness this complex biology.

1.5. Computational Biology

Computational biology or bioinformatics, in the latest (4th) edition of Bioinformatics An Introduction, is defined as “the science of how information is generated, transmitted, received, stored, processed and interpreted in biological systems” or more succinctly, “the application of information science to biology” (159). As it is a rapidly developing field, especially in the last couple decades, many branches have evolved, such as genomics, microbiomics, exposomics and regulation Figure 1.8.

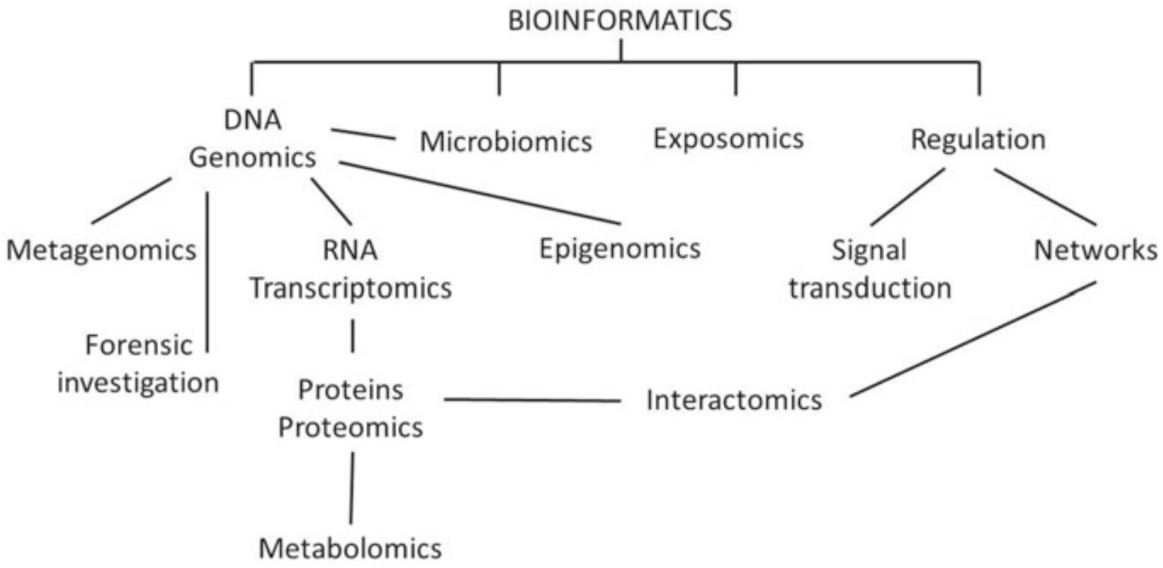


Figure 1.8: A (partial) ontology for bioinformatics from Bioinformatics by Ramsden (6)

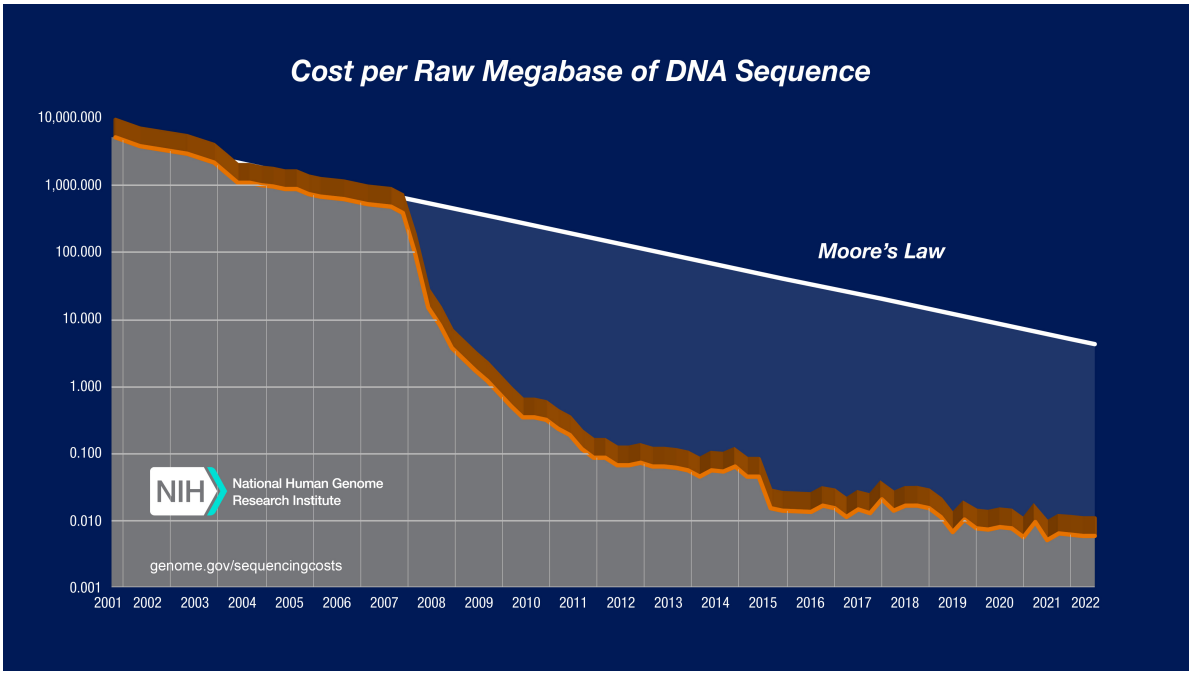
In this chapter, I will explore bioinformatic approaches and tools that were relevant for my project.

1.5.1. Omics

A major catalyst of the rapid advancement of bioinformatics was the development of sequencing technologies. The first sequenced complete genome was Macteriophage MS2 virus in 1976, facilitated by Sanger sequencing methods of the time (160). Following the success, scientists took on more and more complex genomes like chromosome III of yeast in 1993 (161), the entire genome of Haemophilus influenzae in 1995 , and other bacteria with small genomes. It was clear that in order to sequence species with bigger genomes (like eukaryotes), more sophisticated sequencing was needed. The introduction of Shotgun sequencing has allowed the sequencing of the first unicellular eukaryote (Saccharomyces cerevisiae) in 1996 , then the first multicellular eukaryote (Caenorhabditis elegans) in 1998 (162). By the early 2000s, the human chromosome 22 (the shortest autosome) (163), the first insect (Drosophila

melanogaster) (164), first plant (Arabidopsis thaliana) (165), and first mammal (Mus musculus) (166) were sequenced. Finally, the Human Genome Project published an incomplete version of the human genome in 2004 (167).

Genomic sequencing has profoundly affected all areas of biology, from population genetics to phylogenetics through to biochemistry and immunology. Thanks to economic phenomenon of ‘supply and demand’ together with Moore’s Law – cost will decrease based on slow improvements – have decreased the cost of sequencing, making even more accessible (Figure 1.9A). The sharp decrease in Figure 1.9 can be attributed to the introduction of next (or second) generation sequencing (NGS, or also known as massively parallel sequencing or deep sequencing), which reduced sequencing time of the human genome to only a few hours compared to the ten years it originally took, resulting in lower cost per raw megabase as well (Figure 1.9B) (6).



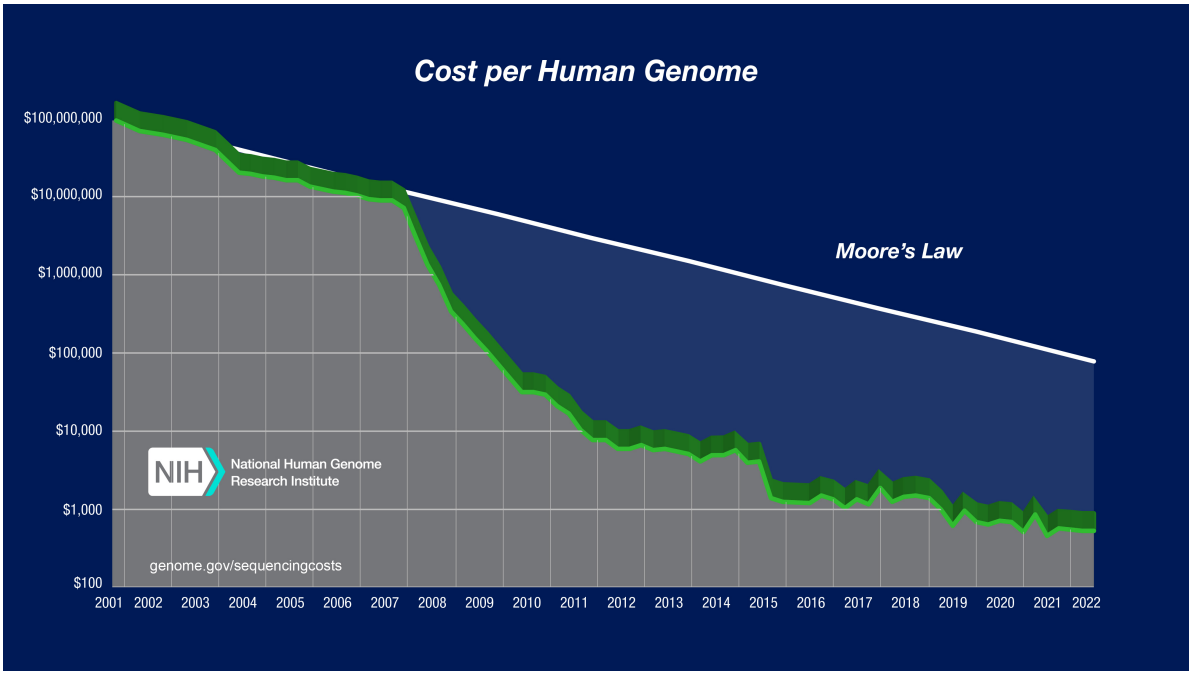


Figure 1.9: Cost of sequencing (A) Cost per Raw Megabase of DNA Sequence; (B) Cost per Human Genome. (7)

Since its first introduction in 2005 (6), many NGS approaches have been developed, some more lasting than others. These slightly differ in attributes such as sequencing read length, cost and data output.

1.5.1.1. Next-Generation Genomics

Generally, in NGS, the process begins with DNA fragmentation followed by DNA end-repair, adapter ligation, surface attachment, and in-situ amplification. The process mostly differs in read length when it comes to various platforms offering sequencing

services. Overall, the most widely used platforms nowadays are Illumina (short-read of second-generation sequencing, or Oxford Nanopore and Pacific Biosciences (PacBio) (long -read or third-generation sequencing).

“Short-read”

Second generation sequencing, otherwise known as “short-read” sequencing, was the technology that introduced massively parallel sequencing, which decreased cost, time and simultaneously making it more accessible to the general public. Short-read sequencing allows the parallel sequencing of 250-800bp following adequate library preparation and sequencing. Good quality library preparation is a pre-requisite of NGS. The procedure slightly differs depending on whether the sequencing is done on DNA (DNAseq) or RNA (RNAseq). There are various types sequencing strategies available depending on the desired outcome; for DNAseq it can include: Whole Genome Sequencing (WGS), Whole Exome Sequencing (WES), Epigenome Sequencing and Targeted Sequencing (TS); for RNAseq it can include: Whole Transcriptome Sequencing (WTS), mRNA sequencing (mRNA-Seq) and small RNA sequencing (smRNA-Seq). For DNAseq, template library preparation can include two approaches: polymerase chain reaction (PCR) or hybridization capture-based approaches. Library construction mainly involves the Library construction mainly involves fragmentation, end-repair, adaptor ligation, and size selection (168).

For RNAseq, sample preparation usually constitutes of three steps: total RNA isolation, target RNA enrichment, and reverse transcription of RNA into complementary DNA (cDNA) (169). In the first step, total RNA isolation serves the purpose of extracting only RNA molecules while minimizing contamination from DNA, proteins, and other cellular components. This step is crucial for obtaining high-quality RNA that accurately represents the biological context. Following isolation, target RNA enrichment is performed to focus on specific RNA types, such as mRNA, by removing ribosomal RNA (which constitutes the bulk of total RNA) using methods like poly-A selection or ribosomal RNA depletion. This enrichment ensures that the sequencing effort is concentrated on the RNA species of interest, increasing the efficiency and effectiveness of the analysis. Finally, the enriched RNA is subjected to reverse transcription, where the enzyme reverse transcriptase synthesizes complementary DNA (cDNA) from the RNA temple. This conversion is essential, as DNA, rather than RNA, is more stable and compatible with sequencing technologies. These steps collectively ensure that the resulting cDNA library is representative of the original RNA population, enabling accurate downstream analysis of gene expression and transcriptome profiling (170, 171).

One of the primary advantages of short-read sequencing is its accuracy in detecting single nucleotide variants (SNVs) and small insertions or deletions, indels. These genetic changes often underlie many diseases, including cancer and inherited disorders. By offering a deep sequencing depth, short-read technologies ensure that even low-frequency variants are identified with confidence. The speed and scalability of short-read sequencing have made it a preferred method for a variety of applications. In genomics, the ability to sequence an entire genome in a matter of days at a relatively low cost has fueled large-scale projects like the 1000 Genome Project. In addition, to whole-genome sequencing, this technology also excels in targeted sequencing and exome sequencing. Targeted sequencing focuses on specific areas of interest within the genome, such as disease-associated genes. This approach is often used in clinical

settings for its efficiency in identifying mutations linked to particular conditions. Exome sequencing, which captures and sequences on the protein-coding regions of the genome, allows for the identification of functional genetic variants that can directly impact protein development and/or function, leading to changes in phenotype. Moreover, RNAseq provides insight into gene expression levels, alternative splicing events, and the presence of novel transcripts. This can highlight active processes in cells particularly in response to condition changes such as stress, development, or disease.

Despite these advantages, short-read sequencing has limitations, particularly in resolving repetitive regions and large structural variants, The short length of reads can create challenges in accurately assembling genomes, especially in regions with complex repetitive sequences. This limitation can lead to gaps or errors in genome assemblies, affecting the analysis of structural variations that may be important in multiple genetic conditions. Many approaches are being developed to improve upon this technology, but one drastic change to the field of sequencing was the introduction of long-read sequencing.

“Long-read”

Long-read next-generation sequencing can generate much longer continuous sequences, roughly around 10,000 base pairs, significantly exceeding short-read technologies’ 50-300 base pair sequences. Companies such as Pacific Biosciences and Oxford Nanopore were the trailblazers of this new approach, which opened the doors for complex genomic investigations, enabling more accurate mapping of repetitive regions, structural variations and previously inaccessible genomic areas.

Long-read technologies excel in resolving structural variants – large genomic alterations that can involve duplications, deletions, inversions or translocations of DNA segments. Structural variants are a major source of genetic diversity and have been implicated in a range of diseases, including cancer, neurological disorders, and developmental abnormalities. This technology addressed head on short-read limitations of dealing with repetitive regions. One of the most significant advantages of long-read sequencing is its ability to improve the assembly of new genomes, The longer reads can span repetitive regions that confound short-read assemblers, leading to more complete and contiguous genome assemblies. This is particularly valuable in de novo, sequencing projects where no reference genome is available. Moreover, long-read sequencing enhances our understanding of epigenetic modifications. Some technologies can detect DNA methylation patterns during the sequencing process itself, offering a simultaneous view of the genetic and epigenetic landscape. In clinical genomics, the precision of long-read sequencing is proving to be very useful. Rare genetic diseases, often caused by intricate and unique genetic variations, can be better diagnosed with the comprehensive view provided by long reads.

Despite its advantages, there are limitations to this technology as well. This technology is currently more expensive than short-read sequencing, although costs are expected to decrease over time as per Moore’s Law. Furthermore, the error rates for individual reads tend to be higher; however, this drawback is often mitigated by improvements in computational algorithms and the sheer length of reads, which enable accurate error corrections through consensus sequencing. Sometimes short-read sequencing can aid in this correction process as well.

In conclusion, next-generation genomics has dramatically transformed (and continuously transforming) the landscape of genetic research, offering unprecedented insights into complex biological systems. Short-read sequencing technologies have provided a foundation with their high throughput and accuracy in detecting single nucleotide variants and small indels, playing a crucial role in large-scale genomic projects and clinical diagnosis. Meanwhile, long-read sequencing has emerged as a powerful complement, enabling the resolution of complex genomic regions and structural variants, while enhancing de novo genome assembly and epigenomic studies. Additionally, RNA sequencing has advanced our understanding of transcriptomes through careful sample preparation, encompassing RNA isolation, enrichment and reverse transcription. Together these technologies have opened new avenues in life sciences and medicine, underscoring the potential for ongoing innovations to further deepen our knowledge of the genetic basis of life. As costs continue to decrease and methodologies improve, the integration of these sequencing approaches promises a future where comprehensive and precise genomic analysis is accessible for diverse research and clinical applications.

1.5.1.2. Proteomics

Proteomics refers to the large-scale study of proteins, which are the key components of cellular processes and functions such as catalysing metabolic reactions, DNA replication, and signal transduction. While genomics and transcriptomics provide insights into potential protein expression, proteomics actually measures the proteins levels of a cell at any given time, reflecting both translational regulation and post-transcriptional modifications. Building on knowledge gained from DNAseq and RNAseq, proteomics helps bridge the gap between the static genetic code and the dynamic protein expressions that drive cellular function. Genomic and transcriptomic data provide a blueprint, but proteomics explores the actual implementation of these blueprints, taking into account regulatory mechanisms and environmental influences that affects protein expression and activity.

Some of its advantages include the comprehensive analysis of protein abundances, modifications interactions and localization within complex biological settings. This allows researchers to gain insights into cellular function, disease mechanisms, and potential therapeutic targets. Moreover, proteomics can identify biomarkers for diseases, leading to improved diagnostics and personalised medicine approaches. However, as any other technology, it has its unique limitations. The complexity and dynamic range of the proteome make it difficult to analyze; proteins exist in a wide range of concentrations and can undergo post-translational modifications that are challenging to detect. Additionally, the sheer number of possible protein isoforms, resulting from alternative splicing and modification, adds another layer of complexity that is not yet fully detectable with current technologies. Proteomics can address a variety of crucial questions, such as identifying the protein composition of cells under different conditions, understanding protein interactions and pathways in disease states, and exploring the impact of genetic variations on protein expression. This can provide valuable insights into disease mechanisms, drug responses, and potential therapeutic targets.

A leading technology of proteomics is mass spectrometry (MS), which has become an indispensable tool for protein identification and quantification. MS works by ionizing protein molecules and measuring their mass-to-charge ratios, allowing for the precise

determination of protein masses. This technology can identify thousands of proteins in a single experiment, providing a comprehensive proteomic profile. Advancements in MS, such as tandem mass spectrometry (MS/MS), have enhanced its ability to sequence and quantify proteins, even in complex mixtures. High resolution MS coupled with chromatography techniques, such as liquid chromatography-mass spectrometry (LC-MS), has significantly improves the sensitivity and accuracy of protein detection. MS-based proteomics can also capture post-translational modification like phosphorylation, ubiquitination, and glycosylation, which are critical to understanding protein function and regulation. Despite these advances, challenges remain, including the need for specialized equipment and expertise, high costs, and the complexity of data analysis. However, ongoing developments in technology and computational methods are steadily overcoming these obstacles.

In summary, proteomics is an important field that extends the insights provided by genomics and transcriptomics into functional biological understanding. By leveraging powerful tools like MS, proteomics has unlocked the potential to analyze proteins comprehensively, offering profound implications for biology, medicine and drug discovery.

1.5.1.3. Imaging

Imaging technologies are crucial in biological research, providing visual insights into the complex architecture and functioning of tissues and cells. Among these techniques, Immunohistochemistry (IHC) staining stands out as a fundamental method for examining protein expression within tissue sections. IHC uses antibodies to detect specific antigens in cells, allowing researchers to visualize the distribution and localization of proteins, which is essential for understanding cellular functions and diagnosing diseases, such as cancer. The analysis of IHC-stained slides has become more sophisticated with digital pathology, particularly through the use of software like QuPath. QuPath is an open-source platform (developed in Edinburgh) that facilitates the annotation and analysis of whole-slide images. It enables quantitative analysis of tissue sections, offering tools to delineate regions of interest, quantify protein expression, and perform complex image analysis tasks with ease. This advancement significantly enhances the efficiency and accuracy of pathological assessments, enabling more detailed and reproducible studies.

The field of imaging is rapidly evolving towards more integrated approaches, such as spatial transcriptomics. This cutting-edge technology combines imaging with genomics, allowing for the visualization of gene expression patterns within the spatial context of tissues. By overlaying spatial gene expression data with IHC images, researchers can gain insights into the molecular underpinnings of tissue organization and disease states. This approach bridges the gap between traditional histopathology and molecular biology, providing a comprehensive understanding of cellular function and interactions within the native tissue environment. Despite the advances in imaging technologies, there are challenges to consider. Traditional imaging methods still rely heavily on the expertise of pathologists for interpretation, which can introduce variability. However, digital tools like QuPath are helping to standardize and automate these analyses, reducing subjective bias and improving diagnostic precision.

In summary, imaging technologies offer another layer to the analysis pipeline of assessing cellular processes. IHC staining remains a staple for tissue analysis,

complemented by digital tools like QuPath for enhanced data extraction and analysis. The emerging field of spatial transcriptomics represents a significant leap forward, offering insights that intertwine cellular and molecular landscapes in unprecedented detail. As these technologies advance and integrate, they promise to deepen our understanding of complex biological systems and disease mechanisms, ultimately improving diagnostic and therapeutic approaches.

1.5.2. Transcriptomic platforms

As RNAseq costs have been decreasing and, therefore, becoming more accessible, the need for skilled bioinformaticians have also increased to correctly interpret the data. One of the most popular tools using Nextflow pipelines for quantification, then using DESeq2 for downstream analysis.

However, it is unlikely that all labs around the world are able to hire or access reliable bioinformaticians. Therefore, it is still a challenge for researchers, specifically those lacking any programming background, to sufficiently analyse data. Therefore, the need for cost-effective, dependable, and easily understandable software has also increased (172). One of these software is the QIAGEN CLC Genomics Workbench (CLC) (159) that is capable of analysing a variety of genomics/transcriptomics related problems in a user-friendly manner such as: whole genome and transcriptome de novo assembly, targeted resequencing analysis, variant calling, ChIP-seq and DNA methylation (159, 172). In this subchapter, I will further explain the underlying mechanisms of CLC and Nextflow-DESeq2 approaches, as well as their advantages and disadvantages.

1.5.2.1. CLC Genomics Workbench

The audience, which is normally non-bioinformatics specialists, that CLC caters for requires no further detail into the underlying algorithmic approach to the analyses resulting in a ‘black box’ like tool. Some parameters are modifiable, which provides sufficient customizability for most standard experiment designs. However, when it comes to comparing to more computationally accessible tools; this ‘black box’ status proves unideal. QIAGEN only provides (in some places) a vague manual. In contrast, the advantage to CLC compared to the Nextflow-DESeq2 approach is that only tool is needed for the entire analysis. The raw sequencing files

CLC provides a number of predetermined workflows and analyses including de novo assembly, read mapping, variant analysis, RNAseq, epigenomic studies, and metagenomics.

Read Mapping and Alignment
CLC supports high-throughout read mapping for aligning sequencing reads to reference genomes. This is a fundamental step for many genomic analyses, such as variant detection and transcriptomics. The software employs a seed-and-extend algorithm, which first identifies short exact matches (seeds) and then extends the alignment using a scoring scheme for mismatches, insertions and deletions. The implementation is highly optimized for speed and memory efficiency, making it suitable for large-scale datasets. It also supports paired-end and single-end read mapping for data derived from technologies like Illumina and PacBio.

De novo Assembly

For projects without a reference genome, CLC offers de novo assembly workflows. It uses de Bruijn graph algorithms, which break reads into k-mers and construct a graph where each k-mer represents a node. Overlapping k-mers are connected by edges, and the graph is traversed to reconstruct contigs. The platform incorporates strategies for resolving ambiguities caused by repetitive regions or sequencing errors, enabling the assembly of high-quality contigs. Applications include genome assembly for non-model organisms and novel pathogen discovery.

Varian Detection and Analysis
Variant analysis workflows allow researchers to identify single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and structural variants. Reads are mapped to the reference genome, and variants are detected using a Bayesian model to calculate the likelihood of a variant at each genomic position. Filters can then be applied based on read depth, quality scores, and variant frequency. The software also includes tools for annotating variants, using integrated or external databases, to determine their potential functional effects.

RNAseq and Transcriptomics
RNAseq workflows in CLC provide comprehensive tools for quantifying gene expression, identifying differential expression and discovering novel transcripts. Read mapping for RNAseq data uses splice-aware algorithms, which recognise exon-intron boundaries and align reads across splice junctions. For expression analysis, reads are assigned to transcripts or genes, and expression levels are quantified as fragments per kilobase of transcript per million mapped reads (FPKM) or transcripts per million (TPM). Differential expression analysis is performed using statistical models such as the negative binomial model (similar to what DESeq2 uses), accounting for variability and overdispersion in RNAseq data.

Epigenomics and Methylation Analysis
The platform also supports workflows for epigenomic studies, including bisulfite sequencing analysis to investigate DNA methylation. Bisulfite-treated reads are aligned to the reference genome using specialized alignment algorithms that account for the C-to-T conversions indicative of methylated cytosines. Methylation levels are quantified at single-base resolution, and regions of differential methylation (DMRs) can be identified using statistical tests.

ChIP-Seq and Peak Calling
For chromatin immunoprecipitation sequencing (ChIP-seq), the software includes workflows for identifying protein-DNA interaction sites. Reads are mapped to the reference genome, and peaks are identified using algorithms such as MACS2 (Model-based Analysis of ChIP-Seq). This method models the background noise and identifies significant enrichment regions where the DNA-binding protein of interest is likely bound.

Metagenomics and Microbial Analysis
The metagenomics workflows in CLC Genomics Workbench allow for taxonomic profiling and functional annotation of microbial communities. The software supports k-mer-based classification for rapid taxonomic assignment and abundance estimation of microbial species in metagenomic samples. Functional annotation can be performed by mapping reads to databases like KEGG or GO terms, enabling the identification of pathways active in the microbial community.

CRISPR and Targeted Editing Analysis
CRISPR analysis workflows are specifically designed to assess the efficiency of genome editing experiments. The software aligns reads to the reference genome and identifies on-target and off-target edits. It calculates editing efficiency by quantifying the proportion of reads with mutations at the target site.

Overall, CLC offers a comprehensive suite of workflows tailored to diverse NGS applications. Its robust algorithms, user-friendly interface and integration options make it a powerful tool for wet-lab researchers. However, an educated comparison is difficult when the algorithms are not open access. In each workflow algorithms seem to be using the same bases as other open access workflows; however, as results differ, a conclusion must be made that somehow these common algorithms are customised to this specific tool, but due to the 'black-box' nature of the programme, I am unsure how.

1.5.2.2. Nextflow and DESeq2

Using computational approaches that require coding knowledge could be a limitation for many "wet lab" scientists. However, developing the analysis pipeline provides a unique control and insight into the data. The modifiable parameters are virtually limitless. As much as this is an advantage for programmers, it causes concerns for reproducibility. To combat this, many workflow managers (173) have been developed, one of which is Nextflow (174).

As per Nextflow's documentation (175), it is "a workflow system for creating scalable, portable, and reproducible workflows." The foundation of the Nextflow language finds its roots in the Unix philosophy, where numerous simple operations can be used like building blocks to build progressively more complex tasks (175). Therefore, Nextflow is also capable to execute many smaller tasks, sometimes consisting of different scripting languages, as part of a bigger pipeline (174, 175). Among its many useful features, Nextflow seamlessly integrates with a variety of widely used execution platforms, including High Performance Computing (HPC) schedulers and cloud providers, as well as popular software tools such as Git (176), Docker (177), and Conda (178). This integration allows the user to comprehensively define a computational pipeline, complete with all its required dependencies, and execute it in almost any computing environment (175) – increasing portability, standardisation, reliability and reproducibility.

Transcriptomics being one of the main disciplines of bioinformatics (among genomics, and proteomics), many analyses have evolved into a 'standard' workflow, which many bioinformaticians repeat, then tweak depending on their data. Programmers tend to refrain from reinventing the wheel, meaning instead of every single bioinformatician writing their pipelines from scratch, they form a community to accelerate the progress, reliability, and efficiency of workflows. Thus, the nf-core framework was created, which is a community-curated pipeline 'library' around the Nextflow programming language (179). As this effort is community driven, theoretically, the pipelines are more likely to be updated with new approaches/packages/updates than privately owned and maintained pipelines. To further highlight the issue, the field of bioinformatics is incredibly rapidly growing, where updates to services could be available as soon as on a monthly basis. As an individual bioinformatician uses more and more packages for any given project, the more difficult it becomes to keep up with updates and

compatibility checks. Hence, nf-core is a useful and reliable effort for the Nextflow community.

Even though, the pipelines are community driven, they are all very well documented according to nf-core's best-practises, which go through an approval process that ensures a pipeline works correctly with appropriate error messages. Somewhat unsurprisingly, the most popular pipeline is RNAseq that can be used to analyse RNA sequencing data obtained from organisms with a reference genome and annotation (180). As this is the pipeline I used in my project, I will now explain in further detail how the pipeline functions.

The input for the nf-core RNAseq pipeline is a 'samplesheet', that states the file names and additional information about the experiment, and FASTQ files. The workflow consists of 5 stages: (1) Pre-processing; (2) Genome alignment and quantification; (3) Pseudo-alignment and quantification; (4) Post-processing; and (5) Final quality control (QC) (180) . The final output is QC report of the samples and a count matrix K, with one row for each gene i and one column for each sample j. The matrix entries K_{ij} indicate the number of sequencing reads mapped to a specific gene in a given sample.

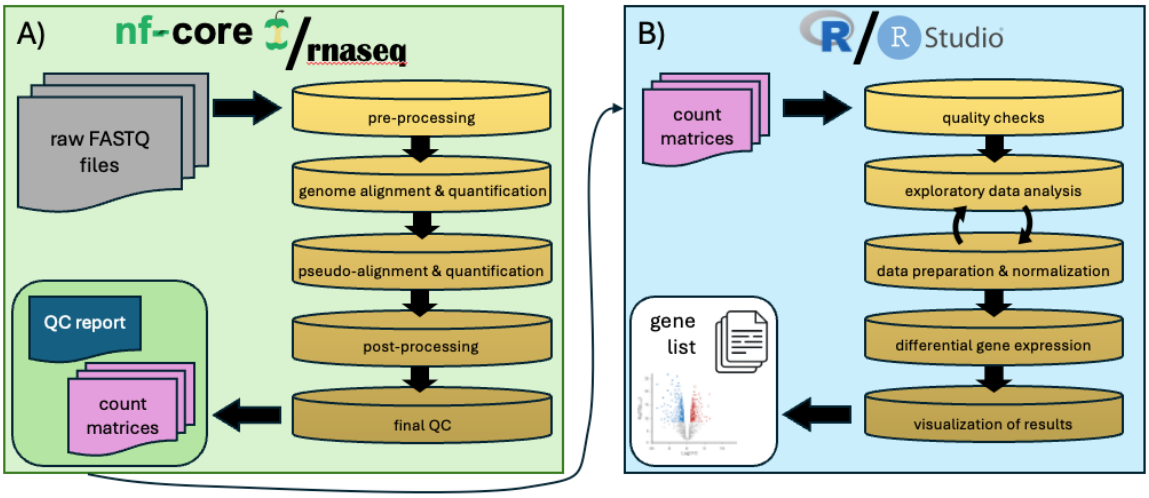


Figure 1.10: RNAseq analysis workflow. ***A)*** *The analysis starts with raw FASTQ files generated by the sequencing technology used. Using the nf-core/RNAseq pipeline, the files go through initial steps like quality check (QC) and trimming. After prepping the data, the pipeline either aligns the reads to a genome and measures transcript levels or uses an alternative pseudo-alignment method depending on the chosen option. The final steps involve formatting the data correctly and verifying results, ending with one more QC step. This stage produces a detailed QC report and count matrices that show gene or transcript expression levels in the samples.* ***B)*** *The count matrices from Panel A are imported into R/RStudio for further statistical and exploratory analysis. This part starts with checking data quality, then explores the data to find patterns or any discrepancies. The data is then prepared and normalized to account for differences in sequencing depth and other technical issues. Analysis is done to find genes with significant expression changes between conditions. The results are visualized with plots like volcano plots or heatmaps to understand their biological importance. The output includes a list of significantly changed genes and figures.*

The count matrices are transferred to a different computational language and platform, most commonly R and RStudio, respectively. For downstream, differential gene expression (DGE) analysis, the two most widely used packages are edgeR (181) and DESeq2 (182). It was noted in many studies, where simulations and performance with sample size differences were compared, that edgeR and DESeq2 perform similarly well. Hence, the selection between the two packages ultimately boiled down to a

matter of personal preference regarding the clarity of the following aspects: (i) the algorithm, (ii) the manuals, and (iii) the setup procedures.

edgeR

edgeR assumes the resulting count matrix (for example, from the Nextflow pipeline) K_{ij} (for ith gene and jth sample), follow a negative binomial (NB) distribution, like so:

$$K_{ij} \sim NB(M_j p_{ig} \Phi_i)$$

where M_j is the library size of sample j, p_{ig} is the relative abundance of gene i in experimental group g to which sample j belongs to, and Φ_i is the dispersion for ith gene. The mean and variance are defined as follows:

$$\mu_{ij} = M_j p_{ig}$$

$$\sigma_{ij}^2 = \mu_{ij}(1 + \mu_{ij}\Phi_i)$$

edgeR estimates genewise dispersions through conditional maximum likelihood, using empirical Bayes method to shrink them toward a consensus value, allowing information sharing among genes. It then performs differential expression analysis using a modified Fisher's exact test for overdispersed data (181).

DESeq2

In contrast, DESeq (183), the predecessor to DESeq2, extended the negative binomial observed in edgeR and included the variance and mean, in a more data-driven way, as follows:

$$\sigma_{ij}^2 = \mu_{ij} + M_j^2 v_{ig},$$

where μ_{ij} is the random error term or "shot noise" and $M_j^2 v_{ig}$ is the variance. Here v_{ig} captures how the raw variance changes with the relative abundance of gene i, hence $v_{ig} = v(p_{ig})$. This assumption is necessary due to the typically limited replicate count, preventing a precise variance estimate for gene i from its corresponding data alone. Therefore, this assumption allows for data pooling among genes of similar expression for variance estimation (183). It then performs a differential expression analysis using, similarly to edgeR, a modified Fisher's exact test, where test statistics are the total counts within group and the sum of those total counts across groups.

In DESeq2, a generalized linear model (GLM) is used to handle more complex relationships between the relative abundance and experimental groups. It employs a logarithmic link between the relative gene abundance and the design matrix as follows:

$$\log_2 p_{ij} = \sum_r x_{jr}\beta_{ir} \ ,$$

with design matrix elements as x_{jr} and coefficients as β_{ir} .In a basic two-group comparison, where control is compared to treatment, the design matrix elements indicate the treatment status of sample j (ie. treated or not). In addition, the GLM model fit returns coefficients for gene expression strength and log2 fold change between control and treatment (182). An empirical Bayes shrinkage approach is used for dispersion estimation, which is normally the parameter (Φ_i) that describes the within-

group variability i.e., the variability between replicates. Accurate estimation of the dispersion is a critical in the correct estimation of differential gene expression. For large datasets, this is normally not an issue; however, for small datasets (two or three replicates), within-group variability can be quite high resulting in highly variably dispersion estimates for each gene. A solution that DESeq2 uses to overcome this is to assume that genes with similar average expression also have similar dispersion (182). Therefore, the dispersion parameter is assumed to follow a log normal prior distribution with its central tendency determined by the mean read count of gene i:

$$\Phi_i \sim N(log\Phi_{trend}(\bar{\mu}_i), \sigma_d^2).$$

Here, $\Phi_{trend}(\bar{\mu}_i)$ is a function of the mean normalized count of gene i, and σ_d^2 is the dispersion around the trend that is shared for all genes. Finally, differential gene expression is tested using a Wald test, which compares the β_{ir} estimate divided by its estimated standard error $SE(\beta_{ir})$ to a standard normal distribution (182).

As DESeq2 has a more comprehensive algorithmic background to deal with a smaller dataset (which I had), and an easier manual and set up, this was the package chosen for downstream analysis (further details about its implementation can be found inn the Methods and relevant chapters).

1.5.3. microRNA databases

As previously highlighted, microRNAs are important post-transcriptional regulatory agents of gene expression; therefore, there was a need to record and catalog these RNA sequences, predict their targets and provide functional annotations. Many resources have been created over the past decade or so; however, their number one issue has been defined as their maintenance. In the small non-coding RNA community, many have claimed to have created a comprehensive database; however, most of them were abandoned after a few years. This has made my search for a reliable database of miRNAs especially challenging. Here, I will highlight my findings and why I settled on using the specific database that I ended up using.

MiRNA databases can broadly be classified based on their primary focus: (1) miRNA sequence repositories, (2) miRNA-mRNA target prediction platforms, (3) experimentally validated interaction databases, and (4) integrated resources combining multiple functionalities.

One of the most comprehensive databases is miRBase, which serves as the standard repository for miRNA sequences across a wide array of species, including animals, plants, and viruses. Initially created in 2002 under the name microRNA Registry, miRBase is updated regularly (although last update was released in 2018) to incorporate newly discovered and experimentally validated miRNAs, making it an essential reference for researchers across the world. Its focus on community contributions ensures it remains a central hub for miRNA identification due to its standardised nomenclature. Each miRNA entry includes detailed information, such as its genomic location, precursor and mature sequences, and links to external resources for further exploration. The database also integrates sequence alignment tools, allowing users to compare newly discovered miRNAs with existing entries. It allows scientists to explore miRNA sequences, family names, and their evolutionary contexts. Despite its extensive reach, miRBase has notable limitations. It primarily serves as a

sequence repository and lacks extensive experimental validation data for miRNA targets. In addition, it may not provide the depth required for studies focused on specialized areas, such as the unique categorisation of miRNAs like medicinal plant derived miRNAs. In addition to miRBase, Rfam is another foundational resource for miRNA related information. This database offers broad coverage of non-coding RNA families, including miRNAs, with a focus on structural data across multiple species. This makes Rfam an ideal candidate for those specifically researching RNA folding, structure and evolutionary relationships and structural conservation – though Rfam's emphasis on all ncRNA types provides a wider scope of RNA data, and, therefore, less specific for certain hypotheses.

For experimentally validated miRNA targets, TarBase and mirTarBase stand out. TarBase, created in 2006, indexes experimentally supported miRNA-target interactions, boasting over one million entries derived from numerous methodologies. This extensive validation makes it highly valuable for those testing predicted interactions (latest update, v8.0 in 2020). Similarly, mirTarBase, established in 2010, focuses on validated interactions confirmed through rigorous experiments like reporter assays and western blotting, offering reliable, confirmed data for researchers. The database is updated regularly, with the latest version, release 9.0, published in 2023. The strength of these databases lies in their emphasis on experimental evidence, whereas databases like miRNEST offer broader integrative data across species but rely more on computational predictions.

Plant-focused studies benefit immensely from databases such as PMRD/PNRD and MepmiRDB (184). PMRD/PNRD provides a substantial collection of plant-specific miRNA data, including expression profiles and targets for a wide variety of species, making it essential for general plant miRNA research. MepmiRDB, on the other hand, specializes in medicinal plants, uniquely catering to researchers interested in the roles of miRNAs in therapeutic pathways. The later catalogs thousands of miRNA sequences belonging to 29 medicinal plant species (ultimately, this database was chosen, although at the time of writing this thesis, this database too has become offline). These databases offer targeted insights specific to plant biology, contrasting with the more general data available in miRBase and miRNEST.

For model organism-specific data, the Arabidopsis Small RNA Project (ASRP) and PmiRKB serve specialized roles. ASRP focuses on small RNAs in Arabidopsis, providing crucial data on small RNA biogenesis pathways in this model plant. PmiRKB extends its focus to include rice, offering functional modules that explore SNPs and miRNA-target interactions. These resources are invaluable for understanding species-specific pathways, particularly in model plants, whereas databases like PMRD offer broader plant-wide data. Additionally, starBase provides a platform for exploring miRNA-interactions validated by techniques like CLIP-seq, supporting detailed interaction analyses. In contrast, sRNAanno and PASmiR focus on plant small RNAs, with sRNAanno covering a broad range of RNA classes beyond miRNAs, and PASmiR offering insights into stress responses in plants.

Ultimately, there are a vast number of resources available for non-coding RNA research. However, the field has diverged so much into such niche areas of research, probably due to funding or manpower support availability, there seem to be no consistent maintenance and updates to these resources.

1.5.4. microRNA prediction algorithms

There are not only a great number of databases available, but also miRNA-target (mostly mRNAs) binding algorithms as well. The interaction between miRNAs and mRNAs is central to numerous biological processes, including development, differentiation and disease pathogenesis. Accurate prediction of miRNA-mRNA interaction is important to understand for this particular post-transcriptional regulatory machinery and for the development of drug targets/agents. Over the past two decades, numerous computational algorithms have been developed to predict these interactions. These methods differ in terms of their underlying methodologies, biological assumptions, computational approaches and prediction accuracies. Here, I will discuss the methodologies, strengths, limitations and practical applications of miRNA-mRNA prediction algorithms, to highlight why a particular algorithm was chosen for this project.

miRNA-mRNA prediction algorithms can broadly be categorized into three main classes: (1) sequence-based algorithms, (2) thermodynamics-based algorithms, and (3) machine learning or artificial intelligence-based algorithms. Each of these classes leverage different biological principles and computational strategies, leading to variations in sensitivity, specificity, and usability.

1.5.4.1. Sequence-based Algorithms

Sequence-based algorithms primarily rely on the Watson-Crick base pairing rules to identify potential binding sites between miRNA and mRNA. The canonical seed region of miRNA, typically spanning nucleotides 2-8 from 5’ end, is a key determinant in these models. Tools such as TargetScan, miRanda, and PITA exemplify sequence-based approaches.

TargetScan is one of the most widely used and extensively validated tools for predicting miRNA-mRNA interactions. It was initially introduced in 2003 and has undergone multiple updates, each enhancing its predicting capabilities by incorporating new features and datasets. The tool relies heavily on sequence-based features, particularly the complementarity of the miRNA seed region with the mRNA 3’ untranslated region (3’ UTR), along with evolutionary conservation as a cornerstone for identifying functional interactions. At its foundation, TargetScan operates by scanning the 3’ UTRs of mRNA sequences for sites that are complementary to the seed region of miRNAs. The seed region is highly conserved and critical for binding. The tool assigns high confidence to interactions where this seed region forms a perfect match with the mRNA. Non-canonical interactions, such as those involving mismatches or G:U wobble pairs, are generally excluded, as they are less likely to result in functional regulation. One of the distinguishing features of TargetScan is its integration of evolutionary conservation. By aligning the 3’ UTRs of orthologous genes across multiple species, TargetScan identifies binding sites that conserved through evolution, under the assumption that conservation implies functional importance. This approach increases the likelihood that predicted interactions are biologically relevant. However, it inherently biases the tool toward identifying conserved interactions, potentially overlooking species-specific miRNA-mRNA pairs that may play significant roles in unique biological contexts. Over time, TargetScan has incorporated additional features to improve prediction accuracy. For instance, later versions account for the position of the miRNA binding site within the 3’ UTR. Studies have shown that binding

sites located closer to the stop codon or in AU-rich regions are more likely to be functional. TargetScan integrates this positional information into its scoring system, providing a more nuanced prediction of interaction likelihood. Another significant enhancement is the inclusion of site accessibility. TargetScan estimates the accessibility of the mRNA binding site by considering the local secondary structure of the mRNA. While this is a relatively basic assessment compared to dedicated tools like PITA, it represents an effort to account for the physical feasibility of miRNA binding. This algorithm also uses a weighted scoring system that combines multiple features such as seed match type (eg. 7mer-m8, 7mer-A1, 8mer), conservation and site position, to rank predicted interactions. This scoring system allows users to prioritize high-confidence interactions for experimental validation. Some of the strengths of TargetScan include: (1) high sensitivity for conserved interaction, (2) comprehensive database, (3) integration of biological context, and (4) user-friendly web-based interface. However, its limitations can include: (1) bias toward conservation, (2) simplistic site accessibility assessment, (3) exclusion of non-canonical interactions, (4) static predictions, (5) false positives, and (6) web-interface. *For my project, the initial search was for an algorithm that I could perhaps improve; however, black box web interfaces and differing programming languages (TargetScan is written in Perl – which I do not know) are great limitations of this endeavour.* Even with all its limitations, TargetScan remains one of the leading algorithms for miRNA-mRNA interaction prediction, combining robust computational methodologies with biological insights to provide reliable predictions. Its emphasis on conservation, seed match specificity, and site position has made it a gold standard for studying conserved regulatory networks. However, its limitations, particularly in capturing non-conserved and non-canonical interactions, underscore the need for complementary tools and experimental validations. Future iterations of TargetScan could benefit from incorporating tissue-specific expression data, medicinal plant information, advanced RNA structure modelling, and machine learning techniques to enhance prediction accuracy and biological relevance. Even with the inclusion of all kinds of features to enhance binding prediction. *This algorithm lacked a key aspect for my project – adaptability for cross-kingdom prediction.*

Other popular sequence-based algorithms are miRanda, PITA, for instance. MiRanda, in contrast to TargetScan, employs sequence complementarity between miRNA and mRNA, followed by a scoring system based on the alignment. It provides a quantitative measure of binding strength but does not always incorporate conservation, making it more flexible in identifying non-conserved targets. The lack of stringent conservation criteria may, however, increase false positive rates, as non-conserved alignments may lack functional relevance.

PITA goes a step further by integrating both sequence complementarity and accessibility of the mRNA secondary structure (therefore could be classed as a hybrid of sequence- and thermodynamics-based algorithm). Unlike TargetScan and miRanda, PITA calculates the change in free energy associated with target site accessibility, thereby factoring in the physical feasibility of interaction. The authors reasoned that site accessibility plays a determinant role in miRNA-mRNA interaction predictions, as demonstrated by the discrepancies between PITA and other algorithms. Their analysis revealed that PITA predicts many highly accessible binding sites that are missed by other methods, while excluding sites with low accessibility that other algorithms often predict. They further argued that genomes appeared to have

evolved to Favor the placement of miRNA target sites in structurally accessible regions, likely to enhance miRNA binding efficiency. This conclusion was supported by their observation that miRNA seeds in four different organisms showed a notable preference for highly accessible regions compared to random genomic locations. Additionally, this preference was even more pronounced for conserved seeds, indicating that the evolutionary positioning of targets in accessible regions is both a selective and functionally significant adaptation. However, PITA's reliance on computation RNA folding algorithms introduces potential biases, as secondary structure predictions are inherently probabilistic. *This algorithm was also limited by the others available at the time as its paper was published in 2007; this method, too, is offline now.*

All in all, sequence-based algorithms are generally efficient and user-friendly, but often suffer from high false positive rates due to the simplicity of their underlying assumptions. They also fail to account for the dynamic nature of miRNA-mRNA interactions, which can vary across tissues and developmental stages.

1.5.4.2. Thermodynamics-based Algorithms

Thermodynamics-based algorithms focus in the stability of miRNA-mRNA duplexes by estimating the Gibbs free energy (ΔG) of hybridization. These methods assume that energetically favourable interactions are more likely to occur in vivo. A prominent example of a thermodynamics-based approach is RNAhybrid. It predicts miRNA-mRNA interactions by finding the most energetically favourable hybridization sites on the mRNA for a given miRNA sequence. It calculates the minimal free energy of hybridization and provides detailed information about the structural configuration of the duplex. While this approach excels at identifying highly stable interactions, it does not consider the biological context, such as miRNA abundance, target site accessibility, or evolutionary conservation. Consequently, RNAhybrid may overestimate the functional relevance in of certain interactions simply because they are energetically favourable. This approach is advantageous in identifying non-canonical interactions, such as those involving bulges or mismatches. However, their lack of integration with biological context and reliance on static RNA secondary structure predictions limit their practical utility in predicting functional interactions under dynamic cellular conditions.

For instance, looking more deeply into the miTAR algorithm reveals a similar approach that PITA was implementing, but with the use of a deep learning algorithm. MiTAR incorporates a detailed assessment of mRNA secondary structure to evaluate the feasibility of miRNA binding given structural alignment in the seed region is fulfilled. This structural perspective adds an additional layer of biological realism to its predictions. The core principle of miTAR lies in its integration of miRNA seed region complementarity with the thermodynamic properties on the 3' UTR of mRNA sequences based on complementarity to the miRNA seed region. It then evaluates the local secondary structure of the mRNA to determine whether the binding site is accessible for interaction. Accessibility is assed using RNA folding algorithms that predict the minimum free energy of mRNA secondary structures. Regions with low free energy and high single-strandedness are considered more accessible, and thus more likely to facilitate miRNA binding. By integrating structural accessibility into its prediction framework, miTAR aims to reduce the number of false positives that arise from purely sequence-based approaches. miTAR also incorporates features such as

conservation of the miRNA binding site across species, albeit to a lesser extent than tools like TargetScan. The primary focus remains on the interplay between sequence complementarity and structural accessibility. miTAR assigns scores to predicted interactions based on these three features. High-scoring interactions are considered more likely to be biologically relevant/possible, providing users with a ranked list of candidates for experimental validation. While miTAR is primarily focused on human miRNAs and mRNA, it also supports predictions for other species. This feature makes it applicable to comparative studies and research in non-human model organism. While conservation is included as a feature, the algorithm does not place much weight on this criterion as tools like TargetScan or PicTar. This limitation may lead to the inclusion of species-specific interactions that are less likely to be conserved across evolutionary timescales. The algorithm is computationally demanding and complex. For large scale studies an HPC is definitely required (*in my case Eddie, Edinburgh University's HPC cluster, was used*). *This algorithm is also still actively maintained, which is why it was ultimately chosen for this project.*

1.5.4.3. Machine Learning-based Algorithms

The recent explosion of artificial intelligence (AI) on top of the already established machine learning (ML) has given a new angle of approach for miRNA-mRNA interaction prediction. These methods leverage large datasets to learn complex patterns and make predictions based on multiple features, including sequence complementarity, conservation, secondary structure, and contextual factors such as tissue0specific expression. Examples of machine-learning-based tools include MirTarget, TargetMiner, and deep learning frameworks like deepTarget.

MirTarget uses features such as seed region complementarity, AU content near the binding site, and mRNA secondary structure to train classification models. By incorporating experimentally validated interaction data as a training set, MirTarget improves prediction accuracy compared to sequence- and thermodynamics-based methods. However, its performance is highly dependent on the quality and size of the training dataset, and the predictions may be biased toward the characteristics of the training data.

TargetMiner employs support vector machines (SVMs) to classify miRNA-mRNA pairs as interacting or non-interacting. This algorithm outperforms traditional methods by integrating diverse features and optimizing decision boundaries for classification. Nevertheless, its reliance on labelled data limits its application in predicting novel interactions where experimental validation is lacking.

Deep learning approaches, such as deepTarget, represent the cutting edge of miRNA-mRNA prediction. These algorithms use neural networks to model intricate relationships between features, enabling the discovery of non-obvious patterns. DeepTarget incorporates multi-layered architectures to predict interactions with high sensitivity and specificity. However, deep learning models are often criticised for their "black-box" nature, as the decision-making process is difficult to interpret. Moreover, these models require extensive computational resources and large, high-quality datasets for training which may not always be available.

The key strengths and limitations of the different algorithms are summarized in terms of sensitivity, specificity, computational efficiency and biological relevance. Sequence-based methods are computationally efficient and easy to implement but often yield high false positive rates due to their simplistic assumptions. Thermodynamics-based approaches provide detailed insights into interaction stability but lack contextual and biological relevance. Machine learning-based methods offer unparalleled accuracy and feature integration but are resource-intensive and highly dependent on the quality of the training data. To combat limitations of individual methodologies, for the benefit of the field, hybrid algorithms should be put into focus for further development that integrate multiple features. For instance, miTAR and PITA had the right direction of development. *Importantly, none of these algorithms take into account potential cross-kingdom binding interactions, which has greatly limited this study. However, based on extensive assessment, miTAR was considered to be the most accessible, recently developed or updated, and algorithmically complex and flexible to handle cross-kingdom inputs.*

1.5.4.4. MirCompare

MirCompare was the computational tool I was supposed to have access to and consequently collaborate to improve as per the industrial collaboration being part of my PhD. However, due to Covid and staffing reasons, this fell through.

MirCompare is a computational tool designed to address a specific niche within the miRNA research community: the identification and comparison of miRNA sequences across different kingdoms. The algorithm was developed to explore cross-kingdom functional homologies between plant and mammalian miRNAs. The tool aims to identify plant miRNAs that “look” like mammalian miRNAs and, therefore, capable of regulating human genes; therefore, leveraging both sequence and seed region homologies. MirCompare employs a scoring algorithm to quantify the alignment between miRNAs from plant and mammalian datasets. The user inputs two fast files one containing plant miRNAs, and the other containing the mammalian miRNAs. The algorithm calculates an alignment score $S_{A,B}$ for each pair of miRNAs A and B :

$$S_{A,B} = \frac{matches_{A,B}}{\max\left(length(A), length(B)\right)},$$

where $matches_{A,B}$ represents the number of matched nucleotides between the two sequences, while $length(A)$ and $length(B)$ denote the lengths of the respective miRNAs. The score is normalized to account for differences in sequence lengths, ensuring a fair comparison. To identify the best alignment for each miRNA pair, MirCompare slides the shorter sequence across the longer one with a step size of one nucleotide and calculates $S_{A,B}$ at each step. The maximum score across all alignments, $r_{A,B}$, is defined as:

$$r_{A,B} = \max_{i}(S_{A,B}),$$

Where i represents the sliding window position. To establish statistical significance, stochastic comparisons are performed using randomized miRNA datasets. These

analyses determine a threshold r-value, typically set at 0.48, which corresponds to a p-value ≤ 0.05. Comparisons with $r_{A,B} \geq 0.48$ are considered statistically significant. An additional filtering step focuses on the seed region (nucleotides 2-8), applying a stricter homology criterion. This dual-layer filtering ensures both overall and seed-specific homology are rigorously evaluated.

In conclusion, MirCompare is robust and unique computational tool for exploring cross-kingdom miRNA-mRNA interactions. By combining precise mathematical modelling with a rigorous filtering workflow, it provides high-confidence predictions of functional homologies. It has been used to study regulatory roles of plant miRNAs such as mol-miR168a from *Moringa* oleifera, which has been shown to downregulate the human SIRT1 gene. There's a user-friendly web interface where the two FASTA files need to be supplied, that contain the miRNA sequences. However, as previously mentioned in other algorithm assessments, this was viewed as a limitation as the tool became a "black-box", once it became clear that access would not be given to this algorithm.

1.6. Aims

The primary goal of this thesis as part of the MRC Precision Medicine PhD programme was to identify high-confidence therapeutic targets in glioblastoma (GBM) suitable for dietary miRNA intervention, leveraging both experimental models and patient-derived tissues to ensure physiological relevance. While cell lines provide controlled systems to study hypoxia-driven responses in glioma stem cells (GSCs), they lack the complexity of the tumour microenvironment. Conversely, tissue samples capture the full spectrum of cellular interactions but introduce confounding variables. By integrating both approaches, this work sought to (1) define hypoxia-induced gene expression patterns in GSCs, (2) validate these findings in primary and recurrent GBM tissues, and (3) identify targetable pathways conserved across models.

To achieve this, I employed a multi-omics framework. Genomic variant analysis (CLCbio) mapped mutations in GSCs, while RNAseq (DESeq2) quantified hypoxia-induced expression changes. These cell-line-derived targets were cross-referenced with tissue RNAseq and proteomics data to filter for clinical relevance. Notably, tissue analysis revealed glioma progression and recurrence-associated factors (e.g., HOX genes, and collagens, respectively) absent in cell lines – an unanticipated but critical finding. Targets were further validated via immunoblotting and IHC (QuPath-quantified) in patient samples. Finally, computational screening of plant miRNAs identified potential regulators of prioritized genes. This dual-model strategy not only highlighted limitations of GSC systems but also uncovered novel targets that may inform future RNA-based therapies.

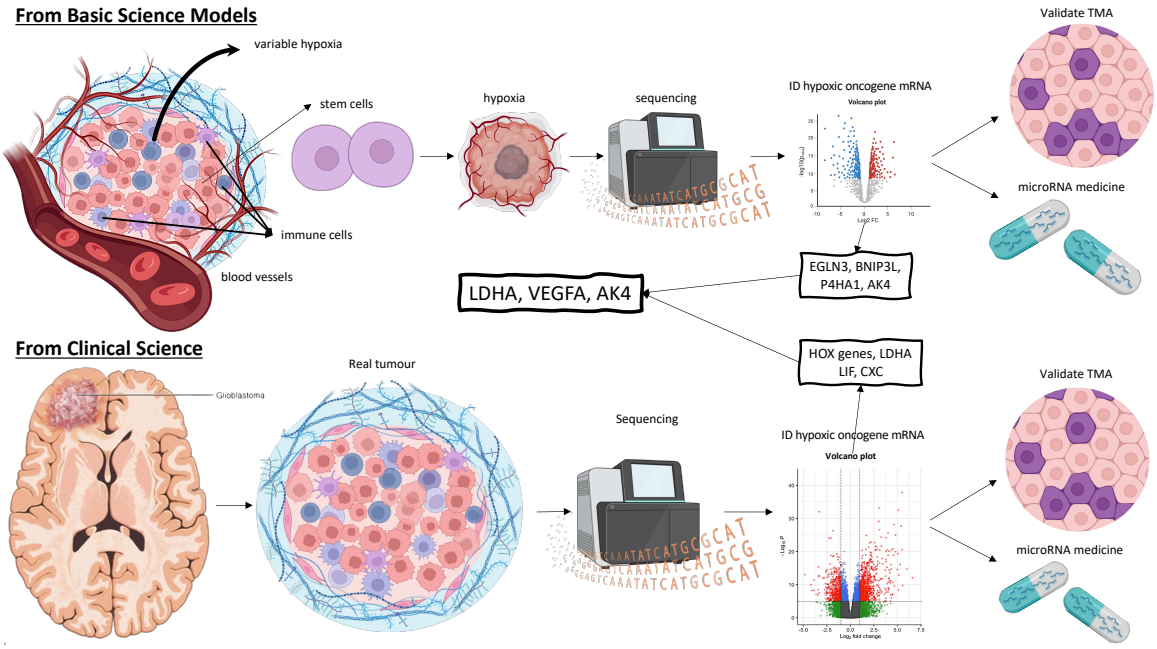


Figure 1.11: Graphical aims

CHAPTER 2. Materials & Methods

2.1. Wet-lab processes

2.1.1. Glioblastoma stem cell lines

Genomic DNA was isolated from patient-derived glioblastoma stem cell (GSC) lines E22 and E27 formerly (G322 and G327, obtained from the GCGR repository) and matched control samples using the ChargeSwitch gDNA Mini Tissue Kit (Invitrogen) according to the manufacturer's protocol. Both GSC lines were derived from patients with primary, grade 4 glioblastoma. The E22 (G322) cell line was established from a patient treated with 30 rounds of radiation therapy without temozolomide (TMZ), while the E27 (G327) line was derived from a patient with 30 rounds of radiation and 7 rounds of TMZ.

Cell culture (179) and hypoxia (180) experiments were performed by previous lab members prior to the start of my project, described in citations. Briefly, cells were cultured under standard (normoxic, 21% O_2) or hypoxic (3% O_2) conditions at 37°C in a dedicated incubator. Mycoplasma testing was routinely performed using Luciferase-based MycoAlert™ kit (Lonza). All extractions were performed by Ashita Singh under standardized conditions to ensure consistency.

2.1.2. Tissue samples

Glioblastoma tumour tissue samples were obtained from patients undergoing craniotomy for suspected glioma at Western General Hospital, Edinburgh, with informed consent provided in accordance with the South East Scotland Research Ethics Committee (REC 20/ES/0061). These samples were generously provided by Dr. Paul Brennan. Our cohort consisted of 18 patient tumours (9 female, 9 male) with a median age of 60 years (range 20-70). Immediately after resection, tumours were snap-frozen and stored at -80°C. For processing, performed by Sofian Al Shboul MD, PhD, frozen tissues were thawed on ice in a Category 2 tissue culture hood and homogenized in two volumes of PamStation12 lysis buffer using mechanical dissection with a plastic tissue smasher followed by repeated pipetting with autoclaved tips to break the tissue. After 30 minutes of incubation on ice, lysates were centrifuged at 13,000 rpm for 25 minutes at 4°C. The resulting supernatants were aliquoted into 5-10 µL volumes, snap-frozen, and stored at -80°C. For immunoblotting, equal volumes of lysate were loaded per gel as indicated in subsequent chapters.

2.1.3. Western blot validation

For all western blot laboratory validation, the following protocols were used. To perform the SDS-PAGE, 4-12% NuPAGE MOPS SDS Running Buffer (20X) (ThermoFisher Scientific, Appendix Table 1) was used to create the buffer, where 50 ml of solution was diluted into 950 ml of diagnosed water to make 1 litre of MOPS SDS Running Buffer. Final concentration is 50 mM MOPS, 50 mM TRIS base, 3.47mM SDS and 1mM EDTA. 15 µL patient sample was loaded onto the SDS-PAGE. To perform the qPCR, the total RNA was isolated from cell pellets with RNeasy Mini Kit (Qiagen) according to the manufacturer´s protocol. Isolated RNA was then used for cDNA synthesis using qScript cDNA SuperMix (QuantaBio) and SureCycler 8800 (Agilent). Afterwards, qPCR was performed using QuantiTect SYBR Green PCR Master Mix (Qiagen) and StepOnePlus™ Real-Time PCR System (Thermo Fisher Scientific).

2.1.4. Immunohistochemistry (IHC) validation

Tissue samples were collected from patients and fixed in blocks. The following laboratory procedures were carried out by Halen Caldwell, Head of Histology Research Service at the IGC and her team following the 'IGC Standard Operation Procedure (SOP) Form (V1)' (Appendix Table 2). Samples were cut while cold using a Microtome set to 4 microns thickness. Cuttings were transferred to warm bath (around 40 °C). Cuttings were collected from bath using positively charged microscope slides at 45° angle. Slides were left to dry overnight or in the oven, ready for IHC staining.

IHC staining procedure

IHC staining was also carried by staff at the IGC's Histology Research Service. Procedure as follows. Antibodies were optimised for selected target genes/proteins for the following genes: MOT4, BNIP3, NDRG1, HOXC10, HOXB3, OAS2, COL1, COL6. For the staining the Leica BOND-III Fully automated IHC/ISH stainer was used. Either IHC staining Protocol F or Modified Protocol F were used for the above samples (Appendix Table 3, Appendix Table 4). Samples were loaded into the staining machine and protocol F was specified for all genes except for COL1, where modified F protocol was used. Following successful staining, samples go in autostainer to dehydrate samples and apply protective cover.

Following successful staining, the resulting samples were scanned using Nanozoomer XR, DAB staining. Images were further analysed in QuPath (more in CHAPTER 7).

2.2. Dry-lab processes

2.2.1. CLC Genomics Workbench

Raw sample files in compressed FASTQ format (fastq.qz) were retrieved following Illumina sequencing. Then the DNA variant analysis and RNA variant analysis workflows were used mostly with the default settings. For more detailed description and figures, see CHAPTER 3.

2.2.2. Nextflow RNAseq pipeline

Raw sample files in compressed FASTQ format (fastq.gz) were obtained for the GBM cell lines 322 and 327 from Prof Ted Hupp. The tissue RNAseq files (from 95 patients: 40 female, 55 male), curtesy of Dr Paul Brennan, were already processed ready for downstream (DESeq2) analysis. The nf-core/rnaseq pipeline (180) was used to analyse the RNA sequencing data obtained from the cell lines to produce a gene expression matric and quality control (QC) report. The nf-core/rnaseq v3.3 pipeline was run on Eddie, the high-performance computing cluster of the University of Edinburgh (185), with the following software and their versions: bedtools v2.30.0 (186), bioconductor-summarizedexperiment v1.20.0 (187), bioconductor-tximeta v1.8.0 (188), DESeq2 v1.28.0 (182), dupradar v1.18.0 (189), FastQC v0.11.9 (190), Nextflow v21.04.3 (174), nf-core/rnaseq v3.3 (180), picard v2.23.9 (191), preseq v3.1.1 (192), qualimap v2.2.2-dev (193), rseqc v3.0.1 (194), salmon v1.4.0 (195), samtools v1.12 (196), star v2.6.1d (197), stringtie v2.1.7 (198), subread v2.0.1 (199), trimgalore v0.6.6 (200), ucsc v377 (201). The pipeline was run using singularity container (202). The pipeline used the following inputs and parameters: samplesheet.csv (Appendix Table 6), sample FASTQ files with reference genome

Homo_sapiens.GRCh38.dna.primary_assembly.fa and Homo_sapiens.GRCh38.104.gtf annotation file (both located on Eddie), read trimming 20, and salmon pseudo aligner. The resulting quantification files were used for downstream analysis conducted in R v4.1.2 (9).

2.2.3. DESeq2 RNAseq pipeline

Here, I will describe the general pipeline used for downstream analysis of samples. Further explanation can be found in subsequent chapters as cell line and tissue analyses slightly differ (CHAPTER 4 and CHAPTER 5, respectively). The pipeline was adapted from https://rpubs.com/BarryDigby/747584, and the DESeq2 manual (203). For this step, R v4.1.2 (9) and RStudio Server (version "Prairie Trillium", for full version name see Appendix 1 - Related to CHAPTER 2 (204), provided by the Bioinformatics Core from the Institute of Genetics and Cancer, University of Edinburgh, were used along with the following packages (for full package list see Appendix 1 - Related to CHAPTER 2: tidyverse (205), dplyr(206), biomaRt (207), tximport (208), gplots (209), org.Hs.eg.db (210), DESeq2 (182), PCAtools (8), EnhancedVolcano (211), ComplexHeatmap (212).

Firstly, metadata was imported along with the count data. Then using biomaRt and tximport, the count files were converted to a gene level abundance matrix containing each sample. Using DESeq2 this matrix was used for the differential gene expression analysis. To follow standard statistical procedures, some exploratory analyses and plots were used to assess the state of the data, if any adjustment (for example sample removal) needs to be made. Then, the differential gene expression was shown using a variety of plots such as volcano plots and heatmaps. In the case of individual genes, boxplots, and analysis of variance (ANOVA) analyses were performed.

CHAPTER 3. Glioblastoma stem cell analysis using DNA and RNA variant detection platforms

The central objective of this thesis was to identify mutated and/or hypoxia-induced genes in glioblastoma stem cell (GSC) models, evaluate whether these genes represent significant outliers in primary GBM tissues, and ultimately select the most physiologically relevant targets for miRNA-based therapeutic design. While targets identified in both GSCs and tissues would be ideal, discrepancies between these systems would highlight critical limitations of current cell line models and underscore the necessity of incorporating tissue-based approaches in future studies. A key focus of this chapter was the identification of potentially oncogenic mutations in GSCs. Although mutated genes are patient-specific, their characterization could inform the development of personalized miRNA therapies targeting GSC populations within tumours. For instance, the interferon locus deletion present in approximately 25% of GBM patients has recently inspired clinical trials exploring engineered macrophage therapies to deliver interferon to tumour sites (213). This example illustrates the therapeutic potential of targeting patient-specific genetic alterations. To systematically address these questions, this chapter employs DNA and RNA variant analysis platforms to investigate:

i. **Baseline genetic landscape:** What DNA variants are present in our glioblastoma stem cell lines under normoxic conditions?
ii. **Functional implications of mutations:** What are the mutated genes in proliferating GSCs and do any of the mutated genes form expression networks that could be targeted using miRNA medicines?
iii. **Clinical relevance:** How do gene regulatory networks derived from GSC models compare to driver pathways identified in matched primary GBM tissue transcriptomes?

The previous chapter laid out the overarching methodology for analysing DNA and RNA sequences from both glioma stem cells (GSCs) and tissues. This chapter delves deeper into exploring the intricate molecular makeup of GSCs, particularly focusing on mutated pathways. To comprehensively understand the underlying mechanisms driving GSC biology, it is imperative to investigate the genetic variations present in their DNA and RNA compared to the reference genome. Thus, it was crucial to identify DNA and RNA variants. By pinpointing these variants, we can uncover potential alterations in key pathways that contribute to a ‘baseline’ GSC behaviour. Additionally, understanding the mutational landscape of GSCs is essential for contextualizing the significance of specific pathways – not just genes – in the cellular response to hypoxia.

To accomplish this, I used the CLCBio workbench, a computational tool, to detect DNA variants in GSCs. This approach helps us understand the specific mutational landscape of our GSC models. Hence, this chapter serves as a fundamental step to establish the ‘baseline’ biology of our glioblastoma stem cell lines before any hypoxic experiments.

3.1. DNA variant analysis using CLC Genomics Workbench

DNA variant calling is a process in bioinformatics used to identify and categorize differences (variants) between a DNA sequence sample and a reference genome (214). These variants can include single nucleotide polymorphisms (SNPs), insertions,

deletions, etc. The typical variant calling process includes sequencing, read mapping or de novo assembly, variant detection, variant calling, annotation, and filtering of possible false positives. Analysing variants across multiple individuals has allowed for better understanding of genetic variations linked to diseases, clinical diagnosis, and evolutionary processes (214). This DNA variant detection is commonly done by using more computationally intensive methods and software such as GATK, Samtools, Nextflow, etc. **For CLCBio, the built-in template workflows allow users, to seamlessly start the otherwise computationally demanding variant calling process, which are shown on the images below (Figure 3.1, Figure 3.2, Figure 3.3).**

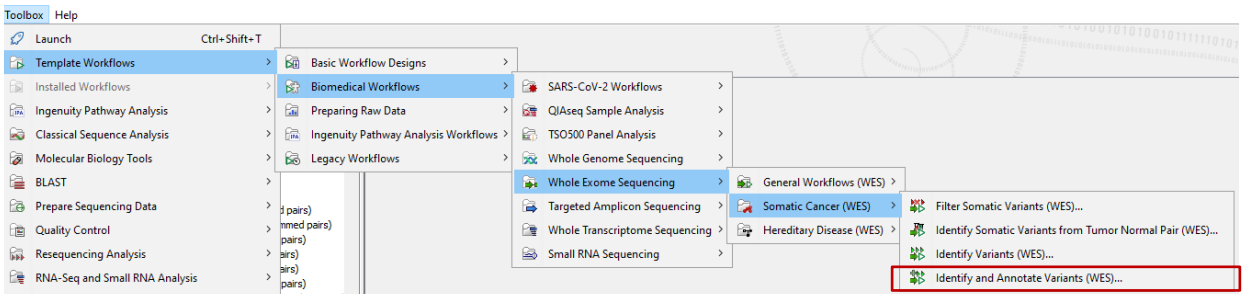


Figure 3.1: Series of steps to show how to start the DNA variant analysis workflow in CLCBio, named "Identify and Annotate Variants (WES)". Firstly, launch the CLC Genomics Workbench software on the computer. Secondly, import the whole exome sequencing (WES) data into the application (not pictured). This typically involves importing the sequencing reads in FASTQ format or aligned reads in BAM format. Then access the workflow through the "Toolbox" menu at the top, click on Template Workflows and navigate to "Identify and Annotate Variants (WES)" as pictured.

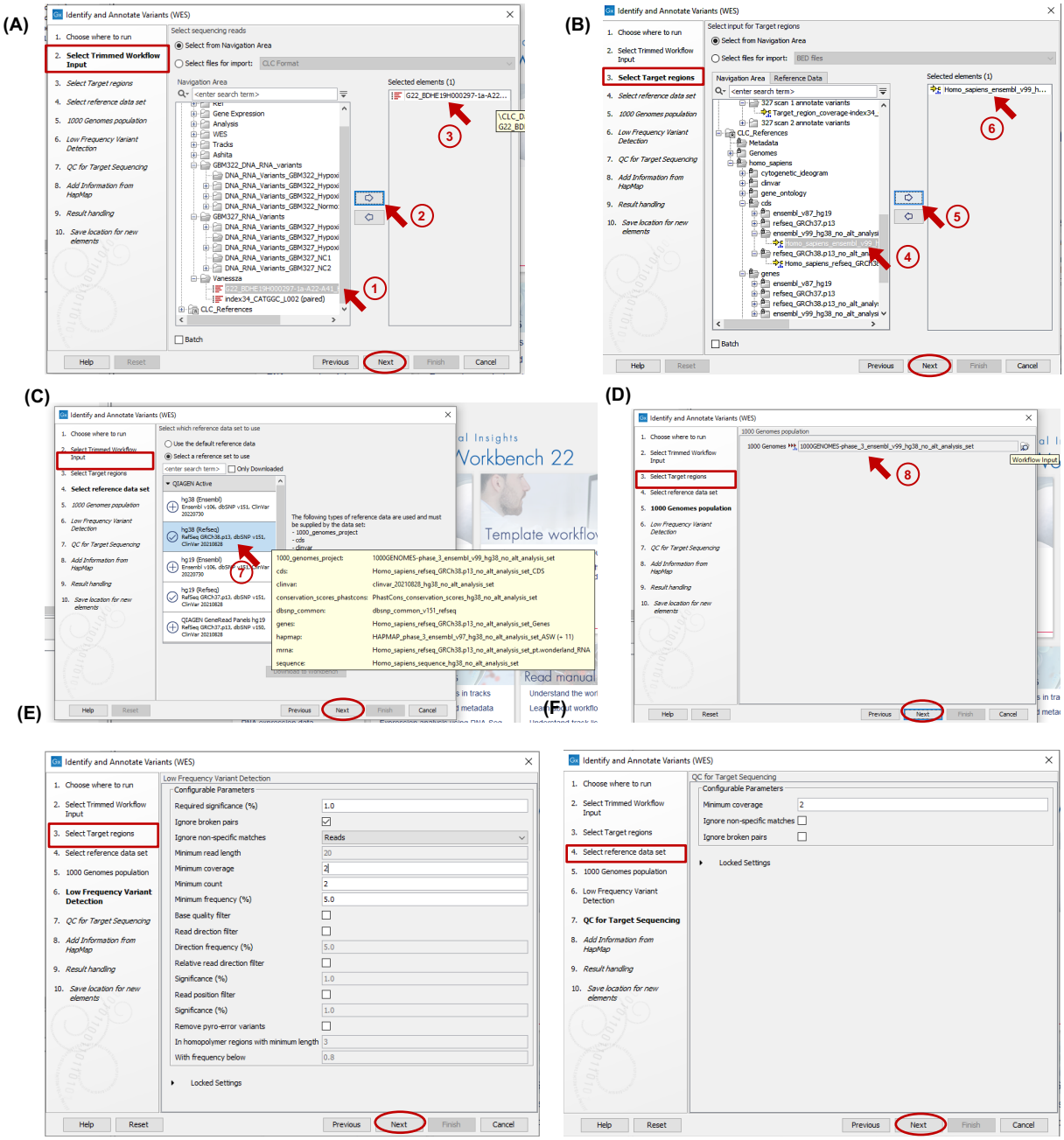


Figure 3.2: Continuation of Figure 3.1 to launch the DNA variant analysis workflow in CLCBio These screenshots show steps in the workflow editor where the user is able configure the workflow parameters. (A) In the "Select Trimmed Workflow Input" step, the user can select the paired end, trimmed reads. In this example, I used GSC327 NC1 by navigating to the right folder, clicking on the sample (1), then clicking on the arrow (2) to move the file to the "Selected elements" panel (3). Click "Next" at the bottom of the window (circled in red) to move to the next step. (B) In the "Select Target regions" step, users are able to tell the workflow if there's a specific region of interest within the genome. This input file is often a BED file format containing the region of interest. Here, as there's no specific are of interest, the whole human genome was given. (4,5,6) (C) In the "Select reference data set" step, the hg38 (RefSeq) reference genome was chosen (7). This is the genome that the samples will be compared to in the variant analysis. The information it contains is described more in detail in the yellow pop-up box. (D) In the "1000 Genomes population" step, the 1000 Genomes population annotation file is selected (8). This contains population level allele frequency information observed in the 1000 Genome Projects. (E) In the "Low Frequency Variant Detection" step, the user is able to further configure certain parameters of the analysis. For example, "required significance", which is a percentage metric, here set to 1% to catch as much information as possible; "Ignore broken pairs" box ticked, which will ignore any broken read pairs that might arise during the assembly step of sequencing; "minimum coverage", which are the only variants in regions covered by at least this many reads are called, "minimum count", which are the only variants that are present in at least this many reads are called, and "minimum frequency", which is the minimum frequency of variants that are present (calculated as count/coverage), are set to 2, 2 and 5%, respectively. The rest of the metrics are left at their default setting.

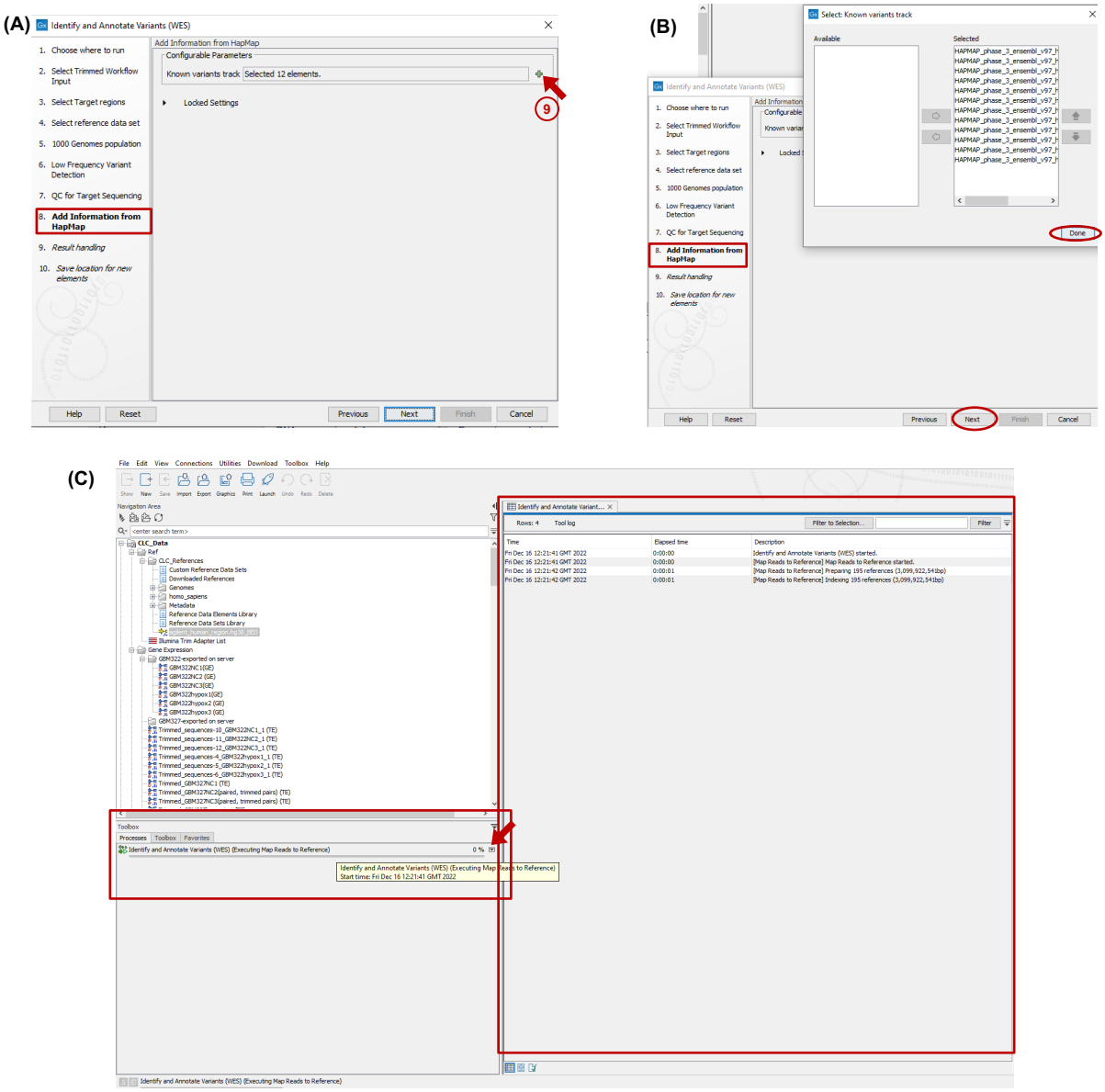


Figure 3.3: Continuation of Figure 3.2 to launch the DNA variant analysis workflow in CLCBio (A) In the step "Add Information from HapMap", the user can provide their variant tracks from HapMap (short for "haplotype map") files. HapMap refers to the International HapMap Project, where the aim is to gather information about variations in human DNA sequences with genes associated to health. A haplotype is a group of DNA variations, or polymorphisms, which are typically inherited together. I can signify either a collection of alleles or a series of single nucleotide polymorphisms (SNPs) situated on a single chromosome. The user can view or edit the HapMap tracks provided for the analysis by clicking the "+" icon (9). (B) Continuing from figure (A), the HapMap tracks open in another window, where the user is able to add or remove HapMap tracks. By clicking "done" at the bottom of the pop-up window, we return to the original window seen in step (A), where we can click "Next" at the bottom to continue saving and launching the analysis. (C) The running confirmation of the DNA variant analysis can be seen on the bottom left "Toolbox" panel under the "Processes" tab (left red box). By clicking the small down pointing arrow (marked with red arrow), we can select to view the information log of the process from the drop-down options, which open to the right of the "Toolbox" panel, where the user can track the progress of the analysis in further detail (right red box).

Once the parameters are chosen from the drop-down tabs and the analysis is launched (Figure 3.1, Figure 3.2, Figure 3.3), CLCBio returns a comprehensive list of genes containing variants (39,843 rows, where each row is a single variant, see Figure 3.5, red arrow) with mutation reads ranging from 5% to 100% occurrence frequency and variant read counts from 2 or higher (data not shown; Table available upon

request). These mutations serve as the reference for identifying mutations present among RNA sequences, which better represent what genes are *actually* being expressed in the cell.

A feature of the software is that individual gene can be manually annotated; a genome browser is part of the results of the CLCBio DNA variant analysis, where the user is able to view and interact with highlighted variants (Figure 3.4). Here, the TP53 gene on chromosome 17 is shown where a single mutation, at the 7,676,156 locus causes a proline (P) to arginine (R) amino acid mutation.

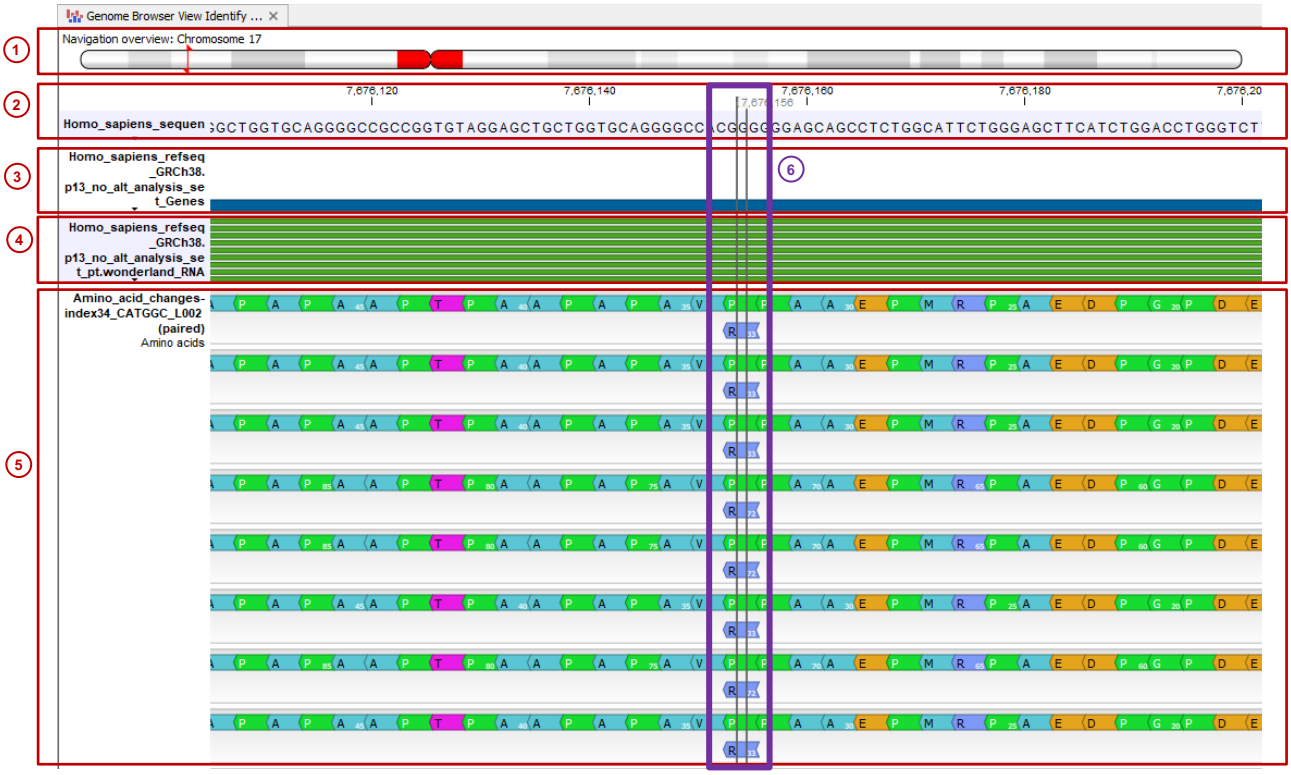


Figure 3.4: Genome browser view of DNA variant analysis by CLCBio Zoomed in view of chromosome 17, more specifically the TP53 gene where an amino acid change has occurred form proline (P) to arginine (R). The navigation overview of chromosome 17 can be seen on the very top of the viewer (1), where a vertical red line signifies our current position on the chromosome. The reference human genome sequence is located (2), where about a 100 bases can be seen along with their exact locus on the genome. Panel (3) contains the gene track (coloured in blue). This information comes from the reference data set provided in step 4 of the workflow in Figure 3.2C. Similarly, panel (4) also comes from the same workflow step; however, here the mRNA sequence coverage is being displayed in green, where each green track is an mRNA sequence. In panel (5) amino acid sequences are shown, using their one letter notation. The vertical purple box (6) highlights the variant of interest, where a variation at the 7,676,156 location causes an amino acid change from P to R. Numbers in some amino acid arrows signify their position in the sequence.

We do not just want to focus on single mutations that create a single amino acid mutation. In another example, one can easily observe multiple nucleotide variants (MNVs) using the genome browser in conjunction with the tabulated variant annotation (Figure 3.5, window sizes are adjustable, example is shown to highlight the variety or information available in the genome browser view). Here, the major histocompatibility complex, class II, DR beta 1 (HLA-DRB1) gene contains two MNV mutations at the same position, which were observed at a combined frequency of 100%, resulting in either an arginine (R) to serine (S) or arginine (R) to phenylalanine (F) amino acid

change. A more comprehensive analysis of the variants detected in the DNA will be discussed below (Figure 3.11) when I compare DNA and RNA variant results.

Figure 3.5: MNV mutation in the HLA-DRB1 gene with two multiple nucleotide variants (MNVs) , that were observed at a combined frequency of 100%. In addition to the genome browser view from the previous image (Figure 3.4), in panel (1) the target region coverage track is located (coloured in yellow). The track is present meaning the region provided in step 3 of the workflow in Figure 3.2B is also present among the provided sequences. In panel (2) the annotated variants found in our sequences are listed for each locus along the genome. Here, there's no additional information yet regarding the frequency, coverage, count etc of the variants, simply that they were found. In panel (3) the information such as indels, 1000 Genome variants, ClinVar variant information and dbSNP variant information come from the hg38 (RefSeq) information that we provided in step 4 of the workflow in Figure 3.2C. In panel (4), the read mappings are shown. This is the step where the workflow aligns the sequencing reads obtained from the whole exome sequencing (WES) to the reference genome. This step is crucial for identifying the genomic locations of the reads and determining their correspondence to specific regions of the reference genome. This is where the workflow obtains information regarding variant count and coverage of a specific locus. The different colours (turquoise, dark blue, red, green, etc) correspond to the type of read or variants (see colour legend in panel 4). Panel 5 shows more detailed information such as chromosome, region, type of variant, reference, allele, count etc. about the MNVs of interest highlighted by grey rectangle in the genome browser view, where a reference TCC mutates to either AAA or AGA. By adding up the frequencies or looking at the tooltip (yellow pop-up box), these MNVs are seen 100% among the sequences we provided.

3.2. RNA variant analysis using CLC Genomics Workbench

We view the DNA tumour barcode, variants, and tumour genome as records of accumulated mutations shaping the tumour's development. However, it's uncertain what percentage of mutant genes are expressed in cell lines derived from patients or in tissue post-surgery. Some tumours exhibit only around 10% mutant gene expression, indicating that 90% of cancer associated genes remain unmutated (66). My aim was to identify which mutant genes are expressed under normal growing conditions and to determine if I could pinpoint mutant signalling pathways using RNA sequencing data. The CLCBio offers several 'Template Workflows'. Here, I used the 'Whole Transcriptome Sequencing' analysis to 'Identify Variants and Expression Values' (v2.101) (Figure 3.6) with settings shown on Figure 3.7 and Figure 3.8.

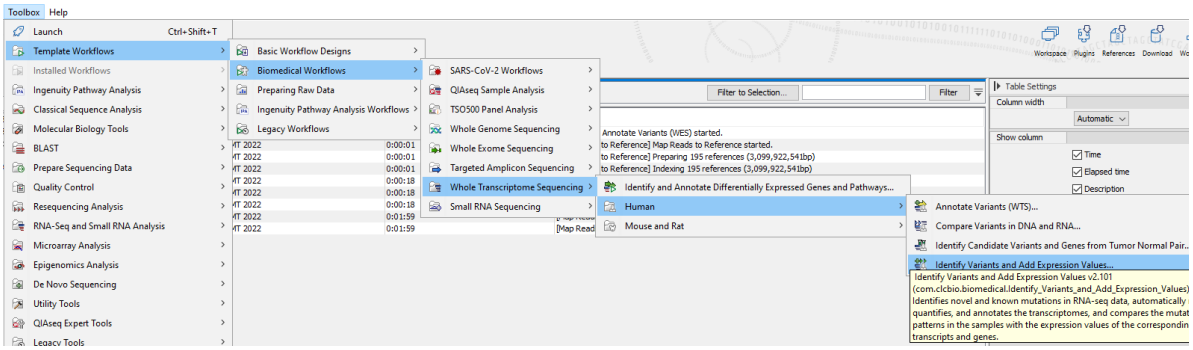


Figure 3.6. Series of steps to show how to start the RNA variant analysis workflow in CLCBio, named "Identify Variants and Add Expression Value". Firstly, launch the CLC Genomics Workbench software on the computer. Secondly, import the trimmed reads into the application (not pictured). Then access the workflow through the "Toolbox" menu at the top, click on Template Workflows and navigate to "Identify Variants and Add Expression Value" as pictured.

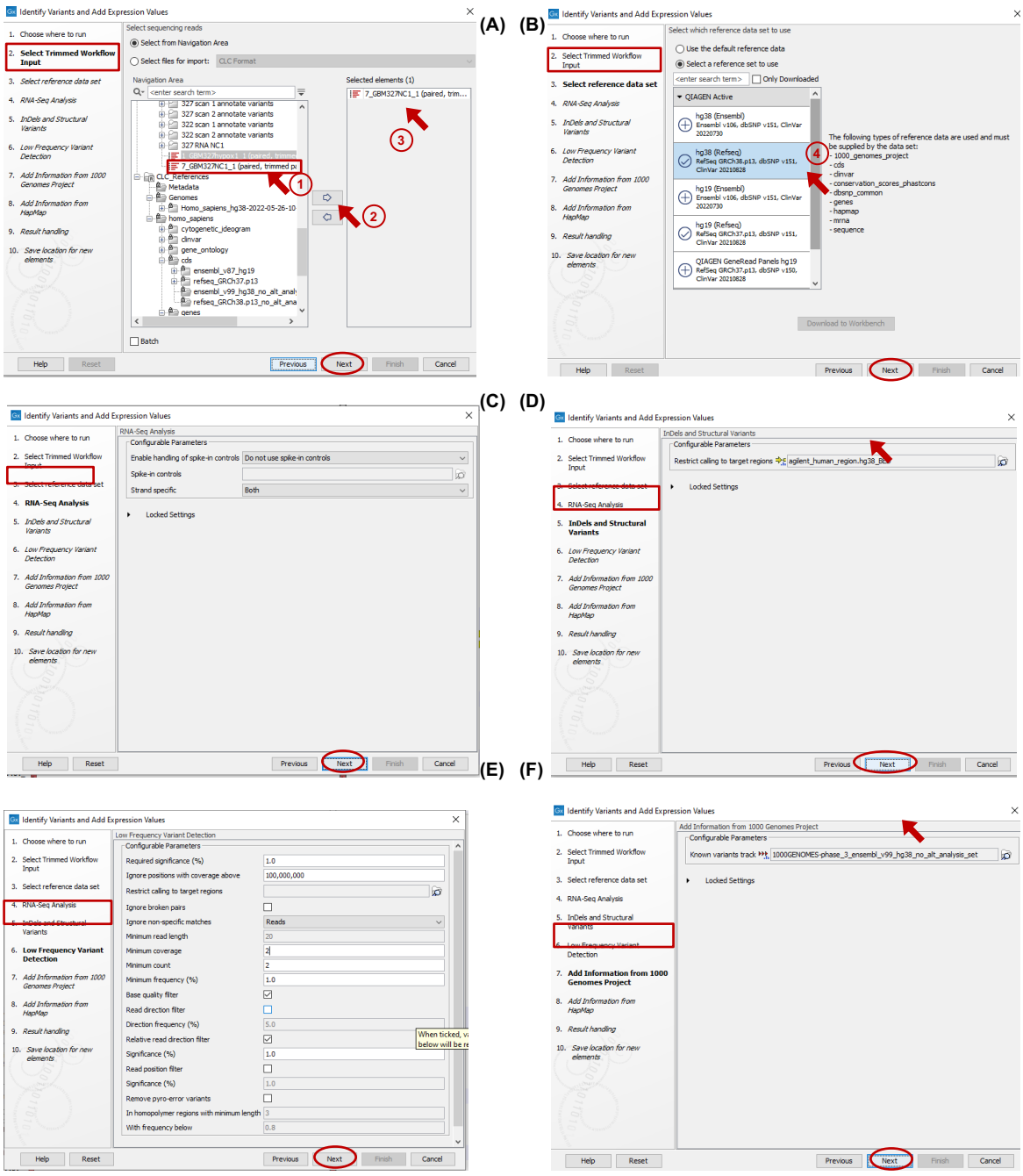


Figure 3.7. Continuation of Figure 3.6 to conduct the RNA variant analysis workflow in CLCBio. These screenshots show the steps in the workflow editor where the user is able to configure the workflow parameters. (A) In the "Select Trimmed Workflow Input" step, the user can select the paired-end, trimmed reads. In this example, I used GSC327 NC1 by navigating to the right folder, clicking on the sample (1), then clicking on the arrow (2) to move the file to the "Selected elements" panel (3). Click "Next" at the bottom of the window (circled in red) to move to the next step. (B) In the "Select reference data set" step, the hg38 (RefSeq) reference genome was chosen (4). This is the genome that the samples will be compared to in the variant analysis. The information provided is described in more detail next to the selection box. (C) In the "RNA-Seq Analysis" step the user can add information about RNA spike-in controls, which are synthetic RNA transcripts of known sequences and quantities, used to measure the accuracy and reliability of the RNA sequencing experiment. (D) Similarly, to the DNA variant analysis, in the "InDels and Structural Variants" step, the user can restrict variant calling to a specific region of interest by providing a BED file with the desired region. (D) In the "Low Frequency Variant Detection" step, the user can further configure certain parameters of the analysis. For example, "required significance", which is a percentage metric, here set to 1% to catch as much information as possible; "Ignore broken pairs" box unticked, which will not ignore any broken read pairs that might arise during the assembly step of sequencing; "minimum coverage", which are the only variants in regions covered by at least this many reads are called, "minimum count", which are the only variants that are present in at least this many reads are called, and "minimum frequency", which is the minimum frequency of variants that are present (calculated as count/coverage), are set to 2, 2 and 1%, respectively. The rest of the metrics are left at their default setting. (E) In the "1000 Genomes population" step, the 1000 Genomes population annotation file is selected (8). This contains population level allele frequency information observed in the 1000 Genome Projects.

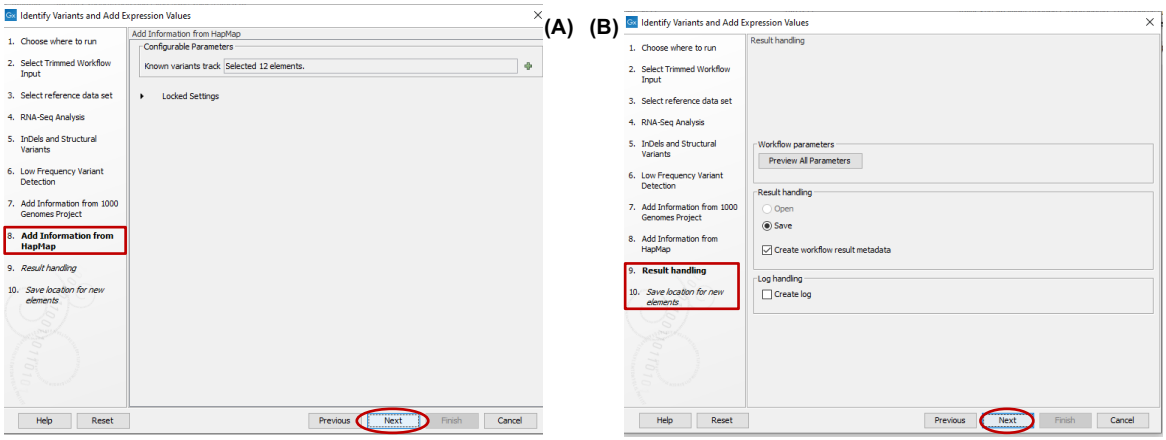


Figure 3.8: Continuation of Figure 3.7 to conduct the RNA variant analysis workflow in CLCBio (A) In the step "Add Information from HapMap", similarly to the DNA variant analysis (for more details see Figure 3.7caption), the user can provide their variant tracks from HapMap (short for "haplotype map") files. (B) Finally, launch and save results of the analysis.

3.3. Variant Analysis of DNA, RNA, and shared mutations

These workflows allow us to compare DNA and RNA variants present in the same cell, GSC327 Normal Control 1. The genome browser makes it easy to visualize scenarios where a mutation occurs in both DNA and RNA variant analyses (Figure 3.9), but also when abnormalities occur, such as when no DNA is detected whereas there's RNA data (Figure 3.10).

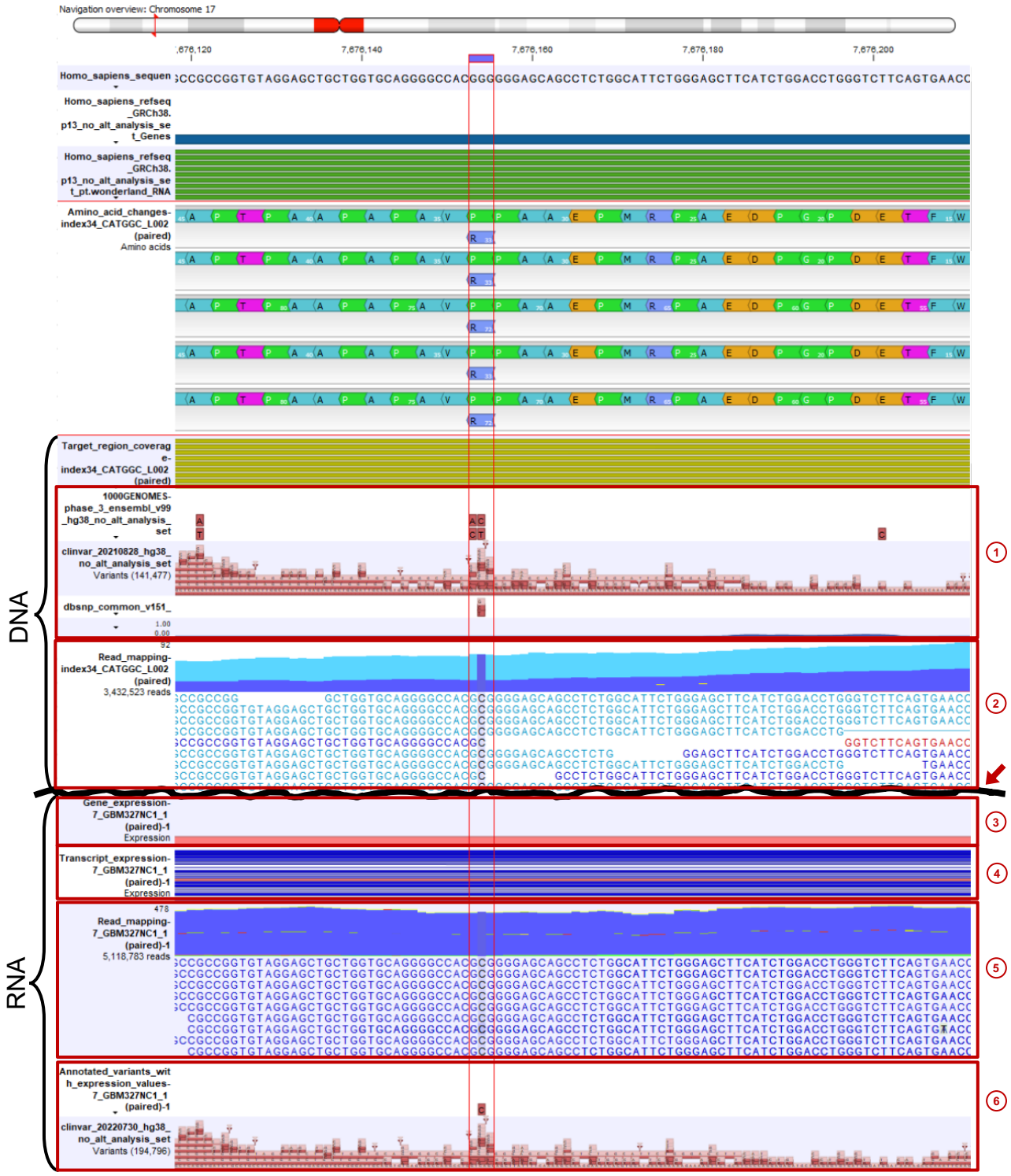


Figure 3.9: TP53 DNA and RNA variant analysis result. Expanded view of Figure 3.4 of the TP53 gene. Here, a black wavy line (pointed out by red arrow) shows the separation where information came from the dan variant analysis or the RNA variant analysis. In addition to previously shown tracks (Figure 3.4, Figure 3.5), in panel (1), variants are shown from the 1000 Genome project, ClinVar and dbSNP data bases. In panel (2), DNA reads are mapped to the reference genome. In panel (3), gene expression coverage is shown by the light red line. In panel (4), the transcript expression coverage is shown. In panel (5), mapped reads are shown from the RNA variant analysis. Finally, in panel (6) additional variant information is shown. Vertical red box highlights the mutation of interest, where a G>C mutation causes an amino acid change from proline (P) to arginine(R). This mutation is present in both analyses.

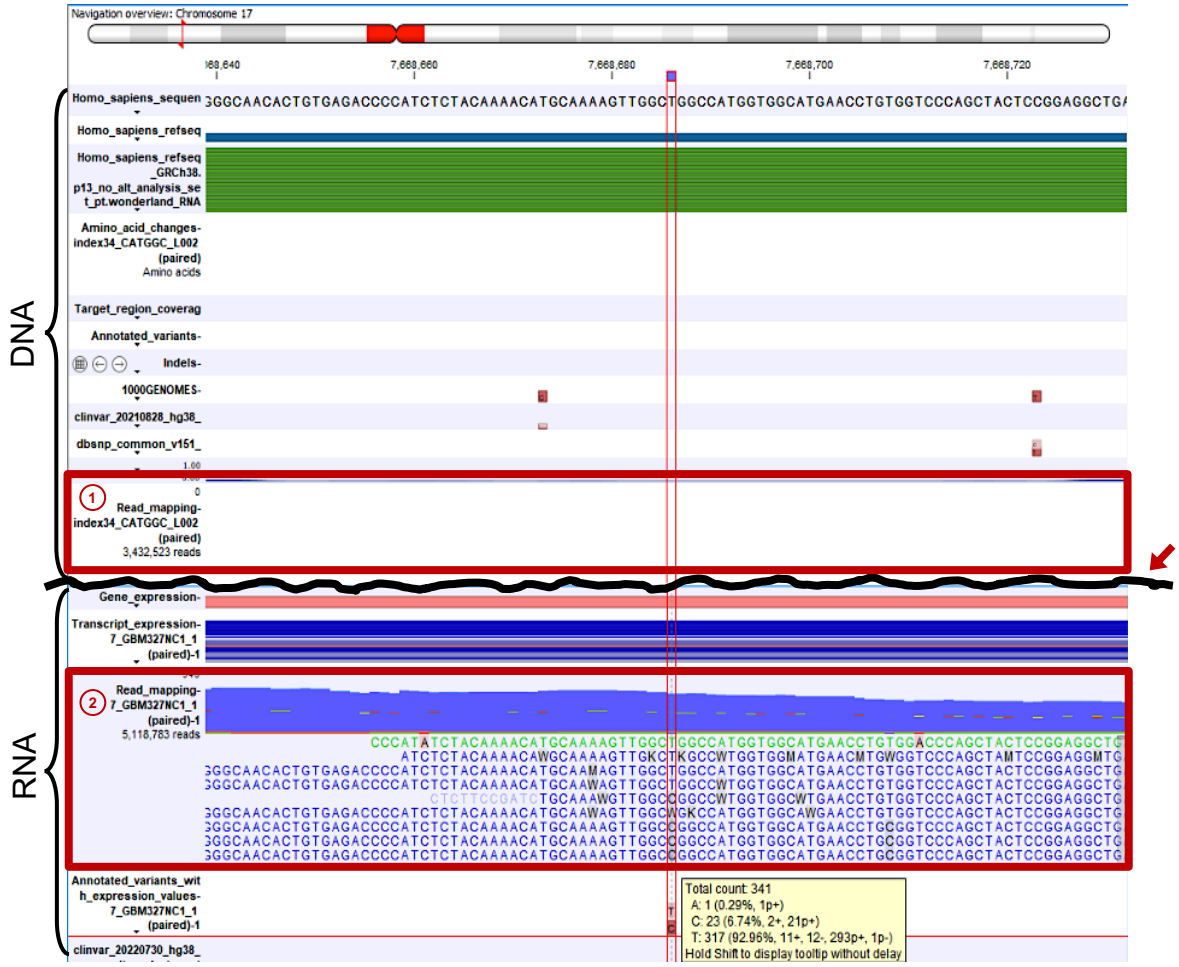


Figure 3.10: Genome browser view of TP53 gene at a different locus This figure contains information (tracks) introduced throughout the chapter (mostly Figure 3.9). In contrast to Figure 3.9, the DNA variant analysis (1) did not detect any DNA at a site whereas the RNA variant analysis (2) detected reads at the same site, marked by vertical red column.

The variant detection workflows not only show an interactive genome browser, but they also return a detailed table containing vast amount of information about the variant. The information can include, but not limited to: chromosome number; region on the reference sequence at which the variant is located; type of variant, single nucleotide variant (SNV), multiple nucleotide variant (MNV), insertion, deletion, replacement; reference sequence at the position of variant; allele; reference allele; count, the number of countable reads supporting the variant; coverage, the fragment coverage at the specific position; and frequency, the ratio between count and coverage. I used this table to compare the number and type of variants returned by CLCBio's variant analysis workflows. I also compared those variants that were present or shared by both DNA and RNA analyses (Figure 3.11). Unsurprisingly, SNVs represent an overwhelming number of the mutations present among both DNA and RNA variants. In the case of DNA variants, I examined the mutations in my cell model compared to the overall frequency mutation in GBM (top 50 genes) (Appendix Figure 2). The data show that 31 genes in my cell model are also seen in GBM genomic databases including COL6A3, EGFR and MUC16. Another expected observation is that the overall number of mutations is significantly more in RNA than DNA (Figure 3.11). This can be explained by the greater number of RNA strands present in the cell

in comparison to the one DNA strand or RNA editing. In addition, errors can occur during sequencing, where DNA data is missing (Figure 3.10).

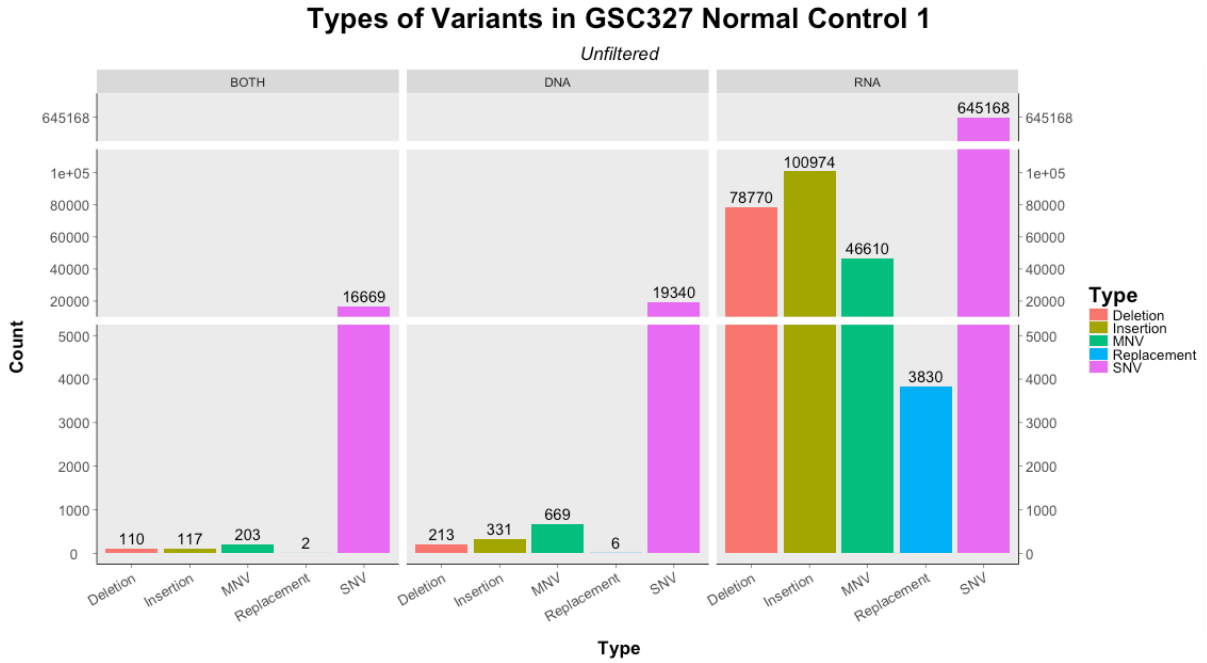


Figure 3.11: Types of variants in GSC327 Normal Control 1 "BOTH" signifying those variants that were shared between the DNA and RNA variant analyses completed by CLCBio. "DNA" and "RNA" represent those DNA and RNA variant analyses respectively. Each type of variant is coloured shown on the x axis, on the y axis their number of count or occurrences is shown. White gaps show a brake in the y axis scale, to better visualize variant types with less counts.

Clinical tumour sequencing poses numerous challenges for detecting somatic mutations. Tumour purity, which indicates the proportion of cancerous cells in a sample, impacts mutation representation, yet pathology's estimates based on light microscopy are notoriously imprecise (215). Somatic mutations occurring at low frequencies due to low tumour cellularity or subclonal mutation architectures are difficult to detect, even with high-depth sequencing. While some mutation callers can be adjusted for low-frequency variant detection, this often increases false positives (215). Additionally, the type of specimen used, such as formalin-fixed, paraffin-embedded (FFPE) samples preferred for histopathological diagnosis, can introduce artifacts from chemical DNA damage. Thus, an effective somatic mutation detection pipeline is needed to address these challenges across diverse clinical tumour samples. In the initial setup of the analysis, I opted for relatively lenient filtering criteria, with parameters such as count and coverage set at low thresholds (e.g., 2), and a required significance threshold of 1 or 5%. This approach aimed to maximize the inclusion of information during data processing, to combat mutation caller challenges. However, during the subsequent data analysis phase, I sought to enhance the rigor of determining significant results. To achieve this, I employed a frequency metric to discern variants with infrequent occurrences and filtered them out accordingly. This adjustment ensured a more stringent selection process, emphasizing the significance of identified variants based on their prevalence within the dataset. Building upon this refinement, further investigation was conducted into variant frequency occurrence, as low frequencies could indicate a sequencing artifact rather than genuine variant

detection (Figure 3.12). The trend for RNA related datapoints to be in a higher count than DNA was a continued trend in this analysis as well.

The observed data reveals an initial spike at the onset of both plots (Figure 3.12A,B), occurring at position 5 for DNA and position 1 for RNA. Notably, this position aligns with the predefined hard cut-off point configured during the workflow setup phase (Figure 3.2E and Figure 3.7E for DNA and RNA, respectively). Such spikes are likely indicative of mostly background noise rather than genuine biological signals. As the initial spike stabilizes and a subsequent spike emerges, it is inferred that this transition marks the onset of genuine data. Consequently, to effectively delineate between noise and meaningful data, a cut-off value of 20 was selected. This value is deemed appropriate given its ability to capture the genuine signal while effectively filtering out noise, thus ensuring the robustness, reliability of the analysis outcomes, enhance stringency in variant calling, and ensure statistical confidence in the identified variants. This threshold helps prioritize variants with frequencies that are significantly different from background noise, thus minimizing false positive calls and improving the reliability of the results. Notably, the ratio between single nucleotide variants (SNVs) and other mutations was maintained at approximately 10% with this cut-off, despite the substantial difference in variant counts observed between DNA and RNA samples. This suggests that the chosen cut-off effectively balances sensitivity and specificity in variant calling while accounting for the inherent differences between DNA and RNA sequencing data.

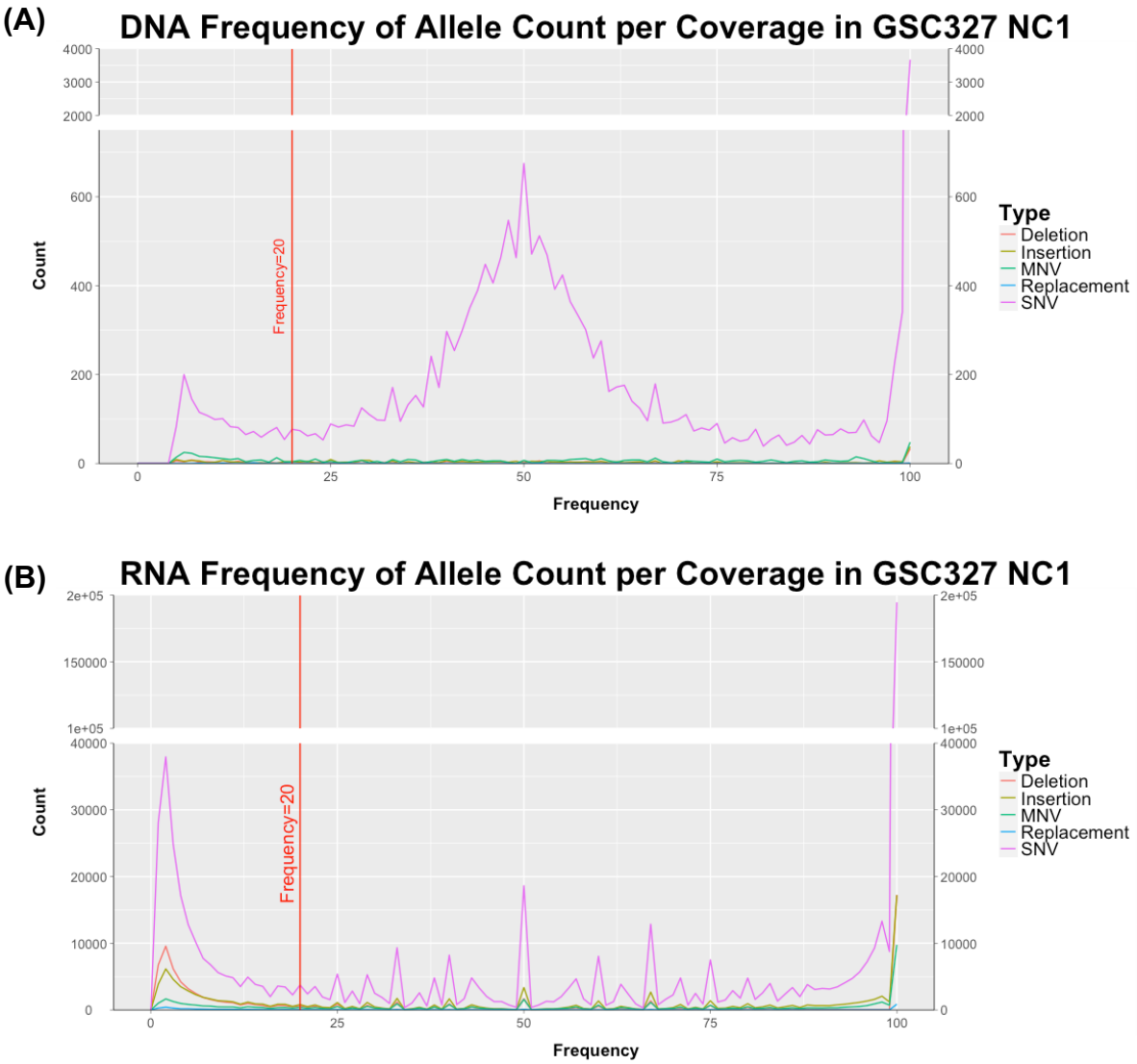


Figure 3.12: DNA and RNA frequency count distribution in GSC327 normal control 1 (A) DNA frequency distribution of allele count per coverage for the different variant types (SNV, MNV, Insertion, Deletion and Replacement) (B) RNA frequency distribution of allele count per coverage for the different variant types (SNV, MNV, Insertion, Deletion and Replacement). Vertical red line marks the frequency 'cut-off' of 20 in both plots.

Additionally, the tabulated variant file also included reference alleles in the overall count of mutations. Therefore, those variants, that were the "baseline" or reference alleles, were removed. The analysis that produced Figure 3.11 was repeated with now redundant reference alleles and low frequency variants filtered out (Figure 3.13), producing a reliable count and number of mutations occurring among both DNA and RNA sequences.

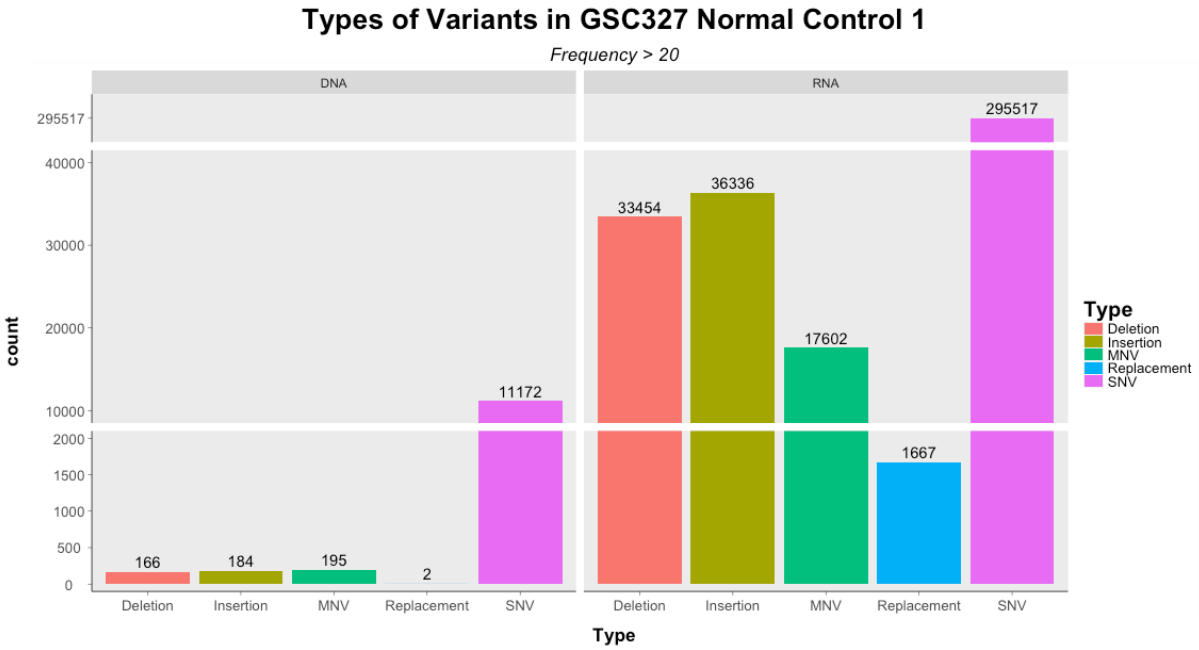


Figure 3.13: Types of variants in GSC327 Normal Control 1 (filtered) "BOTH" column is missing as CLCBio does not provide this metric for DNA-RNA variant comparison. "DNA" and "RNA" represent those DNA and RNA variant analyses respectively. Each type of variant is coloured shown on the x axis, on the y axis their number of count or occurrences is shown. White gaps show a brake in the y axis scale, to better visualize variant types with less counts. Here, reference alleles and variant with lower than '20' frequency have been removed.

3.4. DAVID pathway analysis

After observing the abundance of single nucleotide variants (SNVs) within the cell's DNA and RNA, I proceeded to identify the genes corresponding to these mutations and conducted pathway analysis using DAVID (1, 2). The results of the DNA and RNA pathway analyses largely showed overlapping pathways between filtered and unfiltered analyses (Appendix Figure 3). However, RNA variant pathways did not consistently overlap with DNA variant pathways, likely due to the uneven depth of sequencing between RNA and DNA (Figure 3.14, Appendix Figure 3A,C). The presence of immunity pathways associated with bacteria, viruses, and infections in the pathway analysis was anticipated, given their relevance to cancer biology and tumour microenvironment. However, the unexpected identification of pathways related to Huntington's disease and Parkinson's disease raised intriguing questions. These neurological conditions are primarily characterized by protein misfolding and aggregation within brain cells, leading to neuronal dysfunction and degeneration. The appearance of pathways associated with these neurodegenerative disorders in cancer samples prompts speculation about potential connections between the molecular mechanisms underlying neurodegeneration and carcinogenesis. It raises the intriguing possibility that the protein pathways implicated in Huntington's and Parkinson's diseases may also be dysregulated or mutated in cancer, suggesting a novel feature in cancer biology that warrants further investigation. This unexpected finding highlights the complexity and interconnectedness of biological pathways across different disease contexts and underscores the importance of comprehensive pathway analysis in uncovering novel insights into cancer biology.

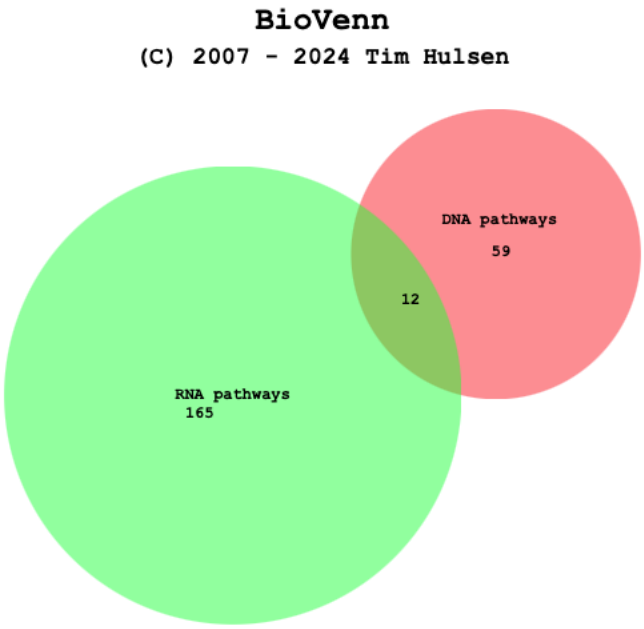


Figure 3.14: Venn diagram of DNA and RNA pathways. This is a comparison of the pathways resulting from SNV gene lists from both DNA and RNA variant analyses. Overall there are notably more pathways present in the RNA analysis than DNA, 165 to 59 unique pathways, respectively, with only 12 pathways appearing in both lists. Full list can be found in the appendix (Appendix Table 7)

Notably, ECM signalling emerged as a pathway with significant variation (Figure 3.15). Among the listed pathways, I singled-out ECM signalling due to its relevance in subsequent tissue RNAseq and proteomics de novo analyses that I have carried out (findings that I later validated using pathological methods (CHAPTER 6 and CHAPTER 7, respectively)). Within this pathway, the COL6 gene family, particularly COL6A3, stood out as one of the highly mutated genes in glioblastoma (Appendix Figure 2) and this is the target I analysed in GBM tissue then validated using IHC (CHAPTER 6 and CHAPTER 7, respectively).

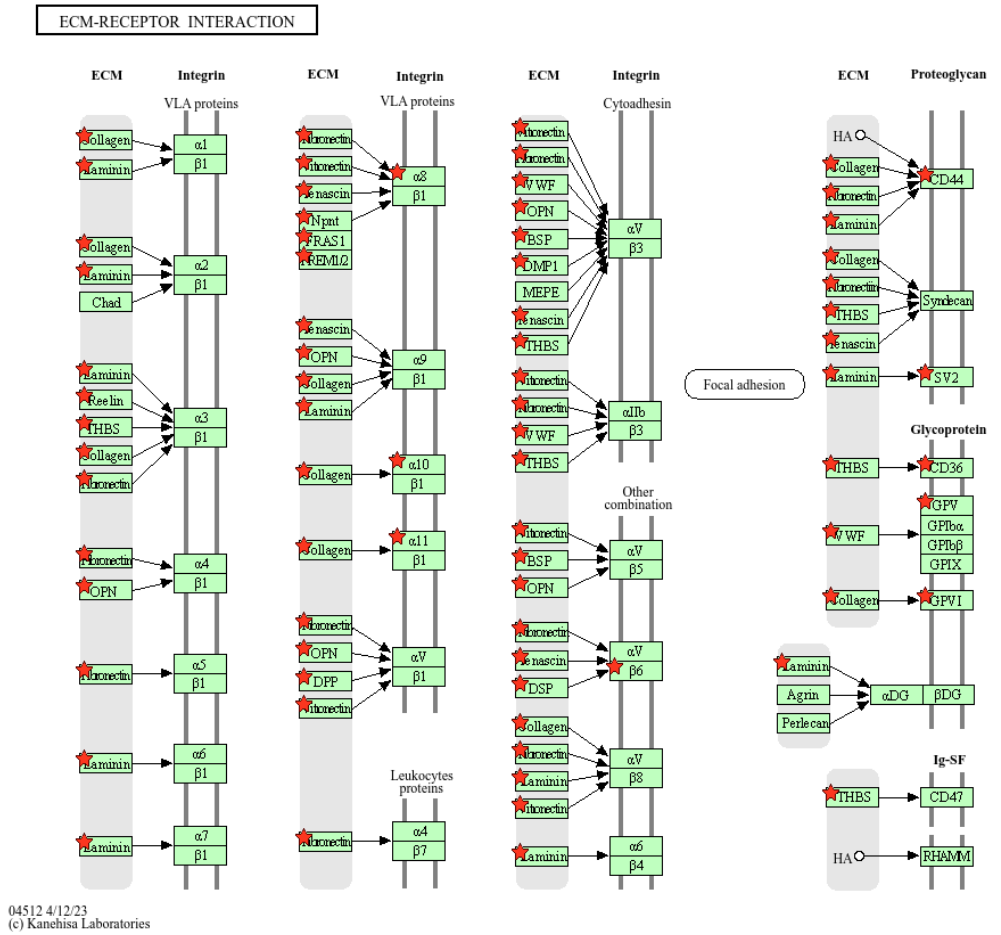


Figure 3.15: ECM-receptor pathway result of DAVID pathway analysis where red stars signify those genes that were present in the final filtered SNV DNA gene list

In conclusion, the utilization of CLC Genomics Workbench for DNA and RNA variant analysis has proven to be a valuable asset in establishing the molecular landscape of glioblastoma (GBM) in under normal (control) conditions and reveal potential therapeutic targets. Leveraging this powerful bioinformatics tool has enabled a comprehensive exploration of the genetic aberrations underlying GBM, shedding light on key mutations and pathways implicated in tumorigenesis and disease progression. Our approach, centered on utilizing CLC Genomics Workbench for variant analysis, has offered a distinct advantage in terms of efficiency and rapid data processing. By swiftly processing and analysing genomic data, we were able to establish a genetic baseline and genomic alterations present in GBM, facilitating a deeper understanding of the disease biology.

Comparison of the cell data with reference datasets from cBioPortal and existing literature on GBM revealed a convergence of mutations in our cell line with those observed in GBM tissues across patients. Notably, certain genes emerged as dominant contributors to GBM pathogenesis, warranting further investigation. Among these genes, TP53 exhibited inconsistencies between DNA and RNA sequencing, with instances where DNA reads were absent at sites where their presence would typically be expected according to the RNA analysis (sequencing is normally done deeper in RNA). This is one of the challenges in clinical tumour sequencing among many others in detecting somatic mutations, where tumour purity, indicative of the proportion of

cancerous cells in a sample, significantly influences mutation representation. However, estimates of tumour purity derived from light microscopy in pathology are notoriously imprecise as well. Moreover, somatic mutations occurring at low frequencies due to factors like low tumour cellularity or subclonal mutation architectures pose difficulties in detection, even with high-depth sequencing. While mutation callers can be adjusted for low-frequency variant detection, this adjustment often leads to increased false positives. Additionally, the use of formalin-fixed, paraffin-embedded (FFPE) samples, preferred for histopathological diagnosis, introduces artifacts from chemical DNA damage. Consequently, there is a pressing need for an effective somatic mutation detection pipeline to address these challenges across diverse clinical tumour samples.

In the initial phase of our analysis, lenient filtering criteria were employed to maximize the inclusion of information during data processing, mitigating challenges posed by mutation callers. However, in subsequent data analysis, a more rigorous approach was adopted to determine significant results. Utilizing a frequency metric, variants with infrequent occurrences were discerned and filtered accordingly, enhancing the selection process's stringency. Further investigation into variant frequency occurrence revealed an initial spike in both DNA and RNA plots, suggestive of background noise rather than genuine biological signals. To effectively distinguish between noise and meaningful data, a cut-off value of 20 was selected, ensuring robustness, reliability, and statistical confidence in the analysis outcomes. This threshold effectively balanced sensitivity and specificity in variant calling, maintaining a consistent ratio between single nucleotide variants (SNVs) and other mutations across DNA and RNA samples.

As an outcome of the consequent pathway analysis, the COL6A3 gene stood out as a highly mutated gene in GBM, implicating the extracellular matrix (ECM) signalling pathway in tumour progression. Notably, RNAseq of primary vs recurrent GBM tissues (CHAPTER 5.2) also identified dysregulation of COL6 family genes, supporting the biological relevance of the COL6A3-mutated GSC model for studying collagen-mediated tumour pathophysiology (currently under further investigation by a new student in the lab). In addition to ECM, other pathways were associated with infection, brain diseases (such as Huntington's and Parkinson's), and thyroid dysfunction. While these pathways present intriguing avenues for further research, it remains unclear which specific genes within these pathways should be prioritized for the development of miRNA-based therapeutics.

To address this gap in knowledge, the next phase of this study will focus on exploring the impact of hypoxia on mRNA expression, whether any hypoxia induced genes are mutated, and identifying potential targets for miRNA-based interventions. Through this iterative genomics-to-transcriptomics analysis of hypoxic GBM cell lines, I aim to identify candidate mRNA targets while concurrently assessing how closely the GSC transcriptional profile mirrors that of primary GBM tissue - a critical validation step for translational relevance.

CHAPTER 4. Identifying hypoxia induced mutated genes and signalling pathways in cell models

Glioblastoma multiforme (GBM), the most aggressive form of glioma, is characterized by hypoxic microenvironments that promote tumour progression and resistance to therapy. Within GBM, glioma stem cells (GSCs) are known to thrive in hypoxic conditions, contributing significantly to tumour growth and recurrence. Understanding the molecular mechanisms underlying hypoxia-induced changes in GSCs is crucial for developing effective therapeutic strategies.

In this chapter, I aim to delve into the intricate genetic landscape of GSCs under hypoxic conditions, focusing on the identification of hypoxia-induced genes and mutated pathways that could serve as potential targets for miRNA-based therapies. Specifically, two primary research questions are addressed:

i. What are the specific genes activated or upregulated in GSCs in response to hypoxia, and how might they contribute to tumour progression?
ii. Do hypoxia-induced genes in GSCs harbour genetic mutations or form functional networks that could be therapeutically targeted using miRNA-based interventions?

By elucidating the hypoxia-induced genetic alterations and pathways in GSCs, this chapter aims to provide valuable insights into novel therapeutic targets for GBM treatment. The findings hold promise for the development of precision medicine approaches tailored to target the unique genetic vulnerabilities of GSCs under hypoxic conditions, ultimately improving patient outcomes in the battle against this devastating disease.

4.1. Differential gene expression analysis using DESeq2

In chapter 2, I explained how the cell line samples were analysed using the RNAseq Nextflow pipeline (2.2.2 Nextflow RNAseq pipeline), and how the downstream analysis used DESeq2 (2.2.3 DESeq2 RNAseq pipeline). In this chapter, I will further describe the details of the RNAseq downstream analysis using DESeq2.

After initial import, the count data was converted to gene level abundance matrices for each sample. Then general exploratory analyses were performed to assess the behaviour of the samples such as principal component analysis (PCA). A PCA is a statistical technique used in data analysis and machine learning to simplify the complexity in high-dimensional data while retaining trends and patterns. It transforms the original variables into a new set of variables, the principal components, which are linear combinations of the original variables. These principal components are ordered by the amount of variance they explain, with the first component explaining the most variance in the data, the second the second most and so on. PCA is useful to visualize the structure of the data and identifying patterns. Nextflow automatically constructs a quality control (QC) document with a number of metrics including PCA and heatmap of sample-to-sample distances, where a clear batch effect was observed between the two cell lines (Figure 4.1) meaning the feature causing dissimilarity between the samples weren't the condition, but rather the samples themselves.

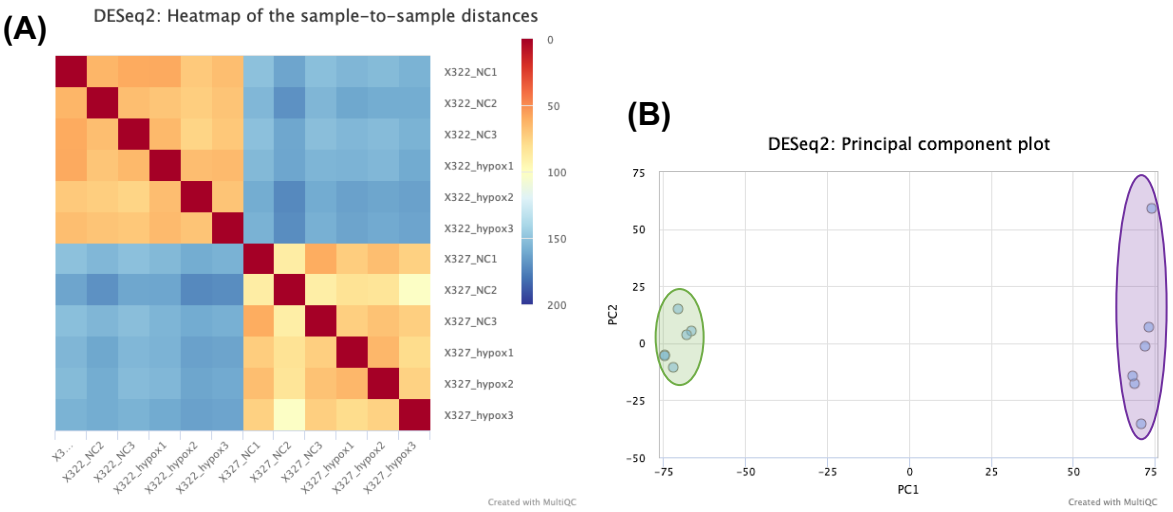


Figure 4.1: MultiQC analysis of glioblastoma stem cell lines (A) Heatmap (using built-in DESeq2) of sample-to-sample distances, where colours represent the degree of similarity or dissimilarity between samples. Blue colours represent high distances, suggesting significant dissimilarity between samples. Lighter shades of blue indicate slightly lower distances, implying moderate similarity. Yellow/Orange/Red colours represent low distances or high similarity between samples. Darker shade of yellow/orange/red indicate higher similarity, suggesting higher similarity between samples. Based on these colours, GSC322 and GSC327 are highly dissimilar. (B) DESeq2: Principal component analysis (PCA) of GSC322 and GSC327 depicting the multidimensional distribution of samples based on their gene expression profiles. Each point represents an individual sample, with the position determined by its principal component scores. The first two principal components, PC1 and PC2, are plotted along the x-axis and y-axis, respectively. The green circled group consists of GSC322 samples and the purple circled group consists of GSC327 samples, further showing the dissimilarity of the two cell lines.

Therefore, even though the sample size was small to begin with, further analysis of the cell lines was done separately. As a result, differences in samples likely due to the conditions of the experiment (control vs hypoxia) could be displayed clearer (Figure 4.2).

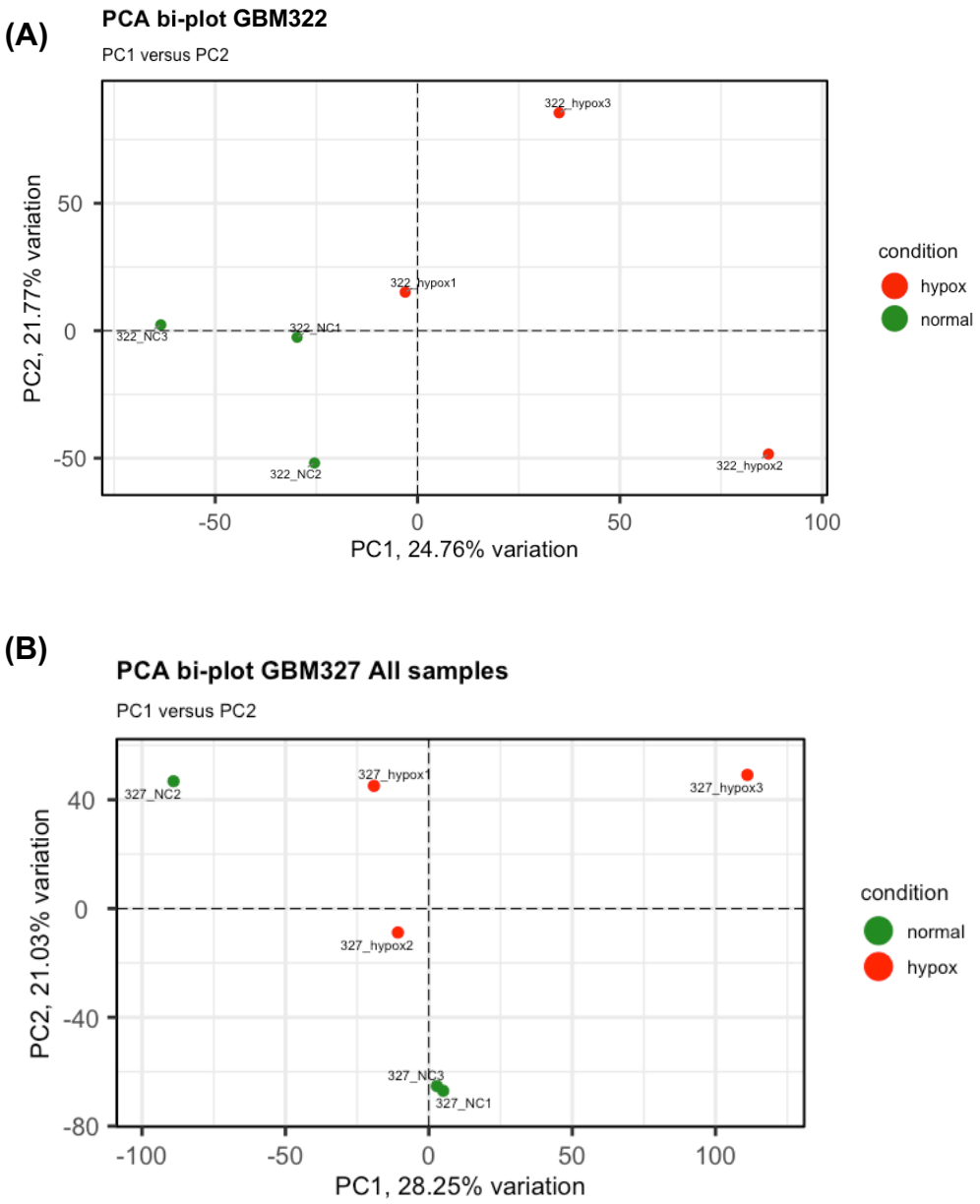


Figure 4.2: Principal component analysis (PCA) of GSC322 and GSC327 ((A) and (B) panels, respectively). Analysis conducted using the "PCAtools" package in R (8, 9). Each point corresponds to a sample coloured by green or red, which corresponds to their normal control (NC) or hypoxic conditions, respectively. The first two principal components, PC1 and PC2, are plotted along the x-axis and y-axis, respectively, stating the percent of variation those PCs explain.

In both cell lines, the principal components (PC1 & PC2) explain about 49% of the variation between the samples, which is normally considered poor; however, this is forgiven due to a few factors. The biggest factor being the small sample size coupled with the nature of the disease. Glioblastoma is a notoriously heterogeneous disease, which is what is observed here. Even in stem cells in a controlled, laboratory environment, the genetic variation between these samples is notable. That being said, GSC322 generally shows a trend of separation between normal and hypoxic cells. On the other hand, unknown artifacts arising during sequencing can also contribute to sample differences like in GSC327. After assessing the MulitQC analysis (returned by

Nextflow) and PCAs conducted separately, GSC327 normal control sample 2 (GSC327_NC2) exhibited anomalous characteristics (see the following panels of the MulitQC analysis, provided as supplementary information document: (i) SALMON DESeq2 PCA plot – 327_NC2 is located somewhat further from the group than the rest of the samples; (ii) Biotype Counts – counts are lower than other samples; (iii) Genomic origin of reads – counts are lower than other samples; (iv) Gene Coverage Profile – lowest coverage among all samples; (v) Mapped reads per contig – doesn't follow the pattern as other samples; as well as its location in Figure 4.2); therefore, it was determined to be an outlier, hence removed. According to the resulting PCA of GSC327, explainability of the variation between samples improved to almost 60% (Figure 4.3).

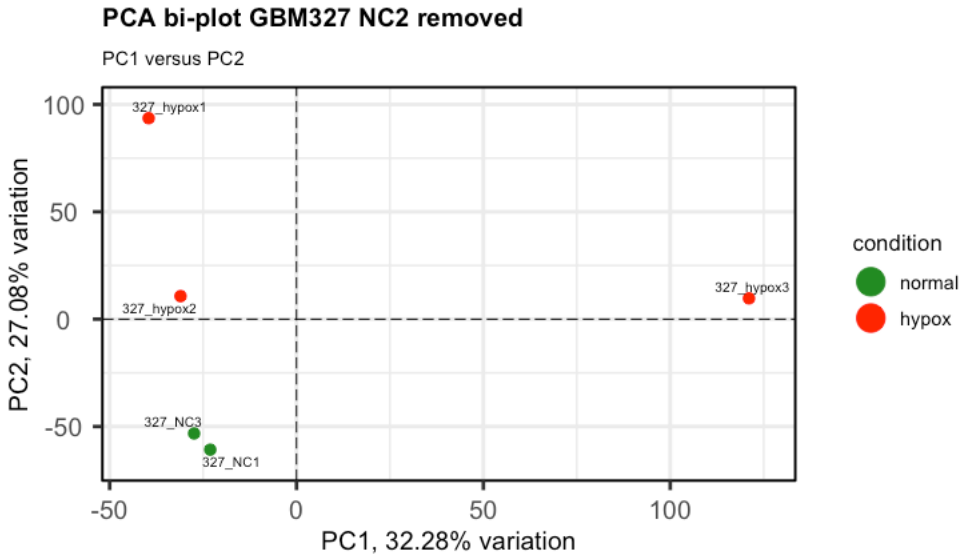


Figure 4.3: Principal component analysis (PCA) of GSC327 PCA without NC2 samples. Analysis conducted using the "PCAtools" package in R (8, 9). Each point corresponds to a sample coloured by green or red, which corresponds to their normal control (NC) or hypoxic conditions, respectively. The first two principal components, PC1 and PC2, are plotted along the x-axis and y-axis, respectively, stating the percent of variation those PCs explain. The overall explainability of the variation between samples has increased to almost 60% compared to Figure 4.2.

Following QC where plots like sample-to-sample distance heatmaps and PCAs were drawn, DESeq2 was used for differential gene expression (DGE) analysis. Although there are many other tools available for DGE, such as edgeR, Limma, SAMseq, CuffDiff2, etc; DESeq2 was chosen as it has a consistently low proportion of false discovery even with small sample sizes (216). DESeq2 was run in accordance with its manual (217) and differentially expressed genes were plotted using the *EnhancedVolcano* (218) package in R. The produced Volcano plots (Figure 4.4A,B) show the difference in expression between normal and hypoxic states of genes. Along the x-axis, the $\log_2$ fold change (FC) is signalling a given gene's over expression (right side of plot) or under expression (left side of plot) in the hypoxic state compared to the normal. A gene's movement on the y-axis is determined by our confidence in its expression value or its p-value (on the $-\log_{10}$ scale). Therefore, the higher a gene is located on the y-axis, the more confident we can be of that result; however, the lower the gene is located, the less confident we are. In summary, the more right and higher the gene is located, the more confidently we can say that the gene is overly expressed.

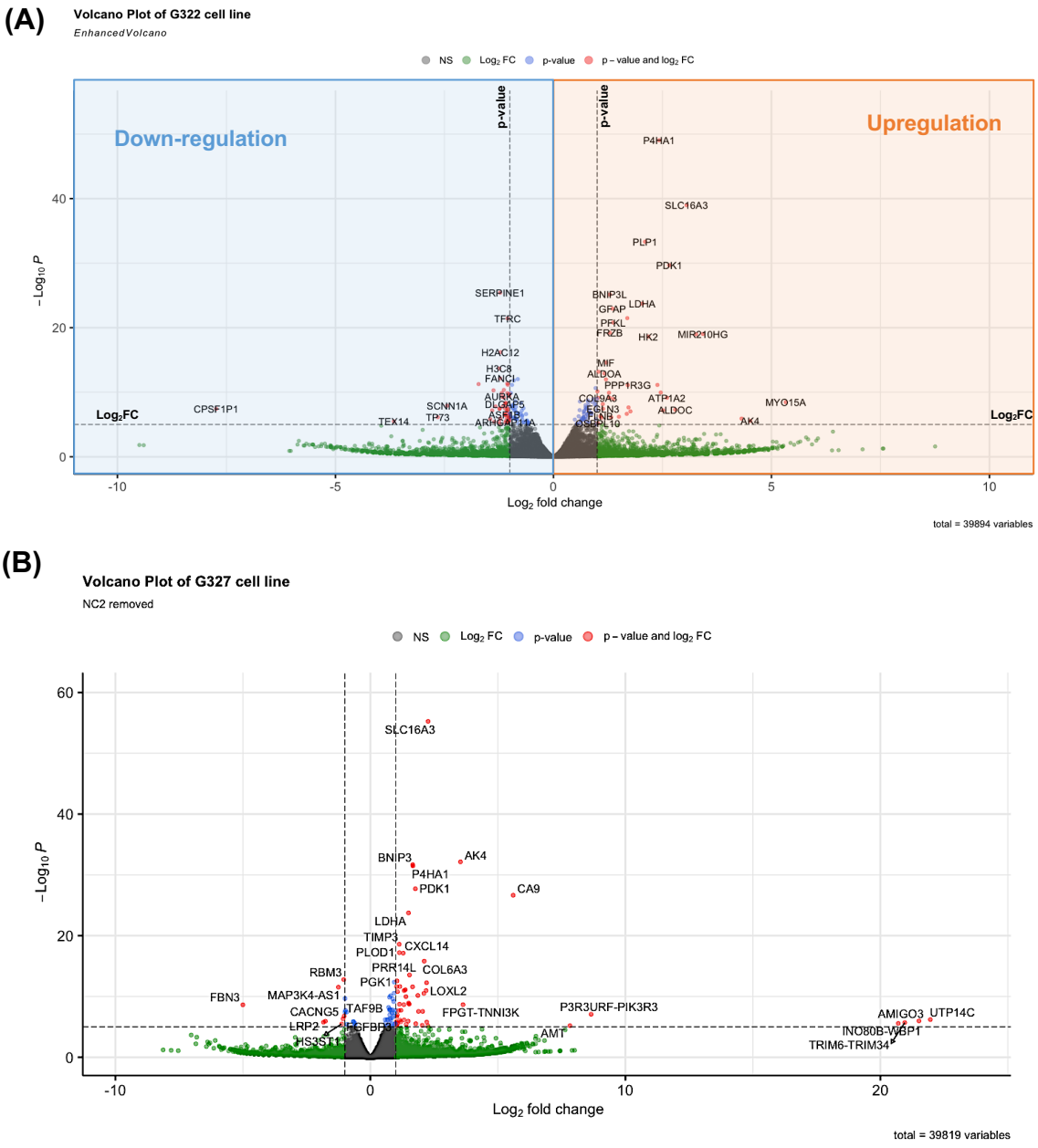


Figure 4.4: Volcano plots of GSC322 and GSC327 differential gene expression. Panel (A) was coloured in to highlight important features of the plot. A volcano plot illustrates, here, illustrates the differential gene expression between two conditions: normoxia and hypoxia. Each point represents a gene, with the x-axis indicating the fold change in expression and the y-axis indicating the statical significance. Genes exhibiting significant upregulation (orange shading) or downregulation (blue shading) are depicted as points located towards the upper corners of the plot. Non-significant (NS) genes are coloured grey; green coloured genes signify their fold change, but not statistically significant; blue colour stands for a significant p-value in the reading, however, non-significant fold change; finally, red points are significant in both fold change and p-value metrics. In panel (A) GSC322 was plotted, and in panel (B) GSC327.

At a first glance, the two volcano plots are quite different in both shape and the containing expressed genes besides a few genes like AK4 (219), LDHA (220) and BNIP3 (221), which are well-known or have been shown as important in glioma or in other cancers. However, as discussed before, the heterogenic nature of glioblastoma explains these differences, hence why the two cell lines were analysed separately.

The upregulation of certain genes within the cell lines highlighted four distinct signalling pathways, identified through DAVID analysis. Firstly, the involvement of genes such as HK2, LDHA, PDK1 and SLC16A3 displayed the significance of the central carbon metabolism pathway in cancer progression. This pathway plays a pivotal role in providing energy and biosynthetic intermediates essential for tumour growth and survival (50). Secondly, the overexpression of EGLN3, LDHA, HK2 and PDK1 highlighted the activation of the HIF1 signalling pathway, a key regulator of cellular responses to a hypoxic environment during and after tumorigenesis. This pathway orchestrates various adaptive mechanisms to facilitate tumour adaptation to low oxygen conditions, promoting angiogenesis, metastasis, and resistance to therapy (37). Additionally, the presence of genes such as AK4, CA9, LDHA, and P4HA1 showed the activation of the metabolic pathway to the introduction of hypoxic environment, which governs diverse cellular processes crucial for tumour progression, including proliferation, survival, and metastasis (219). Finally, the observed upregulation of HK2 and LDHA also highlighted the interplay of the same genes between multiple signalling pathways as these genes also participate in glycolysis/gluconeogenesis (48, 54). These pathways play a central role in meeting the increased energy demands of tumorigenesis and provide and provide essential metabolites for anaerobic respiration, facilitating tumour growth and progression. Collectively, the identification of these upregulated genes in GSCs highlights their pivotal roles in driving key oncogenic pathways implicated in cancer development and progression.

In GSC322, two upregulated genes, MYO15A and MIR210HG were interesting to note fir distinct reasons. Firstly, MYO15A is associated with profound, congenital, neurosensory, nonsyndromal deafness and traditionally unrelated to oncogenic pathways, prompting the question – why is it present? On the other hand, while MIR210HG is associated with varicocele and oncogenic processes such as osteogenic sarcoma, it is a long noncoding RNA, which acts as a negative regulator of miRNA 210 (MIR210). Despite their different roles, the upregulation of these genes highlights the nuanced diversity of glioblastoma cell lines.

One gene in particular, EGLN3, that was over expressed in both samples (Figure 4.5), sparked interest. Upon further examination, a member of the EGLN gene family (EGLN1) was found to play a key role in the hypoxia inducible factor alpha (HIFα) signalling pathway. Interestingly, the role of EGLN3 in the cell was previously uncharacterized, which led me to consider it as a targetable novel hypoxia induced gene. Therefore, it was further analysed under laboratory conditions.

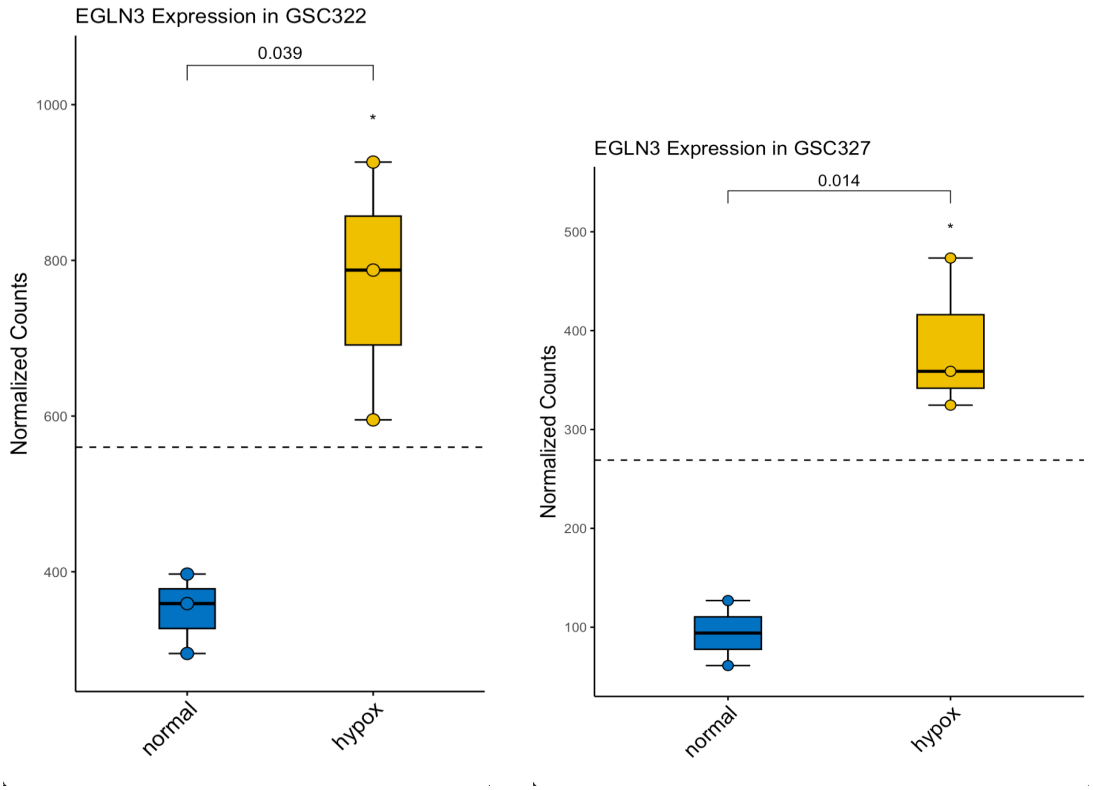


Figure 4.5: Analysis of Variance plot of the differential gene expression of EGLN3 gene in GSC322 and GSC327. Analysis of Variance or ANOVA plot consist of bars representing the mean expression level of EGLN3 in a distinct experimental group, with error bars indicating standard error of the mean. Statistical significance (marked with asterisks and the exact value above the bars) of expression differences is assessed using ANOVA, providing insights into the variability of EGLN3 expression across different cell lines and conditions. In both cases EGLN3, though marginally, upregulated under hypoxic conditions compared to normal.

4.2. Target validation in hypoxic cell line targets

The aim of these experimental assays was to validate whether GSC models express proteins of interest in the basal and/or hypoxic state. The selection process primarily focused on those targets that exhibited significant induction by hypoxia in either of the RNAseq cell line analyses – GSC322, GSC327. These targets underwent PCR, qRT-PCR, and pathway analyses to determine their significance and targetability in the bigger cancer cell signalling cascade picture. Furthermore, if results were significant, targets would be further quantified using immunoblotting.

4.2.1. Target validation of EGLN3 hypoxic cell line target

There are several highly expressed hypoxia-induced genes in both 322 and 327 GSC models (Figure 4.4AB). These genes include AK4, CA9, BNIP3, PDK1, SLC16A3, and many more. One of the most novel genes relating to hypoxia was EGLN (also known as PHD), more specifically EGLN3, which is a member of the prolyl hydroxylase enzyme family that target hypoxia-inducible factor (HIF), hypoxic pathways (3). HIF and its pathways have been in great focus in many areas of research especially cancer research (222). HIF is a heterodimer consisting of an alpha and a beta subunit. The alpha subunit is a basic helix-loop-helix DNA-binding protein that undergoes rapid degradation in normoxic cells (3). HIFβ, unlike HIFα, is independent of oxygen levels; therefore, the subunits are only able to bind when oxygen level declines, as HIFα

subunit becomes more stable (3). The formed HIFα-HIFβ complex undergoes nuclear translocation, binds to hypoxia response elements (HREs), and transcriptionally participates in many gene regulatory processes to adapt to the low oxygen environment (3, 223, 224).

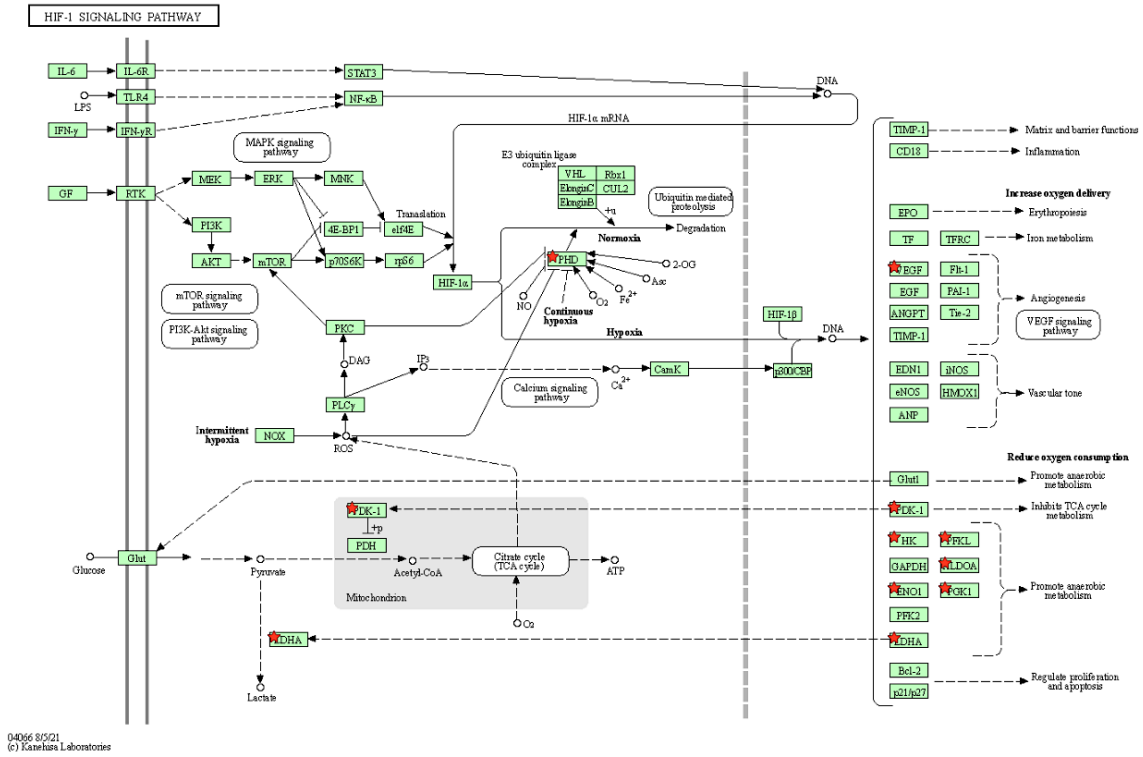


Figure 4.6: HIF-1 signalling pathway highlighting EGLN (PHD) gene family participation.
Red stars indicate genes that were overexpressed in differential gene expression analysis. Pathway analysis done using DAVID (1, 2).

EGLN's 2-oxoglutarate (2OG)-dependent dioxygenases facilitate the HIF prolyl hydroxylation, with the help of iron, oxygen. EGLNs have a somewhat low affinity to oxygen, which makes them an important player in the oxygen sensing machinery in cells because they can detect oxygen level fluctuation in a physiologically relevant concentration (3). Therefore, the EGLN-HIF system has been evolutionarily well conserved; to the point that higher level organisms, such as mammals, produce three different canonical HIF prolyl hydroxylases: ELGN1, 2 and 3 (also known as PHD2, PHD1 and PHD3, respectively) (Table 4.1).

ENZYME NAME	SPECIFIC FEATURES
EGLN1 (PHD2)	• Lowest O_2 affinity (main sensor) • Embryonic knockout is lethal
EGLN2 (PHD1)	• Estrogen-inducible • Transcript unaffected by hypoxia • Potential oncogene
EGLN3 (PHD3)	• Apoptosis regulator • Multiple non-HIF target candidates

Table 4.1: Summary of EGLN isoform functions. Table adapted from Ivan et al (3)

Several studies, with reference to genetic activity of specific isoforms, have uncovered overlapping and specific functions that are particularly influential for physiology, disease and drug discovery (223, 224). The significance of this pathway is highlighted by the identification of genetic variations (polymorphisms) of EGLN1 and EPAS1 among human populations living in extremely high altitudes (3, 225). Although EGLN1 is the primarily recognised sensor for hypoxia and oxygen levels, the other isoforms (EGLN2 and EGLN3) also possess enzymatic properties consistent with their sensing roles and; therefore, contribute to the regulation of HIF in certain contexts. For instance, when EGLN1 is acutely eliminated in the liver of mice, it results in pulsatile activation of the standard HIF target, Erythropoietin (EPO), likely due to compensatory mechanisms involving EGLN2 and EGLN3. However, when all three isoforms are eliminated, sustained and elevated production of EPO in the liver occurs (3, 226).

Therefore, as EGLN3 has not been shown to have a specific role in GBM, I evaluated its expression in the cell models using qRT-PCR and immunoblotting.

Although the RNAseq analysis shows a clear induction of EGLN3 RNA in both cell lines (Figure 4.4A for GSC322, Appendix Figure 4 for GSC327, Figure 4.5), this otherwise observed in the cell models using qRT-PCR (Reverse Transcription–Polymerase Chain Reaction). This is a process where RNA is used for nucleic acid amplification. However, due to RNase activity, RNA is very unstable. To overcome this, reverse transcriptase, which is an RNA-dependent DNA polymerase, is used to generate a more stable, complementary DNA or cDNA. The template RNA is subsequently degraded and PCR amplification proceeds as normal. The primers were chosen using the Primer-BLAST tool by NCBI (227, 228). Picking unique primers for PCR is one of the most important factors for the success of the experiment. Primer-BLAST takes care of the otherwise tedious two-step process, where the primer flanking regions of interests are generated, then they are checked for other potential targets, which without this webtool would be very time consuming (228).

The two primer sets (Pair1 – forward: ATGCCAAGCTACATGGTGG, reverse: ATCTGGTTGCGTAAGAGGGC; Pair 2 – forward: GGGATGCCAAGCTACATGGT, reverse: GTAAGAGGGCTGCACTTCGT) were first evaluated using semi-quantitative PCR (Figure 4.7) in GSC322 model in both the basal and hypoxic states. The RNA used for the primer evaluation comes from the exact samples that were processed for shotgun RNAseq in Chapter 3 and Chapter 4.

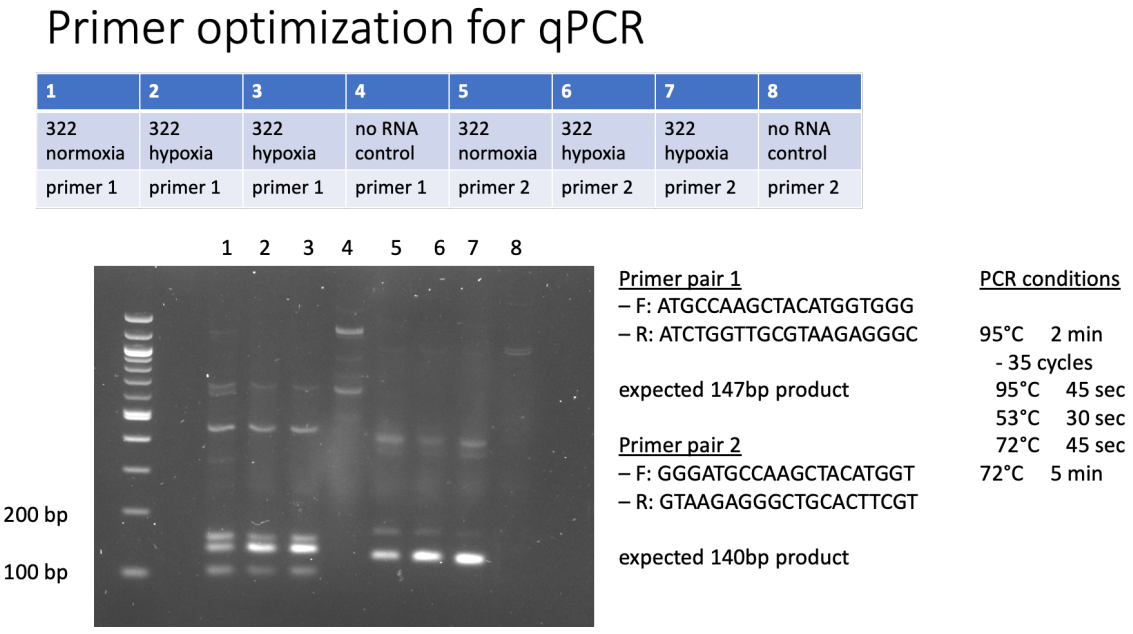


Figure 4.7: Primer optimisation for qPCR

The data shows that both primer pairs 1 and 2 (sequences are in Figure 4.7) could be used to identify the correct sized PCR fragment (140bp, see Figure 4.7, lanes 1-3 and lanes 5-7). The negative control was in lanes 4 and 8 showing an absence of a PCR product. Quantification of western blots were made using ImageJ (229) (Table 4.2).

	Lanes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Normalization	Area	5475.50	5726.62	7193.57	7856.28	7691.03	4435.08	5708.10	1722.67
	Percent	11.95	12.50	15.70	17.15	16.79	9.68	12.46	3.76
	Relative Density	0.70	0.73	0.92	1.00	4.46	2.57	3.31	1.00
Primer	Area	9597.98	14261.10	15731.23	1864.01	10012.45	15288.10	17223.23	847.60
	Percent	11.32	16.81	18.55	2.20	11.80	18.02	20.30	1.00
	Relative Density	5.15	7.65	8.44	1.00	11.81	18.04	20.32	1.00
	Adjusted Density	7.39	10.45	9.22	1.00	2.65	7.01	6.13	1.00

Table 4.2: Quantification of primer selection western blot for qPCR using ImageJ

PCR primer 2 was further evaluated in both GSC322 and GSC327 side-by-side (Figure 4.8). The data shows again that visually, EGLN3 is induced by hypoxia in GSC322 (Figure 4.8, lanes 1 and 2), and it is elevated in GSC327 (Figure 4.8, lanes 3 and 4).

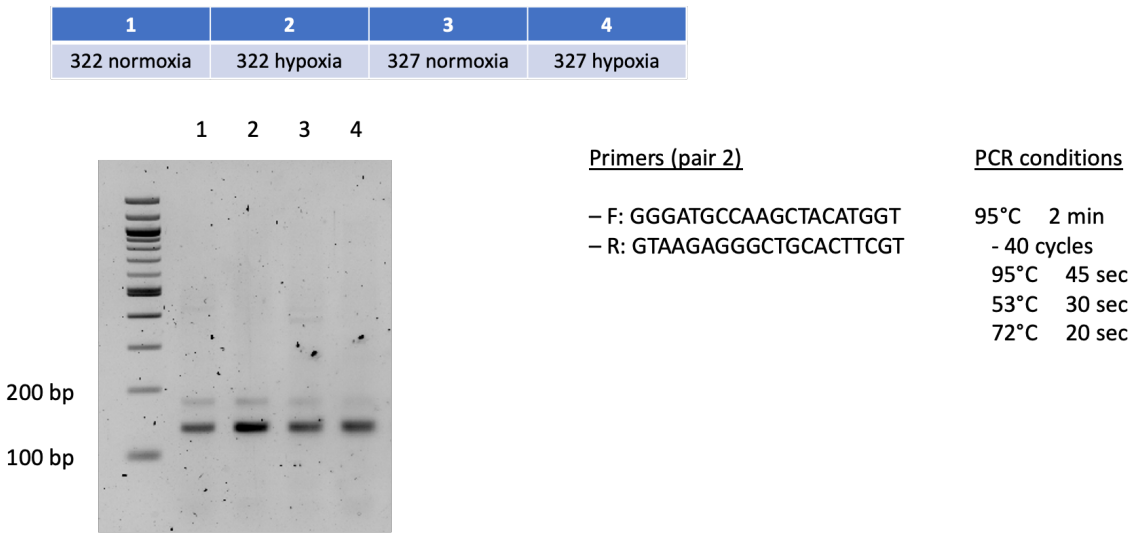


Figure 4.8: Primer 2 evaluation in both GSC 322 and 327

	Lanes	**1**	**2**	**3**	**4**
Normoxia	Area	4623.47	6702.32	5733.00	5224.33
	Percent	20.75	30.08	25.73	23.45
	Relative Density	1.00	1.45	1.00	0.91
Hypoxia	Area	19051.37	34090.95	25288.29	26403.17
	Percent	18.17	32.52	24.12	25.19
	Relative Density	1.00	1.79	1.00	1.04
	Adjusted Density	1.00	1.23	1.00	1.15

Table 4.3: PCR Quantification of primer 2 western blot of GSC322 and GSC327 using ImageJ

To quantify the absolute mRNA levels in the GSC322 and 327 models, qPCR was used (Table 4.3). It's a technique used to amplify and quantify specific DNA sequences, which, using fluorescent light, is monitored real time in a thermal cycler. It rapidly heats and chills samples utilising the physicochemical properties of nucleic acids and DNA polymerase. The PCR process involves temperature changes 25-50 times, with three stages: (i) separation of the DNA double helix at around 95 °C, (ii) primer binding to target at around 50-60 °C, (iii) DNA polymerization at 68-72 °C (230). The actual temperatures for each cycle can be seen on the figure below (Figure 4.9). (detailed methodology can be found in 2.1.3 Western blot validation).

		EGLN3		actin					
		mean C_T	st dev	mean C_T	st dev	ΔC_T	$\Delta\Delta C_T$	$2^{-\Delta\Delta C_T}$	$\log(2^{-\Delta\Delta C_T})$
322	normoxia	27.039	0.250	14.297	0.321	12.742	**0**	**1.0**	**0**
	hypoxia	25.992	0.038	14.490	0.224	11.502	**-1.24**	**2.4**	**0.373**
327	normoxia	27.721	0.057	13.770	0.436	13.951	**0**	**1.0**	**0**
	hypoxia	26.033	0.263	14.525	0.339	11.508	**-2.44**	**5.4**	**0.735**

Primers (pair 2)

– F: GGGATGCCAAGCTACATGGT
– R: GTAAGAGGGCTGCACTTCGT

PCR conditions

95°C 2 min
- 40 cycles
95°C 45 sec
53°C 30 sec
72°C 20 sec

Figure 4.9: qPCR to quantify absolute mRNA in GSC322 and 327.
Each measured in technical replicates.

The qPCR quantification values can be observed on the figure above (Figure 4.9) As before, the two cell lines – GSC322 and GSC327 – were examined for EGLN3 expression against the housekeeping gene, actin. Each sample consisted of triplicate RNA from three pooled biological replicates, processed in technical replicates. The values returned by the qPCR machine, denoted as C_T for cycle threshold, are the cycle numbers where the PCR-generated fluorescence becomes distinguishable from the background noise. These qPCR values for normoxic and hypoxic conditions for GSC322 were 27.039 and 25.992, respectively; and 14.297 and 14.490 for actin, respectively. However, these numbers are not interpretable on their own. The term that represents fold change in this case is $2^{-\Delta\Delta C_T}$. To get there, firstly, ΔC_T needs to be calculated, where Δ refers to delta, the mathematical symbol to describe the difference between two terms or numbers. Therefore, ΔC_T is the difference in C_T values for the gene of interest (EGLN3) and the housekeeping gene (actin) (Equation 4.1).

$$\Delta C_T = C_T\,(gene\ of\ interest) - C_T\,(housekeeping\ gene)$$

$$12.742 = 27.039 - 14.297$$

Equation 4.1: Equation to calculate ΔC_T.
For the exact values see Figure 4.9 ΔC_T column ***Error! Reference source not found.***

So, for each sample, the ΔC_T was calculated (Figure 4.9 ΔC_T column). The next step is to find the difference between hypoxic and normoxic conditions ($\Delta\Delta C_T$). The resulting values of 12.742 ΔC_T for normoxia and 11.502 ΔC_T for hypoxia in GSC322 were substituted in the below equation (Equation 4.2).

$$\Delta\Delta C_T = C_T\,(treatment) - C_T\,(control)$$

$$-1.240 = 11.502 - 12.742$$

Equation 4.2: Equation for ΔΔCT calculation.
Difference between hypoxic and normoxic samples. For the exact values see Figure 4.9 ΔΔCT column.

As normoxia is the control in this experiment, it makes sense that the calculated $\Delta\Delta C_T$ is 0. Finally, to calculate the fold change ($2^{-\Delta\Delta C_T}$), we need to do 2 to the power of

negative $\Delta\Delta C_T$. Resulting in 1.0 for the normoxic samples and 2.4 and 5.4 for GSC322 and GSC327, respectively. However, it is always best to log transform these values before undertaking any further (statistical) analysis as these values ($2^{-\Delta\Delta C_T}$) are likely not normally distributed and heavily skewed. Therefore, the fold change is termed $\log(2^{-\Delta\Delta C_T})$ resulting in 0.373 and 0.735 for hypoxic samples in GSC322 and GSC327, respectively. In conclusion, the data shows that a reasonable induction of EGLN3 by hypoxia was detected using this method.

Despite the observations of EGLN3 mRNA elevation in RNAseq (Figure 4.5), PCR (Figure 4.8) and qPCR (Figure 4.9), EGLN3 protein was not induced by hypoxia in either of the cell models (Figure 4.10). For this western blot experiment, PHD3 Monoclonal Antibody (EG188e/d5 – Invitrogen) was used at 1:500 dilution. Although EGLN3 was not validated as a hypoxia induced protein, I terminated focus on this gene product but still evaluated its expression in GBM tissue (see below).

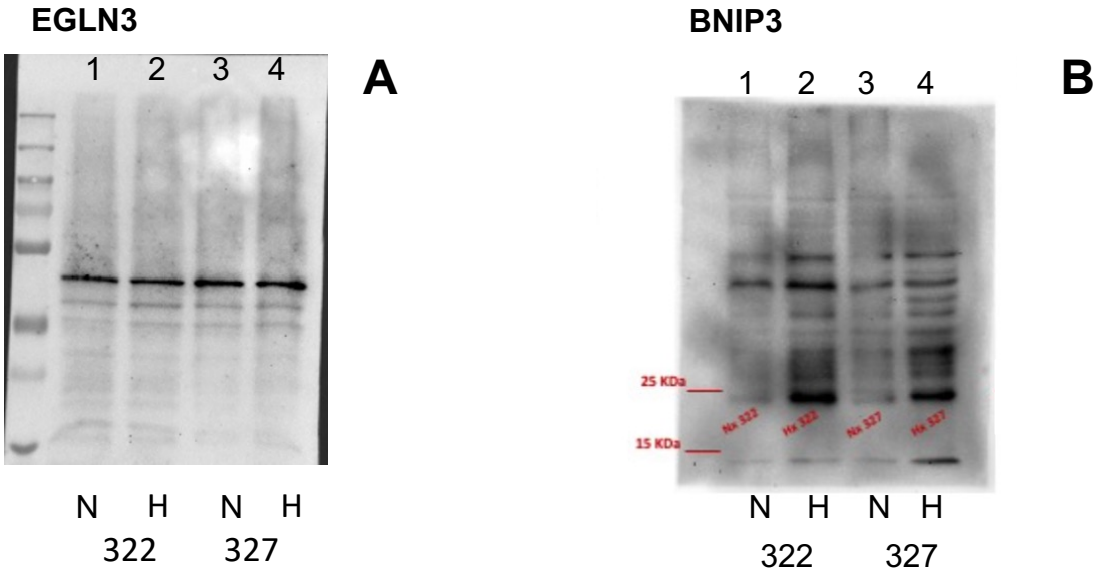


Figure 4.10: Immunoblot of EGLN3 and BNIP3
"N" stands for Normoxia and "H" stands for Hypoxia. A) Immunoblot of EGLN3, demonstrates that EGLN3 does not increase by hypoxia, as there is no intensity change of bands. B) BNIP3 was used as a control, which is a known hypoxia induced protein. There is a clear difference between normoxic and hypoxic samples for both cell lines, as band intensity increases from normoxia to hypoxia. BNIP3 control provided by E. Esposito

4.2.2. Evaluation of NDRG1 levels in GBM tissue

Like EGLN3, which is a relatively uncharacterized member of the proline hydroxylase family, N-myc downstream-regulated gene 1 (NDRG1), also part of the HIF1 signalling cascade, is considered a relatively novel target gene for glioblastoma treatment (231). It is part of the NDRG family, with its expression mainly located in the prostate, kidney, placenta, intestinal tissues and brain, primarily in the cerebral cortex (231, 232). NDRG1 has been shown to be induced by hypoxia in GBM; and correlates inversely with patient survival; therefore, it is considered a cancer suppressor gene (233).

Although NDRG1 was not among the prominent upregulated genes in either cell line, nonetheless, it was upregulated at least $\log_2$FC=1 in both cell lines, meaning NDRG1 was produced twice as more under hypoxia as under normoxic condition. Furthermore,

it was also detected in GBM tissue RNAseq analysis in a later chapter as part of a shared 13 and 39 gene signature set between tissue and the cell lines (GSC322 and GSC327, respectively) (CHAPTER 5.3). This highlights the practicality of using different transcriptomic samples to measure changes in gene expression. To test the validity of these *de novo* observations, I evaluated NDRG1 expression in GSC322 and GSC327 models in the presence and absence of hypoxia.

In conclusion, through RNA sequencing analysis, we uncovered sets of genes induced by hypoxia across our cell models, some were mutually expressed. Their upregulation highlighted four signalling pathways: central carbon metabolism, HIF1 signalling, metabolic pathways, and glycolysis/gluconeogenesis. Notably, the overexpression of EGLN3, alongside its association with the HIFα signalling pathway, suggested its potential as a novel hypoxia-induced gene; however, its activity was not supported by laboratory experiments. Additionally, the upregulation of MYO15A and MIR210HG in GSC322 highlighted the diverse genetic landscape of glioblastoma stem cells and the disease itself.

Some upregulated genes in the differential gene expression (DGE) analysis were found to be mutated in the basal state in GSC327, notably: AK4, LDHA, SLC163, BNIP3 – which was used as a known hypoxic target along the experimental tests. Moreover, our novel target that participates in the HIF signalling cascade, EGLN3, was also mutated at the RNA level. However, NDRG1 emerged as a hypoxia-induced, non-mutant-basal-state gene, that was validated under laboratory conditions. Notably, NDRG1 stands out as an interesting target to take forward, given its induction by Myc. While initial interests leaned towards EGLN3, the subsequent laboratory analysis, has proven NDRG1 to be more promising. In the next steps, I transitioned to patient tissue differential gene expression and subsequent laboratory tissue analyses. Transitioning to tissue analysis was essential to ensure the relevance and accuracy of our findings, given the potential for artifacts in cell lines. This strategic shift justifies the importance of validating findings in clinically relevant contexts, ultimately guiding the identification of the best possible targets for therapeutic miRNA medicines.

Based on these data, I can conclude the following, as per the original aims of this chapter:

i. Mutated and/or hypoxic pathways and targets were identified in GSCs, validated, or invalidated using the cell line only derived data.
ii. These validated data provide a template from which to compare to GBM tissue transcriptomics to determine whether features of GSC cell lines can model tumour tissue.

CHAPTER 5. Target Discovery And Validation in Glioma Tissue

As glioblastoma is such a heterogeneous disease, the complexity of attempting to recreate the tumour microenvironment raises the question whether cell lines are even good models to gain insight into this disease. In CHAPTER 3 and CHAPTER 4, I characterized the genomic and transcriptomic landscapes of GSCs This enabled systematic identification of candidate genes and pathways for subsequent validation in primary tumour tissue. To evaluate the physiological relevance of GSC-derived findings, I performed independent and parallel RNAseq analysis of primary glioblastoma tissues, enabling direct comparison of transcriptional profiles between model systems and native tumours. Tissues were provided by Dr Paul Brennan, a neurosurgeon consultant based at the Royal Infirmary in Edinburgh.

Similarly to the cell lines, glioblastoma tissue samples were sequenced (described in CHAPTER 2) then analysed using DESeq2 (preparation of samples, RNA sequencing and count analysis was done prior to my arrival to the group). As I had no control over the experimental design, there were no “normal” tissues collected in conjunction with tumour sample collection. Therefore, the dataset received only contained the various grade classifications of the tumour samples and some additional patient information such as age, sex, whether tumour was primary or recurrence, etc. (metadata available as supplementary information) The grades of glioblastoma are assigned based on many factors (Table 1.1); however, grades can also be viewed as a disease progression scale. Grade 4 tumours will be more aggressive, more advanced compared to grade 1 tumours. In addition, the number of grade 4 samples heavily outnumbered all other grades. This is likely due to the general poorness of early diagnosis of gliomas. As the focus of the study was the examination of glioblastoma (grade 4), lower grades were combined (grade 1-3) in the comparison to glioblastoma, which provided a better ratio at sample number distributions (grade 1: 1 sample, grade 2: 13 samples, grade 3: 13 samples, grade 4: 93 samples) and, also, highlights the progression of the disease to grade 4. Additionally, in a second assessment, primary tumours were compared to recurrent ones to look for possible biomarkers or driver genes responsible for recurrence of the disease, which is a very prevalent issue, even post-surgery.

5.1. Comparison of low- and high-grade gliomas

In this analysis, low grade (1-3) gliomas were combined and compared against high grade glioma (grade 4, glioblastoma) tissue samples (Appendix Table 8). The pipeline used for the tissue analysis was kept as similar as possible to the pipeline used in the cell line analysis in Chapter 4 to maintain consistency and increase comparability. The data was received as already pre-processed, kallisto counts. The counts were then imported using the `tximport` (234) and `biomaRt` (207) packages in R for DESeq2 differential gene expression (DGE) analysis (full list of loaded packages and versions can be found in Appendix 1 - Related to).

The advantage of using DESeq2 is that the actual differential gene expression step of the analysis is wrapped in a single function, `DESeq`. Then, a table of the results can be accessed using the `results` function that include information such as log2 fold changes and p-values.

As part of the `results` function, a number of additional parameters are customisable. Notably, I specified three arguments beyond the `object` term: `name`, `filterFun`, and `alpha`. It is always advised to specify key arguments of a function to guarantee the correct comparisons are performed. Under the `name` argument, I reinforced the comparison of low vs high grade samples, where 'low' is the control or base level of the comparison.

Secondly, in the DGE analysis, I employed the Independent Hypothesis Weighting (IHW) method for p-value adjustment and independent filtering, utilizing the `filterFun` argument. This approach significantly enhances the control of the false discovery rate (FDR) compared to the default Benjamini-Hochberg (BH) procedure. RNA sequencing (RNAseq) analysis inherently involves multiple hypothesis testing, as it assesses the expression levels of thousands of genes simultaneously to identify those differentially expressed between conditions. Each gene undergoes a hypothesis test to determine if there is a significant change in its expression level. Given the vast number of tests conducted, there is an increased risk of false positives—incorrectly identifying genes as differentially expressed when they are not. To mitigate this risk, methods such as BH and IHW have been developed to control the FDR, ensuring that the proportion of false discoveries is minimized. In contrast to BH, which applies a uniform threshold to all hypotheses and treats them equally without additional context, IHW utilizes covariates or auxiliary information to assign different weights (a non-negative number between zero and one) to each hypothesis. The input of this method is a vector of p-values (just like BH/FDR) and a vector of covariates (either continuous or categorical), in this case it's the sum of read counts per gene across all samples. Essentially, IHW takes each hypothesis test and groups them based on the supplied covariate, then calculates the number of discoveries (where the null hypotheses are rejected) using a series of weights (235, 236). These weights are iteratively adjusted until the method converges on the optimal weights for each group based on the covariate, maximizing the total number of discoveries (235, 236). Additionally, measures are implemented to prevent overfitting and to ensure the method scales efficiently handle millions of comparisons (235, 236). This adaptive approach allows IHW to enhance power and improve the detection of true positives by prioritizing more promising hypotheses based on relevant information. Although the BH procedure is simpler and widely used due to its straightforward implementation, it may lack the efficiency of IHW in scenarios where hypotheses vary in relevance or reliability. Therefore, IHW more effectively controls the FDR, enhancing the reliability of the results and making it a superior choice for improving the accuracy and robustness of my RNAseq analyses.

Lastly, I modified the `alpha` argument to 0.5 from the default 0.1, which the independent filtering of results parameter. It is an additional argument to control filtering of non-significant genes from the RNAseq data. By default, DESeq2 uses the mean of normalized counts as a filter statistic, optimizing the number of adjusted p-values below a specified significance level, denoted as `alpha`. The concept of independent filtering involves setting a threshold on the filter statistic to maximize the number of significant findings, while the genes not meeting this threshold have their adjusted p-values set to `NA`. This approach ensures that only the most promising hypotheses are tested, thereby improving the power of the analysis. The filter threshold value and the number of rejections at each quantile of the filter statistic are provided as metadata in the `results` object, allowing for detailed examination and verification of the filtering process. The value was changed, to allow for a more

balanced and less stringent analysis following the comprehensive FDR filtering provided by IHW.

Following successful DGE, various quality control and exploratory plots were drawn, such as PCA plots, scree plots, and sample-to-sample similarity heatmaps (Appendix Figure 9-Appendix Figure 13). Drawing quality control (QC) plots is an essential step in RNA sequencing analyses following differential gene expression (DGE) analysis to ensure the accuracy and reliability of the results. Several types of QC plots are commonly used, each serving a specific purpose in validating the data and the findings. Principal component analysis (PCA) plots are used to assess the overall variance in the dataset and to ensure that the primary sources of variation are biological rather than technical; therefore, validating the grouping of samples and detecting any outliers or batch effects. In conjunction with PCA plots, scree plots provide additional information about the proportion of total variance explained by each principal component, which is a linear combination of the original variables in a dataset that intends to capture the maximum amount of variance. By examining the scree plot, I can determine the number of principal components that capture the most significant variance in the data, helping to avoid overfitting and focusing on the most informative aspects of the data. Histogram plots of p-values provide insights into the distribution of p-values, helping to identify deviations from the expected uniform distribution (besides a peak at p=0) under the null hypothesis, which may indicate issues with the data or the statistical model used (237). Together, these QC plots were used to provide a comprehensive assessment of the quality of the RNAseq data post-DGE analysis. They help to highlight and address potential issues such as biases, batch effects, and outliers, ensuring that the conclusions drawn from the data are robust and reliable.

Firstly, the p-value histogram (Appendix Figure 9) produced an unexpected second peak around 0.8. Genes with small counts can negatively influence the data, which can be addressed by filtering out genes with small counts. Here, I removed all samples that had less than 6 counts.

Looking at the PCA plot at a first glance, which was drawn by ggplot2 after first transforming the data using the `vst` function, there were three patient samples (806, 898 and 3014) that clustered separately from the other 117 samples. The results of the scree plots (Appendix Figure 11Appendix Figure 13) suggested that the difference between these three samples and the rest explain a considerable amount of variation in the data (PCA overall 56% variance explainability). However, upon closer examination of the samples with regards to patient data, like age, gender, location of the tumour sample or data collection methodology, there was no discrepancy that could have caused the differences in these samples. Which has led me to conclude, as I had no access to the raw sequencing data, that the differences could only be a result of errors that occurred during sequencing, hence these samples should be treated as outliers and removed from further analyses. This decision was also backed up by our collaborator neurosurgeon consultant (Dr Paul Brennan, Royal Infirmary), who provided us with the tissue samples, and had no medical reasoning as to why these samples could differ from the rest.

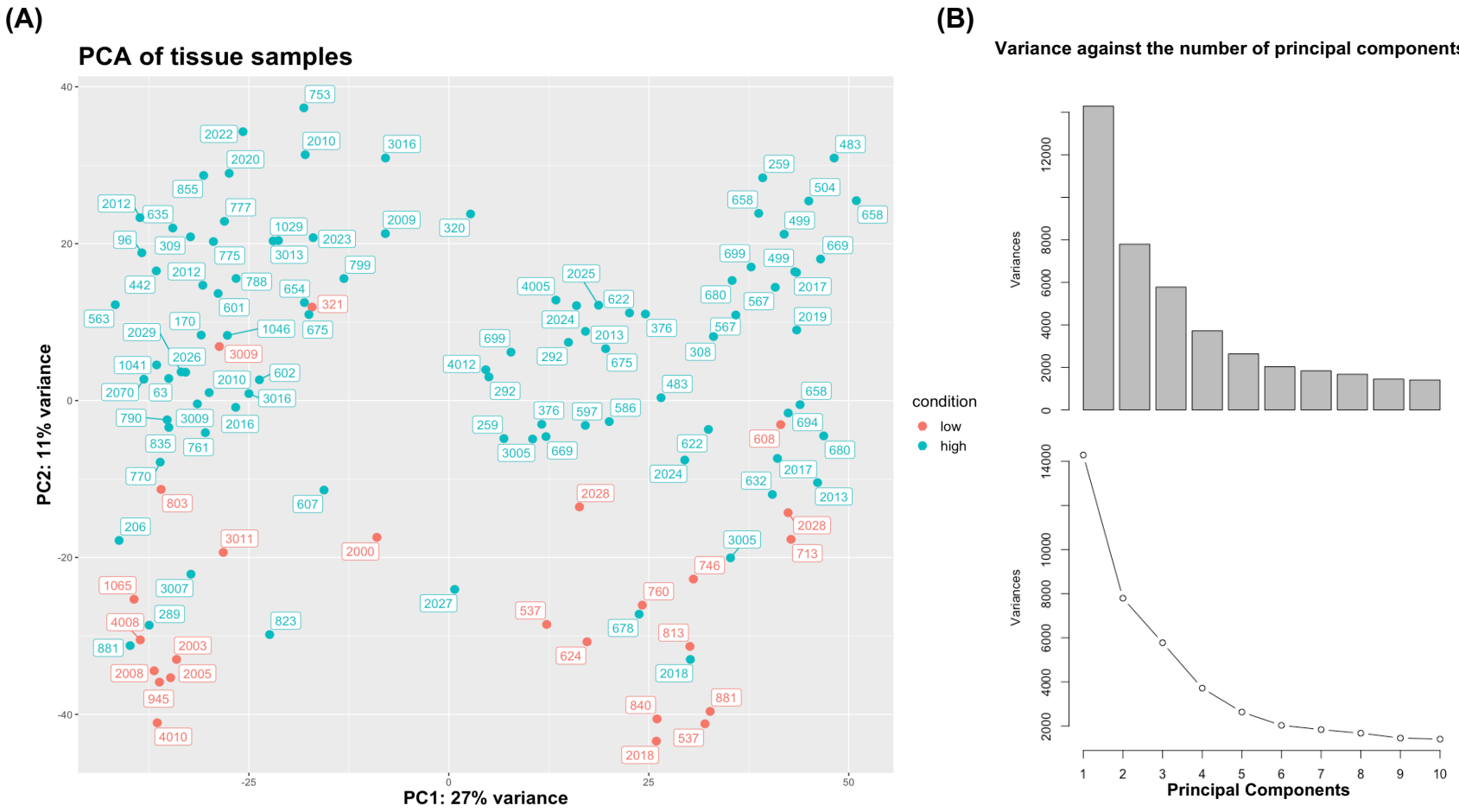


Figure 5.1: Exploratory data analysis plots

(A) Principal Component Analysis (PCA) of glioma tissue samples post-filtering of outliers. Each point represents a patient ID number in the dataset, colour-coded according to their assigned condition (low: grades 1-3, high" grade 4). Some patient IDs may appear twice meaning there was also a recurrent sample taken not just of the primary tumour (full list can be found: Appendix Table 8). This plot is a part of the second round of exploratory data analysis plots, where the intention was to observe how sample relations have changed after some samples were removed. The difference in variance between samples explained by the principal components have lowered to a cumulative 38% (27% and 11% for PC1 and PC2, respectively). In order, to explain the variance, more PCs need to be included like the scree plots (B) indicate as well. These plots display the cumulative variance explained by the principal components (PCs) as a function of the number of components. The x-axis represents the number of PCs, while the y-axis indicates the proportion of total variance explained. The curve demonstrates how adding more principal components incrementally reduces more of the dataset's variance, helping to determine the optimal number of components needed for efficient dimensionality reduction without significant loss of information. The bars indicate, how much additional variation explained by each PC. Where the curve starts to "plateau" or forms an "elbow" shape, indicates the ideal number of PCs. Here, one could argue in favour of either 2 or 6 PCs.

Following the removal of sequencing outlier, the refined PCA (Figure 5.1) reveals a more nuanced structure within the primary data. While the cumulative variance explained by the first two principal components is reduced (38%), the plot now clearly reveals several patient ID points that appear in duplicate (ex. 2018, bottom right of plot; 2012, top left of plot). These pairs represent matched primary and recurrent tumour samples from the same patient. Crucially, in many cases, the recurrent sample clusters mildly closer with samples from other patients than with its own primary counterpart. This pattern strongly suggests the presence of a pronounced technical or biological batch effect associated with sample recurrence, rather than a patient-specific signature dominating the variation. This batch effect likely stems from fundamental biological circumstances, like brain biopsy location, time apart from primary and recurrent sample taken. The biological shift in the tumour microenvironment or genomic evolution between primary and recurrence, compounded by potential technical variances in sampling timing or processing. However, due to lack of both biological and technical metadata, I couldn't verify the source of the batch effect. As a result, I just continued the analysis.

The final output of this analysis is a comprehensive results table containing several informative columns such as: "baseMean", "log2FoldChange", "ifcSE", "stat", "pvalue", "padj", and "weight". The "baseMean" column provides the average expression level of each gene across all samples. The "log2FoldChange" (log_2FC) column shows the fold change in gene expression on the logarithmic (base 2) scale between different conditions. The "ifcSE" column contains the standard error of the log2FC, providing an estimate of the variability on the log2FC. The "stat" column represents the Wald test statistics, which is the log2FC divided by its standard error (lfcSE), this value is then compared to a standard normal distribution to generate a two-tailed p-value. The "pvalue" column shows the probability that the observed difference in gene expression was due to chance. In DESeq2, p-value is calculated using the Wald test, where the hypothesis is that for each gene, there is no differential expression across the two sample groups (low grade vs high grade). Therefore, if the p-value is small ($p<0.05$), the null hypothesis is rejected because there's only a 5% chance that the obtained result is only due to chance. However, when many genes are being tested simultaneously, the error rate increases which is why IHW was also used for p-value correction resulting in the adjusted p-value term or "padj". The "weight" column is included to handle count outliers, with DESeq2 assigning a p-value of NA to those genes that were identified as containing outliers according to the Cook's distance calculation. This results table provides a detailed overview of the DGE analysis, facilitating the identification of genes with significant changes between low and high glioma conditions.

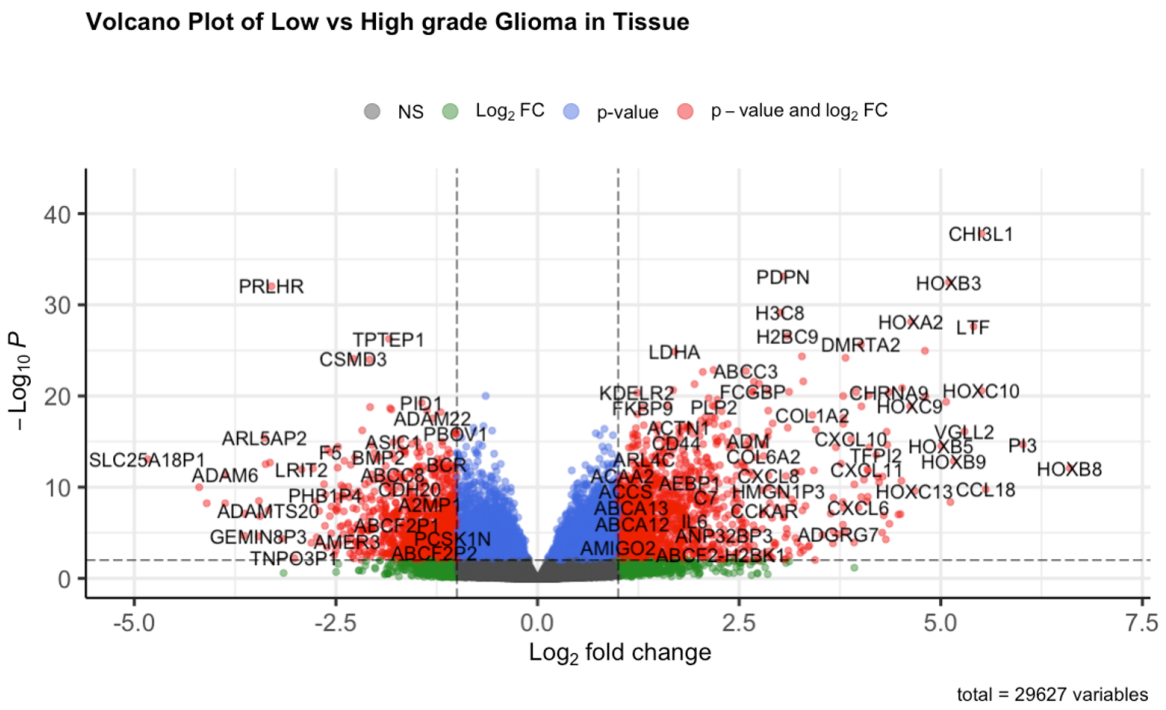


Figure 5.2: Volcano Plot of low vs high grade glioma tissue
Each expressed gene found in glioma tumours were plotted according to their differential expression levels. On x-axis, the expression intensity is shown in the form of Log2 fold change. On the y-axis, the negative log(10) p-value indicates the level of confidence in the fold change for each gene. Genes located on the left side of the plot (genes with negative fold change), are downregulated in this comparison. Conversely, genes located on the right-hand side (positive fold change) are upregulated. The dotted lines indicate filtering cut-off values, vertically, the fold change cut off (FCcutoff=1.0), and horizontally the p-value cut off (pCutoff=0.01). Genes that failed to meet these cut off values were colored green (p-value only), blue (fold change only) and grey (both). However, genes in red have a relatively significant fold change as well as a high confidence level of that result.

To visually interpret the results of DGE analysis, a volcano plot was drawn using the log2FC and padj values for each gene (Figure 5.2). In a volcano plot, the log2FC is plotted on the x-axis, representing the magnitude of change in gene expression between conditions, while the -log10 of the padj is plotted on the y-axis, indicating the statistical significance of the change. This allows for a clear distinction between genes are both highly differentially expressed and statistically significant, typically appearing as points far from the origin in the plot. The presence of several ‘housekeeping’ genes, such as LDHA (238) and PDPN (239) helped validate the results, confirming that the comparison is reliable. Interestingly, the volcano plot revealed unexpected genes and gene families with significant differential expression – homeobox (HOX) gene families (more information in the chapter) – which were investigated further as potential targets.

In summary, the differential gene expression analysis comparing low-grade and high-grade gliomas provided comprehensive insights into the molecular distinctions between these conditions and therefore the genetic profile during the progression of the disease. The use of DESeq2, with customizations such as Independent Hypothesis Weighting (IHW) for p-value adjustment and tailored filtering criteria, ensured robust and reliable results. The detailed table of results, including metrics such as log2 fold change and adjusted p-values, facilitated the identification of significantly differentially expressed genes. Visualizing these results through a volcano plot further highlighted key genes and gene families with notable expression changes,

including well-known glioblastoma genes such as LDHA and PDPN, as well as unexpected candidates like the HOX gene family. The presence of housekeeping genes validated the analysis, while the identification of novel differentially expressed genes have opened new avenues for miRNA target identification.

5.2. Comparison of primary and recurrent glioma tumour gene expression

Even though the primary objective of the project was to identify genes that play an important role during glioblastoma tumorigenesis, a small tangent from this is to observe what happens during tumour recurrence – as it can be considered another form of tumorigenesis. To explore this, we conducted a differential gene expression analysis comparing primary vs recurrent glioma samples. The analysis pipeline was kept as similar as possible to the previous low vs high glioma comparison to maintain consistency and ensure comparability.

As in the previous analysis, the dataset was pre-processed using `kallisto` counts, which were imported with the `tximport` and `biomaRt` packages in R for DESeq2 differential gene expression (DGE) analysis. However, it is important to note that the dataset was skewed regarding the number of grade 4 glioblastoma samples compared to other grades. Additionally, some samples were unpaired, meaning there were only either primary or recurrent samples for some patients. Therefore, for this analysis, only patients with both primary and recurrent samples were considered, ensuring a more accurate and meaningful comparison. This subchapter aims to shed light on the genetic changes that occur during glioma recurrence, providing insights into the mechanisms underlying this critical phase of the disease, which is a frequent occurrence (81).

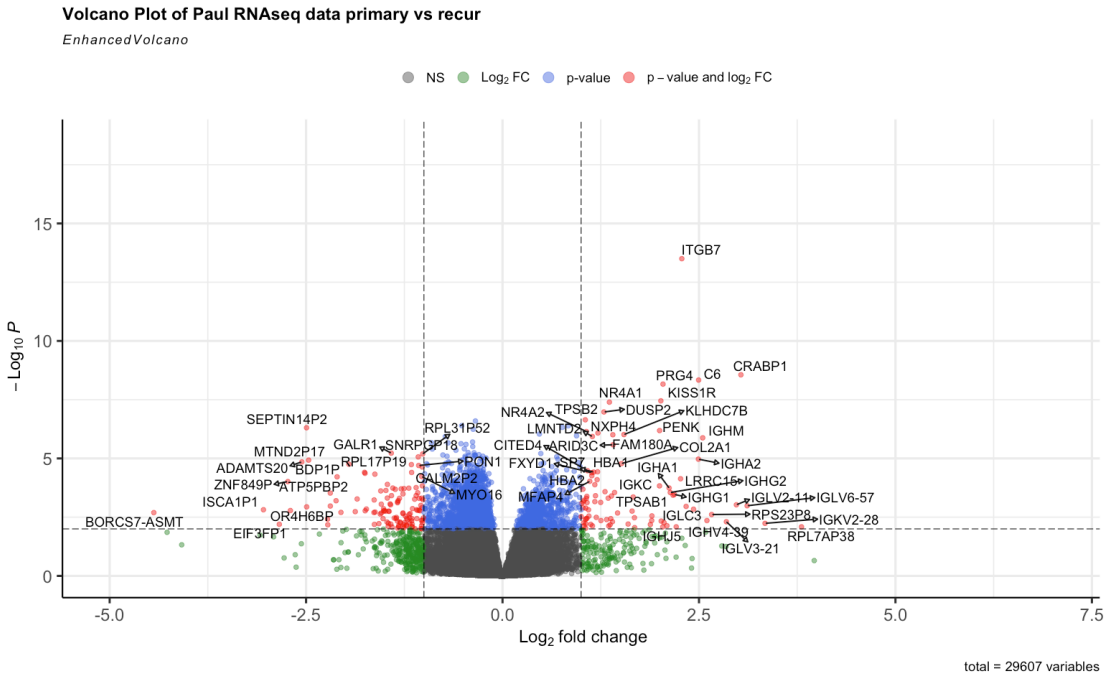


Figure 5.3: Volcano plot of primary vs recurrent tumour tissue comparison
Each expressed gene found in glioma tumours were plotted according to their differential expression levels. On x-axis, the expression intensity is shown in the form of Log2 fold change. On the y-axis, the negative log(10) p-value indicates the level of confidence in the fold change for each gene. Genes located on the left side of the plot (genes with negative fold change), are downregulated in this comparison. Conversely, genes located on the right-hand side (positive fold change) are upregulated. The dotted lines indicate filtering cut-off values, vertically, the fold change cut off (FCcutoff=1.0), and horizontally the p-value cut off (pCutoff=0.01). Genes that failed to meet these cut off values were coloured green (p-value only), blue (fold change only) and grey (both). However, genes in red have a relatively significant fold change as well as a high confidence level of that result.

Following successful patient filtering (now dataset contained 56 samples) and DGE, unexpectedly, these genes were drastically different from the low vs high glioma comparison (Figure 5.3). It was observed that the upregulated genes were predominantly participating in immune-related pathways. This unexpected finding could be attributed to a few characteristics of the tumour microenvironment that highlights the biological and pathological nuances between primary and recurrent GBM gene expression profiles. The recurrent GBM microenvironment is assumed to have undergone significant changes compared to primary tumours, especially in their early stages of tumorigenesis. These changes often involve increased infiltration of immune cells, altered cytokine profiles and a heightened state of immune activity. The prominence of genes related to immune pathways such as immunoglobulins and their regulators (eg. Complement 6 (C6) and Fc Epsilon Receptor II (FCER2)), reflects this evolved microenvironment. Mutations in C6, which are associated with immunodeficiency (240), and the role of FCER2 as a B-cell-specific antigen and a low affinity receptor of IgE (241), showcase the immune dynamics at play in recurrent tumours. Additionally, genes like IGHV1-2, IGHV1-69D, IGHA1, and IGHV3-11, which are part of the immunoglobulin family, further highlight the enhanced immune activity in the recurrent GBM microenvironment. These alterations indicate that the tumour has adapted and modified its environment by engaging various immune mechanisms.

Recurrent tumours are frequently subjected to adaptive immune responses due to prior treatments, which can induce immunogenic cell death and expose tumour

antigens to the immune system. Genes involved in adaptive immune responses, such as those participating in somatic recombination of immune receptors, seem to be upregulated in recurrent GBM. This upregulation suggests ongoing immune monitoring and an attempt by the host immune system to recognise and combat the tumour cells. This heightened immune activity is a critical aspect of the tumour's interaction and survival against the immune system, highlighting the complex immune landscape of recurrent GBM. For instance, the upregulation of IGHV3-74 and IGHV3-48, which are involved in antigen binding, shows the increased adaptive immune response. Treatments for primary GBM, including chemotherapy and radiotherapy, could also have triggered the alteration of the immune landscape of the tumour. These treatments not only target tumour cells but also affect immune cell populations of healthy (neighbouring) cells that alter immune cell functions. The recurrence of GBM often comes with altered immune response, where genes involved in immune regulation become more pronounced. The presence of immunoglobulins and other immune-related genes suggests that the tumour and its surrounding area have adapted to create an immunosuppressive or immune-evasive environment, crucial for tumour survival and progression – marked as Hallmarks of Cancer. This therapeutic intervention impact on the immune system could be a significant factor in the gene expression profiles observed in recurrent GBM. Examples of such genes include ITGB7 and CD164L2, which play roles in immune cell adhesion and signalling.

Alongside immune-related genes, those involved in neuronal development, such as genes associated with fetal cerebrum inhibitory neurons, also show upregulation in recurrent GBM. This can be linked to the neurogenic nature of GBM and their interaction between tumour cells and neuronal elements within the brain. The dual presence of immune and neuronal development pathways highlights the complex interplay between tumour and its microenvironment, where both immune evasion and neural mimicry may facilitate tumour recurrence and aggressiveness. Genes such as NPAS4 and DLX2, which are involved in neuronal development, underscore the aspect of GBM pathology. The upregulation of MAG and PLP1, which are associated with myelin sheath formation – severely modified plasma membrane wrapped around the nerve axon in a spiral fashion (242) – and maintenance further exemplifies the neural component of recurrent GBM.

In conclusion, the dominance of immune-related genes in the primary vs recurrent GBM comparative analysis demonstrates the significant role of the immune system in tumour recurrence. The altered immune landscape, influenced by – probably – prior treatments and the evolving tumour microenvironment, is reflected in the upregulation of genes involved in immune regulation and adaptive immune response. These findings highlight the importance of considering immune agents as therapeutic targets in the battle against GBM recurrence. Additionally, the co-expression of neuronal development genes further emphasizes the complex nature of GBM genetic landscape. However, a limitation to this analysis is the lack of knowledge with regards to the time the samples were taken. The elapsed time between primary and recurrence could be anything. Therefore, we can hypothesise that if the genetic landscape of the recurrent sample is similar to its primary pair, the recurrence, and, therefore, sample extraction, must have occurred soon after primary intervention, for example patient 2018. On the other hand, if the genetic landscape is quite different between primary and recurrence samples, we can theorise that these samples were taken far apart in time from each other like patient 3016. However, this will remain a mystery.

5.3. Gene Expression Landscapes: GSC vs Tissue

Having constructed an expression landscape of both glioblastoma stem cells (GSCs) and tissues samples, their comparison was a crucial next step for advancing our understanding of GBM biology, as well as establishing suitable target candidates. GSCs are a commonly used model to study GBM due to their ability to mimic the stem-like properties and aggressive behaviour of tumour cell in vivo; therefore, are still believed to be an adequate representation of the disease. However, the relevance of GSC models to actual tumour tissue expression has been a subject of ongoing debate. By comparing the highly expressed genes in GSCs with those in tissue samples, I aimed to highlight, both, the similarities and differences between these two models, thereby evaluating the variability and limitations of using GSCs as a proxy for studying GBM. This comparison aims to provide insight into the molecular underpinnings of GBM, identifying genes and pathways that are consistently upregulated across both models. This can help in pinpointing potential therapeutic targets for miRNAs that are relevant in the context of the whole tumour or disease. Furthermore, discrepancies between GSCs and tissue sample genetic profiles may highlight unique aspects of the tumour microenvironment and the complex interactions within the tumour that are not captured by GSC models alone.

As I have established in previous chapters, the two cell lines have very different genetic landscapes; therefore, I expected differences in the number of genes shared between cell lines and tissue samples. However, somewhat unsurprisingly, only just under 50 genes were highly expressed in both tissue and cell line samples (either of the two). As cell lines are regularly used as proxy to model GBM, the expectation would have been to see more shared genes – in reality, on the contrary (Figure 5.4). Firstly, in the comparison between hypoxic GSC322 and tissue genetic landscapes, only 13 genes were mutually upregulated, for example AK4, PDK1, HK2, LDHA, NDRG1 and VEGFA (for full list of genes see Appendix Table 9). These genes are central to energy metabolism, hypoxic response and angiogenesis, underscoring their roles in the metabolic reprogramming and adaptation of GBM to hypoxic conditions (221, 222, 223). Interestingly, the presence of FRZB would suggest Wnt signalling (112) involvement, critical for cell proliferation and stemness in GBM, while BCAT1 indicates the metabolic flexibility required for survival under nutrient and oxygen deprived conditions (238). Notably, COL9A3 was identified among the shared genes – recent evidence classifies it as one of eight cell senescence-associated genes that affect prognosis (243), suggesting its potential role in GBM progression through senescence-related mechanisms. Secondly, in the comparison between GSC327 and tissue genetic landscapes, a broader set of 39 genes were mutually upregulated in tissue samples (for full list of genes see Appendix Table 10). Key genes like AK4, PDK1, LDHA, NDRG1 and VEGFA appeared again, reinforcing their important in cancer hallmark processes such as glycolysis, angiogenesis and hypoxia adaptation. In addition, CA9 (Carbonic Anhydrase 9) and HILPDA, known for their roles in cellular pH regulation (45) and lipid metabolism (44) under hypoxic conditions, were also identified, highlighting the intricate metabolic adaptations GBM cells go through to survive. Interestingly, genes such as LGR6, involved in stem cell signalling (123), and CD68, typically associated with macrophage markers (111), were present, suggesting complex interactions between tumour cells and the immune microenvironment (112). The presence of COL8A2, another collagen gene known to modulate immune responses in the tumour microenvironment (244), further emphasizes the interplay between extracellular matrix components and immune regulation in GBM. The

detection of MAOB (Monoamine Oxidase B) and AOX1 (Aldehyde Oxidase 1) is surprising; these genes are generally associated with neurotransmitter metabolism (137) and xenobiotic detoxification (14), respectively, which are not typically emphasized in GBM research, raising question about their potential roles in the tumour's metabolic landscape.

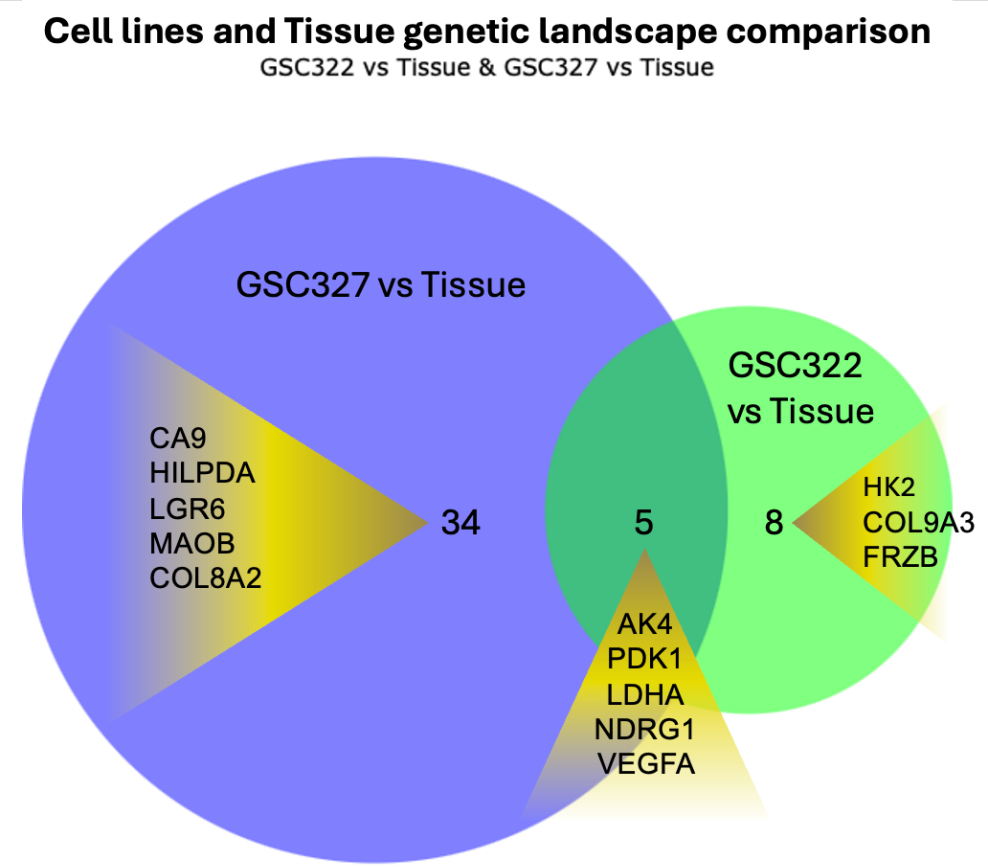


Figure 5.4: Cell line and tissue genetic landscape comparison
Venn diagram illustrating the overlap and divergence in gene expression profiles between glioblastoma stem cell lines (GSC322 and GSC327) and tumour tissue. Differential expression analysis identified 34 genes uniquely altered in GSC327 compared to tissue (blue), 8 genes uniquely altered in GSC322 versus tissue (green), and 5 genes commonly dysregulated in both cell lines relative to tissue (intersection). Highlighted genes within each sector include known hypoxia and metabolic regulators such as CA9, LDHA, and VEGFA, suggesting selective pathway activation across different cell line models. Shared genes in the overlapping region likely reflect conserved core responses to the tumour microenvironment, whereas non-overlapping genes may point to cell line-specific adaptations or limitations in recapitulating in vivo heterogeneity.

The consistent presence of these genes across both comparisons underscores common pathways in GBM pathology, particularly in metabolic reprogramming and adaptation, hypoxia response, and cell signalling. The shared involvement of glycolytic enzymes, hypoxia-induced factors, and angiogenic regulators points to the importance of these processes in maintaining the aggressive phenotype and survival of GBM cells and tumours. The inclusion of unexpected genes like MAOB and AOX1 suggests there may be underexplored areas of GBM biology, such as stress, offering new avenues for research. This comparison thus provides valuable insight into the complex gene expression of GBM stem cells and tissue samples.

5.4. Validation at the protein level

Throughout this iterative process, I picked out some genes/proteins whose presence in cell lines and/or tissue samples was validated under laboratory conditions.

Targets were selected according to the following criteria:

a) Those hypoxia gene targets that were only marginally induced in the GSC models but were highlighted as part of the overlapping gene signature sets with tissue RNAseq. These were evaluated by immunoblotting using whole tissue lysates (GBM).
b) Genes not induced hypoxia from the GSC models; however, were highly expressed when grade 4 and lower grade gliomas were compared in tissue. This class was also evaluated at the immunoblot level using whole tissue lysates (GBM).
c) Those genes that were only induced in the primary vs recurrent experiment.

Based on the tissue immunoblotting results, I would further analyse some of these targets using GBM tissue microarray (TMA) analysis if the antibodies were suitable to detect protein in IHC format – however, in the case of COL6A3, it was too large for immunoblotting; therefore, only pathology analysis through IHC staining was done. I would also carry forward some of these targets for the therapeutic miRNA target discovery part of my project.

5.4.1. Evaluation of NDRG1 in GBM tissue

Even though this was a target picked from the cell lines, upon comparison with tissue, we evaluated in tissue samples as well. Following cell line evaluation, I validated NDRG1 expression in GBM tissues by immunoblotting to determine whether the observations in the cell lines were isolated events. Patient samples are anonymized and numbered as indicated in the annotation in Figure 5.5. There are multiple samples from some patients to evaluate heterogeneity within samples. BNIP3 was included as a known hypoxia induced protein (Figure 5.5B).

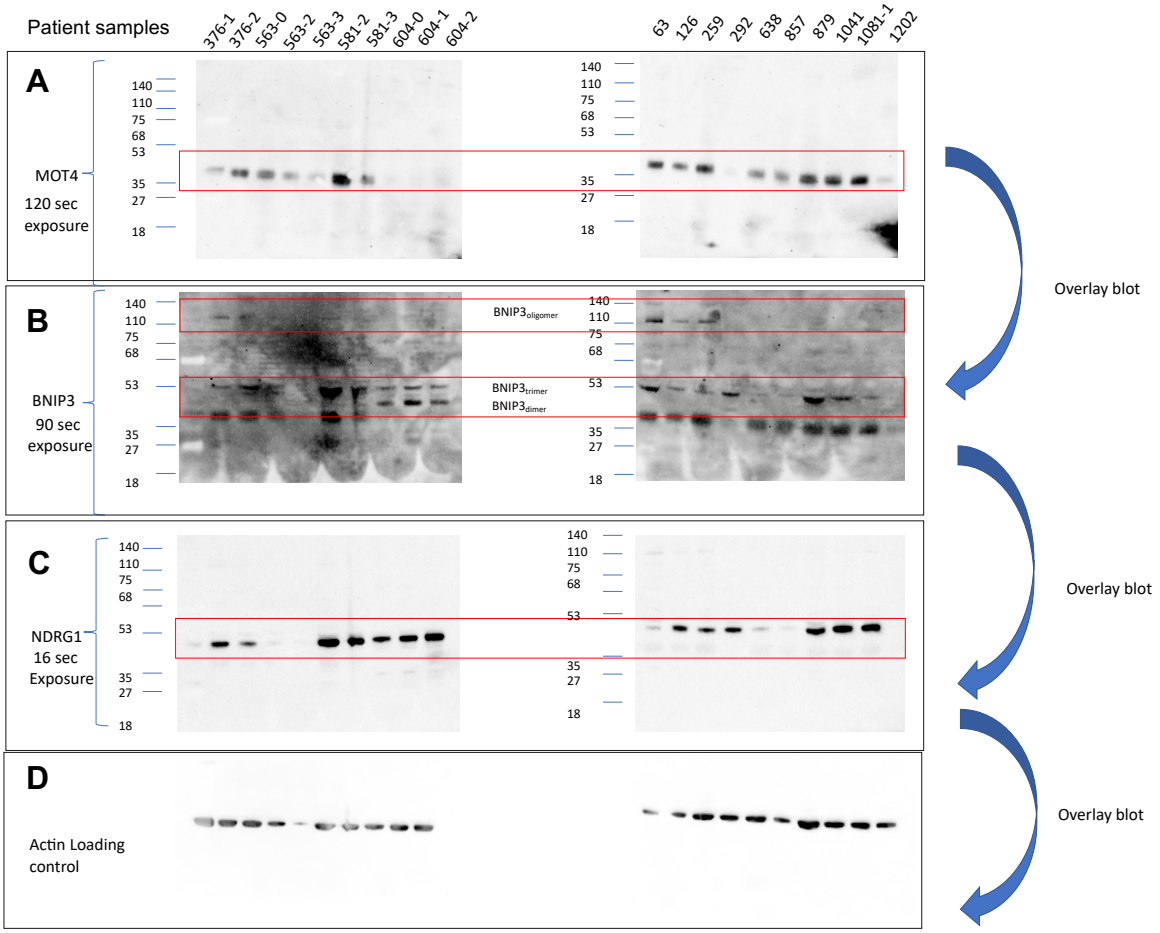


Figure 5.5: Immunoblots of four proteins in glioblastoma patients. (A) MOT4 gene with 120 sec exposure, (B) BNIP3 (hypoxia control) with 90 sec exposure, (C) NDRG1 with 16 sec exposure, and (D) actin loading control.

To initiate our analysis, we will direct our attention to samples 376-1 and 376-2. These samples are from two different parts of the same frozen tumour tissue (lanes 1 and 2 in Figure 5.5). The data shows that BNIP3 is higher and oligomeric in 376-2 vs 376-1 (Figure 5.5A) compared to the actin loading control in the actin panel (Figure 5.5D). NDRG1 was also detected in both patient samples and exhibited higher levels in 376-2 as well (Figure 5.5C lane 2 vs lane 1). The difference in NDRG1 levels is likely due to patient heterogeneity as the actin loading control expression is higher in lane 1 than in lane 2; therefore, the difference is not due to lower protein loading. Heterogeneity was also observed in patient 563: the highest level of expression of NDRG1 was observed in sample 563-0 while other samples were expressing relatively low levels of NDRG1 (Figure 5.5C, lanes 3, 4 and 5). In contrast to these two patient samples, patients 581 and 604 express relatively high levels of NDRG1 in all samples (Figure 5.5C, lanes 6-10, respectively).

The next lanes in Figure 5.5 (lanes 11-20) only contain one sample for each tumour tissue and; therefore, do no address heterogeneity within patients like the previous lanes (Figure 5.5, lanes 1-11), which had multiple samples from different areas of the same tumour. These patient samples also exhibit co-expressive patterns of NDRG1 and BNIP3 in tumour tissues. However, ratio of these two can be ‘reversed’; patient 63 present higher levels of BNIP3 than NDRG1 (lane 12, Figure 5.5B and C). In

addition, patients 126 and 1202 express higher level of NDRG1 than BNIP3 (lanes 12 and 20, Figure 5.5C and B, respectively). These data indicate three overall concepts: (i) NDRG1 is highly expressed in GBM samples; (ii) NDRG1 can exhibit heterogeneity in some patients; (iii) the ratio of NDRG1 and the hypoxic control BNIP3 can be uncoupled, indicating the cancers are biochemically different with respect to these two pathways.

5.4.2. Evaluation of PKM1 and PKM2 in GBM tissue

Similarly to NDRG1, which was part of a gene signature profile consisting of 13 genes from the 'hypoxia-induced' genes in GSCs that overlapped with grade 4 GBM tissue RNAseq (Appendix Table 9, Appendix Table 10), hexokinase and other glycolytic targets are correlating with RNA expression in grade 4 GBM tissue. As such, I also evaluated PKM1 and PKM2 expression in GBM tissue to define how this major glycolytic engine is expressed in GBM (Figure 5.6). Similarly, to the NDRG1 analysis above, patient samples were anonymized and numbered as indicated in the top of Figure 5.6.

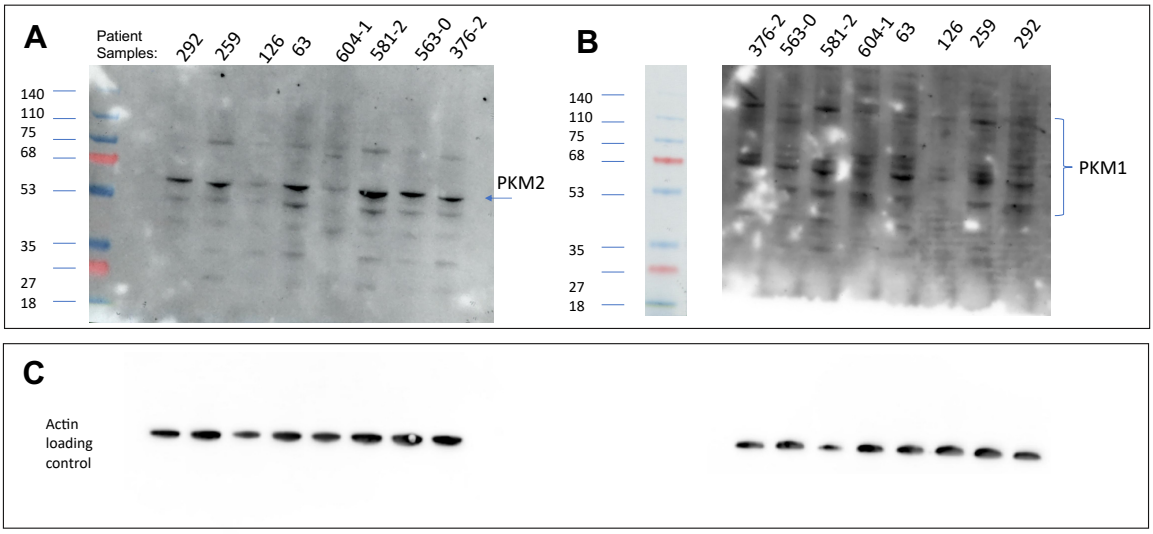


Figure 5.6: PKM1 and PKM2 Immunoblots validation in GBM tissue lysates
(A) PKM2, and molecular weight markers are visible alongside the gel, and a bracket highlights the region corresponding to the expected size of PKM proteins (~58 kDa). (B) Immunoblot showing PKM1-specific antibody. PKM1 is detected in all patient samples with varying band intensities, indicating its expression across GBM tissues. (C) Actin loading control to ensure protein input from all samples.
Although PKM2 is clearly present in GBM lysates, it was not selected for further analysis in the tissue microarray (TMA). Together, these data demonstrate that both PKM1 and PKM2 are expressed in GBM, supporting their role in tumour glycolytic metabolism.

Here, only one sample per patient was used to assess heterogeneity between patients. Firstly, samples 292 and 259 exhibited higher levels of PKM2 than patient 126 (lanes 1,2 and 3, respectively in Figure 5.6A), whereas actin loading control shows a fairly uniform expression level (Figure 5.6B). Patient 604-1 and 376-2 (lanes 5 and 8, respectively in Figure 5.6A) also exhibit a lower level of PKM2 expression than the actin loading control (Figure 5.6B), which is likely due to patient heterogeneity. Patient 581-2, in lane 6, appears to have the highest expression of PKM2 among all patient samples.

The next lanes in Figure 5.6 (lanes 9-16), PKM1 expression was assessed in the same patients as PKM2. This blot is messier than its counterpart meaning the results could be due to a number of artefacts: (i) the protein of interest might have multiple modified forms such as acetylation, ubiquitination, phosphorylation etc., (ii) different splice variants of the protein of interest, (iii) too high primary antibody concentration, (iv) non-purified antibody, (v) target protein forming multimers (245, 246). The white dots appearing here are possibly air bubbles that are trapped against the membrane during transfer. The most intense bands stay around 60-68 kDa, which is the expected molecular weight of PKM1. The other bands could be because of possible ubiquitination. All, except patient 126, exhibit high levels of PKM1. Interestingly, patient 126 also exhibited low levels of PKM2.

The overall conclusion of these data: (i) both PKM1 and PKM2 can exhibit heterogeneity between patient samples, (ii) patient 126 expresses lower levels of PKM1 and PKM2 compared to actin loading control, (iii) PKM1 could be exhibiting multiple modified forms such as acetylation, ubiquitination, phosphorylation, etc.

5.4.3. Evaluation of P4HA1 in GBM tissue

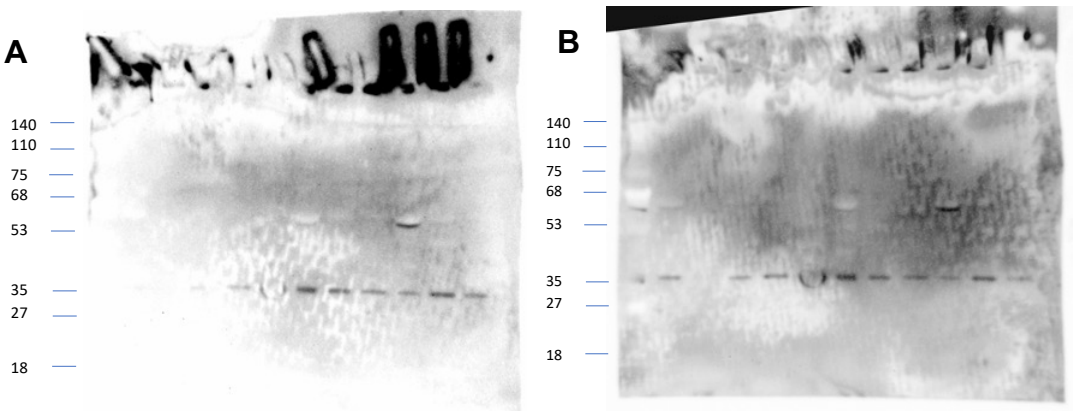


Figure 5.7: Western blot validation of P4HA1
A) First western blot, B) second overlay

Like EGLN3, P4HA1 was differentially overexpressed in both GSC322 and GSC327 models (Figure 4.4) in contrast to the tissue analysis, where it was not differentially expressed (Figure 5.2). So, immunoblotting technique was used to further examine its expression. There are some bubbles present on the blot, which was because the gel was stuck to the membrane; however, the initial figure (Figure 5.7A) was improved with a second overlay overnight (Figure 5.7B). In lane 3, there's no expression of PH4A1, like in all other lanes at 35 kDa. Though there seems to be some additional protein detection at around 68 kDa, which could be other forms of the protein of interest.

In conclusion, the differential gene expression analysis comparing low-grade and high-grade gliomas has provided valuable insights into the molecular changes associated with glioblastoma progression. By employing DESeq2 and incorporating methods such as Independent Hypothesis Weighting (IHW) for p-value adjustment and tailored filtering, the analysis ensured the robustness and reliability of the results. The identification of significantly differentially expression genes, particularly within well-known glioblastoma-associated pathways, highlights the complex genetic landscape

of the disease. Moreover, the unexpected discovery of the HOX gene family among the differentially expressed genes – along with collagen (COL) genes as tissue-specific outliers – suggests potential new avenues for exploration in glioma research. While genes like NDRG1, PKM, and P4HA1 emerged as consistent players across models, the tissue-specific prominence of HOX (more details in the next chapter) and COL genes raises intriguing questions about therapeutic prioritization. Targeting these outliers with miRNA-based approaches could offer higher specificity, as their dysregulation appears more uniquely tied to the native tumour ecosystem. However, the evolutionary conservation of HOX genes and the structural role of collagens may pose delivery challenges or off-target effects that would require careful preclinical evaluation.

This chapter has also provided a comprehensive analysis of gene expression profiles comparing glioma tissue and glioblastoma stem cells (GSCs). The differential genes expression analysis of primary versus recurrent glioma samples revealed significant alterations, particularly the upregulation of immune-related genes in recurrent tumours, which suggest an evolved immune landscape. These findings highlight complex interplay between tumour cells and the immune microenvironment, emphasizing the potential of immune modulation as a therapeutic strategy for recurrent GBM. Furthermore, the upregulation of genes related to neuronal development underlines the neurogenic nature of GBM and its capacity for tumour cell-neuron interaction. This interplay between immune evasion and neural mimicry may facilitate tumour recurrence and aggressiveness.

The comparison of GSCs and tissue samples underscored the limitations of GSC models, with significant differences in the gene expression landscapes between the two. This highlighted the need for careful consideration when extrapolating findings from GSCs to actual tumour biology, as key pathways, such as metabolic reprogramming and angiogenesis, were inconsistently captured across both models. However, shared genes between these models still pointed to crucial aspects of GBM biology, including glycolysis, hypoxia adaptation, and angiogenesis, providing valuable insights for therapeutic targeting.

Finally, the validation of key genes like NDRG1, PKM1, PKM2 and P4HA1 at the protein level through immunoblotting and tissue analysis further confirmed the heterogeneous nature of GBM and the variability in gene expression across patients. While these proteins represent viable miRNA targets due to their consistent detection, the tissue-specific outliers (COL, HOX genes) may offer more selective therapeutic windows, given their stronger association with recurrence and grade progression. Future work could explore whether targeting these outliers with miRNA combinations—while accounting for their potential roles in normal physiology—might strike a balance between efficacy and safety.

These findings collectively refine our understanding of glioma progression and underscore the importance of model-aware therapeutic design, setting the stage for investigations into miRNA-mediated regulation of both conserved and niche pathways.

CHAPTER 6. HOX genes in the Context of Glioma Tumorigenesis

While evaluating the results of the differential gene expression analysis of glioma progression, I observed a striking pattern: the homeobox gene family emerged as prominently upregulated in tumour tissues. Notably, these genes—unlike other candidates identified in our study—were not detected as expression outliers in glioblastoma stem cell (GSC) models, suggesting their dysregulation may be uniquely dependent on the native tissue microenvironment. This makes them particularly compelling candidates for understanding glioma-specific biology that cannot be fully recapitulated *in vitro*.

HOX genes, a specialized subset of homeobox genes, encode transcription factors critical for regulating embryonic development. Highly conserved across species, they orchestrate body patterning and positional identity along the anterior-posterior axis through precise spatiotemporal activation. The human genome contains 39 HOX genes organized into four clusters (HOXA, HOXB, HOXC, and HOXD) on distinct chromosomes, with each cluster's sequential expression during development governing the formation of specific anatomical structures. While homeobox-containing transcription factor complexes exist more broadly, this chapter focuses exclusively on the canonical HOX gene family and their emerging role in glioma pathobiology.

In cancer, HOX genes play a complex dual role, acting as both oncogenes and tumour suppressors depending on the specific gene and cancer type (247). When deregulated, they contribute to cancer progression by promoting cell proliferation, invasion, metastasis, and angiogenesis, and can also lead to resistance to therapies. This dysregulation is a hallmark of many malignancies. For instance, chromosomal translocation involving HOX genes can lead to their constitutive expression in certain leukemias, where specific HOX genes correlated with survival outcomes (248). Similarly, altered HOX gene expression impacting the balance between cell proliferation and differentiation has been identified in lung cancers (249). Pathways regulated by HOX genes are important for controlling a wide array of developmental processes and pathways. Their primary function is to encode transcription factors that bind to specific DNA sequences, thereby regulating the expression of downstream target genes (250). These target genes are involved in several critical biological processes:

- Cell proliferation and survival: HOX genes can influence the cell cycle and cell proliferation, which is essential for tissue growth and regeneration (251). When deregulated in cancer, they can promote uncontrolled cell growth and survival.
- Cell differentiation: They play a role in the differentiation of cells into various specialized types, ensuring the correct formation of tissues and organs (250).
- Apoptosis: HOX genes can also regulate programmed cell death, which is crucial for removing unnecessary or damaged cells during development (34).
- Migration, adhesion and invasion and metastasis: These genes are involved in the regulation of cell migration and adhesion, processes that are vital for the proper positioning of cells during embryogenesis (252). Abnormal HOX gene expression can drive cancer cells to become more invasive and metastatic, a process that involves mechanisms like the epithelial-mesenchymal transition (EMT).

- Angiogenesis: Some HOX genes can induce the formation of new blood vessels to support tumour growth by upregulating secreted angiogenic factors (253).

Furthermore, deregulated HOX gene expression is strongly linked to cancer stem cells (CSCs), which are responsible for initiating cancer, causing recurrence, and spreading to other parts of the body (254). In glioblastoma, specifically, high expression of HOXA9 and HOXA10 is associated with poor survival, and HOXD9 is essential for the proliferation and survival of glioblastoma cells (255).

As I, hopefully, successfully reiterated, the presence of HOX genes is more prominent or 'normal' during the developmental stages of an organism. Then why am I talking about these genes? Their reactivation in glioma represents a hijacking of developmental programs to drive tumorigenesis.

6.1. In silico analysis

On Figure 6.1, the same volcano plot is shown as Figure 5.2; however, here, all the HOX genes are labelled. Interestingly the majority of the subfamily is upregulated across 117 patients.

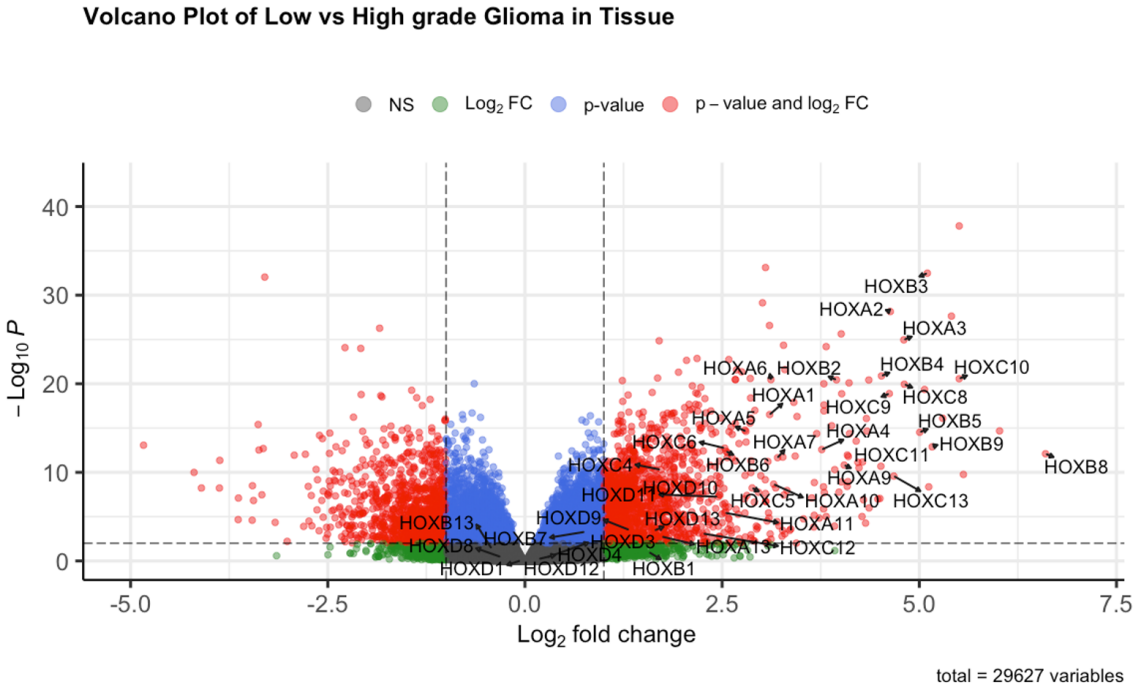


Figure 6.1: Volcano Plot of low vs high grade glioma tissue with HOX genes
Each expressed gene found in glioma tumours were plotted according to their differential expression levels. On x-axis, the expression intensity is shown in the form of Log2 fold change. On the y-axis, the negative log(10) p-value indicates the level of confidence in the fold change for each gene. Genes located on the left side of the plot (genes with negative fold change), are downregulated in this comparison. Conversely, genes located on the right-hand side (positive fold change) are upregulated. The dotted lines indicate filtering cut-off values, vertically, the fold change cut off (FCcutoff=1.0), and horizontally the p-value cut off (pCutoff=0.01). Genes that failed to meet these cut off values were colored green (p-value only), blue (fold change only) and grey (both). However, genes in red have a relatively significant fold change as well as a high confidence level of that result. Labels are only for the HOX gene family members present in the analysis.

More specifically only 6 genes out of the 39 were either not upregulated or the results obtained were not statistically significant (ns) (Appendix Figure 16, Appendix Figure 17). All other genes were clearly upregulated. The only downregulated genes (though not statistically significant) were HOXB13 and HOXD1, and, therefore, anticorrelated with all other HOX gene activation (Figure 6.2). Some interesting patterns were observed in the correlation matrix of HOX genes, showing both positive and negative relationships among these genes. Notably, strong positive correlations were observed within clusters – HOXB, HOXC, and HOXD (Figure 6.2). For example, genes within the HOXB group, such as HOXB3, HOXB4, HOXB5, HOXB6, HOXB7, HOXB8, and HOXB9, exhibit high level of co-expression, suggesting that these genes are functionally linked and may collectively contribute to similar regulatory processes in glioma. A similar pattern is seen within the HOXC cluster, particularly among HOXC4, HOXC5, HOXC6, and HOXC8, indicating their potential involvement in shared biological pathways. The HOXD cluster, including genes like HOXD9, HOXD10, HOXD11, HOXD12, and HOXD13, also shows strong positive correlations, which would reflect their coordinated role in developmental pathways that are seemingly reactivated during glioma progression.

Conversely, several HOX genes, such as HOXD4, HOXD8, HOXA1, HOXA10 display strong negative correlations with other HOX genes. These antagonistic relationships suggest that these genes may act in opposition to the more highly correlated clusters, possibly playing a role in balancing gene expression during tumorigenesis. The heatmap also reveals distinct block of co-expressed genes within the HOXB and HOXC clusters, which may be co-regulated by the same factors, further emphasizing their functional connectivity.

Additionally, certain genes like HOXA13 and HOXD13 exhibit mixed correlation patterns, correlating positively with some genes and negatively with others. Thes suggests that these genes may have more complex regulatory roles or be involved in multiple, distinct pathways within the tumorigenic process. HOXD1, which shows lower correlations with many other HOX genes, may have a more specialized or context-dependent role.

In the context of glioma, the clustered gene expression observed in the HOXB, HOXC, and HOXD groups indicates that these genes are seemingly reactivated during glioma progression, possibly driving tumour growth and maintenance through the reactivation of developmental pathways. The negative correlations observed, particularly with genes like HOXD4 and HOXA1, may reflect a balance between pro-tumorigenic and tumour-suppressive functions, potentially influencing tumour heterogeneity and treatment response. Overall, the heatmap suggests a complex regulatory network among HOX genes in glioma, with both co-regulated clusters and unique gene-specific functions playing separate roles in the disease.

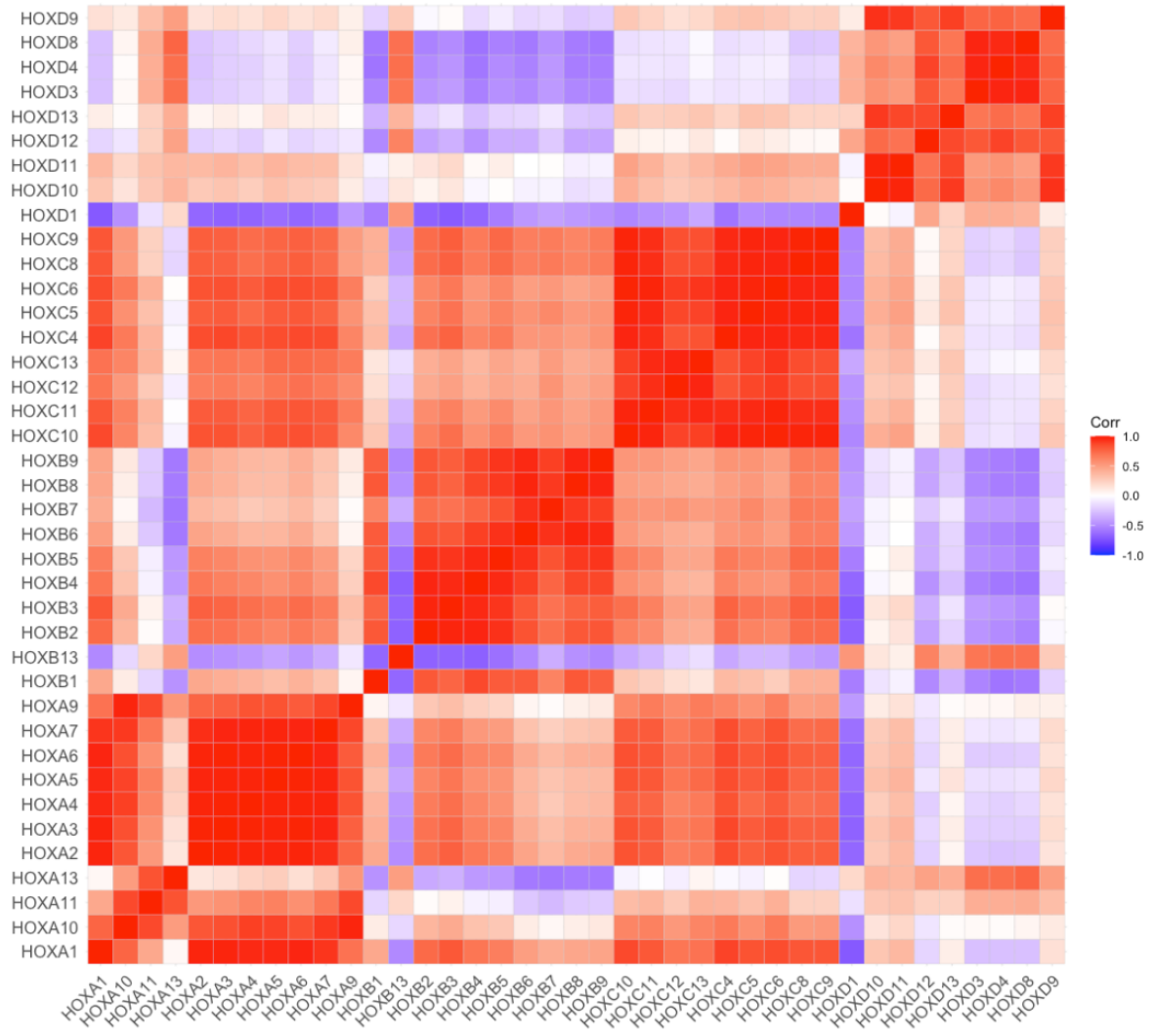


Figure 6.2: Correlation matrix heatmap of HOX gene expression in glioma tissue samples. This heatmap shows the pairwise correlation coefficients between the expression levels of all HOX genes in glioma tissue samples. The colour scale ranges from -1 (strong negative correlation, in blue) to +1 (strong positive correlation, in red), with 0 indicating no correlation (white). The clustered patterns observed within HOXB, HOXC and HOXD gene groups highlight strong positive correlations, suggesting co-regulation and potential involvement in similar biological pathways in glioma. It also corroborates their sequential activation process. In contrast, several HOX genes, such as HOXD4, HOXD8, HOXA1, and HOXA10, show negative correlation with other genes, indication opposing regulatory roles.

To further examine these genes, I wanted to see their expression across the patient cohort. The first heatmap is organised by patient ID, while the second heatmap uses hierarchical clustering to order the data, which helps to reveal patterns in gene expression across the patient cohort. A key observation from these heatmaps is the variability in the expression of specific HOX genes – heterogeneity is one of the main characteristics of glioma – particularly HOXB3 and HOXC10, which were further validated under laboratory conditions. These genes show distinct expression profiles between low- and high-grade gliomas, which highlights their key involvement in tumour progression and aggressiveness. The second heatmap, which is clustered based on the similarity of gene expression profiles, reveals distinct clusters of patients, indicating that specific subsets of HOX genes are consistently upregulated or downregulated in certain patient groups. The hierarchical clustering brings out patterns that are less obvious in the patient ID-ordered heatmap. For instance, genes like HOXA1, HOXA9,

HOXA10, HOXB3, HOXC6, and HOXD8 are seen to cluster together, indicating they may be co-regulated or participate in similar pathways that drive the aggressiveness of high-grade gliomas. The co-regulation is significant because it highlights the potential involvement of these HOX genes in pathways related to cell proliferation, differentiation and tumorigenesis.

HOXB3 consistently shows upregulation in high-grade gliomas across a significant number of patients, as indicated by the intense red regions in both heatmaps. This suggests that HOXB3 may play a role in promoting the more aggressive phenotype seen in high grade glioma tumours.

HOXC10, on the other hand, shows a more variable expression pattern. In some patients, HOXC10 is strongly upregulated in high grade tumours, while in others, it shows little to no change or even downregulation. The variability might suggest that HOXC10's role in glioma progression could be more context-dependent, possibly influenced by specific tumour microenvironments or genetic backgrounds. The HOX gene families' sequential activation may also play a role in this gene's low or even inconsistent activation as it is further 'down the line' than HOXB3. This highly context-dependant property could, therefore, make it a good biomarker for tumour progression. The hierarchical clustering indicated that HOXC10 is part of a distinct gene cluster that may be involved in differing pathways from those regulated by HOXB3, potentially pointing to a role in other aspects of tumour biology such as cell survival or response to hypoxia.

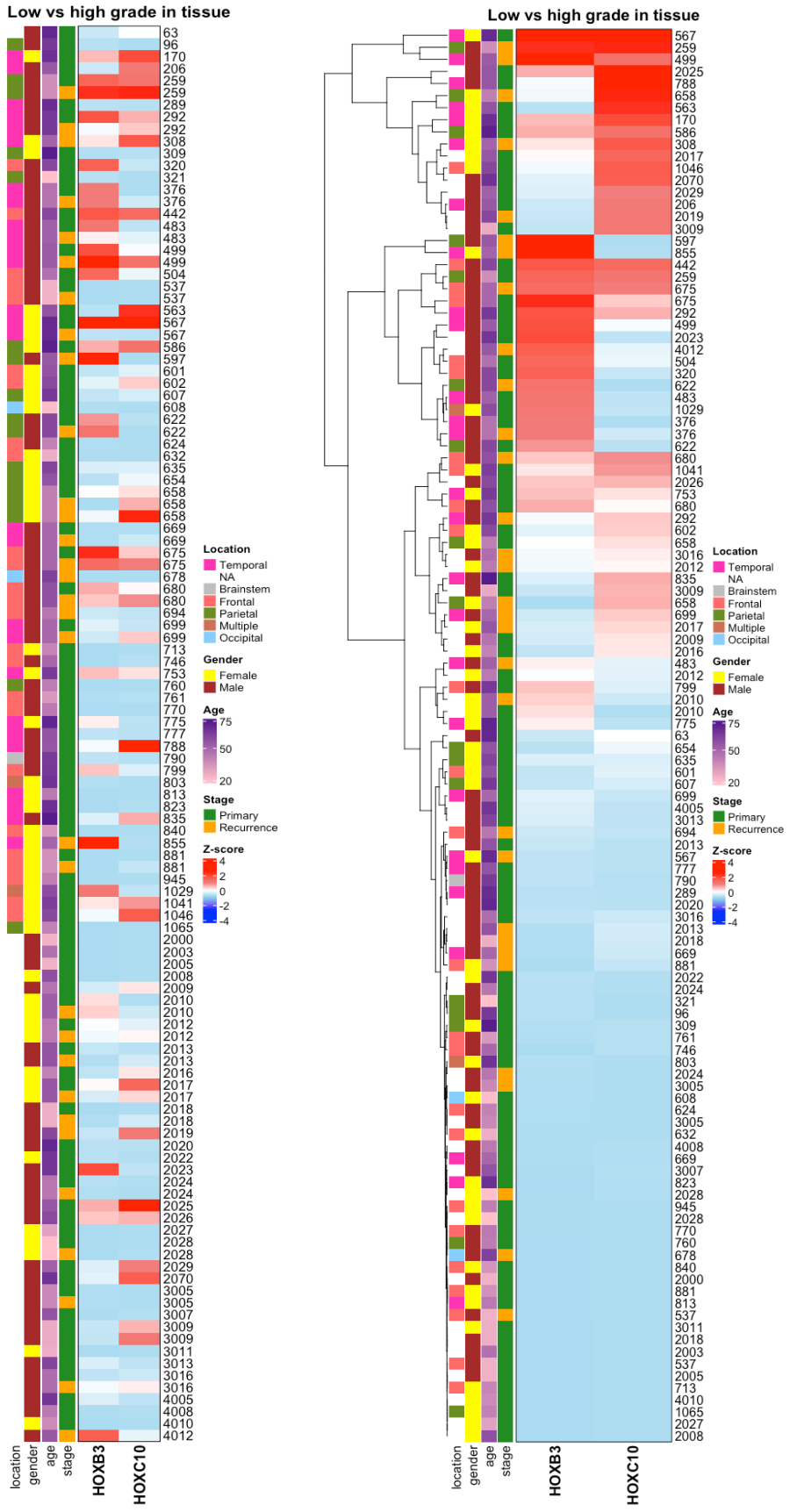


Figure 6.3: Expression Patterns of HOXB3 and HOXC10 in Low vs. High-Grade Gliomas.
The heatmaps depict the expression profiles of HOXB3 and HOXC10 across different glioma patients comparing low- and high-grade glioma tissues. A) The heatmap is organized by hierarchical clustering to visualize patterns of co-expression between the two genes (HOXB3 and HOXC10) across different patient samples. The clustering highlights which samples exhibit similar gene expression profiles based on the z-score scaling of expression data. Strong expression (in red) or reduced expression (in blue) can be compared across various patient characteristics, including tumour location, gender, age, and recurrence status, as displayed by the annotation bars on the left. Age is represented as a gradient from purple (older) to white (younger), while recurrence status is marked in green (primary) or orange (recurrent) for easier interpretation. The dendrogram on the left shows how similar or divergent expression patterns are within patient groups, indicating which patients cluster together based on HOX gene expression. B) The second heatmap presents the same data, but the patients are organized by their IDs, making it easier to observe patterns directly related to individual patients and tumour progression (primary vs. recurrence). The color annotations remain the same, aiding in identifying how expression levels of HOXB3 and HOXC10 change from the primary to recurrent glioma stages in specific patients. The visualization is structured to allow observation of patient-specific differences or consistencies in gene expression between the low- and high-grade stages of glioma.

Correlation with clinical data

In inclusion of clinical data, such as tumour location, patient age, gender and whether the tumour was primary or recurrent, allowed for a more nuanced understanding of how HOXB3 and HOXC10 expression correlated with glioma characteristics. For instance, the expression of HOXB3 is consistently high in tumours located in more aggressive sites, like the frontal and temporal lobes, which are commonly associated with poorer patient outcomes (Figure 6.3) (256, 257). Similarly, HOXC10's variable expression might correlate with specific clinical features, such as tumour recurrence or particular age groups; however, lack of additional data prevented further investigation of this.

The detection of HOX genes upregulation during glioma tumorigenesis is particularly interesting and could be significant for several reasons. Firstly, HOX genes are known to regulate cancer stem cells (123), which are pivotal in driving tumour initiation, progression and recurrence. Glioblastoma stem cells (GSCs) utilize the dysregulation of HOX gene expression to maintain their stemness and self-renewal properties. This aberrant activation of pathways can lead to the proliferation and survival of these cancerous cells. However, these genes were not present in the GSCs analysis in CHAPTER 4. Begs the questions, why? Possibly, the need for survival under hypoxic conditions, which used to mimic the tumour microenvironment, was triggering different pathways first and foremost before focusing on development and growth. In addition, HOX genes are part of a very intricate pathway that are likely to be triggered by many extracellular signals in the context of the tissue/organ, which is why they were not part of the gene signature.

Secondly, HOX genes influence the invasive and metastatic behaviour of cancer cells. Their role in regulating cell migration and adhesion is co-opted by tumour cells to enhance their ability to invade surrounding tissues and spread to distant sites. This capability is particularly detrimental in gliomas, which are known for their aggressive invasion into surrounding brain tissue.

Thirdly, HOX genes can contribute to resistance against anti-cancer drugs, including anti-angiogenic treatments. The involvement of HOX genes in apoptosis and cell survival pathways, a hallmark of cancer, can contribute to the resistance of glioma cells to conventional therapies. By upregulating anti-apoptotic signals and promoting cell survival, HOX genes may help glioma cells evade the cytotoxic effects of treatment, leading to therapy resistance and tumour recurrence (which some patients in this cohort experienced). This resistance could cause a significant hurdle in the effective treatment of glioblastoma. This is another unknown information of this cohort, and, therefore, limitation of the interpretability of results – did they receive any treatment after primary diagnosis, if yes, what?

Finally, the role of HOX genes in neural development makes their dysregulation in gliomas particularly relevant. Gliomas, especially glioblastomas, often hijack development pathways to sustain their growth and malignancy. The reactivation of HOX gene expression in gliomas suggests that these tumours may exploit neurodevelopmental processes to support their pathological behaviour (251, 252, 258).

In conclusion, the identification of upregulated HOX genes in glioma tumorigenesis highlights their potential role as key drivers of this aggressive cancer. In addition, it also highlights key genetic nuances that might be missed while using GSCs to represent the more complex tissue processes in research. Their involvement in critical pathways regulating cell proliferation, differentiation, migration and survival underscore their importance in the malignant transformation and progression of gliomas. Understanding the specific mechanisms by which HOX genes contribute to glioma biology could open new avenues for targeted therapies – especially for my downstream miRNA therapy analysis. From a therapeutic perspective, the dual role of HOX genes makes them complex targets. However, some research suggests that targeting specific HOX genes could be a promising strategy. For example, inhibitors like HXR9 have shown potent activity against glioblastoma stem cells (259). Another therapeutic peptide, HTL-001, which targets HOX gene over-expression in GBM, has shown promise in laboratory studies and is being prepared for clinical trials (260). Preclinical studies also indicate that combining HOX inhibitors with standard treatments, such as radiation therapy, could be a more effective approach, with combinations like HXR9 and radiation showing additive cytotoxic effects *in vitro* and improved tumour growth delay in mouse models (261). However, first, we attempted to validate the presence of (a few of) these genes under laboratory conditions.

6.2. Validation

Under laboratory conditions, the validation of HOXB3 and HOXC10 expression in patient-derived samples corroborated the patterns observed in the heatmaps. The consistent upregulation of HOXB3 in high-grade glioma tumours, alongside its clustering with other proliferative HOX genes, strongly indicates its potential as a reliable biomarker or therapeutic target. In contrast, HOXC10 exhibited highly variable expression across patient samples, emphasizing, the complexity of its role in glioma biology. This variability suggests that personalised strategies may be essential for effective targeting of HOXC10, as its function and relevance likely differ significantly between individuals.

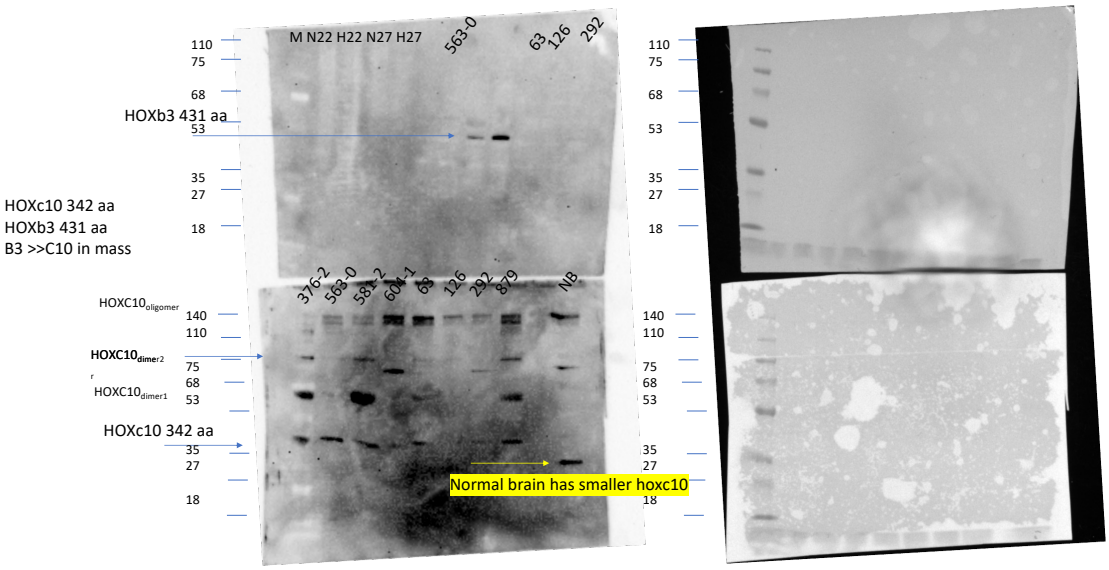


Figure 6.4: Western blot of HOXC10 and HOXB3 expression in patient-derived GBM and normal brain samples
Western blot images show differential protein expression patterns if HOXC10 and HOXB3. The left panel highlights HOXC10 expression across GBM samples, with notable smaller molecular weight band observed in normal brain tissue (yellow annotation), suggesting the presence of an alternative isoform or post-translational modification. In contrast, GBM samples predominantly express a larger HOXC10 band, indicating potential tumour-specific expression of the full-length protein. The right panel includes additional blot of loading controls. Given the full-length HOXC10 is 348 amino acids (~39 kDa) and HOXB3 is 429 amino acids (~45 kDa), the observed size differences support the transcriptomic variability of HOXC10, reinforcing its context-dependent role in GBM.

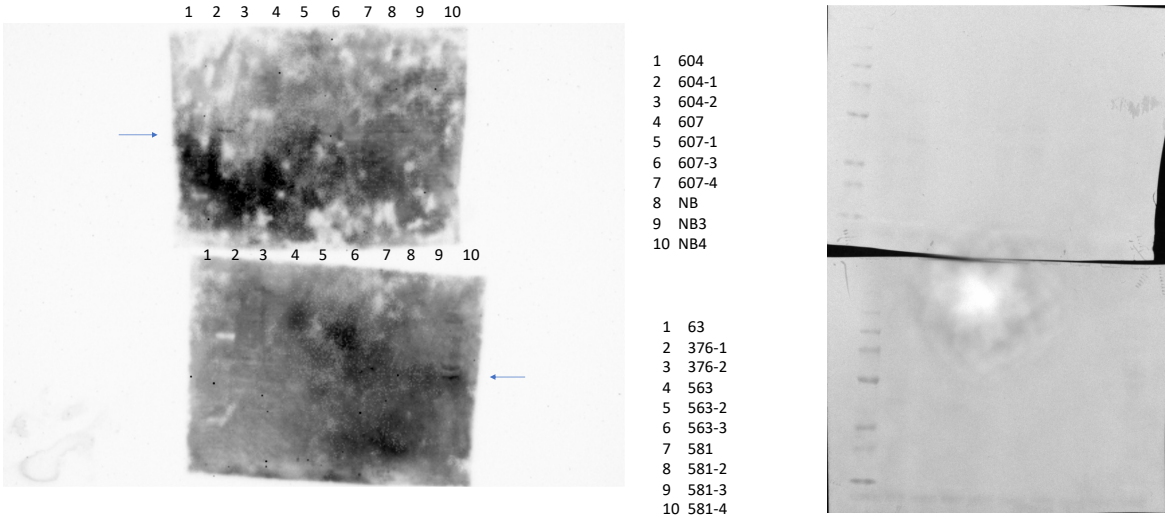


Figure 6.5: Western blot of HOXB3 expression in different regions of the same patient-derived GBM samples Lanes 1-10 correspond to different tumour regions of patients (lanes 1-3, and 4-7, respectively for the top left blot and lanes 1, 2-3, 4-6, 7-10, respectively for the bottom left blot), demonstrating heterogeneous expression of HOXB3 within a single patient's tumour. The observed variation in band intensity and molecular weight (indicated by arrows) suggests the presence of isoform diversity or post-translational modifications. The right panel represents loading controls.

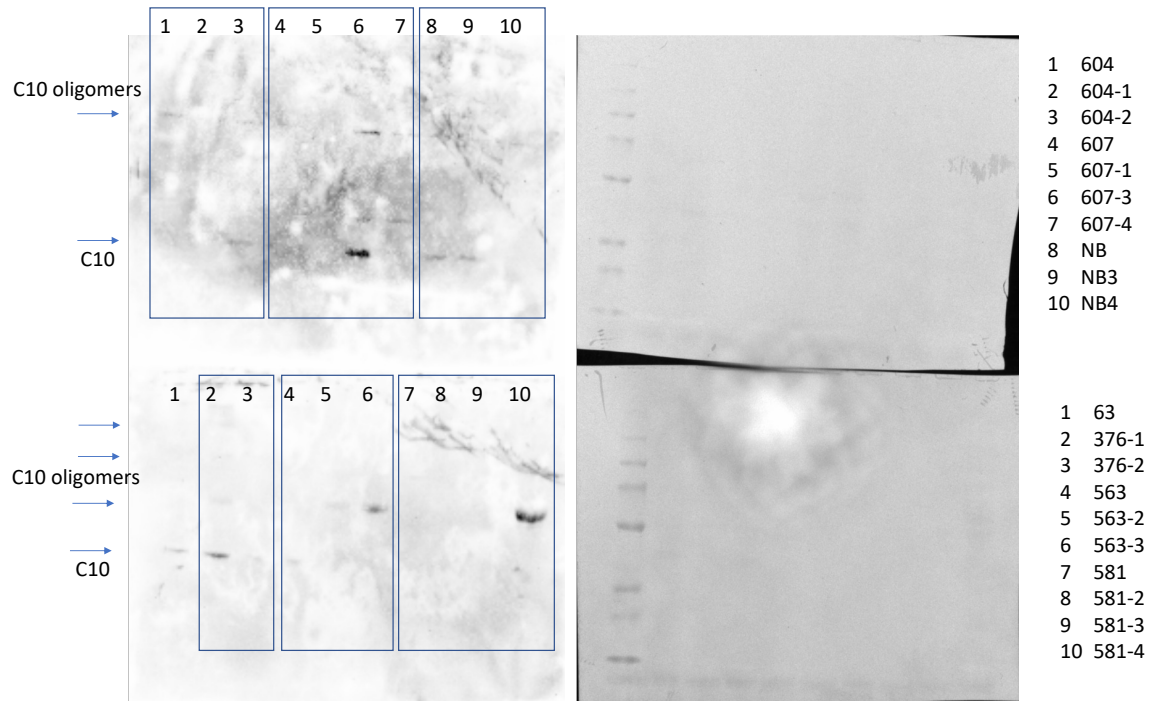


Figure 6.6: Western blot of HOXC10 expression in different regions of the same patient-derived GBM samples
Lanes 1-10 correspond to different tumour regions of patients (lanes 1-3, and 4-7, respectively for the top left blot and lanes 1, 2-3, 4-6, 7-10, respectively for the bottom left blot – blue boxes), demonstrating heterogeneous expression of HOXC10 within a single patient's tumour. Arrows mark the multiple bands at different molecular weights, reinforcing the context-dependent role of HOXC10 in GBMs. The right panel contains loading controls.

The results presented in this chapter show that HOX genes play a significant role in glioma tumorigenesis and reveal some interesting patterns in how these genes behave in glioma tissues. The analysis clearly demonstrated that the majority of the HOX gene family is upregulated in glioma patients, with strong co-expression patterns seen within the HOXB, HOXC, and HOXD subfamily. These patterns suggest that groups of HOX genes may work together to drive tumour progression by reactivating developmental pathways. At the same time, some HOX genes, like HOXD4 and HOXA1, showed negative correlations with others, pointing to potential balancing or opposing roles.

The clinical relevance of these findings was supported by correlations between HOX gene expression and features like tumour location and aggressiveness. For example, HOXB3 was consistently highly expressed in more aggressive tumours and showed a clear link to high-grade gliomas and tumour progression, making it a strong candidate for further research. HOXC10, on the other hand, showed more variable expression across patients, which suggests its role might depend on specific environments or patient genetics. This variability could make it a useful biomarker for tumour progression in certain cases but also means that targeting HOXC10 would likely need to be tailored to individual patients, therefore an ideal candidate for further research specific to the idea of precision medicine.

Validation experiments in the lab confirmed the expression patterns of HOXB3, and HOXC10 in patient-derived samples, supporting the computational findings. HOXB3's consistent high expression strengthens its potential as a target for therapeutic

approaches, while HOXC10's variability highlights the need for more detailed studies to understand its role in glioma biology.

These findings also tie in with the next step of this research, further analysis of patient tissue samples through imaging and the work on miRNA-based therapies, introduced in the introductory chapters. Since HOX genes are highly active in glioma, they could be strong targets for plant-derived miRNAs, which may help disrupt the pathways these genes are involved in. However, this study also highlights some challenges, like the complexity of tumour biology and the gaps in clinical data that could have provided more context for the results.

While genes like NDRG1, PKM, and P4HA1 emerged as consistent players across models, the tissue-specific prominence of HOX and COL genes raises intriguing questions about therapeutic prioritization. NDRG1, a hypoxia-responsive mediator of invasion (233), PKM, with numerous isoforms as metabolic switches (262), and P4HA1, a collagen modifier driving ECM stiffness (108), would all be reliable medicinal miRNA target candidates due to their ubiquitous nature; however, potentially less selective. For example, NDRG1 inhibition might disrupt normal iron metabolism (263). In contrast, the tissue-specific dysregulation of HOX genes (eg. HOXC10's link to GBM progression in my data) and COL6A3 (linked to recurrence in my data) suggests niche roles in tumour adaptation. Targeting these outliers could exploit vulnerabilities unique to the tumour ecosystem – for instance, HOX genes regulate positional identity in development (250), and their aberrant re-expression in GBM might represent a "developmental Achilles' heel". However, their evolutionary conservation raises delivery hurdles: HOX targeting miRNAs could risk cross-reacting with normal neural stem cell pools (264), while COL6A3 suppression could compromise basal lamina integrity in peripheral tissues (265). Theoretically, a tiered approach might balance these risks: (1) prioritize HOX/COL miRNAs for localized delivery (eg. convection-enhanced delivery to minimize off-target effects (266)), (2) multiple miRNA-carrying pills targeting HOX paralogs in GBM to reduce escape variants, and (3) pair NDRG1 or PKM suppression with HOX/COL inhibition for broader pathway coverage. Such strategies would require extensive *in vivo* testing. The trade-off between target ubiquity (NDRG1/PKM) and specificity (HOX/COL) thus hinges on delivery precision and combinatorial design, highlighting the need for extremely high-level miRNA engineering and its delivery to treat GBM.

This chapter establishes HOX genes as biologically significant players in glioma progression, providing a foundation for future investigation. Their absence as expression outliers in GSC models raises important questions about the limitations of current in vitro systems for studying these genes. To address this, we could develop isogenic cell panels with engineered HOX gene expression to systematically evaluate their functional impact on GSC states. An alternative possibility is that the HOX gene signature observed in tumours originates from infiltrating immune or endothelial cells rather than tumour cells themselves—a distinction that would require single-cell resolution studies to resolve.

From a therapeutic perspective, targeting HOX genes with miRNA-based approaches shows promise, but our data suggest such strategies may need to be tested directly in *ex vivo* GBM tissue slices rather than conventional cell line models. Whether through miRNAs or other modalities, elucidating the mechanistic roles of HOX genes in glioma could reveal novel therapeutic opportunities for this devastating disease.

CHAPTER 7. Pathology of Tissue derived RNA targets

The pathological assessment of tissue-derived RNA targets requires careful examination of structural and molecular differences between normoxic and hypoxic tissues. Microscopically, tissues under normoxic physiological conditions (ie. normal levels of oxygen) exhibit well-organised cellular architecture, with preserved matrix integrity. In contrast, hypoxic tissues display distinct adaptive pathological features, including nuclear pleomorphism, cytoplasmic vacuolization, increased mitotic activity, and, in extreme cases, necrotic cores surrounded by viable stressed cells (115). Hypoxic regions are further characterized by aberrant angiogenesis marked by irregular, leaky microvessels and disorganised endothelial proliferation – a compensatory response to oxygen deprivation. Additionally, extracellular matrix remodelling is also evident with increased deposition of fibrillar collagen and elevated expression of hypoxia-responsive factors such as HIF-1a (267).

A diagnostically hallmark feature of hypoxic tissue is the presence of perinecrotic pseudopalisading cells – elongated, densely packed tumour or stromal cells surrounding necrotic foci (85, 115). This phenomenon is particularly prominent in high-grade gliomas, where hypoxia-driven metabolic reprogramming leads to a more aggressive phenotype. In comparison, normal tissues maintain a balanced metabolic state, lacking these morphological alterations. Recognizing these pathological distinctions is critical for identifying the effects of de-novo identified RNA targets, as hypoxia-induced transcriptional changes suggest disease progression and reveal potential therapeutic vulnerabilities. Hypoxia dramatically reshapes glioblastoma's pathology, but the functional roles of our RNA targets—NDRG1, HOXC10, and COL6A3—remain unclear. This chapter uses immunohistochemistry to map their protein expression patterns across tumour tissue, specifically probing whether they localize to hypoxic niches (like pseudopalisading regions or necrotic borders). By linking their spatial distribution to hypoxia-driven pathology, we'll assess whether these molecules contribute to tumour aggressiveness—and whether they represent viable therapeutic targets.

7.1. NDRG1 Validation

N-myc downstream-regulated gene 1 (NDRG1) is a hypoxia-responsive protein implicated in cellular stress responses, differentiation, and tumour progression (231). As a well-established hypoxia-inducible factor, its expression is tightly regulated by HIF-1α under low oxygen conditions, making it a reliable candidate molecular marker of hypoxic tumour regions. In glioblastoma, NDRG1 promotes tumour cell survival during metabolic stress and has been associated with therapy resistance - suggesting its spatial localization within hypoxic niches, such as perinecrotic pseudopalisading cells, may reveal critical vulnerabilities in tumour aggressiveness and progression. In normal brain tissues, NDRG1 predominantly expressed in the cytoplasm of epithelial and mesenchymal cells, with sparse nuclear localization The staining pattern in normoxic tissues is diffuse and weak, reflecting its basal role in cellular homeostasis. However, in hypoxic tissues, NDRG1 expression undergoes a significant upregulation and staining becomes more intense, particularly in perinecrotic regions, where cells exhibit strong cytoplasmic accumulation and occasional perinuclear localization (268). This upregulation aligns with hypoxia-driven activation of HIF-1a and other stress-associated pathways – including the unfolded protein response (UPR) via PERK/ATF4, mTOR-mediated metabolic reprogramming, and NF-κB inflammatory

signalling – reinforcing NDRG1's role in mediating cellular adaptive survival mechanisms.

The intensity and localization of NDRG1 staining may also vary depending on tumour grade and histological subtype. In highly aggressive tumours, aberrant nuclear localization has been reported, potentially linked to altered subcellular trafficking mechanisms under metabolic stress. Moreover, in certain cancers, NDRG1 expression correlates with resistance to chemotherapy (233), further highlighting its role in hypoxia-mediated tumour survival strategies.

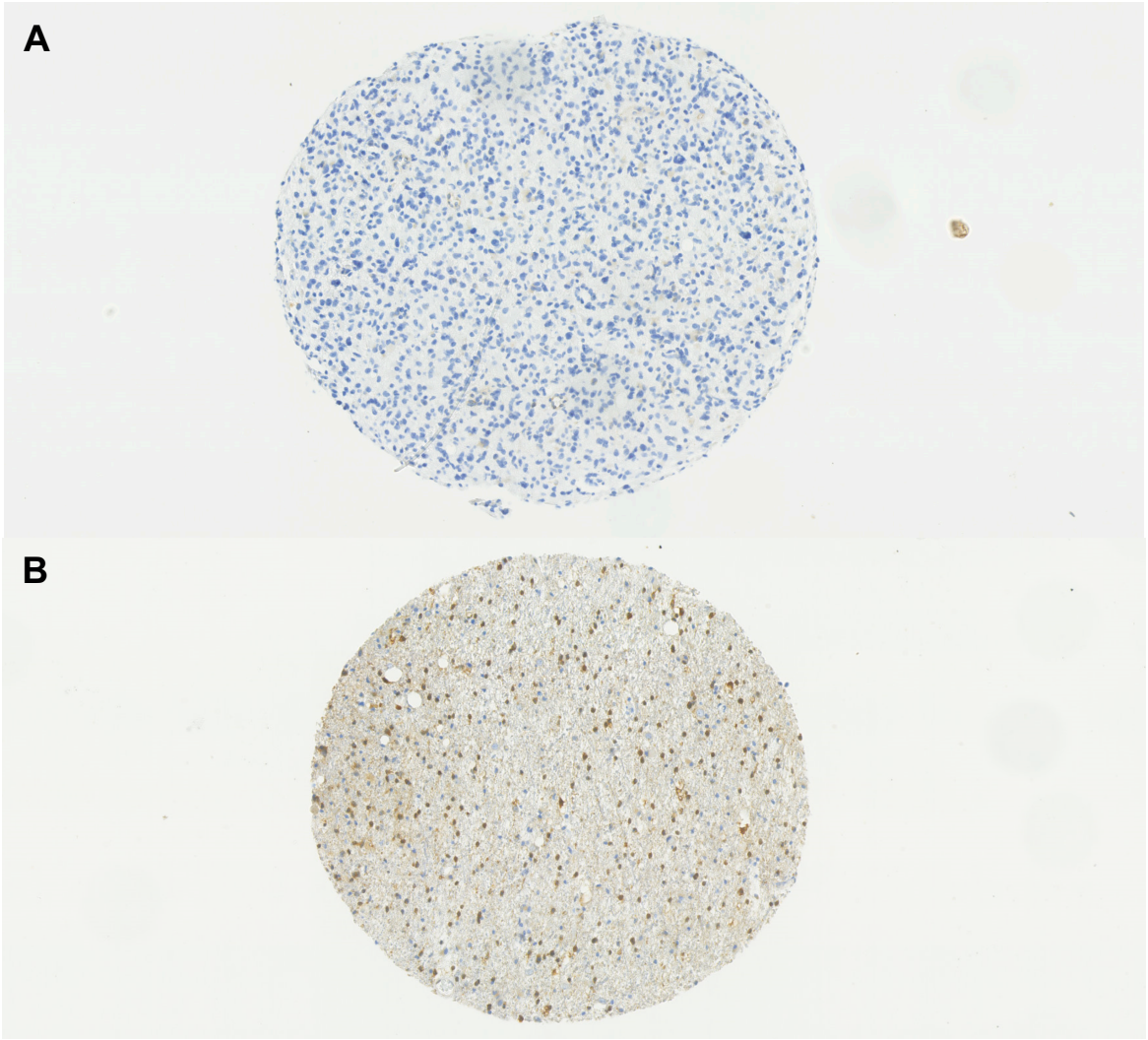


Figure 7.1: Representative NDRG1 Immunohistochemistry in GBM Tissue Microarray Core Immunohistochemistrical (IHC) staining for NDRG1 in human glioblastoma (GBM) tissue microarray (TMA) cores demonstrates the heterogeneity in protein expression on the same TMA. (A) A core with negative staining (blue) shows no detectable NDRG1 signal, indicating an absence of target protein expression. (B) A representative NDRG1-positive core exhibits widespread brown staining, representing convincing NDRG1 expression throughout the tumour tissue. These contrasting examples highlight inter-sample variability in NDRG1 level across the TMA.

In my analysis of NDRG1 expression in GBM, performed on a TMA prepared by the Histology Research Service at the IGC (methods detailed in chapter 2.1.4), the IHC staining revealed pronounced heterogeneity consistent with hypoxia-dependent regulation. Following staining, I digitally scanned all slides using the Nanozoomer XR,

DAB staining and performed semi-quantitative analysis using QuPath software. My evaluation of approximately 70 TMA cores demonstrated that fewer than 10 exhibited robust, diffuse NDRG1 positivity (Figure 7.1, panel B), while the majority showed either absent or only focal, low-intensity staining resembling basal normoxic expression levels (Figure 7.1, panel A). These findings from my digital analysis reflect the spatially restricted nature of hypoxia in GBM, where severe oxygen deprivation – and consequent upregulation of hypoxia-inducible factors like NDRG1 – typically occurs in discrete niches such as perinecrotic zones. The quantitative approach enabled by whole-slide imaging and QuPath analysis revealed that only a small proportion of tumour areas showed strong NDRG1 activation, supporting the concept that metabolic stress and hypoxia are not uniformly distributed throughout GBM tissues. In my assessment of positive cores, NDRG1 localization was predominantly cytoplasmic (confirming reported hypoxia-induced accumulation patterns), with nuclear translocation when the TMA core was positive. These results, obtained through my systematic digital pathology assessment, show that NDRG1 plays a role in some capacity in response to GBM; however, it does not show enough selectiveness and reliability to consider it for therapeutic purposes in light of the molecular heterogeneity present within tumour microenvironments in my study cohort.

7.2. HOXC10 Validation

Following identification of HOXC10 as an RNA target through de novo bulk RNA sequencing analysis of tissue samples, subsequent validation was performed via Western blotting (Figure 6.4, Figure 6.6), which confirmed its presence at the protein level. However, while Western blotting provides evidence for expression, it does not offer spatial resolution regarding tissue localization. Therefore, immunohistochemistry (IHC) was employed to visualize HOXC10 protein distribution within glioma tissue samples.

Localisation and Intensity
HOX genes typically exhibit low expression levels in adult tissues (250), functioning primarily in cellular maintenance. However, HOXC10 has been reported to be re-expressed in several tumour types (251), playing a role in oncogenic pathways. The IHC analysis aimed to assess both the localization and staining intensity of HOXC10 within glioma tissues.

The staining results revealed a predominantly weak signal, making it challenging to determine if staining was specific or a result of background interference. In most tissue cores, HOXC10 localization was observed at the membrane and cytoplasmic regions, rather than a clear nuclear signal. This finding may suggest a context-dependent role of HOXC10, potentially linked to its non-transcriptional functions in tumorigenesis (Figure 7.2).

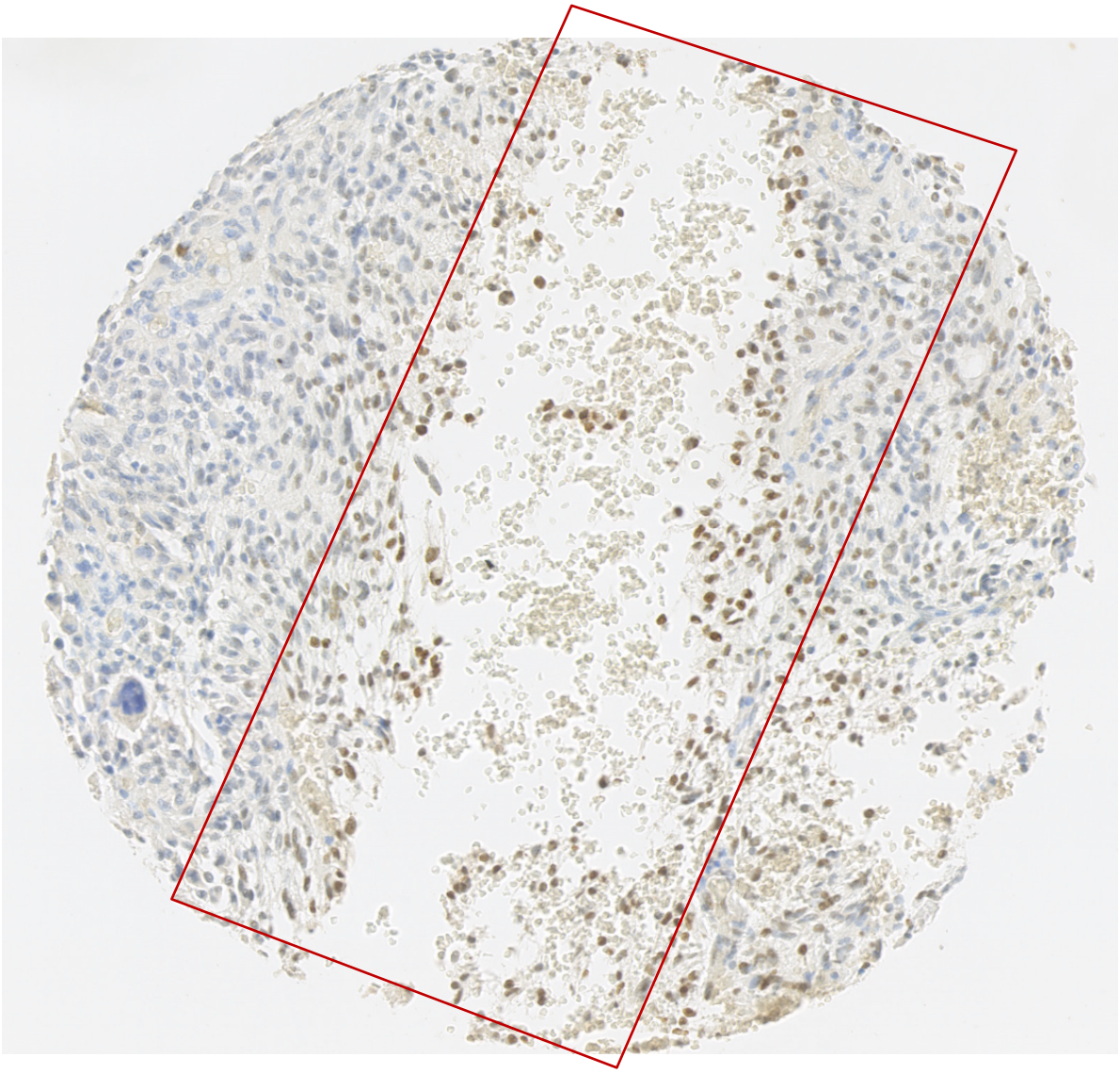


Figure 7.2: Heterogeneous HOXC10 expression in GBM tissue
Immunohistochemical staining shows focal HOXC10 positivity (brown, DAB staining, indicated by red rectangle) with mixed nuclear and cytoplasmic localization, against otherwise predominantly negative tumour regions (stained in blue). The spatially restricted expression pattern could suggest: (1) microenvironment-specific activation of HOXC10-associated pathways, or (2) distinct tumour cell subpopulations with differential HOXC10 regulation.

Regional Variability of HOXC10 Expression
Despite the overall weak signal, specific tissue cores demonstrated notable positive staining:

- Core B13: Displayed a few positively stained cells, particularly in regions adjacent to necrotic areas. Given HOX genes are associated with developmental processes, the presence of HOXC10 in transitioning regions between viable necrotic areas may suggest a role in glioma cell plasticity or tumour progression.
- Core E6: One of the few cores where a distinct nuclear staining pattern was observed. This suggests that in specific tumour environments, HOXC10 may retain its canonical nuclear function, potentially regulating transcriptional programs relevant to gliomagenesis.

- Core E10 (Figure 7.3): An intriguing finding from the IHC analysis was – when staining was distinctive – frequent staining of blood vessels and endothelial cells. While the specificity of this signal requires further validation, it is consistent with prior studies (251) demonstrating that HOXC10 promotes angiogenesis in gliomas. Specifically, HOXC10 overexpression has been reported to enhance angiogenesis via PRMT5 interaction and upregulation of VEGFA expression (252). This observation also suggests HOXC10's involvement in tumour vascular remodelling, aligning with the HOX family's role in developmental pathways and tumour progression; however, suggests different downstream affects for each HOX gene.

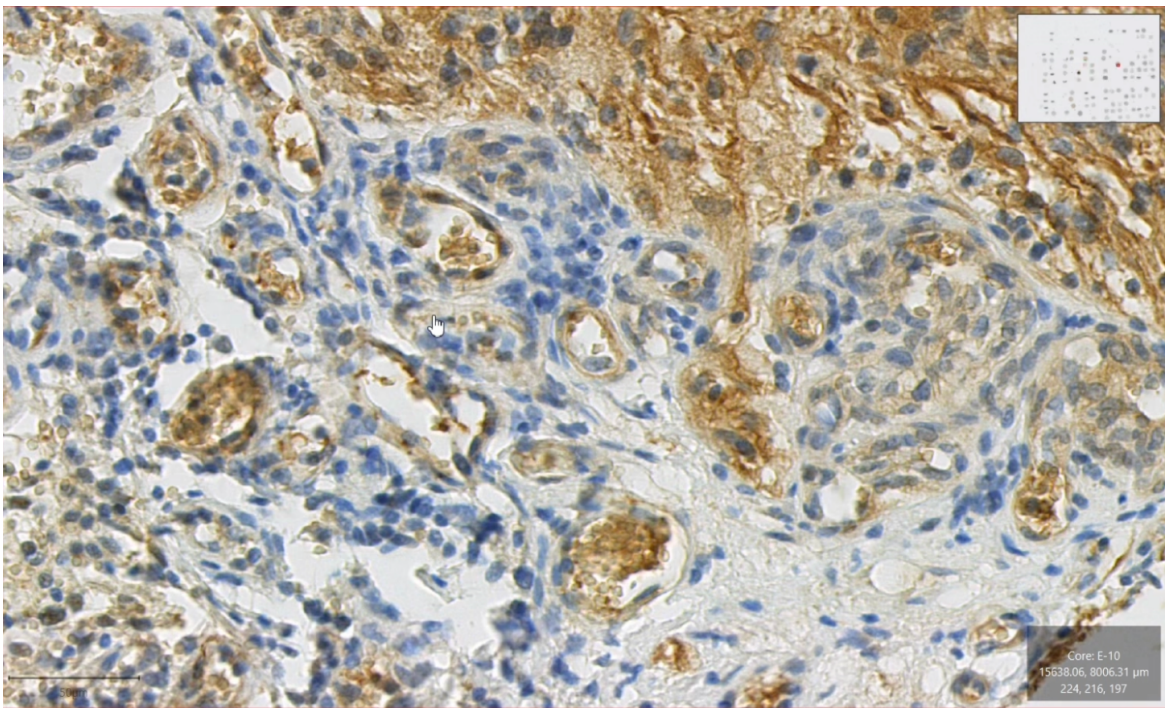


Figure 7.3: High-magnification IHC TMA stained for HOXC10 expression core E10
This representative image depicts a zoomed-in view of TMA core E10, stained to highlight HOXC10 spatial distribution and cellular localization patterns of protein expression. The brown DAB signal indicates positive immunoreactivity, while hematoxylin counterstaining highlights cell nuclei in blue. In this section, HOXC10 signal is distinctly localized to vascular structures, particularly in endothelial cells and cells lining blood vessels, suggesting a potential role in vascular biology or endothelial function. Notably, the staining pattern is predominantly extranuclear, with HOXC10 expression restricted to the cytoplasmic and possibly membranous compartments of these cells, lacking appreciable nuclear localization. This observation may support hypotheses of non-transcriptional roles for HOXC10 in these cellular populations or indicate cell-type specific localization variability.

In conclusion, the IHC analysis of HOXC10 expression in glioma tissues revealed a weak and variable staining pattern, with localized expression in membrane/cytoplasm and occasional nuclear structures and endothelial cells. Additionally, its potential association with vascular structures and endothelial cells aligns with prior findings linking HOXC10 to angiogenesis. Given the heterogeneous expression profile, further experiments – such as co-staining with endothelial markers (eg. CD31, VEGFA) and validation using RNA in situ hybridization (RNAscope) – are suggested to confirm the functional significance of these findings. Moreover, the presence of HOXC10 in transitioning necrotic zones suggests that its re-expression in gliomas may be spatially

and temporally regulated, warranting further investigation into its potential role in tumour hypoxia adaptation and glioma progression.

7.3. COL6A3 as a Recurrence-Associated ECM Remodeller

Bulk RNAseq analysis of primary vs recurrent glioma tumour samples revealed an upregulation of collagen (COL) genes, particularly COL6A, in recurrent tumours. Given its potential role as a relapse biomarker, COL6A was further analysed using IHC tissue microarrays (TMA) (Figure 7.4).

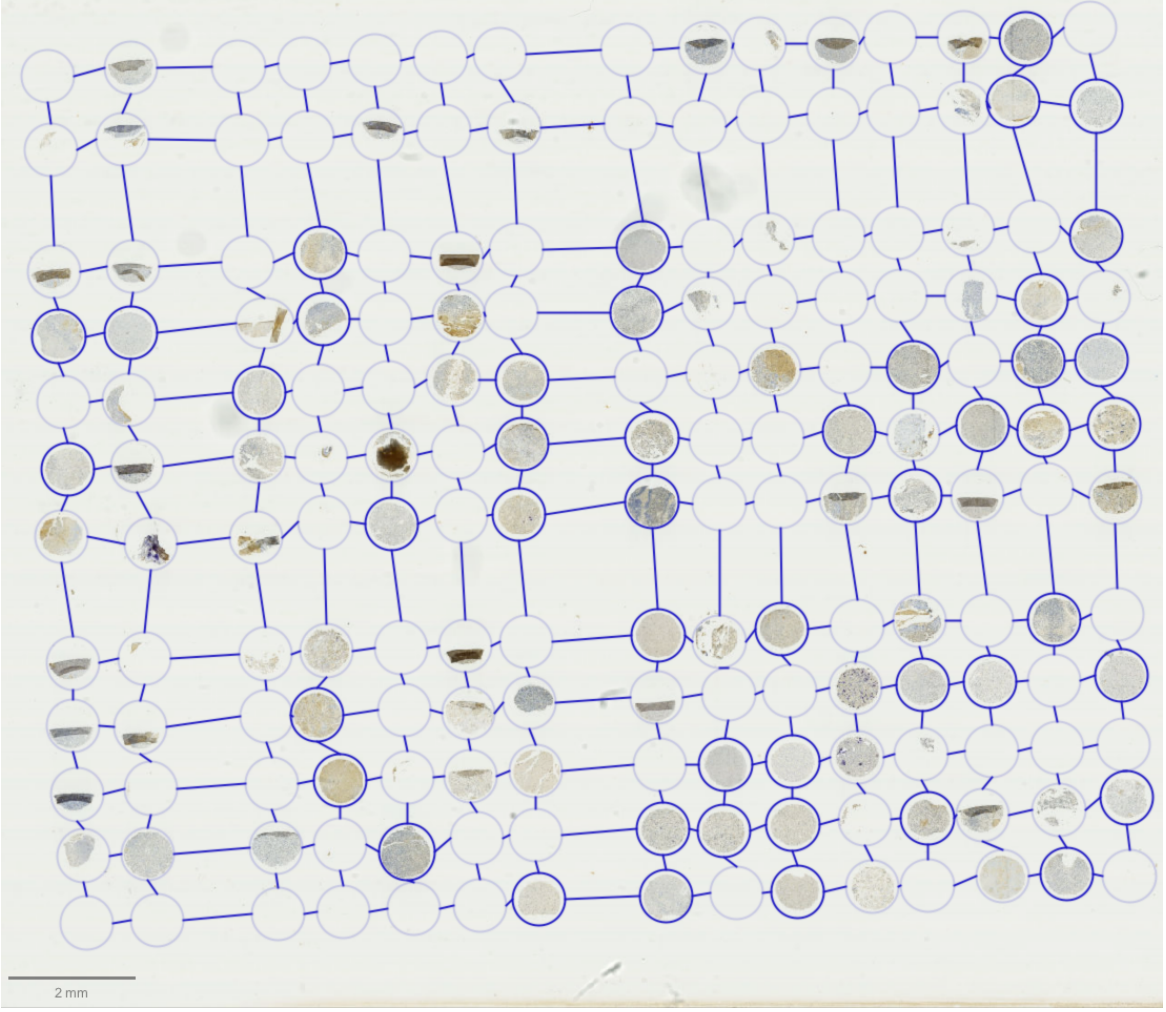


Figure 7.4: Representative Image of a Tissue Microarray (TMA) Slide
This image shows a representative tissue microarray (TMA) layout in QuPath software captured under a Nanozoomer scanner. Each circular spot represents a single tissue core embedded within the paraffin block and sectioned into a slide. The TMA is composed of a grid of tissue cores systematically arrayed in a defined pattern, with connecting lines demarcating their relative positions. The variation in staining intensity (background colour normolized) and morphology across different cores reflects heterogeneity in tissue types, fixation, or staining quality. The scale bar at the bottom left aids the interpretation of size as the user zooms into the image.

Notably, COL6A3 is already established as a prognostic marker in renal cancer according to the Human Protein Atlas (269). While its role in glioblastoma may not be directly established, this precedent strongly suggests potential parallel functions in glioma progression, particularly given the observed enrichment of collagen-related pathways in my recurrent comparison. COL6A staining was assessed alongside

macrophage infiltration, angiogenesis and interferon system activation to evaluate its relevance in the glioma environment. I observed similarly low consistency between IHC antibody staining and RNA expression signal from collagen as it was reported on the Human Protein Atlas for COL6A3. Staining location was cytoplasmic or membranous.

The parallel assessment of CD163+ tumour-associated macrophages (TAMs) and collagen deposition in glioblastoma reveals critical synergies within the tumour microenvironment (TME). CD163, a marker of immunosuppressive M2-like TAMs (111), and fibrillar collagen, a hallmark of desmoplastic ECM remodelling (108), co-localize in hypoxic niches to promote treatment resistance and invasion. This interplay is driven by hypoxia-inducible factors (HIF1a), which simultaneously recruit TAMs through chemokine signalling (115) and stimulate collagen crosslinking via enzymes like P4HA1 (223). Historically, CD163+ macrophage infiltration into glioma tissues correlates with regions of dense collagen networks, suggesting TAMs actively remodel the ECM to facilitate tumour spread (109).

Clinically, their co-enrichment identifies high-risk zones: collagen provides a physical scaffold for tumour cell migration, while TAMs secrete MMPs and TGFβ to further degrade and reorganize the matrix (112). This reciprocal relationship creates a feedforward loop – ECM stiffness enhances macrophage polarization toward pro-tumour phenotypes, which in turn exacerbate fibrosis (121). Notably, similar CD163-collagen crosstalk is observed in renal and pancreatic cancers (Human Protein Atlas), implying conserved mechanisms across malignancies. In GBM, targeting this axis (eg. through CSF1R inhibitors coupled with collagenase) may disrupt stromal barriers to immune infiltration and chemotherapy delivery, though spatial heterogeneity demands precise therapeutic stratification.

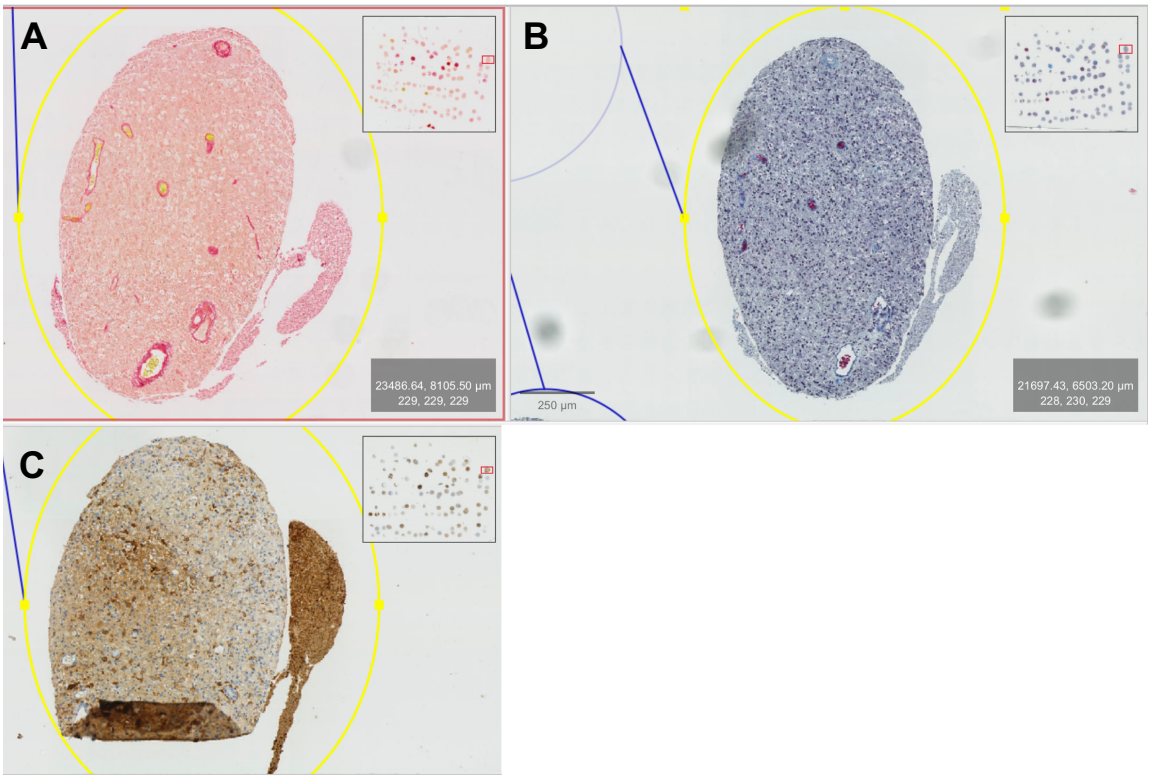


Figure 7.5: IHC stained TMA cored of glioma tissue samples
The 3 cores shown on the different panels represent slices from the same block. (A) H&E-stained core provides morphological context for tissue architecture and cellular organization. (B) Serial section stained for collagen illustrates the distribution of fibrillar extracellular matrix components, with light blue staining highlighting collagen-rich zones. (C) Core stained for CD163, a marker of tumour-associated macrophages (TAMs), reveals strong and diffuse immunoreactivity (score=3), indicative of high macrophage infiltration throughout the tumour microenvironment. The lack of spatial restriction suggests a widespread immunomodulatory presence rather than confinement to stromal or collagen-rich areas.

IHC staining was evaluated using a semi-quantitative scoring system (0 = none, 1 = low, 2 = moderate, 3 = strong intensity) to assess spatial relationships between macrophage infiltration and collagen localization. Analysis revealed robust CD163+ macrophage infiltration (score=3) distributed diffusely throughout the tissue sample, indicating widespread TAM activity (Figure 7.5). The strong but non-localized CD163 signal suggests systemic immune modulation rather than confinement to specific collagen-rich zones, consistent with the pro-invasive role of TAMs in glioma progression (111, 112).

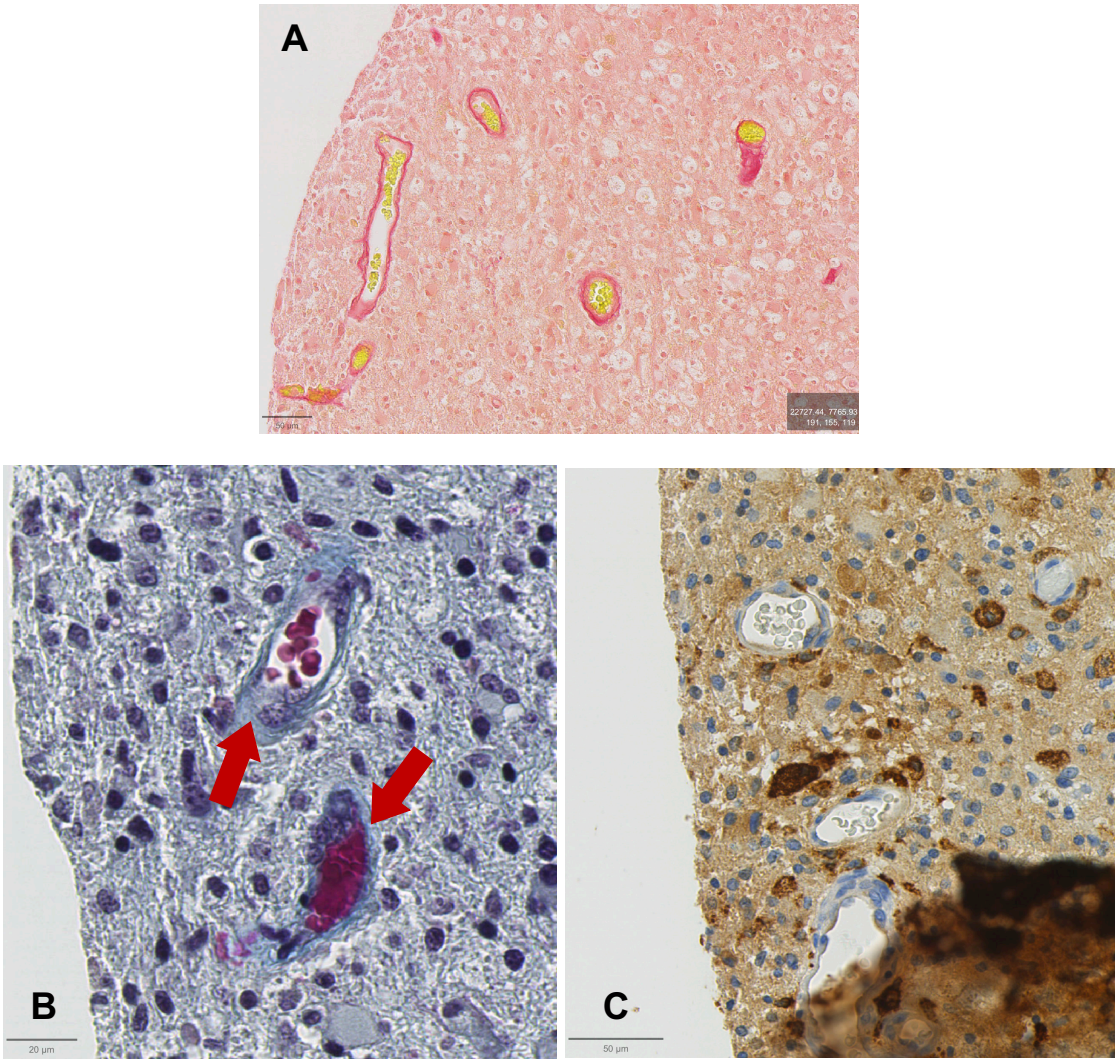


Figure 7.6: Zoomed-in view of glioma TMA Core A1
This figure shows higher-magnification images of the cores from Figure 7.5. (A) H&E staining provides histological context, showing tumour cell density and vasculature within the core. (B) Collagen staining (light blue) highlights regions of dense fibrillar deposition, with red arrows indicating areas where collagen appears to form concentric, sleeve-like arrangements around putative blood vessels. These structures resemble the angiogenic islands described in other malignancies and are consistent with previously reported patterns of COL6A3-enriched perivascular collagen organization. Although endothelial markers were not included in this panel, the anatomical positioning of these collagen sleeves strongly implies vascular coupling. (C) Immunohistochemistry for CD163 shows robust, diffuse staining indicative of widespread tumour-associated macrophage (TAM) infiltration. The lack of specific colocalization with collagen-rich vascular regions suggests that TAMs are distributed more globally across the tumour microenvironment, potentially contributing to broader extracellular matrix remodelling rather than concentrating at perivascular zones. Scale bars denote magnification.

A striking observation was the localization of fibrillar collagen (light blue) in concentric networks around putative vascular structures, creating distinct tubular patterns (marked by red arrow on Figure 7.6) that morphologically resemble the angiogenic islands (Appendix Figure 18) described in other aggressive malignancies (37). While endothelial markers were not stained, the characteristic collagen sleeves – particularly enriched in COL6A3 – strongly suggest organization around blood vessels based on their anatomical distribution and prior reports of collagen-vascular coupling in renal carcinomas (human Protein Atlas). This spatial relationship implies a functional triad: (1) hypoxic stress induces COL6A3-mediated ECM remodelling (223), (2) the resulting

fibrillar networks provide structural guidance for both infiltrating macrophages (112) and potential vascular elements, and (3) TAMs further reinforce this architecture through MMP-mediated matrix processing (108). The recurrence of this pattern in our samples, particularly at invasive margins, underscores COL6A's potential as a therapeutic target for disrupting glioma TME organization.

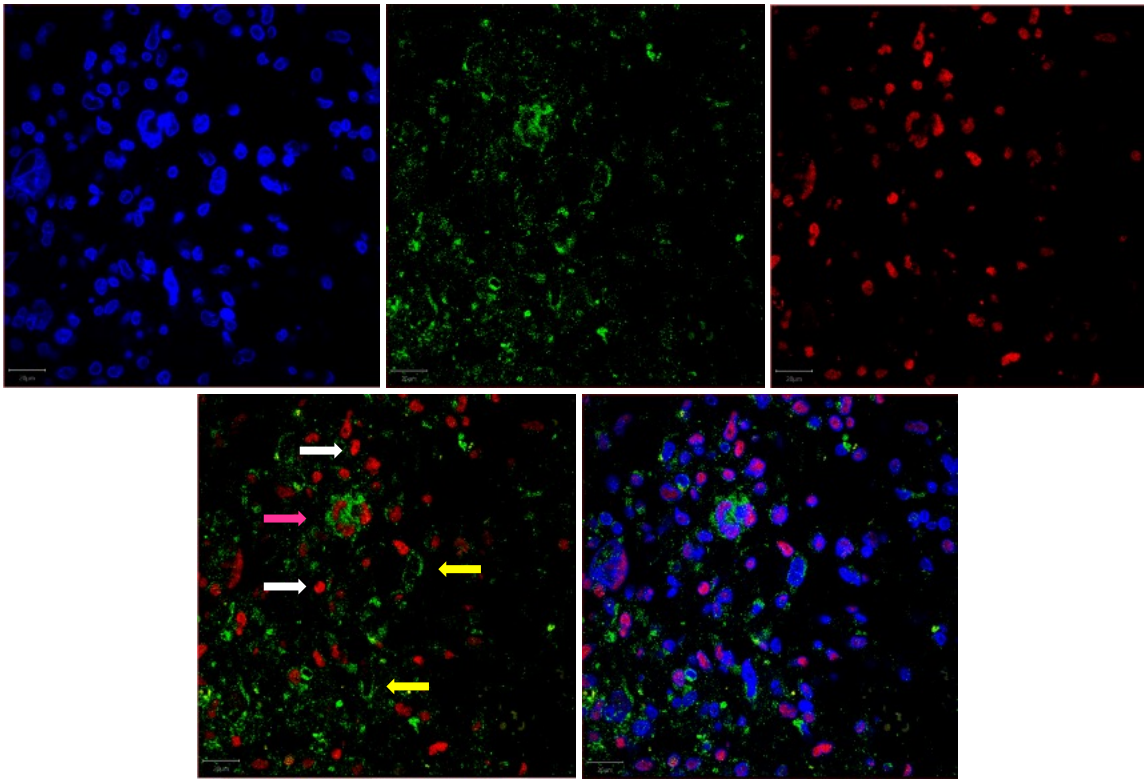


Figure 7.7: Analysis of COL6 expression in relation to SOX2 positive cancer stem cells. Slices were incubated with antibodies to either COL6 or SOX2 and indicated with secondary antibodies that are labelled with green or red. Arrows indicate the position of SOX2 only (red) COL6 (green) or cells co-expressing both. Curtesy of Ashita Singh and Sofian Al Shboul

The incorporation of SOX2 staining, a canonical marker of GSCs was implemented to investigate potential functional interactions between COL6-enriched niches and stem-like populations implicated in tumour recurrence. This analytical approach was selected based on two key considerations: (1) established evidence that SOX2+ GSCs demonstrate therapy resistance and contribute to relapse initiation (123), and (2) the potential for ECM remodelling components such as collagen to participate in protective niche formation for these critical cell populations.

The co-staining analysis revealed a preferential localization of SOX2+ cells within COL6-rich perivascular zones (Figure 7.7), suggesting these remodelled extracellular matrices may function as protective microenvironments for stem-like cell populations. These findings extend previous observations regarding collagen's roles in vascular co-option and macrophage recruitment by demonstrating a tripartite interaction system: hypoxic stress (223) promotes COL6 deposition, which concurrently stabilizes developing vasculature, directs macrophage infiltration (112), and provides anchoring sites for SOX2+ GSCs that drive recurrent disease. Spatial analysis demonstrated a significant association between SOX2+ cells and COL6-ensheathed vascular structures (Figure 7.7), implying potential mechanotransductive signalling between

ECM stiffness properties and stemness pathway activation – a phenomenon documented in other aggressive malignancies (Human Protein Atlas).

From a therapeutic perspective, these findings provide a potential explanation for the particular resilience of COL6-overexpressing recurrent tumours, as their collagen-rich microenvironments appear to simultaneously harbour both pro-angiogenic macrophages and therapy-resistant GSC populations. This suggests that therapeutic strategies targeting COL6A3 – potentially through miRNA-based approaches combined with CSF1R inhibition - might simultaneously disrupt vascular support systems, immune evasion mechanisms, and stem cell survival pathways.

Conclusion and Future Directions

The three key targets validated in this chapter – NDRG1, HOXC10, and COL6A3 – originated from fundamentally different experimental approaches, each providing unique insights into glioblastoma pathogenesis. NDRG1 was initially identified through controlled hypoxic cell line studies, offering a purified view of oxygen-deprived tumour cell responses. This reductionist approach clearly established its role as a HIF-1α-regulated stress effector but lacked the complexity of intact tumour microenvironment interactions. In contrast, HOXC10 emerged from comparative analysis of low- versus high-grade glioma tissues, capturing clinically relevant expression patterns tied to tumour progression and vascular remodelling. While this tissue-based approach revealed important grade-dependent associations, the weak and variable protein expression observed in IHC suggests potential post-transcriptional regulation or technical limitations. Most compellingly, COL6A3 was discovered through direct comparison of primary versus recurrent tumour samples, providing unambiguous clinical relevance to treatment resistance and disease relapse. The interesting spatial organization of COL6A3 around vascular structures and its co-localization with stem cell markers in recurrent tumours offers powerful evidence for its role in creating protective niches, though the observed RNA-protein discordance warrants cautious interpretation.

Despite their distinct origins, these targets converge on a unified hypoxia adaptation axis. NDRG1 operates at the cellular level, mediating intrinsic stress responses through metabolic reprogramming. HOXC10 functions at the vascular interface, potentially guiding angiogenic remodelling. COL6A3 orchestrates microenvironment-wide changes through ECM restructuring and stem cell niche formation. This multi-scale integration—from cell-autonomous survival (NDRG1) to intercellular crosstalk (HOXC10) to structural reorganization (COL6A3)—explains why single-target approaches often fail in glioblastoma. The cell line-derived NDRG1 findings provide mechanistic clarity but require validation in more complex systems. The tissue-based HOXC10 and COL6A3 data offer clinical relevance but benefit from reductionist follow-up studies to isolate specific cellular interactions.

The complementary strengths of these discovery platforms become particularly valuable when considering therapeutic development. NDRG1's consistent hypoxic induction suggests utility as a reliable biomarker, though its ubiquitous expression may limit specificity. HOXC10's vascular association offers potential for targeted delivery but requires improved detection methods. COL6A3's recurrence-linked patterning presents exciting opportunities for niche disruption but poses delivery challenges. Importantly, the limitations of each approach are mitigated when these targets are considered as an integrated network rather than isolated entities. This systems

perspective informs our subsequent miRNA strategy, which aims to concurrently modulate all three pathways to overcome the adaptive resistance that characterizes glioblastoma progression.

CHAPTER 8. Potential Plant Medicinal miRNA Discovery Using Deep Learning

The discovery of plant-derived medicinal microRNAs (miRNAs) has emerged as a fascinating area of research, promising to introduce novel therapeutic approaches for various diseases. This chapter outlines the steps I took to identify *de novo* plant miRNAs with potential medicinal applications, using a deep learning-based tool like miTAR and a database such as MepmirDB. My focus was on identifying miRNAs targeting genes implicated in hypoxic glioblastoma and glioma tissues, specifically using RNA sequencing (RNAseq) datasets and bioinformatic workflows. Despite encountering computational challenges, this analysis yielded thousands of potential miRNA-mRNA pairs, paving the way for experimental validation. The following sections explain the workflow, challenges faced and implications of this research.

8.1. Workflow

8.1.1. Selections of target genes

The first step of this project was to identify potential gene targets for plant-derived miRNAs. These steps were outlined in previous chapters (CHAPTER 4, CHAPTER 5, CHAPTER 6), but briefly, here is a summary. RNAseq data from hypoxic glioblastoma cell lines and glioma tissues provided a comprehensive view of gene expression changes under pathological conditions. Bulk RNAseq analysis of hypoxic glioblastoma cells identified NDRG1, and EGLN3 as significantly upregulated genes. These genes are well-known for their roles in hypoxic responses and tumour progression. In addition, analysis of glioma tissue RNAseq data revealed the HOX gene family, particularly HOXC10 and HOXB3, as interesting targets. These genes are involved in cell differentiation and oncogenesis, making them potential candidates for therapeutic targeting. The 3’ UTR sequences of these genes were retrieved from the NCBI database as inputs for subsequent miRNA-mRNA binding analyses.

8.1.2. Selection of plant miRNAs

To identify plant miRNAs with therapeutic potential, I identified MepmiRDB as the most advantages database to use. The database contained miRNA sequences from 29 different plants that have been associated with medicinal benefits. This number was subsequently reduced to 8 plants (kiwi (ach), sweet oranges (csi), lychee (lch), ginseng (pgi), opium poppy (pso), pomegranate (pgr), raddish (rsa)), as the selection criteria focused on plants with commercial availability and market potential, excluding rare or inedible species with limited practical use. For instance, species like kiwi (*Actinidia chinensis, ach*), were included due to their accessibility, while obscure plant found in the eastern desert of Central Asia like the evergreen broadleaf shrub (*Ammopiptanthus mongolicu, amo*) were excluded. The miRNA sequences for these plants were obtained from the MepmiRDB database as the second input for the subsequent computational analysis.

8.1.3. Computational workflow using miTAR

To predict miRNA-mRNA binding, I used miTAR, a computational tool that integrates miRNA sequence data with mRNA 3’ UTR sequences to assess binding probabilities. The inputs for miTAR were the 3’ UTR sequences of NDRG1, EGLN3, and every member of the HOX gene family, along with the miRNA sequences from the selected

plants (Appendix Figure 23). MiTAR evaluates binding probabilities based on see region complementarity and mRNA secondary structure accessibility.

One significant limitation was the computational demand of miTAR, particularly given the large number of miRNA sequences and target genes in the analysis. Initial attempts to run the algorithm locally were successful with miRNA sequences belonging to one plant only (smaller datasets are typically used during the implementation phase of an algorithm) but failed during the scaling up phase (adding more plant sequences) due to insufficient computational memory resources (RAM) (Appendix Figure 20). To address this, I created a Docker container to manage and standardize the computational environment (Appendix Figure 21-Appendix Figure 22). This initially solved the issue; however, later, failed again when more mRNA sequences were added.

Subsequently, I implemented the Docker container on Eddie, the university's high-performance computing cluster. While this setup successfully ran the analysis, it generated an overwhelming number of results, with millions of miRNA-mRNA pairs predicted. This output underscored the inherent limitation of computational miRNA-mRNA predictions, as miRNAs are small and theoretically capable of binding to many sequences and many parts of a single sequence, often resulting in high rates of false positives.

miRNA-mRNA pairs analysed included:

miRNAs	3' UTR of genes	Number of predicted pairs	
		ALL binding probability	0.99< binding probability
'Edible plants' 2081 sequences from 5 plants	All HOX genes	4,138,845	2,825,616
'Edible plants' 2081 sequences from 5 plants	Genes of interest (NDRG1, EGLN3, P4HA1, PKM, HOXC10, HOXB3, COL6A3)	FAILED TO RUN	
All 106 Actinidia chinensis (ach) miRNAs	GSC322 top 8 genes		1725
'Edible plants' 2081 sequences from 5 plants	Tissue and GSC327 genes overlap (full list in Appendix Table 10)		8,525,697
All 106 Actinidia chinensis (ach) miRNAs	Common upregulated in GSC322	14,973	10,891
All 106 Actinidia chinensis (ach) miRNAs	Common upregulated GSC327	18,071	13,246

Table 8.1: Summary of tested miRNA-gene pairs using miTAR
The miTAR deep-learning prediction binding prediction algorithm was used to predict binding interactions between plant-derived miRNAs from the MepmiRDB database and the 3' UTR sequences of genes of interests identified through de novo analyses. Screening 2,081 miRNAs from five edible plant species (pomegranate, ginseng, lychee, sweet orange, and opium poppy) against all HOX genes predicted 4,138,845 potential binding pairs, of which 2,825,616 showed high-confidence binding probabilities (p>0.99). However, attempts to analyse interactions with specific genes of interest (NDRG1, EGLN3, P4HA1, PKM, HOXC10, HOXB3, COL6A3) failed due to computational limitations. Smaller-scale analyses using only Actinidia chinensis (kiwi) miRNAs produced more manageable results, with 1,725 predicted interactions for the top 8 hypoxia-responsive genes in GSC322 cells. The analysis of tissue and GSC327 overlapping genes yielded 8,525,697 predicted pairs, demonstrating the substantial scale of potential interactions that require stringent filtering for biological relevance.

8.1.4. Filtering strategies

To manage the excess number of predicted interactions, I applied several filtering strategies:

1) **Stringent plant selection:** Upon initial implementation of the algorithm, I used all miRNA sequences belonging to all 29 plants. Due to the demanding computational power needed and the overwhelming number of results, I attempted to reduce the number of miRNA candidates by limiting the analysis to a smaller subset of plants. Plants were first reduced to commercially available and edible plants like kiwi (ach), sweet oranges (csi), lychee (lch), ginseng (pgi), opium poppy (pso), pomegranate (pgr), raddish (rsa). This greatly reduced the number. However, the resulting output was still in the neighbourhood of millions of candidate pairs. In another iteration filtering, I decided to remove kiwi and raddish, as the miRNA samples were taken from the plant's leaves as opposed to its edible part.
2) **Probability distribution analysis:** Next, I examined the distribution of binding probabilities across the predicted pairs. miTAR automatically discards all pairs that have a binding probability of $p < 0.5$. Contrary to expectations of a normal

or a right-skewed distribution, the results displayed a heavily left-skewed distribution with a peak near 1.0 probability. This anomaly is an indicator of potential model overfitting, where the algorithm has learned to fit the training data (eg. known miRNA-mRNA interactions) too closely, including noise or specificities not generalizable to new, unseen plant minra sequences. An overfitted model loses predictive power and, as observed, generates implausibly high-confidence predictions for a vast majority of input pairs, severly compromising its utility for discovery. This suggested a potential issue with the algorithm or its parameterization, as it implied an unusually high confidence for most predictions.

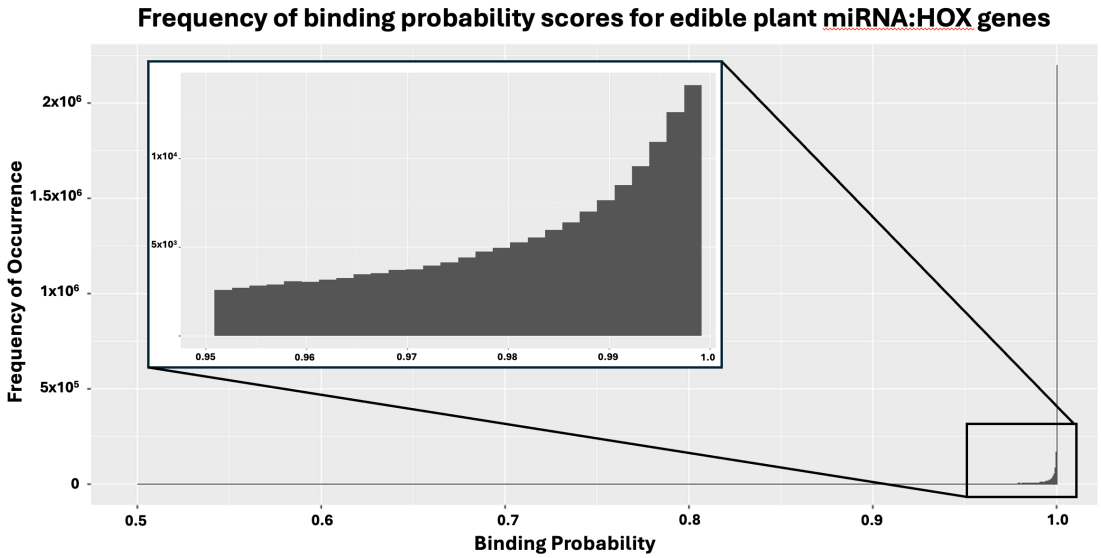


Figure 8.1: Frequency of binding probability scores of plant miRNA and HOX genes

To address the distribution anomaly, I contacted the developers of miTAR. They acknowledged the unusual results and indicated that an update to the algorithm was "in development" at the time. However, no subsequent update has been released to date, leaving the issue unresolved.

Compounding the challenges, after a short while a suspected update to one of Eddie's machine learning packages rendered the miTAR implementation non-functional (Appendix Figure 25). Despite assistance from the Eddie support team, the algorithm could not be restored, preventing me from re-running the analysis to refine the results further. Another capability of the miTAR algorithm is to control the number of target sites per mRNA, which initially was set to 1. However, in subsequent runs of the algorithm, this number would've been increased and tested whether increasing this number would have resulted in less pairs and improved off-target effects.

8.1.5. miTAR results

Despite the challenges, the analysis yielded thousands of potential miRNA-mRNA interactions. While this result highlights the limitations of current computational tools in filtering false positives, it also provides a starting point for experimental validation. Key miRNA candidates targeting NDRG1, EGLN3, HOXC10, and HOXB3 were identified, with high binding probabilities suggesting strong regulatory potential.

8.2. Experimental validation

The extensive list of predicted interactions necessitates experimental validation to identify functional miRNA-mRNA pairs. However, the scale of the dataset poses practical challenges, as validating thousands of interactions would require substantial time and financial resources. Critically, if the predictive model is overfitted, a significant proportion of these high-probability pairs may be false positives, making a naïve, high-throughput validation approach inefficient and costly. Random selection of miRNA candidates for experimental validation is one potential approach, but this strategy risks overlooking the most biologically relevant interactions. Alternatively, prioritizing miRNAs with the highest binding probabilities or targeting multiple genes involved in critical pathways may improve the efficiency of validation efforts. However, given the potential overfitting, this prioritization itself may be biased. A more robust strategy would be to integrate predictions from multiple, independently trained algorithms to build a consensus list, thereby mitigating the risk inherent in any single overfitted model.

Experimental techniques such as luciferase reporter assays, RNA immunoprecipitation, and CRISPR-based approaches could be employed to confirm miRNA-mRNA interactions. Additionally, integrating high-throughput sequencing methods, such as CLIP-seq and RNAseq following miRNA mimic transfections, may provide a broader view of miRNA activity. However, the feasibility of these techniques greatly depended on the lack of time left of my PhD project.

8.3. Discussion

This chapter demonstrates both the promise and challenges of using computational approaches to identify plant-derived miRNAs with therapeutic potential for glioblastoma. While our analysis using miTAR and MepmiRDB successfully predicted thousands of potential miRNA-mRNA interactions targeting key glioma-associated genes, several critical limitations emerged. The most notable was the algorithm's tendency to generate an implausibly high number of predicted interactions with artificially inflated binding probabilities, as evidenced by the left-skewed probability distribution where most predictions clustered near p=1.0. This pattern is strongly suggestive of model overfitting, a fundamental challenge in machine learning where a model performs well on its training data but fails to generalize to new data. In this context, overfitting likely arose from the application of a model trained primarily on animal (especially human) miRNA interactions compared to the distinct sequence and binding context of plant miRNAs, leading to poor generalization and an overestimation of true binding events. This overprediction problem was compounded by technical constraints that prevented analysis of some particularly relevant gene targets like HOXC10 due to computational errors. Our pragmatic decision to focus on commercially viable plant species (like pomegranate and ginseng) while excluding rare plants helped manage the overwhelming data volume but may have limited our ability to discover novel therapeutic miRNAs.

The challenge of overfitting directly impacts the translational potential of this project. The underscores that the raw output of a deep learning predictor cannot be taken at face value and must be interpreted with cation (as always). Future work must address this by: (1) utilizing algorithms specifically trained or fine-tuned on plant miRNA-mRNA interaction data, if available; (2) implementing rigorous cross-validation strategies

during model training to detect and penalize overfitting; and (3) employing ensemble methods that combine predictions from multiple models to improve robustness.

Moving forward, this work suggests three key priorities for the field: first, the development and most importantly, maintenance of robust prediction algorithms that incorporate plant-specific miRNA binding characteristics and better account for biological plausibility; second, a tiered experimental validation approach that progresses from computational predictions to cell-based assays, with an intermediate step of computational consensus-building to filter our likely false positives arising from overfitted models; and third, expansion of screening efforts to include both commercially viable and medicinal plant species to fully explore this untapped therapeutic resource. While the computational challenges were significant, this project provides an important foundation for future work at the intersection of plant biology and glioma therapeutics, highlighting both the potential of plant miRNAs as novel therapeutic agents and the current limitations that must be addressed to realize this potential.

General Discussion and Future Directions

The overarching goal of this project was to integrate multi-omics approaches – RNA sequencing, variant calling, miRNA analysis and immunohistochemistry (IHC) pathology image analysis – to identify key molecular alterations driving glioma progression and recurrence. The results provided a broad molecular landscape, revealing the dysregulation of genes such as HOXC10 and COL6A, alongside evidence of an evolving tumour microenvironment (TME) that supports invasion and recurrence. While these findings contribute valuable insights into glioma basic biology, several methodological limitations and alternative approaches could have been considered to strengthen the project's conclusions and refine future experimental directions.

Variant calling provided the first layer of molecular characterization, revealing potential regulatory mutations in genes associated with the interferon system, suggesting that immune evasion mechanisms are at play in glioma recurrence. However, HOXC10 itself did not harbour recurrent coding mutations, reinforcing the idea that its upregulation is primarily a transcriptional event rather than a consequence of genetic alteration. The presence of regulatory mutations in enhancer and promoter regions of HOX genes raises the possibility of epigenetic dysregulation, which could be further explored through Assay for Transposase-Accessible Chromatin sequencing (ATACseq), or Chromatin Immunoprecipitation Sequencing (ChIPseq) for HOX transcription factors. This would suggest whether chromatin remodelling events facilitate the reactivation of developmental pathways in glioma recurrence. Furthermore, the study relied primarily on whole-exome sequencing (WES), which focuses on protein-coding regions, meaning that important non-coding regulatory elements could have been overlooked. Future studies incorporating whole-genome sequencing (WGS) and, especially, long-read sequencing would allow for a more comprehensive assessment of structural variations and non-coding mutations driving gene expression changes.

RNA sequencing (RNAseq) was an important component of this project, enabling the identification of differentially expressed genes across glioma models and patient-derived tissues. The first phase of RNAseq analysis focused on in vitro glioma cell lines, where bulk RNAseq was performed to identify genes that exhibited significant expression changes under hypoxic conditions – an important characteristic of glioma tumours. Given the well-established role of hypoxia in driving glioma aggressiveness, understanding the molecular response of glioma cells to oxygen deprivation was a critical step in identifying potential therapeutic targets. Among the hypoxia-responsive genes identified, NDRG1 and EGLN3 emerged as promising candidates due to their role and relationship to oxygen sensing and cellular adaptation to hypoxic stress. NDRG1, a stress-responsive gene implicated in tumour suppression, metastasis and differentiation, was significantly upregulated under hypoxia, consistent with its known function as a downstream effector of the hypoxia-inducible factor (HIF) pathway. In contrast, EGLN3, which encodes a prolyl hydroxylase responsible for regulating HOF stability, also showed increased expression but failed to demonstrate reproducible protein validation via Western blotting. The inability to detect EGLN3 at the protein level suggested that neither post-transcriptional regulatory mechanism (eg. miRNA-mediated repression, protein degradation) was at play, or that the detected transcript represented a non-functional, unstable isoform, or the tested antibody was not sufficient. Given these inconsistencies, EGLN3 was excluded from further validation

experiments, while NDRG1 remained a key potential target for downstream IHC analysis to assess its expression in glioma tissues.

While bulk RNAseq provided a broad overview of gene expression changes, single-cell RNA sequencing (scRNAseq) would have significantly enhanced this project by resolving cellular heterogeneity within gliomas. Bulk RNAseq averages expression across cells, masking rare but critical subpopulations (eg. glioma stem cells or immune infiltrating cells) that drive glioma progression and/or recurrence. For example, scRNAseq could have clarified whether HOXC10 upregulation (and generally the HOX gene family) occurs uniformly or is it restricted to specific cell states/types, such as invasive fronts or perivascular niches. Similarly, it would have better characterized the cellular sources of COL6A (tumour cells vs fibroblasts) and identified potential co-expression patterns with immune markers. However, scRNAseq comes with trade-offs: higher cost, computational complexity, and challenges in resolving low-abundance transcripts like non-coding RNAs. Future studies should prioritize scRNAseq for dissecting glioma heterogeneity, particularly in recurrent tumours where subclonal evolution is likely.

The second phase of RNAseq analysis shifted focus from cell lines to patient-derived glioma tissue samples, enabling a broader exploration of gene expression changes associated with tumour grade and recurrence. When comparing low grade vs high grade gliomas, HOX genes emerged as a prominent differentially expressed gene family, with several members – including HOXC10, HOXB3, and HOXA9 – exhibiting significant upregulation in high grade gliomas. Given that HOX genes are typically silenced in adult tissues but play crucial roles in embryonic development and cellular differentiation, their reactivation in gliomas suggests a possible role in maintaining a stem-like, proliferative state that contributes to tumour progression. This aligns with previous literature suggesting that HOX gene re-expression in gliomas is associated with poor prognosis, increased invasiveness, and resistance to therapy. Furthermore, the primary vs recurrent glioma comparison identified collagen-related genes, particularly COL6A, as key targets upregulated in recurrent tumours. COL6A, a major extracellular matrix (ECM) component, plays a pivotal role in modulating the tumour microenvironment and supporting glioma cell survival, particularly under therapy-induced stress conditions. The overexpression of COL6A in recurrent gliomas suggests that ECM remodelling may be an important factor in glioma relapse, potentially providing a protective niche for tumour cells to evade treatment induced apoptosis. These RNAseq findings laid the foundation for subsequent validation experiments, where HOXC10 and COL6A were selected for protein level confirmation via Western blotting and spatial localization using IHC in TMAs.

The IHC validation phase provided spatial context for the RNAseq findings, confirming HOXC10 protein expression in glioma tissues but revealing an unexpected predominantly cytoplasmic and membrane-localized staining pattern rather than the anticipated nuclear localization. This observation raises important questions about non-canonical functions of HOXC10 in glioma, particularly given that HOX genes are traditionally understood as nuclear transcription factors. The possibility that HOXC10 may have a cytoplasmic role, such as interacting with signalling pathways cytoskeletal components, warrants further investigation through proteomic studies and co-immunoprecipitation assays to identify potential interacting partners. Additionally, the presence of HOXC10 in endothelial cells and blood vessels aligns with previous

findings that HOXC10 promotes angiogenesis via PRMT5 interaction and VEGFA upregulation, suggesting that its role in glioma progression extends beyond intrinsic tumour cell functions and may include vascular remodelling and adaptation to hypoxic conditions.

Similarly, COL6A IHC revealed its localization around perivascular structures, suggesting that collagen deposition is an important feature of glioma recurrence. The strong association between COL6A and CD163-positive macrophages highlights a potential tumour-supportive ECM-immune interactions, where collagen rich niches may serve as reservoirs for tumour-associated macrophages (TAMs). This aligns with previous literature suggesting that ECM remodelling facilitates immune cell recruitment, reinforcing a microenvironment conducive to glioma progression. This part of the project would have been benefited from co-staining with endothelial markers (eg. CD31) and hypoxia markers (HIF1a) to better determine the relationship between ECM deposition, vascular remodelling, and immune infiltration. Additionally, given the limitations of traditional IHC, RNAscope-based in situ hybridization could provide higher sensitivity in detecting HOXC10 and COL6A mRNA and stromal compartments.

A major limitation of the IHC analysis was the subjectivity inherent in manual scoring, which introduces observer bias and inter-sample variability. A more robust approach would involve AI- or machine learning-driven computational pathology methods, such as deep-learning-based image analysis tools, to quantify staining intensity, distribution, and co-localization with relevant cellular markers. Integrating computational pathology with spatial transcriptomics and proteomics could significantly enhance the resolution of glioma tissue architecture and the functional relationships between tumour cells, ECM components, and infiltrating immune cells.

This project aimed to provide a scientific base for the potential of using plant miRNA-derived medicines as a therapeutic for glioblastoma by using a computational approach, integrating deep learning (miTAR) with a medicinal plant database (MapmiRDB). While the approach successfully identified thousands of miRNA-mRNA pairs targeting hypoxic response genes (NDRG1, EGLN3) and glioma-associated HOX and COL genes, several critical challenges emerged. First, computational limitations – including memory constraints during scaling and instability of the Docker environment on high-performance computing cluster – restricted the analysis to only 8 commercially viable plant species (eg. pomegranate, ginseng) out of 29 initially considered. Second, miTAR algorithm exhibited fundamental reliability issues, producing a left-skewed binding probability distribution (peak near p=1.0) that suggested overprediction of interactions, a problem acknowledged but unresolved by the developers. These technical hurdles underscore a broader limitation in the field: current miRNA prediction tools prioritize sensitivity over specificity, generating overwhelming false-positive rates that complicate translational applications. Nevertheless, the workflow established here provides a framework for future studies, with three key recommendations: (1) experimental validation should prioritize miRNAs targeting multiple glioma-relevant genes (eg HOXC10+NDRG1) or those from plants with proven bioavailability (eg citrus miRNAs); (2) algorithm maintenance must be a priority if the field aims to advance at such a scale as protein structure prediction (eg. AlphaFold); (3) rare plant species warrant investigation despite practical challenges, as they may harbour unique miRNAs with higher target specificity. While this study

faced constraints in time and validation, it highlights the untapped potential of plant miRNAs as cross-kingdom regulators of glioma pathways and lays groundwork for their therapeutic development.

Given these findings, future studies should prioritize single-cell and spatial multi-omics approaches to refine our understanding of glioma progression at the molecular and cellular levels. Specifically, scRNAseq, spatial transcriptomics and multiplexed IHC could be used to explore cellular heterogeneity, ECM remodelling, and immune cell interactions within the glioma microenvironment. Additionally, functional studies investigating the role of HOXC10 in vascular remodelling and its potential as a therapeutic target could be explored, particularly in the context of angiogenesis inhibitors or HOX gene-targeted therapies. The relationship between COL6A, TAM infiltration, and glioma recurrence also warrants further investigation, as targeting ECM-immune crosstalk may represent a novel therapeutic avenue for preventing glioma relapse.

Overall, this project aimed to provide a comprehensive molecular framework for understanding glioma progression and recurrence, highlighting the reactivation of developmental programs (HOXC10) and ECM remodelling (COL6A), and immune-TME interactions (CD163-positive macrophages) as key drivers of tumour adaptation. Despite the methodological gaps, these findings lay a good base for future studies that incorporate the aforementioned technologies. By addressing these limitations, future research can move towards translating these insights into clinically actionable therapeutic strategies aimed at targeting the molecular and microenvironmental determinants of glioma biology.

References

1. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50(W1):W216-W21.
2. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
3. Ivan M, Kaelin WG, Jr. The EGLN-HIF O2-Sensing System: Multiple Inputs and Feedbacks. Mol Cell. 2017;66(6):772-9.
4. Cancer Stat Facts: Brain and Other Nervous System Cancer: National Cancer Institute (NIH); [Available from: https://seer.cancer.gov/statfacts/html/brain.html.
5. Cancer Stat Facts: Cancer of Any Site NIH: National Cancer Institute [cited 2024. Available from: https://seer.cancer.gov/statfacts/html/all.html.
6. Ramsden J. Bioinformatics: Springer Nature Switzerland AG 2023; 2023.
7. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) 2023 [updated May 16, 2023. Available from: www.genome.gov/sequencingcostsdata.
8. Kevin Blighe AL. PCAtools: PCAtools: Everything Principal Components Analysis. 2021.
9. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2018.
10. Kim HJ, Park JW, Lee JH. Genetic Architectures and Cell-of-Origin in Glioblastoma. Front Oncol. 2020;10:615400.
11. Welcome to Precision Medicine UK 2022 [Available from: https://precisionmedicineuk.com/homepage/.
12. About the Programme ed.ac.uk2022 [updated 22 Nov. 2022. Available from: https://www.ed.ac.uk/usher/precision-medicine/about-the-programme.
13. Office UGH. Genome UK: 2021 to 2022 implementation plan. 2021.
14. Wang Z, Liu X, Ho RL, Lam CW, Chow MS. Precision or Personalized Medicine for Cancer Chemotherapy: Is there a Role for Herbal Medicine. Molecules. 2016;21(7).
15. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision Medicine, AI, and the Future of Personalized Health Care. Clin Transl Sci. 2021;14(1):86-93.
16. Jorgensen AL, Prince C, Fitzgerald G, Hanson A, Downing J, Reynolds J, et al. Implementation of genotype-guided dosing of warfarin with point-of-care genetic testing in three UK clinics: a matched cohort study. BMC Med. 2019;17(1):76.
17. Higgins MJ, Stearns V. Understanding resistance to tamoxifen in hormone receptor-positive breast cancer. Clin Chem. 2009;55(8):1453-5.
18. Hartmaier RJ, Albacker LA, Chmielecki J, Bailey M, He J, Goldberg ME, et al. High-Throughput Genomic Profiling of Adult Solid Tumors Reveals Novel Insights into Cancer Pathogenesis. Cancer Res. 2017;77(9):2464-75.
19. Kim JC, Chan-Seng-Yue M, Ge S, Zeng AGX, Ng K, Gan OI, et al. Transcriptomic classes of BCR-ABL1 lymphoblastic leukemia. Nature Genetics. 2023;55(7):1186-97.
20. Xie N, Shen G, Gao W, Huang Z, Huang C, Fu L. Neoantigens: promising targets for cancer therapy. Signal Transduction and Targeted Therapy. 2023;8(1).
21. Cancer World Health Organization [Available from: https://www.who.int/health-topics/cancer#tab=tab_1.
22. Yin W, Wang J, Jiang L, James Kang Y. Cancer and stem cells. Exp Biol Med (Maywood). 2021;246(16):1791-801.
23. Global Burden of Disease Cancer C, Kocarnik JM, Compton K, Dean FE, Fu W, Gaw BL, et al. Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. JAMA Oncol. 2022;8(3):420-44.
24. Maringe C, Spicer J, Morris M, Purushotham A, Nolte E, Sullivan R, et al. The impact of the COVID-19 pandemic on cancer deaths due to delays in diagnosis in England, UK: a national, population-based, modelling study. Lancet Oncol. 2020;21(8):1023-34.
25. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100(1):57-70.
26. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
27. Montemurro N. Glioblastoma Multiforme and Genetic Mutations: The Issue Is Not Over Yet. An Overview of the Current Literature. J Neurol Surg A Cent Eur Neurosurg. 2020;81(1):64-70.
28. Talmadge JE, Fidler IJ. AACR centennial series: the biology of cancer metastasis: historical perspective. Cancer Res. 2010;70(14):5649-69.
29. Fidler IJ. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited. Nat Rev Cancer. 2003;3(6):453-8.
30. Tang Q, Su Z, Gu W, Rustgi AK. Mutant p53 on the Path to Metastasis. Trends Cancer. 2020;6(1):62-73.
31. Blasco MA. Telomeres and human disease: ageing, cancer and beyond. Nat Rev Genet. 2005;6(8):611-22.
32. Adair TH MJ. Angiogenesis: San Rafael (CA): Morgan & Claypool Life Sciences; 2010.
33. Carmeliet P. VEGF as a key mediator of angiogenesis in cancer. Oncology. 2005;69 Suppl 3:4-10.
34. Morana O, Wood W, Gregory CD. The Apoptosis Paradox in Cancer. Int J Mol Sci. 2022;23(3).
35. S Paget. The distribution of secondary growths in cancer of the breast. Cancer Metastasis Rev. 1889;8(2):98-101.

36. Jin MZ, Jin WL. The updated landscape of tumor microenvironment and drug repurposing. Signal Transduct Target Ther. 2020;5(1):166.
37. Buffa FM, Harris AL, West CM, Miller CJ. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. Br J Cancer. 2010;102(2):428-35.
38. Bhandari V, Li CH, Bristow RG, Boutros PC, Consortium P. Divergent mutational processes distinguish hypoxic and normoxic tumours. Nat Commun. 2020;11(1):737.
39. Bhandari V, Hoey C, Liu LY, Lalonde E, Ray J, Livingstone J, et al. Molecular landmarks of tumor hypoxia across cancer types. Nat Genet. 2019;51(2):308-18.
40. Pan C, Schoppe O, Parra-Damas A, Cai R, Todorov MI, Gondi G, et al. Deep Learning Reveals Cancer Metastasis and Therapeutic Antibody Targeting in the Entire Body. Cell. 2019;179(7):1661-76 e19.
41. Lopez-Soto A, Gonzalez S, Smyth MJ, Galluzzi L. Control of Metastasis by NK Cells. Cancer Cell. 2017;32(2):135-54.
42. Warburg O, Wind F, Negelein E. The Metabolism of Tumors in the Body. J Gen Physiol. 1927;8(6):519-30.
43. Chandel NS, McClintock DS, Feliciano CE, Wood TM, Melendez JA, Rodriguez AM, et al. Reactive oxygen species generated at mitochondrial complex III stabilize hypoxia-inducible factor-1alpha during hypoxia: a mechanism of O2 sensing. J Biol Chem. 2000;275(33):25130-8.
44. Peck B, Schulze A. Lipid Metabolism at the Nexus of Diet and Tumor Microenvironment. Trends Cancer. 2019;5(11):693-703.
45. Webb BA, Chimenti M, Jacobson MP, Barber DL. Dysregulated pH: a perfect storm for cancer progression. Nat Rev Cancer. 2011;11(9):671-7.
46. Ma X, Bi E, Lu Y, Su P, Huang C, Liu L, et al. Cholesterol Induces CD8(+) T Cell Exhaustion in the Tumor Microenvironment. Cell Metab. 2019;30(1):143-56 e5.
47. Zhang X, Feng Y, Liu X, Ma J, Li Y, Wang T, et al. Beyond a chemopreventive reagent, aspirin is a master regulator of the hallmarks of cancer. J Cancer Res Clin Oncol. 2019;145(6):1387-403.
48. Zhou D, Duan Z, Li Z, Ge F, Wei R, Kong L. The significance of glycolysis in tumor progression and its relationship with the tumor microenvironment. Front Pharmacol. 2022;13:1091779.
49. Chandel NS. Glycolysis. Cold Spring Harb Perspect Biol. 2021;13(5).
50. Tanner LB, Goglia AG, Wei MH, Sehgal T, Parsons LR, Park JO, et al. Four Key Steps Control Glycolytic Flux in Mammalian Cells. Cell Syst. 2018;7(1):49-62 e8.
51. Zhang ZJ, Zhang YH, Qin XJ, Wang YX, Fu J. Circular RNA circDENND4C facilitates proliferation, migration and glycolysis of colorectal cancer cells through miR-760/GLUT1 axis. Eur Rev Med Pharmacol Sci. 2020;24(5):2387-400.
52. Guo D, Tong Y, Jiang X, Meng Y, Jiang H, Du L, et al. Aerobic glycolysis promotes tumor immune evasion by hexokinase2-mediated phosphorylation of IkappaBalpha. Cell Metab. 2022;34(9):1312-24 e6.
53. Wu H, Pan L, Gao C, Xu H, Li Y, Zhang L, et al. Quercetin Inhibits the Proliferation of Glycolysis-Addicted HCC Cells by Reducing Hexokinase 2 and Akt-mTOR Pathway. Molecules. 2019;24(10).
54. Uludag D, Bay S, Sucu BO, Savlug Ipek O, Mohr T, Guzel M, et al. Potential of Novel Methyl Jasmonate Analogs as Anticancer Agents to Metabolically Target HK-2 Activity in Glioblastoma Cells. Front Pharmacol. 2022;13:828400.
55. Li S, He P, Wang Z, Liang M, Liao W, Huang Y, et al. RNAi-mediated knockdown of PFK1 decreases the invasive capability and metastasis of nasopharyngeal carcinoma cell line, CNE-2. Cell Cycle. 2021;20(2):154-65.
56. Sun Q, Chen X, Ma J, Peng H, Wang F, Zha X, et al. Mammalian target of rapamycin up-regulation of pyruvate kinase isoenzyme type M2 is critical for aerobic glycolysis and tumor growth. Proc Natl Acad Sci U S A. 2011;108(10):4129-34.
57. William Bateson GM. Mendel's Principles of Heredity - A Defence, with a Translation of Mendel's Original Papers on Hybridisation Cambridge University Press. 1902.
58. T. H. Morgan AHS, H. J. Muller, C. B. Bridges,. The Mechanism of Mendelian Heredity. Nature. 1916;97(2423):117-8.
59. J.D. Watson FHCC. Molecular Structure of Nucleic Acids. Nature. 1953;171:737–8.
60. Stehelin D, Varmus HE, Bishop JM, Vogt PK. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. Nature. 1976;260(5547):170-3.
61. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995;269(5223):496-512.
62. Parada LF, Tabin CJ, Shih C, Weinberg RA. Human EJ bladder carcinoma oncogene is homologue of Harvey sarcoma virus ras gene. Nature. 1982;297(5866):474-8.
63. Santos E, Tronick SR, Aaronson SA, Pulciani S, Barbacid M. T24 human bladder carcinoma oncogene is an activated form of the normal human homologue of BALB- and Harvey-MSV transforming genes. Nature. 1982;298(5872):343-7.
64. Klein G, Klein E. Evolution of tumours and the impact of molecular oncology. Nature. 1985;315(6016):190-5.
65. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177-83.
66. Martinez-Jimenez F, Muinos F, Sentis I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555-72.
67. Schubbert S, Shannon K, Bollag G. Hyperactive Ras in developmental disorders and cancer. Nat Rev Cancer. 2007;7(4):295-308.
68. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941-D7.

69. KRAS - COSMIC 2024 [cited 2024. Available from: https://cancer.sanger.ac.uk/cosmic/census-page/KRAS.
70. Koshland DE, Jr. Molecule of the year. Science. 1993;262(5142):1953.
71. Vogelstein B, Lane D, Levine AJ. Surfing the p53 network. Nature. 2000;408(6810):307-10.
72. TP53 - COSMIC 2024 [cited 2024. Available from: https://cancer.sanger.ac.uk/cosmic/census-page/TP53.
73. PTEN - COSMIC COSMIC2024 [cited 2024. Available from: https://cancer.sanger.ac.uk/cosmic/census-page/PTEN.
74. Cowling VH, Turner SA, Cole MD. Burkitt's lymphoma-associated c-Myc mutations converge on a dramatically altered target gene response and implicate Nol5a/Nop56 in oncogenesis. Oncogene. 2014;33(27):3519-27.
75. Schmitz R, Young RM, Ceribelli M, Jhavar S, Xiao W, Zhang M, et al. Burkitt lymphoma pathogenesis and therapeutic targets from structural and functional genomics. Nature. 2012;490(7418):116-20.
76. Cancer Genome Atlas Research N. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061-8.
77. Narod SA, Iqbal J, Miller AB. Why have breast cancer mortality rates declined? Journal of Cancer Policy. 2015;5:8-17.
78. Fisher B, Costantino JP, Wickerham DL, Redmond CK, Kavanah M, Cronin WM, et al. Tamoxifen for prevention of breast cancer: report of the National Surgical Adjuvant Breast and Bowel Project P-1 Study. J Natl Cancer Inst. 1998;90(18):1371-88.
79. Goss PE, Ingle JN, Ales-Martinez JE, Cheung AM, Chlebowski RT, Wactawski-Wende J, et al. Exemestane for breast-cancer prevention in postmenopausal women. N Engl J Med. 2011;364(25):2381-91.
80. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol. 2021;23(8):1231-51.
81. McKinnon C, Nandhabalan M, Murray SA, Plaha P. Glioblastoma: clinical presentation, diagnosis, and management. BMJ. 2021;374:n1560.
82. Larjavaara S, Mantyla R, Salminen T, Haapasalo H, Raitanen J, Jaaskelainen J, et al. Incidence of gliomas by anatomic location. Neuro Oncol. 2007;9(3):319-25.
83. Gallego Perez-Larraya J, Hildebrand J. Brain metastases. Handb Clin Neurol. 2014;121:1143-57.
84. Lapointe S, Perry A, Butowski NA. Primary brain tumours in adults. The Lancet. 2018;392(10145):432-46.
85. Aldape K, Zadeh G, Mansouri S, Reifenberger G, von Deimling A. Glioblastoma: pathology, molecular mechanisms and markers. Acta Neuropathol. 2015;129(6):829-48.
86. Davis FG, Smith TR, Gittleman HR, Ostrom QT, Kruchko C, Barnholtz-Sloan JS. Glioblastoma incidence rate trends in Canada and the United States compared with England, 1995-2015. Neuro Oncol. 2020;22(2):301-2.
87. Girardi F, Matz M, Stiller C, You H, Marcos Gragera R, Valkov MY, et al. Global survival trends for brain tumors, by histology: analysis of individual records for 556,237 adults diagnosed in 59 countries during 2000-2014 (CONCORD-3). Neuro Oncol. 2023;25(3):580-92.
88. Glioblastoma brain cancer mortality rate in the UK [Internet]. Office for National Statistics. 2022 [cited 30 August 2023]. Available from: https://www.ons.gov.uk/aboutus/transparencyandgovernance/freedomofinformationfoi/glioblastomabraincancermortalityrateintheuk.
89. Vigneswaran K, Neill S, Hadjipanayis CG. Beyond the World Health Organization grading of infiltrating gliomas: advances in the molecular genetics of glioma classification. Ann Transl Med. 2015;3(7):95.
90. Brain, other CNS and intracranial tumours survival statistics Cancer Research UK2024 [updated 2 December 2014; cited 2024 7 March]. Available from: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/brain-other-cns-and-intracranial-tumours/survival#heading-Zero.
91. Killela PJ, Reitman ZJ, Jiao Y, Bettegowda C, Agrawal N, Diaz LA, Jr., et al. TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. Proc Natl Acad Sci U S A. 2013;110(15):6021-6.
92. Komori T. Grading of adult diffuse gliomas according to the 2021 WHO Classification of Tumors of the Central Nervous System. Lab Invest. 2022;102(2):126-33.
93. Han S, Liu Y, Cai SJ, Qian M, Ding J, Larion M, et al. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. Br J Cancer. 2020;122(11):1580-9.
94. Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell. 2010;18(6):553-67.
95. Cancer Genome Atlas Research N, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, et al. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. N Engl J Med. 2015;372(26):2481-98.
96. National Cancer Registration and Analysis Service (NCRAS) Routes to diagnosis [Available from: http://www.ncin.org.uk/publications/routes_to_diagnosis.
97. Walter FM, Penfold C, Joannides A, Saji S, Johnson M, Watts C, et al. Missed opportunities for diagnosing brain tumours in primary care: a qualitative study of patient experiences. Br J Gen Pract. 2019;69(681):e224-e35.
98. HeadSmart Be Brain Tumour A. A new clinical guideline from the Royal College of Paediatrics and Child Health with a national awareness campaign accelerates brain tumor diagnosis in UK children--"HeadSmart: Be Brain Tumour Aware". Neuro Oncol. 2016;18(3):445-54.

99. Wen PY, Weller M, Lee EQ, Alexander BM, Barnholtz-Sloan JS, Barthel FP, et al. Glioblastoma in adults: a Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. Neuro Oncol. 2020;22(8):1073-113.
100. Schiff D, Lee EQ, Nayak L, Norden AD, Reardon DA, Wen PY. Medical management of brain tumors and the sequelae of treatment. Neuro Oncol. 2015;17(4):488-504.
101. Ozawa M, Brennan PM, Zienius K, Kurian KM, Hollingworth W, Weller D, et al. The usefulness of symptoms alone or combined for general practitioners in considering the diagnosis of a brain tumour: a case-control study using the clinical practice research database (CPRD) (2000-2014). BMJ Open. 2019;9(8):e029686.
102. Braganza MZ, Kitahara CM, Berrington de Gonzalez A, Inskip PD, Johnson KJ, Rajaraman P. Ionizing radiation and the risk of brain and central nervous system tumors: a systematic review. Neuro Oncol. 2012;14(11):1316-24.
103. Baan R, Grosse Y, Lauby-Secretan B, El Ghissassi F, Bouvard V, Benbrahim-Tallaa L, et al. Carcinogenicity of radiofrequency electromagnetic fields. Lancet Oncol. 2011;12(7):624-6.
104. Morgan LL, Miller AB, Sasco A, Davis DL. Mobile phone radiation causes brain tumors and should be classified as a probable human carcinogen (2A) (review). Int J Oncol. 2015;46(5):1865-71.
105. International Commission on the Biological Effects of Electromagnetic Fields (ICBE-EMF). Scientific evidence invalidates health assumptions underlying the FCC and ICNIRP exposure limit determinations for radiofrequency radiation: implications for 5G. Environ Health. 2022;21(92).
106. Task Group members. WHO Radiofrequency EMF Health Risk Assessment Monograph (EHC series): WHO; 2023 [Available from: https://www.saferemr.com/2021/09/who-radiofrequency-emf-health-risk.html.
107. Huang YR, Fan HQ, Kuang YY, Wang P, Lu S. The Relationship Between the Molecular Phenotypes of Brain Gliomas and the Imaging Features and Sensitivity of Radiotherapy and Chemotherapy. Clin Oncol (R Coll Radiol). 2024;36(9):541-51.
108. Fawcett JW. The extracellular matrix in plasticity and regeneration after CNS injury and neurodegenerative disease. Prog Brain Res. 2015;218:213-26.
109. Haugland HK, Tysnes BB, Tysnes OB. Adhesion and migration of human glioma cells are differently dependent on extracellular matrix molecules. Anticancer Res. 1997;17(2A):1035-42.
110. Gerritsen JKW, Broekman MLD, De Vleeschouwer S, Schucht P, Nahed BV, Berger MS, et al. Safe surgery for glioblastoma: Recent advances and modern challenges. Neurooncol Pract. 2022;9(5):364-79.
111. Khan F, Pang L, Dunterman M, Lesniak MS, Heimberger AB, Chen P. Macrophages and microglia in glioblastoma: heterogeneity, plasticity, and therapy. J Clin Invest. 2023;133(1).
112. Yang T, Kong Z, Ma W. PD-1/PD-L1 immune checkpoint inhibitors in glioblastoma: clinical studies, challenges and potential. Hum Vaccin Immunother. 2021;17(2):546-53.
113. Han J, Alvarez-Breckenridge CA, Wang QE, Yu J. TGF-beta signaling and its targeting for glioma treatment. Am J Cancer Res. 2015;5(3):945-55.
114. Peng Y, Croce CM. The role of MicroRNAs in human cancer. Signal Transduct Target Ther. 2016;1:15004.
115. Monteiro AR, Hill R, Pilkington GJ, Madureira PA. The Role of Hypoxia in Glioblastoma Invasion. Cells. 2017;6(4).
116. Macharia LW, Muriithi W, Heming CP, Nyaga DK, Aran V, Mureithi MW, et al. The genotypic and phenotypic impact of hypoxia microenvironment on glioblastoma cell lines. BMC Cancer. 2021;21(1):1248.
117. Mathieu J, Zhang Z, Zhou W, Wang AJ, Heddleston JM, Pinna CM, et al. HIF induces human embryonic stem cell markers in cancer cells. Cancer Res. 2011;71(13):4640-52.
118. Musatova OE, Rubtsov YP. Effects of glioblastoma-derived extracellular vesicles on the functions of immune cells. Front Cell Dev Biol. 2023;11:1060000.
119. Wang J, Xu F, Zhu X, Li X, Li Y, Li J. Targeting microRNAs to Regulate the Integrity of the Blood-Brain Barrier. Front Bioeng Biotechnol. 2021;9:673415.
120. Wu W, Klockow JL, Zhang M, Lafortune F, Chang E, Jin L, et al. Glioblastoma multiforme (GBM): An overview of current therapies and mechanisms of resistance. Pharmacol Res. 2021;171:105780.
121. Maggs L, Cattaneo G, Dal AE, Moghaddam AS, Ferrone S. CAR T Cell-Based Immunotherapy for the Treatment of Glioblastoma. Front Neurosci. 2021;15:662064.
122. Weller M, van den Bent M, Preusser M, Le Rhun E, Tonn JC, Minniti G, et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. Nat Rev Clin Oncol. 2021;18(3):170-86.
123. Prager BC, Bhargava S, Mahadev V, Hubert CG, Rich JN. Glioblastoma Stem Cells: Driving Resilience through Chaos. Trends Cancer. 2020;6(3):223-35.
124. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321(5897):1807-12.
125. Nobusawa S, Watanabe T, Kleihues P, Ohgaki H. IDH1 mutations as molecular signature and predictive factor of secondary glioblastomas. Clin Cancer Res. 2009;15(19):6002-7.
126. Kloosterhof NK, Bralten LB, Dubbink HJ, French PJ, van den Bent MJ. Isocitrate dehydrogenase-1 mutations: a fundamentally new understanding of diffuse glioma? Lancet Oncol. 2011;12(1):83-91.
127. Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell. 2017;32(1):42-56 e6.
128. Tang Q, Li L, Wang Y, Wu P, Hou X, Ouyang J, et al. RNA modifications in cancer. Br J Cancer. 2023;129(2):204-21.
129. Brenner S, Jacob F, Meselson M. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. Nature. 1961;190:576-81.

130. Gros F, Hiatt H, Gilbert W, Kurland CG, Risebrough RW, Watson JD. Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli. Nature. 1961;190:581-5.
131. Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, Martin Raff, Keith Roberts, et al. Molecular Biology of the Cell. Sixth Edition ed. New York and Abingdon, UK: Garland Science 2014. 1464 p.
132. Boeckel JN, Jae N, Heumuller AW, Chen W, Boon RA, Stellos K, et al. Identification and Characterization of Hypoxia-Regulated Endothelial Circular RNA. Circ Res. 2015;117(10):884-90.
133. Broughton JP, Lovci MT, Huang JL, Yeo GW, Pasquinelli AE. Pairing beyond the Seed Supports MicroRNA Targeting Specificity. Mol Cell. 2016;64(2):320-33.
134. Lee RC FR, Ambros V,. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75:843–54.
135. the 2024 Nobel Prize in Physiology or Medicine [press release]. 07/10/2024 2024.
136. Damase TR, Sukhovershin R, Boada C, Taraballi F, Pettigrew RI, Cooke JP. The Limitless Future of RNA Therapeutics. Front Bioeng Biotechnol. 2021;9:628137.
137. Tufekci KU, Oner MG, Meuwissen RL, Genc S. The role of microRNAs in human diseases. Methods Mol Biol. 2014;1107:33-50.
138. Paul P, Chakraborty A, Sarkar D, Langthasa M, Rahman M, Bari M, et al. Interplay between miRNAs and human diseases. J Cell Physiol. 2018;233(3):2007-18.
139. Wang J, Chen J, Sen S. MicroRNA as Biomarkers and Diagnostics. J Cell Physiol. 2016;231(1):25-30.
140. Huang W. MicroRNAs: Biomarkers, Diagnostics, and Therapeutics. Methods Mol Biol. 2017;1617:57-67.
141. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116(2):281-97.
142. Jones CH, Androsavich JR, So N, Jenkins MP, MacCormack D, Prigodich A, et al. Breaking the mold with RNA-a "RNAissance" of life science. NPJ Genom Med. 2024;9(1):2.
143. Xie X, Yu T, Li X, Zhang N, Foster LJ, Peng C, et al. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. Signal Transduct Target Ther. 2023;8(1):335.
144. Pardi N, Hogan MJ, Porter FW, Weissman D. mRNA vaccines - a new era in vaccinology. Nat Rev Drug Discov. 2018;17(4):261-79.
145. He Q, Gao H, Tan D, Zhang H, Wang JZ. mRNA cancer vaccines: Advances, trends and challenges. Acta Pharm Sin B. 2022;12(7):2969-89.
146. Magee P, Shi L, Garofalo M. Role of microRNAs in chemoresistance. Ann Transl Med. 2015;3(21):332.
147. Pal AS, Kasinski AL. Animal Models to Study MicroRNA Function. Adv Cancer Res. 2017;135:53-118.
148. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of MicroRNA Biogenesis, Mechanisms of Actions, and Circulation. Front Endocrinol (Lausanne). 2018;9:402.
149. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39(Database issue):D152-7.
150. Shang R, Lee S, Senavirathne G, Lai EC. microRNAs in action: biogenesis, function and regulation. Nature Reviews Genetics. 2023.
151. Richard I. Gregory K-pY, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch & Ramin Shiekhattar. The Microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432:235–40.
152. Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 2003;17(24):3011-6.
153. miRagen Therapeutics Inc. SOLAR: Efficacy and Safety of Cobomarsen (MRG-106) vs. Active Comparator in Subjects With Mycosis Fungoides (SOLAR). ClinicalTrialsgov2019-2020.
154. Foundation ADR. MesomiR 1: A Phase I Study of TargomiRs as 2nd or 3rd Line Treatment for Patients With Recurrent MPM and NSCLC. ClinicalTrialgov2014-2017.
155. Mirna Therapeutics Inc. A Multicenter Phase I Study of MRX34, MicroRNA miR-RX34 Liposomal Injection. ClinicalTrialsgov2013-2017.
156. TransCode Therapeutics. Study of TTX-MC138 in Subjects With Advanced Solid Tumors. ClinicalTrialsgov2024-2027(estimated).
157. InteRNA. First-in-Human Study of INT-1B3 in Patients With Advanced Solid Tumors. ClinicalTrialsgov2020-2023.
158. Gao H, Yang Z, Zhang S, Cao S, Shen S, Pang Z, et al. Ligand modified nanoparticles increases cell uptake, alters endocytosis and elevates glioma distribution and internalization. Sci Rep. 2013;3:2534.
159. QIAGEN CLC Genomics Workbench [Available from: https://digitalinsights.qiagen.com.
160. Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, et al. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature. 1976;260(5551):500-7.
161. Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, et al. The complete DNA sequence of yeast chromosome III. Nature. 1992;357(6373):38-46.
162. Consortium CeS. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282(5396):2012-8.
163. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, et al. The DNA sequence of human chromosome 22. Nature. 1999;402(6761):489-95.
164. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science. 2000;287(5461):2185-95.
165. Arabidopsis Genome I. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000;408(6814):796-815.

166. Mouse Genome Sequencing C, Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420(6915):520-62.
167. International Human Genome Sequencing C. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
168. Podnar J, Deiderick H, Hunicke-Smith S. Next-generation sequencing fragment library construction. Curr Protoc Mol Biol. 2014;107:7 17 1-7 6.
169. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
170. Hrdlickova R, Toloue M, Tian B. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA. 2017;8(1).
171. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. The impact of amplification on differential expression analyses by RNA-seq. Sci Rep. 2016;6:25533.
172. Chia-Hsin Liu YPD. Analysis of RNA Sequencing Data Using CLC Genomics Workbench. In: Keohavong, P., Singh, K., Gao, W. (eds) Molecular Toxicology Protocols. Methods in Molecular Biology. 2020;2102.
173. Wratten L, Wilm A, Goke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. 2021;18(10):1161-8.
174. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316-9.
175. Nextflow documentation 2023 [cited 2023 29 October]. v23.10.0]. Available from: https://www.nextflow.io/docs/latest/index.html.
176. Scott Chacon BS. Pro git: Apress; 2014.
177. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux journal,. 2014;2014(239):2.
178. Anaconda Inc. Anaconda Software Distribution. Anaconda Documentation,. 2020.
179. Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276-8.
180. Harshil Patel PE, Alexander Peltzer, Olga Botvinnik, Gregor Sturm, Denis Moreno, Pranathi Vemuri, silviamorins, Lorena Pantano, Mahesh Binzer-Panchal, nf-core bot, Gavin Kelly, Maxime U. Garcia, FriederikeHanssen, Matthias Zepper, James A. Fellows Yates, Chris Cheshire, rfenouil, Jose Espinosa-Carrasco, marchoeppner, Edmund Miller, Peng Zhou, Sarah Guinchard, Matthias Hörtenhuber, Gisela Gabernet, Christian Mertes, Daniel Straub, Paolo Di Tommaso, Sven F., George Hall. nf-core/rnaseq: nf-core/rnaseq v3.10.1 - Plastered Rhodium Rudolph. 3.10.1 ed: Zenodo; 2023.
181. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139-40.
182. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
183. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106.
184. Yi X, Zhang Z, Ling Y, Xu W, Su Z. PNRD: a plant non-coding RNA database. Nucleic Acids Res. 2015;43(Database issue):D982-9.
185. Edinburgh Compute and Data Facility University of Edinburgh2019 [cited 2023 11 September]. Available from: www.ecdf.ed.ac.uk.
186. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
187. Martin Morgan VO, Jim Hester, Hervé Pagès. SummarizedExperiment: SummarizedExperiment container. 2023.
188. Love MI, Soneson C, Hickey PF, Johnson LK, Pierce NT, Shepherd L, et al. Tximeta: Reference sequence checksums for provenance identification in RNA-seq. PLoS Comput Biol. 2020;16(2):e1007664.
189. Sayols S, Scherzinger D, Klein H. dupRadar: a Bioconductor package for the assessment of PCR artifacts in RNA-Seq data. BMC Bioinformatics. 2016;17(1):428.
190. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data 2010 [Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
191. Picard Toolkit. Broad Institute, GitHub repository: Broad Institute; 2019.
192. Daley T, Smith AD. Predicting the molecular complexity of sequencing libraries. Nat Methods. 2013;10(4):325-7.
193. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.
194. Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012;28(16):2184-5.
195. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14(4):417-9.
196. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).
197. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15-21.
198. Shumate A, Wong B, Pertea G, Pertea M. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLoS Comput Biol. 2022;18(6):e1009730.

199. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30(7):923-30.
200. Krueger F. Trim Galore! Babraham Institute, Github Repository: Babraham Institute; 2012.
201. Nassar LR, Barber GP, Benet-Pages A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. 2023;51(D1):D1188-D95.
202. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS One. 2017;12(5):e0177459.
203. Michael I. Love SA, and Wolfgang Huber. Analyzing RNA-seq data with DESeq2 Bioconductor2023 [cited 2023 21 September]. Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html.
204. RStudio Team. RStudio: Integrated Development Environment for R. RStudio, PBC.; 2020.
205. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686.
206. Hadley Wickham RF, Lionel Henry, Kirill Müller, Davis Vaughan. dplyr: A Grammar of Data Manipulation. 2023.
207. Steffen Durinck YM, Arek Kasprzyk , Sean Davis , Bart (De Moor), Alvis Brazma, Wolfgang Huber. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21:3439--40.
208. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4.
209. Gregory R. Warnes BB, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber, Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz, Bill Venables. gplots: Various R Programming Tools for Plotting Data 2022 [Available from: https://CRAN.R-project.org/package=gplots.
210. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.14.0 ed2021.
211. Kevin Blighe SR, Myles Lewis. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and

labeling. R package version 1.16.0 ed2022.
212. Zuguang Gu RE, Matthias Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 2016.
213. Farina F, Gentner B, Finocchiaro G, Eoli M, Capotondo A, Anghileri E, et al. Genetic Engineering of Hematopoietic Progenitor Stem Cells for Targeted IFN-α Immunotherapy Reprogramming the Solid Tumor Microenvironment: A First-in-Man Study in Glioblastoma Multiforme (NCT03866109). Blood. 2023;142(Supplement 1):4850-.
214. Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, et al. Variant calling and benchmarking in an era of complete human genome sequences. Nat Rev Genet. 2023;24(7):464-83.
215. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91.
216. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform. 2015;16(1):59-70.
217. Michael I. Love SA, and Wolfgang Huber. Analyzing RNA-seq data with DESeq2 Bioconductor2023 [Available from: https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html.
218. Kevin Blighe SR, Myles Lewis,. EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling. 2018.
219. Huang M, Qin X, Wang Y, Mao F. Identification of AK4 as a novel therapeutic target for serous ovarian cancer. Oncol Lett. 2020;20(6):346.
220. Feng Y, Xiong Y, Qiao T, Li X, Jia L, Han Y. Lactate dehydrogenase A: A key player in carcinogenesis and potential target in cancer therapy. Cancer Med. 2018;7(12):6124-36.
221. Jawhari S, Ratinaud MH, Verdier M. Glioblastoma, hypoxia and autophagy: a survival-prone 'menage-a-trois'. Cell Death Dis. 2016;7(10):e2434.
222. Zhang L, Cao Y, Guo X, Wang X, Han X, Kanwore K, et al. Hypoxia-induced ROS aggravate tumor progression through HIF-1alpha-SERPINE1 signaling in glioblastoma. J Zhejiang Univ Sci B. 2023;24(1):32-49.
223. Semenza GL. Hypoxia-inducible factors in physiology and medicine. Cell. 2012;148(3):399-408.
224. Kaelin WG, Jr., Ratcliffe PJ. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. Mol Cell. 2008;30(4):393-402.
225. Bigham AW, Lee FS. Human high-altitude adaptation: forward genetics meets the HIF pathway. Genes Dev. 2014;28(20):2189-204.
226. Minamishima YA, Kaelin WG, Jr. Reactivation of hepatic EPO synthesis in mice after PHD loss. Science. 2010;329(5990):407.
227. [Available from: https://www.ncbi.nlm.nih.gov/tools/primer-blast/.
228. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13:134.
229. Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9(7):671-5.
230. Adams G. A beginner's guide to RT-PCR, qPCR and RT-qPCR. The Biochemist. 2020;42(3):48-53.
231. Nakahara Y, Ito H, Namikawa H, Furukawa T, Yoshioka F, Ogata A, et al. A Tumor Suppressor Gene, N-myc Downstream-Regulated Gene 1 (NDRG1), in Gliomas and Glioblastomas. Brain Sci. 2022;12(4).

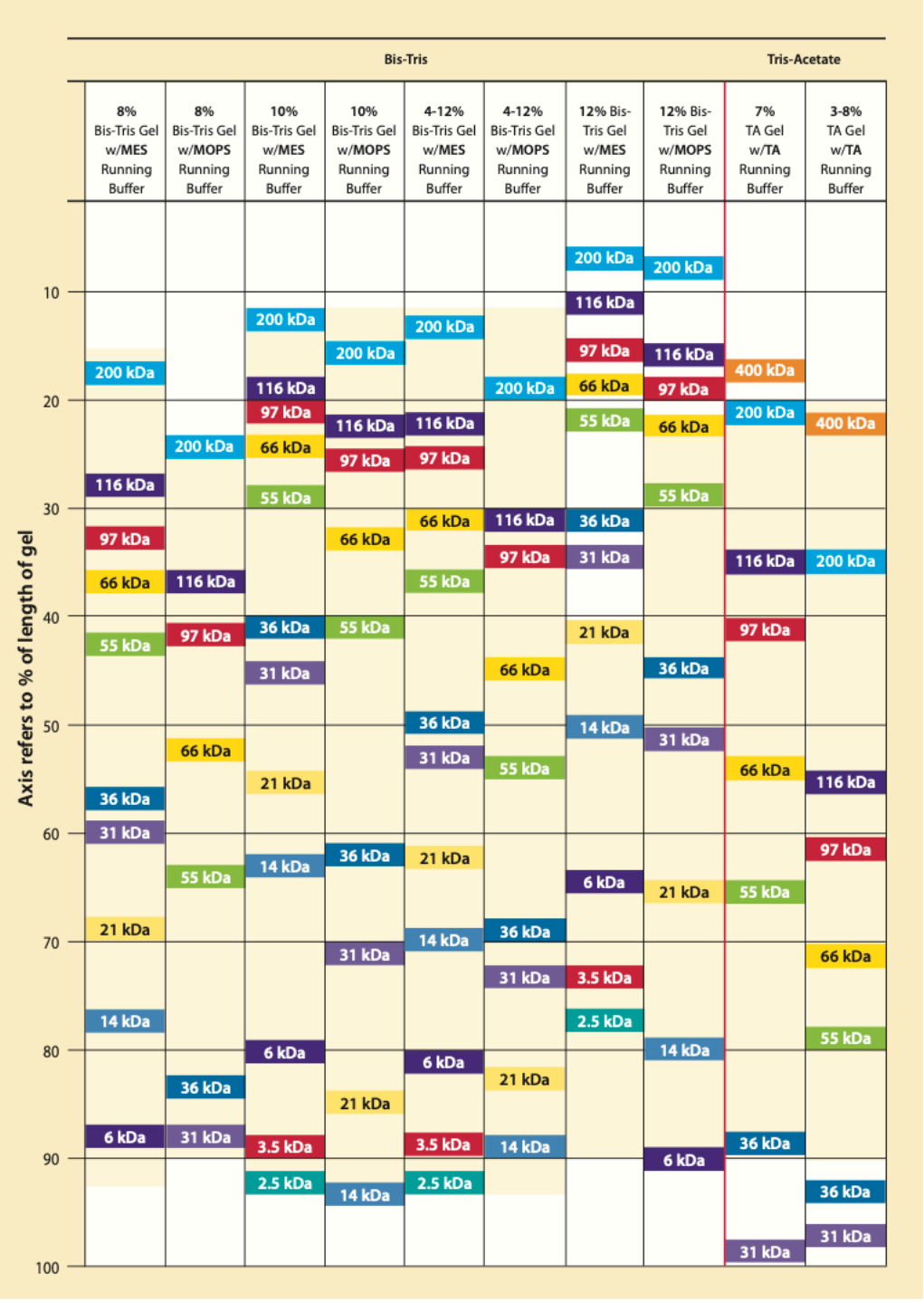
232. Kovacevic Z, Richardson DR. The metastasis suppressor, Ndrg-1: a new ally in the fight against cancer. Carcinogenesis. 2006;27(12):2355-66.
233. Said HM, Safari R, Al-Kafaji G, Ernestus RI, Lohr M, Katzer A, et al. Time- and oxygen-dependent expression and regulation of NDRG1 in human brain cancer cells. Oncol Rep. 2017;37(6):3625-34.
234. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res. 2015;4:1521.
235. CHIMENTI MS. MICHAEL'S BIOINFORMATICS BLOG

A blog about genomics, data science, and analysis [Internet]2017. [cited 2024]. Available from: https://www.michaelchimenti.com/2017/02/beyond-benjamini-hochberg-multiple-test-correction-with-independent-hypothesis-weighting-ihw/.
236. Ignatiadis N, Klaus B, Zaugg JB, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 2016;13(7):577-80.
237. Breheny P, Stromberg A, Lambert J. p-Value Histograms: Inference and Diagnostics. High Throughput. 2018;7(3).
238. Guyon J, Fernandez-Moncada I, Larrieu CM, Bouchez CL, Pagano Zottola AC, Galvis J, et al. Lactate dehydrogenases promote glioblastoma growth and invasion via a metabolic symbiosis. EMBO Mol Med. 2022;14(12):e15343.
239. Modrek AS, Eskilsson E, Ezhilarasan R, Wang Q, Goodman LD, Ding Y, et al. PDPN marks a subset of aggressive and radiation-resistant glioblastoma cells. Front Oncol. 2022;12:941657.
240. complement C6 NIH [Available from: https://www.ncbi.nlm.nih.gov/gene/729#summary.
241. FCER2 Fc epsilon receptor II NIH; [Available from: https://www.ncbi.nlm.nih.gov/gene/2208#summary.
242. Morell P QR. The Myelin Sheath. In: Siegel GJ, Agranoff BW, Albers RW, et al. Philadelphia: Lippincott-Raven1999 [6th:[Available from: https://www.ncbi.nlm.nih.gov/books/NBK27954/.
243. Li H, Wang Z, Sun C, Li S. Establishment of a cell senescence related prognostic model for predicting prognosis in glioblastoma. Front Pharmacol. 2022;13:1034794.
244. Cheng YX, Xiao L, Yang YL, Liu XD, Zhou XR, Bu ZF, et al. Collagen type VIII alpha 2 chain (COL8A2), an important component of the basement membrane of the corneal endothelium, facilitates the malignant development of glioblastoma cells via inducing EMT. J Bioenerg Biomembr. 2021;53(1):49-59.
245. Western blot troubleshooting tips [Website Help Page]. abcam; 2023 [cited 2023 1 August]. Available from: https://www.abcam.com/help/western-blot-troubleshooting-tips#unexpected-or-multiple-bands.
246. Gurgel A. Re: Why do i see multiple bands in my western? ResearchGate2015 [cited 2023 1 August]. Available from: https://www.researchgate.net/post/Why-do-i-see-multiple-bands-in-my-western.
247. Shah N, Sukumar S. The Hox genes and their roles in oncogenesis. Nat Rev Cancer. 2010;10(5):361-71.
248. Alharbi RA, Pettengell R, Pandha HS, Morgan R. The role of HOX genes in normal hematopoiesis and acute leukemia. Leukemia. 2013;27(5):1000-8.
249. Li B, Huang Q, Wei GH. The Role of HOX Transcription Factors in Cancer Predisposition and Progression. Cancers (Basel). 2019;11(4).
250. Rux DR, Wellik DM. Hox genes in the adult skeleton: Novel functions beyond embryonic development. Dev Dyn. 2017;246(4):310-7.
251. Fang J, Wang J, Yu L, Xu W. Role of HOXC10 in Cancer. Front Oncol. 2021;11:684021.
252. Tan Z, Chen K, Wu W, Zhou Y, Zhu J, Wu G, et al. Overexpression of HOXC10 promotes angiogenesis in human glioma via interaction with PRMT5 and upregulation of VEGFA expression. Theranostics. 2018;8(18):5143-58.
253. Miao J, Wang Z, Provencher H, Muir B, Dahiya S, Carney E, et al. HOXB13 promotes ovarian cancer progression. Proc Natl Acad Sci U S A. 2007;104(43):17093-8.
254. Bhatlekar S, Fields JZ, Boman BM. HOX genes and their role in the development of human cancers. J Mol Med (Berl). 2014;92(8):811-23.
255. Costa BM, Smith JS, Chen Y, Chen J, Phillips HS, Aldape KD, et al. Reversing HOXA9 oncogene activation by PI3K inhibition: epigenetic mechanism and prognostic significance in human glioblastoma. Cancer Res. 2010;70(2):453-62.
256. Xu K, Qiu C, Pei H, Mehmood M, Wang H, Li L, et al. Homeobox B3 promotes tumor cell proliferation and invasion in glioblastoma. Oncology Letters. 2018.
257. Zhu S, Yang Z, Zhang Z, Zhang H, Li S, Wu T, et al. HOXB3 drives WNT-activation associated progression in castration-resistant prostate cancer. Cell Death Dis. 2023;14(3):215.
258. Eryi S, Zheng L, Honghua C, Su Z, Han X, Donggang P, et al. HOXC6 Regulates the Epithelial-Mesenchymal Transition through the TGF-beta/Smad Signaling Pathway and Predicts a Poor Prognosis in Glioblastoma. J Oncol. 2022;2022:8016102.
259. Li Z, Huang H, Li Y, Jiang X, Chen P, Arnovitz S, et al. Up-regulation of a HOXA-PBX3 homeobox-gene signature following down-regulation of miR-181 is associated with adverse prognosis in patients with cytogenetically abnormal AML. Blood. 2012;119(10):2314-24.
260. Arunachalam E, Rogers W, Simpson GR, Moller-Levet C, Bolton G, Ismael M, et al. HOX and PBX gene dysregulation as a therapeutic target in glioblastoma multiforme. BMC Cancer. 2022;22(1):400.
261. Daniels TR, Neacato, II, Rodriguez JA, Pandha HS, Morgan R, Penichet ML. Disruption of HOX activity leads to cell death that can be enhanced by the interference of iron uptake in malignant B cells. Leukemia. 2010;24(9):1555-65.

262. Morita M, Sato T, Nomura M, Sakamoto Y, Inoue Y, Tanaka R, et al. PKM1 Confers Metabolic Advantages and Promotes Cell-Autonomous Tumor Cell Growth. Cancer Cell. 2018;33(3):355-67 e7.
263. Menezes SV, Sahni S, Kovacevic Z, Richardson DR. Interplay of the iron-regulated metastasis suppressor NDRG1 with epidermal growth factor receptor (EGFR) and oncogenic signaling. J Biol Chem. 2017;292(31):12772-82.
264. Bielefeld P, Mooney C, Henshall DC, Fitzsimons CP. miRNA-Mediated Regulation of Adult Hippocampal Neurogenesis; Implications for Epilepsy. Brain Plast. 2017;3(1):43-59.
265. Wareham LK, Baratta RO, Del Buono BJ, Schlumpf E, Calkins DJ. Collagen in the central nervous system: contributions to neurodegeneration and promise as a therapeutic target. Mol Neurodegener. 2024;19(1):11.
266. Mehta AM, Sonabend AM, Bruce JN. Convection-Enhanced Delivery. Neurotherapeutics. 2017;14(2):358-71.
267. Gilkes DM, Semenza GL, Wirtz D. Hypoxia and the extracellular matrix: drivers of tumour metastasis. Nat Rev Cancer. 2014;14(6):430-9.
268. Weiler M, Blaes J, Pusch S, Sahm F, Czabanka M, Luger S, et al. mTOR target NDRG1 confers MGMT-dependent resistance to alkylating chemotherapy. Proc Natl Acad Sci U S A. 2014;111(1):409-14.
269. COL6A3 THE HUMAN PROTEIN ATLAS [Available from: https://www.proteinatlas.org/ENSG00000163359-COL6A3/cancer.

Appendices

Appendix 1 - Related to CHAPTER 2



Appendix Table 1: Migration patterns of protein standards on NuPAGE Novex Gels from ThermoFisher Scientific (https://www.thermofisher.com/order/catalog/product/NP0001

INSTITUTE OF GENETICS & CANCER

IGC Standard Operating Procedure (SOP) Form (V1)

Any Procedure MUST be accompanied by a relevant Risk Assessment:
https://www.ed.ac.uk/health-safety/biosafety/forms/risk-assessments

Ref: 157

IGC-SOP037

This document is written to supplement risk assessments (RAs) for activities being carried out in the facility/Lab and will outline how the task/method is to be undertaken in a safe manner. Users must also read, understand, and acknowledge all RAs, relevant to the work to be carried out, in Q-Pulse.

A. Title:	Microtomy of Paraffin Embedded Tissues		
B. Author:	Carrie Cunningham	**C. Location of Activity:**	Lab 1.02
D. Revision No:	1	**E. Date:**	3rd October 2023

F. Method *(Clear step by step details of the procedure is required in this section, including preparation of working space and returning the working space to a safe condition).*

1. **Please refer to risk assessments IGMM 112 and IGMM 071 before using this piece of equipment**

2. **Ice Plates**

   Blocks should be placed on an ice plate to cool before sectioning. Plastic containers filled with water and stored frozen in the freezer are used for this.

   After use these ice trays should be wiped clean with some paper towel, refilled and placed back into the freezer

3. **Water Baths**

   The water bath is filled with water. Check the temperature dial is at correct setting (it should be 35-450C). After use, the water is poured out into the lab sink and the inside dried with paper towel.

4. **Disposable blades**

   The blades should be carefully inserted into the microtome and the clamp holding it tightened, only clean the blade with a brush to remove trimmings during use to avoid injury. **ALWAYS** remove the final blade before cleaning down the microtome with histoclear.

5. **Microtomes**

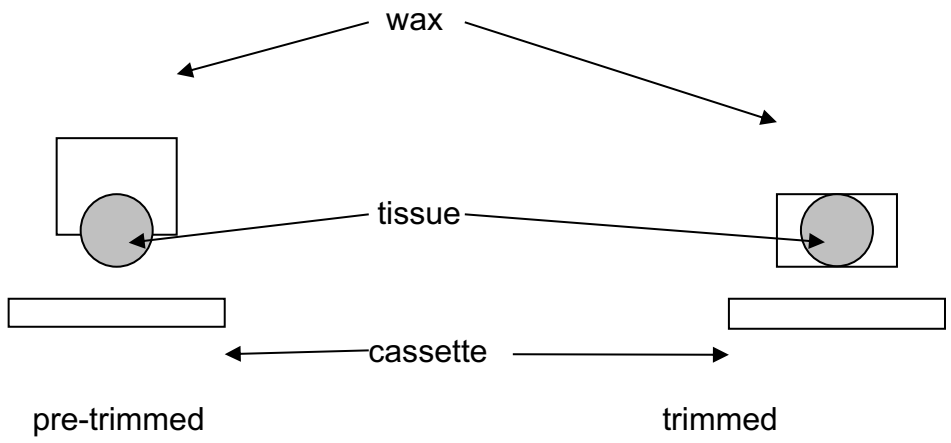
   The lab has 6 Leica RM2235 microtomes. They are bench mounted and adapted to use disposable blades.

6. **Block trimming-** before sectioning can begin excess wax has to be taken from the surface of the cutting face. Make sure the safety catch is on (main handle on the right of the microtome is locked in) and place the block on the chuck/cassette holder tightly (front of microtome wax side facing out towards the blade).

7. Position the block close behind the blade edge using the adjuster handle on the left hand side of the microtome and then the one on the main right one once unlocked.

8. Using the dial on the front right hand side select the thickness of the section required.

9. Continue rotating the microtome handle until the surface wax has been removed and the full face of the tissue can be seen, repeat for each block.



10. **Sectioning-** Once trimmed, lock the main handle and remove the block from the chuck and place back on the ice until needed.

11. Move the blade along to a new area, previously unused and retighten the clamp holding it in place.

12. Reattach the block to the chuck and realign the block to the blade edge. Release the lock on the main handle and rotate until a full face section in achieved.

13. Dispose of any unwanted/curled/broken sections with a brush or forceps and keep rotating the handle in a steady continuous motion to cut a section or ribbon of sections.

14. Remove the desired sections with a brush or forceps carefully from the blade, whilst avoiding the blade edge and float the section using a gentle lying motion across the surface of the water bath.

15. The waters heat will cause the section to expand slightly and flatten out.

16. Inspect the section on the water with the surface of the block to check all the tissue areas are accounted for, if not the block will need to be trimmed in a little more.

17. Any small folds can be gently teased out with the brush or forceps, larger folds or tears in sections cannot be fixed and therefore those sections should be discarded.

18. **Picking Up Sections –** once a satisfactory section has been obtained it must be picked up onto the appropriately labelled slide. To pick up a section hold the slide at the top and immerse it in the water close to the section floating.

19. Slope the slide away from the section and then advance the slide gently until it is under the

section.

20. Lift the slide slowly out of the water underneath the section and it will adhere to the slide as it is lifted from the water bath. Make sure the section is picked up on the correct side of the slide and is placed centrally.

21. Once out of the water lay the slide on the rim of the bath to dry or prop up on the side of it to allow the excess water to come off.

22. Place the slide into a staining rack and repeat with the process again as necessary with any remaining blocks/slides.

23. Put slides in either a 37oC oven or a 55oC oven to dry for several hours/overnight. **ONLY** use a 37oC oven if immuno-staining the slides after.

24. **Problem solving- Surface Decalcification**

    Occasionally, blocks are received which have small areas of calcification that will not section.

    To enable a section to be taken from these blocks they should be placed in a jar of surface decalcifying agent for a period of half an hour or more depending on the amount of calcium present. This will decalcify a small amount of calcium close to the cutting surface of the block.

    Do not leave blocks in surface decalcifying agent for long periods of time as this can cause tissue damage.

25. **Tissue Softener**

    Some tissue blocks are composed of hard dense material that does not section well.

    If this problem is encountered then the block can be placed into tissue softening solution (Mollifex™) for a few hours.

    Blocks should not be left in this solution for long periods of time as this can cause tissue damage.

26. **Embedding Problems**

    If a block is found to be too thin, or if when sectioning the block becomes loose, or the wax cracks, the block should be taken for re-embedding before a section is cut.

G. Waste *(For example, what waste will be generated and how will the waste be collected/disposed of)*

The excess wax cut offs should be emptied into the yellow waste stream after every use.

Biological waste (fixed tissue) via yellow waste stream.

All blue roll used cleaning up water spills/baths/ice trays/histoclear disposed of into yellow waste stream.

Used Microtome blades are collected into the bottom of microtome blade dispenser and when filled with all "used" blades (no new ones coming out from the top) dispose of dispenser into sharps bins:-

Do not place the sharps bin on the floor in the lab.

Do not fill the sharps bin past the recommended fill line.

Contact tech support for the collection of sharps bins for safe disposal in incinerator.

H. Emergency Procedures (*Information to include in this SOP should already be detailed in the risk assessment for the work activity, please copy and paste the relevant information here.)*

In case of injury, dial 88 (Reception) to summon a First Aider. Advise Local Safety Coordinator and Health and Safety Manager.

All accidents, incidents and near misses must be reported using the online form:

Accident / incident reporting | The University of Edinburgh

If any shortcomings, deviations, additional hazards and/or risks are identified associated with this work activity, please contact the Health and Safety Team immediately healthsafety@igc.ed.ac.uk

Appendix Table 2: IGC Standard Operation Procedure (SOP) Form (V1), Microtomy of Paraffin Embedded Tissues

Step N°	Reagent	Supplier	Inc. (min)
1	Peroxide Block	Leica Microsystems	5:00
5	MARKER	Leica Microsystems	15:00
9	Post Primary	Leica Microsystems	8:00
13	Polymer	Leica Microsystems	8:00
17	Mixed DAB Refine	Leica Microsystems	0:00
18	Mixed DAB Refine	Leica Microsystems	10:00
22	Hametoxylin	Leica Microsystems	5:00

Appendix Table 3: Leica Microsystems BOND-III Bond Polymer Refine IHC Protocol F

Step N°	Reagent	Supplier	Inc. (min)
1	Peroxide Block	Leica Microsystems	5:00
5	Biotin Block 1	Leica Microsystems	20:00
9	MARKER	Leica Microsystems	30:00
13	Polymer	Leica Microsystems	10:00
17	Mixed DAB Refine	Leica Microsystems	0:00
18	Mixed DAB Refine	Leica Microsystems	10:00
22	Hametoxylin	Leica Microsystems	5:00

Appendix Table 4: Leica Microsystems BOND-III Bond Polymer Refine IHC Protocol MODIFIED F

sample	fastq_1	fastq_2	strandedness	condition	cell_line
322_hypox1	4_GBM322hypox1_1.fastq.gz	4_GBM322hypox1_2.fastq.gz	reverse	hypox	322
322_hypox2	5_GBM322hypox2_1.fastq.gz	5_GBM322hypox2_2.fastq.gz	reverse	hypox	322
322_hypox3	6_GBM322hypox3_1.fastq.gz	6_GBM322hypox3_2.fastq.gz	reverse	hypox	322
322_NC1	10_GBM322NC1_1.fastq.gz	10_GBM322NC1_2.fastq.gz	reverse	normal	322
322_NC2	11_GBM322NC2_1.fastq.gz	11_GBM322NC2_2.fastq.gz	reverse	normal	322
322_NC3	12_GBM322NC3_1.fastq.gz	12_GBM322NC3_2.fastq.gz	reverse	normal	322
327_hypox1	1_GBM327hypox1_1.fastq.gz	1_GBM327hypox1_2.fastq.gz	reverse	hypox	327
327_hypox2	2_GBM327hypox2_1.fastq.gz	2_GBM327hypox2_2.fastq.gz	reverse	hypox	327
327_hypox3	3_GBM327hypox3_1.fastq.gz	3_GBM327hypox3_2.fastq.gz	reverse	hypox	327
327_NC1	7_GBM327NC1_1.fastq.gz	7_GBM327NC1_2.fastq.gz	reverse	normal	327
327_NC2	8_GBM327NC2_1.fastq.gz	8_GBM327NC2_2.fastq.gz	reverse	normal	327
327_NC3	9_GBM327NC3_1.fastq.gz	9_GBM327NC3_2.fastq.gz	reverse	normal	327

Appendix Table 5: Samplesheet.csv for Nextflow nf-core/RNAseq analysis containing sample name, forward and reverse files (fastq_1, fastq_2), strandedness of sequencing, the condition that sample belongs to and the cell line the sample is from.

sample	fastq_1	fastq_2	strandedness	condition	cell_line
322_hypox1	4_GBM322hypox1_1.fastq.gz	4_GBM322hypox1_2.fastq.gz	reverse	hypox	322
322_hypox2	5_GBM322hypox2_1.fastq.gz	5_GBM322hypox2_2.fastq.gz	reverse	hypox	322
322_hypox3	6_GBM322hypox3_1.fastq.gz	6_GBM322hypox3_2.fastq.gz	reverse	hypox	322
322_NC1	10_GBM322NC1_1.fastq.gz	10_GBM322NC1_2.fastq.gz	reverse	normal	322
322_NC2	11_GBM322NC2_1.fastq.gz	11_GBM322NC2_2.fastq.gz	reverse	normal	322
322_NC3	12_GBM322NC3_1.fastq.gz	12_GBM322NC3_2.fastq.gz	reverse	normal	322
327_hypox1	1_GBM327hypox1_1.fastq.gz	1_GBM327hypox1_2.fastq.gz	reverse	hypox	327
327_hypox2	2_GBM327hypox2_1.fastq.gz	2_GBM327hypox2_2.fastq.gz	reverse	hypox	327
327_hypox3	3_GBM327hypox3_1.fastq.gz	3_GBM327hypox3_2.fastq.gz	reverse	hypox	327
327_NC1	7_GBM327NC1_1.fastq.gz	7_GBM327NC1_2.fastq.gz	reverse	normal	327
327_NC2	8_GBM327NC2_1.fastq.gz	8_GBM327NC2_2.fastq.gz	reverse	normal	327
327_NC3	9_GBM327NC3_1.fastq.gz	9_GBM327NC3_2.fastq.gz	reverse	normal	327

Appendix Table 6: Samplesheet.csv for Nextflow nf-core/RNAseq analysis containing sample name, forward and reverse files (fastq_1, fastq_2), strandedness of sequencing, the condition that sample belongs to and the cell line the sample is from.

Full RStudio server version: RStudio Server 2022.02.0+443 "Prairie Trillium" Release (9f7969398b90468440a501cf065295d9050bb776, 2022-02-16) for RHEL 8 Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/16.5.2 Safari/605.1.15.

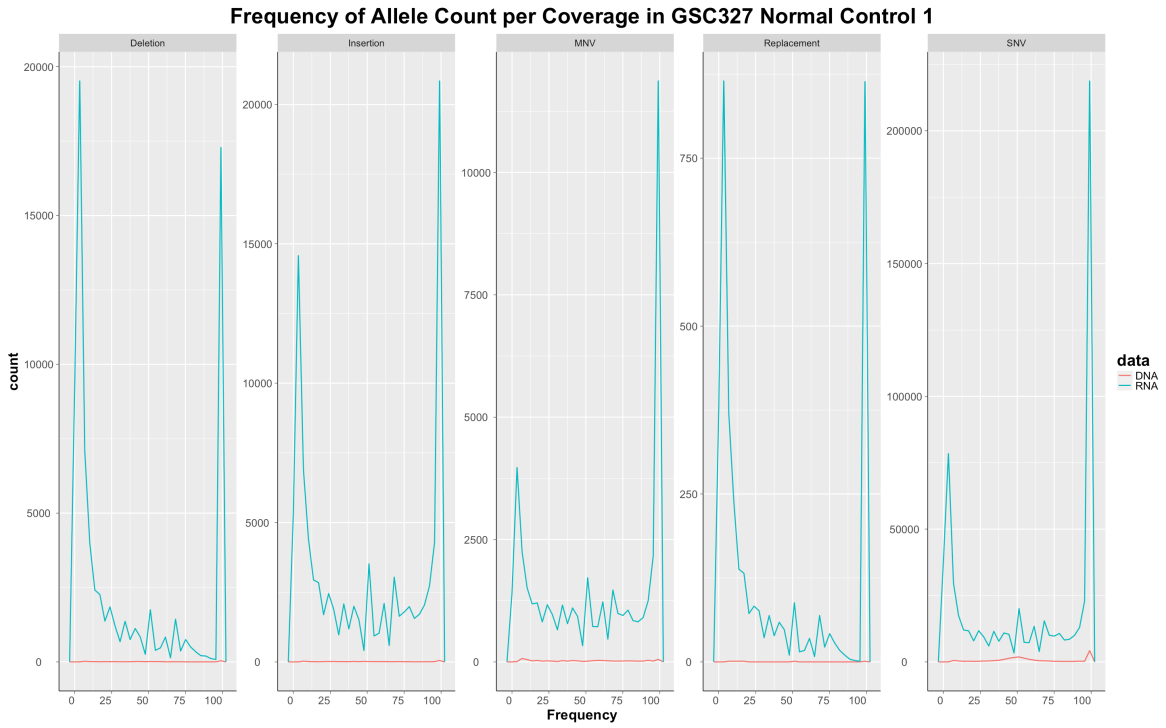
Full list of packages and their versions (or sessionInfo):

```
R version 4.1.2 (2021-11-01)
Platform: x86_64-redhat-linux-gnu (64-
   bit)
Running under: AlmaLinux 8.5 (Arctic
   Sphynx)

Matrix products: default
BLAS/LAPACK: /usr/lib64/libopenblas-
   r0.3.12.so

locale:
 [1] LC_CTYPE=en_GB.UTF-8
 [2] LC_NUMERIC=C
 [3] LC_TIME=en_GB.UTF-8
 [4] LC_COLLATE=en_GB.UTF-8
 [5] LC_MONETARY=en_GB.UTF-8
 [6] LC_MESSAGES=en_GB.UTF-8
 [7] LC_PAPER=en_GB.UTF-8
 [8] LC_NAME=C
 [9] LC_ADDRESS=C
[10] LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_GB.UTF-8
[12] LC_IDENTIFICATION=C

attached base packages:
[1] grid      stats4    stats
[4] graphics  grDevices utils
[7] datasets  methods   base

other attached packages:
 [1] fgsea_1.20.0
 [2] circlize_0.4.15
 [3] pheatmap_1.0.12
 [4] rlist_0.4.6.2
 [5] ggnewscale_0.4.8
 [6] ggpubr_0.6.0
 [7] ComplexHeatmap_2.10.0
 [8] IHW_1.22.0
 [9] BiocManager_1.30.20
[10] PCAtools_2.6.0
[11] RColorBrewer_1.1-3
[12] apeglm_1.16.0
```

```
[13] DT_0.26
[14] DESeq2_1.34.0
[15] SummarizedExperiment_1.24.0
[16] MatrixGenerics_1.6.0
[17] matrixStats_0.63.0
[18] GenomicRanges_1.46.1
[19] GenomeInfoDb_1.30.1
[20] org.Hs.eg.db_3.14.0
[21] AnnotationDbi_1.56.2
[22] IRanges_2.28.0
[23] S4Vectors_0.32.4
[24] Biobase_2.54.0
[25] BiocGenerics_0.40.0
[26] gplots_3.1.3
[27] rhdf5_2.38.1
[28] tximport_1.22.0
[29] readxl_1.4.2
[30] lubridate_1.9.2
[31] forcats_1.0.0
[32] stringr_1.5.0
[33] purrr_1.0.1
[34] readr_2.1.4
[35] tidyr_1.3.0
[36] tibble_3.2.1
[37] tidyverse_2.0.0
[38] dplyr_1.1.1
[39] biomaRt_2.50.3
[40] EnhancedVolcano_1.12.0
[41] ggrepel_0.9.2
[42] ggplot2_3.4.2

loaded via a namespace (and not
   attached):
  [1] utf8_1.2.2
  [2] tidyselect_1.2.0
  [3] RSQLite_2.2.20
  [4] htmlwidgets_1.6.1
  [5] BiocParallel_1.28.3
  [6] munsell_0.5.0
  [7] ScaledMatrix_1.2.0
  [8] codetools_0.2-18
  [9] withr_2.5.0
 [10] colorspace_2.0-3
 [11] filelock_1.0.2
 [12] ggalt_0.4.0
 [13] knitr_1.41
 [14] rstudioapi_0.14
 [15] ggsignif_0.6.4
 [16] Rttf2pt1_1.3.12
 [17] slam_0.1-50
 [18] bbmle_1.0.25
 [19] GenomeInfoDbData_1.2.7
 [20] lpsymphony_1.17.0
 [21] bit64_4.0.5
 [22] coda_0.19-4
 [23] vctrs_0.6.1
 [24] generics_0.1.3
 [25] xfun_0.38
 [26] timechange_0.1.1
 [27] BiocFileCache_2.2.1
 [28] R6_2.5.1
 [29] doParallel_1.0.17
 [30] ggbeeswarm_0.7.1
 [31] clue_0.3-63
 [32] rsvd_1.0.5
 [33] locfit_1.5-9.7
 [34] bitops_1.0-7
 [35] rhdf5filters_1.6.0
 [36] cachem_1.0.6
 [37] DelayedArray_0.20.0
 [38] scales_1.2.1
 [39] beeswarm_0.4.0
 [40] gtable_0.3.1
 [41] beachmat_2.10.0
 [42] ash_1.0-15
 [43] rlang_1.1.0
 [44] genefilter_1.76.0
 [45] GlobalOptions_0.1.2
 [46] splines_4.1.2
 [47] rstatix_0.7.2
 [48] extrafontdb_1.0
 [49] broom_1.0.4
 [50] yaml_2.3.6
 [51] reshape2_1.4.4
 [52] abind_1.4-5
 [53] backports_1.4.1
 [54] extrafont_0.19
 [55] tools_4.1.2
 [56] ellipsis_0.3.2
 [57] Rcpp_1.0.9
 [58] plyr_1.8.8
 [59] sparseMatrixStats_1.6.0
 [60] progress_1.2.2
 [61] zlibbioc_1.40.0
 [62] RCurl_1.98-1.9
 [63] prettyunits_1.1.1
 [64] GetoptLong_1.0.5
 [65] cowplot_1.1.1
 [66] cluster_2.1.4
 [67] magrittr_2.0.3
 [68] data.table_1.14.6
 [69] mvtnorm_1.1-3
 [70] evaluate_0.19
 [71] hms_1.1.2
 [72] xtable_1.8-4
 [73] XML_3.99-0.13
 [74] emdbook_1.3.12
 [75] gridExtra_2.3
 [76] shape_1.4.6
 [77] compiler_4.1.2
 [78] bdsmatrix_1.3-6
 [79] maps_3.4.1
 [80] KernSmooth_2.23-20
 [81] crayon_1.5.2
```

```
 [82] htmltools_0.5.4
 [83] tzdb_0.3.0
 [84] geneplotter_1.72.0
 [85] DBI_1.1.3
 [86] dbplyr_2.3.2
 [87] proj4_1.0-12
 [88] MASS_7.3-58.1
 [89] rappdirs_0.3.3
 [90] Matrix_1.5-3
 [91] car_3.1-2
 [92] cli_3.6.1
 [93] parallel_4.1.2
 [94] pkgconfig_2.0.3
 [95] numDeriv_2016.8-1.1
 [96] xml2_1.3.3
 [97] foreach_1.5.2
 [98] annotate_1.72.0
 [99] vipor_0.4.5
[100] dqrng_0.3.0
[101] XVector_0.34.0
[102] digest_0.6.31
[103] Biostrings_2.62.0
[104] rmarkdown_2.19
[105] fastmatch_1.1-3
[106] cellranger_1.1.0
[107] DelayedMatrixStats_1.16.0
[108] curl_4.3.3
[109] gtools_3.9.4
[110] rjson_0.2.21
[111] lifecycle_1.0.3
[112] Rhdf5lib_1.16.0
[113] carData_3.0-5
[114] fansi_1.0.3
[115] pillar_1.8.1
[116] lattice_0.20-45
[117] ggrastr_1.0.1
[118] KEGGREST_1.34.0
[119] fastmap_1.1.0
[120] httr_1.4.4
[121] survival_3.4-0
[122] glue_1.6.2
[123] fdrtool_1.2.17
[124] png_0.1-8
[125] iterators_1.0.14
[126] bit_4.0.5
[127] stringi_1.7.8
[128] blob_1.2.3
[129] BiocSingular_1.10.0
[130] caTools_1.18.2
[131] memoise_2.0.1
[132] irlba_2.3.5.1
```

Appendix 2 - Related to CHAPTER 3



Appendix Figure 1: Frequency of allele count per coverage in GSC327 normal control 1 for both DNA and RNA by variant type.

Mutated Genes (397 profiled samples)			
Gene	# Mut	#	Freq
PTEN	138	133	33.5%
TP53	149	125	31.5%
TTN	202	101	25.4%
EGFR	117	94	23.7%
MUC16	103	61	15.4%
FLG	70	53	13.4%
NF1	65	46	11.6%
RYR2	52	43	10.8%
PIK3R1	41	39	9.8%
SPTA1	45	38	9.6%
RB1	39	38	9.6%
PIK3CA	42	38	9.6%
ATRX	47	37	9.3%
SYNE1	41	34	8.6%
MUC17	40	29	7.3%
LRP2	41	29	7.3%
PCLO	35	28	7.1%
PKHD1	36	27	6.8%
HMCN1	39	27	6.8%
OBSCN	53	26	6.5%
COL6A3	35	26	6.5%
IDH1	25	25	6.3%
AHNAK2	36	25	6.3%
DNAH2	39	24	6.0%
DNAH5	36	24	6.0%
USH2A	34	22	5.5%
FAT2	26	22	5.5%
FLG2	32	22	5.5%
HYDIN	31	21	5.3%
CFAP47	22	21	5.3%
LAMA1	27	21	5.3%
APOB	29	21	5.3%
HRNR	25	21	5.3%
RELN	28	20	5.0%
AHNAK	34	20	5.0%
RIMS2	27	20	5.0%
RYR3	33	19	4.8%
KMT2C	27	19	4.8%
KEL	20	19	4.8%
GALNT17	19	19	4.8%
CNTNAP2	20	19	4.8%
HSPG2	22	18	4.5%
DNAH11	30	18	4.5%
DOCK5	22	18	4.5%
ADGRV1	31	18	4.5%
TCHH	21	18	4.5%
DNAH3	36	18	4.5%
DNAH8	26	18	4.5%
DNAH9	23	18	4.5%
SDK1	21	18	4.5%
STAG2	19	18	4.5%
MROH2B	21	17	4.3%
LZTR1	17	17	4.3%
MXRA5	22	17	4.3%
FAT4	28	17	4.3%
GRIN2A	21	16	4.0%
DSP	19	16	4.0%
KIF2B	18	16	4.0%
KDR	20	16	4.0%
MACF1	30	16	4.0%

Appendix Figure 2: Most frequently occurring mutated genes in Glioblastoma according to CBioPortal.

(A) **Functional Annotation Chart**

Help and Manual

Current Gene List: DNA SNV genes (unfiltered)
Current Background: Homo sapiens
4998 DAVID IDs

Options

Rerun Using Options | Create Sublist

68 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Olfactory transduction	RT		266	5.3	9.7E-72	3.4E-69
	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		147	2.9	2.7E-15	4.7E-13
	KEGG_PATHWAY	Taste transduction	RT		50	1.0	1.6E-12	1.8E-10
	KEGG_PATHWAY	Protein digestion and absorption	RT		49	1.0	2.2E-8	1.9E-6
	KEGG_PATHWAY	Staphylococcus aureus infection	RT		45	0.9	1.6E-7	1.0E-5
	KEGG_PATHWAY	Calcium signaling pathway	RT		92	1.8	1.8E-7	1.0E-5
	KEGG_PATHWAY	Complement and coagulation cascades	RT		41	0.8	3.5E-7	1.7E-5
	KEGG_PATHWAY	Bile secretion	RT		41	0.8	1.1E-6	4.7E-5
	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT		34	0.7	5.5E-6	2.1E-4
	KEGG_PATHWAY	ECM-receptor interaction	RT		39	0.8	8.8E-6	3.1E-4
	KEGG_PATHWAY	Nicotine addiction	RT		22	0.4	2.2E-5	7.0E-4
	KEGG_PATHWAY	Retinol metabolism	RT		31	0.6	3.6E-5	1.0E-3
	KEGG_PATHWAY	Ascorbate and aldarate metabolism	RT		18	0.4	3.7E-5	1.0E-3
	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		34	0.7	4.2E-5	1.1E-3
	KEGG_PATHWAY	Pentose and glucuronate interconversions	RT		20	0.4	5.0E-5	1.2E-3
	KEGG_PATHWAY	Steroid hormone biosynthesis	RT		28	0.6	1.1E-4	2.5E-3
	KEGG_PATHWAY	Hematopoietic cell lineage	RT		39	0.8	1.5E-4	3.0E-3
	KEGG_PATHWAY	Estrogen signaling pathway	RT		49	1.0	2.9E-4	5.6E-3
	KEGG_PATHWAY	Circadian entrainment	RT		37	0.7	4.7E-4	8.5E-3
	KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts	RT		28	0.6	1.2E-3	2.1E-2
	KEGG_PATHWAY	Inflammatory mediator regulation of TRP channels	RT		36	0.7	1.2E-3	2.1E-2
	KEGG_PATHWAY	GnRH secretion	RT		26	0.5	1.4E-3	2.3E-2
	KEGG_PATHWAY	GABAergic synapse	RT		33	0.7	1.7E-3	2.6E-2
	KEGG_PATHWAY	Salivary secretion	RT		34	0.7	1.9E-3	2.8E-2
	KEGG_PATHWAY	Tyrosine metabolism	RT		17	0.3	2.2E-3	3.0E-2
	KEGG_PATHWAY	Porphyrin metabolism	RT		19	0.4	2.6E-3	3.5E-2
	KEGG_PATHWAY	Pancreatic secretion	RT		36	0.7	2.7E-3	3.5E-2
	KEGG_PATHWAY	Cell adhesion molecules	RT		51	1.0	2.8E-3	3.6E-2
	KEGG_PATHWAY	Serotonergic synapse	RT		39	0.8	3.9E-3	4.6E-2
	KEGG_PATHWAY	Renin secretion	RT		26	0.5	4.7E-3	5.4E-2
	KEGG_PATHWAY	ABC transporters	RT		19	0.4	4.8E-3	5.4E-2
	KEGG_PATHWAY	Morphine addiction	RT		32	0.6	5.2E-3	5.5E-2
	KEGG_PATHWAY	Graft-versus-host disease	RT		18	0.4	5.2E-3	5.5E-2
	KEGG_PATHWAY	Mineral absorption	RT		23	0.5	6.6E-3	6.8E-2
	KEGG_PATHWAY	Oxytocin signaling pathway	RT		48	1.0	8.0E-3	7.8E-2
	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT		84	1.7	8.1E-3	7.8E-2
	KEGG_PATHWAY	cAMP signaling pathway	RT		66	1.3	8.3E-3	7.8E-2
	KEGG_PATHWAY	Aldosterone synthesis and secretion	RT		33	0.7	9.2E-3	8.5E-2
	KEGG_PATHWAY	Vascular smooth muscle contraction	RT		42	0.8	1.2E-2	1.1E-1
	KEGG_PATHWAY	Motor proteins	RT		57	1.1	1.2E-2	1.1E-1
	KEGG_PATHWAY	Long-term depression	RT		22	0.4	1.4E-2	1.2E-1
	KEGG_PATHWAY	Insulin secretion	RT		29	0.6	1.5E-2	1.2E-1
	KEGG_PATHWAY	Drug metabolism - other enzymes	RT		27	0.5	1.9E-2	1.5E-1
	KEGG_PATHWAY	Amphetamine addiction	RT		24	0.5	1.9E-2	1.5E-1

(B) **Functional Annotation Chart**

Help and Manual

Current Gene List: dna_snv_filtered
Current Background: Homo sapiens
4840 DAVID IDs

Options

Rerun Using Options | Create Sublist

71 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Olfactory transduction	RT		260	5.4	7.9E-71	2.7E-68
	KEGG_PATHWAY	Neuroactive ligand-receptor interaction	RT		146	3.0	2.5E-16	4.4E-14
	KEGG_PATHWAY	Taste transduction	RT		50	1.0	3.5E-13	4.1E-11
	KEGG_PATHWAY	Protein digestion and absorption	RT		48	1.0	2.1E-8	1.8E-6
	KEGG_PATHWAY	Staphylococcus aureus infection	RT		45	0.9	4.9E-8	3.4E-6
	KEGG_PATHWAY	Complement and coagulation cascades	RT		41	0.8	1.2E-7	6.3E-6
	KEGG_PATHWAY	Calcium signaling pathway	RT		90	1.9	1.3E-7	6.3E-6
	KEGG_PATHWAY	Bile secretion	RT		41	0.8	3.8E-7	1.7E-5
	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT		34	0.7	2.3E-6	8.9E-5
	KEGG_PATHWAY	ECM-receptor interaction	RT		39	0.8	3.5E-6	1.2E-4
	KEGG_PATHWAY	Nicotine addiction	RT		22	0.5	1.2E-5	3.8E-4
	KEGG_PATHWAY	Metabolism of xenobiotics by cytochrome P450	RT		34	0.7	1.9E-5	5.5E-4
	KEGG_PATHWAY	Ascorbate and aldarate metabolism	RT		18	0.4	2.2E-5	5.9E-4
	KEGG_PATHWAY	Pentose and glucuronate interconversions	RT		20	0.4	2.8E-5	7.1E-4
	KEGG_PATHWAY	Retinol metabolism	RT		30	0.6	5.0E-5	1.2E-3
	KEGG_PATHWAY	Hematopoietic cell lineage	RT		39	0.8	6.3E-5	1.4E-3
	KEGG_PATHWAY	Steroid hormone biosynthesis	RT		27	0.6	1.7E-4	3.5E-3
	KEGG_PATHWAY	Circadian entrainment	RT		37	0.8	2.2E-4	4.2E-3
	KEGG_PATHWAY	Estrogen signaling pathway	RT		48	1.0	2.3E-4	4.2E-3
	KEGG_PATHWAY	Inflammatory mediator regulation of TRP channels	RT		36	0.7	6.1E-4	1.1E-2
	KEGG_PATHWAY	Chemical carcinogenesis - DNA adducts	RT		28	0.6	6.4E-4	1.1E-2
	KEGG_PATHWAY	GnRH secretion	RT		26	0.5	8.1E-4	1.3E-2
	KEGG_PATHWAY	GABAergic synapse	RT		33	0.7	9.0E-4	1.4E-2
	KEGG_PATHWAY	Salivary secretion	RT		34	0.7	9.8E-4	1.4E-2
	KEGG_PATHWAY	Serotonergic synapse	RT		39	0.8	1.9E-3	2.7E-2
	KEGG_PATHWAY	Pancreatic secretion	RT		35	0.7	2.8E-3	3.6E-2
	KEGG_PATHWAY	Renin secretion	RT		26	0.5	2.8E-3	3.6E-2
	KEGG_PATHWAY	Morphine addiction	RT		32	0.7	2.9E-3	3.6E-2
	KEGG_PATHWAY	ABC transporters	RT		19	0.4	3.1E-3	3.7E-2
	KEGG_PATHWAY	Graft-versus-host disease	RT		18	0.4	3.5E-3	4.0E-2
	KEGG_PATHWAY	Oxytocin signaling pathway	RT		48	1.0	3.8E-3	4.2E-2
	KEGG_PATHWAY	Cell adhesion molecules	RT		49	1.0	3.8E-3	4.2E-2
	KEGG_PATHWAY	Tyrosine metabolism	RT		16	0.3	4.3E-3	4.4E-2
	KEGG_PATHWAY	Cytokine-cytokine receptor interaction	RT		83	1.7	4.3E-3	4.4E-2
	KEGG_PATHWAY	Porphyrin metabolism	RT		18	0.4	4.6E-3	4.6E-2
	KEGG_PATHWAY	Aldosterone synthesis and secretion	RT		33	0.7	5.2E-3	5.0E-2
	KEGG_PATHWAY	cAMP signaling pathway	RT		65	1.3	5.3E-3	5.0E-2
	KEGG_PATHWAY	Motor proteins	RT		57	1.2	5.7E-3	5.2E-2
	KEGG_PATHWAY	Insulin secretion	RT		29	0.6	8.9E-3	7.8E-2
	KEGG_PATHWAY	Long-term depression	RT		22	0.5	9.2E-3	7.8E-2
	KEGG_PATHWAY	Mineral absorption	RT		22	0.5	9.2E-3	7.8E-2
	KEGG_PATHWAY	Vascular smooth muscle contraction	RT		41	0.8	1.1E-2	8.9E-2
	KEGG_PATHWAY	Drug metabolism - other enzymes	RT		27	0.6	1.2E-2	9.5E-2
	KEGG_PATHWAY	Long-term potentiation	RT		23	0.5	1.7E-2	1.4E-1

(C) **Functional Annotation Chart**

Help and Manual

Current Gene List: RNA SNV genes (unfiltered)
Current Background: Homo sapiens
19696 DAVID IDs

Options

Rerun Using Options | Create Sublist

177 chart records — Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Herpes simplex virus 1 infection	RT		456	2.3	4.3E-21	1.5E-18
	KEGG_PATHWAY	Metabolic pathways	RT		1253	6.4	1.4E-18	2.5E-16
	KEGG_PATHWAY	Endocytosis	RT		234	1.2	2.2E-17	2.6E-15
	KEGG_PATHWAY	Salmonella infection	RT		229	1.2	1.8E-14	1.6E-12
	KEGG_PATHWAY	Autophagy - animal	RT		157	0.8	1.7E-13	1.1E-11
	KEGG_PATHWAY	Amyotrophic lateral sclerosis	RT		321	1.6	1.9E-13	1.1E-11
	KEGG_PATHWAY	Pathways of neurodegeneration - multiple diseases	RT		409	2.1	6.9E-13	3.5E-11
	KEGG_PATHWAY	Cell cycle	RT		148	0.8	8.7E-12	3.9E-10
	KEGG_PATHWAY	Huntington disease	RT		269	1.4	4.9E-11	1.9E-9
	KEGG_PATHWAY	Ubiquitin mediated proteolysis	RT		134	0.7	8.4E-11	3.0E-9
	KEGG_PATHWAY	Axon guidance	RT		167	0.8	1.7E-10	5.3E-9
	KEGG_PATHWAY	Protein processing in endoplasmic reticulum	RT		157	0.8	1.8E-10	5.3E-9
	KEGG_PATHWAY	Cellular senescence	RT		145	0.7	2.9E-10	8.0E-9
	KEGG_PATHWAY	Alzheimer disease	RT		327	1.7	1.8E-9	4.7E-8
	KEGG_PATHWAY	Colorectal cancer	RT		84	0.4	4.4E-9	1.0E-7
	KEGG_PATHWAY	ErbB signaling pathway	RT		83	0.4	5.9E-9	1.3E-7
	KEGG_PATHWAY	Neurotrophin signaling pathway	RT		112	0.6	6.5E-9	1.4E-7
	KEGG_PATHWAY	Polycomb repressive complex	RT		82	0.4	7.8E-9	1.6E-7
	KEGG_PATHWAY	Mitophagy - animal	RT		98	0.5	1.5E-8	2.7E-7
	KEGG_PATHWAY	Nucleocytoplasmic transport	RT		102	0.5	2.2E-8	3.9E-7
	KEGG_PATHWAY	Human papillomavirus infection	RT		281	1.4	5.6E-8	9.5E-7
	KEGG_PATHWAY	Bacterial invasion of epithelial cells	RT		75	0.4	6.0E-8	9.7E-7
	KEGG_PATHWAY	Focal adhesion	RT		179	0.9	7.8E-8	1.1E-6
	KEGG_PATHWAY	Chronic myeloid leukemia	RT		74	0.4	8.0E-8	1.1E-6
	KEGG_PATHWAY	Pancreatic cancer	RT		74	0.4	8.0E-8	1.1E-6
	KEGG_PATHWAY	Parkinson disease	RT		229	1.2	1.0E-7	1.3E-6
	KEGG_PATHWAY	Lysosome	RT		121	0.6	1.0E-7	1.3E-6
	KEGG_PATHWAY	Human T-cell leukemia virus 1 infection	RT		193	1.0	2.5E-7	3.2E-6
	KEGG_PATHWAY	Small cell lung cancer	RT		87	0.4	2.7E-7	3.3E-6
	KEGG_PATHWAY	Spinocerebellar ataxia	RT		129	0.7	3.7E-7	4.4E-6
	KEGG_PATHWAY	Renal cell carcinoma	RT		67	0.3	6.0E-7	6.8E-6
	KEGG_PATHWAY	mTOR signaling pathway	RT		139	0.7	7.1E-7	7.9E-6
	KEGG_PATHWAY	Inositol phosphate metabolism	RT		70	0.4	1.3E-6	1.4E-5
	KEGG_PATHWAY	Oocyte meiosis	RT		118	0.6	1.6E-6	1.6E-5
	KEGG_PATHWAY	MAPK signaling pathway	RT		253	1.3	1.6E-6	1.6E-5
	KEGG_PATHWAY	Nucleotide excision repair	RT		61	0.3	3.2E-6	3.2E-5
	KEGG_PATHWAY	Pathways in cancer	RT		429	2.2	3.5E-6	3.3E-5
	KEGG_PATHWAY	Pathogenic Escherichia coli infection	RT		171	0.9	3.5E-6	3.3E-5
	KEGG_PATHWAY	Yersinia infection	RT		122	0.6	4.2E-6	3.9E-5
	KEGG_PATHWAY	Hippo signaling pathway	RT		138	0.7	4.7E-6	4.2E-5
	KEGG_PATHWAY	Choline metabolism in cancer	RT		90	0.5	5.4E-6	4.7E-5
	KEGG_PATHWAY	Proteoglycans in cancer	RT		176	0.9	5.6E-6	4.7E-5
	KEGG_PATHWAY	Shigellosis	RT		209	1.1	5.8E-6	4.8E-5
	KEGG_PATHWAY	Phosphatidylinositol signaling system	RT		89	0.5	6.8E-6	5.5E-5

(D) **Functional Annotation Chart**

Help and Manual

Current Gene List: RNA SNV genes (filtered)
Current Background: Homo sapiens
18711 DAVID IDs

Options

Rerun Using Options | Create Sublist

177 chart records — Download File

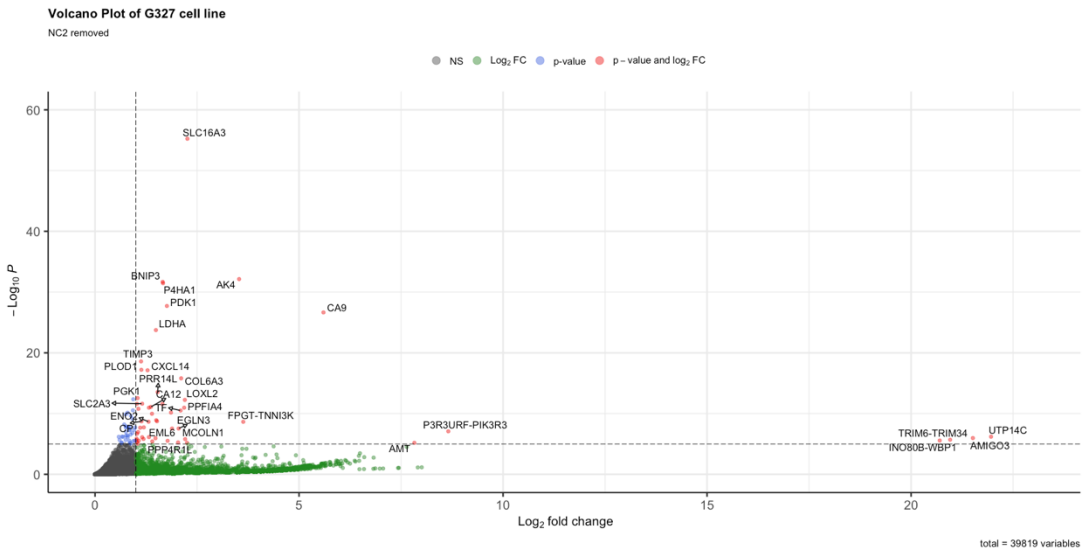
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Herpes simplex virus 1 infection	RT		440	2.4	4.8E-20	1.7E-17
	KEGG_PATHWAY	Metabolic pathways	RT		1200	6.4	1.5E-17	2.6E-15
	KEGG_PATHWAY	Endocytosis	RT		226	1.2	6.4E-16	7.6E-14
	KEGG_PATHWAY	Salmonella infection	RT		222	1.2	5.4E-14	4.8E-12
	KEGG_PATHWAY	Autophagy - animal	RT		150	0.8	3.3E-11	2.4E-9
	KEGG_PATHWAY	Pathways of neurodegeneration - multiple diseases	RT		390	2.1	4.4E-11	2.6E-9
	KEGG_PATHWAY	Ubiquitin mediated proteolysis	RT		131	0.7	5.5E-11	2.8E-9
	KEGG_PATHWAY	Cell cycle	RT		143	0.8	7.3E-11	3.2E-9
	KEGG_PATHWAY	Amyotrophic lateral sclerosis	RT		303	1.6	2.7E-10	1.1E-8
	KEGG_PATHWAY	Axon guidance	RT		160	0.9	3.3E-9	1.2E-7
	KEGG_PATHWAY	Neurotrophin signaling pathway	RT		109	0.6	9.0E-9	2.9E-7
	KEGG_PATHWAY	ErbB signaling pathway	RT		81	0.4	1.0E-8	3.0E-7
	KEGG_PATHWAY	Huntington disease	RT		254	1.4	1.6E-8	4.0E-7
	KEGG_PATHWAY	Focal adhesion	RT		175	0.9	1.6E-8	4.0E-7
	KEGG_PATHWAY	Bacterial invasion of epithelial cells	RT		74	0.4	1.9E-8	4.4E-7
	KEGG_PATHWAY	Nucleocytoplasmic transport	RT		99	0.5	4.6E-8	1.0E-6
	KEGG_PATHWAY	Cellular senescence	RT		137	0.7	6.3E-8	1.3E-6
	KEGG_PATHWAY	Human papillomavirus infection	RT		271	1.4	6.7E-8	1.3E-6
	KEGG_PATHWAY	Polycomb repressive complex	RT		79	0.4	8.8E-8	1.6E-6
	KEGG_PATHWAY	Chronic myeloid leukemia	RT		72	0.4	1.8E-7	3.1E-6
	KEGG_PATHWAY	Small cell lung cancer	RT		85	0.5	2.2E-7	3.7E-6
	KEGG_PATHWAY	Renal cell carcinoma	RT		66	0.4	2.4E-7	4.0E-6
	KEGG_PATHWAY	Colorectal cancer	RT		80	0.4	2.6E-7	4.0E-6
	KEGG_PATHWAY	Inositol phosphate metabolism	RT		69	0.4	4.4E-7	6.6E-6
	KEGG_PATHWAY	Alzheimer disease	RT		307	1.6	7.4E-7	1.1E-5
	KEGG_PATHWAY	Oocyte meiosis	RT		115	0.6	9.2E-7	1.3E-5
	KEGG_PATHWAY	Pancreatic cancer	RT		71	0.4	9.8E-7	1.3E-5
	KEGG_PATHWAY	Spinocerebellar ataxia	RT		124	0.7	1.4E-6	1.8E-5
	KEGG_PATHWAY	Pathways in cancer	RT		414	2.2	1.4E-6	1.8E-5
	KEGG_PATHWAY	Lysine degradation	RT		60	0.3	1.6E-6	1.9E-5
	KEGG_PATHWAY	Pathogenic Escherichia coli infection	RT		166	0.9	1.9E-6	2.2E-5
	KEGG_PATHWAY	Mitophagy - animal	RT		92	0.5	2.7E-6	3.0E-5
	KEGG_PATHWAY	Protein processing in endoplasmic reticulum	RT		144	0.8	3.2E-6	3.5E-5
	KEGG_PATHWAY	Thyroid hormone signaling pathway	RT		106	0.6	3.4E-6	3.5E-5
	KEGG_PATHWAY	Shigellosis	RT		202	1.1	4.5E-6	4.6E-5
	KEGG_PATHWAY	Parkinson disease	RT		216	1.2	5.4E-6	5.3E-5
	KEGG_PATHWAY	Hippo signaling pathway	RT		133	0.7	8.1E-6	7.8E-5
	KEGG_PATHWAY	Choline metabolism in cancer	RT		87	0.5	9.8E-6	9.1E-5
	KEGG_PATHWAY	mTOR signaling pathway	RT		132	0.7	1.0E-5	9.1E-5
	KEGG_PATHWAY	Adherens junction	RT		83	0.4	1.0E-5	9.1E-5
	KEGG_PATHWAY	Fatty acid metabolism	RT		54	0.3	1.1E-5	9.1E-5
	KEGG_PATHWAY	Human T-cell leukemia virus 1 infection	RT		182	1.0	1.1E-5	9.1E-5
	KEGG_PATHWAY	Progesterone-mediated oocyte maturation	RT		90	0.5	1.2E-5	9.6E-5
	KEGG_PATHWAY	Prostate cancer	RT		86	0.5	1.3E-5	9.9E-5

Appendix Figure 3: KEGG pathways of SNV mutation containing genes of both unfiltered and filtered DNA (A-B) and RNA (C-D) from the GSC327 NC1 variant analyses. KEGG pathway results were determined using DAVID. Full list can be found in Appendix Table 7

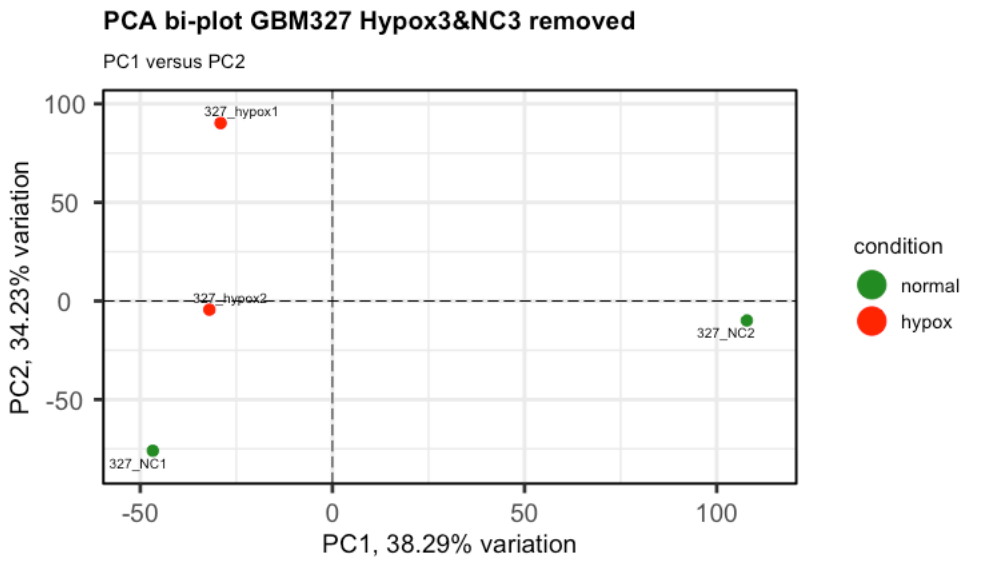
DNA ONLY	DNA & RNA	RNA ONLY		
Olfactory transduction	Calcium signaling pathway	Herpes simplex virus 1 infection	Efferocytosis	Proteasome
Neuroactive ligand-receptor interaction	ECM-receptor interaction	Metabolic pathways	Insulin signaling pathway	Breast cancer
Taste transduction	Inflammatory mediator regulation of TRP channels	Endocytosis	Longevity regulating pathway	Terpenoid backbone biosynthesis
Protein digestion and absorption	GnRH secretion	Salmonella infection	Human immunodeficiency virus 1 infection	Sphingolipid metabolism
Staphylococcus aureus infection	Oxytocin signaling pathway	Autophagy - animal	Hepatocellular carcinoma	Notch signaling pathway
Complement and coagulation cascades	Motor proteins	Pathways of neurodegeneration - multiple diseases	Peroxisome	Fluid shear stress and atherosclerosis
Bile secretion	Long-term depression	Ubiquitin mediated proteolysis	AGE-RAGE signaling pathway in diabetic complications	Central carbon metabolism in cancer
Drug metabolism - cytochrome P450	Long-term potentiation	Cell cycle	Viral life cycle - HIV-1	Bladder cancer
Nicotine addiction	Glutamatergic synapse	Amyotrophic lateral sclerosis	Nucleotide excision repair	Autophagy - other
Metabolism of xenobiotics by cytochrome P450	Gap junction	Axon guidance	Endometrial cancer	Fatty acid elongation
Ascorbate and aldarate metabolism	GnRH signaling pathway	Neurotrophin signaling pathway	Hepatitis B	Cushing syndrome
Pentose and glucuronate interconversions	Arrhythmogenic right ventricular cardiomyopathy	ErbB signaling pathway	Non-small cell lung cancer	Hedgehog signaling pathway
Retinol metabolism		Huntington disease	Proteoglycans in cancer	Kaposi sarcoma-associated herpesvirus infection
Hematopoietic cell lineage		Focal adhesion	Phospholipase D signaling pathway	Fatty acid degradation
Steroid hormone biosynthesis		Bacterial invasion of epithelial cells	Fc gamma R-mediated phagocytosis	Gastric cancer
Circadian entrainment		Nucleocytoplasmic transport	Non-alcoholic fatty liver disease	Melanogenesis
Estrogen signaling pathway		Cellular senescence	Epstein-Barr virus infection	Circadian rhythm
Chemical carcinogenesis - DNA adducts		Human papillomavirus infection	RNA degradation	Prolactin signaling pathway
GABAergic synapse		Polycomb repressive complex	Purine metabolism	One carbon pool by folate
Salivary secretion		Chronic myeloid leukemia	Rap1 signaling pathway	Glycosphingolipid biosynthesis - ganglio series
Serotonergic synapse		Small cell lung cancer	Pyrimidine metabolism	Hepatitis C
Pancreatic secretion		Renal cell carcinoma	Tight junction	2-Oxocarboxylic acid metabolism
Renin secretion		Colorectal cancer	Propanoate metabolism	Thyroid cancer
Morphine addiction		Inositol phosphate metabolism	Glioma	Biosynthesis of nucleotide sugars
ABC transporters		Alzheimer disease	Relaxin signaling pathway	Biosynthesis of cofactors
Cell adhesion molecules		Oocyte meiosis	Various types of N-glycan biosynthesis	Oxidative phosphorylation
Graft-versus-host disease		Pancreatic cancer	p53 signaling pathway	DNA replication
Cytokine-cytokine receptor interaction		Spinocerebellar ataxia	Carbon metabolism	Vasopressin-regulated water reabsorption
Tyrosine metabolism		Pathways in cancer	Human cytomegalovirus infection	Mismatch repair
Porphyrin metabolism		Lysine degradation	Thermogenesis	Mannose type O-glycan biosynthesis
Aldosterone synthesis and secretion		Pathogenic Escherichia coli infection	HIF-1 signaling pathway	Biosynthesis of unsaturated fatty acids
cAMP signaling pathway		Mitophagy - animal	Prion disease	Adipocytokine signaling pathway
Insulin secretion		Protein processing in endoplasmic reticulum	Hippo signaling pathway - multiple species	Other types of O-glycan biosynthesis
Vascular smooth muscle contraction		Thyroid hormone signaling pathway	Apoptosis	Apelin signaling pathway
Mineral absorption		Shigellosis	Insulin resistance	Glycerophospholipid metabolism
Drug metabolism - other enzymes		Parkinson disease	N-Glycan biosynthesis	PI3K-Akt signaling pathway
Primary immunodeficiency		Hippo signaling pathway	Longevity regulating pathway - multiple species	Selenocompound metabolism
Caffeine metabolism		Choline metabolism in cancer	SNARE interactions in vesicular transport	Fc epsilon RI signaling pathway
Platelet activation		mTOR signaling pathway	Platinum drug resistance	Adrenergic signaling in cardiomyocytes
Inflammatory bowel disease		Adherens junction	T cell receptor signaling pathway	Amino sugar and nucleotide sugar metabolism
Amphetamine addiction		Fatty acid metabolism	Chemical carcinogenesis - reactive oxygen species	Biosynthesis of amino acids
Carbohydrate digestion and absorption		Human T-cell leukemia virus 1 infection	Apoptosis - multiple species	C-type lectin receptor signaling pathway
Phototransduction		Progesterone-mediated oocyte maturation	Growth hormone synthesis, secretion and action	Glucagon signaling pathway
Ovarian steroidogenesis		Phosphatidylinositol signaling system	Vibrio cholerae infection	Cysteine and methionine metabolism
Antigen processing and presentation		Prostate cancer	Protein export	Parathyroid hormone synthesis, secretion and action
Amoebiasis		Regulation of actin cytoskeleton	Wnt signaling pathway	Pyruvate metabolism
Phagosome		Yersinia infection	Base excision repair	Signaling pathways regulating pluripotency of stem cells
Thyroid hormone synthesis		Lysosome	Basal transcription factors	
Viral protein interaction with cytokine and cytokine receptor		MAPK signaling pathway	Citrate cycle (TCA cycle)	
B cell receptor signaling pathway		Nucleotide metabolism	Ras signaling pathway	
Endocrine and other factor-regulated calcium reabsorption		Epithelial cell signaling in Helicobacter pylori infection	Dopaminergic synapse	
Arachidonic acid metabolism		Fanconi anemia pathway	Valine, leucine and isoleucine degradation	
Type II diabetes mellitus		AMPK signaling pathway	PD-L1 expression and PD-1 checkpoint pathway in cancer	
Vitamin digestion and absorption		Sphingolipid signaling pathway	Leukocyte transendothelial migration	
Mucin type O-glycan biosynthesis		Endocrine resistance	Toxoplasmosis	
Cortisol synthesis and secretion		EGFR tyrosine kinase inhibitor resistance	VEGF signaling pathway	
Cocaine addiction		mRNA surveillance pathway	Diabetic cardiomyopathy	
Autoimmune thyroid disease		Homologous recombination	TNF signaling pathway	
Folate biosynthesis		FoxO signaling pathway	Acute myeloid leukemia	

Appendix Table 7: Full list of KEGG pathway analysis

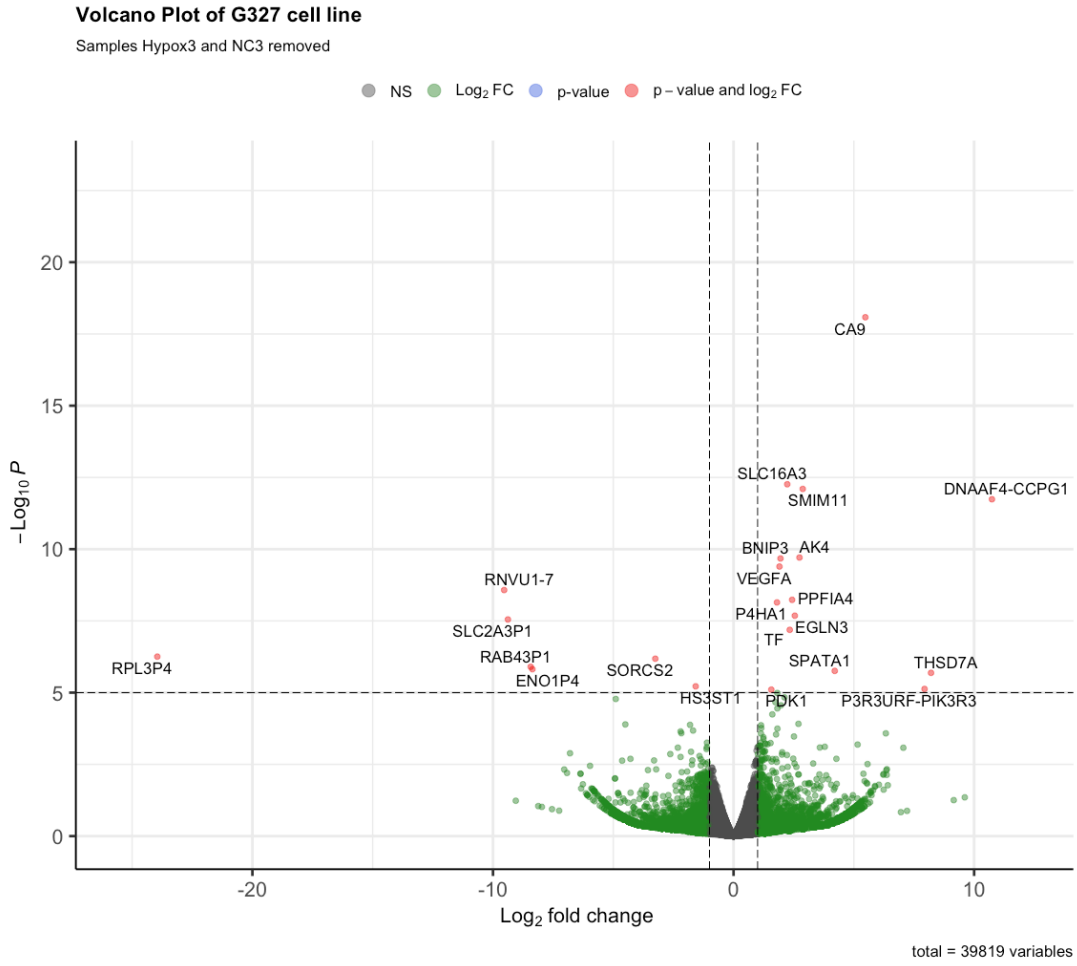
Appendix 3 - Related to CHAPTER 4



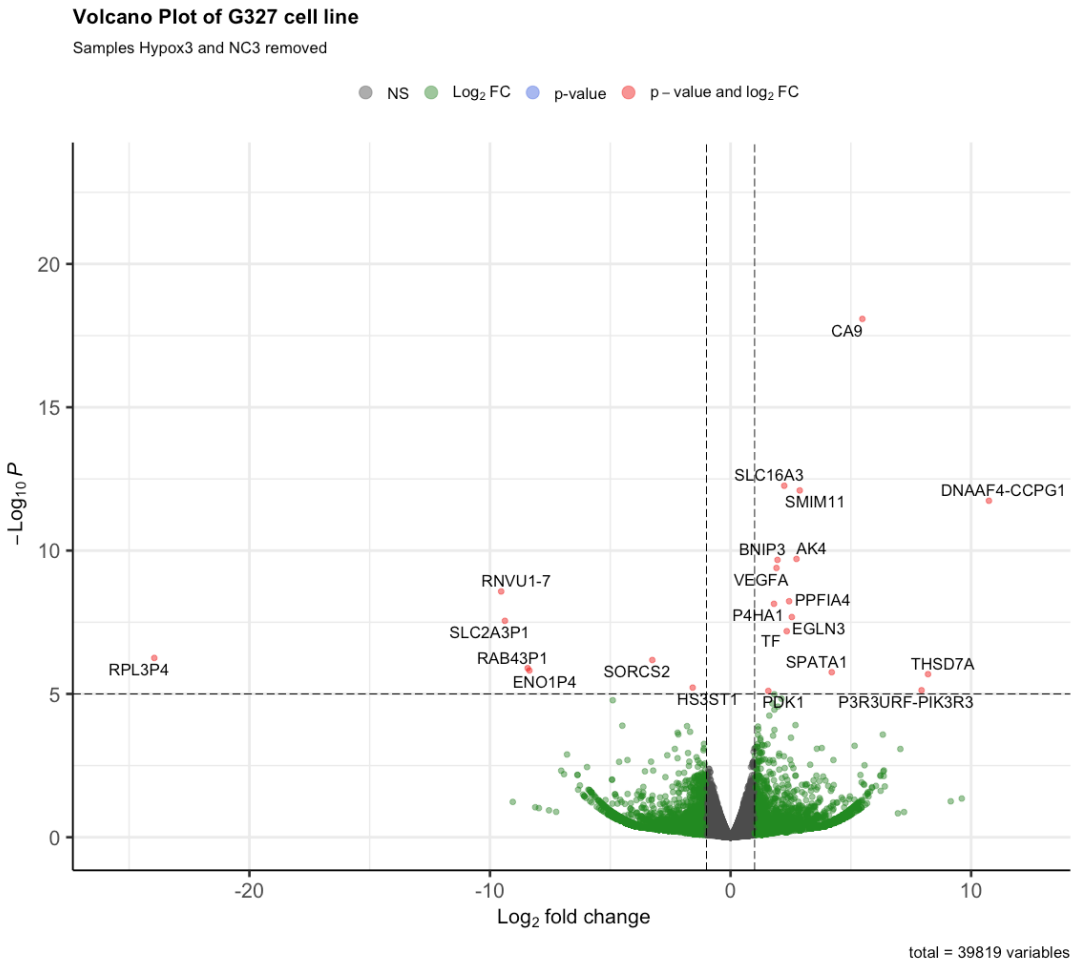
Appendix Figure 4: 327 Volcano plot, NC2 removed, overexpressed view



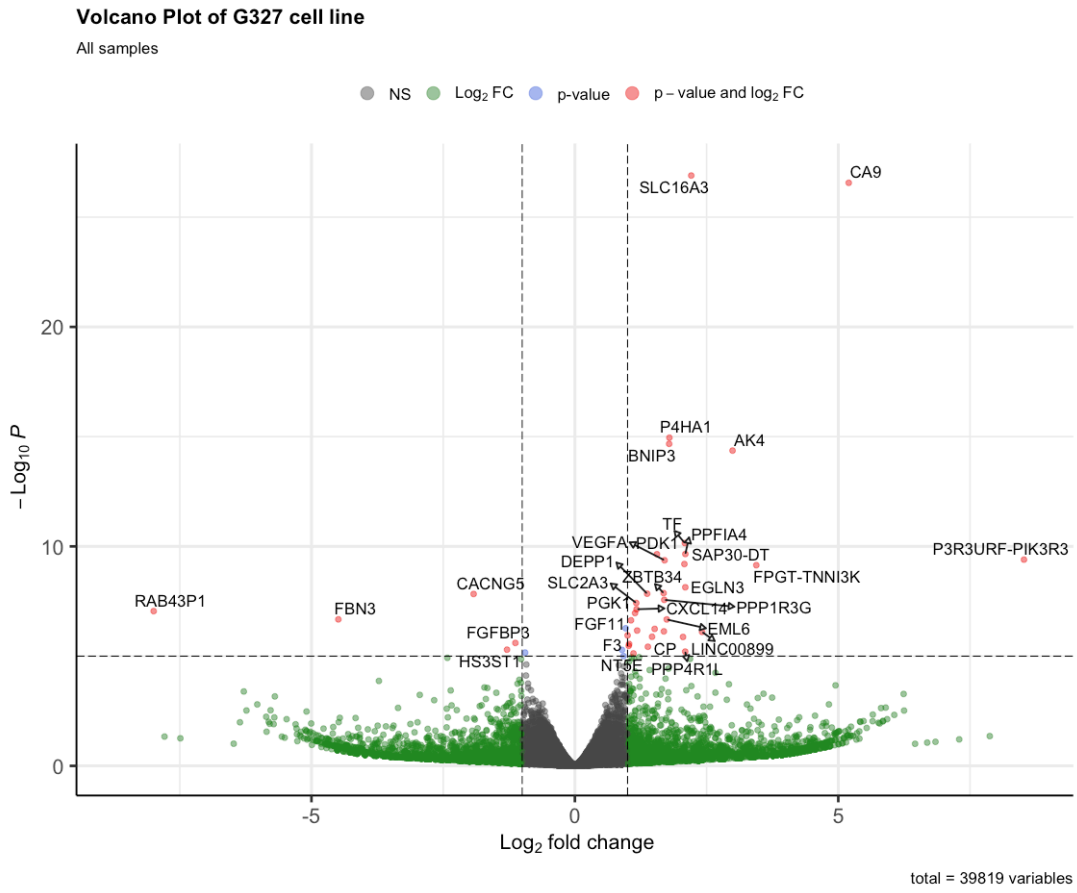
Appendix Figure 5: 327 PCA plot (Hypox3 & NC3 removed)



Appendix Figure 6: 327 Volcano plot Hypox3 & NC3 removed



Appendix Figure 7: 327 Volcano plot Hypox3 & NC3 removed



Appendix Figure 8: 327 Volcano plot (all samples)

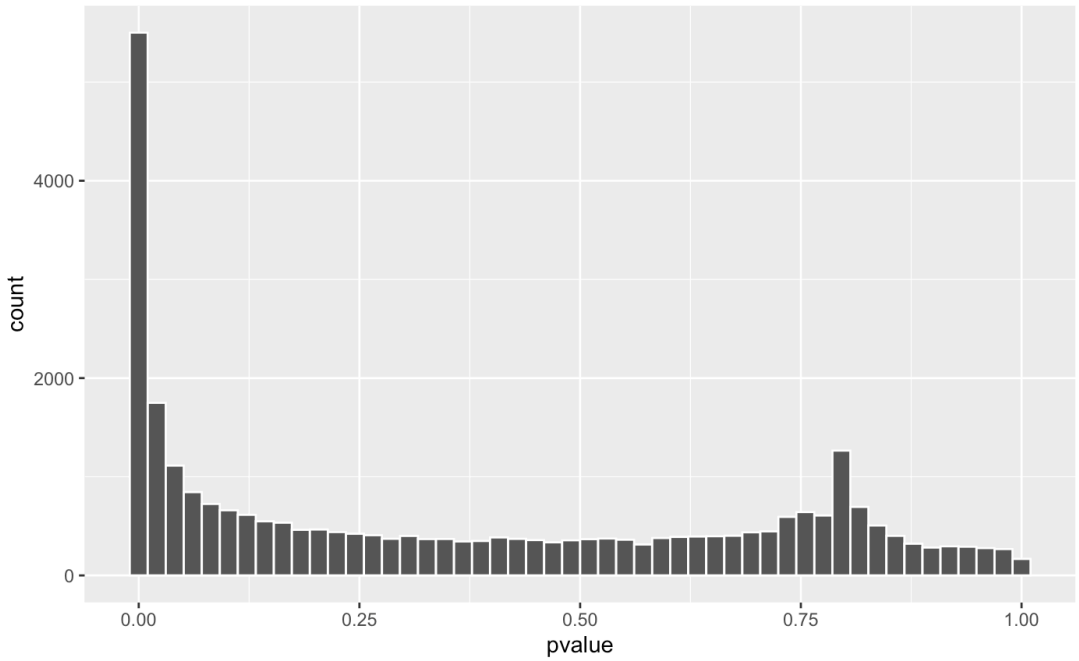
Appendix 4 - Related to CHAPTER 5

Patient ID	Stage
63	Primary
96	Primary
170	Primary
206	Primary
259	Primary, recurrence
289	Primary
292	Primary, recurrence
308	Primary, recurrence
309	Primary
320	Primary
376	Primary, recurrence
442	Primary
483	Primary, recurrence
499	Primary, recurrence
504	Primary
563	Primary
567	Primary
586	Primary
597	Primary, recurrence
601	Primary
602	Primary
607	Primary
622	Primary, recurrence
632	Primary

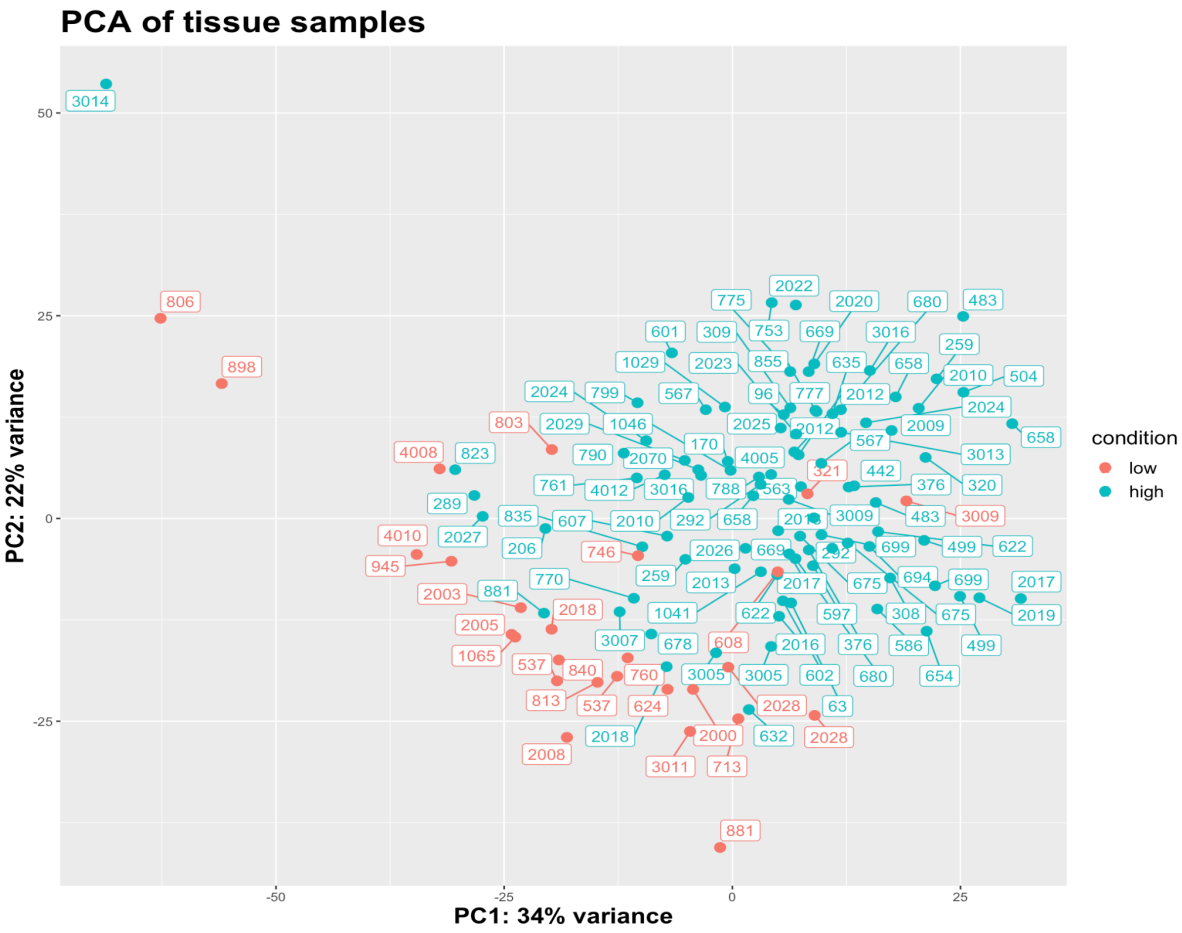
Patient ID	Stage
635	Primary
654	Primary
658	Primary, recurrence
669	Primary, recurrence
675	Primary, recurrence
678	Primary, recurrence
680	Primary, recurrence
694	Primary, recurrence
699	Primary, recurrence
753	Primary
761	Primary
770	Primary
775	Primary
777	Primary
788	Primary
790	Primary
799	Primary
823	Primary
835	Primary
855	Primary, recurrence
881	Primary, recurrence
1029	Primary
1041	Primary
1046	Primary

Patient ID	Stage
2009	Primary, recurrence
2010	Primary, recurrence
2012	Primary, recurrence
2013	Primary, recurrence
2016	Primary
2017	Primary, recurrence
2018	Primary, recurrence
2019	Primary, recurrence
2020	Primary
2022	Primary
2023	Primary
2024	Primary, recurrence
2025	Primary
2026	Primary
2027	Primary
2029	Primary
2070	Primary
3005	Primary, recurrence
3007	Primary
3009	Primary
3013	Primary
3016	Primary, recurrence
4005	Primary
4012	Primary, recurrence

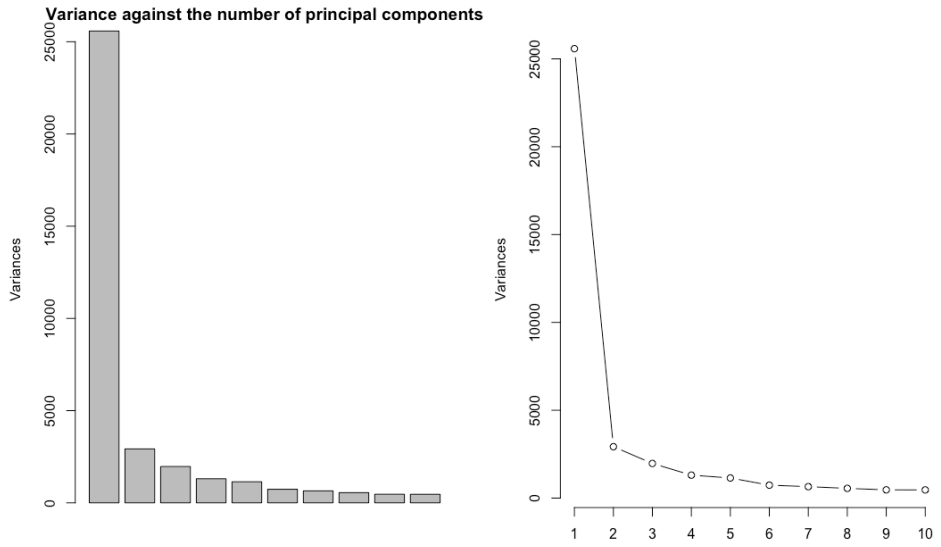
Appendix Table 8: Patient tissue sample types. Metadata of patient tissue samples whether there were both primary and recurrent samples taken and sequenced or just primary tissue samples were sequenced and used for this analysis.



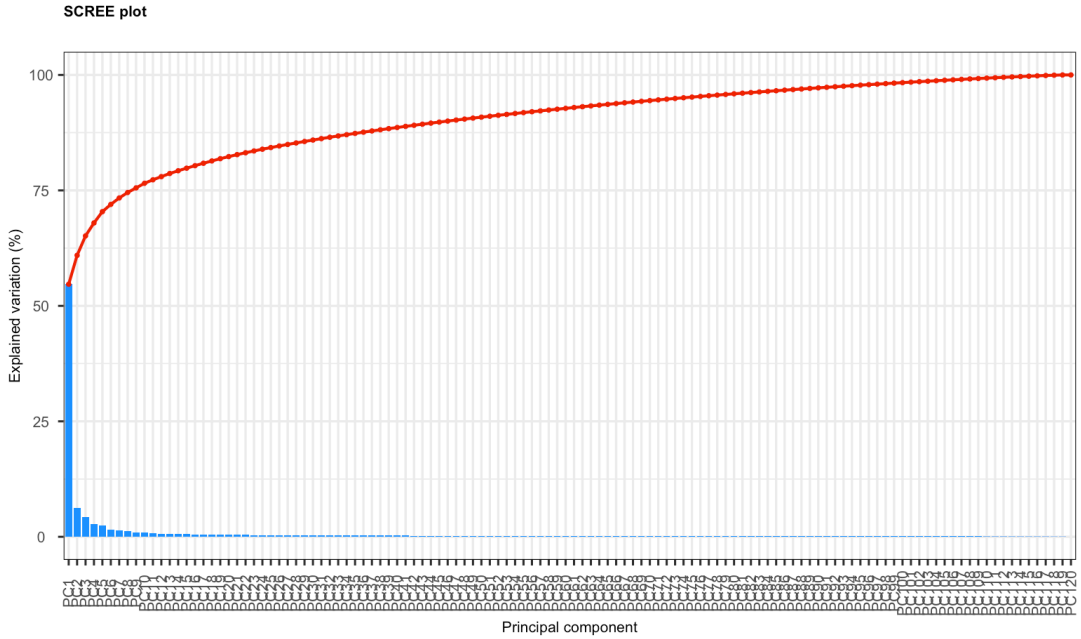
Appendix Figure 9: Histogram plot of p-values. This plot provides insight into the distribution of p-values reported by the differential gene expression analysis (DESeq function). Under the null hypothesis, besides a peak at p=0, a uniform distribution is expected. Here, there is an unexpected peak at 0.8. To address this, gene counts were filtered as genes with low counts can negatively influence the data. Any gene with less than 6 counts were filtered out.



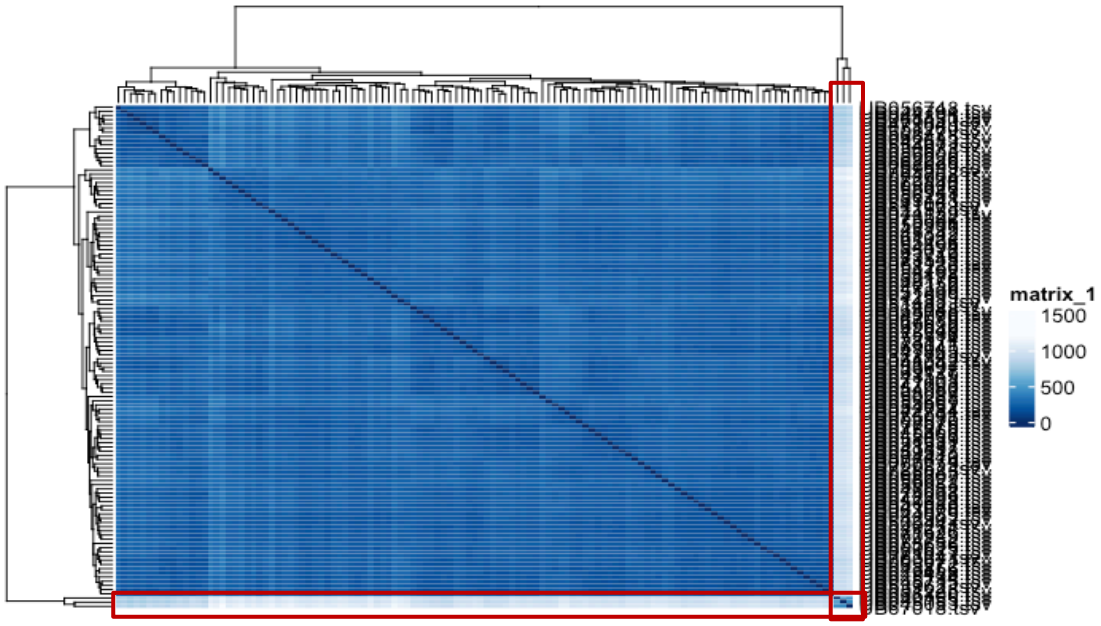
Appendix Figure 10: Initial Principal Component Analysis (PCA) of glioma tissue samples post-differential gene expression analysis from RNA sequencing data. Each point represents a sample in the dataset, color-coded according to their assigned condition (Low: grades 1-3, high: grade 4). This plot is part of the first round of exploratory data analysis plots, where we can observe that three samples (806, 898 and 3014) are located on the far-right hand side of the plot. The difference in variance between these samples and the rest of the samples is explained by the first principal component 54.41% (PC1), whereas PC2 only explains 6.25% of the data variation.



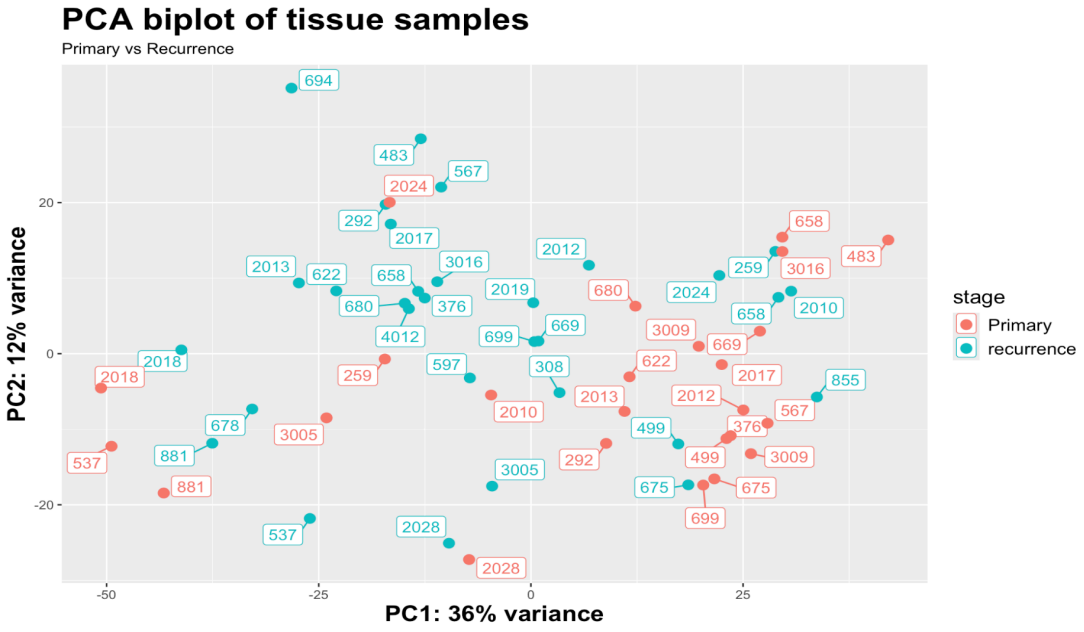
Appendix Figure 11: Initial Variance explained by Principal Components. This plot is a continuation of Appendix Figure 10 and it displays the cumulative variance explained by the principal components as a function of the number of components. The x-axis represents the number of principal components, while the y-axis indicates the proportion of total variance explained. The curve demonstrates how adding more principal components incrementally captures more of the dataset's variance, helping to determine the optimal number of components needed for efficient dimensionality reduction without significant loss of information. PC1 explains the majority of the variance in the data.



Appendix Figure 12: Initial SCREE plot of principal component. This scree plot (as a continuation of Appendix Figure 11) illustrates the eigenvalues associated with each principal component, plotted against the component number. The x-axis shows the principal component number, while the y-axis shows the eigenvalue magnitude. The plot reveals how much variance is explained by each principal component. Typically, a sharp decline (blue bars or inverse for the red line), or "elbow," indicates the optimal number of components to retain, as additional components contribute progressively less to the total variance. This helps in determining the number of principal components that should be considered for effective dimensionality reduction.



Appendix Figure 13: Initial Sample-to-sample correlation heatmap. This heatmap illustrates the correlation coefficients between pairs of all 120 tissue samples across the dataset. Each cell represents the correlation between two samples, with darker blue colours indicating higher correlations. Notably, three samples exhibit strong correlations with each other, as evidenced by the dark cluster in the bottom left corner (intersection of red boxes) but show no significant correlation with the remaining 117 samples. This distinct clustering suggests a high degree of similarity among these three samples, setting them apart from the rest of the dataset.



Appendix Figure 14: Principal Component Analysis (PCA) of glioma tissue samples post-differential gene expression analysis from RNA sequencing data. Each point represents a sample in the dataset, color-coded according to their assigned condition (Primary, recurrence). This plot is part of the first round of exploratory data analysis plots. The difference in variance between these samples and the rest of the samples is explained by the first principal component 36% (PC1), whereas PC2 only explains 12% of the data variation.

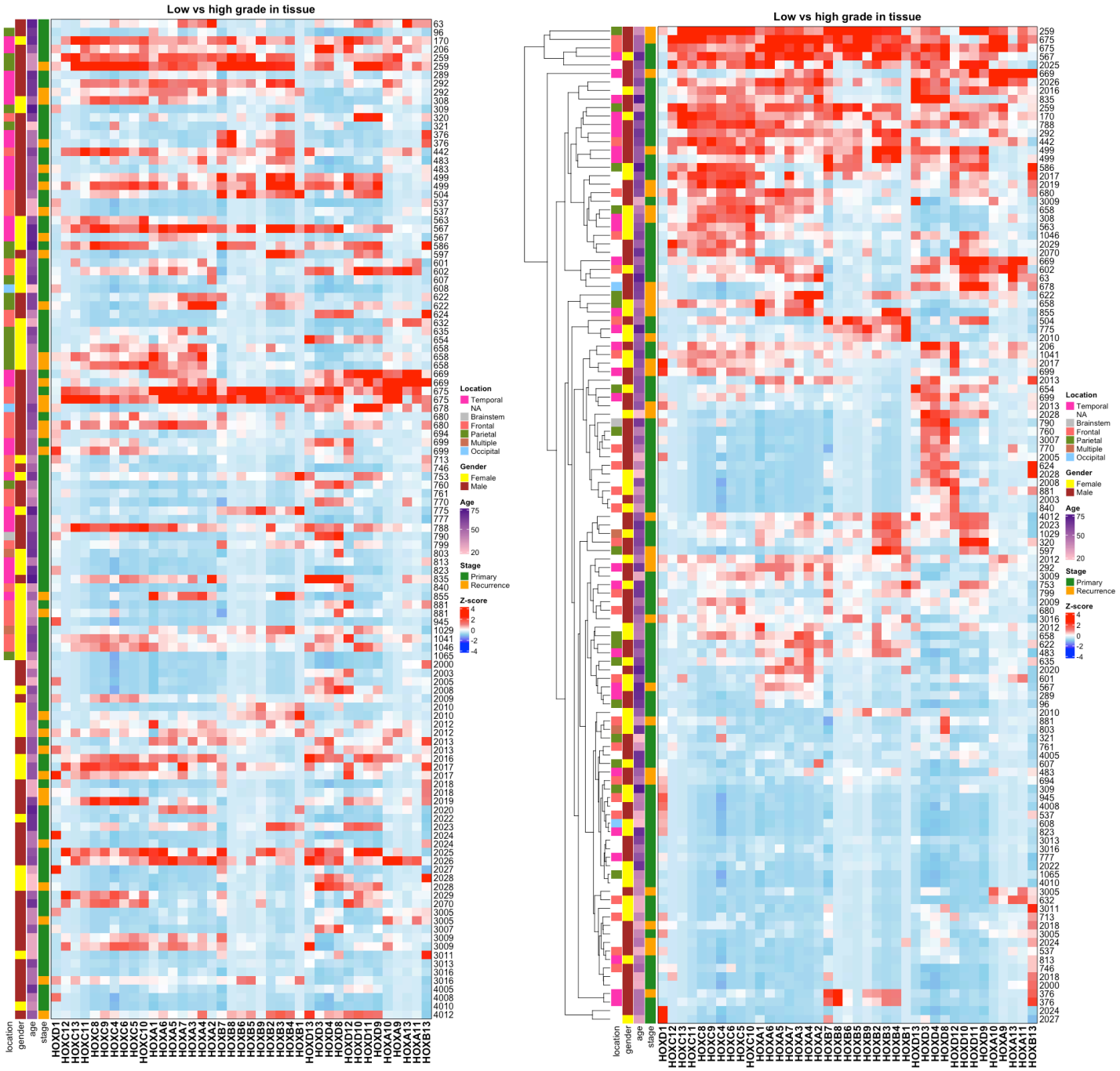
	HGNC symbol	GSC 322		Gene name	Tissue	
		log2 FC	p-value		log2 FC	p-value
1	AK4	4.51316085	2.78E-06	adenylate kinase 4	1.11663589	3.06E-11
2	PDK1	2.6566326	2.09E-30	pyruvate dehydrogenase kinase 1	1.28445251	3.79E-11
3	HK2	2.17566105	2.36E-19	hexokinase 2	1.47646083	3.26E-12
4	C4orf47	2.04755236	0.00373684	chromosome 4 open reading frame 47	1.23420344	2.72E-08
5	LDHA	2.03633883	1.87E-24	lactate dehydrogenase A	1.70374573	1.38E-25
6	NDRG1	1.72213077	2.27E-08	N-myc downstream regulated 1	1.02373283	1.71E-05
7	VEGFA	1.6943133	3.31E-22	vascular endothelial growth factor A	2.04391794	3.67E-10
8	SH3D21	1.50115755	6.25E-07	SH3 domain containing 21	1.09513863	4.55E-08
9	PGAM2	1.32556967	7.63E-07	phosphoglycerate mutase 2	1.26733254	0.00506644
10	FRZB	1.29365586	6.07E-20	frizzled related protein	1.24086878	2.49E-05
11	ZNF442	1.17424151	0.00049056	zinc finger protein 442	1.07005997	3.39E-09
12	COL9A3	1.0248528	7.34E-10	collagen type IX alpha 3 chain	1.16623393	0.00013277
13	BCAT1	1.01033198	0.00024036	branched chain amino acid transaminase 1	1.26031734	1.32E-08

Appendix Table 9: Full list of shared upregulated genes in the comparison between GSC322 and tissue genetic landscape. There were 13 shared genes with fold change (FC) higher than a threshold of 1, and p-value lower than 0.05.

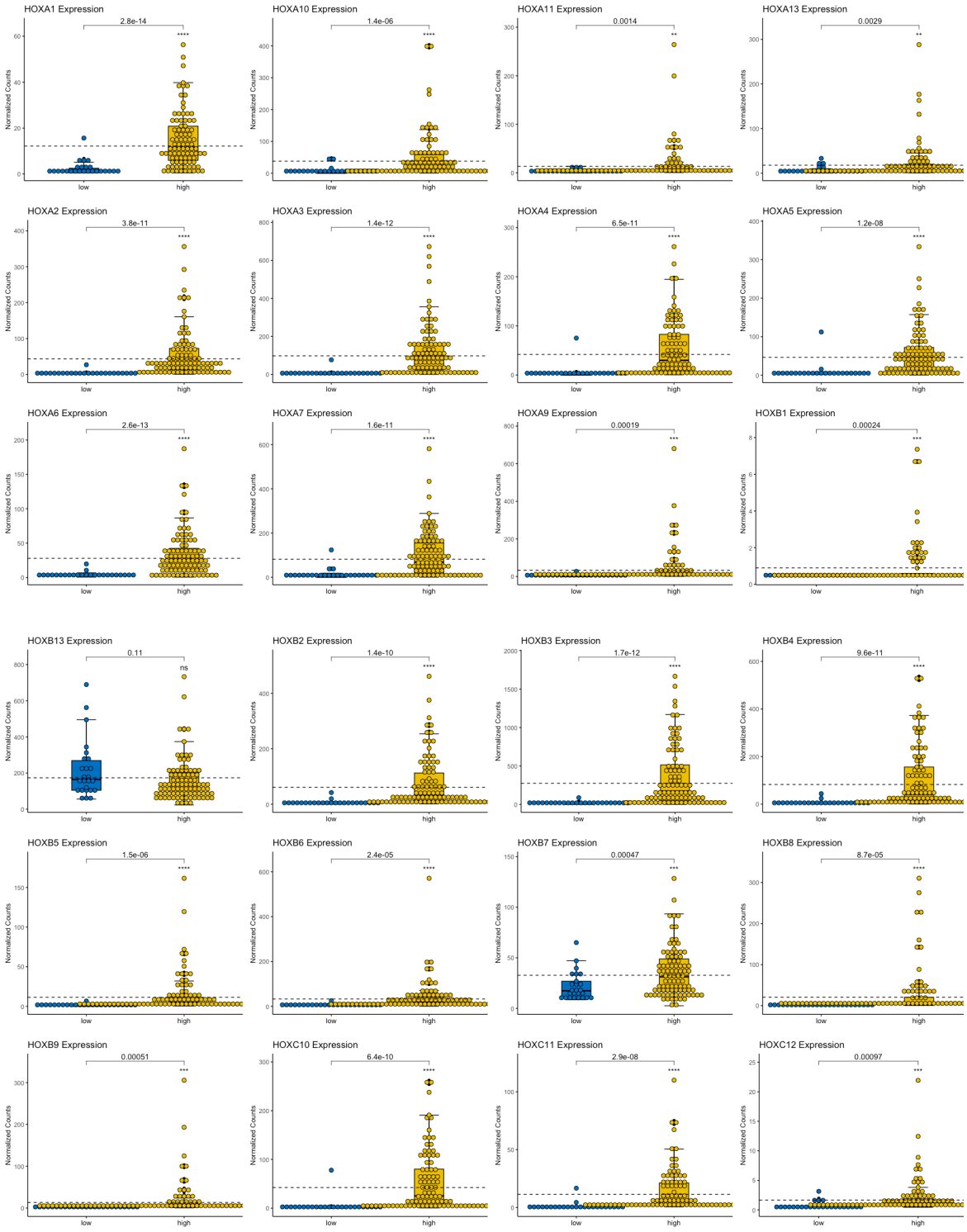
	HGNC Symbol	GSC327		Gene name	Tissue	
		log2 FC	p-value		log2 FC	p-value
1	LIMS4	7.63844367	3.20E-05	LIM zinc finger domain containing 4	1.2451889	0.00029443
2	C6orf118	5.89882776	0.02193402	chromosome 6 open reading frame 118	1.32735524	1.55E-06
3	CA9	5.59897167	3.11E-27	carbonic anhydrase 9	1.83202539	3.03E-06
4	LGR6	5.24861022	0.0352538	leucine rich repeat containing G protein-coupled receptor 6	1.13334679	0.00024206
5	EMILIN3	5.16086214	0.03759297	elastin microfibril interfacer 3	1.76901688	2.22E-05
6	LSP1	4.76630555	0.01373641	lymphocyte specific protein 1	1.10360864	4.86E-08
7	SP6	3.73661891	0.01108408	Sp6 transcription factor	1.67028988	4.92E-08
8	AK4	3.53189064	1.01E-32	adenylate kinase 4	1.11663589	3.06E-11
9	AK4P1	3.11067363	0.00571836	Adenylate Kinase 4 Pseudogene 1	1.10299974	0.00113493
10	FGFBP2	2.63463126	0.00710845	fibroblast growth factor binding protein 2	1.74292342	0.00020524
11	MT1A	2.53942848	0.04457652	metallothionein 1A	1.78209692	5.14E-05
12	COL6A3	2.11085285	1.64E-16	collagen type VI alpha 3 chain	2.25507545	2.32E-08
13	COL1A1	2.10536615	0.00160115	collagen type I alpha 1 chain	4.10907949	8.02E-21
14	AOX1	2.07498314	0.00663375	aldehyde oxidase 1	1.17574769	2.06E-06
15	PDK1	1.76201454	1.08E-28	pyruvate dehydrogenase kinase 1	1.28445251	3.79E-11
16	WIPF3	1.67899023	0.01160027	WAS/WASL interacting protein family member 3	1.31324904	0.00057779
17	VEGFA	1.65208983	2.83E-12	vascular endothelial growth factor A	2.04391794	3.67E-10
18	LENG9	1.56348976	0.02480653	leukocyte receptor cluster member 9	1.07507116	3.42E-10
19	LDHA	1.49285315	1.00E-24	lactate dehydrogenase A	1.70374573	1.38E-25
20	CCN1	1.48100052	0.03695574	cellular communication network factor 1	1.38090901	3.44E-06
21	SCNN1B	1.43969866	3.08E-05	sodium channel epithelial 1 subunit beta	1.80443285	3.46E-05
22	MAOB	1.40717558	0.04152025	monoamine oxidase B	1.33097313	1.05E-07
23	CA12	1.3692853	7.39E-12	carbonic anhydrase 12	1.70761051	8.50E-09
24	NDRG1	1.36586023	0.00417575	N-myc downstream regulated 1	1.02373283	1.71E-05
25	LOX	1.3255241	0.00039617	lysyl oxidase	2.91487384	1.02E-17
26	HSPG2	1.32176633	1.03E-11	heparan sulfate proteoglycan 2	1.24318392	4.86E-08
27	CXCL14	1.28735103	4.36E-18	C-X-C motif chemokine ligand 14	1.27904806	2.14E-05
28	ANG	1.24976828	0.03777488	angiogenin	1.62733276	5.33E-08
29	SPAG4	1.23461293	0.00022334	sperm associated antigen 4	1.28804618	8.64E-10
30	FMOD	1.18859142	1.50E-06	fibromodulin	2.26072604	3.83E-13
31	SLC2A3	1.15610119	2.02E-12	solute carrier family 2 member 3	1.1036304	1.04E-08
32	COL8A2	1.14353191	9.84E-05	collagen type VIII alpha 2 chain	1.39364466	6.09E-09
33	CP	1.14065269	1.73E-09	ceruloplasmin	1.18901349	1.16E-05
34	CHI3L2	1.08450166	0.01029586	chitinase 3 like 2	2.47951221	7.45E-17
35	TYMS	1.06404613	0.00113281	thymidylate synthetase	1.27034038	4.65E-12
36	MXRA5	1.04208168	0.00572401	matrix remodeling associated 5	2.2949873	1.38E-09
37	CD68	1.02163707	0.00805566	CD68 molecule	1.13268304	3.08E-06
38	ADM	1.01939196	1.27E-06	adrenomedullin	2.6150725	7.66E-16
39	HILPDA	1.01619565	2.21E-06	hypoxia inducible lipid droplet associated	1.76755573	5.99E-14

Appendix Table 10: Full list of shared upregulated genes in the comparison between GSC327 and tissue genetic landscape. There were 39 shared genes with fold change (FC) higher than a threshold of 1, and p-value lower than 0.05.

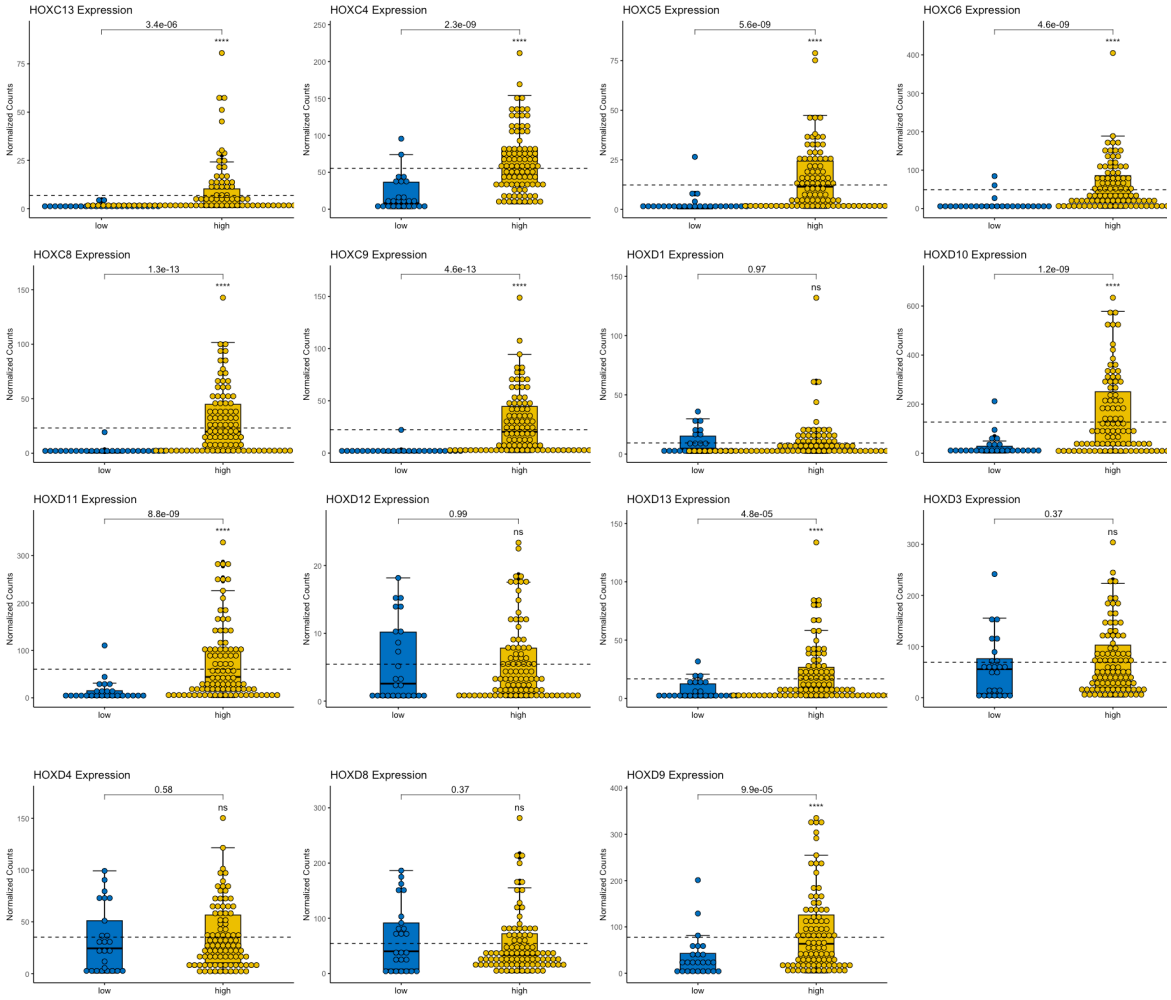
Appendix 5 - Related to CHAPTER 6



Appendix Figure 15: Differential Expression of HOX genes in low vs high grade gliomas
A) Heatmap showing the differential expression of HOX genes across a cohort of 117 glioma patient samples, organized by patient ID. The colour gradient represents the Z-score of gene expression, with red indicating upregulation and blue indicating downregulation. Patient metadate is provided on the left, including location of sample taken, patient gender, age, and whether the tumour was primary or recurrent. B) Hierarchically clustered heatmap of the same dataset, which groups patients and HOX genes based on similarities in expression patterns. This clustering reveals distinct expression patterns and relationships among HOX genes across different patients, emphasizing the variability in expression between different genes, where clusters can be observed.

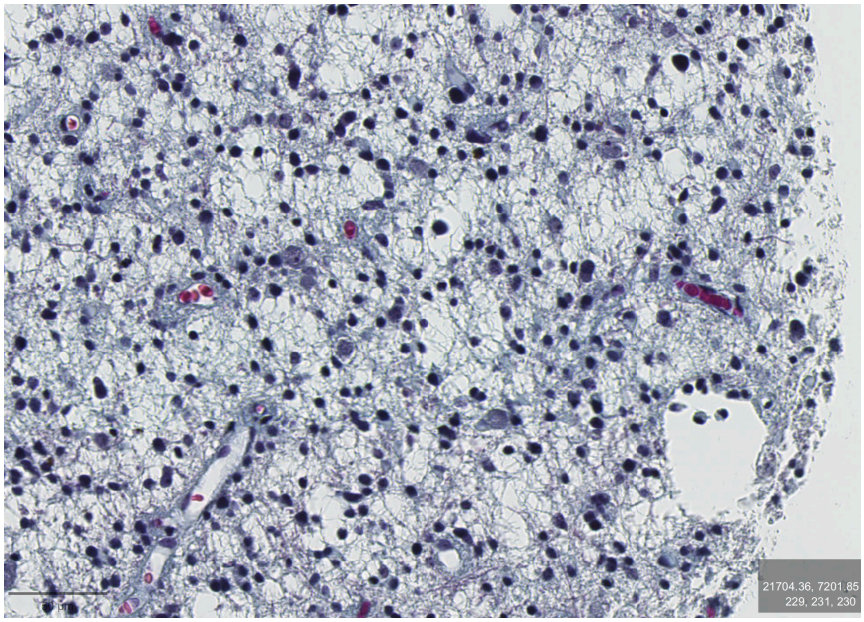


Appendix Figure 16: ANOVA of all HOX genes in low vs high grade tissue.

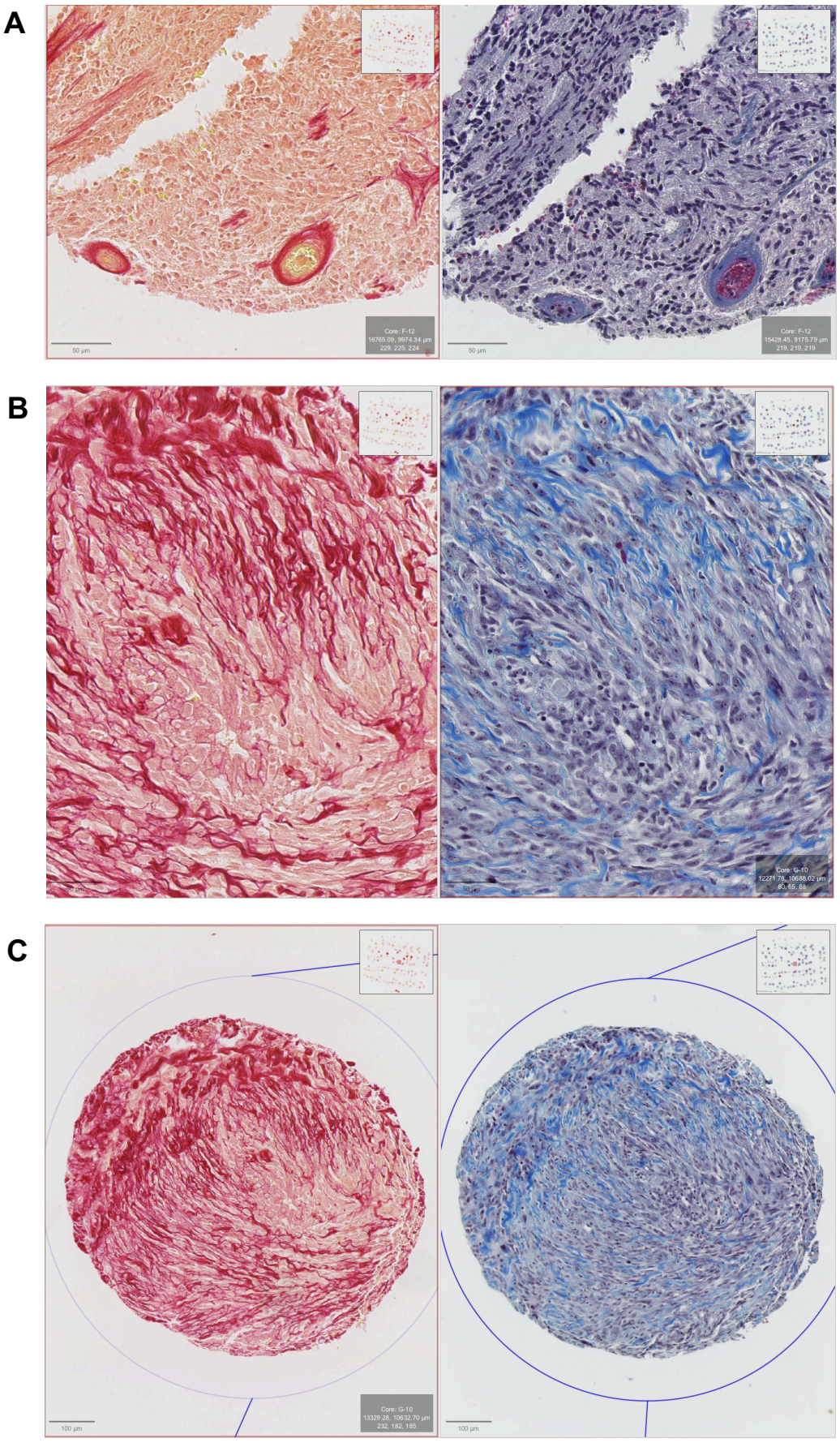


Appendix Figure 17: ANOVA of all HOX genes in low vs high grade tissue. Continuation of Appendix Figure 16.

Appendix 6 - Related to CHAPTER 7

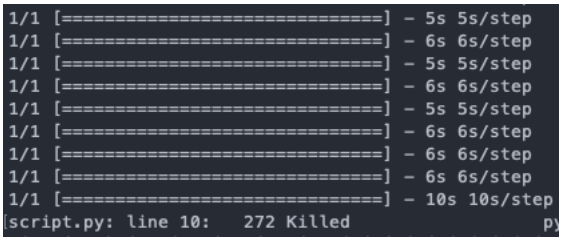


Appendix Figure 18: Collagen TMA staining Core A2
10 nanotubes forming angiogenic islands



Appendix Figure 19: H&E and collagen TMA staining of two cores
(A) Core F12 shows collagen localization around the blood vessels.
(B-C) Core G10 shows diffuse staining of collagen. This was only observed in very few cores.

Appendix 7 - Related to CHAPTER 8



Appendix Figure 20: Example error of miTAR algorithm when RAM is too small.

Build image from Dockerfile and copy the image ID.

```
1 > docker image build .

2 > docker image ls

3 REPOSITORY    TAG        IMAGE ID       CREATED          SIZE

4 <none>        <none>     0351a16f3299   18 seconds ago   2.31GB
```

Appendix Figure 21: Building docker image from Dockerfile
Line 1: Build the docker image in the current directory.
Line 2: Docker to list the locally available container images.
Line 3: Print the resulting headers (Repository – the name of the image; Tag – the version or variant of the image; Image ID – a unique identifier to represent the image; Created – the time elapsed since the image was created; Size – the storage size of the image)
Line 4: Answers to the requested list in line 2.

Using the image ID, run the container:

```
1 > docker container run -it 0351a16f3299 bash

2 (env) root@65c60c7b45f9:/#
```

Appendix Figure 22: Initializing docker container
Line 1: "docker container run" – Creates and starts a new container from a specified image. "-it" These flags enable interactive mode. "0351a16f3299" – the image ID from which the container is created. "bash" - the command executed inside the container. In this case, it starts a Bash shell, giving the user direct control over the container's environment.
Line 2: Starting prompt of the interactive session of the container.

Initialize docker container and navigate to the folder with the data and create a bash script to run the miRNA prediction algorithm: miTAR.

```
1 #!/bin/sh
2
3 #Job
4 python  predict_multimiRmultimRNA.py \
5         -i1 mydata/genes_of_interest.fa \
6         -i2 mydata/plant_csv/mirnas_filtered.fa \
7         -o myresults/lab_7_genes_edible_mirna.fa \
8         -s 22 \
9         -p 1 \
10        -ns 1
```

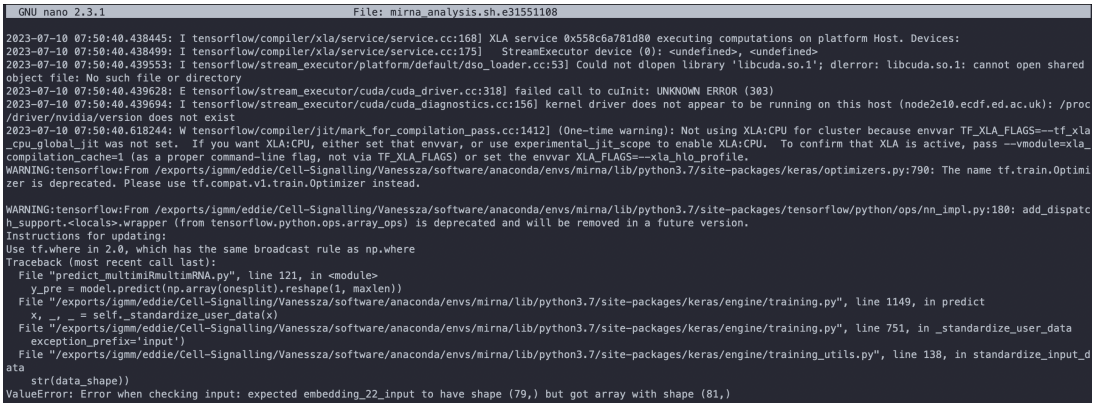
Appendix Figure 23: Example code to run miTAR
Line 4: Run python script
Line5: Input one: 3'UTR sequences of chosen mRNAs in FASTA format
Line 6: Input 2: miRNA sequences of interest in FASTA format
Line 7: Output file name and location
Line 8: "-s" – step; the length between each two fragments
Line 9: "-p" probability
Line 10: "-ns" – number of target sites per gene

Common upregulated in GSC322	AK4, ALDOC, ANKRD20A1, ATP10B, ATP1A2, BNIP3, BNIP3L, C4orf47, CRYAB, EGLN3, ENO2, FAM162A, FLNB, FRZB, GALNT16, KCNJ9, KLF11, L1CAM, LDHA, LRRC51, MAGEE2, MCHR1, MGAT3, MT1G, NAT8L, NDRG1, P4HA1, PDK1, PLN, PLP1, PPP1R3G, Prss53, SCRG1, SELL, SH3D21, SLC16A3, TEKT3, VEGFA, VWA3A
Common upregulated GSC327	ADM, ARMCX4, BNIP3, CA12, CAVIN2, CCN2, COL6A3, COL8A2, CXCL14, DEPP1, DOC2A, EGLN3, ENO2, FMOD, FRMPD4, HAPLN1, HILPDA, HSF4, IFITM1, LDHA, LOXL2, NDRG1, NEK9, NT5E, P3H2, P4HA1, PADI2, PDGFB, PGK1, PPP1R3C, RASSF7, SCNN1B, SLC16A3, PAG4, TEK, SSC4D, TIMP3, TMEM130, VAMP1, VEGFA
GSC322 top 8 genes	ATP1A2, BNIP3, LDHA, LYN, PDK1, PGK1, PLP1, VWA3A

Appendix Table 11: Full list of tested genes for miRNA binding prediction with miTAR

```
>MeP-ach-miR1|ENSG00000180818|HOXC10|protein_coding|5|ENST00000303460.5|ENST00000303460|31|1|0.9999796
CCCCUUACUUCGGACCAGGNNNNNNNUUCCCGCUCCUUCCUCUCCCCGCCCCUCCUCCCUUUGUGCCUGGUGAUAUAUU
```

Appendix Figure 24: Example output of miTAR
Line 1: Name of miRNA, Ensembl ID of gene, HGNC name of gene, type of gene, version number, transcript version, transcript number, number of pairs identified, instance of pairing, probability
Line 2: miRNA sequence – separated by Ns – mRNA sequence



Appendix Figure 25: miTAR algorithm error on Eddie due to suspected update of Tensorflow